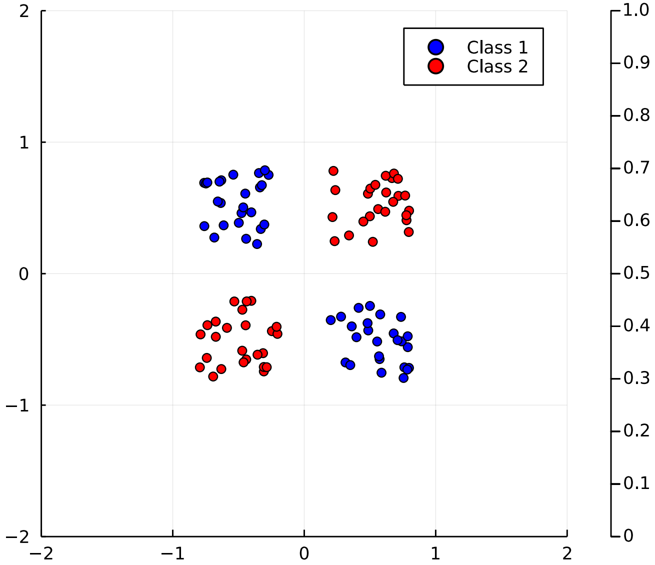
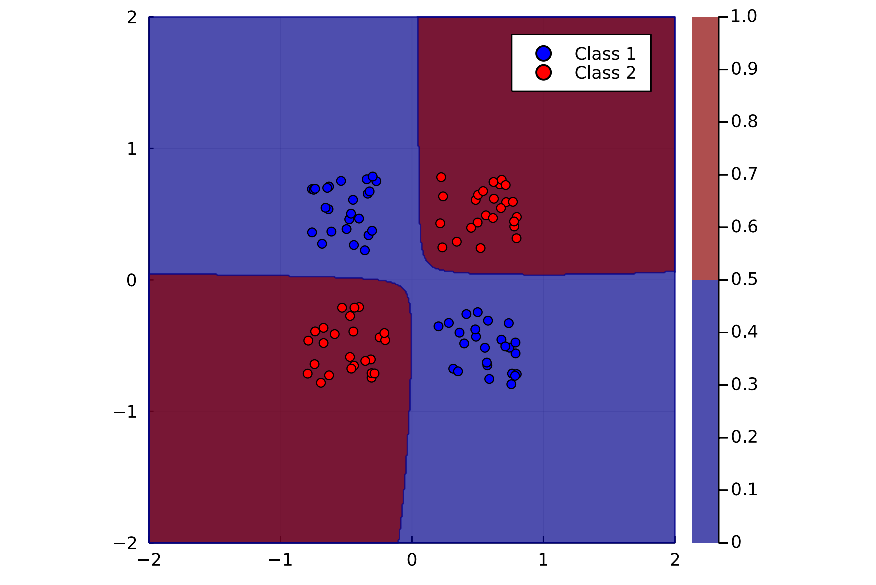
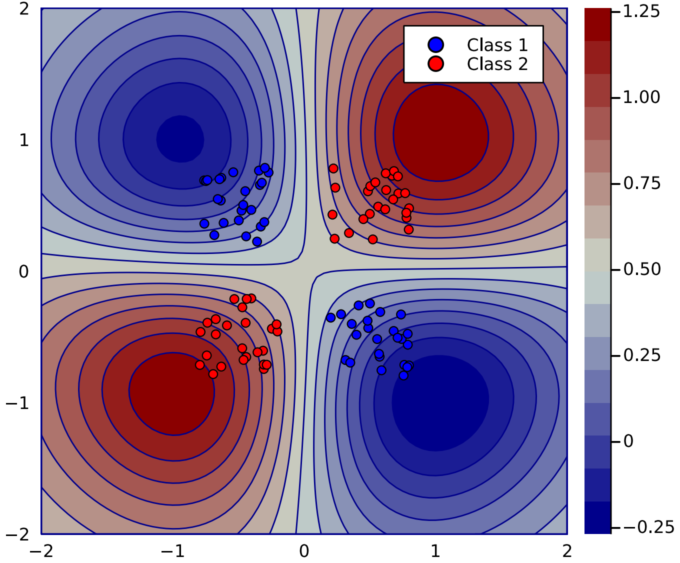


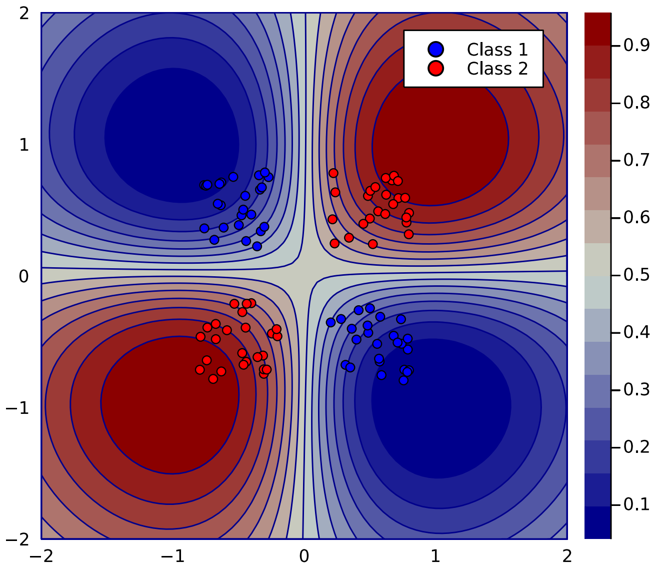
How can we model this?



SVM classification



Gaussian process regression



Gaussian process **classification**

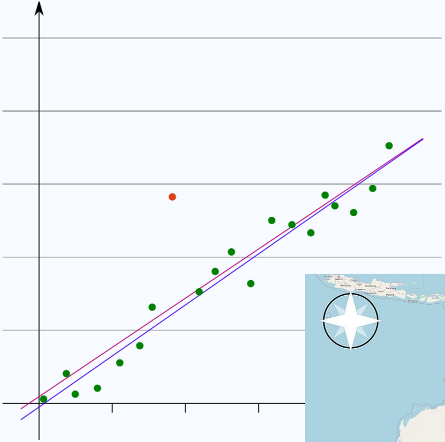
Gaussian processes for non-Gaussian likelihoods

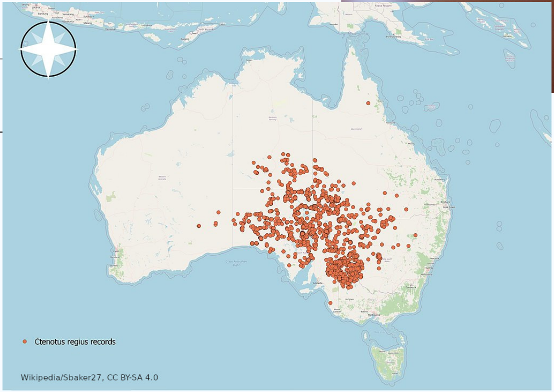
ST John
ti.john @ aalto.fi

Finnish Center for Artificial Intelligence
& Aalto University

Gaussian Process Summer School 2022, 13 September 2022

Not Gaussian noise





Overview

Outline:

1. **Gaussian processes with Gaussian likelihood**
2. What is the likelihood? Connecting observations and Gaussian process prior
3. Non-Gaussian likelihoods: what happens to the posterior?
4. How to approximate the intractable
5. Comparison

+ *Intuitive* understanding
+ Learning the language

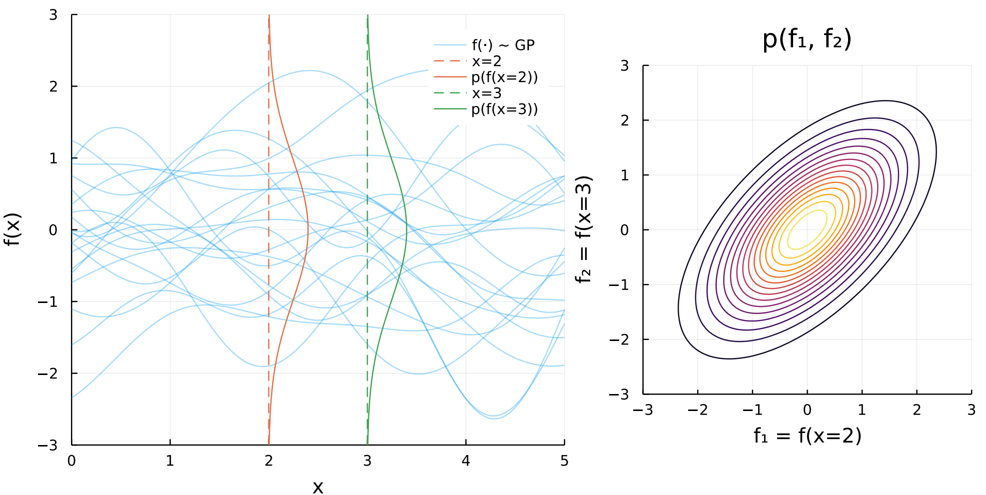
– In-depth expertise
– Lots of maths

Setting the scene

Gaussian process $f(\cdot)$

Distribution over *functions*

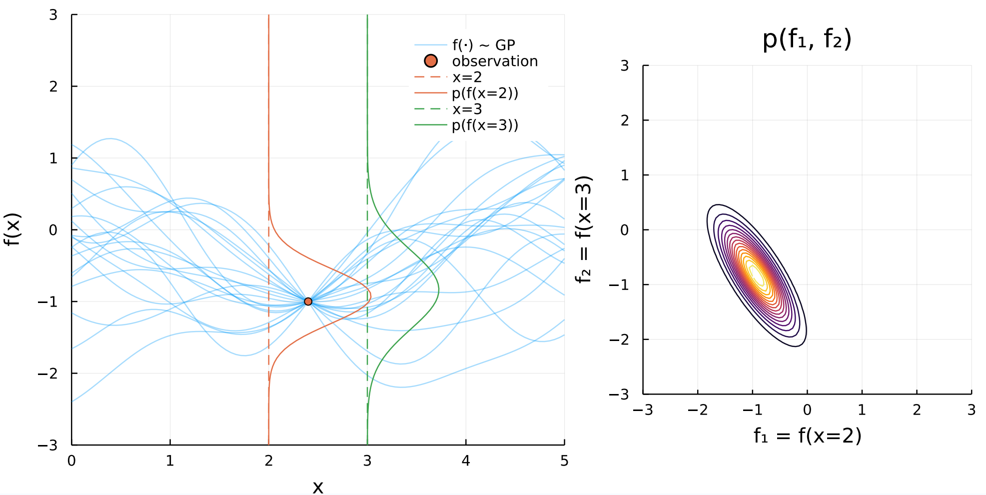
Marginals are Gaussian (mean and covariance)



infinitecuriosity.org/vizgp

Gaussian process conditioned on observation

Distribution over *functions*
Marginals are Gaussian (mean and covariance)



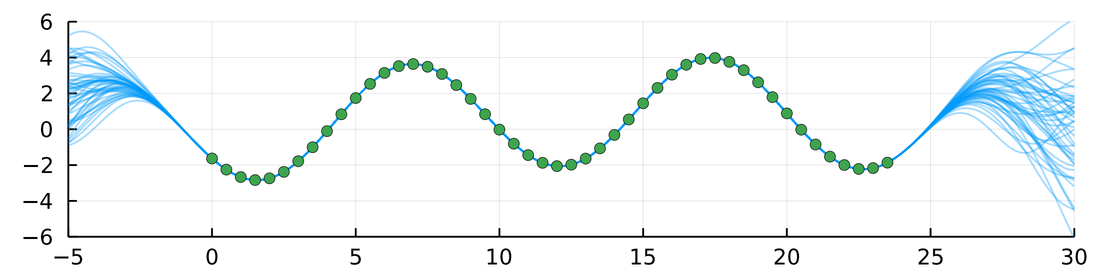
infinitecuriosity.org/vizgp

exact conditioning

Without noise model, we interpolate observations:

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

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

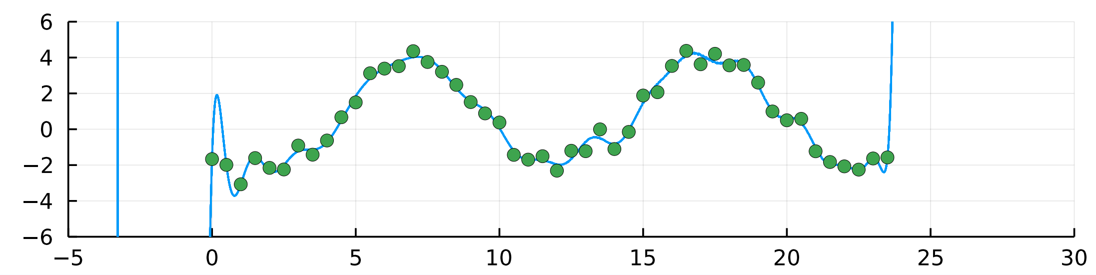


exact conditioning

Without noise model, we interpolate observations:

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

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

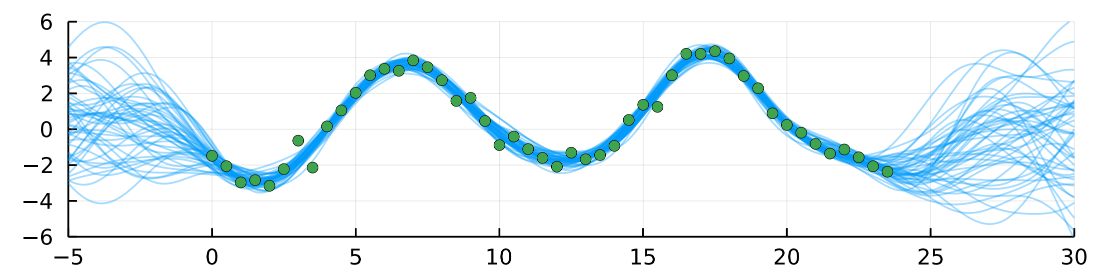


Gaussian noise model

Gaussian additive noise model, written two ways:

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

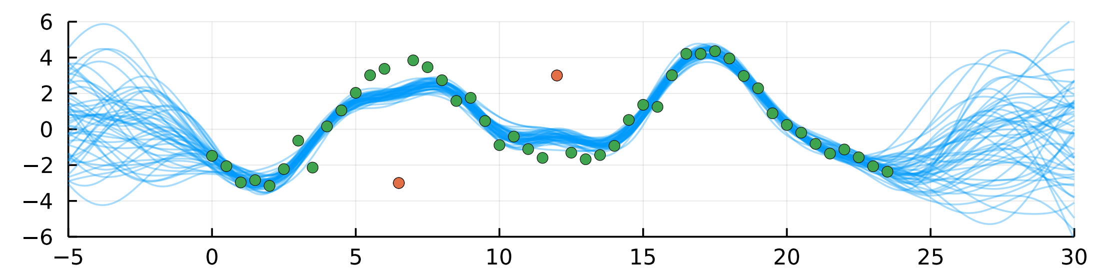
$$p(y \mid f) = \mathcal{N}(y \mid f, \sigma^2_{\text{noise}})$$



misspecified Gaussian noise model

Gaussian additive noise model, written two ways:

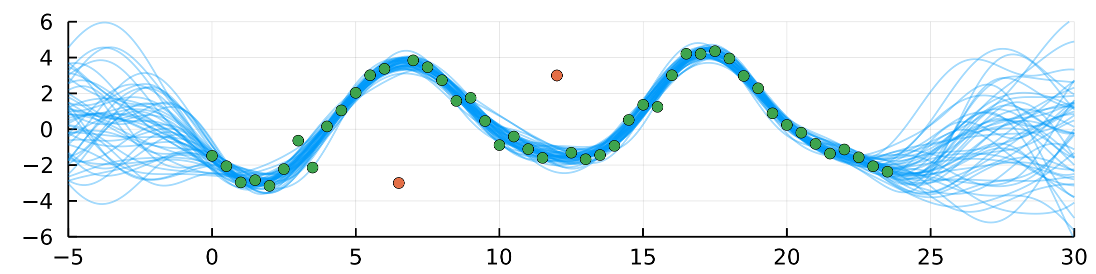
$$y(x) = f(x) + \epsilon, \qquad \epsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_{\text{noise}})$$
$$p(y \mid f) = \mathcal{N}(y \mid f, \sigma^2_{\text{noise}})$$



heavy-tailed noise model

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

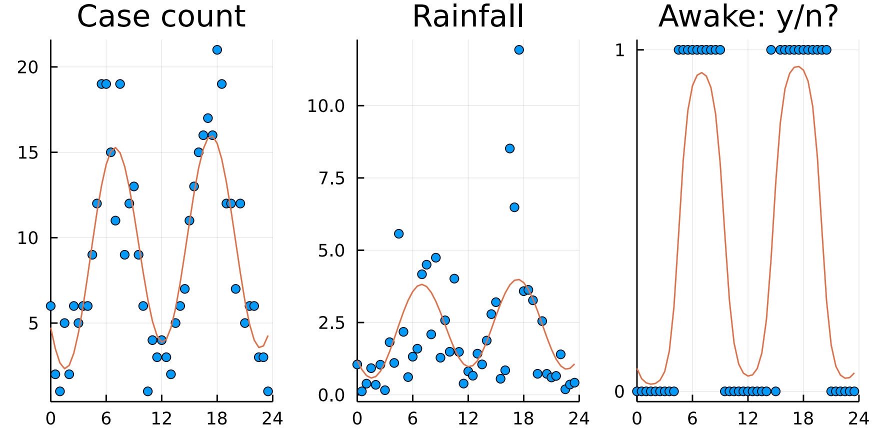
$$p(y \mid f) = \mathcal{N}(y \mid f, \sigma^2_{\text{noise}})$$



Outline

Likelihood

Non-Gaussian observations



latent functional relationship

$$p(y_n \mid f(x_n))$$

Likelihood

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

factorizing

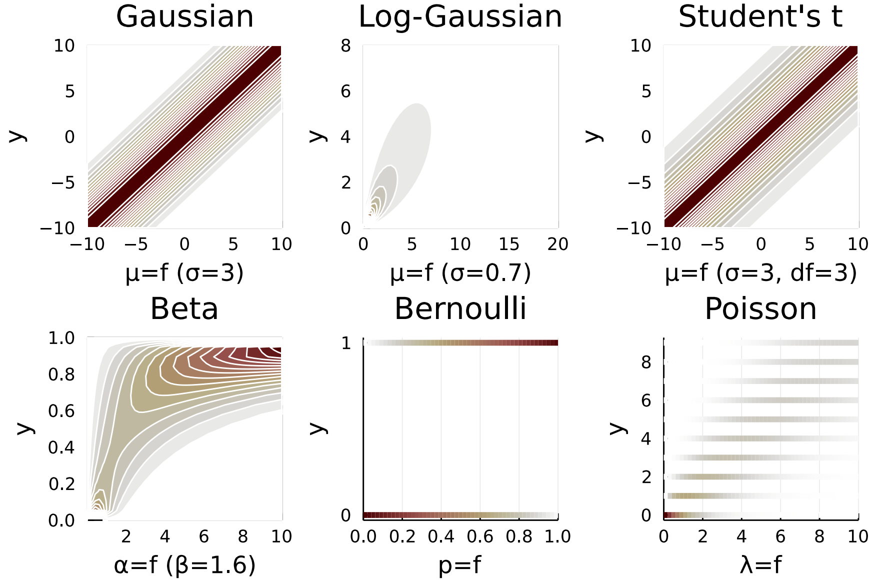
Let's consider the individual (marginal, 1D) likelihood term:

$$p(y \mid f)$$

Function of two arguments:

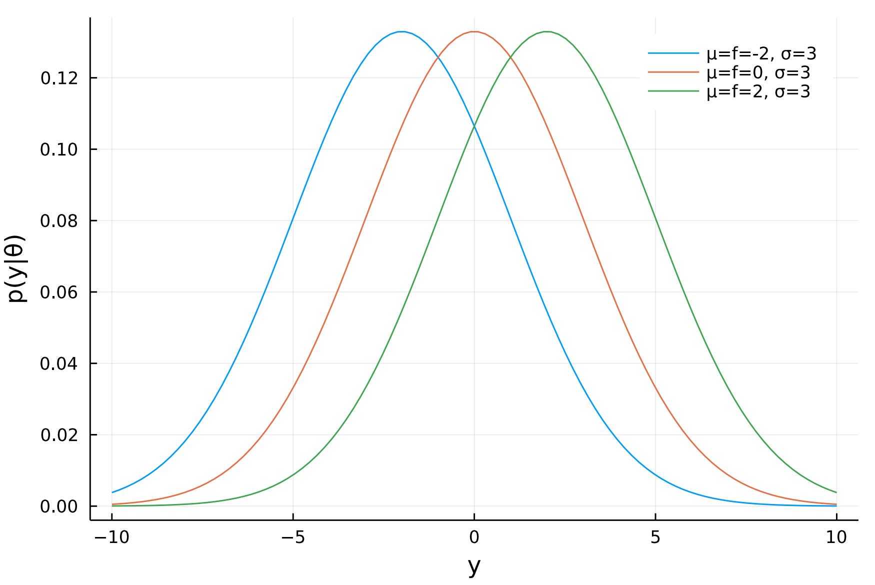
$$y \mapsto p(y \mid f), \qquad f \mapsto p(y \mid f)$$

$p(y\,|\,f)$

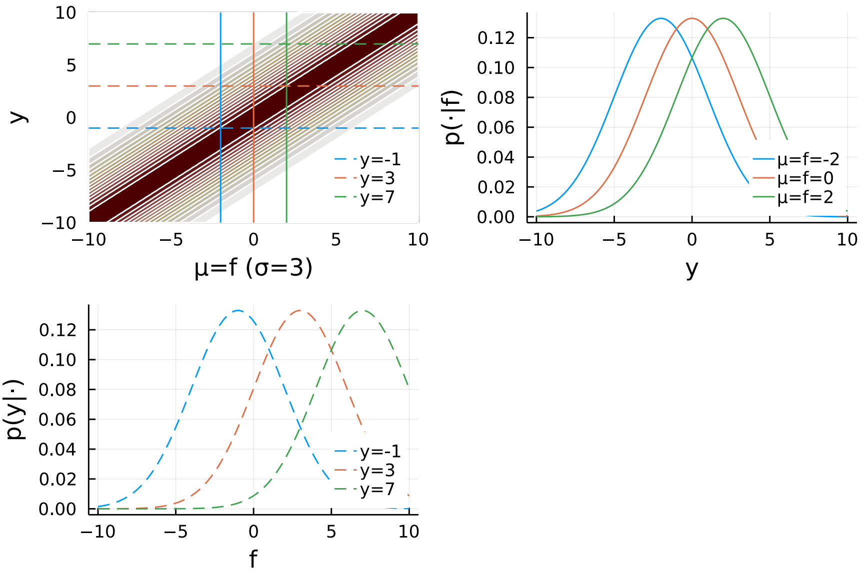


Gaussian
Log-Gaussian
Student's t
Beta
Bernoulli
Poisson
y
μ=f (σ=3)
μ=f (σ=0.7)
μ=f (σ=3, df=3)
α=f (β=1.6)
p=f
λ=f

