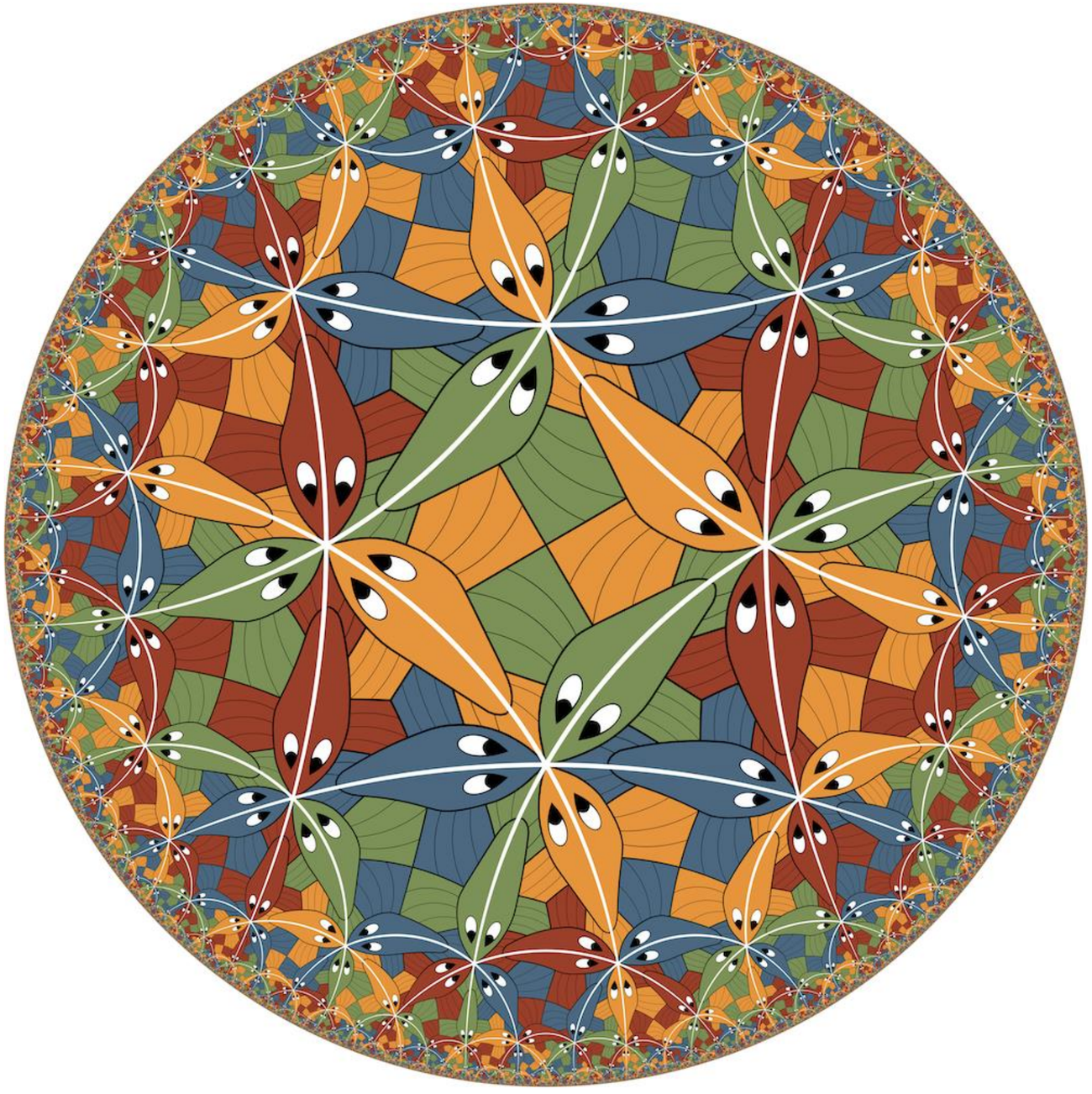
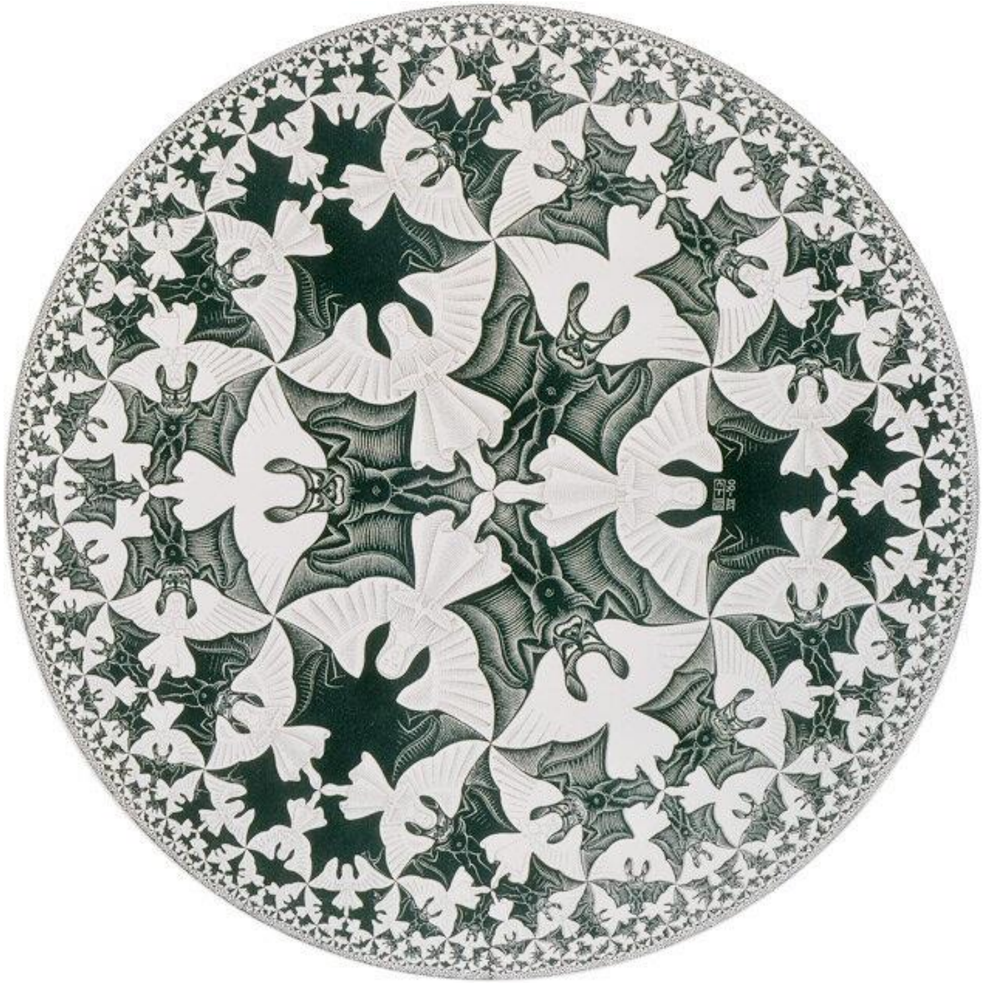
Neural networks and **information geometry**

