How is sensory information processed?

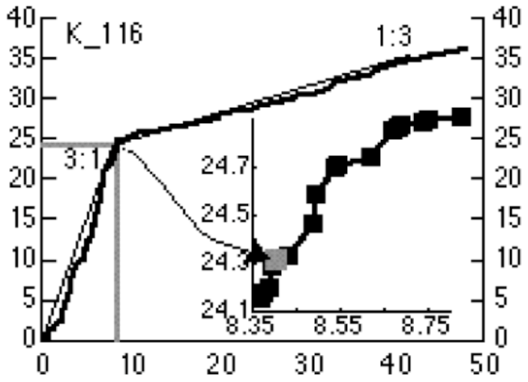
Ilya Nemenman

KITP, UCSB

Advances in statistical learning theory leave us with many possible designs of learning machines. But which of them are implemented by brains, metabolic and genetic networks, and other biological information processors? We analyze how abstract Bayesian learners would perform on different data and discuss possible experiments that can determine which learning–theoretic computation is performed by a particular organism.

Determining a model. . .

. . . used by an animal to represent the world may, in principle, be done by measuring the speed of approach of some behavioral property to its asymptotic value. But. . .



(Gallistel et al., 2001)

A rat learns the distribution of rewards at two different targets and adjusts its visitation rates as the reward rates change. The reward rates and the ratio of the cumulative visit durations is shown for a specific subject as a function of time. Note the speed with which the animal responds, and also note the fluctuations.

- learning is often too fast to observe transients
- learning is often too noisy to estimate asymptotic values [such noise may manifest the stochastic nature of learning in animals (Seung, 2003)]

So what can we do? Remember the *Fluctuation–Dissipation theorem*: transient response (dissipation) is related to steady state properties (r.m.s. fluctuations).

To derive similar relations, we first need to understand. . .

Part 1. . . . Different learning scenarios

Bayesian learning

(MacKay, 1992; Balasubramanian, 1997; Bialek, IN, & Tishby, 2001)

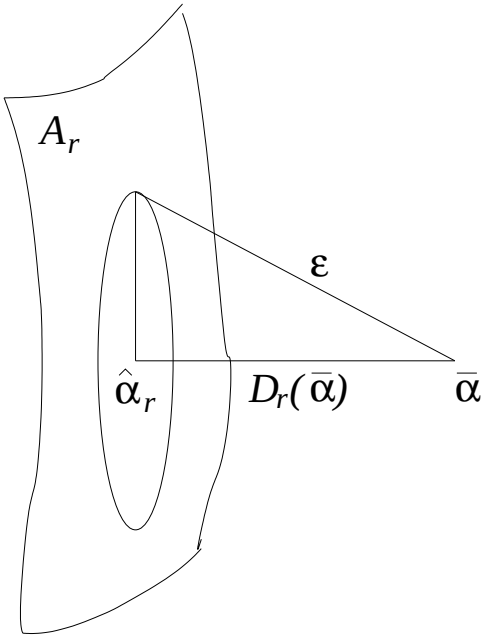
$$P(x) \quad \xrightarrow{\text{i.i.d.}} \quad \{x_i\}_{i=1}^{N} \quad \xrightarrow{???} \quad Q(x|\hat{\boldsymbol{\alpha}})|_{\boldsymbol{\alpha}\in A,\mathcal{P}(\boldsymbol{\alpha})}$$

$$\hat{\boldsymbol{\alpha}} = \arg \min_{\boldsymbol{\alpha}\in A,\mathcal{P}(\boldsymbol{\alpha})} D_{\mathsf{KL}}[P||Q(x|\boldsymbol{\alpha})]$$

$$= \arg \min_{\boldsymbol{\alpha}\in A,\mathcal{P}(\boldsymbol{\alpha})} D_{\mathsf{KL}}(\bar{\boldsymbol{\alpha}}||\hat{\boldsymbol{\alpha}})$$

Properties of learning of $\hat{\boldsymbol{\alpha}}$ depend on *model density*

$$\rho(\epsilon;\bar{\boldsymbol{\alpha}}) = \int d\boldsymbol{\alpha}\, \mathcal{P}(\boldsymbol{\alpha})\, \delta\left[D_{\mathsf{KL}}(\bar{\boldsymbol{\alpha}}||\boldsymbol{\alpha}) - \epsilon\right]$$



- large $\rho(\epsilon \to 0) \Longrightarrow$ higher probability to be correct
- consistency? $\Longleftrightarrow$ $\rho(\epsilon \to 0) \neq 0$
- does the learning speed depend on ρ?

