UNIVERSITÄT BONN

Juergen Gall

Classification

MA-INF 2213 - Advanced Computer Vision

SS25

Literature

Chapter 9 Classification Models

S. Prince. **Computer Vision: Models, Learning, and Inference.** Cambridge University Press 2012

Example application: Dogs vs. Cats

Learn relationship

Probabilistic: Given feature **x,** model class $w \in \{c_1, c_2, \ldots, c_k\}$ as probability distribution Pr(w|**x**)

How to model Pr(**w**|**x**)?

- Choose an appropriate form for prior Pr(**w**)
- Parameterize a function of type **w** = f(**x; θ)**

Learning algorithm: learn parameters θ from training data **x**,w

Inference algorithm: just evaluate **Pr(**w**|x)**

Similar to regression, but w has changed!

Logistic Regression

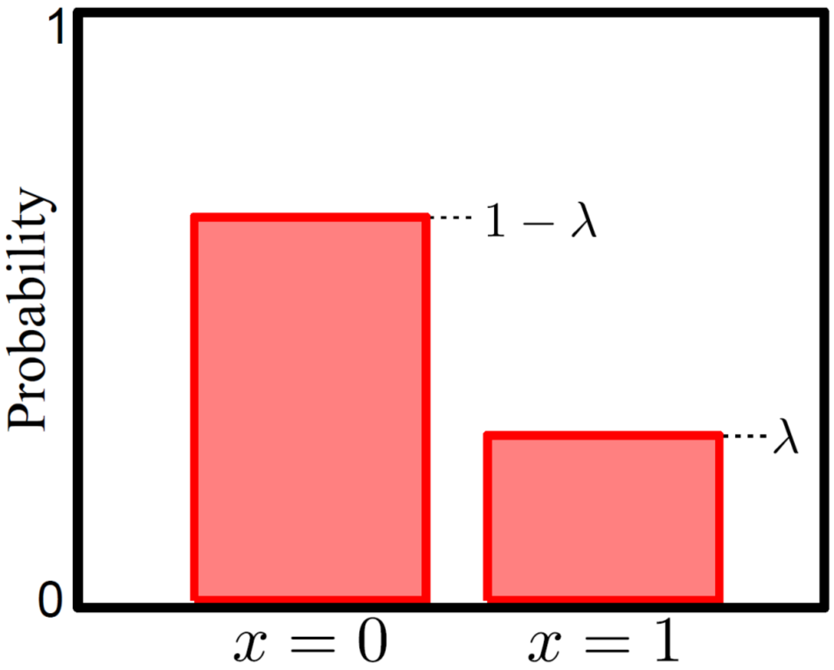
Consider two class problem.

- Choose Bernoulli distribution over world.
- Make parameter λ a function of x

$$Pr(w|\phi_0, \boldsymbol{\phi}, \mathbf{x}) \quad = \quad \text{Bern}_w\left[\text{sig}[a]\right]$$

Simon J.D. Prince

Bernoulli Distribution



$$\begin{aligned} Pr(x=0) &= 1-\lambda \\ Pr(x=1) &= \lambda. \end{aligned}$$

or

$$Pr(x) = \lambda^x (1-\lambda)^{1-x}$$

For short we write:

$$Pr(x) = \text{Bern}_x[\lambda]$$

Bernoulli distribution describes situation where only two possible outcomes x=0/x=1 or failure/success

Takes a single parameter $\lambda \in [0, 1]$

Logistic Regression

Consider two class problem.

- Choose Bernoulli distribution over world.
- Make parameter λ a function of x

$$Pr(w|\phi_0, \boldsymbol{\phi}, \mathbf{x}) \quad = \quad \text{Bern}_w\left[\text{sig}[a]\right]$$

Model activation with a linear function

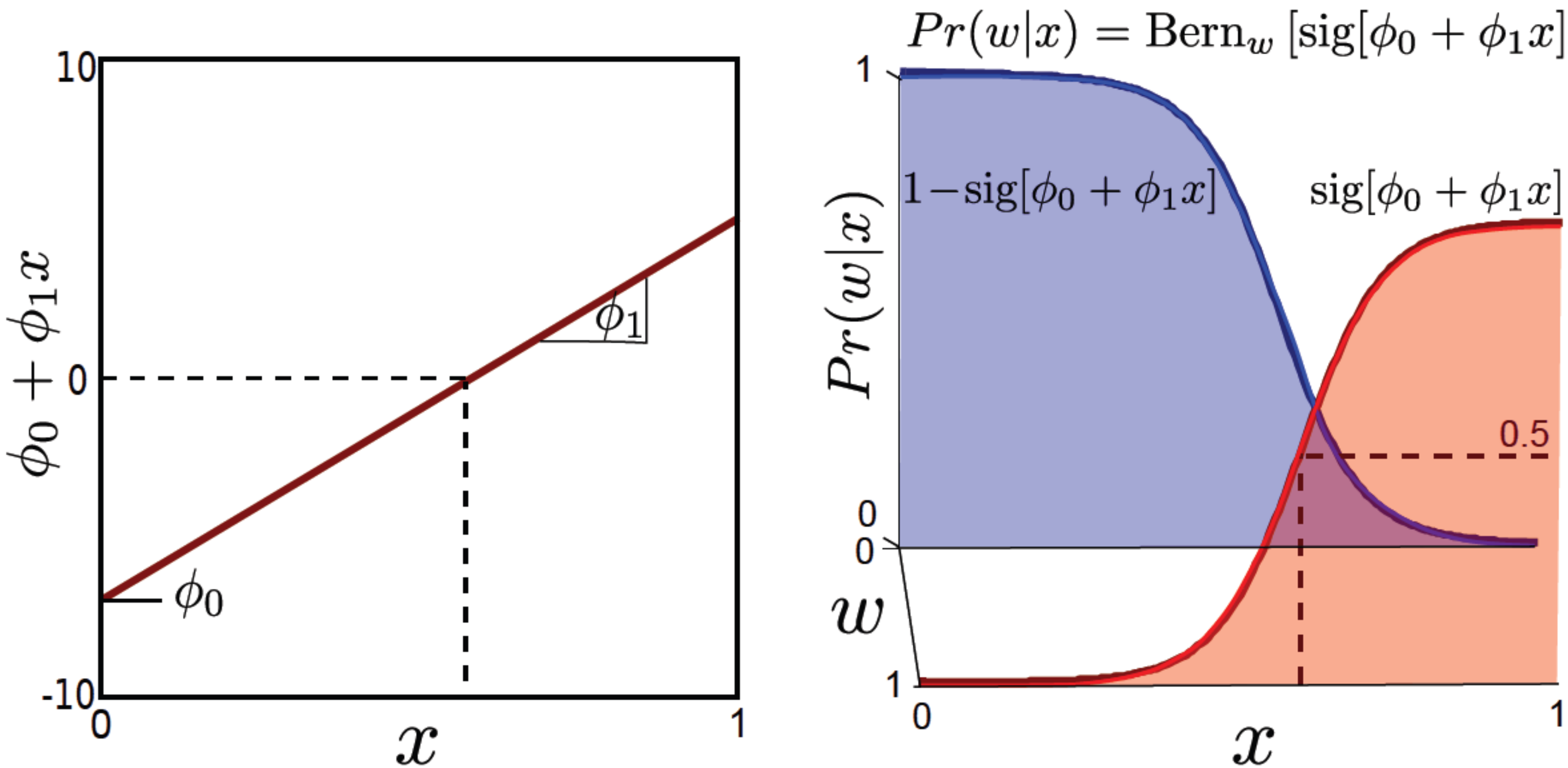
$$a = \phi_0 + \boldsymbol{\phi}^T \mathbf{x}$$

creates number between $[-\infty, \infty]$. Maps to $[0, 1]$ with

$$\text{sig}[a] = \frac{1}{1 + \exp[-a]}$$

Simon J.D. Prince

Logistic Regression



Two parameters $\boldsymbol{\theta} = \{\phi_0, \phi_1\}$

Learning by standard methods (ML,MAP, Bayesian)
Inference: Just evaluate Pr(w|x)

Simon J.D. Prince

Neater Notation

$$Pr(w|\phi_0, \boldsymbol{\phi}, \mathbf{x}) \quad = \quad \text{Bern}_w\left[\text{sig}[a]\right]$$