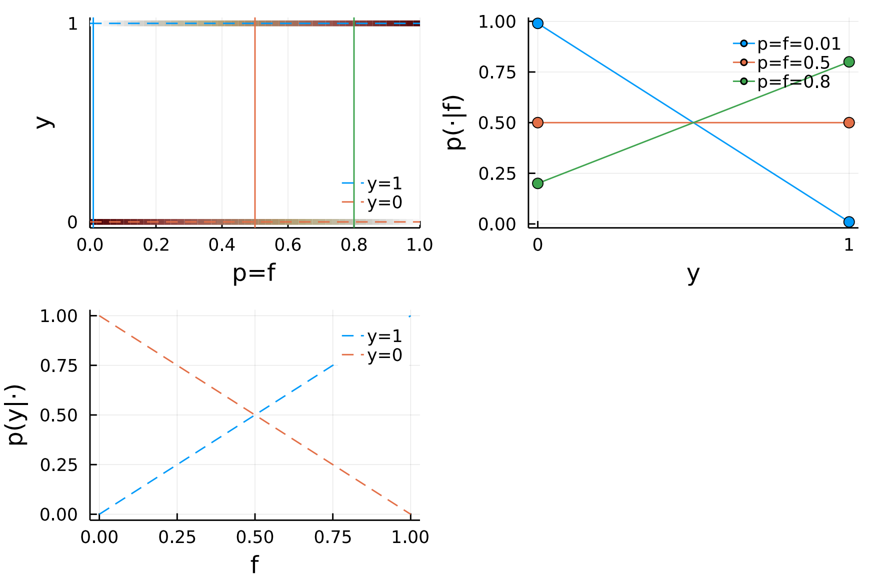
$p(y \mid f)$: Gaussian



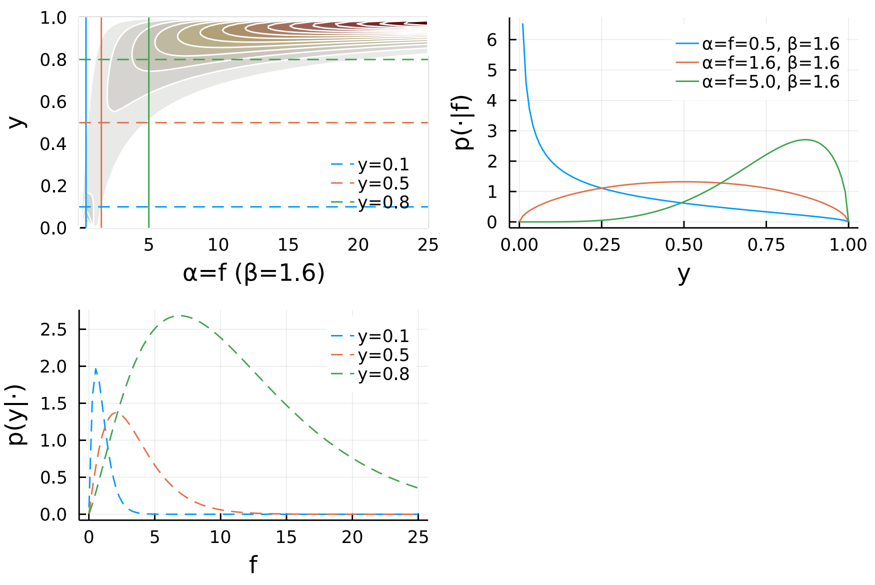
$p(y \mid f)$: Gaussian



$p(y \mid f)$: Bernoulli

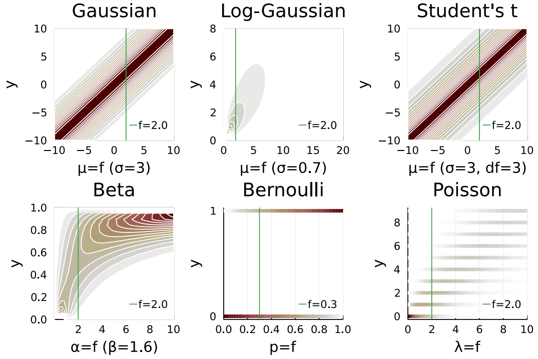


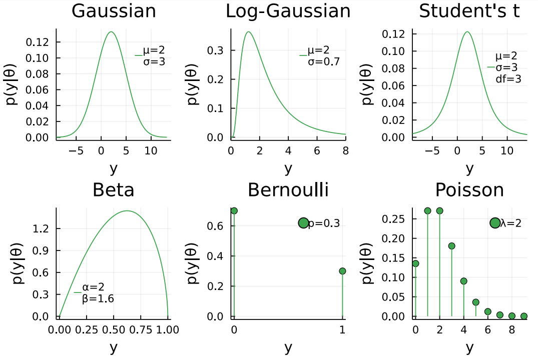
$p(y \mid f)$: Beta



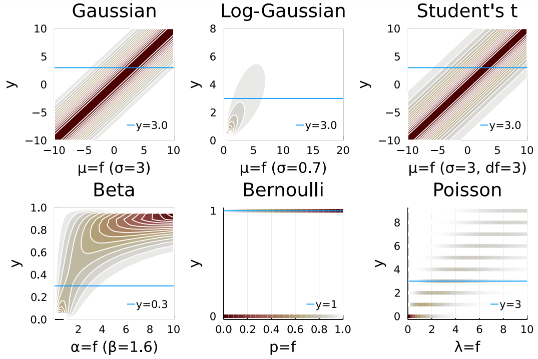
1.0
0.8
0.6
0.4
0.2
0.0
y
5
10
15
20
25
α=f (β=1.6)
y=0.1
y=0.5
y=0.8
p(·|f)
6
5
4
3
2
1
0
0.00
0.25
0.50
0.75
1.00
y
α=f=0.5, β=1.6
α=f=1.6, β=1.6
α=f=5.0, β=1.6
p(y|·)
2.5
2.0
1.5
1.0
0.5
0.0
0
5
10
15
20
25
f
y=0.1
y=0.5
y=0.8

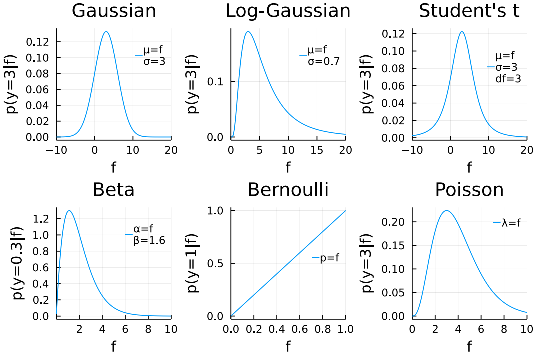
$p(y \mid f)$: distribution of observation y — f fixed





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





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

Two aspects of likelihoods:

1. link functions
2. log-concavity

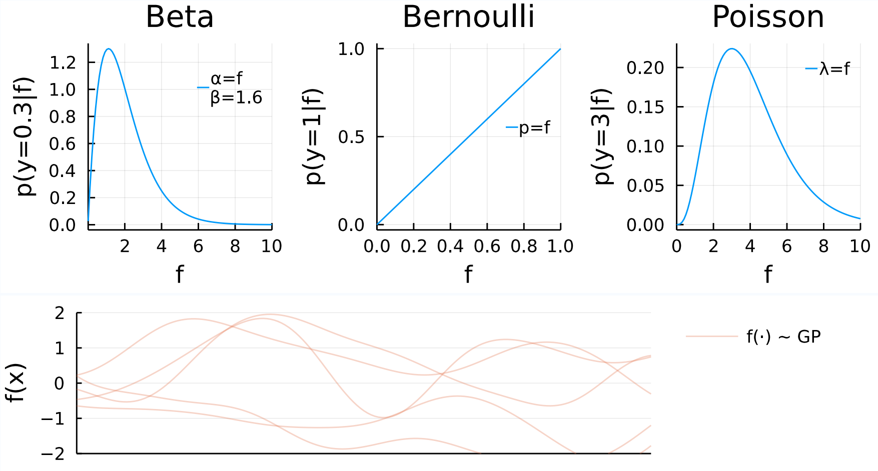
Link functions

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

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

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

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



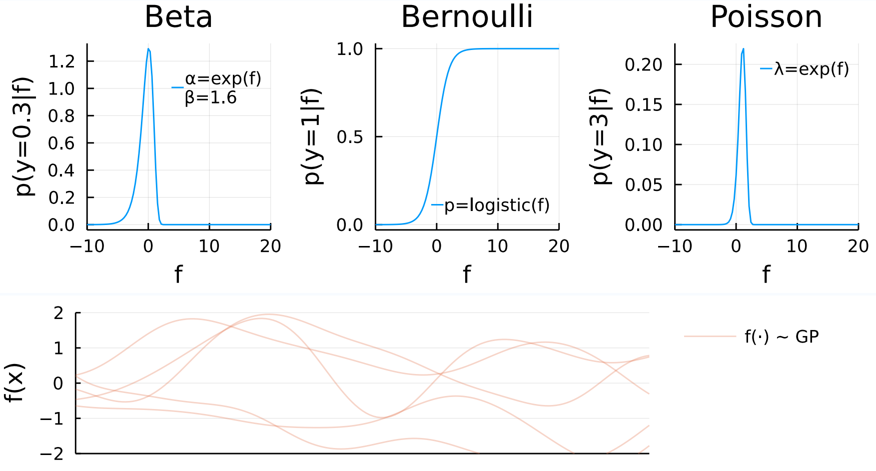
Link functions

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

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

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

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



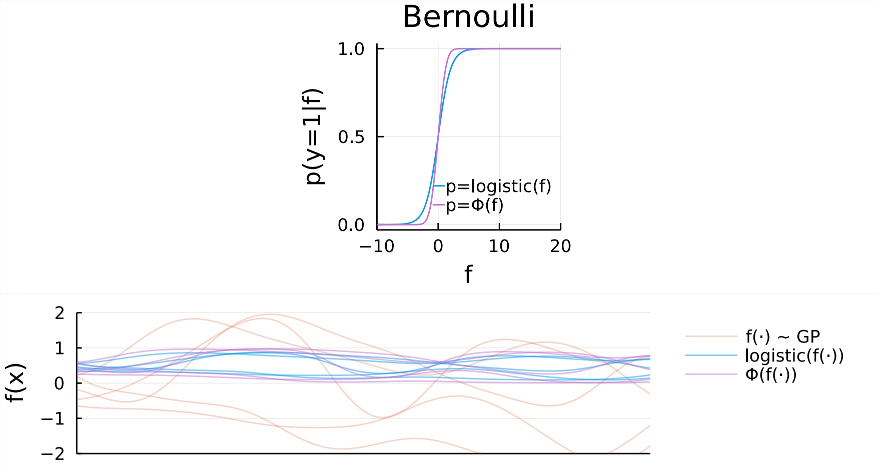
Link functions

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

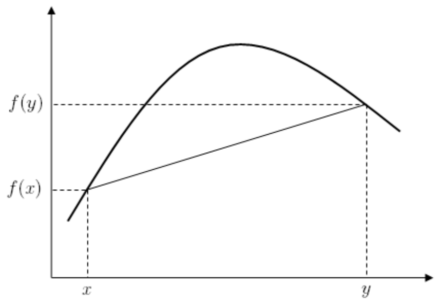
$$f \sim \mathcal{N} \qquad \in (-\infty \ldots \infty)$$

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

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

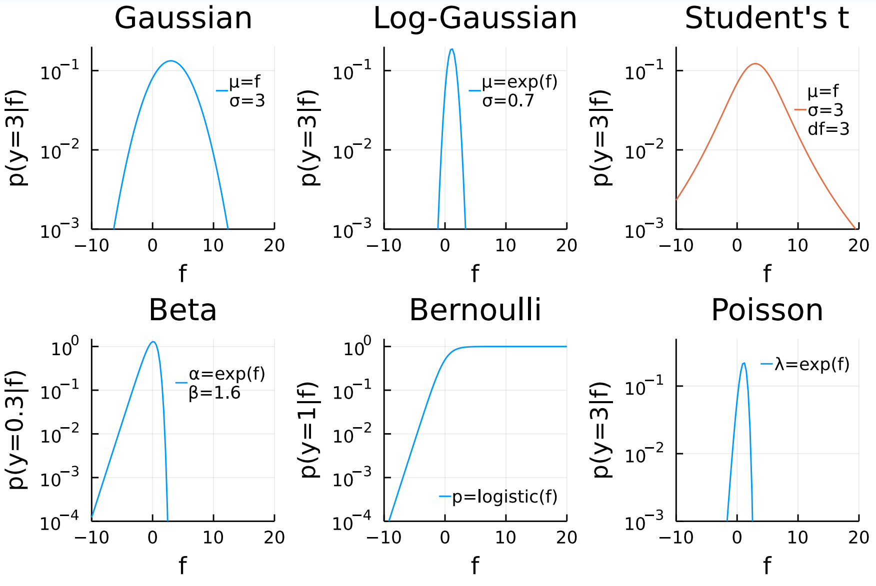


(Log-)concavity



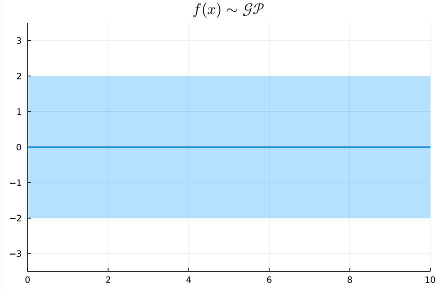
$$f\left(\alpha x + (1-\alpha)y\right) \geq \alpha f(x) + (1-\alpha)f(y)$$

Log-concavity of likelihoods



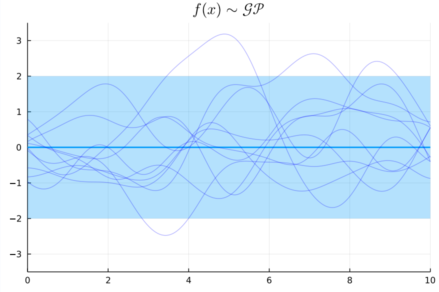
Back to GPs...

Functional prior $p(f)$



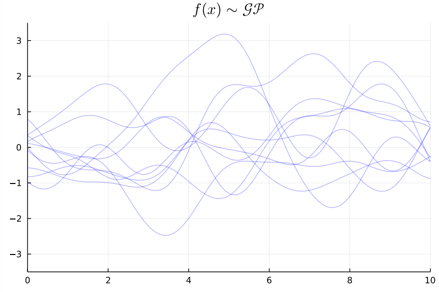
Back to GPs...

