Bayesian linear regression

Nadia Chirkova

Centre of Deep Learning and Bayesian Methods
Moscow, Russia

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

p(B|A)yesgroup.ru

Plan

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

Plan

- $\boxed{\text{Linear regression: reminder}}$
- 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

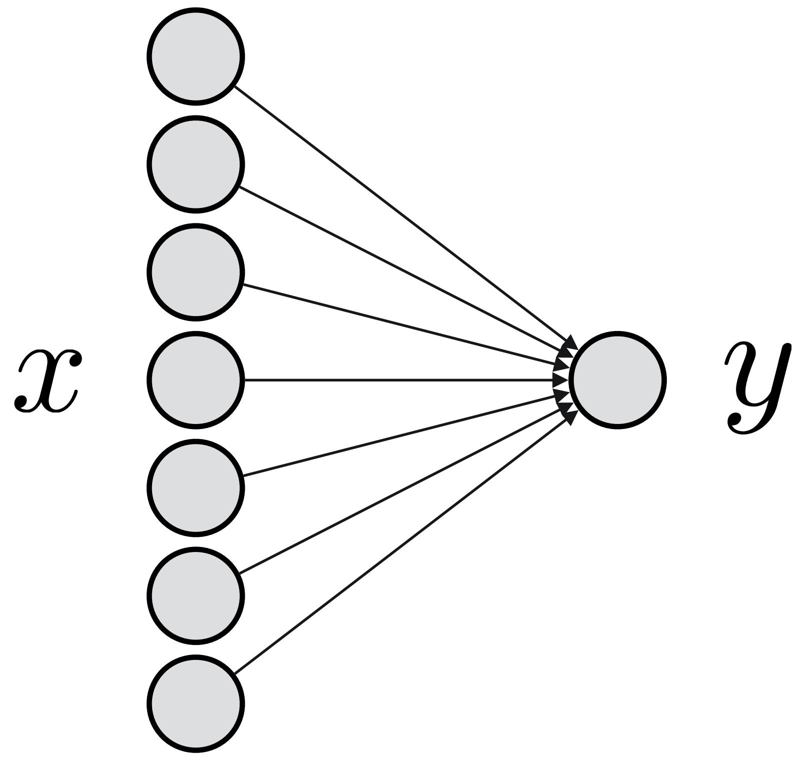
N — number of objects

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

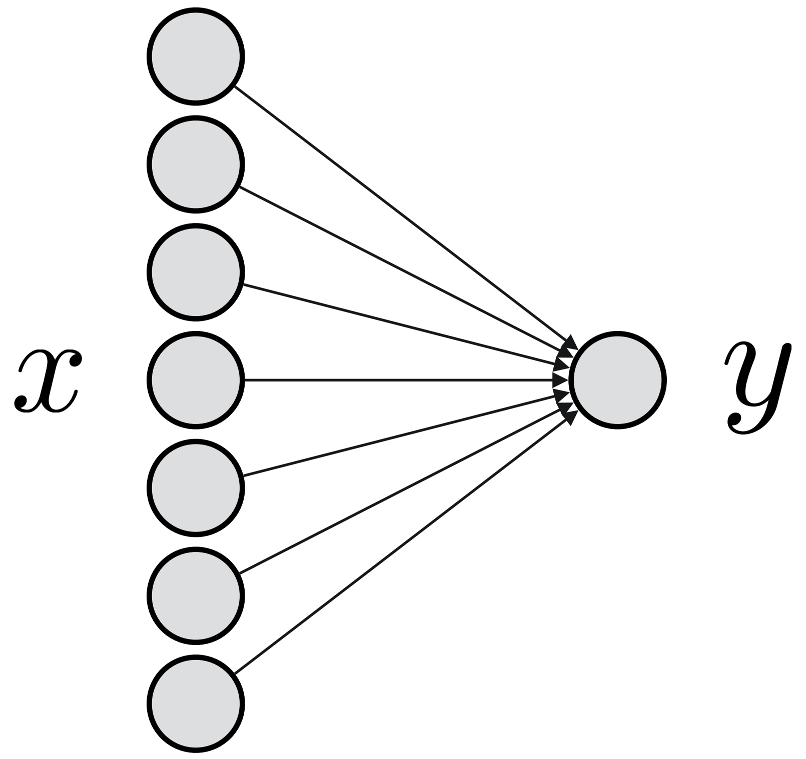
N — number of objects

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

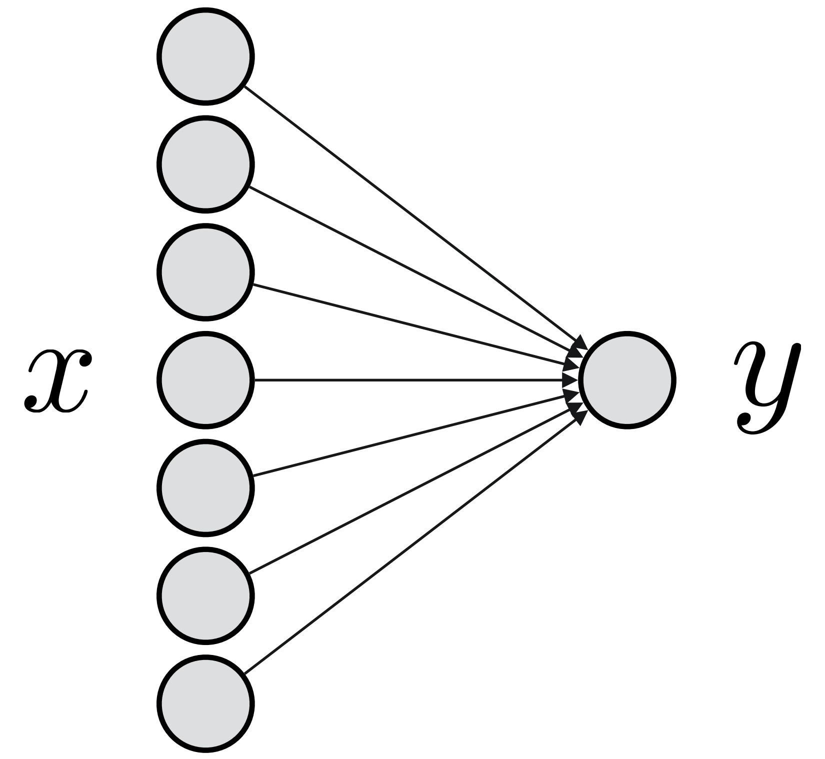
N — number of objects

d — number of features

Model:

$$Xw \approx Y$$

$$x_i^T w \approx y_i$$



linear model
with weights w

Training:

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

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

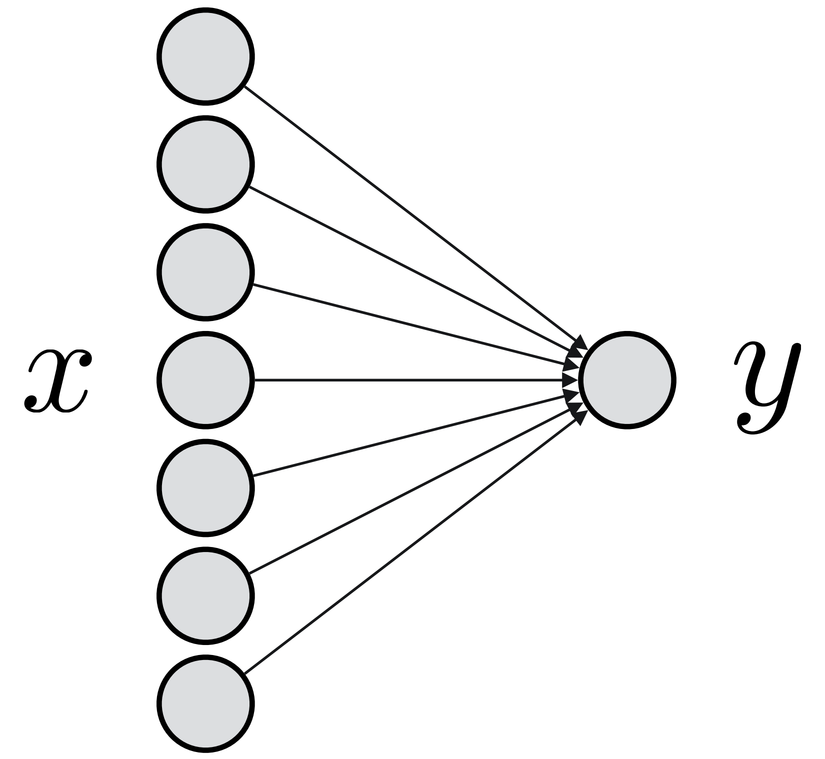
N — number of objects

d — number of features

Model:

$Xw \approx Y$

$x_i^T w \approx y_i$



linear model
with weights w

Training:

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

Prediction on a new object x_* :

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

Linear regression: training

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

Convex function:

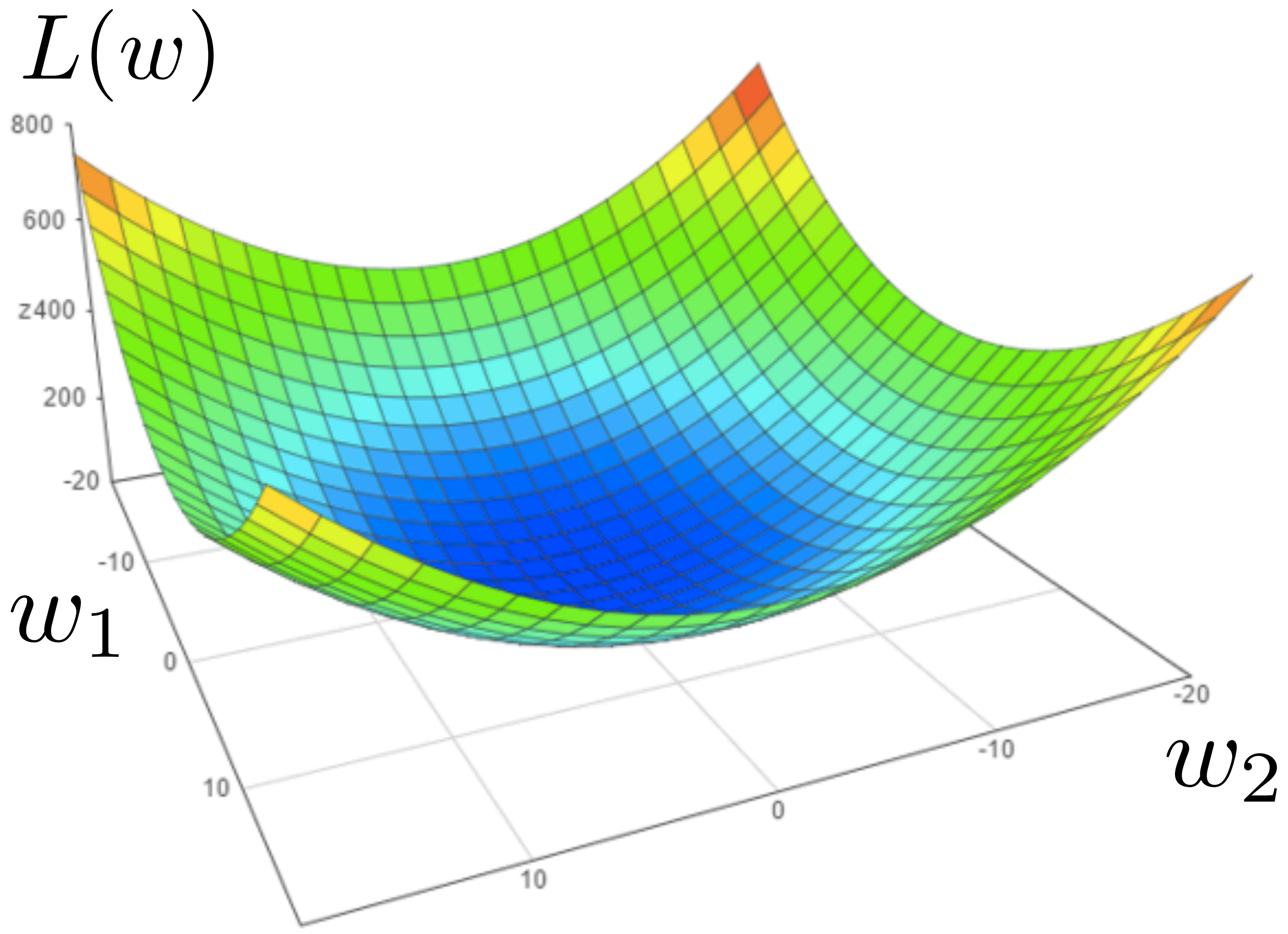
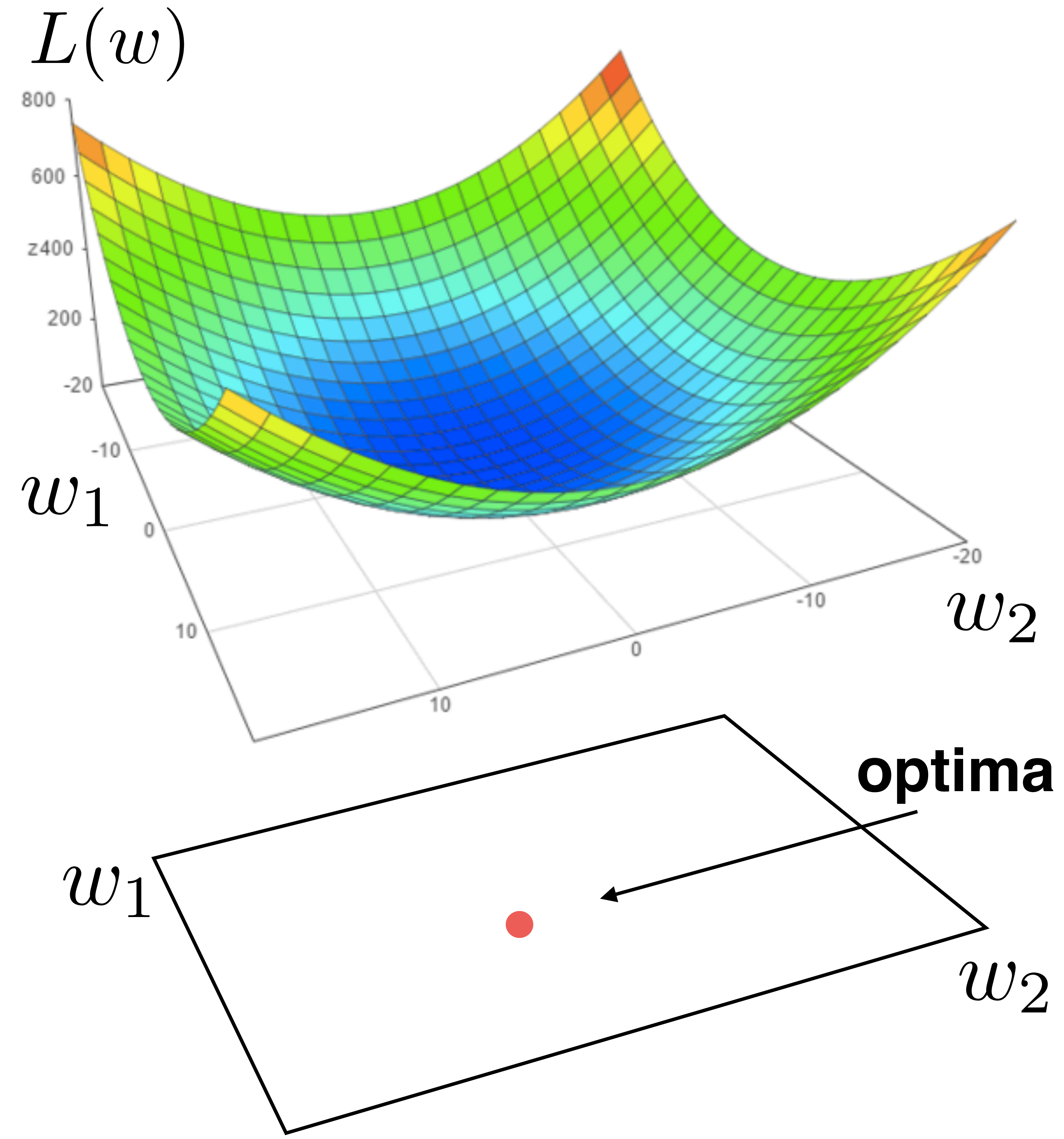