To make notation easier to handle, we

- Attach a 1 to the start of every data vector

$$\mathbf{x}_i \leftarrow [1 \quad \mathbf{x}_i^T]^T$$

- Attach the offset to the start of the gradient vector $\boldsymbol{\phi}$

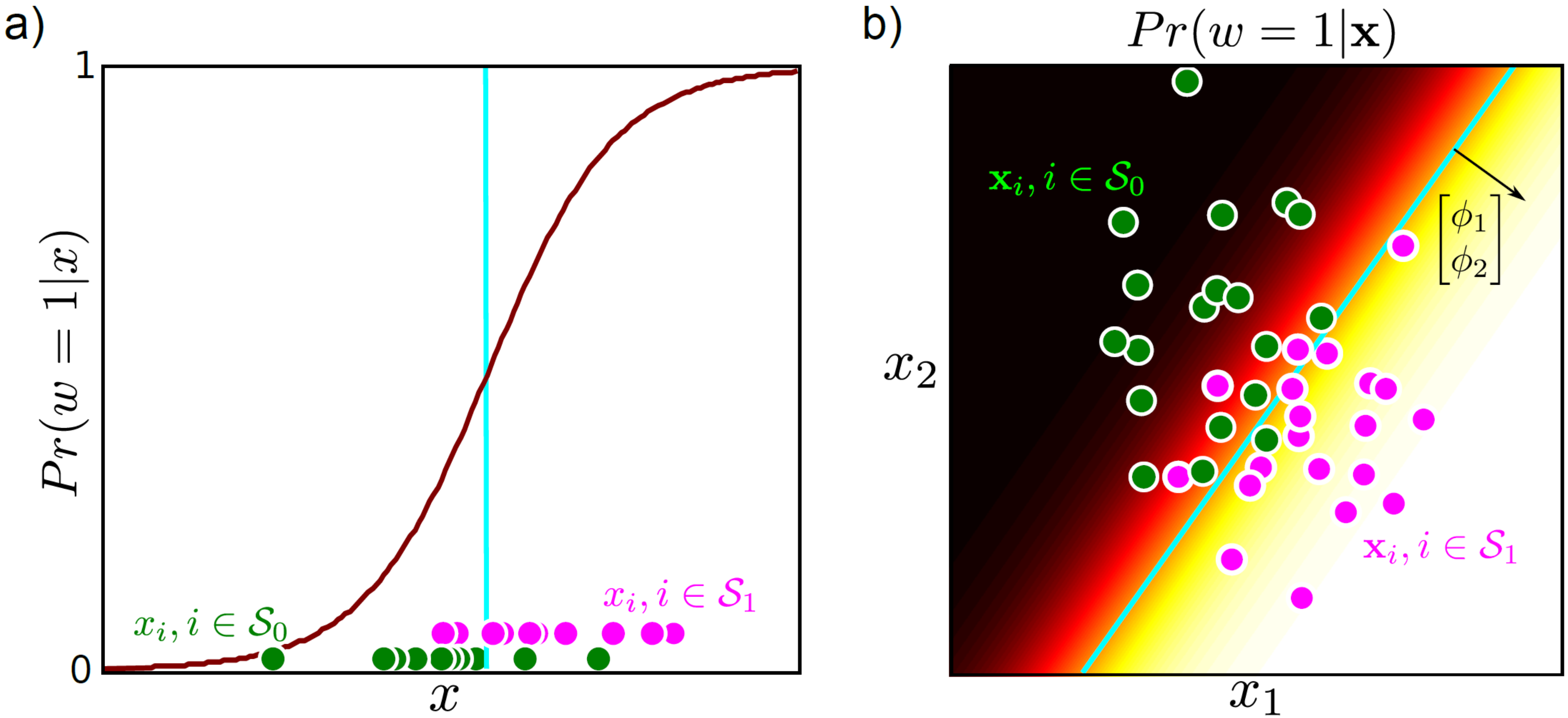
$$\boldsymbol{\phi} \leftarrow [\phi_0 \quad \boldsymbol{\phi}^T]^T$$

New model:

$$Pr(w|\boldsymbol{\phi}, \mathbf{x}) = \text{Bern}_w\left[\frac{1}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}]}\right]$$

Simon J.D. Prince

Logistic regression



$$Pr(w|\boldsymbol{\phi}, \mathbf{x}) = \text{Bern}_w \left[\frac{1}{1 + \exp[-\boldsymbol{\phi}^T \mathbf{x}]} \right]$$

Simon J.D. Prince

Maximum Likelihood

$$
\begin{aligned}
Pr(\mathbf{w}|\mathbf{X}, \phi) &= \prod_{i=1}^{I} \lambda^{w_i}(1-\lambda)^{1-w_i} \\
&= \prod_{i=1}^{I} \left(\frac{1}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}\right)^{w_i} \left(\frac{\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}\right)^{1-w_i}
\end{aligned}
$$

Take logarithm

$$
L = \sum_{i=1}^{I} w_i \log\left[\frac{1}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}\right] + \sum_{i=1}^{I}(1-w_i)\log\left[\frac{\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}\right]
$$

Take derivative:

$$
\frac{\partial L}{\partial \boldsymbol{\phi}} = -\sum_{i=1}^{I}\left(\frac{1}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]} - w_i\right)\mathbf{x}_i = -\sum_{i=1}^{I}\left(\mathrm{sig}[a_i] - w_i\right)\mathbf{x}_i
$$

Simon J.D. Prince

Derivatives

$$\frac{\partial L}{\partial \boldsymbol{\phi}} = -\sum_{i=1}^{I}\left(\frac{1}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]} - w_i\right)\mathbf{x}_i = -\sum_{i=1}^{I}\left(\text{sig}[a_i] - w_i\right)\mathbf{x}_i$$

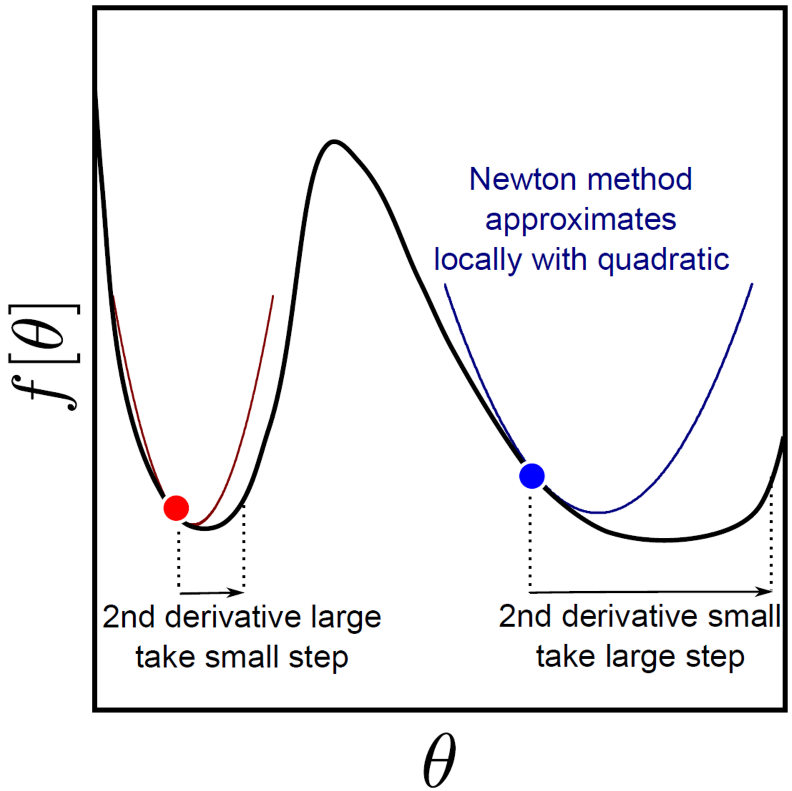
Unfortunately, there is no closed form solution– we cannot get an expression for ϕ in terms of x and w

Use iterative non-linear optimization

Simon J.D. Prince

Newton's Method

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[f[\boldsymbol{\theta}] \right] \qquad \boldsymbol{\theta}^{[t+1]} = \boldsymbol{\theta}^{[t]} - \lambda \left(\frac{\partial^2 f}{\partial \boldsymbol{\theta}^2} \right)^{-1} \frac{\partial f}{\partial \boldsymbol{\theta}}$$



Matrix of second derivatives is called the Hessian.

If positive definite, then convex

Simon J.D. Prince

Optimization for Logistic Regression

$$\phi^{[t]} = \phi^{[t-1]} + \alpha \left(\frac{\partial^2 L}{\partial \phi^2} \right)^{-1} \frac{\partial L}{\partial \phi}$$

Derivatives of log likelihood:

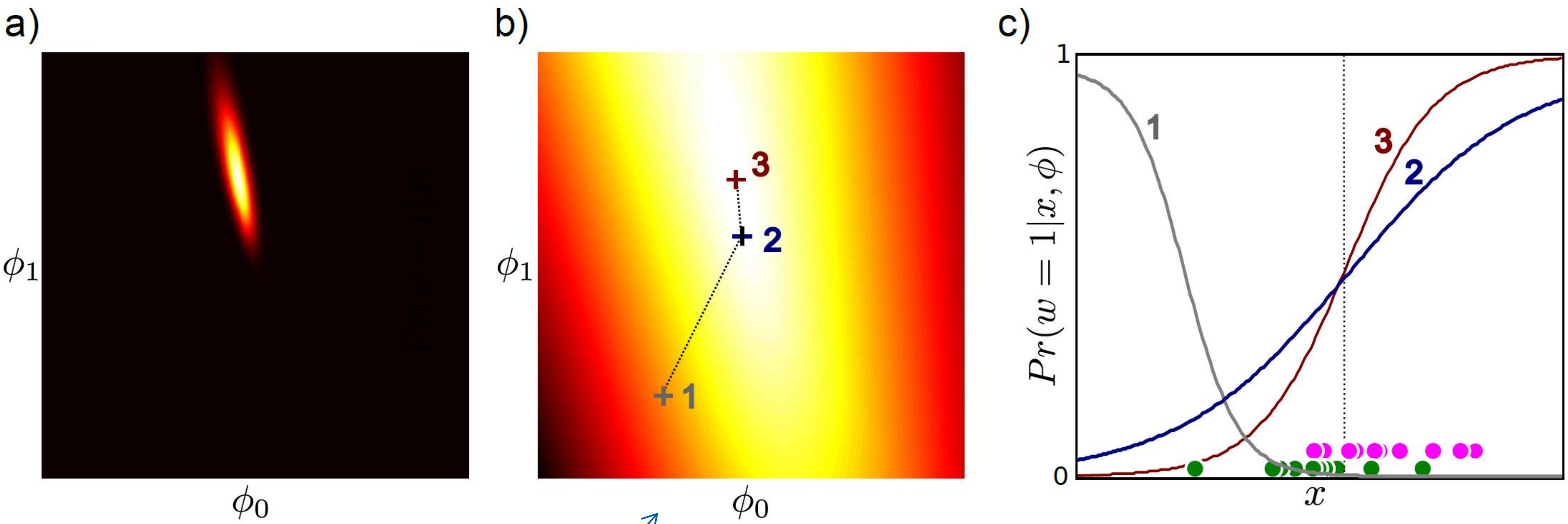
$$\frac{\partial L}{\partial \phi} = -\sum_{i=1}^{I} (\mathrm{sig}[a_i] - w_i)\mathbf{x}_i$$

$$\frac{\partial^2 L}{\partial \phi^2} = -\sum_{i=1}^{I} \mathrm{sig}[a_i](1 - \mathrm{sig}[a_i])\mathbf{x}_i\mathbf{x}_i^T$$

Positive definite!

