

Double descent

Lucca Frachelle

abril 2026

Al aumentar la complejidad de un modelo, esperamos que el error de prueba primero baje y después suba: la conocida curva en U. El **double descent** introduce otro escenario. Cerca del umbral en el que el modelo puede ajustar exactamente los datos de entrenamiento, el error se dispara; con más parámetros, puede volver a bajar.

En un caso concreto — **regresión polinómica en una dimensión**, con una base de Legendre y soluciones de norma mínima — se puede conectar lo que aparece en las curvas con el álgebra del problema: la pseudoinversa y los valores singulares de la matriz de diseño.

Dos objetivos que chocan

Tenemos datos de entrenamiento $(x_i, y_i)_{i=1}^n$. En el ejemplo central, cada x_i es un escalar en $[-1, 1]$; en la discusión general conviene pensar $x_i \in \mathbb{R}^p$ como vector de covariables y $y_i \in \mathbb{R}$ como respuesta. Disponemos además de un conjunto de prueba $(x_j^*, y_j^*)_{j=1}^{n^*}$ para medir lo que importa en aplicaciones. En muchos modelos aparece, de forma explícita o implícita, el objetivo de **interpoliar** el entrenamiento,

$$\min_{\theta} \sum_{i=1}^n (f_{\theta}(x_i) - y_i)^2,$$

mientras que lo que realmente importa en aplicaciones es la **generalización**: que $f_{\theta}(x_j^*)$ se parezca a y_j^* en puntos nuevos. Esa tensión aparece en regresión lineal, en modelos basados en kernels, en redes neuronales cuando el entrenamiento minimiza el error cuadrático hasta casi cero, etc.

La historia clásica del sesgo-varianza sugiere que, al aumentar la complejidad, en algún momento el error de test empeora: la curva típica en la literatura pedagógica tiene forma de **U** o de tazón. El fenómeno de **double descent** muestra que, en ciertos regímenes (sobre todo con interpolación y reglas de elección de solución como la de **norma mínima**), la curva de error de test puede volver a **bajar** después de un pico dramático cerca del umbral donde el modelo pasa a poder interpolar. No es magia: es consecuencia de cómo se resuelve el problema lineal subyacente y de qué direcciones en el espacio de features quedan mal identificadas cuando la matriz de diseño está casi en rango deficiente.

Polinomios de Legendre como features (1D)

En el ejemplo numérico principal, x vive en $[-1, 1]$ y usamos la base de **polinomios de Legendre** $\{P_k\}_{k \geq 0}$, ortogonal en $[-1, 1]$ con peso 1:

$$\int_{-1}^1 P_k(x) P_{\ell}(x) dx = 0, \quad k \neq \ell.$$

Por ejemplo $P_0(x) = 1$, $P_1(x) = x$ y $P_3(x) = \frac{1}{2}(5x^3 - 3x)$. En lugar de $(1, x, x^2, \dots)$ tomamos

$$\tilde{\phi}_P(x_i) = (P_0(x_i), \dots, P_{P-1}(x_i))^T \in \mathbb{R}^P$$

como vector de covariables. La matriz de diseño tiene una fila por observación de entrenamiento:

$$\mathbf{X}_P = \begin{bmatrix} \tilde{\phi}_P(x_1)^T \\ \vdots \\ \tilde{\phi}_P(x_n)^T \end{bmatrix} \in \mathbb{R}^{n \times P}.$$

Eso reduce el mal condicionamiento típico de las potencias crudas $(1, x, x^2, \dots)$ cuando P crece: las columnas de monomios en una malla fija suelen volverse casi colineales, lo que encarece numéricamente la inversión de $X^\top X$. La base de Legendre no elimina por completo el mal condicionamiento en el **límite** $P \approx n$, pero ordena el problema de un modo más estable para el experimento que mostramos abajo.

Experimento: función verdadera, ruido, nodos y umbral

La función “verdadera” del ejemplo es

$$y(x) = 2x + \cos(25x),$$

y observamos $y_i = y(x_i) + \epsilon_i$ con ruido gaussiano $\epsilon_i \sim \mathcal{N}(0, 0.2^2)$, independiente entre observaciones.

Dónde se evalúa la función. Los valores de x no se toman en una malla equiespaciada en $(-1, 1)$. Con una malla uniforme, la matriz de Legendre evaluada en muchos puntos cercanos puede acercarse peligrosamente al rango deficiente de modo “accidental”, y el umbral $P = n$ deja de coincidir tan limpiamente con el fenómeno de interpolación **en la práctica numérica**. Por eso fijamos los x en nodos tipo **Chebyshev** en $(-1, 1)$ (transformación coseno de índices discretos), que reparten mejor los puntos hacia los extremos del intervalo y son el estándar en interpolación polinómica.

Train y test. Partimos la muestra en mitad entrenamiento y mitad prueba, pero **antes** permutamos los índices al azar. Si usáramos siempre “la primera mitad” como train, en una malla ordenada los puntos de entrenamiento podrían concentrarse en una parte del intervalo (por ejemplo solo $x < 0$), la matriz $\mathbf{X}_P$ sería atípica y las conclusiones visuales se confundirían con un artefacto de muestreo. La permutación hace que train y test vean soporte comparable en $[-1, 1]$.

El **umbral de interpolación** es $P = n$: tantas columnas (features) como filas (puntos de train) en $\mathbf{X}_P$. Para $P > n$, en rango fila completo, existe una infinidad de β que interpolan exactamente los y_i del train; entre todas ellas se elige la de **norma euclídea mínima**, que es la que entrega la pseudo-inversa de Moore–Penrose en ese régimen.

El ajuste queda entonces:

$$\hat{\beta}_P^{\min} = \mathbf{X}_P^+ \mathbf{y},$$

con la forma cerrada $(\mathbf{X}_P^T \mathbf{X}_P)^{-1} \mathbf{X}_P^T$ si $P \leq n$ (rango columna completo) y $\mathbf{X}_P^T (\mathbf{X}_P \mathbf{X}_P^T)^{-1}$ si $P > n$ (rango fila completo, régimen interpolante).

La siguiente figura muestra el MSE en entrenamiento y en prueba frente a P (escala logarítmica en ambos ejes). El patrón típico del double descent aparece de inmediato: error alto con pocos parámetros, un **pico** cerca de $P = n$, y luego descenso o estabilización cuando P supera claramente a n . En el tramo $P < n$ marcamos además un P^* que minimiza el MSE de test en **esta** realización fija del experimento (línea vertical punteada verde), para contrastarlo con el umbral $P = n$ (gris).

El train baja monótonamente cuando el modelo gana flexibilidad; el test no: cerca de $P = n$ el sistema está mal condicionado y pequeñas perturbaciones (ruido, geometría de los datos) se amplifican. Más allá del umbral, la elección de la solución de norma mínima vuelve a imponer una forma “suave” al coeficiente y el error de test puede recuperarse.

