Introduction to Machine Learning

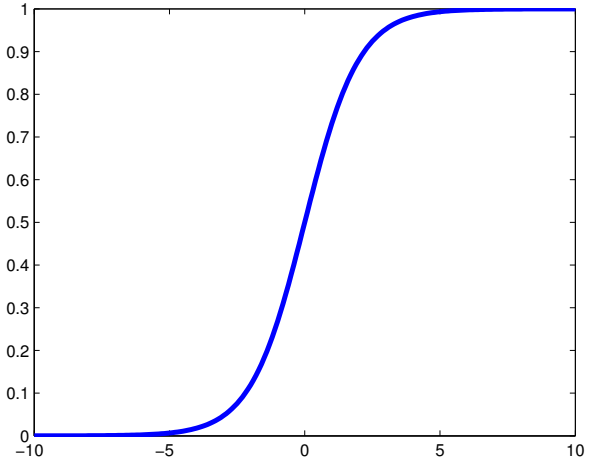
Brown University CSCI 1950-F, Spring 2012
Prof. Erik Sudderth

Lecture 11:
ML & MAP Estimation for Logistic Regression
Bayesian Prediction via Laplace Approximations

Many figures courtesy Kevin Murphy's textbook,
Machine Learning: A Probabilistic Perspective

Logistic Regression

$$p(y_i \mid x_i, w) = \mathrm{Ber}(y_i \mid \mathrm{sigm}(w^T \phi(x_i)))$$



- Linear discriminant analysis:

$$\phi(x_i) = [1, x_{i1}, x_{i2}, \ldots, x_{id}]$$

- Quadratic discriminant analysis:

$$\phi(x_i) = [1, x_{i1}, \ldots, x_{id}, x_{i1}^2, x_{i1}x_{i2}, x_{i2}^2, \ldots]$$

- Can derive weights from Gaussian generative model if that happens to be known, but more generally:
  - Choose any convenient feature set $\phi(x)$
  - Do discriminative Bayesian learning:

$$p(w \mid x, y) \propto p(w) \prod_{i=1}^{N} \mathrm{Ber}(y_i \mid \mathrm{sigm}(w^T \phi(x_i)))$$

Learning via Optimization

ML Estimate: $\hat{w} = \arg\min_w -\sum_i \log p(y_i \mid x_i, w)$

MAP Estimate: $\hat{w} = \arg\min_w -\log p(w) - \sum_i \log p(y_i \mid x_i, w)$

Gradient vectors:

$$f : \mathbb{R}^M \to \mathbb{R}$$
$$\nabla_w f : \mathbb{R}^M \to \mathbb{R}^M$$
$$(\nabla_w f(w))_k = \frac{\partial f(w)}{\partial w_k}$$

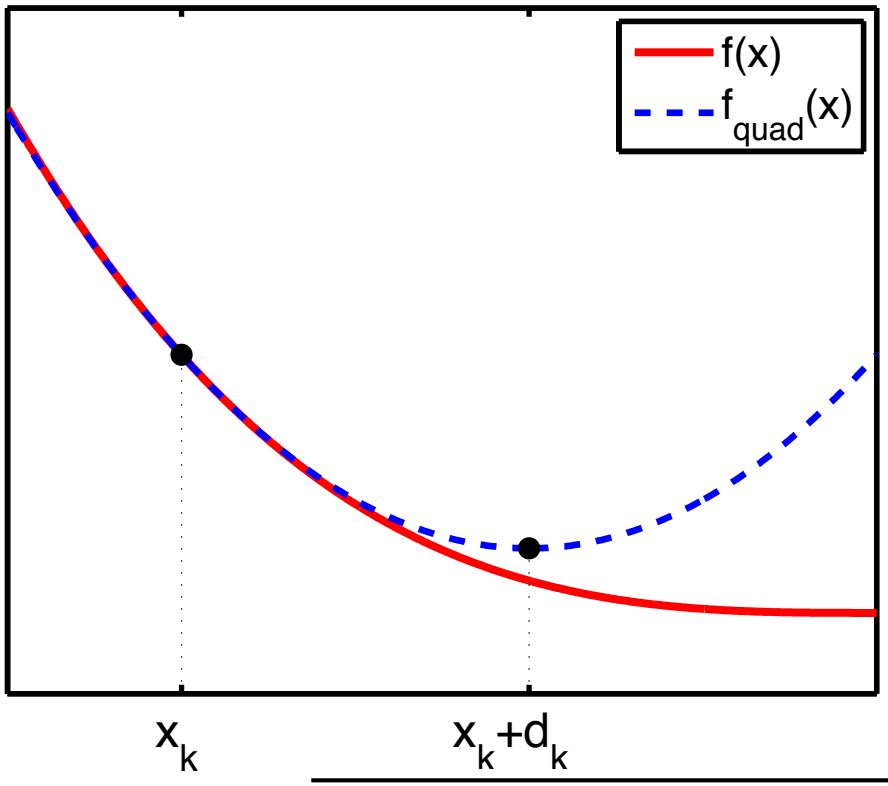
Hessian matrices:

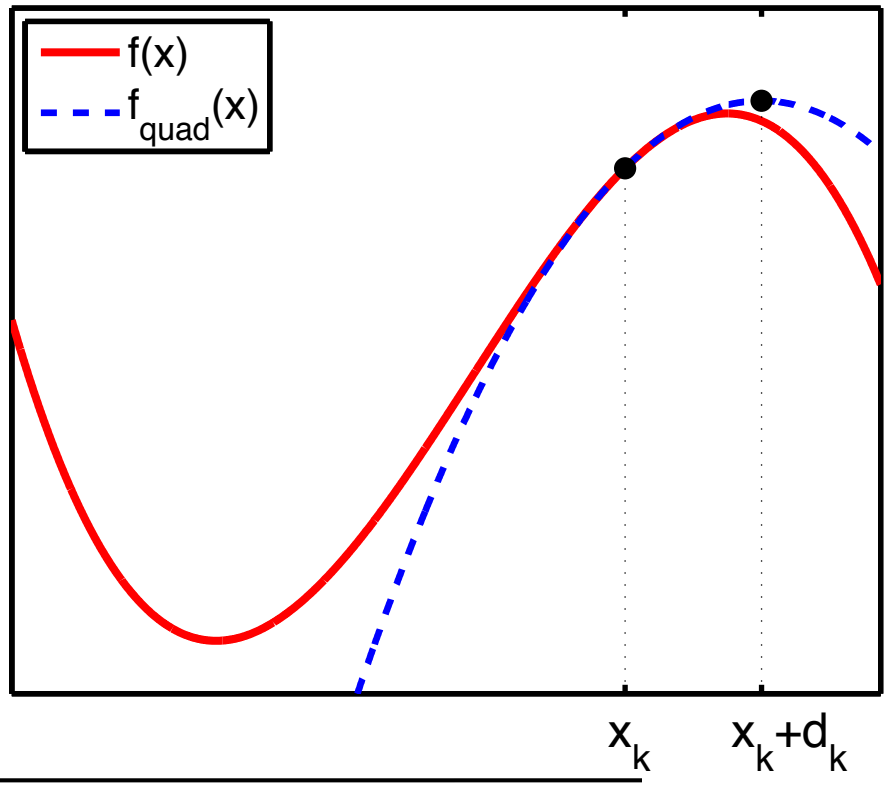
$$\nabla_w^2 f : \mathbb{R}^M \to \mathbb{R}^{M \times M}$$
$$(\nabla_w f(w))_{k,\ell} = \frac{\partial^2 f(w)}{\partial w_k \partial w_\ell}$$

Iterative Optimization of Smooth Functions:

- Initialize somewhere, use gradients to take steps towards better (by convention, smaller) values
- For convex objectives, there is a unique global optimum

Newton's Method

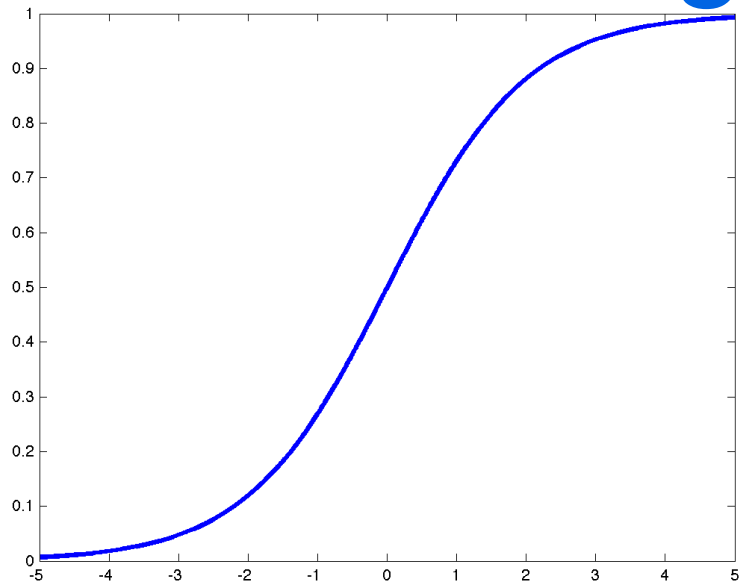




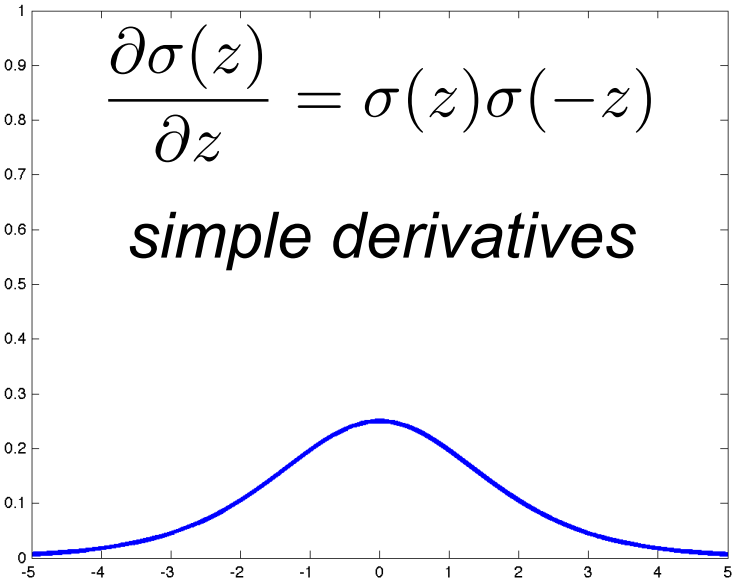
Algorithm 6.1: Newton's method for minimizing a strictly convex function

Initialize $\boldsymbol{\theta}_0$;
for $k = 1, 2, \ldots$ *until convergence* **do**
  Evaluate $\mathbf{g}_k = \nabla f(\boldsymbol{\theta}_k)$;
  Evaluate $\mathbf{H}_k = \nabla^2 f(\boldsymbol{\theta}_k)$;
  Solve $\mathbf{H}_k \mathbf{d}_k = -\mathbf{g}_k$ for $\mathbf{d}_k$;
  Use line search to find stepsize η_k along $\mathbf{d}_k$;
  $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \eta_k \mathbf{d}_k$;

The Logistic Function

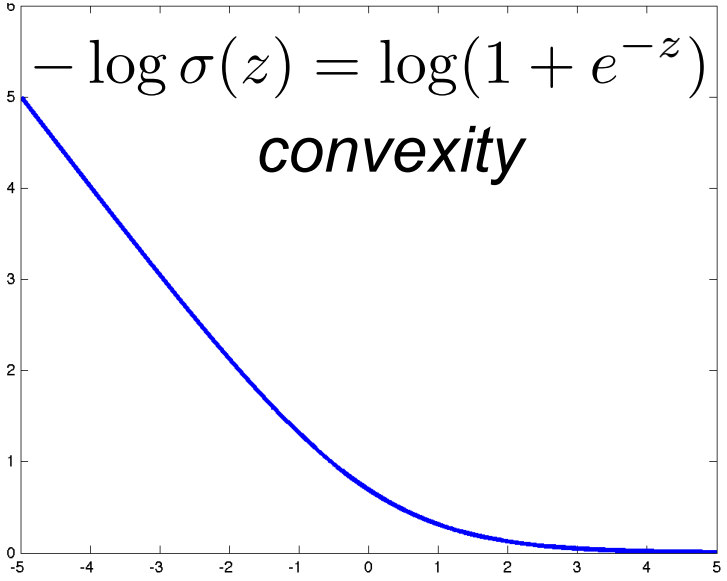


$$\sigma(z) = \text{sigm}(z) = \frac{1}{1+e^{-z}}$$



$$\frac{\partial \sigma(z)}{\partial z} = \sigma(z)\sigma(-z)$$

simple derivatives



$$-\log \sigma(z) = \log(1+e^{-z})$$

convexity

- Symmetry:

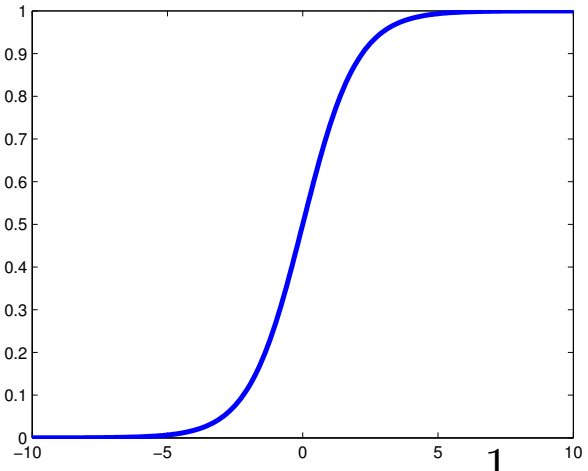
$$\sigma(-z) = 1 - \sigma(z)$$

- Log-odds ratio:

$$\log \frac{\sigma(z)}{1-\sigma(z)} = z$$

Logistic Regression ML: Formulation

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \mathrm{Ber}(y_i|\mathrm{sigm}(\mathbf{w}^T\mathbf{x}_i))$$
$$\phi(x_i) = x_i$$
$$\mu_i = \mathrm{sigm}(\mathbf{w}^T\mathbf{x}_i)$$
$$\mathbf{S} \triangleq \mathrm{diag}(\mu_i(1-\mu_i))$$



$$\sigma(z) = \mathrm{sigm}(z) = \frac{1}{1+e^{-z}}$$

Goal: Minimize negative conditional log-likelihood

$$f(w) = -\log p(y \mid x, w) = -\sum_{i=1}^{N}[y_i \log \mu_i + (1-y_i)\log(1-\mu_i)]$$

$$\mathbf{g} = \frac{d}{d\mathbf{w}} f(\mathbf{w}) = \sum_i (\mu_i - y_i)\mathbf{x}_i = \mathbf{X}^T(\boldsymbol{\mu} - \mathbf{y})$$

$$\mathbf{H} = \frac{d}{d\mathbf{w}} \mathbf{g}(\mathbf{w})^T = \sum_i (\nabla_{\mathbf{w}} \mu_i)\mathbf{x}_i^T = \sum_i \mu_i(1-\mu_i)\mathbf{x}_i\mathbf{x}_i^T = \mathbf{X}^T\mathbf{S}\mathbf{X}$$

Logistic Regression ML: Optimization

- Gradient descent: Apply using derivative formulas below
- Newton's method: *Iteratively Reweighted Least Squares (IRLS)*

$$w_{k+1} = w_k + (X^T S_k X)^{-1} X^T (y - \mu_k) \qquad \mu_k = \sigma(X w_k)$$

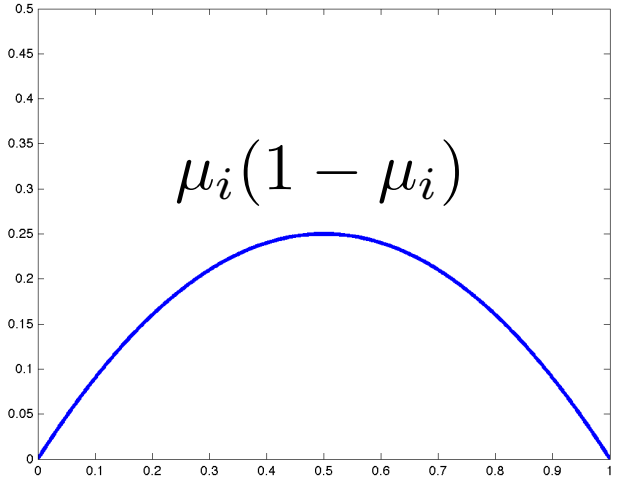
$$\mathbf{S} \triangleq \mathrm{diag}(\mu_i(1-\mu_i))$$

- Weighted least squares problems:

$$J(w) = \sum_{i=1}^{N} \lambda_i (y_i - w^T x_i)^2$$

$$\hat{w} = (X^T \Lambda X)^{-1} X^T \Lambda y$$

$$\Lambda = \mathrm{diag}(\lambda_i)$$



$$\mathbf{g} = \frac{d}{d\mathbf{w}} f(\mathbf{w}) = \sum_i (\mu_i - y_i)\mathbf{x}_i = \mathbf{X}^T(\boldsymbol{\mu} - \mathbf{y})$$

$$\mathbf{H} = \frac{d}{d\mathbf{w}} \mathbf{g}(\mathbf{w})^T = \sum_i (\nabla_{\mathbf{w}} \mu_i)\mathbf{x}_i^T = \sum_i \mu_i(1-\mu_i)\mathbf{x}_i\mathbf{x}_i^T = \mathbf{X}^T\mathbf{S}\mathbf{X}$$

Logistic Regression ML: Degeneracies

$$\log \frac{p(y_i \mid x_i, w)}{1 - p(y_i \mid x_i, w)} = w^T \phi(x_i)$$

$$w_k > 0 : \text{increasing } \phi_k(x_i) \text{ makes } y_i = 1 \text{ more likely}$$

$$w_k < 0 : \text{increasing } \phi_k(x_i) \text{ makes } y_i = 1 \text{ less likely}$$

