IMPERIAL

Bayesian Model Selection:

Selecting the “best” model in a systematic way

SA AIMS

18th Feb, 2026

Lloyd Fung

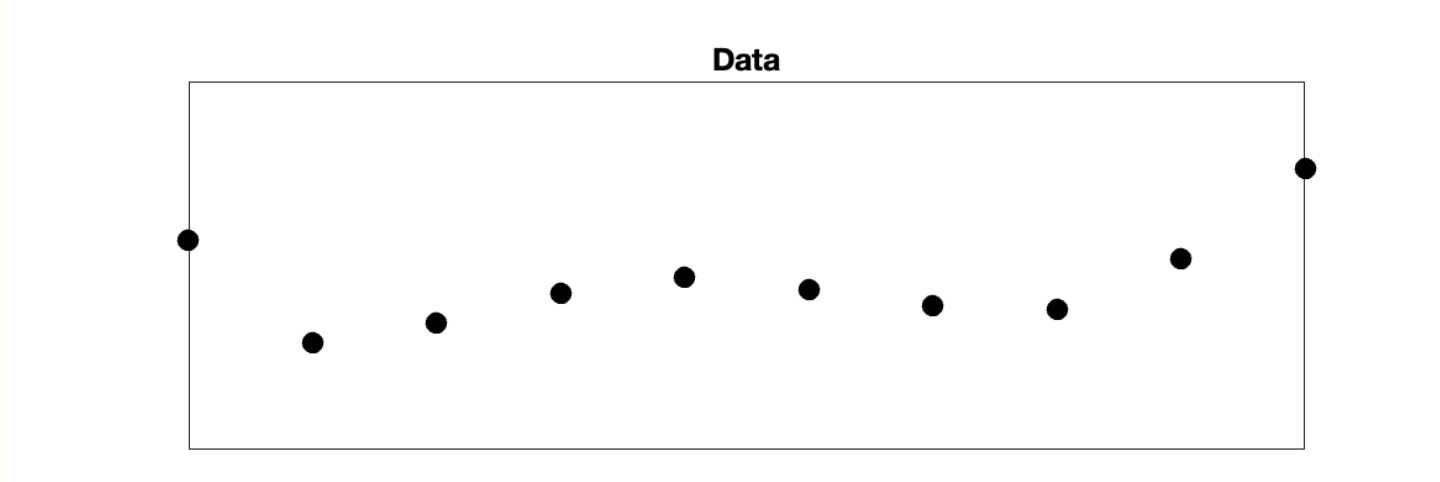
Research Fellow

Imperial College London

Learning Outcome

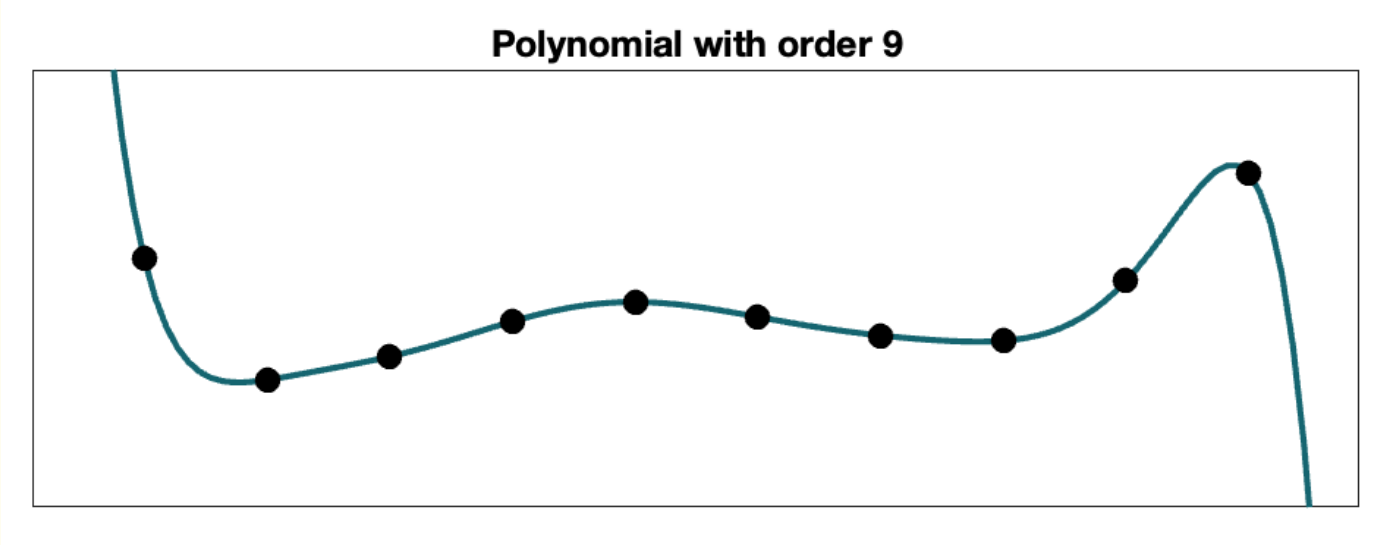
- Understand the principle of Occam razor and its Bayesian interpretation
- Gain an intuition to the scaling of prior and likelihood with data and noise
- Able to interpret Information Criterions from the Bayesian evidence perspective
- (Understand the mechanism of sparsification from Bayes' point of view)
- (Applying sparsifying prior to promote sparsity in model creation)

What line would you put through these datapoints?



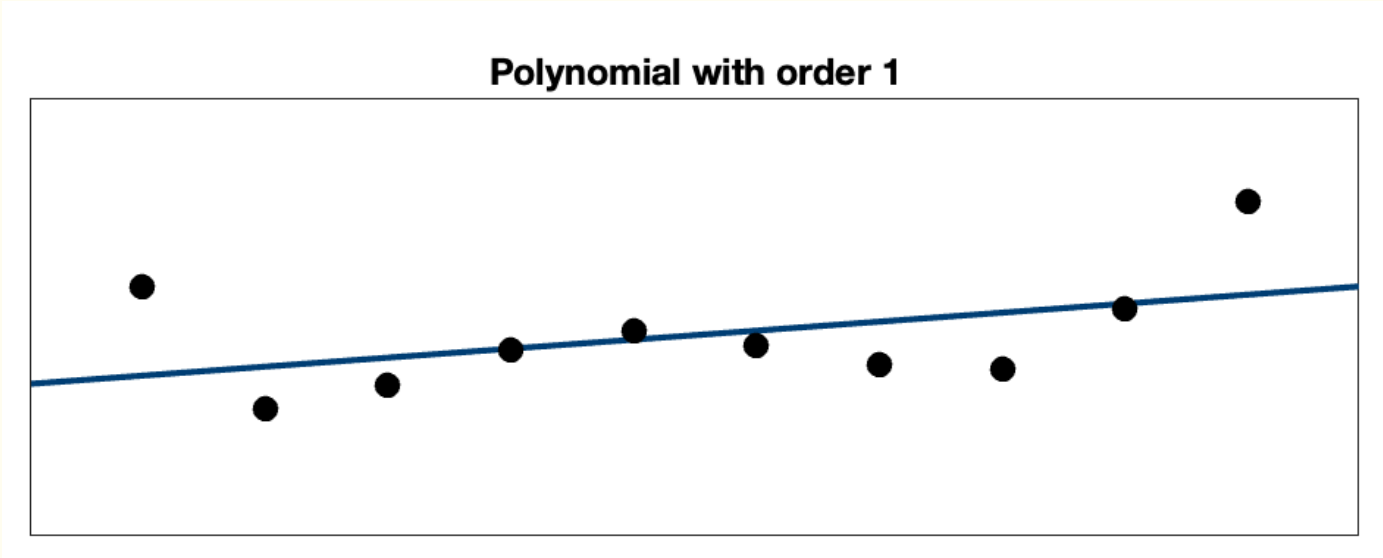
Taken from Prof. Matthew Juniper's lecture on Bayesian Data Assimilation

What line would you put through these datapoints?



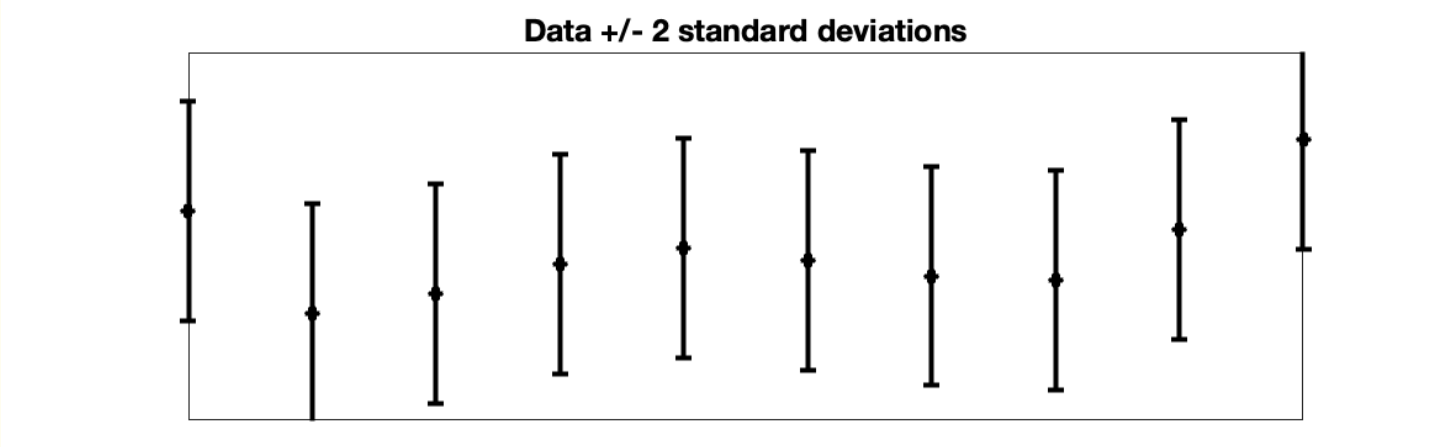
Taken from Prof. Matthew Juniper's lecture on Bayesian Data Assimilation

What line would you put through these datapoints?



