DD2434 - Advanced Machine Learning

Approximative Inference

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

Royal Institute of Technology

November 12th, 2015

Last Lecture

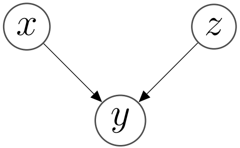
- Representation Learning
  - Same story as before
  - Priors even more important
  - Factor Analysis
  - PCA as a latent variable model
  - GP-LVM

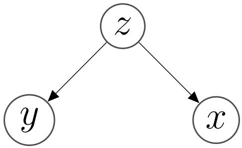
Introduction

Recap

Approximative Inference

Graphical Models[1]





$p(x, y, z) = p(y|x, z)p(x)p(z)$

$p(x, y, z) = p(y|z)p(x|z)p(z)$

$$p(\{x_i\}_{i=1}^{N}) = \prod_{i=1}^{N} p(x_i|\mathrm{pa}_i) \quad (1)$$

[1]Bishop 2006, pp. 8.0, 8.1.

Latent Variable Models[2]

Machine Learning

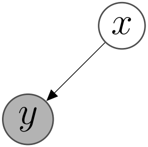
- What is our task?
- $p(y)$
- Latent variables
- Generative Model
- Explaining away

[2] Bishop 2006, p. 364.

Latent Variable Models[2]

Machine Learning

- What is our task?
- $p(y)$
- Latent variables
- Generative Model
- Explaining away

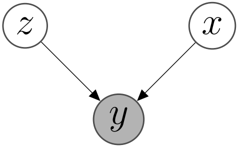


[2]Bishop 2006, p. 364.

Latent Variable Models[2]

Machine Learning

- What is our task?
- $p(y)$
- Latent variables
- Generative Model
- Explaining away

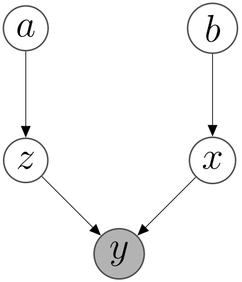


[2] Bishop 2006, p. 364.

Latent Variable Models[2]

Machine Learning

- What is our task?
- $p(y)$
- Latent variables
- Generative Model
- Explaining away

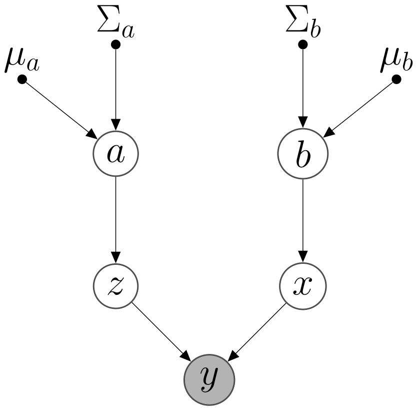


[2] Bishop 2006, p. 364.

Latent Variable Models[2]

Machine Learning

- What is our task?
- $p(y)$
- Latent variables
- Generative Model
- Explaining away

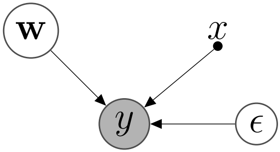


[2] Bishop 2006, p. 364.

Latent Variable Models[2]

Machine Learning

- What is our task?
- $p(y)$
- Latent variables
- Generative Model
- Explaining away



$$\mathbf{y}_i = \mathbf{w}\mathbf{x}_i + \boldsymbol{\epsilon} \tag{2}$$

[2] Bishop 2006, p. 364.

Latent Variable Models[2]

Latent Variables[a]

[a] Bishop 2006, p. 366.

"The primary role of the latent variable is to allow a complicated distribution over the observed variables be represented in terms of a model constructed from a simpler (typically exponential family) conditional distribution."

[2] Bishop 2006, p. 364.

Factor Analysis

$$
\begin{aligned}
p(\mathbf{y}_i|\boldsymbol{\theta}) &= \int p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta})p(\mathbf{x}_i)\mathrm{d}\mathbf{x}_i = \int \mathcal{N}(\mathbf{W}\mathbf{x}_i + \boldsymbol{\mu}, \boldsymbol{\Psi})\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\
&= \mathcal{N}(\mathbf{W}\boldsymbol{\mu}_0 + \boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\boldsymbol{\Sigma}_0\mathbf{W}^{\mathrm{T}}) \\
&= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^{\mathrm{T}}) \qquad (3)
\end{aligned}
$$

- $\mathbf{X}$ and $\mathbf{W}$ are related
- Integrate out $\mathbf{X}$
  - pick $\boldsymbol{\mu}_0 = 0$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$
- Low dimensional density model of $\mathbf{Y}$
  - rank of $\mathbf{W}\mathbf{W}^{\mathrm{T}}$ dimensionality of $\mathbf{X}$

