CS-E4895 Gaussian Processes
Lecture 6: Classification

Arno Solin

Aalto University

Tuesday 14.3.2023

based on slides by ST John and Michael Riis Andersen

Roadmap for today

1. Beyond Gaussian noise
   - Classification vs. regression
   - Other likelihoods
   - Looking at the likelihood in more detail
2. Inference for arbitrary likelihoods
   - Posterior predictive distribution
   - Why is the posterior intractable?
3. Approximating the intractable
   - Gaussian approximations
   - Laplace approximation
   - Minimising divergences
   - Variational inference
4. Conclusion

Beyond Gaussian noise

Regression vs. classification

- Response variable $\boldsymbol{y}$ is continuous in regression problems

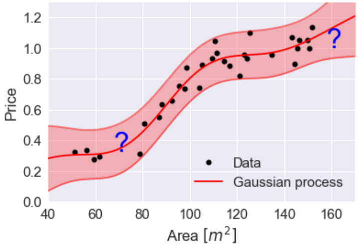
$$y_n \in \mathbb{R}$$

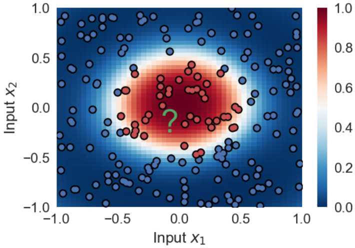
- Response variable $\boldsymbol{y}$ is discrete in classification problems

$$y_n \in \{c_1, c_2, \ldots, c_K\}$$

- Classification problems

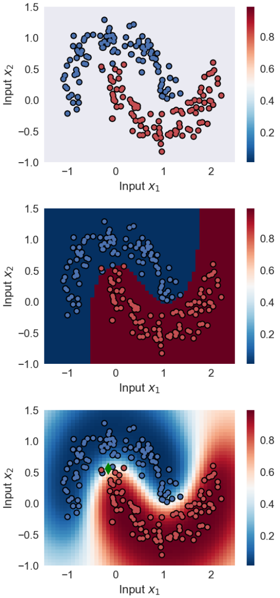
$$\begin{aligned}
\boldsymbol{X} &= \text{images}, & y_n &\in \{\text{cat}, \text{dog}\} \\
\boldsymbol{X} &= \text{X-ray scan}, & y_n &\in \{\text{tumor}, \text{no tumor}\} \\
\boldsymbol{X} &= \text{images of digits}, & y_n &\in \{0, 1, 2, \ldots, 9\} \\
\boldsymbol{X} &= \text{emails}, & y_n &\in \{\text{spam}, \text{not spam}\}
\end{aligned}$$





Why Gaussian processes for classification?

- Complex decision boundaries
  1. Non-linear boundary
  2. Can learn complexity of decision boundary from data
- Probabilistic classification
  1. How would you classify the green point?
  2. We want to model the uncertainty



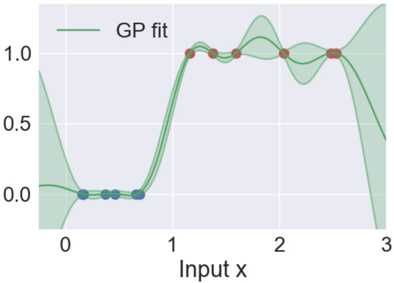
Why don't we use regression models for classification?

- We focus on binary classification: $y_n \in \{0, 1\}$ or $y_n \in \{-1, 1\}$
- Given a data set $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$, we want to model

$$p(y_n = +1 \mid \boldsymbol{x}_n)$$

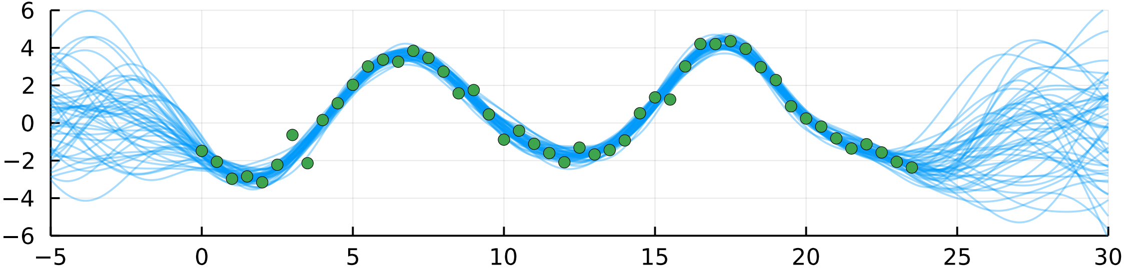
- What's wrong with simply using the GP regression model with labels: $y_n \in \{0, 1\}$:

$$p(y_n = +1 \mid \boldsymbol{x}_n) = f(\boldsymbol{x}_n)$$



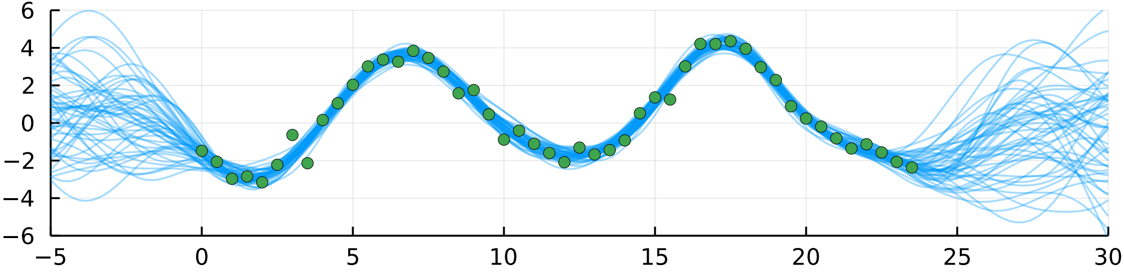
Recap: Gaussian noise model

$$y(x) = f(x) + \epsilon, \qquad \epsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_{\text{noise}})$$



Recap: Gaussian noise model

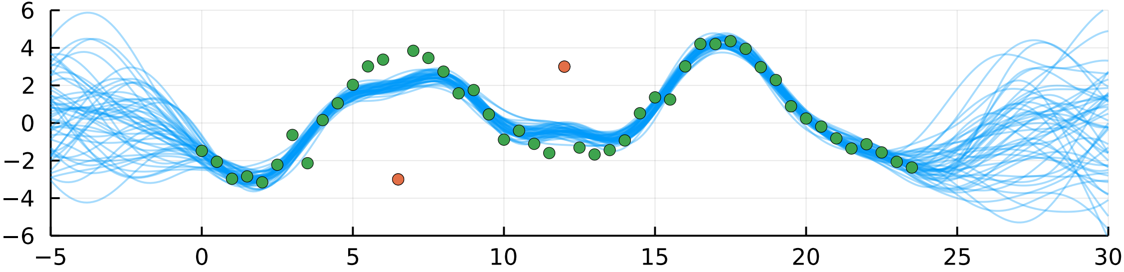
$$y(x) = f(x) + \epsilon, \qquad \epsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_{\text{noise}})$$
$$p(y \mid f) = \mathcal{N}(y \mid f, \sigma^2_{\text{noise}})$$



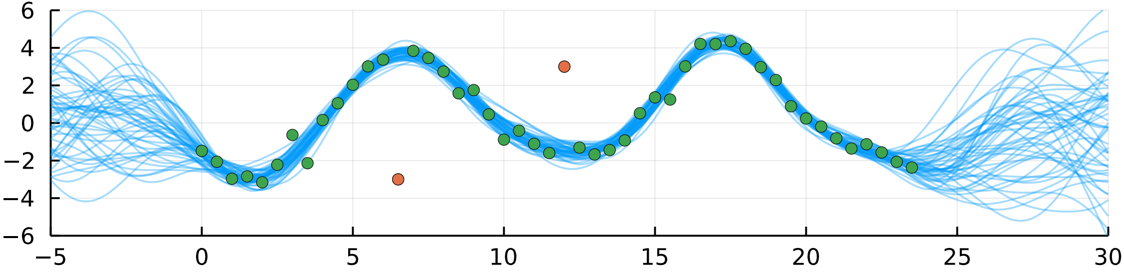
Misspecified Gaussian noise model

$$y(x) = f(x) + \epsilon, \qquad \epsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_{\text{noise}})$$

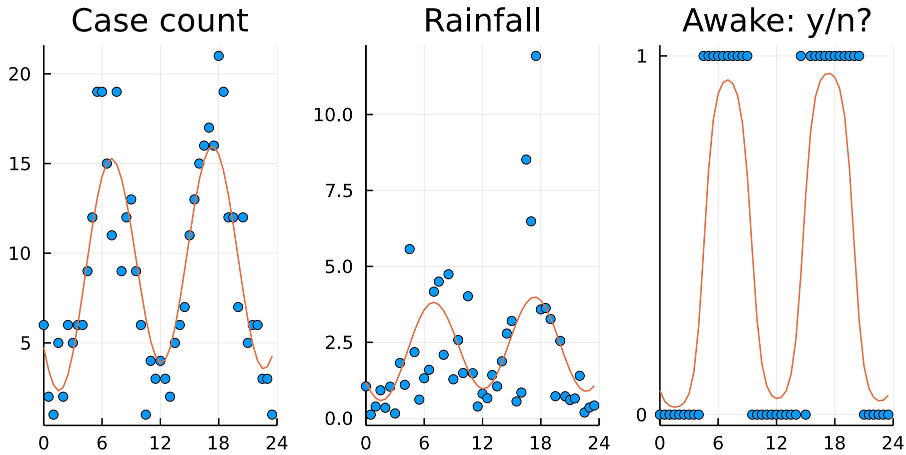
$$p(y \mid f) = \mathcal{N}(y \mid f, \sigma^2_{\text{noise}})$$



Heavy-tailed noise model



Non-Gaussian observations



Latent functional relationship

$$p\big(y_n \,|\, f(x_n)\big)$$

Likelihood

$$p(\mathbf{y} \mid \mathbf{f}) = \prod_{n=1}^{N} p(y_n \mid f_n); \qquad f_n = f(x_n)$$

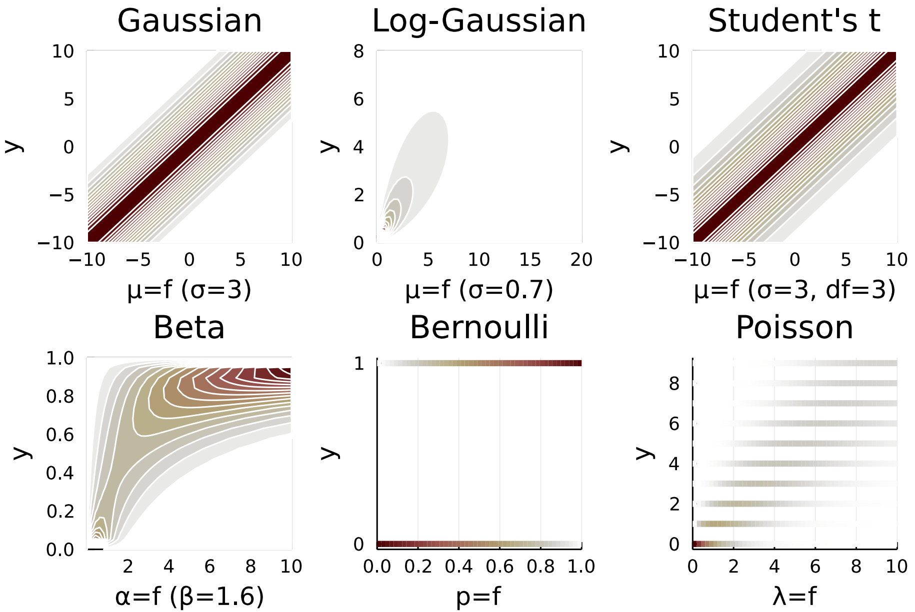
factorizing

$$p(y \mid f)$$

Function of two arguments:

