Deep Learning Basics
Lecture 3: Regularization I

Princeton University COS 495

Instructor: Yingyu Liang

What is regularization?

- In general: any method to prevent overfitting or help the optimization

- Specifically: additional terms in the training optimization objective to prevent overfitting or help the optimization

Review: overfitting

Overfitting example: regression using polynomials

$$t = \sin(2\pi x) + \epsilon$$

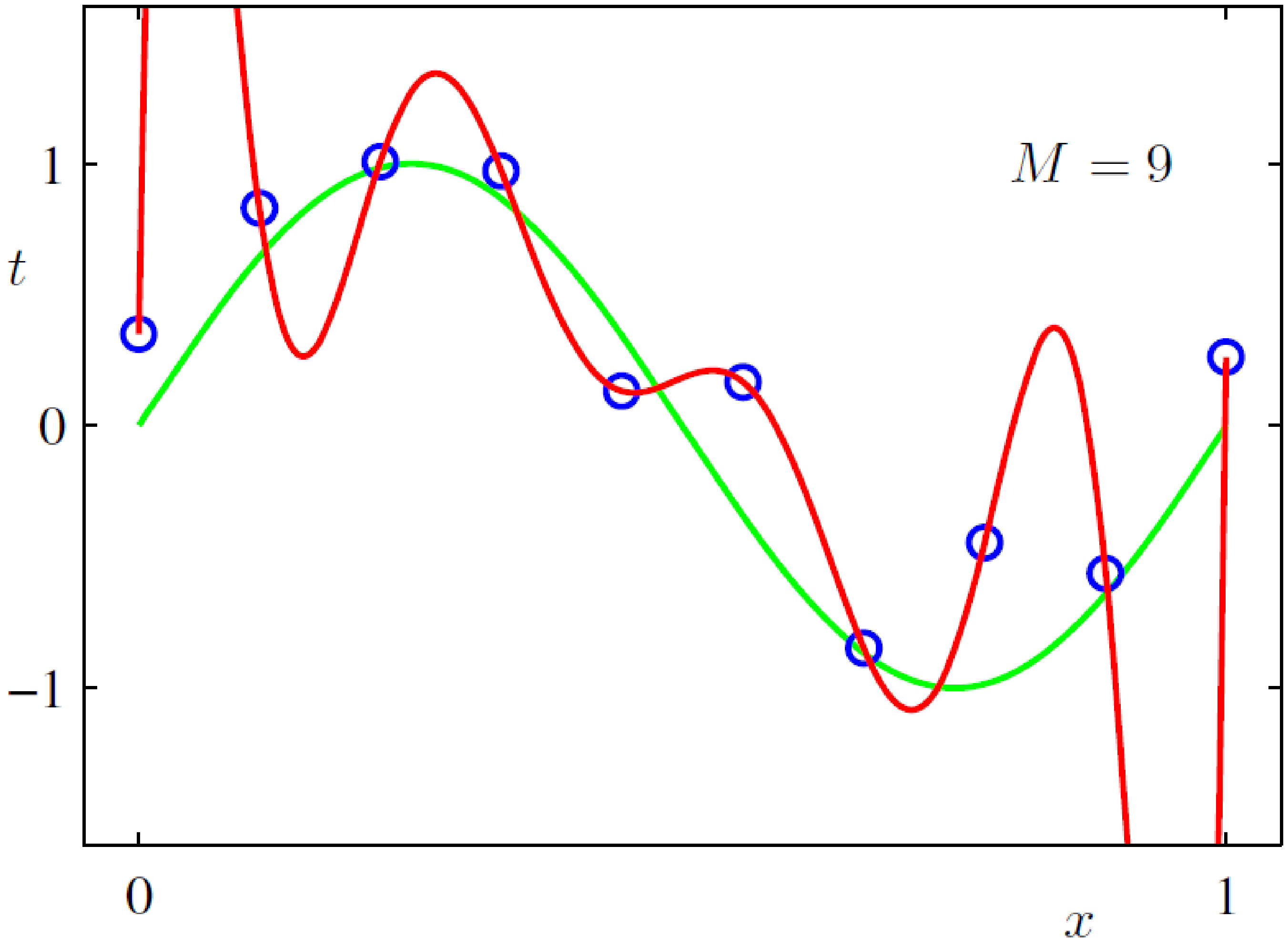


Figure from *Machine Learning and Pattern Recognition*, Bishop

Overfitting example: regression using polynomials

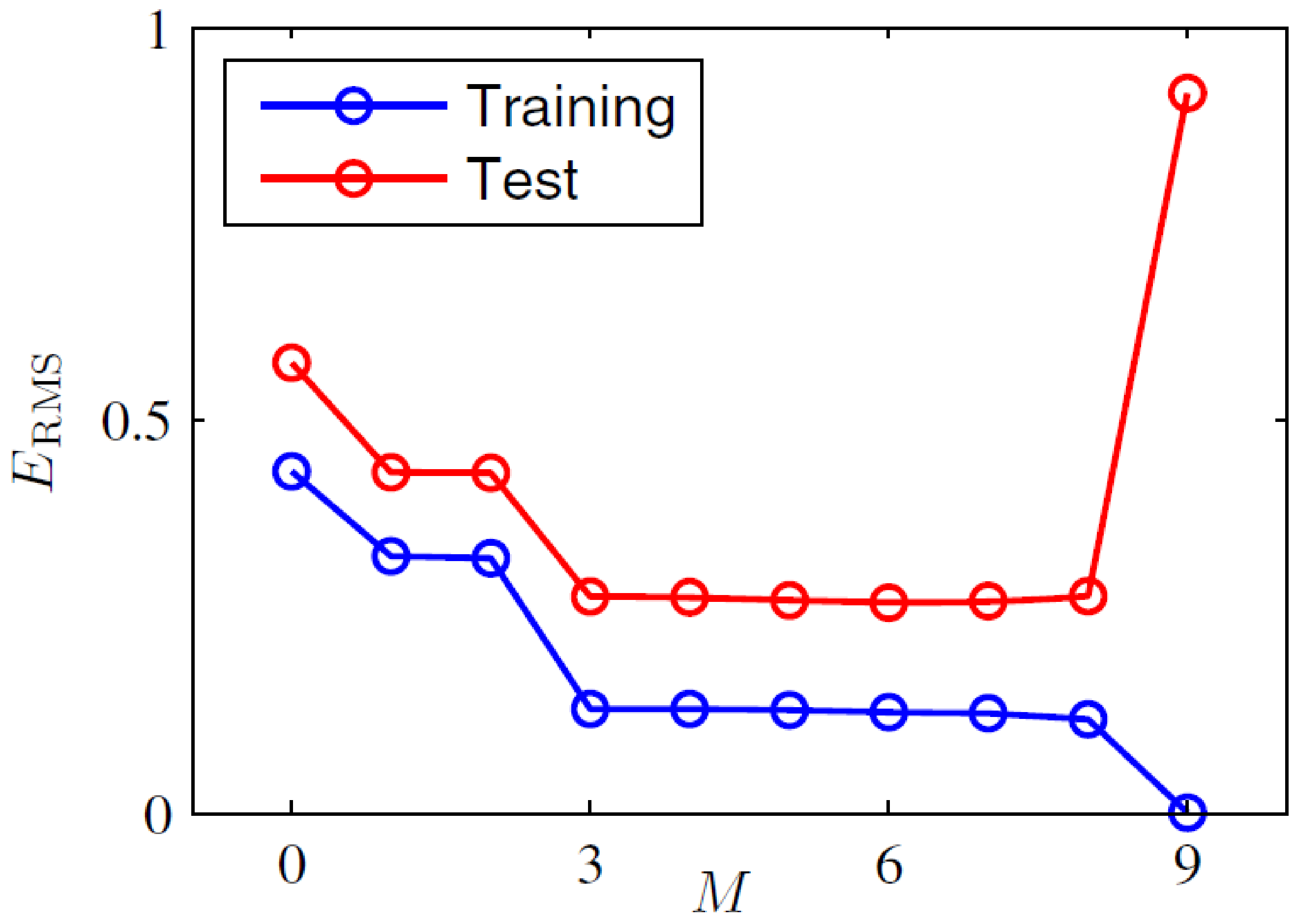


Figure from *Machine Learning and Pattern Recognition*, Bishop

Overfitting

- Empirical loss and expected loss are different

- Smaller the data set, larger the difference between the two
- Larger the hypothesis class, easier to find a hypothesis that fits the difference between the two
  - Thus has small training error but large test error (overfitting)

Prevent overfitting

- Larger data set helps
- Throwing away useless hypotheses also helps

- Classical regularization: some principal ways to constrain hypotheses
- Other types of regularization: data augmentation, early stopping, etc.

