ISTA 421/521
Introduction to Machine Learning

Lecture 17: Estimation: The Laplace Approximation

Clay Morrison

clayton@sista.arizona.edu

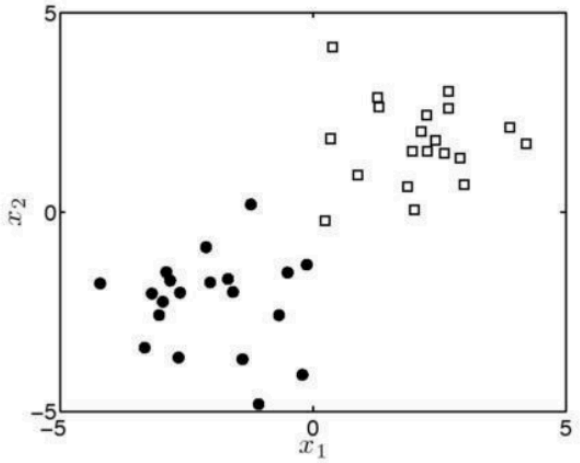
Gould-Simpson 819

Phone 621-6609

24 October 2014

Binary Classification!

- A very common type of problem
- Model 1: Binary Logistic Regression



two attributes (x_1 and x_2)

binary target, $t = \{0, 1\}$

t = 0 are dark circles
t = 1 are white squares

The Binary Logistic Likelihood

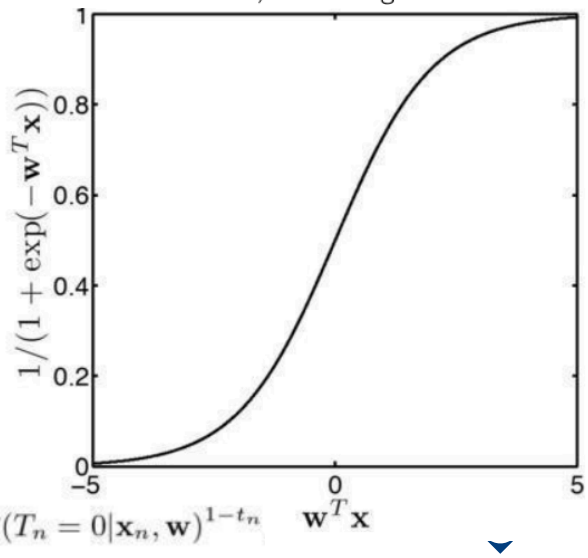
$$p(\mathbf{t}|\mathbf{X},\mathbf{w}) = \prod_{n=1}^{N} P(T_n = t_n|\mathbf{x}_n,\mathbf{w})$$

The Sigmoid function

$$P(T_n = 1|\mathbf{x}_n,\mathbf{w}) = \frac{1}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)}$$

Linear component

As $\mathbf{w}^\mathsf{T}\mathbf{x}$ increases, the value converges to 1 as it decreases, it converges to 0.



When target is 0:

$$\begin{aligned} P(T_n = 0|\mathbf{x}_n,\mathbf{w}) &= 1 - P(T_n = 1|\mathbf{x}_n,\mathbf{w}) \\ &= 1 - \frac{1}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)} \\ &= \frac{\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)}. \end{aligned}$$

Combine both into a single probability function

$$P(T_n = t_n|\mathbf{x}_n,\mathbf{w}) = P(T_n = 1|\mathbf{x}_n,\mathbf{w})^{t_n} P(T_n = 0|\mathbf{x}_n,\mathbf{w})^{1-t_n}$$

Likelihood:

$$p(\mathbf{t}|\mathbf{X},\mathbf{w}) = \prod_{n=1}^{N} \left(\frac{1}{1+\exp(-\mathbf{w}^\top\mathbf{x}_n)}\right)^{t_n} \left(\frac{\exp(-\mathbf{w}^\top\mathbf{x}_n)}{1+\exp(-\mathbf{w}^\top\mathbf{x}_n)}\right)^{1-t_n}$$

Prior:

$$p(\mathbf{w}|\sigma^2) = \mathcal{N}(0,\sigma^2\mathbf{I})$$

Once we have the Posterior…

… can predict the response (class) of new objects by taking the expectation with respect to this density:

$$P(t_{\text{new}} = 1|\mathbf{x}_{\text{new}},\mathbf{X},\mathbf{t}) = \mathbf{E}_{p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2)}\left\{\frac{1}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_{\text{new}})}\right\}$$

Problem: the posterior is not in a standard form.