Characterizing learning performance

(Bialek, IN, & Tishby, 2001)

Generalization error, *fluctuation determinant*:

$$\mathcal{D}(\bar{\boldsymbol{\alpha}}; N) = -\log \int d\epsilon\, \rho(\epsilon; \bar{\boldsymbol{\alpha}}) \mathrm{e}^{-N\epsilon}$$

Universal learning curve:

$$\begin{aligned} \Lambda(\bar{\boldsymbol{\alpha}}; N) &= \langle D_{\mathsf{KL}}(\bar{\boldsymbol{\alpha}} || \hat{\boldsymbol{\alpha}}) \rangle_N = \frac{d\mathcal{D}(\bar{\boldsymbol{\alpha}}; N)}{dN} \\ \Lambda(N) &= \int d\bar{\boldsymbol{\alpha}}\, \mathcal{P}(\bar{\boldsymbol{\alpha}}) \Lambda(\bar{\boldsymbol{\alpha}}; N) \end{aligned}$$

Remember for panel (6):

$$\lim_{\hat{\boldsymbol{\alpha}} \to \bar{\boldsymbol{\alpha}}} D_{\mathsf{KL}}(\bar{\boldsymbol{\alpha}} || \hat{\boldsymbol{\alpha}}) = \frac{1}{2} (\bar{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}})^T \mathcal{F} (\bar{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}})$$

($\mathcal{F}$ — Fisher information matrix)

Learning a model $\mathcal{P}(\alpha)$ with wrong (atypical) assumptions $\mathcal{R}(\alpha)$:

$$\Lambda(N; \mathcal{P}, \mathcal{R}) = \int d\bar{\boldsymbol{\alpha}}\, \mathcal{P}(\bar{\boldsymbol{\alpha}}) \Lambda_{\mathcal{R}}(\bar{\boldsymbol{\alpha}}; N)$$

Examples

Finite states, $\mathcal{P} = \sum_{i=1}^{M} \mathcal{P}_i \, \delta(a_i)$, $Q(x|a_i)$

$$\Lambda(a_i; N) \approx c_i \exp[-N d_i]$$

$K < \infty$ continuous parameters, $\mathcal{P}(\boldsymbol{\alpha})$, $Q(x|\boldsymbol{\alpha})$

$$\Lambda(\bar{\boldsymbol{\alpha}}; N) \approx \frac{K}{2N}$$

Complete models, able to represent and learn *any* probability distribution:

*Nested (*n*)*		*QFT (*q*)*
$\Lambda(\mathsf{q},\mathsf{n}) \propto \left(\frac{\log N}{N\ell}\right)^{1-1/2\eta}$	target typical in $\mathcal{P}_{\mathsf{QFT}}$	$\Lambda(\mathsf{q},\mathsf{q}) \propto \left(\frac{1}{N\ell}\right)^{1-1/2\eta}$
$\Lambda(\mathsf{n},\mathsf{n}) \propto \frac{K^*}{N}$	target typical in $\mathcal{P}_{\mathsf{nest}}$	$\Lambda(\mathsf{n},\mathsf{q}) \propto \left(\frac{1}{N\ell}\right)^{1-1/2\eta}$

Remark: nested model is never much worse than the QFT one, and sometimes it is much better if the environment has well defined "important" directions, to which the animal is tuned, not necessarily precisely (this is characteristic of natural signals; e.g., the most important parameter in statistics of visual scenes is the light intensity).

Part 2. Determining the model

For a fixed target asymptotically: $\frac{\partial \Lambda}{\partial N} = -\zeta_N \Lambda^\nu$

ν	model
1	finite states
2	finite parameter or nested
> 2	QFT

If target changes and

- the animal is fast in noticing the change (Fairhall et al., 2001; Gallistel et al., 2001), and changes are small then the equation still holds.
- changes happen mostly along direction α_1 (e.g., only reward rates change, but not the functional forms of distributions), then $\Lambda \propto (\widehat{\alpha}_1 - \bar{\alpha}_1)^2 \equiv \Delta^2$ (easy to generalize for more directions).
- animal does fixed number of observations per unit time, then $dN \propto dt$.