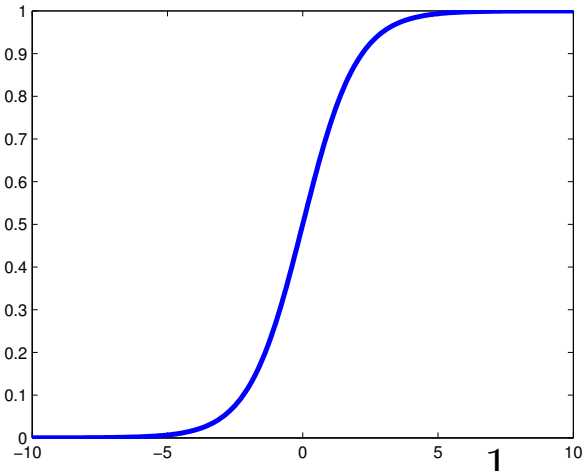
$$w_k = 0 : \text{changing } \phi_k(x_i) \text{ has no impact on } y_i$$

Pathological estimates that poorly generalize:

- If the training data is *linearly separable* in the feature space, conditional ML takes $||w|| \to \infty$ (threshold function)
- If a single feature happens to distinguish the two classes, that feature will be given infinite weight
- If multiple features are constant $(\phi_k(x_i) = c_k)$ across the training data, no unique optimum (numerical problems)

Logistic Regression: Gaussian MAP

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \mathrm{Ber}(y_i|\mathrm{sigm}(\mathbf{w}^T\mathbf{x}_i))$$
$$\phi(x_i) = x_i$$
$$\mu_i = \mathrm{sigm}(\mathbf{w}^T\mathbf{x}_i)$$
$$p(w) = \mathcal{N}(w \mid 0, \alpha^{-1}I)$$



$$\sigma(z) = \mathrm{sigm}(z) = \frac{1}{1+e^{-z}}$$

Goal: Minimize negative conditional log posterior

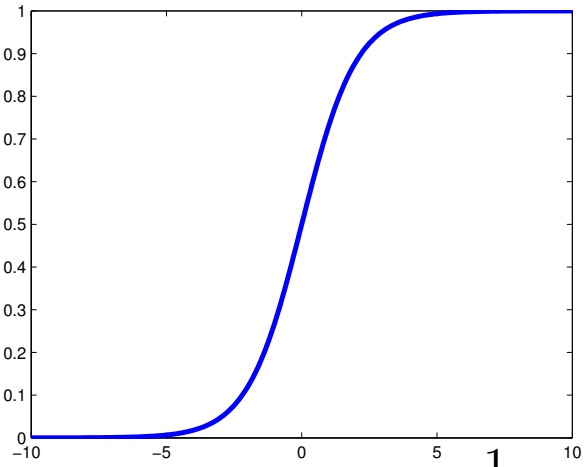
$$\bar{f}(w) = -\log p(y \mid x, w) - \log p(w) = -\sum_{i=1}^{N}[y_i \log \mu_i + (1-y_i)\log(1-\mu_i)] + \frac{\alpha}{2}w^T w$$

$$\bar{f}(w) = f(w) + \frac{\alpha}{2}w^T w$$
$$\nabla_w \bar{f}(w) = \nabla_w f(w) + \alpha w$$
$$\nabla_w^2 \bar{f}(w) = \nabla_w^2 f(w) + \alpha I$$

→ *simple modification to gradients for any model*

Logistic Regression: Bayes Prediction

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \text{Ber}(y_i|\text{sigm}(\mathbf{w}^T\mathbf{x}_i))$$
$$\phi(x_i) = x_i$$
$$\mu_i = \text{sigm}(\mathbf{w}^T\mathbf{x}_i)$$
$$p(w) = \mathcal{N}(w \mid 0, \alpha^{-1}I)$$



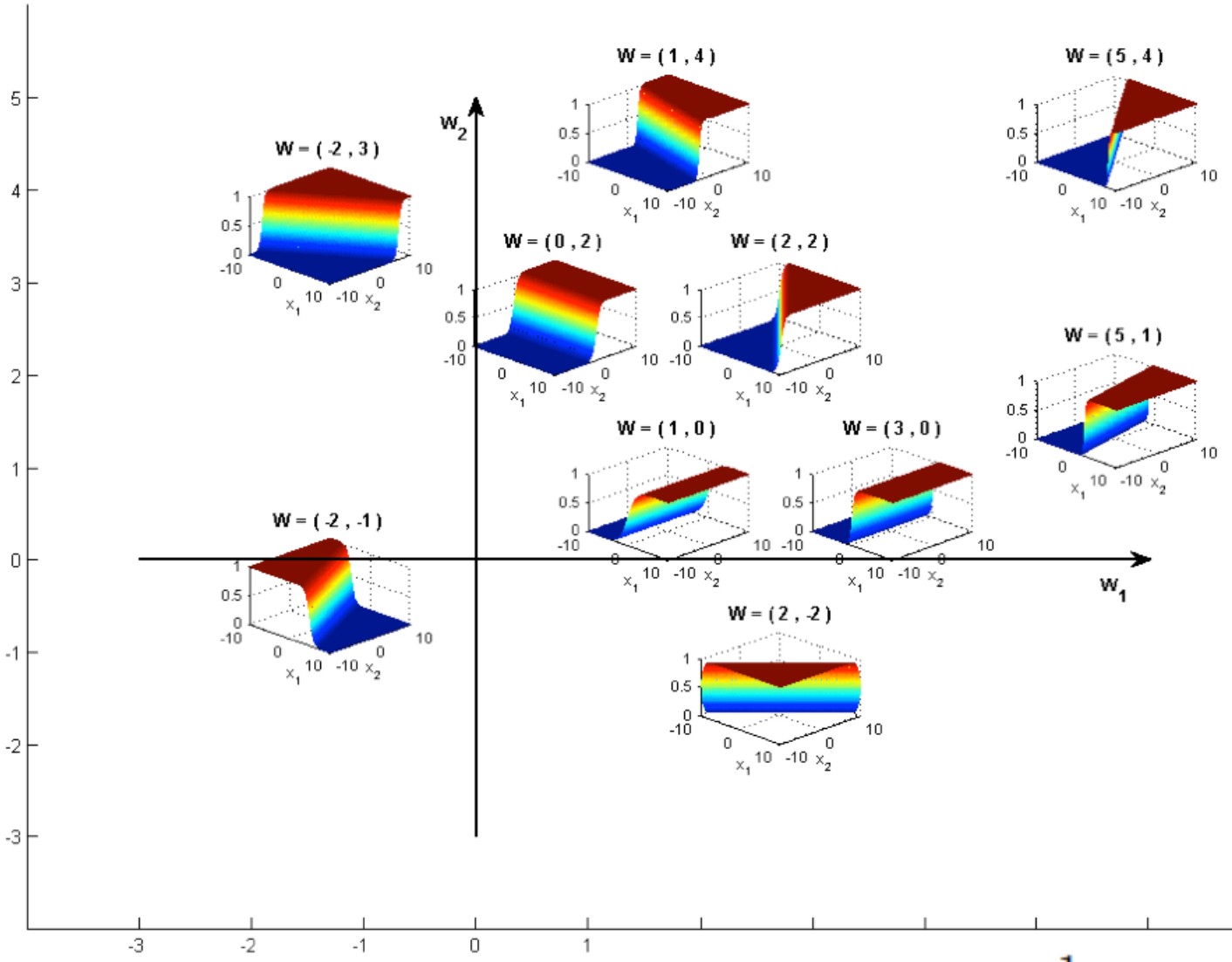
$$\sigma(z) = \text{sigm}(z) = \frac{1}{1+e^{-z}}$$

