The Abdus Salam International Centre for Theoretical Physics
50th Anniversary 1964–2014

2584-18

Spring College on the Physics of Complex Systems

26 May - 20 June, 2014

Theoretical Neuroscience:
Supervised Learning and Information Theory

Sara Solla
Northwestern University
USA

Strada Costiera, 11 - 34151 - Trieste - Italy • Tel. +39 0402240111 • Fax. +39 040224163 • sci_info@ictp.it • www.ictp.it

Theoretical Neuroscience: Supervised Learning and Information Theory

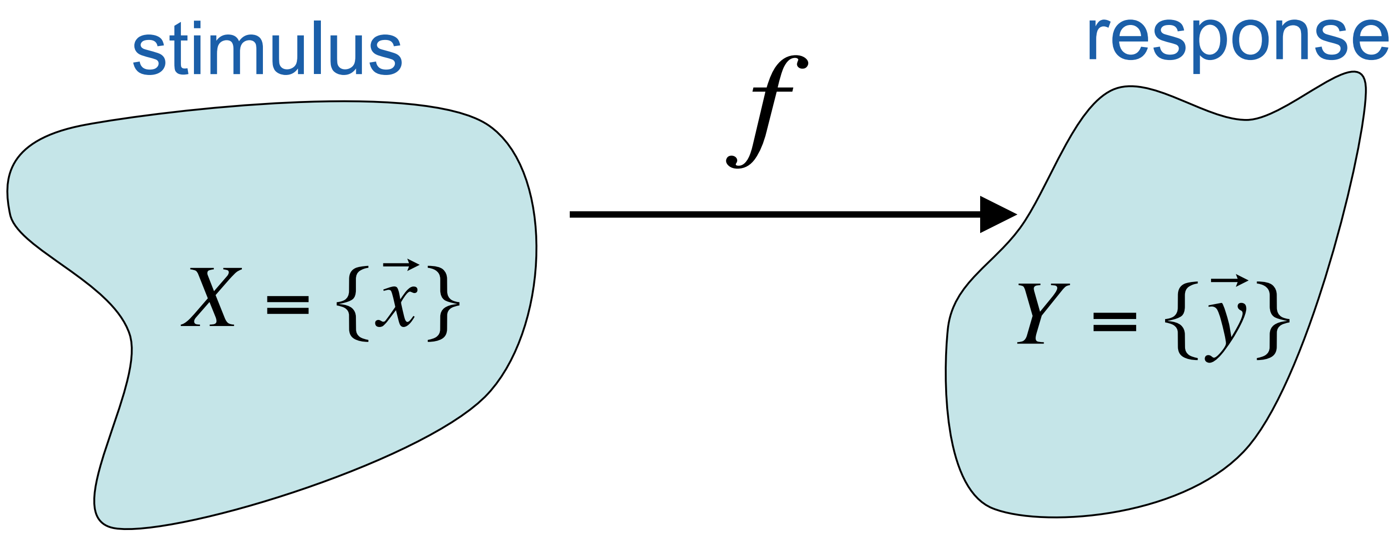
ICTP, Trieste, June 2014

Sara A. Solla
Department of Physiology
Department of Physics and Astronomy
Northwestern University

What is Learning?

Learning is an entropy reduction process!

Input-Output Maps



$$\vec{x} = \{x_1, x_2, \ldots, x_n\} \rightarrow \vec{y} = \{y_1, y_2, \ldots, y_s\}$$

