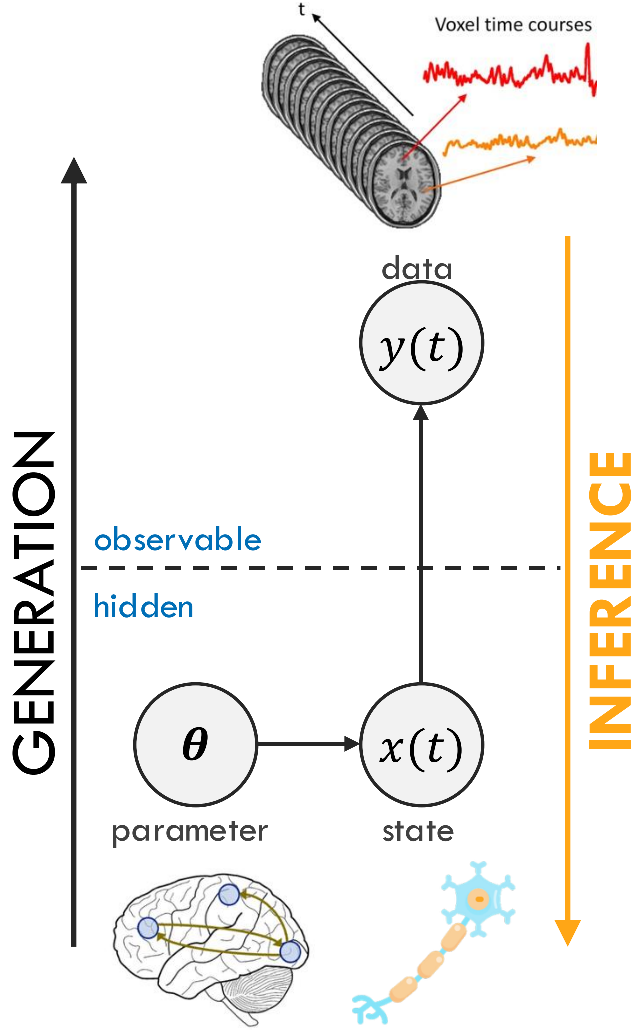
Fitting a Model: Maximum Likelihood Estimation (MLE)

Herman Galioulline

Recap: generative modeling



Last talk:

✓ Building a model
✓ Simulating data

This talk:

? Fitting the model to observed data

Recap: Bayesian workflow

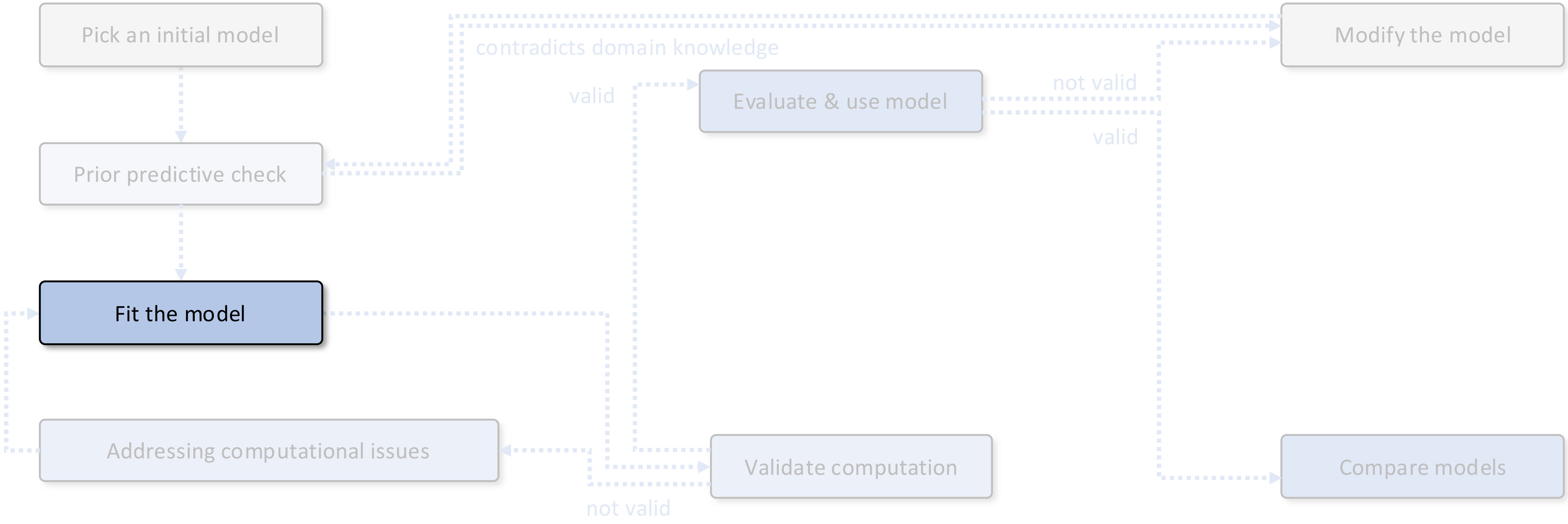


Figure from Gelman et al., 2020, *arXiv*

MLE: maximum likelihood estimator

Principle:

Find the parameters θ for which the **acquired data Y is most likely** under the model m.

$$\theta_{MLE} = \underset{\theta \in \Theta}{\arg\max}\ \underbrace{p(Y \mid \theta, m)}_{\text{Likelihood}}$$

where

$$p(Y \mid \theta, m) = p(y_{1 \dots T} \mid \theta, m)$$

m	model
Θ	parameter space
θ	model parameters
θ_{MLE}	MLE estimate of θ
Y	observed dataset
y_t	single observation
T	number of trials

Example: slot machines

Understand how people learn to maximise their rewards in a case where the most rewarding choice is initially unknown.

vs.

$p(\ |\ ) = 0.8$

$p(\ |\ ) = 0.2$

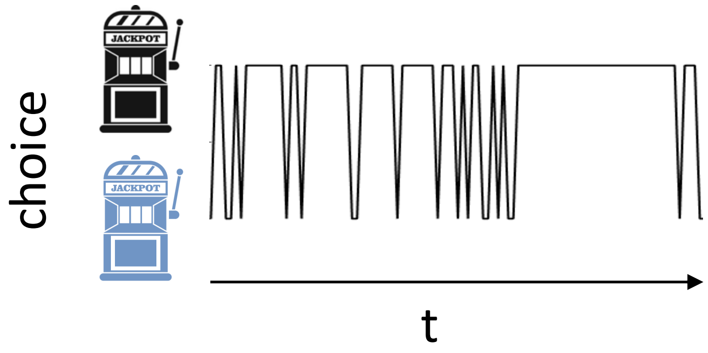
Experiment

Observations

Dataset:

Choice y_t in each trial t

$$Y = (y_1, \dots, y_T)$$



Based on slides by Alex Hess

Wilson & Collins, 2019, *eLife*

Specifying the likelihood function

Model 1
Random choice

$$p_t^1 = b$$
$$p_t^2 = 1 - b$$

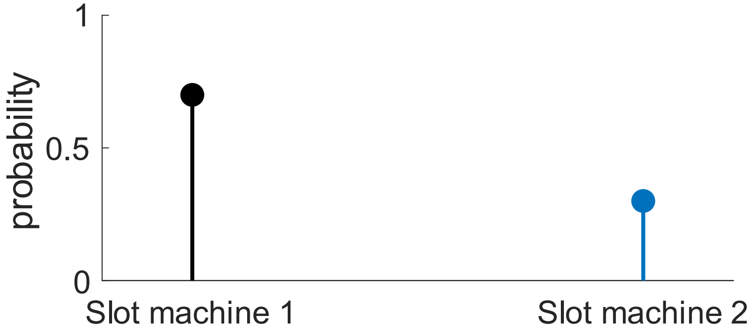
$$0 \le b \le 1$$

$$\boldsymbol{\theta} = \{\boldsymbol{b}\}$$

For a single trial t:

$$p(y_t \mid \theta, m) = \theta^{y_t}(1-\theta)^{(1-y_t)}$$
$$= Bernoulli$$

$$p(1 \mid 0.9, m_1) = 0.9^1(1-0.9)^{(1-1)} = 0.9$$
$$p(0 \mid 0.9, m_1) = 0.9^0(1-0.9)^{(1-0)} = 0.1$$



$Y = \{y_1, \ldots, y_T\}$

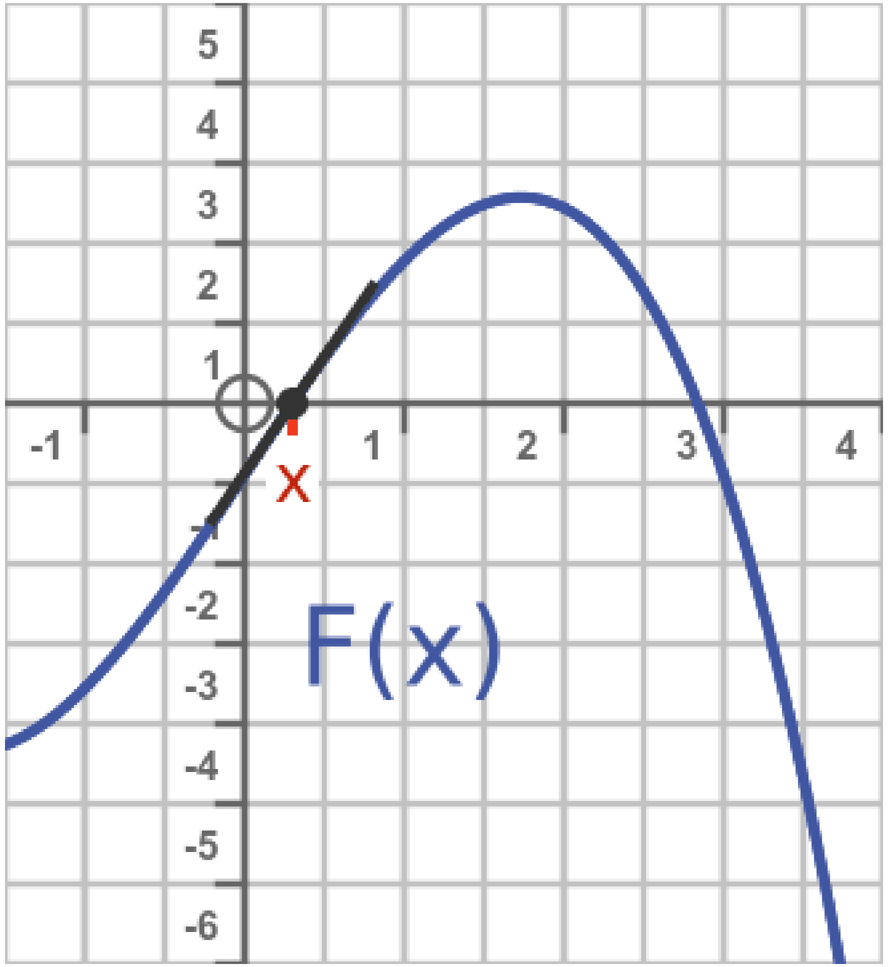
Assume trial independence

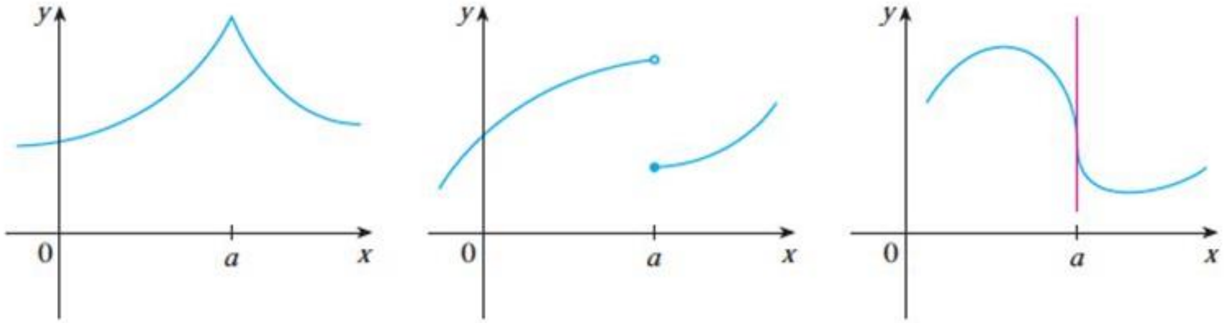
For all trials 1 … T:

$$p(Y \mid \theta, m) = p(y_{1\ldots T} \mid \theta, m) = \prod_{t=1}^{T} \theta^{y_t}(1-\theta)^{(1-y_t)}$$

$$0.9 * 0.1 * 0.1 * \cdots * 0.9$$

Math reminder: first-order derivative





(a) A corner

(b) A discontinuity

(c) A vertical tangent

https://graphicmaths.com/pure/integration/fundamental-theorem-calculus/

Maximising the likelihood function

Likelihood function

$$p(Y \mid \theta, m) = \prod_{t=1}^{T} p(y_t \mid \theta, m)$$

$$\underset{\theta \in \Theta}{\arg\max}$$

Analytical solution

Is the likelihood tractable?
Is the likelihood differentiable?

→ Solve $\frac{d}{d\theta} p(Y \mid \theta, m) \stackrel{!}{=} 0$ and find maximum

Numerical solution

Use numerical optimisation routines available in different software (MATLAB, Python, Julia, etc.)

