DD2434 - Advanced Machine Learning

My best friend the Gaussian

Carl Henrik Ek
{chek}@csc.kth.se

Royal Institute of Technology

November 21, 2014

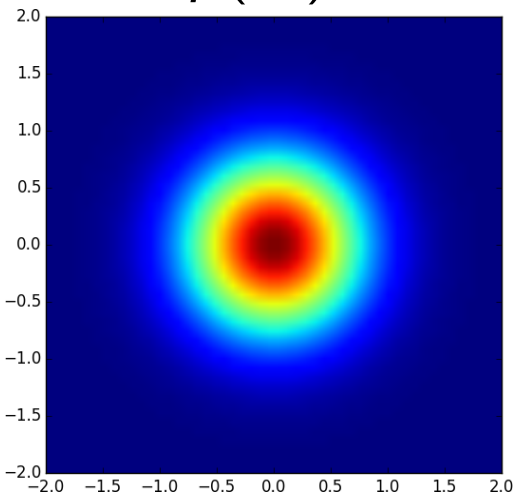
Last Lecture

- General Probabilistic Modelling
  - Probabilistic objects
  - Marginalisation
- Kernels
  - Dual linear regression
  - Implications for modelling

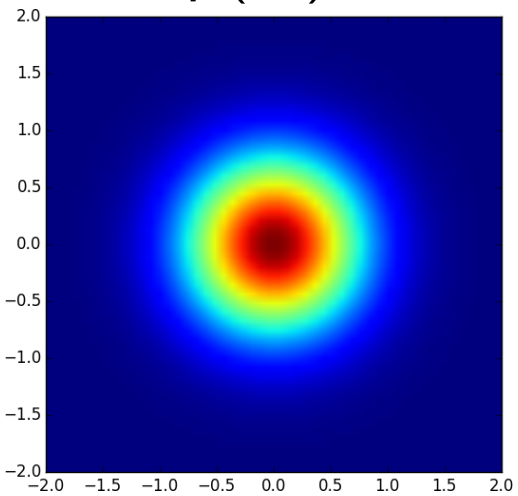
Toolbox

1. Formulate prediction error likelihood
   - Does the likelihood have structure?
2. Formulate belief of model in prior
   - Does the prior have structure
3. Marginalise irrelevant variables
4. Choose model based on *evidence* $p(\mathcal{D}|\mathcal{M})$

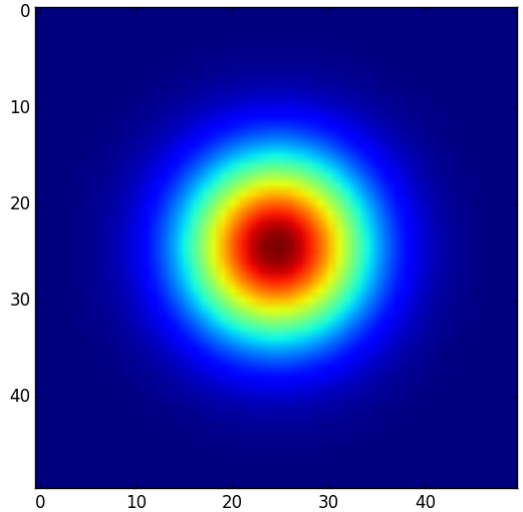
$p(\mathbf{W})$



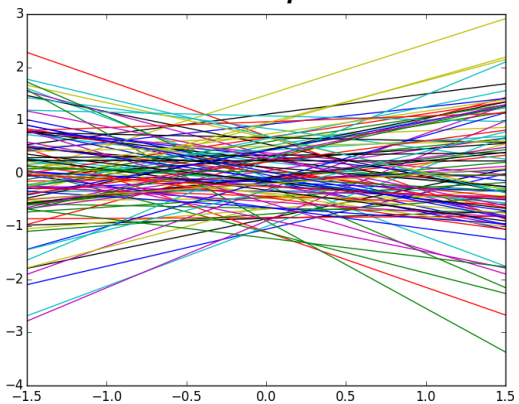
$p(\mathbf{W})$



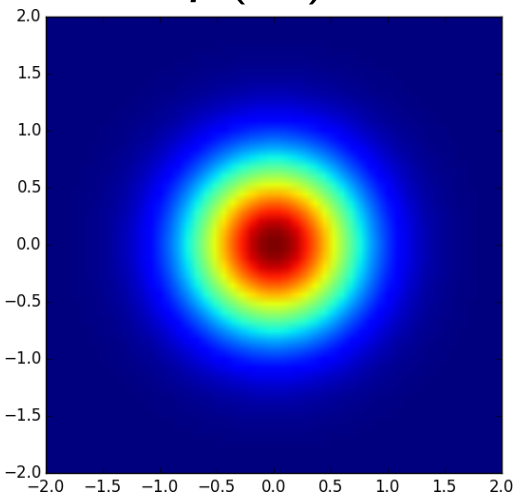
$p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$



