# Fuzzy Interface with Prior Concepts and Non–convex Regularization

Jörg C. Lemm

Universität Münster
Institut für Theoretische Physik I
Wilhelm–Klemm–Str.9
D–48149 Münster, Germany
Phone: (+49)251/83–34922
Fax (+49)251/83–36328
Email: lemm@uni-muenster.de

## Introduction

Generalization is based on the possibility to guarantee constancy of probability distributions and of dependencies between variables. For functions of continuous variables control of dependencies corresponds formally to an infinite amount of data (prior information), Nevertheless, prior knowledge can be "measured" by posterior control. Practically, prior information is most times available only in fuzzy (e.g. verbal) form. The approach presented here includes prior information (continuous data) as explicit as possible, quite analogous to the treatment of training data.

# Bayesian approach

Expected risk minimization

$$\boxed{\begin{aligned} h^* &= \mathrm{argmin}_h\, r(h,f) \\ r(h,f) &= \int df^0\, p(f^0|f) r(h,f^0) \end{aligned}}$$

$$r(h,f^0) = \int dx \int dy\, p(x) p(y|x,f^0) l(x,y,h)$$

$x$ independent, $y$ dependent variables
$f^0$ pure state, $f$ state of knowledge,
$h$ hypothesis, action, $l$ loss function

Approximation methods for sums/integrals
• Sampling (Monte Carlo)
• High temperature expansion (annealed app, moment or cumulant generating function)
• Low temperature expansion (saddle point app.)

Approximation problems:
Identify $h$ with $f^0$, choose $l(x,y,h) = \ln p(y|x,f^0)$

Saddle point approximation

$$\boxed{\begin{aligned} h^* &= \mathrm{argmax}_h\, p(h|f) \\ &= \mathrm{argmin}_h (-\ln p(h|f)) \end{aligned}}$$

$$p(h|f(D)) = \frac{p(D|h)p_0(h)}{p(D)} \propto \prod_i p(y_i|x_i,h) p_0(h)$$

# Quadratic concepts

**Definition 1 (quadratic concept)**

$$\boxed{d^2[h] = \langle h - t|O|h - t\rangle}$$

$h = h(x)$ *hypothesis function*,
$t = t(x)$ *template (prototype, data) function*,
$O = P_t O_t P_t$, $P_t$ = *projector*,
$O_t = W^T W = (I - S)^T (I - S)$, *symmetric, positive definite.*

**Example 1 (Data: Mean square error)**

$$\boxed{\langle h - t|O|h - t\rangle = (y_i - h(x_i))^2}$$

$t(x) = y_i$, $O(x, x') = \delta(x - x')\delta(x - x_i)$.

**Example 2 (Smoothness, kinetic energy)**

$$\boxed{\langle h - t|O|h - t\rangle = -\langle h|\Delta|h\rangle}$$

$t(x) \equiv 0$, (zero template)
$O(x, x') = -\Delta(x, x') = -\delta(x - x') \sum_{l=1}^{d} \frac{\partial^2 h(x)}{\partial x_l^2}$
(negative $d$–dimensional Laplacian).
This is an example of general symmetries $S$, i.e. with $Sh(x) = h(Tx)$ ($T$ invertible transformation). For Lie groups $S = e^{\theta s}$ with generator $s$ one finds $\lim_{\theta \to 0}(I - S)/\theta = s$.

# AND: Linear regularization

**Proposition 1 (Probabilistic AND)**

$$E = \sum_i^N d_i^2[h]/2 = d^2[h]/2 + E_0$$

with square distance

$$d^2[h] = \langle h - \bar{t}|O|h - \bar{t}\rangle$$

from template average

$$\bar{t} = O^{-1}\tilde{t} \text{ with } \tilde{t} = \sum_i^N O_i t_i \text{ and } O = \sum_i O_i$$

and $h$–independent minimal energy

$$E_0 = \frac{1}{2}\left(\sum_i^N \langle t_i|O_i|t_i\rangle - \langle \bar{t}|O|\bar{t}\rangle\right).$$

The stationarity equation is linear and reads

$$0 = \tilde{t} - Oh \Rightarrow h = \bar{t}.$$

**Example 3 (Radial Basis Functions)**

$$E = d_0^2[h] + \sum_{i=1}^N d_i^2[h],$$
$$O_i(x, x') = \delta(x - x')\delta(x - x_i),$$
$$t_i = y_i \text{ (data)},$$
$$O_0(x, x') = \sum_{m=0}^{\infty}(-1)^m \frac{\sigma^{2m}}{m!2^m}\Delta^m,$$
$$t_0 \equiv 0 \text{ (smoothness)}.$$

