Universidade Federal de Minas Gerais
Programa de Pós-graduação em Engenharia Elétrica

Illya Kokshenev

Aprendizado Multi-objetivo de Redes RBF e de Máquinas de Kernel

Belo Horizonte/MG — Junho 2010

Illya Kokshenev

Aprendizado Multi-objetivo de Redes RBF e de Máquinas de Kernel

Tese apresentada à banca examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais, como parte dos requisitos necessários à obtenção do grau de doutor em Engenharia Elétrica.

Orientador: Prof. Dr. Antônio P. Braga

Belo Horizonte/MG — Junho 2010

Resumo estendido

Avanços recentes em aprendizagem de máquina contemplam ampla variedade de tarefas de inteligência computacional, que têm sido aplicadas em vários problemas modernos de engenharia, economia, biomedicina, dentre outros. Tarefas como reconhecimento de padrões, previsão de séries temporais, controle adaptativo e detecção de falhas podem ser formuladas em termos da busca de dependências escondidas em observações empíricas. Tal busca desempenha um papel central no contexto de aprendizagem de máquina, e corresponde ao problema de aprendizagem supervisionada.

Considerando que as observações empíricas são geralmente induzidas por variáveis de entrada não observadas, cujas propriedades são desconhecidas, a aprendizagem supervisionada deve ser tratada como um processo não-determinístico nas condições de incerteza. Isto, por sua vez, faz com que o complexo metodológico da aprendizagem de máquina seja difundido para outros campos, tais como estatística, programação matemática, teorias de informação e tomada de decisões.

Uma visão abrangente do problema de aprendizagem supervisionada é dada pela teoria de aprendizagem estatística (SLT) [Vapnik, 1998], que estabelece os princípios de minimização empírica e estrutural do risco (ERM e SRM, correspondentemente). A implementação destes princípios encontra-se nos bem conhecidos conceitos de aprendizagem, tais como redes de regularização [Poggio and Girosi, 1990] e aprendizagem Bayesiana [Neal, 1996], cuja combinação com as máquinas de vetores de suporte (SVM) [Cortes and Vapnik, 1995] e métodos de kernel modernos [Scholkopf, 1999] representam o estado da arte em aprendizagem de máquina.

Com o recente desenvolvimento da otimização evolucionária, tem havido crescente interesse na aplicação do conceito de *Pareto-optimality* para estender as capacidades dos algoritmos e modelos aprendizagem. Esses conceitos tem sido aplicados no desenvolvimento de métodos de aprendizagem de máquina multi-objetivo (MOML) [Jin, 2006], em que a aprendizagem é considerada como um processo de tomada de decisão no ambiente de múltiplos critérios conflitantes.

Do ponto de vista multi-objetivo, a aprendizagem supervisionada pode ser representada como um problema de tomada de decisão entre dois objetivos conflitantes: minimização de erro de treinamento (risco empírico) e complexidade de modelo. Dentro da abordagem tradicional (mono-objetivo), este processo corresponde à seleção de modelos pelo princípio SRM. Esta visão quanto ao problema de aprendizagem supervisionada apresenta-se como o objeto de pesquisa deste trabalho.

A presente tese foi estruturada a partir de seis capítulos e três apêndices que apresentam uma introdução do tema, uma análise sistemática dos fundamentos teóricos conhecidos, desenvolvimento da metodologia proposta, suas aplicações e resultados dos testes realizados, além das conclusões finais.

O **Capítulo 1** apresenta uma introdução aos problemas do presente trabalho, com as suas motivações, justificativas iniciais e objetivos desta pesquisa.

As máquinas de aprendizagem modernas, tais como SVM, são baseadas no princípio de regularização e, portanto, representam problemas convexos cujas soluções únicas podem ser obtidas de maneira eficiente por meio da programação não-linear. Entretanto, suas extensões aos mais amplos espaços de hipóteses (e.g., com introdução de parâmetros de kernel), tradicionalmente, são efetuadas no nível de seleção de modelo através de uma busca no espaço de múltiplos hiper-parâmetros. Essa abordagem de extensão, no entanto, não corresponde ao princípio de SRM, que apresenta-se como uma busca unidimensional de equilíbrio entre o erro e complexidade de modelo. Por outro lado, a extensão correspondente ao princípio de SRM é possível com a abordagem multi-objetivo. Entretanto, devido a não-convexidade dos problemas de otimização associados, que são NP-completos, a aplicação de técnicas de soluções aproximadas de programação global é requerida. Este fato explica porque a maioria das soluções propostas da área MOML (e.g., [Liu. and Kadirkamanathan, 1995; Hatanaka and Uosaki, 2003; Jin, Okabe, and Sendhoff, 2004; Bevilacqua, Mastronardi, Menolascina, Pannarale, and Pedone, 2006; Yen, 2006; Kondo, Hatanaka, and Uosaki, 2006]) são orientadas para as técnicas de otimização evolucionária, prestando pouca atenção à implementação dos princípios fundamentais de aprendizagem estatística.

Como alternativa, nos trabalhos [Teixeira, Braga, Takahashi, and Saldanha, 2000; Costa, Braga, Menezes, Teixeira, and Parma, 2003; Costa and Braga, 2006] foi desenvolvida a abordagem chamada MOBJ, para a aprendizagem multi-objetivo das redes perceptron de múltiplas camadas (MLP). Neste caso, com o objetivo de controle da generalização conforme [Bartlett, 1997], o erro de treinamento foi minimizado junto com a norma Euclidiana dos pesos de rede através das técnicas de programação não-linear de maneira determinística (não-evolucionária). Contudo, devido a não-convexidade do problema tratado, a abordagem MOBJ desenvolvida nos trabalhos supracitados pode apresentar soluções fracamente não-dominadas.

Conforme os resultados publicados, ambas as abordagens de multi-objetivo (evolucionária e MOBJ) demonstram um bom potencial, enquanto a conexão com SRM e utilização das capacidades de programação não-linear mostram as vantagens do MOBJ. Então, sua certa evolução em direção à garantia de *Pareto-optimality* de soluções e a implementação de SRM determina um caminho de desenvolvimento para uma nova abordagem eficiente.

Por apresentar fundamentos teóricos associados com máquinas de kernel e regularização, as redes de funções da base radial (RBF) foram escolhidas para o desenvolvimento dos novos conceitos, métodos, e modelos de aprendizagem multi-objetivo neste trabalho.

O **Capítulo 2** apresenta os conceitos teóricos da aprendizagem estatística, redes de regularização e RBF, máquinas de kernel e as suas interligações dentro de um conceito unificado.

Do ponto de vista estatístico, dada uma função de perda $l(x, y, f(x))$ como uma medida de erro de classificação ou regressão, e o conjunto de N observações

$$Z_{\text{tr}}^N := \Big\{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1 \ldots N\Big\},$$

i.i.d. de acordo com a distribuição desconhecida $P(x, y)$, o problema de aprendizagem supervisionada é formulado como minimização do funcional de risco esperado

$$R[f] := \int_{\mathcal{X}\times\mathcal{Y}} l(x, y, f(x))\partial P(x, y) = E[l(x, y, f(x))], \tag{1}$$

sobre o espaço de hipóteses Ω. O espaço Ω representa uma classe de funções $f : \mathcal{X} \to \mathcal{Y}$, suportadas pela máquina de aprendizagem, que mapeiam as observações do domínio entrada $\mathcal{X}$ para domínio de saída $\mathcal{Y}$. Devido a indisponibilidade de $P(x, y)$, o funcional (1) não pode ser minimizado diretamente, mas somente a sua aproximação empírica

$$R_{\text{emp}}[f] := \frac{1}{N}\sum_{i=1}^{N} l(x_i, y_i, f(x_i)), \tag{2}$$

disponível através do conjunto de observações Z_{tr}^N. Entretanto, conforme [Vapnik and Chervonenkis, 1989], a minimização do (2) que corresponde ao princípio de minimização do risco empírico (ERM), leva a uma estimação consistente do mínimo do risco esperado somente quando a convergência de $R_{\text{emp}}[f]$ para $R[f]$ é uniforme, assumindo a condição de que a capacidade da classe Ω seja limitada. Como mostra o resultado de análise da convergência, existe um limite superior do risco esperado na forma

$$R[f] \leq R_{\text{emp}}[f] + \Psi\left(R_{\text{emp}}[f], N, \Omega, \eta\right) \tag{3}$$

onde Ψ determina o intervalo de confiança que se mantém com a probabilidade maior do que $1 - \eta$. É possível demonstrar que o risco empírico é uma função decrescente da capacidade de Ω, e crescente para o intervalo de confiança (o mesmo fenômeno é conhecido como o dilema *bias-variance* [Geman, Bienenstock, and Doursat, 1992]). Então, existe um espaço Ω de certa capacidade que garante o menor risco esperado através do limite (3). Esta idéia constitui a base do princípio indutivo de minimização estrutural do risco (SRM), que considera a construção de uma estrutura de conjuntos aninhados,

$$\emptyset \subset \Omega_1 \subset \Omega_2 \subset \ldots \subset \Omega$$

na ordem de crescimento das suas capacidades. Posteriormente, a hipótese final a ser escolhida corresponde ao mínimo de R_{emp} em um determinado subconjunto Ω_i, cuja complexidade leva ao menor risco empírico, sendo este garantido por seu limite superior. Desta maneira, o problema de aprendizagem supervisionado é visualizado como uma busca de equilíbrio entre o mínimo de risco empírico (menor erro de treinamento) e menor capacidade de espaço de hipóteses de máquina de aprendizagem (complexidade do modelo).

É possível mostrar que o problema do aprendizado supervisionado, tratado como um problema *ill-posed* com o método de regularização [Tikhonov, 1943], corresponde à implementação do princípio de SRM na forma de minimização do funcional risco regularizado

$$R_{\text{reg}}[f] := R_{\text{emp}}[f] + \lambda Q[f], \tag{4}$$

onde $Q[f] = \|Df\|^2$ é o termo estabilizador (regularizador) baseado em um operador linear diferencial D, e λ é o parâmetro da regularização. Como foi mostrado em [Poggio and Girosi, 1990] e [Scholkopf, Herbrich, Smola, and Williamson, 2001], no caso mais geral, o mínimo global de (4) está contido no espaço de Hilbert do kernel reprodutivo [Aronszajn, 1950] (RKHS) $\mathcal{H}_k$. Os elementos do RKHS são funções que admitem uma expansão

$$f(x) = \sum_i \alpha_i k(x, x_i), \tag{5}$$

onde

$$k(x, x') := \langle \Phi(x), \Phi(x') \rangle = \langle \widetilde{x}, \widetilde{x}' \rangle$$

é a função de kernel definido positivo, correspondente ao produto escalar das imagens de seus argumentos através de um mapeamento não-linear $\Phi_k : \mathcal{X} \to \mathcal{H}_k$, e associada com operador autoadjunto $\widetilde{D}D$ como sua função de Green. Assim, o termo Q através do operador D unicamente determina o kernel k e seu espaço RKHS $\mathcal{H}_k$ associado de funções (5). É possível mostrar que as funções (5) podem ser evoluídas como produto escalar

$$f(x) = \langle k(x, \cdot), f \rangle_{\mathcal{H}_k} \tag{6}$$

em RKHS ou na forma geral

$$f(x) = \left\langle \widetilde{x}, \widetilde{f} \right\rangle \tag{7}$$

em qualquer outro espaço de Hilbert, isomórfico ao $\mathcal{H}_k$, onde

$$\widetilde{f} := \sum_i \alpha_i \widetilde{x}_i,$$

é a imagem da hipótese f em seu espaço característico induzido por Φ_k. Desta maneira, uma classe de funções não-lineares (5) que contêm soluções do problema de aprendizagem, colocada na forma da regularização (4), pode ser representada no

espaço de hipóteses lineares $\mathcal{H}_k$, cujos associados hiperplanos de separação (decisão) $\left\langle \widetilde{x}, \widetilde{f} \right\rangle = 0$ são determinados através dos coeficientes de expansão α_i, $i = 1, \ldots, N$. Isto, por sua vez, permite uma extensão dos algoritmos e modelos lineares de aprendizado a uma ampla variedade de funções não-lineares de maneira eficiente através de kernels.

Assim, a escolha do kernel corresponde à determinação do espaço característico para uma máquina de aprendizagem linear. No ponto de vista da regularização, a penalização por termo regularizador Q implica a suavidade da função através das propriedades espectrais do operador D, garantindo a solução única do problema com a restrição de capacidade do conjunto de hipóteses. No espaço característico $\mathcal{H}_k$ ou outro espaço de Hilbert isomórfico dele, o termo

$$Q[f] = \|Df\|^2 = \|\widetilde{f}\|^2 = \|f\|_k^2$$

também corresponde ao quadrado de comprimento do vetor normal $\widetilde{f}$ do hiperplano de separação, determinando o inverso de margem geométrica dele. Desta forma, a regularização se relaciona com o conceito de maximização da margem, que é uma base da classe de algoritmos, bem-conhecidos como SVM [Cortes and Vapnik, 1995].

Na interpretação Bayesiana de minimização do risco regularizado (4), a escolha da função de perda $l(x, y, f(x))$, o valor do parâmetro de regularização λ, e o termo $Q[f]$ (chamado de prior) correspondem ao fornecimento de informações a priori sobre o modelo de ruído, sua variância, e distribuição de probabilidade das hipóteses [Girosi, Jones, and Poggio, 1993], respectivamente. Assim, considerando uma escolha a priori de $l(x, y, f(x))$ e $Q[f]$ (ou seu correspondente kernel k) a solução do problema de aprendizagem pela minimização de (4) pode ser unicamente determinada pelo escolha de λ como

$$f_\lambda = \mathrm{KM}(Z_{\mathrm{tr}}^N, R_{\mathrm{emp}}, \lambda Q[\cdot]), \tag{8}$$

onde KM é o resultado de um algoritmo de kernel genérico que fornece o extremo de (4) dado um conjunto de treinamento Z_{tr}^N e os termos R_{emp} e Q. Desta maneira, o problema se reduz à estimação do hiper-parâmetro λ pelo processo de seleção de modelo

$$\lambda = \arg \min_{\lambda \in \mathbb{R}^+} \zeta\left(f_\lambda\right) \tag{9}$$

através de um critério ζ, que também corresponde a implementação de princípio SRM quando ζ é uma estimativa do risco esperado.

Devido à incerteza associada com R_{emp} e Q, de maneira geral, as suas escolhas são efetuadas junto com λ através de uma busca estendida

$$(\theta_R, \theta_Q) = \arg \min_{(\theta_R, \theta_Q) \in \Theta} \zeta\left(f_{\theta_R, \theta_Q}\right) \tag{10}$$

no espaço Θ de múltiplos hiperparâmetros θ_R e θ_Q correspondentes à função de perda e prior Q (incluindo o parâmetro de regularização), respectivamente, que determinam a hipótese estendida na forma

$$f_{\theta_R,\theta_Q} = \mathrm{KM}(Z^N_{\mathrm{tr}}, R_{\mathrm{emp}\theta_R}[\cdot], Q_{\theta_Q}[\cdot]),$$

semelhante a (8). Na prática, é comum escolher $l(x, y, f(x))$ empiricamente, quando λ e os hiperparâmetros de Q (que são os parâmetros de kernel) são estimados através de (10). Neste caso, somente o processo da estimação de λ corresponde à escolha da capacidade do espaço de hipóteses de uma estrutura aninhada em $\mathcal{H}_k$ induzida com Q, que corresponde a uma implementação do SRM, enquanto a estimação dos hiperparâmetros do prior Q é considerada como um nível de inferência mais alto [Guyon, Saffari, Dror, and Cawley, 2010].

Por outro lado, a parametrização do prior pode ser considerada como uma extensão do espaço de hipóteses da máquina de aprendizagem ao conjunto de múltiplos RKHSs onde, conforme princípio SRM, somente uma busca unidimensional é necessária para determinar o equilíbrio entre o risco empírico e complexidade do modelo. Assim, a busca (10) é considerada redundante cujo espaço pode ser reduzido, reduzindo também a incerteza do problema de aprendizagem. Isto levará ao aumento de confiança dos parâmetros estimados e, como sua conseqüência, o aumento de qualidade da generalização obtida. Este ponto justifica a necessidade de desenvolvimento dos novos métodos e modelos de aprendizagem.

O **Capítulo 3** introduz os conceitos básicos de otimização multi-objetivo e determina os elementos principais da abordagem desenvolvida, tais como formulação do problema de otimização e o método determinístico da sua solução.

Visualizando o problema de aprendizado supervisionado como um processo de decisão multicritério, é possível formular um procedimento geral para solução do problema de acordo com o esquema seguinte: avaliar o conjunto de todas alternativas Pareto-ótimas e tomar decisão de escolha de um único elemento. O primeiro passo corresponde à redução da região de incerteza do domínio do problema que, no caso de objetivos conflitantes, representa o *trade-off*: aumento do nível de satisfação de um objetivo exige a redução do nível de satisfação do outro.

No contexto de SEM, em domínio de um RKHS $\mathcal{H}_k$, os objetivos conflitantes são o risco empírico R_{emp} e o prior Q (que desempenha o papel da medida de complexidade do modelo), enquanto a região do *trade-off* corresponde ao conjunto Pareto [Pareto, 1896] do problema de minimização bi-objetivo

$$\min_{f\in\mathcal{H}_k} \phi[f] = (R_{\mathrm{emp}}[f], Q[f]), \tag{11}$$

onde ϕ é o vetor-funcional minimizado.

Formalmente, o conjunto Pareto de um problema de minimização multi-objetivo de ϕ no domínio Ω é dado por conjunto não-dominado,

$$\mathcal{P}(\Omega, \phi) := \left\{ x \in \Omega \mid \forall x' \in \Omega : x \overset{\phi}{\preceq} x' \right\}, \tag{12}$$

onde a relação $x \overset{\phi}{\preceq} x'$ significa "x domina x' com relação a ϕ" e corresponde à uma relação de ordem lexicográfica das imagens de x e x' sobre ϕ.

No contexto do problema bi-critério (11), é possível mostrar que obtenção dos elementos Pareto-òtimos $\mathcal{P}(\mathcal{H}_k, \phi)$ pelo método de soma ponderada [Geoffrion, 1968] equivalente a minimização do risco regularizado (4) para os todos valores de $\lambda \in \mathbb{R}^+$, enquanto a decisão correspondente ao mínimo de ζ em $\mathcal{P}(\mathcal{H}_k, \phi)$ é equivalente a seleção de modelo na forma (9). Na mesma maneira, é possível superar a desvantagem da busca estendida (10) com a abordagem multi-objetivo, implementando o princípio de SRM em um espaço de hipóteses estendido de múltiplos RKHSs com uma busca dentro do correspondente conjunto Pareto.

Especialmente, introduzindo uma família de kernels

$$K \subset \left\{ k \in \mathbb{R}^{\mathcal{X}^2} \right\}$$

ao seu correspondente espaço de hipótese

$$\mathcal{H}_K := \bigcup_{k \in K} \mathcal{H}_k, \tag{13}$$

induzido pela união dos associados RKHSs, é possível substituir (10) com o procedimento multi-objetivo

$$f_{\text{mobj}} = \arg \min_{f \in \mathcal{P}(\mathcal{H}_K, \phi)} \zeta[f], \tag{14}$$

que corresponde à uma máquina de aprendizagem com a capacidade de escolha de kernel (e seus hiperparâmetros) implementando o princípio de SRM. A forma proposta de solução do problema de aprendizagem pode considerada como uma evolução da abordagem chamado MOBJ, desenvolvida nos trabalhos recentes [Teixeira, Braga, Takahashi, and Saldanha, 2000; Costa, Braga, Menezes, Teixeira, and Parma, 2003; Costa and Braga, 2006].

Contudo, para colocar a abordagem MOBJ proposta em (14) na prática, é necessário resolver dois problemas.

Primeiramente, precisa-se definir certa medida de complexidade Q no espaço estendido $\mathcal{H}_K$. Infelizmente, como é demonstrado neste trabalho, para um caso geral de família de kernels K, a medida de complexidade não pode ser um prior e, por isso,

é necessário um tratamento especial na sua derivação. Este problema está sendo tratado nos Capítulos 4 e 5 com duas abordagens diferentes, cada uma correspondente ao seu resultado independente.

Em seguida, é necessário desenvolver um método de obtenção das hipóteses Pareto-ótimas $\mathcal{P}(\mathcal{H}_K, \phi)$ de uma forma determinística. Devido à não-convexidade do domínio $\mathcal{H}_K$ e, conseqüentemente, do problema multi-objetivo, a sua solução não pode ser obtida pela aplicação do método de soma ponderada cuja aplicação é adequada somente para problemas estritamente convexos [Das and Dennis, 1997]. Contudo, a aplicação do método ϵ-restrito [Haimes, Lasdon, and Wismer, 1971; Chankong and Haimes, 1983] é possível, mas não eficiente devido à necessidade de busca dos mínimos globais. Então, como solução foi proposto um método de decomposição dos conjuntos não dominados que permite decompor o problema MOBJ em conjunto de sub-problemas convexos na forma,

$$\mathcal{P}(\mathcal{H}_K, \phi) = \mathcal{P}\left(\bigcup_{k \in K} \mathcal{P}(\mathcal{H}_k, \phi), \phi\right), \tag{15}$$

que possibilita reconstruir o conjunto Pareto $\mathcal{P}(\mathcal{H}_K, \phi)$ no domínio global através dos conjuntos não-dominados $\mathcal{P}(\mathcal{H}_k, \phi)$, $k \in K$ cujos elementos podem ser obtidos pela programação convexa minimizando uma certa forma de (4). Na prática, aproximando K com o número finito de elementos, o método de decomposição permite gerar um subconjunto finito de hipóteses Pareto-ótimos $\mathcal{P}(\mathcal{H}_K, \phi)$ em um tempo garantido na forma determinística.

Assim, a abordagem proposta pode ser considerada como um conceito generalizado de aprendizagem supervisionada para uma classe geral de máquinas de kernel na forma de um algoritmo MOBJ, baseado em procedimento de seleção de modelo multi-objetivo (14) que implementa o princípio SRM dado uma certa medida de complexidade, e cujos resultados podem ser obtidos de maneira eficiente e determinística através da decomposição (15).

No **Capítulo 4**, a medida de complexidade proposta é baseada no conceito de suavidade que, em combinação com a abordagem MOBJ desenvolvida, leva ao algoritmo de aprendizagem para redes RBF.

Em regularização, a complexidade da hipótese é refletida pelo termo regularizador, que no domínio Fourier corresponde ao filtro de passa-alta, implicando certo grau de suavidade à função f. A suavidade de uma função, no entanto, pode ser determinada de maneira explícita, fora do contexto do regularizador (prior, ou seu correspondente kernel). Uma medida de complexidade pode ser obtida para o espaço de hipóteses arbitrário, inclusive para $\mathcal{H}_K$ induzido por uma família K.

Neste trabalho, uma análise de suavidade baseada na norma no espaço de Sobolev é feita para a classe de hipóteses em $\mathcal{H}_k$, associados com os modelos RBF $f : \mathbb{R}^n \to \mathbb{R}$,

dados pela expansão

$$f(x) = \sum_{i=1}^{m} \alpha_i k_\sigma(x, c_i),$$

onde c_i são os centros das funções RBF $k_\sigma(x, c_i) = k_\sigma(x - c_i) = \kappa(\frac{x-c_i}{\sigma})$ de largura σ. Em particular, é mostrado que a norma $\|f\|_{q,p}$ no espaço de Sobolev $\mathbb{W}_{q,p}$ possui o limite superior

$$\|f\|_{q,p} \leq \|\alpha\|_1 \cdot \|k_\sigma\|_{q,p}, \tag{16}$$

onde $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)^T$ é o vetor de coeficientes da expansão (pesos) da rede RBF correspondente ao f. Baseado em uma modificação de (16), a medida de complexidade

$$Q_{\text{rbf}}[f] = \sigma^{-\frac{n}{p}} \|\alpha\|_1 \sum_{|s|=q} \|D^s k_\sigma\|_p, \tag{17}$$

foi proposta, onde

$$D^s := \frac{\partial^{|s|}}{\partial x_1^{s_1} \partial x_2^{s_2} \cdots \partial x_n^{s_n}}$$

é o operador diferencial generalizado sobre o espaço $\mathbb{R}^{\mathbb{R}^n}$, dado pelo multi-índice $s \in \mathbb{Z}^n$.

Supondo que as funções k_σ são Gaussianas, i.e., $\kappa(u) = \exp(-\frac{1}{2}\|u\|^2)$, e a ordem do diferencial da norma de Sobolev $q = 2$, é possível mostrar que a medida de complexidade (16) se reduz à forma simples

$$Q_{\text{rbf}}[f] = \frac{\|\alpha\|_1}{\sigma^2}. \tag{18}$$

Tal forma permite descrever o problema de aprendizagem MOBJ no espaço de hipóteses

$$F := \Big\{ f : \mathbb{R}^n \to \mathbb{R} \mid f(x) = \sum_i \alpha_i k_\sigma(x - c_i) \Big\}, \tag{19}$$

correspondente a todas as possíveis redes RBF com $m \in \mathbb{N}$ funções bases, centros $c_i \in \mathbb{R}^n$, larguras $\sigma \in \mathbb{R}^+$ e pesos $\alpha_i \in \mathbb{R}$. Para obter uma aproximação finita do conjunto Pareto

$$\mathcal{P}(F, \phi), \quad \phi[f] = (R_{\text{emp}}[f], Q_{\text{rbf}}[f]) \tag{20}$$

de maneira eficiente, a redução do domínio F ao

$$\widetilde{F} = \bigcup_{\sigma \in S_\sigma} F_{\sigma, C_M},$$

foi proposta, onde F_{σ,C_M} são os elementos de F associados às redes RBF cujos centros correspondem a padrões distintos do conjunto de treinamento. Assim, aplicando a

decomposição (15) ao $\widetilde{F}$ é possível obter uma aproximação determinística de (20) na forma

$$\mathcal{P}\Big(\widetilde{F},\phi\Big) = \mathcal{P}\left(\bigcup_{\sigma\in S_\sigma} \mathcal{P}(F_{\sigma,C_M},\phi),\phi\right), \tag{21}$$

onde $S_\sigma = (\sigma_j)_j$ é um *grid* de larguras, cuja quantidade de elementos determina a qualidade da aproximação de (20).

Como as estruturas (camadas escondidas) das redes RBF associadas com F_{σ,C_M} são iguais, o problema de busca dos elementos não-dominados $\mathcal{P}(F_{\sigma,C_M},\phi)$ é convexo. Assumindo uma função de perda quadrática $l(x,y,f(x)) = (y-f(x))^2$, este problema pode ser resolvido no domínio $\mathbb{R}^M$ minimizando

$$R_{\text{reg}}(\alpha) = \|Y - H\alpha\|^2 + \lambda\|\alpha\|_1, \tag{22}$$

onde $Y = (y_1, y_2, \ldots, y_N)^T$ é o vetor das saídas desejadas do conjunto de treinamento Z^N_{tr} e $H = \{k_\sigma(x_i, c_j)\}$, $i = 1,\ldots,N$, $j = 1,\ldots,M$ é a matriz $N \times M$ da camada escondida.

Sabe-se que (22) correspondente à regressão LASSO [Tibshirani, 1996], cujas soluções são esparsas e seu caminho de regularização (minimizadores de (22) correspondentes a todos $\lambda \in \mathbb{R}^+$) é uma curva linear em trechos no $\mathbb{R}^M$. O caminho de regularização do LASSO, que no contexto do presente problema representa o conjunto não-dominado $\mathcal{P}(F_{\sigma,C_M},\phi)$, pode ser inteiramente calculado através do algoritmo LARS [Efron, Hastie, Johnstone, and Tibshirani, 2004]. Assim, a aproximação do conjunto Pareto (20) pode ser obtida diretamente de (21) através dos correspondentes resultados de múltiplas execuções do LARS. Assim, a utilização do conjunto (21) em um esquema (14) da abordagem MOBJ, levou ao algoritmo MOBJ-RBF desenvolvido, que é capaz de determinar os pesos, larguras, centros e quantidades de funções-bases das redes RBF de acordo com principio SRM.

Em combinação com os critérios de informação AIC [Akaike, 1974] e BIC [Schwarz, 1978], adaptados para seleção de modelo no contexto de regressões LASSO, o algoritmo proposto demonstrou alto desempenho de generalização para diversos problemas comumente usados como *benchmark*.

O **Capítulo 5** apresenta uma extensão multi-objetivo do conceito de maximização da margem para o contexto do espaço de hipóteses $\mathcal{H}_K$, associado aos múltiplos RKHSs.

Como foi mostrado na análise préliminar do problema, a escolha da medida de complexidade em $\mathcal{H}_K$ como o prior $Q[f] = \|f\|^2_k$ (relacionado com a largura da margem geométrica num subespaço $\mathcal{H}_k$ correspondente), não é adequada para o caso de uma família de kernels K arbitrária. Assim, os conjuntos aninhados

$$\Omega_i := \{f \in \mathcal{H}_K \mid \|f\|^2_k \leq \epsilon_i\},$$

induzidos pela medida de complexidade $Q[f] = \|f\|_k^2$, não necessariamente correspondem à ordem das suas capacidades. Além disso, a ordem induzida é altamente dependente da normalização dos kernels em K, enquanto o espaço das funções, representadas por $\mathcal{H}_K$, continua sendo independente. Desta maneira, a organização de K influencia o conjunto de hipóteses Pareto-ótimas sem influenciar suas propriedades da generalização. Assim, $\|f\|_k^2$ não representa uma medida de complexidade adequada em $\mathcal{H}_K$.

Do ponto de vista dos espaços de característica, a violação do SRM com a escolha $Q[f] = \|f\|_k^2$ é causada pelas diferenças de suas topologias, que levam à incompatibilidade das métricas. Assim, o problema pode ser resolvido por certa equiparação dos espaços. Especialmente, uma das técnicas de equiparação desenvolvida é a normalização dos tamanhos dos vetores característicos, que leva à seguinte medida de complexidade normalizada

$$Q_{\text{norm}}[f] := \sum_i \sum_j \alpha_i \alpha_j \sqrt{k(x_i, x_i) k(x_j, x_j)} k(x_i, x_j). \tag{23}$$

No caso específico, quando os vetores característicos associados ao kernel k têm comprimentos iguais (e.g., como no caso de kernel RBF Gaussiano), o valor do $Q_{\text{norm}}[f]$ corresponde ao quadrado de norma RKHS de uma hipótese f', cujo kernel induz vetores característicos unitários, equivalente ao f. No caso do kernel geral, a hipótese f' não necessariamente é equivalente a f, mas a medida $Q_{\text{norm}}[f]$ é invariante aos escalamentos dos espaços de característica associados com f.

De fato, é possível demonstrar que a capacidade do espaço RKHS é influenciada não somente pelos comprimentos de vetores característicos, mas também pela sua topologia angular. Por exemplo, é fácil verificar que os kernels suaves induzem vetores característicos fracamente angulados. Assim, a equiparação dos espaços de característica a partir da normalização dos vetores característicos pode não ser suficiente para representar as complexidades das hipóteses na forma adequada. Por isso, foi proposta outra abordagem de equiparação, baseada na idéia de representar todas as hipóteses em único espaço de características de forma equivalente. Esta técnica denominada equalização, é formulada a partir do princípio a seguir.

Dado um mapeamento $\Phi^\circ : \mathcal{X} \to \mathcal{H}^\circ$ fixo ao espaço de referência $\mathcal{H}^\circ$ assumimos que é possível representar qualquer kernel $k \in K$ na forma do produto escalar em $\mathcal{H}^\circ$

$$k(x, x') = \langle \Phi^\circ(x), \Phi_k^*(x') \rangle_{\mathcal{H}^\circ} = \langle \mathring{x}, \mathring{x}' \rangle \text{ para os todos } (x, x') \in \mathcal{X}^2, \tag{24}$$

onde $\Phi_k^*(x') : \mathcal{X} \to \mathcal{H}^\circ$ é um mapeamento auxiliar associado com k. Assim, qualquer hipótese $f \in \mathcal{H}_K$ pode ser representada na forma equivalente pelo produto escalar no

espaço de características de referência como

$$f(x) = \left\langle \mathring{x}, \overset{*}{f} \right\rangle \tag{25}$$

através da sua imagem auxiliar

$$\overset{*}{f} = \sum_i \alpha_i \overset{*}{x}_i. \tag{26}$$

Na forma da hipótese (7), ambos componentes do produto escalar são dependentes do mapeamento Φ_k. Em contrapartida, a forma equivalente (25) considera os vetores característicos de referência $\mathring{x}$, independentes do kernel e comuns para todas as hipóteses. Assim, os elementos $\mathcal{H}_K$ podem ser analisados num espaço de características equalizado, com as suas propriedades expressas em termos das correspondentes imagens auxiliares $\overset{*}{f}$.

