Non Gaussian Likelihoods

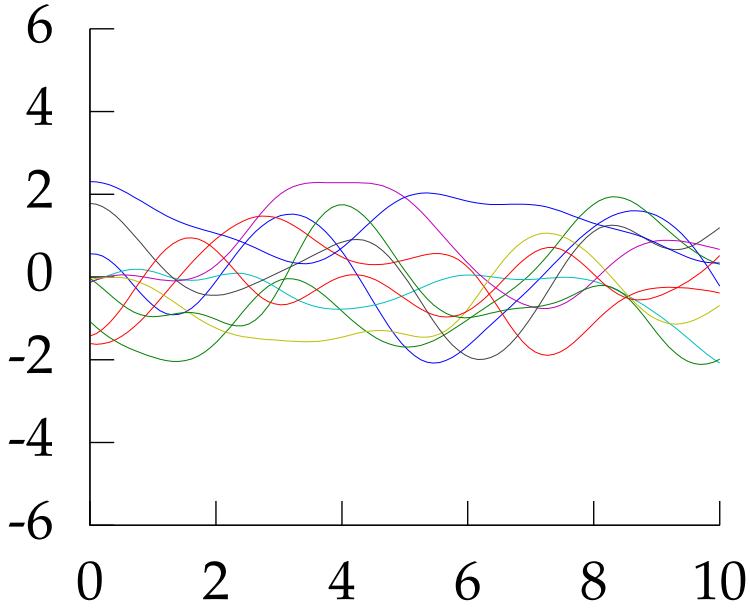
Neil Lawrence

GPRS
13th February 2014

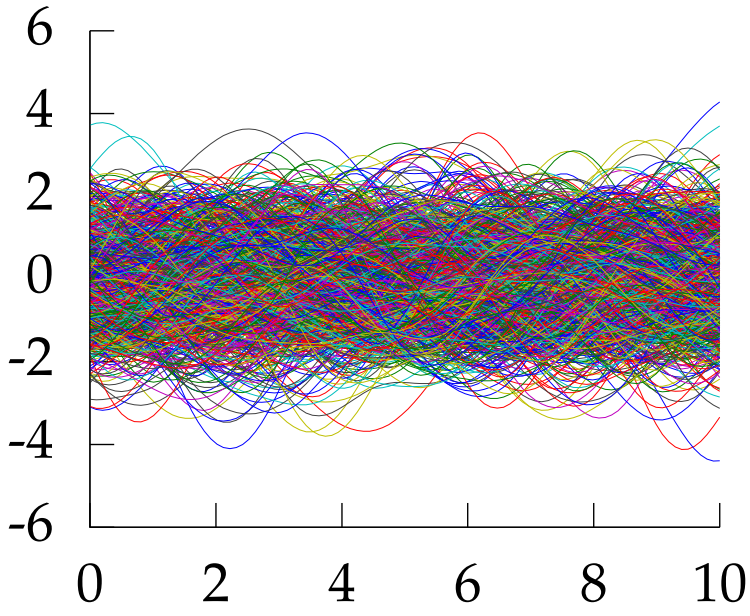
Outline

Outline

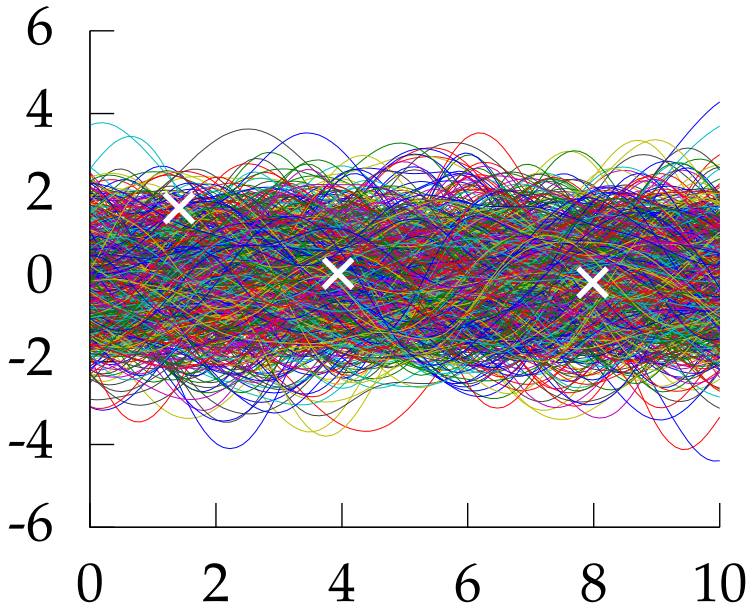
Gaussian Processes



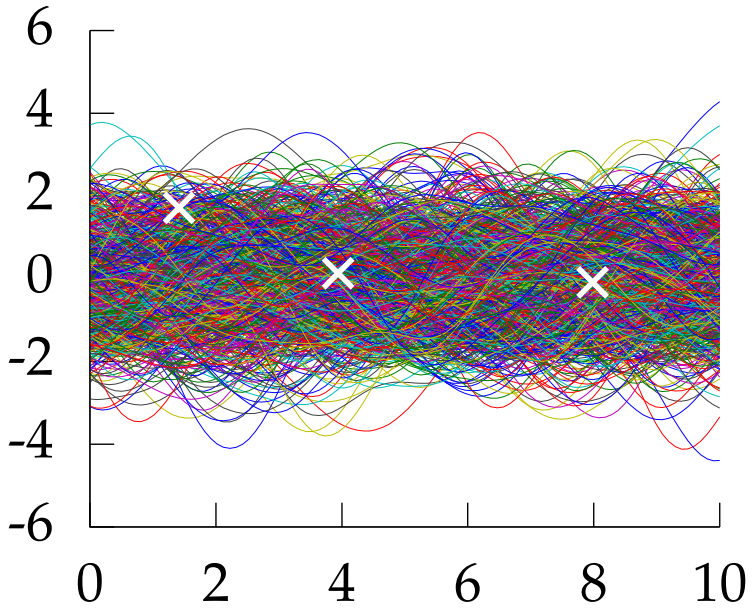
Gaussian Processes

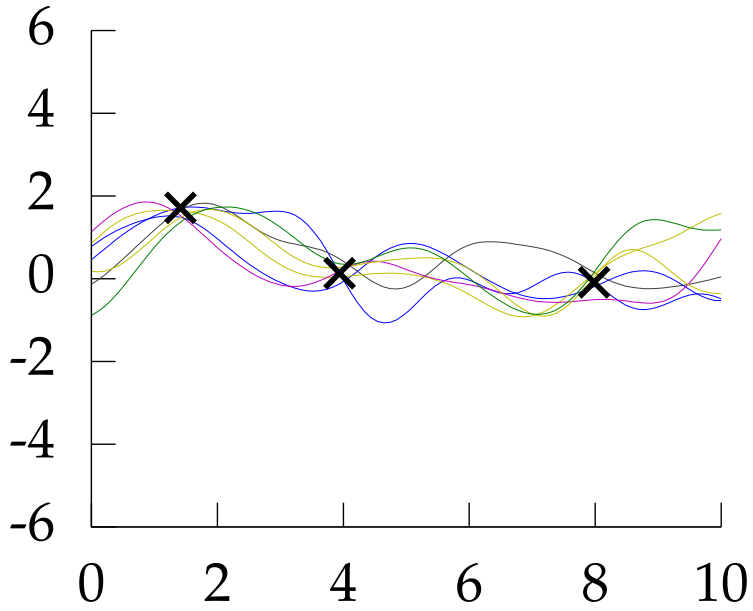


Gaussian Processes



Gaussian Processes





GP Regression

Analytical tractability of the posterior distribution is assured:

- Gaussian prior:

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K_{ff}})$$

- Gaussian likelihood:

$$\prod_{i=1}^{n} p(y_i|f_i) = \mathcal{N}\left(\mathbf{y}|\mathbf{f}, \sigma_i^2\mathbf{I}\right)$$

- Gaussian posterior:

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

Bernoulli Distribution

- A mathematical switch allows us to write a probability table as a function.

$$P(Y = 1) = \pi$$

$$P(Y = 0) = (1 - \pi)$$

- Write as a function

$$P(Y = y) = \pi^y (1 - \pi)^{1-y}$$

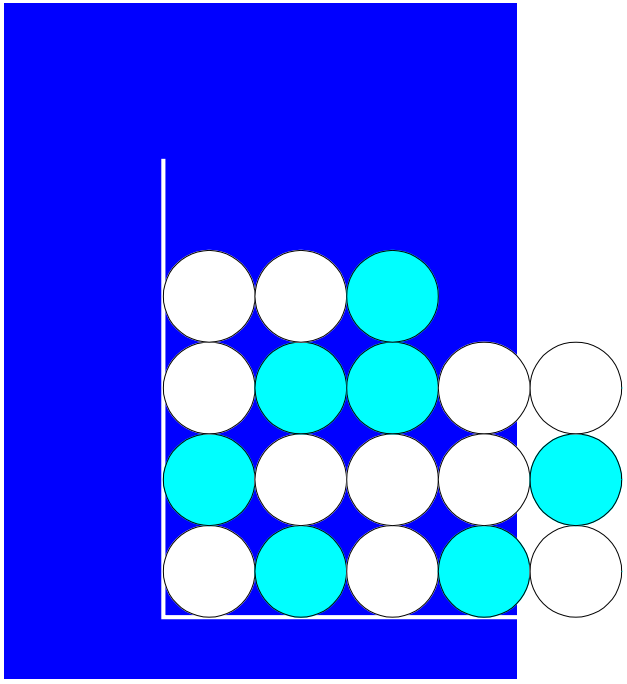
- Can think of this construction as a "mathematical switch". Known as the Bernoulli distribution.
- Widely used in classification algorithms: π parameter is made to be dependent on "inputs".

Binomial Distribution

- Generalization of Bernoulli to multiple trials.
- Jakob Bernoulli: black and red balls in an urn. Proportion of red is π.
- Sample with replacement. Binomial gives the distribution of number of reds, y, from S extractions

$$P(y|\pi, S) = \frac{S!}{y!(S-y)!}\pi^y(1-\pi)^{(S-y)}$$

- Mean is given by $S\pi$ and variance $S\pi(1-\pi)$.



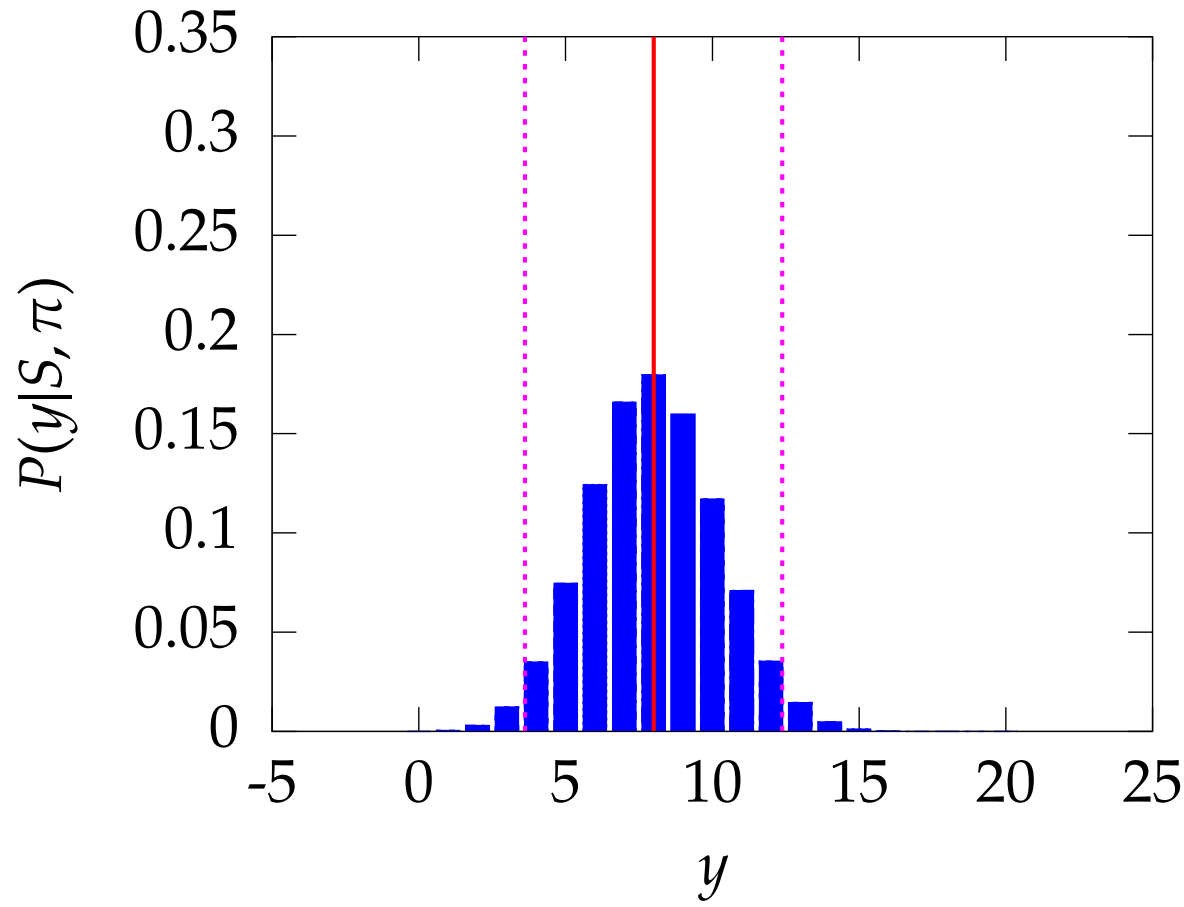


Figure : The binomial distribution for $\pi = 0.4$ and $S = 20$. Mean is shown as red line, 2 standard deviations are magenta.

The Gamma Density

- Density over positive real values.

$$p(y|a,b) = \frac{b^a}{\Gamma(a)} y^{a-1} \exp(-by)$$
$$= \mathcal{G}\left(y|\mu,\sigma^2\right)$$

- Mean is $\frac{a}{b}$ and variance is $\frac{a}{b^2}$.
- Also available in multivariate as the Wishart (positive definite matrices).

Gamma PDF I

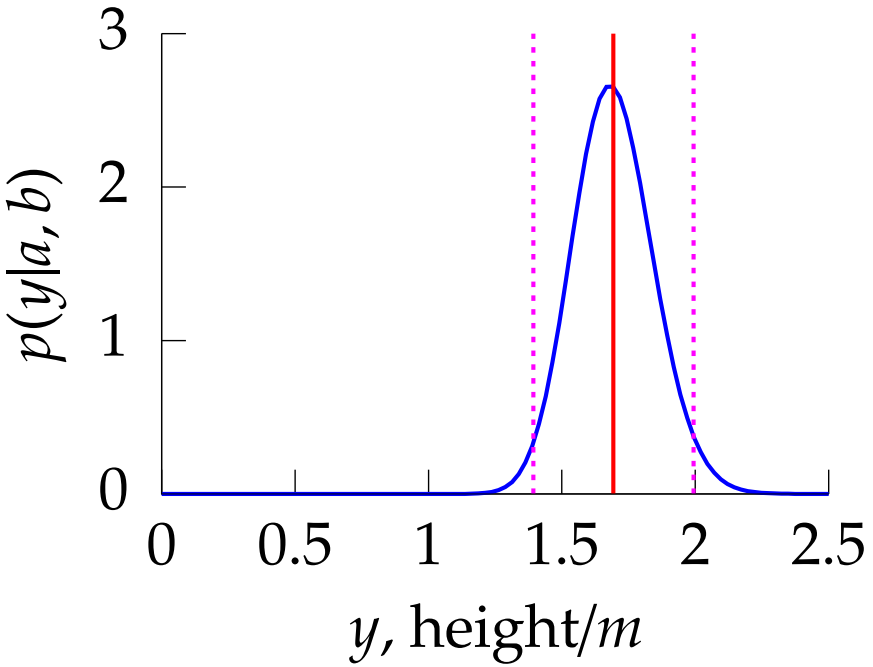


Figure : The Gamma PDF with $a = 127$ and $b = 75$. Here it represents the heights of a population of students and constrains them positive.