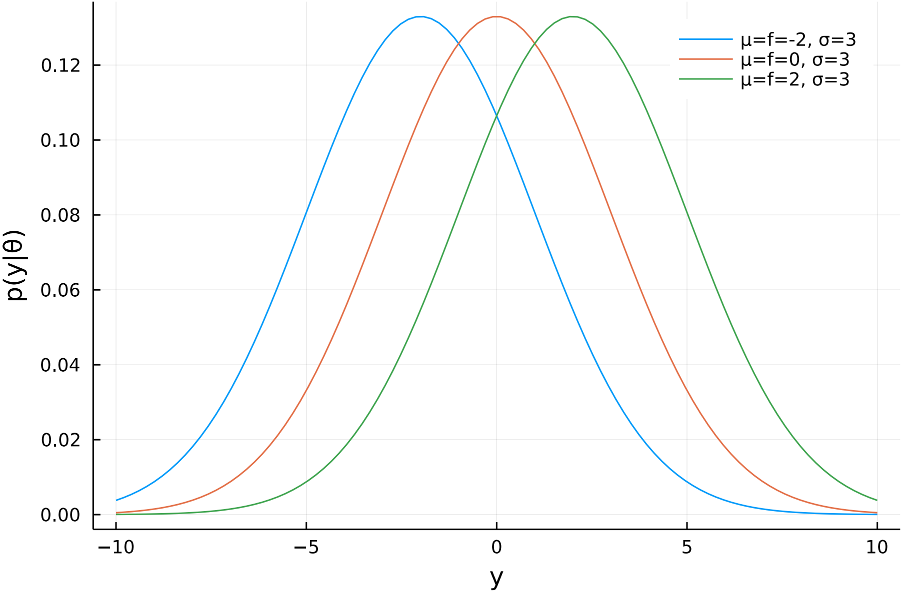
$$y \mapsto p(y \mid f), \qquad f \mapsto p(y \mid f)$$

$p(y \mid f)$

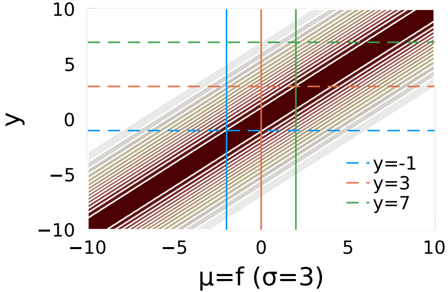


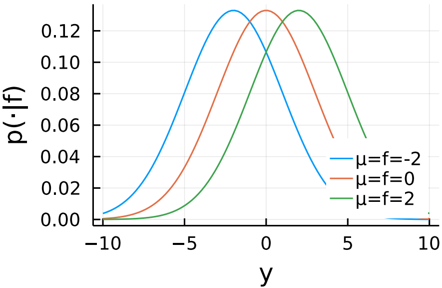
Gaussian
Log-Gaussian
Student's t
10
5
0
−5
−10
y
−10 −5 0 5 10
μ=f (σ=3)
8
6
4
2
0
0 5 10 15 20
μ=f (σ=0.7)
μ=f (σ=3, df=3)
Beta
Bernoulli
Poisson
1.0
0.8
0.6
0.4
0.2
0.0
2 4 6 8 10
α=f (β=1.6)
1
0
0.0 0.2 0.4 0.6 0.8 1.0
p=f
8
6
4
2
0
0 2 4 6 8 10
λ=f

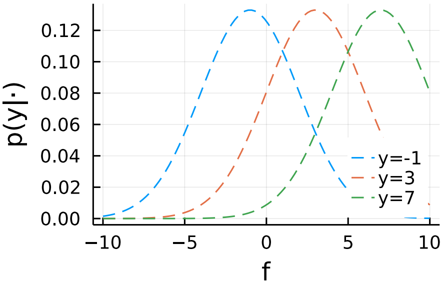
$p(y \mid f)$: Gaussian



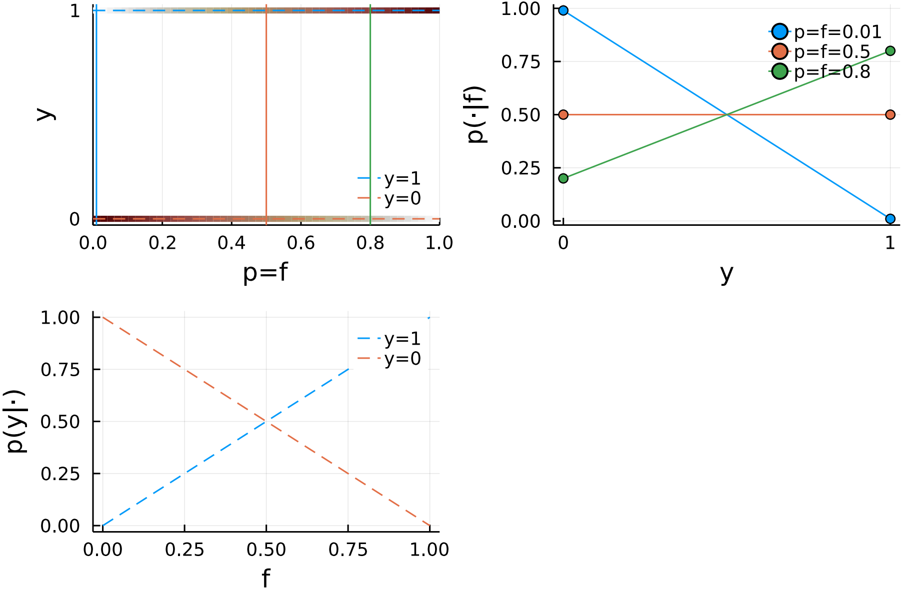
$p(y \mid f)$: Gaussian



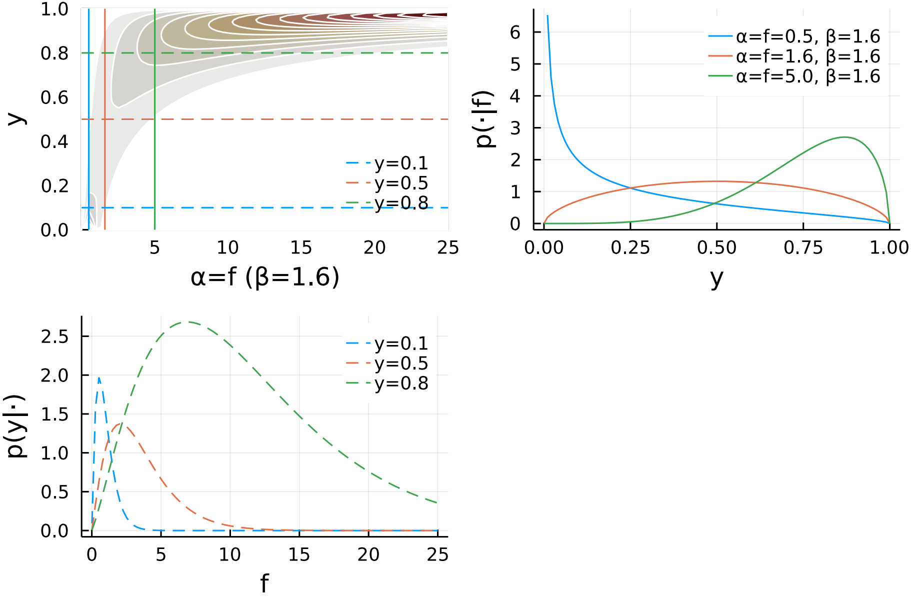




$p(y \mid f)$: Bernoulli

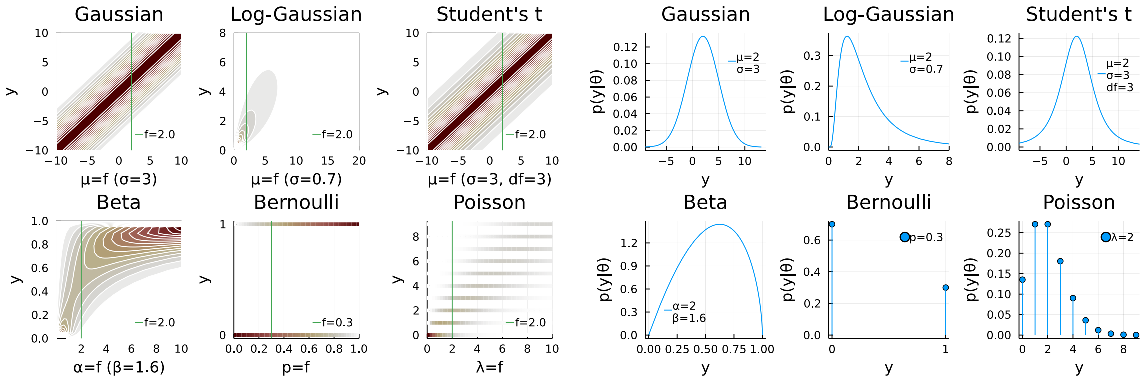


$p(y \mid f)$: Beta

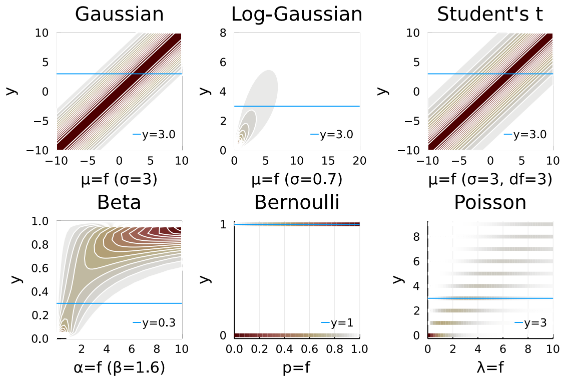


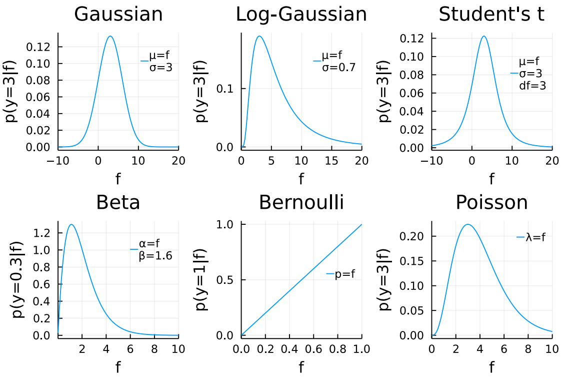
1.0
0.8
0.6
0.4
0.2
0.0
y
5
10
15
20
25
α=f (β=1.6)
y=0.1
y=0.5
y=0.8
6
5
4
3
2
1
0
p(·|f)
0.00
0.25
0.50
0.75
1.00
y
α=f=0.5, β=1.6
α=f=1.6, β=1.6
α=f=5.0, β=1.6
2.5
2.0
1.5
1.0
0.5
0.0
p(y|·)
0
5
10
15
20
25
f
y=0.1
y=0.5
y=0.8

$p(y \mid f)$: distribution of observation $\quad f$ fixed



$p(y \mid f)$: likelihood of parameter y fixed





$p(y \mid f)$: likelihood of parameter $\qquad y$ fixed

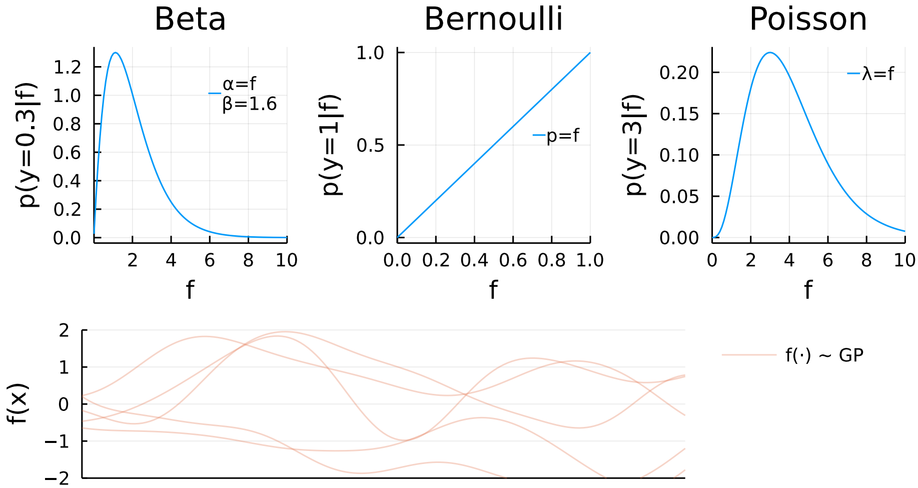
Two important aspects of likelihoods:

1. link functions
2. log-concavity

Link functions

$$\mathbb{E}[y] = \theta \in (0 \ldots \infty)$$

$$f \sim \mathcal{N} \qquad \in (-\infty \ldots \infty)$$



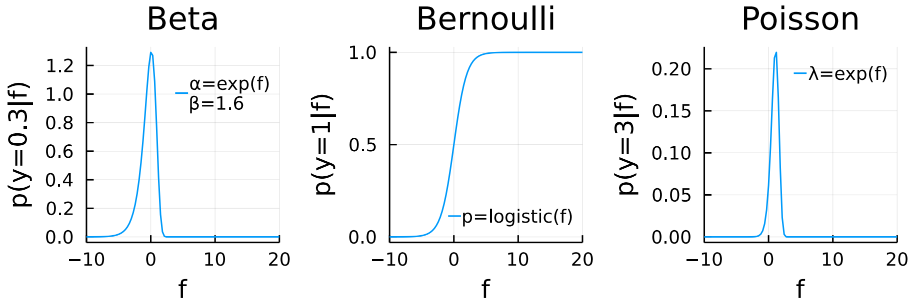
Link functions

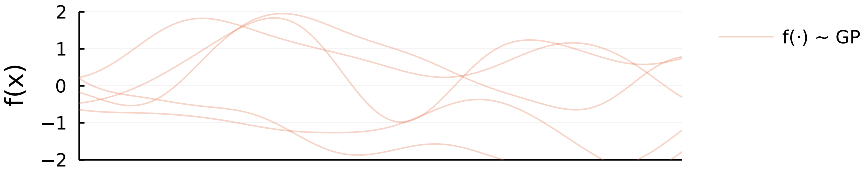
$$\mathbb{E}[y] = \theta \in (0 \ldots \infty)$$

$$f \sim \mathcal{N} \qquad \in (-\infty \ldots \infty)$$

$$\text{link}(\theta) = f$$

$$\theta = \text{invlink}(f)$$





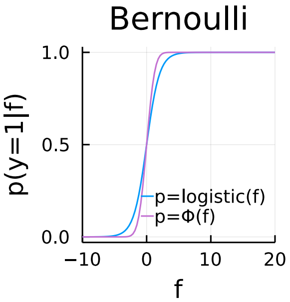
Link functions

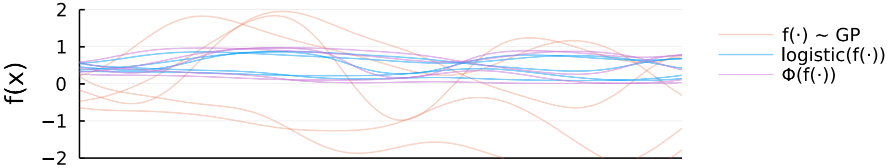
$$\mathbb{E}[y] = \theta \in (0 \ldots \infty)$$

$$f \sim \mathcal{N} \qquad \in (-\infty \ldots \infty)$$

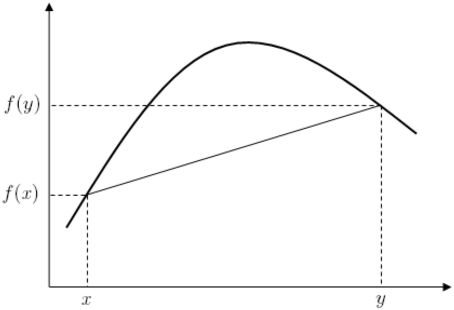
$$\text{link}(\theta) = f$$

$$\theta = \text{invlink}(f)$$



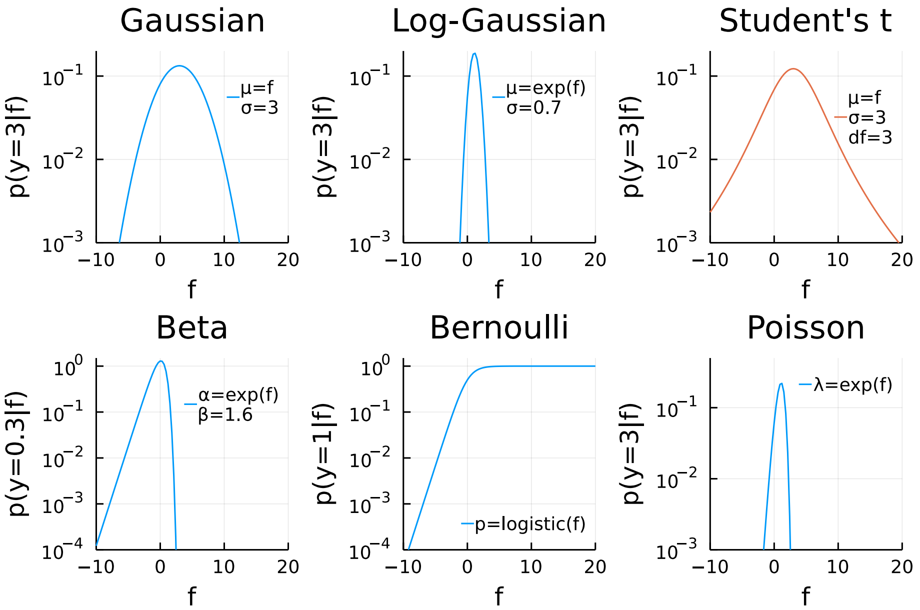


(Log-)concavity



$$f\big(\alpha x + (1-\alpha)y\big) \geq \alpha f(x) + (1-\alpha) f(y)$$

Log-concavity of likelihoods

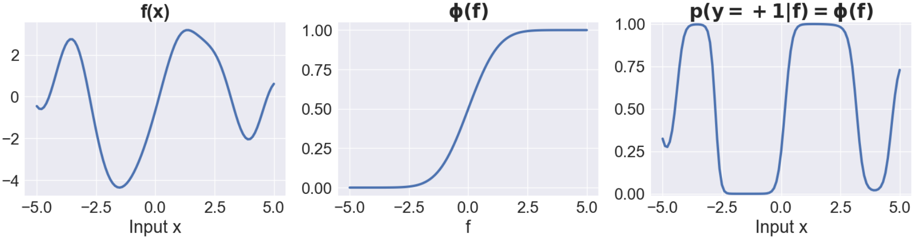


Section 2

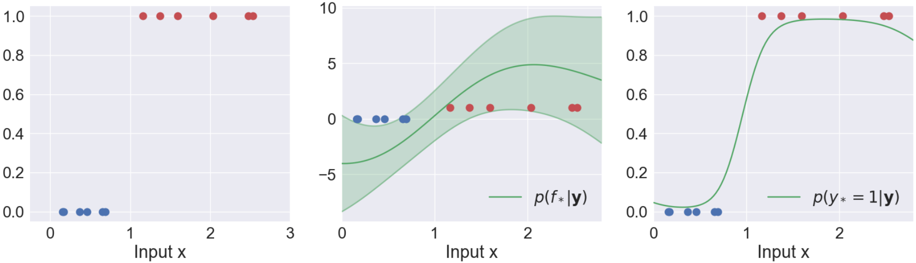
Inference for arbitrary likelihoods

1. Beyond Gaussian noise
2. Inference for arbitrary likelihoods
   - Posterior predictive distribution
   - Why is the posterior intractable?
3. Approximating the intractable
4. Conclusion

- We map the unknown function $f(x)$ through the squashing function



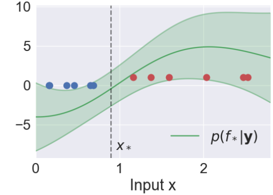
- Example re-visited



Predictive distribution at new test point $\boldsymbol{x}_*$

1. Joint model:

$$p(\boldsymbol{y}, \boldsymbol{f}) = p(\boldsymbol{y} \mid \boldsymbol{f}) p(\boldsymbol{f}) = \prod_{n=1}^{N} p(y_n \mid f_n) \mathcal{N}(\boldsymbol{f} \mid \boldsymbol{0}, \boldsymbol{K})$$

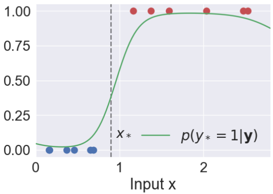


2. Posterior distribution at training points:

$$p(\boldsymbol{f} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{f}) p(\boldsymbol{f})}{p(\boldsymbol{y})} \approx q(\boldsymbol{f})$$

3. Posterior of f_* for new test point $\boldsymbol{x}_*$:

$$p(f_* \mid \boldsymbol{y}) = \int p(f_* \mid \boldsymbol{f}) \, p(\boldsymbol{f} \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{f} \approx \int p(f_* \mid \boldsymbol{f}) \, q(\boldsymbol{f}) \, \mathrm{d}\boldsymbol{f}$$



4. Predictive distribution

$$p(y_* \mid \boldsymbol{y}) = \int p(y_* \mid f_*) p(f_* \mid \boldsymbol{y}) \, \mathrm{d}f_*$$

Analytically intractable distributions!

Posterior predictions

At new point x^*:

$$p(f^* \,|\, x^*, \mathbf{x}, \mathbf{y}) = \int p(f^* \,|\, x^*, \mathbf{x}, \mathbf{f})\, p(\mathbf{f} \,|\, \mathbf{x}, \mathbf{y})\, \mathrm{d}\mathbf{f}$$

At training data:

$$p(\mathbf{f} \,|\, \mathbf{x}, \mathbf{y}) = \frac{p(\mathbf{f} \,|\, \mathbf{x}) \prod_{n=1}^{N} p(y_n \,|\, f(x_n))}{\int p(\mathbf{f}' \,|\, \mathbf{x}) \prod_{n=1}^{N} p(y_n \,|\, f'(x_n))\, \mathrm{d}\mathbf{f}'}$$

$$p(\mathbf{f} \,|\, \mathbf{y}) = \frac{1}{Z} p(\mathbf{f}) \prod_{n=1}^{N} p(y_n \,|\, f_n)$$

$$Z = p(\mathbf{y} \,|\, \mathcal{M}) = \int p(\mathbf{f} \,|\, \mathcal{M}) \prod_{n=1}^{N} p(y_n \,|\, f_n, \mathcal{M})\, \mathrm{d}\mathbf{f}$$

"marginal likelihood" or "evidence" given model $\mathcal{M}$

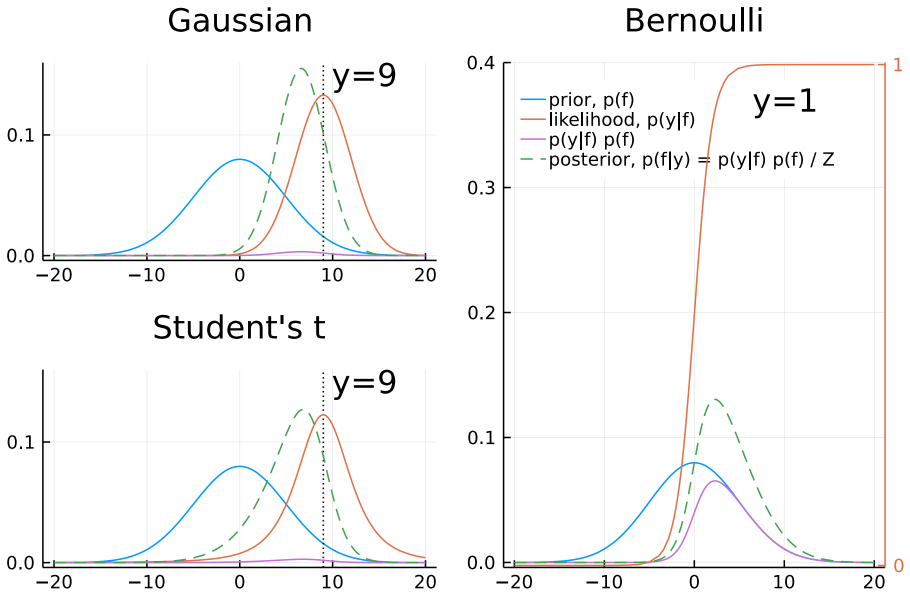
Posterior at training points

$$p(\mathbf{f} \mid \mathbf{y}) = \frac{1}{Z} p(\mathbf{f}) \prod_{n=1}^{N} p(y_n \mid f_n)$$

Gaussian (process) prior $p(f(\cdot))$ ...

