Reproducing Kernel Hilbert Space

Consider a Mercer kernel $K(\underline{x}, \cdot)$ where $\underline{x} \in \mathcal{H}$ and $\mathcal{F}$ be any vector space of all real valued functions of $\underline{x}$ generated by $K(\underline{x}, \cdot)$

Suppose we pick two functions $f(\cdot)$ and $g(\cdot)$ from $\mathcal{F}$ (I.P. space)

$$f(\underset{\uparrow \underline{x}}{\cdot}) = \sum_{i=1}^{\ell} a_i K(\underline{x}_i, \underset{\uparrow \underline{x}}{\cdot})$$

for all

$$\underline{x}_i, \underline{\tilde{x}}_j \in \mathcal{H}$$

IIIly $\underline{x}$

$$g(\cdot) = \sum_{j=1}^{n} b_j K(\underline{\tilde{x}}_j, \cdot)$$

Consider the bilinear form

Defn

$$\langle f, g \rangle = \sum_{i=1}^{\ell} \sum_{j=1}^{n} a_i b_j K(\underline{x}_i, \tilde{\underline{x}}_j)$$

$$= \underline{a}^T K \underline{b}$$

Gram matrix / Kernel matrix

$$\langle K(\underline{x}_i, \cdot), K(\underline{x}_j, \cdot) \rangle = K(\underline{x}_i, \underline{x}_j)$$

One element of the Gram matrix

We can rewrite $\langle f, g \rangle$ as

$$\langle f, g \rangle = \sum_{i=1}^{\ell} a_i \underbrace{\sum_{j=1}^{n} b_j K(\underline{x}_i, \tilde{\underline{x}}_j)}_{g(\underline{x}_i)}$$

$$\left(\because K(\underline{x}_i, \tilde{\underline{x}}_j) = K(\tilde{\underline{x}}_j, \underline{x}_i) \right)$$

$$= \sum_{i=1}^{\ell} a_i \, g(\underline{x}_i)$$

$|||^{ly}$ $\langle f, g \rangle = \sum_{j=1}^{n} b_j f(\tilde{\underline{x}}_j)$

Properties

1) Symmetry: For all fns f and $g \in \mathcal{F}$ the term $\langle f, g \rangle$ is symmetric

i.e., $\langle f, g \rangle = \langle g, f \rangle$

2) Scaling and distribution

For any pair of constants c and d and any set of functions f, g and $h \in \mathcal{F}$

$$\langle (cf + dg), h \rangle = c \langle f, h \rangle + d \langle g, h \rangle$$

3) Squared norm

For any real valued fn $f \in \mathcal{F}$

$$\|f\|^2 = \langle f, f \rangle$$

$$= \underline{a}^T K \underline{a}$$

(non negative definite)

$$\|f\|^2 \geq 0$$

4) Reproducing Kernel Property

Suppose $g(\cdot) = K(\underline{x}, \cdot)$

$$\langle f, K(\underline{x}, \cdot) \rangle = \sum_{i=1}^{\ell} a_i \, K(\underline{x}, \underline{x}_i)$$

$(\because$ Symmetry$)$

$K(\underline{x}, \underline{x}_i) = K(\underline{x}_i, \underline{x})$

$$= \sum_{i=1}^{\ell} a_i \, K(\underline{x}_i, \underline{x})$$

$$= f(\underline{x})$$

Mercer Kernel reproduces $f(\cdot)$ (Reproducing Kernel)

1) For every $\underline{x}_i \in \mathcal{X}$, $K(\underline{x}, \underline{x}_i)$ as a function of $\underline{x}$ $\in \mathcal{F}$

2) Satisfies reproducing property

Mercer kernel $\Longrightarrow$ Reproducing kernel

Reproducing kernel space
Complete $\Longrightarrow$ Reproducing kernel Hilbert space

Representer Theorem

Theorem: Any function defined in a RKHS can be represented as a linear combination of Mercer kernel functions.

Proof: Define a space $\mathcal{H}$ to represent RKHS induced by a Mercer kernel $K(\underline{x}, \cdot)$.
Given any real valued fn $f(\cdot) \in \mathcal{H}$, we could decompose $f(\cdot)$ into 2 components lying in $\mathcal{H}$.

