MCAP as Regularization

$$W \leftarrow \arg\max_{W} \sum_{l} \underbrace{\ln P(Y^l|X^l,W) - \frac{\lambda}{2}||W||^2}$$

$$\frac{\partial l(W)}{\partial w_i} = \sum_l X_i^l (Y^l - \hat{P}(Y^l = 1|X^l, W)) - \lambda w_i$$

- Weight update rule:

$$w_i \leftarrow w_i + \eta \sum_l X_i^l (Y^l - \hat{P}(Y^l = 1|X^l, W)) - \eta\lambda w_i$$

  - Quadratic penalty: drives weights towards zero
  - Adds a negative linear term to the gradients

Penalizes high weights, like we did in linear regression