Nonlinear Regularization with Prior Concepts

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
WWW: http://planck.uni-muenster.de:8080/~lemm

*Parts of this work have been done at the Center for Biological and Computational Learning, Massachusetts Institute of Technology, Cambridge MA, 02139

AMAI'98

Learning (deterministic toy example)

Training data: e.g. $h(x_1) = 5$

Test data: $h(x_2) =?$

Dependencies (prior information, rules):
e.g. $h(x_1) - h(x_2) = 2$ ("smoothness")

Generalization: $h(x_2) = 7$

$\Rightarrow$

Generalization requires:

- ☞ (algorithmic aspect) reformulation of knowledge to find answers for test set given training data and dependencies.

- ☞ (control aspect) control of dependencies between test and training data (e.g. stationary joint distribution, prior information = posterior control)

Explicit priors

Generalization is essentially based on dependencies (prior information). Often prior information is implicit, hidden in the parameterization of $h(x)$ or the learning algorithm.

To work with priors as explicit as possible we aim in formulating priors directly *in terms of the function values* $h(x)$ itself.

This leads to the framework of *regularization theory* or the theory of *random processes and random fields*.

Two problems related to "measurement of dependencies"

1. Often there exist formally an *infinite* number of potential test data (e.g. $x \in \mathbb{R}$). Then typical regularization terms like

$$\int dx \left(\frac{dh(x)}{dx} \right)^2$$

   (have to) depend on an infinite number of function values $h(x)$. How can an infinite amount of data be empirically measured?

2. Dependencies are often controlled by *humans*. (E.g. object classes are implicitly defined by selecting typical examples, like faces, chairs, cars, typical symptoms, ..., for training and test set.) How can complex data generating processes be modelled?

Measurement of prior information = posterior control of dependencies

Infinity is not a real obstacle for empirical measurement of dependencies. As the number of actually measured (local) test data is always finite (for the same practical reasons as for training data) those *dependencies must only be active for the finite number of times test data are actually measured*. For example:

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.

"Fuzzy Interface"

Models of complex generating processes are constructed by combining building blocks ("*quadratic prior concepts*") with AND and (X)OR–like combinations. to define the regularization (error) functional E.

AND: classical linear regularization theory (Tikhonov, 1963)

$$\boxed{E = \text{data AND smoothness}}$$

OR: non–convex error surface, regularization theory (nonlinear also in function values, not only in parameters.)

$$\boxed{\begin{array}{c} E = \text{(brown left eye AND brown right eye)} \\ \text{OR (blue left eye AND blue right eye)} \end{array}}$$

Different variants for eyes (e.g. brown vs. blue, open vs. closed, but also translations, deformations, ...) are combined by OR.

Bayesian approach

Expected risk minimization

$$\boxed{\begin{array}{l} h^* = \text{argmin}_h\, r(h,f) \\ r(h,f) = \int df^0\, p(f^0|f) r(h,f^0) \end{array}}$$

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

$$\boxed{\begin{array}{l} h^* = \text{argmax}_h\, p(h|f) \\ = \text{argmin}_h(-\ln p(h|f)) \end{array}}$$

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

Quadratic concepts

Definition 1 (quadratic concept)

$$\boxed{d^2[h] = \langle h - t|O|h - t\rangle,\ p \propto e^{-\beta d^2/2}}$$

$h = h(x)$ *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)

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

$t(x) \equiv 0$, (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)

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

with square distance

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

from template average

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

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

Example 3 (Radial Basis Functions)

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

Solving in data space

$$\boxed{E[h] = \langle h - t_D | O_D | h - t_D \rangle + \langle h - t_0 | O_0 | h - t_0 \rangle}$$

$\langle h - t_D | O_D | h - t_D \rangle = \sum_i^n (y_i - h(x_i))^2 +$ c

The stationarity equation

$$0 = O_D(h - t_D) + O_0(h - t_0)$$

yields $\boxed{h = O_0^{-1} a + t_0}$, with $a = O_D(t_D - h)$
For known covariance $\Sigma = O_0^{-1}$ it is sufficient to solve

$$\boxed{a = [I + O_D O_0^{-1}]^{-1} [O_D(t_D - t_0)]}$$

in the n–dimensional space of functions defined on data points x_i.

Examples include kernel methods like Radial Basis Functions or splines.

