Capítulo 2

Cálculo de Matrices de Regularización

"No basta dar pasos que un día puedan conducir hasta la meta, sino que cada paso ha de ser una meta, sin dejar de ser un paso"
Anónimo.

Para comenzar

La regularización es una técnica ampliamente utilizada para trabajar con problemas mal situados y mal condicionados. Su aplicación ha sido explorada sobre una variedad de áreas diferentes, incluyendo inferencia bayesiana, análisis funcional, optimización, análisis numérico y sistemas conexionistas. Tras exportar el enunciado probabilista del problema de aprendizaje a partir de ejemplos hacia un contexto de aproximación de funciones multivariables, se enuncian diferentes situaciones sobre el conjunto de datos empíricos que propicían la introducción de técnicas de regularización. Esta metodología permite aumentar la estabilidad numérica de los resultados si es introducida como un elemento de penalización en forma cuadrática, así como asegurar la unicidad de la solución si es utilizada siguiendo el sentido teórico de Tikhonov. Mientras la primera opción se pondría en práctica mediante la elección *a priori* de la forma de la función aproximadora, por lo que se ha definido como visión bayesiana, la segunda versión transforma el problema en un proceso variacional.

La demostración de que las ANNs basadas en una estructura de funciones de base radial, RBFNN, poseen la propiedad de *mejor aproximación* en el sentido de regulari-

zación de Tikhonov, ha permitido mostrar en algunos casos la equivalencia funcional entre las dos visiones, bayesiana y variacional, si son utilizados como elementos de regularización funcionales definidos en teoría de 'splines'. Tal equivalencia permite reducir el cálculo de la matriz de regularización a un proceso directo y exacto, evitando los procedimientos iterativos que hasta el momento se hacían necesarios.

2.1 Introducción

El problema general de aprendizaje a partir de ejemplos fue establecido en la página 18 del Capítulo 1 como el desarrollo de una probabilidad condicionada sobre espacios de probabilidad cuya función de distribución es desconocida. Utilizando la terminología propia de la teoría de aproximación de funciones multivariables, el problema puede ser formulado como aquel de hallar una función f^{opt}, a partir de un conjunto admisible de funciones $\mathcal{LM}$, que permita la mejor representación, léase aproximación, de la totalidad del conjunto finito de datos empíricos.

La idoneidad de la aproximación se mide habitualmente substituyendo el funcional de riesgo $R(\omega)$ por el error de salida mímimo-cuadrático estándar — MSE, del inglés *Mean Squared Error* — , lo que supone la aplicación del principio inductivo ERM. La *propiedad 'clásica' de mejor aproximación* en el sentido del principio ERM conduce a una correlación exacta entre los ejemplos de entrada y salida del conjunto de datos, *la función interpolación.*

Sin embargo, en las implementaciones prácticas ya se comentó como la interpolación puede no ser la mejor aproximación ya que el conjunto de datos empíricos suele acumular pequeños errores de medida — ruido en las observaciones —; en otras ocasiones puede darse el caso que el conjunto de datos de trabajo sea muy pequeño en relación al espacio sobre el que se quiere realizar la aproximación; finalmente, es posible que algunas regiones del espacio de aproximación sean no observables y se carezca de datos sobre los que realizar la aproximación. Las funciones minimizadoras del criterio MSE, en estos casos de observaciones con ruido, no eliminarán los errores de medida asociados a los datos. Con la intención de generar nuevos modelos más robustos se hace necesario definir un nuevo funcional de riesgo.

En cuanto a la posibilidad de trabajo con un tamaño reducido o nulo del conjunto de datos en algunas regiones del espacio de aproximación, sería deseable que la función hallada tuviera unas características que pudieran servir para definirla como

una función 'suave', ya que este tipo de funciones son robustas y hacen posible la *propiedad de generalización*, aquella que permite a la función aproximadora retornar una salida razonable para un nuevo dato de entrada cuya salida no ha sido observada. Por tanto, para tratar con un conjunto de datos deficiente o incompleto, algunas restricciones de suavidad deberían ser añadidas al funcional de riesgo para obtener una solución regularizada.

Regularizar, en una definición general, significa que algunas restricciones son aplicadas durante el proceso de construcción de la función aproximadora con la finalidad de reducir el error de generalización. Las restricciones de suavidad son las más comunes para regularizar una solución, pero cualquier información *a priori* sobre la relación entrada-salida puede ser aplicada. Por ejemplo, en los modelos paramétricos las restricciones pueden ser seleccionadas sobre los valores de los parámetros con el objetivo de obtener modelos más tolerantes a fallos. Otra posibilidad consiste en la utilización de restricciones de escasez, es decir forzar la reducción en el número de bloques de construcción del modelo en presencia de un conjunto de datos sobre-representativos o redundantes. Finalmente, también pueden ser consideradas restricciones de estabilidad y convexidad.

