

# Bayesian linear regression

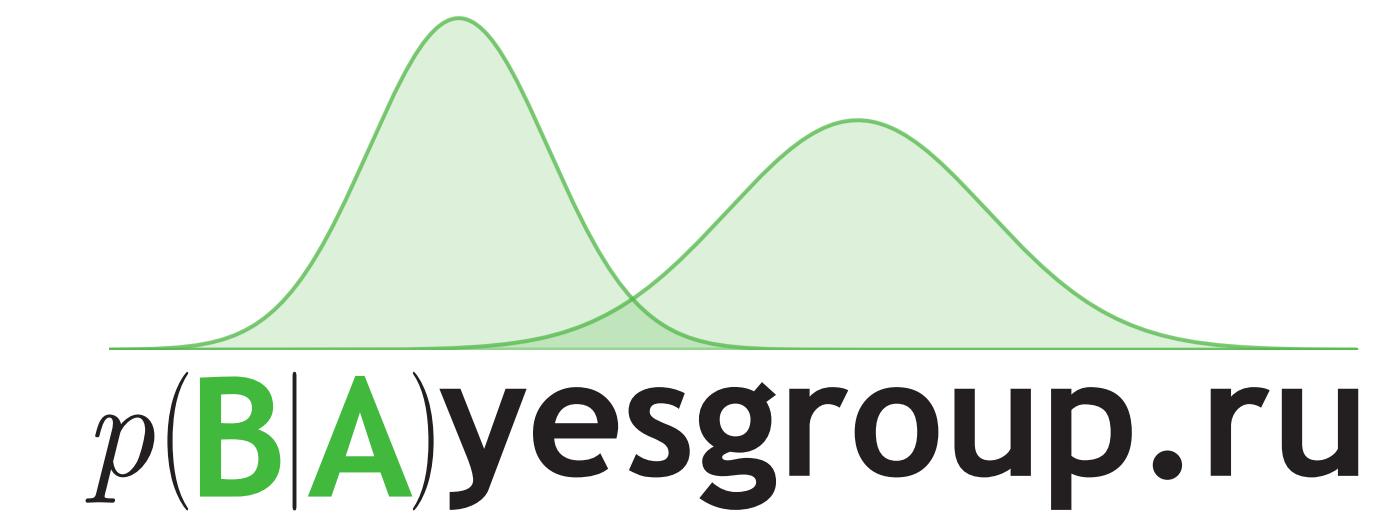
Nadia Chirkova

Higher School of Economics, Samsung-HSE Laboratory  
Moscow, Russia



NATIONAL RESEARCH  
UNIVERSITY

SAMSUNG  
Research



# Plan

- Linear regression: reminder
- Bayesian linear regression:
  - model definition
  - training
  - prediction

# Plan

- Linear regression: reminder
- Bayesian linear regression:
  - model definition
  - training
  - prediction

# Linear regression: reminder

Given:

$X \in \mathbb{R}^{N \times d}$  — input data

$Y \in \mathbb{R}^N$  — target values

$N$  — number of objects

$d$  — number of features

# Linear regression: reminder

Given:

$X \in \mathbb{R}^{N \times d}$  — input data

$Y \in \mathbb{R}^N$  — target values

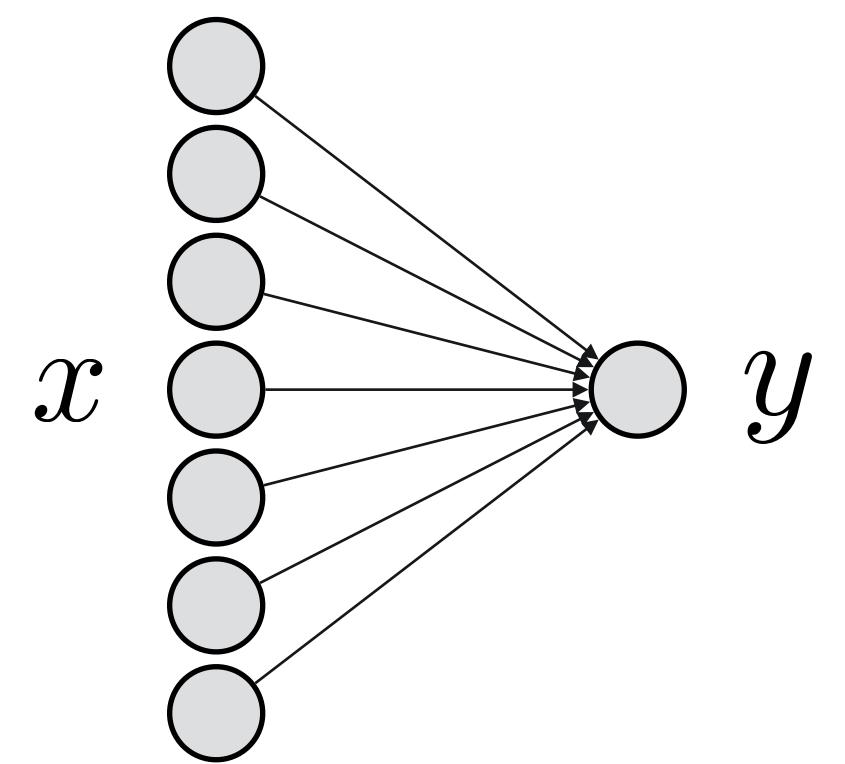
$N$  — number of objects

$d$  — number of features

Model:

$$Xw \approx Y$$

$$x_i^T w \approx y_i$$



linear model  
with weights  $w$

# Linear regression: reminder

Given:

$X \in \mathbb{R}^{N \times d}$  — input data

$Y \in \mathbb{R}^N$  — target values

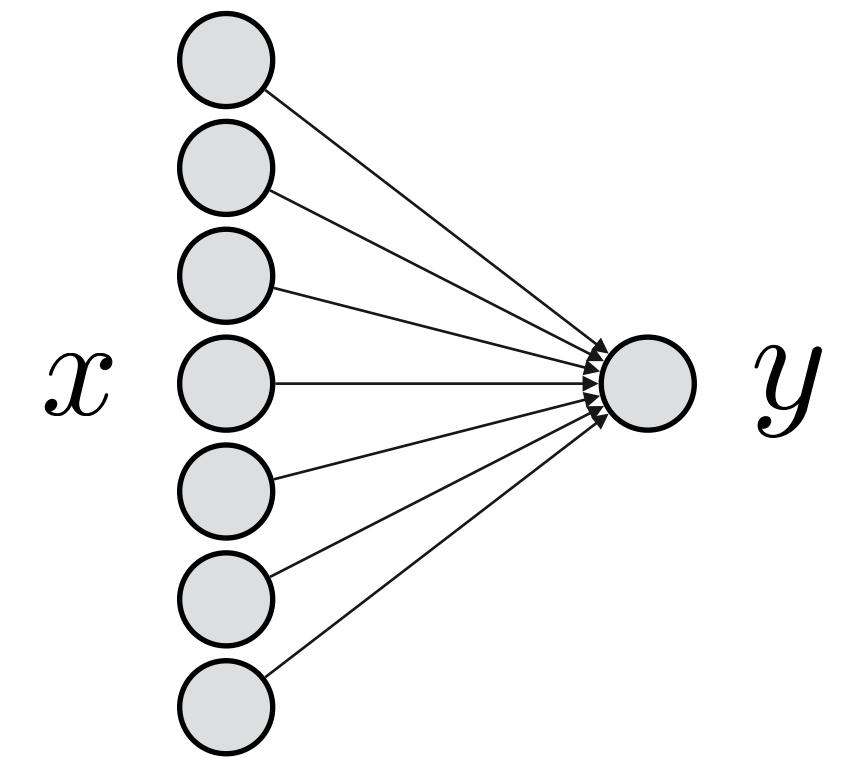
$N$  — number of objects

$d$  — number of features

Model:

$$Xw \approx Y$$

$$x_i^T w \approx y_i$$



linear model  
with weights  $w$

Applications:

- bioinformatics
- physics
- economics
- text processing
- search engines ...

...

# Linear regression: reminder

Given:

$X \in \mathbb{R}^{N \times d}$  — input data

$Y \in \mathbb{R}^N$  — target values

$N$  — number of objects

$d$  — number of features

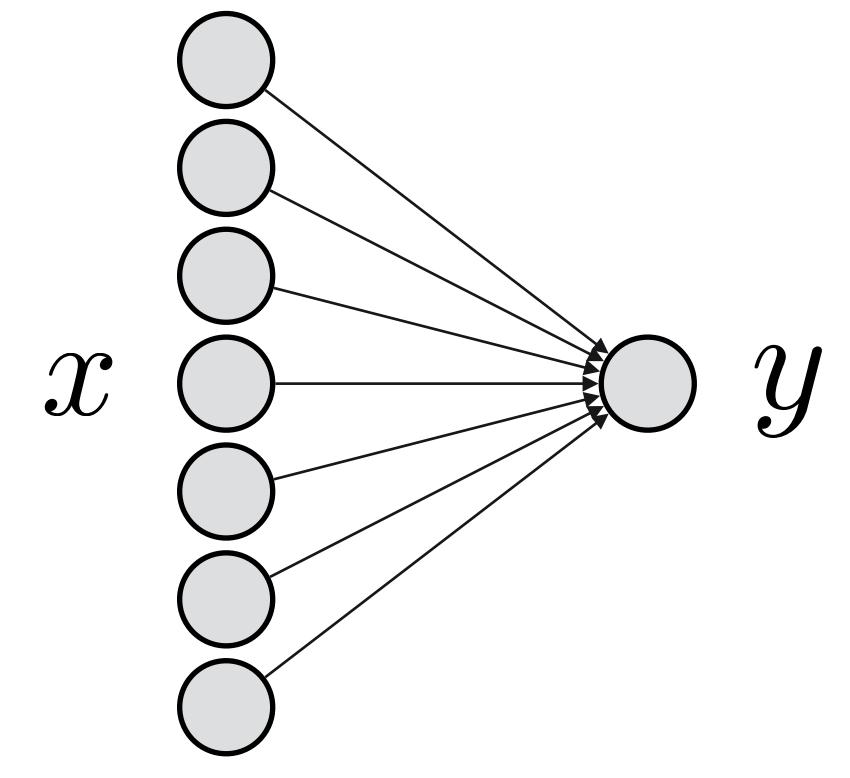
Training:

$$\frac{1}{N} \sum_{i=1}^N (x_i^T w - y_i)^2 \rightarrow \min_{w \in \mathbb{R}^d}$$

Model:

$$Xw \approx Y$$

$$x_i^T w \approx y_i$$



linear model  
with weights  $w$

Prediction on a new object  $x_*$ :

$$a(x_*) = x_*^T w$$

# Linear regression: reminder

Given:

$X \in \mathbb{R}^{N \times d}$  — input data

$Y \in \mathbb{R}^N$  — target values

$N$  — number of objects

$d$  — number of features

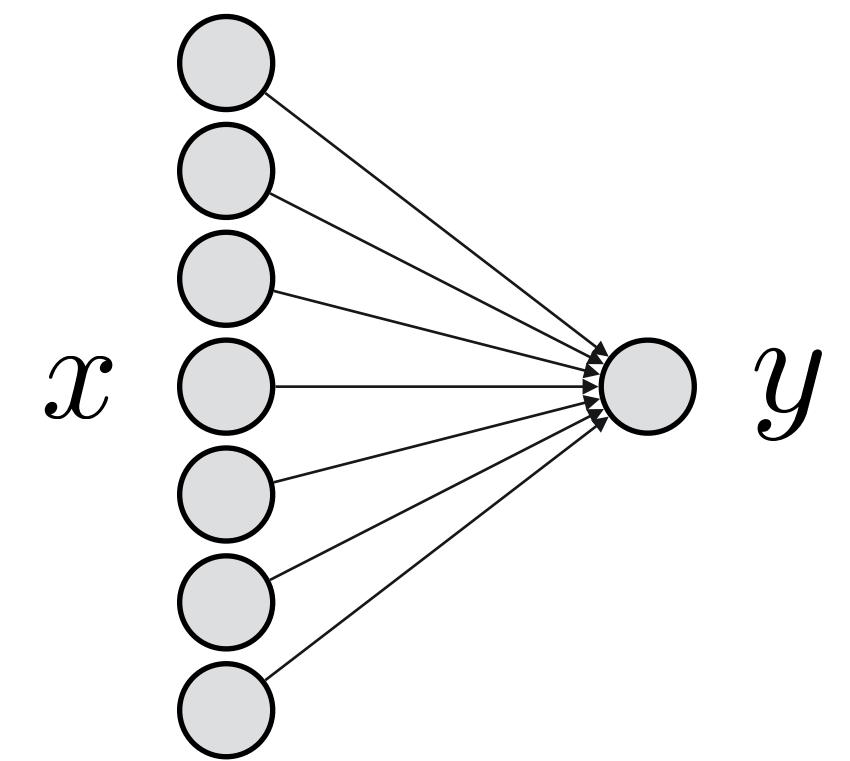
Training:

$$\frac{1}{N} \|Xw - Y\|^2 \rightarrow \min_{w \in \mathbb{R}^d}$$

Model:

$$Xw \approx Y$$

$$x_i^T w \approx y_i$$



linear model  
with weights  $w$

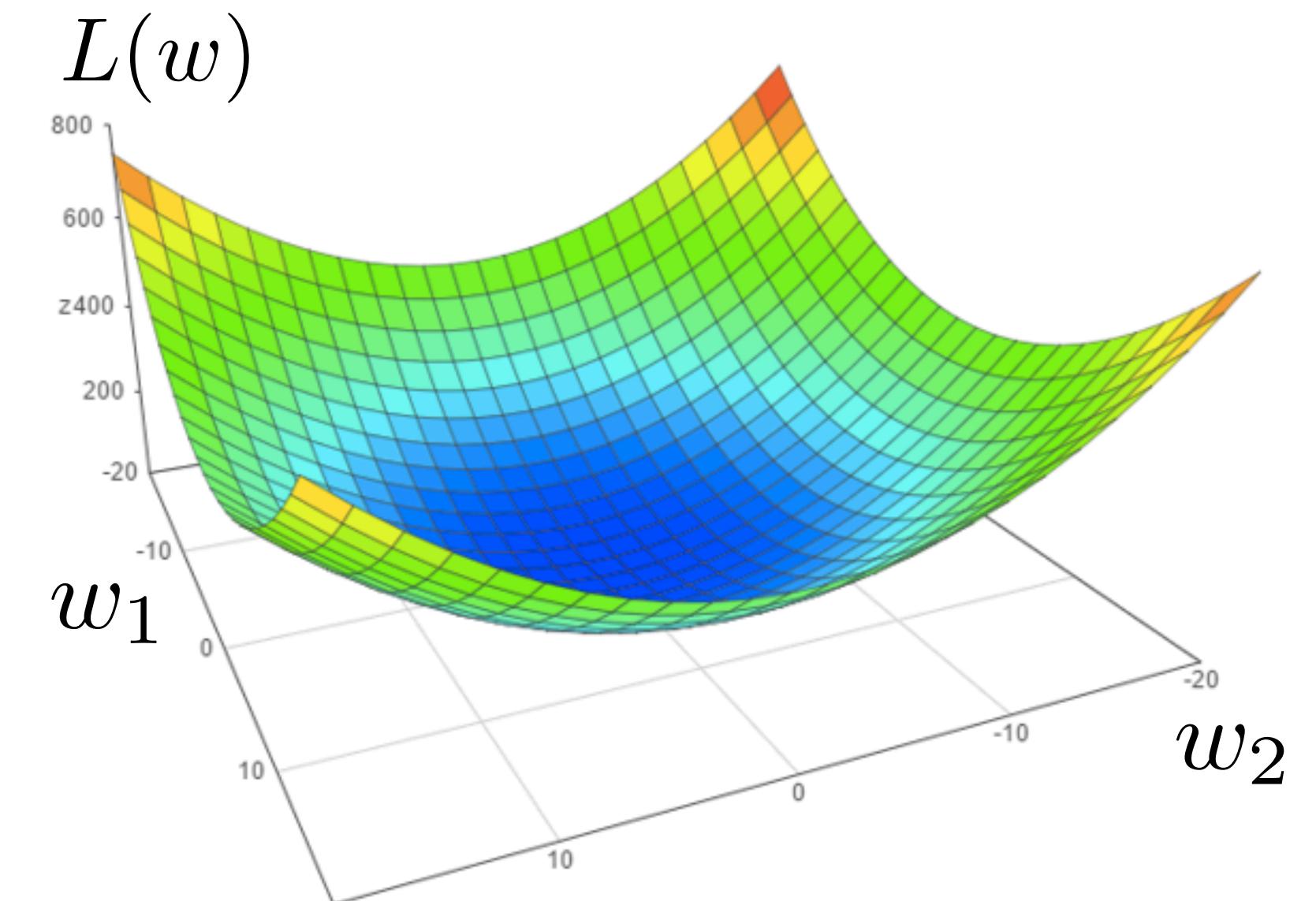
Prediction on a new object  $x_*$  :

$$a(x_*) = x_*^T w$$

# Linear regression: training

$$L(w) = \frac{1}{N} \|Xw - Y\|^2 \rightarrow \min_{w \in \mathbb{R}^d}$$

Convex function:



# Linear regression: training

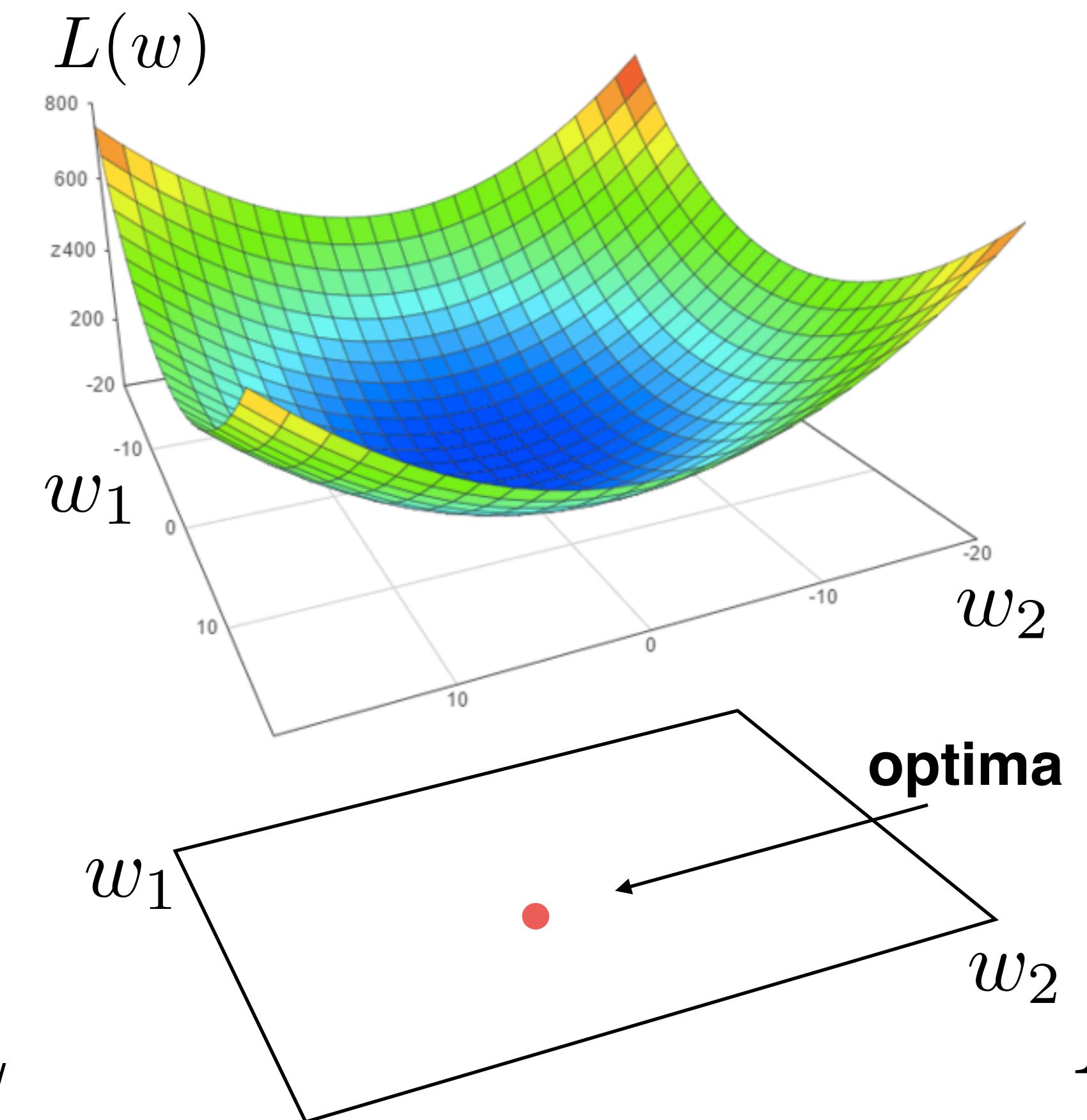
$$L(w) = \frac{1}{N} \|Xw - Y\|^2 \rightarrow \min_{w \in \mathbb{R}^d}$$

Optimal weights:

$$w_{ML} = (X^T X)^{-1} X^T Y$$

— if  $\text{rank}(X^T X) = d$ ,  
otherwise infinite number of solutions

Convex function:



# Linear regression: regularization

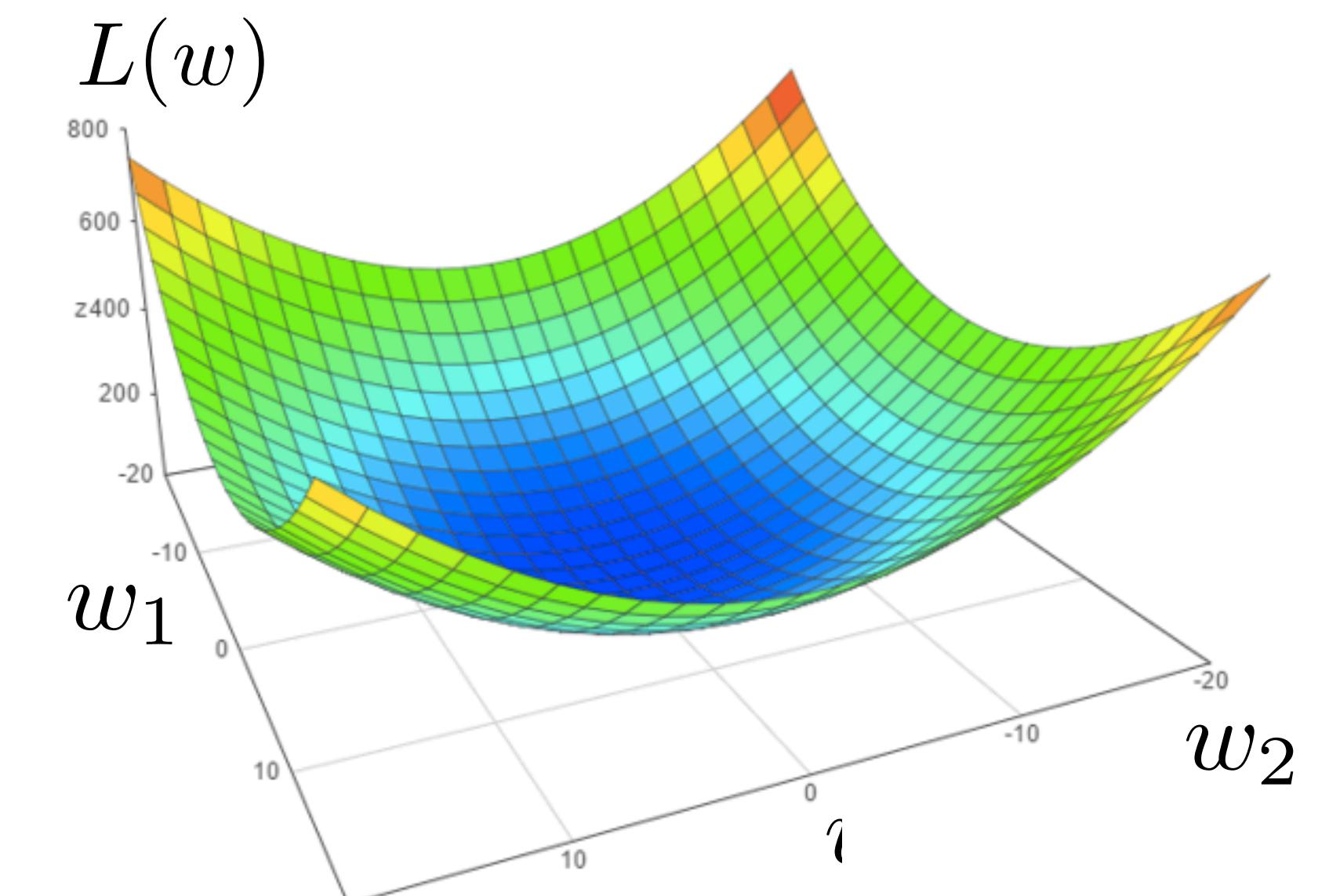
$$L(w) = \frac{1}{N} \|Xw - Y\|^2 + \lambda \|w\|^2 \rightarrow \min_{w \in \mathbb{R}^d} \quad \lambda > 0$$

Optimal weights:

$$w_{MP} = (X^T X + \lambda I)^{-1} X^T Y$$

- Always unique solution
- Preventing overfitting

Strongly convex function:



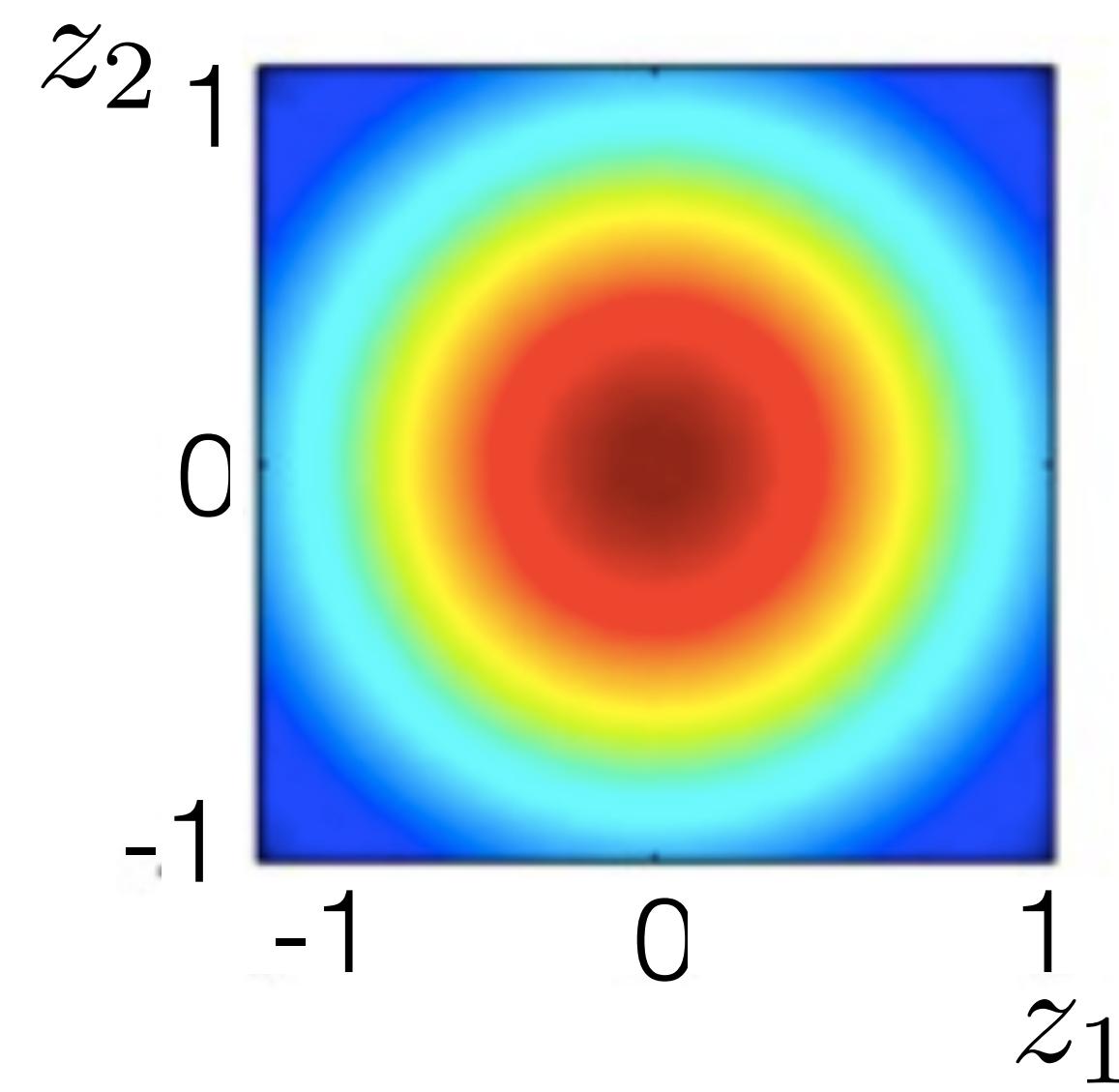
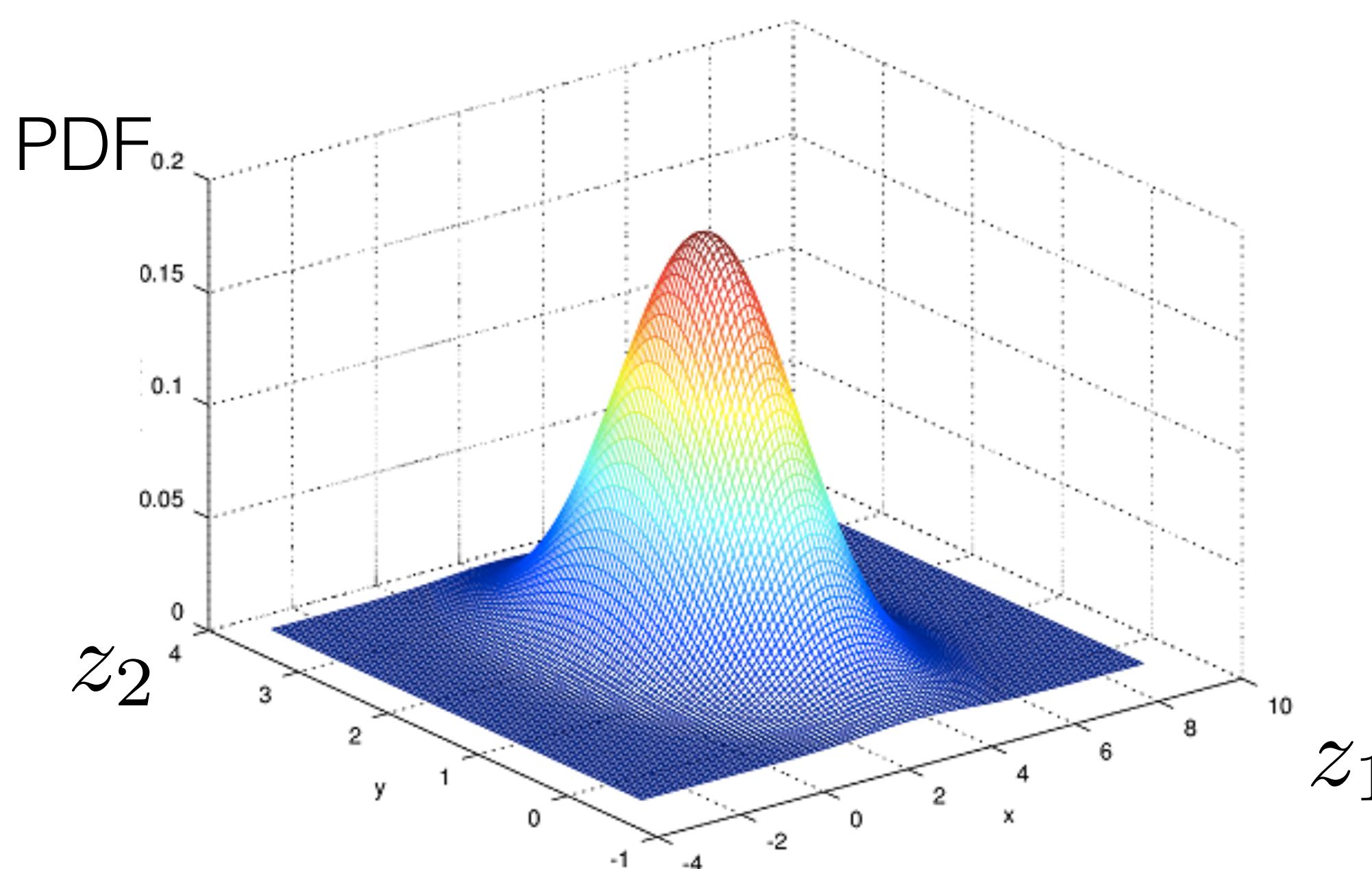
# Plan

- Linear regression: reminder
- Bayesian linear regression:
  - model definition
  - training
  - prediction

