# Laplace’s method and GP classification

Week 5b - Statistical ML / Thang Bui / ANU / 2025 S1 - Thanks to Yuan-Sen Ting for many slides

# Plan for today and next Wed

Bishop's textbook: chapter 4.4-4.5, 6.4.5-6.4.6
GP book: http://gaussianprocess.org/gpml/chapters/
chapter 3.1-3.4

# Plan for today and next Wed

Bishop's textbook: chapter 4.4-4.5, 6.4.5-6.4.6
GP book: http://gaussianprocess.org/gpml/chapters/
chapter 3.1-3.4

- Bayesian Logistic Regression

  - the challenge to computing a non-Gaussian predictive distribution

# Plan for today and next Wed

Bishop's textbook: chapter 4.4-4.5, 6.4.5-6.4.6
GP book: http://gaussianprocess.org/gpml/chapters/
chapter 3.1-3.4

- Bayesian Logistic Regression

  - the challenge to computing a non-Gaussian predictive distribution

- Laplace Approximation for Bayesian Logistic Regression

# Plan for today and next Wed

Bishop's textbook: chapter 4.4-4.5, 6.4.5-6.4.6
GP book: http://gaussianprocess.org/gpml/chapters/
chapter 3.1-3.4

- Bayesian Logistic Regression

  - the challenge to computing a non-Gaussian predictive distribution

- Laplace Approximation for Bayesian Logistic Regression
- Gaussian Process Classification

# Plan for today and next Wed

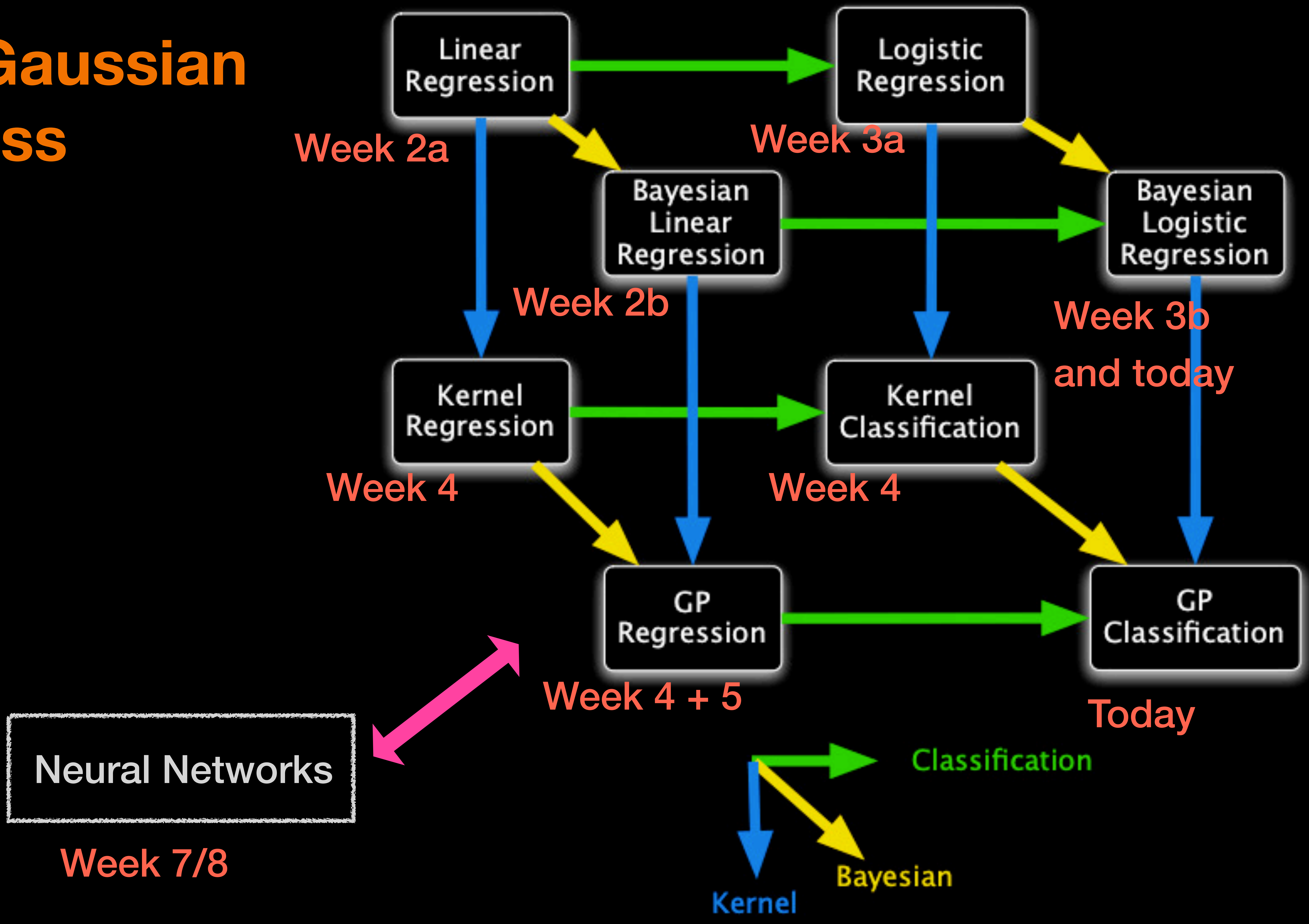
Bishop's textbook: chapter 4.4-4.5, 6.4.5-6.4.6
GP book: http://gaussianprocess.org/gpml/chapters/
chapter 3.1-3.4

- Bayesian Logistic Regression

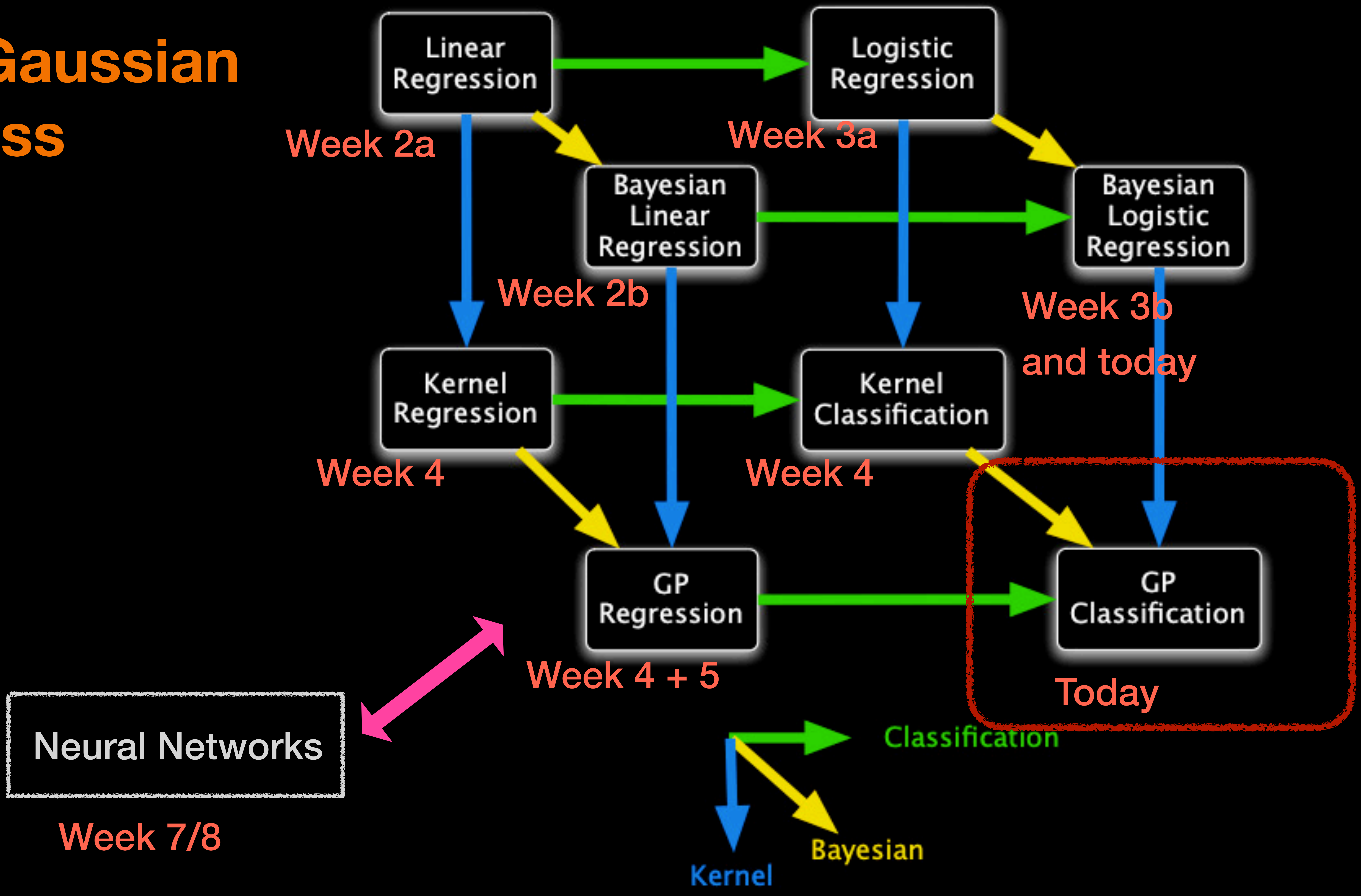
  - the challenge to computing a non-Gaussian predictive distribution

- Laplace Approximation for Bayesian Logistic Regression
- Gaussian Process Classification
- Laplace Approximation for Gaussian Process Classification

# Why Gaussian Process



# Why Gaussian Process



# Last Time on Bayesian Linear Regression

Predictive distribution

$$p(y^* \,|\, \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{\mathbf{p}(\mathbf{y}^* \,|\, \mathbf{x}^*, \theta)}_{\text{Gaussian}} \, \underbrace{\mathbf{p}(\theta \,|\, \mathbf{y}, \mathbf{X})}_{\text{Gaussian}} \, \mathrm{d}\theta$$

Recall in Bayesian linear regression

$$p(\theta \,|\, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\theta; \mathbf{m}, \mathbf{S})$$

# Recap : Logistic Regression

$$p(y = 1 \mid \theta, x) = g_\theta(\mathbf{x}) = \underbrace{\sigma}_{\text{Sigmoid}}\left(\theta^T \phi(\mathbf{x})\right)$$

Bishop eq 4.87

Given training data $\{\mathbf{x_n}, \mathbf{y_n}\}, y_n \in \{0,1\}$

Likelihood

$$p(\mathbf{Y} \mid \theta, \mathbf{X}) = \prod_{n=1}^{N} g_n^{y_n} (1 - g_n)^{(1-y_n)}$$

Bishop eq 4.89

# Recap : Logistic Regression

## Likelihood

$$p(\mathbf{y} \mid \theta, \mathbf{X}) = \prod_{n=1}^{N} g_n^{y_n} (1 - g_n)^{1-y_n} = \prod_{n=1}^{N} \sigma(\theta^T \boldsymbol{\phi}(\mathbf{x_n}))^{y_n} \left\{ 1 - \sigma(\theta^T \boldsymbol{\phi}(\mathbf{x_n}) \right\}^{1-y_n}$$

# Recap : Logistic Regression

Likelihood

$$p(\mathbf{y} \,|\, \theta, \mathbf{X}) = \prod_{n=1}^{N} g_n^{y_n} (1 - g_n)^{1-y_n} = \prod_{n=1}^{N} \sigma(\theta^T \boldsymbol{\phi}(\mathbf{x_n}))^{y_n} \left\{1 - \sigma(\theta^T \boldsymbol{\phi}(\mathbf{x_n})\right\}^{1-y_n}$$

Prior

$$p(\theta) = \mathcal{N}(\theta; \mathbf{m_0}, \mathbf{S_0})$$

Bishop eq 4.140

# Recap : Logistic Regression

Likelihood

$$p(\mathbf{y} \mid \theta, \mathbf{X}) = \prod_{n=1}^{N} g_n^{y_n} (1 - g_n)^{1-y_n} = \prod_{n=1}^{N} \sigma(\theta^T \boldsymbol{\phi}(\mathbf{x_n}))^{y_n} \left\{ 1 - \sigma(\theta^T \boldsymbol{\phi}(\mathbf{x_n})) \right\}^{1-y_n}$$

Prior

$$p(\theta) = \mathcal{N}(\theta; \mathbf{m_0}, \mathbf{S_0})$$

Bishop eq 4.140

Posterior

$$p(\theta \mid \mathbf{y}, \mathbf{X}) \propto \mathcal{N}(\mathbf{m_0}, \mathbf{S_0}) \cdot \prod_{n=1}^{N} g_n^{y_n} (1 - g_n)^{1-y_n}$$

# Bayesian Logistic Regression

Predictive distribution

$$p(y^* = 1 \,|\, \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \mathbf{p}(\mathbf{y}^* = \mathbf{1} \,|\, \mathbf{x}^*, \theta)\, \mathbf{p}(\theta \,|\, \mathbf{y}, \mathbf{X})\; \mathrm{d}\theta$$

# Bayesian Logistic Regression

Predictive distribution

$$p(y^* = 1 \,|\, \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{\mathbf{p}(\mathbf{y}^* = \mathbf{1} \,|\, \mathbf{x}^*, \theta)}_{\text{Sigmoid}} \, \mathbf{p}(\theta \,|\, \mathbf{y}, \mathbf{X}) \, \mathrm{d}\theta$$

$$g^* = \sigma\left(\theta^T \phi(\mathbf{x}^*)\right)$$

# Bayesian Logistic Regression

Predictive distribution

$$p(y^* = 1 \,|\, \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{\mathbf{p}(\mathbf{y}^* = \mathbf{1} \,|\, \mathbf{x}^*, \theta)}_{\text{Sigmoid}} \, \underbrace{\mathbf{p}(\theta \,|\, \mathbf{y}, \mathbf{X})}_{\text{Non-Gaussian}} \, \mathrm{d}\theta$$

$$g^* = \sigma\left(\theta^T \phi(\mathbf{x}^*)\right)$$

$$\propto \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0) \cdot \prod_{n=1}^{N} g_n^{y_n} \{1 - g_n\}^{1-y_n}$$

# Bayesian Logistic Regression

Predictive distribution

$$p(y^* = 1 \,|\, \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{\mathbf{p}(\mathbf{y}^* = \mathbf{1} \,|\, \mathbf{x}^*, \theta)}_{\text{Sigmoid}} \, \underbrace{\mathbf{p}(\theta \,|\, \mathbf{y}, \mathbf{X})}_{\text{Non-Gaussian}} \, \mathrm{d}\theta$$

Not analytically tractable

$$g^* = \sigma\left(\theta^T \phi(\mathbf{x}^*)\right)$$

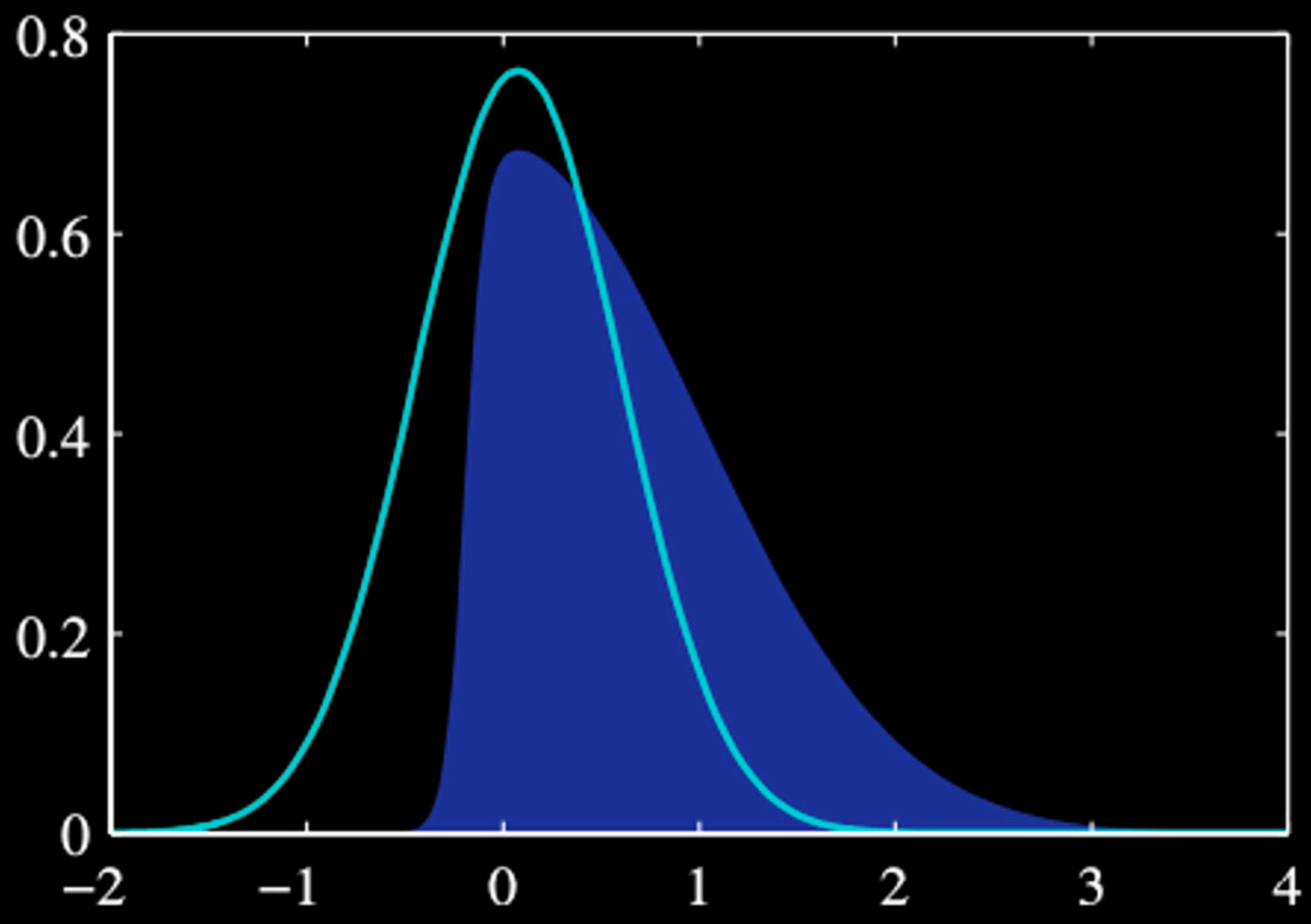
$$\propto \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0) \cdot \prod_{n=1}^{N} g_n^{y_n} \{1 - g_n\}^{1-y_n}$$

Let's first "Gaussianise" this part
using Laplace Approximation

# Laplace Approximation in 1D

- Goal: find a Gaussian distribution $q(z)$ that approximates the PDF $p(z)$

- Idea: find out $q(z)$ which centers at the mode of $p(z)$ with the same "curvature" at that maximum point

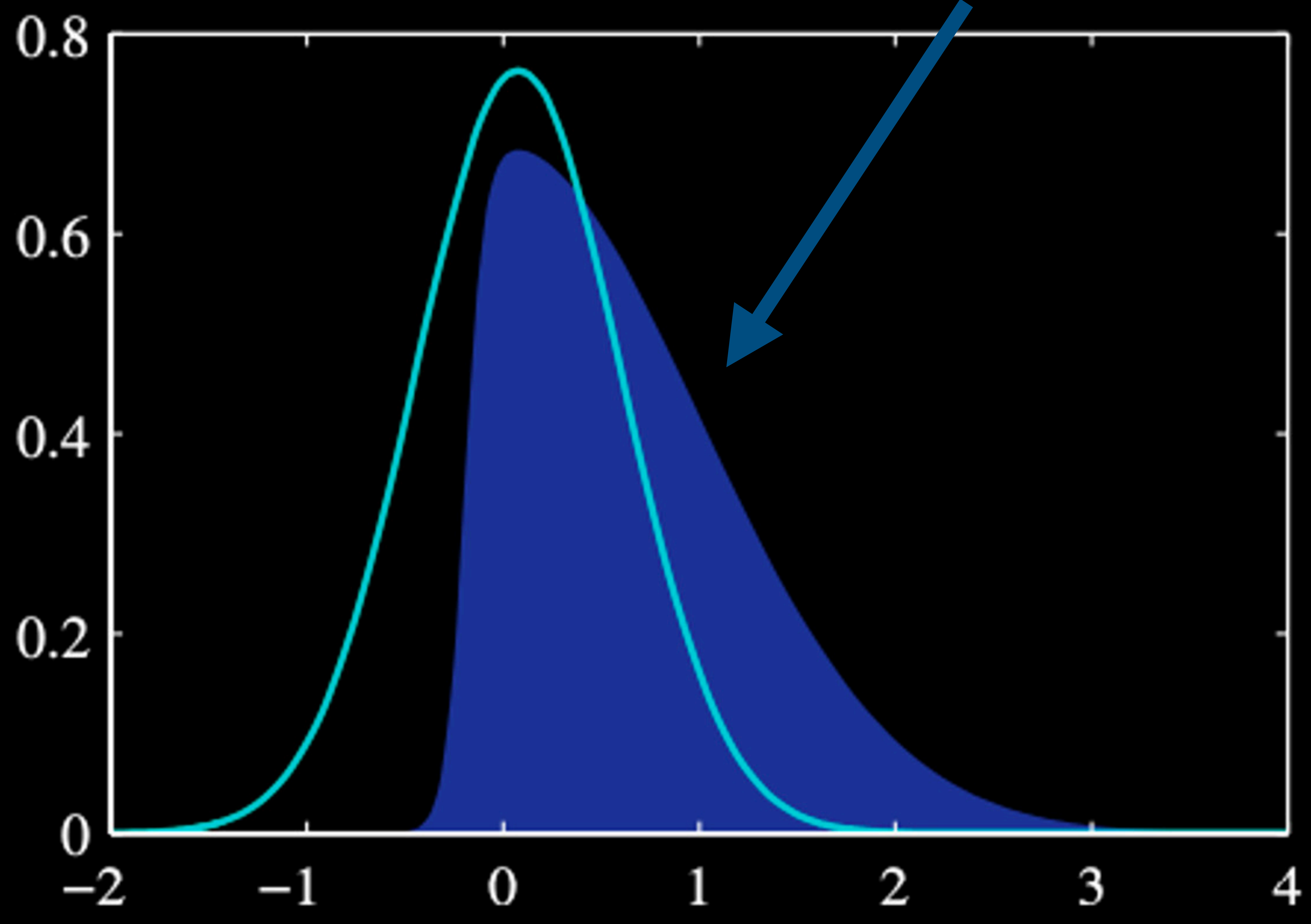


# Laplace Approximation in 1D

- Goal: find a Gaussian distribution $q(z)$ that approximates the PDF $p(z)$

- Idea: find out $q(z)$ which centers at the mode of $p(z)$ with the same "curvature" at that maximum point

