EE488 Special Topics in EE <Deep Learning and AlphaGo>

Sae-Young Chung

Lecture 7

September 27, 2017

Chap. 7 Regularization for DL

- L^2 regularization
- L^1 regularization
- Dataset augmentation

L^2 Regularization

- L^2 regularization (ridge regression)

$$\tilde{J}(\mathbf{w};\mathbf{X},\mathbf{y}) = J(\mathbf{w};\mathbf{X},\mathbf{y}) + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

Usually we regularize weights only (there are far fewer bias terms and thus they do not contribute much to overfitting)

- Gradient

$$\nabla_{\mathbf{w}}\tilde{J}(\mathbf{w};\mathbf{X},\mathbf{y}) = \nabla_{\mathbf{w}}J(\mathbf{w};\mathbf{X},\mathbf{y}) + \alpha\mathbf{w}$$

- If J is quadratic near $\mathbf{w}^*$, i.e., $J(\mathbf{w}) = J(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w}-\mathbf{w}^*)^T H(\mathbf{w}-\mathbf{w}^*)$, then

$$\nabla_{\mathbf{w}}\tilde{J}(\mathbf{w};\mathbf{X},\mathbf{y}) = H(\mathbf{w}-\mathbf{w}^*) + \alpha\mathbf{w}$$

- At minimum, we have

$$\tilde{\mathbf{w}} = (H+\alpha I)^{-1}H\mathbf{w}^*$$

- If $H = Q\Lambda Q^T$, then

$$\tilde{\mathbf{w}} = (Q\Lambda Q^T + \alpha I)^{-1}Q\Lambda Q^T\mathbf{w}^* = Q(\Lambda+\alpha I)^{-1}\Lambda Q^T\mathbf{w}^*$$

L^2 Regularization

- Consider the following constrained optimization

$$\min_{\mathbf{w}:\ \frac{1}{2}\|\mathbf{w}\|^2 \le \gamma} J(\mathbf{w}; \mathbf{X}, \mathbf{y})$$

- Lagrangian: $L(\mathbf{w}, \mu) = J(\mathbf{w}; \mathbf{X}, \mathbf{y}) + \mu(\|\mathbf{w}\|^2/2 - \gamma)$
- KKT conditions

  1. $\mu \mathbf{w} + \nabla_{\mathbf{w}} J(\mathbf{w}; \mathbf{X}, \mathbf{y}) = 0$
  2. $\|\mathbf{w}\|^2 \le \gamma$, $\mu \ge 0$
  3. $\mu(\|\mathbf{w}\|^2/2 - \gamma) = 0$

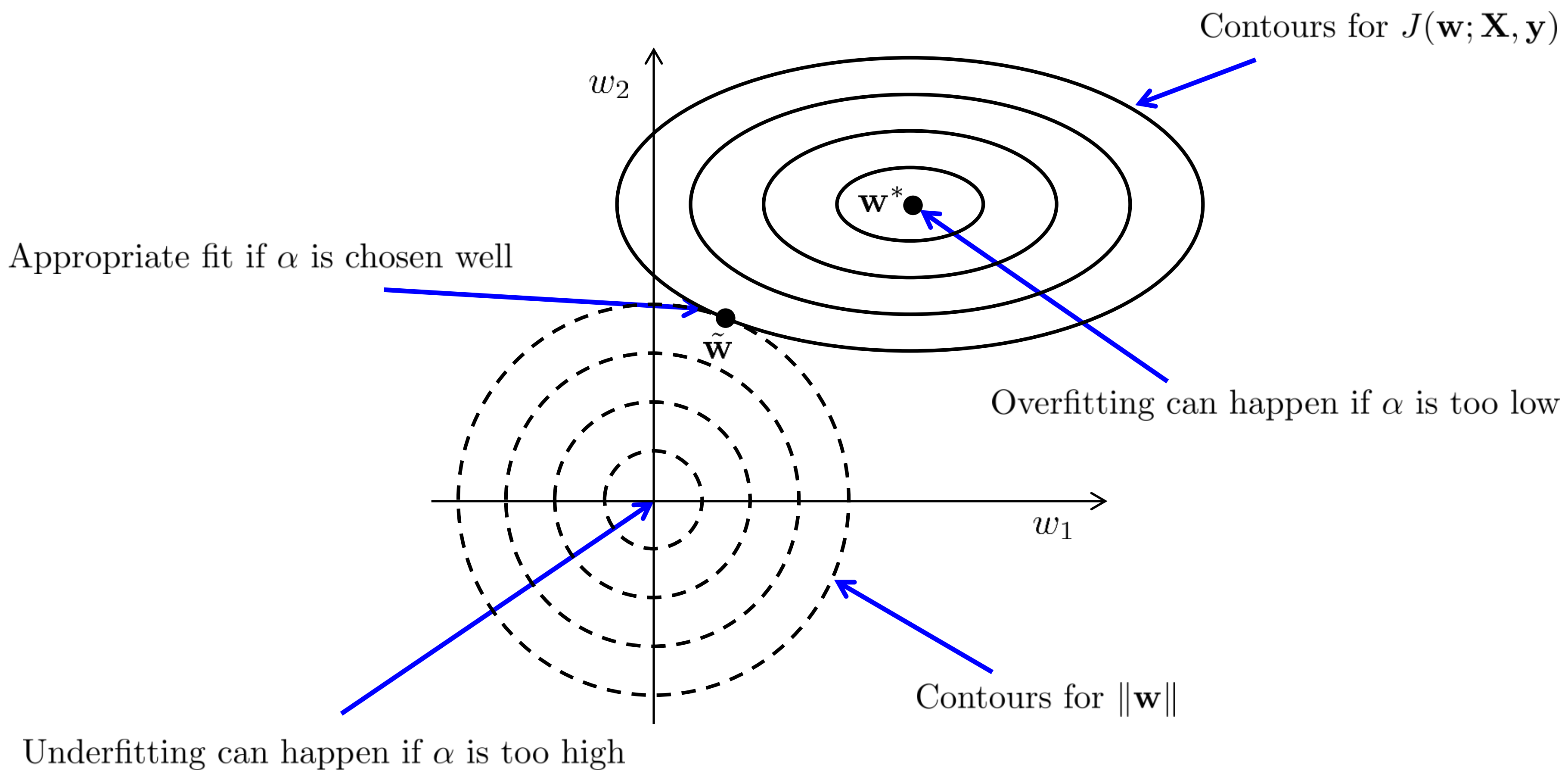
- Assume $\|\mathbf{w}\|^2/2 = \gamma$, i.e., the inequality constraint is active, then the KKT conditions is simplified as

