Lecture 14: Bayesian Linear Regression

Professor Ilias Bilionis

Probabilistic interpretation of leas squares - Estimating the measurement noise

PREDICTIVE
SCIENCE LABORATORY

Reminder: Generalized linear model and least squares fit

$x_{1:N} = (x_1, \dots, x_N)$; $y_{1:N} = (y_1, \dots, y_N)$

$(w_1, \dots, w_m)$

$$y = w_1 \varphi_1(x) + \dots + w_m \varphi_m(x) = \sum_{j=1}^{m} w_j \varphi_j(x) = \underline{\varphi}^T(x)\, \underline{w}$$

features

basis functions

$(\varphi_1(x), \dots, \varphi_m(x))$

$(y_1, \dots, y_N)$

$$\min L(\underline{w}) = \sum_{i=1}^{N} \left(y_i - \underline{\varphi}^T(x_i)\, \underline{w} \right)^2$$

$$\nabla_{\underline{w}} L(\underline{w}) = 0 \Rightarrow \underline{\underline{\Phi}}^T \underline{\underline{\Phi}} \cdot \underline{w} = \underline{\underline{\Phi}}^T \cdot \underline{y}$$

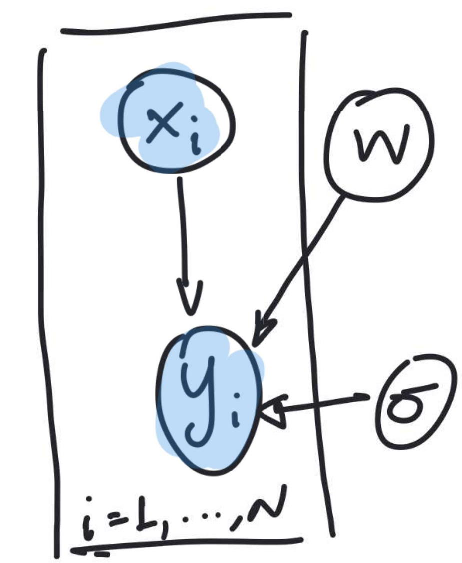
Design matrix $N \times m$

$$\underline{\underline{\Phi}} = \begin{pmatrix} \varphi_1(x_1) & \dots & \varphi_m(x_1) \\ \vdots & & \vdots \\ \varphi_1(x_N) & & \varphi_m(x_N) \end{pmatrix}$$

Open questions

- How do I quantify the measurement noise?
- How do we avoid overfitting?
- How do I quantify epistemic uncertainty induced by limited data?
- How do I choose any remaining parameters?
- How do I choose which basis functions to keep?

Probabilistic interpretation



(plate notation)

Prior: $\underline{w} \sim p(\underline{w})$

$\sigma \sim p(\sigma)$

Likelihood:

$$y_i | x_i, \underline{w}, \sigma^2 \sim N(\underline{\phi}^T(x_i)\underline{w}, \sigma^2)$$

$$p(y_{1:N} | x_{1:N}, \underline{w}, \sigma^2) = \prod_{i=1}^{N} p(y_i | x_i, \underline{w}, \sigma^2)$$

$$= \prod_{i=1}^{N} (2\pi)^{-1/2} \sigma^{-1} \exp\left\{-\frac{(y_i - \underline{\phi}^T(x_i)\underline{w})^2}{2\sigma^2}\right\}$$

$$= (2\pi)^{-N/2} \sigma^{-N} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \underline{\phi}^T(x_i)\underline{w})^2\right\}$$

Posterior $\propto$ Likelihood $\cdot$ Prior

$$p(\underline{w}, \sigma | x_{1:N}, y_{1:N}) \propto \boxed{p(y_{1:N} | x_{1:N}, \underline{w}, \sigma^2)}\, \underbrace{p(\underline{w})}_{1}\, \underbrace{p(\sigma)}_{1}$$

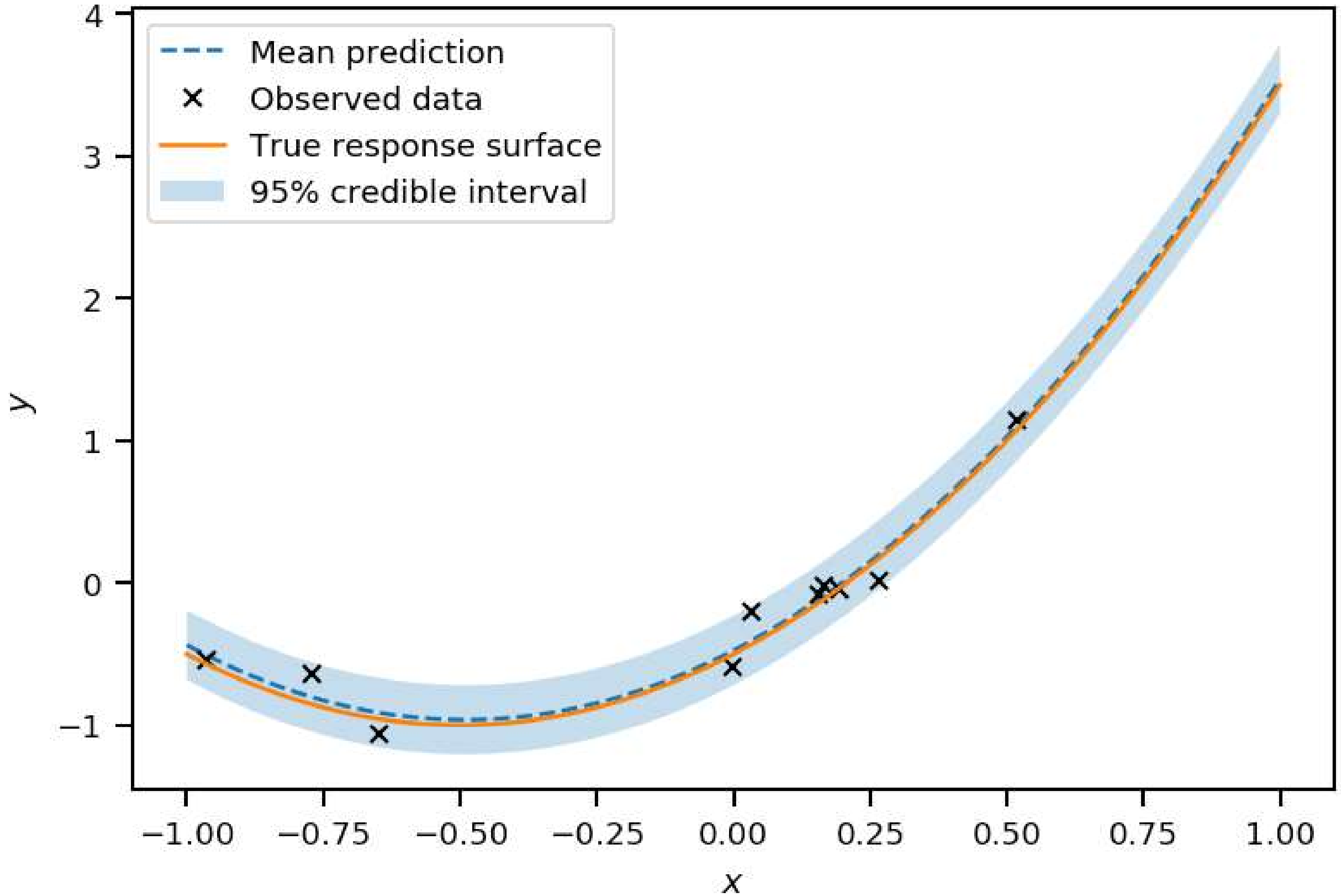
Maximum likelihood estimate of weights yields least squares

$$\log p(y_{1:N} \mid x_{1:N}, \underline{w}, \sigma) = -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\underbrace{\sum_{i=1}^{N}\left(y_i - \underline{\varphi}^T(x_i)\underline{w}\right)^2}_{= L(\underline{w})}$$