Image from https://www.globalsoftwaresupport.com/linear-regression/

Linear regression: training

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

Convex function:



Optimal weights:

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

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

Image from https://www.globalsoftwaresupport.com/linear-regression/

Linear regression: regularization

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

$$\lambda > 0$$

Optimal weights:

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

- Always unique solution
- Preventing overfitting

Strongly convex function:

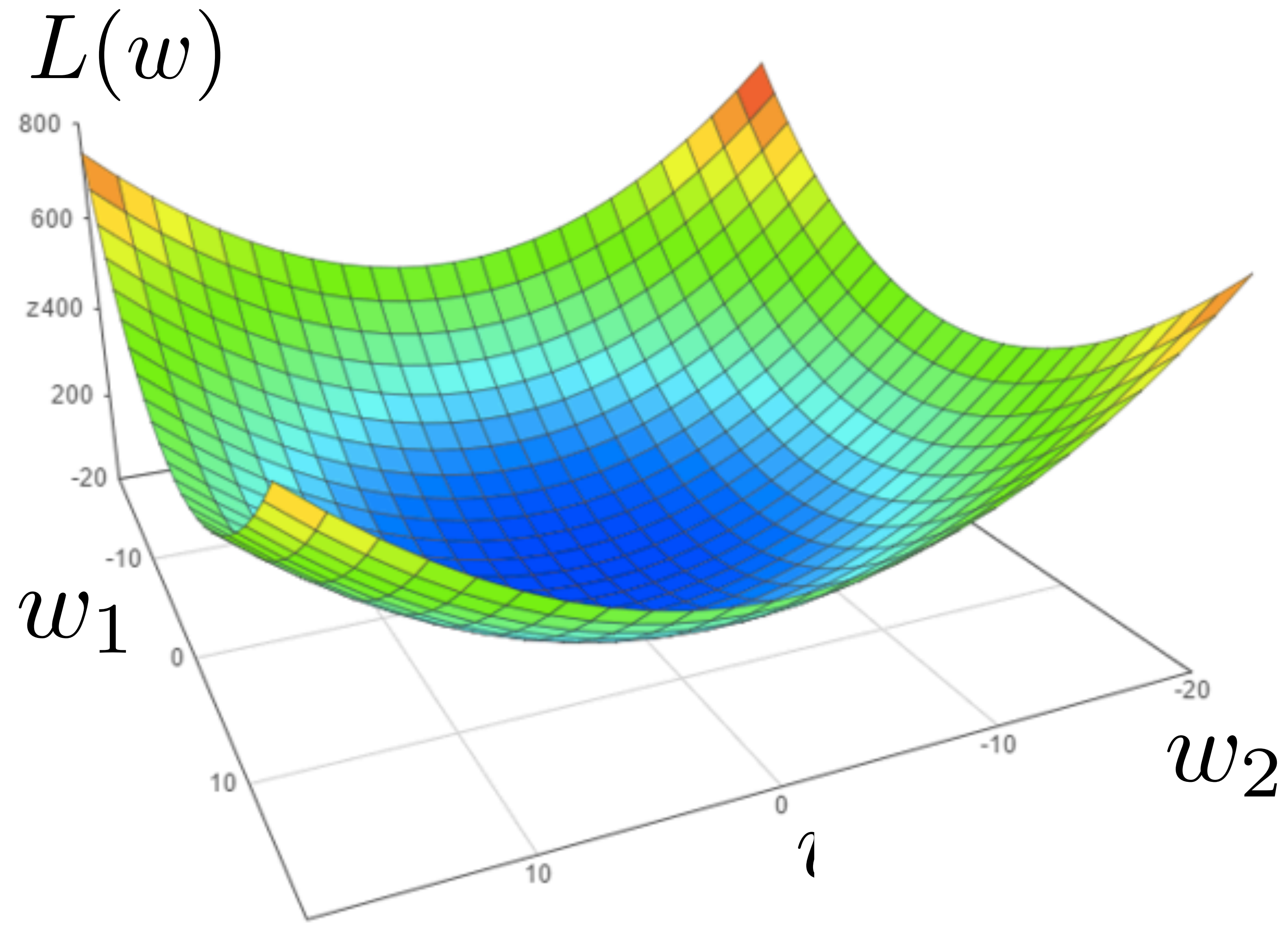


Image from https://www.globalsoftwaresupport.com/linear-regression/

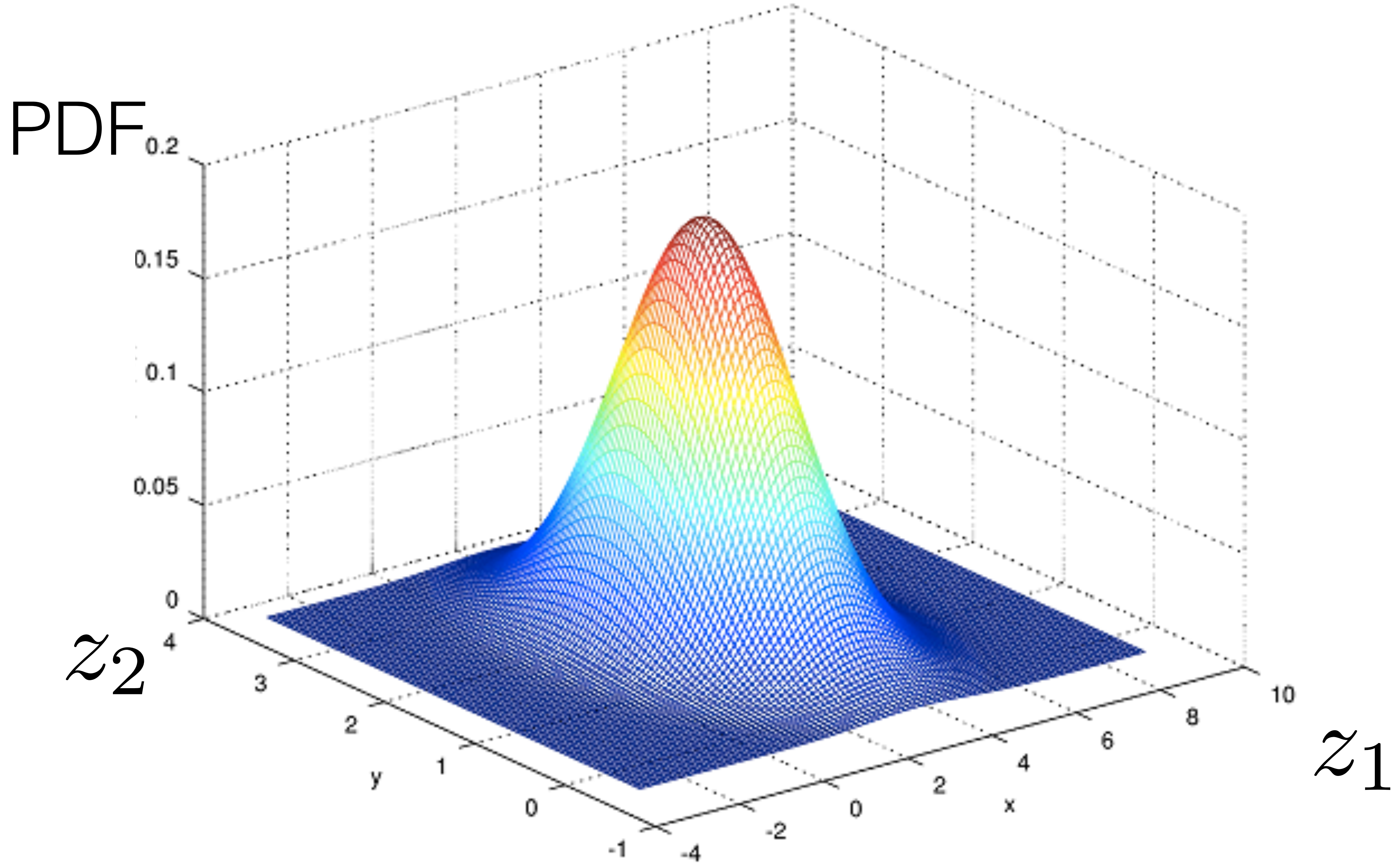
Plan

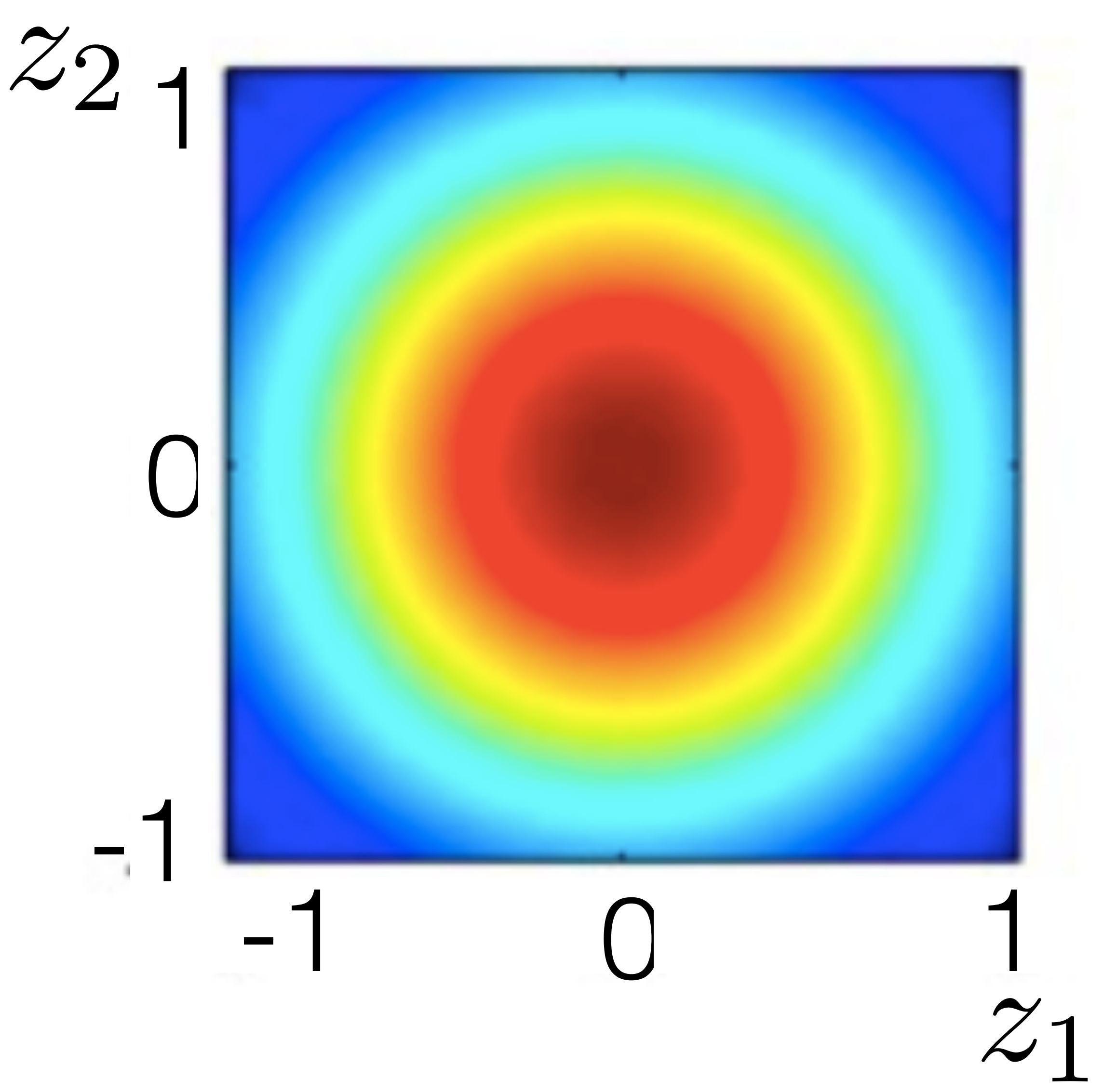
- Linear regression: reminder
- Bayesian linear regression:
    - model definition
    - training
    - prediction

Multivariate normal (Gaussian) distribution

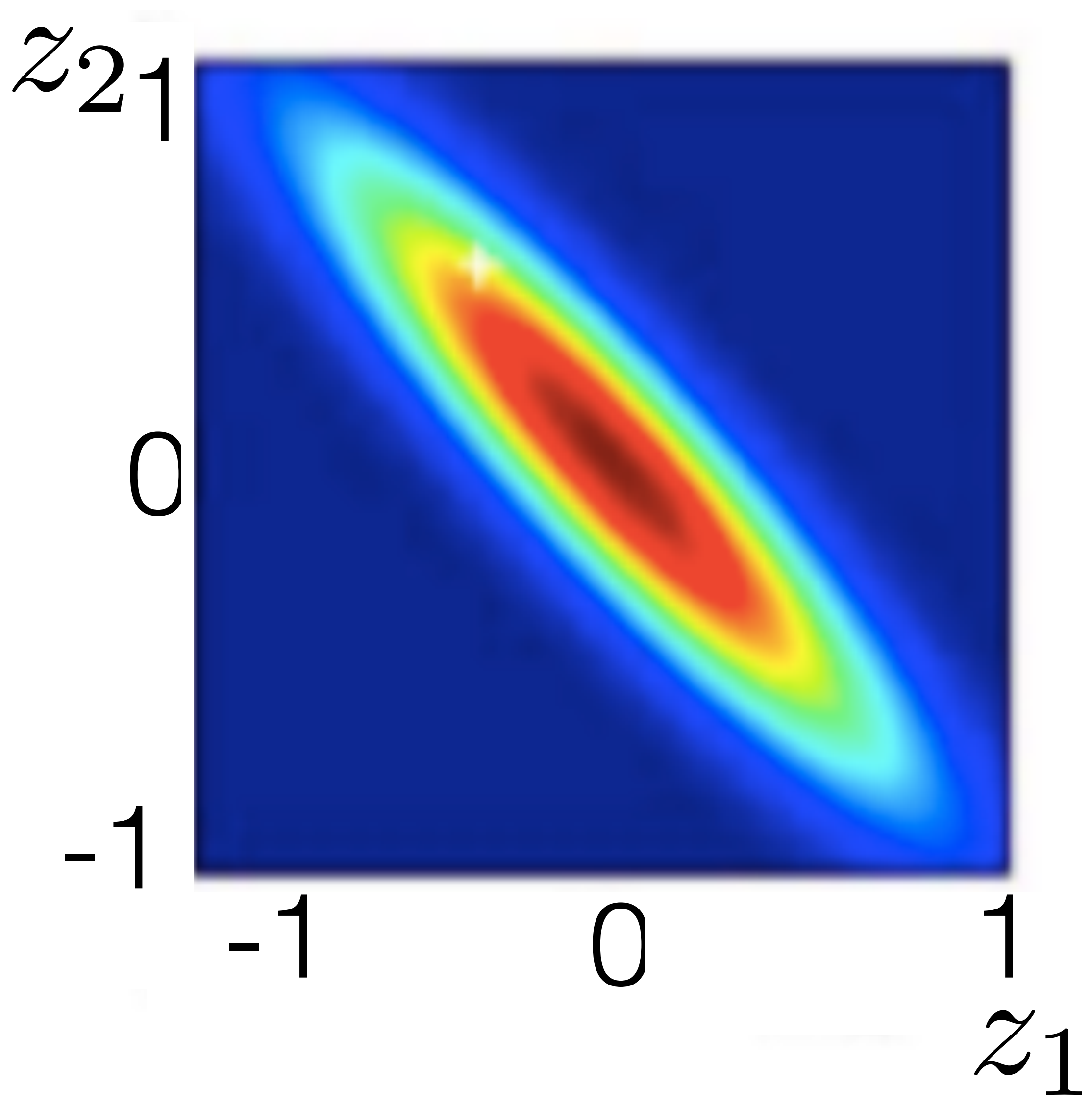
$$\mathcal{N}(\mu, \Sigma) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp(-\frac{1}{2}(z-\mu)^T \Sigma^{-1}(z-\mu)),$$

$$z \in \mathbb{R}^d$$
$$\mu \in \mathbb{R}^d$$
$$\Sigma \in \mathbb{R}^{d \times d}$$





diagonal Σ



non-diagonal Σ

Images from https://medium.com/ming-learns-thing/machine-learning-bayesian-linear-regression-f160c4eaef99, C. Bishop. Pattern Recognition and Machine Learning