Then for $d\bar{\alpha}_1/dt \equiv v_{\bar{\alpha}}$,

$$\frac{d\Delta}{dt} = -\zeta\, \mathrm{sign}(\Delta)\, |\Delta|^{2\nu-1} - v_{\bar{\alpha}}$$

Examples...

...of various *slow* changes of one parameter for small Λ. If animal does not extrapolate:

$v_{\bar{\alpha}} = \mathcal{A} = \text{const}$:

$$\lim_{t\to\infty} \Delta(t) \equiv\to \Delta_\infty = -\left(\frac{\mathcal{A}}{\zeta}\right)^{1/(2\nu-1)}$$

$v_{\bar{\alpha}} = \mathcal{A}\omega \cos\omega t$:

$$\lim_{t\to\infty} \langle|\Delta|^{4\nu-2}\rangle = \frac{(\mathcal{A}\omega)^2}{2\zeta^2}$$

Animal may anticipate changes in $\bar{\alpha}_1$. To remove this possibility, take $\langle v_{\bar{\alpha}}(t) v_{\bar{\alpha}}(t')\rangle = \Omega\,\delta(t-t')$:

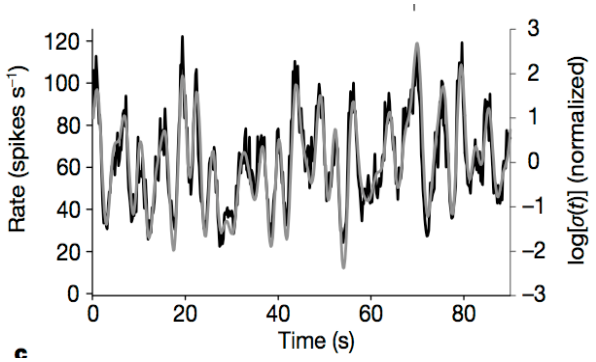
$$\lim_{t\to\infty} \Delta_{\text{rms}} = \left\{\nu^{1/\nu}\frac{\Gamma\left(\frac{3}{2\nu}\right)}{\Gamma\left(\frac{1}{2\nu}\right)}\right\}^{1/2}\left(\frac{\Omega}{\zeta}\right)^{1/(2\nu)}.$$

Remark: This is an analog of FDT.

Remark: If learning in animals is intrinsically stochastic, similar fluctuations will be present even for stationary targets. However, since the distribution of intrinsic fluctuations is unknown, they cannot be used to distinguish learning models.

To distinguish the model

1. Select a class of targets learnable by the animal and a response that signifies the animal's current guess.



A fly is subjected to the angular velocity signal with a variable standard deviation. Instantaneous firing rate is a good measure of the instantaneous standard deviation.

(Fairhall et al., 2001)

2. Estimate intrinsic fluctuations and change the target within the class (deterministically and/or stochastically) so that $\Delta \gg \Delta_{\text{intrinsic}}$.

3. Vary change parameters and estimate ν from the relation between them and statistics of Δ.

4. If $\nu > 2$, animal uses QFT-type models.

5. If $\nu = 2$, add "dimensions" to the target class. If at some point the animal fails to learn — animal uses finite parameter models; otherwise nested models.

6. If $\nu = 1$, finite state model is used.

7. Exotic, in-between, values of ν are also possible.

Bibliography

1. V Balasubramanian. Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neur. Comp.*, 9:349–368, 1997.
2. W Bialek, C Callan, and S Strong. Field theories for learning probability distributions. *Phys. Rev. Lett.*, 77:4693–4697, 1996.
3. W Bialek, I Nemenman, and N Tishby. Predictability, complexity, and learning. *Neur. Comp.*, 13:2409–2463, 2001.
4. A Fairhall, G Lewen, W Bialek, and R de Ruyter van Steveninck. Efficiency and ambiguity in an adaptive neural code. *Nature*, 412:787–792, 2001.
5. CR Gallistel, T Mark, AP King, and P Latham. The rat approximates an ideal detector of changes in rates of reward: Implications for the law of effect. *J. Exper. Psych.: Animal Behav. Proc.*, 27:354–372, 2001.
6. D MacKay. Bayesian interpolation. *Neur. Comp.*, 4:415–448, 1992.
7. I Nemenman and W Bialek. Occam factors and model-independent bayesian learning of continuous distributions. *Phys. Rev. E*, 65:026137, 2002.
8. HS Seung. Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron*, 40:1063–1073, 2003.