The first component $f_{\parallel}(\cdot)$ is contained in the span of the kernel fns $K(\underline{x}_1, \cdot), K(\underline{x}_2, \cdot) \ldots$

$$f_{\parallel}(\cdot) = \sum_{i=1}^{\ell} a_i K(\underline{x}_i, \cdot) \quad \text{——— (1)}$$

The second component is orthogonal to the span of the kernel fns; $f_{\perp}(\cdot)$

$$f(\cdot) = f_{\parallel}(\cdot) + f_{\perp}(\cdot) \quad \text{——— (2)}$$

$$f(\cdot) = \sum_{i=1}^{\ell} a_i K(\underline{x}_i, \cdot) + f_\perp(\cdot) \quad \text{--- (3)}$$

From the reproducing property

$$f(\underline{x}_j) = \langle f(\cdot), K(\underline{x}_j, \cdot) \rangle_{\mathcal{H}} \quad \text{--- (4)}$$

Plug in (3) into (4)

$$f(\underline{x}_j) = \left\langle \left[\sum_{i=1}^{\ell} a_i K(\underline{x}_i, \cdot) + f_\perp(\cdot) \right] K(\underline{x}_j, \cdot) \right\rangle$$

$$f(\underline{x}_j) = \left\langle \sum_{i=1}^{\ell} a_i K(\underline{x}_i, \cdot), K(\underline{x}_j, \cdot) \right\rangle + \underbrace{\left\langle f_\perp(\cdot), K(\underline{x}_j, \cdot) \right\rangle}_{0}$$

$$= \sum_{i=1}^{\ell} a_i K(\underline{x}_i, \underline{x}_j)$$

$$\left(\because K(\underline{x}_i, \underline{x}_j) = \langle K(\underline{x}_i, \cdot), K(\underline{x}_j, \cdot) \rangle \right.$$

Mercer Kernel functions

Generalized Applicability

Theorem: $f(\underline{x}_j) = \sum_{i=1}^{\ell} a_i K(\underline{x}_i, \underline{x}_j)$ is the minimizer of the regularized empirical risk

$$\mathcal{E}(f) = \underbrace{\frac{1}{2N}\sum_{i=1}^{N}\left(d(n) - f(x(n))\right)^2}_{\text{Std error}} + \underbrace{\Omega\left(\|f\|_{\mathcal{H}}\right)}_{\text{regularizing fn.}}$$

($\Omega(\cdot)$ must be a non decreasing fn.)

$f(\cdot)$ unknown

$(x(n), d(n))$ Data pairs, $n = 1, \ldots, N$

Proof :

Step 1 : Let $f_\perp$ denote the orthogonal complement to the span of the kernel fns $\{K(x_i, \cdot)\}_{i=1}^{\ell}$

Now, every fn can be expressed as a kernel expansion on the training + $f_\perp$

$$\Omega\left(\|f\|_{\mathcal{H}}\right) = \Omega\left(\left\|\sum_{i=1}^{\ell} a_i K(x_i, \cdot) + f_\perp(\cdot)\right\|_{\mathcal{H}}\right)$$

Introduce

$$\tilde{\Omega}\left(\|f\|_{\mathcal{H}}^2\right) = \Omega\left(\|f\|_{\mathcal{H}}\right)$$

$$\tilde{\Omega}\left(\|f\|_{\mathcal{H}}^2\right) = \tilde{\Omega}\left(\left\|\sum_{i=1}^{\ell} a_i k(\underline{x}_i, \cdot) + f_\perp(\cdot)\right\|_{\mathcal{H}}^2\right)$$

Step 2 : Apply Pythagorus theorem

$$\tilde{\Omega}\left(\|f\|^2_{\mathcal{H}}\right) = \tilde{\Omega}\left(\left\|\sum_{i=1}^{\ell} a_i K(\underline{x}_i,\cdot)\right\|^2_{\mathcal{H}} + \|f_\perp(\cdot)\|^2_{\mathcal{H}}\right)$$

