3F8: Inference

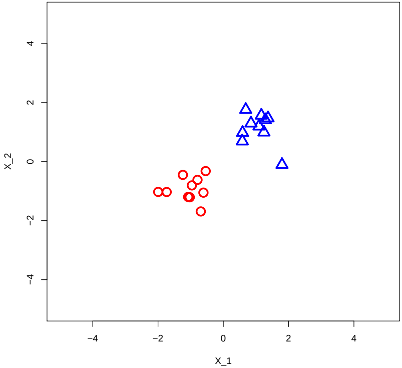
Bayesian Linear Classification

José Miguel Hernández–Lobato and Richard E. Turner
Department of Engineering
University of Cambridge

Lent Term

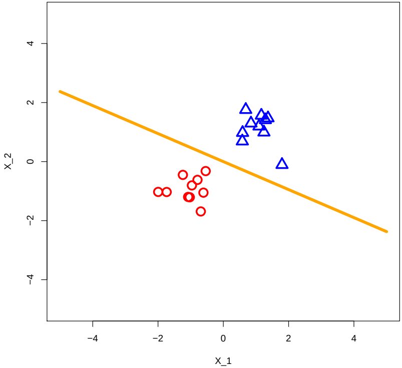
In high dimensions (many basis functions) the data will be **linearly separable**.

Many **w** fit the data equally well. This can lead to **overfitting problems**.



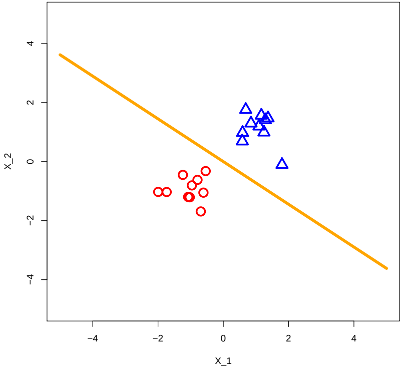
In high dimensions (many basis functions) the data will be **linearly separable**.

Many **w** fit the data equally well. This can lead to **overfitting problems**.



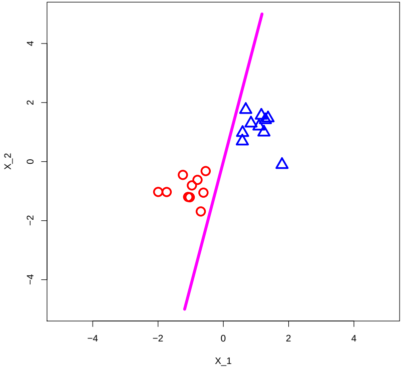
In high dimensions (many basis functions) the data will be **linearly separable**.

Many **w** fit the data equally well. This can lead to **overfitting problems**.



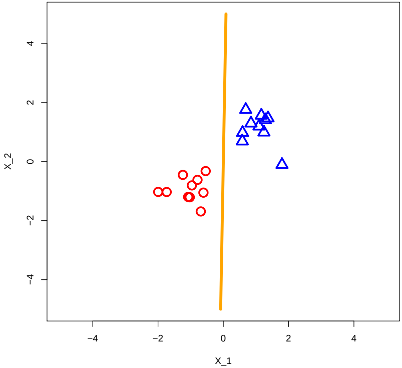
In high dimensions (many basis functions) the data will be **linearly separable**.

Many **w** fit the data equally well. This can lead to **overfitting problems**.



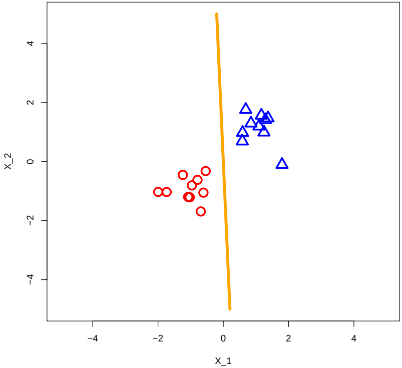
In high dimensions (many basis functions) the data will be **linearly separable**.