Factor Analysis

$$\tilde{\mathbf{W}} = \mathbf{W}\mathbf{R} \tag{4}$$

$$p(\mathbf{y}_i|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\mathbf{R}\mathbf{R}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}) \tag{5}$$

$$= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^{\mathrm{T}}) \tag{6}$$

(7)

Identifiability

- The marginal likelihood is invariant to a rotation
  - no unique solution
  - model is the same but interpretation tricky

Factor Analysis

$$\mathbf{W}_{ML} = \text{argmax}_{\mathbf{W}} p(\mathbf{Y}|\boldsymbol{\theta}) \tag{8}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \tag{9}$$

$$\mathbf{W}_{ML} = \mathbf{U}_q (\Lambda - \sigma^2 \mathbf{I})^{\frac{1}{2}} \tag{10}$$

$$\mathbf{S} = \mathbf{U} \Lambda \mathbf{U}^{\mathrm{T}} \tag{11}$$

Probabilistic PCA

- Dimensions of $\mathbf{Y}$ independent given $\mathbf{X}$
  - $\mathbf{W}$ orthogonal matrix $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}$

Factor Analysis

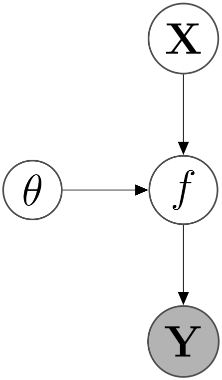
Summary

- Factor Analysis is a linear continous latent variable model
- Solution not unique
- PCA is Factor Analysis with two assumptions
  - factor loadings orthogonal $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}$
  - noise free case $\epsilon = \lim_{\sigma^2 \to 0} \sigma^2 \mathbf{I}$

Gaussian Process Latent Variable Models

GP-LVM

- General co-variance function (Ex. SE)
- $\mathbf{X}$ appears non-linearly in relation to $\mathbf{Y}$
- Marginalisation of $\mathbf{X}$ intractable



Gaussian Process Latent Variable Models

$$\text{argmax}_{\mathbf{X},\theta} p(\mathbf{Y}|\mathbf{X},\theta)p(\mathbf{X}) \tag{12}$$

$$p(\mathbf{Y}|\mathbf{X},\theta) = \int p(\mathbf{Y}|\mathbf{f})p(\mathbf{f}|\mathbf{X},\theta)\mathrm{d}\mathbf{f} \tag{13}$$

$$p(\mathbf{X}) = \mathcal{N}(\mathbf{0},\mathbf{I}) \tag{14}$$

- $\mathbf{GP}$-prior sufficiently regularises objective
- Need to set dimensionality of $\mathbf{X}$

Introduction

Recap

Approximative Inference

Being Bayesian

- You know nothing for certain
- All parameters should have prior distributions
- Decisions should be made from posterior
- Posterior is reached through Bayes Rule

Being Bayesian

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \tag{15}$$

$$p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{16}$$

- posterior distribution requires us to compute evidence $p(\mathbf{y})$
- integral often intractable
- approximate computation

Being Bayesian

$$\hat{\boldsymbol{\theta}} = \text{argmax} p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{17}$$

- Maximum a Posteriori estimation (MAP)
- point estimate as mode of posterior
- simple
- but does not communicate uncertainty well

Outline

- Stochastic Approximations
  - Sampling
  - Exact if infinite computational resources
  - Hedvig will do this
- Deterministic Approximations
  - Analytical approximations of posterior
  - Laplace approximation or mode matching
  - Variational Bayes

Laplace Approximation[3]

- MAP finds the mode of the posterior
- Does not describe the region around the mode particularly well because there is no averaging happening
- Laplace Approximation,
  - use the MAP estimate
  - approximate the posterior with an analytic form around its mode

[3] Bishop 2006, pp. 213-216,

Laplace Approximation[3]

Algorithm

1. Approximate posterior with Gaussian
2. Find MAP estimate
3. Make Taylor Expansion around mode

[3] Bishop 2006, pp. 213-216,

Laplace Approximation[3]

$$t(\boldsymbol{\theta}) = \log\left(p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right) = \log\left(p(\mathbf{y}|\boldsymbol{\theta})\right) + \log\left(p(\boldsymbol{\theta})\right) \tag{18}$$

$$= \sum_{i=1}^{N} \log\left(p(\mathbf{y}_i|\boldsymbol{\theta})\right) + \log\left(p(\boldsymbol{\theta})\right) \tag{19}$$

[3] Bishop 2006, pp. 213-216,

Laplace Approximation[3]

$$t(\boldsymbol{\theta}) = \log\left(p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right) = \log\left(p(\mathbf{y}|\boldsymbol{\theta})\right) + \log\left(p(\boldsymbol{\theta})\right) \tag{20}$$

$$= \sum_{i=1}^{N} \log\left(p(\mathbf{y}_i|\boldsymbol{\theta})\right) + \log\left(p(\boldsymbol{\theta})\right) \tag{21}$$

