Learning continuous distributions: Simulations with field theoretic priors

Ilya Nemenman[1,2], William Bialek[1]

[1]NEC Research Institute
[2]Princeton University

August 27, 2003

Bayesian model selection

(MacKay 1992, Balasubramanian 1996)

$$P(x) \xrightarrow{\text{i.i.d.}} X = \{x_1 \cdots x_N\}$$

$$\downarrow$$

$$A \text{ or } B?$$

Model family A	Model family B
$Q_A(x\|\boldsymbol{\alpha})$	$Q_B(x\|\boldsymbol{\beta})$
$\dim\boldsymbol{\alpha} = K_A$	$\dim\boldsymbol{\beta} = K_B$
$\mathcal{P}_A(\boldsymbol{\alpha})$, $Pr(A)$	$\mathcal{P}_B(\boldsymbol{\beta})$, $Pr(B)$

Solution: Maximize

$$\begin{aligned} P(A|X) &= \frac{P(X|A)Pr(A)}{P(X)} \\ &= \frac{Pr(A)\int d\boldsymbol{\alpha}\mathcal{P}(\boldsymbol{\alpha})P(X|\boldsymbol{\alpha})}{P(X|A)Pr(A)+P(X|B)Pr(B)} \end{aligned}$$

Large N expansion is almost always valid

$$\log P(A|X) \to \sum_i \underbrace{\log Q_A(x_i|\boldsymbol{\alpha}_{\text{ML}})}_{\text{goodness of fit}} - \underbrace{\frac{K_A}{2}\log N}_{\substack{\text{generalization error} \\ \text{fluctuations, complexity} \\ \text{weak dependence on priors}}} + \ldots$$

Bayesian inference penalizes for complexity (large K).

Fight between the goodness of fit and the complexity selects an optimal model family.

This is a Bayesian analogue of the MDL principle.

Does this generalize to infinite–dimensional models?

Bayesian learning for $K \to \infty$

(Bialek, Callan, Strong 1996)

Finite	Infinite
$\boldsymbol{\alpha}$	$\phi(x) = -\log \ell_0 Q(x)$
$\mathcal{P}(\boldsymbol{\alpha})$	$\mathcal{P}[Q] \propto \exp\left[-\frac{\ell^{2\eta-1}}{2} \underbrace{\int dx (\partial_x^\eta \phi)^2}_{\text{smoothness penalty}} \right]$
$\{A, K_A\}$	$\{\ell, \eta(?)\}$ – index continuum of families
$Pr(A)$	$Pr(\ell, \eta(?))$

Fix ℓ and η:

$$
\begin{aligned}
P[Q|X] &= \frac{P(X|Q)\mathcal{P}[Q]}{P(X)} \\
\langle Q \rangle &= \frac{\int [dQ]\, \mathcal{P}[Q]\; Q(x) \prod_{i=1}^N Q(x_i)}{\int [dQ]\, P[Q]\; \prod_{i=1}^N Q(x_i)} \\
&= \frac{\langle Q(x) Q(x_1) \cdots Q(x_N) \rangle^0}{\underbrace{\langle Q(x_1) \cdots Q(x_N) \rangle^0}}
\end{aligned}
$$

Correlation function in a QFT defined by $\mathcal{P}[Q]$

Correlation functions:

$$\begin{aligned}
\text{C. F.} &\equiv \int [dQ]\mathcal{P}[Q] \prod_{i=1}^{N} Q(x_i) \\
&= \int [d\phi]\, \frac{1}{\ell_0^N}\, \mathrm{e}^{-S[\phi]}\, \delta\left[\int dx \frac{1}{\ell_0}\mathrm{e}^{-\phi} - 1\right] \\
\underbrace{S[\phi]}_{\text{action}} &= \underbrace{\frac{\ell}{2}\int dx (\partial_x^\eta \phi)^2}_{\text{kinetic term}} + \underbrace{\sum_i \phi(x_i)}_{\text{random potential}}
\end{aligned}$$

Large N approximation for $\eta = 1$

ML (classical, saddle point) solution dominates

converges to $-\log \ell_0 P(x)$

changes on scale $\delta x \sim \sqrt{\ell / N P(x)}$

$$\ell \partial_x^2 \phi_{\mathsf{cl}}(x) + \tfrac{N}{\ell_0} \mathsf{e}^{-\phi_{\mathsf{cl}}(x)} = \textstyle\sum_j \delta(x - x_j)$$

$$\text{C. F.} \approx (1/\ell_0)^N \mathsf{e}^{-S_{\mathsf{eff}}[\phi_{\mathsf{cl}}(x)]}$$

$$S_{\mathsf{eff}}[\phi_{\mathsf{cl}}] = \underbrace{\frac{\ell}{2} \int dx (\partial \phi_{\mathsf{cl}})^2}_{\text{prior, smoothness}} + \underbrace{\sum \phi_{\mathsf{cl}}(x_i)}_{\text{goodness of fit}}$$

$$+ \underbrace{\frac{1}{2} \sqrt{\frac{N}{\ell \ell_0}} \int dx \mathsf{e}^{-\phi_{\mathsf{cl}}(x)/2}}_{\text{fluctuations, complexity, error}}$$

How do we measure performance?

