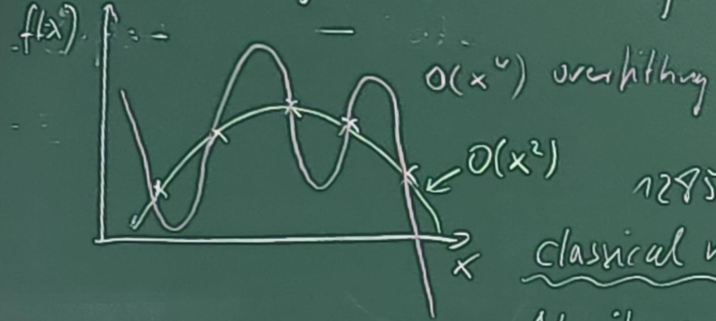
## external validation of SBI (against real data, contd.)

- **model misspecification detection**: is $X^{obs}$ an outlier to the simulation?
  if yes $\rightarrow$ reject $X^{obs}$ and answer "I don't know"

- exploit the feature detection network:
  - add loss $\kappa \, \mathrm{MMD}\left(p(h(x)) \,|\, \mathcal{N}(0, \mathbb{I})\right)$ to pull summary/feature distr. towards standard normal
  - reject $X^{obs}$ if $h(X^{obs})$ is an outlier of $\mathcal{N}(0, \mathbb{I})$

- **model comparison & selection** (between competing theories)
  - measuring training error might not be enough, because overfitting might occur




- need trade-off between model accuracy & complexity
  (intuition: simpler models generalize better)
  1285 – 1347: "Occam's razor": prefer simplest theory

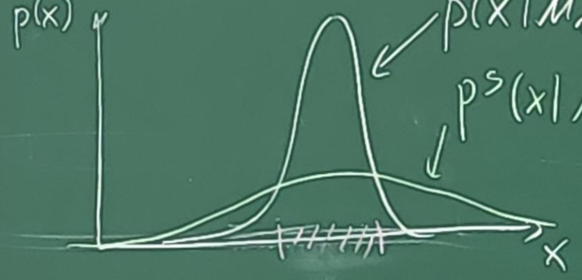
classical model selection criteria:

Akaike criterion $\mathrm{AIC} = 2\,\mathbb{E}[\mathrm{NLL}] + 2\,\mathrm{size}$ ← # model paramet. (e.g. # dimensions of linear model)

Bayesian information crit $BIC = 2\,\mathbb{E}[NLL] + \log(N) \cdot \text{size}$

dataset size

- in Bayesian inverence, model complexity is penalized automatically




$p^s(X|M_1) \leftarrow$ simple model

$p^s(x|M_2) \leftarrow$ complex model

$M_1, \dots, M_L$ competing simulations

in the scope of the simple model |++++|, the marginal density for $M_1$ is much higher than for $M_2$ due to normalization of probs

$\Rightarrow$ if $x^{obs} \in$ |++|, during training of SBI, it typically came from $M_1 \Rightarrow$ automatically prefer $M_1$ during inference as well

noise outsourcing

forward model: $\underbrace{p(M)}_{\text{uniform}(1,L)} \cdot \underbrace{p(Y|M) \cdot p(\eta|M,Y) \cdot \delta(X - \phi_M(Y,\eta))}_{= p(X|Y,M)\ \text{likelihood}}$

simulation M

- - =

marginal density $p(X|M) = \int p(X|Y,M)\, p(Y|M)\, dy$

$\Rightarrow$ model comparison ①: Bayes factor $= \dfrac{p(X = X^{obs} | M_l)}{p(X = X^{obs} | M_{l'})} \begin{cases} > 1 & \text{prefer } M_l \\ < 1 & \text{prefer } M_{l'} \end{cases}$

posterior for model preferences

$$p(M|X) = \frac{p(X|M)\, p(M)}{p(X)} \qquad p(X) = \sum_{l=1}^{L} p(M=l)\, p(X|M=l)$$

$\Rightarrow$ model comparison ②: posterior odds $\dfrac{p(M_l | X = X^{obs})}{p(M_{l'} | X = X^{obs})}$ equal to Bayes factor if $p(M) = \text{uniform}(1, L)$

- practical alg.: train a standard softmax classifier for $p(M|X)$

- comparison thresholds $\dfrac{p(M_l | X^{obs})}{p(M_{l'} | X^{obs})} = r \begin{cases} \frac{1}{3} < r < 3 & \text{: no significant differens} \\ 3 < r < 10 & \text{: substantial evidence for } M_l \\ 10 < r < \cancel{30}\ 100 & \text{strong — // —} \\ \cancel{30}\ 100 < r & \text{overwhelming — // —} \\ \text{likewise } \frac{1}{30}, \frac{1}{10}, \frac{1}{3} & \text{: evidence for } M_{l'} \end{cases}$

## external validation alg.

given competing theories $M_1, \dots, M_L$

[ epidemiology: different # compartments, priors, observation uncertainty etc ]

