

Prediction with Gaussian Processes: Basic Ideas and Theoretical Perspectives

Chris Williams



School of Informatics, University of Edinburgh, UK

Overview

- Gaussian Process Priors over Functions
- GP regression, classification
- Consistency
- Learning curves
- PAC-Bayesian bounds for GP classifiers

Bayesian prediction

- Define a prior over functions
- Observe data, obtain a posterior distribution over functions

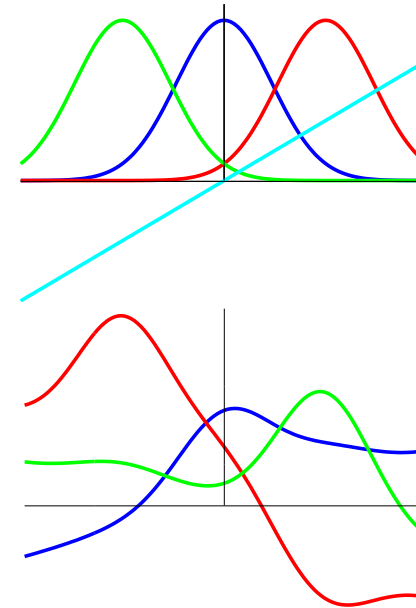
$$P(f|D) \propto P(f)P(D|f)$$

posterior \propto prior \times likelihood

- Make predictions by averaging predictions over the posterior $P(f|D)$
- Averaging mitigates *overfitting*

Bayesian Linear Regression

$$f(\mathbf{x}) = \sum_i w_i \phi_i(\mathbf{x}) \quad \mathbf{w} \sim N(0, \Sigma)$$



Samples from the prior

Gaussian Processes: Priors over functions

- For a stochastic process $f(\mathbf{x})$, mean function is

$$\mu(\mathbf{x}) = E[f(\mathbf{x})].$$

Assume $\mu(\mathbf{x}) \equiv 0 \forall \mathbf{x}$

- Covariance function

$$k(\mathbf{x}, \mathbf{x}') = E[f(\mathbf{x})f(\mathbf{x}')].$$

- Forget those weights! We should be thinking of defining priors over functions, not weights.
- Priors over function-space can be defined directly by choosing a covariance function, e.g.

$$k(\mathbf{x}, \mathbf{x}') = \exp(-w|\mathbf{x} - \mathbf{x}'|)$$

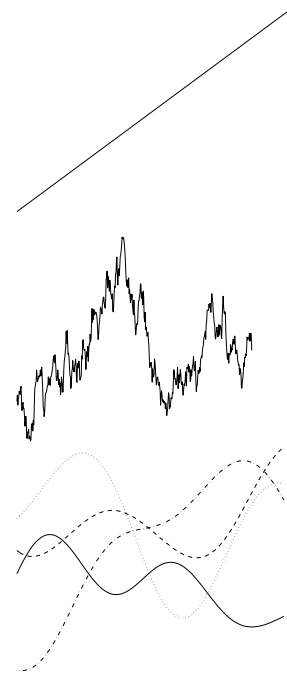
- Gaussian processes are stochastic processes defined by their mean and covariance functions.

Examples of GPs

- $\sigma_0^2 + \sigma_1^2 x x'$

- $\exp -|x - x'|$

- $\exp -(x - x')^2$



Connection to feature space

A Gaussian process prior over functions can be thought of as a Gaussian prior on the coefficients $\mathbf{w} \sim N(0, \Lambda)$ where

$$f(\mathbf{x}) = \sum_{i=1}^{N_F} w_i \phi_i(\mathbf{x}) = \mathbf{w} \cdot \Phi(\mathbf{x})$$

$$\Phi(\mathbf{x}) = \begin{pmatrix} \phi_1(\mathbf{x}) \\ \phi_2(\mathbf{x}) \\ \vdots \\ \phi_{N_F}(\mathbf{x}) \end{pmatrix}$$

In many interesting cases, $N_F = \infty$

Choose $\Phi(\cdot)$ as eigenfunctions of the kernel $k(\mathbf{x}, \mathbf{x}')$ wrt $p(\mathbf{x})$ (Mercer)

$$\int k(\mathbf{x}, \mathbf{y}) p(\mathbf{x}) \phi_i(\mathbf{x}) d\mathbf{x} = \lambda_i \phi_i(\mathbf{y})$$

