Non-Gaussian likelihoods for Gaussian Processes

Alan Saul

PROWLER.IO

Outline

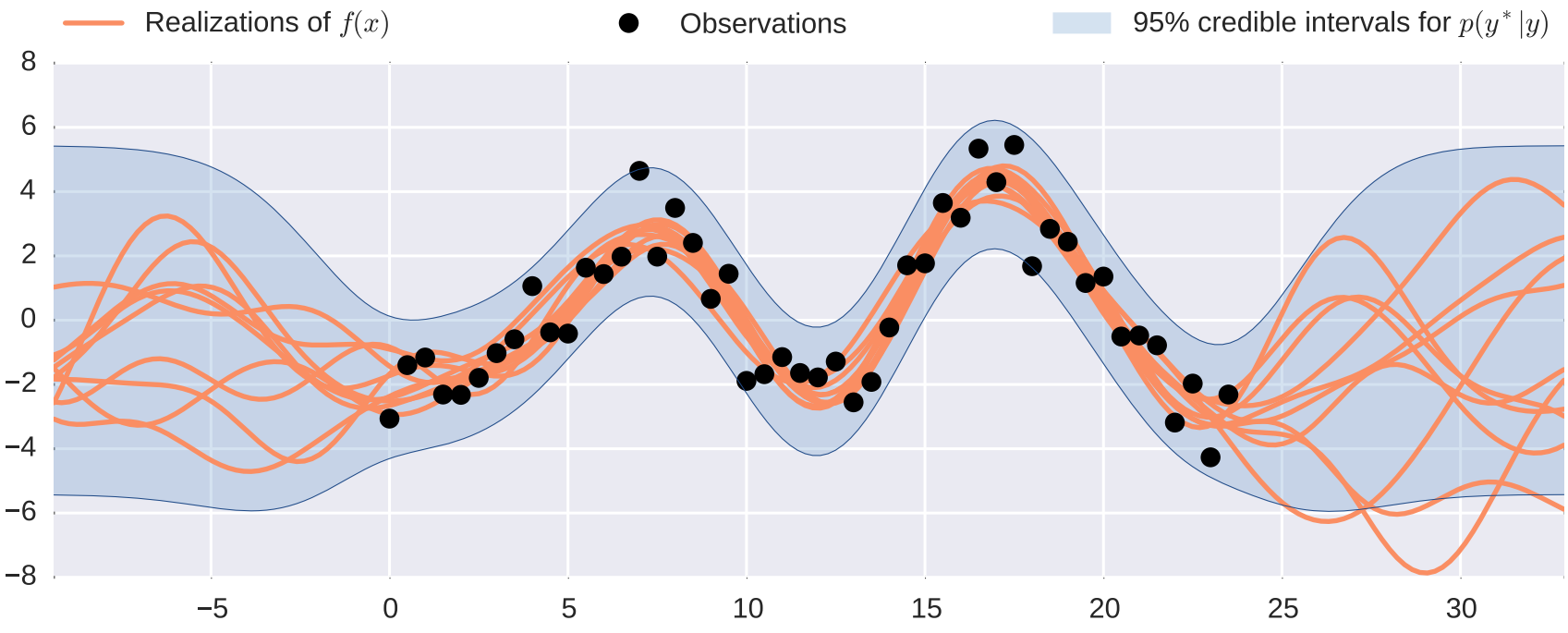
- Motivation
- Laplace approximation
- KL method
- Expectation Propagation
- Comparing approximations

GP regression - recap so far

Model the observations as a distorted version of the process $\mathbf{f}_i = f(\mathbf{x}_i)$:

$$\mathbf{y}_i \sim \mathcal{N}\left(f(\mathbf{x}_i), \sigma^2\right)$$

f is a non-linear function, in our case we assume it is latent, and is assigned a Gaussian process prior.



GP regression setting

So far we have assumed that the latent values, $\mathbf{f}$, have been corrupted by Gaussian noise. Everything remains analytically tractable.

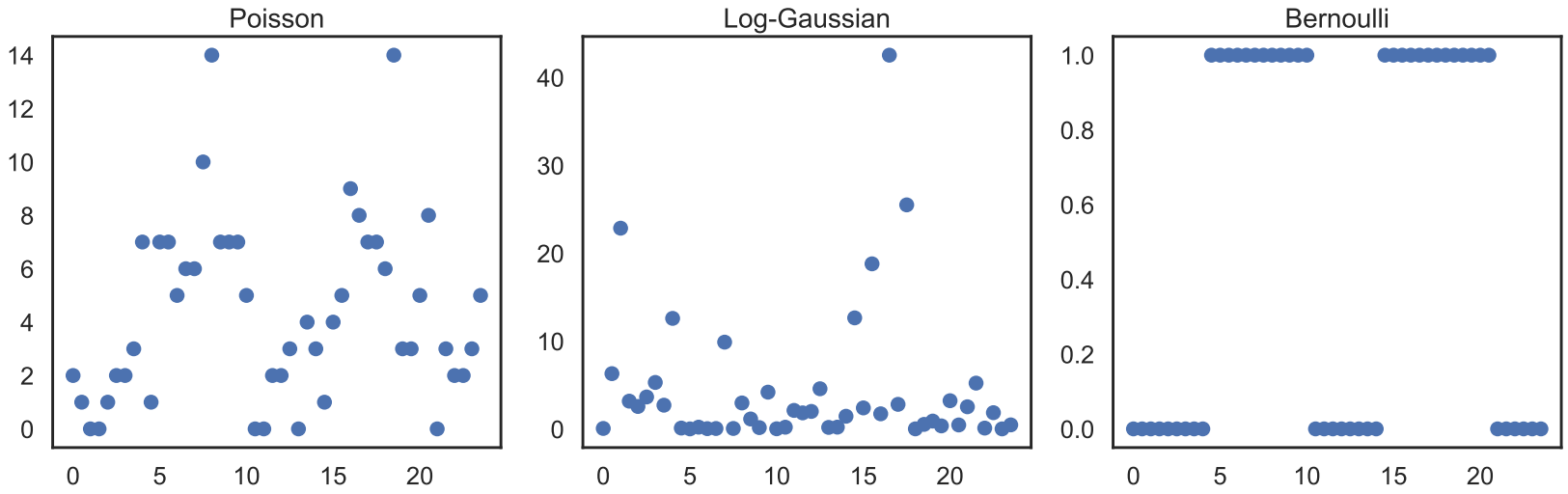
Gaussian Prior: $$\mathbf{f} \sim \mathcal{GP}(\mathbf{0}, \mathbf{K_{ff}}) = p(\mathbf{f})$$

Gaussian likelihood: $$\mathbf{y} \sim \mathcal{N}\left(\mathbf{f}, \sigma^2\mathbf{I}\right) = \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{f}_i)$$

Gaussian posterior: $$p(\mathbf{f}|\mathbf{y}) \propto \mathcal{N}\left(\mathbf{y}|\mathbf{f}, \sigma^2\mathbf{I}\right)\mathcal{N}\left(\mathbf{f}|\mathbf{0}, \mathbf{K_{ff}}\right)$$

Motivation

- You have been given some data you wish to model.
- You believe that the observations are connected through some underlying unknown function.
- You know from your understanding of the data generation process, that the observations are not Gaussian.
- You still want to learn, as best as possible, what is the unknown function being used, and make predictions.

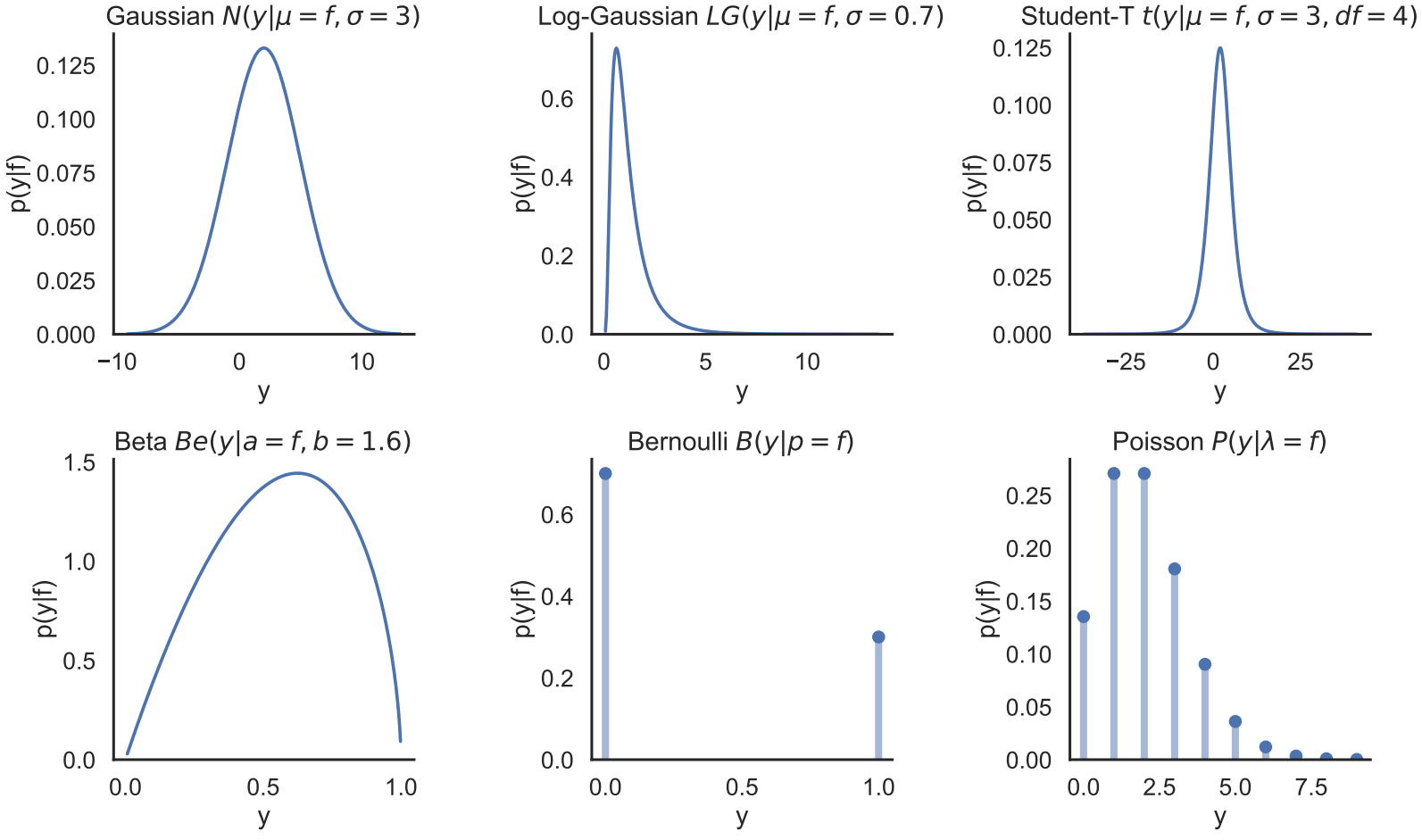


Likelihood

- $p(\mathbf{y}|\mathbf{f})$ is the probability that we would see some random variables, $\mathbf{y}$, if we knew the latent function values $\mathbf{f}$, which act as parameters.
- Given the observed values for $\mathbf{y}$ are fixed, it can also be seen as the likelihood that some latent function values, $\mathbf{f}$, would give rise to the observed values of $\mathbf{y}$. Note this is a *function* of $\mathbf{f}$, and doesn't integrate to 1 in $\mathbf{f}$.
- So far assumed that the distortion of the underlying latent function, $\mathbf{f}$, that gives rise to the observed data, $\mathbf{y}$, is independent and normally distributed.
- This is often not the case, count data, binary data, etc.

Likelihood

$p(y|f)$ with fixed f



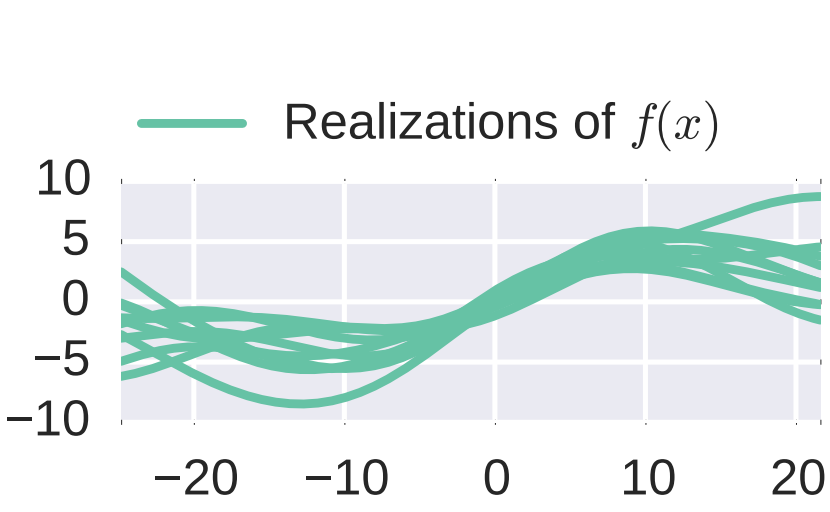
Likelihood

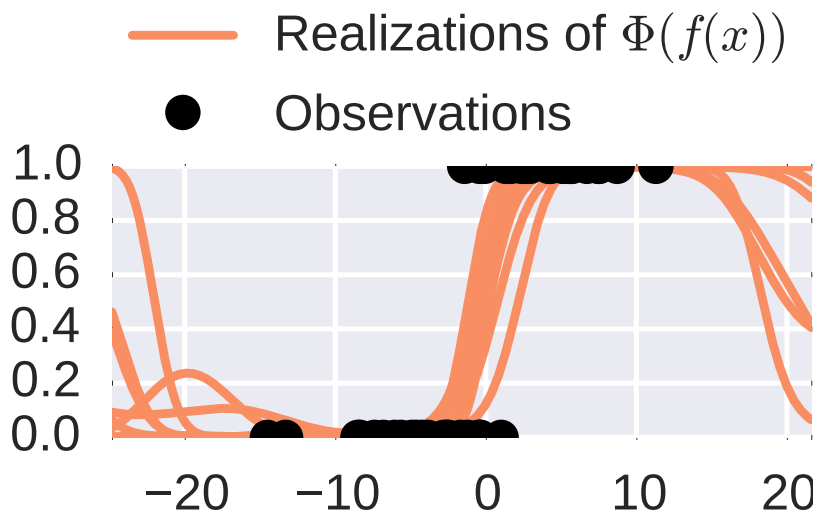
$p(y|f)$ as a function of f MAKE PLOT OF LIKELIHOOD AS A FUNCTION OF F, SHOULD SHOW IT NOT NORMALISING

Binary example

- Binary outcomes for $\mathbf{y}_i$, $\mathbf{y}_i \in [0, 1]$.
- Model the probability of $\mathbf{y}_i = 1$ with transformation of GP, with Bernoulli likelihood.
- Probability of 1 must be between 0 and 1, thus use squashing transformation, $\lambda(\mathbf{f}_i) = \Phi(\mathbf{f}_i)$.

$$p(\mathbf{y}_i|\lambda(\mathbf{f}_i)) = \begin{cases} \lambda(\mathbf{f}_i), & \text{if } \mathbf{y}_i = 1 \\ 1 - \lambda(\mathbf{f}_i), & \text{if } \mathbf{y}_i = 0 \end{cases}$$

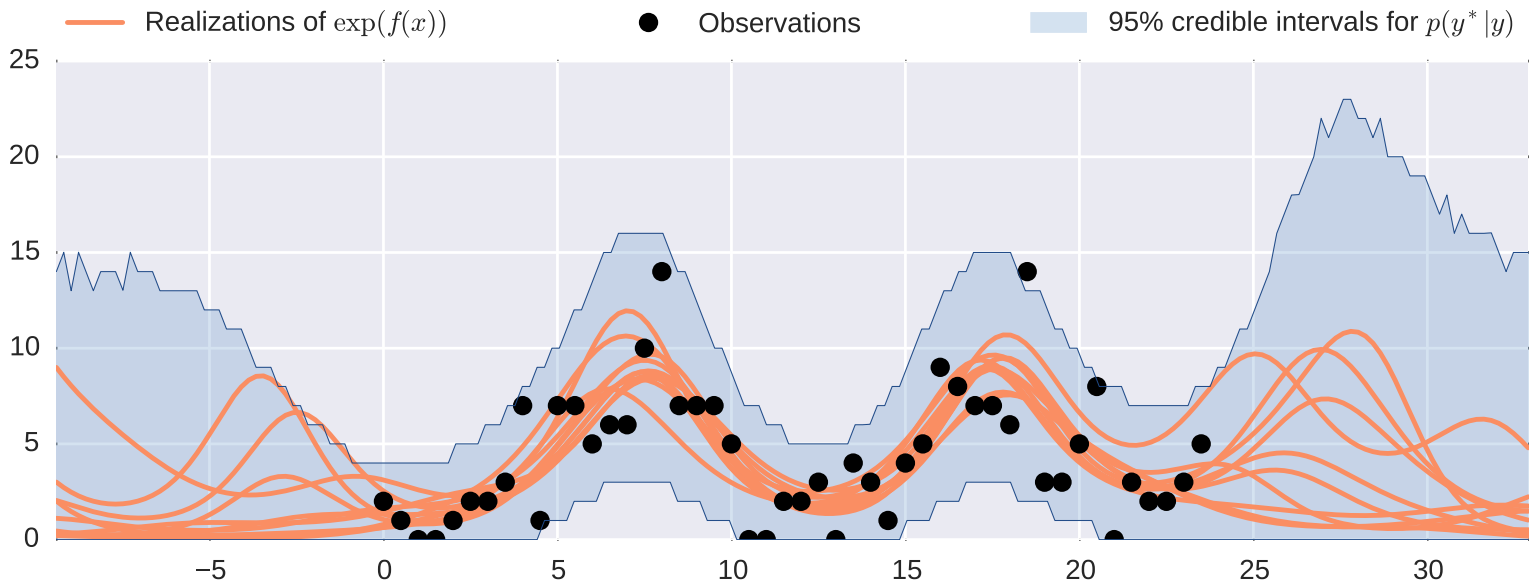




Count data example

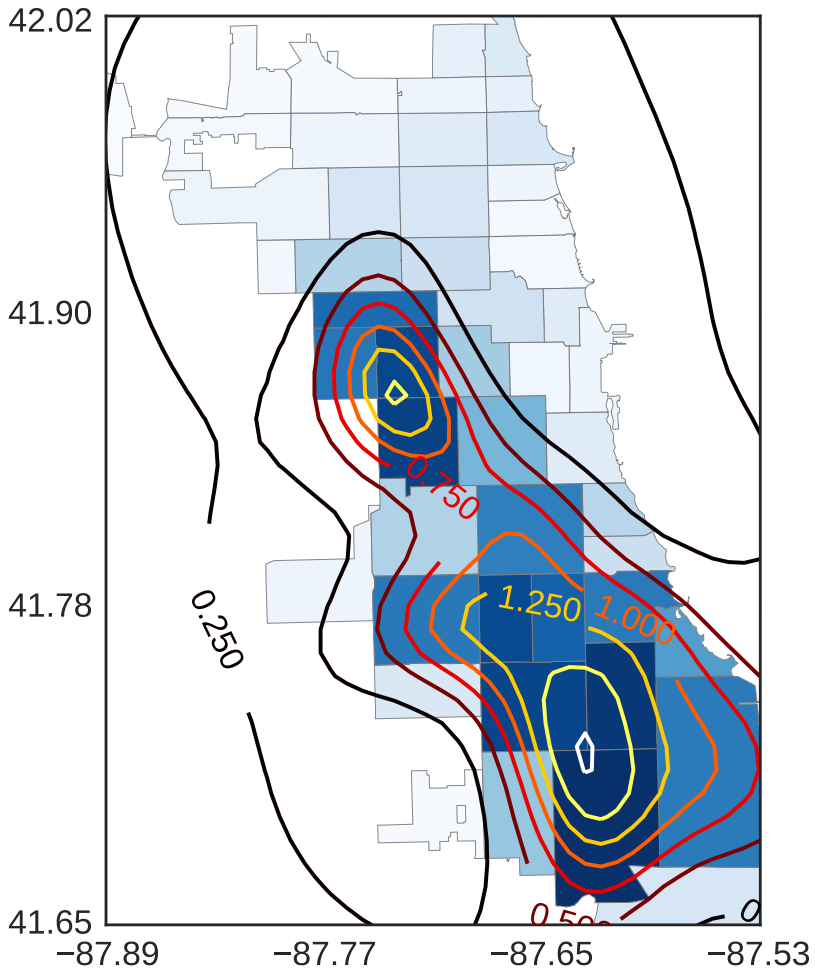
- Non-negative and discrete values only for $\mathbf{y}_i$, $\mathbf{y}_i \in \mathbb{N}$.
- Model the *rate* or *intensity*, λ, of events with a transformation of a Gaussian process.
- Rate parameter must remain positive, use transformation to maintain positiveness $\lambda(\mathbf{f}_i) = \exp(\mathbf{f}_i)$ or $\lambda(\mathbf{f}_i) = \mathbf{f}_i^2$

$$\mathbf{y}_i \sim \text{Poisson}(\mathbf{y}_i|\lambda_i = \lambda(\mathbf{f}_i)) \qquad \text{Poisson}(\mathbf{y}_i|\lambda_i) = \frac{\lambda_i^{\mathbf{y}_i}}{!\mathbf{y}_i} e^{-\lambda_i}$$



Application example

- Chicago crime counts.
- Same Poisson likelihood.
- 2D-input to kernel.



