



UNIVERSIDAD
DE GRANADA

Visión por Computador

Temario y Cuestiones de Examen — Temas 1, 2 y 3

Ismael Sallami Moreno

Recursos Ingeniería Informática y Ade

Licencia

Este trabajo está bajo una Licencia Creative Commons BY-NC-ND 4.0.

Permisos: Se permite compartir, copiar y redistribuir el material en cualquier medio o formato.

Condiciones: Es necesario dar crédito adecuado, proporcionar un enlace a la licencia e indicar si se han realizado cambios. No se permite usar el material con fines comerciales ni distribuir material modificado.



Visión por Computador

Ismael Sallami Moreno

<https://elblogdeismael.github.io>

Índice general

1	Introducción y Fundamentos de la Visión por Computador	7
1.1	Tema 1: Concepto, Disciplinas Afines y Estado del Arte	7
2	Representación, Filtrado Espacial y Detección de Bordes	13
2.1	Tema 2: Fundamentos del Procesado Bidimensional y Operadores Diferenciales	13
3	Cuestiones y Problemas Resueltos	21
3.1	Tema 2: Representación, Filtrado Espacial y Detección de Bordes	21
4	Cuestiones y Autoevaluación de Machine Learning	39
4.1	Tema 3: Introducción al Aprendizaje Automático	39

Introducción y Fundamentos de la Visión por Computador

1.1 Tema 1: Concepto, Disciplinas Afines y Estado del Arte

La visión por computador constituye una de las disciplinas troncales de la inteligencia artificial. Su propósito formal consiste en dotar a sistemas computacionales de la capacidad de interpretar, modelar y reconstruir el mundo físico tridimensional a partir de señales visuales bidimensionales o secuencias de vídeo. Como señalan Russell y Norvig (2021), la visión no representa un módulo aislado, sino un componente indispensable para superar el test de Turing completo en entornos físicos, ya que exige interacción perceptiva directa con objetos y personas.

Desde una perspectiva operativa, el procesamiento visual invierte el flujo clásico de la computación gráfica. Mientras los gráficos por computador transforman representaciones geométricas y propiedades de iluminación en proyecciones bidimensionales de píxeles, la visión por computador se enfrenta al problema inverso: reconstruir la física, la geometría y el significado semántico de la escena a partir de una matriz discreta de mediciones ópticas.

1.1.1 1. Definición Formal y Modelos Jerárquicos

Diversos autores canónicos han delimitado el alcance de la disciplina:

- **Richard Szeliski (2022):** Define la visión computacional como el intento de describir el mundo observado en una o varias imágenes y reconstruir sus propiedades intrínsecas, tales como la forma volumétrica, las fuentes de iluminación y las distribuciones espectrales de color.
- **Milan Sonka, Vaclav Hlavac y Roger Boyle (2015):** Enfatizan el objetivo de duplicar las capacidades del sistema visual biológico mediante la percepción electrónica y la comprensión semántica de escenas del mundo real.
- **David Forsyth y Jean Ponce (2012):** Describen la disciplina como la aplicación de métodos estadísticos, geométricos y de teoría del aprendizaje para desenredar los datos ópticos y extraer descripciones estructuradas del mundo.

Niveles de Procesamiento Visual

Para estructurar la complejidad de las tareas, la literatura establece una jerarquía en función del nivel de abstracción representacional. Siguiendo la taxonomía clásica de Gonzalez y Woods (2018),

contrastada con los esquemas de Forsyth y Ponce (2012) y Sonka et al. (2015), distinguimos tres estratos bien diferenciados:

1. **Bajo Nivel (*Low-Level Vision* o *Early Vision*):** Las operaciones toman como entrada imágenes matriciales y devuelven como salida imágenes procesadas. Carecen de conocimiento contextual semántico sobre el contenido. Incluye el preprocesamiento para reducción de ruido, realce de contraste, operaciones morfológicas, transformaciones en el dominio de la frecuencia y detección de bordes primitivos.
2. **Nivel Medio (*Mid-Level Vision*):** La entrada sigue siendo una señal visual, pero la salida consiste en atributos y entidades estructuradas extraídas de los píxeles (esquinas, contornos continuos, regiones homogéneas o identidades preliminares). Comprende la segmentación semántica, el seguimiento de trayectorias (*tracking*), el emparejamiento de características locales y el agrupamiento perceptual.
3. **Alto Nivel (*High-Level Vision* o *Image Understanding*):** Opera sobre las representaciones intermedias para dotar de significado semántico al conjunto de elementos detectados, emulando procesos cognitivos complejos. Agrupa el reconocimiento y categorización de objetos, la estimación de posturas humanas, el registro volumétrico de imágenes, la navegación autónoma y el razonamiento visual compositivo.

1.1.2 2. Disciplinas Conexas y Límites Epistemológicos

El análisis visual converge en una intersección multidisciplinar donde conviven cuatro ramas del conocimiento:

- **Procesamiento Digital de Imágenes (*Digital Image Processing*):** Se centra en la manipulación de la señal de imagen para mejorar su calidad para observación humana o prepararla para almacenamiento y transmisión eficiente (filtrado, restauración, compresión JPEG).
- **Gráficos por Computador (*Computer Graphics*):** Proceso directo (*forward model*) que parte de modelos 3D matemáticos, cámaras virtuales y parámetros de reflectancia para sintetizar matrices de píxeles realistas.
- **Visión por Computador (*Computer Vision*):** Proceso inverso (*inverse problem*) que toma las matrices de intensidad captadas por sensores físicos e infiere las configuraciones 3D, propiedades de los materiales y relaciones funcionales.
- **Reconocimiento de Patrones y Aprendizaje Automático (*Pattern Recognition / Machine Learning*):** Proporciona el marco probabilístico, algebraico y estadístico para inferir modelos de decisión a partir de vectores de características visuales o directamente de tensores de datos crudos (*end-to-end*).

1.1.3 3. Las Dos Grandes Vertientes Metodológicas

A lo largo del curso y de la práctica investigadora contemporánea, la visión por computador se estructura en dos aproximaciones complementarias:

- **La Vertiente Geométrica (Modelado Físico y Proyectivo):** Se apoya en la geometría proyectiva, la óptica física y la estimación lineal y robusta. Formula la relación entre el espacio euclídeo $\mathbb{R}^3$ y el plano sensor $\mathbb{R}^2$ mediante coordenadas homogéneas y matrices de proyección de cámara. Resuelve problemas fundamentales como la geometría epipolar (matriz esencial

E y matriz fundamental F), la reconstrucción tridimensional a partir de múltiples vistas (*Structure from Motion*, SfM), el ajuste de haces (*Bundle Adjustment*) y la estimación robusta con RANSAC.

- **La Vertiente de Aprendizaje (*Data-Driven Vision*):** Trata de inferir patrones visuales a partir de grandes volúmenes de datos empíricos. Abarca desde los clasificadores clásicos basados en descriptores artesanales (*handcrafted features* como SIFT, HOG o LBP) combinados con SVM o árboles de decisión, hasta los paradigmas modernos de aprendizaje profundo:
 - *Supervised Learning*: Entrenamiento supervisado con etiquetas densas (*bounding boxes*, máscaras de segmentación).
 - *Self-Supervised Learning*: Extracción de representaciones visuales invariantes a partir de secuencias de vídeo o tareas pretextuales sin supervisión humana directa.
 - *Fine-Tuning y RLHF*: Ajuste supervisado y alineación con refuerzo humano para modelos multimodales visión-lenguaje (VLMs).
 - *Modelos Generativos*: Difusión probabilística y redes antagónicas (GANs) para síntesis y restauración visual.

Un elemento determinante en el éxito contemporáneo de la vertiente de aprendizaje ha sido la disponibilidad de conjuntos de datos masivos. Mientras ImageNet (Deng et al., 2009) supuso la chispa del aprendizaje profundo convolucional con más de 14 millones de imágenes anotadas, conjuntos a escala de internet como LAION-5B (Schuhmann et al., 2022), con más de 5.800 millones de pares de imagen y texto, han permitido el entrenamiento de modelos fundacionales multimodales capaces de generalización visual de disparo cero (*zero-shot*).

1.1.4 4. Por Qué la Visión por Computador es un Problema Difícil

Para los seres humanos, la visión resulta intuitiva y aparentemente inmediata, pues más del 50 % de la corteza cerebral de los primates está involucrada directa o indirectamente en el procesamiento visual. Sin embargo, en términos computacionales se trata de un problema formalmente mal condicionado (*ill-posed inverse problem*).

1. **Pérdida de Dimensionalidad (Ambigüedad 3D a 2D):** La proyección de una escena tridimensional sobre una superficie sensora bidimensional destruye la coordenada de profundidad. Infinitas configuraciones tridimensionales de distinta escala y distancia pueden proyectar exactamente la misma configuración de intensidades sobre la imagen (el principio del trampantojo o las ilusiones anamórficas de Ames).
2. **Variabilidad Fotométrica y de Iluminación:** El valor registrado en cada píxel es el producto de la radiancia de la fuente de luz incidente por la función de reflectancia bidireccional (BRDF) de la superficie del material. Variaciones en la posición del foco solar, sombras arrojadas o reflexiones especulares alteran los valores RGB de un mismo objeto de forma mucho más drástica que un cambio radical de categoría.
3. **Variabilidad Geométrica y de Punto de Vista:** La pose de la cámara respecto al objeto introduce rotaciones fuera del plano, deformaciones afines y distorsiones de perspectiva que desalinean por completo la disposición de las intensidades en la matriz.
4. **Oclusión Parcial y Desorden de Fondo (*Clutter*):** En entornos naturales, los objetos raras veces aparecen aislados sobre fondos planos. Frecuentemente se intersecan, quedan parcialmente ocultos por otros elementos o presentan texturas indistinguibles de los patrones de fondo.

5. **Variabilidad Intra-Clase y Deformabilidad:** Objetos pertenecientes a la misma categoría semántica (por ejemplo, sillas o animales) presentan una infinita variabilidad de formas, colores, dimensiones y configuraciones cinemáticas no rígidas.

1.1.5 5. Hitos Históricos Fundamentales

La disciplina ha transitado por diversas etapas conceptuales desde sus inicios a mediados del siglo XX:

- **1966 — El Proyecto de Verano de Marvin Minsky (MIT):** Al inicio de la IA, Minsky encargó al estudiante de grado Gerald Sussman “conectar una cámara a un ordenador y hacer que describiera lo que veía” como proyecto de verano de unas diez semanas. La anécdota ilustra cómo se subestimó profundamente la complejidad de la percepción biológica y la visión de bajo nivel.
- **Década de 1970 — El Mundo de los Bloques y Procesado Primitivo:** Investigaciones centradas en entornos sintéticos controlados (*blocks world* de Roberts) y extracción de aristas lineales mediante operadores heurísticos (Roberts, Sobel, Prewitt).
- **1982 — El Paradigma Computacional de David Marr:** Marr publica su obra seminal *Vision*, proponiendo que la visión es un sistema de procesamiento de información estructurado en tres niveles (computacional, algorítmico y de implementación física) que transita secuencialmente desde el esbozo primario (*primal sketch*), pasando por el esbozo 2.5D (orientación de superficies y profundidad relativa centrada en el observador), hasta el modelo 3D centrado en el objeto.
- **Décadas de 1980 y 1990 — Madurez Geométrica y Filtros Óptimos:** Formulación del detector de bordes óptimo de Canny (1986), espacios de escala (*scale-space theory*) de Witkin y Koenderink, la Laplaciana de Gaussiana y la consolidación de la visión multivista con Hartley, Zisserman y Faugeras.
- **Años 2000 — Ingeniería de Características Locales Robustas:** Aparición de detectores invariantes a escala y rotación como SIFT (Lowe, 2004), SURF y descriptores de gradientes orientados HOG (Dalal y Triggs, 2005), combinados con modelos de bolsas de palabras visuales (*Bag of Visual Words*) y el detector en tiempo real de Viola-Jones (2001) para caras.
- **2012 — La Revolución del Aprendizaje Profundo (AlexNet):** Krizhevsky, Sutskever y Hinton reducen drásticamente el error de clasificación en el reto ImageNet mediante una red convolucional profunda entrenada en GPUs, desplazando los enfoques artesanales por representaciones aprendidas directamente de los datos.
- **Actualidad — Modelos Fundacionales y Transformers Visuales:** Generalización de arquitecturas Vision Transformer (ViT), modelos multimodales contrastivos (CLIP), modelos de segmentación universales (Segment Anything - SAM) y modelos generativos de difusión probabilística.

1.1.6 6. Limitaciones Contemporáneas y Retos Abiertos

A pesar de que los modelos modernos superan con frecuencia el rendimiento medio humano en tareas cerradas y específicas de laboratorio (como la clasificación a 1.000 clases de ImageNet), presentan fragilidades estructurales cuando se enfrentan a situaciones reales:

1. **Inconsistencia de Sentido Común Físico:** Los modelos actuales carecen de un modelo intuitivo del mundo físico (gravedad, soporte, permanencia de los objetos, masa o fricción). Se

basan en correlaciones estadísticas de alta dimensión y pueden segmentar un objeto flotando en el vacío o predecir interpenetraciones de materia que violan las leyes físicas elementales.

2. **Vulnerabilidad ante Ataques Adversarios (*Adversarial Perturbations*):** Modificaciones imperceptibles para el ojo humano añadidas a una imagen pueden inducir a una red neuronal profunda a clasificar un vehículo escolar como un avestruz con una certidumbre superior al 99 %, demostrando que las representaciones neuronales difieren notablemente de los mecanismos perceptivos de la corteza biológica.
3. **Dependencia Extrema de Datos y Huella Energética:** El entrenamiento de redes profundas requiere millones de muestras etiquetadas y consumos energéticos considerables. Un ser humano es capaz de asimilar un concepto visual nuevo con uno o dos ejemplos (*few-shot learning*), mientras que un modelo de visión suele necesitar miles de iteraciones de optimización.
4. **Opacidad y Falta de Interpretabilidad (*Black-Box*):** En aplicaciones críticas como el diagnóstico biomédico o la conducción autónoma, la ausencia de trazabilidad explicativa dificulta garantizar el comportamiento del sistema ante condiciones de fallo o distribuciones no vistas (*out-of-distribution*).
5. **Sesgos Algorítmicos y Alucinaciones:** Los modelos replican y amplifican los sesgos socio-demográficos, geográficos y culturales presentes en los datos masivos de internet. Asimismo, los modelos generativos y los decodificadores visuales modernos pueden inventar texturas o detalles inexistentes (*alucinaciones visuales*), comprometiendo su uso en contextos periciales y forenses.

Representación, Filtrado Espacial y Detección de Bordes

2.1 Tema 2: Fundamentos del Procesado Bidimensional y Operadores Diferenciales

El presente capítulo aborda los principios matemáticos y algorítmicos que gobiernan la manipulación de la señal de imagen en el dominio espacial. Comprende desde la naturaleza física de la formación de la imagen y su discretización digital, pasando por la teoría del filtrado lineal y no lineal, hasta los operadores de gradiente y el detector de bordes óptimo de Canny.

2.1.1 1. Formación y Representación de la Imagen Digital

Una imagen en el plano continuo puede modelarse como una función bidimensional de intensidad lumínica $f(x, y)$, donde (x, y) denotan las coordenadas espaciales en el plano de proyección y el valor escalar $f(x, y)$ representa la irradiancia o energía luminosa por unidad de área recibida por el sensor.

Desde el punto de vista radiométrico, la energía captada en cada punto de la escena viene determinada por dos componentes primordiales:

$$f(x, y) = i(x, y) \cdot r(x, y)$$

donde $i(x, y) \in (0, \infty)$ representa la componente de iluminación incidente procedente de las fuentes luminosas del entorno, y $r(x, y) \in [0, 1]$ es el coeficiente de reflectancia intrínseco de los materiales de la superficie observada.

Muestreo y Cuantización

Para su procesamiento en arquitecturas digitales, la señal continua $f(x, y)$ debe someterse a dos etapas sucesivas de discretización:

1. **Muestreo Espacial (*Sampling*):** Discretización de las coordenadas espaciales (x, y) sobre una rejilla regular de $M \times N$ elementos, dando lugar a los píxeles (*picture elements*). Si el periodo de muestreo no cumple el teorema de Nyquist-Shannon respecto a las frecuencias espaciales más altas de la escena óptica, se producen artefactos de solapamiento espectral (*aliasing*).

2. **Cuantización de Amplitud (*Quantization*):** Mapeo del rango continuo de irradiancia a un conjunto finito de L niveles discretos. Habitualmente se emplean representaciones de 8 bits por canal ($L = 2^8 = 256$, en el rango entero $[0, 255]$), o representaciones normalizadas en coma flotante $\in [0, 1, 0]$.

En imágenes multicanal o a color, la señal se representa mediante una función vectorial:

$$\mathbf{I}(x, y) = \begin{bmatrix} I_R(x, y) \\ I_G(x, y) \\ I_B(x, y) \end{bmatrix}$$

La conversión estándar a escala de grises combina los canales ponderando la sensibilidad fototópica de la fovea humana según la recomendación ITU-R BT.601:

$$Y(x, y) = 0,299 \cdot I_R(x, y) + 0,587 \cdot I_G(x, y) + 0,114 \cdot I_B(x, y)$$

2.1.2 2. Filtrado Espacial Lineal: Correlación y Convolución

El filtrado en el dominio espacial consiste en sustituir el valor de cada píxel de la imagen de entrada por una combinación ponderada de las intensidades de sus píxeles vecinos dentro de una ventana de soporte local. La matriz de ponderaciones que define los coeficientes de dicha ventana recibe el nombre de máscara, filtro o *kernel* h de tamaño $(2k + 1) \times (2k + 1)$.