The numerator is fine: just calc prior and likelihood at observations, then multiply.

It's the denominator that is the problem: can't integrate…

$$Z^{-1} = p(\mathbf{t}|\mathbf{X},\sigma^2) = \int p(\mathbf{t}|\mathbf{X},\mathbf{w})p(\mathbf{w}|\sigma^2)\, d\mathbf{w}$$

$$g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2) = p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\sigma^2)$$

$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \sigma^2) = Z^{-1} g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2)$$

$$Z^{-1} = p(\mathbf{t}|\mathbf{X}, \sigma^2) = \int p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\sigma^2)\, d\mathbf{w}$$

Our Options

1. Find the single value of **w** that corresponds to the highest value of the posterior. As $g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2)$ is proportional to the posterior, a maximum of *g* will also correspond to a maximum of the posterior. Z^{-1} is not a function of **w**. MAP
2. Approximate $p(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2)$ with some other density that we can compute analytically.
3. Sample directly from the posterior $p(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2)$, knowing only $g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2)$

Method 1: MAP point estimate

- While we cannot derive a direct analytic posterior density that we can compute, we can computer something proportional to it:

$$g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2) = p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\sigma^2)$$

- We will find the value of **w** that maximizes *g*
- This will correspond to the value at the maximum of the posterior.
- This will be the most likely value $\hat{\mathbf{w}}$ under the posterior.

Using Newton-Raphson for MAP

$$\log g(\mathbf{w}; \mathbf{X}, \mathbf{t}) = \log p(\mathbf{t}|\mathbf{X}, \mathbf{w}) + \log p(\mathbf{w}|\sigma^2)$$

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$$

$$\mathbf{w}' = \mathbf{w} - \left(\frac{\partial^2 \log g(\mathbf{w}; \mathbf{X}, \mathbf{t})}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}} \right)^{-1} \frac{\partial \log g(\mathbf{w}; \mathbf{X}, \mathbf{t})}{\partial \mathbf{w}}$$

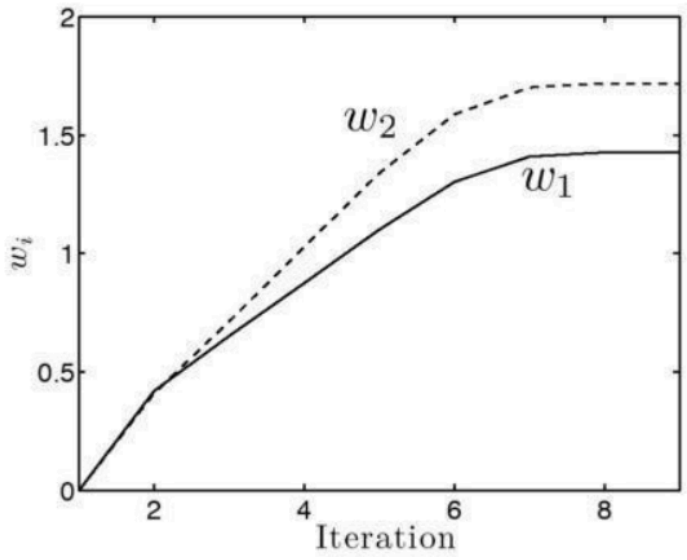
Point is maximum if Hessian is negative definite (as we did with max likelihood)

Estimating w

Starting from:

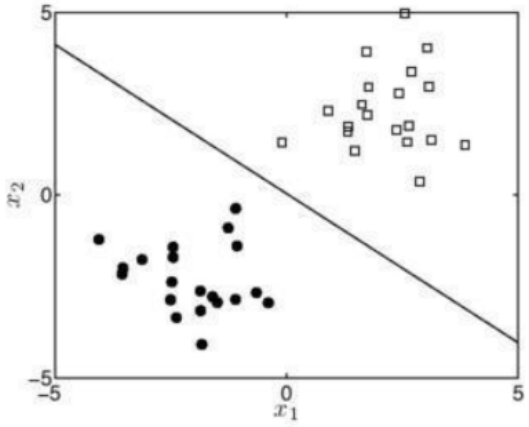
$$\mathbf{w} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