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:

$$p(Y|X, w) = \prod_{i=1}^{N} \mathcal{N}(y_i|x_i^T w, 1) =$$

$$= \mathcal{N}(Y|Xw, I)$$

- 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:

$$p(Y | X, w) = \prod_{i=1}^{N} \mathcal{N}(y_i | x_i^T w, 1) =$$

$$= \mathcal{N}(Y | Xw, I)$$

- 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:

$$p(Y|X, w) = \prod_{i=1}^{N} \mathcal{N}(y_i|x_i^T w, 1) =$$

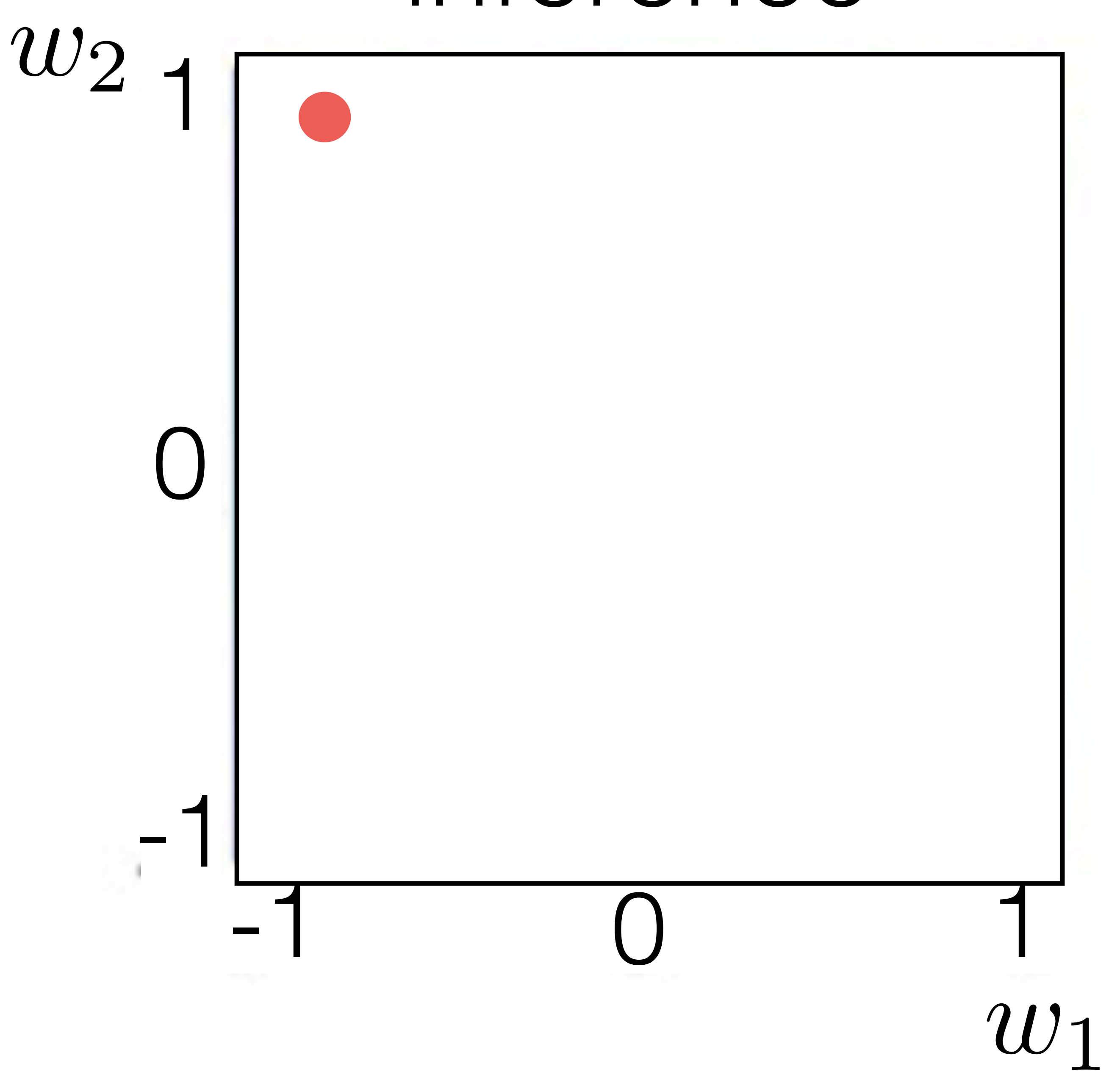
$$= \mathcal{N}(Y|Xw, I)$$

- conjugate prior:

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

Options of training Bayesian linear regression

Maximum likelihood inference

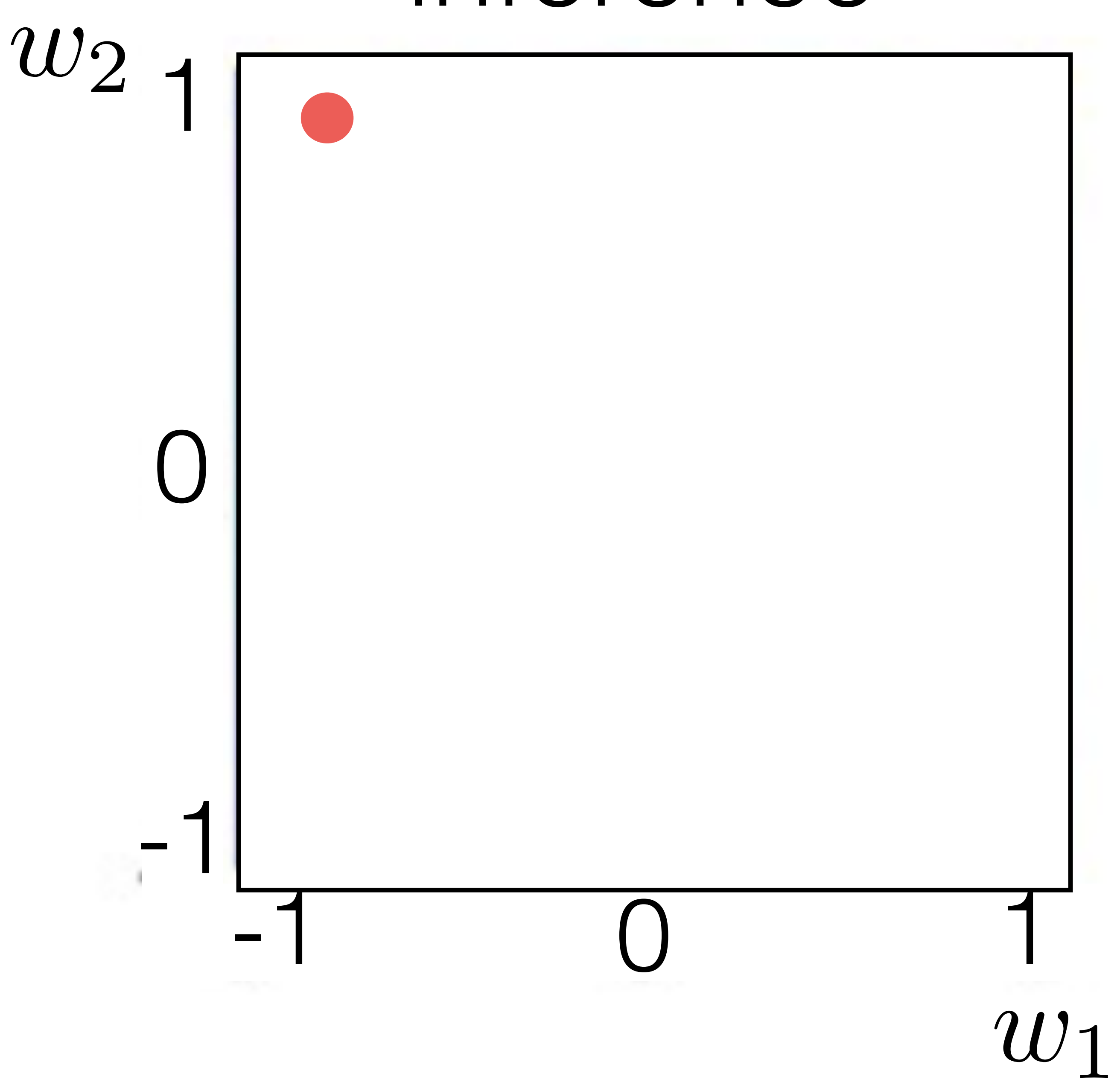


$$p(Y|X, w) \to \max_w$$

Image draft from C. Bishop. Pattern Recognition and Machine Learning

Options of training Bayesian linear regression

Maximum likelihood inference



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

Prior distribution

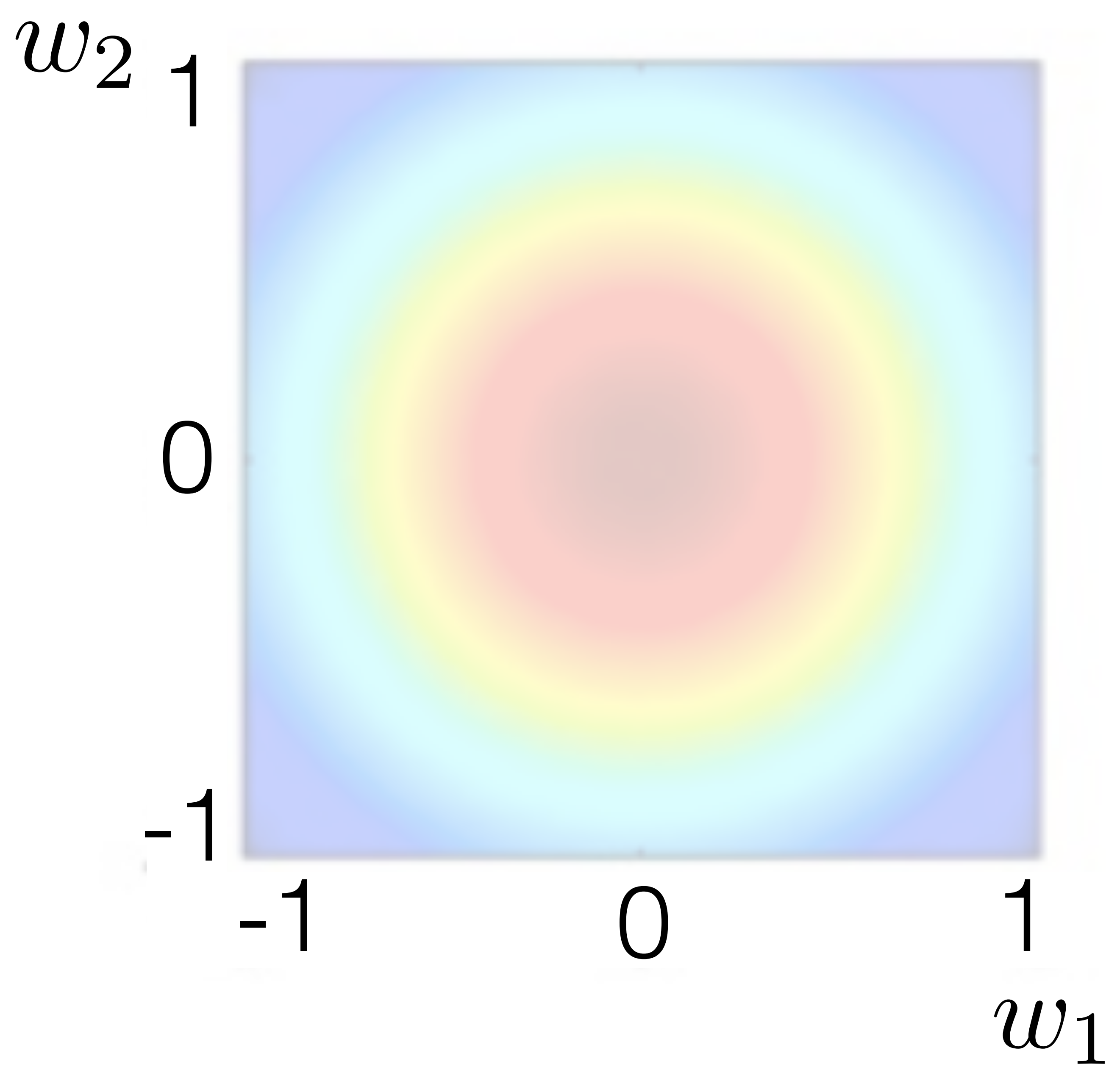
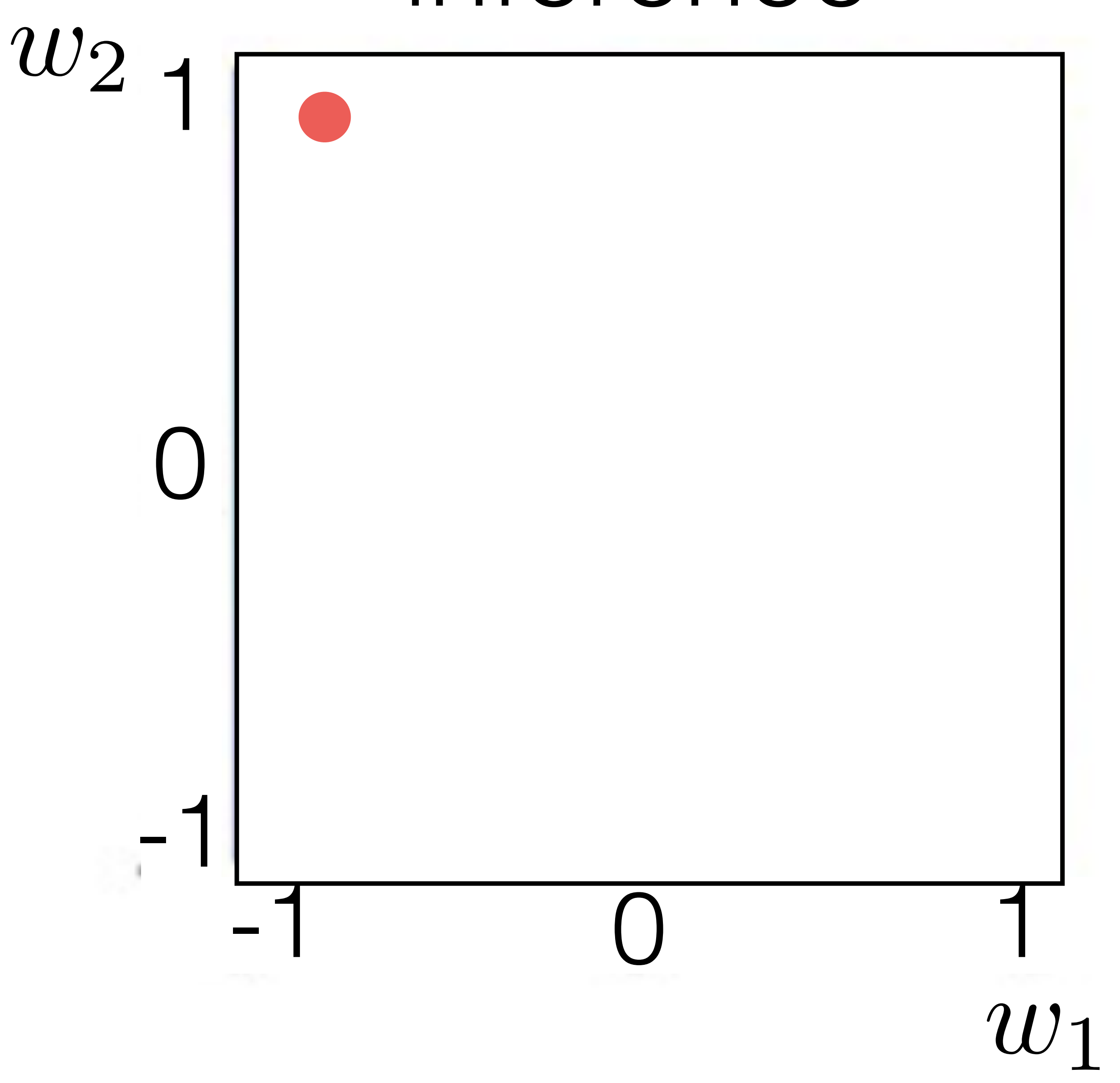