→ Implement $p(Y \mid \theta, m)$ and find the maximum

Maximising the likelihood function

Analytical solution

$$p(Y \mid \theta, m) = \prod_{t=1}^{T} \theta^{y_t}(1-\theta)^{(1-y_t)}$$

Likelihood function for random choice model

1.

$$\log p(Y \mid \theta, m) = \log \prod_{t=1}^{T} \theta^{y_t}(1-\theta)^{(1-y_t)}$$

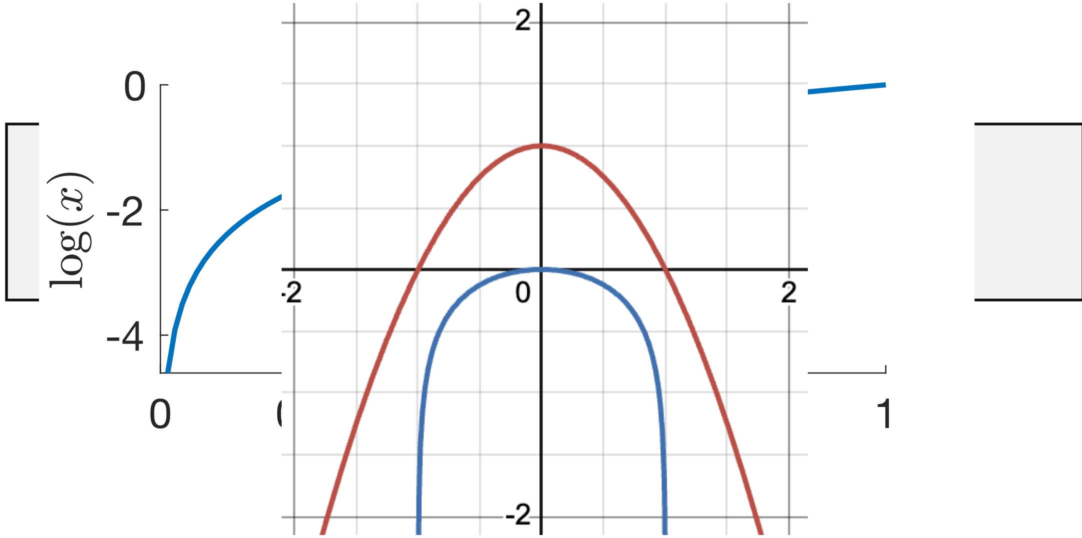
$$= \sum_{t=1}^{T} \log \theta^{y_t}(1-\theta)^{(1-y_t)}$$

2.

$$= \sum_{t=1}^{T} y_t \log(\theta) + (1-y_t)\log(1-\theta)$$

1. Change product to sum by log-transformation:

$$\log\left(\prod_t x_t\right) = \sum_t \log x_t$$



Maximising the likelihood function

Analytical solution

$$\sum_{t=1}^{T} y_t \log(\theta) + (1-y_t)\log(1-\theta)$$

$$= \sum_{t=1}^{T} y_t \log(\theta) + \sum_{t=1}^{T} (1-y_t)\log(1-\theta)$$

$$= \log(\theta)\sum_{t=1}^{T} y_t + \log(1-\theta)\sum_{t=1}^{T}(1-y_t)$$

$$\frac{d}{d\theta}\left[\log(\theta)\sum_{t=1}^{T} y_t + \log(1-\theta)\sum_{t=1}^{T}(1-y_t)\right] \overset{!}{=} 0$$

$$\frac{d}{d\theta}\log(\theta)\sum_{t=1}^{T} y_t + \frac{d}{d\theta}\log(1-\theta)\sum_{t=1}^{T}(1-y_t) \overset{!}{=} 0$$

$$\frac{1}{\theta}\sum_{t=1}^{T} y_t - \frac{1}{1-\theta}\sum_{t=1}^{T}(1-y_t) \overset{!}{=} 0$$

$$\frac{1-\theta}{\theta(1-\theta)}\sum_{t=1}^{T} y_t - \frac{\theta}{\theta(1-\theta)}\sum_{t=1}^{T}(1-y_t) \overset{!}{=} 0$$

$$\frac{1}{\theta(1-\theta)}\left[(1-\theta)\sum_{t=1}^{T} y_t - \theta\sum_{t=1}^{T}(1-y_t)\right] \overset{!}{=} 0$$

$$(1-\theta)\sum_{t=1}^{T} y_t - \theta\sum_{t=1}^{T}(1-y_t) \overset{!}{=} 0$$

$$\sum_{t=1}^{T} y_t - \theta\sum_{t=1}^{T} y_t - \theta\sum_{t=1}^{T} 1 + \theta\sum_{t=1}^{T} y_t \overset{!}{=} 0$$

Maximising the likelihood function
Analytical solution

$$\sum_{t=1}^{T} y_t - \theta \sum_{t=1}^{T} y_t - \theta \sum_{t=1}^{T} 1 + \theta \sum_{t=1}^{T} y_t \overset{!}{=} 0$$

$$\sum_{t=1}^{T} y_t - \theta \sum_{t=1}^{T} 1 \overset{!}{=} 0$$

$$\sum_{t=1}^{T} y_t - \theta T \overset{!}{=} 0$$

$$\sum_{t=1}^{T} y_t \overset{!}{=} \theta T$$

MLE estimate

$$\boxed{\theta_{MLE} = \frac{1}{T} \sum_{t=1}^{T} y_t}$$

$\downarrow$

Max. likelihood estimate is arithmetic mean of data!

Advantages of maximum likelihood estimation

1. **Easy to compute:** simple optimisation problem

Advantages of maximum likelihood estimation

1. **Easy to compute:** simple optimisation problem

2. **Interpretable:** often θ_{MLE} is intuitively interpretable wrt. model parameters (see MLE of random choice model)

Advantages of maximum likelihood estimation

1. **Easy to compute:** simple optimisation problem

2. **Interpretable:** often θ_{MLE} is intuitively interpretable wrt. model parameters (see MLE of random choice model)

3. **Asymptotic properties:** consistency (true parameter value recovered) and efficiency (lowest possible parameter variance)

Advantages of maximum likelihood estimation

1. **Easy to compute:** simple optimisation problem

2. **Interpretable:** often θ_{MLE} is intuitively interpretable wrt. model parameters (see MLE of random choice model)

3. **Asymptotic properties:** consistency (true parameter value recovered) and efficiency (lowest possible parameter variance)

4. **Invariant to reparameterisation:** if $\theta_{MLE} = \mathrm{MLE}(\theta)$ for $\theta \in \Theta$, then $g(\theta_{MLE}) = \mathrm{MLE}(\mathrm{g}(\theta))$ for $g: \mathbb{R} \to \mathbb{R}$

Limitations of maximum likelihood estimation

1. **Point estimate:** θ_{MLE} is a point estimate → no representation of uncertainty

Limitations of maximum likelihood estimation

1. **Point estimate:** θ_{MLE} is a point estimate → no representation of uncertainty

2. **Existence & uniqueness:** The MLE might not be unique or even non-existent, depending on properties of the likelihood function and parameter space

Limitations of maximum likelihood estimation

1. **Point estimate:** θ_{MLE} is a point estimate → no representation of uncertainty

2. **Existence & uniqueness:** The MLE might not be unique or even non-existent, depending on properties of the likelihood function and parameter space

3. **Overfitting:** MLE is limited to a finite set of observed datapoints

Limitations of maximum likelihood estimation

3. **Overfitting:** MLE is limited to a finite set of observed datapoints, unlike in the case of the asymptotic property