A análise do paradigma das hipóteses no espaço de referência mostrou que a margem geométrica do hiperplano de separação corresponde ao inverso do comprimento do vetor $\mathring{g}$, que representa a projeção ortogonal de $\overset{*}{f}$ no espaço linear $S^\circ := \text{span}\{\mathring{x}_i\}_{i=1}^N$, gerado pelas combinações lineares dos vetores característicos de referência. Por outro lado, foi demonstrado que a introdução da medida de complexidade baseada na largura da margem, tal como $\|\mathring{g}\|^2$ ou $\|\overset{*}{f}\|^2$, leva a degenerações do espaço de hipóteses da máquina de aprendizagem e, por isso, não representa a escolha aceitável. Esta conclusão justificou a necessidade do desenvolvimento de uma extensão do conceito de margem geométrica a uma propriedade de hiperplano de separação mais geral.

Seguindo uma conhecida interpretação de margem larga como a estabilidade paramétrica do modelo, o critério de estabilidade *leave-one-out* do hiperplano de separação foi proposto. Particularmente, dado um hiperplano de separação pelo seu vetor normal f na forma da combinação linear

$$f = \sum_{j=1}^{N} \alpha_j x_j \tag{27}$$

dos N vetores-suporte x_j, $j = 1, \ldots, N$, a sua estabilidade é definida pelo critério

$$E(f) := \sum_{i=1}^{N} e_i^2, \tag{28}$$

onde

$$e_i^2 = \|f - g^{(i)}\|^2 \tag{29}$$

é obtido por mínimos quadrados da distância entre o vetor original f e a sua aproximação

$$g^{(i)} := \sum_{j=1, j\neq i}^{N} \beta_j^{(i)} x_j$$

considerando os $N-1$ vetores-suporte, após a exclusão do i-ésimo. A partir da transformação realizada, foi mostrado que o critério (28) possui a forma compacta

$$E(f) = \sum_{i=1}^{N} \frac{\alpha_i^2}{d_i}, \tag{30}$$

onde $(d_1, d_2, \ldots, d_N) = \mathrm{diag}(G^{-1})$ são os elementos diagonais do inverso da matriz de Gram $G = X^T X$, associada com os vetores-suporte $X = (x_1, x_2, \ldots, x_N)$.

É possível mostrar, que o critério proposto (30) está diretamente ligado a norma $\|f\|^2$ (que reflete a largura da margem geométrica) e, além disso, considera o volume de informação mútua na descrição do vetor f, no sistema de vetores-suporte dado por X. Desta forma, o critério (30) pode ser interpretado como uma medida de complexidade dentro do conceito de *minimum description length* (MDL) [Wallace and Boulton, 1968], sendo uma medida do volume de informação necessário para descrição do modelo de um classificador, representado por f, com certa precisão.

Como o critério (30) também é dependente da métrica num espaço de Hilbert, associado com f, a medida de complexidade baseada em (30) para o espaço de hipóteses $\mathcal{H}_K$ exige a equiparação dos correspondentes RKHSs. Assim, aplicando a técnica de equalização desenvolvida, foi proposta a medida de complexidade de referência

$$Q_{\mathrm{ref}}[f] := \sum_{i=1}^{N} \frac{\alpha_i^2}{d_i}, \tag{31}$$

que é baseada no critério de estabilidade do hiperplano de separação dos padrões em S° associado com $f \in \mathcal{H}_k$. Em (31), os elementos diagonais

$$(d_1, d_2, \ldots, d_N) = \mathrm{diag}\big(G_k^{-1} G_{k^\circ} G_k^{-1}\big)$$

são calculados a partir das matrizes de Gram G_k^{-1} e G_{k°, associadas com o kernel k e o kernel de referência k°, respectivamente.

O kernel de referência k° é associado com Φ° através dos produtos escalares dos vetores característicos $\mathring{x}_i$. Por isso, a sua escolha influencia na medida de complexidade Q_{ref}. Se por um lado, a escolha de um Φ° deve garantir a existência dos mapeamentos auxiliares Φ_k^* para todos os elementos da família K, fornecendo a possibilidade de representação equivalente de todas as hipóteses em $\mathcal{H}_K$ na forma (25); por outro lado,

é possível mostrar que a projeção ortogonal $\mathring{g}$ de $\overset{*}{f}$ no espaço linear S° não depende do mapeamento Φ_k^*. Apesar disso, a análise feita no Apêndice A mostra a existência dos mapeamentos auxiliares e a abordagem prática para suas derivações, inclusive nas formas fechadas dos kernels associados. Utilizando estes resultados, foi proposta a escolha universal de mapeamento de referência Φ° com o kernel

$$k^\circ(x_i, x_j) = \delta_{ji} = \begin{cases} 1, & i = j \\ 0, & \text{se contrário} \end{cases}$$

sendo a função delta de Kronecker, que leva a uma forma prática de medida Q_{ref} para qualquer $\mathcal{H}_K$.

Para confirmar a adequação da abordagem desenvolvida, um algoritmo MOBJ no domínio de soluções de SVM, baseado no esquema (14) com as medidas de complexidade propostas (23) e (31) foi comparado com a busca (10) pelo *grid* equivalente, numa série de experimentos extensivos. A análise estatística dos resultados obtidos com experimentos confirmou as propriedades esperadas da medida Q_{ref}, juntamente com a redundância das buscas exaustivas (10) no espaço de hiperparâmetros, como foi teoricamente previsto no Capítulo 2. Este resultado justifica a necessidade de desenvolvimentos futuros da abordagem MOBJ.

O **Capítulo 6** conclui o presente trabalho resumindo os resultados teóricos e práticos obtidos, junto com a discussão das perspectivas de desenvolvimentos futuros.

Os resultados dos Capítulos 3 e 4 desse trabalho foram apresentados em congressos nacionais e internacionais, e também publicados em periódicos revisados. Os resultados de Capítulo 5 encontram-se em fase de preparação para submissão.

Produção bibliográfica

I. Kokshenev and A. P. Braga. Complexity bounds of radial basis functions and multi-objective learning. Em *ESANN*, páginas 73–78, 2007.

I. Kokshenev and A. P. Braga. A multi-objective approach to RBF network learning. *Neurocomputing*, 71(7-9):1203–1209, 2008a.

I. Kokshenev and A. P. Braga. A multi-objective learning algorithm for rbf neural network. Em *10th Brazilian Symposium on Neural Networks, SBRN 2008*, páginas 9–14, 2008b.

I. Kokshenev and A. P. Braga. An efficient multi-objective learning algorithm for RBF neural network. *Neurocomputing*, 2010. (aceito)

Federal University of Minas Gerais (UFMG)
Post-graduate Programme in Electrical Engineering (PPGEE)

Multi-objective Learning of Radial-Basis Function Networks and Kernel Machines

Illya Kokshenev

A dissertation thesis submitted to the examination committee in partial fulfillment of the requirements for the degree of doctor in Electrical Engineering.

Supervisor: Prof. Dr. Antônio P. Braga.

Belo Horizonte — June 2010

To my parents

Acknowledgements

I am most grateful to my supervisor, Antônio de Padua Braga, Dr., Professor of the Electrical Engineering Department and Head of the Computational Intelligence Laboratory (LITC) at the Federal University of Minas Gerais (UFMG), for his valuable guidance, advice, discussion, and support throughout this work.

I am gratefully indebted to Yevgeniy Bodyanskiy, Dr. Sci., Professor of the Artificial Intelligence Department and Head of the Control Systems Research Laboratory (CSRL) at the Kharkiv National University of Radio Electronics, Ukraine. His wide knowledge and encouraging guidance have been of great value for me since my very first steps in science and during my bachelor's and master's graduation.

I owe my most sincere gratitude to Hani Camille Yehia, Dr., Professor and ex-Coordinator of the Post-Graduate Programme of the Electrical Engineering Department, Head of the Center for Research on Speech, Acoustics, Language and Music (CEFALA) at the UFMG, for his enthusiasm, openness and welcome in my application for doctorate programme.

My sincere thanks are due to Petr Ekel, Dr. Sci., Professor of the the Post-Graduate Programme of the Electrical Engineering Department at the Pontifical Catholic University (PUC) of Minas Gerais, for his heplful advice and support in a number of ways since my first days in Brazil.

I am thankful to many of my colleagues from LITC and CSRL for their enthusiasm in discussion of ideas, exchange of experience, and assistance.

I also thank the National Council for Scientific and Technological Development of Brazil (CNPq), for their generous financial support.

Last, but not least, I wish to thank my family: my father for sharing his scientific and life experience, my mother for her encouragement in all my efforts, Natalia for time she spared for me during several years, and Janaina for her inestimable support in finalizing this work.

Abstract

As known from statistical learning theory, the training error and complexity of a model must be simultaneously minimized and yet certainly balanced for a valid generalization. Modern learning algorithms, such as support vector machines, achieve this goal by means of regularization and kernel methods, whose combination provides possibilities for analysis and construction of efficient nonlinear learning machines.

In such algorithms, due to the non-convexity of the learning problem when the kernel is not fixed, the choice of the kernel is commonly addressed using sophisticated techniques of model selection, in a manner, different from the original idea of balance between the error and complexity. In contrast, the search of balance between the error and complexity in non-convex learning problems can be treated within the multi-objective framework, by viewing the supervised learning as a decision process in the environment of two conflicting goals. However, modern methods of multi-objective learning are focused on evolutionary optimization, paying a few attention to implementation of key learning principles.

This work develops a multi-objective approach to supervised learning as an extension of the traditional (single-objective) concepts, such as regularization and margin maximization, to the cases of non-convex hypothesis spaces, induced with multiple kernels. In the proposed learning scheme, approximate solutions to generally non-convex problems are obtained from their decompositions into the subsets of convex subproblems, where the application of deterministic nonlinear programming is efficient. Aiming for implementation of the principle of structural risk minimization, there are several complexity measures derived, each one inducing a particular multi-objective algorithm.

In particular, the proposed smoothness-based complexity measure for the Gaussian radial-basis function (RBF) networks led to an efficient multi-objective algorithm, which is capable of finding the weights, widths, locations, and quantities of basis functions in a deterministic manner. In combination with the Akaike and Bayesian information criteria, the developed algorithm demonstrates a high generalization efficiency on several synthetic and real-world benchmark problems. Aiming to extend the concept of margin maximization to supervised learning with multiple kernels, the techniques of feature normalization and equalization were proposed. The further analysis shows the necessity in extension of the concept of margin to the more general property of a separation hyperplane, such as its stability. As the result, the proposed stability-based complexity measure, which reliability has been experimentally confirmed, allows a construction of multi-objective algorithms for arbitrary classes of kernels.

Keywords: *multi-objective*, *model selection*, *regularization*, *kernel machines*, *radial-basis functions*.

Resumo

Conforme a teoria de aprendizagem estatística, o erro de treinamento e a complexidade de modelos de aprendizado devem ser certamente equilibrados para uma generalização válida, além de serem minimizados. Os algoritmos de aprendizagem modernos, tais como máquinas de vetores de suporte, atingem esta meta por meio da regularização e dos métodos de kernel. A sua combinação permite de maneira eficiente analisar e construir máquinas de aprendizagem não-lineares.

Nestes algoritmos, devido à não-convexidade do problema de aprendizagem quando o kernel não é fixo, a escolha do kernel é efetuada por meio das técnicas sofisticadas de seleção de modelos, diferentemente da ideia original de equilíbrio entre o erro e a complexidade. Por outro lado, a busca de equilíbrio entre o erro e a complexidade de problemas não-convexos pode ser tratada de maneira multi-objetiva, considerando a aprendizagem supervisionada como o processo de decisão no ambiente de dois objetivos conflitantes. Contudo, métodos modernos de aprendizagem multi-objetiva são voltados à otimização evolucionária, prestando pouca atenção à implementação dos princípios fundamentais de aprendizagem estatística.

Neste trabalho foi desenvolvida uma abordagem multi-objetiva de aprendizagem supervisionada baseada na extensão dos conceitos tradicionais, tais como regularização e maximização de margem, aos casos de espaços de hipótese não-convexos, induzidos com múltiplos kernels. No esquema de aprendizagem proposto, as soluções aproximadas dos problemas, geralmente não-convexos, sao obtidos por meio de certa decomposição em conjuntos de sub-problemas convexos, nos quais a programação não linear pode ser eficientemente aplicada de maneira determinística. Com o objetivo de implementação do princípio de minimização do risco estrutural, várias medidas de complexidade foram propostas, induzindo os correspondentes algoritmos multi-objetivos.

Entretanto, a medida de complexidade baseada em suavidade para as redes de função da base radial (RBF) permitiu a construção de um algoritmo multi-objetivo, com a sua capacidade de definição dos pesos, larguras, centros e quantidades de funções-bases. Em combinação com os critérios de informação de Akaike e Bayes, o algoritmo proposto demonstrou um alto desempenho de generalização em vários problemas-testes de natureza diversa. Com o objetivo de extensão do conceito de maximização de margem ao aprendizagem supervisionada com múltiplos kernels, as técnicas de normalização e equalização dos espaços de características foram propostas. As suas análises mostraram a necessidade de formulação de conceito de margem com uma característica mais geral de hiperplano de separação, tal como sua estabilidade. Como resultado, a medida de complexidade baseada no critério de estabilidade desenvolvido, cuja adequação foi confirmada com experimentos, permite a construção de algoritmos multi-objetivos para as classes de kernel arbitrários.

Palavras-chave: *multi-objetivo*, *seleção de modelo*, *regularização*, *máquinas de kernel*, *redes RBF*.

Contents

List of Figures

List of Tables

Abbreviations and symbols

Abbreviations

AIC	Akaike information criterion
ANN	artificial neural network
BIC	Bayessian information criterion
CV	cross-validation
GCV	generalized cross-validation
GRN	generalized regularization network
i.i.d.	independent and identically distributed
LASSO	least absolute shrinkage and selection operator
LARS	least angle regression shrinkage
MAP	maximum *a posteriori* probability
MLP	multilayer perceptron
MOBJ	bi-objective supervised learning (concept, method, or algorithm)
MOML	multi-objective machine learning
MSE	mean squared error
MVE	minimum of validation error
NRMSE	normalized root mean squared error
OLS	ordinary least squares
p.d.f.	probability density function
PCA	principal component analysis
RBF	radial-basis function
RKHS	reproducing kernel Hilbert space
RN	regularization network
RMSE	root mean squared error

ROFS	regularized orthogonal forward selection
SV	support vector
SVM	SV machine
SLT	statistical learning theory
SRM	structural risk minimization
VC	Vapnik-Chervonenkis (-dimension or -theory)

Important Symbols

b	bias parameter
c	usually a constant
c_i	center (prototype) of i-th basis function
D	linear differential operator
$\widetilde{D}$	adjoint of D
df$[f]$	effective number of degrees of freedom of hypothesis f
e_i	the i-th error, usually $e_i = y_i - \hat{y}_i$
$E[x]$	expected value of random variable x
$E(f)$	leave-one-out stability criterion (hyperplane error)
$f(x)$	hypothesis, discriminant, or regression function
$\widetilde{f}$	feature space representation of hypothesis $f(x) = \left\langle \widetilde{x}, \widetilde{f} \right\rangle$
$\mathcal{F}$	hypothesis space and of RBF networks
$F[\kappa](\omega)$	Fourier transform of κ on $\mathbb{R}^n$
G_k	Gram matrix associated with kernel k
$G(x, x_i)$	Green's function to $\widetilde{D}D$
$G(N)$	growth function, upper bound on Vapnik's entropy
h	VC-dimension
H	design or kernel matrix
$\mathcal{H}$	Hilbert space
$\mathcal{H}_k$	RKHS associated with k
i	index variable or imaginary unit
I	identity matrix
$k(\cdot, \cdot)$	kernel function
$k_\sigma(x - c)$	convolution kernel with the width σ, usually a Gaussian RBF

K	family of available kernels
ℓ_p	space of $p-$summable sequences
L_p	space of $p-$integrable functions
m	number of basis functions (hidden units) or support vectors
n	dimensionality of input space
N	number of training samples (length of the training set)
$\mathbb{N}$	set of natural numbers
$\mathcal{O}(g(x))$	asymptotic growth rate of $g(x)$ (Bachmann-Landau notation)
$\mathcal{P}(\Omega, \phi)$	nondominated or Pareto set of Ω w.r.t. ϕ
$\Pr\{X\}$	probability of event X
$p(x, y)$	joint probability density function of distribution $P(x, y)$
$P(x, y)$	joint probability distribution of random variables x and y
$Q[f]$	regularization stabilizer, prior, or complexity measure of f
$R[\cdot]$	expected risk functional
$\mathbb{R}$	field of real numbers
$\mathbb{R}^+$	set of non-negative real numbers
$R_{\text{emp}}[\cdot]$	empirical risk functional
$R_{\text{reg}}[\cdot]$	regularized risk functional
$\text{tr}(H)$	trace of the square matrix H
X	set of input patterns
$\mathcal{X}$	space of input patterns, usually $\mathbb{R}^n$
x_i	i-th training pattern from X
$\widetilde{x}_i$	i-th feature, usually is the shorthand for $\widetilde{x}_i = \Phi(x_i)$
$\mathring{x}_i$	i-th reference feature, $\mathring{x}_i = \Phi^\circ(x_i)$
$\overset{*}{x}_i$	i-th auxiliary feature, $\overset{*}{x}_i = \Phi^*(x_i)$
Y	training vector of target values
$\mathcal{Y}$	space of target values
y_i	i-th training target value from Y
$\hat{y}_i$	response to i-th training input pattern
$\hat{Y}$	vector of model response to the set input training patterns
$\mathbb{Z}$	set of integers
Z_{tr}^N	training set of N pairs from $\mathcal{X} \times \mathcal{Y}$

α, α_i	vector of expansion coefficients (weights) and its component
γ	bandwidth of the kernel
$\delta_x(\cdot)$	Dirac's delta function centered at x, $\delta_x(x') = \delta(x - x')$
ϵ	constraint parameter
$\zeta[\cdot]$	model selection criterion
θ, Θ	vector and domain of hyperparameters, $\theta \in \Theta$
$\kappa(\cdot)$	translation-invariant kernel generating function
λ	regularization parameter
λ_j	j-th eigenvalue
$\mu(\cdot)$	effective feature vector
$\nu(\cdot)$	effective support vector
$\varrho[f]$	geometrical margin of the hyperplane associated with f in the feature space
$\rho\,(\Omega, \phi)$	Pareto front of Ω w.r.t. ϕ
σ	width of the kernel
$\varphi_j(\cdot)$	j-th eigenfunction
ϕ	vector-objective functional, usually $\phi[f] = (R_{\text{emp}}[f], Q[f])$
$\Phi(\cdot)$	non-linear feature map
$\Phi^{\circ}(\cdot)$	reference feature map
$\Phi^{*}(\cdot)$	auxiliary feature map
ω	angular frequency
Ω	domain of the multi-objective problem (decision space)
$\langle \cdot, \cdot \rangle$	dot product
$\| \cdot \|$	2−norm or Euclidean norm in ℓ_2 or L_2
$\| \cdot \|_p$	p−norm
$\| \cdot \|_k$	norm in RKHS of k, where k is a kenel
$\preceq$	lexicographical order relation
$\overset{\phi}{\preceq}$, $\overset{\phi}{\prec}$	weak and strict dominance relations w.r.t. ϕ, respectively
hull$\{X\}$	convex hull of X
span$\{X\}$	linear span of X
sign$(\cdot)$	signum function

Chapter 1

Introduction

In the middle of the twentieth century, the innovative work [McCulloch and Pitts, 1943] brought a new way of understanding and modeling of cognitive processes with the connectionist approach, known as artificial neural networks (ANNs). With the rapid development of computers in the mid-80s, ANNs receive much attention from scientists as a powerful computational intelligence tool, whose evolution furthered the development of machine learning as a discipline.

Recent advances in machine learning embrace a wide range of computational intelligence tasks, whose ever-growing demand arises from a variety of modern problems in engineering, economics, and bio-medicine. Most of these tasks are associated with induction of models from data. For instance, such tasks as pattern recognition, time-series prediction, adaptive control, and fault detection may be formulated in terms of a search for hidden dependencies in empirical observations. Within the machine learning framework, the latter corresponds to a setting of the supervised learning problem, which plays a central role in the discipline.

Since the empirical observations are induced by objects whose properties are generally unknown, the solution to a supervised learning problem requires dealing with nondeterministic processes under conditions of uncertainty. That in turn relates the methodology complex of machine learning to other fields, such as statistics, mathematical programming, decision, and information theories.

A comprehensive look at learning is given by the well-known Statistical Learning Theory (SLT) [Vapnik, 1998], which establishes the methodology of consistent learning by the key principles of empirical and structural risk minimization (ERM and SRM). These principles are closely related to regularization [Poggio and Girosi, 1990] and Bayesian learning [Neal, 1996], whose combination with the support vector (SV) machines [Cortes and Vapnik, 1995] and modern kernel methods [Scholkopf, 1999] represents the state-of-the-art machine learning framework.

Within the context of supervised learning in its traditional formulation, one ensures the consistency of learning by controlling the training error (empirical risk) and model complexity (capacity of the hypothesis class) via minimization of a certain loss function. In such a way, one implements the principle of SRM by maintaining the error and complexity in a certain balance, determined with the choice of hyperparameters (e.g., regularization parameter). The latter provides *a priori* information about the solution to the learning algorithm, which is usually unavailable due to uncertainty. Hence, the problem of hyperparameter estimation arises at the next level of inference, referred to as model selection.

With the recent development of evolutionary optimization, an increasing interest has been seen in application of the Pareto-optimality concept to machine learning aiming to extend the capabilities of existing learning models and algorithms. This approach led to the development of multi-objective machine learning (MOML) [Jin, 2006] methods, where learning is viewed as a decision process within the environment of multiple and competitive goals, representing trade-offs. Within the MOML framework, the uncertainty of the supervised learning problem is represented with the trade-off region between minimization of the error and complexity objectives, whereas the decision towards a single solution corresponds to model selection.

1.1 Motivation and goals

When there is a single hyperparameter, both the single-objective and multi-objective concepts are equivalent and correspond to the same implementation of the SRM principle. When there is more than one hyperparameter, the traditional approach

requires several levels of estimation, as recently depicted in [Guyon, Saffari, Dror, and Cawley, 2010] with the unified multi-level inference model of learning. For example, when estimation of the regularization parameter stays at the second level after the model parameters, the estimation of kernel parameters [1] corresponds to the third level of multi-level inference hierarchy. The drawback of a such multi-level scheme is that only transformation of the uncertainty occurs between levels, instead of its reduction, leading to the necessity in exhaustive search within the space of hyperparameters. The typical example is the grid search techniques, widely applied for selection of hyperparameters in combination with the diverse validation criteria.

From the SRM point of view, an introduction of multiple hyperparameters (e.g., kernel parameters) can be considered as an extension of hypothesis space of a learning machine (space of available models), whereas the learning goals (minimum error and model complexity) remain the same. Hence, in the MOML formulation the supervised learning problem remains bi-objective that always requires only two levels of inference: estimation of model parameters and finding the balance between the error and complexity.

It is noteworthy that a similar approach in a single-objective form is also possible. In particular, the so-called multiple kernel learning has been recently developed in [Bach, Lanckriet, and Jordan, 2004; Micchelli and Pontil, 2005; Ong, Smola, and Williamson, 2005], where the variety of kernels is represented by a single hyper-prior, instead of the set of hyperparameters. In such a formulation, the learning problem is convex and solved by a certain form of regularization, where only a single hyperparameter determines the regularization strength. This approach demonstrates implementation of the SRM on the hypothesis space associated with multiple kernels, however, relies on computationally heavy optimization to ensure sparsity of solutions while maintaining the convexity of the problem.

In contrast to regularization, the multi-objective formulation of supervised learning is not limited to convex problems, whereas both permit the SRM implementation. Moreover, for convex problems it can be shown that both approaches are equivalent. Consequently, a unification of supervised learning with the methodology of MOML

[1] Also known as kernel selection, in regularization and SV learning, and selection of the prior, in Bayesian learning.

can be viewed as a generalized learning framework. However, when allowing generally non-convex multi-objective problems, one has to deal with NP-completeness, addressing them with approximation techniques. One of the earliest approaches of such multi-objective treatment of supervised learning was developed in [Liu. and Kadirkamanathan, 1995], where a genetic algorithm was used for finding of approximate solutions to the problem of multi-objective minimization of the training error of a neural network, along with several norms of its weights, playing the role of complexity measures. The later developments [Hatanaka and Uosaki, 2003; Jin, Okabe, and Sendhoff, 2004; Bevilacqua, Mastronardi, Menolascina, Pannarale, and Pedone, 2006; Yen, 2006; Kondo, Hatanaka, and Uosaki, 2006] followed similar ideas, relying on evolutionary programming as the means of approximate solutions of generally non-convex multi-objective problems. Alternatively, application of nonlinear programming techniques has been shown in [Teixeira, Braga, Takahashi, and Saldanha, 2000; Costa, Braga, Menezes, Teixeira, and Parma, 2003; Costa and Braga, 2006] for multi-objective supervised learning of multilayer perceptron (MLP) networks, where their complexities were expressed by the Euclidean norms of their weights, aiming to control the generalization according to [Bartlett, 1997]. Such an approach, called MOBJ, concerns treatment of multi-objective problem on the non-evolutionary basis, taking advantage of deterministic learning algorithms.

The above examples of both evolutionary and non-evolutionary (MOBJ) multi-objective approaches demonstrated good results in practice. Although the evolutionary MOML algorithms are focused on optimization techniques, their theoretical basis is mostly heuristic and weakly connected to learning concepts. On the other hand, the existing MOBJ approach addresses multi-objective problem with regularization-like procedures, whose connections to the SRM could be revealed, but the Pareto-optimality of resulted solutions is weak due to the non-convexity of learning problems addressed with convex programming. Therefore, the evolution of ideas of the MOBJ towards their theoretical groundings from both the SRM and Pareto-optimality points of view represents a novel perspective in the field of MOML.

As known, kernel methods provide the methodology for analysis and construction of various learning machines, such as SV machines, regularization networks (RN), and radial-basis function (RBF) networks; in the aspects of statistical learning, informa-

tion, and optimization theories. Consequently, the above learning machines or their modifications are the appropriate choice for the detailed multi-objective analysis and extension to the SRM-consistent MOBJ framework. Finally, the presented above line of motivations gives rise to the following goals of the current work:

- Develop the methodology of the generalized MOBJ framework for supervised learning, aimed to implement the SRM in a general, deterministic (non-evolutionary) scheme, taking advantages of the nonlinear programming;
- Determine and implement the components of MOBJ learning algorithm for particular neural network architectures;
- Demonstrate reliability of the proposed theoretical basis in practice and outline its further development.

1.2 Thesis outline

The dissertation thesis has the following structure.

Chapter 2 provides overview and systematic analysis of existing results on the statistical and regularization learning, their connections and applications to the RBF networks and kernel machines. The unified view on supervised learning with kernel machines concludes the chapter by a discussion of the problem of estimation of hyperparameters, motivating further developments.

Chapter 3 starts with the formal introduction of the principle of Pareto-optimality, then passes to the formulation of concepts of the bicriteria supervised learning (MOBJ), which play a central role in the current research. The complexity measure and method of convex decomposition are introduced as the key elements of the proposed generalized learning framework, resulting in the scheme of the MOBJ algorithm. The chapter ends with the conclusion of necessity in a special complexity measure, providing two alternative approaches for its derivation. The following two chapters represent two corresponding studies, addressing the derivation of complexity measures with different approaches.

Chapter 4 follows the smoothness-based approach to the complexity measure, dealing with the idea of its expression via the properties of functions in Sobolev spaces. In application to the hypothesis space of RBF networks with Gaussian basis functions, the proposed complexity measure induces the MOBJ-RBF algorithm, which is capable of finding of efficient solutions to the supervised learning problem determining the weights, widths, centers and quantities of the basis functions in a deterministic and computationally-efficient manner. A special attention is paid to adaptation of the information criteria for model selection, that provide a high generalization performance to the MOBJ algorithm at almost no computational costs. The capabilities of the proposed MOBJ algorithm are demonstrated in a series of benchmark tests.

Chapter 5 aims to extend the concept of margin maximization to multi-objective learning in the context of multiple kernels by means of the derivation of corresponding complexity measure. The chapter starts with the detail analysis of the argument, earlier formulated in Chapter 3, that a traditional definition of the margin via the norm in a feature space is not a valid complexity measure in the multi-kernel context, from the SRM point of view. Consequently, new techniques of normalization and equalization of feature spaces are proposed, demonstrating the necessity in further extension of the concept of geometrical margin. The corresponding extension is proposed using the stability interpretation of the margin maximization. Its formalization leads to the development of new complexity measure, based on the stability properties of separation hyperplanes. The theoretical results are examined in the extensive experiment, whose statistical analysis confirms the reliability of the proposed MOBJ approach.

Chapter 6 concludes the thesis, summarizing results and outlining their possible ways of further development.

Chapter 2

Theoretical background

Starting from the basic concepts of machine learning, this chapter provides a brief introduction to the fundamental learning principles and several common learning techniques, serving as a point of reference for further developments. In detail, the subjects of the chapter are covered in [Vapnik, 1995; Haykin, 1999; Scholkopf and Smola, 2001].

2.1 Introducton

The classical machine learning discipline distinguishes several scenarios of the learning process: supervised, unsupervised, semi-supervised, reinforcement, and transductive. In spite of principal differences, they all can be viewed in a general actor-environment scheme, where the learner (actor) interacts with a learning object (environment) by means of observations. Commonly, when observations are represented by certain mathematical objects, the learner's goal is to find a *hypothesis* function, which provides a certain response to input observations. In the above scheme, the scenario of interaction and the sought hypothesis depends on the kind of learning task. Tasks, such as pattern recognition, function approximation, prediction, control, and filtering are usually associated with the scenario of supervised leaning, lying in the scope of current work.

In supervised learning, the learner's aim is an induction from observations. Namely, given a finite *training set* of the observed inputs and corresponding target outputs

of the learning object, the learner provides a hypothesis function, whose response to an unseen input observation (not from the training set) predicts a response of the learning object. In other words, the learner performs generalization of the training data with the hypothesis.

Most supervised learning tasks are reduced to classification (pattern recognition) or regression. Moreover, the former can be seen as a specific case of the latter. In the case of classification, the input observation consists of characteristics of the object (*features*), whose class is denoted by the *label* and contained in the corresponding target output. Hence, a good hypothesis is supposed to classify previously unseen objects by providing the correct label of their class. In the case of regression, the target response is usually a real-valued scalar or vector and, thus, the hypothesis is supposed to reproduce (approximate) the unknown function, from which the training set was sampled.

The learner is usually represented by a certain algorithm, referred to as *learning machine*. Learning machines can be distinguished by the arsenal of available hypothesis functions they implement and the way these functions are implemented. The former determines a *hypothesis space* of a learning machine, whereas the latter splits learning algorithms into several classes.

Lazy algorithms perform generalization of the training set at the moment of evaluation of a hypothesis function and, consequently, require significant computational resources each time the response is produced. The opposite, *eager* algorithms, generalize the data at the training phase. Also, the algorithms can be split into instance- and model-based. Instance-based algorithms rely on comparison of unseen observations with the training set and, thus, are also lazy. Such algorithms often require memorization of the training set, e.g., the well-known k-nearest neighbor ($k-$NN) [Fix and Hodges, 1951] algorithm. Model-based algorithms are usually eager and generalize the data by means of a model. The common examples of model-based algorithms are decision trees [Quinlan, 1986], adaptive liner elements (Adalines) [Widrow and Stearns, 1985], multilayer neural networks [Haykin, 1999], and kernel machines [Hofmann, Schöolkopf, and Smola, 2008].

