OR: Ambiguous priors

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

$t_1(x)$ and $t_2(x)$ represent two (alternative) mean functions (templates, prototypes).

I. Finite temperature/mixture regularization

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

$d_D^2 = \sum_i (y_i - h(x_i))^2$, $d_i^2 = -\left\langle h - t_i \left| \mathbf{K}_0 \right| h - t_i \right\rangle$

The stationarity equation $h = p_1^M t_1 + p_2^M t_2$

is equivalent to the mean field equation of the ferromagnet

$$\boxed{h = t + \tanh\left[\frac{\beta}{2}\left(E_2 - E_1\right)\right] \frac{t_1 - t_2}{2}}$$

II. Landau–Ginzburg regularization

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

(g strictly monotone) with stationarity equation

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

$K_D(x, x') = \delta(x - x') \sum_i \delta(x - x_i)$

$D(x) = \sum_n \delta(x - x_i) y_i / (\sum_n \delta(x - x_i))$

Mixture model

Mixture model with N_i components $p(h) = \sum_i^{N_i} p(h,i)$ $= \sum_i^{N_i} p_i\, p(h|i)$, with mixing probabilities p_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

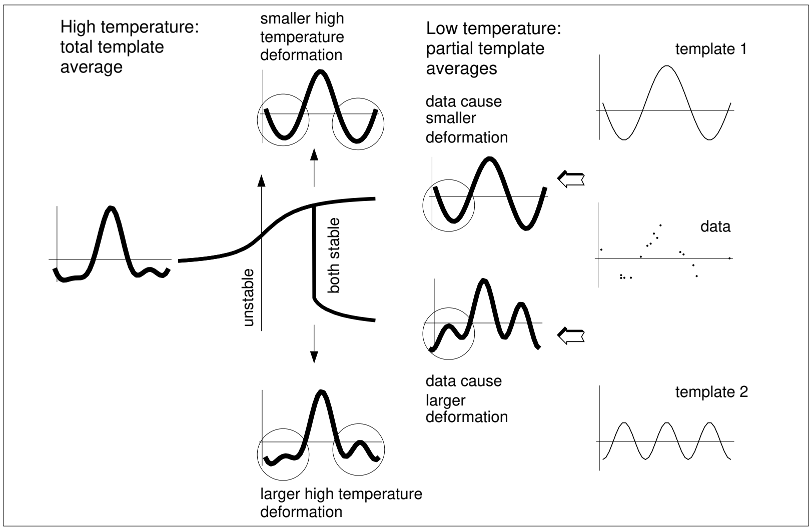
$$\boxed{E^M[h] = -\ln\left(\sum_i^{N_i} e^{-\beta E_i[h] - be_i - \ln Z_i(\beta)}\right),}$$

with $E_i = \frac{1}{2}\sum_{j_i}^{n_{j_i}} d_{ij_i}^2$.

The (nonlinear) stationarity condition becomes

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

with $K^M = \sum_i p(h,i) K_i$, $t^M = \sum_i p(h,i) K_i \bar{t}_i = \sum_i p(h,i)\tilde{t}_i$, with component template averages $\bar{t}_i$. (Finite Gaussian mixture models can also be integrated analytically.)



Linear problems:

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

Iteration procedures = learning

$$\boxed{h^{(k+1)} = h^{(k)} - \eta A^{-1} \nabla E(h^{(k)})}$$

For example:

$$h^{(k+1)} = h^{(k)}k + \eta_k A_k^{-1} (t^{(k)}[h^{(k)}] - K^{M,(k)}[h^{(k)}]h^{(k)})$$

Optimization for example by quasi–Newton or conjugate gradient methods. (But also parallel approaches like genetic algorithms can be used.) Some simple choices are:

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)
Prior–relaxation	$\mathbf{A} = \mathbf{K}$ (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	$\mathbf{A} = \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{\mathrm{M}^2}{2\tilde{\sigma}^2}\right)^k = e^{\frac{\mathrm{M}^2}{2\tilde{\sigma}^2}}$

Hessian, mixture model

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

Hessian, Landau–Ginzburg

$$H^L = \sum_i \sum_{j_i} M_{ij_i} K_{ij_i}$$
$$+ \sum_i \sum_{j_i} \sum_{k_i} M_{ij_ik_i} | K_{ij_i} h - t_{ij_i} \rangle\langle K_{ij_i} h - t_{ij_i} |,$$
$$M_{ij_ik_i} = \prod_{j_i \neq l_i \neq k_i} \beta_F^{n_{j_i}-1} d_{il_i}^2[h]$$