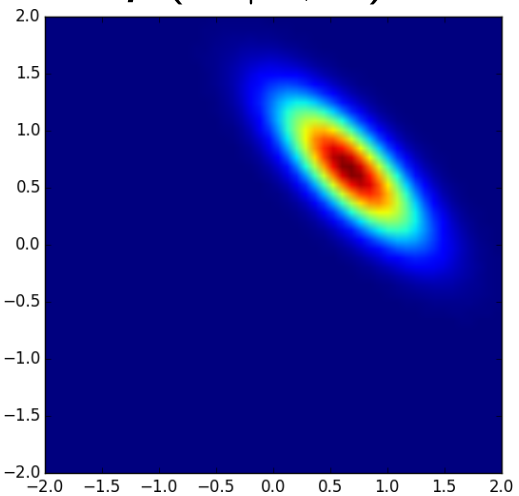
$W_{samples}$



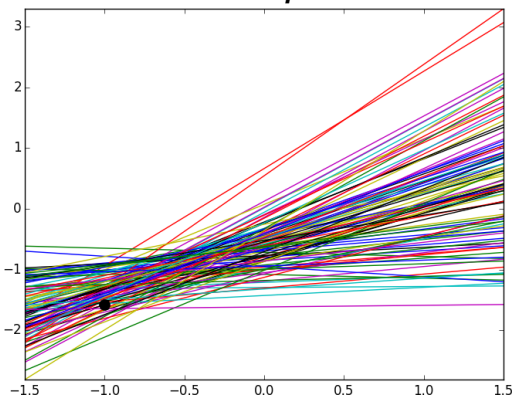
$p(\mathbf{W})$



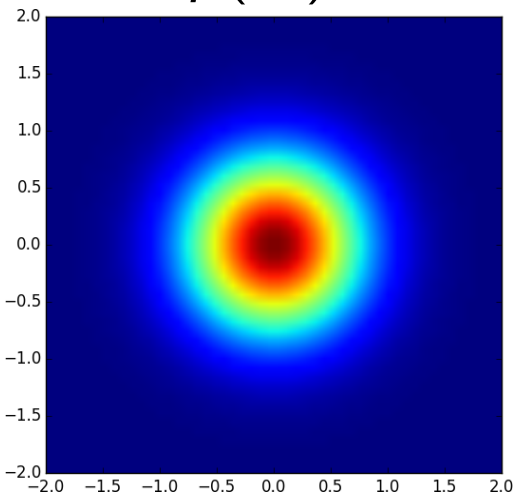
$p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$



