Tema 2 – Optimización y Regularización (Parte 4)

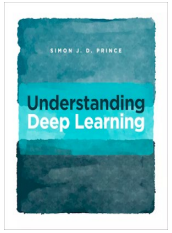
Aprendizaje Automático II - Grado en Inteligencia Artificial
Universidad Rey Juan Carlos

Iván Ramírez Díaz
ivan.ramirez@urjc.es

José Miguel Buenaposada Biencinto
josemiguel.buenaposada@urjc.es

Bibliografía

- Understanding Deep Learning. Capítulo 9.



Regularización

- ¿Por qué aparece un error de generalización entre datos de entrenamiento y test?
  - Sobreajuste: el modelo describe peculiaridades estadísticas.
  - Subauste: El modelo no tiene restricciones en áreas sin datos de entrenamiento.
- Regularización = métodos para reducir el error de generalización
- Técnicamente: añadir términos a la función de coste
- Coloquialmente: cualquier método para reducir ese error.

2.6 Regularización

- Regularización explícita
- Regularización implícita
- Parar antes (early stopping)
- Aumento de datos (data augmentation)

2.6 Regularización

- Regularización explícita
- Regularización implícita
- Parar antes (early stopping)
- Aumento de datos (data augmentation)

Regularización explícita

- Función de coste estándar:

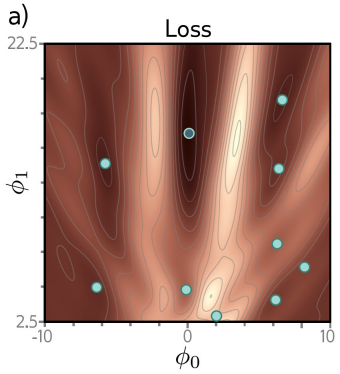
$$\hat{\boldsymbol{\phi}} = argmin_{\boldsymbol{\phi}} \; J(\boldsymbol{\phi})$$
$$= argmin_{\boldsymbol{\phi}} \left[\sum_{i=1}^{N} L(f(\mathbf{x}_i, \boldsymbol{\phi}), \mathbf{y}_i) \right]$$

- La regularización añade un término extra

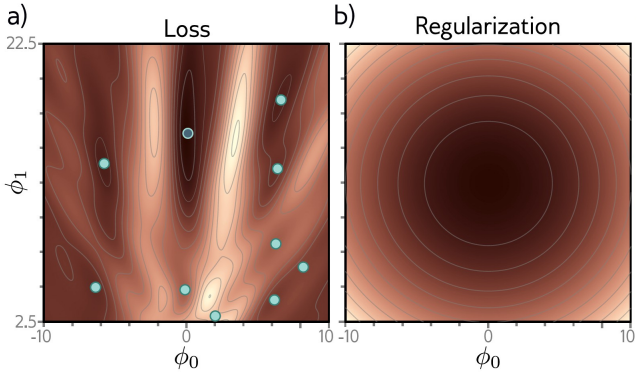
$$\hat{\boldsymbol{\phi}} = argmin_{\boldsymbol{\phi}} \left[\sum_{i=1}^{N} L(f(\mathbf{x}_i, \boldsymbol{\phi}), \mathbf{y}_i) + \lambda \cdot g[\boldsymbol{\phi}] \right]$$

  - g[$\boldsymbol{\Phi}$] Prefiere/desalienta algunos valores en los parámetros
  - λ>0 controla la fuerza de la regularización

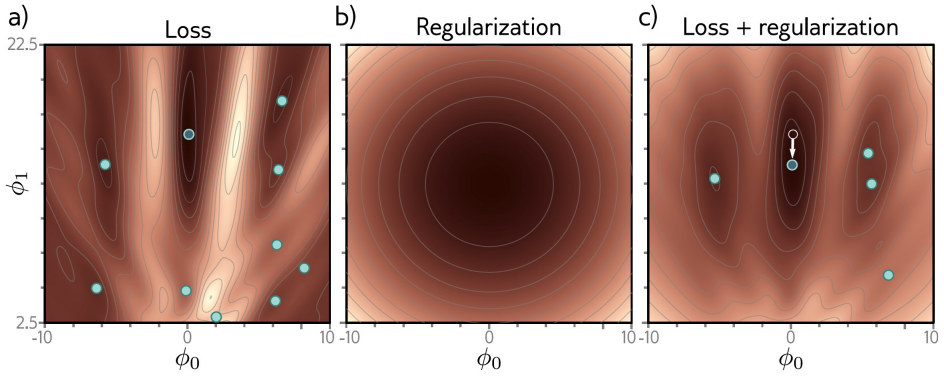
Regularización explícita



Regularización explícita



Regularización explícita



Interpretación probabilística

- Máxima verosimilitud:

$$\hat{\boldsymbol{\phi}} = argmáx_{\boldsymbol{\phi}} \left[\prod_{i=1}^{N} Pr(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\phi}) \right]$$

- La regularización es un *a priori* sobre los parámetros

$$\hat{\boldsymbol{\phi}} = argmáx_{\boldsymbol{\phi}} \left[\prod_{i=1}^{N} Pr(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\phi}) \cdot Pr(\boldsymbol{\phi}) \right]$$