Taken from Prof. Matthew Juniper's lecture on Bayesian Data Assimilation

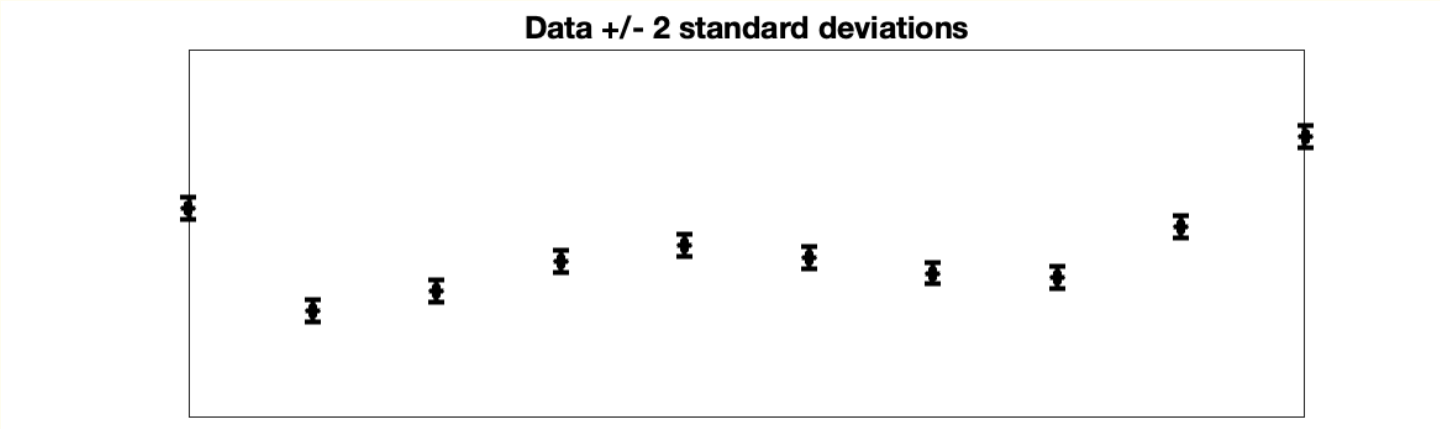
What line would you put through these datapoints?



Juniper (2022)

Taken from Prof. Matthew Juniper's lecture on Bayesian Data Assimilation

What line would you put through these datapoints?



Juniper (2022)

Taken from Prof. Matthew Juniper's lecture on Bayesian Data Assimilation

Occam Razor

“When faced with two (or more) possible explanations, the simpler one is the one most likely to be true.”

(Paraphrasing) **William of Ockham**

We want as simple a model as possible (that still fits the data).

Principle of parsimony in modelling

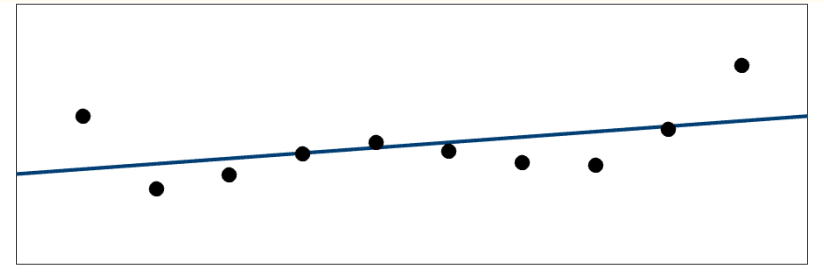
“Truth is much too complicated to allow anything but approximations.”

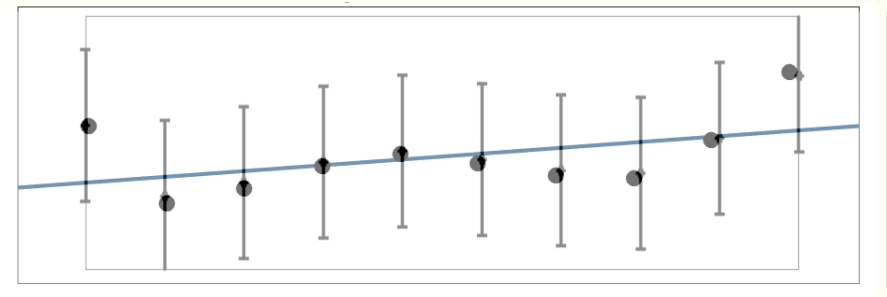
John von Neumann

We want to balance **model complexity** against **accuracy.**

What we want is… parsimonious model !

- **Ockham:** We want as **simple** a model as possible
- **von Neumann**: We want to balance **model complexity** against **accuracy**





How does Bayes help us?

Bayesian perspective
in the pursue of parsimonious models

Bayes' Theorem

$$P(\boldsymbol{\xi}|D) = \frac{P(D|\boldsymbol{\xi})P(\boldsymbol{\xi})}{P(D)}$$

posterior

likelihood prior

evidence

Just a normalising constant?

$D = \{\boldsymbol{x}, \boldsymbol{y}\}$ Data

$\boldsymbol{\xi}$ Fitting parameters

Bayes' Theorem

$$P\left(\boldsymbol{\xi}_{\mathcal{M}_i}|D,\mathcal{M}_i\right) = \frac{P\left(D|\boldsymbol{\xi}_{\mathcal{M}_i},\mathcal{M}_i\right)P(\boldsymbol{\xi}_{\mathcal{M}_i}|\mathcal{M}_i)}{P(D|\mathcal{M}_i)}$$

likelihood — prior

posterior

evidence

Key 'score' to compare btw models !

Model selection under Bayes

$$\underset{\mathcal{M}_i}{\operatorname{argmax}}\, P(\mathcal{M}_i|D)$$

where

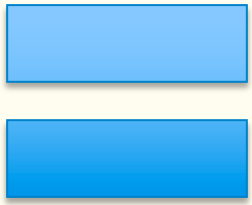
$$P(\mathcal{M}_i|D) \propto P(D|\mathcal{M}_i)P(\mathcal{M}_i)$$

$D = \{\boldsymbol{x}, \boldsymbol{y}\}$ Data

$\mathcal{M}_i$ Hypothesis (model basis)

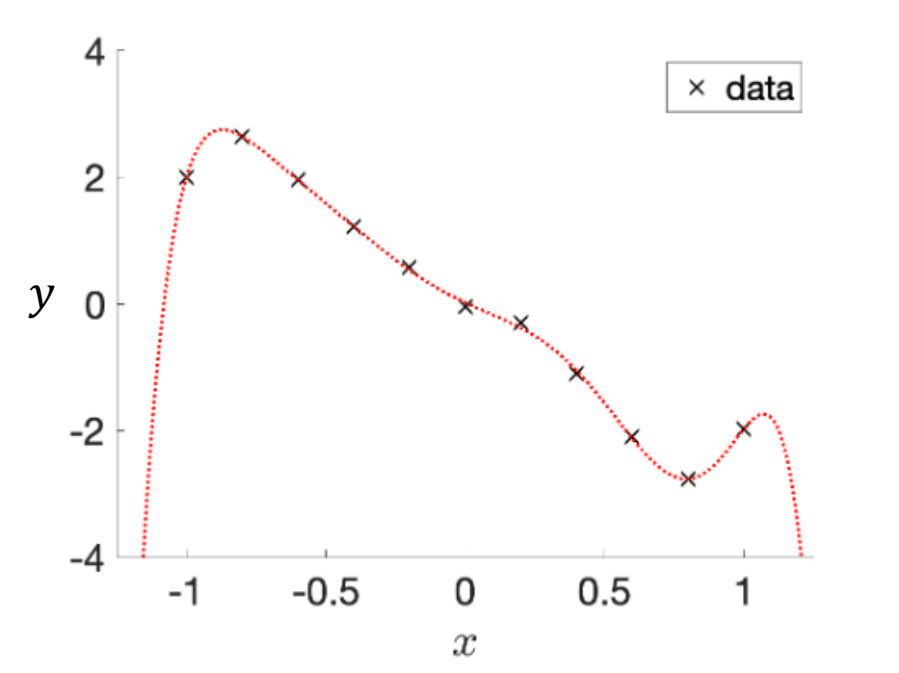
$\boldsymbol{\xi}_{\mathcal{M}_i}$ Fitting parameters

Model with max. $P(D|\mathcal{M}_i)$



Most parsimonious model

An example with linear regression – Model 1 (Complex)

