Introduction to Machine Learning

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

Lecture 12:
Laplace Approximation for Logistic Regression
Exponential Families
Generalized Linear Models

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

Bayesian Logistic Regression

Posterior Predictive Distribution

$$p(y_{\text{test}} \mid x_{\text{test}}, y_{\text{train}}, x_{\text{train}}) = \int_{\Theta} p(y_{\text{test}} \mid x_{\text{test}}, \theta) p(\theta \mid y_{\text{train}}, x_{\text{train}})\, d\theta$$

- No closed form for logistic regression, must approximate.

Posterior Parameter Estimation

$$p(y_{\text{test}} \mid x_{\text{test}}, y_{\text{train}}, x_{\text{train}}) \approx p(y_{\text{test}} \mid x_{\text{test}}, \hat{\theta})$$

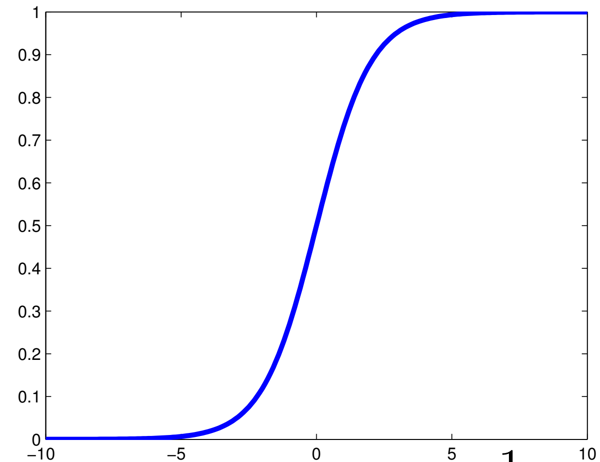
MAP: $$\hat{\theta} = \arg\max_{\theta}\ \log p(\theta) + \sum_i \log p(y_i \mid x_i, \theta)$$

ML: $$\hat{\theta} = \arg\max_{\theta}\ \sum_i \log p(y_i \mid x_i, \theta)$$

- Gradient algorithms can be used to optimize both objectives
- Convexity guarantees there is a single, global optimum

Logistic Regression: Bayes Prediction

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



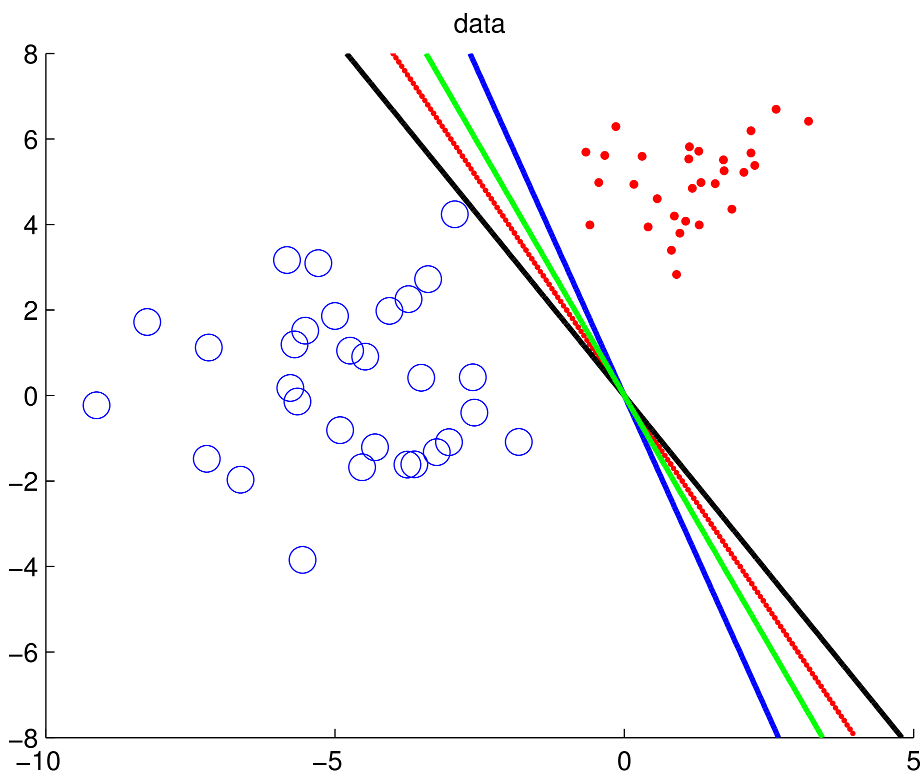
$$\sigma(z) = \mathrm{sigm}(z) = \frac{1}{1+e^{-z}}$$