Image draft from C. Bishop. Pattern Recognition and Machine Learning

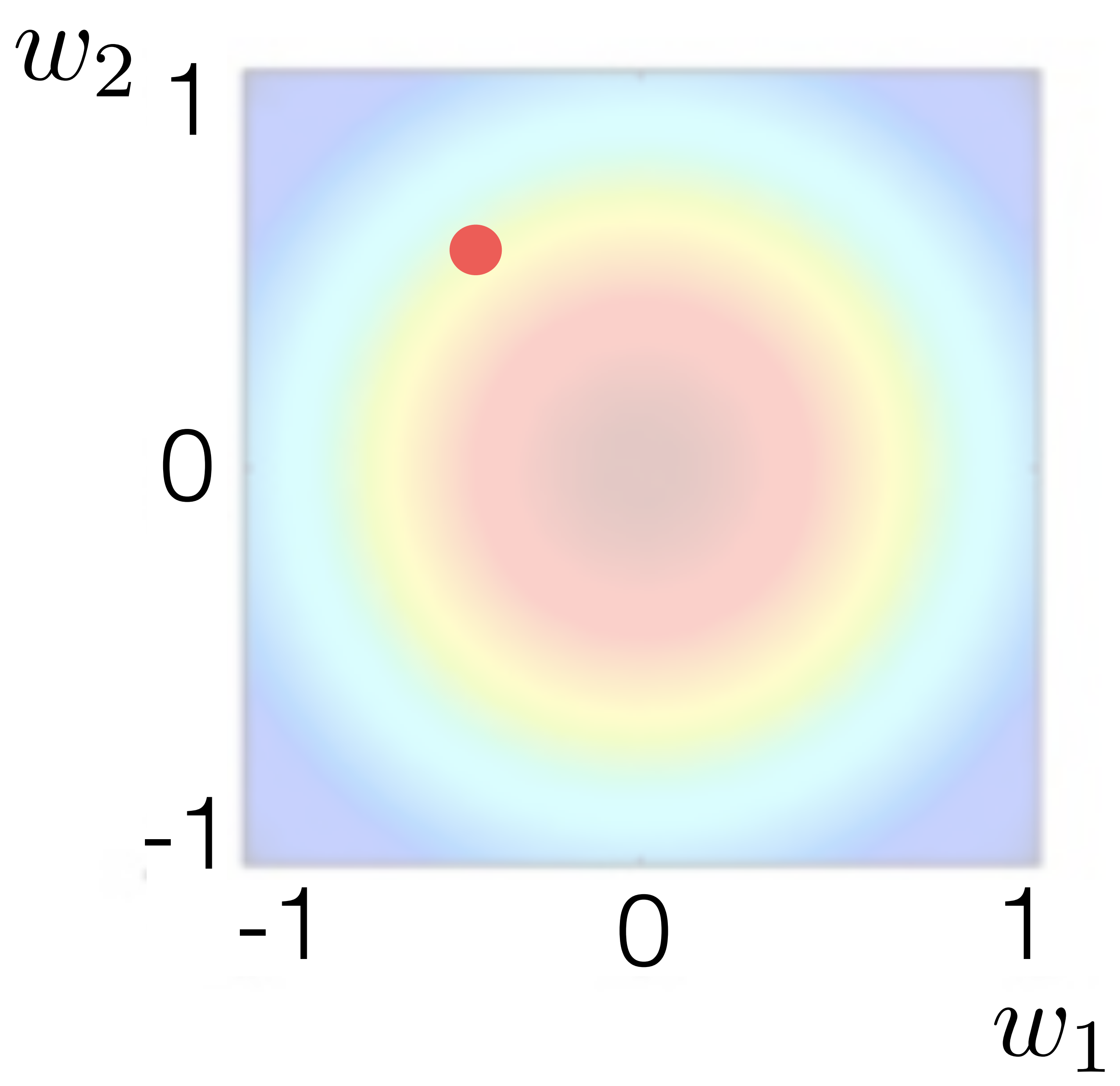
Options of training Bayesian linear regression

Maximum likelihood inference



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

Maximum posterior inference

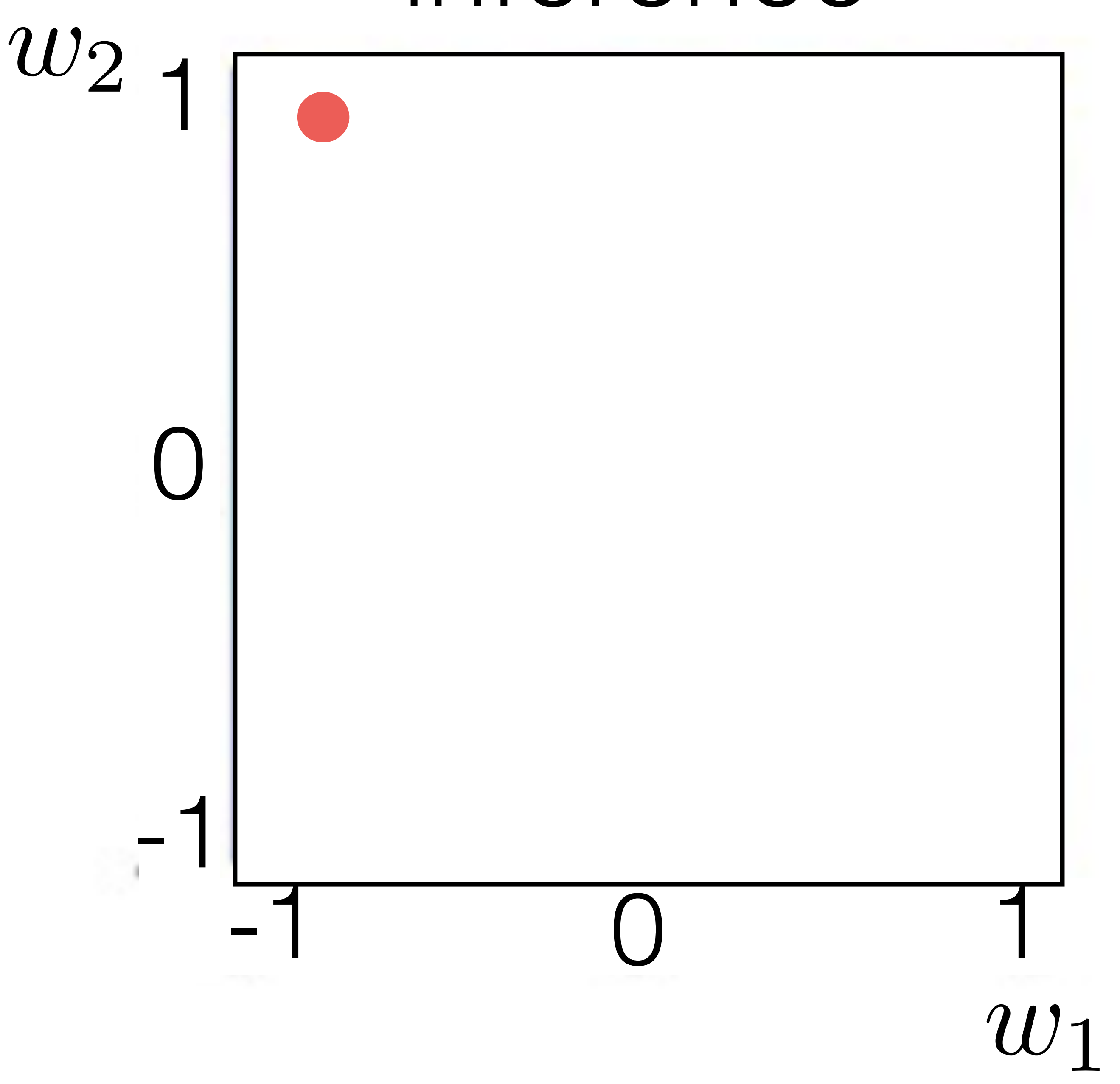


$$p(Y|X, w)p(w) \to \max_w$$

Image draft from C. Bishop. Pattern Recognition and Machine Learning

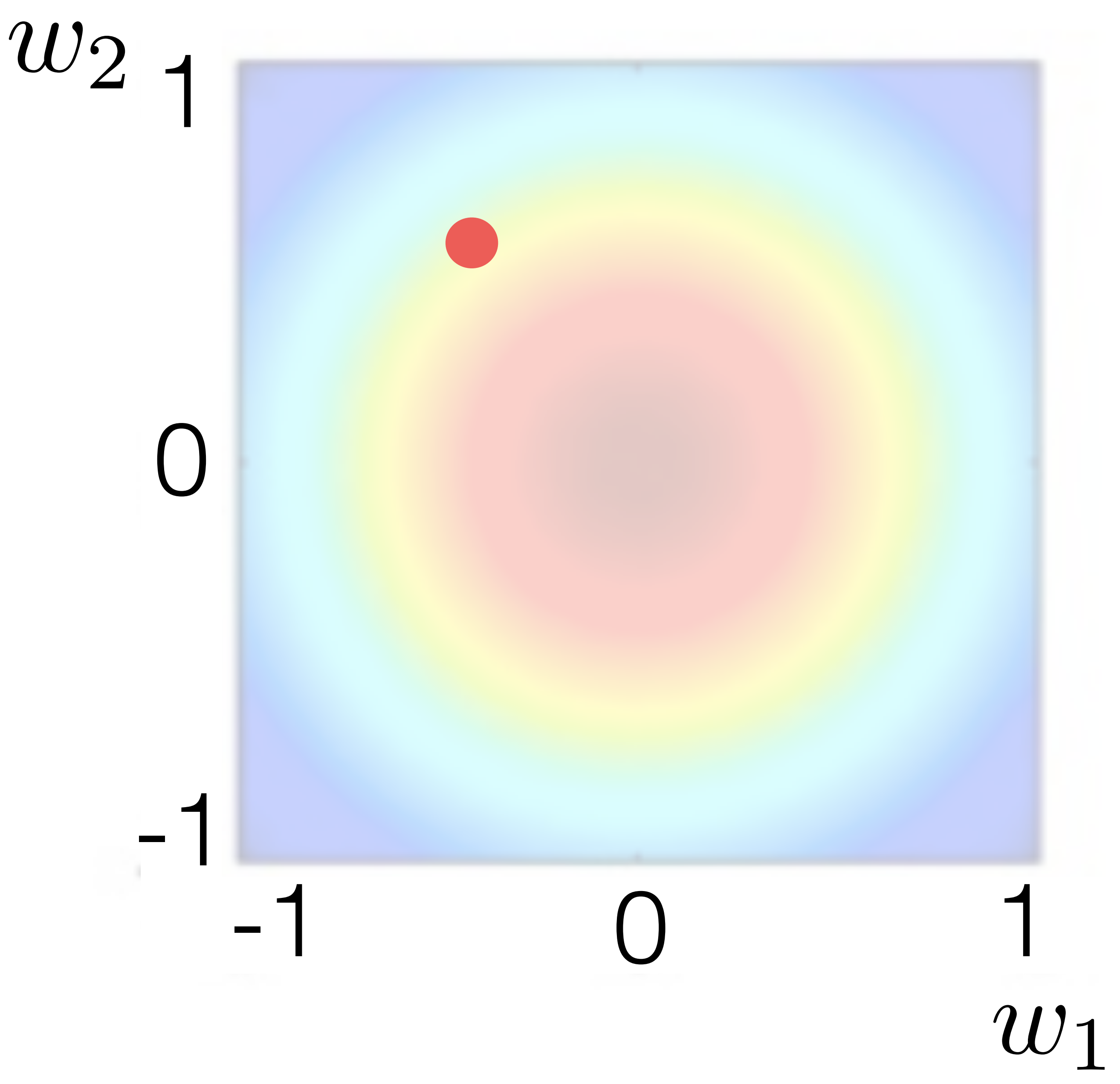
Options of training Bayesian linear regression