Many **w** fit the data equally well. This can lead to **overfitting problems**.



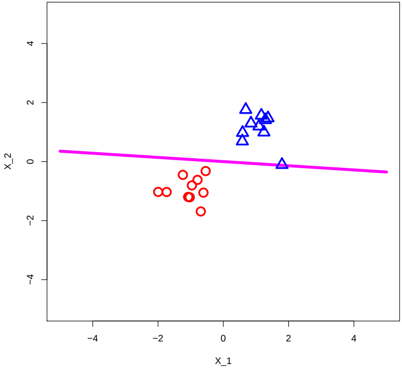
In high dimensions (many basis functions) the data will be **linearly separable**.

Many **w** fit the data equally well. This can lead to **overfitting problems**.



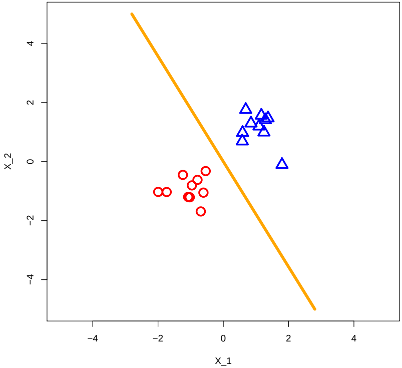
In high dimensions (many basis functions) the data will be **linearly separable**.

Many **w** fit the data equally well. This can lead to **overfitting problems**.



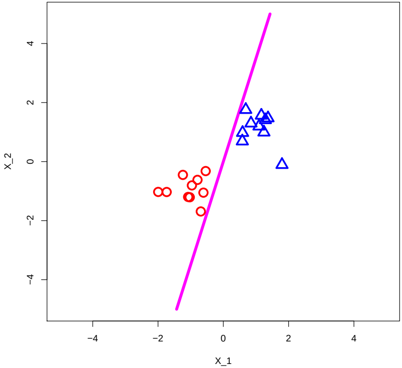
In high dimensions (many basis functions) the data will be **linearly separable**.

Many **w** fit the data equally well. This can lead to **overfitting problems**.



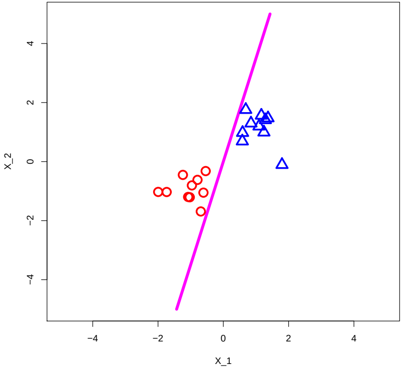
In high dimensions (many basis functions) the data will be **linearly separable**.

Many **w** fit the data equally well. This can lead to **overfitting problems**.



In high dimensions (many basis functions) the data will be **linearly separable**.

Many $\mathbf{w}$ fit the data equally well. This can lead to **overfitting problems**.



Solution: Bayesian inference.

The prior on $\mathbf{w}$ is

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \lambda^{-1}\mathbf{I})\,.$$

The posterior on $\mathbf{w}$ is

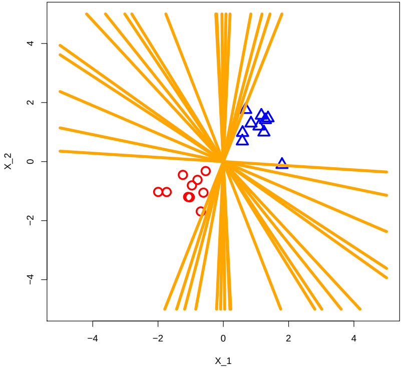
$$p(\mathbf{w}|\mathbf{y}, \widetilde{\mathbf{X}}) \propto \left[\prod_{n_1}^{N} \sigma(y_n \mathbf{w}^{\mathrm{T}} \widetilde{\mathbf{x}}_n)\right] p(\mathbf{w})\,,$$

The predictive distribution is

$$p(y_\star|\widetilde{\mathbf{x}}_\star, \mathbf{y}, \widetilde{\mathbf{X}}) = \int \sigma(y_\star \mathbf{w}^{\mathrm{T}} \widetilde{\mathbf{x}}_\star) p(\mathbf{w}|\mathbf{y}, \widetilde{\mathbf{X}})\, d\mathbf{w}\,,$$

In high dimensions (many basis functions) the data will be **linearly separable**.