Gamma PDF I

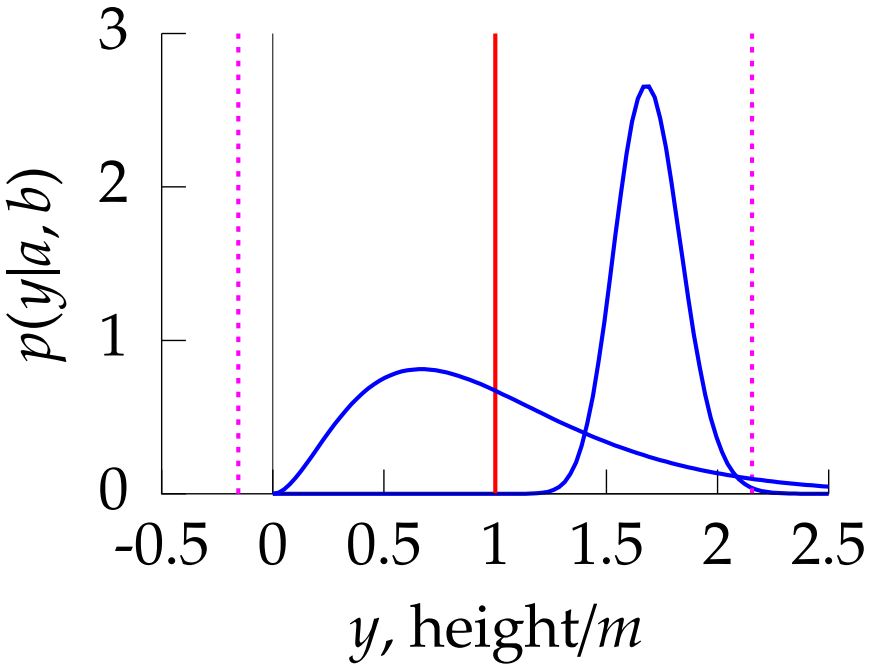


Figure : The Gamma PDF with $a = 127$ and $b = 75$ alongside a Gamma PDF with $a = 3$ and $b = 3$.

Categorical Distribution

Multiple outcomes, example: die roll.

die role	probability	$\mathbf{y}$
1	π_1	[1 0 0 0 0 0]
2	π_2	[0 1 0 0 0 0]
3	π_3	[0 0 1 0 0 0]
4	π_4	[0 0 0 1 0 0]
5	π_5	[0 0 0 0 1 0]
6	π_6	[0 0 0 0 0 1]

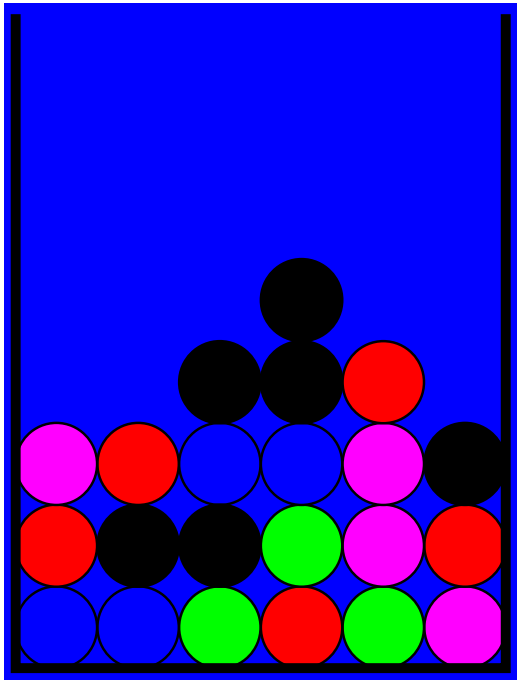
$$P(\mathbf{y}) = \prod_{i=1}^{k} \pi_i^{y_i}$$

Multinomial Distribution

- Generalization of categorical to multiple trials.
- Generalization of binomial to multiple outcomes. Proportion of each colour ball is now π_i.
- Sample with replacement. Multinomial gives the distribution of number of each of k different balls, y, from S extractions

$$P(y|\pi, S) = \frac{S!}{\prod_{i=1}^{k} y_i!} \prod_{i=1}^{k} \pi_i^{y_i}$$

- Mean for each colour is given by $S\pi_i$ and variance $S\pi_i(1 - \pi_i)$.



Distributions as Functions

- Probability distribution with a simple table can be limiting.
- The Poisson Distribution — a distribution a a function
- First published by **Siméon Denis Poisson** (1781-1840) in 1837.
- Defined over the space of all non-negative integers.
- This set is countably infinite: impossible to summarise in a table!
- The Poisson distribution is therefore defined as

$$P(y|\mu) = \frac{\mu^y}{y!} \exp(-\mu). \tag{1}$$

where y is any integer from 0 to ∞, and μ is a parameter of the distribution.

A Poisson with $\mu = 2$

- To work out the probability of y in a Poisson $\mu = 2$ we can start filling a table.
- The values in a table are computed from (1)

y	0	1	2	...
$P(y)$	0.135	0.271	0.271	...

Table : Some values for the Poisson distribution with $\mu = 2$.

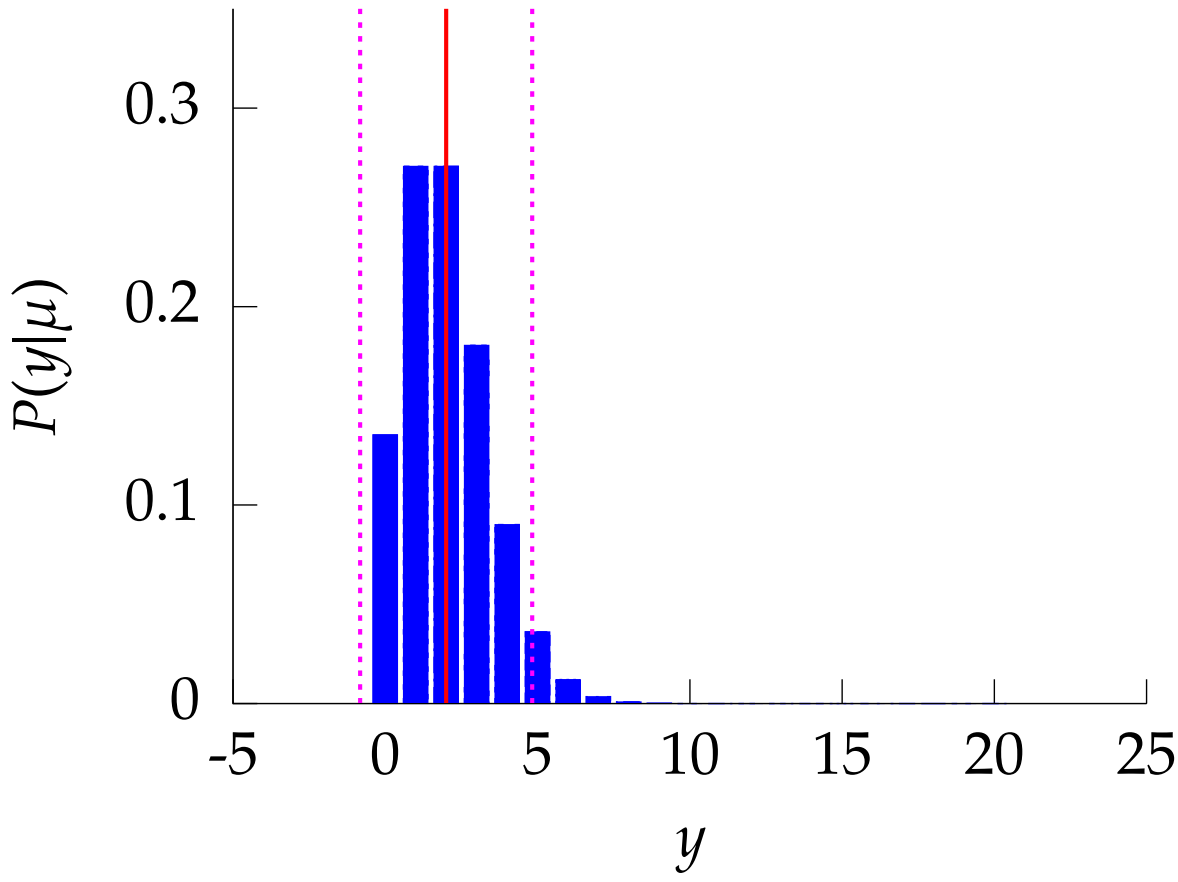


Figure : The Poisson distribution for $\mu = 2$. Mean is given by μ (red line), standard deviation is given by $\sqrt{\mu}$ (magenta lines show 2 standard deviations).

Gaussian Noise

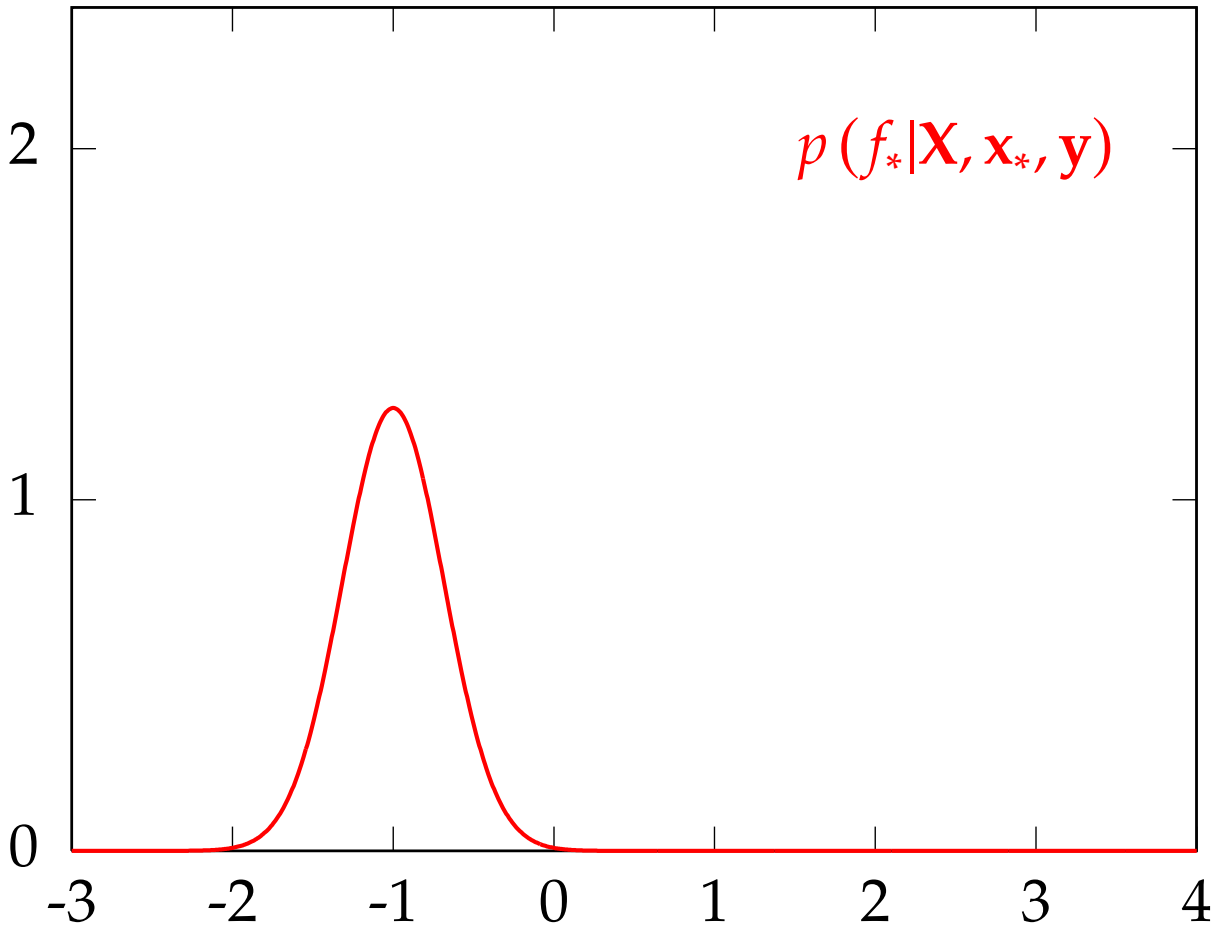


Figure : Inclusion of a data point with Gaussian noise.