… qué sabes acerca de los parámetros antes de ver los datos

Interpretación probabilística

- Máxima verosimilitud:

$$\hat{\boldsymbol{\phi}} = argmáx_{\boldsymbol{\phi}} \left[\prod_{i=1}^{N} Pr(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\phi}) \right]$$

- La regularización es un *a priori* sobre los parámetros minimizando el -log:

$$\hat{\boldsymbol{\phi}} = argmín_{\boldsymbol{\phi}} -\log\left[\prod_{i=1}^{N} Pr(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\phi}) \cdot Pr(\boldsymbol{\phi}) \right]$$

Interpretación probabilística

- Máxima verosimilitud:

$$\hat{\boldsymbol{\phi}} = argmáx_{\boldsymbol{\phi}} \left[\prod_{i=1}^{N} Pr(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\phi}) \right]$$

- La regularización es un *a priori* sobre los parámetros minimizando el -log:

$$\hat{\boldsymbol{\phi}} = argmín_{\boldsymbol{\phi}} \left[\left(-\sum_{i=1}^{N} \log Pr(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\phi}) \right) - \log Pr(\boldsymbol{\phi}) \right]$$

Equivalencia de la regularización y a priori

- Regularización explícita:

$$\hat{\boldsymbol{\phi}} = argmin_{\boldsymbol{\phi}} \left[\sum_{i=1}^{N} L(f(\mathbf{x}_i, \boldsymbol{\phi}), \mathbf{y}_i) + \lambda \cdot g[\boldsymbol{\phi}] \right]$$

- La regularización es un *a priori* sobre los parámetros minimizando el -log:

$$\hat{\boldsymbol{\phi}} = argmín_{\boldsymbol{\phi}} \left[\left(-\sum_{i=1}^{N} \log Pr(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\phi}) \right) - \log Pr(\boldsymbol{\phi}) \right]$$

- Equivalencia:

$$\lambda \cdot g[\boldsymbol{\phi}] = -\log[Pr(\boldsymbol{\phi})]$$

Regularización L2 (repaso)

- La más común en Deep Learning es la L2
- Favorece obtener parámetros más pequeños

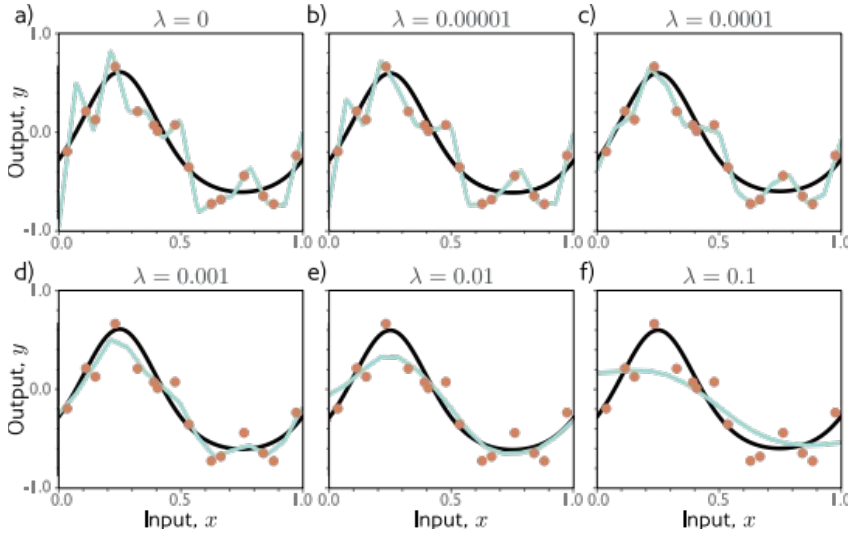
$$\hat{\boldsymbol{\phi}} = argmín_{\boldsymbol{\phi}} \left[J(\boldsymbol{\phi}, \{\mathbf{x}_i, \mathbf{y}_i\}) + \lambda \cdot \sum_j \phi_j^2 \right]$$

- También llamada regularización de Tichonov, ridge regression
- En redes de neuronas, usualmente únicamente aplicada a los pesos (no al sesgo del modelo lineal) y se le llama weight decay.

Por qué ayuda la regularización L2

- Ayuda a que los parámetros no se ajusten completamente a los datos (sobreajuste).
- Ayuda a tener suavidad fuera de los datos de entrenamiento.