① create synthetic training data

$$TS_\ell = \left\{ \left( Y_{\ell i} \sim p(Y \mid M=\ell),\ X_n \sim p(X \mid Y_{n\ell}, M=\ell) \right) \right\}_{n=1}^{N_\ell} \qquad TS = \{ TS_1, \dots, TS_L \}$$

② train a separate SBI model for each $TS_\ell$ with MMD, so that $p(h_\ell(X)) = \mathcal{N}(0, \mathbb{I})$

③ perform internal validation for each $\ell$, redesign $SBI_\ell$ until successful

④ train a softmax classifier $p(M \mid X)$ using combined TS

$$\hat{p}(M \mid X) = \arg\min_p \frac{1}{L} \sum_{\ell=1}^{L} \frac{1}{N_\ell} \sum_{n=1}^{N_\ell} - \log p(M=\ell \mid X = X_{n\ell})$$

cross-entropy loss

⑤ external validation given $X^{obs}$:

ⓐ model misspecification detection: $M^{in} = \{ \ell : h_\ell(X^{obs}) \text{ is inlier of } \mathcal{N}(0, \mathbb{I}) \}$

ⓑ compute logits of model classifier $s_\ell$ (penultimate layer, before softmax)

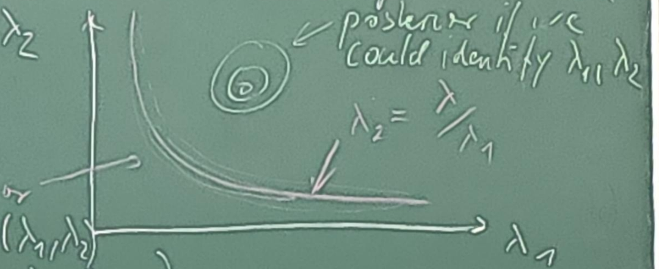
ⓒ define classifier $p(M \mid X^{obs}) = \mathrm{softmax}(s_\ell : \ell \in M^{in})$

ⓓ model comparison by posterior odds

# parameter degeneracy

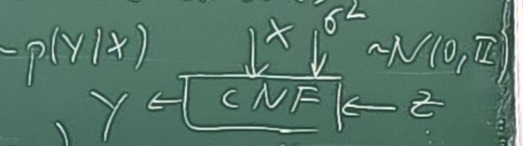
- sometimes, some elements in $Y$ cannot be fully identified from $X$
  $\Rightarrow$ correlations in the posteriors
- example – epidemiologist write SIR equations in terms of natural / conceptional parameters
  - $\lambda_1$ average number of people a healthy person meets per day
  - $\lambda_2$ fraction of dangerous meetings leading to transmission
  - in SIR eq., we always have $\lambda_1 \cdot \lambda_2$ $\Rightarrow$ we cannot distinguish them
  - but, we can infer $\lambda = \lambda_1 \cdot \lambda_2$

$\Rightarrow$ if we still use $\lambda_1$ & $\lambda_2$ $\Rightarrow$ posterior shows the dependency



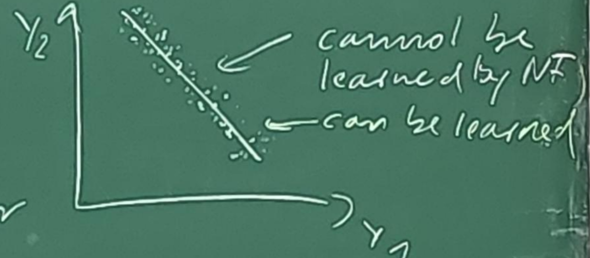

actual posterior
(infinitely many pairs $(\lambda_1, \lambda_2)$ for fixed $\lambda = \lambda_1 \cdot \lambda_2$ )

- here, cause of degeneracy is easy to spot, but generally difficult and hard to distinguish from bad convergence of neural networks
- cNF struggle when $p(Y|X)$ is degenerate, because
  - code distribution $\mathcal{N}(0, \mathbb{I})$ has $D$ dimensions
  - but $p(Y|X)$ has $< D$ dimensions ($\hat{=}$ degeneracy)



theorem: $\underbrace{\text{bijective transforms}}_{\text{e.g. NF}}$ are only possible if dimension does not change

- trick to learn a good approximation: Soft Flow [Kim et al. 2020]
  - idea: add $\underbrace{\text{noise}}_{\text{D-dimension Gaussian } N(0, \sigma^2 \mathbb{I})}$ to $Y_1$ from TS to make it D-dimensional




  - vary $\sigma^2$ during training according to $\sigma^2 \sim \underbrace{p(\sigma^2)}_{\text{some prior}}$
  - tell NF about current value of $\sigma^2$
    (additional condition $p(Y \mid X, \sigma^2)$)

    $\Rightarrow$ network learns to generate data with given amount of noise $\sigma^2$
  - at inference time, make $\sigma^2 \to \sigma^2_{min}$ ($\sigma^2_{min}$ smallest prior value during training)

$\Rightarrow$ line $\lambda_2 = \lambda / \lambda_1$ becomes as narrow as possible

(ideally, do inference $\sigma^2 \to 0$, but practical NF saturate at some finite $\sigma^2_{min}$)

[do not confuse with NoiseNet adds noise to X
Soft Flow — " — Y ]