Make Taylor Expansion around MAP parameter estimate $\hat{\boldsymbol{\theta}}$

$$t(\boldsymbol{\theta}) = t(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathrm{T}} \frac{\delta t(\boldsymbol{\theta})}{\delta \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} + \frac{1}{2!}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathrm{T}} \frac{\delta^2 t(\boldsymbol{\theta})}{\delta^2 \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \ldots$$

$$\approx t(\hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathrm{T}} H(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \tag{22}$$

[3] Bishop 2006, pp. 213-216,

Laplace Approximation[3]

Make Taylor Expansion around MAP parameter estimate $\hat{\boldsymbol{\theta}}$

$$
\begin{aligned}
t(\boldsymbol{\theta}) &= t(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathrm{T}} \frac{\delta t(\boldsymbol{\theta})}{\delta \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} + \frac{1}{2!}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathrm{T}} \frac{\delta^2 t(\boldsymbol{\theta})}{\delta^2 \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \ldots \\
&\approx t(\hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathrm{T}} H(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \qquad (23)
\end{aligned}
$$

Linear term disappears as $\hat{\boldsymbol{\theta}}$ is MAP estimate

$$
H(\hat{\boldsymbol{\theta}}) = \frac{\delta^2 \log p(\boldsymbol{\theta}|\mathbf{y})}{\delta^2 \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \qquad (24)
$$

Hessian is the second derivatives of the true posterior

[3] Bishop 2006, pp. 213-216,

Laplace Approximation[3]

Compute evidence using approximation

$$\log p(\mathbf{y}) = \log \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \log \int e^{t(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta} \tag{25}$$

$$\approx t(\hat{\boldsymbol{\theta}}) + \frac{1}{2}|2\pi H^{-1}| \tag{26}$$

$$= \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}) + \log p(\hat{\boldsymbol{\theta}}) + \frac{d}{2}\log 2\pi - \frac{1}{2}\log|H| \tag{27}$$

[3] Bishop 2006, pp. 213-216,

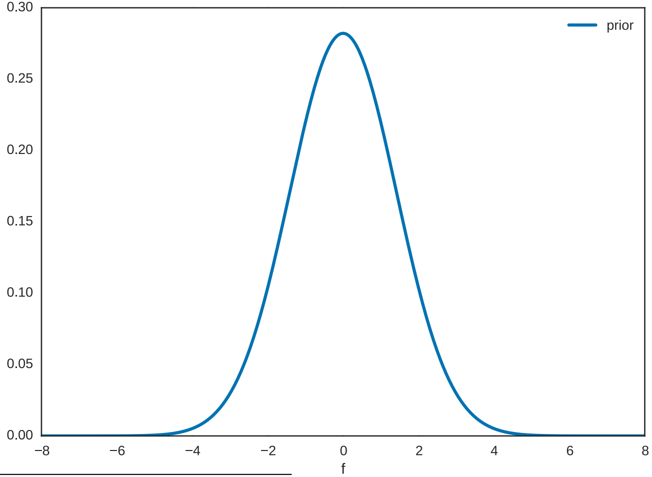
Laplace Approximation[3]

$$p(\mathbf{y}) \approx p(\mathbf{y}|\hat{\boldsymbol{\theta}})p(\hat{\boldsymbol{\theta}})|2\pi H^{-1}|^{\frac{1}{2}} \tag{28}$$

- Likelihood at MAP point
- Penalty term of prior
- Hessian takes into account local curvature around MAP point
- Gaussian approximation to the posterior

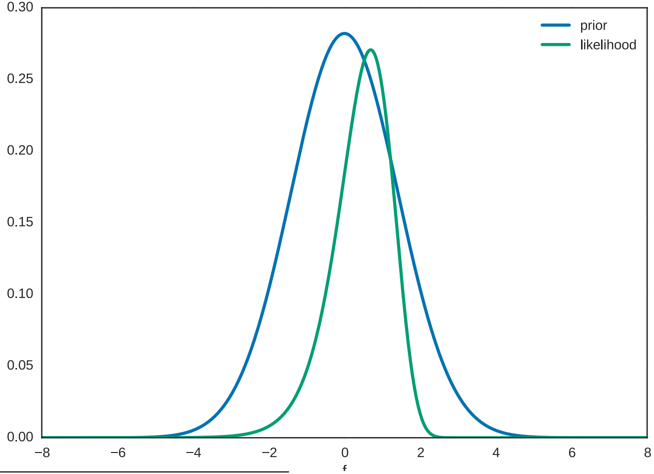
[3] Bishop 2006, pp. 213-216,

Laplace Approximation[3]