Part 3. Appendix

Finite parameter models

$$\rho(\epsilon;\bar{\boldsymbol{\alpha}}) \approx \mathcal{P}(\bar{\boldsymbol{\alpha}}|r)\frac{2\pi^{K/2}}{\Gamma(K/2)}\frac{\epsilon^{(K-2)/2}}{\sqrt{\det \mathcal{F}_K}}$$

Nested models

$\{A_r\}_{r=1}^{R}$, $\mathcal{P}_{\text{nest}}(r)$, $K(r)$, $\mathcal{P}(\boldsymbol{\alpha}|r)$, $Q(x|\boldsymbol{\alpha})$:

$$\begin{aligned}
\mathcal{P}_{\text{nest}}(\mathbf{r}) &\propto r^{-\gamma}\,,\ \gamma > 1,\ \ R \to \infty \\
\mathcal{P}(\boldsymbol{\alpha}|r) &= \prod_{\mu=1}^{R} \mathcal{P}(\alpha_\mu|r) \\
\mathcal{P}(\alpha_\mu|r) &= \begin{cases} p(\alpha_\mu)\,, & \mu \le K(r) \\ \delta(\alpha_\mu)\,, & \mu > K(r) \end{cases} \\
p(\alpha_\mu) &= \mathcal{N}(0,\sigma_\mu^2) \\
\sigma_\mu &= c r_\mu^{-\beta},\ \beta \ge 0
\end{aligned}$$

$$\langle \delta\alpha_\mu^2 \rangle = \sigma_\mu^2 \sum_{r \ge r_\mu} \mathcal{P}(r) \sim \mu^{-\beta} \sum_{r=\mu}^{\infty} r^{-\gamma} \sim \mu^{-\beta-\gamma+1}$$

$$\rho(\epsilon;\bar{\boldsymbol{\alpha}}) = \sum_{r:\, D_r(\bar{\boldsymbol{\alpha}})\le\epsilon} \mathcal{P}(r)\mathcal{P}(\widehat{\boldsymbol{\alpha}}_r|r)\frac{2\pi^{K(r)/2}}{\Gamma[K(r)/2]} \frac{[\epsilon^2 - D_r^2(\bar{\boldsymbol{\alpha}})]^{[K(r)-2]/4}}{\sqrt{\det \mathcal{F}_{K(r)}}}$$

Bayesian model selection finds a posteriori dominant r^*, models with $r > r^*$ exponentially inhibited.

Example: $x \in [0,1)$. $A_r \cup \phi(x|\boldsymbol{\alpha}) \equiv -\log Q(x|\boldsymbol{\alpha})$

$$\phi = \alpha_0 + \sum_{\mu=1}^{r}\left(\alpha_\mu^+ \cos 2\pi\mu x + \alpha_\mu^- \sin 2\pi\mu x\right)$$

$$\alpha_0 = \log \int dx\, \mathrm{e}^{-\sum_{\mu=1}^{r}\left(\alpha_\mu^+ \cos 2\pi\mu x + \alpha_\mu^- \sin 2\pi\mu x\right)}$$

Always consistent for $\beta \ge 0$, $\gamma > 1$; $r^* \lesssim N/\log N$, and Fourier$[Q_r^*] \approx$ Fourier$[\frac{1}{N}\sum \delta(x - x_i)]$.

QFT models

(Bialek et al, 1996; Nemenman and Bialek, 2002)

$$\mathcal{P}_{\mathrm{QFT}}[\phi(x)] = \frac{1}{\mathcal{Z}} \exp\left[-\frac{\ell^{2\eta-1}}{2}\int dx \left(\frac{\partial^\eta \phi}{\partial x^\eta}\right)^2\right] \delta\left[\frac{1}{l_0}\int dx\, \mathrm{e}^{-\phi(x)} - 1\right]$$

$$\rho(\epsilon;\bar{\phi}) \approx A[\bar{\phi}]\,\epsilon^{\xi} \exp\left[-\frac{B[\bar{\phi}]}{\ell\epsilon^{1/(2\eta-1)}}\right]$$

$$\langle(\delta\alpha_\mu^\pm)^2\rangle = \frac{2}{\ell^{2\eta-1}}\frac{1}{(2\pi\mu)^{2\eta}}, \qquad \mu > 0$$

$\eta \to 1/2$ – the most complex learning problem.

Bayesian model selection: "nesting" of QFT models, integrate over ℓ. Produces *correct* dominant ℓ even for some incorrect η.

Focus on *correct* η, ℓ.