

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE MATEMÁTICA

Estudio de las Funciones de Pérdida L_p en Redes Neuronales Profundas

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL
MATEMÁTICO

MARTÍN CONTRERAS PACHECO

PROFESOR GUÍA:
Ronny Vallejos

MIEMBROS DE LA COMISIÓN:
Nixon Jerez
Felipe Osorio



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: ___ Estudio de las Funciones de Pérdida Lp en Redes Neuronales Profundas ___

Nombre del candidato(a): ___ Martin Contreras Pacheco ___

Carrera / Grado: ___ Ingeniería Civil Matemática ___

Campus: ___ Casa Central ___ **Departamento:** ___ Matemática ___

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, ___ Ronny Vallejos ___, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por (**marcar una opción**):

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

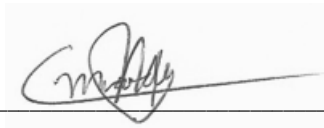
Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis: Ronny Vallejos

Fecha: ___ 14/10/2026 ___ **Firma:** 

Estudiante o Candidato(a):

Fecha: ___ 13/10/2026 ___ **Firma:** 

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Para mi familia y mi pareja, por su apoyo, cariño y
acompañamiento a lo largo de este proceso, y por ser parte
fundamental del camino que me ha llevado hasta este momento.

Agradecimientos

A mis padres, quienes me han apoyado desde que decidí estudiar esta carrera. Desde el primer momento confiaron en que sería capaz de lograrlo, incluso cuando yo mismo llegué a dudarlo. Su confianza, apoyo y preocupación constante fueron fundamentales para seguir adelante durante los momentos más difíciles y, finalmente, llegar hasta este punto.

A mi pareja, a quien conocí durante este proceso y sin cuyo apoyo este camino habría sido mucho más difícil. Gracias por brindarme comprensión y apoyo en los momentos en que más lo necesité, por escucharme cuando las cosas no salían como esperaba y por motivarme día a día a seguir adelante para alcanzar este objetivo. Su compañía durante esta última etapa fue especialmente importante y una parte fundamental de este proceso.

A mis compañeros, con quienes nos apoyamos mutuamente a lo largo de todo este proceso. Si bien nos conocimos recién en tercer año, desde entonces fuimos avanzando juntos durante el resto de la carrera. Compartimos largas jornadas de estudio, trabajos, preocupaciones y momentos de estrés, pero también muchas experiencias y buenos momentos. Haber podido recorrer esta etapa junto a ellos hizo más llevaderos los momentos difíciles.

Por último, agradezco al profesor Ronny Vallejos por aceptarme como su alumno, por su tutela y guía durante el desarrollo de esta memoria, y especialmente por su paciencia y disposición a lo largo de todo este proceso. Sus consejos y orientación fueron fundamentales para llevar adelante este trabajo y alcanzar su término.

Resumen

Esta memoria estudia la familia de funciones de pérdida L_p en el entrenamiento de redes neuronales profundas, situándola dentro del marco de la teoría estadística de la decisión de Wald y Berger, donde cada valor de p induce un estimador de Bayes distinto: la media ($p = 2$), la mediana ($p = 1$) y la moda ($p \rightarrow 0$). A partir de esta caracterización teórica, y del análisis de robustez mediante funciones de influencia y continuidad Lipschitz, se evalúa experimentalmente el desempeño de las pérdidas L_p con $p \in \{2, 1, 0,5, 0,25, 0\}$ mediante simulaciones de Monte Carlo ($K = 100$ réplicas) en dos escenarios: regresión con contaminación de outliers y clasificación binaria con ruido simétrico de etiquetas, comparando el desempeño mediante MSE y accuracy, respectivamente, y contrastando las diferencias con pruebas de Wilcoxon pareadas. Los resultados muestran que, en regresión, L_2 es óptima en ausencia de contaminación pero se deteriora más de cien veces con solo un 10 % de outliers, mientras que las pérdidas no convexas $L_{0,5}$ y $L_{0,25}$ permanecen prácticamente estables, sugiriendo una región de robustez estadísticamente indistinguible para $p \in (0,25, 0,5)$. En clasificación, se confirma la jerarquía de robustez $L_0 > L_{0,5} > L_1 > L_2$ para niveles de ruido moderados, pero se identifica un fenómeno no anticipado en la literatura: dicha jerarquía se invierte cuando el ruido de etiquetas supera el 50 %, ya que las pérdidas con menor p terminan ajustándose a la señal mayoritaria, que en ese régimen es la incorrecta. En conjunto, estos resultados confirman que la elección de la función de pérdida constituye una decisión estadística con consecuencias directas sobre la robustez y la eficiencia de los modelos entrenados bajo datos potencialmente contaminados.

Palabras clave: Funciones de pérdida L_p , Robustez estadística, Redes neuronales profundas, Contaminación con outliers, Ruido de etiquetas, Teoría de decisiones, Monte Carlo.

Abstract

This thesis studies the family of L_p loss functions in the training of deep neural networks, situating it within the framework of statistical decision theory developed by Wald and Berger, where each value of p induces a different Bayes estimator: the mean ($p = 2$), the median ($p = 1$), and the mode ($p \rightarrow 0$). Building on this theoretical characterization, together with a robustness analysis based on influence functions and Lipschitz continuity, the performance of the L_p losses for $p \in \{2, 1, 0,5, 0,25, 0\}$ is evaluated experimentally through Monte Carlo simulations ($K = 100$ replicas) under two scenarios: regression with outlier contamination and binary classification with symmetric label noise, comparing performance via MSE and accuracy, respectively, and assessing the differences with paired Wilcoxon tests. The results show that, in regression, L_2 is optimal in the absence of contamination but deteriorates by more than a hundredfold with only 10 % outlier contamination, while the non-convex losses $L_{0,5}$ and $L_{0,25}$ remain practically stable, revealing a region of statistically indistinguishable robustness for $p \in (0,25, 0,5)$. In classification, the robustness hierarchy $L_0 > L_{0,5} > L_1 > L_2$ is confirmed for moderate noise levels, but a phenomenon not anticipated in the reviewed literature is identified: this hierarchy reverses once label noise exceeds 50 %, as losses with smaller p end up fitting the majority signal, which in that regime is the incorrect one. Taken together, these results confirm that the choice of loss function constitutes a statistical decision with direct consequences for the robustness and efficiency of models trained on potentially contaminated data.

Keywords: L_p loss functions, Statistical robustness, Deep neural networks, Outlier contamination, Label noise, Decision theory, Monte Carlo simulation.

Índice general

1. Introducción	6
2. Marco Teórico y Estado del Arte	9
2.1. Redes neuronales profundas	9
2.2. Fundamentos de la teoría de decisiones estadísticas	11
2.3. La familia de funciones de pérdida L_p	12
2.4. Robustez estadística: función de influencia y conexión distribucional	15
2.5. Estado del arte sobre funciones de pérdida robustas	18
3. Experimentos Numéricos	23
3.1. Diseño experimental	23
3.2. Experimentos de Monte Carlo	24
3.2.1. Clasificación con ruido de etiquetas	24
3.2.2. Regresión con contaminación con outliers	25
3.3. Resultados	26
3.3.1. Clasificación: ruido de etiquetas	26
3.3.2. Regresión: contaminación con outliers	30
3.3.3. Regresión con datos reales: el conjunto Airfoil Self-Noise	35
4. Conclusiones y Trabajo Futuro	41
5. Anexo	46
5.1. Implementación computacional	46
5.2. Reproducibilidad y semillas	48

Capítulo 1

Introducción

Las redes neuronales profundas se han consolidado como una de las herramientas más importantes del aprendizaje automático moderno, demostrando un desempeño sobresaliente en una amplia variedad de problemas de regresión y clasificación. El entrenamiento de estos modelos depende de manera crítica de la función de pérdida utilizada, ya que esta cuantifica la discrepancia entre las predicciones generadas por la red y los valores observados, guiando así el proceso de ajuste de parámetros mediante algoritmos de optimización basados en descenso de gradiente.

A pesar de su relevancia, la elección de la función de pérdida suele abordarse como una decisión secundaria, motivada principalmente por consideraciones computacionales o por convenciones ampliamente aceptadas dentro de la literatura. Sin embargo, esta elección posee profundas implicancias tanto estadísticas como prácticas, pues determina la sensibilidad del modelo frente a observaciones atípicas, ruido en las etiquetas y desviaciones respecto de los supuestos probabilísticos subyacentes.

Tradicionalmente, dos funciones de pérdida han dominado gran parte de las aplicaciones de aprendizaje supervisado: el error cuadrático medio (*Mean Squared Error*, MSE) y el error absoluto medio (*Mean Absolute Error*, MAE). El MSE destaca por sus propiedades analíticas favorables, tales como su diferenciabilidad y su carácter óptimo bajo supuestos gaussianos para los errores. No obstante, penaliza cuadráticamente las discrepancias grandes entre predicción y observación, lo que lo vuelve particularmente sensible a la presencia de valores atípicos.

Por otro lado, el MAE ofrece una mayor robustez frente a observaciones extremas debido a su crecimiento lineal. Además, posee propiedades de acotamiento que permiten establecer garantías teóricas de generalización [10]. Sin embargo, su falta de diferenciabilidad en el origen introduce desafíos adicionales para la optimización numérica. Asimismo, desde una perspectiva estadística, el MAE resulta óptimo cuando los errores siguen distribuciones de tipo Laplace.

Tanto el MSE como el MAE pueden interpretarse como casos particulares de una familia más amplia de funciones de pérdida parametrizadas por un exponente $p \geq 0$, conocidas como funciones de pérdida L_p . Esta familia permite transitar gradualmente entre distintos niveles de robustez y sensibilidad a errores grandes. En particular, valores de p menores que uno inducen pérdidas no convexas que reducen considerablemente la influencia de observaciones extremas, lo que ha motivado un creciente interés por su estudio en contextos donde los datos

presentan contaminación o ruido.

Desde una perspectiva más general, la elección de una función de pérdida puede analizarse dentro del marco de la teoría estadística de la decisión, donde las pérdidas determinan la noción de riesgo asociada a un procedimiento de estimación o aprendizaje. Bajo este enfoque, distintas funciones de pérdida conducen a diferentes reglas óptimas y a diferentes compromisos entre eficiencia estadística y robustez. En consecuencia, comprender el comportamiento de las funciones L_p no solo constituye un problema de interés computacional, sino también una cuestión fundamental desde el punto de vista teórico.

El objetivo principal de esta memoria es evaluar la eficacia y la robustez de distintas funciones de pérdida L_p en el entrenamiento de redes neuronales profundas bajo escenarios de contaminación de datos. Particularmente, se estudiará el efecto de la contaminación con *outliers* en problemas de regresión y del ruido de etiquetas en problemas de clasificación, comparando el desempeño de diversas elecciones del parámetro p .

Para alcanzar este objetivo general, se plantean los siguientes objetivos específicos:

- Caracterizar analíticamente las funciones de pérdida L_p dentro del marco de la teoría estadística de la decisión, estudiando sus propiedades matemáticas y estadísticas.
- Analizar las implicancias teóricas de distintas elecciones del parámetro p en términos de robustez, convexidad, consistencia y comportamiento frente a observaciones atípicas.
- Evaluar experimentalmente el desempeño de distintas funciones L_p implementadas en redes neuronales profundas bajo condiciones controladas de contaminación, tanto en tareas de regresión como de clasificación.
- Comparar cuantitativamente el impacto de cada función de pérdida sobre métricas de rendimiento y estudiar la significancia estadística de las diferencias observadas.

Cabe precisar el grado de novedad de cada uno de los cuatro objetivos respecto de la literatura existente. Los dos primeros objetivos, de naturaleza teórica, constituyen principalmente una labor de síntesis y articulación de resultados ya establecidos: la caracterización de L_p dentro de la teoría de la decisión se apoya directamente en el marco de Wald (1950) y Berger (1985), mientras que el análisis de robustez, convexidad y comportamiento frente a valores atípicos se fundamenta en la teoría clásica de la función de influencia de Hampel (1974) y en resultados más recientes sobre pérdidas robustas continuas, como los de Barron (2019). El aporte de esta memoria en estos dos objetivos consiste en integrar estos marcos, originalmente desarrollados de forma independiente, en una presentación unificada centrada específicamente en la familia L_p con $p \in [0, 2]$.

Los dos últimos objetivos, de naturaleza experimental, presentan un carácter más propio. Si bien existen antecedentes que evalúan casos particulares de la familia L_p bajo condiciones de contaminación, como el análisis de Ghosh et al. (2017) sobre el MAE frente al ruido de etiquetas o el de Robitzsch (2023) sobre la familia completa de exponentes p en modelos de ecuaciones estructurales, ninguno de estos trabajos evalúa la grilla completa de valores de p considerados aquí en redes neuronales profundas y de manera simultánea en dos regímenes de contaminación distintos, contaminación con *outliers* en regresión y ruido de etiquetas en clasificación, bajo un mismo diseño experimental controlado. Asimismo, la comparación cuantitativa mediante pruebas de Wilcoxon pareadas aplicadas sistemáticamente sobre réplicas

de Monte Carlo para establecer la significancia estadística de las diferencias entre funciones de pérdida, en lugar de limitarse a la comparación de valores promedio, constituye un aporte metodológico propio de esta memoria dentro de este contexto específico.

El presente documento se organiza de la siguiente manera. El Capítulo 2 desarrolla el marco teórico y la revisión del estado del arte, integrando elementos de teoría estadística de la decisión, aprendizaje estadístico, propiedades matemáticas de las funciones L_p y los principales antecedentes sobre funciones de pérdida robustas en aprendizaje profundo. El Capítulo 3 presenta los experimentos numéricos: el diseño experimental y la justificación metodológica, la implementación computacional de los estudios de simulación de Monte Carlo (clasificación con ruido de etiquetas, regresión sintética con contaminación con *outliers*, y una réplica del experimento de regresión sobre el conjunto de datos real Airfoil Self-Noise), y los resultados obtenidos. El Capítulo 4 sintetiza los principales hallazgos de esta memoria, discute sus limitaciones y propone líneas de trabajo futuro.

Capítulo 2

Marco Teórico y Estado del Arte

2.1. Redes neuronales profundas

Una red neuronal artificial es un modelo computacional compuesto por capas de unidades de procesamiento (neuronas) conectadas entre sí mediante parámetros ajustables. Dada una entrada $x \in \mathbb{R}^d$, la red aplica una secuencia de transformaciones afines seguidas de funciones de activación no lineales para producir una salida $f_\theta(x)$, donde θ denota el conjunto de pesos y sesgos de la red. La profundidad de la red (el número de capas) le otorga capacidad para aprender representaciones jerárquicas de los datos, lo que explica su éxito en problemas complejos de percepción y predicción [3].

Una red neuronal *feedforward* de profundidad L es una función $f_\theta : \mathbb{R}^{n_0} \rightarrow \mathbb{R}^{n_L}$ definida como la composición de L transformaciones sucesivas. Dado un vector de entrada $a^{(0)} = x \in \mathbb{R}^{n_0}$, la salida de la capa l -ésima, para $l = 1, \dots, L$, se calcula como

$$z^{(l)} = W^{(l)}a^{(l-1)} + b^{(l)}, \quad a^{(l)} = \sigma^{(l)}(z^{(l)}), \quad (2.1)$$

donde $W^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$ es la matriz de pesos, $b^{(l)} \in \mathbb{R}^{n_l}$ es el vector de sesgos, $\sigma^{(l)}$ es la función de activación de la capa (aplicada componente a componente) y n_l es el número de unidades de la capa l . El vector de parámetros de la red es $\theta = \{W^{(l)}, b^{(l)}\}_{l=1}^L$, y la salida de la red es $f_\theta(x) = a^{(L)}$. Las capas $l = 1, \dots, L-1$ se denominan *capas ocultas*, y la última, $l = L$, *capa de salida*.

Una función de activación es una función $\sigma : \mathbb{R} \rightarrow \mathbb{R}$, generalmente no lineal, que se aplica componente a componente sobre la salida afín $z^{(l)}$ de cada capa. Las funciones de activación más utilizadas en la práctica son:

- **Sigmoide:** $\sigma(z) = \frac{1}{1 + e^{-z}}$, con recorrido $(0, 1)$, históricamente asociada al desarrollo del algoritmo de retropropagación [17] y utilizada en esta memoria como activación de salida en el experimento de clasificación binaria (Sección 3).
- **Tangente hiperbólica:** $\sigma(z) = \tanh(z)$, con recorrido $(-1, 1)$.
- **Unidad lineal rectificadora (ReLU):** $\sigma(z) = \max(0, z)$, propuesta por Nair y Hinton [19] como alternativa a la sigmoide y la tangente hiperbólica para mitigar el problema del gradiente evanescente en redes profundas, y empleada como activación de las capas ocultas en ambos experimentos de esta memoria.

La no linealidad de σ es indispensable para que la profundidad de la red aporte capacidad expresiva adicional: si todas las activaciones fuesen la identidad, la composición de L transformaciones afines de la forma (2.1) sería a su vez una única transformación afín, $f_\theta(x) = W'x + b'$ para ciertos W', b' , independientemente del valor de L [3]. Es la alternancia entre transformaciones lineales y no linealidades puntuales la que permite a la red aproximar funciones no lineales de x .

