Nonlinear Regularization

Jörg Lemm

Center for Biological and Computational Learning
Massachusetts Institute of Technology

Learning:

Data: e.g. $x_1 = 5$

Dependencies, e.g. $x_2 - x_1 = 2$ ("smoothness")

"Generalization", e.g. $x_2 = 7$

Generalization requires

1. reformulation (algorithmic aspect) and
2. control or identification (practical aspect)

of dependencies between test and training data.

Because every generalization is due to such "dependencies" we try to treat them as empirically measurable/ controllable data, just like standard "local" data. Therefore, instead of using a parametrization which induces non-empirical dependencies, we start with independent function values $f(x)$ and include dependencies as explicit as possible.

Two problems related to the "measurement of dependencies":

1. Terms like $\int dx \left(\frac{df(x)}{dx}\right)^2$ depend on an *infinite* number of x (local data).

2. Dependencies are often controlled by *humans* (E.g. objects classes are implicitly defined by selecting typical example objects, like faces, chairs, $\cdots$, which may, for example, include drawing of such objects or only realistic images.)

"Measurement/control/identification of dependencies":

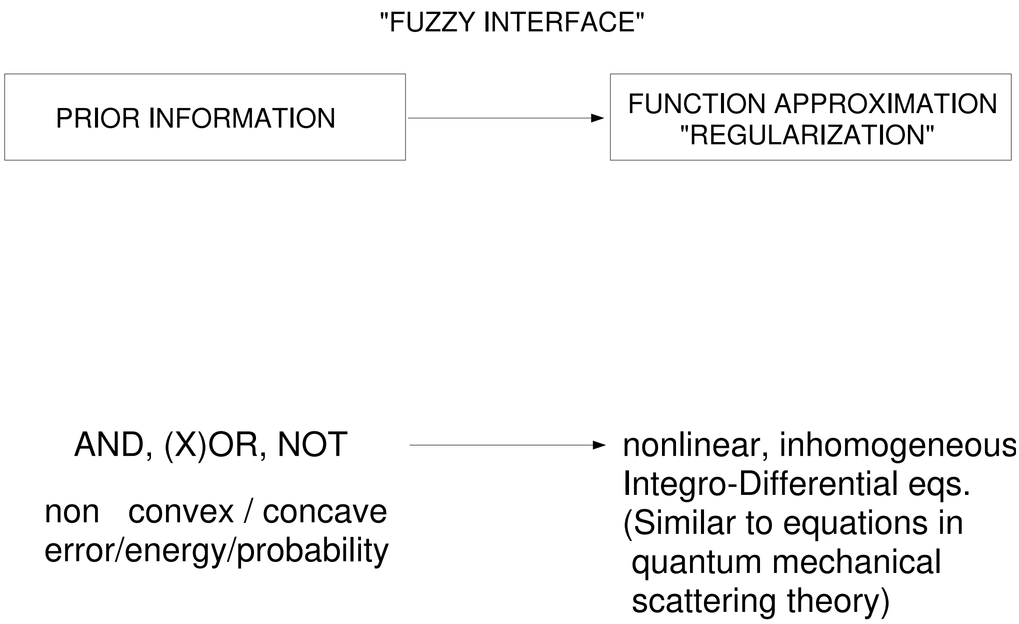
1. Infinity is not a real obstacle for empirical determination of dependencies. As the number of (local) test data is always finite (for the same practical reason as the number of training data is finite) those *dependencies must only be active for the finite number of test data*:

   a. *Bounds* are easily implemented by using measurement devices with *cut–off*.

   b. Approximate or exact *symmetries* are implemented by *averaging* ("coarse graining") or *input noise*. (E.g. smoothness for infinitesimal translations.)

   Both are realistic assumptions.