  1. $\mu \mathbf{w} + \nabla_{\mathbf{w}} J(\mathbf{w}; \mathbf{X}, \mathbf{y}) = 0$ (*)
  2. $\mu \ge 0$

- (*) is the same as the necessary condition for local minimum for

$$J(\mathbf{w}; \mathbf{X}, \mathbf{y}) + \frac{\mu}{2} \mathbf{w}^T \mathbf{w}$$

L^2 Regularization

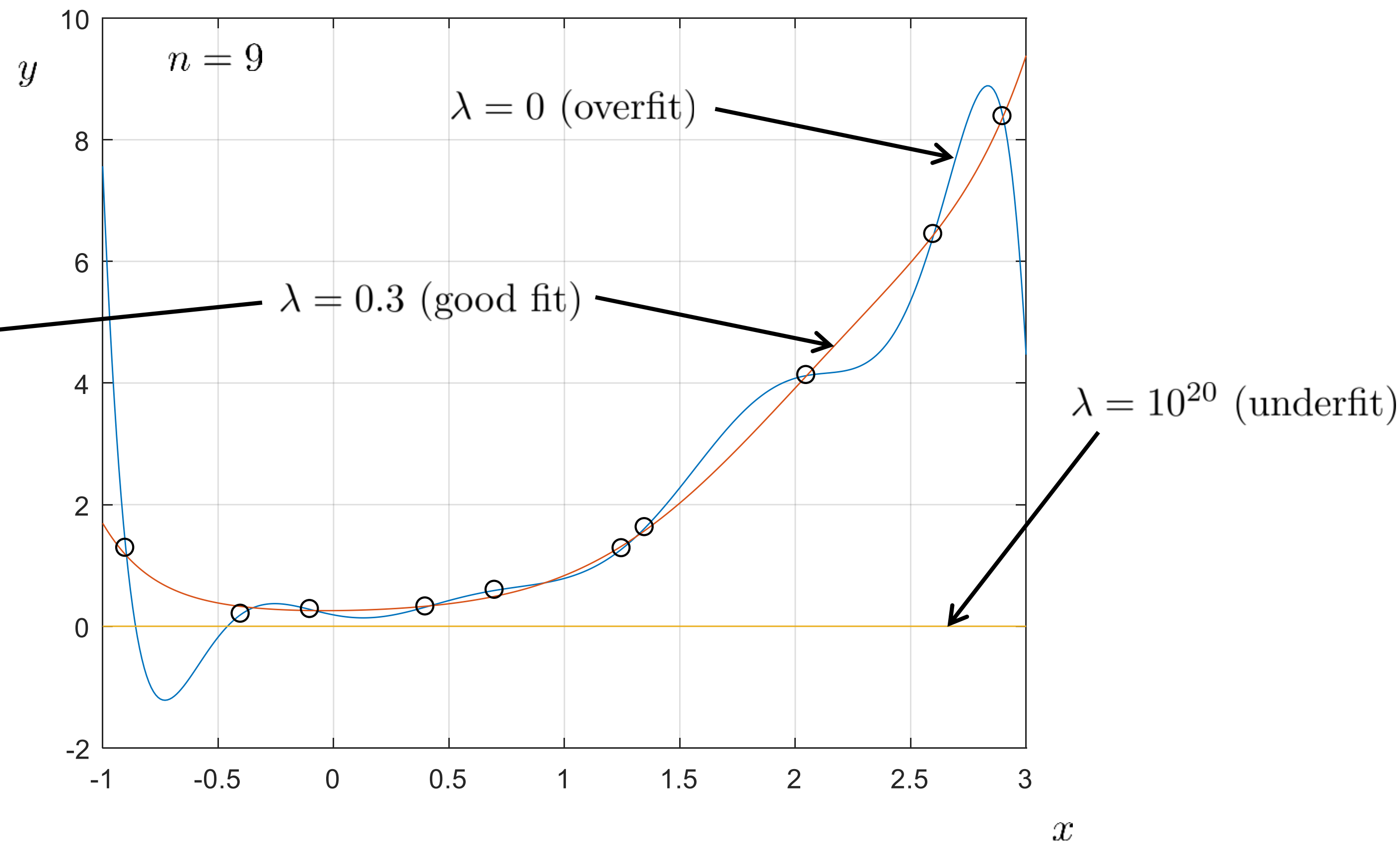


Recap – Regularization Example

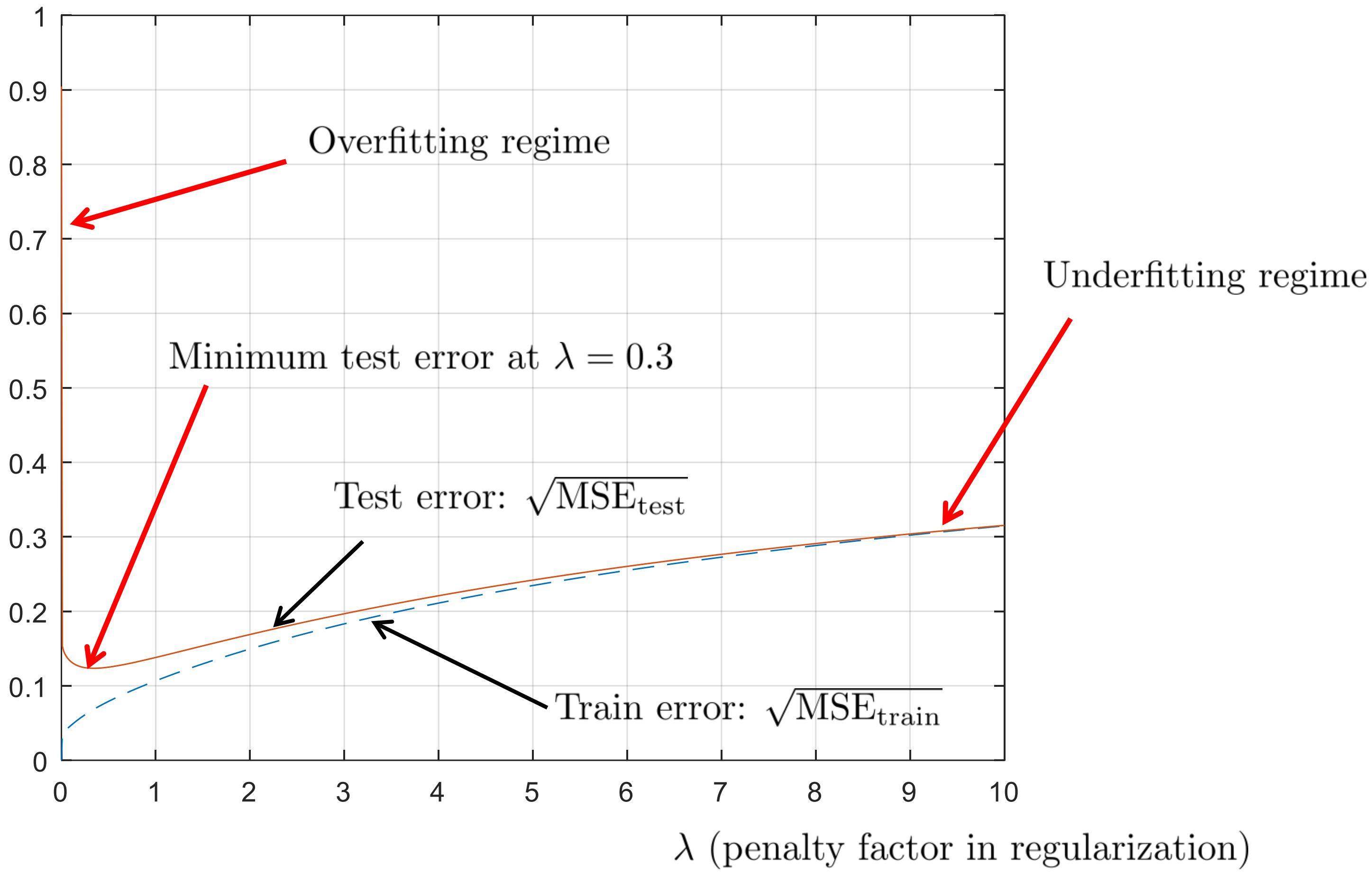
- To reduce the generalization error, we can penalize higher model complexity.

e.g., find $\mathbf{w}$ that minimizes $J(\mathbf{w}) = \text{MSE}_{\text{train}} + \lambda \mathbf{w}^T \mathbf{w}$

$$\begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \\ \theta_7 \\ \theta_8 \\ \theta_9 \end{pmatrix} = \begin{pmatrix} 0.2572 \\ 0.0297 \\ 0.3784 \\ -0.1229 \\ 0.3298 \\ -0.1311 \\ 0.2400 \\ -0.1999 \\ 0.0590 \\ -0.0056 \end{pmatrix}$$



Recap – Regularization Example



L^1 Regularization

- L^1 regularization

$$\tilde{J}(\mathbf{w}; \mathbf{X}, \mathbf{y}) = J(\mathbf{w}; \mathbf{X}, \mathbf{y}) + \alpha \|\mathbf{w}\|_1$$

- If J is quadratic near $\mathbf{w}^*$, i.e., $J(\mathbf{w}) = J(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^T H (\mathbf{w} - \mathbf{w}^*)$, then