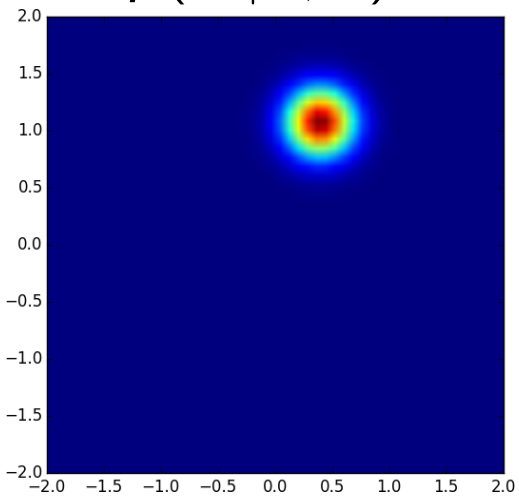
$W_{samples}$



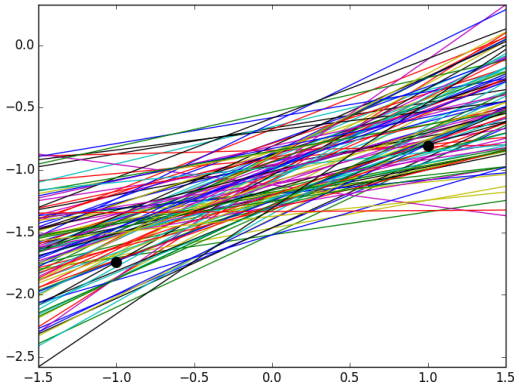
$p(\mathbf{W})$



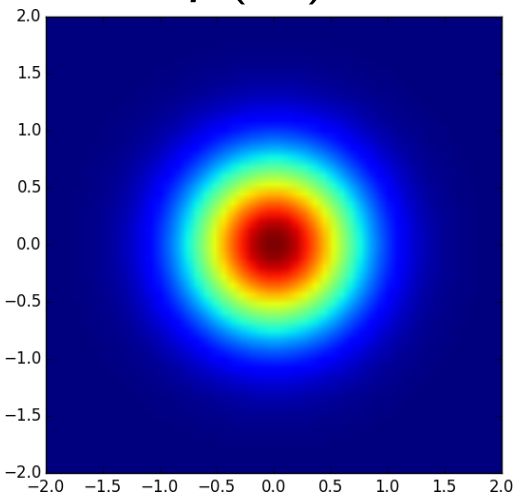
$p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$



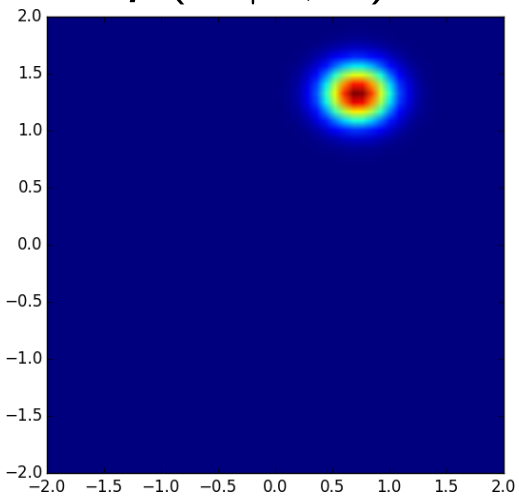
$W_{samples}$



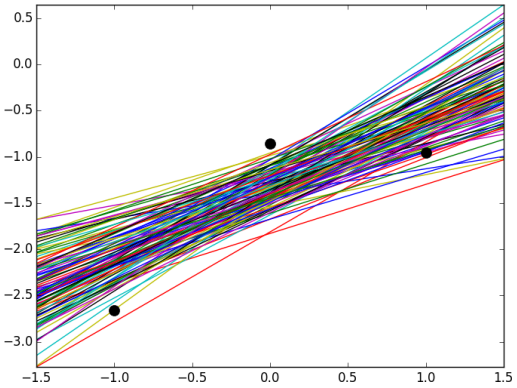
$p(\mathbf{W})$



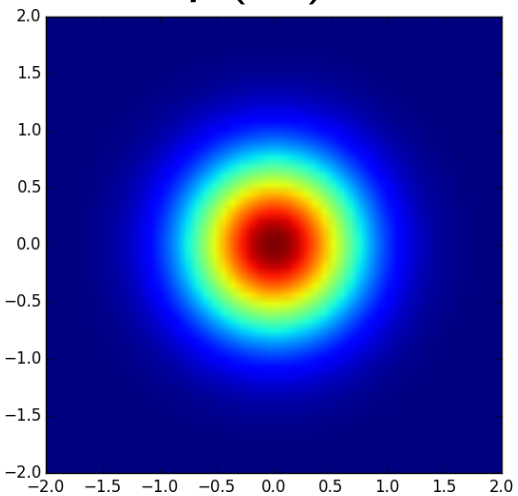
$p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$



