ISTA 421 + INFO 521
Introduction to Machine Learning

Lecture 17: Estimation II: The Laplace Approximation

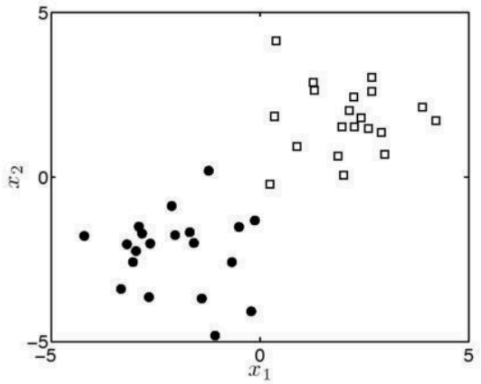
Clay Morrison
claytonm@email.arizona.edu
Harvill 437A
Phone 621-6609

25 October 2017

1

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

2

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

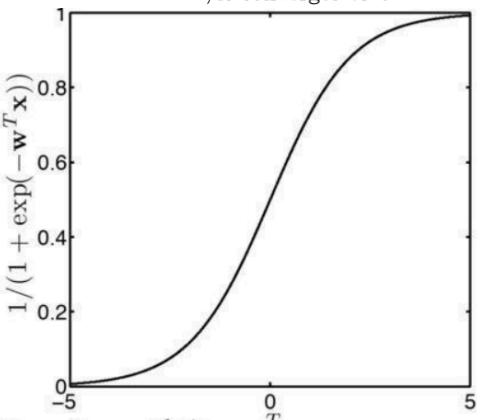
$$P(T_n = 1|\mathbf{x}_n,\mathbf{w}) = \frac{1}{1+\exp(-\mathbf{w}^\top\mathbf{x}_n)}$$

Linear component

When target is 0:

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

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



Combine both into a single probability function

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

In Sum: Problems Optimizing Logistic Regression
Our analytic optimization tools (to date) fail

- **Problem 1**: cannot directly compute max likelihood (or MAP) – cannot isolate **w**

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

- **Problem 2**: cannot compute full Bayes because cannot integrate Bayes denominator (marginal likelihood)

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

Numerator of Bayes Theorem

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

marginal Likelihood (denominator)

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

The Posterior

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

Our Options

1. Find the single value of **w** that corresponds to the highest value of the likelihood or posterior. As *g* 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**
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 $g(\mathbf{w};\mathbf{X},\mathbf{t},\sigma^2)$

5

Method 1: ML or 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* (or that maximizes just the likelihood)
- 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.

6

Using Newton-Raphson for MAP

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

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

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

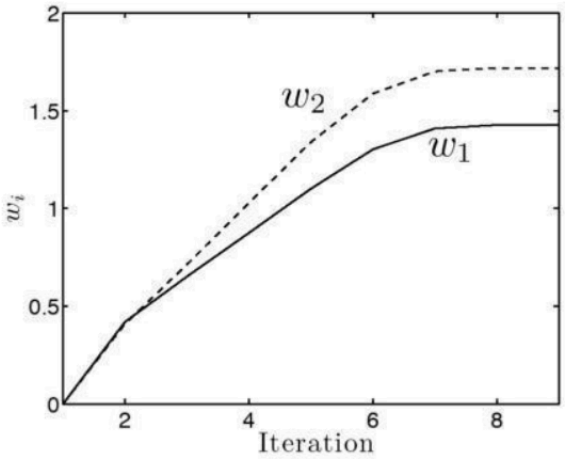
Point is maximum if Hessian is negative definite (as we did with max likelihood)

Estimating w

Starting from:

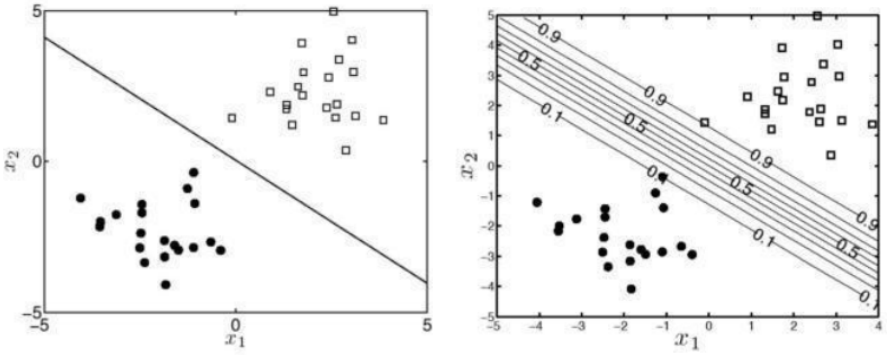
$$\mathbf{w} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\sigma^2 = 10$$