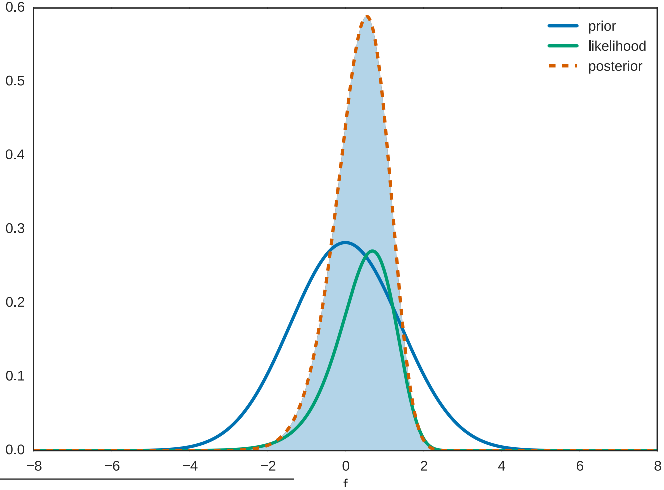
[3] Bishop 2006, pp. 213-216,Images courtesy by Alan Saul

Laplace Approximation[3]



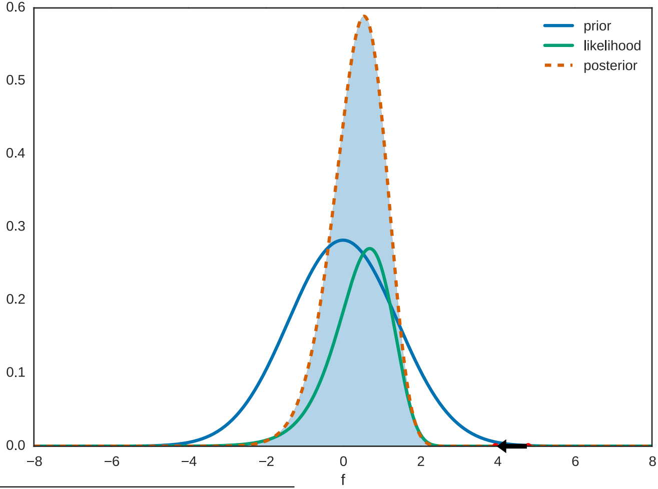
[3] Bishop 2006, pp. 213-216,Images courtesy by Alan Saul

Laplace Approximation[3]



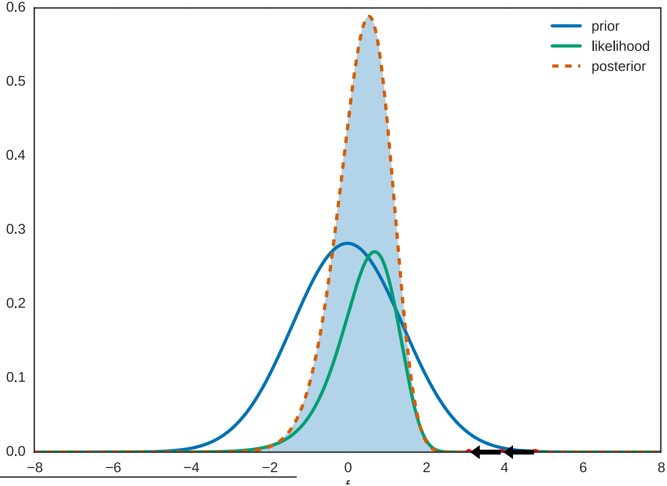
[3]Bishop 2006, pp. 213-216,Images courtesy by Alan Saul

Laplace Approximation[3]



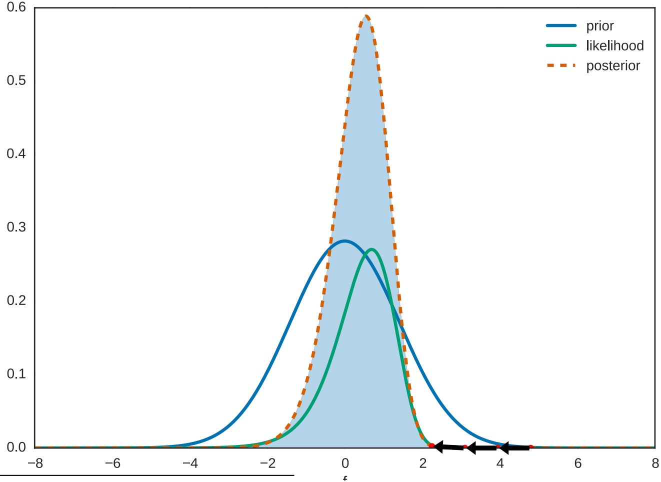
[3] Bishop 2006, pp. 213-216,Images courtesy by Alan Saul

Laplace Approximation[3]