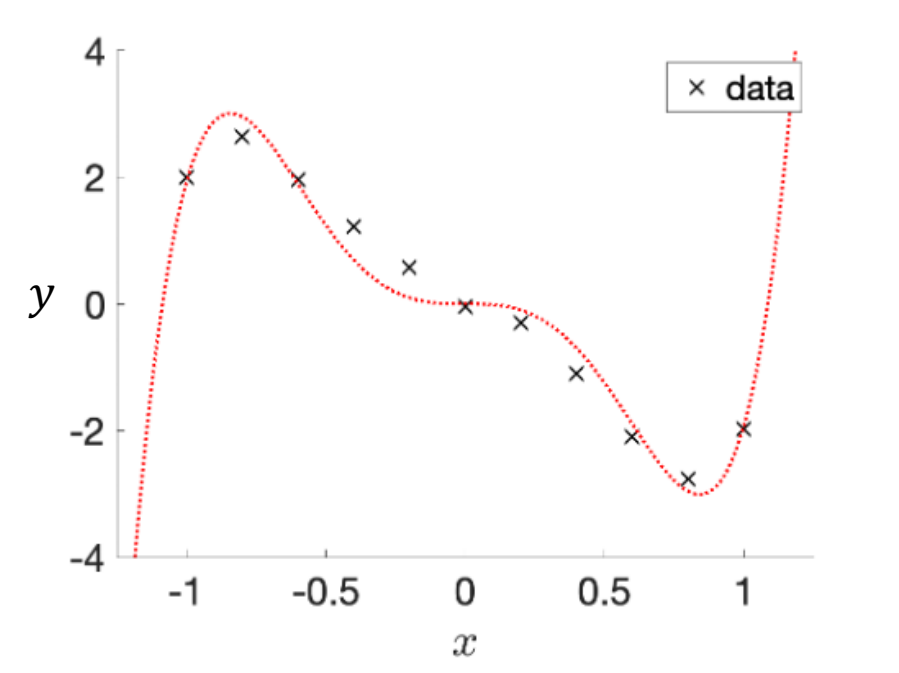


target = Polynomial Functions · Learnt Coefficients

$\mathbf{y} = \boldsymbol{\Theta}(\boldsymbol{x}) \cdot \boldsymbol{\xi}$

$\begin{bmatrix} \mathbf{y} \end{bmatrix} = \begin{bmatrix} 1 & x & x^2 & x^3 & x^4 & x^5 & x^6 & x^7 & x^8 & \cdots \end{bmatrix} \begin{bmatrix} \xi \end{bmatrix}$

An example with linear regression – Model 2 (parsimonious)



target = Polynomial Functions · Learnt Coefficients

$\mathbf{y} = \boldsymbol{\Theta}(\boldsymbol{x}) \cdot \boldsymbol{\xi}$

$\mathbf{y} = \begin{bmatrix} 1 & x & x^2 & x^3 & x^4 & x^5 & x^6 & x^7 & x^8 & \cdots \end{bmatrix} \begin{bmatrix} \cdot \\ \cdot \\ \xi \end{bmatrix}$

Why is max. evidence = Occam Razor?
An example with linear regression

$D = \{\boldsymbol{x}, \boldsymbol{y}\}$ Data

$\mathcal{M}_i$ Model

$\boldsymbol{\Theta}_{\mathcal{M}}(\boldsymbol{x})$ Model Basis

$\boldsymbol{\xi}$ Fitting parameters

Bayes' rule

$$P(\boldsymbol{\xi}|D,\mathcal{M}_i) = \frac{P(D|\boldsymbol{\xi},\mathcal{M}_i)P(\boldsymbol{\xi}|\mathcal{M}_i)}{P(D|\mathcal{M}_i)}$$

posterior; likelihood; prior; evidence

Gaussian Linear model:
$\boldsymbol{y} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{\Theta}_{\mathcal{M}}(\boldsymbol{x}) \cdot \boldsymbol{\xi}, \sigma^2\mathbf{I})$

$$-\ln P(\boldsymbol{\xi}|D,\mathcal{M}_i) = \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{\Theta}_{\mathcal{M}}(\boldsymbol{x}) \cdot \boldsymbol{\xi}\|_F^2 + \frac{1}{2\sigma_p^2}\|\boldsymbol{\xi}\|_2^2 + normalise$$

error

posterior; likelihood; Prior (L_2 regularisation)

Maximum a posteriori (MAP):

$$\boldsymbol{\xi}_{MAP} = \underset{\boldsymbol{\xi}}{\operatorname{argmax}}\, P(\boldsymbol{\xi}|D,\mathcal{M}_i)$$

=

Regularised Regression
(Regularisation ↔ prior)

Why is max. evidence = Occam Razor?
An example with linear regression

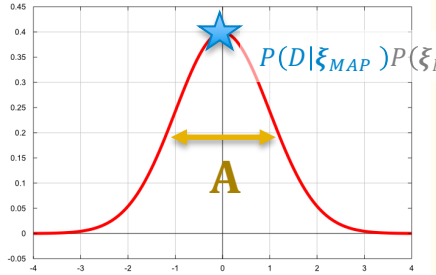
$$\mathbf{A} = \sigma_p^{-2}\mathbf{I} + \sigma^{-2}\boldsymbol{\Theta}_{\mathcal{M}}{}^{\mathrm{T}}\boldsymbol{\Theta}_{\mathcal{M}}$$

Assuming the prior to all models $P(\mathcal{M}_i)$ is uniform:

$$P(\mathcal{M}_i|D) \propto P(D|\mathcal{M}_i) = \int_{\xi} P(D|\boldsymbol{\xi}, \mathcal{M}_i) P(\boldsymbol{\xi}|\mathcal{M}_i) d\boldsymbol{\xi}$$

← Difficult to eval.

Posterior:



$D = \{\boldsymbol{x}, \boldsymbol{y}\}$ Data

$\mathcal{M}_i$ Model

$\boldsymbol{\xi}$ Fitting parameters

$\mathbf{A}$ Parameter Covariance

Why is max. evidence = Occam Razor?
An example with linear regression

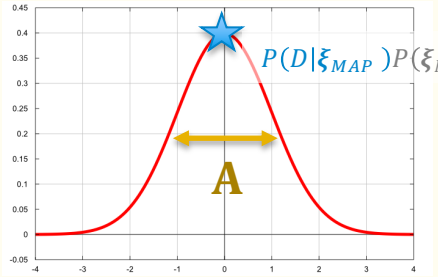
$$\mathbf{A} = \sigma_p^{-2}\mathbf{I} + \sigma^{-2}\boldsymbol{\Theta}_{\mathcal{M}}{}^{\mathrm{T}}\boldsymbol{\Theta}_{\mathcal{M}}$$

Assuming the prior to all models $P(\mathcal{M}_i)$ is uniform:

$$P(\mathcal{M}_i|D) \propto P(D|\mathcal{M}_i) = \int_{\xi} P(D|\xi, \mathcal{M}_i) P(\xi|\mathcal{M}_i) d\xi$$

Occam Razor!

Posterior:



$$= P(D|\xi_{MAP}, \mathcal{M}_i) \quad P(\xi_{MAP}|\mathcal{M}_i) \quad \left[\det\left(\frac{\mathbf{A}}{2\pi}\right)\right]^{-1/2}$$

Likelihood @ MAP — Prior on ξ @ MAP — Uncertainty in ξ around MAP

$\uparrow \#\xi \Rightarrow$
$\uparrow likelihood$

$\uparrow \#\xi \Rightarrow$
$\downarrow prior$

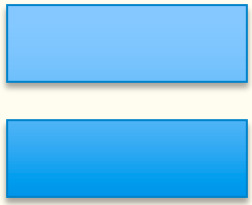
$\uparrow \#\xi \Rightarrow$
$\uparrow rank(\mathbf{A}) \Rightarrow$
$\downarrow |\mathbf{A}|^{-1/2}$

How well $\mathcal{M}_i$ fit

Penalise models with too many param.

$D = \{\boldsymbol{x}, \boldsymbol{y}\}$ Data
$\mathcal{M}_i$ Model
$\boldsymbol{\xi}$ Fitting parameters
$\mathbf{A}$ Parameter Covariance

Model with max. $P(D|\mathcal{M}_i)$



Indeed!

Most parsimonious model

Computing Evidence – Gaussian case

Aim: $\underset{\mathcal{M}_i}{\operatorname{argmax}} P(\mathcal{M}_i|D)$

$$P(\mathcal{M}_i|D) \propto P(D|\mathcal{M}_i) = \int_{\xi} P(D|\boldsymbol{\xi}, \mathcal{M}_i) P(\boldsymbol{\xi}|\mathcal{M}_i) d\boldsymbol{\xi}$$

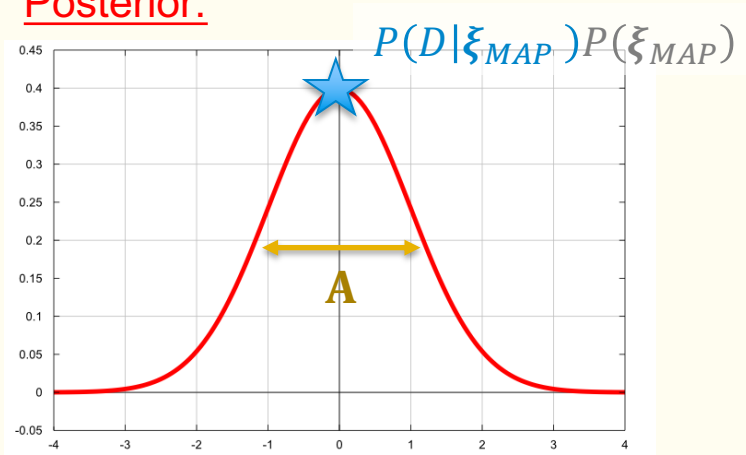
$$= P(D|\boldsymbol{\xi}_{MAP}, \mathcal{M}_i) \quad P(\boldsymbol{\xi}_{MAP}|\mathcal{M}_i) \quad \left[\det\left(\frac{\mathbf{A}}{2\pi}\right)\right]^{-1/2}$$

Likelihood @ MAP — Prior on $\boldsymbol{\xi}$ @ MAP — Uncertainty in $\boldsymbol{\xi}$ around MAP

Occam Factor

Obtained from linear regression

Posterior:



Penalizes models with too many param.