Example

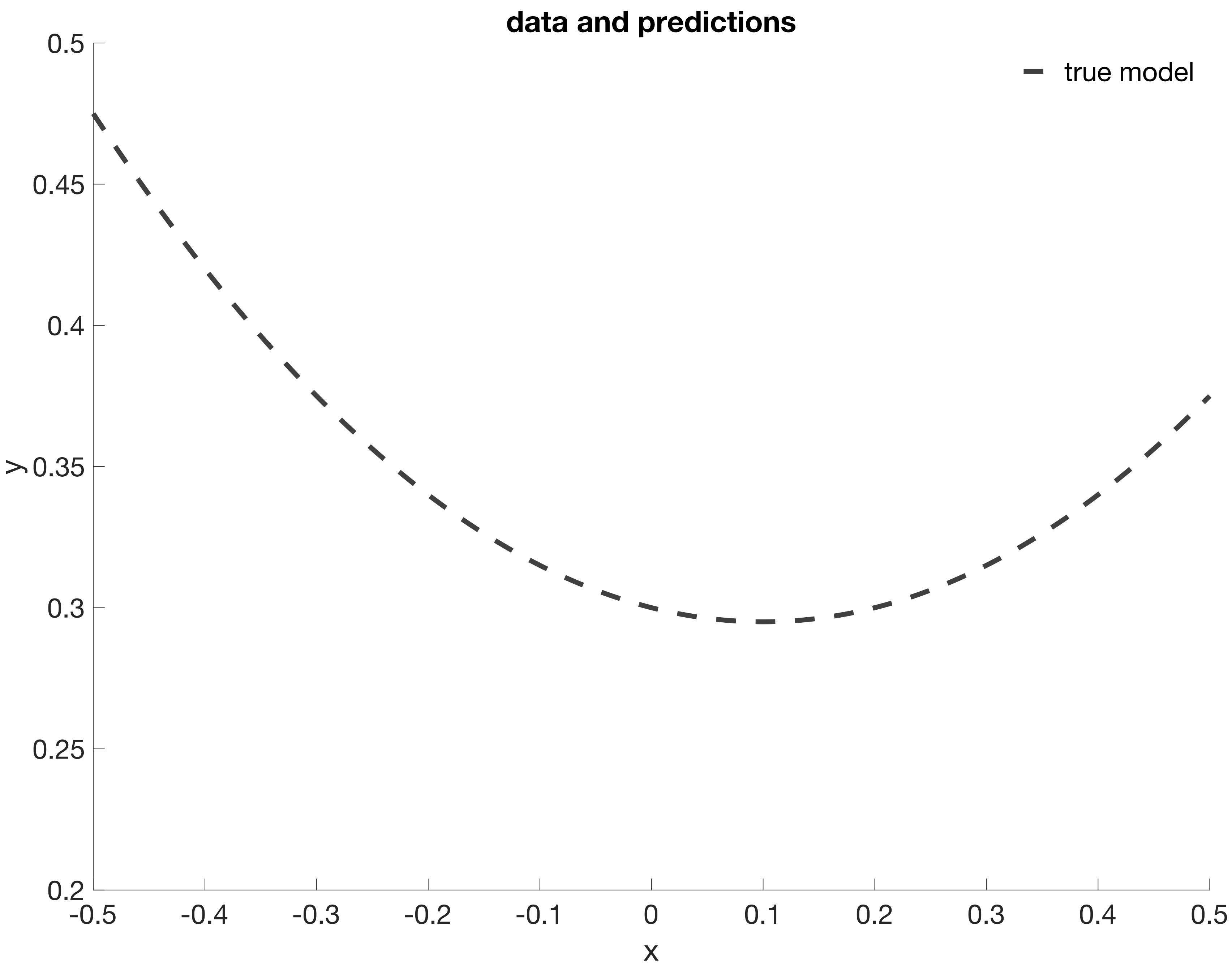
Polynomial model of order P with Gaussian noise

$$y = \theta_0 + \theta_1 x + \cdots + \theta_P x^P + \epsilon$$
$$= \boldsymbol{x\theta} + \epsilon$$

