Generalized RBF n/w

The 1-1 correspondence of a training sample \( x_i \) and an associated Green's function \( G(x, x_i) \) can pose many problems.

1) It can be prohibitively expensive \( n/w \) when \( N \) becomes large.

2) For determining the unknown weights from the hidden to the output layer, we need to solve a linear matrix equation, requiring matrix inverse operations \( \sim O(N^3) \).
**Require:** We need an approx. of the regularized soln.

**Idea:** Approximate the regularized soln. in a lower dim. space using a sub-optimal soln.

\[ F^*_s(x) = \sum_{i=1}^{m_1} w_i p_i(x) \]
$\{ y_i(x) \}_{i=1}^{m_1}$ is a new set of basis functions

$y_i(x) = G \left( \| x - t_i \| \right); \quad i = 1, \ldots, m_1$

@ center $t_i = x_i$

$f^*(x) = \sum_{i=1}^{m_1} w_i \cdot G \left( \| x - t_i \| \right)$

where $t_i$'s have to be determined
Let us formulate the functional

$$\mathcal{E}(F^*) = \sum_{i=1}^{N} \left( \sum_{j=1}^{m_i} w_{ji} G \left( \| x_i - t_j \| \right) \right)^2 + \alpha \| D F^* \|$$

Can be expanded as

$$\left\| d - G \omega \right\|$$
where \( d = \begin{bmatrix} d_1 & d_2 & \ldots & d_N \end{bmatrix}^T \)

\( w = \begin{bmatrix} w_1 & \ldots & w_{m_1} \end{bmatrix}^T \)

Observe this detail

\[
G_1 = \begin{bmatrix}
G_1(x_1, t_1) & \ldots & G_N(x_1, t_1) \\
\vdots & \ddots & \vdots \\
G_1(x_N, t_1) & \ldots & G_N(x_N, t_1)
\end{bmatrix}_{N \times m_1}
\]
\( \mathbf{G} \) is of size \( N \times m_1 \), (rectangular matrix)

Evaluating \( \| \mathbf{D} \mathbf{F}^* \|^2 = \left( \mathbf{D} \mathbf{F}^* \right)^\top \mathbf{D} \mathbf{F}^* \) \( \mathcal{H} \)

\[
\begin{align*}
\mathbf{D} \mathbf{F}^* & = \sum_{i=1}^{m_1} w_i \mathbf{G}_i(x, b_i), \\
& = \mathbf{W}^\top \mathbf{G}_0 \mathbf{W} \\
& = \mathbf{W}^\top \mathbf{D} \mathbf{F}^* \mathbf{W} \\
& \text{(Regularization constraint)}
\end{align*}
\]
\[ G_0 = \begin{bmatrix} G(t_1, t_1) & \cdots & G(t_1, t_m) \\ \vdots & \ddots & \vdots \\ G(t_m, t_1) & \cdots & G(t_m, t_m) \end{bmatrix} \quad \text{(Square!)} \]

Homework: The minimization of \( \mathbb{E}(F^*) \) w.r.t. \( w \)

\[
(G^T G + \lambda G_0) w = G^T d
\]
If $\lambda \to 0$ (i.e., no regularization)

$$w = \sqrt{(G^T G)^{-1}} G^T d$$

(min. norm soln / pseudo inverse soln to the least squares problem when $m_1 < N$)
Weighted norm of data point
\[ \left\| \overset{(m_0 \times 1)}{X} \right\|_C^2 = (C^{T} CX)^T CX = \overset{m_0 \times m_0 \text{ norm}}{x} \]

\[ \text{max}_{i} \sum_{i=1}^{m_1} w_i \ G \left( \left\| x - \hat{t}_i \right\|_C \right) \]
For the Gaussian case,

\[ G_1\left(\|x - t_i\|_c\right) = \exp\left(-\frac{(x - t_i)^\top S^{-1} (x - t_i)}{2}\right) \]

\[ S^{-1} = C^\top C \]
XOR Problem Revisited

We will consider RBF n/w as a special case of the Green’s n/w.