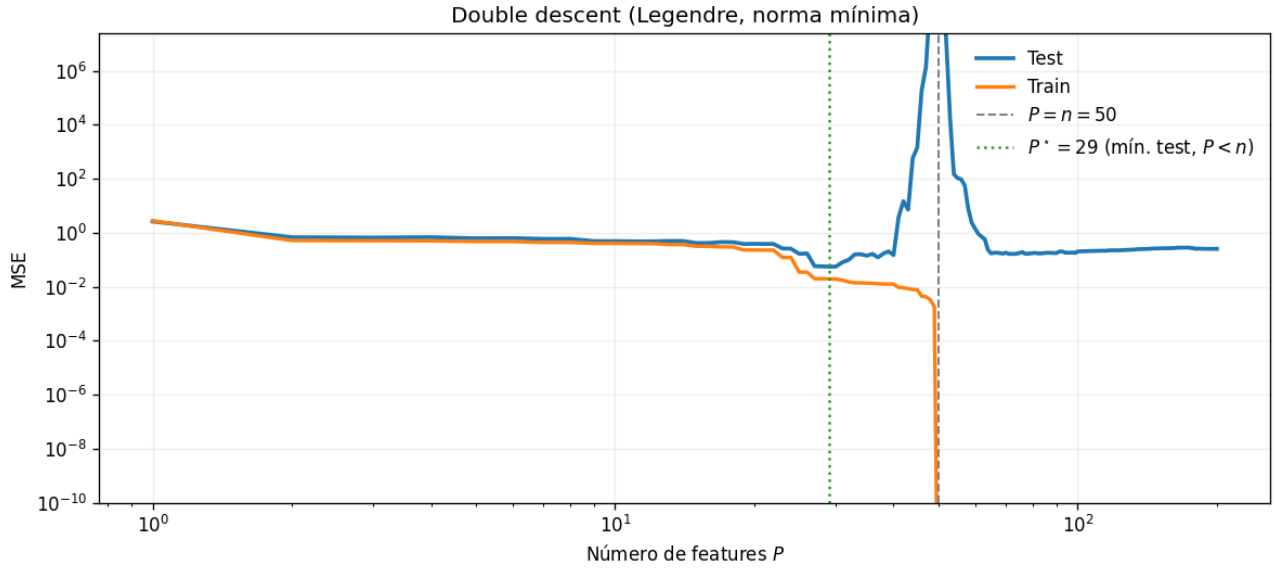


Figura 1: MSE en train y en test frente al número de features P . Línea gris: $P = n$; verde: P^* que minimiza el MSE de test con $P < n$ en esta muestra.

Qué muestran las curvas ajustadas

El panel siguiente usa **siempre los mismos** puntos de entrenamiento y cambia solo P . Así se ve qué significa el gráfico anterior en términos de la función $\hat{y}(x)$ frente a la verdadera $y(x) = 2x + \cos(25x)$.

1. **Pocos parámetros** (P pequeño): el modelo es rígido; la onda de alta frecuencia $\cos(25x)$ casi no entra y el ajuste queda suave pero sesgado respecto de la verdad.
2. **Mejor P antes del umbral** ($P^* < n$): suele aparecer un compromiso razonable entre capturar estructura y no inflar varianza; la curva se acerca más a $y(x)$.
3. **Umbral** ($P = n$): el sistema cuadrado es delicado; direcciones con valores singulares muy pequeños amplifican el ruido y el ajuste puede oscilar de forma dramática entre los puntos de entrenamiento.
4. **Muchos parámetros con min-norm** ($P \gg n$): el modelo interpola el train pero la regularización implícita de la norma mínima tiende a producir predicciones más suaves fuera de los datos de entrenamiento que en el caso “en el borde” $P = n$, lo que en este ejemplo puede **reducir** de nuevo el error de test.

Esa secuencia es la narrativa visual que acompaña la curva en forma de doble descenso: el pico en $P \approx n$ no es un artefacto del dibujo, sino que corresponde a ajustes casi singulares que explotan entre los datos, mientras que el régimen $P \gg n$ con norma mínima puede verse mucho más benigno a ojo.

Pseudo-inversa de Moore–Penrose (lenguaje unificado)

Para una matriz de diseño $X \in \mathbb{R}^{N \times P}$ (aquí N es el número de puntos de entrenamiento) y el vector de respuestas $Y \in \mathbb{R}^N$, la solución que unifica los dos regímenes (cuando los rangos son completos en el sentido habitual) es

$$\hat{\beta} = X^+ Y,$$

donde X^+ es la **pseudo-inversa de Moore–Penrose**. En forma cerrada, fuera de los casos límite de rango deficiente, se recuerda como