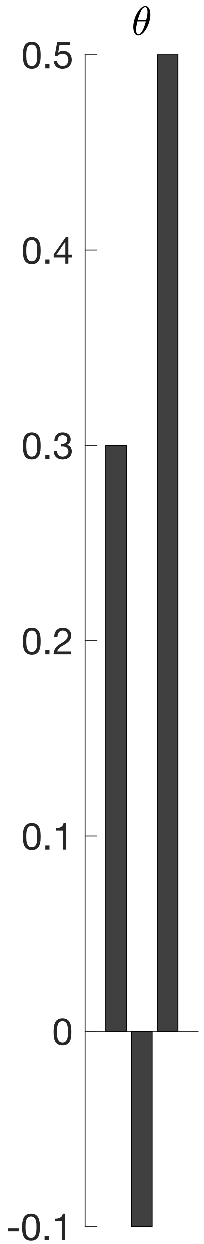
where

$\epsilon \sim \mathcal{N}(0, \sigma^2)$

Higher order => more degrees of freedom

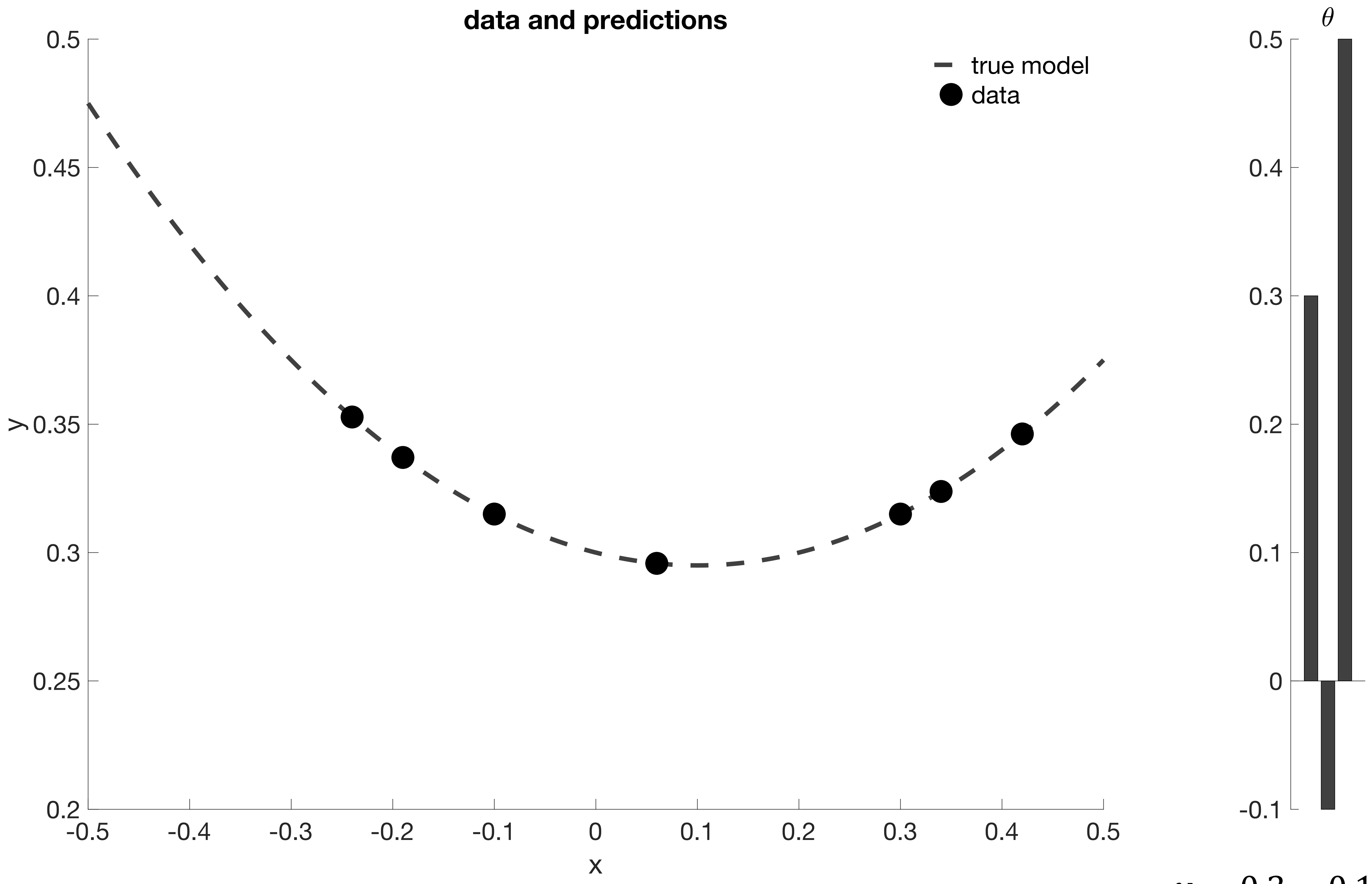


data and predictions
true model
y
x
0.5
0.45
0.4
0.35
0.3
0.25
0.2
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5