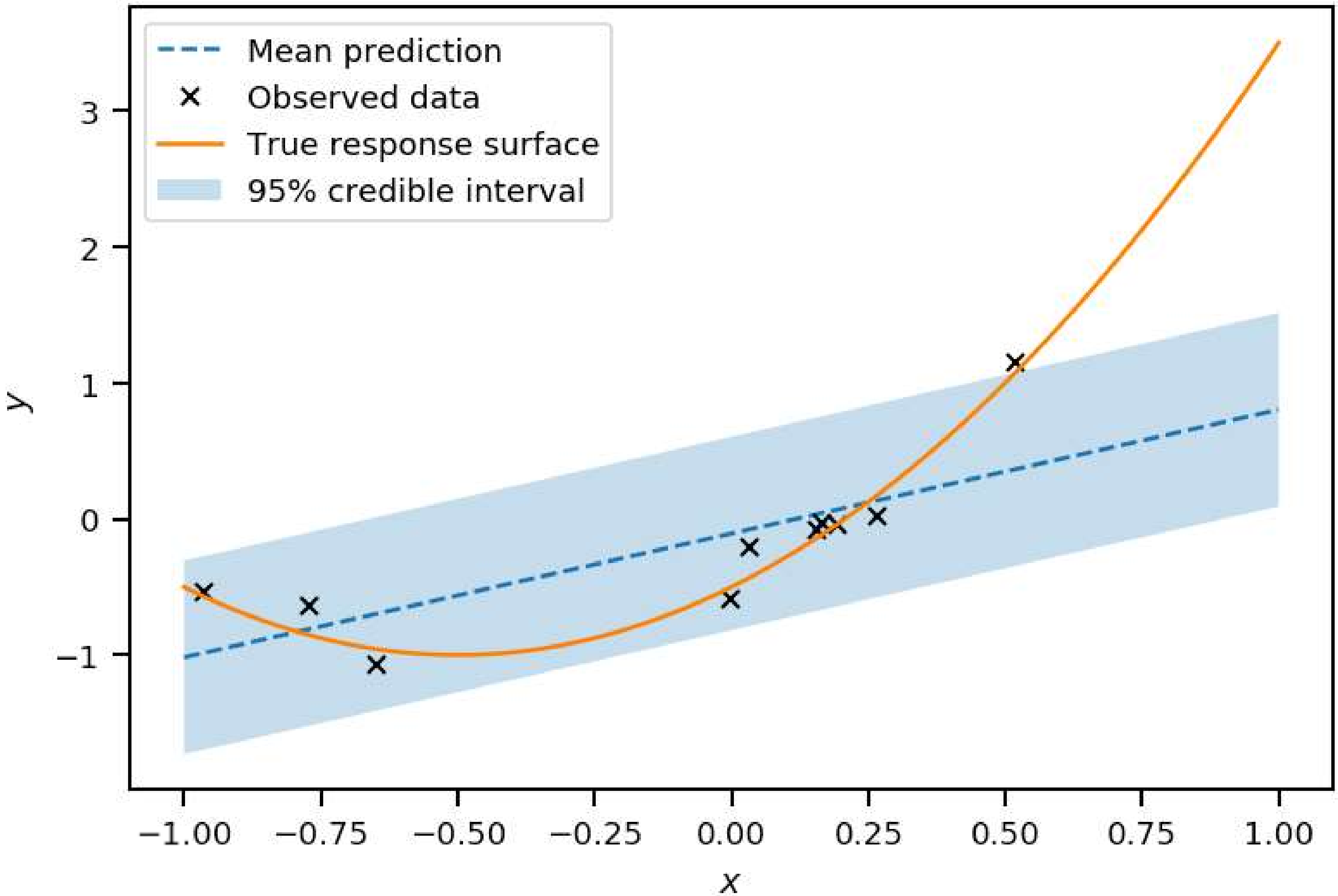
$$\max_{\underline{w}} \log \text{like} \equiv \min_{\underline{w}} L(\underline{w})$$

$$\underline{\underline{\Phi}}^T \underline{\underline{\Phi}}\, \underline{w} = \underline{\underline{\Phi}}^T \underline{y}$$

Example: Degree 2 polynomial



Example: Degree 1 polynomial

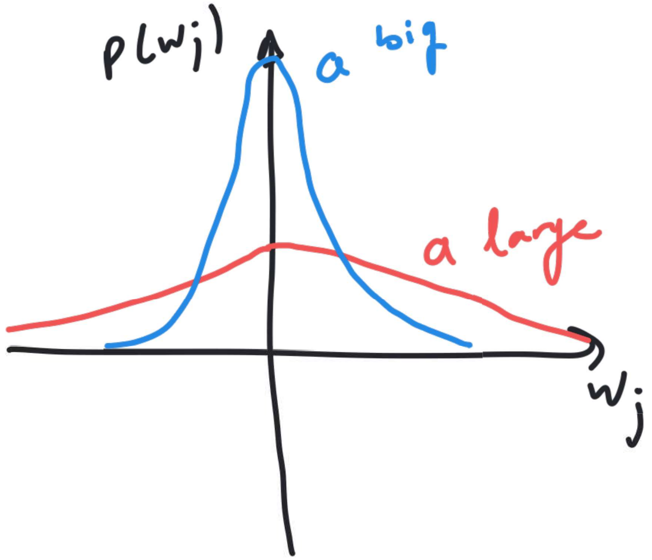


PREDICTIVE
SCIENCE LABORATORY

Open questions

- How do I quantify the measurement noise?
- How do we avoid overfitting?
- How do I quantify epistemic uncertainty induced by limited data?
- How do I choose any remaining parameters?
- How do I choose which basis functions to keep?

Gaussian prior on weights



$$w_j \sim N(0, \underbrace{\alpha^{-1}}_{\text{precision}}) \quad \text{precision} = \frac{1}{\text{variance}}$$

$$p(\underline{w}) = \prod_{j=1}^{m} p(w_j)$$

$$\propto \exp\left\{-\frac{\alpha}{2}\sum_{j=1}^{m} w_j^2\right\}$$

Maximum a posteriori estimate

posterior $\propto$ likelihood $\times$ prior

$$p(\underline{w} \mid x_{1:N}, y_{1:N}, \sigma) \propto p(y_{1:N} \mid x_{1:N}, \sigma^2)\, p(\underline{w})$$