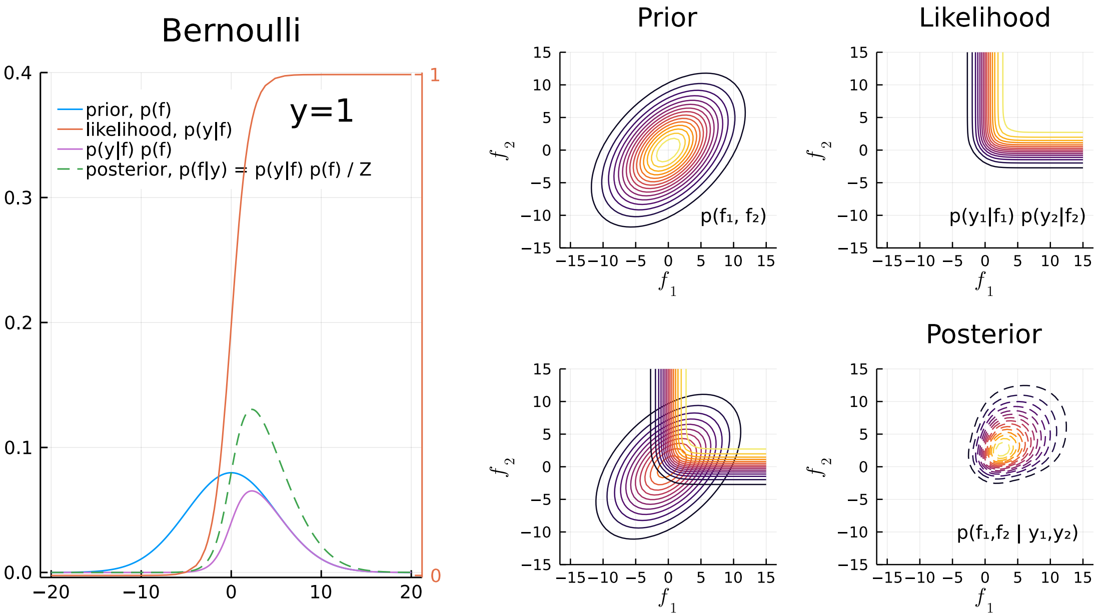
- & Gaussian likelihood: conjugate case $\rightarrow$ posterior Gaussian
- & non-Gaussian $p(y \mid f) \rightarrow p(\mathbf{f} \mid \mathbf{y})$ also non-Gaussian, intractable

1D examples



Gaussian
y=9
0.1
0.0
−20
−10
0
10
20
Student's t
y=9
0.1
0.0
−20
−10
0
10
20
Bernoulli
0.4
0.3
0.2
0.1
0.0
1
0
y=1
prior, p(f)
likelihood, p(y|f)
p(y|f) p(f)
posterior, p(f|y) = p(y|f) p(f) / Z
−20
−10
0
10
20

Bernoulli example in 2D



Posterior for N observations

$$p(\mathbf{f} \mid \mathbf{y}) = \frac{p(\mathbf{f}) \prod_{n=1}^{N} p(y_n \mid f_n)}{\int p(\mathbf{f}') \prod_{n=1}^{N} p(y_n \mid f_n') \, \mathrm{d}\mathbf{f}'}$$

$$\begin{aligned} f_1 &= f(x_1) \\ f_2 &= f(x_2) \\ &\vdots \\ f_N &= f(x_N) \end{aligned}$$

Summary so far

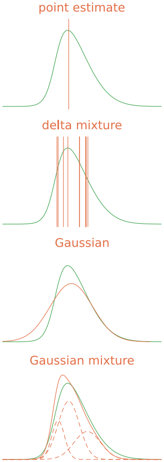
- What is the likelihood $p(y \mid f)$?
- When is it non-Gaussian?
- Why does the posterior $p(f \mid y)$ become intractable?

Section 3

Approximating the intractable

Approximating distributions

- Delta distribution
  - Point estimate
- Mixture of delta distributions
  - Markov Chain Monte Carlo (MCMC)
  - Neural network ensembles...
- **Gaussian distribution**
  - Laplace
  - Variational Bayes/Variational Inference (VB / VI)
  - Expectation Propagation (EP), PowerEP, ...
- Mixture of Gaussians
- ...



Approximating the exact posterior with Gaussian

Approximating the posterior at observations:

$$p(\mathbf{f} \mid \mathbf{y}) \approx q(\mathbf{f}) = \mathcal{N}(\mathbf{f} \mid \mu = ?, \Sigma = ?)$$

Predictions at new points:

$$p(f^* \mid x^*, \mathbf{y}) \approx q(f^*) = \int p(f^* \mid x^*, \mathbf{f}) \, q(\mathbf{f}) \, \mathrm{d}\mathbf{f}$$

Demo

What does this mean for Gaussian processes?

Choosing μ and Σ for $q(\mathbf{f})$

$$p(\mathbf{f} \mid \mathbf{y}) \approx q(\mathbf{f}) = \mathcal{N}(\mathbf{f} \mid \mu =?, \Sigma =?)$$

locally: match mean & variance at point

globally: minimise divergence

Laplace approximation

Variational inference (VI)

Expectation Propagation (EP)

Laplace approximation

Idea: log of Gaussian pdf = quadratic polynomial

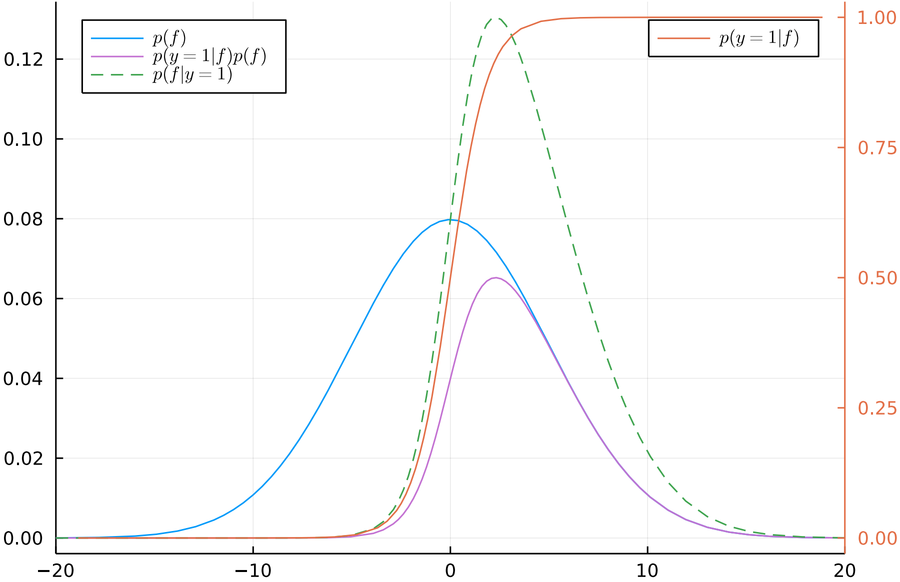
$$p_{\mathcal{N}}(\mathbf{f}) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{f} - \mu)^\top \Sigma^{-1} (\mathbf{f} - \mu)\right)$$

Quadratic polynomial through approximation:
2nd-order Taylor expansion of log of $h(f) = p(y \mid f)p(f)$ at $\hat{f}$

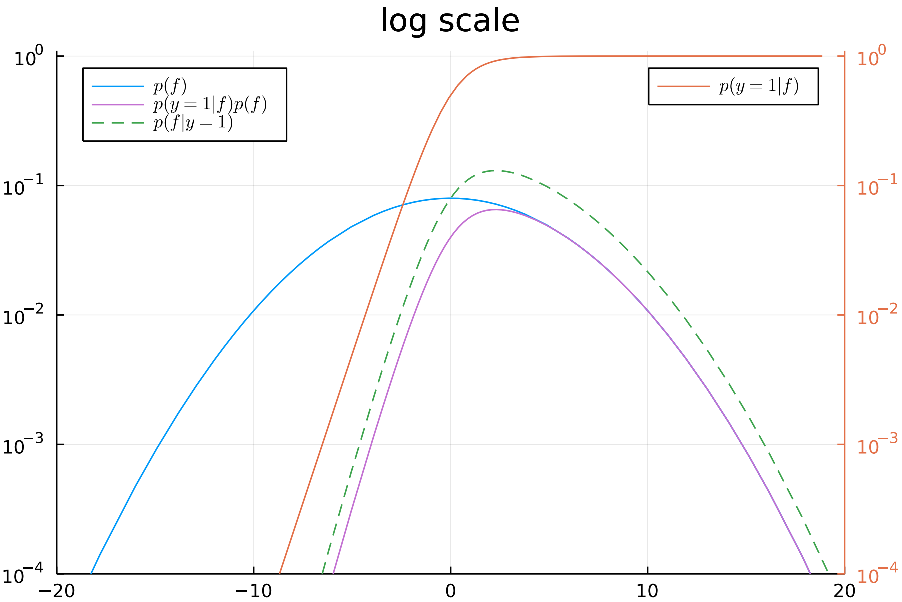
$$g(x + \delta) \approx g(x) + \left(\frac{\mathrm{d}g}{\mathrm{d}x}(x)\right)\delta + \frac{1}{2!}\left(\frac{\mathrm{d}^2 g}{\mathrm{d}x^2}(x)\right)\delta^2$$

1. Find **mode** of posterior
   2nd-order gradient optimisation (e.g. Newton's method)
2. Match **curvature** (Hessian) at mode

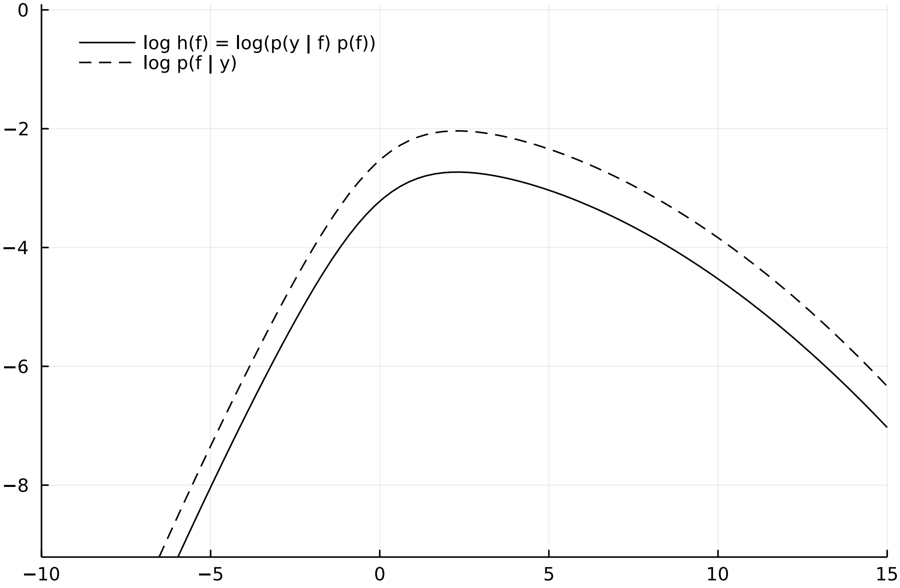
$$p(f \mid y) = \frac{1}{Z} p(y \mid f) p(f)$$



