

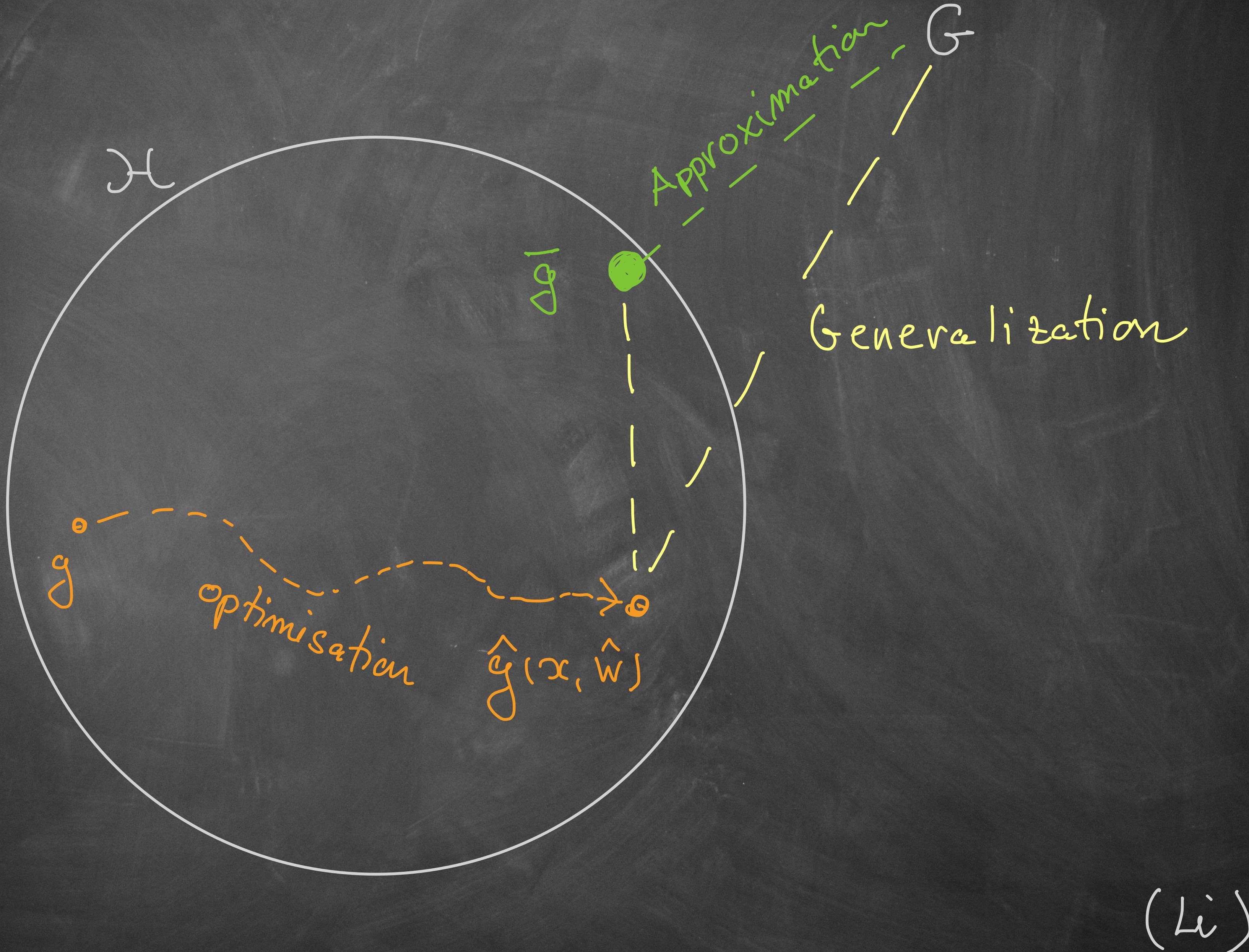
# **Mathematics of AI**

**Error sources, bias-variance, and regularisation for deep learning**

**Prince, *Understanding Deep Learning* & Bishop, *PRML* & Goodfellow et al, *Deep Learning***

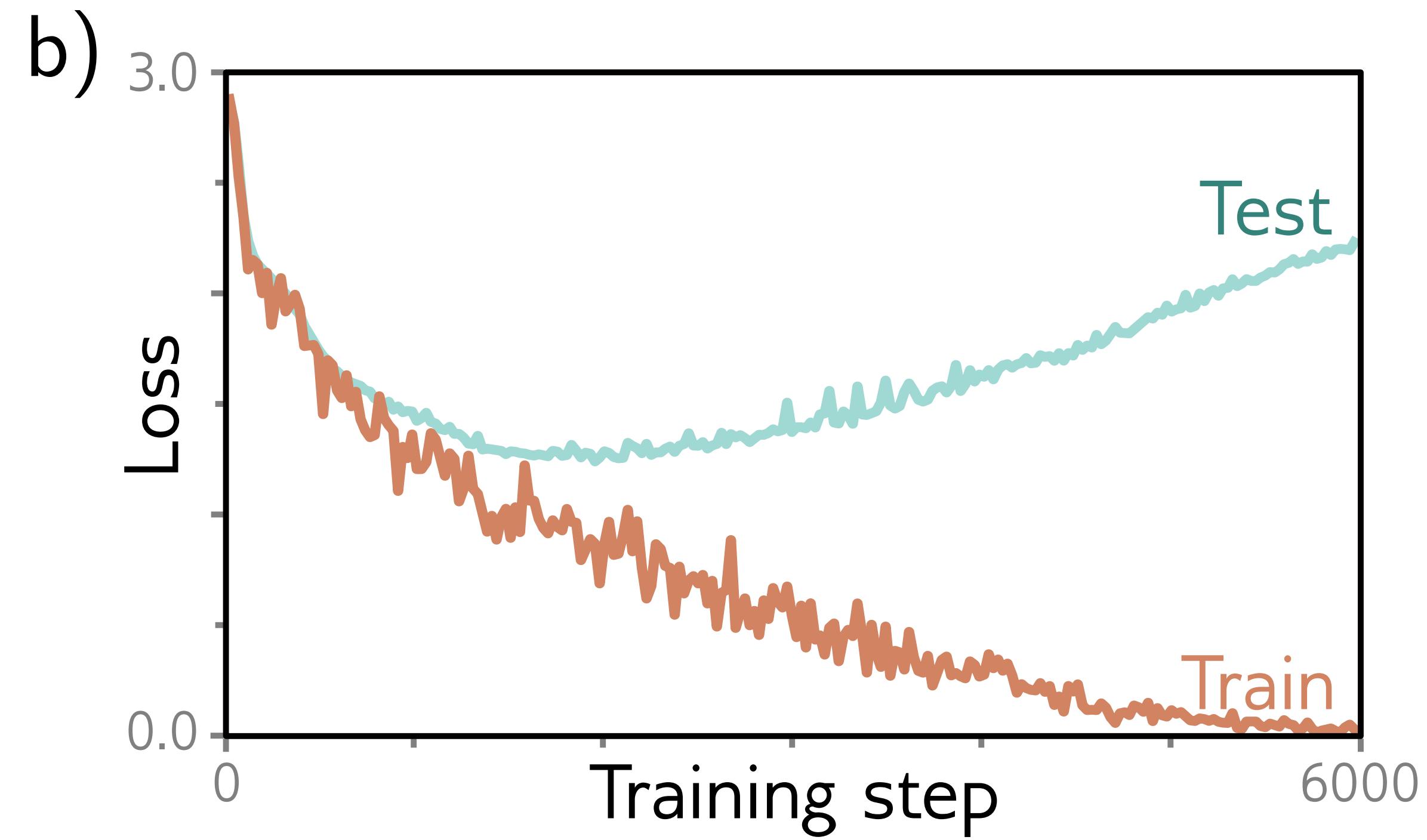
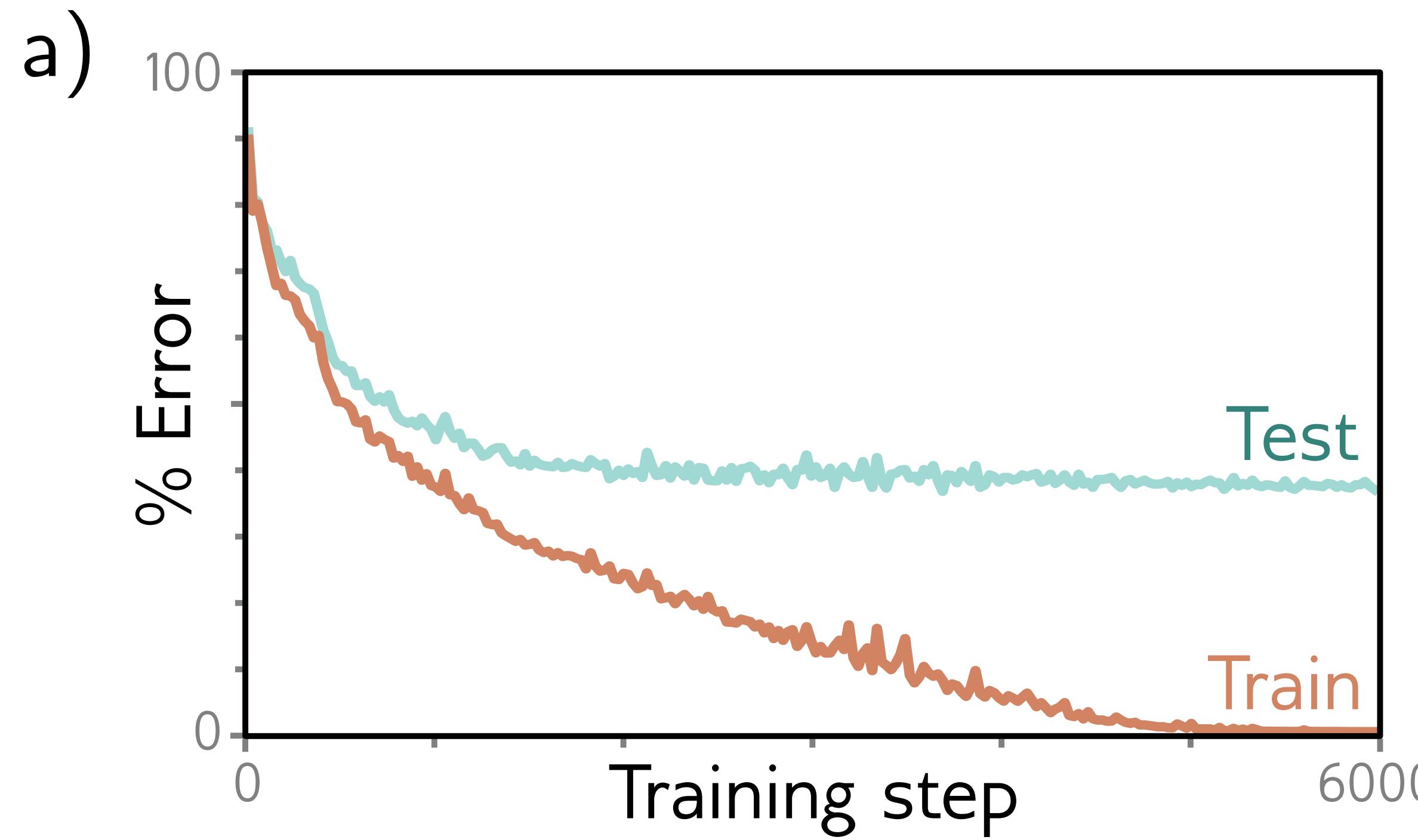
SIMON J. D. PRINCE