Set $f_\perp(\cdot) = 0$ for optimality

$$\geq \tilde{\Omega}\left(\left\|\sum_{i=1}^{\ell} a_i K(\underline{x},\cdot)\right\|^2_{\mathcal{H}}\right)$$

$$\tilde{\Omega}\left(\|f\|^2_{\mathcal{H}}\right) = \tilde{\Omega}\left(\left\|\sum_{i=1}^{\ell} a_i K(\underline{x},\cdot)\right\|^2_{\mathcal{H}}\right)$$

Step 3 : In light of monotonicity

$$\Omega\left(\|f\|_{\mathcal{H}}\right) = \Omega\left(\left\|\sum_{i=1}^{\ell} a_i\, k(\underline{x}_i, \cdot)\right\|\right)$$

$\Longrightarrow$ For fixed $a_i \in \mathbb{R}$, the representer theorem is also a minimizer of the regularizing fn $\Omega(\|f\|_{\mathcal{H}})$ provided monotonicity is satisfied!

MOTIVATION TO REGULARIZATION THEORY

Often, in machine learning problems, we encounter situations where problems are not well-posed.

For example, when the # of data points in the training samples >> # of degrees of freedom, the problem is over determined.

One may fit misleading variations in the data!

Learning is a sort of multi-D mapping (f), and can be viewed as a problem of hypersurface reconstruction given a set of sparse points

Now, given X (domain) and Y (range) that are metric spaces, related by a fixed but unknown mapping

$$f : X \rightarrow Y$$

The problem of reconstructing f is well-posed if it satisfies the following:

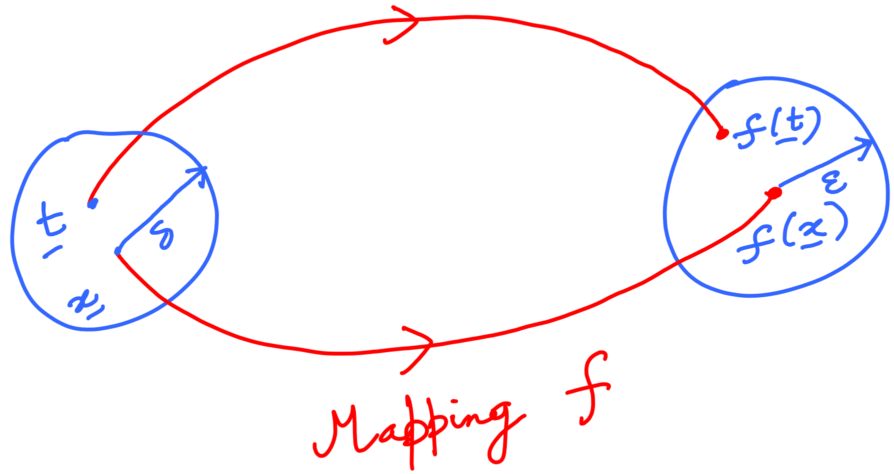
a) Existence: For every input vector $\underline{x} \in X$, $\exists$ a $\underline{y} = f(\underline{x})$, $\underline{y} \in Y$

b) Uniqueness: For any pair of input vectors $\underline{x}, \underline{t} \in X$

$f(\underline{x}) = f(\underline{t})$ iff $\underline{x} = \underline{t}$

c) Continuity: For any $\varepsilon > 0$, $\exists$ $\delta = \delta(\varepsilon)$ /

$d(\underline{x}, \underline{t}) < \delta \Rightarrow d(f(\underline{x}), f(\underline{t})) < \varepsilon$



How can one make an ill-posed problem, well-posed?

<u>SOLN</u> : Regularization (Tikhonov)

Consider the following problem

Input signal : $\underline{x}_i \in \mathbb{R}^{m_0}$ $\quad i = 1, \dots, N$

Desired signal : $d_i \in \mathbb{R}$ $\quad i = 1, \dots, N$

Data

$$\{\underline{x}_i, d_i\}_{i=1}^{N}$$

Let the approximating function be $F(\underline{x})$

$$\mathcal{E}_s(F) = \frac{1}{2} \sum_{i=1}^{N} \left(d_i - F(\underline{x}_i)\right)^2$$

(Approximation error)

Introduce the regularization term that depends on the geometry of the problem

$$\mathcal{E}_c^{(Reg)}(F) = \frac{1}{2} \|DF\|^2$$

Linear differential operator D

Choice of 'D' is problem dependent!