Teorema 2.1.1 (Aproximación universal, versión informal). *Sea σ una función de activación continua, acotada y no constante. Entonces, para toda función continua $g : K \rightarrow \mathbb{R}$ definida sobre un subconjunto compacto $K \subset \mathbb{R}^d$ y todo $\epsilon > 0$, existe una red neuronal feedforward con una única capa oculta de ancho finito n_1 y activación σ tal que*

$$\sup_{x \in K} |f_\theta(x) - g(x)| < \epsilon. \quad (2.2)$$

Este resultado clásico garantiza que, en principio, una red suficientemente ancha con una sola capa oculta puede aproximar cualquier función continua sobre un dominio compacto con precisión arbitraria. Sin embargo, el teorema no especifica el ancho n_1 necesario ni garantiza que dicha red sea aprendible mediante descenso de gradiente a partir de datos finitos; en la práctica, aumentar la profundidad L en lugar del ancho suele ser más eficiente en número de parámetros para representar funciones jerárquicamente estructuradas [3]. Esta es la motivación teórica detrás del uso de arquitecturas profundas en los experimentos de esta memoria.

El aprendizaje consiste en ajustar θ para minimizar una función objetivo que mide la calidad de las predicciones sobre un conjunto de entrenamiento $\{(x_i, y_i)\}_{i=1}^N$. Esta función puede escribirse como un promedio sobre el conjunto de entrenamiento:

$$J(\theta) = \mathbb{E}_{(x,y) \sim \hat{p}_{\text{data}}} [\mathcal{L}(f(x; \theta), y)], \quad (2.3)$$

donde $\mathcal{L}$ es la función de pérdida, $f(x; \theta)$ es la salida predicha cuando la entrada es x , y $\hat{p}_{\text{data}}$ es la distribución empírica de los datos. Idealmente se preferiría optimizar sobre la verdadera distribución generadora p_{data} ; sin embargo, en la práctica esta distribución es desconocida y solo se dispone de un conjunto finito de muestras. La forma de convertir esto nuevamente en un problema de optimización es minimizar la pérdida esperada sobre el conjunto de entrenamiento, lo que se conoce como **minimización del riesgo empírico**:

$$\hat{J}(\theta) = \mathbb{E}_{(x,y) \sim \hat{p}_{\text{data}}} [\mathcal{L}(f(x; \theta), y)] = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(f(x^{(i)}; \theta), y^{(i)}). \quad (2.4)$$

En lugar de optimizar directamente el riesgo verdadero, se minimiza el riesgo empírico y se espera que el riesgo verdadero disminuya también. Este proceso se realiza típicamente mediante el algoritmo de retropropagación (*backpropagation*), originalmente formalizado por Rumelhart, Hinton y Williams [17], que calcula los gradientes de $\hat{J}$ respecto a θ aplicando la regla de la cadena sobre la composición de capas de la Ecuación (2.1), y el descenso de gradiente estocástico (SGD) o variantes adaptativas como Adam [6], que actualizan iterativamente los parámetros en la dirección que reduce la pérdida.

El objetivo final es reducir el error esperado de generalización. Si se conociera la verdadera distribución $p_{\text{data}}(x, y)$, minimizar el riesgo sería simplemente una tarea de optimización. Al

no conocerla, surge el problema del aprendizaje automático: generalizar a partir de muestras finitas.

Dos técnicas adicionales, empleadas en la arquitectura de ambos experimentos de esta memoria, merecen definición formal por su rol en la comparación entre funciones de pérdida.

Definición 2.1.1 (Normalización por lotes). *Dado un mini-lote de activaciones $\mathcal{B} = \{z_1, \dots, z_m\}$ correspondientes a una misma unidad de una capa, la normalización por lotes (Batch Normalization), propuesta por Ioffe y Szegedy [20], transforma cada activación como*

$$\hat{z}_i = \frac{z_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}, \quad \tilde{z}_i = \gamma \hat{z}_i + \beta, \quad (2.5)$$

donde $\mu_{\mathcal{B}}$ y $\sigma_{\mathcal{B}}^2$ son la media y varianza empíricas del lote, $\epsilon > 0$ es una constante de estabilidad numérica, y γ, β son parámetros entrenables que permiten a la red recuperar la escala y el desplazamiento originales si resulta óptimo hacerlo. Como se señala en el Anexo, esta técnica es particularmente relevante en el diseño experimental de esta memoria porque estabiliza la escala de las activaciones independientemente de la función de pérdida utilizada, evitando que las comparaciones entre pérdidas L_p se vean confundidas por efectos de escala en la propagación del gradiente.

Definición 2.1.2 (Dropout). *El dropout, propuesto por Srivastava et al. [21], es una técnica de regularización que, durante el entrenamiento, desactiva aleatoriamente cada unidad de una capa con probabilidad $1 - q$, de forma independiente para cada unidad y cada paso de entrenamiento, donde $q \in (0, 1]$ es la probabilidad de retención. Formalmente, la activación regularizada de la unidad j en la capa l es $\tilde{a}_j^{(l)} = r_j \cdot a_j^{(l)}$, con $r_j \sim \text{Bernoulli}(q)$ independientes. Esto evita la coadaptación excesiva entre unidades y reduce el sobreajuste, especialmente relevante en los escenarios de alta contaminación estudiados en esta memoria, donde el riesgo de que la red memorice observaciones ruidosas es mayor.*

Como se puede apreciar, una parte fundamental del trabajo a la hora de conseguir un buen rendimiento en aprendizaje automático gira en torno a la función de pérdida. La elección de $\mathcal{L}$ no es trivial: determina qué tipo de errores el modelo penaliza más, qué distribución se asume implícitamente sobre los residuos, y cuán robusto será el modelo ante irregularidades en los datos. Como se desarrollará en las secciones siguientes, esta elección tiene un fundamento estadístico profundo que conecta el aprendizaje automático con la teoría de decisiones.

2.2. Fundamentos de la teoría de decisiones estadísticas

La teoría estadística de la decisión, desarrollada formalmente por Wald [7] y extendida por Berger [1], proporciona el marco conceptual más riguroso para entender el rol de las funciones de pérdida en el aprendizaje estadístico. En este marco, el problema de estimación se formula como un juego entre un estadístico que elige una regla de decisión y la naturaleza, que controla el valor del parámetro de interés.

Formalmente, sea Θ el espacio de parámetros, $\mathcal{A}$ el espacio de acciones y $L : \Theta \times \mathcal{A} \rightarrow [0, \infty)$ una función de pérdida que asigna un costo a cada par (parámetro verdadero, acción

tomada). Una **regla de decisión** $\delta(x)$ es una función desde el espacio muestral $\mathcal{X}$ hacia el espacio de acciones $\mathcal{A}$: dado el valor observado $X = x$, la acción tomada es $\delta(x)$ [1].

La evaluación de una regla de decisión desde la perspectiva frecuentista se realiza a través de su **función de riesgo**, definida como la pérdida esperada ante distintas realizaciones de los datos:

$$R(\theta, \delta) = E_{\theta} [L(\theta, \delta(X))] = \int L(\theta, \delta(x)) dP_X(x | \theta). \quad (2.6)$$

Desde una perspectiva bayesiana, si $\pi^*(\theta)$ es la distribución de probabilidad creída para θ al momento de tomar una decisión, típicamente la distribución posterior luego de observar los datos, se define la **pérdida esperada bayesiana** de una acción a como [1]:

$$\rho(\pi^*, a) = E_{\pi^*} [L(\theta, a)]. \quad (2.7)$$

Ambas nociones de riesgo son complementarias: la primera evalúa el desempeño promedio de una regla ante distintas muestras para un θ fijo, mientras que la segunda promedia sobre la incertidumbre en θ dado un conjunto de datos observado.

Una regla de decisión δ_1 es **R-mejor** que δ_2 si $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ para todo $\theta \in \Theta$, con desigualdad estricta para algún θ . Una regla δ^* es **admisible** si no existe ninguna R-mejor que ella [1]. Esta noción de admisibilidad permite comparar formalmente distintas elecciones de función de pérdida, ya que cambiar L cambia tanto el riesgo como el conjunto de reglas admisibles. Nótese que en general no existirá una única regla que domine a todas las demás para todo θ : las funciones de riesgo suelen cruzarse, siendo distintas reglas preferibles en distintas regiones del espacio de parámetros.

En el contexto del aprendizaje estadístico, cada función de pérdida $\mathcal{L}_p$ define una regla de decisión δ_p , el algoritmo de entrenamiento que minimiza $\mathcal{L}_p$, con su propia función de riesgo $R(\theta, \delta_p)$. La minimización del riesgo empírico es una aproximación muestral al riesgo esperado verdadero, y la brecha entre ambos corresponde al error de generalización. Esta conexión tiene implicancias directas: distintas funciones de pérdida no son equivalentes, sino que codifican distintos supuestos sobre la distribución de los errores y distintos criterios de optimalidad.

Un resultado fundamental es que la elección de la función de pérdida determina implícitamente el estimador de Bayes bajo una distribución a priori dada. Bajo pérdida cuadrática $L(\theta, a) = (\theta - a)^2$, el estimador de Bayes es la media posterior; bajo pérdida absoluta $L(\theta, a) = |\theta - a|$, es la mediana posterior [1]. Esta correspondencia conecta directamente las pérdidas L_p con propiedades distribucionales específicas, como se desarrolla en la siguiente sección.

2.3. La familia de funciones de pérdida L_p

Antes de introducir la familia general L_p , conviene revisar las funciones de pérdida más utilizadas en la literatura estadística y de aprendizaje automático, dado que constituyen casos particulares de dicha familia.

La **pérdida cuadrática** $L(\theta, a) = (\theta - a)^2$ es la más ampliamente usada en problemas de estimación y regresión. En el contexto de redes neuronales, su uso como función de entrenamiento es conveniente por su diferenciabilidad y por el hecho de que, bajo supuestos de ruido

gaussiano en los datos, corresponde al estimador de máxima verosimilitud [2]. Sin embargo, la penalización cuadrática amplifica los errores grandes de forma desproporcionada, lo que la hace especialmente sensible a valores atípicos.

La **pérdida absoluta** $L(\theta, a) = |\theta - a|$ es un caso particular de la pérdida lineal asimétrica con $K_0 = K_1$. A diferencia de la pérdida cuadrática, penaliza los errores de forma proporcional a su magnitud, lo que la hace más robusta ante observaciones extremas. Su principal desventaja computacional es la no diferenciabilidad en el origen, que requiere tratamiento especial en métodos basados en gradiente.

La **pérdida 0-1**, definida como $L(\theta, a_i) = \mathbb{I}\{\theta \notin \Theta_i\}$, asigna costo cero a las decisiones correctas y costo unitario a las incorrectas. Su función de riesgo corresponde directamente a la probabilidad de error de clasificación, lo que la hace el criterio natural en problemas de clasificación. Sin embargo, su discontinuidad la hace intratable para optimización mediante gradiente.

Estas tres pérdidas pueden verse como casos particulares de la siguiente familia parametrizada:

$$\rho_p(r) = |r|^p, \quad p \geq 0, \quad (2.8)$$

donde $r = \theta - a$ denota el residuo, y que denominaremos *familia de funciones L_p* . Cada valor de p define un criterio de optimalidad distinto y, como consecuencia, lleva a distintos estimadores cuando se minimiza el riesgo esperado $R_p(t) = \mathbb{E}[|Y - t|^p]$. Los tres casos más relevantes son:

- **MSE** ($p = 2$):

$$\frac{\partial}{\partial t} R_2(t) = -2 \mathbb{E}[Y - t] = 0 \implies t^* = \mathbb{E}[Y].$$

El estimador óptimo bajo riesgo cuadrático es la **media** (o la esperanza condicional $\mathbb{E}[Y | X]$).

- **MAE** ($p = 1$):

$$\frac{\partial}{\partial t} R_1(t) = \mathbb{E}[-\text{sgn}(Y - t)] = 0 \implies P(Y \leq t) = P(Y \geq t) \implies t^* = \text{Mediana}(Y).$$

El estimador óptimo bajo riesgo absoluto es la **mediana** (o la mediana condicional).

- **Pérdida 0-1** ($p \rightarrow 0$):

$$R_{0-1}(t) = P(Y \neq t) = 1 - P(Y = t) \implies t^* = \arg \max_t P(Y = t).$$

El estimador óptimo bajo pérdida 0-1 es la **moda**, lo que en clasificación resulta en el clasificador de Bayes.

Desde el punto de vista computacional, ρ_p no es diferenciable en $r = 0$ para $p \leq 1$. Para hacerla tratable mediante descenso de gradiente, se utiliza la versión suavizada propuesta en [11]:

$$\rho_{p,\varepsilon}(x) = (x^2 + \varepsilon)^{p/2}, \quad (2.9)$$

donde $\varepsilon > 0$ es un parámetro de suavizado. Robitzsch [11] recomienda $\varepsilon = 10^{-3}$ para $p > 0$ y $\varepsilon = 10^{-2}$ para $p = 0$, valores que equilibran la aproximación a la función original con la estabilidad numérica del proceso de optimización.

En cuanto a sus propiedades analíticas, para $p \geq 1$ la función ρ_p es convexa, lo que garantiza la existencia de mínimos globales únicos y la validez de las cotas de error de estimación derivadas en la teoría del aprendizaje estadístico [8]. Para $0 < p < 1$, la función es cóncava, introduciendo múltiples mínimos locales que dificultan la optimización pero otorgan mayor robustez ante valores atípicos.

El caso límite $p \rightarrow 0$ da origen a la pérdida L_0 :

$$\rho_0(x) = \begin{cases} 0 & x = 0, \\ 1 & x \neq 0. \end{cases} \quad (2.10)$$

Minimizar esta función equivale a minimizar el número de residuos no nulos, lo que la hace teóricamente deseable bajo supuestos de dispersión (*sparsity*). Como L_0 no es diferenciable en ningún punto, se utiliza la aproximación diferenciable propuesta por O'Neill y Burke, citada en [11]:

$$\phi_\varepsilon(x) = \frac{x^2}{x^2 + \varepsilon}. \quad (2.11)$$

Esta función actúa como puente entre L_2 (cuando ε es grande) y L_0 (cuando ε es pequeño), lo que facilita su uso en esquemas de optimización iterativa con ε decreciente.

Cabe destacar que la parametrización suavizada $\rho_{p,\varepsilon}$ empleada en este trabajo coincide, salvo notación, con la denominada pérdida de *Charbonnier generalizada*, utilizada originalmente en tareas de estimación de flujo óptico y profundidad, y posteriormente incorporada por Barron [12] como caso particular de una función de pérdida robusta con dos parámetros continuos: uno de forma, análogo a p , y uno de escala, análogo a ε . Dicha función unifica, mediante un único parámetro de forma α , un amplio espectro de pérdidas clásicas: la pérdida L_2 ($\alpha = 2$), la pérdida de Charbonnier o pseudo-Huber ($\alpha = 1$, equivalente a L_1 suavizada), la pérdida de Cauchy o Lorentziana ($\alpha \rightarrow 0$), la pérdida de Geman–McClure ($\alpha = -2$) y la pérdida de Welsch o Leclerc ($\alpha \rightarrow -\infty$), extendiendo así el espectro de robustez más allá de los valores $p \in [0, 2]$ considerados en esta memoria. Barron [12] demuestra que, para $\alpha \leq 1$, la derivada de esta pérdida se vuelve *redescendiente*: su magnitud decrece a medida que $|x|$ crece más allá de la escala c , de modo que el gradiente inducido por un residuo suficientemente grande se atenúa en lugar de crecer, replicando en un marco continuo y diferenciable en todo punto el comportamiento de robustez creciente que se observa al disminuir p dentro de la familia L_p . Interpretando esta pérdida como el negativo del logaritmo de una densidad de probabilidad, se obtiene además una familia de distribuciones que generaliza simultáneamente a la normal ($\alpha = 2$) y a la Cauchy ($\alpha = 0$), lo que permite tratar el parámetro de forma como una variable libre, optimizada junto con los pesos de la red, de manera que el grado de robustez de la pérdida se adapte automáticamente durante el entrenamiento sin necesidad de fijar p o ε manualmente [12]. Esta perspectiva adaptativa constituye una extensión natural del enfoque desarrollado en esta memoria, y se retoma como línea de trabajo futuro en el Capítulo 4.

En la Figura 2.1 se puede observar el comportamiento de las funciones L_p para diferentes valores de p con $\varepsilon = 0,001$.

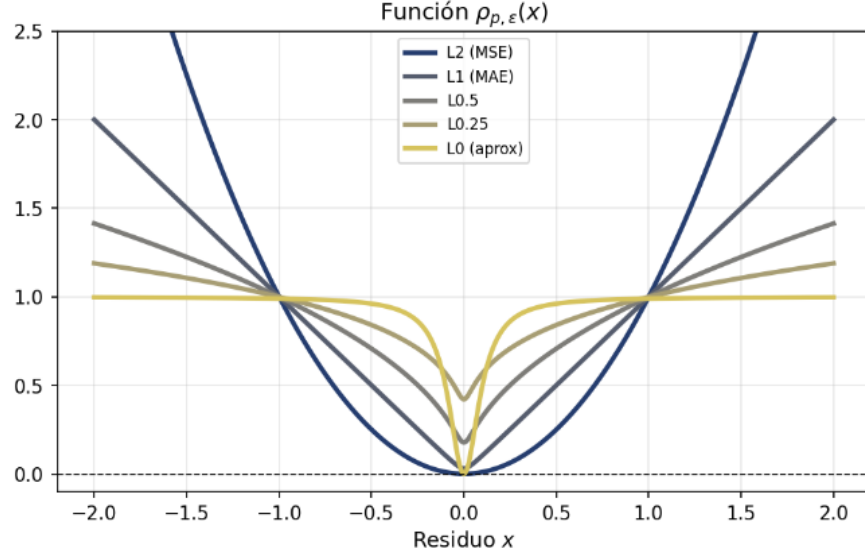


Figura 2.1: Comportamiento de las funciones L_p suavizadas para distintos valores de p .