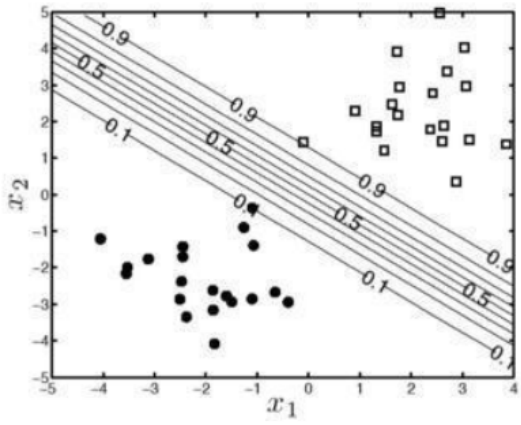
$$\sigma^2 = 10$$



Using w to compute prob of response

$$P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \widehat{\mathbf{w}}) = \frac{1}{1 + \exp(-\widehat{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_{\text{new}})}$$

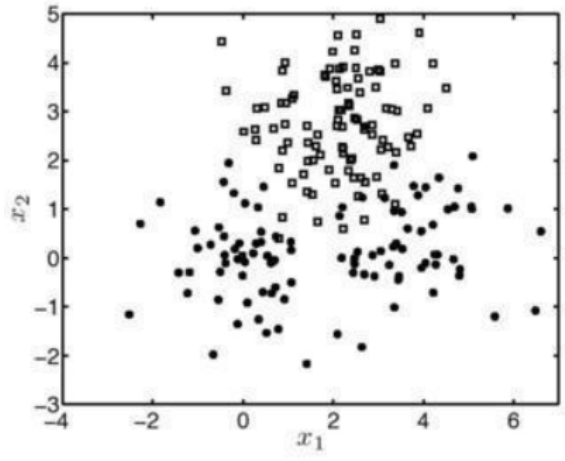


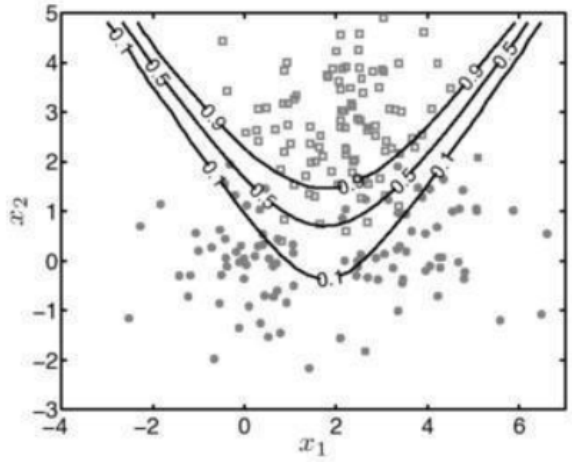


Nonlinear Decision Functions

$$\log\left(\frac{P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{w})}{P(T_{\text{new}} = 0|\mathbf{x}_{\text{new}}, \mathbf{w})}\right) = w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2$$