Functional prior $p(f)$



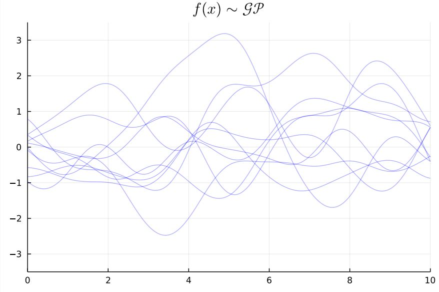
Back to GPs…

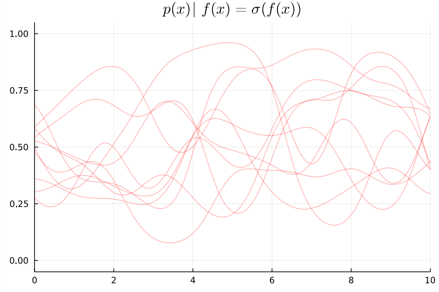
Functional prior $p(f)$



Back to GPs...

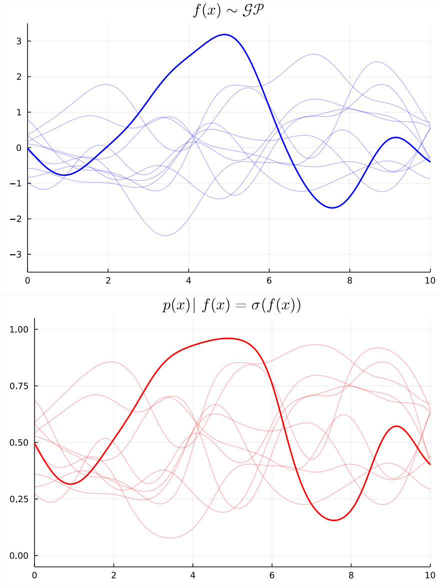
Joint (generative) model: $p(y, f) = p(y \mid f)p(f)$





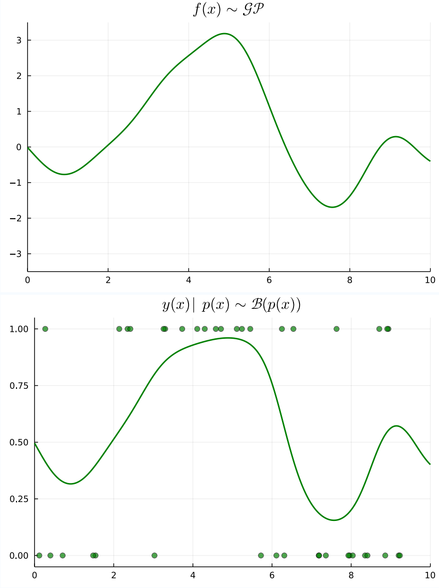
Back to GPs...

Joint (generative) model: $p(y, f) = p(y \mid f)p(f)$



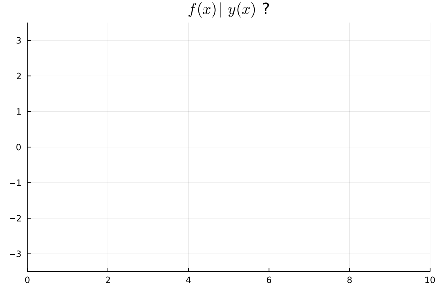
Back to GPs...

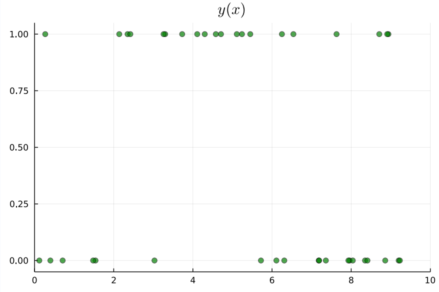
Joint (generative) model: $p(y, f) = p(y \mid f)p(f)$



Back to GPs...

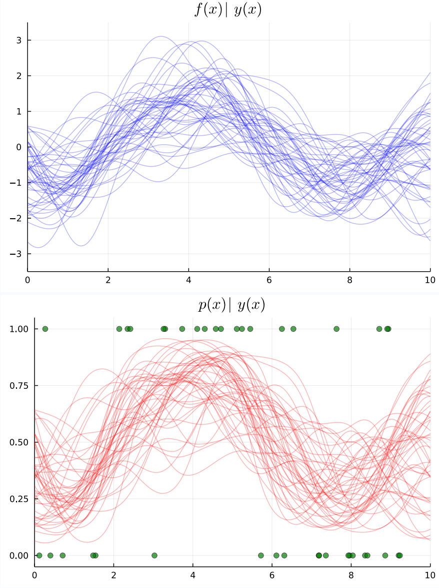
Posterior: $p(f \mid y) = p(y \mid f)p(f)\,/\,p(y)$





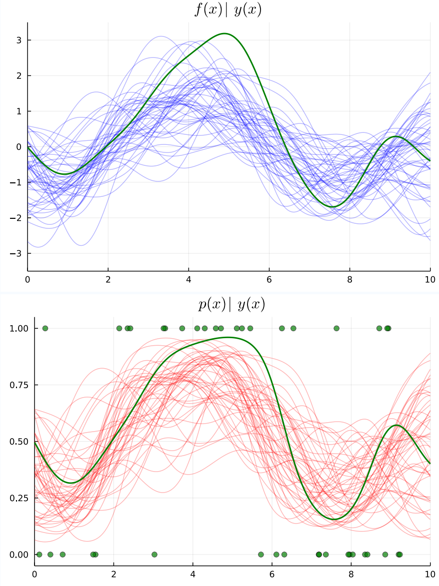
Back to GPs…

Posterior: $p(f \mid y) = p(y \mid f)p(f)/p(y)$



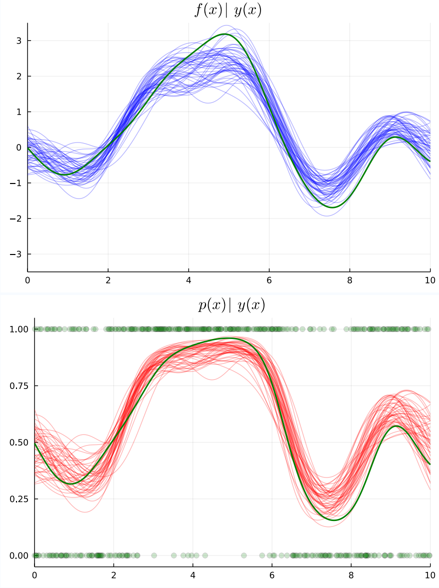
Back to GPs...

Posterior: $p(f \mid y) = p(y \mid f)p(f)/p(y)$



Back to GPs…

Posterior: $p(f \mid y) = p(y \mid f)p(f)/p(y)$ for more data



Outline

✓ Gaussian processes with Gaussian likelihood

✓ What is the likelihood? Connecting observations and Gaussian process prior

3. **Non-Gaussian likelihoods: what happens to the posterior?**
4. How to approximate the intractable
5. Comparison

Posterior

Likelihood

$$p(\mathbf{y} \,|\, \mathbf{f})$$

Joint distribution

$$p(\mathbf{y}, \mathbf{f}) = p(\mathbf{y} \,|\, \mathbf{f}) p(\mathbf{f})$$

Posterior

$$\mathbf{f} \mapsto p(\mathbf{f} \,|\, \mathbf{y}) = \frac{p(\mathbf{y} \,|\, \mathbf{f}) p(\mathbf{f})}{p(\mathbf{y})}$$

$$\mathbf{y} \mapsto \big(\mathbf{f} \mapsto p(\mathbf{f} \,|\, \mathbf{y})\big)$$

Posterior predictions

At new point x^*:

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

At training data:

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

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

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

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

Posterior

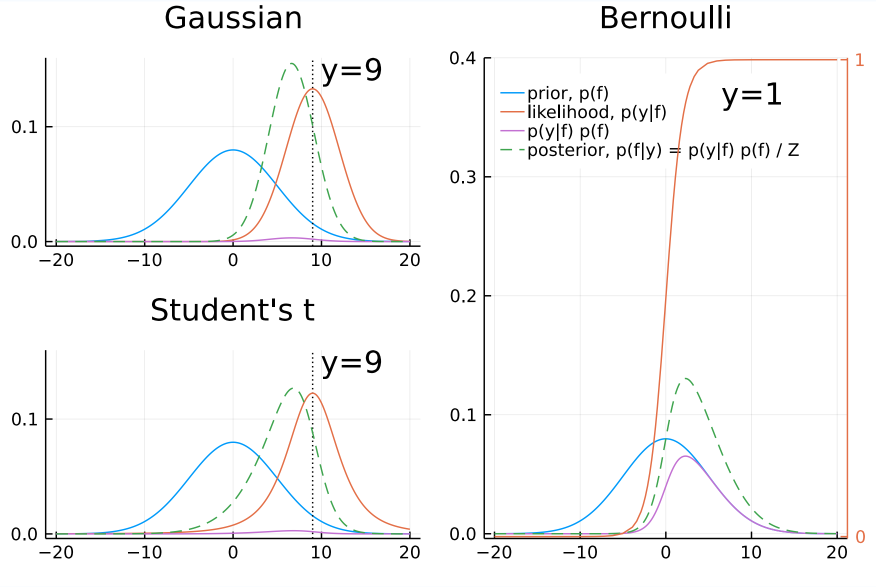
$$p(\mathbf{f} \,|\, \mathbf{y}) = \frac{1}{Z} p(\mathbf{f}) \prod_{n=1}^{N} p(y_n \,|\, f_n)$$

Gaussian (process) prior $p(f(\cdot))$ ... $p(\mathbf{f}) = \mathcal{N}(\mathbf{f} \,|\, \mathbf{0}, \mathrm{K})$

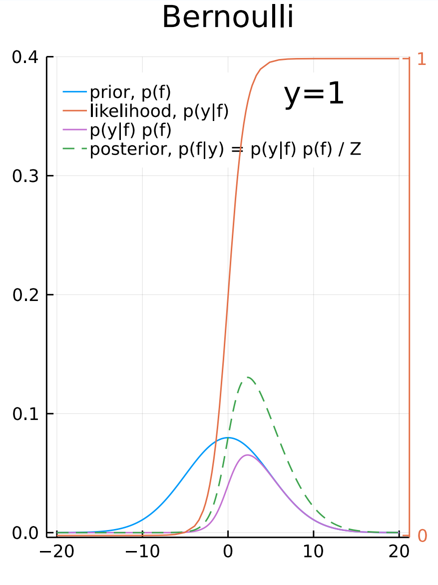
& Gaussian likelihood: conjugate case $\rightarrow$ posterior Gaussian

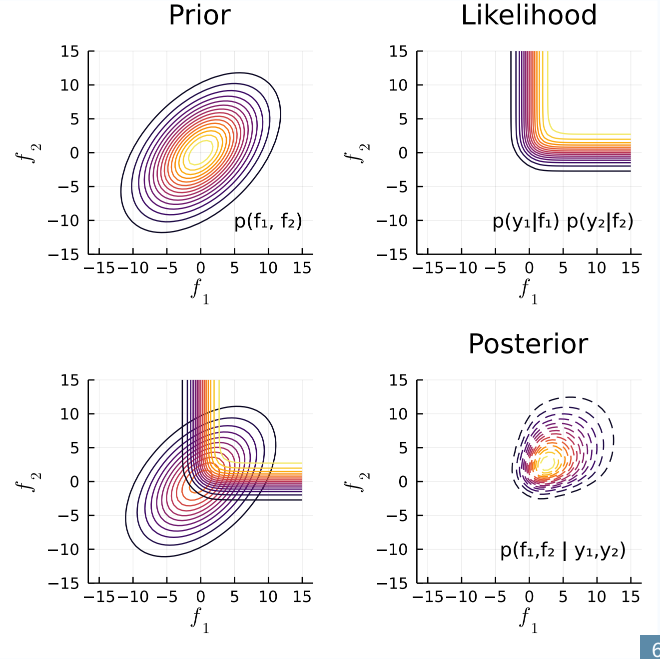
& non-Gaussian $p(y|f)$ $\rightarrow p(\mathbf{f} \,|\, \mathbf{y})$ also non-Gaussian, intractable

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
y=1
prior, p(f)
likelihood, p(y|f)
p(y|f) p(f)
posterior, p(f|y) = p(y|f) p(f) / Z
0.4
0.3
0.2
0.1
0.0
1
0
−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?

Questions?! :)

Outline

✓ Gaussian processes with Gaussian likelihood

✓ What is the likelihood? Connecting observations and Gaussian process prior

✓ Non-Gaussian likelihoods: what happens to the posterior?

4. **How to approximate the intractable**

5. Comparison

Approximations

- Joint model:

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

- Posterior distribution at training points:

$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{p(\mathbf{y})} \approx q(\mathbf{f})$$

- Posterior of f^* for new test point $\mathbf{x}^*$:

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

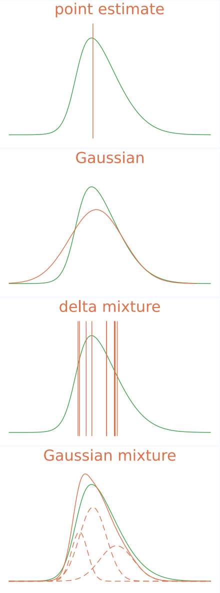
- Predictive distribution

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

Analytically intractable distributions!