2.4. Robustez estadística: función de influencia y conexión distribucional

La robustez de un estimador ante contaminación en los datos puede cuantificarse formalmente mediante la función de influencia, introducida por Hampel [4] como herramienta fundamental de la estadística robusta. Dado un estimador T y una distribución F , la función de influencia mide el efecto asintótico de una contaminación puntual en x_0 :

$$IF(x_0; T, F) = \lim_{\varepsilon \rightarrow 0^+} \frac{T((1 - \varepsilon)F + \varepsilon\delta_{x_0}) - T(F)}{\varepsilon}, \quad (2.12)$$

donde δ_{x_0} es la masa de Dirac centrada en x_0 . Un estimador es robusto si su función de influencia es acotada, es decir, si ningún valor extremo puede perturbar arbitrariamente al estimador.

Para el estimador de mínimos cuadrados (asociado a la pérdida L_2) la función de influencia es

$$IF(x_0; \mu, F) = x_0 - \mu, \quad (2.13)$$

que es lineal y no acotada. Esto implica que un único *outlier* de magnitud arbitraria puede desplazar ilimitadamente al estimador. En contraste, para la mediana (asociada a la pérdida L_1) la función de influencia es:

$$IF(x_0; \text{mediana}, F) = \frac{\mathbb{I}\{x_0 \leq m\} - 1/2}{f(m)}, \quad (2.14)$$

que es acotada independientemente del valor de x_0 . Esta propiedad fundamenta teóricamente la mayor robustez del MAE ante valores atípicos.

Para completar el espectro, analizamos la pérdida 0-1, minimizada por la **moda**. Asumiendo una distribución F continua, unimodal y con densidad f dos veces diferenciable, la

función de influencia de la moda se aproxima como:

$$IF(x_0; \text{moda}, F) = -\frac{f'(x_0)}{f''(\theta_0)}, \quad (2.15)$$

donde θ_0 es la verdadera moda y $f''(\theta_0) < 0$. A diferencia del MSE (influencia lineal) y el MAE (influencia acotada y constante por tramos), la función de influencia de la pérdida 0-1 posee una propiedad *redescendiente*: a medida que x_0 se aleja hacia las colas de la distribución, $f'(x_0) \rightarrow 0$, de modo que $IF(x_0) \rightarrow 0$. Los *outliers* extremos son asintóticamente ignorados, lo que confiere a la moda el máximo punto de ruptura teórico, aunque a expensas de la eficiencia estadística.

Esta distinción tiene además una interpretación distribucional directa. La maximización de la verosimilitud bajo una distribución gaussiana de errores es matemáticamente equivalente a minimizar el MSE, mientras que bajo una distribución de Laplace es equivalente a minimizar el MAE [10]. La distribución de Laplace posee colas más pesadas que la gaussiana, asignando mayor probabilidad a errores grandes y, en consecuencia, penalizándolos menos durante la optimización. De manera análoga, optimizar la pérdida 0-1 es equivalente a realizar MLE asumiendo que los residuos siguen una distribución insensible a la magnitud espacial del error, típicamente modelada como una mezcla de masa de Dirac concentrada en cero y un fondo probabilístico constante. En esta formulación, un error de magnitud ε es igualmente penalizado que uno de magnitud 10ε , justificando que el modelo se concentre exclusivamente en maximizar la densidad de aciertos exactos [5]. Más en general, la familia L_p puede interpretarse como el espacio de pérdidas inducido por distribuciones de error con colas progresivamente más pesadas a medida que p decrece, lo que explica la graduación continua de robustez observada al variar p .

Las tres expresiones anteriores no son fórmulas aisladas, sino instancias particulares de un mismo marco teórico: la teoría de M-estimadores. Si un funcional T queda definido implícitamente como la solución $\theta = T(F)$ de una ecuación de estimación de la forma

$$\int \psi(y, \theta) dF(y) = 0, \quad (2.16)$$

donde ψ es una función de puntuación (*score function*) asociada a la pérdida considerada, entonces, bajo condiciones de regularidad, la función de influencia de T admite la expresión general [4]:

$$IF(x_0; T, F) = \frac{\psi(x_0, \theta)}{-\int \frac{\partial}{\partial \theta} \psi(y, \theta) dF(y)} \Bigg|_{\theta=T(F)}. \quad (2.17)$$

El numerador captura el efecto *directo* de la observación contaminante x_0 sobre la ecuación de estimación, mientras que el denominador (la derivada esperada de ψ , cambiada de signo) normaliza dicho efecto por la curvatura de la función objetivo en torno al valor verdadero.

Para $p = 2$, la condición de primer orden de la pérdida cuadrática da $\psi(y, \theta) = y - \theta$, con derivada constante $\partial\psi/\partial\theta = -1$; sustituyendo en (2.17) se recupera exactamente la (2.13), $IF(x_0; \mu, F) = x_0 - \mu$. La ausencia de curvatura en ψ es precisamente lo que impide que el

denominador de (2.17) “amortigüe” observaciones extremas, de modo que la influencia crece linealmente y sin cota con x_0 .

Para $p = 1$, la función de puntuación es $\psi(y, \theta) = \text{sign}(y - \theta)$, cuya derivada generalizada es $\partial\psi/\partial\theta = -2\delta(y - \theta)$, con δ la delta de Dirac; evaluando (2.17) en $\theta = m = \text{Mediana}(F)$ se obtiene

$$IF(x_0; \text{mediana}, F) = \frac{\text{sign}(x_0 - m)}{2f(m)}, \quad (2.18)$$

expresión equivalente, salvo el signo de la convención adoptada para ψ , a la (2.14). A diferencia del caso $p = 2$, aquí ψ es acotada por construcción, lo que acota directamente la influencia,

$$|IF(x_0; \text{mediana}, F)| \leq 1/(2f(m)),$$

para todo x_0 ; sin embargo, esta cota no decae con la distancia a m , sino que permanece constante en magnitud, $\pm 1/(2f(m))$, para cualquier x_0 suficientemente alejado del centro de la distribución. La mediana es, en la terminología de Hampel, un estimador de influencia acotada pero no redescendiente.

El caso $p \rightarrow 0$, correspondiente a la moda, no admite una derivación directa a partir de (2.17), porque el funcional moda no está definido por una ecuación de puntuación suave sobre y , sino como el maximizador de la densidad, $T(F) = \arg \max_t f_F(t)$, sujeto a la condición de primer orden $f'(\theta) = 0$. Bajo el modelo de contaminación de Hampel $F_\varepsilon = (1 - \varepsilon)F + \varepsilon\delta_{x_0}$, la masa puntual δ_{x_0} no posee densidad en el sentido usual; la derivación rigurosa aproxima δ_{x_0} mediante una sucesión de densidades suaves concentradas en x_0 y toma el límite cuando su ancho de banda tiende a cero [4]. Bajo el supuesto de que f es dos veces diferenciable y que θ_0 es la verdadera moda ($f'(\theta_0) = 0$, $f''(\theta_0) < 0$), este procedimiento conduce a

$$IF(x_0; \text{moda}, F) = -\frac{f'(x_0)}{f''(\theta_0)}, \quad (2.19)$$

que recupera (2.15). Intuitivamente, la expresión indica que una masa puntual en x_0 desplaza la moda en una dirección y magnitud proporcionales a la pendiente local de la densidad en x_0 : si x_0 se ubica en una zona donde f crece hacia la moda, la contaminación “atrae” la moda hacia x_0 ; si f es aproximadamente plana (como ocurre en las colas de una distribución unimodal, donde $f'(x_0) \rightarrow 0$), la influencia se anula. A diferencia de la mediana, aquí la influencia no solo está acotada, sino que *decae a cero* en las colas: es este comportamiento *redescendiente* el que le confiere a la moda el máximo punto de ruptura teórico mencionado en el texto.

El comportamiento redescendiente de la moda, sin embargo, no es exclusivo de la pérdida 0-1: una propiedad relacionada, aunque de naturaleza distinta, puede establecerse formalmente para el MAE a través de su continuidad Lipschitz.

La propiedad de continuidad Lipschitz del MAE, demostrada formalmente por Qi et al. [10], refuerza esta robustez desde la perspectiva de la generalización. Si el MAE es 1-Lipschitz, entonces la diferencia de objetivos ante ruido aditivo η en la entrada satisface $|h(x + \eta) - h(x)| \leq \Lambda \|\eta\|_2$, donde Λ es la constante de Lipschitz de la red. Esta cota superior formal no existe para el MSE, que no es Lipschitz continuo [10]. Esta distinción resulta directamente pertinente para el diseño experimental de esta memoria: en el experimento de regresión (Capítulo 3), la contaminación con *outliers* introduce precisamente el tipo de perturbación

puntual y de gran magnitud frente a la cual la cota Lipschitz del MAE garantiza un control formal de la degradación del riesgo, mientras que el MSE carece de dicha garantía, lo cual es consistente con la fragilidad empírica observada para L_2 ante *outliers* en la Sección 3.3.

Esta noción de control formal del gradiente admite además una generalización continua más allá del caso particular $p = 1$.

Barron [12] formaliza la propiedad redescendiente descrita anteriormente para un continuo de valores del parámetro de forma. Acotando la magnitud del gradiente de su pérdida general, muestra que dicha magnitud nunca supera $1/c$ cuando el parámetro de forma es menor o igual a uno, y que a medida que dicho parámetro tiende a $-\infty$ el gradiente asociado a un residuo cuya magnitud supera $3c$ se anula casi por completo. Este resultado formaliza, para toda la familia continua de pérdidas robustas (de la cual L_p constituye un caso particular relevante) la intuición desarrollada a partir de la función de influencia: cuanto menor es el parámetro que controla la forma de la pérdida, mayor es el grado en que el proceso de optimización ignora las observaciones extremas durante el entrenamiento. Trasladado a la grilla discreta $p \in \{2, 1, 0.5, 0.25, 0\}$ empleada en los experimentos de esta memoria, este resultado sugiere que la ganancia de robustez observada al disminuir p no es un artefacto particular de las funciones L_p construidas en la Sección 2.3, sino la manifestación de un fenómeno más general propio de las pérdidas con gradiente acotado y decreciente. Esto respalda la interpretación, retomada en el Capítulo 4, de que el parámetro de forma p podría en principio tratarse como una cantidad continua y adaptable en lugar de un hiperparámetro discreto fijo.

Los fundamentos matemáticos y estadísticos expuestos en las secciones anteriores permiten situar la familia L_p dentro de un panorama más amplio de investigación sobre robustez en aprendizaje profundo. La siguiente sección revisa los antecedentes que, desde la teoría del aprendizaje estadístico, el aprendizaje profundo bajo ruido de etiquetas y la estadística robusta aplicada a modelos multivariados, resultan más directamente relevantes para los objetivos de esta memoria.

2.5. Estado del arte sobre funciones de pérdida robustas

El análisis comparativo de funciones de pérdida en aprendizaje automático ha avanzado desde perspectivas complementarias: la teoría estadística del aprendizaje, el aprendizaje profundo bajo condiciones adversariales, y la estadística robusta aplicada a modelos multivariados. Los trabajos que se revisan a continuación representan las contribuciones más directamente relevantes para los objetivos de esta memoria.

Rosasco et al. [8] estudian el impacto de la elección de la función de pérdida desde la teoría del aprendizaje estadístico. Formalmente, para una función de pérdida convexa V y un espacio de hipótesis $\mathcal{H}_R$, descomponen el error de generalización del minimizador empírico f_D respecto del minimizador verdadero $f_0 = \arg \min_f I[f]$, con $I[f] = \int V(f(x), y) dP(x, y)$, como

$$I[f_D] - I[f_0] = (I[f_D] - I[f_R]) + (I[f_R] - I[f_0]), \quad (2.20)$$

donde f_R es el minimizador del riesgo empírico sobre $\mathcal{H}_R$. El primer término, el error de estimación, se acota mediante números de cobertura de $\mathcal{H}_R$, y la velocidad de dicha cota depende explícitamente de propiedades de V (su módulo de continuidad y su comportamiento

cerca del origen). Bajo una hipótesis de convexidad, satisfecha por todas las pérdidas comúnmente usadas, demuestran que distintas funciones de pérdida generan distintas tasas de convergencia del error de estimación, siendo la *hinge loss* la de mejor tasa entre las pérdidas convexas analizadas, mientras que la pérdida cuadrática converge significativamente más lento. Además, prueban el siguiente resultado de consistencia: si $V(w, y) = V(wy)$ es convexa y decreciente en un entorno de 0, entonces el signo del minimizador del riesgo esperado coincide con la solución de Bayes,

$$\text{sgn}(f_0(x)) = \begin{cases} +1 & \text{si } p(1 | x) > p(-1 | x), \\ -1 & \text{si } p(1 | x) < p(-1 | x), \end{cases} \quad (2.21)$$

lo que constituye una propiedad de consistencia compartida por toda la familia de pérdidas convexas. Este trabajo establece el marco dentro del cual pueden compararse las pérdidas L_p con $p \geq 1$, y delimita también el límite de ese marco: las pérdidas con $p < 1$ no son convexas y quedan fuera de sus garantías teóricas, ya que la Ecuación (2.21) exige convexidad.

Ghosh et al. [9] abordan el problema de robustez ante ruido de etiquetas en redes neuronales profundas. Bajo el modelo de ruido no uniforme, denotan por $\eta_{\mathbf{x}}$ la probabilidad de que la etiqueta de $\mathbf{x}$ sea corrompida y por $\bar{\eta}_{\mathbf{x}j}$ la probabilidad de que sea reasignada a la clase j , de modo que el riesgo- L bajo datos ruidosos se escribe como

$$R_L^\eta(f) = \mathbb{E}_{\mathcal{D}} \left[(1 - \eta_{\mathbf{x}}) L(f(\mathbf{x}), y_{\mathbf{x}}) + \sum_{j \neq y_{\mathbf{x}}} \bar{\eta}_{\mathbf{x}j} L(f(\mathbf{x}), j) \right]. \quad (2.22)$$

Definen formalmente la tolerancia al ruido como la propiedad de que el minimizador de R_L^η coincida con el minimizador de R_L bajo etiquetas limpias, y demuestran que esta propiedad se satisface para ruido simétrico o uniforme ($\eta_{\mathbf{x}} = \eta$ constante) si L satisface la condición de **simetría**

$$\sum_{i=1}^k L(f(\mathbf{x}), i) = C, \quad \forall \mathbf{x} \in \mathcal{X}, \forall f, \quad (2.23)$$

para alguna constante C , siempre que la tasa de ruido satisfaga $\eta < \frac{k-1}{k}$, donde k es el número de clases (Teorema 1 de [9]). Demuestran que el MAE satisface la Ecuación (2.23), mientras que el MSE y la entropía cruzada no lo hacen. Los experimentos en MNIST, CIFAR-10 y cuatro *datasets* de texto muestran que el MAE mantiene precisiones estables bajo tasas de ruido de etiquetas de hasta el 80 %, mientras que la entropía cruzada colapsa con apenas un 40 % de ruido.

Qi et al. [10] complementan el trabajo anterior desde la perspectiva de la regresión vector-a-vector con DNNs. Recuerdan que una función f es β -Lipschitz continua si $\|f(\mathbf{x}) - f(\mathbf{y})\|_p \leq \beta \|\mathbf{x} - \mathbf{y}\|_p$ para todo $\mathbf{x}, \mathbf{y}$, y demuestran que, para las pérdidas

$$\mathcal{L}_{MAE}(S, S^*) = \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}_i - \mathbf{y}_i\|_1, \quad \mathcal{L}_{MSE}(S, S^*) = \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}_i - \mathbf{y}_i\|_2^2, \quad (2.24)$$

el MAE es exactamente 1-Lipschitz continuo (por la desigualdad triangular sobre $\|\cdot\|_1$), mientras que el MSE no admite tal cota. Esta propiedad es la que permite acotar el error

de estimación $\mathcal{T} = \sup_{f_v \in \mathbb{F}} |\mathcal{L}(f_v) - \hat{\mathcal{L}}(f_v)|$ mediante la complejidad de Rademacher empírica $\hat{\mathcal{R}}_S(\mathbb{H}) := \mathbb{E}_{\sigma} \left[\sup_{h \in \mathbb{H}} \frac{1}{N} \sum_{i=1}^N \sigma_i h(\mathbf{x}_i) \right]$ de la clase de funciones aprendidas, obteniendo

$$\mathcal{T} \leq \hat{\mathcal{R}}_S(\mathbb{L}) \leq \hat{\mathcal{R}}_S(\mathbb{F}), \quad (2.25)$$

cota que no tiene equivalente para el MSE. De esta propiedad se sigue que la diferencia de objetivos ante ruido aditivo está formalmente acotada para redes entrenadas con MAE, lo que constituye una garantía de robustez ante perturbaciones adversarias. A nivel experimental, una DNN entrenada con MAE supera consistentemente a una entrenada con MSE en las métricas PESQ (2,93 vs. 2,85) y STOI (0,8509 vs. 0,8317) en el corpus Edinburgh Noisy Speech.