Maximum likelihood inference



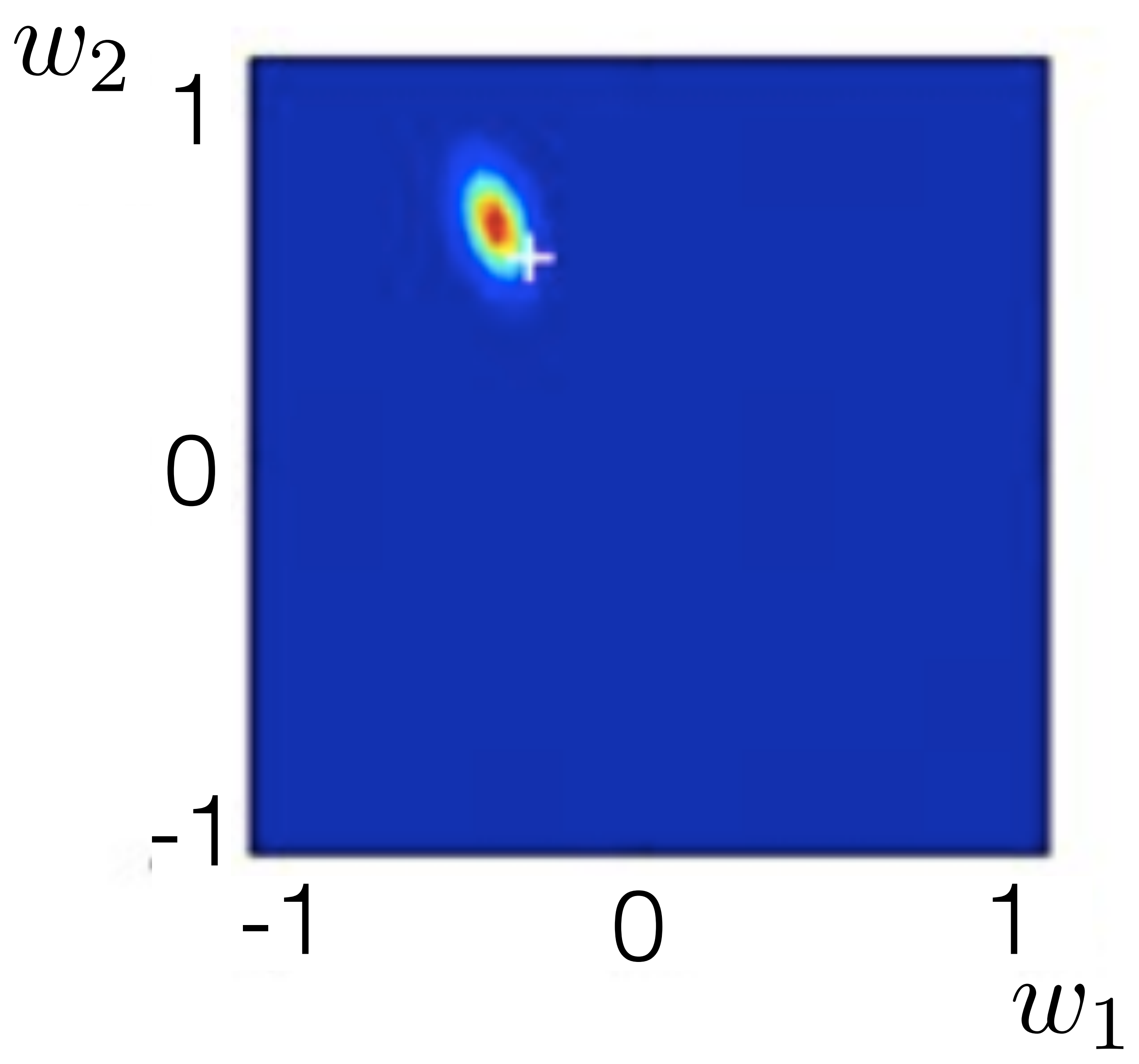
$$p(Y|X,w) \rightarrow \max_w$$

Maximum posterior inference



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

Full Bayesian inference: posterior distribution



$$p(w|X,Y) \propto$$
$$\propto p(Y|X,w)p(w)$$

Image draft from C. Bishop. Pattern Recognition and Machine Learning

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

Modified slide from Ekaterina Lobacheva's presentation

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

Modified slide from Ekaterina Lobacheva's presentation

Bayesian linear regression: training (ML)

Maximum likelihood inference: $\log p(Y|X, w) \to \max\limits_{w}$

Bayesian linear regression: training (ML)

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

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

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

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

Bayesian linear regression: training (ML)

Maximum likelihood inference: $\log p(Y|X,w) \to \max\limits_{w}$

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

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

$$\Updownarrow$$

$$\|Y - Xw\|^2 \to \min_{w}$$ **— 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

Modified slide from Ekaterina Lobacheva's presentation

Bayesian linear regression: training (MP)

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

$$\underset{\displaystyle\max_w}{\underset{\downarrow}{p(w|X,Y)}} = \frac{p(Y|X,w)p(w)}{p(Y|X)}$$

Bayesian linear regression: training (MP)

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

$$\underset{\max_w}{\underset{\downarrow}{p(w|X,Y)}} = \frac{\boxed{p(Y|X,w)p(w)}}{\boxed{p(Y|X)}} \to \max_w$$

does not depend on weights

Bayesian linear regression: training (MP)

Maximum posterior inference: $\log[p(Y|X,w)p(w)] \to \max_{w}$

Bayesian linear regression: training (MP)

Maximum posterior inference: $\log[p(Y|X,w)p(w)] \to \max\limits_{w}$

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

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

Bayesian linear regression: training (MP)

Maximum posterior inference: $\log[p(Y|X,w)p(w)] \to \max\limits_{w}$

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

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

$$\Updownarrow$$

$$\|Y - Xw\|^2 + \frac{1}{\alpha}\|w\|^2 \to \min_{w}$$

— we already know the solution!

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

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

Modified slide from Ekaterina Lobacheva's presentation

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:

$$p(Y|X, w) = \prod_{i=1}^{N} \mathcal{N}(y_i|x_i^T w, 1) =$$

$$= \mathcal{N}(Y|Xw, I)$$

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

$$p(w|X,Y) \propto p(Y|X,w)p(w) \propto$$

$$\text{Const} \cdot \exp\left(-\frac{1}{2}(Y - Xw)^T(Y - Xw)\right)\exp\left(-\frac{1}{2\alpha}w^T w\right) =$$

Bayesian linear regression: training

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

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

$$p(w|X,Y) \propto p(Y|X,w)p(w) \propto$$

$$\mathrm{Const} \cdot \exp\left(-\frac{1}{2}(Y - Xw)^T(Y - Xw)\right)\exp\left(-\frac{1}{2\alpha}w^T w\right) =$$

$$\mathrm{Const} \cdot \exp\underbrace{\left(-\frac{1}{2}w^T(X^T X + \frac{1}{\alpha}I)w + w^T X^T Y\right)}_{\text{quadratic form w.r.t weights}}$$

quadratic form w.r.t weights ➡ **normal distribution**

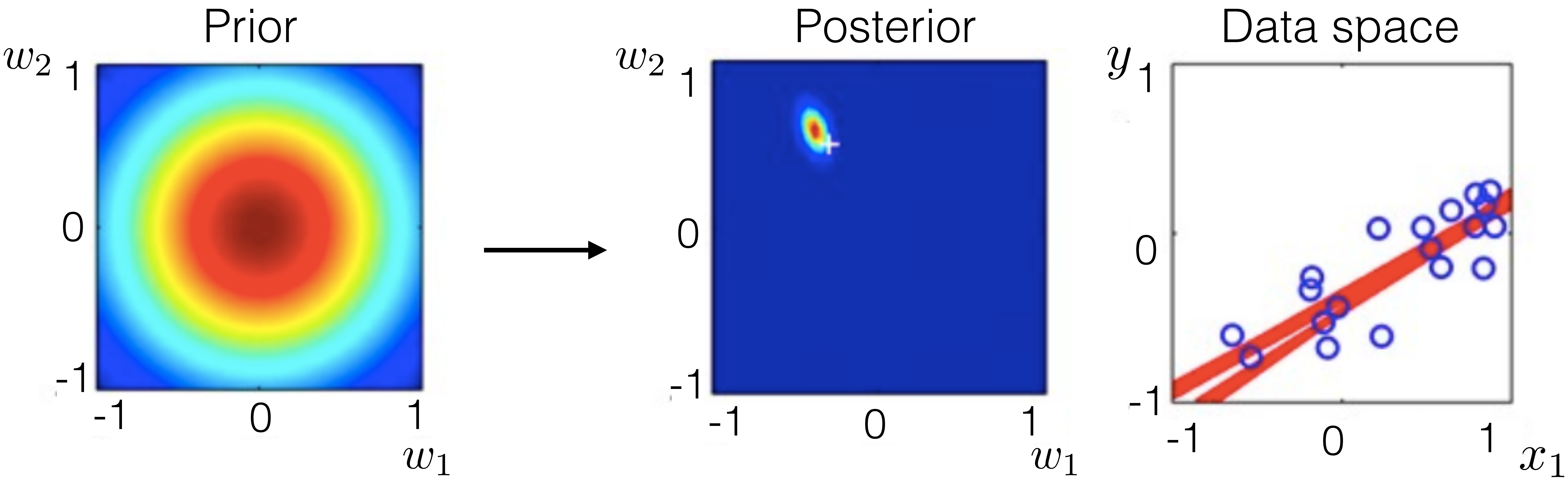
Bayesian linear regression: training

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

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

$$\boxed{\begin{gathered} p(w|X,Y) = \mathcal{N}(w|w_{MP}, \Sigma) \\ w_{MP} = (X^T X + \tfrac{1}{\alpha} I)^{-1} X^T Y \\ \Sigma = \left(X^T X + \frac{1}{\alpha} I\right)^{-1} \end{gathered}}$$

Training visualization



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

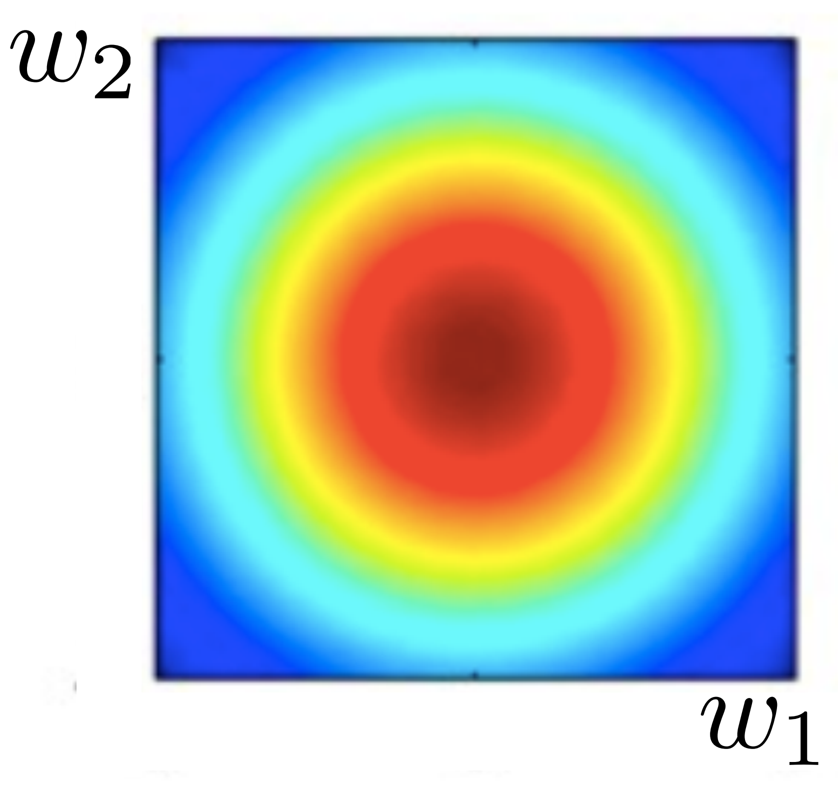
$$p(w|X,Y) = \mathcal{N}(w|w_{MP}, \Sigma)$$

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

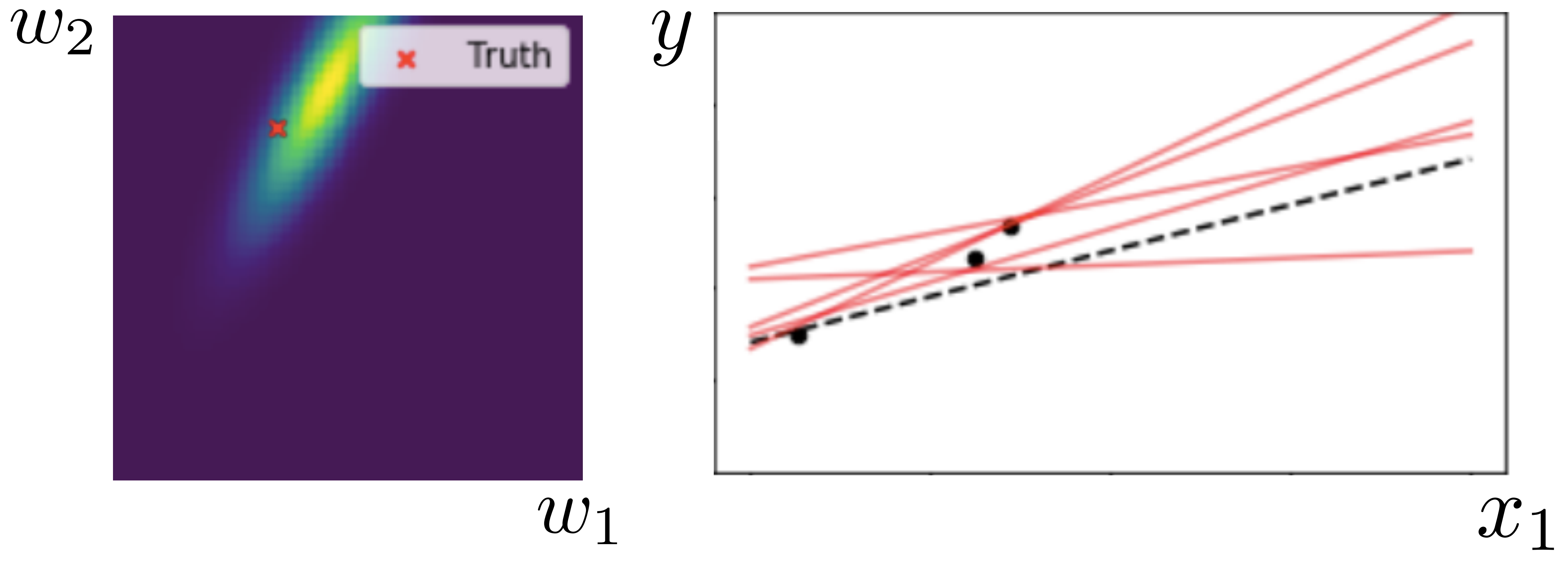
Image from C. Bishop. Pattern Recognition and Machine Learning