$\| \cdot \|$ is the norm over which the function space belongs to.

$$\mathcal{E}(F) = \mathcal{E}_s(F) + \lambda\, \mathcal{E}_c(F)$$

$$= \underbrace{\frac{1}{2}\sum_{i=1}^{N}\left(d_i - F(\underline{x}_i)\right)^2}_{\text{approx. error}} + \frac{1}{2}\lambda \underbrace{\|DF\|^2}_{\text{regularization term}}$$

$\mathcal{E}(F)$ is also called the "Tikhonov functional"

$\lambda \to 0$; unconstrained

$\lambda \to \infty$; x_i's are unreliable

Choose λ in between $(0, \infty)$

Normalize / $\lambda \in (0, 1)$

i.e., a fraction

Now $\quad F_\lambda(x) = \min\limits_{\lambda,\, \underline{w}} \mathcal{E}(F)$ (min. Tikhonov functional)

parameter in $F(\cdot)$

Consider the standard error term differential

$$d\mathcal{E}_s(F, h) = \left[\frac{d}{d\beta}\mathcal{E}_s(F + \beta h)\right]_{\beta=0}$$

$h(\underline{x})$ is a fixed function of 'x'

$$d\left(\mathcal{E}(F,h)\right) = d\left(\mathcal{E}_s(F,h)\right) + \lambda\, d\left(\mathcal{E}_c(F,h)\right) = 0$$

$$d\left(\mathcal{E}_s(F,h)\right) = \frac{1}{2}\frac{d}{d\beta}\sum_{i=1}^{N}\left[d_i - F(\underline{x}_i) - \beta h(\underline{x}_i)\right]^2$$

$$= -\sum_{i=1}^{N}\left[d_i - F(\underline{x}_i) - \beta h(\underline{x}_i)\right] h(\underline{x}_i)\Bigg|_{\beta=0}$$

$$= -\sum_{i=1}^{N}\left(d_i - F(\underline{x}_i)\right) h(\underline{x}_i)$$

$$= -\left\langle h, \left(\underline{d} - F(\underline{x})\right)\delta(\underline{x}-\underline{x}_i)\right\rangle$$

III) By doing it over the regularization term

$$d(\mathcal{E}_c(F,h)) = \frac{d}{d\beta}\,\mathcal{E}_c(F+\beta h)\Big|_{\beta=0}$$

$$= \frac{1}{2}\frac{d}{d\beta}\int_{\mathbb{R}^{m_0}} \big(D(F+\beta h)\big)^2 \, d\underline{x}\,\Big|_{\beta=0}$$

$$= \int_{\mathbb{R}^{m_0}} D(F+\beta h)\cdot Dh \; d\underline{x}\,\Big|_{\beta=0}$$

$$= \int_{\mathbb{R}^{m_0}} DF\cdot Dh\; d\underline{x} \; = \langle DF, Dh\rangle_{\mathcal{H}}$$

Euler-Lagrange equation

Given a linear differential operator D, we can find a uniquely determined adjoint operator by $\tilde{D}$ / for any pair of functions $u(\underline{x})$ and $v(\underline{x})$ that are sufficiently differentiable (upto a certain degree) & satisfy proper boundary conditions

$$\int_{\mathbb{R}^m} u(\underline{x})\, D\, v(\underline{x})\, d\underline{x} = \int_{\mathbb{R}^m} v(\underline{x})\, \tilde{D}\, u(\underline{x})\, d\underline{x}$$

D is a matrix.