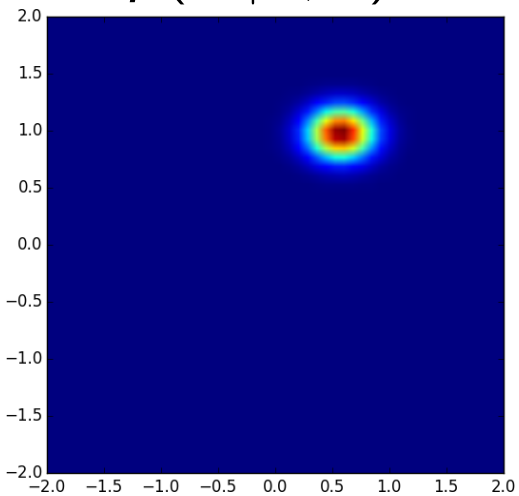
$W_{samples}$



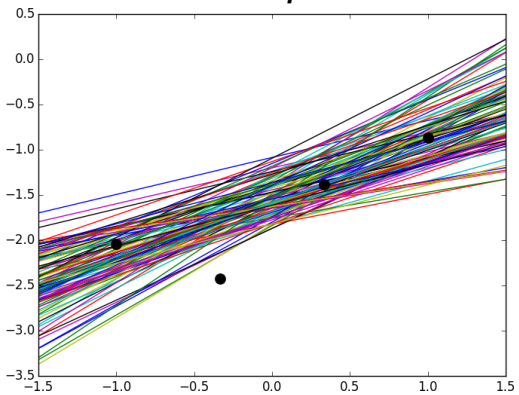
$p(\mathbf{W})$



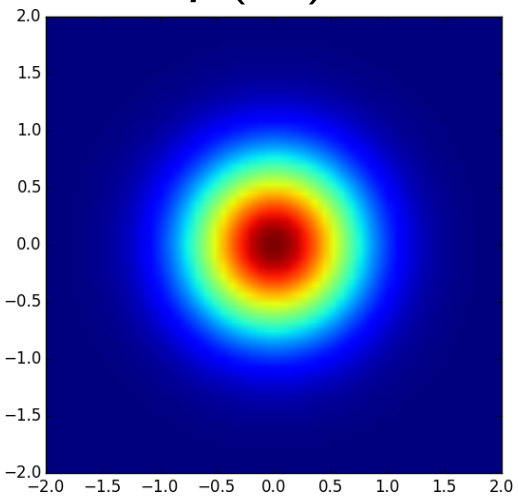
$p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$



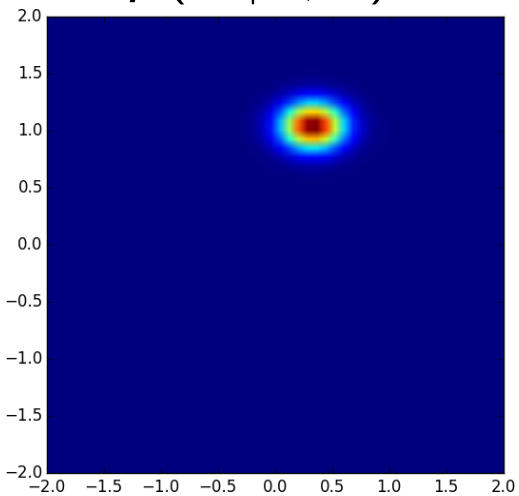
$W_{samples}$



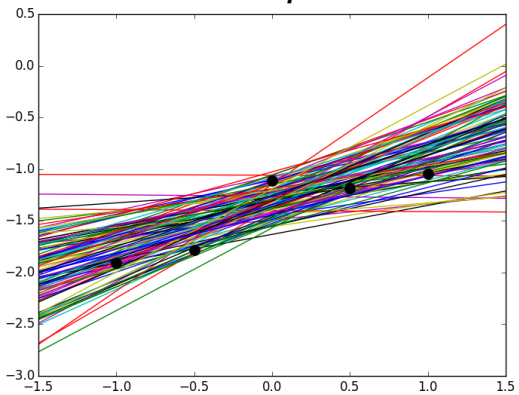
$p(\mathbf{W})$



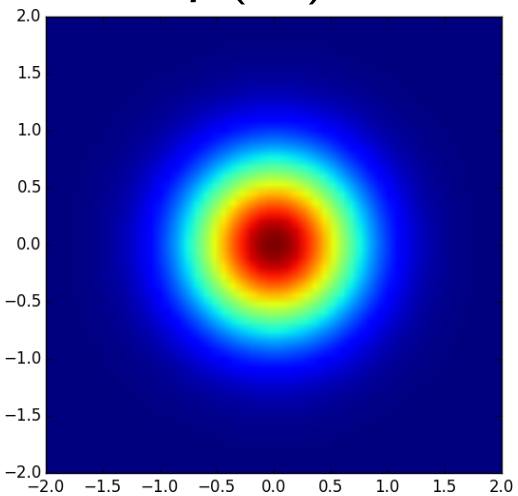
$p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$