Goal: Find true posterior predictive distribution, integrating over posterior uncertainty in weights

$$p(y|\mathbf{x}, \mathcal{D}) = \int p(y|\mathbf{x}, \mathbf{w})p(\mathbf{w}|\mathcal{D})d\mathbf{w}$$

- The posterior distribution of the weight vector, under the logistic regression likelihood, is not a member of any standard, parametric family of distributions
- There is no closed form expression for marginal likelihood

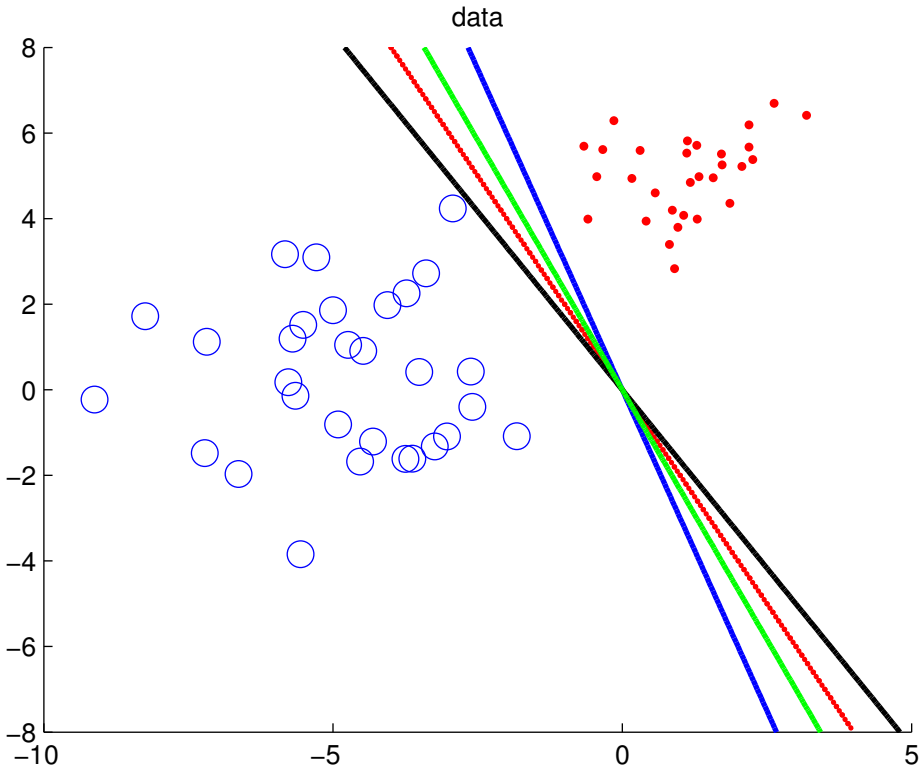
Logistic Regression Intuition



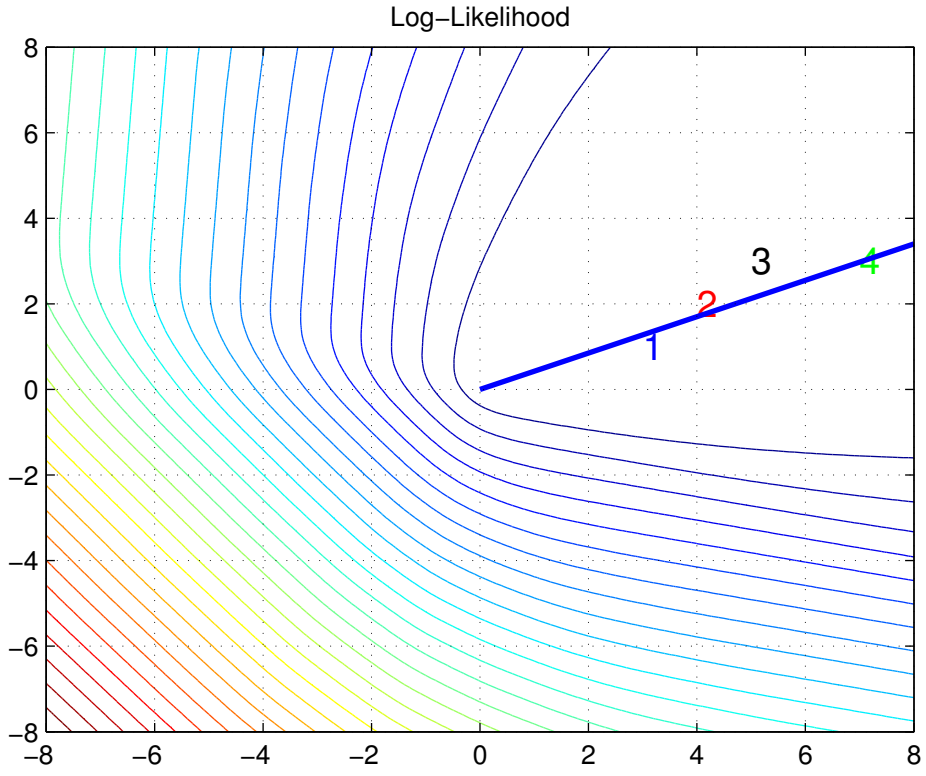
$$p(y|\mathbf{x}, \mathbf{w}) = \mathrm{Ber}(y|\mathrm{sigm}(\mathbf{w}^T\mathbf{x}))$$

