Learning for vision I

Karel Zimmermann
http://cmp.felk.cvut.cz/~zimmerk/

Vision for Robotics and Autonomous Systems
https://cyber.felk.cvut.cz/vras/

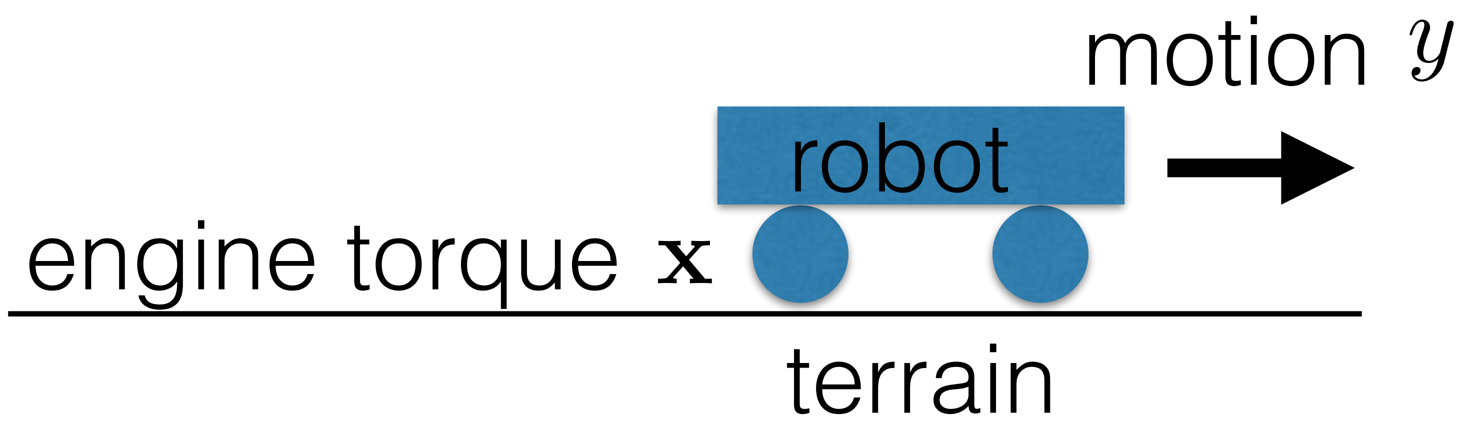
Center for Machine Perception
https://cmp.felk.cvut.cz

Department for Cybernetics
Faculty of Electrical Engineering
Czech Technical University in Prague

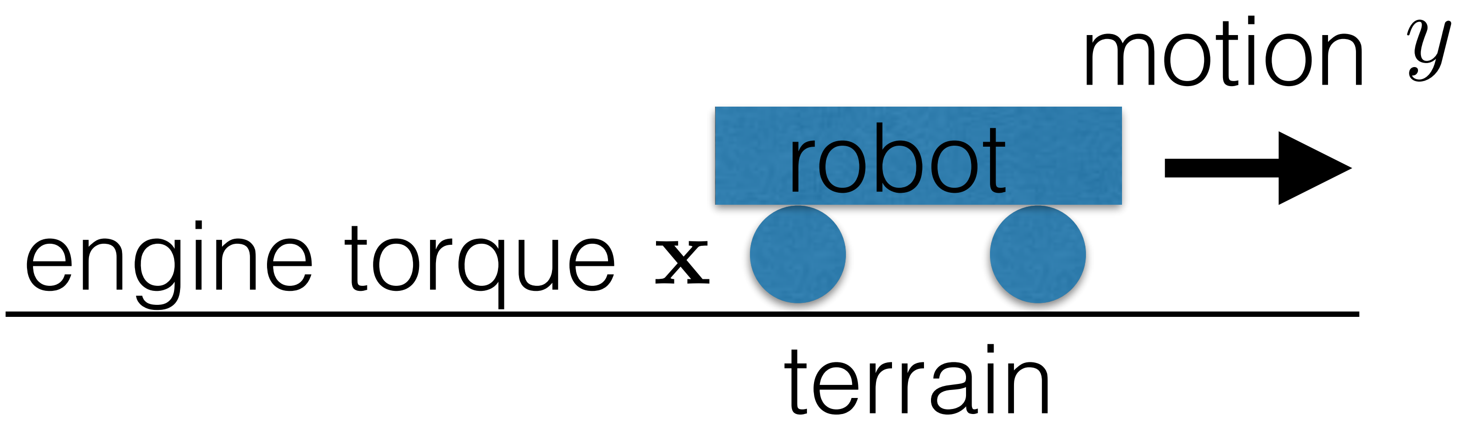
Outline

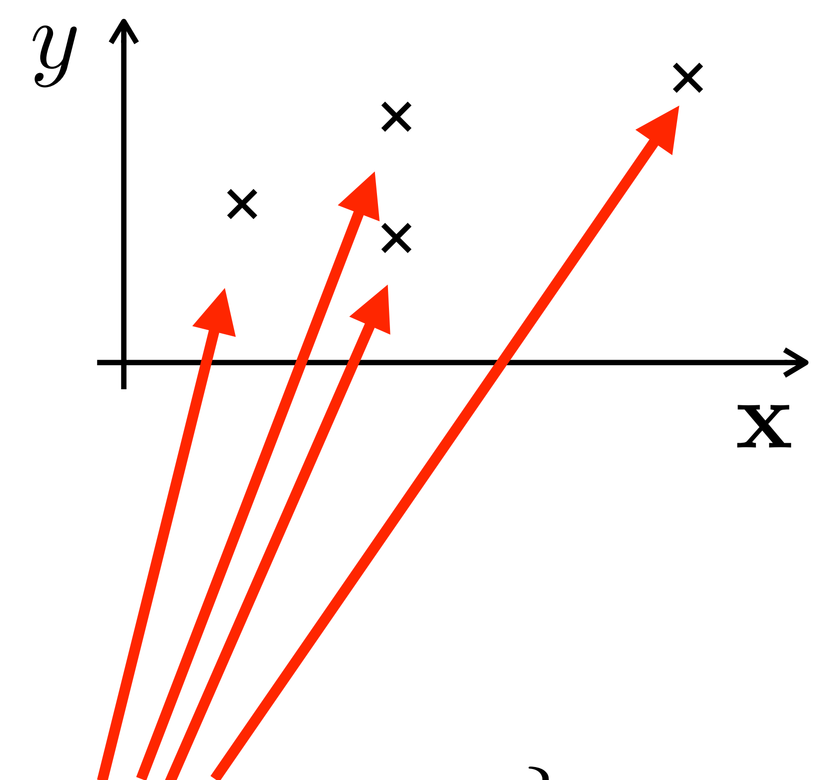
- Pre-requisites: linear algebra, Bayes rule
- MAP/ML estimation, prior and overfitting
- Linear regression
- Linear classification

- Fast summary of Maximum A-Posteriori estimation of parameters of a probability distribution
- Motivation example: estimation of a motion model



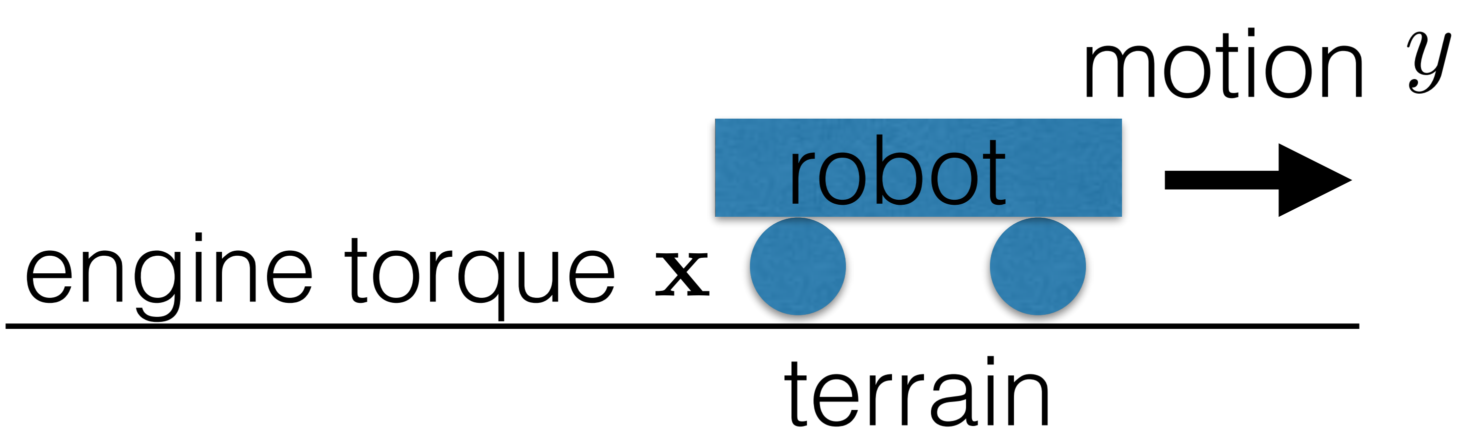
- Fast summary of Maximum A-Posteriori estimation of parameters of a probability distribution
- Motivation example: estimation of a motion model

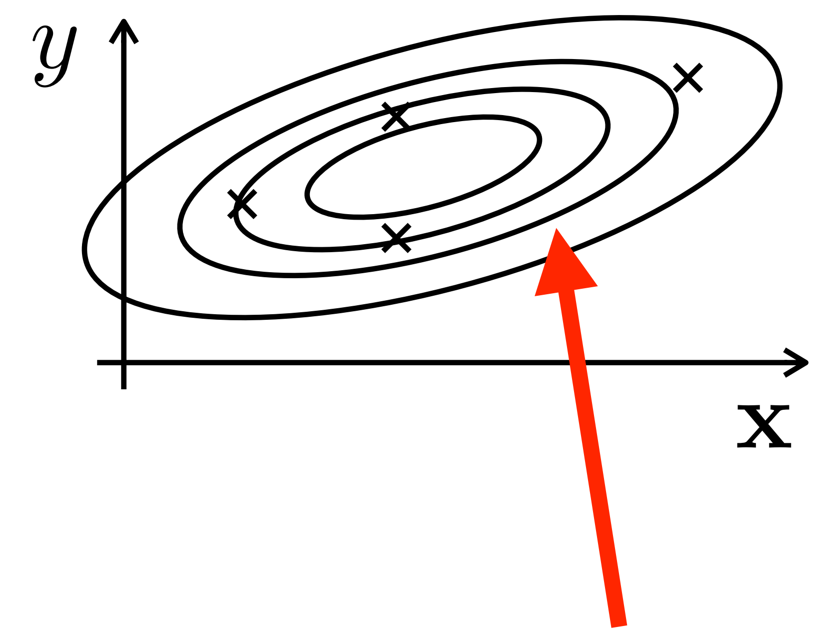




$$\mathcal{D} = \{\mathbf{x}_1, y_1 \ldots \mathbf{x}_N, y_N\}$$

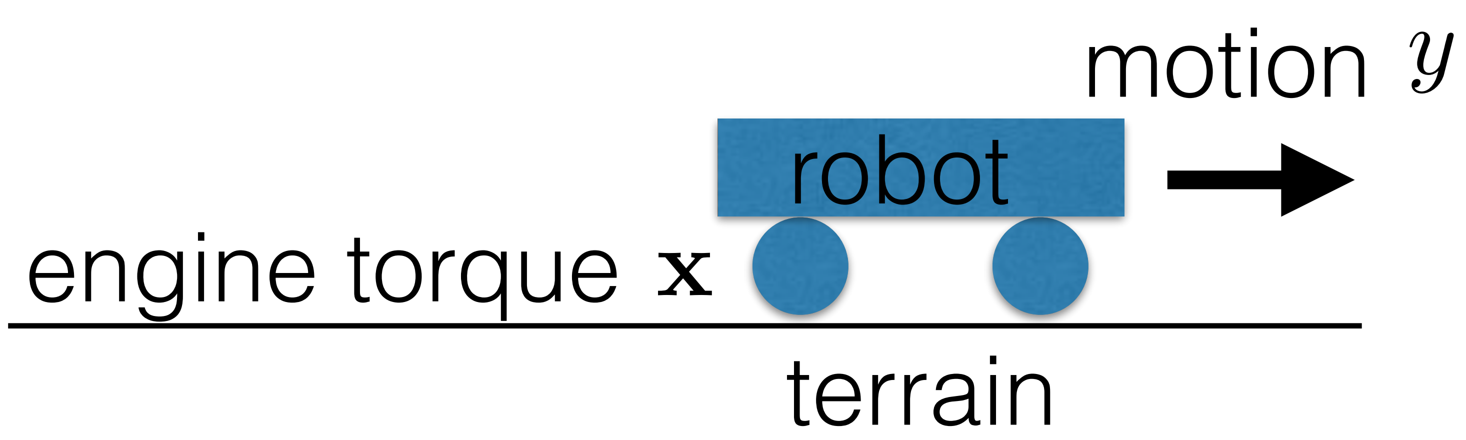
- Fast summary of Maximum A-Posteriori estimation of parameters of a probability distribution
- Motivation example: estimation of a motion model

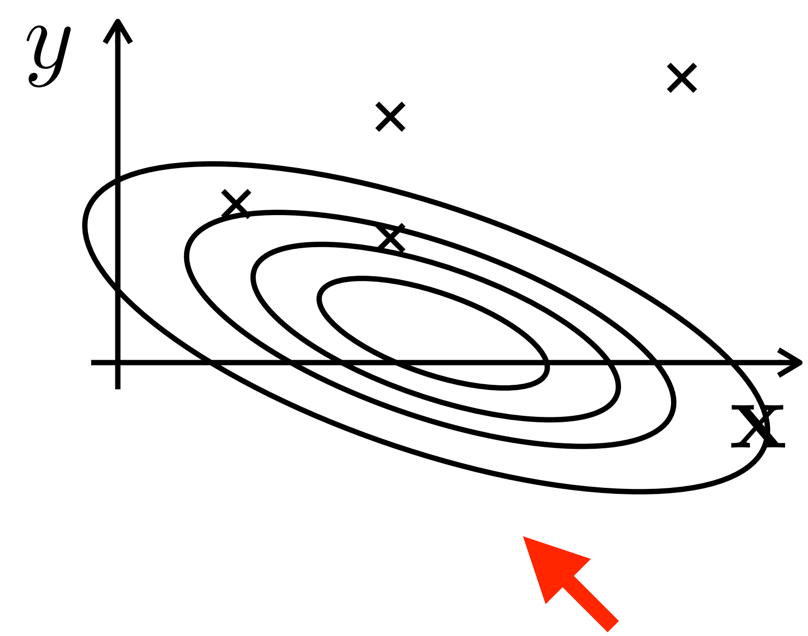




- We search for parameters $\mathbf{w}$ of motion model $p(y|\mathbf{x}, \mathbf{w})$ given i.i.d. measurements $\mathcal{D} = \{\mathbf{x}_1, y_1 \ldots \mathbf{x}_N, y_N\}$