2. Verbal statements have to be translated in quantitative models, e.g. by a "*Fuzzy interface*".


"FUZZY INTERFACE"
PRIOR INFORMATION
FUNCTION APPROXIMATION
"REGULARIZATION"
AND, (X)OR, NOT
non convex / concave
error/energy/probability
nonlinear, inhomogeneous
Integro-Differential eqs.
(Similar to equations in
quantum mechanical
scattering theory)

"AND"

Probabilities:

$$p(\mathrm{AND}_i A_i) = \prod_i p(A_i), \quad \text{for independent } A_i,$$

Log–probabilities, $p = e^L$, $L = \ln p$:

$$\boxed{L(\mathrm{AND}_i A_i) = \sum_i L(A_i)}, \quad \text{for independent } A_i.$$

Linear in $L(A_i)$, e.g. $L(A_i) = -a(y_i - \bar{y}(x_i))^2 + b$.

"OR"

Probabilities:

$$p(\mathrm{OR}_i A_i) = \sum_i p(A_i), \quad \text{for } p(A_i \text{ AND } A_j) = 0, i \neq j,$$

Log–probabilities $p = e^L$, $L = \ln p$:

$$\boxed{L(\mathrm{OR}_i A_i) = \ln \sum_i e^{L(A_i)}}$$

Nonlinear in $L(A_i)$.

"DICTIONARY":

Probability: $p_i = e^{L_i} = e^{-\beta E_i + c(\beta)} = \frac{Z_i}{Z}$

Log–probability: $L = \ln p$

(negative information, bit–number)

Energy or *Error*: E

(shifted, β–scaled probability for elementary events, average of random variable for non–elementary events, canonical statistics of exponential family)

Free energy:

$$\boxed{F(\beta) = \frac{c(\beta)}{\beta} = -\frac{\ln Z(\beta)}{\beta} = -\frac{1}{\beta}\ln\sum_i e^{-\beta E_i}}$$

(cumulant generating function, shifted, β–scaled probability for non–elementary events)

Normalization constant or *partition sum*:

$$Z = \sum_i e^{-\beta E_i}$$

(unnormalized probability, moment generating function)

Inverse temperatur:

$$\beta = \frac{1}{T}$$

(Lagrange parameter (vector) for generalized Boltzmann–Gibbs distribution, canonical parameter of exponential family, regularization constant)

Realization of "(X)OR" = system at "finite temperature"

$$\boxed{F(\mathrm{OR}_i A_i) = \ln \sum_i e^{-\beta E(A_i)}}$$

Possible approximations:

high temperatur expansion (in β), first order: annealed approximation = AND, F cumulant generating function with respect to mixture probabilities.

low temperatur expansion, (restricting to most probable events), for continuous x *saddle point approximation*: maximum posterior approximation, empirical risk minimization (from Bayesian point of view), mean field equations, "classical physics", fluctuations around extremum: Gaussian (free field or random phase) approximation.

LINEAR REGULARIZATION

$$L(\bar{y}) = -\beta E(\bar{y}) + c(\beta) = -\frac{\beta}{2}\sum_{j=0}\Delta_j + c = -\frac{\beta}{2}\tilde{\Delta} + c$$

Quadratic energy terms (free field or Gaussian process)

$$\boxed{\Delta_j = ||\bar{y} - T_j||_{\mathcal{O}} =< \bar{y} - T^j|\mathcal{O}^j|\bar{y} - T^j >}$$

representing a canonical $\mathcal{O}$–distance to *template* T_j. (E.g. for RBF (Poggio & Girosi, 1990; Girosi, Jones, & Poggio, 1995), $\mathcal{O}^j = \Sigma_{m=0}^{\infty}(-1)^m \frac{\sigma^{2m}}{m!2^m}\nabla^{2m}$)

$$\tilde{\Delta}(\bar{y}) = \bar{\Delta}(\bar{y}) + E_{\mathcal{O}}$$

with minimal error or ground state energy

$$E_{\mathcal{O}} = N_j\,\mathrm{VAR}^{\mathcal{O}}(T^j) = \sum_j^{N_j} <T^j|\mathcal{O}^j|T^j> - <\overline{T}^{\mathcal{O}}|\mathcal{O}|\overline{T}^{\mathcal{O}}>$$

and distance from the "template average"

$$\bar{\Delta} =<\bar{y}-\overline{T}^{\mathcal{O}}|\mathcal{O}|\bar{y}-\overline{T}^{\mathcal{O}}>, \quad \overline{T}^{\mathcal{O}} = \mathcal{O}^{-1}\sum_j \mathcal{O}^j T^j \quad \mathcal{O} = \sum_j \mathcal{O}^j.$$