Correlación Cruzada (Cross-Correlation)

La correlación cruzada bidimensional desplaza directamente el kernel h sobre la imagen I sin modificar la orientación relativa de sus índices:

$$(I \otimes h)(x, y) = \sum_{u=-k}^k \sum_{v=-k}^k I(x + u, y + v) h(u, v)$$

La correlación cruzada mide el grado de coincidencia o similitud local entre la estructura del kernel y el vecindario del píxel. Se emplea extensamente en técnicas de emparejamiento de plantillas (*template matching*). Sin embargo, la correlación carece de la propiedad asociativa y conmutativa respecto a la combinación sucesiva de filtros.

Convolución Lineal Bidimensional (Convolution)

La convolución lineal invierte espacialmente el kernel en ambos ejes coordenados (equivalente a una rotación de 180°) antes de calcular el producto escalar ponderado:

$$(I * h)(x, y) = \sum_{u=-k}^k \sum_{v=-k}^k I(x - u, y - v) h(u, v) = \sum_{u=-k}^k \sum_{v=-k}^k I(x + u, y + v) h(-u, -v)$$

La inversión de índices dota a la convolución de propiedades algebraicas fundamentales en el análisis de sistemas lineales e invariantes a la traslación (LTI):

- **Conmutatividad:** $I * h = h * I$.
- **Asociatividad:** $I * (h_1 * h_2) = (I * h_1) * h_2$. Permite prefusionar múltiples filtros en un único kernel equivalente mediante $h_{eq} = h_1 * h_2$, reduciendo el coste computacional.
- **Distributividad respecto a la suma:** $I * (h_1 + h_2) = (I * h_1) + (I * h_2)$.
- **Equivariancia ante traslación:** Si T_Δ es un operador de desplazamiento espacial puro tal que $T_\Delta(I)(x, y) = I(x - x_0, y - y_0)$, entonces $T_\Delta(I * h) = (T_\Delta I) * h$.

Tratamiento de Condiciones de Frontera

Al evaluar la convolución cerca de los bordes exteriores de la matriz de tamaño $M \times N$, el soporte del kernel sobresale de las dimensiones válidas de la imagen. Existen cuatro estrategias analíticas para resolver estas discontinuidades de frontera:

1. **Relleno con ceros (Zero Padding):** Asume que todos los píxeles fuera del dominio de la imagen tienen intensidad nula. Introduce una caída artificial de energía que oscurece el perímetro de la imagen filtrada.
2. **Replicación de borde (Clamp / Border Replicate):** Extiende el valor del píxel frontera más próximo de forma constante. Evita contrastes bruscos espurios.
3. **Reflexión o Simetría (Reflect / Mirroring):** Refleja la señal hacia el interior en torno al eje de la frontera. Preserva la continuidad de las derivadas de primer orden y minimiza artefactos de alta frecuencia.
4. **Periódica o Toroidal (Wrap around):** Conecta el borde superior con el inferior y el izquierdo con el derecho. Es la condición implícita en la convolución circular mediante la Transformada Rápida de Fourier (FFT).

2.1.3 3. Filtros Lineales de Suavizado: Box Filter frente al Filtro Gaussiano

El suavizado espacial (*smoothing*) tiene por objeto atenuar las fluctuaciones rápidas debidas a ruido aditivo independiente, actuando como un filtro paso bajo (*low-pass filter*).

Filtro de Media (Box Filter)

El filtro de caja asigna ponderaciones uniformes a todas las posiciones del vecindario dentro de un soporte $K \times K$:

$$h_{\text{box}}(u, v) = \frac{1}{K^2}, \quad \forall (u, v) \in \left[-\frac{K-1}{2}, \frac{K-1}{2} \right]^2$$

Aunque resulta computacionalmente trivial, presenta dos desventajas analíticas graves: - Asigna el mismo peso a un píxel contiguo al centro que a un píxel ubicado en los vértices del vecindario. - Su respuesta en el dominio frecuencial es una función sinc bidimensional ($\text{sinc}(u)\text{sinc}(v)$), caracterizada por lóbulos secundarios prominentes que introducen frecuencias espurias y artefactos visuales en forma de bandas rectangulares (*ringing* y efecto de cuadrícula).

El Filtro Gaussiano y sus Propiedades Analíticas

La función gaussiana continua bidimensional isotrópica centrada en el origen se formula como:

$$G_{\sigma}(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right)$$

donde σ es la desviación típica que parametriza la escala o anchura efectiva del filtro. Sus propiedades distintivas son:

- **Isotropía (Simetría Circular):** La respuesta depende exclusivamente de la distancia euclídea al origen $r = \sqrt{x^2 + y^2}$. Suaviza con igual intensidad en todas las direcciones del plano.
- **Respuesta Frecuencial Óptima:** La transformada de Fourier de una gaussiana en el espacio es otra gaussiana en la frecuencia. Carece de lóbulos secundarios, lo que garantiza que no introduce artefactos oscilatorios ni altas frecuencias espurias.
- **Teorema de Convolución de Gaussians:** La convolución sucesiva de dos filtros gaussianos con parámetros σ_1 y σ_2 equivale a aplicar una única gaussiana con varianza acumulada:

$$G_{\sigma_1} * G_{\sigma_2} = G_{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

- **Separabilidad Computacional:** La función gaussiana bidimensional se descompone en el producto exacto de dos gaussianas unidimensionales independientes:

$$G_{\sigma}(x, y) = \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right)\right) \cdot \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{y^2}{2\sigma^2}\right)\right) = G_{\sigma}(x) \cdot G_{\sigma}(y)$$

En virtud de la asociatividad de la convolución:

$$I * G_{\sigma}(x, y) = (I * G_{\sigma}(x)) * G_{\sigma}(y)$$

Esta descomposición reduce la complejidad algorítmica de aplicar un kernel de $K \times K$ píxeles de $O(K^2)$ operaciones por píxel a tan solo $O(2K)$, permitiendo aceleraciones de más de un orden de magnitud en kernels medianos y grandes.

Para discretizar el kernel en una matriz de tamaño $(2k + 1) \times (2k + 1)$, se trunca habitualmente el soporte a una distancia de $k = \lceil 3\sigma \rceil$, abarcando el 99,73 % de la masa probabilística, y se normalizan los coeficientes resultantes dividiendo entre la suma total para garantizar ganancia unitaria en continua (preservación del nivel medio de iluminación).

2.1.4 4. Filtros No Lineales y Transformaciones Estadísticas

Cuando el ruido de la imagen no es de naturaleza gaussiana aditiva, los operadores lineales no consiguen desacoplar el ruido de las aristas genuinas.

Filtro de Mediana (Median Filter)

El filtro de mediana sustituye el píxel central por el valor mediano estadístico de las intensidades contenidas en su ventana local:

$$I_{\text{med}}(x, y) = \text{mediana}(\{I(x + u, y + v) \mid -k \leq u, v \leq k\})$$

Es especialmente efectivo frente a ruido de tipo impulso (*sal y pimienta*). Debido a que la mediana es un estimador estadístico robusto de orden, descarta valores extremos atípicos sin promediarlos, preservando con gran fidelidad la nitidez de los bordes rectilíneos y contrastados.

Umbralización Óptima: Algoritmo de Otsu

La umbralización global (*thresholding*) segmenta una imagen I en dos clases (fondo y primer plano) mediante un valor de corte T :

$$B(x, y) = \begin{cases} 1 & \text{si } I(x, y) \geq T \\ 0 & \text{si } I(x, y) < T \end{cases}$$

El algoritmo de Otsu (1979) calcula de forma analítica el umbral óptimo T^* maximizando la varianza entre clases $\sigma_B^2(T)$, o equivalentemente minimizando la varianza intraclase $\sigma_W^2(T)$:

$$\sigma_B^2(T) = \omega_0(T)\omega_1(T)[\mu_0(T) - \mu_1(T)]^2$$

donde $\omega_0(T)$ y $\omega_1(T)$ son las probabilidades acumuladas de cada clase bajo el histograma normalizado, y $\mu_0(T), \mu_1(T)$ son sus medias respectivas. Este criterio no requiere parámetros heurísticos y converge evaluando directamente las frecuencias acumuladas del histograma.

2.1.5 5. Detección de Bordes y Operadores Diferenciales

Un borde o arista (*edge*) en una imagen se define como una discontinuidad local significativa en la función de intensidad luminosa, originada comúnmente por saltos de profundidad geométrica, cambios en la orientación normal de la superficie o variaciones en la reflectancia del material.

Operadores de Gradiente de Primer Orden

Para una función continua $f(x, y)$, el gradiente espacial es un campo vectorial bidimensional orientado en la dirección de máxima variación de intensidad:

$$\nabla f(x, y) = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}$$

En el dominio discreto de la imagen digital, las derivadas parciales se aproximan mediante diferencias finitas: - Diferencia hacia adelante: $\frac{\partial I}{\partial x} \approx I(x + 1, y) - I(x, y)$ - Diferencia centrada: $\frac{\partial I}{\partial x} \approx \frac{I(x+1, y) - I(x-1, y)}{2}$

Los núcleos de derivación combinan la diferenciación en un eje con el suavizado en el eje ortogonal para amortiguar el ruido:

– **Filtro de Sobel (3x3):**

$$S_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix}, \quad S_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix}$$

El operador de Sobel es separable en el producto de una máscara de suavizado binomial $[1, 2, 1]^T$ y un operador de diferencia finita centrada $[-1, 0, 1]$.

A partir de los mapas de respuesta $g_x = I * S_x$ y $g_y = I * S_y$, se obtienen la magnitud escalar y el ángulo de orientación del gradiente en cada píxel:

$$M(x, y) = \|\nabla I(x, y)\| = \sqrt{g_x^2(x, y) + g_y^2(x, y)}$$

$$\theta(x, y) = \text{atan2}(g_y(x, y), g_x(x, y)) \in [-\pi, \pi]$$

La magnitud mide la fuerza del contraste local, mientras que $\theta(x, y)$ es el vector normal ortogonal a la dirección tangente a lo largo del contorno del borde.

Operadores de Segundo Orden: La Laplaciana y Laplaciana de Gaussiana (LoG)

La Laplaciana es un operador diferencial isotrópico de segundo orden definido como la divergencia del gradiente:

$$\nabla^2 f(x, y) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$$

Los puntos donde se localizan los máximos de pendiente en la función de intensidad corresponden a los cruces por cero (*zero-crossings*) de la segunda derivada. No obstante, la diferenciación de segundo orden amplifica drásticamente el ruido de alta frecuencia. Por este motivo, David Marr y Ellen Hildreth (1980) propusieron prefiltrar la imagen con un núcleo gaussiano antes de aplicar el operador de Laplace:

$$\nabla^2(I * G_\sigma) = I * (\nabla^2 G_\sigma)$$

El núcleo combinado resultante es la Laplaciana de Gaussiana (LoG), formulada como:

$$\text{LoG}_\sigma(x, y) = \nabla^2 G_\sigma(x, y) = \frac{1}{\pi\sigma^4} \left(\frac{x^2 + y^2}{2\sigma^2} - 1 \right) \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right)$$

Este filtro adopta el característico perfil de “*sombrero mexicano*” (*Mexican hat*). Asimismo, puede aproximarse de forma computacionalmente eficiente mediante la Diferencia de Gaussianas (DoG):

$$\text{DoG}(x, y) = G_{\sigma_1}(x, y) - G_{\sigma_2}(x, y) \approx (\sigma_1 - \sigma_2) \nabla^2 G_\sigma$$

2.1.6 6. El Detector de Bordes de Canny

John F. Canny formalizó en 1986 una formulación analítica de la detección de bordes a partir de tres criterios de optimalidad matemática:

1. **Baja Tasa de Error:** Debe detectar todas las aristas genuinas minimizando falsos positivos provocados por ruido y falsos negativos causados por contrastes débiles.
2. **Localización Exacta:** Los puntos detectados deben situarse lo más próximos posible al verdadero centro físico del borde.
3. **Respuesta Única:** No debe generar múltiples detecciones espurias adyacentes para una única transición física de borde (*aristas de un solo píxel de espesor*).

Para satisfacer estas condiciones de forma rigurosa, el algoritmo de Canny se ejecuta en cuatro fases secuenciales:

- **Fase 1: Suavizado Gaussiano:** La imagen de entrada se convoluciona con un filtro gaussiano de escala σ para eliminar componentes de ruido estocástico de alta frecuencia:

$$I_{\sigma} = I * G_{\sigma}$$

- **Fase 2: Cálculo del Gradiente:** Se aplican operadores diferenciales (habitualmente Sobel) para calcular las componentes ortogonales del gradiente g_x y g_y . Se obtienen la magnitud $M(x, y)$ y la dirección $\theta(x, y)$, cuantizando este ángulo en cuatro orientaciones discretas principales: horizontal (0°), diagonal principal (45°), vertical (90°) y diagonal secundaria (135°).
- **Fase 3: Supresión de No Máximos (*Non-Maximum Suppression*, NMS):** Para garantizar la respuesta única y reducir el espesor del contorno a un solo píxel, se examina si la magnitud $M(x, y)$ es un máximo local respecto a sus dos vecinos situados a lo largo de la dirección ortogonal al borde $\theta(x, y)$:

$$M_{\text{NMS}}(x, y) = \begin{cases} M(x, y) & \text{si } M(x, y) \geq M(x + d_x, y + d_y) \text{ y } M(x, y) \geq M(x - d_x, y - d_y) \\ 0 & \text{en caso contrario} \end{cases}$$

Todos los píxeles que no superan a sus vecinos directos en la línea del gradiente se anulan a cero, produciendo crestas de arista perfectamente delgadas.

- **Fase 4: Umbralización con Histéresis y Conectividad (*Hysteresis Thresholding*):** Un único umbral global introduce aristas rotas si es demasiado alto o contornos ruidosos espurios si es demasiado bajo. Canny introduce dos umbrales complementarios, T_{alto} y T_{bajo} (habitualmente fijados en relación $T_{\text{alto}}/T_{\text{bajo}} \approx 2 : 1$ o $3 : 1$):
 - Los píxeles con magnitud $M_{\text{NMS}} \geq T_{\text{alto}}$ se declaran bordes firmes (*aristas fuertes*).
 - Los píxeles con $M_{\text{NMS}} < T_{\text{bajo}}$ se descartan definitivamente como ruido.
 - Los píxeles con $T_{\text{bajo}} \leq M_{\text{NMS}} < T_{\text{alto}}$ (*aristas débiles*) solo se aceptan como bordes legítimos si se encuentran conectados topológicamente (mediante vecindad de 8) con algún píxel perteneciente a una arista fuerte.

Cuestiones y Problemas Resueltos

3.1 Tema 2: Representación, Filtrado Espacial y Detección de Bordes

Este capítulo recoge la resolución analítica, matemática y conceptual de las cuestiones teóricas, problemas de diseño de kernels y baterías de autoevaluación planteadas al cierre del Tema 2 (dividido en los bloques 2.1 *Image Representation and Filtering* y 2.2 *Gaussian Filtering and Edges*) del programa oficial de Visión por Computador en la Universidad de Granada.

Las deducciones se fundamentan en el tratamiento formal del procesado de señal bidimensional y la teoría clásica de visión computacional, referenciando las obras canónicas de Szeliski (2022), Gonzalez y Woods (2018), Birchfield (2018) y los artículos seminales de Otsu (1979) y Canny (1986).

3.1.1 Bloque I: Cuestiones Teóricas sobre Filtros Lineales

Las siguientes cuestiones corresponden a los ejercicios propuestos en las diapositivas 118 y 119 del bloque temático 2.1.

CUESTIÓN 1: Equivariancia de la convolución frente a rotación y escalado

Enunciado:

¿Es la convolución equivariante frente a la rotación y frente al escalado?

Respuesta breve:

La convolución estándar **no** es equivariante a la rotación ni al escalado; únicamente lo es ante traslaciones. Para la rotación requeriría un kernel estrictamente isotrópico (como una Gaussiana circular), y para el escalado una adaptación dinámica del tamaño del filtro o el uso de pirámides multiescala.

Resolución analítica:

Un operador o transformación T es equivariante respecto a un grupo de transformaciones $g \in G$ si aplicar la transformación a la entrada y luego operar produce exactamente el mismo resultado que operar primero sobre la entrada y transformar posteriormente la salida:

$$T(g \cdot I) = g \cdot (T(I))$$

Analizamos cada caso por separado:

– **1. Equivariancia a la traslación (comportamiento de referencia)**

La convolución continua en $\mathbb{R}^2$ de una imagen $I(\mathbf{x})$ con un kernel $K(\mathbf{x})$ se define como:

$$(I * K)(\mathbf{x}) = \int_{\mathbb{R}^2} I(\boldsymbol{\tau}) K(\mathbf{x} - \boldsymbol{\tau}) d\boldsymbol{\tau}$$

Si aplicamos una traslación espacial pura $\mathbf{x}_0$, de modo que $I'(\mathbf{x}) = I(\mathbf{x} - \mathbf{x}_0)$, la respuesta del filtro es:

$$(I' * K)(\mathbf{x}) = \int_{\mathbb{R}^2} I(\boldsymbol{\tau} - \mathbf{x}_0) K(\mathbf{x} - \boldsymbol{\tau}) d\boldsymbol{\tau}$$

Mediante el cambio de variable $\boldsymbol{\mu} = \boldsymbol{\tau} - \mathbf{x}_0$ ($d\boldsymbol{\mu} = d\boldsymbol{\tau}$):

$$(I' * K)(\mathbf{x}) = \int_{\mathbb{R}^2} I(\boldsymbol{\mu}) K((\mathbf{x} - \mathbf{x}_0) - \boldsymbol{\mu}) d\boldsymbol{\mu} = (I * K)(\mathbf{x} - \mathbf{x}_0)$$

La convolución es intrínsecamente equivariante ante traslaciones. El valor detectado se desplaza exactamente en la misma magnitud espacial que la escena de entrada.