$D = \{\boldsymbol{x}, \boldsymbol{y}\}$ Data

$\mathcal{M}_i$ Model

$\boldsymbol{\xi}$ Fitting parameters

$\mathbf{A}$ Parameter Covariance

$$\mathbf{A} = \sigma_p^{-2}\mathbf{I} + \sigma^{-2}\boldsymbol{\Theta}_{\mathcal{M}}{}^{\mathrm{T}}\boldsymbol{\Theta}_{\mathcal{M}}$$

Approximating Evidence – Laplace Method

Aim: $\underset{\mathcal{M}_i}{\operatorname{argmax}} P(\mathcal{M}_i|D)$

$$P(\mathcal{M}_i|D) \propto P(D|\mathcal{M}_i) = \int_{\xi} \underbrace{P(D|\boldsymbol{\xi},\mathcal{M}_i)P(\boldsymbol{\xi}|\mathcal{M}_i)}_{\propto P(\boldsymbol{\xi}|D)} d\boldsymbol{\xi}$$

Difficult to eval.

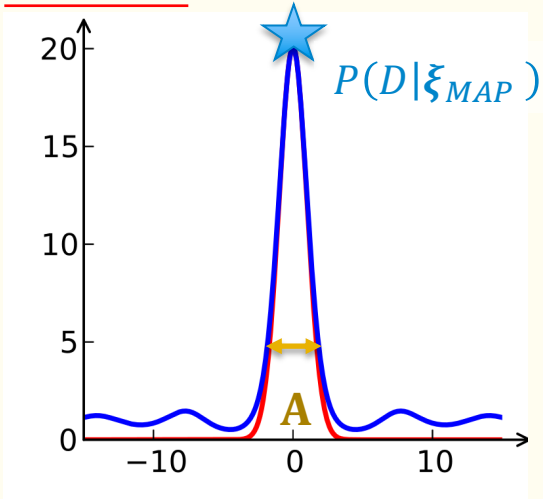
$$\approx P(D|\boldsymbol{\xi}_{MAP},\mathcal{M}_i) \quad P(\boldsymbol{\xi}_{MAP}|\mathcal{M}_i) \quad \left[\det\left(\frac{\mathbf{A}}{2\pi}\right)\right]^{-1/2}$$

Likelihood @ MAP — Prior on $\boldsymbol{\xi}$ @ MAP — Uncertainty in $\boldsymbol{\xi}$ around MAP

Occam Factor

Obtained from variational inference

Posterior:



Penalizes models with too many param.

$$\mathbf{A} = -\left.\frac{d^2 \ln P(D|\boldsymbol{\xi},\mathcal{M}_i)P(\boldsymbol{\xi}|\mathcal{M}_i)}{d\xi^2}\right|_{\xi_{MAP}}$$

$D = \{\boldsymbol{x}, \boldsymbol{y}\}$ Data

$\mathcal{M}_i$ Model

$\boldsymbol{\xi}$ Fitting parameters

$\mathbf{A}$ Parameter Uncertainty

Scaling of Likelihood and Occam Factor and BIC

$\mathbf{A}$ is a $k \times k$ matrix scales with n

$$\ln P(D|\mathcal{M}_i) \approx \underbrace{\ln P(D|\boldsymbol{\xi}_{MAP}, \mathcal{M}_i)}_{\sim O(n)} + \underbrace{\ln P(\boldsymbol{\xi}_{MAP}|\mathcal{M}_i)}_{\sim O(k)} + \underbrace{\ln \left[\det\left(\frac{\mathbf{A}}{2\pi}\right)\right]^{-1/2}}_{\sim O(-\frac{k}{2}\ln\frac{n}{2\pi})}$$

Likelihood @ MAP

Occam Factor

When $n \to \infty \gg k$,

$$-2 \ln P(D|\mathcal{M}_i) \approx -2 \ln \hat{L} + k \ln n + O(1)$$

$$BIC = k \ln n - 2 \ln \hat{L}$$

$\therefore$ Minimise BIC $\approx$ Maximise Evidence

n	Number of data
k	Number of parameters
D	Data
$\mathcal{M}_i$	Model
$\boldsymbol{\xi}$	Fitting parameters
$\mathbf{A}$	Parameter Covariance

Model Sparsification through priors

Selecting a model in a set of models defined by combinations of terms

The Role of Prior in Regression

$D = \{\boldsymbol{x}, \boldsymbol{y}\}$ Data

$\mathcal{M}_i$ Model

$\boldsymbol{\Theta}_{\mathcal{M}}(\boldsymbol{x})$ Model Basis

$\boldsymbol{\xi}$ Fitting parameters

Bayes' rule

likelihood prior

$$P(\boldsymbol{\xi}|D,\mathcal{M}_i) = \frac{P(D|\boldsymbol{\xi},\mathcal{M}_i)P(\boldsymbol{\xi}|\mathcal{M}_i)}{P(D|\mathcal{M}_i)}$$

posterior

evidence

Gaussian Linear model:
$\boldsymbol{y} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{\Theta}_{\mathcal{M}}(\boldsymbol{x}) \cdot \boldsymbol{\xi}, \sigma^2\mathbf{I})$

error

$$-\ln P(\boldsymbol{\xi}|D,\mathcal{M}_i) = \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{\Theta}_{\mathcal{M}}(\boldsymbol{x}) \cdot \boldsymbol{\xi}\|_F^2 + \gamma\|\boldsymbol{\xi}\| + normalise$$

posterior likelihood Prior (Regularisation)

Maximum a posteriori (MAP):

$$\boldsymbol{\xi}_{MAP} = \underset{\boldsymbol{\xi}}{\operatorname{argmax}}\, P(\boldsymbol{\xi}|D,\mathcal{M}_i)$$

=

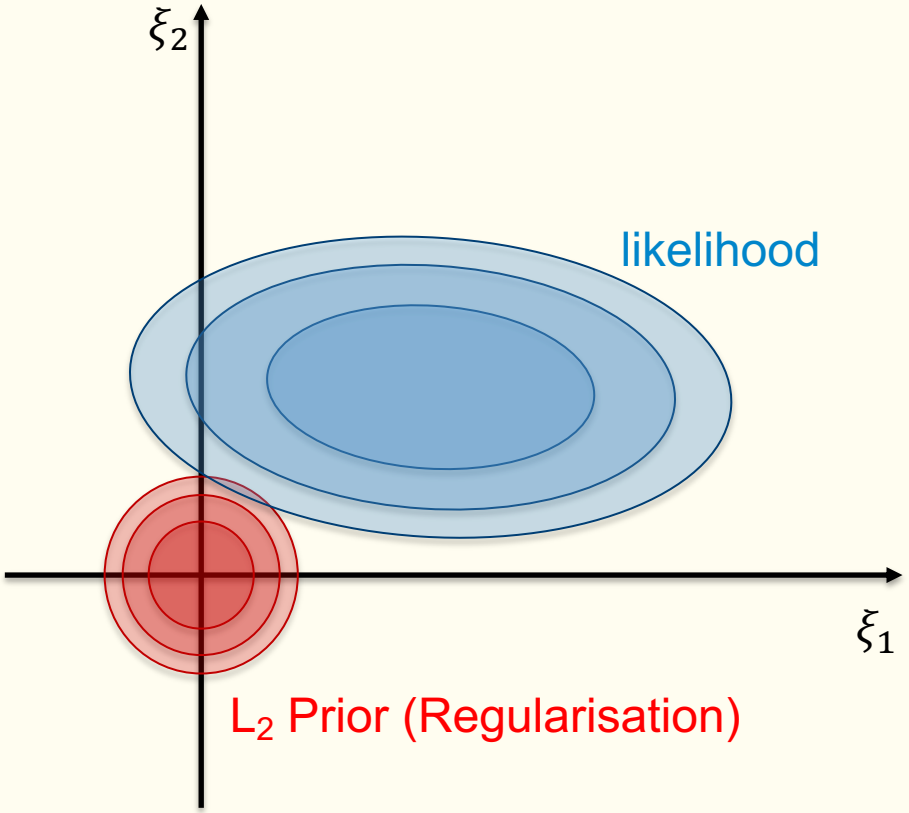
Regularised Regression
(Regularisation ↔ prior)