Bill Oxbury

6 May 2025

School of Mathematical Sciences | Lancaster University





$$w = \frac{z - i}{z + i}$$

$$z = -i\frac{w + 1}{w - 1}$$

Univariate normal distribution parameter space $z = \mu + i\sigma$

Fisher metric on a model family

In the space of distributions, use **Kullback-Leibler divergence** as local distance:

$$\mathrm{KL}(\theta||\theta+\delta) = \int p(x|\theta)\log\frac{p(x|\theta)}{p(x|\theta+\delta)}dx$$

Fisher metric on a model family

In the space of distributions, use **Kullback-Leibler divergence** as local distance:

$$\mathrm{KL}(\theta||\theta+\delta) = \int p(x|\theta)\log\frac{p(x|\theta)}{p(x|\theta+\delta)}dx$$

By taking Taylor expansion

$$\log p(x|\theta+\delta) \approx \log p(x|\theta) + \delta^T\nabla_\theta \log p(x|\theta) + \frac{1}{2}\delta^T\nabla_\theta^2 \log p(x|\theta)\delta$$

we get a symmetric approximation

$$\mathrm{KL}(\theta||\theta+\delta) \approx \frac{1}{2}\delta^T I(\theta)\delta + o(||\delta||^2)$$

where

$$I(\theta) = -\mathbb{E}[\nabla_\theta^2 \log p(x|\theta)] = \mathbb{E}[\nabla_\theta \log p(x|\theta)\cdot(\nabla_\theta \log p(x|\theta))^T]$$

is called the **Fisher Information Matrix**.

Fisher metric on a model family

Normal distribution $\theta = (\mu, \sigma)$

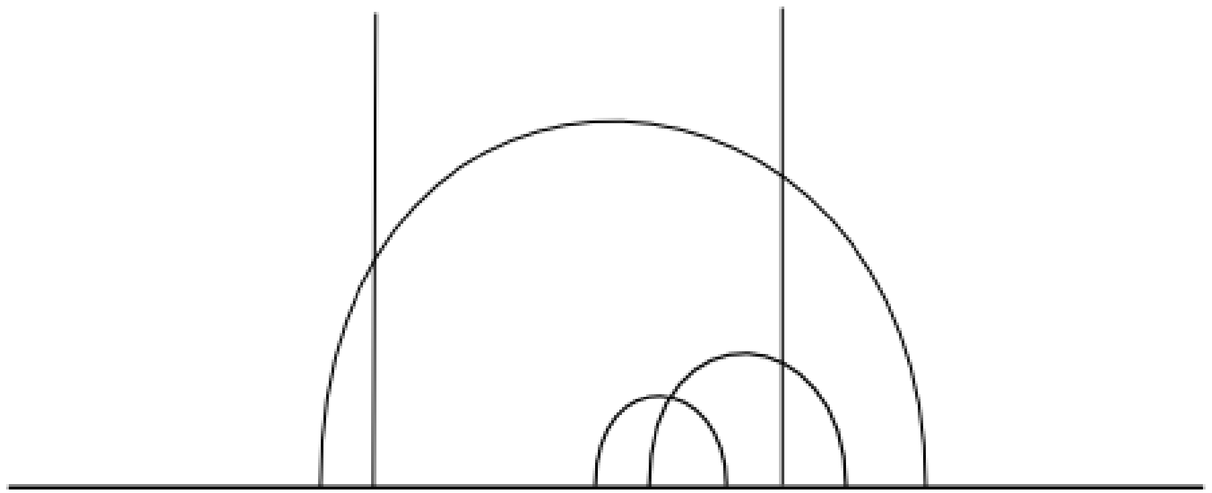
$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Differentiating gives FIM

$$I(\mu, \sigma) = \begin{pmatrix} \frac{1}{\sigma^2} & \\ & \frac{2}{\sigma^2} \end{pmatrix}$$

corresponding to hyperbolic metric:

$$ds^2 = \frac{d\mu^2 + 2d\sigma^2}{\sigma^2}$$



Fisher metric on a model family

How does this geometry look for more complex model families such as neural networks?

We can think about two practical applications:

1. Natural gradient descent
2. Model selection

Natural gradient descent

Standard gradient descent update is

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L(\theta)$$

This comes from solving the constrained optimization problem

$$\min_\delta L(\theta + \delta) \quad \text{s.t.} \quad ||\delta|| \leq \epsilon$$

Better is to solve

$$\min_\delta L(\theta + \delta) \quad \text{s.t.} \quad \mathrm{KL}(p(x|\theta) \,||\, p(x|\theta + \delta)) \approx \delta^T I(\theta) \delta \leq \epsilon$$