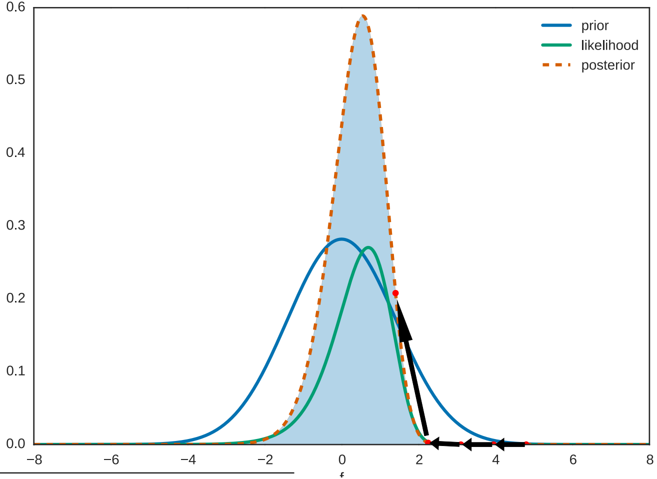
[3]Bishop 2006, pp. 213-216,Images courtesy by Alan Saul

Laplace Approximation[3]



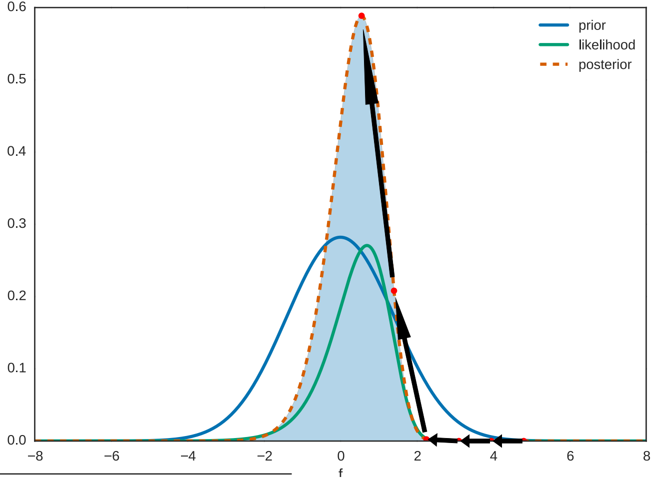
[3] Bishop 2006, pp. 213-216,Images courtesy by Alan Saul

Laplace Approximation[3]



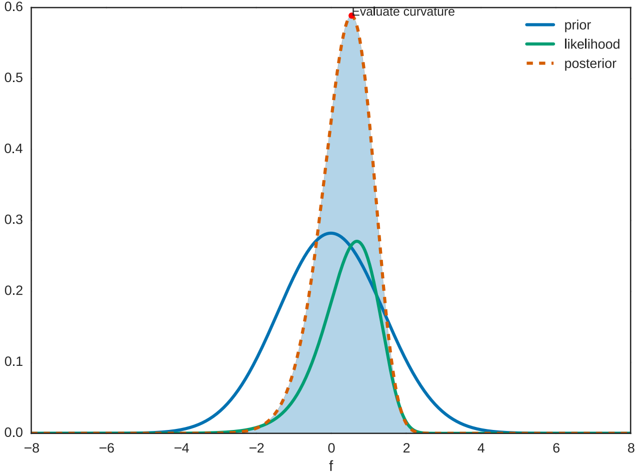
[3] Bishop 2006, pp. 213-216,Images courtesy by Alan Saul

Laplace Approximation[3]



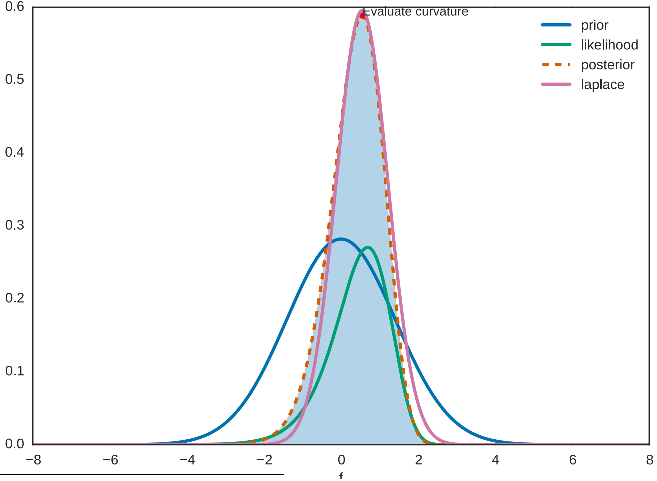
[3]Bishop 2006, pp. 213-216,Images courtesy by Alan Saul

Laplace Approximation[3]



[3]Bishop 2006, pp. 213-216,Images courtesy by Alan Saul

Laplace Approximation[3]



[3] Bishop 2006, pp. 213-216,Images courtesy by Alan Saul

Laplace Approximation[3]

- Good
  - Easy (assuming we can compute second derivatives)
  - When we know that posterior is well characterized by its mode
- Bad
  - Local
  - Hard to know how good the approximation is

[3] Bishop 2006, pp. 213-216,

Variational Bayes[4]

- Want to reach posterior $p(\boldsymbol{\theta}|\mathbf{y})$
- Approximate true posterior with simple distribution $q(\boldsymbol{\theta})$
  - $q(\boldsymbol{\theta})$ has a set of parameters
- Formulate optimisation problem
- What is the objective function?