- Smoother processes have faster decay of eigenvalues
- Modified Bessel (MB) covariance function

$$C_r(h) = \left(\sum_{k=0}^r a_k h^k \right) e^{-|h|} \quad h = |x - x'|/\ell$$

$$S(\omega) \propto \frac{1}{(1 + \omega^2 \ell^2)^r}$$

- MB_r process is $r - 1$ times mean-square differentiable
- $r \rightarrow \infty$ gives Gaussian RBF kernel

Gaussian process regression

Dataset $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^n$, Gaussian likelihood $p(y_i|f_i) \sim N(0, \sigma^2)$

$$\bar{f}(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}, \mathbf{x}_i)$$

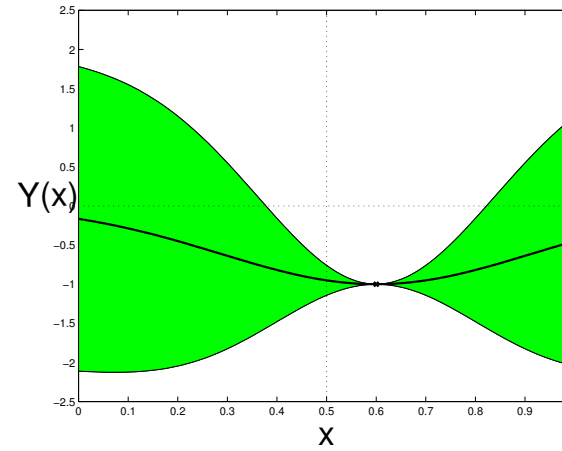
where

$$\boldsymbol{\alpha} = (K + \sigma^2 I)^{-1} \mathbf{y}$$

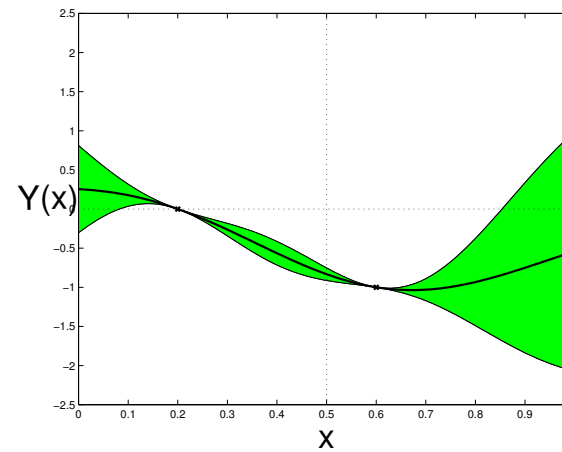
$$\text{var}(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}^T(\mathbf{x})(K + \sigma^2 I)^{-1} \mathbf{k}(\mathbf{x})$$

in time $O(n^3)$, with $\mathbf{k}(\mathbf{x}) = (k(\mathbf{x}, \mathbf{x}_1), \dots, k(\mathbf{x}, \mathbf{x}_n))^T$

After 1 observation:



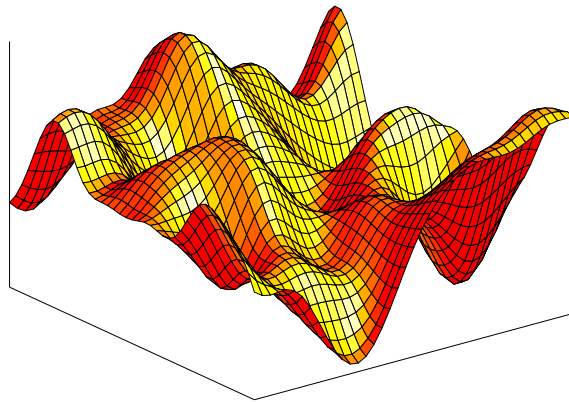
After 2 observations:



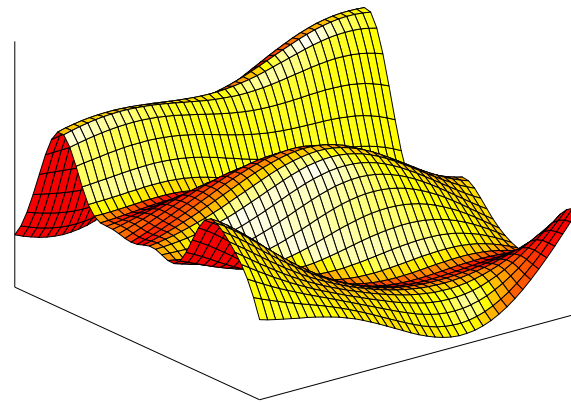
- Approximation methods can reduce $O(n^3)$ to $O(nm^2)$ for $m \ll n$
- GP regression is competitive with other kernel methods (e.g. SVMs)
- Can use non-Gaussian likelihoods (e.g. Student-t)

Adapting kernel parameters

$$k(\mathbf{x}^i, \mathbf{x}^j) = v_0 \exp -\frac{1}{2} \sum_{l=1}^d w_l (x_l^i - x_l^j)^2$$



$w_1 = 5.0$ $w_2 = 5.0$



$w_1 = 5.0$ $w_2 = 0.5$

- For GPs, the marginal likelihood (aka Bayesian evidence) $\log P(\mathbf{y}|\theta)$ can be optimized wrt the kernel parameters $\theta = (v_0, \mathbf{w})$
- For GP regression $\log P(\mathbf{y}|\theta)$ can be computed exactly

$$\log P(\mathbf{y}|\theta) = -\frac{1}{2} \log |K + \sigma^2 I| - \frac{1}{2} \mathbf{y}^T (K + \sigma^2 I)^{-1} \mathbf{y} - \frac{n}{2} \log 2\pi$$

Regularization

- $\bar{f}(\mathbf{x})$ is the (functional) minimum of

$$J[f] = \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2 + \frac{1}{2} \|f\|_{\mathcal{H}}^2$$

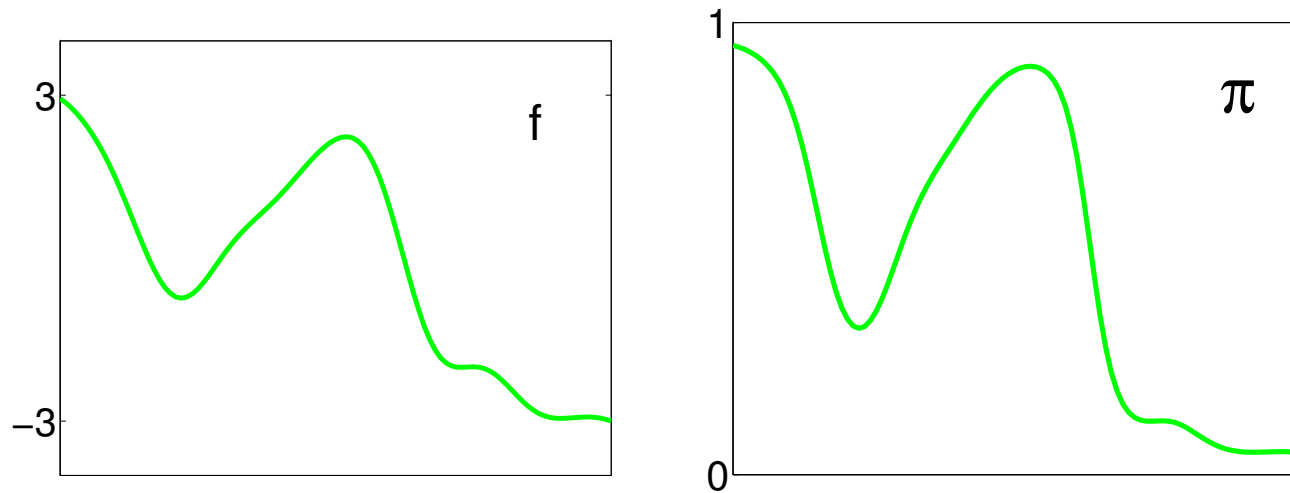
(1st term = $-\log$ -likelihood, 2nd term = $-\log$ -prior)

- However, the regularization framework does not yield predictive variance or marginal likelihood

Previous work

- Wiener-Kolmogorov prediction theory (1940's)
- Splines (Kimeldorf and Wahba, 1971; Wahba 1990)
- ARMA models for time-series
- Kriging in geostatistics (for 2-d or 3-d spaces)
- Regularization networks (Poggio and Girosi, 1989, 1990)
- Design and Analysis of Computer Experiments (Sacks et al, 1989)
- Infinite neural networks (Neal, 1995)