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

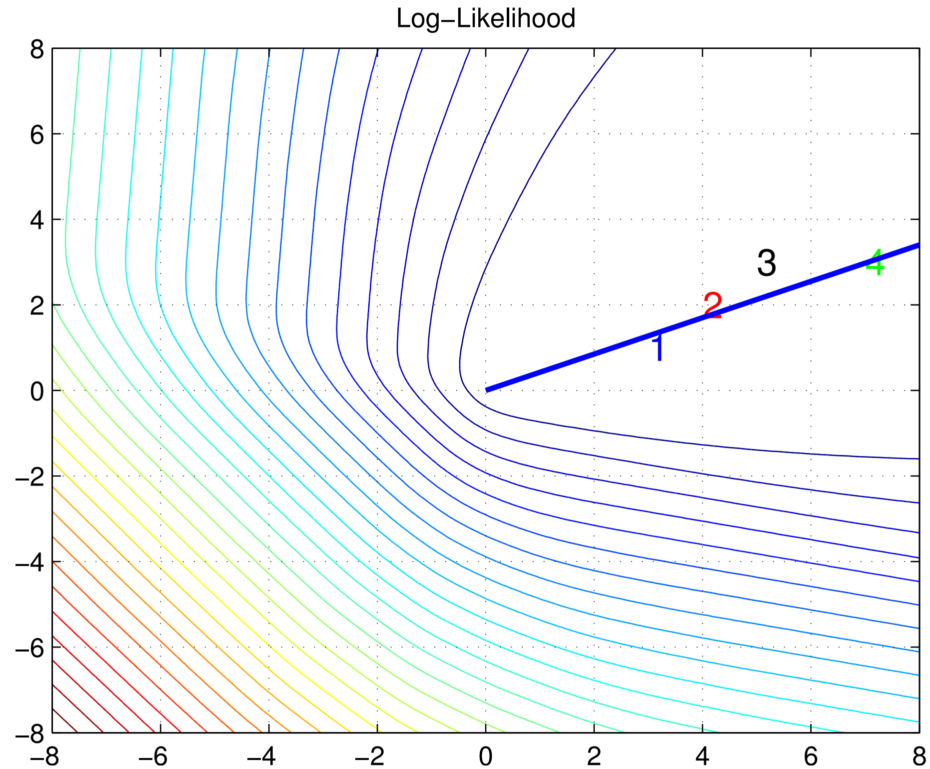
$$p(y|\mathbf{x}, \mathcal{D}) = \int p(y|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathcal{D}) d\mathbf{w}$$