Robitzsch [11] analiza la familia completa de funciones L_p con $0 < p \leq 1$ y una aproximación diferenciable de L_0 en modelos de ecuaciones estructurales, evaluando cada estimador $\hat{\theta}$ mediante el sesgo y la raíz del error cuadrático medio sobre réplicas de simulación,

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta_0, \quad \text{RMSE}(\hat{\theta}) = \sqrt{\mathbb{E}[(\hat{\theta} - \theta_0)^2]}, \quad (2.26)$$

con θ_0 el valor poblacional verdadero. A diferencia de los trabajos anteriores, la robustez aquí se mide ante errores de especificación del modelo (*model misspecification*) más que ante contaminación con *outliers* o ruido de etiquetas. Sus hallazgos más relevantes para esta memoria son tres: el parámetro de suavizado ε , definido para esta memoria en la Ecuación (2.13), tiene un impacto significativo sobre $\text{Bias}(\hat{\theta})$ y $\text{RMSE}(\hat{\theta})$, con $\varepsilon = 10^{-3}$ óptimo para $p \leq 0,5$ y $\varepsilon = 10^{-2}$ necesario para estabilidad de L_0 ; la función L_0 supera a las demás en sesgo y RMSE cuando la muestra es suficientemente grande ($N \geq 1000$), pero para $N = 250$ la función $L_{0,5}$ es preferible; y los errores estándar derivados por el método delta exhiben coberturas estadísticas aceptables para todos los estimadores considerados. Como se retoma en la discusión del Capítulo 3, esta sensibilidad del desempeño de L_0 al tamaño muestral y al valor de ε resulta relevante para interpretar los resultados experimentales de regresión de esta memoria.

Más allá de la caracterización original de Ghosh et al. [9], una línea reciente de investigación en clasificación profunda ha buscado generalizar y sistematizar el diseño de funciones de pérdida robustas al ruido de etiquetas, aportando perspectivas complementarias a la familia L_p estudiada en esta memoria.

Zhou et al. [13] demuestran que, más allá de la condición de simetría de la Ecuación (2.23), cualquier función de pérdida puede volverse tolerante al ruido de etiquetas si se restringe el espacio de hipótesis de modo que la salida de la red se aproxime a un vector *one-hot*. Dado que imponer esta restricción de forma exacta produce gradientes nulos en casi todo punto, los autores proponen una *regularización dispersa* que resuelve, en su lugar, el problema de minimización de riesgo restringido

$$\min_{f \in \mathcal{H}} R_L(f) \quad \text{s.a.} \quad \|f\|_q \leq \gamma, \quad 0 \leq q \leq 1, \quad (2.27)$$

que en la práctica se entrena como el problema penalizado $\sum_{i=1}^N L(f(\mathbf{x}_i), \tilde{y}_i) + \lambda \|f(\mathbf{x}_i)\|_q^q$. Este resultado es relevante para la presente memoria porque desplaza el rol de la familia L_p : en lugar de emplearse directamente como función de pérdida sobre los residuos, aparece

aquí como regularizador que induce dispersión (*sparsity*) en las predicciones, reforzando la interpretación de L_0 como preferencia por soluciones dispersas discutida en la Sección 2.3.

Feng et al. [14] abordan el mismo problema desde una perspectiva distinta, expandiendo la entropía cruzada categórica $L_{CCE}(f(\mathbf{x}), y) = -\log f_y(\mathbf{x})$ como una serie de Taylor centrada en $f_y(\mathbf{x}) = 1$,

$$L_{CCE}(f(\mathbf{x}), y) = \sum_{i=1}^{\infty} \frac{(1 - f_y(\mathbf{x}))^i}{i}, \quad (2.28)$$

y truncando dicha serie en el orden t para definir la *Taylor Cross Entropy* (TCE),

$$L_{t-CE}(f(\mathbf{x}), y) = \sum_{i=1}^t \frac{(1 - f_y(\mathbf{x}))^i}{i}, \quad t \in \mathbb{N}^+. \quad (2.29)$$

Los autores demuestran que, cuando $t = 1$, $L_{t-CE} = 1 - f_y(\mathbf{x}) = \frac{1}{2}L_{MAE}$, es decir, el MAE es exactamente la aproximación de primer orden de esta serie; cuando $t \rightarrow \infty$, L_{t-CE} recupera la entropía cruzada exacta. Esta reformulación provee una justificación adicional, complementaria a la condición de simetría de Ghosh et al. [9], de por qué el MAE resulta más robusto que la entropía cruzada frente al ruido de etiquetas: al truncar la serie en un orden bajo, se limita explícitamente la penalización que la pérdida asigna a las observaciones cuya predicción $f_y(\mathbf{x})$ está muy alejada de 1.

Liu y Guo [15] proponen una familia alternativa de pérdidas, las funciones de *peer loss*, que logran tolerancia al ruido de etiquetas sin requerir conocimiento previo de la tasa de ruido. Dada una muestra $(\mathbf{x}_n, \tilde{y}_n)$ con etiqueta potencialmente ruidosa, y dos muestras “par” $(\mathbf{x}_{n_1}, \tilde{y}_{n_1})$, $(\mathbf{x}_{n_2}, \tilde{y}_{n_2})$ elegidas al azar del conjunto de entrenamiento (con $n_1 \neq n_2 \neq n$), la *peer loss* se define como

$$\ell_{\text{peer}}(f(\mathbf{x}_n), \tilde{y}_n) = \ell(f(\mathbf{x}_n), \tilde{y}_n) - \ell(f(\mathbf{x}_{n_1}), \tilde{y}_{n_2}), \quad (2.30)$$

donde el primer término evalúa la predicción del clasificador sobre su propia etiqueta ruidosa, y el segundo término, construido deliberadamente a partir de instancias no correspondientes, penaliza a la red por ajustarse en exceso al patrón de ruido, sin requerir estimar e_{-1}, e_{+1} (las tasas de error de las etiquetas). A diferencia de las pérdidas L_p y de la condición de simetría de la Ecuación (2.23), la robustez de ℓ_{peer} en la Ecuación (2.30) no depende de la forma funcional de la pérdida base ℓ , sino de esta corrección relativa entre observaciones, lo que sugiere que la familia L_p estudiada en esta memoria constituye una entre varias estrategias posibles para lograr tolerancia al ruido de etiquetas, y no la única vía disponible.

Ma et al. [16] generalizan aún más la condición de simetría, demostrando que cualquier función de pérdida L puede normalizarse como

$$L_{\text{norm}}(f(\mathbf{x}), y) = \frac{L(f(\mathbf{x}), y)}{\sum_{j=1}^K L(f(\mathbf{x}), j)}, \quad (2.31)$$

de modo que L_{norm} satisface automáticamente la condición de simetría (Ecuación (2.23), con $C = 1$) y es, por tanto, tolerante al ruido si $\eta < \frac{K-1}{K}$. Sin embargo, identifican que las pérdidas robustas así obtenidas, incluyendo el propio MAE, sufren de un problema de *underfitting* en conjuntos de datos complejos. Para resolverlo, proponen el marco de *Active Passive Loss*

(APL), que combina una componente “activa” L_{Active} (que maximiza la probabilidad de la clase etiquetada) con una componente “pasiva” $L_{Passive}$ (que minimiza la probabilidad de las clases restantes),

$$L_{APL} = \alpha \cdot L_{Active} + \beta \cdot L_{Passive}, \quad \alpha, \beta > 0, \quad (2.32)$$

logrando robustez sin sacrificar completamente la capacidad de ajuste. Este hallazgo resulta particularmente relevante para interpretar los resultados de la presente memoria: el compromiso entre robustez y capacidad de ajuste, observado en el Capítulo 3 en el desempeño ligeramente inferior de $L_{0,5}$ y L_0 respecto de L_2 en ausencia de contaminación, es documentado de forma independiente por Ma et al. para el caso de clasificación con entropía cruzada, lo que sugiere que se trata de una tensión general entre robustez y eficiencia estadística, y no de una particularidad del diseño experimental de este trabajo.

Finalmente, desde una perspectiva distinta, orientada al aprendizaje profundo general y la visión computacional, Barron [12] propone una función de pérdida robusta con dos parámetros continuos (de forma y de escala) que generaliza simultáneamente las pérdidas de Cauchy/Lorentziana, Geman-McClure, Welsch/Leclerc, la Charbonnier generalizada (equivalente, salvo notación, a la suavización $\rho_{p,\varepsilon}$ de la Ecuación (2.12), empleada en esta memoria) y L_2 . Su principal contribución, discutida con la notación completa en la Sección 2.3, es mostrar que, al interpretar dicha pérdida como el negativo del logaritmo de una densidad de probabilidad, el parámetro de forma puede tratarse como una variable libre optimizada junto con los pesos de la red, permitiendo que el grado de robustez de la función de pérdida se adapte automáticamente durante el entrenamiento sin ajuste manual. Barron demuestra experimentalmente que esta adaptatividad mejora el desempeño en tareas de síntesis de imágenes, estimación de profundidad, registro y *clustering*, lo que constituye evidencia adicional, en dominios distintos al de esta memoria, de que el parámetro que controla la robustez de la pérdida, análogo al p de la familia L_p , no admite en general un valor óptimo único y fijo, sino que su valor preferible depende del nivel de contaminación de los datos, en línea con los resultados experimentales presentados en el Capítulo 3.

Con este panorama teórico y de antecedentes como base, el capítulo siguiente traduce estas consideraciones en un diseño experimental concreto, orientado a contrastar empíricamente el comportamiento de la familia L_p bajo contaminación con *outliers* en regresión y ruido de etiquetas en clasificación.

Capítulo 3

Experimentos Numéricos

3.1. Diseño experimental

Este capítulo describe el diseño experimental, la implementación computacional y los resultados del estudio. El objetivo central es evaluar de manera controlada cómo varía la robustez del entrenamiento de redes neuronales profundas al modificar el parámetro p de la función de pérdida L_p , tanto ante contaminación con *outliers* en la variable respuesta como ante ruido de etiquetas en problemas de clasificación. La estrategia general consiste en contaminar exclusivamente los conjuntos de entrenamiento manteniendo los conjuntos de prueba completamente limpios, de modo que las métricas de evaluación reflejen únicamente la calidad de generalización del modelo y no el efecto directo de las perturbaciones.

Se consideran tres experimentos principales: dos de regresión, donde la contaminación con *outliers* afecta directamente a la variable respuesta y , y uno de clasificación, donde el ruido corrompe las etiquetas de clase. En ambos casos se comparan funciones L_p para $p \in \{2, 1, 0.5, 0.25, 0\}$ y se analizan las diferencias en desempeño a medida que la tasa de contaminación aumenta.

La función L_0 no es diferenciable, por lo que se utiliza la aproximación diferenciable

$$\phi_\varepsilon(r) = \frac{r^2}{r^2 + \varepsilon},$$

con $\varepsilon = 0.01$. Para $p > 0$ se emplea la suavización

$$\rho_{p,\varepsilon}(r) = (r^2 + \varepsilon)^{p/2},$$

con $\varepsilon = 0.001$. Ambas formulaciones son diferenciables en todo $\mathbb{R}$ y convergen puntualmente a sus respectivas formas no suavizadas cuando $\varepsilon \rightarrow 0$. Los valores de ε adoptados siguen la recomendación de Robitzsch [11], quien muestra que $\varepsilon = 10^{-3}$ ofrece un buen equilibrio entre aproximación y estabilidad numérica para $p \leq 0.5$, mientras que L_0 requiere un suavizado mayor ($\varepsilon = 10^{-2}$) para evitar gradientes degenerados cerca del origen.

Para contrastar el desempeño de las distintas funciones de pérdida bajo cada condición experimental se empleó, en todos los experimentos, el test de rangos con signo de Wilcoxon aplicado de forma pareada entre cada par de pérdidas. Esta elección responde a tres características del diseño experimental. En primer lugar, dentro de cada réplica Monte Carlo

todas las funciones de pérdida se entrenan y evalúan sobre el mismo conjunto de datos y la misma partición entrenamiento/validación/prueba, de modo que las observaciones comparadas (Acc_{p_i} y Acc_{p_j} , o MSE_{p_i} y MSE_{p_j}) están naturalmente pareadas por réplica y no son independientes entre sí; el test de Wilcoxon está diseñado precisamente para este tipo de comparación pareada, a diferencia de pruebas para muestras independientes. En segundo lugar, no existe garantía de que la distribución del *accuracy* o del MSE a través de las réplicas sea aproximadamente normal (de hecho, los *boxplots* presentados en la Sección 3.3 muestran asimetrías y colas pesadas, particularmente bajo niveles altos de contaminación), lo que hace inapropiado un test paramétrico como el t pareado, cuya validez depende de dicho supuesto. El test de Wilcoxon, al basarse en los rangos de las diferencias pareadas y no en sus magnitudes, no requiere normalidad y resulta más apropiado en este contexto. En tercer lugar, el test de Wilcoxon presenta una menor sensibilidad a observaciones extremas que el test t pareado, al basarse en los rangos de las diferencias en lugar de sus magnitudes.

Todos los estudios de simulación se basan en $K = 100$ réplicas Monte Carlo por condición experimental (combinación de función de pérdida y nivel de contaminación).

3.2. Experimentos de Monte Carlo

3.2.1. Clasificación con ruido de etiquetas

El primer experimento replica la metodología de Ghosh et al. [9] en un problema de clasificación binaria bajo ruido simétrico de etiquetas. Se utiliza el *dataset load_digits* de scikit-learn, que contiene imágenes de dígitos escritos a mano de 8×8 píxeles (64 variables). Se considera el subconjunto de los dígitos 0 y 1, obteniendo aproximadamente 360 muestras balanceadas.

El ruido de etiquetas se introduce de forma simétrica, reemplazando, con probabilidad *rate*, cada etiqueta del conjunto de entrenamiento por una etiqueta de la clase contraria. Se consideran cuatro tasas de ruido de etiquetas: {0 %, 20 %, 40 %, 60 %}. El conjunto de prueba se mantiene con etiquetas correctas. La red tiene arquitectura [64, 256, 128, 64, 1] con *dropout* = 0,3, y se comparan funciones L_p para $p \in \{2, 1, 0,5, 0,25, 0\}$. El desempeño se mide mediante *accuracy* sobre el conjunto de prueba limpio, definida de la siguiente manera.

Sea $\{(x_i, y_i)\}_{i=1}^{n_{\text{test}}}$ el conjunto de prueba, con etiquetas $y_i \in \{0, 1\}$ no contaminadas, y sea $\hat{y}_i = \mathbb{I}\{f_\theta(x_i) \geq 0,5\}$ la clase predicha a partir de la salida sigmoide de la red f_θ . La exactitud (*accuracy*) sobre el conjunto de prueba se define como

$$\text{Acc} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbb{I}\{\hat{y}_i = y_i\}. \quad (3.1)$$

En ambos experimentos principales el entrenamiento se realizó con el optimizador Adam [6]. En clasificación se utilizó una tasa de aprendizaje $\eta = 5 \times 10^{-4}$ y regularización L_2 con penalización $\lambda = 10^{-4}$, un tamaño de mini-lote de 32 y un máximo de 300 épocas con parada temprana de paciencia 30 (criterio: pérdida de validación). La división de datos fue 64 %/16 %/20 % (entrenamiento/validación/prueba), con estratificación por clase. Al igual que en el experimento de regresión, la partición del conjunto de datos se realizó antes de introducir el ruido de etiquetas y previamente a cualquier procesamiento adicional, de modo

que no existe fuga de información entre los distintos subconjuntos: el ruido se inyecta únicamente sobre las etiquetas del conjunto de entrenamiento ya separado, y el conjunto de prueba permanece completamente libre de contaminación en todos los escenarios. Dado que los valores de píxel del conjunto `load_digits` se encuentran naturalmente acotados en el rango $[0, 16]$, no fue necesario aplicar una estandarización adicional a las variables predictoras.

3.2.2. Regresión con contaminación con outliers

En el segundo experimento se estudia el impacto de la función de pérdida en un problema de regresión afectado por contaminación con *outliers*. Se utiliza un *dataset* sintético generado mediante `make_regression` de la librería `scikit-learn`, compuesto por $N = 2000$ muestras con 15 variables numéricas, 10 de ellas realmente informativas, y ruido gaussiano de desviación estándar 0,5 (la semilla se fija para garantizar la reproducibilidad; véase el Anexo, Sección 5.2). Este diseño del experimento permite controlar exactamente la estructura del problema y ajustar la cantidad de contaminación de forma precisa.

La contaminación con *outliers* se inserta únicamente en el conjunto de entrenamiento, reemplazando una fracción de las respuestas y_i por valores desplazados según

$$y_i^* = y_i \pm 10 \hat{\sigma}_y,$$

donde el signo se elige de forma aleatoria con igual probabilidad, y $\hat{\sigma}_y$ es la desviación estándar de las respuestas en el conjunto de entrenamiento original, previo a la contaminación. Se consideran las siguientes tasas de contaminación con *outliers*: $\{0\%, 10\%, 20\%, 30\%\}$. El conjunto de prueba se mantiene limpio en todos los casos y la red utilizada tiene arquitectura $[15, 128, 64, 32, 1]$ con $dropout = 0,2$, y se entrena con cada función L_p con $p \in \{2, 1, 0,5, 0,25, 0\}$.