This leads to natural gradient descent update

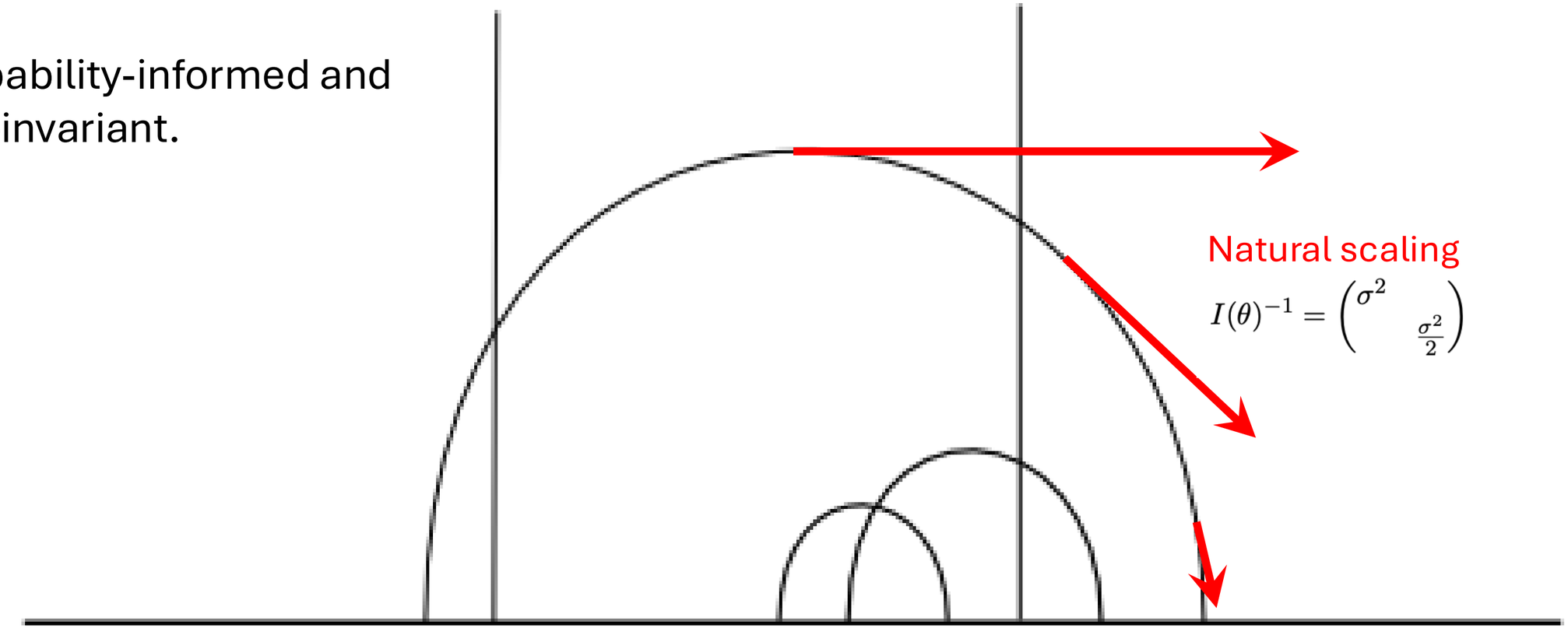
$$\theta_{t+1} = \theta_t - \eta \widetilde{\nabla}_\theta L(\theta), \qquad \text{where } \widetilde{\nabla}_\theta = \mathbf{I}(\theta)^{-1} \nabla_\theta$$

Natural gradient descent

Gradient vectors are scaled by the inverse FIM

$$\widetilde{\nabla}_\theta = \mathbf{I}(\theta)^{-1}\nabla_\theta$$

This is probability-informed and parameter-invariant.

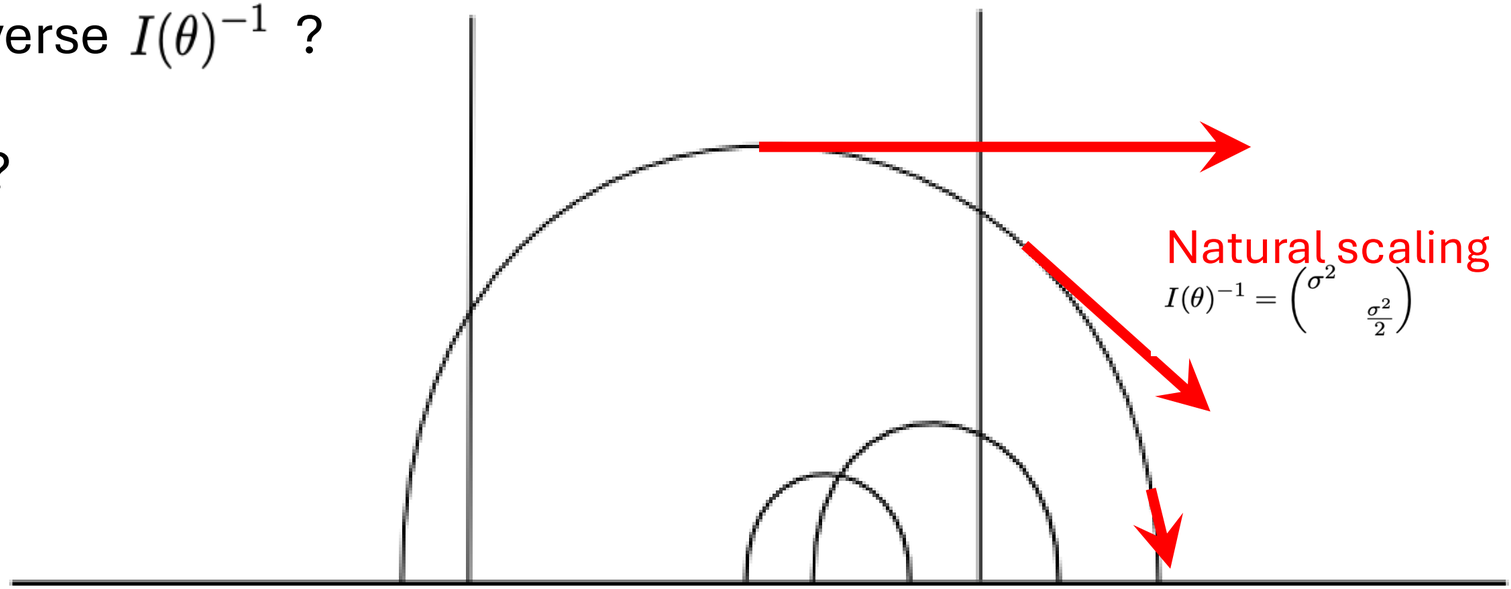


Natural gradient descent

For **neural networks** life is less simple.

$\theta \in \mathbb{R}^k$ where $k \sim 10^6 \nearrow 10^{10}$

- How do we compute the $k \times k$ matrix $I(\theta)$?
- How do we compute the inverse $I(\theta)^{-1}$?
- What if $I(\theta)^{-1}$ doesn't exist?