Approximating distributions

- delta distribution
  - point estimate
- **Gaussian distribution**
  - Laplace
  - Variational Bayes/Variational Inference (VB / VI)
  - Expectation Propagation (EP)
- mixture of delta distributions
  - Markov Chain Monte Carlo (MCMC)
- mixture of Gaussians
- ...



Gaussian approximations

Approximating the exact posterior with Gaussian

Approximating the posterior at observations:

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

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?

tinyurl.com/nongaussian-inference-viz-v1

Choosing μ and Σ for $q(\mathbf{f})$

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

locally: match mean & variance at point

globally: minimise divergence

Laplace approximation

Variational Bayes (VB)

Expectation Propagation (EP)

Laplace approximation

Laplace approximation

Idea: log of Gaussian pdf = quadratic polynomial

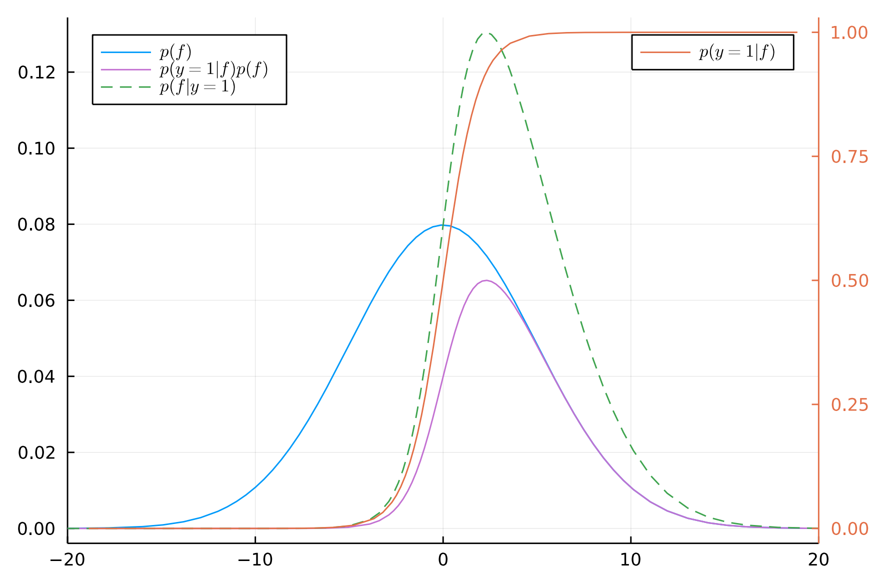
$$p_{\mathcal{N}}(\mathbf{f}) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{f}-\mu)^\top\Sigma^{-1}(\mathbf{f}-\mu)\right)$$

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

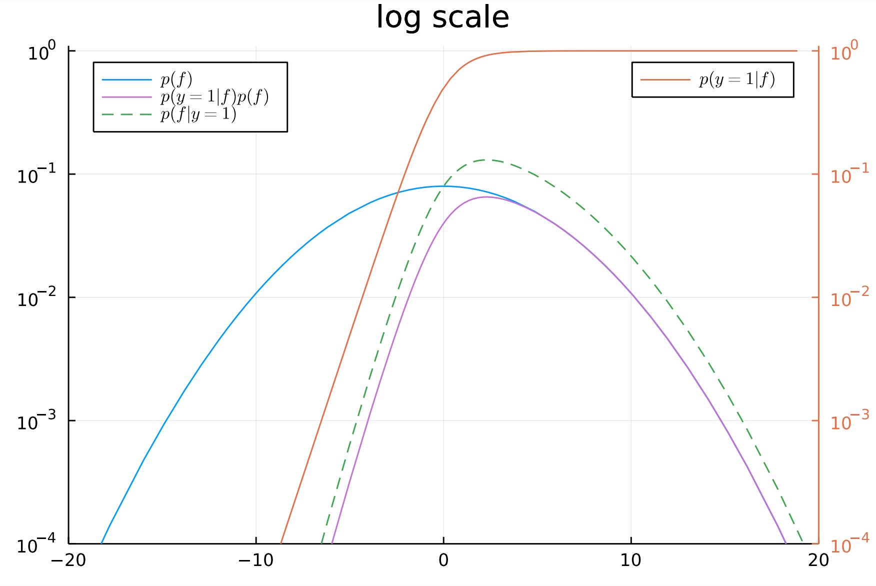
$$g(x+\delta) \approx g(x) + \left(\frac{\mathrm{d}g}{\mathrm{d}x}(x)\right)\delta + \frac{1}{2!}\left(\frac{\mathrm{d}^2 g}{\mathrm{d}x^2}(x)\right)\delta^2$$

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

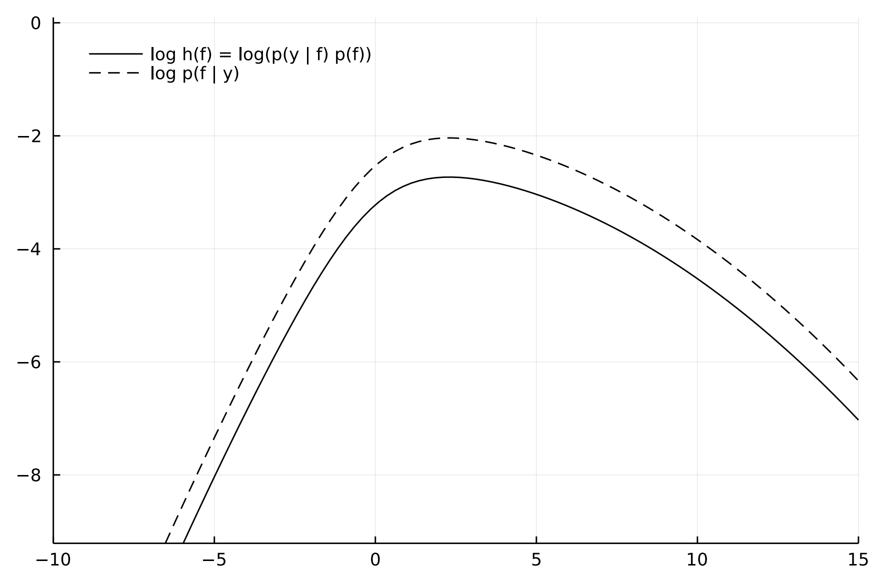
$p(f \mid y) = \frac{1}{Z} p(y \mid f) p(f)$



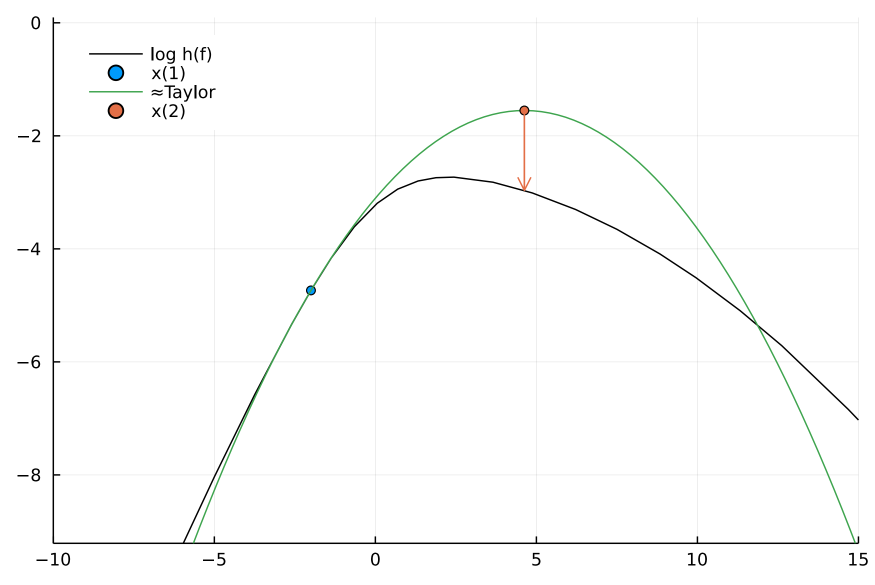
$\log p(f \mid y) = -\log Z + \log p(y \mid f) + \log p(f)$



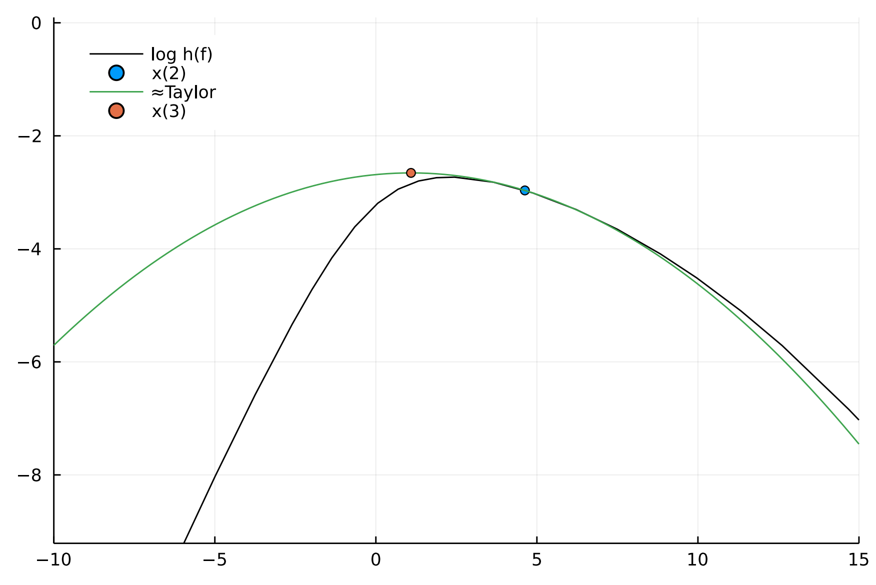
$\log p(f \mid y) = -\log Z + \log h(f)$