Different views of regularization

Regularization as hard constraint

- Training objective

$$\min_{f} \hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} l(f, x_i, y_i)$$

$$\text{subject to:} \quad f \in \boldsymbol{\mathcal{H}}$$

- When parametrized

$$\min_{\theta} \hat{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(\theta, x_i, y_i)$$

$$\text{subject to:} \quad \theta \in \Omega$$

Regularization as hard constraint

- When Ω measured by some quantity R

$$\min_{\theta} \hat{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(\theta, x_i, y_i)$$

$$\text{subject to: } R(\theta) \leq r$$

- Example: l_2 regularization

$$\min_{\theta} \hat{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(\theta, x_i, y_i)$$

$$\text{subject to: } ||\theta||_2^2 \leq r^2$$

Regularization as soft constraint

- The hard-constraint optimization is equivalent to soft-constraint

$$\min_{\theta} \hat{L}_R(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(\theta, x_i, y_i) + \lambda^* R(\theta)$$

  for some regularization parameter $\lambda^* > 0$

- Example: l_2 regularization

$$\min_{\theta} \hat{L}_R(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(\theta, x_i, y_i) + \lambda^* ||\theta||_2^2$$

Regularization as soft constraint

- Showed by Lagrangian multiplier method

$$\mathcal{L}(\theta, \lambda) := \hat{L}(\theta) + \lambda[R(\theta) - r]$$

- Suppose θ^* is the optimal for hard-constraint optimization

$$\theta^* = \operatorname*{argmin}_{\theta} \max_{\lambda \geq 0} \mathcal{L}(\theta, \lambda) := \hat{L}(\theta) + \lambda[R(\theta) - r]$$

- Suppose λ^* is the corresponding optimal for max

$$\theta^* = \operatorname*{argmin}_{\theta} \mathcal{L}(\theta, \lambda^*) := \hat{L}(\theta) + \lambda^*[R(\theta) - r]$$

Regularization as Bayesian prior

- Bayesian view: everything is a distribution
- Prior over the hypotheses: $p(\theta)$
- Posterior over the hypotheses: $p(\theta \mid \{x_i, y_i\})$
- Likelihood: $p(\{x_i, y_i\} | \theta)$

- Bayesian rule:

$$p(\theta \mid \{x_i, y_i\}) = \frac{p(\theta) p(\{x_i, y_i\} | \theta)}{p(\{x_i, y_i\})}$$

Regularization as Bayesian prior

- Bayesian rule:

$$p(\theta \mid \{x_i, y_i\}) = \frac{p(\theta)p(\{x_i, y_i\}|\theta)}{p(\{x_i, y_i\})}$$

- Maximum A Posteriori (MAP):

$$\max_{\theta} \log p(\theta \mid \{x_i, y_i\}) = \max_{\theta} \underbrace{\log p(\theta)}_{\text{Regularization}} + \underbrace{\log p(\{x_i, y_i\} \mid \theta)}_{\text{MLE loss}}$$

Regularization as Bayesian prior

- Example: l_2 loss with l_2 regularization

$$\min_{\theta} \hat{L}_R(\theta) = \frac{1}{n}\sum_{i=1}^{n}(f_\theta(x_i) - y_i)^2 + \lambda^* ||\theta||_2^2$$

- Correspond to a normal likelihood $p(x, y \mid \theta)$ and a normal prior $p(\theta)$

Three views

- Typical choice for optimization: soft-constraint

$$\min_{\theta} \hat{L}_R(\theta) = \hat{L}(\theta) + \lambda R(\theta)$$

- Hard constraint and Bayesian view: conceptual; or used for derivation

Three views

- Hard-constraint preferred if
  - Know the explicit bound $R(\theta) \leq r$
  - Soft-constraint causes trapped in a local minima with small θ
  - Projection back to feasible set leads to stability