La elección de una magnitud de $10 \hat{\sigma}_y$ para desplazar las observaciones contaminadas no es arbitraria, sino que responde a las siguientes consideraciones metodológicas.

En primer lugar, esta magnitud es consistente con el modelo de contaminación de error grueso (*gross-error model*) de Hampel [4], subyacente al análisis de robustez desarrollado en la Sección 2.3: dicho marco concibe la contaminación como una masa puntual ubicada en una región de probabilidad esencialmente nula bajo la distribución generadora de los datos limpios, y no como una realización extrema pero plausible de dicha distribución. Bajo el supuesto aproximadamente gaussiano de la variable respuesta, razonable en este experimento dado que $y = X\beta + \eta$ con $\eta \sim \mathcal{N}(0, 0,5^2)$ y variables predictoras gaussianas, la probabilidad de observar un valor a más de 10 desviaciones estándar de la media es del orden de 10^{-23} , es decir, prácticamente nula bajo el proceso generador limpio. Esto garantiza que cada observación contaminada sea inequívocamente atribuible al mecanismo de contaminación introducido y no a la variabilidad natural del ruido gaussiano de desviación estándar 0,5 ya presente en el modelo.

En segundo lugar, al tratarse de un desplazamiento determinista y de magnitud fija, se evita que la severidad de la contaminación varíe de forma incidental entre réplicas Monte Carlo, lo que permite aislar el efecto de la *tasa* de contaminación, el factor experimental de interés, del efecto de su *magnitud*, que se mantiene constante en todos los escenarios.

La métrica de evaluación es el error cuadrático medio (MSE) calculado sobre el conjunto de prueba limpio. El optimizador utilizado fue Adam con $\eta = 10^{-3}$ y $\lambda = 10^{-4}$. El tamaño

de mini-lote fue de 64, con un máximo de 300 épocas y parada temprana de paciencia 25. La división fue aproximadamente 72 %/13 %/15 % (entrenamiento/validación/prueba). Tanto las variables predictoras como la variable respuesta y fueron estandarizadas a media cero y varianza unitaria; siguiendo la práctica estándar para evitar fugas de información (*data leakage*), los parámetros de estandarización se calcularon exclusivamente a partir del conjunto de entrenamiento, una vez separado de validación y prueba, y ya con la contaminación insertada, y la misma transformación, con esos parámetros fijos, se aplicó posteriormente a los conjuntos de validación y de prueba. De este modo, ninguna estadística calculada sobre los datos de validación o prueba influye en la transformación aplicada a dichos conjuntos.

3.3. Resultados

3.3.1. Clasificación: ruido de etiquetas

El objetivo de este experimento fue estudiar la robustez de diferentes funciones de pérdida ante la presencia de ruido de etiquetas en el conjunto de entrenamiento, siguiendo la metodología propuesta por Ghosh et al. [9]. El ruido de etiquetas constituye uno de los problemas más relevantes en el aprendizaje supervisado, dado que en la práctica las clases observadas suelen presentar errores de anotación, medición o clasificación humana. Desde un punto de vista estadístico, este fenómeno puede interpretarse como una forma de contaminación del mecanismo generador de datos, donde la variable observada deja de coincidir con la verdadera clase latente. Con esa finalidad, se consideraron funciones de pérdida dentro de la familia L_p con $p \in \{0, 0.25, 0.5, 1, 2\}$ bajo los niveles de ruido de etiquetas $\{0\%, 20\%, 40\%, 60\%\}$, midiendo el desempeño mediante el *accuracy* sobre un conjunto de prueba limpio.

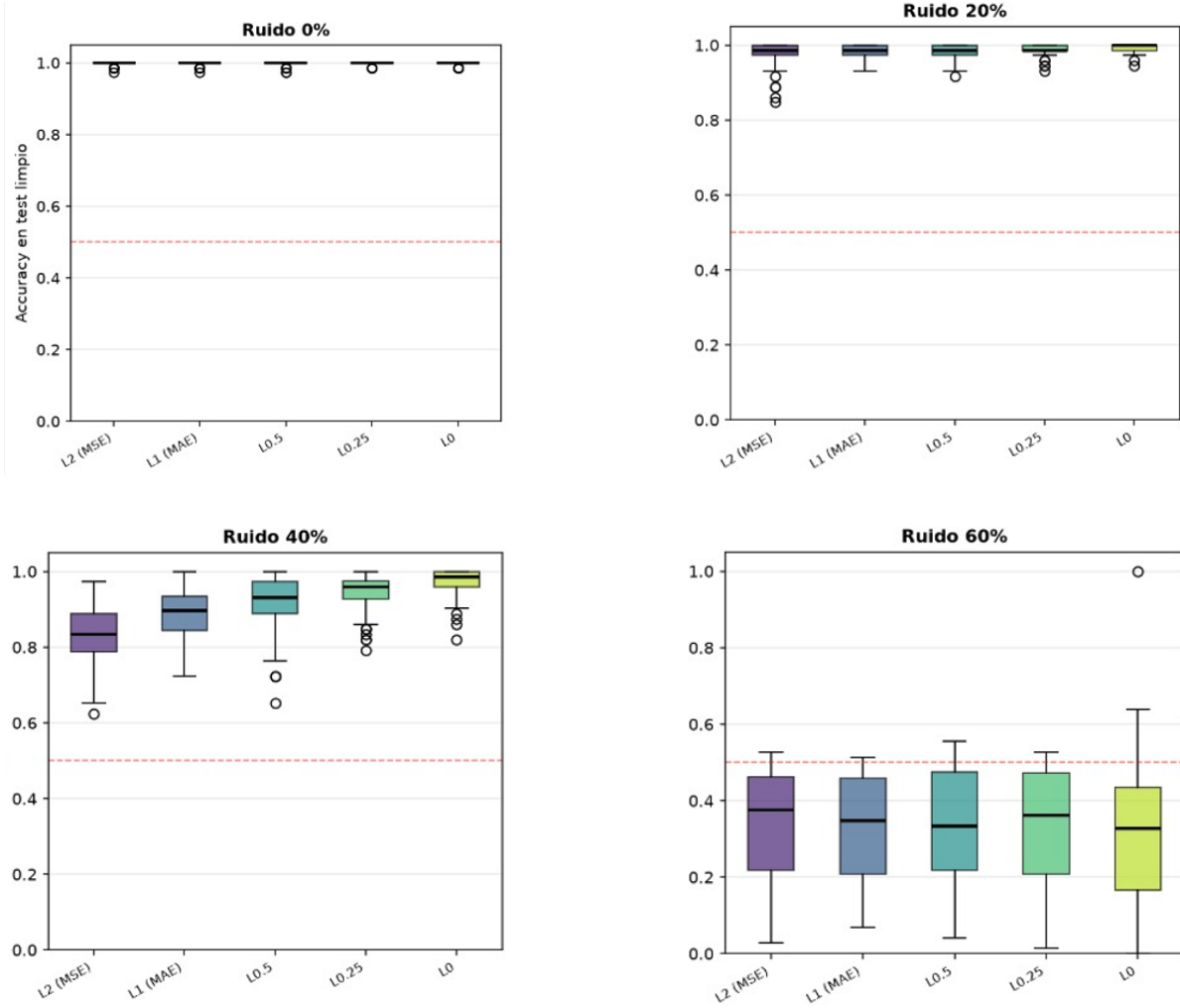


Figura 3.1: Distribucion del Accuracy en test.

La Figura 3.1 muestra la distribución completa de las 100 réplicas Monte Carlo para cada combinación de pérdida y nivel de ruido de etiquetas, permitiendo evaluar simultáneamente precisión y estabilidad. Cuando no existe ruido de etiquetas (0%), todas las pérdidas alcanzan prácticamente un desempeño perfecto, con promedios de $L_2 = 0,9985$, $L_1 = 0,9983$, $L_{0,5} = 0,9988$, $L_{0,25} = 0,9992$ y $L_0 = 0,9997$. En este caso, los *boxplots* muestran cajas extremadamente comprimidas cerca del valor máximo posible, indicando que el problema es fácilmente separable y que todas las funciones de pérdida son capaces de aprender adecuadamente la frontera de decisión; las diferencias observadas son tan pequeñas que carecen de relevancia práctica.

Cuando un quinto de las etiquetas son incorrectas (20% de ruido), en cambio, comienza a observarse una separación entre los métodos: los promedios pasan a ser $L_2 = 0,9775$, $L_1 = 0,9785$, $L_{0,5} = 0,9842$, $L_{0,25} = 0,9892$ y $L_0 = 0,9944$. Aunque todas las pérdidas siguen mostrando un *accuracy* elevado, empieza a emerger una leve jerarquía dominada por aquellas pérdidas con un menor valor de p , es decir $L_0 > L_{0,5} > L_1 > L_2$. Estas pérdidas con menor p presentan además una menor dispersión respecto a las clásicas, lo que indica mayor estabilidad entre réplicas. A medida que la proporción de etiquetas incorrectas se aproxima a

la mitad (ruido del 40 %), se obtienen desempeños promedio de 0,8276, 0,8846, 0,9196, 0,9429 y 0,9728 para L_2 , L_1 , $L_{0,5}$, $L_{0,25}$ y L_0 , respectivamente. La brecha entre las funciones deja de ser marginal y la diferencia en robustez se torna más notoria: L_0 conserva una exactitud cercana al 97 %, lo que implica que incluso cuando casi la mitad de las etiquetas son erróneas, el clasificador sigue recuperando adecuadamente la estructura verdadera del problema.

Finalmente, el escenario de 60 % de ruido de etiquetas constituye un caso extremo en el que más de la mitad de las observaciones poseen etiquetas incorrectas. En estas condiciones, todos los métodos sufren un deterioro importante, registrando *accuracies* promedio de $L_2 = 0,3406$, $L_1 = 0,3282$, $L_{0,5} = 0,3333$, $L_{0,25} = 0,3328$ y $L_0 = 0,3072$, todos por debajo de la línea de referencia del clasificador aleatorio (50 %). Nótese que en este régimen la jerarquía de robustez se invierte respecto a niveles de ruido menores: las pérdidas con menor p exhiben los desempeños promedio más bajos. Este comportamiento es consistente con la teoría: cuando la mayoría de las etiquetas son incorrectas, la señal correcta es minoritaria y los modelos que descartan agresivamente observaciones discrepantes terminan aprendiendo más de la señal errónea. En este régimen, la información estadística útil prácticamente desaparece y ningún método puede recuperar la frontera de decisión de manera confiable.

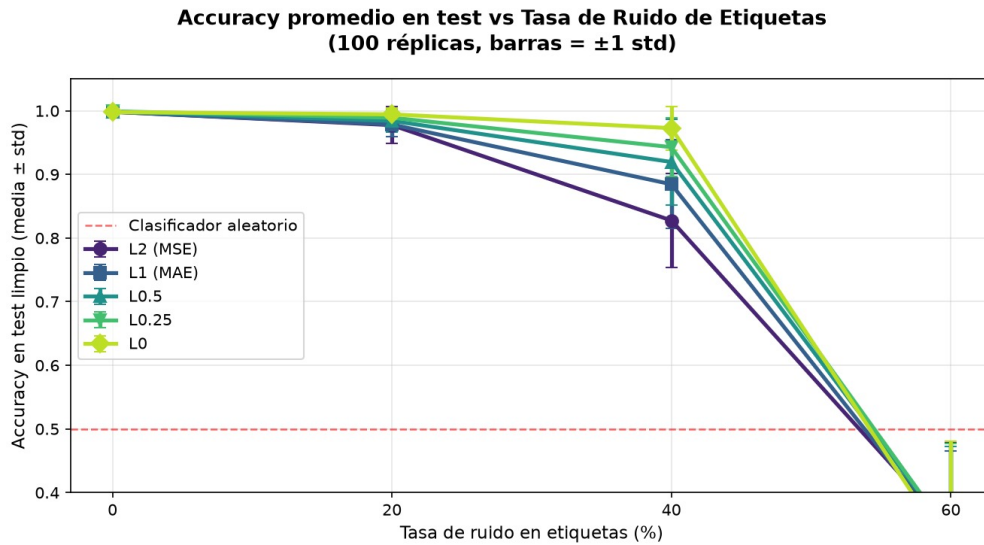


Figura 3.2: *Accuracy* promedio en prueba en función de la tasa de ruido de etiquetas.

La Figura 3.2 sintetiza esta evolución del *accuracy* promedio en función del nivel de ruido de etiquetas de manera continua, complementando los niveles discretos examinados en los *boxplots* y permitiendo distinguir tres fases del proceso. Durante una primera fase de ruido bajo (entre 0 % y 20 %), todas las curvas permanecen cercanas a uno y la degradación es mínima, lo que indica que los clasificadores poseen una capacidad natural para absorber pequeñas cantidades de ruido de etiquetas sin deteriorar significativamente su desempeño. En una segunda fase, de ruido moderado (entre 20 % y 40 %), las curvas comienzan a separarse progresivamente entre sí, en la misma dirección de la jerarquía ya identificada en los *boxplots*, lo que constituye evidencia visual adicional de la robustez diferencial entre pérdidas. Por último, al entrar en una fase de ruido extremo (alcanzando el 60 %), todas las curvas experimentan una caída abrupta, que puede interpretarse como una transición de fase estadística:

hasta cierto nivel de contaminación existe suficiente información correcta para identificar la estructura del problema, pero una vez superado un umbral crítico la mayoría de las etiquetas inducen información errónea y el aprendizaje se vuelve prácticamente imposible para todos los métodos.

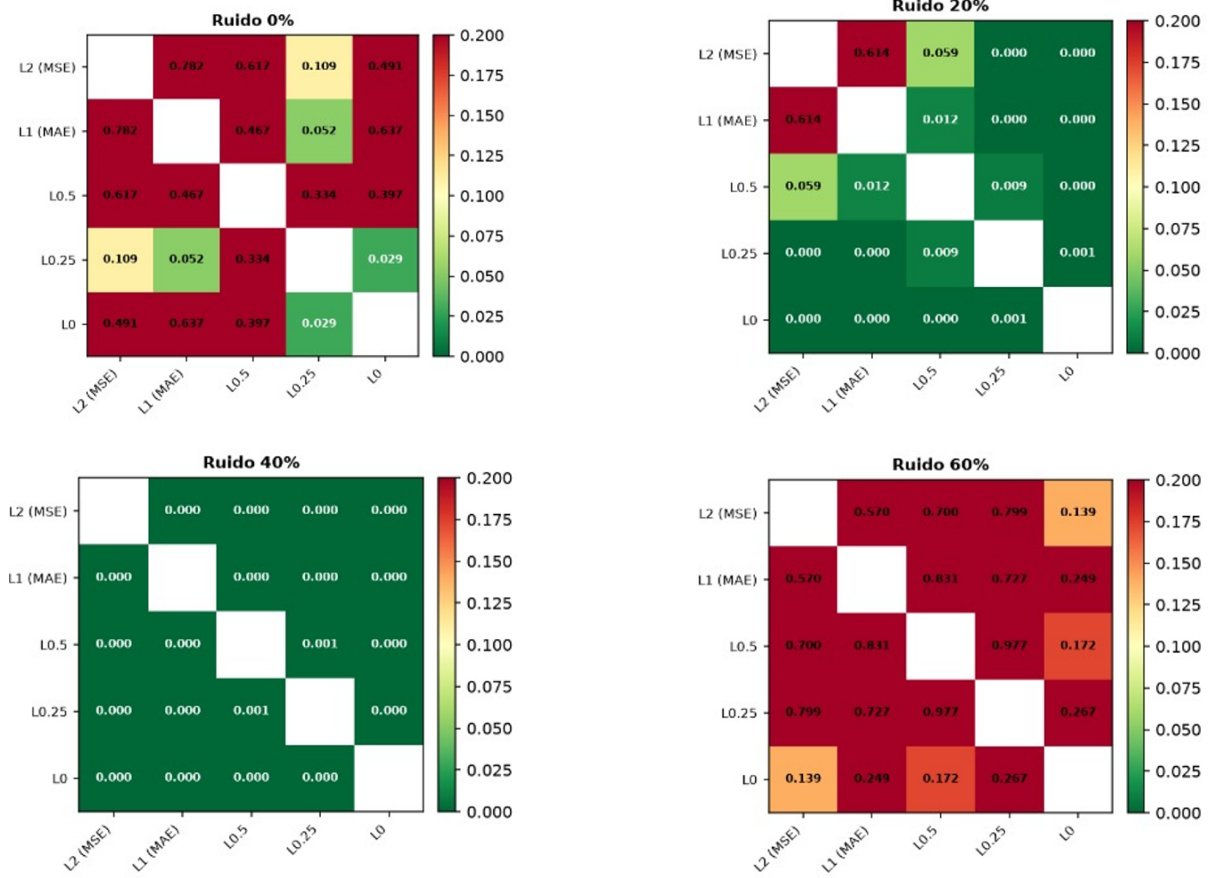


Figura 3.3: P-valores Test de Wilcox Pareado.