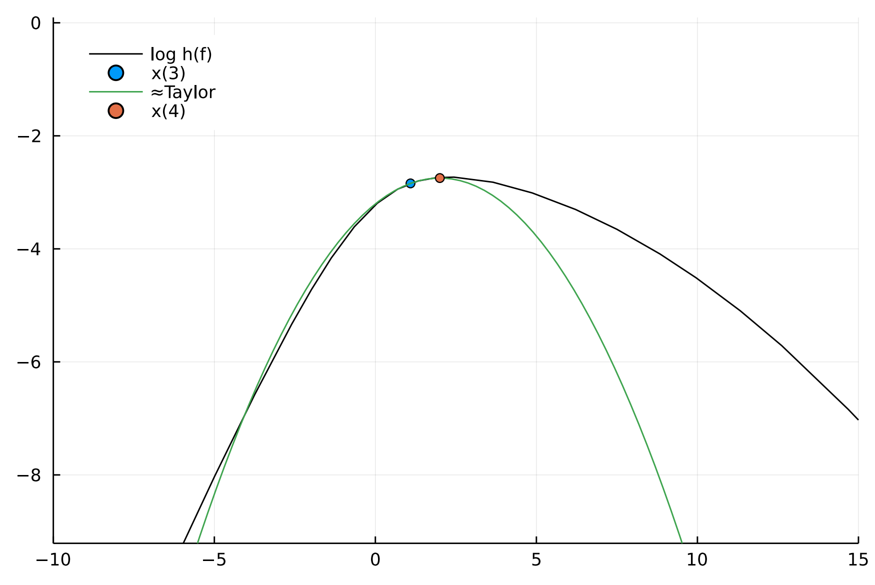
Newton's method



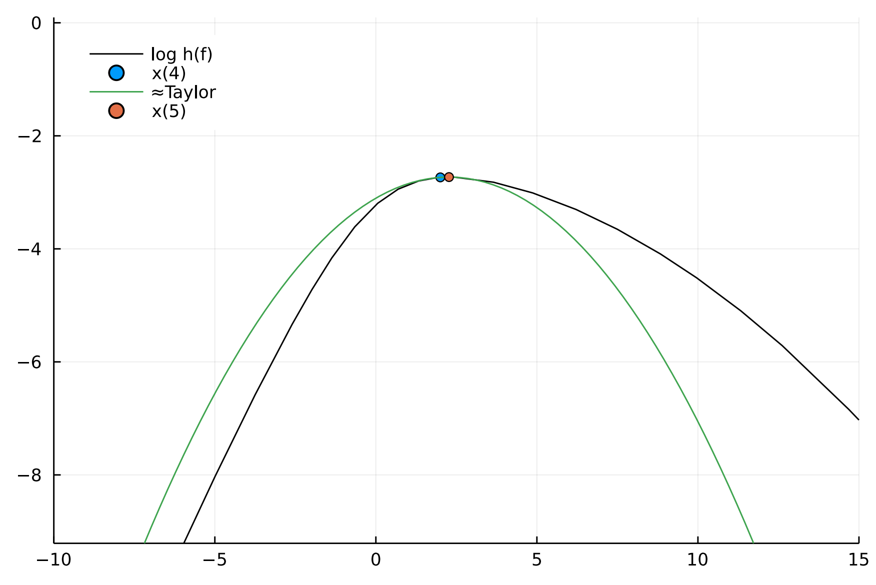
Newton’s method



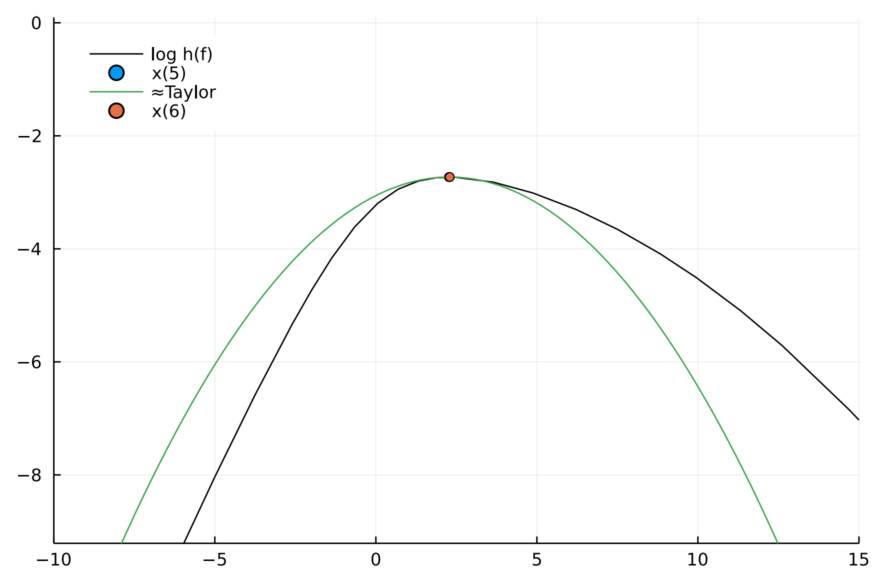
Newton's method



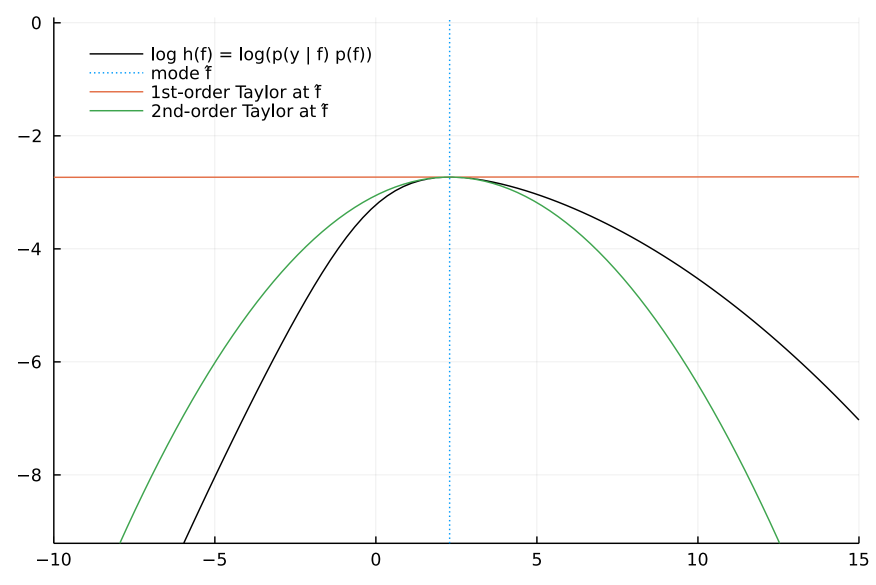
Newton's method



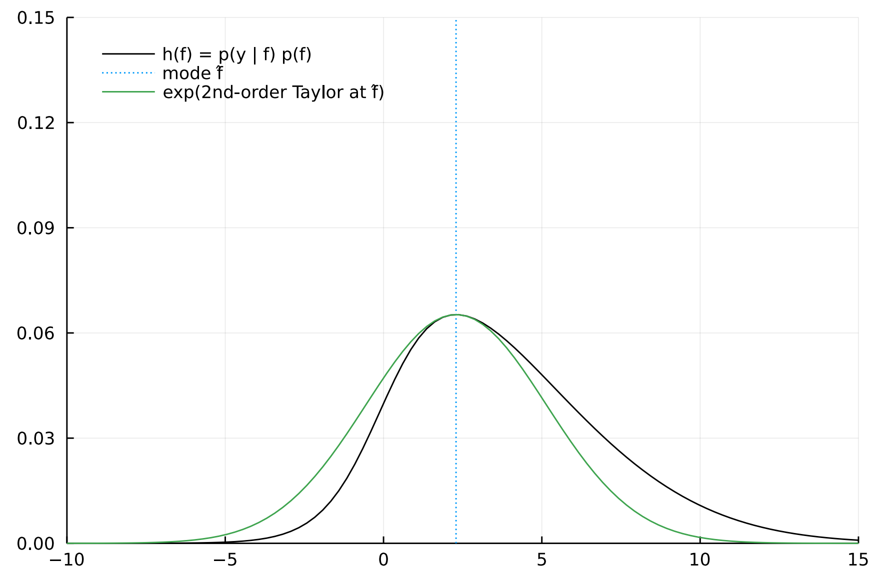
Newton's method



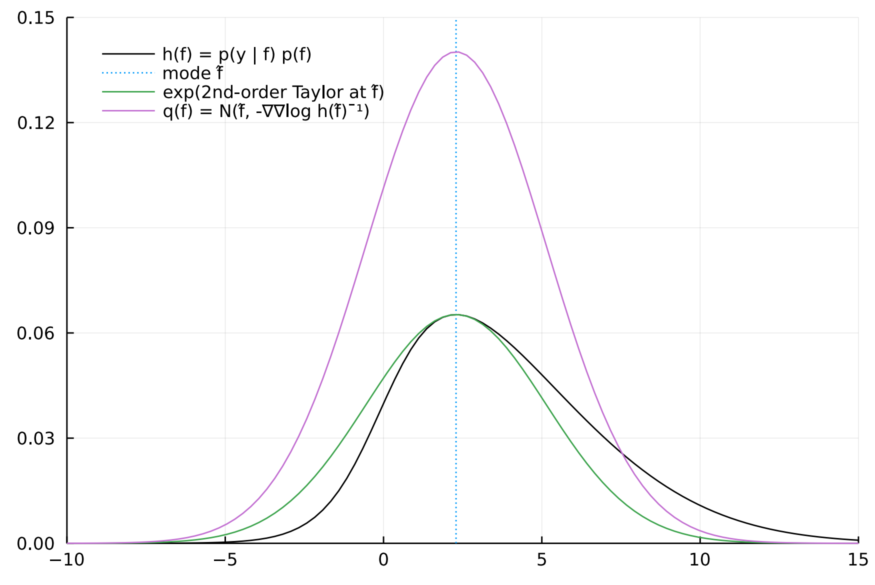
$\log p(f \mid y) + \log Z = \log h(f) \approx \mathcal{O}(f^2)$



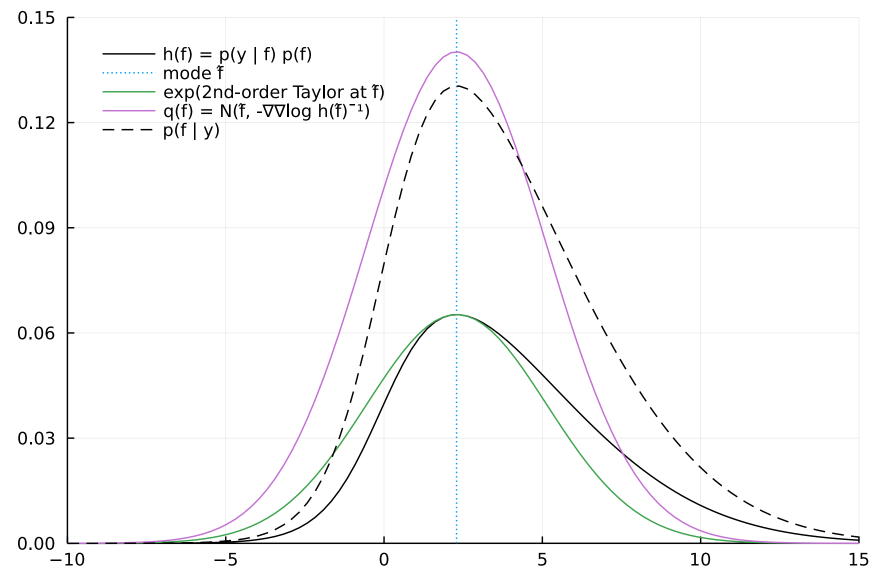
$p(f \mid y)\, Z \approx \exp(\mathcal{O}(f^2))$



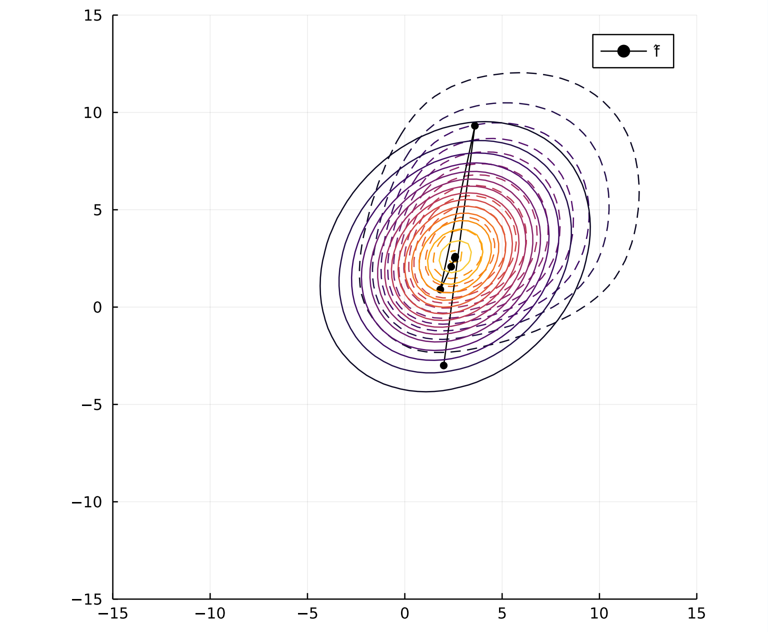
$$p(f \mid y) \approx \mathcal{N}(f \mid \hat{f}, -(\mathrm{d}^2 \log h / \mathrm{d}f^2)^{-1})$$



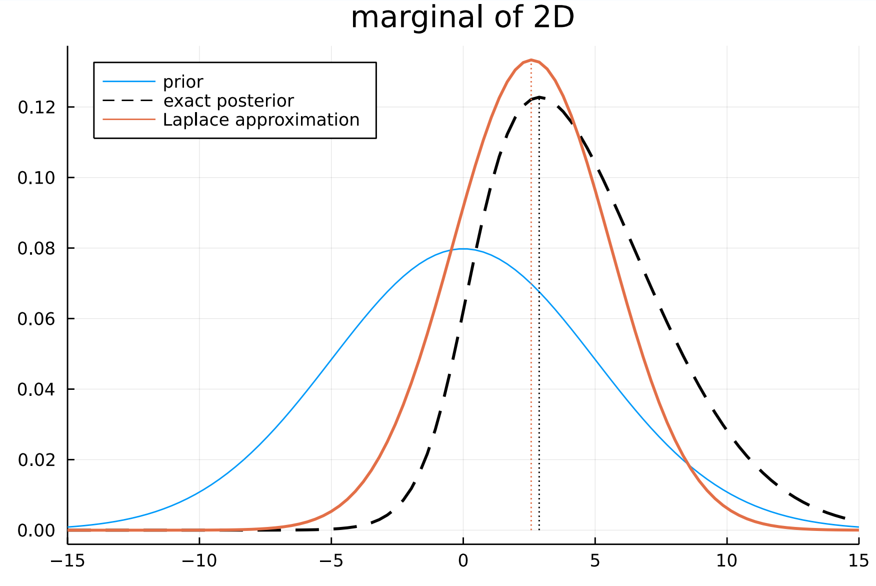
$$p(f \mid y) \approx \mathcal{N}(f \mid \hat{f}, -(\mathrm{d}^2 \log h / \mathrm{d}f^2)^{-1}) = q(f)$$



Laplace in 2D example



Laplace in 2D: marginals



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 [1]

Choosing μ and Σ for $q(\mathbf{f})$

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

locally: match mean & variance at point

globally: minimise divergence

Laplace approximation

Variational Bayes (VB)

Expectation Propagation (EP)

Minimising divergences

Kullback–Leibler (KL) divergence

"Relative entropy", "information gain" *from* q *to* p

$$D_{\mathsf{KL}}(p\|q) = \mathrm{KL}[p(x)\|q(x)] = \mathbb{E}_{x\sim p}\left[\log\frac{p(x)}{q(x)}\right] = \int p(x)\left[\log\frac{p(x)}{q(x)}\right]\mathrm{d}x$$

- non-symmetric: $\mathrm{KL}[p\|q] \neq \mathrm{KL}[q\|p]$
- positive: $\mathrm{KL} \geq 0$ (Gibbs' inequality)
- minimum: $\mathrm{KL}[p\|q] = 0 \Leftrightarrow q = p$.