Application example

MORE COMPLEX MOTIVATING APPLICATION. MNIST MULTICLASS

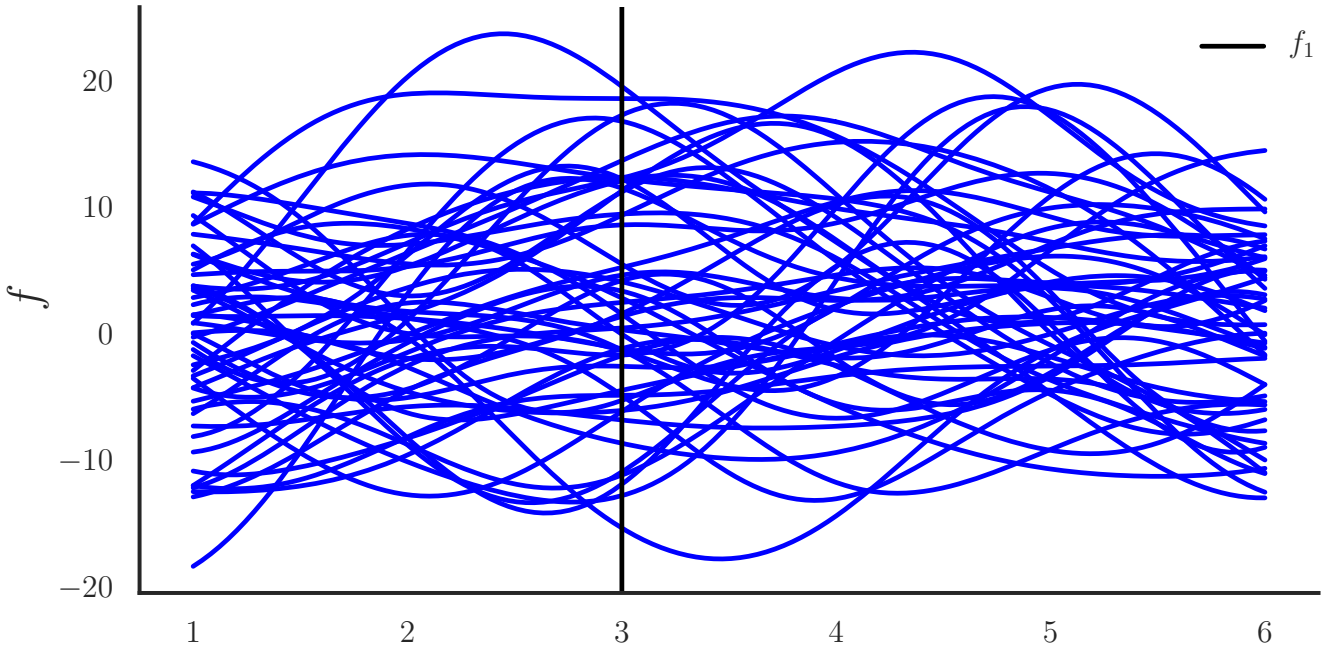
Non-Gaussian posteriors

- Exact computation of posterior is no longer analytically tractable due to non-conjugate Gaussian process prior to non-Gaussian likelihood, $p(\mathbf{y}|\mathbf{f})$.

$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{f}) \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{f}_i)}{\int p(\mathbf{f}) \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{f}_i)\, d\mathbf{f}}$$

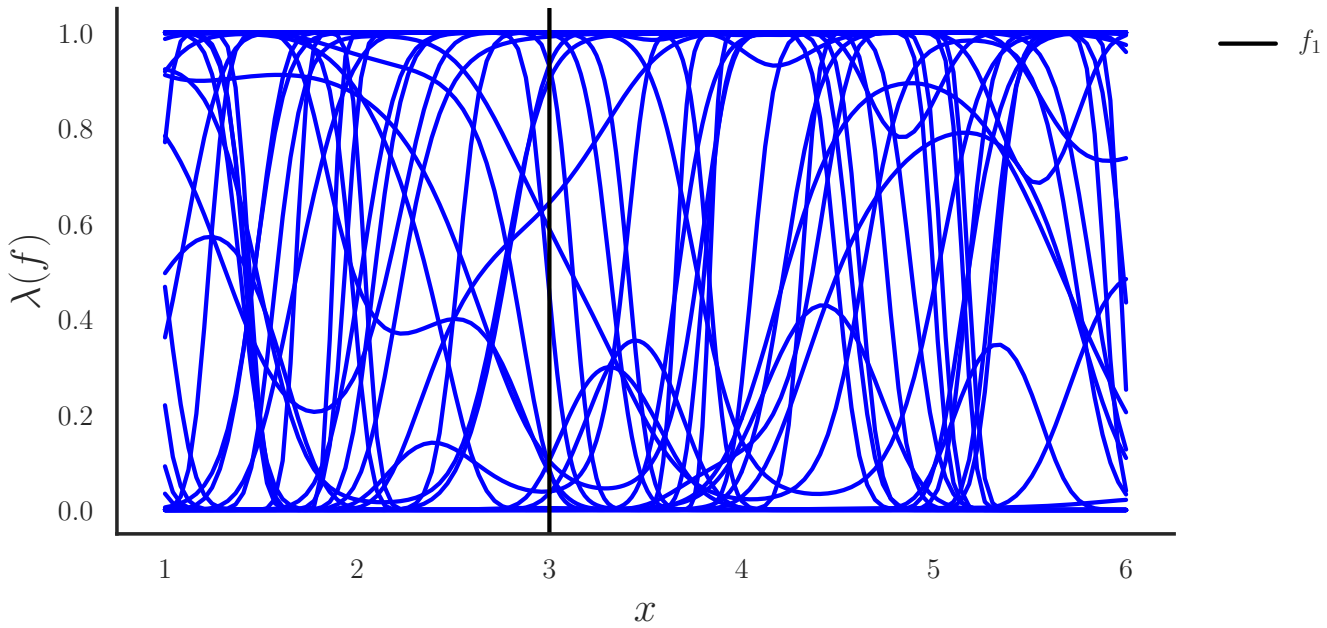
Non-Gaussian posteriors illustrated

- Why is it so difficult?
- Consider one observation, $y_1 = 1$, at input x_1.
- Can normalise easily with numerical integration, $\int p(y_1 = 1|f_1)p(f_1)df_1$.



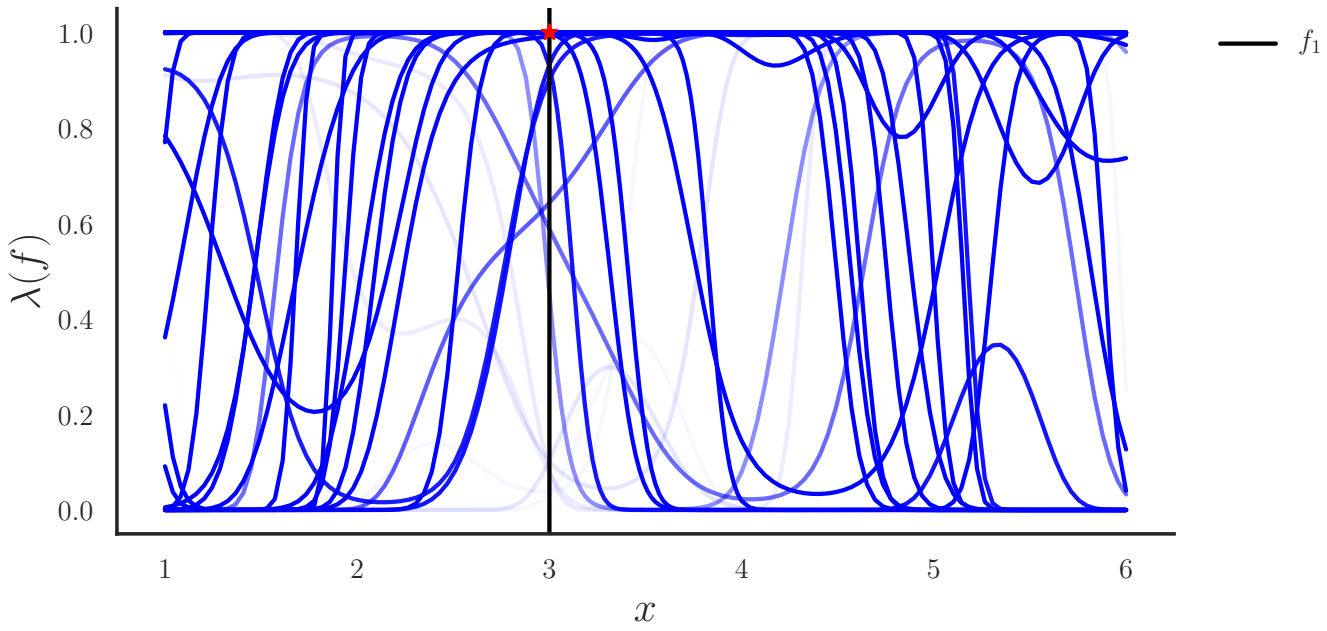
Non-Gaussian posteriors illustrated

- Why is it so difficult?
- Consider one observation, $y_1 = 1$, at input x_1.
- Can normalise easily with numerical integration, $\int p(y_1 = 1|f_1)p(f_1)df_1$.

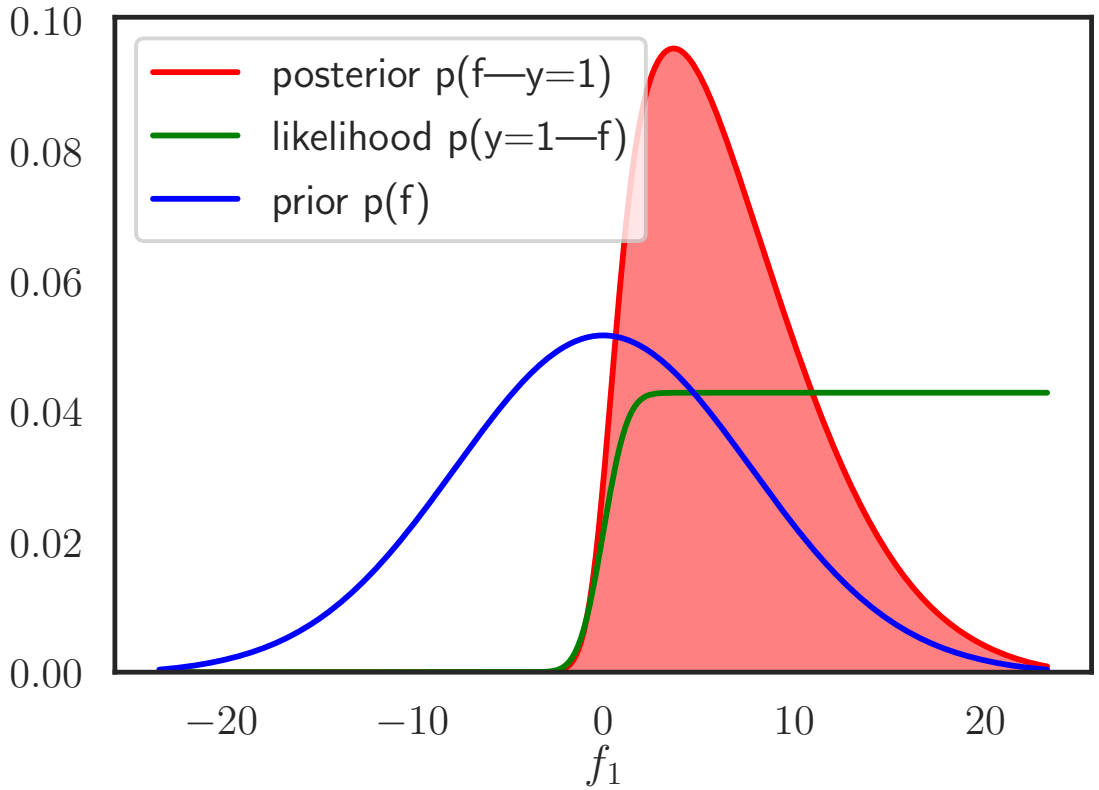


Non-Gaussian posteriors illustrated

- Why is it so difficult?
- Consider one observation, $y_1 = 1$, at input x_1.
- Can normalise easily with numerical integration, $\int p(y_1 = 1|f_1)p(f_1)df_1$.

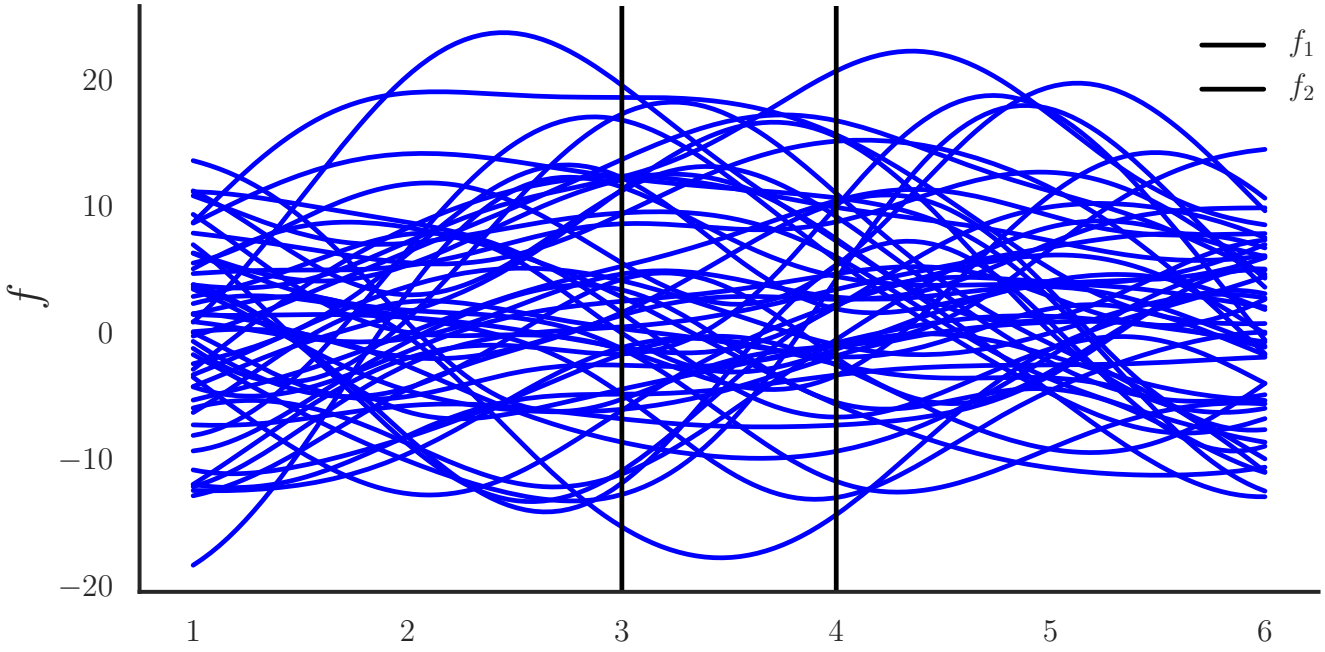


Non-Gaussian posteriors illustrated



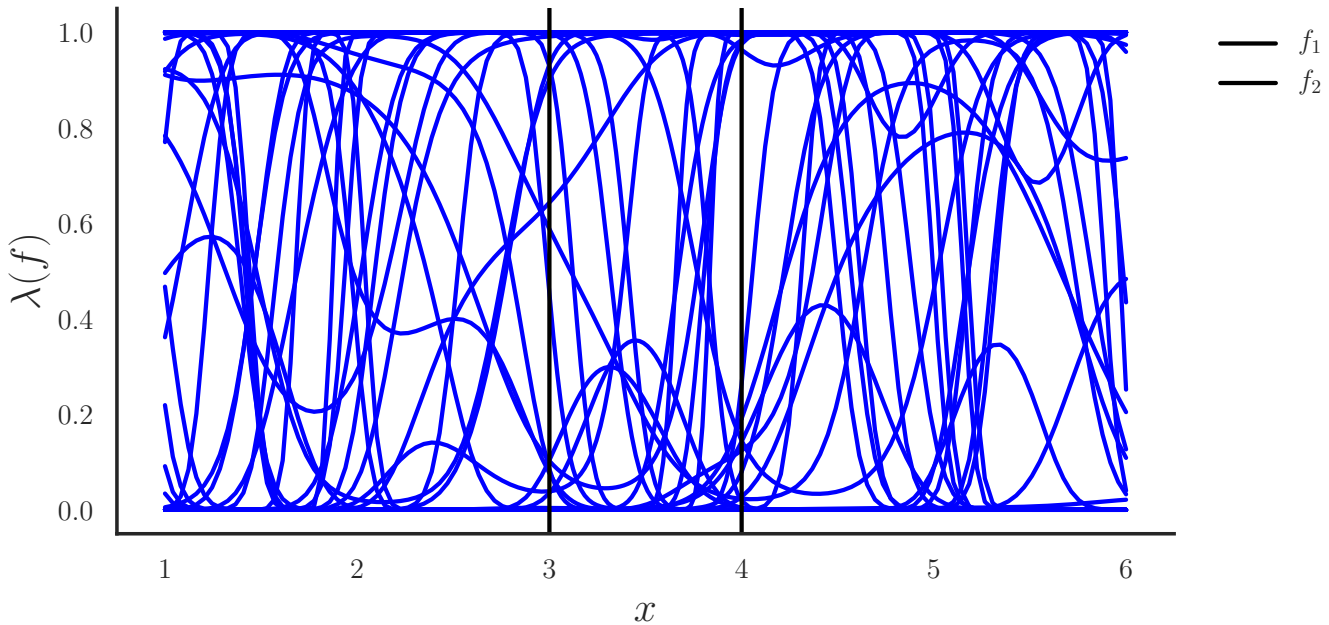
Non-Gaussian posteriors illustrated

- Now consider two observations, $y_1 = 1$ and $y_2 = 1$ at x_1 and x_2
- Need to calculate the joint posterior, $p(\mathbf{f}|\mathbf{y}) = p(f_1, f_2|y_1 = 1, y_2 = 1)$.
- Requires 2D integral $p(y_1 = 1, y_2 = 1|f_1, f_2)p(f_1, f_2)df_1df_2$.



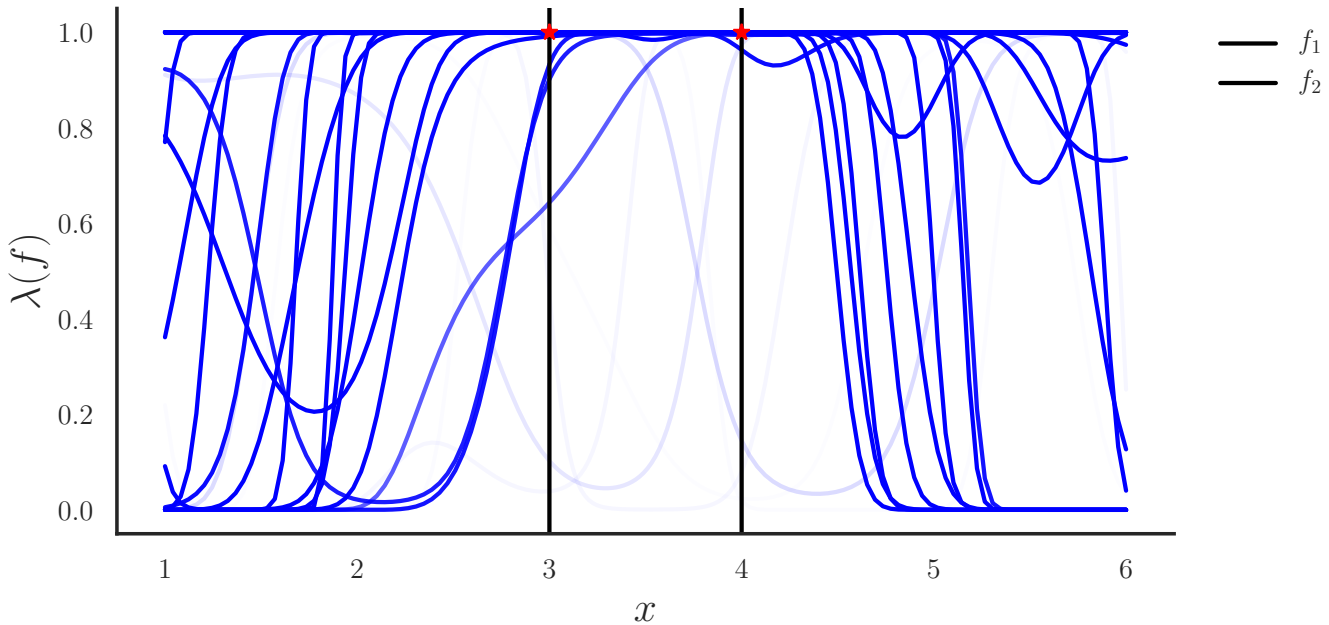
Non-Gaussian posteriors illustrated

- Now consider two observations, $y_1 = 1$ and $y_2 = 1$ at x_1 and x_2
- Need to calculate the joint posterior, $p(\mathbf{f}|\mathbf{y}) = p(f_1, f_2|y_1 = 1, y_2 = 1)$.
- Requires 2D integral $p(y_1 = 1, y_2 = 1|f_1, f_2)p(f_1, f_2)df_1df_2$.