$$p(z) \propto \exp(-z^2/2)\,\overbrace{\sigma}^{\text{Sigmoid}}(20z+4)$$

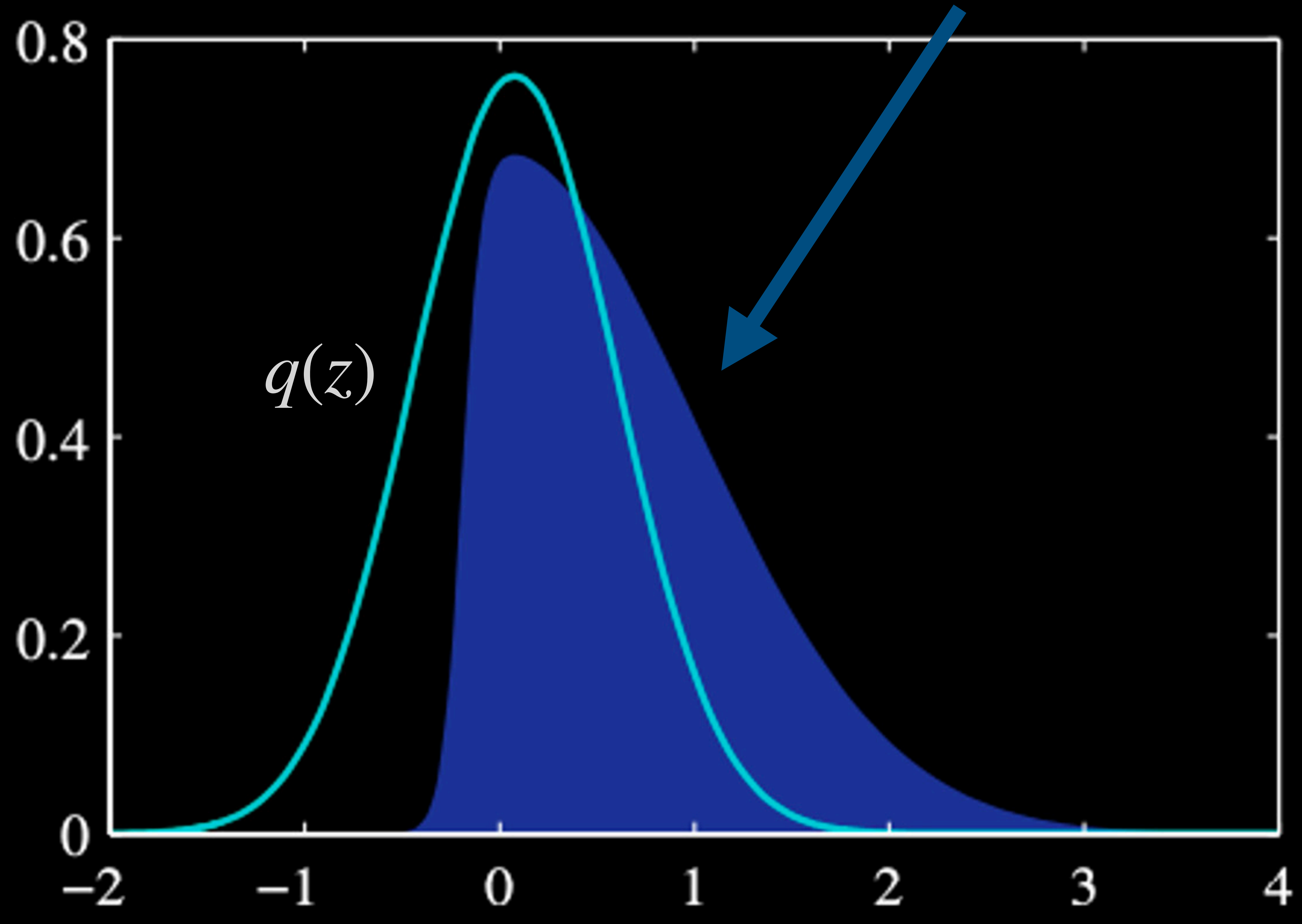


# Laplace Approximation in 1D

$$p(z) \propto \exp(-z^2/2)\, \overbrace{\sigma}^{\text{Sigmoid}}(20z + 4)$$

- Goal: find a Gaussian distribution $q(z)$ that approximates the PDF $p(z)$

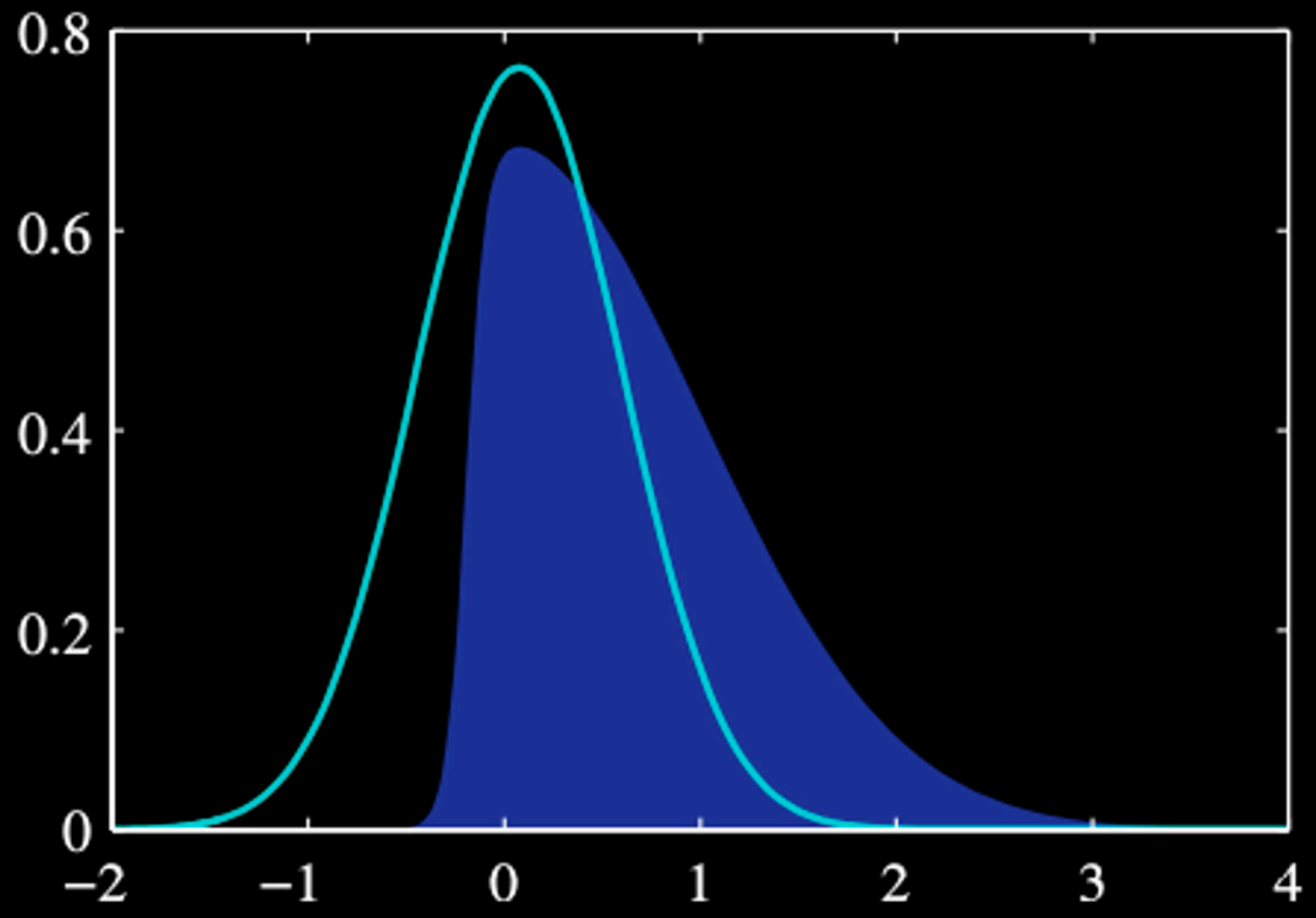
- Idea: find out $q(z)$ which centers at the mode of $p(z)$ with the same "curvature" at that maximum point



# Laplace Approximation in 1D

Let $p(z) = \frac{1}{Z} f(z)$ where $Z = \int f(z)\,\mathrm{d}z$

Bishop eq 4.125

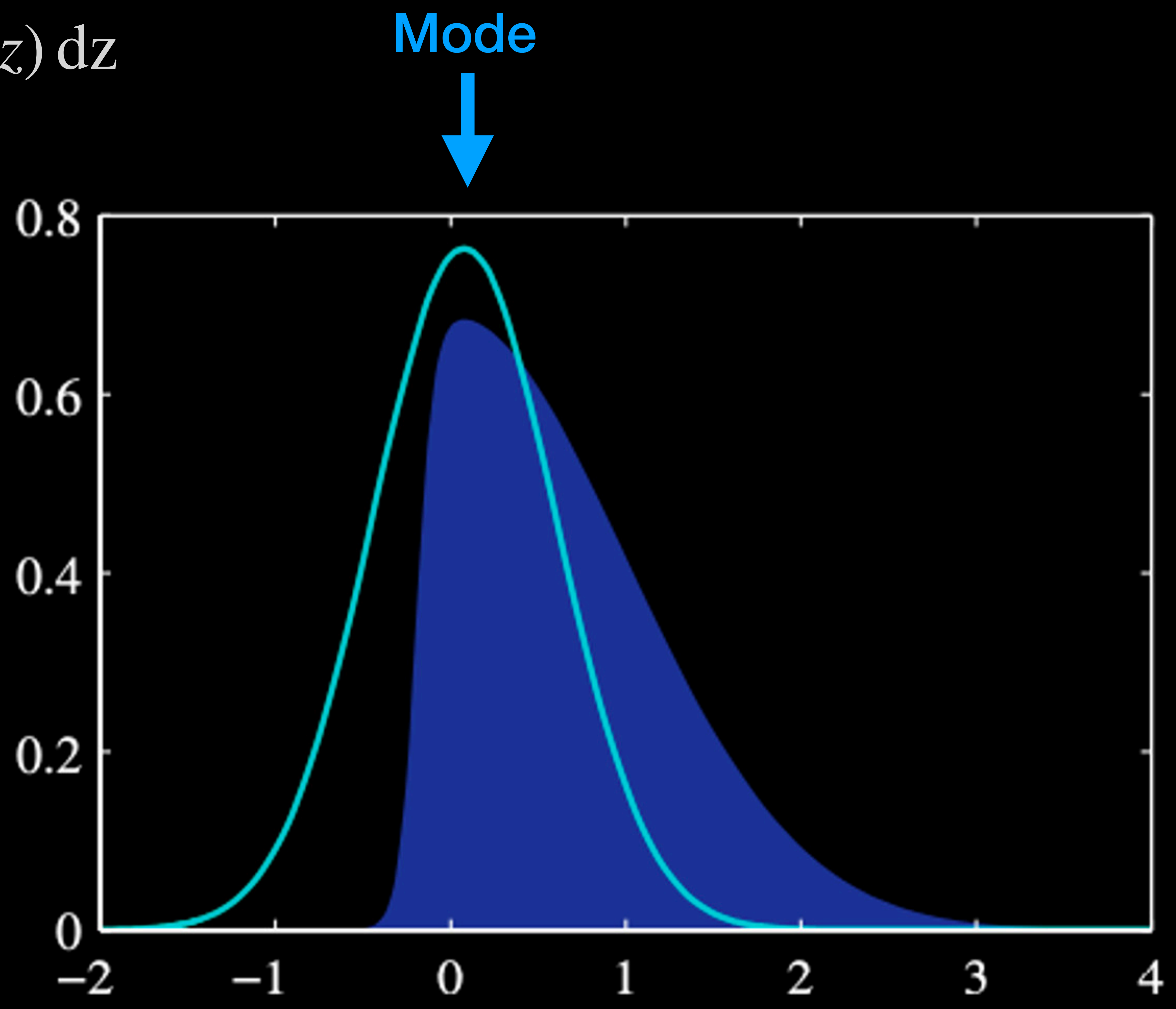


# Laplace Approximation in 1D

Let $p(z) = \frac{1}{Z} f(z)$ where $Z = \int f(z)\,\mathrm{d}z$

Bishop eq 4.125

Find the mode



# Laplace Approximation in 1D

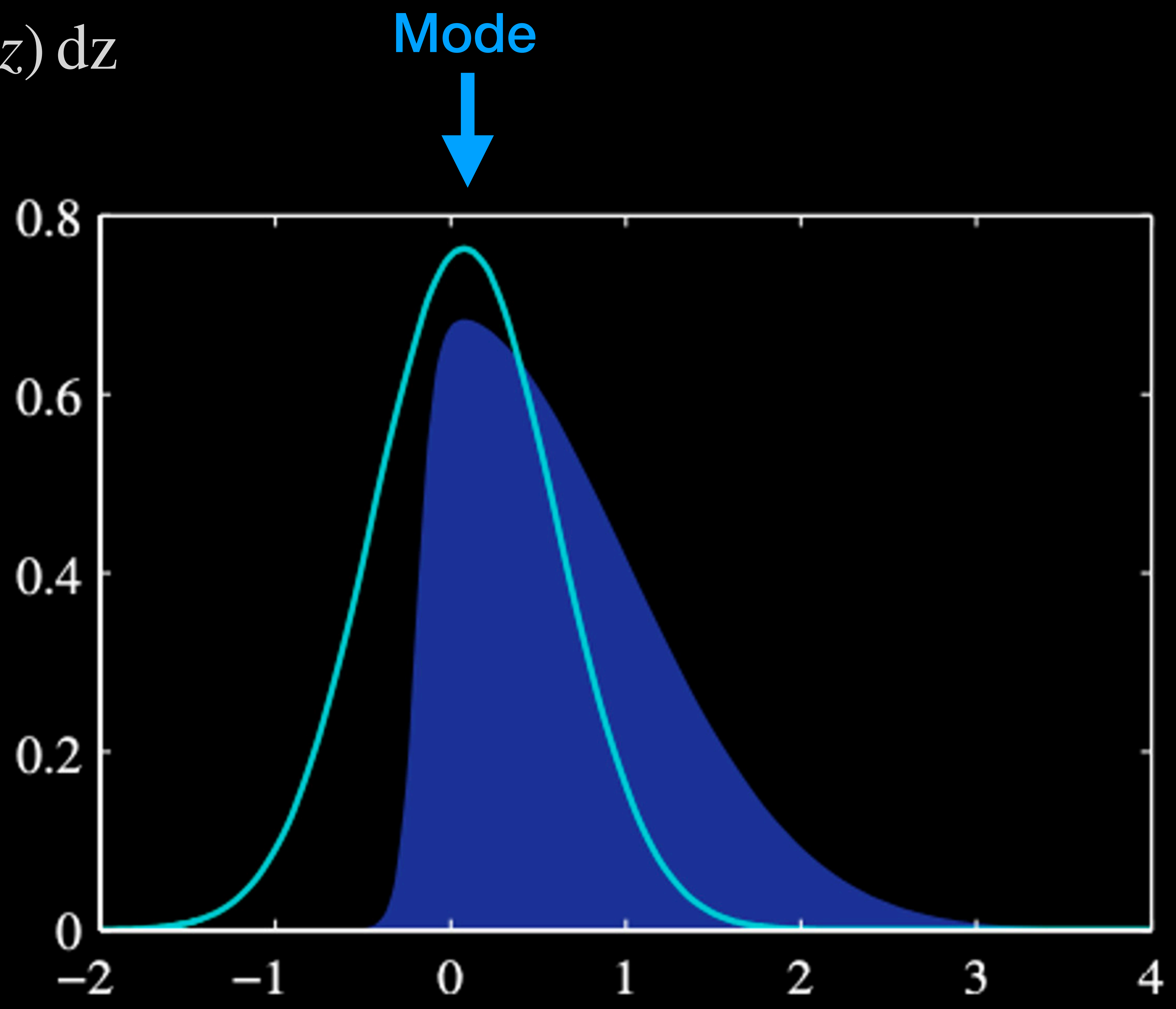
Let $p(z) = \frac{1}{Z} f(z)$ where $Z = \int f(z) \, \mathrm{d}z$

Bishop eq 4.125

Find the mode

$$\left. \frac{\mathrm{d}f(z)}{\mathrm{d}z} \right|_{z=z_0} = 0$$

Bishop eq 4.126

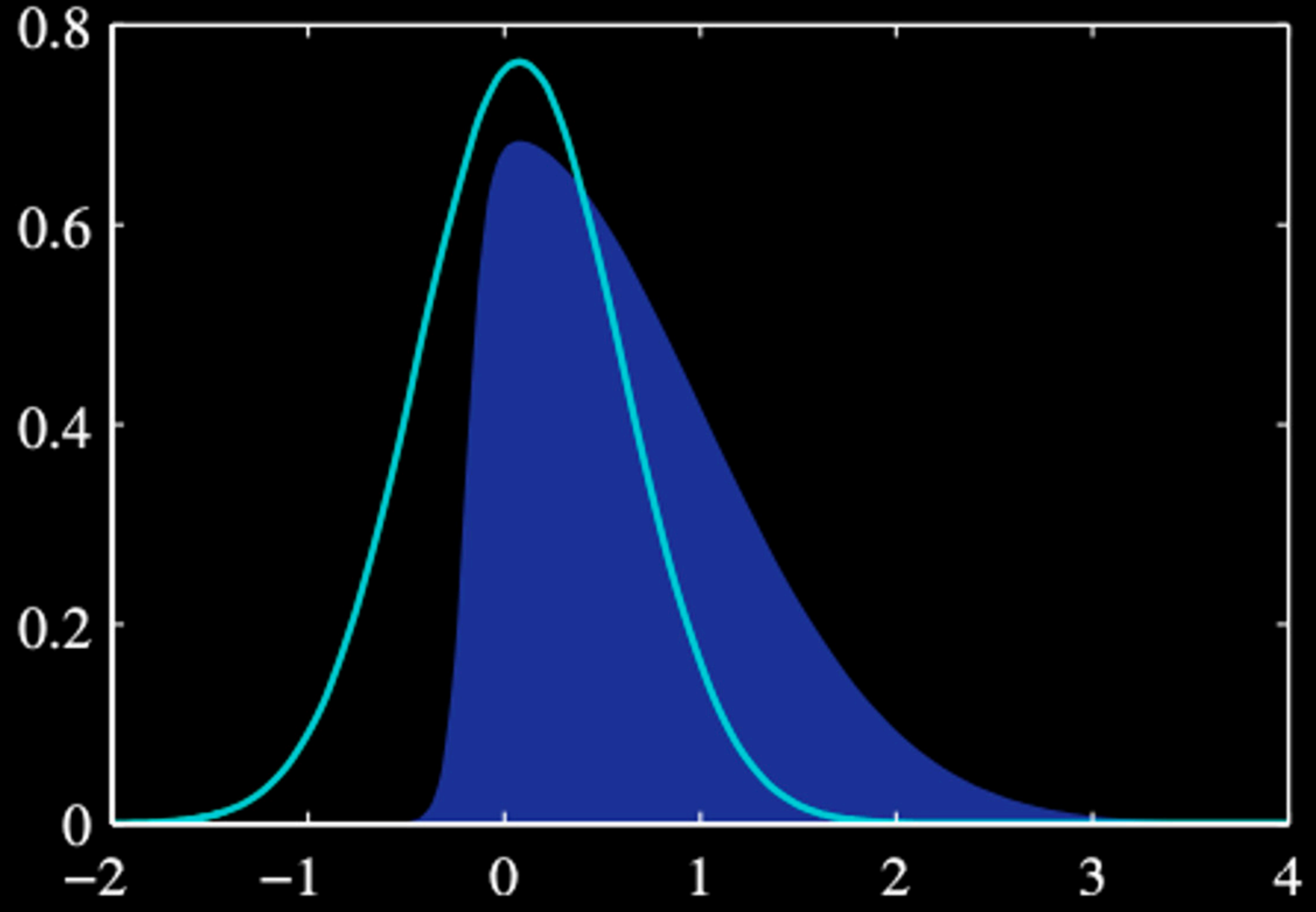


# Laplace Approximation in 1D

Let $p(z) = \frac{1}{Z} f(z)$ where $Z = \int f(z)\,\mathrm{d}z$

Bishop eq 4.125

Taylor expansion at $z_0$



# Laplace Approximation in 1D

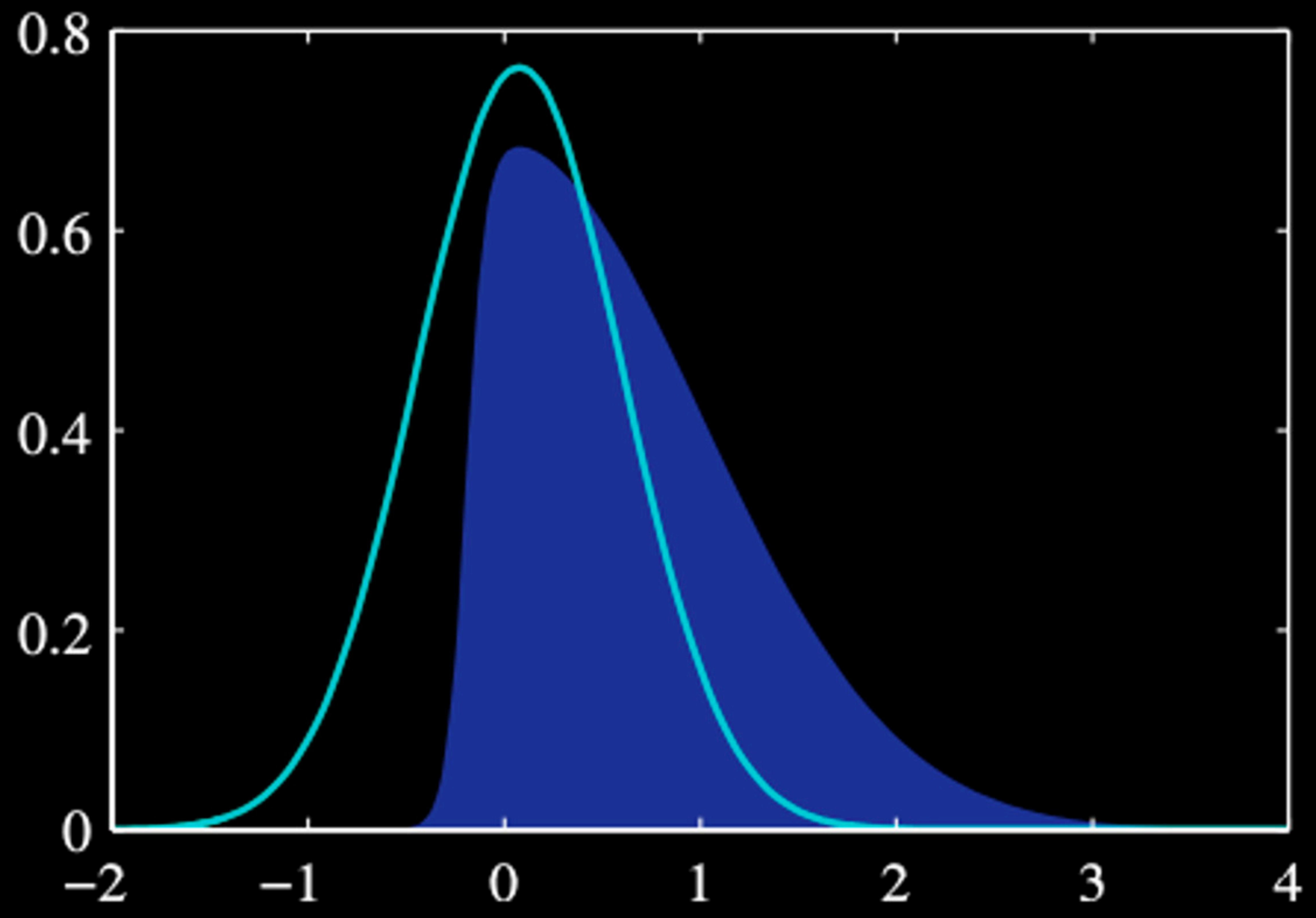
Let $p(z) = \frac{1}{Z} f(z)$ where $Z = \int f(z)\,\mathrm{d}z$