- Bayesian view preferred if
  - Know the prior distribution

Some examples

Classical regularization

- Norm penalty
  - l_2 regularization
  - l_1 regularization

- Robustness to noise

l_2 regularization

$$\min_{\theta} \hat{L}_R(\theta) = \hat{L}(\theta) + \frac{\alpha}{2}||\theta||_2^2$$

- Effect on (stochastic) gradient descent
- Effect on the optimal solution

Effect on gradient descent

- Gradient of regularized objective

$$\nabla \hat{L}_R(\theta) = \nabla \hat{L}(\theta) + \alpha\theta$$

- Gradient descent update

$$\theta \leftarrow \theta - \eta\nabla \hat{L}_R(\theta) = \theta - \eta\, \nabla \hat{L}(\theta) - \eta\alpha\theta = (1 - \eta\alpha)\theta - \eta\, \nabla \hat{L}(\theta)$$

- Terminology: weight decay

Effect on the optimal solution

- Consider a quadratic approximation around θ^*

$$\hat{L}(\theta) \approx \hat{L}(\theta^*) + (\theta - \theta^*)^T \nabla \hat{L}(\theta^*) + \frac{1}{2}(\theta - \theta^*)^T H (\theta - \theta^*)$$

- Since θ^* is optimal, $\nabla \hat{L}(\theta^*) = 0$

$$\hat{L}(\theta) \approx \hat{L}(\theta^*) + \frac{1}{2}(\theta - \theta^*)^T H (\theta - \theta^*)$$

$$\nabla \hat{L}(\theta) \approx H(\theta - \theta^*)$$

Effect on the optimal solution

- Gradient of regularized objective

$$\nabla \hat{L}_R(\theta) \approx H(\theta - \theta^*) + \alpha\theta$$

- On the optimal θ_R^*

$$0 = \nabla \hat{L}_R(\theta_R^*) \approx H(\theta_R^* - \theta^*) + \alpha\theta_R^*$$

$$\theta_R^* \approx (H + \alpha I)^{-1} H\theta^*$$

Effect on the optimal solution

- The optimal

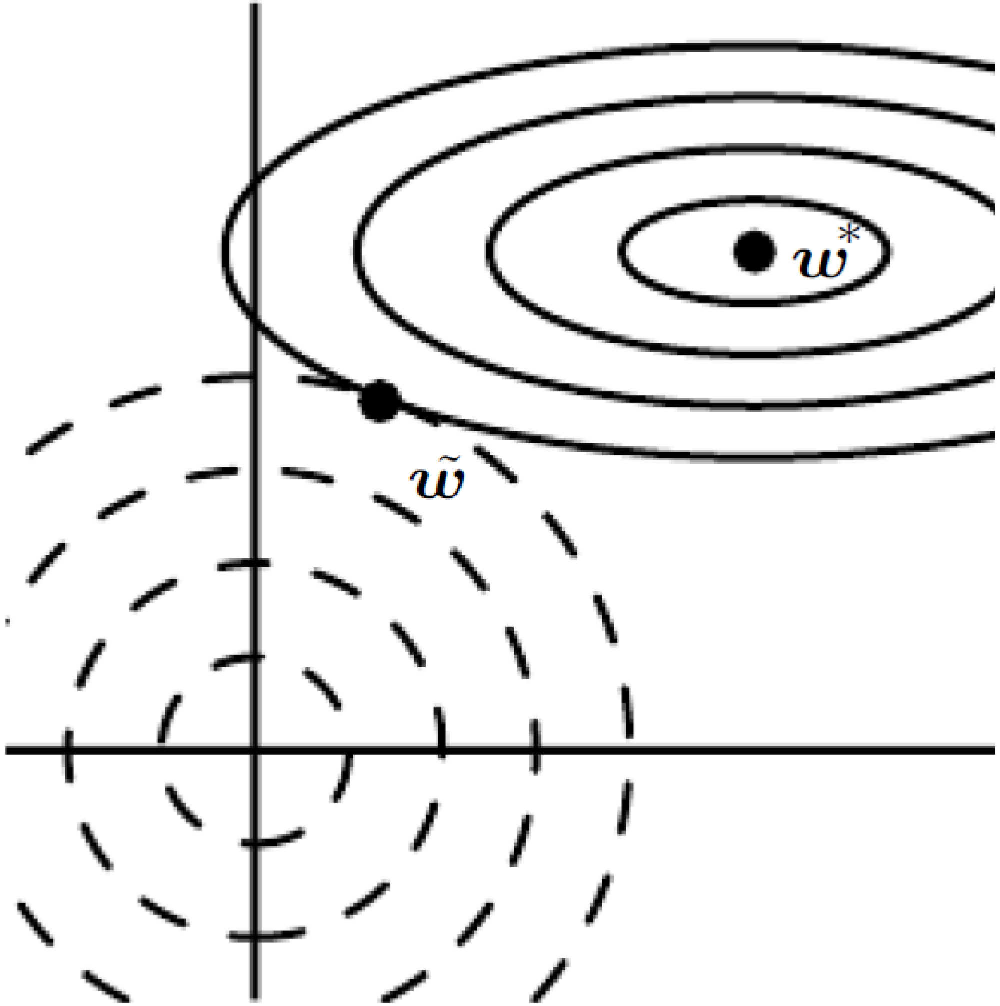
$$\theta_R^* \approx (H + \alpha I)^{-1} H \theta^*$$