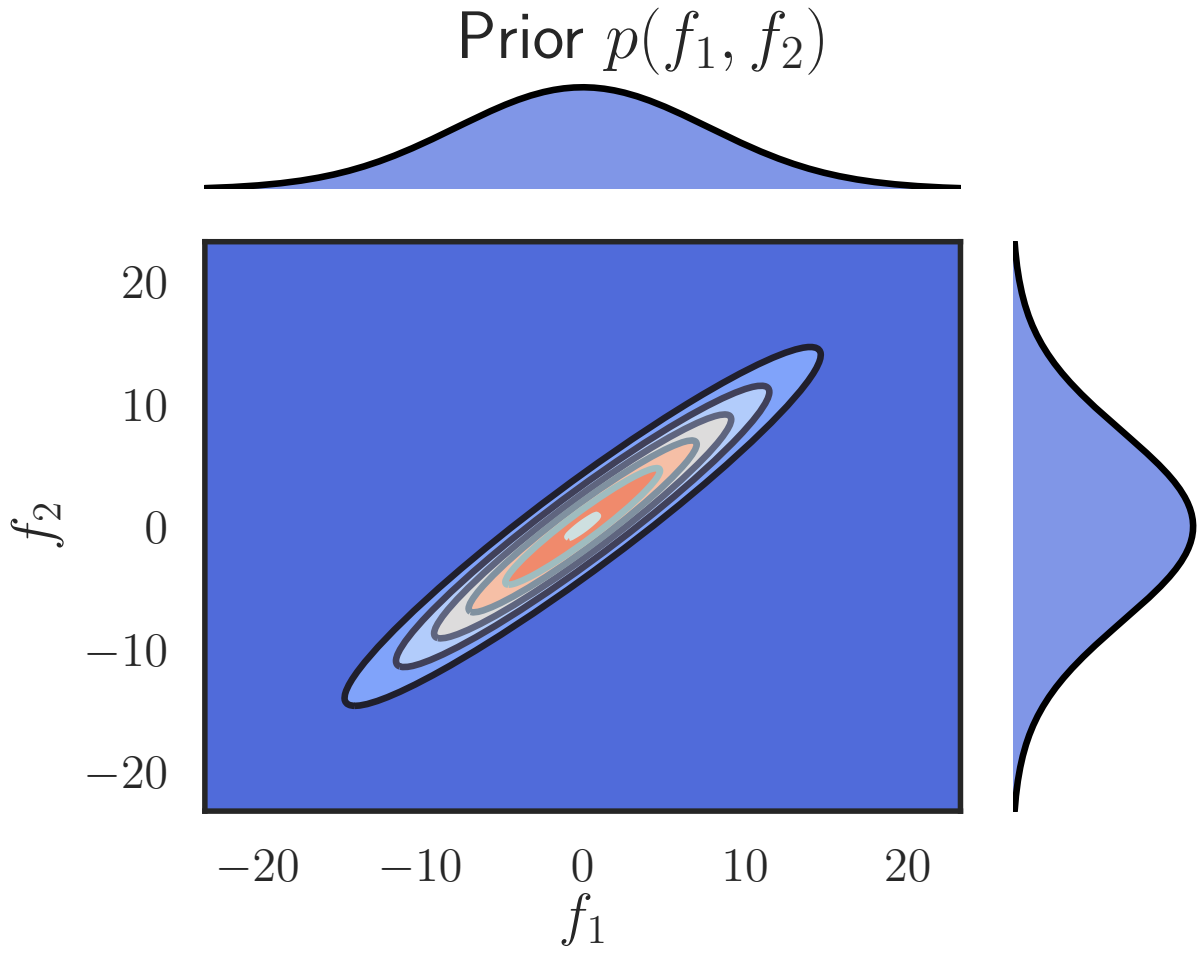
Non-Gaussian posteriors illustrated

- Now consider two observations, $y_1 = 1$ and $y_2 = 1$ at x_1 and x_2
- Need to calculate the joint posterior, $p(\mathbf{f}|\mathbf{y}) = p(f_1, f_2|y_1 = 1, y_2 = 1)$.
- Requires 2D integral $p(y_1 = 1, y_2 = 1|f_1, f_2)p(f_1, f_2)df_1df_2$.



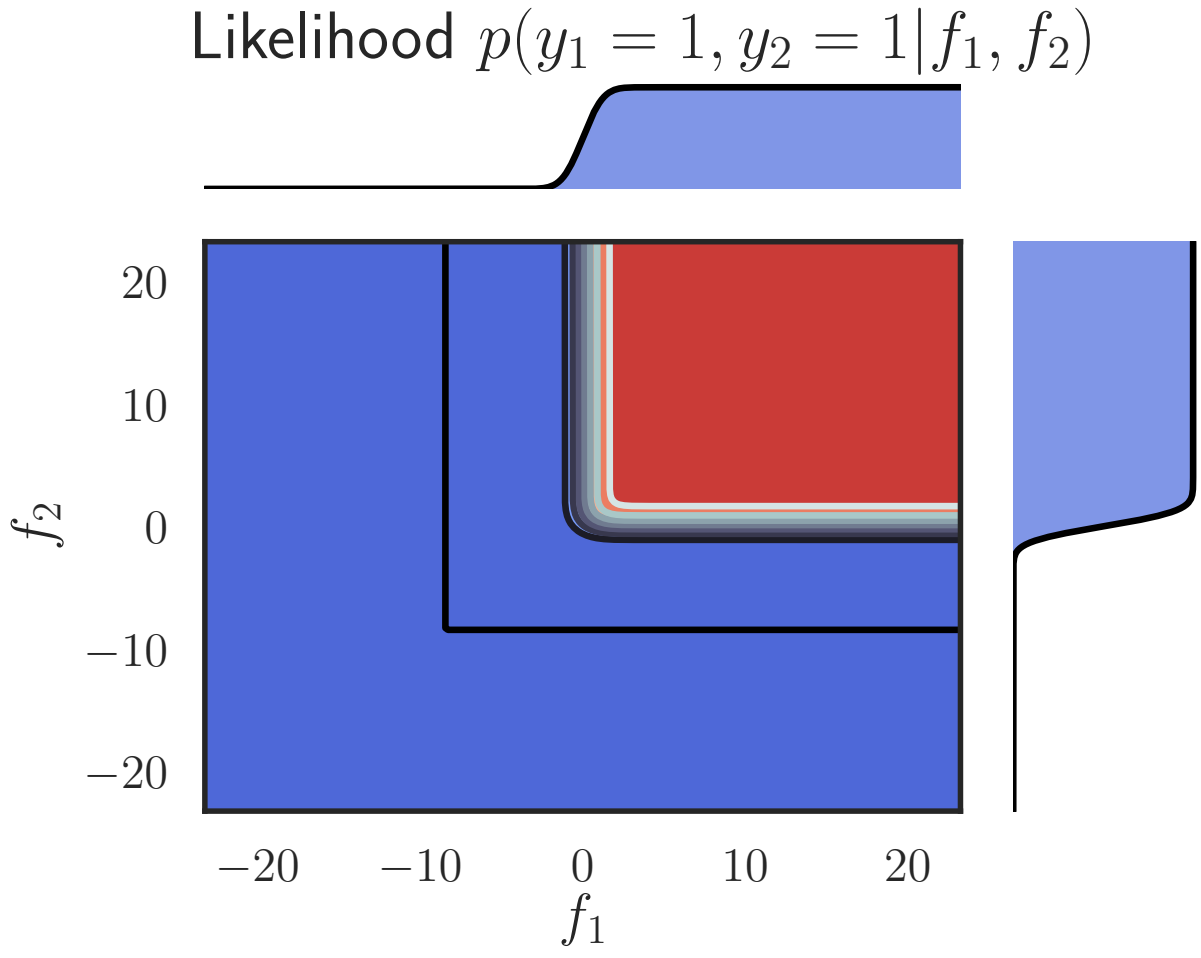
Non-Gaussian posteriors illustrated

- To find the true posterior values, we need to perform a two dimensional integral.
- Still possible, but things are getting more difficult quickly.



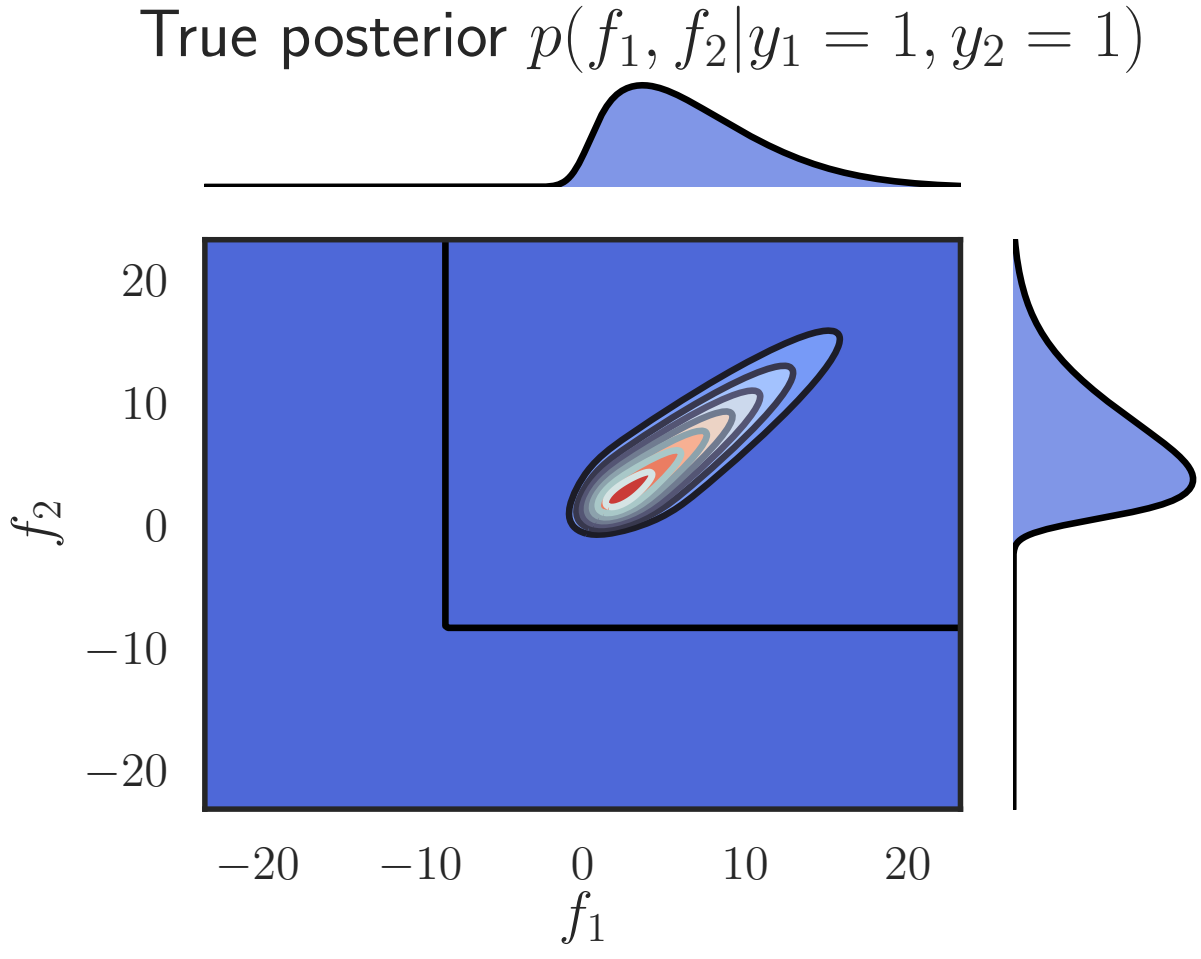
Non-Gaussian posteriors illustrated

- To find the true posterior values, we need to perform a two dimensional integral.
- Still possible, but things are getting more difficult quickly.



Non-Gaussian posteriors illustrated

- To find the true posterior values, we need to perform a two dimensional integral.
- Still possible, but things are getting more difficult quickly.



Approaches to handling non-Gaussian posteriors

Generally fall into two areas:

- Sampling methods that obtain samples of the posterior.
- Approximation of the posterior with something of known form.

Today we will focus on the latter. This leads to more scalable methods, however this comes at the expense of them no longer being exact in the limit of infinite computation time.

IMAGE WITH TRUE NON-GAUSSIAN 2D POSTERIOR, SAMPLES TAKEN FROM THE TRUE POSTERIOR, AND A GAUSSIAN APPROXIMATION

Why do we want an analytical form?

We have said you can't get a nice analytical form because we can't figure out the normaliser. Why is this a problem though? What do we want to do with this posterior?

Non-Gaussian posterior approximation

- Various methods to make a Gaussian approximation, $q(\mathbf{f}) \approx p(\mathbf{f}|\mathbf{y}) = \mathcal{N}(\mathbf{f}|\mu = ?, C = ?)$.
- Allows simple predictions

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

Approximation validity

WHY IS IT OKAY TO APPROXIMATE A GAUSSIAN PROCESS WITH A GAUSSIAN AT TRAINING POINTS ONLY? HOW DOES IT HELP US EVERYWHERE ELSE?

Methods overview

Given choice of Gaussian approximation of posterior. How do we choose the parameter values μ and C?

There a number of different methods in which to choose how to set the parameters of our Gaussian approximation.

Parameters effect

ILLUSTRATION OF WHAT HAPPENS WHEN ADJUSTING THE MEAN AND COVARIANCE PARAMETERS TO THE CONDITIONAL GAUSSIAN

How to choose the parameters?

Two approaches that we might take:

- Is to match the mean and variance at some point, for example the mode.
- Attempt to minimise some divergence measure between the approximate distribution and the true distribution.

- Laplace takes the former
- KL-method takes the latter
- EP kind of takes the latter

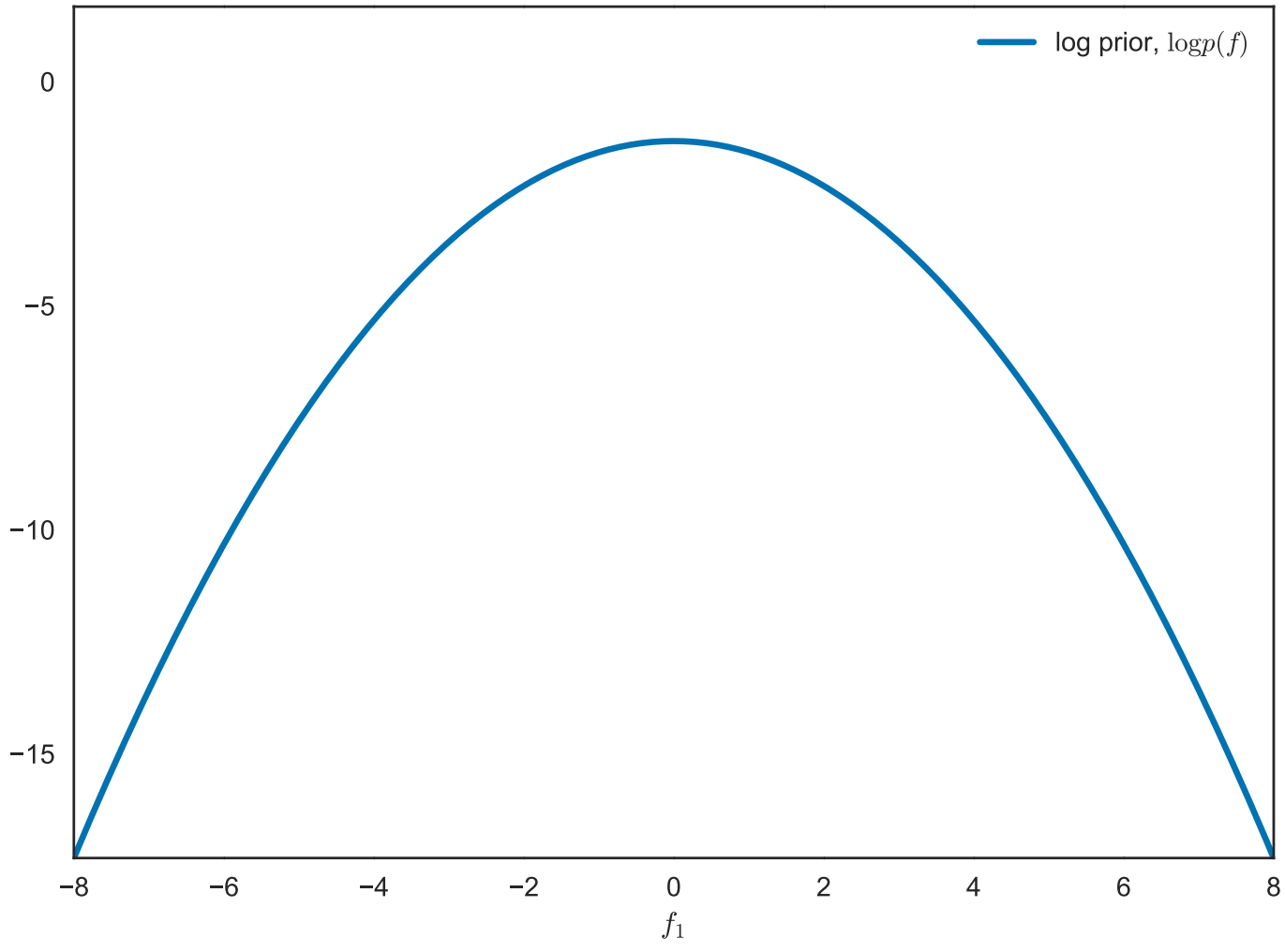
Laplace approximation

- Log of a Gaussian distribution, $q(\mathbf{f}) = \mathcal{N}\left(\mathbf{f}|\hat{\mu}, \hat{C}\right)$, is a quadratic function of $\mathbf{f}$.
- Idea of Laplace approximation is to find the values of the quadratic function that match at the mode of the log posterior.
- It is an approximation of the true posterior, based on the curvature at a single point, the mode.
- Many of you will know this as a second-order Taylor expansion around the modal value.
- The first and second derivatives of the form of the log-posterior, at the mode, will match the derivatives of the approximate Gaussian at this same point.

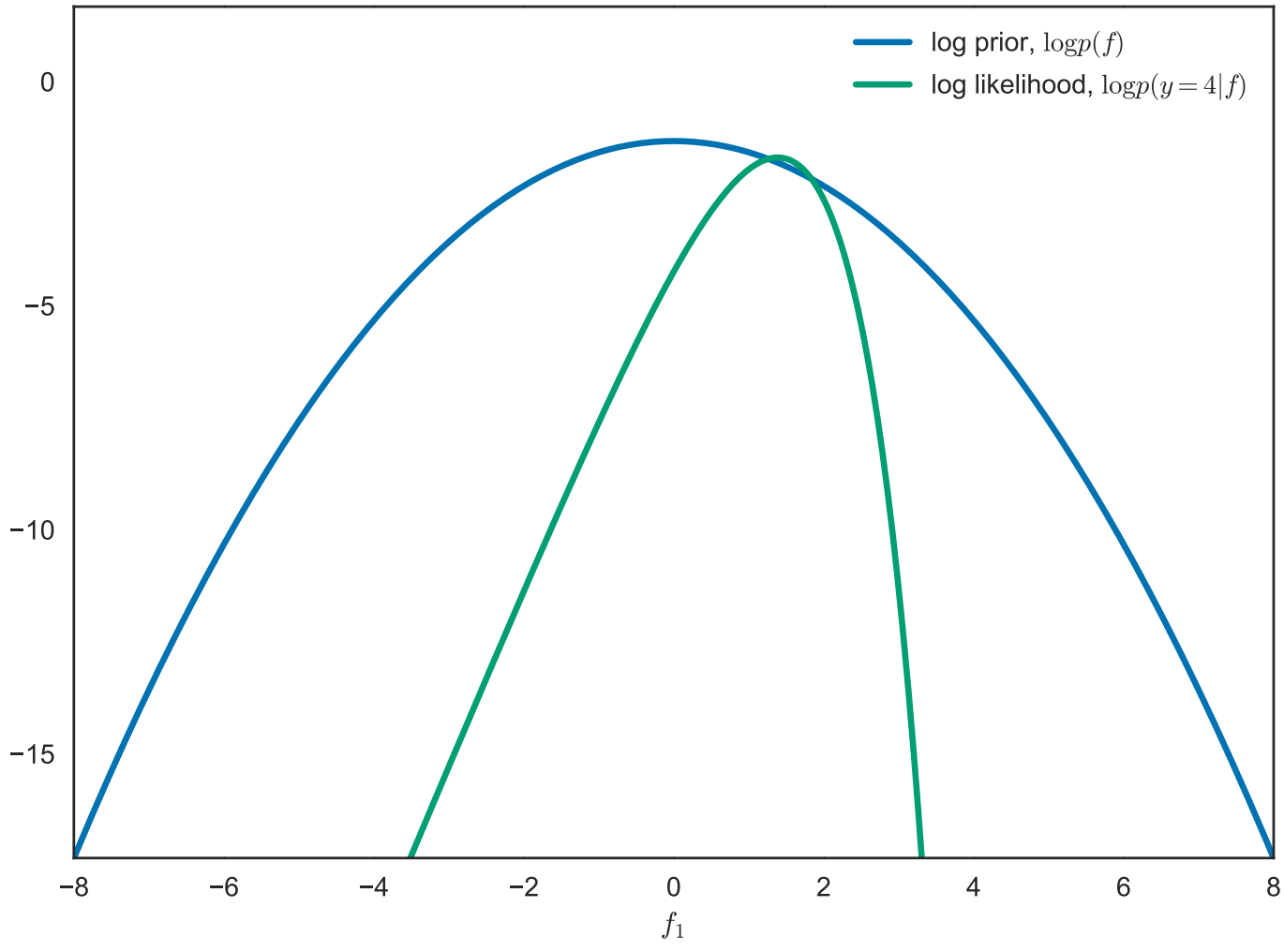
Laplace approximation - algorithm

- Find the mode of the true log posterior, via Newton's method.
- Use second order Taylor expansion around this modal value.
  - i.e obtain curvature at this point.
- Form Gaussian approximation setting the mean equal to the posterior mode, $\hat{\mathbf{f}}$, and matching the curvature.
- $p(\mathbf{f}|\mathbf{y}) \approx q(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{C}) = \mathcal{N}\left(\mathbf{f}|\hat{\mathbf{f}}, (\mathbf{K}_{\mathbf{ff}}^{-1} + \boldsymbol{W})^{-1}\right)$
- $\boldsymbol{W} \triangleq -\frac{d^2 \log p(\mathbf{y}|\hat{\mathbf{f}})}{d\hat{\mathbf{f}}^2}$.
- For factorizing likelihoods (most), $\boldsymbol{W}$ is diagonal.