Many $\mathbf{w}$ fit the data equally well. This can lead to **overfitting problems**.



Solution: Bayesian inference.

The prior on $\mathbf{w}$ is

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \lambda^{-1}\mathbf{I})\,.$$

The posterior on $\mathbf{w}$ is

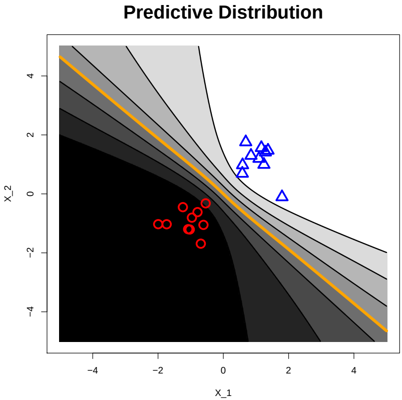
$$p(\mathbf{w}|\mathbf{y}, \widetilde{\mathbf{X}}) \propto \left[\prod_{n_1}^{N} \sigma(y_n \mathbf{w}^\mathrm{T} \widetilde{\mathbf{x}}_n)\right] p(\mathbf{w})\,,$$

The predictive distribution is

$$p(y_\star|\widetilde{\mathbf{x}}_\star, \mathbf{y}, \widetilde{\mathbf{X}}) = \int \sigma(y_\star \mathbf{w}^\mathrm{T} \widetilde{\mathbf{x}}_\star) p(\mathbf{w}|\mathbf{y}, \widetilde{\mathbf{X}})\, d\mathbf{w}\,,$$

In high dimensions (many basis functions) the data will be **linearly separable**.

Many $\mathbf{w}$ fit the data equally well. This can lead to **overfitting problems**.



Solution: Bayesian inference.

The prior on $\mathbf{w}$ is

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \lambda^{-1}\mathbf{I})\,.$$

The posterior on $\mathbf{w}$ is

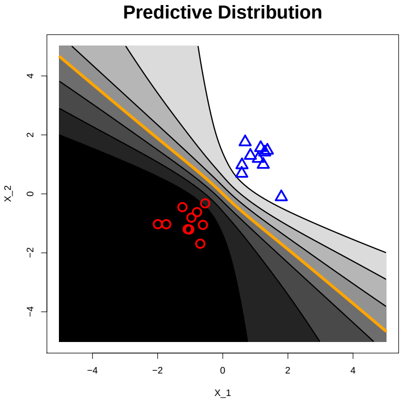
$$p(\mathbf{w}|\mathbf{y}, \widetilde{\mathbf{X}}) \propto \left[\prod_{n_1}^{N} \sigma(y_n \mathbf{w}^{\mathrm{T}} \widetilde{\mathbf{x}}_n)\right] p(\mathbf{w})\,,$$

The predictive distribution is

$$p(y_\star|\widetilde{\mathbf{x}}_\star, \mathbf{y}, \widetilde{\mathbf{X}}) = \int \sigma(y_\star \mathbf{w}^{\mathrm{T}} \widetilde{\mathbf{x}}_\star) p(\mathbf{w}|\mathbf{y}, \widetilde{\mathbf{X}})\, d\mathbf{w}\,,$$

In high dimensions (many basis functions) the data will be **linearly separable**.

Many $\mathbf{w}$ fit the data equally well. This can lead to **overfitting problems**.



Difference w.r.t. logistic regression: **higher uncertainty** far from data.

Solution: Bayesian inference.