La técnica de regularización puede ser aplicada con diferentes metodologías de aproximación de funciones, léase métodos de aprendizaje, incluyendo sistemas conexionistas, 'wavelets', regresión estadística, 'splines',... En el presente Capítulo la aproximación será realizada por medio de redes neuronales artificiales, ANN, en concreto una red neuronal de función de base radial — RBFNN, del inglés *Radial Basis Function Neural Networks* — [Powell, 1987]. El conjunto de funciones admisibles por esta clase de ANN tiene la forma

$$\mathcal{LM}_{RBF} = \left\{ f\left(\mathbf{x}, \omega\right) = \mathbf{G}\left(\mathbf{x}_i, \mathbf{x}\right) \cdot \omega + p_k\left(\mathbf{x}\right) : \mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d, \omega \in \Omega \right\} , \tag{2.1}$$

donde $\mathbf{G}\left(\mathbf{x}_i, \mathbf{x}\right)$ es una combinación de funciones de base radial, $p_k\left(\mathbf{x}\right)$ es una función polinomial de grado k, y los centros $\mathbf{x}_i$'s son un subconjunto de los vectores de entrada del conjunto de entrenamiento $\mathcal{T}$. Una gran ventaja que aporta este tipo de modelos RBFNN es su entrenamiento relativamente simple ya que la salida, obviando el polinomio que acompaña, es una combinación lineal de los pesos ω**.**

Es en este sentido, cuando la forma de las funciones solución es elegida de antemano para aplicar técnicas de regularización estándares o numéricas, en el que se ha definido esta aproximación al problema como visión bayesiana. Por contra, la visión variacional es el resultado de la aplicación de la técnica de regularización de Tikhonov al problema de aprendizaje que tiene cualquier tipo de función como posible

solución. No conlleva ninguna otra implicación la definición de ambas visiones puesto que el calificativo de *bayesiana* toma el nombre de un proceso de inferencia mientras que la visión denominada *variacional* lo toma como definición de un problema de optimización sobre espacios de funciones[1] .

2.2 Regularización

Definición 2.1. *La solución de una ecuación* $y = f(x)$ *es* estable en el sentido de Hadamard *si una pequeña variación en las observaciones de las salidas significa tan solo un pequeño cambio en la solución:*

$$\forall\, |\delta| < M, \quad \exists\, |\varepsilon| < N \quad : \quad y + \delta = f(x + \varepsilon) \ .$$

Definición 2.2. *El problema de solucionar una ecuación* $y = f(x)$ *es* bien situado en el sentido (duro) de Hadamard *si la solución de la ecuación:*

- *existe*
- *es única*
- *es estable*

La condición necesaria de estabilidad está relacionada con el requerimiento de robustez. Algunos autores reemplazan esta tercera condición por la de continuidad — problema bien situado en el sentido débil de Hadamard — . Sin embargo, la continuidad es una condición necesaria pero no suficiente para asegurar la estabilidad. Un problema bien situado en el sentido débil de Hadamard puede ser mal condicionado, no necesariamente robusto contra ruido [Bertero et al., 1988].

Los problemas extraídos de condiciones de la vida real deben tratar con conjuntos de datos discretos y por tanto en el caso lineal se reducen a la inversión de una matriz. La no unicidad y la inestabilidad numérica pueden provocar efectos similares sobre ellos. La regularización puede ser definida sobre ambos requerimientos no bien situados.

Teorema 2.3. *Sea*

$$\mathcal{T} = \left\{ (\mathbf{x}_p, y_p) \in \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R} \right\}_{p=1}^{\ell} , \tag{2.2}$$

[1] Agradezco la participación del Doctor Junbin Gao del ISIS Group de la Universidad de Southampton, UK en la discusión sobre las apreciaciones de la exposición [Català and Angulo, 2000].