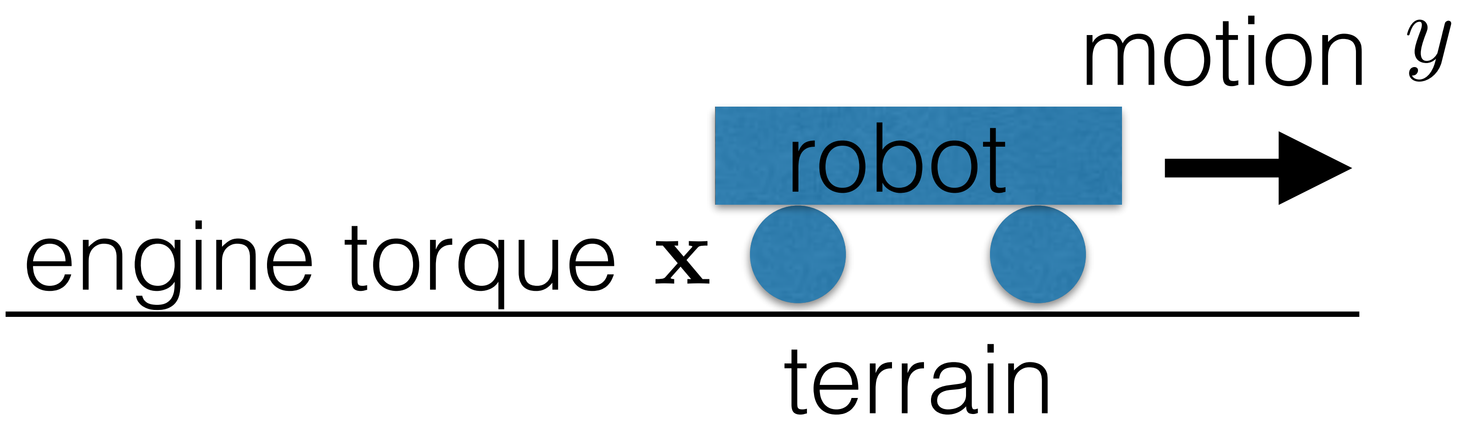
- Fast summary of Maximum A-Posteriori estimation of parameters of a probability distribution
- Motivation example: estimation of a motion model

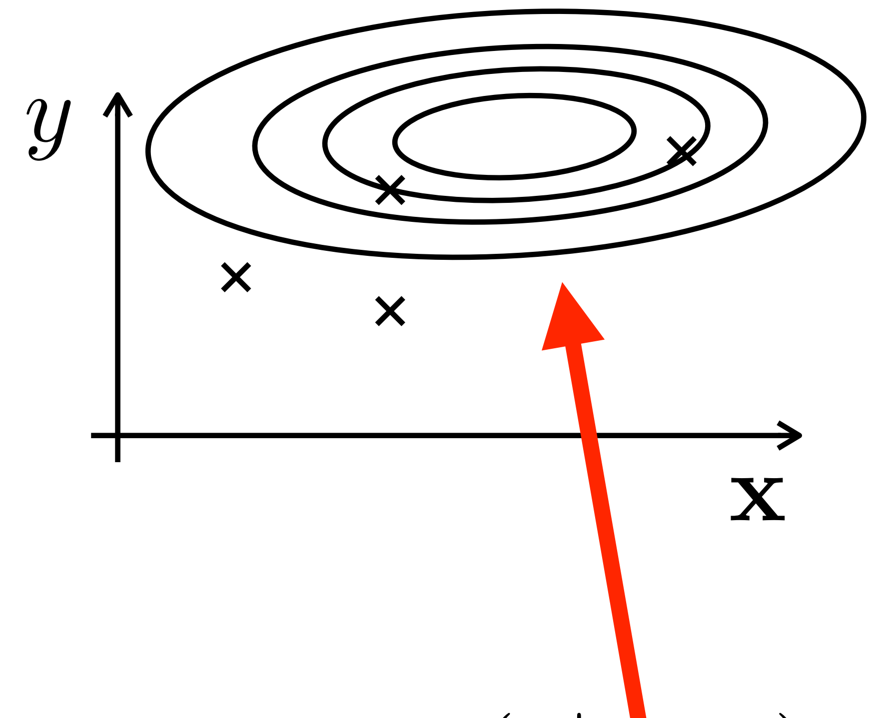




- We search for parameters $\mathbf{w}$ of motion model $p(y|\mathbf{x}, \mathbf{w})$ given i.i.d. measurements $\mathcal{D} = \{\mathbf{x}_1, y_1 \ldots \mathbf{x}_N, y_N\}$

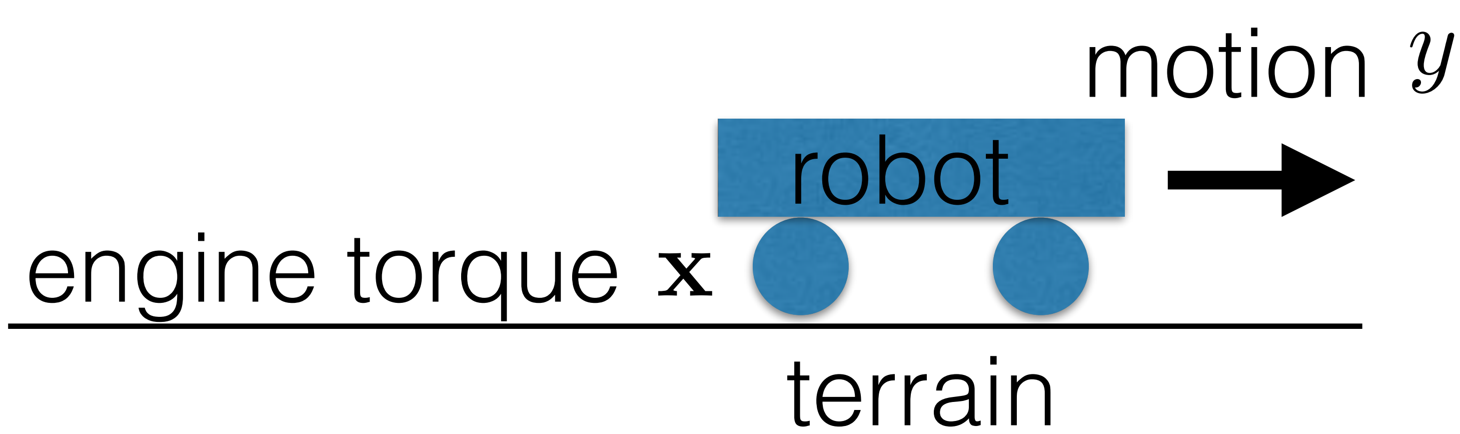
- Fast summary of Maximum A-Posteriori estimation of parameters of a probability distribution
- Motivation example: estimation of a motion model

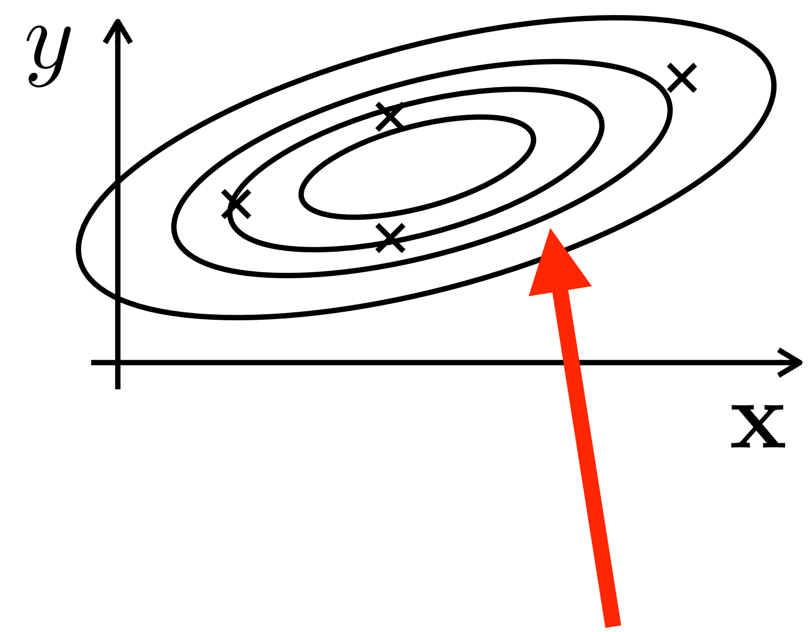




- We search for parameters $\mathbf{w}$ of motion model $p(y|\mathbf{x}, \mathbf{w})$ given i.i.d. measurements $\mathcal{D} = \{\mathbf{x}_1, y_1 \ldots \mathbf{x}_N, y_N\}$

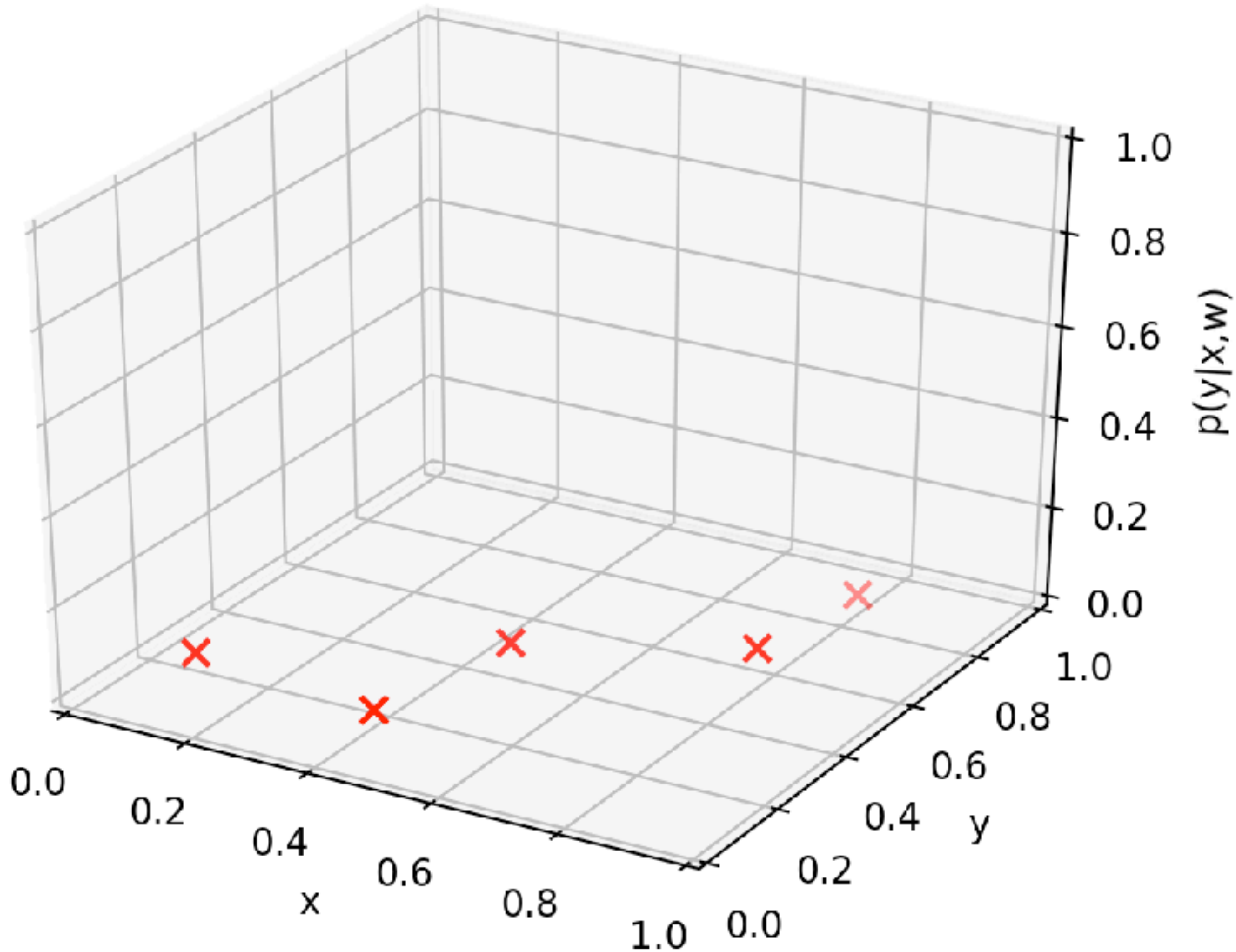
- Fast summary of Maximum A-Posteriori estimation of parameters of a probability distribution
- Motivation example: estimation of a motion model