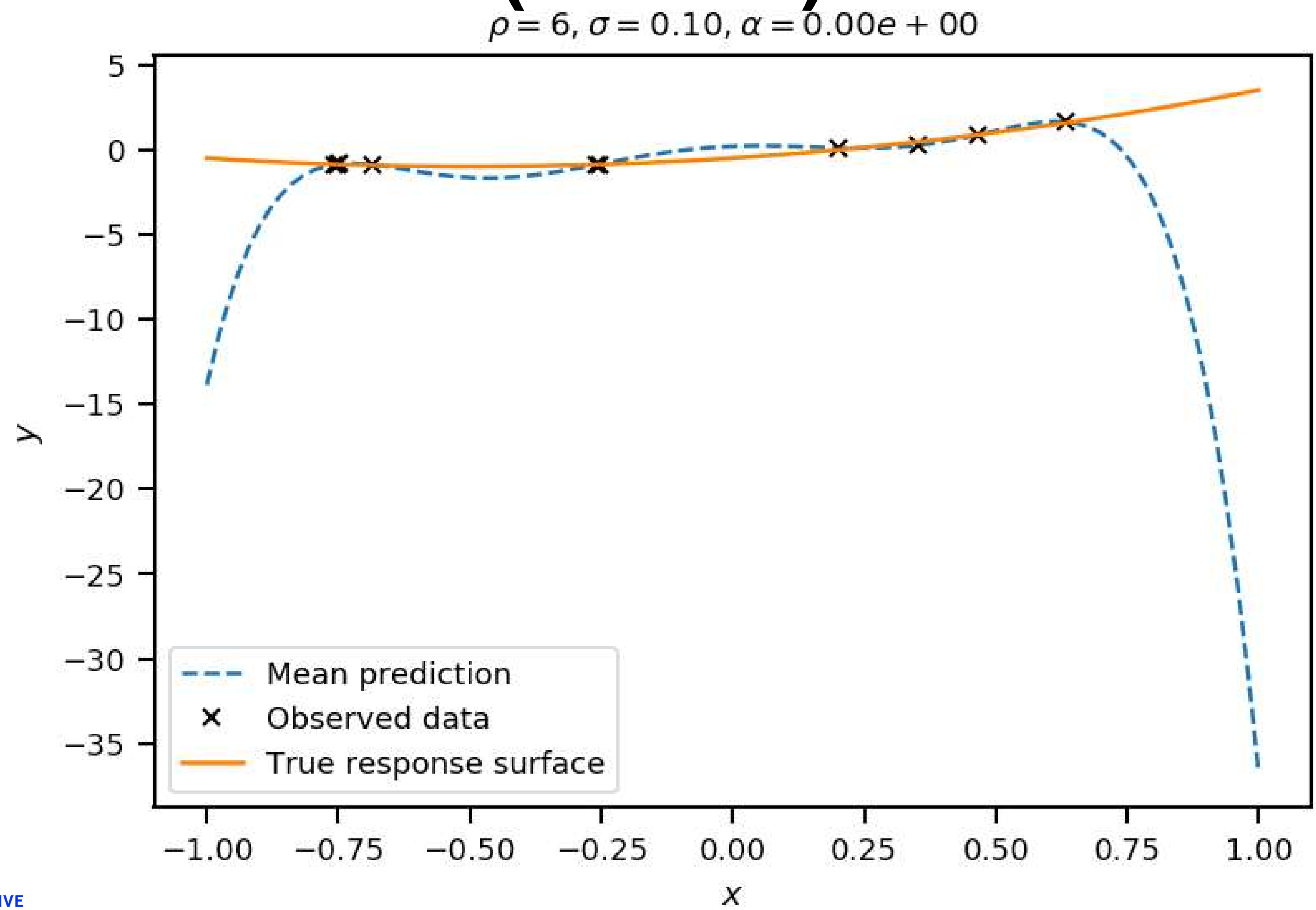
$$\max_{w} \log \text{post} = \log \text{like} + \log p(w)$$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - \underline{\varphi}^T(x_i)\underline{w}\right)^2 - \frac{\alpha}{2}\sum_{j=1}^{m} w_j^2 + \text{const.}$$

$$\min_{\underline{w}} \underbrace{\sum_{i=1}^{N} \left(y_i - \underline{\varphi}^T(x_i)\underline{w}\right)^2}_{\text{SE}} + \underbrace{\frac{\alpha}{2}\sum_{j=1}^{m} w_j^2}_{\text{Regularization term}}$$

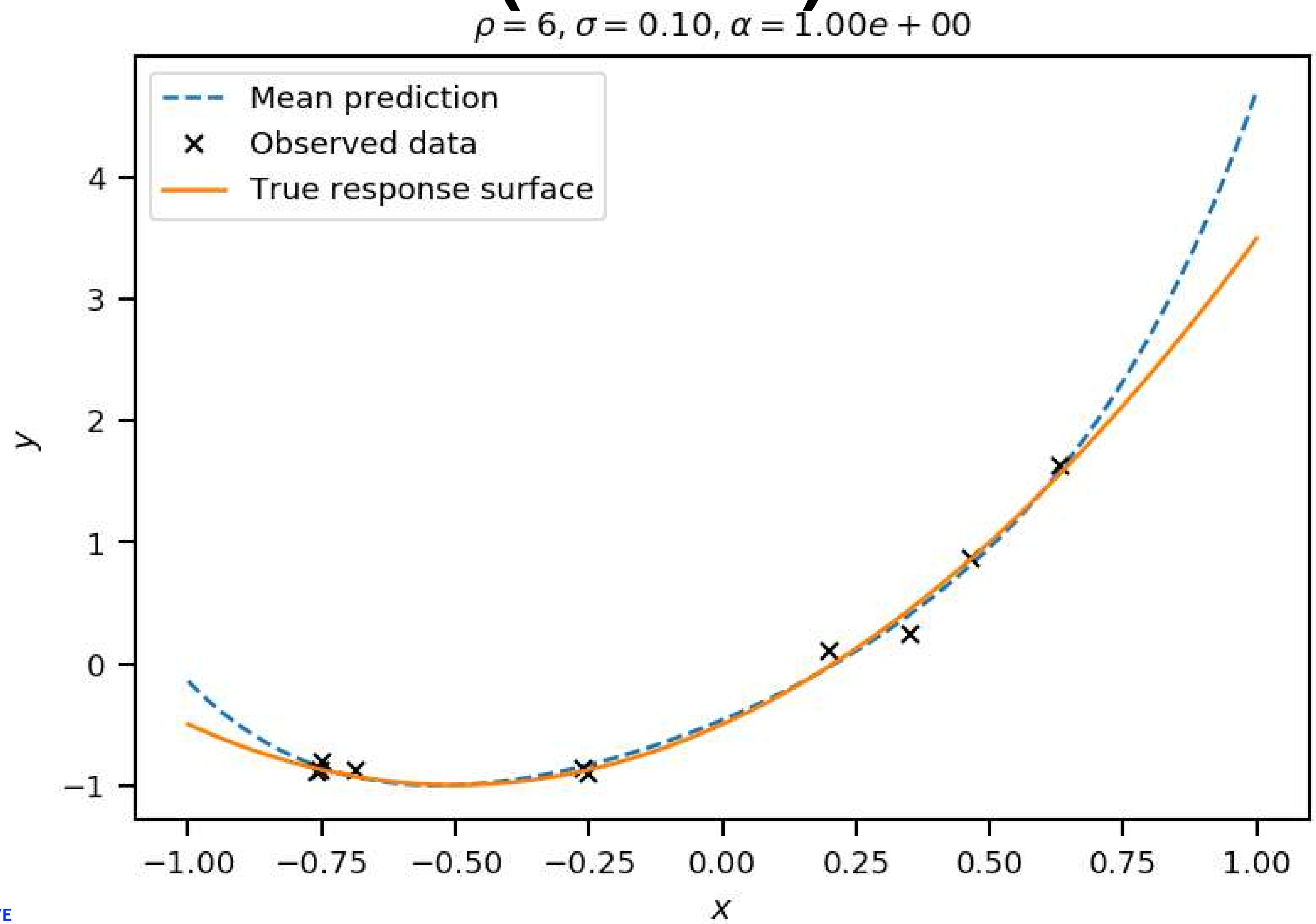
PREDICTIVE SCIENCE LABORATORY

Example: Degree 6 polynomial ($\alpha = 0$)