un conjunto de entrenamiento definido sobre un espacio de salida real. Si se interpreta el problema de aprendizaje a partir de ejemplos en el caso lineal (como por ejemplo en las RBFNN) como aquel de hallar la función óptima de la forma

$$f(\mathbf{x}, \omega) = \mathbf{G} \cdot \omega = \mathbf{y}\,, \tag{2.3}$$

con $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_\ell)^\top \in \mathbb{R}^\ell \times \mathbb{R}^d$ *,* $\mathbf{G} = G(\mathbf{x}_i, \mathbf{x}_j) \in \mathbb{R}^\ell \times \mathbb{R}^\ell$ *,* $\omega = (\omega_1, \cdots, \omega_\ell)^\top \in \mathbb{R}^\ell \times \mathbb{R}$ *,* $\mathbf{y} = (y_1, \cdots, y_\ell)^\top \in \mathbb{R}^\ell \times \mathbb{R}$ *, minimizando el funcional de riesgo empírico estándar (1.8) con función de coste (1.6)*

$$R_{emp}^{L_2}(\omega) = \frac{1}{\ell} \cdot \sum_{i=1}^{\ell} (y_i - f(\mathbf{x}_i))^2\,, \tag{2.4}$$

entonces el vector de pesos óptimos

$$\omega^{MP} = \mathbf{R}^{-1} \cdot \mathbf{p}\,, \tag{2.5}$$

es la solución pseudo inversa de Moore-Penrose, donde $\mathbf{R} = \mathbf{G}^\top \cdot \mathbf{G} \in \mathbb{R}^\ell \times \mathbb{R}^\ell$ *es la matriz de auto-correlación y* $\mathbf{p} = \mathbf{G}^\top \cdot \mathbf{y} \in \mathbb{R}^\ell \times \mathbb{R}$ *es la matriz de correlación cruzada.*

2.2.1 Regularización para la Estabilidad Numérica

A menudo, la matriz de auto-correlación $\mathbf{R}$ está mal condicionada y es dificultoso y largo hallar de forma eficiente el vector de pesos óptimos ω^{MP} en (2.5). La regularización es un método popular [Bertero et al., 1988], [Marroquin et al., 1987], [Wahba, 2000] usado para restringir la optimización de los pesos que se realiza penalizando el funcional de riesgo empírico mediante la adición de alguna distribución *a priori* sobre los pesos. El funcional de riesgo es penalizado mediante un *funcional error penalizador*, E^ℓ, obteniéndose el funcional de riesgo regularizado idéntico a (1.12)

$$R_{emp}^{reg}(\omega) = R_{emp}(\omega) + \lambda \cdot E^\ell\,. \tag{2.6}$$

La penalización E^ℓ es usualmente representada mediante una función cuadrática en los pesos,

$$E^\ell = \omega^\top \cdot \mathbf{K} \cdot \omega\,, \tag{2.7}$$

por lo que la solución al funcional de riesgo (2.6) resulta en

$$\omega^* = (\mathbf{R} + \lambda \cdot \mathbf{K})^{-1} \cdot \mathbf{p}\,, \tag{2.8}$$

que generalmente conduce a problemas mejor condicionados. La matriz $\mathbf{K}$ es denominada *matriz de regularización* y es aquella que impone la distribución *a priori* de los

pesos. El *parámetro de regularización*, λ, controla el compromiso entre la adecuación de la función buscada al conjunto de datos y el cumplimiento de las restricciones impuestas por el funcional penalizador. La elección de este parámetro depende de forma crítica del problema particular que se esté trabajando [Bossley, 1993].

2.2.2 Regularización de Tikhonov para la Unicidad

Se mostró con anterioridad que en el marco de la teoría de aproximación de funciones, el problema de aprendizaje a partir de ejemplos es mal situado en el sentido que el conjunto de datos no es suficiente para reconstruir de forma única una aplicación en las regiones donde no existen datos disponibles. Se necesitará, pues, realizar algunas suposiciones *a priori* sobre la aplicación con la finalidad de convertir el problema en bien situado.

La regularización de Tikhonov reemplaza el problema de aproximación original con el problema variacional de hallar la superficie que minimiza un funcional de riesgo consistente en dos términos, al igual que en la expresión (2.6), donde en esta ocasión el error de penalización, E^{ℓ}, es representado por

$$E^{\ell} = \|Pf\|^2 = \int_{\mathbb{R}^d} |Pf|^2 \, d\mathbf{x} = \int_{\mathbb{R}^d} f(\mathbf{x}) \, \widehat{P}Pf(\mathbf{x}) \, d\mathbf{x} \,, \tag{2.9}$$

con P un operador restricción, denominado estabilizador o *regularizador*, siendo usualmente un operador diferencial, y $\widehat{P}$ el operador adjunto.

Se observa que los funcionales de riesgo (1.12) y (2.6) tienen gran similaridad en ambos desarrollos regularizadores. Sería deseable tomar ventaja de estas similaridades para trasladar información entre las dos aproximaciones. Las RBFNN son un conjunto de funciones indicado para implementar la teoría de regularización puesto que es un modelo lineal sobre los pesos. Por otra parte, esta formulación puede ser fácilmente extendida a casos con múltiples funciones penalizadoras y parámetros de regularización.