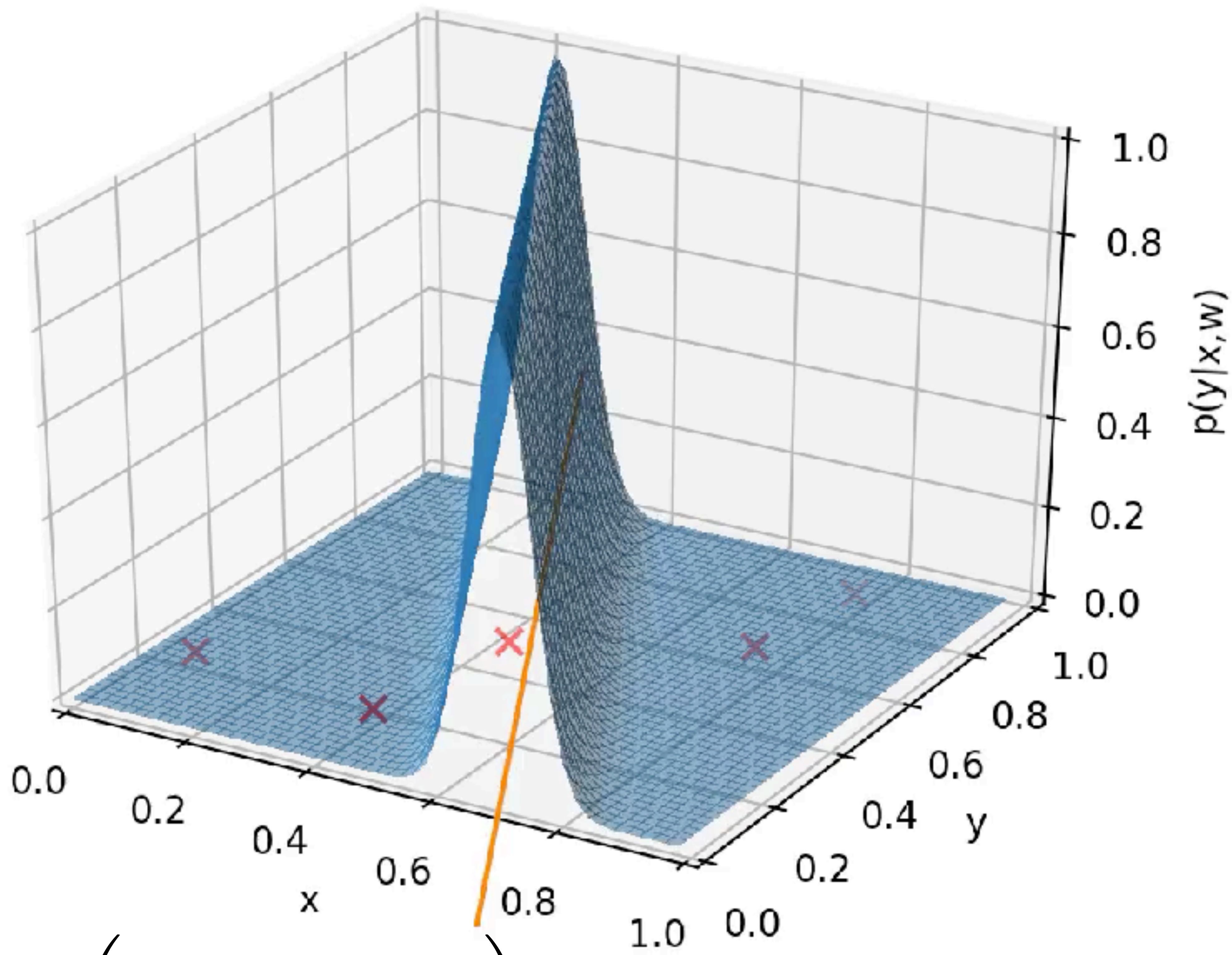


- We search for parameters $\mathbf{w}$ of motion model $p(y|\mathbf{x}, \mathbf{w})$ given i.i.d. measurements $\mathcal{D} = \{\mathbf{x}_1, y_1 \dots \mathbf{x}_N, y_N\}$

$$p(y|\mathbf{x}, \mathbf{w}) \sim \mathcal{N}_y(w_1 x + w_0,\ \sigma^2)$$

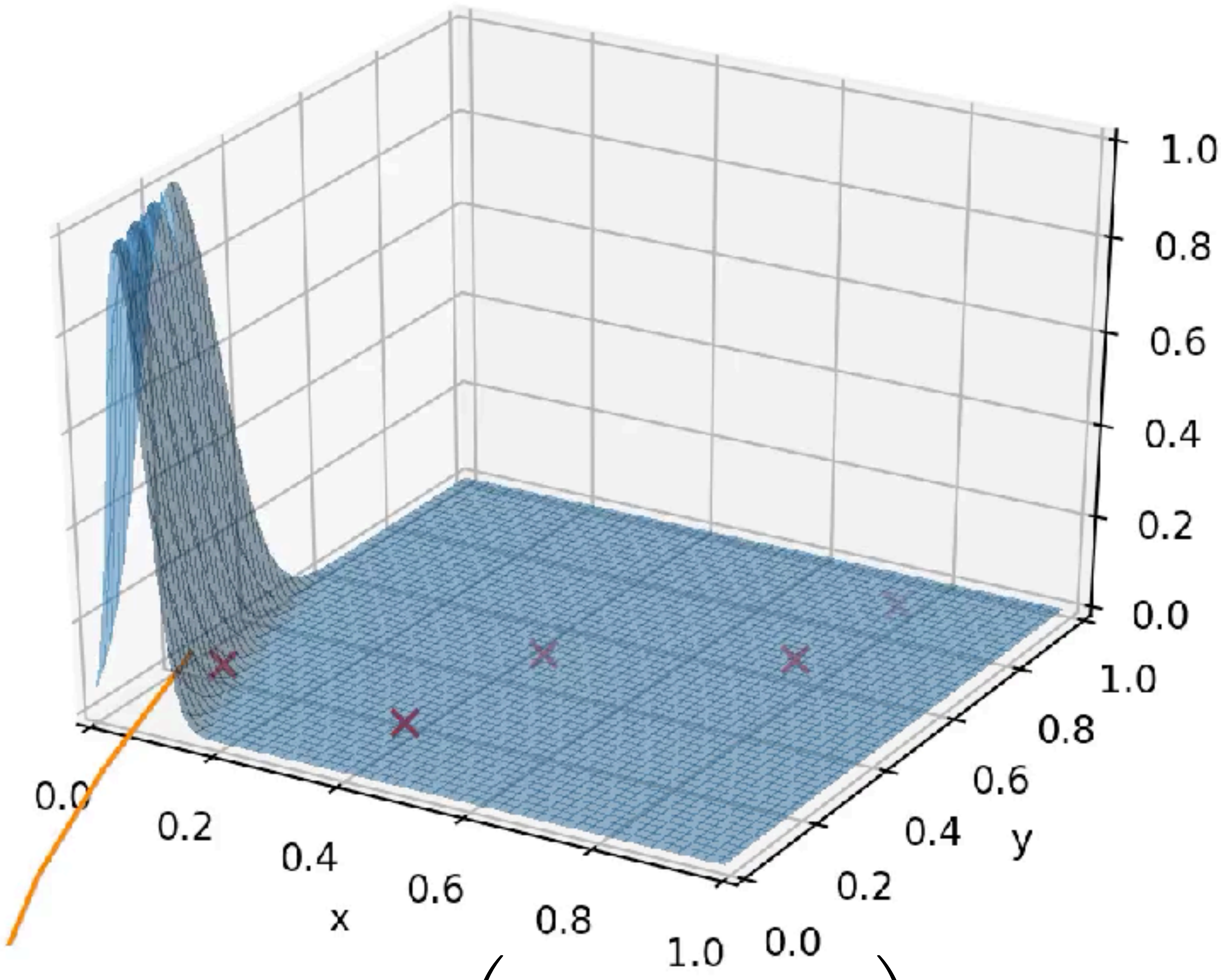


$$p(y|\mathbf{x}, \mathbf{w}) \sim \mathcal{N}_y(w_1 x + w_0,\ \sigma^2)$$



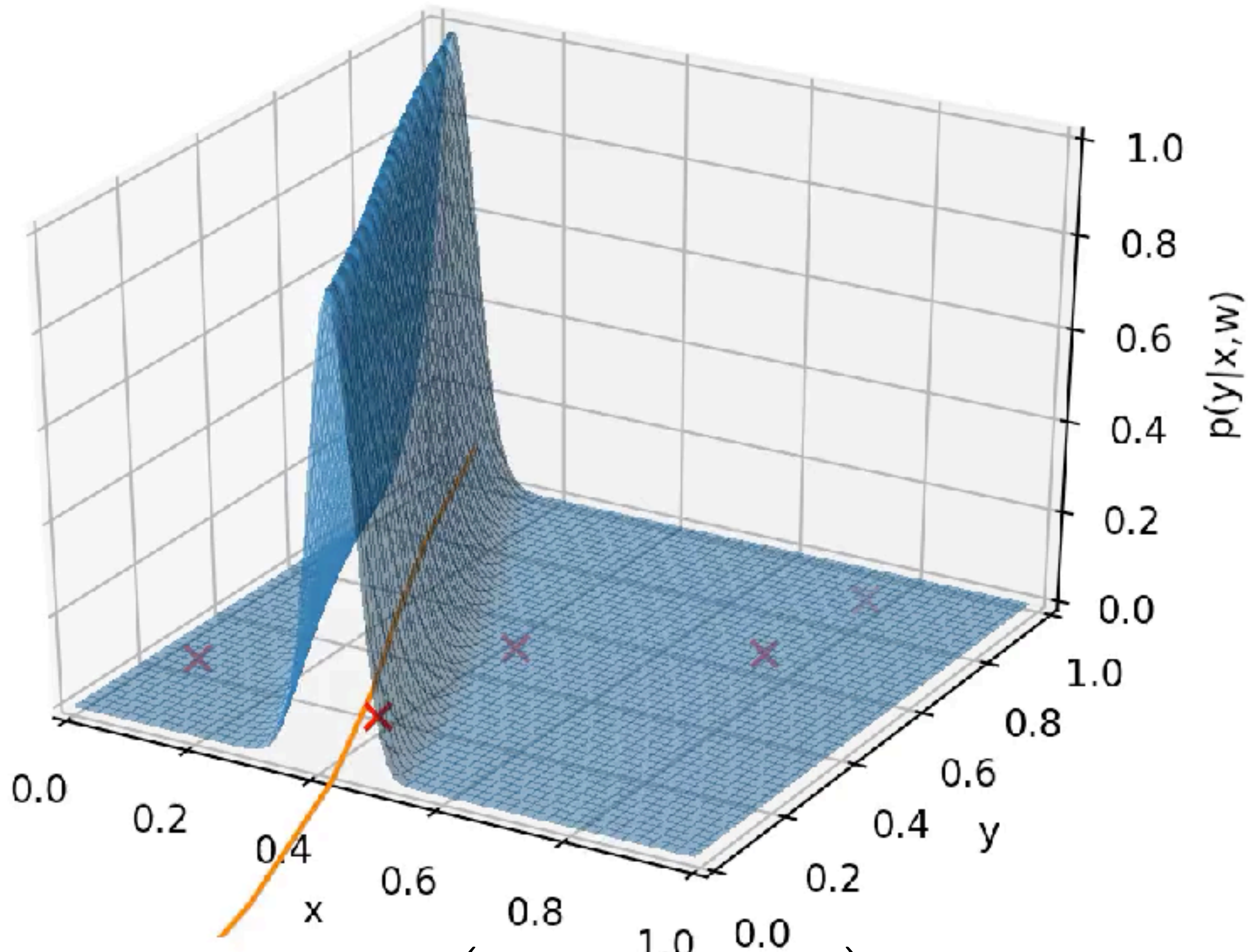
$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \left(\prod_i p(y_i|\mathbf{x}_i, \mathbf{w}) \right) = \arg\min_{\mathbf{w}} \sum_i (w_1 x_i + w_0 - y_i)^2$$

$$p(y|\mathbf{x},\mathbf{w}) \sim \mathcal{N}_y(w_2x^2 + w_1x + w_0,\ \sigma^2)$$



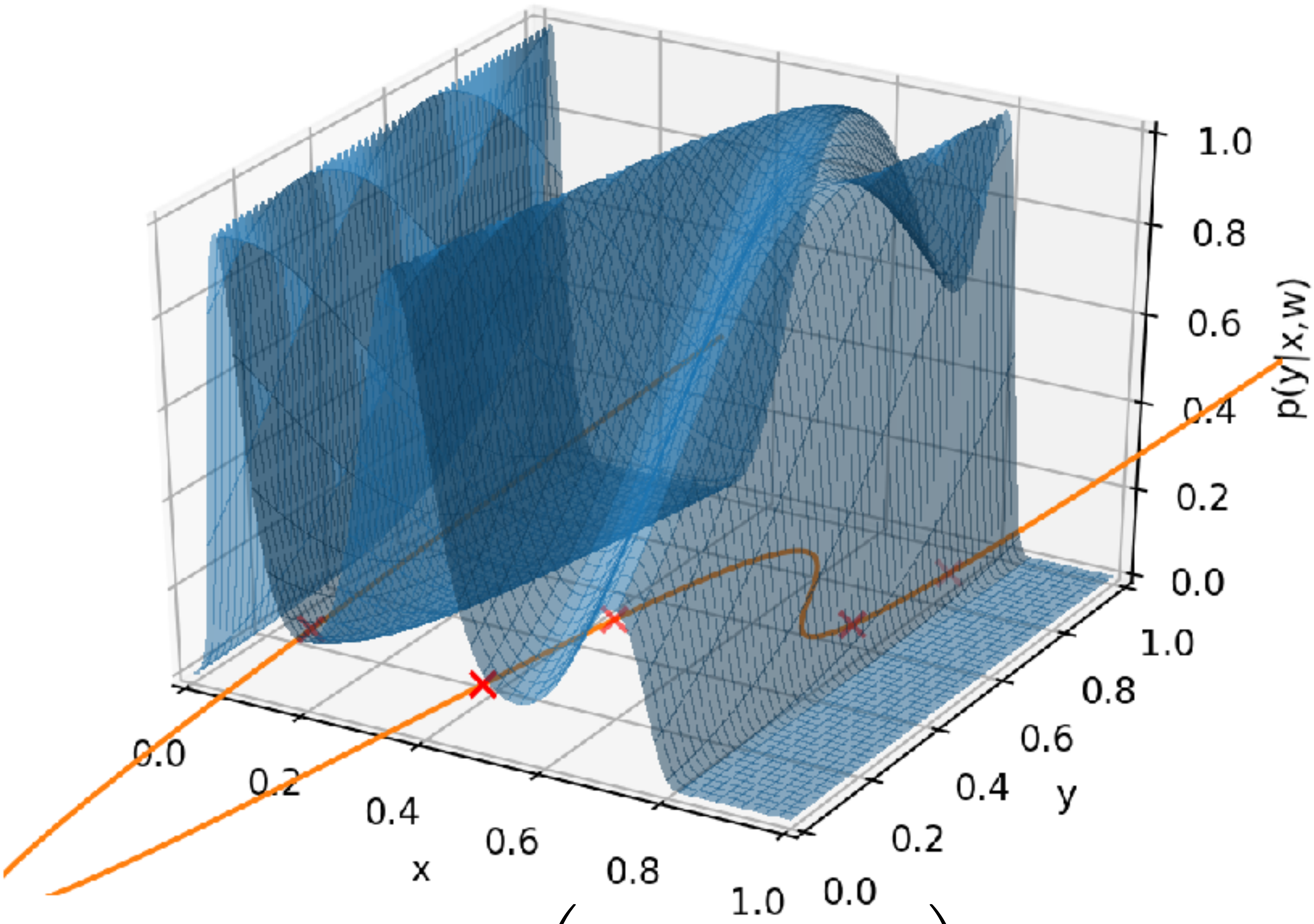
$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \left(\prod_i p(y_i|\mathbf{x}_i,\mathbf{w}) \right)$$

$$p(y|\mathbf{x},\mathbf{w}) \sim \mathcal{N}_y(w_4x^4 + w_3x^3 + w_2x^2 + w_1x + w_0,\ \sigma^2)$$