Demo: KL between two Gaussians

tinyurl.com/nongaussian-inference-viz-v1

Minimising divergences

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

1. $\min \mathrm{KL}[q(\mathbf{f}) \| p(\mathbf{f} \mid \mathbf{y})]$: **Variational Bayes**
2. $\min \mathrm{KL}[p(\mathbf{f} \mid \mathbf{y}) \| q(\mathbf{f})]$: Expectation Propagation

Variational Bayes (VB)
Variational Inference (VI)

Variational inference: the big picture

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

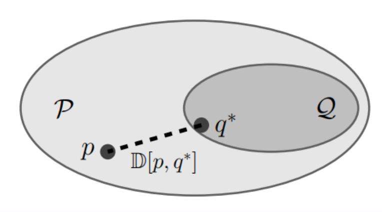
1. Define some "simple" family of distributions $\mathcal{Q}$.
2. Define some way to compute a "distance" $\mathbb{D}[p, q]$ between intractable distribution p and each distribution $q \in \mathcal{Q}$

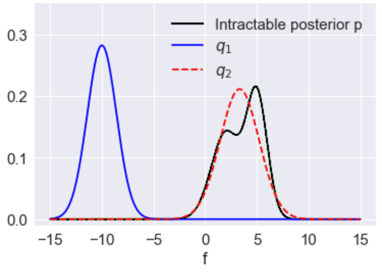
$$\mathbb{D}[p, q_1] > \mathbb{D}[p, q_2]$$

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





Variational Bayes (VB)

$$q(\mathbf{f}) = \mathcal{N}(\mu, \Sigma)$$

$$\underset{\mu,\Sigma}{\operatorname{argmin}} \ \mathrm{KL}\left[q(\mathbf{f}) \| p(\mathbf{f} \mid \mathbf{y})\right]$$

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

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

$$\log p(\mathbf{y}) = \int q(\mathbf{f})\left[\log p(\mathbf{y}\mid\mathbf{f})\right]\mathrm{d}\mathbf{f} - \mathrm{KL}[q(\mathbf{f})\|p(\mathbf{f})] + \mathrm{KL}[q(\mathbf{f})\|p(\mathbf{f}\mid\mathbf{y})]$$

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

$$\log p(\mathbf{y}) = \int q(\mathbf{f}) \left[\log p(\mathbf{y} \mid \mathbf{f}) \right] \mathrm{d}\mathbf{f} - \mathrm{KL}[q(\mathbf{f}) \| p(\mathbf{f})] + \mathrm{KL}[q(\mathbf{f}) \| p(\mathbf{f} \mid \mathbf{y})]$$

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

Lower bound on the (log-)evidence $p(\mathbf{y})$: ELBO

Likelihood term

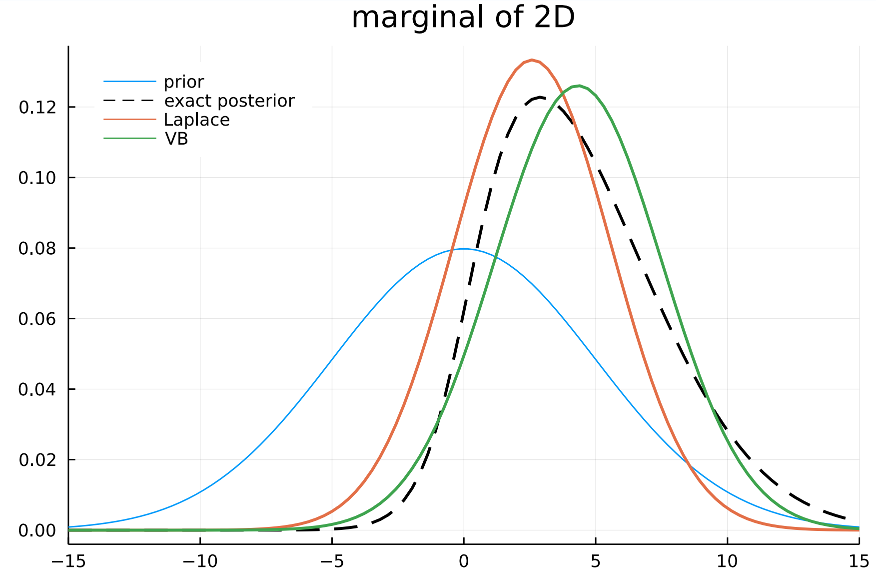
Integral separates for a factorizing likelihood:

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

Evaluating the 1D integrals:

- analytic for some (e.g. Exponential, Gamma, Poisson)
- numerically, for example Gauss–Hermite quadrature
- Monte Carlo (e.g. multi-class classification)

Marginals



marginal of 2D
prior
exact posterior
Laplace
VB
0.00
0.02
0.04
0.06
0.08
0.10
0.12
−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)

\+ minimises a true lower bound $\rightarrow$ convergence

– underestimates variance

Minimising divergences

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

✓ $\min \mathrm{KL}[q(\mathbf{f}) \| p(\mathbf{f} \mid \mathbf{y})]$: Variational Bayes

2. $\min \mathrm{KL}[p(\mathbf{f} \mid \mathbf{y}) \| q(\mathbf{f})]$: **Expectation Propagation**

Expectation Propagation (EP)

Expectation Propagation

Can we minimise KL divergence in the “right” direction?

$$q(\mathbf{f}) = \underset{\mu,\Sigma}{\operatorname{argmin}}\ \mathrm{KL}\left[p(\mathbf{f}\,|\,\mathbf{y}) \| q(\mathbf{f})\right]$$

Exact posterior:

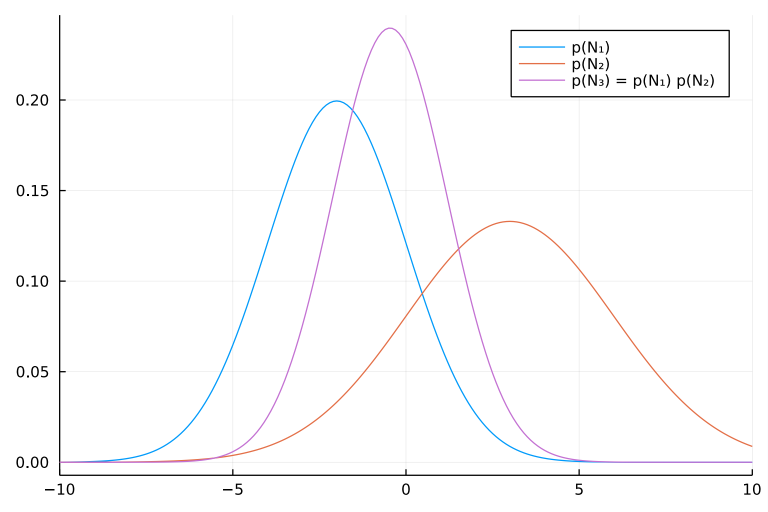
$$p(\mathbf{f}\,|\,\mathbf{y}) \propto p(\mathbf{f}) \prod_{n=1}^{N} p(y_n\,|\,f_n)$$

Approximate posterior:

$$q(\mathbf{f}) \propto p(\mathbf{f}) \prod_{n=1}^{N} t_n(f_n)$$

$$t_n = Z_n \mathcal{N}(f_n\,|\,\tilde{\mu}_n, \tilde{\sigma}_n^2)$$

Multiplying and dividing Gaussians



Adding and subtracting natural (canonical) parameters

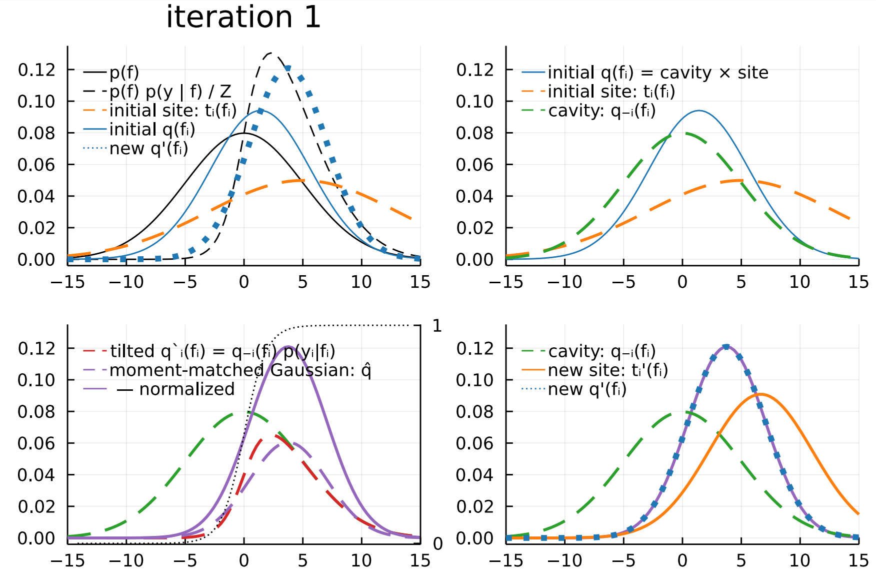
Expectation Propagation iterations

$$\text{“}\min \mathrm{KL}[p(\mathbf{f}\,|\,\mathbf{y})\|q(\mathbf{f})]\text{”} \qquad\qquad q(\mathbf{f}) \propto p(\mathbf{f}) \prod_{n=1}^{N} \underbrace{t_n(f_n)}_{\text{site} \,\propto \mathcal{N}(f_n)}$$

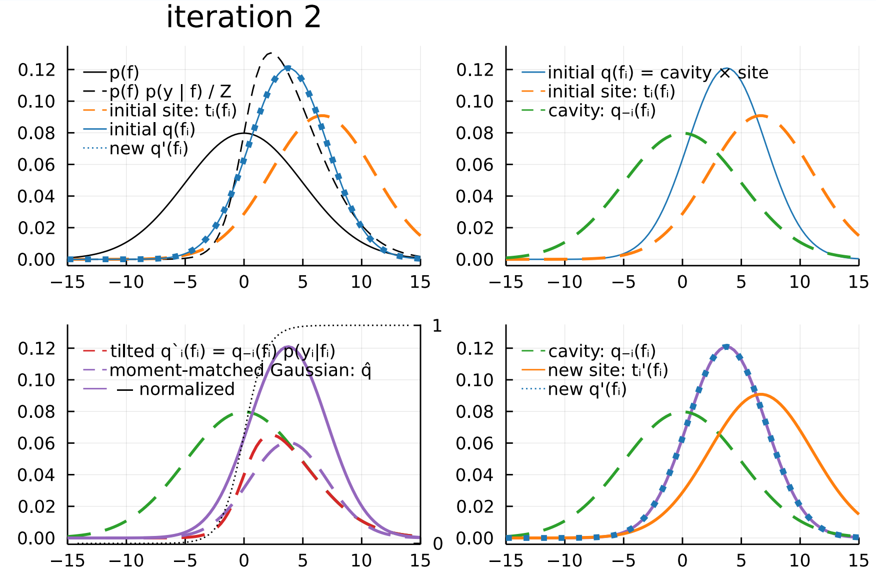
For each site n:

1. marginalize $\int q(\mathbf{f})\, \mathrm{d}\{f_{j\neq n}\} = q(f_n) \quad \not\propto t_n(f_n)$
2. improve local approximation: $\min \mathrm{KL}[q(f_n)\frac{p(y_n\,|\,f_n)}{t_n(f_n)} \| q(f_n)\frac{t'_n(f_n)}{t_n(f_n)}]$
   - 2.1 *cavity* distribution $q_{-n}(f_n) = \frac{q(f_n)}{t_n(f_n)} \quad \Leftrightarrow \quad q(f_n) = q_{-n}(f_n) t_n(f_n)$
   - 2.2 *tilted* distribution $q_{\backslash n}(f_n) = q_{-n}(f_n) p(y_n\,|\,f_n)$
   - 2.3 $\operatorname{argmin} \mathrm{KL}[q_{-n}(f_n)p(y_n\,|\,f_n)\|\hat{q}]$ by moment-matching
   - 2.4 update site: $t'_n(f_n) = \frac{\hat{q}}{q_{-n}(f_n)} \quad \Leftrightarrow \quad \hat{q} = q_{-n}(f_n)\, t'_n(f_n)$
3. compute new $q'(\mathbf{f})$ (rank-1 update)