Gaussian Noise

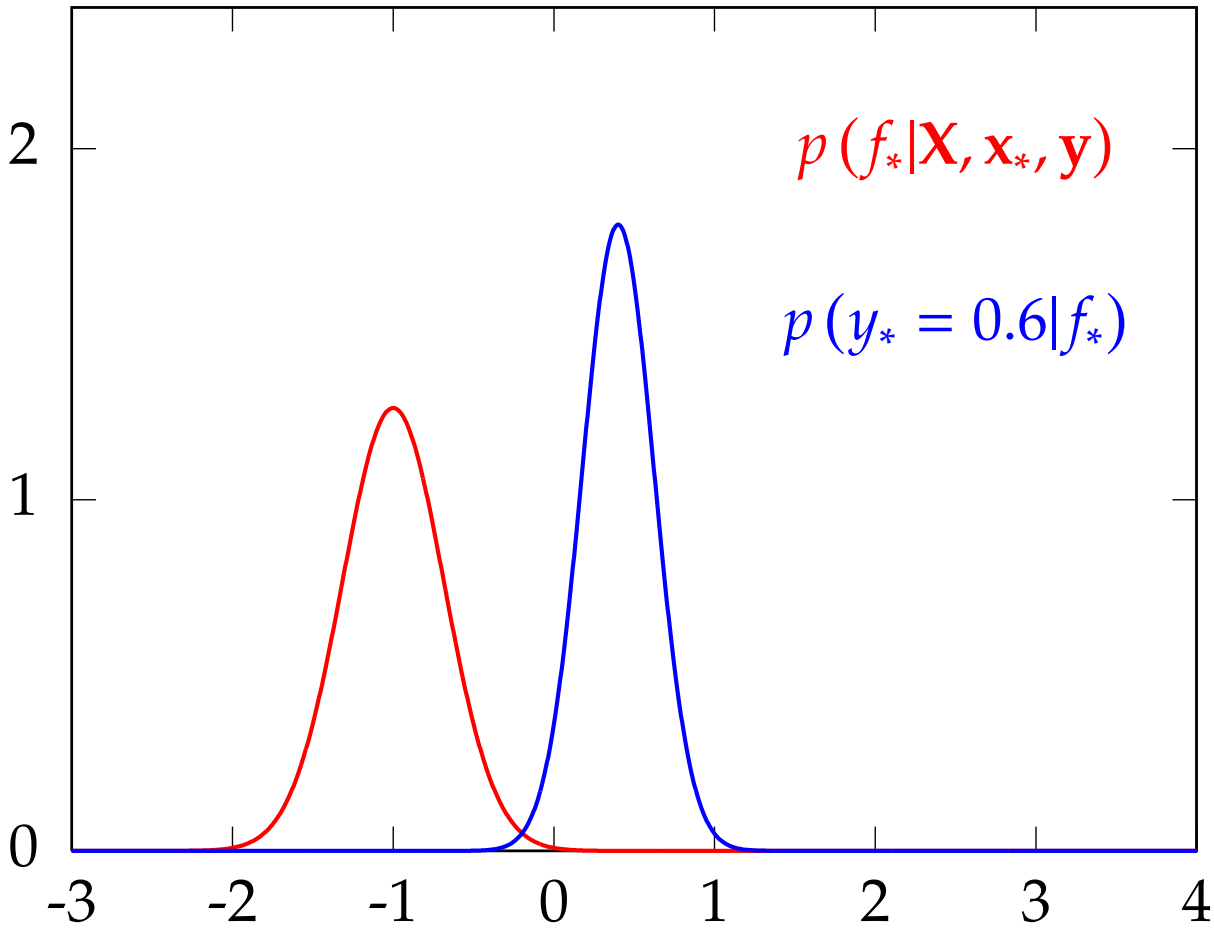


Figure : Inclusion of a data point with Gaussian noise.

Gaussian Noise

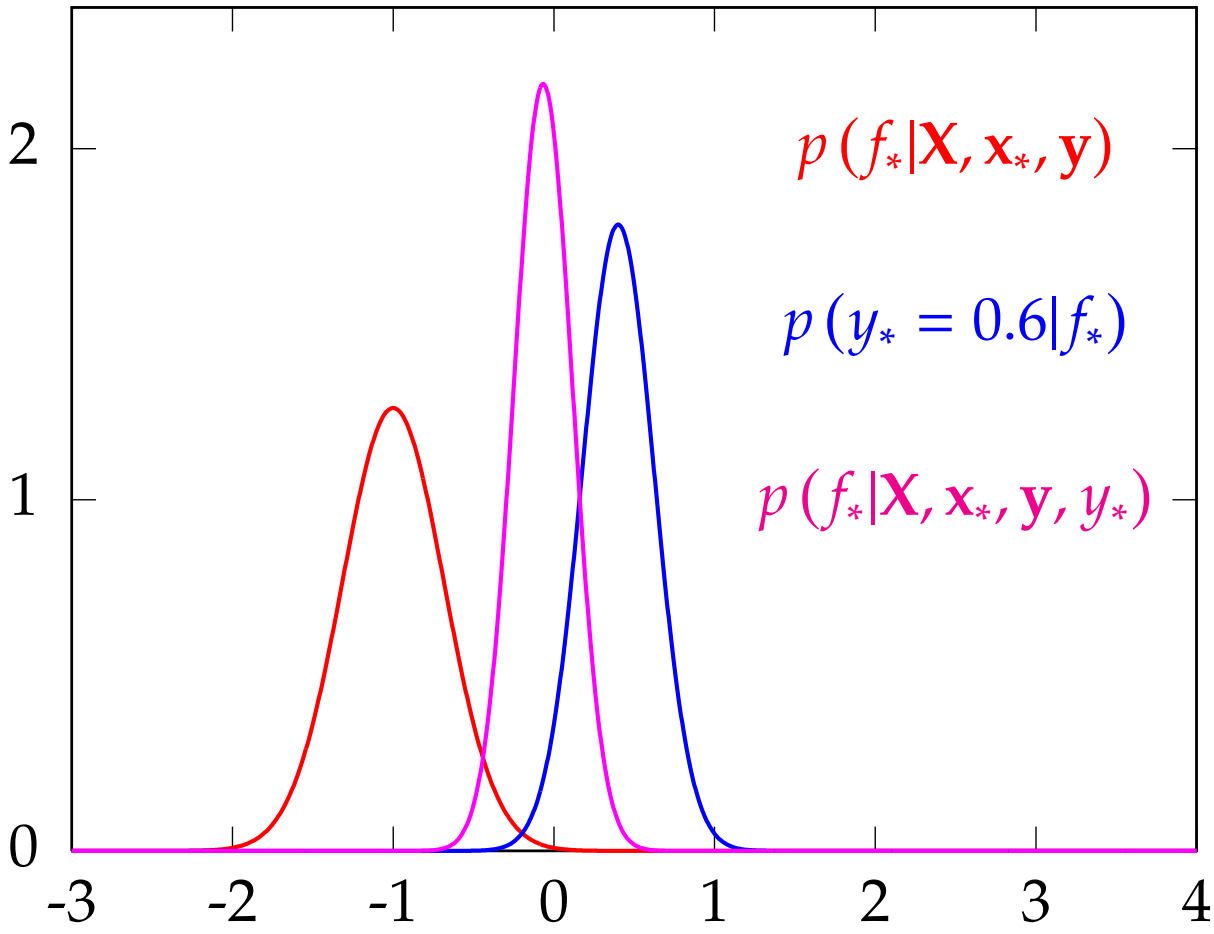


Figure : Inclusion of a data point with Gaussian noise.

Classification Noise Model

Probit Noise Model

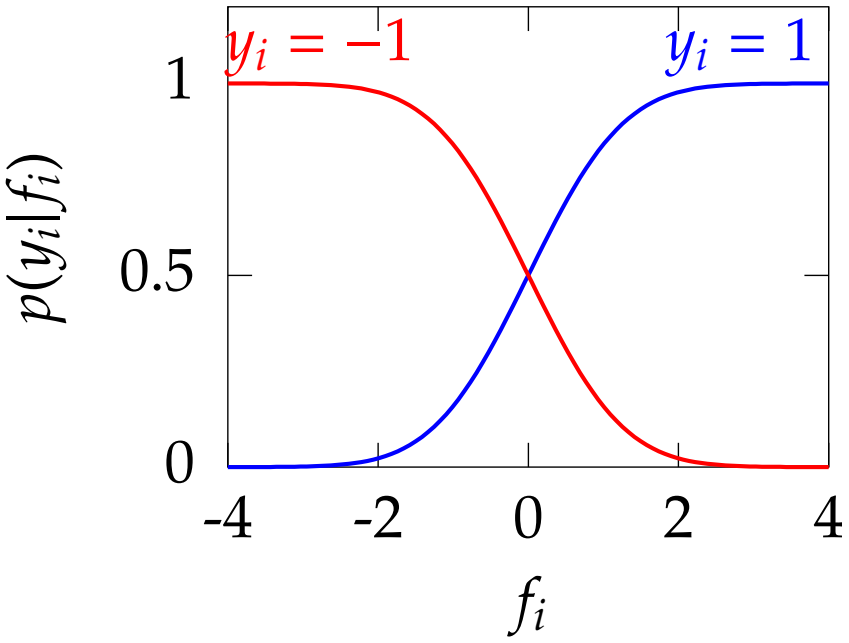


Figure : The probit model (classification). The plot shows $p(y_i|f_i)$ for different values of y_i. For $y_i = 1$ we have

$p(y_i|f_i) = \phi(f_i) = \int_{-\infty}^{f_i} \mathcal{N}(z|0,1)\,dz.$

Ordinal Noise Model

Ordered Categories

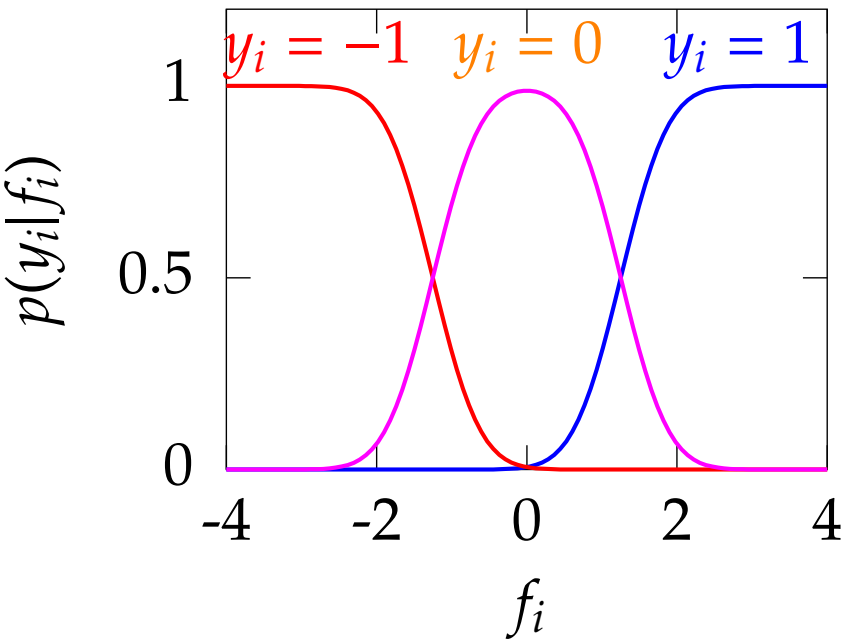


Figure : The ordered categorical noise model (ordinal regression). The plot shows $p(y_i|f_i)$ for different values of y_i. Here we have assumed three categories.

Null Category Noise Model

Classification with a Missing Category

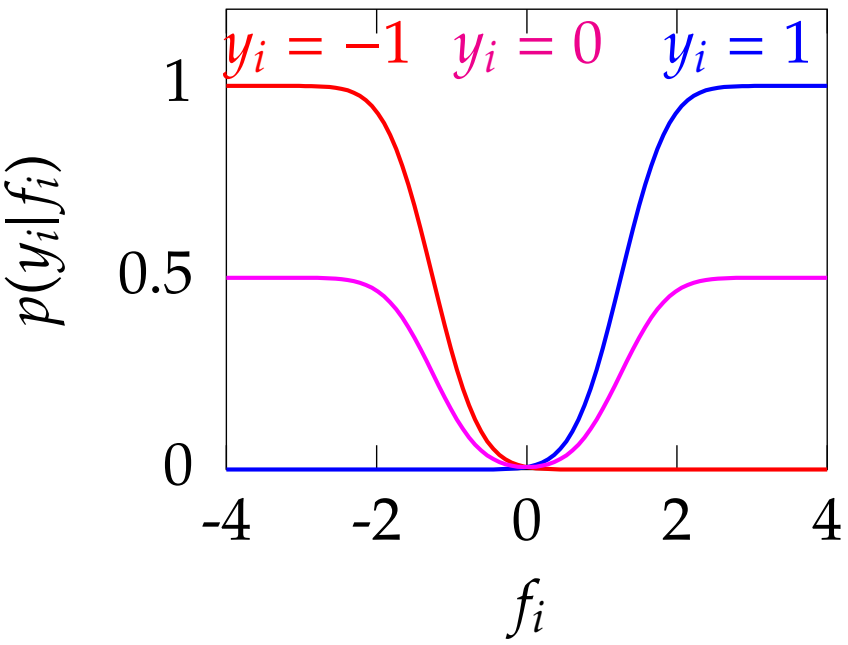


Figure : The null category noise model (semi-supervised learning). The plot shows $p(y_i|f_i)$ for different values of y_i. Here we have assumed three categories.

Non-linear Response Functions

- Non Gaussian likelihood:

$$p(y_i|f_i) = \Phi(f_i)$$

- Exact computation of the posterior is no longer possible analytically.

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

Outline

Link Functions

- Take the output of our function, $f(\cdot)$ use as:
  - Success probability in binomial distribution.
  - Rate function in Poisson likelihood.
  - shape parameter of Gamma distribution.
- Problem: $f(\cdot)$ defined over real line.
- Needs to be squashed down to 0-1 or constrained positive.

Link Functions

- Log link function, model the log rate.

$$\log \lambda(\mathbf{x}) = f(\mathbf{x})$$

- Logit link function, model the log odds.

$$\frac{\log \pi(\mathbf{x})}{\log(1 - \pi(\mathbf{x}))} = f(\mathbf{x})$$

Generative Model

- From a generative perspective we often naturally think of the inverse link:

$$\lambda(\mathbf{x}) = \exp(f(\mathbf{x}))$$

$$\pi(\mathbf{x}) = \frac{1}{1 + \exp(-f(\mathbf{x}))}$$

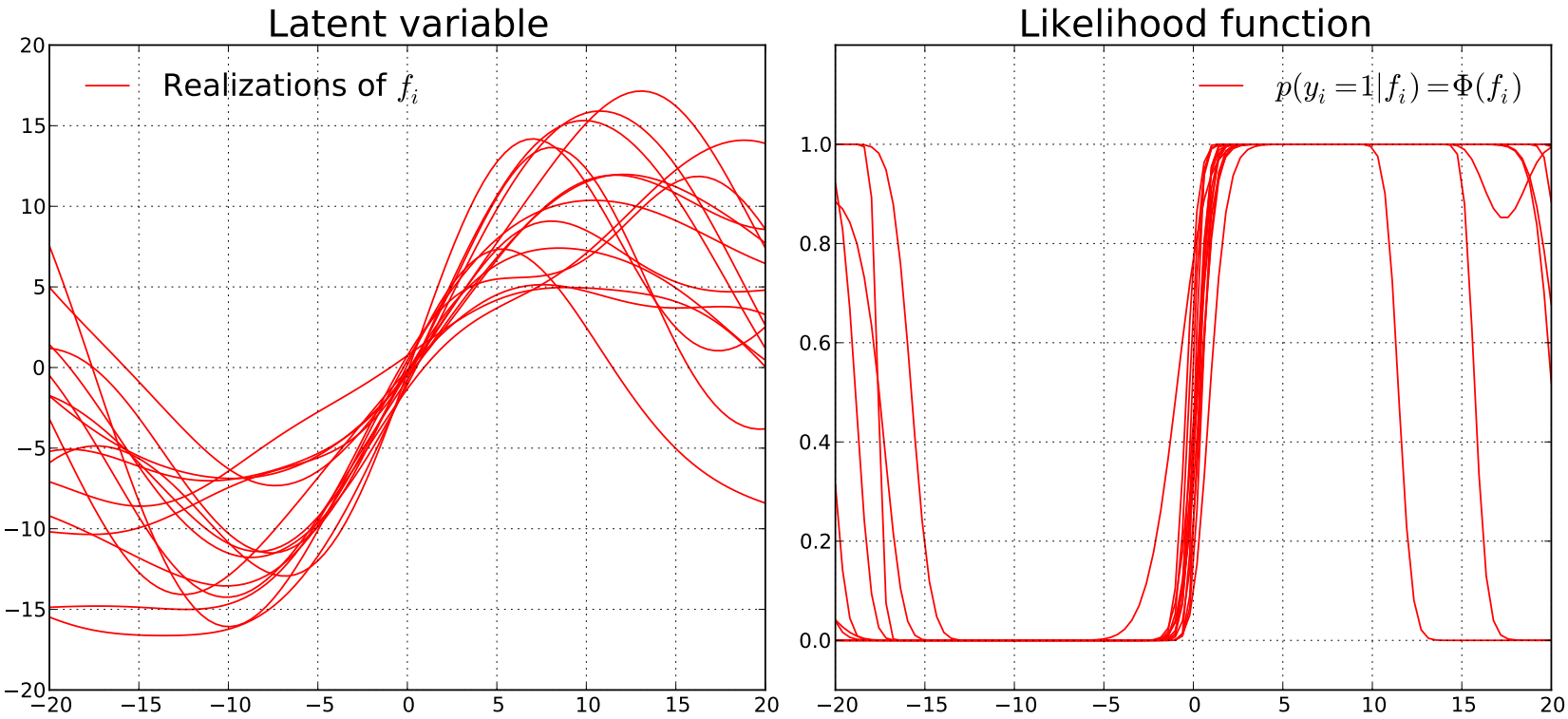
- Can make some assumptions of the link function clearer. For example log additive link function:

$$\log \lambda(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x})$$

is a product of functions:

$$\lambda(\mathbf{x}) = \exp(f_1(\mathbf{x}))\exp(f_2(\mathbf{x}))$$

Example: Logit/Probit Link Function



Outline

Laplace Approximation

- Second order Taylor expansion at mode of log likelihood.
- First suggested by Laplace for his English dice example.
- How Laplace independently (of de Moivre) reinvented the Gaussian density.