The prior on $\mathbf{w}$ is

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \lambda^{-1}\mathbf{I})\,.$$

The posterior on $\mathbf{w}$ is

$$p(\mathbf{w}|\mathbf{y}, \widetilde{\mathbf{X}}) \propto \left[\prod_{n_1}^{N} \sigma(y_n \mathbf{w}^{\mathrm{T}} \widetilde{\mathbf{x}}_n)\right] p(\mathbf{w})\,,$$

The predictive distribution is

$$p(y_\star|\widetilde{\mathbf{x}}_\star, \mathbf{y}, \widetilde{\mathbf{X}}) = \int \sigma(y_\star \mathbf{w}^{\mathrm{T}} \widetilde{\mathbf{x}}_\star) p(\mathbf{w}|\mathbf{y}, \widetilde{\mathbf{X}})\, d\mathbf{w}\,,$$

Problem:

Integration with respect to $p(\mathbf{w}, \mathbf{y}|\widetilde{\mathbf{X}}) = \left[\prod_{n_1}^{N} \sigma(y_n \mathbf{w}^\mathsf{T} \widetilde{\mathbf{x}}_n)\right] p(\mathbf{w})$ over $\mathbf{w}$ is intractable (**how do you know?**), but this is required to compute

- The normalization constant of the posterior distribution $p(\mathbf{w}|\mathbf{y}, \widetilde{\mathbf{X}})$.
- The predictive distribution $p(y_\star|\widetilde{\mathbf{x}}_\star, \mathbf{y}, \widetilde{\mathbf{X}})$.

Problem:

Integration with respect to $p(\mathbf{w}, \mathbf{y}|\widetilde{\mathbf{X}}) = \left[\prod_{n_1}^{N} \sigma(y_n \mathbf{w}^\mathrm{T} \widetilde{\mathbf{x}}_n)\right] p(\mathbf{w})$ over $\mathbf{w}$ is intractable (**how do you know?**), but this is required to compute

- The normalization constant of the posterior distribution $p(\mathbf{w}|\mathbf{y}, \widetilde{\mathbf{X}})$.
- The predictive distribution $p(y_\star|\widetilde{\mathbf{x}}_\star, \mathbf{y}, \widetilde{\mathbf{X}})$.

Solution:

Use **approximate Bayesian inference**. Different approaches are possible:

1. Draw a sequence of asymptotically unbiased samples from $p(\mathbf{w}|\mathbf{y}, \widetilde{\mathbf{X}})$ (Monte Carlo methods).
2. Approximate $p(\mathbf{w}|\mathbf{y}, \widetilde{\mathbf{X}})$ with a simpler distribution (e.g. a Gaussian) (Deterministic approximate inference).

What are the trade-offs between these options?

Problem:

Integration with respect to $p(\mathbf{w}, \mathbf{y}|\widetilde{\mathbf{X}}) = \left[\prod_{n_1}^{N} \sigma(y_n \mathbf{w}^{\mathrm{T}} \widetilde{\mathbf{x}}_n)\right] p(\mathbf{w})$ over $\mathbf{w}$ is intractable (**how do you know?**), but this is required to compute

- The normalization constant of the posterior distribution $p(\mathbf{w}|\mathbf{y}, \widetilde{\mathbf{X}})$.
- The predictive distribution $p(y_\star|\widetilde{\mathbf{x}}_\star, \mathbf{y}, \widetilde{\mathbf{X}})$.

Solution:

Use **approximate Bayesian inference**. Different approaches are possible:

1. Draw a sequence of asymptotically unbiased samples from $p(\mathbf{w}|\mathbf{y}, \widetilde{\mathbf{X}})$ (Monte Carlo methods).
2. Approximate $p(\mathbf{w}|\mathbf{y}, \widetilde{\mathbf{X}})$ with a simpler distribution (e.g. a Gaussian) (**Deterministic approximate inference**).

What are the trade-offs between these options?

The Laplace approximation

Fits a **Gaussian** (why Gaussian?) approximation to the posterior.

The univariate case:

Consider a scalar continuous variable w with

$$p(w|\mathcal{D}) = \frac{1}{Z} f(w) \,,$$

where $f(w) = p(w, \mathcal{D})$ for some data $\mathcal{D}$ and $Z = \int f(w) dw$.