Find $\widehat{\mathbf{w}}$ by MAP



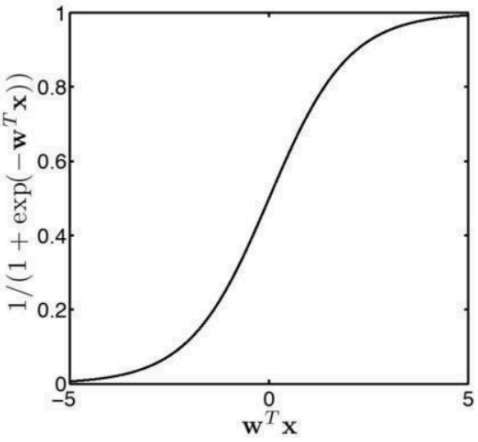


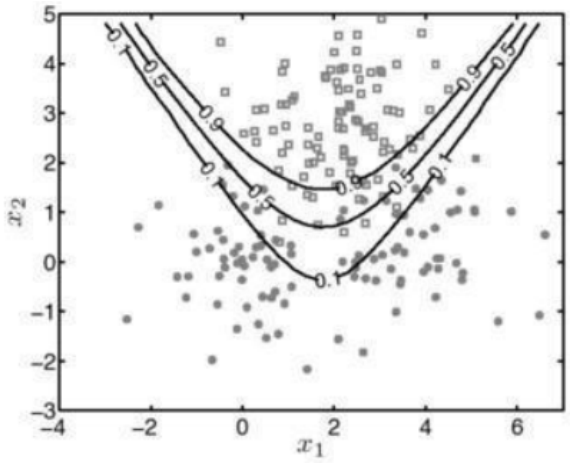
Nonlinear Decision Functions

BUT: Not a model of our uncertainty

$$\log\left(\frac{P(T_{\text{new}}=1|\mathbf{x}_{\text{new}},\mathbf{w})}{P(T_{\text{new}}=0|\mathbf{x}_{\text{new}},\mathbf{w})}\right) = w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2$$

Find $\widehat{\mathbf{w}}$ by MAP





Method 2:

The Laplace* Approximation

- **The Idea**: approximate the density of interest with a Gaussian.
- (Recall that the Gaussian is used quite often in statistics to approximate other distributions!)
- However, **keep in mind**: our predictions will only be as good as our approximation – if the true posterior is not very Gaussian, then our predictions will be easy to compute but not very useful.

*Following the note in the book: the Machine Learning community has come to refer to the method this way, but this is elsewhere referred to as **saddle-point approx.**, and in statistics, the Laplace approx. is something different.

The Gaussian

- The Gaussian density is defined by its mean and (co)variance

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\}$$

- We need to find suitable values for these parameters.
- To construct this approximation:
  - First, Suppose we knew the highest value of our posterior, $\hat{\mathbf{w}}$.
  - We can approximate the posterior using a Taylor expansion around the maximum $\hat{\mathbf{w}}$.

The Taylor Expansion

- A way of approximating a function 'near' some value $\hat{\mathbf{w}}$. (based on characteristics of *f* at $\hat{\mathbf{w}}$)
- The approximation will diverge from the true function as we move away from $\hat{\mathbf{w}}$.

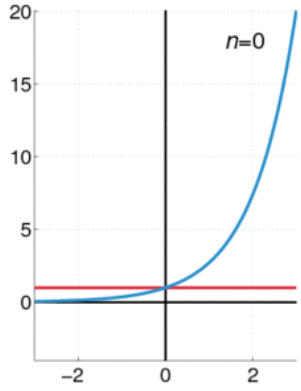
$$f(\mathbf{w}) = f(\hat{\mathbf{w}}) + \frac{f'(\hat{\mathbf{w}})}{1!}(\mathbf{w}-\hat{\mathbf{w}}) + \frac{f''(\hat{\mathbf{w}})}{2!}(\mathbf{w}-\hat{\mathbf{w}})^2 + \frac{f'''(\hat{\mathbf{w}})}{3!}(\mathbf{w}-\hat{\mathbf{w}})^3 + \dots$$

$$= \sum_{n=0}^{\infty} \frac{(\mathbf{w}-\hat{\mathbf{w}})^n}{n!} \left.\frac{\partial^n f(\mathbf{w})}{\partial \mathbf{w}^n}\right|_{\hat{\mathbf{w}}}$$

$$\sin(x) \approx x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!}.$$

$$\exp(w) = \exp(\hat{w}) + \frac{w}{1!}\exp(\hat{w}) + \frac{w^2}{2!}\exp(\hat{w}) + \dots$$

When $\hat{w} = 0$: $\exp(w) = 1 + \frac{w}{1!} + \frac{w^2}{2!} + \frac{w^3}{3!} + \dots$



$$p(\mathbf{t}|\mathbf{X},\mathbf{w}) = \prod_{n=1}^{N} \left(\frac{1}{1+\exp(-\mathbf{w}^\top \mathbf{x}_n)}\right)^{t_n} \left(\frac{\exp(-\mathbf{w}^\top \mathbf{x}_n)}{1+\exp(-\mathbf{w}^\top \mathbf{x}_n)}\right)^{1-t_n} \qquad p(\mathbf{w}|\sigma^2) = \mathcal{N}(0, \sigma^2 \mathbf{I})$$

