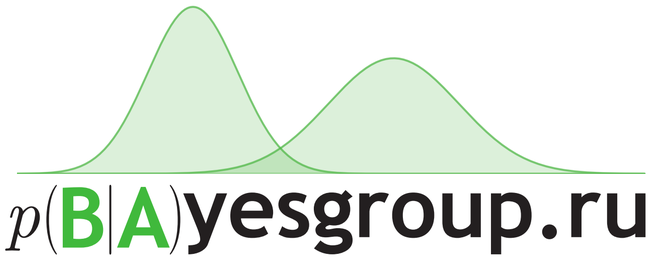
Bayesian linear regression

Nadezhda Chirkova

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

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

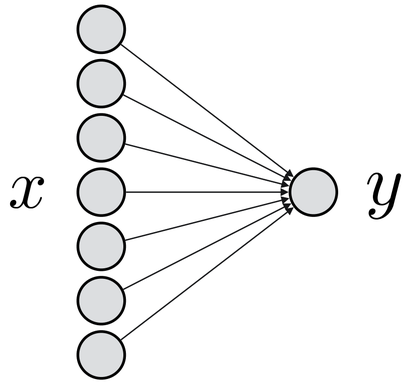
N — number of objects

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

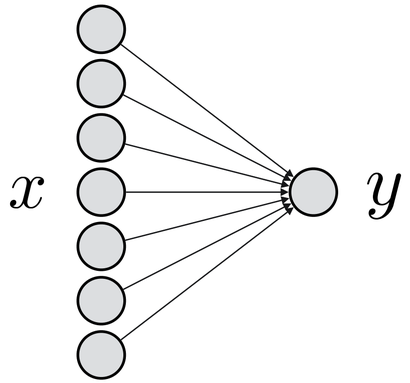
N — number of objects

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

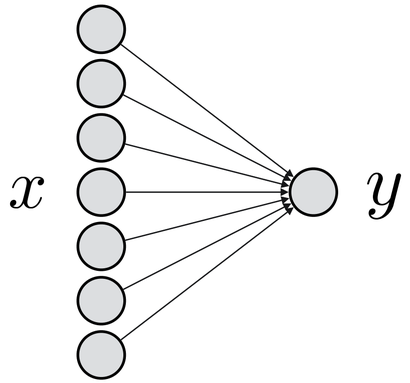
N — number of objects

d — number of features

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

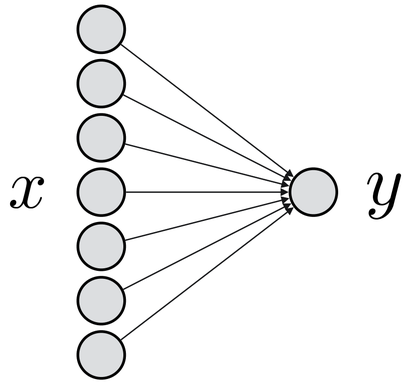
N — number of objects

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:

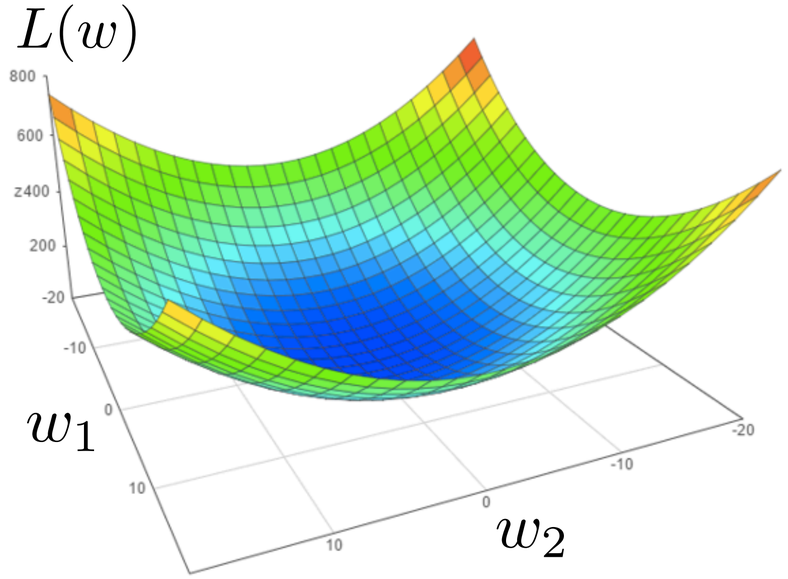


Image from https://www.globalsoftwaresupport.com/linear-regression/

Linear regression: training

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

Optimal weights:

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

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

Convex function:

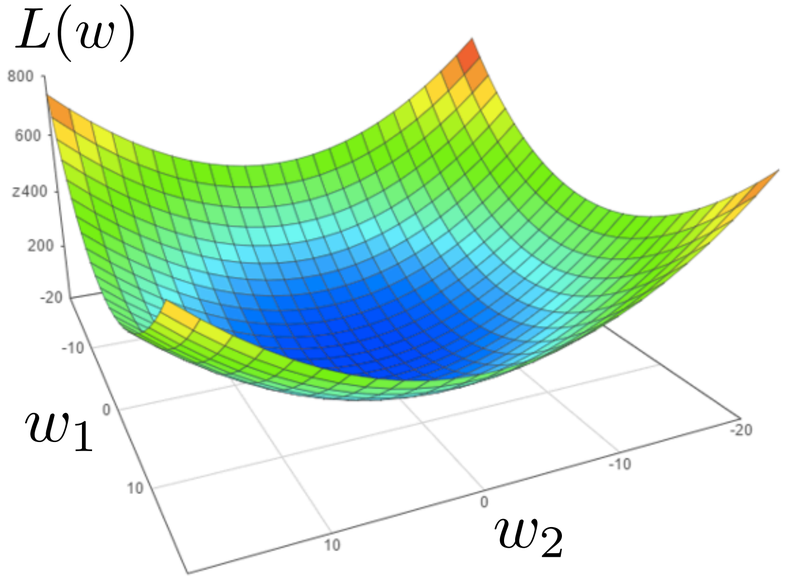


Image from https://www.globalsoftwaresupport.com/linear-regression/

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:

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

- Always unique solution
- Preventing overfitting

Strongly convex function:

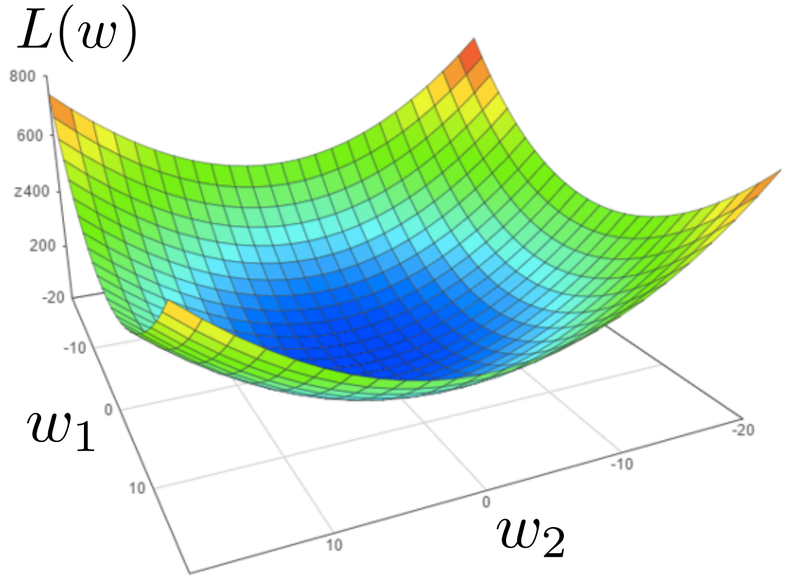


Image from https://www.globalsoftwaresupport.com/linear-regression/

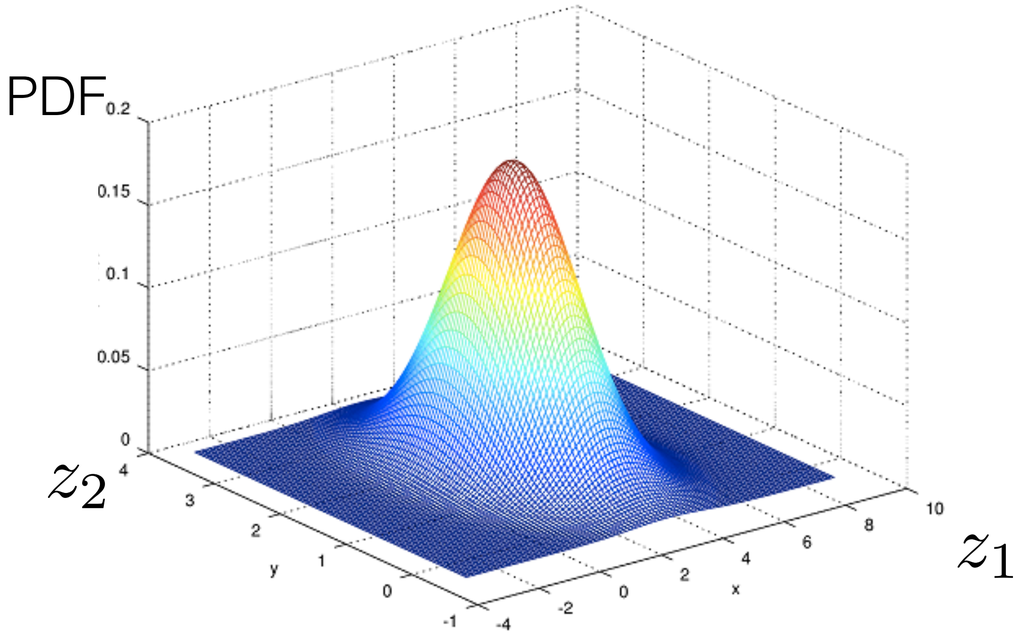
Plan

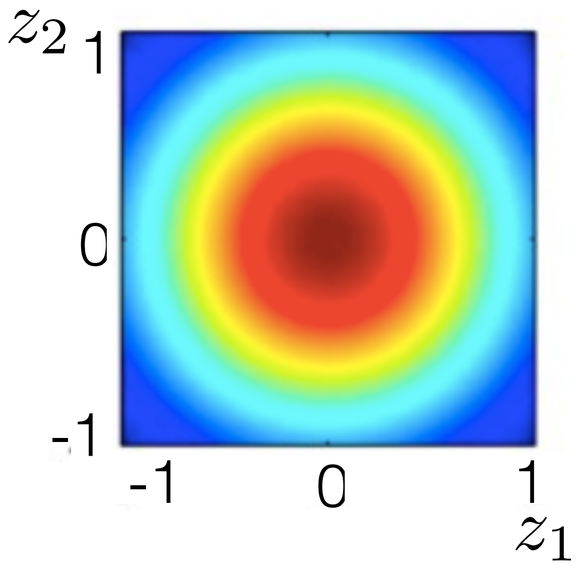
- Linear regression: reminder
- Bayesian linear regression:
    - model definition
    - training
    - prediction