Training: increasing amount of data

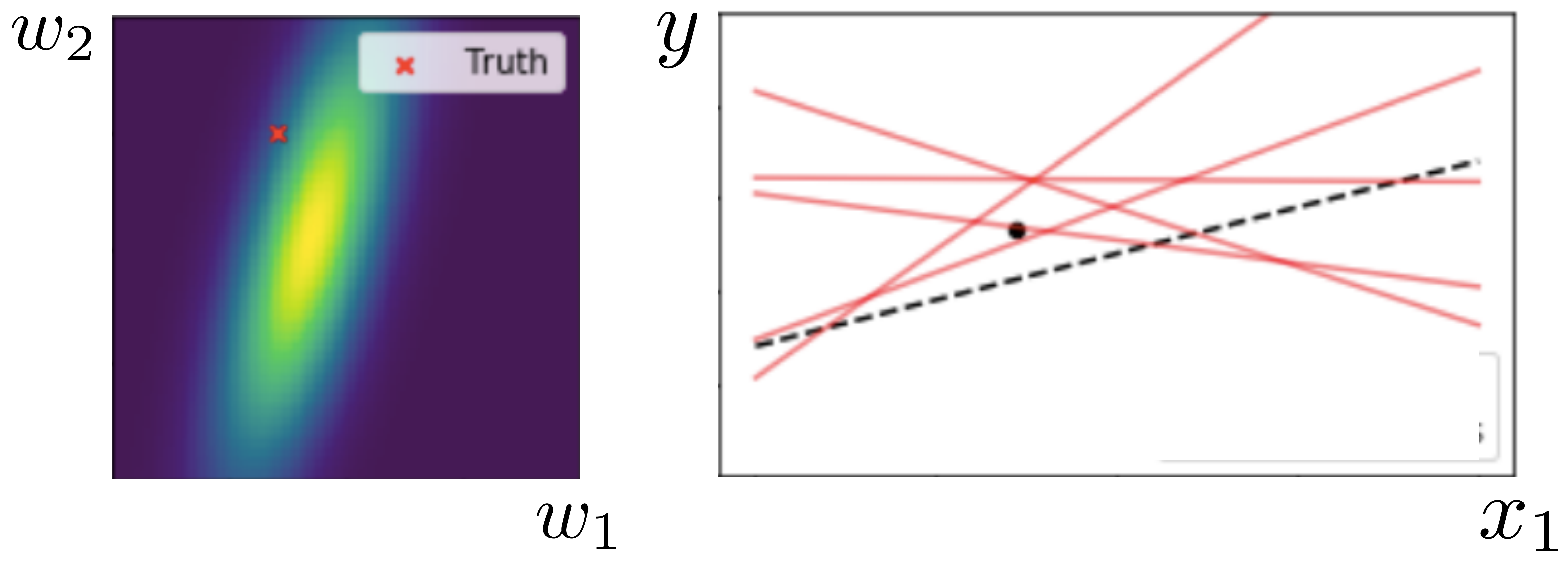
Prior:



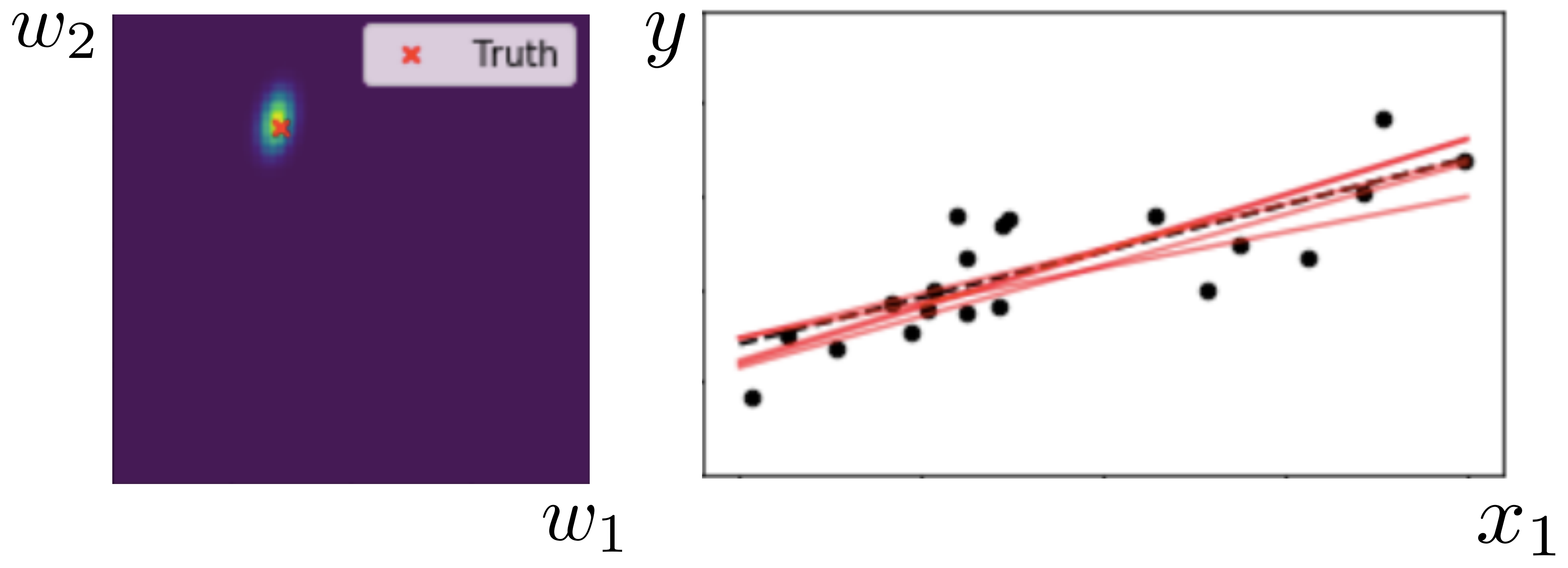
3 data points (N=3):



1 data point (N=1):



20 data points (N=20):



Images from http://krasserm.github.io/2019/02/23/bayesian-linear-regression/

Bayesian linear regression

Given:

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

$Y \in \mathbb{R}^{N}$ — target values

Model:

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

Training:

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

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

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

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

Modified slide from Ekaterina Lobacheva's presentation

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

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

Modified slide from Ekaterina Lobacheva's presentation

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

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

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

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

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

$$\mathbb{D}y = \mathbb{E}(y - \mathbb{E}y)(y - \mathbb{E}y)^T = \mathbb{E}(Ax + \epsilon - A\mu)(Ax + \epsilon - A\mu)^T = \mathbb{E}\big(A(x-\mu) + \epsilon\big)\big(A(x-\mu) + \epsilon\big)^T =$$

$$= \mathbb{E}A(x-\mu)(x-\mu)^T A^T + 2\mathbb{E}\epsilon\mathbb{E}A(x-\mu) + \mathbb{E}\epsilon\epsilon^T =$$

$$= A\Sigma A^T + \Gamma$$

Convolution of normal distributions

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

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

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

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

Bayesian linear regression: prediction

$$p(y_*|x_*, X, Y) = \int p(y_*|x_*, w)p(w|X, Y)dw =$$

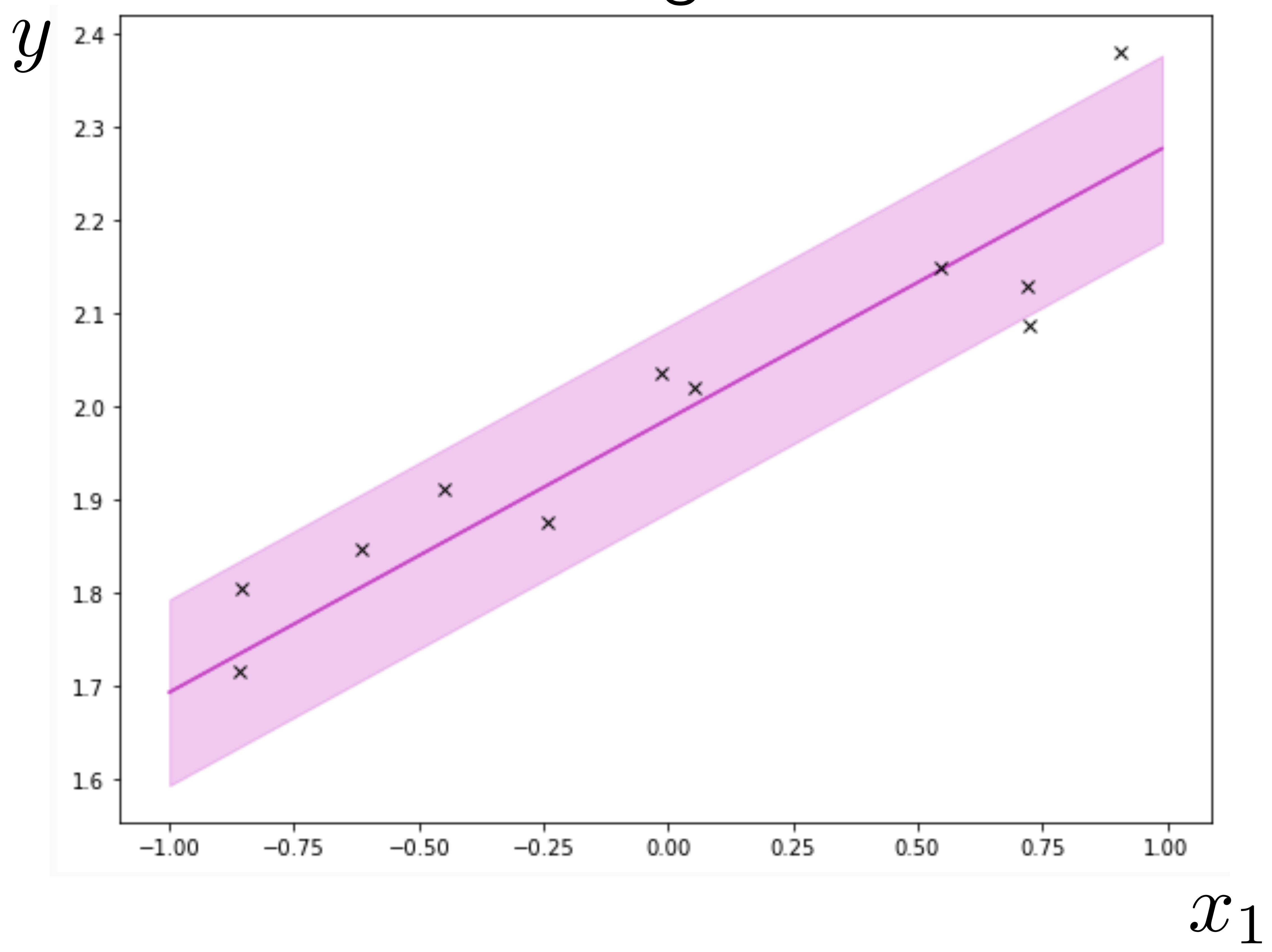
$$\int \mathcal{N}(y_*|x_*^T w, 1)\mathcal{N}(w|w_{MP}, \Sigma)dw =$$