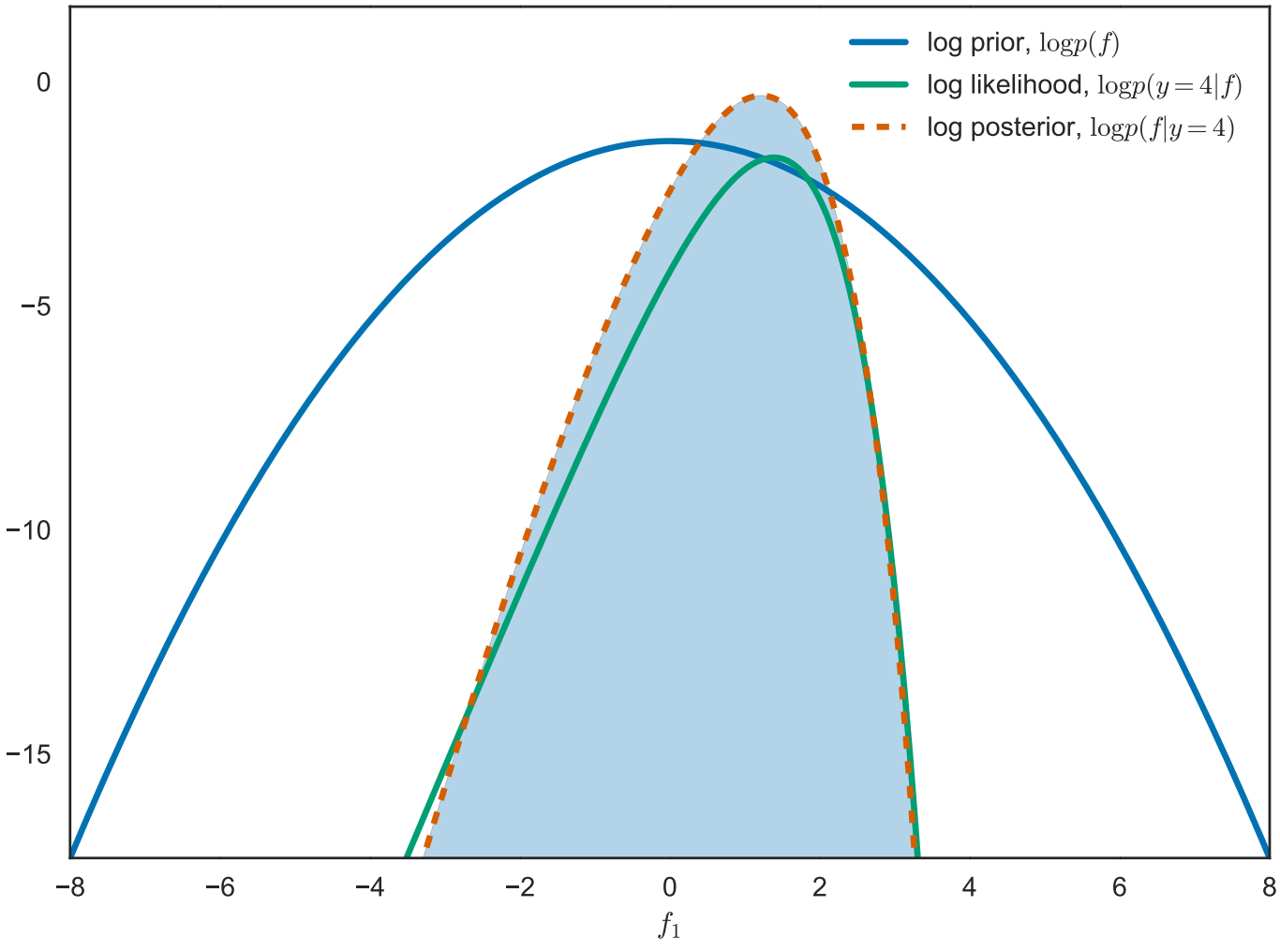
Visualization of Laplace



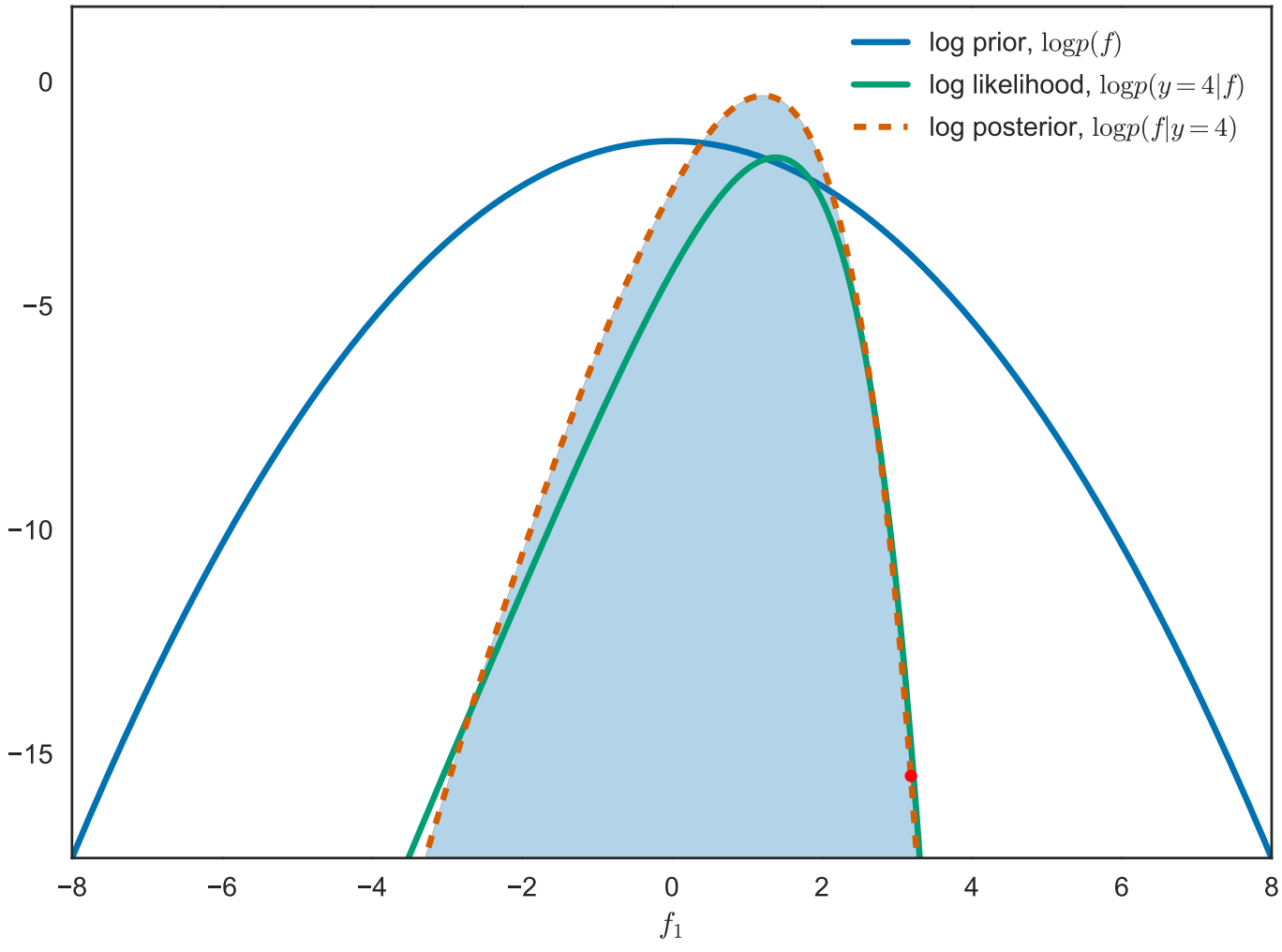
Visualization of Laplace



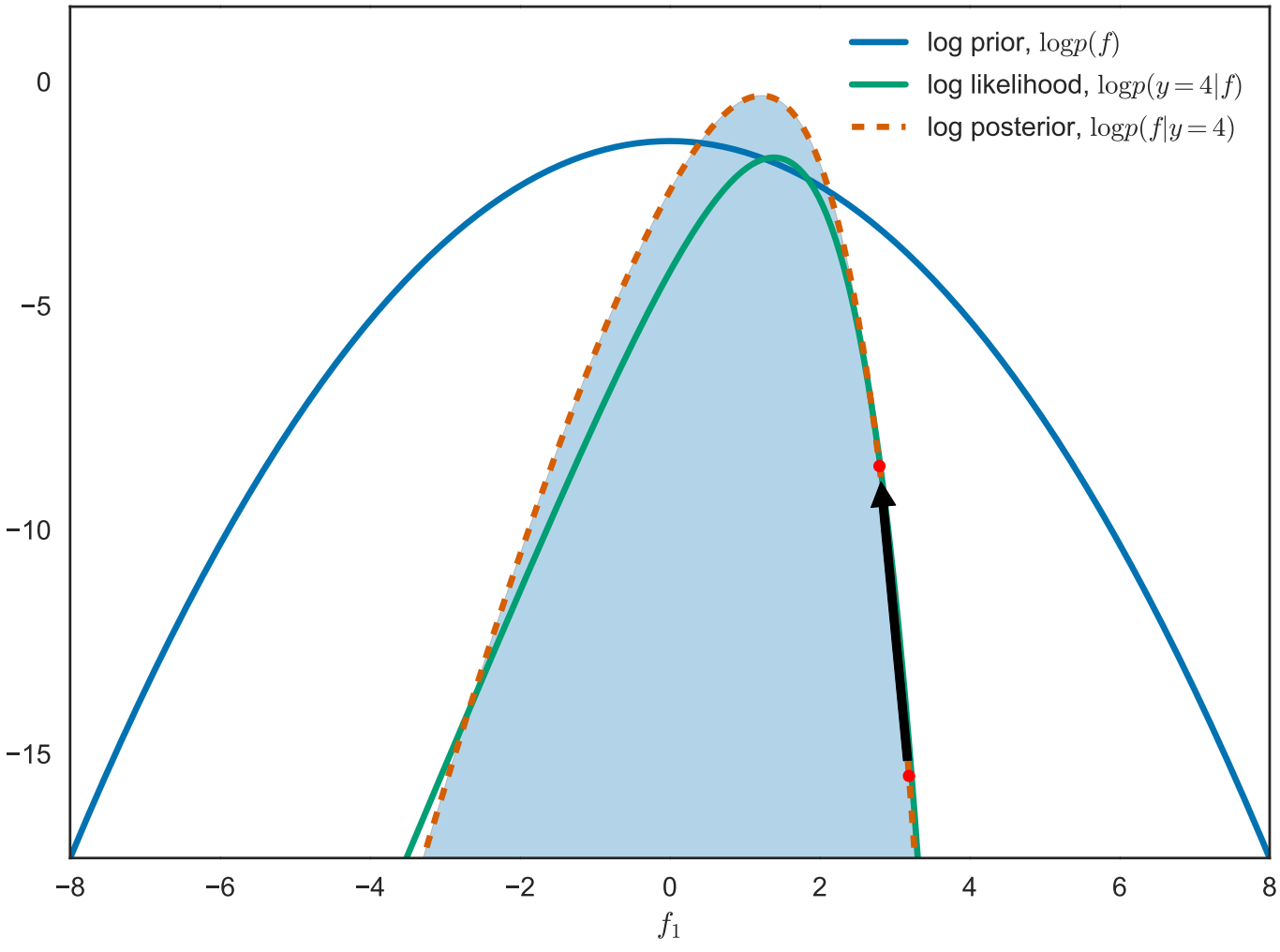
Visualization of Laplace



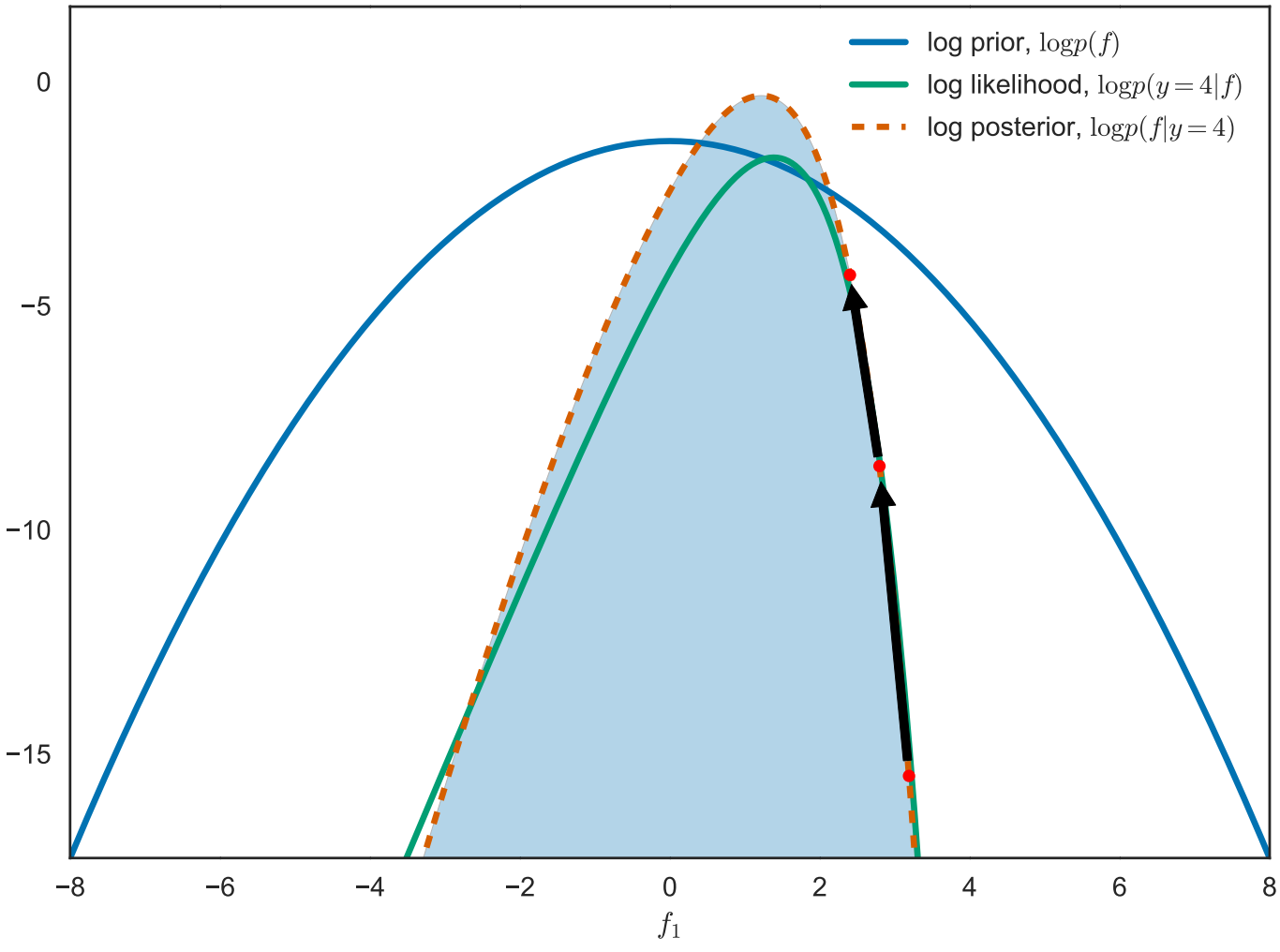
Visualization of Laplace



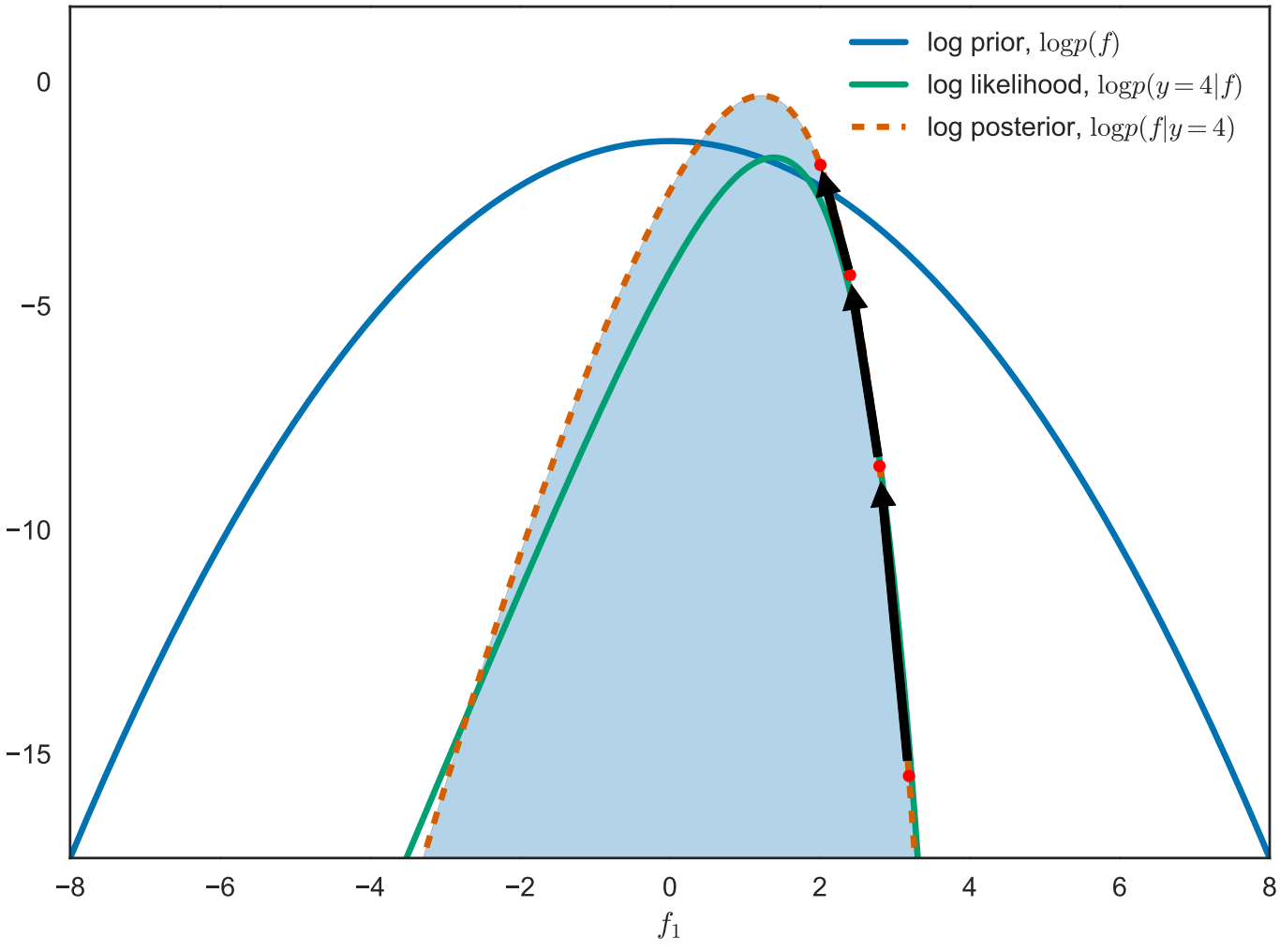
Visualization of Laplace



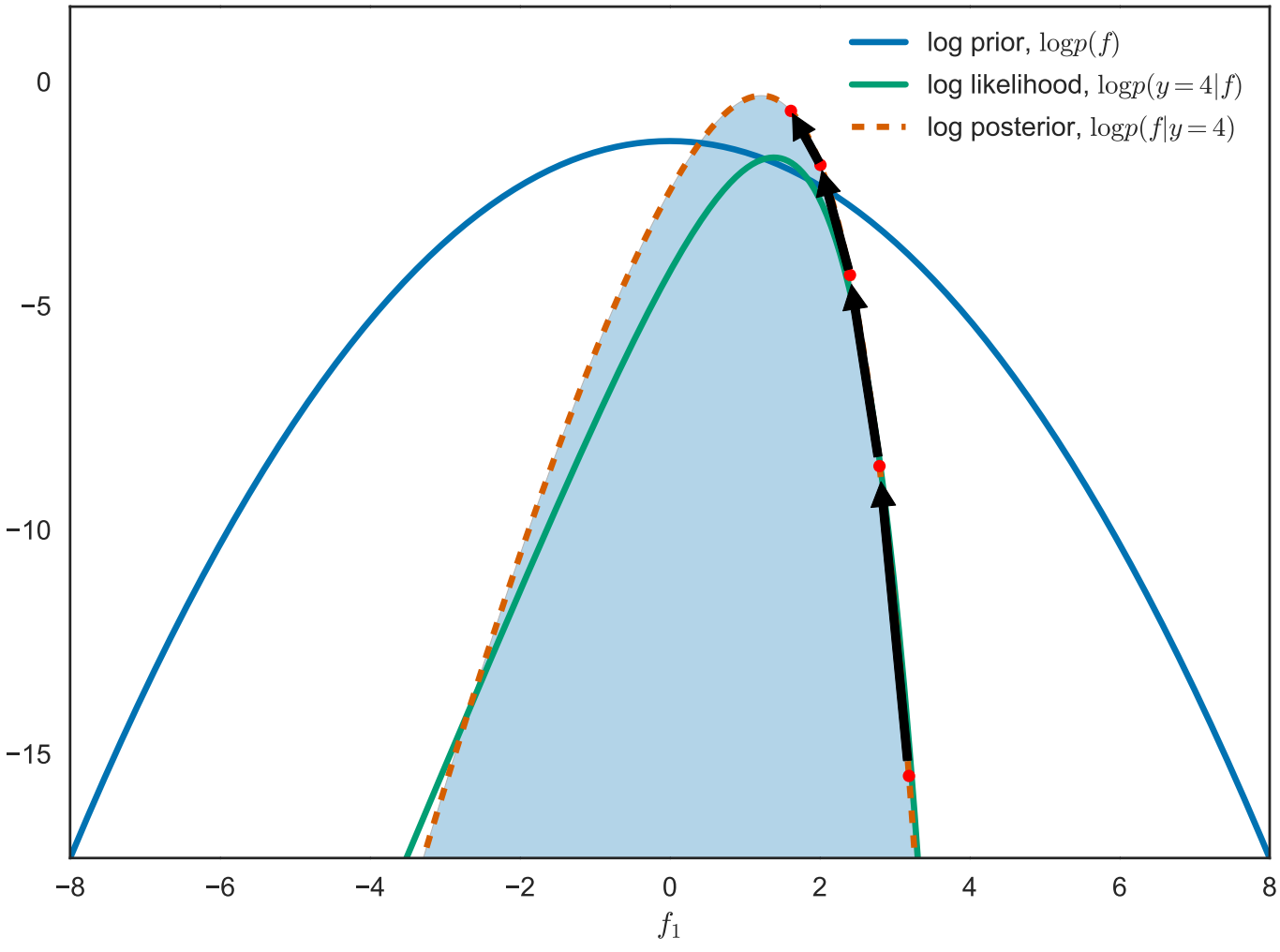
Visualization of Laplace