Bishop eq 4.125

Taylor expansion at $z_0$

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2} A (z - z_0)^2$$

Bishop eq 4.127



# Laplace Approximation in 1D

Let $p(z) = \frac{1}{Z} f(z)$ where $Z = \int f(z)\, \mathrm{d}z$

Bishop eq 4.125

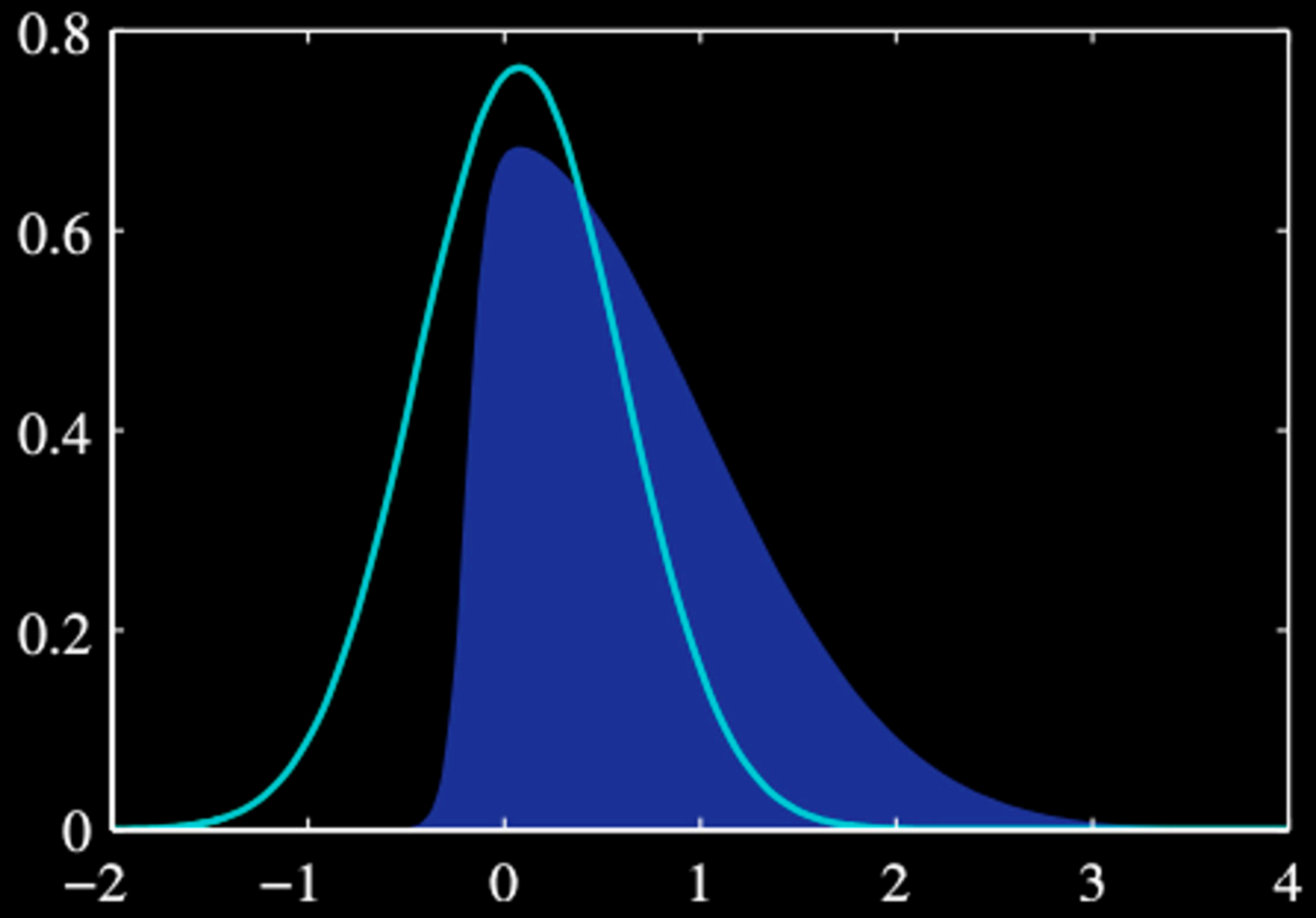
Taylor expansion at $z_0$

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2} A (z - z_0)^2$$

Bishop eq 4.127

$$A = -\frac{\mathrm{d}^2}{\mathrm{d}z^2} \ln f(z) \bigg|_{z=z_0}$$

Bishop eq 4.128



# Laplace Approximation in 1D

Let $p(z) = \frac{1}{Z} f(z)$ where $Z = \int f(z)\,\mathrm{d}z$

Bishop eq 4.125

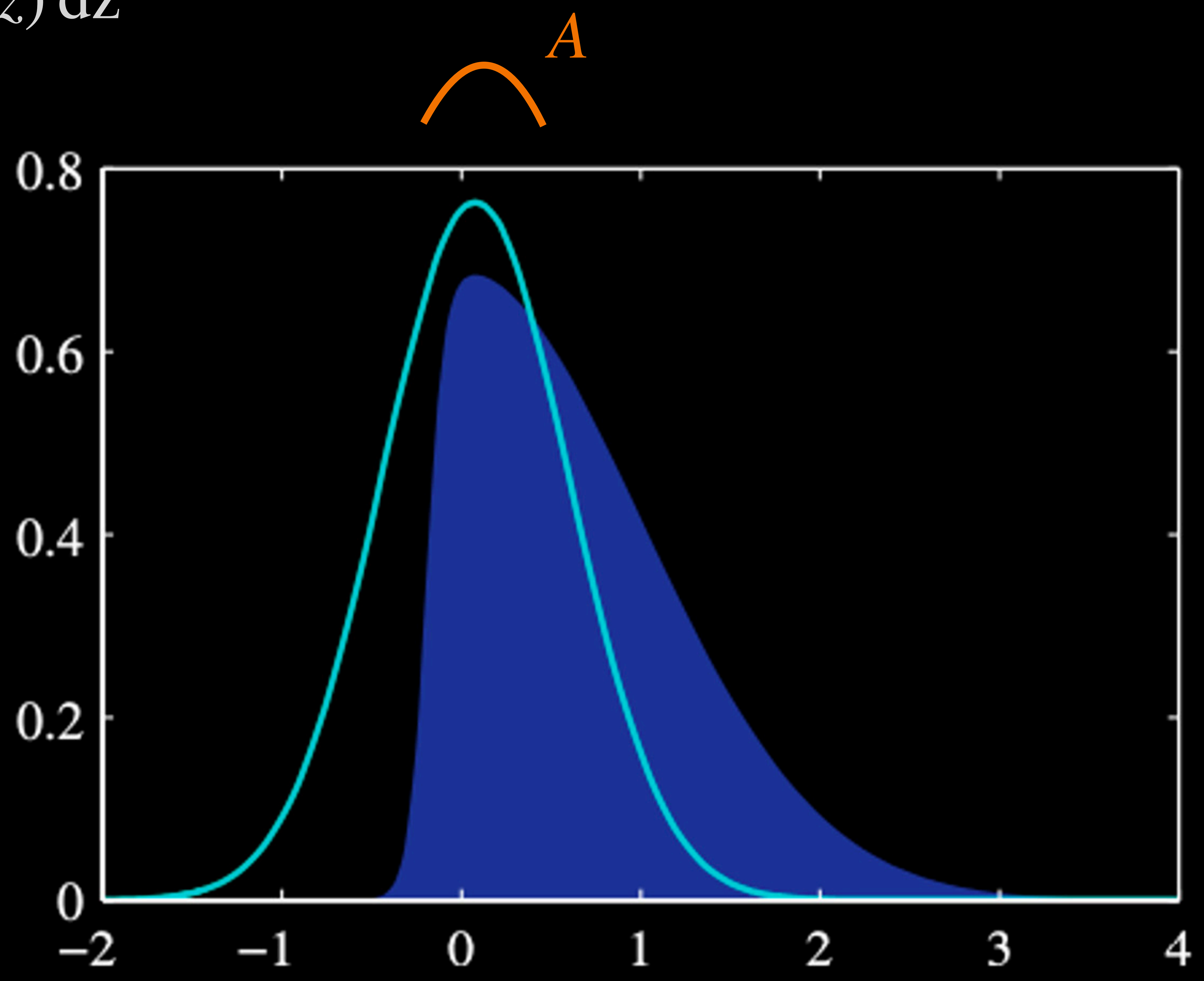
Taylor expansion at $z_0$

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2} A (z - z_0)^2$$

Bishop eq 4.127

$$A = -\frac{\mathrm{d}^2}{\mathrm{d}z^2} \ln f(z) \Big|_{z=z_0}$$

Bishop eq 4.128



# Laplace Approximation in 1D

Let $p(z) = \frac{1}{Z} f(z)$ where $Z = \int f(z)\,\mathrm{dz}$

Bishop eq 4.125

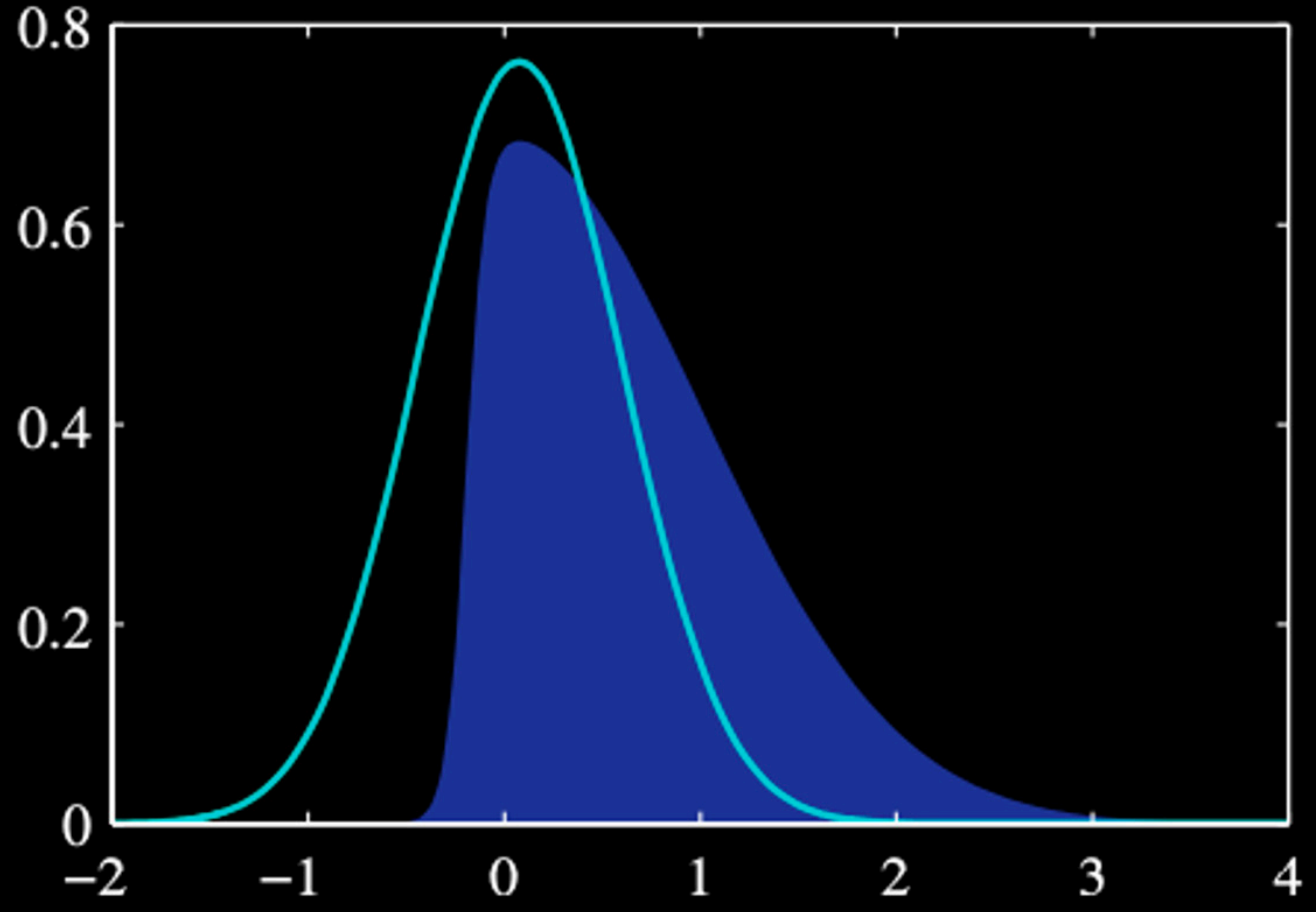
Taylor expansion at $z_0$

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2} A (z - z_0)^2$$

Bishop eq 4.127

$$f(z) \simeq f(z_0) \exp\left\{ -\frac{A}{2}(z - z_0)^2 \right\}$$

Bishop eq 4.129



# Laplace Approximation in 1D

Let $p(z) = \frac{1}{Z} f(z)$ where $Z = \int f(z)\, \mathrm{dz}$

Bishop eq 4.125

Taylor expansion at $z_0$

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2} A (z - z_0)^2$$

Bishop eq 4.127

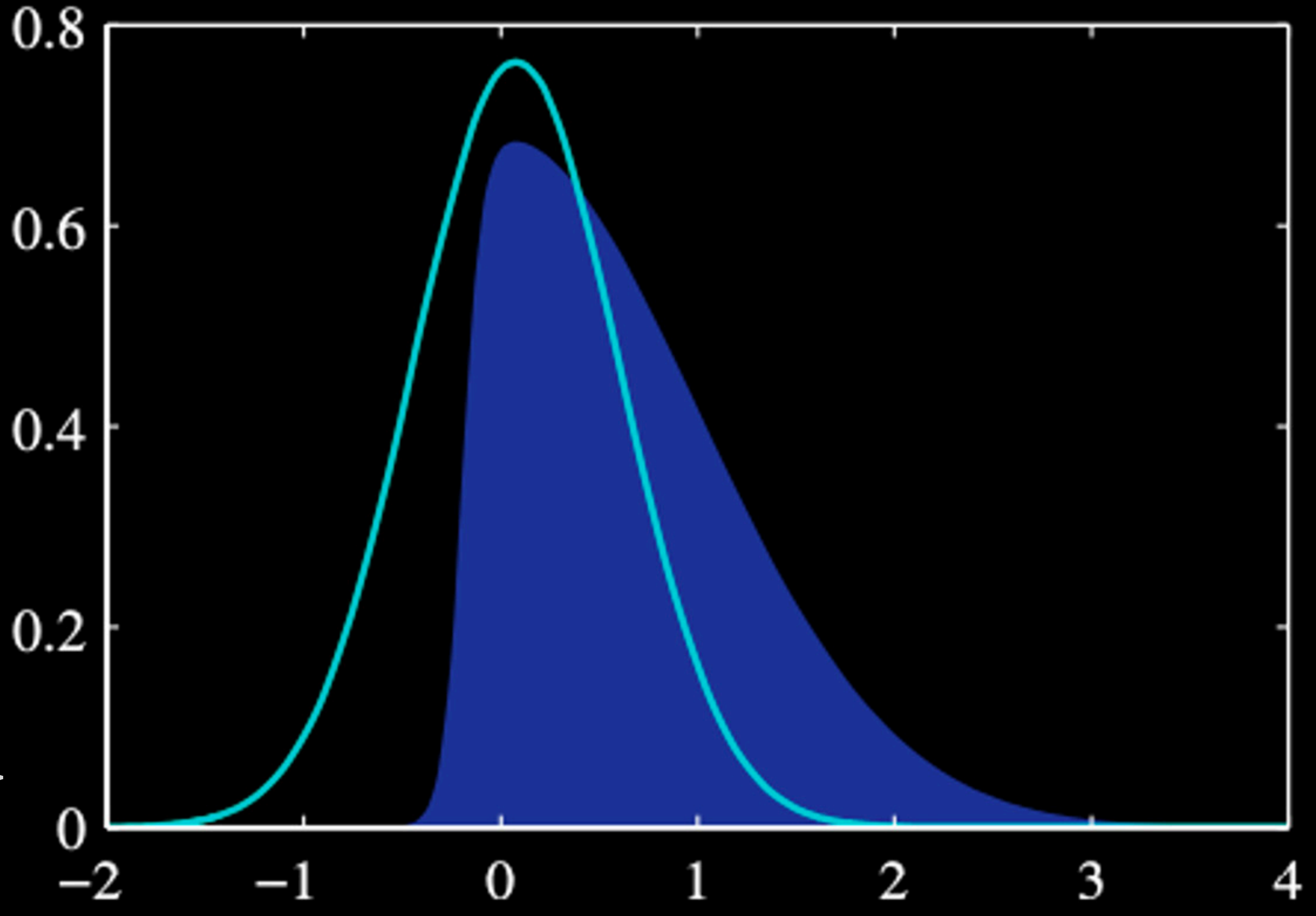
$$f(z) \simeq f(z_0) \exp\left\{ -\frac{A}{2} (z - z_0)^2 \right\}$$

Bishop eq 4.129

Normalize

$$\Rightarrow q(z) = \left( \frac{A}{2\pi} \right)^{1/2} \exp\left\{ -\frac{A}{2} (z - z_0)^2 \right\}$$

Bishop eq 4.130



# Laplace Approximation in 1D

Mode $\left.\frac{\mathrm{d}f(z)}{\mathrm{d}z}\right|_{z=z_0} = 0$

$$A = -\left.\frac{\mathrm{d}^2}{\mathrm{d}z^2}\ln f(z)\right|_{z=z_0}$$

Let $p(z) = \frac{1}{Z} f(z)$ where $Z = \int f(z)\,\mathrm{d}z$

Bishop eq 4.125

Taylor expansion at $z_0$

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2}A(z-z_0)^2$$

Bishop eq 4.127

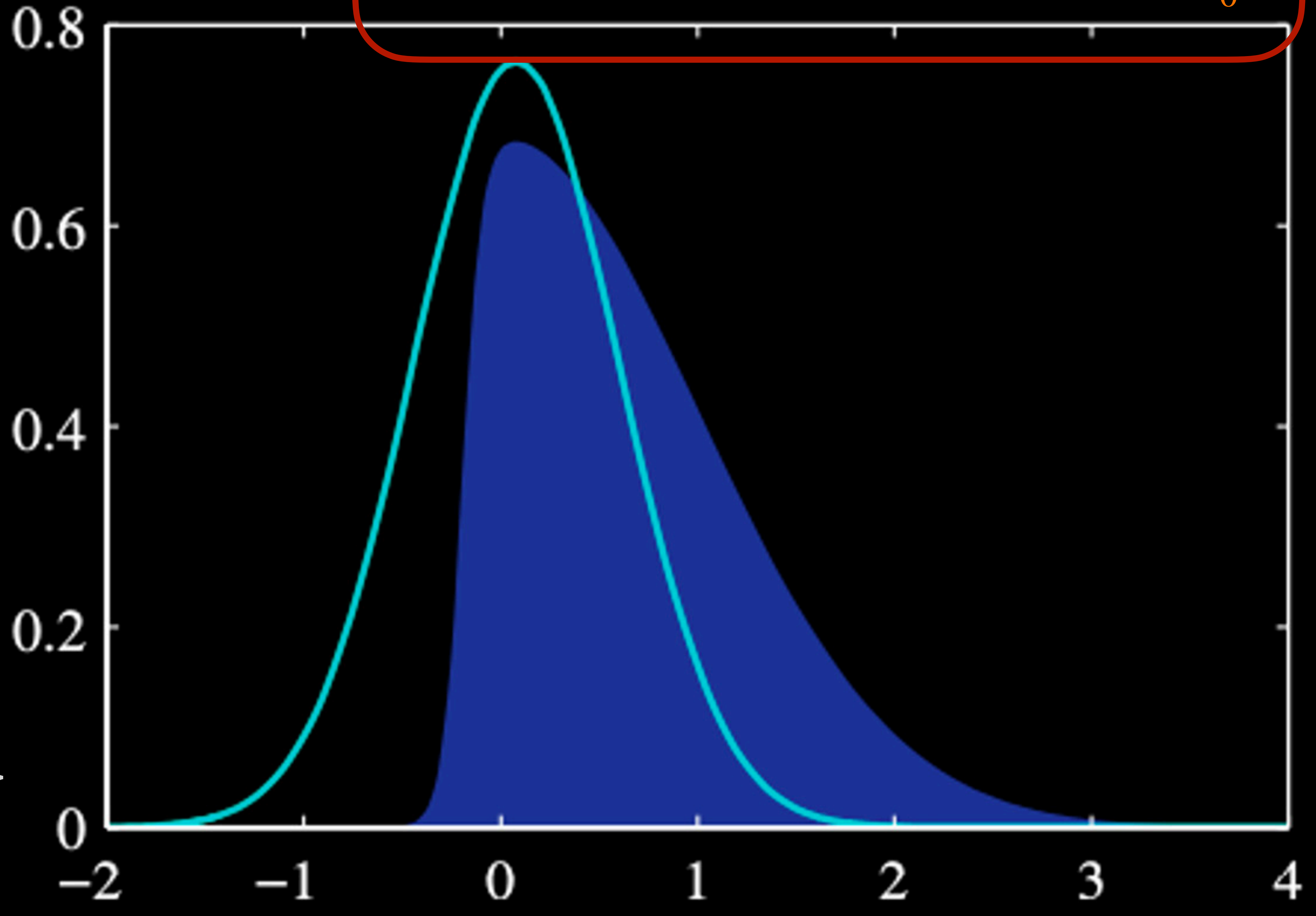
$$f(z) \simeq f(z_0)\exp\left\{-\frac{A}{2}(z-z_0)^2\right\}$$

Bishop eq 4.129

Normalize

$$\Rightarrow \quad q(z) = \left(\frac{A}{2\pi}\right)^{1/2}\exp\left\{-\frac{A}{2}(z-z_0)^2\right\}$$

Bishop eq 4.130



Pros:

Cons:

Pros: ✓

Cons: ✗

- Assume that $\mathbf{z}$ spans the entire support of $\mathbb{R}^d$, might need to perform a change of variable if, e.g., $\mathbf{z} \in \mathbb{R}^{+d}$

Pros:

Cons:

- Assume that $\mathbf{z}$ spans the entire support of $\mathbb{R}^d$, might need to perform a change of variable if, e.g., $\mathbf{z} \in \mathbb{R}^{+d}$
- The approximation is purely based on local information around $\mathbf{z}_0$ , disregarding the global information about $p(\mathbf{z})$

Pros: ✓

- Normalization constant $Z$ for $p(\mathbf{z})$ is not needed

---

Cons: ✗

- Assume that $\mathbf{z}$ spans the entire support of $\mathbb{R}^d$, might need to perform a change of variable if, e.g., $\mathbf{z} \in \mathbb{R}^{+d}$

- The approximation is purely based on local information around $\mathbf{z_0}$ , disregarding the global information about $p(\mathbf{z})$

Pros: ✓

- Normalization constant $Z$ for $p(\mathbf{z})$ is not needed

- Central Limit Theorem ensures that the posterior is better approximated by a Gaussian as the number of observed data increases

---

Cons: ✗

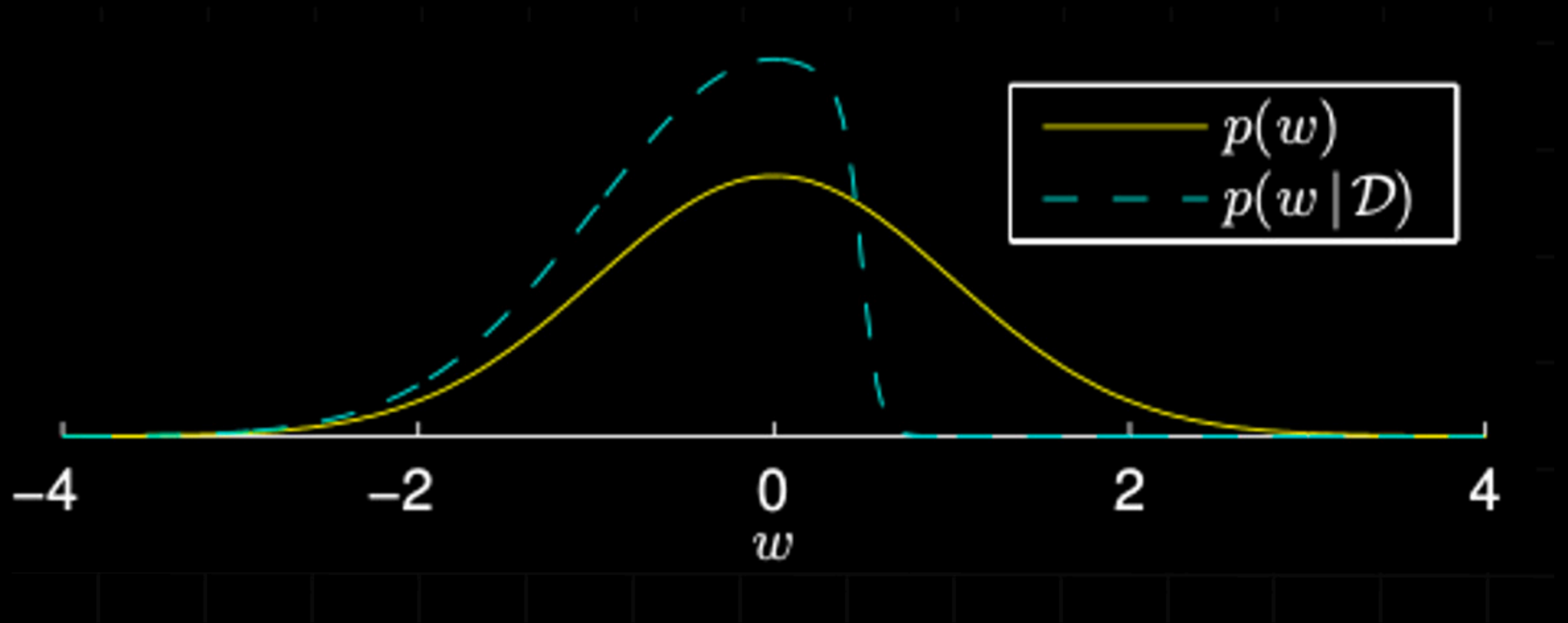
- Assume that $\mathbf{z}$ spans the entire support of $\mathbb{R}^d$, might need to perform a change of variable if, e.g., $\mathbf{z} \in \mathbb{R}^{+d}$