Consider the pair of Gaussian functions

$$G_i (\|x - t_i\|) = \exp \left(- \|x - t_i\|^2 \right)$$

$i = 1, 2$

Let us choose centers @ $t_1$ and $t_2$

$$t_1 = [1 \ 1]^T \quad t_2 = [0 \ 0]^T$$

At the moment, forget optimization over $t_1, t_2$
\[ y(x) = \sum_{i=1}^{2} w_i \cdot G_1(||x - t_i||) + b \]

\[ y(x, j) = d_j \quad j = 1, 2, 3, 4; \forall \ z = 1, 2 \]

\[ x_j \in \mathbb{R}^2 \]

<table>
<thead>
<tr>
<th>( x_j )</th>
<th>( d_j )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 1)</td>
<td>0</td>
</tr>
<tr>
<td>(0, 1)</td>
<td>1</td>
</tr>
<tr>
<td>(0, 0)</td>
<td>0</td>
</tr>
<tr>
<td>(1, 0)</td>
<td>1</td>
</tr>
</tbody>
</table>

Solution to the XOR problem
\[ G = \begin{bmatrix}
1 & 0.1353 & 1 \\
0.367 & 0.3678 & 1 \\
0.1253 & 1 & 1 \\
0.3678 & 0.3678 & 1 \\
\end{bmatrix} \quad 4 \times 3 \]

\[ d = \begin{bmatrix}
0 \\
1 \\
0 \\
1 \\
\end{bmatrix}^{T} \quad 4 \times 1 \]

\[ w = \begin{bmatrix}
\omega \\
w \\
b \\
\end{bmatrix}^{T} \quad 3 \times 1 \]

4 data points \quad 2 centers \quad t_1, t_2
Solve: 

\[ w = \left( G_1^T G_1 \right)^{-1} G_1^T d \]

(without regularization)

Plugging \( G_1, d \) into \( w \) from numerical computations done earlier:

\[
\begin{bmatrix}
  -2.5 \\
  -2.5 \\
  +2.84
\end{bmatrix}
\]
Structural risk minimization

Suppose we have the non-linear regression model

\[ d = f(x) + \varepsilon \]

\( f(\cdot) \) is unknown

Consider the ensemble-averaged cost

\[ J_{\text{act}}(\mathcal{F}) = \mathbb{E}_{x,d} \left( \frac{1}{2} (d - f(x))^2 \right) \]

\( \mathbb{E}_{x,d} \) joint expectation

\[ \hat{f}^* = \mathbb{E}(d | x) \] minimizes \( J_{\text{act}}(\mathcal{F}) \)
\(^*\) requires the knowledge of joint pdf of \(x\) and \(d\) 

Suppose we bring in a neural network and make a first approximation

\[ f(x) = F(x; W) \]

\[ J(W) = E_{x, d} \left[ \frac{1}{2} (d - F(x; W))^2 \right] \]

Let \( \hat{W}^* = \arg \min_W J(W) \)
\[ J(\hat{\mathbf{w}}^*) > J_{act}(\hat{\mathbf{x}}^*) \] 

This is the 1st level of approximation.

Consider the time averaged energy function

\[ \mathcal{E}_{av}(N; \mathbf{w}) = \frac{1}{2N} \sum_{i=1}^{N} (d(i) - F(x_c(i); \hat{\mathbf{w}}))^2 \]

The minimizer of \( \mathcal{E}_{av}(N; \mathbf{w}) \) is

\[ \hat{\mathbf{w}}_N = \arg\min_{\mathbf{w}} \mathcal{E}_{av}(N; \mathbf{w}) \]
\[
J(\hat{\omega}_N) \geq J(\hat{\omega}^*) \geq J_{\text{act}}(\hat{f}^*)
\]

\text{time averaged cost opt.}

Excess error:

\[
\frac{J(\hat{\omega}_N) - J_{\text{act}}(\hat{f}^*)}{J(\hat{\omega}_N) - J(\hat{\omega}^*) + J(\hat{\omega}^*) - J_{\text{act}}(\hat{f}^*)}
\]

\text{depends on the size of the data N}

\text{conditional mean over } \mathbb{E} C(.)
For e.g., for a single hidden layer MLP, the capacity of the learning machine is governed by the size of the hidden layer.

Consider a family of nested approximating functions

\[ F_k = \left\{ F(x; w) \mid w \in \mathbb{R}^k \right\} \]

such that \( F_1 \subseteq F_2 \subseteq \cdots \subseteq F_k \)

\( F_k \) is a measure of the machine capacity.
1) Before opt. is reached, the machine capacity is too small for the details within the data.

2) After opt. is reached, the machine capacity is too large for the details within the data.
Bias - Variance Dilemma

Consider the functional approximation problem.

We have a data set $D$ and an associated mapping $f : X \rightarrow Y$. 'f' is unknown here!

We need to get a good estimate of 'f' from the data set $D$. Get $g_D(x)$ close to $f$ in some sense.

Also, typically, one can have several data sets in the learning example. Given different sets $\{D_i\}$, one can arrive at various estimates of 'f'.

$y = f(x)$, $f$ scalar in label associations
$\rightarrow$ vector
An example shall help us visualize.

Suppose I have 3 data points from a parabola above.

If I need to fit a line, i.e., \( y = mx + c \) form.