Multivariate normal (Gaussian) distribution

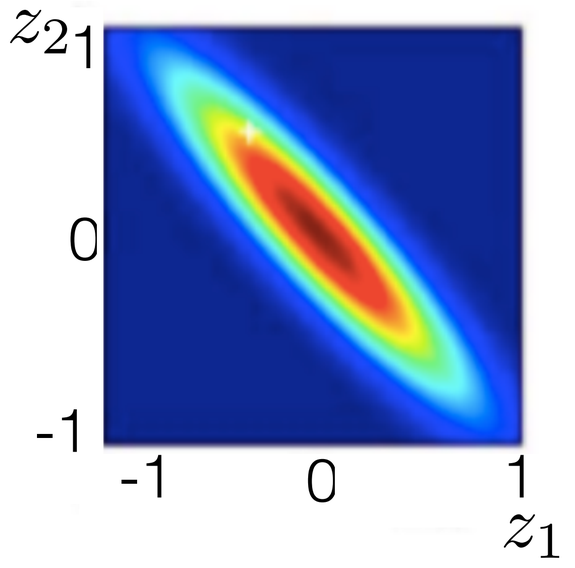
$$\mathcal{N}(z|\mu,\Sigma) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp(-\tfrac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)),$$

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





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:

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

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:

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

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

- prior:

$p(w) = \mathcal{N}(w|\mu_0, \Sigma_0)$ — conjugate

μ_0, Σ_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:

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

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

- prior:

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

Training methods: summary

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

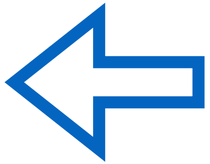
Approximation		Inference
Exact	$p(w\|X, Y)$	Full Bayesian inference
Parametric	$p(w\|X, Y) \approx q(w\|\lambda)$	Parametric Var. Inference
Delta function	$p(w\|X, Y) \approx \delta(w_{MP})$	Max. posterior inference
No prior	w_{ML}	Max. likelihood inference

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

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

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

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

Bayesian linear regression: training (ML)

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

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

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

$$\Updownarrow$$

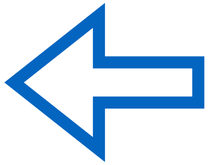
$$\|Y - Xw\|^2 \to \min_w$$ — **we already know the solution!**

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

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

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

$$\downarrow$$

$$\max_{w}$$

Bayesian linear regression: training (MP)

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

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

$$\downarrow$$

$$\max_w$$

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

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

Bayesian linear regression: training (MP)

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

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

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

$$\Updownarrow$$

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

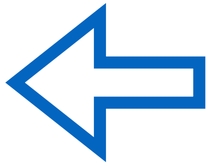
— we already know the solution!

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

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: training

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

Likelihood and prior are conjugate → 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: $p(w) = \mathcal{N}(w|0, \alpha I),\ \alpha > 0$

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

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

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

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

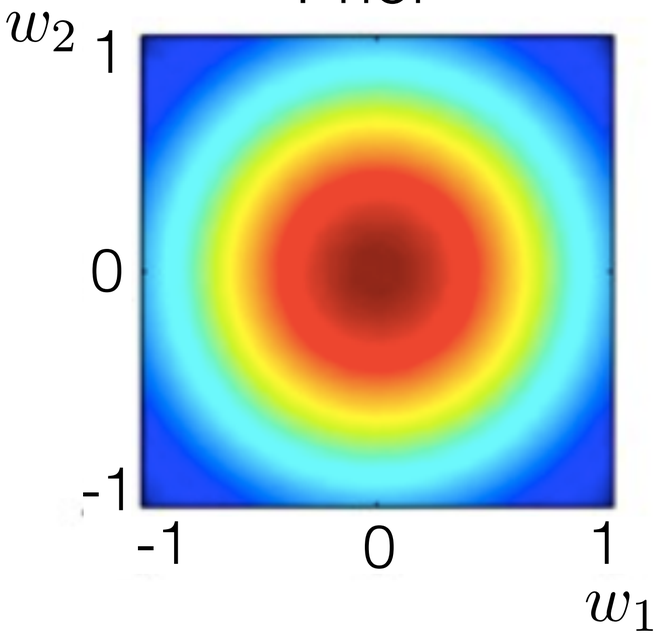
$$p(w|X,Y) = \mathcal{N}(w|w_{MP}, \Sigma)$$