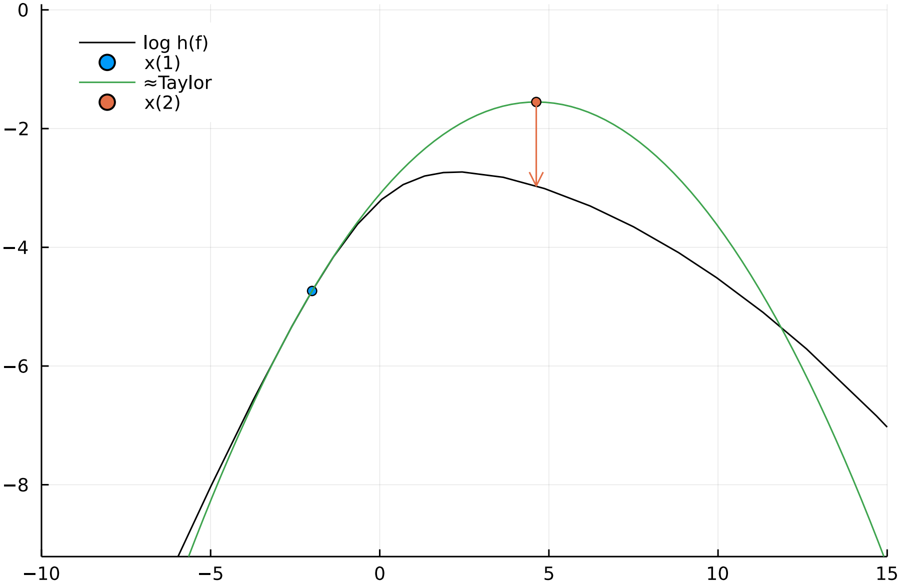
$$\log p(f \mid y) = -\log Z + \log p(y \mid f) + \log p(f)$$



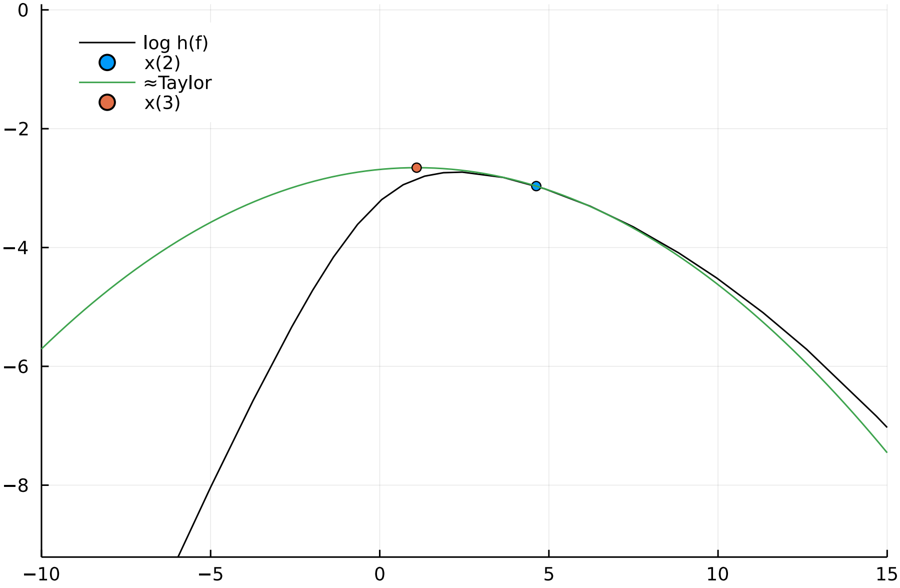
$\log p(f \mid y) = -\log Z + \log h(f)$