In practice, the learning object is usually nondeterministic and, generally, nonstationary with unknown properties. The learning process in such conditions becomes

a difficult task and requires dealing with several kinds of uncertainty. Within the model-based approach, the uncertainty is usually divided into structural and parametrical components, giving rise to the problems of *model selection* and *parameter estimation*, respectively.

2.2 Elements of statistical learning theory

Statistical Learning Theory (SLT) [Vapnik, 1995, 1998], also known as VC theory due to Vladimir Vapnik and Alexey Chervonenkis, is an essential machine learning framework, which is based on statistical interpretation of a learning process. As its major contribution, SLT establishes fundamental definitions and principles of learning, addressing the problems of parameter estimation and model selection in the nondeterministic environment.

2.2.1 Problem setting

Let the hypothesis space Ω be a class of functions $f : \mathcal{X} \to \mathcal{Y}$, which map input observations from $\mathcal{X}$ into the target space $\mathcal{Y}$, and let the scalar loss function $l(x, y, f(x))$ stand for the measure of regression or classification errors. Then, the fundamental problem of supervised learning can be stated as the minimization of the expected error over Ω. Specifically, one seeks in Ω for the minimizer of the expected risk functional

$$R[f] := \int_{\mathcal{X}\times\mathcal{Y}} l(x, y, f(x))\partial P(x, y) = E[l(x, y, f(x))], \tag{2.1}$$

where the joint probability distribution $P(x, y)$ describing the learning object is unknown, but only the training set

$$Z_{\text{tr}}^N := \Big\{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1 \dots N\Big\}$$

of N samples i.i.d. with respect to $P(x, y)$ is given.

2.2.2 Regression as density estimation

Let the problem of regression given by the squared error loss function $l(x, y, f(x)) = (y - f(x))^2$. Then, the risk functional (2.1) can be rewritten as

$$R[f] = \int_{\mathcal{X}} \int_{\mathcal{Y}} (y - f(x))^2 p(x, y) \partial y \partial x,$$

where $p(x, y)$ is the joint probability density function. Introducing the conditional expectation $g(x) = E[y|x] = \int_{\mathcal{Y}} y p(y|x) \partial y$, one may write

$$\begin{aligned}
R[f] = & \int_{\mathcal{X}} \int_{\mathcal{Y}} \left(y - g(x) + g(x) - f(x)\right)^2 p(x, y) \partial y \partial x \\
= & \int_{\mathcal{X}} \int_{\mathcal{Y}} \left(y - g(x)\right)^2 p(x, y) \partial y \partial x + \int_{\mathcal{X}} \int_{\mathcal{Y}} \left(f(x) - g(x)\right)^2 p(x, y) \partial y \partial x \\
& -2 \int_{\mathcal{X}} \int_{\mathcal{Y}} \left(y - g(x)\right) \left(f(x) - g(x)\right) p(x, y) \partial y \partial x \\
= & E\left[\left(y - g(x)\right)^2\right] + E\left[\left(f(x) - g(x)\right)^2\right] \\
& -2 \int_{\mathcal{X}} \left(f(x) - g(x)\right) p(x) \int_{\mathcal{Y}} \left(y - g(x)\right) p(y|x) \partial y \partial x.
\end{aligned}$$

As seen, the term

$$\int_{\mathcal{Y}} (y - g(x)) p(y|x) \partial y = \int y p(y|x) \partial y - g(x) = 0$$

vanishes, hence the expected risk turns to be the sum of the two expectation terms

$$R[f] = E[(y - g(x))^2] + E[(f(x) - g(x))^2],$$

where the former depends only on the unknown p.d.f. and the latter depends on the hypothesis f. Finally, one can conclude that

$$f^{\circ}(x) = E[y|x]$$

is the unique minimizer of $R[f]$. Consequently, the problem of supervised learning in current regression setting is equivalent to the problem of estimation of the conditional expectation $E[y|x]$ (or conditional density $p(y|x)$) from the set of empirical observations. It is noteworthy that the global minimum of the expected risk

$R[f^\circ] = E[(y - g(x))^2]$ is a constant of a particular learning problem, commonly interpreted as the variance of sampling noise.

2.2.3 Empirical risk minimization

The uncertainty of the distribution $P(x, y)$ prevents explicit minimization of the expected risk $R[f]$. However, approximation of $P(x, y)$ with the empirical distribution, available from the training set Z_{tr}^N, allows one to minimize the empirical risk

$$R_{\text{emp}}[f] = \frac{1}{N} \sum_{i=1}^{N} l(x_i, y_i, f(x_i)), \tag{2.2}$$

instead.

Obviously, the empirical risk converges to the expected risk as the number of observations grows infinitely large, i.e.,

$$\lim_{N \to \infty} R_{\text{emp}}[f] = R[f]. \tag{2.3}$$

This basic assumption lies in the basis of the empirical risk minimization (ERM) principle. However, the fundamental study [Vapnik and Chervonenkis, 1989] claims that the fact of convergence (2.3) is not sufficient for a consistent learning with ERM, since the convergence of the empirical minimum $R_{\text{emp}}(f^*|N)$ to $R(f^\circ)$ is not uniform. As the result, the ERM is proved to be consistent *iff* the empirical risk converges uniformly to the expected risk in the worst-case scenario

$$\lim_{N \to \infty} \Pr\left\{ \sup_{f \in \Omega} |R[f] - R_{\text{emp}}[f]| > \varepsilon \right\} = 0, \text{ for all } \varepsilon > 0. \tag{2.4}$$

The condition (2.4) is also known as the nontrivial consistency principle, which is schematically demonstrated in Fig. 2.1.

The necessary and sufficient conditions of consistency are given in [Vapnik and Chervonenkis, 1968] by the concept of the growth functions. Namely, (2.4) holds *iff*

$$\lim_{N \to \infty} \frac{G(N)}{N} = 0, \tag{2.5}$$

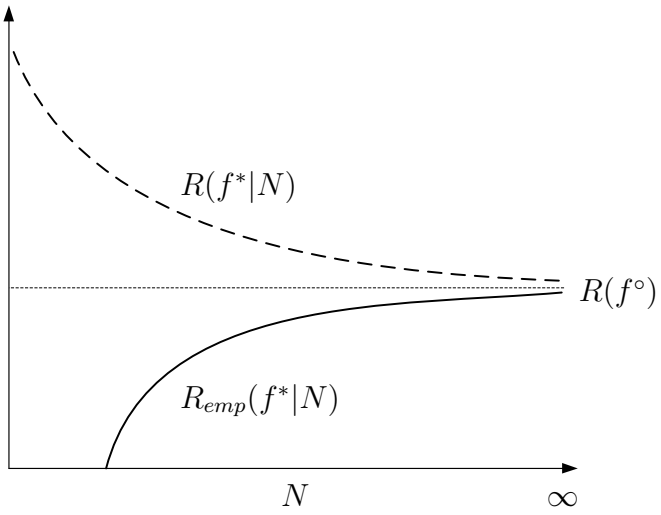

Figure 2.1: Demonstration of the non-trivial consistency principle of empirical risk minimization.

where $G(N)$ is the growth function. For instance, if $l(x, y, f(x))$ is the indicator[1] function, the growth function is then defined as

$$G(N) := \ln \max_{Z^N \in \mathcal{X}^N \times \mathcal{Y}^N} D(Z^N),$$

where $D(Z^N)$ is the number of all possible dichotomies (shatterings) of the set Z^N of N observations by the loss function $l(x, y, f(x))$ on Ω. The growth function $G(N)$ is the upper bound of the term $\ln E[D(Z^N_{\text{tr}})]$, called Vapnik's entropy, where the expectation is taken with respect to the unknown distribution $P(x, y)$. Vapnik's entropy represents the capacity of a learning machine with respect to $P(x, y)$, whereas $G(N)$ provides its distribution-independent bound.

It is easy to see that $D(Z^N_{\text{tr}})$ is bounded from above with 2^N, which is the maximum possible number of dichotomies of N samples. Consequently, for a particular learning machine, there exists such a positive constant h, that for all $N \leq h$ the identity $G(N) = N \ln 2$ holds. It means that the learning machine, after training on Z^N_{tr} of length $N \leq h$ with arbitrary labels, will classify all samples correctly, inducing false generalizations. This situation is called overfitting, which must be avoided by increasing N (a larger data-set) or by reducing h (smaller capacity of the hypothesis

[1]For example, in the problem of binary classification with labels $\mathcal{Y} := \{-1, 1\}$ the error loss function $l(x, y, f(x)) = \frac{1}{2}(1 - yf(x))$ is indicator.

space). As shown in [Vapnik and Chervonenkis, 1989], the growth of $G(N)$ after $N > h$ must slow down, which is sufficient for satisfying the condition (2.5). Therefore it can be seen that (2.5) is satisfied when there exists a finite h. The value of h is also known as VC-dimension and represents the distribution-independent measure of capacity of a learning machine.

Another interpretation of the consistency of ERM can be given within the context of Shannon's sampling problem [Shannon, 1949], where the problem of supervised learning can be viewed as the reconstruction of the unknown signal f, which is sampled N times with a certain frequency (see e.g., [Vapnik, 1995]). Then, the Nyquist-Shannon sampling theorem states that the reconstruction is possible when the sampling frequency (controlled by N) is sufficiently large for a given signal bandwidth (capacity of the a learning machine).

2.2.4 Bounds on uniform convergence

Several different learning machines may consistently implement ERM, providing different solutions to a particular learning problem. However, the value of the minimized empirical risk does not reflect quality of the achieved generalization, while the true value of the expected risk cannot be computed. Therefore, within the SLT framework, the quality of generalization is evaluated through the analysis of the uniform convergence results in $\Pr\{\sup_{f\in\Omega} |R[f] - R_{\text{emp}}[f]| > \varepsilon\}$ namely, as the form

$$R[f] \leq R_{\text{emp}}[f] + \Psi\left(R_{\text{emp}}[f], N, \Omega, \eta\right) \tag{2.6}$$

of the upper bound on the expected risk. Here η determines the confidence level, such that (2.6) holds with the probability of at least $1 - \eta$, and the term Ψ is the corresponding confidence interval. The bound (2.6) is also known as a generalization bound, since it tells us how well the expected risk is approximated.

Generalization bounds can be expressed in a variety of ways, depending on settings. For instance, if the classification problem is set for indicator loss functions, the VC-bound is given by the confidence interval

$$\Psi_{VC} = \frac{\varepsilon}{2}\left(1 + \sqrt{1 + \frac{4R_{\text{emp}}[f]}{\varepsilon}}\right),$$

where $\varepsilon = a_1 \left(h(\ln(a_2 \frac{N}{h}) + 1) - \ln(\frac{\eta}{4})\right)/N$. Here the coefficients $0 < a_1 \leq 4$ and $0 < a_2 \leq 2$ must be empirically chosen [Vapnik, 1995]. The VC-bounds demonstrate that a good generalization can be achieved when the ratio $\frac{N}{h}$ is large enough, so that ε is small and, as the consequence, a small empirical risk implies a small expected risk as well.

Generalization bounds can be also formulated without a certain dimension-like measure of capacity. For instance, an extension of Vapnik's results for arbitrary hypothesis space Ω and bounded loss function $l(x, y, f(x)) \leq \tau$ is also possible within the concept of metric entropy. For example, [Alon, Cesa-Bianchi, Ben-david, and Haussler, 1997] proposed the bound

$$\Pr\left\{\sup_{f\in\Omega} |R[f] - R_{\text{emp}}[f]| > 6\varepsilon\tau\right\} \leq 12NE\left[\mathcal{N}\left(\varepsilon, \Omega, \ell_\infty^{X^{2N}}\right)\right] \exp\left(-\varepsilon^2 N\right), \quad (2.7)$$

where $\mathcal{N}\left(\varepsilon, \Omega, \ell_\infty^{Z_{\text{tr}}^{2N}}\right)$ denotes the $\varepsilon-$covering number[2] of the hypothesis class Ω with ℓ^∞ metric on Ω w.r.t. double sample [3]. In (2.7), the expectation is taken with respect to the unknown distribution of the training sample. Hence, one usually constructs the distribution-independent form of (2.7) with

$$\mathcal{N}^m(\varepsilon, \Omega) = \sup_{x_1 \ldots x_m \in \mathcal{X}} \mathcal{N}\left(\varepsilon, \Omega, \ell_\infty^{X^m}\right)$$

under the assumption

$$E\left[\mathcal{N}\left(\varepsilon, \Omega, \ell_\infty^{X^{2N}}\right)\right] \leq \mathcal{N}^{2N}(\varepsilon, \Omega), \quad (2.8)$$

giving rise to the practical bound

$$\Pr\left\{\sup_{f\in\Omega} |R_{\text{emp}}[f] - R[f]| > 6\varepsilon\tau\right\} \leq 12N\mathcal{N}^{2N}(\varepsilon, \Omega) \exp\left(-\varepsilon^2 N\right). \quad (2.9)$$

Then, the confidence interval in the form (2.6) is $\Psi(R_{\text{emp}}[f], N, \Omega, \eta) = 6\varepsilon\tau$, which

[2] Given the set Ω and the metric $d(\cdot, \cdot)$ on it, the $\varepsilon-$covering number $\mathcal{N}(\varepsilon, X, d)$ is the minimum number of ε-balls sufficient to cover entire Ω

[3] The ℓ^∞ matric on functions in Ω on double sample id $d(f, f') = \|f - f'\|_{\ell_\infty^{X^{2N}}} = \max_{i=1\ldots 2N} |f(x_i) - f'(x_i)|$

can be calculated by solving

$$\eta = 12N\mathcal{N}^{2N}(\varepsilon, \Omega)\exp\left(-\varepsilon^2 N\right)$$

with respect to ε, given the confidence probability level $1 - \eta$.

In order to take advantage of (2.9), the confidence interval $6\varepsilon\tau$ must be evidently lower than the trivial bound $|R_{\text{emp}}[f] - R[f]| \leq \tau$, hence one is interested in $\varepsilon \leq \frac{1}{6}$. Also, the bound (2.9) is meaningful only when $\eta \leq 1$, since η is a probability. Then, it is easy to see that in the optimistic case with the minimum possible covering number $\mathcal{N}^{2N}(\varepsilon, \Omega) = 1$ the bound (2.9) becomes useful, starting with the length $N > 294$ of the training set. Under realistic conditions, when the bound (2.8) is not tight and the desired confidence level η is small enough, the minimum required sample size may be sufficiently larger than that available. Similar situation may occur with the VC-bounds, where the precise estimation of the VC-dimension is difficult. This is a significant limitation of the application of the uniform convergence bounds approach in practice.

2.2.5 Structural risk minimization

In spite of the practical difficulties of application of bounds on uniform-convergence, the analysis of generalization bounds (2.6) led to a fundamental approach for controlling the generalization properties of a learning machine by minimizing both the empirical risk and confidence interval terms. This idea lies at the basis of the inductive learning principle, called the structural risk minimization (SRM), developed by Vapnik [Vapnik, 1982].

In its original setting, SRM relies on the construction of the structure of nested hypothesis subsets

$$\emptyset \subset \Omega_1 \subset \Omega_2 \subset \ldots \subset \Omega,$$

following the order of their learning capacity, i.e., the VC-dimensions h_i of the corresponding subclases Ω_i that must satisfy the order $0 \leq h_1 \leq h_2 \leq \ldots \leq h$. Then, the minimizers $f_1, f_2, \ldots, f$ of the empirical risk R_{emp} over the corresponding elements of the structure represent alternatives, from which one selects the final hypothesis in accordance with the lowest risk bound, ensuring minimum of the guaranteed risk.

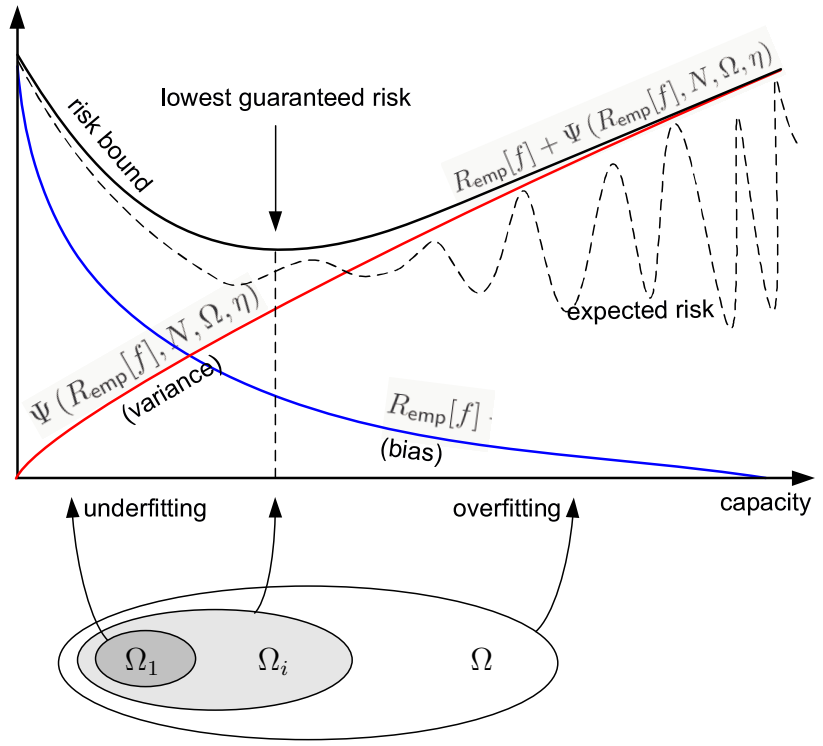

Figure 2.2: Illustration of the structural risk minimization principle

As seen from the general form of the bounds (2.6), the empirical risk decreases and the term $\Psi\left(R_{\text{emp}}[f], N, \Omega, \eta\right)$, playing the role of the confidence interval, increases simultaneously with the growth of capacity and *vice versa*. This paradigm depicts the main trade-off of learning, where minimization of the one of the learning objectives leads to increase of another. Similar phenomenon is also demonstrated with the known bias-variance dilemma [Geman, Bienenstock, and Doursat, 1992], where behavior of the empirical risk and capacity of the hypothesis class corresponds to a tradeoff between the bias and variance of the expected risk, the same as corresponding components in (2.6). Hereby, aiming to reach the best generalization to the given data-set, one should perform the search for a certain equilibrium between the empirical risk and capacity of the hypothesis class. This principle is usually put in practice with the application of model selection techniques, discussed below in 2.5.2.

2.3 Radial-basis function networks

The radial-basis function (RBF) networks are inspired by a kind of biological neurons, whose responses are concentrated in a narrow band of the input signal range. The earliest developments of the RBF architecture and its learning algorithm appear in [Moody and Darken, 1989]. The RBF networks are closely related to the non-parametric and kernel regression techniques [Nadaraya, 1964] and also proved to be universal approximators [Park and Sandberg, 1991], as an alternative to the popular multilayer perceptrons (MLP).

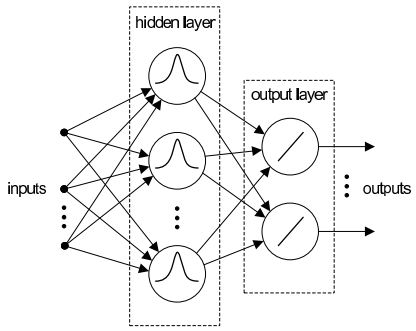

Figure 2.3: Architecture of the RBF network

2.3.1 Architecture

Analogous to MLP, the feed-forward architecture of an RBF network is composed of two layers: the hidden layer of non-linear units and linear output layer (Fig. 2.3). The n-input single-output[4] RBF network of m hidden units implements the function $f : \mathbb{R}^n \to \mathbb{R}$ with

$$f(x) = \sum_{i=1}^{m} \alpha_i k(\|x - c_i\|) + b, \tag{2.10}$$

where $k(\|x - c_i\|)$ is the i-th radial-basis function with the center parameter c_i (sometimes called prototype or centroid); α_i is the weight (or expansion coefficient) and b is the bias parameter.[5] Usually, basis functions are also parametrized. For instance,

[4]Without loss of generality, hereafter only single-output architectures are considered.

[5]For the sake of simplicity, the term $+b$ is omitted. As common, it is considered as the weight, corresponding to the unit basis function a constant output. However, when such representation is not convenient, parameter b is estimated separately from the weights (see e.g., 4.3.3 for details).

the most common are the Gaussian basis functions

$$k_\sigma(\|x - c_i\|) = \exp\left(-\frac{\|x - c_i\|^2}{2\sigma^2}\right), \tag{2.11}$$

with the width parameter σ.

As known, the RBF architecture is connected with Cover's theorem [Cover, 1965], which states that given any nonlinearly-separable training set of patterns in $\mathbb{R}^n$, there exists such a nonlinear map $\phi : \mathbb{R}^n \to \mathbb{R}^m$, $m \geq n$, that the images of training patterns under ϕ are linearly-separable. In other words, there is a hyperplane in $\mathbb{R}^m$ that separates N training samples in an arbitrary way. Within the context of RBF architecture, the hidden layer performs a nonlinear mapping from $\mathbb{R}^n$ into $\mathbb{R}^m$, whereas the vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)^T$ of expansion coefficients determines the separation hyperplane. Such connection with Cover's theorem immediately gives a clue for the choice $m \geq n$ of the number of basis functions. However, there is another perspective on Cover's theorem, further discussed in 2.4.1.

2.3.2 Connection with kernel regression

As shown earlier in 2.2.2, the problem of regression with the squared loss reduces to the problem of estimation of the conditional expectation

$$f(x) = E[y|x] = \int yp(y|x)\partial y,$$

which implies estimation of the unknown density $p(y|x)$. Applying the Bayes' rule, one can express the conditional density via the joint p.d.f. as $p(y|x) = \frac{p(x,y)}{p(x)}$, giving rise to the alternative form of the optimal regression function:

$$f(x) = p(x)^{-1} \int_{\mathcal{Y}} yp(x, y)\partial y. \tag{2.12}$$

Now, the unknown densities can be estimated by means of non-parametric methods, such as the Parzen-Rosenblatt [Parzen, 1962] estimator. The estimates of $p(x)$ and $p(x, y)$ are

$$\hat{p}(x) = \frac{1}{N\sigma_x} \sum_{i=1}^{N} k\left(\frac{x - x_i}{\sigma_x}\right)$$

and

$$\hat{p}(x,y) = \frac{1}{N\sigma_x\sigma_y}\sum_{i=1}^{N} k\left(\frac{x-x_i}{\sigma_x}\right) k\left(\frac{y-y_i}{\sigma_y}\right),$$

respectively, where σ_x, σ_y are the smoothing (width) constants and $k(\cdot)$, $\int k(x)\partial x = 1$ is the density kernel. Substitution of the unknown densities by their corresponding estimates in (2.12) leads to the estimated regression function

$$\hat{f}(x) = \hat{p}(x)^{-1}\int_{\mathcal{Y}} y\hat{p}(x,y)\partial y = \hat{p}(x)^{-1}\frac{1}{N\sigma_x\sigma_y}\sum_{i=1}^{N} k\left(\frac{x-x_i}{\sigma_x}\right)\int_{\mathcal{Y}} yk\left(\frac{y-y_i}{\sigma_y}\right)\partial y,$$

or

$$\hat{f}(x) = \frac{\sum_{i=1}^{N} y_i k\left(\frac{x-x_i}{\sigma_x}\right)}{\sum_{i=1}^{N} k\left(\frac{x-x_i}{\sigma_x}\right)}, \tag{2.13}$$

after simplification.[6]

After choosing the density kernel k to be translation- and rotation-invariant and introducing the RBF function $k_\sigma(\|x\|) = k(\frac{x}{\sigma})$, one recovers the RBF network

$$\hat{f}(x) = \sum_{i=1}^{N} \alpha_i k_\sigma(\|x - x_i\|)$$

from the estimate (2.13). Here, the centers of the basis functions correspond to the training patterns and the weights correspond to the relation

$$\alpha_i = y_i\left(\sum_{i=1}^{N} k_\sigma(\|x - x_i\|)\right)^{-1}.$$

Moreover, introducing the normalized basis functions

$$k'_\sigma(\|x\|) = k_\sigma(\|x\|)\left(\sum_{i=1}^{N} k_\sigma(\|x - x_i\|)\right)^{-1}$$

the regression function

$$\hat{f}(x) = \sum_{i=1}^{N} y_i k'_\sigma(\|x - x_i\|)$$

[6]The identity $\int_{\mathcal{Y}} yk\left(\frac{y-y_i}{\sigma_y}\right)\partial y = \sigma_y y_i$ is straightforward, since $\int k(x)\partial x = 1$.

receives the form of the Nadaraya-Watson [Nadaraya, 1964] estimator, which also corresponds to the normalized RBF network [Xu, Krzyzak, and Yuille, 1994].

2.3.3 Regularization networks

Another theoretical basis of RBF networks arises from the regularization framework, where the supervised learning is viewed as an ill-posed (underdetermined) curve-fitting problem in a high-dimensional space: the training set is sparse in $\mathbb{R}^n$ and therefore there are infinitely many interpolations possible. In this case, the method of regularization [Tikhonov, 1943] treats such a problem by a completion with a certain prior knowledge to "stabilize" the solutions, so that a unique optimal solution is guaranteed and the problem becomes well-posed.

In applications to neural networks, Tikhonov's regularization approach has been developed in [Poggio and Girosi, 1990] where the assumption of smoothness of the regression function f is used to stabilize solutions. In particular, the empirical risk functional (2.2) is extended by the regularized risk functional

$$R_{\text{reg}}[f] := \sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda\|Df\|^2. \tag{2.14}$$

Here the first term of the sum (2.14) stands for the approximation error and corresponds to $NR_{\text{emp}}[f]$ with the squared loss. The second term $\lambda\|Df\|^2$ in (2.14) is the regularization term, often called as stabilizer, where λ is the regularization parameter and D is a linear differential operator.

The results of [Poggio and Girosi, 1990] show that a unique minimizer of (2.14) is given by the expansion

$$f(x) = \sum_{i=1}^{N} \alpha_i G(x_i, x) \tag{2.15}$$

of Green's functions $G(x_i, x)$, associated with $\widetilde{D}D$ [7], with the expansion coefficients α_i, $i = 1, \ldots, N$, satisfying

$$\alpha_i = \frac{1}{\lambda}(y_i - f(x_i)). \tag{2.16}$$

[7] Functions satisfying $\widetilde{D}DG(x_i, \cdot) = \delta_{x_i}$, where $\delta_{x_i}(x) = \delta(x_i - x)$ is Dirac's delta function, centered at x_i, and $\widetilde{D}$ is the adjoint operator to D

Note, since $\widetilde{D}D$ is self-adjoint, the Green's functions are symmetric, i.e $G(x_i, x) = G(x, x_i)$.

Combination of (2.15) with (2.16) leads to the system of linear equations, whose solution with respect to the vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)^T$ of the expansion coefficients is

$$\alpha = (G + \lambda I)^{-1} Y, \tag{2.17}$$

where $G = G(x_i, x_j)$ is the $N \times N$ Green's matrix, $Y = (y_1, y_2, \ldots, y_N)^T$ is the target response vector and I is the identity matrix of corresponding size. One can show that given an arbitrary square matrix G there exist such positive λ, that $G + \lambda I$ is invertible, hence the solution is always possible at a certain regularization strength.

The effect of regularization depends on properties of the differential operator D, which determines Green's functions. In detail, the smoothing effect of regularization in frequency domain is demonstrated in [Girosi, Jones, and Poggio, 1995] via the Fourier analysis of D. Also, some properties of D directly determine the class of Green's functions. For instance, if D is translational- and rotational-invariant, the Green's functions $G(x_i, x) = G(\|x_i - x\|)$ are radial-basis. For example, one can show that the differential operator

$$(Df)(\cdot) = \sum_{|r|=n}^{\infty} \frac{\sigma^{2n}}{n! 2^n} \frac{\partial^n}{\partial^{r_1} x_1 \partial^{r_2} x_2 \cdots \partial^{r_n} x_n}$$

induces the Gaussian RBF (2.11) with the width σ. That, in turn, demonstrates that regularization with the Gaussian functions implies smoothness by penalizing all partial derivatives of f up to infinite order.

As seen from (2.15) and (2.16), the minimizer of the regularized risk (2.14) is expressed in terms of Green's functions and does not require evaluation of D, hence the solution to the regularization can be found for any symmetric $G(x_i, x)$, which admits the existence of the corresponding differential operator D.

Taking a closer look at the stabilizer term, it is straightforward to show that

$$\begin{aligned}\|Df\|^2 = \langle Df, Df\rangle &= \left\langle f, \widetilde{D}Df \right\rangle \\ &= \left\langle \sum_{i=1}^{N} \alpha_i G(x_i, \cdot), \sum_{j=1}^{N} \alpha_j \delta_{x_j} \right\rangle \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j \left\langle G(x_i, \cdot), \delta_{x_j} \right\rangle \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j G(x_i, x_j) = \alpha^T G \alpha,\end{aligned} \tag{2.18}$$

from where one concludes that with the growth of regularization strength λ, the expansion coefficients are getting more penalized and their magnitudes decrease. Hence, an infinite λ results in a flat regression function. On the other hand, as seen from (2.17), if G is invertible, the regression function f interpolates the training set, which corresponds to the known interpolation technique [Powell, 1985; Light, 1992]. It worth mentioning that according to Michelli's theorem [Michelli, 1986], the matrix G is invertible for certain classes of RBF functions, if the input vectors x_i (which are the centers of RBF functions) are distinct.

With the radial-basis Green's functions the expansion (2.15) corresponds to the architecture of the RBF network, while the regularization solution (2.17) determines its weights. The combination of the RBF architecture with such a learning procedure into a learning machine, is often called *regularization network* (RN). The key advantage of RN in contrast to other neural networks is the possibility of direct computation of the unique and optimal solution to the learning problem.

RNs are deeply related to the theory of reproducing kernels [Aronszajn, 1950], which serves as a bridge connecting RBF networks to the class of kernel machines (2.4) and also establishes their relation with the SRM [Girosi, Jones, and Poggio, 1995; Niyogi and Girosi, 1996]. These relations are discussed further in 2.5.1.

2.3.4 Generalized regularization networks

RNs were shown to be highly effective, but not always computationally affordable. This is due to the requirement of Green's matrix inversion, which computation becomes prohibitively expensive and numerically instable for large N. To overcome this

obstacle Poggio and Girosi [1990] proposed a method of the generalized regularization network (GRN) as an approximation of the RN (2.15), where the structure of an RBF network is allowed to be incomplete, i.e., the number of basis functions $m \leq N$.

Assuming that m basis functions are previously determined, the output response vector of the RBF network can be written in matrix form

$$\hat{Y} = H\alpha, \tag{2.19}$$

where $H = \{h_{ji}\}$ is the $N \times m$ design matrix with the elements $h_{ji} = k(\|x_j - c_i\|)$. Now, the regularized risk functional (2.10) can be rewritten with respect to the vector of expansion coefficients α, namely

$$\begin{aligned} R_{\text{reg}}(\alpha) =& \|Y - \hat{Y}\|^2 + \lambda \|Df\|^2 \\ =& \|Y - H\alpha\|^2 + \lambda \alpha^T G_0 \alpha \\ =& (Y - H\alpha)^T (Y - H\alpha) + \lambda \alpha^T G_0 \alpha \\ =& Y^T Y - 2Y^T H\alpha + \alpha^T (H^T H + \lambda G_0)\alpha, \end{aligned} \tag{2.20}$$

where G_0 is the $m \times m$ Green's matrix. The minimizer of (2.20) is then a solution of the system of linear equations

$$\frac{\partial R_{\text{reg}}(\alpha)}{\partial \alpha} = -2H^T Y + 2(H^T H + \lambda G_0)\alpha = 0,$$

leading to the optimal vector of expansion coefficients

$$\alpha = (H^T H + \lambda G_0)^{-1} H^T Y, \tag{2.21}$$

which coincides with the solution to a linear regularization problem [Tikhonov, 1963].

The choice of $\lambda = 0$ leads to the common pseudo-inversion form of the ordinary least squares (OLS) estimate. A closely related estimate called *ridge regression* can be recovered from (2.21) when $G_0 = I$ is assumed. In this case, the regularization problem corresponds to penalization of the squared Euclidean norm $\|\alpha\|^2$ of the weights. Such an approach corresponds to Tikhonov's regularization in the space of expansion coefficients, also known as coefficient-based regularization, where the stabilizer $\|\alpha\|^2$ is not connected with any Green's functions, but also imposes smoothness.