2.3 RBFNN Regularizadas

Definición 2.4. *Se define una* RBFNN regularizada *como una función RBFNN estándar tal como se define en (2.1) asociada al funcional de riesgo modificado* $R^{reg}_{emp}(\omega)$ *dado por la expresión (2.6).*

El error de penalización, E^{ℓ}, estará representado en esta ocasión por una función cuadrática de los pesos definida en (2.7). Por motivos de brevedad en la nomenclatura este tipo de redes suele recibir el nombre de Redes de Regularización — RN, del inglés *Regularization Networks* —.

Las diferencias entre la solución regularizada y aquella no regularizada se centran en el término añadido a la medida del error mínimo cuadrático. El vector de pesos óptimo puede ser calculado mediante métodos directos como la solución pseudo inversa de Moore-Penrose, o por medio de procedimientos iterativos como los procesos de aprendizaje de una ANN. En cualquier caso, la contribución del nuevo término al riesgo total depende de forma crítica de la elección del parámetro de regularización λ y de la matriz de regularización $\mathbf{K}$.

En la literatura, el mayor esfuerzo investigador ha estado dirigido al cálculo sobre la intensidad de regularización que debe poseer una solución, es decir, cuál es el parámetro regularizador apropiado para un problema específico. K. M. Bossley [Bossley, 1993] describe diferentes técnicas útiles. Obviamente, un estudio sobre el grado de regularización para un problema particular es necesario, sin embargo el hecho de que la función a ser aproximada sea una función desconocida provoca que este trabajo comparativo en busca de la determinación del mejor parámetro para un problema específico posea siempre un alto grado de incertidumbre motivada por el ruido en las observaciones o la ausencia de datos. La voluntad de obtener una precisión cada vez mayor en el proceso de aproximación a una función desconocida ruidosa mediante la determinación de un parámetro exacto es un propósito poco realista.

Otras buenas características del comportamiento de la función aproximadora (robustez, tolerancia a fallos, arquitectura de dimensión reducida, bajo costo computacional ...) podrían ser obtenidas si fuera utilizado un apropiado término penalizador. Los problemas en esta dirección surgen desde la pregunta sobre cómo expresar estas características en forma de un regularizador específico, o, alternativamente, determinar cuál es la apropiada matriz de regularización e incluso preguntarse si el término penalizador es una función cuadrática de los pesos o no. Tradicionalmente, un marco estadístico es utilizado para trasladar la usual restricción de suavidad en forma de distribución *a priori* de los pesos en el término penalizador, aunque el uso de transformadas de Fourier también ha sido estudiado [Girosi et al., 1993]. Dos casos bien conocidos del uso de la inferencia bayesiana sobre RBFNN regularizadas se describen en los siguientes apartados.

2.3.1 Regularización de Orden Cero desde una Visión Bayesiana (Ridge Regression)

Es esta la forma más común de regularización debido a su bajo coste computacional. El término penalizador se expresa como

$$E^{\ell} = \int_{\mathbb{R}^d} |f(\mathbf{x}, \omega)|^2 \, p(\mathbf{x}) \, d\mathbf{x} \,, \tag{2.10}$$

donde $p(\mathbf{x})$ es la función densidad de probabilidad. El valor de E^{ℓ}, representando el tamaño esperado de la salida, puede ser aproximado por una función cuadrática de los pesos donde la matriz de regularización $\mathbf{K}$ es la matriz identidad

$$\mathbf{K} = \mathbf{Id} \,. \tag{2.11}$$

La minimización de E^{ℓ} tiene un efecto similar a aquel de reducir parámetros superfluos en la solución. En el contexto de trabajo de las redes neuronales podría ser comparado a un proceso de poda.

2.3.2 Regularización de Segundo Orden desde una Visión Bayesiana (Regresión no lineal)

Con la intención de asumir que la función es suave, se define como término penalizador

$$E^{\ell} = \int_{\mathbb{R}^d} \left| \frac{\partial^2 f(\mathbf{x}, \omega)}{\partial \mathbf{x}^2} \right| p(\mathbf{x}) \, d\mathbf{x} \,. \tag{2.12}$$

En este caso, E^{ℓ} representa la curvatura esperada de la salida. Puede ser aproximado por una función cuadrática de los pesos donde la matriz de regularización $\mathbf{K}$ es una función cuadrada de la curvatura

$$\mathbf{K} = \mathbf{\Psi}(curvatura) \,. \tag{2.13}$$

Esta forma de regularización es muy popular, pero es siempre necesario realizar rudas aproximaciones para obtener la matriz $\mathbf{K}$ puesto que el cálculo de derivadas de segundo orden de funciones base multidimensionales es computacionalmente intensivo. Así, a la hora de utilizar una perspectiva bayesiana, la complejidad computacional limita su aplicabilidad.