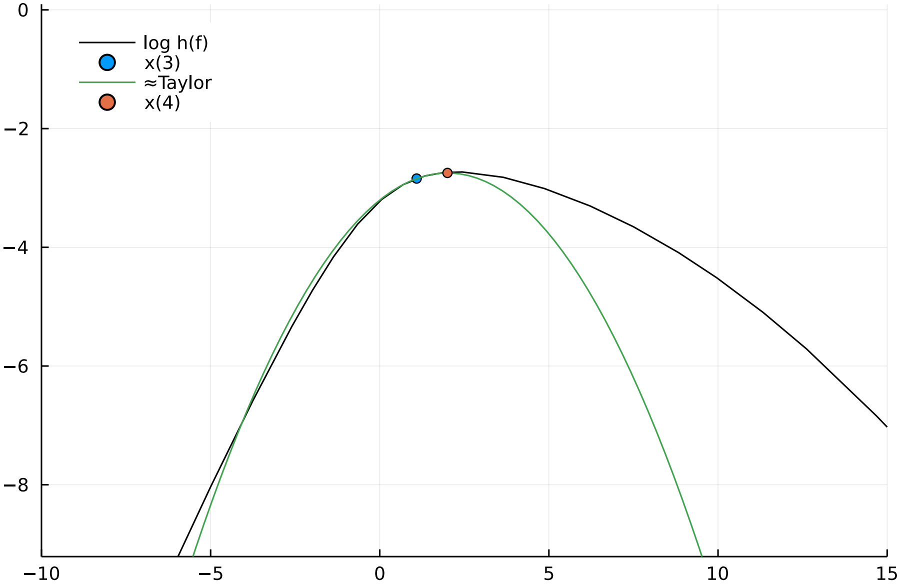
Newton's method



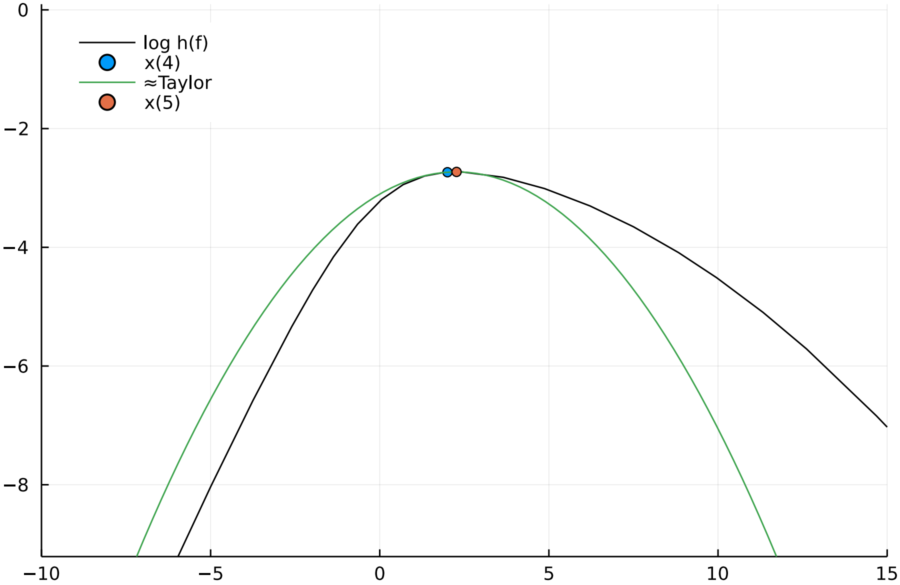
Newton's method



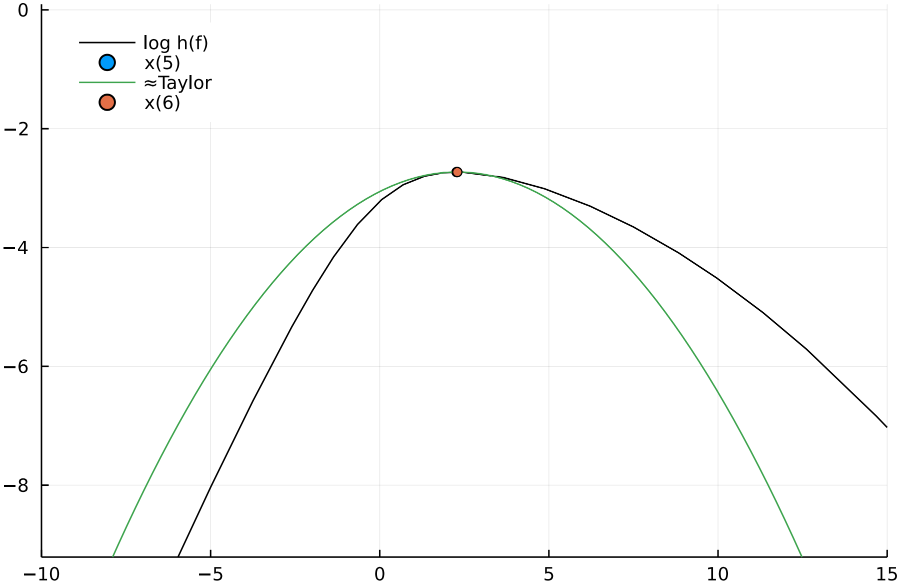
Newton's method



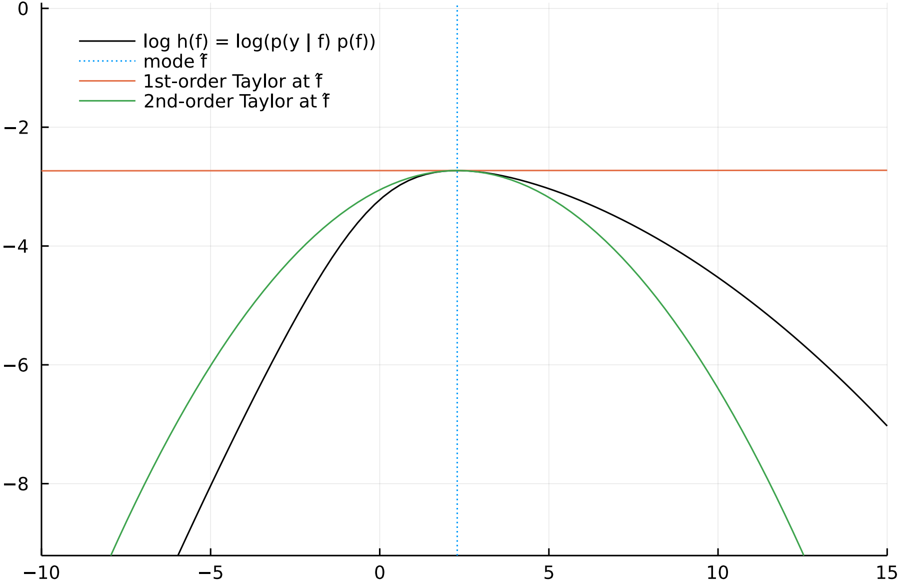
Newton's method



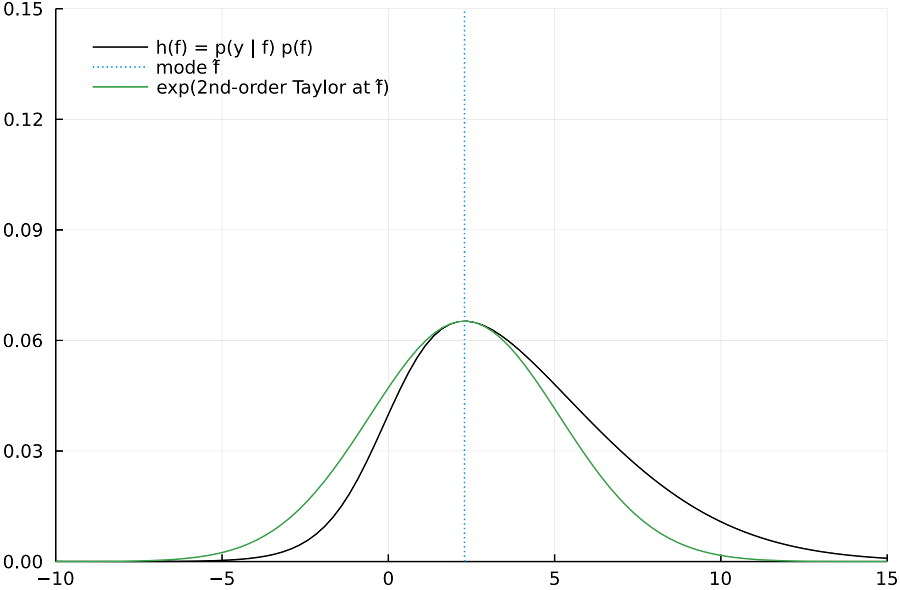
Newton's method



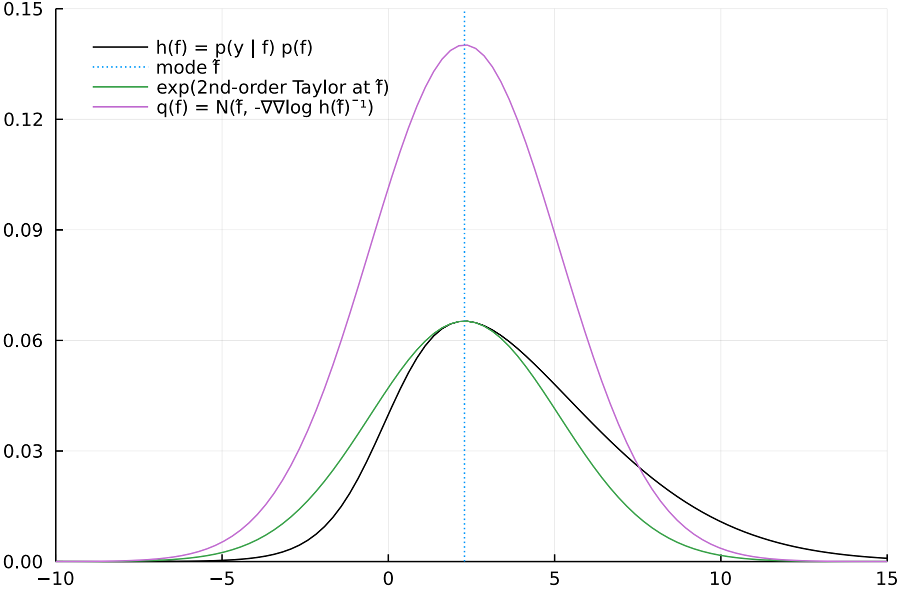
$\log p(f \mid y) + \log Z = \log h(f) \approx \mathcal{O}(f^2)$



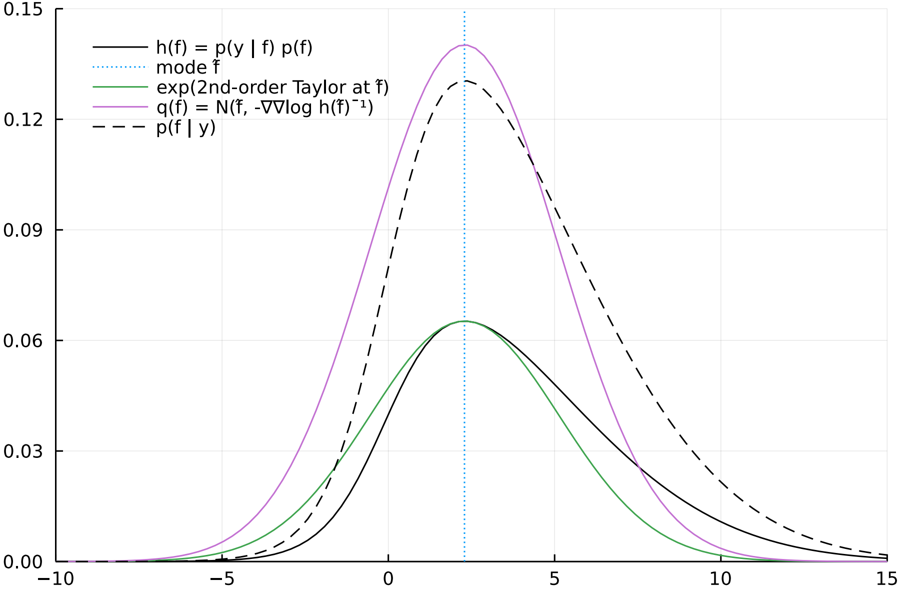
$p(f \mid y)\, Z \approx \exp(\mathcal{O}(f^2))$



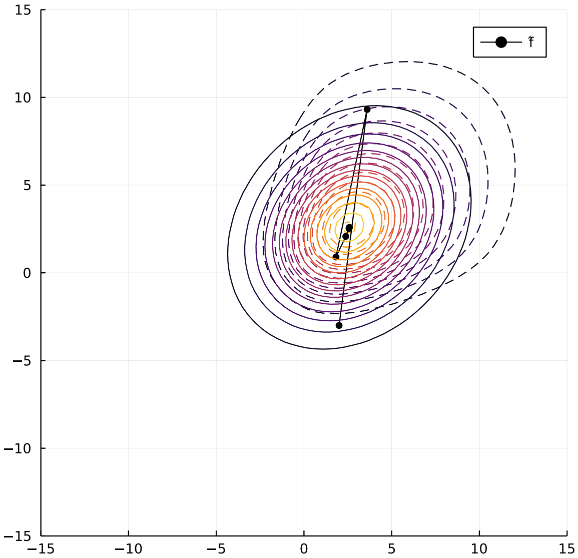
$p(f \mid y) \approx \mathcal{N}(f \mid \hat{f}, -(\mathrm{d}^2 \log h/\mathrm{d} f^2)^{-1})$



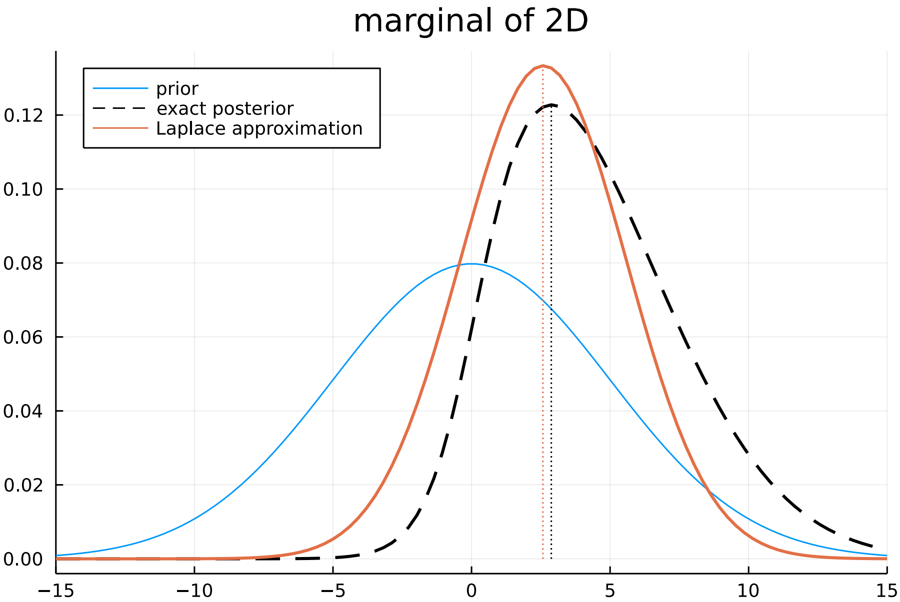
$$p(f \mid y) \approx \mathcal{N}(f \mid \hat{f}, -(\mathrm{d}^2 \log h / \mathrm{d} f^2)^{-1}) = q(f)$$



Laplace in 2D example



Laplace in 2D: marginals



Marginal likelihood approximation (I)

- Finally, we need the marginal likelihood in order to do model selection

$$\begin{aligned} p(\boldsymbol{y}) &= \int p(\boldsymbol{y} \mid \boldsymbol{f}) p(\boldsymbol{f}) \, \mathrm{d}\boldsymbol{f} \\ &= \int \exp\left[\log p(\boldsymbol{y} \mid \boldsymbol{f}) + \log p(\boldsymbol{f})\right] \, \mathrm{d}\boldsymbol{f} \end{aligned}$$

- Let's define of $\psi(\boldsymbol{f}) = \log h(\mathbf{f}) = \log\big(p(\mathbf{y} \mid \mathbf{f}) p(\mathbf{f})\big)$

$$\psi(\boldsymbol{f}) = \log p(\boldsymbol{y} \mid \boldsymbol{f}) + \log p(\boldsymbol{f}) = \log p(\boldsymbol{y} \mid \boldsymbol{f}) - \frac{N}{2} \log(2\pi) - \frac{1}{2} \log|\boldsymbol{K}| - \frac{1}{2} \boldsymbol{f}^\top \boldsymbol{K}^{-1} \boldsymbol{f}$$

- Second order Taylor approximation around the mode $\hat{\boldsymbol{f}}$

$$\psi(\boldsymbol{f}) \approx \psi\left(\hat{\boldsymbol{f}}\right) - \frac{1}{2}\left(\boldsymbol{f} - \hat{\boldsymbol{f}}\right)^\top \boldsymbol{A} \left(\boldsymbol{f} - \hat{\boldsymbol{f}}\right)$$

- Substituting back

$$p(\boldsymbol{y}) \approx q(\boldsymbol{y}) = \int \exp\left[\psi\left(\hat{\boldsymbol{f}}\right) - \frac{1}{2}\left(\boldsymbol{f} - \hat{\boldsymbol{f}}\right)^\top \boldsymbol{A} \left(\boldsymbol{f} - \hat{\boldsymbol{f}}\right)\right] \mathrm{d}\boldsymbol{f}$$

Marginal likelihood approximation (II)

- We have

$$p(\boldsymbol{y}) \approx q(\boldsymbol{y}) = \int \exp\left[\psi\left(\hat{\boldsymbol{f}}\right) - \frac{1}{2}\left(\boldsymbol{f} - \hat{\boldsymbol{f}}\right)^\top \boldsymbol{A}\left(\boldsymbol{f} - \hat{\boldsymbol{f}}\right)\right] \mathrm{d}\boldsymbol{f}$$
$$= \int \exp\left[\psi\left(\hat{\boldsymbol{f}}\right)\right] \exp\left[-\frac{1}{2}\left(\boldsymbol{f} - \hat{\boldsymbol{f}}\right)^\top \boldsymbol{A}\left(\boldsymbol{f} - \hat{\boldsymbol{f}}\right)\right] \mathrm{d}\boldsymbol{f}$$

- $\exp\left[\psi\left(\hat{\boldsymbol{f}}\right)\right]$ does not depend on $\boldsymbol{f}$:

$$p(\boldsymbol{y}) \approx q(\boldsymbol{y}) = \exp\left[\psi\left(\hat{\boldsymbol{f}}\right)\right] \int \exp\left[-\frac{1}{2}\left(\boldsymbol{f} - \hat{\boldsymbol{f}}\right)^\top \boldsymbol{A}\left(\boldsymbol{f} - \hat{\boldsymbol{f}}\right)\right] \mathrm{d}\boldsymbol{f}$$

- The integral evaluates to the normalization constant of a Gaussian

$$p(\boldsymbol{y}) \approx q(\boldsymbol{y}) = \exp\left[\psi\left(\hat{\boldsymbol{f}}\right)\right] (2\pi)^{\frac{N}{2}} \left|\boldsymbol{A}^{-1}\right|^{\frac{1}{2}}$$

- We substitute in the expression for $\exp\left[\psi\left(\hat{\boldsymbol{f}}\right)\right]$:

$$q(\boldsymbol{y}) = \exp\left[\log p(\boldsymbol{y} \mid \hat{\boldsymbol{f}}) - \frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{K}| - \frac{1}{2}\hat{\boldsymbol{f}}^\top \boldsymbol{K}^{-1}\hat{\boldsymbol{f}}\right] (2\pi)^{\frac{N}{2}} \left|\boldsymbol{A}^{-1}\right|^{\frac{1}{2}}$$

