

# Lecture 25: Deep neural networks continued

Professor Ilias Bilonis

## Regularization through parameter penalties

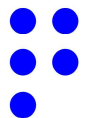
# Regularization terms in loss functions

$$J(\theta) = \underbrace{L(\theta)}_{\text{loss funct.}} + \underbrace{\lambda}_{\text{hyper-parameter}} \underbrace{R(\theta)}_{\text{regularization parameter}} + \dots$$

$$+ R(\theta) = \|\theta\|_2^2 = \sum_i \theta_i^2$$

$$+ R(\theta) = \|\theta\|_1 = \sum_i |\theta_i|$$

⋮



# Bayesian interpretation of regularization

$$\max_{\theta} p(y_{1:n} | x_{1:n}, \theta) \Rightarrow \min_{\theta} L(\theta)$$

prior over weights

$$p(\theta) \Rightarrow p(\theta | x_{1:n}, y_{1:n}) \propto p(y_{1:n} | x_{1:n}, \theta) p(\theta)$$

MAP of  $\theta$ :  $\max_{\theta} \log p(\theta | x_{1:n}, y_{1:n})$

$$J(\theta) = -\log p(\theta | x_{1:n}, y_{1:n}) = \underbrace{-\log p(y_{1:n} | x_{1:n}, \theta)}_{L(\theta)} - \underbrace{\log p(\theta)}_{\lambda R(\theta)}$$

Gaussian prior:  $\theta \sim \mathcal{N}(0, \lambda^{-1})$

$$-\log p(\theta) = -\log \mathcal{N}(\theta | 0, \lambda^{-1})$$

$$= -\lambda \|\theta\|_2^2 + \text{const.}$$

$$\Rightarrow R(\theta) = \|\theta\|_2^2$$

