TEMA 5: SOBREAJUSTE Y REGULARIZACIÓN.

SOBREAJUSTE

+ Nos sobreajustamos a una muestra ↔ ↓↓ E_{in} → ↑↑ E_{out}.

* ¿Por qué sucede?

→ Según dimensión-VC: $E_{out} \leq E_{in} + \Omega(N, \mathcal{H}, \delta)$ (¿Tenemos suficientes ejemplos?; ¿Hemos elegido $\mathcal{H}$ muy compleja?)

→ Según sesgo-varianza:

$$E_{\mathcal{D}}[E_{out}(g^{(\mathcal{D})})] = \sigma^2 + \text{bias} + \text{varianza}$$

* σ^2 = "Error estocástico": error presente en datos entrenamiento.

* Bias = "Error determinista": error que comete la mejor función de nuestro $\mathcal{H}$ "g" estimando f (objetivo).

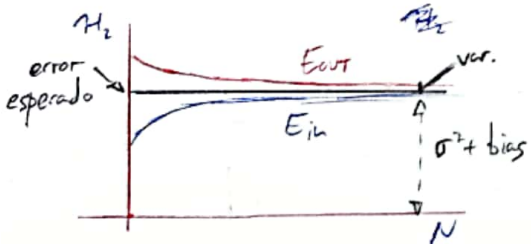
↳ Para mejorarlo → Aumentar $\mathcal{H}$ → ↑↑ varianza (intercambio).

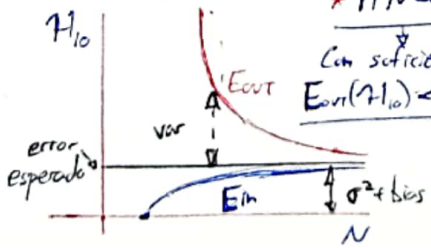
* Si tenemos una $\mathcal{H}$ y $\mathcal{D}$ (muestra) fijas → ruido $\begin{cases} \text{determinista} \\ \text{estocástico} \end{cases}$ → indistinguibles.

* Por ejemplo, consideremos 2 casos; con $N=15$:

1. f = Polinomio grado 10 + ruido estocástico (σ^2).
   → Si intentamos ajustar polinomios:
   * Con grado 2 → ↑ E_{in} (debido a bias), pero ↓ E_{out} (↓varianza).
   * Con grado 10 → ↓ E_{in} (≈ 0), pero ↑↑↑ E_{out} (↑↑ sobreajuste, debido al error estocástico, "aprendimos el error").
2. f = Polinomio de grado 50. (sin ruido). (↓↓ var)
   * Con grado 2 → ↑ E_{in} (bias), ↓ E_{out} (varianza baja).
   * Con grado 10 → ↓↓ E_{in}, pero ↑↑↑ E_{out} (↑↑ sobreajuste, ya que ↑↑ variabilidad → hay error determinista).

* En ambos casos, estas son las curvas de aprendizaje:





* ↑↑ N → ↓↓ var

Con suficiente N, $E_{out}(\mathcal{H}_{10}) < E_{out}(\mathcal{H}_2)$

PROTECCIONES CONTRA SOBREAJUSTE

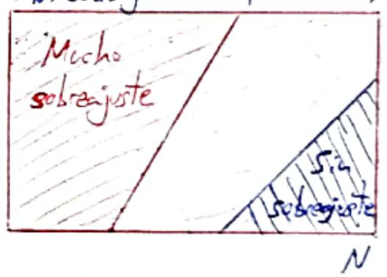
+ ¿Cómo decidimos la complejidad de $\mathcal{H}$?
  * Hay ruido en la muestra.
  * La selección de $h \in \mathcal{H}$ se realiza con ERM o SRM.
+ Solución 1: sesgo inductivo.
  * Limitamos $\mathcal{H}$ para que no aprenda el ruido.
  * Problema: Restringimos también la capacidad de $\mathcal{H}$ para estimar f.
+ Solución 2: regularización.
  * Añadimos a la función de error otra componente ≈ complejidad.
  * Así → Compromiso entre ajuste y complejidad / variabilidad.

INFLUENCIA DEL RUIDO EN SOBREAJUSTE

+ Ruido estocástico (σ^2)

$\mathcal{H}_{10}$: Sobreajuste (comparando $\mathcal{H}_{10}$ y $\mathcal{H}_2$).

* ↑↑ σ^2 → ↑↑ sobreajuste (aprendemos ruido).

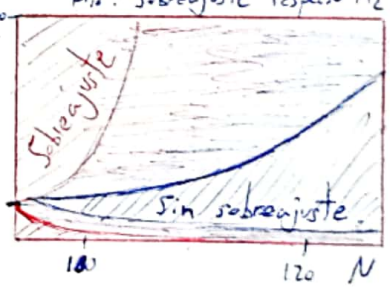


* ↑↑ N → ↓ sobreajuste (si hay suficientes ej., entonces el ruido se contrarresta, pues tiene media = 0 → aprendemos).