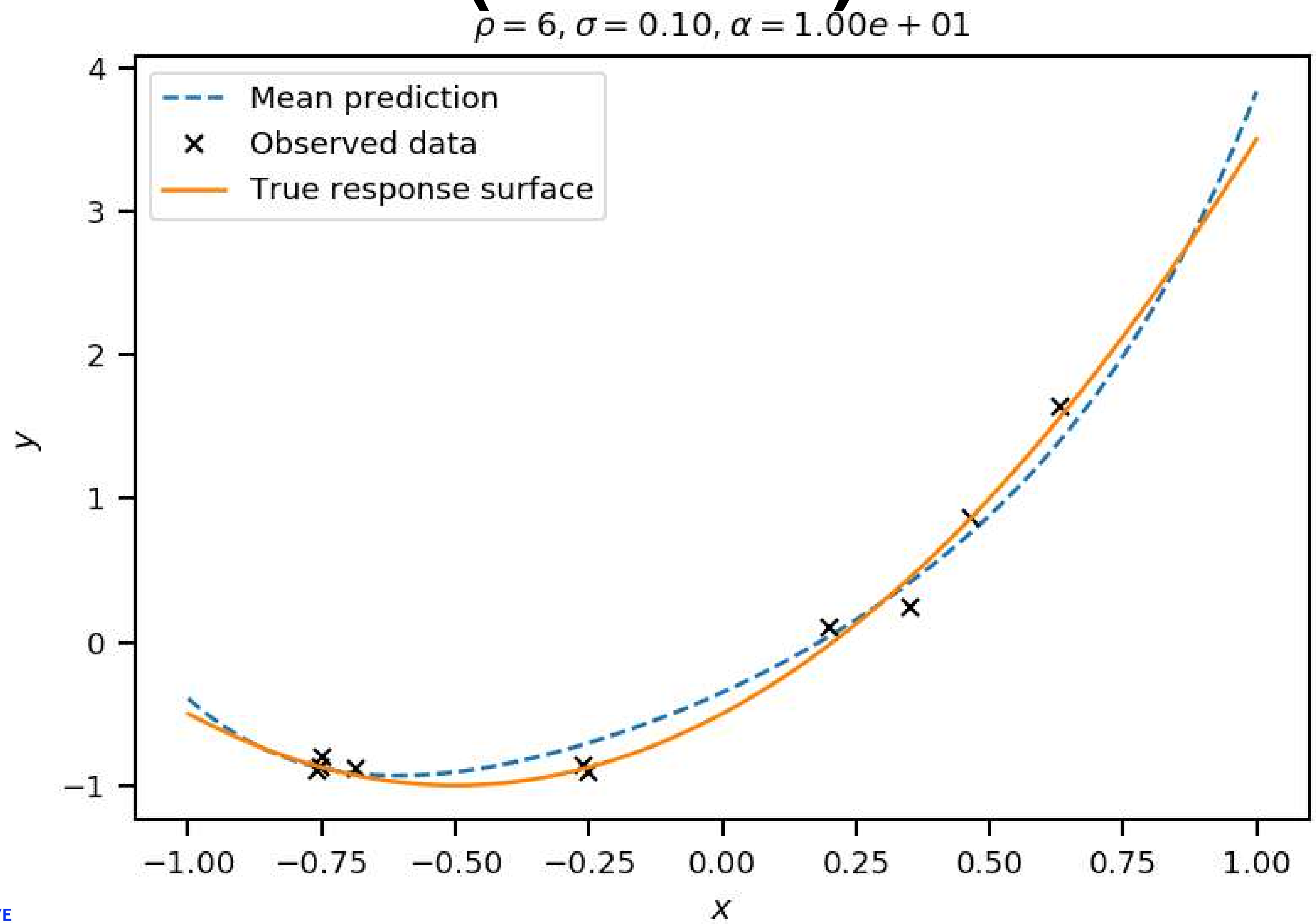
PREDICTIVE SCIENCE LABORATORY

Example: Degree 6 polynomial $(\alpha = 1)$



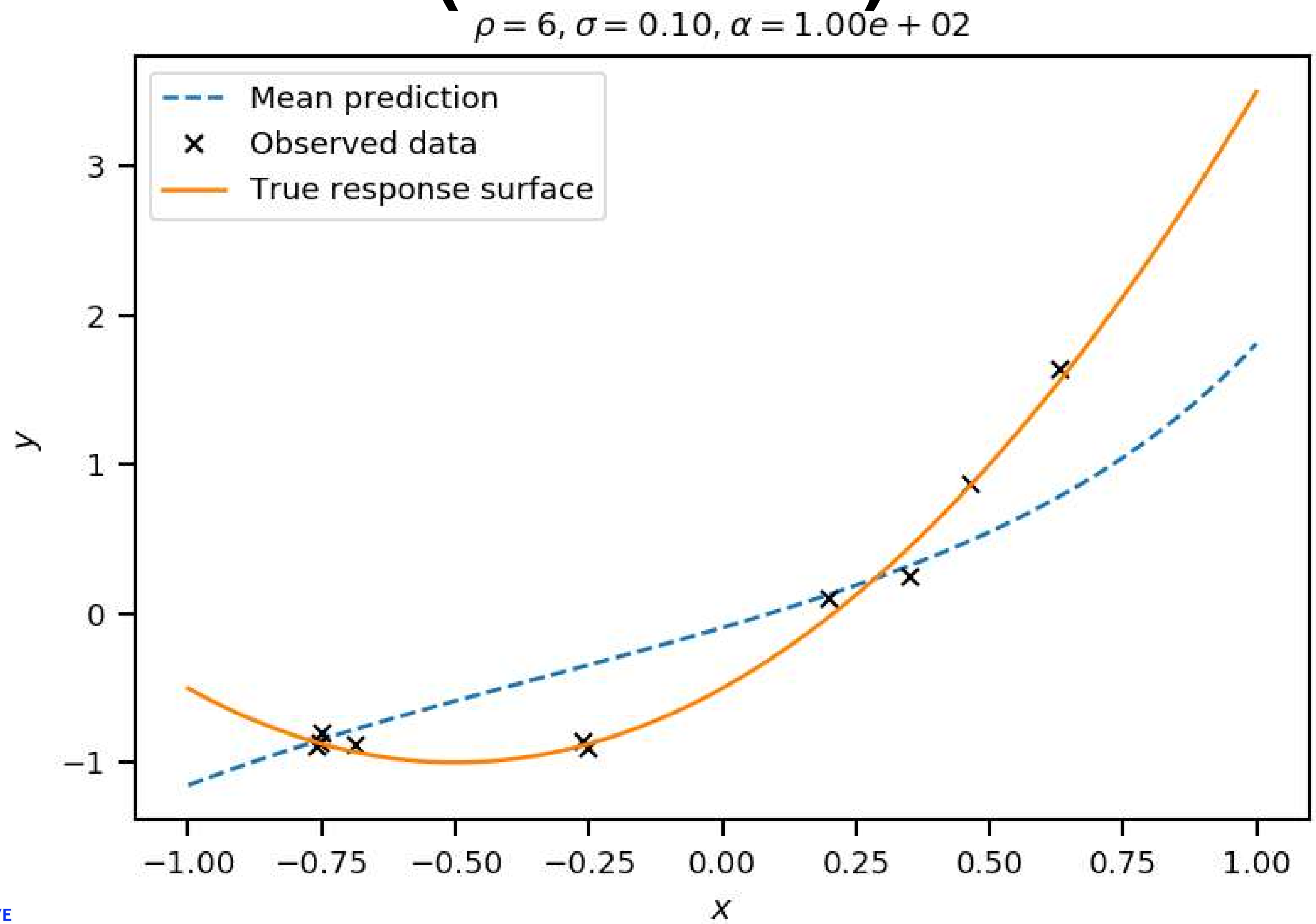
PREDICTIVE SCIENCE LABORATORY