# OR: Mixture regularization

**Proposition 2 (Probabilistic OR)**
Implementing the probabilistic OR (for exclusive events) by a mixture (or finite temperature) model

$$p(h) = \sum_i^{N_i} p(h,i) = \sum_i^{N_i} p_i\, p(h|i),$$
$$p(h|i) = e^{-\beta E_i}/Z_i,$$
$$Z_i = \int dh\, e^{-\beta E_i},$$
$$p_i = e^{-b e_i}/z,\; z = \sum_i e^{-b e_i}$$

yields

$$\boxed{E^M[h] = \ln\left(\sum_i^{N_i} e^{-\beta E_i[h] - b e_i - \ln Z_i(\beta)}\right),}$$

$$\text{with } E_i = (1/2) \sum_{j_i}^{n_{j_i}} d_{ij_i}^2.$$

The (nonlinear) stationarity condition becomes

$$\boxed{0 = t^M[h] - O^M[h]h,}$$

with

$$O^M = \sum_i p(h,i) O_i,$$
$$t^M = \sum_i p(h,i) O_i \bar{t}_i = \sum_i p(h,i) \tilde{t}_i.$$

# OR: Landau–Ginzburg regularization

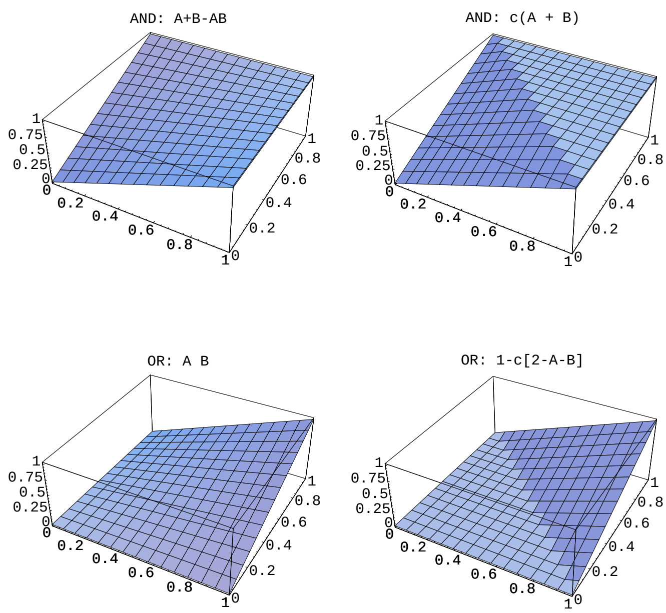
**Proposition 3 (“Fuzzy” OR)**
A product implementation of OR yields models with polynomial interactions

$$\boxed{E^F = \frac{1}{2}\sum_{i=1}^{n_i} \beta_F^{n_{j_i}-1} \prod_{j_i=1}^{n_{j_i}} d_{ij_i}^2,}$$

with (nonlinear) stationarity equation

$$\boxed{0 = t^F[h] - O^F[h]h.}$$

$$\begin{aligned} O^F[h] &= \sum_i \sum_{j_i} M_{ij_i}[h] O_{ij_i}, \\ t^F[h] &= \sum_i \sum_{j_i} M_{ij_i}[h] O_{ij_i} t_{ij_i}, \\ M_{ij_i}[h] &= \beta_F^{n_{j_i}-1} \prod_{k_i \neq j_i} d_{ik_i}^2[h]. \end{aligned}$$



# Iteration procedures = learning

$$h^{k+1} = h^k + \eta A^{-1}(t^k[h^k] - O^k[h^k]h^k)$$

| Learning algorithm | Learning matrix |
|---|---|
| Gradient | $\mathbf{A} = \mathbf{I}$ (Identity) |
| Jacobi | $\mathbf{A} = \mathbf{D}$ (diagonal) |
| Gauss–Seidel | $\mathbf{A} = \mathbf{D} + \mathbf{L}$ (low.triang.) |
| Newton | $\mathbf{A} = -\mathbf{H}$ (neg. Hessian) |
| $\mathbf{O}$–relaxation | $\mathbf{A} = \mathbf{O}$ (EM for mixture) |
| Massive relaxation | $\mathbf{A} = \mathbf{A}_0 + m^2\mathbf{I}$ |
| Linear boundary | $\mathbf{A} = \mathbf{A}_{II} + \mathbf{A}_{IB} + \mathbf{A}_{BB}$ |
| Gaussian | $\mathbf{A} = \sum_{k=0}^{\infty} \frac{1}{k!}\left(\frac{\mathrm{M}^2}{2\tilde{\sigma}^2}\right)^k = e^{\frac{\mathrm{M}^2}{2\tilde{\sigma}^2}}$ |