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

x_* — new object

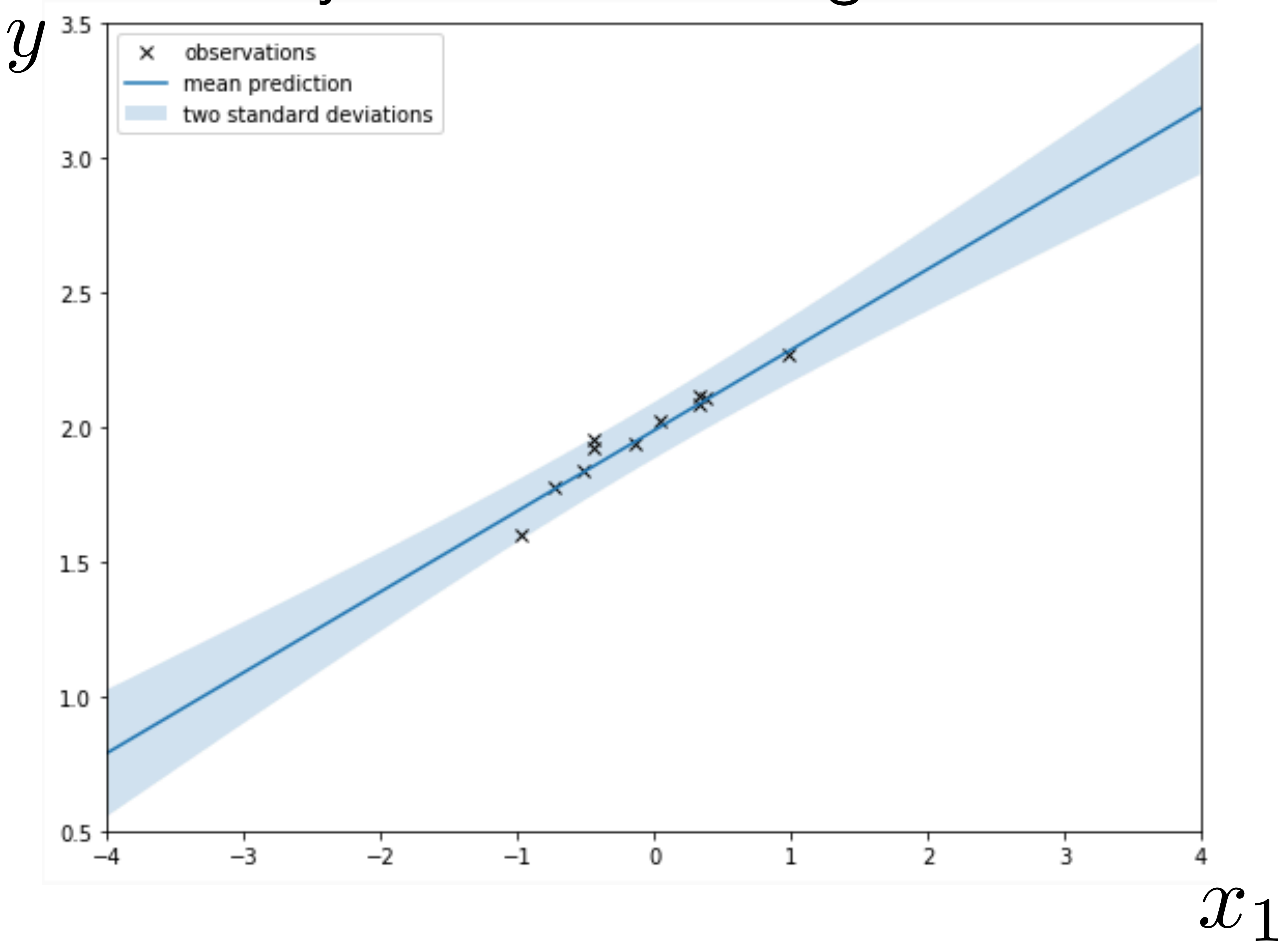
Prediction visualization

Linear regression



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

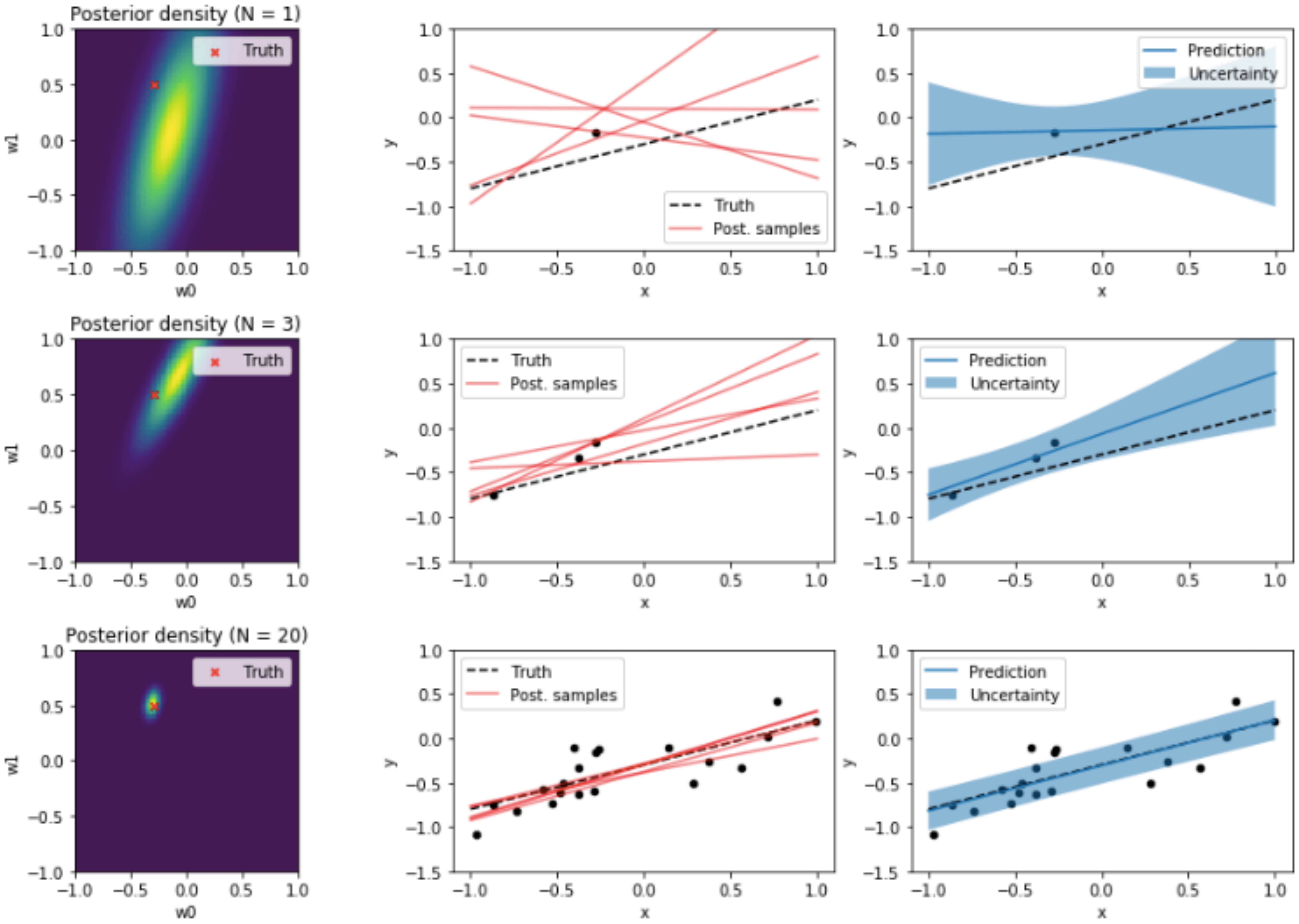
Bayesian linear regression



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

Image from https://jessicastringham.net/2018/01/10/bayesian-linreg-plots/

Prediction: increasing amount of data



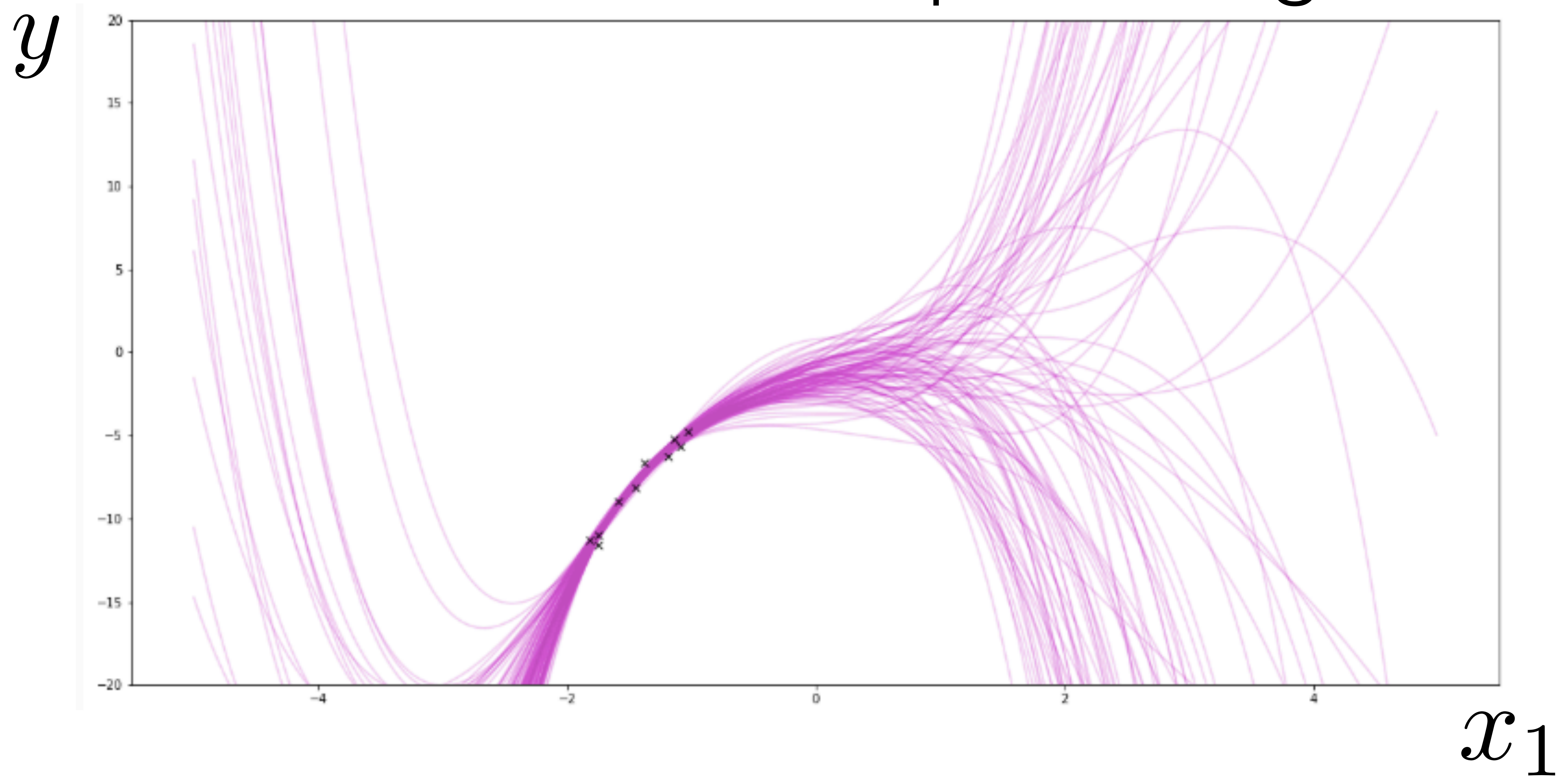
Images from http://krasserm.github.io/2019/02/23/bayesian-linear-regression/

Prediction: polynomial features

Modify training data: add polynomial features

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

Prediction with sampled weights



Mean ± Std of predictions

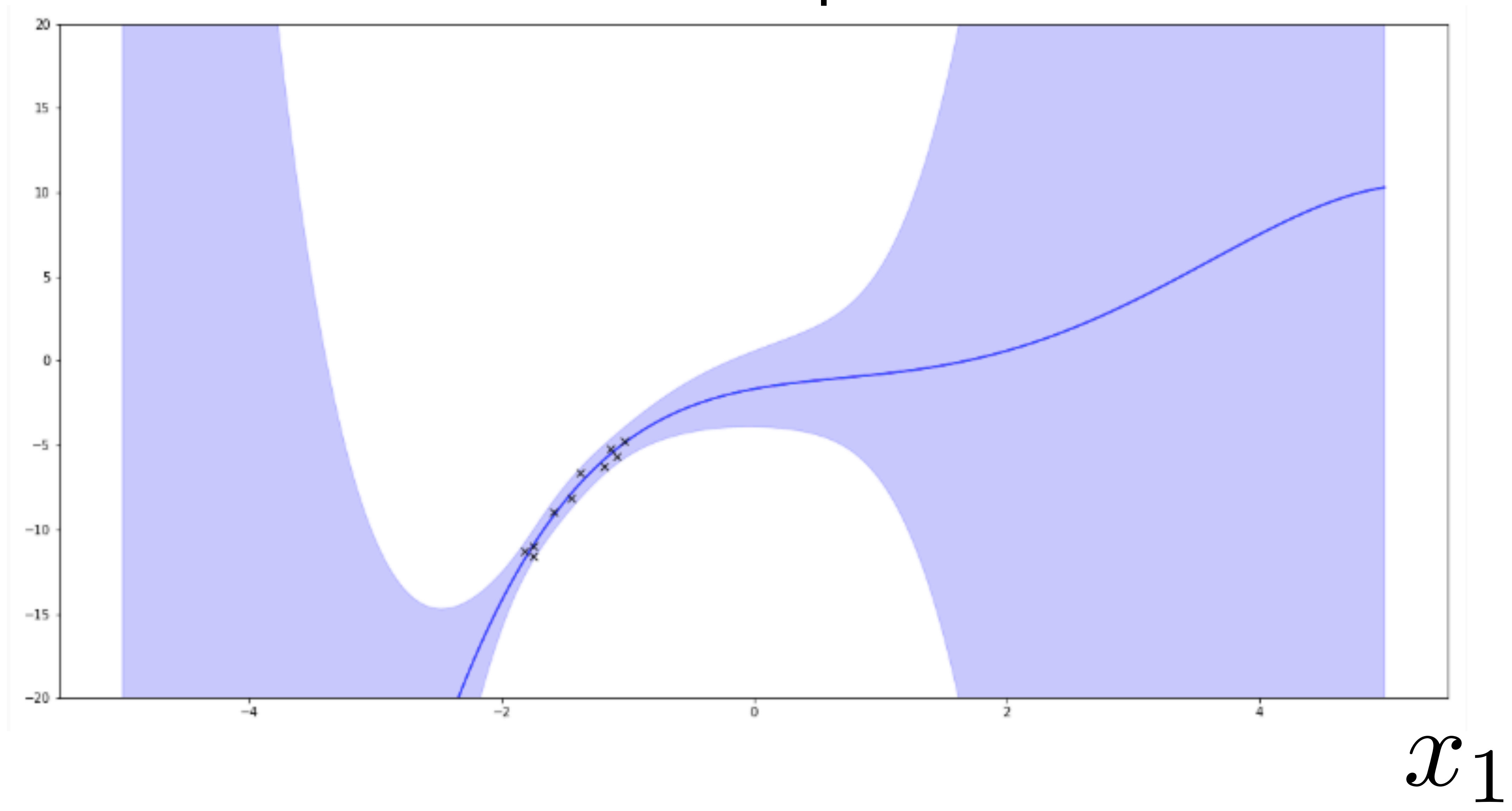


Image from https://jessicastringham.net/2018/01/10/bayesian-linreg-plots/

Bayesian linear regression

Given:

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

$Y \in \mathbb{R}^{N}$ — target values

Model:

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

Training:

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

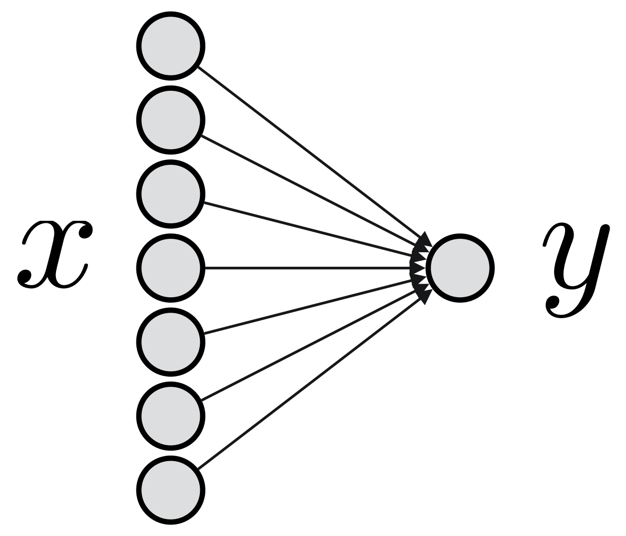
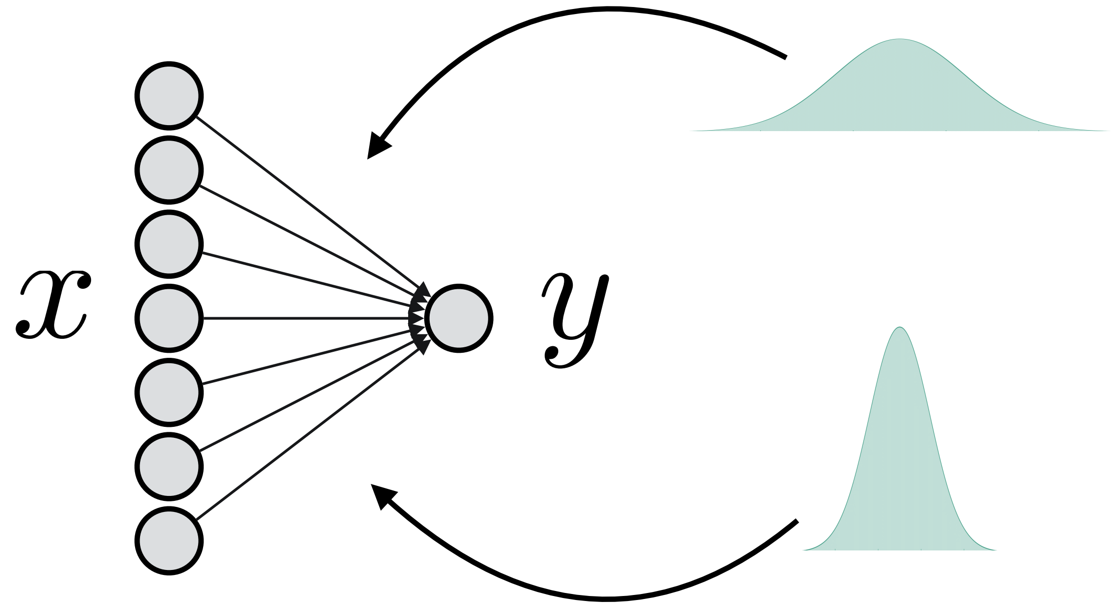
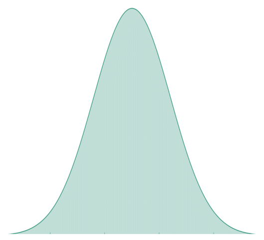
$$w_{MP} = (X^T X + \tfrac{1}{\alpha} I)^{-1} X^T Y$$