- 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 Likelihood



Linearly Separable Data

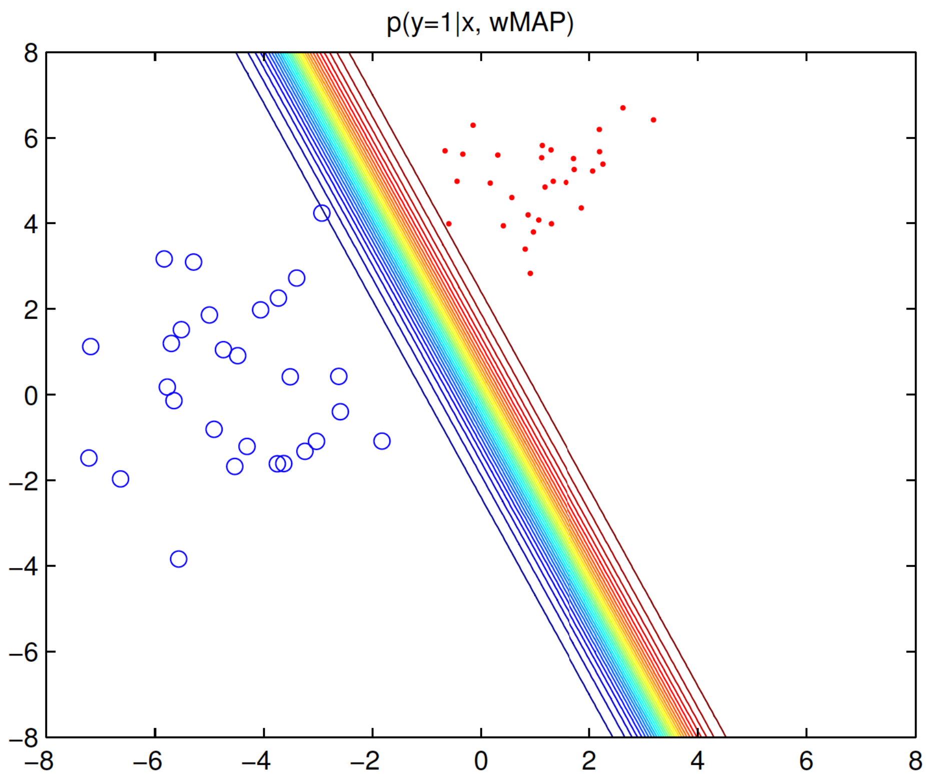


Log-likelihood Function

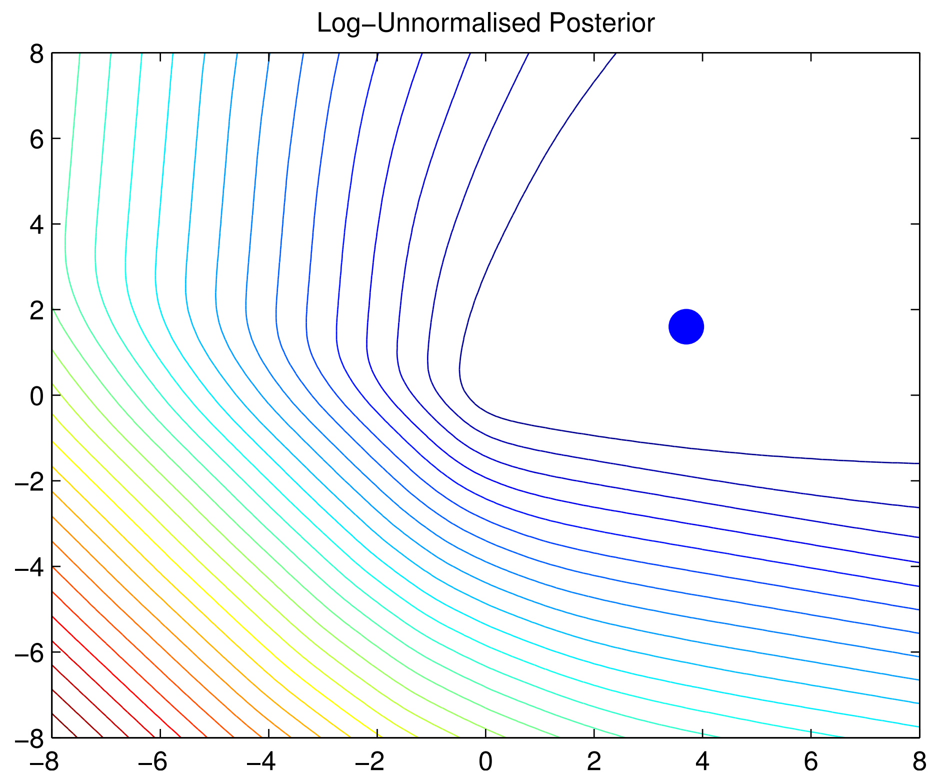
$$f(w) = -\log p(y \mid x, w) = -\sum_{i=1}^{N} [y_i \log \mu_i + (1 - y_i) \log(1 - \mu_i)]$$