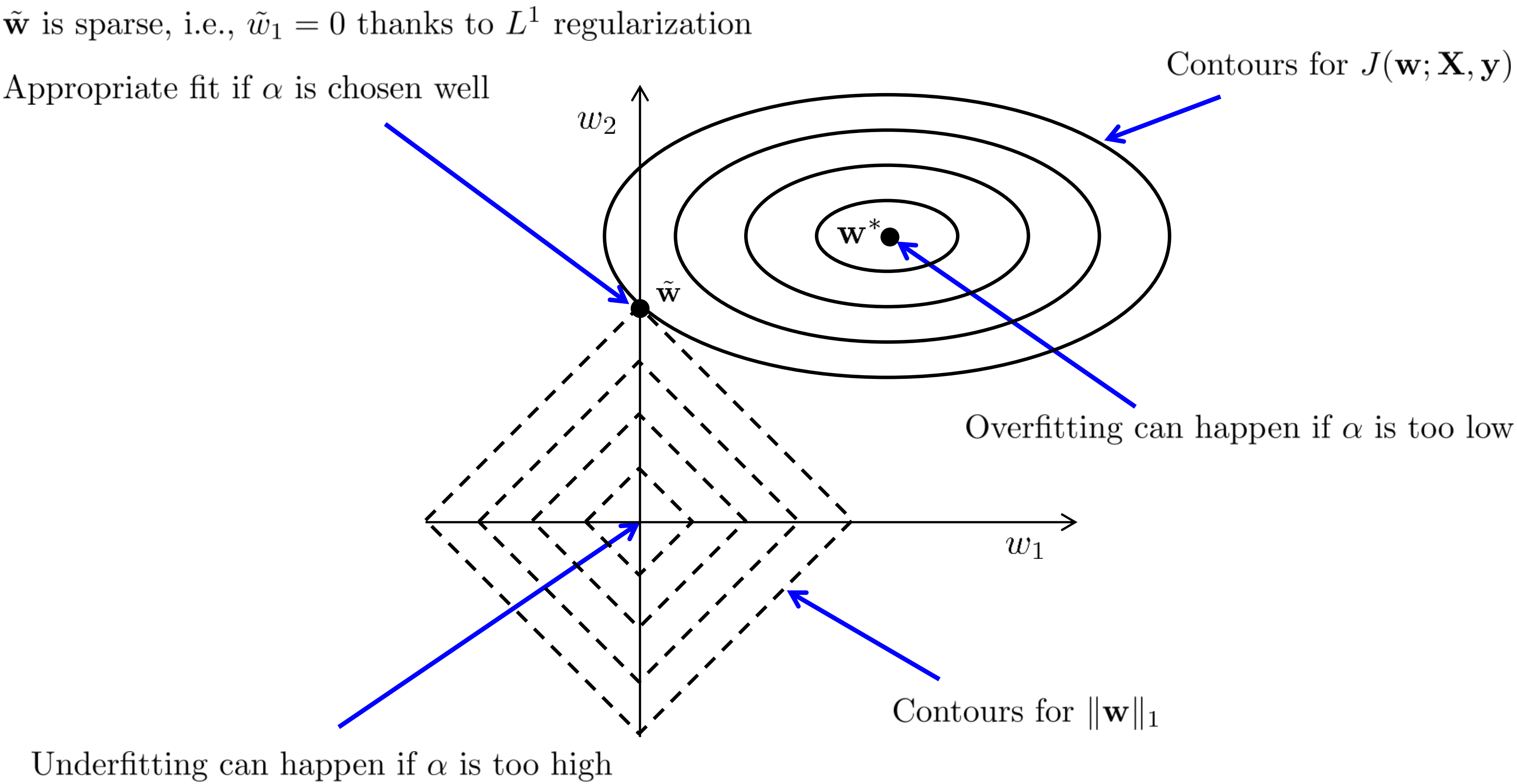
$$\tilde{J}(\mathbf{w}; \mathbf{X}, \mathbf{y}) = J(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^T H (\mathbf{w} - \mathbf{w}^*) + \alpha \|\mathbf{w}\|_1$$

- Assume H is diagonal, then the analytical solution minimizing the above is given by

$$\tilde{w}_i = \operatorname{sign}(w_i^*) \max\left\{ |w_i^*| - \frac{\alpha}{H_{i,i}}, 0 \right\}$$

- $\tilde{w}_i = 0$ if $|w_i^*| \le \frac{\alpha}{H_{i,i}}$
- L^1 regularization tends to give a more sparse solution than L^2 regularization. Sparse solution is good since it means some parameters can be set to zero, which simplifies computations.
- Related topic: Lasso (least absolute shrinkage and selection operator)