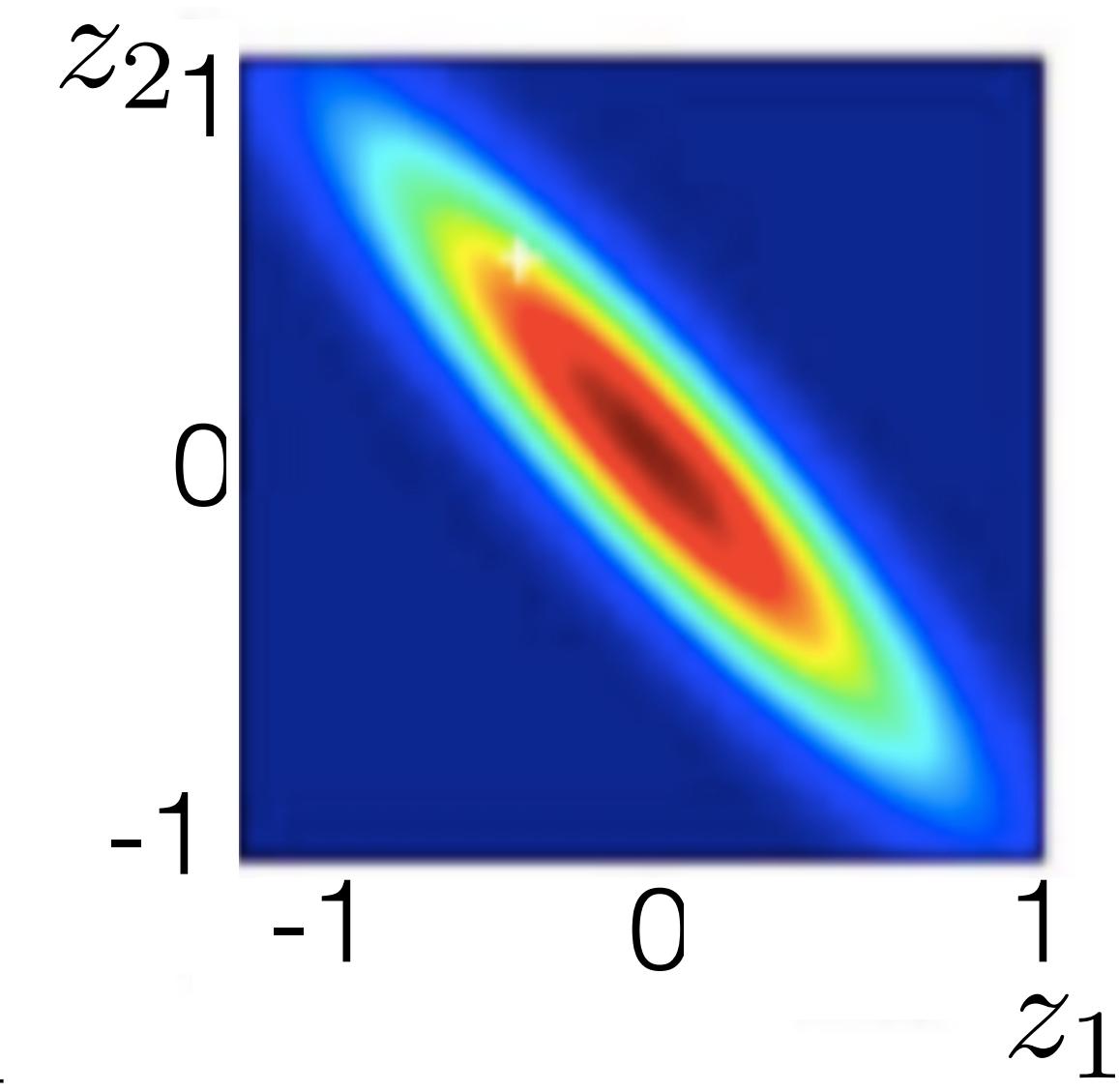
# Multivariate normal (Gaussian) distribution

$$\mathcal{N}(\mu, \Sigma) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2}(z - \mu)^T \Sigma^{-1} (z - \mu)\right),$$

$$\begin{aligned} z &\in \mathbb{R}^d \\ \mu &\in \mathbb{R}^d \\ \Sigma &\in \mathbb{R}^{d \times d} \end{aligned}$$



diagonal  $\Sigma$



non-diagonal  $\Sigma$

# Bayesian linear regression

Given:

$X \in \mathbb{R}^{N \times d}$  — input data

$Y \in \mathbb{R}^N$  — target values

$N$  — number of objects

$d$  — number of features

Model:

$$p(Y, w|X) = p(Y|X, w)p(w)$$

**how does target Y  
depend on input X?**

**what weights w  
do we expect?**

# Bayesian linear regression

Given:

$X \in \mathbb{R}^{N \times d}$  — input data

$Y \in \mathbb{R}^N$  — target values

$N$  — number of objects

$d$  — number of features

Model:

$$p(Y, w|X) = p(Y|X, w)p(w)$$

- likelihood:

$$\begin{aligned} p(Y|X, w) &= \prod_{i=1}^N \mathcal{N}(y_i | \underbrace{x_i^T w}_1, 1) = \\ &= \mathcal{N}(Y | \underbrace{Xw}_I, I) \end{aligned}$$

- prior ?

# Bayesian linear regression

Given:

$X \in \mathbb{R}^{N \times d}$  — input data

$Y \in \mathbb{R}^N$  — target values

$N$  — number of objects

$d$  — number of features

Model:

$$p(Y, w|X) = p(Y|X, w)p(w)$$

- likelihood:

$$\begin{aligned} p(Y|X, w) &= \prod_{i=1}^N \mathcal{N}(y_i|x_i^T w, 1) = \\ &= \mathcal{N}(Y|Xw, I) \end{aligned}$$

- conjugate prior:

$$p(w) = \mathcal{N}(w|0, \alpha I), \alpha > 0$$

# Bayesian linear regression

Given:

$X \in \mathbb{R}^{N \times d}$  — input data

$Y \in \mathbb{R}^N$  — target values

$N$  — number of objects

$d$  — number of features

Training? Prediction?

Model:

$$p(Y, w|X) = p(Y|X, w)p(w)$$

- likelihood:

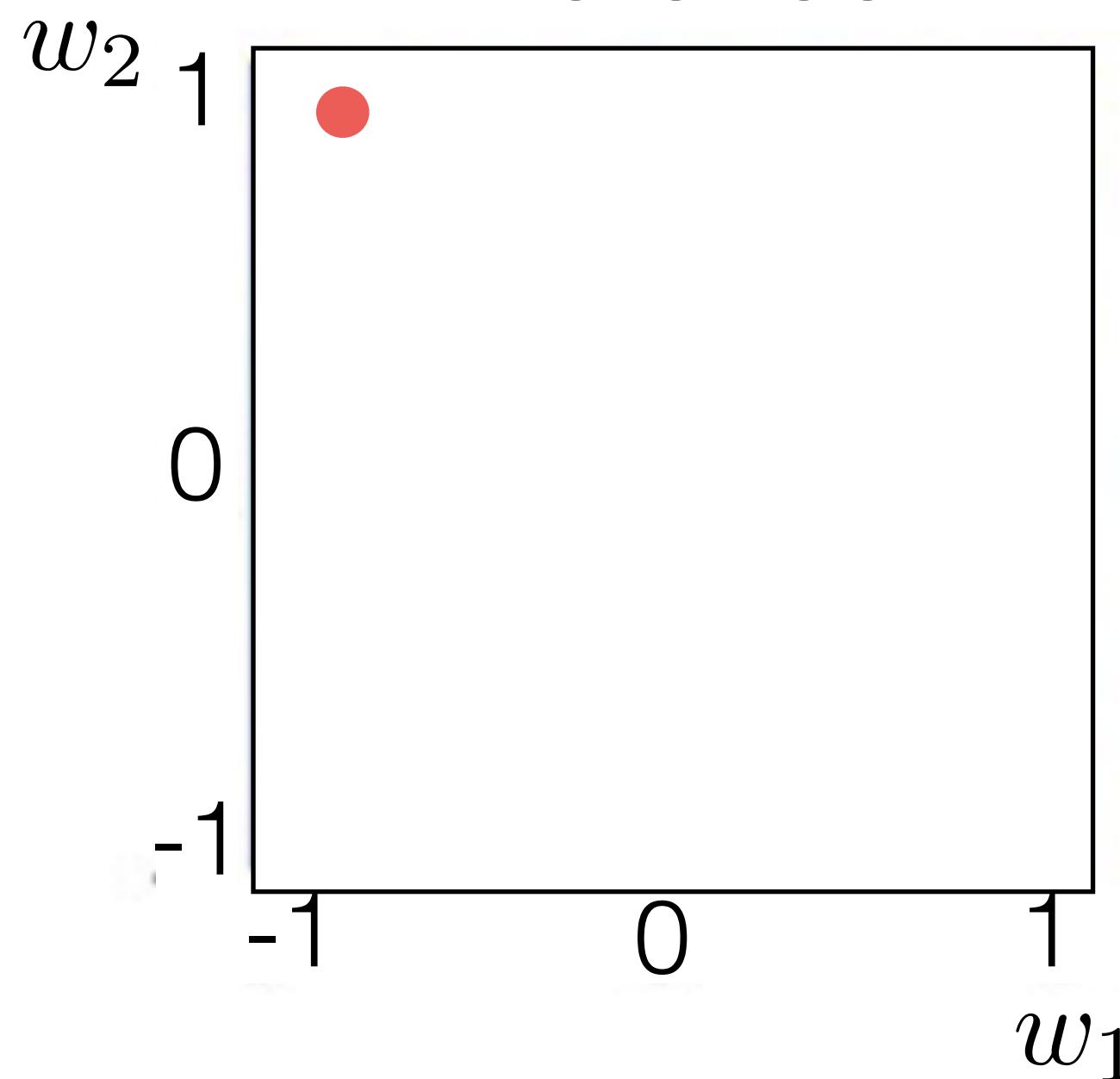
$$\begin{aligned} p(Y|X, w) &= \prod_{i=1}^N \mathcal{N}(y_i|x_i^T w, 1) = \\ &= \mathcal{N}(Y|Xw, I) \end{aligned}$$

- conjugate prior:

$$p(w) = \mathcal{N}(w|0, \alpha I), \alpha > 0$$

# Options of training Bayesian linear regression

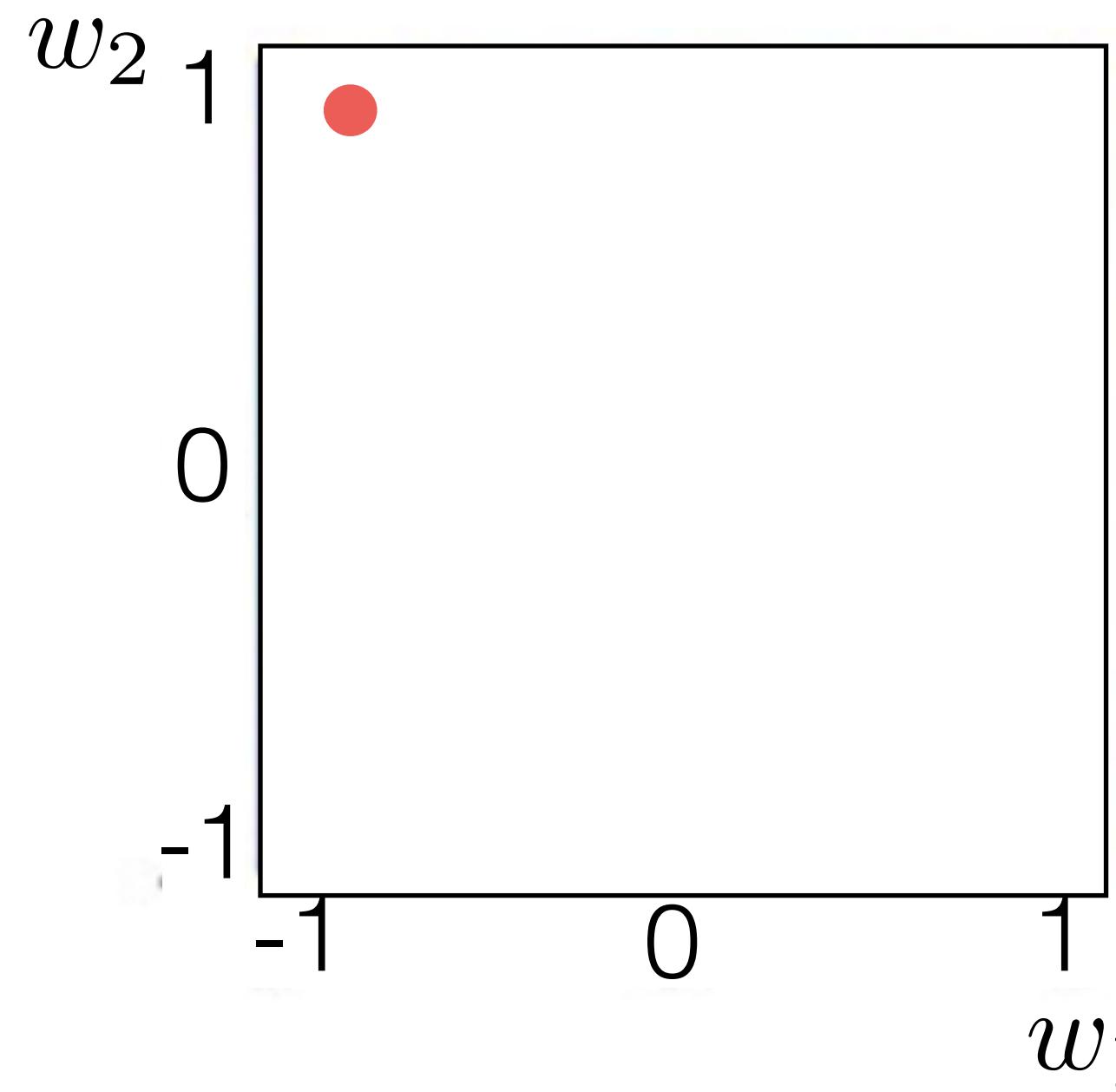
Maximum likelihood  
inference



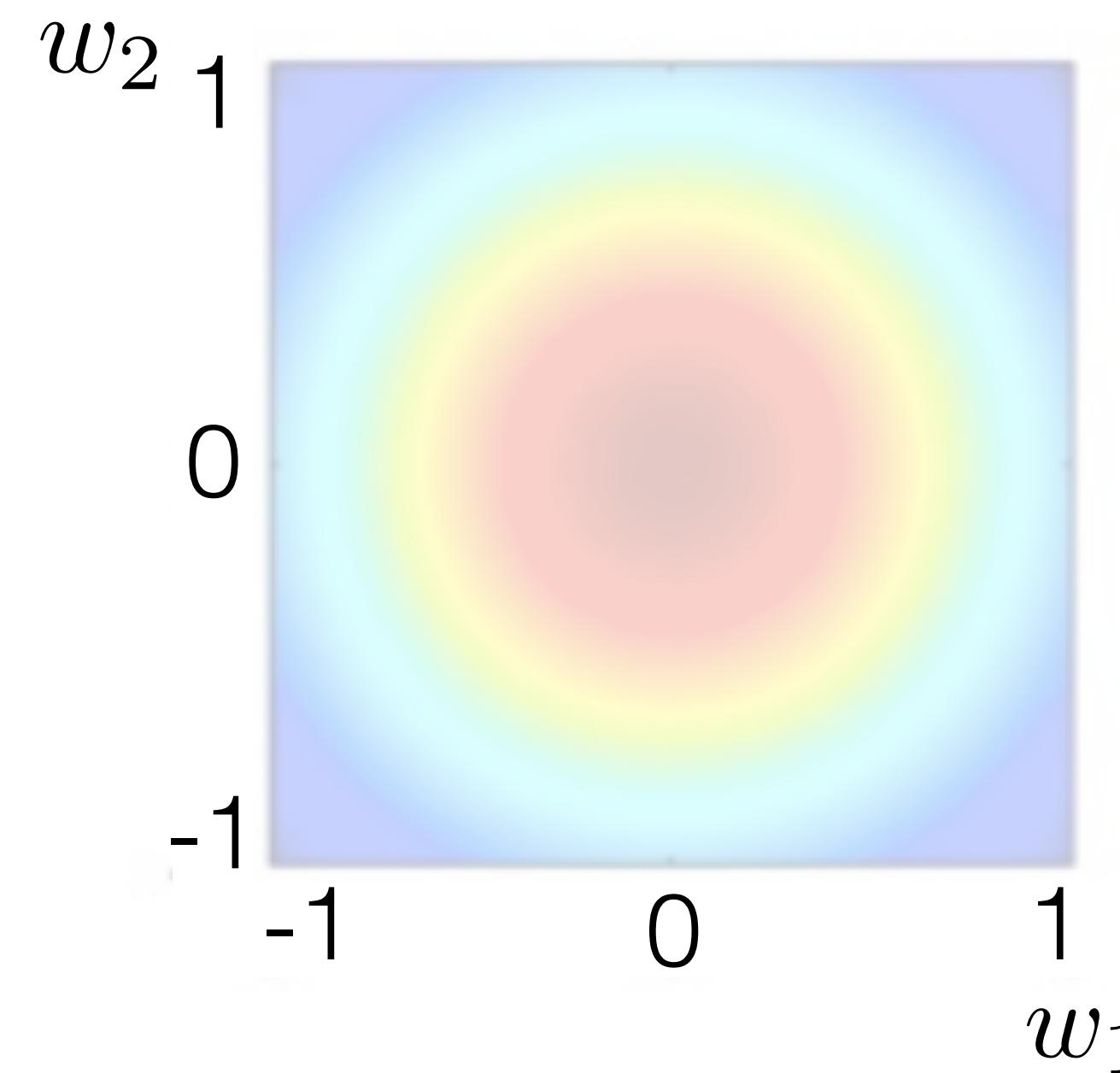
$$p(Y|X, w) \rightarrow \max_w$$

# Options of training Bayesian linear regression

Maximum likelihood  
inference



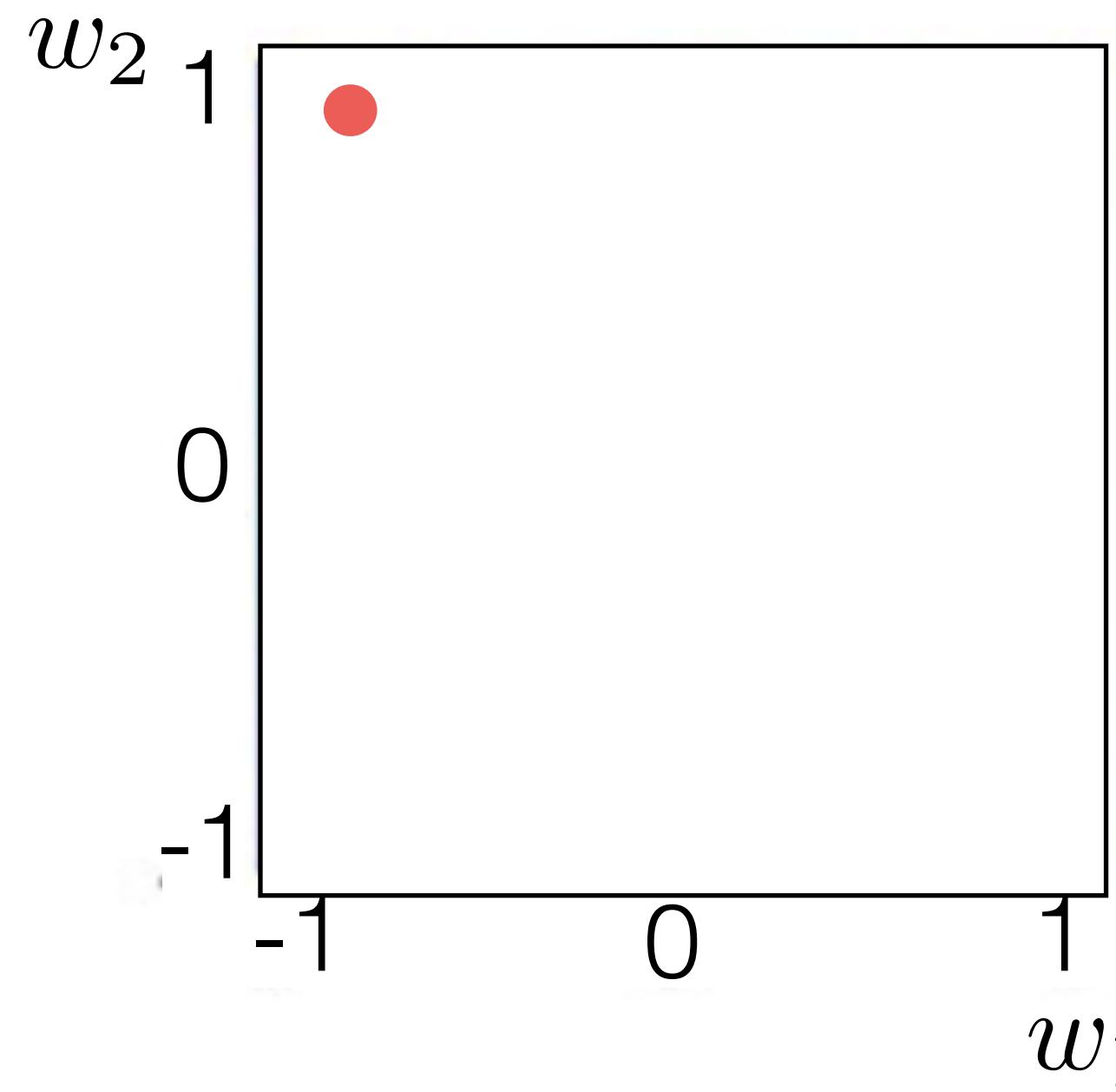
Prior distribution



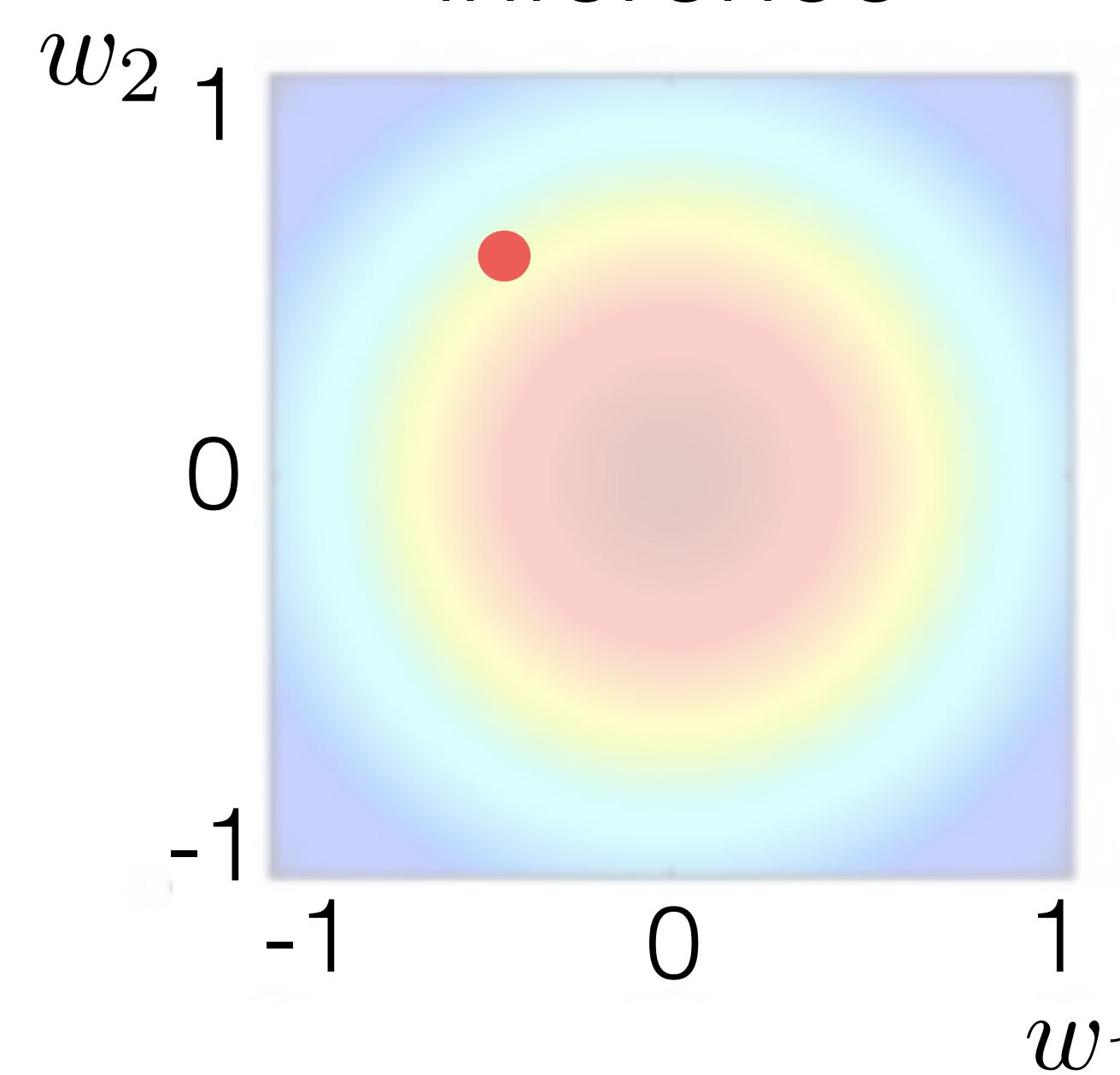
$$p(Y|X, w) \rightarrow \max_w$$

# Options of training Bayesian linear regression

Maximum likelihood  
inference



Maximum posterior  
inference

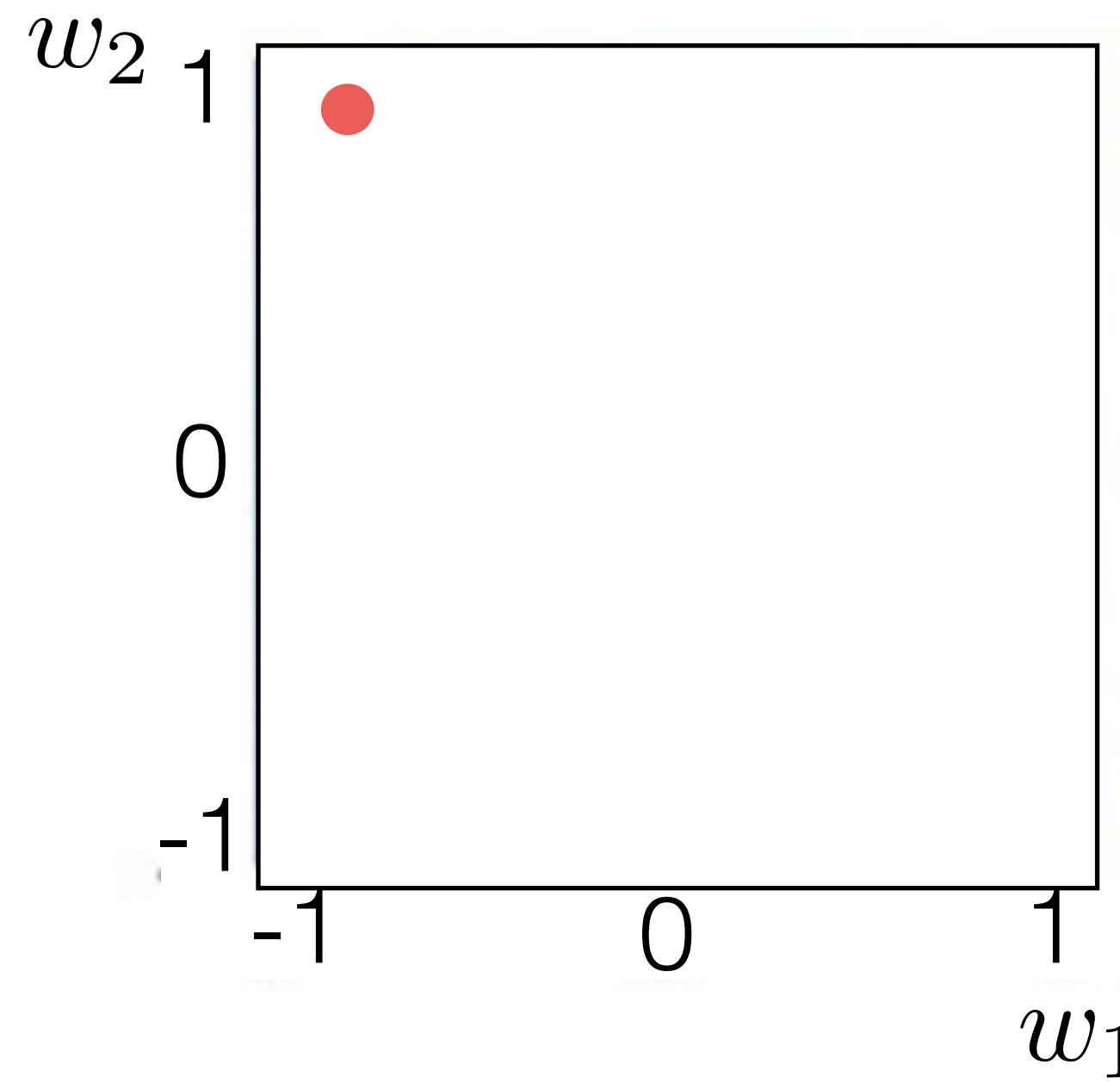


$$p(Y|X, w) \rightarrow \max_w$$

$$p(Y|X, w)p(w) \rightarrow \max_w$$

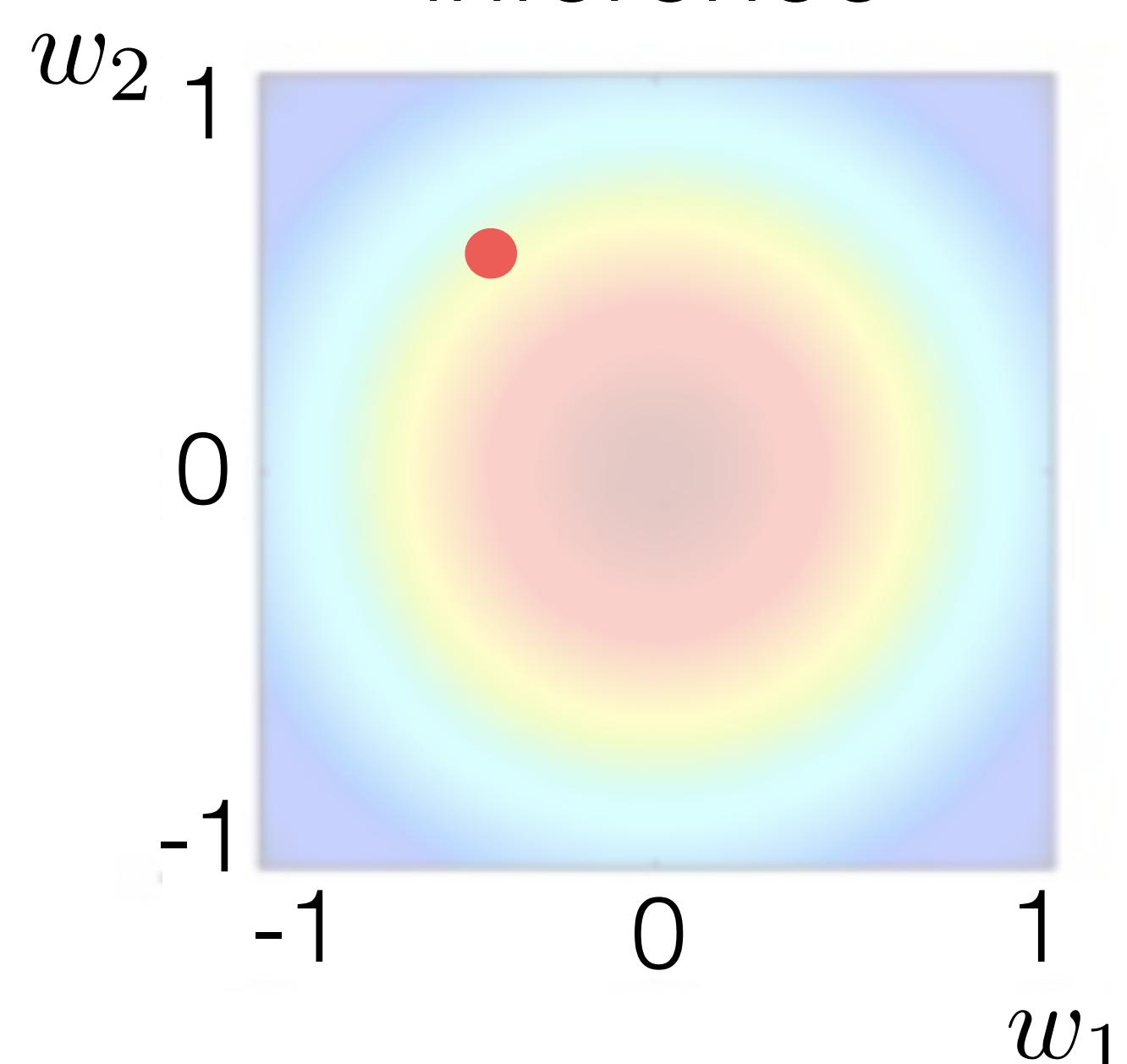
# Options of training Bayesian linear regression

Maximum likelihood  
inference



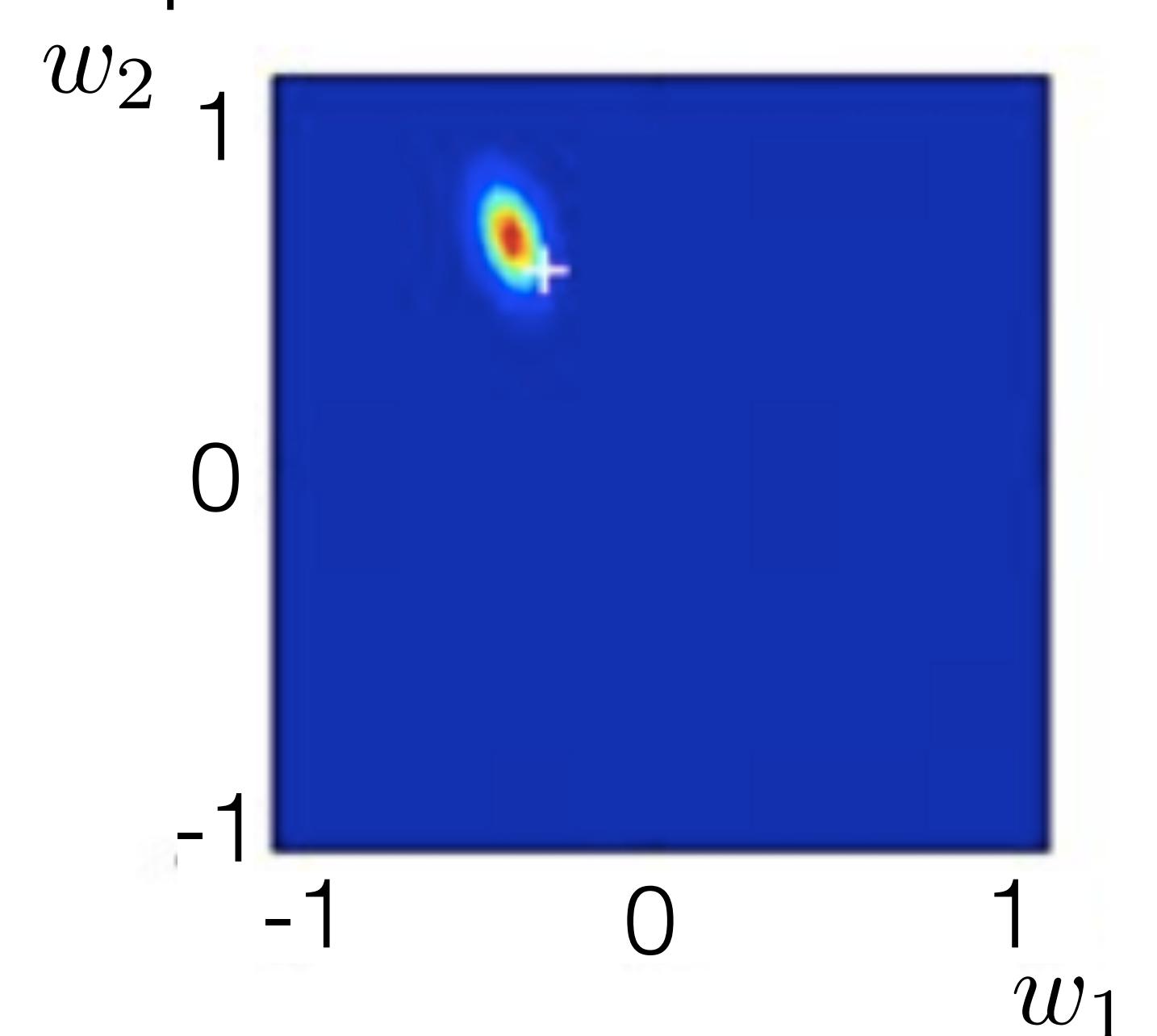
$$p(Y|X, w) \rightarrow \max_w$$

Maximum posterior  
inference



$$p(Y|X, w)p(w) \rightarrow \max_w$$

Full Bayesian inference:  
posterior distribution



$$\begin{aligned} p(w|X, Y) &\propto \\ &\propto p(Y|X, w)p(w) \end{aligned}$$

# Training methods: summary

Probabilistic model:  $p(Y, w|X)$

We want to compute:  $p(w|X, Y)$

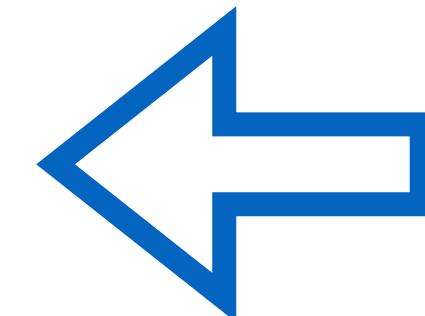
Approximation		Inference
Exact	$p(w X, Y)$	Full Bayesian inference
Parametric	$p(w X, Y) \approx q(w \lambda)$	Parametric Var. Inference
Delta function	$p(w X, Y) \approx \delta(w_{MP})$	Max. posterior inference
No prior	$w_{ML}$	Max. likelihood inference