The choice of parameter λ, similarly to the case of RN, plays a key role in generalization properties of RBF networks and has its direct connection with the principle SRM, further discussed in 2.5.2.

2.3.5 Overview of learning strategies

The traditional concept of regularization addresses only estimation of the parameters of the linear output layer, assuming the selection of basis functions and their parameters to be done *a priori* by the choice of the corresponding differential operator. Then, given a fixed design matrix H whose columns are regressors, the techniques of model selection for linear regression are supposed to be employed for estimation of the regularization parameter λ (see 2.5.2 for further discussion). However, the content of H, which is the structure of RBF network, also determines its generalization properties and must be therefore considered during the model selection process.

Within the concept of linear regression, the structure of RBF network can be determined with the techniques of subset selection, which assumes a selection of certain regressors from H into the model instead of all. An example of such approach is the regularized forward selection (RFS) [Orr, 1993] and its computationally efficient analogue, the regularized orthogonal least squares (ROLS) [Chen, Chng, and Alkadhimi, 1996]. The latter is based on the technique of orthogonal least squares [Chen, Cowan, and Grant, 1991], which consists of the sequential orthogonalization of H, providing information of the next regressor that must be included (or excluded) at each step. Such an approach led to development of a variety of the forward selection and backward elimination algorithms for RBF networks.

As an alternative to the subset selection, other strategies exist, not directed at minimization of the training error by selection of centers. First of all, there are various heuristic rules, such as random selection of centers from the training set [Lowe, 1989]. Another universal heuristics are based on filling the hyperbox of the input space with a grid of centers, however this approach suffers from the known "curse of dimensionality" due to the exponential blowup of the covariance matrix $H^T H$. In a combination with the above procedures for the selection of centers, the widths of basis function can be tuned either by heuristical rules (e.g., as the distance between centers) or by gradient procedures, such as the conventional back-propagation algorithm

and its second-order extensions, e.g., the Levenberg-Marquardt algorithm [Shepherd, 1997], widely used for training of MLPs.

The known equivalence of RBF networks to certain classes of fuzzy-inference systems [Wang, 1992; Jang and Sun, 1993; Jang, Sun, and Mizutani, 1997] allows application of the so-called hybrid learning [Moody and Darken, 1989]. Within this concept, the hidden layer, corresponding to a system of fuzzy rules, is constructed in the unsupervised manner by means of clustering algorithms, such as the popular k-means [MacQueen, 1967] and its fuzzy c-means (FCM) [Bezdek, 1981] analogue. Moreover, among the the sophisticated clustering algorithms are the subtractive clustering [Chiu, 1994], mountain clustering [Yager and Filev, 1994], and robust clustering [Bodyanskiy, **Kokshenev**, Gorshkov, and Kolodyazhniy, 2006; Bodyanskiy, Gorshkov, **Kokshenev**, and Kolodyazhniy, 2010]. Existing recurrent modifications of the latter clustering algorithms, in combination with the recurrent least squares procedures, allow the application of RBF networks to the class of evolving systems, as mechanisms of nonlinear identification in the on-line mode, which can be used for the adaptive control, filtering, and prediction under non-stationary signal conditions.

2.4 Kernel machines

A deeper study of regularization learning within the functional-analytic framework makes a bridge to kernel methods. With the growing popularity of SV machines, the kernel methods received much attention in the machine learning community and brought the understanding of learning process into a new level.

2.4.1 Kernel trick

Consider the class of learning machines, whose hypothesis space consists of functions $f : \mathcal{H} \to \mathbb{R}$, such that

$$f(\widetilde{x}) = \langle \widetilde{x}, w \rangle \,, \tag{2.22}$$

where $\widetilde{x}$ is the vector of features in a dot product space $\mathcal{H}$ and $w \in \mathcal{H}$ is the vector of model parameters (weights). With a slight adaptation of (2.22), such as addition of the bias term $+b$ and application of the sign$(\cdot)$ function, one recovers the hypothesis of

a linear classifier, where the vector w determines a normal of the decision hyperplane $\langle \widetilde{x}, w \rangle + b = 0$ and $-b$ determines its shift from the origin.

Now, assume that there exists a nonlinear feature map $\Phi : \mathcal{X} \to \mathcal{H}$ such that the feature vector $\widetilde{x} = \Phi(x)$ is the image, corresponding to the input observation from $\mathcal{X}$. In the view of Cover's theorem [Cover, 1965], the observations from $\mathcal{X}$ might be linearly-separable in $\mathcal{H}$ if its dimensionality is sufficiently large. Accordingly, the hypotheses (2.22) extended on $\mathcal{X}$ by means of Φ in the form $f(x) = \langle \Phi(x), w \rangle$ may correspond to a nonlinear learning machine, constructed on the basis of a learning algorithm for the class of linear functions (2.22) on $\mathcal{H}$.

However, adopting such direct extension, the learning algorithm becomes computationally unaffordable as it suffers from the "curse of dimensionality", when the required dimensionality of $\mathcal{H}$ (number of features) is high. Instead of imposing limitations on dimensionality of $\mathcal{H}$ and reducing the power of a learning machine, let us allow $\mathcal{H}$ to be a general (possibly infinite-dimensional) Hilbert space, endowed with the norm $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$. Next, consider the form

$$f(x) = \left\langle \widetilde{x}, \widetilde{f} \right\rangle \tag{2.23}$$

of nonlinear hypothesis, where the vector of weights

$$\widetilde{f} = \sum_i \alpha_i \widetilde{x}_i \tag{2.24}$$

lies within the $\operatorname{span}\{\widetilde{x}_1, \widetilde{x}_2, \ldots\}$ of a countable set of feature vectors. Using the bilinear property of the dot product, the nonlinear function (2.23) can be rewritten in the form

$$\begin{aligned} f(x) &= \left\langle \widetilde{x}, \widetilde{f} \right\rangle = \left\langle \widetilde{x}, \sum_i \alpha_i \widetilde{x}_i \right\rangle \\ &= \sum_i \alpha_i \left\langle \widetilde{x}, \widetilde{x}_i \right\rangle \\ &= \sum_i \alpha_i k(x, x_i), \end{aligned} \tag{2.25}$$

where computation of the dot product is encapsulated within the kernel function

$$k(x, x') := \langle \widetilde{x}, \widetilde{x}' \rangle = \langle \Phi(x), \Phi(x') \rangle . \tag{2.26}$$

Such a form of hypothesis representation with a kernel, often called "kernel trick", induces a class of learning algorithms referred to as *kernel machines*.

The existence of the closed form of a kernel function permits one to construct computationally affordable learning algorithms with up to infinite-dimensional feature spaces. Hence, the feature space is usually determined *a priori* by specifying a certain kernel function. In this case, the feature map Φ does not need to be calculated explicitly and only the function k instead. However, a proper choice of k must ensure the existence of the underlying feature map Φ, satisfying the so-called positive definiteness condition [8]

$$\sum_i \sum_j \alpha_i \alpha_j k(x_i, x_j) \geq 0, \quad \text{for all } \alpha_i \in \mathbb{R} \text{ and } x_i \in \mathbb{X}. \tag{2.27}$$

Assuming N feature vectors in the expansion (2.24), the condition (2.27) can be also given in the matrix form

$$\alpha^T G_k \alpha \geq 0 \quad \text{for all } \alpha \in \mathbb{R}^N \text{ and } x_i \in \mathbb{X},$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)^T$ is vector of expansion parameters and

$$G_k := \Big\{ k(x_i, x_j) \Big\},$$

is the square $N \times N$ Gram matrix [9] associated with k, from where one can infer that k is positive definite *iff* the Gram matrix G_k is non-negative definite.

With the traditional meaning of the dot product, the value of the kernel $k(x, x')$ receives interpretation of the correlation-based similarity measure between the features, associated with the observations x and x'. Therefore, many linear algorithms can be extended to a wide class of nonlinear functions, i.e., "kernelized", by means of the kernel trick with G_k, playing the role of the correlation matrix. For example, in such manner, the method of principal component analysis (PCA) [Pearson, 1901] transforms into the kernel-PCA [Schölkopf, Smola, and Müller, 1998].

[8] The condition is given for real-valued kernels.

[9] The Gram (or Grammian) matrix is the matrix of all possible dot products within a set of vectors. In the current context, dot products are computed with the kernel function k, that is why the matrix G_k sometimes referred in literature to as kernel matrix.

The equivalence of the hypothesis (2.25) to the function implemented by the architecture of RBF network allows one to include RBF networks and some of their learning algorithms into the class of kernel machines, when radial-basis functions play the role of translation- and rotation-invariant positive definite kernel.

2.4.2 Feature maps

The map Φ determines a content of features available for a kernel machine and, thus, strongly influences its generalization properties. Feature maps are not unique. Indeed, there are infinitely many feature maps satisfying (2.26) for a given k and $\mathcal{X}$ [Minh, Niyogi, and Yao, 2006]. The theory of reproducing kernels [Aronszajn, 1950] and Mercer's theorem [Mercer, 1909] establish two fundamental views on feature maps and spaces of positive definite kernels.

For a given positive definite kernel k, there is a Hilbert space $\mathcal{H}_k \subset \mathbb{R}^{\mathcal{X}}$, whose dot product is characterized by the reproducing property

$$\langle k(x,\cdot), f\rangle_{\mathcal{H}_k} = f(x), \tag{2.28}$$

where $k(x,\cdot)$ and f are the vectors in $\mathcal{H}_k$. As proved in [Aronszajn, 1950], the space $\mathcal{H}_k$, called the reproducing kernel Hilbert space (RKHS), contains expansions (2.25) with the dot product

$$\begin{aligned}\langle f, f'\rangle_{\mathcal{H}_k} &:= \left\langle \sum_i \alpha_i k(x_i,\cdot), \sum_j \alpha'_j k(x'_j,\cdot) \right\rangle_{\mathcal{H}_k} \\ &= \sum_{i,j} \alpha_i \alpha'_j k(x_i, x'_j) = \alpha^T G_k \alpha',\end{aligned} \tag{2.29}$$

and is unique for a given k.

It follows directly from (2.28) that

$$\langle k(x,\cdot), k(x',\cdot)\rangle_{\mathcal{H}_k} = k(x,x'), \tag{2.30}$$

which immediately yields the feature map

$$\begin{aligned}\Phi &: \mathcal{X} \to \mathbb{R}^{\mathcal{X}} \\ \Phi(x)(\cdot) &:= k(x, \cdot)\end{aligned}$$

into the RKHS $\mathcal{H}_k$ and, also, the symmetry property $k(x, x') = k(x', x)$ of the kernel.

Another view via Mercer's theorem provides the representation of a positive definite kernel in a uniformly convergent series

$$k(x, x') = \sum_j \lambda_j \psi_j(x) \psi_j(x'), \tag{2.31}$$

where $(\lambda_j)_j$ and $(\psi_j)_j$ are sequences of the corresponding non-zero eigenvalues for the integral operator

$$(T_k f)(\cdot) := \int_{\mathcal{X}} k(\cdot, x) f(x) \partial x. \tag{2.32}$$

As known, the operator T_k is compact for the symmetric real positive definite $k \in L_\infty(\mathcal{X}^2)$ on the non-empty closed subset $\mathcal{X} \subset \mathbb{R}^n$. Thus, the eigenspectrum of T_k is discrete and there exists a countable set of real positive eigenvalues $(\lambda_j)_j \in \ell_1$, corresponding to the orthonormal basis $(\psi_j)_j$ of square-integrable functions (see e.g., [König, 1986] for proofs). Consequently, the series (2.31) can be written in the form of the dot product (2.26) in ℓ_2, with the corresponding feature map

$$\begin{aligned}&\Phi : \mathcal{X} \to \ell_2, \\ &\Phi(x) := \left(\sqrt{\lambda_j} \psi_j(x)\right)_j .\end{aligned} \tag{2.33}$$

Note, that $\tilde{f} = f$ within the context of the RKHS. Also, the identity $f(x) = \left\langle \tilde{x}, \tilde{f} \right\rangle$ is irrelevant for the specification of a particular feature map associated with k. Hence, due to an obvious isometry between different feature spaces, associated with the same kernel, it is convenient to omit their specification, by the introduction of the notation

$$\|f\|_k = \sqrt{\left\langle \widetilde{f}, \widetilde{f} \right\rangle} = \sqrt{\alpha^T G_k \alpha} \tag{2.34}$$

of the norm of f in $\mathcal{H}_k$ or any other feature space associated with k.

2.4.3 Regularization in RKHS

Recall the setting of the regularization problem (see 2.3.3) and the form $f(x) = \sum_{i=1}^{N} \alpha_i G(x_i, x)$ of its unique solution, where $G(x_i, x)$ is a Green's function, corresponding to the linear operator $\widetilde{D}D$ and satisfying the condition

$$(\widetilde{D}DG(x', \cdot))(x) = \delta_{x'}(x), \text{ for all } (x, x') \in \mathcal{X}. \tag{2.35}$$

Now, with the identity (2.35) it is straightforward to show that

$$G(x, x') = \langle G(x, \cdot), \delta'_x \rangle = \left\langle G(x, \cdot), \widetilde{D}DG(x', \cdot) \right\rangle = \langle DG(x, \cdot), DG(x', \cdot) \rangle$$

or $G(x, x') = \langle \Phi(x), \Phi(x') \rangle$ with $\Phi(x) = DG(x', \cdot)$. Therefore, the Green's function $G(x, x')$ is a positive definite kernel. Connection of Green's function of $\widetilde{D}D$ with the kernel k leads to a conclusion that RN, in fact, is a kernel machine, whose hypothesis space is the RKHS of k. Moreover, one can identify the regularization term $\|Df\|^2$ with the squared RKHS norm of f, namely,

$$\begin{aligned} \|Df\|^2 = \langle Df, Df \rangle = \left\langle f, \widetilde{D}Df \right\rangle &= \left\langle \sum_{i=1}^{N} \alpha_i k(x_i, \cdot), \sum_{j=1}^{N} \alpha_j \delta_{x_j} \right\rangle \\ &= \sum_{i=1} \sum_{j=1} \alpha_i \alpha_j k(x_i, x_j) = \|f\|_k^2, \end{aligned} \tag{2.36}$$

which can also be seen after identifying the Gram matrix G_k in (2.29) with Green's matrix in (2.18).

Hence, the problem of minimization of regularized risk functional (2.14) can be viewed as the form of Tikhonov's regularization [Tikhonov, 1943] in $\mathcal{H}_k$:

$$\min_{f \in \mathcal{H}_k} R_{\text{reg}}[f] = \sum_{i=1}^{N} (y_i - \langle \widetilde{x}_i, f \rangle)^2 + \lambda \|f\|_k^2, \tag{2.37}$$

giving another interpretation of the regularization term, namely, the "size" of the hypothesis f in a feature space.

Note that according to the unique solution of regularization problem (2.15), the solution to (2.37) for the given training set Z_{tr}^N is the expansion of the corresponding feature vectors $\widetilde{x}_i$, $i = 1, \ldots, N$ with at most N nonzero coefficients.

2.5 A big picture

2.5.1 Unified learning framework

The well-known representer theorem in its modern setting [Scholkopf, Herbrich, Smola, and Williamson, 2001] allows a generalization of regularization learning to the case of arbitrary convex loss function $l(y, x, f(x))$ and the penalty term $\eta(\|f\|_k)$, when stating that the minimizer of $R_{\text{reg}}[f] = R_{\text{emp}}[f] + \eta(\|f\|_k)$ is unique and belongs to $\mathcal{H}_k$, where $\eta(\cdot)$ is a strictly monotonically increasing function. Therefore, a general regularization-based kernel machine can be given with the algorithm, solving the minimization problem

$$\min_{f \in \mathcal{H}_k} R_{\text{reg}}[f] = R_{\text{emp}}[f] + \lambda Q[f], \tag{2.38}$$

where $Q[f] = \eta(\|f\|_k)$.

Formally, the problem (2.38) can be equivalently represented in the form of constrained minimization

$$\min_{f \in \mathcal{H}_k} Q[f] \quad \text{s.t.} \quad R_{\text{emp}}[f] \leq \xi, \tag{2.39}$$

where the parameter ξ has a one-to-one correspondence with λ. Such a representation of (2.38) induces the class of SV machines[10], where $Q[f] = \|f\|_k^2$ corresponds to the inverse of the geometrical margin of a decision hyperplane. In particular, a choice of the so-called hinge loss function $l(x, y, f(x)) = \max(0, 1 - yf(x))$ or the $\varepsilon-$insensitive loss function $l(x, y, f(x)) = |y - f(x)|_\varepsilon$, one implements the algorithms of soft-margin SV classification (C-SVC) or SV regression (ε-SVR), respectively, developed in [Cortes and Vapnik, 1995]. Thereby, the learning problem (2.38) with $Q[f] = \|f\|_k^2$ simultaneously represents both concepts of regularization and margin maximization.

Another equivalent constrained form of (2.38) and (2.39)

$$\min_{f \in \mathcal{H}_k} R_{\text{emp}}[f] \quad \text{s.t.} \quad Q[f] \leq \epsilon, \tag{2.40}$$

[10] In practice, the problem (2.39) is usually solved in its dual form, where the constraint ξ is transformed into a sum of slack variables, whose penalization strength is controlled by another regularization parameter C.

directly illustrates the principle of structural risk minimization (SRM) [Vapnik, 1995], where the empirical risk R_{emp} is minimized within the structure of nested subsets $\Omega_i := \left\{ f \in \mathcal{H}_k : Q[f] < \epsilon_i \right\}$ in the RKHS of k, induced by $Q[f]$. In particular case of $Q[f] = \|f\|_k^2$, the empirical risk is minimized within the structure of nested balls in the RKHS of k, whose radii naturally represent their capacity.

The equivalence between the principles of regularization, margin maximization and SRM is widely discussed in literature (see e.g., [Evgeniou, Pontil, and Poggio, 2000]). The generalization of the above concepts is also possible within the Bayesian framework, which leads to the probabilistic (Bayesian) interpretations of $R_{\text{emp}}[f]$ and $Q[f]$.

Consider the maximum *a posteriori* probability (MAP) estimate of the random variable f from its noised sample Z_{tr}^N. Introducing the posterior probability $\Pr\{ f \mid Z_{\text{tr}}^N \}$, one seeks for the estimate

$$\hat{f} = \arg\max_f \Pr\{ f \mid Z_{\text{tr}}^N \}. \tag{2.41}$$

Application of Bayes' theorem allows one to substitute the posterior probability term in (2.41) by the proportion

$$\Pr\{ f \mid Z_{\text{tr}}^N \} \propto \Pr\{ Z_{\text{tr}}^N | f \} \Pr\{f\}, \tag{2.42}$$

where $\Pr\{ Z_{\text{tr}}^N | f \}$ and $\Pr\{f\}$ are the likelihood and prior probabilities, respectively. Assuming that Z_{tr}^N is sampled from the underlying f with a normal zero mean and σ^2 variance noise, the likelihood term is

$$\Pr\{ Z_{\text{tr}}^N \mid f \} = \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - f(x_i))^2 \right) = \exp\left(-\frac{N}{2\sigma^2} R_{\text{emp}}[f] \right),$$

which is the well-known connection between the maximum likelihood estimation and ERM. Now, assuming also that a prior distribution is known and given in the form $\Pr\{f\} = \exp(-Q[f])$, the estimate (2.41) corresponds to the minimizer of

$$-\ln\left(\Pr\{ Z_{\text{tr}}^N | f \} \Pr\{f\} \right) = \frac{N}{2\sigma^2} R_{\text{emp}}[f] + Q[f],$$

where, after multiplication by $\frac{2\sigma^2}{N}$, one recovers the regularization functional (2.38) with $\lambda = \frac{2\sigma^2}{N}$ (see e.g., [Girosi, Jones, and Poggio, 1993]). Within this context, the penalty term $Q[f]$, commonly referred to as *prior*, supplies *a priori* information about the distribution of f, whereas the parameter λ reflects *a priori* knowledge of the noise variance.

2.5.2 Hyperparameters and model selection

The combination of the concepts discussed above represents a unified framework for construction and analysis of modern learning algorithms, such as the state-of-the-art SV machines. Within this framework, the uncertainty of the learning problem can be split into the structural and parametrical parts, associated with the choice of the prior Q and regularization parameter λ, respectively.

Although both of them determine the generalization performance of a learning machine and must be addressed with model selection, only selection of λ corresponds to the principle of SRM. Therefore, within the SRM context the empirical risk R_{emp} and prior Q functionals are assumed to be specified *a priori*, whereas the minimizer of (2.38) for the given training data-set Z_{tr}^N is uniquely determined by λ in the form

$$f_\lambda = \text{KM}(Z_{\text{tr}}^N, R_{\text{emp}}, \lambda Q[\cdot]), \tag{2.43}$$

where KM is the generic kernel algorithm, solving the learning problem (2.38) in the one of its equivalent forms. In this setting, instead of expansion coefficients, the hypothesis f_λ is defined in terms of a single *hyperparameter* λ.

For implementation of the SRM, the selection of λ must correspond to the best estimate of the expected risk. Such estimates, of course, are not limited to generalization bounds (see 2.2.4), but also include a variety of techniques for selection of regularization parameters as well as the techniques for linear regression, available from statistics. Among them are the generalized cross-validation (GCV) estimate [Golub, Heath, and Wahba, 1979], Akaike (AIC) [Akaike, 1974] and Bayesian (BIC) [Schwarz, 1978] information criteria, and other heuristic criteria, such as L-curve [Hansen and O'Leary, 1993].

In general, model selection can be formulated with the concept of minimization of a certain criterion ζ over some set of hypotheses. Using the hyperparameter notation (2.43), the estimation of λ can be therefore formulated with the procedure

$$\lambda = \arg\min_{\lambda} \zeta(f_\lambda). \tag{2.44}$$

Depending on setting, more hyperparameters may appear. For instance, the classical SV regression technique requires *a priori* determination of ε, the hyperparameter of the ε-insensitive loss function, giving rise to the hypothesis

$$f_{\lambda,\varepsilon} = \mathrm{KM}(Z_{\mathrm{tr}}^N, R_{\mathrm{emp}_\varepsilon}[\cdot], \lambda Q[\cdot])$$

of the ε-SVR machine having two hyperparameters.

In the above example, hyperparameters control properties of the solution within the same hypothesis space $\mathcal{H}_k$, while the kernel k, as well as its corresponding prior Q, remains fixed. For obvious reasons, however, the introduction of kernel parameters, such as the kernel width σ (or its equivalent[11] γ), is essential in practice. In this case, a kernel machine extends to the multi-kernel context and the problem of *kernel selection* arises as a part of the whole model selection process.

A direct extension of the hypothesis space of a learning machine to the multi-kernel context by parametrization of the prior cannot be considered within a single regularization scheme. This is due to the problem becoming generally non-convex: there is already a unique optimal solution, associated with each choice of the prior. Therefore, kernel parameters are usually treated as additional hyperparameters resulting in the general form

$$f_{\theta_R,\theta_Q} = \mathrm{KM}(Z_{\mathrm{tr}}^N, R_{\mathrm{emp}_{\theta_R}}[\cdot], Q_{\theta_Q}[\cdot]) \tag{2.45}$$

of the hypothesis, where θ_R and θ_Q are the vectors of hyperparameters associated with the loss function and the prior (including the regularization parameter), respectively.

Within the context of multiple hyperparameters, including the parameters of the prior, the problem of model selection cannot be addressed with criteria for linear

[11] In SV machine literature, the Gaussian RBF kernel is commonly defined as $k(x, x') = \exp(-\gamma\|x - x\|^2)$, where γ determines the bandwidth of kernel or scaling factor applied to the input space.

regression and requires the application of another risk estimates in the minimization scheme

$$(\theta_R, \theta_Q) = \arg \min_{(\theta_R,\theta_Q)\in\Theta} \zeta \left(f_{\theta_R,\theta_Q} \right), \tag{2.46}$$

were Θ is the combined space of hyperparameters θ_R and θ_Q associated with the loss function and the prior, respectively. Since Θ involves one or more hyperparameters different to λ, the scheme (2.46) does not implement the SRM on the whole hypothesis space, associated with Θ. Therefore, estimation of the hyperparameters, such as parameters of the prior Q, corresponds to another level of inference [Guyon, Saffari, Dror, and Cawley, 2010], for which the unbiased risk estimates are required.

2.5.3 Validation techniques

As known, the validation techniques can be used for the estimation of parameters of arbitrary learning algorithms. For example, one can consider the minimization of the validation error (MVE) criterion

$$\zeta^{\text{val}}_{Z_{\text{val}}}[f] = \frac{1}{N_v} \sum_{i=1}^{N_v} l(x_i, y_i, f(x_i)), \tag{2.47}$$

calculated for observations from the separate validation data-set $Z_{\text{val}} = \{(x_i, y_i)\}_{i=1}^{N_v}$. However, the MVE criterion becomes biased, suffering from a loss of representability when the validation set is not large enough or is not i.i.d. from the same distribution as the training set.

Another validation technique, not requiring an additional data-set, is the T-fold cross-validation (CV). Basically, it consists of splitting of the training set Z^N_{tr} into T non-overlapping subsets Z_l, $l = 1 \dots T$ of equal length and computing of the validation error on each subset, after training on the rest $T - 1$. Formally, the T-fold CV procedure for the selection of hyperparameters in Θ can be described in the extended form of (2.46), namely

$$(\theta_R, \theta_Q) = \arg \min_{(\theta_R,\theta_Q)\in\Theta} \sum_{l=1}^{T} \zeta^{\text{val}}_{Z_l} \left(\text{KM}(Z^N_{\text{tr}} \backslash Z_l, R_{\text{emp}}\theta_R[\cdot], Q_{\theta_Q}[\cdot]) \right). \tag{2.48}$$

If the training set contains N observations, the N-fold CV, also known as the leave-one-out error estimate, is the unbiased estimate of the expected risk on $N-1$ samples [Luntz and Brailovsky, 1969]. That is why CV is more reliable than a single validation pass (2.47). However as seen from (2.46), N-fold CV requires N runs of the learning algorithm per each combination of hyperparameters, ultimately leading to a significant computational load on large data-sets. Therefore, applications of the less accurate but faster 5-fold or 10-fold CV are most common in practice.

2.5.4 Overview of kernel selection techniques

The problem of kernel selection is commonly approached by selection of the corresponding hyperparameters in the T-fold CV scheme (2.48), where the space of hyperparameters Θ is usually approximated on multi-dimensional grid Θ_{grid}. Although such grid search procedures are the most used in practice, they become computationally unacceptable when the number of kernel parameters is more than several, due to the exponential growth of the number of grid elements and high computational requirements of the CV.

One of the alternatives to such exhaustive search by CV is the application of optimization techniques in combination with the computationally affordable approximates of the leave-one-out error. For instance, the proposed in [Chapelle and Vapnik, 1999; Chapelle, Vapnik, Bousquet, and Mukherjee, 2002; Keerthi, 2002] techniques of kernel selection for SV machines optimize the radius/margin bound [Vapnik and Chapelle, 2000] on leave-one-out error with gradient procedures. However, this approach may suffer from both the non-convexity of the leave-one-out error with respect to hyperparameters and biasedness of the radius/margin bound.

The novel techniques [Bach, Lanckriet, and Jordan, 2004; Micchelli and Pontil, 2005] and [Ong, Smola, and Williamson, 2005] brought the concept in kernel selection by the formulation of the problem in a convex form, referred to as multiple kernel learning. In particular, the kernel associated with the hypothesis is allowed to be an arbitrary convex combination of predetermined basis kernels. In terminology of [Ong, Smola, and Williamson, 2005] such a learning problem with multiple kernels is viewed as a regularization with the corresponding *hyperprior*, which solution depends on a single regularization parameter. The solution of such regularization problems are

associated with the quadratically-constrained quadratic programs (QCQP), whose efficient solutions are yet to be found.

2.6 Discussion and further motivation

The unified framework for kernel machines provides a methodology for the construction and analysis of efficient learning algorithms with various neural architectures, including the RBF and certain cases of MLP networks[12]. As discussed in section 2.5, when the learning problem is given by the empirical risk (loss function) and prior functionals, the method of regularization determines a unique and optimal solution in correspondence with the choice of the regularization parameter. Such solution simultaneously corresponds to the maxima of geometrical margin and *a posteriori* probability, and, in combination with the model selection procedures for determination of regularization parameter, implements the principle of SRM.

Nevertheless, a further extension of the hypothesis space by multiple kernels requires a selection of kernel hyperparameters within the model selection scheme (2.46), which does not implement the SRM for the whole problem, but only at the level of regularization learning with a fixed prior. On the other hand, the trade-off between empirical risk and learning capacity can be resolved with a search along a single dimension, according to the SRM. Therefore, minimization of the model selection criterion over the complete space of hyperparameters is redundant. Since even an unbiased estimate of the expected risk involves some sort of uncertainty, the redundance of the search adds extra degrees of freedom into the problem of hyperparameter estimation, thereby reducing the precision.

Multiple kernel learning as the regularization with a convex hyperprior demonstrates one of the possibilities SRM on a hypothesis space of multiple kernels, by the search along the dimension of a single hyperparameter. However, the SRM itself is not restricted to the class of convex problems. The above facts being combined together serve as a motivation for development of the general learning framework, which provides an SRM-consistent learning scheme for possibly non-convex problems of learning with multiple kernels.

[12]The MLP with a linear output layer corresponds to the kernel architecture with the sigmoidal kernel $k(x, x') = \tanh(\langle x, x' \rangle + b)$, which is positive definite for certain choices of b.

Chapter 3

Multi-objective learning

The present chapter plays a central role in this study connecting the apparatus of statistical and regularization learning with the methodology of multi-objective machine learning (MOML). The first part of the chapter briefly introduces basic concepts of multi-objective optimization, which are necessary for the development of a multi-objective framework for supervised learning. Second part, deals with the methodology and main components of a non-evolutionary multi-objective algorithm for supervised learning, aimed at the construction of the multi-kernel and SRM-consistent learning machines.

3.1 Introduction

In a conventional (single-objective) setting of the learning problem, the solution is usually assumed to be a single hypothesis as the result of optimization of a single cost function, whose hyperparameters determine the point of balance between the empirical risk and capacity of the hypothesis class *a priori*. For example, in regularization learning, the trade-off between the goodness of fit and smoothness of hypothesis function is resolved by the choice of the regularization hyperparameter. Since the choice of hyperparameters is uncertain at the level of error minimization, higher level strategies are employed (e.g., diverse model selection techniques). In contrast, the MOML addresses problems of trade-offs between two or more learning goals dealing with multiple objective functions explicitly [Jin, 2006; Jin and Sendhoff, 2008].

Within the multi-objective framework, the learning problem is formulated as a multi-criteria decision process, where a resolution of uncertainty is carried out through the analysis of all possible alternatives (outcomes) and its reduction to their efficient subset. Specifically, the solution of a general multi-criteria problem includes the following steps:

1. Evaluation of all efficient outcomes, representing the trade-off of the problem;
2. Decision towards a specific efficient outcome with respect to the information supplied by the *decisor*.

In the context of the regularization example, the results of the first step correspond to the minimizers of the regularized risk for all possible choices of the regularization parameter, whereas the second step corresponds to the application of a model selection criterion, playing the role of decisor. In this particular case, the results of the multi-criteria decision process are equivalent to those of the model selection procedure (2.44), discussed in the previous chapter.

In its general setting, determination of efficient outcomes in the environment of multiple goals leads to the multi-objective optimization problem, whose solution is based on the so-called Pareto-optimality principle originally formulated by Pareto [1896].