L^1 Regularization

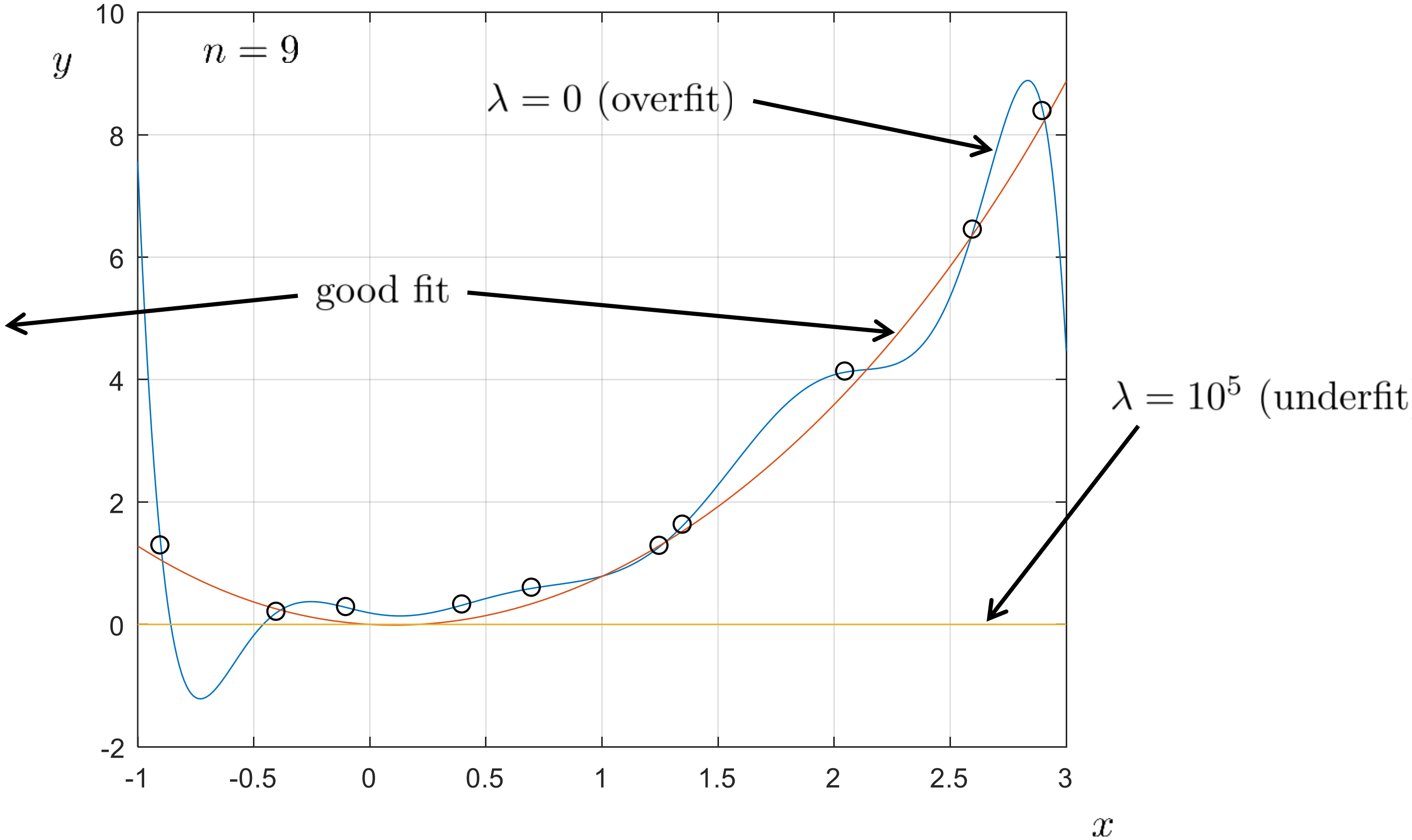


L^1 Regularization Example

- L^1 regularization

e.g., find $\mathbf{w}$ that minimizes $J(\mathbf{w}) = \mathrm{MSE}_{\mathrm{train}} + \lambda \|\mathbf{w}\|_1$

$$\begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \\ \theta_7 \\ \theta_8 \\ \theta_9 \end{pmatrix} = \begin{pmatrix} 0 \\ -0.2128 \\ 1.0172 \\ -0.0341 \\ 0.0156 \\ -0.0013 \\ 0.0001 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$