# Understanding Deep Learning



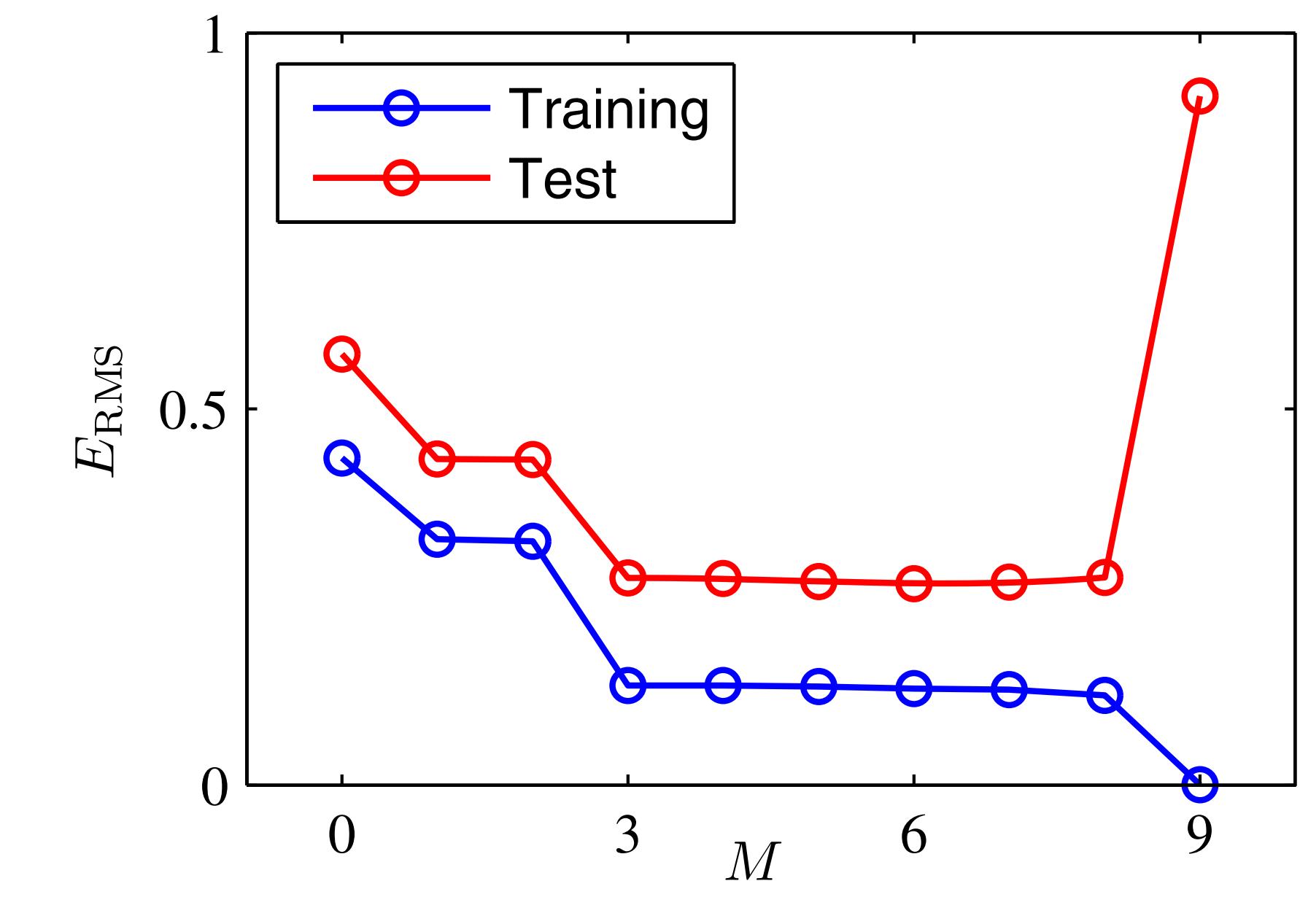
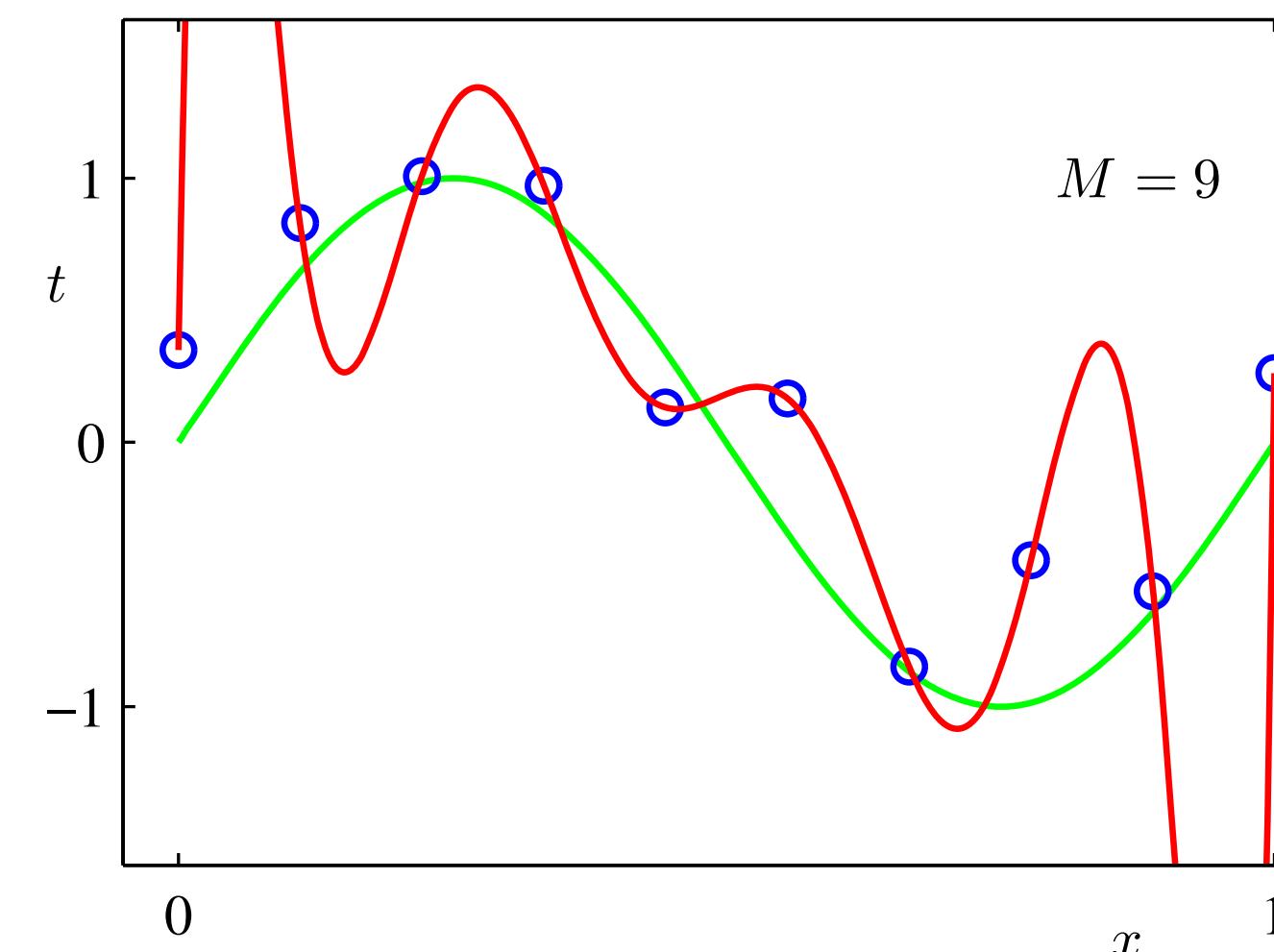
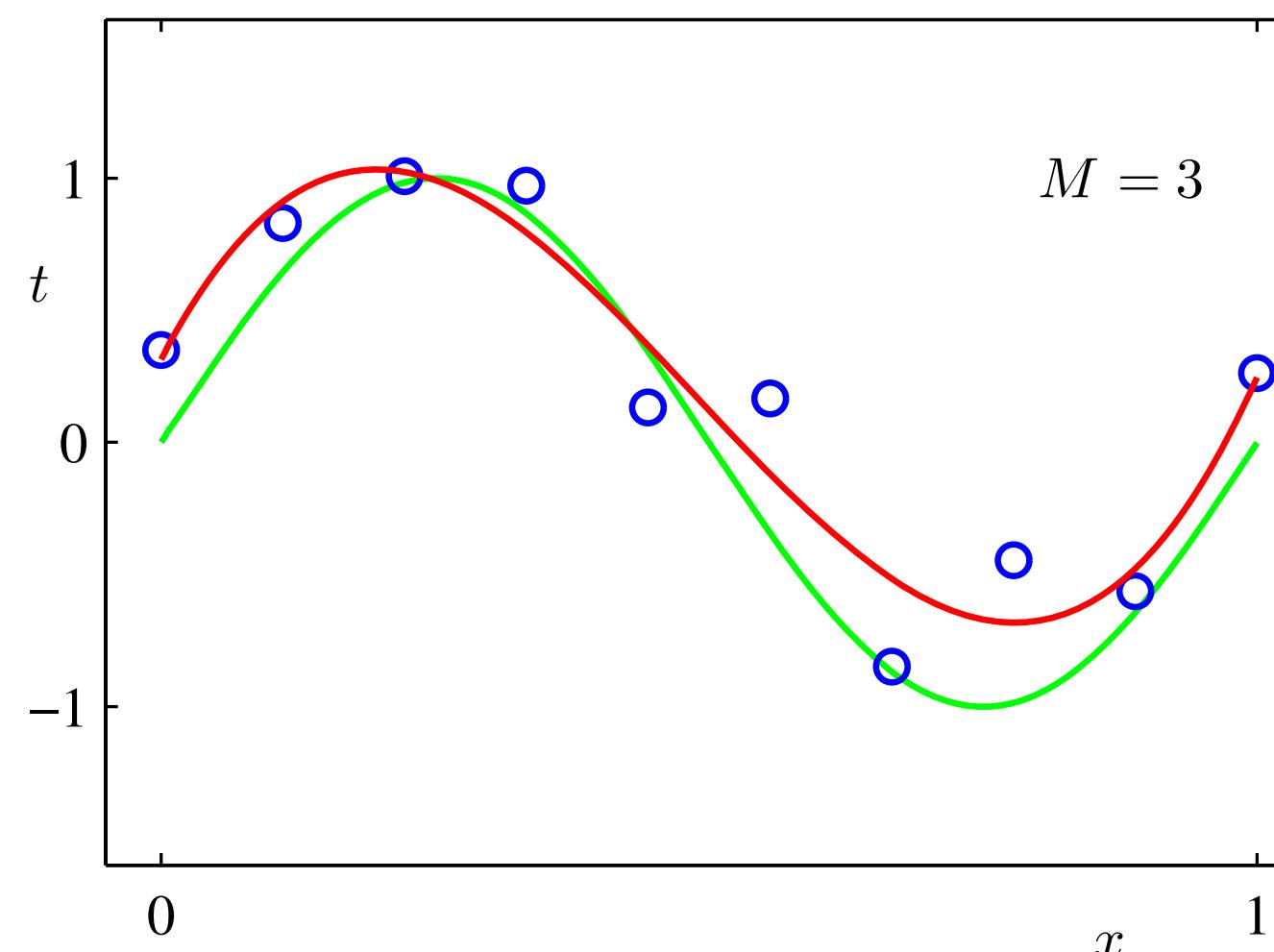
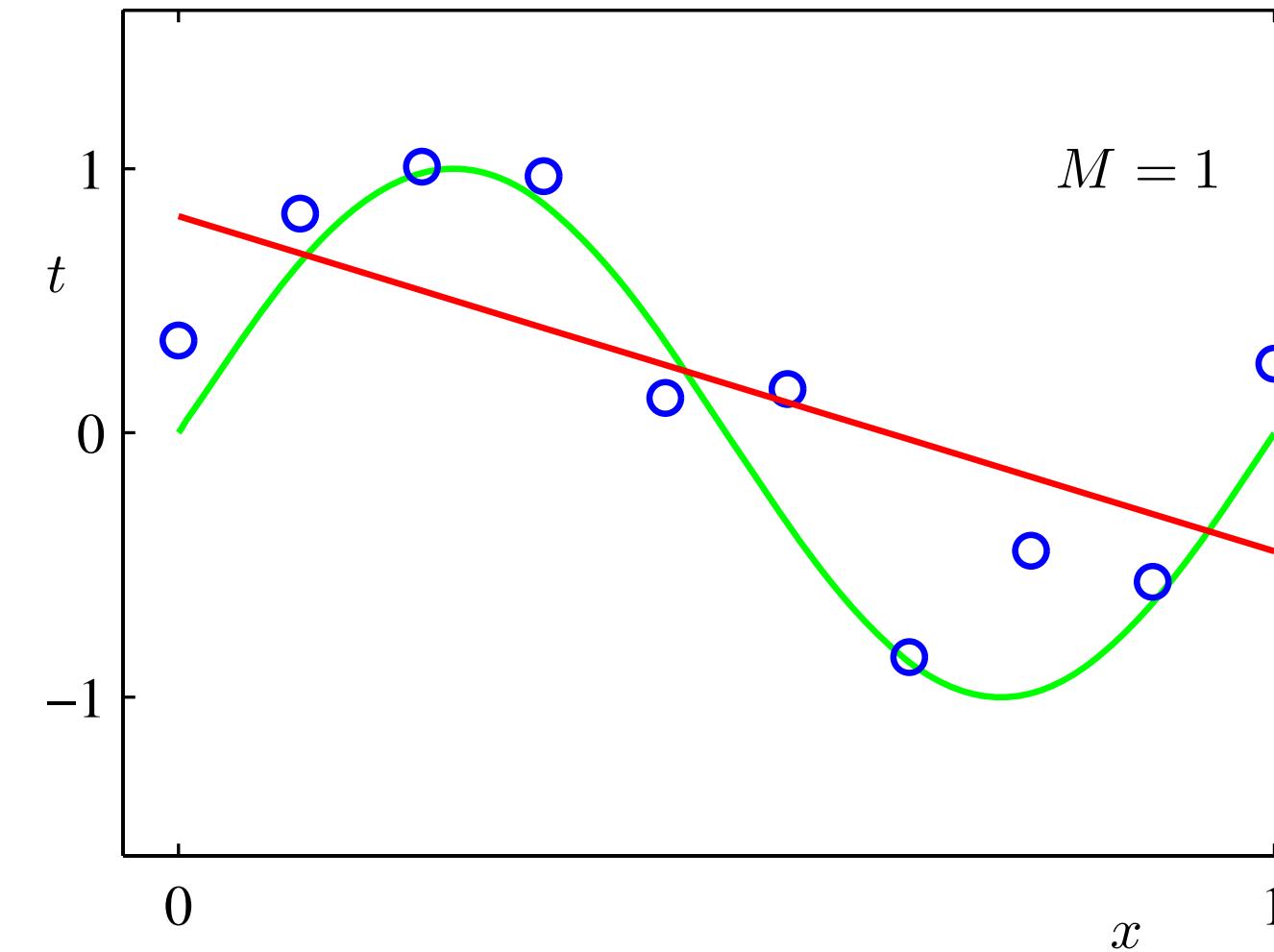
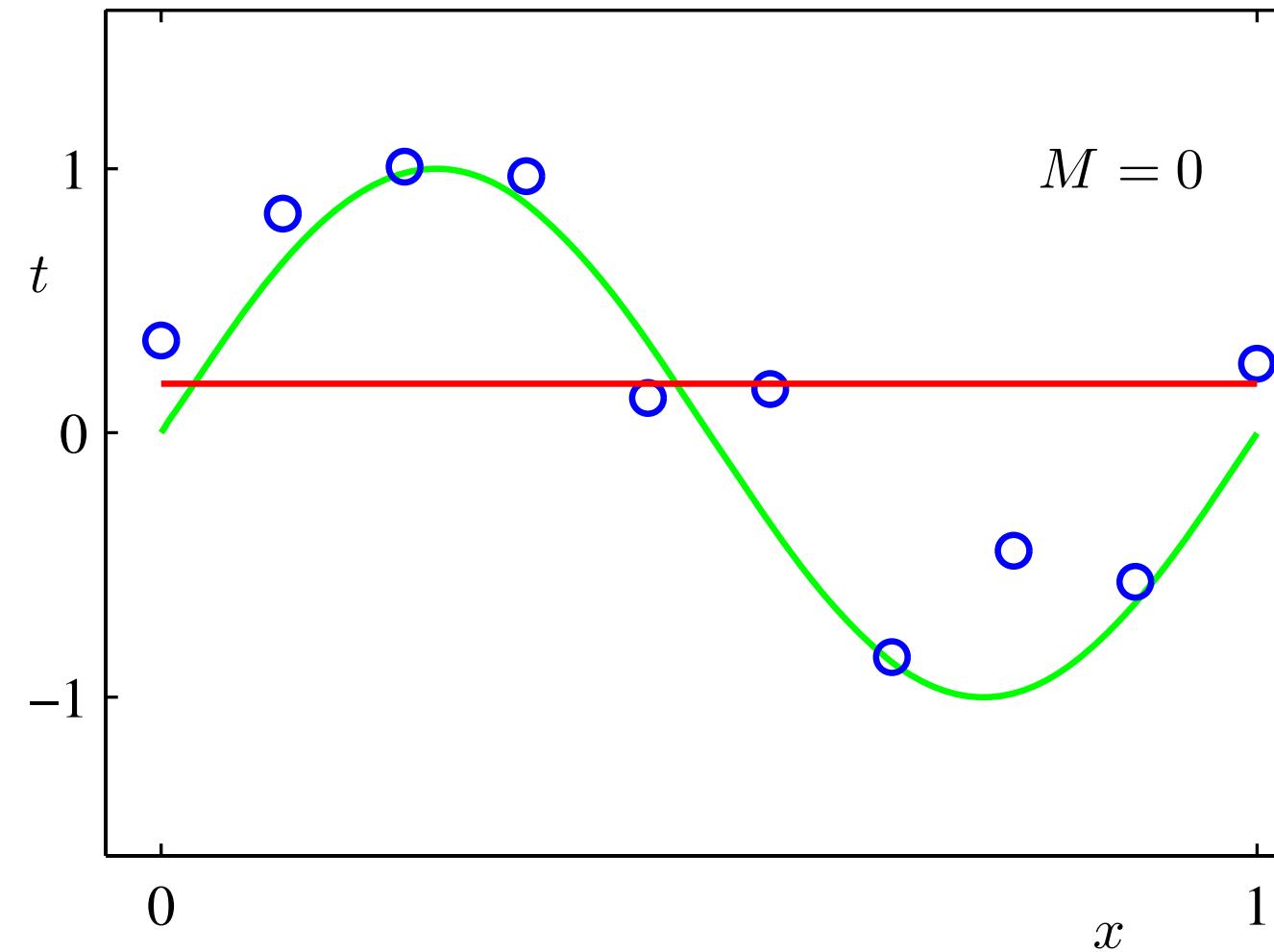
# Bit of a problem with over-training