Simon J.D. Prince

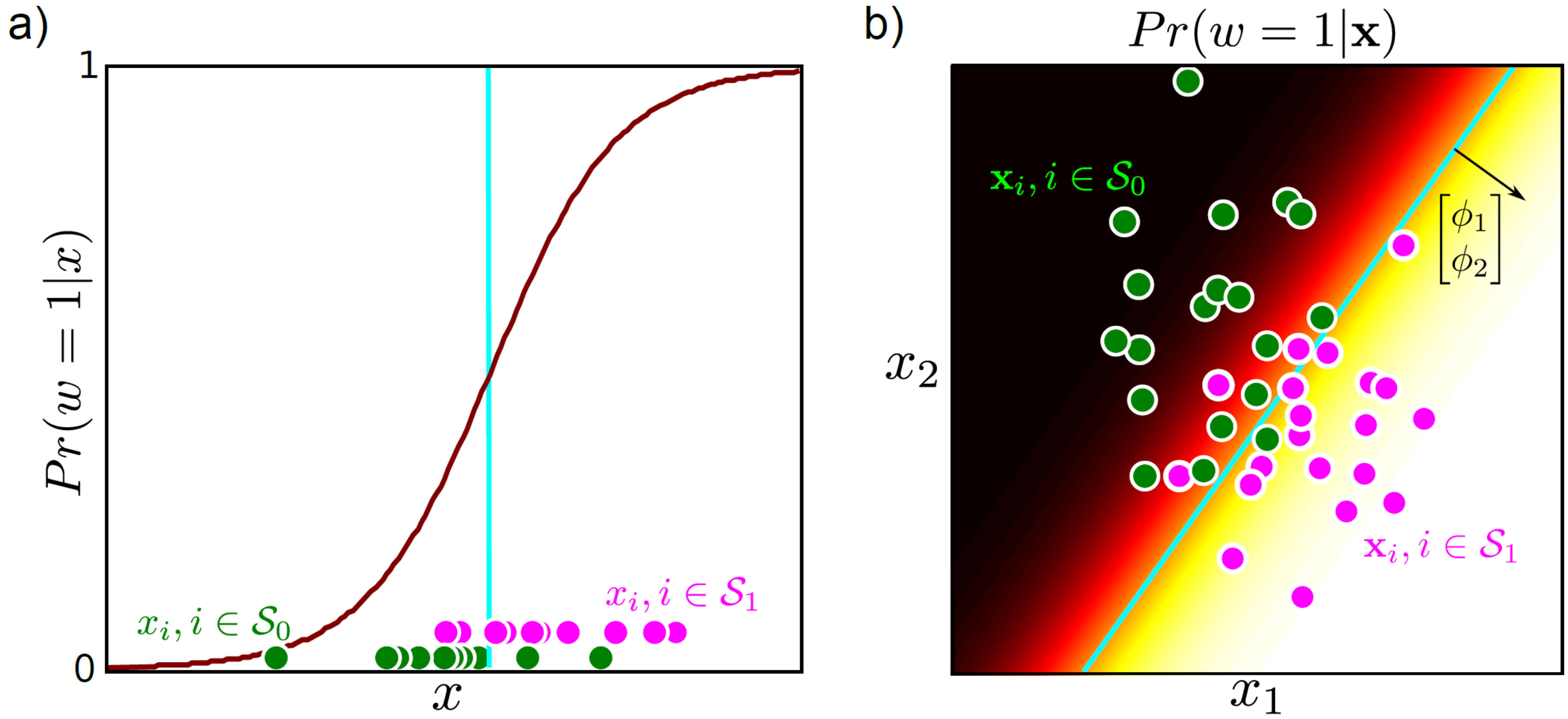
$$Pr(\mathbf{w}|\mathbf{X},\boldsymbol{\phi}) = \prod_{i=1}^{I}\left(\frac{1}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}\right)^{w_i}\left(\frac{\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}\right)^{1-w_i}$$



$$L = \sum_{i=1}^{I} w_i \log\left[\frac{1}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}\right] + \sum_{i=1}^{I}(1-w_i)\log\left[\frac{\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}\right]$$

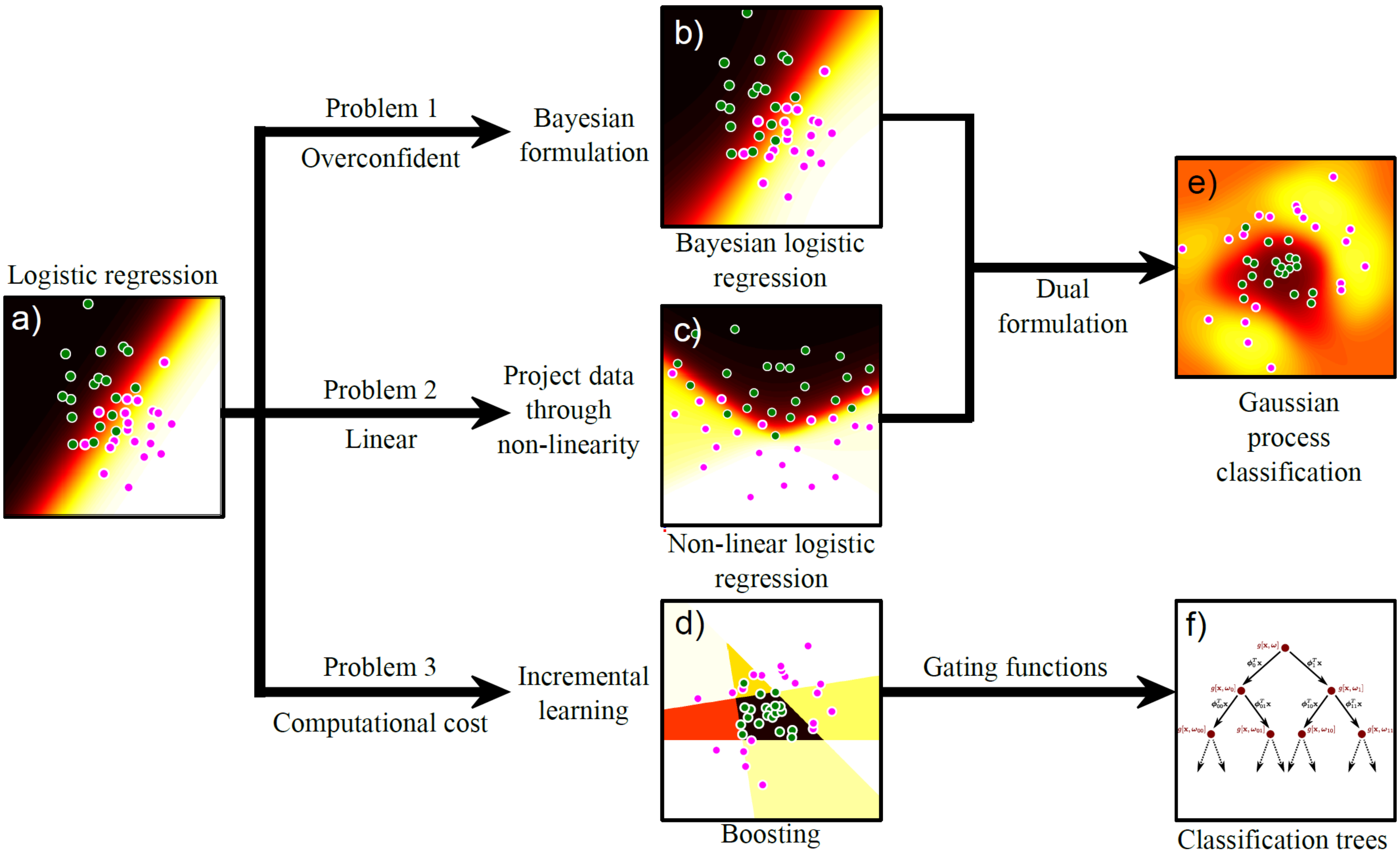
Simon J.D. Prince

Maximum likelihood fits



$$Pr(w|\boldsymbol{\phi}, \mathbf{x}) = \text{Bern}_w \left[\frac{1}{1+\exp[-\boldsymbol{\phi}^T \mathbf{x}]} \right]$$

Simon J.D. Prince



Simon J.D. Prince

Bayesian Logistic Regression

Likelihood:

$$Pr(\mathbf{w}|\mathbf{X},\phi) = \prod_{i=1}^{I}\left(\frac{1}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}\right)^{w_i}\left(\frac{\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}\right)^{1-w_i}$$

Prior:

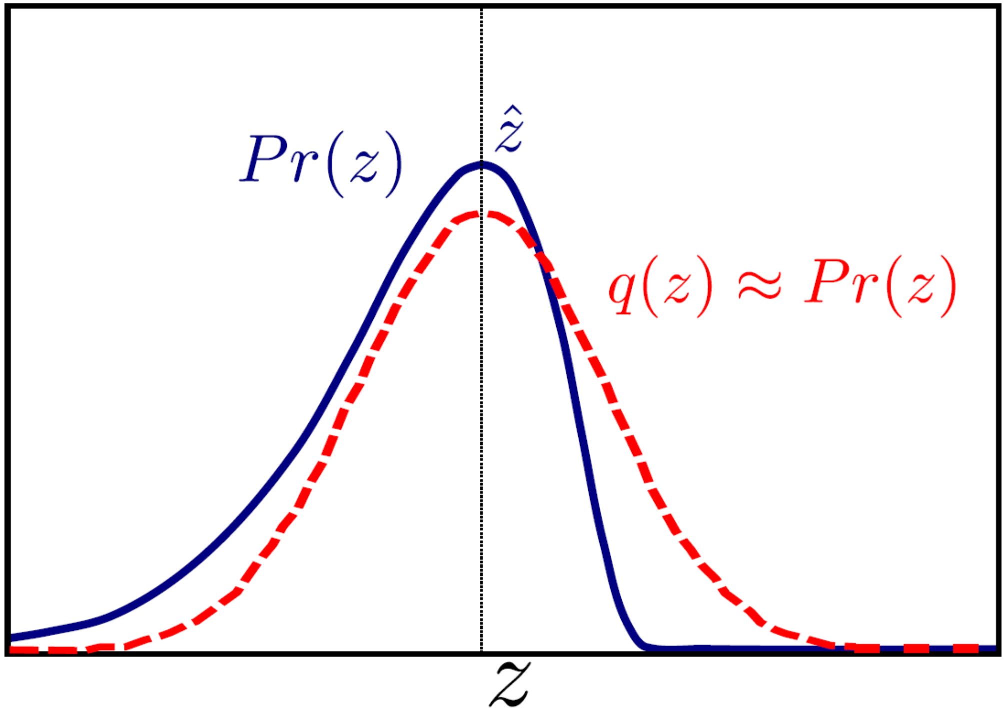
$$Pr(\boldsymbol{\phi}) = \text{Norm}_{\boldsymbol{\phi}}[\mathbf{0},\sigma_p^2\mathbf{I}]$$