$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \left(\prod_i p(y_i|\mathbf{x}_i,\mathbf{w}) \right)$$

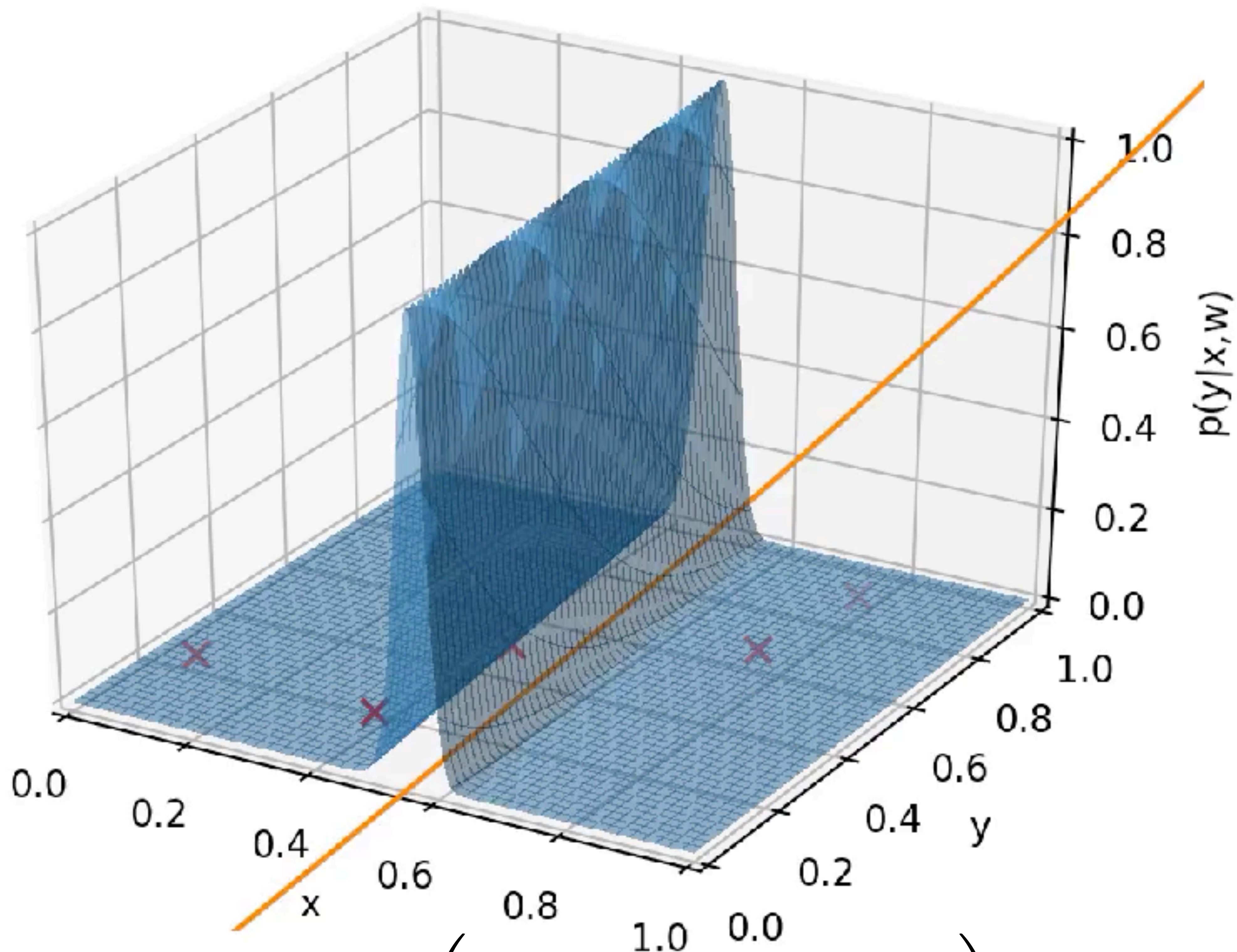
$$p(y|\mathbf{x}, \mathbf{w}) \sim \mathcal{N}_y(f(\mathbf{x}, \mathbf{w}), \sigma^2)$$



$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \left(\prod_i p(y_i|\mathbf{x}_i, \mathbf{w}) \right)$$

$$p(y|\mathbf{x}, \mathbf{w}) \sim \mathcal{N}_y(f(\mathbf{x}, \mathbf{w}), \sigma^2)$$

$$p(\mathbf{w}) \sim \mathcal{N}_w(\mathbf{0},\ \sigma^2)$$



$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \left(\prod_i p(y_i|\mathbf{x}_i, \mathbf{w}) p(\mathbf{w}) \right)$$

- We search for parameters $\mathbf{w}$ of motion model $p(y|\mathbf{x}, \mathbf{w})$ given i.i.d. measurements $\mathcal{D} = \{\mathbf{x}_1, y_1 \dots \mathbf{x}_N, y_N\}$

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathcal{D}) = \arg\max_{\mathbf{w}} \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

- We search for parameters $\mathbf{w}$ of motion model $p(y|\mathbf{x}, \mathbf{w})$ given i.i.d. measurements $\mathcal{D} = \{\mathbf{x}_1, y_1 \dots \mathbf{x}_N, y_N\}$

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathcal{D}) = \arg\max_{\mathbf{w}} \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

$$= \arg\max_{\mathbf{w}} p(\mathcal{D}|\mathbf{w})p(\mathbf{w}) = \arg\max_{\mathbf{w}} p(\mathbf{x}_1, y_1 \dots \mathbf{x}_N, y_N|\mathbf{w})p(\mathbf{w})$$

- We search for parameters $\mathbf{w}$ of motion model $p(y|\mathbf{x}, \mathbf{w})$ given i.i.d. measurements $\mathcal{D} = \{\mathbf{x}_1, y_1 \ldots \mathbf{x}_N, y_N\}$

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathcal{D}) = \arg\max_{\mathbf{w}} \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

$$= \arg\max_{\mathbf{w}} p(\mathcal{D}|\mathbf{w})p(\mathbf{w}) = \arg\max_{\mathbf{w}} p(\mathbf{x}_1, y_1 \ldots \mathbf{x}_N, y_N|\mathbf{w})p(\mathbf{w})$$

i.i.d.

$$= \arg\max_{\mathbf{w}} \left(\prod_i p(\mathbf{x}_i, y_i|\mathbf{w}) \right) p(\mathbf{w})$$

- We search for parameters $\mathbf{w}$ of motion model $p(y|\mathbf{x}, \mathbf{w})$ given i.i.d. measurements $\mathcal{D} = \{\mathbf{x}_1, y_1 \dots \mathbf{x}_N, y_N\}$

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathcal{D}) = \arg\max_{\mathbf{w}} \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

$$= \arg\max_{\mathbf{w}} p(\mathcal{D}|\mathbf{w})p(\mathbf{w}) = \arg\max_{\mathbf{w}} p(\mathbf{x}_1, y_1 \dots \mathbf{x}_N, y_N|\mathbf{w})p(\mathbf{w})$$

$$= \arg\max_{\mathbf{w}} \left(\prod_i p(\mathbf{x}_i, y_i|\mathbf{w})\right) p(\mathbf{w})$$

$$= \arg\max_{\mathbf{w}} \left(\prod_i p(y_i|\mathbf{x}_i, \mathbf{w})p(\mathbf{x}_i)\right) p(\mathbf{w})$$

- We search for parameters $\mathbf{w}$ of motion model $p(y|\mathbf{x},\mathbf{w})$ given i.i.d. measurements $\mathcal{D} = \{\mathbf{x}_1, y_1 \dots \mathbf{x}_N, y_N\}$

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathcal{D}) = \arg\max_{\mathbf{w}} \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

$$= \arg\max_{\mathbf{w}} p(\mathcal{D}|\mathbf{w})p(\mathbf{w}) = \arg\max_{\mathbf{w}} p(\mathbf{x}_1, y_1 \dots \mathbf{x}_N, y_N|\mathbf{w})p(\mathbf{w})$$

$$= \arg\max_{\mathbf{w}} \left(\prod_i p(\mathbf{x}_i, y_i|\mathbf{w})\right) p(\mathbf{w})$$

$$= \arg\max_{\mathbf{w}} \left(\prod_i p(y_i|\mathbf{x}_i, \mathbf{w})p(\mathbf{x}_i)\right) p(\mathbf{w})$$

$$= \arg\max_{\mathbf{w}} \left(\sum_i \log(p(y_i|\mathbf{x}_i, \mathbf{w})) + \log p(\mathbf{x}_i)\right) + \log p(\mathbf{w})$$

- We search for parameters $\mathbf{w}$ of motion model $p(y|\mathbf{x}, \mathbf{w})$ given i.i.d. measurements $\mathcal{D} = \{\mathbf{x}_1, y_1 \ldots \mathbf{x}_N, y_N\}$

$$= \arg\max_{\mathbf{w}} \left(\sum_i \log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) + \log p(\mathbf{w})$$

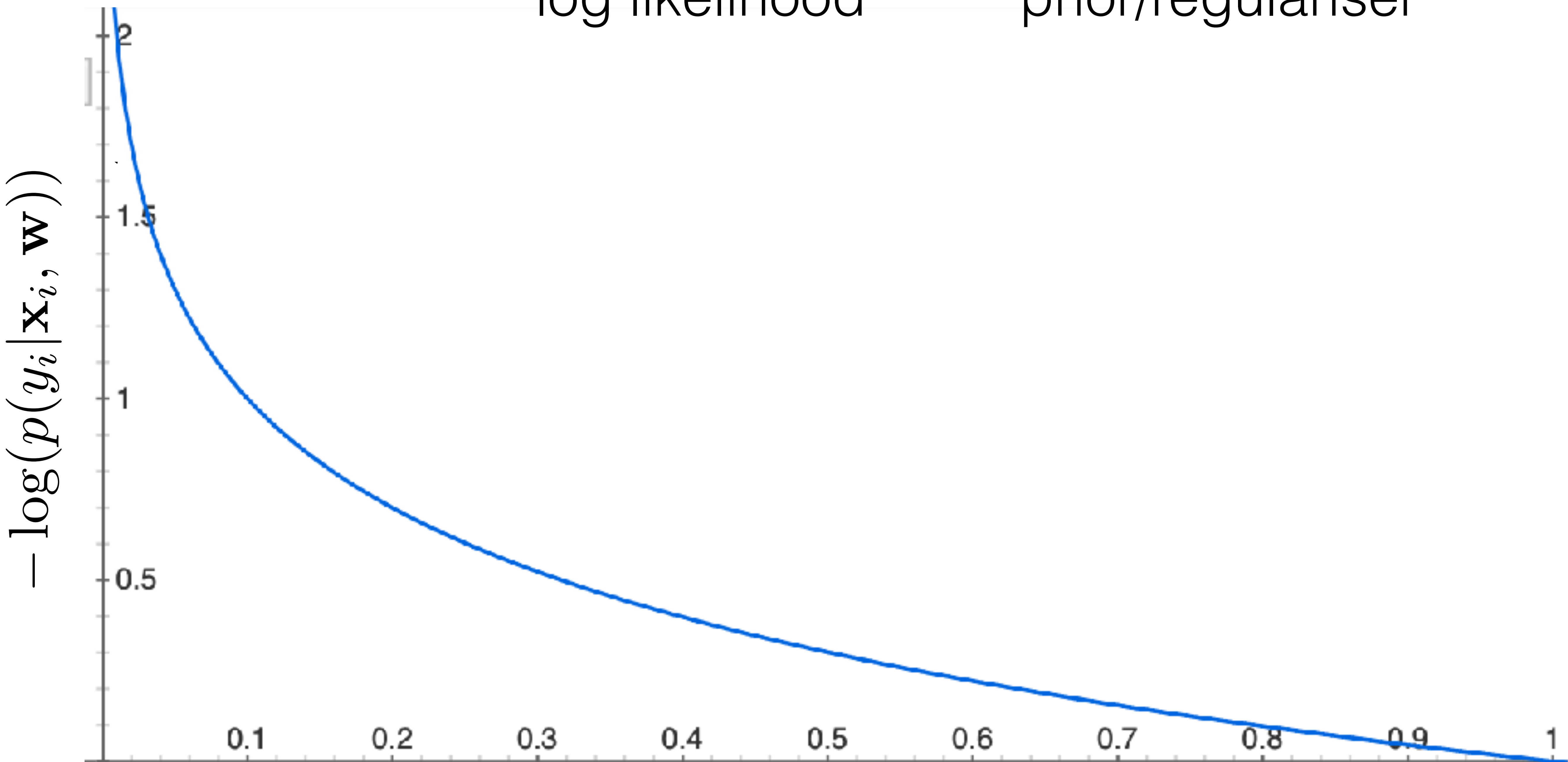
$$= \arg\min_{\mathbf{w}} \left(\sum_i - \log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) + (-\log p(\mathbf{w}))$$

log likelihood prior/regulariser

$$= \arg\max_{\mathbf{w}} \left(\prod_i p(y_i|\mathbf{x}_i, \mathbf{w}) p(\mathbf{w}) \right)$$