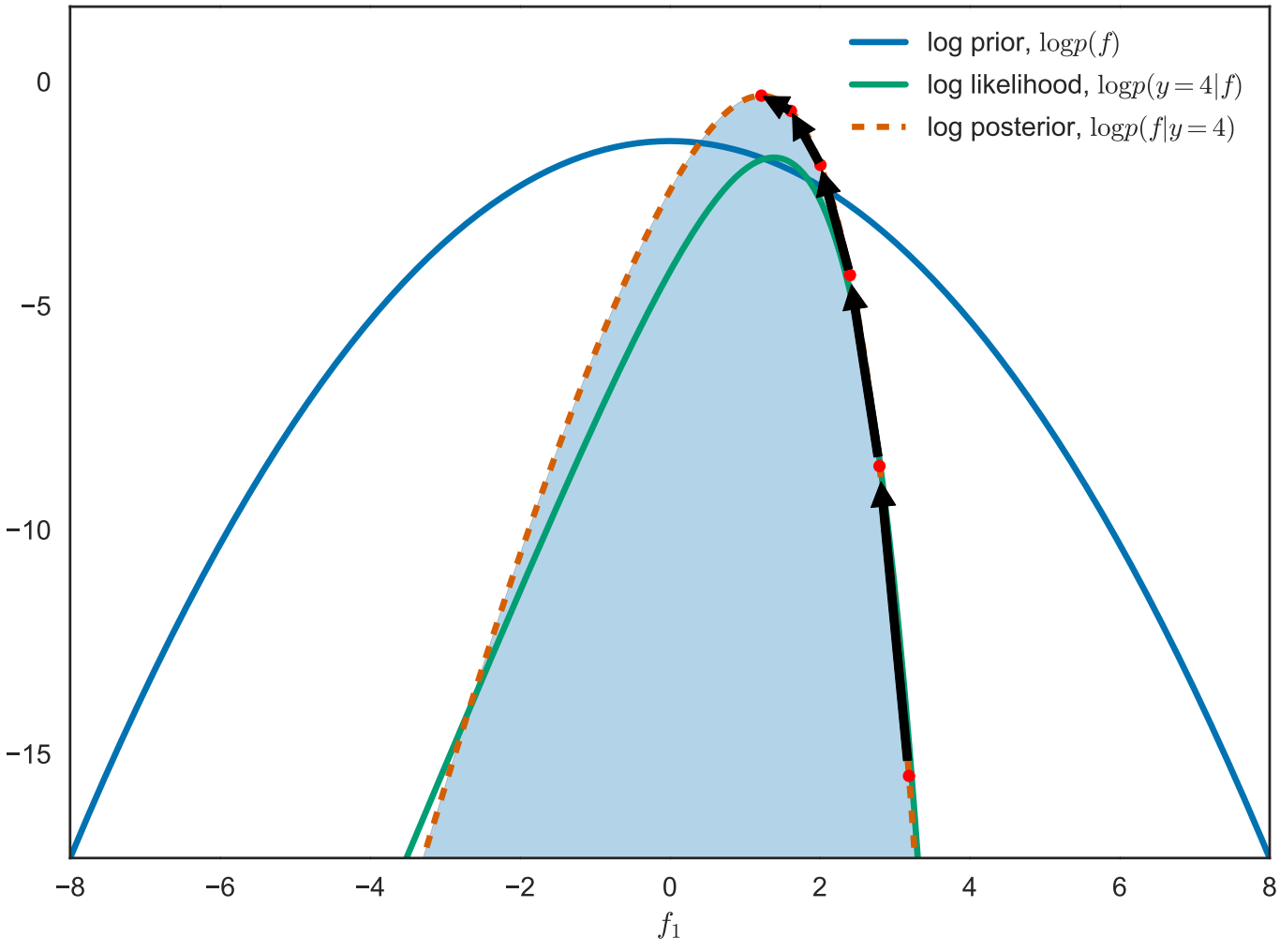
Visualization of Laplace



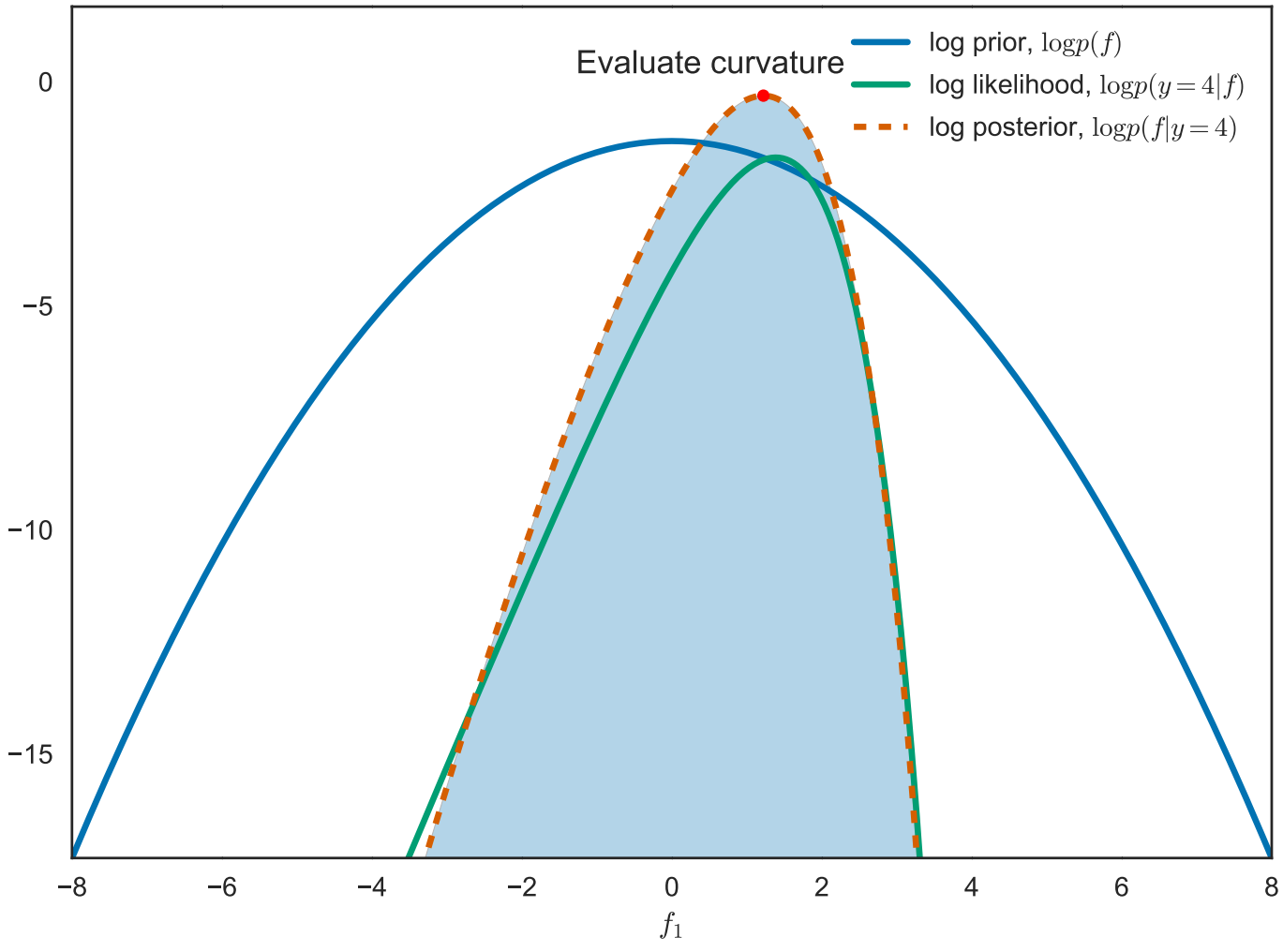
Visualization of Laplace



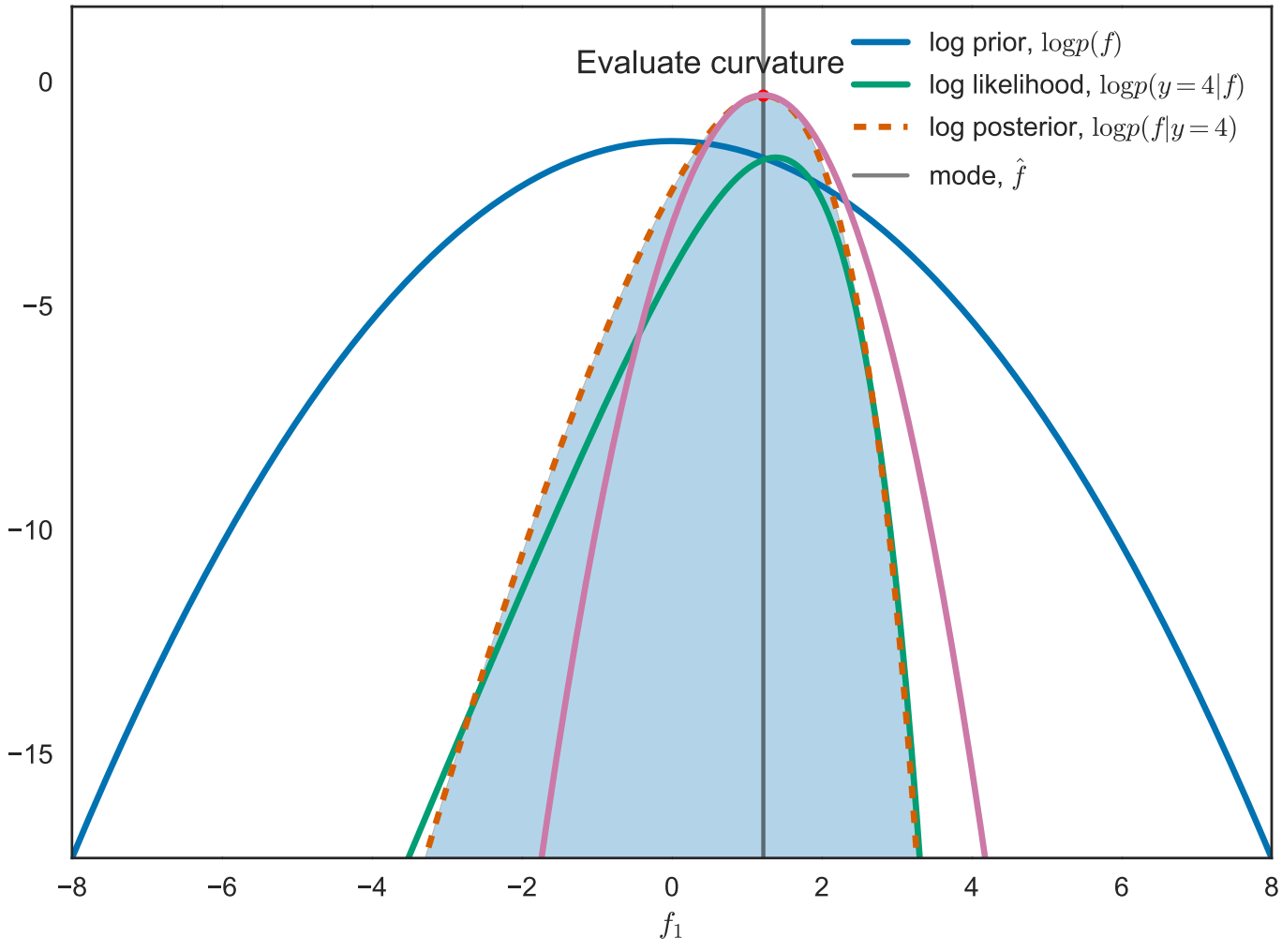
Visualization of Laplace



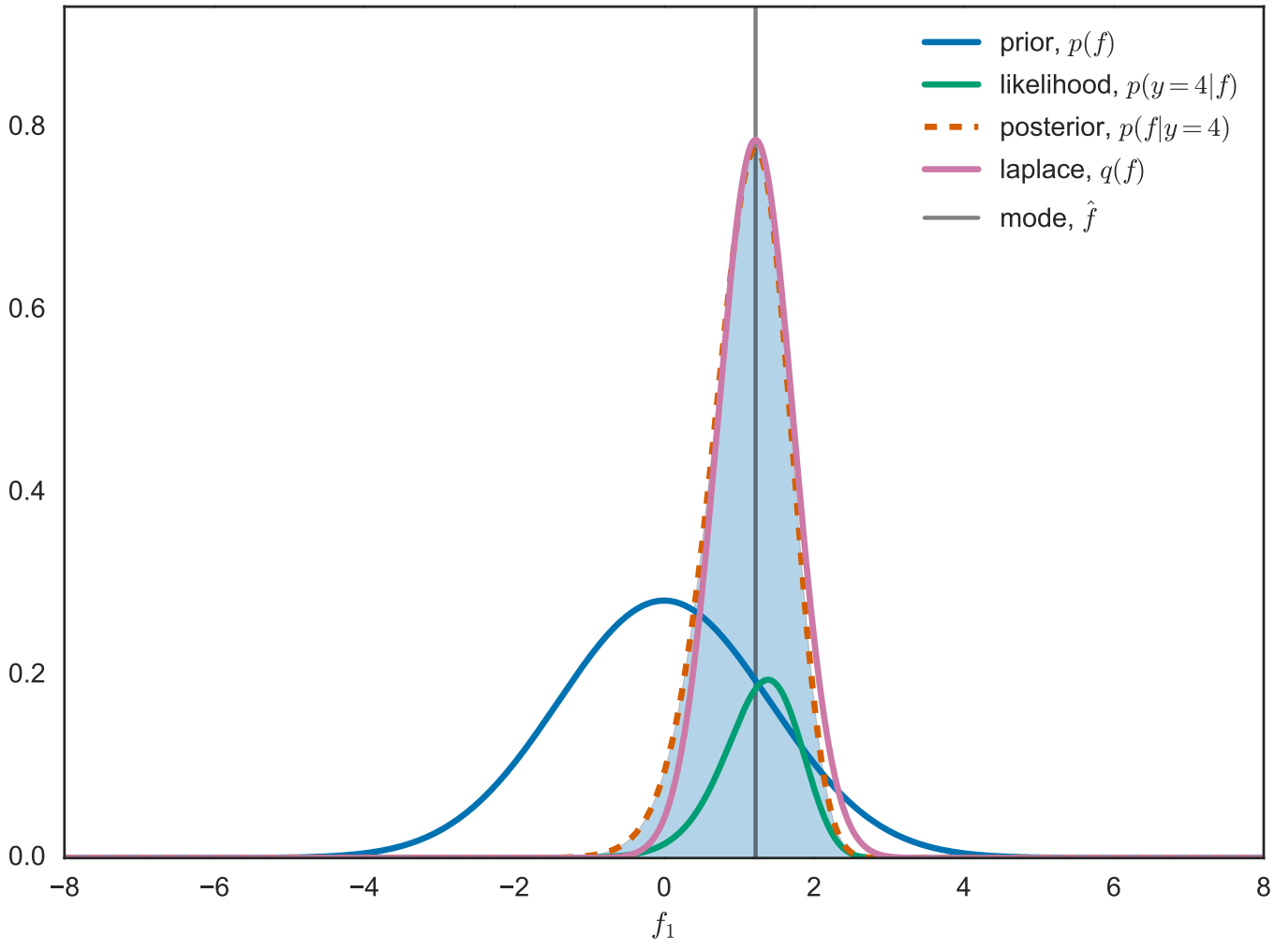
Visualization of Laplace



Visualization of Laplace



Visualization of Laplace



Visualise of Laplace - Bernoulli

MAKE A VISUALISATION OF THE LAPLACE ON THE BERNOULLI TO ILLUSTRATE WHERE IT FALLS DOWN.