– **2. Equivariancia a la rotación**

Sea R_θ una matriz ortogonal de rotación de ángulo θ en $\mathbb{R}^2$. Definimos la imagen rotada como $I_{rot}(\mathbf{x}) = I(R_\theta^{-1}\mathbf{x})$. Si convolucionamos esta imagen con el kernel estático K :

$$(I_{rot} * K)(\mathbf{x}) = \int_{\mathbb{R}^2} I(R_\theta^{-1}\boldsymbol{\tau}) K(\mathbf{x} - \boldsymbol{\tau}) d\boldsymbol{\tau}$$

Efectuando el cambio de variable $\boldsymbol{\mu} = R_\theta^{-1}\boldsymbol{\tau}$, donde $\boldsymbol{\tau} = R_\theta\boldsymbol{\mu}$ y el jacobiano satisface $|\det(R_\theta)| = 1$:

$$(I_{rot} * K)(\mathbf{x}) = \int_{\mathbb{R}^2} I(\boldsymbol{\mu}) K(\mathbf{x} - R_\theta\boldsymbol{\mu}) d\boldsymbol{\mu}$$

Por otra parte, si calculáramos la convolución sobre la imagen original y rotáramos posteriormente el mapa de salida resultante, obtendríamos:

$$[R_\theta(I * K)](\mathbf{x}) = (I * K)(R_\theta^{-1}\mathbf{x}) = \int_{\mathbb{R}^2} I(\boldsymbol{\mu}) K(R_\theta^{-1}\mathbf{x} - \boldsymbol{\mu}) d\boldsymbol{\mu}$$

Multiplicando el argumento de K por R_θ aprovechando la linealidad de la transformación:

$$K(R_\theta^{-1}\mathbf{x} - \boldsymbol{\mu}) = K(R_\theta^{-1}(\mathbf{x} - R_\theta\boldsymbol{\mu}))$$

Para que ambas expresiones coincidan para cualquier imagen I , es estrictamente necesario que:

$$K(\mathbf{x} - R_\theta\boldsymbol{\mu}) = K(R_\theta^{-1}(\mathbf{x} - R_\theta\boldsymbol{\mu})) \quad \forall \mathbf{x}, \boldsymbol{\mu}$$

Esto solo se cumple si el kernel es rotacionalmente invariante (isotrópico), es decir, dependiente exclusivamente del radio euclídeo $\|\mathbf{x}\|$ respecto al origen, como sucede en una Gaussiana circular $G(\mathbf{x}) = \frac{1}{2\pi\sigma^2} e^{-\frac{\|\mathbf{x}\|^2}{2\sigma^2}}$ o en un Laplaciano de Gaussiana (LoG).

Para cualquier kernel direccional común (como los operadores de Sobel, Prewitt, Gabor o derivadas orientadas), el kernel mantiene fija su orientación respecto al sistema de coordenadas de la matriz.

Si la imagen rota 45° , un detector horizontal continuará diferenciando filas horizontales en lugar de seguir la inclinación de la estructura rotada. En consecuencia, en el caso general la convolución **no es equivariante a la rotación**.

– 3. Equivariancia al escalado

Sea una transformación de escala isotrópica por un factor escalar $s > 0$, de forma que $I_{esc}(\mathbf{x}) = I(s\mathbf{x})$. Al filtrar con un kernel de tamaño fijo K :

$$(I_{esc} * K)(\mathbf{x}) = \int_{\mathbb{R}^2} I(s\boldsymbol{\tau})K(\mathbf{x} - \boldsymbol{\tau}) d\boldsymbol{\tau}$$

Con el cambio $\boldsymbol{\mu} = s\boldsymbol{\tau}$ ($d\boldsymbol{\tau} = s^{-2}d\boldsymbol{\mu}$):

$$(I_{esc} * K)(\mathbf{x}) = \frac{1}{s^2} \int_{\mathbb{R}^2} I(\boldsymbol{\mu})K\left(\mathbf{x} - \frac{\boldsymbol{\mu}}{s}\right) d\boldsymbol{\mu}$$

Sin embargo, filtrar primero y escalar después la respuesta generaría:

$$[S_s(I * K)](\mathbf{x}) = (I * K)(s\mathbf{x}) = \int_{\mathbb{R}^2} I(\boldsymbol{\mu})K(s\mathbf{x} - \boldsymbol{\mu}) d\boldsymbol{\mu}$$

Ambas integrales difieren sustancialmente: el kernel K posee una huella de píxeles finita fija. Si reducimos la imagen a la mitad, los patrones visuales ocupan menos píxeles y el filtro cubrirá el doble de contexto semántico; si la ampliamos, cubrirá la mitad. Para mantener la equivariancia sería imperativo redimensionar dinámicamente el soporte y la varianza del propio filtro ($K_s(\mathbf{x}) = \frac{1}{s^2}K(\mathbf{x}/s)$).

Por tanto, con máscaras fijas, la convolución estándar **no es equivariante al escalado**. Esta limitación motivó históricamente la teoría del espacio-escala (*scale-space representation*) de Witkin y Lindeberg, así como las pirámides Gaussianas y Laplacianas.

CUESTIÓN 2: Equivalencia entre correlación cruzada y convolución rotando la salida

Enunciado:

¿Es la correlación equivalente a la convolución si rotamos su salida 180 grados? Es decir, ¿se cumple la igualdad:

$$\text{rotate180}(\text{Correlation}(I, K)) == \text{Convolution}(I, K)?$$

Respuesta breve:

No. Rotar la salida invierte todo el marco de coordenadas de la imagen resultante, lo que equivale a convolucionar la imagen original invertida. Para que la correlación sea idéntica a la convolución, lo que debe rotarse 180° es el **kernel** antes de operar ($I * K = I \otimes \text{rot}_{180}(K)$).

Resolución analítica:

La respuesta matemática es **no**. Rotar la salida tras haber calculado la correlación no equivale a convolucionar la imagen original.

Analicemos las definiciones discretas bidimensionales centradas en el origen. Para simplificar la notación y sin pérdida de generalidad, consideramos dominios simétricos indexados por números enteros:

– **1. Correlación cruzada ($I \otimes K$):**

$$(I \otimes K)(x, y) = \sum_u \sum_v I(x + u, y + v) K(u, v)$$

– **2. Convolución ($I * K$):**

$$(I * K)(x, y) = \sum_u \sum_v I(x - u, y - v) K(u, v)$$

Efectuando el cambio de índices mudos $u' = -u$ y $v' = -v$, la convolución se expresa de manera idéntica a una correlación pero invirtiendo los ejes del kernel:

$$(I * K)(x, y) = \sum_{u'} \sum_{v'} I(x + u', y + v') K(-u', -v') = [I \otimes \text{rotate180}(K)](x, y)$$

Esta es la relación canónica fundamental: **la convolución equivale a la correlación cruzada cuando se rota el kernel 180 grados antes de operar.**

– **3. Efecto de rotar la salida de la correlación:**

Sea $C(x, y) = (I \otimes K)(x, y)$. La operación $\text{rotate180}(C)$ aplicada sobre la imagen de salida invierte simultáneamente ambas coordenadas espaciales:

$$C_{\text{rot180}}(x, y) = C(-x, -y)$$

Sustituyendo $(-x, -y)$ en la definición de la correlación:

$$C(-x, -y) = \sum_u \sum_v I(-x + u, -y + v) K(u, v)$$

Haciendo el cambio de variables $u' = -u$ y $v' = -v$:

$$C(-x, -y) = \sum_{u'} \sum_{v'} I(-(x + u'), -(y + v')) K(-u', -v')$$

Definiendo la imagen de entrada rotada 180 grados como $I_{\text{rot180}}(x, y) = I(-x, -y)$, reconocemos que la expresión anterior es:

$$C(-x, -y) = \sum_{u'} \sum_{v'} I_{\text{rot180}}(x + u', y + v') K(-u', -v') = [\text{rotate180}(I) * K](x, y)$$

Rotar la imagen resultante de la correlación equivale a haber convolucionado la imagen original rotada 180 grados con el kernel original sin rotar. Toda la escena resultante queda orientada boca abajo y con el eje horizontal invertido. Para obtener la convolución $I * K$, el giro de 180 grados debe realizarse obligatoriamente sobre el soporte del **filtro** K con carácter previo al cálculo, nunca sobre la imagen de respuesta generada.

CUESTIÓN 3: Linealidad del umbralizado (*Thresholding*)

Enunciado:

¿Puede implementarse el umbralizado mediante un filtro lineal? Dicho de otro modo, ¿es el umbralizado una operación lineal?

Respuesta breve:

No. El umbralizado es una operación estrictamente no lineal (función escalón de Heaviside). Viola tanto la homogeneidad ($T(\alpha I) \neq \alpha T(I)$) como la aditividad ($T(I_1 + I_2) \neq T(I_1) + T(I_2)$) del principio de superposición, por lo que no puede implementarse como un filtro lineal ni mediante convolución.

Resolución analítica:

La respuesta categórica es **no**. El umbralizado es una operación estrictamente **no lineal**.

Para que un operador $T : \mathcal{I} \rightarrow \mathcal{I}$ sea lineal en el espacio vectorial de imágenes, debe satisfacer rigurosamente el **principio de superposición**, descompuesto en dos condiciones necesarias y suficientes:

1. **Homogeneidad:** $T(\alpha I) = \alpha T(I)$ para cualquier escalar $\alpha \in \mathbb{R}$.
2. **Aditividad:** $T(I_1 + I_2) = T(I_1) + T(I_2)$ para cualquier par de imágenes I_1, I_2 .

Consideremos la definición estándar del operador de umbralizado con valor umbral $\theta > 0$, que mapea la intensidad de cada píxel a un valor binario:

$$T_\theta(I)(x, y) = \begin{cases} 1 & \text{si } I(x, y) \geq \theta \\ 0 & \text{si } I(x, y) < \theta \end{cases}$$

Demostramos la violación de ambas propiedades mediante contraejemplos analíticos directos:

– **Violación de la homogeneidad:**

Tomemos un píxel con intensidad $I(x, y) = 0,5\theta$. Al aplicar el umbral:

$$T_\theta(I)(x, y) = 0$$

Multipliquemos la imagen de entrada por el factor de escala escalar $\alpha = 2$. La nueva intensidad es $(\alpha I)(x, y) = 2 \cdot (0,5\theta) = \theta$. Evaluando el operador:

$$T_\theta(\alpha I)(x, y) = 1$$

Sin embargo, multiplicar la salida por el escalar produce:

$$\alpha \cdot T_\theta(I)(x, y) = 2 \cdot 0 = 0$$

Dado que $1 \neq 0$, se incumple la homogeneidad: $T_\theta(\alpha I) \neq \alpha T_\theta(I)$.

– **Violación de la aditividad:**

Consideremos dos imágenes I_1 e I_2 que en la posición (x, y) presentan valores $I_1(x, y) = 0,6\theta$ e $I_2(x, y) = 0,6\theta$. Evaluando por separado:

$$T_\theta(I_1)(x, y) = 0 \quad \text{y} \quad T_\theta(I_2)(x, y) = 0 \implies T_\theta(I_1)(x, y) + T_\theta(I_2)(x, y) = 0$$

Sumando ambas imágenes previamente a la entrada, la intensidad combinada es $(I_1 + I_2)(x, y) =$

$1, 2\theta \geq \theta$. Por tanto:

$$T_\theta(I_1 + I_2)(x, y) = 1 \neq 0$$

El umbralizado modela una función escalón de Heaviside desplazada, cuya derivada contiene una delta de Dirac. Introduce componentes espectrales inexistentes en la señal original y carece de representación matricial o convolucional mediante máscaras espaciales.

CUESTIÓN 4: Diseño analítico de filtro de contraste local con 4-vecinos

Enunciado:

¿Cuál de los siguientes filtros 3×3 devuelve un valor positivo si el promedio de los cuatro vecinos contiguos (4-conectados) es menor que el centro, y un valor negativo (o cero) en caso contrario?

$$(a) \begin{bmatrix} 0 & 1/4 & 0 \\ 1/4 & -1 & 1/4 \\ 0 & 1/4 & 0 \end{bmatrix} \quad (b) \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 2 \end{bmatrix} \quad (c) \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

Respuesta breve:

El filtro correcto es el **(c)** (con centro $+4$ y los cuatro vecinos ortogonales con peso -1). Su respuesta es $4(I_c - \bar{I}_{adj})$, estrictamente positiva si y solo si el centro supera la media de los vecinos. Corresponde al Laplaciano discreto negativo ($-\nabla^2$), que actúa como detector de máximos locales y picos de contraste.

Resolución analítica:

Modelamos matemáticamente la condición exigida en el enunciado. Sea $I_c = I(x, y)$ la intensidad del píxel central y definamos el conjunto de sus cuatro vecinos ortogonales $N_4(x, y) = \{I(x, y - 1), I(x - 1, y), I(x + 1, y), I(x, y + 1)\}$.

El promedio aritmético de dichos cuatro vecinos es:

$$\bar{I}_{adj} = \frac{1}{4} \sum_{(u,v) \in N_4} I(u, v)$$

La condición específica que la respuesta R debe ser estrictamente positiva si y solo si $\bar{I}_{adj} < I_c$:

$$\bar{I}_{adj} < I_c \iff I_c - \bar{I}_{adj} > 0$$

Multiplicando la inecuación por la constante positiva 4:

$$4(I_c - \bar{I}_{adj}) > 0 \iff 4I_c - \sum_{(u,v) \in N_4} I(u, v) > 0$$

Evaluamos la respuesta de convolución de cada uno de los filtros candidatos sobre el entorno local:

– **Filtro (a):**

$$R_a = \frac{1}{4} \sum_{(u,v) \in N_4} I(u, v) - 1 \cdot I_c = \bar{I}_{adj} - I_c = -(I_c - \bar{I}_{adj})$$

Si el promedio de los vecinos es menor que el centro ($\bar{I}_{adj} < I_c$), la diferencia $I_c - \bar{I}_{adj}$ es positiva, por lo que R_a toma un valor **negativo**. Produce exactamente el signo opuesto al buscado.

– **Filtro (b):**

Es un operador diferencial asimétrico orientado en el eje vertical (variante de Sobel con error en el extremo inferior derecho donde figura un 2 en lugar de 1). No evalúa la relación isotrópica entre el centro y sus vecinos, sino un gradiente direccional con peso nulo en la fila central.

– **Filtro (c):**

$$R_c = 4I_c - 1 \cdot \sum_{(u,v) \in N_4} I(u,v) = 4(I_c - \bar{I}_{adj})$$

Analizando su signo: 1. Si $\bar{I}_{adj} < I_c \implies I_c - \bar{I}_{adj} > 0 \implies R_c > 0$ (valor estrictamente positivo).
2. Si $\bar{I}_{adj} \geq I_c \implies I_c - \bar{I}_{adj} \leq 0 \implies R_c \leq 0$ (valor negativo o cero).

La respuesta es unívocamente la **(c)**.

Desde el punto de vista del procesamiento digital, el filtro (c) representa la aproximación en diferencias finitas del **Laplaciano discreto negativo** ($-\nabla^2 I$). El Laplaciano clásico mide la divergencia del gradiente: en un máximo local (donde la intensidad central sobresale de su entorno inmediato), la curvatura es cóncava ($\nabla^2 I < 0$), por lo que $-\nabla^2 I$ produce una respuesta positiva máxima que detecta picos luminosos y realza detalles finos.

3.1.2 Bloque II: Problema Analítico del Kernel Gaussiano

Este ejercicio corresponde al problema cuantitativo propuesto en la diapositiva 85 del bloque temático 2.2.

Enunciado general:

Imaginemos un kernel Gaussiano bidimensional de tamaño 9×9 .

APARTADO 1: Determinación de σ equivalente bajo la cota $k \geq 3\sigma$

Pregunta:

¿Cuál sería el valor equivalente de σ , asumiendo que k debe ser el entero más pequeño mayor o igual que 3σ ?

Respuesta breve:

Para un kernel 9×9 , el radio es $k = 4$. Al cumplirse $k = \lceil 3\sigma \rceil = 4$, el rango admisible es $1 < \sigma \leq 4/3$, resultando en el límite exacto de diseño que cubre el soporte $[-3\sigma, +3\sigma]$ el valor $\sigma = 4/3 \approx 1,333$.

Resolución:

Un kernel digital bidimensional simétrico centrado en el origen posee dimensiones impares $N \times N$, donde $N = 2k + 1$ y k denota el radio espacial o semiancho del filtro en píxeles (las coordenadas recorren el intervalo $[-k, +k]$ tanto en x como en y).