2.4 Problema Variacional de la Regularización

Teorema 2.5 ([Poggio and Girosi, 1989]). *Sea $\mathcal{T}$ el conjunto de datos empíricos con salida real definido en (2.2) a ser aproximado por una función f,* en forma general. *La solución al problema de regularización de Tikhonov, es decir, la minimización de (2.6) siendo el funcional error penalizador E^{ℓ} representado por la expresión (2.9), es dada por la fórmula:*

$$f(\mathbf{x}, \omega) = \mathbf{G}(\mathbf{x}_i, \mathbf{x}) \cdot \omega + p_k(\mathbf{x}) \,, \tag{2.14}$$

donde

- $\mathbf{G}(\mathbf{x}_i, \mathbf{x})$ *satisface la siguiente ecuación diferencial distribucional* — función de Green —
$$\widehat{P}PG(\xi, \mathbf{x}) = \delta(\mathbf{x} - \xi) \,, \tag{2.15}$$
- $p_k(\mathbf{x})$ *es una función polinomial del espacio núcleo definido por el regularizador de E^{ℓ} ,*
- ω^{Tik}*, vector de pesos óptimo, satisface el sistema lineal*
$$(\mathbf{G} + \lambda \cdot \mathbf{Id}) \cdot \omega = \mathbf{y} \,. \tag{2.16}$$

El presente teorema demuestra que una RBFNN — obsérvese la equidad entre la expresión de las funciones admisibles en (2.1) y el resultado obtenido en (2.14) — posee la propiedad de mejor aproximador en el sentido variacional utilizando regularización de Tikhonov si las funciones base elegidas en la definición de $\mathbf{G}$ cumplen la ecuación de Green (2.15). Además la ecuación lineal (2.16) permite obtener de forma efectiva los pesos óptimos solución. Sin embargo, a pesar de ser una demostración constructiva, presenta inconvenientes computacionales en su aplicación. La función que minimiza el funcional de riesgo está explicitada sobre ℓ coeficientes, siendo ℓ el número de ejemplos en el conjunto de datos $\mathcal{T}$, por lo que en el caso que ℓ sea una cantidad elevada, la computación de los coeficientes de la expansión puede convertirse en un operación con gran consumo de tiempo: su complejidad crece polinomialmente con ℓ, del orden $O(\ell^3)$, puesto que debe ser invertida una matriz $\ell \times \ell$ [Broomhead and Lowe, 1988].

Una forma eficiente de reducir la complejidad del problema es utilizar una función aproximada, $\widetilde{f}(\mathbf{x}, \omega)$, sobre una base de elementos más reducida,

$$\widetilde{f}(\mathbf{x}, \omega) = \sum_{r=1}^{c \ll \ell} \omega_r \cdot G(\mathbf{t}_r, \mathbf{x}) \,, \tag{2.17}$$

donde el vector con los coeficientes reales ω_r y los centros $\mathbf{t}_r \in \mathbb{R}^d$ son incógnitas que podrán ser halladas imponiendo la condición de que el conjunto de parámetros $\{\omega_r, \mathbf{t}_r\}_{r=1}^{c}$ debe ser tal que minimice el funcional de riesgo $R_{emp}^{reg}(\omega, \mathbf{t})$.

La forma explícita del sistema de ecuaciones a ser satisfecho por los coeficientes depende del operador restricción específico utilizado. En el presente estudio se tomarán como penalizadores generales aquellos definidos por Duchon y Meinguet con la intención de generalizar la teoría de 'splines' de una a varias dimensiones, funcionales rotacionalmente invariantes de la siguiente forma [Poggio and Girosi, 1990],

$$\|P_m f\|^2 = \sum_{i_1+\cdots+i_m}^{n} \int_{\mathbb{R}^d} \|\partial_{i_1,\cdots,i_m} f(\mathbf{x})\|^2 \, d\mathbf{x}\,, \tag{2.18}$$

donde

$$O^m = \partial_{i_1,\cdots,i_m} = \frac{\partial^m}{\partial \mathbf{x}_{i_1} \cdots \partial \mathbf{x}_{i_m}}\,, \quad m \geq 1\,. \tag{2.19}$$