# Training methods: summary

Probabilistic model:  $p(Y, w|X)$

We want to compute:  $p(w|X, Y)$

Approximation		Inference
Exact	$p(w X, Y)$	Full Bayesian inference
Parametric	$p(w X, Y) \approx q(w \lambda)$	Parametric Var. Inference
Delta function	$p(w X, Y) \approx \delta(w_{MP})$	Max. posterior inference
No prior	$w_{ML}$	Max. likelihood inference



# Bayesian linear regression: training (ML)

Maximum likelihood inference:  $\log p(Y|X, w) \rightarrow \max_w$

# Bayesian linear regression: training (ML)

Maximum likelihood inference:  $\log p(Y|X, w) \rightarrow \max_w$

Likelihood:  $p(Y|X, w) = \mathcal{N}(Y|Xw, I)$

$$\log p(Y|X, w) = \text{Const} - \frac{1}{2} \|Y - Xw\|^2 \rightarrow \max_w$$

$$* z^T z = \|z\|^2$$

# Bayesian linear regression: training (ML)

Maximum likelihood inference:  $\log p(Y|X, w) \rightarrow \max_w$

Likelihood:  $p(Y|X, w) = \mathcal{N}(Y|Xw, I)$

$$\log p(Y|X, w) = \text{Const} - \frac{1}{2} \|Y - Xw\|^2 \rightarrow \max_w$$

$\Updownarrow$

$$\|Y - Xw\|^2 \rightarrow \min_w \quad \text{— we already know the solution!}$$

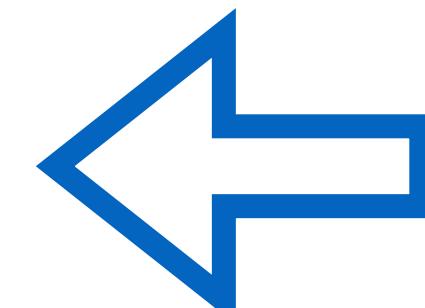
$$w_{ML} = (X^T X)^{-1} X^T Y$$

# Training methods: summary

Probabilistic model:  $p(Y, w|X)$

We want to compute:  $p(w|X, Y)$

Approximation		Inference
Exact	$p(w X, Y)$	Full Bayesian inference
Parametric	$p(w X, Y) \approx q(w \lambda)$	Parametric Var. Inference
Delta function	$p(w X, Y) \approx \delta(w_{MP})$	Max. posterior inference
No prior	$w_{ML}$	Max. likelihood inference



# Bayesian linear regression: training (MP)

Maximum posterior inference:  $\log[p(Y|X, w)p(w)] \rightarrow \max_w$

$$p(w|X, Y) = \frac{p(Y|X, w)p(w)}{p(Y|X)}$$

↓  
 $\max_w$

# Bayesian linear regression: training (MP)

Maximum posterior inference:  $\log[p(Y|X, w)p(w)] \rightarrow \max_w$

$$p(w|X, Y) = \frac{p(Y|X, w)p(w)}{p(Y|X)} \rightarrow \max_w$$

↓  
 $\max_w$

**does not  
depend on  
weights**

# Bayesian linear regression: training (MP)

Maximum posterior inference:  $\log[p(Y|X, w)p(w)] \rightarrow \max_w$

# Bayesian linear regression: training (MP)

Maximum posterior inference:  $\log[p(Y|X, w)p(w)] \rightarrow \max_w$

Likelihood:  $p(Y|X, w) = \mathcal{N}(Y|Xw, I)$       Prior:  $p(w) = \mathcal{N}(w|0, \alpha I)$ ,  $\alpha > 0$

$$\log[p(Y|X, w)p(w)] = \text{Const} - \frac{1}{2}\|Y - Xw\|^2 - \frac{1}{2\alpha}\|w\|^2 \rightarrow \max_w$$

# Bayesian linear regression: training (MP)

Maximum posterior inference:  $\log[p(Y|X, w)p(w)] \rightarrow \max_w$

Likelihood:  $p(Y|X, w) = \mathcal{N}(Y|Xw, I)$       Prior:  $p(w) = \mathcal{N}(w|0, \alpha I)$ ,  $\alpha > 0$

$$\log[p(Y|X, w)p(w)] = \text{Const} - \frac{1}{2}\|Y - Xw\|^2 - \frac{1}{2\alpha}\|w\|^2 \rightarrow \max_w$$

$$\Updownarrow$$
$$\|Y - Xw\|^2 + \frac{1}{\alpha}\|w\|^2 \rightarrow \min_w$$

— we already know the solution!

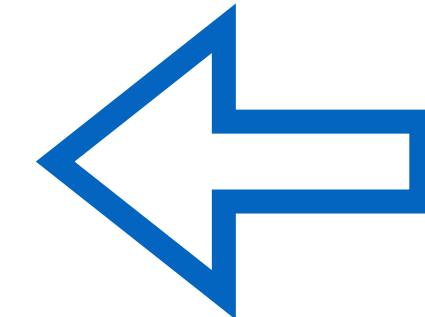
$$w_{MP} = (X^T X + \frac{1}{\alpha} I)^{-1} X^T Y$$

# Training methods: summary

Probabilistic model:  $p(Y, w|X)$

We want to compute:  $p(w|X, Y)$

Approximation		Inference
Exact	$p(w X, Y)$	Full Bayesian inference
Parametric	$p(w X, Y) \approx q(w \lambda)$	Parametric Var. Inference
Delta function	$p(w X, Y) \approx \delta(w_{MP})$	Max. posterior inference
No prior	$w_{ML}$	Max. likelihood inference



# Bayesian linear regression

Given:

$X \in \mathbb{R}^{N \times d}$  — input data

$Y \in \mathbb{R}^N$  — target values

$N$  — number of objects

$d$  — number of features

Training? Prediction?

Model:

$$p(Y, w|X) = p(Y|X, w)p(w)$$

- likelihood:

$$\begin{aligned} p(Y|X, w) &= \prod_{i=1}^N \mathcal{N}(y_i|x_i^T w, 1) = \\ &= \mathcal{N}(Y|Xw, I) \end{aligned}$$

- conjugate prior:

$$p(w) = \mathcal{N}(w|0, \alpha I), \alpha > 0$$

# Bayesian linear regression: training

Full Bayesian inference:  $p(w|X, Y)$

Likelihood and prior are conjugate  $\rightarrow$  posterior is normal

# Bayesian linear regression: training

Full Bayesian inference:  $p(w|X, Y)$

Likelihood:  $p(Y|X, w) = \mathcal{N}(Y|Xw, I)$       Prior:  $p(w) = \mathcal{N}(w|0, \alpha I)$ ,  $\alpha > 0$

# Bayesian linear regression: training

Full Bayesian inference:  $p(w|X, Y)$

Likelihood:  $p(Y|X, w) = \mathcal{N}(Y|Xw, I)$       Prior:  $p(w) = \mathcal{N}(w|0, \alpha I)$ ,  $\alpha > 0$

$$p(w|X, Y) \propto p(Y|X, w)p(w) \propto \\ \text{Const} \cdot \exp\left(-\frac{1}{2}(Y - Xw)^T(Y - Xw)\right) \exp\left(-\frac{1}{2\alpha}w^Tw\right) =$$

# Bayesian linear regression: training

Full Bayesian inference:  $p(w|X, Y)$

Likelihood:  $p(Y|X, w) = \mathcal{N}(Y|Xw, I)$       Prior:  $p(w) = \mathcal{N}(w|0, \alpha I)$ ,  $\alpha > 0$

$$\begin{aligned} p(w|X, Y) &\propto p(Y|X, w)p(w) \propto \\ &\text{Const} \cdot \exp\left(-\frac{1}{2}(Y - Xw)^T(Y - Xw)\right) \exp\left(-\frac{1}{2\alpha}w^Tw\right) = \\ &\text{Const} \cdot \exp\left(-\frac{1}{2}\underbrace{w^T}_{\text{quadratic form w.r.t weights}}(X^TX + \frac{1}{\alpha}I)w + \underbrace{w^TX^TY}_{\text{normal distribution}}\right) \end{aligned}$$

quadratic form w.r.t weights → normal distribution

# Bayesian linear regression: training

Full Bayesian inference:  $p(w|X, Y)$

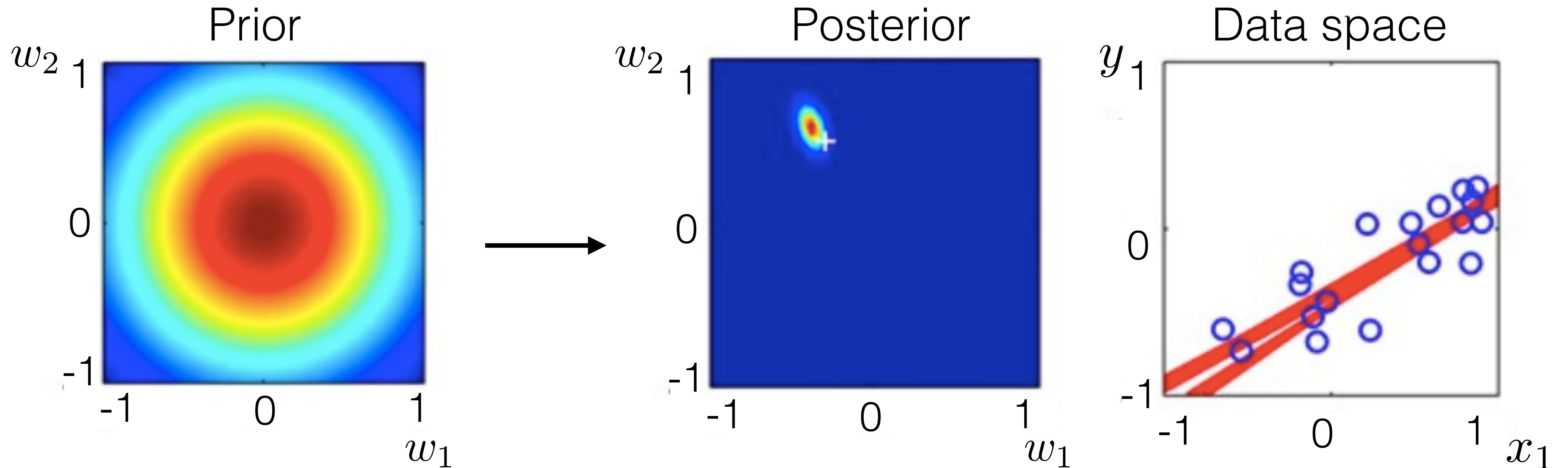
Likelihood:  $p(Y|X, w) = \mathcal{N}(Y|Xw, I)$       Prior:  $p(w) = \mathcal{N}(w|0, \alpha I)$ ,  $\alpha > 0$

$$p(w|X, Y) = \mathcal{N}(w|w_{MP}, \Sigma)$$

$$w_{MP} = (X^T X + \frac{1}{\alpha} I)^{-1} X^T Y$$

$$\Sigma = \left( X^T X + \frac{1}{\alpha} I \right)^{-1}$$

# Training visualization



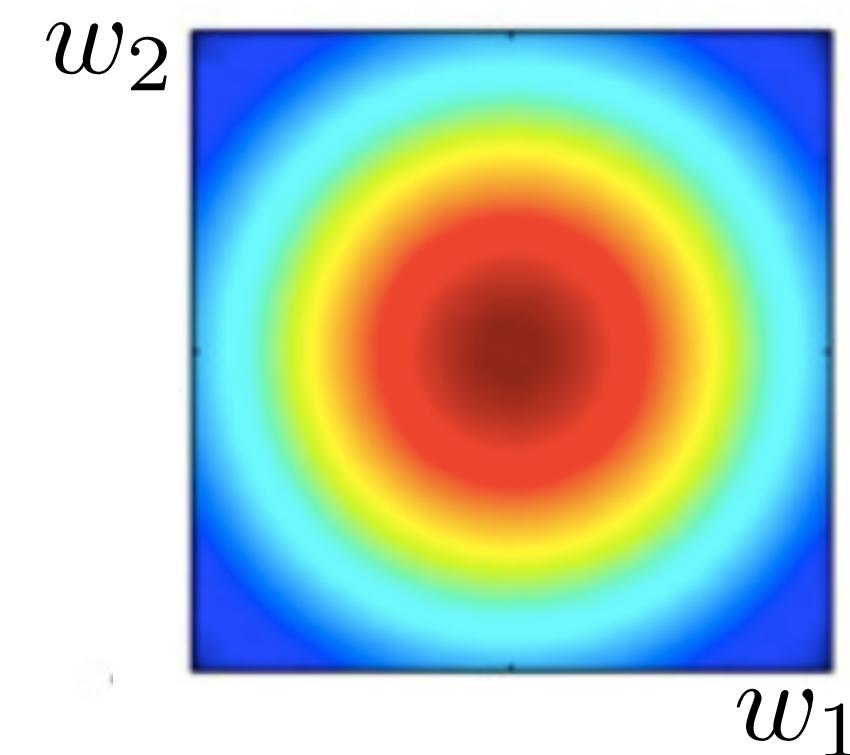
$$p(w) = \mathcal{N}(w|0, \alpha I)$$

$$\begin{aligned} p(w|X, Y) = \\ \mathcal{N}(w|w_{MP}, \Sigma) \end{aligned}$$

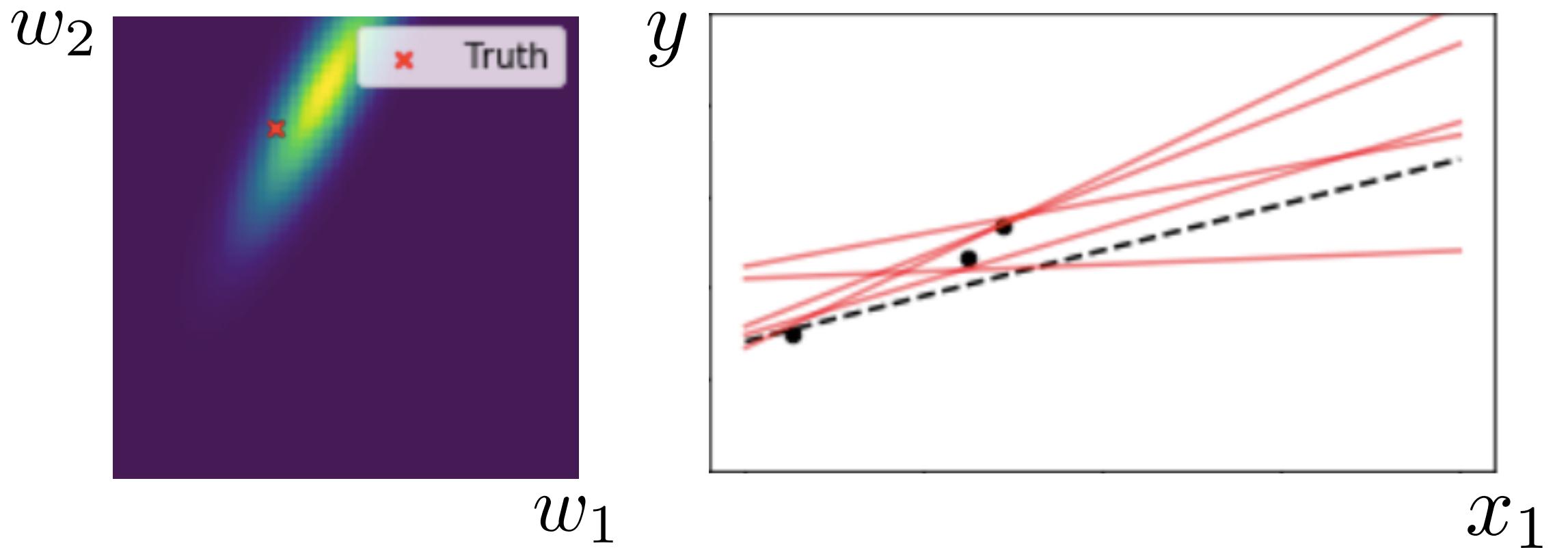
$$\begin{aligned} p(y|x, w) = \\ \mathcal{N}(y|w_1 + w_2 x_1, 1) \end{aligned}$$