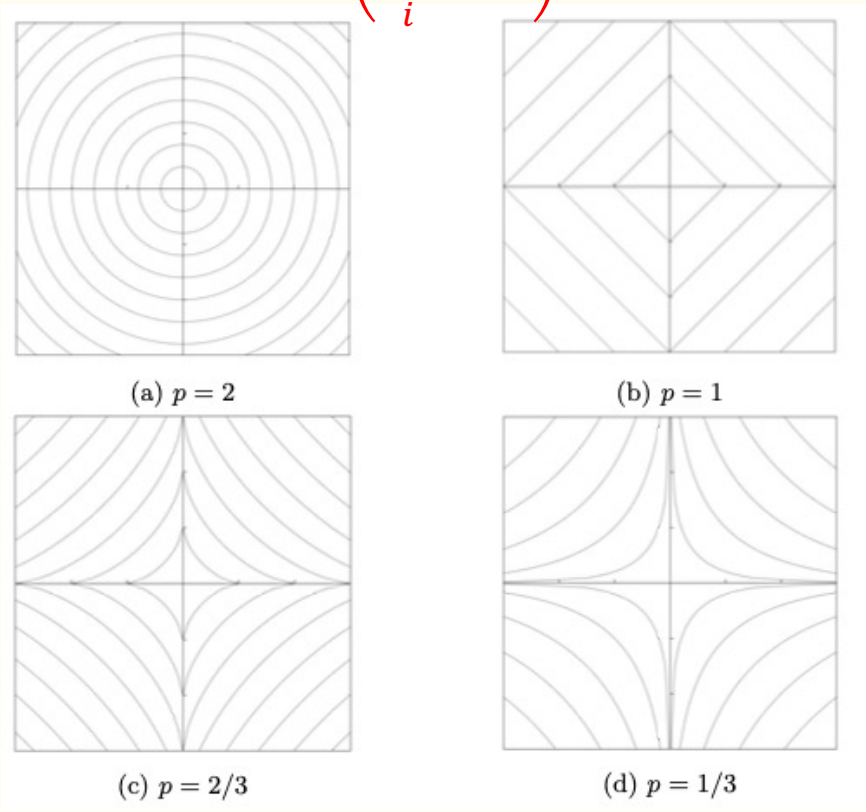
Types of prior

Regression	Loss	Regulariser	Prior
Ridge	$\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{\Theta}_{\mathcal{M}}(\boldsymbol{x})\cdot\boldsymbol{\xi}\|_F^2+\frac{\gamma}{2}\|\boldsymbol{\xi}\|_2^2$	$\frac{1}{2}\|\boldsymbol{\xi}\|_2^2=\frac{1}{2}\sum_i \xi_i^2$	Gaussian (L2) $\xi\sim\mathcal{N}(0,1)$
LASSO	$\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{\Theta}_{\mathcal{M}}(\boldsymbol{x})\cdot\boldsymbol{\xi}\|_F^2+\gamma\|\boldsymbol{\xi}\|_1$	$\|\boldsymbol{\xi}\|_1=\sum_i \|\xi_i\|$	Laplace (L1) $\xi\sim Laplace(0,1)$
Sequential Threshold	$\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{\Theta}_{\mathcal{M}}(\boldsymbol{x})\cdot\boldsymbol{\xi}\|_F^2+\gamma\|\boldsymbol{\xi}\|_0$	$\|\boldsymbol{\xi}\|_0$ $=\#\ of\ nonzero$	(No formal prior, L0)
Spike and Slab	(Ridge with $\|\boldsymbol{\xi}\|_2^2$ + loss for λ)	$\xi\|\lambda\sim\mathcal{N}(0,1)\lambda+\mathcal{N}(0,\epsilon^2)(1-\lambda)$ $\lambda\sim Ber(\pi)$	
Normal-Gamma	(Ridge with $\|\boldsymbol{\xi}\|_2^2$ + loss for α)	$\xi\|\alpha\sim\mathcal{N}(0,diag(\alpha^{-1}))$ $\alpha\sim Gamma(k,\theta)$	

Prior as sparsifier

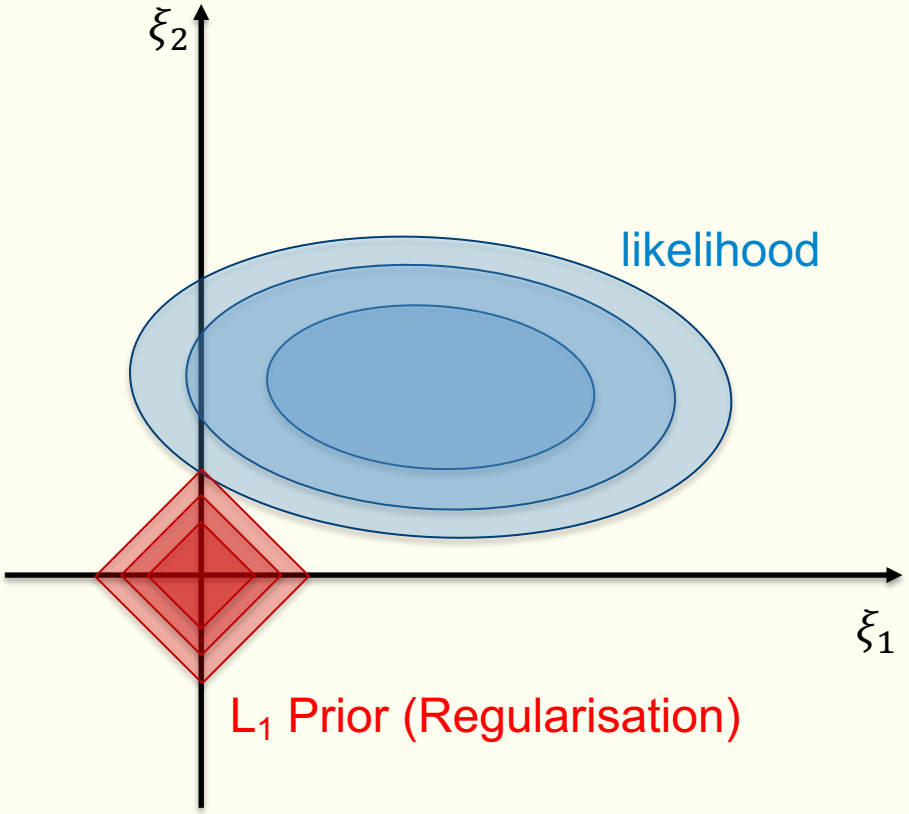
$$\|\boldsymbol{\xi}\|_{\boldsymbol{p}} = \left(\sum_i |\xi_i|^p\right)^{1/p}$$