## Example 4 (Hessian, finite temperature)

$$\begin{aligned} H^M &= -\beta \textstyle\sum_i p(h,i) O_i \\ &+\beta^2 \textstyle\sum_i p(h,i) |O_i h - t_i\rangle\langle O_i h - t_i|) \\ &-\beta^2 \textstyle\sum_{ij} p(h,i) |O_i h - t_i\rangle\langle O_j h - t_j| p(h,j). \end{aligned}$$

## Example 5 (Hessian, Landau–Ginzburg)

$$\begin{aligned} H^L &= \textstyle\sum_i \sum_{j_i} M_{ij_i} O_{ij_i} \\ &+ \textstyle\sum_i \sum_{j_i} \sum_{k_i} M_{ij_ik_i} |O_{ij_i}h - t_{ij_i}\rangle\langle O_{ij_i}h - t_{ij_i}| \end{aligned}$$

with $M_{ij_ik_i} = \prod_{j_i \neq l_i \neq k_i} \beta_F^{n_{j_i}-1} d^2_{il_i}[h]$

# A numerical example

(data AND $t_1$) OR (data AND $t_2$)

Finite temperature/mixture regularization

$$\boxed{\begin{aligned} p^M &= p_1^M + p_2^M \propto e^{-\beta E} = e^{-\beta E_1} + e^{-\beta E_2} \\ &= e^{-\frac{\beta}{2}(d_D^2 + d_1^2)} + e^{-\frac{\beta}{2}(d_D^2 + d_2^2)} \end{aligned}}$$

$d_D^2 = \sum_i (y_i - h(x_i))^2$
$d_i^2 = -\langle h - t_i | \Delta | h - t_i \rangle$(*non–zero templates* !)
The stationarity equation $h = p_1^M \bar{t}_1 + p_2^M \bar{t}_2$
has the ferromagnetic form

$$\boxed{h = \bar{t} + \tanh\left[\frac{\beta}{2}(E_2 - E_1)\right] \frac{\bar{t}_1 - \bar{t}_2}{2}}$$

Landau–Ginzburg regularization

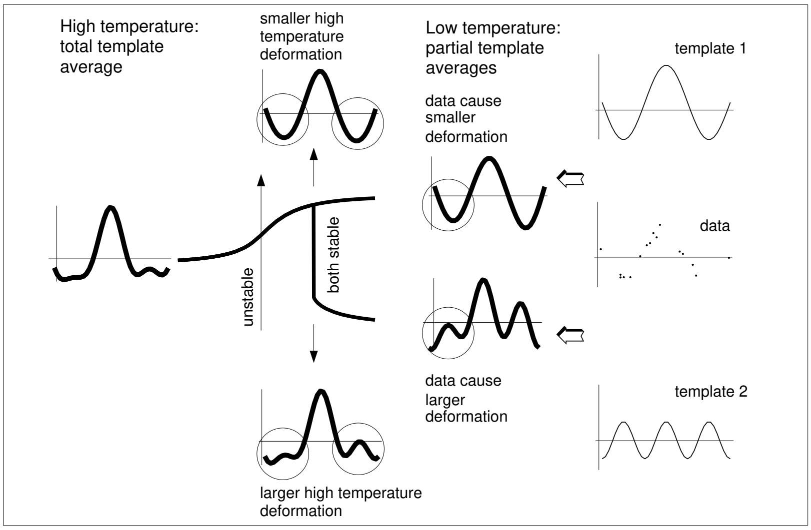
$$\boxed{E = g(d_D^2 + \beta\, d_1^2\, d_2^2)}$$

($g$ strictly monotone)
with stationarity equation

$$\boxed{\begin{aligned} &\left[O_D - \beta(d_1^2 + d_2^2)\Delta\right] h \\ &= O_D D - \beta\Delta(d_1^2\, t_1 + d_2^2\, t_2) \end{aligned}}$$

