1 What is Regularization?

- **Regularization** is a set of techniques used to reduce the complexity of a machine learning model and prevent it from overfitting the training data.
- Regularization adds a **penalty** to the model's **loss function**, discouraging it from fitting the training data too perfectly.
- Regularization works by adding a **regularization term** to the loss function (which the model tries to minimize).

1.1 Example

This is the original loss function (for example):

$$cost(W) = \frac{1}{2N} \sum_{i=1}^{N} (y(X^n, W) - t^n)^2 \tag{1}$$

After adding one of regularization types (ridge for example) it will be:

$$cost(W) = \frac{1}{2N} \sum_{i=1}^{N} \left(y(X^i, W) - t^i\right)^2 + \sum_{\mathbf{i=1}}^{\mathbf{M}} \left(\frac{\lambda}{\mathbf{2}} \mathbf{W_i^2}\right) \tag{2}$$

The new term is the **Regularization Term**.

Given one of the weights W_j, the partial derivative will be:

$$\frac{\partial cost(W)}{\partial W_j} = \frac{1}{N} \sum_{i=1}^{N} \left(y(X^i, W) - t^n\right)^2 * X_j^i + \lambda W_j \tag{3}$$

2 Ridge Regression

- We also call it L2 Reguralization.
- Adds the squared magnitude of all weights to the loss function:

$$\textbf{Penalty} = \sum_{i=1}^{M} w_j^2 \tag{4}$$

- It encourages **smaller weights**.
- It keeps all features, but shrinks their influence.
- **Notice** that we started with **i=1**, which means we don't penalize the intercept at W_0.

3 Ridge Regression

- Adds the **absolute value** of weights to the loss function:

$$\textbf{Penalty} = \sum_{j}^{M} |w_j| \tag{5}$$

- Encourages **sparsity** — sets some weights to zero.
- Can be used for **feature selection**.