$$\vec{y} = f(\vec{x})$$

Input-Output Modules

$$\vec{y} = f_{\vec{W}}(\vec{x})$$

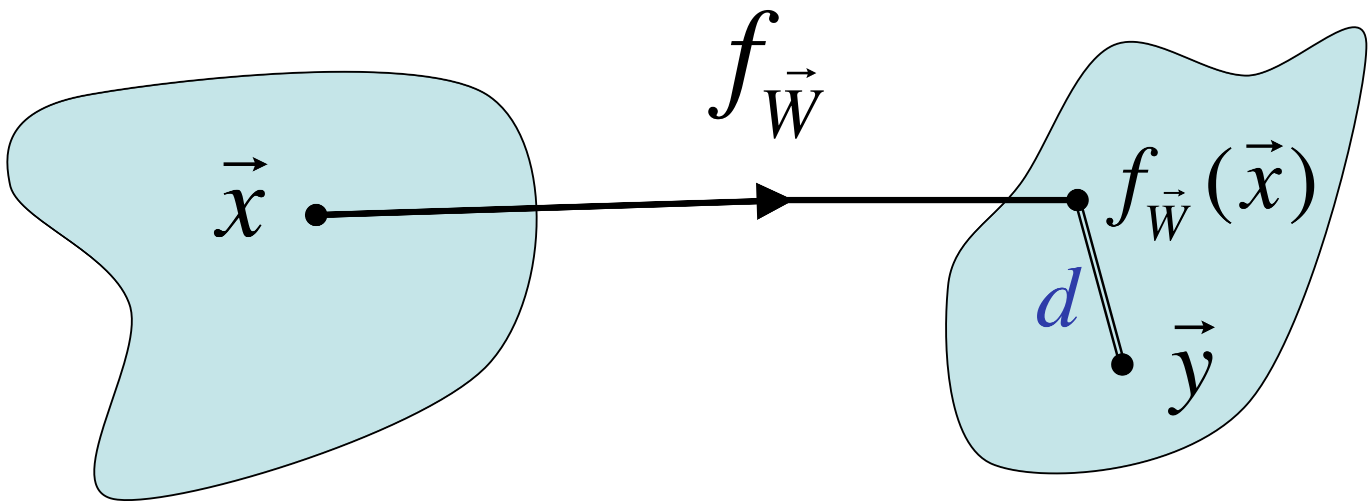
$$\vec{x} \longrightarrow \boxed{\vec{W}} \longrightarrow \vec{y}$$

What specifies the value of the parameters $\vec{W}$?

Data: $$\vec{\xi}^{\mu} = (\vec{x}^{\mu},\ \vec{y}^{\mu}) \qquad 1 \le \mu \le m$$

Examples of the desired map: input-output pairs

Learning from Examples



Given an example $(\vec{x}, \vec{y})$ of the desired map, the error made by a specific module $\vec{W}$ on this example is:

$$E(\vec{W} \mid \vec{x}, \vec{y}) = d\left(\vec{y}, f_{\vec{W}}(\vec{x})\right)$$

Learning Error

Given a training set of size m:

$$\vec{\xi}^{\mu} = (\vec{x}^{\mu},\ \vec{y}^{\mu}),\quad 1 \le \mu \le m,$$

construct a cost function that measures the average error over the training set, the learning error:

$$E_L(\vec{W}) = (1/m)\sum_{\mu=1}^{m} E(\vec{W} \mid \vec{x}^{\mu}, \vec{y}^{\mu})$$

Most learning algorithms are based finding the $\vec{W}^*$ that minimize this learning error, i.e., back-propagation.

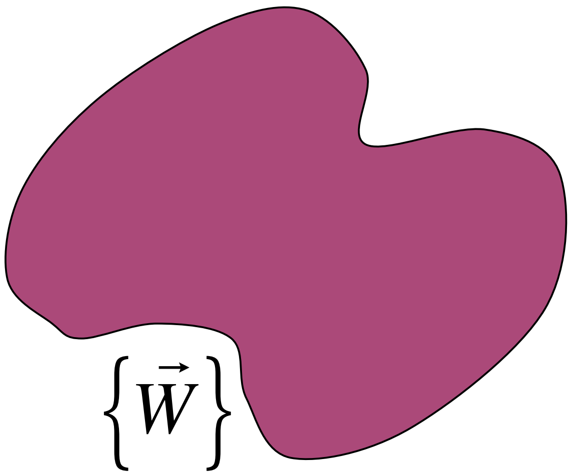
Rumelhart, Hinton, Williams, 1986

Configuration Space

For each example $\vec{\xi}^{\mu} = (\vec{x}^{\mu},\ \vec{y}^{\mu})$ in the training set, define a masking function:

$$\Theta(\vec{W}, \vec{\xi}^{\mu}) = 1 \quad \text{if} \quad f_{\vec{W}}(\vec{x}^{\mu}) = \vec{y}^{\mu}$$

$$\Theta(\vec{W}, \vec{\xi}^{\mu}) = 0 \quad \text{if} \quad f_{\vec{W}}(\vec{x}^{\mu}) \neq \vec{y}^{\mu}$$