## MNIST-1D

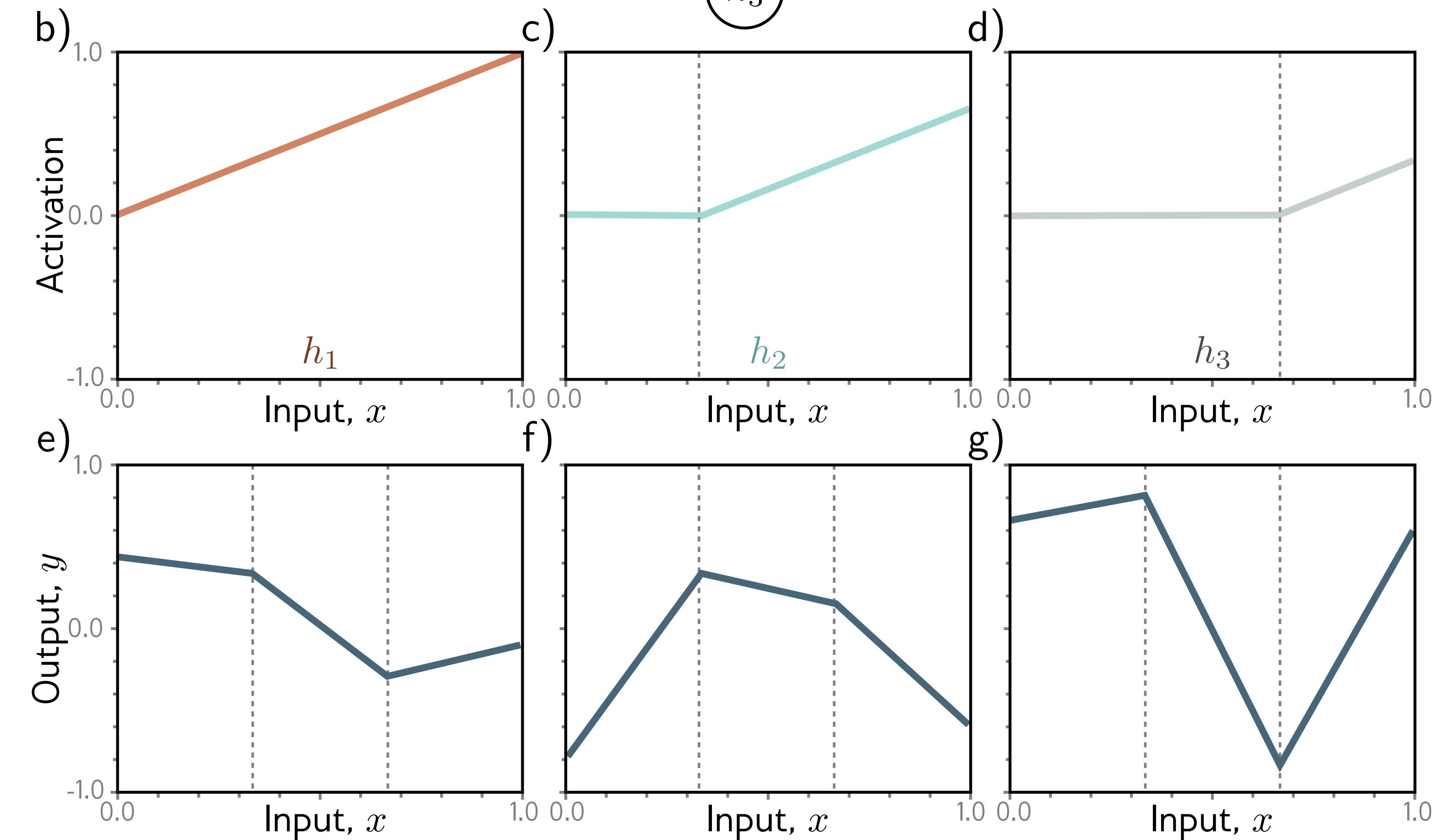
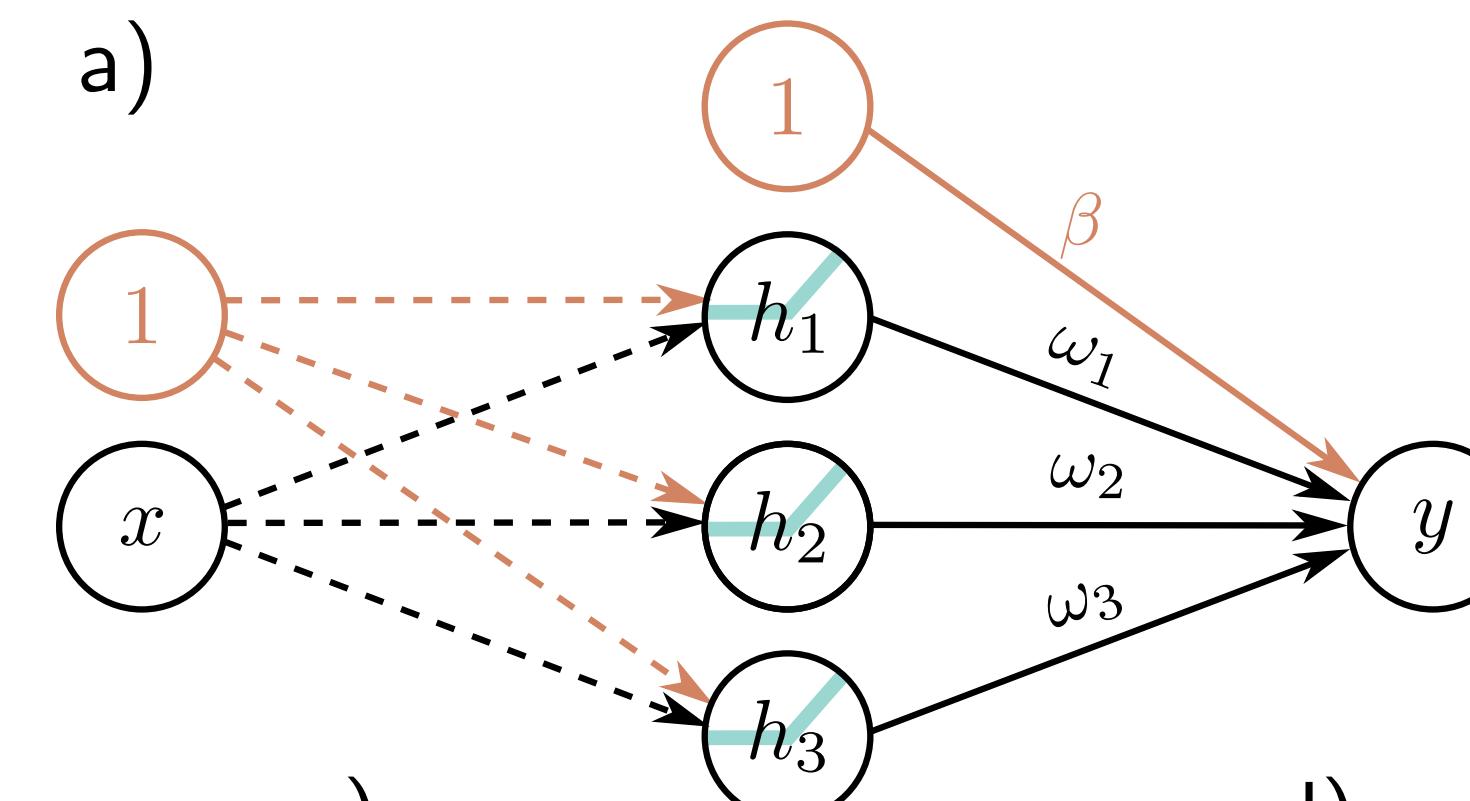
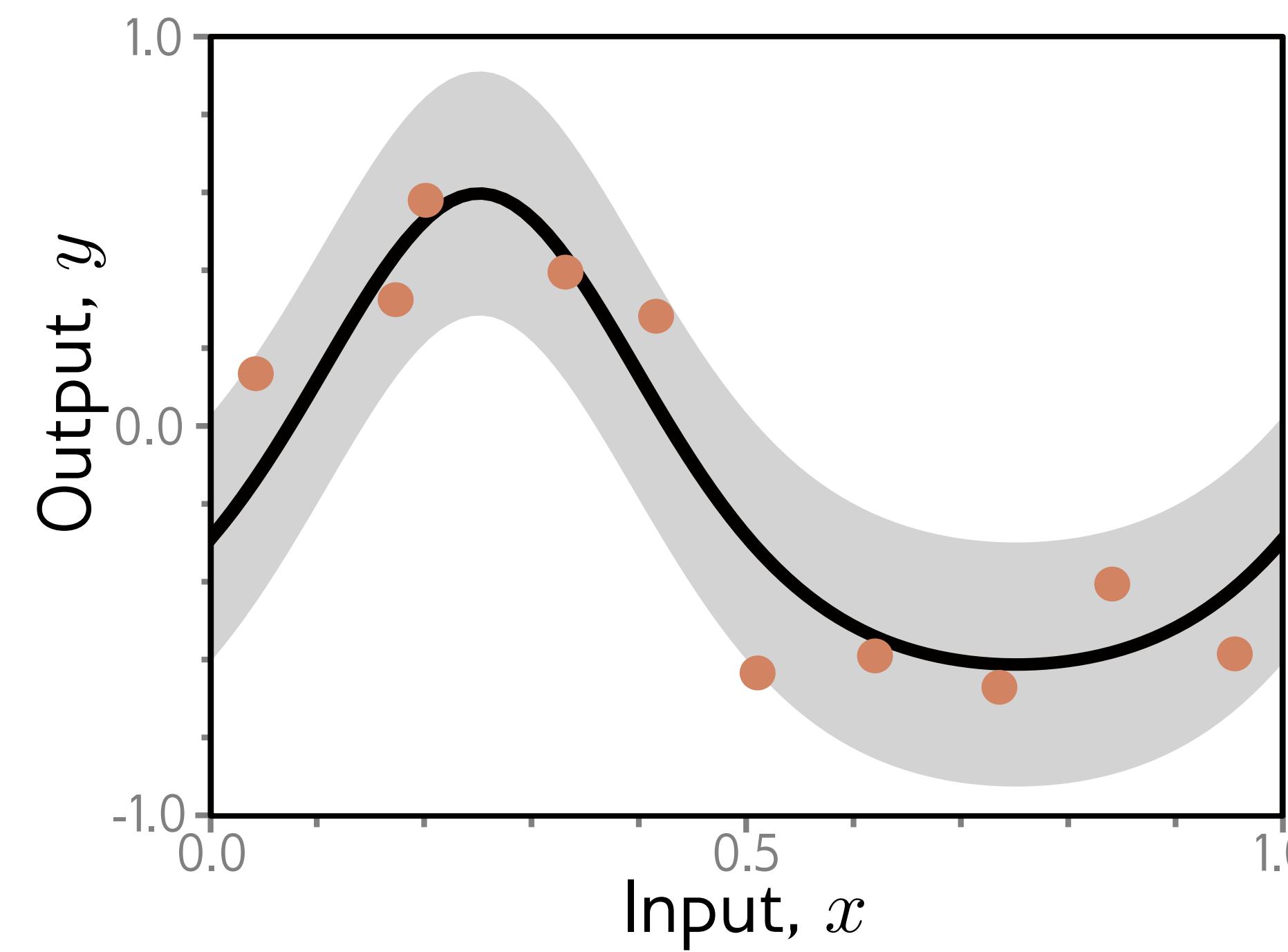


# Overfitting

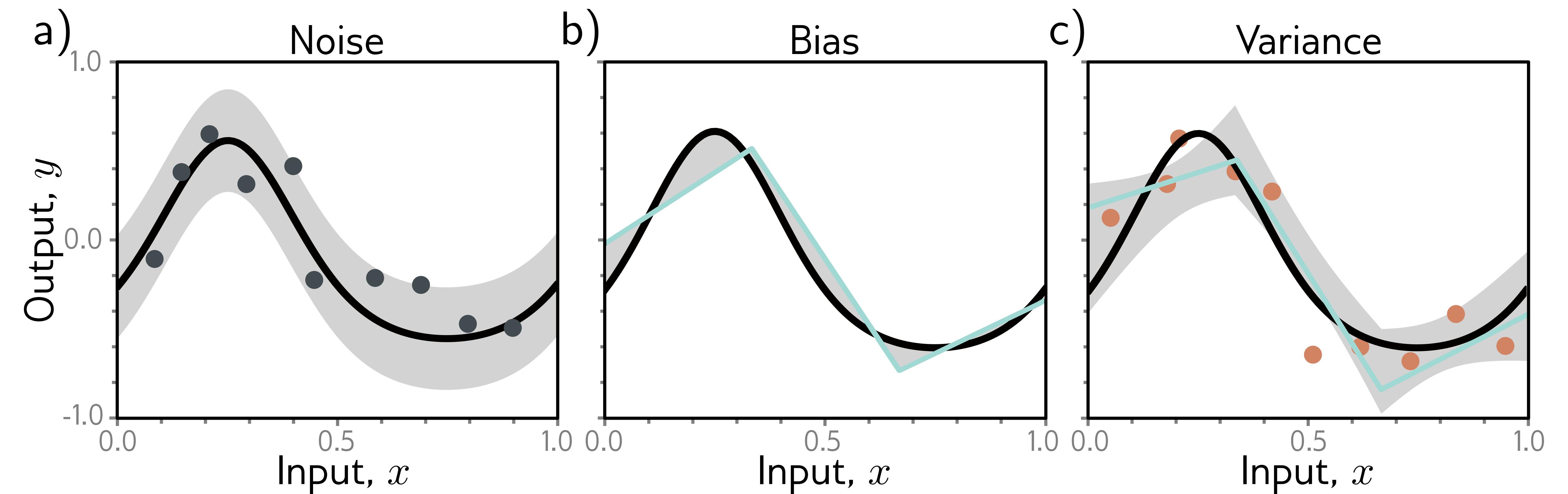
When too many parameters are barely enough