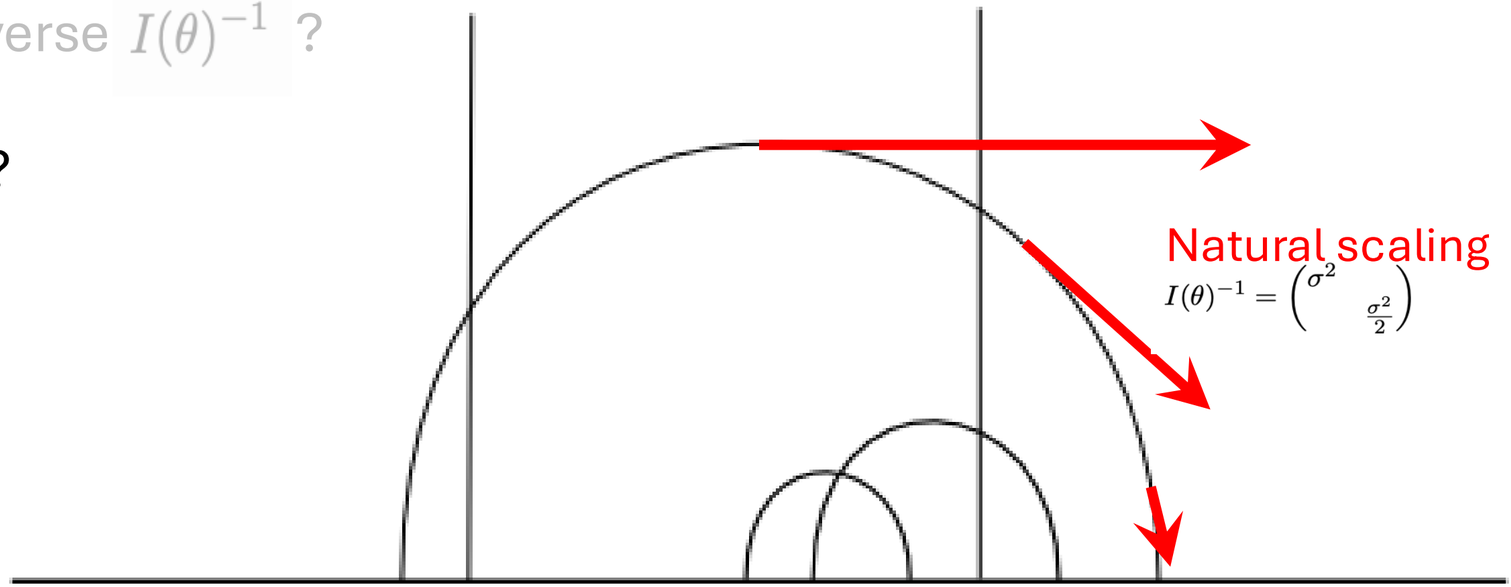


Natural gradient descent

For **neural networks** life is less simple.

$\theta \in \mathbb{R}^k$ where $k \sim 10^6 \nearrow 10^{10}$

- How do we compute the $k \times k$ matrix $I(\theta)$?
- How do we compute the inverse $I(\theta)^{-1}$?
- What if $I(\theta)^{-1}$ doesn't exist?



A toy example

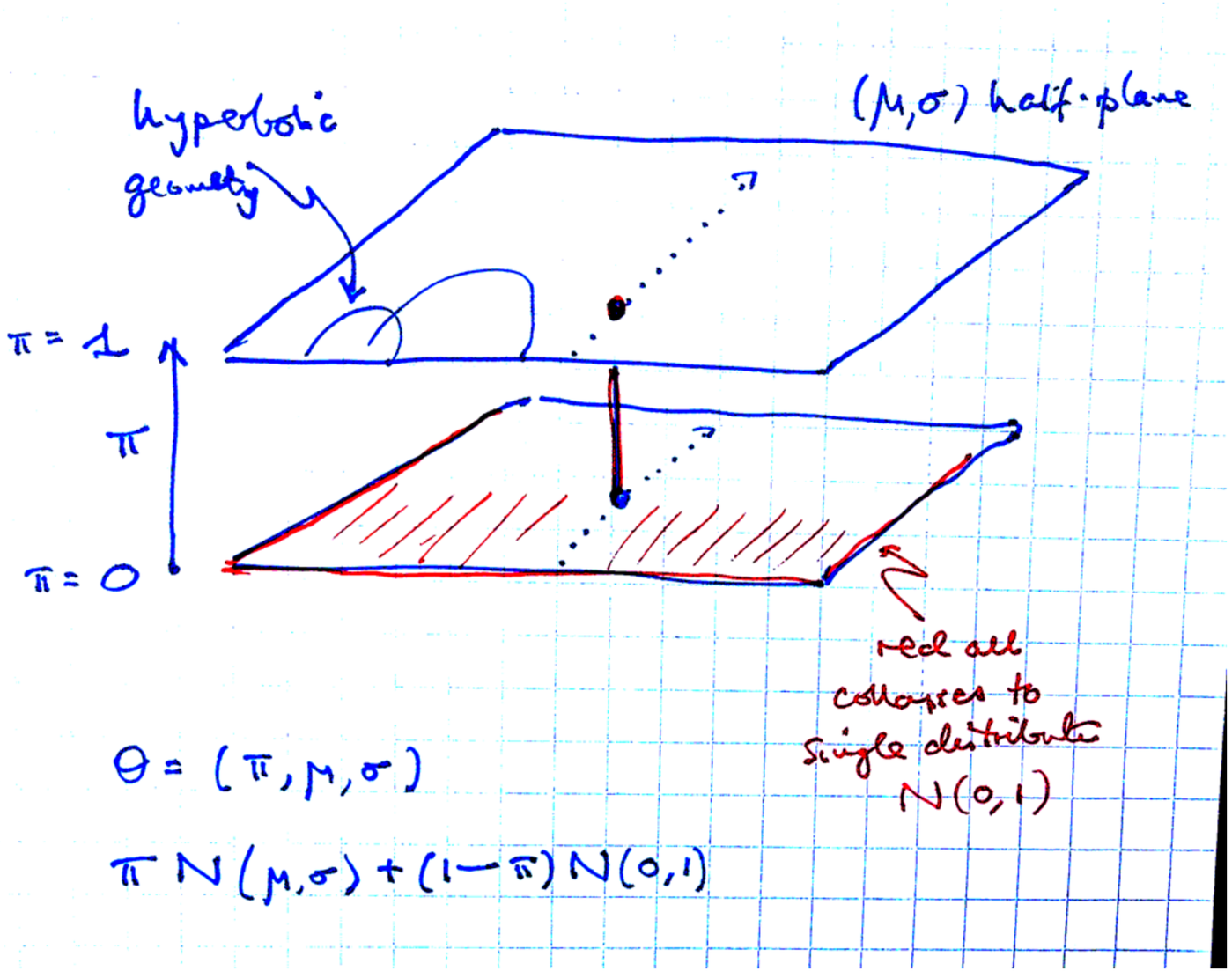
Consider 3-parameter (for ease of visualisation) mixture model

$$\pi\mathcal{N}(\mu,\sigma)+(1-\pi)\mathcal{N}(0,1)$$

The FIM

$$I(\theta)=-\mathbb{E}\left[\nabla^2_\theta \log p(x\mid\theta)\right]$$

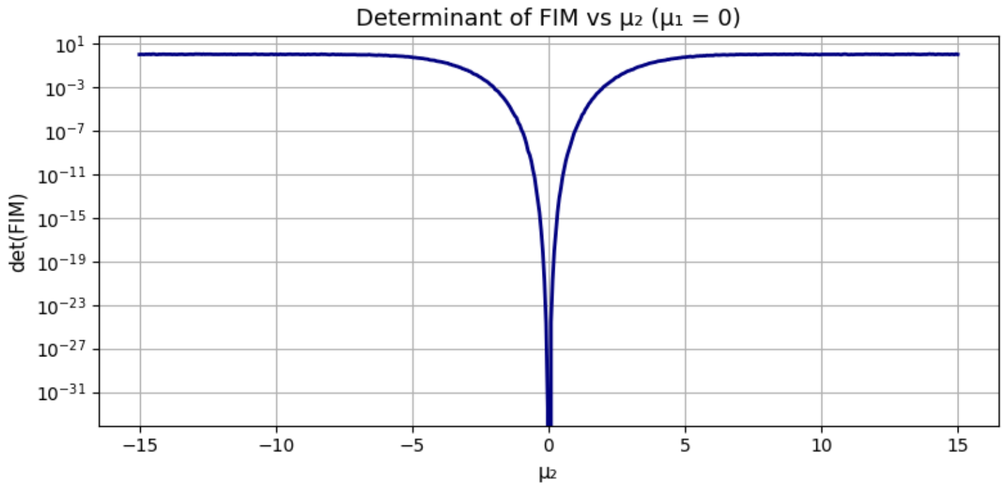
can be estimated by Monte Carlo for any $\theta=(\pi,\mu,\sigma)$