The Laplace approximation

Fits a **Gaussian** (why Gaussian?) approximation to the posterior.

The univariate case:

Consider a scalar continuous variable w with

$$p(w|\mathcal{D}) = \frac{1}{Z} f(w) \,,$$

where $f(w) = p(w, \mathcal{D})$ for some data $\mathcal{D}$ and $Z = \int f(w) dw$.

Let $q(w) = \mathcal{N}(w|m, v)$ be the Gaussian approximation. How can we adjust the parameters m and v so that q is similar to $p(w|\mathcal{D})$?

The Laplace approximation

Fits a **Gaussian** (why Gaussian?) approximation to the posterior.

The univariate case:

Consider a scalar continuous variable w with

$$p(w|\mathcal{D}) = \frac{1}{Z} f(w)\,,$$

where $f(w) = p(w, \mathcal{D})$ for some data $\mathcal{D}$ and $Z = \int f(w) dw$.

Let $q(w) = \mathcal{N}(w|m, v)$ be the Gaussian approximation. How can we adjust the parameters m and v so that q is similar to $p(w|\mathcal{D})$?

A choice for m can be the **MAP solution**. We find a **mode** w_{MAP} of $p(w|\mathcal{D})$:

$$\left.\frac{df(w)}{dw}\right|_{w=w_{\text{MAP}}} = 0$$

Any **optimization algorithm** can be used for this purpose.

The Laplace approximation

Given $m = w_{\text{MAP}}$, **what should the value of v be?**

The Laplace approximation

Given $m = w_{\text{MAP}}$, **what should the value of v be?**

We consider a **truncated Taylor expansion** of $\log f(w)$ center at w_{MAP}:

$$\log f(w) \approx \log f(w_{\text{MAP}}) - \frac{1}{2} a (w - w_{\text{MAP}})^2 , \quad a = -\frac{d^2}{dw^2} \log f(w) \Big|_{w = w_{\text{MAP}}}$$

The Laplace approximation

Given $m = w_{\text{MAP}}$, **what should the value of v be?**

We consider a **truncated Taylor expansion** of $\log f(w)$ center at w_{MAP}:

$$\log f(w) \approx \log f(w_{\text{MAP}}) - \frac{1}{2} a (w - w_{\text{MAP}})^2 \,, \quad a = -\frac{d^2}{dw^2} \log f(w) \Big|_{w = w_{\text{MAP}}}$$

Taking the exponential we obtain:

$$f(w) \approx f(w_{\text{MAP}}) \exp\left\{ -\frac{a}{2} (w - w_{\text{MAP}})^2 \right\} = \tilde{q}(w) \quad q(w) = \mathcal{N}(w | w_{\text{MAP}}, a^{-1}) \,,$$

The exponentiated truncated series is **Gaussian**. Very easy to normalize!

The Laplace approximation

Given $m = w_{\text{MAP}}$, **what should the value of v be?**

We consider a **truncated Taylor expansion** of $\log f(w)$ center at w_{MAP}:

$$\log f(w) \approx \log f(w_{\text{MAP}}) - \frac{1}{2}a(w - w_{\text{MAP}})^2\,, \quad a = -\frac{d^2}{dw^2}\log f(w)\Big|_{w=w_{\text{MAP}}}$$

Taking the exponential we obtain:

$$f(w) \approx f(w_{\text{MAP}})\exp\left\{-\frac{a}{2}(w - w_{\text{MAP}})^2\right\} = \tilde{q}(w) \quad q(w) = \mathcal{N}(w|w_{\text{MAP}}, a^{-1})\,,$$

The exponentiated truncated series is **Gaussian**. Very easy to normalize!

The approximation of the **normalizer** of $p(w|\mathcal{D})$ is $Z \approx f(w_{\text{MAP}})\sqrt{\frac{2\pi}{a}}$.