# Training: increasing amount of data

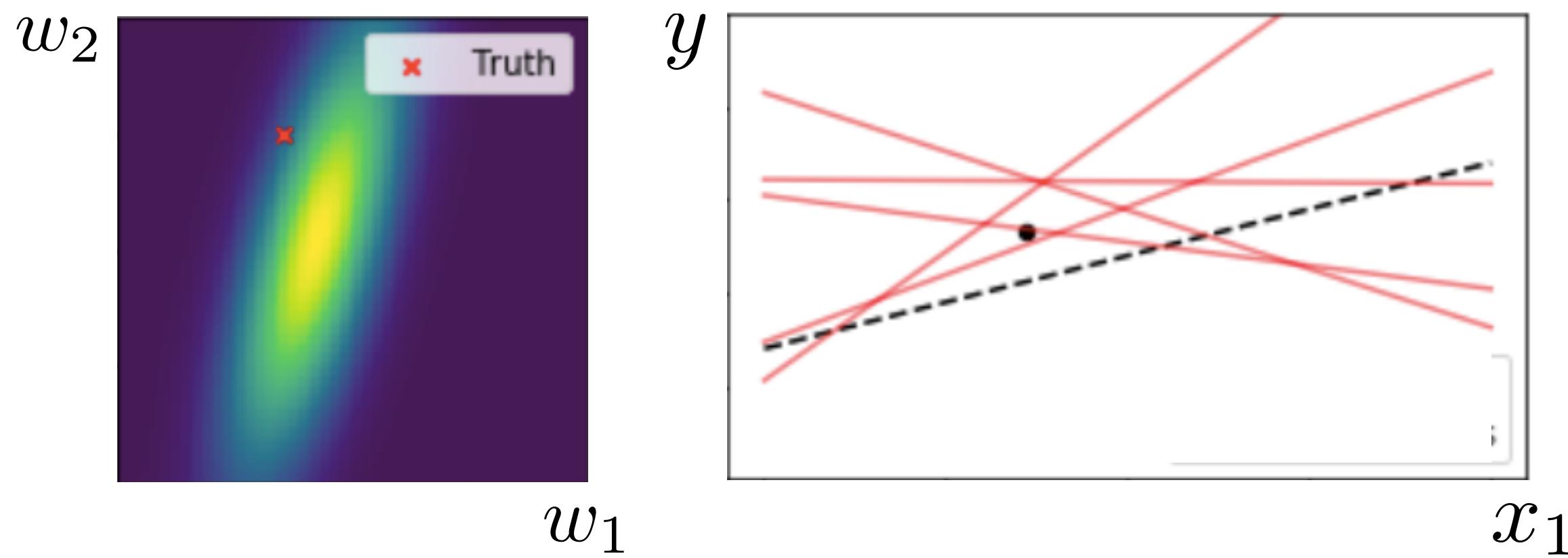
Prior:



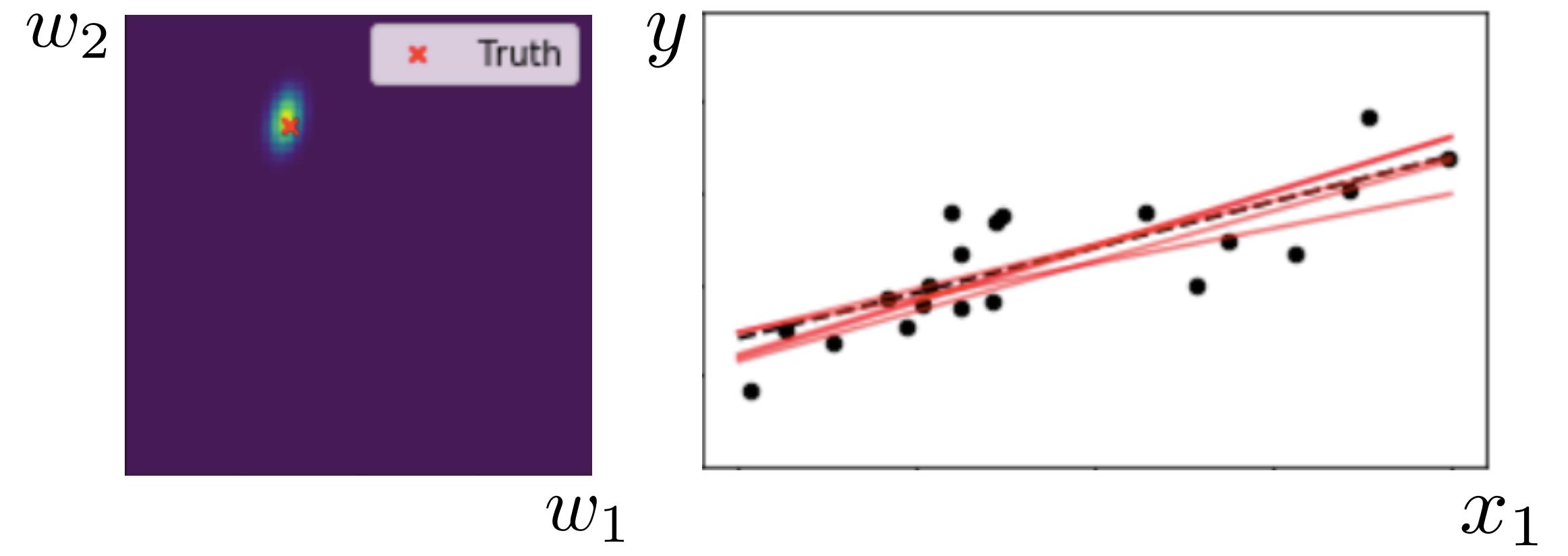
3 data points ( $N=3$ ):



1 data point ( $N=1$ ):



20 data points ( $N=20$ ):



# Bayesian linear regression

Given:

$X \in \mathbb{R}^{N \times d}$  — input data

$Y \in \mathbb{R}^N$  — target values

Model:

$$\begin{aligned} p(Y, w|X) &= p(Y|X, w)p(w) = \\ &= \mathcal{N}(Y|Xw, I)\mathcal{N}(w|0, \alpha I) \end{aligned}$$

Training:

$$p(w|X, Y) = \mathcal{N}(w|w_{MP}, \Sigma)$$

$$w_{MP} = (X^T X + \frac{1}{\alpha} I)^{-1} X^T Y$$

$$\Sigma = \left( X^T X + \frac{1}{\alpha} I \right)^{-1}$$

Prediction?

# Full Bayesian inference

**Training stage:**

$$p(w|X, Y) = \frac{p(Y|X, w)p(w)}{\int p(Y|X, \tilde{w})p(\tilde{w})d\tilde{w}}$$



**Testing stage:**

$$p(y_*|x_*, X, Y) = \int p(y_*|x_*, w)p(w|X, Y)dw = \mathbb{E}_{p(w|X, Y)}p(y_*|x_*, w)$$

$x_*$  — new object

# Full Bayesian inference

**Testing stage:**

$$p(y_*|x_*, X, Y) = \int p(y_*|x_*, w)p(w|X, Y)dw = \mathbb{E}_{p(w|X, Y)}p(y_*|x_*, w)$$

$x_*$  — new object

$$p(y_*|x_*, w) = \mathcal{N}(y_*|w^T x_*, 1)$$

$$p(w|X, Y) = \mathcal{N}(w|w_{MP}, \Sigma)$$

# Convolution of normal distributions

$$p(x) = \mathcal{N}(x|\mu, \Sigma), \quad p(y|x) = \mathcal{N}(y|Ax, \Gamma), \quad p(y) - ?$$

# Convolution of normal distributions

$$p(x) = \mathcal{N}(x|\mu, \Sigma), \quad p(y|x) = \mathcal{N}(y|Ax, \Gamma), \quad p(y)-?$$

$$y = Ax + \epsilon, \quad \epsilon \sim \mathcal{N}(\epsilon|0, \Gamma)$$

$$\mathbb{E}y-? \quad \mathbb{D}y-?$$

$$\mathbb{E}y = \mathbb{E}[Ax + \epsilon] = A\mathbb{E}x = A\mu$$

$$\begin{aligned}\mathbb{D}y &= \mathbb{E}(y - \mathbb{E}y)^2 = \mathbb{E}(a^T x + \epsilon - a^T \mu)^2 = \mathbb{E}(a^T(x - \mu) + \epsilon)^2 = \\ &= \mathbb{E}a^T(x - \mu)(x - \mu)^T a + 2\mathbb{E}\epsilon\mathbb{E}a^T(x - \mu) + \mathbb{E}\epsilon^2 = \\ &= a^T \Sigma a + \Gamma\end{aligned}$$

# Convolution of normal distributions

$$p(x) = \mathcal{N}(x|\mu, \Sigma), \quad p(y|x) = \mathcal{N}(y|Ax, \Gamma), \quad \Rightarrow$$

$$p(y) = \mathcal{N}(y|a^T \mu, a^T \Sigma a + \Gamma)$$

**Testing stage:**

$$p(y_*|x_*, X, Y) = \int p(y_*|x_*, w)p(w|X, Y)dw = \mathbb{E}_{p(w|X, Y)}p(y_*|x_*, w)$$

$x_*$  — new object

$$p(y_*|x_*, w) = \mathcal{N}(y_*|w^T x_*, 1) \quad p(w|X, Y) = \mathcal{N}(w|w_{MP}, \Sigma)$$

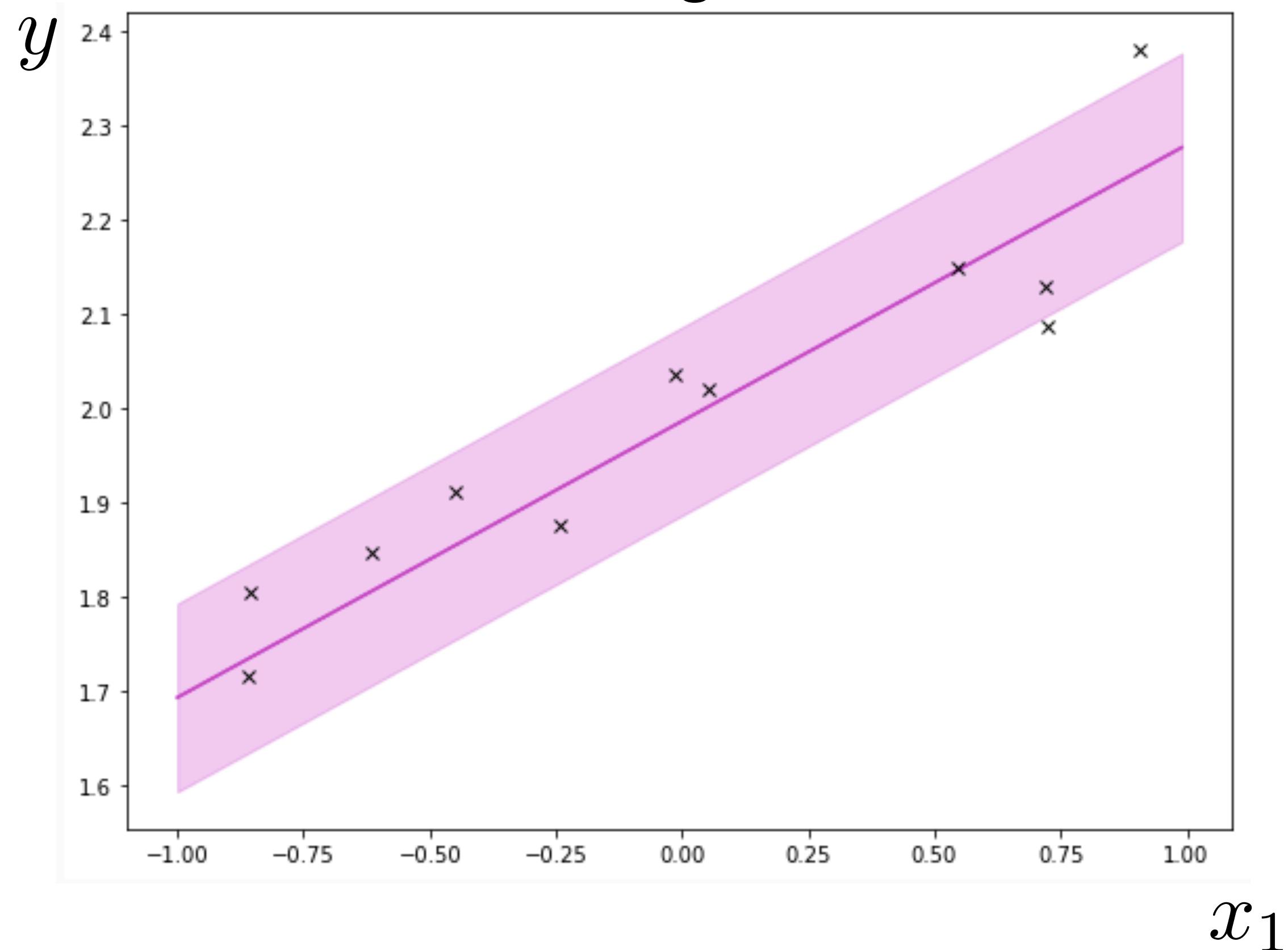
# Bayesian linear regression: prediction

$$\begin{aligned} p(y_*|x_*, X, Y) &= \int p(y_*|x_*, w)p(w|X, Y)dw = \\ &\int \mathcal{N}(y_*|x_*^T w, 1)\mathcal{N}(w|w_{MP}, \Sigma)dw = \\ &\mathcal{N}(y_*|x_*^T w_{MP}, 1 + x_*^T \Sigma x_*) \end{aligned}$$