Esta definición puede ser extendida de forma natural al caso $m = 0$ si se define

$$\partial_{i_1,\cdots,i_{m=0}} f(\mathbf{x}) = f(\mathbf{x})\ . \tag{2.20}$$

2.5 Similaridades entre las Visiones Bayesiana y de Tikhonov

La visión de la inferencia bayesiana que permite describir las RBFNN regularizadas trata a la función solución con una descripción similar a aquella de la visión variacional, tal como fue apuntado anteriormente al comparar las expresiones (2.1) y (2.14). Si el estabilizador P es rotacionalmente invariante entonces la solución regularizada es dada por una expansión en funciones de base radial

$$G(\xi, \mathbf{x}) = G(\|\mathbf{x} - \xi\|)\ , \tag{2.21}$$

y la aproximación de Tikhonov lleva a una expansión en RBFs. Para determinar una equivalencia completa entre ambas visiones o aproximaciones se hace necesario realizar:

- una comparación entre la solución del vector de pesos óptimo en (2.8) con la solución de Tikhonov del vector de pesos óptimo (2.16);

- una comparación entre los funcionales de riesgo, es decir cómo puede ser representada la matriz de regularización $\mathbf{K}$.

Iniciando el estudio con el funcional de riesgo modificado $R_{emp}^{reg}(\omega)$, si la función solución f en forma general es reemplazada por la función aproximada $\widetilde{f}$ sobre una base reducida, el funcional penalizador E^{ℓ} resulta en

$$\begin{aligned} E^{\ell} = \|P_m f\|^2 &= \sum_{i_1+\cdots+i_m}^{n} \int_{\mathbb{R}^d} \left\| \partial_{i_1,\cdots,i_m} \widetilde{f}(\mathbf{x},\omega) \right\|^2 d\mathbf{x} = \\ &= \int_{\mathbb{R}^d} \widetilde{f}(\mathbf{x},\omega)\, \widehat{O^m} O^m \widetilde{f}(\mathbf{x},\omega)\, d\mathbf{x} = \\ &= \int_{\mathbb{R}^d} \widetilde{f}(\mathbf{x},\omega)\, \widehat{O^m} O^m \sum_{r=1}^{c} \omega_r \cdot G(\mathbf{t}_r,\mathbf{x})\, d\mathbf{x} = \\ &= \int_{\mathbb{R}^d} \left(\sum_{s=1}^{c} \omega_s \cdot G(\mathbf{t}_s,\mathbf{x}) \right) \left(\sum_{r=1}^{c} \omega_r \cdot \delta(\mathbf{x}-\mathbf{t}_r) \right) d\mathbf{x} = \\ &= \sum_{r,s=1}^{c} \omega_r \cdot \omega_s \cdot G(\mathbf{t}_r,\mathbf{t}_s)\ . \end{aligned} \tag{2.22}$$

Definiendo una $\ell \times c$–matriz $\mathbf{G}$ como $(\mathbf{G})_i^r = G(\mathbf{t}_r,\mathbf{x}_i)$ y una $c \times c$–matriz cuadrada simétrica $(\mathbf{g})_r^s = G(\mathbf{t}_r,\mathbf{t}_s)$, se puede reescribir (2.6) como

$$R_{emp}^{reg}(\omega) = \omega^{\top} \left(\mathbf{G}^{\top}\mathbf{G} + \lambda \mathbf{g} \right) \omega - 2\omega^{\top}\mathbf{G}^{\top}\mathbf{y} + \mathbf{y}^{\top}\mathbf{y}\,, \tag{2.23}$$

que posee forma cuadrática sobre los coeficientes. Se deduce así que, para todo conjunto de centros fijado, el vector de pesos óptimo se obtiene de la expresión

$$\omega^* = \left(\mathbf{G}^{\top}\mathbf{G} + \lambda \mathbf{g} \right)^{-1} \mathbf{G}^{\top}\mathbf{y}\,. \tag{2.24}$$

Si $\mathbf{R} = \mathbf{G}^{\top}\mathbf{G}$, $\mathbf{p} = \mathbf{G}^{\top}\mathbf{y}$, y $\mathbf{K} = \mathbf{g}$, entonces (2.24) puede reescribirse como

$$\omega^* = (\mathbf{R} + \lambda \cdot \mathbf{K})^{-1} \cdot \mathbf{p}\,, \tag{2.25}$$

observándose que la expresión obtenida para los coeficientes es equivalente a aquella obtenida en la regularización para la estabilidad numérica (2.8). Más aún, para cualquier función de Green G asociada a un estabilizador P, se ha obtenido una expresión explícita y directa para el cálculo de la matriz de regularización $\mathbf{K}$.