Laplace Approximation

$$\log p(\mathbf{f}|\mathbf{y}) = \log p(\mathbf{y}|\mathbf{f}) + \log p(\mathbf{f}) + \text{const}$$

$$\log p(\mathbf{f}|\mathbf{y}) = \log p(\mathbf{y}|\mathbf{f}) - \frac{1}{2}\mathbf{f}^\top \mathbf{K}_{\mathbf{ff}}^{-1}\mathbf{f}$$

- Find MAP estimate $\hat{\mathbf{f}}$. This is mean of Gaussian approximation.
- Find Hessian of this system.
- Covariance of approximation is $-\mathbf{H}^{-1}$.

$$\mathbf{H} = \left(\frac{\mathrm{d}^2 \log p(\mathbf{y}|\mathbf{f})}{\mathrm{d}f_i \mathrm{d}f_j}\right)_{ij} - \mathbf{K}_{\mathbf{ff}}^{-1}$$

Outline

Expectation Propagation: General Case

- Exact (intractable) posterior:

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

- EP posterior approximation:

$$q(\mathbf{f}|\mathbf{y}) = \frac{\prod_{i=1}^{K} t_i(f_i)}{Z_{EP}}$$

Expectation Propagation: Gaussian Approximation

Consider the special case:

$$p(y_i|f_i) \approx t_i(f_i) = Z_i \mathcal{N}\left(\tilde{\mu}_i|f_i, \tilde{\sigma}_i^2\right)$$

Here Z_i is a scaling factor so t_i is unnormalized.

If

$$p(\mathbf{f}) \sim \mathcal{N}\left(\mathbf{0}, \mathbf{K_{ff}}\right).$$

No approximation needed.

EP Posterior Approximation

$$q(\mathbf{f}|\mathbf{y}) = \frac{\prod_{i=1}^{n} t(f_i)p(\mathbf{f})}{Z_{EP}} = \mathcal{N}\left(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

Site functions provide "fake Gaussian observations" with target value $\hat{\mu}_i$ and observation variance $\hat{\sigma}_i^2$.

$$Z_{EP} = \prod_{i=1}^{n} Z_i \int \prod_{i=1}^{n} \mathcal{N}\left(\hat{\mu}_i|f_i, \hat{\sigma}_i^2\right) p(\mathbf{f})\mathrm{d}\mathbf{f}$$

EP Posterior Approximation

$$q(\mathbf{f}|\mathbf{y}) = \frac{\prod_{i=1}^{n} Z_i \mathcal{N}\left(\hat{\mu}_i | f_i, \hat{\sigma}_i^2\right) p(\mathbf{f})}{Z_{EP}} = \mathcal{N}\left(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

Site functions provide "fake Gaussian observations" with target value $\hat{\mu}_i$ and observation variance $\hat{\sigma}_i^2$.

$$Z_{EP} = \prod_{i=1}^{n} Z_i \int \prod_{i=1}^{n} \mathcal{N}\left(\hat{\mu}_i | f_i, \hat{\sigma}_i^2\right) p(\mathbf{f}) \mathrm{d}\mathbf{f}$$

Site approximations

- Given initial site approximations: $t_j(f_j)$ for $j \neq i$.
- Need to set

$$t_i(f_i) \approx p(y_i|f_i)$$

$$p(y_i|f_i)p(\mathbf{f})\prod_{j\neq i} t_j(f_j) \approx p(\mathbf{f})\prod_{j=1}^{n} t_j(f_j)$$

$$p(y_i|f_i)\int p(\mathbf{f})\prod_{j\neq i} t_j(f_j)\mathrm{d}f_{j\neq i} \approx \int p(\mathbf{f})\prod_{j=1}^{n} t_j(f_j)\mathrm{d}f_{j\neq i}$$

$$p(y_i|f_i)q_{\setminus i}(f_i) \approx \mathcal{N}\left(f_i|\hat{\mu}_i, \hat{\sigma}_i^2\right)\hat{Z}_i$$

Cavity Distribution

$$q_{\setminus i}(f_i) = \frac{\prod_{j \neq i} t(f_j) p(\mathbf{f})}{\int \prod_{j \neq i} t(f_j) p(\mathbf{f})} \mathrm{d}\mathbf{f}$$

Tilted Distribution

$$\hat{p}_i(f_i|y_i) = \frac{p(y_i|f_i)q_{\setminus i}(f_i)}{\hat{Z}}$$

where

$$\hat{Z}_i = \int p(y_i|f_i)q_{\setminus i}(f_i)\mathrm{d}f_i$$

Minimization of the KL divergence

$$\hat{\mu}_i, \hat{\sigma}_i = \text{argmin}_{\hat{\mu}_i, \hat{\sigma}_i} \text{KL}\left(\frac{p(y_i|f_i) q_{\setminus i}(f_i)}{\hat{Z}} \| \mathcal{N}\left(f_i | \hat{\mu}_i, \hat{\sigma}_i^2\right) \right)$$

This is the KL between *tilted distribution* and *marginal of approximation*.
Since the approximation is Gaussian, KL is minimal when:

- $\hat{\mu}_i = \langle f_i \rangle_{p(y_i|f_i) q_{\setminus i}(f_i)}$
- $\hat{\sigma}_i^2 = \langle f_i \rangle^2_{p(y_i|f_i) q_{\setminus i}(f_i)} - \tilde{\mu}_i^2$

Scale of Site Approximation

- Since the approximation is un-normalized, we set scale as follows:

$$\hat{Z}_i = \int p(y_i|f_i) q_{\setminus i}(f_i) \mathrm{d}f_i$$

Classification Noise Model

Probit Noise Model

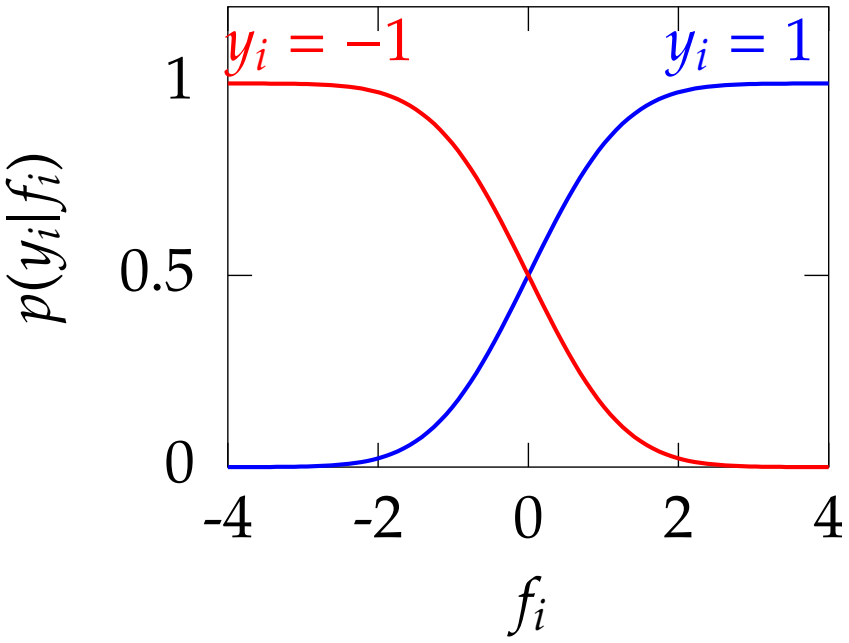


Figure : The probit model (classification). The plot shows $p(y_i|f_i)$ for different values of y_i. For $y_i = 1$ we have

$p(y_i|f_i) = \phi(f_i) = \int_{-\infty}^{f_i} \mathcal{N}(z|0,1)\,dz.$

Classification

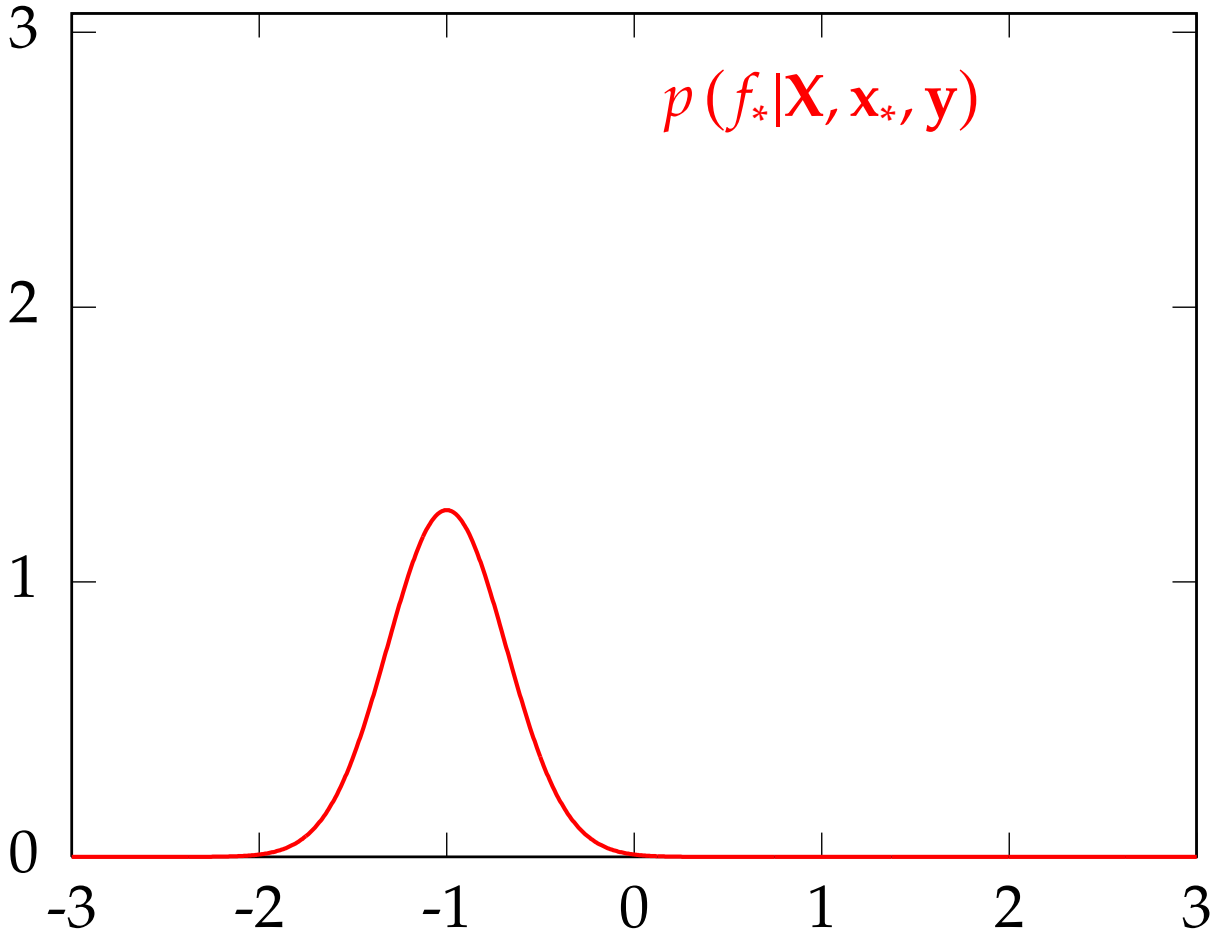


Figure : An EP style update with a classification noise model.

Classification

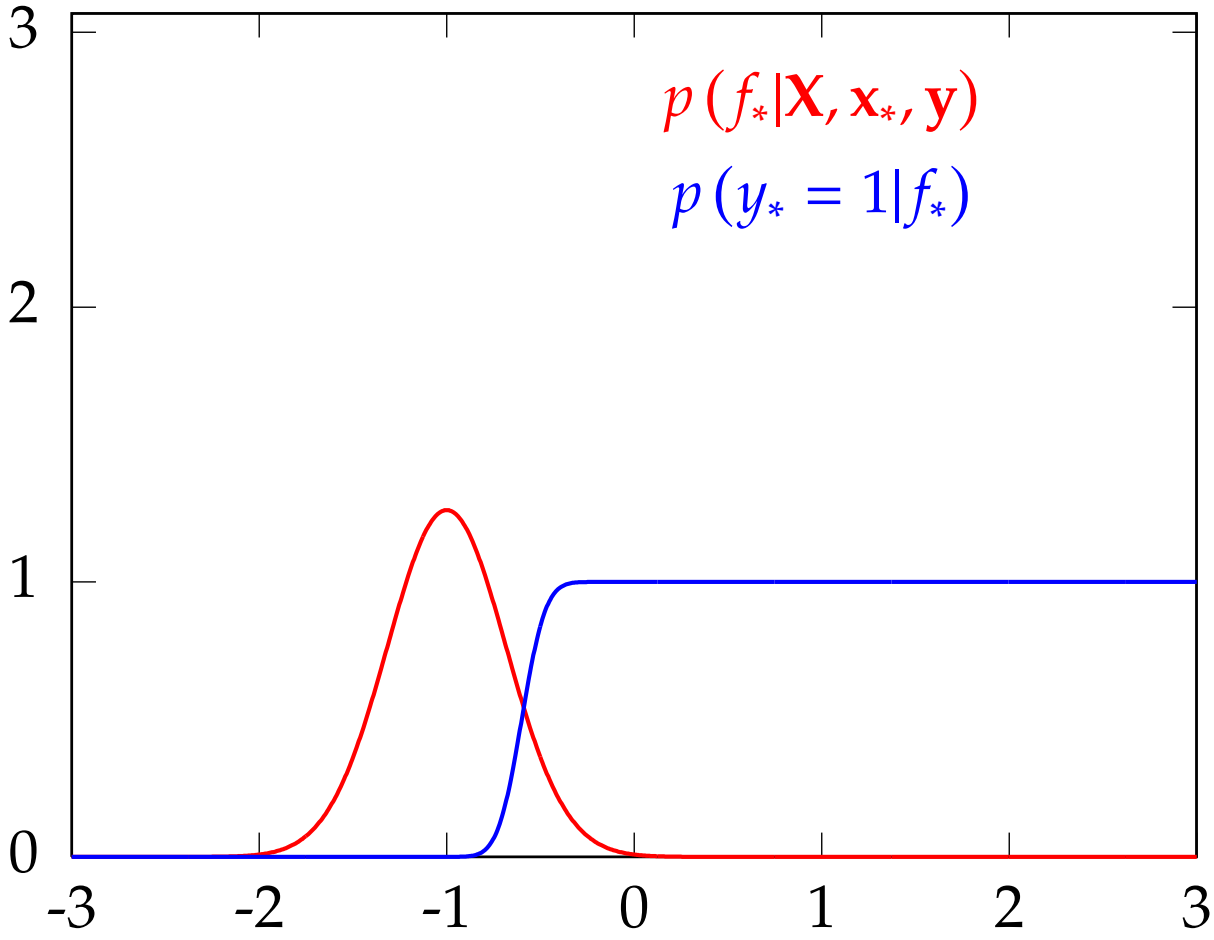


Figure : An EP style update with a classification noise model.