Marginal likelihood approximation (III)

- Taking the log of $q\left(\boldsymbol{y}\right)$

$$\begin{aligned}\log q(\boldsymbol{y}) &= \log p(\boldsymbol{y}|\hat{\boldsymbol{f}}) - \frac{N}{2}\log\left(2\pi\right) - \frac{1}{2}\log|\boldsymbol{K}| - \frac{1}{2}\hat{\boldsymbol{f}}^\top\boldsymbol{K}^{-1}\hat{\boldsymbol{f}} + \frac{N}{2}\log\left(2\pi\right) + \frac{1}{2}\log\left|\boldsymbol{A}^{-1}\right| \\ &= \log p(\boldsymbol{y}\mid\hat{\boldsymbol{f}}) - \frac{1}{2}\log|\boldsymbol{K}| - \frac{1}{2}\hat{\boldsymbol{f}}^\top\boldsymbol{K}^{-1}\hat{\boldsymbol{f}} + \frac{1}{2}\log\left|\boldsymbol{A}^{-1}\right|\end{aligned}$$

- We can now use the fact that $|\boldsymbol{A}^{-1}| = |\boldsymbol{A}|^{-1}$ to get

$$\log q(\boldsymbol{y}) = \log p(\boldsymbol{y}\mid\hat{\boldsymbol{f}}) - \frac{1}{2}\hat{\boldsymbol{f}}^\top\boldsymbol{K}^{-1}\hat{\boldsymbol{f}} - \frac{1}{2}\log|\boldsymbol{K}| - \frac{1}{2}\log\left|\boldsymbol{A}\right|$$

- Finally, recall that $\boldsymbol{A} = \boldsymbol{K}^{-1} + \boldsymbol{W}$

$$\log q(\boldsymbol{y}) = \log p(\boldsymbol{y}\mid\hat{\boldsymbol{f}}) - \frac{1}{2}\hat{\boldsymbol{f}}^\top\boldsymbol{K}^{-1}\hat{\boldsymbol{f}} - \frac{1}{2}\log|\boldsymbol{K}| - \frac{1}{2}\log\left|\boldsymbol{K}^{-1} + \boldsymbol{W}\right|$$

- We optimize $\log q(\boldsymbol{y})$ using gradient based methods to choose hyperparameters

Laplace approximation: important properties

- Find mode: Newton's method
- Match curvature (Hessian) at mode
- "Point estimate++"

+ Simple, fast

– Poor approximation if mode is not representative (e.g., Bernoulli)

– May not converge for non-log-concave likelihoods

Choosing μ and Σ for $q(\mathbf{f})$

$$p(\mathbf{f} \mid \mathbf{y}) \approx q(\mathbf{f}) = \mathcal{N}(\mathbf{f} \mid \mu =?, \Sigma =?)$$

locally: match mean & variance at point

globally: minimise divergence

Laplace approximation

Variational inference (VI)

Expectation Propagation (EP)

Why variational inference

- General framework for approximate Bayesian inference
- Many recent applications in the machine learning literature:
  1. GPs with non-Gaussian likelihoods (today)
  2. GPs for big data (tomorrow)
  3. Deep Gaussian processes (next week)
  4. Variational autoencoders (VAEs)
  5. ...

Variational inference: The big picture

Recipe for approximating intractable distribution $p \in \mathcal{P}$

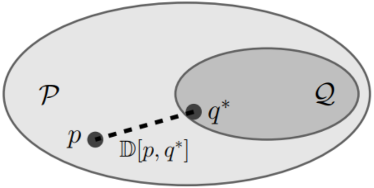
1. Define some "simple" family of distributions $\mathcal{Q}$.
2. Define some way to compute a "distance" $\mathbb{D}[p, q]$ between intractable distribution p and each distribution $q \in \mathcal{Q}$

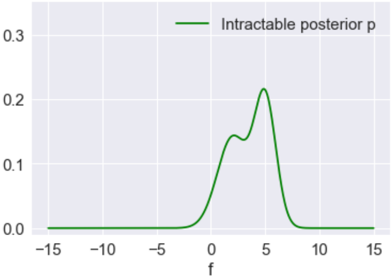
$$\mathbb{D}[p, q_1] > \mathbb{D}[p, q_2]$$

3. Search for $q \in \mathcal{Q}$ such that $\mathbb{D}[p, q]$ is minimized

$$q^* = \arg\min_{q \in \mathcal{Q}} \mathbb{D}[p, q]$$

4. Use q^* as an approximation of p





Here we will always choose $\mathcal{Q}$ to be the set of multivariate Gaussian distributions.

How to “measure distances” between distributions?

Here: *Kullback-Leibler divergence*

$$\mathbb{D}[p,q] := \mathrm{KL}[q \,\|\, p] = \int q(\boldsymbol{f}) \log \frac{q(\boldsymbol{f})}{p(\boldsymbol{f})} \,\mathrm{d}\boldsymbol{f} = \mathbb{E}_q \left[\log \frac{q(\boldsymbol{f})}{p(\boldsymbol{f})} \right]$$

Important properties:

1. Non-symmetric: $\mathrm{KL}[q \,\|\, p] \neq \mathrm{KL}[p \,\|\, q]$
2. Positive: $\mathrm{KL} \geq 0$ (Gibbs’ inequality)
3. Minimum: $\mathrm{KL}[q \,\|\, p] = 0 \iff q \equiv p$.

Variational inference: Minimizing $\mathrm{KL}[q(\mathbf{f}) \,\|\, p(\mathbf{f} \mid \mathbf{y})]$

$$
\begin{aligned}
&\mathrm{KL}[q(\mathbf{f}) \,\|\, p(\mathbf{f} \mid \mathbf{y})] \\
&= \int q(\mathbf{f}) \Big[\log \frac{q(\mathbf{f})}{p(\mathbf{f} \mid \mathbf{y})} \Big] \mathrm{d}\mathbf{f} = \int q(\mathbf{f}) \Big[\log q(\mathbf{f}) - \log p(\mathbf{f} \mid \mathbf{y}) \Big] \mathrm{d}\mathbf{f} \\
&= \int q(\mathbf{f}) \Big[\log q(\mathbf{f}) - \log p(\mathbf{f}) - \log p(\mathbf{y} \mid \mathbf{f}) + \log p(\mathbf{y}) \Big] \mathrm{d}\mathbf{f} \\
&= \int q(\mathbf{f}) \Big[\log \frac{q(\mathbf{f})}{p(\mathbf{f})} \Big] \mathrm{d}\mathbf{f} - \int q(\mathbf{f}) \Big[\log p(\mathbf{y} \mid \mathbf{f}) \Big] \mathrm{d}\mathbf{f} + \log p(\mathbf{y}) \\
&= \mathrm{KL}[q(\mathbf{f}) \,\|\, p(\mathbf{f})] - \int q(\mathbf{f}) \Big[\log p(\mathbf{y} \mid \mathbf{f}) \Big] \mathrm{d}\mathbf{f} + \log p(\mathbf{y}) \\
\log p(\mathbf{y}) &= \int q(\mathbf{f}) \Big[\log p(\mathbf{y} \mid \mathbf{f}) \Big] \mathrm{d}\mathbf{f} - \mathrm{KL}[q(\mathbf{f}) \,\|\, p(\mathbf{f})] + \mathrm{KL}[q(\mathbf{f}) \,\|\, p(\mathbf{f} \mid \mathbf{y})]
\end{aligned}
$$

Variational inference: Minimizing $\mathrm{KL}[q(\mathbf{f}) \,\|\, p(\mathbf{f} \mid \mathbf{y})]$ by bounding

$$\log p(\mathbf{y}) = \underbrace{\int q(\mathbf{f}) \big[\log p(\mathbf{y} \mid \mathbf{f})\big] \mathrm{d}\mathbf{f} - \mathrm{KL}[q(\mathbf{f}) \,\|\, p(\mathbf{f})]}_{\mathcal{L}[q]} + \underbrace{\mathrm{KL}[q(\mathbf{f}) \,\|\, p(\mathbf{f} \mid \mathbf{y})]}_{\geq 0}$$

$$\geq \int q(\mathbf{f}) \big[\log p(\mathbf{y} \mid \mathbf{f})\big] \mathrm{d}\mathbf{f} - \mathrm{KL}[q(\mathbf{f}) \,\|\, p(\mathbf{f})] = \mathcal{L}[q]$$

- $\log p(\mathbf{y})$ is a constant
- $\mathcal{L}[q]$ does not depend on $p(\mathbf{f} \mid \mathbf{y})$
- $\mathcal{L}[q] \leq \log p(\mathbf{y})$, so $\mathcal{L}[q]$ is *lower bound* on marginal log likelihood
- Maximizing $\mathcal{L}[q]$ is equivalent to minimizing $\mathrm{KL}[q(\mathbf{f}) \,\|\, p(\mathbf{f} \mid \mathbf{y})]$

Key take-away: we can fit variational approximation q by optimizing $\mathcal{L}$

Variational inference: ELBO

$$\log p(\mathbf{y}) \geq \mathcal{L}[q] = \underbrace{\int q(\mathbf{f}) \big[\log p(\mathbf{y} \mid \mathbf{f}) \big] \mathrm{d}\mathbf{f}}_{\text{data fit}} - \underbrace{\mathrm{KL}[q(\mathbf{f}) \,\|\, p(\mathbf{f})]}_{\text{regularization}}$$

$\mathcal{L}[q]$ often called the *Evidence Lower Bound* (ELBO)

- To approximate $p(\boldsymbol{f} \mid \boldsymbol{y})$, use $q(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f} \mid \boldsymbol{m}, \boldsymbol{S})$
- Define $\boldsymbol{\lambda} = \{\boldsymbol{m}, \boldsymbol{S}\}$, then we can write $\mathcal{L}[q] = \mathcal{L}[\boldsymbol{\lambda}]$
- In practice, we optimize $\mathcal{L}[\boldsymbol{\lambda}]$ using gradient-based methods

Likelihood term

Integral separates for a factorizing likelihood:

$$\int q(\mathbf{f}) \left[\log p(\mathbf{y} \mid \mathbf{f}) \right] \mathrm{d}\mathbf{f}$$
$$= \sum_{n=1}^{N} \int q(f_n) \left[\log p(y_n \mid f_n) \right] \mathrm{d} f_n$$

Sum over 1D integrals

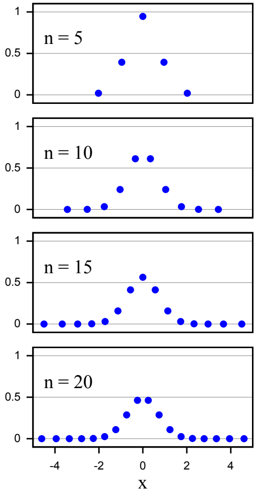
Each integral is a Gaussian expectation of the log likelihood

- Analytic for some (e.g., Exponential, Gamma, Poisson)
- Fast and accurate to approximate numerically (e.g., Gauss–Hermite quadrature)
- Monte Carlo (e.g., multi-class classification)

Take away #2: We can tractably optimize the bound for non-Gaussian likelihoods

Gauss–Hermite Quadrature

$$\int q(f_n)\left[\log p(y_n \mid f_n)\right]\mathrm{d}f_n, \qquad q(f_n) = \mathcal{N}(m_n, S_n)$$



Gauss–Hermite quadrature can be applied:

$$\mathbb{E}_{q(f_n)}[\log p(y_n \mid f_n)] \approx \sum_{j=1}^{C} w_j \log p(y_n \mid f_j),$$

