# Regularization

## 1. Overfitting

- Underfitting, or high bias, is when the form of our hypothesis function h maps **poorly** to the trend of the data.
- Overfitting, or high variance, is caused by a hypothesis function that fits the available data but does not generalize well to predict new data.

### Addressing

- Reduce the number of features (manually select, model selection).
- Regularization (keep all the features, but reduce the magnitude of parameters).

### 2. Regularized Linear Regression

$$\min_{\theta} \frac{1}{2m} \left[ \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2 \right], \text{ where } h_{\theta}(x) = \theta^T x$$

• Normal equation

$$\theta = (X^TX + \lambda \cdot L)^{-1}X^Ty \text{where } L = \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$$

Prove.

$$\left\{ \begin{array}{ll} \frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{k=1}^m (\theta^T x^{(k)} - y^{(k)}) x_j^{(k)} & (j=0) \\ \frac{\partial}{\partial \theta_i} J(\theta) = \frac{1}{m} \sum_{k=1}^m (\theta^T x^{(k)} - y^{(k)}) x_j^{(k)} + \frac{\lambda}{m} \theta_j & (j\in N^+) \end{array} \right. \\ \Rightarrow \nabla_{\theta} J(\theta) = \frac{1}{m} (X^T X \theta - X^T y) + \frac{\lambda}{m} L \theta_j +$$

## 3. Regularized Logistic Regression

$$\min_{\theta} \left[ -\frac{1}{m} \sum_{i=1}^{m} \left( y^{(i)} \mathrm{log} h_{\theta}(x^{(i)}) + (1-y^{(i)}) \mathrm{log} (1-h_{\theta}(x^{(i)})) \right) + \frac{\lambda}{2m} \sum_{j=1}^{n} \theta_{j}^{2} \right]$$

### 4. MLE & MAP

#### **Preliminaries**

- Assume data are generated via  $d \sim p(d; \theta)$
- $D = \{d^{(i)}\}_{i=1,2,\cdots,m}$ , where  $d^{(i)}$  is i.i.d. (independent of others and same distribution).
- Goal: Estimate parameter  $\theta$  that best models the data.

## Maximum Likelihood Estimation (MLE)

- $\begin{array}{l} \bullet \ \ \text{Likelihood:} \ L(\theta) = p(D;\theta) = \prod_{i=1}^m p(d^{(i)};\theta) \\ \bullet \ \ \text{MLE typically maximizes the $\log -$likelihood} \ l(\theta). \\ \bullet \ \ \theta_{MLE} = \arg\max_{\theta} \ \sum_{i=1}^m \log p(d^{(i)};\theta) \\ \end{array}$

## Maximum-a-Posteriori Estimation (MAP)

- Posterior probability of  $\theta$  is  $p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$   $p(\theta)$  is prior probability of  $\theta$ , where p(D) is probability of the data.
- MAP usually maximizes the  $\log$  of the posteriori probability  $\theta_{MAP} = \arg\max_{\theta} \ \left( \log p(\theta) + \sum_{i=1}^{m} \log p(d^{(i)}|\theta) \right)$

## Linear Regression

### 1. MLE

- $$\begin{split} \bullet & \text{ Suppose } y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}, \text{ where } \epsilon \sim \mathcal{N}(0, \delta^2) \\ \bullet & p(x) = \frac{1}{\sigma \sqrt{2\pi}} \, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ \bullet & \theta_{MLE} = \arg\min_{\theta} \sum_{i=1}^m (y^{(i)} \theta^T x^{(i)})^2 \end{split}$$

#### 2. MAP

- $$\begin{split} \bullet & \text{ Suppose } \theta \sim \mathcal{N}(0, \lambda^2 I) \\ \bullet & \log p(\theta) = n \! \log \! \frac{1}{\sqrt{2\pi}\lambda} \frac{\theta^T \theta}{2\lambda^2} \\ \bullet & \theta_{MAP} = \arg \min_{\theta} \left\{ \sum_{i=1}^m (y^{(i)} \theta^T x^{(i)})^2 + \frac{\theta^T \theta}{2\lambda^2} \right\} \end{split}$$

## 3. MLE vs MAP

- MLE (unregularized solution) vs MAP (regularized solution)
- The prior distribution acts as a regularizer in MAP estimation

#### Logistic Regression

Similar conclusion as above.