Normalization, Regularization

$$L(w) = \frac{1}{N_x} \sum_x div(f(x; w), D(x))$$

$$\hat{w} = \arg\min_w L(w)$$

$$\nabla_w L(w) = \frac{1}{N_x} \sum_x \underbrace{\nabla_w div(f(x; w), D(x))}$$

Computed using backpropagation

Gradient Descent

$$W_k = W_{k-1} - \eta \nabla_w L(w)^T$$

- Batch Methods: consider all training points before making an update
- Online Methods: present training instances incrementally to make quick updates
  - (need to shrink learning rate to converge)
  - (greater variance than batch method)