[4] Bishop 2006, pp. 462-470

Variational Bayes[4]

KL-divergence

$$\mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})) = \int q(\boldsymbol{\theta})\log\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})}\mathrm{d}\boldsymbol{\theta} \tag{29}$$

- Always positive
- Non-symmetric
- Part of a class of functionals called α-divergence
  - encapsulate several different inference algorithms

[4] Bishop 2006, pp. 462-470

Variational Bayes[4]

$$
\begin{aligned}
\mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})) &= \int q(\boldsymbol{\theta})\log\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})}\mathrm{d}\boldsymbol{\theta} = \left\{p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y},\boldsymbol{\theta})}{p(\mathbf{y})}\right\} \\
&= \int q(\boldsymbol{\theta})\log\frac{q(\boldsymbol{\theta})p(\mathbf{y})}{p(\mathbf{y},\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta} = \\
&= \int q(\boldsymbol{\theta})\log\frac{q(\boldsymbol{\theta})}{p(\mathbf{y},\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta} + \underbrace{\int q(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}_{=1}\log p(\mathbf{y}) \\
&= \int q(\boldsymbol{\theta})\log\frac{q(\boldsymbol{\theta})}{p(\mathbf{y},\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta} + \log p(\mathbf{y}) \qquad (30)
\end{aligned}
$$

[4] Bishop 2006, pp. 462-470

Variational Bayes[4]

$$\mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})) = \int q(\boldsymbol{\theta})\log\frac{q(\boldsymbol{\theta})}{p(\mathbf{y},\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta} + \log p(\mathbf{y}) \tag{31}$$

$$\Rightarrow \log p(\mathbf{y}) = \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})) + \underbrace{\int q(\boldsymbol{\theta})\log\frac{p(\mathbf{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta}}_{\mathcal{L}(\boldsymbol{\theta})} \tag{32}$$

- KL-divergence always positive
- $\mathcal{L}(\boldsymbol{\theta})$ lower bound on $\log p(\mathbf{y})$

[4]Bishop 2006, pp. 462-470

Variational Bayes[4]

Lower bound,

$$\mathcal{L}(\boldsymbol{\theta}) = \int q(\boldsymbol{\theta}) \log \frac{p(\mathbf{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} \tag{33}$$

$$= \int q(\boldsymbol{\theta}) \log p(\mathbf{y}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} - \int q(\boldsymbol{\theta}) \log q(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$$

$$\mathbb{E}_{q(\boldsymbol{\theta})} \left[\log p(\mathbf{y}, \boldsymbol{\theta}) \right] = \int q(\boldsymbol{\theta}) \log p(\mathbf{y}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \tag{34}$$

$$\mathrm{H}(q(\boldsymbol{\theta})) = \int q(\boldsymbol{\theta}) \log q(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \tag{35}$$

- Maximise expected value of joint distribution with as "informative" distribution as possible

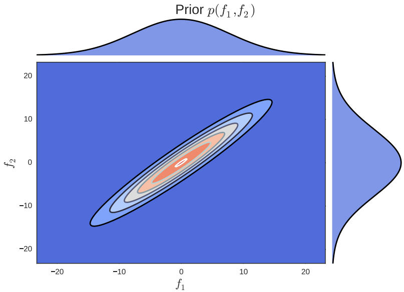
[4] Bishop 2006, pp. 462-470

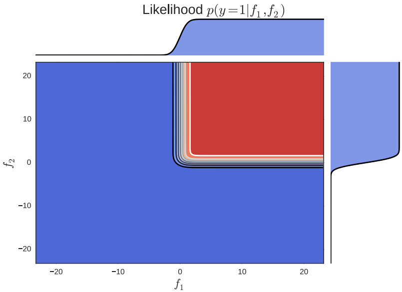
Variational Bayes[4]

- Good
  - fast
  - principled, we know how well bad we are doing
- Bad
  - how to design proposal distribution
- KL-divergence is not symmetric if we instead minimise $\text{KL}(p||q)$ we get Expectation Propagation

[4]Bishop 2006, pp. 462-470

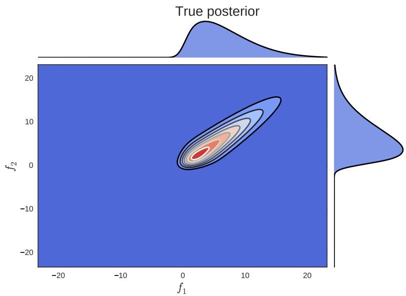
Comparison between approximations[5]