GP prediction for classification problems



Squash through logistic (or erf) function

- Likelihood

$$-\log P(y_i|f_i) = \log(1 + e^{-y_i f_i})$$

- Integrals can't be done analytically

- Find *maximum a posteriori* value of $P(\mathbf{f}|\mathbf{y})$ (Williams and Barber, 1997)
- Expectation-Propagation (Minka, 2001; Opper and Winther, 2000)
- MCMC methods (Neal, 1997)

MAP Gaussian process classification

To obtain the MAP approximation to the GPC solution, we find $\hat{\mathbf{f}}$ that maximizes the convex function

$$\Psi(\mathbf{y}) = -\sum_{i=1}^n \log(1 + e^{-y_i f_i}) - \frac{1}{2} \mathbf{f}^T K^{-1} \mathbf{f} + c$$

The optimization is carried out using the Newton-Raphson iteration

$$\mathbf{f}^{new} = K(I + WK)^{-1}(W\mathbf{f} + (\mathbf{t} - \boldsymbol{\pi}))$$

where $W = \text{diag}(\pi_1(1 - \pi_1), \dots, \pi_n(1 - \pi_n))$ and $\pi_i = \sigma(\hat{f}_i)$. Basic complexity is $O(n^3)$

For a test point \mathbf{x}_* we compute $\bar{f}(\mathbf{x}_*)$ and the variance, and make the prediction as

$$P(\text{class 1} | \mathbf{x}_*, \mathcal{D}) = \int \sigma(f_*) p(f_* | \mathbf{y}) df_*$$

SVMs

1-norm soft margin classifier has the form

$$f(\mathbf{x}) = \sum_{i=1}^n y_i \alpha_i^* k(\mathbf{x}, \mathbf{x}_i) + w_0^*$$

where $y_i \in \{-1, 1\}$ and α^* optimizes the quadratic form

$$Q(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n t_i t_j \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j)$$

subject to the constraints

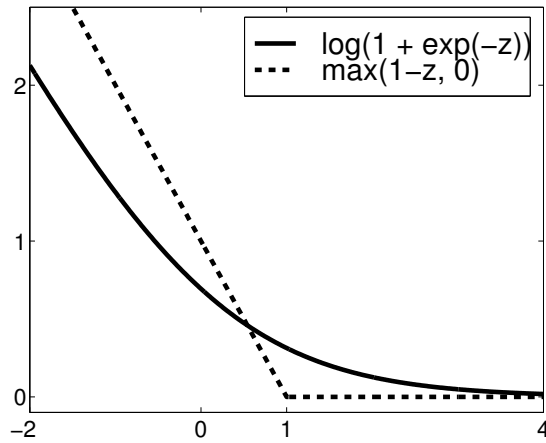
$$\sum_{i=1}^n y_i \alpha_i = 0$$

$$C \geq \alpha_i \geq 0, \quad i = 1, \dots, n$$

This is a *quadratic programming* problem. Can be solved in many ways, e.g. with interior point methods, or special purpose algorithms such as SMO.

Basic complexity is $O(n^3)$.

- Define $g_\sigma(z) = \log(1 + e^{-z})$
- SVM classifier is similar to GP classifier, but with g_σ replaced by $g_{SVM}(z) = [1 - z]_+$ (Wahba, 1999)



- Note that the MAP solution using g_σ solution is not sparse, but gives a probability output

Consistency

- Risk

$$R_L(f) = \int L(y, f(\mathbf{x}))d\mu(\mathbf{x}, y)$$

- $\eta(\mathbf{x})$ minimizes $R_L(f)$

- A procedure that returns $f_{\mathcal{D}}$ is consistent for a given measure $\mu(\mathbf{x}, y)$ and loss function L if

$$R_L(f_{\mathcal{D}}) \rightarrow R_L(\eta) \quad \text{as } n \rightarrow \infty$$

- If the procedure is consistent for all Borel probability measures $\mu(\mathbf{x}, y)$ it is said to be universally consistent