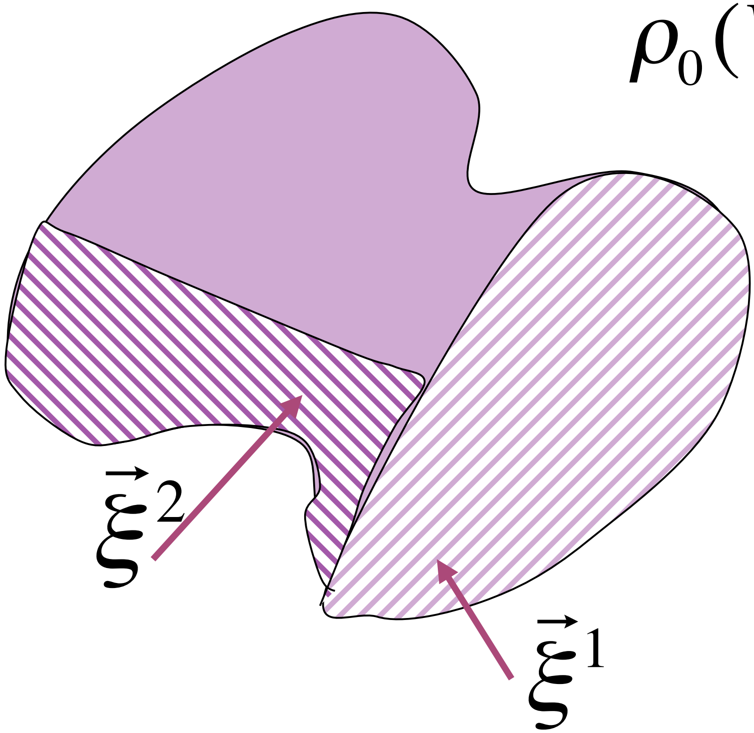


Prior $\rho_0(\vec{W})$

Normalization:

$$\int \rho_0(\vec{W})\, d\vec{W} = 1$$

Error-Free Learning



$$\rho_0(\vec{W}) \Rightarrow$$

$$\rho_0(\vec{W})\Theta(\vec{W},\vec{\xi}^1) \Rightarrow$$

$$\rho_0(\vec{W})\Theta(\vec{W},\vec{\xi}^1)\Theta(\vec{W},\vec{\xi}^2)$$

Masking: $Z_m = \int d\vec{W}\,\rho_0(\vec{W})\prod_{\mu=1}^{m}\Theta(\vec{W},\vec{\xi}^\mu)$

Contraction: $Z_m \le Z_{m-1} \le \ldots \le Z_1 \le Z_0 = 1$

Learning from Noisy Data

Consider the error on the μth example:

$$E(\vec{W}\,|\,\vec{\xi}^{\mu}) = d\left(\vec{y}^{\mu}, f_{\vec{W}}(\vec{x}^{\mu})\right)$$

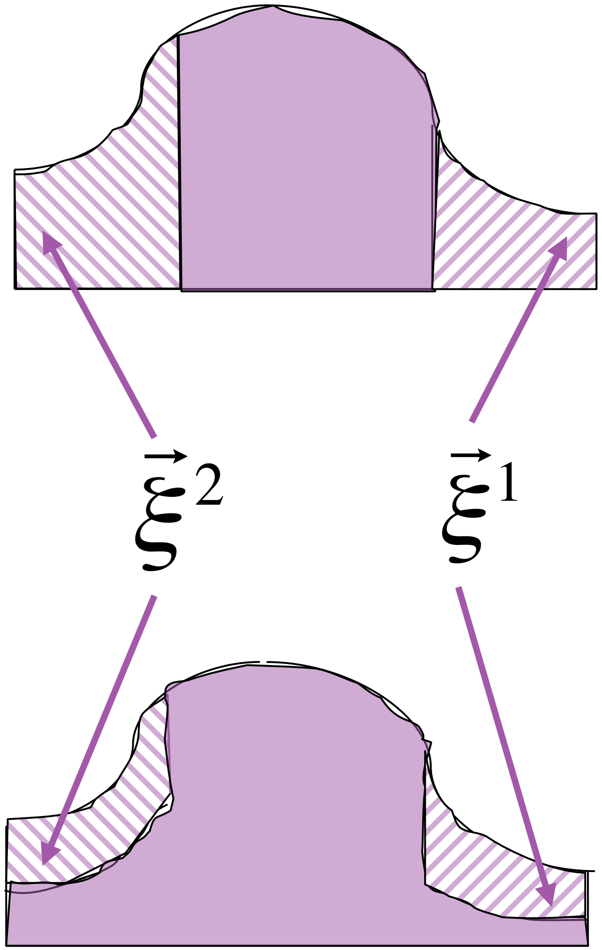
If $f_{\vec{W}}(\vec{x}^{\mu}) = \vec{y}^{\mu}$, $E(W\,|\,\vec{\xi}^{\mu}) = 0 \Rightarrow \Theta(\vec{W}, \vec{\xi}^{\mu}) = 1$

If $f_{\vec{W}}(\vec{x}^{\mu}) \neq \vec{y}^{\mu}$, instead of setting $\Theta(\vec{W}, \vec{\xi}^{\mu}) = 0$

introduce a survival probability:

$$\Theta(\vec{W}, \vec{\xi}^{\mu}) \rightarrow \exp\left(-\beta E(\vec{W}\,|\,\vec{\xi}^{\mu})\right)$$

Hard vs Soft Masking



Hard masking: configurations incompatible with the data are eliminated.