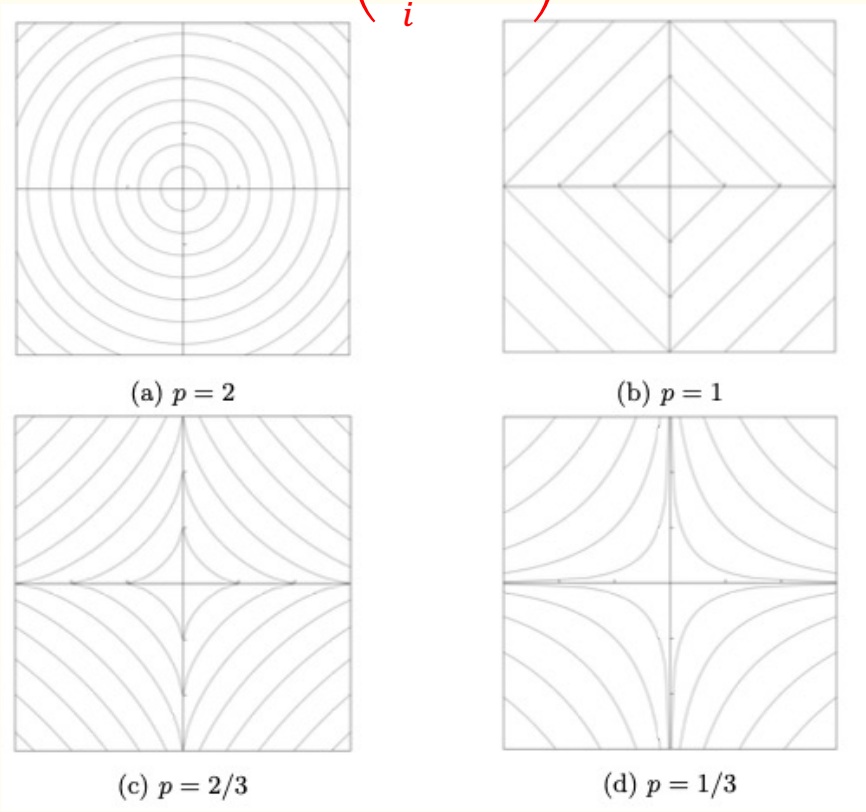




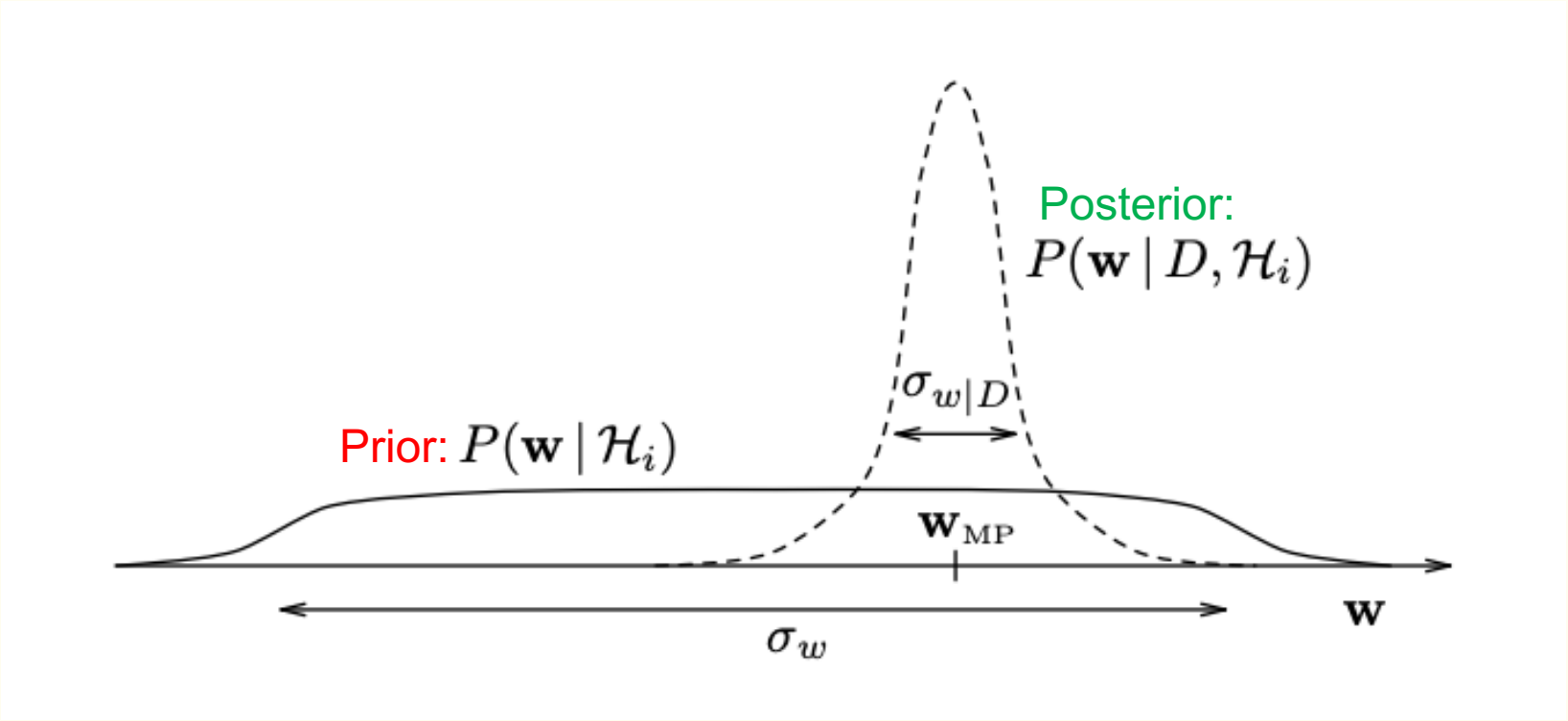
Prior as sparsifier

$$\|\boldsymbol{\xi}\|_{\boldsymbol{p}} = \left(\sum_i |\xi_i|^p\right)^{1/p}$$





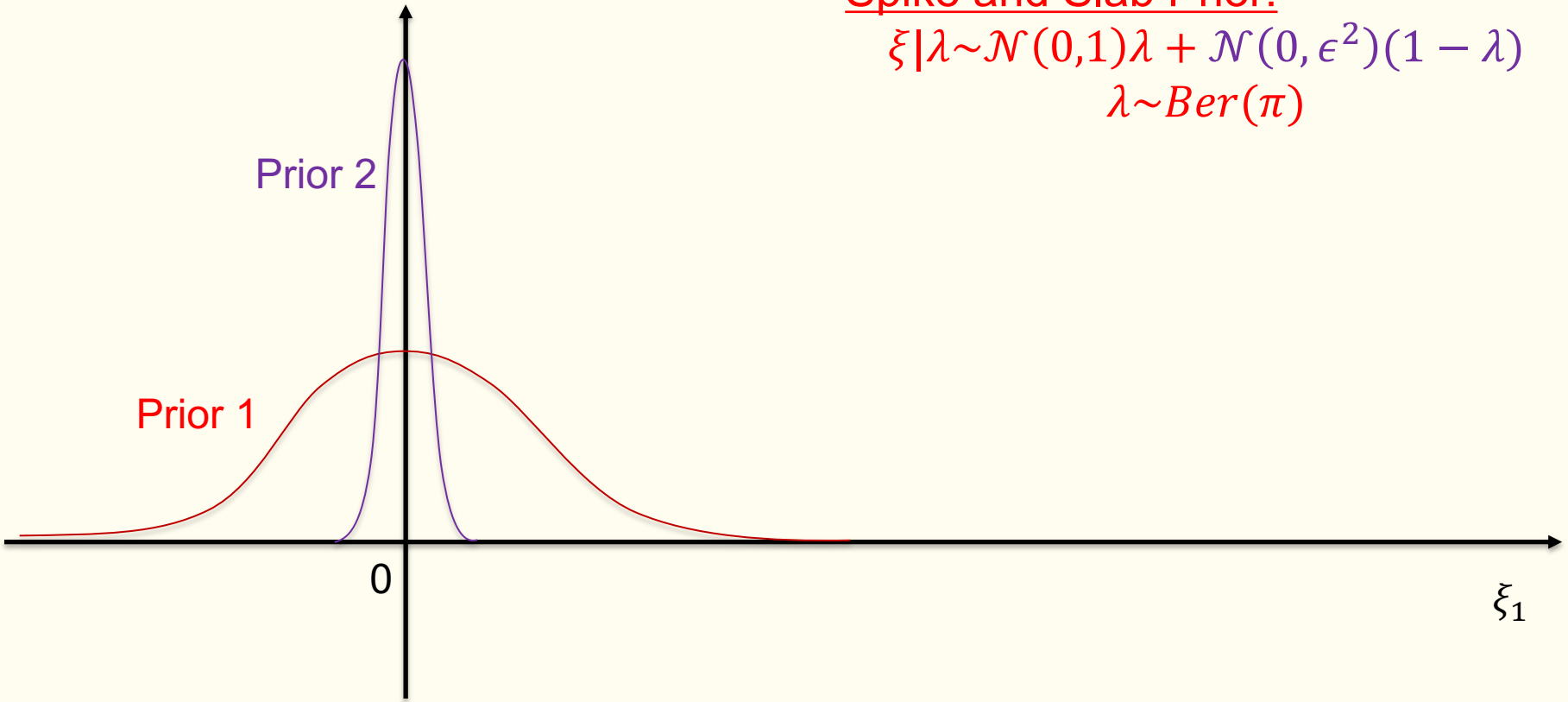
Prior as sparsifier – Hierarchical Prior



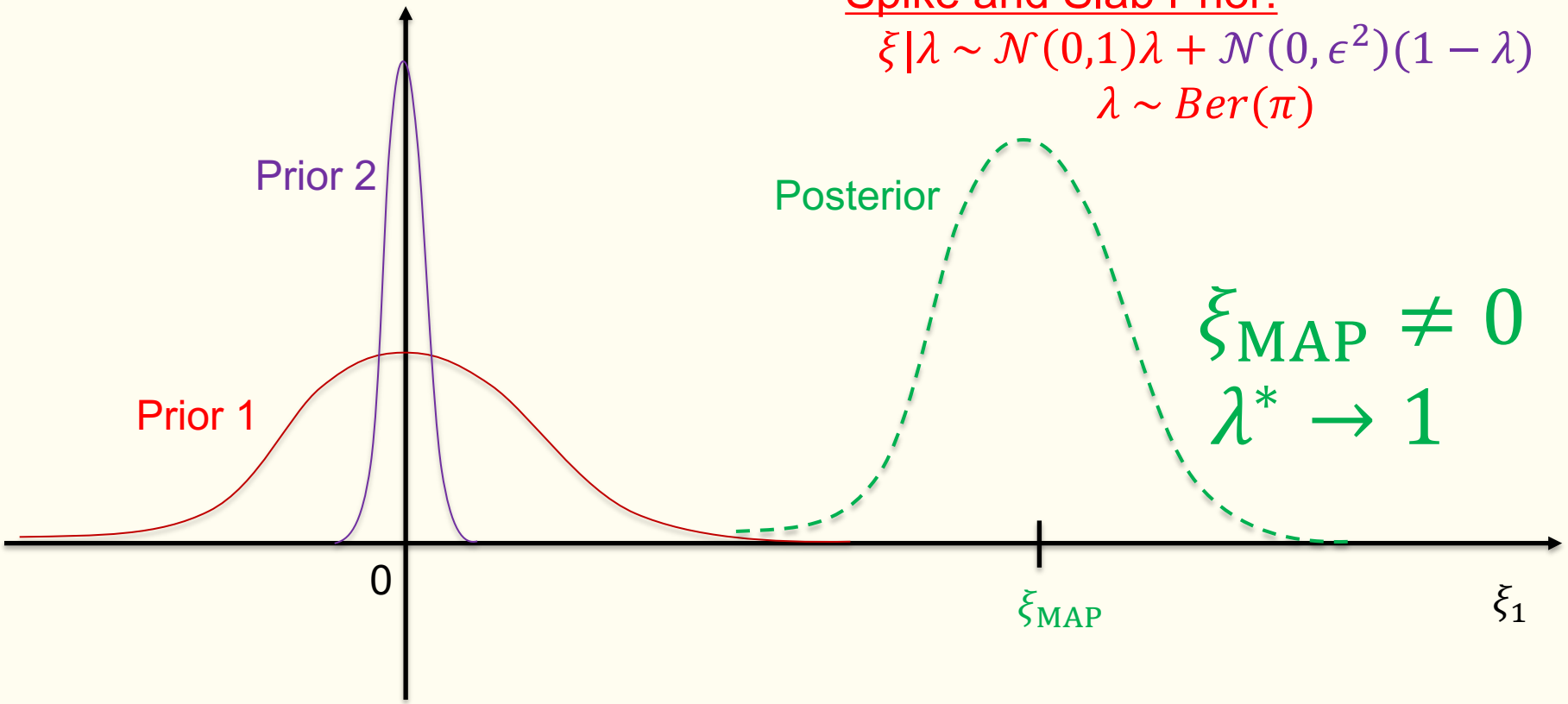
Prior as sparsifier – Hierarchical Prior