Proving consistency

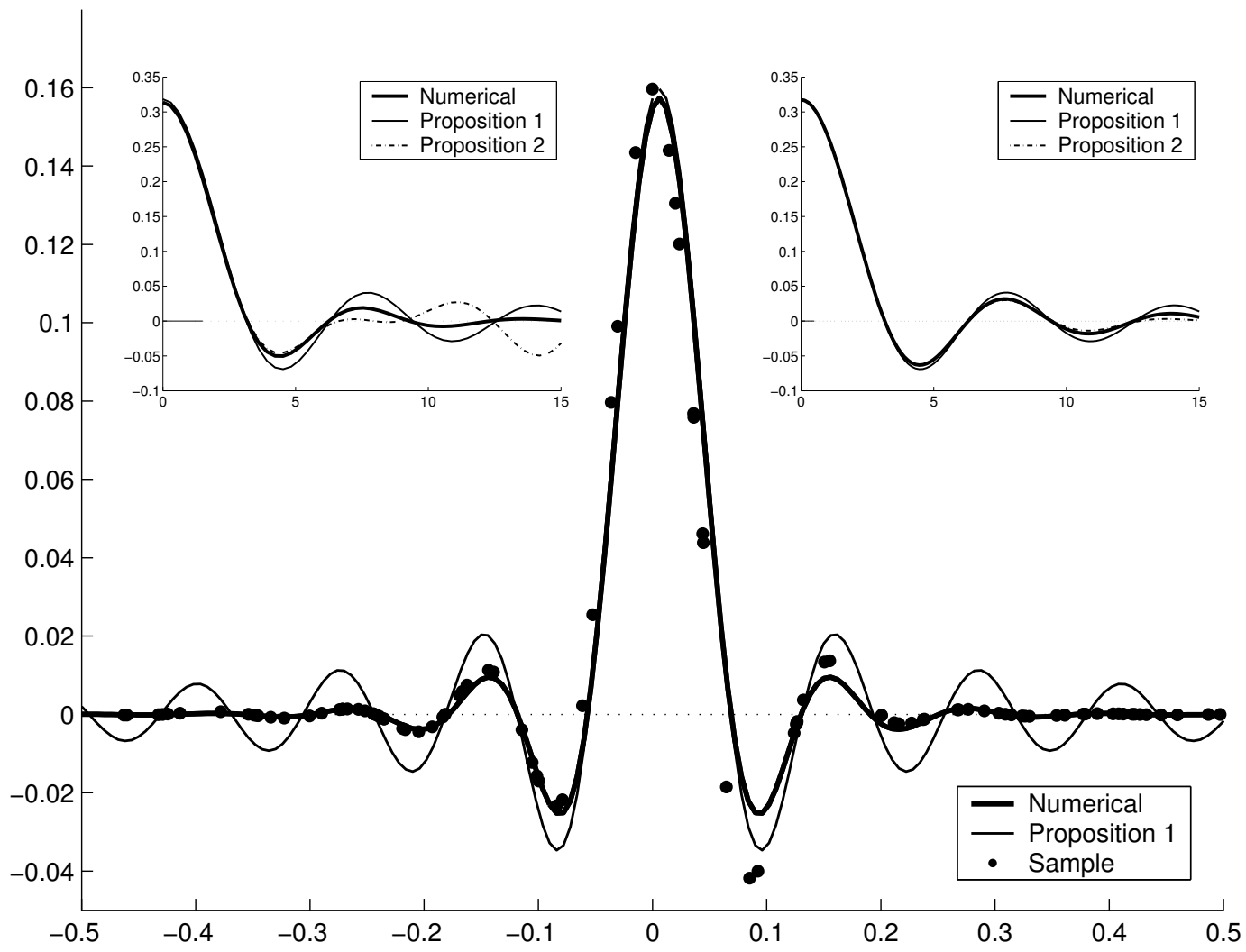
- Kernel smoother with bandwidth h : $h \rightarrow 0$ and $nh^D \rightarrow \infty$.
Equivalent kernel analysis for GP regression
- Regularized risk approaches
- KL divergence view (?)