Example: Degree 6 polynomial $(\alpha = 10)$



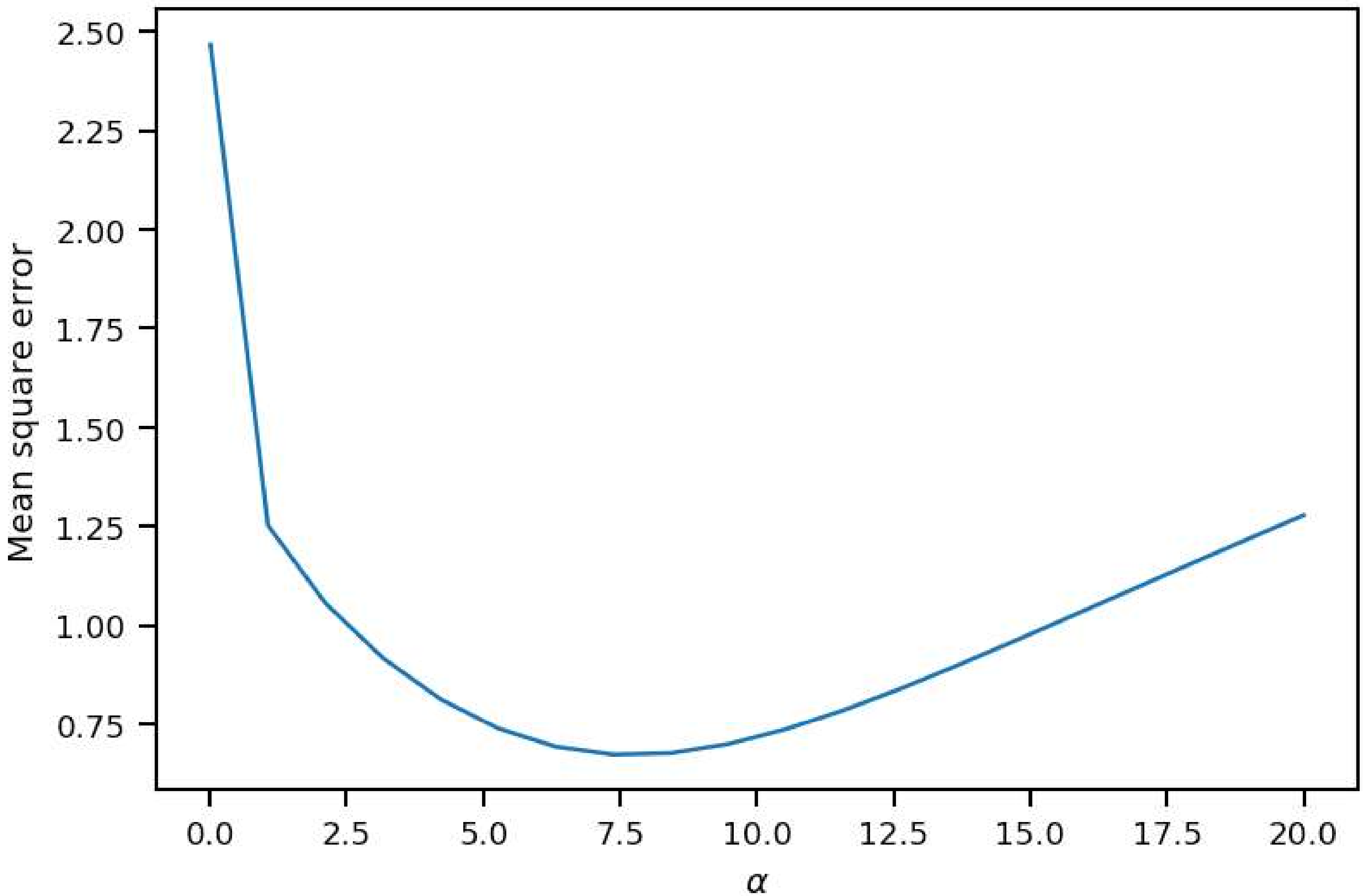
PREDICTIVE
SCIENCE LABORATORY

Example: Degree 6 polynomial ($\alpha = 100$)