$$\mathrm{sigm}(\eta) := \frac{1}{1+\exp(-\eta)} = \frac{e^{\eta}}{e^{\eta}+1}$$

Logistic Regression Likelihood



Linearly Separable Data



Log-likelihood Function

$$f(w) = -\log p(y \mid x, w) = -\sum_{i=1}^{N} [y_i \log \mu_i + (1 - y_i) \log(1 - \mu_i)]$$

$$\mu_i = \text{sigm}(\mathbf{w}^T \mathbf{x}_i)$$

Laplace (Gaussian) Approximations

- Perform Taylor expansion of posterior *energy function*:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z}e^{-E(\boldsymbol{\theta})} \qquad E(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}, \mathcal{D})$$

$$Z = p(\mathcal{D})$$

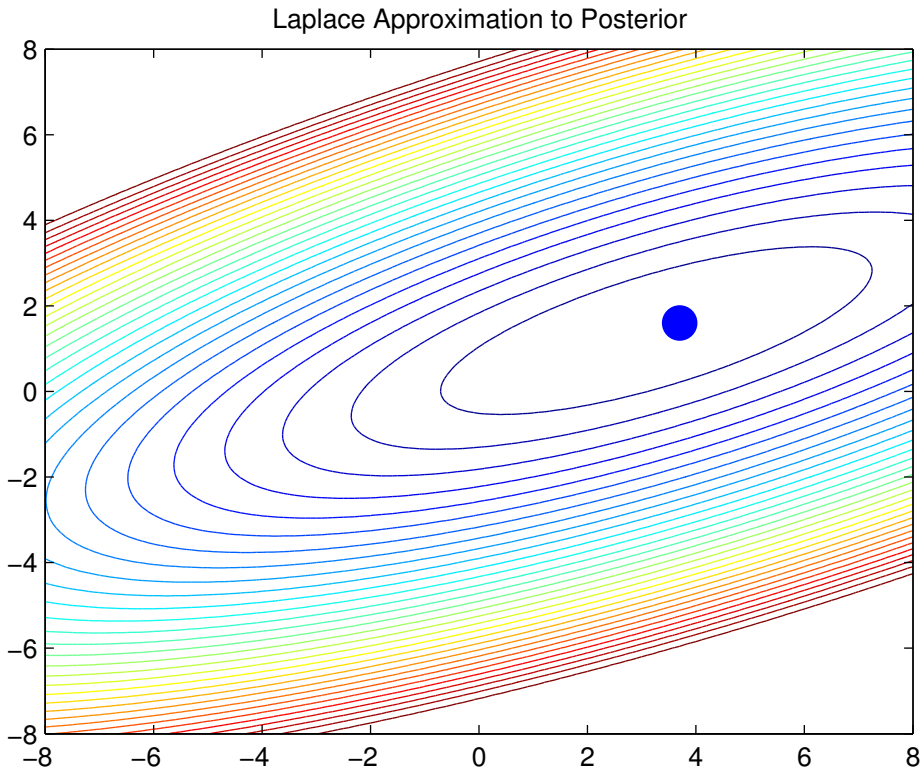
$$E(\boldsymbol{\theta}) \approx E(\boldsymbol{\theta}^*) + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbf{g} + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbf{H} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)$$

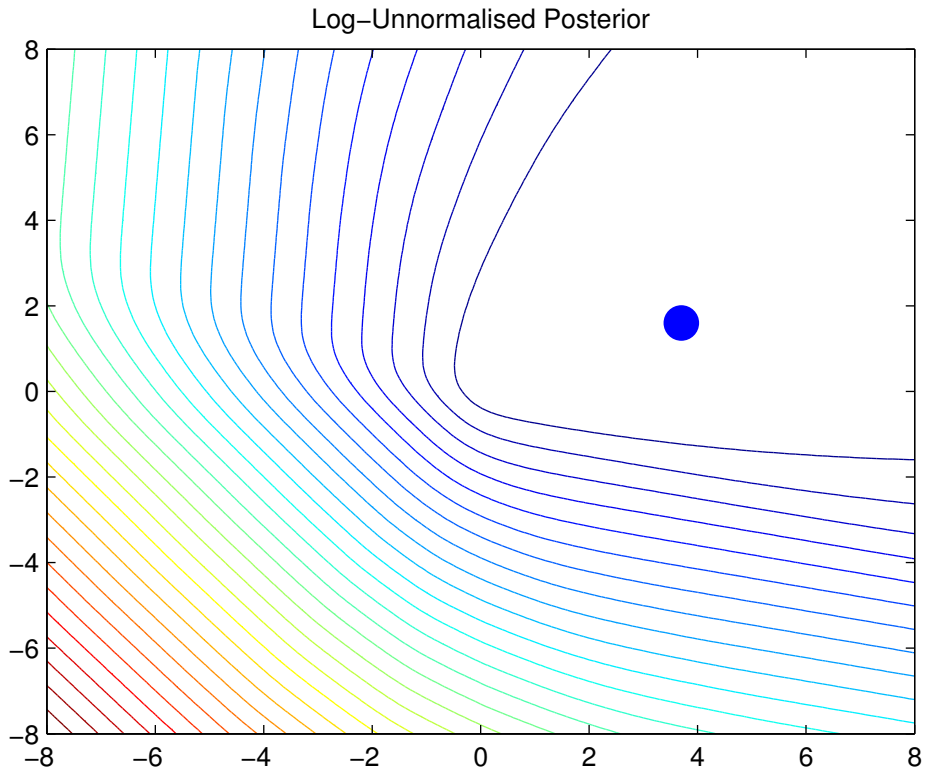
$$\mathbf{g} \triangleq \nabla E(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}^*} \qquad \mathbf{H} \triangleq \frac{\partial^2 E(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}^*}$$