Equivalent kernel

$$\bar{f}(\mathbf{x}_*) = \mathbf{h}^\top(\mathbf{x}_*)\mathbf{y}, \quad \mathbf{h}(\mathbf{x}_*) = (K + \sigma^2 I)^{-1} \mathbf{k}(\mathbf{x}_*)$$

- $\mathbf{h}(\mathbf{x}_*)$ is known as the weight function. It is hard to understand as we need the matrix inverse of $K + \sigma^2 I$, and K depends on location of training inputs
- Idealize: consider observations as “smeared out” across input space, ρ data points per unit length/area/volume, hence replace $J[f]$ by

$$J_\rho[f] = \frac{\rho}{2\sigma^2} \int (\eta(\mathbf{x}) - f(\mathbf{x}))^2 d\mathbf{x} + \frac{1}{2} \|f\|_{\mathcal{H}}^2$$



- For stationary kernels we can use Fourier analysis to obtain minimum of $J_\rho(f)$ as

$$f(\mathbf{x}_*) = \int h(\mathbf{x}_* - \mathbf{x})\eta(\mathbf{x})d\mathbf{x}$$

where

$$\tilde{h}(\mathbf{s}) = \frac{S(\mathbf{s})}{S(\mathbf{s}) + \sigma^2/\rho} = \frac{1}{1 + \sigma^2/(\rho S(\mathbf{s}))}$$

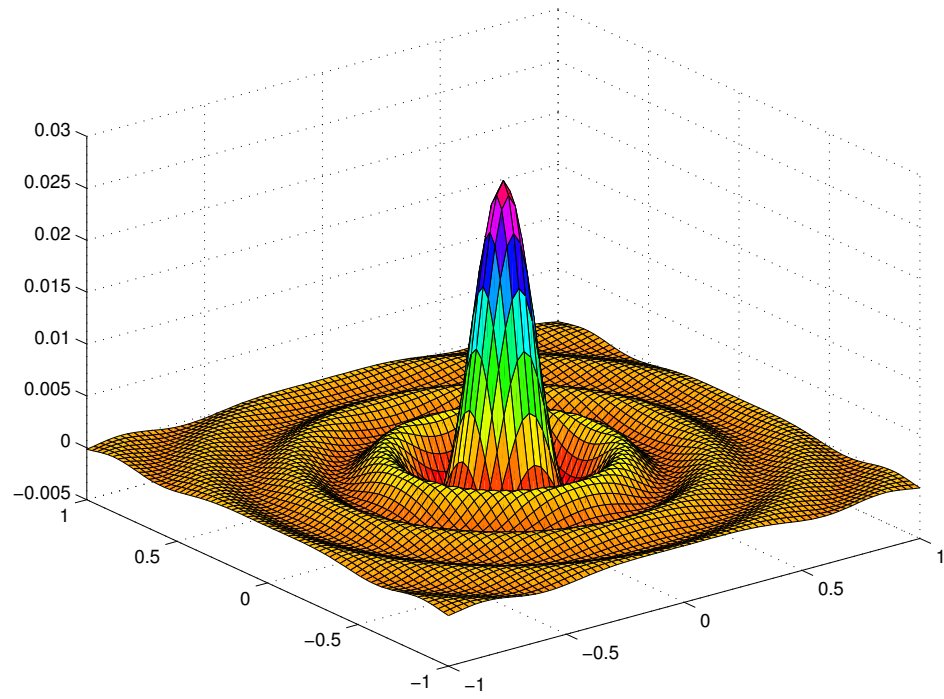
- $h(\mathbf{x})$ is known exactly only for a few kernels, e.g. some splines and the Ornstein-Uhlenbeck process.

- For RBF kernel $k(r) = \exp(-r^2/2\ell^2)$, ($r^2 \equiv |\mathbf{x} - \mathbf{x}'|^2$) we can obtain an approximate result

$$h(r) = (s_c/r)^{D/2} J_{D/2}(2\pi s_c r)$$

where $J_\nu(z)$ is a Bessel function and $\exp(2\pi^2\ell^2 s_c^2) = \rho\sigma^{-2}(2\pi\ell^2)^{D/2}$ (Sollich and Williams, 2004)

- bandwidth of EK $\sim (\log(\rho))^{-1/2}$, so very slow decay with ρ



Regularized Risk

Consider minimization of

$$J_\lambda[f] = \frac{\lambda_n}{2} \|f\|_{\mathcal{H}}^2 + \frac{1}{n} \sum_{i=1}^n L(y_i, f(\mathbf{x}_i))$$