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

MAP Prediction Rule



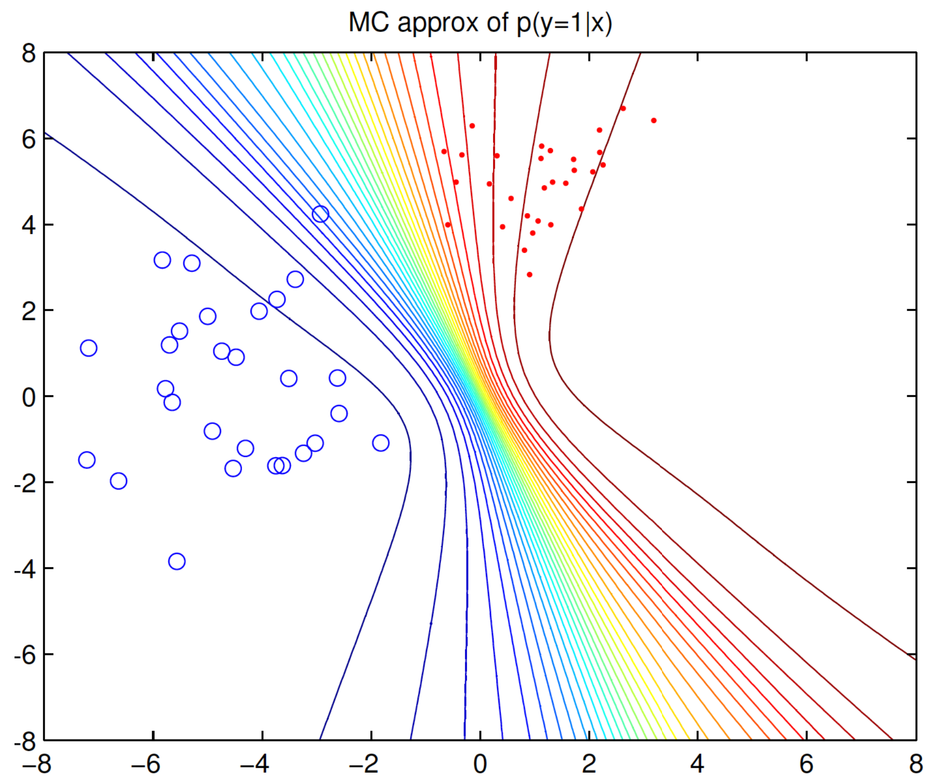
Prediction from MAP Estimate



Log Posterior Distribution

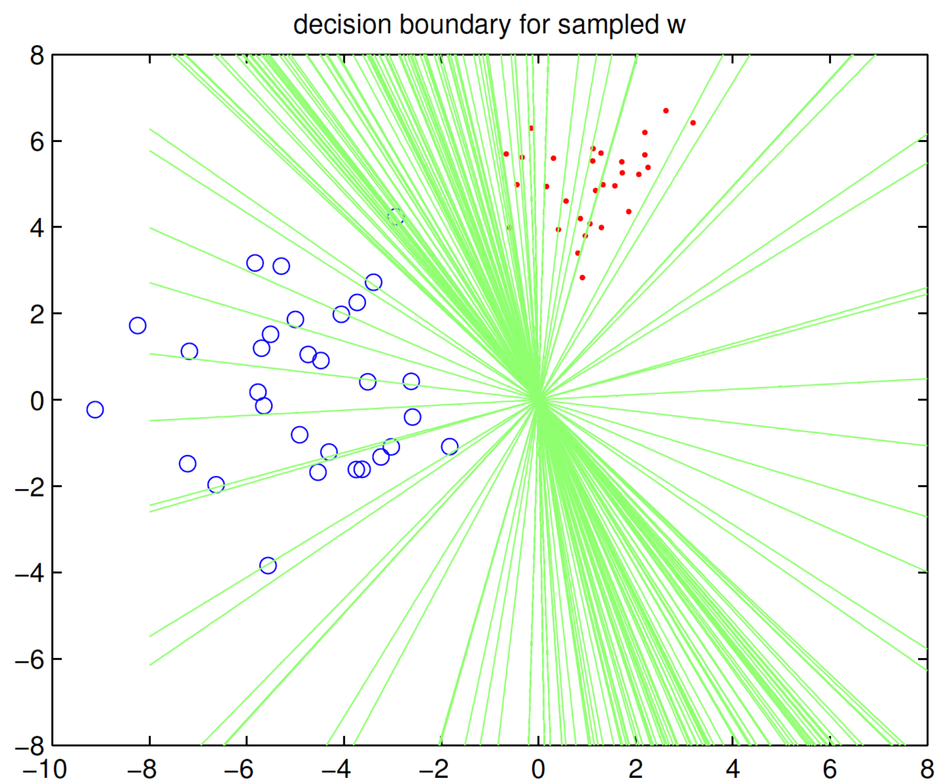
$$\hat{w} = \arg\max_{w} \; \log p(w) + \sum_i \log p(y_i \mid x_i, w)$$