Classification

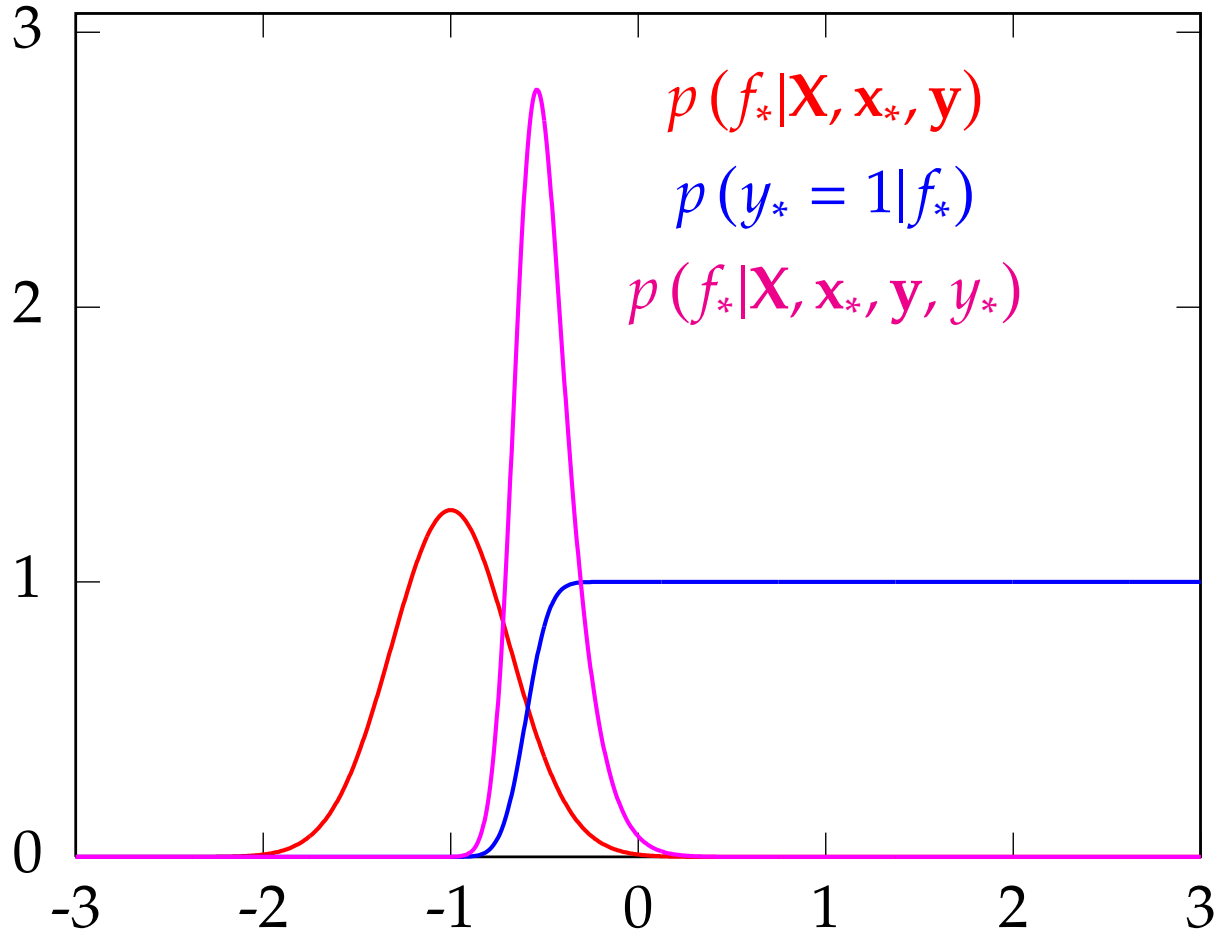


Figure : An EP style update with a classification noise model.

Classification

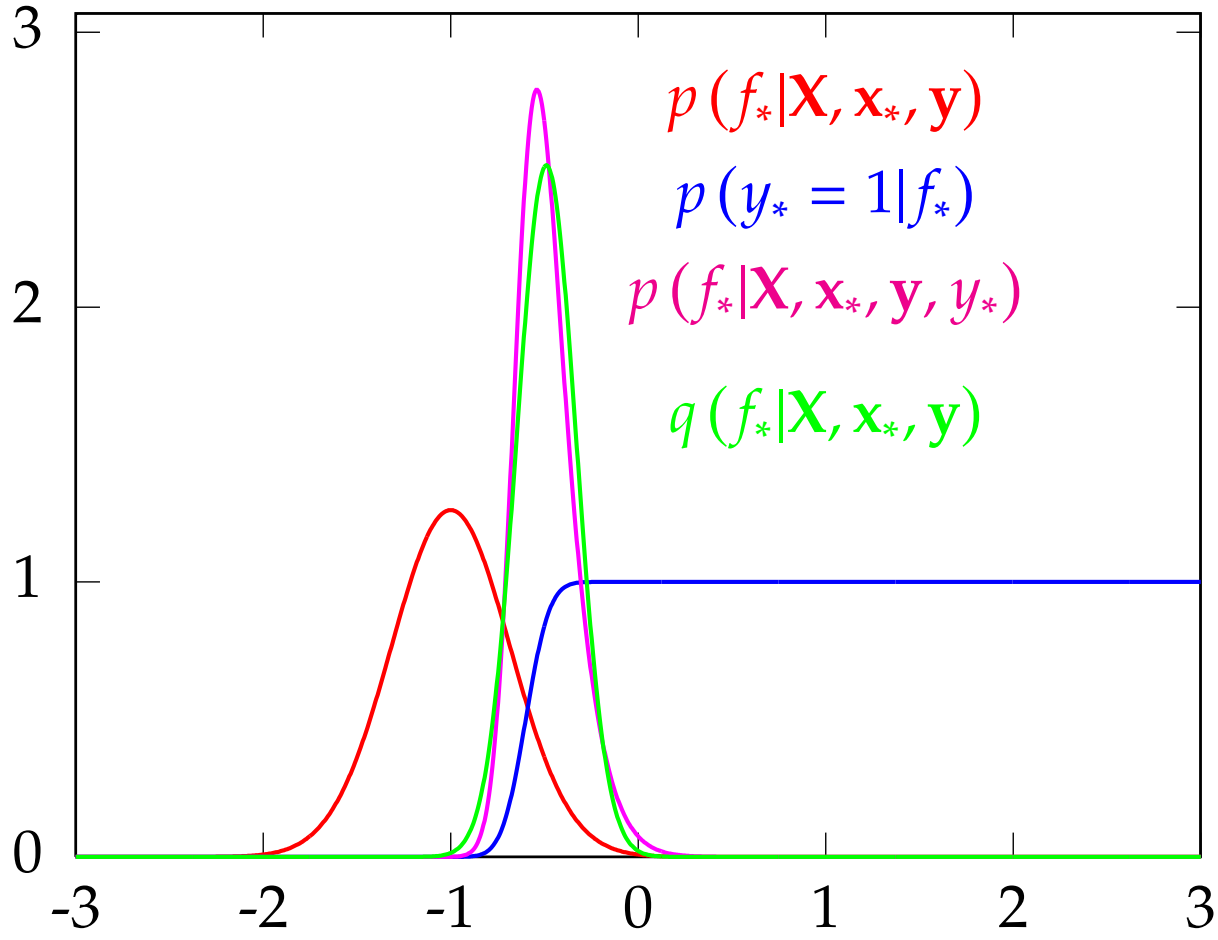


Figure : An EP style update with a classification noise model.

Ordinal Noise Model

Ordered Categories

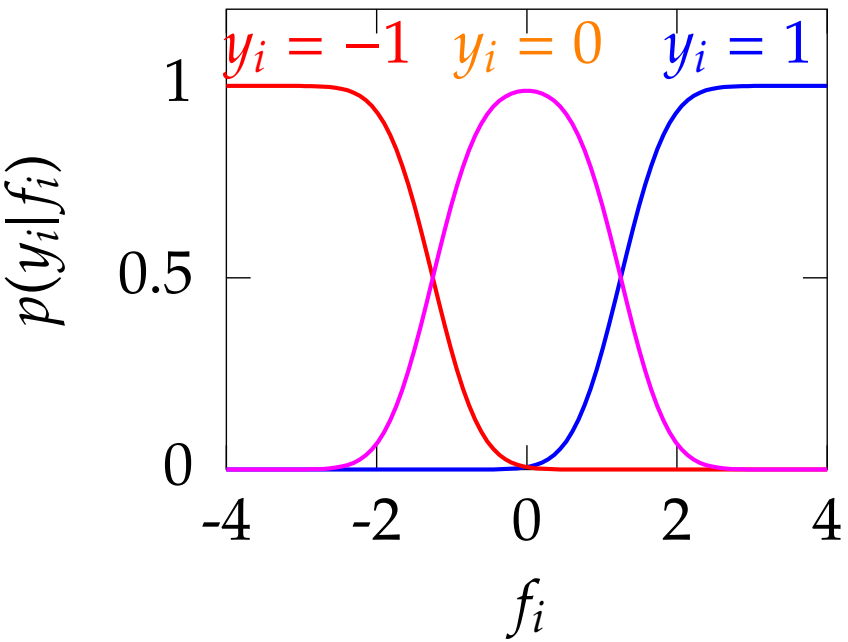


Figure : The ordered categorical noise model (ordinal regression). The plot shows $p(y_i|f_i)$ for different values of y_i. Here we have assumed three categories.

Ordinal Regression

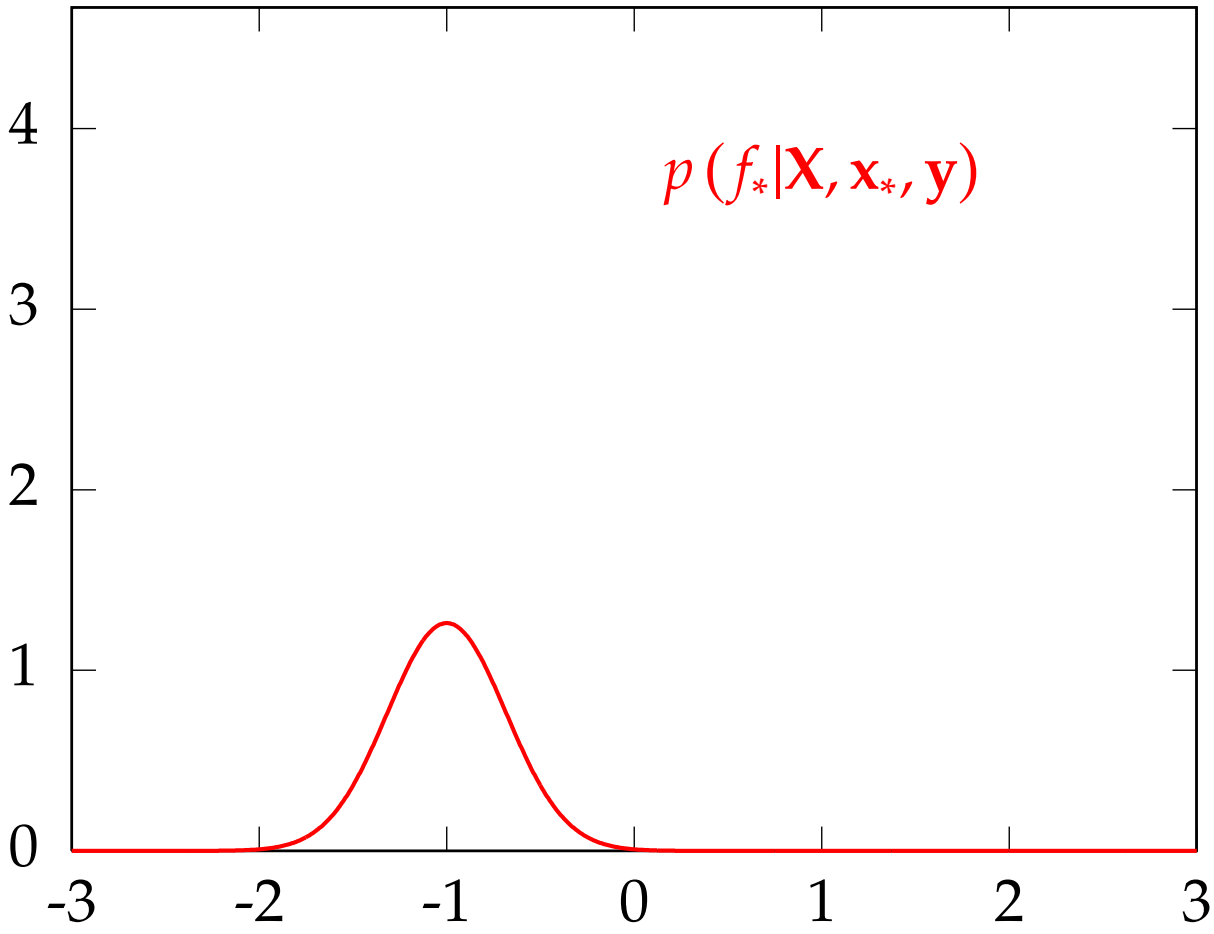


Figure : An EP style update with an ordered category noise model.