PREDICTIVE
SCIENCE LABORATORY

Mean square error over a validation dataset as a function of α



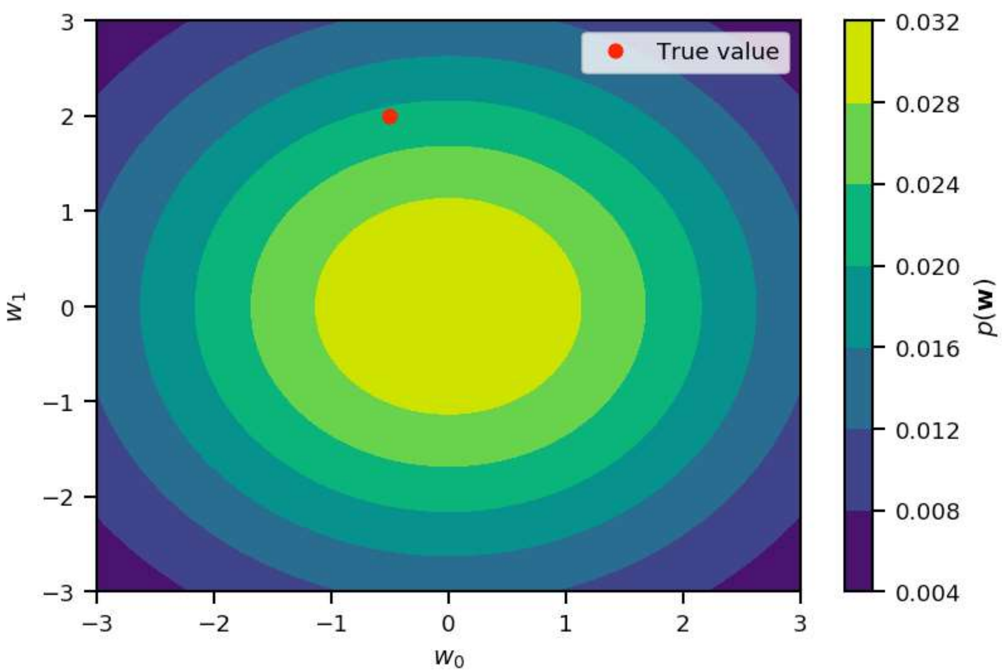
PREDICTIVE SCIENCE LABORATORY

Open questions

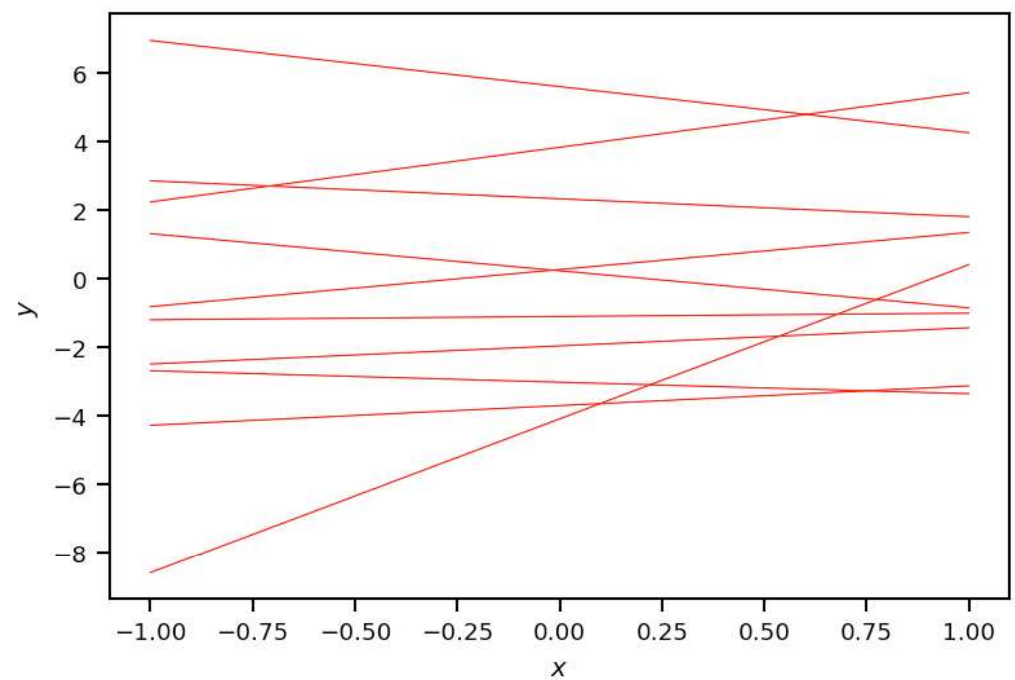
- How do I quantify the measurement noise?
- How do we avoid overfitting?
- How do I quantify epistemic uncertainty induced by limited data?
- How do I choose any remaining parameters?
- How do I choose which basis functions to keep?

Weight prior (linear regression)

Weight space



Model space



Weight posterior

posterior $\propto$ likelihood $\times$ prior

$$p(\underline{w} \mid x_{1:N}, y_{1:N}, \sigma) \propto p(y_{1:N} \mid x_{1:N}, \sigma, \underline{w}) \cdot p(\underline{w})$$

$$= N\left(y_{1:N} \mid \underline{\underline{\Phi}}\, \underline{w}, \sigma^2 \underline{\underline{I}}_N\right) \times N\left(\underline{w} \mid 0, \alpha^{-1} \underline{\underline{I}}_m\right)$$

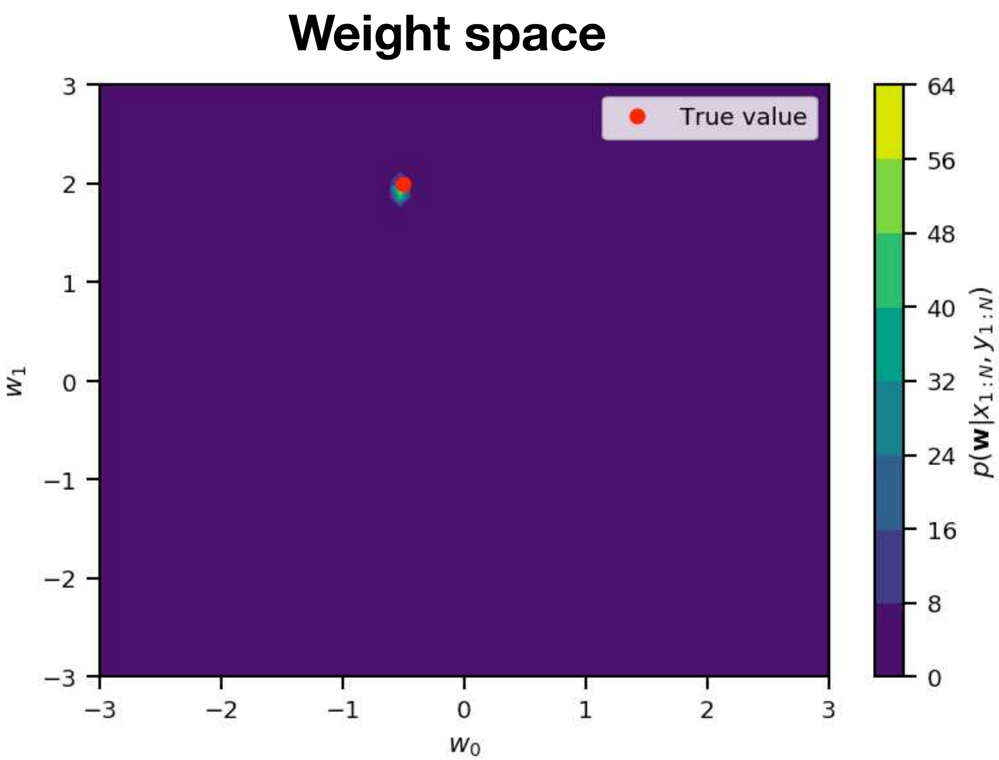
$$\underline{\underline{\Phi}}\, \underline{w} = \begin{pmatrix} \underline{\varphi}^T(x_1)\, \underline{w} \\ \vdots \\ \underline{\varphi}^T(x_N)\, \underline{w} \end{pmatrix}$$

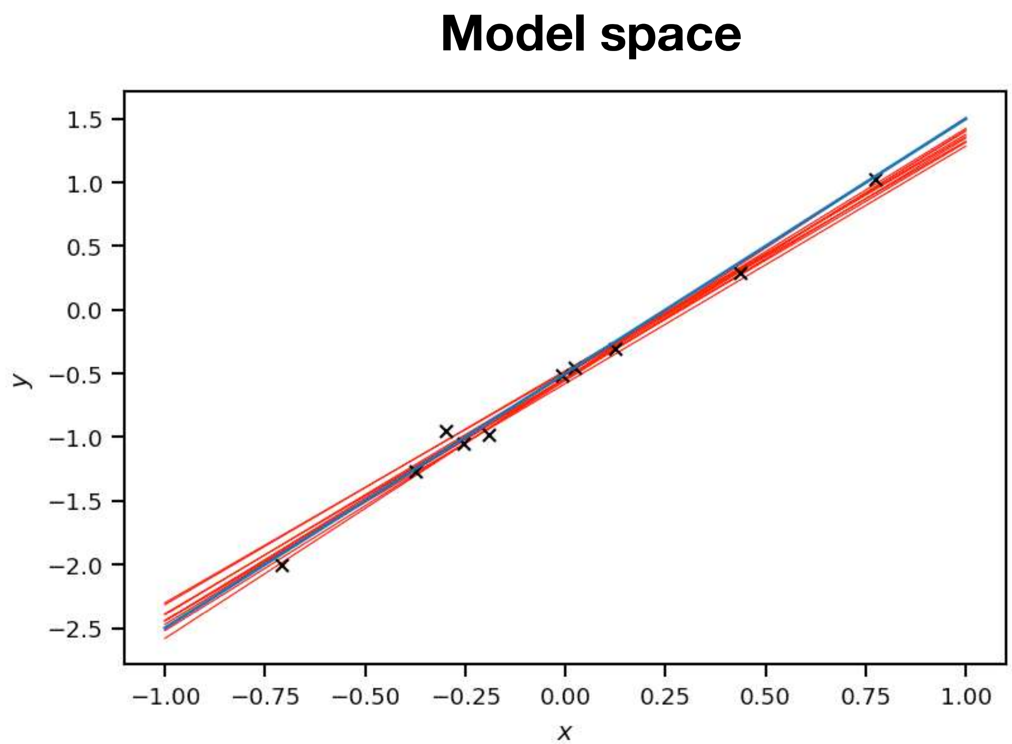
$$= N\left(\underline{w} \mid \underline{m}, \underline{\underline{S}}\right)$$