True Predictive Marginal Distribution



Numerical Averaging over Monte Carlo Samples

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



Samples from Posterior

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

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

Laplace (Gaussian) Approximations

- Perform Taylor expansion of posterior *energy function*:

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

$$\begin{aligned} E(\boldsymbol{\theta}) &= -\log p(\boldsymbol{\theta}, \mathcal{D}) \\ Z &= p(\mathcal{D}) \end{aligned}$$

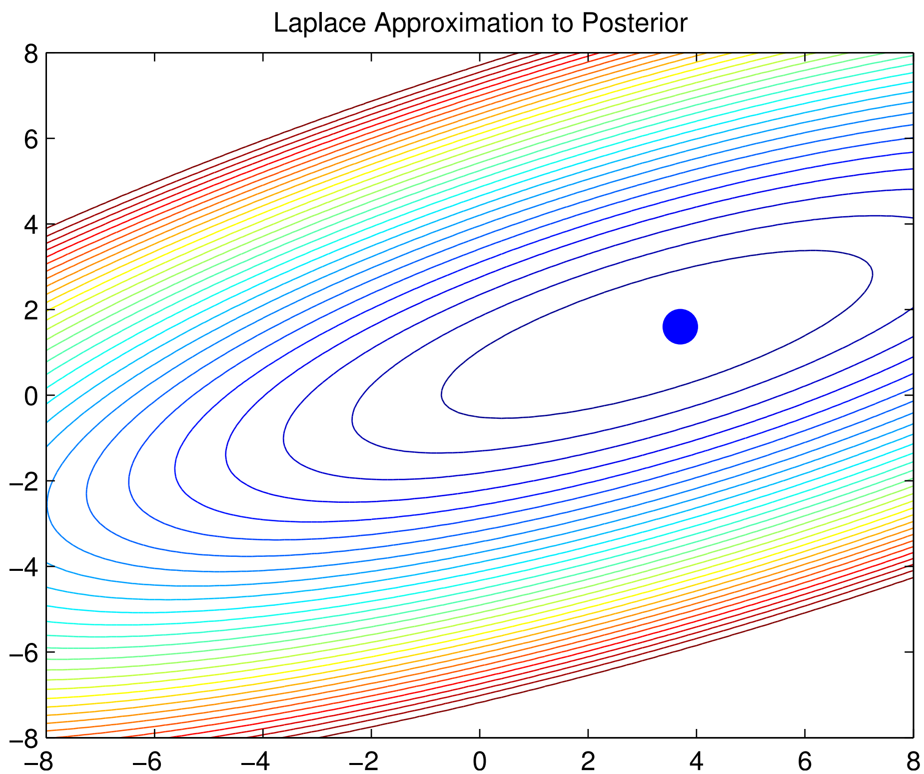
$$E(\boldsymbol{\theta}) \approx E(\boldsymbol{\theta}^*) + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbf{g} + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$$