E.g. data term

$$\Delta_0 = \Delta_D =< \bar{y} - \bar{D}|\mathcal{N}^D|\bar{y} - \bar{D} > .$$

Solution:

$$\boxed{\bar{y} = \overline{T}^{\mathcal{O}}}$$

FINITE TEMPERATURE REGULARIZATION

Implementing OR by a mixture model, i.e. system at finite temperature with β interpolating between OR and AND:

$$\boxed{L = L^M = \ln\left(\sum_i^{N_i} e^{-\frac{\beta}{2}\sum_{j_i}^{N_{j_i}} \Delta_{i,j_i}+c_i(\beta)}\right) = \ln\left(\sum_i^{N_i} e^{-\frac{\beta}{2}\tilde{\Delta}_i+c_i(\beta)}\right)}$$

Stationarity equation:

$$\mathbf{O}^M \bar{y} = \mathbf{t}^M$$

$$\mathbf{O}^M(\bar{y}) = \overline{\mathcal{O}}^z = \frac{\sum_i z_i \mathcal{O}^i}{z} = \frac{\mathcal{O}^z}{z},$$

$$\mathbf{t}^M(\bar{y}) = \overline{T}^z = \frac{\sum_i z_i T^i}{z} = \frac{T^{z,\mathcal{O}}}{z},$$

$$z_i = e^{-\frac{\beta}{2}\sum_{j_i}\Delta_{i,j_i}(\bar{y})+c_i}, \;\; z = \sum_i z_i,$$

$$\mathcal{O}^i = \sum_{j_i} \mathcal{O}^{ij_i}, \;\; T^i = \sum_{j_i} \mathcal{O}^{ij_i} T^{ij_i} = \mathcal{O}^i\,\overline{T}^{\mathcal{O}^i}.$$

INTRODUCTORY TWO TEMPLATE MODEL

Standard data term

$$\mathcal{O}^D = \lambda_D \mathcal{N}^D,$$

with two templates

$$T_1 = -\sin\left((x-1)/\alpha\right) - b_1,$$

$$T_2 = \sin^2\left((x-1)/\alpha\right) - b_2,$$

and "smoothness" distance

$$\mathcal{O}^s = \lambda_0 \mathcal{O}_0 + \lambda_2 \mathcal{O}_2 + \lambda_4 \mathcal{O}_4$$

with

$$\mathcal{O}_0(x, x') = \delta(x - x'),$$

$$\mathcal{O}_2(x, x') = \delta(x - x') \frac{d^2}{dx^2},$$

$$\mathcal{O}_4(x, x') = \delta(x - x') \frac{d^4}{dx^4}.$$

FINITE TEMPERATURE REGULARIZATION: TWO TEMPLATES

Mixture model:

$$e^L = p \propto z = e^{-\beta\frac{1}{2}(\Delta_D+\Delta_1)+c} + e^{-\beta\frac{1}{2}(\Delta_D+\Delta_2)+c}$$

$$= e^{-\frac{\beta c}{2}\frac{\tilde{\Delta}_1+\tilde{\Delta}_2}{2}} 2\cosh\left(\beta(\tilde{\Delta}_2 - \tilde{\Delta}_1)/4\right).$$

Stationarity equation:

$$\bar{y} = (z_1/z)\overline{T}^{\mathcal{O}^1} + (z_2/z)\overline{T}^{\mathcal{O}^2}$$

$$= \overline{T}^{\mathcal{O}} + \tanh\left(\beta(\tilde{\Delta}_2 - \tilde{\Delta}_1)/4\right)(\overline{T}^{\mathcal{O}^1} - \overline{T}^{\mathcal{O}^2})/2,$$

with solutions in the one–dimensional subspace spanned by the two component averages. This is the form of the form of the famous ferromagnetic equation

$$\bar{y} = \tanh(\beta\bar{y})$$

.