With $u(\underline{x}) \triangleq DF(\underline{x})$

and $\underline{v}(\underline{x}) \triangleq h(\underline{x})$

$$d\mathcal{E}_c(F, h) = \int_{\mathbb{R}^m} \underbrace{h(\underline{x})}_{v(\underline{x})} \tilde{D} \underbrace{DF(\underline{x})}_{u(\underline{x})} \, d\underline{x}$$

$$= \left\langle h(\underline{x}), \tilde{D} D F \right\rangle_{\mathcal{H}}$$

Interpret this as an inner product

With the inclusion of a regularization parameter,

$$d\mathcal{E}(F, h) = \left\langle h, \left[\overbrace{\tilde{D}DF}^{\text{regulatory}} - \frac{1}{\lambda} \sum_{i=1}^{N} (d_i - F)\, \delta_{x_i} \right] \right\rangle_{\mathcal{H}}$$

Fréchet differential

desired value

approximating fn

$$\lambda \in (0, \infty)$$

$d\mathcal{E}(F, h)$ is zero for every $h(\underline{x})$ in $\mathcal{H}$ space

$$\text{iff} \quad \tilde{D}DF - \frac{1}{\lambda} \sum_{i=1}^{N} (d_i - F)\, \delta_{x_i} = 0$$

i.e.,
$$\tilde{D}DF_{\lambda}(\underline{x}) = \frac{1}{\lambda} \sum_{i=1}^{N} \left(d_i - F(\underline{x}_i)\right) \delta(\underline{x} - \underline{x}_i) \qquad (A)$$

Green's Function

Eq^n (A) represents a partial differential eq^n in the approximating function F, whose solution is of interest.

Let $G(\underline{x}, \underline{\xi})$ be a function of $\underline{x}$ and $\underline{\xi}$.

(Green's function)

Some argument

satisfying certain properties.

For a given linear differential operator L, $G(\underline{x}, \underline{\xi})$ satisfies the following properties: (Courant & Hilbert)

1) For a fixed $\underline{\xi}$, $G(\underline{x}, \underline{\xi})$ is a function of $\underline{x}$ satisfying the boundary conditions

2) Except @ $\underline{x} = \underline{\xi}$, the derivatives of $G(\underline{x}, \underline{\xi})$ w.r.t $\underline{x}$ are all continuous; the # of derivatives is determined by L

3) $L\, G(\underline{x}, \underline{\xi}) = 0$ everywhere except @ $\underline{x} = \underline{\xi}$, where it is singular.

$$L G(\underline{x}, \underline{\xi}) = \delta(\underline{x} - \underline{\xi}) \text{ @ } \underline{x} = \underline{\xi} \text{ exists}$$

The function $G(\underline{x}, \underline{\xi})$ is called the Green's function of operator L.

(Similar to the inverse of a matrix eqn!)

Let $\varphi(\underline{x})$ be a continuous / piecewise continuous function of $\underline{x} \in \mathbb{R}^m$, then

Claim: $F(\underline{x}) = \int_{\mathbb{R}^m} G(\underline{x}, \underline{\xi})\, \varphi(\underline{\xi})\, d\underline{\xi}$ is a solution to $L\, F(\underline{x}) = \varphi(\underline{x})$

Let us verify the validity!

$$LF(\underline{x}) = L\underbrace{\int_{\mathbb{R}^m} G(\underline{x},\underline{\xi})\,\varphi(\underline{\xi})\,d\underline{\xi}}_{F(\underline{x})}$$

$$= \int_{\mathbb{R}^m} L\,G(\underline{x},\underline{\xi})\,\varphi(\underline{\xi})\,d\underline{\xi}$$

From the property of Green's functions.

$$= \int_{\mathbb{R}^m} \delta(\underline{x}-\underline{\xi})\,\varphi(\underline{\xi})\,d\underline{\xi} = \varphi(\underline{x})$$

$$LF(\underline{x}) = \varphi(\underline{x})$$

Let us look into the regularization problem

$$L = \tilde{D} D$$

$$\varphi(\underline{\xi}) = \frac{1}{\lambda} \sum_{i=1}^{N} \left(d_i - F(\underline{x}_i)\right) \delta(\underline{x}_i - \underline{\xi})$$