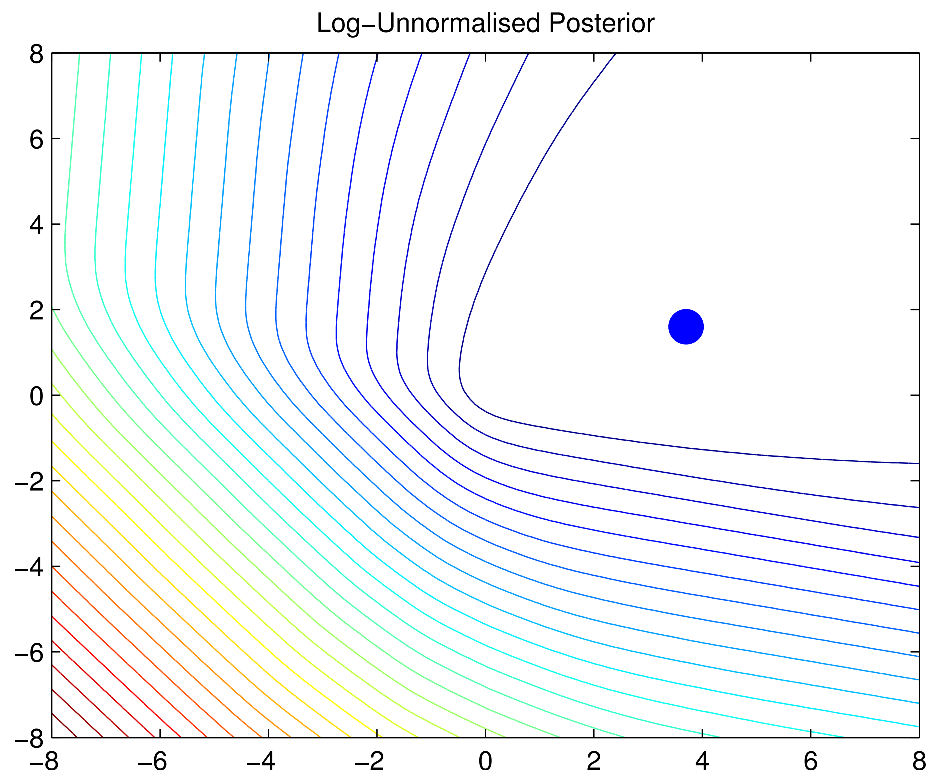
$$\mathbf{g} \triangleq \nabla E(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}^*} \qquad \mathbf{H} \triangleq \frac{\partial^2 E(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Big|_{\boldsymbol{\theta}^*}$$

- 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

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

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

Exponential Families of Distributions

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\theta}) &= \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] \\ &= h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})] \end{aligned} \qquad \begin{aligned} Z(\boldsymbol{\theta}) &= \int_{\mathcal{X}^m} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] d\mathbf{x} \\ A(\boldsymbol{\theta}) &= \log Z(\boldsymbol{\theta}) \end{aligned}$$

$\phi(x) \in \mathbb{R}^d \longrightarrow$ fixed vector of *sufficient statistics* (features), specifying the family of distributions

$\theta \in \Theta \longrightarrow$ unknown vector of *natural parameters*, determine particular distribution in this family

$Z(\theta) > 0 \longrightarrow$ normalization constant or *partition function*, ensuring this is a valid probability distribution

$h(x) > 0 \longrightarrow$ *reference measure* independent of parameters (for many models, we simply have $h(x) = 1$)

To ensure this construction is valid, we take

$$\Theta = \{\theta \in \mathbb{R}^d \mid Z(\theta) < \infty\}$$

Why the Exponential Family?

$$
\begin{aligned}
p(\mathbf{x}|\boldsymbol{\theta}) &= \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] \qquad & Z(\boldsymbol{\theta}) &= \int_{\mathcal{X}^m} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] d\mathbf{x} \\
&= h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})] & A(\boldsymbol{\theta}) &= \log Z(\boldsymbol{\theta})
\end{aligned}
$$