LANDAU–GINZBURG REGULARIZATION

Instead of combining "free models" one may directly specify nonconvave ("interacting") log–probabilities by fuzzy logical methods.

$$\boxed{L = L^I = -g\left(\frac{1}{2}\sum_i \prod_{j_i} \Delta_{i,j_i}\right)}$$

Stationarity equation:

$$\mathbf{O}^I = \sum_i \sum_{j_i} M^{ij_i} \mathcal{O}^{ij_i},$$

$$\mathbf{t}^I = \sum_i \sum_{j_i} M^{ij_i} \mathcal{O}^{ij_i} T^{ij_i},$$

$$M^{ij_i} = M^{ij_i}(\bar{y}) = \prod_{k_i \neq j_i} \Delta_{ik_i}.$$

LANDAU–GINZBURG REGULARIZATION: TWO–TEMPLATE MODEL

Product as simplest form of a “fuzzy OR”:

$$L^I = -g((1-\gamma)\Delta_D + \gamma\Delta_1\Delta_2),$$

with inverse–temperature–like convex parameter

$$\beta^I = (\gamma/1-\gamma)$$

The stationarity equation is according to

$$\mathbf{O}^I = \mathcal{O}^D + \beta^I(\Delta_1 + \Delta_2)\mathcal{O}^s$$

$$\mathbf{t}^I = \mathcal{O}^D \bar{D} + \beta^I \mathcal{O}^s(\Delta_2 T^1 + \Delta_1 T^2)$$

a cubic equation in $\bar{y}$ with, depending on β^I, either one or two stable solutions. This corresponds to the phenomenological Landau–Ginzburg treatment of phase transitions.

ITERATION PROCEDURES = LEARNING

The stationarity equations are in both models in general inhomogeneous integral equations. As nonlinear equations they have to be solved by iteration ("learning"), where an iteration procedure is defined by selecting some operator $\mathcal{A}$ (usually positive definite, possibly i and $\bar{y}^i$–dependent) and a relaxation factor η according to

$$\bar{y}^{i+1} = \bar{y}^i - \eta \mathcal{A}^{-1}(\mathbf{O}\bar{y}^i - \mathbf{t}).$$

Two extreme possibilities:

1. Gradient (on L surface) $\mathcal{A} = \mathcal{I}$ (identity)
$$\bar{y}^{i+1} = \bar{y}^i - \eta(\mathbf{O}\bar{y}^i - \mathbf{t})$$
2. full relaxation $\mathcal{A} = \mathbf{O}$ (for mixture model equivalent to EM)
$$\bar{y}^{i+1} = \bar{y}^i - \eta(\bar{y}^i - \mathbf{O}^{-1}\mathbf{t})$$

Intermediate methods, are for example Gaussian $\mathcal{A}^{-1}$ or multigrid methods.

CONCLUSIONS:

AIM: a general "prior/fuzzy interface" for function approximation/classification (especially for continuous x). Interactive generation of "templates" and relevant "template distances".

1. *Practical aspect*: Measuring/controlling/identifying dependencies between training and test data ("*prior knowledge*") is crucial for any generalization. There still seems to be much potential for learning algorithms in including dependencies as explicit as possible the formalism, especially if "humans/experts" are integrated as "measurement devices".

2. *Theoretical aspect*: It is possible to include general prior knowledge in a theoretical formalism. In general, also in function approximation the error/-probability/energy surfaces are non-convex/non–concave (XOR–like). The resulting saddle point equations are *nonlinear* (in the component (log–)probabilities or in the regression function).

3. *Technical aspect*: A temperature-like parameter allows to interpolate between AND and OR, i.e. to change the degree of non–convexity/ non–concavity. This creates typical *phase transition/bifurcation* behavior. There is a huge freedom in choosing the iteration ("learning") procedures which should/have to be adapted to the dependency structure.