The Laplace approximation

Given $m = w_{\text{MAP}}$, **what should the value of v be?**

We consider a **truncated Taylor expansion** of $\log f(w)$ center at w_{MAP}:

$$\log f(w) \approx \log f(w_{\text{MAP}}) - \frac{1}{2} a (w - w_{\text{MAP}})^2 \,, \quad a = -\frac{d^2}{dw^2} \log f(w) \Big|_{w = w_{\text{MAP}}}$$

Taking the exponential we obtain:

$$f(w) \approx f(w_{\text{MAP}}) \exp\left\{ -\frac{a}{2} (w - w_{\text{MAP}})^2 \right\} = \tilde{q}(w) \quad q(w) = \mathcal{N}(w | w_{\text{MAP}}, a^{-1}) \,,$$

The exponentiated truncated series is **Gaussian**. Very easy to normalize!

The approximation of the **normalizer** of $p(w|\mathcal{D})$ is $Z \approx f(w_{\text{MAP}}) \sqrt{\frac{2\pi}{a}}$.

Only defined if $a > 0$: f must have **negative second derivative** at w_{MAP}.
Beware of saddle points when finding w_{MAP}!

Example

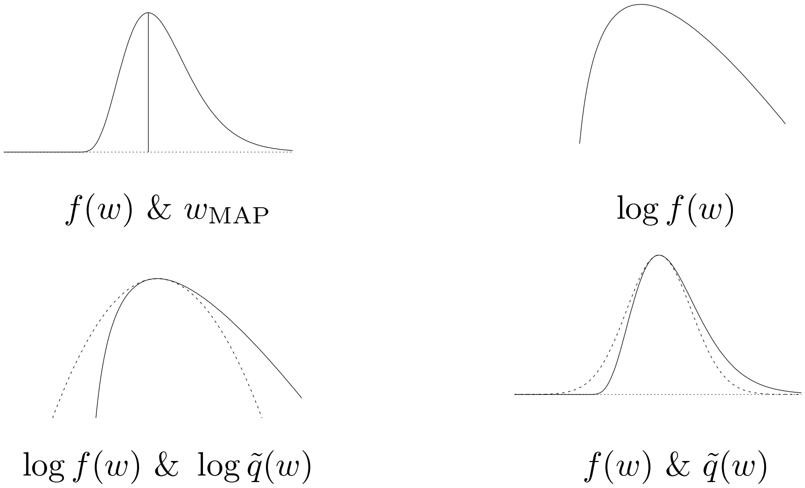


Figure: David J.C. MacKay. Information Theory, Inference, and Learning Algorithms. 2003.

The multi-variate case:

Now $\mathbf{w}$ is a d-dimensional vector and $p(\mathbf{w}|\mathcal{D}) = \frac{1}{Z} f(\mathbf{w})$.

The same principle can be applied. The truncated Taylor series is

$$\log f(\mathbf{w}_{\text{MAP}}) \approx \log f(\mathbf{w}_{\text{MAP}}) - \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^{\mathsf{T}} \mathbf{A} (\mathbf{w} - \mathbf{w}_{\text{MAP}}) \,,$$

where $\mathbf{A}$ is the **Hessian** of $-\log f(\mathbf{w})$ at $\mathbf{w}_{\text{MAP}}$, $\mathbf{A} = -\nabla^{\mathsf{T}} \nabla \log f(\mathbf{w})\Big|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}}$.

The multi-variate case:

Now $\mathbf{w}$ is a d-dimensional vector and $p(\mathbf{w}|\mathcal{D}) = \frac{1}{Z} f(\mathbf{w})$.

The same principle can be applied. The truncated Taylor series is

$$\log f(\mathbf{w}_{\mathrm{MAP}}) \approx \log f(\mathbf{w}_{\mathrm{MAP}}) - \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\mathrm{MAP}})^{\mathrm{T}} \mathbf{A} (\mathbf{w} - \mathbf{w}_{\mathrm{MAP}}),$$

where $\mathbf{A}$ is the **Hessian** of $-\log f(\mathbf{w})$ at $\mathbf{w}_{\mathrm{MAP}}$, $\mathbf{A} = -\nabla^{\mathrm{T}} \nabla \log f(\mathbf{w})\Big|_{\mathbf{w}=\mathbf{w}_{\mathrm{MAP}}}$.