$$\underline{\underline{S}} = \left(\sigma^{2} \underline{\underline{\Phi}}^T \underline{\underline{\Phi}} + \alpha \underline{\underline{I}}_m\right)^{-1}$$

$$\underline{m} = \sigma^{-2} \underline{\underline{S}}^{-1} \cdot \underline{\underline{\Phi}}^T \underline{y}.$$

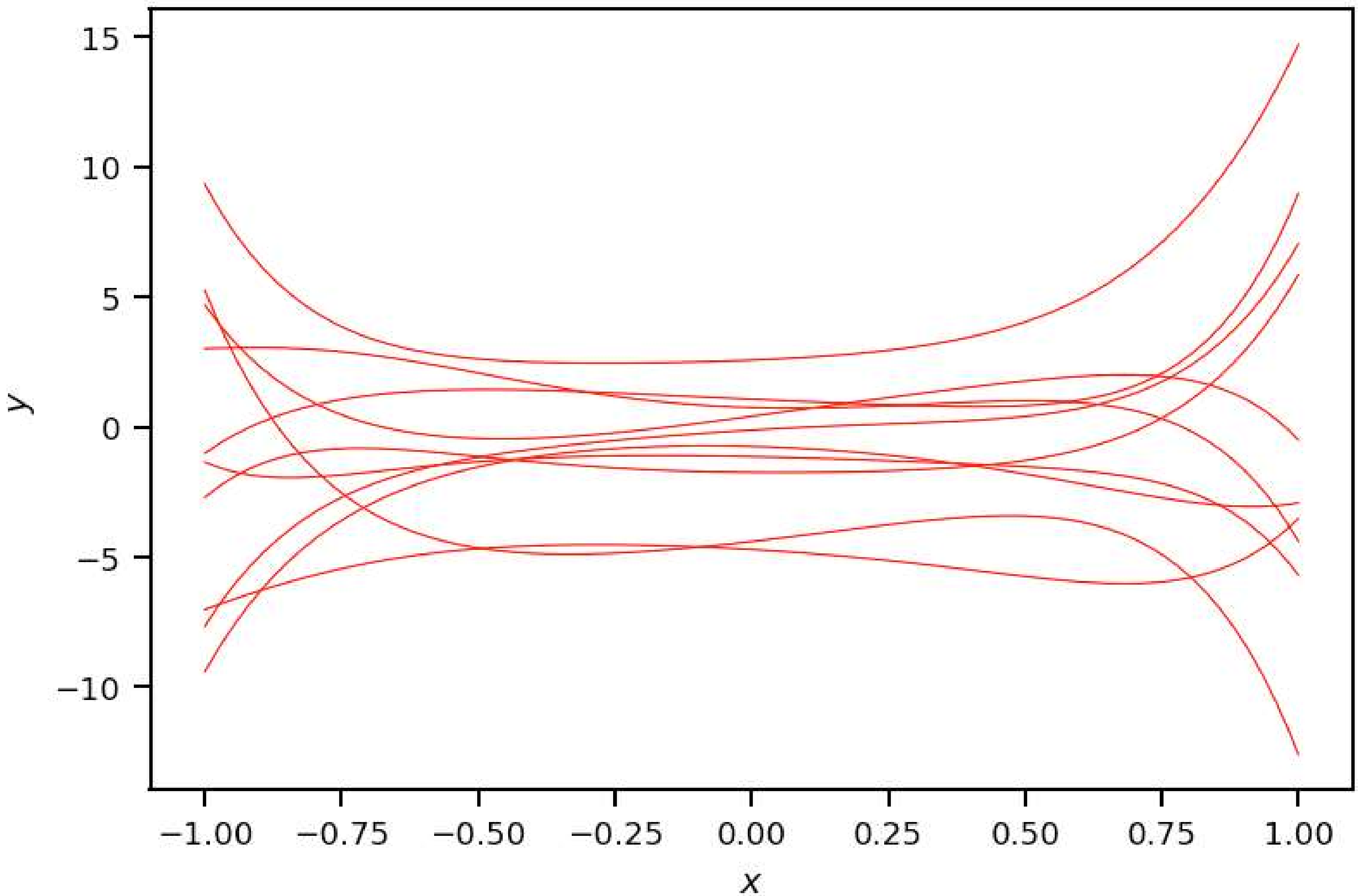
Weight posterior (linear regression)