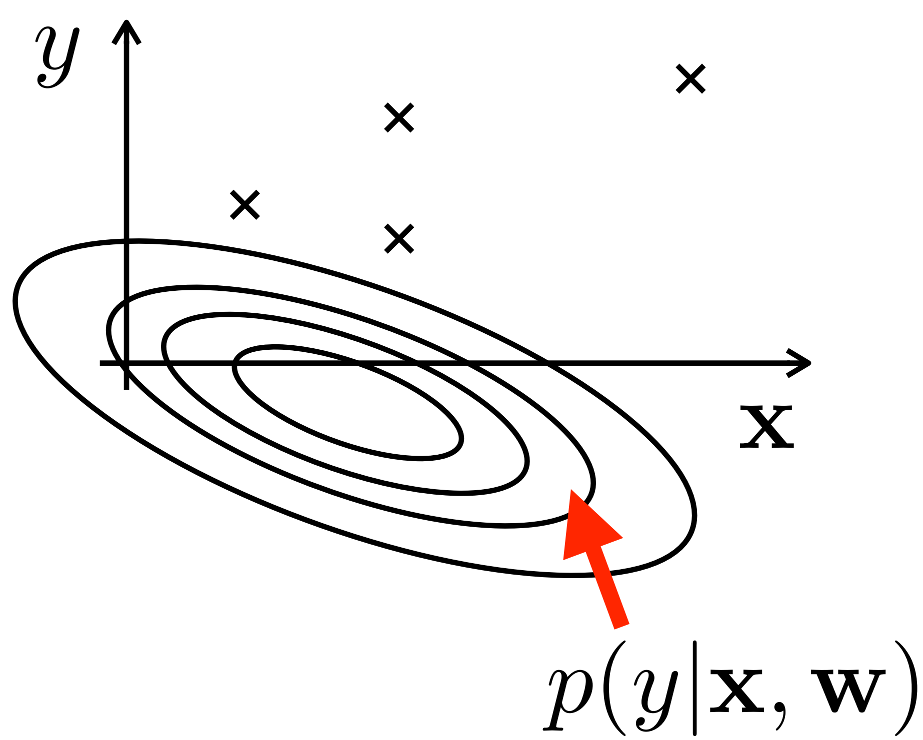
$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) + (-\log p(\mathbf{w}))$$

log likelihood prior/regulariser



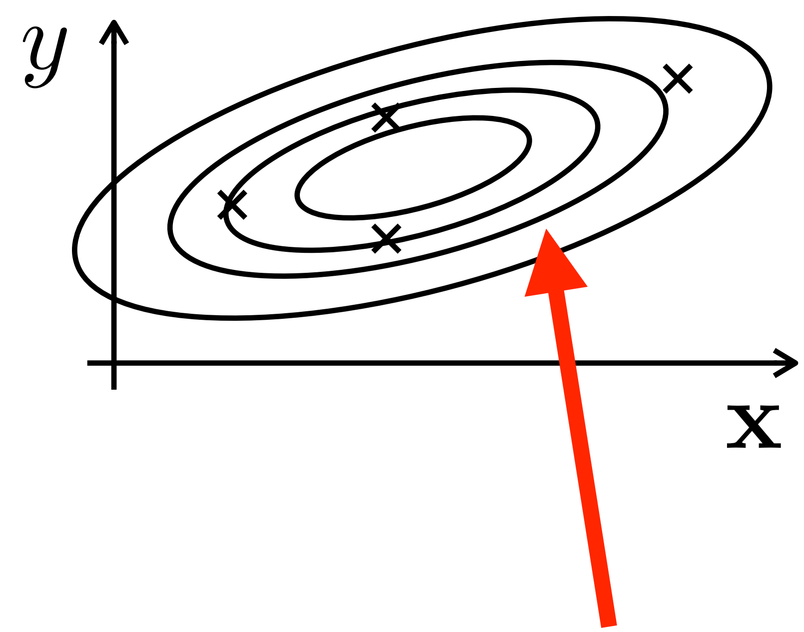
$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) + (-\log p(\mathbf{w}))$$

log likelihood prior/regulariser



$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) + (-\log p(\mathbf{w}))$$

log likelihood   prior/regulariser

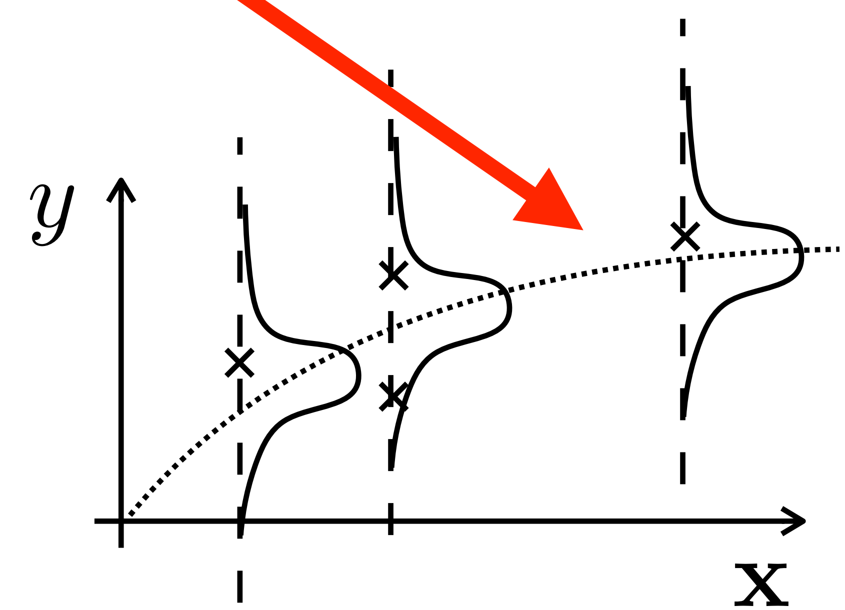


$p(y|\mathbf{x}, \mathbf{w})$

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) + (-\log p(\mathbf{w}))$$

log likelihood prior/regulariser

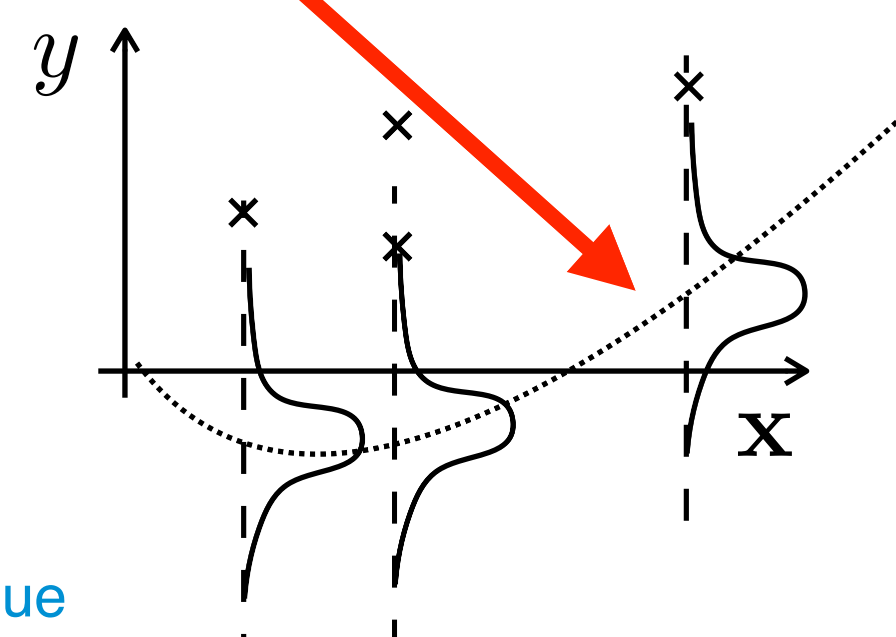
- **Regression:** $p(y|\mathbf{x}, \mathbf{w}) \sim \mathcal{N}_y(f(\mathbf{x}, \mathbf{w}), \sigma^2)$



$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) + (-\log p(\mathbf{w}))$$

log likelihood prior/regulariser

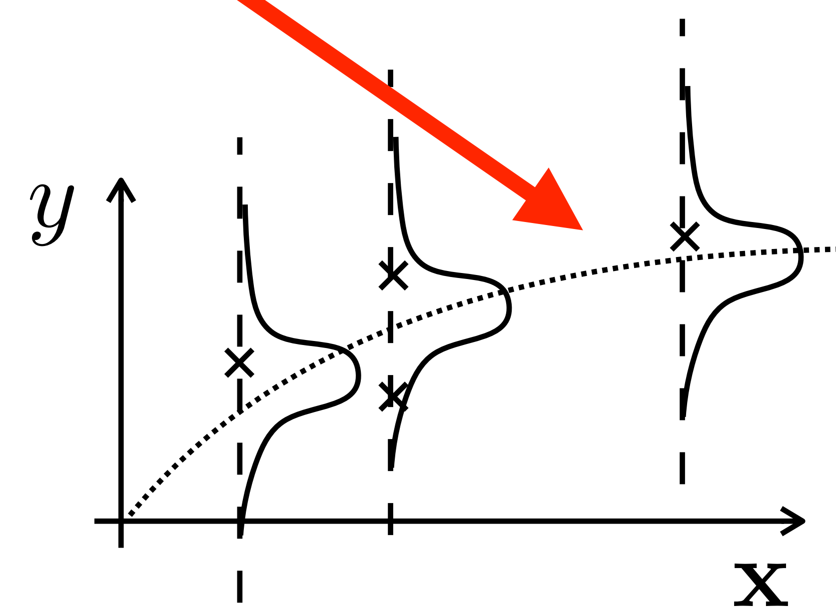
- **Regression:** $p(y|\mathbf{x}, \mathbf{w}) \sim \mathcal{N}_y(f(\mathbf{x}, \mathbf{w}), \sigma^2)$



$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) + (-\log p(\mathbf{w}))$$

log likelihood prior/regulariser

- **Regression:** $p(y|\mathbf{x}, \mathbf{w}) \sim \mathcal{N}_y(f(\mathbf{x}, \mathbf{w}), \sigma^2)$

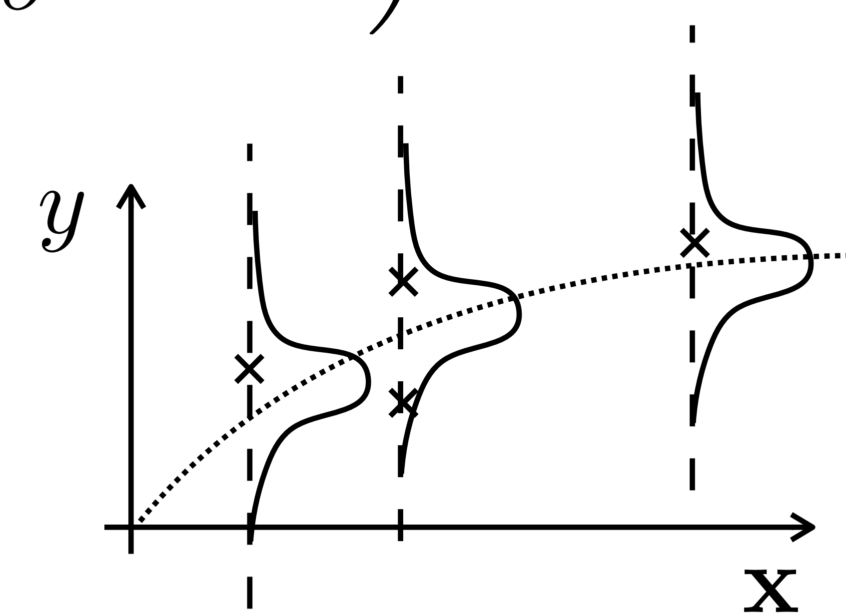


$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) + (-\log p(\mathbf{w}))$$

log likelihood — prior/regulariser

- **Regression:** $p(y|\mathbf{x}, \mathbf{w}) \sim \mathcal{N}_y(f(\mathbf{x}, \mathbf{w}), \sigma^2)$
- Probability of observing y_i when measuring $\mathbf{x}_i$ is

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(f(\mathbf{x}_i, \mathbf{w}) - y_i)^2}{2\sigma^2}\right)$$



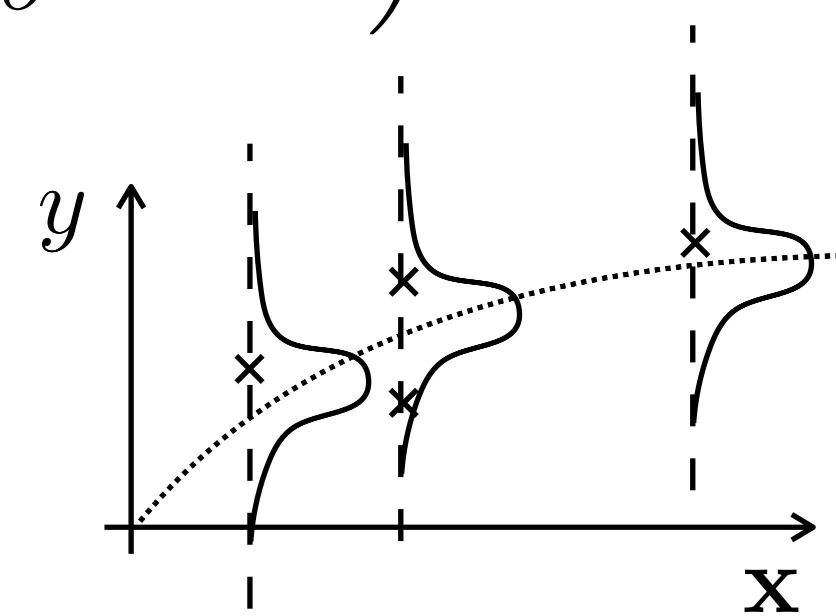
$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) \cdot$$

log likelihood prior/regulariser

- **Regression:** $p(y|\mathbf{x}, \mathbf{w}) \sim \mathcal{N}_y(f(\mathbf{x}, \mathbf{w}), \sigma^2)$
- Probability of observing y_i when measuring $\mathbf{x}_i$ is