- Suppose we expand around a posterior mode θ^*:
  - Gradient will be zero
  - Hessian will (for many priors) be positive definite

$$\begin{aligned}
\hat{p}(\boldsymbol{\theta}|\mathcal{D}) &\approx \frac{1}{Z}e^{-E(\boldsymbol{\theta}^*)} \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbf{H} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right] \\
&= \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\theta}^*, \mathbf{H}^{-1}) \\
p(\mathcal{D}) &\approx e^{-E(\boldsymbol{\theta}^*)}(2\pi)^{D/2}|\mathbf{H}|^{-\frac{1}{2}}
\end{aligned}$$

Laplace Approximation of LR Posterior





Laplace (Gaussian) Approximation

$$p(\mathbf{w}|\mathcal{D}) \approx \mathcal{N}(\mathbf{w}|\hat{\mathbf{w}}, \mathbf{H}^{-1})$$

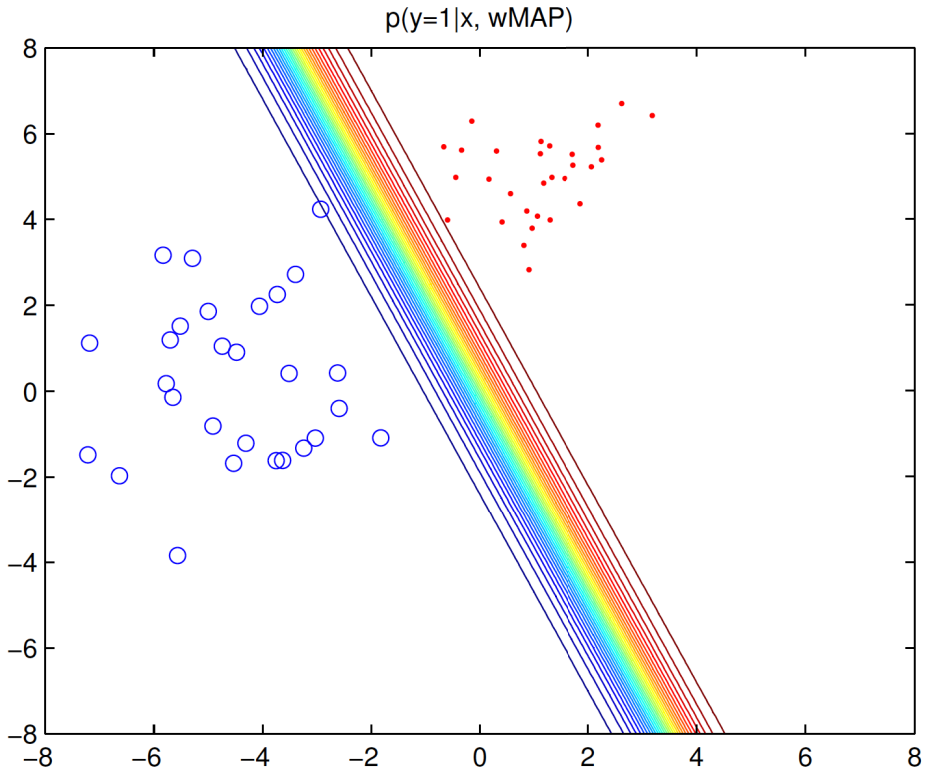
$$\mathbf{H} = -\nabla^2 E(\mathbf{w})|_{\hat{\mathbf{w}}}$$

Log Posterior Distribution

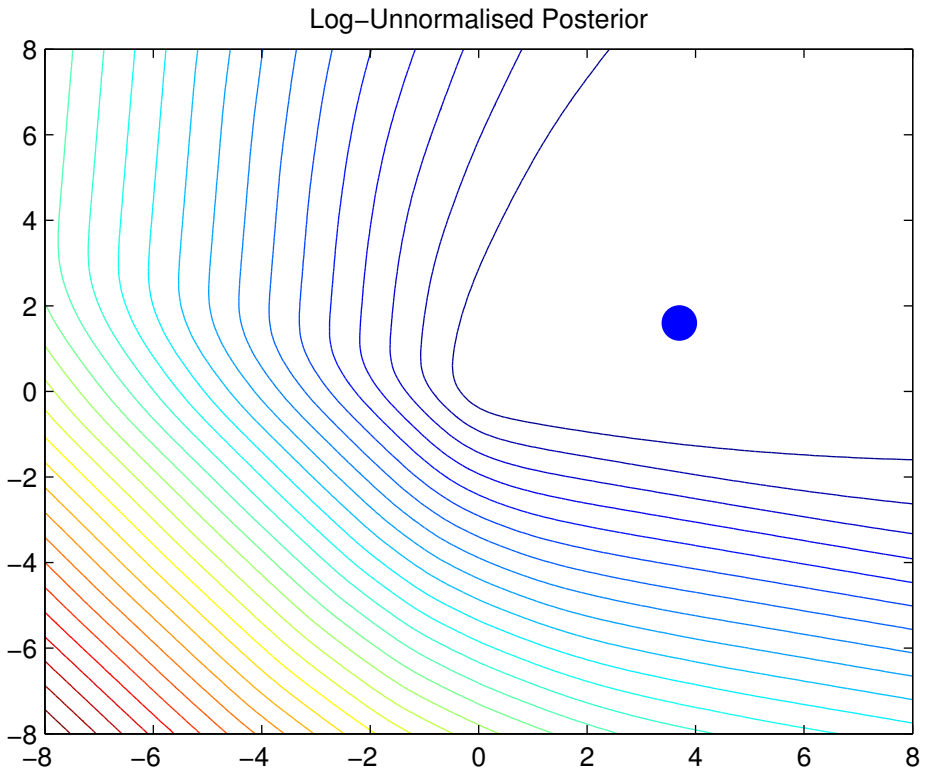
$$E(\mathbf{w}) = -(\log p(\mathcal{D}|\mathbf{w}) + \log p(\mathbf{w}))$$