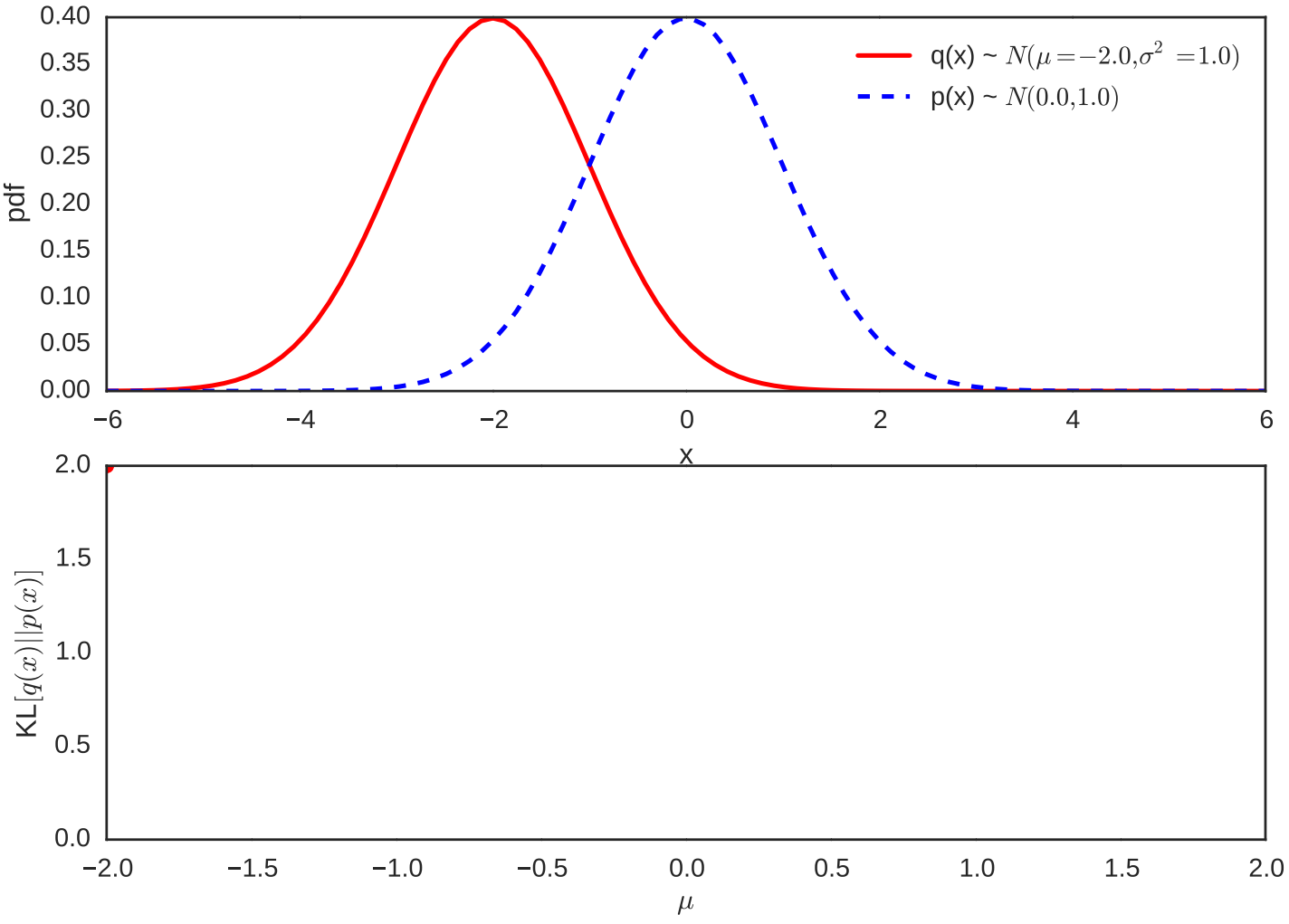
KL-method

- Make a Gaussian approximation, $q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{C})$, as similar possible to true posterior, $p(\mathbf{f}|\mathbf{y})$.
- Treat $\boldsymbol{\mu}$ and $\boldsymbol{C}$ as variational parameters, effecting quality of approximation.
- Define a divergence measure between two distributions, KL divergence, $\mathrm{KL}\left(q(\mathbf{f}) \,||\, p(\mathbf{f}|\mathbf{y})\right)$.
- Minimize this divergence between the two distributions (Nickisch and Rasmussen, 2008), with respect to $\boldsymbol{\mu}$ and $\boldsymbol{C}$ and other parameters.

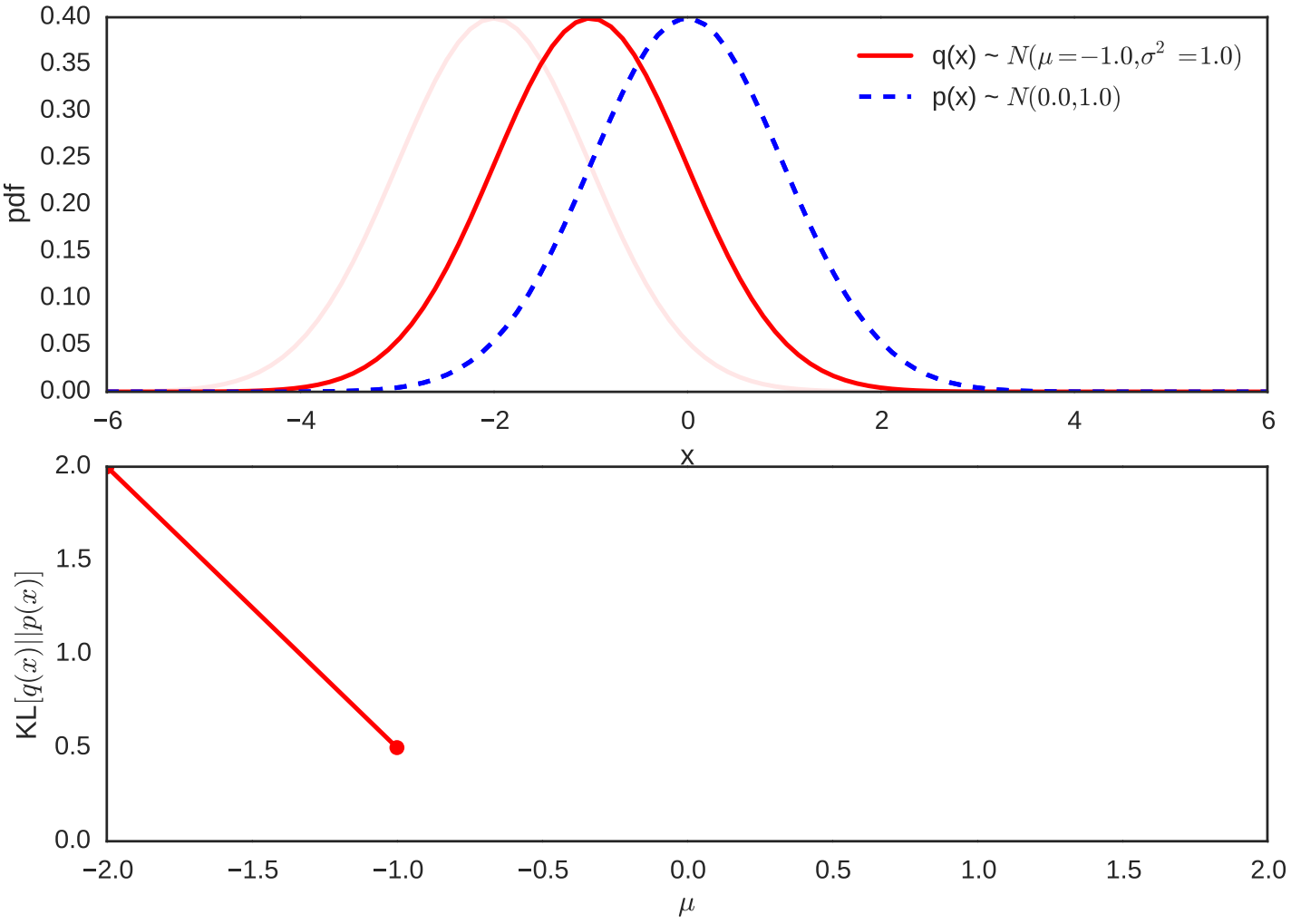
KL divergence

- General for any two distributions $q(\mathbf{x})$ and $p(\mathbf{x})$.
- $\mathrm{KL}\left(q(\mathbf{x}) \| p(\mathbf{x})\right)$ is the average additional amount of information required to specify the values of $\mathbf{x}$ as a result of using an approximate distribution $q(\mathbf{x})$ instead of the true distribution, $p(\mathbf{x})$.
- $\mathrm{KL}\left(q(\mathbf{x}) \| p(\mathbf{x})\right) = \left\langle \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\rangle_{q(\mathbf{x})}$
- Always 0 or positive, not symmetric.
- Lets look at how it changes with response to changes in the approximating distribution.