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

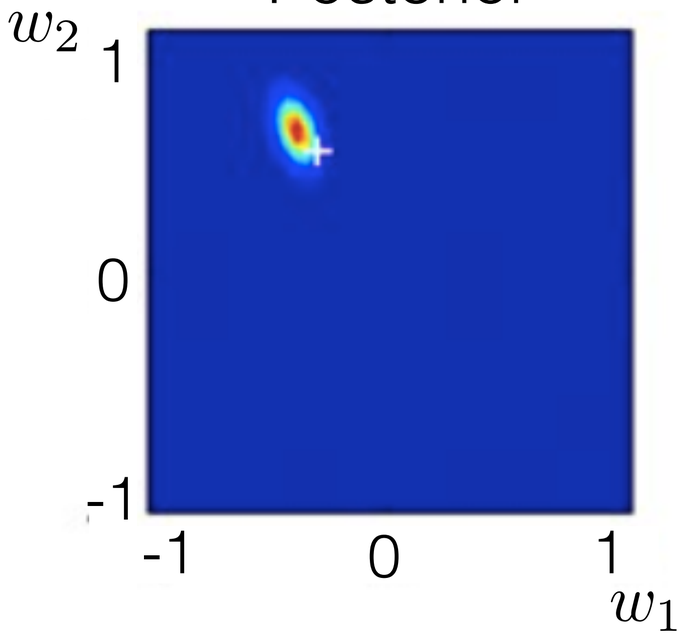
$$\Sigma = X^T X + \tfrac{1}{\alpha} I$$

Training visualization

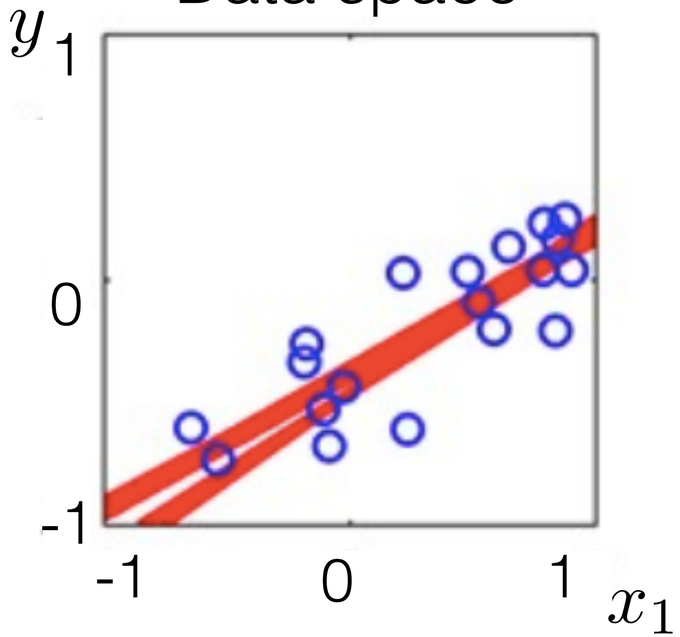
Prior



Posterior



Data space



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

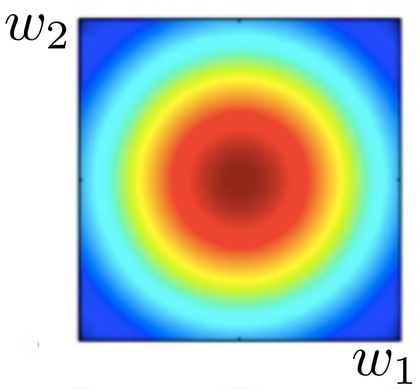
$$p(w|X, Y) = \mathcal{N}(w|w_{MP}, \Sigma)$$

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

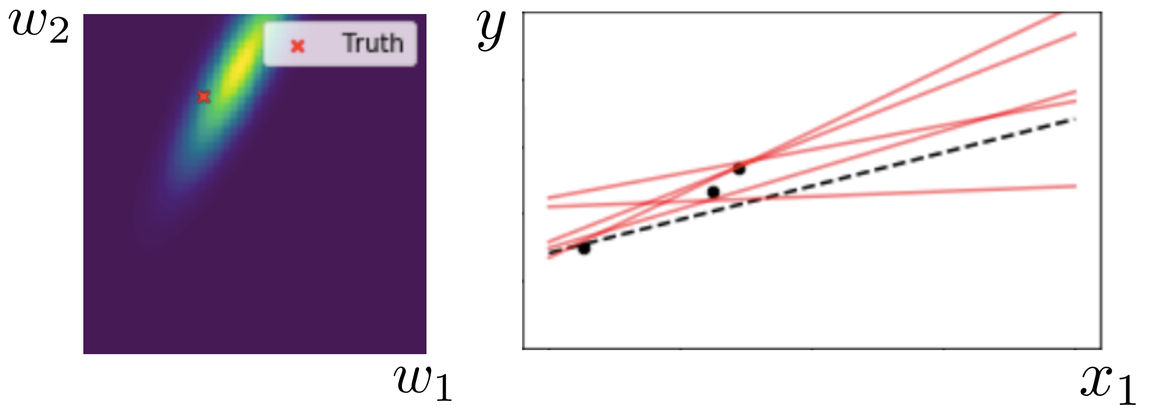
Image from C. Bishop. Pattern Recognition and Machine Learning