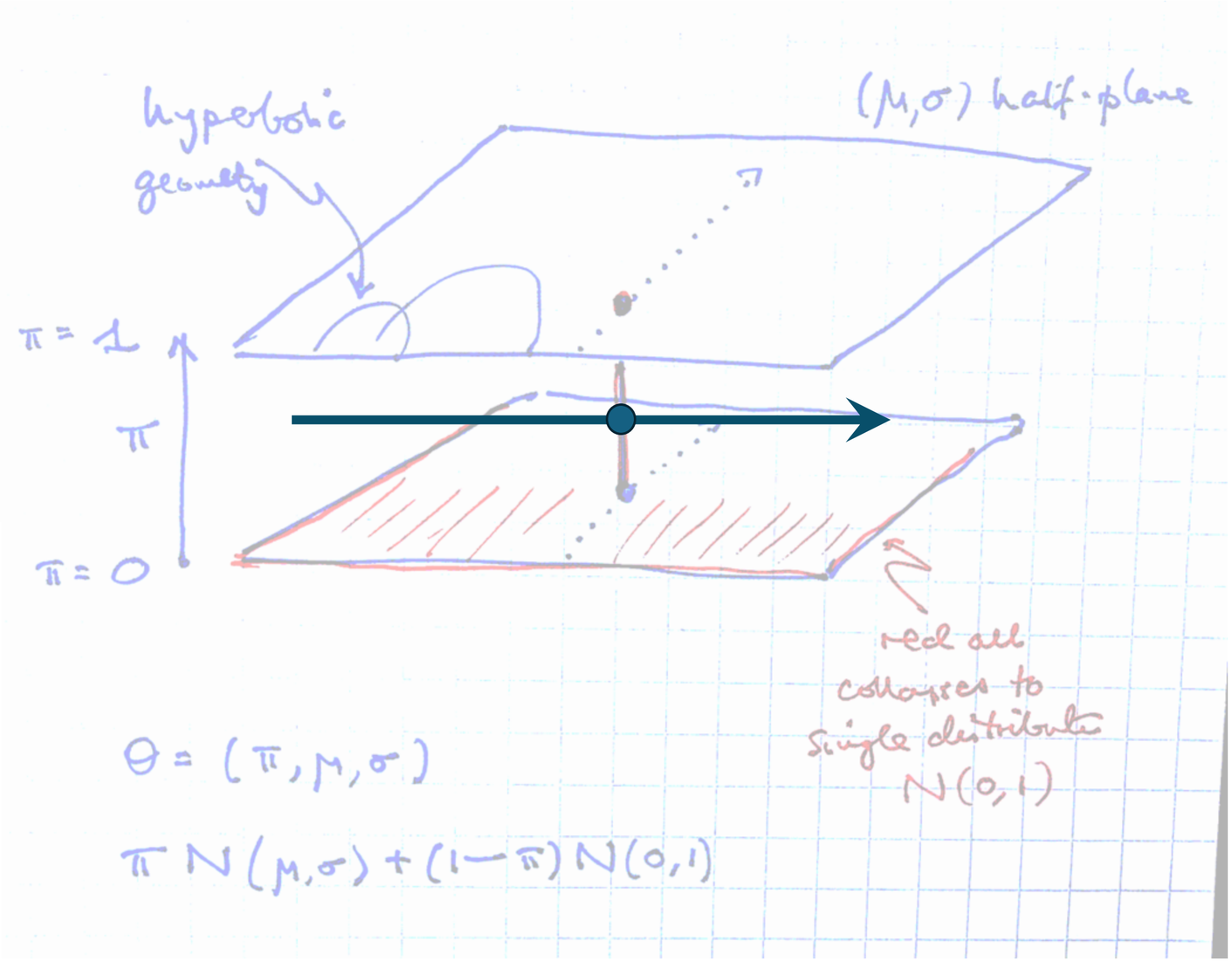


A toy example

Hold $\pi = 0.5, \sigma = 1$.

Plot $\det I(\theta)$ with μ:

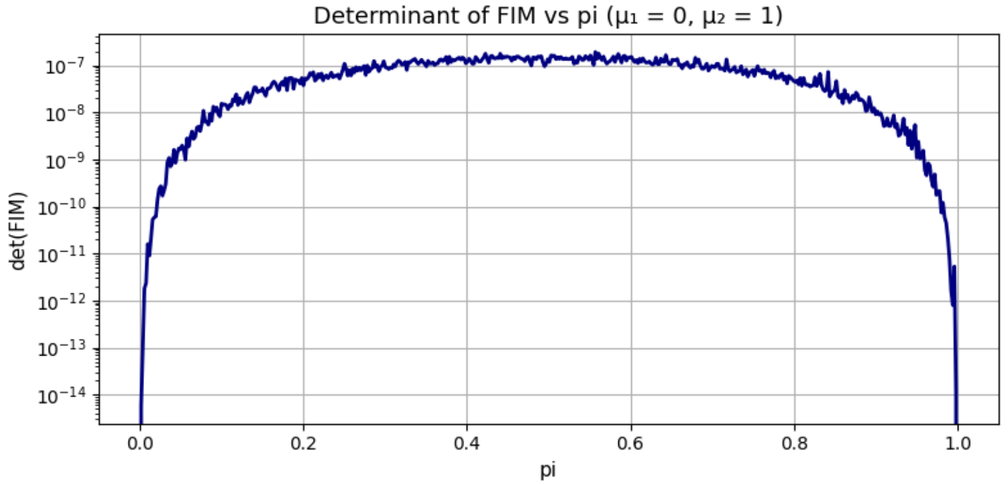


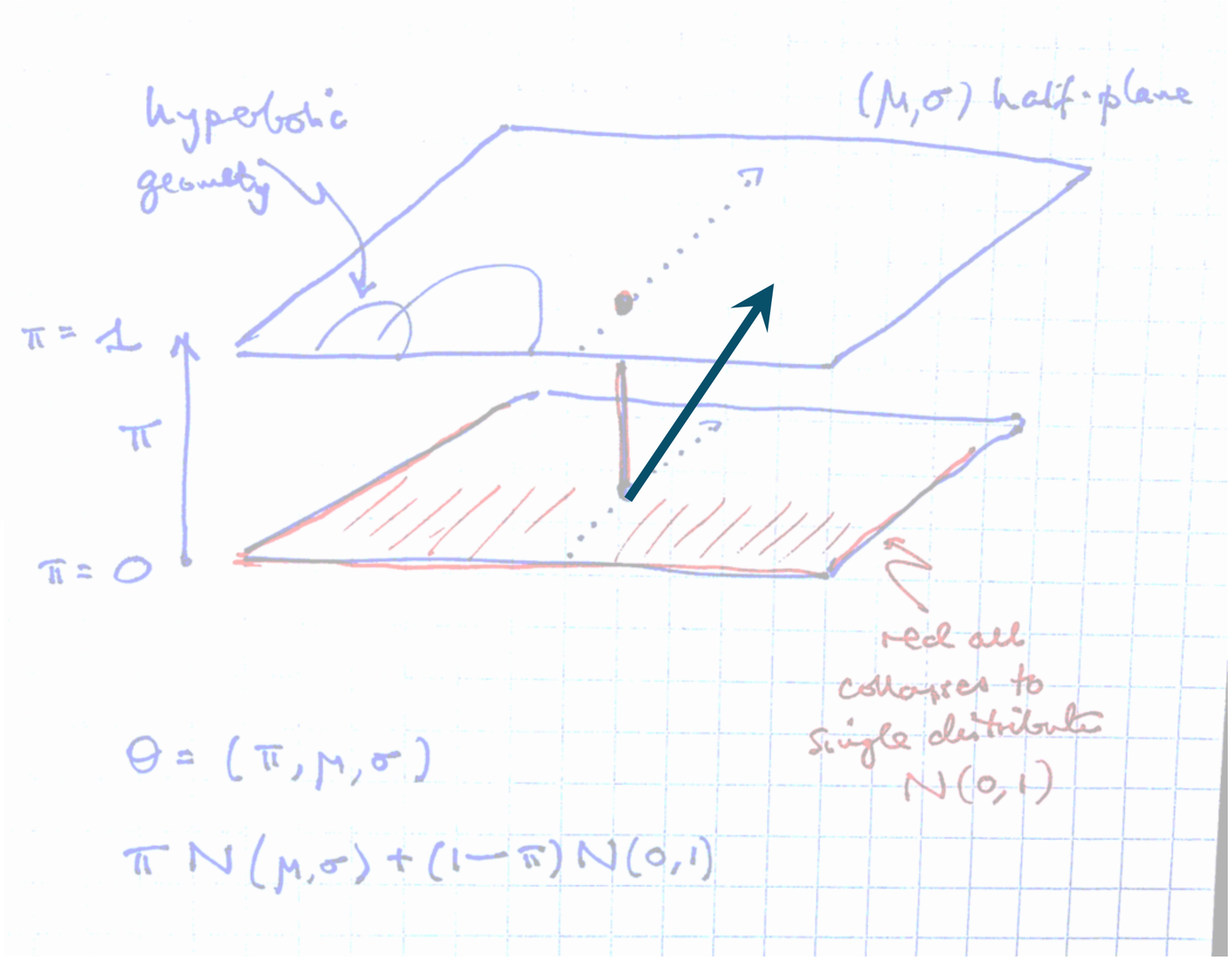


A toy example

Hold $\mu = 1, \sigma = 1$.

Plot $\det I(\theta)$ with π:





Bayesian model selection

Model family with parameter space $\Theta \subset \mathbb{R}^k$

Training data $D = \{x_1, \ldots, x_n\}$

Model fitting means minimizing $L(\theta) = -\mathbb{E} \log p(x|\theta)$

Empirically $L_D(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i|\theta)$

Bayes posterior is

$$p(\theta \mid D) = \frac{1}{Z_D} \phi(\theta) e^{-nL_D(\theta)}$$

Bayesian model selection

Bayes posterior

$$p(\theta \mid D) = \frac{1}{Z_D}\phi(\theta)e^{-nL_D(\theta)}$$

Denominator is called the model evidence:

$$Z_D = \int_\Theta e^{-nL_D(\theta)}\phi(\theta)d\theta$$

Irrelevant at model fitting but is the likelihood of the model class (e.g. neural network architecture) given the data.

Needed at model selection.

Bayesian model selection

Claim: asymptotically

$$\log Z_D = -\mathrm{BIC} + O(1), \quad n \to \infty$$

where BIC is the Bayesian Information Criterion

$$\mathrm{BIC} = nL_D(\theta_{\mathrm{MP}}) + \frac{k}{2}\log n$$

(MP means ‘maximum posterior’.)

Model selection by minimizing BIC penalises $k = \dim \Theta$

Yet neural network models can generalise well with extremely high dimension.

Singular models

A model family is called regular if

- the map $\Theta \longrightarrow \mathcal{P}$ is injective
- Fisher matrix is nonsingular $\det I(\theta) \neq 0$ for all $\theta \in \Theta$

Otherwise it's called strictly singular.

Machine Learning's dirty secret:
most useful model families (*mixture models, HMMs, neural networks,* ...) are strictly singular!

And BIC is only valid for regular models!

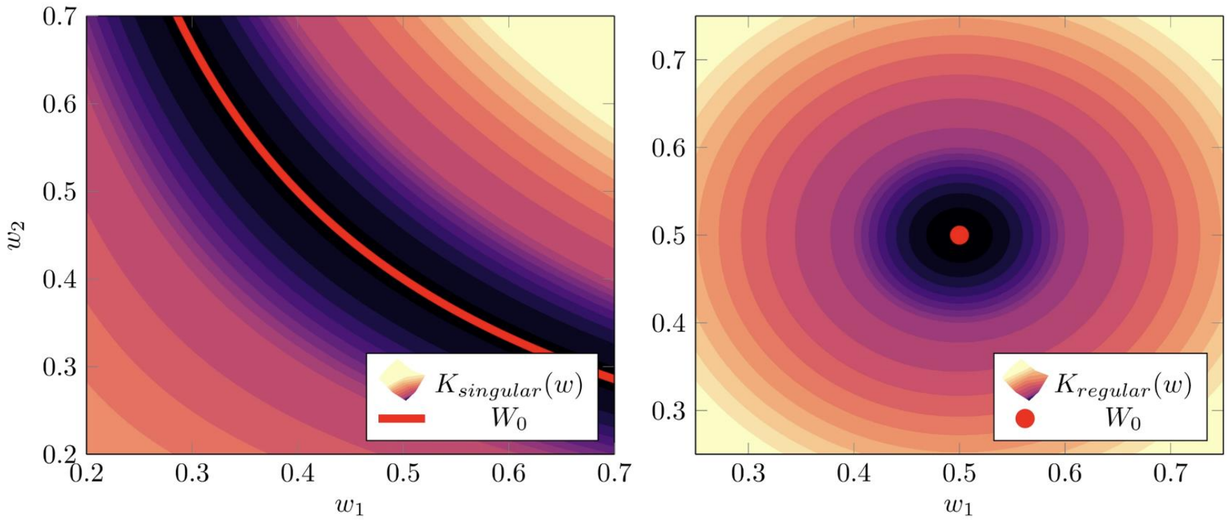


Image source: ref [3]