- Suppose H has eigen-decomposition $H = Q \Lambda Q^T$

$$\theta_R^* \approx (H + \alpha I)^{-1} H \theta^* = Q(\Lambda + \alpha I)^{-1} \Lambda Q^T \theta^*$$

- Effect: rescale along eigenvectors of H

Effect on the optimal solution



Notations:
$\theta^* = w^*, \theta_R^* = \widetilde{w}$

Figure from *Deep Learning*, Goodfellow, Bengio and Courville

l_1 regularization

$$\min_{\theta} \hat{L}_R(\theta) = \hat{L}(\theta) + \alpha||\theta||_1$$

- Effect on (stochastic) gradient descent
- Effect on the optimal solution

Effect on gradient descent

- Gradient of regularized objective

$$\nabla \hat{L}_R(\theta) = \nabla \hat{L}(\theta) + \alpha\, \mathrm{sign}(\theta)$$

where sign applies to each element in θ

- Gradient descent update

$$\theta \leftarrow \theta - \eta \nabla \hat{L}_R(\theta) = \theta - \eta\, \nabla \hat{L}(\theta) - \eta\alpha\, \mathrm{sign}(\theta)$$

Effect on the optimal solution

- Consider a quadratic approximation around θ^*

$$\hat{L}(\theta) \approx \hat{L}(\theta^*) + (\theta - \theta^*)^T \nabla \hat{L}(\theta^*) + \frac{1}{2}(\theta - \theta^*)^T H (\theta - \theta^*)$$

- Since θ^* is optimal, $\nabla \hat{L}(\theta^*) = 0$

$$\hat{L}(\theta) \approx \hat{L}(\theta^*) + \frac{1}{2}(\theta - \theta^*)^T H (\theta - \theta^*)$$

Effect on the optimal solution

- Further assume that H is diagonal and positive $(H_{ii} > 0, \forall i)$
  - not true in general but assume for getting some intuition
- The regularized objective is (ignoring constants)