Para un kernel de tamaño 9×9 :

$$2k + 1 = 9 \implies 2k = 8 \implies k = 4$$

La condición impuesta en el enunciado establece que k es el menor número entero mayor o igual que 3σ , lo cual se expresa formalmente mediante la función techo (*ceiling*):

$$k = \lceil 3\sigma \rceil = 4$$

Por definición de la función techo, esto impone el rango de valores admisibles para 3σ :

$$3 < 3\sigma \leq 4 \implies 1 < \sigma \leq \frac{4}{3}$$

En el límite estricto de diseño óptimo donde el soporte del filtro trunca la distribución en exactamente 3σ a cada lado del centro:

$$3\sigma = 4 \implies \sigma = \frac{4}{3} \approx 1,3333$$

Con este valor de σ , el intervalo de discretización $[-4, 4]$ cubre el rango $[-3\sigma, +3\sigma]$, garantizando que el filtro contenga más del 99,73 % del volumen de la campana Gaussiana continua y minimizando los artefactos de discontinuidad en los bordes del kernel.

APARTADO 2: Tamaño equivalente tras 3 aplicaciones sucesivas mediante soporte discreto

Pregunta:

Si quisiéramos aplicar este mismo kernel de 9×9 un total de 3 veces consecutivas, ¿cuál sería el tamaño del kernel equivalente resultante?

Nota incluida en la diapositiva: Al aplicar el mismo kernel de tamaño N un total de k_{rep} veces, el tamaño del filtro equivalente viene dado por $N_{eq} = N + (k_{rep} - 1)(N - 1)$.

Respuesta breve:

Por la regla del soporte algebraico discreto $N_{eq} = N + (k_{rep} - 1)(N - 1)$, aplicando el filtro de tamaño $N = 9$ tres veces consecutivas ($k_{rep} = 3$), el tamaño del kernel discreto equivalente resultante es $9 + 2 \times 8 = \mathbf{25 \times 25}$.

Resolución:

La fórmula suministrada proviene de la propiedad del soporte algebraico de la convolución discreta. Si convolucionamos dos señales discretas de soportes finitos L_1 y L_2 , la longitud del soporte resultante es $L_1 + L_2 - 1$.

Al convolucionar $k_{rep} = 3$ veces un kernel con lado de dimensión $N = 9$:

- Primera convolución: tamaño $N = 9$.
- Segunda convolución: $(N) + (N - 1) = 9 + 8 = 17$.
- Tercera convolución: $(17) + (N - 1) = 17 + 8 = 25$.

Aplicando la expresión compacta:

$$N_{eq} = 9 + (3 - 1)(9 - 1) = 9 + 2 \times 8 = 9 + 16 = 25$$

El kernel discreto resultante equivalente que sintetiza las tres etapas en una sola pasada posee una dimensión de 25×25 .

APARTADO 3: Tamaño equivalente por efecto de desenfoque continuo Gaussiano ($\sigma\sqrt{N_{rep}}$)

Pregunta:

Si quisiéramos aplicar este mismo kernel de 9×9 un total de 3 veces consecutivas, ¿cuál sería el tamaño del kernel equivalente utilizando la propiedad de que la Gaussiana es cerrada bajo la convolución (es decir, convolucionar N_{rep} veces consecutivas con desviación σ equivale a una única convolución con una Gaussiana de desviación $\sigma\sqrt{N_{rep}}$)? En este caso, buscamos el filtro que proporcione el efecto de desenfoque (*blurring*) equivalente.

Respuesta breve:

Por la propiedad reproductiva de adición de varianzas Gaussianas, $\sigma_{eq} = \sigma\sqrt{3} = \frac{4}{\sqrt{3}} \approx 2,309$. Aplicando la regla $k_{eq} = \lceil 3\sigma_{eq} \rceil = \lceil 4\sqrt{3} \rceil = 7$, el tamaño de filtro con desenfoque equivalente es $2(7) + 1 = 15 \times 15$.

Resolución:

La función de densidad Gaussiana satisface la propiedad reproductiva bajo la operación de convolución continua. Si convolucionamos dos Gaussianas independientes con desviaciones estándar σ_1 y σ_2 :

$$G_{\sigma_1}(\mathbf{x}) * G_{\sigma_2}(\mathbf{x}) = G_{\sigma_{eq}}(\mathbf{x}) \quad \text{donde} \quad \sigma_{eq}^2 = \sigma_1^2 + \sigma_2^2$$

Al aplicar $N_{rep} = 3$ convoluciones consecutivas con la misma desviación estándar σ :

$$\sigma_{eq}^2 = \sigma^2 + \sigma^2 + \sigma^2 = 3\sigma^2 \implies \sigma_{eq} = \sigma\sqrt{3}$$

Tomando el valor de $\sigma = 4/3$ obtenido en el Apartado 1:

$$\sigma_{eq} = \frac{4}{3}\sqrt{3} = \frac{4}{\sqrt{3}} \approx 2,3094$$

Para diseñar un nuevo filtro discreto que proporcione exactamente este mismo efecto visual de suavizado respetando la regla inicial ($k_{eq} = \lceil 3\sigma_{eq} \rceil$):

$$3\sigma_{eq} = 3 \cdot \frac{4}{\sqrt{3}} = 4\sqrt{3} \approx 4 \times 1,73205 = 6,9282$$

Calculando el radio entero mínimo necesario:

$$k_{eq} = \lceil 6,9282 \rceil = 7$$

Calculamos la dimensión total del filtro simétrico:

$$N'_{eq} = 2k_{eq} + 1 = 2(7) + 1 = 15$$

Por consiguiente, el tamaño del kernel equivalente en cuanto a desenfoque es de 15×15 .

Discusión técnica comparativa (25 vs 15):

Existe una aparente discrepancia entre el resultado del Apartado 2 (25×25) y el del Apartado 3

(15×15). Su justificación radica en la naturaleza de la discretización:

- El cálculo algebraico 25×25 determina el soporte exterior estricto de convolucionar tres matrices truncadas. Sin embargo, los coeficientes en las coronas exteriores (para radios de distancia entre 8 y 12 píxeles respecto al centro) son producto de multiplicar valores ya infinitesimales de las colas de la campana. La suma de todos los pesos en esa franja externa representa menos del 0,05 % de la masa energética total del filtro.
- Al emplear la propiedad continua y trincar la Gaussiana equivalente en el umbral del 99,73 % ($3\sigma_{eq}$), un kernel de 15×15 captura prácticamente la totalidad de la función. Convolucionar con una matriz de 15×15 requiere $15^2 = 225$ operaciones por píxel (o $2 \times 15 = 30$ si se descompone de forma separable 1D), mientras que una matriz ingenua de 25×25 exige $25^2 = 625$ multiplicaciones (o $2 \times 25 = 50$). El tratamiento analítico mediante $\sigma\sqrt{3}$ permite reducir el coste computacional en un 64 % produciendo una salida indistinguible perceptualmente.

3.1.3 Bloque III: Cuestiones de Examen y Autoevaluación

Resolución exhaustiva de los siete quizzes oficiales de examen correspondientes a las diapositivas 86 a 89 del bloque temático 2.2.

QUIZ 1: Composición de filtros complejos (Diapositiva 86)

Enunciado:

¿Qué propiedad o teorema relativo a la convolución es el que, principalmente, permite la creación/-composición de filtros más complejos a partir de filtros más sencillos?

- a) Asociativa
- b) Conmutativa
- c) Distributiva
- d) Teorema de la Convolución

Respuesta correcta: **a) Asociativa**

Respuesta breve:

Opción **a) Asociativa**. Permite componer y preconvolucionar filtros en cascada ($K = K_1 * K_2$) antes de aplicarlos a la imagen en una sola pasada, base matemática que permite derivar analíticamente la Gaussiana ($\frac{\partial}{\partial x}(I * G) = I * \frac{\partial G}{\partial x}$).

Justificación analítica:

La propiedad asociativa establece que, dadas una imagen I y dos filtros espaciales K_1 y K_2 :

$$[I * K_1] * K_2 = I * [K_1 * K_2]$$

Esta identidad permite convolucionar previamente los dos núcleos elementales entre sí para sintetizar un único operador compuesto $K_{comp} = K_1 * K_2$ de forma estática (*offline*). Al procesar la imagen, basta realizar una única convolución directa $I * K_{comp}$ en lugar de filtrar secuencialmente toda la escena en dos etapas sucesivas.

El ejemplo fundamental en visión por computador es el operador de **Derivada de la Gaussiana** (DoG o gradiente suavizado): en lugar de suavizar la imagen completa con una Gaussiana G y luego derivar numéricamente el resultado ($D * (G * I)$), la propiedad asociativa permite derivar analíticamente la función Gaussiana primero y aplicar directamente el kernel precalculado:

$$\frac{\partial}{\partial x}(I * G) = I * \left(\frac{\partial G}{\partial x} \right)$$

Análisis de las opciones incorrectas:

- **b) Conmutativa** ($I * K = K * I$): Garantiza que el orden de los factores no altera el resultado, pero no permite la agrupación y prefusión de operadores.
- **c) Distributiva** ($I * (K_1 + K_2) = I * K_1 + I * K_2$): Rige la adición de filtros en paralelo, no la aplicación en cascada o composición jerárquica.
- **d) Teorema de la Convolución** ($\mathcal{F}\{I * K\} = \mathcal{F}\{I\} \cdot \mathcal{F}\{K\}$): Transforma la convolución en multiplicación puntual en el dominio frecuencial, pero la propiedad que rige el diseño algorítmico y la composición de máscaras espaciales en el dominio directo es la asociatividad.

QUIZ 2: Identificación del filtro de realce (Diapositiva 86)

Enunciado:

¿Qué tipo de filtro considera que es el siguiente?

$$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

- a) Suavizado Gaussiano
- b) Laplaciana de la Gaussiana
- c) Sharpening
- d) Ninguna de las anteriores

Respuesta correcta: **c) Sharpening**

Respuesta breve:

Opción **c) Sharpening**. El filtro descompone en Identidad más Laplaciano negativo ($I - \nabla^2 I$, máscara de desenfoque). Sus pesos suman 1 para no alterar el brillo en zonas homogéneas, mientras que su centro dominante 5 frente a vecinos -1 realza las altas frecuencias (bordes).

Justificación analítica:

Descomponemos matricialmente el kernel dado K :

$$K = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} = \delta(x, y) - \nabla^2$$

donde $\delta(x, y)$ representa el operador identidad (filtro delta de Dirac centrado que reproduce el píxel original intacto) y el segundo término es el Laplaciano discreto estándar con signo negativo ($-\nabla^2$).

La acción del filtro sobre un píxel I_c con vecinos ortogonales N_4 es:

$$(I * K)(x, y) = 5I_c - \sum_{(u,v) \in N_4} I(u, v) = I_c + \left(4I_c - \sum_{(u,v) \in N_4} I(u, v) \right) = I_c - \nabla^2 I$$

Esta formulación corresponde exactamente a la técnica clásica de **máscara de desenfoque** (**Unsharp Masking**) o **Sharpening**: 1. En regiones homogéneas (zonas planas), los vecinos tienen la misma intensidad que el centro ($I_c \approx \bar{I}_{adj}$). La suma de los coeficientes de la máscara es $5 - 4 = 1$, por lo que la ganancia es unitaria y la luminosidad media de la imagen no se altera. 2. En bordes o transiciones bruscas, la segunda derivada $-\nabla^2 I$ detecta la discontinuidad y suma un valor proporcional a la curvatura local, intensificando el salto de contraste y haciendo que los perfiles visuales parezcan más nítidos.

Análisis de las opciones incorrectas:

- **a) Suavizado Gaussiano:** Falsa. Los filtros de suavizado actúan como promedios ponderados paso bajo, por lo que todos sus coeficientes son estrictamente no negativos y normalizados para sumar 1.
- **b) Laplaciana de la Gaussiana (LoG):** Falsa. Los núcleos LoG son operadores derivados puros de segundo orden cuya suma de coeficientes es estrictamente cero (anulan áreas uniformes y solo responden a bordes).

QUIZ 3: Sigma equivalente tras tres etapas de suavizado (Diapositiva 87)

Enunciado:

Si suavizamos una imagen tres veces consecutivas con un filtro Gaussiano σ , ¿cuál sería el sigma equivalente que nos permitiría obtener el mismo resultado suavizando una única vez?

- a) $\sigma\sqrt{3}$
- b) $\sqrt{3}\sigma$
- c) $3\sqrt{\sigma}$
- d) Ninguna de las anteriores

Respuesta correcta: **a)** $\sigma\sqrt{3}$

Respuesta breve:

Opción **a)** $\sigma\sqrt{3}$. La convolución continua de filtros Gaussianos suma algebraicamente sus varianzas ($\sigma_{eq}^2 = \sigma^2 + \sigma^2 + \sigma^2 = 3\sigma^2$), lo que proporciona una desviación estándar equivalente resultante de $\sigma_{eq} = \sqrt{3\sigma^2} = \sigma\sqrt{3}$.

Justificación analítica:

La función Gaussiana unidimensional centrada viene dada por:

$$g(x; \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$$

Su transformada de Fourier continua es asimismo otra función Gaussiana:

$$\mathcal{F}\{g(x; \sigma)\}(\omega) = e^{-\frac{\sigma^2 \omega^2}{2}}$$

Por el teorema de convolución, convolucionar en el dominio espacial tres veces consecutivas con la misma desviación estándar σ equivale a multiplicar algebraicamente sus espectros de Fourier:

$$\mathcal{F}\{g * g * g\}(\omega) = e^{-\frac{\sigma^2 \omega^2}{2}} \cdot e^{-\frac{\sigma^2 \omega^2}{2}} \cdot e^{-\frac{\sigma^2 \omega^2}{2}} = e^{-\frac{3\sigma^2 \omega^2}{2}} = e^{-\frac{\sigma_{eq}^2 \omega^2}{2}}$$

Igualando los exponentes:

$$\sigma_{eq}^2 = 3\sigma^2 \implies \sigma_{eq} = \sqrt{3\sigma^2} = \sigma\sqrt{3}$$

El mismo principio aplica directamente a Gaussianas bidimensionales isotrópicas por su propiedad de separabilidad cartesiana ($G(x, y; \sigma) = g(x; \sigma)g(y; \sigma)$).

Análisis de las opciones incorrectas:

- **b) $\sqrt{3}\sigma$:** Falsa. En esta opción la raíz cuadrada abarca tanto al 3 como al parámetro σ . Desde una perspectiva de análisis dimensional, si σ se mide en píxeles, $\sqrt{\sigma}$ tendría unidades de píxeles^{1/2}, lo cual es físicamente absurdo.
- **c) $3\sqrt{\sigma}$:** Falsa. Padece del mismo error dimensional y confunde la suma de varianzas con la multiplicación lineal del factor de escala.

QUIZ 4: Análisis comparativo de Canny y Otsu (Diapositiva 87)

Enunciado:

¿Cuál de las siguientes afirmaciones sobre los algoritmos de Canny y Otsu es correcta?

- a) El detector de bordes de Canny se comporta como un filtro lineal dado que emplea filtrado Gaussiano para suavizar/emborronar la imagen y, a continuación, emplea filtrado lineal para calcular el gradiente.
- b) El umbral óptimo en el algoritmo de Otsu se obtiene minimizando la varianza inter-clase o maximizando la varianza intra-clase.
- c) En el algoritmo de Canny obtendremos bordes más discontinuos/fragmentados si reducimos el umbral bajo.
- d) Las tres respuestas anteriores son erróneas.

Respuesta correcta: **d) Las tres respuestas anteriores son erróneas**

Respuesta breve:

Opción **d) Las tres respuestas anteriores son erróneas**. Canny no es lineal (incorpora supresión de no máximos e histéresis); Otsu maximiza la varianza inter-clase (no la minimiza); y reducir el umbral bajo en Canny genera bordes más continuos y conectados, no más fragmentados.

Justificación analítica:

Demostramos la falsedad individual de cada una de las proposiciones anteriores:

- **Refutación de la afirmación a:**

El detector de Canny consta de cuatro etapas: 1. Suavizado Gaussiano (lineal). 2. Cálculo de gradiente por derivadas direccionales (lineal). 3. **Supresión de no máximos (Non-Maximum Suppression, NMS)**: Operación de adelgazamiento que compara la magnitud de cada píxel con

sus dos vecinos en la dirección normal al gradiente y anula el píxel si no es el máximo local estricto. Es una función de conmutación no lineal condicionada. 4. **Umbralización con histéresis:** Clasifica píxeles en fuertes, débiles y fondo mediante dos umbrales y preserva los débiles únicamente si forman una cadena conectada de 8-vecindad con un borde fuerte. La conectividad topológica y el umbralizado son operaciones fuertemente no lineales.

En consecuencia, el detector de Canny en su conjunto es un sistema **no lineal**.

– **Refutación de la afirmación b:**

En la formulación matemática de Otsu (1979), la varianza total de las intensidades de la imagen σ_T^2 es constante e invariante respecto al umbral t , satisfaciendo la descomposición:

$$\sigma_T^2 = \sigma_W^2(t) + \sigma_B^2(t)$$

donde $\sigma_W^2(t)$ es la varianza intra-clase (*within-class variance*, dispersión interna de los dos grupos) y $\sigma_B^2(t)$ es la varianza inter-clase (*between-class variance*, distancia cuadrática entre las medias ponderadas de ambos grupos).

Para encontrar la separación que maximiza la discriminación estadística entre fondo y objeto, Otsu busca el umbral óptimo t^* mediante:

$$t^* = \arg \max_t \sigma_B^2(t) \quad \Longleftrightarrow \quad t^* = \arg \min_t \sigma_W^2(t)$$

La opción b invierte deliberadamente el criterio al enunciar: “minimizando la varianza inter-clase o maximizando la intra-clase”, lo cual generaría la peor segmentación posible.