Soft masking: configurations are attenuated by a factor exponentially controlled by the error made on the data.

Learning with Uncertainty

$$\rho_0(\vec{W}) \quad \Rightarrow \quad \rho_0(\vec{W})\exp\left(-\beta E(\vec{W}\,\middle|\,\vec{\xi}^1)\right) \quad \Rightarrow$$

$$\rho_0(\vec{W})\exp\left(-\beta E(\vec{W}\,\middle|\,\vec{\xi}^1)\right)\exp\left(-\beta E(\vec{W}\,\middle|\,\vec{\xi}^2)\right)$$

$$Z_m = \int d\vec{W}\,\rho_0(\vec{W})\prod_{\mu=1}^{m}\exp\left(-\beta E(\vec{W}\,\middle|\,\vec{\xi}^\mu)\right)$$

$$Z_m = \int d\vec{W}\,\rho_0(\vec{W})\exp\left(-m\beta E_L(\vec{W})\right)$$

with learning error: $E_L(\vec{W}) = (1/m)\sum_{\mu=1}^{m} E(\vec{W}\,|\,\vec{\xi}^\mu)$

Gibbs Distribution

The ensemble of all possible modules is described by the prior density $\rho_0(\vec{W})$. The ensemble of trained modules is described by the posterior density $\rho_m(\vec{W})$:

$$\rho_m(\vec{W}) = \frac{1}{Z_m}\rho_0(\vec{W})\exp\left(-\beta m E_L(\vec{W})\right)$$

Note that $\int d\vec{W}\,\rho_m(\vec{W}) = 1$, and that the partition function Z_m provides the normalization constant. Note also that this distribution arises from without invoking specific algorithms for exploring the configuration space $\{\vec{W}\}$.

Natural Statistics

Training data $\vec{\xi} = (\vec{x},\ \vec{y})$ is drawn from a distribution $\tilde{P}(\vec{\xi}) = \tilde{P}(\vec{x},\vec{y}) = \tilde{P}(\vec{y} \mid \vec{x})\,\tilde{P}(\vec{x})$

$\tilde{P}(\vec{x})$ describes the region of interest input space

$\tilde{P}(\vec{y} \mid \vec{x})$ describes the functional dependence

Thermodynamics of Learning

The partition function

$$Z_m = \int d\vec{W}\, \rho_0(\vec{W}) \exp\left(-\beta \sum_{\mu=1}^{m} E(\vec{W} \middle| \vec{\xi}^{\mu}) \right)$$

depends on the specific set of data points $D = \left\{ \vec{\xi}^{\mu} \right\}$ drawn from $\tilde{P}(\vec{\xi})$. The associated free energy

$$F = -(1/\beta) \left\langle \left\langle \ln Z_m \right\rangle \right\rangle_D$$

follows from averaging over all possible data sets of size m. The average learning error follows from the usual thermodynamic derivative:

$$E_L = -\frac{1}{m} \frac{\partial}{\partial \beta} \left\langle \left\langle \ln Z_m \right\rangle \right\rangle_D$$