# A toy example

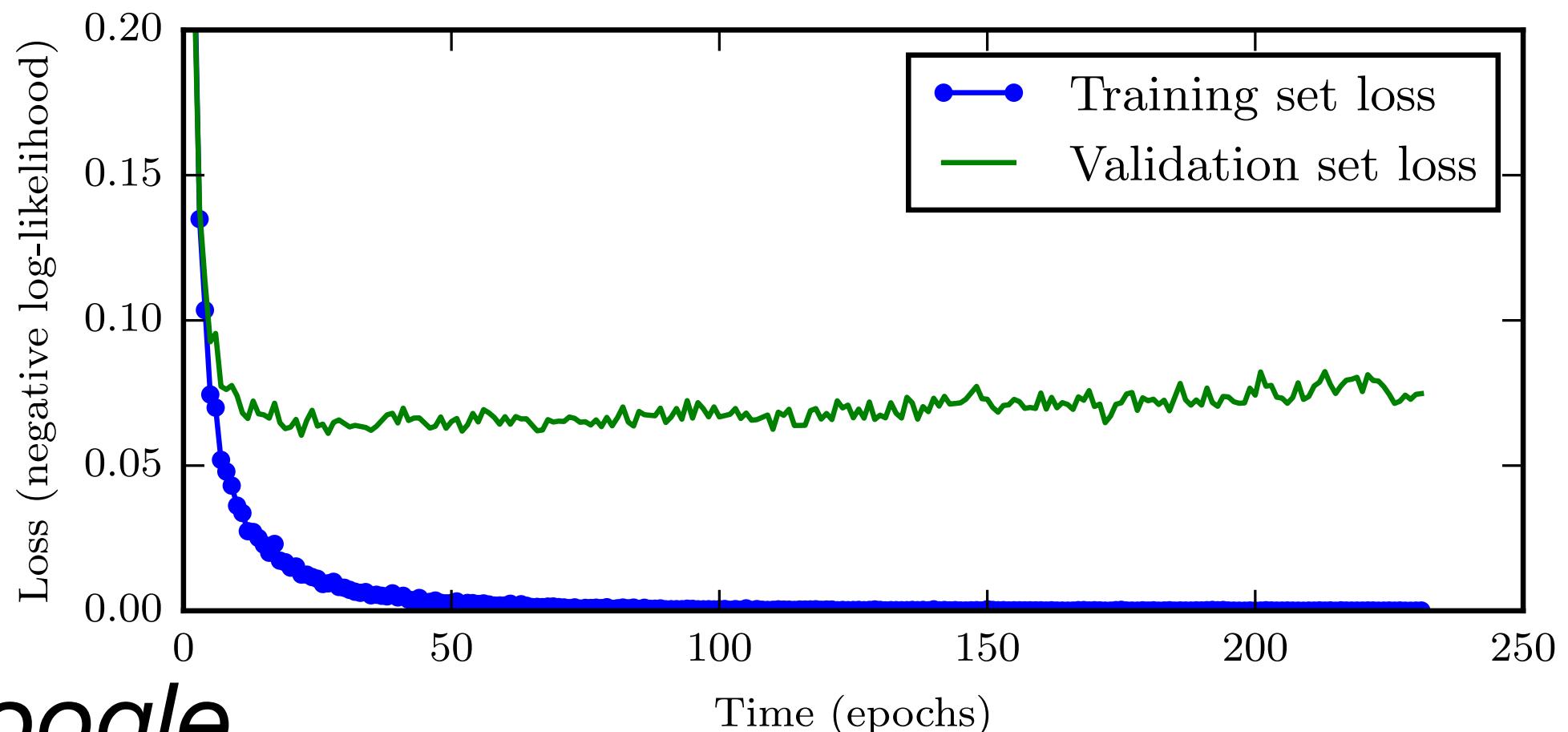


# Three sources of error

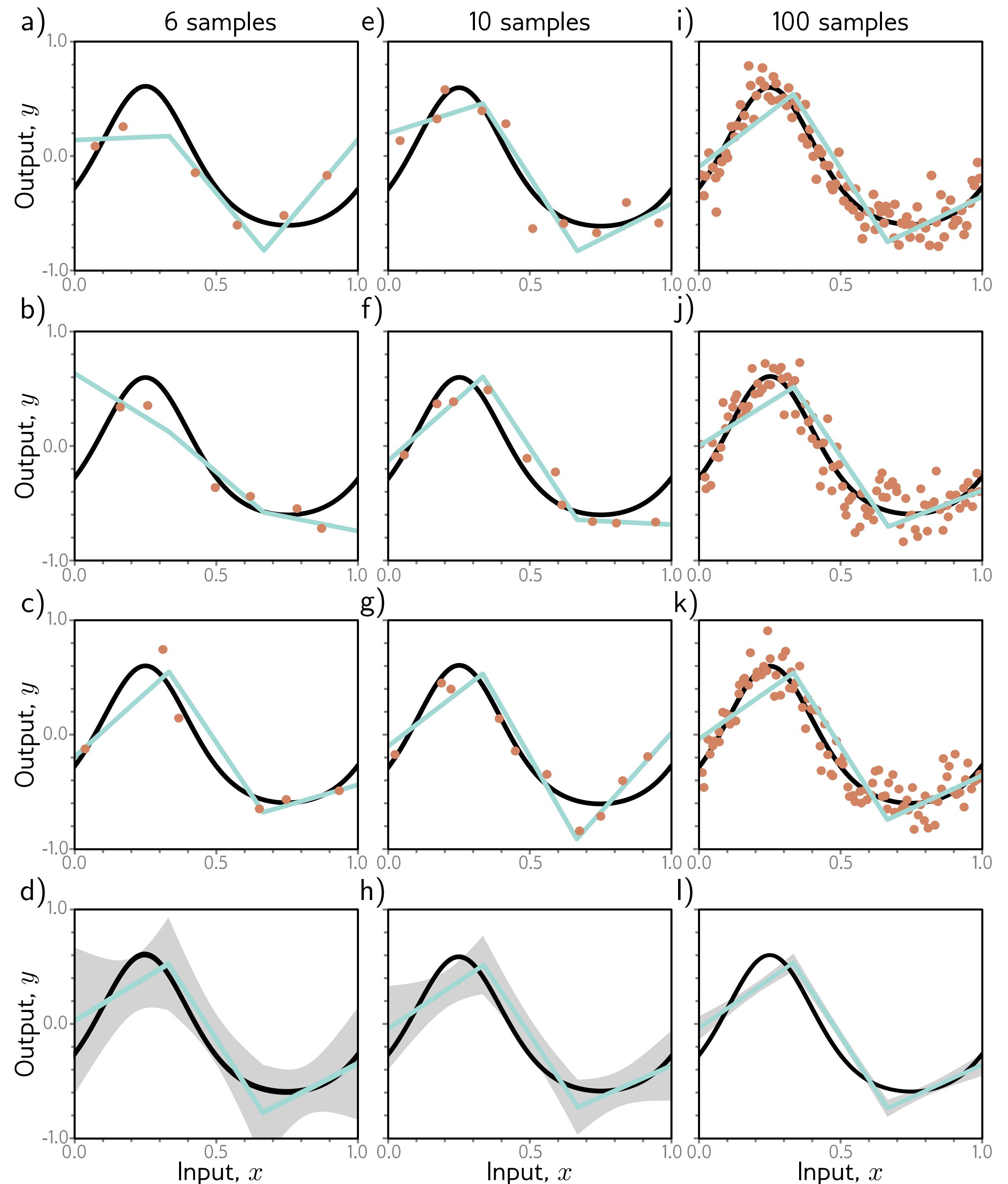


# Reducing error And regularisation

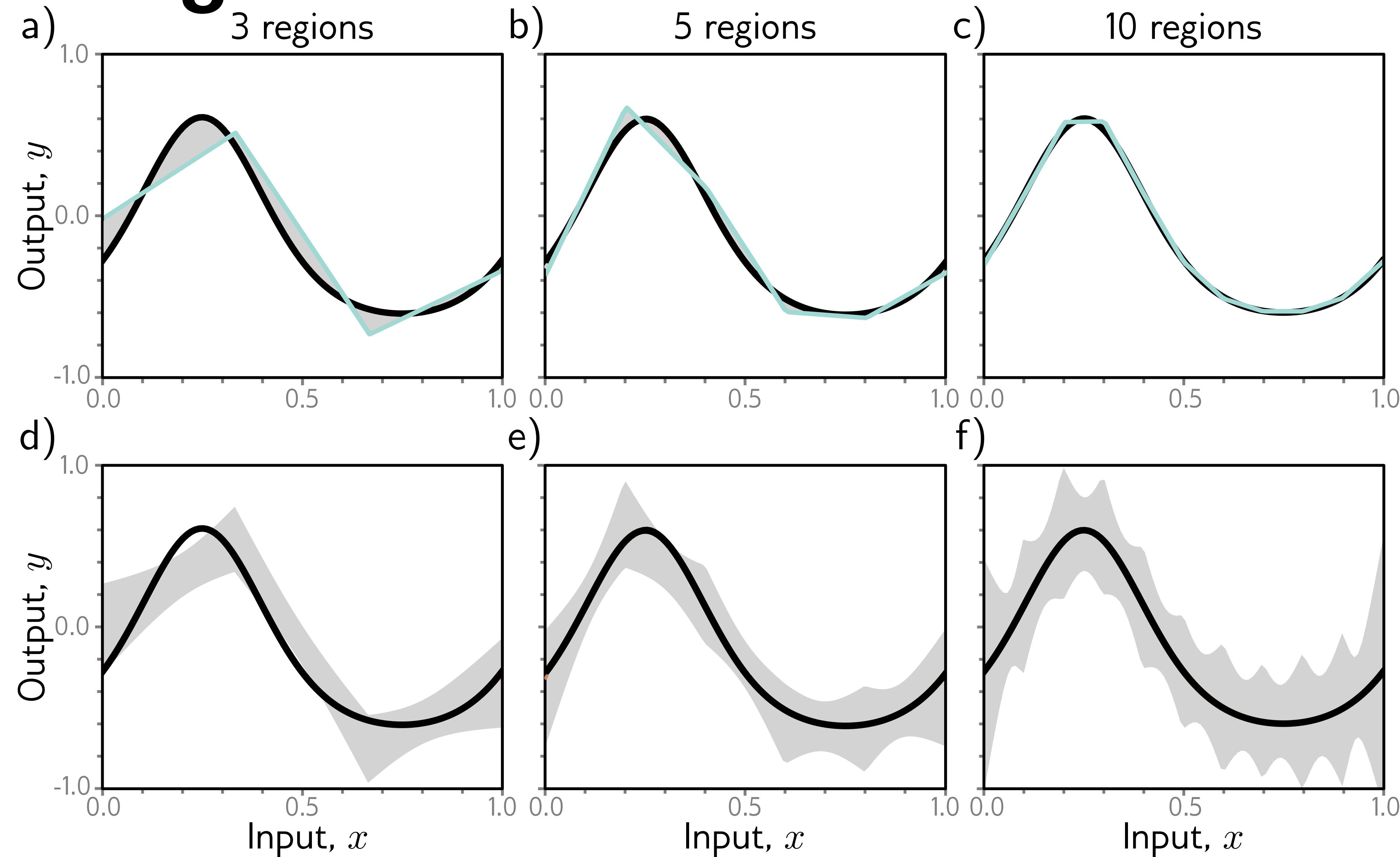
- **Variance:** “Just get more data” – *Guy from Google*
- **Bias:** “We need to go deeper”
- Early stopping, i.e., “cheating”
- ... (bagging, multi-task learning, ...)
- Sparse representations! (cf. compressive sensing...)
- Dropout