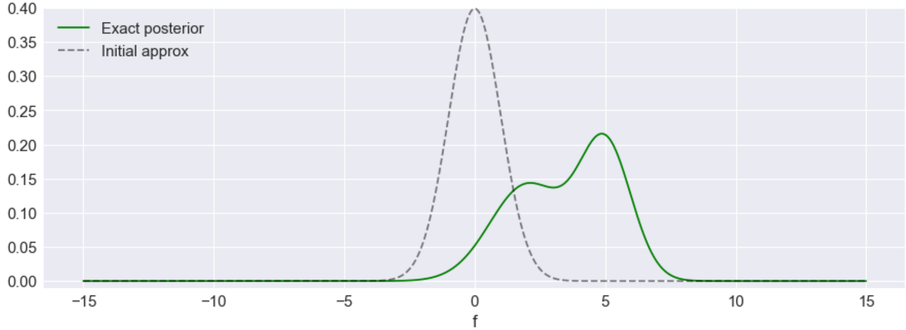
$$w_j = \frac{2^{C-1} C! \sqrt{\pi}}{C^2 [H_{C-1}(f_j)]^2}$$

Gauss–Hermite is exact if $\log p(y_n \mid f_n)$ is polynomial of order less than C

1D Toy example I

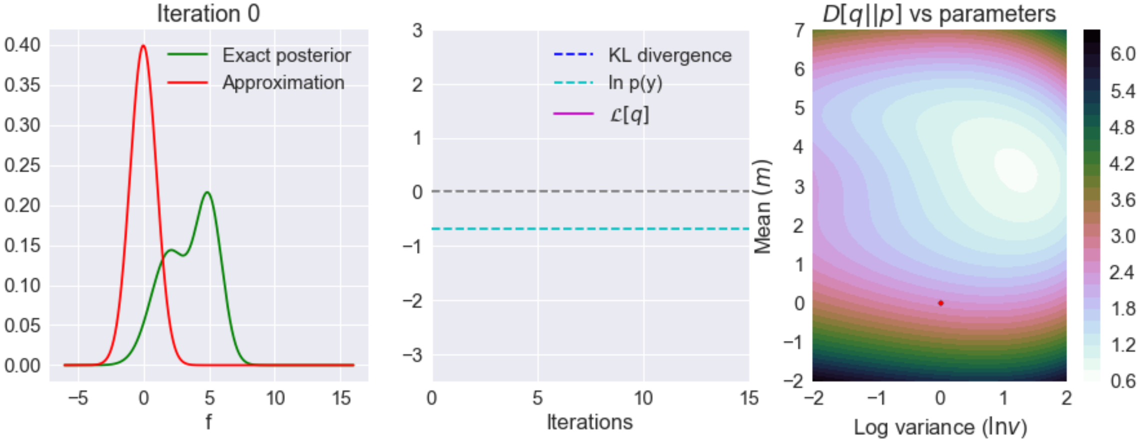
Assume model $p(y, f)$ with some intractable posterior $p(f \mid y)$

- Variational approximation for $p(f \mid y)$
- In 1D: $\mathcal{Q}$ is the set of univariate Gaussians, i.e. $q_\lambda(x) = \mathcal{N}(x \mid m, v)$, and $\boldsymbol{\lambda} = \{m, v\}$
- Initialization: $q(f) = \mathcal{N}\left(f \mid 0, 1\right)$

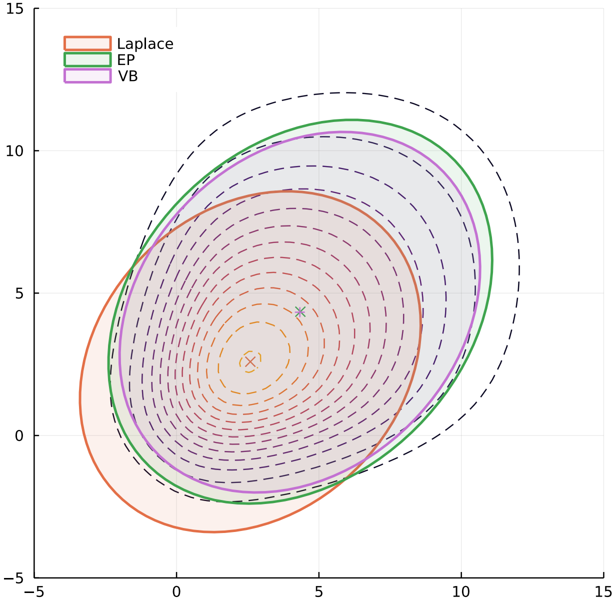


1D Toy example II

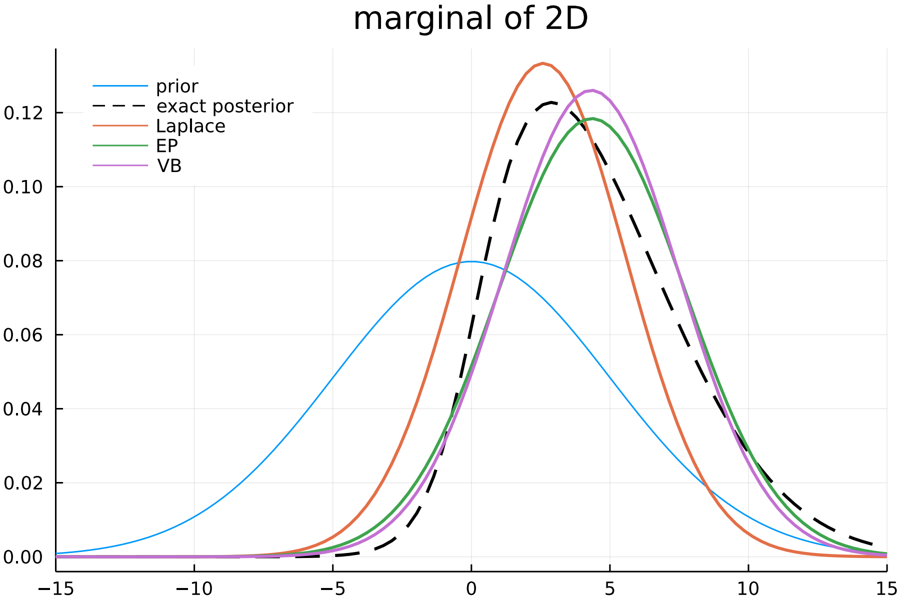
- Gradient ascent: $\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i + \eta \nabla_{\boldsymbol{\lambda}} \mathcal{L}\left[\boldsymbol{\lambda}\right]$
- $\log p(\boldsymbol{y}) = \mathcal{L}\left[\boldsymbol{\lambda}\right] + \mathbb{D}\left[q_\lambda(\boldsymbol{f}) \,\|\, p(\boldsymbol{f} \mid \boldsymbol{y})\right] \geq \mathcal{L}\left[\boldsymbol{\lambda}\right]$



Comparison 2D



Marginals



marginal of 2D
prior
exact posterior
Laplace
EP
VB
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−15
−10
−5
0
5
10
15

Variational Bayes: Important properties

- Principled: directly minimising divergence from true posterior
- Mode-seeking (e.g., multi-modal posterior: fits just one, if q is unimodal)

+ Minimises a true lower bound $\rightarrow$ convergence

− Underestimates variance

Section 4

Conclusion

Posterior distribution for f_*

- Now we know how to compute the approximate posterior $q(\boldsymbol{f} \mid \boldsymbol{y})$
- Let's now consider the posterior distribution for f_*

$$
\begin{aligned}
p(f_* \mid y) &= \int p\left(f_* \mid \boldsymbol{f}\right) p\left(\boldsymbol{f} \mid \boldsymbol{y}\right) \,\mathrm{d}\boldsymbol{f} \\
&= \int \mathcal{N}\left(f_* \mid \boldsymbol{k}_{f_* \boldsymbol{f}} \boldsymbol{K}^{-1} \boldsymbol{f}, k_{f_* f_*} - \boldsymbol{k}_{f_* \boldsymbol{f}} \boldsymbol{K}^{-1} \boldsymbol{k}_{f_* \boldsymbol{f}}^\top\right) p\left(\boldsymbol{f} \mid \boldsymbol{y}\right) \,\mathrm{d}\boldsymbol{f} \\
&\approx \int \mathcal{N}\left(f_* \mid \boldsymbol{k}_{f_* \boldsymbol{f}} \boldsymbol{K}^{-1} \boldsymbol{f}, k_{f_* f_*} - \boldsymbol{k}_{f_* \boldsymbol{f}} \boldsymbol{K}^{-1} \boldsymbol{k}_{f_* \boldsymbol{f}}^\top\right) \mathcal{N}\left(\boldsymbol{f} \mid \hat{\boldsymbol{f}}, \boldsymbol{A}^{-1}\right) \,\mathrm{d}\boldsymbol{f} \\
&= \mathcal{N}\left(f_* \mid \underbrace{\boldsymbol{k}_{f_* \boldsymbol{f}} \boldsymbol{K}^{-1} \hat{\boldsymbol{f}}}_{\mu_*}, \underbrace{k_{f_* f_*} - \boldsymbol{k}_{f_* \boldsymbol{f}} \left(\boldsymbol{K} + \boldsymbol{W}^{-1}\right)^{-1} \boldsymbol{k}_{f_* \boldsymbol{f}}^\top}_{\sigma_*^2}\right) \\
&= \mathcal{N}\left(f_* \mid \mu_*, \sigma_*^2\right)
\end{aligned}
$$

Predictive distribution

- Using the (approximate) posterior $q(f_*)$, we can compute $p(y_* \mid \boldsymbol{y})$

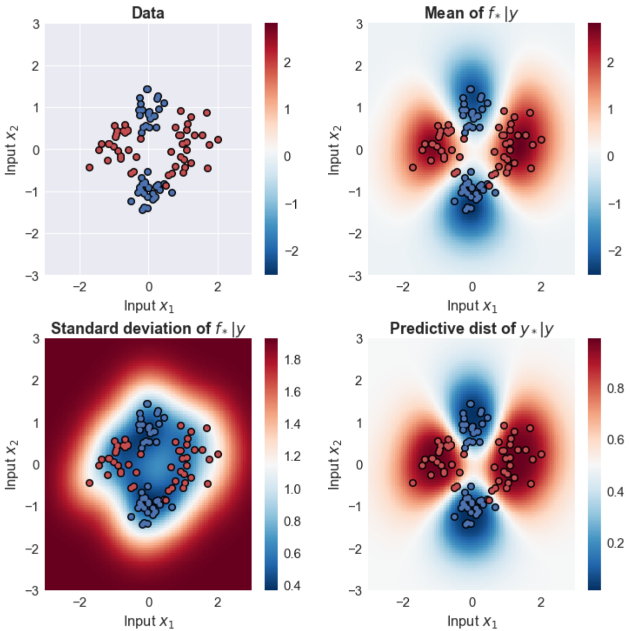
$$
\begin{aligned}
p(y_* = 1 \mid \boldsymbol{y}) &= \int p(y_* \mid f_*) p(f_* \mid \boldsymbol{y}) \, \mathrm{d}f_* \\
&= \int \phi\left(y_* \cdot f_*\right) p(f_* \mid \boldsymbol{y}) \, \mathrm{d}f_* \\
&\approx \int \phi\left(y_* \cdot f_*\right) q\left(f_*\right) \, \mathrm{d}f_* \\
&= \int \phi\left(y_* \cdot f_*\right) \mathcal{N}\left(f_* \mid \mu_*, \sigma_*^2\right) \, \mathrm{d}f_* \\
&= \phi\left(\frac{\mu_*}{\sqrt{1+\sigma_*^2}}\right)
\end{aligned}
$$

Question

- What can we say about predictive distributions for y_* when μ_* is positive? (or negative?)
- How does uncertainty of posterior distribution of f_* influence the predictions for y_*? What happens as σ_*^2 approaches ∞?

Gaussian process classification example

- Non-linear classification problem
- $N = 100$ data points
- Squared exponential kernel
- Hyperparameters are chosen by optimizing $\mathcal{L}[q]$



End of today's lecture

- GPs can be used for all kinds of response variables
- Likelihood with parameters modulated by latent GP
- Non-Gaussian likelihood: non-Gaussian posterior, inference no longer exact
- Approximations: We covered Laplace and Variational Inference