Using w to compute prob of response

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

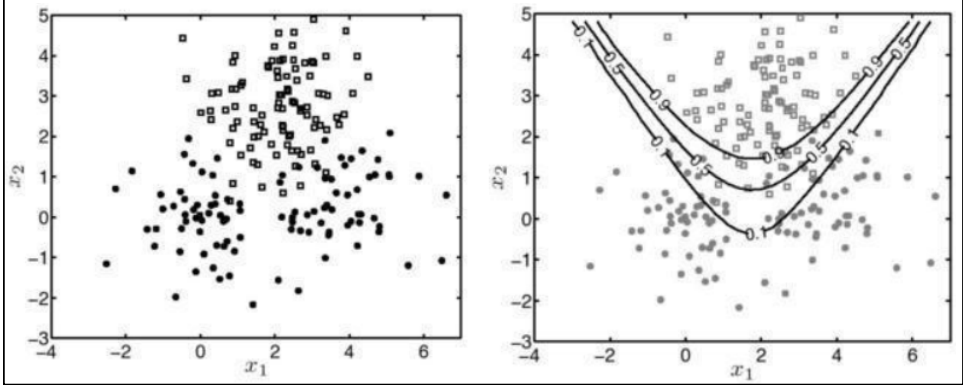


9

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



BUT: Not a model of our uncertainty

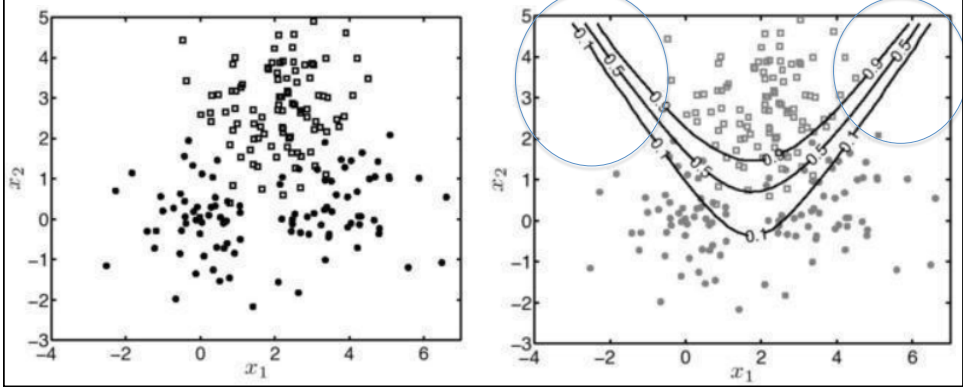
We are being Bayesian about **w**, so put prior on expected values of **w**, and that affects what **w** values are considered good

However, this is NOT a model of our posterior uncertainty in **w**: we just selected MAP estimate of **w**.

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

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

Shouldn't we be *more* uncertain where we don't have data?



BUT: Not a model of our uncertainty

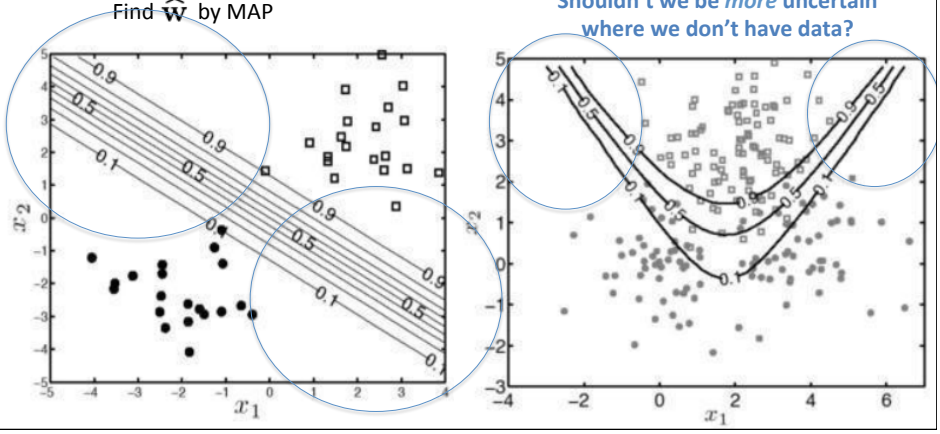
We are being Bayesian about **w**, so put prior on expected values of **w**, and that affects what **w** values are considered good

However, this is NOT a model of our posterior uncertainty in **w**: we just selected MAP estimate of **w**.

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

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

Shouldn't we be *more* uncertain where we don't have data?