Apply Bayes' rule:

$$Pr(\phi|\mathbf{X},\mathbf{w}) = \frac{Pr(\mathbf{w}|\mathbf{X},\phi)Pr(\phi)}{Pr(\mathbf{w}|\mathbf{X})}$$

Simon J.D. Prince

Laplace Approximation



Approximate posterior distribution with normal

- Set mean to MAP estimate
- Set covariance to match that at MAP estimate (actually: get 2nd derivatives to agree)

Simon J.D. Prince

Laplace Approximation

Find MAP solution by optimizing

$$L = \sum_{i=1}^{I} \log[Pr(w_i|\mathbf{x}_i, \boldsymbol{\phi})] + \log[Pr(\boldsymbol{\phi})]$$

using Newton's method:

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\phi}} &= -\sum_{i=1}^{I} (\text{sig}[a_i] - w_i)\mathbf{x}_i - \frac{\boldsymbol{\phi}}{\sigma_p^2} \\
\frac{\partial^2 L}{\partial \boldsymbol{\phi}^2} &= -\sum_{i=1}^{I} \text{sig}[a_i](1 - \text{sig}[a_i])\mathbf{x}_i\mathbf{x}_i^T - \frac{1}{\sigma_p^2}
\end{aligned}$$

Simon J.D. Prince

Laplace Approximation

Find MAP solution by optimizing

$$L = \sum_{i=1}^{I} \log[Pr(w_i|\mathbf{x}_i, \phi)] + \log[Pr(\phi)]$$

Approximate with normal

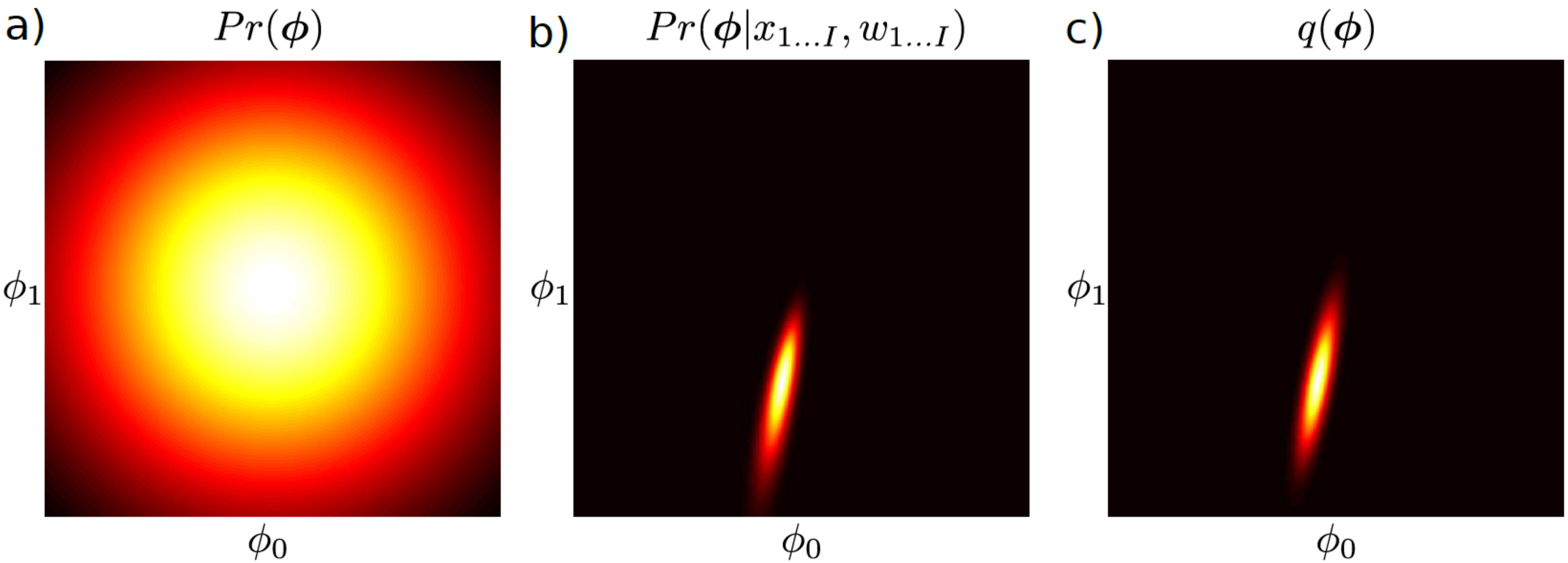
$$Pr(\phi|\mathbf{X}, \mathbf{w}) \approx q(\phi) = \text{Norm}_{\phi}[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$$

where

$$\begin{aligned} \boldsymbol{\mu} &= \hat{\phi} \\ \boldsymbol{\Sigma} &= -\left(\frac{\partial^2 L}{\partial \phi^2}\right)^{-1}\Bigg|_{\phi=\hat{\phi}} \end{aligned}$$

Simon J.D. Prince

Laplace Approximation



Prior

Actual posterior

Approximated

Simon J.D. Prince

Inference

$$\begin{aligned}
Pr(w^*|\mathbf{x}^*, \mathbf{X}, \mathbf{w}) &= \int Pr(w^*|\mathbf{x}^*, \boldsymbol{\phi}) Pr(\boldsymbol{\phi}|\mathbf{X}, \mathbf{w}) d\boldsymbol{\phi} \\
&\approx \int Pr(w^*|\mathbf{x}^*, \boldsymbol{\phi}) q(\boldsymbol{\phi}) d\boldsymbol{\phi}.
\end{aligned}$$

Can re-express in terms of activation $a = \boldsymbol{\phi}^T \mathbf{x}^*$

$$Pr(w^*|\mathbf{x}^*, \mathbf{X}, \mathbf{w}) \approx \int Pr(w^*|a) Pr(a) da$$

Using transformation properties of normal distributions

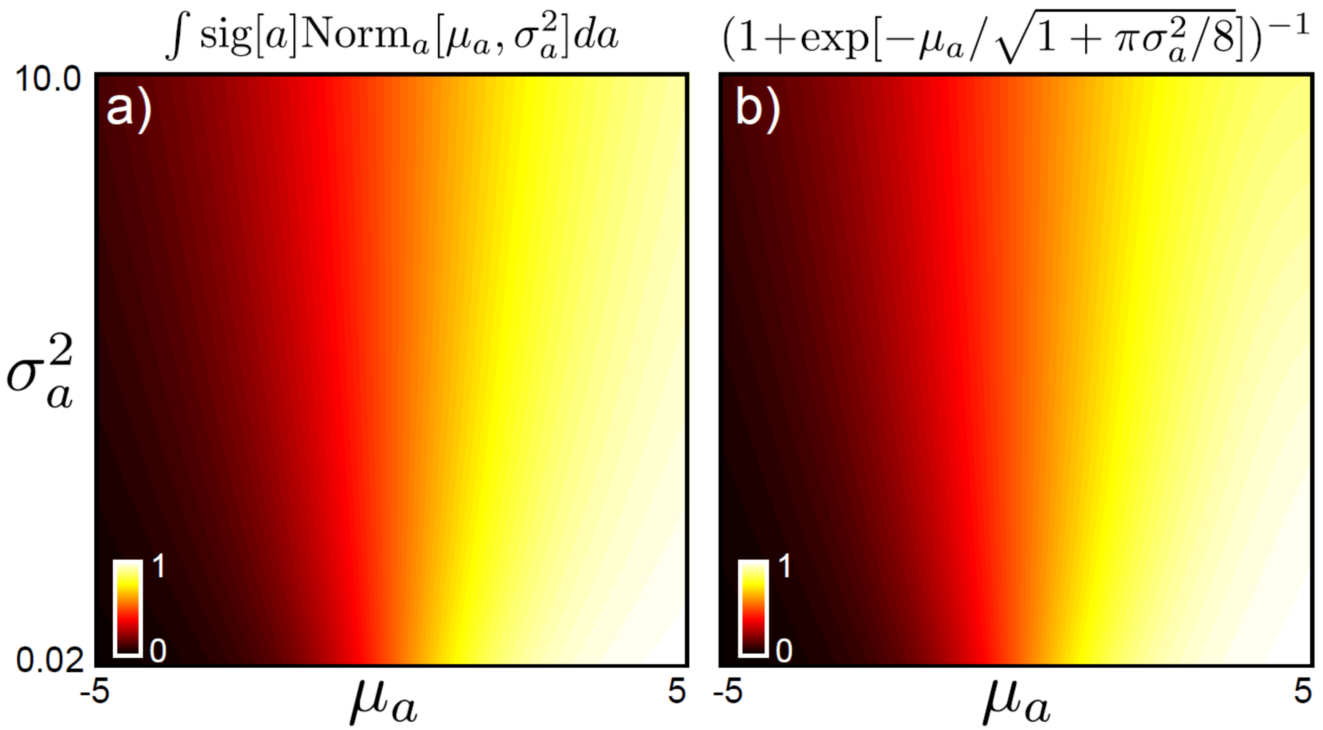
$$\begin{aligned}
Pr(a) = Pr(\boldsymbol{\phi}^T \mathbf{x}^*) &= \text{Norm}_a[\boldsymbol{\mu}^T \mathbf{x}^*, \mathbf{x}^{*T} \boldsymbol{\Sigma} \mathbf{x}^*] \\
&= \text{Norm}_a[\mu_a, \sigma_a^2],
\end{aligned}$$