Singular models

Let's justify the claim. We needed asymptotic approximation for the model evidence

$$Z_D = \int_\Theta e^{-nL_D(\theta)}\phi(\theta)d\theta$$

Start with Taylor approximation

$$L_D(\theta) \approx L_D(\theta_{\mathrm{MP}}) + \frac{1}{2}(\theta - \theta_{\mathrm{MP}})I(\theta_{\mathrm{MP}})(\theta - \theta_{\mathrm{MP}})^T$$

Equivalent to approximating the posterior as a normal distribution in a neighbourhood of θ_{MP} with covariance matrix $I(\theta_{\mathrm{MP}})^{-1}$

When we substitute into the integral, it turns out that, assuming $\det I(\theta_{\mathrm{MP}}) \neq 0$, we can infer

$$Z_D = e^{-nL_D(\theta_{\mathrm{MP}})}\left(\frac{2\pi}{n}\right)^{k/2}\frac{\phi(\theta_{\mathrm{MP}})}{\sqrt{\det I(\theta_{\mathrm{MP}})}}$$

Singular models

Sketch proof: Start from *Gauss integral*

$$\int_{\mathbb{R}^k} e^{-(c/2)||w||^2} dw = \left(\frac{2\pi}{c}\right)^{k/2}$$

Using this, the *Laplace approximation* says suppose $f : \mathbb{R}^k \to \mathbb{R}$ has nondegenerate Hessian at the origin $\det \nabla^2 f(0) \neq 0$, and we're interested in

$$Z(n) = \int_{\mathbb{R}^k} e^{-nf(w)} dw$$

Then asymptotically as $n \to \infty$,

$$Z(n) \approx e^{-nf(0)} n^{-k/2} \sqrt{\frac{(2\pi)^k}{\det \nabla^2 f(0)}}$$

The formula for Z_D follows by applying Laplace to $f(w) = L_D(w - \theta_{\mathrm{MP}})$.

Generalising BIC

Arnol'd, Gusein-Zade, Varchenko (1985): a general $Z_D = \int_\Theta e^{-nL_D(\theta)}\phi(\theta)d\theta$ has asymptotic expansion

$$Z_D \approx e^{-nL_D(\theta_{\mathrm{MP}})} \cdot Cn^{-\lambda}(\log n)^{\nu-1} \quad \text{as } n \to \infty$$

for some constants $C \in \mathbb{R}$, $\lambda \in \mathbb{Q}$, $\nu \in \mathbb{N}$

So for model selection we should use

$$\mathrm{WBIC} = nL_D(\theta_{\mathrm{MP}}) + \lambda \log n - (\nu - 1) \log \log n$$

So what do (λ, ν) mean and how do we calculate them?

Generalising BIC

Watanabe (~2009):

$\lambda \leq \frac{k}{2}$ with equality iff the model family is **regular**;

$\lambda \leq \frac{1}{2}\operatorname{codim} \Theta_0$ the preimage in Θ of $p(\cdot \mid \theta_{\mathrm{MP}})$ with equality if Θ_0 is **minimally singular**

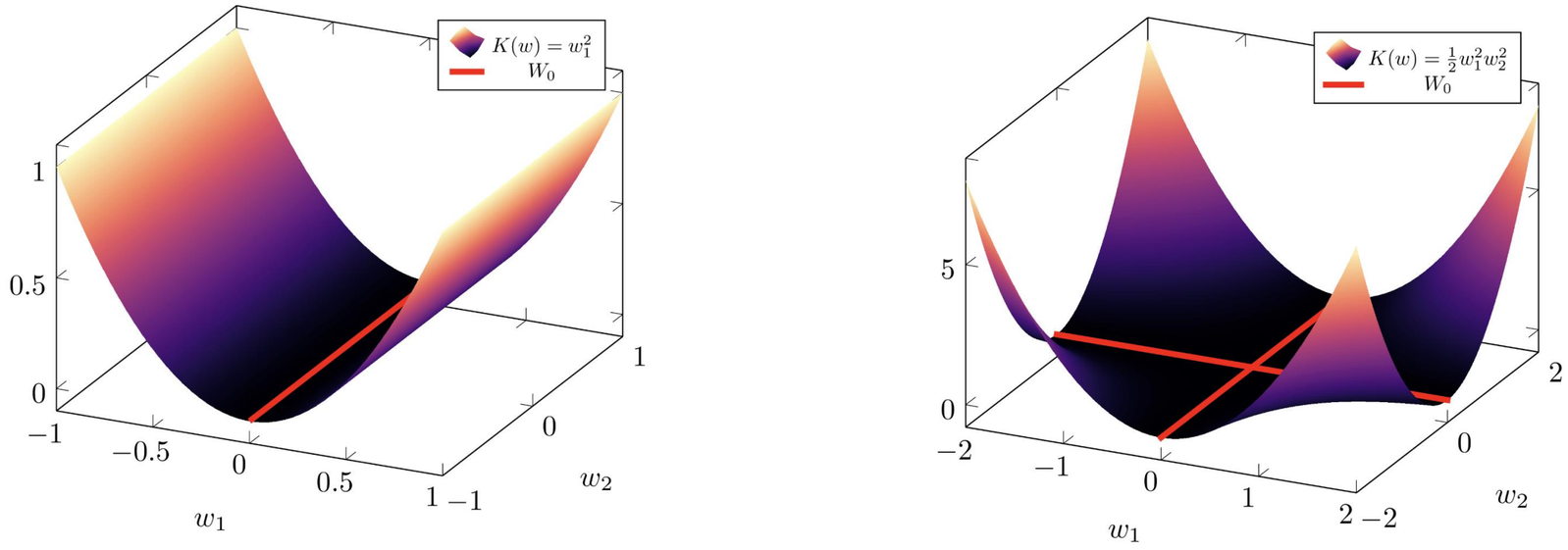


Image source: ref [3]

Generalising BIC

Monomial case (Arnol'd, Gusein-Zade, Varchenko):