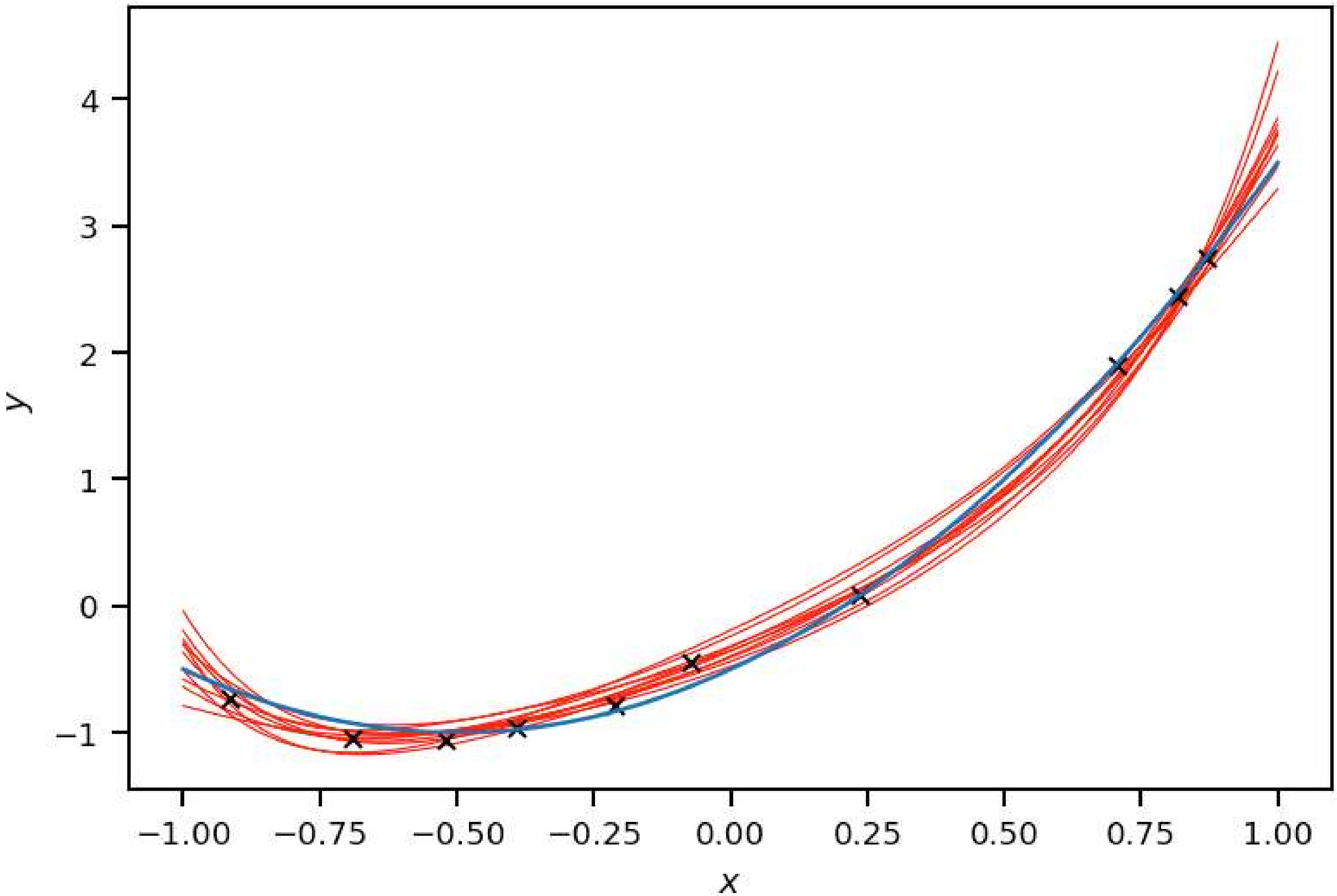


PREDICTIVE
SCIENCE LABORATORY

Example: 7th degree polynomial (prior)



Example: 7th degree polynomial (posterior)



PREDICTIVE SCIENCE LABORATORY

Point-predictive distribution

$$p(y \mid \underline{x}, x_{1:N}, y_{1:N}, a, \sigma^2) = ?$$

$$p(\underline{w} \mid x_{1:N}, y_{1:N}, a, \sigma^2) = \checkmark \quad ; \quad p(y \mid x, \underline{w}, \sigma^2) = \checkmark$$

Sum Rule : $p(A \mid I) = \sum_i p(A \mid B_i, I)\, p(B_i \mid I)$

$$p(B_1 \text{ or } B_2 \text{ or } \dots) = 1, \quad p(B_i, B_j) = 0$$

$$\underbrace{p}_{A}(y \mid \underbrace{x, x_{1:N}, y_{1:N}, a, \sigma^2}_{I}) = \sum_i p(y \mid B_i, I)\, p(B_i \mid I)$$

$$= \int p(y \mid \underline{w}, I)\, p(\underline{w} \mid I)\, d\underline{w}$$

$$p(\underline{w} \mid x_{1:N}, y_{1:N}, a, \sigma^2)$$

$$p(y \mid \underline{w}, x, \cancel{x_{1:N}, y_{1:N}}, \cancel{a}, \sigma^2) = p(y \mid x, \underline{w}, \sigma^2)$$

$$= \int p(y \mid x, \underline{w}, \sigma^2)\, p(\underline{w} \mid x_{1:N}, y_{1:N}, a, \sigma^2)\, d\underline{w}$$

$$= N\left(y \mid \underline{\varphi}^T(x)\underline{m}, \underbrace{\underline{\varphi}^T(x)\underline{\underline{S}}\,\underline{\varphi}(x)}_{\text{epistemic}} + \sigma^2\right)$$

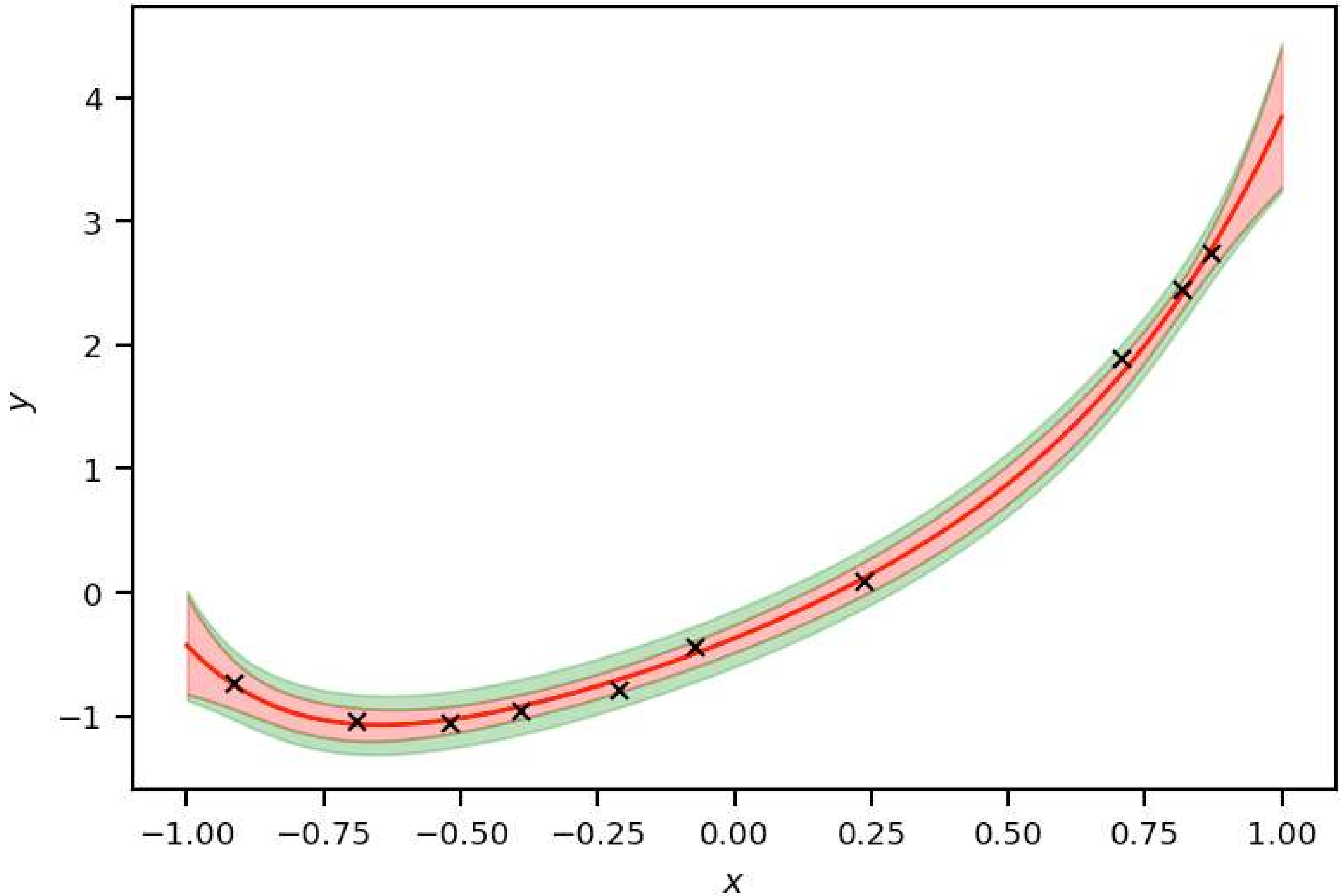
post. mean of weights

post. cov. of weights

noise variance

PREDICTIVE SCIENCE LABORATORY

Example: Separating epistemic and aleatory uncertainties



PREDICTIVE SCIENCE LABORATORY