Simon J.D. Prince

Approximation of Integral

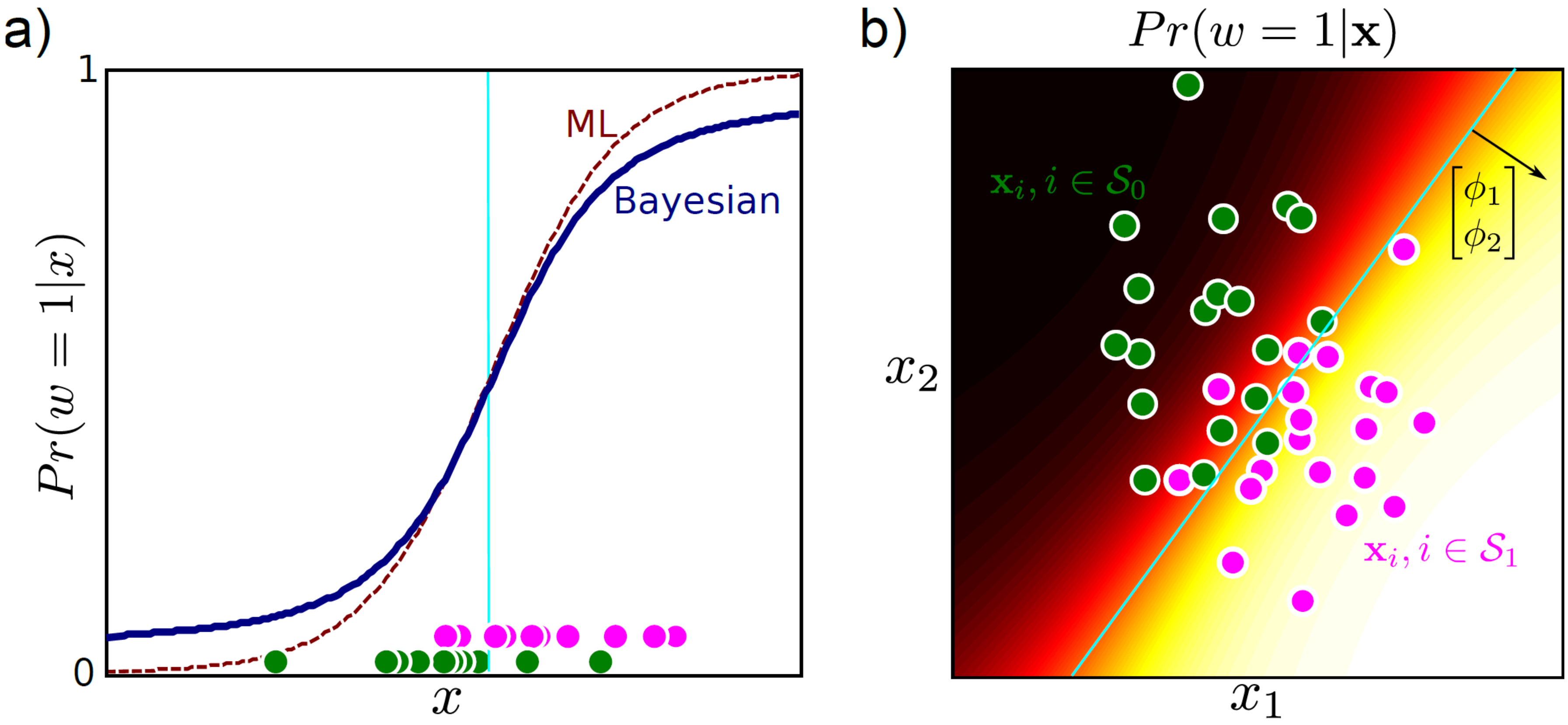
(Or perform numerical integration on *a* – which is 1D)

$$\int Pr(w^*|a)\text{Norm}_a[\mu_a, \sigma_a^2]da \approx \frac{1}{1+\exp[-\mu_a/\sqrt{1+\pi\sigma_a^2/8}]}$$



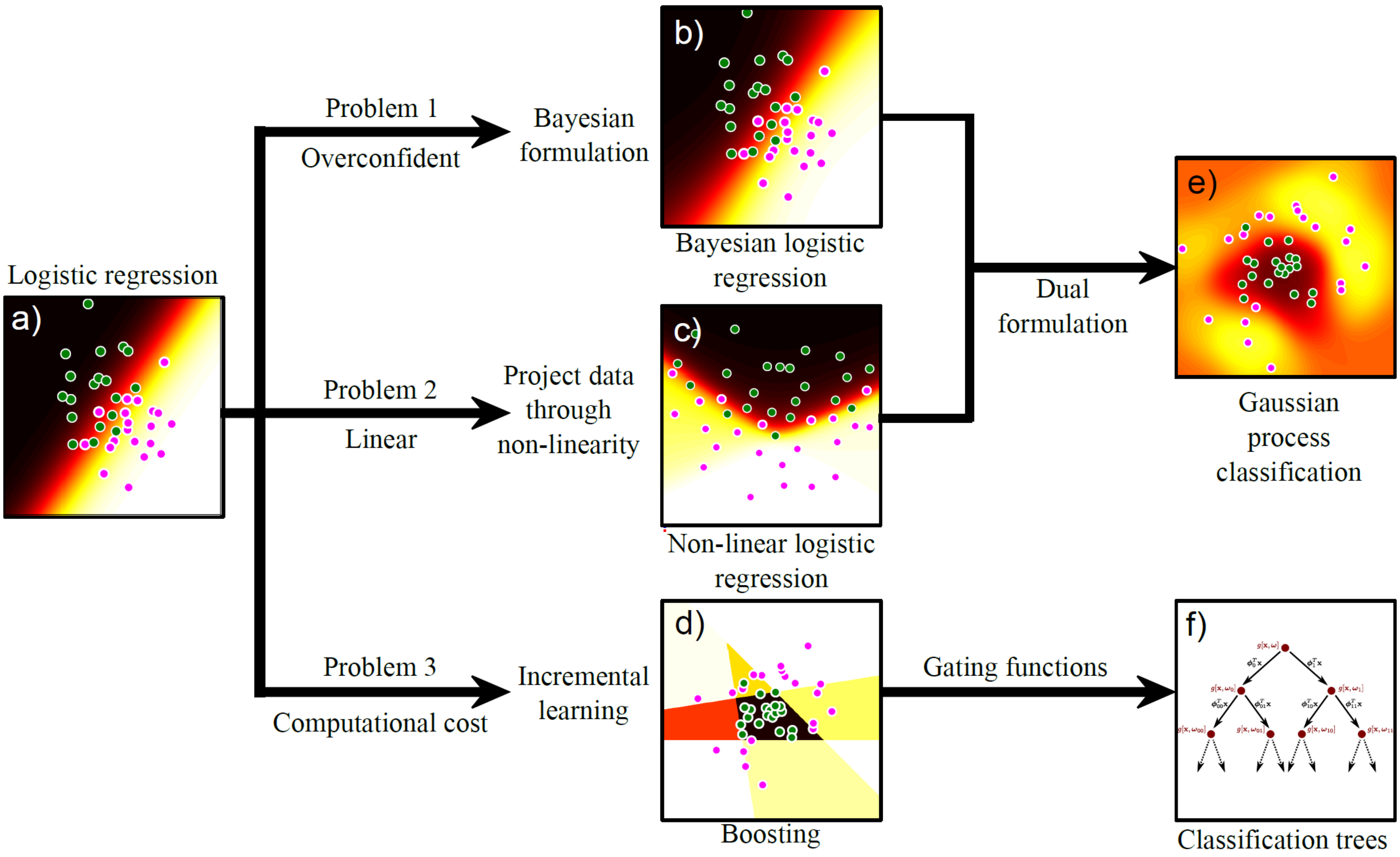
Simon J.D. Prince

Bayesian Solution



Simon J.D. Prince

UNIVERSITÄT BONN



Simon J.D. Prince

Non-linear logistic regression

Same idea as for regression.

- Apply non-linear transformation

$$\mathbf{z} = \mathbf{f}[\mathbf{x}]$$

- Build model as usual

$$\begin{aligned} Pr(w=1|\mathbf{x},\phi) &= \text{Bern}_w\left[\text{sig}[\phi^T\mathbf{z}]\right] \\ &= \text{Bern}_w\left[\text{sig}[\phi^T\mathbf{f}[\mathbf{x}]]\right] \end{aligned}$$

Simon J.D. Prince

Non-linear logistic regression

UNIVERSITÄT BONN

Example transformations:

- Arc tan functions of projections: $z_k = \arctan[\boldsymbol{\alpha}_k^T \mathbf{x}]$
- Radial basis functions: $z_k = \exp\left[-\frac{1}{\lambda_0}(\mathbf{x} - \boldsymbol{\alpha}_k)^T(\mathbf{x} - \boldsymbol{\alpha}_k)\right]$

Simon J.D. Prince

Non-linear logistic regression

Example transformations:

- Arc tan functions of projections: $z_k = \arctan[\boldsymbol{\alpha}_k^T \mathbf{x}]$
- Radial basis functions: $z_k = \exp\left[-\frac{1}{\lambda_0}(\mathbf{x}-\boldsymbol{\alpha}_k)^T(\mathbf{x}-\boldsymbol{\alpha}_k)\right]$

Fit using optimization (also transformation parameters **α**):

$$\boldsymbol{\theta} = [\boldsymbol{\phi}^T, \boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \ldots, \boldsymbol{\alpha}_K^T]^T \qquad a_i = \boldsymbol{\phi}^T \mathbf{f}[\mathbf{x}_i]$$

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{I} (w_i - \text{sig}[a_i]) \frac{\partial a_i}{\partial \boldsymbol{\theta}}$$

$$\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2} = \sum_{i=1}^{I} \text{sig}[a_i](\text{sig}[a_i]-1) \frac{\partial a_i}{\partial \boldsymbol{\theta}} \frac{\partial a_i}{\partial \boldsymbol{\theta}}^T + (w_i - \text{sig}[a_i]) \frac{\partial^2 a_i}{\partial \boldsymbol{\theta}^2}$$