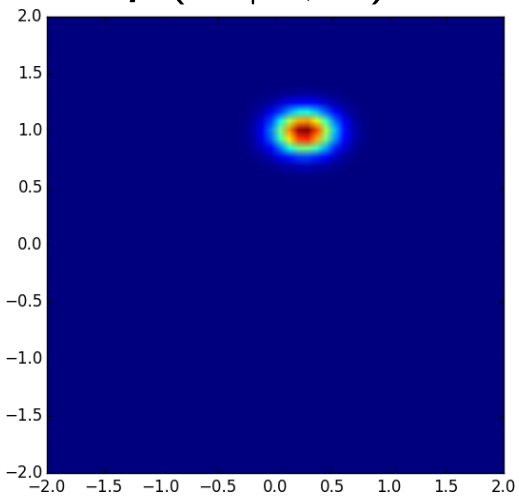
$W_{samples}$



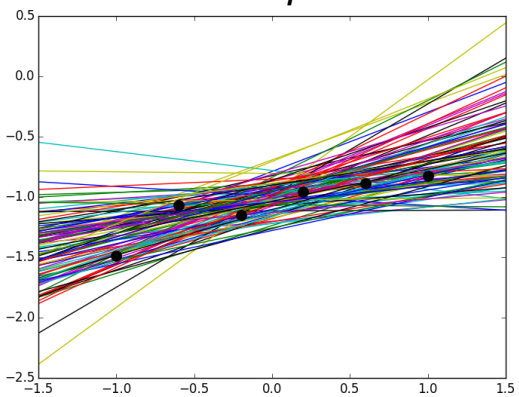
$p(\mathbf{W})$



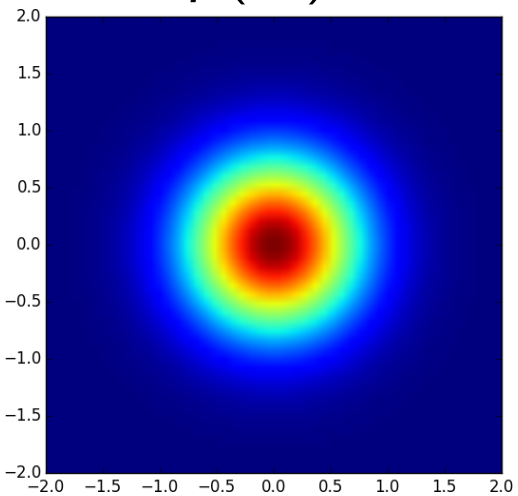
$p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$



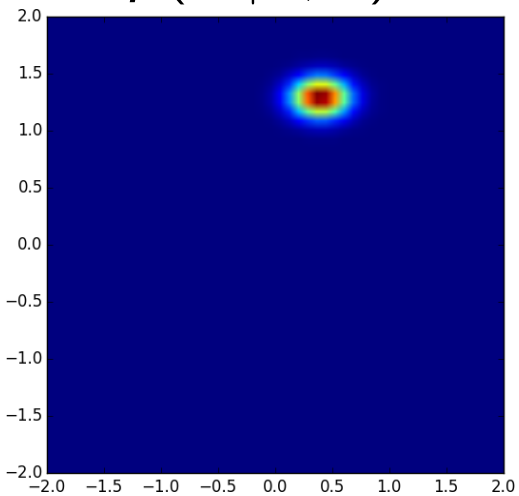
$W_{samples}$



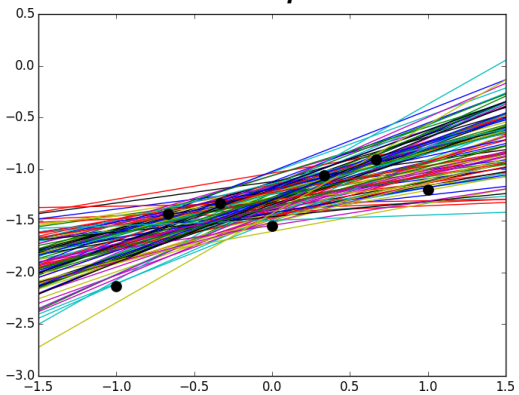
$p(\mathbf{W})$



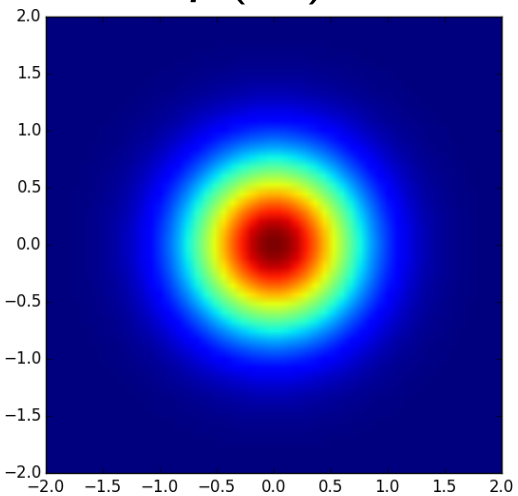
$p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$