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(f(\mathbf{x}_i, \mathbf{w}) - y_i)^2}{2\sigma^2} \right)$$

- Let us substitute it into the loss function (ignore prior for now)



$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right)$$

log likelihood prior/regulariser

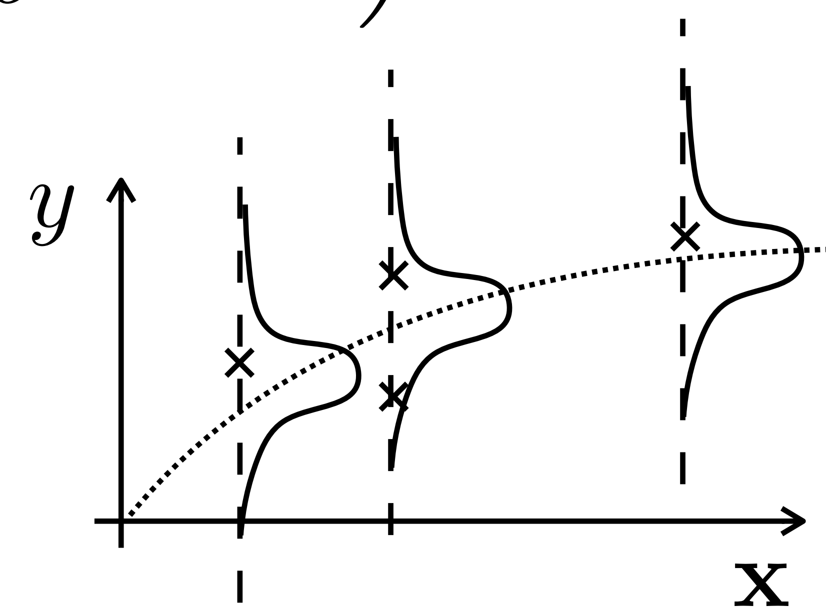
- **Regression:** $p(y|\mathbf{x},\mathbf{w}) \sim \mathcal{N}_y(f(\mathbf{x},\mathbf{w}),\sigma^2)$
- Probability of observing y_i when measuring $\mathbf{x}_i$ is

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(f(\mathbf{x}_i,\mathbf{w}) - y_i)^2}{2\sigma^2}\right)$$

- which yields well known L2 loss

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_i (f(\mathbf{x}_i,\mathbf{w}) - y_i)^2$$

- Especially $f(\mathbf{x},\mathbf{w}) = \mathbf{w}^\top \overline{\mathbf{x}}$



$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right)$$

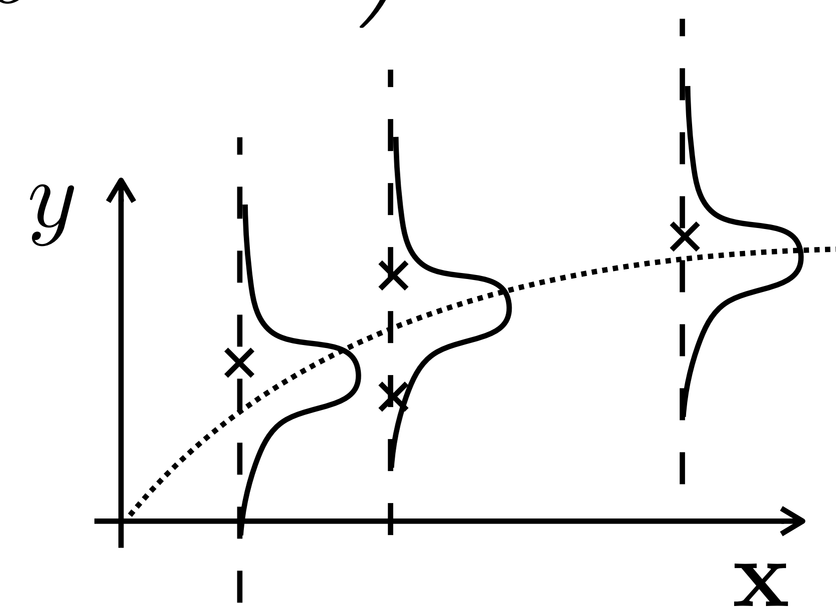
log likelihood prior/regulariser

- **Regression:** $p(y|\mathbf{x}, \mathbf{w}) \sim \mathcal{N}_y(f(\mathbf{x}, \mathbf{w}), \sigma^2)$
- Probability of observing y_i when measuring $\mathbf{x}_i$ is

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(f(\mathbf{x}_i, \mathbf{w}) - y_i)^2}{2\sigma^2} \right)$$

- which yields well known L2 loss

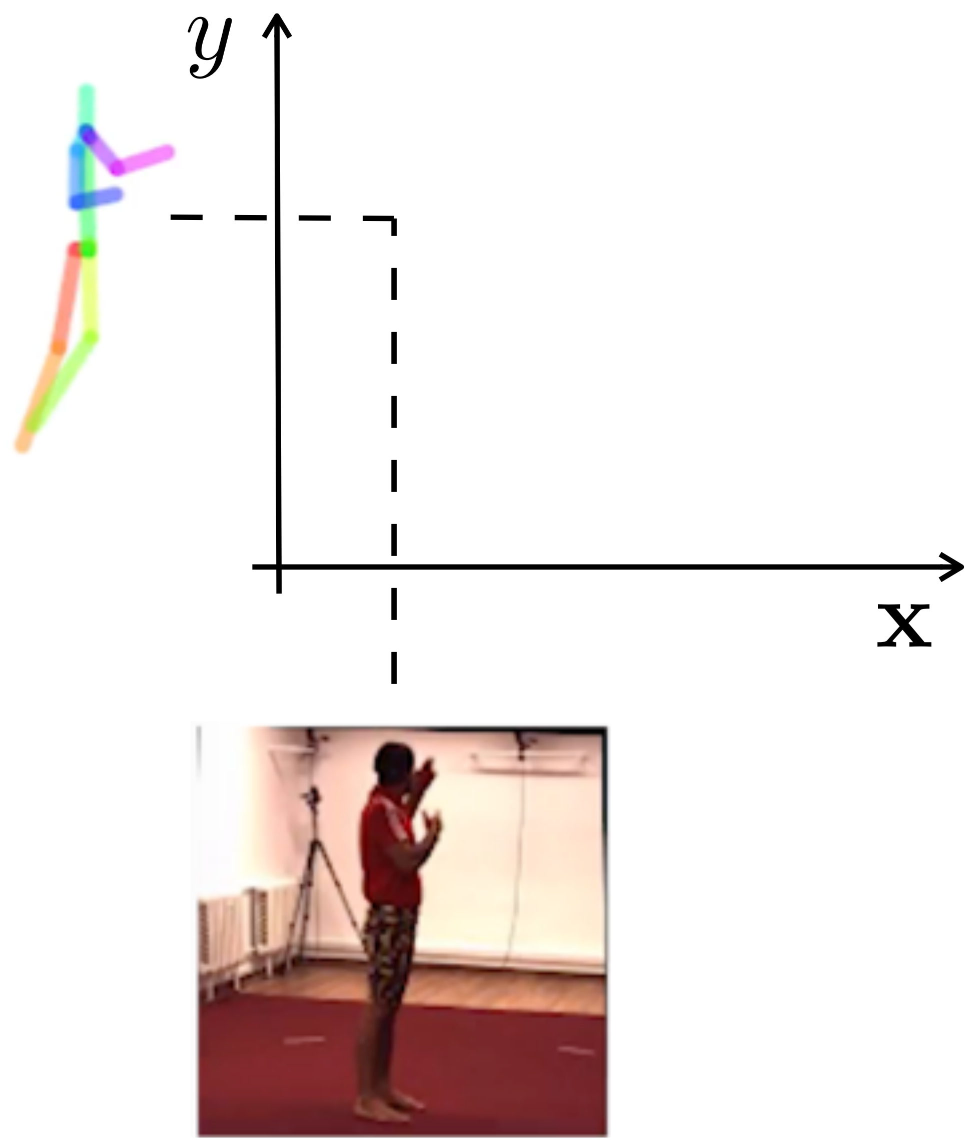
$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_i (f(\mathbf{x}_i, \mathbf{w}) - y_i)^2$$



- Especially $f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^\top \overline{\mathbf{x}}$ yields closed-form solution

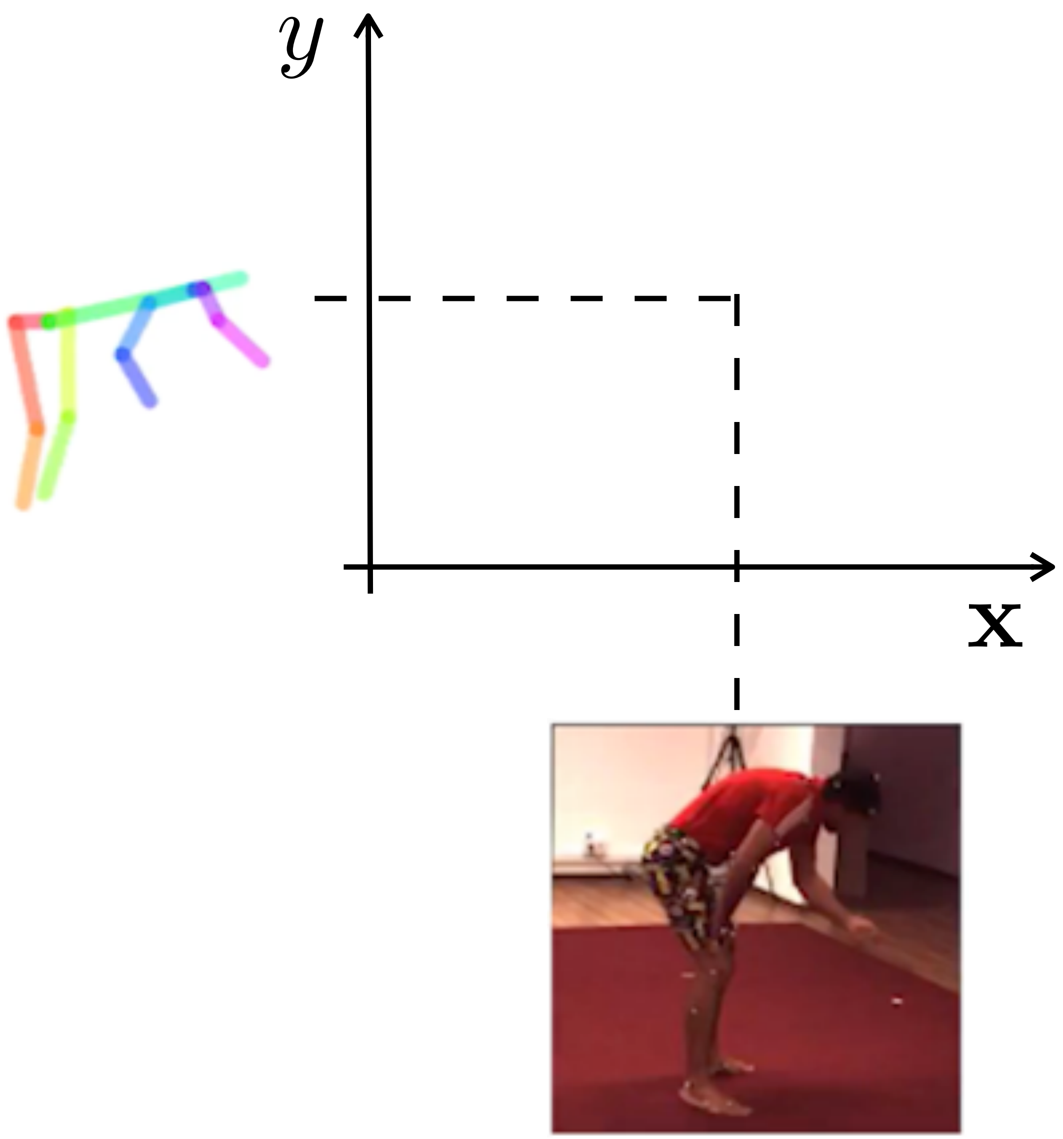
Other examples discussed during the course