- The approximation is purely based on local information around $\mathbf{z}_0$ , disregarding the global information about $p(\mathbf{z})$

# Why Laplacian Approximation Works Well

Recall

$$p(\theta \,|\, \mathbf{y}, \mathbf{X}) \propto \mathcal{N}(\mathbf{m_0}, \mathbf{S_0}) \cdot \prod_{n=1}^{N} \sigma(\theta^T \boldsymbol{\phi}(\mathbf{x_n}))^{y_n} \left\{ 1 - \sigma(\theta^T \boldsymbol{\phi}(\mathbf{x_n})) \right\}^{1-y_n}$$

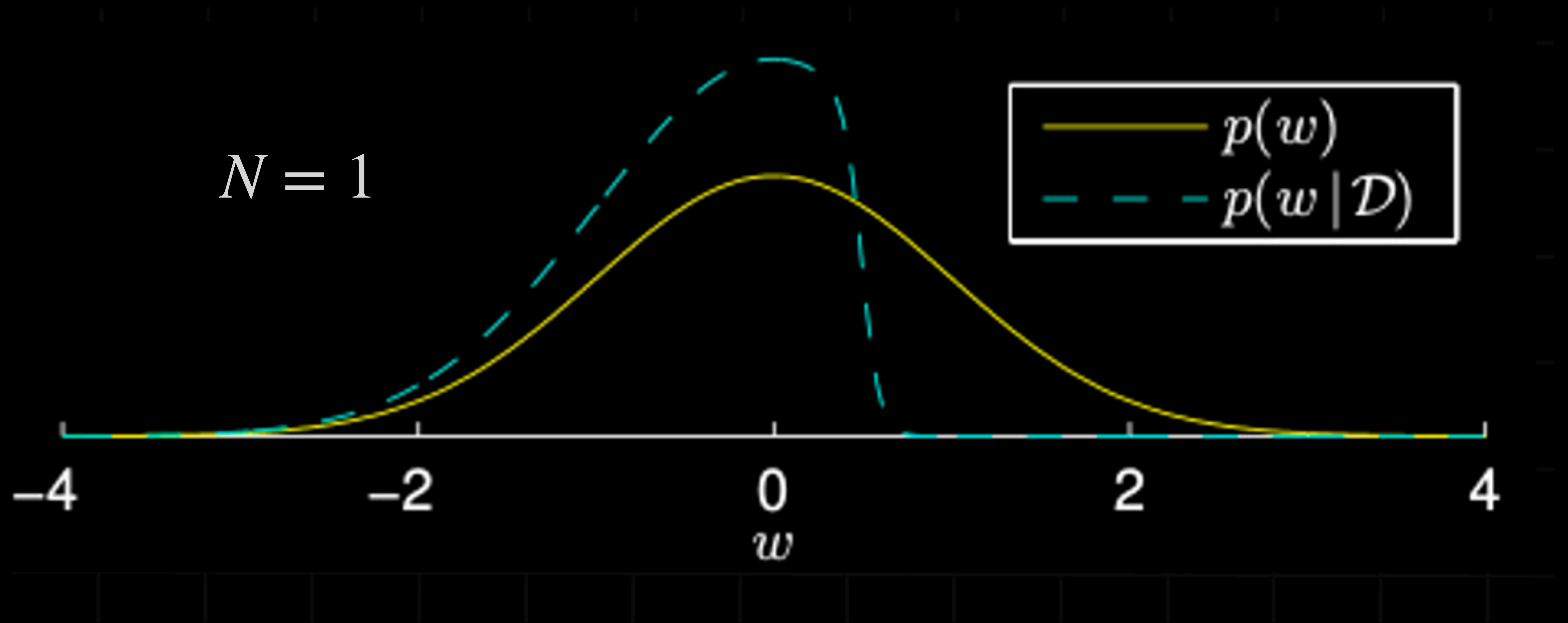



Example courtesy of Edinburgh MLPR course
https://www.inf.ed.ac.uk/teaching/courses/mlpr/2016/notes/w8a_bayes_logistic_regression_laplace.pdf

# Why Laplacian Approximation Works Well

Recall

$$p(\theta\,|\,\mathbf{y}, \mathbf{X}) \propto \mathcal{N}(\mathbf{m_0}, \mathbf{S_0}) \cdot \prod_{n=1}^{N} \sigma(\theta^T \boldsymbol{\phi}(\mathbf{x_n}))^{y_n} \left\{1 - \sigma(\theta^T \boldsymbol{\phi}(\mathbf{x_n})\right\}^{1-y_n}$$



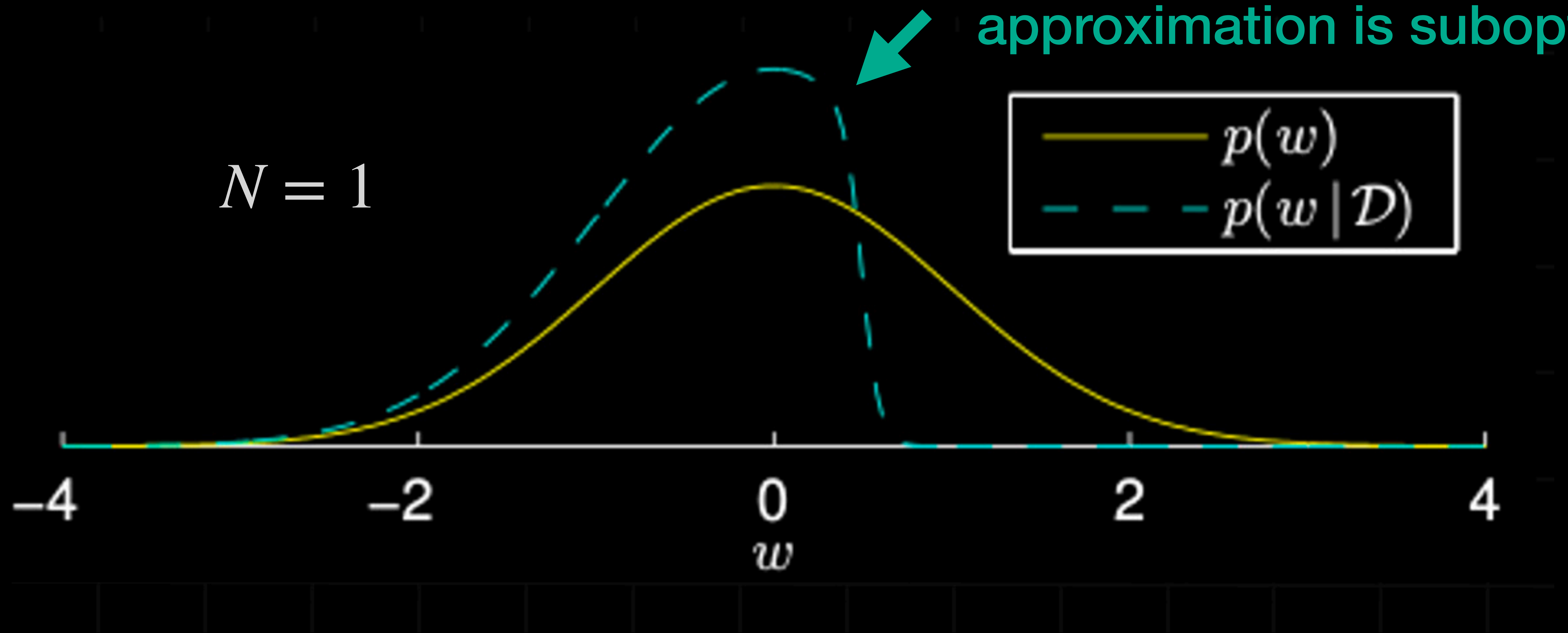

Example courtesy of Edinburgh MLPR course
https://www.inf.ed.ac.uk/teaching/courses/mlpr/2016/notes/w8a_bayes_logistic_regression_laplace.pdf

# Why Laplacian Approximation Works Well

Recall

$$p(\theta \,|\, \mathbf{y}, \mathbf{X}) \propto \mathcal{N}(\mathbf{m_0}, \mathbf{S_0}) \cdot \prod_{n=1}^{N} \sigma(\theta^T \boldsymbol{\phi}(\mathbf{x_n}))^{y_n} \{1 - \sigma(\theta^T \boldsymbol{\phi}(\mathbf{x_n}))\}^{1-y_n}$$

Posterior : Gaussian approximation is suboptimal



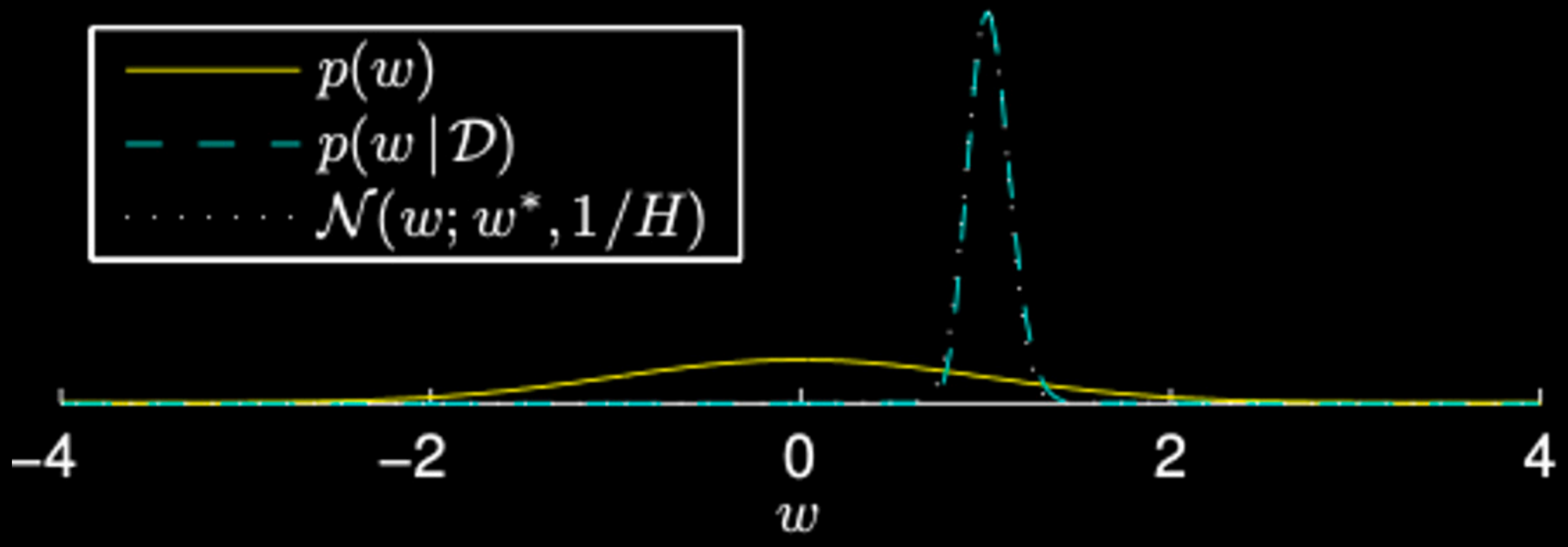

Example courtesy of Edinburgh MLPR course
https://www.inf.ed.ac.uk/teaching/courses/mlpr/2016/notes/w8a_bayes_logistic_regression_laplace.pdf

# Why Laplacian Approximation Works Well

Recall

$$p(\theta \,|\, \mathbf{y}, \mathbf{X}) \propto \mathcal{N}(\mathbf{m_0}, \mathbf{S_0}) \cdot \prod_{n=1}^{N} \sigma(\theta^T \boldsymbol{\phi}(\mathbf{x_n}))^{y_n} \left\{1 - \sigma(\theta^T \boldsymbol{\phi}(\mathbf{x_n})\right\}^{1-y_n}$$

$N = 500$

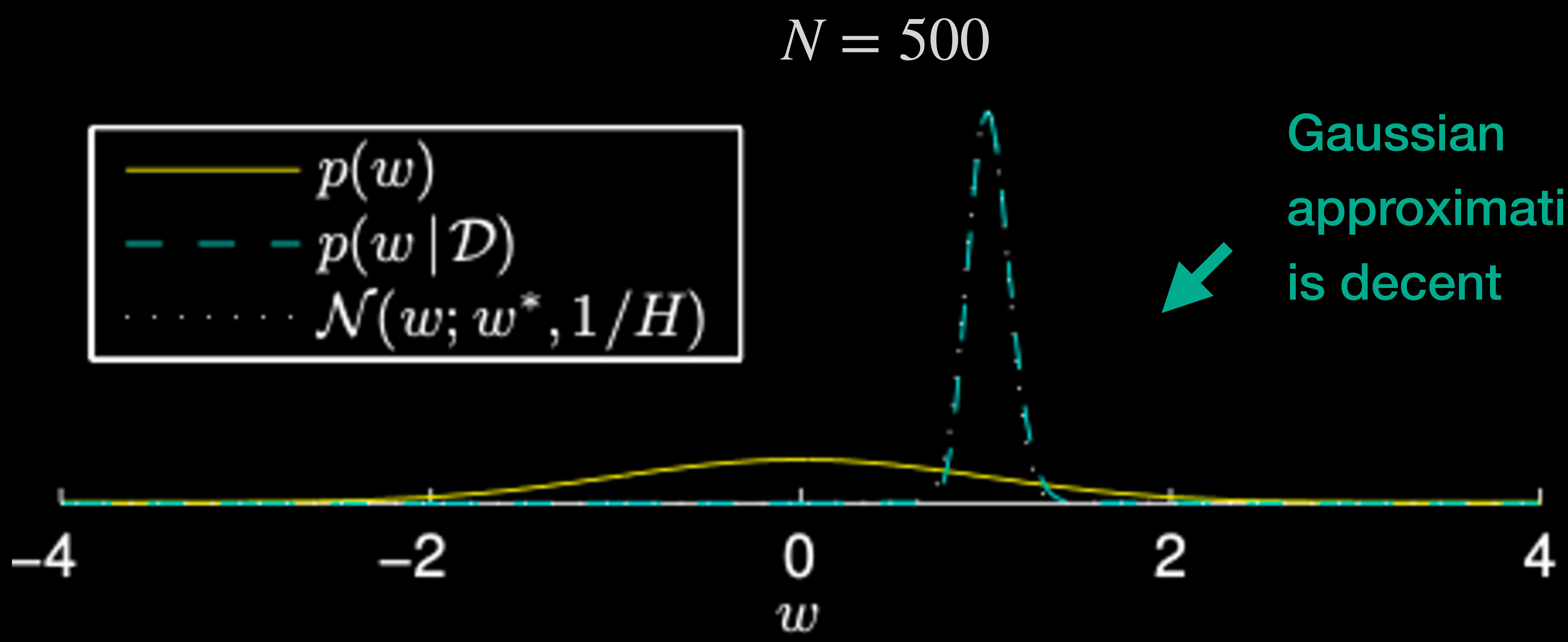



Example courtesy of Edinburgh MLPR course
https://www.inf.ed.ac.uk/teaching/courses/mlpr/2016/notes/w8a_bayes_logistic_regression_laplace.pdf

# Why Laplacian Approximation Works Well

Recall

$$p(\theta \,|\, \mathbf{y}, \mathbf{X}) \propto \mathcal{N}(\mathbf{m_0}, \mathbf{S_0}) \cdot \prod_{n=1}^{N} \sigma(\theta^T \boldsymbol{\phi}(\mathbf{x_n}))^{y_n} \{1 - \sigma(\theta^T \boldsymbol{\phi}(\mathbf{x_n}))\}^{1-y_n}$$




Gaussian approximation is decent

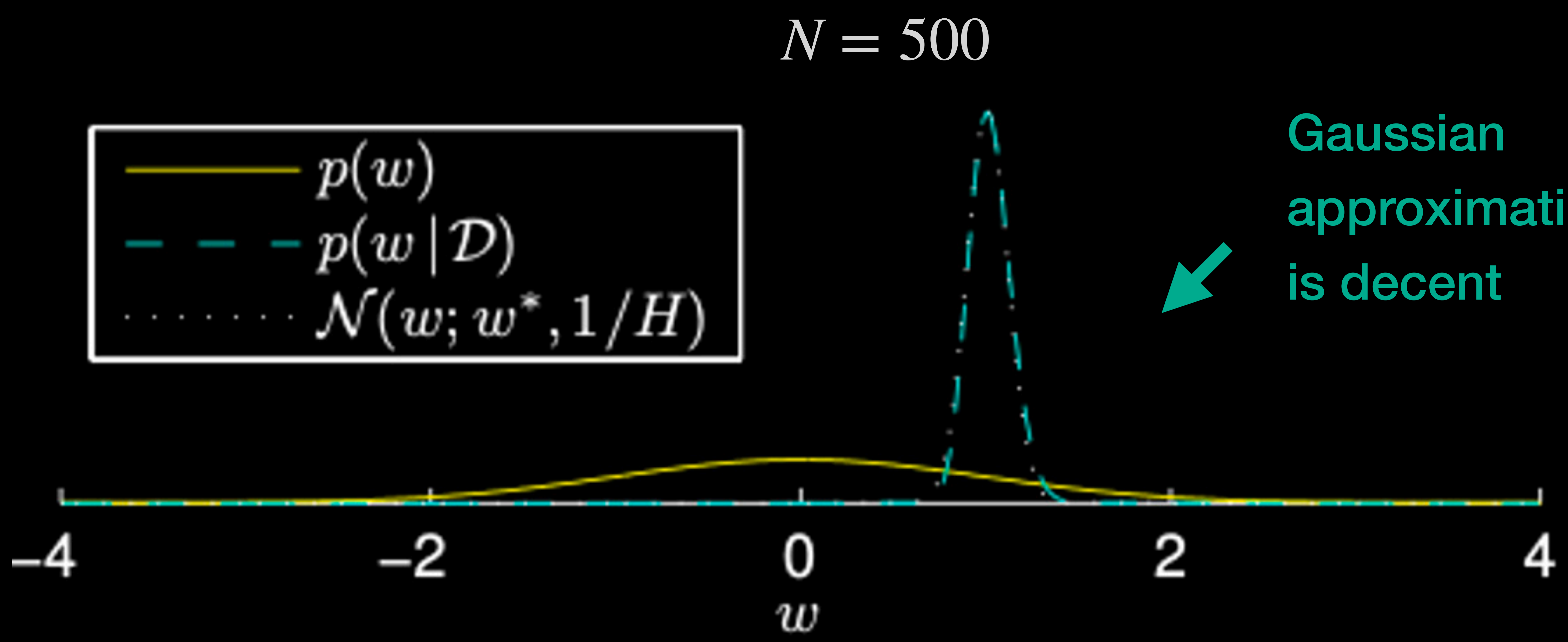
Example courtesy of Edinburgh MLPR course
https://www.inf.ed.ac.uk/teaching/courses/mlpr/2016/notes/w8a_bayes_logistic_regression_laplace.pdf

# Why Laplacian Approximation Works Well

Recall

$$p(\theta \,|\, \mathbf{y}, \mathbf{X}) \propto \mathcal{N}(\mathbf{m_0}, \mathbf{S_0}) \cdot \prod_{n=1}^{N} \sigma(\theta^T \boldsymbol{\phi}(\mathbf{x_n}))^{y_n} \left\{1 - \sigma(\theta^T \boldsymbol{\phi}(\mathbf{x_n})\right\}^{1-y_n}$$

This function is concave and have a unique maximum (exercise)




Gaussian approximation is decent

Example courtesy of Edinburgh MLPR course
https://www.inf.ed.ac.uk/teaching/courses/mlpr/2016/notes/w8a_bayes_logistic_regression_laplace.pdf

# Laplace Approximation in Higher Dimensions

Stationary point - local maximum :

$$\nabla f(\mathbf{z})\Big|_{\mathbf{z}=\mathbf{z}_0} = 0$$

Hessian matrix :

$$\underbrace{\mathbf{A}}_{\mathbb{R}^{m\times m},\, m \text{ number of dimensions}} = -\nabla\nabla \ln f(\mathbf{z})\Big|_{\mathbf{z}=\mathbf{z}_0}$$

Bishop eq 4.132

Multivariate Gaussian - provided that $A$ is positive definite, i.e. $\mathbf{z}_0$ is a local maximum

$$q(\mathbf{z}) = (2\pi)^{-M/2}|\mathbf{A}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\} = \mathcal{N}(\mathbf{z}|\mathbf{z}_0, \mathbf{A}^{-1})$$

Bishop eq 4.134, GP Book eq 3.11

# Laplace Approximation for Bayesian Logistic Regression

$$p(\theta \,|\, \mathbf{y}, \mathbf{X}) \simeq \mathbf{q}(\theta) = \mathcal{N}(\theta \,|\, \theta_{\mathbf{MAP}}, \mathbf{S})$$

Bishop eq 4.144

$$p(\theta \,|\, \mathbf{y}, \mathbf{X}) \propto \mathcal{N}(\mathbf{m_0}, \mathbf{S_0}) \cdot \prod_{n=1}^{N} g_n^{y_n} \{1 - g_n\}^{1-y_n}$$

# Laplace Approximation for Bayesian Logistic Regression

$$p(\theta \,|\, \mathbf{y}, \mathbf{X}) \simeq \mathbf{q}(\theta) = \mathcal{N}(\theta \,|\, \theta_{\mathbf{MAP}}, \mathbf{S})$$

Bishop eq 4.144

$$\mathbf{S}^{-1} = -\nabla\nabla \ln f(\mathbf{z})\Big|_{\mathbf{z}=\mathbf{z_0}}$$

Bishop eq 4.132

$$p(\theta \,|\, \mathbf{y}, \mathbf{X}) \propto \mathcal{N}(\mathbf{m_0}, \mathbf{S_0}) \cdot \prod_{n=1}^{N} g_n^{y_n} \{1 - g_n\}^{1-y_n}$$

# Laplace Approximation for Bayesian Logistic Regression

$$p(\theta \,|\, \mathbf{y}, \mathbf{X}) \simeq \mathbf{q}(\theta) = \mathcal{N}(\theta \,|\, \theta_{\mathbf{MAP}}, \mathbf{S})$$

Bishop eq 4.144

$$\mathbf{S}^{-1} = -\nabla\nabla \ln f(\mathbf{z})\Big|_{\mathbf{z}=\mathbf{z}_0}$$

Bishop eq 4.132

$$p(\theta \,|\, \mathbf{y}, \mathbf{X}) \propto \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0) \cdot \prod_{n=1}^{N} g_n^{y_n} \{1 - g_n\}^{1-y_n}$$

$$-\ln p(\theta \,|\, \mathbf{y}, \mathbf{X}) = \frac{1}{2}(\theta - \mathbf{m}_0)^T \mathbf{S}_0^{-1}(\theta - \mathbf{m}_0) - \sum_{n=1}^{N} [y_n \ln g_n + (1 - y_n)\ln(1 - g_n)] + \text{const}$$

Bishop eq 4.142

# Laplace Approximation for Bayesian Logistic Regression

$$p(\theta \mid \mathbf{y}, \mathbf{X}) \simeq \mathbf{q}(\theta) = \mathcal{N}(\theta \mid \theta_{\mathbf{MAP}}, \mathbf{S})$$

Bishop eq 4.144

$$\mathbf{S}^{-1} = -\nabla\nabla \ln f(\mathbf{z})\Big|_{\mathbf{z}=\mathbf{z_0}}$$

Bishop eq 4.132

$$p(\theta \mid \mathbf{y}, \mathbf{X}) \propto \mathcal{N}(\mathbf{m_0}, \mathbf{S_0}) \cdot \prod_{n=1}^{N} g_n^{y_n} \{1 - g_n\}^{1-y_n}$$

Function of $w$, $y_n = \sigma\left(\mathbf{w^T}\phi(\mathbf{x_n})\right)$

$$-\ln p(\theta \mid \mathbf{y}, \mathbf{X}) = \frac{1}{2}(\theta - \mathbf{m_0})^T \mathbf{S_0}^{-1}(\theta - \mathbf{m_0}) - \sum_{n=1}^{N} [y_n \ln \overbrace{g_n} + (1 - y_n)\ln(1 - \overbrace{g_n})] + \text{const}$$

Bishop eq 4.142

# Laplace Approximation for Bayesian Logistic Regression

$$p(\theta \,|\, \mathbf{y}, \mathbf{X}) \simeq \mathbf{q}(\theta) = \mathcal{N}(\theta \,|\, \theta_{\mathbf{MAP}}, \mathbf{S})$$

Bishop eq 4.144

$$\mathbf{S}^{-1} = -\nabla\nabla \ln f(\mathbf{z})\Big|_{\mathbf{z}=\mathbf{z_0}}$$