OR: Two template example (I)

$$\boxed{(\text{data AND } t_1) \text{ OR } (\text{data AND } t_2)}$$

$t_1(x)$ and $t_2(x)$ represent two (alternative) function templates (prototypes).

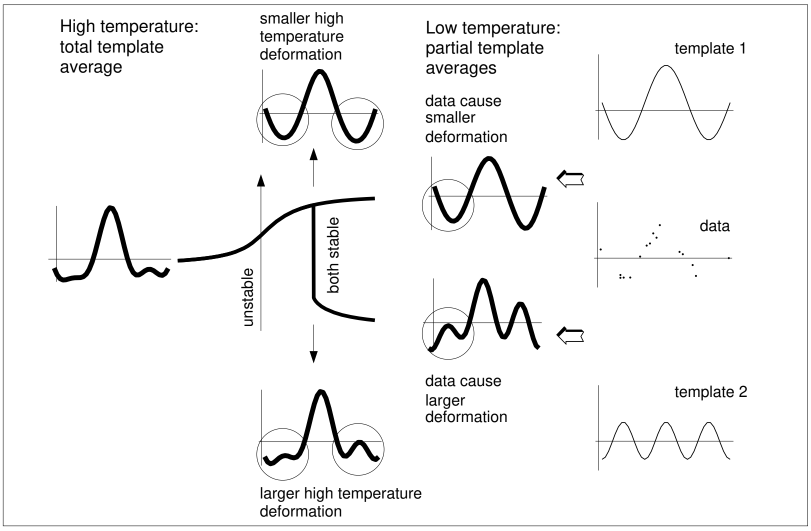
I. Finite temperature/mixture regularization

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

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

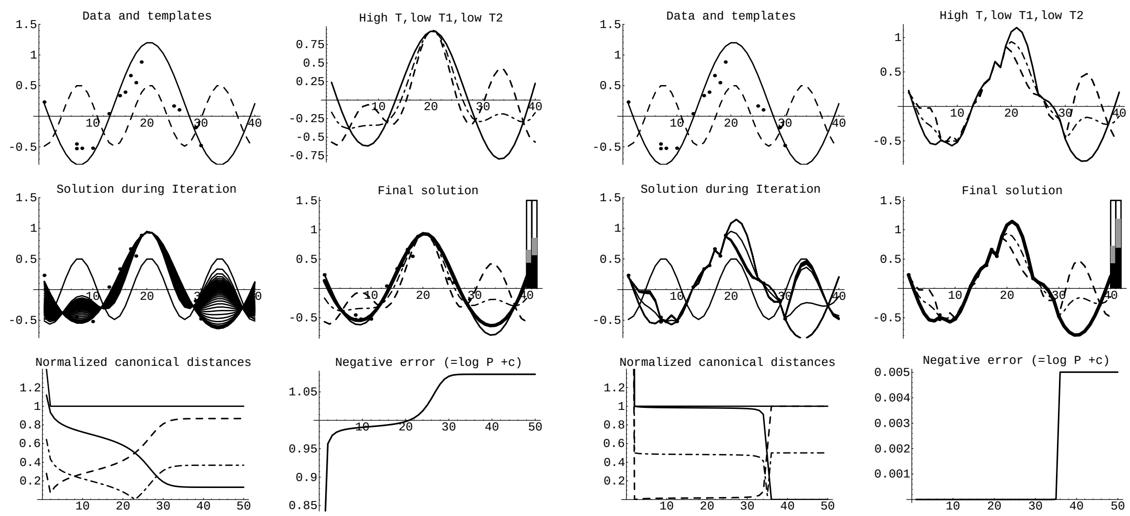
$$\boxed{h = \bar{t} + \tanh\left[\frac{\beta}{2}(E_2 - E_1)\right] \frac{\bar{t}_1 - \bar{t}_2}{2}}$$

Temperature dependence



Linear problems:

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



OR: Two template example (II)

$$\boxed{(\text{data AND } t_1) \text{ OR } (\text{data AND } t_2)}$$

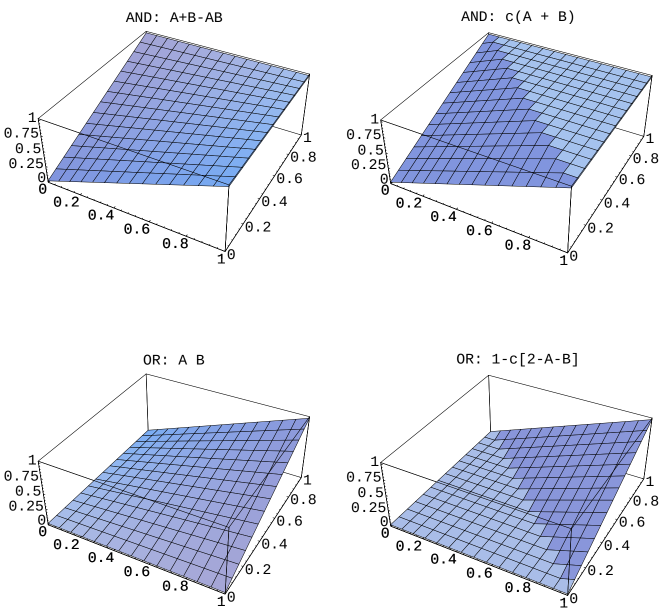
II. Landau–Ginzburg regularization

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

(g strictly monotone)
with stationarity equation

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

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



OR: Mixture regularization

Proposition 2 (Probabilistic OR)

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

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

yields a spin glas like error functional

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

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

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

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

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)

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

Example 5 (Hessian, Landau–Ginzburg)

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

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

Support vector machine(Vapnik, 1995)

$$\boxed{\begin{array}{l} E[h] = \langle h - t_D | O_D | h - t_D \rangle + \langle h | O_0 | h \rangle \\ = \sum_i^n (y_i - h(x_i))^2 + \int dx\, dx'\, h(x) O_0(x, x') h(x') \end{array}}$$

Eigenfunction decomposition of covariance operator

$$O_0 = \Sigma^{-1} = \sum_k \lambda_k \Phi_k \Phi_k^T \qquad (1)$$

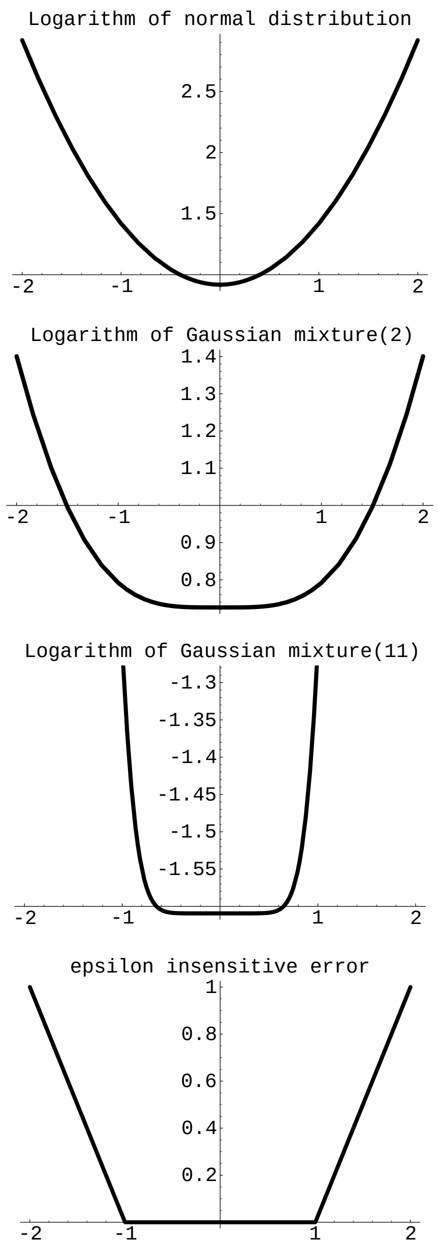
$$h(x) = \sum_k n_k \Phi_k(x) \qquad (2)$$

yields

$$\boxed{E[h] = \sum_i \left(\sum_k n_k \Phi_k(x_i) - y_i \right)^2 + \sum_k \lambda_k |n_k|^2}$$

Robust statistics: replace mean square error on data by an OR–like combination like a mixture model (e.g. at a (second order) phase transition) or ϵ–insensitive error (SVM, quadratic programming problem).