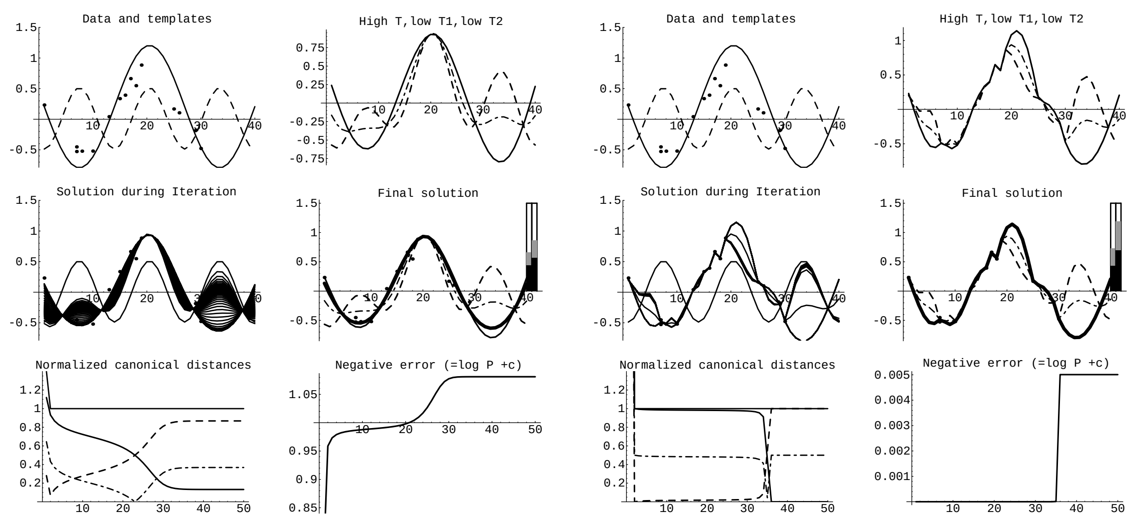
$O_D(x, x') = \delta(x - x') \sum_i \delta(x - x_i)$
$D(x) = \sum_n \delta(x - x_i) y_i / (\sum_n \delta(x - x_i))$

# Temperature dependence




Linear problems:

- High temperature limit
- (two) low temperature limits,
- classical regularization ($t_1 = t_2 \equiv 0$).



# Continuous transformations

OR–like combinations describe variations of templates. Often those are continuously parameterized (translations, scaling, $\cdots$).

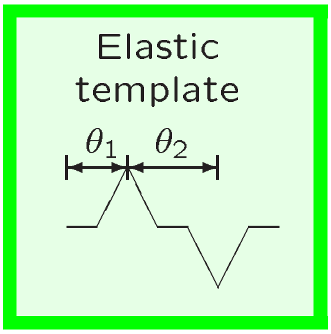
$$p^M \propto \int d\theta\, p(\theta) \sum_i e^{-\frac{\beta}{2}\sum_{j_i} \langle h - t(\theta) | O_{j_i} | h - t(\theta) \rangle + \beta F}$$

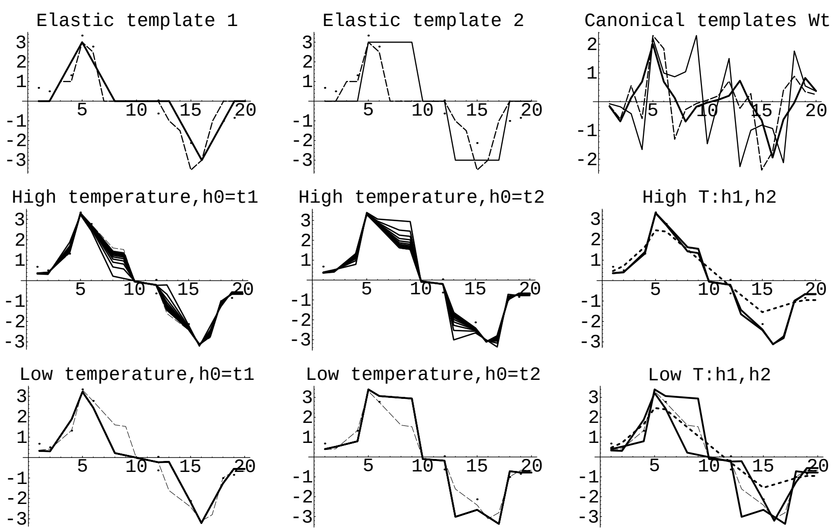
Now treat

☞ $\int d\theta$ by saddle point approximation

☞ $\sum_i$ exactly

Algorithm:
1. Find optimal $\theta^*$ in data space
2. Solve for $h^*$ for fixed $\theta^*$
3. Adapt $\theta^*$ for fixed $h^*$
4. Iterate till convergence.