$x_*$  — new object

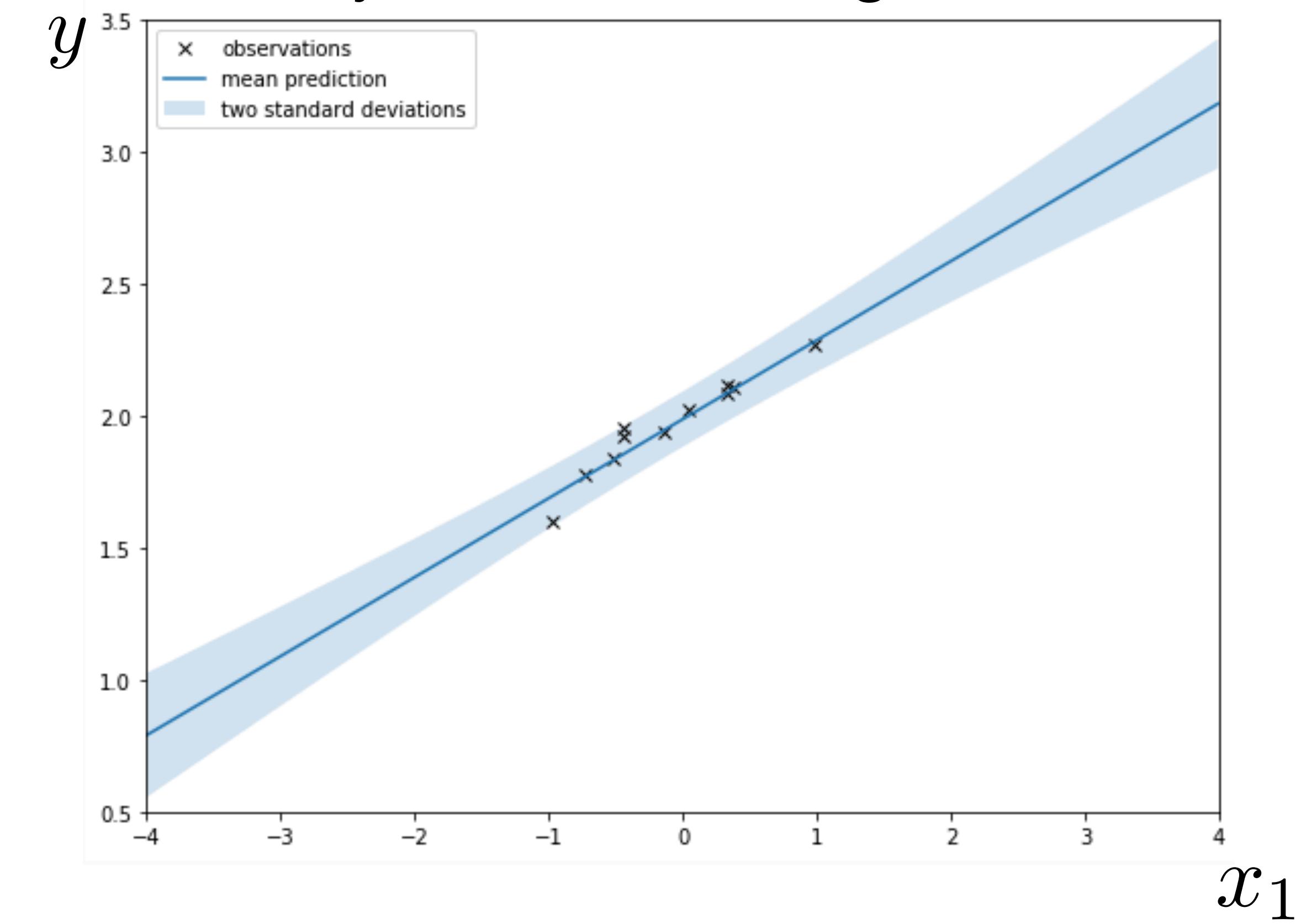
# Prediction visualization

Linear regression



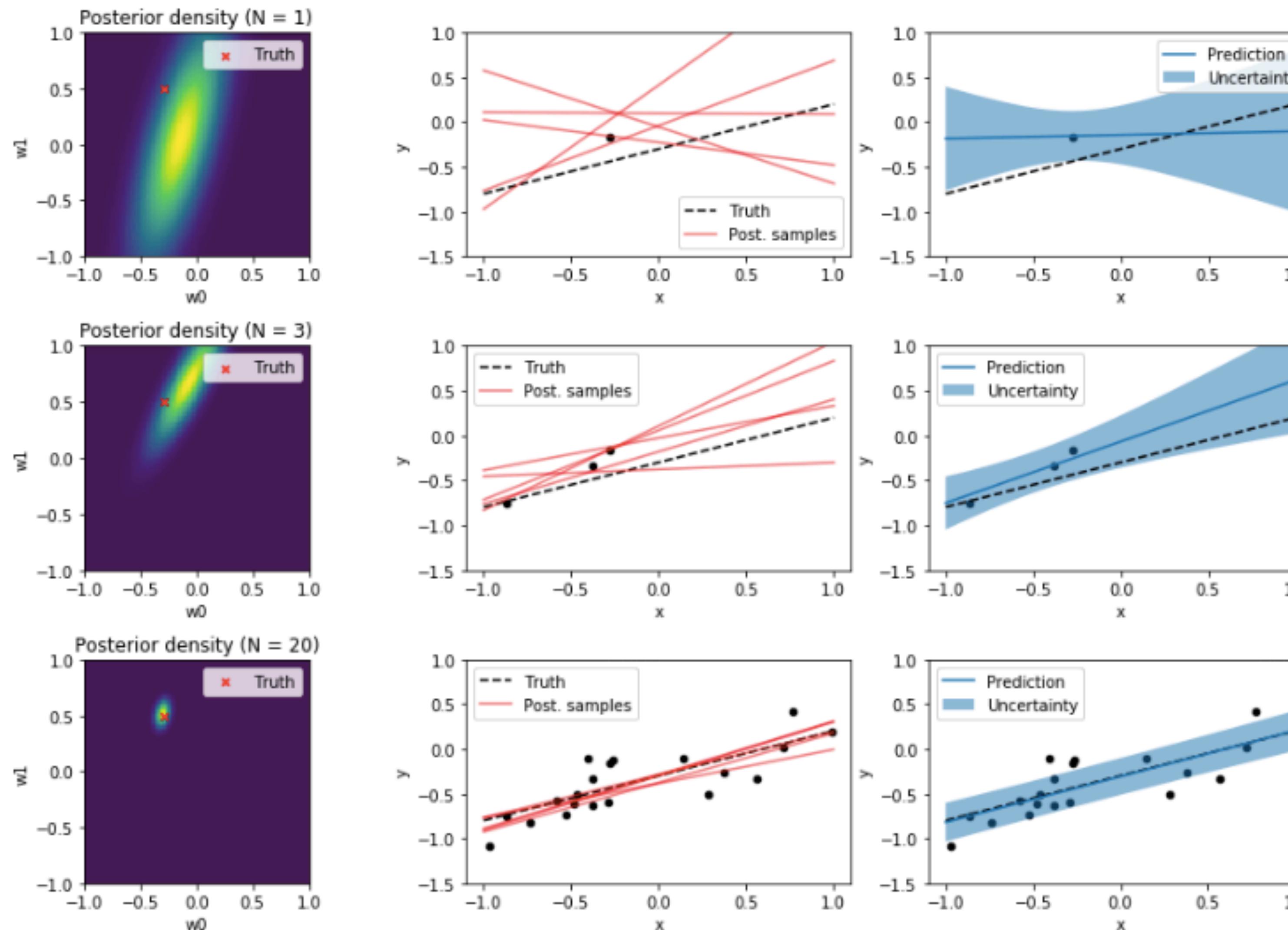
$$\mathcal{N}(y_* | x_*^T w_{MP}, 1)$$

Bayesian linear regression



$$\mathcal{N}(y_* | x_*^T w_{MP}, 1 + x_*^T \Sigma x_*)$$

# Prediction: increasing amount of data

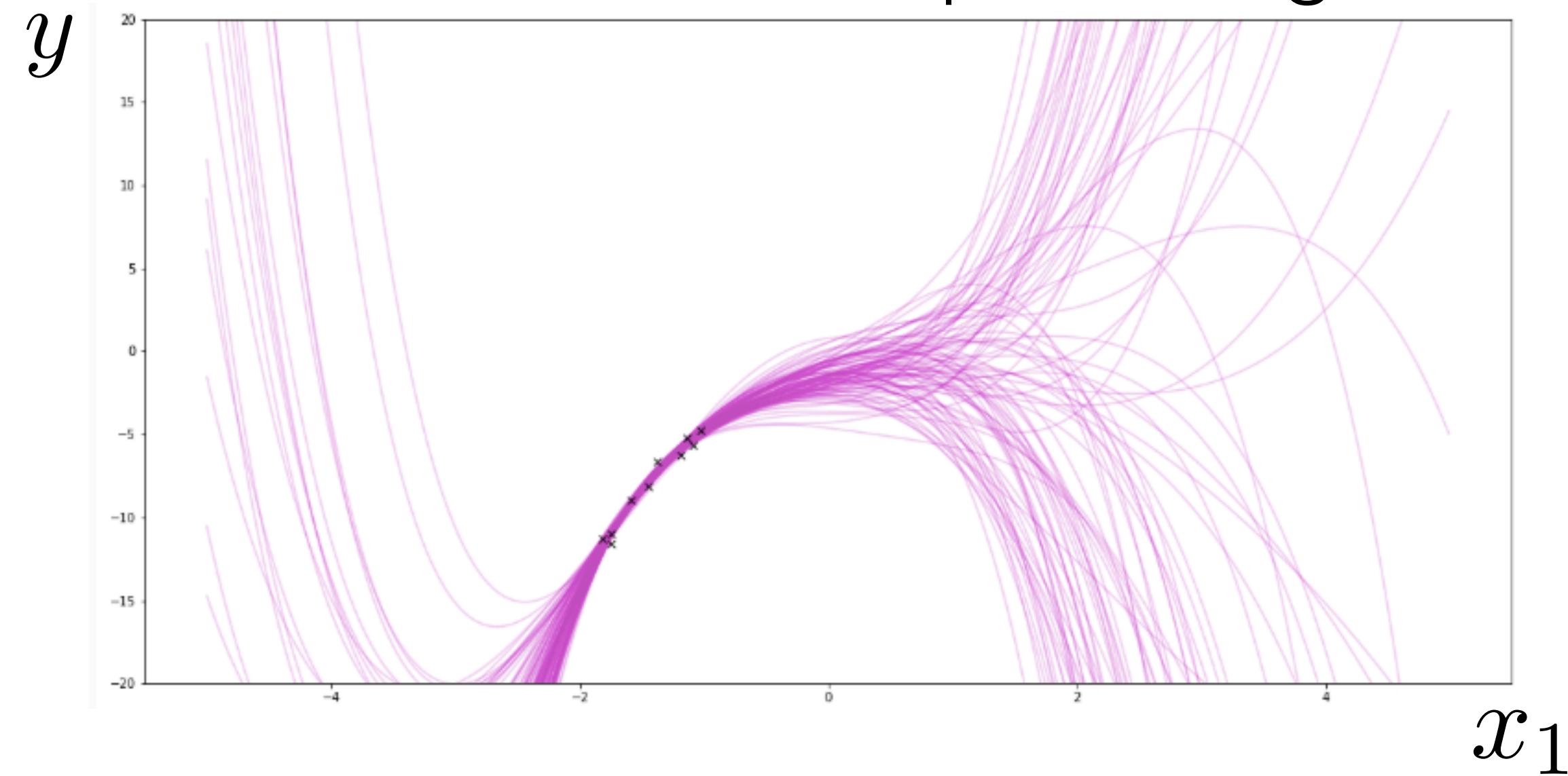


# Prediction: polynomial features

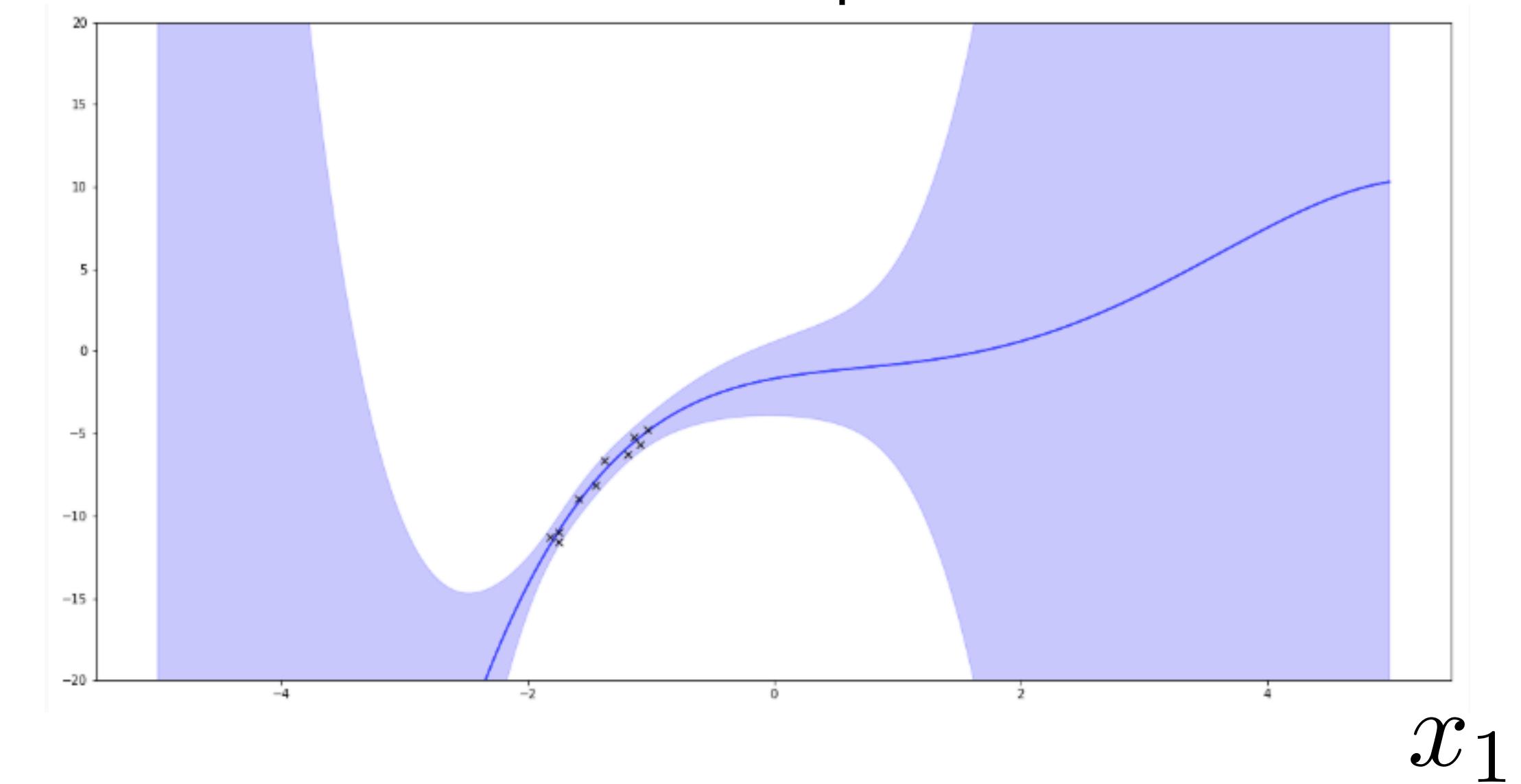
Modify training data: add polynomial features

$$p(y|x, w) = \mathcal{N}(y|w_1 + w_2\underbrace{x_1}_{} + w_3\underbrace{x_1^2}_{} + \dots w_6\underbrace{x_1^5}_{}, 1)$$

Prediction with sampled weights



Mean  $\pm$  Std of predictions



# Bayesian linear regression

Given:

$X \in \mathbb{R}^{N \times d}$  — input data

$Y \in \mathbb{R}^N$  — target values

Model:

$$\begin{aligned} p(Y, w|X) &= p(Y|X, w)p(w) = \\ &= \mathcal{N}(Y|Xw, I)\mathcal{N}(w|0, \alpha I) \end{aligned}$$

Training:

$$p(w|X, Y) = \mathcal{N}(w|w_{MP}, \Sigma)$$

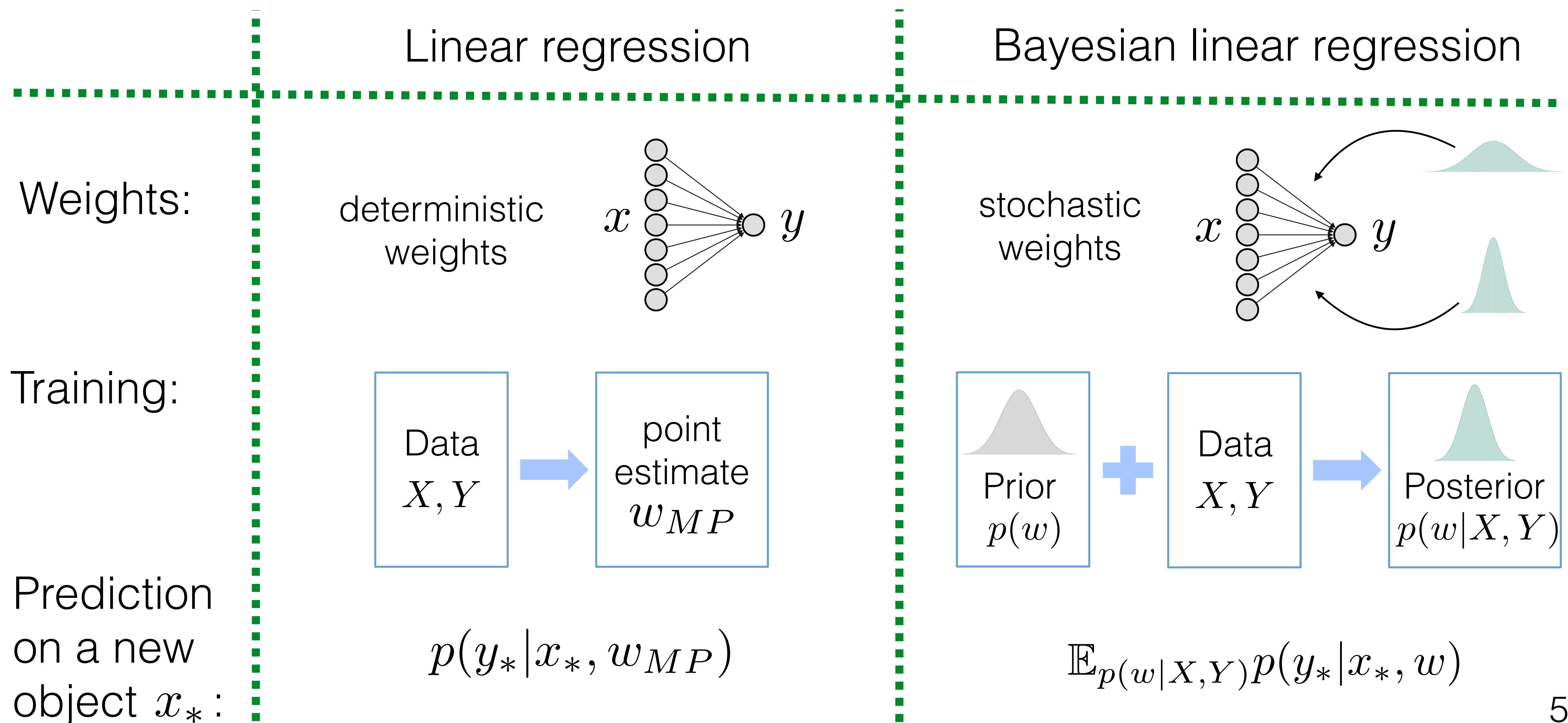
$$w_{MP} = (X^T X + \frac{1}{\alpha} I)^{-1} X^T Y$$

$$\Sigma = \left( X^T X + \frac{1}{\alpha} I \right)^{-1}$$

Prediction on a new object  $x_*$ :

$$\begin{aligned} p(y_*|x_*, X, Y) &= \\ &= \mathcal{N}(y_*|x_*^T w_{MP}, 1 + x_*^T \Sigma x_*) \end{aligned}$$