Expectation Propagation in 1D



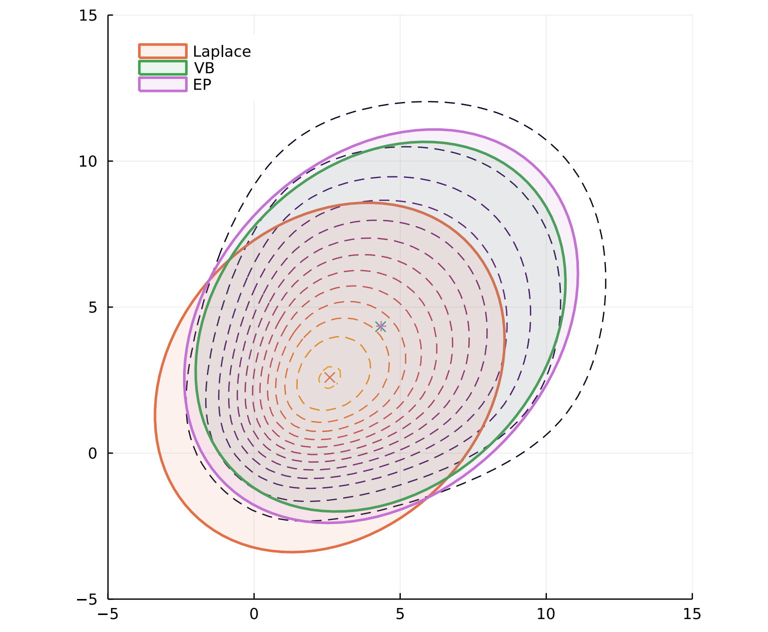
Expectation Propagation in 1D



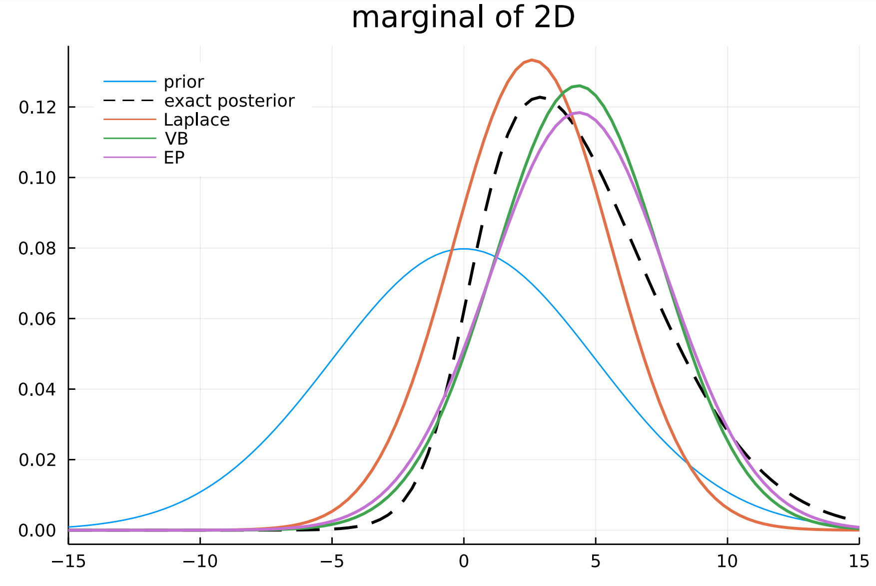
Demo: EP in 2D

tinyurl.com/nongaussian-inference-viz-v1

Comparison 2D



Marginals


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

Expectation Propagation: important properties

- multiple passes required to converge
- moment-matching (e.g. covering multiple modes)
- + effective for classification
- – not guaranteed to converge
- – updates may be invalid (non-log-concave likelihoods) [2]

Outline

✓ Gaussian processes with Gaussian likelihood

✓ What is the likelihood? Connecting observations and Gaussian process prior

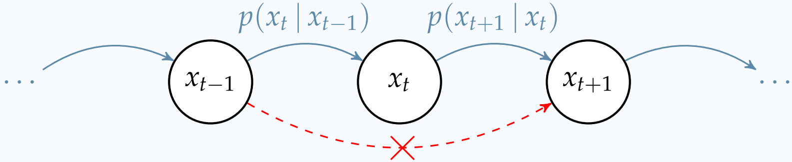
✓ Non-Gaussian likelihoods: what happens to the posterior?

4. **How to approximate the intractable**
    - ✓ with Gaussians
        - Laplace
        - Variational Bayes
        - Expectation Propagation
    - 4.2 **with samples: MCMC**

5. Comparison

Markov Chain Monte Carlo

Markov Chain



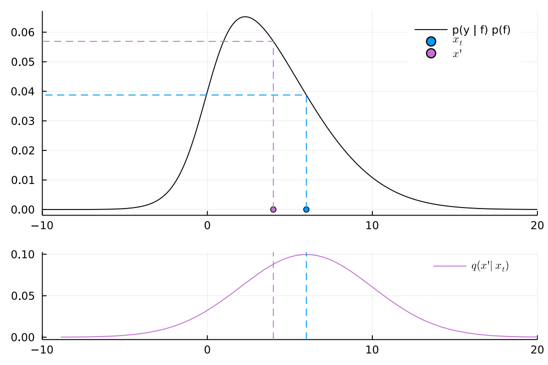
- Samples $x_1, \dots, x_T$
- "Markov" = 1-step history
- $x_{t+1} \sim p(x_{t+1} \mid x_t)$, independent of $x_{t-1}, \dots, x_1$

Markov Chain Monte Carlo (MCMC)

Generate samples $\{x_t\} \sim p(f \mid y)$

Requires:

- *unnormalized* posterior
  $h(f) = p(y \mid f) p(f)$
- Markov proposal $q(x' \mid x_t)$
- initial x_0



In each iteration t:

1. Random proposal $x' \sim q(x' \mid x_t)$
2. Acceptance probability $\frac{h(x')}{h(x_t)}$ → ensures sampling from $p(f \mid y)$

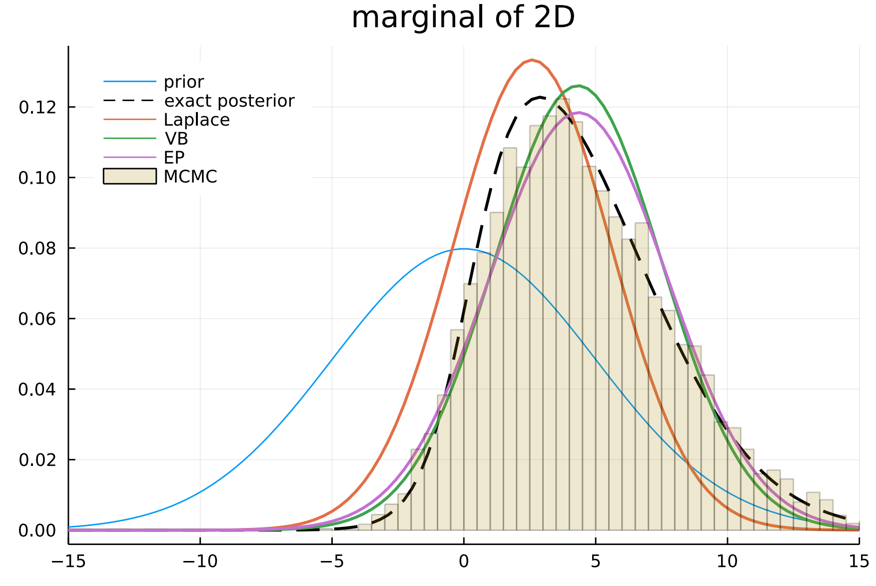
   accept: $x_{t+1} = x'$ reject: copy $x_{t+1} = x_t$

   $h(x') > h(x_t)$: always accepts → climbs uphill

Demo: MCMC in 2D

tinyurl.com/nongaussian-inference-viz-v1

Marginals


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

MCMC: important properties

- burn-in
- acceptance ratio
- auto-correlation, effective sample size (ESS); thinning to save memory
- mixing and multiple chains ($\hat{R}$)
- better proposals (HMC, NUTS) $\rightarrow$ use robust implementations

+ very accurate (gold-standard)

– very slow, predictions require keeping all (thinned) samples around

Michael Betancourt's `betanalpha.github.io/writing/`

MCMC: robust implementations

- Stan
- PyMC3
- Pyro & NumPyro
- TensorFlow Probability (GPflow)
- Turing.jl

Outline

- ✓ Gaussian processes with Gaussian likelihood
- ✓ What is the likelihood? Connecting observations and Gaussian process prior
- ✓ Non-Gaussian likelihoods: what happens to the posterior?
- ✓ How to approximate the intractable
  - ✓ with Gaussians
    - Laplace
    - Variational Bayes
    - Expectation Propagation
  - ✓ with samples: MCMC

5. **Comparison**

Comparison

Comparison

MCMC

- samples
- gold standard
- slow

Laplace

- $\mathcal{N}$ = curvature at mode
- simple & fast
- often poor approximation

Variational Bayes

- $\mathcal{N}$ minimises $\mathrm{KL}[q(\mathbf{f})\|p(\mathbf{f}\mid\mathbf{y})]$
- principled, any likelihood
- underestimates variance

Expectation Propagation

- $\mathcal{N}$ matches marginal moments
- good calibration in classification
- may not converge

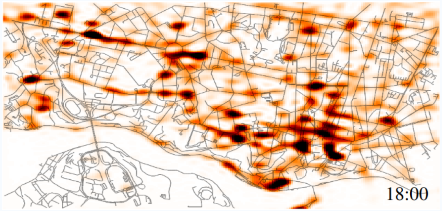
What we did not cover…

- More complex likelihoods (heteroskedastic, zero-inflated, multi-stage…)
- Marginal likelihood approximations for hyperparameter learning [3]
- How parametrisation affects Gaussianity of $p(\mathbf{f} \mid \mathbf{y})$
- Connections between EP and VB (“PowerEP”, CVI dual parameterization) [4, 5]
- Other divergences, generalised VI, …
- Combinations of MCMC and variational methods
- Augmenting likelihood with auxiliary variable
  $\rightarrow$ conditionally conjugate model [6]

Take-away

We can...

- create **richer models** with likelihoods beyond the Gaussian
- **learn latent functions** that form the connection between data points
- handle the non-Gaussian posterior with **approximations**
- **trade off** speed, accuracy, and ease-of-use



References I

Marcelo Hartmann and Jarno Vanhatalo.
Laplace approximation and natural gradient for Gaussian process regression with heteroscedastic student-t model.
Statistics and Computing, 29(4):753–773, October 2018.

Pasi Jylänki, Jarno Vanhatalo, and Aki Vehtari.
Robust Gaussian process regression with a student-t likelihood.
Journal of Machine Learning Research, 12(99):3227–3257, 2011.

Hannes Nickisch and Carl Edward Rasmussen.
Approximations for binary Gaussian process classification.
Journal of Machine Learning Research, 9(67):2035–2078, 2008.

Thang D. Bui, Josiah Yan, and Richard E. Turner.
A unifying framework for Gaussian process pseudo-point approximations using Power Expectation Propagation.
Journal of Machine Learning Research, 18(104):1–72, 2017.

References II

Vincent Adam, Paul Chang, Mohammad Emtiyaz E Khan, and Arno Solin.
Dual parameterization of sparse variational gaussian processes.
In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 11474–11486, 2021.

Théo Galy-Fajou, Florian Wenzel, and Manfred Opper.
Automated augmented conjugate inference for non-conjugate Gaussian process models, 2020.

James Hensman, Nicolo Fusi, and Neil D. Lawrence.
Gaussian processes for big data.
UAI, 2013.

Malte Kuss and Carl Edward Rasmussen.
Assessing approximate inference for binary Gaussian process classification.
Journal of Machine Learning Research, 6(57):1679–1704, 2005.

Alan Saul.
Gaussian process based approaches for survival analysis, 2017.

References III

Aki Vehtari, Andrew Gelman, Tuomas Sivula, Pasi Jylänki, Dustin Tran, Swupnil Sahai, Paul Blomstedt, John P. Cunningham, David Schiminovich, and Christian P. Robert.
Expectation Propagation as a way of life: A framework for Bayesian inference on partitioned data.
Journal of Machine Learning Research, 21(17):1–53, 2020.

Will Penny.
Bayesian inference course: Variational inference, 2013.