Para determinar si estas diferencias observadas son estadísticamente significativas o producto de la variabilidad Monte Carlo, se aplicó el test de rangos con signo de Wilcoxon pareado descrito en la Sección 3, cuyos resultados se ilustran en la Figura 3.3. La hipótesis nula evaluada es la igualdad de las distribuciones de *accuracy* para cada par de pérdidas, H_0 : la distribución de Acc_{p_i} coincide con la de Acc_{p_j} . Al analizar el escenario sin ruido de etiquetas, a excepción de $p(L_0, L_{0.25}) = 0,029$, todos los p -valores restantes resultan superiores a 0,05 (por ejemplo, $p(L_2, L_1) = 0,7815$, $p(L_2, L_{0.5}) = 0,6171$ y $p(L_{0.5}, L_0) = 0,3971$), de modo que no existe evidencia estadística suficiente para afirmar diferencias entre los distintos métodos, lo cual es esperable dado que el desempeño es prácticamente perfecto para las funciones en ausencia de contaminación.

Al incrementar el ruido de etiquetas al 20 %, en cambio, comienzan a aparecer más diferencias estadísticamente significativas: $p(L_2, L_{0.25}) = 0,0003$ y $p(L_2, L_0) < 0,0001$, lo que permite concluir que $L_{0.25}$ y L_0 superan significativamente a L_2 ; asimismo, $p(L_{0.25}, L_0) < 0,0001$ constituye evidencia estadística de la superioridad de L_0 sobre $L_{0.25}$ a este nivel de ruido.

Finalmente, bajo un ruido de etiquetas del 40 %, los p -valores son extremadamente pequeños, indicando que las diferencias en *accuracy* no son atribuibles al azar, y los resultados permiten ordenar estrictamente las funciones de pérdida por su desempeño como $L_0 > L_{0,5} > L_1 > L_2$. Esto evidencia empíricamente la teoría expuesta en el Capítulo 2 y sugiere de forma contundente que, dentro del rango de contaminación moderada estudiado, a mayor valor de p menor es la robustez ante el ruido de etiquetas.

3.3.2. Regresión: contaminación con outliers

El objetivo de este experimento fue evaluar el comportamiento de distintas funciones de pérdida L_p frente a la contaminación con *outliers* en problemas de regresión supervisada. Para ello se compararon las pérdidas L_2 (MSE), L_1 (MAE), $L_{0,5}$, $L_{0,25}$ y L_0 , mediante las aproximaciones diferenciables descritas en la Sección 3, utilizando redes neuronales profundas entrenadas sobre conjuntos de datos sintéticos contaminados con diferentes proporciones de *outliers*. La evaluación se realizó mediante un esquema de Monte Carlo con 100 réplicas independientes por condición experimental: en cada réplica se generó un nuevo conjunto de datos, se insertó contaminación con *outliers* en el conjunto de entrenamiento y posteriormente se evaluó el desempeño sobre un conjunto de prueba libre de contaminación, empleando como criterio de comparación el error cuadrático medio (MSE) en prueba.

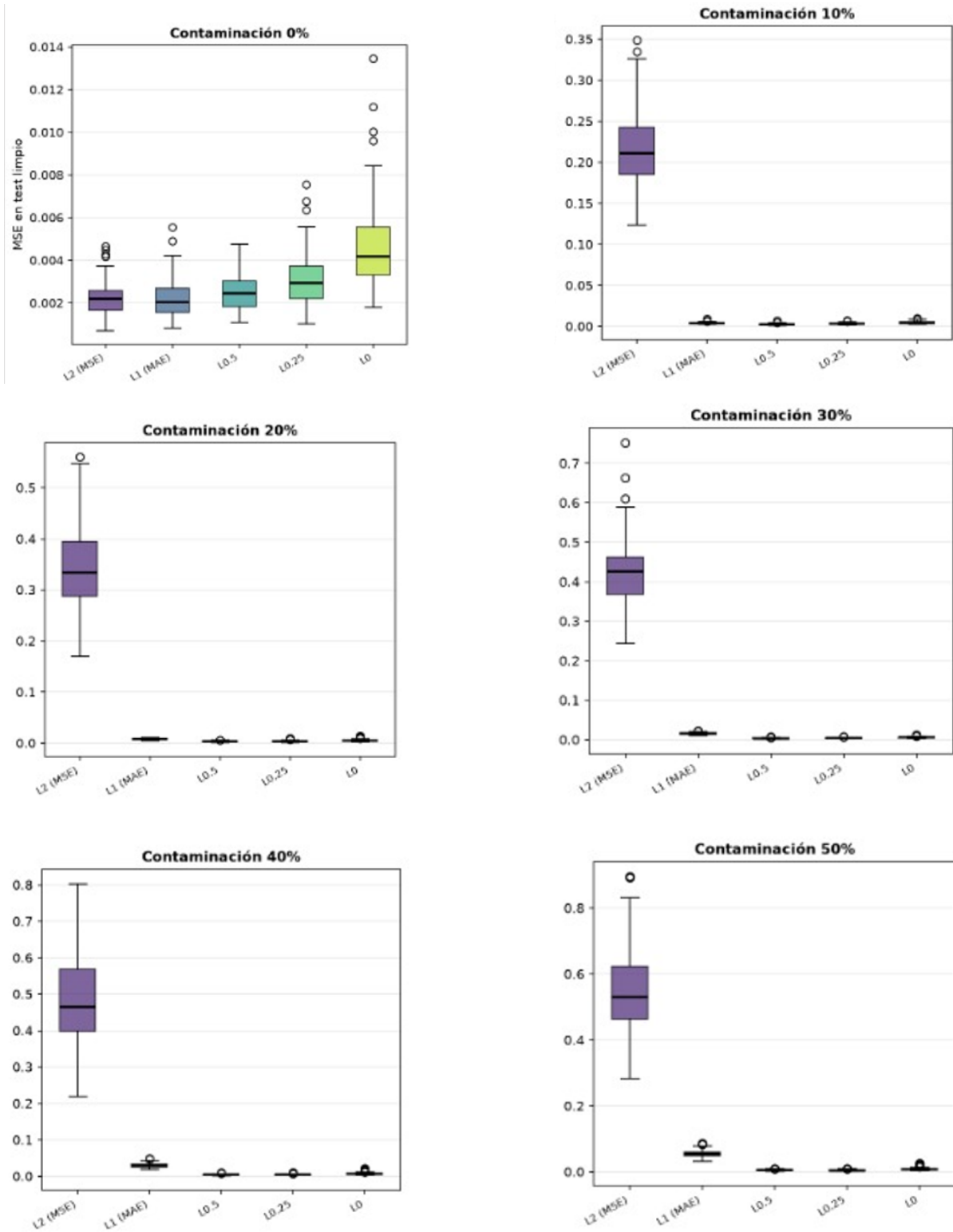


Figura 3.4: Distribucion MSE en test.

El análisis del desempeño, visualizado en la Figura 3.4, revela comportamientos contrastantes dependiendo del nivel de contaminación con *outliers*. En ausencia de contaminación, la

función de pérdida L_2 obtiene el mejor desempeño promedio, alcanzando un MSE medio de 0,00224, resultado consistente con la teoría clásica de regresión, ya que la pérdida cuadrática es óptima cuando los errores siguen distribuciones aproximadamente gaussianas y no existen observaciones extremas. Bajo este mismo escenario limpio, la pérdida L_1 presenta un desempeño prácticamente equivalente (0,00221), mientras que las pérdidas no convexas $L_{0,5}$ y $L_{0,25}$ muestran un resultado ligeramente inferior (errores cercanos a 0,0027). La aproximación a L_0 presenta, en este escenario limpio, el peor desempeño de toda la familia, con un MSE que duplica aproximadamente al de L_2 .

Este resultado contrasta de manera notoria con lo observado en el experimento de clasificación, donde L_0 resultó ser, en promedio, una de las pérdidas con mejor desempeño incluso en ausencia de contaminación. Esta discrepancia no cuenta con una explicación definitiva a partir de los resultados obtenidos en esta memoria, por lo que a continuación se plantean dos hipótesis plausibles y complementarias entre sí, que deben entenderse como posibles vías explicativas y no como conclusiones demostradas.

Una primera hipótesis considera la forma del gradiente de la aproximación suavizada $\phi_\varepsilon(r) = r^2/(r^2 + \varepsilon)$ empleada para L_0 . A diferencia de $\rho_{p,\varepsilon}$ con $p > 0$, cuyo gradiente crece de forma monótona (aunque saturada) con la magnitud del residuo, el gradiente de ϕ_ε es no monótono: se anula tanto cuando el residuo r es muy pequeño como cuando es muy grande, alcanzando su magnitud máxima únicamente en una vecindad intermedia de orden $\sqrt{\varepsilon}$. En el escenario de regresión sin contaminación, donde el objetivo del entrenamiento es precisamente reducir los residuos hacia valores cercanos a cero a medida que el modelo converge, esta forma funcional podría resultar desfavorable: conforme los residuos se acercan al régimen de alta precisión, el gradiente de ϕ_ε se atenúa fuertemente, lo que dificultaría el ajuste fino del modelo y podría dejar a la red estancada en una región subóptima del espacio de parámetros. En el experimento de clasificación, en cambio, el residuo relevante (la diferencia entre la salida sigmoide de la red y la etiqueta binaria) está naturalmente acotado en un rango reducido, más compatible con la escala fija utilizada para ϕ_ε , lo que podría evitar en mayor medida este problema de gradiente evanescente.

Una segunda hipótesis, no necesariamente excluyente de la anterior, es que el valor $\varepsilon = 0,01$ empleado para L_0 en ambos experimentos (seleccionado siguiendo la recomendación general de Robitzsch [11]) no constituya el valor óptimo para el contexto particular de esta regresión. Robitzsch [11] reporta precisamente que la estabilidad de L_0 depende críticamente de la elección de ε y del tamaño muestral disponible, documentando escenarios en los que $L_{0,5}$ resulta preferible a L_0 para tamaños muestrales moderados. Dado que la escala de los residuos de regresión en este experimento, tras la estandarización de la variable respuesta, difiere de la escala de los residuos binarios del experimento de clasificación, es plausible que un valor de ε distinto al aquí utilizado hubiese producido un comportamiento diferente para L_0 en regresión. En conjunto, ambas hipótesis sugieren que el desempeño de la aproximación a L_0 no depende únicamente de sus propiedades de robustez estadística frente a la contaminación, sino también, de forma no trivial, de la interacción entre la escala de los residuos del problema y el parámetro de suavizado ε utilizado durante la optimización.

Al incorporar un 10 % de contaminación con *outliers*, la dinámica cambia drásticamente: el MSE promedio de la pérdida L_2 se dispara desde 0,00224 hasta 0,21725, un aumento superior a cien veces, ya que los errores grandes se elevan al cuadrado y unos pocos *outliers* contribuyen de forma desproporcionada a la función objetivo, desviando el ajuste del mode-

lo. Las pérdidas robustas, por el contrario, mantienen errores muy cercanos a los del escenario limpio: $L_1 = 0,00403$, $L_{0,5} = 0,00268$ y $L_{0,25} = 0,00322$, $L_0 = 0,00467$ y los *boxplots* evidencian además una dispersión considerablemente menor para estas pérdidas, indicando un comportamiento robusto consistente entre réplicas. A medida que la contaminación con *outliers* aumenta al 20 % y 30 %, el rendimiento de L_2 continúa disminuyendo de forma sostenida ($MSE_{L_2} = 0,34079$ y $0,42081$, respectivamente), creciendo aproximadamente en un orden de magnitud por cada 10 % de contaminación añadida. En contraparte, las pérdidas $L_{0,5}$ y $L_{0,25}$ permanecen notablemente estables, pasando de 0,00308 a 0,00326 y de 0,00354 a 0,00370, respectivamente, y sus cajas mantienen una amplitud reducida incluso bajo estos altos niveles de contaminación, lo que subraya no solo un mejor desempeño promedio, sino una acotada variabilidad experimental.

Al alcanzar niveles de contaminación con *outliers* del 40 % y 50 %, el desempeño de L_2 continúa deteriorándose, llegando a valores promedio de 0,48161 y 0,54108, respectivamente. Le sigue en desempeño la función L_1 , con un valor promedio de 0,05458 bajo un 50 % de contaminación, mientras que las tres pérdidas restantes ($L_{0,5}$, $L_{0,25}$ y L_0) presentan valores inferiores a 0,01 en todos los casos.

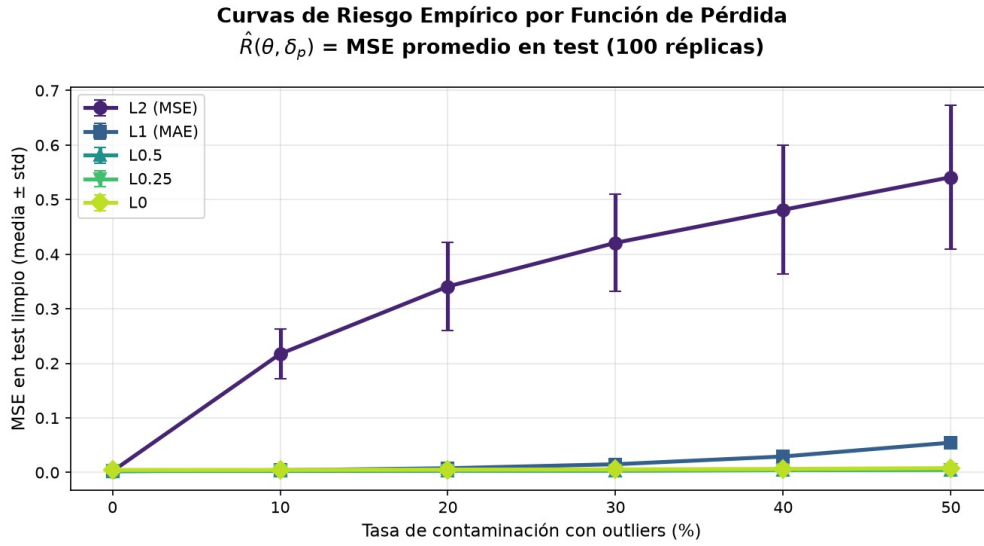


Figura 3.5: Curvas de riesgo empírico estimado mediante MSE en prueba en función de la tasa de contaminación con *outliers*.

La Figura 3.5 presenta las curvas de riesgo de generalización estimado mediante MSE sobre datos limpios, formalmente

$$\hat{R}_{MC}(p) = \frac{1}{50} \sum_{k=1}^{50} MSE_k(p),$$

donde $MSE_k(p)$ corresponde al error obtenido por la pérdida L_p en la réplica k . Este gráfico refuerza lo discutido anteriormente: conforme aumenta el nivel de contaminación con *outliers*, el riesgo de generalización para la pérdida L_2 crece de forma acelerada, evidenciando su extrema sensibilidad ante dicha contaminación, mientras que las curvas correspondientes a las

pérdidas robustas se mantienen prácticamente horizontales, confirmando la escasa influencia que ejerce la contaminación con *outliers* sobre su desempeño predictivo.

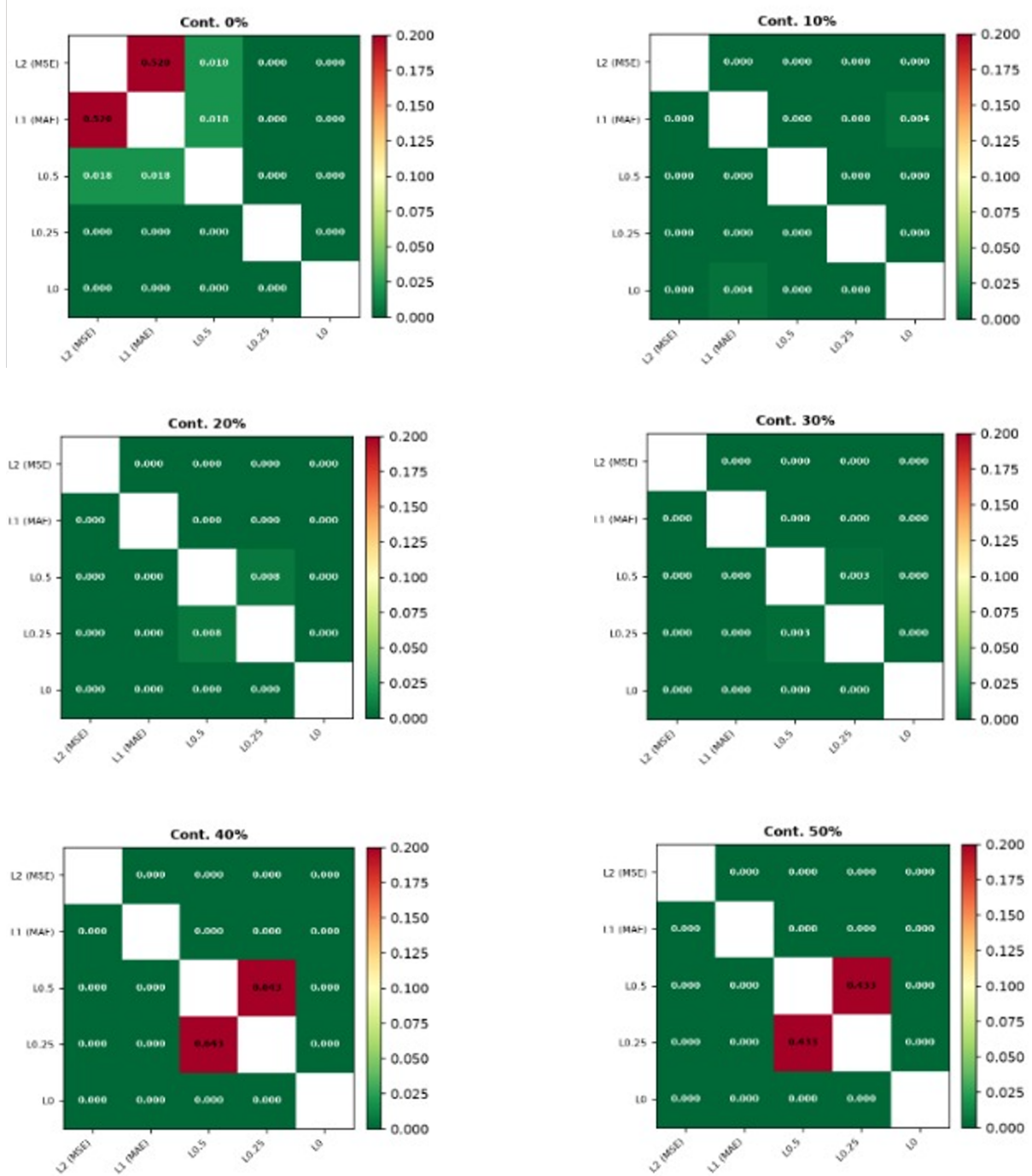


Figura 3.6: P-valores Test de Wilcoxon Pareado.