# Reducing variance

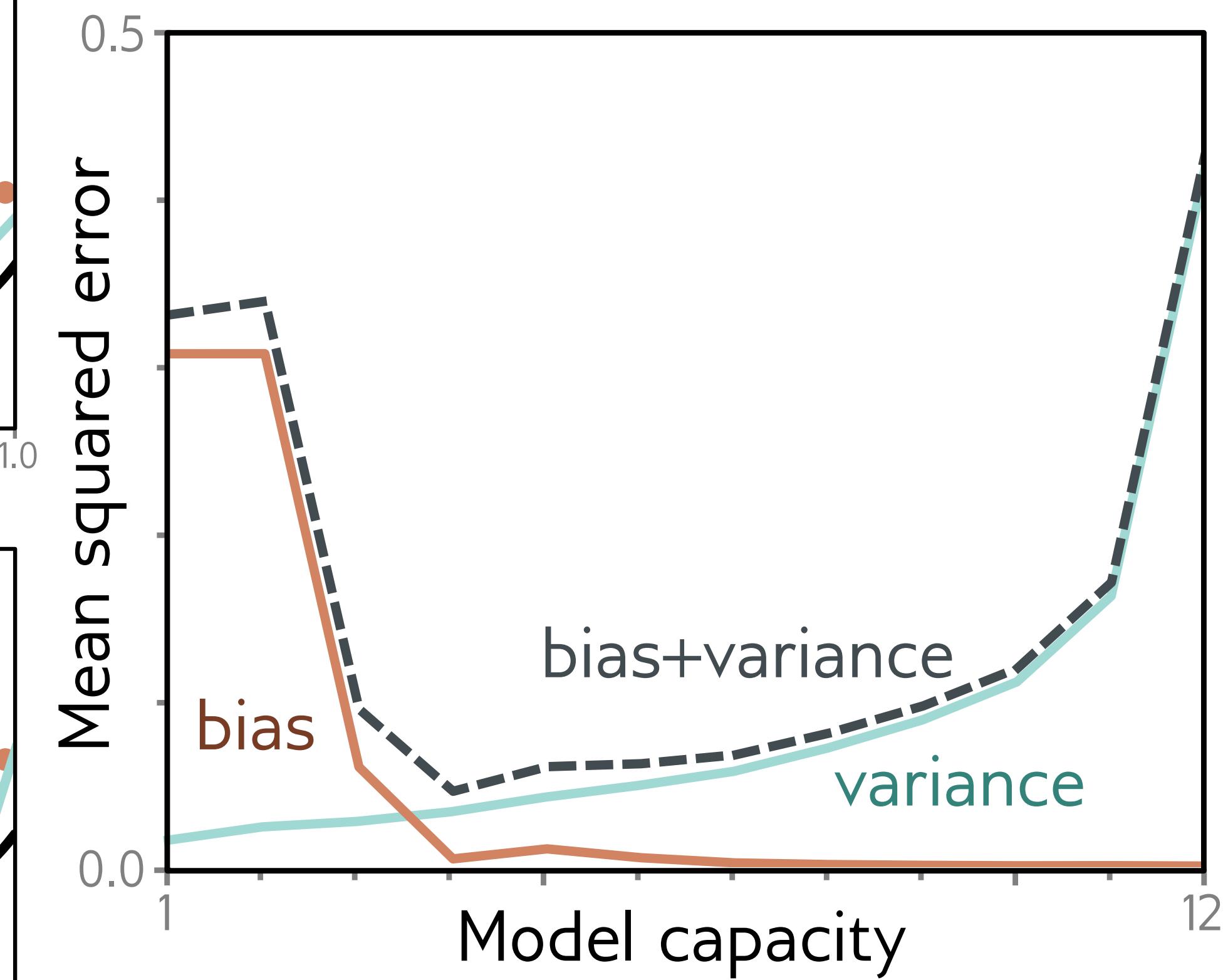
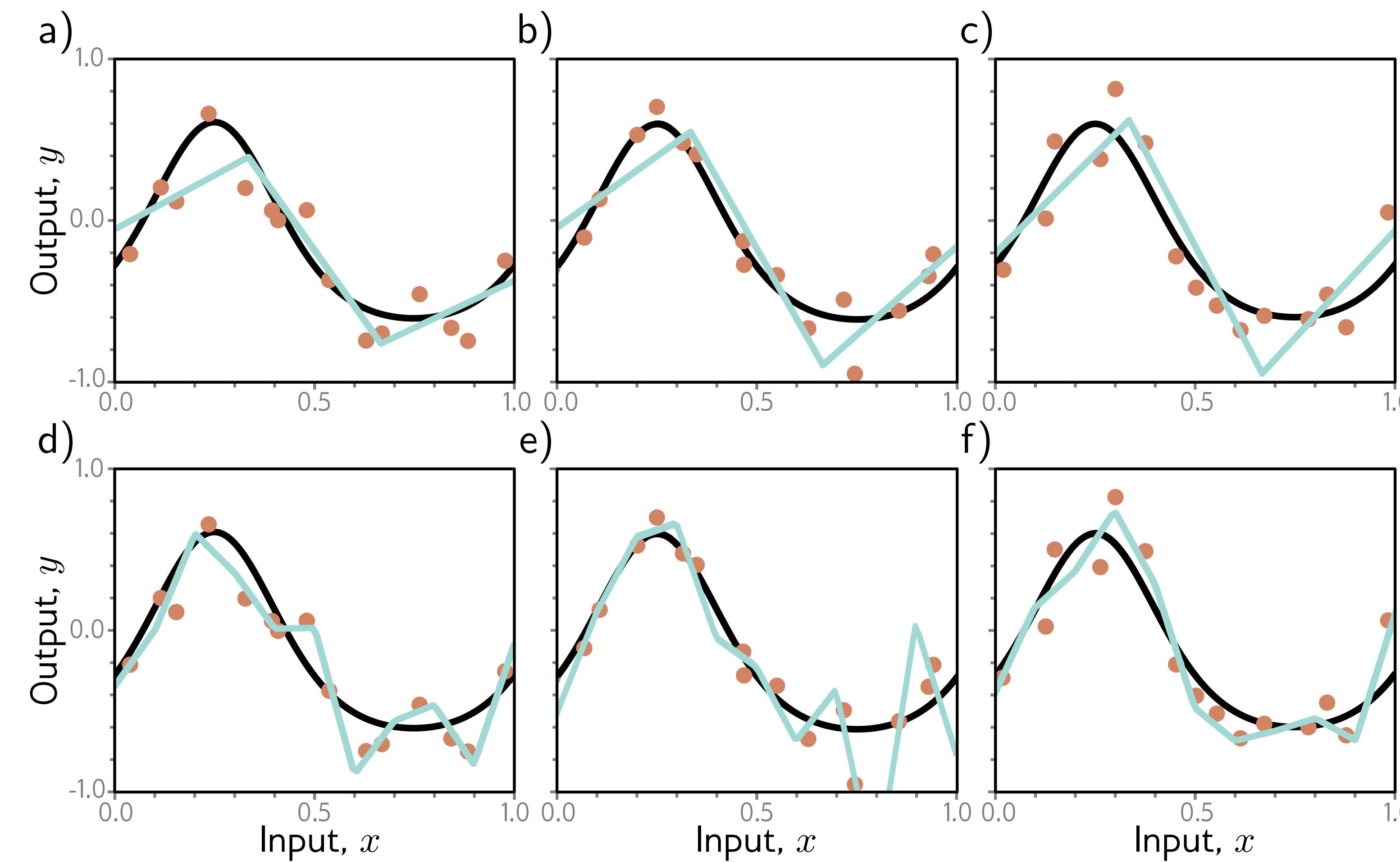