For $x \in [0, L)$ the *universal* learning curve is

$$\Lambda(N) \to \langle D_{\mathsf{KL}}(P || Q_{\mathsf{cl}}) \rangle^0_{\{x_i\}} \sim \sqrt{\frac{L}{\ell N}}$$

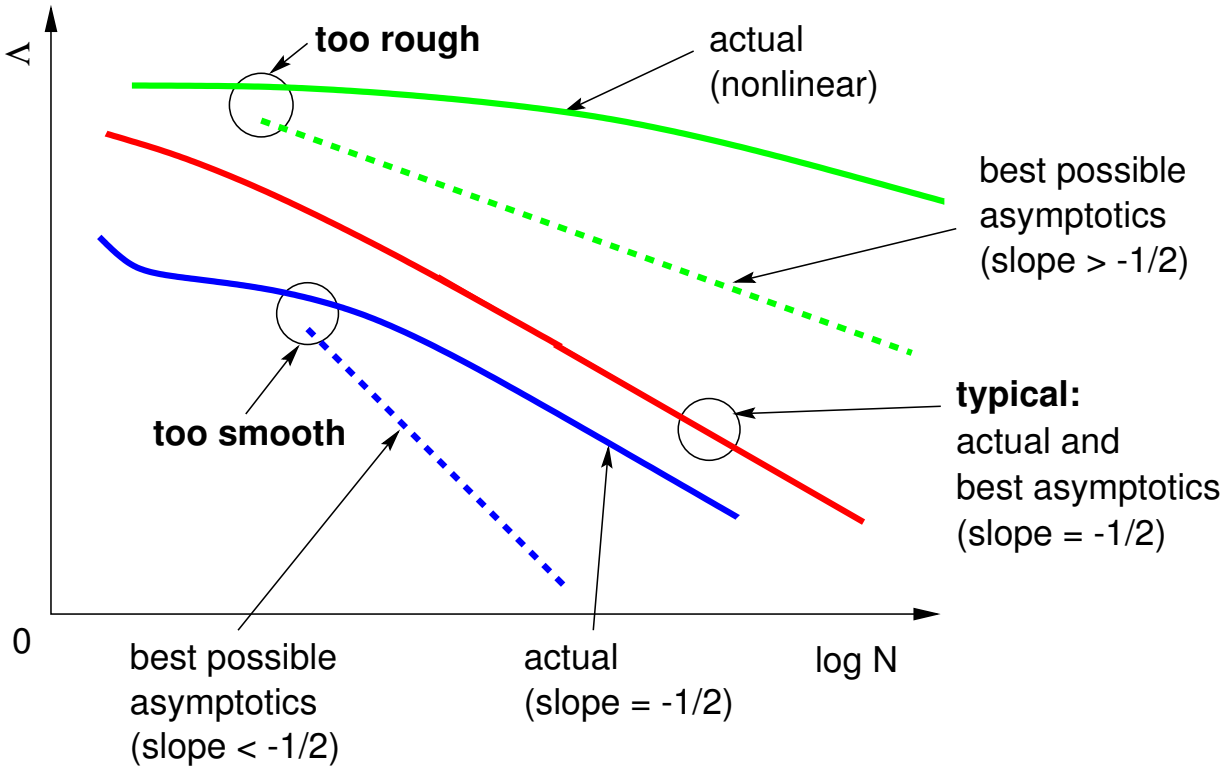
For a different η:

$$\Lambda(N) \sim \left(\frac{L}{\ell} \right)^{1/2\eta} N^{1/2\eta - 1}$$

Learning curves for fixed ℓ, $\eta = 1$

Learner's assumptions $\mathcal{P}_{\ell,\eta=1}[Q]$
Actual target distribution $\mathcal{P}'_{\ell_a,\eta_a}[Q]$

$\eta = \eta_a$, $\ell = \ell_a$	learning typical cases, $\mathcal{P} = \mathcal{P}'$
$\eta = \eta_a$, $\ell \neq \ell_a$	marginal outliers of $\mathcal{P}$
$\eta > \eta_a$	extremely rough outliers
$\eta < \eta_a$	extremely smooth outliers



No overfits!

Smoothness scale selection

Allow a prior over ℓ, but keep $\eta = 1$

$$\text{C. F.} \to \langle \text{C. F.} \rangle_\ell = \int d\ell \; Pr(\ell) \; \mathrm{e}^{-S_{\text{eff}}[\phi_{\text{cl}}(\phi,\ell)]}$$

$$S_{\text{eff}}[\phi_{\text{cl}}] = \underbrace{\text{smoothing} + \text{data}}_{\text{grows with } \ell} + \underbrace{\text{fluctuations}}_{\text{grows with } 1/\ell}$$

Some ℓ^* *always* dominates the C. F. and $\langle Q \rangle$!

What is ℓ^* for η_a and ℓ_a?

If $\eta = \eta_a$, then $\ell^* = \ell_a$. Otherwise:

$0.5 < \eta_a \leq 1.5$	$1.5 < \eta_a$
data $>$ smoothing	smoothing $>$ data
$\ell^* \sim N^{(\eta_a - 1)/\eta_a}$ $\Lambda \sim N^{1/2\eta_a - 1}$	$\ell^* \sim N^{1/3}$ $\Lambda \sim N^{-2/3}$
best possible performance	better, but not best performance

Averaging over ℓ and allowing $\ell^* = \ell^*(N)$ deals with *qualitatively* wrong smoothness $\eta_a \neq 1$!

MDL analogies.

Bayesian smoothness (model) selection works for nonparametric models!

Open questions

- constant factor or constant summand?
- what to do with $\eta_a > 1.5$?
- reparameterization invariance
- information theoretic meaningful priors
- higher dimensions

There is hope that all of this problems are resolvable in a single formulation.