Entropy of Learning

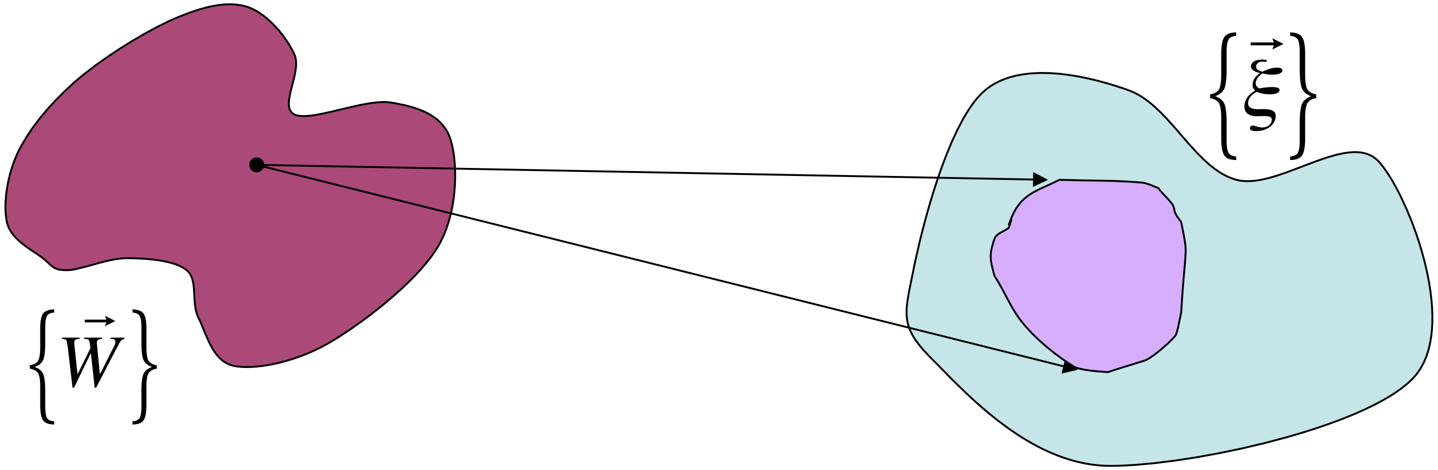
The entropy follows from $F = m\,E_L - (1/\beta)\,S$

For the learning process, this results in:

$$S = -\int d\vec{W}\,\rho_m(\vec{W})\ln\left[\frac{\rho_m(\vec{W})}{\rho_0(\vec{W})}\right] = -\,D_{KL}\left[\rho_m \middle| \rho_0\right]$$

The entropy of learning is minus the Kullback-Leibler distance between the posterior $\rho_m(\vec{W})$ and the prior $\rho_0(\vec{W})$, and it measures the amount of information gained. The distance between posterior and prior increases monotonically with the size m of the training set.

Maximum Likelihood Learning



$P(\vec{\xi}\,|\,\vec{W})$: distribution induced through hypothesis $\vec{W}$

$\tilde{P}(\vec{\xi})$: true distribution

Likelihood of the data:

$$\mathcal{L}(\vec{W}) = P(D\,|\,\vec{W}) = P(\vec{\xi}^{\,1}, \vec{\xi}^{\,2}, \ldots, \vec{\xi}^{\,m}\,|\,\vec{W}) = \prod_{\mu=1}^{m} P(\vec{\xi}^{\,\mu}\,|\,\vec{W})$$

BUT: what is the form of $P(\vec{\xi}\,|\,\vec{W})$?

Learning Coherence

Two approaches to learning:

•Minimize the error on the data:

$$E_L(\vec{W}) = \sum_{\mu=1}^{m} E(\vec{W} \mid \vec{\xi}^{\mu})$$

•Maximize the likelihood of the data:

$$\mathcal{L}(\vec{W}) = \prod_{\mu=1}^{m} P(\vec{\xi}^{\mu} \mid \vec{W})$$

Require that these two approaches be coherent!

$$P(\vec{\xi} \mid \vec{W}) = \frac{1}{z(\beta)} \exp\left(-\beta E(\vec{W} \mid \vec{\xi})\right)$$

(Appendix)

Bayesian Learning

We now compute the likelihood of the data: $P(D|\vec{W}) =$

$$\prod_{\mu=1}^{m} P(\vec{\xi}^{\mu}|\vec{W}) = \frac{1}{z(\beta)^m}\exp\left(-\beta\sum_{\mu=1}^{m}E(\vec{\xi}^{\mu}|\vec{W})\right) = \frac{1}{z(\beta)^m}\exp\left(-\beta m E_L(\vec{W})\right)$$

Bayesian inversion: $$P(\vec{W}|D) = \frac{P(D|\vec{W}) * P(\vec{W})}{P(D)}$$

Gibbs distribution:

$$\rho_m(\vec{W}) = \frac{1}{Z_m}\rho_0(\vec{W})\exp\left(-\beta m E_L(\vec{W})\right)$$

Bayes ⟺ Gibbs

Prior: $$P(\vec{W}) \Leftrightarrow \rho_0(\vec{W})$$

Posterior: $$P(\vec{W}|D) \Leftrightarrow \rho_m(\vec{W})$$

Likelihood: $$P(D|\vec{W}) \Leftrightarrow \frac{1}{z(\beta)^m} \exp\left(-\beta m E_L(\vec{W})\right)$$

Evidence: $$P(D) \Leftrightarrow \frac{1}{z(\beta)^m} Z_m$$

where $$P(D) = \int d\vec{W}\, P(D|\vec{W}) P(\vec{W})$$

The normalization constant $z(\beta)$ plays a role in the evaluation of prediction errors (has the brain acquired a good model of the world?)