Spike and Slab Prior:

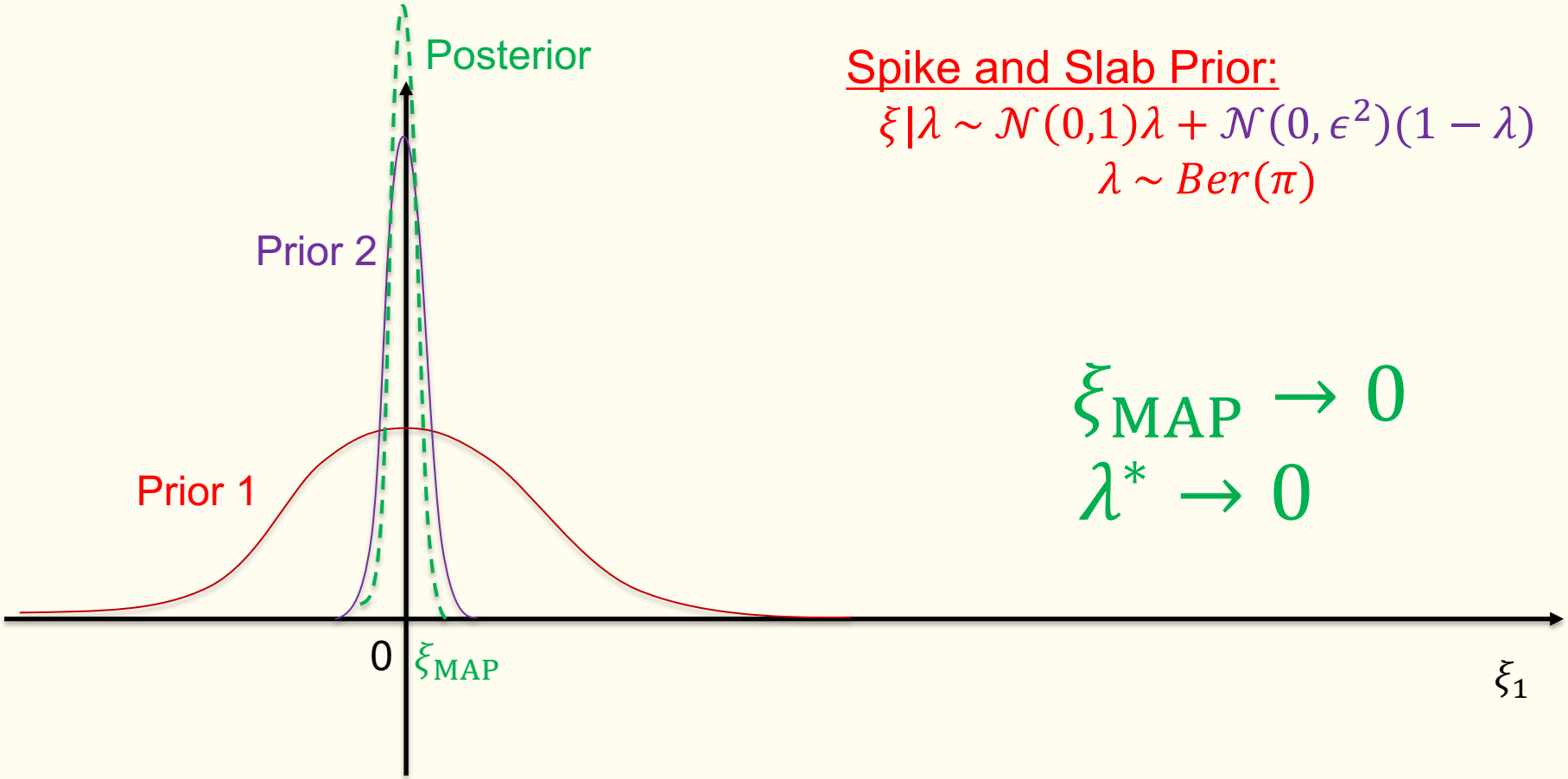
$$\xi|\lambda \sim \mathcal{N}(0,1)\lambda + \mathcal{N}(0,\epsilon^2)(1-\lambda)$$
$$\lambda \sim Ber(\pi)$$



Prior as sparsifier – Hierarchical Prior



Prior as sparsifier – Hierarchical Prior

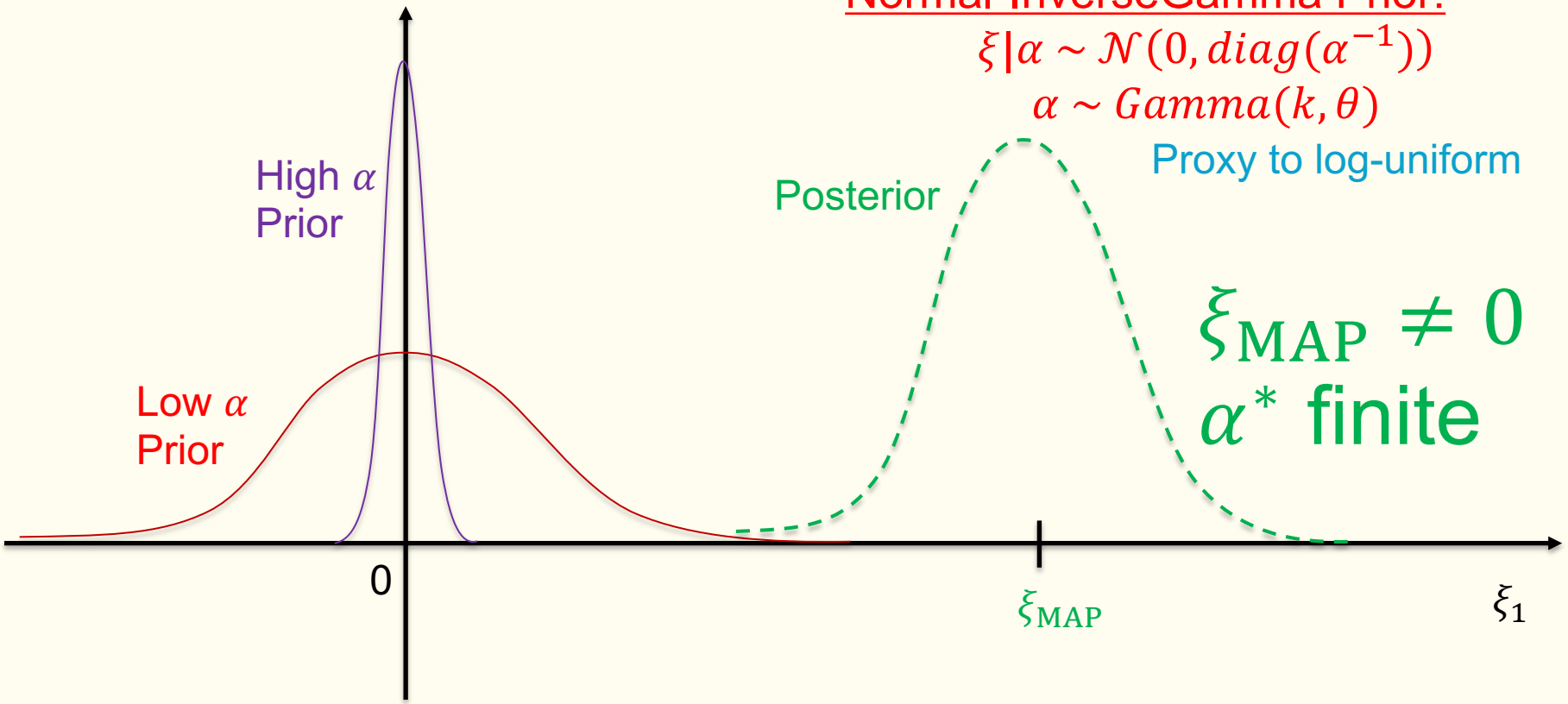


Spike and Slab Prior:

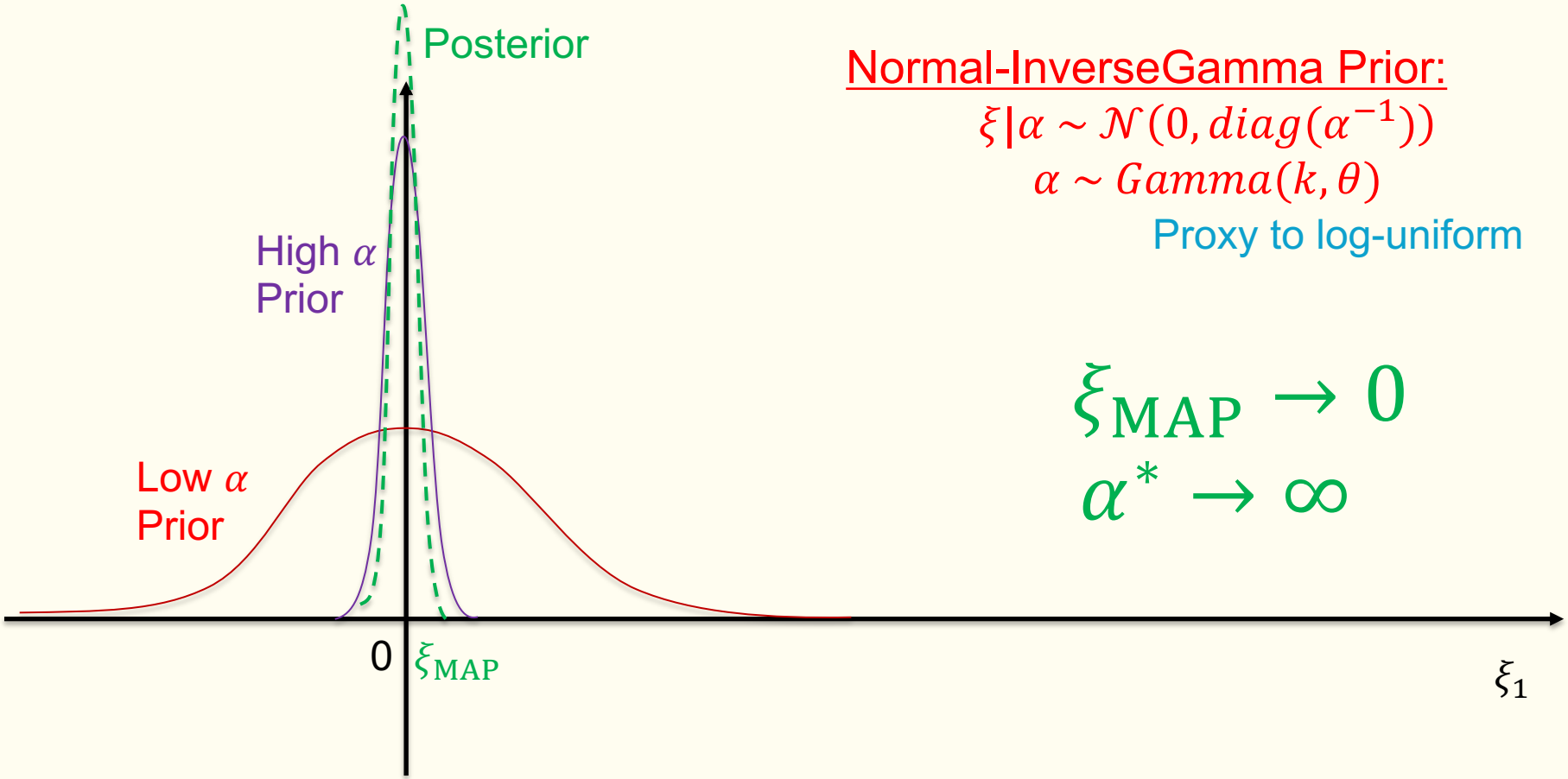
$$\xi|\lambda \sim \mathcal{N}(0,1)\lambda + \mathcal{N}(0,\epsilon^2)(1-\lambda)$$
$$\lambda \sim Ber(\pi)$$

$$\xi_{\mathrm{MAP}} \to 0$$
$$\lambda^* \to 0$$

Prior as sparsifier – Hierarchical Prior



Prior as sparsifier – Hierarchical Prior



Normal-InverseGamma Prior:

$$\xi|\alpha \sim \mathcal{N}(0, diag(\alpha^{-1}))$$

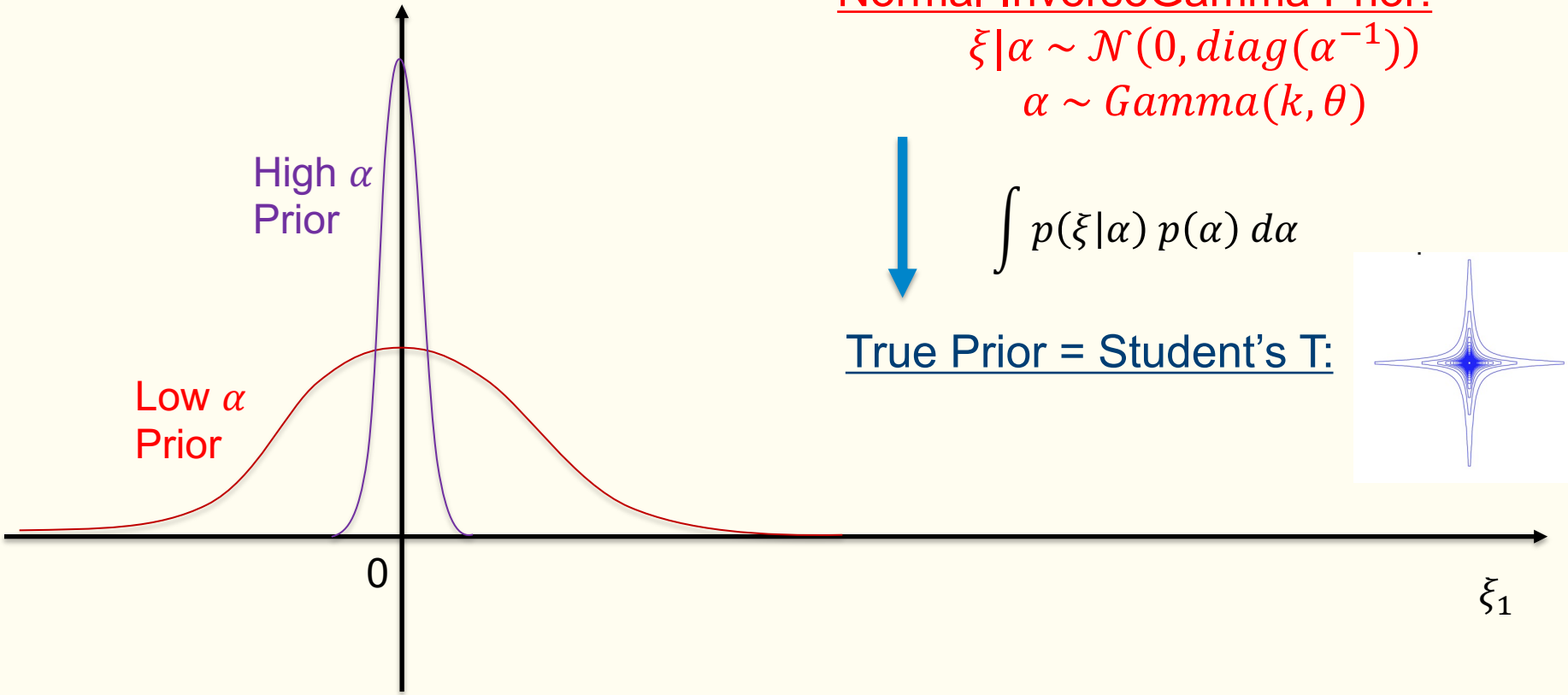
$$\alpha \sim Gamma(k, \theta)$$

Proxy to log-uniform

$$\xi_{\mathrm{MAP}} \to 0$$

$$\alpha^* \to \infty$$

Prior as sparsifier – Hierarchical Prior



Normal-InverseGamma Prior:

$$\xi|\alpha \sim \mathcal{N}(0, diag(\alpha^{-1}))$$
$$\alpha \sim Gamma(k, \theta)$$

$$\int p(\xi|\alpha)\, p(\alpha)\, d\alpha$$

True Prior = Student's T:

Final remark on sparsifying priors VS Bayesian evidence

- Selecting model by Bayesian evidence is to formally think about

$$P(\mathcal{M}_i|D) = P(D|\mathcal{M}_i)P(\mathcal{M}_i)$$

  - FORMAL!
  - Fundamentally Bayesian, Fundamentally probabilistic
  - Can deal with heterogeneous datasets (parameters are marginalised out)

- Sparsifying (selecting) model by sparsifying prior
  - Rooted in “regression” typed problem
  - (not as) formal
  - More popular (because they’re “faster”?)

Tutorial: 1D Polynomial Regression

Data Generation

- Generate {x,y} data pair from some randomly selected sum of polynomial terms
- Add Gaussian noise to y with known variance σ^2

Regression

- Perform linear regression (Frequentist: OLS; Bayesian: Ridge) on a polynomial library
- Find Evidence
- Remove terms in the library and redo regression; Find the combination of terms (model) that give max evidence

Experimentation

- How does changing σ^2 in the input of the regression changes the model selection?
  - How does that correspond to the intuition of Occam Razor?
- How can one further optimise the prior variance (a hyperparameter)?