Exact mean of predictive density

Mean of predictive density:

$$\bar{y} = \int dy\, y\, p(y|x, f) = \sum_j b_j\, \bar{t}_j$$

with component mean

$$\bar{t}_j = t_j + \mathbf{K}_j^{-1}\mathbf{K}_{t_D}(j)(t_D - t_j)$$

and mixture coefficients

$$b_j = \frac{p(j)p(y_D|x_D, D_0, j)}{\sum_j p(j)p(y_D|x_D, D_0, j)}$$

and predictive density for component j

$$p(y_D|x_D, D_0, j) =$$

$$e^{-\frac{\beta}{2}\left\langle t_D - t_j \middle| \mathbf{K}_{t_D}(j) \middle| t_D - t_j \right\rangle + \frac{1}{2}\ln\det(\frac{\beta}{2\pi}\mathbf{K}_{t_D}(j))}.$$

Hereby

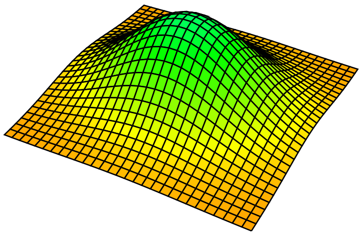
$$\mathbf{K}_{t_D}(j) = (\mathbf{K}_D^{-1} + \mathbf{K}_{j,DD}^{-1})^{-1}$$

is a $\tilde{n} \times \tilde{n}$–matrix in data space and $\mathbf{K}_{j,DD}^{-1}$ the projection of the prior covariance of component j into data space.

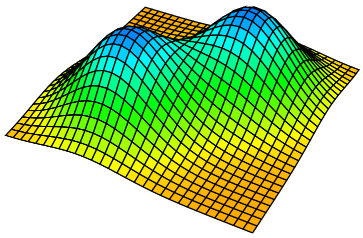
J.L., ICANN99; J.L., physics/9912005; J.L., Uhlig, Weiguny, quant-ph/0005122

Explicit modeling of complex priors with prior mixtures, hyperparameters, and hyperfields

Gaussian process

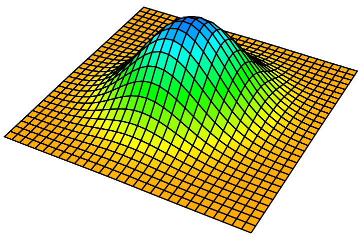


mixture of two Gaussians

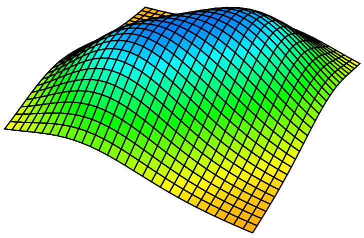


low temperature case

mixture of two Gaussians

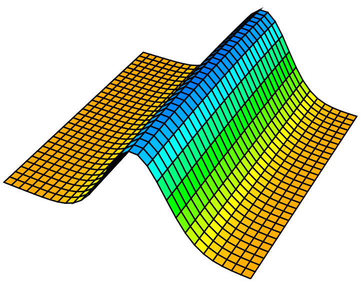


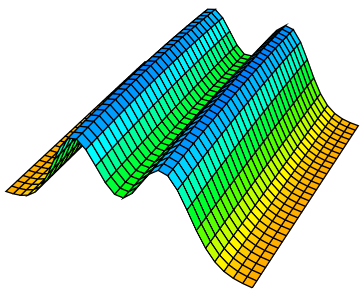
high temperature case



near “critical point”

hyperparameter θ or hyperfiels $\theta(x)$

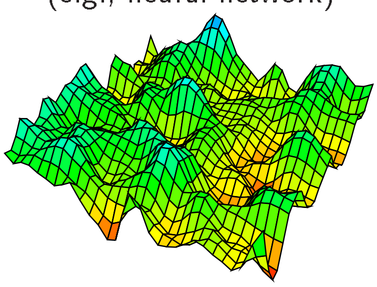




mixture with hyperparameters/–fields

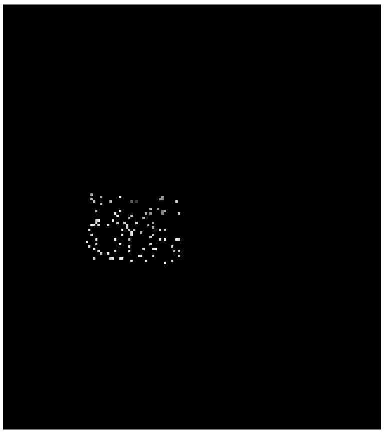
typical error surface of a parametric model