(note $\lambda_n = 1/n$ above), and its smoothed version

$$J_{\lambda,n}[f] = \frac{\lambda_n}{2} \|f\|_{\mathcal{H}}^2 + \int L(y, f) d\mu(\mathbf{x}, y)$$

$$|R_L(f_{\mathcal{D}}) - R_L(\eta)| \leq |R_L(f_{\mathcal{D}}) - R_L(f_n)| + |R_L(f_n) - R_L(\eta)|$$

approximation and estimation errors

Some results

- Cox (1984). Spline regularizer $\|f\|_m^2 = \sum_{k=0}^m \|O^k f\|^2$ and that the sampling becomes becomes uniformly dense in $\Omega \subset \mathbb{R}^D$, then for $p < m$ and $m > 3d/2$

$$\mathbb{E}\|f_{\mathcal{D}} - \eta\|_p^2 = O(\lambda_n^{(m-p)/m} + n^{-1} \lambda_n^{-(2p+D)/2m})$$

Best rate is obtained for $\lambda_n = Cn^{-2m/(2m+D)}$ giving

$$\mathbb{E}\|f_{\mathcal{D}} - \eta\|_p^2 = O(n^{-2(m-p)/(2m+D)})$$

(Note, rate for λ_n is slower than $1/n$)

- Further results by O'Sullivan and Cox (1990) on splines, generalized to regression, classification, density estimation etc

- Zhang (2004): Classification with logistic loss, non-degenerate RKHS, $\mu(\mathbf{x}, y)$ sufficiently regular and $\lambda_n \rightarrow 0$ but $n\lambda_n \rightarrow \infty$: convergence to Bayes optimal loss as $n \rightarrow \infty$. (Note, rate for λ_n is slower than $1/n$)
- Steinwart (2003) Similar results with various decay rates on λ_n depending on the smoothness of the kernel
- Question: can one obtain consistency results for GPR, GPC with $\lambda_n = 1/n$?

KL divergence view

- The mean of a finite-dimensional distribution on \mathbb{R}^D can be found by minimizing the KL divergence between $p(\mathbf{x})$ and $N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ wrt $\boldsymbol{\mu}$ for *any* fixed positive-definite $\boldsymbol{\Sigma}$
- Does the same argument hold for the infinite dimensional case?

Generalization Error and Learning Curves

- Draw a target function from a GP. Choose n locations and add output noise to create a dataset $\mathcal{D} = (X, \mathbf{y})$.
- Use a prediction method to produce an estimate $f_{\mathcal{D}}$ of f ; measure *generalization error* as

$$E_{\mathcal{D}}^g(f) = \int L(f(\mathbf{x}_*), f_{\mathcal{D}}(\mathbf{x}_*)) d\mu(\mathbf{x}_*)$$

- Average over the choice of f to give

$$E^g(X) = \int E_{\mathcal{D}}^g(f) dP(f)$$

(this is not too hard for GP priors and regression with squared loss)

$$\begin{aligned}
E^g(X) &= \int \mathbb{E}[(f(\mathbf{x}_*) - \mathbf{k}_1^\top(\mathbf{x}_*)K_{1,y}^{-1}\mathbf{y})^2]d\mu(\mathbf{x}_*) \\
&= \int k_0(\mathbf{x}_*, \mathbf{x}_*)d\mu(\mathbf{x}_*) - 2\text{tr} \left(K_{1,y}^{-1} \int \mathbf{k}_0(\mathbf{x}_*)\mathbf{k}_1^\top(\mathbf{x}_*)d\mu(\mathbf{x}_*) \right) \\
&\quad + \text{tr} \left(K_{1,y}^{-1}K_{0,y}K_{1,y}^{-1} \int \mathbf{k}_1(\mathbf{x}_*)\mathbf{k}_1^\top(\mathbf{x}_*)d\mu(\mathbf{x}_*) \right).
\end{aligned}$$

- Average over X to obtain

$$E^g(n) = \int E^g(X)d\mu(\mathbf{x}_1) \dots d\mu(\mathbf{x}_n)$$

- $E^g(n)$ can rarely be computed exactly; bounds and approximations are needed
- Instead of averaging over X it may be of interest to optimize: this is *optimal experimental design*