Bishop eq 4.132

$$p(\theta \,|\, \mathbf{y}, \mathbf{X}) \propto \mathcal{N}(\mathbf{m_0}, \mathbf{S_0}) \cdot \prod_{n=1}^{N} g_n^{y_n} \{1 - g_n\}^{1-y_n}$$

Function of $w$, $y_n = \sigma\left(\mathbf{w^T}\phi(\mathbf{x_n})\right)$

$$-\ln p(\theta \,|\, \mathbf{y}, \mathbf{X}) = \frac{1}{2}(\theta - \mathbf{m_0})^T \mathbf{S_0}^{-1}(\theta - \mathbf{m_0}) - \sum_{n=1}^{N} [y_n \ln \overbrace{g_n} + (1 - y_n)\ln(1 - \overbrace{g_n})] + \text{const}$$

Bishop eq 4.142

$$\mathbf{S^{-1}} = -\nabla_\theta \nabla_\theta \mathbf{p}(\theta \,|\, \mathbf{t}, \mathbf{X}) =$$

# Laplace Approximation for Bayesian Logistic Regression

$$p(\theta | \mathbf{y}, \mathbf{X}) \simeq \mathbf{q}(\theta) = \mathcal{N}(\theta | \theta_{\mathbf{MAP}}, \mathbf{S})$$

Bishop eq 4.144

$$\mathbf{S}^{-1} = -\nabla\nabla \ln f(\mathbf{z})\Big|_{\mathbf{z}=\mathbf{z}_0}$$

Bishop eq 4.132

$$p(\theta | \mathbf{y}, \mathbf{X}) \propto \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0) \cdot \prod_{n=1}^{N} g_n^{y_n} \{1 - g_n\}^{1-y_n}$$

Function of $w$, $y_n = \sigma\left(\mathbf{w}^{\mathbf{T}}\phi(\mathbf{x_n})\right)$

$$-\ln p(\theta | \mathbf{y}, \mathbf{X}) = \frac{1}{2}(\theta - \mathbf{m}_0)^T \mathbf{S}_0^{-1}(\theta - \mathbf{m}_0) - \sum_{n=1}^{N} [y_n \ln \overbrace{g_n} + (1 - y_n)\ln(1 - \overbrace{g_n})] + \text{const}$$

Bishop eq 4.142

$$\mathbf{S}^{-1} = -\nabla_\theta \nabla_\theta \mathbf{p}(\theta | \mathbf{t}, \mathbf{X}) = \quad \mathbf{S}_0^{-1}$$

# Laplace Approximation for Bayesian Logistic Regression

$$p(\theta \mid \mathbf{y}, \mathbf{X}) \simeq \mathbf{q}(\theta) = \mathcal{N}(\theta \mid \theta_{\mathbf{MAP}}, \mathbf{S})$$

Bishop eq 4.144

$$\mathbf{S}^{-1} = -\nabla\nabla \ln f(\mathbf{z})\Big|_{\mathbf{z}=\mathbf{z_0}}$$

Bishop eq 4.132

$$p(\theta \mid \mathbf{y}, \mathbf{X}) \propto \mathcal{N}(\mathbf{m_0}, \mathbf{S_0}) \cdot \prod_{n=1}^{N} g_n^{y_n} \{1 - g_n\}^{1-y_n}$$

Function of $w$, $y_n = \sigma\left(\mathbf{w}^{\mathbf{T}}\phi(\mathbf{x_n})\right)$

$$-\ln p(\theta \mid \mathbf{y}, \mathbf{X}) = \frac{1}{2}(\theta - \mathbf{m_0})^T \mathbf{S_0}^{-1}(\theta - \mathbf{m_0}) - \sum_{n=1}^{N} [y_n \ln \overbrace{g_n} + (1 - y_n)\ln(1 - \overbrace{g_n})] + \text{const}$$

Bishop eq 4.142

$$\downarrow \nabla_\theta$$

$$\sum_{n=1}^{N} (g_n - y_n)\boldsymbol{\phi_n}$$

$$\downarrow \nabla_\theta$$

$$\mathbf{S}^{-1} = -\nabla_\theta \nabla_\theta \mathbf{p}(\theta \mid \mathbf{t}, \mathbf{X}) = \mathbf{S_0^{-1}} + \sum_{n=1}^{N} g_n(1 - g_n)\boldsymbol{\phi}_n \boldsymbol{\phi}_n^T$$

Bishop eq 4.143

# Predictive Distribution for Bayesian Logistic Regression

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{\mathbf{p}(\mathbf{y}^* \mid \mathbf{x}^*, \theta)}_{\text{Sigmoid}} \, \underbrace{\mathbf{p}(\theta \mid \mathbf{y}, \mathbf{X})}_{\approx \mathcal{N}(\theta;\, \theta_{\mathbf{MAP}}, \mathbf{S})} \, \mathrm{d}\theta$$

# Predictive Distribution for Bayesian Logistic Regression

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{\mathbf{p}(\mathbf{y}^* \mid \mathbf{x}^*, \theta)}_{\text{Sigmoid}}\, \underbrace{\mathbf{p}(\theta \mid \mathbf{y}, \mathbf{X})}_{\approx \mathcal{N}(\theta;\, \theta_{\mathbf{MAP}}, \mathbf{S})}\, \mathrm{d}\theta$$

Still analytically intractable 😢

# Probit Function

Recall that, the Sigmoid “logistic” function is used in logistic regression because it falls out naturally from the idea of log odd and can be regarded as generalised linear regression

# Probit Function

Recall that, the Sigmoid “logistic” function is used in logistic regression because it falls out naturally from the idea of log odd and can be regarded as generalised linear regression

Probit function (closely related to the erf function)

$$\Phi(x) \equiv \int_{-\infty}^{x} \mathcal{N}(x \mid 0,1)\,\mathrm{d}x$$

Bishop eq 4.114

# Probit Function

Recall that, the Sigmoid “logistic” function is used in logistic regression because it falls out naturally from the idea of log odd and can be regarded as generalised linear regression

Probit function (closely related to the erf function)

$$\Phi(x) \equiv \int_{-\infty}^{x} \mathcal{N}(x \,|\, 0, 1) \, \mathrm{d}x$$

Bishop eq 4.114

is a close analog, and yet makes the previous integral possible

$$\int \Phi(\lambda a) \, \mathcal{N}(a \,|\, \mu, \sigma^2) \, \mathrm{d}a = \Phi\left( \frac{\mu}{\sqrt{\lambda^{-2} + \sigma^2}} \right)$$

Bishop eq 4.152

**Proof :** $\lambda = 1/\omega$

$$\int \Phi\left(\frac{a}{\omega}\right) \mathcal{N}(a \,|\, \mu, \sigma^2) \,\mathrm{d}a = \Phi\left(\frac{\mu}{\sqrt{\omega^2 + \sigma^2}}\right)$$

**Proof :** $\lambda = 1/\omega$

$$\int \Phi\left(\frac{a}{\omega}\right) \mathcal{N}(a \,|\, \mu, \sigma^2)\, \mathrm{d}a = \Phi\left(\frac{\mu}{\sqrt{\omega^2 + \sigma^2}}\right)$$

Let

$$X \sim \mathcal{N}(0,\omega^2), \quad Y \sim \mathcal{N}(\mu, \sigma^2)$$

**Proof :** $\lambda = 1/\omega$

$$\int \Phi\left(\frac{a}{\omega}\right) \mathcal{N}(a \,|\, \mu, \sigma^2)\, \mathrm{d}a = \Phi\left(\frac{\mu}{\sqrt{\omega^2 + \sigma^2}}\right)$$

Let

$$X \sim \mathcal{N}(0, \omega^2), \quad Y \sim \mathcal{N}(\mu, \sigma^2) \;\Rightarrow\; Z \equiv X - Y \sim \mathcal{N}(-\mu, \omega^2 + \sigma^2)$$

**Proof :** $\lambda = 1/\omega$

$$\int \Phi\left(\frac{a}{\omega}\right) \mathcal{N}(a \,|\, \mu, \sigma^2)\, \mathrm{d}a = \Phi\left(\frac{\mu}{\sqrt{\omega^2 + \sigma^2}}\right)$$

Let

$$X \sim \mathcal{N}(0,\omega^2), \quad Y \sim \mathcal{N}(\mu, \sigma^2) \;\Rightarrow\; Z \equiv X - Y \sim \mathcal{N}(-\mu, \omega^2 + \sigma^2)$$

On the one hand

$$P(X \leq Y) = P(Z \leq 0) = \Phi\left(\frac{\mu}{\sqrt{\omega^2 + \sigma^2}}\right)$$

**Proof :** $\lambda = 1/\omega$

$$\int \Phi\left(\frac{a}{\omega}\right) \mathcal{N}(a \,|\, \mu, \sigma^2)\,\mathrm{d}a = \Phi\left(\frac{\mu}{\sqrt{\omega^2 + \sigma^2}}\right)$$

Let

$$X \sim \mathcal{N}(0, \omega^2), \quad Y \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow Z \equiv X - Y \sim \mathcal{N}(-\mu, \omega^2 + \sigma^2)$$

On the one hand

$$P(X \leq Y) = P(Z \leq 0) = \Phi\left(\frac{\mu}{\sqrt{\omega^2 + \sigma^2}}\right)$$

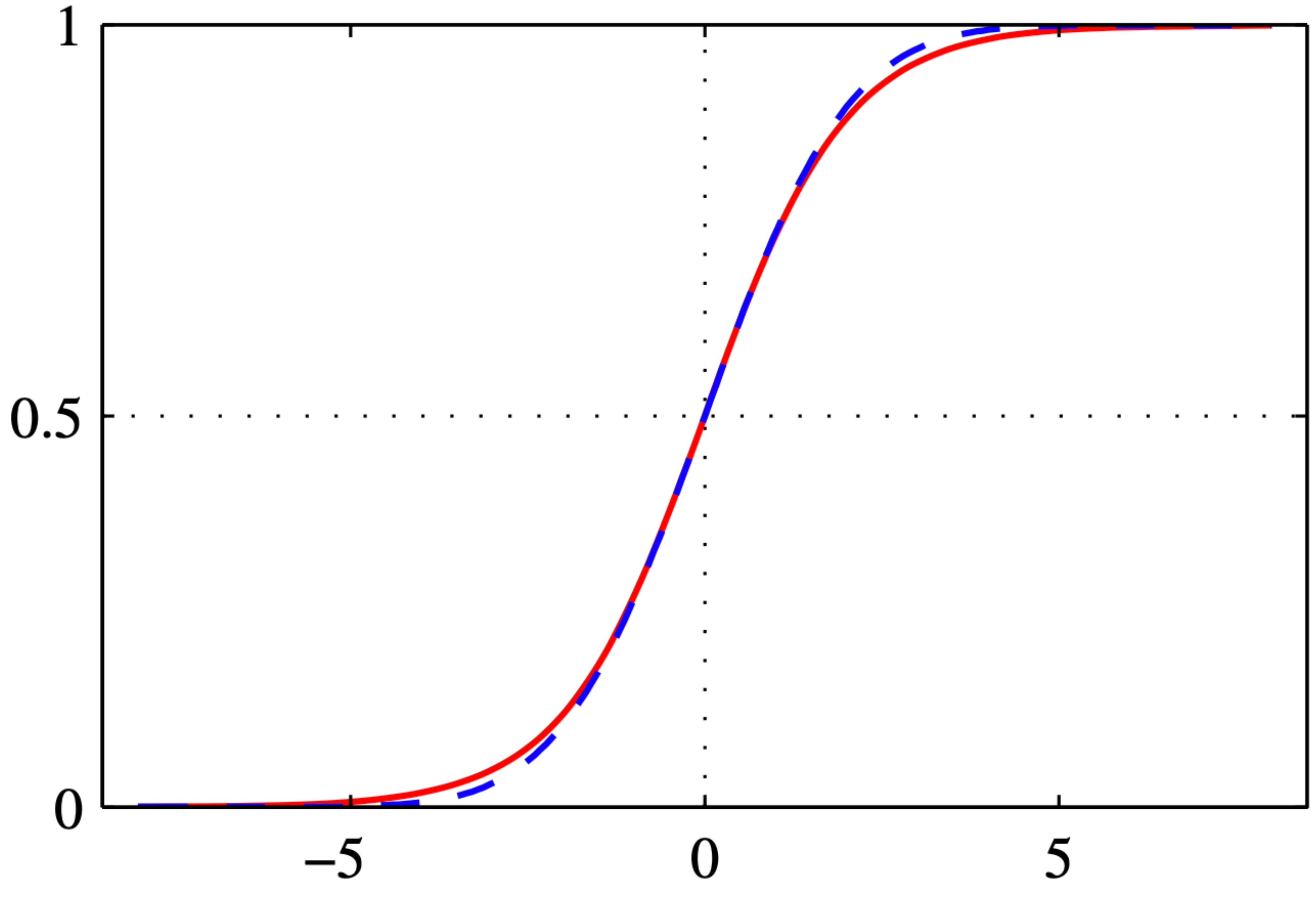
On the other hand, integrating all possible $Y$

$$P(X \leq Y) = \int_{-\infty}^{\infty} P(X \leq a) \mathcal{N}(a \,|\, \mu, \sigma^2)\,\mathrm{d}a = \int_{-\infty}^{\infty} \Phi\left(\frac{a}{\omega}\right) \mathcal{N}(a \,|\, \mu, \sigma^2)\,\mathrm{d}a$$

# Approximating Sigmoid Function with Probit Function

$$\sigma(a) \simeq \Phi(\lambda a), \quad \text{with} \quad \lambda^2 = \pi/8$$

**Figure 4.9** Plot of the logistic sigmoid function $\sigma(a)$ defined by (4.59), shown in red, together with the scaled probit function $\Phi(\lambda a)$, for $\lambda^2 = \pi/8$, shown in dashed blue, where $\Phi(a)$ is defined by (4.114). The scaling factor $\pi/8$ is chosen so that the derivatives of the two curves are equal for $a = 0$.



# Approximating Sigmoid Function with Probit Function

$$\sigma(a) \simeq \Phi(\lambda a), \quad \text{with} \quad \lambda^2 = \pi/8$$

$$\int \sigma(a)\mathcal{N}(a \,|\, \mu, \sigma^2)\mathrm{d}a \simeq \sigma(\kappa(\sigma^2)\,\mu)$$

$$\kappa(\sigma^2) = (1 + \pi\sigma^2/8)^{-1/2}$$

Bishop eq 4.153

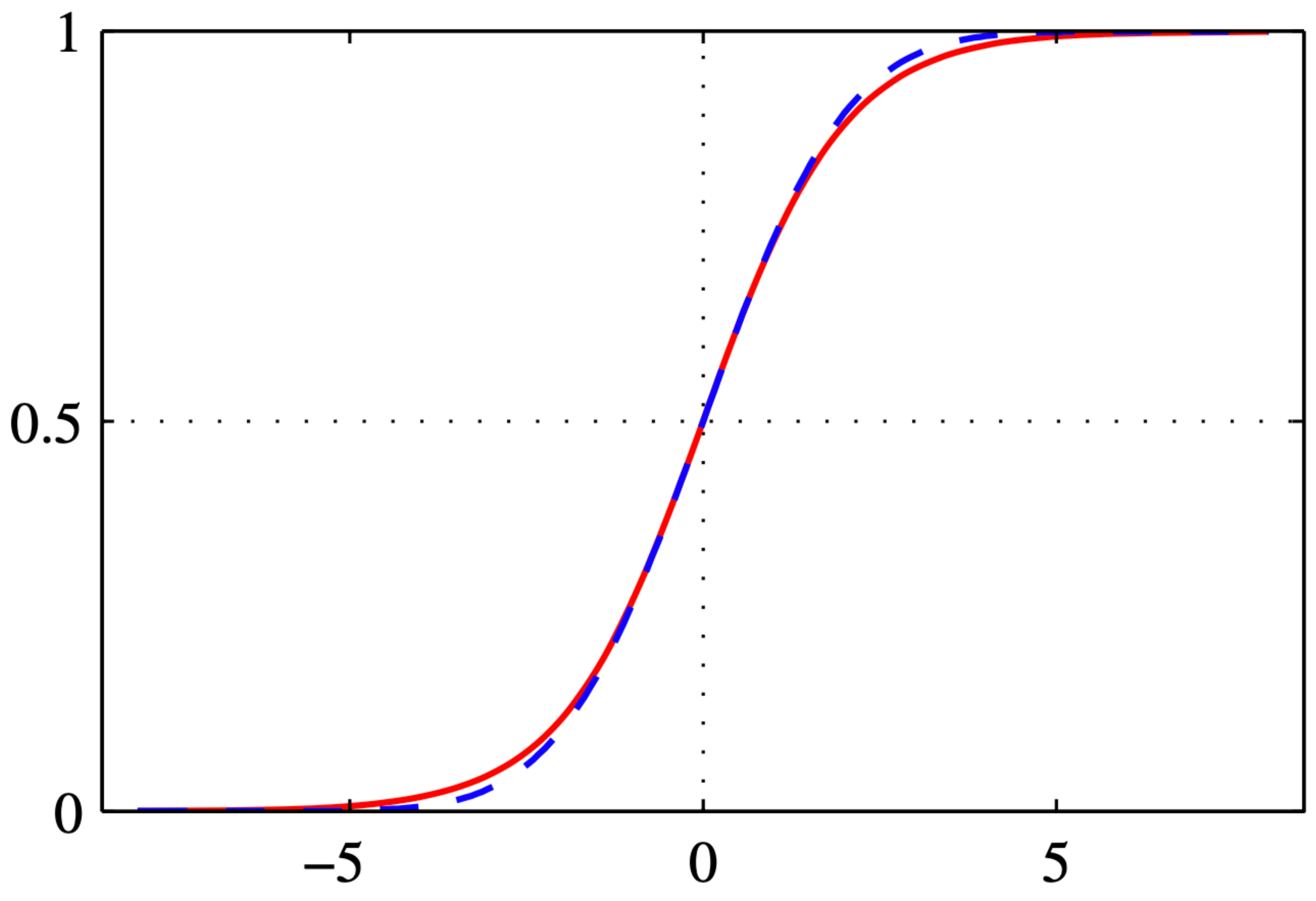



**Figure 4.9** Plot of the logistic sigmoid function $\sigma(a)$ defined by (4.59), shown in red, together with the scaled probit function $\Phi(\lambda a)$, for $\lambda^2 = \pi/8$, shown in dashed blue, where $\Phi(a)$ is defined by (4.114). The scaling factor $\pi/8$ is chosen so that the derivatives of the two curves are equal for $a = 0$.

# Predictive distribution for Bayesian Logistic Regression

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{\mathbf{p}(\mathbf{y}^* \mid \mathbf{x}^*, \theta)}_{\sigma\left(\theta^T \phi(\mathbf{x}^*)\right)} \, \underbrace{\mathbf{p}(\theta \mid \mathbf{y}, \mathbf{X})}_{\approx q(\theta) = \mathcal{N}(\theta \mid \theta_{\mathbf{MAP}}, \mathbf{S})} \, \mathrm{d}\theta$$

Let

$$a = \theta^T \phi(\mathbf{x}^*)$$

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \sigma(a) \, p(a) \, \mathrm{d}a$$

# Predictive distribution for Bayesian Logistic Regression

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{\mathbf{p}(\mathbf{y}^* \mid \mathbf{x}^*, \theta)}_{\sigma\left(\theta^T \phi(\mathbf{x}^*)\right)} \underbrace{\mathbf{p}(\theta \mid \mathbf{y}, \mathbf{X})}_{\approx q(\theta) = \mathcal{N}(\theta \mid \theta_{\mathbf{MAP}}, \mathbf{S})} \, \mathrm{d}\theta$$

Let

$$a = \theta^T \phi(\mathbf{x}^*)$$

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \sigma(a) \underbrace{p(a)}_{\text{Also Gaussian}} \, \mathrm{d}a$$

$$\mathcal{N}(a \mid \mu_a, \sigma_a^2)$$

# Predictive distribution for Bayesian Logistic Regression

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{\mathbf{p}(\mathbf{y}^* \mid \mathbf{x}^*, \theta)}_{\sigma\left(\theta^T \phi(\mathbf{x}^*)\right)} \underbrace{\mathbf{p}(\theta \mid \mathbf{y}, \mathbf{X})}_{\approx q(\theta) = \mathcal{N}(\theta \mid \theta_{\mathbf{MAP}}, \mathbf{S})} \, \mathrm{d}\theta$$

Let

$$a = \theta^T \phi(\mathbf{x}^*)$$

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \sigma(a) \underbrace{p(a)}_{\text{Also Gaussian}} \, \mathrm{d}a$$

$$\mathcal{N}(a \mid \mu_a, \sigma_a^2)$$

$$\mu_a = \theta_{\mathrm{MAP}}^T \phi(\mathbf{x}^*)$$

$$\sigma_a^2 = \phi^T(\mathbf{x}^*) \, \mathbf{S} \, \phi(\mathbf{x}^*)$$

Bishop eq 4.147, 4.151

# Predictive distribution for Bayesian Logistic Regression

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{\mathbf{p}(\mathbf{y}^* \mid \mathbf{x}^*, \theta)}_{\sigma\left(\theta^T \phi(\mathbf{x}^*)\right)} \underbrace{\mathbf{p}(\theta \mid \mathbf{y}, \mathbf{X})}_{\approx q(\theta) = \mathcal{N}(\theta \mid \theta_{\mathbf{MAP}}, \mathbf{S})} \, \mathrm{d}\theta$$

Let

$$a = \theta^T \phi(\mathbf{x}^*)$$

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \sigma(a) \underbrace{p(a)}_{\text{Also Gaussian } \mathcal{N}(a \mid \mu_a, \sigma_a^2)} \mathrm{d}a$$

$$\mu_a = \theta_{\mathrm{MAP}}^T \phi(\mathbf{x}^*)$$

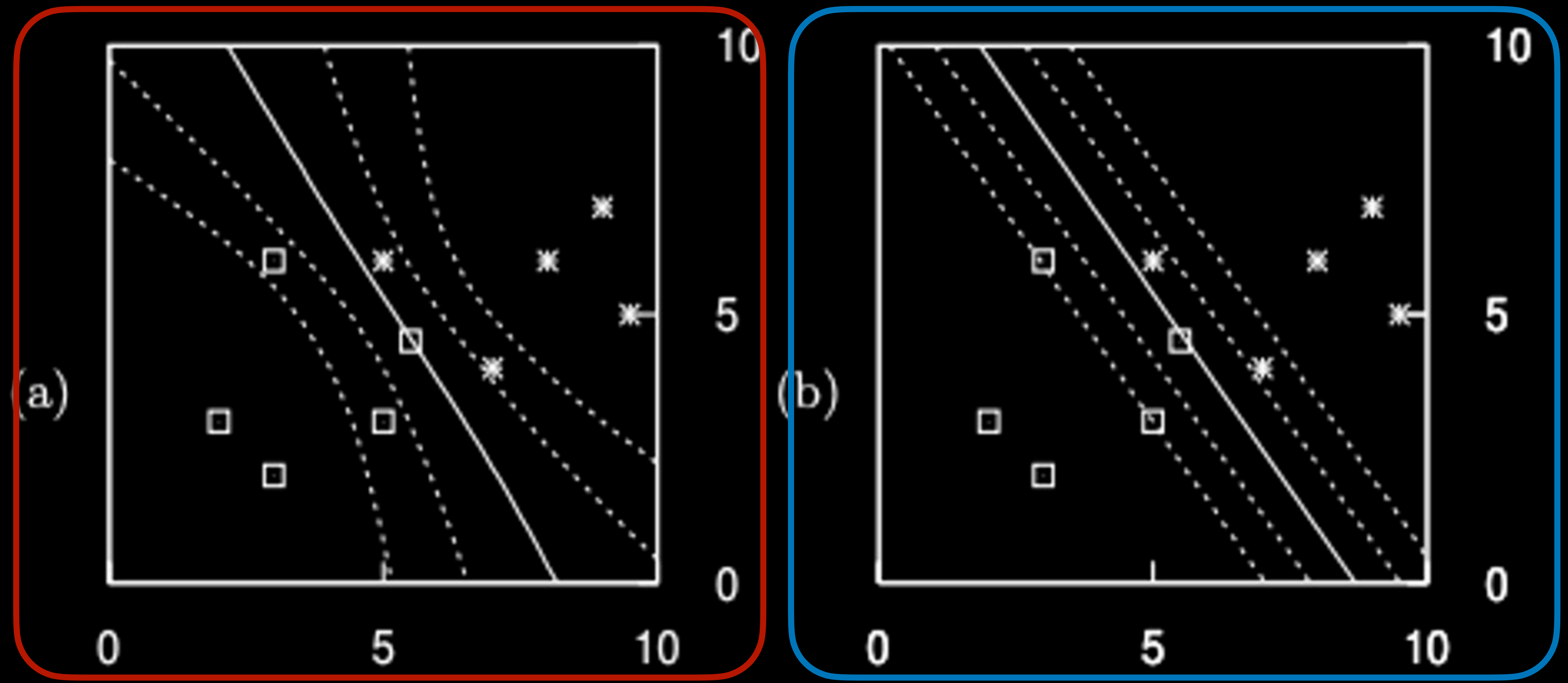
$$\sigma_a^2 = \phi^T(\mathbf{x}^*) \, \mathbf{S} \, \phi(\mathbf{x}^*)$$