$\downarrow$ Plug in

$$F_\lambda(\underline{x}) = \int_{\mathbb{R}^m} G(\underline{x}, \underline{\xi}) \, \varphi(\underline{\xi}) \, d\underline{\xi}$$

$$F_\lambda(\underline{x}) = \int_{\mathbb{R}^m} G(\underline{x}, \underline{\xi}) \left\{ \frac{1}{\lambda} \sum_{i=1}^{N} [d_i - F(\underline{x}_i)]\, \delta(\underline{x}_i - \underline{\xi}) \right\} d\underline{\xi}$$

$$= \frac{1}{\lambda} \sum_{i=1}^{N} [d_i - F(\underline{x}_i)] \underbrace{\int_{\mathbb{R}^m} G(\underline{x}, \underline{\xi})\, \delta(\underline{x}_i - \underline{\xi})\, d\underline{\xi}}_{G(\underline{x}, \underline{x}_i)}$$

$$F_\lambda(\underline{x}) = \frac{1}{\lambda} \sum_{i=1}^{N} (d_i - F(\underline{x}_i))\, G(\underline{x}, \underline{x}_i)$$

The minimizing function to the regularization problem is a linear superposition of N- green functions.

The points $\underline{x}_i$ represent the centers of the expansion and $\left(d_i - F(\underline{x}_i)\right)/\lambda$ represent the weights of the expansion

$\left\{ G(\underline{x}, \underline{x}_i) \right\}_{i=1}^{N}$ centered @ $\underline{x} = \underline{x}_i$ constitute the basis of a subspace of smooth fn. where the soln to the regularization problem lies

How do we determine the Coeffts (w_i) ?

Let $w_i \triangleq \frac{1}{\lambda}\left[d_i - F(\underline{x}_i)\right]$; $i = 1, \ldots, N$

continuous

$$F_\lambda(\underline{x}) = \sum_{i=1}^{N} w_i \, G(\underline{x}, \underline{x}_i) \quad \text{—— (1)}$$

(# of Green's functions = # of data points)

Evaluate (1) @ $\underline{x}_j$; $j = 1, \ldots, N$

data points

Matrix / Vector Notations

Let $\underline{F_\lambda} \triangleq [F_\lambda(\underline{x}_1) \quad \cdots \quad F_\lambda(\underline{x}_N)]^T$

$\underline{d} \triangleq [d_1 \quad \cdots \quad d_N]^T$; $\underline{w} = [w_1 \quad \cdots \quad w_N]^T$

$$G \triangleq \begin{bmatrix} G(\underline{x}_1, \underline{x}_1) & \cdots & G(\underline{x}_1, \underline{x}_N) \\ \vdots & \ddots & \\ G(\underline{x}_N, \underline{x}_1) & & G(\underline{x}_N, \underline{x}_N) \end{bmatrix}_{N \times N}$$

(Gram matrix)

Writing in matrix form,

$$\underline{w} = \frac{1}{\lambda}\left[\underline{d} - \underline{F}_\lambda\right] \implies \underline{F}_\lambda = \underline{d} - \lambda \underline{w}$$

$$G := \left[G(\underline{x}_i, \underline{x}_j) \right]$$

$$\underline{F}_\lambda = G \underline{w}$$

$$\therefore \left(G + \lambda I \right) \underline{w} = \underline{d}$$

But, the adjoint of the linear differential operator L

$\tilde{L} = L \implies$ Green's fns are symmetric!

$$G(\underline{x}_i, \underline{x}_j) = G(\underline{x}_j, \underline{x}_i)$$

However, all functions in the null space of D are invisible to the smoothing term / regulatory constraint $\|DF\|^2$ and is problem dependent.

The RBF happens to be a special case of Green's function that is translationally and rotationally invariant

i.e., if $G(\underline{x}, \underline{x}_i) = G(\|\underline{x} - \underline{x}_i\|)$

For RBF,

$$F_\lambda(\underline{x}) = \sum_{i=1}^{N} w_i \, G(\|\underline{x} - \underline{x}_i\|)$$

(Linear function space) (depends on data!)