La Figura 3.6 resume los resultados del test de Wilcoxon pareado aplicado a cada par de funciones de pérdida, evaluando la hipótesis nula de igualdad funcional, H_0 : la distribución de MSE_{p_i} coincide con la de MSE_{p_j} , frente a la alternativa de que difieren. En ausencia de

contaminación, las comparaciones entre L_2 y L_1 arrojan un p -valor de 0,5202, por lo que no existe evidencia estadística suficiente para afirmar diferencias significativas entre estos pares, confirmando empíricamente que L_2 y L_1 ofrecen desempeños equivalentes sin contaminación.

En presencia de contaminación con *outliers*, en cambio, todas las comparaciones presentan $p < 0,001$, lo que indica diferencias significativas entre todas las funciones, a excepción de los casos con 40 % y 50 % de contaminación, donde la comparación entre las pérdidas $L_{0,5}$ y $L_{0,25}$ arroja p -valores superiores al umbral de significancia habitual (0,6425 y 0,4331, respectivamente). En conclusión, no hay evidencia para afirmar que una de estas dos pérdidas supere a la otra, lo que sugiere la existencia de una región robusta $p \in (0,25, 0,5)$ dentro de la cual el comportamiento predictivo frente a la contaminación con *outliers* se vuelve esencialmente idéntico.

3.3.3. Regresión con datos reales: el conjunto Airfoil Self-Noise

Los dos experimentos anteriores emplearon conjuntos de datos sintéticos y semisintéticos, lo que permitió controlar con precisión el mecanismo de contaminación pero deja abierta la pregunta, retomada como limitación en el Capítulo 4, de si la jerarquía de robustez observada se mantiene sobre datos reales. Para explorar esta cuestión se replicó el diseño experimental de la Sección 3 sobre el conjunto de datos real Airfoil Self-Noise de la librería ucimlrepo de Python, empleando el mismo esquema de contaminación con *outliers* en la variable respuesta, las mismas cinco funciones de pérdida L_p con $p \in \{2, 1, 0,5, 0,25, 0\}$, las tasas de contaminación $\{0 \%, 10 \%, 20 \%, 30 \%, 40 \%, 50 \%\}$ y el mismo esquema de evaluación mediante MSE en prueba y test de Wilcoxon pareado sobre 100 réplicas Monte Carlo por condición.

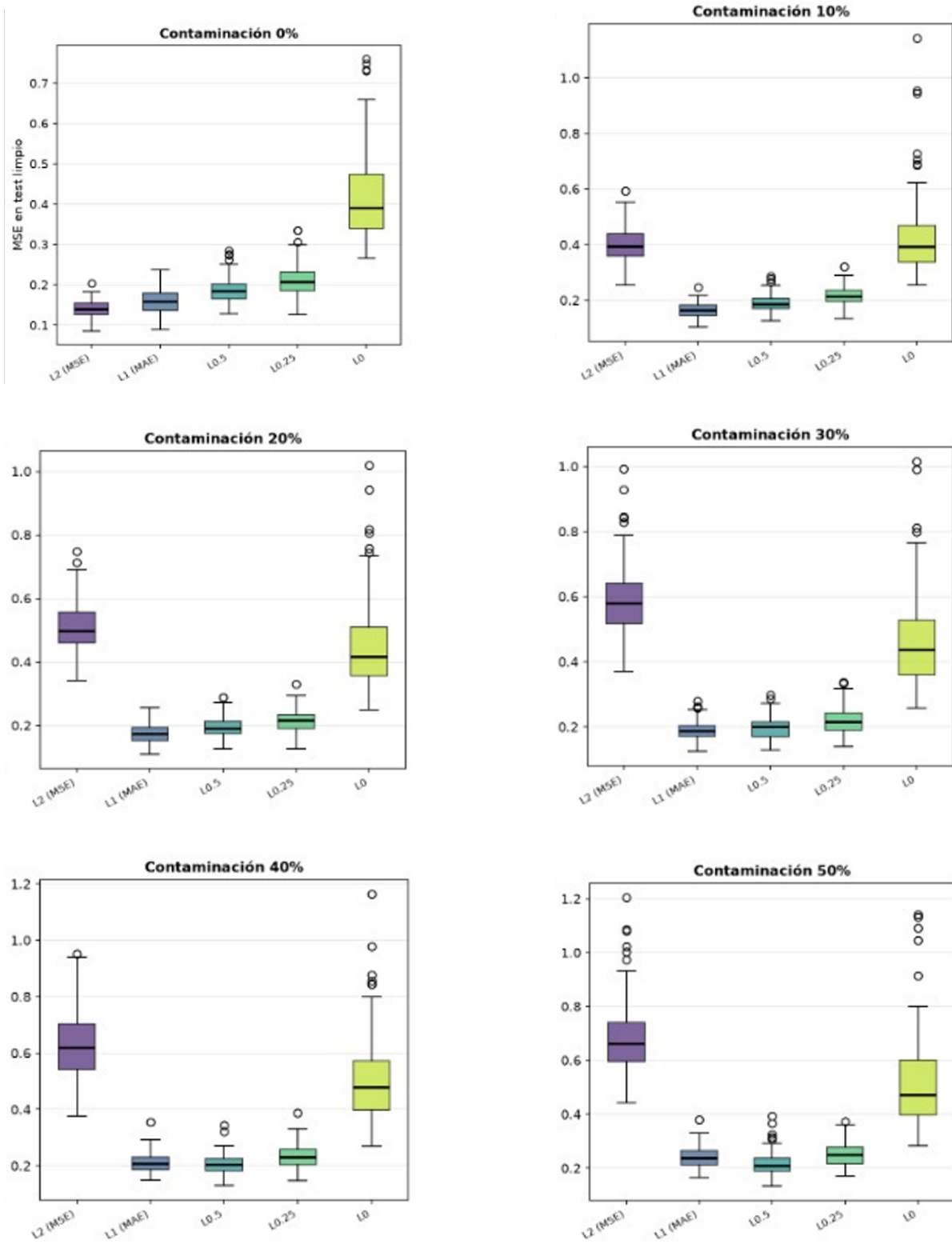


Figura 3.7: Distribución del MSE en test Airfoil Self-Noise.

La Figura 3.7 presenta la distribución del MSE en prueba obtenida en este experimento. En ausencia de contaminación, L_2 vuelve a presentar el mejor desempeño, con un MSE

promedio de 0,1394, seguido por L_1 , $L_{0,5}$, $L_{0,25}$ y L_0 , con valores de 0,1571, 0,1860, 0,2117 y 0,4217, respectivamente; todas las diferencias entre pares resultan estadísticamente significativas ($p < 0,001$). Este resultado replica sobre datos reales el patrón ya observado en el experimento sintético de la Sección anterior: sin contaminación, L_2 (utilizada además como métrica de evaluación) logra el menor error, y L_0 presenta el desempeño más pobre de la familia, aquí con un MSE aproximadamente tres veces mayor que el de L_2 , de modo que en condiciones ideales no se observa ventaja alguna en reducir p .

Al introducir un 10 % de contaminación, el MSE de L_2 aumenta de 0,1394 a 0,4000, mientras que L_1 apenas se mueve hasta 0,1644, y $L_{0,5}$ y $L_{0,25}$ permanecen igualmente estables en 0,1881 y 0,2168. El test de Wilcoxon confirma diferencias significativas entre L_1 , $L_{0,5}$ y $L_{0,25}$, pero no entre L_2 y L_0 ($p = 0,6599$) pese a la diferencia observada entre sus medias, lo que subraya nuevamente que las medias por sí solas no bastan para establecer diferencias entre métodos y justifica el uso de comparaciones pareadas sobre las réplicas. Con un 20 % de contaminación la ventaja de las funciones robustas se acentúa: L_2 alcanza 0,5112 frente a 0,1732, 0,1943 y 0,2143 para L_1 , $L_{0,5}$ y $L_{0,25}$, mientras que L_0 se sitúa en 0,4519, sin resultar competitiva; todas las comparaciones son significativas ($p < 0,001$), de modo que L_1 constituye la alternativa con mejor desempeño a este nivel. El patrón se mantiene con un 30 % de contaminación, donde L_2 continúa deteriorándose hasta 0,5914 y L_1 obtiene el menor error (0,1885), seguida por $L_{0,5}$ (0,1972) y $L_{0,25}$ (0,2198); todas las diferencias resultan significativas, aunque la que separa a L_1 de $L_{0,5}$ es comparativamente pequeña en magnitud ($p = 0,0009$).

A partir de un 40 % de contaminación se observa un cambio en la función con menor MSE: $L_{0,5}$ pasa a liderar con 0,2069, muy cerca de L_1 (0,2103), diferencia que aquí deja de ser significativa ($p = 0,0968$), mientras que las diferencias respecto de L_2 , $L_{0,25}$ y L_0 sí lo son. Este nivel puede interpretarse como una zona de transición en la que las pérdidas no convexas comienzan a mostrar una ventaja más clara frente a L_1 . Finalmente, bajo un 50 % de contaminación $L_{0,5}$ mantiene el menor MSE promedio (0,2172), seguida por L_1 (0,2412) y $L_{0,25}$ (0,2502), mientras que L_0 (0,5234) y en particular L_2 (0,6840) presentan los peores desempeños; en este escenario todas las comparaciones por pares resultan estadísticamente significativas.

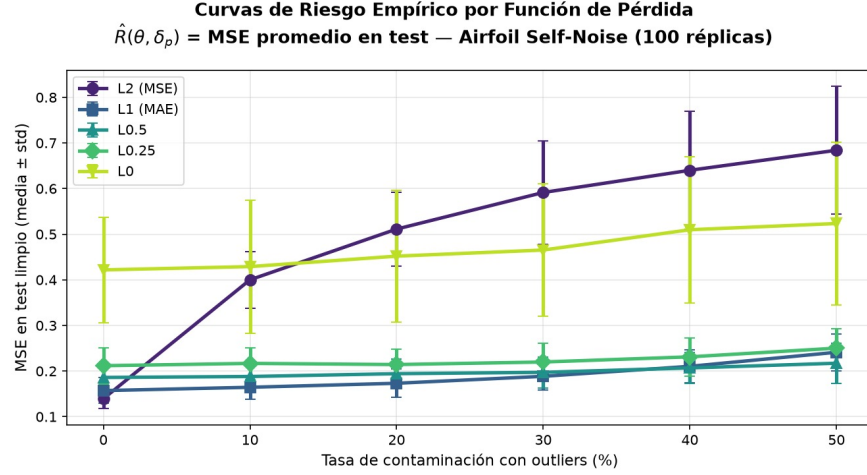


Figura 3.8: Curvas de riesgo empírico estimado mediante MSE en prueba en función de la tasa de contaminación con *outliers*, conjunto Airfoil Self-Noise.

La Figura 3.8 resume esta evolución mediante las curvas de riesgo empírico. En términos de variabilidad, L_2 exhibe un aumento progresivo de su desviación estándar, desde 0,0218 sin contaminación hasta 0,1399 con un 50 % de contaminación, y L_0 presenta una dispersión igualmente elevada en todos los escenarios, alcanzando 0,1785 en el nivel máximo; en contraste, L_1 , $L_{0,5}$ y $L_{0,25}$ mantienen desviaciones estándar considerablemente menores, lo que se traduce en cajas más compactas y una menor influencia de valores extremos en todos los niveles de contaminación.

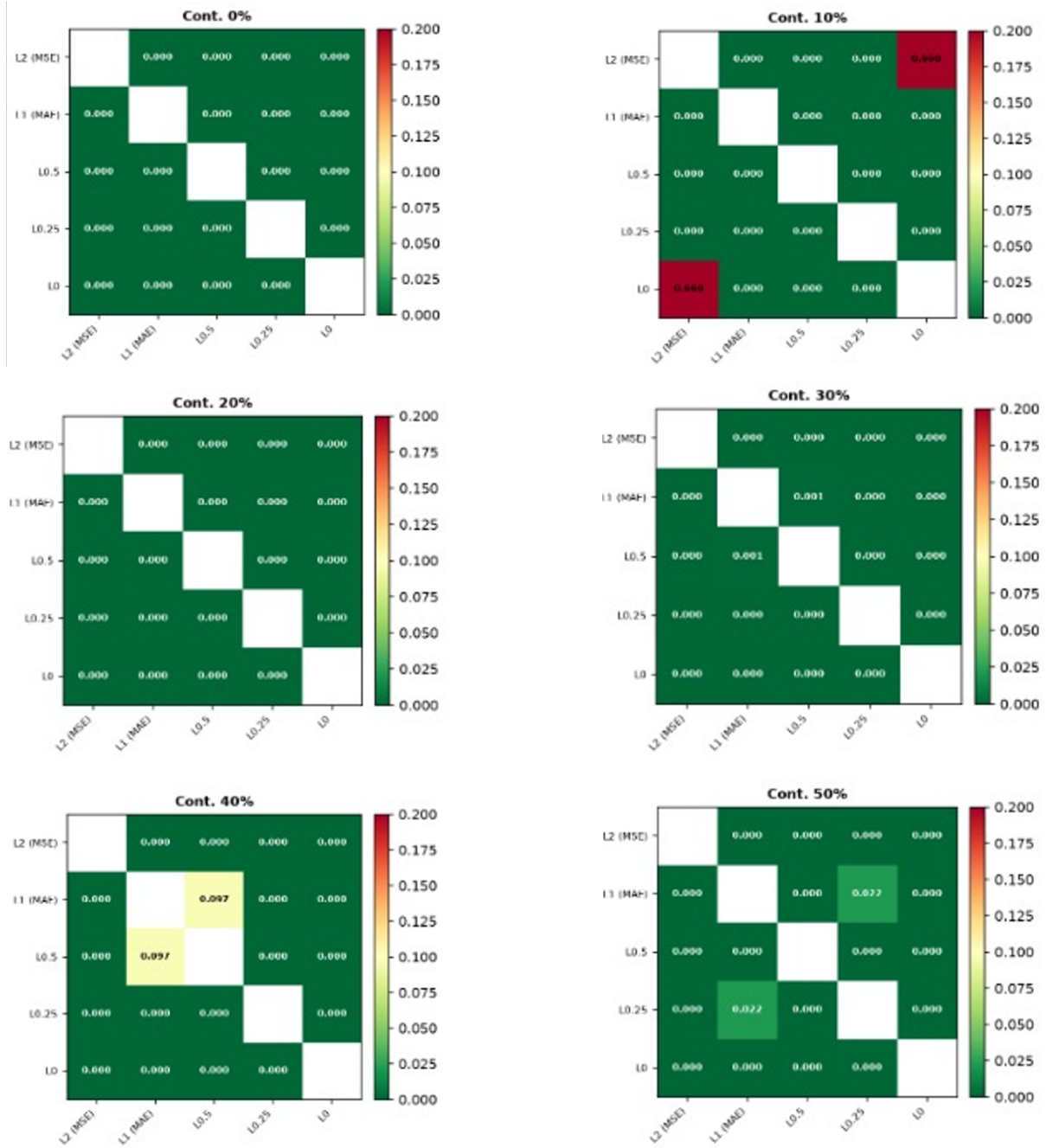


Figura 3.9: P-valores del test de Wilcoxon pareado Airfoil Self-Noise.

La Figura 3.9 presenta los p -valores del test de Wilcoxon pareado para cada nivel de contaminación. El comportamiento global puede resumirse a través del cambio en la función con menor MSE a medida que aumenta la contaminación: L_2 resulta óptima sin contaminación, L_1 domina entre el 10 % y el 30 %, y $L_{0.5}$ se convierte en la alternativa preferible a partir del 40 %. A diferencia del experimento sintético, donde la región de indistinguibilidad estadística se ubicó entre $L_{0.5}$ y $L_{0.25}$, en el conjunto Airfoil Self-Noise dicha región se desplaza hacia el par L_1 - $L_{0.5}$ en los niveles intermedios y altos de contaminación, y $L_{0.25}$ no logra en ningún escenario superar a L_1 ni a $L_{0.5}$. En conjunto, este tercer experimento confirma sobre

datos reales el compromiso general entre desempeño bajo datos limpios y robustez frente a la contaminación ya documentado en el experimento sintético: L_2 es preferible sin *outliers* pero se deteriora rápidamente en su presencia, mientras que L_1 y $L_{0,5}$ ofrecen el mejor equilibrio entre precisión y robustez a medida que la contaminación aumenta, sin que la reducción del parámetro p produzca una mejora monótona del desempeño.