Given \( D_1 \), I can get \( L_1 \).

\( D_2 \), \( L_2 \).

\( \vdots \).

Qn: What if I just want to approximate the parabola by just a scalar? say, c.
Intuition

Space of functions having just one function \( g \)

\[ g \sim f \]

\[ \Rightarrow \text{There is bias } g(x) - f(x) \]

But no variance since we have just one function

Subset of functions \( s \)

We have a pool of functions \( \{ g_i \}_{i=1}^N \)

\( \{ g_i(x) \} \) agree with \( f \) on the training data sets \( \{ D_i \} \)

The \( \langle \{ g_i \} \rangle \) is \( \ll \) (much less)

\[ \Rightarrow \text{Bias is } \ll \text{ but the Variance is more since we have } > 1 \text{ function} \]
Having seen that there is 'bias' and 'variance' in the error averaged over the data sets corresponding to the choice in the pool of functions available, this gives rise to a trade off in the bias & variance given the generalization problem

⇒ Bias - Variance dilemma
Role of Bias/variance

Suppose we have a higher model complexity (due to fitting noisy samples)

Bias is less but variance is more

$\overline{g}(x) = \frac{1}{N} \sum_{i=1}^{N} g_i(x)$ (Sample Mean)

$\text{bias}(x) = \overline{g}(x) - f(x)$
$\text{var}(x) = E_{D|x}[L(g_D(x), \overline{g}(x))^2]$
Let us work out the analysis

\[ y = f(x) + \epsilon \]

\( \epsilon \) is a random variable, with mean 0, and variance \( \sigma^2 \) and statistically independent of \( f \) and approximating function \( g_D(x) \) for a data set \( D \)

We are interested in

\[ E_{D,x} \left[ (g_D(x) - y)^2 \right] \]
Plugging (A) in (B)

$$E_{D, X} \left[ (g_D(x) - f(x) - \varepsilon)^2 \right]$$

Exchanging expectations, assuming if can be done

$$E_X \left[ E_{D|x} \left[ (g_D(x) - f(x) - \varepsilon)^2 \right] \right]$$

Let us expand the terms
\begin{align*}
&E_x \left[ E_{D|x} \left( g_D^2(x) + f^2(x) + \varepsilon^2 - 2 g_D(x) f(x) \right) \right] \\
&\text{Since expectation is linear, we shall evaluate some of terms towards simplification} \\
&E_{D|x} \left[ f^2(x) \right] = f^2(x) \\
&E_{D|x} \left[ \varepsilon^2 \right] = \sigma_x^2 \\
&E_{D|x} \left[ g_D(x) \varepsilon \right] = 0 \quad (\text{Statistically independent})
\end{align*}
\[ E_{D|x}[f(x) \varepsilon] = 0 \quad \text{(Same as earlier)} \]

Define \( E_{D|x}[g_D(x)] \triangleq \bar{g}(x) \)

Let us simplify I

\[ E_x \left[ E_{D|x}(g_D^2(x)) - \bar{g}^2(x) + \bar{g}^2(x) + f^2(x) - 2 \bar{g}(x)f(x) + \sigma_x^2 \right] \]

\[ E_x \left[ \bar{g}^2(x) - 2 \bar{g}(x)f(x) + f^2(x) \right] = E_x \left[ \bar{g}(x) - f(x) \right]^2 \text{biase}(x) \]
- Variance + bias + \sigma^2 \quad E_{x,d}(\varepsilon^2) = \sigma^2
Estimation of regularization parameter

Consider the N.L. reg. problem

\[ d_i = f(x_i) + \varepsilon_i \quad i = 1, \ldots, N \]

\( f(.\) is unknown

\( \varepsilon_i \) is drawn from a zero mean white process

with \( E(\varepsilon_i \varepsilon_k) = \begin{cases} \sigma^2 & i = k \\ 0 & \text{else} \end{cases} \)
Goal: Recover \( f(x_i) \) given \( \{ (x_i, d_i) \}_{i=1}^N \).