3.1.1 Principle of Pareto-optimality

Formally, the unconstrained r-objective optimization problem on the domain Ω is given by the vector-objective function $\phi : \Omega \to \mathbb{R}^r$. Without loss of generality, assume that all components of ϕ are aimed at minimization. Then, the problem of minimization of ϕ on Ω can be denoted as

$$\min_{x \in \Omega} \phi(x) = (\phi_1(x), \phi_2(x), \ldots, \phi_r(x)).$$

Introducing the relation $x \overset{\phi}{\preceq} x'$ of weak dominance on Ω, the solution to the above problem can be given by the nondominated set set

$$\mathcal{P}(\Omega, \phi) := \left\{ x \in \Omega \mid \forall x' \in \Omega : x \overset{\phi}{\preceq} x' \right\} \tag{3.1}$$

of elements in Ω. Here, $x \overset{\phi}{\preceq} x'$ holds *iff* vectors $\phi(x)$ and $\phi(x')$ follow the lexicographical order, i.e., $\phi(x) \preceq \phi(x')$, where $\preceq$ is the lexicographical order relation on $\mathbb{R}^r$. In the current minimization setting, the lexicographical relation $u \preceq v$, $u \in \mathbb{R}^r$, $v \in \mathbb{R}^r$, holds *iff* all components of the difference vector $v - u$ are nonnegative [1]. In other words, one should read $x \overset{\phi}{\preceq} x'$ as "x dominates or is equivalent to x' with respect to ϕ", implying that x is equivalent or preferable to x' from the point of view of ϕ. Given the weak dominance relation, it is also possible to define the strict form $x \overset{\phi}{\prec} x'$, which holds *iff* $x \overset{\phi}{\preceq} x'$ and $x' \overset{\phi}{\npreceq} x$ simultaneously.

The set (3.1) is often called the Pareto set having Pareto-optimal elements. Note that a Pareto set is nondominated, whereas the opposite may not hold, unless the nondominance is global on whole domain Ω. All Pareto-optimal elements can be considered equivalent with respect to ϕ, since any improvement by the one of the objectives requires a certain loss at the rest. In other words, Pareto-optimal elements are efficient outcomes representing the trade-off, which play a central role in solutions of the game and multicriteria problems [Karlin, 1959; Liu, Yang, and Whidborne, 2003; Jahn, 2004].

The space $\mathbb{R}^r$ is called objective space and the image of the Pareto set in it referred to as Pareto frontier (or front). The latter stands for the "spectrum" of all possible trade-off outcomes and can be denoted as

$$\rho\left(\Omega, \phi\right) := \Big\{ p \in \mathbb{R}^r \mid \exists x \in \mathcal{P}(\Omega, \phi) : \phi(x) = p \Big\},$$

or

$$\rho\left(\Omega, \phi\right) = \phi(\mathcal{P}(\Omega, \phi))$$

with a simpler notation. The relation between Pareto set and its frontier is schematically demonstrated in Fig. 3.1 for the bi-dimensional case of $r = 2$. The Pareto front is bounded with the axis-parallel hyperbox with vertices corresponding to r extrema points, including the so-called ideal point (also known as "utopia" point). The extrema points $y_i^\circ = \phi(x_i^\circ)$, $i = 1, \ldots, r$ correspond to the independent minimizers $x_i^\circ = \arg\min_{x \in \Omega} \phi_i(x)$ of each objective function, whereas the ideal point

[1] Note that under the lexicographical relation $u \in \mathbb{R}^r$ and $v \in \mathbb{R}^r$ is a totally ordered space, hence the relation $\preceq$ is transitive (if $u \preceq v'$ and $v' \preceq v$, then $u \preceq v$), antisymmetric (if $u \preceq v$ and $v \preceq u$ then $u = v$) and total ($u \preceq v$ either $v \preceq u$), and is also has a number of linear properties (if $u \preceq v$ then $u - v \preceq 0$, $-u \succeq -v$, etc.)

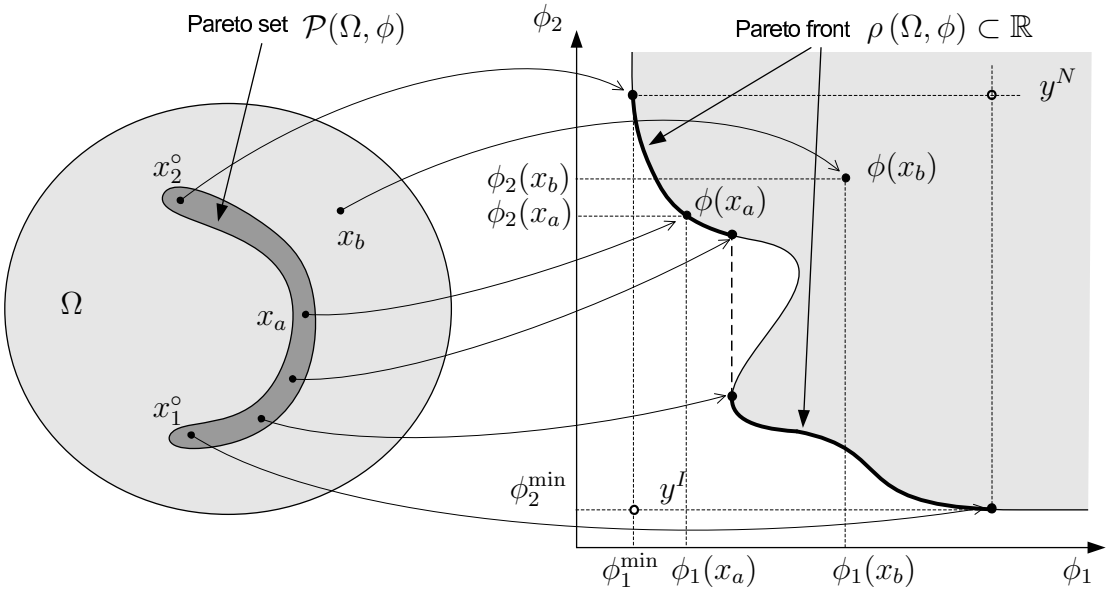

Figure 3.1: Schematic illustration of the Pareto-optimality principle and relations between the problem domain (left) and the objective space (right) on example of the following elements: x_a is Pareto-optimal; x_b is dominated; $x_1^\circ = \arg\min_{x\in\Omega} \phi_1(x)$ and $x_2^\circ = \arg\min_{x\in\Omega} \phi_2(x)$ are the extrema; y^I and y^N are the ideal and nadir points, respectively.

$y^I = (\phi_1(x_1^\circ), \phi_2(x_2^\circ), \ldots, \phi_r(x_r^\circ))$ represents the unreachable minimum, bounding the Pareto front from below. As an opposite to the ideal point y^I, the so-called nadir point y^N corresponds to the main-diagonal reflection of y^I that bounds the Pareto front from above. In case of $r = 2$, the nadir point is simply given by the vector $y^N = (\phi_1(x_2^\circ), \phi_2(x_1^\circ))$.

Denoting the image of Ω under ϕ as $\phi(\Omega)$, one can give an alternative definition of the Pareto front in terms of the lexicographical infinum of $\phi(\Omega)$. Depending on the topology of $\phi(\Omega)$, the Pareto front may be non-convex and generally noncompact.

The convexity of the Pareto front is a commonly desired property. According to the one of possible definitions, the Pareto front is said to be convex *iff* all its elements belong to the boundary of the convex hull of $\phi(\Omega)$, i.e., when the following holds:

$$\rho(\Omega, \phi) \subseteq \partial\text{hull}\{\phi(\Omega)\}. \tag{3.2}$$

It is straightforward to show that the Pareto front belongs to the boundary of $\phi(\Omega)$, i.e., $\rho(\Omega, \phi) \subseteq \partial\phi(\Omega)$. On the other hand, a convex image $\phi(\Omega)$ implies the identity

$\partial\text{hull}\{\phi(\Omega)\} = \partial\phi(\Omega)$. Therefore, the convexity of $\phi(\Omega)$ is the sufficient condition for obtaining of the convex Pareto front, which, in turn, is achieved by the convex ϕ on a convex domain Ω.

3.1.2 Basic scalarization techniques

Solution of the multi-objective problem with the *a posteriori* decisor requires finding all elements of $\mathcal{P}(\Omega, \phi)$. The direct application of the Pareto-optimality principle by means of pairwise comparison (or sorting) of the elements in Ω for obtaining the whole Pareto set is possible only when Ω is finite and computationally affordable when the number of elements is relatively small. When Ω is the open subset of a certain vector space (discrete or continuous), the multi-objective problem can be solved with the conventional (single-objective) apparatus of mathematical programming by means of the so-called scalarization techniques. One of the scalarization approaches is the minimization of a certain aggregate objective function. For instance, the popular weighted-sum method assumes the minimization of the convex combination (convolution) of the objectives ϕ in the form

$$x_w^\circ = \arg\min_{x\in\Omega} \langle w, \phi(x)\rangle \,, \tag{3.3}$$

providing the Pareto-optimal solution $x_w^\circ \in \mathcal{P}(\Omega, \phi)$, where $w \in \mathbb{R}^r$ is the vector of non-negative weights (usually normalized such that $\|w\|_1 = 1$). However, the known limitation of the weighted-sum method (see e.g, [Das and Dennis, 1997]) is that solutions corresponding to non-convex parts of the Pareto front are unreachable. Moreover, in order to guarantee the uniqueness of the solution of (3.3), the weighted-sum $w^T\phi(x)$ must be strictly convex on Ω.

There is a number of theorems relating the weighted-sum method with the convexity of $\phi(\Omega)$ and $\rho\,(\Omega, \phi)$ (see e.g., [Ehrgott, 2005], ch. 3). The known example is [Geoffrion, 1968], stating that the solution of (3.3) is unique and strictly Pareto-optimal for arbitrary non-negative weight vector $w \in \mathbb{R}^r$ *iff* $\phi(\Omega)$ is strictly convex. Therefore, one infers that application of the weighted-sum method is suitable only for convex multi-objective problems, i.e., when the convex combination $w^T\phi(x)$ and consequently, the Pareto front are strictly convex.

Another common scalarization method is the so-called ϵ-constraint [Haimes, Lasdon, and Wismer, 1971; Chankong and Haimes, 1983], which consists in transformation of a multi-objective problem into the form of the constrained minimization

$$\begin{aligned} x^{\circ}_{\epsilon} &= \arg\min_{x\in\Omega} \phi_s(x),\ s.t. \\ &\phi_i(x) \leq \epsilon_i, \\ &i = 1, \ldots, s-1, s+1, \ldots r \end{aligned} \tag{3.4}$$

of an arbitrary objective function from ϕ subject to $r-1$ constraints on the others, where $\epsilon = (\epsilon_1, \ldots, \epsilon_{s-1}, \epsilon_{s+1}, \ldots, \epsilon_r)$ stands for is the constraint vector. The solutions of (3.3) associated with admissible choices of ϵ are Pareto-optimal. In contrast with the weighted-sum method, the convexity of the problem is not a limitation of the ϵ-constraint method. However for a generally non-convex ϕ, the procedure (3.4) requires global optimization, which brings a multi-objective problem into the class of NP-complete problems.

3.1.3 Overview of approximate methods

In most practical cases, solutions to multi-objective problems cannot be drawn analytically but only numerically. However, the Pareto set may contain infinitely many elements and therefore a complete evaluation of its elements cannot be achieved. Hence, the multi-objective problem should be solved approximately with the representative finite subset of Pareto-optimal elements. The representativeness here means a uniform-like coverage of the whole Pareto set with a sufficient number of solutions. Formally, one can approximate $\mathcal{P}(\Omega, \phi)$ by its ε-cover found for a certain metric on Ω. However in practice, when ϕ is smooth, one usually finds an empirically large number of points, evenly-spaced on the Pareto front. Such an approximate Pareto-optimal set can be found directly by solving the weighted-sum (3.3) or ϵ-constraint (3.4) problems for the finite number of parameter choices on a certain multidimensional grid.

Another kind of approximation is needed when the optimization problem can not be solved exactly. These are mostly the cases when the problem is non-convex and its exact global solution is not computationally affordable due to NP-completeness. Therefore, the exploration of evolutionary optimization techniques, such as genetic algorithms (GA) and particle swarm optimization (PSO) became popular tools for

solving multi-objective problems [Forrest, 1993; Fonseca and Fleming, 1995; Hanne; Tan, Lee, and Khor, 2002]. In contrast to the finite-set approximation with exact Pareto-optimal solutions, the evolutionary approximations are nondeterministic[2] and result in populations of nondominated solutions close to the Pareto-optimum.

3.2 MOBJ: bicriteria supervised learning

Consolidation of the theoretical canvas of Chapter 2 with the multi-objective approach exhibits the supervised learning problem as the bicriteria decision process of minimization of the empirical risk and capacity of a hypothesis class. As mentioned in the above motivation in section 1.1, such a view on supervised learning for MLP networks, called MOBJ, has been demonstrated in [Teixeira, Braga, Takahashi, and Saldanha, 2000; Costa, Braga, Menezes, Teixeira, and Parma, 2003; Costa and Braga, 2006], by addressing the problem of multi-objective optimization with the techniques of nonlinear programming.

In fact, the concept of the MOBJ learning can be generally extended to a hypothesis space with arbitrary objective functions, defined on it. However from the SLT point of view, they must reflect the empirical risk and hypothesis capacity, respectively, in order to ensure the consistency of the MOBJ as the method of supervised learning. This claim defines the principal direction of current research.

3.2.1 Generalized learning concept

Coming back to the generalized form of regularization (2.38) discussed in 2.5, one can rewrite its equivalent MOBJ formulation as

$$\min_{f \in \mathcal{H}_k} \phi[f] = (R_{\text{emp}}[f], Q[f]). \tag{3.5}$$

It can be readily shown that the solution of (3.5) with the weighted-sum method (3.3) resembles the regularized risk functional (2.38) and exactly reproduces the set of its optimal solutions corresponding to all choices of the regularization parameter λ. It

[2]Commonly, evolutionary algorithms involve randomization to maintain the diversity of populations, which brings the nondeterministic nature to the solutions.

is not surprising that solution of (3.5) with the $\epsilon-$constraint (3.4) scheme leads to the forms of margin maximization (2.39) and SRM (2.40). Similarly, the usage of the model selection criterion in place of the decisor for choosing the final solution from $\mathcal{P}(\mathcal{H}_k, \phi)$ leads to the procedure (2.44) for estimation of the regularization parameter (see 2.5.2).

While the application of regularization learning in its weighed-sum form is limited only to strictly convex functionals R_{emp} and Q, the $\epsilon-$constraint solutions of the problem (3.5) follow the principles of margin maximization and SRM beyond the assumption of convexity. Therefore, one can consider the MOBJ as an extension of regularization on generally non-convex problems.

Following the motivation of section 2.6, a model selection search (2.46) within the space of multiple hyperparameters must be reduced to the search within the Pareto optimal set of the corresponding MOBJ problem, which follows the principle of SRM independently on the convexity of the problem and number of hyperparameters (the MOBJ problem remains bicriteria).

For the construction of such a MOBJ algorithm, one should first determine a corresponding decision space (domain) of the problem. Aiming for construction of a learning machine which is capable of automatic kernel selection (otherwise the problem is strictly convex and can be therefore directly solved with the regularization or its equivalents), let us introduce the family

$$K \subset \left\{ k \in \mathbb{R}^{\mathcal{X}^2} \right\}$$

of available kernels. Now, the hypothesis space of the multi-kernel machine can be given by the union

$$\mathcal{H}_K := \bigcup_{k \in K} \mathcal{H}_k \tag{3.6}$$

of the multiple RKHSs induced with the elements of K. Next, the model selection procedure (2.46) for kernel selection can be substituted by the MOBJ procedure

$$f_{\text{mobj}} = \arg \min_{f \in \mathcal{P}(\mathcal{H}_K, \phi)} \zeta[f], \tag{3.7}$$

where the model selection criterion ζ plays the role of the *a posteriori*[3] decisor and the components of the vector-functional ϕ stand for the empirical risk R_{emp} and a certain measure Q of the capacity of a hypothesis class. Within the context of MOBJ, the measure Q, usually called a *complexity measure*, induces the capacity order on $\mathcal{H}_K$ according to the SRM.

Note that in contrast to the conventional model selection procedure (2.46), the MOBJ procedure (3.7) minimizes the model selection criterion only over the Pareto-optimal subset of $\mathcal{H}_K$, which corresponds to reduction of the uncertainty.

3.2.2 Complexity measure and priors

In the MOBJ setting (3.5), corresponding to the learning machine with a fixed kernel, the prior functional Q stands for the complexity measure on $\mathcal{H}_k$. In this case, as outlined above in section 3.2.1, the equivalence between MOBJ, SRM, regularization, and margin maximization (with model selection) holds. In contrast, kernel selection (2.46) involves the parametrized prior $Q_{\theta_Q}[\cdot]$, whose hyperparameters θ_Q determine the kernel. However, the MOBJ procedure (3.7) must be endowed with a fixed complexity measure $Q : \mathcal{H}_K \to \mathbb{R}$ on the whole problem domain $\mathcal{H}_K$.

From the SRM point of view, the complexity measure Q must induce the nested subsets of $\mathcal{H}_K$, in the order of their learning capacity. Therefore, Q must be a kernel-invariant measure and, generally, cannot be a prior (such as the RKHS norm). This claim can be demonstrated by the following counterexample. Suppose that the family of kernels consists of two different kernels $K = \{k_1, k_2\}$ and assume the complexity measure to be the corresponding prior, i.e., the squared RKHS norm $Q[f] = \|f\|_k^2$. Hence, the complexities of hypotheses $f_1 \in \mathcal{H}_{k_1}$ and $f_2 \in \mathcal{H}_{k_2}$ are $\|f_1\|_{k_1}^2$ and $\|f_2\|_{k_2}^2$, respectively. At this point, one can already note that the norms $\|\cdot\|_{k_1}$ and $\|\cdot\|_{k_2}$ are metrics in different spaces and therefore the order relation $\|f_1\|_{k_1}^2 \leq \|f_2\|_{k_2}^2$ may be meaningless, unless a certain normalization is applied. Now, let us assume that kernels are linearly dependent such that $k_1(x, x') = c \cdot k_2(x, x')$, $c > 0$ and the hypotheses f_1

[3] The decision is taken after all Pareto-optimal elements are found and thus is posterior.

and f_2 are equivalent, i.e.,

$$\begin{aligned} f_1(x) &= \sum_i \alpha_i k_1(x, x') \\ &= \sum_i \alpha_i' k_2(x, x') = f_2(x), \quad \text{for all } x \in \mathcal{X}. \end{aligned}$$

This in turn yields $\alpha_i' = c \cdot \alpha_i$ and

$$\begin{aligned} \|f_2\|_{k_2}^2 &= \sum_i \sum_j \alpha_i' \alpha_j' k_2(x, x') \\ &= \sum_i \sum_j c^2 \alpha_i \alpha_j \frac{1}{c} k_1(x, x') \\ &= c \|f_1\|_{k_1}^2, \end{aligned} \tag{3.8}$$

depicting the contradiction:

$$f_1 = f_2, \quad Q[f_1] \neq Q[f_2]. \tag{3.9}$$

Consequently, the above expressions (3.8) and (3.9) demonstrate that there are infinitely many choices of c, such that $Q[f_1] \leq Q[f_2]$ either $Q[f_1] \geq Q[f_2]$ holds, whereas the hypotheses f_1 and f_2 are equivalent. Therefore, the squared RKHS norm or any other prior (which is a monotonic function of the RKHS norm by definition) cannot be a complexity measure on $\mathcal{H}_K$, if K contains arbitrary kernels. This conclusion is further extended in 5.1.1 to the case of families of linearly-independent kernels.

Accordingly, one should derive the complexity measure on the basis of a certain capacity concept, instead of using multiple priors.

3.2.3 Method of convex decomposition

The result of the MOBJ procedure (3.7) critically depends on the solution of the underlaying multi-objective problem, which, assuming that the functionals $\phi[f] = (R_{\text{emp}}[f], Q[f])$ are properly chosen, consists of finding the elements of $\mathcal{P}(\mathcal{H}_K, \phi)$. However, the application of nonlinear programming techniques for finding $\mathcal{P}(\mathcal{H}_K, \phi)$ is hampered by the fact that the decision space $\mathcal{H}_K$ is not a linear vector space.

Moreover, one can immediately conclude that $\mathcal{H}_K$ is non-convex since the convex combination

$$\alpha f_1 + (1-\alpha) f_2$$

of two elements $f_1 \in \mathcal{H}_{k_1}$ and $f_2 \in \mathcal{H}_{k_2}$ is not contained in $\mathcal{H}_K$, whereas $k_1 \neq k_2$, $k_1 \in K$, and $k_2 \in K$. Consequently, even when R_{emp} and Q are convex, the problem remains to be generally non-convex.

In order to avoid a global programming, it is proposed to decompose the problem into multiple convex subproblems in accordance with the following lemma:

Lemma 3.2.1 *(Decomposition of nondominated sets): Let the decision space be the union of l subsets $\Omega = \bigcup_{i=1}^{l} \Omega_i$, then the identity*

$$\mathcal{P}(\Omega, \phi) = \mathcal{P}\left(\bigcup_{i=1}^{l} \mathcal{P}(\Omega_i, \phi)\right)$$

holds.

Proof Define $P_i := \mathcal{P}(\Omega_i, \phi)$ for short and choose an arbitrary $x \in \mathcal{P}(\Omega, \phi)$. Since $x \in \Omega$ and $\Omega = \bigcup_{i=1}^{l} \Omega_i$, there exists at least one subset Ω_j that contains x, i.e., $x \in \Omega_j$. Also, since x is globally nondominated on Ω and $\Omega_j \subseteq \Omega$, x is nondominated on Ω_j as well. Thus $x \in P_j$ and, consequently, $x \in \bigcup_{i=1}^{l} P_i$. Hence, $\mathcal{P}(\Omega, \phi) \subseteq \bigcup_{i=1}^{l} P_i$.

Now, consider $x' \in \mathcal{P}\left(\bigcup_{i=1}^{l} P_i, \phi\right)$. Since x' is nondominated on $\bigcup_{i=1}^{l} P_i$, same holds for any subset of $\bigcup_{i=1}^{l} P_i$, including $\mathcal{P}(\Omega, \phi)$. Thus, the nondominated set $\mathcal{P}\left(\bigcup_{i=1}^{l} P_i, \phi\right)$ must contain $\mathcal{P}(\Omega, \phi)$. Since $\mathcal{P}(\Omega, \phi)$ in nondominated on Ω, same holds for its nondominated superset.■

The application of the above lemma to the union (3.6) allows one to decompose the Pareto set with the relation

$$\mathcal{P}(\mathcal{H}_K, \phi) = \mathcal{P}\left(\bigcup_{k \in K} \mathcal{P}(\mathcal{H}_k, \phi), \phi\right) \tag{3.10}$$

into a number of nondominated subsets $\mathcal{P}(\mathcal{H}_k, \phi)$, $k \in K$. Therefore, a generally non-convex problem of finding $\mathcal{P}(\mathcal{H}_K, \phi)$ can be split into the subproblems of finding

$\mathcal{P}(\mathcal{H}_k, \phi)$. The latter, in fact, are defined on convex $\mathcal{H}_k$ domains, and can be solved by means of the weighted-sum method, if R_{emp} and Q are strictly convex on $\mathcal{H}_k$, for all $k \in K$.

In such a way, assuming a finite number of kernels in K, one can directly find the Pareto set $\mathcal{P}(\mathcal{H}_K, \phi)$ from (3.10) with the elements of $\mathcal{P}(\mathcal{H}_k, \phi)$, $k \in K$ found exactly by means of minimization of the corresponding regularized risk functionals (2.38). Therefore, a combination of (3.10) with the general MOBJ procedure (3.7) allows one to construct an efficient and deterministic MOBJ algorithm that takes advantage of convex optimization.

3.3 Summary

Supervised learning and its fundamental concepts such as regularization, margin maximization, and the SRM are generalized into a common MOBJ scenario of the decision process within the environment of two conflicting goals. While the regularization method is restricted to the class of convex learning problems, the method of MOBJ can be efficiently applied to the non-convex cases as well, while preserving the implementation of the SRM.

For instance, endowed with a certain complexity measure, the MOBJ procedure implements the SRM within the multi-kernel context, in contrast to the conventional kernel selection approach. In this case, the non-convex bicriteria optimization problem can be efficiently solved with the proposed method of decomposition, taking advantage of convex programming for finding of exact solutions, in contrast to non-deterministic approximation via evolutionary programming. Such the multi-kernel MOBJ algorithm can be constructed from the following components:

1. Hypothesis space $\mathcal{H}_K$ induced with the family of available kernels K;
2. Empirical risk functional R_{emp} associated with a convex error loss function;
3. Complexity measure Q on $\mathcal{H}_K$ (convex on $\mathcal{H}_k$, for all $k \in K$);
4. Procedure for finding the elements of nondominated set $\mathcal{P}(\mathcal{H}_k, \phi)$;
5. Model selection criterion (decisor) ζ on $\mathcal{H}_K$.

In view of the discussion in section 3.2.2, all elements from the above list can be borrowed from existing single-objective kernel algorithms, except the complexity measure Q, which cannot be a prior associated with a variable kernel but also must induce the correct order on $\mathcal{H}_K$. Consequently, the complexity measure is a cornerstone of the MOBJ algorithm whereas there is no off-the-shelf receipt for its derivation.

Within the context of learning with a single kernel, both the smoothness and margin capacity concepts are equivalent since associated with the same complexity measure, playing the role of the prior determining the kernel. The extension to the context of multiple available kernels requires a definition of the complexity measure of a higher level. In this case, the capacity concepts may remain connected (i.e., hypotheses with a large margin represent smooth functions), but no longer equivalent. Therefore, the sought complexity measure can be derived on the basis of a single concept of capacity, represented with the smoothness or margin.

Chapter 4

Multi-objective algorithm for RBF networks

The smoothness of functions can be expressed explicitly by means of a certain measure of curvature. Following this general formulation of smoothness, the capacity of arbitrary hypotheses classes can be measured explicitly in the manner, invariant to their parametrization and structure. The development of such idea for the hypothesis space of RBF networks is presented in current chapter. As a result, the smoothness-based complexity measure and the corresponding MOBJ algorithm are proposed. The algorithm is capable of finding efficient solutions to the supervised learning problem by determining the weights, widths, centers and quantities of basis functions in a deterministic and computationally-efficient manner. [1]

4.1 Smoothness-based complexity measure

There is no unique definition of the smoothness of a function. In fact, smoothness can be viewed in many perspectives. Most of them are associated with the oscillatory nature or with the curvature of a function. The former perspective commonly leads to the analysis in the frequency domain, whereas the latter can be addressed with the notion of weak-differentiability and Sobolev spaces [Adams and Fournier, 2003].

[1] This chapter contains an extended version of results, published in [**Kokshenev** and Braga, 2007, 2008a,b, 2010]

4.1.1 Sobolev spaces and smoothness

The Sobolev space $\mathbb{W}_{q,p}(\Omega)$ on the open domain $\Omega \subset \mathbb{R}^n$ is the subspace of functions in $L_p(\Omega)$, whose all weak partial derivatives up to order q are also in $L_p(\Omega)$. Let us introduce the generalized partial differential operator

$$D^s := \frac{\partial^{|s|}}{\partial x_1^{s_1} \partial x_2^{s_2} \cdots \partial x_n^{s_n}}$$

on $\mathbb{R}^{\mathbb{R}^n}$, where $s \in \mathbb{Z}^n$ is the nonnegative multi-index, i.e., $s = (s_1, s_2, \ldots, s_n)$, $|s| = s_1 + s_2 + \ldots + s_n$. Then, the Sobolev space can be denoted as

$$\mathbb{W}_{q,p}(\Omega) := \left\{ f \in \mathbb{R}^{\Omega} \middle| \sum_{|s| \le q} \|D^s f\|_p < \infty \right\}$$

with the norm

$$\|f\|_{q,p} = \sum_{|s| \le q} \|D^s f\|_p, \tag{4.1}$$

or its equivalent[2]

$$\|f\|'_{q,p} = \|f\|_p + \sum_{|s|=q} \|D^s f\|_p. \tag{4.2}$$

One can show that $\mathbb{W}_{q,p}(\Omega)$ is complete under the norms (4.1)-(4.2) for all $1 \le q < \infty$ and thus is a Banach space. Moreover, in the case of $p = 2$, the space $\mathbb{W}_{q,2}$ becomes a Hilbert space of smooth square-integrable functions with the dot product

$$\langle f, g \rangle_{q,2} := \sum_{|s| \le q} \langle D^s f, D^s g \rangle_{L_2}.$$

In fact, as shown in [Girosi, Jones, and Poggio, 1993], the Sobolev space $\mathbb{W}_{2,2}$ is the RKHS associated with the Laplacian kernel, which also corresponds to the regularization term $Q[f] = \|f\|_{2,2}$.

However, aiming for the kernel-invariant complexity measure, the reproducing properties of the Sobolev spaces are intentionally omitted. Instead, assuming that the hypotheses' functions are smooth, i.e., belong to a certain $\mathbb{W}_{q,p}$, one can consider the complexity measure on the basis of the Sobolev norm $\|\cdot\|_{q,p}$.

[2]Since $D^0 f = f$, the condition $\sum_{|s|=q} \|D^s f\|_p < \infty$ is sufficient for all partial derivatives of lower order to be in L_p.

4.1.2 Bounds on smoothness

Assume that the hypothesis $f : \mathbb{R}^n \to \mathbb{R}$ is given with the expansion

$$f(x) = \sum_{i=1}^{m} \alpha_i k_\sigma(x, c_i)$$

of m translation invariant basis functions $k_\sigma(x, c_i) = k_\sigma(x - c_i) = \kappa(\frac{x-c_i}{\sigma})$, with center and width parameters c_i and σ, respectively. Here $\kappa : \mathbb{R}^n \to \mathbb{R}$ is the generating function which induces the family of available basis functions, usually positive definite kernels.

Assuming $f \in \mathbb{W}_{q,p}$ and using the Minkowski inequality [Minkowski, 1953] it is straightforward to show that the Sobolev norm of f, being a certain sum of L_p norms, is bounded from above with

$$\|f\|_{q,p} \leq \sum_{i=1}^{m} \|\alpha_i k_\sigma(\cdot, c_i)\|_{q,p},$$

and the bound is tight. On the other hand, one can show that

$$\|\alpha_i k_\sigma(\cdot, c_i)\|_{q,p} = |\alpha_i| \cdot \|k_\sigma(\cdot, c_i)\|_{q,p}$$

and the term $\|k_\sigma(\cdot, c_i)\|_{q,p}$ is invariant to c_i due to the translational invariance of k_σ. Hence, the term $\|k_\sigma(\cdot, c_i)\|_{q,p} = \|k_\sigma\|_{q,p}$ depends only on k_σ and can be taken out from the sum, giving rise to the bound

$$\|f\|_{q,p} \leq \|\alpha\|_1 \cdot \|k_\sigma\|_{q,p}. \tag{4.3}$$

In the above expression (4.3), the 1-norm $\|\alpha\|_1$ reflects the size of the expansion parameters (neural network's weights) and therefore is expected to be a part of the complexity measure. However, the term $\|k_\sigma\|_{q,p}$ includes at least one L_p norm term which, unfortunately, grows as $\mathcal{O}(\sigma^n)$ while the opposite trend of the complexity measure is expected: by increasing the width of radial-basis functions one is dealing with a smoother class of functions, thus its complexity must decrease. Therefore, the complexity measure should reflect a curvature of f in terms of $\|f\|_{q,p}$ while suppressing its growth with σ. Such an adaptation is possible using the relation of (4.3) to the

"size" of the basis function in place of the complexity measure. Namely, one can denote the general smoothness-based complexity measure as

$$Q[f] := \|\alpha\|_1 \frac{\|k_\sigma\|_{q,p}}{\|k_\sigma\|_p}. \tag{4.4}$$

The combination of (4.4) with the definition (4.2) of the equivalent Sobolev norm, results in the complexity measure

$$Q_{\text{rbf}}[f] := \|\alpha\|_1 \frac{\sum_{|s|=q} \|D^s k_\sigma\|_p}{\|k_\sigma\|_p}. \tag{4.5}$$

It is straightforward to show that

$$\|k_\sigma\|_p = \sigma^{\frac{n}{p}} \|\kappa\|_p,$$

where the term $\|\kappa\|_p$ does not depend on the hypothesis' parameters and, thus, can be treated as a constant for a given learning problem. Hence, the complexity measure (4.5) can be rewritten in the simplified form

$$Q_{\text{rbf}}[f] = \sigma^{-\frac{n}{p}} \|\alpha\|_1 \sum_{|s|=q} \|D^s k_\sigma\|_p. \tag{4.6}$$

Note, that the sum of remaining L_p terms in (4.6) is associated with the derivatives of order q, representing the curvature, whereas the multiplier $\sigma^{-\frac{n}{p}}$ provides normalization.