Bishop eq 4.147, 4.151

$$\simeq \sigma\left(\left(1 + \frac{\pi \sigma_a^2}{8}\right)^{-1/2} \mu_a\right)$$

Bishop eq 4.153

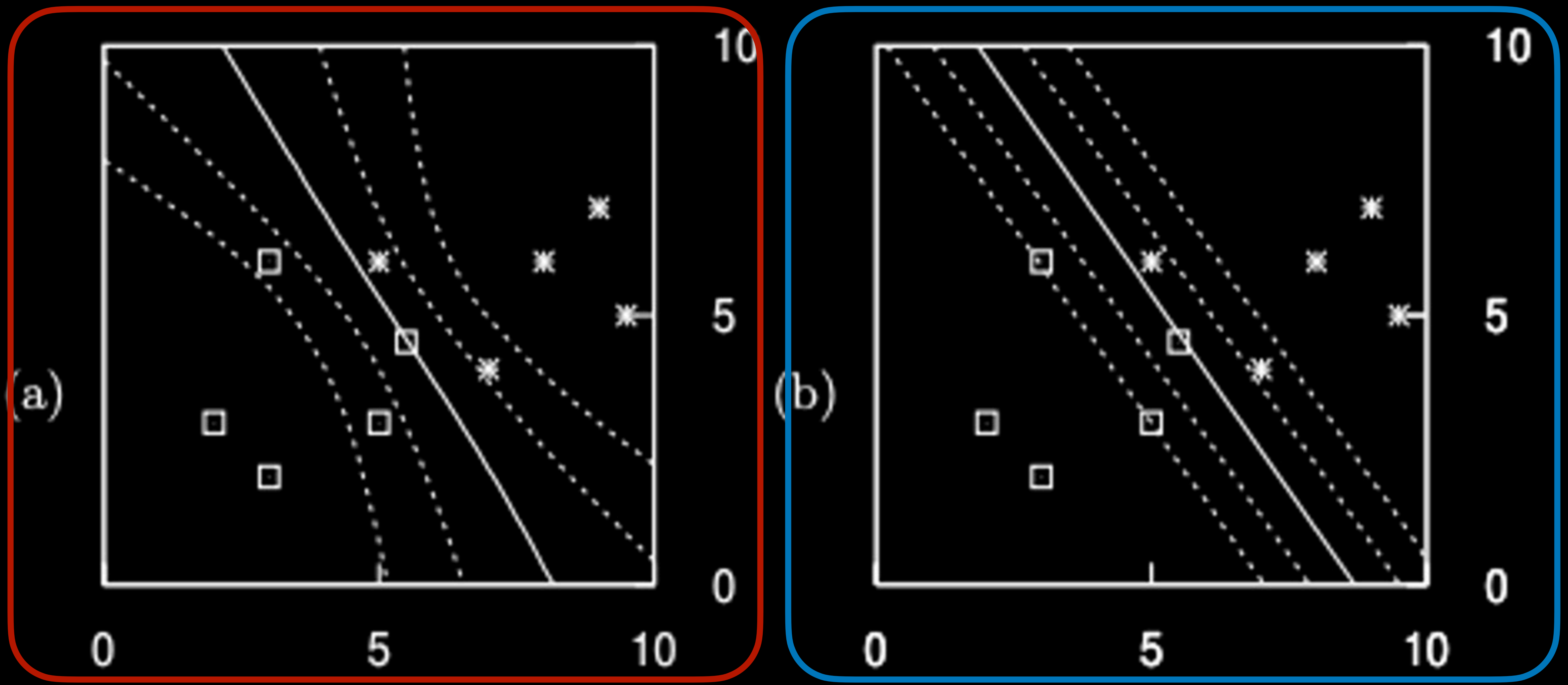
# Predictive Distribution with Laplace Approximation




Again, the axes are the input features $x_1$ and $x_2$. The right hand figure shows $P(y=1 \mid \mathbf{x}, \mathbf{w}^*)$ for some fitted weights $\mathbf{w}^*$. No matter how these fitted weights are chosen, the contours have to be linear. The parallel contours mean that the uncertainty of predictions falls at the same rate when moving away from the decision boundary, no matter how far we are from the training inputs.

Example courtesy of Edinburgh MLPR course
https://www.inf.ed.ac.uk/teaching/courses/mlpr/2016/notes/w8a_bayes_logistic_regression_laplace.pdf

# Predictive Distribution with Laplace Approximation



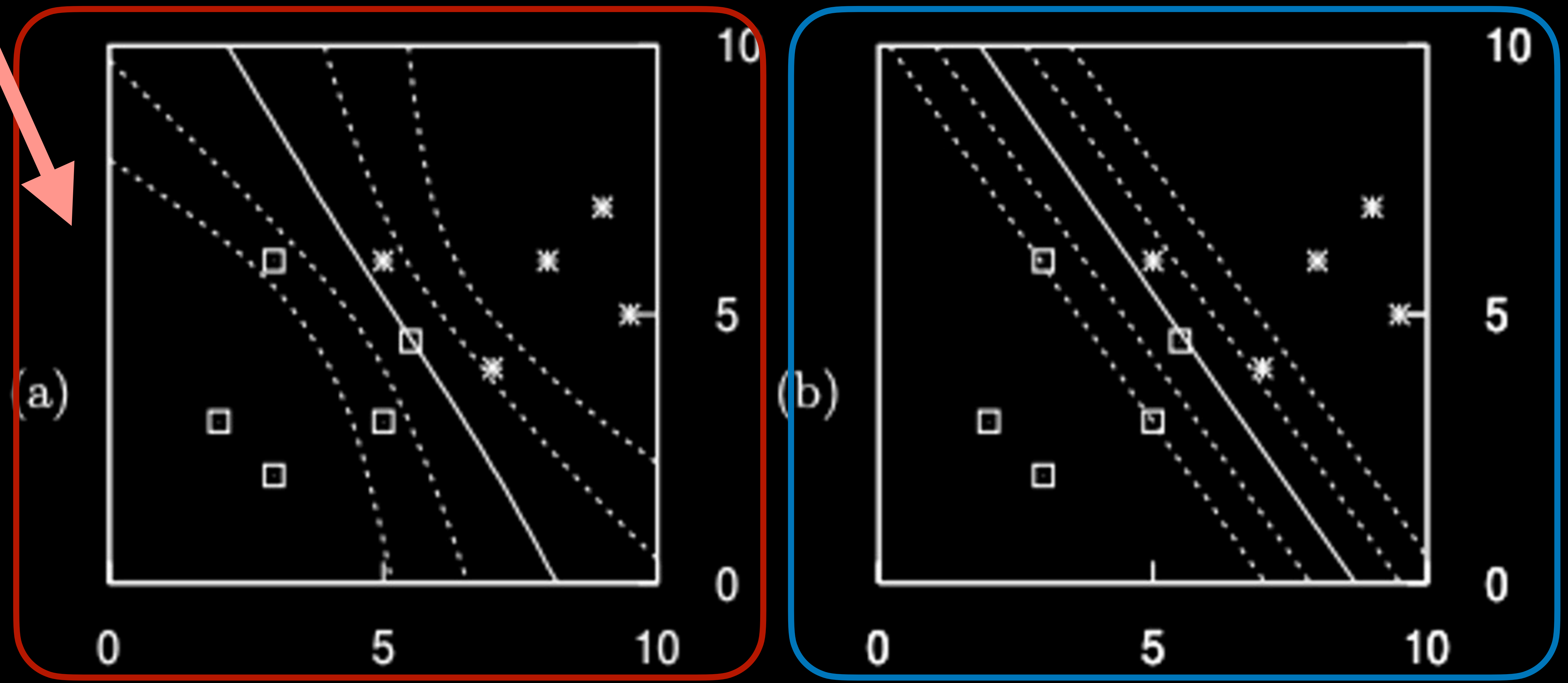

$p(\mathbf{y}^* \,|\, \mathbf{x}^*, \theta_{\text{MAP}})$

Example courtesy of Edinburgh MLPR course
https://www.inf.ed.ac.uk/teaching/courses/mlpr/2016/notes/w8a_bayes_logistic_regression_laplace.pdf

Again, the axes are the input features $x_1$ and $x_2$. The right hand figure shows $P(y=1 \,|\, \mathbf{x}, \mathbf{w}^*)$ for some fitted weights $\mathbf{w}^*$. No matter how these fitted weights are chosen, the contours have to be linear. The parallel contours mean that the uncertainty of predictions falls at the same rate when moving away from the decision boundary, no matter how far we are from the training inputs.

# Predictive Distribution with Laplace Approximation

$p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y})$




$p(\mathbf{y}^* \mid \mathbf{x}^*, \theta_{\text{MAP}})$

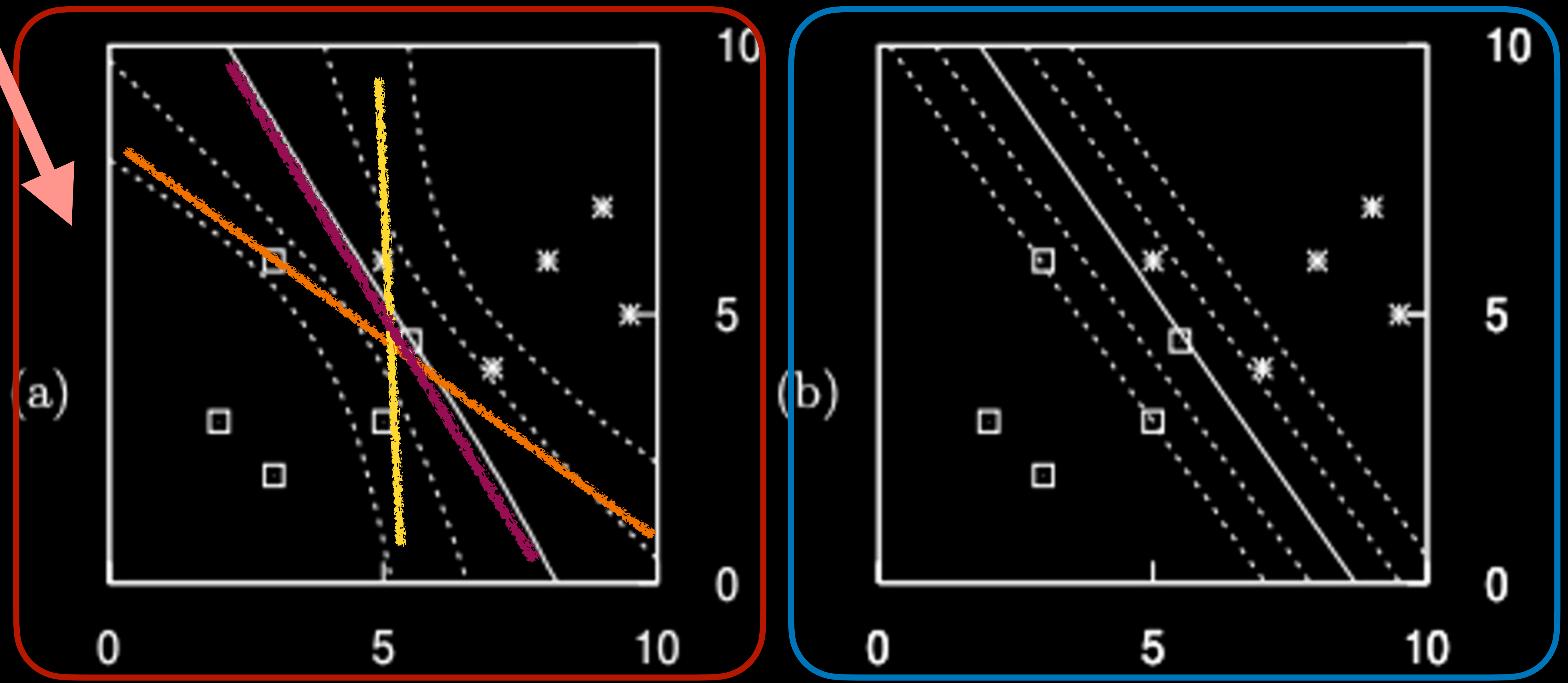
Example courtesy of Edinburgh MLPR course https://www.inf.ed.ac.uk/teaching/courses/mlpr/2016/notes/w8a_bayes_logistic_regression_laplace.pdf

Again, the axes are the input features $x_1$ and $x_2$. The right hand figure shows $P(y=1 \mid \mathbf{x}, \mathbf{w}^*)$ for some fitted weights $\mathbf{w}^*$. No matter how these fitted weights are chosen, the contours have to be linear. The parallel contours mean that the uncertainty of predictions falls at the same rate when moving away from the decision boundary, no matter how far we are from the training inputs.

# Predictive Distribution with Laplace Approximation

$p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y})$

$p(\mathbf{y}^* \mid \mathbf{x}^*, \theta_{\mathrm{MAP}})$




Example courtesy of Edinburgh MLPR course
https://www.inf.ed.ac.uk/teaching/courses/mlpr/2016/notes/w8a_bayes_logistic_regression_laplace.pdf

Again, the axes are the input features $x_1$ and $x_2$. The right hand figure shows $P(y = 1 \mid \mathbf{x}, \mathbf{w}^*)$ for some fitted weights $\mathbf{w}^*$. No matter how these fitted weights are chosen, the contours have to be linear. The parallel contours mean that the uncertainty of predictions falls at the same rate when moving away from the decision boundary, no matter how far we are from the training inputs.

# Gaussian Process for Classification

Recall in Gaussian Process Regression

$$p(\mathbf{f} \mid \mathbf{X}) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K})$$

# Gaussian Process for Classification

Recall in Gaussian Process Regression

$$p(\mathbf{f} \mid \mathbf{X}) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K})$$

we have $y_n \in \{0,1\}$ in classification. We use the same approach as in logistic classification

# Gaussian Process for Classification

Recall in Gaussian Process Regression

$$p(\mathbf{f} \mid \mathbf{X}) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K})$$

we have $y_n \in \{0,1\}$ in classification. We use the same approach as in logistic classification

Key idea — using GP as a intermediate step

# Gaussian Process for Classification

Recall in Gaussian Process Regression

$$p(\mathbf{f}\,|\,\mathbf{X}) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K})$$

we have $y_n \in \{0,1\}$ in classification. We use the same approach as in logistic classification

Key idea — using GP as a intermediate step

Recall in logistic regression

$$g(\mathbf{x}) = \sigma\left(\theta^{\mathbf{T}}\phi(\mathbf{x})\right)$$

# Gaussian Process for Classification

Recall in Gaussian Process Regression

$$p(\mathbf{f} \mid \mathbf{X}) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K})$$

we have $y_n \in \{0,1\}$ in classification. We use the same approach as in logistic classification

Key idea — using GP as a intermediate step

And the "squash" it with a sigmoid function

$$g(\mathbf{x}) = \sigma\big(f(\mathbf{x})\big)$$

$$p(y_n \mid f, \mathbf{x}_n) = \sigma\left(f(\mathbf{x_n})\right)^{y_n}\left(1 - \sigma(f(\mathbf{x_n}))\right)^{1-y_n}$$

Bishop eq 6.73

Recall in logistic regression

$$g(\mathbf{x}) = \sigma\big(\theta^{\mathbf{T}}\phi(\mathbf{x})\big)$$

# Gaussian Process for Classification

Recall in Gaussian Process Regression

$$p(\mathbf{f}\,|\,\mathbf{X}) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K})$$

we have $y_n \in \{0,1\}$ in classification. We use the same approach as in logistic classification

Key idea — using GP as a intermediate step

And the "squash" it with a sigmoid function

$$g(\mathbf{x}) = \sigma\big(f(\mathbf{x})\big)$$

$$p(y_n\,|\,f, \mathbf{x}_n) = \sigma\left(\boxed{f(\mathbf{x_n})}\right)^{y_n}\left(1 - \sigma(f(\mathbf{x_n}))\right)^{1-y_n}$$

Bishop eq 6.73

Recall in logistic regression

$$g(\mathbf{x}) = \sigma\left(\boxed{\theta^{\mathbf{T}}\phi(\mathbf{x})}\right)$$

Substituting the linear model with Gaussian Process

# Gaussian Process for Classification

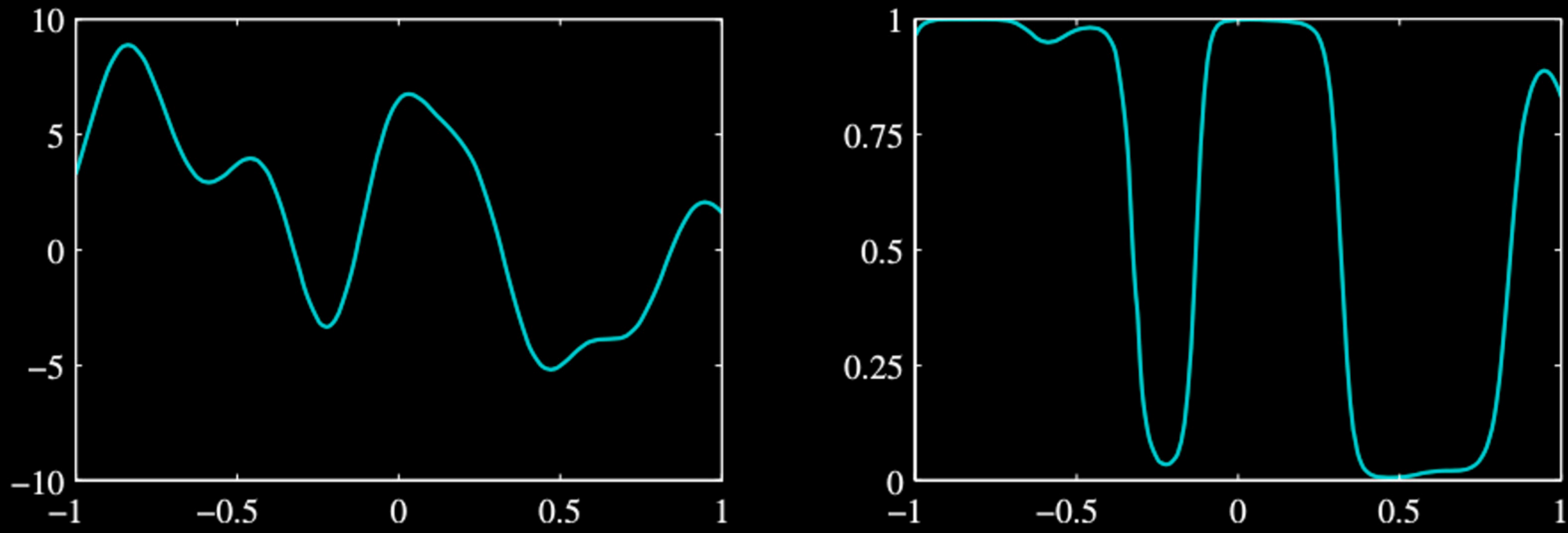



**Figure 6.11** The left plot shows a sample from a Gaussian process prior over functions $a(\mathbf{x})$, and the right plot shows the result of transforming this sample using a logistic sigmoid function.

# Gaussian Process for Classification

Input Variable

$$[\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n]$$

# Gaussian Process for Classification

Input Variable

$$[\mathbf{x_1}, \mathbf{x_2}, \ldots \mathbf{x_n}]$$

Gaussian
Process

Intermediate Variable

$$\mathbf{f} = [f_1, f_2, \ldots f_n]$$

# Gaussian Process for Classification

Input Variable

$$[\mathbf{x_1}, \mathbf{x_2}, \ldots \mathbf{x_n}]$$

Gaussian Process

Response Function, e.g. Sigmoid

Intermediate Variable

$$\mathbf{f} = [f_1, f_2, \ldots f_n]$$

Target Variable

$$\mathbf{y} = [y_1, y_2, \ldots y_n]$$

# Gaussian Process for the Intermediate Variable

Bishop eq 6.74

$$p(\mathbf{f}\,|\,\mathbf{X}) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \underbrace{\mathbf{K}}_{\text{Kernel (X, X)}})$$

Bishop eq 6.75

$$\mathbf{K}_{nm} = k(\mathbf{x}_n, \mathbf{x}_m \,|\, \underbrace{\boldsymbol{\theta}}_{\text{Kernel hyperparameters}})$$

# Gaussian Process for the Intermediate Variable

Bishop eq 6.74

$$p(\mathbf{f}\,|\,\mathbf{X}) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \underbrace{\mathbf{K}}_{\text{Kernel (X, X)}})$$

Bishop eq 6.75

$$\mathbf{K}_{nm} = k(\mathbf{x}_n, \mathbf{x}_m\,|\,\underbrace{\boldsymbol{\theta}}_{\text{Kernel hyperparameters}})$$

---

Recall from Gaussian Process Regression that

$$\boxed{p(f^*\,|\,\mathbf{x}^*, \mathbf{X}, \mathbf{y})} = \mathcal{N}(f^*\,|\,m(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$

# Gaussian Process for the Intermediate Variable

Bishop eq 6.74

$$p(\mathbf{f}\,|\,\mathbf{X}) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \underbrace{\mathbf{K}}_{\text{Kernel (X, X)}})$$

Bishop eq 6.75

$$\mathbf{K}_{nm} = k(\mathbf{x}_n, \mathbf{x}_m\,|\,\underbrace{\boldsymbol{\theta}}_{\text{Kernel hyperparameters}})$$

---

Recall from Gaussian Process Regression that

$$\boxed{p(f^*\,|\,\mathbf{x}^*, \mathbf{X}, \mathbf{y})} = \mathcal{N}(f^*\,|\,m(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$

$$m(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{X})\,(k(\mathbf{X}, \mathbf{X})^{-1} + \sigma^2 \mathrm{I}_N)^{-1}\,\mathbf{y}$$

$$\sigma^2(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})\,(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I})^{-1}\,k(\mathbf{X}, \mathbf{x}^*)$$

Bishop eq 6.78

# Laplace Approximation for Gaussian Process Classification

$$\ln p(\mathbf{f} \mid \mathbf{X}, \mathbf{y}) = \ln p(\mathbf{y} \mid \mathbf{f}, \mathbf{X}) + \ln p(\mathbf{f} \mid \mathbf{X}) + \text{const}\,.$$

# Laplace Approximation for Gaussian Process Classification

$$\ln p(\mathbf{f} \mid \mathbf{X}, \mathbf{y}) = \ln p(\mathbf{y} \mid \mathbf{f}, \mathbf{X}) + \ln p(\mathbf{f} \mid \mathbf{X}) + \text{const}\,.$$

$$\mathbf{y}^T\mathbf{f} - \sum_{n=1}^{N} \ln(1 + e^{f_n})$$

Bishop eq 6.79

# Laplace Approximation for Gaussian Process Classification

$$\ln p(\mathbf{f} \mid \mathbf{X}, \mathbf{y}) = \ln p(\mathbf{y} \mid \mathbf{f}, \mathbf{X}) + \ln p(\mathbf{f} \mid \mathbf{X}) + \text{const}\,.$$

$$\mathbf{y}^T\mathbf{f} - \sum_{n=1}^{N} \ln(1 + e^{f_n})$$

Bishop eq 6.79

$$-\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f} - \frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{K}|$$

Bishop eq 6.80, GP Book eq 3.12

# Laplace Approximation for Gaussian Process Classification

$$\ln p(\mathbf{f}\,|\,\mathbf{X},\mathbf{y}) = \ln p(\mathbf{y}\,|\,\mathbf{f},\mathbf{X}) + \ln p(\mathbf{f}\,|\,\mathbf{X}) + \text{const}\,.$$

$$\boxed{\mathbf{y}^T\mathbf{f}} - \boxed{\sum_{n=1}^{N}\ln(1+e^{f_n})} \qquad -\boxed{\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f}} - \frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{K}|$$