– **Refutación de la afirmación c:**

En el umbralizado por histéresis de Canny, el umbral alto T_{high} inicia contornos con alta confianza, mientras que el umbral bajo T_{low} actúa como cota de corte inferior para extender dichos contornos a través de zonas donde el gradiente se debilita.

Si **reducimos** T_{low} , admitimos gradientes más débiles para prolongar las cadenas de bordes, permitiendo cerrar discontinuidades y logrando contornos **más continuos y conectados**. Los bordes se volverían más discontinuos o fragmentados si *aumentáramos* T_{low} (o T_{high}).

Al ser a, b y c estrictamente falsas, la opción válida es la **d**.

QUIZ 5: Propiedades del gradiente y derivadas espaciales (Diapositiva 88)

Enunciado:

¿Cuál de las afirmaciones siguientes sobre procesamiento de imagen es cierta?

- a) Las derivadas en X detectan bordes horizontales, y las derivadas en Y detectan bordes verticales.
- b) La magnitud del gradiente puede tener valores negativos, dado que el gradiente puede hacer referencia a pendientes positivas o negativas.
- c) El gradiente en una imagen es mayor en las regiones planas que en las zonas con bordes.
- d) Ninguna de las anteriores es correcta.

Respuesta correcta: **d) Ninguna de las anteriores es correcta**

Respuesta breve:

Opción **d) Ninguna de las anteriores es correcta**. Las derivadas en X detectan bordes verticales (no horizontales); la magnitud euclídea del gradiente es una norma no negativa ($\|\nabla I\| \geq 0$); y en zonas planas el gradiente es nulo, alcanzando su máximo en los bordes.

Justificación analítica:

Examinamos la falsedad de las opciones planteadas:

– **Refutación de la afirmación a:**

La derivada parcial respecto a X , $\frac{\partial I}{\partial x}$, mide la tasa de variación de intensidad a lo largo del eje horizontal (de izquierda a derecha). Por definición geométrica, un cambio brusco horizontal se produce al atravesar un **borde vertical**. De forma análoga, $\frac{\partial I}{\partial y}$ mide el cambio vertical al atravesar un **borde horizontal**. Las correspondencias están cruzadas en el enunciado.

– **Refutación de la afirmación b:**

El vector gradiente espacial se define como $\nabla I = \left[\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y} \right]^T$. Aunque las derivadas parciales individuales pueden tomar valores negativos (indicando pendientes descendentes de intensidad), la **magnitud del gradiente** es su norma euclídea:

$$\|\nabla I\| = \sqrt{\left(\frac{\partial I}{\partial x}\right)^2 + \left(\frac{\partial I}{\partial y}\right)^2}$$

Por propiedades de la métrica euclídea, la norma es una cantidad real no negativa ($\|\nabla I\| \geq 0$). El signo o sentido de la transición queda codificado en el ángulo de orientación del gradiente ($\theta = \text{atan2}(I_y, I_x)$), jamás en su módulo.

– **Refutación de la afirmación c:**

En una región plana o uniforme de una imagen, la intensidad es prácticamente constante ($I(x, y) \approx c$). Sus derivadas parciales son nulas ($\partial I/\partial x \approx 0, \partial I/\partial y \approx 0$), lo que da lugar a una magnitud de gradiente próxima a cero. El gradiente alcanza sus valores más elevados precisamente en las fronteras entre regiones (bordes).

Al ser todas las anteriores afirmaciones erróneas, la única alternativa válida es la **d**.

QUIZ 6: Equivariancia de la convolución bidimensional (Diapositiva 88)

Enunciado:

¿Cuáles de las siguientes afirmaciones relativas a la operación de convolución son correctas?

- a) La convolución es equivariante con la traslación.
- b) La convolución no es equivariante con el escalado.
- c) La convolución no es equivariante con la rotación.
- d) Todas las anteriores son correctas.

Respuesta correcta: **d) Todas las anteriores son correctas**

Respuesta breve:

Opción **d)** **Todas las anteriores son correctas.** La convolución espacial es equivariante ante traslaciones, pero no lo es ante escalados (el filtro tiene tamaño fijo en píxeles) ni ante rotaciones (los kernels direccionales no giran con la imagen).

Justificación analítica:

Esta cuestión sintetiza los fundamentos matemáticos demostrados en la Cuestión 1 del Bloque I:

1. **Equivariancia con la traslación (Proposición a):** Es una propiedad intrínseca de los sistemas lineales e invariantes en el espacio (LSI). Desplazar la entrada una distancia $\mathbf{x}_0$ desplaza idénticamente la respuesta: $(I(\cdot - \mathbf{x}_0) * K)(\mathbf{x}) = (I * K)(\mathbf{x} - \mathbf{x}_0)$.
2. **Falta de equivariancia con el escalado (Proposición b):** El soporte de la máscara K está definido sobre una cuadrícula fija de píxeles. Si la imagen se amplía o reduce, el filtro muestrea una escala espacial distinta de la escena, requiriendo pirámides espaciales para compensarlo.
3. **Falta de equivariancia con la rotación (Proposición c):** Para un kernel direccional general (como Sobel o Prewitt), rotar la imagen no rota la orientación preferente del filtro, por lo que la convolución estándar no conmuta con rotaciones espaciales en el caso general.

Al ser las tres proposiciones verdaderas, la opción correcta es la **d**.

QUIZ 7: Reconocimiento del perfil de un kernel 1D (Diapositiva 89)

Enunciado:

La siguiente figura representa el perfil de un kernel 1D. ¿Qué tipo de filtro representa?

(Gráfica: Función unidimensional continua y suave, simétrica respecto a $x = 0$, no negativa en todo su soporte $[-7, 7]$, con un único valor máximo unitario en el origen y decaimiento asintótico hacia cero en los extremos).

- a) Un kernel de suavizado Gaussiano.
- b) Un kernel de primeras derivadas.
- c) Un kernel de segundas derivadas.
- d) Ninguno de los anteriores.

Respuesta correcta: **d) Ninguno de los anteriores**

Respuesta breve:

Opción **d)** **Ninguno de los anteriores.** Aunque el perfil muestra una campana simétrica positiva, no corresponde a un kernel de suavizado Gaussiano operativo porque **no está normalizado** (la suma de sus coeficientes supera con creces la unidad, $\sum K[n] \approx 6,32 \gg 1$, lo que multiplicaría la energía y saturaría la imagen) y porque en el origen $x = 0$ presenta una **cúspide puntiaguda no suave** en lugar de la tangente horizontal plana ($g'(0) = 0$) propia de una campana de Gauss.

Justificación analítica:

El análisis detallado de la gráfica revela por qué no encaja con ninguna de las tres primeras categorías:

1. **Ausencia de normalización para filtrado de suavizado:** Todo kernel de suavizado

(filtro paso bajo) debe estar estrictamente normalizado para conservar el nivel medio de gris y la energía lumínica de la escena: $\sum_{n=-k}^k K[n] = 1$ (o $\int_{-\infty}^{\infty} K(x) dx = 1$ en continuo). En la gráfica observamos que el valor máximo central es $K[0] = 1,0$ y los valores adyacentes son marcadamente positivos ($K[\pm 1] \approx 0,93$, $K[\pm 2] \approx 0,73$, $K[\pm 3] \approx 0,50$, etc.), sumando aproximadamente:

$$\sum_{n=-7}^7 K[n] \approx 1,0 + 2(0,93 + 0,73 + 0,50 + 0,28 + 0,13 + 0,05 + 0,02) \approx 6,28$$

Si se aplicase este filtro sobre una imagen digital, multiplicaría las intensidades de los píxeles por más de seis, quemando las altas luces y provocando una saturación total. Un filtro Gaussiano real de suavizado discretizado escala su amplitud dividiendo entre la suma total de sus coeficientes (o incorpora el factor analítico $\frac{1}{\sqrt{2\pi}\sigma}$), por lo que el valor central $K[0]$ en una ventana de radio 7 suele situarse en torno a $0,15 - 0,25$, jamás en $1,0$.

2. **Geometría del vértice en el origen (cúspide frente a derivada nula):** La función Gaussiana continua $g(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$ es de clase C^∞ (infinitamente diferenciable y suave). Su primera derivada analítica se anula idénticamente en el origen:

$$\frac{d}{dx}g(x) = -\frac{x}{\sigma^2}g(x) \implies \left. \frac{d}{dx}g(x) \right|_{x=0} = 0$$

Esto exige que el vértice de cualquier campana de Gauss presente una tangente horizontal perfectamente plana en su cumbre. Al examinar con detalle la ampliación de la diapositiva 89, la curva converge hacia $x = 0$ con una pendiente finita no nula a ambos lados, formando una cúspide o vértice anguloso (perfil característico de distribuciones exponenciales bilaterales tipo Laplace $e^{-|x|/\lambda}$ o funciones triangulares), no la curvatura cóncava diferenciable de una Gaussiana.

3. Incompatibilidad con operadores derivados:

- **Primeras derivadas (opción b):** Un operador de gradiente o primera derivada es una función **impar y antisimétrica** ($K(-x) = -K(x)$), con un lóbulo positivo y otro negativo cuya suma neta debe ser idénticamente cero para no responder ante regiones planas uniformes. La gráfica es estrictamente par y positiva.
- **Segundas derivadas (opción c):** Un operador de segunda derivada (como el Laplaciano o la Laplaciana de Gaussiana, LoG) presenta una morfología de sombrero mexicano (*Mexican hat*), con un pico central flanqueado por lóbulos laterales de signo opuesto cuya integral global se anula ($\sum K[n] = 0$).

Al no tratarse de un kernel de derivadas ni cumplir los requisitos matemáticos ni de normalización de un kernel de suavizado Gaussiano, la opción rigurosa y correcta es la **d**.

Análisis de las opciones descartadas:

- **a) Kernel de suavizado Gaussiano:** Falsa por carecer de normalización unitaria (amplificaría la energía por más de 6) y por presentar un vértice puntiagudo sin tangente horizontal nula en el origen.
- **b) Kernel de primeras derivadas:** Falsa porque las primeras derivadas son antisimétricas impares y suman cero.
- **c) Kernel de segundas derivadas:** Falsa porque las segundas derivadas requieren lóbulos negativos de compensación con integral nula.

Cuestiones y Autoevaluación de Machine Learning

4.1 Tema 3: Introducción al Aprendizaje Automático

Este capítulo recoge la resolución analítica, matemática y conceptual de la batería oficial de autoevaluación correspondiente al bloque temático 3.1 (*(A Very Brief) Introduction to Machine Learning*) de la asignatura Visión por Computador, impartida por el profesor Pablo Mesejo en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada.

El aprendizaje automático proporciona las herramientas para que un sistema aprenda patrones complejos directamente a partir de datos visuales. La colección de 32 cuestiones aborda los fundamentos de la disciplina: paradigmas de supervisión, formulación de modelos lineales y redes neuronales, dinámica del descenso de gradiente, prevención del sobreajuste mediante regularización, protocolos de validación rigurosos y fenómenos contemporáneos como el doble descenso. Las deducciones se apoyan en los manuales de referencia del curso: Abu-Mostafa et al. (2012), Bishop (2006), Hastie et al. (2009), Murphy (2022) y Szeliski (2022).

4.1.1 Bloque I: Concepto, Paradigmas de Aprendizaje y Modelos Básicos

QUIZ 1: Idea central del aprendizaje automático (Diapositiva 4)

Enunciado:

What is the central idea of machine learning?

- A. Writing explicit rules to describe the data
- B. Using data plus learning algorithms to uncover underlying patterns
- C. Building symbolic reasoning systems
- D. None of the above

Respuesta correcta: **B. Using data plus learning algorithms to uncover underlying patterns**

Respuesta breve:

El aprendizaje automático invierte el paradigma clásico de la programación: en lugar de codificar manualmente reglas deterministas, alimenta un algoritmo de optimización con observaciones para que infiera de forma autónoma los patrones subyacentes.

Justificación teórica y analítica:

Como formalizan Abu-Mostafa et al. (2012) en *Learning from Data* y François Chollet (2018), la programación tradicional toma como entrada reglas lógicas codificadas a mano y datos brutos para generar respuestas computadas. El aprendizaje automático toma como entrada datos empíricos y respuestas observadas para que un algoritmo de aprendizaje descubra la función o conjunto de reglas subyacentes:

$$\text{Datos} + \text{Algoritmo de Aprendizaje} \longrightarrow \text{Modelo} / \text{Patrones}$$

El objetivo operativo consiste en estimar una hipótesis $h \in \mathcal{H}$ que aproxime con fidelidad una función objetivo desconocida $f : \mathcal{X} \rightarrow \mathcal{Y}$.

Análisis de las opciones descartadas:

- **A. Writing explicit rules to describe the data:** Describe la programación imperativa y los sistemas expertos basados en reglas cableadas a mano, incapaces de escalar ante variaciones visuales complejas.
- **C. Building symbolic reasoning systems:** Caracteriza la inteligencia artificial simbólica clásica (sistemas basados en lógica de primer orden y grafos de conocimiento), que no aprenden de distribuciones estadísticas continuas.
- **D. None of the above:** La opción B sintetiza con exactitud la definición matemática del campo.

QUIZ 2: Condiciones de inaplicabilidad del aprendizaje automático (Diapositiva 4)*Enunciado:*

When can't we apply machine learning algorithms?

- A. There is no pattern to learn, only randomness
- B. We can describe our problem mathematically
- C. We have no data whatsoever
- D. We have data, but we know it's poorly annotated

Respuesta correcta: A, B y C representan las condiciones canónicas de inaplicabilidad; la opción D es la excepción donde el aprendizaje sí es viable

Respuesta breve:

Siguiendo el criterio de Abu-Mostafa et al. (2012), el aprendizaje automático no debe ni puede aplicarse si el fenómeno es puro ruido aleatorio (A), si disponemos de una fórmula determinista cerrada (B) o si carecemos por completo de datos (C). En cambio, disponer de datos con anotaciones ruidosas o deficientes (D) sí permite aplicar técnicas específicas como aprendizaje débilmente supervisado (*weakly supervised learning*).

Justificación teórica y analítica:

Abu-Mostafa et al. (2012, Cap. 1) formulan tres condiciones necesarias para que un problema requiera y admita aprendizaje automático:

1. Existe un patrón subyacente.
2. No es posible describirlo mediante una fórmula matemática exacta de forma analítica cerrada.
3. Se dispone de observaciones empíricas (datos).

Al examinar las opciones ante la pregunta de cuándo **no** podemos aplicar aprendizaje automático:

- Si no hay patrón y solo existe aleatoriedad estocástica pura (opción A), cualquier algoritmo se limitará a ajustar ruido espurio, derivando en un sobreajuste absoluto sin capacidad predictiva.
- Si el fenómeno cuenta con una derivación física o matemática exacta (opción B), como calcular el tiro parabólico en el vacío o liquidar una tabla de retenciones fiscales fijas, la solución cerrada es exacta, computacionalmente óptima y trivialmente verificable, haciendo innecesario e ineficiente recurrir a aproximaciones estadísticas.
- Si no existen datos (opción C), la formulación empírica $D = \{(\mathbf{x}_i, y_i)\}$ queda vacía y ningún parámetro puede actualizarse.
- Por el contrario, la opción D describe un escenario donde **sí** se aplica aprendizaje automático: el aprendizaje débilmente supervisado (*weakly supervised learning*) y el aprendizaje con etiquetas ruidosas (*learning with label noise*), apoyados en funciones de pérdida robustas o destilación de conocimiento.

QUIZ 3: Definición de aprendizaje autosupervisado (Diapositiva 8)

Enunciado:

Which statement best describes self-supervised learning?

- A. It uses human-provided labels for every training example.
- B. It learns a policy exclusively from rewards received from an environment.
- C. It derives supervisory signals or training targets from the data itself.
- D. It necessarily requires only one labeled example per class.

Respuesta correcta: C. It derives supervisory signals or training targets from the data itself

Respuesta breve:

El aprendizaje autosupervisado extrae la señal de entrenamiento a partir de la propia estructura intrínseca del dato sin intervención humana, formulando tareas pretextuales como predecir parches enmascarados, rotaciones o contrastar vistas aumentadas.

Justificación teórica y analítica:

En el aprendizaje autosupervisado (*Self-Supervised Learning, SSL*), el sistema convierte datos no etiquetados en problemas supervisados mediante tareas pretexto (*pretext tasks*). Por ejemplo, en visión por computador, modelos como SimCLR, MoCo o Masked Autoencoders (MAE) ocultan parches de una imagen y entrenan la red para reconstruirlos, o contrastan representaciones de distintas transformaciones geométricas de la misma escena. La función de pérdida se calcula comparando la predicción del modelo con la porción retenida u oculta del propio dato, prescindiendo de etiquetas manuales.

Análisis de las opciones descartadas:

- **A. It uses human-provided labels for every training example:** Definición canónica del aprendizaje totalmente supervisado (*Supervised Learning*).
- **B. It learns a policy exclusively from rewards received from an environment:** Describe el aprendizaje por refuerzo (*Reinforcement Learning*).
- **D. It necessarily requires only one labeled example per class:** Define el régimen de aprendizaje con un único ejemplo (*one-shot learning*), una variante de *few-shot learning*.

QUIZ 4: Identificación de paradigma a partir de vídeo demostrativo (Diapositivas 8 y 13)

Enunciado:

Observe the video and tell the machine learning paradigm that you think it's more likely being used? (Vídeo: DQN Breakout, DeepMind, <https://www.youtube.com/watch?v=TmPfTpjtdgg>)

- A. Supervised Learning
- B. Reinforcement Learning
- C. Unsupervised Learning
- D. Self-supervised Learning

Respuesta correcta: **B. Reinforcement Learning**

Respuesta breve:

El vídeo muestra el agente DQN (*Deep Q-Network*) de DeepMind aprendiendo a jugar a Atari Breakout mediante interacción con el entorno, recibiendo la puntuación del juego como recompensa escalar para optimizar su política de acciones.