Simon J.D. Prince

Linear logistic regression (recall)

$$\phi^{[t]} = \phi^{[t-1]} + \alpha \left(\frac{\partial^2 L}{\partial \phi^2} \right)^{-1} \frac{\partial L}{\partial \phi}$$

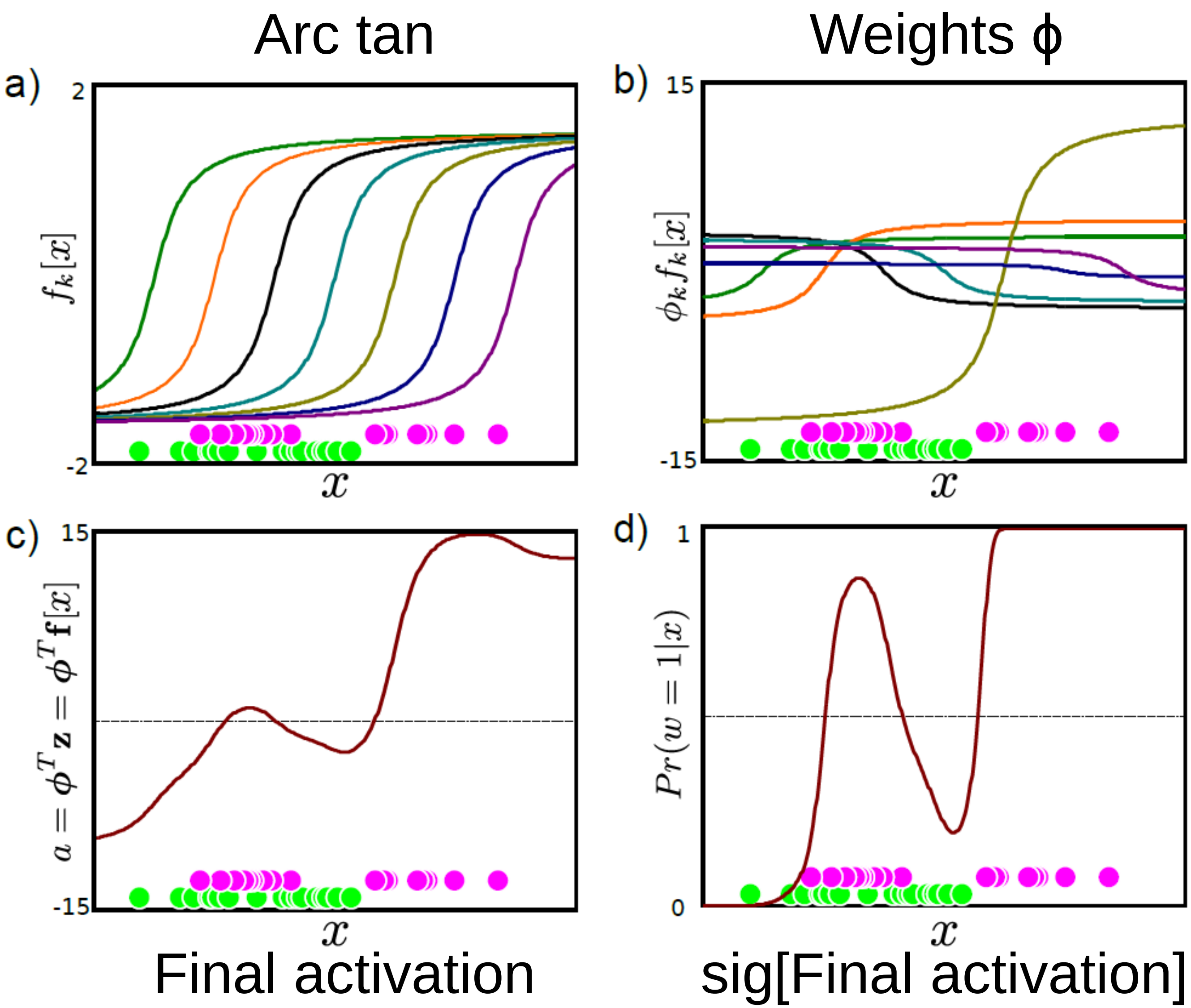
Derivatives of log likelihood:

$$\frac{\partial L}{\partial \phi} = -\sum_{i=1}^{I} (\text{sig}[a_i] - w_i)\mathbf{x}_i$$

$$\frac{\partial^2 L}{\partial \phi^2} = -\sum_{i=1}^{I} \text{sig}[a_i](1 - \text{sig}[a_i])\mathbf{x}_i \mathbf{x}_i^T$$