2.5.1 Equivalencia entre las Visiones Bayesiana y Variacional para la Regularización de Orden Cero

Proposición 2.6 ([Angulo and Català, 1998]). *La regularización de orden cero es equivalente a la regularización de Tikhonov si el operador que define la regularización es el funcional extendido de Duchon-Meinguet O^m con $m = 0$ y las funciones de base radial son funciones delta de Dirac.*

Demostración. Puede observarse que en este caso la función de Green asociada al regularizador satisface

$$G(\xi, \mathbf{x}) = \delta(\mathbf{x} - \xi) \,, \tag{2.26}$$

por lo que la solución aproximada toma la forma

$$\widetilde{f}(\mathbf{x}, \omega) = \sum_{r=1}^{c} \omega_r \cdot \delta(\mathbf{x} - \mathbf{t}_r) \,. \tag{2.27}$$

Utilizando la expresión (2.8) o (2.24) se obtiene el vector de pesos óptimo, donde

$$\mathbf{K} = (\mathbf{g})_r^s = \delta(\mathbf{t}_r - \mathbf{t}_s) = \mathbf{Id} \,, \tag{2.28}$$

de manera que el término penalizador en el funcional de riesgo puede expresarse como

$$\begin{aligned} E^\ell &= \left\| P_0 \widetilde{f} \right\|^2 = \int_{\mathbb{R}^d} \left\| O^0 \widetilde{f}(\mathbf{x}) \right\|^2 d\mathbf{x} = \\ &= \int_{\mathbb{R}^d} \left(\sum_{r=1}^{c} \omega_r \cdot G(\mathbf{x} - \mathbf{t}_r) \right) \left(\sum_{s=1}^{c} \omega_s \cdot G(\mathbf{x} - \mathbf{t}_s) \right) d\mathbf{x} = \\ &= \sum_{i=1}^{c} \omega_i^2 = \omega^\top \cdot \omega = \omega^\top \cdot \mathbf{K} \cdot \omega \,, \quad \text{siendo } \mathbf{K} = \mathbf{Id} \,. \end{aligned} \tag{2.29}$$

Estos resultados son equivalentes a aquellos obtenidos en la visión bayesiana del marco de la regularización. □

2.5.2 Equivalencia entre las Visiones Bayesiana y Variacional para la Regularización de Segundo Orden

Proposición 2.7 ([Angulo and Català, 1998]). *La regularización de segundo orden es equivalente a la regularización de Tikhonov si el operador que define la regularización es el funcional de Duchon-Meinguet O^m con $m = 2$ y las funciones de base radial son las bien conocidas funciones 'splines thin-plate' multidimensionales.*

Demostración. Puede observarse que en este caso la función de Green asociada al regularizador satisface

$$\widehat{O^2}O^2G\left(\xi,\mathbf{x}\right)=\delta\left(\mathbf{x}-\xi\right)\,, \tag{2.30}$$

cuya solución es

$$G\left(\mathbf{r}\right)=\begin{cases}\left\|\mathbf{r}\right\|^{4-d}\cdot\ln\left(\left\|\mathbf{r}\right\|\right) & \text{si } d=\dot{2}\\ \left\|\mathbf{r}\right\|^{4-d} & \text{si } d\neq\dot{2}\end{cases}\,, \tag{2.31}$$

La función solución aproximada toma la forma

$$\widetilde{f}\left(\mathbf{x},\omega\right)=\sum_{r=1}^{c}\omega_r\cdot G\left(\mathbf{t}_r,\mathbf{x}\right)+p_1\left(\mathbf{x}\right)\,. \tag{2.32}$$

Utilizando la expresión (2.8) o (2.24) se obtiene el vector de pesos óptimo, donde

$$\mathbf{K}=\left(\mathbf{g}\right)_r^s=G\left(\mathbf{t}_r-\mathbf{t}_s\right)=\mathbf{g}\,, \tag{2.33}$$

de manera que el término penalizador en la función de coste puede expresarse como[2]

$$\begin{aligned}E^\ell&=\left\|P_2\widetilde{f}\right\|^2=\int_{\mathbb{R}^d}\left\|O^2\widetilde{f}\left(\mathbf{x}\right)\right\|^2d\mathbf{x}=\\&=\int_{\mathbb{R}^d}\widetilde{f}\left(\mathbf{x}\right)\widehat{O^2}O^2\left(\sum_{r=1}^{c}\omega_r\cdot G\left(\mathbf{t}_r,\mathbf{x}\right)+p_1\left(\mathbf{x}\right)\right)d\mathbf{x}=\\&=\int_{\mathbb{R}^d}\left(\sum_{s=1}^{c}\omega_s\cdot G\left(\mathbf{t}_s,\mathbf{x}\right)+p_1\left(\mathbf{x}\right)\right)\left(\sum_{r=1}^{c}\omega_r\cdot\delta\left(\mathbf{x}-\mathbf{t}_r\right)\right)d\mathbf{x}=\\&=\left(\sum_{s=1}^{c}\omega_s\cdot G\left(\mathbf{t}_s,\mathbf{t}_r\right)+p_1\left(\mathbf{t}_r\right)\right)\left(\sum_{r=1}^{c}\omega_r\right)=\\&=\omega^\top\cdot\mathbf{g}\cdot\omega=\omega^\top\cdot\mathbf{K}\cdot\omega\,,\quad\text{siendo }\mathbf{K}=\mathbf{g}\,.\end{aligned} \tag{2.34}$$