Intuición visual: misma muestra de entrenamiento

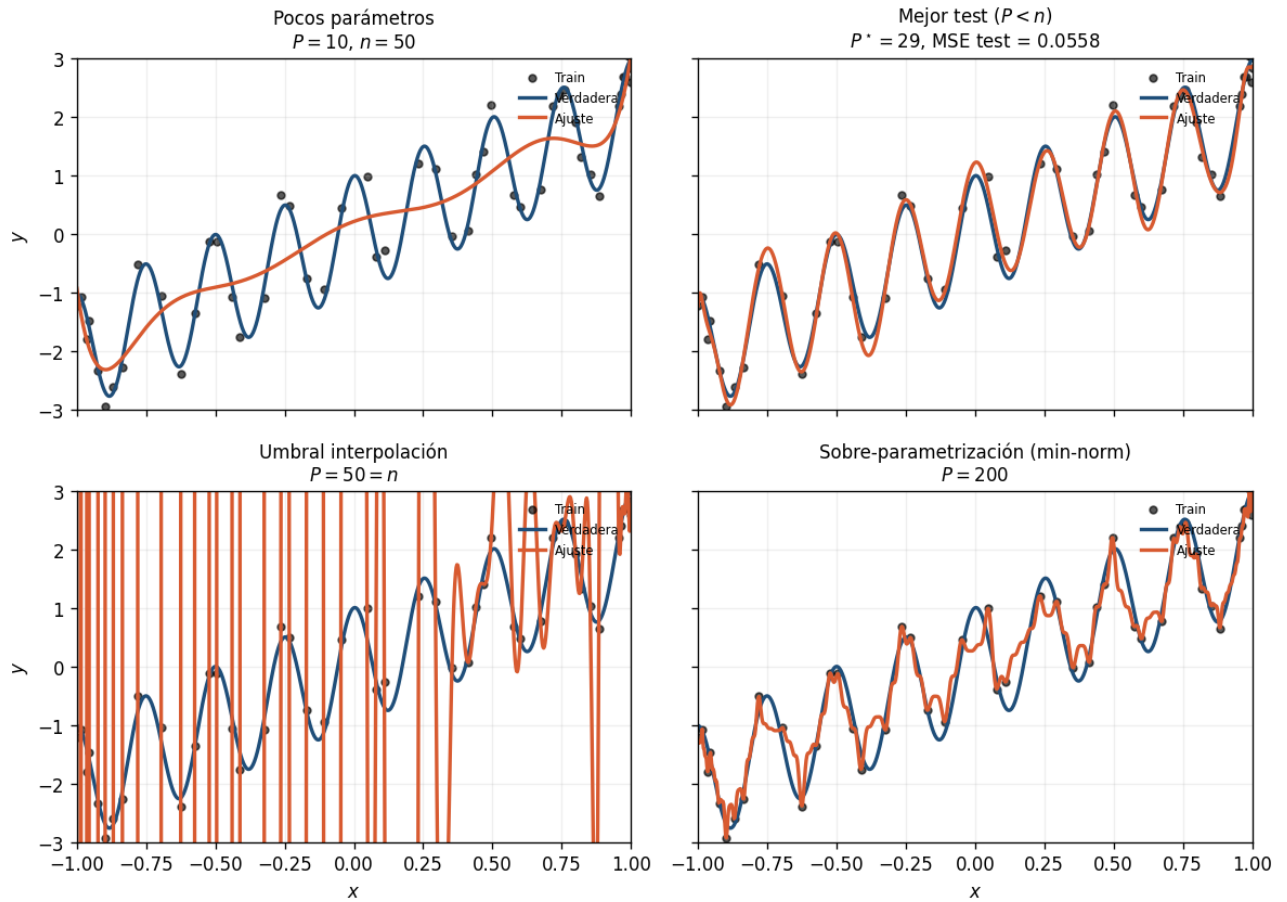


Figura 2: Misma muestra de entrenamiento; cuatro valores de P : pocos parámetros, óptimo $P^* < n$, umbral $P = n$, y P grande con solución de norma mínima.

$$X^+ = \begin{cases} (X^\top X)^{-1} X^\top & \text{si } P \leq N \text{ y } X \text{ tiene rango columna completo,} \\ X^\top (X X^\top)^{-1} & \text{si } P \geq N \text{ y } X \text{ tiene rango fila completo.} \end{cases}$$

En el régimen **sobre-parametrizado interpolante** ($P > N$ y rango fila completo), X^+Y es exactamente el vector $\hat{\beta}$ de **norma euclídea mínima** entre todos los que satisfacen $X\beta = Y$:

$$\hat{\beta} = \arg \min_{\beta} \|\beta\|_2^2 \quad \text{sueto a } X\beta = Y.$$

Cómo predecimos en un punto de test

Sea $\mathbf{x}_{\text{test}} \in \mathbb{R}^P$ el vector de features del mismo tipo que las filas de X (en nuestro ejemplo, valores de Legendre hasta orden $P - 1$). Entonces

$$\hat{y}_{\text{test,under}} = \mathbf{x}_{\text{test}}^\top (X^\top X)^{-1} X^\top Y \quad (P < N, \text{ rango columna completo}),$$

$$\hat{y}_{\text{test,over}} = \mathbf{x}_{\text{test}}^\top X^\top (X X^\top)^{-1} Y \quad (P > N, \text{ rango fila completo}).$$

En la práctica, conviene obtener $\hat{\beta}$ y las predicciones con métodos basados en la **descomposición en valores singulares** u otras factorizaciones estables, en lugar de invertir matrices de Gram de forma directa cuando P o N son grandes. Las fórmulas anteriores son la referencia algebraica que conecta el experimento con el análisis del error.

Modelo de referencia y error de predicción

Para analizar el error sin asumir que el mundo es lineal, introducimos notación: escribimos

$$Y = X\beta^* + E,$$

donde β^* representa los coeficientes del **mejor predictor lineal** en un sentido ideal y E recoge todo lo que esa clase lineal no puede explicar (ruido, o componentes no lineales). El valor “ideal” en test es $y_{\text{test}}^* = \mathbf{x}_{\text{test}} \cdot \beta^*$. En general β^* no tiene que ser los coeficientes de un “mundo lineal verdadero”; es una convención analítica que permite separar la parte del error que se parece a **varianza** (sensible al ruido E) de la parte que se parece a **sesgo** (sensible a la incapacidad del predictor lineal de representar la verdad).

Descomposición en valores singulares

Escribimos la SVD de X en la forma reducida que conviene al rango R :

$$X = U \Sigma V^\top,$$

con $U \in \mathbb{R}^{N \times R}$, $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_R)$, $V \in \mathbb{R}^{P \times R}$ y valores singulares **estrictamente positivos** $\sigma_1 \geq \dots \geq \sigma_R > 0$. Las columnas u_r de U y v_r de V son ortonormales en sus espacios. La pseudo-inversa actúa sobre la parte de Y que cae en el espacio columna de X mediante factores $1/\sigma_r$; por eso los **valores singulares pequeños** son el corazón técnico del pico cerca de $P \approx n$.

Caso $P < N$ (sub-parametrizado): derivación del error

Partimos de la fórmula del estimador de mínimos cuadrados (coincide con X^+Y en rango columna completo). Sustituyendo $Y = X\beta^* + E$,

$$\begin{aligned}
\hat{y}_{\text{test}} &= \mathbf{x}_{\text{test}}^\top (X^\top X)^{-1} X^\top Y \\
&= \mathbf{x}_{\text{test}}^\top (X^\top X)^{-1} X^\top (X\beta^* + E) \\
&= \mathbf{x}_{\text{test}}^\top (X^\top X)^{-1} X^\top X\beta^* + \mathbf{x}_{\text{test}}^\top (X^\top X)^{-1} X^\top E \\
&= \underbrace{\mathbf{x}_{\text{test}}^\top \beta^*}_{= y_{\text{test}}^*} + \mathbf{x}_{\text{test}}^\top (X^\top X)^{-1} X^\top E.
\end{aligned}$$

Por tanto el error en test respecto del mejor predictor lineal en la notación adoptada es

$$\hat{y}_{\text{test}} - y_{\text{test}}^* = \mathbf{x}_{\text{test}}^\top (X^\top X)^{-1} X^\top E.$$

Sustituyendo $X = U\Sigma V^\top$ y usando las identidades estándar para $(X^\top X)^{-1} X^\top$, se obtiene la forma modal

$$\hat{y}_{\text{test}} - y_{\text{test}}^* = \mathbf{x}_{\text{test}}^\top V \Sigma^+ U^\top E = \sum_{r=1}^R \frac{1}{\sigma_r} (\mathbf{x}_{\text{test}} \cdot v_r) (u_r \cdot E),$$

donde $\Sigma^+ = \text{diag}(1/\sigma_1, \dots, 1/\sigma_R)$. Es decir, el error es una **suma sobre modos** del operador lineal que mapea el ruido (y el residual mal modelado) del espacio de las observaciones al espacio de predicción.