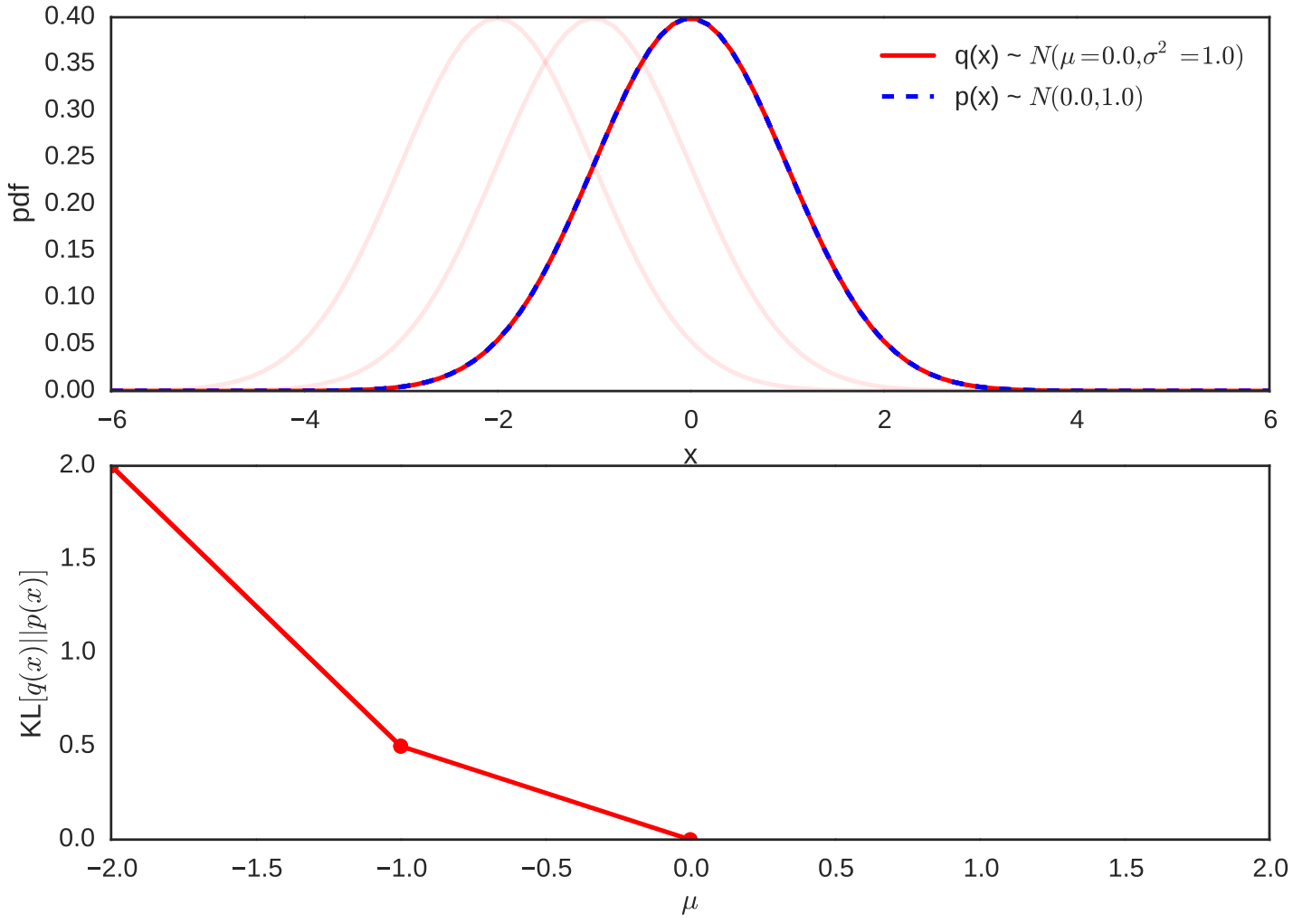
KL varying mean



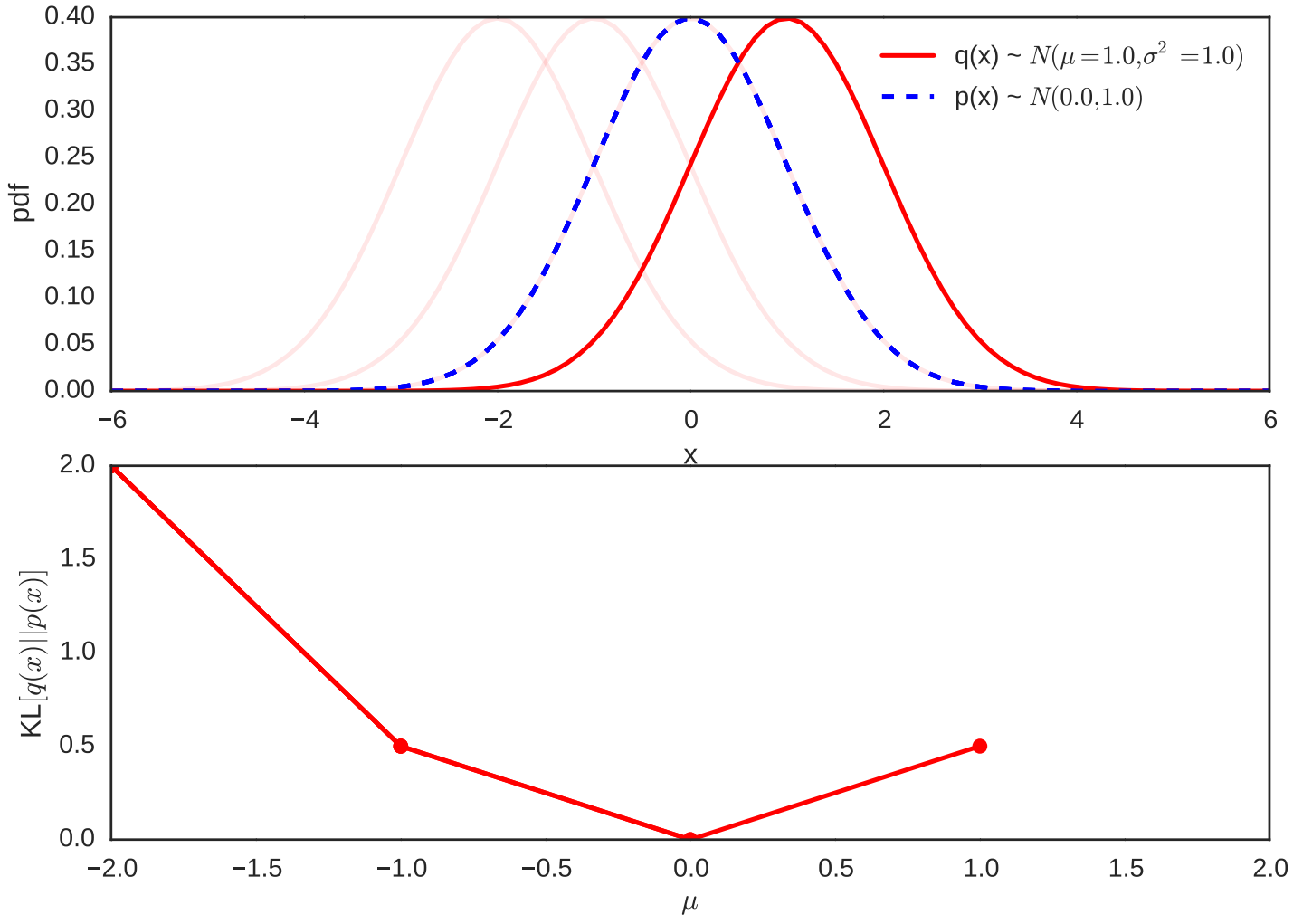
KL varying mean



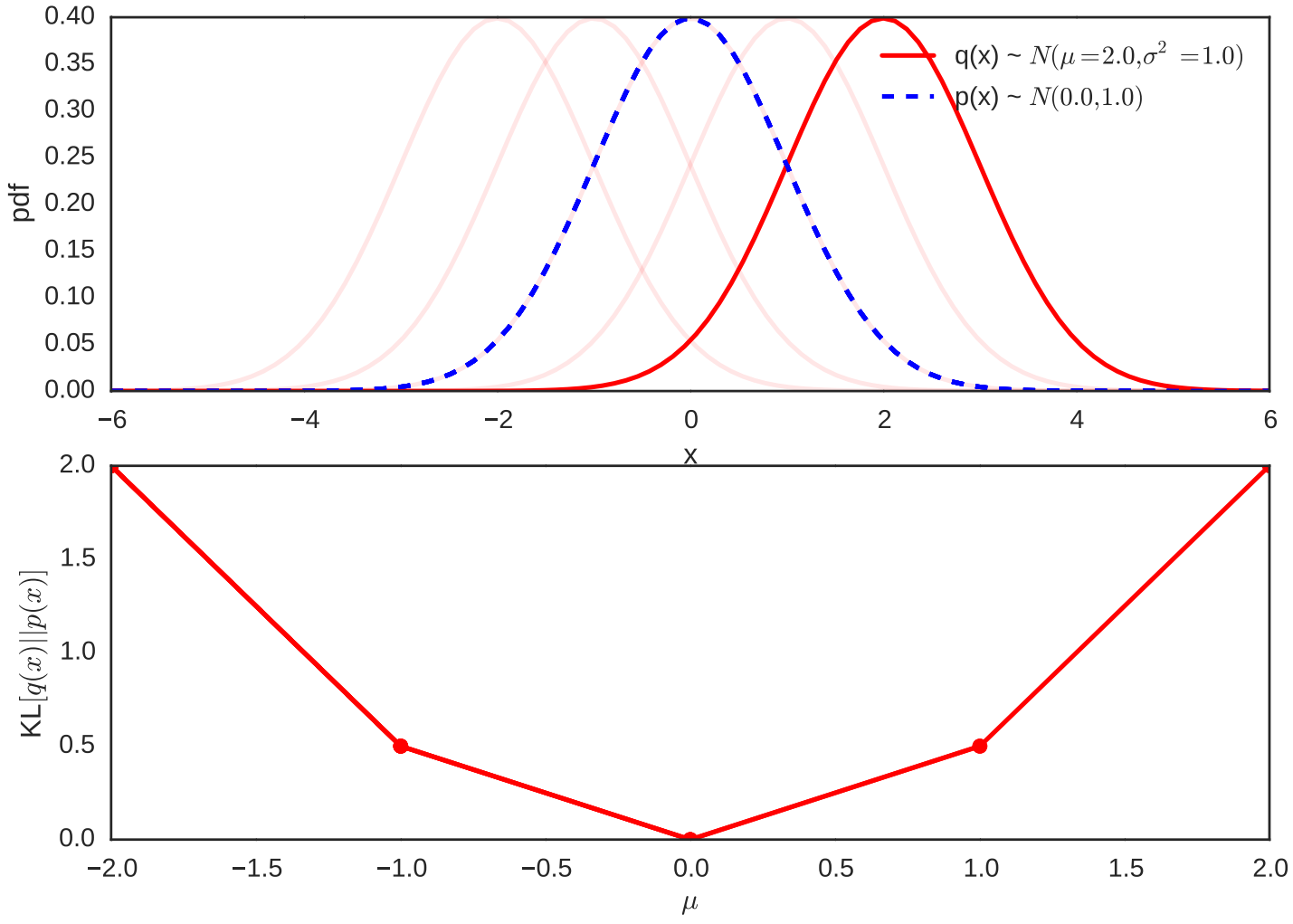
KL varying mean



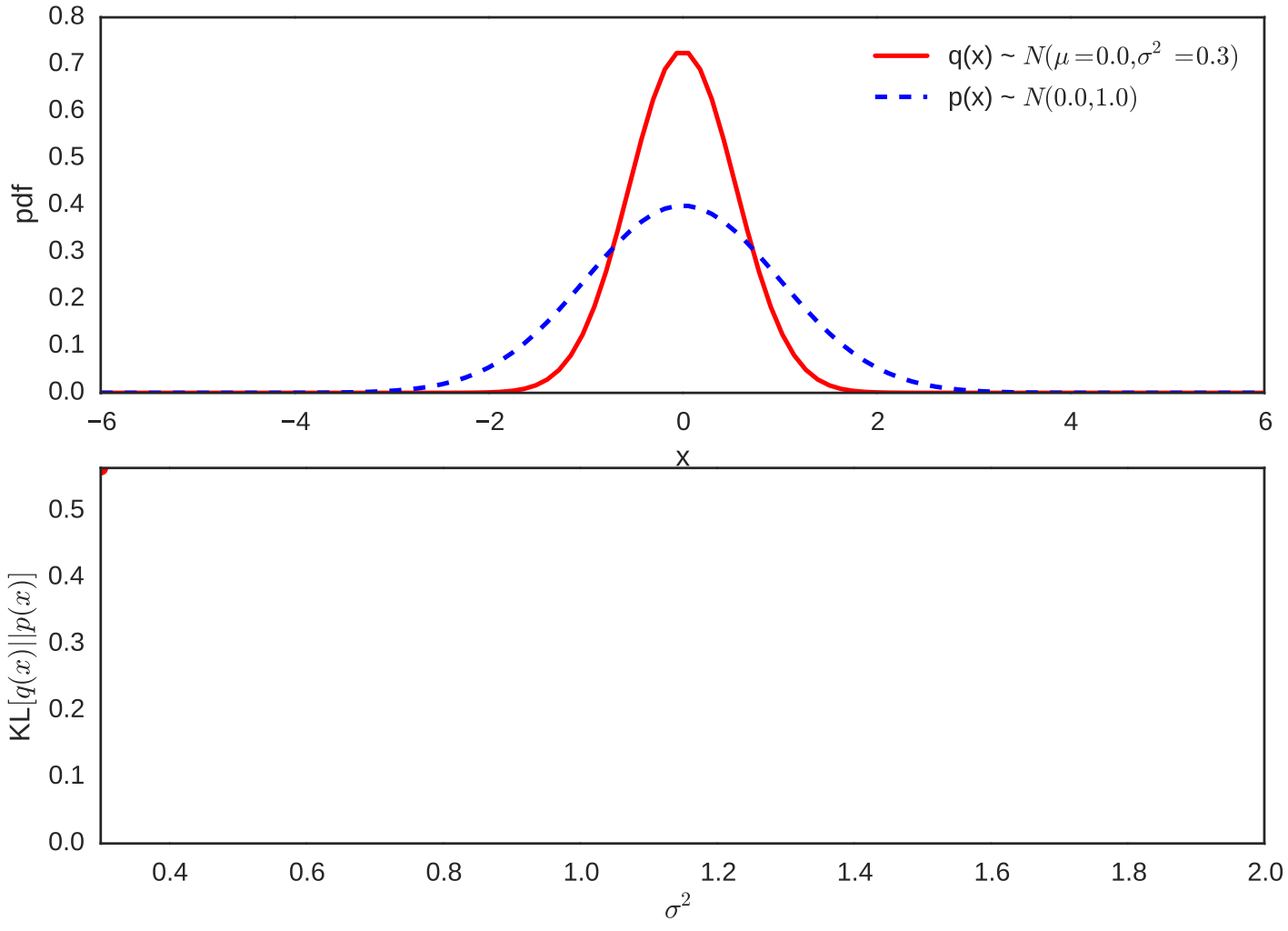
KL varying mean



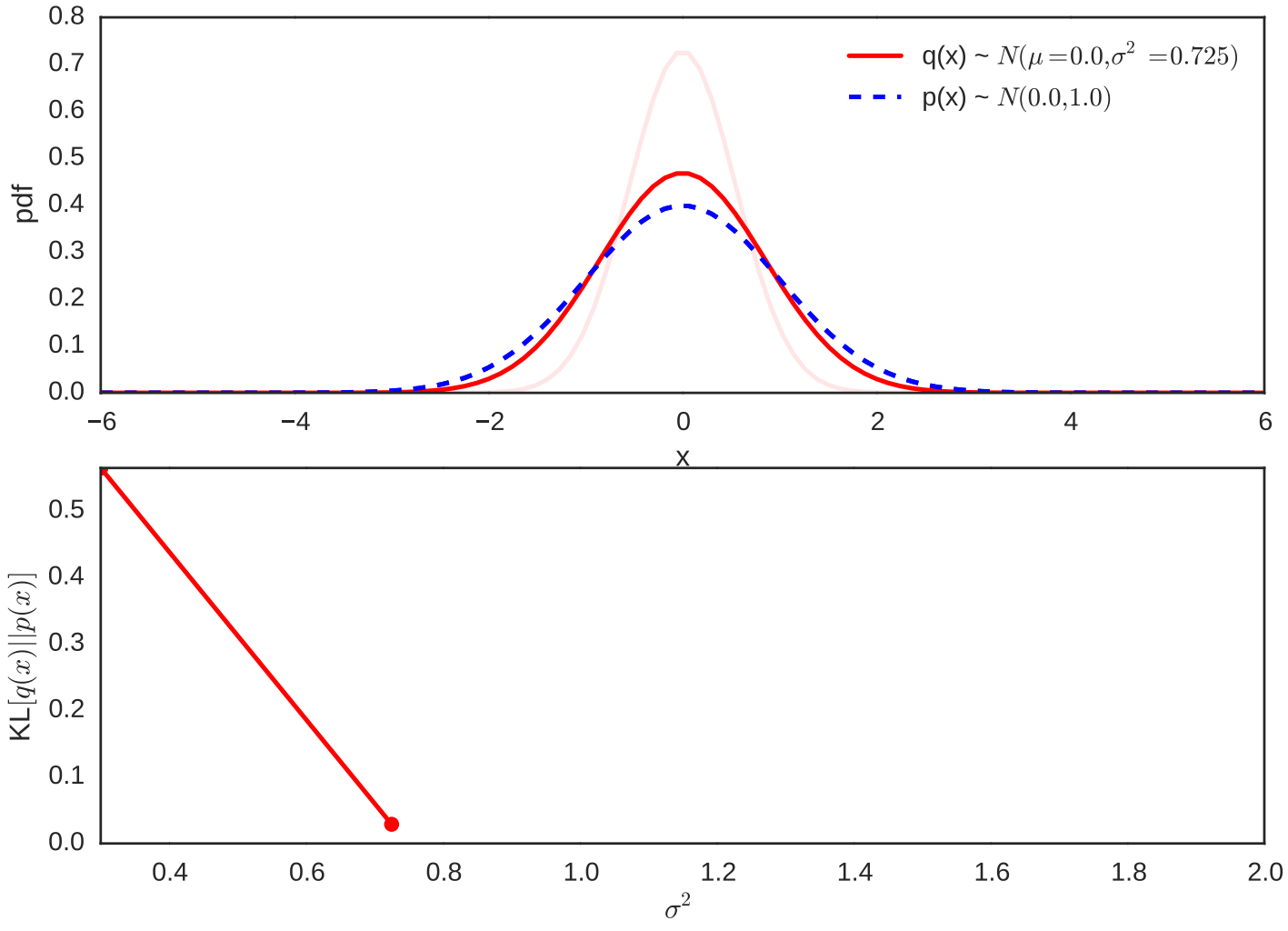
KL varying mean



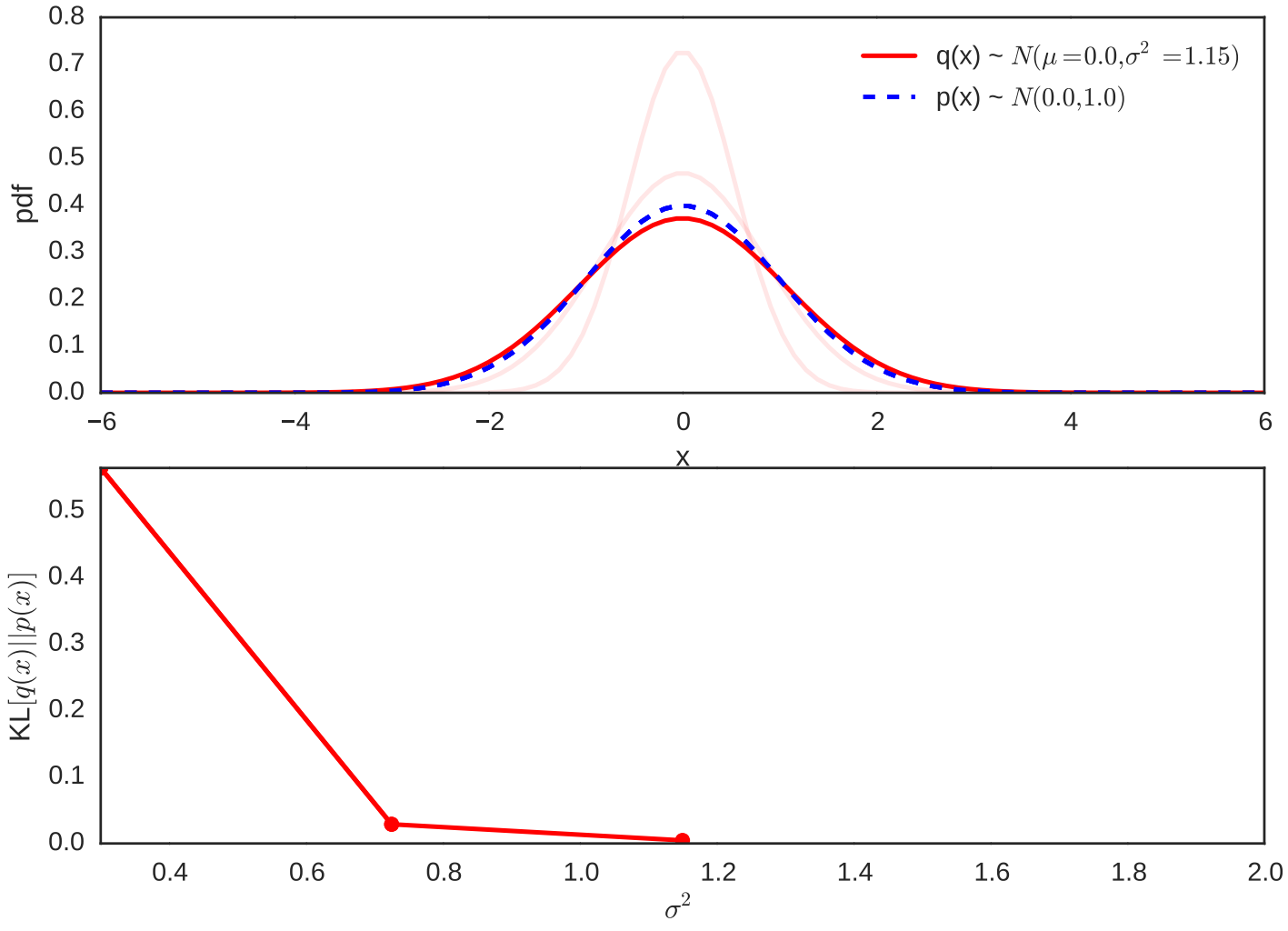
KL varying variance



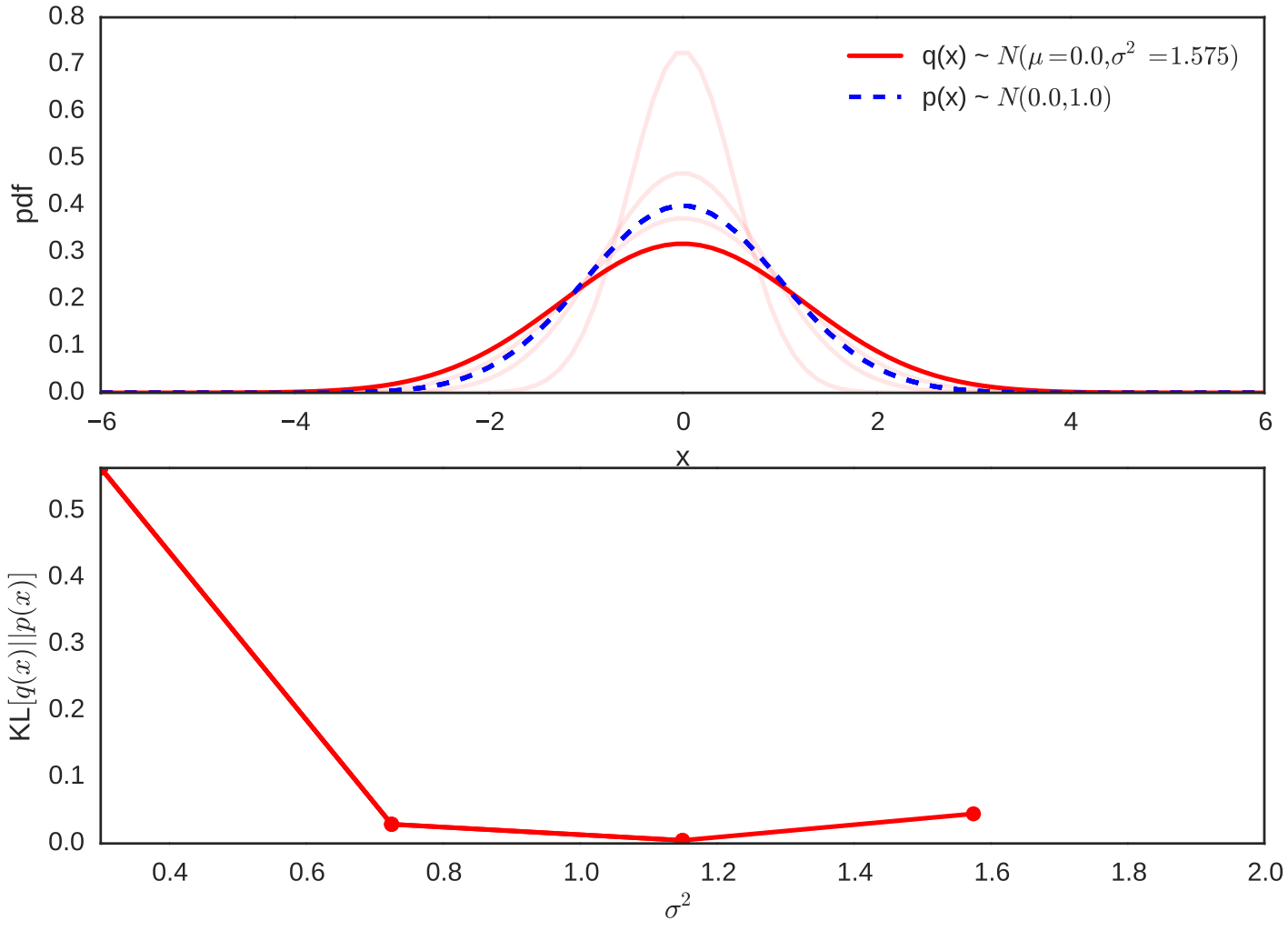
KL varying variance



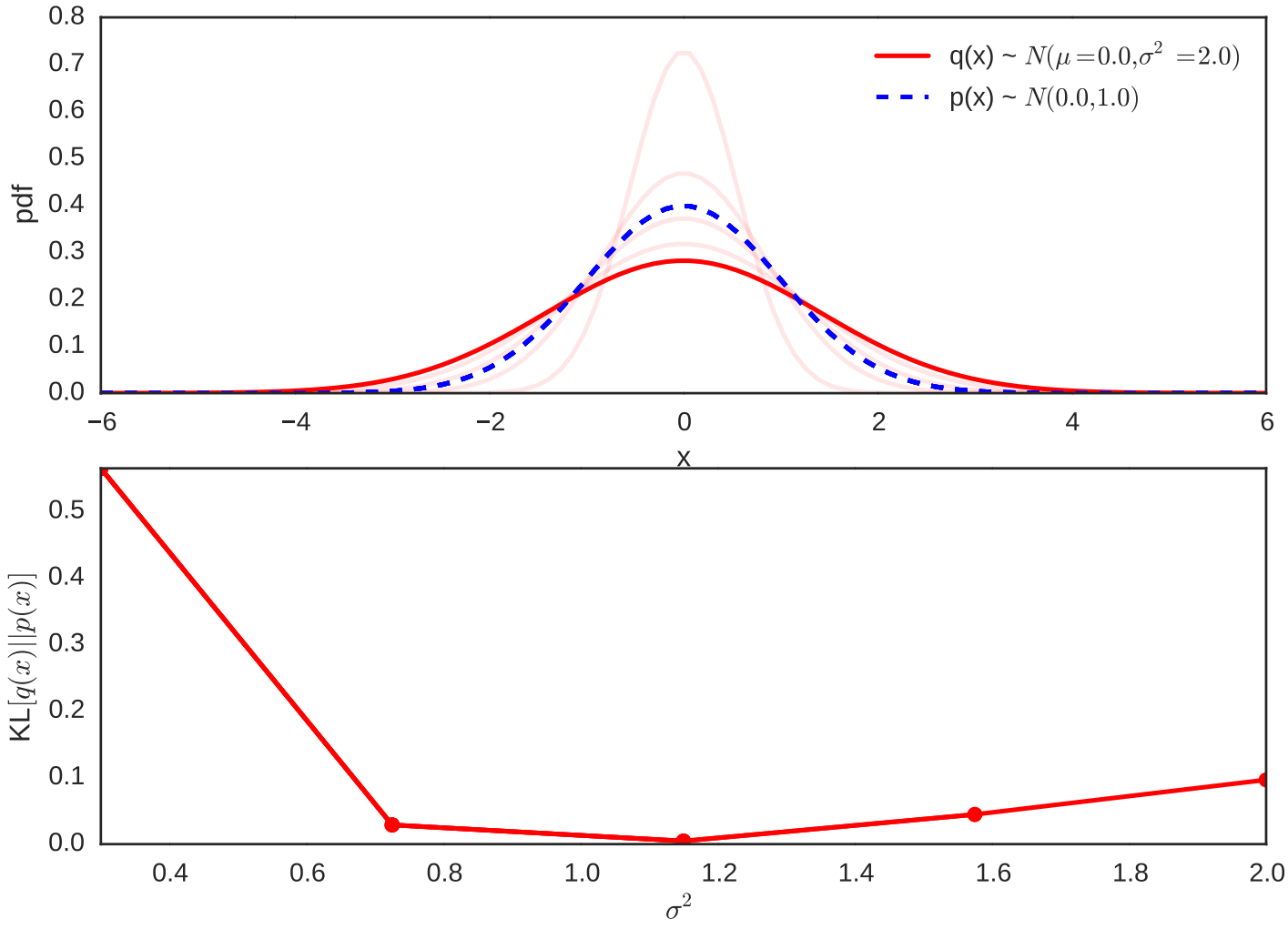
KL varying variance



KL varying variance



KL varying variance



Optimisation illustration

SINCE YOU HAVE SEEN WHAT THE KL DIVERGENCE BETWEEN TWO DISTRIBUTIONS. CONSIDER WHAT WOULD HAPPEN IF YOU COULD OPTIMISE ILLUSTRATION OF WHAT IS BEING OPTIMISED, MAYBE SEVERAL SLIDES ON THIS SIMILAR TO JAMES

KL-method derivation

- Assume Gaussian approximate posterior, $q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{C})$.
- True posterior using Bayes rule, $p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{p(\mathbf{y})}$.
- Cannot compute the KL divergence as we cannot compute the true posterior, $p(\mathbf{f}|\mathbf{y})$.

$$
\begin{aligned}
\mathrm{KL}\left(q(\mathbf{f}) \,||\, p(\mathbf{f}|\mathbf{y})\right) &= \left\langle \log \frac{q(\mathbf{f})}{p(\mathbf{f}|\mathbf{y})} \right\rangle_{q(\mathbf{f})} \\
&= \left\langle \log \frac{q(\mathbf{f})}{p(\mathbf{f})} - \log p(\mathbf{y}|\mathbf{f}) + \log p(\mathbf{y}) \right\rangle_{q(\mathbf{f})} \\
&= \mathrm{KL}\left(q(\mathbf{f}) \,||\, p(\mathbf{f})\right) - \left\langle \log p(\mathbf{y}|\mathbf{f}) \right\rangle_{q(\mathbf{f})} + \log p(\mathbf{y}) \\
\log p(\mathbf{y}) &= \left\langle \log p(\mathbf{y}|\mathbf{f}) \right\rangle_{q(\mathbf{f})} - \mathrm{KL}\left(q(\mathbf{f}) \,||\, p(\mathbf{f})\right) + \mathrm{KL}\left(q(\mathbf{f}) \,||\, p(\mathbf{f}|\mathbf{y})\right)
\end{aligned}
$$

KL-method derivation

$$\begin{aligned}\log p(\mathbf{y}) &= \left\langle \log p(\mathbf{y}|\mathbf{f}) \right\rangle_{q(\mathbf{f})} - \mathrm{KL}\left(q(\mathbf{f}) \,\|\, p(\mathbf{f})\right) + \mathrm{KL}\left(q(\mathbf{f}) \,\|\, p(\mathbf{f}|\mathbf{y})\right) \\ &\geq \left\langle \log p(\mathbf{y}|\mathbf{f}) \right\rangle_{q(\mathbf{f})} - \mathrm{KL}\left(q(\mathbf{f}) \,\|\, p(\mathbf{f})\right)\end{aligned}$$

- Tractable terms give lower bound on $\log p(\mathbf{y})$ as $\mathrm{KL}\left(q(\mathbf{f}) \,\|\, p(\mathbf{f}|\mathbf{y})\right)$ always positive.
- Adjust variational parameters $\boldsymbol{\mu}$ and $\boldsymbol{C}$ to make tractable terms as large as possible, thus $\mathrm{KL}\left(q(\mathbf{f}) \,\|\, p(\mathbf{f}|\mathbf{y})\right)$ as small as possible.
- $\left\langle \log p(\mathbf{y}|\mathbf{f}) \right\rangle_{q(\mathbf{f})}$ with factorizing likelihood can be done with a series of n 1 dimensional integrals.
- In practice, can reduce the number of variational parameters by reparameterizing $\boldsymbol{C} = (\mathbf{K_{ff}} - 2\Lambda)^{-1}$ by noting that the bound is constant in off diagonal terms of $\boldsymbol{C}$.

KL validity

WHY IS IT OKAY TO DEFINE THE KL DIVERGENCE BETWEEN TWO INFINITE PROCESSES BASED ON ONLY TRAINING POINTS?

Expectation Propagation

$$p(\mathbf{f}|\mathbf{y}) \propto p(\mathbf{f}) \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{f}_i)$$

$$q(\mathbf{f}|\mathbf{y}) \triangleq \frac{1}{Z_{ep}} p(\mathbf{f}) \prod_{i=1}^{n} t_i(\mathbf{f}_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \Sigma)$$

$$t_i \triangleq \tilde{Z}_i \mathcal{N}\left(\mathbf{f}_i|\tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\sigma}}_i^2\right)$$

- Individual likelihood terms, $p(\mathbf{y}_i|\mathbf{f}_i)$, replaced by independent local likelihood functions, t_i.
- Uses an iterative algorithm to update t_i's.