Bishop eq 6.80, GP Book eq 3.12

Bishop eq 6.79

$$\nabla_{\mathbf{f}}\ln p(\mathbf{f}\,|\,\mathbf{X},\mathbf{y}) = \boxed{\mathbf{y}} - \boxed{\sigma(\mathbf{f})} - \boxed{\mathbf{K}^{-1}\mathbf{f}} = 0$$

Bishop eq 6.81, GP Book eq 3.13, 3.15

# Laplace Approximation for Gaussian Process Classification

$$\ln p(\mathbf{f}\,|\,\mathbf{X},\mathbf{y}) = \ln p(\mathbf{y}\,|\,\mathbf{f},\mathbf{X}) + \ln p(\mathbf{f}\,|\,\mathbf{X}) + \text{const}\,.$$

$$\boxed{\mathbf{y}^T\mathbf{f}} - \boxed{\sum_{n=1}^{N} \ln(1+e^{f_n})}$$

Bishop eq 6.79

$$\boxed{-\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f}} - \frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{K}|$$

Bishop eq 6.80, GP Book eq 3.12

$$\nabla_{\mathbf{f}} \ln p(\mathbf{f}\,|\,\mathbf{X},\mathbf{y}) = \boxed{\mathbf{y}} - \boxed{\sigma(\mathbf{f})} - \boxed{\mathbf{K}^{-1}\mathbf{f}} = 0$$

Bishop eq 6.81, GP Book eq 3.13, 3.15

$$\Rightarrow \quad \mathbf{f}_{\text{MAP}} = \mathbf{K}(\mathbf{y} - \sigma(\mathbf{f}_{\text{MAP}}))$$

Bishop eq 6.84, GP Book eq 3.17

# Laplace Approximation for Gaussian Process Classification

$$\ln p(\mathbf{f}\,|\,\mathbf{X},\mathbf{y}) = \ln p(\mathbf{y}\,|\,\mathbf{f},\mathbf{X}) + \ln p(\mathbf{f}\,|\,\mathbf{X}) + \text{const}\,.$$

$$\boxed{\mathbf{y}^T\mathbf{f}} - \boxed{\sum_{n=1}^{N} \ln(1+e^{f_n})}$$

Bishop eq 6.79

$$-\boxed{\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f}} - \frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{K}|$$

Bishop eq 6.80, GP Book eq 3.12

$$\nabla_{\mathbf{f}} \ln p(\mathbf{f}\,|\,\mathbf{X},\mathbf{y}) = \boxed{\mathbf{y}} - \boxed{\sigma(\mathbf{f})} - \boxed{\mathbf{K}^{-1}\mathbf{f}} = 0$$

Bishop eq 6.81, GP Book eq 3.13, 3.15

$$\Rightarrow \quad \mathbf{f}_{\mathrm{MAP}} = \mathbf{K}(\mathbf{y} - \sigma(\mathbf{f}_{\mathrm{MAP}}))$$

Bishop eq 6.84, GP Book eq 3.17

**Implicit function of $\mathbf{f}_{\mathrm{MAP}}$, e.g. solve with Newton's method**

See GP Book eq 3.18-3.19, Algorithm 3.1

# Laplace Approximation for Gaussian Process Classification

$$\ln p(\mathbf{f} \,|\, \mathbf{X}, \mathbf{y}) = \ln p(\mathbf{y} \,|\, \mathbf{f}, \mathbf{X}) + \ln p(\mathbf{f} \,|\, \mathbf{X}) + \text{const}\,.$$

$\Rightarrow$ $$\nabla_{\mathbf{f}} \ln p(\mathbf{f} \,|\, \mathbf{X}, \mathbf{y}) = \mathbf{y} - \sigma(\mathbf{f}) - \mathbf{K}^{-1}\mathbf{f} = 0$$

# Laplace Approximation for Gaussian Process Classification

$$\ln p(\mathbf{f} \,|\, \mathbf{X}, \mathbf{y}) = \ln p(\mathbf{y} \,|\, \mathbf{f}, \mathbf{X}) + \ln p(\mathbf{f} \,|\, \mathbf{X}) + \text{const}\,.$$

$$\Rightarrow \quad \nabla_{\mathbf{f}} \ln p(\mathbf{f} \,|\, \mathbf{X}, \mathbf{y}) = \mathbf{y} - \boxed{\sigma(\mathbf{f})} - \boxed{\mathbf{K}^{-1}\mathbf{f}} = 0$$

$$\Rightarrow \quad -\nabla_{\mathbf{f}} \nabla_{\mathbf{f}} \ln p(\mathbf{f} \,|\, \mathbf{X}, \mathbf{y}) \Big|_{\mathbf{f}_{\text{MAP}}} = \boxed{\text{diag}\left\{\sigma(\mathbf{f}_{\text{MAP}})(1 - \sigma\left(\mathbf{f}_{\text{MAP}}\right))\right\}} + \boxed{\mathbf{K}^{-1}} \equiv \mathbf{H}$$

Bishop eq 6.82, GP Book eq 3.14, 3.15

# Laplace Approximation for Gaussian Process Classification

$$\ln p(\mathbf{f} \mid \mathbf{X}, \mathbf{y}) = \ln p(\mathbf{y} \mid \mathbf{f}, \mathbf{X}) + \ln p(\mathbf{f} \mid \mathbf{X}) + \text{const}\,.$$

$$\Rightarrow \quad \nabla_{\mathbf{f}} \ln p(\mathbf{f} \mid \mathbf{X}, \mathbf{y}) = \mathbf{y} - \boxed{\sigma(\mathbf{f})} - \boxed{\mathbf{K}^{-1}\mathbf{f}} = 0$$

Factorizes into a diagonal matrix because the sample is i.i.d

$$\Rightarrow \quad -\nabla_{\mathbf{f}} \nabla_{\mathbf{f}} \ln p(\mathbf{f} \mid \mathbf{X}, \mathbf{y}) \Big|_{\mathbf{f}_{\mathrm{MAP}}} = \boxed{\mathrm{diag}\left\{\sigma(\mathbf{f}_{\mathrm{MAP}})\left(1 - \sigma\left(\mathbf{f}_{\mathrm{MAP}}\right)\right)\right\}} + \boxed{\mathbf{K}^{-1}} \equiv \mathbf{H}$$

Bishop eq 6.82, GP Book eq 3.14, 3.15

# Laplace Approximation for Gaussian Process Classification

$$\ln p(\mathbf{f} \mid \mathbf{X}, \mathbf{y}) = \ln p(\mathbf{y} \mid \mathbf{f}, \mathbf{X}) + \ln p(\mathbf{f} \mid \mathbf{X}) + \text{const}\,.$$

$$\Rightarrow \nabla_{\mathbf{f}} \ln p(\mathbf{f} \mid \mathbf{X}, \mathbf{y}) = \mathbf{y} - \boxed{\sigma(\mathbf{f})} - \boxed{\mathbf{K}^{-1}\mathbf{f}} = 0$$

Factorizes into a diagonal matrix because the sample is i.i.d

$$\Rightarrow -\nabla_{\mathbf{f}} \nabla_{\mathbf{f}} \ln p(\mathbf{f} \mid \mathbf{X}, \mathbf{y}) \Big|_{\mathbf{f}_{\mathrm{MAP}}} = \boxed{\mathrm{diag}\left\{\sigma(\mathbf{f}_{\mathrm{MAP}})(1 - \sigma\left(\mathbf{f}_{\mathrm{MAP}}\right))\right\}} + \boxed{\mathbf{K}^{-1}} \equiv \mathbf{H}$$

Bishop eq 6.82, GP Book eq 3.14, 3.15

Laplace Approximation $\quad q(\mathbf{f}) \simeq p(\mathbf{f} \mid \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mathbf{f}; \mathbf{f}_{\mathrm{MAP}}, \mathbf{H}^{-1})$

Bishop eq 6.86

# One More Layer Up …

$$p(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int p(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{f})\, p(\mathbf{f} \mid \mathbf{X}, \mathbf{y})\, \mathrm{d}\mathbf{f}$$

# One More Layer Up …

$$p(f^* \,|\, \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int p(f^* \,|\, \mathbf{x}^*, \mathbf{X}, \mathbf{f})\, \underbrace{p(\mathbf{f} \,|\, \mathbf{X}, \mathbf{y})}_{\mathcal{N}(\mathbf{f};\, \mathbf{f}_{\mathrm{MAP}},\, \mathbf{H}^{-1})}\, \mathrm{d}\mathbf{f}$$

$$\mathbf{f}_{\mathrm{MAP}} = k(\mathbf{X}, \mathbf{X})(\mathbf{y} - \sigma(\mathbf{f}_{\mathrm{MAP}}))$$

$$\mathbf{H} = \mathrm{diag}\left\{\sigma(\mathbf{f}_{\mathrm{MAP}})(1 - \sigma(\mathbf{f}_{\mathrm{MAP}}))\right\} + k(\mathbf{X}, \mathbf{X})^{-1}$$

# One More Layer Up …

$$p(f^* \,|\, \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{p(f^* \,|\, \mathbf{x}^*, \mathbf{X}, \mathbf{f})}_{\mathcal{N}(f^*;\, m(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))} \underbrace{p(\mathbf{f} \,|\, \mathbf{X}, \mathbf{y})}_{\mathcal{N}(\mathbf{f};\, \mathbf{f}_{\mathrm{MAP}}, \mathbf{H}^{-1})} \, \mathrm{d}\mathbf{f}$$

$$m(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{X})\, k(\mathbf{X}, \mathbf{X})^{-1}\, \mathbf{f}$$

$$\sigma^2(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})\, k(\mathbf{X}, \mathbf{X})^{-1}\, k(\mathbf{X}, \mathbf{x}^*)$$

$$\mathbf{f}_{\mathrm{MAP}} = k(\mathbf{X}, \mathbf{X})(\mathbf{y} - \sigma(\mathbf{f}_{\mathrm{MAP}}))$$

$$\mathbf{H} = \mathrm{diag}\left\{\sigma(\mathbf{f}_{\mathrm{MAP}})(1 - \sigma(\mathbf{f}_{\mathrm{MAP}}))\right\} + k(\mathbf{X}, \mathbf{X})^{-1}$$

# One More Layer Up …

$$p(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{p(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{f})}_{\mathcal{N}(f^*;\, m(\mathbf{x}^*),\, \sigma^2(\mathbf{x}^*))} \underbrace{p(\mathbf{f} \mid \mathbf{X}, \mathbf{y})}_{\mathcal{N}(\mathbf{f};\, \mathbf{f}_{\mathrm{MAP}},\, \mathbf{H}^{-1})} \,\mathrm{d}\mathbf{f} \qquad = \mathcal{N}(\mathbf{f}^*; c, d^2)$$

$$m(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{X})\, k(\mathbf{X}, \mathbf{X})^{-1}\, \mathbf{f}$$

$$\sigma^2(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})\, k(\mathbf{X}, \mathbf{X})^{-1}\, k(\mathbf{X}, \mathbf{x}^*)$$

$$\mathbf{f}_{\mathrm{MAP}} = k(\mathbf{X}, \mathbf{X})(\mathbf{y} - \sigma(\mathbf{f}_{\mathrm{MAP}}))$$

$$\mathbf{H} = \mathrm{diag}\left\{\sigma(\mathbf{f}_{\mathrm{MAP}})(1 - \sigma(\mathbf{f}_{\mathrm{MAP}}))\right\} + k(\mathbf{X}, \mathbf{X})^{-1}$$

# One More Layer Up …

$$p(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{p(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{f})}_{\mathcal{N}(f^*; m(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))} \underbrace{p(\mathbf{f} \mid \mathbf{X}, \mathbf{y})}_{\mathcal{N}(\mathbf{f}; \mathbf{f}_{\mathrm{MAP}}, \mathbf{H}^{-1})} \, \mathrm{d}\mathbf{f} \qquad = \mathcal{N}(\mathbf{f}^*; c, d^2)$$

$$m(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{X})\, k(\mathbf{X}, \mathbf{X})^{-1}\, \mathbf{f}$$

$$\sigma^2(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})\, k(\mathbf{X}, \mathbf{X})^{-1}\, k(\mathbf{X}, \mathbf{x}^*)$$

$$\mathbf{f}_{\mathrm{MAP}} = k(\mathbf{X}, \mathbf{X})(\mathbf{y} - \sigma(\mathbf{f}_{\mathrm{MAP}}))$$

$$\mathbf{H} = \mathrm{diag}\left\{\sigma(\mathbf{f}_{\mathrm{MAP}})(1 - \sigma(\mathbf{f}_{\mathrm{MAP}}))\right\} + k(\mathbf{X}, \mathbf{X})^{-1}$$

Bishop eq 2.115

# One More Layer Up …

$$p(f^* \,|\, \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{p(f^* \,|\, \mathbf{x}^*, \mathbf{X}, \mathbf{f})}_{\mathcal{N}(f^*;\, m(\mathbf{x}^*),\, \sigma^2(\mathbf{x}^*))} \underbrace{p(\mathbf{f} \,|\, \mathbf{X}, \mathbf{y})}_{\mathcal{N}(\mathbf{f};\, \mathbf{f}_{\text{MAP}},\, \mathbf{H}^{-1})} \, \mathrm{d}\mathbf{f} \qquad = \mathcal{N}(\mathbf{f}^*; c, d^2)$$

$$m(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{X})\, k(\mathbf{X}, \mathbf{X})^{-1}\, \mathbf{f}$$

$$\sigma^2(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})\, k(\mathbf{X}, \mathbf{X})^{-1}\, k(\mathbf{X}, \mathbf{x}^*)$$

$$\mathbf{f}_{\text{MAP}} = k(\mathbf{X}, \mathbf{X})(\mathbf{y} - \sigma(\mathbf{f}_{\text{MAP}}))$$

$$\mathbf{H} = \operatorname{diag}\left\{\sigma(\mathbf{f}_{\text{MAP}})(1 - \sigma(\mathbf{f}_{\text{MAP}}))\right\} + k(\mathbf{X}, \mathbf{X})^{-1}$$

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}, \mathbf{H}^{-1})$$

$$p(y \,|\, \mathbf{x}) = \mathcal{N}(\mathbf{y} \,|\, \mathbf{A}\mathbf{x} + b, \sigma^2)$$

Bishop eq 2.115

# One More Layer Up …

$$p(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{p(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{f})}_{\mathcal{N}(f^*;\, m(\mathbf{x}^*),\, \sigma^2(\mathbf{x}^*))} \underbrace{p(\mathbf{f} \mid \mathbf{X}, \mathbf{y})}_{\mathcal{N}(\mathbf{f};\, \mathbf{f}_{\mathrm{MAP}},\, \mathbf{H}^{-1})} \,\mathrm{d}\mathbf{f} \qquad = \mathcal{N}(\mathbf{f}^*; c, d^2)$$

$$m(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{X})\, k(\mathbf{X}, \mathbf{X})^{-1}\, \mathbf{f}$$

$$\sigma^2(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})\, k(\mathbf{X}, \mathbf{X})^{-1}\, k(\mathbf{X}, \mathbf{x}^*)$$

$$\mathbf{f}_{\mathrm{MAP}} = k(\mathbf{X}, \mathbf{X})(\mathbf{y} - \sigma(\mathbf{f}_{\mathrm{MAP}}))$$

$$\mathbf{H} = \mathrm{diag}\left\{\sigma(\mathbf{f}_{\mathrm{MAP}})(1 - \sigma(\mathbf{f}_{\mathrm{MAP}}))\right\} + k(\mathbf{X}, \mathbf{X})^{-1}$$

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{H}^{-1})$$

$$p(y \mid \mathbf{x}) = \mathcal{N}(\mathbf{y} \mid \mathbf{A}\mathbf{x} + b, \sigma^2)$$

$$\Rightarrow \quad p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \sigma^2 + \mathbf{A}\mathbf{H}^{-1}\mathbf{A}^T)$$

Bishop eq 2.115

# One More Layer Up ...

$$p(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{p(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{f})}_{\mathcal{N}(f^*;\, m(\mathbf{x}^*),\, \sigma^2(\mathbf{x}^*))}\, \underbrace{p(\mathbf{f} \mid \mathbf{X}, \mathbf{y})}_{\mathcal{N}(\mathbf{f};\, \mathbf{f}_{\mathrm{MAP}},\, \mathbf{H}^{-1})}\, \mathrm{d}\mathbf{f} \qquad = \mathcal{N}(\mathbf{f}^*; c, d^2)$$

$$m(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{X})\, k(\mathbf{X}, \mathbf{X})^{-1} \boxed{\mathbf{f}}$$

$$\sigma^2(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})\, k(\mathbf{X}, \mathbf{X})^{-1}\, k(\mathbf{X}, \mathbf{x}^*)$$

$$\mathbf{f}_{\mathrm{MAP}} = \boxed{k(\mathbf{X}, \mathbf{X})(\mathbf{y} - \sigma(\mathbf{f}_{\mathrm{MAP}}))}$$

$$\mathbf{H} = \mathrm{diag}\left\{\sigma(\mathbf{f}_{\mathrm{MAP}})(1 - \sigma(\mathbf{f}_{\mathrm{MAP}}))\right\} + k(\mathbf{X}, \mathbf{X})^{-1}$$

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boxed{\boldsymbol{\mu}}, \mathbf{H}^{-1})$$

$$p(y \mid \mathbf{x}) = \mathcal{N}(\mathbf{y} \mid \boxed{\mathbf{A}\mathbf{x}} + b, \sigma^2)$$

$$\Rightarrow\quad p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \boxed{\mathbf{A}\boldsymbol{\mu} + \mathbf{b}}, \sigma^2 + \mathbf{A}\mathbf{H}^{-1}\mathbf{A}^T)$$

Bishop eq 2.115

# One More Layer Up ...

$$p(f^* \,|\, \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{p(f^* \,|\, \mathbf{x}^*, \mathbf{X}, \mathbf{f})}_{\mathcal{N}(f^*;\, m(\mathbf{x}^*),\, \sigma^2(\mathbf{x}^*))} \underbrace{p(\mathbf{f} \,|\, \mathbf{X}, \mathbf{y})}_{\mathcal{N}(\mathbf{f};\, \mathbf{f}_{\mathrm{MAP}},\, \mathbf{H}^{-1})} \, \mathrm{d}\mathbf{f} \qquad = \mathcal{N}(\mathbf{f}^*; c, d^2)$$

$$m(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{X})\, k(\mathbf{X}, \mathbf{X})^{-1} \boxed{\mathbf{f}}$$

$$\sigma^2(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})\, k(\mathbf{X}, \mathbf{X})^{-1}\, k(\mathbf{X}, \mathbf{x}^*)$$

$$\mathbf{f}_{\mathrm{MAP}} = \boxed{k(\mathbf{X}, \mathbf{X})(\mathbf{y} - \sigma(\mathbf{f}_{\mathrm{MAP}}))}$$

$$\mathbf{H} = \mathrm{diag}\left\{\sigma(\mathbf{f}_{\mathrm{MAP}})(1 - \sigma(\mathbf{f}_{\mathrm{MAP}}))\right\} + k(\mathbf{X}, \mathbf{X})^{-1}$$

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \,|\, \boxed{\boldsymbol{\mu}}, \mathbf{H}^{-1})$$

$$p(y \,|\, \mathbf{x}) = \mathcal{N}(\mathbf{y} \,|\, \boxed{\mathbf{A}\mathbf{x}} + b, \sigma^2)$$

$$\Rightarrow \quad p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \,|\, \boxed{\mathbf{A}\boldsymbol{\mu} + \mathbf{b}}, \sigma^2 + \mathbf{A}\mathbf{H}^{-1}\mathbf{A}^T)$$

Bishop eq 2.115

$$\Rightarrow \quad c = k(\mathbf{x}^*, \mathbf{X})\big(\mathbf{y} - \sigma(\mathbf{f}_{\mathrm{MAP}})\big)$$

Bishop eq 6.87,
GP Book eq. 3.21

# One More Layer Up …

$$p(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{p(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{f})}_{\mathcal{N}(f^*;\, m(\mathbf{x}^*),\, \sigma^2(\mathbf{x}^*))} \underbrace{p(\mathbf{f} \mid \mathbf{X}, \mathbf{y})}_{\mathcal{N}(\mathbf{f};\, \mathbf{f}_{\mathrm{MAP}},\, \mathbf{H}^{-1})} \, \mathrm{d}\mathbf{f} \qquad = \mathcal{N}(\mathbf{f}^*; c, d^2)$$

$$m(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{X})\, k(\mathbf{X}, \mathbf{X})^{-1} \mathbf{f}$$

$$\sigma^2(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})\, k(\mathbf{X}, \mathbf{X})^{-1}\, k(\mathbf{X}, \mathbf{x}^*)$$

$$\mathbf{f}_{\mathrm{MAP}} = k(\mathbf{X}, \mathbf{X})(\mathbf{y} - \sigma(\mathbf{f}_{\mathrm{MAP}}))$$

$$\mathbf{H} = \mathrm{diag}\{\sigma(\mathbf{f}_{\mathrm{MAP}})(1 - \sigma(\mathbf{f}_{\mathrm{MAP}}))\} + k(\mathbf{X}, \mathbf{X})^{-1}$$

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{H}^{-1})$$

$$p(y \mid \mathbf{x}) = \mathcal{N}(\mathbf{y} \mid \mathbf{A}\mathbf{x} + b, \sigma^2)$$

$$\Rightarrow p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \sigma^2 + \mathbf{A}\mathbf{H}^{-1}\mathbf{A}^T)$$

Bishop eq 2.115

$\Rightarrow$

$$c = k(\mathbf{x}^*, \mathbf{X})\big(\mathbf{y} - \sigma(\mathbf{f}_{\mathrm{MAP}})\big)$$

Bishop eq 6.87,
GP Book eq. 3.21

# One More Layer Up ...

$$p(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \underbrace{p(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{f})}_{\mathcal{N}(f^*; m(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))} \underbrace{p(\mathbf{f} \mid \mathbf{X}, \mathbf{y})}_{\mathcal{N}(\mathbf{f}; \mathbf{f}_{\text{MAP}}, \mathbf{H}^{-1})} \, d\mathbf{f} \qquad = \mathcal{N}(\mathbf{f}^*; c, d^2)$$

$$m(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{X})\, k(\mathbf{X}, \mathbf{X})^{-1} \mathbf{f}$$

$$\sigma^2(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})\, k(\mathbf{X}, \mathbf{X})^{-1}\, k(\mathbf{X}, \mathbf{x}^*)$$

$$\mathbf{f}_{\text{MAP}} = k(\mathbf{X}, \mathbf{X})(\mathbf{y} - \sigma(\mathbf{f}_{\text{MAP}}))$$

$$\mathbf{H} = \operatorname{diag}\{\sigma(\mathbf{f}_{\text{MAP}})(1 - \sigma(\mathbf{f}_{\text{MAP}}))\} + k(\mathbf{X}, \mathbf{X})^{-1}$$

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{H}^{-1})$$

$$p(y \mid \mathbf{x}) = \mathcal{N}(\mathbf{y} \mid \mathbf{A}\mathbf{x} + b, \sigma^2)$$

$$\Rightarrow p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \sigma^2 + \mathbf{A}\mathbf{H}^{-1}\mathbf{A}^T)$$

Bishop eq 2.115

$\Rightarrow$

$$c = k(\mathbf{x}^*, \mathbf{X})\big(\mathbf{y} - \sigma(\mathbf{f}_{\text{MAP}})\big)$$

Bishop eq 6.87,
GP Book eq. 3.21

$$d^2 = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X}) \Big( \operatorname{diag}\{\sigma(\mathbf{f}_{\text{MAP}})(1 - \sigma(\mathbf{f}_{\text{MAP}}))\}^{-1} + k(\mathbf{X}, \mathbf{X}) \Big)^{-1} k(\mathbf{X}, \mathbf{x}^*)$$

Bishop eq 6.88,
GP Book eq 3.22

# Predictive Distribution for Gaussian Process Classification

Finally, $p(y^* = 1 \,|\, x^*, \mathbf{X}, \mathbf{y}) = \displaystyle\int p(y^* = 1 \,|\, f^*, \mathbf{x}^*)\, p(f^* \,|\, \mathbf{x}^*, \mathbf{X}, \mathbf{y})\, \mathrm{d}f^*$

GP Book eq 3.10

# Predictive Distribution for Gaussian Process Classification

Finally, $p(y^* = 1 \,|\, x^*, \mathbf{X}, \mathbf{y}) = \displaystyle\int p(y^* = 1 \,|\, f^*, \mathbf{x}^*)\, p(f^* \,|\, \mathbf{x}^*, \mathbf{X}, \mathbf{y})\, \mathrm{d}f^*$