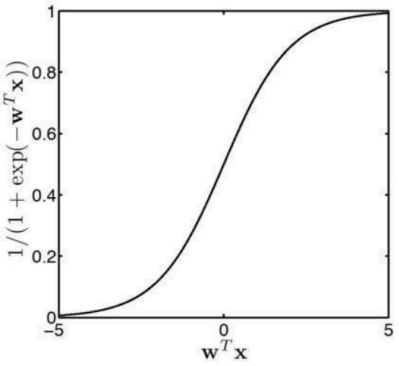
BUT: Not a model of our uncertainty

We are being Bayesian about **w**, so put prior on expected values of **w**, and that affects what **w** values are considered good

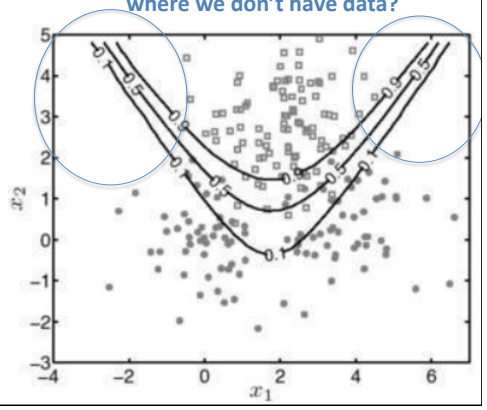
However, this is NOT a model of our posterior uncertainty in **w**: we just selected MAP estimate of **w**.

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

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



Shouldn't we be *more* uncertain where we don't have data?



Numerator of Bayes Theorem

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

marginal Likelihood (denominator)

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

The Posterior

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

Our Options

1. Find the single value of **w** that corresponds to the highest value of the likelihood or posterior. As *g* 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**
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 2:
The Laplace* Approximation

- **The Idea**: approximate the density of interest (some function $f(\hat{\mathbf{w}})$) with a *Gaussian*.
- This can approximate the posterior uncertainty.
- (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 approximation**, and in statistics, "Laplace smoothing" is something different. (additive smoothing)

The Gaussian

- The Gaussian density is defined by its mean $\boldsymbol{\mu}$ and (co)variance $\boldsymbol{\Sigma}$

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

The Gaussian

- The Gaussian density is defined by its mean $\boldsymbol{\mu}$ and (co)variance $\boldsymbol{\Sigma}$

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

- **Goal:** find suitable values for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ derived from the to-be-modeled distribution $f(\hat{\mathbf{w}})$.

17

The Gaussian

- The Gaussian density is defined by its mean $\boldsymbol{\mu}$ and (co)variance $\boldsymbol{\Sigma}$

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

- **Goal:** find suitable values for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ derived from the to-be-modeled distribution $f(\hat{\mathbf{w}})$.
- To construct this approximation:
  - First, Suppose we knew the highest value (mode) of our posterior f at $\hat{\mathbf{w}}$.

18

The Gaussian

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

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

- **Goal:** find suitable values for **μ** and **Σ** derived from the to-be-modeled distribution *f*(**ŵ**).
- To construct this approximation:
  - First, Suppose we knew the highest value (mode) of our posterior *f* at **ŵ**.
  - Second, we can approximate the posterior using a **Taylor expansion** around the maximum **ŵ**.

The Taylor Expansion

- A way of approximating a function 'near' some value **ŵ**. (based on characteristics of *f* at **ŵ**)
- The approximation will diverge from the true function as we move away from **ŵ**.

ŵ is the location we're approximating from
w is new value we evaluate *f* at, from the perspective of **ŵ**

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

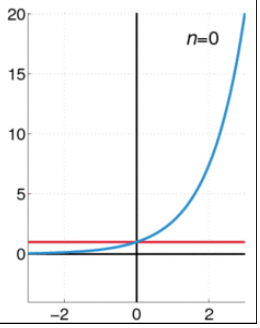
$$= \sum_{n=0}^{\infty} \frac{(\mathbf{w}-\hat{\mathbf{w}})^n}{n!} \frac{\partial^n f(\mathbf{w})}{\partial \mathbf{w}^n}\bigg|_{\hat{\mathbf{w}}}$$

$$\sin(w) \approx w - \frac{w^3}{3!} + \frac{w^5}{5!} - \frac{w^7}{7!}, \quad \text{where } \hat{w} = 0$$

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

When **ŵ** = 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

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

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

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

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

Evaluate this at $\hat{\mathbf{w}}$ equal to the peak of *g*!