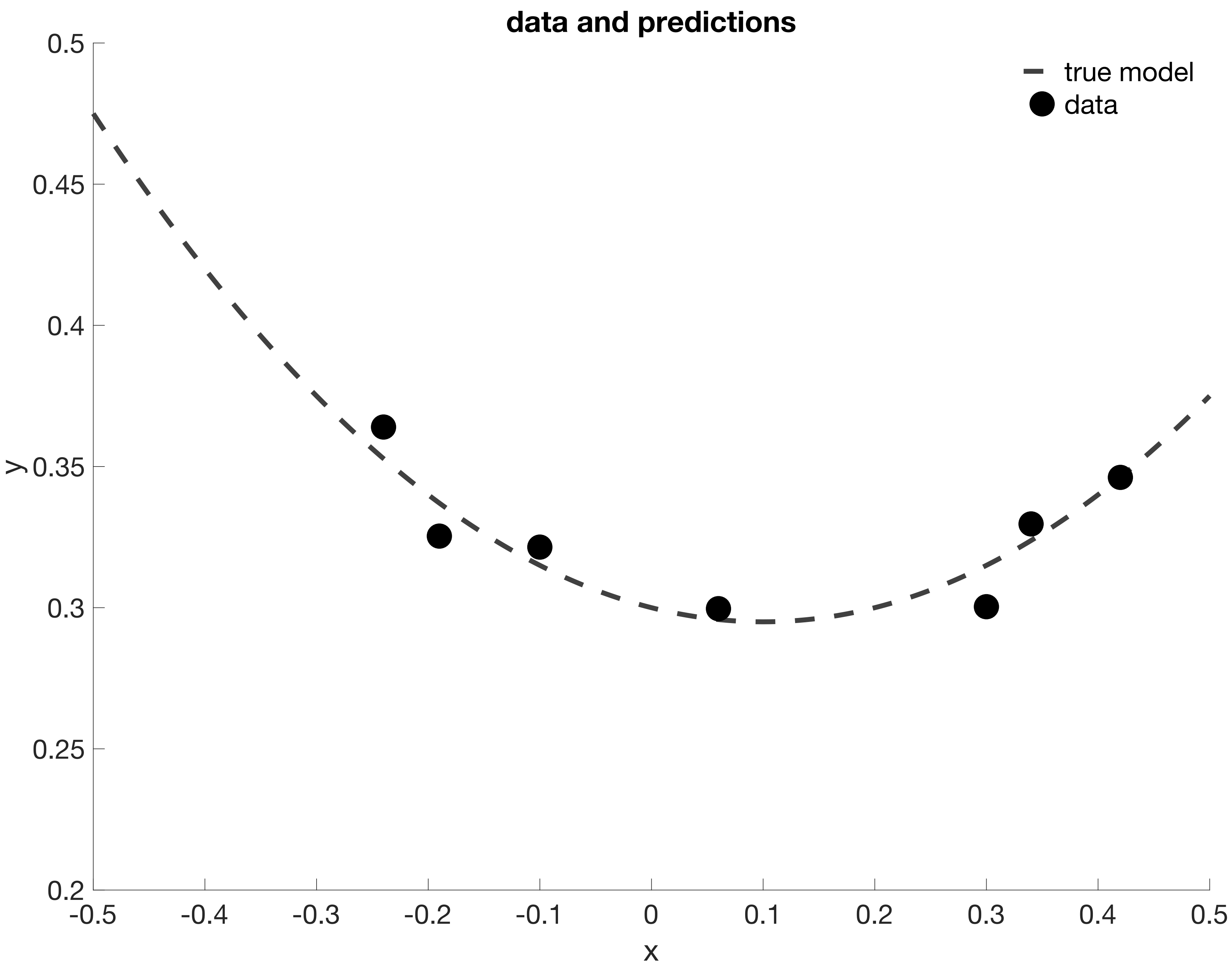
θ
0.5
0.4
0.3
0.2
0.1
0
-0.1

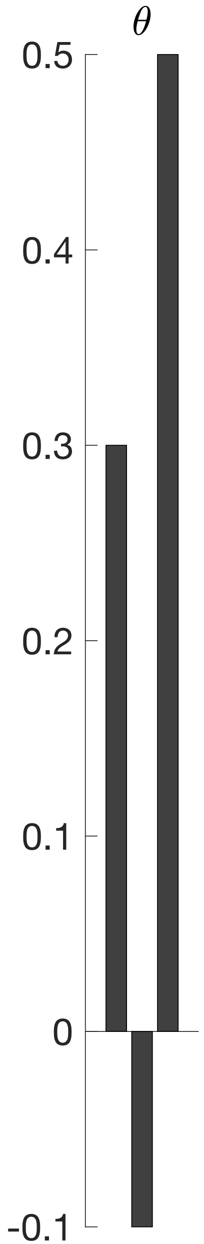
$$y = 0.3 - 0.1x + 0.5x^2$$



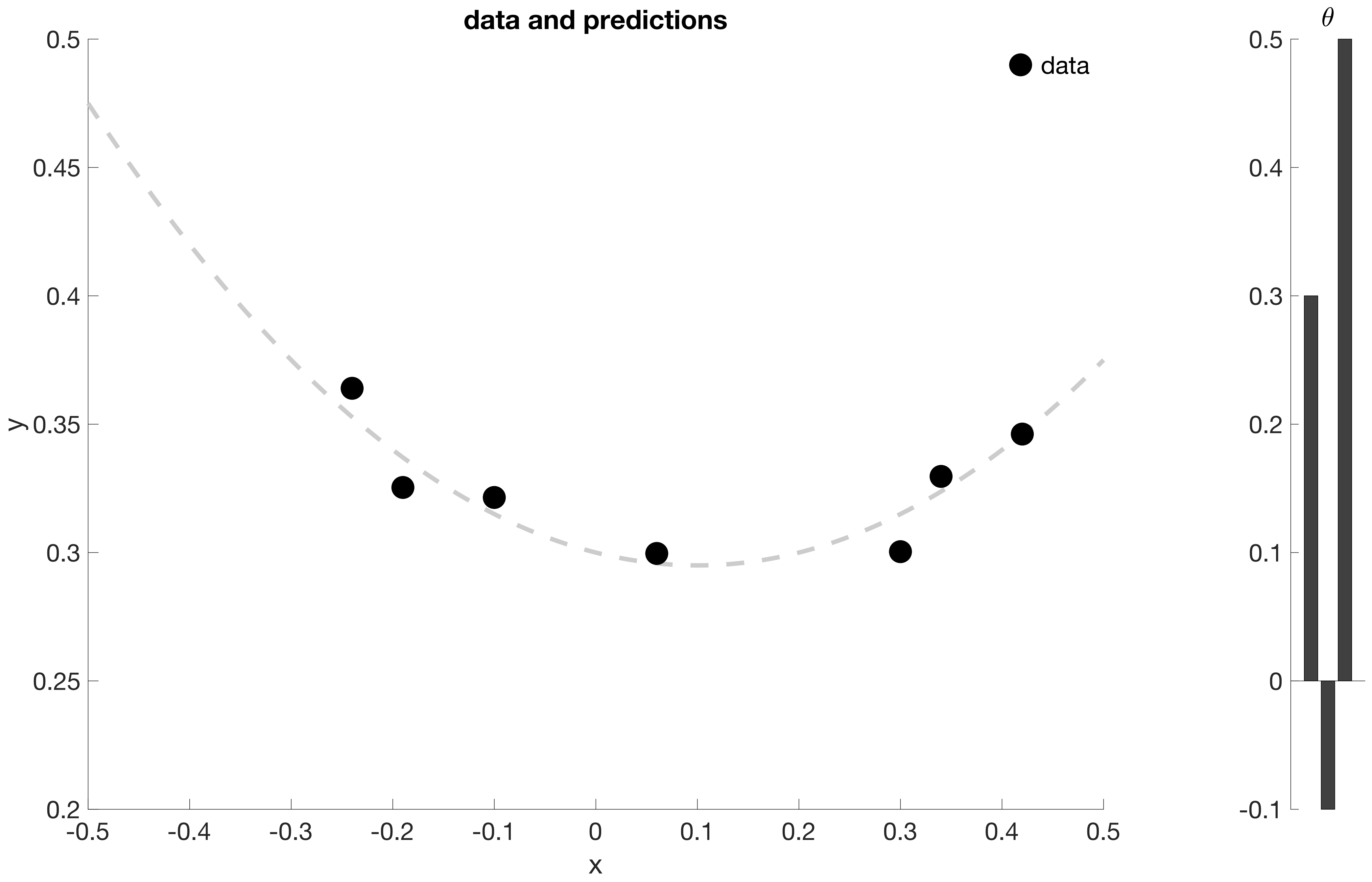
data and predictions
true model
data
y
x
0.5
0.45
0.4
0.35
0.3
0.25
0.2
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
θ
0.5
0.4
0.3
0.2
0.1
0
-0.1