(e.g., neural network)

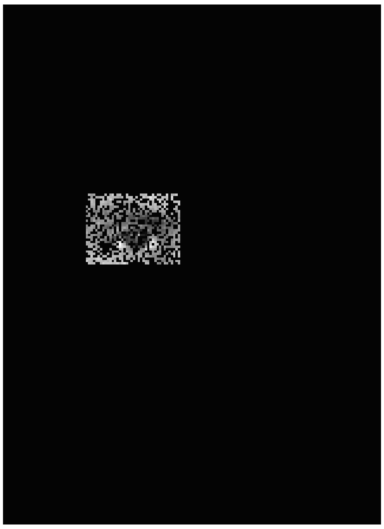


Bayesian image reconstruction

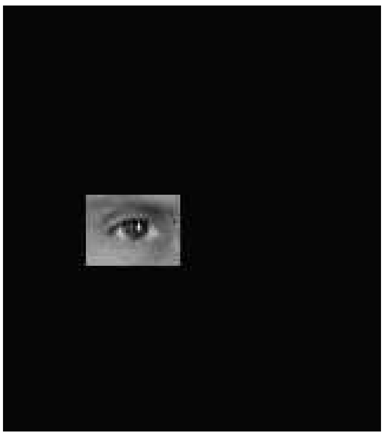
Example data:



100 data points



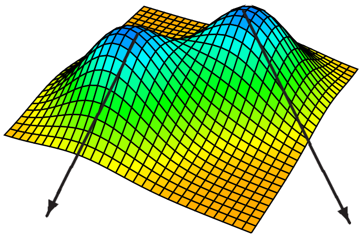
1000 data points



10000 data points

J.L., ICANN99 (physics/9911077) [J.L., EUFIT'98 + FNS'98]

Priors as mixtures of Gaussian prozesses



t_1  or  t_2

$$p(h) \propto \sum_{j=1}^{m} a_j \; e^{-\frac{1}{2}\left\langle h-t_j \middle| \mathbf{K}_j \middle| h-t_j \right\rangle}$$

Expectation of predictive density

$$\bar{y}(x) = \int dy \; y \, p(y|x,D) = \sum_j b_j \, \bar{t}_j$$

with component solution $\bar{t}_j = t_j + \mathbf{K}_j^{-1}\mathbf{K}_{t_D}(j)(t_D - t_j)$ and mixture coefficient

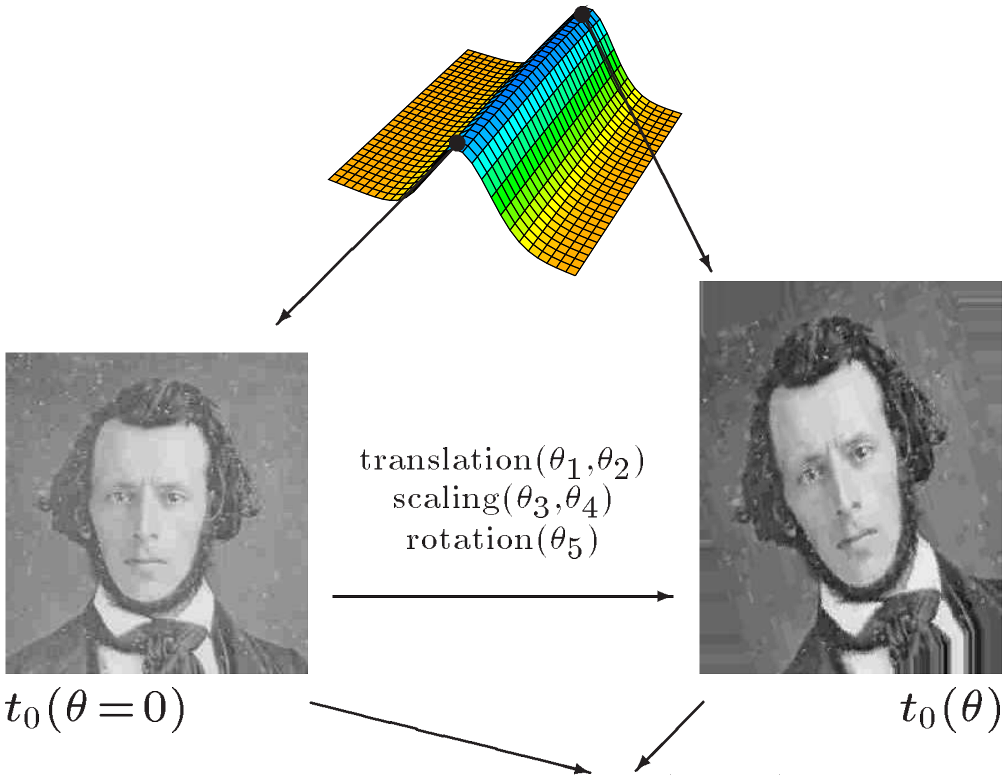
$$b_j = \frac{p(j)p(y_D|x_D,D_0,j)}{\sum_j p(j)p(y_D|x_D,j)},$$