En la descomposición algebraica anterior, para $P < N$ **solo** aparece el canal que pasa por E en la expresión cerrada del error respecto de y_{test}^* ; el término de **sesgo de parametrización** $(X^\top (XX^\top)^{-1} X - I)\beta^*$ es propio del régimen $P > N$.

Caso $P > N$ (sobre-parametrizado, min-norm): derivación

Usamos la predicción $\hat{y}_{\text{test,over}} = \mathbf{x}_{\text{test}}^\top X^\top (XX^\top)^{-1} Y$ y otra vez $Y = X\beta^* + E$:

$$\begin{aligned}
\hat{y}_{\text{test,over}} &= \mathbf{x}_{\text{test}}^\top X^\top (XX^\top)^{-1} (X\beta^* + E) \\
&= \mathbf{x}_{\text{test}}^\top X^\top (XX^\top)^{-1} X\beta^* + \mathbf{x}_{\text{test}}^\top X^\top (XX^\top)^{-1} E.
\end{aligned}$$

Recordando $y_{\text{test}}^* = \mathbf{x}_{\text{test}}^\top \beta^*$, el error es

$$\begin{aligned}
\hat{y}_{\text{test,over}} - y_{\text{test}}^* &= \mathbf{x}_{\text{test}}^\top (X^\top (XX^\top)^{-1} X - I)\beta^* \\
&\quad + \mathbf{x}_{\text{test}}^\top X^\top (XX^\top)^{-1} E.
\end{aligned}$$

El segundo sumando admite la misma reexpresión modal que en $P < N$:

$$\mathbf{x}_{\text{test}}^\top X^\top (XX^\top)^{-1} E = \sum_{r=1}^R \frac{1}{\sigma_r} (\mathbf{x}_{\text{test}} \cdot v_r) (u_r \cdot E).$$

Así

$$\hat{y}_{\text{test,over}} - y_{\text{test}}^* = \mathbf{x}_{\text{test}}^\top (X^\top (XX^\top)^{-1} X - I)\beta^* + \sum_{r=1}^R \frac{1}{\sigma_r} (\mathbf{x}_{\text{test}} \cdot v_r) (u_r \cdot E).$$

Aparece un **término extra** que no está en el caso $P < N$: el que lleva $(X^\top (XX^\top)^{-1} X - I)\beta^*$. En estadística se asocia al **sesgo** respecto de β^* en el sentido siguiente: con $P > N$, los N datos solo identifican un subespacio de dimensión a lo sumo N dentro de $\mathbb{R}^P$. La matriz $X^\top (XX^\top)^{-1} X$ es la **proyección ortogonal** sobre el subespacio fila de X (equivalentemente, sobre el span de las columnas de $X^\top$ en $\mathbb{R}^P$). La componente de β^* **ortogonal** a ese subespacio no puede recuperarse del

entrenamiento; al comparar con $\mathbf{x}_{\text{test}}^\top \beta^*$ completo, esa componente “faltante” genera error sistemático en test.

El **segundo término** (la suma en r) es análogo al de $P < N$ y es el que conecta con el **pico** del double descent cuando ciertos σ_r son muy pequeños.

Resumen compacto de las dos fórmulas

$P < N$:

$$\hat{y}_{\text{test}} - y_{\text{test}}^* = \sum_{r=1}^R \frac{1}{\sigma_r} (\mathbf{x}_{\text{test}} \cdot v_r) (u_r \cdot E).$$

$P > N$:

$$\hat{y}_{\text{test,over}} - y_{\text{test}}^* = \mathbf{x}_{\text{test}}^\top (X^\top (XX^\top)^{-1} X - I) \beta^* + \sum_{r=1}^R \frac{1}{\sigma_r} (\mathbf{x}_{\text{test}} \cdot v_r) (u_r \cdot E).$$

¿Por qué explota el error?

En el caso $P < N$, el error de predicción en test (y por tanto el **error cuadrático** en test) descompone la incertidumbre en modos singulares. Con la SVD $X = U\Sigma V^\top$ y valores singulares no nulos $\sigma_1, \dots, \sigma_R$, ya vimos

$$\hat{y}_{\text{test}} - y_{\text{test}}^* = \sum_{r=1}^R \frac{1}{\sigma_r} (\mathbf{x}_{\text{test}} \cdot v_r) (u_r \cdot E). \quad (1)$$

La ecuación (1) es **central**: el error depende de una **interacción** entre tres cantidades (en el modo r -ésimo):

1. **Cuánto “varían” las columnas de entrenamiento de X en cada dirección.** Más formalmente: los **inversos** de los valores singulares **no nulos** de la matriz de diseño X , es decir $1/\sigma_r$.
2. **Cuánto y en qué direcciones varían las covariables de test $\mathbf{x}_{\text{test}}$ en relación con el entrenamiento.** Más formalmente: cómo $\mathbf{x}_{\text{test}}$ se proyecta sobre los vectores singulares **derechos** V de X , es decir $\mathbf{x}_{\text{test}} \cdot v_r$.
3. **En qué medida el mejor modelo posible dentro de la clase puede alinear la variación en X con los objetivos Y .** Más formalmente: cómo los **residuos** E del mejor predictor lineal en la clase (lo que, desde la clase, son errores “insalvables”) se proyectan sobre los vectores singulares **izquierdos** U de X , es decir $u_r \cdot E$.

Usamos términos como “variar” y “varianza” para sugerir el vínculo con la noción estadística de varianza.¹ El **producto** de esas tres cantidades determina cuánto aporta el modo singular r al error de predicción.

Cuándo los tres factores se vuelven extremos

El **double descent** aparece cuando esos tres factores son **simultáneamente** desfavorables:

¹Si centráramos las columnas de X para obtener $\bar{X}$ y escribiéramos $\bar{X} = \bar{U}\bar{\Sigma}\bar{V}^\top$, entonces $\bar{X}^\top \bar{X}$ sería la **matriz de covarianza empírica** y su descomposición espectral $\bar{X}^\top \bar{X} = \bar{V}\bar{\Sigma}^2\bar{V}^\top$: cada v_r sería un eje ortogonal de variación y σ_r^2 la varianza empírica en esa dirección. **Sin centrar**, $X^\top X$ es el **momento de segundo orden** empírico, no la covarianza; por eso hablar de “varianza” es un ligero abuso terminológico, pero la intuición correcta es la misma: cuánto “se mueven” las entradas de X en cada dirección y si ese movimiento se **correlaciona** con Y a través del residuo. Centrar no es esencial para el fenómeno de double descent; omitir el centrado evita confundir al lector pedagógicamente.

- (i) En el entrenamiento hay varianza **pequeña pero no nula** en alguna dirección singular (equivalente: σ_r muy pequeño).
- (ii) Desde la perspectiva de la clase de modelos, el residuo E tiene **gran** proyección sobre ese modo: la relación entre **descriptores** de entrenamiento y respuestas es frágil en esa dirección.
- (iii) El vector de test **presenta variación importante** a lo largo de ese mismo modo: $\mathbf{x}_{\text{test}} \cdot \mathbf{v}_r$ es grande.