Simon J.D. Prince

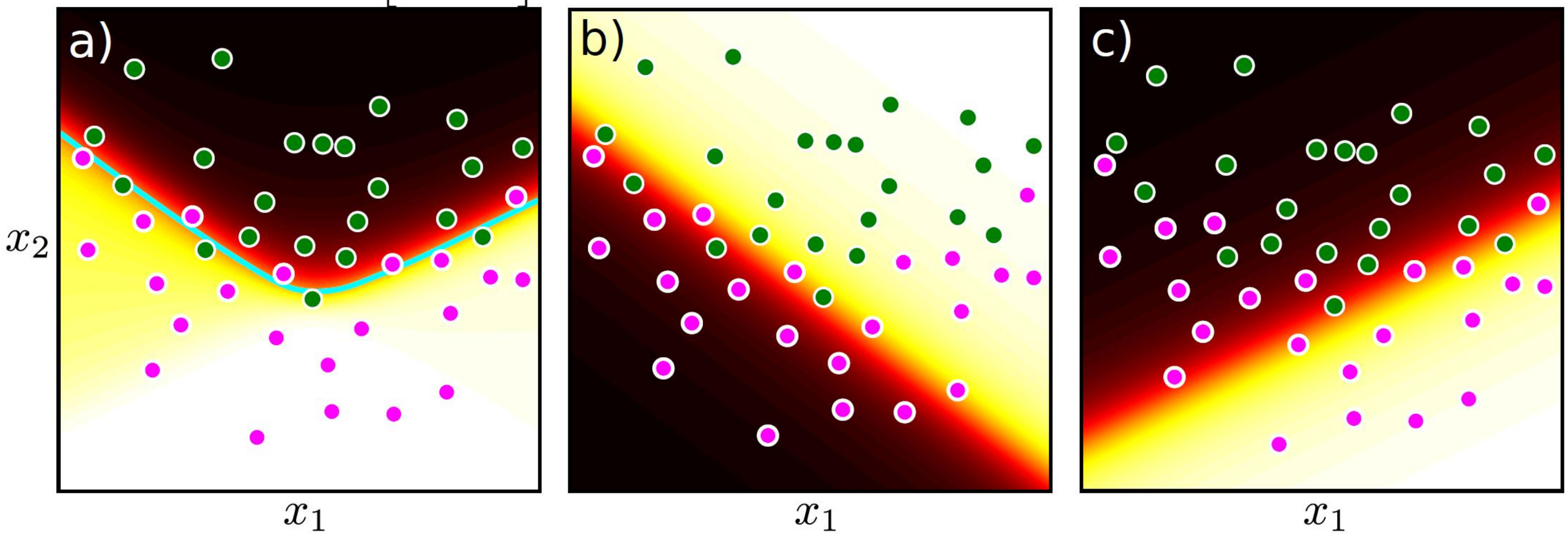
Non-linear logistic regression in 1D



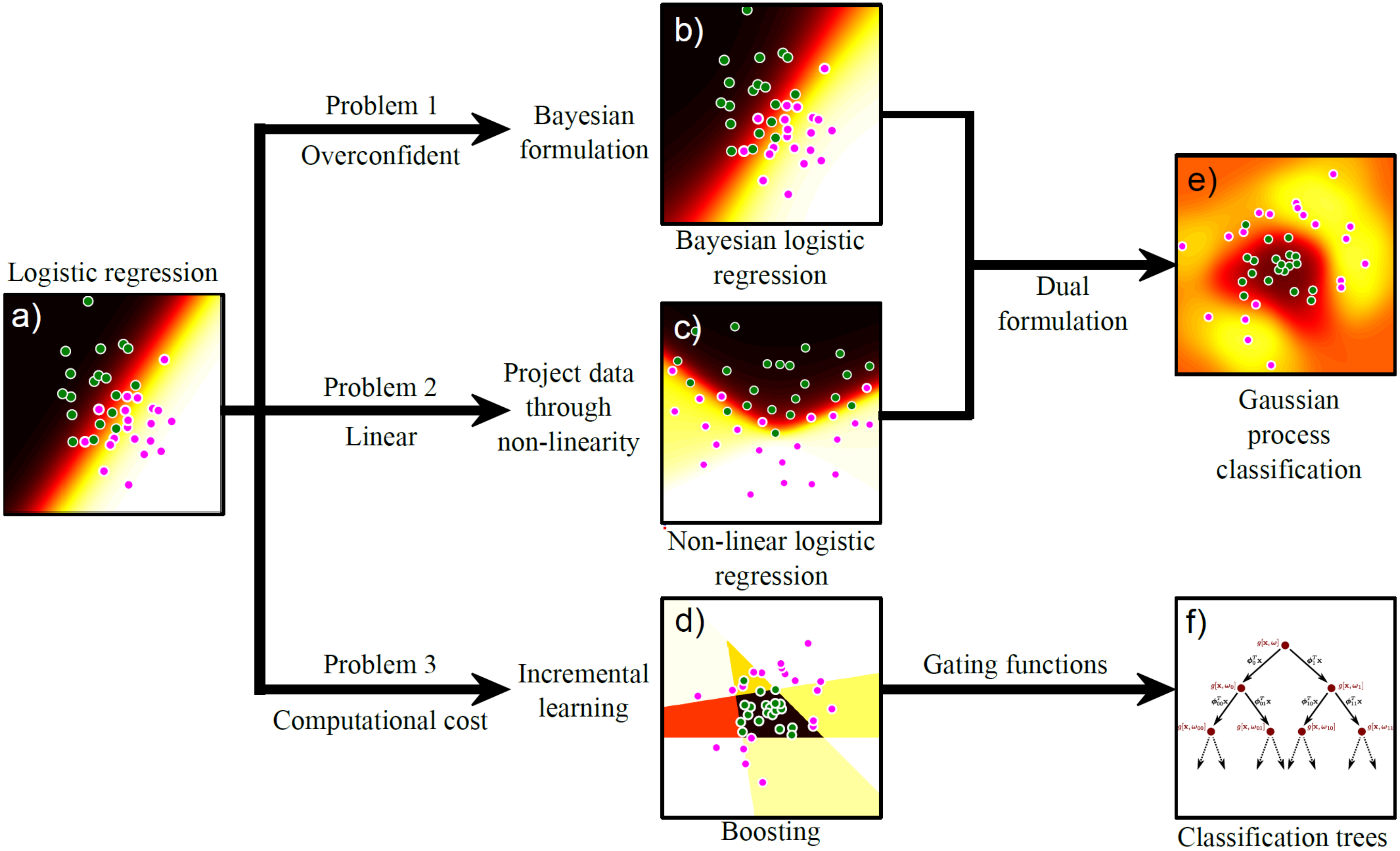
Simon J.D. Prince

Non-linear logistic regression in 2D

$$Pr(w\!=\!1|\mathbf{x}) = \text{sig}\left[\boldsymbol{\phi}^T\mathbf{f}[\mathbf{x}]\right] \qquad f_1[\mathbf{x}] = \arctan\left[\boldsymbol{\alpha}_1^T\mathbf{x}\right] \qquad f_2[\mathbf{x}] = \arctan\left[\boldsymbol{\alpha}_2^T\mathbf{x}\right]$$



Simon J.D. Prince



Simon J.D. Prince

Dual Logistic Regression

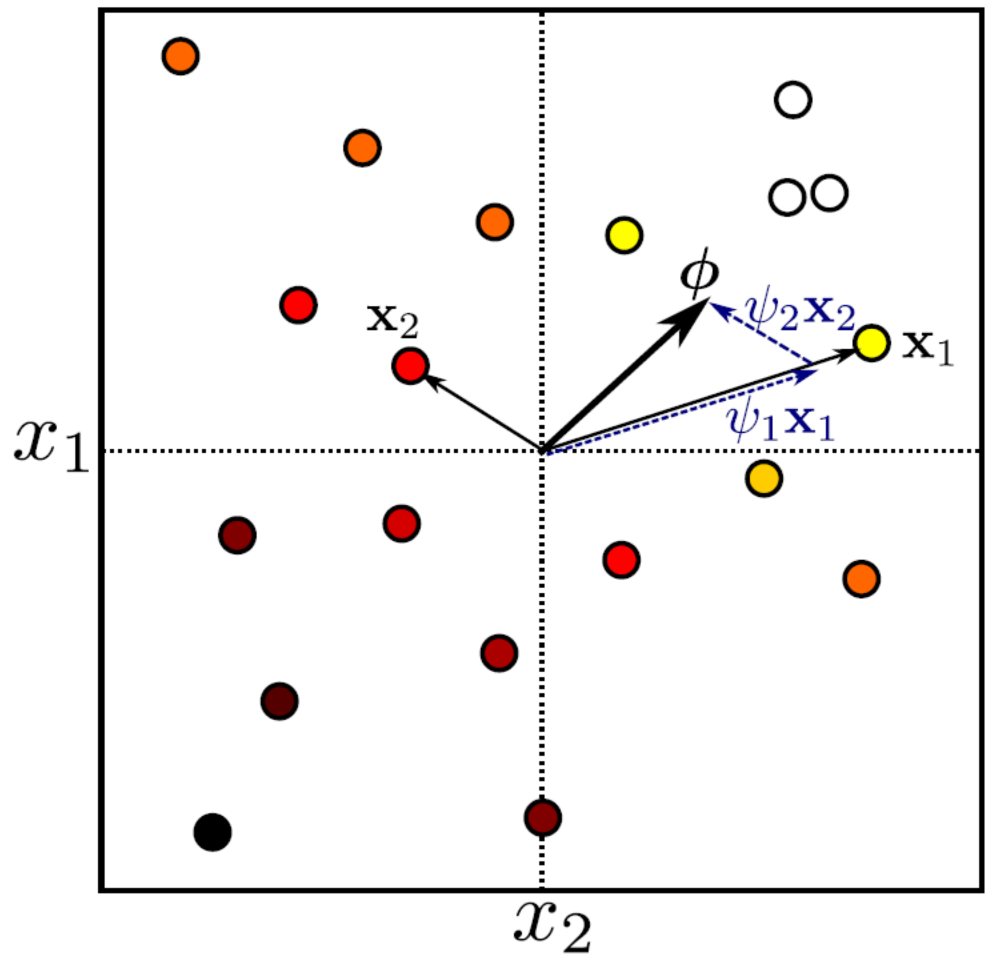
KEY IDEA:

Gradient $\boldsymbol{\Phi}$ is just a vector in the data space

Can represent as a weighted sum of the data points

$$\phi = \mathbf{X}\psi$$

Now solve for $\boldsymbol{\Psi}$. One parameter per training example.



Simon J.D. Prince

Maximum Likelihood

Likelihood

$$Pr(\mathbf{w}|\mathbf{X},\boldsymbol{\psi}) = \prod_{i=1}^{I} \text{Bern}_{w_i}\left[\text{sig}[a_i]\right] = \prod_{i=1}^{I} \text{Bern}_{w_i}\left[\text{sig}[\boldsymbol{\psi}^T\mathbf{X}^T\mathbf{x}_i]\right]$$

Derivatives

$$\frac{\partial L}{\partial \boldsymbol{\psi}} = -\sum_{i=1}^{I}\left(\text{sig}[a_i] - w_i\right)\mathbf{X}^T\mathbf{x}_i$$

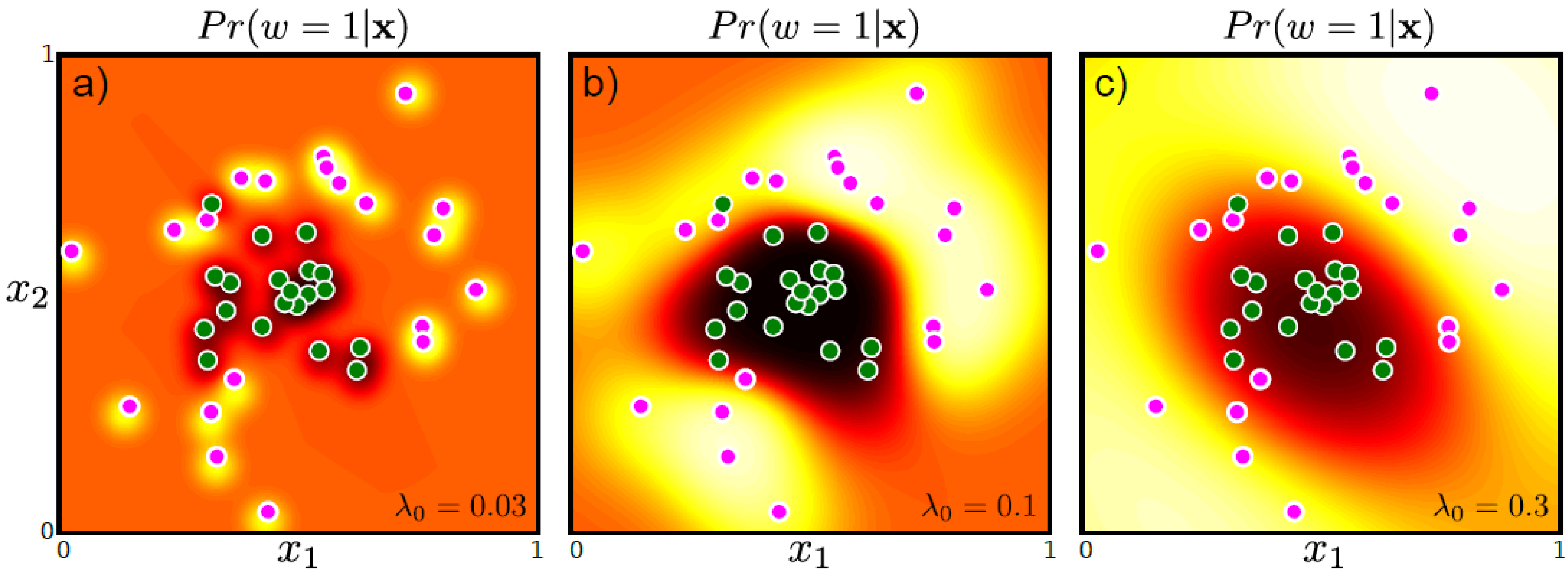
$$\frac{\partial^2 L}{\partial \boldsymbol{\psi}^2} = -\sum_{i=1}^{I}\text{sig}[a_i]\left(1 - \text{sig}[a_i]\right)\mathbf{X}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{X}$$

Depend only depend on inner products!