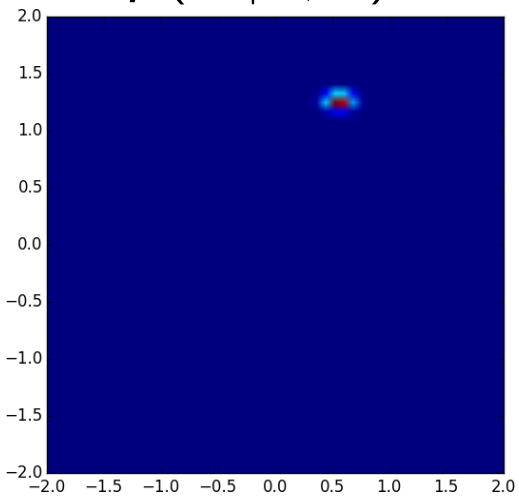
$W_{samples}$



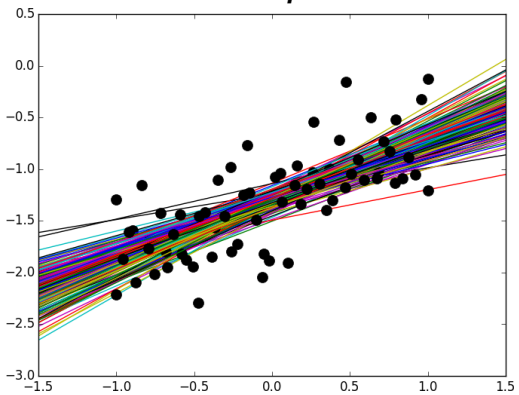
$p(\mathbf{W})$



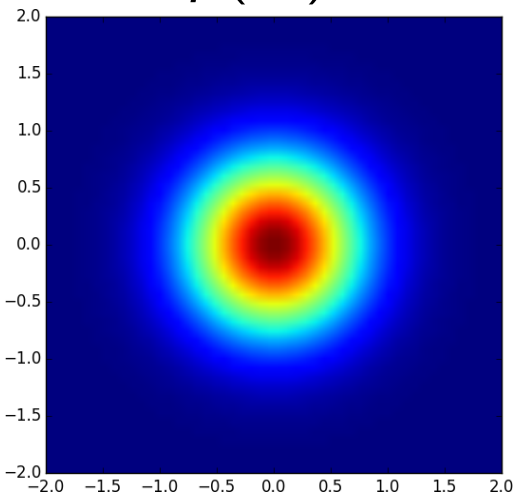
$p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$



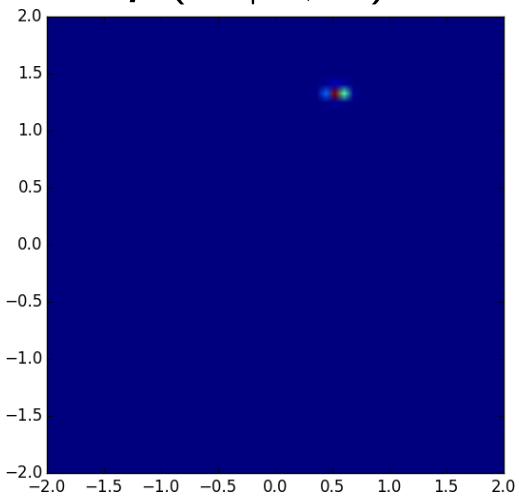
$W_{samples}$



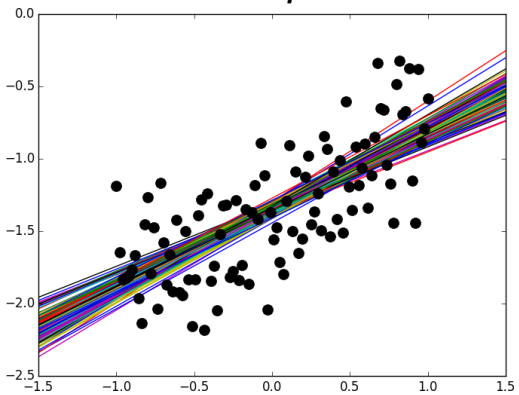
$p(\mathbf{W})$



$p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$



$W_{samples}$



Posterior Distribution[1]

$$p(\mathbf{X}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{p(\mathbf{Y})} \tag{1}$$

$$p(\mathbf{X}|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X})p(\mathbf{X}) \tag{2}$$