Si (i) y (ii) coinciden, los coeficientes efectivos a lo largo de ese modo suelen estar **mal estimados**. Si además (iii) entra por un $\mathbf{x}_{\text{test}}$ con gran proyección, el modelo se ve obligado a **extrapolar** fuerte en una dirección poco vista en entrenamiento, donde además la relación predicción–respuesta era ya propensa a error: el **error cuadrático en test** puede **explotar**.

Por qué la explosión ocurre cerca del umbral de interpolación

El factor $1/\sigma_r$ es el que se vuelve más **peligroso** al acercarse al umbral **desde cualquier régimen** de parametrización: los valores singulares casi nulos se vuelven más probables.

Una explicación **probabilística** profunda pasa por la ley de **Marchenko–Pastur** (matrices aleatorias): el menor valor singular no nulo tiende a situarse en un régimen crítico junto al borde de interpolación. Por ser técnica, aquí priorizamos la intuición geométrica: **cuánta dispersión hemos observado en cada dirección ortogonal** del espacio de covariables al ir añadiendo muestras (Figura 3).

Supongamos un **único** dato de entrenamiento $\mathbf{x}_1$. Mientras no sea el vector cero, ese punto define **una** dirección de variación no trivial: ganamos información sobre la dispersión de la distribución **en esa dirección**; en todas las ortogonales la varianza empírica es **exactamente cero**, y el ajuste lineal no puede usarlas. Con un **segundo** dato $\mathbf{x}_2$, de nuevo hay variación, pero parte de $\mathbf{x}_2$ suele proyectarse sobre $\mathbf{x}_1$; la dirección **compartida** acumula más evidencia, pero la segunda dirección ortogonal de variación queda **peor** identificada. Por eso, con dos muestras, el **menor** valor singular no nulo de la matriz de datos es en **probabilidad** más pequeño que con una sola.

A medida que crece N y nos acercamos al umbral $N = P = D$ (en el panel de Figura 3, $D = 3$), la probabilidad de que cada dato nuevo aporte varianza fuerte en una dirección **nueva** y ortogonal a todo lo observado **cae**. En el umbral, para que el N -ésimo dato **no** añada un valor singular pequeño-pero-no-nulo harían falta a la vez: (1) existir una dimensión en la que **ninguno** de los $N - 1$ datos anteriores hubiera variado, y (2) que el N -ésimo dato varíe **mucho** solo en esa dimensión: es **poco probable**. Más allá del umbral, con más datos, la varianza en cada dimensión covariable se estima mejor y el menor valor singular no nulo se **aleja** de cero. Eso conecta el pico de error con el peor condicionamiento espectral de X .

La Figura 4 muestra en datos reales que el **pico** de MSE de test y el **mínimo** del menor valor singular no nulo de X ocurren **juntos** en el umbral: evidencia empírica del vínculo entre $1/\sigma_r$ y el error.

Generalización en regresión lineal sobre-parametrizada

Puede sorprender que, en el régimen sobre-parametrizado, **tres** conjuntos de datos exhiban error cuadrático de test relativamente bajo (California Housing, Diabetes, Student–Teacher) mientras que **uno** (expectativa de vida de la OMS, *WHO Life Expectancy*) no. La clave es que, para $P > N$, el error de predicción incluye un **término de sesgo** que **no aparece** cuando $P < N$.

El término de sesgo y la “visibilidad” de las covariables

El objetivo del modelo lineal es **correlacionar** fluctuaciones en las covariables con fluctuaciones en la respuesta. Con **más parámetros que datos** ($P > N$), en cada punto de $\mathbb{R}^P$ el ajuste solo puede “ver” fluctuaciones en un subespacio de dimensión a lo sumo N : las $P - N$ direcciones restantes son **invisibles** para el entrenamiento. Por eso se pierde información sobre la relación lineal óptima β^* y crece el error de predicción en régimen sobre-parametrizado.

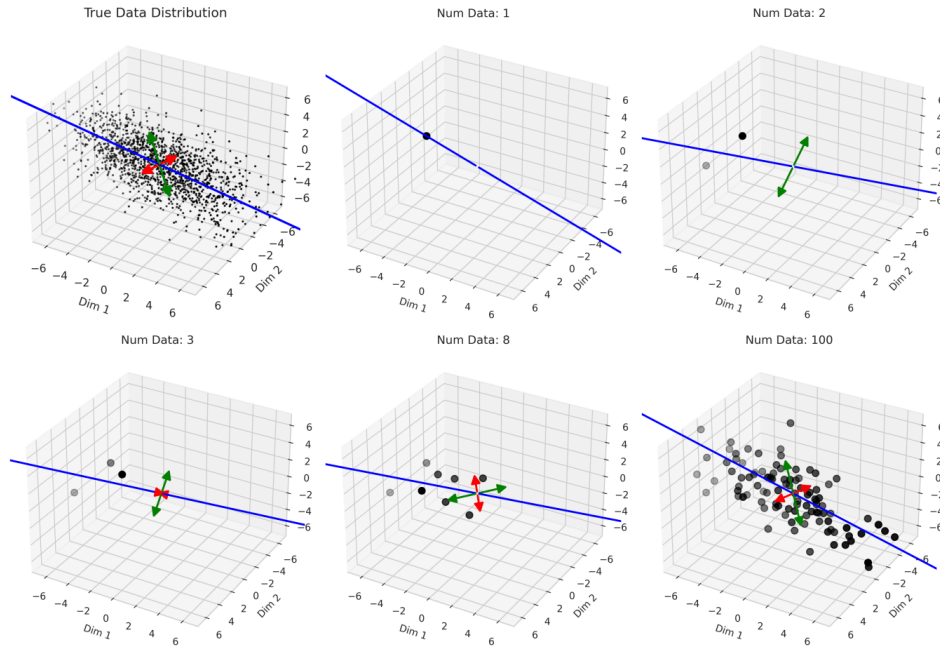


Figura 3: intuición geométrica ($D = 3$). Distribución verdadera y muestras con $N = 1, 2, 3, 8, 100$; ejes principales (azul, verde, rojo).