Expectation Propagation

1. From the approximate current posterior, $q(\mathbf{f}|\mathbf{y})$, leave out one of the local likelihoods, t_i, and marginalise $\mathbf{f}_j$ where $j \neq i$, giving rise to the marginal *cavity distribution*, $q_{-i}(\mathbf{f}_i)$.
2. Combine resulting cavity distribution, $q_{-i}(\mathbf{f}_i)$, with exact likelihood contribution, $p(\mathbf{y}_i|\mathbf{f}_i)$, giving non-Gaussian un-normalized distribution, $\hat{q}(\mathbf{f}_i) \triangleq p(\mathbf{y}_i|\mathbf{f}_i)q_{-i}(\mathbf{f}_i)$.
3. Choose a un-normalized Gaussian approximation to this distribution, $\mathcal{N}\left(\mathbf{f}_i|\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\sigma}}_i^2\right)\hat{Z}_i$, by finding moments of $\hat{q}(\mathbf{f}_i)$.
4. Replace parameters of t_i with those that produce the same moments as this approximation.
5. Choose another i and start again. Repeat to convergence.

Expectation Propagation - in math

Step 1. First choose a local likelihood contribution, i, to leave out, and find the marginal cavity distribution,

$$q(\mathbf{f}|\mathbf{y}) \propto p(\mathbf{f}) \prod_{j=1}^{n} t_j(\mathbf{f}_j) \rightarrow \frac{p(\mathbf{f}) \prod_{j=1}^{n} t_j(\mathbf{f}_j)}{t_i(\mathbf{f}_i)} \rightarrow p(\mathbf{f}) \prod_{j \neq i}^{n} t_j(\mathbf{f}_j)$$

$$\rightarrow \int p(\mathbf{f}) \prod_{j \neq i} t_j(\mathbf{f}_j) \, d\mathbf{f}_{j \neq i} \triangleq q_{-i}(\mathbf{f}_i)$$

Step 2. $p(\mathbf{y}_i|\mathbf{f}_i) q_{-i}(\mathbf{f}_i) \triangleq \hat{q}(\mathbf{f}_i)$

Step 3. $\hat{q}(\mathbf{f}_i) \approx \mathcal{N}\left(\mathbf{f}_i|\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\sigma}}_i^2\right) \hat{Z}_i$

Step 4: Compute parameters of $t_i(\mathbf{f}_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2)$ making moments of $q(\mathbf{f}_i)$ match those of $\hat{Z}_i \mathcal{N}\left(\mathbf{f}_i|\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\sigma}}_i^2\right)$.

Expectation Propgation - illustrated

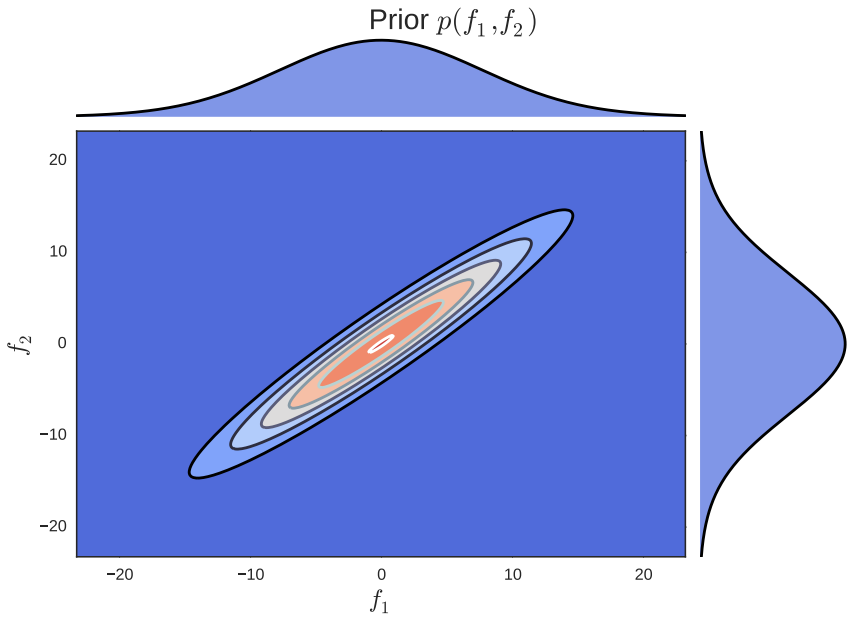
ILLUSTRATION OF EP. Titlted distribution, cavity distribution, etc.

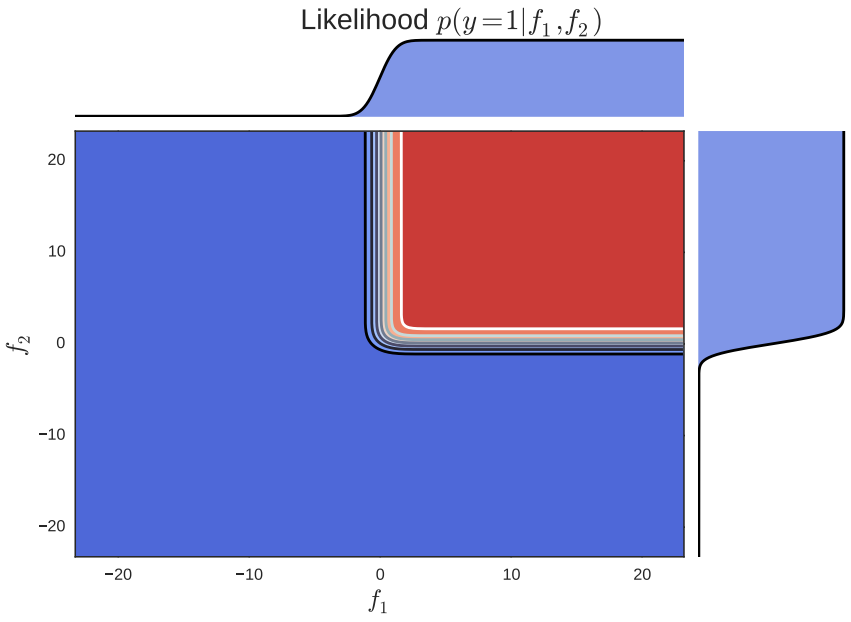
Sampling

ALTHOUGH NOT OUR FOCUS, WHY IS IT DIFFICUILT AND WHAT METHODS ARE CURRENTLY USED?

HOW CAN SAMPLES (HOWEVER THEY ARE OBTAINED) BE USED TO COMPUTE QUANTITIES OF INTEREST?

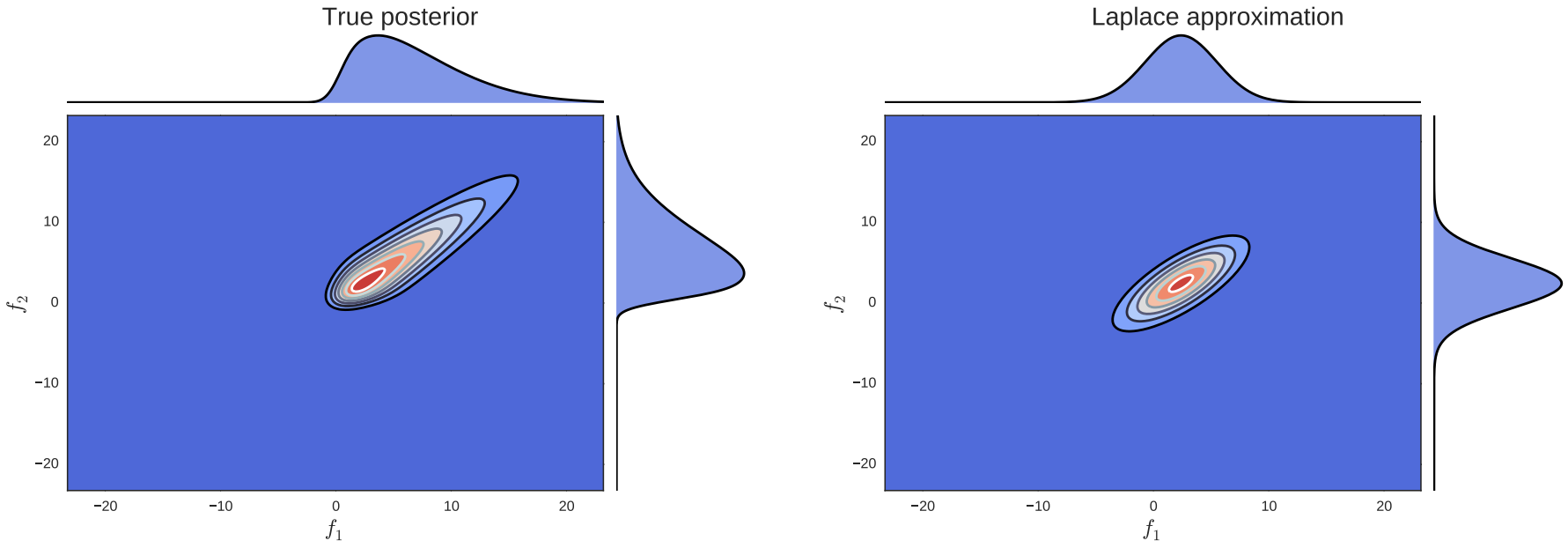
Comparing posterior approximations





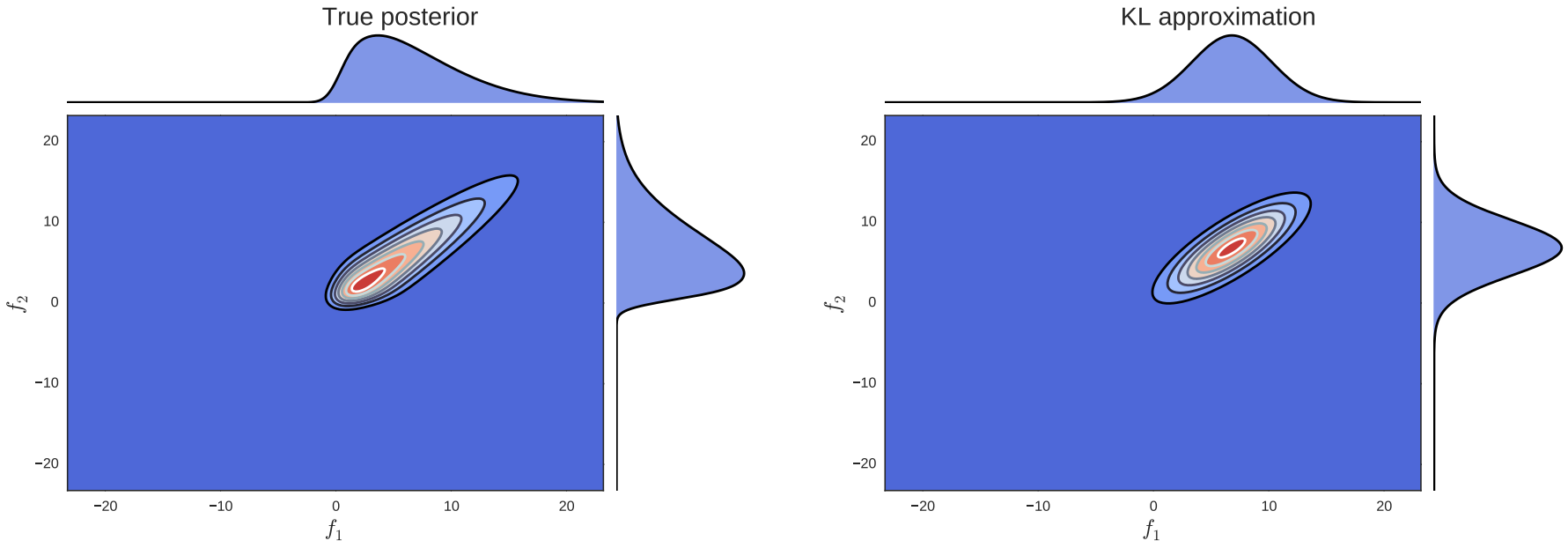
- Gaussian prior between two function values $\{\mathbf{f}_1, \mathbf{f}_2\}$, at $\{\mathbf{x}_1, \mathbf{x}_2\}$ respectively.
- Bernoulli likelihood, $\mathbf{y}_1 = 1$ and $\mathbf{y}_2 = 1$.

Comparing posterior approximations



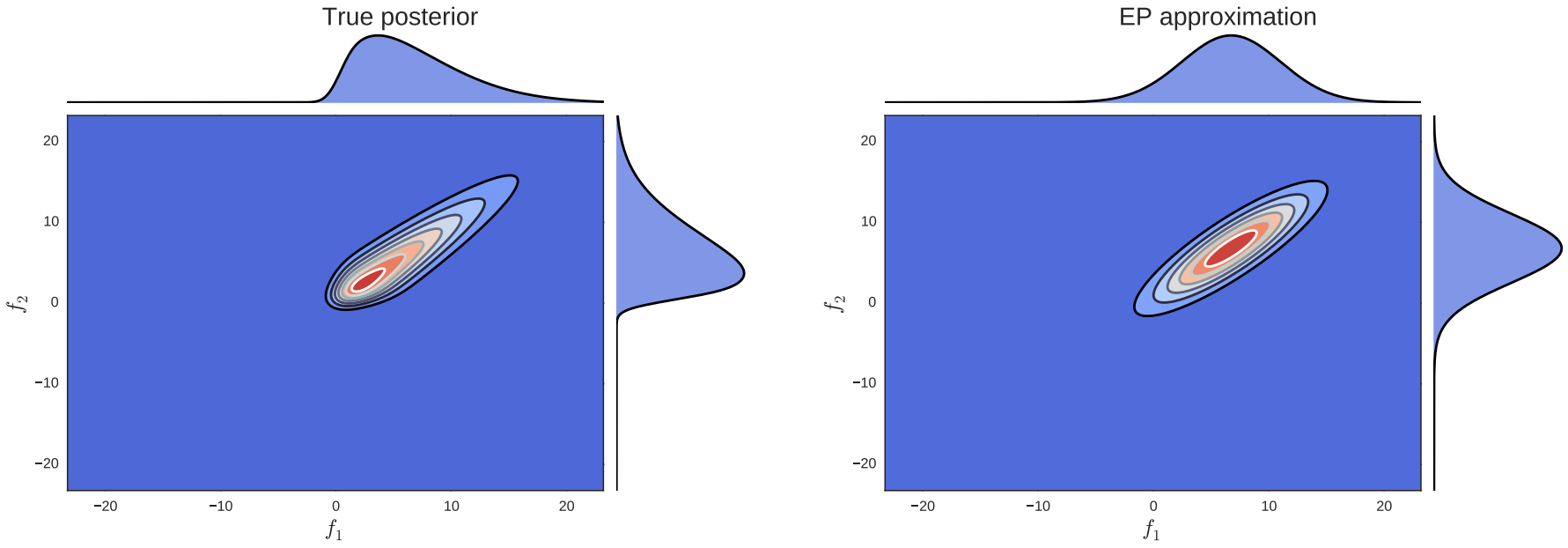
- $p(\mathbf{f}|\mathbf{y}) \propto \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{p(\mathbf{y})}$
- True posterior is non-Gaussian.
- Laplace approximates with a Gaussian at the mode of the posterior.

Comparing posterior approximations



- True posterior is non-Gaussian.
- KL approximate with a Gaussian that has minimal KL divergence, $\mathrm{KL}\left(q(\mathbf{f}) \,\|\, p(\mathbf{f}|\mathbf{y})\right)$.
- This leads to distributions that avoid regions in which $p(\mathbf{f}|\mathbf{y})$ is small.
- It has a large penality for assigning density where there is none.

Comparing posterior approximations

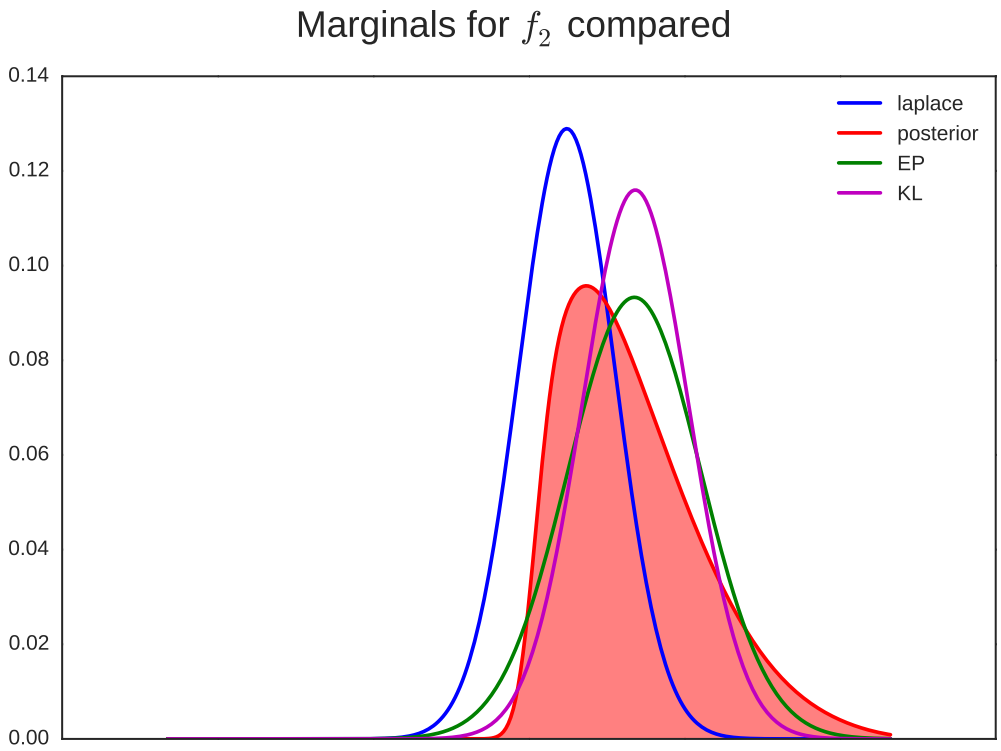


- True posterior is non-Gaussian.
- EP tends to try and put density where $p(\mathbf{f}|\mathbf{y})$ is large
- Cares less about assigning density density where there is none. Contrasts to KL method.

EP vs KL

ILLUSTRATE ON MULTI-MODAL THE EFFECT OF USING KL pq vs KL qp, i.e. BISHOP PLOT

Comparing posterior marginal approximations



- Laplace: Poor approximation.
- KL: Avoids assigning density to areas where there is none, at the expense of areas where there is some (right tail).
- EP: Assigns density to areas with density, at the expense of areas where there is none (left tail).

Pros - Cons - When - Laplace

Laplace approximation

- Pros
  - Very fast.
- Cons
  - Poor approximation if the mode does not well describe the posterior, for example Bernoulli likelihood.
- When
  - When the posterior *is* well characterized by its mode, for example Poisson.

Pros - Cons - When - KL

KL method

- Pros
  - Principled in that it we are directly optimizing a measure of divergence between an approximation and true distribution.
  - Lends itself to sparse extensions.
- Cons
  - Requires factorizing likelihoods to avoid n dimensional integral.
  - As seen, can result in underestimating the variance, i.e. becomes overconfident.
- When
  - Applicable to a range of likelihoods, but is known in some cases to underestimate variance, might need to be careful if you wish to be conservative with predictive uncertainty.
  - In conjunction with sparse methods.

Pros - Cons - When - EP

EP method

- Pros
  - Very effective for certain likelihoods (classification).
  - Also lends itself to sparse approximations.
- Cons
  - Standard algorithm is slow though possible to extend to sparse case.
  - Convergence issues for some likelihoods.
  - Must be able to match moments.
- When
  - Binary data (Nickisch and Rasmussen, 2008; Kuß, 2006), perhaps with truncated likelihood (censored data) (Vanhatalo et al., 2015).
  - In conjunction with sparse methods.

Pros - Cons - When - MCMC

MCMC methods

- Pros
  - Theoretical limit gives true distribution
- Cons
  - Can be very slow
- When
  - If time is not an issue, but exact accuracy is.
  - If you are unsure whether a different approximation is appropriate, can be used as a “ground truth”

Conclusion

MANY TASKS REQUIRE NON-GAUSSIAN OBSERVATION MODELS SOMETIMES BESPOKE LIKELIHOODS NEED TO BE CONSTRUCTED NON-GAUSSIAN LIKELIHOODS CAUSE COMPLICATIONS IN OUR FRAMEWORK DIFFERENT WAYS TO DEAL WITH THE PROBLEM, MANY ARE BASED ON GAUSSIAN APPROXIMATIONS DIFFERENT METHODS HAVE THEIR OWN ADVANTAGES AND DISADVANTAGES

Questions

Thanks for listening.

Any questions?

Bonus

HETEROSCEDASTIC LIKELIHOODS

References I

Hensman, J., Matthews, A. G. D. G., and Ghahramani, Z. (2015). Scalable variational gaussian process classification. In *In 18th International Conference on Artificial Intelligence and Statistics*, pages 1–9, San Diego, California, USA.

Kuß, M. (2006). *Gaussian Process Models for Robust Regression, Classification, and Reinforcement Learning*. PhD thesis, TU Darmstadt.

Nickisch, H. and Rasmussen, C. E. (2008). Approximations for Binary Gaussian Process Classification. *Journal of Machine Learning Research*, 9:2035–2078.

Vanhatalo, J., Riihimaki, J., Hartikainen, J., Jylanki, P., Tolvanen, V., and Vehtari, A. (2015). Gpstuff. `http://mloss.org/software/view/451/`.