# Reducing bias



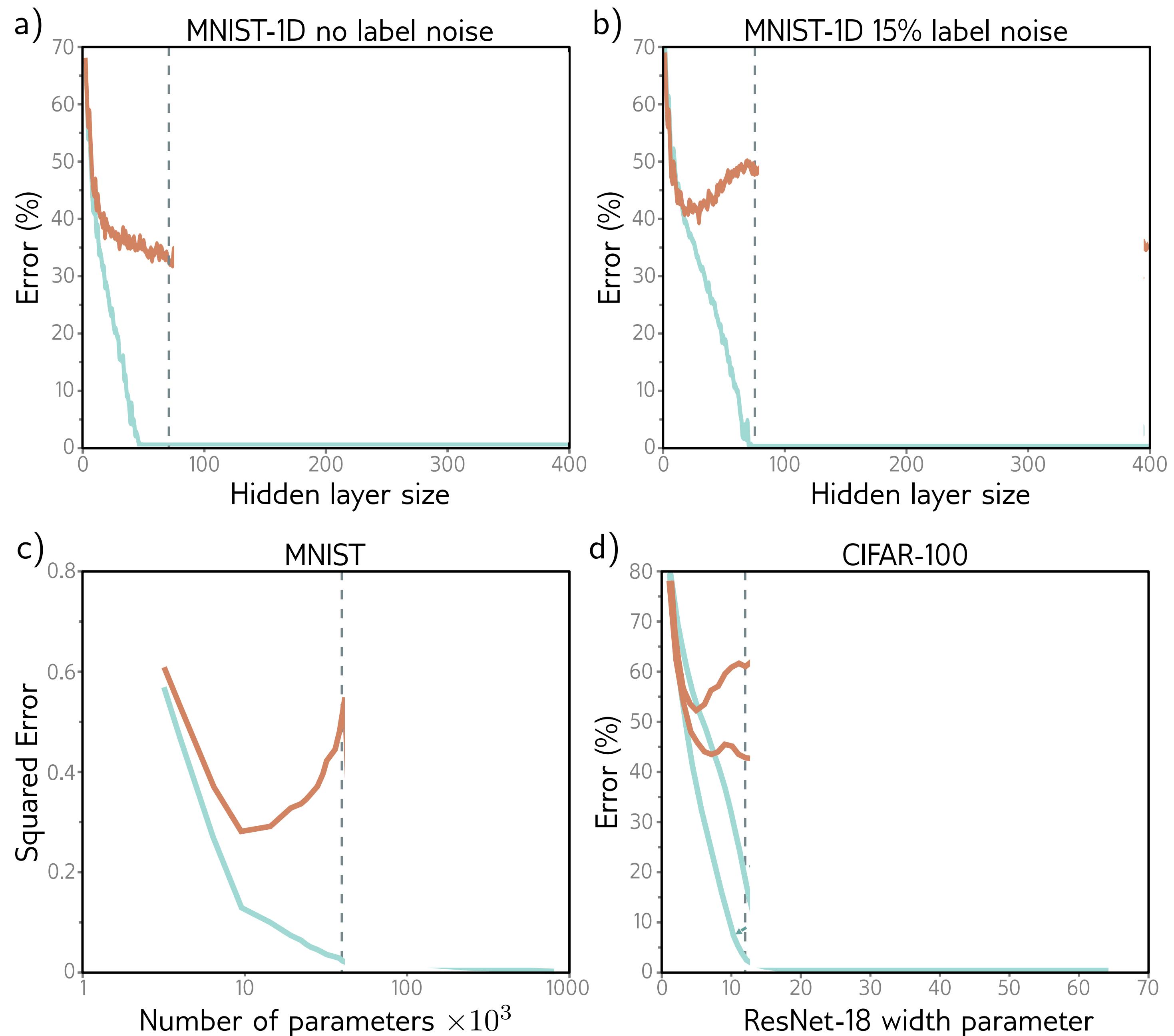
# Overfitting

## Bias-variance tradeoff



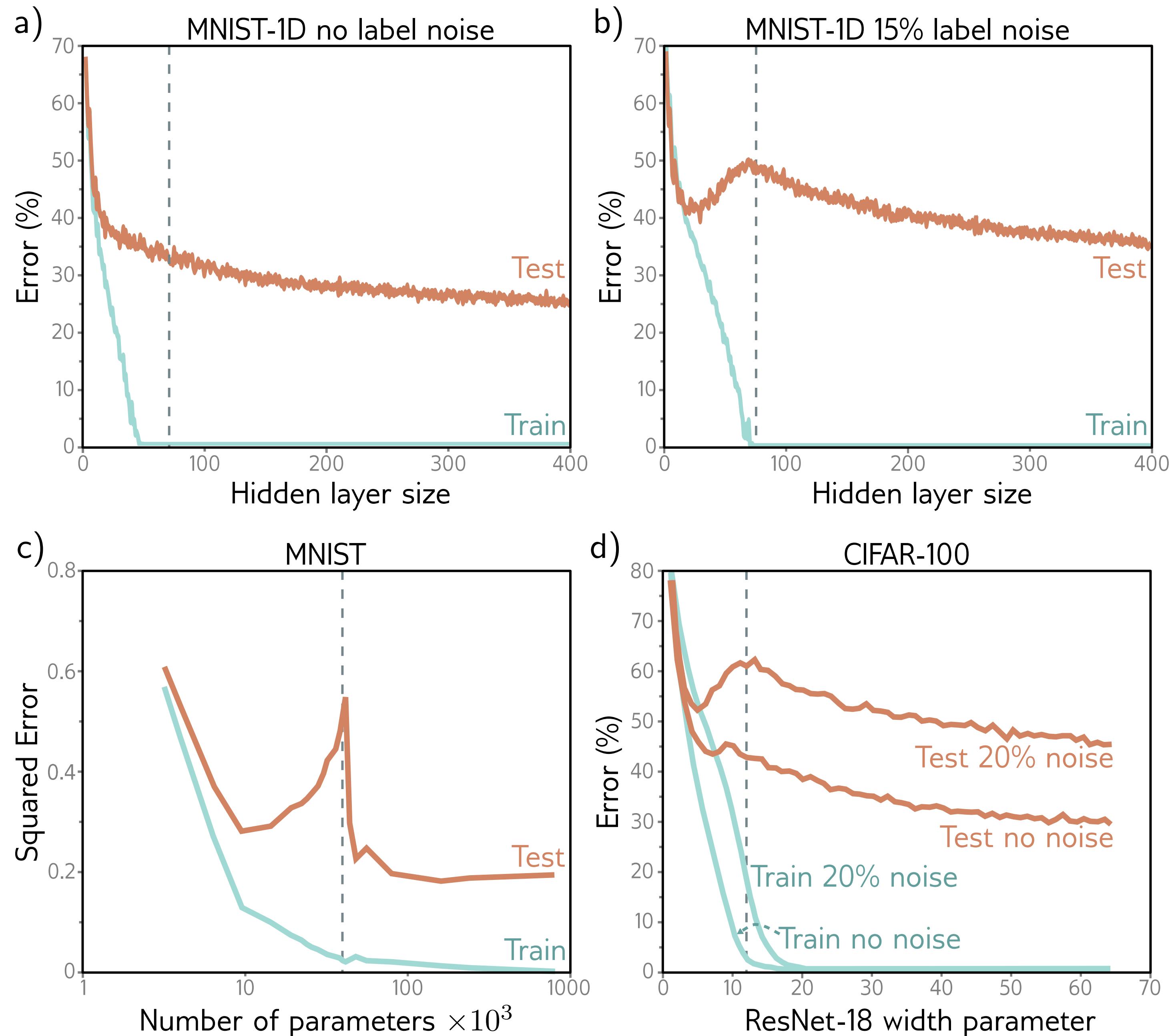
# Double descent

Test  
Train



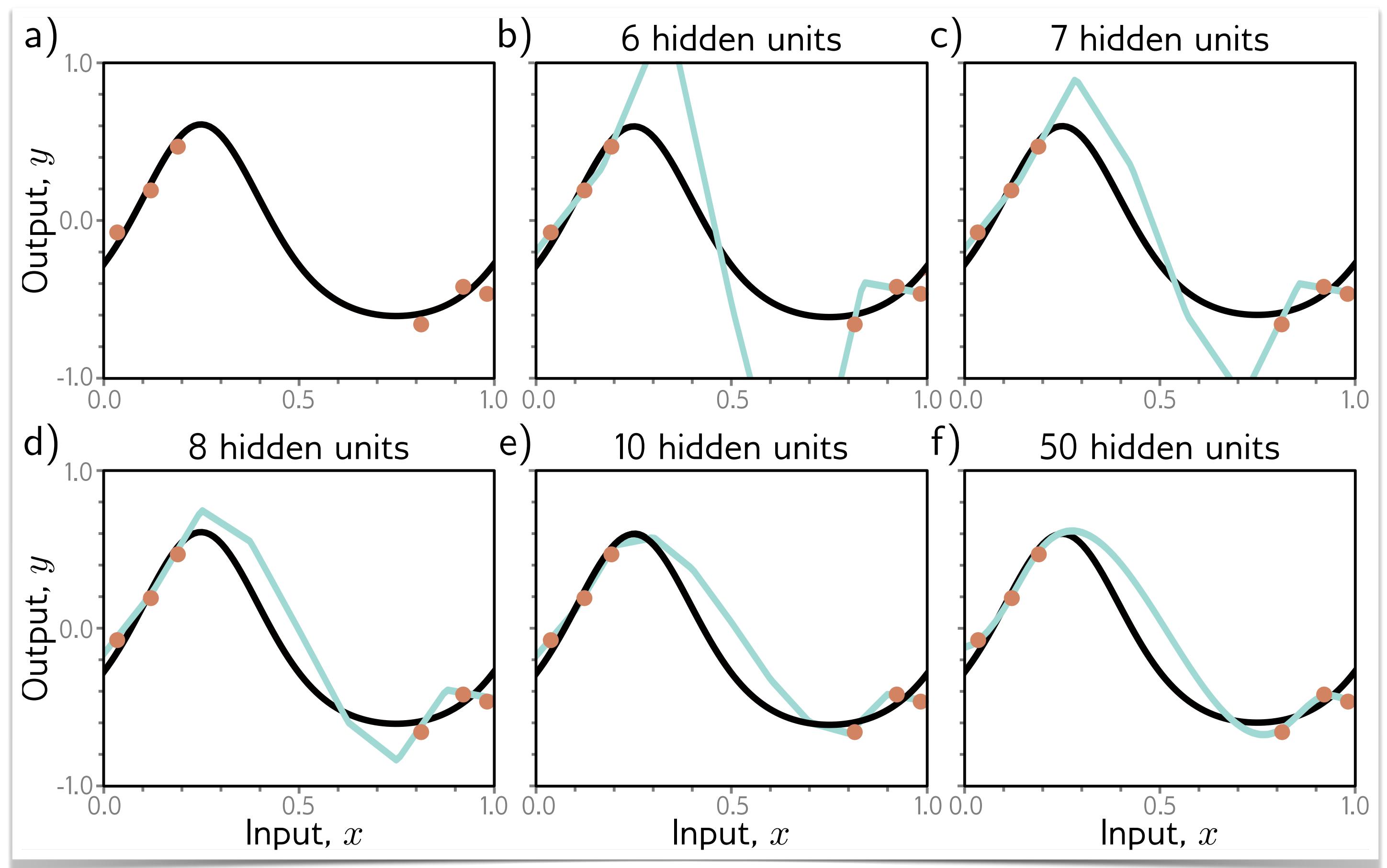
# Double descent

Test  
Train

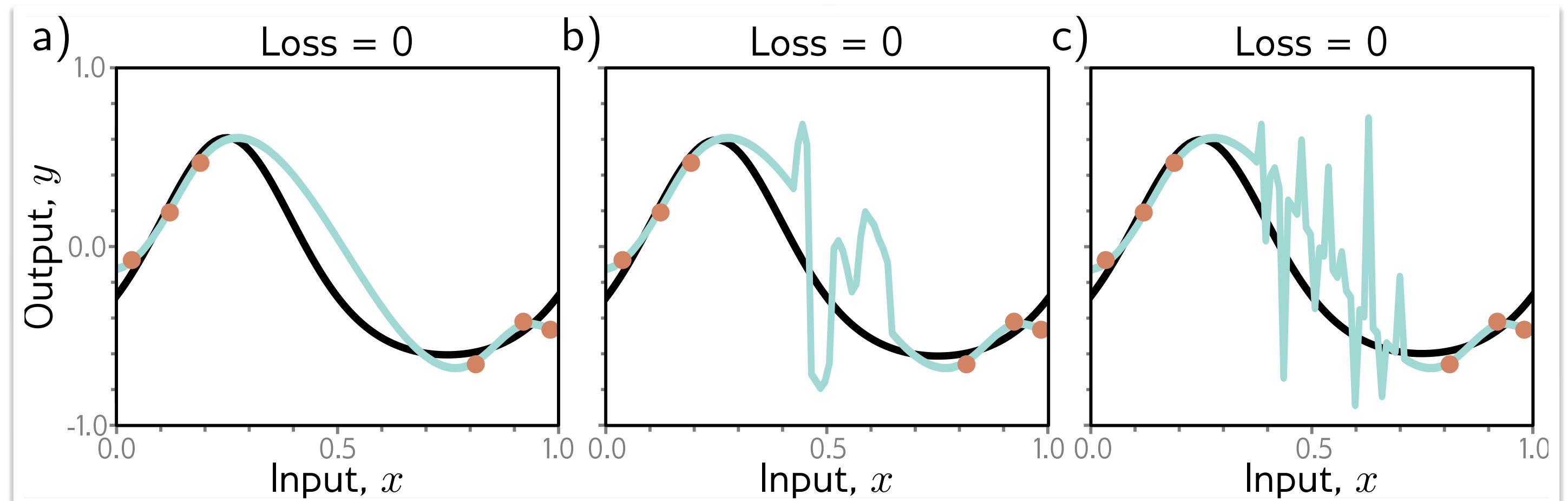


# Why?

## Capacity == smoothness



## Regularisation == smoothness



# Regularisation

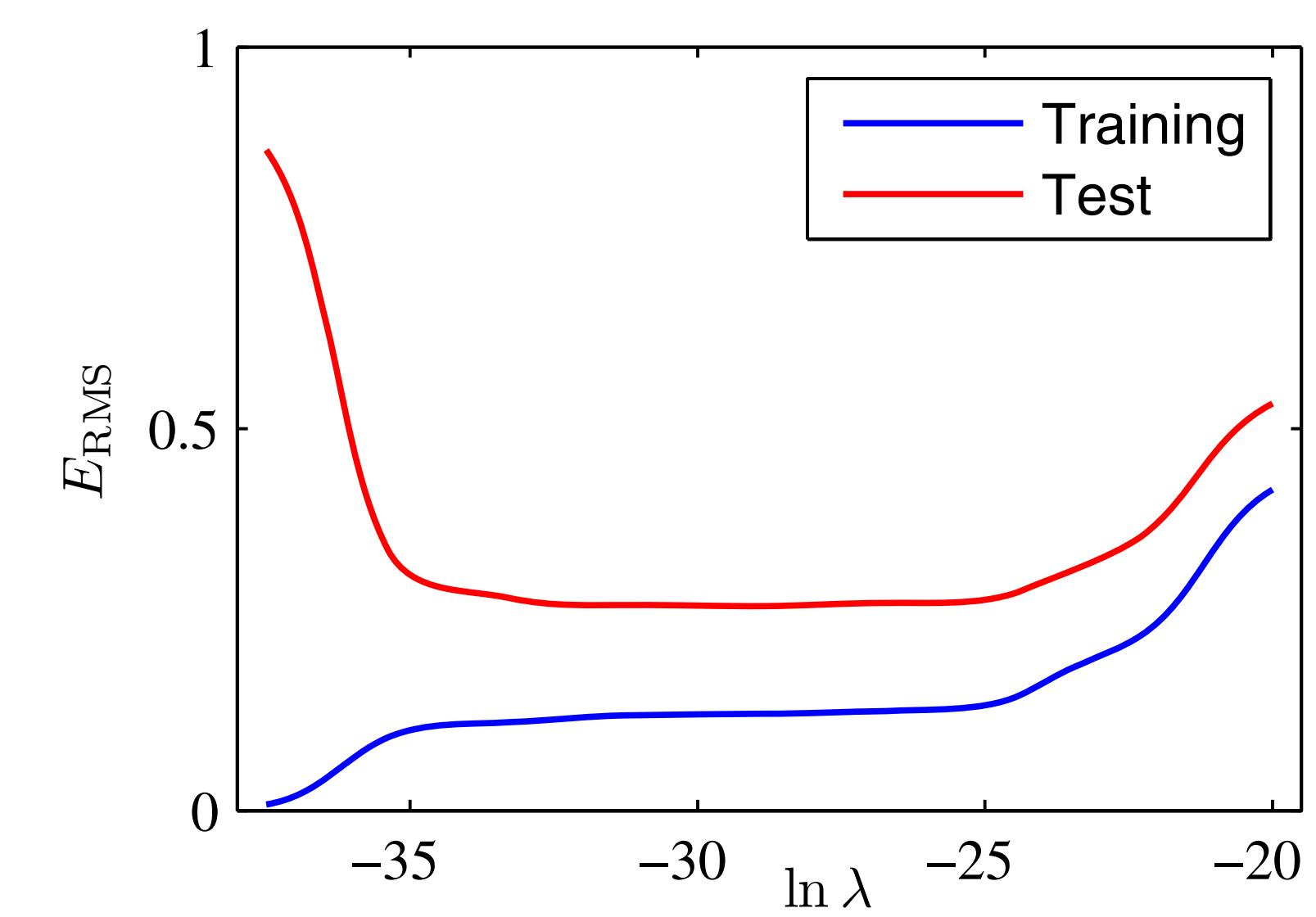
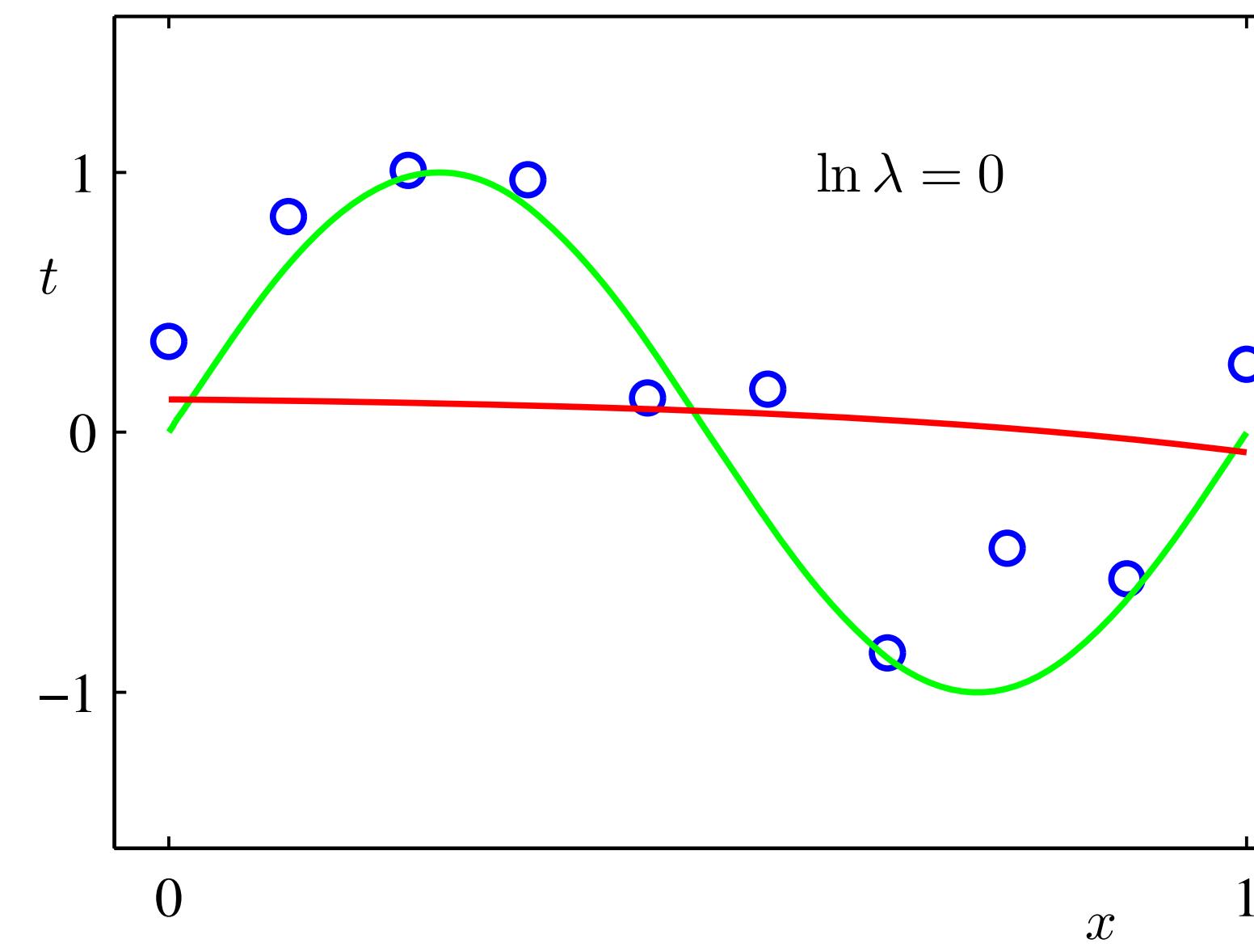
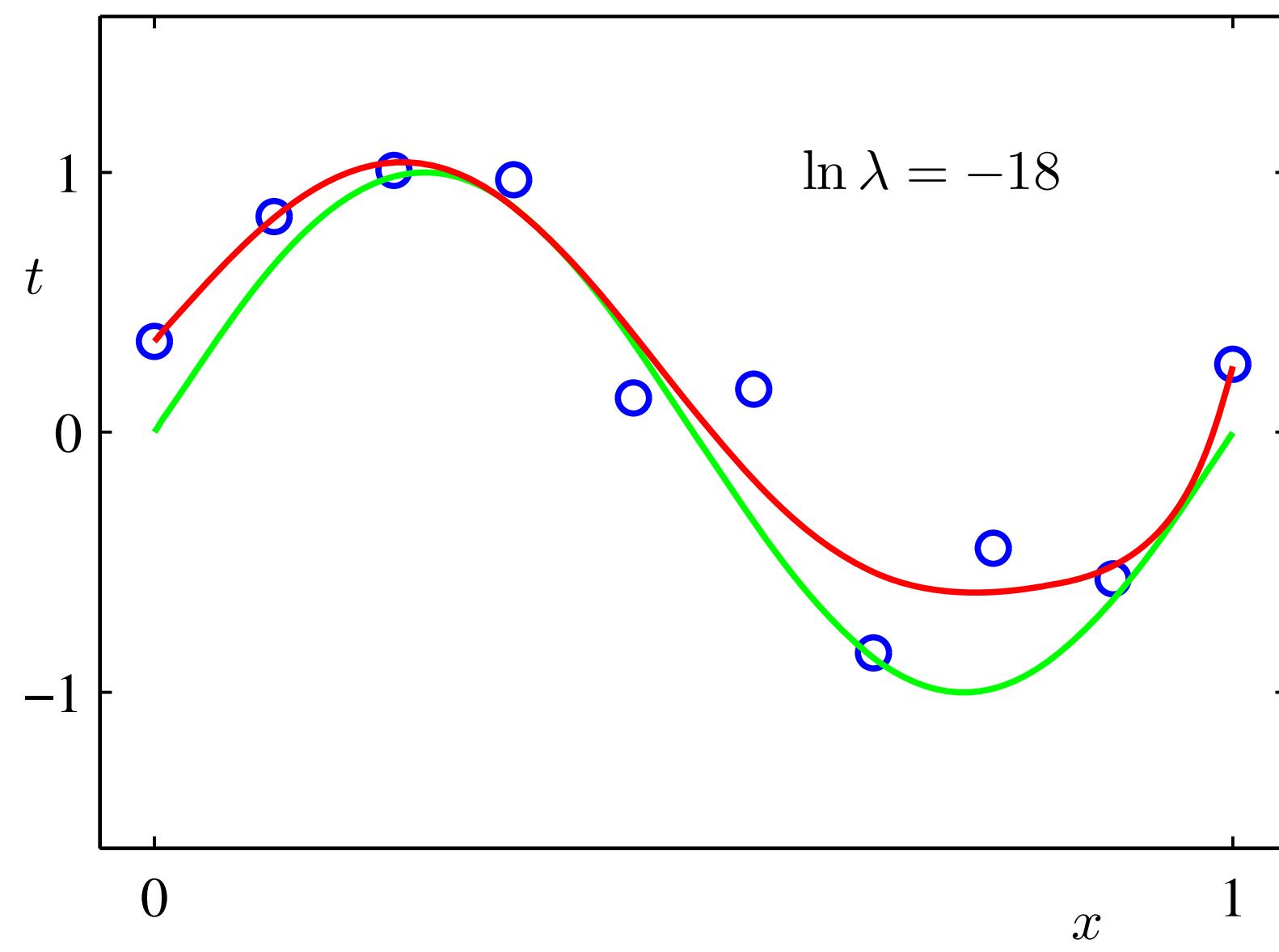
# Regularisation