Ordinal Regression

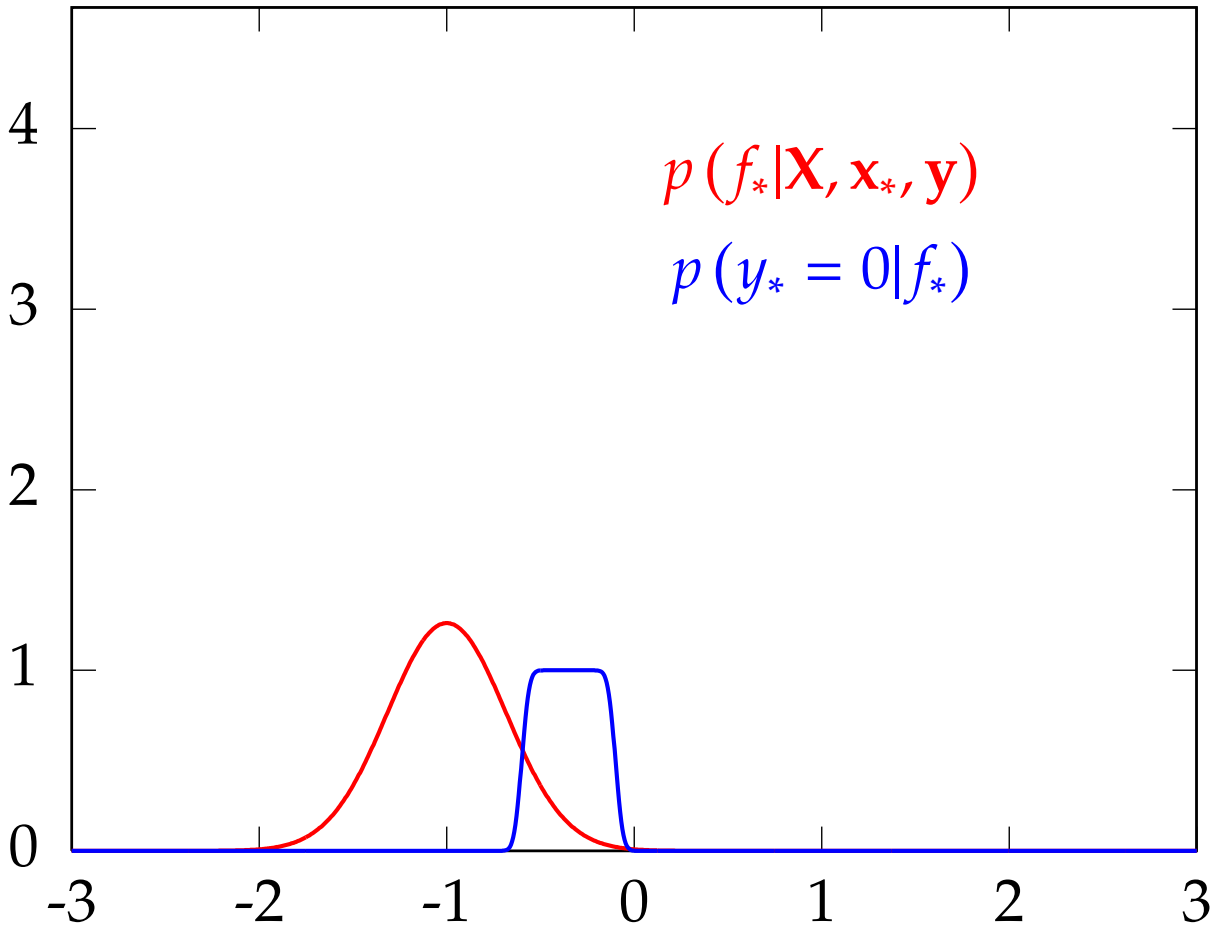


Figure : An EP style update with an ordered category noise model.

Ordinal Regression

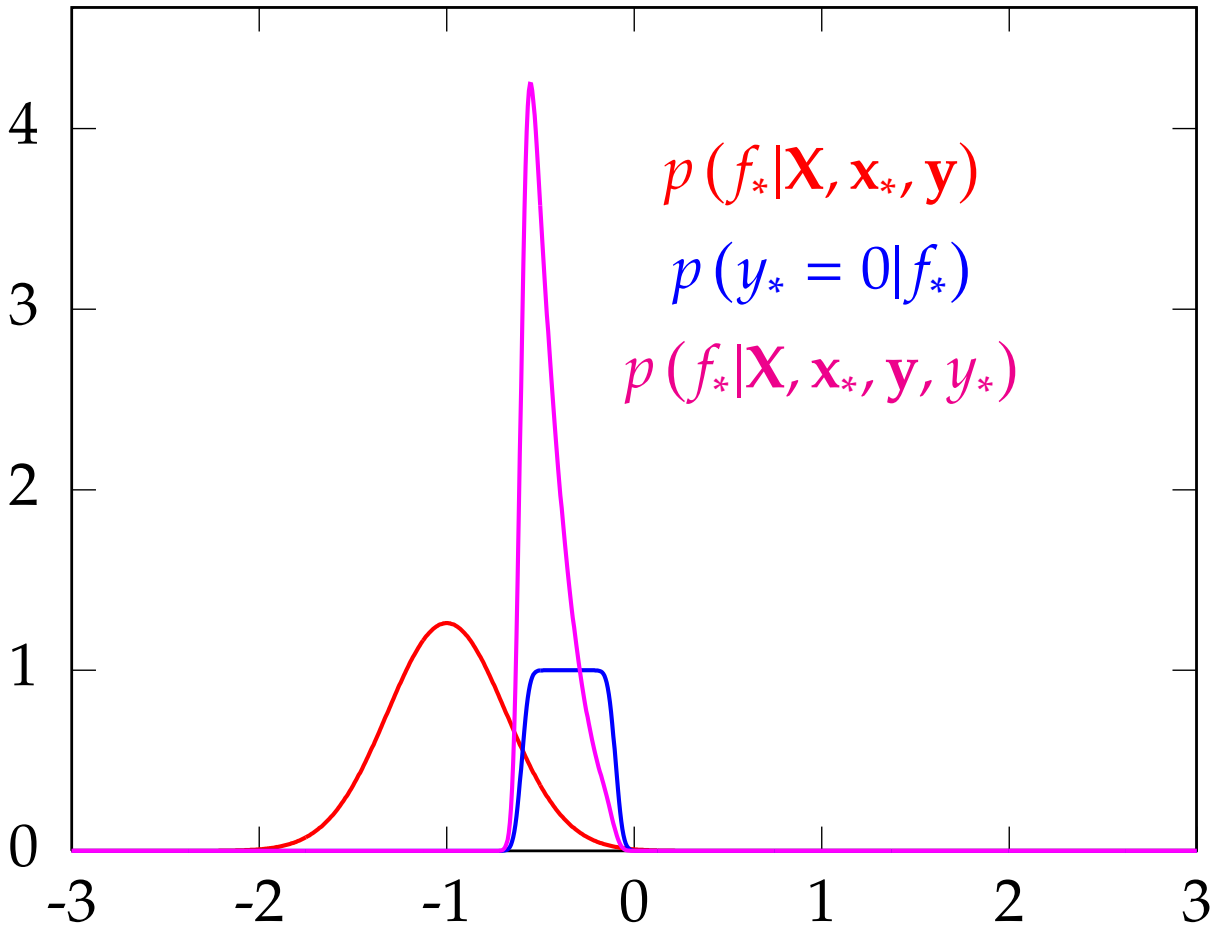


Figure : An EP style update with an ordered category noise model.

Ordinal Regression

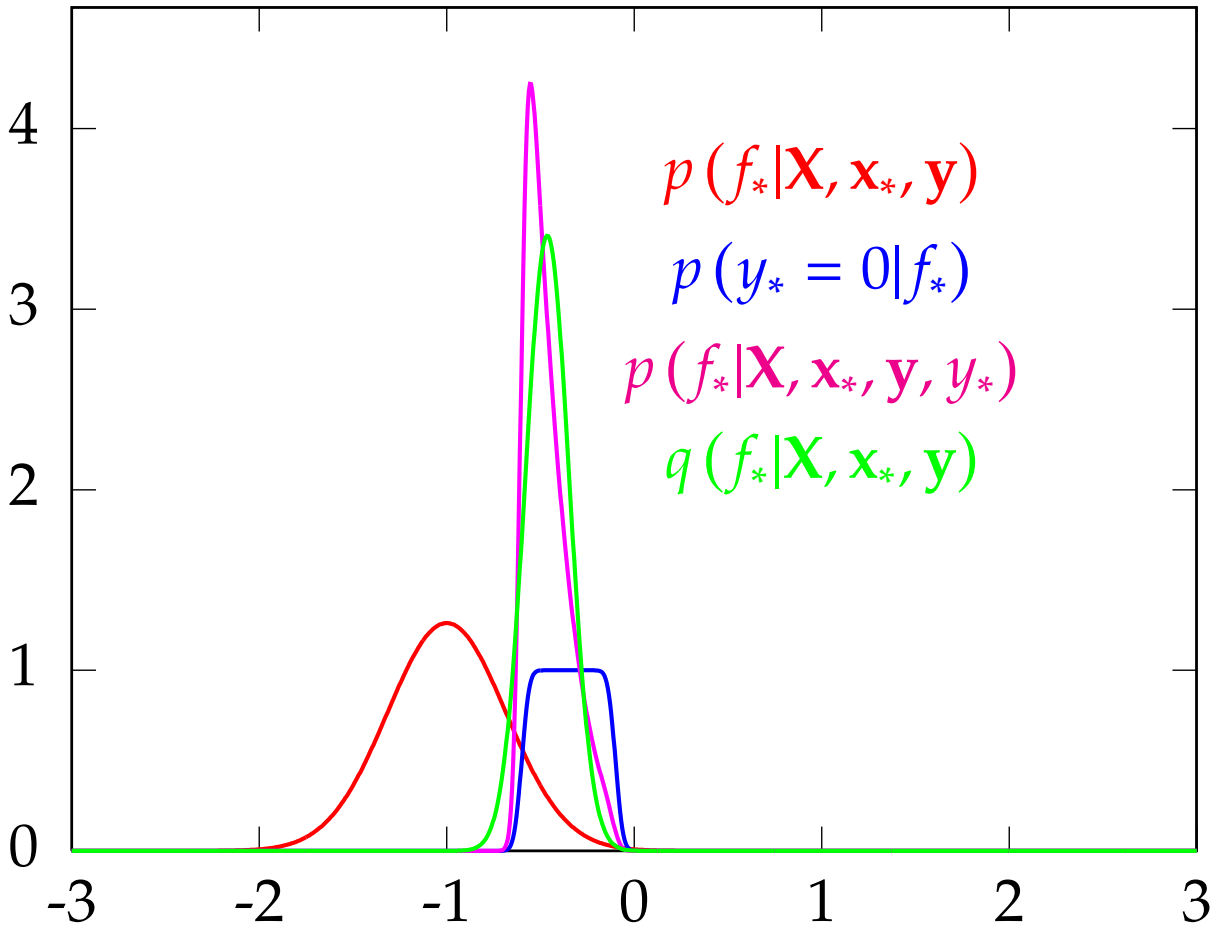


Figure : An EP style update with an ordered category noise model.

Null Category Noise Model

Classification with a Missing Category

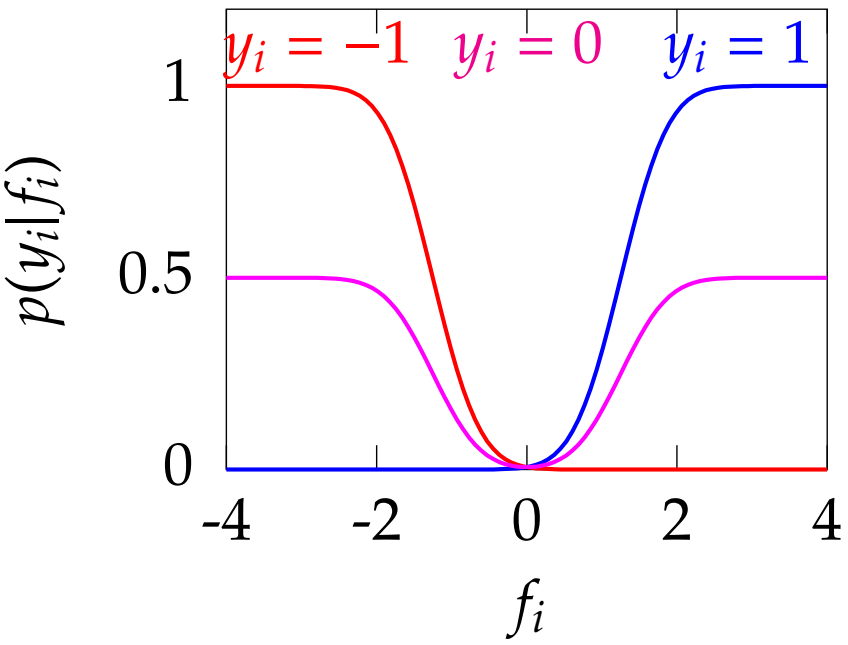


Figure : The null category noise model (semi-supervised learning). The plot shows $p(y_i|f_i)$ for different values of y_i. Here we have assumed three categories.

Semi-supervised Learning

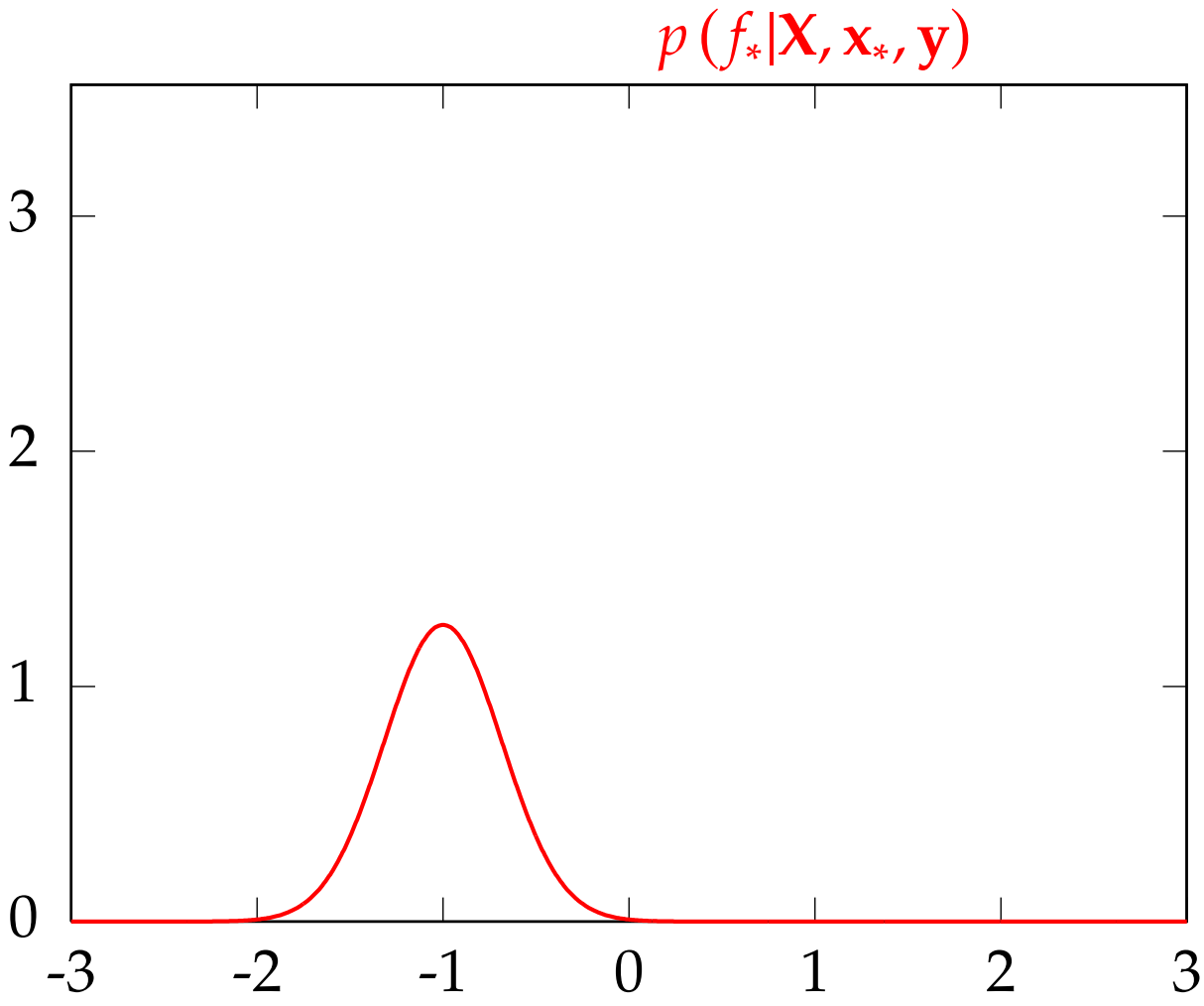


Figure : An EP style update with an null category noise model.