$$y = 0.3 - 0.1x + 0.5x^2$$

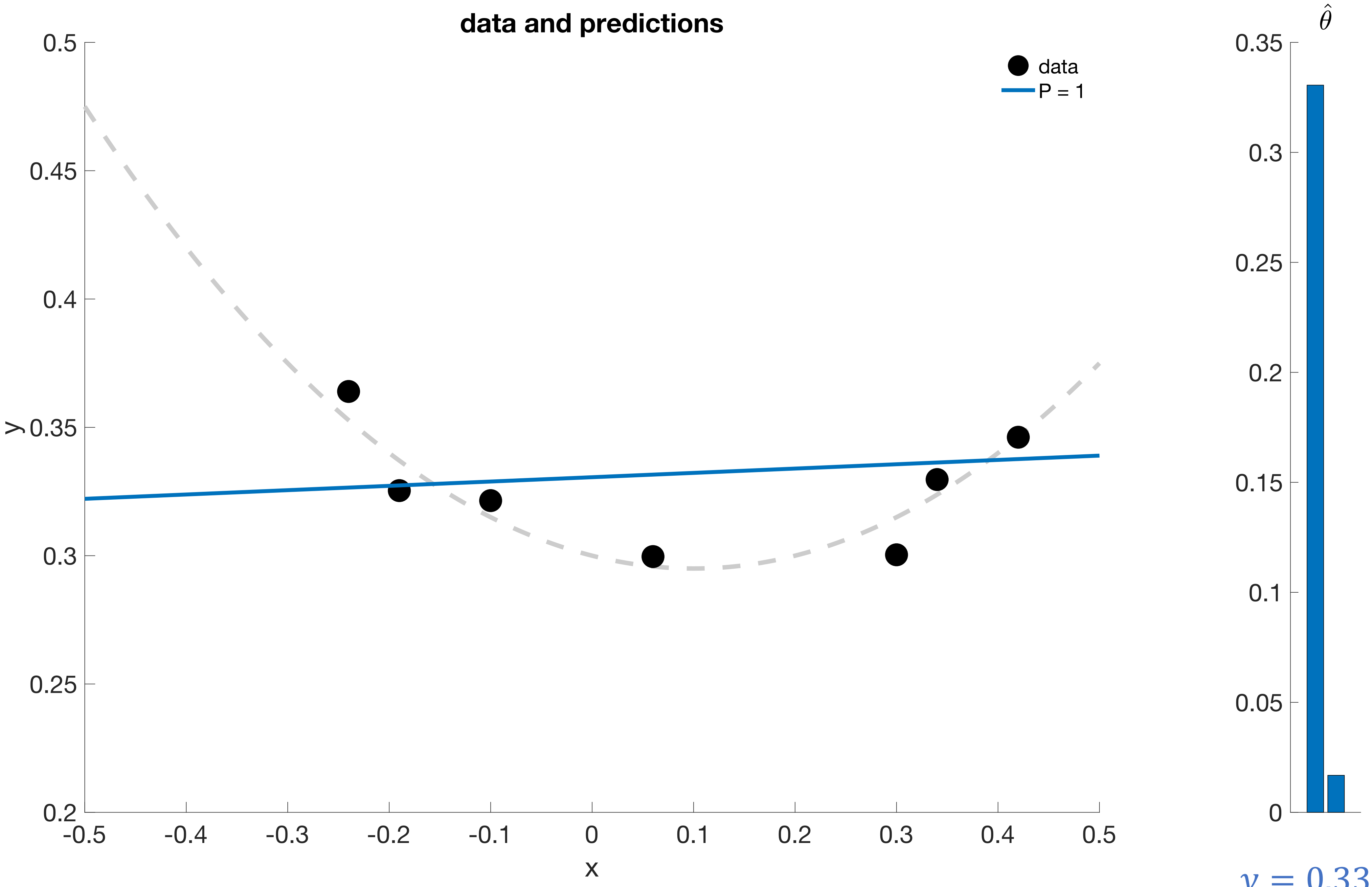




$$y = 0.3 - 0.1x + 0.5x^2 + \epsilon$$

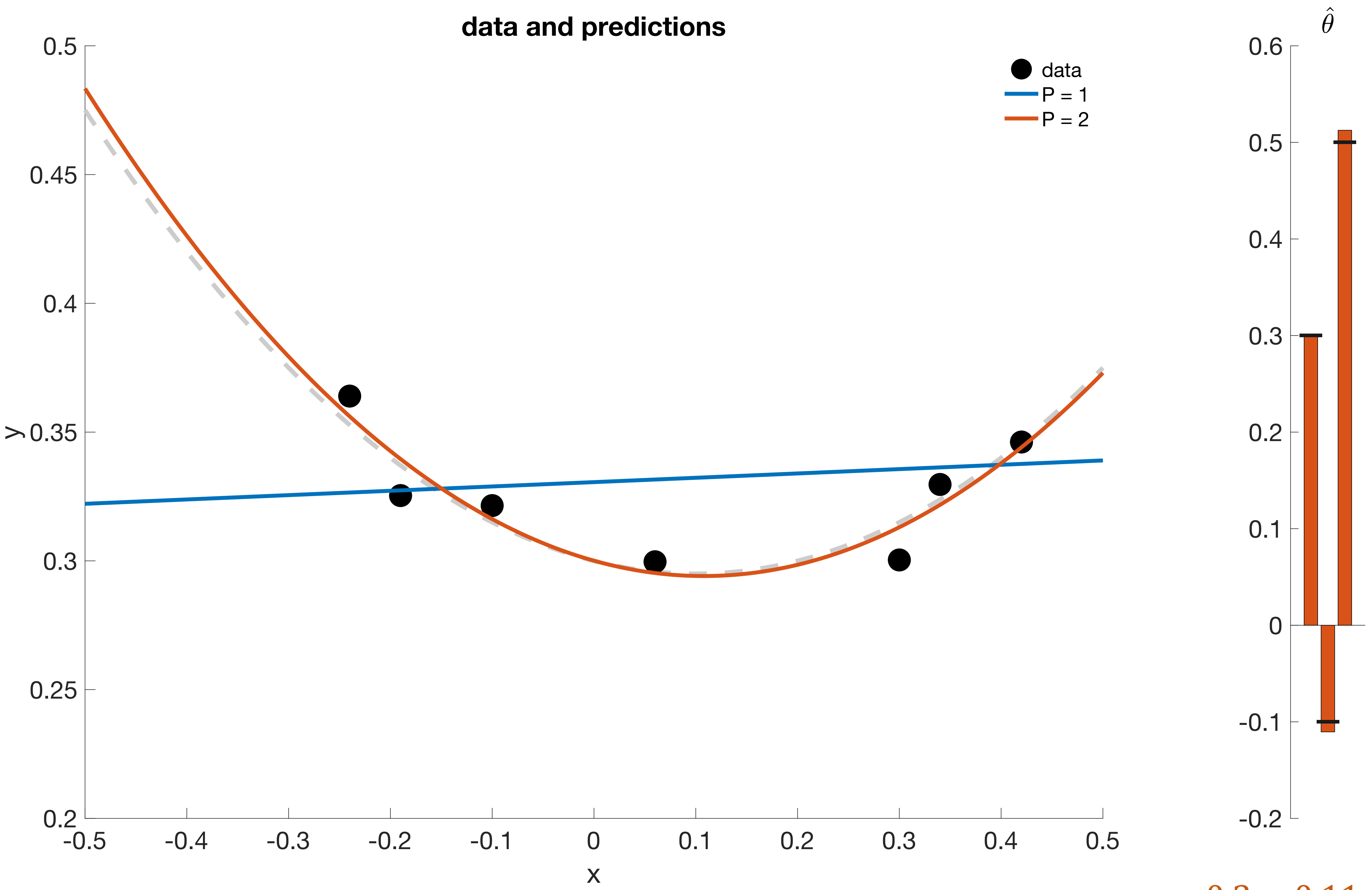


data and predictions
data
y
x
0.5
0.45
0.4
0.35
0.3
0.25
0.2
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
θ
0.5
0.4
0.3
0.2
0.1
0
-0.1