- In traditional statistics: “explicit terms in the loss function to favour certain parameter choices” (e.g., L1, L2, ridge regression, LASSO, ...)
- In deep learning: “any strategy that improves generalisation”

# Regularisation

$$\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

(ridge regression)



# Regularisation

## In general

L1 (LASSO):  $\tilde{J}(\mathbf{w}; \mathbf{X}, \mathbf{y}) = \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w} + J(\mathbf{w}; \mathbf{X}, \mathbf{y}),$   $w_i = \text{sign}(w_i^*) \max \left\{ |w_i^*| - \frac{\alpha}{H_{i,i}}, 0 \right\}.$

L2 (ridge):  $\tilde{J}(\mathbf{w}; \mathbf{X}, \mathbf{y}) = \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w} + J(\mathbf{w}; \mathbf{X}, \mathbf{y}),$  (note bias-variance)

# Regularisation

## In general

L1 (LASSO):  $\tilde{J}(\mathbf{w}; \mathbf{X}, \mathbf{y}) = \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w} + J(\mathbf{w}; \mathbf{X}, \mathbf{y}), \quad w_i = \text{sign}(w_i^*) \max \left\{ |w_i^*| - \frac{\alpha}{H_{i,i}}, 0 \right\}.$

L2 (ridge):  $\tilde{J}(\mathbf{w}; \mathbf{X}, \mathbf{y}) = \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w} + J(\mathbf{w}; \mathbf{X}, \mathbf{y}),$  (note bias-variance)

This is constrained  
optimisation!  
(i.e., KKT)

$$\min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} \max_{\boldsymbol{\alpha}, \boldsymbol{\alpha} \geq 0} -f(\mathbf{x}) + \sum_i \lambda_i g^{(i)}(\mathbf{x}) + \sum_j \alpha_j h^{(j)}(\mathbf{x}).$$

# Weight Decay as Constrained Optimization

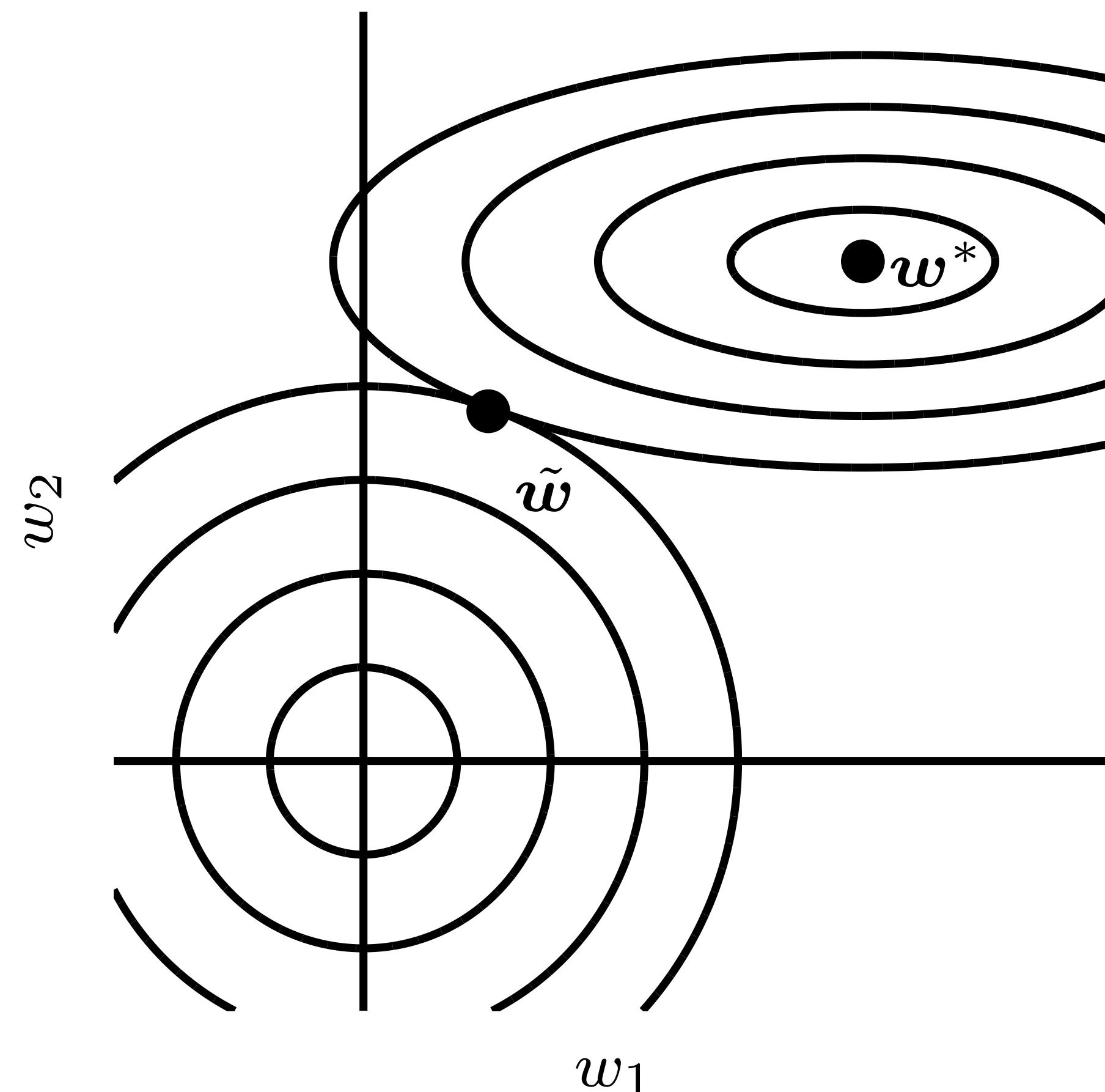


Figure 7.1

(Goodfellow 2016)

the problem is not convex (but some other regularity conditions hold), then these KKT conditions still characterise a local minima.

**Theorem 4.12** (KKT: Slater's condition version). *Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be a convex, continuously differentiable function on the set  $\Omega = \{\mathbf{x} \in \mathbb{R}^n : g_i(\mathbf{x}) \leq 0, \text{ for all } i = 1, \dots, m\}$ , where the  $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$  are convex and continuously differentiable for all  $i = 1, \dots, m$ , such that  $\text{int } \Omega \neq \emptyset$ . A necessary and sufficient condition that  $\mathbf{x}^* \in \Omega$  minimises  $f$  over  $\Omega$  is that there exists a solution to the system*

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i \nabla g_i(\mathbf{x}^*) = \mathbf{0}, \quad (4.7a)$$

$$\lambda_i g_i(\mathbf{x}^*) = 0 \text{ for all } i = 1, 2, \dots, m, \quad (4.7b)$$

$$\mathbf{g}(\mathbf{x}^*) \leq \mathbf{0}, \quad (4.7c)$$

$$\boldsymbol{\lambda} \geq \mathbf{0}. \quad (4.7d)$$

Section 4.2.1 proves this theorem.

The condition (4.7b) is sometimes called a complementary slackness condition because it expresses that either the constraint is active (and the Lagrange multiplier is potentially positive) or the Lagrange multiplier is zero (and potentially the constraint is not active).

The conditions (4.7) are called the **KKT conditions**. Conditions (4.7a) and (4.7b) may be equivalently replaced with

$$\nabla f(\mathbf{x}^*) + \sum_{i \in I(\mathbf{x}^*)} \lambda_i \nabla g_i(\mathbf{x}^*) = \mathbf{0}, \quad (4.8)$$

and for all  $i \notin I(\mathbf{x}^*)$  we set  $\lambda_i = 0$ .

**Theorem 4.12** is an extension of the Lagrange multiplier idea. That is, for the active constraints  $I(\mathbf{x}^*)$ <sup>5</sup> we would write a new objective function

$$h(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{i \in I(\mathbf{x}^*)} \lambda_i g_i(\mathbf{x}),$$

including Lagrange multipliers as in (4.4) for the active constraints, and then  $\nabla h$  is just the left-hand side of condition (4.8). So we are basically solving a problem including Lagrange multipliers to enforce the appropriate set of constraints (the active ones).

The only additional information we get from Theorem 4.12 is that the Lagrange multipliers  $\lambda_i$  must be positive or zero. Figures 4.15 and 4.16 illustrate the positivity requirements: the constraints  $I(\mathbf{x}^*)$  are active, so the point  $\mathbf{x}^*$  is in the intersection of the curves  $g_i(\mathbf{x}) = 0$ , for  $i \in I(\mathbf{x}^*)$ . We know that the feasible region is inside the set  $\Omega$ , that is that  $g_i(\mathbf{x}) \leq 0$ , and also that  $g_i(\mathbf{x})$  is increasing in the direction  $\nabla g_i(\mathbf{x})$ . Since  $g_i(\mathbf{x}^*) = 0$ , then  $\nabla g_i(\mathbf{x}^*)$  points outside the set  $\Omega$ , equivalently  $-\nabla g_i(\mathbf{x}^*)$  points inside. Likewise, if  $\mathbf{x}^*$  is a minimum, the objective function must increase in any direction pointing into the region. However, what the theorem says is even stronger: it says that

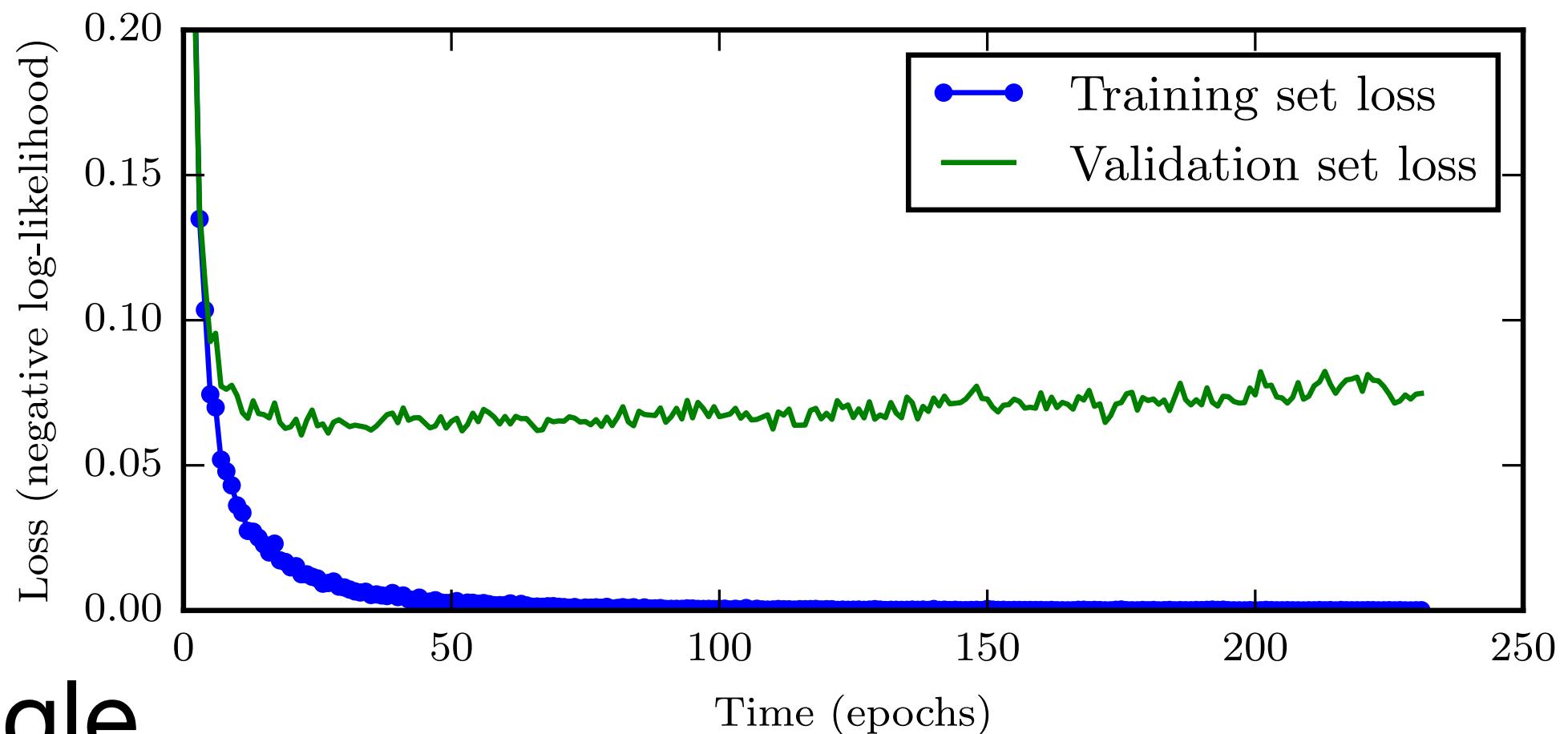
$$\nabla f(\mathbf{x}^*) = - \sum_{i \in I(\mathbf{x}^*)} \lambda_i \nabla g_i(\mathbf{x}^*),$$

---

<sup>5</sup>If  $I(\mathbf{x}^*) = \emptyset$  then  $\mathbf{x}^*$  is not on any boundary, and the KKT condition just reverts to the familiar condition for an unconstrained problem; that is,  $\nabla f(\mathbf{x}^*) = \mathbf{0}$ .