Generalization Ability

Consider a new point $\vec{\xi}$ not part of the training data $D = \{\vec{\xi}^1, \vec{\xi}^2, \ldots, \vec{\xi}^m\}$. What is the likelihood of this test point?

$$P(\vec{\xi}\,|\,D) = \int d\vec{W}\; P(\vec{\xi}\,|\,\vec{W})P(\vec{W}\,|\,D)$$

with: $$P(\vec{\xi}\,|\,\vec{W}) = \frac{1}{z(\beta)}\exp\left(-\beta\, E(\vec{W}\,|\,\vec{\xi})\right)$$

and: $$P(\vec{W}\,|\,D) = \rho_m(\vec{W}) = \frac{1}{Z_m}\rho_0(\vec{W})\exp\left(-\beta\sum_{\mu=1}^{m} E(\vec{W}\,|\,\vec{\xi}^{\mu})\right)$$

Generalization Ability

$$P(\vec{\xi}|D) = \int d\vec{W}\, P(\vec{\xi}|\vec{W})P(\vec{W}|D) =$$

$$= \frac{1}{z(\beta)Z_m}\int d\vec{W}\,\rho_0(\vec{W})\exp\left(-\beta\sum_{\mu=1}^{m+1}E(\vec{W}|\vec{\xi}^{\mu})\right)$$

Where $\vec{\xi}^{m+1} = \vec{\xi}$: the test point appears as if it had been added to the training set. Thus:

$$P(\vec{\xi}|D) = \frac{Z_{m+1}}{z(\beta)Z_m}$$

Generalization Error

The generalization error is defined through the ln of the likelihood of the test point $\vec{\xi}$:

$$P(\vec{\xi}\,|D) = \frac{Z_{m+1}}{z(\beta) Z_m} \quad \Rightarrow \quad E_G = -\frac{1}{\beta}\left[\ln \frac{Z_{m+1}}{Z_m} - \ln z(\beta)\right]$$

For large m, the difference between $(\ln \mathrm{Z}_{m+1})$ and $(\ln \mathrm{Z}_{\mathrm{m}})$ can be approximated by a derivative with respect to m. Then $(\ln \mathrm{Z})$ is averaged over all possible data sets of size m, to obtain:

$$E_G = -\frac{1}{\beta}\frac{\partial}{\partial m}\left\langle\left\langle \ln Z_m \right\rangle\right\rangle_D + \frac{1}{\beta}\ln z(\beta)$$