Assuming Gaussian units,

$$F_\lambda(\underline{x}) = \sum_{i=1}^{N} w_i \exp\left(-\frac{1}{2\sigma_i^2} \left\| \underline{x} - \underline{x}_i \right\|^2\right)$$

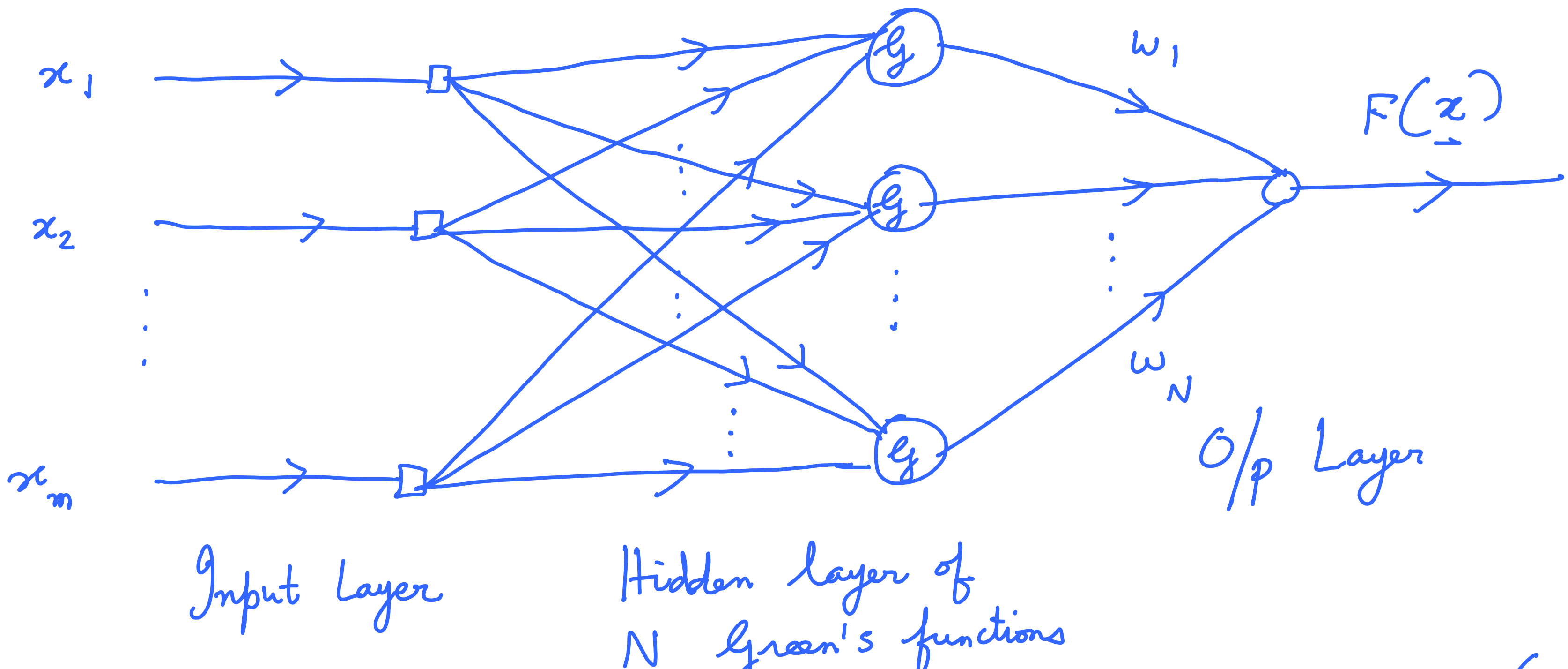
usual weight

Regularization Networks

The idea of Green's fns $G(\underline{x}, \underline{x}_i)$ centered @ $\underline{x}_i$ gives us a feel of the n/w structure