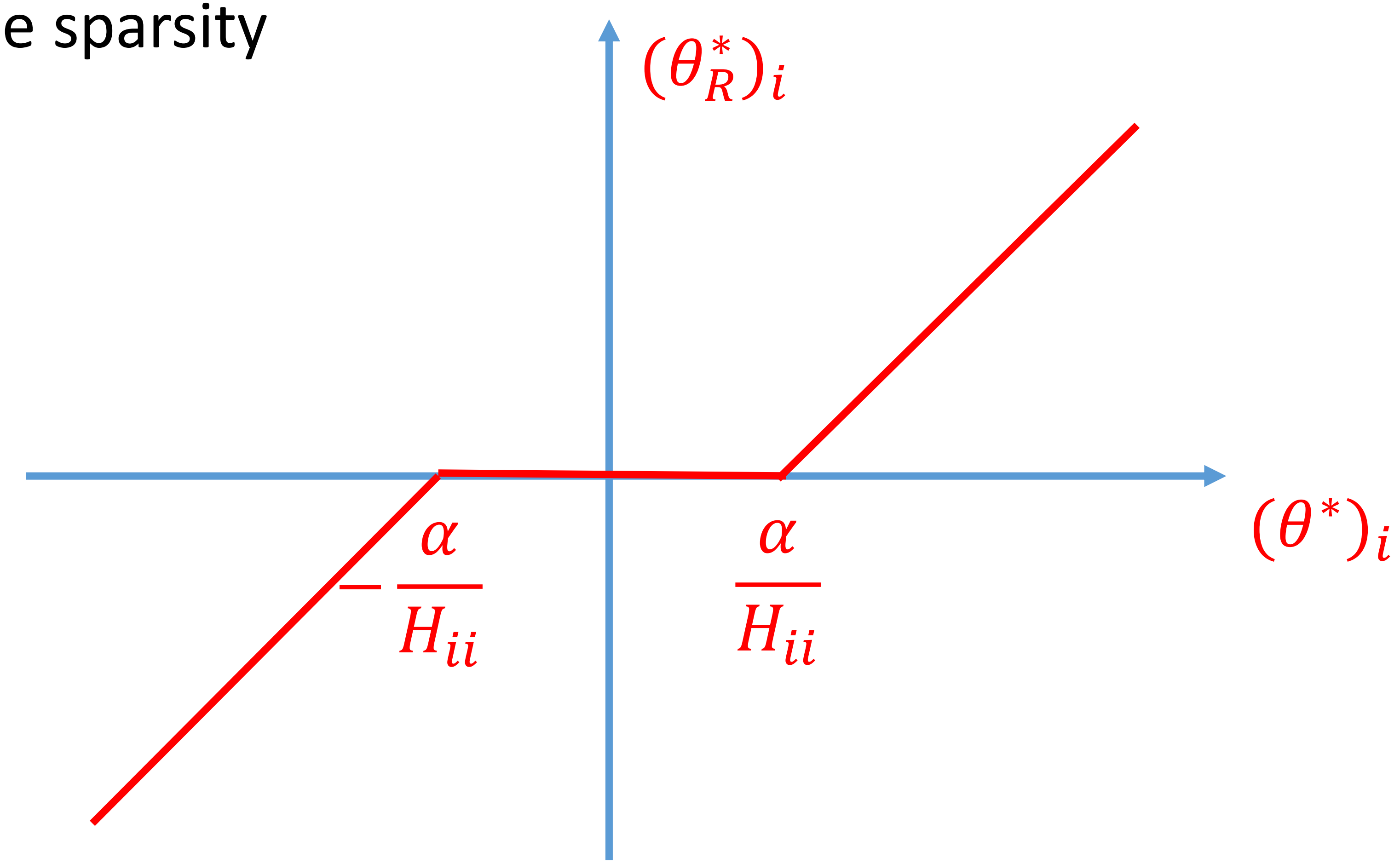
$$\hat{L}_R(\theta) \approx \sum_i \frac{1}{2} H_{ii}(\theta_i - \theta_i^*)^2 + \alpha\, |\theta_i|$$

- The optimal θ_R^*

$$(\theta_R^*)_i \approx \begin{cases} \max\left\{\theta_i^* - \dfrac{\alpha}{H_{ii}}, 0\right\} & \text{if } \theta_i^* \geq 0 \\ \min\left\{\theta_i^* + \dfrac{\alpha}{H_{ii}}, 0\right\} & \text{if } \theta_i^* < 0 \end{cases}$$

Effect on the optimal solution

- Effect: induce sparsity



Effect on the optimal solution

- Further assume that H is diagonal
- Compact expression for the optimal θ_R^*

$$(\theta_R^*)_i \approx \operatorname{sign}(\theta_i^*) \max\{|\theta_i^*| - \frac{\alpha}{H_{ii}}, 0\}$$

Bayesian view

- l_1 regularization corresponds to Laplacian prior

$$p(\theta) \propto \exp(\alpha \sum_i |\theta_i|)$$

$$\log p(\theta) = \alpha \sum_i |\theta_i| + \text{constant} = \alpha ||\theta||_1 + \text{constant}$$