Training: increasing amount of data

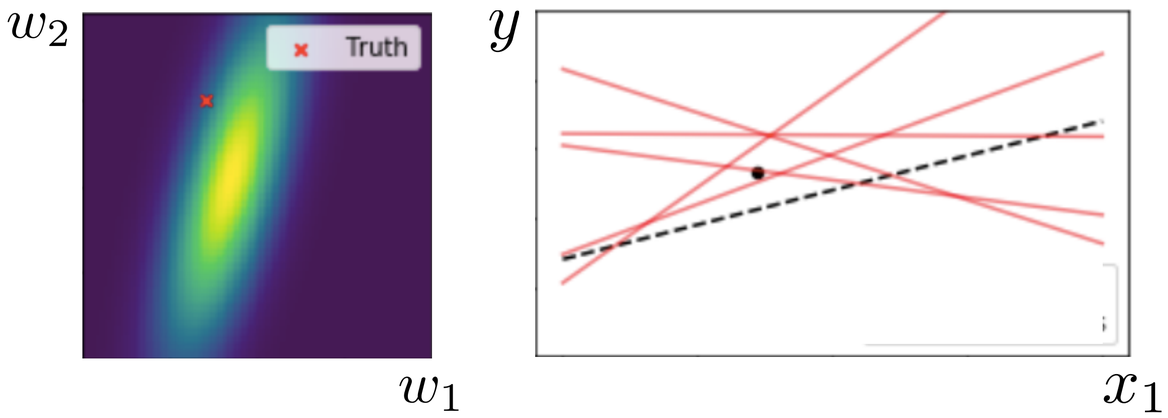
Prior:



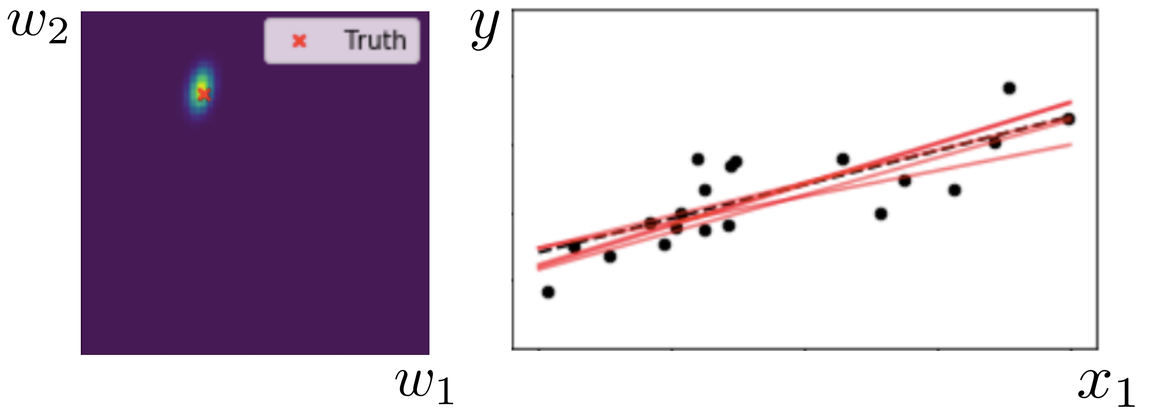
3 data points (N=3):



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:

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

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

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

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

Bayesian linear regression: prediction

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

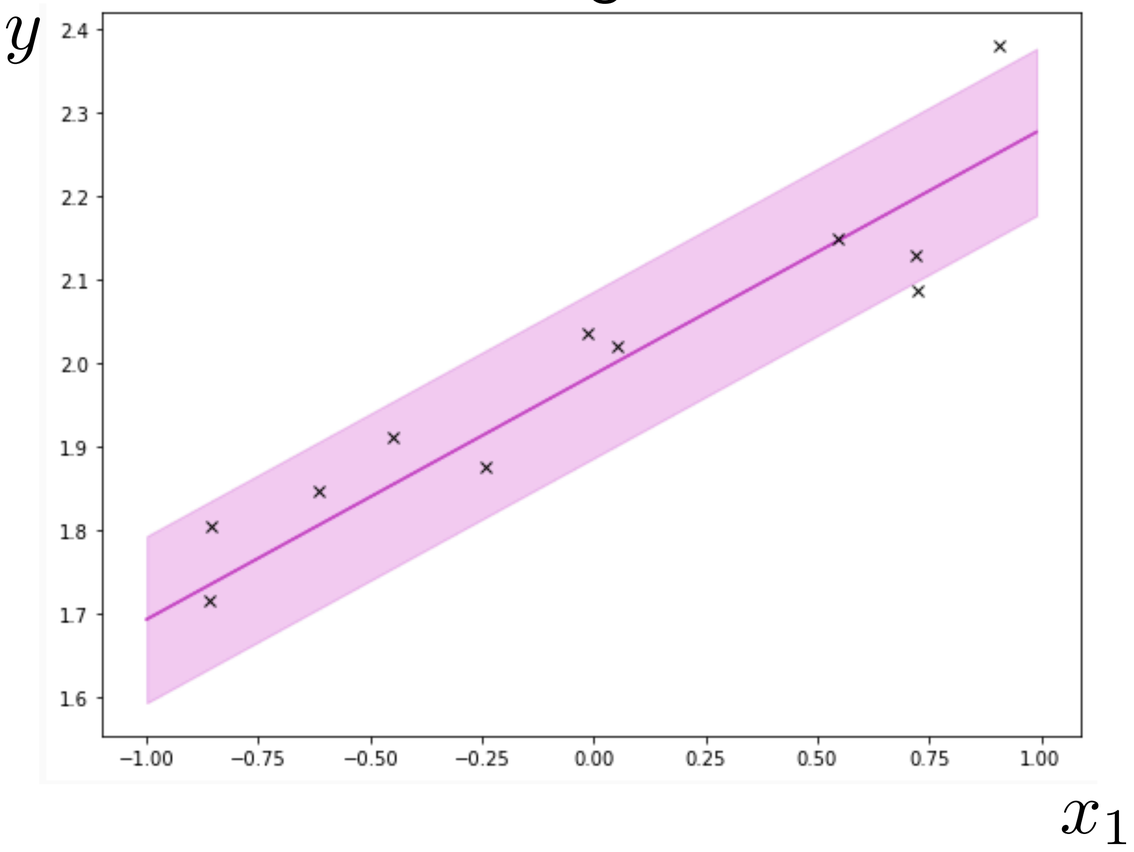
$$\int \mathcal{N}(y_*|x_*^T w, 1)\mathcal{N}(w|w_{MP}, \Sigma)dw =$$