- For squared loss $E^g(X)$ can be readily expressed in terms matrices etc, but averaging over X is usually hard (except for Markovian processes on the line)
- Alternative is to work in eigenbasis, with $k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^N \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{x}')$. Let Λ be the spectrum. Then (Oppor and Vivarelli, 1999)

$$E^g(n) \geq \sigma^2 \sum_{i=1}^N \frac{\lambda_i}{\sigma^2 + n\lambda_i}$$

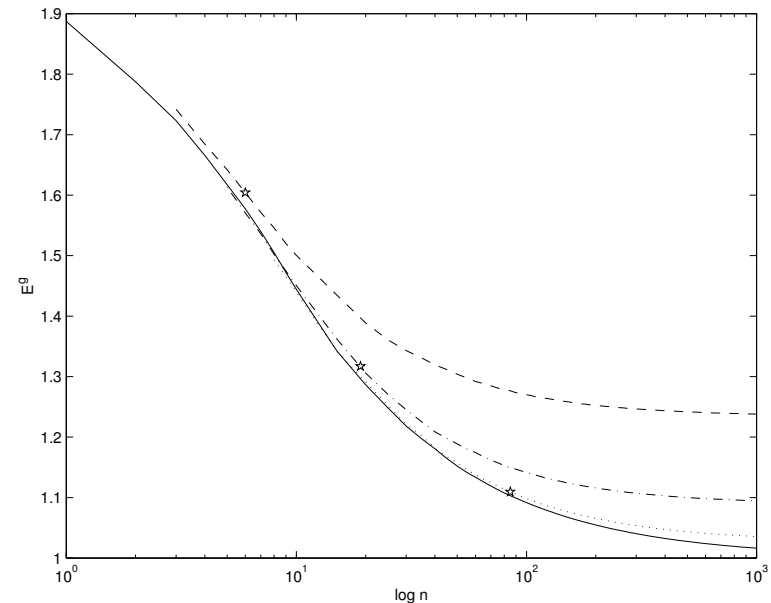
- Notice that we add σ^2/n to the bound for each eigenvalue where $\lambda_i \gg \sigma^2/n$. Hence the decay rate of $E^g(n)$ is slower than $1/n$ as more eigenfunctions “come in to play” as n increases.

- Ritter (2000) gives asymptotic results $O((n^{-(2r+1)/(2r+2)})$ for covariance functions obeying Sacks-Ylvisaker conditions of order r in 1-d with optimal sampling. Sollich (2002) obtains the same result for random designs. [S-Y conditions of order $r \Rightarrow \lambda_i \propto i^{-(2r+2)}$ asymptotically].
- Sollich (2002) has analyzed the mismatched case, and the case where the spectrum of the covariance function can be adapted on the basis of training data
- Malzahn and Opper (2002) have considered more general techniques for learning curves, e.g. for GP classification techniques

Finite dimensional approximation

$$\tilde{y}(x) = \sum_{i=1}^M w_i \phi_i(x)$$

- Can choose M as a function of n ; we expect that predictions will be close up to $n \sim \frac{\sigma_v^2}{\lambda_M}$
- Work with Gianni Ferrari Treccate and Manfred Opper



PAC-Bayesian bounds for GP classifiers

M. Seeger (2001)

- We have a prior distribution $P(\mathbf{w})$ and a posterior distribution $Q(\mathbf{w}|S)$
- McAllester's PAC-Bayesian Theorem:

$$Pr_S\{gen(S, Q) - emp(S, Q) \geq \epsilon(\delta, S, K, Q)\} < \delta$$

where

$$\text{emp}(S, Q) = \frac{1}{n} \sum_{i=1}^n \Pr_{y_i \sim Q(y_i | x_i^S, S)} \{\text{sgn } y_i \neq t_i^S\}$$

$$\text{gen}(S, Q) = E_{(x_*, t_*)} [\Pr_{y_* \sim Q(y_* | x_*, S)} \{\text{sgn } y_* \neq t_*\}]$$

$$\epsilon(\delta, S, K, Q) = \sqrt{\frac{D(Q \| P) + \log \delta^{-1} + \log n + 2}{2n - 1}}$$

- For a GP classifier with Laplace approximation, $D(Q||P)$ can be computed in $O(n^3)$ (extension to infinite dimensional feature spaces)
- Experiments: MNIST data, classify 2s vs 3s, RBF kernel
- For $n = 5000$, obtain non-trivial result
 $emp = 0.0204$, $gen = 0.021$, $bound = 0.1433$ for $\delta = 0.05$. Even tighter for relative entropy Chernoff bound

Conclusion

- GPs are a competitive practical method, but also are simple enough to give rise to nice theoretical problems