4.1.3 Second order curvature of Gaussian RBF

Let us consider the family of Gaussian functions choosing the generator function to be $\kappa(u) = \exp(-\frac{1}{2}\|u\|^2)$ and the differential order $q = 2$, corresponding to the common definition of curvature with second order derivatives. The second order partial derivatives of k_σ with respect to i-th and j-th coordinates are given by

$$\frac{\partial^2 k_\sigma(x)}{\partial x_i x_j} = \frac{1}{\sigma^2} k_\sigma(x) \cdot \begin{cases} \frac{x_i^2}{\sigma^2} - 1, & i = j, \\ \frac{x_i x_j}{\sigma^2} & \text{otherwise} \end{cases} = \frac{1}{\sigma^2} k_\sigma(x) \left(\frac{x_i x_j}{\sigma^2} - \delta_{ij} \right),$$

where δ_{ij} is the Kronecker delta. Hence, one can expand Sobolev's term in (4.6) as

$$\begin{aligned}\sum_{|s|=2} \|D^s k_\sigma\|_p =& \sum_{i=1}^{n}\sum_{j=1}^{n} \left\| \frac{\partial^2 k_\sigma(x)}{\partial x_i x_j} \right\|_p \\ =& \frac{1}{\sigma^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \left(\int_{\mathbb{R}^n} \left| k_\sigma(x) \left(\frac{x_i x_j}{\sigma^2} - \delta_{ij} \right) \right|^p \partial x \right)^{\frac{1}{p}},\end{aligned}$$

which yields

$$\sum_{|s|=2} \|D^s k_\sigma\|_p = \frac{1}{\sigma^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \left(\sigma^n \int_{\mathbb{R}^n} \kappa^p(u) \left(u_i u_j - \delta_{ij} \right)^p \partial u \right)^{\frac{1}{p}},$$

after switching the variable of integration to $u = \frac{x}{\sigma}$. Since in current setting κ is symmetric in all directions, the above expression reduces to

$$\sum_{|s|=2} \|D^s k_\sigma\|_p = \sigma^{\frac{n}{p}-2} C(n,p), \tag{4.7}$$

where

$$C(n,p) = (n^2 - n) \left(\int_{\mathbb{R}^n} \left(\kappa(u) u_i u_j \right)^p \partial u \right)^{\frac{1}{p}} + n \left(\int_{\mathbb{R}^n} \kappa(u) \left(u_i u_j - \delta_{ji} \right)^p \partial u \right)^{\frac{1}{p}}$$

is the multiplier depending only on κ, p, and n, which all are fixed in the context of a given learning problem. Hence, $C(n,p)$ is a constant and thus is irrelevant for the measure of complexity. It is noteworthy that the measure is irrelevant on p.

Finally, the combination of (4.7) with (4.6) (omitting $C(n,p)$) yields the simple expression of the complexity measure

$$Q_{\text{rbf}}[f] = \frac{\|\alpha\|_1}{\sigma^2} \tag{4.8}$$

for the class of hypothesis of RBF networks with the Gaussian basis functions. As expected, $Q_{\text{rbf}}[f]$ is a decreasing function of σ and increasing function the weights' magnitudes.

4.2 Pareto set of RBF networks

The derived complexity measure (4.8) for the class of RBF networks with Gaussian basis functions, in fact, is not a regularization stabilizer (prior) covered by the representer theorem (see. 2.5.1).

Hence, even though the empirical risk R_{emp} and complexity measure Q_{rbf} are strictly convex on the RKHS $\mathbb{H}_{k_\sigma}$ (associated with basis function of particular width σ), the minimizer of $R_{\text{emp}}[f] + \lambda Q_{\text{rbf}}[f]$ (which is, in turn, the element of $\mathcal{P}(\mathbb{H}_k, \phi)$) is the expansion

$$f(x) = \sum_{i=1}^{m} \alpha_i k_\sigma(x - c_i),$$

where the number of basis functions m and their centers c_i, $i = 1, \ldots, m$ do not necessary correspond to N training samples, same as in case of GRN (see 2.3.4). Therefore, one should start solving the multi-objective problem with a consideration of the general hypothesis space of RBF networks, containing all possible numbers of basis functions and locations of their centers.

4.2.1 Problem setting

Consider the general hypothesis class of RBF networks, implementing the functions

$$F := \Big\{ f : \mathbb{R}^n \to \mathbb{R} \mid f(x) = \sum_i \alpha_i k_\sigma(x - c_i) \Big\}, \tag{4.9}$$

where $m \in \mathbb{N}$, $c_i \in \mathbb{R}^n$, $\sigma \in \mathbb{R}^+$, and $\alpha_i \in \mathbb{R}$. It is assumed that the class F contains functions corresponding to all possible RBF networks, starting with the empty $m = 0$ and finishing up by infinitely large number of the Gaussian basis functions.

Given the mean squared error

$$R_{\text{emp}}[f] = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - f(x_i) \right)^2,$$

playing the role of the empirical risk functional on the training set Z_{tr}^N and the complexity measure (4.8), one aims to find elements of the Pareto set

$$\mathcal{P}(F, \phi), \quad \phi[f] = (R_{\text{emp}}[f], Q_{\text{rbf}}[f]) \tag{4.10}$$

of the corresponding MOBJ problem (see 3.2.1 for details).

It can be shown that F is the subset of the multi-kernel hypothesis space $\mathcal{H}_K$ induced with the family $K = \left\{k_\sigma | \sigma \in \mathbb{R}^+\right\}$ of Gaussian kernels. Then following the decomposition idea developed in 3.2.3, the MOBJ problem can be split into a number of subproblems on $\mathcal{H}_{k_\sigma}$ for various σ. In particular, it is possible to represent F with the union

$$F = \bigcup_{\sigma \in \mathbb{R}^+} \bigcup_{m \in \mathbb{N}} \bigcup_{C_m \in \mathbb{R}^{n \times m}} F_{\sigma_j, C_m}, \tag{4.11}$$

where

$$F_{\sigma_j, C_m} = \left\{ f \in F \mid \sigma = \sigma_j,\, (c_1, c_2, \ldots, c_m) = C_m \right\}$$

is the subset of all possible RBF networks with the given $n \times m$ centroid matrix C_m of m basis functions of width σ_j. Hence, all the RBF networks in F_{σ_j, C_m} correspond to the same design matrix and thus both R_{emp} and Q_{rbf} are strictly convex with respect to the vector of weight $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)^T$, while the rest of hypothesis parameters are fixed. Therefore, the application of the decomposition (3.10) to (4.11) allows one to reconstruct the Pareto set (4.10) from the elements of $\mathcal{P}\big(F_{\sigma_j, C_m}, \phi\big)$, whose corresponding weight vectors in $\mathbb{R}^m$ can be found by means of convex optimization.

4.2.2 Refinement of the hypothesis space

Even though the whole problem of finding (4.10) is decomposed into a number of smaller and convex subproblems on $\mathbb{R}^m$, the application of decomposition (3.10) requires carrying out the union over the infinite number of subsets, corresponding to a Cartesian of $\sigma \in \mathbb{R}^+$, $m \in \mathbb{N}$, and $C_m \in \mathbb{R}^{n \times m}$. While the approximation of $\sigma \in \mathbb{R}^+$ with a finite grid permits one to cover the range of widths in a representative way, the similar approach to approximation of $C_m \in \mathbb{R}^{n \times m}$ is not affordable due to a high dimensionality of its domain. On the other hand, it can be always assumed that input observations in the training set localized within the closed subset of $\mathbb{R}^n$. Hence, the choice of centers outside the closure of input observations is likely to lead to inefficient (dominated) hypotheses. Therefore, the domain of C_m can be significantly reduced to proper subsets associated with potentially nondominated hypotheses. This reduction is done in two steps: first, by fixing the number of basis functions m, and then fixing the centroid matrix.

Intuitively, it does make sense to limit $m \leq M$ by the number of distinct input patterns $M \neq N$. In fact, when all patterns are distinct, $M = N$ radial basis functions are sufficient for an arbitrary approximation of Z_{tr}^N, including its interpolation (see 2.3.3 for details). The next straightforward conclusion is that

$$\bigcup_{C_m \in \mathbb{R}^{n \times m}} F_{\sigma_j, C_m} \subseteq \bigcup_{C_M \in \mathbb{R}^{n \times M}} F_{\sigma_j, C_M}$$

holds when $m \leq M$ since $\mathbb{R}^{n \times m} \subseteq \mathbb{R}^{n \times M}$ allows to consider only the cases of exactly M basis functions, since all smaller networks are contained within F_{σ_j, C_m}. Therefore, the reduction of basis functions to the fixed number $m = M$ does not reduce the representability of the hypothesis space F.

Next, one can consider the approximation of the domain of C_M with the set of distinct input vectors from the training set. Let the centroid matrix

$$X_M = (x_1, x_2, \ldots, x_M)$$

consist of M distinct training input vectors and the centroid matrix

$$C_M = (c_1, c_2, \ldots, c_M)$$

contain at least one vector $c_i \notin X_M$, different from the training set. Assume that the hypothesis f, associated with C_M and a certain σ, is nondominated in F_{σ, C_M}, i.e.,

$$f \in \mathcal{P}(F_{\sigma, C_M}, \phi), \quad \phi[f] = (R_{\text{emp}}[f], Q_{\text{rbf}}[f]).$$

Formally, the choice of the centroid matrix X_M would be preferable to C_M, if another hypothesis $f' \in \mathcal{P}(F_{\sigma, X_M}, \phi)$ dominating f, i.e., $f' \overset{\phi}{\preceq} f$, exists.

Unfortunately, the existence of f' cannot be shown analytically for an arbitrary learning problem. Otherwise, this would be equivalent to the proof of the representer theorem for the case of coefficient-based regularization with Q_{rbf}. However, it can be demonstrated that the hypotheses in F_{σ, X_M} likely dominate those in F_{σ, X_C} for the particular case when all training patterns are distinct, i.e. $M = N$.

Let f' be equivalent to f with respect to the empirical risk, i.e., $R_{\text{emp}}[f'] = R_{\text{emp}}[f]$. Then, it is sufficient to consider the identity

$$H_X \alpha' = H_C \alpha, \tag{4.12}$$

where H_X and H_C are the design matrices associated with f' and f, and α' and α are the corresponding vectors of expansion coefficients, respectively. Since the basis functions are Gaussian, it is straightforward to show that $\text{tr}(H_X) = N$ since all diagonal elements of H_X are unit due to identity $k_\sigma(0) = 1$. Then, since there exists $c_i \notin X_M$, H_C has at least one diagonal element $k_\sigma(x_i - c_i) < 1$ and thus $\text{tr}(H_C) \leq N$. Therefore, the traces of design matrices satisfy

$$\text{tr}(H_C) \leq \text{tr}(H_X).$$

Since the magnitudes of α' and α grow inversely proportional to the eigenvalues of the corresponding design matrices, one infers that the situation when

$$\|\alpha'\|_1 < \|\alpha\|_1$$

may occur more likely than the opposite, implying that $f' \overset{\phi}{\preceq} f$. Similar conclusion can be extended to the case of $M < N$ with a consideration of singular values of the corresponding design matrices. However the above argument is sufficient to consider the choice of the centroid matrix

$$C_M = (x_1, x_2, \ldots, x_M) = X_M$$

to be an efficient heuristic strategy.

The above considerations combined together lead to the refinement of the hypothesis space F to the smaller and finite subset of hypotheses

$$\widetilde{F} = \bigcup_{\sigma \in S_\sigma} F_{\sigma, C_M},$$

where $S_\sigma = (\sigma_j)_j$ is the one-dimensional sequence (grid) of widths and C_M is the centroid matrix of M distinct training vectors from Z^N_{tr}. Finally, the application of

the decomposition (3.10) for $\widetilde{F}$ yields the finite-set approximation

$$\mathcal{P}\Big(\widetilde{F}, \phi\Big) = \mathcal{P}\Bigg(\bigcup_{\sigma \in S_\sigma} \mathcal{P}(F_{\sigma, C_M}, \phi), \phi\Bigg) \tag{4.13}$$

of $\mathcal{P}(F, \phi)$, where the length of S_σ (number of elements of the grid) and its elements control the approximation quality.

Note that at this point other heuristic strategies for selection of centers also can be used. For instance, the application of clustering procedures (see 2.3.5) may be useful in practice for further approximation of X_M having a smaller number of basis functions then M, especially on large data-sets. Therefore, it is hereafter assumed that the center vectors in C_M belong to a distinct subset input vectors from the training set.

4.3 Learning algorithm

After the objective functions are defined and the learning problem is decomposed into a finite number of convex bi-objective subproblems, one requires the computational procedure for finding the elements of $\mathcal{P}(F_{\sigma, C_M}, \phi)$ for the construction of the MOBJ algorithm for RBF networks.

4.3.1 Convex subproblem

The domain of the optimization subproblem $\mathcal{P}(F_{\sigma, C_M}, \phi)$ contains the RBF networks of M Gaussian basis functions with σ widths centered at distinct training input vectors (or their certain subset), corresponding to the common $N \times M$ design matrix $H = \{h_{ij}\}$ with the elements $h_{ij} = k_\sigma(x_i - x_j)$.[3] Therefore, the RBF networks can be represented only by their corresponding $M \times 1$ vectors of weights $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_M)^T$, leading to the representation of the bi-objective problem $\mathcal{P}(F_{\sigma, C_M}, \phi)$ on $\mathbb{R}^M$ with $\mathcal{P}\big(\mathbb{R}^M, \phi\big)$. Here, the complexity measure (4.8) is a function of α

$$Q_{\mathrm{rbf}}(\alpha) = \frac{\|\alpha\|_1}{\sigma^2}, \tag{4.14}$$

[3] Without loss of generality, it is assumed here that the training set is ordered, such that the first M input vectors are distinct.

and the mean squared error on $\mathbb{R}^M$ can be given in the matrix form

$$R_{\text{emp}}(\alpha) = \|Y - H\alpha\|^2, \tag{4.15}$$

where $Y = (y_1, y_2, \ldots, y_N)^T$ is the vector of target values from the training set Z_{tr}^N.

Due to the obviously strict convexity of the both objective functions, the minimizers of

$$R_{\text{reg}}(\alpha) = \|Y - H\alpha\|^2 + \lambda\|\alpha\|_1, \tag{4.16}$$

corresponding to all $\lambda \in \mathbb{R}^+$, are the sought Pareto-optimal elements of $\mathcal{P}\big(\mathbb{R}^M, \phi\big)$ and therefore is a corresponding solution of the convex subproblem associated with $\mathcal{P}(F_{\sigma,C_M}, \phi)$.

In fact, (4.16) is the well-known form of linear regularization with $1-$norm penalty term, also known as the least absolute shrinkage and selection operator (LASSO) regression [Tibshirani, 1996], which posses several useful properties.

4.3.2 Regularization path of the LASSO

There is a certain similarity between the LASSO and ridge regressions, as both are shrinking least squares estimators. As known, the Euclidean norm shrinkage operator used in ridge regression (see e.g, 2.3.4), also corresponds to Tikhonov's regularization. In the LASSO regression, the shrinkage operator is the $1-$norm operator, which not only penalizes the length of the coefficient vector but also implies its sparsity.

Geometrically, the sparsity is explained by the minimizers of (4.16) located at the edges of the 1-norm restriction polytope, whose coordinates are zeros for certain dimensions. As the restriction strength increases with λ thereby reducing the polytope size, more weights shrink to zero. Within the statistical framework, the LASSO regression is viewed as the subset selection process [Efron, Hastie, Johnstone, and Tibshirani, 2004], where the regressors (columns of the design matrix H) are sequentially selected in the order of their decreasing correlation with Y until reaching a certain stopping criterion. In this interpretation, the sparsity of the LASSO solutions is associated with discarded regressors. Therefore, the sparse solutions, associated with large λ resulting in small complexity Q_{rbf}, simultaneously imply smaller number of the basis functions than M, naturally reduce the structure of RBF networks.

Another remarkable property of the LASSO solutions is their piecewise-linear regularization path [Park and Hastie, 2007]. The regularization path of (4.16) is the set of its minimizers for all choices of $\lambda \in \mathbb{R}^+$, which is in turn the set of weight (coefficient) vectors associated with the Pareto-optimal solutions $\mathcal{P}(F_{\sigma,C_M}, \phi)$ of the corresponding convex subproblem.

Since the regularization path is piecewise-linear, all its elements can be represented exactly by a finite sequence of its nodes. The latter can be efficiently calculated by means of the so-called least angle regression shrinkage (LARS) algorithm developed in [Efron, Hastie, Johnstone, and Tibshirani, 2004]. Given the $N \times M$ design matrix H of regressors and the $N \times 1$ target vector Y, the LARS procedure

$$(p_j)_j = \mathrm{LARS}(H, Y) \tag{4.17}$$

computes the sequence $(p_j)_j$ of vectors in $\mathbb{R}^M$, corresponding to nodes of the piecewise-linear path. The first node of the path is the null vector $p_0 = 0$ and the last element is the OLS solution $p_{ols} = (H^T H)^{-1} H^T Y$. The connection of the piecewise-linear regularization path of the LASSO with the front of Pareto optimal solutions is demonstrated schematically in Fig. 4.1 for the two-dimensional case.

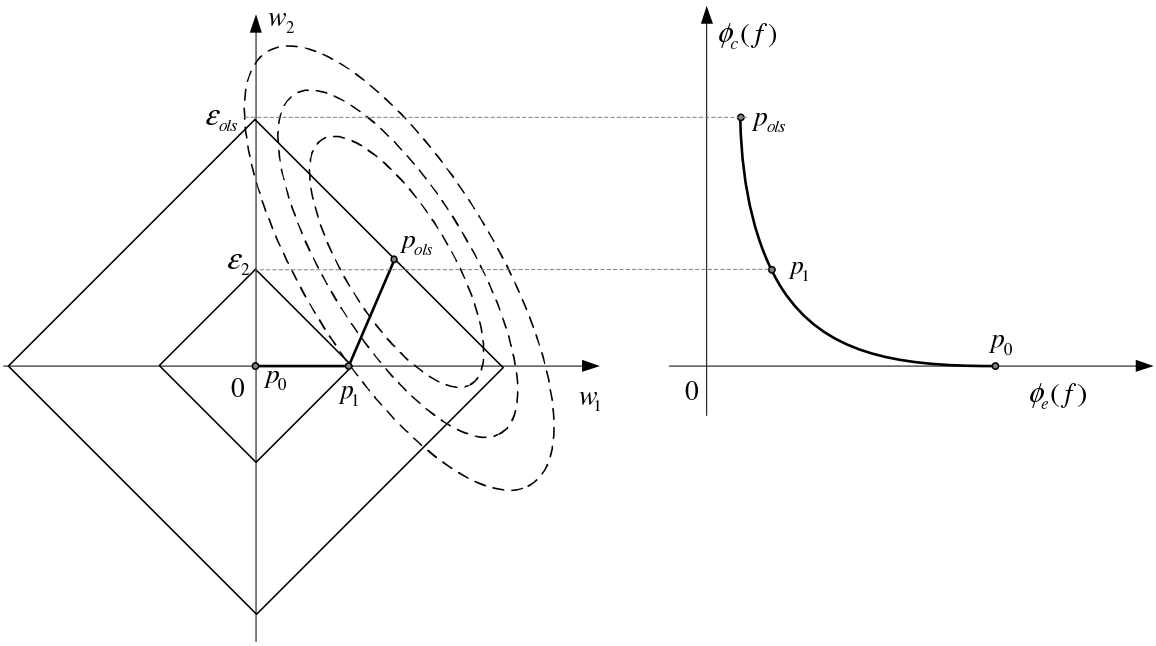

Figure 4.1: Schematic demonstration of the the LASSO regularization path (left) and the corresponding Pareto front (right).

Any element of the regularization path can be calculated exactly from the linear interpolation between the corresponding pair of nodes. In particular, the complete Pareto set of the convex subproblem $\mathcal{P}(F_{\sigma,C_M}, \phi)$ in the domain $\mathbb{R}^M$ can be expressed in terms of the sequence $(p_j)_j$, namely,

$$\mathcal{P}\big(\mathbb{R}^M, (R_{\text{emp}}, Q_{\text{rbf}})\big) = \Big\{\beta p_j + (1-\beta)p_{j+1}\Big|\beta \in [0,1], (p_j, p_{j+1}) \in (p_j)_j\Big\}, \quad (4.18)$$

which can be determined entirely with a single run of the LARS algorithm.

According to the definitions (4.14) and (4.15), the objective functions $Q_{\text{rbf}}(\alpha)$ and $R_{\text{emp}}(\alpha)$ are linear and quadratic functions of α, respectively. Hence, the entire Pareto front curve in $\mathbb{R}^2$ can also be determined exactly via interpolations of the corresponding orders. Introducing the precomputed vectors

$$\rho_0 = \begin{bmatrix} Y^T Y \\ \|p_j\|_1 \end{bmatrix}, \quad (4.19)$$

$$\rho_1(p_j, p_{j+1}) = \begin{bmatrix} -2Y^T H(p_{j+1} - p_j) \\ \|p_{j+1}\|_1 - \|p_j\|_1 \end{bmatrix}, \quad (4.20)$$

and

$$\rho_2(p_j, p_{j+1}) = \begin{bmatrix} \|H(p_{j+1} - p_j)\|^2 \\ 0 \end{bmatrix} \quad (4.21)$$

which are parameters of the quadratic segments, the Pareto front curve associated with the convex subproblem $\mathcal{P}(F_{\sigma,C_M}, \phi)$, is given by

$$\rho\left(\mathbb{R}^M, \phi\right) = \Big\{\rho_0 + \beta\rho_1(p_j, p_{j+1}) + \beta^2\rho_2(p_j, p_{j+1})\Big|\beta \in [0,1], (p_j, p_{j+1}) \in (p_j)_j\Big\}. \quad (4.22)$$

In such a way, both the Pareto set (4.18) and its frontier (4.22) of the convex subproblem are determined in terms of the results of the LARS algorithm for their further incorporation into the global Pareto set by means of decomposition, associated with the MOBJ problem.

4.3.3 Treating the bias parameter

In the very beginning of current development, the class of functions (4.9) of RBF networks is introduced without the bias term ($+b$), which is of great practical importance.

Since the bias parameter does not affect smoothness, the complexity measure Q_{rbf} obviously remains independent on it. However, the introduction of the bias parameter affects the empirical risk R_{emp}, influencing the mean squared error. Hence, after the introduction of the bias parameter b as the additional parameter of hypothesis one should provide its estimation, that may need rewriting the whole chain of development. Instead, it can be shown that $b = 0$ always corresponds to the optimal least squares estimate of the bias when the data (design matrix H and the target vector Y) are centered. Consequently, regression procedures can be applied to the centered data without the bias parameter, whose value can be restored later.

Suppose the minimizer of (4.16) is found for the centered design matrix

$$\overline{H} = H - U\mu(H)$$

and centered target vector

$$\overline{Y} = Y - U\mu(Y).$$

Here, $\mu(H)$ is the $1 \times M$ mean row vector of H, $\mu(Y)$ is the scalar mean of Y, and $U = (1, 1, \ldots, 1)^T$ is the $N \times 1$ vector of units. For arbitrary α, the mean squared error $R_{\text{emp}}(\alpha)$ of the hypothesis response

$$\widetilde{Y} = H\alpha + b$$

can be rewritten as

$$\begin{aligned} R_{\text{emp}}(\alpha) &= \left\| Y - \widetilde{Y} \right\|^2 = \|Y - H\alpha - b\|^2 \\ &= \left\| \overline{Y} - \overline{H}\alpha \right\|^2 = \|Y - H\alpha - U\mu(Y) + U\mu(H)\alpha\|^2, \end{aligned}$$

providing the bias parameter

$$b = \mu(Y) - \mu(H)\alpha. \tag{4.23}$$

Consequently, when the centered matrices $\overline{H}$ and $\overline{Y}$ are passed to the LARS procedure (4.17) and the bias parameters are later restored with (4.23) for the elements of the regularization path, the resulting hypothesis' parameters α and b are minimizers of

$$R_{\text{reg}}(\alpha) = \|Y - H\alpha - b\|^2 + \lambda\|\alpha\|_1. \tag{4.24}$$

This in turn corresponds to the solution of particular convex subproblem $\mathcal{P}(F_{\sigma,C_M}, \phi)$ obtained for the class of hypotheses functions

$$f(x) = \sum_i \alpha_i k_\sigma(x - c_i) + b$$

with the bias parameter b.

4.3.4 MOBJ-RBF algorithm

Before combining the components of the MOBJ algorithm together, it is necessary to introduce several additional parameters. First, one should define the grid for a search along σ, that can be an arbitrary finite discrete sequence $(\sigma_j)_j$. For example, the exponentially spaced grids, such as

$$\log_2 \sigma_j = \left(1 - \frac{j}{R_\sigma}\right) \log_2 \sigma_{\min} + \frac{j}{R_\sigma} \log_2 \sigma_{\max},\ j = 0 \dots R_\sigma \tag{4.25}$$

are common. Here R_σ determines the size (resolution) of the grid and $\sigma_{\min}$ and $\sigma_{\max}$ determine its range. While R_σ characterizes only a resolution and thus has a weak influence on the results, both $\sigma_{\min}$ and $\sigma_{\max}$ must be determined more carefully, in order to cover the whole range of Pareto-efficient solutions or at least its important part[4], associated with the distribution of distances in the particular data-set. The range parameters $\sigma_{\min}$ and $\sigma_{\max}$ can be selected either empirically or using a certain heuristic rule. For instance, one can select $\sigma_{\min}$ and $\sigma_{\max}$ proportional to the minimum and maximum distances between distinct training input vectors, respectively. Also, one can consider the more robust heuristic, based on the empirical distribution of distances in the data-set. For instance, the selection of $\sigma_{\min}$ and $\sigma_{\max}$ according to

[4]Usually, one is not interested in evaluation of extreme solutions. Instead, the solutions near the ideal point are likely to represent good decisions [Forrest, 1993].

10th and 90th percentiles of the distance distribution, respectively, provides an outlier resistant estimate for width range.

The matrix of centers must contain distinct training vectors. However, in order to ensure numerical conditionality of design matrices, it is necessary to consider ε-distinct vectors instead, where ε is a small constant. For example, the choice of

$$\varepsilon = \sigma_{\max}\sqrt{-2\log(1-l)}$$

guarantees that any two Gaussian basis functions do not overlap at a level higher than $1-l$, and thus the elements of the design matrix are l-distinct.

The second useful parameter is the maximum complexity limit $Q_{\max}$. The regularization path of LASSO regression spreads from the empty solution (null weights) until reaching the most complex one, OLS estimate. The OLS solution is not regularized and, thus, is usually ill-conditioned and associated with infinitely large weights, making the LARS algorithm numerically instable at the end of regularization path. Since the solutions that, close to the ill-conditioned OLS point, unlikely represent a good generalization, they can be discarded without a loss of representability of the hypothesis space. Therefore, the stopping condition $\|\alpha\|_1 > \sigma^2 Q_{\max}$ must be considered within the LARS procedure in order to ensure numerical safety of the results and also to reduce the time of computation. The value of $Q_{\max}$ also can be selected empirically (note that the norm of the weights $\|\alpha\|_1 \leq \|Y\|_1$ is likely to be sufficient for approximations of Y with the Gaussian functions) or increased progressively, until reaching the critical numerical precision of the results.

Aiming at application of the central idea of decomposition (4.13), the search of nondominated elements within the union of $\mathcal{P}\big(F_{\sigma_j,C_M}\big)$, $j = 1,\ldots,R_\sigma$ generally requires one to find all intersections of the corresponding piecewise-quadratic Pareto fronts (4.22). As the result, the final Pareto front of the MOBJ problem is also piecewise-quadratic. However, due to numerical nature of the most model selection criteria, there is no benefit in dealing with analytical representation of the Pareto front with a number of continuous segments. Instead, it is sufficient to introduce the linear grid $(q_i)_i \in [0, Q_{\max}]$ of complexity levels with the resolution R_Q. Then, the nondominated elements of the whole problem can be computed by means of finding the minimum of the training error R_{emp}, within the equicomplex elements, while

the Pareto-optimal elements corresponding to the points $(r_{ij}, q_i) \in \rho\left(F_{\sigma_j, C_M}\right)$ can be found from (4.22) for each j-th subproblem and i-th complexity value.

Finally, gathering up the above considerations, the MOBJ-RBF algorithm can be stated as follows:

1. Initialization: given the training set $Z_{\text{tr}}^N = \left\{< x_i, y_i >\right\}$, and the parameters $Q_{\max}$, R_Q and R_σ do:
   (a) Find the $(M \times n)$-matrix C_M of ε-distinct input vectors x_i from Z_{tr}^N.
   (b) Determine the range of widths selecting $\sigma_{\min}$ and $\sigma_{\max}$ by the corresponding estimates, e.g., $\sigma_{\min} = \min_{i \neq j} \|x_i - x_j\|$ and $\sigma_{\max} = \max_{i \neq j} \|x_i - x_j\|$.
   (c) Calculate the mean $\mu(Y)$ and centered target vector $\overline{Y} = Y - \mu(Y)$.
2. For $j = 0 \ldots R_\sigma$, calculate the corresponding elements of the grid σ_j (e.g., using (4.25)), and find the elements of the Pareto set $\mathcal{P}(F_{\sigma_j, C_M})$ of the j-th subproblem:
   (a) Calculate the $(N \times M)$ design matrix H for N input patterns from Z_{tr}^N, the centroid matrix C_M and width parameter σ_j of the Gaussian RBF functions k_σ.
   (b) Calculate the mean row vector $\mu(H)$ and the centered design matrix $\overline{H}$.
   (c) Find the sequence $(p_l)_l = \text{LARS}(\overline{H}, \overline{Y})$ of the piecewise-linear LASSO regularization path, until the element $\|p_k\|_1 > \sigma_j^2 Q_{\max}$ is reached.
   (d) Compute the curve parameters (4.19), (4.20), and (4.21) of quadratic segments of the piecewise-quadratic Pareto front (4.22).
3. Combine solutions of the convex subproblems. For $i = 1 \ldots R_Q$ and the corresponding complexity magnitudes on the grid, e.g., $q_i = \frac{Q_{\max}}{R_Q} i$ do:
   (a) Find the weights of the element $f_{ij} \in \mathcal{P}(F_{\sigma_j, C_M})$ from (4.18), such that $Q_{\text{rbf}}[f_{ij}] = q_i$ and its corresponding mean squared error $r_{ij} = R_{\text{emp}}[f_{ij}]$ by means of quadratic interpolation within the corresponding segment of (4.22).
   (b) Find i-th globally-nondominated element $f_i = \arg\min_{j=0 \ldots R_\sigma} q_i$.

The set of nondominated elements f_i, $i = 0 \dots R_Q$ is the sought approximation of the Pareto-set of the whole problem.

4. Restore the bias parameter using (4.23) for the resulted hypotheses f_i and apply the model selection criterion ζ to determine the final solution

$$f_{\text{mobj}} = \arg \min_{i=0\dots R_Q} \zeta(f_i).$$

4.4 Model selection criteria

The proposed MOBJ algorithm approximates the Pareto set containing a wide spectrum of RBF networks: from zero up to M basis functions; from sharp ($\sigma_{\min}$) to smooth ($\sigma_{\max}$) Gaussian functions. Hence, generalization properties of the final solution substantially depend on model selection criterion, that plays the role of decisor.

The model selection criteria such as the MVE (2.47) can be applied directly (see 2.5.3). More reliable but computationally hard T-fold CV procedure (2.48) can be also adopted in the form

$$f^{\text{cv}}_{\text{mobj}} = \arg \min_{i=0\dots R_Q} \sum_{l=1}^{T} \zeta^{\text{val}}_{Z_l} \left[\text{LASSO}(H_{il}, Y, \sigma_i^2 Q_{\text{rbf}}[f_i]) \right], \tag{4.26}$$

where Z_l is $l-$th non-overlapping subset of the training set Z^N_{tr}. Here in (4.26), H_{il} corresponds to the design matrix of the hidden layer of i-th Pareto-optimal hypothesis f_i, calculated for input vectors of the training set $Z^N_{\text{tr}} \backslash Z_l$, σ_i is the width of the Gaussian basis functions associated with f_i. The results of the procedure $\text{LASSO}(H, Y, \epsilon)$ in (4.26) correspond to the hypotheses with the same basis functions as f_i, but another weights α and bias b, minimizing $\|Y - H\alpha - b\|^2$ subject to the Lasso constraint $\|\alpha\|_1 \leq \epsilon$. In fact, such adaptation of the CV is built on the assumption that the widths of basis functions of nondominated solutions may not change drastically after an exclusion of any of T subsets Z_l from the training set. Otherwise, it would require T runs of the entire MOBJ algorithm for the direct computation of CV.