Taking the exponential we obtain a **multi-variate Gaussian** approximation:

$$f(\mathbf{w}) \approx f(\mathbf{w}_{\mathrm{MAP}}) \exp\left\{ -\frac{1}{2}(\mathbf{w} - \mathbf{w}_{\mathrm{MAP}})^{\mathrm{T}} \mathbf{A} (\mathbf{w} - \mathbf{w}_{\mathrm{MAP}})) \right\} = \tilde{q}(\mathbf{w}),$$

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}}, \mathbf{A}^{-1}), \quad Z \approx f(\mathbf{w}_{\mathrm{MAP}})(2\pi)^{d/2} |\mathbf{A}|^{-1/2}.$$

The multi-variate case:

Now $\mathbf{w}$ is a d-dimensional vector and $p(\mathbf{w}|\mathcal{D}) = \frac{1}{Z}f(\mathbf{w})$.

The same principle can be applied. The truncated Taylor series is

$$\log f(\mathbf{w}_{\text{MAP}}) \approx \log f(\mathbf{w}_{\text{MAP}}) - \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^{\text{T}}\mathbf{A}(\mathbf{w} - \mathbf{w}_{\text{MAP}})\,,$$

where $\mathbf{A}$ is the **Hessian** of $-\log f(\mathbf{w})$ at $\mathbf{w}_{\text{MAP}}$, $\mathbf{A} = -\,\nabla^{\text{T}}\nabla \log f(\mathbf{w})\big|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}}$.

Taking the exponential we obtain a **multi-variate Gaussian** approximation:

$$f(\mathbf{w}) \approx f(\mathbf{w}_{\text{MAP}}) \exp\left\{-\frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^{\text{T}}\mathbf{A}(\mathbf{w} - \mathbf{w}_{\text{MAP}}))\right\} = \tilde{q}(\mathbf{w})\,,$$

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{MAP}}, \mathbf{A}^{-1})\,, \quad Z \approx f(\mathbf{w}_{\text{MAP}})(2\pi)^{d/2}|\mathbf{A}|^{-1/2}\,.$$

Many optimization methods can return the Hessian at the solution $\mathbf{w}_{\text{MAP}}$.

The multi-variate case:

Now $\mathbf{w}$ is a d-dimensional vector and $p(\mathbf{w}|\mathcal{D}) = \frac{1}{Z} f(\mathbf{w})$.

The same principle can be applied. The truncated Taylor series is

$$\log f(\mathbf{w}_{\text{MAP}}) \approx \log f(\mathbf{w}_{\text{MAP}}) - \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^{\text{T}} \mathbf{A} (\mathbf{w} - \mathbf{w}_{\text{MAP}}) \,,$$

where $\mathbf{A}$ is the **Hessian** of $-\log f(\mathbf{w})$ at $\mathbf{w}_{\text{MAP}}$, $\mathbf{A} = -\nabla^{\text{T}} \nabla \log f(\mathbf{w})\big|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}}$.

Taking the exponential we obtain a **multi-variate Gaussian** approximation:

$$f(\mathbf{w}) \approx f(\mathbf{w}_{\text{MAP}}) \exp\left\{ -\frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^{\text{T}} \mathbf{A} (\mathbf{w} - \mathbf{w}_{\text{MAP}})) \right\} = \tilde{q}(\mathbf{w}) \,,$$

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{MAP}}, \mathbf{A}^{-1}) \,, \quad Z \approx f(\mathbf{w}_{\text{MAP}})(2\pi)^{d/2} |\mathbf{A}|^{-1/2} \,.$$

Many optimization methods can return the Hessian at the solution $\mathbf{w}_{\text{MAP}}$.