$$p(y_D|x_D,j) = e^{-\frac{\beta}{2}\left\langle t_D-t_j \middle| \mathbf{K}_{t_D}(j) \middle| t_D-t_j \right\rangle + \frac{1}{2}\ln\det(\frac{\beta}{2\pi}\mathbf{K}_{t_D}(j))}.$$

The $\tilde{n} \times \tilde{n}$ matrix $\mathbf{K}_{t_D}(j) = (\mathbf{K}_D^{-1} + \mathbf{K}_{j,DD}^{-1})^{-1}$ is defined in the space of data points, $\mathbf{K}_{j,DD}^{-1}$ being the projection of the prior covariance $\mathbf{K}_j^{-1}$ into that space.

J.L., ICANN99 (physics/9911077) [J.L., EUFIT'98 + FNS'98]

Flexible priors using hyperparameters



$$p(h) \propto \int d\theta \; a(\theta) \; e^{-\frac{1}{2}\left\langle h - t_0(\theta) \middle| \mathbf{K}_0(\theta) \middle| h - t_0(\theta) \right\rangle}$$

Expectation of predictive density (for combination with mixture model)

$$\bar{y}(x) = \int dy \, y \, p(y|x,D) = \sum_j \int d\theta \; b_j(\theta) \, \bar{t}_j(\theta)$$

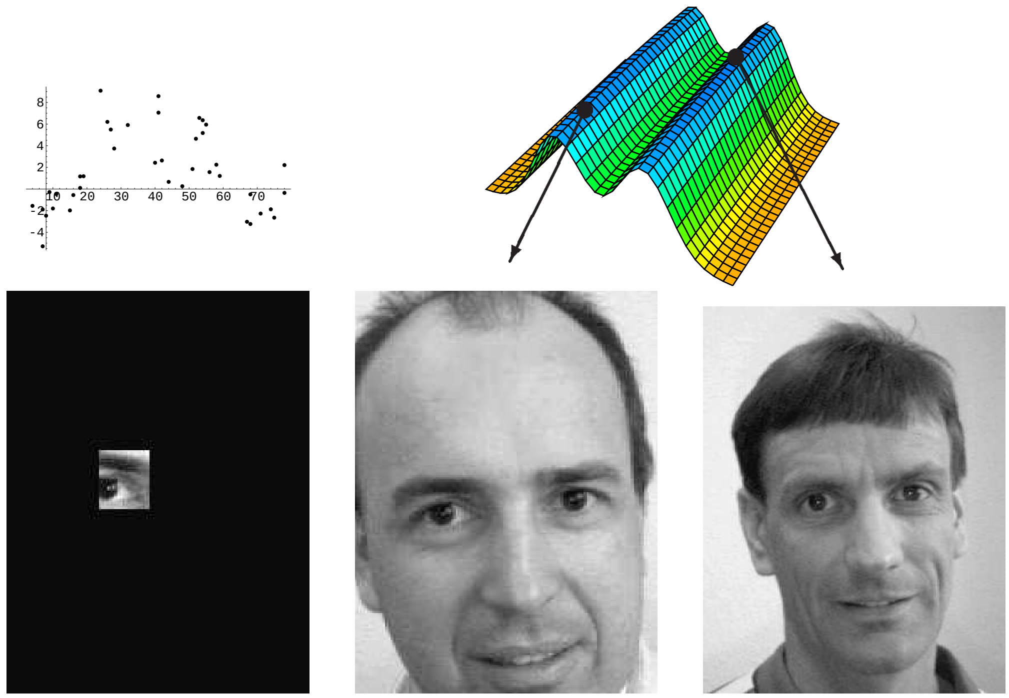
with mixture coefficients

$$b_j(\theta) = \frac{p(\theta,j)\, p(y_D|x_D,\theta,j)}{\sum_j \int d\theta \, p(\theta,j)\, p(y_D|x_D,\theta,j)}$$