Unfortunately, the application of approximation of the leave-one-out error proposed in [Bousquet and Elisseeff, 2002] is not adequate in the current context, since the LASSO regression, being sparse, is not uniformly-stable [Xu, Caramanis, and

Mannor, 2008]. However, the application of the almost computationally-free information criterion such as AIC [Akaike, 1974] or BIC [Schwarz, 1978] is possible, since both rely on the balance of degrees of freedom and the likelihood of estimates, independently on the ordering or structure of the candidate models [Burnham and Anderson, 1998].

In a general form of AIC and BIC, one seeks the minimizers of

$$\mathrm{AIC}[f] := 2\mathrm{df}[f] - 2\ln L[f]$$

and

$$\mathrm{BIC}[f] := \ln(N)\mathrm{df}[f] - 2\ln L[f],$$

respectively, within the set of competitive solutions. Here $\mathrm{df}[f]$ is the effective number of degrees of freedom and $L[f]$ is the maximized likelihood function for the given model f. As suggested in [Burnham and Anderson, 1998], the second-order correction

$$\mathrm{AIC}_c[f] = \mathrm{AIC}[f] + \frac{2\mathrm{df}[f](\mathrm{df}[f]+1)}{N - \mathrm{df}[f] - 1} \tag{4.27}$$

is necessary for preventing the AIC from overfitting on short data-sets (e.g., when $\frac{N}{\mathrm{df}[f]} < 40$).

In the ridge regression, the unbiased estimate of $\mathrm{df}[f]$ corresponds to the trace of the inverse of covariance matrix of regressors. In the case of LASSO regression, the unbiased estimate $\hat{\mathrm{df}}[f] = m_r$ for $\mathrm{df}[f]$ has been proved in [Zou, Hastie, and Tibshirani, 2007], where m_r is the number of non-zero weights. Since the spectrum of Pareto-optimal RBF networks may contain large models such that $\frac{N}{\mathrm{df}[f]} \simeq 1$, the corrected AIC_c (4.27) should be always used.

In our setting, the generalized form

$$\zeta(f,\tau) = \tau m_r - 2\ln L[f], \tag{4.28}$$

of the information criterion can be introduced, and the AIC and BIC model selection criteria are denoted hereafter as

$$\zeta^{\mathrm{AIC}}[f] = \zeta\left(f, \frac{2N}{N - m_r - 1}\right) \tag{4.29}$$

and

$$\zeta^{\mathrm{BIC}}[f] = \zeta\left(f, \ln(N)\right), \tag{4.30}$$

respectively.

In (4.28), the likelihood $L[f]$ depends on a particular settings of the learning problem and, thus, should be treated separately for regression and classification cases.

4.4.1 Regression

In regression tasks, the additive noise model

$$y_i = f^{\circ}(x_i) + e_i, \quad i = 1, \ldots, N, \tag{4.31}$$

is commonly considered, where $f^{\circ}(x_i)$ is an unknown true hypothesis (regression function) and y_i is the output observation from the training data-set, corresponding to x_i. Here, the random noise components $e_i \sim (0, \sigma_{ns}^2)$ are assumed to be independent and normally distributed with the variance σ_{ns}^2.

As known, the maximized log-likelihood functional

$$\ln L[f] = -\frac{1}{2} N \ln(2\pi\sigma_{ns}^2) - \frac{1}{2\sigma_{ns}^2} \sum_{i=1}^{N} e_i^2 \tag{4.32}$$

corresponds to the minimum of the sum of squared errors $\sum_{i=1}^{N} e_i^2$ (SSE), $e_i = y_i - f(x_i)$, associated with the given candidate hypothesis f. The combination of (4.32) with (4.28) yields the criterion

$$\zeta_{\mathrm{reg}}(f, \tau) = \tau m_r + N \ln(2\pi\sigma_{ns}^2) + \frac{1}{\sigma_{ns}^2} \sum_{i=1}^{N} e_i^2,$$

or

$$\zeta_{\mathrm{reg}}(f, \tau) = \tau\sigma_{ns}^2 m_r + R_{\mathrm{emp}}[f],$$

after multiplication by σ_{ns}^2 and omitting the constant terms.

The noise variance σ_{ns}^2 is usually unknown, however, its unbiased estimate

$$\hat{\sigma}_{ns}^2 = \frac{1}{N - \mathrm{df}[f]} \sum_{i=1}^{N} e_i^2 = \frac{R_{\mathrm{emp}}[f]}{N - m_r}$$

can be used. Then, the information criterion based on the variance estimate $\hat{\sigma}^2_{ns}$ is

$$\zeta'_{\text{reg}}(f,\tau) = \left(\frac{\tau m_r}{N-m_r}+1\right) R_{\text{emp}}[f]. \tag{4.33}$$

4.4.2 Classification

Consider the binary classification task with the labels $\{-1,+1\}$. In this case, the true hypothesis $f^\circ(x_i)$ and its observed output response y_i are binary. The assumption of the additive noise $e_i = y_i - f^\circ(x_i)$ distributed among possible states $\{-2,0,2\}$ contradicts its independence, leading to the distribution of y_i among the values $\{-3,-1,0,1,+3\}$. Consequently, the assumption (4.31) is not valid for the case of classification, i.e., the noise is not additive in this case.

On the other hand, the likelihood function of the classification error can be explicitly written as

$$L[f] = \eta^{E_c[f]}(1-\eta)^{N-E_c[f]},$$

where η is the probability of incorrectly labeled (misclassified) observation and $E_c[f]$ is the apparent number of misclassifications produced by the candidate hypothesis f. Substitution of the likelihood function into (4.28) yields the criterion

$$\zeta_{\text{cls}}(f,\tau) = \tau m_r - 2E_c[f]\ln(\eta) - 2(N-E_c[f])\ln(1-\eta).$$

Similarly to the regression case, one passes from the unknown probability η to its unbiased estimate $\hat{\eta} = \frac{E_c[f]}{N-m_r}$, which, after a simplification and removing a constant term, yields the practical criterion

$$\zeta'_{\text{cls}}(f,\tau) = \tau m_r + 2E_c[f]\ln\left(\frac{N-m_r}{E_c[f]}-1\right). \tag{4.34}$$

4.5 Experiments

This section provides the experimental study of the proposed MOBJ-RBF algorithm and its properties. In the first part, the MOBJ algorithm is tested on synthetic datasets, whose distributions are known. The last part demonstrates the results of the MOBJ algorithm for several real-world data benchmarks compared to the results of several existing learning algorithms.

4.5.1 Twin spiral

The "twin spiral" is the classical two-dimensional problem for learning machines [Lang and Witbrock, 1988]. In the original setting (experiment 1), the data-set consists of 194 training patterns lying on two non-intersecting spirals (97 points each), representing a problem of high nonlinear separability of patterns. Since the problem is two-dimensional, the generalization properties of obtained solutions can be evaluated from the visual analysis of the shape of separation surfaces.

The training set is sparse and almost all patterns are needed for the correct reconstruction of the spiral with radial-basis functions. Hence, in order to verify the ability of the MOBJ algorithm to generalize the data with sparse models, other test (experiment 2) is performed on the larger data-set of 582 samples, in which the original training set is replicated three times with addition of a small amount of Gaussian noise.

In both experiments the same settings of the MOBJ algorithm were used: the range parameters $\sigma_{\min} = 0.2$, $\sigma_{\max} = 2$, and $Q_{\max} = 1000$ and the resolutions $R_\sigma = 100$ and $R_Q = 500$. Five nondominated elements were evaluated for comparison: the overfitting and underfitting solutions with complexities $Q_{\mathrm{rbf}}[f] \approx 1000$ and $Q_{\mathrm{rbf}}[f] \approx 100$, respectively; the minima of the AIC (4.29), BIC (4.29) and 10-fold CV (4.26).

The Pareto fronts and selected solutions are shown in Fig. 4.2 and Fig. 4.3 for the original (experiment 1) and redundant (experiment 2) training sets, along with the corresponding separation surfaces. The distributions of the magnitudes of model selection criteria along the Pareto sets obtained in experiments 1 and 2 are plotted in Fig. 4.4.

In Table 4.1, numerical results for both the experiments are shown. Here the column m_r corresponds to the numbers of basis functions with non-zero weights, i.e., the apparent sizes of obtained RBF networks; Q_{rbf}, σ, and $\|w\|_1$ are the values of complexity measure, widths of the basis functions and sizes of the weight vectors of corresponding solutions. The last column represents the root mean squared errors (RMSE) on the corresponding training sets.

As seen from results for both the experiments the AIC and BIC solutions are identical and demonstrate good generalization. On the other hand, the 10-fold CV fails in the experiment 1, whereas in the experiment 2 the results are close to AIC

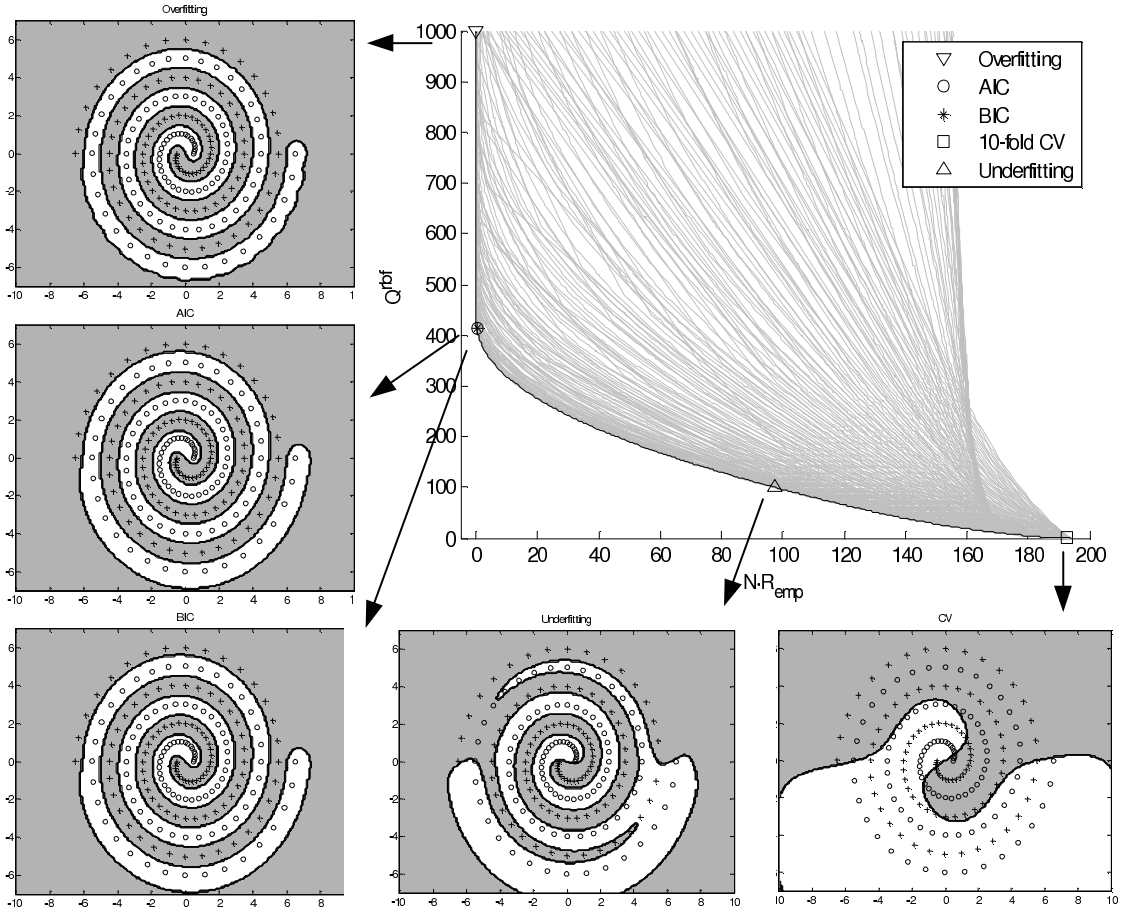

Figure 4.2: *Twin spiral* experiment 1: 194 samples.

Table 4.1: Twin-spiral benchmark results

Experiment	Solution	Properties				RMSE
		m_r	Q_{rbf}	σ	$\|w\|_1$	
1	Overfitting	183	997.04	0.50	244.71	0.0005
	AIC, BIC	153	413.09	0.67	185.61	0.0335
	10-fold CV	14	0.09	2.00	0.37	0.9967
	Underfitting	145	101.18	0.67	45.46	0.7064
2	Overfitting	261	996.60	0.51	256.25	0.4201
	AIC, BIC	165	473.58	0.63	185.08	0.4917
	10-fold CV	176	508.46	0.61	189.68	0.4828
	Underfitting	101	100.29	0.70	49.46	0.7774

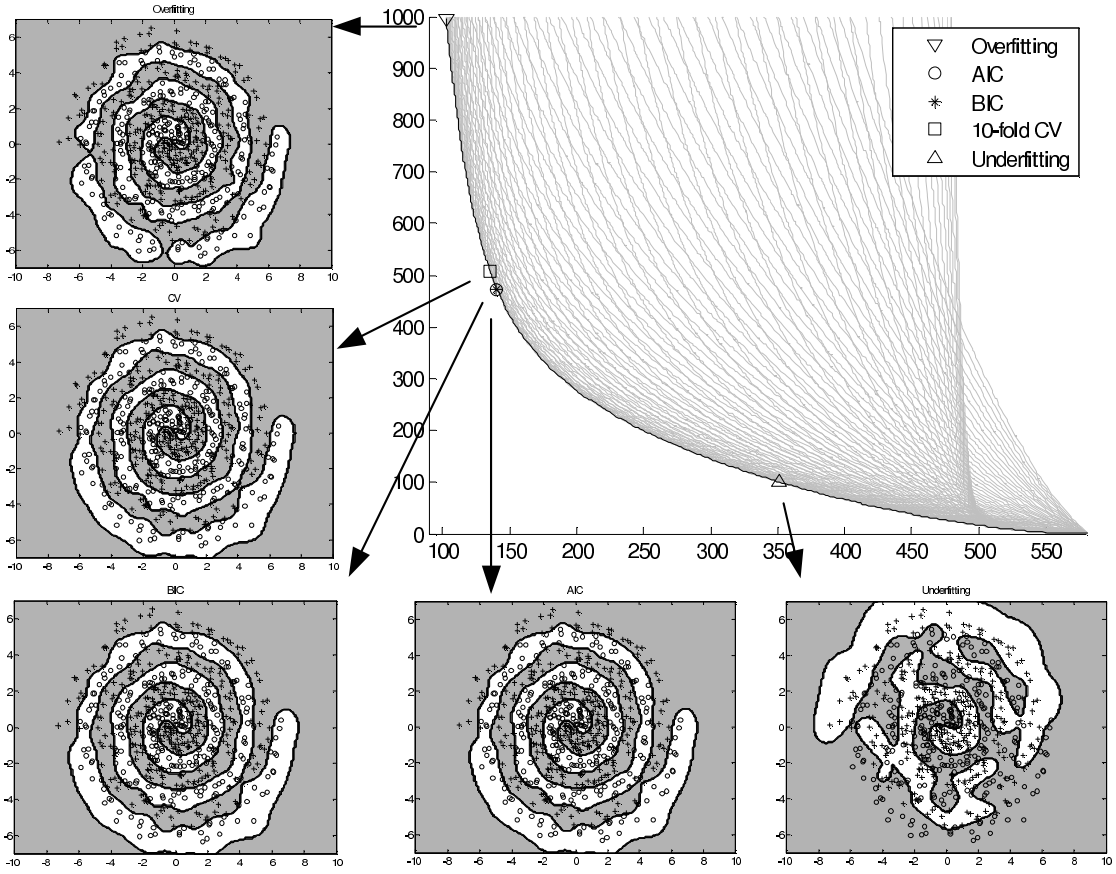

Figure 4.3: *Twin spiral* experiment 2: 582 samples with noise.

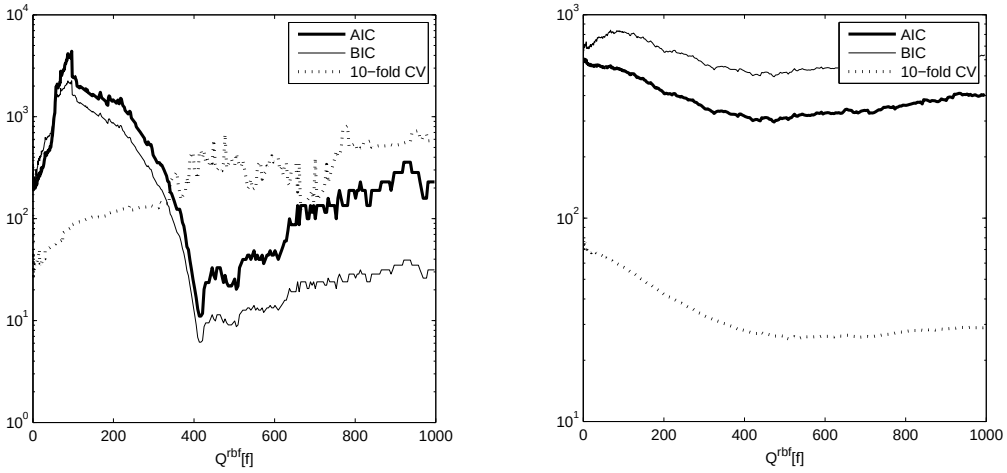

Figure 4.4: Distribution of the values of model selection criteria along the Pareto sets in experiments 1 (left) and 2 (right).

and BIC. Moreover, the quantitative results indicate that the properties of solutions (number of basis functions, their width, and complexity) found by AIC and BIC remain similar in both experiments. This fact demonstrates that the MOBJ algorithm detects that the data of the both experiments are i.i.d. from the same distribution, provided by similar models independently to the training set.

The underfitting solutions are seen to provide false generalizations in both cases, whereas the overfitting effects are weak. This behavior is specific for this particular problem, since good solutions are expected to be complex (i.e., include most part of the training set). The hypotheses suffering from stronger overfitting exist, but they are not revealed in the results of the MOBJ algorithm since being dominated.

4.5.2 Noised *sinc* regression

Another experiment with synthetical data is aimed to demonstrate of the influence of length and level of noise disturbance on the generalization capabilities of the MOBJ algorithm, endowed with various model selection criteria.

In the experiment, the training and validation sets are non-overlapping and have equal numbers of samples, generated from the equation

$$y_i = \frac{\sin(\pi x_i)}{\pi x_i} + e_i \tag{4.35}$$

for the range $x \in (0, 4\pi]$, where $e_i \sim (0, \sigma_n^2 s)$ is the normally distributed random noise.

For each combination of three data-set sizes $N \in \{50, 100, 200\}$ and three noise variances $\sigma_{ns} \in \{0.1, 0.2, 0.4\}$, the $3 \times 3 \times 100$ training and validation sets were generated: each of 100 realizations of Gaussian noise was used to produce 9 pairs of non-overlapping training and validation sets for all combinations of three sizes and three variances. The generated data were split in such a manner that the training and validation sets, corresponding to the same noise realizations, were nested: the training set of the length N is the concatenation of the training and validation sets of the lengths $N/2$. Such a nested partitioning were made with the purpose of the further comparison of validation and information criteria in conditions of equal amount of information supplied to a learning algorithm.

The 1000 test samples of (4.35) without noise were used for estimation of the true (test) regression error of obtained solutions. For comparative reasons, the regularized orthogonal forward selection (ROFS) [Orr, 1999] algorithm for RBF networks was used. Although ROFS is based on single-objective approach, it acts similar to the MOBJ algorithm: the weights and centers of the basis functions are determined by the forward selection, width of Gaussian radial basis functions and regularization strength are determined by the grid search, in accordance with a certain model selection criterion.

In the MOBJ algorithm, the settings $R_\sigma = 100$, $R_Q = 1000$ and $Q_{\max} = 150$ were used in all tests. The range parameters also remain fixed with the values $\sigma_{\min} = 0.05$ and $\sigma_{\max} = 6$ for all tests. In order to maintain equivalence of the hypothesis spaces of MOBJ and ROFS, the same sequence of widths, calculated from (4.25), was supplied to both the algorithms.

In experiment, the MOBJ algorithm was endowed with the AIC, BIC, and MVE model selection criteria and compared to the ROFS with the BIC and generalized cross-validation (GCV). The validation sets were used only with the MVE criterion (MOBJ-MVE), whereas only the information available from the training sets were used for obtaining the rest of the solutions (MOBJ-AIC, MOBJ-BIC, ROFS-BIC, and ROFS-GCV).

In Table 4.2, the test errors and its dispersions averaged over 100 noise realizations, are listed for all combinations of data-set sizes and noise variances (best values are bold). The fragments of Pareto fronts and final solutions associated with a single noise realization are demonstrated on Fig. 4.5 for all combinations of experiment settings.

The synthetically generated validation sets can be considered representative, since they are i.i.d. from the same distributions as the training sets and have equal lengths. This explains why the MOBJ-MVE solutions are ranked first, whereas the MOBJ-AIC and MOBJ-BIC share the second place. However, since the data-sets were nested, one can compare the results of MOBJ-AIC and MOBJ-BIC with those of MOBJ-MVE under the conditions of equal amount of information passed to the learning algorithm. In this case, for example, one should compare the results of the MOBJ-MVE from the first column ($N = 50$) with those of MOBJ-AIC and MOBJ-BIC from the second column ($N = 100$). Such a comparison is made under assumption of equivalent

Table 4.2: Results for the *sinc* regression benchmark: test NRMSE$\times 10^2$ (mean) and its standard deviation (std.)

σ_{ns}	Method	$N = 50$		$N = 100$		$N = 200$	
		mean	std.	mean	std.	mean	std.
0.1	MOBJ-AIC	2.76	0.723	1.84	0.544	1.29	0.386
	MOBJ-BIC	2.72	0.684	1.86	**0.524**	1.32	0.38
	MOBJ-MVE	**2.51**	**0.624**	**1.78**	0.545	**1.2**	**0.301**
	ROFS-GCV	8.97	4.03	5.97	2.4	3.87	1.39
	ROFS-BIC	9.79	4.48	4.46	3	1.41	0.41
0.2	MOBJ-AIC	4.25	1.23	2.81	0.786	1.97	0.572
	MOBJ-BIC	4.09	1.15	2.77	0.782	1.98	0.516
	MOBJ-MVE	**3.89**	**1.04**	**2.64**	**0.758**	**1.8**	**0.443**
	ROFS-GCV	11.6	2.93	9.1	1.67	6.75	1.17
	ROFS-BIC	11.5	3.23	6.66	3.22	2.51	0.902
0.4	MOBJ-AIC	5.49	1.76	3.54	1.06	2.53	0.759
	MOBJ-BIC	5.4	1.8	3.53	1.16	2.44	**0.565**
	MOBJ-MVE	**5.11**	**1.61**	**3.51**	**1.05**	**2.28**	0.619
	ROFS-GCV	12.3	2.42	10.5	1.64	8.54	0.887
	ROFS-BIC	12.2	2.84	8.66	3.03	3.95	1.69

amount of information available from 100 samples supplied for MOBJ-AIC, MOBJ-BIC, and MOBJ-MVE, however the latter learning machine uses only one half (50 samples) for training and another half for validation. Surprisingly, the performance comparison of MOBJ-MVE with MOBJ-AIC and MOBJ-BIC on double sample (the values from the next column in Table 4.2) demonstrates a strong advantage of the information criteria. Consequently, the experiment shows that using all the available data for training and model selection with an information criterion is often more reliable than a consumption of a part of the training data for validation.

The results in Fig. 4.5 also demonstrate a trend of the Pareto fronts to become irregular with the growth in uncertainty level: the Pareto fronts are mostly non-convex below the main diagonal of Fig. 4.5 (small N or large σ_{ns}) and convex above (large N or small σ_{ns}).

4.5.3 Wisconsin breast cancer

Another well-known benchmark contains real data obtained from microscopic examination of clinical patients. The Wisconsin breast cancer data-set [Asuncion and

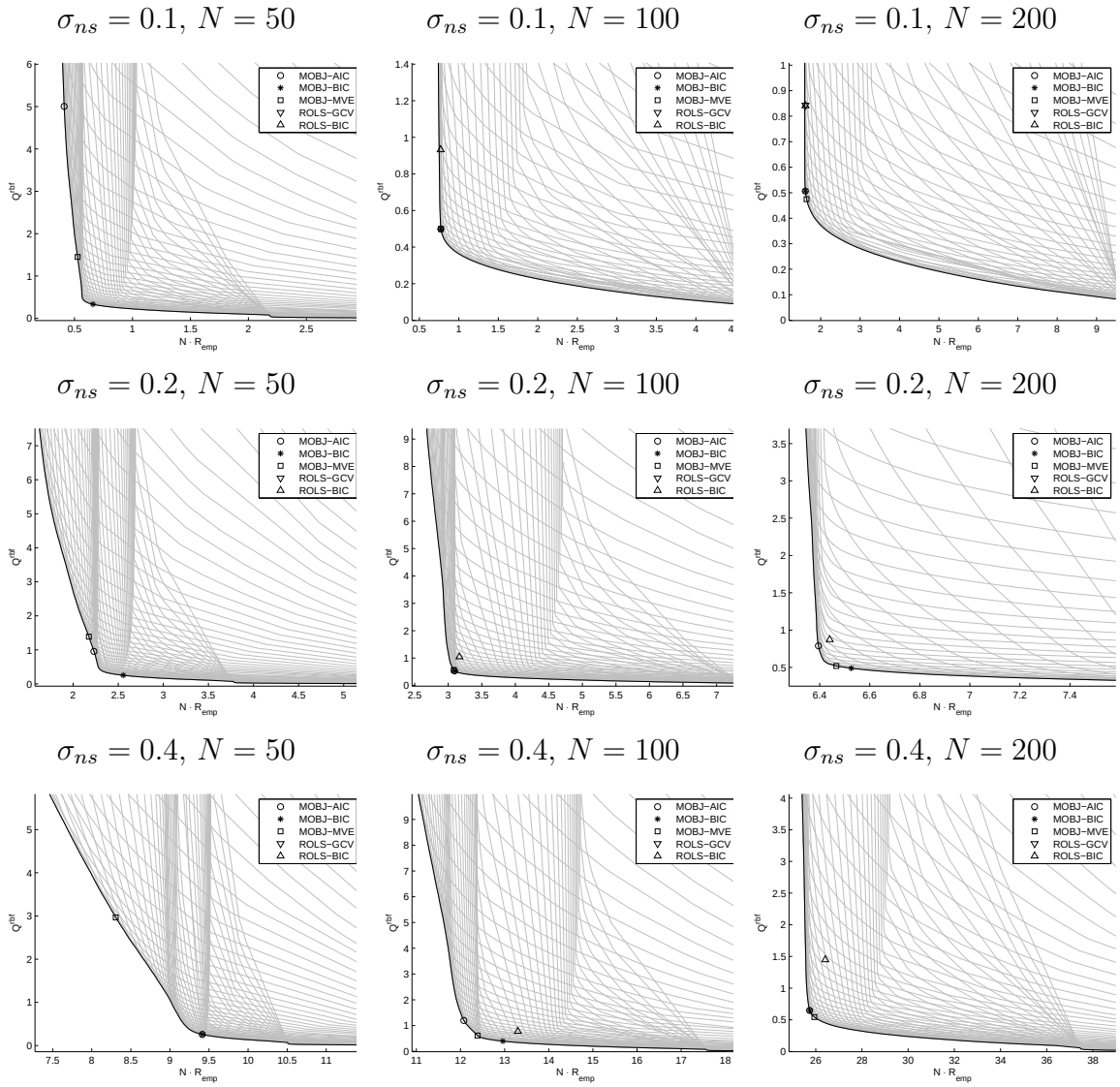

Figure 4.5: The fragments of Pareto fronts from the *sinc* experiment, corresponding to a particular noise realization.

Newman, 2007] contains 699 patterns describing the benign and malignant (cancer) tissue classes, via 9 real-valued attributes. In order to maintain the comparability and reproductivity of the results, the data-set partitions from the Proben1 benchmark set [Prechelt, 1994] were used, where 350 samples are selected for training, 175 for validation, and 174 for test.

For all partitions the parameters of the MOBJ algorithm $Q_{\max} = 100$, $R_\sigma = 50$, and $R_Q = 250$ were used. The ranges of widths were independently calculated for

each partition according to the relations $\sigma_{\min} = \frac{1}{3}d_5$ and $\sigma_{\max} = 3d_{95}$, where d_p is the $p-$th percentile of the distance distribution of input patterns in the corresponding training set.

The classification results were obtained for each one of three different data partitions: *cancer1*, *cancer2*, and *cancer3*. Correspondingly, 251, 247, and 239 distinct prototypes were found from the training sets. The MOBJ algorithm was tested with the AIC, BIC, MVE, and 10-fold CV model selection criteria were taken in the same manner as in previous experiments. In addition, the results are compared with the standard C-SVC algorithm with of the Gaussian kernel implemented in LIBSVM [Chang and Lin, 2001], whose hyperparameters were estimated with the common grid search and 10-fold CV technique. The widths of Gaussian kernel were selected from the same grid (4.25), used in the MOBJ algorithm. Hence, the hypothesis spaces of both the learning machines are equivalent. The regularization parameter C of the SVM was selected from the exponentially spaced grid in range $C \in [2^{-8}, 2^{12}]$ with the resolution of 50 steps.

The results for all three partitions are shown in Table 4.3, where the classification accuracy on the corresponding test sets is demonstrated with the numbers of false-positive (F_p) and false-negative (F_n) classifications, along with total correct classification rates. The description of the columns applies from Table 4.1 of the previous experiment, while the values of m_r associated with the SVM correspond to the numbers of the obtained support vectors. The fragments of Pareto fronts and the obtained solutions are shown in Fig. 4.6 for all data partitions.

As seen from the results, MOBJ solutions are almost as accurate as the solutions found by the standard SVM technique. In fact, the performances of the MOBJ-CV solutions are superior, while the SVM-CV solutions were obtained by the exhaustive search with the same model selection criterion in the same hypothesis space. This fact experimentally supports the claim given in section 2.6 about the redundancy of the exhaustive model selection, from the SRM point of view.

The precision of model selection with the information criteria AIC and BIC is seen to be slightly lower than the that of CV, however the resulting models found with AIC and BIC are surprisingly small (only 1 to 5 basis functions), in contrast to the significantly larger models, when compared to the other methods.

Table 4.3: Wisconsin breast cancer benchmark results

Part.	Method	Properties				Accuracy		
		m_r	Q_{rbf}	σ	$\|\alpha\|_1$	F_p	F_n	Total (%)
*cancer*1	MOBJ-AIC,BIC	2	0.93	5.94	32.7	4	0	97.7
	MOBJ-MVE	45	80.3	0.56	24.8	3	0	**98.3**
	MOBJ-CV	7	4.32	0.91	3.58	3	0	**98.3**
	SVM-CV	93	9.17	1.35	16.76	4	0	97.7
*cancer*2	MOBJ-AIC	5	1.16	5.70	37.8	1	5	**96.6**
	MOBJ-BIC	1	0.81	5.70	26.3	4	5	94.8
	MOBJ-MVE	16	8.37	0.81	5.43	3	4	96.0
	MOBJ-CV	9	4.65	0.89	3.67	3	4	96.0
	SVM-CV	69	77.1	0.60	27.8	2	5	96.0
*cancer*3	MOBJ-AIC	5	3.18	1.01	3.26	6	2	95.4
	MOBJ-BIC	2	0.89	6	2.24	8	2	94.25
	MOBJ-MVE	10	3.97	0.92	3.34	5	2	96.0
	MOBJ-CV	26	29.9	0.68	13.9	4	2	**96.6**
	SVM-CV	92	11.3	0.92	9.50	5	2	96.0

4.5.4 *Abalone* data-set

The *abalone* data-set represents the benchmark problem of mid-to-large scale (4177 samples) available from the UCI [Asuncion and Newman, 2007] repository. The problem consists of prediction of the ages of abalones by 8 physical measurements (7 scalar and 1 categorial). In the data-set, the ages are represented with integers from 1 to 29 years. Therefore, the problem can be viewed as a regression task.