Quadratic error, Gaussian mixtures and ϵ–insensitive error



Continuous transformations

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

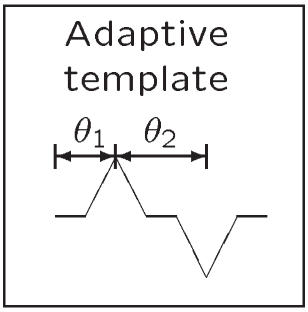
$$\boxed{p^M \propto \int d\theta\, p(\theta) \sum_i e^{-\frac{\beta}{2}\sum_{j_i}\langle h-t(\theta)|O_{j_i}|h-t(\theta)\rangle+\beta F}}$$

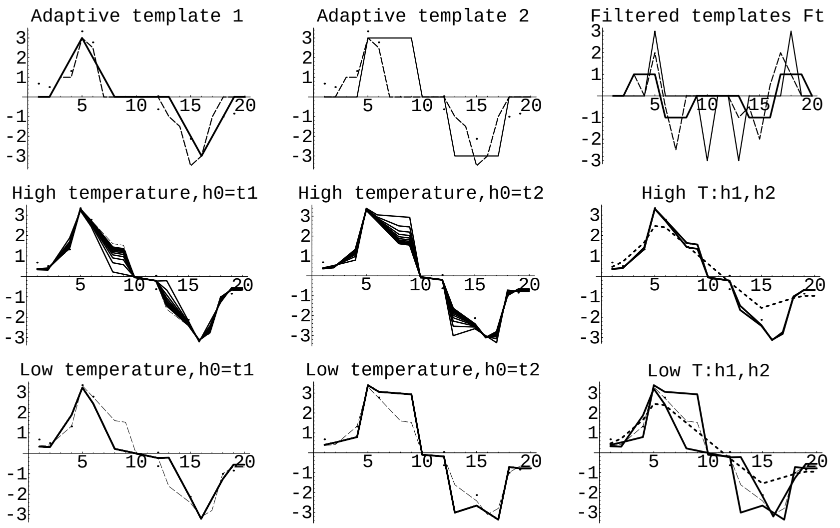
Now treat

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

☞ $\sum_i$ exactly

Algorithm:
1. Find optimal θ^* in data space
2. Solve for h^* for fixed θ^*
3. Adapt θ^* 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 ϕ to P

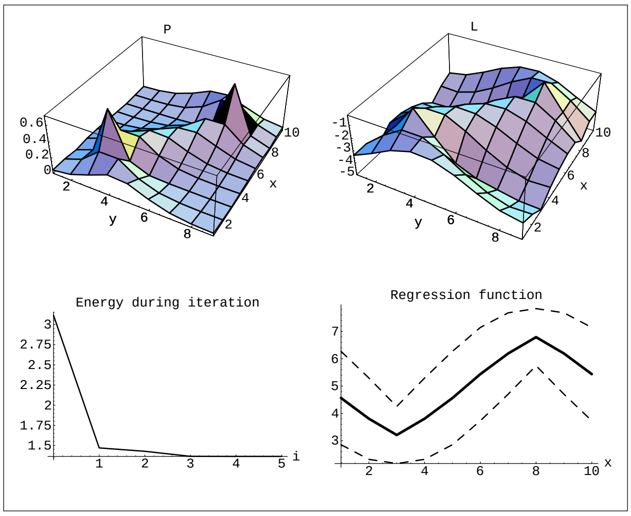
Λ_X Lagrange multiplier function

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

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

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

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



Open problems

For non–Gaussian processes the calculations cannot be performed in an n–dimensional space. (n number of training data).

For low dimensional systems (time series, spatial processes, image reconstruction) the resulting minimization problem is feasible, but in higher dimensions saddle point approximations and sampling become extremely difficult.

(Similar to lattice field theories in some cases mean field approximations may become more accurate in higher dimensions.)

Summary

1. Measuring *prior information* can be seen as (posterior) control of dependencies between training and test data.

2. A method has been presented to construct error (energy) functionals by OR and AND–like combinations of quadratic building blocks.

3. This allows to implement prior information explicitly in terms of the function values itself. Prior information may also be collected in form of *verbal statements* of experts.

4. Non–convex, OR–like combinations yield non–linear stationarity equations and represent a. systems at *finite temperature* (mixture models) or b. systems with *interactions* (Landau–Ginzburg models).