Regularización L2



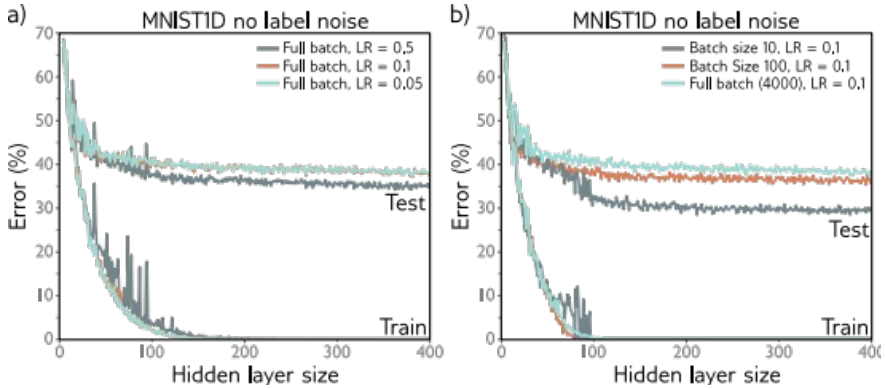
2.6 Regularización

- Regularización explícita
- Regularización implícita
- Parar antes (early stopping)
- Aumento de datos (data augmentation)

Regularización implícita

- El descenso de gradiente evita áreas donde los gradientes son muy grandes.
- SGD prefiere que todos los mini-batches tengan gradientes parecidos.
- Depende del learning rate – quizá esa es la razón por la que tener un learning rate más grande generaliza mejor.

Regularización implícita



- Rendimiento mejor con:
  - learning rates más grandes,
  - mini-batch más pequeños.

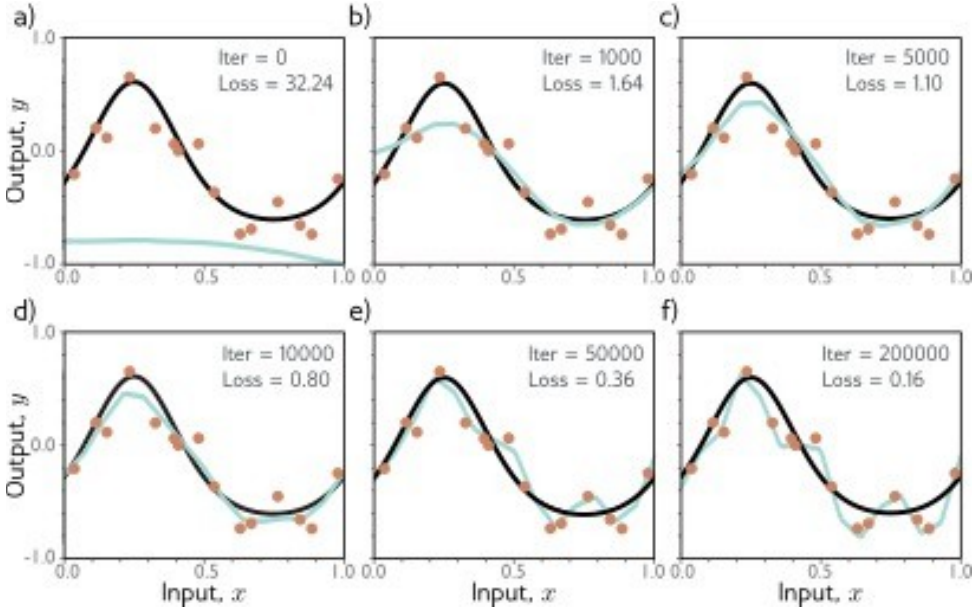
2.6 Regularización

- Regularización explícita
- Regularización implícita
- Parar antes (early stopping)
- Aumento de datos (data augmentation)

Early stopping

- Si paramos pronto, los pesos no “tienen tiempo” de sobreajustar al ruido
- Los pesos comienzan pequeños, no tienen tiempo de hacerse grandes
- Reduce la complejidad efectiva del modelo
- Efecto parecido a la regularización L2
- No necesita de re-entrenamiento.

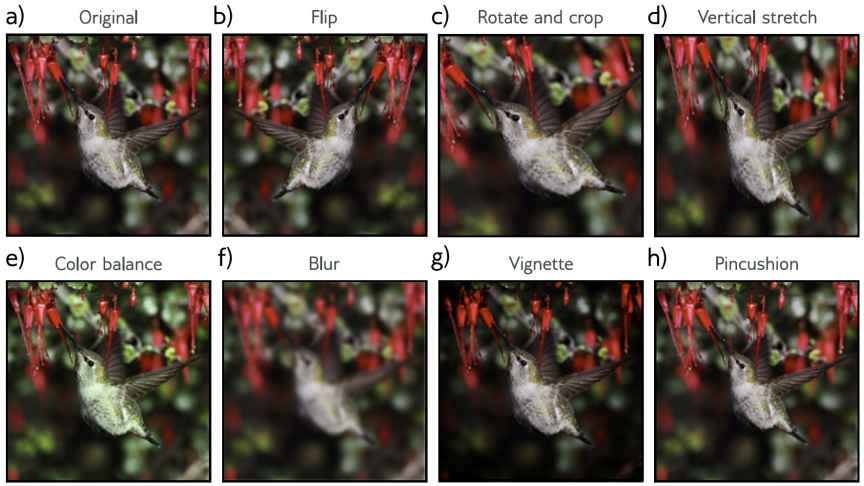
Early stopping



2.6 Regularización

- Regularización explícita
- Regularización implícita
- Parar antes (early stopping)
- Aumento de datos (data augmentation)

Aumento de Datos (data augmentation)



Métodos de regularización