- Many standard distributions are in this family, and by studying exponential families, we study them all simultaneously
- Explains similarities among learning algorithms for different models, and makes it easier to derive new algorithms:
  - ML estimation takes a simple form for exponential families: *moment matching* of sufficient statistics
  - Bayesian learning is simplest for exponential families: they are the only distributions with *conjugate priors*
- They have a *maximum entropy* interpretation: Among all distributions with certain moments of interest, the exponential family is the most random (makes fewest assumptions)

Examples of Exponential Families

$$
\begin{aligned}
p(\mathbf{x}|\boldsymbol{\theta}) &= \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] \qquad & Z(\boldsymbol{\theta}) &= \int_{\mathcal{X}^m} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] d\mathbf{x} \\
&= h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})] & A(\boldsymbol{\theta}) &= \log Z(\boldsymbol{\theta})
\end{aligned}
$$

- Bernoulli and binomial (2 classes) $\phi(x) = \mathbb{I}(x = 1) = x$
- Categorical and multinomial (K classes)

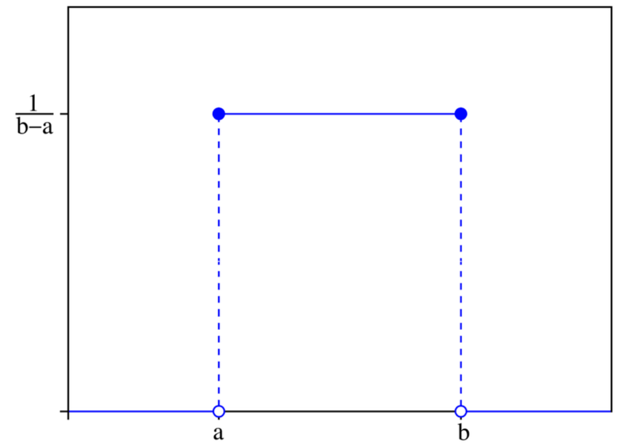
$$\phi(x) = [\mathbb{I}(x = 1), \ldots, \mathbb{I}(x = K - 1)]$$

- Scalar Gaussian $\phi(x) = [x, x^2]$
- Multivariate Gaussian $\phi(x) = [x, xx^T]$
- Poisson $h(x) = \frac{1}{x!}, \phi(x) = x$
- Dirichlet and beta
- Gamma and exponential
- …