Constrained Optimization

- Norm penalty as a constraint

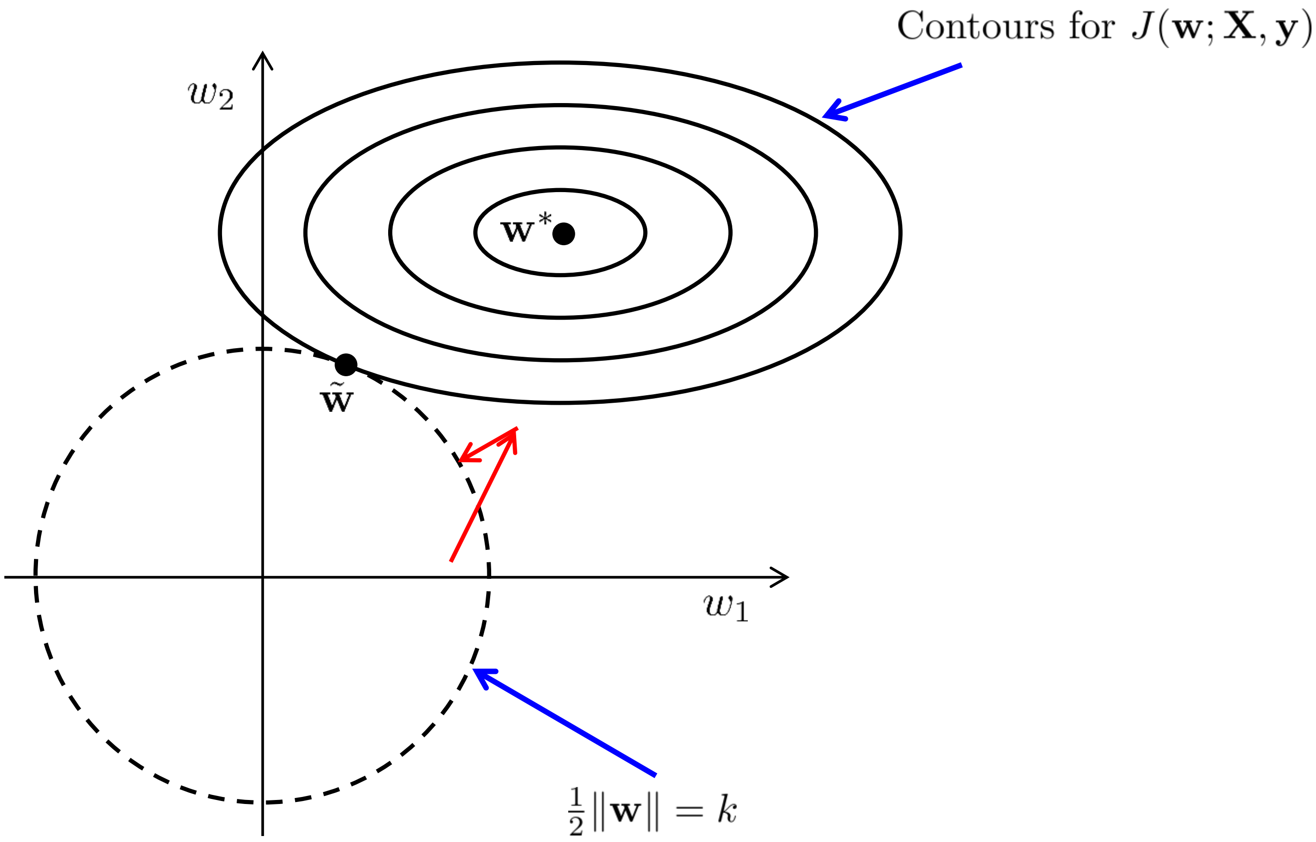
$$\min J(\boldsymbol{\theta}; \mathbf{X}, \mathbf{y}), \quad \text{subj. to } \Omega(\boldsymbol{\theta}) \leq k$$

- E.g., $\Omega(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|^2$.
- Solutions
  - Barrier method
  - Penalty method
  - Re-projection

Re-projection

- Re-projection
  - If $\Omega(\boldsymbol{\theta}) > k$ during GD, project $\boldsymbol{\theta}$ back to the nearest point satisfying $\Omega(\boldsymbol{\theta}) \leq k$
  - Regularization kicks in only when $\Omega(\boldsymbol{\theta}) > k$
  - Can prevent overflow of weights
  - Unlike L^2 (or L^1) regularization, it is now possible to handle multiple inequality constraints simultaneously and explicitly, e.g., can be used to limit the norm of each column of a weight matrix

Re-projection



Dataset Augmentation

- Having more data can reduce overfitting problem, but costly
- Dataset augmentation for images
  - Translation, rotation, scaling, color variation
- Injecting noise
  - Adding noise at the input
  - Adding noise at hidden layer units
  - Dropout ($\sim$ multiplicative noise)
  - Adding noise at the output, label smoothing
  - Adding noise to weights