Justificación teórica y analítica:

El experimento seminal de Mnih et al. (Nature, 2015) introdujo las redes neuronales profundas en el marco del aprendizaje por refuerzo (*Deep Q-Networks*). El agente observa el estado actual s_t (los últimos cuatro fotogramas de la pantalla en escala de grises), ejecuta una acción $a_t \in \{\text{izquierda, derecha, disparo, espera}\}$ guiada por una política ϵ -codiciosa, y recibe una recompensa r_t (incremento en el marcador del juego) junto con el nuevo estado s_{t+1} . La red aprende a estimar la función de valor óptima $Q^*(s, a)$ mediante la ecuación de Bellman:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right]$$

A partir del episodio 600, el agente descubre de forma autónoma la estrategia óptima: perforar un túnel lateral en los bloques para que la bola rebote en el techo a alta velocidad acumulando puntos de forma continuada.

Análisis de las opciones descartadas:

- **A. Supervised Learning:** Requeriría un conjunto de datos con millones de trayectorias previamente etiquetadas por jugadores expertos con la acción exacta a tomar en cada milisegundo.
- **C. Unsupervised Learning:** Agruparía estados o reduciría dimensionalidad sin noción de recompensa ni optimización de toma de decisiones secuenciales.
- **D. Self-supervised Learning:** Aunque la red convolucional extrae características visuales de los píxeles, la dinámica que guía la toma de decisiones y el aprendizaje temporal es un proceso de decisión de Markov gobernado por refuerzo.

QUIZ 5: Elementos actualizados durante el entrenamiento (Diapositiva 10)

Enunciado:

What do models “learn”/update in most cases?

- A. Features
- B. Labels
- C. Weights/parameters
- D. Only hyperparameters

Respuesta correcta: **C. Weights/parameters**

Respuesta breve:

Los modelos paramétricos aprenden ajustando iterativamente sus pesos y sesgos internos mediante gradiente descendente para minimizar el error empírico.

Justificación teórica y analítica:

En modelos de regresión, clasificadores lineales y redes neuronales, el conjunto de parámetros entrenables se denota como $\mathbf{w} \in \mathbb{R}^d$ o $\boldsymbol{\theta}$. El entrenamiento consiste en resolver un problema de optimización continua:

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}; \mathcal{D}_{train})$$

Los pesos modulan la importancia y combinación de las señales de entrada a lo largo de las capas.

Análisis de las opciones descartadas:

- **A. Features:** Son las variables o atributos de entrada extraídos de los datos. En arquitecturas profundas las representaciones intermedias emergen de los datos, pero lo que se actualiza físicamente mediante los gradientes son los pesos de los filtros.
- **B. Labels:** Constituyen la verdad terreno fija (*ground truth*), permaneciendo inmutables durante el ajuste.
- **D. Only hyperparameters:** Los hiperparámetros (tasa de aprendizaje η , factor de regularización λ , número de capas o tamaño de lote) no se optimizan por descenso de gradiente estándar, sino que se configuran externamente antes del entrenamiento o se ajustan mediante búsqueda sobre el conjunto de validación.

QUIZ 6: Modelos sin fase de entrenamiento explícita (Diapositiva 10)

Enunciado:

Which model has no proper “training” phase in the traditional sense and mainly memorizes data?

- A. Logistic regression
- B. k-Nearest Neighbors (KNN)
- C. Support Vector Machine
- D. Decision Tree

Respuesta correcta: **B. k-Nearest Neighbors (KNN)**

Respuesta breve:

KNN es un clasificador perezoso (*lazy learner*) basado en instancias: no ajusta ningún vector de parámetros durante el entrenamiento, limitándose a almacenar el conjunto de datos para calcular distancias en tiempo de inferencia.

Justificación teórica y analítica:

A diferencia de los modelos ansiosos (*eager learners*), el algoritmo de los k vecinos más cercanos no construye una función discriminante explícita durante una etapa de entrenamiento previa. Su coste de entrenamiento es $O(1)$ (o el tiempo requerido para indexar los datos en una estructura tipo *KD-Tree* o *Ball-Tree*). Toda la carga computacional se difiere a la fase de inferencia, donde, para cada punto de consulta $\mathbf{x}_q$, debe calcular la distancia métrica frente a todos los puntos almacenados:

$$d(\mathbf{x}_q, \mathbf{x}_i) = \|\mathbf{x}_q - \mathbf{x}_i\|_2$$

La predicción final se obtiene mediante votación mayoritaria entre los k puntos más próximos en clasificación, o mediante el promedio de sus valores en regresión.

Análisis de las opciones descartadas:

- **A. Logistic regression:** Requiere optimización numérica mediante descenso de gradiente o algoritmos cuasi-Newton para hallar el vector de coeficientes $\mathbf{w}$.
- **C. Support Vector Machine:** Resuelve un problema de programación cuadrática convexa para encontrar el hiperplano de margen máximo y los multiplicadores de Lagrange asociados a los vectores de soporte.
- **D. Decision Tree:** Construye de forma voraz una jerarquía de divisiones ortogonales evaluando la ganancia de información o la reducción del índice de impureza de Gini en cada nodo.

QUIZ 7: Modelos discriminativos frente a generativos (Diapositiva 9)

Enunciado:

Discriminative vs. Generative: which pairing is correct (assuming inputs X and outputs Y)?

- A. Discriminative $\rightarrow p(X)$; Generative $\rightarrow p(Y|X)$
- B. Discriminative $\rightarrow p(Y|X)$; Generative $\rightarrow p(X)$ or $p(X, Y)$
- C. Discriminative $\rightarrow p(X, Y)$; Generative $\rightarrow p(Y|X)$
- D. Both model $p(X)$ only

Respuesta correcta: **B. Discriminative $\rightarrow p(Y|X)$; Generative $\rightarrow p(X)$ or $p(X, Y)$**

Respuesta breve:

Los modelos discriminativos aprenden la frontera de decisión modelando la probabilidad condicional de las clases dadas las entradas $p(Y|X)$, mientras que los modelos generativos aprenden la distribución conjunta de las variables $p(X, Y)$ o la distribución marginal de los datos $p(X)$.

Justificación teórica y analítica:

Siguiendo la formalización canónica de Ng y Jordan (NeurIPS, 2001) y Murphy (2022):

- **Enfoque discriminativo:** Modela directamente la frontera entre clases estimando $p(Y|X)$ (ejemplos: regresión logística, redes neuronales clásicas) o una función de asignación determinista $f(X)$ sin probabilidades explícitas (ejemplo: SVM dura). No gasta capacidad representacional en modelar cómo se distribuyen las entradas X en el espacio.
- **Enfoque generativo:** Modela la distribución conjunta $p(X, Y) = p(X|Y)p(Y)$. Al capturar cómo se generan los datos para cada clase, puede aplicar la regla de Bayes para inferir la probabilidad posterior:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

Además, un modelo generativo permite muestrear sintéticamente nuevos datos artificiales $\mathbf{x}_{nuevo} \sim p(X)$ (ejemplos: Naive Bayes, Modelos de Mezclas Gaussianas, Variational Autoencoders y Modelos de Difusión).

Análisis de las opciones descartadas:

- **A y C:** Intercambian las distribuciones o invierten el orden epistemológico de modelado.
- **D:** Limitarse a modelar exclusivamente $p(X)$ corresponde a la estimación de densidad no supervisada, ignorando las etiquetas del problema discriminativo.

QUIZ 8: Señal de supervisión en aprendizaje por refuerzo (Diapositivas 8 y 13)*Enunciado:*

In reinforcement learning, the learner mainly uses?

- A. Exact labels for each state/situation
- B. A reward or penalty signal to learn a policy
- C. No label or feedback at all
- D. None of the above

Respuesta correcta: **B. A reward or penalty signal to learn a policy**

Respuesta breve:

En aprendizaje por refuerzo no existen respuestas correctas prescritas para cada estado; el agente recibe recompensas o castigos escalares derivados de sus acciones y aprende una política que maximiza la recompensa acumulada esperada.

Justificación teórica y analítica:

El paradigma de aprendizaje por refuerzo modela la interacción entre un agente y un entorno formulado como un Proceso de Decisión de Markov (MDP), formalizado mediante la tupla $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. El objetivo no consiste en ajustar una salida a una etiqueta supervisada instantánea y , sino en encontrar una política $\pi(a|s)$ que maximice el retorno acumulado descontado a lo largo del horizonte temporal:

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

donde $\gamma \in [0, 1)$ representa el factor de descuento temporal.

Análisis de las opciones descartadas:

- **A. Exact labels for each state/situation:** Corresponde al aprendizaje supervisado clásico o al aprendizaje por imitación (*behavioral cloning*), donde un experto humano suministra la acción perfecta para cada fotograma.
- **C. No label or feedback at all:** El aprendizaje no supervisado carece de retroalimentación; en RL existe una retroalimentación evaluativa continua mediante la recompensa escalar r_t .

4.1.2 Bloque II: Modelos Lineales, Optimización y Funciones de Pérdida

QUIZ 9: Formulación del descenso de gradiente (Diapositivas 12 a 15)*Enunciado:*

Machine learning training via gradient descent is framed as?

- A. A combinatorial enumeration problem
- B. An optimization problem minimizing a loss
- C. A constraint satisfaction problem
- D. A symbolic regression problem

Respuesta correcta: **B. An optimization problem minimizing a loss**

Respuesta breve:

El ajuste paramétrico mediante gradiente descendente formula el aprendizaje como la minimización continua y diferenciable de una función de coste sobre el espacio de pesos.

Justificación teórica y analítica:

El marco matemático predominante en el aprendizaje continuo es la minimización del riesgo empírico (*Empirical Risk Minimization*, ERM). Dados N ejemplos $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, se define una función de pérdida $\mathcal{L}(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N \ell(f(\mathbf{x}_i; \mathbf{w}), y_i)$. Al ser $\mathcal{L}$ diferenciable con respecto al vector de pesos $\mathbf{w}$, el algoritmo actualiza los parámetros en la dirección opuesta al vector gradiente:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}^{(t)})$$

donde $\eta > 0$ es la tasa de aprendizaje.

Análisis de las opciones descartadas:

- **A. A combinatorial enumeration problem:** La optimización combinatoria explora configuraciones discretas (como la búsqueda en grafos o el problema del viajante), computacionalmente intratable para espacios continuos de millones de parámetros.
- **C. A constraint satisfaction problem:** Propio de la programación con restricciones (CSP), donde se busca cualquier asignación de variables discretas que cumpla un conjunto de cláusulas lógicas, sin noción de gradiente de descenso continuo.
- **D. A symbolic regression problem:** Técnica basada en algoritmos genéticos que busca la expresión matemática analítica exacta del árbol sintáctico, no los pesos numéricos continuos de una arquitectura fija.

QUIZ 10: Correspondencia gráfica de modelos lineales clásicos (Diapositiva 11)

Enunciado:

Which model matches each graphical representation? (Referencia: Y. Abu-Mostafa, Caltech CS156, Slide 10). Señal acumulada de entrada: $s = \sum_{i=0}^d w_i x_i$.

- A. Linear classification (Perceptron)
- B. Linear regression
- C. Logistic regression

1. $h(\mathbf{x}) = \text{sign}(s)$
2. $h(\mathbf{x}) = s$
3. $h(\mathbf{x}) = \theta(s)$ con función logística suave

Respuesta correcta: 1-A, 2-B, 3-C

Respuesta breve:

El Perceptrón aplica la función signo no diferenciable para clasificar en clases binarias discretas (1-A); la regresión lineal utiliza la identidad manteniendo la salida real continua (2-B); y la regresión logística aplica la sigmoide suave para generar probabilidades en el intervalo $[0, 1]$ (3-C).

Justificación teórica y analítica:

Los tres modelos comparten idéntico núcleo lineal: el cálculo de la suma ponderada o producto escalar $s = \mathbf{w}^T \mathbf{x} = \sum_{i=0}^d w_i x_i$. La diferencia radica exclusivamente en la función de activación final aplicada a la señal s :

1. **Perceptrón de Rosenblatt (1 → A):** Aplica la función escalón duro o signo:

$$h(\mathbf{x}) = \text{sign}(s) \in \{-1, +1\}$$

Produce una partición estricta mediante un hiperplano separator en el espacio de características.

2. **Regresión Lineal (2 → B):** Aplica la función identidad:

$$h(\mathbf{x}) = s \in \mathbb{R}$$

La salida es un número real continuo sin acotar, optimizado habitualmente mediante el error cuadrático medio.

3. **Regresión Logística (3 → C):** Aplica la función de activación sigmoide o logística $\theta(s)$:

$$h(\mathbf{x}) = \theta(s) = \frac{1}{1 + e^{-s}} \in (0, 1)$$

Interpreta la salida como la probabilidad condicionada posterior de la clase positiva $P(y = 1|\mathbf{x})$. Al ser suave y diferenciable, permite calcular analíticamente el gradiente de la entropía cruzada.

QUIZ 11: Configuración canónica de pérdidas y funciones de activación de salida (Diapositivas 11 y 38)

Enunciado:

Fill the table with the most common configuration of loss function and output units activation function.

Problema	Función de Pérdida (<i>Loss Function</i>)	Activación de Salida (<i>Output Activation</i>)
Regresión	(a)	(d)
Clasificación Binaria	(b)	(e)
Clasificación Multiclase	(c)	(f)

Opciones disponibles:

- 1 – Sigmoid
- 2 – Mean Squared Error
- 3 – Binary Crossentropy
- 4 – Softmax
- 5 – Identity/Linear
- 6 – Categorical Crossentropy

Respuesta correcta: (a)=2, (b)=3, (c)=6, (d)=5, (e)=1, (f)=4

Respuesta breve:

- Regresión: Mean Squared Error (2) con activación Identidad (5).
- Clasificación Binaria: Binary Crossentropy (3) con activación Sigmoide (1).
- Clasificación Multiclase: Categorical Crossentropy (6) con activación Softmax (4).

Justificación teórica y analítica:

El emparejamiento entre la función de activación de la capa de salida y la función de pérdida no es arbitrario; se deriva del principio de estimación de máxima verosimilitud (*Maximum Likelihood Estimation*, MLE):

- **Regresión continua:** Asumiendo ruido gaussiano homocedástico $\epsilon \sim \mathcal{N}(0, \sigma^2)$, maximizar la verosimilitud equivale a minimizar el error cuadrático medio:

$$\mathcal{L}_{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

La activación debe ser la función identidad $g(s) = s$ para permitir predecir cualquier valor real en $(-\infty, +\infty)$.

- **Clasificación binaria:** Modelada bajo una distribución de Bernoulli con $y \in \{0, 1\}$. La activación sigmoide acota la predicción $\hat{y} = \sigma(s) \in (0, 1)$, y la pérdida de entropía cruzada binaria penaliza la discrepancia logarítmica:

$$\mathcal{L}_{BCE} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

- **Clasificación multiclase:** Modelada bajo una distribución categórica sobre K clases mutuamente excluyentes. La función Softmax normaliza las salidas garantizando $\sum_{k=1}^K \hat{y}_k = 1$:

$$\hat{y}_k = \frac{e^{s_k}}{\sum_{j=1}^K e^{s_j}}$$

Se combina con la entropía cruzada categórica $\mathcal{L}_{CCE} = -\sum_{k=1}^K y_k \log(\hat{y}_k)$, cuya derivada respecto a los logits de entrada adopta la forma lineal limpia $\frac{\partial \mathcal{L}}{\partial s_k} = \hat{y}_k - y_k$, evitando el desvanecimiento prematuro del gradiente.

QUIZ 13: Función del término de sesgo (*bias*) (Diapositivas 10 y 11)

Enunciado:

What role does the bias term play in linear models and neural nets?

- A. Forces the decision boundary through the origin
- B. Shifts decision boundaries and adds flexibility
- C. Prevents overfitting
- D. Normalizes inputs

Respuesta correcta: **B. Shifts decision boundaries and adds flexibility**

Respuesta breve:

El término de sesgo w_0 traslada el hiperplano de decisión alejándolo del origen de coordenadas, otorgando la flexibilidad necesaria para separar distribuciones no centradas.

Justificación teórica y analítica:

En una frontera de decisión lineal definida por $\mathbf{w}^T \mathbf{x} + b = 0$, los pesos $\mathbf{w} = (w_1, \dots, w_d)$ determinan el vector normal que define la orientación angular del hiperplano. Si el sesgo $b = w_0$ fuese idénticamente cero, la ecuación quedaría restringida a $\mathbf{w}^T \mathbf{x} = 0$, obligando al hiperplano a cruzar forzosamente por el origen $(0, 0, \dots, 0)$. El parámetro b actúa como una traslación rígida: la distancia euclídea

ortogonal desde el origen hasta el hiperplano es exactamente $\frac{|b|}{\|\mathbf{w}\|_2}$.

Análisis de las opciones descartadas:

- **A. Forces the decision boundary through the origin:** Es justo lo contrario; prescindir del sesgo es lo que fuerza a cruzar el origen.
- **C. Prevents overfitting:** Añadir un parámetro libre aumenta ligeramente la capacidad del modelo; la prevención del sobreajuste se logra mediante regularización o parada temprana.
- **D. Normalizes inputs:** La normalización es una transformación de preprocesamiento sobre los datos $\mathbf{x}$, no un parámetro entrenable del modelo.