Non-Exponential Families

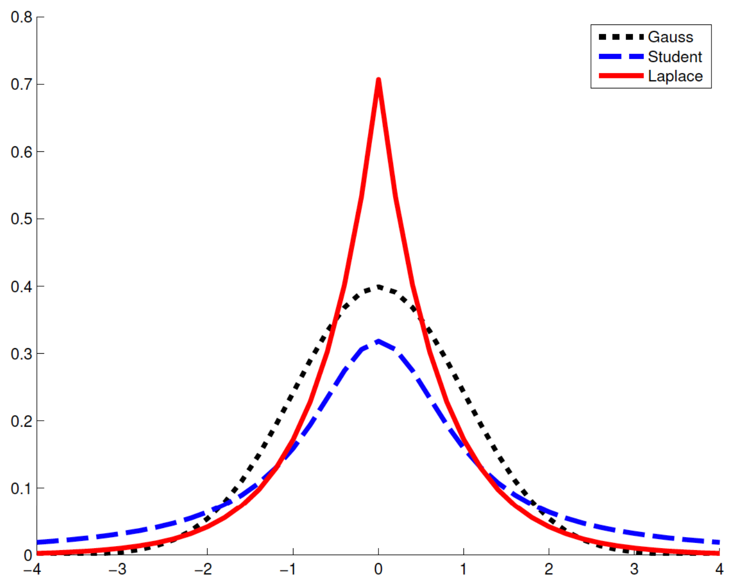
- Uniform distribution

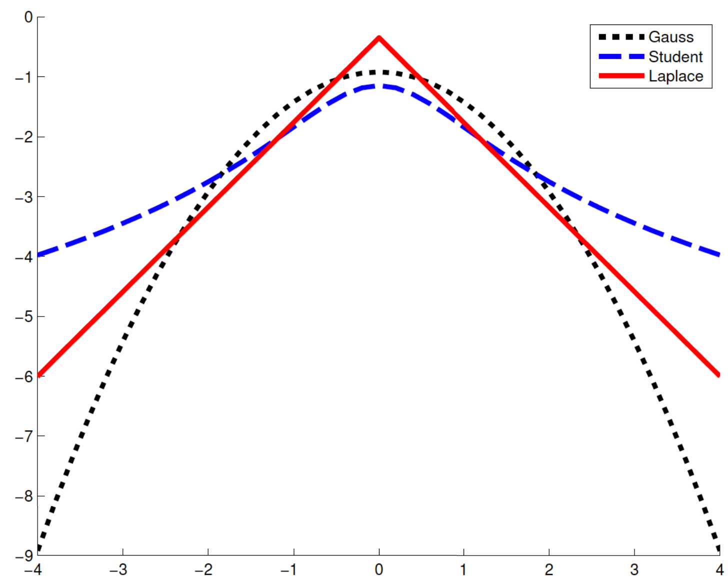
$$\text{Unif}(x \mid a, b) = \frac{1}{b-a} \mathbb{I}(a \leq x \leq b)$$



- Laplace and Student-t distributions

$$\text{Lap}(x \mid \mu, \lambda) = \frac{\lambda}{2} \exp(-\lambda |x - \mu|)$$





Log Partition Function

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\theta}) &= \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] \\ &= h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})] \end{aligned} \qquad \begin{aligned} Z(\boldsymbol{\theta}) &= \int_{\mathcal{X}^m} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] d\mathbf{x} \\ A(\boldsymbol{\theta}) &= \log Z(\boldsymbol{\theta}) \end{aligned}$$

- Derivatives of log partition function have an intuitive form:

$$\nabla_\theta A(\theta) = \mathbb{E}_\theta[\phi(x)]$$

$$\nabla^2_\theta A(\theta) = \mathrm{Cov}_\theta[\phi(x)] = \mathbb{E}_\theta[\phi(x)\phi(x)^T] - \mathbb{E}_\theta[\phi(x)]\mathbb{E}_\theta[\phi(x)]^T$$

- Important consequences for learning with exponential families:
  - Finding gradients is equivalent to finding expected sufficient statistics, or *moments*, of some current model
  - The Hessian is positive definite so $A(\theta)$ is convex
  - Learning is a convex problem: No local optima!

Learning in Exponential Families

$$
\begin{aligned}
p(\mathbf{x}|\boldsymbol{\theta}) &= \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] & Z(\boldsymbol{\theta}) &= \int_{\mathcal{X}^m} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] d\mathbf{x} \\
&= h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})] & A(\boldsymbol{\theta}) &= \log Z(\boldsymbol{\theta})
\end{aligned}
$$

- For *maximum likelihood* estimation, we find the *unique* set of parameters which satisfy:

$$
\mathbb{E}_\theta[\phi(x)] = \frac{1}{N} \sum_{i=1}^{N} \phi(x_i)
$$

- Special cases we've seen: *Categorical, Gaussian, …*
- For *Bayesian* estimation, there are convenient properties:
  - Except for a few "odd" exceptions, exponential families are the only distributions with *conjugate priors*
  - Leads to more tractable posteriors and marginal likelihoods
  - There is a simple formula for constructing these priors: *Beta-Bernoulli, Dirichlet-categorical, Gaussian-Gaussian, …*

Generalized Linear Models

- General framework for modeling non-Gaussian data with linear prediction, using exponential families:
  - Construct instance-specific natural parameters:

$$\theta_i = w^T \phi(x_i)$$

  - Observation comes from exponential family:

$$p(y_i \mid x_i, w) = \exp\left\{y_i \theta_i - A(\theta_i)\right\}$$

- Special cases: linear regression and logistic regression
- ML and MAP estimation is generally straightforward
- Many possible extensions:
  - *Multivariate responses* with more parameters (biggest difficulty is notation and indexing)
  - *Link functions* to allow more flexibility in how $(w, x_i) \rightarrow \theta_i$