3D pose regression



Other examples discussed during the course

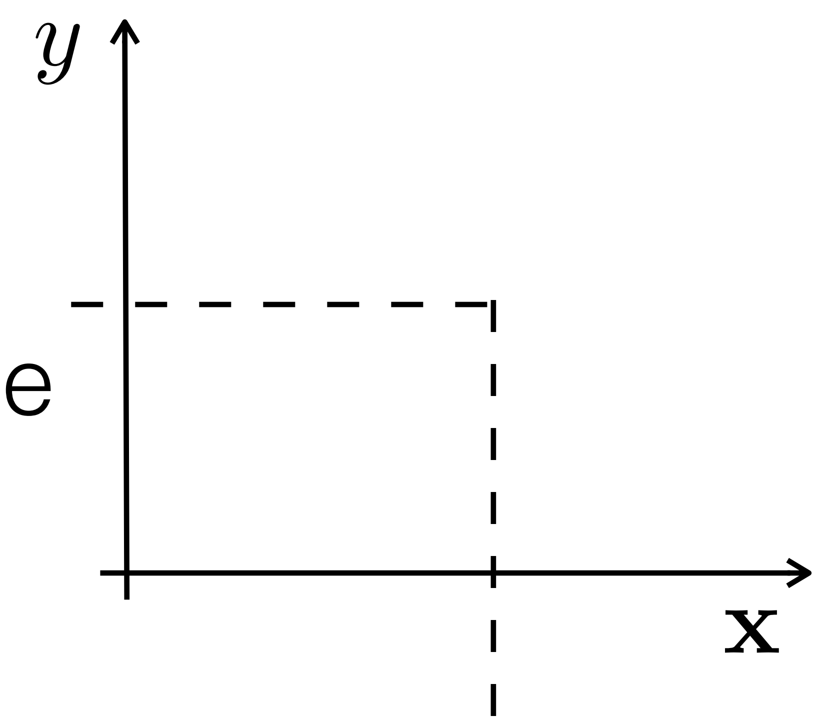
3D pose regression



Other examples discussed during the course

Camera calibration

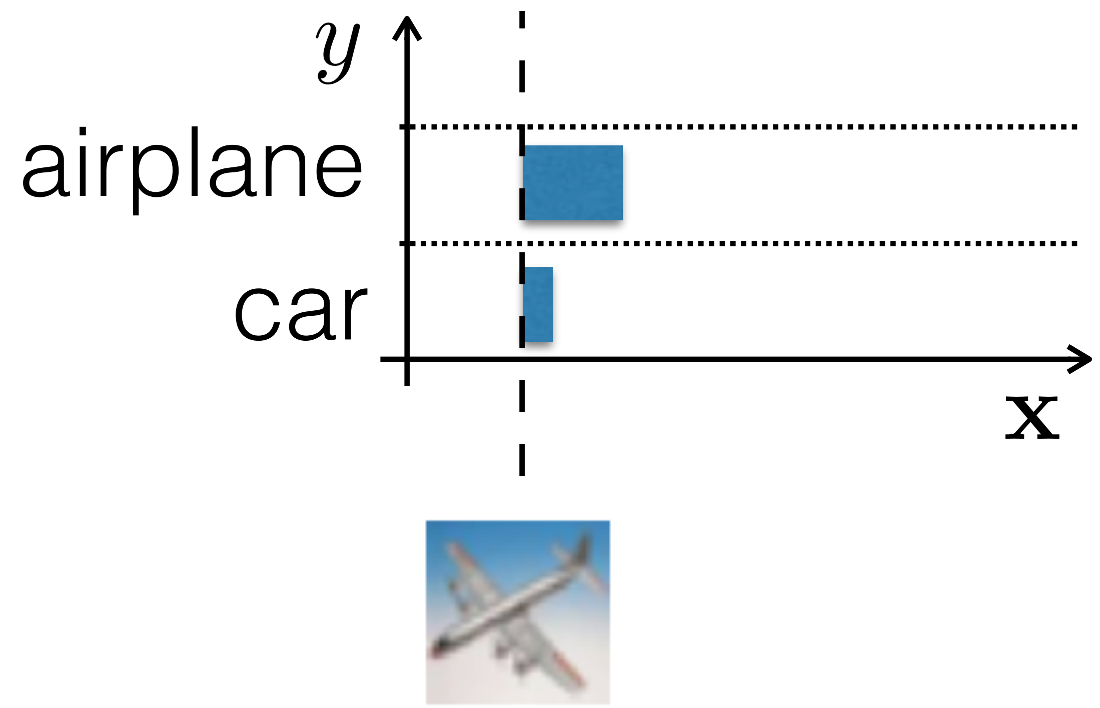
2D point
on camera image plane



3D point (e.g. lidar measurement)

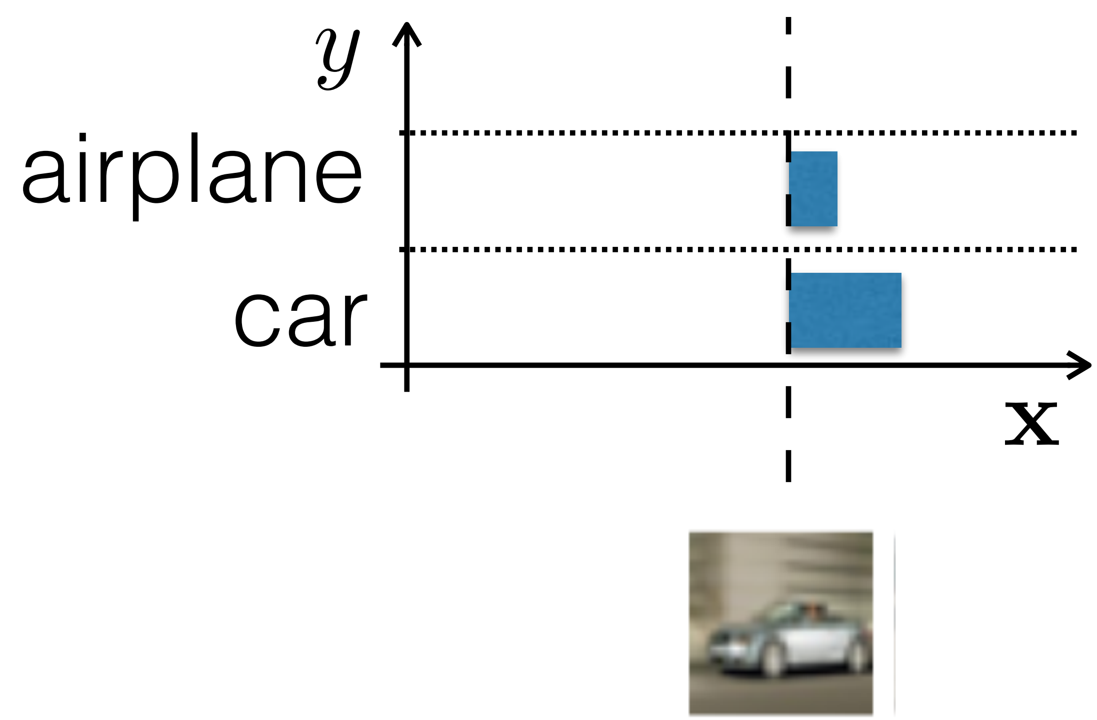
Other examples discussed during the course

Two-class object classification from RGB images



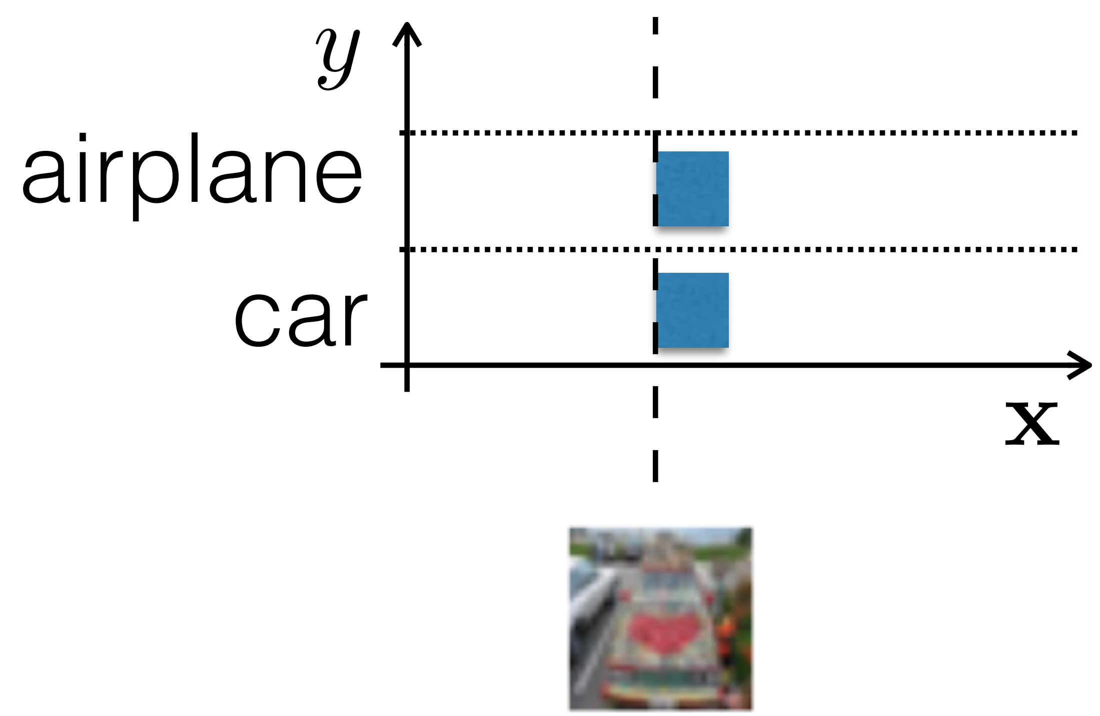
Other examples discussed during the course

Two-class object classification from RGB images



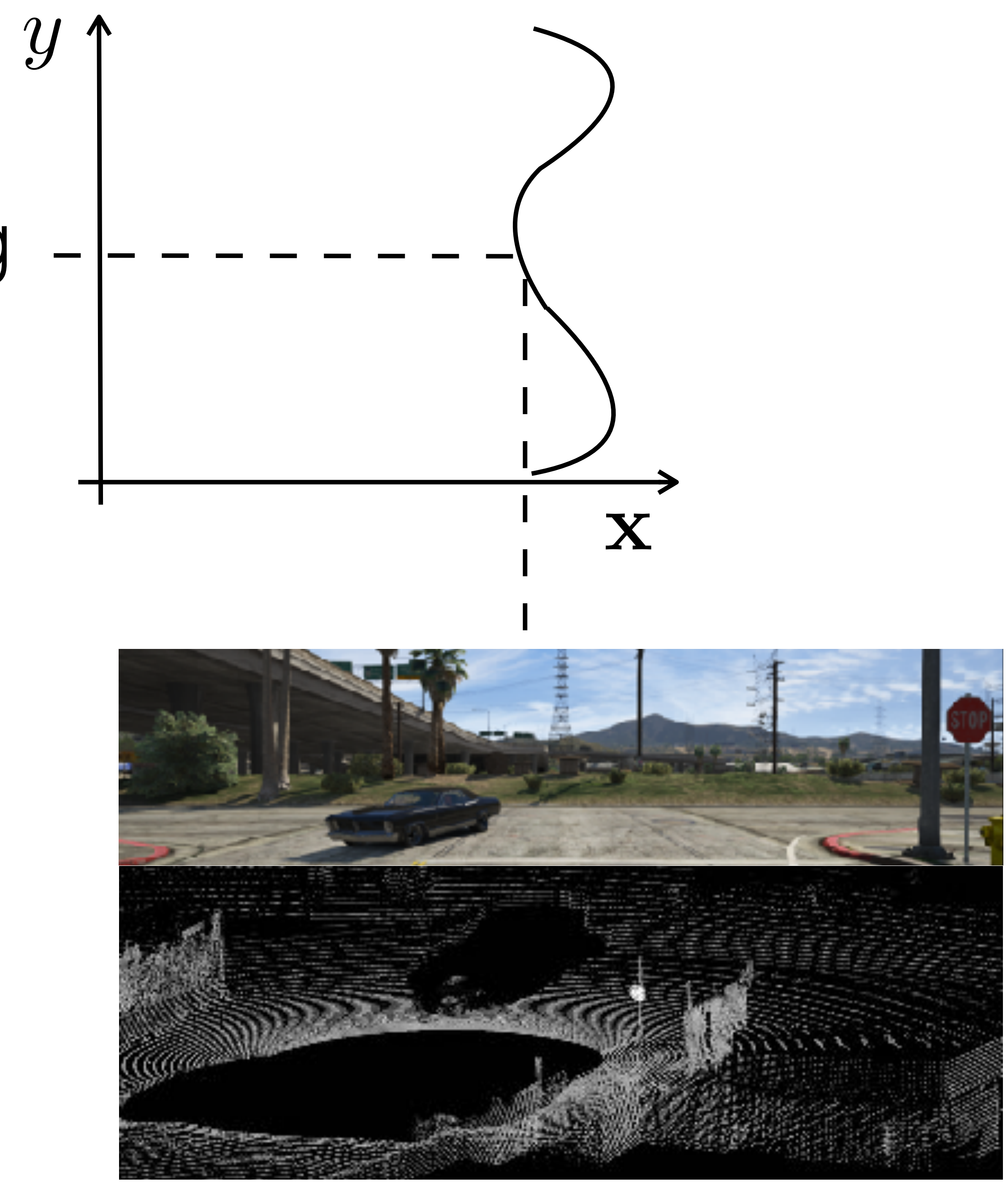
Other examples discussed during the course