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

x_* — new object

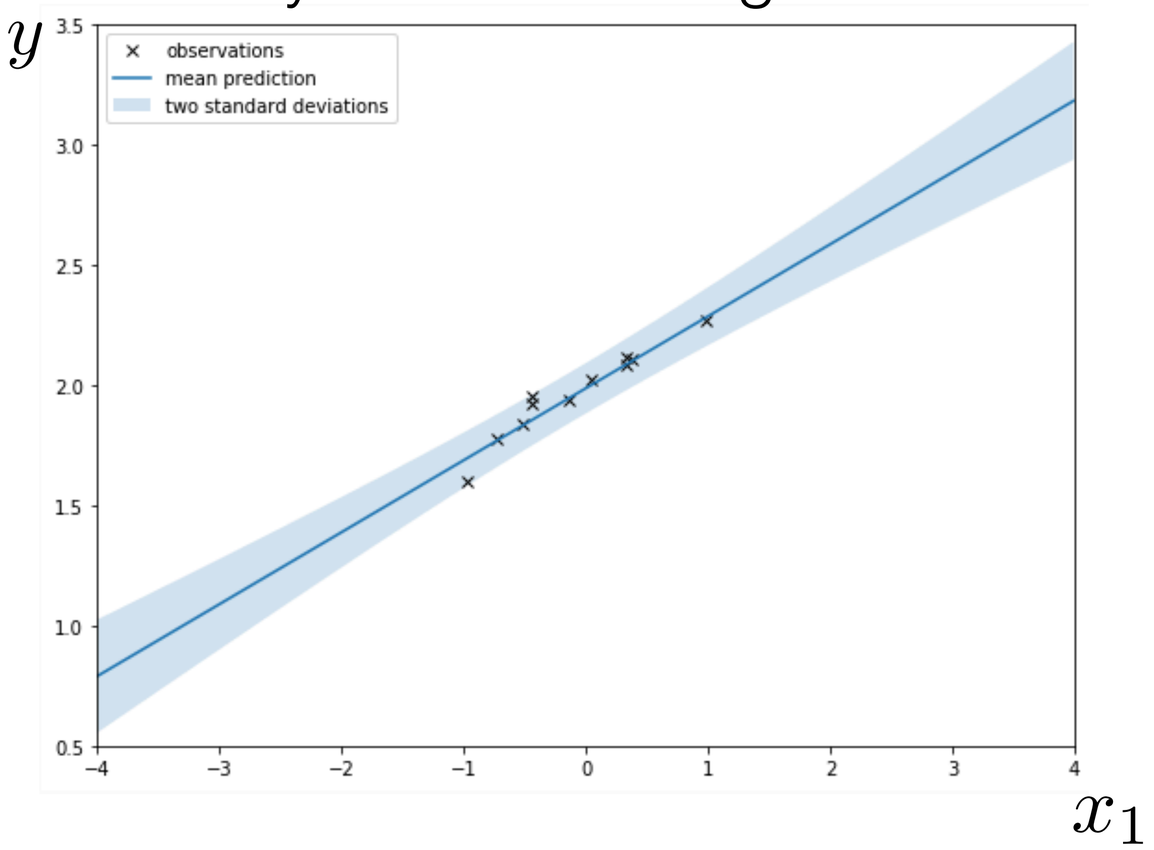
Prediction visualization

Linear regression



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

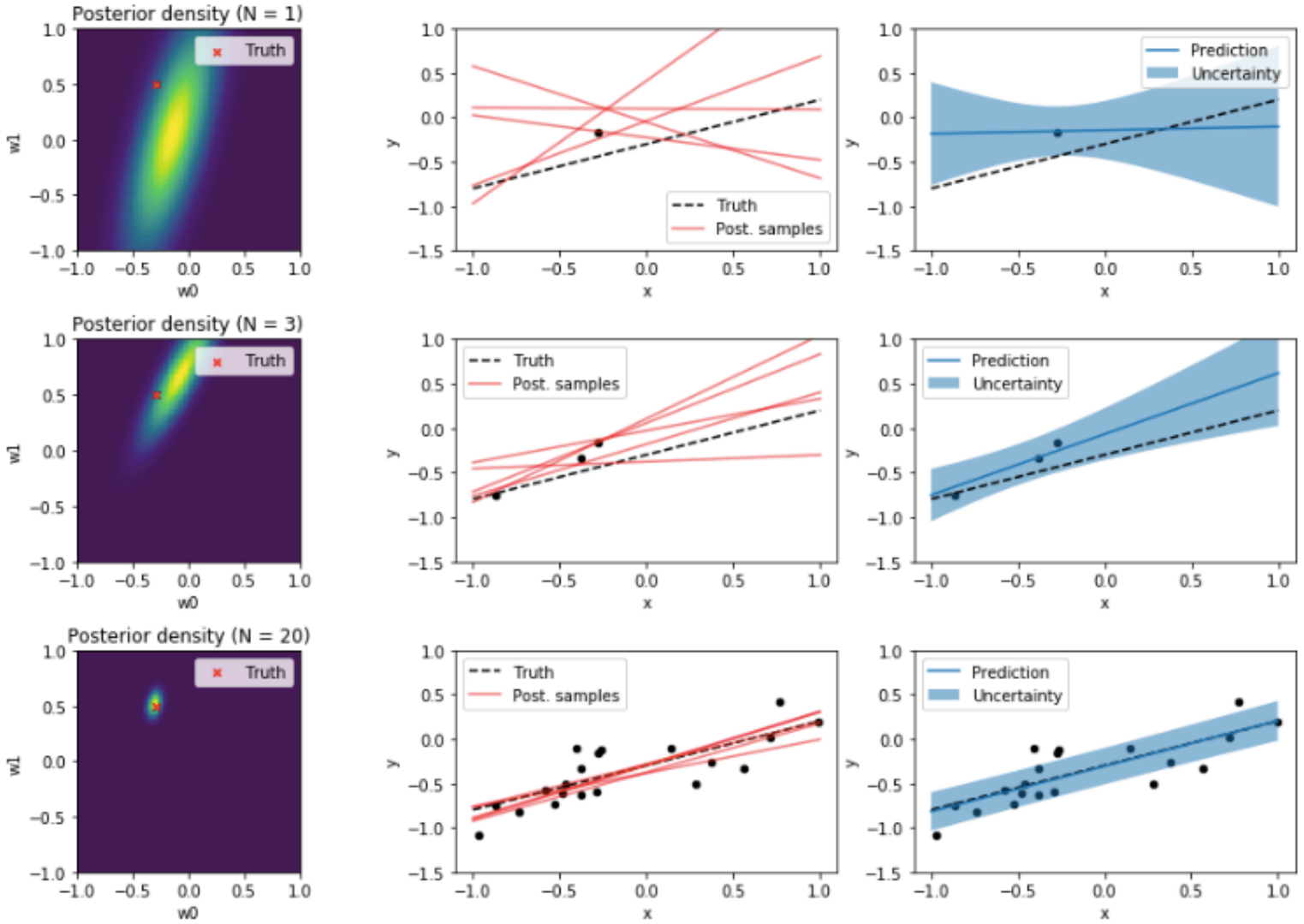
Bayesian linear regression