Semi-supervised Learning

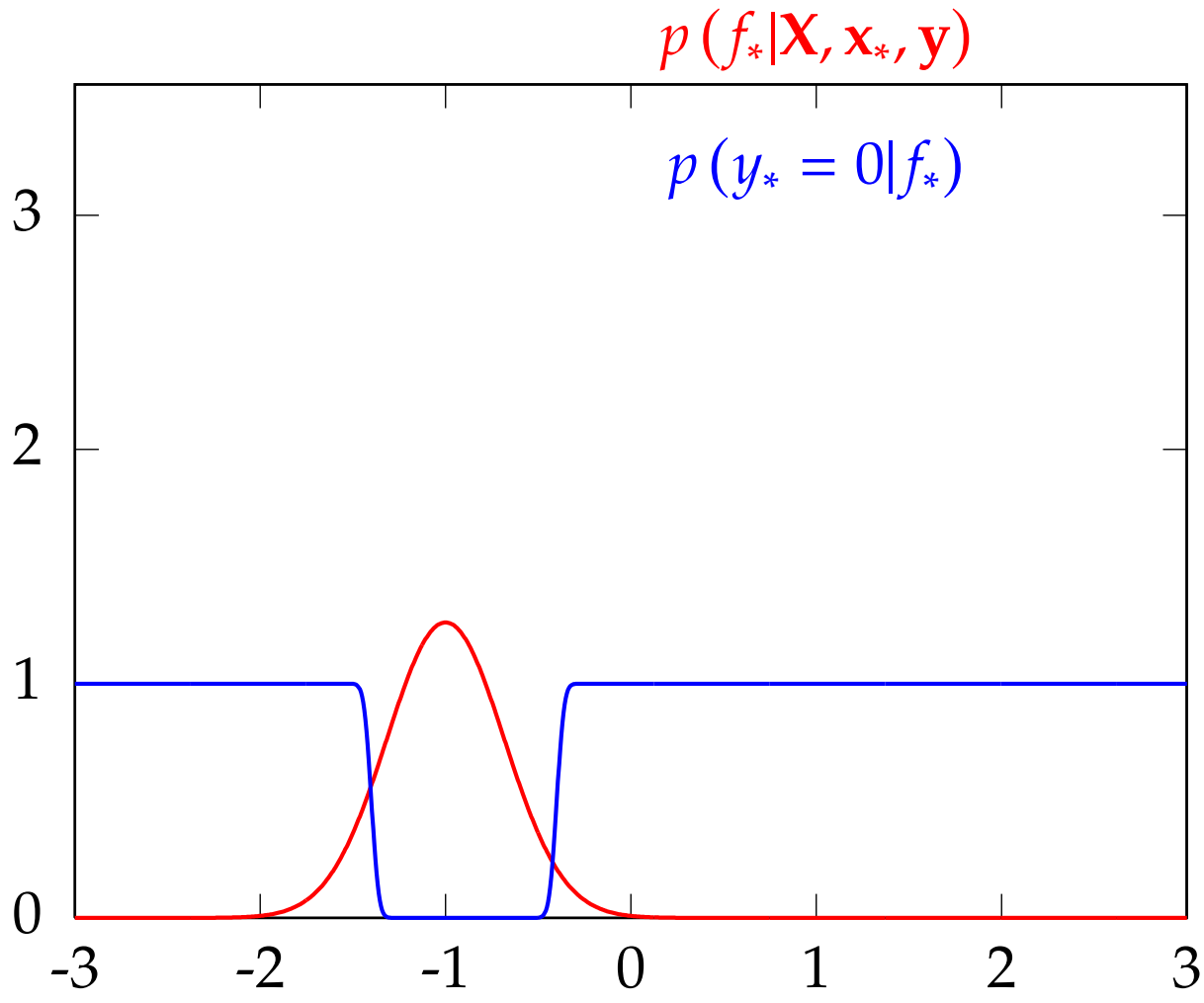


Figure : An EP style update with an null category noise model.

Semi-supervised Learning

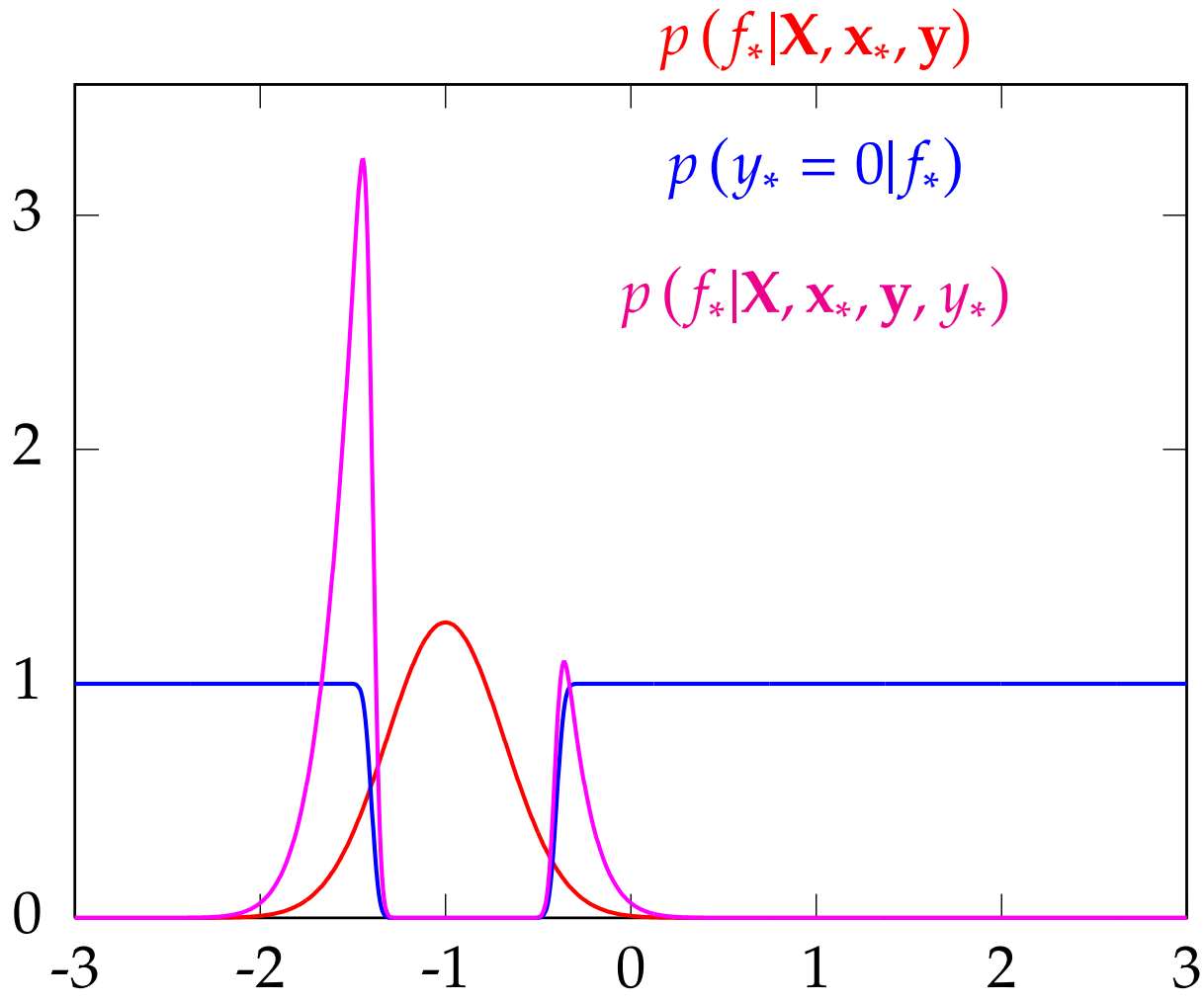


Figure : An EP style update with an null category noise model.

Semi-supervised Learning

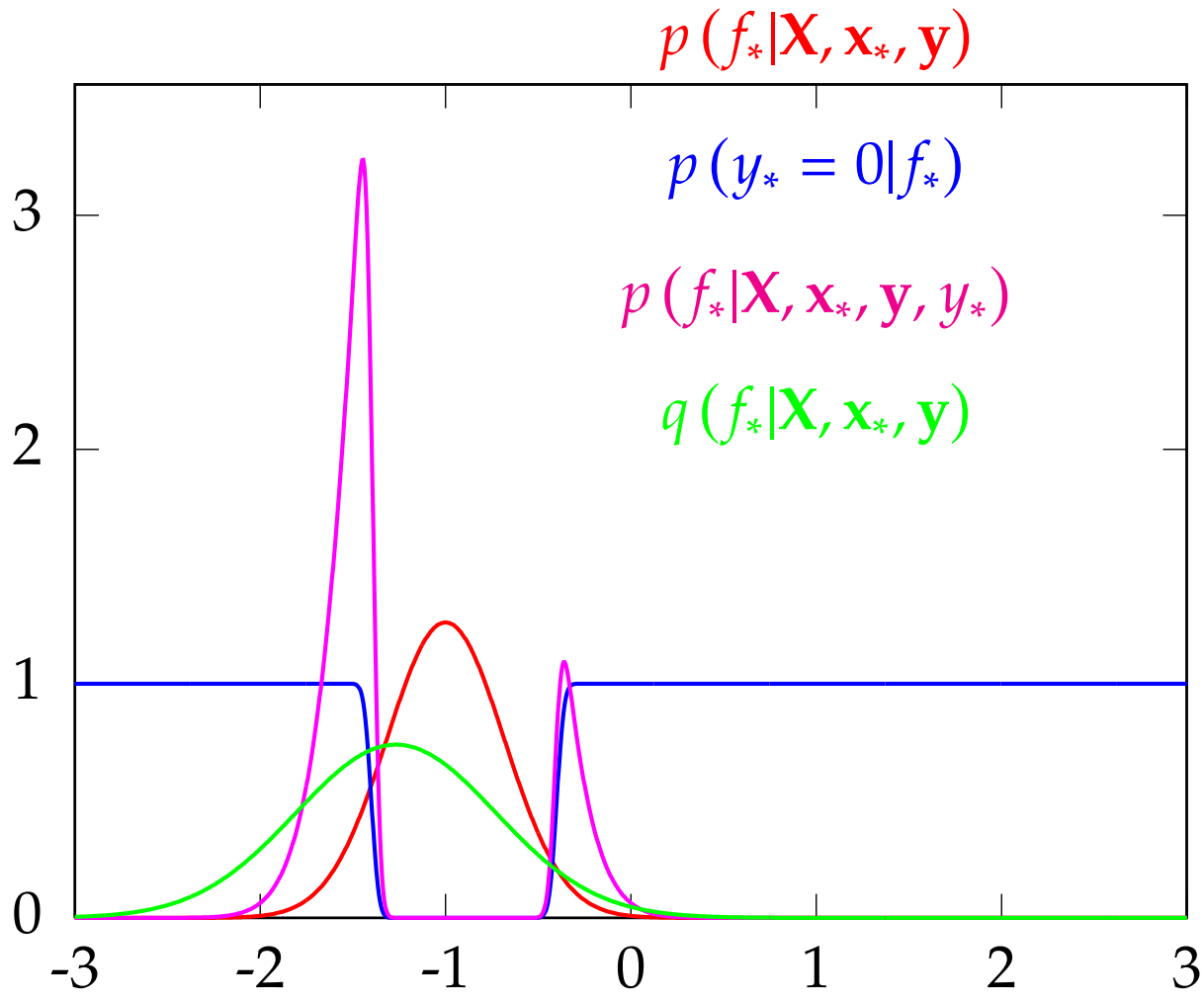


Figure : An EP style update with an null category noise model.