Simon J.D. Prince

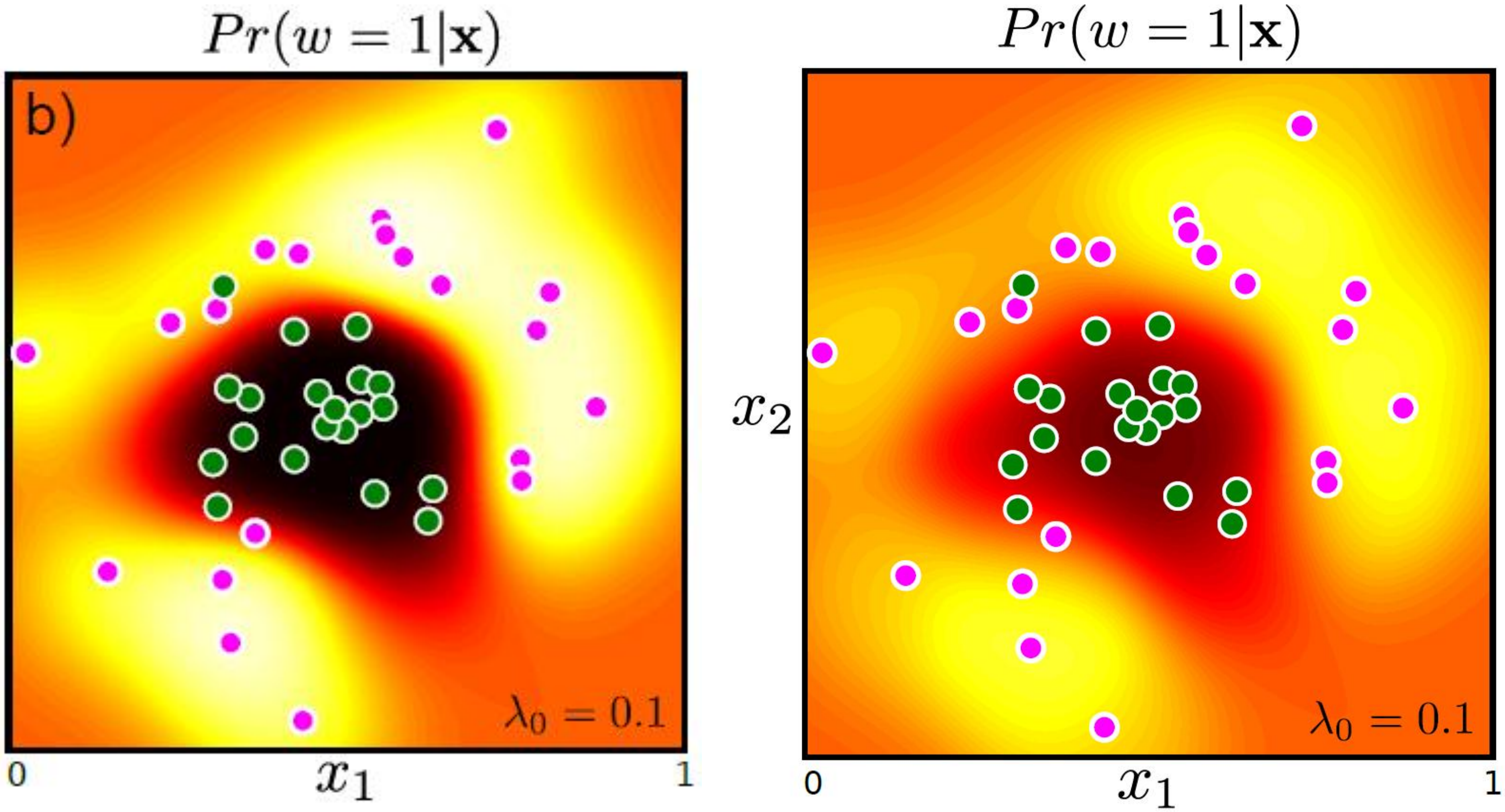
Kernel Logistic Regression

$$k[\mathbf{x}_i, \mathbf{x}_j] = \exp\left[-0.5\left(\frac{(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j)}{\lambda_0^2}\right)\right]$$



Simon J.D. Prince

ML vs. Bayesian



Bayesian case is known as Gaussian process classification

Simon J.D. Prince

Relevance vector classification

Apply sparse prior to dual variables (dual logistic regression):

$$Pr(\boldsymbol{\psi}) = \prod_{i=1}^{I} \text{Stud}_{\boldsymbol{\psi}_i}[0, 1, \nu]$$

As before, write as marginalization of dual variables:

$$\begin{aligned} Pr(\boldsymbol{\psi}) &= \prod_{i=1}^{I} \int \text{Norm}_{\psi_i}\left[0, \frac{1}{h_i}\right] \text{Gam}_{h_i}\left[\frac{\nu}{2}, \frac{\nu}{2}\right] dh_i \\ &= \int \text{Norm}_{\boldsymbol{\psi}}[0, \mathbf{H}^{-1}] \prod_{i=1}^{I} \text{Gam}_{h_i}[\nu/2, \nu/2] \, d\mathbf{H}. \end{aligned}$$

Simon J.D. Prince

Relevance vector classification

UNIVERSITÄT BONN

Apply sparse prior to dual variables:

$$Pr(\boldsymbol{\psi}) = \int \text{Norm}_{\boldsymbol{\psi}}[0, \mathbf{H}^{-1}] \prod_{i=1}^{I} \text{Gam}_{h_i}[\nu/2, \nu/2] \, d\mathbf{H}.$$

Gives likelihood:

$$\begin{aligned} &Pr(\mathbf{w}|\mathbf{X}) \\ &\quad = \int Pr(\mathbf{w}|\mathbf{X}, \boldsymbol{\psi}) Pr(\boldsymbol{\psi}) \, d\boldsymbol{\psi} \\ &\quad = \int\int \prod_{i=1}^{I} \text{Bern}_{w_i}\Big[\text{sig}[\boldsymbol{\psi}^T \mathbf{K}[\mathbf{X}, \mathbf{x}_i]]\Big] \text{Norm}_{\boldsymbol{\psi}}[0, \mathbf{H}^{-1}] \, \text{Gam}_{h_i}[\nu/2, \nu/2] \, d\mathbf{H} d\boldsymbol{\psi}. \end{aligned}$$

Simon J.D. Prince

Relevance vector classification

Laplace approximation:

$$Pr(\mathbf{w}|\mathbf{X}) \approx$$
$$\int \prod_{i=1}^{I} (2\pi)^{I/2} |\mathbf{\Sigma}|^{0.5} \mathrm{Bern}_{w_i}[\mathrm{sig}[\boldsymbol{\mu}^T \mathbf{K}[\mathbf{X}, \mathbf{x}_i]]] \mathrm{Norm}_{\boldsymbol{\mu}}[0, \mathbf{H}^{-1}] \mathrm{Gam}_{h_i}\left[\frac{\nu}{2}, \frac{\nu}{2}\right] d\mathbf{H}$$

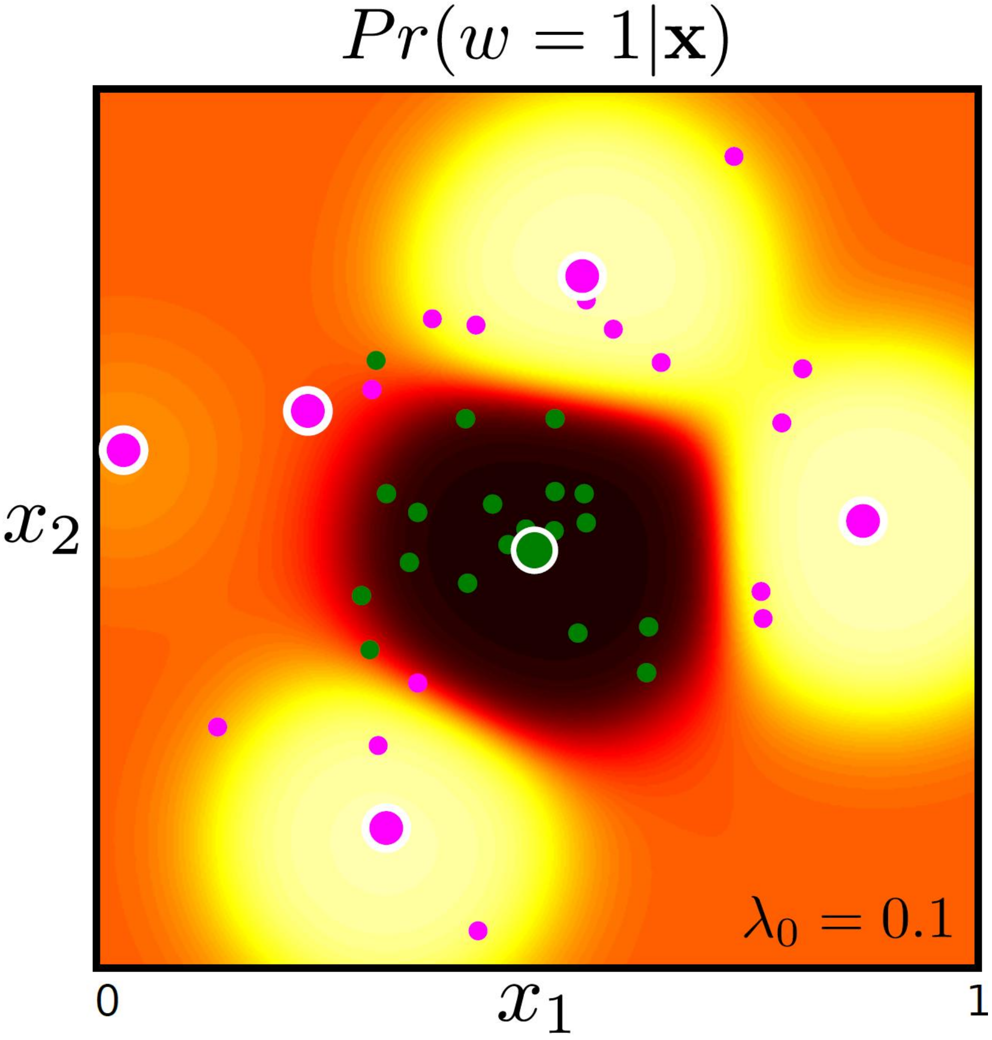
Second approximation:

$$Pr(\mathbf{w}|\mathbf{X}) \approx$$
$$\max_{\mathbf{H}} \left[\prod_{i=1}^{I} (2\pi)^{I/2} |\mathbf{\Sigma}|^{0.5} \mathrm{Bern}_{w_i}[\mathrm{sig}[\boldsymbol{\mu}^T \mathbf{K}[\mathbf{X}, \mathbf{x}_i]]] \mathrm{Norm}_{\boldsymbol{\mu}}[0, \mathbf{H}^{-1}] \mathrm{Gam}_{h_i}\left[\frac{\nu}{2}, \frac{\nu}{2}\right]\right]$$

To solve, alternately update hidden variables in **H** and mean and variance of Laplace approximation.

Simon J.D. Prince

Relevance vector classification



The final solution only depends on a very small number of examples – efficient

Simon J.D. Prince

UNIVERSITÄT BONN