Let \( \tilde{f}_\lambda(x) \) be the regularized estimate of \( f(x) \) for some regularization parameter \( \lambda \).

\[
\mathcal{E}(\tilde{f}) = \frac{1}{2} \sum_{i=1}^N (d_i - \tilde{f}(x_i))^2 + \frac{\lambda}{2} \| DF(x) \|^2
\]

Tikhonov functional

Fidelity to data

Smoothness constraint
Averaged Square error

Let \( R(\lambda) \) denote the averaged square error over a given data between \( f(\lambda) \) pertaining to the model and the approximating function \( f_2(\lambda) \) pertaining to the representation of the solution for some \( \lambda \) over the training data.
\[ R(\lambda) = \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - F_\lambda(x_i))^2 \]

\[ F_\lambda(x_k) = \sum_{i=1}^{N} a_{k,i}(\lambda) d_i \]  

(Linear combination)

Observe the detail pertaining to the data point \( x_k \)
\[ F_{\alpha} = \begin{bmatrix} F_{\alpha}(x_1) & \cdots & F_{\alpha}(x_N) \end{bmatrix}^T \]

\[ A(\alpha) = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix} \]

\[ d = \begin{bmatrix} d_1 & \cdots & d_N \end{bmatrix}^T \]

\[ R(\alpha) = \frac{1}{N} \left\| \frac{1}{N} f - F_{\alpha} \right\|^2 \]

("\|\cdot\|" is \(L^2\) norm)
\[ f = \begin{bmatrix} f(x_1) & \cdots & f(x_n) \end{bmatrix}^T \]

Simplifying,

\[ R(x) = \frac{1}{n} \left\| f - A(x) d \right\|^2 \]

\[ d = f + \xi \]

\[ \xi = \begin{bmatrix} \xi_1 & \cdots & \xi_n \end{bmatrix}^T \]

Plug (2) in (1)

Plug (2) in (1)
\[ R(\lambda) = \frac{1}{N} \left\| \mathbf{f} - A(\lambda) (\mathbf{f} + \varepsilon) \right\|^2 \]

\[ = \frac{1}{N} \left\| \mathbf{f} - A(\lambda) \mathbf{f} - A(\lambda) \varepsilon \right\|^2 \]

\[ = \frac{1}{N} \left\| (I - A(\lambda)) \mathbf{f} - A(\lambda) \varepsilon \right\|^2 \]

Let us expand (3)
\[ R(\lambda) = \frac{1}{N} \left\| (I - A(\lambda))^\frac{1}{2} \right\|^2 \]

\[ - \frac{2}{N} \mathbb{E}^T A(\lambda) (I - A(\lambda))^\frac{1}{2} \]

\[ + \frac{1}{N} \mathbb{E}^T A(\lambda)^T \mathbb{E} \]

We need \( \mathbb{E}(R(\lambda)) \) (\( \mathbb{E}(\text{Middle Term}) = 0 \))

\[ \mathbb{E} \left( \frac{1}{N} \left\| (I - A(\lambda))^\frac{1}{2} \right\|^2 \right) = \frac{1}{N} \left\| (I - A(\lambda))^\frac{1}{2} \right\|^2 \]
Consider

\[ E \left( \left\| A(\alpha) \varepsilon \right\|^2 \right) \]

\[ = E \left[ \varepsilon^T A(\alpha)^T A(\alpha) \varepsilon \right] \]

\[ = \text{tr} \left[ E \left( \varepsilon^T A(\alpha)^T A(\alpha) \varepsilon \right) \right] \]

\[ = E \left[ \text{tr} \left( \varepsilon^T A^T(\alpha) A(\alpha) \varepsilon \right) \right] \]

\[ \text{(\because \text{tr (Scalar)} = \text{Scalar})} \]

\[ \text{(\because \text{exchanging \ E (\cdot) \ with \ tr (\cdot)})} \]
\[
E \left[ \operatorname{tr} \left( A^T(\lambda) A(\lambda) \varepsilon \varepsilon^T \right) \right] = \operatorname{tr} (AB) = \operatorname{tr} (BA)
\]

\[
= \operatorname{tr} (A^T(\lambda) A(\lambda)) E \left[ \varepsilon \varepsilon^T \right]
\]

\[
= \sigma^2 \operatorname{tr} (A^T(\lambda) A(\lambda))
\]

\[
E \left( \| A(\lambda) \varepsilon \|_2^2 \right) = \sigma^2 \operatorname{tr} (A^T(\lambda) A(\lambda))
\]
\[ E(R(\lambda)) = \frac{1}{N} \left\| (I - A(\lambda)f \right\|^2 \\
+ \frac{\sigma^2}{N} \text{tr} \left\| A^T(\lambda) A(\lambda) \right\| \]

But we still have a problem!

\[ E(R(\lambda)) \text{ is still a fn of } f(\cdot) \text{ which is unknown!} \]
A reasonable estimate of $\hat{R}(\lambda)$ is given by

$$
\hat{R}(\lambda) = \frac{1}{N} \left\| (I - A(\lambda)) d \right\|^2
+ \frac{\sigma^2}{N} \text{tr} \left( A^2(\lambda) \right)
+ \frac{\sigma^2}{N} \text{tr} \left( (I - A(\lambda))^2 \right)
$$

depends on $\lambda$