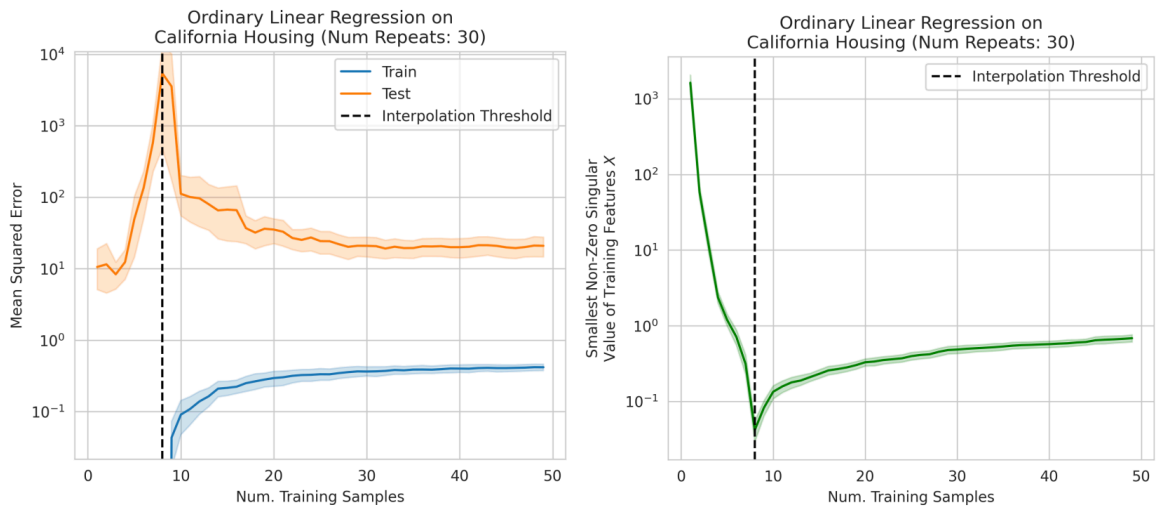


Figura 4: double descent en regresión lineal ordinaria (California Housing).

Sea $P = X^\top(XX^\top)^{-1}X$ la **proyección ortogonal** sobre el espacio fila de X (en $\mathbb{R}^P$, el span de las filas de la matriz de diseño). Para un covariable de test $\mathbf{x}_{\text{test}}$ definimos la **representación interna** inducida por el modelo lineal como la proyección

$$\hat{\mathbf{x}}_{\text{test}} := P \mathbf{x}_{\text{test}} = X^\top(XX^\top)^{-1}X \mathbf{x}_{\text{test}}.$$

La parte del error atribuible al sesgo respecto de β^* (la componente que ya aparecía al descomponer $\hat{y}_{\text{test,over}} - y_{\text{test}}^*$) puede escribirse como

$$\mathbf{x}_{\text{test}}^\top (X^\top(XX^\top)^{-1}X - I_P) \beta^* = (\hat{\mathbf{x}}_{\text{test}} - \mathbf{x}_{\text{test}})^\top \beta^*. \quad (2)$$

Es decir, el sesgo es el **producto interno** entre (1) el vector que va del dato de test a su proyección sobre el espacio “visible” por el entrenamiento, $\hat{\mathbf{x}}_{\text{test}} - \mathbf{x}_{\text{test}}$, y (2) los coeficientes ideales β^* de la clase. Si β^* es casi **ortogonal** a la componente “no vista”, el sesgo es pequeño; si β^* **aline**a con direcciones mal identificadas por X , el sesgo puede ser grande.

Lejos del umbral de interpolación, la **varianza** suele explicar poca parte de la discrepancia entre el modelo sobre-parametrizado y el ideal; buena parte de la discrepancia proviene entonces del **sesgo** ((2)). Esa misma expresión admite una lectura geométrica sorprendente: la regresión lineal sobre-parametrizada hace **aprendizaje de representaciones** —reemplaza $\mathbf{x}_{\text{test}}$ por su proyección $\hat{\mathbf{x}}_{\text{test}}$ en el subespacio aprendido con los datos de entrenamiento.

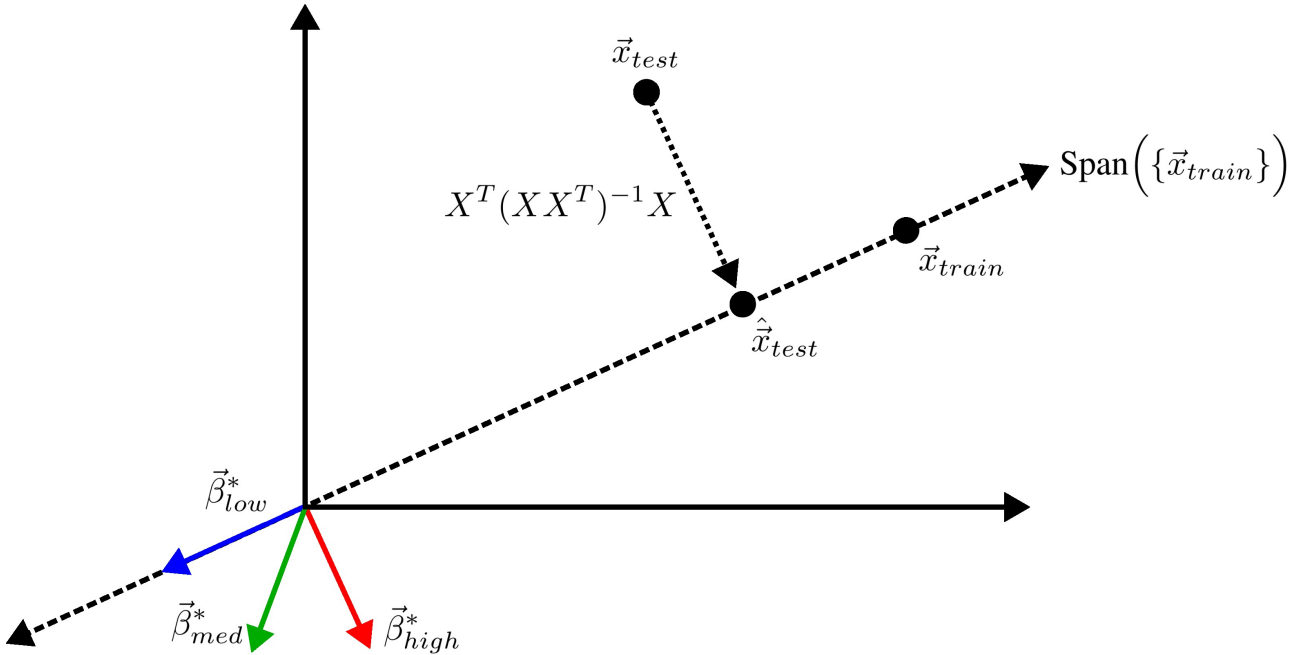


Figura 5: Geometría de la generalización en OLR sobre-parametrizada.

Sesgo al cuadrado en test en cuatro conjuntos de datos

Intuitivamente, un modelo sobre-parametrizado **generaliza bien** cuando sus **representaciones** (las proyecciones $\hat{\mathbf{x}}$) retienen la información que β^* necesita para predecir (Figura 5, Figura 6). Donde β^* tiene componente fuerte en direcciones que el entrenamiento no puede ver, el término (2) —y por tanto el error de test— se dispara aunque la varianza del estimador sea manejable.