# Density estimation

The stationarity equations for density estimation are nonlinear from the beginning due to a non–quadratic data term $\sum_i \ln P_i$ and the normalization condition.

$$\boxed{-\beta E_\phi = \langle \ln P[\phi] | N \rangle - \frac{1}{2} \langle \phi | \mathbf{O} | \phi \rangle - \langle P[\phi] | \Lambda_X \rangle,}$$

$P[\phi]$ (nonlinear) mapping from $\phi$ to $P$

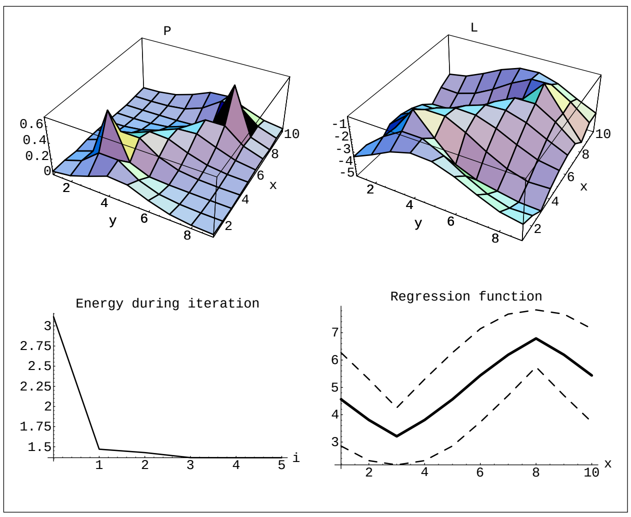
$\Lambda_X$ Lagrange multiplier function

$$N(x, y) = \sum_i \delta(x - x_i)\delta(y - y_i)$$

$$\boxed{\begin{aligned} 0 &= \mathbf{P}'[\phi] \left( \mathbf{P}^{-1}[\phi] N - \Lambda_X \right) - \mathbf{O}\phi \\ \Lambda_X &= \mathbf{I}_X \left( N - \mathbf{P}\mathbf{P}'^{-1} \mathbf{O}\, \phi \right) \end{aligned}}$$

$$\mathbf{P}(x, y; x', y') = \delta(x - x')\delta(y - y') P[\phi](x, y)$$

$$\mathbf{P}'[\phi](x, y; x', y') = \delta P[\phi](x', y') / \delta\phi(x, y)$$



# References


Bishop, C.M. (1995) Training with noise is equivalent to Tikhonov regularization. *Neural Computation* **7** (1), 108–116.

Girosi, F., Jones, M., & Poggio, T.: Regularization Theory and Neural Networks Architectures. *Neural Computation* **7** (2), 219–269, 1995

Itzykson, C.& Drouffe, J–M: Statistical field theory (vol 1 & vol 2). Cambridge University Press, 1989.

Klir, G.J. & Yuan, B. (eds.): Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems. World Scientific, 1996.

Lemm, J.C.: Inhomogeneous Random Phase Approximation for Nuclear and Atomic Reactions. *Annals of Physics* **244** (1), 136–200 & 201–238, 1995.

Lemm, J.C. (1996) Prior Information and Generalized Questions. A.I. Memo No. 1598, C.B.C.L. Paper No. 141, Massachusetts Institute of Technology.

Poggio, T. & Girosi, F.: Networks for Approximation and Learning. Proceedings of the IEEE, Vol 78, No. 9, 1990.

Rose, K., Gurewitz, E., & Fox, G.C.: Statistical mechanics and phase transitions in clustering. *Phys. Rev. Lett.* **65**, 945–948, 1990.

Tikhonov, A.N.: Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.* **4**, 1035–1038, 1963.

Wahba, G.: Spline Models for Observational Data. Philadelphia: SIAM, 1990.

Yuille, A.L. & Kosowski, J.J.: Statistical Physics Algorithm That Converge. *Neural Computation* **6** (3), 341–356, 1994.