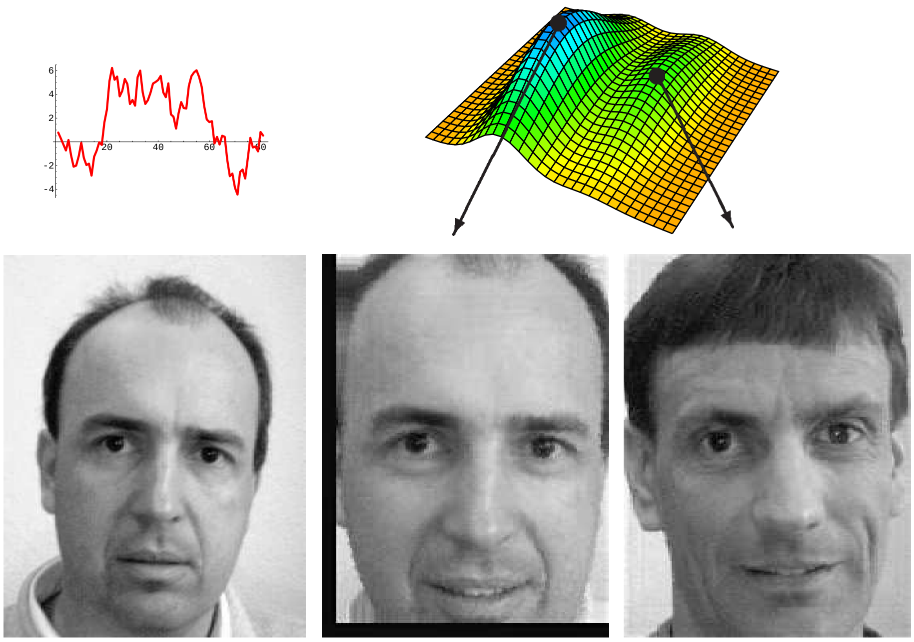
The MAP–Stationarity equation for θ becomes

$$\begin{aligned} 0 \quad = \quad & \sum_j \Bigg[\left\langle \frac{\partial t_j(\theta)}{\partial \theta} \middle| \mathbf{K}_{t_D}(\theta,j) \middle| t_j(\theta) - t_D \right\rangle \\ & + \frac{1}{2} \left\langle t_D - t_j(\theta) \middle| \frac{\partial \mathbf{K}_{t_D}(\theta,j)}{\partial \theta} \middle| t_D - t_j(\theta) \right\rangle \\ & - \mathrm{Tr}\left(\mathbf{K}_{t_D}^{-1}(\theta,j) \frac{\partial \mathbf{K}_{t_D}(\theta,j)}{\partial \theta} \right) - \frac{1}{p(\theta,j)} \frac{\partial p(\theta,j)}{\partial \theta} \Bigg] \end{aligned}$$

Image reconstruction: Low–temperature results

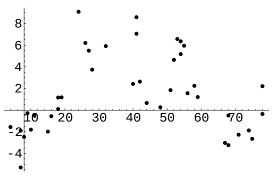


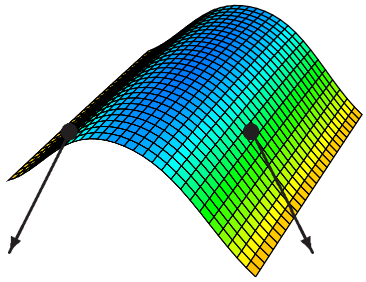
data $\quad t_1 \quad t_2$

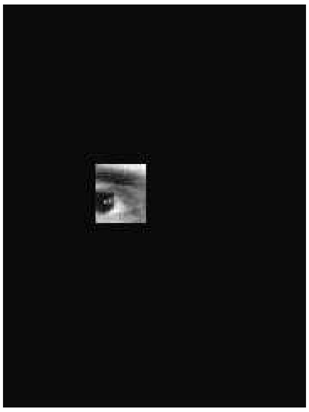


original $\quad$ reconstructed $\quad$ local maximum

Image reconstruction: High–temperature results



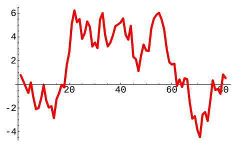


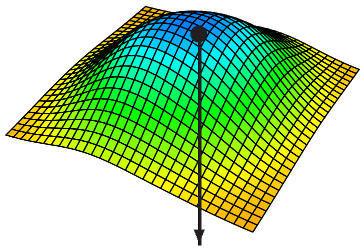


data

t_1

t_2





original

high–temperature reconstruction