Hacia las redes: *lazy training*, NTK y el prisma de la interpolación

Las cuentas anteriores son **lineales** a propósito: matriz de diseño $\mathbf{X}_P$ (Legendre) y solución de norma mínima. Para trasladar el mismo lenguaje a redes no lineales se usa un régimen en que w permanece en

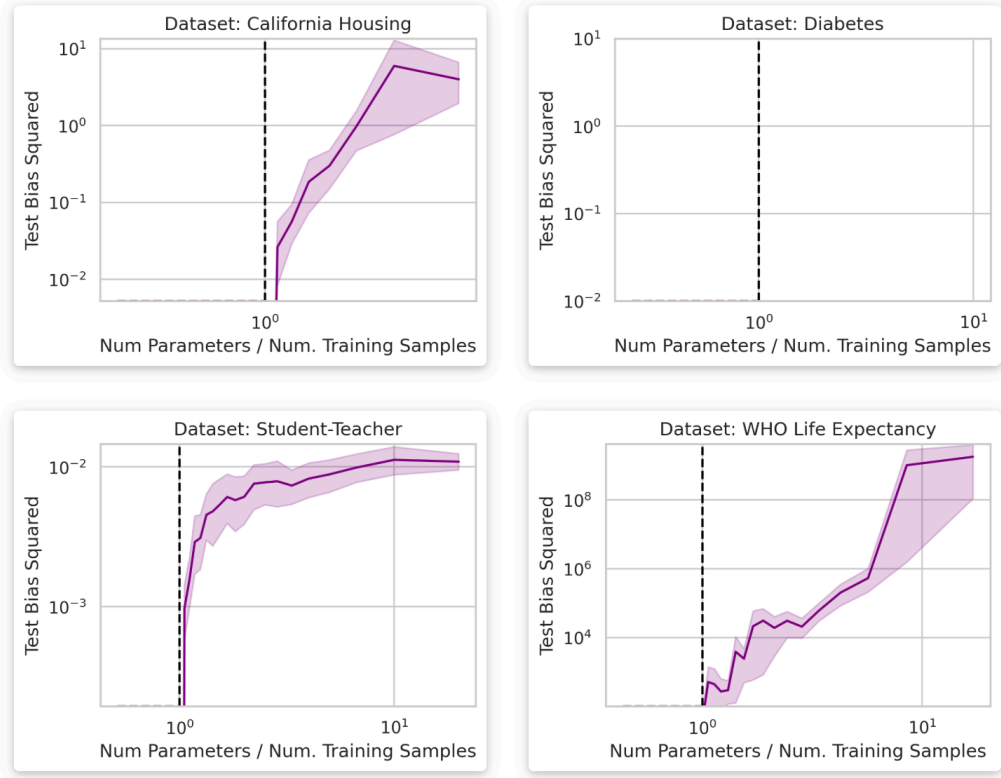


Figura 6: Sesgo al cuadrado en test para modelos sobre-parametrizados (cuatro conjuntos).

un entorno de la inicialización w_0 , de modo que $f(w, x)$ admite una aproximación de **orden uno** en w ; eso es el marco de **lazy training** / linealización. La exposición enlaza con la visión de interpolación y sobre-parametrización de Belkin (2021); las demostraciones que siguen son el caso **linealizado finito** (riguroso) y el puente al **kernel tangente** (límite de ancho y flujo de gradiente).

Linealización de Taylor y *feature matrix*

Sea $f : \mathbb{R}^d \times \mathcal{X} \rightarrow \mathbb{R}$ de clase $\mathcal{C}^1$ en w . Fijamos w_0 y escribimos el desarrollo de Taylor con resto integral de orden uno:

$$f(w, x) = f(w_0, x) + \nabla_w f(w_0, x)^\top (w - w_0) + R(w, w_0, x),$$

con $\|R(w, w_0, x)\| = o(\|w - w_0\|)$ cuando $w \rightarrow w_0$. El modelo **linealizado** es

$$f_{\text{lin}}(w, x) := f(w_0, x) + \phi(x)^\top (w - w_0), \quad \phi(x) := \nabla_w f(w_0, x) \in \mathbb{R}^d.$$

Definimos la **matriz Jacobiana de entrenamiento**

$J \in \mathbb{R}^{n \times d}$ por filas $J_i = \phi(x_i)^\top$, y los vectores $f_0 := (f(w_0, x_1), \dots, f(w_0, x_n))^\top$, $y := (y_1, \dots, y_n)^\top$. Entonces

$$f_{\text{lin}}(w) := (f_{\text{lin}}(w, x_1), \dots, f_{\text{lin}}(w, x_n))^\top = f_0 + J(w - w_0).$$

Con $u := w - w_0$ y $r := y - f_0$ (vector de **residuos al punto linealizado**), interpolar los datos en el modelo linealizado es

$$Ju = r. \quad (\star)$$

La matriz J es el análogo matricial de apilar features: en Legendre, las filas eran $\tilde{\phi}_P(x_i)^\top$; aquí las filas son gradientes $\nabla_w f(w_0, x_i)^\top$. El régimen *lazy* es aquel en que $\|w - w_0\|$ es tan pequeño (o el ancho tan grande) que $f \approx f_{\text{lin}}$ a lo largo de la optimización; entonces el análisis de $(\star)$ es el sustituto riguroso del modelo polinómico finito.

Neural Tangent Kernel: definición y matriz de Gram

Definición. El **kernel tangente neuronal** (NTK) en w_0 es la función positiva semidefinida

$$\Theta(x, x') := \langle \nabla_w f(w_0, x), \nabla_w f(w_0, x') \rangle = \phi(x)^\top \phi(x').$$

Sobre la muestra $\{x_i\}_{i=1}^n$, la **matriz** $\Theta \in \mathbb{R}^{n \times n}$ dada por $\Theta_{ij} = \Theta(x_i, x_j)$ es la **matriz de Gram** de los vectores $\phi(x_i)$:

$$\Theta = JJ^\top.$$

Observación (flujo de gradiente en el límite ancho). Jacot et al. (2018) prueban que, bajo inicialización aleatoria adecuada y ancho $\rightarrow \infty$, Θ converge a un límite determinista y, en el entrenamiento continuo de GD sobre pérdida cuadrática, la evolución de las predicciones en los datos queda gobernada por un kernel que coincide con ese límite (**NTK fijo**). En la práctica, “NTK estático” significa que Θ no varía apreciablemente durante el entrenamiento, de modo que la dinámica es la de un modelo lineal en el espacio de funciones inducido por Θ . El condicionamiento de Θ (valores propios pequeños) es el paralelo espectral al mal condicionamiento de J o de $\mathbf{X}_P$ en el ejemplo de Legendre.

Proposición (GD lineal y norma mínima)

Trabajamos con el modelo linealizado f_{lin} y la pérdida cuadrática

$$L(w) = \frac{1}{2} \|f_{\text{lin}}(w) - y\|_2^2 = \frac{1}{2} \|Ju - r\|_2^2, \quad u = w - w_0.$$

Lema 1 (gradiente y subespacio). Se cumple $\nabla_u L = J^\top(Ju - r)$. Por tanto, cada paso de GD,

$$u_{t+1} = u_t - \eta J^\top(Ju_t - r),$$

satisface $u_{t+1} - u_t \in \text{Im}(J^\top) = \text{span}\{\phi(x_1), \dots, \phi(x_n)\}$. Si $u_0 = 0$, entonces $u_t \in \text{Im}(J^\top)$ para todo t .

