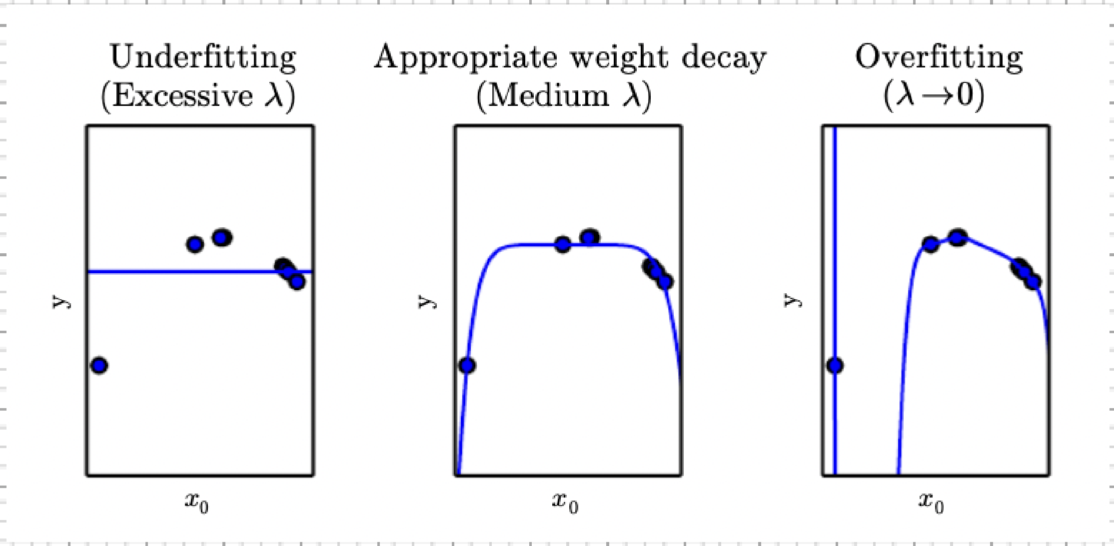
Regularización L2 (Penalización de parámetros)

Limitar la capacidad o flexibilidad del modelo.

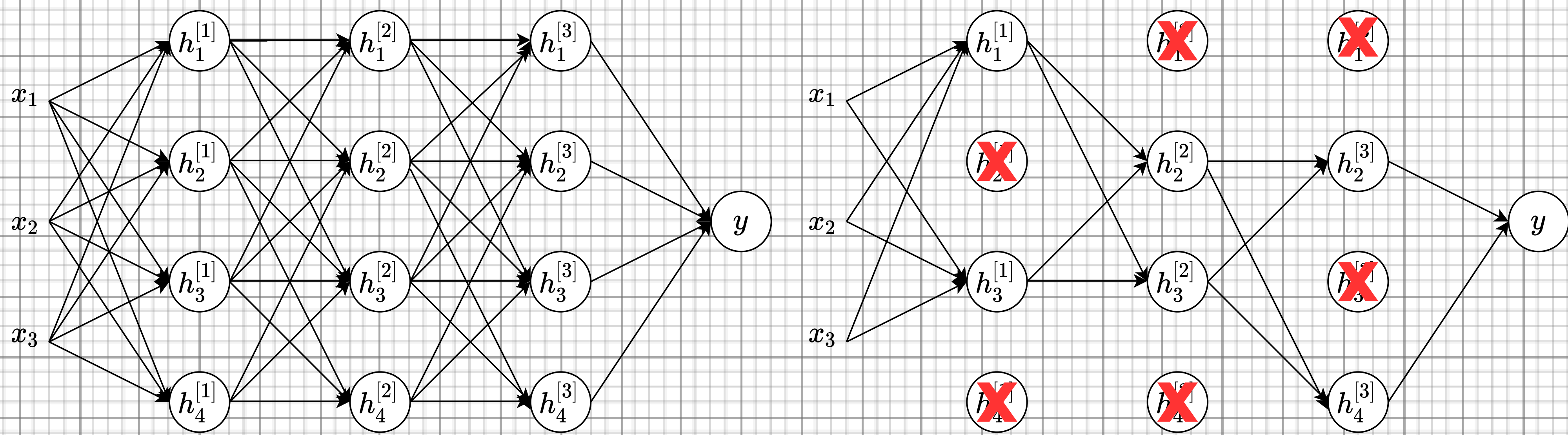
$$\hat{J}(\theta; X, Y) = J(\theta; X, Y) + \lambda\Omega(\theta) \qquad \theta = \{\mathbf{W}, \mathbf{b}\} \qquad \Omega(\theta) = \frac{1}{2}\|\mathbf{W}\|_2^2$$

$$\mathbf{W}_{t+1} = (1 - \eta\lambda)W_t - \eta\nabla_{\mathbf{W}} J(\mathbf{W}_t; X, Y)$$



Dropout

Este método un implica añadir ruido mientras se calcula $z^{[\ell]}$ de cada capa intermedia $\ell \in \{1, 2, \cdots, L-2, L-1\}$. Esta técnica consiste en dejar fuera (*drop out)* cierta cantidad de neuronas de cada capa durante el entrenamiento.



Dropout

Consideremos un vector (o arreglo) μ que especifica que unidades debemos aniquilar al aplicar dropout en cada una de las capas. Esto modifica en cada época los parámetros que recibe la función objetivo, en este caso usamos

objetivo: $\min_\theta (\max_\theta) \mathbb{E}_\mu \left[J(\theta, \mu) \right]$

$$\mu = \left\{ \mu^{[1]}, \mu^{[2]}, \cdots, \mu^{[\ell]}, \cdots, \mu^{[L]} \right\}$$

$$\mu^{[\ell]} = \begin{bmatrix} \mu_1^{[\ell]} \\ \vdots \\ \mu_{n_\ell}^{[\ell]} \end{bmatrix}$$

$$\mu_x^{[\ell]} \sim \text{Bernoulli}\left(p^{[\ell]} \right)$$

$$h_x^{[\ell]\mu} = \begin{cases} 0 & \text{con proba } p^{[\ell]} \\ \frac{h_x^{[\ell]}}{1-p^{[\ell]}} & \text{e.o.c} \end{cases} \quad \Longrightarrow \quad \sum_{\mu_x^{[\ell]}} p\left(\mu_x^{[\ell]\mu} \right) h_x^{[\ell]} \to \mathbb{E}_\mu \left[h_x^{[\ell]\mu} \right]$$