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

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

Prediction: increasing amount of data



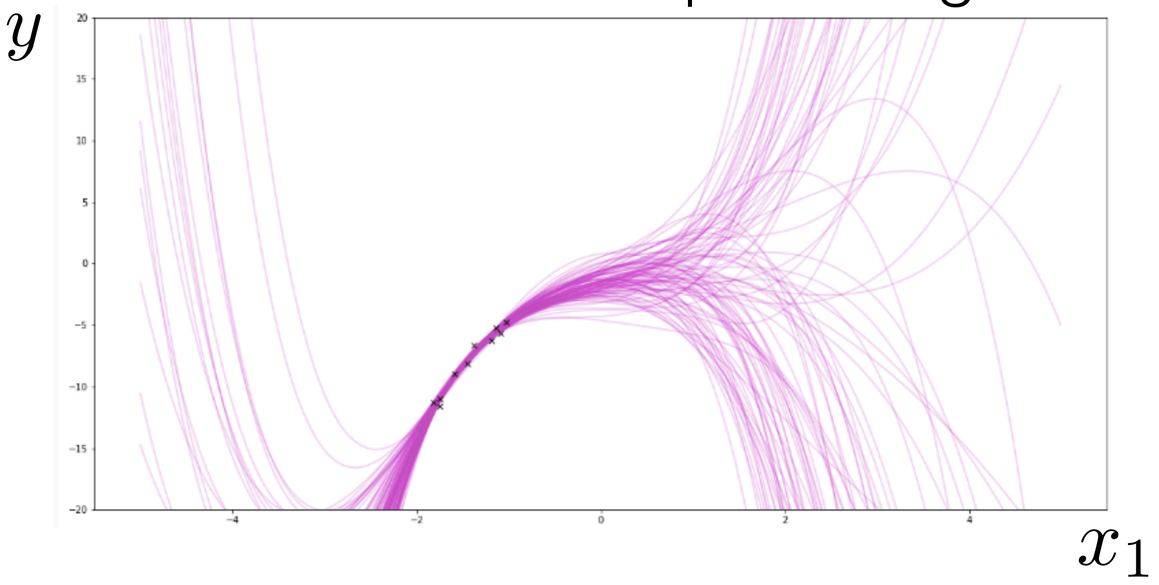
Images from http://krasserm.github.io/2019/02/23/bayesian-linear-regression/