QUIZ 31: Linealidad respecto a los pesos entrenables (Diapositiva 10)

Enunciado:

Which of the following models is linear with respect to its trainable weights?

- A. $y = w_0 + w_1x_1 + w_2x_2^2$
- B. $y = w_0 + w_1^2x_1 + w_2x_2$
- C. $y = w_0 + w_1w_2x_1 + w_3x_2$
- D. $y = w_0 + \sin(w_1x_1) + w_2x_2$

Respuesta correcta: **A.** $y = w_0 + w_1x_1 + w_2x_2^2$

Respuesta breve:

Un modelo es lineal si depende linealmente de sus parámetros entrenables $\mathbf{w}$, con independencia de que las variables de entrada hayan sufrido transformaciones no lineales como x_2^2 .

Justificación teórica y analítica:

En teoría del aprendizaje estadístico (Hastie et al., 2009; Bishop, 2006), la definición formal de modelo lineal exige que la hipótesis pueda expresarse como una combinación lineal de funciones base fijas $\phi(\mathbf{x})$:

$$f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T \phi(\mathbf{x}) = \sum_{j=0}^M w_j \phi_j(\mathbf{x})$$

En la opción A, definiendo el vector de características no lineales como $\phi(\mathbf{x}) = [1, x_1, x_2^2]^T$, la salida es un producto escalar directo $\mathbf{w}^T \phi(\mathbf{x})$. El gradiente $\nabla_{\mathbf{w}} y = \phi(\mathbf{x})$ no depende de los pesos, lo que garantiza que el problema de mínimos cuadrados posea una solución analítica convexa en forma cerrada (las ecuaciones normales).

Análisis de las opciones descartadas:

- **B.** $y = w_0 + w_1^2x_1 + w_2x_2$: Presenta una dependencia cuadrática con respecto a w_1 , perdiendo la linealidad en los parámetros y generando una superficie de error no lineal respecto a w_1 .
- **C.** $y = w_0 + w_1w_2x_1 + w_3x_2$: Contiene el producto bilineal entre dos pesos (w_1w_2), rompiendo la linealidad.
- **D.** $y = w_0 + \sin(w_1x_1) + w_2x_2$: El peso w_1 está atrapado dentro de la función trascendente trigonométrica no lineal seno.

QUIZ 32: Linealidad de una red de dos capas con activaciones lineales (Diapositivas 10 y 11)

Enunciado:

Consider a two-layer neural network with identity activations and no biases: $y = w_2(w_1x)$.

- A. Linear in x and linear in (w_1, w_2) .
- B. Linear in x , but nonlinear in (w_1, w_2) .
- C. Nonlinear in x , but linear in (w_1, w_2) .
- D. Nonlinear in both x and (w_1, w_2) .

Respuesta correcta: **B. Linear in x , but nonlinear in (w_1, w_2)**

Respuesta breve:

Al colapsar las funciones identidad, la salida es un múltiplo escalar directo de la entrada $y = (w_2w_1)x$ (lineal en x), pero depende del producto multilíneal de los pesos w_2w_1 (no lineal en los parámetros).

Justificación teórica y analítica:

Analizamos la relación de dependencia respecto a cada término:

- **Respecto a la variable de entrada x :** Agrupando los escalares, tenemos $y = c \cdot x$, donde $c = w_2w_1$. Cumple estrictamente las propiedades de aditividad $f(x + x') = f(x) + f(x')$ y homogeneidad $f(\alpha x) = \alpha f(x)$, por lo que es estrictamente lineal en x .
- **Respecto al vector conjunto de parámetros $\mathbf{w} = (w_1, w_2)$:** La función $g(w_1, w_2) = w_2w_1x$ es una forma bilíneal de segundo grado. Si duplicamos simultáneamente ambos pesos ($w'_1 = 2w_1, w'_2 = 2w_2$), la salida se multiplica por cuatro ($2w_2 \cdot 2w_1x = 4w_2w_1x \neq 2y$), violando la homogeneidad de grado 1.

Este principio demuestra por qué las redes neuronales profundas lineales, a pesar de colapsar a transformaciones lineales globales, exhiben dinámicas de entrenamiento no lineales y paisajes de optimización no convexos con puntos de silla (Saxe et al., 2014).

4.1.3 Bloque III: Sobreajuste, Dinámica de Optimización y Curvas de Aprendizaje

QUIZ 12: Definición y manifestación del sobreajuste (Diapositiva 50)

Enunciado:

Overfitting is most closely associated with:

- A. Too simple a model for a complex task
- B. Model memorizing training data and performing poorly on new data
- C. Too high the learning rate
- D. Infinite training time

Respuesta correcta: **B. Model memorizing training data and performing poorly on new data**

Respuesta breve:

El sobreajuste ocurre cuando un modelo con excesiva capacidad memoriza el ruido y las peculiaridades locales del conjunto de entrenamiento, perdiendo su habilidad para generalizar sobre datos no observados.

Justificación teórica y analítica:

Desde la perspectiva del compromiso sesgo-varianza (*bias-variance trade-off*), el sobreajuste (*overfitting*) se manifiesta cuando el modelo exhibe una varianza excesivamente elevada. El error de

entrenamiento (E_{in}) decrece hasta valores cercanos a cero, pero el error fuera de muestra (E_{out} o error de test) se degrada drásticamente:

$$E_{in} \ll E_{out}$$

El algoritmo ajusta tanto la señal subyacente legítima como el ruido estocástico inherente a las muestras recogidas.

Análisis de las opciones descartadas:

- **A. Too simple a model for a complex task:** Define el subajuste (*underfitting* o sesgo elevado), donde el modelo carece de capacidad para capturar la tendencia de los datos.
- **C. Too high the learning rate:** Provoca divergencia u oscilaciones numéricas en el optimizador, no necesariamente sobreajuste estadístico.
- **D. Infinite training time:** Aunque un tiempo prolongado puede favorecer el sobreajuste si la capacidad lo permite, no lo define conceptualmente.

QUIZ 14: Efecto del valor de la tasa de aprendizaje η (Diapositivas 16 y 21)

Enunciado:

Which learning rate matches each graphical representation? (Referencia: Y. Abu-Mostafa, Caltech CS156, Slide 21). Regla de actualización: $w_j := w_j - \eta \frac{\partial E_{in}(\mathbf{w})}{\partial w_j}$.

- A. η too small
- B. η too large
- C. η just right

1. Pasos diminutos descendiendo con extrema lentitud por la parábola.
2. Rebotes bruscos de un extremo a otro de las paredes del valle, aumentando la oscilación.
3. Saltos amplios pero convergentes que reducen el error con rapidez hacia el vértice mínimo.

Respuesta correcta: 1-A, 2-B, 3-C

Respuesta breve:

Un valor de η demasiado pequeño avanza a ritmo lento y costoso (1-A); uno excesivamente grande rebota entre las paredes del error y arriesga la divergencia (2-B); y un valor equilibrado desciende con rapidez y estabilidad al mínimo (3-C).

Justificación teórica y analítica:

La dinámica del descenso de gradiente depende de la curvatura local de la función de coste, descrita por los valores propios de la matriz Hessiana $\mathbf{H}$:

- **1 $\rightarrow$ A (η demasiado pequeña):** Cada actualización resta un vector gradiente escalado por una fracción reducida. Aunque la estabilidad matemática está asegurada en funciones convexas Lipschitz-continuas ($\eta < 2/\lambda_{max}$), el avance resulta ineficiente y susceptible de quedar atrapado en mesetas prolongadas.
- **2 $\rightarrow$ B (η demasiado grande):** Cuando $\eta > 2/\lambda_{max}$, el paso supera la anchura del valle parabólico. El optimizador evalúa el gradiente en puntos con pendientes opuestas cada vez mayores, generando oscilaciones divergentes donde $E_{in} \rightarrow \infty$.
- **3 $\rightarrow$ C (η adecuada):** El valor de paso aprovecha la curvatura local para recortar distancias de forma sustancial en las primeras etapas y aproximarse con precisión al mínimo.

QUIZ 21: Objetivo primordial del aprendizaje automático (Diapositivas 43 y 50)*Enunciado:*

What is our actual goal in machine learning?

- A. Minimizing training loss
- B. Reaching zero gradient
- C. Generalizing to unseen data
- D. Maximizing parameter count

Respuesta correcta: **C. Generalizing to unseen data***Respuesta breve:*

Minimizar la pérdida de entrenamiento es solo un medio instrumental; la meta genuina del aprendizaje automático es la generalización, logrando que el modelo rinda con precisión ante distribuciones y datos nunca vistos.

Justificación teórica y analítica:

Abu-Mostafa et al. (2012) destacan que cualquier memorización trivial puede reducir a cero la pérdida de entrenamiento sin aportar valor funcional. El objetivo teórico y práctico consiste en garantizar que el error fuera de muestra $E_{out} = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}}[\ell(f(\mathbf{x}), y)]$ sea lo menor posible. Las cotas clásicas de generalización como la dimensión VC acotan esta diferencia:

$$E_{out} \leq E_{in} + \Omega(N, \mathcal{H})$$

donde Ω cuantifica el coste de complejidad del espacio de hipótesis.

Análisis de las opciones descartadas:

- **A. Minimizing training loss:** Representa un sustituto instrumental (*proxy*), susceptible de inducir sobreajuste si se persigue de forma ciega.
- **B. Reaching zero gradient:** Señala un punto estacionario de optimización (que puede ser un mal mínimo local, un punto de silla o un punto sobreajustado).
- **D. Maximizing parameter count:** Aumentar parámetros sin control satura el modelo y encarece el cómputo sin garantizar una mejora predictiva.

QUIZ 26: Interpretación formal de curvas de aprendizaje clásicas (Diapositiva 55)*Enunciado:*

In the classical learning-curve picture, which of the following pairs of graphs is correct?

- Pareja (a):
 - Gráfica izquierda: Abscisas = # training examples. Error de entrenamiento arranca bajo y asciende asintóticamente; error de test arranca alto y descende asintóticamente hacia un límite común.
 - Gráfica derecha: Abscisas = # parameters. Error de entrenamiento descende monótonamente hacia cero; error de test describe una curva en “U” (desciende y luego sube por sobreajuste).
- Pareja (b): Intercambia los ejes horizontales entre ambas gráficas.

Respuesta correcta: **Pareja (a)***Respuesta breve:*

La pareja (a) describe con precisión la teoría estadística clásica: al aumentar los ejemplos, las curvas

de entrenamiento y test convergen asintóticamente; al aumentar los parámetros, el error de test forma una curva en U gobernada por el compromiso sesgo-varianza.

Justificación teórica y analítica:

El comportamiento de las curvas de aprendizaje se apoya en dos regímenes complementarios:

1. **Error frente al número de datos de entrenamiento (N):** Con pocas muestras (N pequeño), el modelo ajusta con facilidad todos los datos ($E_{train} \approx 0$), pero no representa la distribución general (E_{test} muy elevado). Al incorporar más datos, resulta más costoso ajustar cada punto individual y E_{train} sube ligeramente, mientras que E_{test} desciende al mejorar la representatividad estadística. Ambas curvas convergen hacia una asíntota determinada por el sesgo inherente del modelo y el ruido irreducible de Bayes.
2. **Error frente a la capacidad del modelo (número de parámetros):** Con pocos parámetros, el modelo sufre de alto sesgo (subajuste) y ambos errores son elevados. Al dotarlo de mayor capacidad, el error disminuye hasta un umbral óptimo. Si la capacidad continúa creciendo sin regularización, el error de entrenamiento se anula por memorización mientras que el error de test repunta bruscamente debido a la varianza.

QUIZ 29: Regularización L1 (Lasso) frente a L2 (Ridge) (Diapositivas 51 y 52)

Enunciado:

Which statement about L1 and L2 regularization is generally correct?

- A. L1 tends to encourage sparse weights; L2 tends to reduce weight magnitudes without necessarily setting them to zero.
- B. L1 and L2 always yield exactly the same optimal parameters.
- C. L2 tends to encourage sparse weights; L1 tends to reduce weight magnitudes without necessarily setting them to zero.
- D. Regularization guarantees perfect generalization.

Respuesta correcta: A. L1 tends to encourage sparse weights; L2 tends to reduce weight magnitudes without necessarily setting them to zero

Respuesta breve:

La penalización L1 induce dispersión forzando muchos pesos a cero exacto (selección implícita de características), mientras que L2 reduce proporcionalmente la magnitud de todos los pesos atenuando su impacto sin anularlos del todo.

Justificación teórica y analítica:

La regularización añade una penalización sobre los parámetros a la función de pérdida original: $\mathcal{L}_{reg}(\mathbf{w}) = \mathcal{L}_0(\mathbf{w}) + \lambda \mathcal{R}(\mathbf{w})$.

- **Regularización L1 (Lasso, $\mathcal{R}(\mathbf{w}) = \sum_j |w_j|$):** Las curvas de nivel de la norma ℓ_1 forman un poliedro rómbico con vértices afilados sobre los ejes de coordenadas. Cuando las curvas de nivel elípticas de la pérdida contactan con la región de restricción, el punto de tangencia se sitúa con alta probabilidad en una esquina, anulando por completo las componentes $w_j = 0$ correspondientes a variables prescindibles.
- **Regularización L2 (Ridge / Weight Decay, $\mathcal{R}(\mathbf{w}) = \frac{1}{2} \sum_j w_j^2$):** La región de restricción es una bola hipersférica suave sin aristas. La fuerza de contracción es directamente proporcional al peso ($\frac{\partial \mathcal{R}}{\partial w_j} = w_j$), comprimiendo las magnitudes de forma uniforme pero sin fijarlas en cero de forma analítica estricta.

Análisis de las opciones descartadas:

- **B:** L1 y L2 resuelven problemas geométricos distintos con diferentes soluciones óptimas.
- **C:** Invierte las propiedades de ambas regularizaciones.
- **D:** Ningún método garantiza una generalización perfecta, ya que subsisten el error irreducible de Bayes y el sesgo de representación.

QUIZ 30: El fenómeno del doble descenso (*Double Descent*) (Diapositiva 54)

Enunciado:

What does the double-descent phenomenon show about model size and generalization?

- A. Test error must always increase when a model has more parameters.
- B. The smallest model always generalizes best.
- C. Test error can rise and then fall again as model capacity increases.
- D. Training error must increase monotonically with model size.

Respuesta correcta: **C. Test error can rise and then fall again as model capacity increases**

Respuesta breve:

El fenómeno del doble descenso demuestra que, al sobrepasar el umbral de interpolación donde el modelo ajusta por completo los datos de entrenamiento, un incremento adicional en los parámetros puede hacer que el error de test vuelva a descender.

Justificación teórica y analítica:

Identificado formalmente por Belkin et al. (PNAS, 2019) y ampliado en modelos profundos por Nakkiran et al. (ICLR, 2020), el doble descenso reconcilia la teoría estadística tradicional con la práctica empírica de las redes neuronales:

1. **Régimen clásico (subparametrizado, $p < N$):** Al aumentar los parámetros p , el error de test describe la clásica forma de U, alcanzando un pico crítico en el **umbral de interpolación** ($p \approx N$), donde el modelo apenas tiene grados de libertad suficientes para anular el error de entrenamiento forzando pesos extremos.
2. **Régimen moderno (sobrepametrizado, $p \gg N$):** Una vez sobrepasado dicho límite, el espacio de soluciones que consiguen interpolar perfectamente los datos ($E_{train} = 0$) pasa a tener infinitas dimensiones. Los algoritmos basados en gradiente eligen de manera implícita la solución de mínima norma $\|\mathbf{w}\|_2$, lo que suaviza la función de interpolación y reduce de nuevo el error de generalización fuera de muestra.

Análisis de las opciones descartadas:

- **A:** Desmentida por el comportamiento del régimen sobrepametrizado.
- **B:** Un modelo excesivamente pequeño sufre de subajuste severo.
- **D:** El error de entrenamiento decrece de forma monótona conforme se incorporan más parámetros.

4.1.4 Bloque IV: Protocolos de Validación Experimental y Fuga de Datos

QUIZ 15: Forma extrema de validación cruzada (Diapositiva 44)

Enunciado:

Which validation scheme is the most extreme form of cross-validation?

- A. 5-fold CV
- B. Leave-one-out (LOO)
- C. Hold-out
- D. Bootstrap only

Respuesta correcta: **B. Leave-one-out (LOO)**

Respuesta breve:

Leave-one-out (LOO) representa el caso límite de k -fold cross-validation donde el número de particiones iguala al total de datos disponibles ($k = N$).

Justificación teórica y analítica:

En una validación cruzada con k particiones, el conjunto se subdivide en k subconjuntos disjuntos. En la variante extrema *Leave-One-Out* (LOO-CV), fijamos $k = N$: en cada una de las N iteraciones se entrena el modelo con $N - 1$ ejemplos y se valida sobre la única muestra retenida. Proporciona una estimación casi insesgada del error de generalización, pero a cambio de un coste computacional de entrenar el modelo N veces consecutivas.

Análisis de las opciones descartadas:

- **A. 5-fold CV:** Configuración intermedia estándar con $k = 5$, donde cada fold retiene el 20 % de las observaciones.
- **C. Hold-out:** Partición estática única (típicamente 70/30 u 80/20), sin rotación cruzada de subconjuntos.
- **D. Bootstrap:** Muestreo con reemplazamiento (*bagging*), diferente en formulación a la partición exhaustiva determinista de LOO.