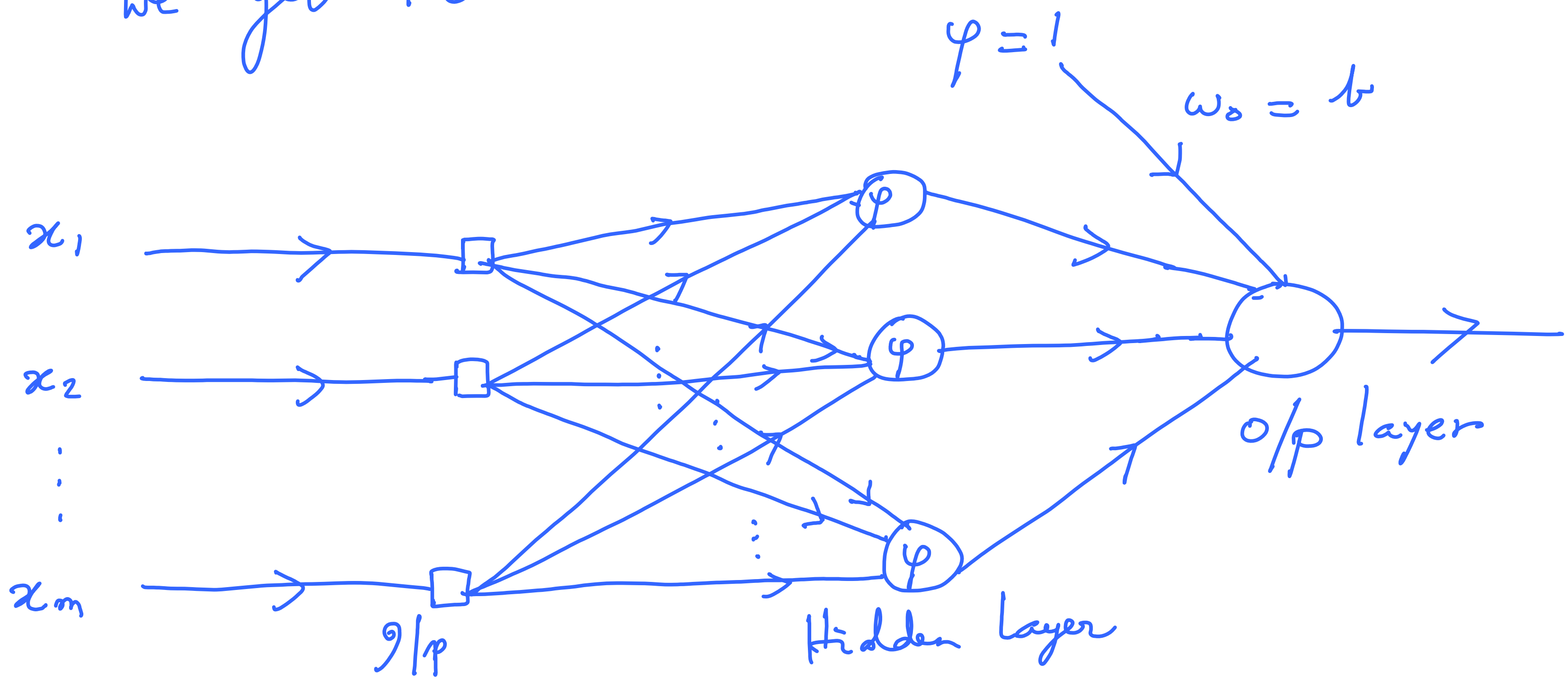
1) One hidden unit for each data point $\underline{x}_i$, $i = 1, \ldots, N$. The o/p of the hidden unit is $G(\underline{x}, \underline{x}_i)$.

2) The o/p of the n/w is $F(\underline{x})$ by combining the o/p of the Green's functions



For the ith hidden unit, the o/p is $G(\underline{x}, \underline{x}_i)$

By imposing certain constraints such as (+ve definite property) and making $G(\cdot)$ to be rotationally invariant, we get the Gaussian form used in RBF n/ws.



3 desirable properties for regularization n/ws from approximation theory perspective

1) It is a universal approximator; approx. any multivariate continuous fn very well.

2) Since the approx. scheme is derived from regularization theory & linear in the unknown coeffts, the unknown non-linear function can be always be approx. through an appropriate choice of the coeffts.

3) The soln. computed by a regularization n/w is optimal, and based on minimizing a functional