# Putting everything together



# Putting everything together

	Linear regression	Bayesian linear regression
Weights:	deterministic weights	stochastic weights
Training:	$w_{MP} = (X^T X + \frac{1}{\alpha} I)^{-1} X^T Y$	$p(w X, Y) = \mathcal{N}(w w_{MP}, \Sigma)$ $w_{MP} = (X^T X + \frac{1}{\alpha} I)^{-1} X^T Y$ $\Sigma = (X^T X + \frac{1}{\alpha} I)^{-1}$
Prediction on a new object $x_*$ :	$\mathcal{N}(y_*   x_*^T w_{MP}, 1)$	$\mathcal{N}(y_*   x_*^T w_{MP}, 1 + x_*^T \Sigma x_*)$

# Choosing hyperparameters in Bayesian approach

Given:

$X \in \mathbb{R}^{N \times d}$  — input data

$Y \in \mathbb{R}^N$  — target values

Model:

$$\begin{aligned} p(Y, w|X) &= p(Y|X, w)p(w) = \\ &= \mathcal{N}(Y|Xw, I)\mathcal{N}(w|0, \alpha I) \end{aligned}$$

Training:

$$p(w|X, Y) = \mathcal{N}(w|w_{MP}, \Sigma)$$

$$w_{MP} = (X^T X + \frac{1}{\alpha} I)^{-1} X^T Y$$

$$\Sigma = X^T X + \frac{1}{\alpha} I$$

Prediction on a new object  $x_*$ :

$$\begin{aligned} p(y_*|x_*, X, Y) &= \\ &= \mathcal{N}(y_*|x_*^T w_{MP}, 1 + x_*^T \Sigma x_*) \end{aligned}$$

**Any hyperparameters?**

# Choosing hyperparameters in Bayesian approach

Given:

$X \in \mathbb{R}^{N \times d}$  — input data

$Y \in \mathbb{R}^N$  — target values

Model:

$$\begin{aligned} p(Y, w|X) &= p(Y|X, w)p(w) = \\ &= \mathcal{N}(Y|Xw, I) \mathcal{N}(w|0, \alpha I) \end{aligned}$$

Training:

$$p(w|X, Y) = \mathcal{N}(w|w_{MP}, \Sigma)$$

$$w_{MP} = (X^T X + \frac{1}{\alpha} I)^{-1} X^T Y$$

$$\Sigma = X^T X + \frac{1}{\alpha} I$$

Prediction on a new object  $x_*$ :

$$\begin{aligned} p(y_*|x_*, X, Y) &= \\ &= \mathcal{N}(y_*|x_*^T w_{MP}, 1 + x_*^T \Sigma x_*) \end{aligned}$$

**Any hyperparameters?**

# Choosing hyperparameters in Bayesian approach

Likelihood:  $p(Y|X, \theta)$

Prior:  $p(\theta|\alpha)$

Training:

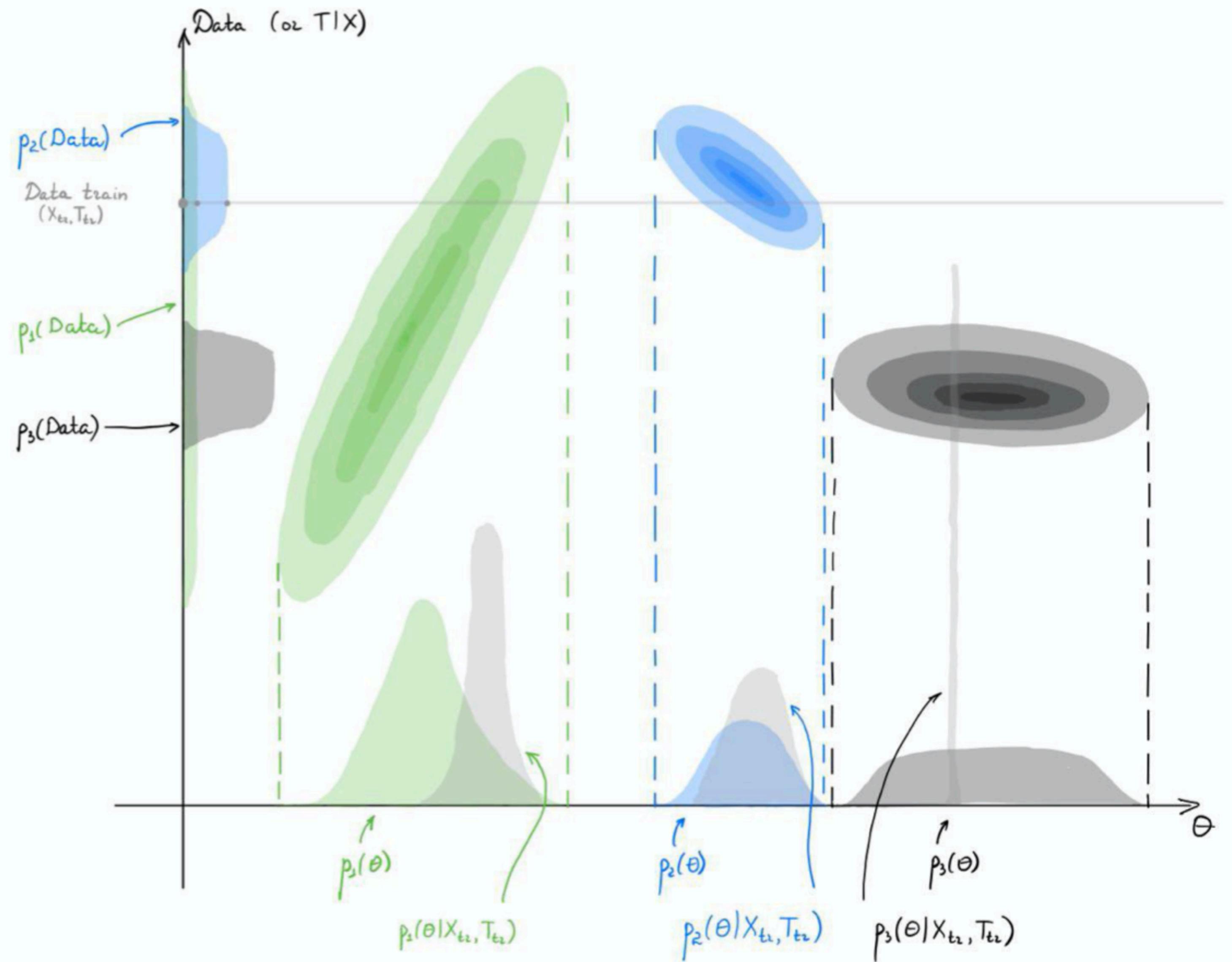
$$p(\theta|X, Y, \alpha) = \frac{p(Y|X, \theta)p(\theta|\alpha)}{\int p(Y|X, \tilde{\theta})p(\tilde{\theta}|\alpha)d\tilde{\theta}}$$

$$\text{Posterior} = \frac{\text{Likelihood} \cdot \text{Prior}}{\text{Evidence}}$$

Maximum evidence:

$$\int p(Y|X, \tilde{\theta})p(\tilde{\theta}|\alpha)d\tilde{\theta} \rightarrow \max_{\alpha}$$

# Maximum evidence



# Relevance vector machine

Given:

$X \in \mathbb{R}^{N \times d}$  — input data

$Y \in \mathbb{R}^N$  — target values

Model:

$$\begin{aligned} p(Y, w | X) &= p(Y | X, w)p(w) = \\ &= \mathcal{N}(Y | Xw, I)\mathcal{N}(w | 0, \mathbf{A}) \end{aligned}$$

$$A = \begin{pmatrix} \alpha_1^2 & & & & 0 \\ & \ddots & & & \\ & & \ddots & & \\ 0 & & & \ddots & \\ & & & & \alpha_d^2 \end{pmatrix}$$

$A - ?$

# Relevance vector machine

Given:

$X \in \mathbb{R}^{N \times d}$  — input data

$Y \in \mathbb{R}^N$  — target values

Maximum evidence:

$$\int \mathcal{N}(Y|Xw, I) \mathcal{N}(w|0, A) dw \rightarrow \max_A$$

Model:

$$p(Y, w|X) = p(Y|X, w)p(w) = \\ = \mathcal{N}(Y|Xw, I)\mathcal{N}(w|0, A)$$

$$A = \begin{pmatrix} \alpha_1^2 & & & & \\ & \ddots & & & 0 \\ & & \ddots & & \\ 0 & & & \ddots & \\ & & & & \alpha_d^2 \end{pmatrix}$$

$A - ?$

# Relevance vector machine

Given:

$X \in \mathbb{R}^{N \times d}$  — input data

$Y \in \mathbb{R}^N$  — target values

Model:

$$\begin{aligned} p(Y, w | X) &= p(Y | X, w)p(w) = \\ &= \mathcal{N}(Y | Xw, I)\mathcal{N}(w | 0, \mathbf{A}) \end{aligned}$$

Training:

Repeat:

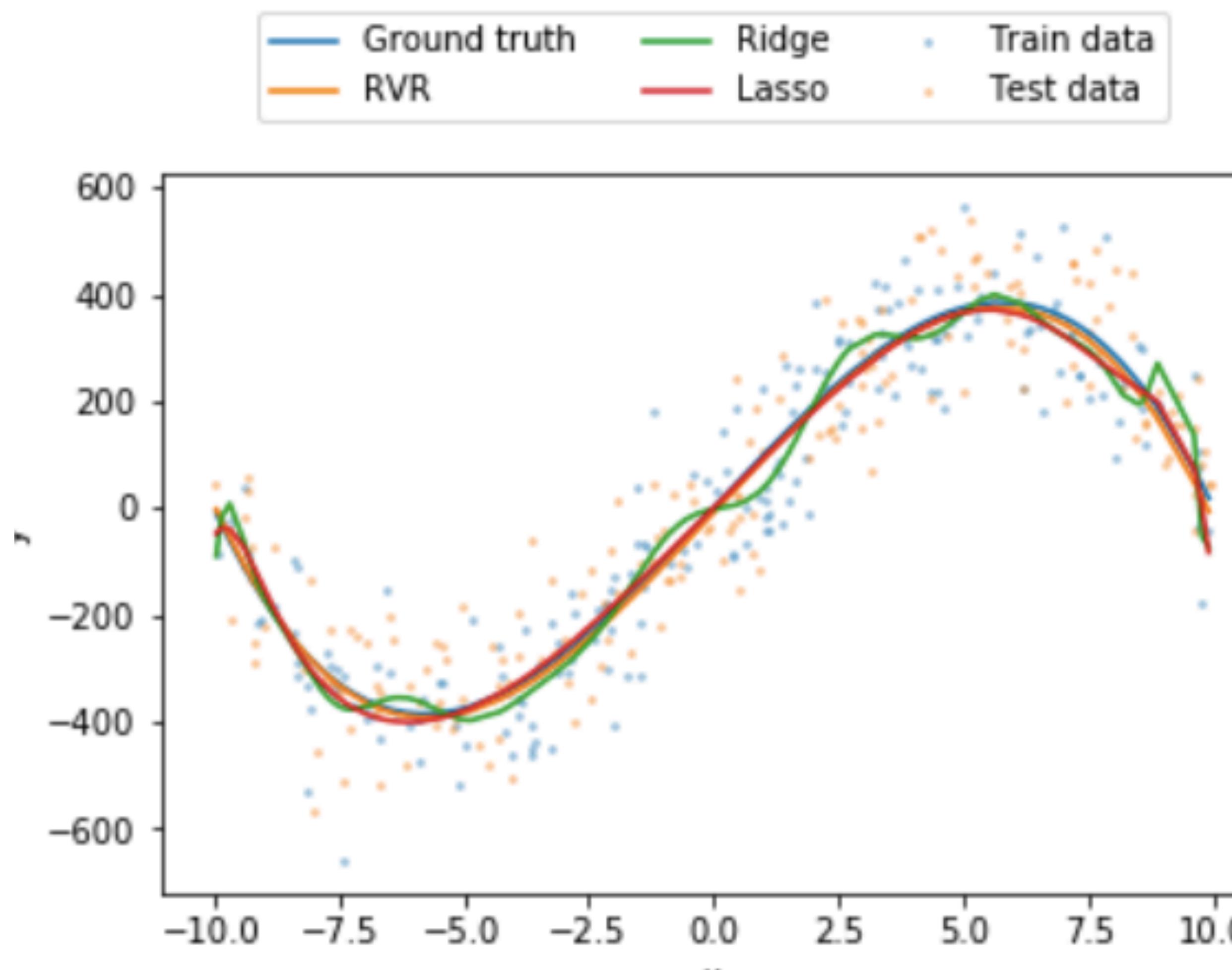
$$\tilde{\alpha}_i^{new} = \frac{1 - \tilde{\alpha}_i^{old}\Sigma_{ii}}{w_{MP,i}^2}$$

$$\begin{aligned} \alpha_i^{new} &= (\tilde{\alpha}_i^{new})^{-2} \\ \Sigma &= \left( X^T X + \frac{1}{\alpha} I \right)^{-1} \end{aligned}$$

$$w_{MP} = (X^T X + A^{new})^{-1} X^T Y \quad \Sigma = (X^T X + A^{new})^{-1}$$

# Relevance vector machine: visualization

$\alpha_i \rightarrow 0 \Rightarrow i$ -й вес всегда равен 0 (нерелевантный признак)



Relevance Vector Regression

Features remaining: 4 / 21

Train error: 9052.79670618

Test error: 9793.46413262

Ridge Regression

Features remaining: NA (no sparsity)

Train error: 8090.96629035

Test error: 12256.7842487

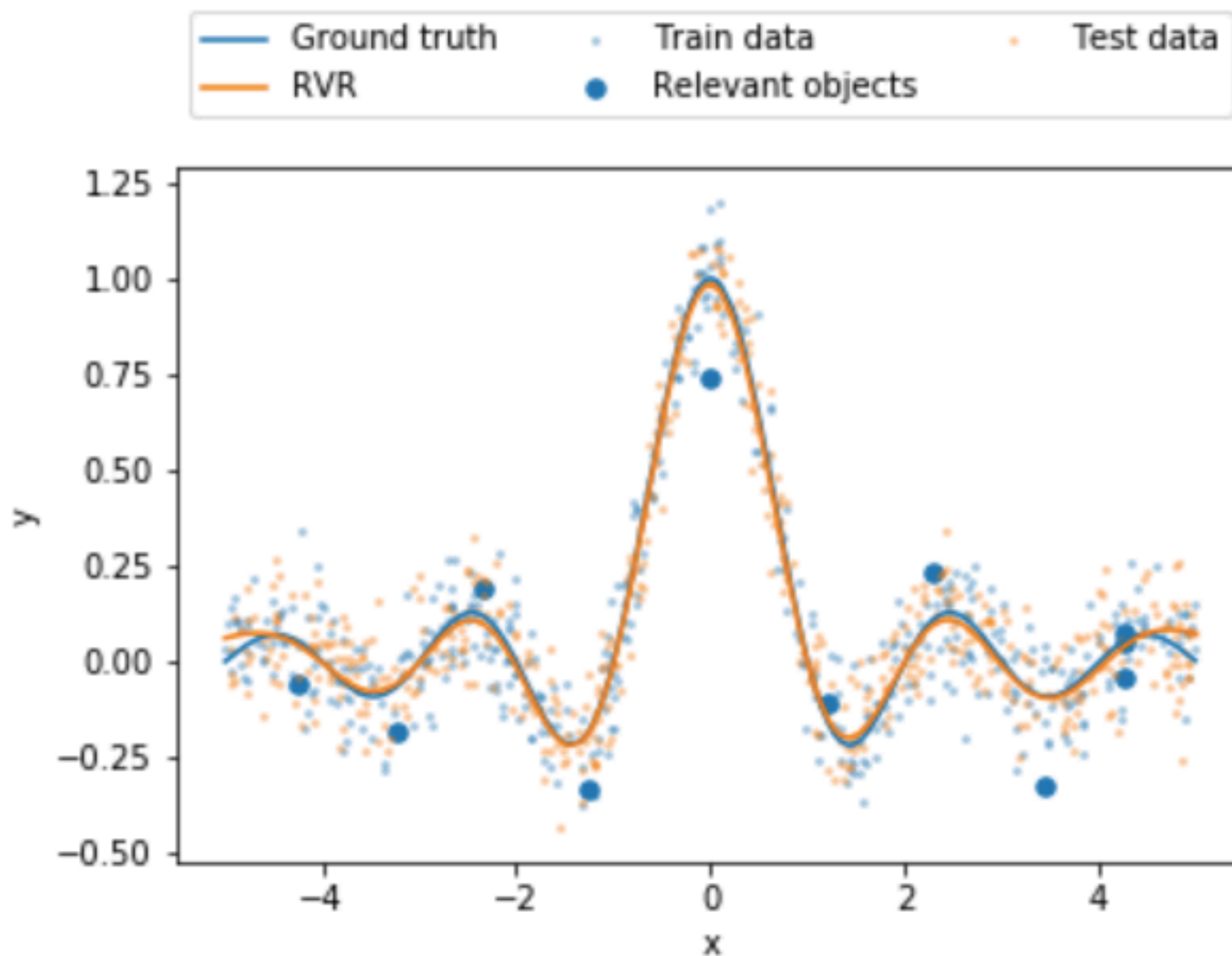
Lasso Regression

Features remaining: 19 / 21

Train error: 8941.22847976

Test error: 10237.8405027

# Relevance vector machine: RBF-features



RBF - kernel:

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$