- The posterior and the prior are in the same *family*
- Relationship with all **three** terms
- How to multiply two different Gaussians

[1] Murphy 2012, pp. 4.3.4 & 4.4.3

Posterior Distribution[1]

$$p(\mathbf{X}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{p(\mathbf{Y})} \tag{3}$$

$$p(\mathbf{X}|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X})p(\mathbf{X}) \tag{4}$$

- The posterior and the prior are in the same *family*
- Relationship with all **three** terms
- How to multiply two different Gaussians

[1] Murphy 2012, pp. 4.3.4 & 4.4.3

Conditional Distribution[2]

$$p(\mathbf{X}_1, \mathbf{X}_2) = p(\mathbf{X}_1|\mathbf{X}_2)p(\mathbf{X}_2) \tag{5}$$

- We often have the joint distribution
- Want to make predictions of one variable
- Conditional distribution

[2] Murphy 2012, pp. 4.3.4 & 4.4.3

Marginalisation[3]

$$p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{W}, \mathbf{X}) p(\mathbf{W}) \mathrm{d}\mathbf{W} \tag{6}$$

- Average according to belief and how well the model fits the observations
- "Pushes" uncertain belief in parameters (in this case) through to the observations
- How to marginalise two Gaussians with each other

[3] Murphy 2012, p. 4.3.1

Marginalisation[3]

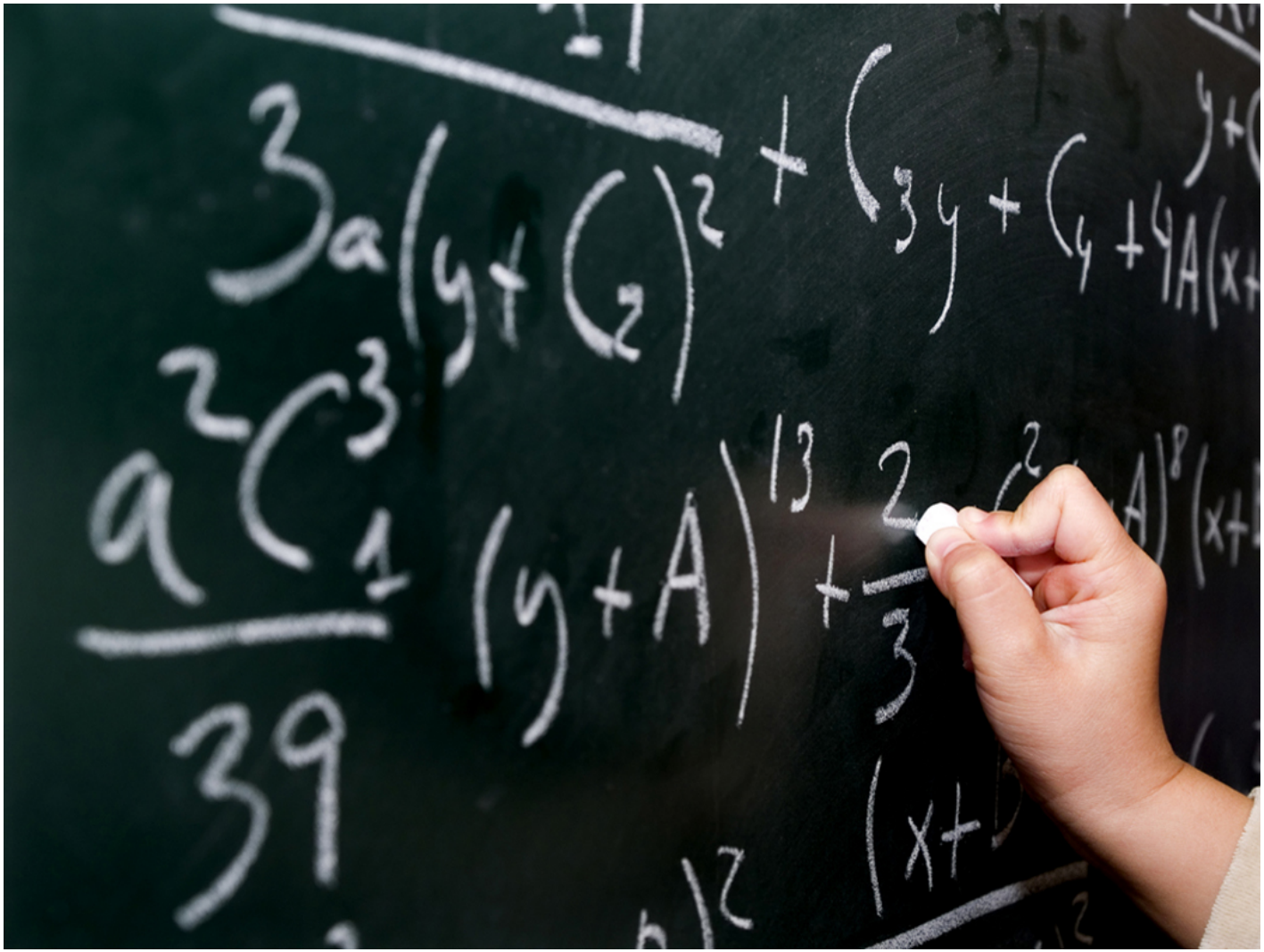
$$p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{W}, \mathbf{X}) p(\mathbf{W}) \mathrm{d}\mathbf{W} \tag{7}$$