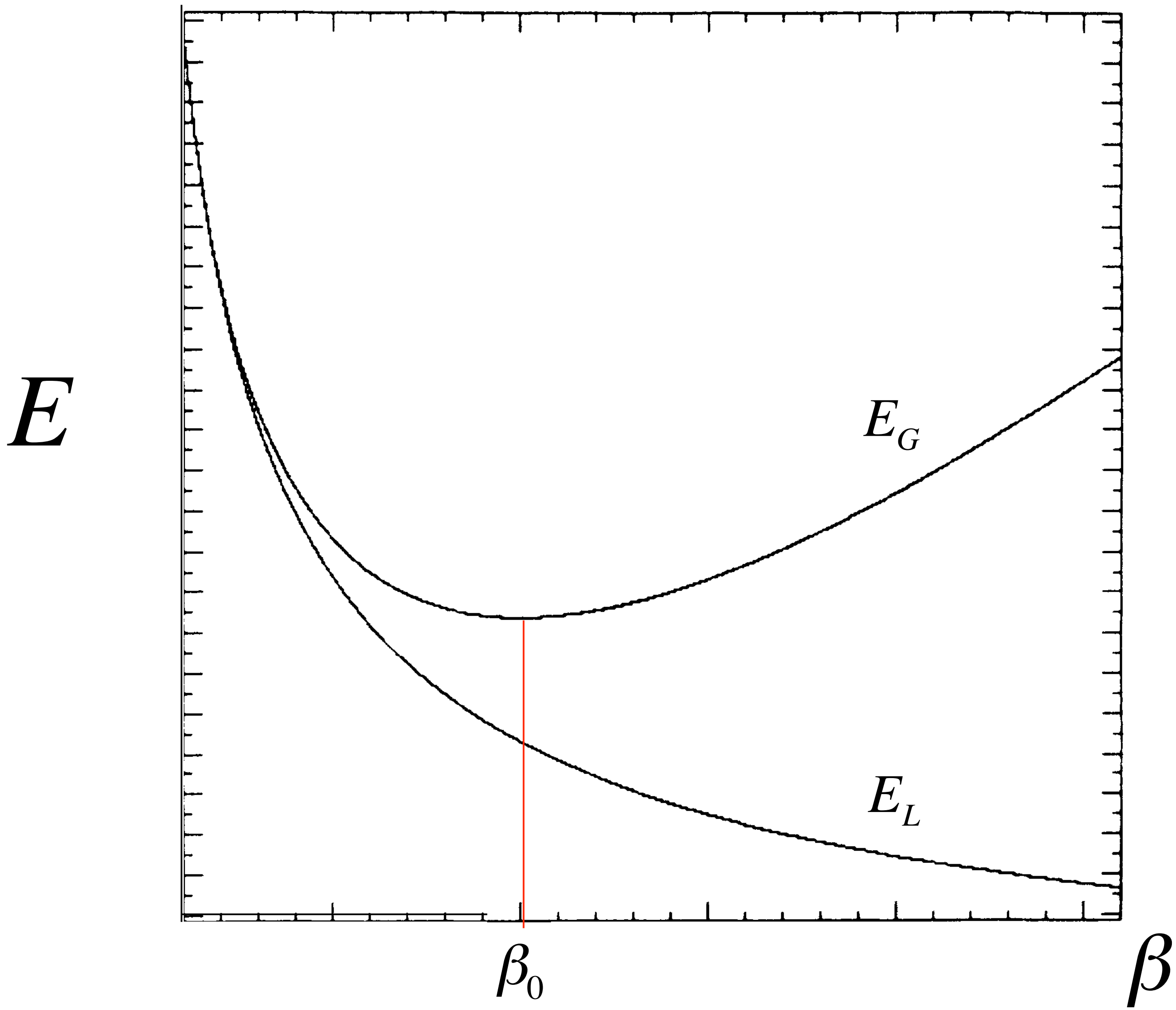
Learning vs Generalization

Two thermodynamic derivatives:

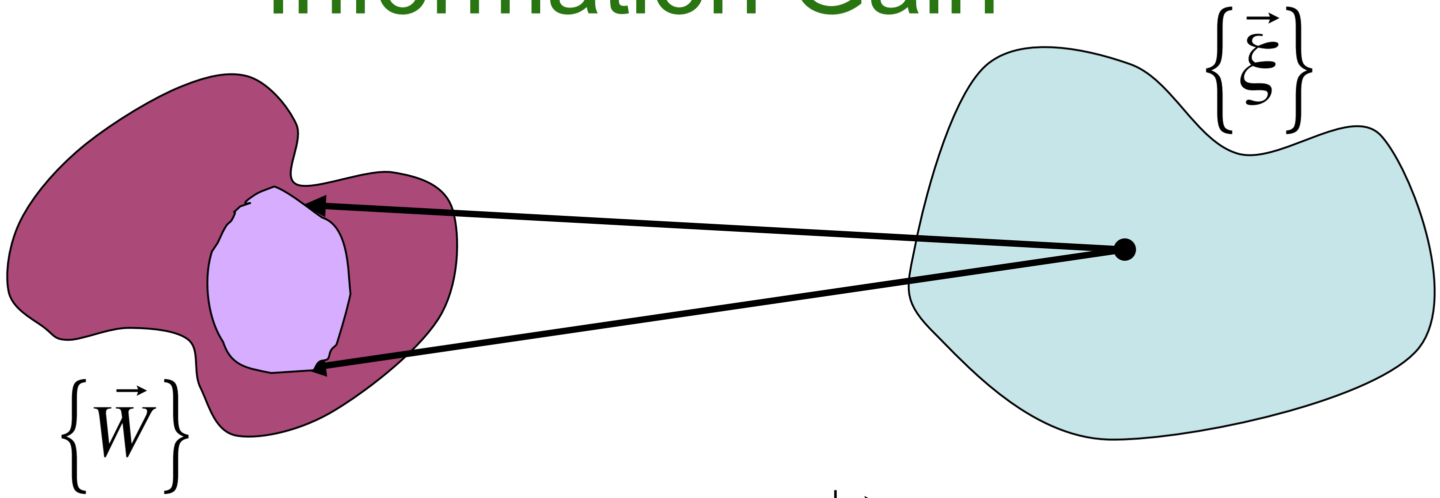
$$E_L = -\frac{1}{m}\frac{\partial}{\partial \beta}\left\langle\left\langle \ln Z_m \right\rangle\right\rangle_D$$

$$E_G = -\frac{1}{\beta}\frac{\partial}{\partial m}\left\langle\left\langle \ln Z_m \right\rangle\right\rangle_D + \frac{1}{\beta}\ln z(\beta)$$