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, as we'll see in a moment…)

$$\log g(w; \mathbf{X}, \mathbf{t}, \sigma^2) \approx \log g(\widehat{w}; \mathbf{X}, \mathbf{t}, \sigma^2) - \frac{v}{2}(w-\widehat{w})^2 \qquad v = -\frac{\partial^2 \log g(w; \mathbf{X}, \mathbf{t}, \sigma^2)}{\partial w^2}\bigg|_{\widehat{w}}$$

21

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

Approximating *g* using the Taylor Expansion

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

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

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

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 univariate Gauss.
(*K* is the normalizing constant):

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

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

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

Multivariate version:

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

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

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

22

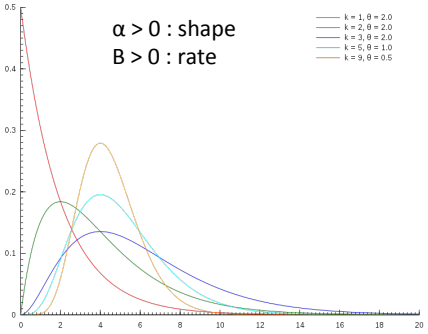
The Gamma Distribution

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 can be 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

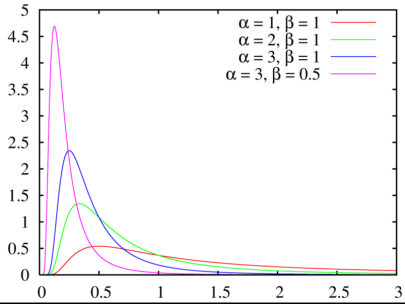
α > 0 : shape
B > 0 : rate



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 of, 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: $\hat{y} = \frac{\alpha-1}{\beta}$
- 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 $y = \hat{y}$

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

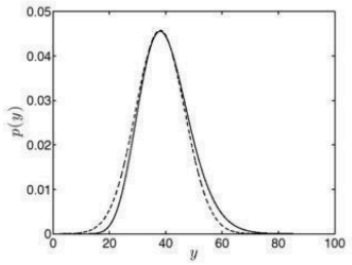
Laplace Approximation Example 1

- We know the true density of the gamma:

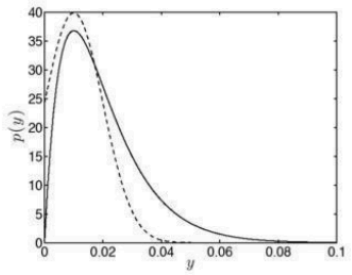
$$p(y|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1}\exp\{-\beta y\}$$

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

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



(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 logistic regression model
- Recall that we already calculated the Hessian for Newton-Raphson (last lecture)

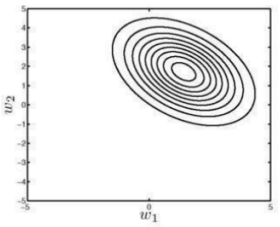
$$\frac{\partial^2 \log g(\mathbf{w};\mathbf{X},\mathbf{t})}{\partial \mathbf{w}\partial \mathbf{w}^{\mathsf{T}}} = -\frac{1}{\sigma^2}\mathbf{I} - \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^{\mathsf{T}} P_n(1-P_n).$$

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

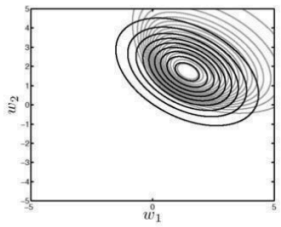
$$\sigma^2 = 1/v \qquad \mu = \widehat{w}$$

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

Laplace Approximation Example 2

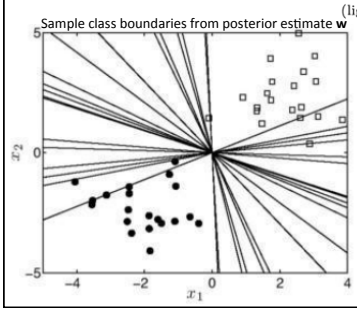


(a) Laplace approximation to the posterior

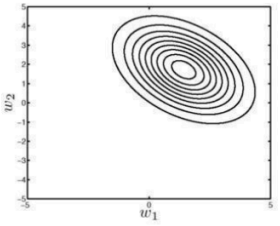


(b) Laplace approximation to the posterior and the true unnormalised posterior (lighter lines)

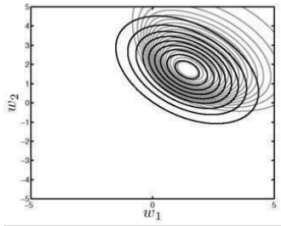
Sample class boundaries from posterior estimate **w**



Laplace Approximation Example 2



(a) Laplace approximation to the posterior



We can also use samples of **w** to estimate class probability at each point: $\mathbf{w}_s \leftarrow \mathcal{N}(\mu, \mathbf{\Sigma})$

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

Sample class boundaries from posterior estimate **w**