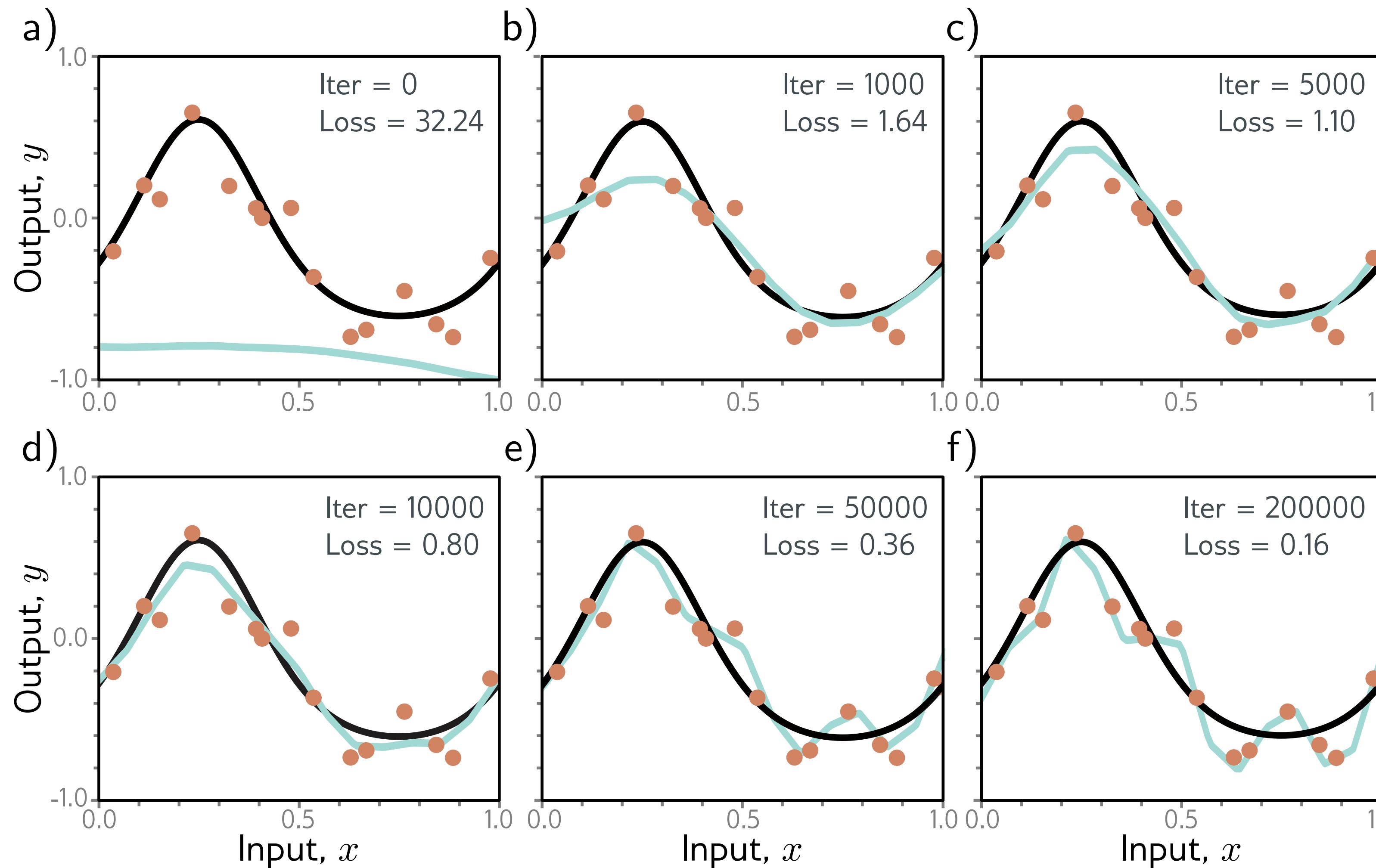
# Reducing error And regularisation

- Variance: “Just get more data” – Guy from Google
- Bias: “We have to go deeper”
- **Early stopping, i.e., “cheating”** (see Goodfellow Fig 7.4)
- ... (**bagging, multi-task learning, ...**)
- **Sparse representations! (cf. compressive sensing...)**
- **Dropout**



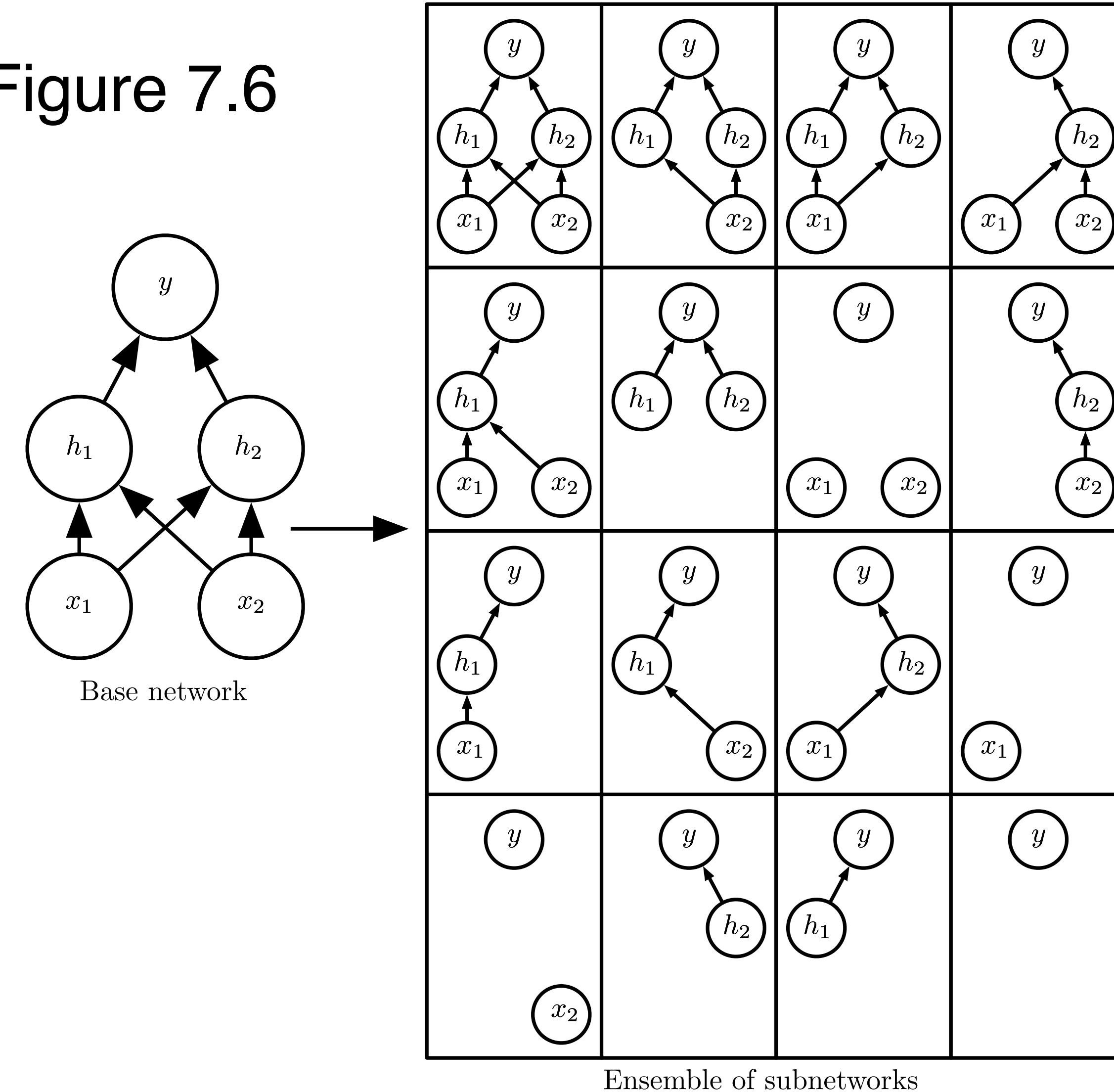
# Early stopping

## On our toy model



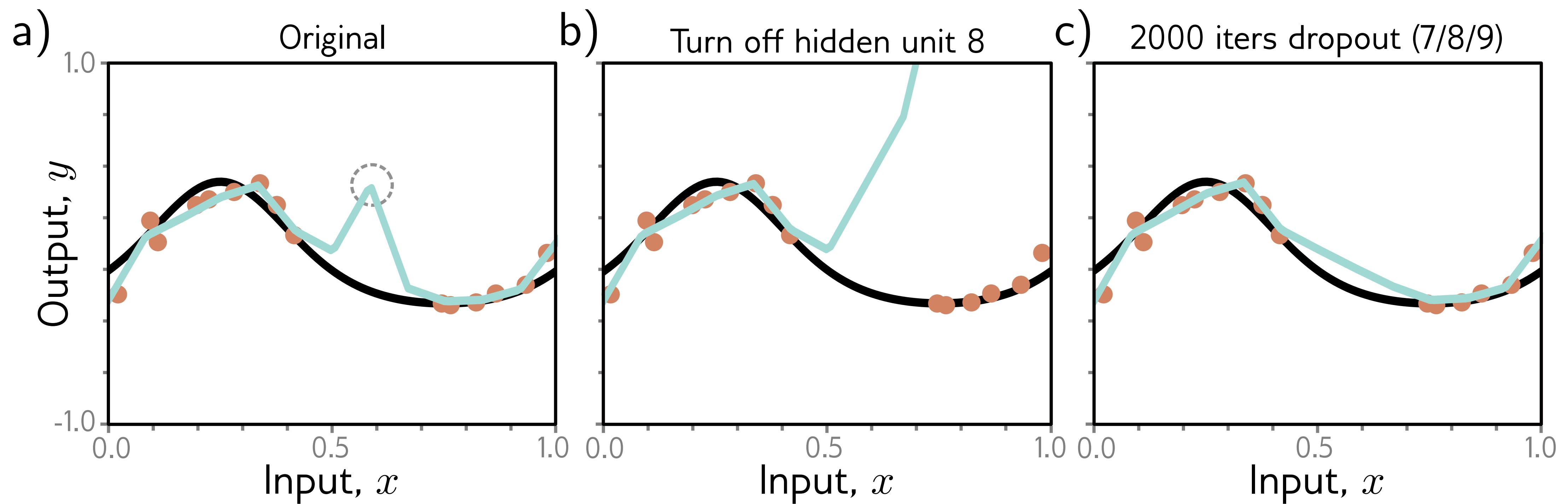
# Dropout

Figure 7.6



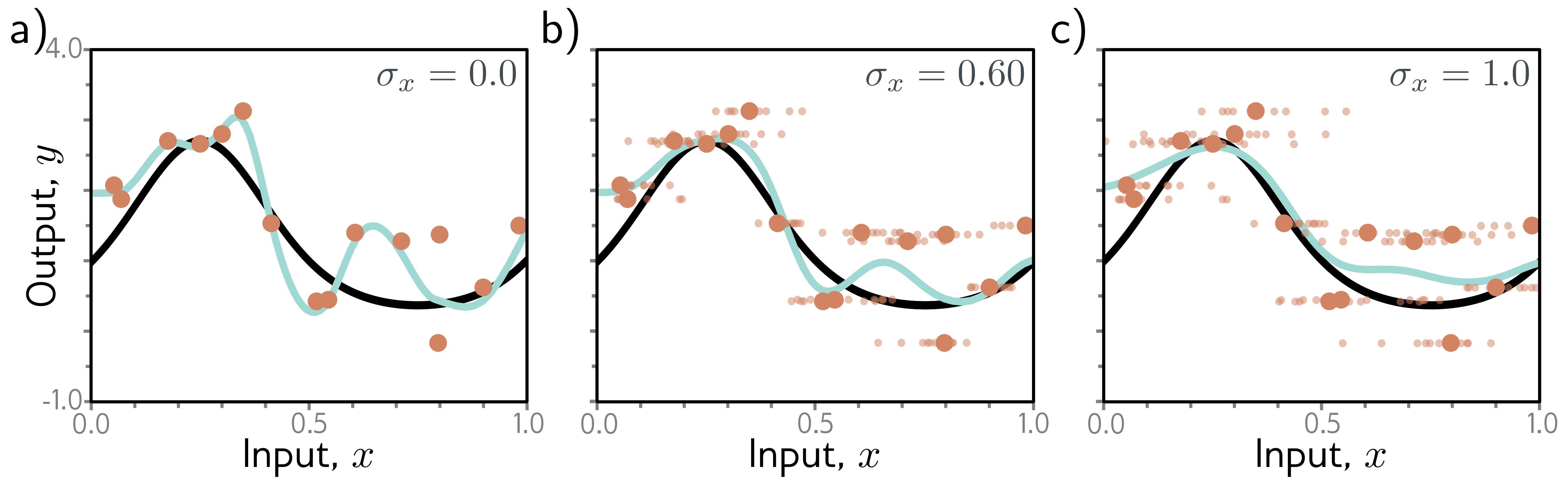
# Dropout

## On our toy model



# Input noise

## On our toy model



Make function smoother

Explicit L2  
regularization

Early  
stopping

Apply noise  
to inputs  
  
Apply noise  
to outputs  
(label smoothing)

Increase data

Data  
augmentation

Multi-task  
learning

Transfer  
learning

Ensembling

Bayesian  
approach

Dropout

Implicit  
regularization

Apply noise  
to weights

Combine multiple models

Find wider minima