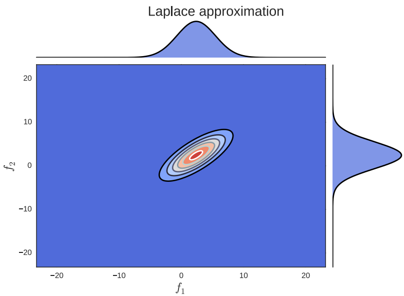




[5]Images courtesy by Alan Saul

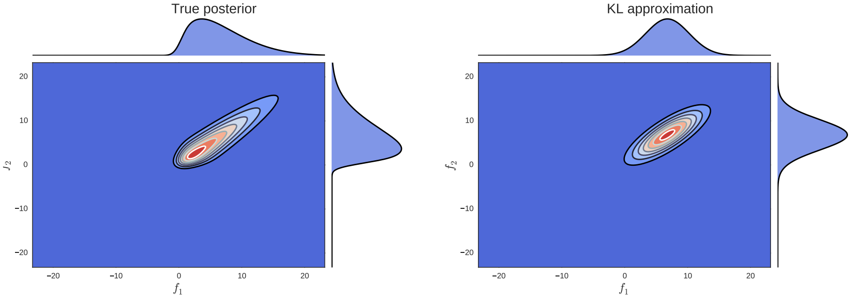
Comparison between approximations[5]





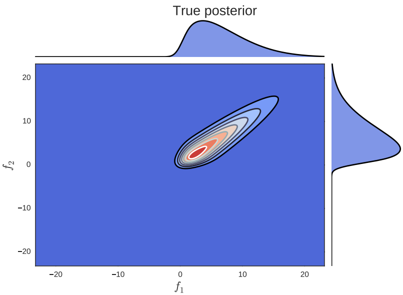
[5] Images courtesy by Alan Saul

Comparison between approximations[5]



[5] Images courtesy by Alan Saul

Comparison between approximations[5]



[5]Images courtesy by Alan Saul

Summary

- A Bayesian makes decision from the posterior distribution
  - combines data and belief
- computing posterior often requires intractable integrals
- to proceed we use approximations
- now you got a flavour of how this is done
- This is current very active research

Summary

- A Bayesian makes decision from the posterior distribution
  - combines data and belief
- computing posterior often requires intractable integrals
- to proceed we use approximations
- now you got a flavour of how this is done
- This is current very active research

End of Part 1

- 4 lectures
- Dealing with Uncertainty is the key
- Probabilities are useful means for representing certainty
  - probabilistic objects
- Priors are the key to learning
  - two different examples
  - parametrisation
- What is the tasks we need to solve
- I have been very abstract on purpose to focus on understanding learning

"It's true there's been a lot of work on trying to apply statistical models to various linguistic problems. I think there have been some successes, but a lot of failures. There is a notion of success which I think is novel in the history of science. It interprets success as approximating unanalyzed data."

[Noam Chomsky]

What do you need to do?

- Translate to your own problems/data
- How have you solved problems before, think of the assumptions you made
- What are sensible priors/likelihoods/structures
- What assumptions can you make?
- Don't be afraid of being abstract, when you get too close to the problem you often make assumptions that you are not aware of
- Get your hands dirty, i.e. develop your own priors for developing models
- Form your own opinion and disagree with me

What do you need to do?

- Translate to your own problems/data
- How have you solved problems before, think of the assumptions you made
- What are sensible priors/likelihoods/structures
- What assumptions can you make?
- Don't be afraid of being abstract, when you get too close to the problem you often make assumptions that you are not aware of
- Get your hands dirty, i.e. develop your own priors for developing models
- Form your own opinion and disagree with me

What do you need to do?

- Translate to your own problems/data
- How have you solved problems before, think of the assumptions you made
- What are sensible priors/likelihoods/structures
- What assumptions can you make?
- Don't be afraid of being abstract, when you get too close to the problem you often make assumptions that you are not aware of
- Get your hands dirty, i.e. develop your own priors for developing models
- Form your own opinion and disagree with me

Take home message

- Machine learning is really simple, it should be as even Carl have learnt quite a few things in life
- Formulating learning so that it can be externalised might be very hard and really involved but that is just labour
  - you can always find someone to do this work
- Make assumptions, lots of them, that is the basis of learning, but be aware of them

Take home message

- Machine learning is really simple, it should be as even Carl have learnt quite a few things in life
- Formulating learning so that it can be externalised might be very hard and really involved but that is just labour
  - you can always find someone to do this work
- Make assumptions, lots of them, that is the basis of learning, but be aware of them