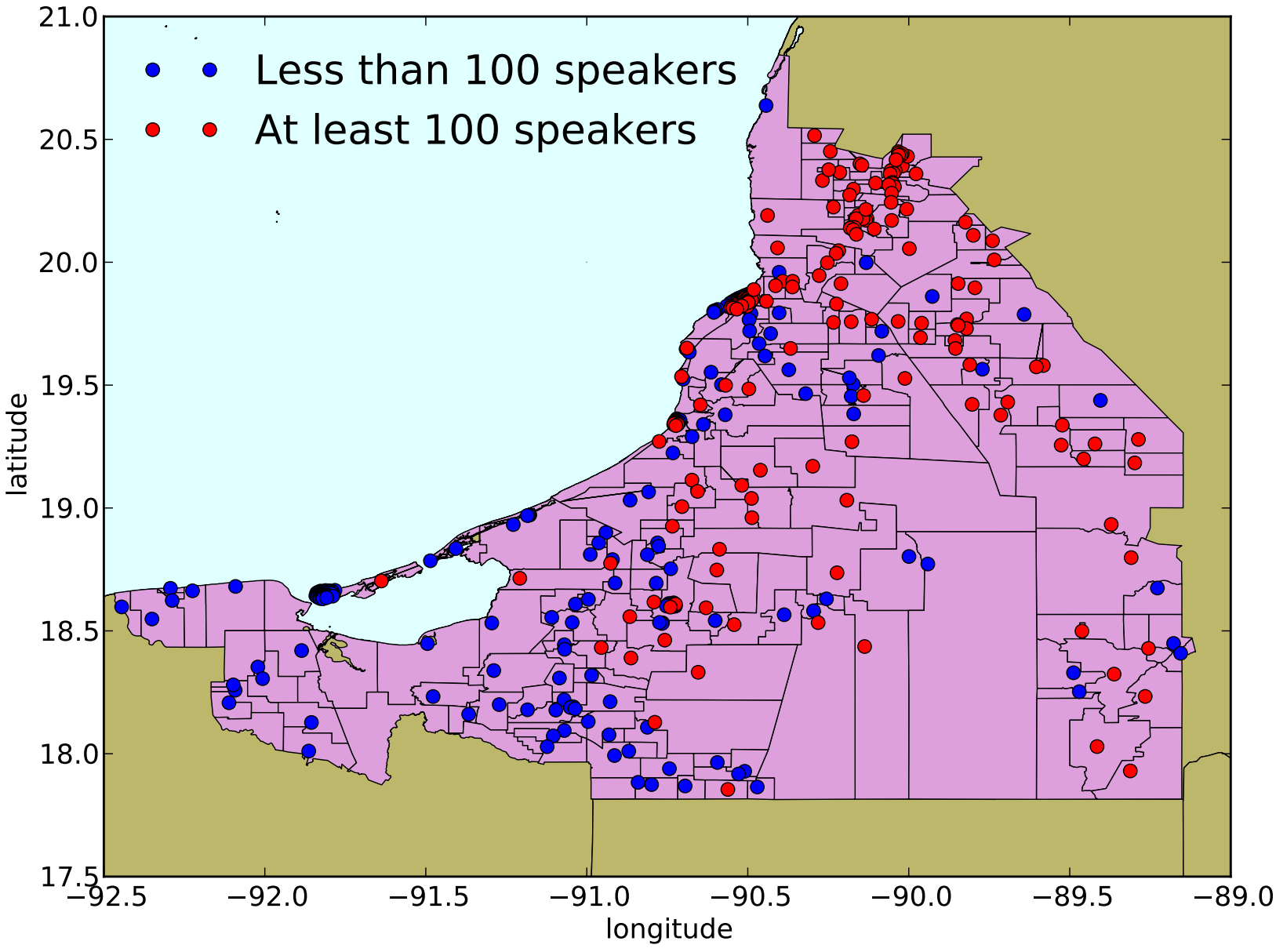
Predictions

- Predictive distribution of $q(f_*|\mathbf{y})$ is also Gaussian:

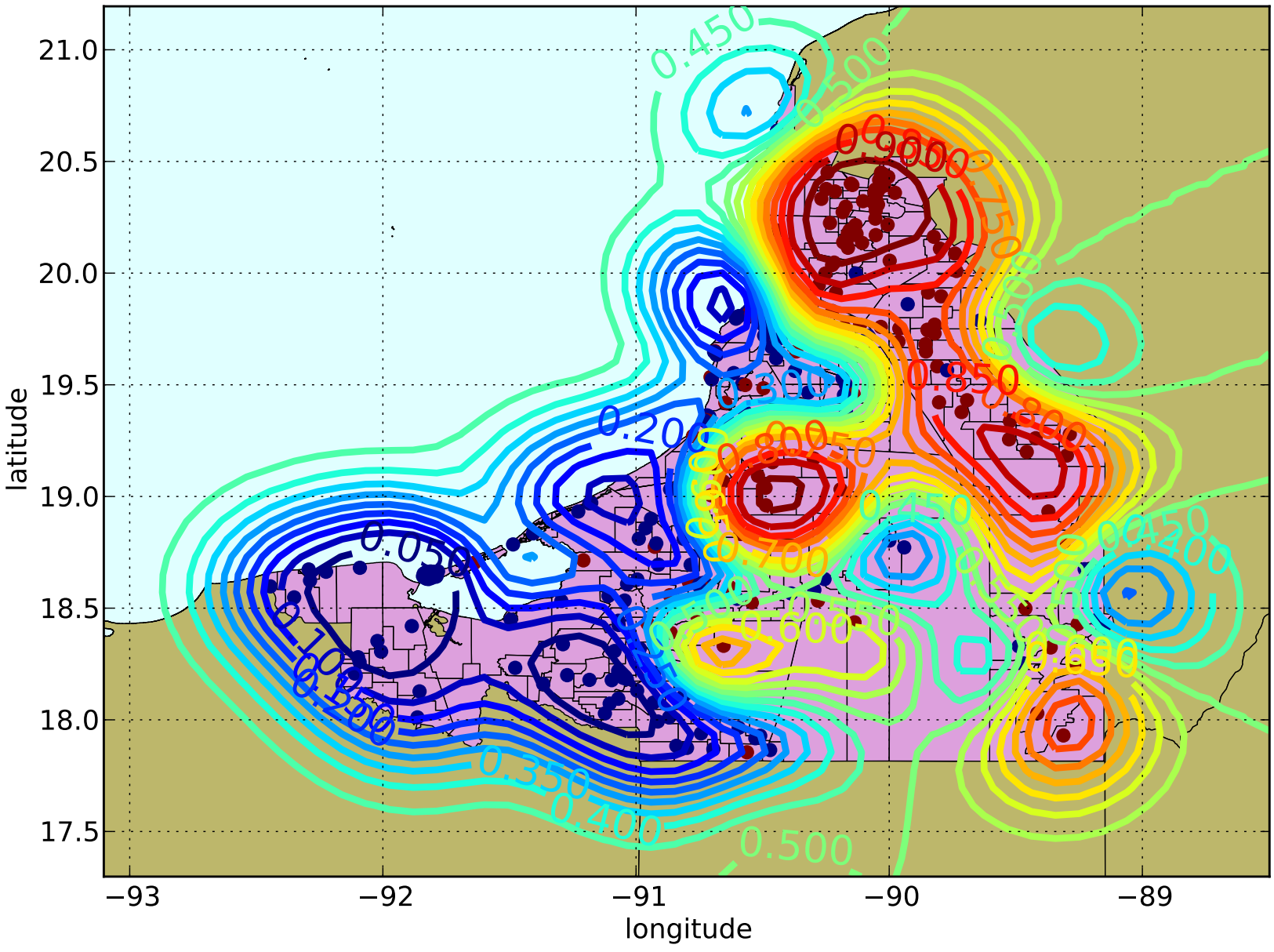
$$\langle f_* \rangle_{q(f_*|\mathbf{y})} = \mathbf{k}_*^\top \left(\mathbf{K_{ff}} + \tilde{\mathbf{\Sigma}} \right)^{-1} \tilde{\boldsymbol{\mu}}$$

$$\text{var}\left(f_*\right) = k_{**} - \mathbf{k}_*^\top \left(\mathbf{K_{ff}} + \tilde{\mathbf{\Sigma}} \right)^{-1} \mathbf{k}_*$$

Example: People who speak an indigenous language



Example: People who speak an indigenous language



Outline

Posterior variance update

- Complexity is dominated by the computation of the posterior covariance:

$$\boldsymbol{\Sigma} = \left(\mathbf{K}_{\mathbf{ff}}^{-1} + \tilde{\boldsymbol{\Sigma}}^{-1}\right)^{-1}$$

Sparse EP

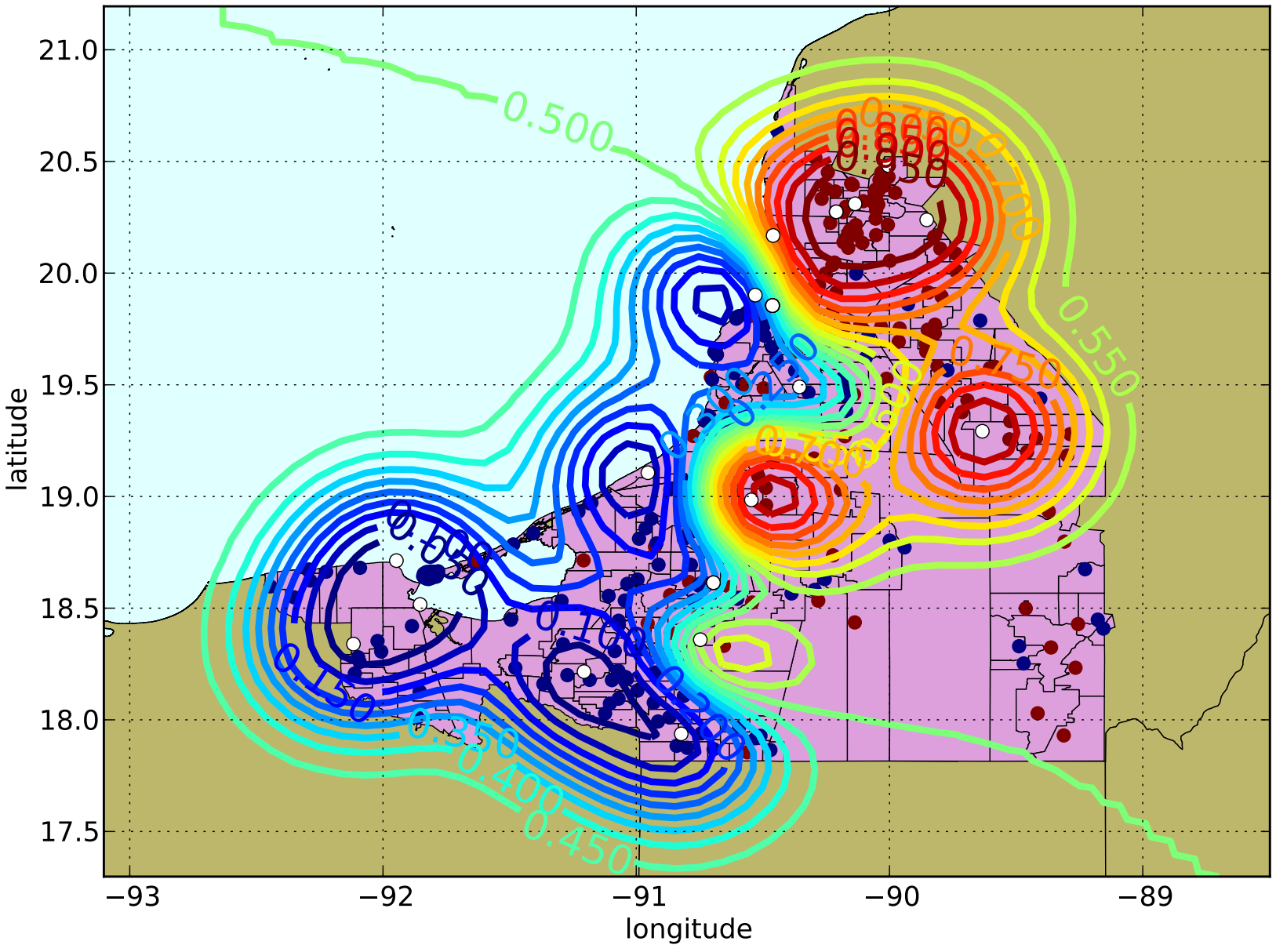
- $q(\mathbf{f}|\mathbf{y})$ is computed as before, but an sparse approximation is used instead of the exact covariance $\mathbf{K_{ff}}$.
- FITC approximation: $O(nm^2)$

$$\mathbf{K_{ff}} \approx \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{K_{uf}} + \mathrm{diag}\left(\mathbf{K_{ff}} - \mathbf{Q_{ff}}\right)$$

- DTC approximation: $O(nm^2)$

$$\mathbf{K_{ff}} \approx \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{K_{uf}}$$

EP-FITC (generalized FITC)

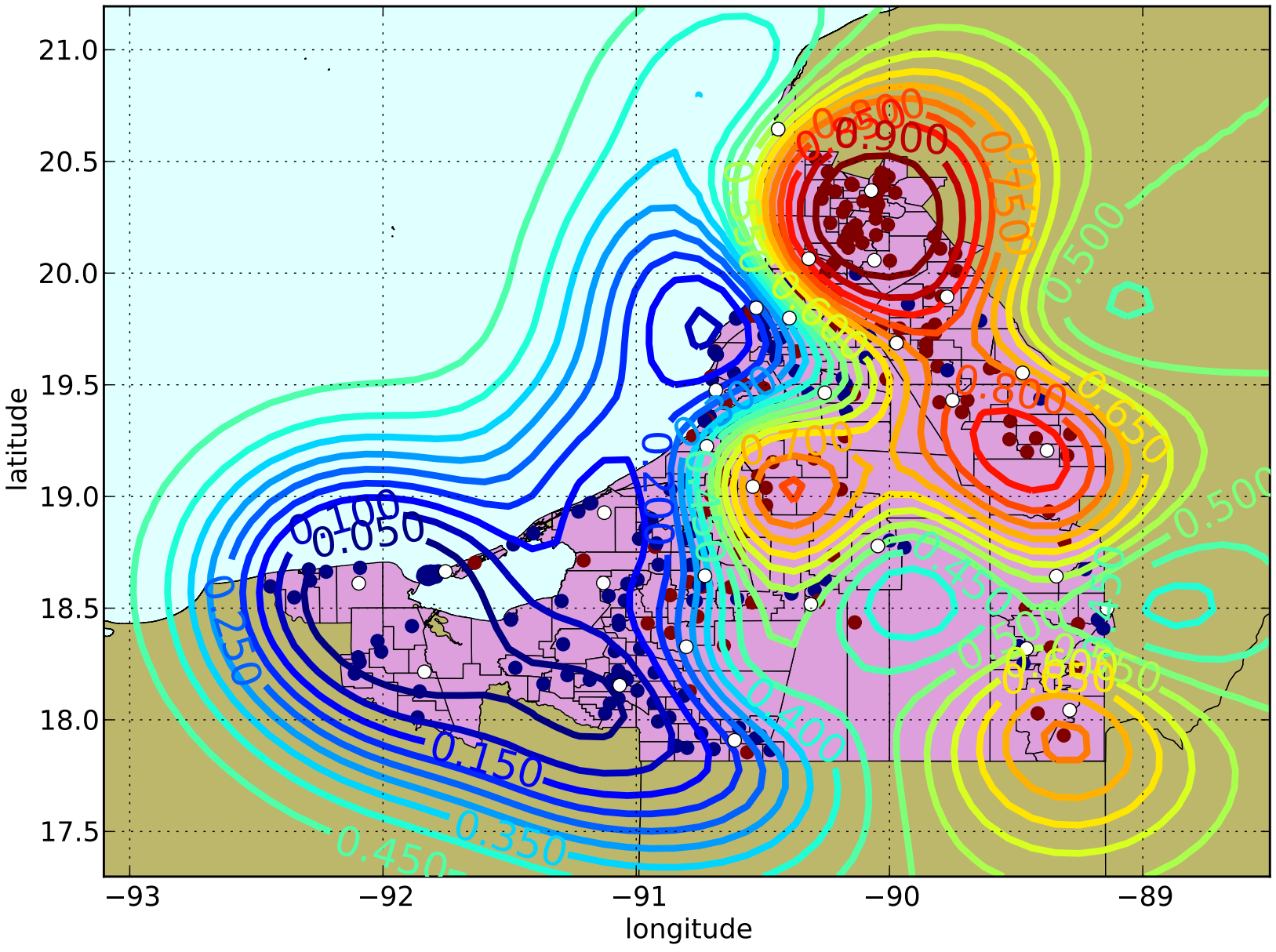


EP-DTC

Compatible with sparse variational approach:

$$\mathcal{L} = \log \mathcal{N}\left(\tilde{\mu}|\mathbf{0}, \mathbf{Q_{ff}} + \tilde{\mathbf{\Sigma}}\right) - \frac{1}{2}\mathrm{tr}\left((\mathbf{K_{ff}} - \mathbf{Q_{ff}})\tilde{\mathbf{\Sigma}}^{-1}\right) - Z_{EP}$$

Sparse variational + EP-DTC



References I