Learning vs Generalization



Information Gain



$P(\vec{W}) = \rho_0(\vec{W})$: prior distribution

$P(\vec{W}|\vec{\xi})$: distribution induced by example $\vec{\xi}$

The entropy difference $\Delta H = H_{P(\vec{W})} - \left\langle \left\langle H_{P(\vec{W}|\vec{\xi})} \right\rangle \right\rangle_{P(\vec{\xi})}$

can be shown to be equal to the mutual information between the $\underbrace{\{\vec{W}\}}_{\text{the brain}}$ space and the $\underbrace{\{\vec{\xi}\}}_{\text{the world}}$ space.

Appendix.1

Require that the minimization of the learning error:

$$E_L(\vec{W}) = \sum_{\mu=1}^{m} E(\vec{W} \mid \vec{\xi}^{\mu})$$

guarantees the maximization of the likelihood:

$$\mathcal{L}(\vec{W}) = \prod_{\mu=1}^{m} P(\vec{\xi}^{\mu} \mid \vec{W})$$

Given a training set $\left(\vec{\xi}^{1}, \vec{\xi}^{2}, \ldots, \vec{\xi}^{m}\right)$, these two functions need to be related:

$$\mathcal{L}\left(\vec{W}\right) = \Phi\left(E_L\left(\vec{W}\right)\right)$$

Appendix.2

Take a derivative on both sides with respect to one of the points in the training set, $\vec{\xi}_j$:

$$\frac{\partial \mathcal{L}\left(D\middle|\vec{W}\right)}{\partial \vec{\xi}_j} = \mathcal{L}\left(D\middle|\vec{W}\right)\frac{1}{P\left(\vec{\xi}_j\middle|\vec{W}\right)}\frac{\partial P\left(\vec{\xi}_j\middle|\vec{W}\right)}{\partial \vec{\xi}_j} =$$

$$= \Phi' \frac{\partial E\left(\vec{W}\middle|\vec{\xi}_j\right)}{\partial \vec{\xi}_j}$$

This leads to:

$$\frac{\Phi'}{\Phi} = \frac{\dfrac{1}{P\left(\vec{\xi}_j\middle|\vec{W}\right)}\dfrac{\partial P\left(\vec{\xi}_j\middle|\vec{W}\right)}{\partial \vec{\xi}_j}}{\dfrac{\partial E\left(\vec{W}\middle|\vec{\xi}_j\right)}{\partial \vec{\xi}_j}}$$

Appendix.3

While the left-hand side of the equation depends on the full training set $\left(\vec{\xi}^1, \vec{\xi}^2, ..., \vec{\xi}^m\right)$, the right-hand side depends only on $\vec{\xi}^j$. The only way for this equality to hold for all values of $\left(\vec{\xi}^1, \vec{\xi}^2, ..., \vec{\xi}^m\right)$ is for both sides to be actually independent of the data, and thus equal to a constant:

$$\frac{\dfrac{1}{P\left(\vec{\xi}_j \middle| \vec{W}\right)} \dfrac{\partial P\left(\vec{\xi}_j \middle| \vec{W}\right)}{\partial \vec{\xi}_j}}{\dfrac{\partial E\left(\vec{W} \middle| \vec{\xi}_j\right)}{\partial \vec{\xi}_j}} = -\beta$$

Appendix.4

The equation
$$\frac{1}{P\left(\vec{\xi}_j\middle|\vec{W}\right)}\frac{\partial P\left(\vec{\xi}_j\middle|\vec{W}\right)}{\partial\vec{\xi}_j} = -\beta\frac{\partial E\left(\vec{W}\middle|\vec{\xi}_j\right)}{\partial\vec{\xi}_j}$$

leads to
$$P\left(\vec{\xi}_j\middle|\vec{W}\right) \propto \exp\left(-\beta E\left(\vec{W}\middle|\vec{\xi}_j\right)\right)$$

The normalized probability distribution is:
$$P\left(\vec{\xi}\middle|\vec{W}\right) = \frac{1}{z(\beta)}\exp\left(-\beta E\left(\vec{W}\middle|\vec{\xi}\right)\right)$$

with $z(\beta) = \int d\vec{\xi}\,\exp\left(-\beta E\left(\vec{W}\middle|\vec{\xi}\right)\right)$

Since the equation that determines $P\left(\vec{\xi}\middle|\vec{W}\right)$ is first order, there is only one constant of integration: β. For $\beta > 0$, E minima will correspond to P maxima.