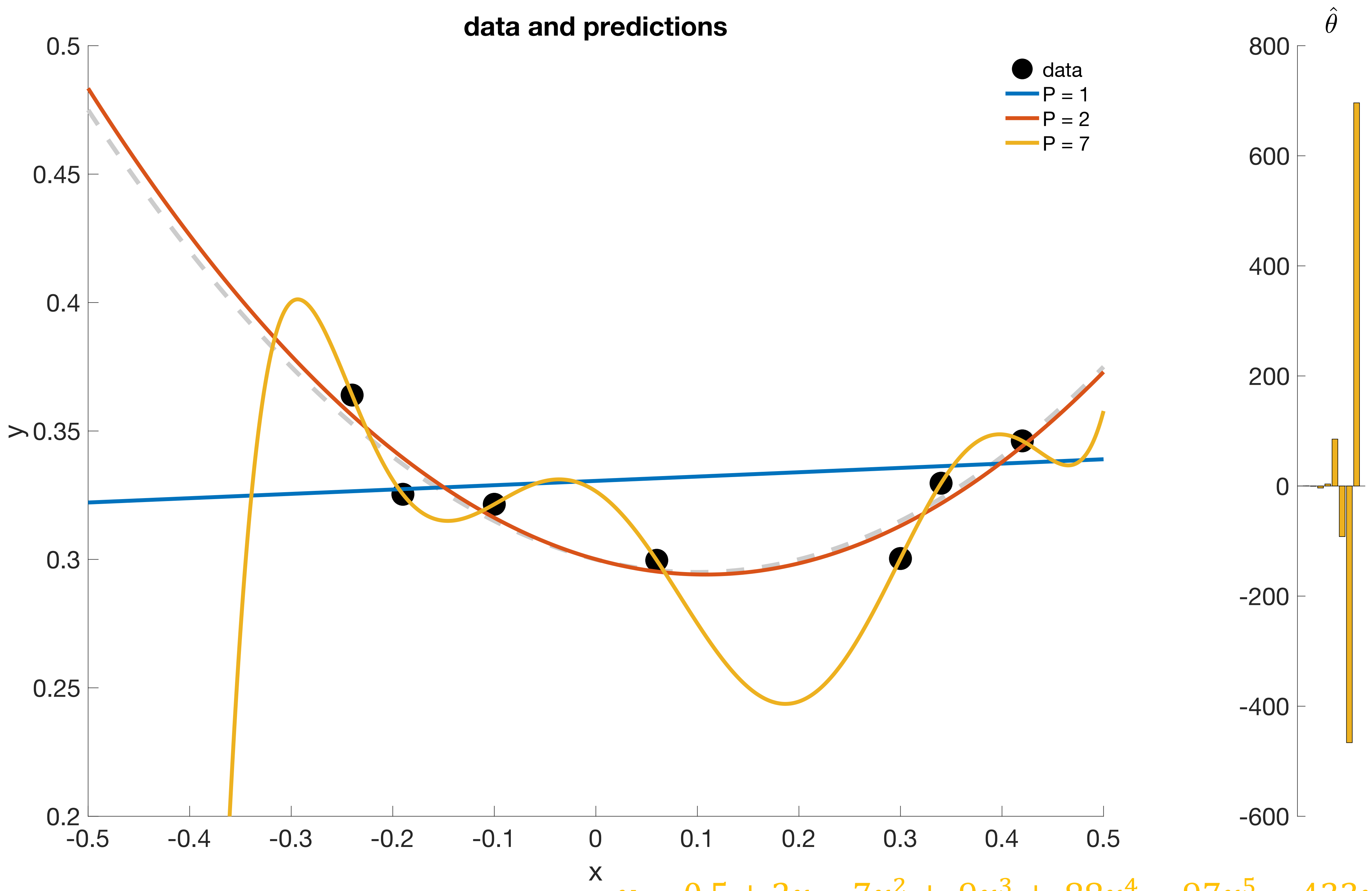


data and predictions
data
P = 1
0.5
0.45
0.4
0.35
0.3
0.25
0.2
y
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
x
$\hat{\theta}$
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0

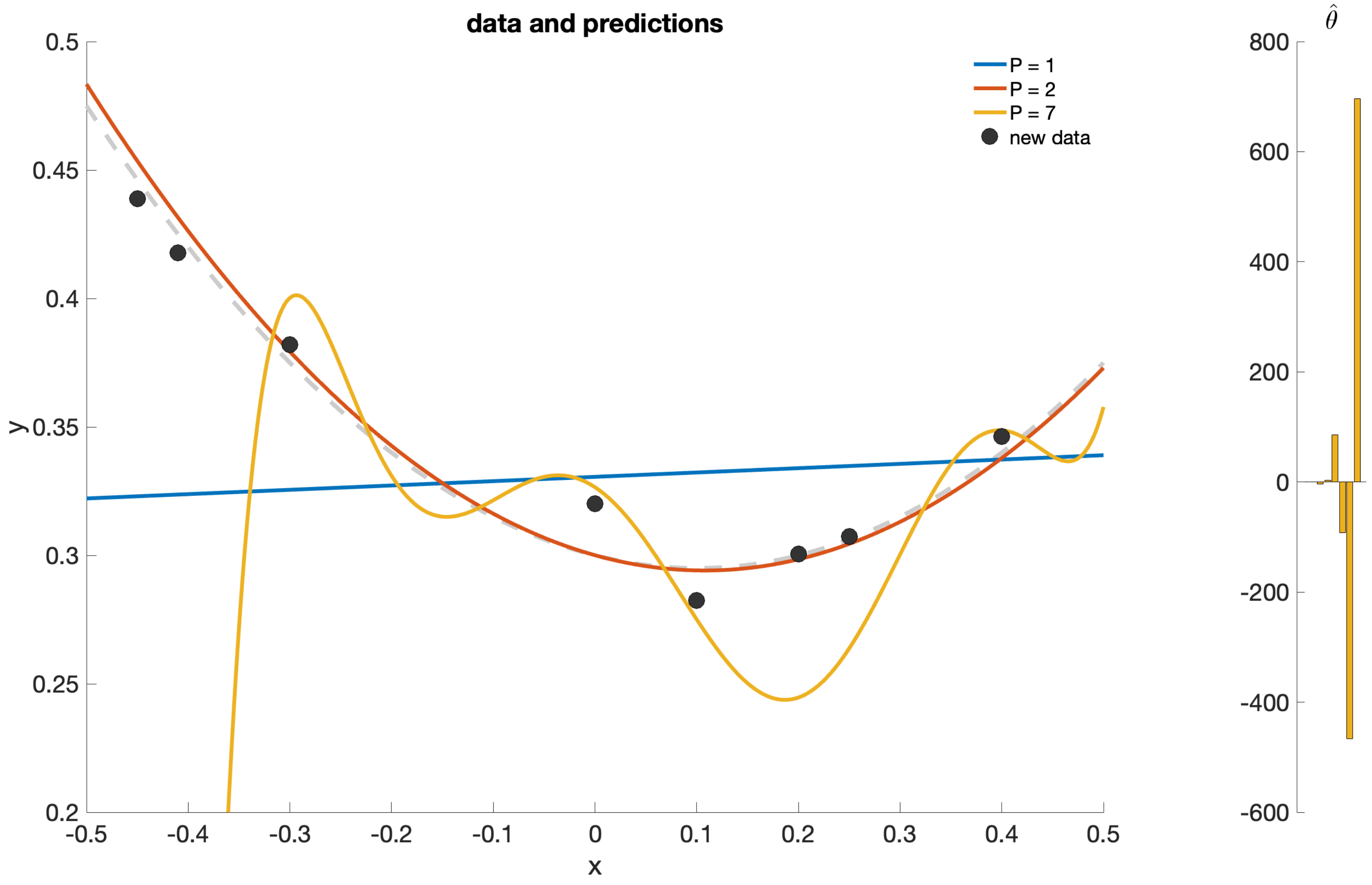
$$y = 0.33 + 0.01x$$



$$y = 0.3 - 0.11x + 0.52x^2$$



$$y = 0.5 + 2x - 7x^2 + 9x^3 + 88x^4 - 97x^5 - 433x^6 + 702x^7$$


data and predictions
P = 1
P = 2
P = 7
new data
y
x
0.5
0.45
0.4
0.35
0.3
0.25
0.2
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
$\hat{\theta}$
800
600
400
200
0
-200
-400
-600

Maximum likelihood vs. full Bayesian inference

Bayesian statistics:

Likelihood

Prior

Posterior

$$p(\theta \mid y, m) = \frac{p(y \mid \theta, m)\, p(\theta, m)}{p(y \mid m)}$$

Model evidence

Maximum-a-posteriori (**MAP**) estimation:

→ Point estimate of the posterior (no uncertainty)

→ Under a flat prior MAP=MLE (estimates the joint)

Variational Bayes (**VB**), sampling-based (**MCMC**) techniques

→ Full posterior densities

Acknowledgement

Special thanks to my TNU colleagues

Social Evening Tomorrow: Bouldering after last lecture

QUESTIONS?

hermanG@ethz.ch

https://github.com/computational-psychiatry-course/cpc2025

https://www.linkedin.com/in/hermangal