Approximating *g* using the Taylor Expansion

$$g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2) = p(\mathbf{t}|\mathbf{X},\mathbf{w}) p(\mathbf{w}|\sigma^2)$$

$$\log g(\mathbf{w}; \mathbf{X}, \mathbf{t}) = \log p(\mathbf{t}|\mathbf{X},\mathbf{w}) + \log p(\mathbf{w}|\sigma^2)$$

$$\sum_{n=0}^{\infty} \frac{(\mathbf{w}-\hat{\mathbf{w}})^n}{n!} \left.\frac{\partial^n f(\mathbf{w})}{\partial \mathbf{w}^n}\right|_{\hat{\mathbf{w}}}$$

$$\log g(w; \mathbf{X}, \mathbf{t}, \sigma^2) \approx \log g(\hat{w}; \mathbf{X}, \mathbf{t}, \sigma^2) + \left.\frac{\partial \log g(w; \mathbf{X}, \mathbf{t}, \sigma^2)}{\partial w}\right|_{\hat{w}} \frac{(w-\hat{w})}{1!} + \left.\frac{\partial^2 \log g(w; \mathbf{X}, \mathbf{t}, \sigma^2)}{\partial w^2}\right|_{\hat{w}} \frac{(w-\hat{w})^2}{2!} + \ldots$$

Evaluate this at $\hat{w}$ equal to the peak!

At this point, the first derivative is 0, by definition, so can eliminate 1st-order term
Also eliminate all terms after 2nd order (they make the approximation better,
but don't achieve what we're trying to do mathematically…)

$$\log g(w; \mathbf{X}, \mathbf{t}, \sigma^2) \approx \log g(\hat{w}; \mathbf{X}, \mathbf{t}, \sigma^2) - \frac{v}{2}(w-\hat{w})^2$$

$$v = -\left.\frac{\partial^2 \log g(w; \mathbf{X}, \mathbf{t}, \sigma^2)}{\partial w^2}\right|_{\hat{w}}$$

$$p(\mathbf{t}|\mathbf{X},\mathbf{w}) = \prod_{n=1}^{N} \left(\frac{1}{1+\exp(-\mathbf{w}^\top \mathbf{x}_n)}\right)^{t_n} \left(\frac{\exp(-\mathbf{w}^\top \mathbf{x}_n)}{1+\exp(-\mathbf{w}^\top \mathbf{x}_n)}\right)^{1-t_n} \qquad p(\mathbf{w}|\sigma^2) = \mathcal{N}(0, \sigma^2 \mathbf{I})$$

Approximating *g* using the Taylor Expansion

$$g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2) = p(\mathbf{t}|\mathbf{X},\mathbf{w}) p(\mathbf{w}|\sigma^2)$$

$$\log g(\mathbf{w}; \mathbf{X}, \mathbf{t}) = \log p(\mathbf{t}|\mathbf{X},\mathbf{w}) + \log p(\mathbf{w}|\sigma^2)$$

$$\sum_{n=0}^{\infty} \frac{(\mathbf{w}-\hat{\mathbf{w}})^n}{n!} \left.\frac{\partial^n f(\mathbf{w})}{\partial \mathbf{w}^n}\right|_{\hat{\mathbf{w}}}$$

Recall, the univariate Gaussian:

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(w-\mu)^2\right\}$$

The log of the univ. G
(K is the normalizing constant):

$$\log(K) - \frac{1}{2\sigma^2}(w-\mu)^2$$

$$\sigma^2 = 1/v \qquad \mu = \hat{w}$$

This is the Laplace approximation!
We approximate the posterior with
a Gaussian that has its
mean at the posterior **mode** ($\hat{w}$),
variance inversely proportional to
the curvature of the posterior (*g''*)
at its mode.