Two-class object classification from RGB images



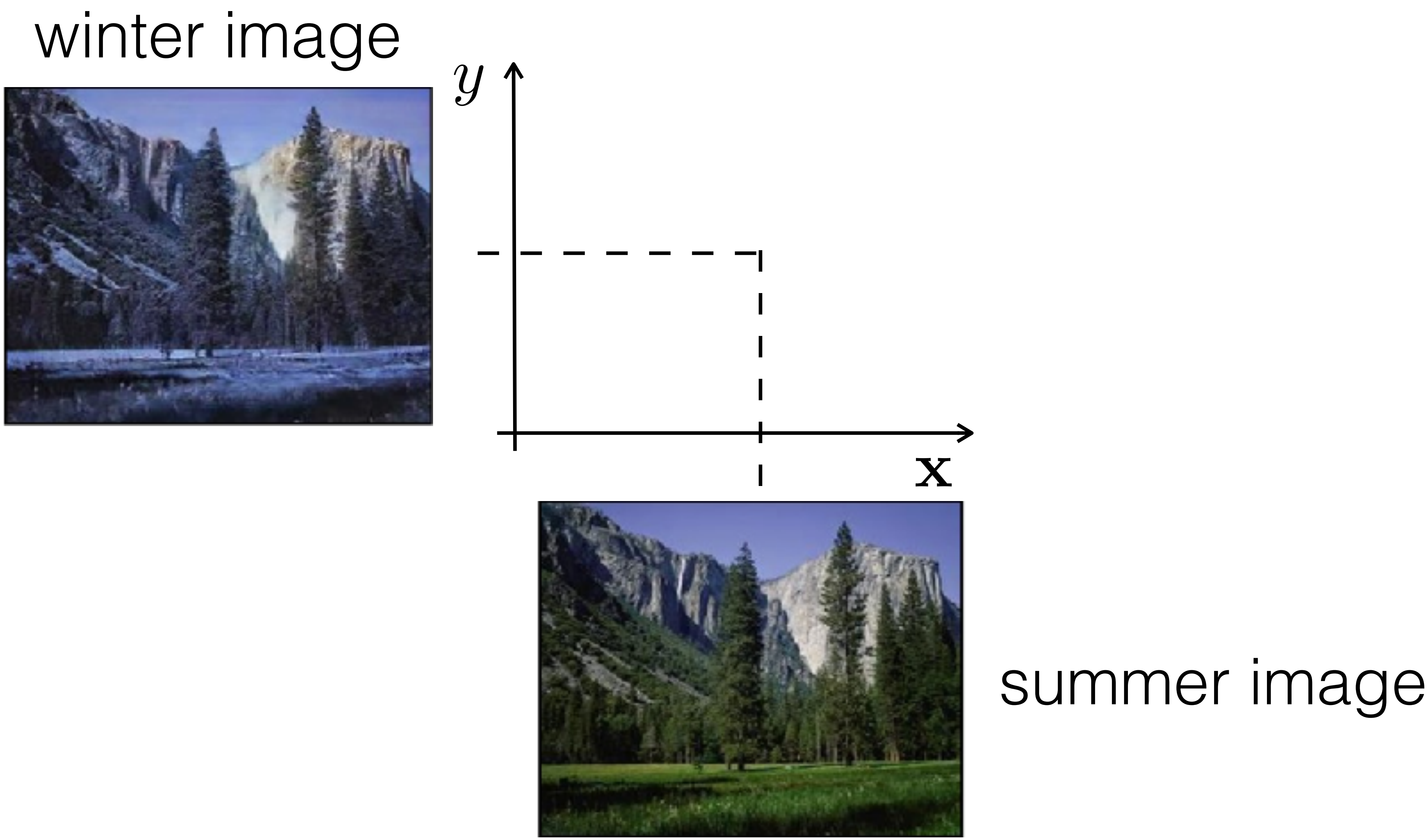
Other examples discussed during the course

Reactive control



Other examples discussed during the course

Generative networks



Other examples discussed during the course

- "x" and/or "y" could be high-dimensional

- Assuming Gaussian noise is in many cases myopic
  - Pose regression left/right hand is often indistinguishable
  - Right/left avoiding of an obstacle should be replaced by a mean (center).
  - Coloring of grayscale images is also obviously not gaussian

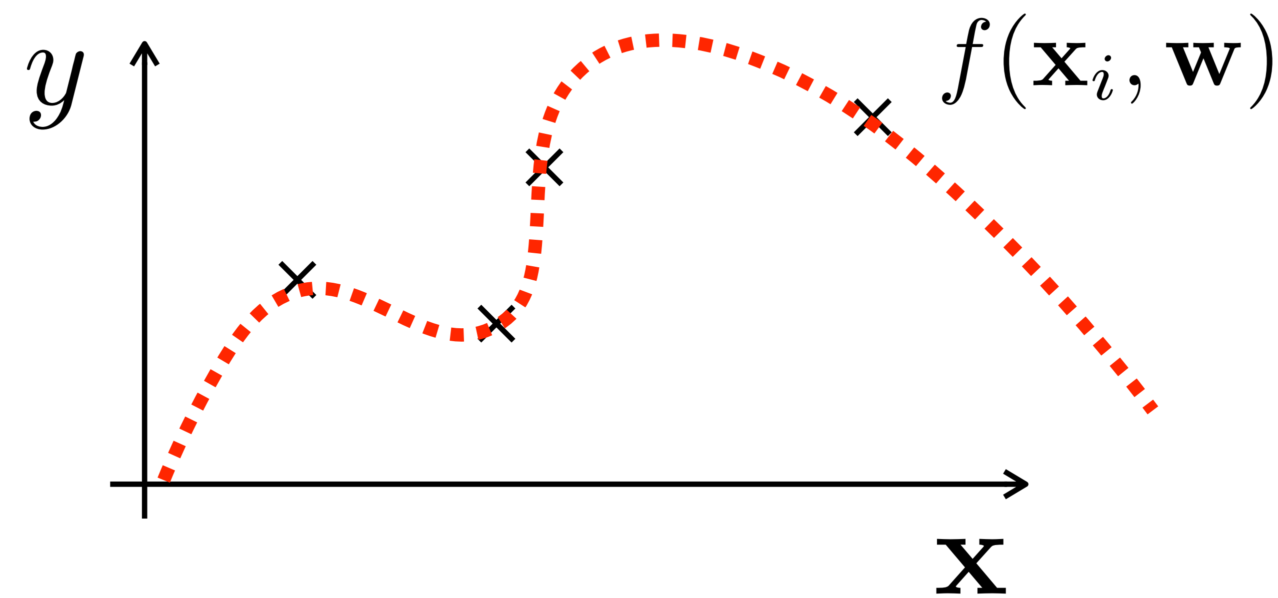
- Linear function is obviously insufficient in many cases => more complex models needed.

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i - \log(p(y_i | \mathbf{x}_i, \mathbf{w})) \right) \cdot \quad \boxed{}$$

log likelihood prior/regulariser

- **Prior** is important:

no prior, powerful f => overfitting



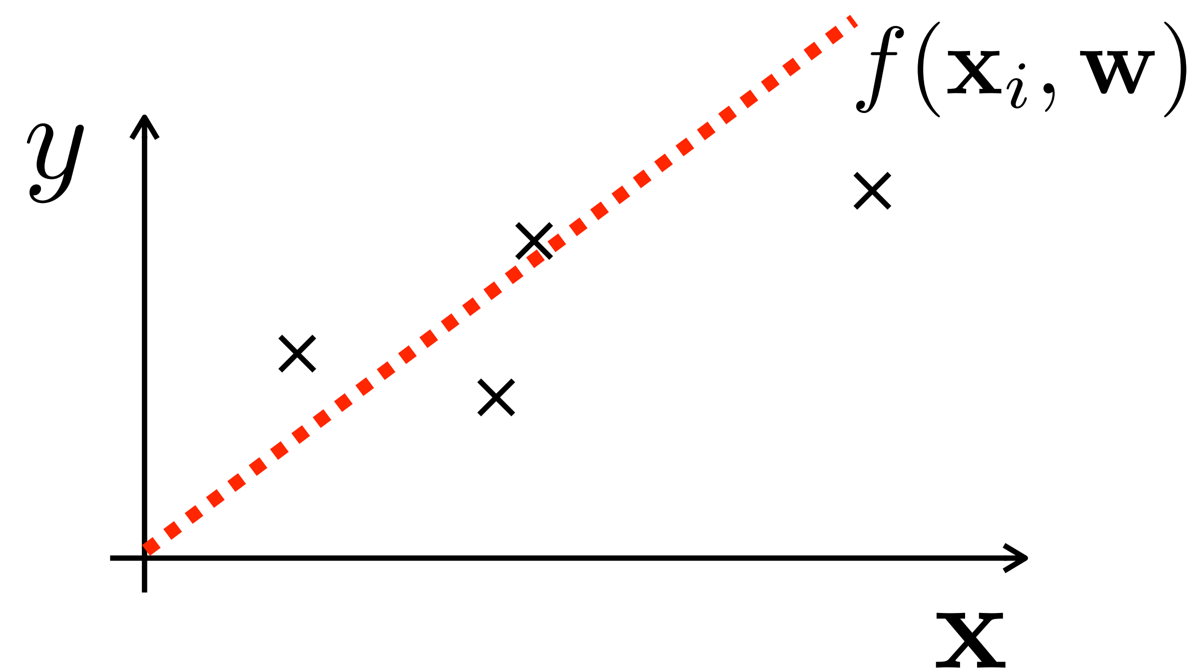
$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right)$$

log likelihood

prior/regulariser

- **Prior** is important:

no prior, simple f => underfitting

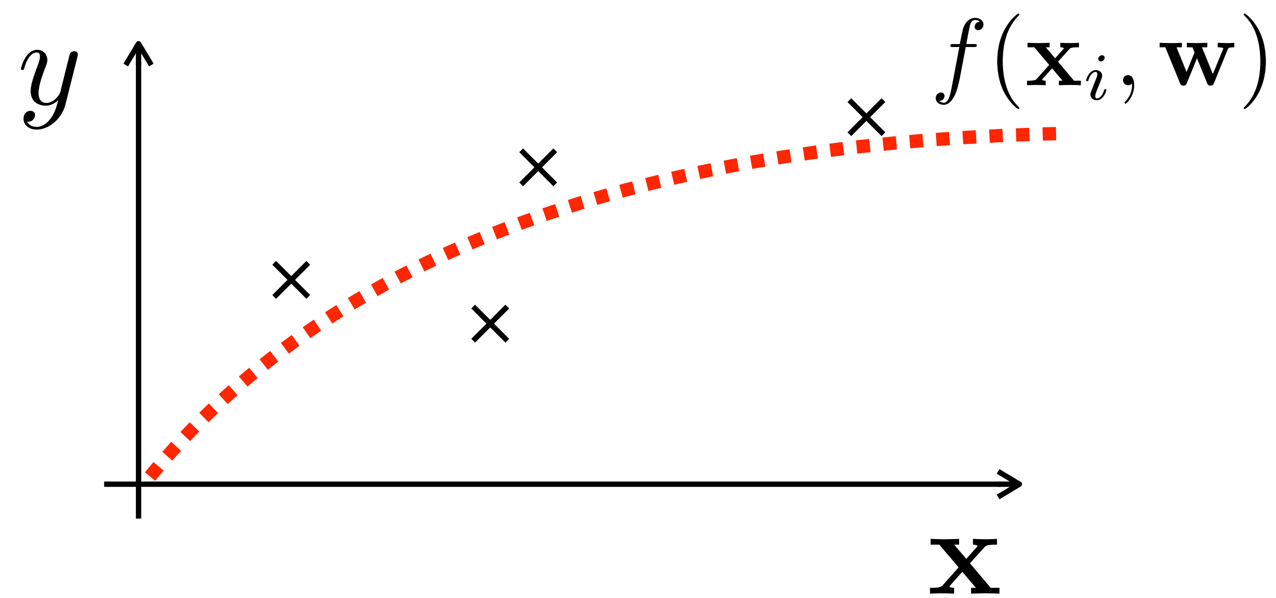


$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) + (-\log p(\mathbf{w}))$$

log likelihood   prior/regulariser

- **Prior** is important:

good prior



$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) + (-\log p(\mathbf{w}))$$

log likelihood prior/regulariser

- **Prior** is important:
  - Any prior knowledge restricts class of functions $f(\mathbf{x}_i, \mathbf{w})$ (e.g. for the class of linear functions the probability of non-zero weight for higher degrees monomials is zero)

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) + (-\log p(\mathbf{w}))$$

log likelihood prior/regulariser

- **Prior** is important:
  - Any prior knowledge restricts class of functions $f(\mathbf{x}_i, \mathbf{w})$ (e.g. for the class of linear functions the probability of non-zero weight for higher degrees monomials is zero)
  - Gaussian prior $p(\mathbf{w}) \sim \mathcal{N}_{\mathbf{w}}(\mathbf{0}, \lambda\mathbb{I})$ yields L2 regularization (it adds eye matrix to least squares)

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) + (-\log p(\mathbf{w}))$$

log likelihood prior/regulariser

- **Prior** is important:
  - Any prior knowledge restricts class of functions $f(\mathbf{x}_i, \mathbf{w})$ (e.g. probability of non-zero weight for higher degrees monomials is zero)
  - Gaussian prior $p(\mathbf{w}) \sim \mathcal{N}_{\mathbf{w}}(\mathbf{0}, \lambda\mathbb{I})$ yields L2 regularization (it adds eye matrix to least squares)
  - Regression with L1 regularization is known as Lasso

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) + (-\log p(\mathbf{w}))$$

log likelihood prior/regulariser

- **Prior** is important:
  - Any prior knowledge restricts class of functions $f(\mathbf{x}_i, \mathbf{w})$ (e.g. probability of non-zero weight for higher degrees monomials is zero)
  - Gaussian prior $p(\mathbf{w}) \sim \mathcal{N}_{\mathbf{w}}(\mathbf{0}, \lambda \mathbb{I})$ yields L2 regularization (it adds eye matrix to least squares)
  - Regression with L1 regularization is known as Lasso
  - Well chosen prior partially reduces overfitting
  - Occam's Razor

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i - \log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) + (-\log p(\mathbf{w}))$$

loss function prior/regulariser

William of Ockham (1287-1347)

leprechauns can be involved in any explanation

https://en.wikipedia.org/wiki/Occam%27s_razor

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) + (-\log p(\mathbf{w}))$$

log likelihood prior/regulariser

- It is very important to avoid any "*not-well justified leprechauns*" in the model, otherwise any learning (parameter estimations) may suffer from too complex explanations => overfitting

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) + (-\log p(\mathbf{w}))$$

log likelihood prior/regulariser

- It is very important to avoid any "*not-well justified leprechauns*" in the model, otherwise any learning (parameter estimations) may suffer from too complex explanations => overfitting
- Consequently we study different phenomenas
  - animal cortex structure (for ConvNets)
  - geometry of rigid motion (for robot/scene motion or DKT)
  - projective transformation of pinhole cameras

  to create as simple (i.e.leprechauns-free) model as possible

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i - \log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right)$$

ML estimate

log likelihood

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sum_i -\log(p(y_i|\mathbf{x}_i, \mathbf{w})) \right) + (-\log p(\mathbf{w}))$$

ML estimate

log likelihood

prior/regulariser

MAP estimate

Conclusions

- Explained regression as MAP/ML estimator
- Discussed under/overfitting and regularisations

Competencies required for the test T1

- Derive MAP/ML estimate for regression,
- Compute L2-loss,
- Understand difference between loss, likelihood and prior
- Understand role of prior in underfitting/overfitting.