Suppose $L_D(\theta) = \theta^\kappa$ and $\phi(\theta) = \theta^\tau$ where $\kappa, \tau \in \mathbb{Z}_{\geq 0}$

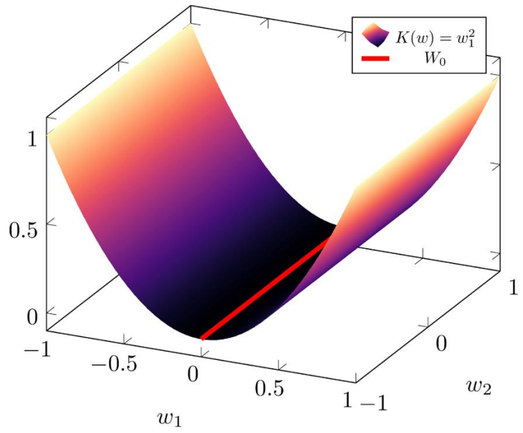
Then

$$Z_D \approx C \cdot n^{-\lambda} (\log n)^{\nu - 1}$$

where

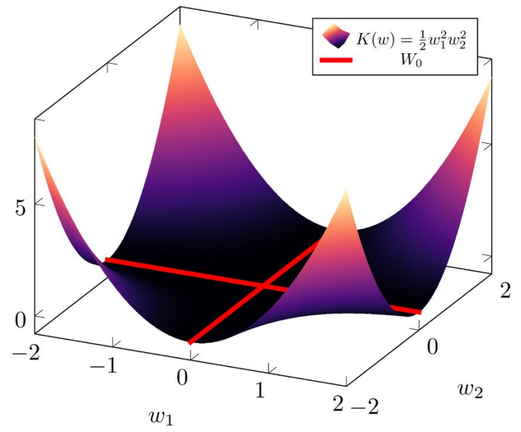
$$\lambda = \min_i \frac{\tau_i + 1}{\kappa_i}$$

$$\nu = \text{multiplicity of this minimum}$$



$L(x, y) = x^2$

$(\lambda, \nu) = (\frac{1}{2}, 1)$



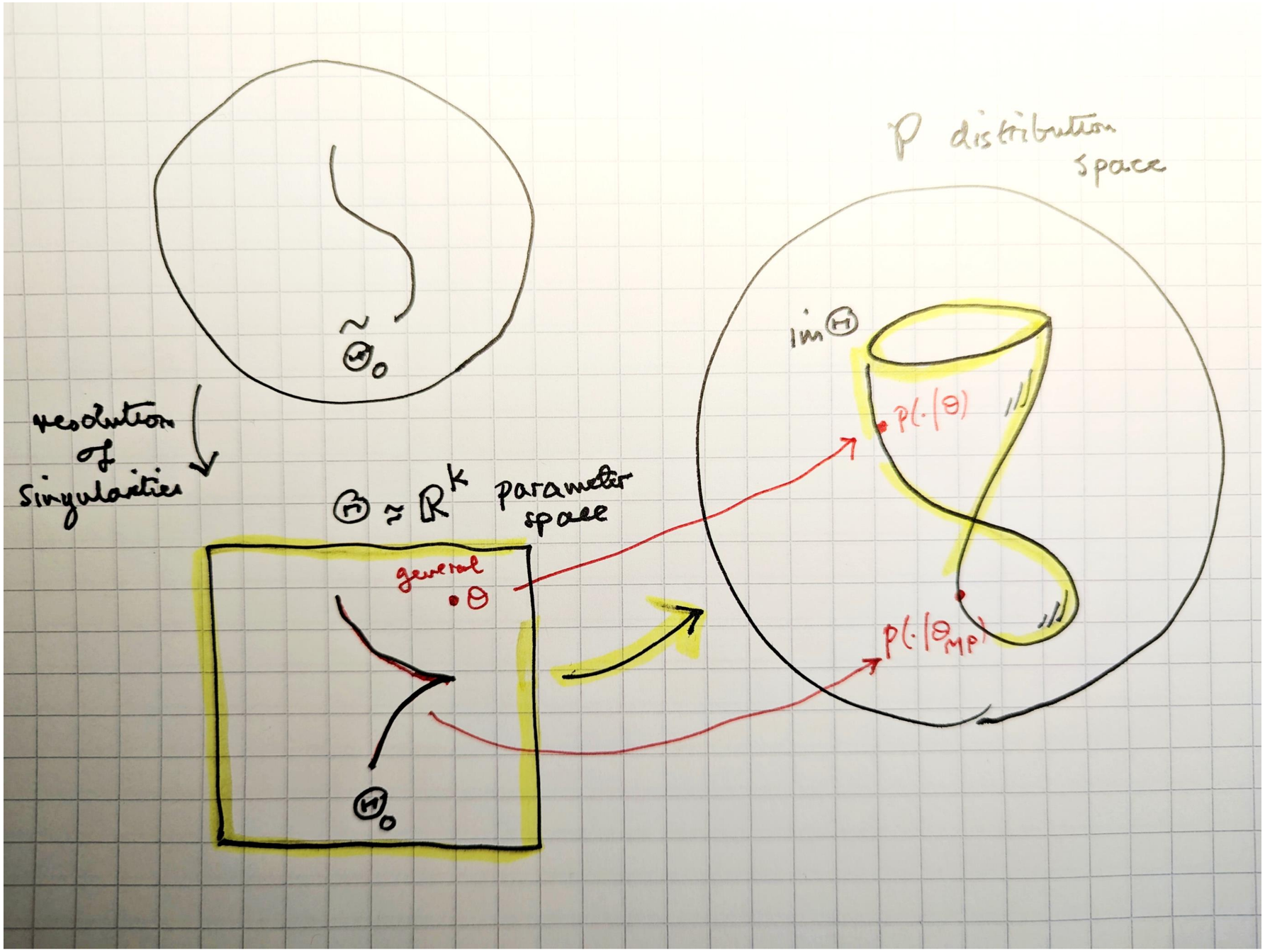
$L(x, y) = (xy)^2$

$(\lambda, \nu) = (\frac{1}{2}, 2)$

Image source: ref [3]

Resolution of singularities (Hironaka)

We can always reduce to the monomial case by blowing up.



Useful references

1. J. Martens, New Insights and Perspectives on the Natural Gradient Method, *Journal of Machine Learning Research 21 (2020) 1-76*
2. D. Murfet et al, Deep Learning is Singular and That's Good, *arXiv:2010.11560* (2020)
3. L. Carroll, Distilling Singular Learning Theory, *AI Alignment Forum blog* (2023)
4. E. Lau et al, The Local Learning Coefficient: A Singularity-Aware Complexity Measure, *arXiv:2308.12108* (2024)