□

2.5.3 Simplicidad Computacional del Nuevo Método

La matriz de regularización $\mathbf{K}$ para la regularización de orden cero es la matriz identidad. Esta forma de regularización puede ser implementada con la adición de una matriz muy simple, una relativamente ineficiente operación computacional. En este caso, la aproximación de Tikhonov es idéntica a aquella que utiliza inferencia bayesiana, por lo que es computacionalmente barata.

[2] En la penúltima igualdad, si $c=\ell$ y $\{\mathbf{x}_i\}_{i=1}^{\ell}=\{\mathbf{t}_r\}_{r=1}^{\ell}$ entonces el término $\sum\omega_r p_1\left(\mathbf{t}_r\right)$ es cero; en el caso general, esta condición es un nuevo sistema lineal de ecuaciones a ser cumplido por el polinomio multivariable de grado 1.

Hasta el momento, emplear funciones densidad de probabilidad en la regularización de segundo orden significaba que el proceso de cálculo de la matriz de regularización fuera costoso al ser necesario calcular las derivadas segundas de funciones base multidimensionales. El uso de la regularización de Tikhonov permite identificar de forma exacta la matriz $\mathbf{K}$ y los cálculos no son complicados al ser elegida como función de base radial aquella específica cumpliendo la función de Green. El mayor gasto computacional con el nuevo método pasa a ser el tiempo de evaluación del funcional error penalizador a partir del cálculo del valor de las funciones de base radial sobre los centros.

2.6 En Resumen

En el presente Capítulo el principal resultado mostrado ha sido la equivalencia entre las visiones bayesiana a la teoría de la regularización y la regularización de Tikhonov dentro del marco de la teoría de aproximación de funciones cuando son empleadas redes neuronales de función de base radial. La equivalencia ha sido empleada para poder obtener de una *forma directa* la expresión de la matriz de regularización $\mathbf{K}$, evitando el proceso de cálculo que hasta el momento resultaba tan caro desde el punto de vista de tiempo computacional. En particular se ha demostrado que la regularización de segundo orden puede ser obtenida desde un marco de trabajo variacional. El tan conocido problema de la 'explosión de la dimensionalidad' es el principal escollo a superar cuando se aplica la metodología de la regularización puesto que han de ser calculadas derivadas en un espacio multidimensional. Tradicionalmente se realizan aproximaciones más o menos precisas para calcular la matriz Hessiana de derivadas de orden dos por lo que la precisión final se ve mermada. El nuevo planteamiento ofrece una fácil y eficiente expresión para el cálculo de esta matriz.

La herramienta clave para la transferencia de información entre las visiones bayesiana y variacional es el operador regularizador P. Este operador funcional debe ser capaz de transferir las restricciones — tolerancia a fallos, suavidad, estabilidad, o cualquier otro tipo de comportamiento — y a su vez su función de Green asociada (2.15) debe tener una solución razonable. Una importante tarea matemática a realizar es la de buscar operadores que sean adecuados a las citadas características.

El principal lastre que arrastra la aplicación de métodos de regularización en la aproximación de funciones multidimensional sobre un conjunto grande de patrones de aprendizaje es la falta de capacidad en reducir la dimensionalidad del problema tratado. Intentando reducir esta explosión de dimensionalidad, una aproximación ha sido planteada expandiendo la solución sobre un conjunto reducido de funciones de base radial. El área que trata este problema general es conocida como aproximación escasa. A pesar de los prometedores estudios iniciales basados en técnicas de regularización que se han llevado a cabo, su aplicación se ha visto relegada por el éxito

obtenido por las máquinas de soporte vectorial basadas en la teoría del aprendizaje estadístico [Girosi, 1998].

Finalmente, señalar que la técnica de regularización puede ser generalizada de forma evidente mediante la adición y combinación de cualquier tipo de término penalizador e incluso podría imponerse que algunas restricciones pudieran ser de tipo no lineal.