+ Ruido determinista. (Q_f = complejidad de f = grado de f).

$\mathcal{H}_{10}$: Sobreajuste respecto $\mathcal{H}_2$

* ↑↑ Q_f → ↑ sobreajuste.



* ↑↑ N → ↓ sobreajuste.

* Problema: $\mathcal{H}_{10}$, cuando $Q_f > 10$, no es capaz de ajustarse a f. Además, como N es finito → $\mathcal{H}_{10}$ trata de aprender la muestra (aunque no sea capaz de adaptarse al f real) → Sobreajuste.
↳ (Con ruido, tanto estocástico como determinista). (Aprende patrón inexistente)

REGULARIZACIÓN

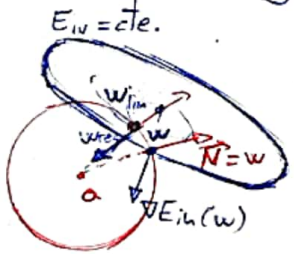
→ Para ello → Heurística.

+ Consiste en restringir $\mathcal{H}$ → ↓↓ E_{out}.
+ Tiene cierta justificación teórica:
  * VC → ↓↓ d_{vc}, por lo que hay mejor generalización.
  * Sesgo-Varianza → Aumenta un poco el sesgo (limitamos $\mathcal{H}$, por lo que empeora la mejor solución probablemente), pero reduce mucho la varianza.
+ También tiene una intuición práctica: evitar la influencia de ruidos:
  * El ruido determinista suele ser irregular.
  * El ruido estocástico tiene alta frecuencia (muchos datos están afectados).

→ Por tanto, limitamos $\mathcal{H}$ para que sea simple / regular.

WEIGHT DECAY - DECAIMIENTO DE PESOS

+ Consiste en medir la complejidad de una hipótesis h por el tamaño de sus pesos.
+ Por tanto → Restringimos $\mathcal{H}$: $\sum_{i=0}^{h} w_i^2 \leq C$ → cte. (↓↓ d_{vc}).
  * Nuevo problema: $\min_w E_{in}(w)$, tal que $w^T w \leq C$. --→ w_{reg}.
  * ¿Cómo calcularlo? (Aprox. gráfica)



→ Círculo rojo = espacio de w's que satisface $w^T w \leq C$.
→ Elipse = puntos en el espacio con el mismo E_{in} que w.
→ w_{lin} es la solución óptima (sin restricciones).

Caso 1: si w_{lin} cumple la restricción → $w_{reg} = w_{lin}$.

Caso 2: si w_{lin} no cumple la restricción:

→ $w_{reg} \in$ Círculo rojo // $w_{reg}^T w_{reg} = C$. (para acercarnos lo máx. posible a w_{lin}).

→ Tiene que cumplirse que $-\nabla E_{in}(w_{reg})$, que es la dirección hacia donde se minimiza el error, coincida con la normal al círculo ($C = w$, ya que centro = 0). Si: no → $-\nabla E_{in}(w)$ tiene componente en dir. tangente → ↓↓ E_{in} al mov. en $\mathbf{0}$.

ERROR AUMENTADO

+ Dado w, definimos su error aumentado:

$$E_{aug}(w, \lambda, \Omega) = E_{in}(w) + \frac{\lambda}{N}\Omega(w)$$

↳ Función de reg. usada (Ej. weight decay: $\Omega(w) = w^T w$).
↳ Fuerza de la restricción de la regularización (↑↑ λ → ↓↓ d_{vc}).

+ Esta surge de que:
  * $\min_w E_{in}(w)$, tq $\underbrace{w^T w \leq C}_{\text{weight decay}}$ $\overset{\text{equivale a}}{\longleftrightarrow}$ $\min_w E_{in}(w) + \lambda_c w^T w$, $\lambda_c > 0$ →

usando multiplicadores de Lagrange → $\min_w \{E_{in}(w) + \lambda(w^T w - C)\}$ → $E_{aug} = E_{in}(w) + \lambda w^T w$ (sin restricción).

* Por tanto, minimizar E_{aug} resuelve el problema con regularización.