Demostración. La primera afirmación es la regla de la cadena; la segunda es definición de GD; la tercera, inducción sobre t . $\square$

Lema 2 (complemento ortogonal fijo). Sea $P_{\text{row}} = J^+J$ la proyección ortogonal sobre $\text{Im}(J^\top)$ y $P_{\text{null}} = I - P_{\text{row}}$ sobre $\ker(J)$. Si $u_0 = 0$, entonces $P_{\text{null}}u_t = 0$ para todo t (la componente en $\ker(J)$ permanece nula).

Demostración. $J^\top v \in \text{Im}(J^\top)$ para todo v , luego $P_{\text{null}}(u_{t+1} - u_t) = 0$. Con $u_0 = 0$ se obtiene $P_{\text{null}}u_t = 0$. $\square$

Proposición. Supongamos $\text{rango}(J) = n$ (interpolación posible y J de rango fila completo). Entonces existe una única solución u^* de **norma euclídea mínima** tal que $Ju^* = r$, dada por $u^* = J^+r = J^\top(JJ^\top)^{-1}r$. Si el GD con paso η suficientemente pequeño converge desde $u_0 = 0$, su límite es u^* .

Esquema de demostración. El conjunto $\{u : Ju = r\}$ es un subespacio afín paralelo a $\ker(J)$. Entre ellos, el de norma mínima es el único ortogonal a $\ker(J)$, es decir, en $\text{Im}(J^\top)$; la fórmula de J^+ es estándar. El flujo de GD en problemas cuadráticos convexos converge al único mínimo global u^* en

$\text{Im}(J^\top)$ cuando $u_0 = 0$ y η está por debajo del inverso de la constante de Lipschitz de ∇L (típicamente $\eta < 2/\lambda_{\text{máx}}(J^\top J)$). Ese mínimo global coincide con el interpolador de norma mínima por Lema 2. $\square$

Corolario (RKHS). Si identificamos el espacio de funciones lineales en las features $\phi(x)$ con el RKHS asociado al kernel Θ , la solución de norma mínima en parámetros u corresponde a la función de **norma de RKHS mínima** que interpola en los puntos de entrenamiento (en el subespacio generado por $\Theta(\cdot, x_i)$), en analogía directa con el caso Legendre y la pseudo-inversa.

Condición de Polyak–Łojasiewicz y variantes

Definición (PL). Una función L diferenciable satisface la **condición PL** con constante $\mu > 0$ en un conjunto Ω si

$$\|\nabla L(w)\|_2^2 \geq 2\mu (L(w) - L^*) \quad \forall w \in \Omega,$$

donde L^* es el mínimo global de L en Ω (a veces se escribe la forma equivalente $\|\nabla L\|^2 \geq \mu L$ cuando $L^* = 0$).

Proposición (convergencia lineal bajo PL). Si L es β -suave (gradiente Lipschitz) y satisface PL con constante μ , el GD con $\eta \leq 1/\beta$ verifica

$$L(w_t) - L^* \leq (1 - \eta\mu)^t (L(w_0) - L^*).$$

Demostración (esquema). Descenso estándar usando PL y suavidad; ver Polyak (1963) o Karimi et al. para la constante exacta. $\square$

En redes, L **no** es convexa en w ; la variante **PL*** o hipótesis de tipo PL en el régimen interpolante (Belkin (2021), §4) postula que, con sobre-parametrización suficiente, en una región que contiene trayectorias razonables de GD se tiene control del tipo PL **hacia el conjunto de mínimos globales** (error nulo en entrenamiento). La consecuencia conceptual: no hace falta convexidad global para explicar convergencia a interpolación; basta una geometría que impida estancarse en valles con pérdida estrictamente positiva cuando existen mínimos con $L = 0$.

Prisma de la interpolación y *benign overfitting*

Formalmente, la **navaja de Occam** clásica penaliza la **cardinalidad** de parámetros o la dimensión del modelo. En el régimen interpolante moderno, la complejidad se mide en otra métrica: **norma en el RKHS** (o norma de los parámetros en la capa lineal final bajo NTK) en lugar del recuento de parámetros.

Belkin (2021) enfatiza que, con muchos grados de libertad, un predictor puede interpolar datos ruidosos descomponiendo la solución en una parte “suave” (baja norma en el kernel efectivo) más correcciones de alta frecuencia que encajan el ruido pero tienen norma pequeña en la métrica adecuada; la **señal** puede preservarse en la componente suave. Por eso el error de test puede **descender** al aumentar la capacidad: hay más maneras de ser simultáneamente interpolante y de norma controlada en el espacio de funciones relevante —fenómeno distinto del sesgo-varianza con un solo mínimo.

Síntesis. Del mismo modo que en Legendre la pseudo-inversa selecciona, entre los w con $Jw = y$, el de norma $\|w - w_0\|$ mínima (con w_0 el origen del modelo linealizado), el **NTK** describe, en el límite de red ancha, un espacio de Hilbert de funciones donde el descenso de gradiente desde $w_0 \approx 0$ aproxima la interpolación de **norma mínima** en ese espacio: el puente entre el álgebra finita de $\mathbf{X}_P$ y el comportamiento cualitativo del aprendizaje profundo.

Resumen

- El **double descent** aparece aquí en **regresión polinómica 1D** con base de **Legendre**, ruido gaussiano $\mathcal{N}(0, 0.2^2)$, nodos de Chebyshev, reparto train/test aleatorio y solución de **norma mínima** (pseudo-inversa de Moore–Penrose en el régimen interpolante).
- $P = n$ es el umbral donde el modelo lineal puede interpolar el entrenamiento; cerca de ese borde, valores singulares pequeños de $\mathbf{X}_P$ **amplifican** componentes del ruido y del residual en la forma $\sum_r (\mathbf{x}_{\text{test}} \cdot v_r)(u_r \cdot E)/\sigma_r$.
- Para $P > N$, además entra un término de **sesgo** respecto de β^* asociado a la proyección $X^\top (XX^\top)^{-1} X$: solo una parte de β^* es identificable desde N ecuaciones en P incógnitas.
- El pico y la recuperación del error de test se entienden con mayor nitidez en el lenguaje de la **SVD** y la pseudo-inversa que solo con una curva genérica “sesgo vs. varianza”.

Belkin, Mikhail. 2021. «Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation». <https://arxiv.org/abs/2105.14368>.

Jacot, Arthur, Franck Gabriel, y Clément Hongler. 2018. «Neural tangent kernel: Convergence and generalization in neural networks». En *Advances in Neural Information Processing Systems*. Vol. 31.