Capítulo 4

Conclusiones y Trabajo Futuro

Esta memoria abordó el estudio de la familia de funciones de pérdida L_p en redes neuronales profundas desde una doble perspectiva: teórica, a través de la teoría estadística de la decisión y la estadística robusta, y experimental, mediante simulaciones de Monte Carlo en problemas de regresión y clasificación bajo contaminación controlada de los datos. Los resultados obtenidos permiten dar respuesta a los objetivos planteados en el Capítulo 2. La caracterización analítica de la familia L_p dentro del marco de Wald [7] y Berger [1] mostró que cada valor de p induce un estimador de Bayes distinto (la media para $p = 2$, la mediana para $p = 1$ y la moda para $p \rightarrow 0$) y que estas diferencias no son meramente formales, sino que se traducen en funciones de influencia con propiedades marcadamente distintas: no acotada y lineal para L_2 , acotada para L_1 , y redescendiente para la aproximación a L_0 . Esta jerarquía teórica, reforzada por el resultado de Barron [12] sobre pérdidas de forma continua, anticipó de manera consistente el comportamiento observado experimentalmente.

Los experimentos de regresión con contaminación con *outliers* confirmaron de manera contundente la fragilidad de L_2 ante este tipo de contaminación: su error de generalización se multiplicó por más de cien veces al introducir un 10 % de *outliers*, mientras que las pérdidas no convexas $L_{0,5}$ y $L_{0,25}$ permanecieron prácticamente estables incluso bajo un 30 % de contaminación. Las pruebas de Wilcoxon pareadas mostraron además evidencia de una región de robustez estadísticamente indistinguible para $p \in (0,25, 0,5)$, sugiriendo que, dentro de cierto rango, el valor exacto del exponente resulta menos relevante que el hecho de que la pérdida sea no convexa. En ausencia de contaminación, sin embargo, L_2 y L_1 resultaron óptimas, y las pérdidas no convexas exhibieron una ligera pérdida de eficiencia estadística; en particular, la aproximación a L_0 mostró en ese escenario limpio el peor desempeño de toda la familia, un resultado que, según se discutió en el Capítulo 3, podría atribuirse al carácter no monótono del gradiente de su aproximación suavizada ϕ_ε y a su interacción con la escala de los residuos de regresión. Este compromiso entre robustez y eficiencia replica, en el dominio de la regresión, el fenómeno documentado por Ma et al. [16] para clasificación con entropía cruzada. El experimento adicional sobre el conjunto real Airfoil Self-Noise (Sección 3.3.3) mostró que este mismo compromiso persiste sobre datos reales, aunque con matices: allí L_1 , y no $L_{0,5}$ o $L_{0,25}$, domina en el rango intermedio de contaminación, y la región de indistinguibilidad estadística se desplaza hacia el par L_1 - $L_{0,5}$.

El experimento de clasificación con ruido de etiquetas, por su parte, replicó y extendió la metodología de Ghosh et al. [9], confirmando que la jerarquía de robustez $L_0 > L_{0,5} > L_1 > L_2$

se mantiene de forma estadísticamente significativa para niveles de ruido moderados (20 % y 40 %). No obstante, se identificó un fenómeno no anticipado explícitamente en la literatura revisada: cuando el ruido de etiquetas supera el 50 %, la jerarquía de robustez se invierte, y las pérdidas con menor p , que descartan más agresivamente las observaciones discrepantes, terminan aprendiendo preferentemente de la señal mayoritaria, que en ese régimen es la señal incorrecta. Este resultado matiza la recomendación general de preferir siempre valores pequeños de p ante la sospecha de contaminación, y sugiere que la robustez de las pérdidas L_p con p pequeño es beneficiosa únicamente mientras la señal correcta permanece mayoritaria en los datos de entrenamiento. En conjunto, la evidencia teórica y experimental reunida en esta memoria confirma que la elección de la función de pérdida no es una decisión secundaria del proceso de entrenamiento, sino una decisión estadística con consecuencias directas sobre la robustez, la eficiencia y la validez de las conclusiones que se extraen de un modelo entrenado sobre datos potencialmente contaminados.

Estos hallazgos deben interpretarse, sin embargo, a la luz de las limitaciones del diseño experimental empleado. En primer lugar, los experimentos principales se basaron en conjuntos de datos sintéticos y en el conjunto `load_digits`, lo que permitió controlar con precisión el mecanismo de contaminación. El experimento adicional sobre el conjunto real Airfoil Self-Noise (Sección 3.3.3) ofrece una primera evidencia de que el compromiso entre desempeño bajo datos limpios y robustez frente a *outliers* se replica en datos reales de regresión, aunque queda abierta la pregunta de si la jerarquía de robustez observada se mantiene también en clasificación con ruido de etiquetas real, y en conjuntos de datos de mayor escala y complejidad, donde tanto el ruido de etiquetas como la contaminación con *outliers* surgen de procesos de recolección y anotación imperfectos en lugar de ser inducidos experimentalmente. En segundo lugar, el parámetro p se trató en todo momento como un hiperparámetro fijo, seleccionado mediante una búsqueda sobre una grilla discreta; Barron [12] muestra que tratar el parámetro de forma como una variable libre, optimizada junto con los pesos de la red mediante la maximización de la verosimilitud, permite que la robustez de la pérdida se adapte automáticamente al nivel de contaminación presente en los datos, y esta perspectiva adaptativa no fue explorada en la presente memoria. Asimismo, el marco de *Active Passive Loss* propuesto por Ma et al. [16], que combina una pérdida robusta con una componente complementaria que preserva la capacidad de ajuste, sugiere una vía concreta para mitigar la pérdida de eficiencia observada en esta memoria para $L_{0.5}$ y L_0 en ausencia de contaminación, tanto en clasificación como en regresión, mientras que la estrategia de regularización ℓ_p sobre la salida de la red propuesta por Zhou et al. [13] ofrece una vía alternativa para incorporar la noción de dispersión asociada a L_0 sin necesidad de emplearla directamente como función de pérdida sobre los residuos. Finalmente, esta memoria se restringió a dos formas de contaminación (*outliers* en la variable respuesta y ruido de etiquetas), dejando fuera otros regímenes relevantes, como los errores de especificación del modelo estudiados por Robitzsch [11], y se apoyó en las cotas de convergencia de Rosasco et al. [8], que solo cubren formalmente el caso convexo ($p \geq 1$) y dejan sin garantías teóricas explícitas a las pérdidas L_p con $p < 1$ utilizadas en gran parte de los experimentos; el desarrollo de cotas de generalización específicas para esta región no convexa constituye un problema teórico abierto de relevancia directa para justificar con mayor rigor los resultados empíricos aquí presentados. Abordar estas limitaciones, mediante un mayor número de réplicas, conjuntos de datos reales de clasificación, robustez adaptativa, esquemas de tipo *Active Passive Loss* o regularización

dispersa, escenarios de mala especificación del modelo, y fundamentos teóricos para pérdidas no convexas, constituye el conjunto de líneas de trabajo futuro más directamente motivadas por los resultados de esta memoria.

Bibliografía

- [1] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springer, New York, 1985.
- [2] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.
- [3] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, Cambridge, MA, 2016.
- [4] F. R. Hampel, *The Influence Curve and Its Role in Robust Estimation*, Journal of the American Statistical Association, vol. 69, no. 346, pp. 383–393, 1974.
- [5] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, New York, 2009.
- [6] D. P. Kingma, J. Ba, *Adam: A Method for Stochastic Optimization*, Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2015.
- [7] A. Wald, *Statistical Decision Functions*, Wiley, New York, 1950.
- [8] L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, A. Verri, *Are Loss Functions All the Same?*, Technical Report DISI-TR-03-07, University of Genova, 2003.
- [9] A. Ghosh, H. Kumar, P. S. Sastry, *Robust Loss Functions under Label Noise for Deep Neural Networks*, Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-17), pp. 1919–1925, 2017.
- [10] J. Qi, J. Du, S. M. Siniscalchi, X. Ma, C.-H. Lee, *On Mean Absolute Error for Deep Neural Network Based Vector-to-Vector Regression*, IEEE Signal Processing Letters, vol. 27, pp. 1485–1489, 2020. DOI: 10.1109/LSP.2020.3016837.
- [11] A. Robitzsch, *L_0 and L_p Loss Functions in Model-Robust Estimation of Structural Equation Models*, Psych, vol. 5, pp. 1122–1139, 2023.
- [12] J. T. Barron, *A General and Adaptive Robust Loss Function*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4331–4339, 2019.
- [13] X. Zhou, X. Liu, C. Wang, D. Zhai, J. Jiang, X. Ji, *Learning with Noisy Labels via Sparse Regularization*, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 72–81, 2021.

- [14] L. Feng, S. Shu, Z. Lin, F. Lv, L. Li, B. An, *Can Cross Entropy Loss Be Robust to Label Noise?*, Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20), 2020.
- [15] Y. Liu, H. Guo, *Peer Loss Functions: Learning from Noisy Labels without Knowing Noise Rates*, Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR vol. 119, 2020.
- [16] X. Ma, H. Huang, Y. Wang, S. Romano, S. Erfani, J. Bailey, *Normalized Loss Functions for Deep Learning with Noisy Labels*, Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR vol. 119, 2020.
- [17] D. E. Rumelhart, G. E. Hinton, R. J. Williams, *Learning Representations by Back-Propagating Errors*, Nature, vol. 323, pp. 533–536, 1986.
- [18] K. Hornik, M. Stinchcombe, H. White, *Multilayer Feedforward Networks are Universal Approximators*, Neural Networks, vol. 2, no. 5, pp. 359–366, 1989.
- [19] V. Nair, G. E. Hinton, *Rectified Linear Units Improve Restricted Boltzmann Machines*, Proceedings of the 27th International Conference on Machine Learning (ICML), pp. 807–814, 2010.
- [20] S. Ioffe, C. Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*, Proceedings of the 32nd International Conference on Machine Learning (ICML), PMLR vol. 37, pp. 448–456, 2015.
- [21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*, Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.

Capítulo 5

Anexo

5.1. Implementación computacional

La implementación se realizó en Python 3.11 usando PyTorch como *framework* de aprendizaje profundo. Las librerías auxiliares son `scikit-learn` para generación de datos y pre-procesamiento, y `matplotlib` para visualización.

Las funciones de pérdida se implementan como subclases de `nn.Module`, lo que permite que PyTorch calcule automáticamente los gradientes mediante diferenciación automática (*autograd*), sin necesidad de derivar manualmente las expresiones del gradiente de cada $\rho_{p,\varepsilon}$.

```
1 class LpLoss(nn.Module):
2     """
3     Familia de funciones de perdida Lp suavizadas.
4
5     Para p > 0: rho_{p,eps}(r) = mean( (r^2 + eps)^(p/2) )
6     Para p = 0: phi_eps(r)      = mean( r^2 / (r^2 + eps) )
7
8     Parametro eps automatico segun p:
9         p > 0 -> eps = 1e-3 (Robitzsch 2023)
10        p = 0 -> eps = 1e-2 (Robitzsch 2023, Seccion 6.2)
11    """
12    def __init__(self, p: float, eps: float = None):
13        super().__init__()
14        self.p = p
15        self.eps = 1e-2 if p == 0.0 else 1e-3
16
17    def forward(self, y_pred, y_true):
18        r = y_pred - y_true
19        if self.p == 0.0:
20            return torch.mean(r**2 / (r**2 + self.eps))
21        else:
22            return torch.mean((r**2 + self.eps) ** (self.p / 2))
```

Listing 5.1: Implementación de la familia L_p en PyTorch

La arquitectura de red utilizada en ambos experimentos es un perceptrón multicapa profundo (*Deep MLP*) con capas ocultas ReLU, normalización por lotes (`BatchNorm1d`) y regularización por *dropout*. La normalización por lotes es especialmente importante aquí porque

distintas funciones de pérdida producen distribuciones de activaciones distintas durante el entrenamiento; sin normalización, la comparación entre funciones estaría confundida con efectos de escala. La salida es lineal para regresión y sigmoide para clasificación binaria.

```

1 class DeepMLP(nn.Module):
2     def __init__(self, layer_sizes, task='regression',
3                 dropout=0.2):
4         super().__init__()
5         self.task = task
6         layers = []
7         for i in range(len(layer_sizes) - 1):
8             layers.append(nn.Linear(layer_sizes[i],
9                                     layer_sizes[i+1]))
10            is_last = (i == len(layer_sizes) - 2)
11            if not is_last:
12                layers.append(nn.BatchNorm1d(layer_sizes[i+1]))
13                layers.append(nn.ReLU())
14                if dropout > 0:
15                    layers.append(nn.Dropout(dropout))
16            self.network = nn.Sequential(*layers)
17
18        def forward(self, x):
19            out = self.network(x)
20            if self.task == 'classification':
21                out = torch.sigmoid(out)
22            return out

```

Listing 5.2: Arquitectura del MLP profundo

En la implementación real, cada tarea instancia su propia arquitectura con los hiperparámetros que le corresponden: el experimento de regresión usa `dropout=0.2`, mientras que el de clasificación usa `dropout=0.3`, dada la mayor tendencia al sobreajuste observada en el conjunto de menor tamaño (*digits* 0 vs. 1). El listado anterior omite el parámetro `task` por simplicidad expositiva; en el código, esta bifurcación (salida lineal vs. sigmoide) está resuelta mediante dos clases separadas, una por experimento.

El entrenamiento utiliza el optimizador Adam con decaimiento de pesos y *early stopping* basado en la pérdida de validación. El *early stopping* es necesario porque distintas funciones de pérdida tienen velocidades de convergencia distintas; un número fijo de épocas favorecería artificialmente a algunas funciones sobre otras.

```

1 class Trainer:
2     def __init__(self, model, loss_fn, lr=1e-3, patience=20):
3         self.model = model
4         self.loss_fn = loss_fn
5         self.optimizer = optim.Adam(model.parameters(),
6                                     lr=lr, weight_decay=1e-4)
7         self.patience = patience
8
9     def fit(self, train_loader, val_loader, epochs=300):
10        best_val, best_state, wait = float('inf'), None, 0
11        for epoch in range(1, epochs + 1):
12            self.model.train()
13            for X_b, y_b in train_loader:

```

```

14         self.optimizer.zero_grad()
15         loss = self.loss_fn(self.model(X_b), y_b)
16         loss.backward()
17         self.optimizer.step()
18     val_loss = self._eval(val_loader)
19     if val_loss < best_val - 1e-6:
20         best_val = val_loss
21         best_state = {k: v.clone() for k, v in
22                       self.model.state_dict().items()}
23         wait = 0
24     else:
25         wait += 1
26         if wait >= self.patience:
27             break
28     self.model.load_state_dict(best_state)

```

Listing 5.3: Bucle de entrenamiento con early stopping

El fragmento anterior sintetiza, con fines expositivos, el bucle de entrenamiento que en la implementación real está escrito de forma inline (no como clase reutilizable) en cada script de experimento. Los valores de *learning rate* y paciencia utilizados efectivamente difieren entre tareas:

Experimento	<i>Learning rate</i>	Paciencia (épocas)
Regresión sintética	10^{-3}	25
Clasificación (dígitos)	5×10^{-4}	30

5.2. Reproducibilidad y semillas

Todos los experimentos de esta memoria son completamente reproducibles a partir de una semilla maestra asociada a cada réplica Monte Carlo. Para la réplica k (con $k = 0, 1, \dots, K-1$, $K = 100$), la semilla utilizada se calcula como

$$\text{seed}_k = 17k + 3.$$

Esta expresión se aplica de manera idéntica en los tres experimentos (regresión sintética, clasificación y regresión sobre Airfoil Self-Noise), y su único propósito es generar semillas bien separadas entre réplicas consecutivas, evitando solapamientos accidentales en los generadores pseudoaleatorios de `numpy` y `scikit-learn`.

La semilla seed_k controla dos etapas independientes del proceso de generación de datos de cada réplica:

- **Generación y partición de los datos limpios.** En regresión, seed_k se pasa como `random_state` tanto a `make_regression` (que genera X e y) como a las dos llamadas sucesivas de `train_test_split` que producen las particiones de entrenamiento, validación y prueba. En clasificación, dado que los datos de `load_digits` se cargan una única vez para todas las réplicas, seed_k se emplea únicamente en las dos llamadas de `train_test_split` estratificado que generan las particiones. En ambos casos, reutilizar el mismo seed_k en las dos llamadas de `train_test_split` garantiza que la partición

completa (entrenamiento/validación/prueba) sea idéntica cada vez que se ejecuta la réplica k .

- **Inserción de la contaminación.** La contaminación (*outliers* en regresión, ruido de etiquetas en clasificación) se controla mediante un generador `numpy.random.RandomState` independiente, inicializado con un desplazamiento fijo sobre seed_k : $\text{seed}_k + 1000$ para los *outliers* en regresión, y $\text{seed}_k + 2000$ para el ruido de etiquetas en clasificación. Este generador determina tanto los índices de las observaciones contaminadas como, en el caso de los *outliers*, el signo del desplazamiento aplicado a cada una. El uso de un desplazamiento distinto para cada mecanismo de contaminación asegura que la partición de los datos y la contaminación introducida sean estadísticamente independientes entre sí, evitando que un mismo valor de semilla determine simultáneamente ambos procesos.