GP Book eq 3.10

$\mathcal{N}(f^*; c, d^2)$

# Predictive Distribution for Gaussian Process Classification

Finally, $p(y^* = 1 \,|\, x^*, \mathbf{X}, \mathbf{y}) = \displaystyle\int \underbrace{p(y^* = 1 \,|\, f^*, \mathbf{x}^*)}_{\sigma(f^*)} \, \underbrace{p(f^* \,|\, \mathbf{x}^*, \mathbf{X}, \mathbf{y})}_{\mathcal{N}(f^*;\, c,\, d^2)} \, \mathrm{d}f^*$

GP Book eq 3.10

# Predictive Distribution for Gaussian Process Classification

Finally, $p(y^* = 1 \,|\, x^*, \mathbf{X}, \mathbf{y}) = \int p(y^* = 1 \,|\, f^*, \mathbf{x}^*) \, p(f^* \,|\, \mathbf{x}^*, \mathbf{X}, \mathbf{y}) \, \mathrm{d}f^*$

GP Book eq 3.10

$\sigma(f^*)$ $\qquad$ $\mathcal{N}(f^*; c, d^2)$

Recall with Probit approximation

$$\int \sigma(a) \mathcal{N}(a \,|\, \mu, \sigma^2) \mathrm{d}a \simeq \sigma(\kappa(\sigma^2)\, \mu)$$

$$\kappa(\sigma^2) = (1 + \pi\sigma^2/8)^{-1/2}$$

Bishop eq 4.153

# Predictive Distribution for Gaussian Process Classification

Finally, $p(y^* = 1 \,|\, x^*, \mathbf{X}, \mathbf{y}) = \displaystyle\int p(y^* = 1 \,|\, f^*, \mathbf{x}^*)\, p(f^* \,|\, \mathbf{x}^*, \mathbf{X}, \mathbf{y})\, \mathrm{d}f^*$

GP Book eq 3.10

$\sigma(f^*)$ $\qquad$ $\mathcal{N}(f^*; c, d^2)$

$$p(y^* = 1 \,|\, x^*, \mathbf{X}, \mathbf{y}) = \sigma\Big((1 + \pi d^2/8)^{-1/2}\, c\Big)$$

$$c = k(\mathbf{x}^*, \mathbf{X})\big(\mathbf{y} - \sigma(\mathbf{f}_{\mathrm{MAP}})\big)$$

Bishop eq 6.87, GP Book eq 3.21

$$d^2 = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})\Big(\mathrm{diag}\{\sigma(\mathbf{f}_{\mathrm{MAP}})(1 - \sigma(\mathbf{f}_{\mathrm{MAP}}))\}^{-1} + k(\mathbf{X}, \mathbf{X})\Big)^{-1} k(\mathbf{X}, \mathbf{x}^*)$$

Bishop eq 6.88, GP Book eq 3.22

**Recall with Probit approximation**

$$\int \sigma(a)\mathcal{N}(a \,|\, \mu, \sigma^2)\mathrm{d}a \simeq \sigma(\kappa(\sigma^2)\,\mu)$$

$$\kappa(\sigma^2) = (1 + \pi\sigma^2/8)^{-1/2}$$

Bishop eq 4.153

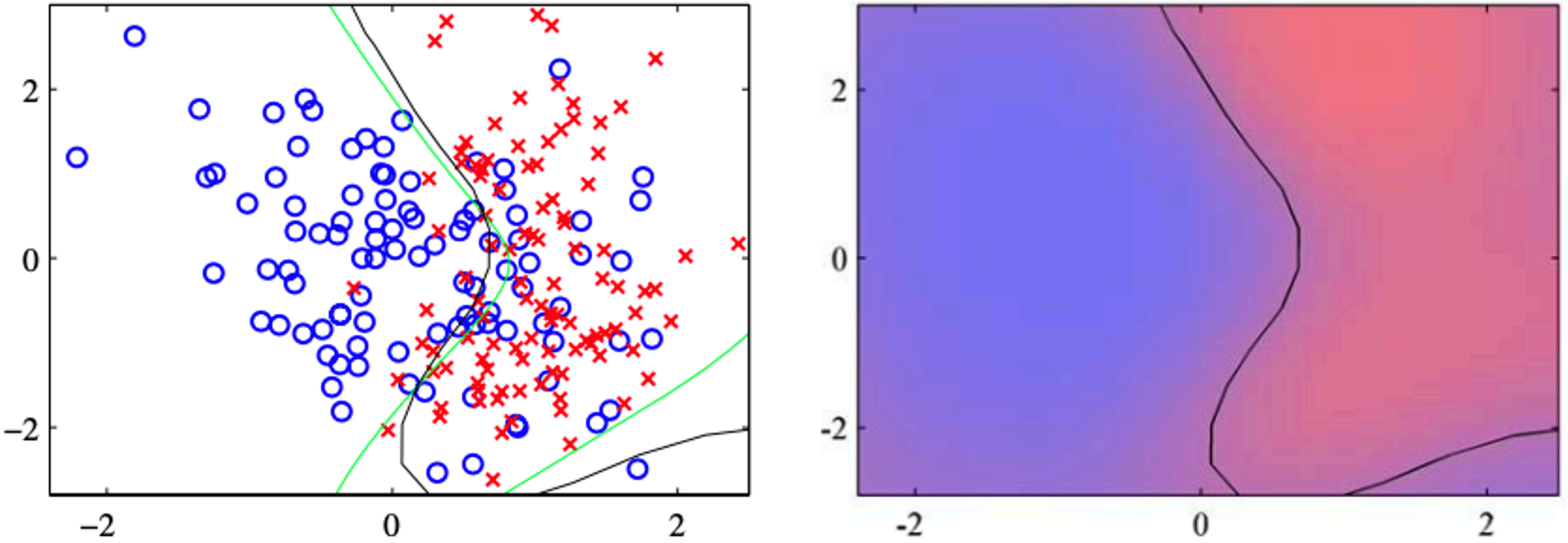



**Figure 6.12** Illustration of the use of a Gaussian process for classification, showing the data on the left together with the optimal decision boundary from the true distribution in green, and the decision boundary from the Gaussian process classifier in black. On the right is the predicted posterior probability for the blue and red classes together with the Gaussian process decision boundary.

# Recap: Learning Hyperparameters in GP Regression

$$\underbrace{p(\mathbf{y} \mid \mathbf{X}, \overbrace{\theta}^{\text{Kernel hyperparameter}})}_{\text{the marginal likelihood}} = \mathcal{N}(\mathbf{y}; \mathbf{0}, \underbrace{\mathbf{K}(\mathbf{X}, \mathbf{X})}_{\hat{\mathbf{K}}} + \sigma^2 \mathrm{I})$$

$$\ln p(\mathbf{y} \mid \mathbf{X}, \theta) = -\sum_{\mathbf{i}} \ln \mathbf{L}_{\mathbf{ii}} - \frac{\mathbf{1}}{\mathbf{2}} \mathbf{y}^{\mathrm{T}} \alpha - \frac{\mathbf{N}}{\mathbf{2}} \ln(\mathbf{2}\pi)$$

$L =$ Cholesky $\hat{\mathbf{K}}$

$\boldsymbol{\alpha} = \mathbf{L}^{\mathbf{T}} \setminus (\mathbf{L} \setminus \mathbf{y})$

# Recap: Learning Hyperparameters in GP Regression

$$\underbrace{p(\mathbf{y} \mid \mathbf{X}, \overbrace{\theta}^{\text{Kernel hyperparameter}})}_{\text{the marginal likelihood}} = \mathcal{N}(\mathbf{y}; \mathbf{0}, \underbrace{\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I}}_{\hat{\mathbf{K}}})$$

$$\ln p(\mathbf{y} \mid \mathbf{X}, \theta) = -\frac{1}{2} \ln |\hat{\mathbf{K}}| - \frac{1}{2} \mathbf{y}^{\mathrm{T}} \hat{\mathbf{K}}^{-1} \mathbf{y} - \frac{\mathrm{N}}{2} \ln(2\pi)$$

(Bishop eq 6.69)

$$\ln p(\mathbf{y} \mid \mathbf{X}, \theta) = -\sum_{\mathrm{i}} \ln \mathbf{L}_{\mathrm{ii}} - \frac{1}{2} \mathbf{y}^{\mathrm{T}} \alpha - \frac{\mathrm{N}}{2} \ln(2\pi)$$

$L =$ Cholesky $\hat{\mathbf{K}}$

$\boldsymbol{\alpha} = \mathbf{L}^{\mathrm{T}} \backslash (\mathbf{L} \backslash \mathbf{y})$

# Recap: Learning Hyperparameters in GP Regression

$$\underbrace{p(\mathbf{y} \mid \mathbf{X}}_{\text{the marginal likelihood}}, \overbrace{\theta}^{\text{Kernel hyperparameter}}) = \mathcal{N}(\mathbf{y}; \mathbf{0}, \underbrace{\mathbf{K}(\mathbf{X}, \mathbf{X})}_{\hat{\mathbf{K}}} + \sigma^2 \mathrm{I})$$

$$\ln p(\mathbf{y} \mid \mathbf{X}, \theta) = -\frac{1}{2} \ln |\hat{\mathbf{K}}| - \frac{1}{2} \mathbf{y}^{\mathrm{T}} \hat{\mathbf{K}}^{-1} \mathbf{y} - \frac{\mathbf{N}}{2} \ln(2\pi)$$

(Bishop eq 6.69)

Learning through gradient descent

$$\frac{\partial}{\partial x} \ln |\mathbf{A}| = \mathrm{Tr}\left(\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial \mathrm{x}}\right)$$

$$\frac{\partial}{\partial x}(\mathbf{A}^{-1}) = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial \mathbf{x}} \mathbf{A}^{-1}$$

(Bishop eq C.21, C.22)

$$\ln p(\mathbf{y} \mid \mathbf{X}, \theta) = -\sum_{\mathbf{i}} \ln \mathbf{L}_{\mathbf{ii}} - \frac{1}{2} \mathbf{y}^{\mathrm{T}} \alpha - \frac{\mathbf{N}}{2} \ln(2\pi)$$

$L =$ Cholesky $\hat{\mathbf{K}}$

$\boldsymbol{\alpha} = \mathbf{L}^{\mathrm{T}} \backslash (\mathbf{L} \backslash \mathbf{y})$

# Recap: Learning Hyperparameters in GP Regression

$$\underbrace{p(\mathbf{y} \mid \mathbf{X}}_{\text{the marginal likelihood}}, \overbrace{\theta}^{\text{Kernel hyperparameter}}) = \mathcal{N}(\mathbf{y}; \mathbf{0}, \underbrace{\mathbf{K}(\mathbf{X}, \mathbf{X})}_{\hat{\mathbf{K}}} + \sigma^2 \mathrm{I})$$

$$\ln p(\mathbf{y} \mid \mathbf{X}, \theta) = -\frac{1}{2} \ln |\hat{\mathbf{K}}| - \frac{1}{2} \mathbf{y}^{\mathrm{T}} \hat{\mathbf{K}}^{-1} \mathbf{y} - \frac{\mathbf{N}}{2} \ln(2\pi)$$

(Bishop eq 6.69)

$$\frac{\partial}{\partial x} \ln |\mathbf{A}| = \mathrm{Tr}\left(\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial \mathrm{x}}\right)$$

$$\frac{\partial}{\partial x}(\mathbf{A}^{-1}) = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial \mathbf{x}} \mathbf{A}^{-1}$$

(Bishop eq C.21, C.22)

Learning through gradient descent

$$\frac{\partial}{\partial \theta_i} \ln p(\mathbf{y} \mid \mathbf{X}, \theta) = -\frac{1}{2} \mathrm{Tr}\left(\hat{\mathbf{K}}^{-1} \frac{\partial \hat{\mathbf{K}}}{\partial \theta_{\mathrm{i}}}\right) + \frac{1}{2} \mathbf{y}^{\mathrm{T}} \hat{\mathbf{K}}^{-1} \frac{\partial \hat{\mathbf{K}}}{\partial \theta_{\mathrm{i}}} \hat{\mathbf{K}}^{-1} \mathbf{y}$$

(Bishop eq 6.70)

$$\ln p(\mathbf{y} \mid \mathbf{X}, \theta) = -\sum_{\mathrm{i}} \ln \mathbf{L}_{\mathrm{ii}} - \frac{1}{2} \mathbf{y}^{\mathrm{T}} \alpha - \frac{\mathbf{N}}{2} \ln(2\pi)$$

$L =$ Cholesky $\hat{\mathbf{K}}$

$\boldsymbol{\alpha} = \mathbf{L}^{\mathrm{T}} \backslash (\mathbf{L} \backslash \mathbf{y})$

# Marginal Likelihood for GP Classification and BIC

$$p(\mathbf{y} \,|\, \mathbf{X}, \theta) = \int p(\mathbf{y} \,|\, \mathbf{f}, \mathbf{X}, \theta)\, p(\mathbf{f} \,|\, \mathbf{X}, \theta)\, \mathrm{d}\mathbf{f}$$

$$\boxed{p(\mathbf{f} \,|\, \mathbf{X}, \theta) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K})}$$

Bishop eq 6.89, GP Book eq 3.30

# Marginal Likelihood for GP Classification and BIC

$$p(\mathbf{y}\,|\,\mathbf{X},\theta) = \int p(\mathbf{y}\,|\,\mathbf{f},\mathbf{X},\theta)\,\underbrace{p(\mathbf{f}\,|\,\mathbf{X},\theta)}_{\propto\, p(\mathbf{f}\,|\,\mathbf{y},\mathbf{X},\theta)\ \simeq\ \mathcal{N}(\mathbf{a}\,|\,\mathbf{a}_{\mathrm{MAP}},\,\mathbf{H}^{-1})}\,\mathrm{d}\mathbf{f}$$

$$\boxed{p(\mathbf{f}\,|\,\mathbf{X},\theta) = \mathcal{N}(\mathbf{f};\mathbf{0},\mathbf{K})}$$

Bishop eq 6.89, GP Book eq 3.30

# Marginal Likelihood for GP Classification and BIC

$$p(\mathbf{y}\,|\,\mathbf{X},\theta) = \int \underbrace{p(\mathbf{y}\,|\,\mathbf{f},\mathbf{X},\theta)\,p(\mathbf{f}\,|\,\mathbf{X},\theta)}_{\propto\, p(\mathbf{f}\,|\,\mathbf{y},\mathbf{X},\theta)\ \simeq\ \mathcal{N}(\mathbf{a}\,|\,\mathbf{a}_{\mathrm{MAP}},\,\mathbf{H}^{-1})}\,\mathrm{d}\mathbf{f}$$

$$\boxed{p(\mathbf{f}\,|\,\mathbf{X},\theta) = \mathcal{N}(\mathbf{f};\mathbf{0},\mathbf{K})}$$

Bishop eq 6.89, GP Book eq 3.30

For a scaled Gaussian, the “normalising” constant

Bishop eq 4.135, 4.137, GP Book eq 3.31

$$p(\mathbf{y}\,|\,\mathbf{X},\theta) =$$

$$\int -\frac{1}{2}(\mathbf{f}-\mathbf{f}_{\mathrm{MAP}})^T\,\mathbf{H}\,(\mathbf{f}-\mathbf{f}_{\mathrm{MAP}})\,\mathrm{d}\mathbf{f}$$

# Marginal Likelihood for GP Classification and BIC

$$p(\mathbf{y}\,|\,\mathbf{X},\theta) = \int \underbrace{p(\mathbf{y}\,|\,\mathbf{f},\mathbf{X},\theta)\,p(\mathbf{f}\,|\,\mathbf{X},\theta)}_{\propto\, p(\mathbf{f}\,|\,\mathbf{y},\mathbf{X},\theta)\ \simeq\ \mathcal{N}(\mathbf{a}\,|\,\mathbf{a}_{\mathrm{MAP}},\,\mathbf{H}^{-1})}\,\mathrm{d}\mathbf{f}$$

$$\boxed{p(\mathbf{f}\,|\,\mathbf{X},\theta) = \mathcal{N}(\mathbf{f};\mathbf{0},\mathbf{K})}$$

Bishop eq 6.89, GP Book eq 3.30

For a scaled Gaussian, the "normalising" constant

Bishop eq 4.135, 4.137, GP Book eq 3.31

$$p(\mathbf{y}\,|\,\mathbf{X},\theta) = p(\mathbf{y}\,|\,\mathbf{f}_{\mathrm{MAP}},\mathbf{X},\theta)\,p(\mathbf{f}_{\mathrm{MAP}}\,|\,\mathbf{X},\theta)\ \int -\frac{1}{2}(\mathbf{f}-\mathbf{f}_{\mathrm{MAP}})^T\,\mathbf{H}\,(\mathbf{f}-\mathbf{f}_{\mathrm{MAP}})\,\mathrm{d}\mathbf{f}$$

# Marginal Likelihood for GP Classification and BIC

$$p(\mathbf{y}\,|\,\mathbf{X},\theta) = \int \underbrace{p(\mathbf{y}\,|\,\mathbf{f},\mathbf{X},\theta)\,p(\mathbf{f}\,|\,\mathbf{X},\theta)}_{\propto\, p(\mathbf{f}\,|\,\mathbf{y},\mathbf{X},\theta)\ \simeq\ \mathcal{N}(\mathbf{a}\,|\,\mathbf{a}_{\mathrm{MAP}},\,\mathbf{H}^{-1})}\,\mathrm{d}\mathbf{f}$$

$$\boxed{p(\mathbf{f}\,|\,\mathbf{X},\theta) = \mathcal{N}(\mathbf{f};\mathbf{0},\mathbf{K})}$$

Bishop eq 6.89, GP Book eq 3.30

For a scaled Gaussian, the “normalising” constant

Bishop eq 4.135, 4.137, GP Book eq 3.31

$$p(\mathbf{y}\,|\,\mathbf{X},\theta) = p(\mathbf{y}\,|\,\mathbf{f}_{\mathrm{MAP}},\mathbf{X},\theta)\,p(\mathbf{f}_{\mathrm{MAP}}\,|\,\mathbf{X},\theta) \int -\frac{1}{2}(\mathbf{f}-\mathbf{f}_{\mathrm{MAP}})^T\,\mathbf{H}\,(\mathbf{f}-\mathbf{f}_{\mathrm{MAP}})\,\mathrm{d}\mathbf{f}$$

Bishop eq 6.90

$$\ln p(\mathbf{y}\,|\,\mathbf{X},\theta) = \ln p(\mathbf{y}\,|\,\mathbf{f}_{\mathrm{MAP}},\mathbf{X},\theta) + \ln p(\mathbf{f}_{\mathrm{MAP}}\,|\,\mathbf{X},\theta) - \frac{1}{2}\ln|\mathbf{H}| + \frac{N}{2}\ln 2\pi$$

# Marginal Likelihood for GP Classification and BIC

$$p(\mathbf{y}\,|\,\mathbf{X},\theta) = \int p(\mathbf{y}\,|\,\underbrace{\mathbf{f},\mathbf{X},\theta)\,p(\mathbf{f}\,|\,\mathbf{X},\theta)}_{\propto\, p(\mathbf{f}\,|\,\mathbf{y},\mathbf{X},\theta)\ \simeq\ \mathcal{N}(\mathbf{a}\,|\,\mathbf{a}_{\mathrm{MAP}},\,\mathbf{H}^{-1})}\,\mathrm{d}\mathbf{f}$$

$$\boxed{p(\mathbf{f}\,|\,\mathbf{X},\theta) = \mathcal{N}(\mathbf{f};\mathbf{0},\mathbf{K})}$$

Bishop eq 6.89, GP Book eq 3.30

For a scaled Gaussian, the “normalising” constant

Bishop eq 4.135, 4.137, GP Book eq 3.31

$$p(\mathbf{y}\,|\,\mathbf{X},\theta) = p(\mathbf{y}\,|\,\mathbf{f}_{\mathrm{MAP}},\mathbf{X},\theta)\,p(\mathbf{f}_{\mathrm{MAP}}\,|\,\mathbf{X},\theta)\int -\frac{1}{2}(\mathbf{f}-\mathbf{f}_{\mathrm{MAP}})^T\,\mathbf{H}\,(\mathbf{f}-\mathbf{f}_{\mathrm{MAP}})\,\mathrm{d}\mathbf{f}$$

Bishop eq 6.90

$$\ln p(\mathbf{y}\,|\,\mathbf{X},\theta) = \boxed{\ln p(\mathbf{y}\,|\,\mathbf{f}_{\mathrm{MAP}},\mathbf{X},\theta)} + \ln p(\mathbf{f}_{\mathrm{MAP}}\,|\,\mathbf{X},\theta) - \frac{1}{2}\ln|\mathbf{H}| + \frac{N}{2}\ln 2\pi$$

Goodness of fit at the optimal parameter

# Marginal Likelihood for GP Classification and BIC

$$p(\mathbf{y}\,|\,\mathbf{X},\theta) = \int \underbrace{p(\mathbf{y}\,|\,\mathbf{f},\mathbf{X},\theta)\,p(\mathbf{f}\,|\,\mathbf{X},\theta)}_{\propto\, p(\mathbf{f}\,|\,\mathbf{y},\mathbf{X},\theta)\;\simeq\;\mathcal{N}(\mathbf{a}\,|\,\mathbf{a}_{\mathrm{MAP}},\,\mathbf{H}^{-1})}\,\mathrm{d}\mathbf{f}$$

$$\boxed{p(\mathbf{f}\,|\,\mathbf{X},\theta) = \mathcal{N}(\mathbf{f};\mathbf{0},\mathbf{K})}$$

Bishop eq 6.89, GP Book eq 3.30

For a scaled Gaussian, the "normalising" constant

Bishop eq 4.135, 4.137, GP Book eq 3.31

$$p(\mathbf{y}\,|\,\mathbf{X},\theta) = p(\mathbf{y}\,|\,\mathbf{f}_{\mathrm{MAP}},\mathbf{X},\theta)\,p(\mathbf{f}_{\mathrm{MAP}}\,|\,\mathbf{X},\theta)\int -\frac{1}{2}(\mathbf{f}-\mathbf{f}_{\mathrm{MAP}})^T\,\mathbf{H}\,(\mathbf{f}-\mathbf{f}_{\mathrm{MAP}})\,\mathrm{d}\mathbf{f}$$

Bishop eq 6.90

$$\ln p(\mathbf{y}\,|\,\mathbf{X},\theta) = \boxed{\ln p(\mathbf{y}\,|\,\mathbf{f}_{\mathrm{MAP}},\mathbf{X},\theta)} + \ln p(\mathbf{f}_{\mathrm{MAP}}\,|\,\mathbf{X},\theta)\boxed{-\frac{1}{2}\ln|\mathbf{H}| + \frac{N}{2}\ln 2\pi}$$

Goodness of fit at the optimal parameter

Complexity / degree of freedom "penalty"

# What we have learned

Bishop's textbook: chapter 4.4-4.5, 6.4.5-6.4.6
GP book: http://gaussianprocess.org/gpml/chapters/
chapter 3.1-3.4

Bishop's textbook: chapter 4.4-4.5, 6.4.5-6.4.6
GP book: http://gaussianprocess.org/gpml/chapters/
chapter 3.1-3.4

# What we have learned

- Bayesian Logistic Regression

  - the challenge to computing a non-Gaussian predictive distribution

# What we have learned

Bishop's textbook: chapter 4.4-4.5, 6.4.5-6.4.6
GP book: http://gaussianprocess.org/gpml/chapters/
chapter 3.1-3.4

- Bayesian Logistic Regression

  - the challenge to computing a non-Gaussian predictive distribution

- Laplace Approximation for Bayesian Logistic Regression

Bishop's textbook: chapter 4.4-4.5, 6.4.5-6.4.6
GP book: http://gaussianprocess.org/gpml/chapters/
chapter 3.1-3.4

# What we have learned

- Bayesian Logistic Regression

  - the challenge to computing a non-Gaussian predictive distribution

- Laplace Approximation for Bayesian Logistic Regression

- Gaussian Process Classification

## What we have learned

Bishop's textbook: chapter 4.4-4.5, 6.4.5-6.4.6
GP book: http://gaussianprocess.org/gpml/chapters/
chapter 3.1-3.4

- Bayesian Logistic Regression

  - the challenge to computing a non-Gaussian predictive distribution

- Laplace Approximation for Bayesian Logistic Regression
- Gaussian Process Classification
- Laplace Approximation for Gaussian Process Classification