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

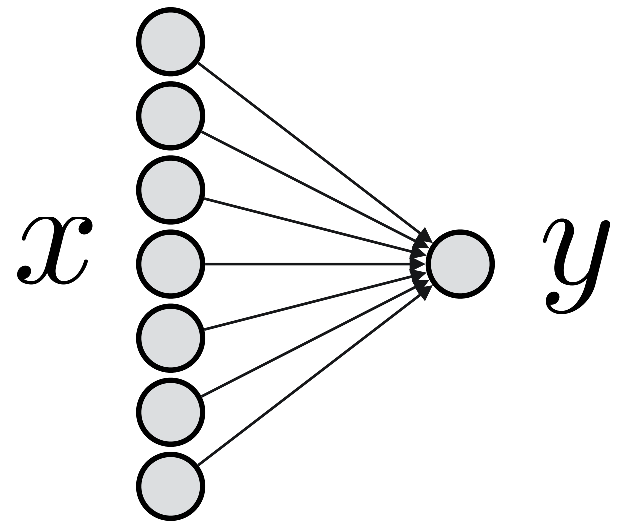
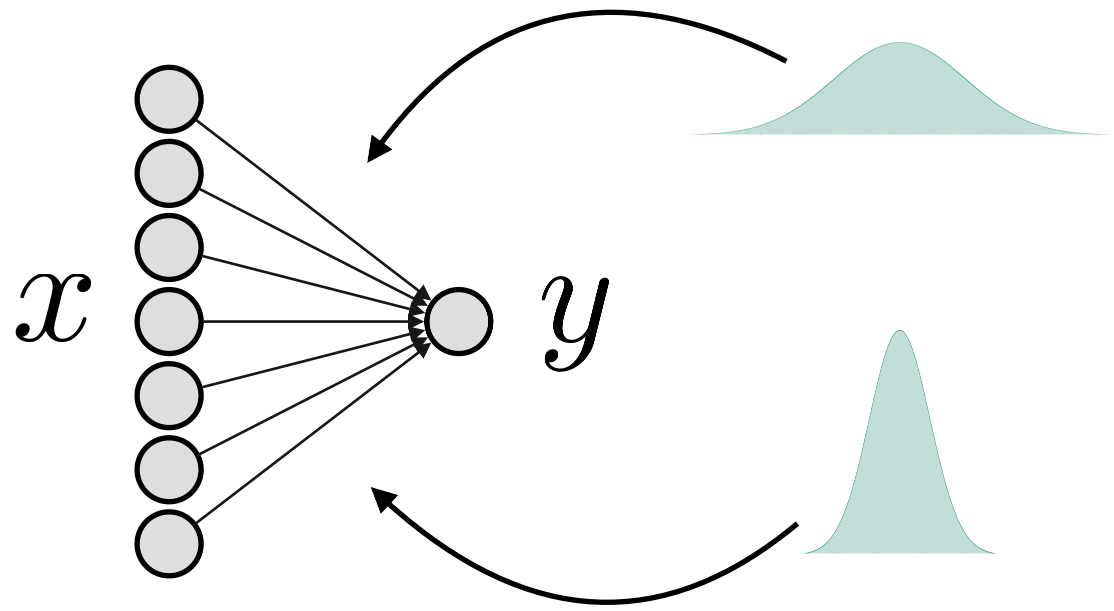
Prediction on a new object x_* :

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

Putting everything together

	Linear regression	Bayesian linear regression
Weights:	deterministic weights 	stochastic weights 
Training:	Data X, Y → point estimate w_{MP}	 Prior $p(w)$ + Data X, Y →  Posterior $p(w \mid X, Y)$
Prediction on a new object x_* :	$p(y_* \mid x_*, w_{MP})$	$\mathbb{E}_{p(w \mid X,Y)} p(y_* \mid x_*, w)$

Putting everything together

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

Choosing hyperparameters in Bayesian approach

Given:

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

$Y \in \mathbb{R}^{N}$ — target values

Model:

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

Training:

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

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

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

Prediction on a new object x_* :

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

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:

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

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_* :

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

Any hyperparameters?

Choosing hyperparameters in Bayesian approach

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

Prior: $p(\theta|\boxed{\alpha})$

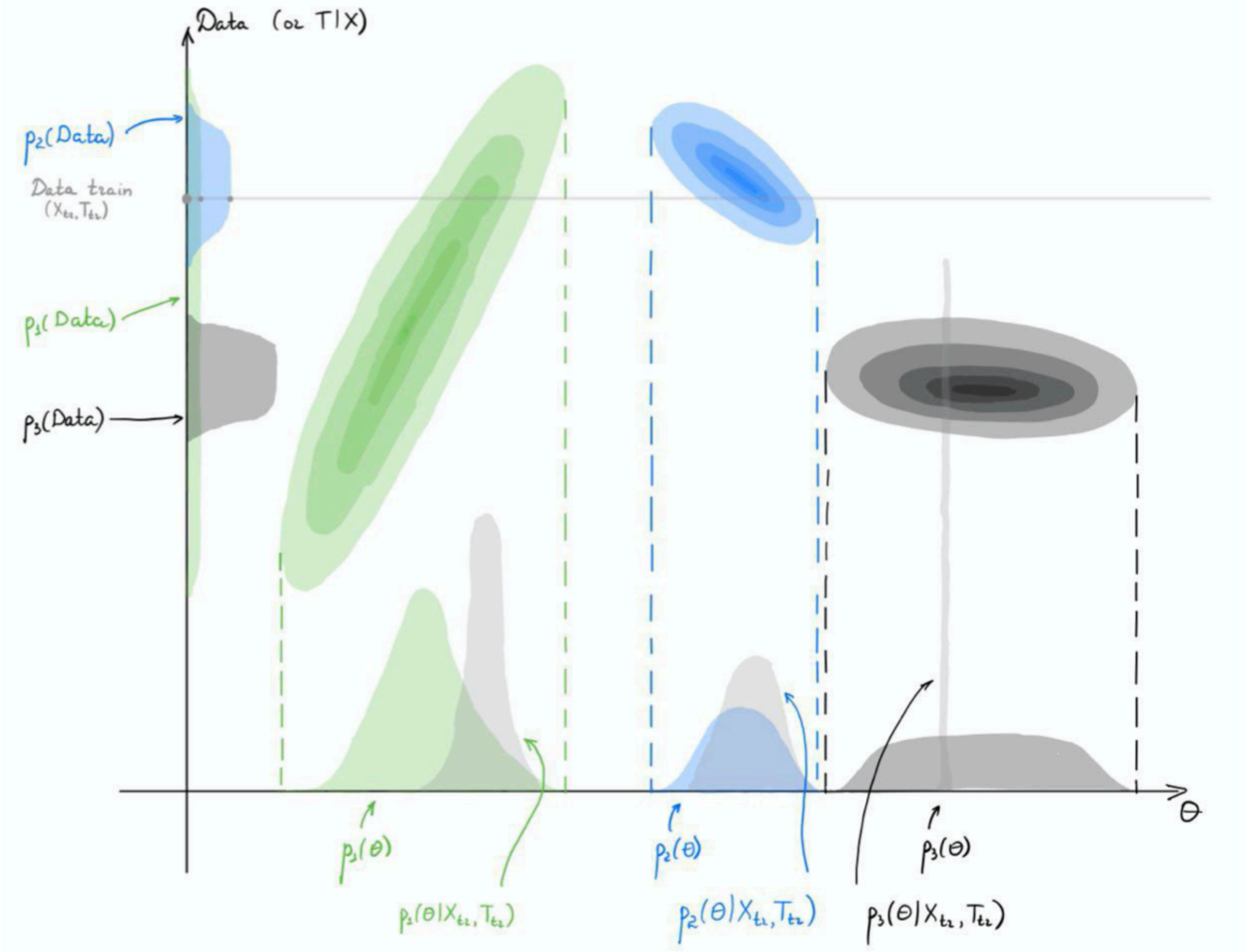
Training:

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

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

Maximum evidence: $\int p(Y|X,\tilde{\theta})p(\tilde{\theta}|\alpha)d\tilde{\theta} \to \max_{\alpha}$

Maximum evidence



Relevance vector machine

Given:

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

$Y \in \mathbb{R}^{N}$ — target values

Model:

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

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

A–?

Relevance vector machine

Given:

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

$Y \in \mathbb{R}^{N}$ — target values

Model:

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

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

A—?

Maximum evidence:

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

Relevance vector machine

Given:

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

$Y \in \mathbb{R}^{N}$ — target values

Model:

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

Training:

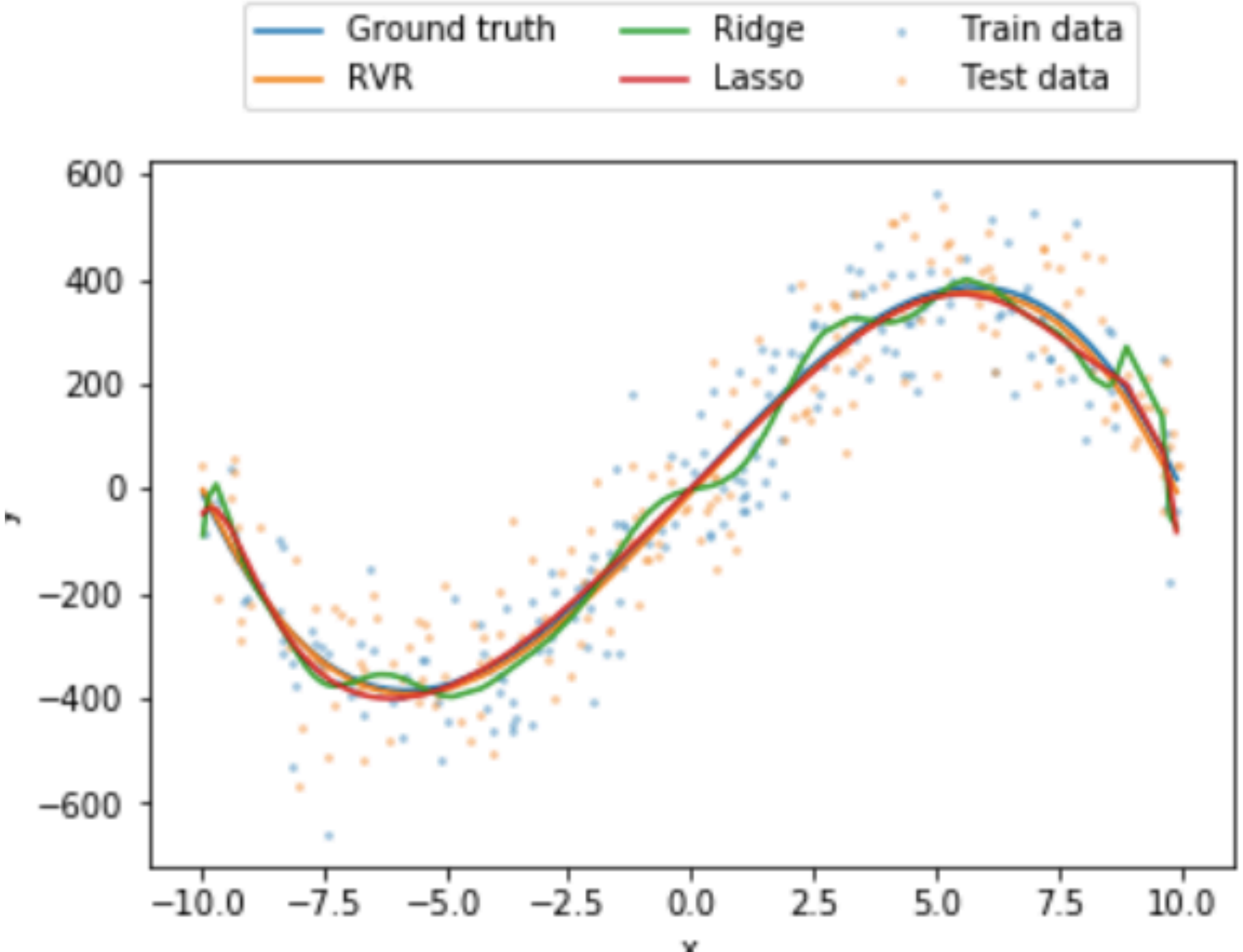
Repeat:

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

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

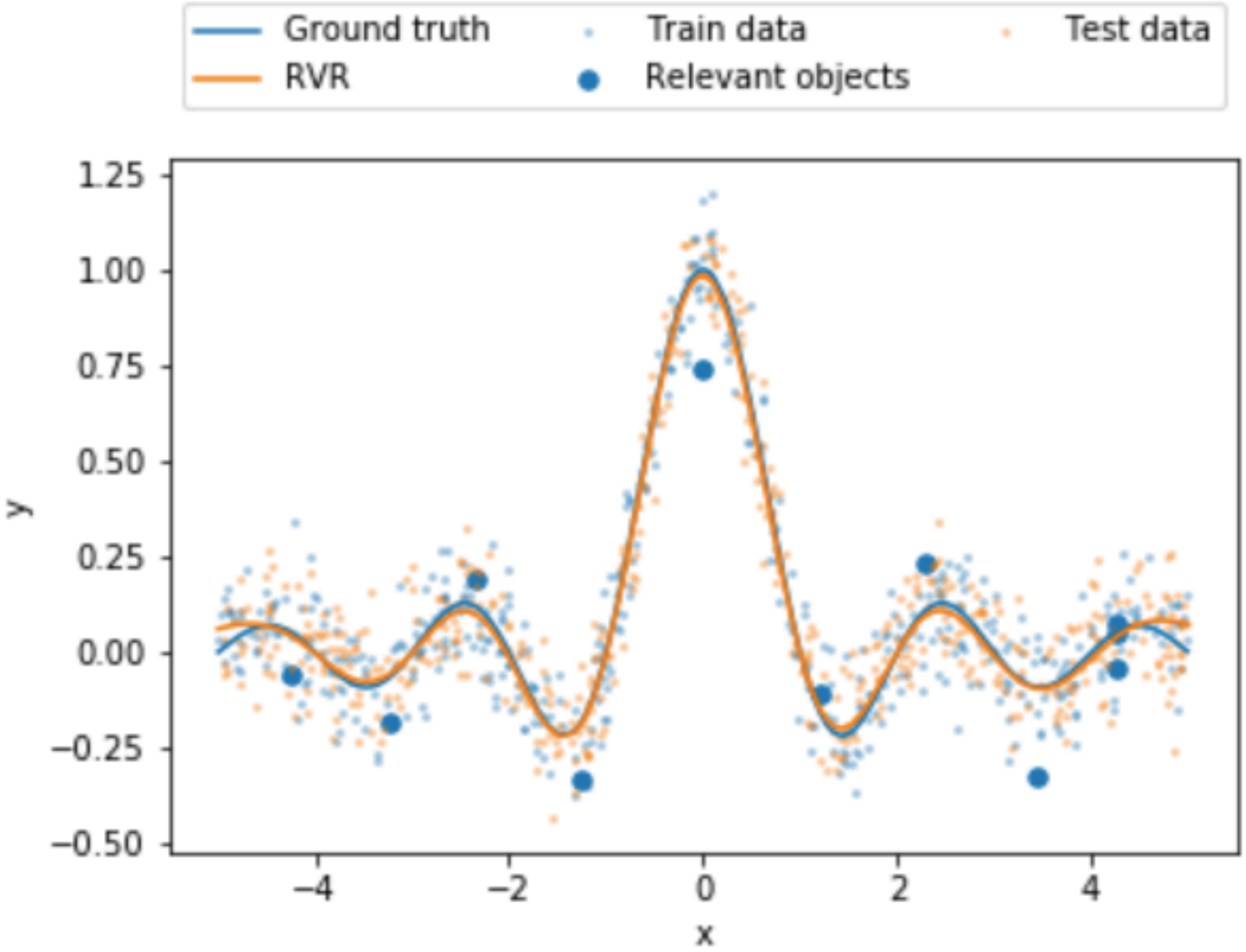
Relevance vector machine: visualization

$\alpha_i \to 0 \Rightarrow i$ -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:

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