- Average according to belief and how well the model fits the observations
- "Pushes" uncertain belief in parameters (in this case) through to the observations
- How to marginalise two Gaussians with each other

[3] Murphy 2012, p. 4.3.1

Outline

- Multivariate Gaussians
  - Mahalanobis distance
  - Marginals
  - Posterior
  - Conditionals
- Murphy 2012, Ch. 4
- Pointers: Matrix derivatives



Posteriors of parameters[4]

$$p(\mathbf{X}|\mu, \Sigma) = \mathcal{N}(\mathbf{X}|\mu, \Sigma) \tag{8}$$
$$p(\mu|\mathbf{X}, \Sigma) = ? \tag{9}$$
$$p(\Sigma|\mathbf{X}, \mu) = ? \tag{10}$$

- We can infer parameters of the model from data using posteriors
- Exactly the same framework

[4] Murphy 2012, pp. 4.6.1 & 4.6.2

Posteriors of parameters[4]

$$p(\mathbf{X}|\mu, \Sigma) = \mathcal{N}(\mathbf{X}|\mu, \Sigma) \tag{11}$$
$$p(\mu|\mathbf{X}, \Sigma) = ? \tag{12}$$
$$p(\Sigma|\mathbf{X}, \mu) = ? \tag{13}$$

- We can infer parameters of the model from data using posteriors
- Exactly the same framework

[4] Murphy 2012, pp. 4.6.1 & 4.6.2

Learning

$$\frac{\partial r}{\partial p} = \begin{matrix} \frac{\partial r_1}{\partial p_{a1}} & \dots & \frac{\partial r_1}{\partial p_{aN}} & \frac{\partial r_1}{\partial p_{b1}} & \dots & \frac{\partial r_1}{\partial p_{bM}} & \frac{\partial r_1}{\partial p_{c1}} & \dots & \frac{\partial r_1}{\partial p_{cQ}} \\ \frac{\partial r_2}{\partial p_{a1}} & \dots & \frac{\partial r_2}{\partial p_{aN}} & \frac{\partial r_2}{\partial p_{b1}} & \dots & \frac{\partial r_2}{\partial p_{bM}} & \frac{\partial r_2}{\partial p_{c1}} & \dots & \frac{\partial r_2}{\partial p_{cQ}} \\ \frac{\partial r_3}{\partial p_{a1}} & \dots & \frac{\partial r_3}{\partial p_{aN}} & \frac{\partial r_3}{\partial p_{b1}} & \dots & \frac{\partial r_3}{\partial p_{bM}} & \frac{\partial r_3}{\partial p_{c1}} & \dots & \frac{\partial r_3}{\partial p_{cQ}} \\ \frac{\partial r_4}{\partial p_{a1}} & \dots & \frac{\partial r_4}{\partial p_{aN}} & \frac{\partial r_4}{\partial p_{b1}} & \dots & \frac{\partial r_4}{\partial p_{bM}} & \frac{\partial r_4}{\partial p_{c1}} & \dots & \frac{\partial r_4}{\partial p_{cQ}} \end{matrix} \tag{28}$$

- Gradient based learning
- Often compute gradients of matrices
- “The matrix cookbook” - Petersen and Pedersen 2012

Next Time

Lecture 2

- November 25th 8-10 E2
- Continue with Kernels
  - relation to co-variance
- Non-parametric Regression
  - Gaussian Processes
  - Conditional & Marginal Gaussians

e.o.f.

References I

Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 0262018020, 9780262018029.

KB Petersen and MS Pedersen. "The matrix cookbook". In: *Technical University of Denmark* (Nov. 2012).