QUIZ 16: Estrategia de evaluación para conjuntos reducidos y cómputo ligero (Diapositiva 44)

Enunciado:

With very small datasets and computationally inexpensive training, which evaluation strategy allows us to robustly estimate the performance of our model and, at the same time, allows us to maximize the number of training examples?

- A. 5-fold CV
- B. Leave-one-out (LOO)
- C. Hold-out
- D. No validation

Respuesta correcta: **B. Leave-one-out (LOO)**

Respuesta breve:

Con pocos datos y entrenamiento rápido, Leave-one-out maximiza los datos disponibles para entrenar ($N - 1$ muestras por partición) y ofrece una estimación robusta y determinista del rendimiento.

Justificación teórica y analítica:

Cuando N es reducido (por ejemplo, entre 30 y 100 muestras), realizar una partición fija tipo *Hold-out* (como dejar el 20 % para test) priva al modelo de datos valiosos para aprender y genera métricas de test con varianza elevada dependientes del azar del corte. Por contra, LOO-CV utiliza el 100 % de la información disponible para evaluar y entrena cada modelo con prácticamente todo el conjunto ($N - 1$), minimizando el sesgo pesimista de estimación. Dado que el modelo es

computacionalmente barato, entrenar N veces se completa en pocos segundos.

Análisis de las opciones descartadas:

- **A. 5-fold CV:** Entrena con el 80 % de los datos en cada partición, desechando una fracción mayor de datos frente al $(N - 1)/N$ de LOO.
- **C. Hold-out:** Desperdicia una parte sustancial de datos para entrenamiento y sufre de alta varianza estadística con muestras pequeñas.
- **D. No validation:** Carecer de validación impide cuantificar el rendimiento predictivo del sistema.

QUIZ 17: Estrategia de evaluación para modelos y datos masivos (Diapositiva 44)

Enunciado:

When dealing with very large models and datasets, which evaluation strategy is the most commonly used?

- A. 5-fold CV
- B. Leave-one-out (LOO)
- C. Hold-out
- D. No validation

Respuesta correcta: **C. Hold-out**

Respuesta breve:

En presencia de grandes volúmenes de datos y modelos pesados como redes neuronales profundas, la partición simple *Hold-out* (entrenamiento, validación y test) es el estándar indiscutible por razones de viabilidad computacional.

Justificación teórica y analítica:

Entrenar un modelo de aprendizaje profundo contemporáneo sobre millones de imágenes (por ejemplo, en ImageNet) o miles de millones de tokens requiere días de cómputo sobre clústeres de GPUs. Multiplicar dicho coste por 5 o 10 veces mediante validación cruzada resulta inviable en la práctica. Dado que el tamaño de muestra N es muy elevado, la ley de los grandes números asegura que un subconjunto de validación fijo retenido del 10 % contenga suficientes muestras para que la varianza de la métrica sea despreciable.

Análisis de las opciones descartadas:

- **A y B:** Multiplican el coste de computación por k o por N , inasumible en modelos a gran escala.
- **D:** Evaluar sobre entrenamiento sin conjunto de prueba incurre en ceguera operativa frente al sobreajuste.

QUIZ 18: Concepto de contaminación de datos (*Data Snooping*) (Diapositivas 44 a 46)

Enunciado:

“Data snooping” refers to:

- A. Using only training data for all decisions
- B. Any leakage of test information into modeling decisions
- C. Hyperparameter tuning on the validation set
- D. Random initialization of weights

Respuesta correcta: **B. Any leakage of test information into modeling decisions**

Respuesta breve:

El fenómeno de *Data Snooping* (o *Data Leakage*) describe cualquier filtración, directa o indirecta, de información procedente del conjunto de test hacia las decisiones de diseño, modelado o preprocesamiento del sistema.

Justificación teórica y analítica:

Abu-Mostafa et al. (2012, Principio 2) formalizan que si un conjunto de datos influye en cualquier fase del aprendizaje (selección de modelo, normalización, reducción de dimensionalidad o selección de hiperparámetros), su capacidad para evaluar objetivamente el error fuera de muestra queda comprometida. El rendimiento en test deja de ser una estimación insesgada de E_{out} y pasa a estar sesgado al alza de forma optimista:

E_{test} pasa a actuar como un E_{in} encubierto

Análisis de las opciones descartadas:

- **A. Using only training data for all decisions:** Es la práctica metodológica correcta para evitar sesgos.
- **C. Hyperparameter tuning on the validation set:** Es el procedimiento normativo estándar; la validación existe precisamente para ajustar hiperparámetros sin tocar el conjunto de test.
- **D. Random initialization of weights:** Es una técnica estándar de ruptura de simetría en optimización, independiente de la fuga de información entre particiones.

QUIZ 19: Situaciones prácticas de *Data Snooping* (Diapositivas 45 y 46)

Enunciado:

Which of the following situations could be considered data snooping:

- A. Someone calculates the mean and standard deviation to standardize the data using all the examples (i.e. prior to any data partition).
- B. We visualize the full dataset, and we have the impression that it's linearly separable. Consequently, we decide to use the Perceptron Learning Algorithm (PLA).
- C. We have 3 candidate hypotheses (SVM, MLP, RF), and we decide which one is the best by looking at its performance in the test set.
- D. All of the above

Respuesta correcta: **D. All of the above**

Respuesta breve:

Las tres acciones contaminan el modelado: calcular estadísticas sobre la totalidad de los datos filtra distribuciones futuras (A), inspeccionar visualmente el conjunto de test condiciona la elección del algoritmo (B) y comparar modelos finales sobre test convierte la prueba en una variable de decisión (C).

Justificación teórica y analítica:

Analizamos la transgresión metodológica de cada supuesto:

- **Situación A:** Calcular $\mu = \frac{1}{N} \sum x_i$ y σ sobre toda la base de datos antes de particionar propaga la escala y dispersión de las muestras de test dentro de las entradas de entrenamiento. Al escalar, el modelo de entrenamiento recibe información matemática derivada de datos que

en un entorno real aún no habrían ocurrido.

- **Situación B (Ejemplo de Abu-Mostafa, Diapositiva 45):** Inspeccionar el espacio completo contamina el ojo del investigador. Decidir usar el Perceptrón porque la nube completa parece separable implica que las etiquetas y posiciones de los puntos de test han sesgado el espacio de hipótesis $\mathcal{H}$ escogido.
- **Situación C (Diapositiva 46):** Evaluar SVM, MLP y Random Forest en test y seleccionar el que mejor resultado obtenga equivale a utilizar el conjunto de prueba como conjunto de validación. El test final debe ejecutarse sobre una única hipótesis final congelada.

QUIZ 20: Función de pérdida frente a métrica de evaluación (Diapositiva 43)

Enunciado:

Which pair below correctly distinguishes loss function vs evaluation metric?

- A. Loss: measures generalization; Metric: drives optimization
- B. Loss: optimized during training; Metric: used to evaluate performance
- C. Loss: computed only on test; Metric: only on train
- D. Both are always identical

Respuesta correcta: **B. Loss: optimized during training; Metric: used to evaluate performance**

Respuesta breve:

La función de pérdida es el objetivo matemático continuo y diferenciable que guía el descenso de gradiente durante el ajuste; la métrica de evaluación mide el rendimiento del sistema en términos comprensibles para la aplicación real.

Justificación teórica y analítica:

La distinción responde a condicionantes matemáticos y operacionales:

- **Función de pérdida (*Loss function*):** Diseñada para optimización numérica. Debe ser suave y casi con seguridad diferenciable respecto a los pesos $\mathbf{w}$ (por ejemplo, MSE, Huber Loss o Entropía Cruzada) para permitir el cálculo de gradientes por propagación hacia atrás.
- **Métrica de evaluación (*Evaluation metric*):** Diseñada para reflejar el objetivo del problema (por ejemplo, tasa de acierto *accuracy*, F1-Score, mAP en detección de objetos o BLEU en traducción). Estas métricas son escalonadas, discretas o no convexas, careciendo de derivadas informativas y no pudiendo optimizarse directamente mediante descenso de gradiente.

QUIZ 23: Coeficientes de escalado aplicables a los datos de test (Diapositivas 44 y 46)

Enunciado:

Which coefficients should we use to scale the test data?

- A. Those calculated in test
- B. Those calculated before in training
- C. It's the same

Respuesta correcta: **B. Those calculated before in training**

Respuesta breve:

Para transformar los datos de test deben utilizarse obligatoriamente los mismos parámetros de centrado y dispersión (media y desviación típica) calculados exclusivamente sobre el conjunto de

entrenamiento.

Justificación teórica y analítica:

Sea $\mathbf{x}_{test}$ una muestra del conjunto de prueba. Su estandarización se rige por:

$$\mathbf{x}_{test}^{escalado} = \frac{\mathbf{x}_{test} - \boldsymbol{\mu}_{train}}{\boldsymbol{\sigma}_{train}}$$

El conjunto de test emula el entorno de producción o inferencia en el mundo real, donde las muestras futuras llegan individualmente o en pequeños lotes sin permitir calcular una media o varianza poblacional fiable. Si recalculásemos $\boldsymbol{\mu}_{test}$ y $\boldsymbol{\sigma}_{test}$, estaríamos aplicando una proyección espacial distinta a la que utilizó el clasificador durante su ajuste, desalineando las fronteras aprendidas.

Análisis de las opciones descartadas:

- **A. Those calculated in test:** Introduce fuga de datos (*data snooping*) y genera distribuciones desacopladas con respecto al entrenamiento.
- **C. It's the same:** En muestras finitas $\boldsymbol{\mu}_{train} \neq \boldsymbol{\mu}_{test}$, por lo que los resultados numéricos difieren sensiblemente.

QUIZ 28: Monitorización en parada temprana (*Early Stopping*) (Diapositivas 44 y 53)

Enunciado:

When using early stopping to choose the training epoch, which dataset should normally be monitored?

- A. The test set, to directly optimize the reported final score.
- B. The validation set, while keeping the test set untouched.
- C. Only the training set, because validation is unnecessary.
- D. All data merged together after every epoch.

Respuesta correcta: **B. The validation set, while keeping the test set untouched**

Respuesta breve:

La parada temprana debe monitorizar la pérdida o métrica en el conjunto de validación, reservando el conjunto de test totalmente intacto para la evaluación final imparcial.

Justificación teórica y analítica:

La parada temprana (*early stopping*) funciona como una técnica de regularización que selecciona un hiperparámetro: el número óptimo de épocas de entrenamiento T^* . La pérdida de entrenamiento descende de forma continuada a lo largo del tiempo, mientras que la pérdida de validación descende hasta un punto mínimo y comienza a repuntar en cuanto el modelo sobreajusta. El criterio detiene el entrenamiento cuando la pérdida de validación deja de mejorar durante una paciencia fijada:

$$T^* = \arg \min_t \mathcal{L}_{val}(t)$$

Monitorizar el conjunto de test para tomar la decisión de frenar convertiría ese conjunto en una partición de entrenamiento implícita, invalidando su condición de estimador insesgado.

4.1.5 Bloque V: Preprocesamiento de Datos, Codificación y Arquitecturas

QUIZ 22: Propósito fundamental del escalado de variables (Diapositivas 47 y 48)

Enunciado:

Feature/variable scaling primarily aims to:

- A. Make variables integer-valued
- B. Put variables on similar ranges to speed optimization
- C. Reduce the need for cross-validation
- D. Put variables on similar ranges to better understand their meaning

Respuesta correcta: **B. Put variables on similar ranges to speed optimization**

Respuesta breve:

El escalado homogeneiza los rangos numéricos de las entradas, transformando curvas de nivel elípticas alargadas e inestables en contornos circulares regulares que aceleran y estabilizan la convergencia del descenso de gradiente.

Justificación teórica y analítica:

Cuando dos variables de entrada x_1 y x_2 poseen escalas dispares (por ejemplo, $x_1 \in [0, 1]$ y $x_2 \in [0, 100\,000]$), la matriz Hessiana de la función de coste presenta un número de condición $\kappa(\mathbf{H}) = \lambda_{\max}/\lambda_{\min} \gg 1$. En consecuencia, las curvas de nivel del error son elipses muy estrechas. El gradiente apunta casi ortogonal a la dirección del mínimo óptimo, provocando oscilaciones violentas en el eje de mayor pendiente y un avance lento en el eje plano.

Al aplicar estandarización ($z = \frac{x-\mu}{\sigma}$) o escalado Min-Max ($z = \frac{x-x_{\min}}{x_{\max}-x_{\min}}$), los contornos se tornan cuasi-esféricos, permitiendo que el gradiente apunte en línea recta hacia el mínimo y admitiendo tasas de aprendizaje η más altas. Además, en métodos basados en distancias (KNN, SVM) previene que la variable con magnitudes grandes monopolice la norma euclídea.

QUIZ 24: Cálculo de iteraciones de optimización por época (Diapositiva 41)

Enunciado:

Assume standard mini-batch gradient descent, processing every example once per epoch. If our training set contains 10000 images and the batch size is 8, how many optimization iterations are there per epoch?

- A. 1
- B. 10
- C. 1000
- D. None of the above

Respuesta correcta: **D. None of the above**

Respuesta breve:

El número exacto de iteraciones requeridas para procesar las 10000 imágenes en lotes de 8 es $10000/8 = 1250$ iteraciones por época, cifra no incluida en las opciones A, B o C.

Justificación teórica y analítica:

Por definición, una época (*epoch*) concluye cuando el algoritmo de optimización ha procesado cada

ejemplo del conjunto de entrenamiento exactamente una vez. Si el conjunto contiene N ejemplos y cada lote (*mini-batch*) procesa B muestras simultáneamente, el número de iteraciones (pasos de actualización del vector de pesos $\mathbf{w}$) por época es:

$$\text{Iteraciones por época} = \left\lceil \frac{N}{B} \right\rceil = \frac{10000}{8} = 1250$$

Al ser 1250 una división exacta sin resto y diferir sustancialmente de 1, 10 y 1000, la respuesta correcta es la opción D.

QUIZ 25: Elección de modelo según la naturaleza de los datos y sesgo inductivo (Diapositiva 57)

Enunciado:

Which statement about model choice and data type is correct?

- A. Deep learning is unsuitable for sequential data
- B. For tabular data, tree-based models often have the right inductive bias
- C. KNN is specifically designed to exploit spatial patterns in images
- D. An autoencoder is usually best for graphs

Respuesta correcta: **B. For tabular data, tree-based models often have the right inductive bias**

Respuesta breve:

Los modelos basados en árboles de decisión (como Random Forest, XGBoost y LightGBM) poseen el sesgo inductivo idóneo para datos tabulares heterogéneos, manejando escalas dispares, valores faltantes y fronteras ortogonales de forma natural.

Justificación teórica y analítica:

Cada familia de modelos asume hipótesis estructurales previas (*inductive bias*) sobre el tipo de función subyacente:

- **Datos tabulares y modelos de árboles:** Las tablas contienen variables categóricas y numéricas continuas con distribuciones asimétricas y relaciones complejas sin correlación espacial. Los árboles particionan el espacio con hiperplanos ortogonales a los ejes, siendo invariantes a transformaciones monótonas en las entradas y robustos frente a escalas desordenadas.
- **Datos con estructura espacial (imágenes):** El sesgo inductivo idóneo es la invariancia a traslaciones y la localidad espacial, explotadas de manera natural por los campos receptivos de las Redes Convolucionales (CNN) y la atención visual de los Vision Transformers (ViT), no por KNN.
- **Datos secuenciales (texto, series temporales):** Se procesan de forma habitual mediante arquitecturas recurrentes (LSTM, GRU) y mecanismos de autoatención (*Transformers*), desmintiendo rotundamente la opción A.
- **Grafos y datos relacionales:** La arquitectura diseñada para explotar su topología son las Redes Neuronales sobre Grafos (*Graph Neural Networks*, GNN), no los autoencoders convencionales.

QUIZ 27: Codificación de variables categóricas no ordinales (Diapositiva 47)

Enunciado:

For a non-ordinal categorical feature (e.g., color = red, green, blue), which encoding usually avoids

imposing an arbitrary numerical order?

- A. Assign the integers 0, 1 and 2 to the categories.
- B. Use one-hot encoding, with one indicator per category.
- C. Assign 10, 20 and 30, then standardize the resulting numbers.
- D. Apply cosine/sine transformation.

Respuesta correcta: **B. Use one-hot encoding, with one indicator per category**

Respuesta breve:

La codificación *one-hot* representa cada categoría como un vector binario ortogonal independiente, evitando introducir relaciones métricas o de orden artificiales entre elementos cualitativos.

Justificación teórica y analítica:

Si codificamos una variable no ordinal como el color asignando números enteros (rojo = 0, verde = 1, azul = 2), un modelo lineal o basado en distancias asumirá que:

$$\text{azul} > \text{verde} > \text{rojo} \quad \text{y que} \quad d(\text{rojo}, \text{azul}) = 2 \cdot d(\text{rojo}, \text{verde})$$

Esta jerarquía numérica carece de justificación física o semántica y distorsiona el aprendizaje.

La codificación *one-hot* (*One-Hot Encoding*) proyecta la variable a un espacio vectorial de dimensión K (donde K es el número de categorías), asignando un vector base canónico con un único 1 y el resto ceros:

$$\text{rojo} \mapsto [1, 0, 0]^T, \quad \text{verde} \mapsto [0, 1, 0]^T, \quad \text{azul} \mapsto [0, 0, 1]^T$$

Todos los vectores resultan ortogonales entre sí y equidistantes:

$$\|\mathbf{v}_i - \mathbf{v}_j\|_2 = \sqrt{2} \quad \forall i \neq j$$

eliminando cualquier ordenación preferente arbitraria.