Prediction: polynomial features

Modify training data: add polynomial features

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

Prediction with sampled weights



Mean ± Std of predictions

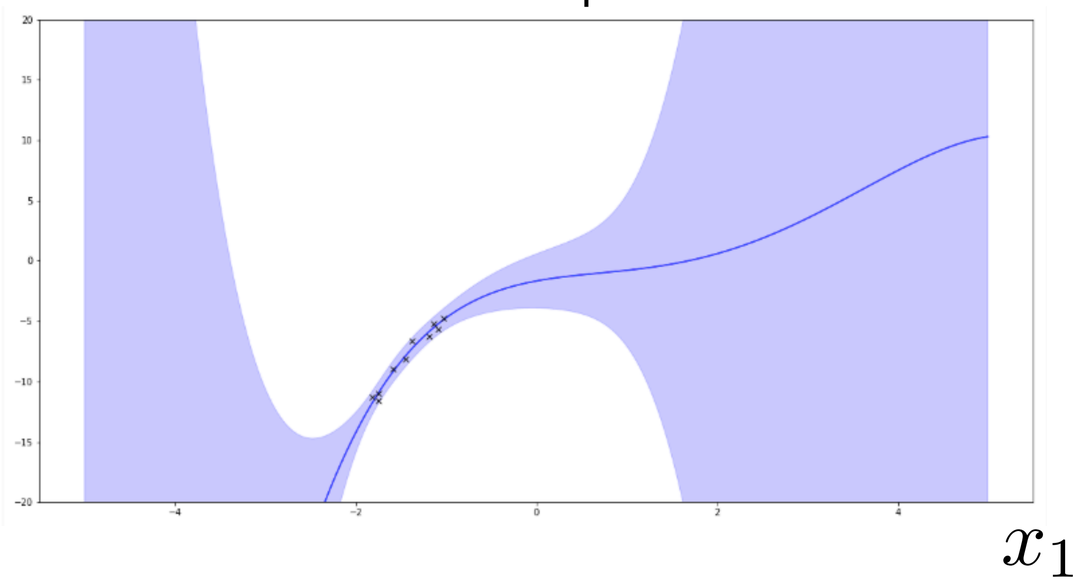


Image from https://jessicastringham.net/2018/01/10/bayesian-linreg-plots/

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:

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

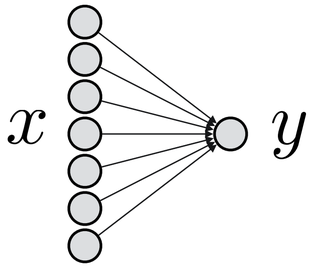
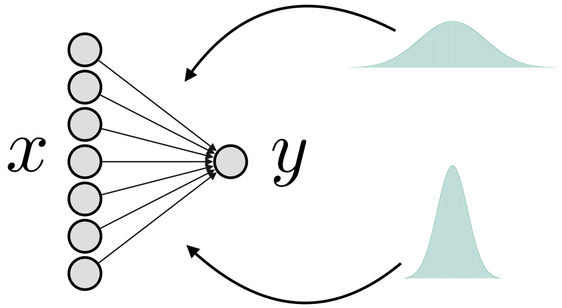
$$w_{MP} = (X^T X + \frac{1}{\alpha} I)^{-1} X^T Y$$

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

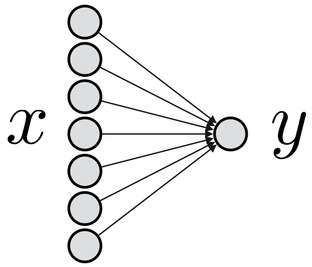
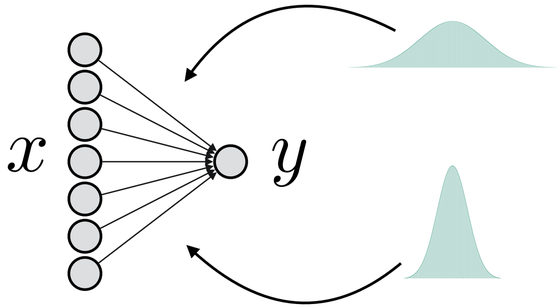
Prediction on a new object x_* :

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

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 $p(w\|X, Y)$ 
Prediction on a new object x_*:	$p(y_*\|x_*, w_{MP})$	$\mathbb{E}_{p(w\|X,Y)} p(y_*\|x_*, w)$

Putting everything together

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

Summary

- Usual training of linear regression is equivalent to ML / MP Bayesian inference
- We can perform full Bayesian inference for linear regression, and obtain weight variance and covariance, in addition to mean values
- Bayesian regression provides more informative predictive uncertainty