Dropout

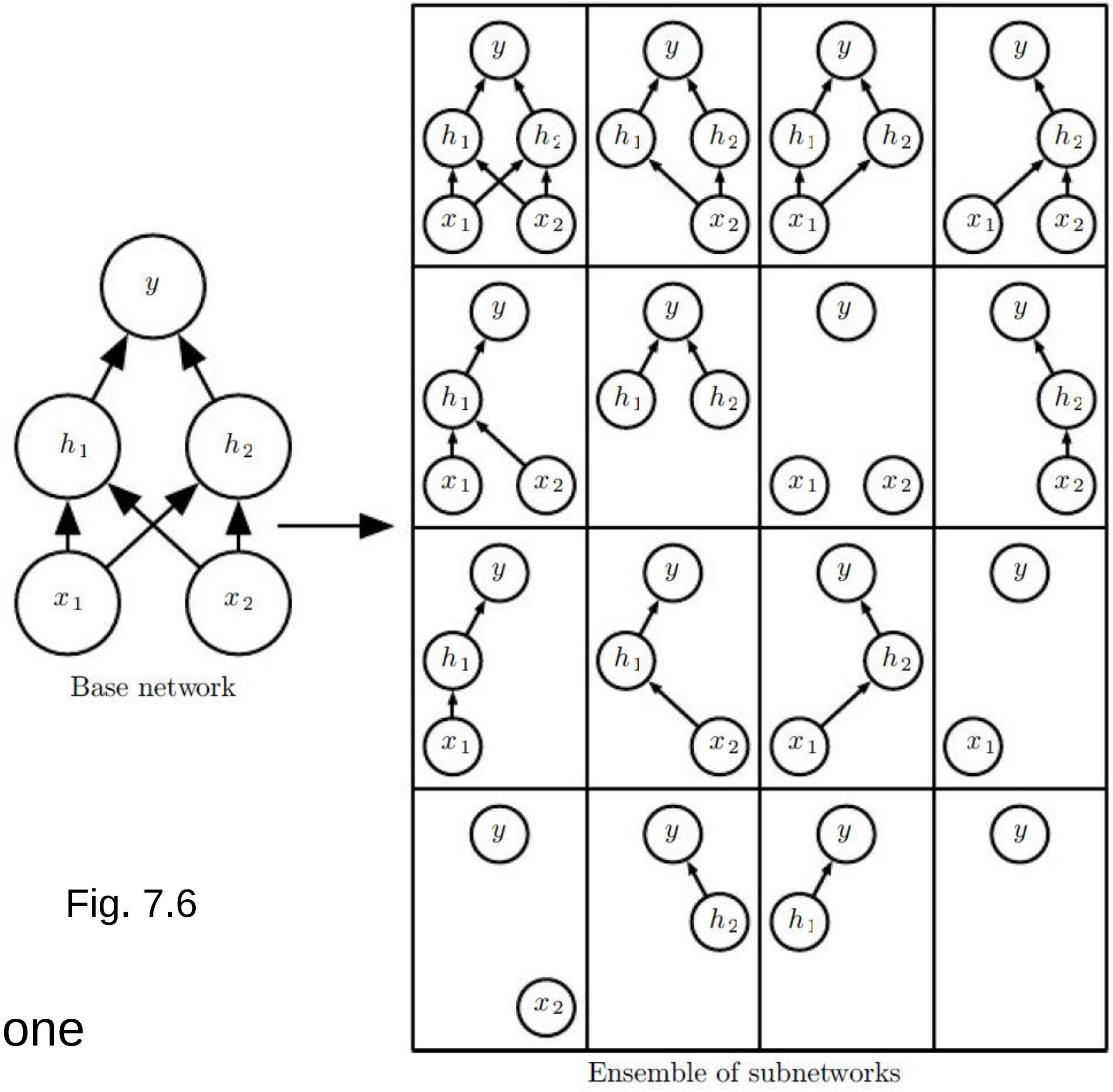


Fig. 7.6

Weight scaling can be done

Adversarial Training

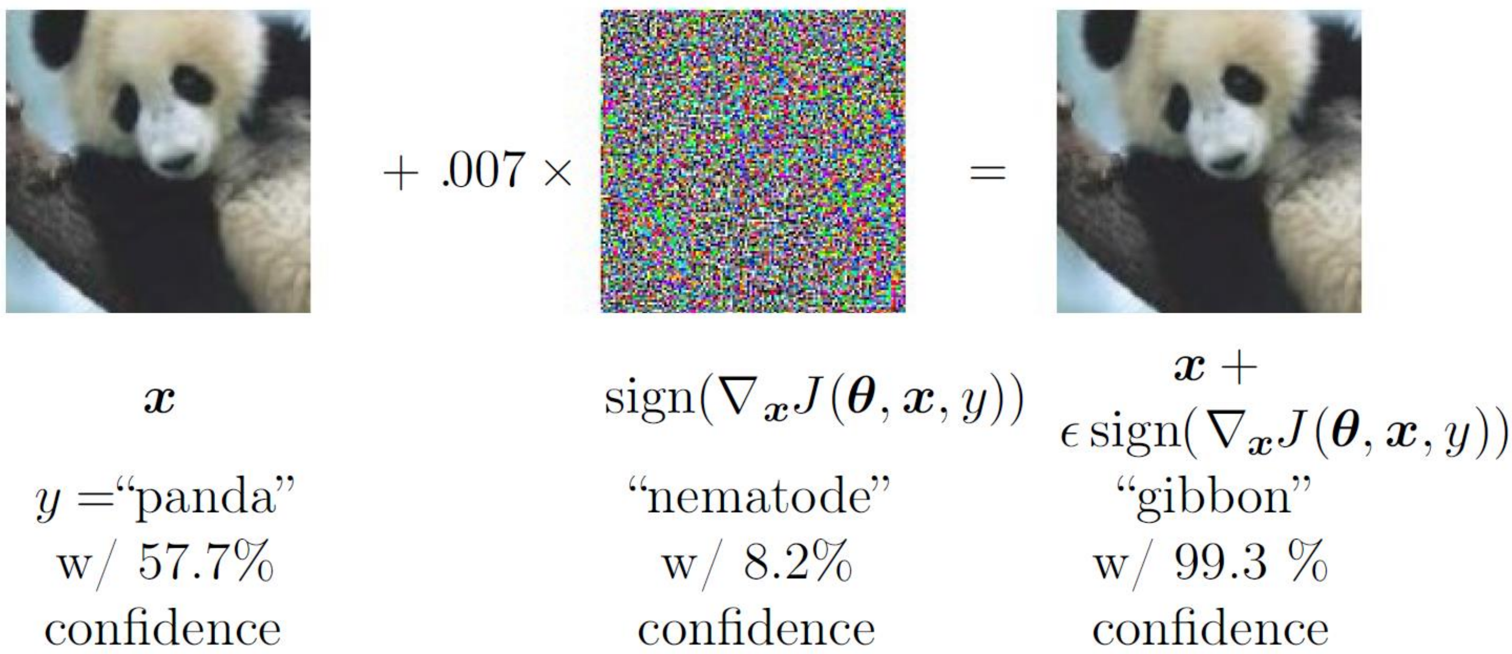
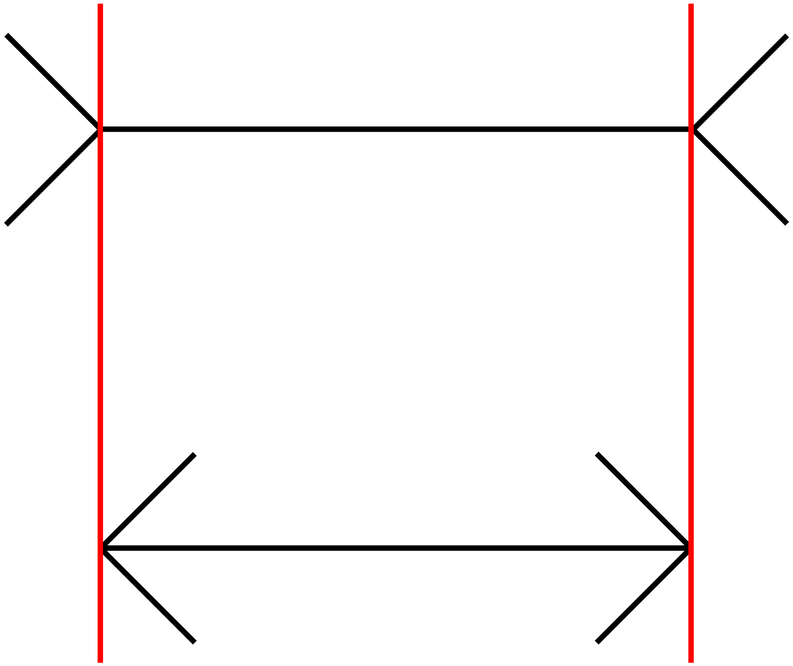
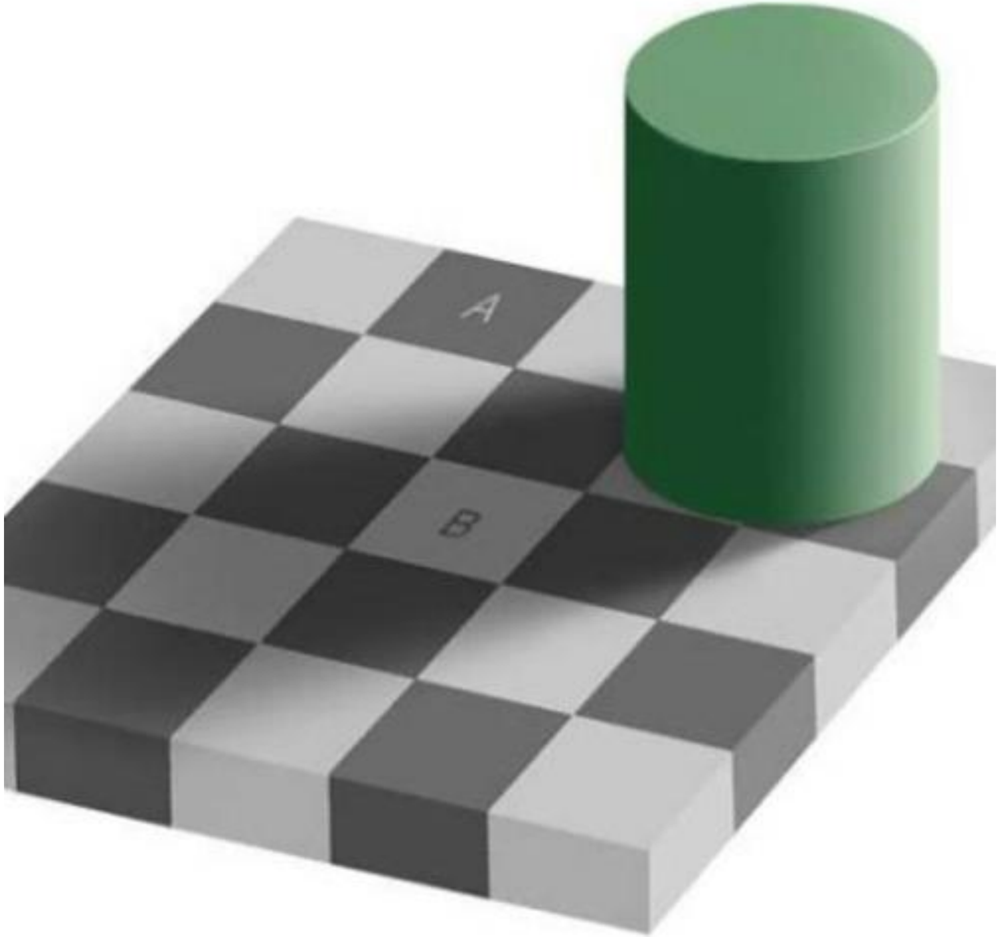


Fig. 7.8




A
B


gettyimages
Claudio Alecce / EyeEm
510427521

ABCDEFGHI
JKLMNOPQR
STUVWXYZ
DEPRTONLS

Witthoft N, Winawer J. Synesthetic colors determined by having colored refrigerator magnets in childhood. Cortex. 2006 Feb;42(2):175-83.

Other Topics

- Multi-task learning
- Early stopping
- Parameter tying and parameter sharing
- Bagging
- Ensemble methods
- Tangent prop