Take home message

- Machine learning is really simple, it should be as even Carl have learnt quite a few things in life
- Formulating learning so that it can be externalised might be very hard and really involved but that is just labour
  - you can always find someone to do this work
- Make assumptions, lots of them, that is the basis of learning, but be aware of them

Take home message

- Machine learning is really simple, it should be as even Carl have learnt quite a few things in life
- Formulating learning so that it can be externalised might be very hard and really involved but that is just labour
  - you can always find someone to do this work
- Make assumptions, lots of them, that is the basis of learning, but be aware of them

Carls rant

Carls rant

- Intelligence is the ability to reason under uncertainty
- If we have absolute knowledge we do not need intelligence reasoning(Laplace)
- Absolute knowledge - all data
- which interesting problems do we have that for?
- no priors (or not formulated priors) makes us headless chickens we become too naïve
- when we need a lot of data to solve a simple problem you should be worried
- ML is a young field, the important thing is not that it can but why it can, otherwise development will stop

Carls rant

- Intelligence is the ability to reason under uncertainty
- If we have absolute knowledge we do not need intelligence reasoning(Laplace)
- Absolute knowledge - all data
- which interesting problems do we have that for?
- no priors (or not formulated priors) makes us headless chickens we become too naïve
- when we need a lot of data to solve a simple problem you should be worried
- ML is a young field, the important thing is not that it can but why it can, otherwise development will stop

Carls rant

- Intelligence is the ability to reason under uncertainty
- If we have absolute knowledge we do not need intelligence reasoning(Laplace)
- Absolute knowledge - all data
- which interesting problems do we have that for?
- no priors (or not formulated priors) makes us headless chickens we become too naïve
- when we need a lot of data to solve a simple problem you should be worried
- ML is a young field, the important thing is not that it can but why it can, otherwise development will stop

Carls rant

- Intelligence is the ability to reason under uncertainty
- If we have absolute knowledge we do not need intelligence reasoning(Laplace)
- Absolute knowledge - all data
- which interesting problems do we have that for?
- no priors (or not formulated priors) makes us headless chickens we become too naïve
- when we need a lot of data to solve a simple problem you should be worried
- ML is a young field, the important thing is not that it can but why it can, otherwise development will stop

Carls rant

- Intelligence is the ability to reason under uncertainty
- If we have absolute knowledge we do not need intelligence reasoning(Laplace)
- Absolute knowledge - all data
- which interesting problems do we have that for?
- no priors (or not formulated priors) makes us headless chickens we become too naïve
- when we need a lot of data to solve a simple problem you should be worried
- ML is a young field, the important thing is not that it can but why it can, otherwise development will stop

Carls rant

- Intelligence is the ability to reason under uncertainty
- If we have absolute knowledge we do not need intelligence reasoning(Laplace)
- Absolute knowledge - all data
- which interesting problems do we have that for?
- no priors (or not formulated priors) makes us headless chickens we become too naïve
- when we need a lot of data to solve a simple problem you should be worried
- ML is a young field, the important thing is not that it can but why it can, otherwise development will stop

Carls rant

- Intelligence is the ability to reason under uncertainty
- If we have absolute knowledge we do not need intelligence reasoning(Laplace)
- Absolute knowledge - all data
- which interesting problems do we have that for?
- no priors (or not formulated priors) makes us headless chickens we become too naïve
- when we need a lot of data to solve a simple problem you should be worried
- ML is a young field, the important thing is not that it can but why it can, otherwise development will stop

Next Time

Lecture 4

- November 13th 15-19 V1
- Variational Bayes example
- Help session for assignment
  - anything else that you want me to do
  - within the realm of me keeping some decency

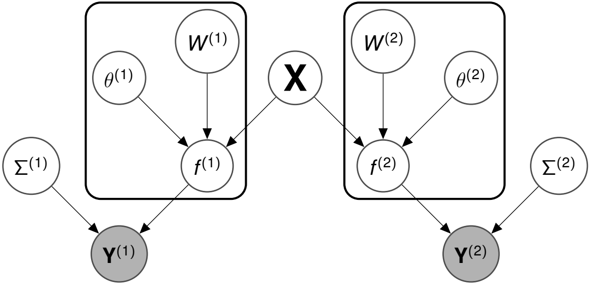
Next Time

Lecture 4

- November 13th 15-19 V1
- Variational Bayes example
- Help session for assignment
  - anything else that you want me to do
  - within the realm of me keeping some decency

e.o.f.

My Research



$W^{(1)}$
$\theta^{(1)}$
$\mathbf{X}$
$W^{(2)}$
$\theta^{(2)}$
$\Sigma^{(1)}$
$f^{(1)}$
$f^{(2)}$
$\Sigma^{(2)}$
$\mathbf{Y}^{(1)}$
$\mathbf{Y}^{(2)}$

References I

Christopher M Bishop. *Pattern recognition and machine learning*. 2006. URL: http://www.library.wisc.edu/selectedtocs/bg0137.pdf.