+ E_{aug} puede interpretarse como una aproximación a E_{out} más cercana que E_{in}:
  * $E_{out} \equiv E_{in} + \underbrace{\Omega(\mathcal{H})}$; $E_{AUG} = E_{in} + \underbrace{\Omega(h)}_{\text{Estima generalización del error}} \cdot \frac{\lambda}{N}$

REGRESIÓN REGULARIZADA

+ Objetivo: minimizar $E_{in}(w)$ sobre la hipótesis $\mathcal{H}$ restringida.
  * Nota: emplearemos $\Omega(w) = w^T w$.
+ $w_{reg} = \min_w E_{AUG}(w) = \min_w \underbrace{\|Zw - y\|^2}_{E_{in}(w)} + \frac{\lambda}{N}\underbrace{\|w\|^2}_{\Omega}$.
+ $\nabla_w E_{aug}(w) = \nabla_w (E_{in}(w) + \lambda w^T w) = 0$ → $w = w_{reg}$ (mínimo).

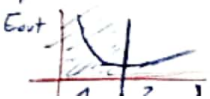
→ $= \underbrace{2Z^T(Zw - y)}_{\nabla E_{in}(w) = -\frac{2}{N}\cdot} + \lambda w^T = 0$ (¡coincide con intuición geométrica!)

→ $w_{reg} = (Z^T Z + \lambda I)^{-1} Z^T y$

↳ Si $\lambda = 0$ → $w_{reg} = w_{lin}$ } ↑↑ λ → ↓↓ C.
↳ Si $\lambda \to \infty$ → $w_{reg} \to 0$.

* ¿Cómo calculamos λ?
→ Empíricamente (ensayo y error). } - Si es muy bajo → Sobreajuste. (Ver)
- Si es muy alto → ¡No se ajusta! (Bias)

↳ Validación.

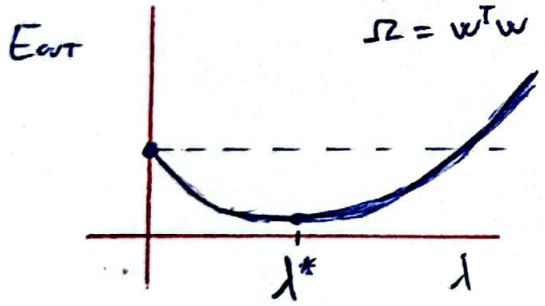


ELECCIÓN DE REGULARIZADOR

+ Regulador óptimo → Aquel que nos lleve hacia f_0 (desconocida).
  * Por tanto, nos podemos guiar por ir a hipótesis "suaves" y "simples", ya que evitan el ruido (que no es suave normalm.).
+ Ejemplos de Ω:
  - Weight decay
  - $\sum_{q=0}^{Q} |w_q| \leq C$ → Útil para hacer pesos = 0.
  - $\sum_{q=0}^{Q} \gamma_q w_q^2 \leq C$ → Da más importancia a ciertos q.
+ Si elegimos Ω inadecuado → En validación, $\lambda = 0$ (sin reg.).

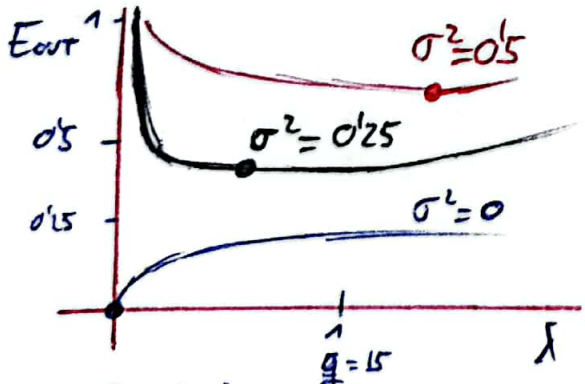
INFLUENCIA DE λ - RELACIÓN λ - RUIDO.

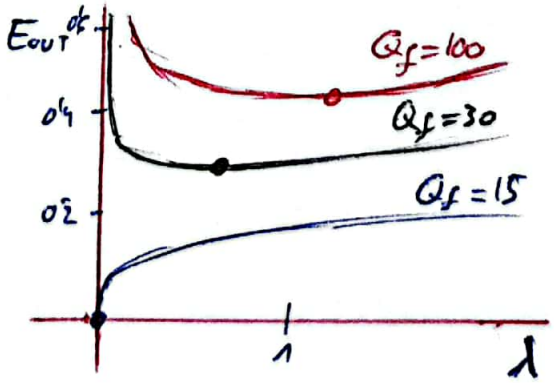
+ Como decíamos, debemos ajustar λ en cada problema:



  * Para $\lambda \approx 0$ → Sobreajuste
  * Para $\lambda \uparrow\uparrow$ → Infraajuste
  * Hay que buscar λ^* de forma experimental.

+ La cantidad de ruido influye en la elección de λ:





  * $\Omega(w) = \sum_{q=0} w_q^2$
  * Tanto si es error estocástico o determinista, cuanto más error haya → Mayor λ necesitaremos para combatirlo.
  * Si no hubiera error de ningún tipo → No necesitamos regularización.