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} E(\mathbf{w})$$

MAP Prediction Rule



Prediction from MAP Estimate

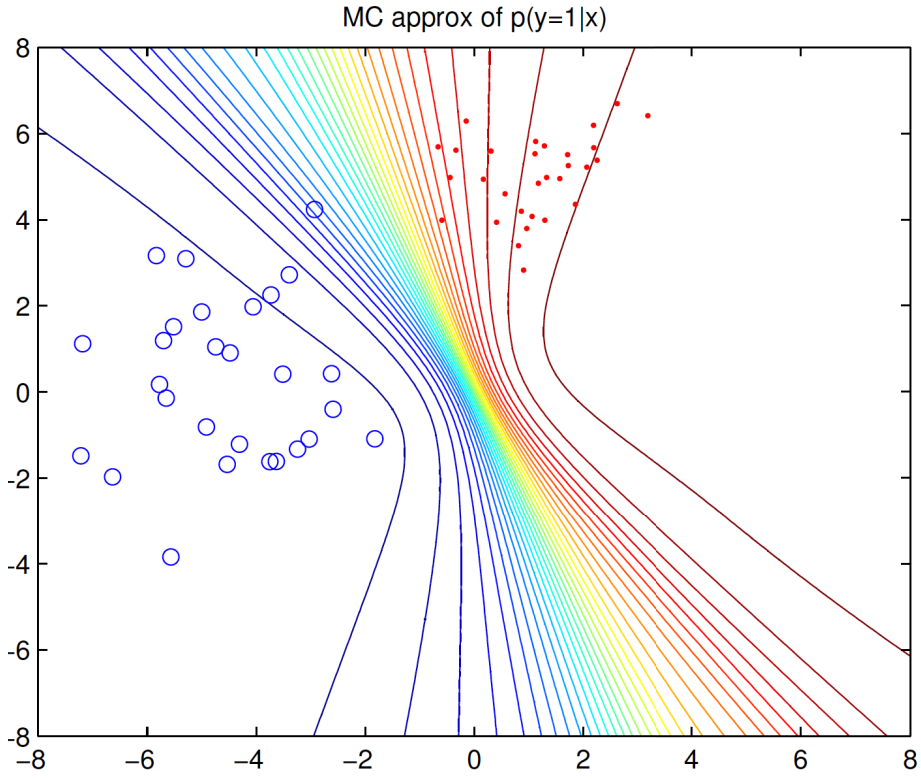


Log Posterior Distribution

$$E(\mathbf{w}) = -(\log p(\mathcal{D}|\mathbf{w}) + \log p(\mathbf{w}))$$

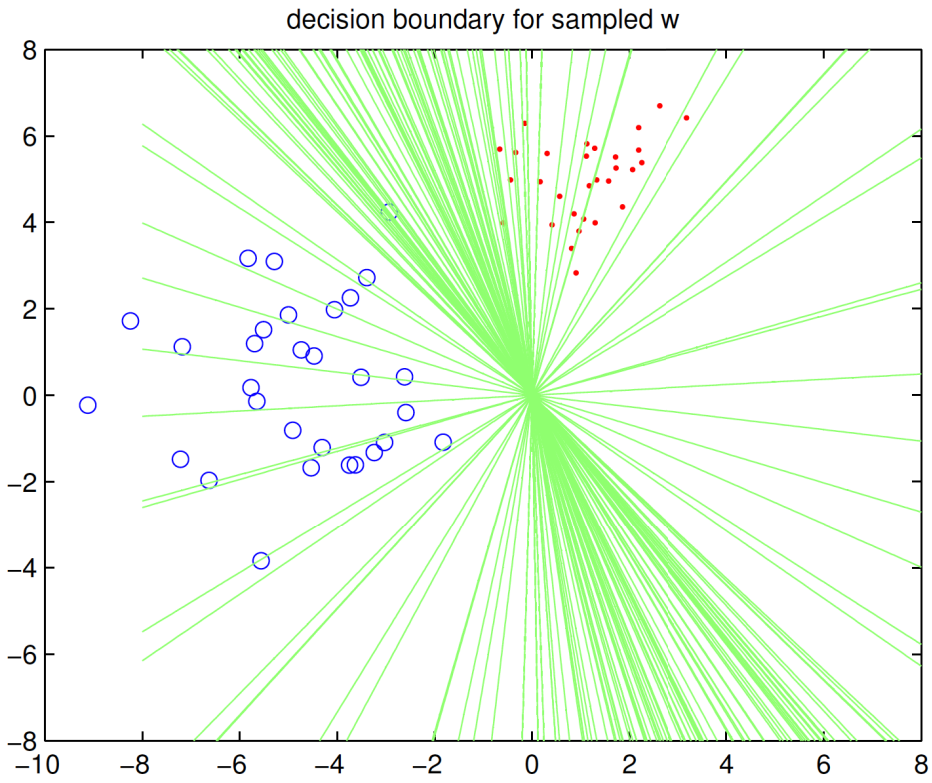
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} E(\mathbf{w})$$

True Predictive Marginal Distribution



Numerical Averaging over Monte Carlo Samples

$$p(y|\mathbf{x},\mathcal{D}) = \int p(y|\mathbf{x},\mathbf{w})p(\mathbf{w}|\mathcal{D})d\mathbf{w}$$



Samples from Posterior

$$p(y=1|\mathbf{x},\mathcal{D}) \approx \frac{1}{S}\sum_{s=1}^{S}\text{sigm}((\mathbf{w}^s)^T\mathbf{x})$$

$$\mathbf{w}^s \sim p(\mathbf{w}|\mathcal{D})$$