Only defined if $\mathbf{A}$ **positive definite**. **Beware of saddle points!**

Example: probit regression

The posterior distribution is:

$$p(\mathbf{w}|\mathcal{D}) \propto p(\mathbf{y}|\mathbf{w}, \widetilde{\mathbf{X}})p(\mathbf{w}) = f(\mathbf{w}), \quad \log f(\mathbf{w}) = \log p(\mathbf{y}|\mathbf{w}, \widetilde{\mathbf{X}}) + \log p(\mathbf{w}).$$

Let $\Phi(\cdot)$ be the standard Gaussian cdf. Then, we have that:

$$\log f(\mathbf{w}) = \sum_{n=1}^{N} \log \Phi(y_n \mathbf{w}^{\mathrm{T}} \widetilde{\mathbf{x}}_n) - \frac{1}{2}\lambda \mathbf{w}^{\mathrm{T}}\mathbf{w} + \text{const}.$$

Let $\mathbf{w}_{\text{MAP}}$ be the maximizer of f. Computing the negative Hessian at $\mathbf{w}_{\text{MAP}}$:

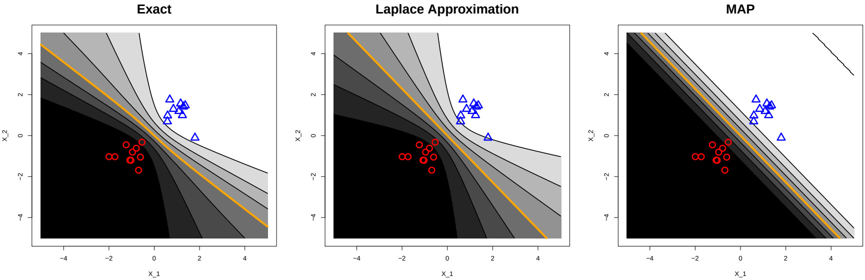
$$\mathbf{A} = \sum_{n=1}^{N} \left[v_n(s_n + v_n)\widetilde{\mathbf{x}}_n \widetilde{\mathbf{x}}_n^{\mathrm{T}} \right] + \lambda \mathbf{I}, \quad v_n = \frac{\mathcal{N}(s_n|0,1)}{\Phi(s_n)}, \quad s_n = y_n \mathbf{w}_{\text{MAP}}^{\mathrm{T}} \widetilde{\mathbf{x}}_n.$$

We then have $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{MAP}}, \mathbf{A}^{-1})$ and $Z \approx f(\mathbf{w}_{\text{MAP}})\sqrt{\frac{(2\pi)^d}{|\mathbf{A}|}}$.

An approx. **predictive distribution** is obtained by replacing $p(\mathbf{w}|\mathcal{D})$ with $q(\mathbf{w})$:

$$p(y_\star|\widetilde{\mathbf{x}}_\star, \mathbf{y}) \approx \int p(y_\star|\widetilde{\mathbf{x}}_\star, \mathbf{w}) q(\mathbf{w}) d\mathbf{w} = \int \Phi(y_\star \mathbf{w}^{\mathrm{T}} \widetilde{\mathbf{x}}_\star) \mathcal{N}(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}}, \mathbf{A}^{-1}) d\mathbf{w}\,,$$

$$= \Phi\left(\frac{y_\star \mathbf{w}_{\mathrm{MAP}}^{\mathrm{T}} \widetilde{\mathbf{x}}_\star}{\sqrt{\widetilde{\mathbf{x}}_\star^{\mathrm{T}} \mathbf{A}^{-1} \widetilde{\mathbf{x}}_\star + 1}}\right)\,.$$



Uncertainty increases in regions where there is **no data**.

The MAP predictive uncertainty is **constant** along the decision border.

Considerations for the Laplace approximation

- The Hessian can be approximated by numerical differences.
- Multiple possible solutions on multi-modal distributions, one per mode.
- Often, the posterior is more and more Gaussian as $N \to \infty$.
- Does not work with discrete random variables or discrete likelihoods.
- Fails to capture global properties, only considers curvature at the mode.
- A change of variables changes the solution (a defect or an opportunity).