Multivariate version:

$$\boldsymbol{\mu} = \hat{\mathbf{w}}, \quad \boldsymbol{\Sigma}^{-1} = -\left.\left(\frac{\partial^2 \log g(\mathbf{w}; \mathbf{X}, \mathbf{t})}{\partial \mathbf{w} \partial \mathbf{w}^\top}\right)\right|_{\hat{\mathbf{w}}}$$

$$\log g(w; \mathbf{X}, \mathbf{t}, \sigma^2) \approx \log g(\hat{w}; \mathbf{X}, \mathbf{t}, \sigma^2) - \frac{v}{2}(w-\hat{w})^2$$

$$v = -\left.\frac{\partial^2 \log g(w; \mathbf{X}, \mathbf{t}, \sigma^2)}{\partial w^2}\right|_{\hat{w}}$$

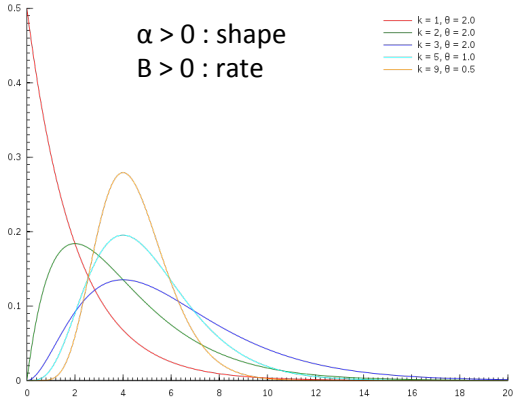
The Gamma Distribution

We will use this as an example function that we'll estimate using Laplace estimation. The Gamma dist. is very flexible, can be made to look very Gaussian (e.g., nearly symmetric) but also skewed. So we can observe what it looks like when Laplace estimation is good as well as what it looks like when the estimation is not so good.

$$p(y|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp\{-\beta y\}$$

α=k, B= 1/θ

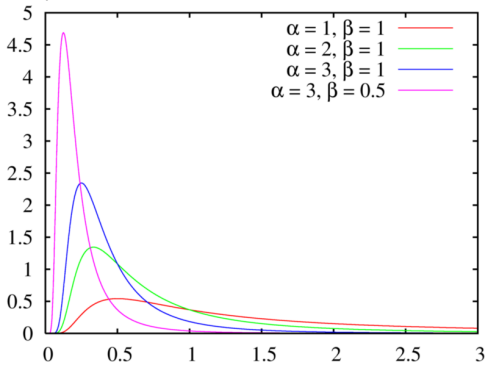
α > 0 : shape
B > 0 : rate



Side note…
The Inverse Gamma:

$$f(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right)$$

Is conjugate to, and therefore popular to use in modeling the "scale" (i.e., variance, σ^2) parameter of Gaussian



Laplace Approximation Example 1

- We know the true density of the gamma:

$$p(y|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp\{-\beta y\}$$

- We'll use this example so we can tell how good *or bad* the approximation is.
- Analytic expression for gamma mode: $\widehat{y} = \frac{\alpha-1}{\beta}$
- To find the variance, σ^2, take 2nd derivative of log $p(y|\alpha,\beta)$:

$$\log p(y|\alpha,\beta) = \alpha \log \beta - \log(\Gamma(\alpha)) + (\alpha-1)\log y - \beta y$$

$$\frac{\partial \log p(y|\alpha,\beta)}{\partial y} = \frac{\alpha-1}{y} - \beta$$

$$\frac{\partial^2 \log p(y|\alpha,\beta)}{\partial y^2} = -\frac{\alpha-1}{y^2}.$$

σ^2 is the negative inverse of the second derivative, evaluated at $y = \widehat{y}$

$$\sigma^2 = \frac{\widehat{y}^2}{\alpha-1} = \frac{\alpha-1}{\beta^2}$$

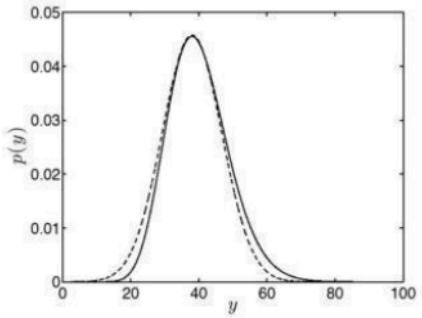
Laplace Approximation Example 1

- We know the true density of the gamma:

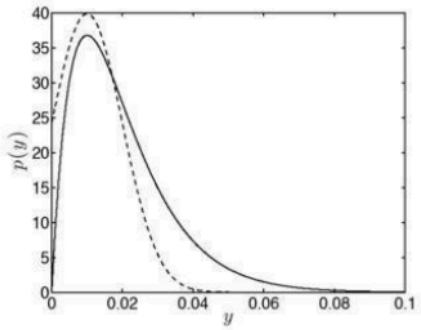
$$p(y|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1}\exp\{-\beta y\}$$

$$\widehat{y} = \frac{\alpha-1}{\beta}$$

$$\sigma^2 = \frac{\widehat{y}^2}{\alpha-1} = \frac{\alpha-1}{\beta^2}$$



(a) $p(y|\alpha,\beta)$ (solid line) and approximating Gaussian (dashed line) for $\alpha = 20,\ \beta = 0.5$



(b) $p(y|\alpha,\beta)$ (solid line) and approximating Gaussian (dashed line) for $\alpha = 2,\ \beta = 100$

Laplace Approximation Example 2

- Return to the binary response model
- Recall that we did calculate the Hessian for Newton-Raphson

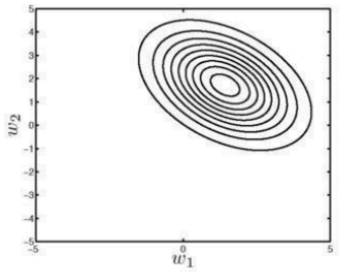
$$\frac{\partial^2 \log g(\mathbf{w};\mathbf{X},\mathbf{t})}{\partial \mathbf{w}\partial \mathbf{w}^\mathsf{T}} = -\frac{1}{\sigma^2}\mathbf{I} - \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^\mathsf{T} P_n(1-P_n).$$

- And we used Newton-Raphson to estimate the mode, $\hat{\mathbf{w}}$

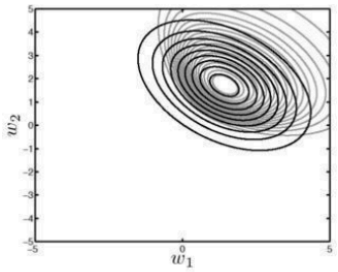
$$\sigma^2 = 1/v \qquad \mu = \widehat{w}$$

$$\boldsymbol{\mu} = \widehat{\mathbf{w}}, \quad \boldsymbol{\Sigma}^{-1} = -\left(\frac{\partial^2 \log g(\mathbf{w};\mathbf{X},\mathbf{t})}{\partial \mathbf{w}\partial \mathbf{w}^\mathsf{T}}\right)\Big|_{\widehat{\mathbf{w}}}$$

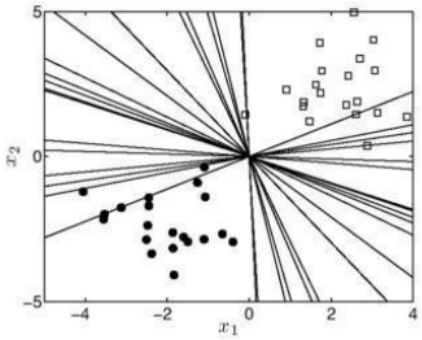
Laplace Approximation Example 2



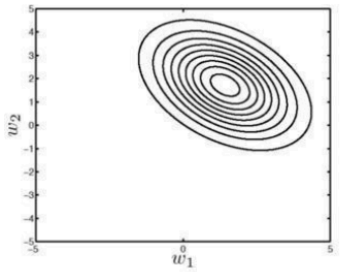
(a) Laplace approximation to the posterior



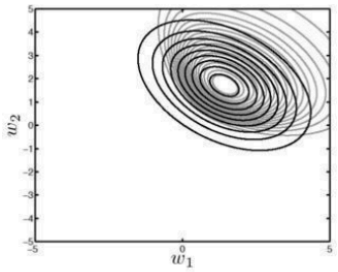
(b) Laplace approximation to the posterior and the true unnormalised posterior (lighter lines)



Laplace Approximation Example 2



(a) Laplace approximation to the posterior



$$\mathbf{w}_s \leftarrow \mathcal{N}(\mu, \boldsymbol{\Sigma})$$

$$P(T_{\text{new}} = 1 | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}, \sigma^2) = \frac{1}{N_s} \sum_{s=1}^{N_s} \frac{1}{1 + \exp(-\mathbf{w}_s^{\mathsf{T}} \mathbf{x}_{\text{new}})}$$