For the comparability of results, the settings of the benchmark were reproduced exactly from [Meyer, Leisch, and Hornik, 2003], by using the same 100 data-set partitions of the Abalone data-set[5]. Each partition consists of 90% of non-overlapping samples for training and 10% for test generated with 10-fold CV from 10 random permutations of the data. The settings of the MOBJ algorithm were $Q_{\max} = 6000$, $R_Q = 250$, and $R_\sigma = 50$, whereas the parameters $\sigma_{\min}$ and $\sigma_{\max}$ were automatically determined from the percentile distribution of distances, same way as described in the experiment 4.5.3. Each data partition contained approximately $M \approx 2900$ distinct patterns, whose percentile distribution of distances resulted in the following values of the range parameters: $\sigma_{\min} \approx 0.07$ and $\sigma_{\max} \approx 5$. The AIC, BIC, and 3-fold CV

[5]The data-set partitions available at `http://www.ci.tuwien.ac.at/~meyer/benchdata`

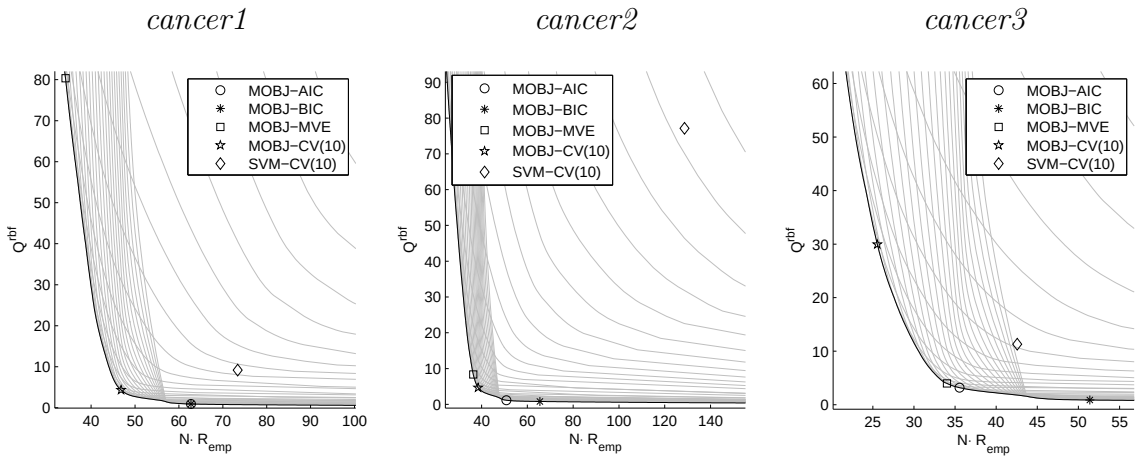

Figure 4.6: Experiment results for the Wisconsin breast cancer data-set.

model selection criteria were used with the MOBJ algorithm, in comparison with the results of the SVM with 3-fold CV, published in [Meyer, Leisch, and Hornik, 2003].

The median average values of solution properties, training, and test errors obtained with the MOBJ algorithm are given in Table 4.4 in comparison with the corresponding test error measurements of the SVM from [Meyer, Leisch, and Hornik, 2003]. As the measure of stability, the standard deviations of the test error over 100 cases are included.

Table 4.4: Abalone data-set results: median values of solution parameters and test RMSE

Method	Average properties				Train RMSE	Test RMSE	
	m_r	Q_{rbf}	σ	$\|\alpha\|_1$		median (mean)	std.
MOBJ-AIC	67	5811	0.33	641.71	4.22	**4.45** (**4.46**)	0.55
MOBJ-BIC	26	1433	1.16	1665.79	4.43	4.56 (4.58)	0.53
MOBJ-CV	63	5294	0.37	698.80	4.24	4.47 (4.48)	0.55
SVM-CV	–	–	–	–	–	4.48 (4.51)	–

The results in Table 4.4 show that MOBJ-AIC achieves better generalization performance. However, a high dispersion prevents the drawing of conclusions about significance levels. Hence, the performance of all presented methods can be treated as equivalent.

4.5.5 Discussion

The benchmark results of the developed MOBJ algorithm have confirmed in practice its theoretically expected properties such as good generalization properties of the Pareto-optimal hypotheses, existence of the non-convex Pareto fronts, and the sparsity of solutions.

The analysis of the randomized experiments (see Tables 4.2 and 4.4) as well as the non-randomized (Table 4.3) demonstrate different properties of the model selection criteria. The performance of validation criteria (MVE or CV) remains stable in most conditions, however the resulting solutions are usually complex (with respect to the complexity measure) and contain large number of basis functions. The information criteria AIC and BIC seen more conservative, their indicated solutions usually tend to have low complexity and small number of basis functions. The AIC and BIC solutions are often close to each other and sometimes coincide, especially on short data-sets. On a larger data-sets the BIC becomes more conservative and sometimes leads to underfitting, whereas the AIC demonstrates the increasing accuracy and competes with the validation techniques.

The ROFS algorithm in most cases has demonstrated significantly inferior performance and usually overfitting solutions. The SVM combined with CV shows a good and stable generalization performance. However, despite of the known sparsity of the large-margin classifiers, the SVM solutions required significantly larger numbers of the support vectors than the corresponding MOBJ solutions showing equivalent generalization performance. Consequently, the MOBJ-CV solutions are much sparser, even relying on the same model selection criterion as used with the SVM.

Even though MOBJ-CV, generally, requires less retrainings than SVM-CV (since the former performs CV only for nondominated subset of solutions), the application of CV remains computationally expensive with the MOBJ algorithm, especially on large data-sets. In contrast, the accuracy of the AIC increases with the length of the data-set whereas its application remains almost free of computational costs. Hence, the most critical part of the MOBJ algorithm (from the point of view of computational performance) is the procedure for calculation of regularization path of the LASSO regression.

As known, the LARS algorithm requires $\mathcal{O}(N^3)$ [Zou, Hastie, and Tibshirani, 2007] operations for obtaining of the complete regularization path, which is too high for applications to the large-scale problems, say, $N > 5000$. Noteworthy, the same complexity order is required for the single fit by least-squares or solution of the margin-maximization problem in the classical SVM algorithm, since both are the particular forms of the so-called Gaussian process. Unfortunately, the problem cannot be exactly solved in a time shorter than $\mathcal{O}(N^3)$. Thus, switching to approximate but faster solutions of the sparse Gaussian process, e.g., [Snelson and Ghahramani, 2007] might be an efficient adaptation of the MOBJ algorithm to the large-scale problems.

Indeed, in the context of the MOBJ algorithm, its computational complexity can also be controlled by discarding the unnecessary parts of regularization paths by choosing the proper value of the complexity limit $Q_{\max}$. Since Q_{rbf} is closely related to the number of non-zero coefficients, the value of $Q_{\max}$ also limits the maximum number of the basis functions (say, $m_r^{\max}$). In this case, if the LARS procedure stops after reaching maximum complexity, it takes $\mathcal{O}(m_r^{\max} N^2)$ to find the necessary part of the regularization path. Consequently, by the appropriate choice of $Q_{\max}$ the computational complexity the MOBJ algorithm reduces to $\mathcal{O}(N^2)$.

While the parameter R_σ directly determines how many regularization pathes will be calculated, the complexity resolution R_Q affects mainly the number of evaluations of the model selection criterion. In the case when AIC or BIC are employed, the choice of R_Q has almost no influence on the overall time of execution. For instance, computation of the solution with MOBJ-AIC algorithm took less than one minute for the *cancer1* data-set ($N = 350$) and about 15 minutes for the *abalone* data-set ($N = 3759$) [6].

In contrast with the reported in [Jin, Okabe, and Sendhoff, 2004; Gonzalez, Rojas, Ortega, H., Fernandez, and Diaz, 2003; Jin and Sendhoff, 2008] results of the evolutionary MOML algorithms, the Pareto front approximations of the MOBJ algorithm consist of smooth curves instead of populations of scattered points. Such an approach to the multi-objective problem overcomes the difficulties of the uniform distribution of solutions along the Pareto front, usually occurring in population-based

[6] The execution times corresponds to the pure-MATLAB single-threaded implementation of the MOBJ algorithm on a convectional desktop PC.

evolutionary algorithms. Also, it is possible to control the approximation quality directly by increasing the resolution R_σ. Moreover, with slight modifications of the proposed algorithm one can increase the approximation quality progressively in the neighborhood of the region of interest, or for the entire Pareto front.

4.6 Summary

The developed MOBJ algorithm performs a broad search within the hypothesis space of RBF networks and efficiently renders a wide spectrum of Pareto-optimal solutions. The MOBJ algorithm implements the idea of decomposition of a generally non-convex multi-objective problem into its convex parts, such that globally nondominated solutions can be found in a deterministic way, employing one of the computationally-efficient convex optimization procedures. In contrast to the common nondeterministic evolutionary MOML algorithms, the proposed MOBJ algorithm is capable of finding exact solutions of the multi-criteria problem within a guaranteed time.

The MOBJ algorithm, endowed with the information model selection criteria demonstrated the possibility of reaching a high generalization performance without necessity of the time- and data-consuming validation steps.

The parameters of the algorithm only determine the resolution and range of the search, not affecting the generalization properties of final solutions, if both are sufficiently large. In particular, overestimation of ranges or resolutions leads only to the computational overhead, which can be reduced using certain heuristic strategies. Also, with a certain form of implementation of the algorithm, the range and resolution may be progressively increased during the execution time, until reaching a desired precision of the Pareto set approximation.

A high stability and generalization performance has been demonstrated on the synthetic and real-world benchmarks. In comparison with other methods, such as the ROFS and SVM, the MOBJ algorithm has shown to be an efficient and competitive multi-objective tool for supervised learning.

Chapter 5

Multi-objective extension of margin maximization

In contrast to the smoothness, having a variety of formulations for arbitrary classes of functions, the concept of margin maximization belongs to the paradigm of separation of observations with a hyperplane. This in turn means a nonbreakable connection of the measure of margin with a particular feature space. Aiming for extension of the principle of margin maximization to learning machines with multiple feature spaces, this chapter addresses the problem of finding the corresponding complexity measure for the MOBJ algorithm. Starting with the argument that a direct extension of the measure of margin is not valid from the SRM point of view, the ideas of matching of feature spaces and extension of the measure of margin are developed. As the result, different complexity measures are proposed and their theoretical properties are extensively tested.

5.1 Introduction

As discussed earlier in Chapter 3, the MOBJ algorithm must be endowed with the complexity measure, such that nested hypothesis subsets are ordered by their learning capacity. Otherwise, as follows from the SRM, the generalization performance of a learning machine may be poor.

When the hypothesis space of a learning machine is the single RKHS $\mathcal{H}_k$ associated with a particular kernel k, the choice of the complexity measure on $\mathcal{H}_k$ as its squared norm $Q[f] = \|f\|_k^2$ simultaneously satisfies several principles of learning (see 3.2.1 for details). Namely, the known geometrical margin interpretation of $\|f\|_k$ shows the equivalence of the margin maximization principle to regularization and SRM (see 2.5.1).

Dealing with the extended hypothesis space $\mathcal{H}_K$, induced by a family K of multiple kernels, the conventional kernel selection techniques lead to hyperparameter estimation procedures, discussed in 2.5.4. As claimed in 2.6, such procedures do not implement the principle of SRM on the complete hypothesis space $\mathcal{H}_K$, but only on its single-kernel subsets $\mathcal{H}_k$, $k \in K$. From the SRM point of view, such a solution scheme is redundant. Hence, its implementation at the level of complete hypothesis space corresponds to a smaller region of efficient decisions, which implies a reduction of uncertainty.

The possibility of such uncertainty reduction and simultaneous implementation of both the SRM and margin maximization principles at the level of the complete multi-kernel hypothesis space $\mathcal{H}_K$ has been demonstrated by the suggested MOBJ procedure (3.7). In order to put this learning machine into practice, one has to specify a valid complexity measure on $\mathcal{H}_K$, which matches the idea of margin maximization. However, as claimed in 3.2.2, the validity of the traditional margin inverse (or prior) $Q[f] = \|f\|_k^2$ in place of the complexity measure on $\mathcal{H}_K$ is questionable.

5.1.1 Why $\|f\|_k^2$ is not a valid complexity measure on arbitrary hypothesis space?

It was shown in 3.2.2 that there exist infinitely many equivalent hypotheses of different RKHS norms, namely, for any hypothesis $f \in \mathcal{H}_k$ there exists such hypothesis $f' \in \mathcal{H}_{k'}$, that $f'(x) = f(x)$, for all $x \in \mathcal{X}$ and

$$\|f'\|_{k'}^2 = c\|f\|_k^2, \tag{5.1}$$

where the kernels k' and k satisfy $c \cdot k'(x, x') = k(x, x')$, $c > 0$. Nevertheless, the learning capacities of f and f' are equal, since they equivalently represent the same

function. Thus, a valid complexity measure Q must satisfy $Q[f'] = Q[f]$. Obviously, such the identity does not hold with $Q[f] = \|f\|_k^2$ due to (5.1). Moreover, if the class of kernels K contains all possible scalings of a particular kernel k, the capacity of any nested hypothesis subset $\Omega_i := \left\{ f \in \mathcal{H}_K : \|f\|_k^2 < \epsilon_i \right\}$ is broadly unlimited, since any function in $\mathcal{H}_K$ can be represented with arbitrary small $\|f\|_k^2$. One can see that the principle of SRM is violated.

Now, assume that K does not include scaled kernels, i.e., its elements are linearly independent. This can be seen as the result of a certain normalization of K. For example, the family of Gaussian kernels

$$k(x, x') = (\sqrt{2\pi}\sigma)^{-n} \exp\left(-\frac{\|x - x'\|^2}{\sigma^2} \right)$$

is normalized in L_1, i.e., $\|k(x, \cdot)\|_1 = 1$. In this case, the hypothesis space $\mathcal{H}_K$ does not contain equivalent functions produced by different scalings of the same kernel (except null). Hence, the capacity of the subset Ω_i now can be considered as limited with $\|f\|_k^2 < \epsilon_i$. Therefore, the results of margin maximization (i.e., minimization of $\|\cdot\|_k^2$) on $\mathcal{H}_K$ depend on the normalization of K. Moreover, it can be shown that a certain normalization of K may *a priori* predetermine the subset of dominated hypotheses $\mathcal{H}_K$, regardless of the training set. In order to show this, let us consider the subset of hypotheses

$$S_{\hat{Y}} := \left\{ f \in \mathcal{H}_K \mid (f(x_1), f(x_2), \ldots, f(x_N))^T = \hat{Y} \right\},$$

whose elements produce the same response vector $\hat{Y}$ for the given training set[1] Z_{tr}.

Obviously, all elements of $S_{\hat{Y}}$ are also associated with the same value of empirical risk R_{emp}, since the latter is a function of Z_{tr} and $\hat{Y}$. Hence, if the vector $\hat{Y}$ is such that $S_{\hat{Y}}$ contains the minimizer f of R_{emp} within the subset Ω_i, then f belongs to the set $\mathcal{P}(\mathcal{H}_K, (R_{\text{emp}}, \|\cdot\|_k^2))$ of Pareto-optimal hypotheses of the corresponding MOBJ problem and is a candidate solution to the learning problem. Consequently, f is also the minimizer of $\|\cdot\|_k^2$ on $S_{\hat{Y}}$ (otherwise, f is dominated).

Since all kernels in K are linearly independent, the vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)^T$ of expansion coefficients of any element of $S_{\hat{Y}}$ is uniquely determined by the associated

[1] Without loss of generality, it is assumed hereafter that all N input observations in Z_{tr} are distinct

kernel from the equation

$$G_k \alpha = \hat{Y}, \tag{5.2}$$

where G_k is the Gram matrix associated with k on the training set Z_{tr}. Solving equation (5.2) one can show that

$$\|f\|_k^2 = \hat{Y}^T G_k^{-1} \hat{Y}$$

for all $f \in S_{\hat{Y}}$, from where it follows that $\|f\|_k^2 / \|\hat{Y}\|^2$ is bounded with the eigenvalues of the inverse of Gram matrix G_k^{-1}. Consequently, the following bounds

$$\frac{1}{\lambda_k^{\max}} \|\hat{Y}\|^2 \leq \|f\|_k^2 \leq \frac{1}{\lambda_k^{\min}} \|\hat{Y}\|^2 \tag{5.3}$$

are tight, holding for all $f \in S_{\hat{Y}}$, where $\lambda_k^{\max}$ and $\lambda_k^{\min}$ are the largest and smallest eigenvalues of G_k, respectively.

The analysis of the bounds (5.3) leads to a conclusion that the hypothesis $f' \in S_{\hat{Y}}$, associated with the eigenvalue $\lambda_{k'}^{\max}$, is dominated if there exists another hypothesis $f \in S_{\hat{Y}}$, whose associated eigenvalues $\lambda_k^{\min}$ are larger than $\lambda_{k'}^{\max}$. In other words, there is a subset of kernels whose associated hypotheses are always dominated, independently to their generalization properties. Such kind of "masking" of the hypothesis space depends on normalization of K, allowing one to scale the kernel (along with the eigenvalues of Gram matrix) in arbitrary manner. While the set of functions represented by $\mathcal{H}_K$ remains to be invariant to normalization, their generalization properties remain constant as well. That means, again, that $Q[f] = \|f\|_k^2$ is not a valid complexity measure on $\mathcal{H}_K$.

For example, if the family of Gaussian kernels is normalized in L_1, it can be shown that the hypothesis associated with the kernel of infinitely small width σ (that is Dirac's delta function) will dominate over all other hypotheses since all eigenvalues of its Gram matrix are infinitely large with $\sigma \to 0$. In practice, it will result in the almost diagonal Gram matrix and the corresponding solution will be characterized by strong overfitting and no generalization.

Consequently, the complexity measure based on the RKHS norm $\|\cdot\|_k$ may lead to the degeneracy of Pareto set and poor generalization, when applied to the learning problems on the space of hypotheses $\mathcal{H}_K$ induced with arbitrary K. Therefore, aiming

to derive the margin-based complexity measure one has to consider a special treatment of the idea of margin maximization, needing to be extended on the hypothesis space of multiple RKHSs.

5.2 Feature normalization

The drawback of using the RKHS norm in the complexity measure on arbitrary $\mathcal{H}_K$ can be viewed as the consequence of different topologies of feature spaces, associated with the elements of K. Since RKHSs are different, their norms are incompatible. Consequently, it is naturally to consider a certain technique for their matching, for instance, by developing a scale-invariant metric.

Let us assume that feature spaces match if the lengths of feature vectors corresponding to the same observations are equal. Such an assumption directly leads to the metric, invariant to scaling due to normalization of features. In particular, given the hypothesis $f(x) = \sum_i \alpha_i k(x, x_i)$, $f \neq 0$, let us denote its complexity via the squared RKHS norm of another equivalent hypothesis $f'(x) = \sum_{i=1}^{N} \alpha'_i k'(x, x_i)$, associated with the scaled kernel $k'(x, x') = c^{-1}k(x, x')$ and coefficients $\alpha'_i = c\alpha_i$, such that the feature vectors associated with k' are unit. Assuming that all feature vectors induced with k have equal lengths (e.g., k is Gaussian), the scaling constant

$$c = \|\widetilde{x}\| \cdot \|\widetilde{x}'\| = \sqrt{k(x,x)k(x',x')} = k(x,x)$$

(recall the shorthand notation of $\Phi(x)_k$ is $\widetilde{x}$) satisfies the above conditions and in combination with (5.1) leads to the sought complexity measure

$$Q[f] = \|f'\|^2_{k'} = k(x,x)\|f\|^2_k. \tag{5.4}$$

The extension of (5.4) to the general case of kernels, whose feature lengths are not equal (e.g., polynomial kernels), requires finding the equivalent hypothesis f' with such α' and k', that

$$\sqrt{k'(x,x)} = \|\Phi_{k'}(x)\| = 1$$

and

$$f'(x) = f(x), \quad \text{for all } x \in \mathcal{X}. \tag{5.5}$$

Since the set of functions, spanned with $\mathcal{H}_k$, is unique to a given kernel, the spans of $\mathcal{H}_k$ and $\mathcal{H}'_k$ must coincide in order to satisfy (5.5). Hence, there must exist a linear operator $D : \mathcal{H}_k \to \mathcal{H}_{k'}$ satisfying

$$\|D\Phi_k(x)\| = 1$$

for all $x \in \mathcal{X}$. Obviously, such a linear operator D may not exist for arbitrary non-linear feature map $\Phi_k(x)$. Consequently, the disagreement with the equivalence condition (5.5) must be allowed in the general case, but certain properties of f must be preserved in f'.

5.2.1 Effective support vectors

Let the hypothesis f' be given as the sum

$$\widetilde{f}' = \sum_i \nu_i \tag{5.6}$$

of vectors ν_i, referred to as *effective support vectors*. Similarly, denote μ_i to be *effective feature vector*, associated with i-th observation. Next, associate f' to f by assuming that both ν_i and μ_i are collinear to $\widetilde{x}_i$ and, thus, $f' \in \mathcal{H}_k$.

Instead of the equivalence condition (5.5), impose the restriction on the effective vectors such that

$$\langle \nu_i, \mu_i \rangle = \alpha_i \langle \widetilde{x}_i, \widetilde{x}_i \rangle \tag{5.7}$$

holds for all $i = 1, \ldots, N$ and denote

$$Q[f] = \|\widetilde{f}'\|^2 = \sum_i \sum_j \langle \nu_i, \nu_j \rangle \tag{5.8}$$

to be the complexity measure.

Such definition of f' does not uniquely determine $Q[f]$ in terms of f, but allows a certain freedom in the choice of the lengths of ν_i and μ_i. For instance, the choice $\mu_i = \widetilde{x}_i$ in (5.7) (effective feature vectors equal to real feature vectors) yields $\nu_i = \alpha_i \widetilde{x}_i$. In this case, the original hypothesis f is recovered from (5.6), while (5.8) corresponds to its squared RKHS norm, i.e., $\widetilde{f} = \widetilde{f}'$ and $Q[f] = \|f\|_k^2$. However, different choices of μ_i will result in other definitions of the complexity measure.

5.2.2 Normalized complexity measure

Aiming to normalize feature spaces, consider the effective feature vectors

$$\mu_i = \frac{\widetilde{x}_i}{\|\widetilde{x}_i\|},$$

of unit lengths. Accordingly, the expression

$$\nu_i = \alpha_i \|\widetilde{x}_i\| \cdot \widetilde{x}_i$$

directly follows from the condition (5.7) and in combination with the definition (5.8) yields the *normalized complexity measure*

$$\begin{aligned} Q_{\text{norm}}[f] :=& \sum_i \sum_j \alpha_i \alpha_j \|\widetilde{x}_i\| \cdot \|\widetilde{x}_j\| \, \langle \widetilde{x}_i, \widetilde{x}_j \rangle \\ =& \sum_i \sum_j \alpha_i \alpha_j \sqrt{k(x_i, x_i) k(x_j, x_j)} k(x_i, x_j). \end{aligned} \tag{5.9}$$

Introducing the kernel

$$k_\nu(x_i, x_j) = \sqrt{k(x_i, x_i) k(x_j, x_j)} k(x_i, x_j),$$

the complexity measure (5.9) can be rewritten in the compact form

$$Q_{\text{norm}}[f] = \alpha^T G_{k_\nu} \alpha, \tag{5.10}$$

similar to $\|f\|_k^2$, where G_{k_ν} is the Gram matrix associated with k_ν, calculated for the training input observations.

It is easy to see that the complexity measure (5.4) is a particular case of Q_{norm}, when all features associated with k have equal lengths. As before, the equivalence condition (5.5) holds in this case. In case of arbitrary kernel k, the equivalence condition (5.5) may not hold, however it is straightforward to show that Q_{norm}, as well as its particular case (5.4), remains invariant to equivalent hypotheses produced by scaling a kernel.

Note that the usage of Q_{norm} in practice does not necessary involve additional calculations. For instance, if the hypothesis set $\mathcal{H}_K$ is induced with the family of

Gaussian kernels $k(x, x') = \exp(-\gamma\|x - x'\|^2)$, whose associated feature vectors are already unit for any choice of $\gamma > 0$, the normalized complexity measure reduces to the squared RKHS norm $Q_{\text{norm}}[f] = \|f\|_k^2$. This is the particular case of the hypothesis space $\mathcal{H}_K$, when the RKHS norm is, in fact, a valid complexity measure, from the point of view of feature normalization.

5.2.3 Radius/margin interpretation

The classical definition of the geometrical margin of a hypothesis f associated with k is

$$\varrho[f] = \frac{2}{\|f\|_k},$$

which is the distance between the hyperplanes $f(x) = -1$ and $f(x) = 1$ in the RKHS of k. Again, assume the particular case of k, whose associated feature vectors lie on a sphere of radius $R = \|\widetilde{x}\| = \sqrt{k(x, x)}$. Aiming at the scale-invariant measure of margin, one can consider the relation of the margin to the radius of the sphere via the expression

$$\frac{\varrho[f]}{R} = \frac{2}{\sqrt{k(x, x)}\|f\|_k} = 2Q_{\text{norm}}[f]^{\frac{1}{2}}. \tag{5.11}$$

On the one hand, the squared inverse of the relation (5.11) is the particular case of normalized complexity measure (5.9). On the other hand, the relation (5.11) resembles the well-known radius/margin generalization bound, proposed in [Vapnik and Chapelle, 2000] for hard-margin SVM, where R must be the radius of the smallest sphere, enclosing the feature vectors associated with the training obvervations.

5.3 Feature equalization

The approach of feature normalization implemented in Q_{norm} (5.9) matches different feature spaces regardless of distributions of the feature vectors. The same is known as the drawback of the radius/margin bound. In fact, generalization properties of hypotheses depend on the orientation of feature vectors, as well as on their lengths. For example, a class of functions spanned in the RKHS, whose feature vectors uniformly fill the sphere of radius R, represent higher learning capacity than a class of

functions associated with RKHS, where feature vectors are concentrated only in a small region of the same sphere. Such a conclusion becomes obvious after observing that the latter feature space is associated with smoother kernel due to smaller angles between feature vectors.

Therefore, in order to include the information of the angular topology of a feature space into the complexity measure, the following matching technique is proposed: map all hypotheses from $\mathcal{H}_K$ into a certain *reference space* of features, maintaining the equivalence between hypotheses by their dot product representation. The equivalence of the mapped hypotheses to their originals is treated in terms of (5.5). Since the reference space remains common for all elements of $\mathcal{H}^K$, their feature spaces already match within its context.

Following this idea, a learning machine with multiple feature spaces associated with K can be equivalently represented in the context of a single feature space. For the obvious reasons, such a matching approach can be called as feature equalization.

5.3.1 Reference and auxiliary maps

Consider the feature map Φ_k associated with the kernel $k \in K$ and the fixed *reference map* $\Phi^\circ : \mathcal{X} \to \mathcal{H}^\circ$, which is common for all kernels in the context of learning problem. Let there exist an *auxiliary map* $\Phi_k^* : \mathcal{X} \to \mathcal{H}^\circ$ associated with a particular $k \in K$, such that the identity

$$k(x, x') = \langle \Phi_k(x), \Phi_k(x') \rangle = \langle \Phi^\circ(x), \Phi_k^*(x') \rangle_{\mathcal{H}^\circ} \tag{5.12}$$

holds for all $(x, x') \in \mathcal{X}^2$. Introducing the shorthand images $\mathring{x}$ and $\overset{*}{x}$ of the observation x under the reference Φ° and auxiliary Φ_k^* maps, respectively, the identity (5.12) can be rewritten in a compact form

$$k(x, x') = \langle \widetilde{x}, \widetilde{x}' \rangle = \langle \mathring{x}, \overset{*}{x}' \rangle_{\mathcal{H}^\circ} . \tag{5.13}$$

Assuming the existence of Φ_k^* satisfying (5.12) for all $k \in K$, all hypotheses $f \in \mathcal{H}_K$ can be rewritten in terms of the dot product in the reference space $\mathcal{H}^\circ$. In particular,

one can rewrite the traditional dot product form of the hypothesis function

$$f(x) = \left\langle \widetilde{x}, \widetilde{f} \right\rangle \tag{5.14}$$

in terms of the dot product in $\mathcal{H}^\circ$ as follows:

$$\begin{aligned} f(x) &= \left\langle \widetilde{x}, \sum_i \alpha_i \widetilde{x}_i \right\rangle = \\ &= \left\langle \overset{\circ}{x}, \sum_i \alpha_i \overset{*}{x}_i \right\rangle_{\mathcal{H}^\circ}, \end{aligned}$$

or

$$f(x) = \left\langle \overset{\circ}{x}, \overset{*}{f} \right\rangle_{\mathcal{H}^\circ}, \tag{5.15}$$

after introducing the auxiliary image

$$\overset{*}{f} = \sum_i \alpha_i \overset{*}{x}_i \tag{5.16}$$

of the hypothesis f in $\mathcal{H}^\circ$.

Since both the Φ° and Φ_k^* are maps into a dot product space, there exist reference and auxiliary kernels

$$k^\circ(x, x') = \langle \Phi^\circ(x), \Phi^\circ(x') \rangle$$

and

$$k_k^*(x, x') = \langle \Phi_k^*(x), \Phi_k^*(x') \rangle$$

associated with them, respectively. Hereafter, the explicit specifications of the dot product domain $\mathcal{H}^\circ$ is omitted for the sake of shortness.

Therefore, $\overset{*}{f}$ is the equivalent representation of f in the reference space $\mathcal{H}^\circ$. Note, that the form (5.15) is similar to (5.14) and the image $\overset{*}{f}$ is the expansion of the auxiliary vectors with the same coefficients α_i as the original hypothesis $\widetilde{f}$. However, the evaluation of $f(x)$ in $\mathcal{H}^\circ$ is achieved with the dot product of the images $\overset{\circ}{x}$ and $\overset{*}{f}$, associated with different feature maps.

As an independent branch of the current research, theoretical framework for the derivation of auxiliary maps and their associated kernels is developed in Appendix A. In particular, the existence of auxiliary maps and compact forms of their associated

kernels is demonstrated in practice for the classes of Gaussian RBF and polynomial kernels.

5.3.2 A closer look at the reference space

Unlike the regular feature map of the kernel k, the reference map Φ° is fixed and responsible only for mapping input observations. Hence, the properties of the hypothesis f and its associated feature space are represented with $\overset{*}{f}$. For instance, one can already show that $\|\overset{*}{f}\|^2$ (the squared length of auxiliary image of f) is invariant to the scaling of k. For instance, consider again the case of two equivalent hypotheses produced by scaling: $\widetilde{f} = \sum_i \alpha_i \Phi_k(x_i)$ and $\widetilde{f}' = \sum_i \alpha'_i \Phi'_k(x_i)$, such that $k'(x, x') = c^{-1}k(x, x')$ and $\alpha'_i = c \cdot \alpha_i$, $c > 0$. Then, the identity

$$k'(x, x') = \langle \Phi^\circ(x), \Phi^*_{k'}(x) \rangle = c^{-1} k(x, x')$$

yields $\Phi^*_{k'}(x) = c^{-1}\Phi^*_k(x)$ and, thus, the identity

$$\|\overset{*}{f}'\|^2 = \sum_i \sum_j c^2 \alpha_i \alpha_j \left\langle c^{-1}\Phi^*_k(x_i), c^{-1}\Phi^*_k(x_j) \right\rangle \quad = \|\overset{*}{f}\|^2$$

holds. Therefore, the auxiliary images of equivalent hypotheses coincide in the reference space.

In the conventional feature space representation of the hypothesis f, the support vectors [2] and the feature vectors of training observations are the images under the same feature map $\Phi_k(x)$. Thus, the support vectors and, consequently, their linear expansion $\widetilde{f}$ belong to the span $S := \mathrm{span}\{\widetilde{x}_i\}_{i=1}^N$ of N images of distinct input observations x_i, $i = 1, \ldots, N$ from the training set. In contrast, within the reference space representation of f, the images $\overset{*}{x}_i$ of the corresponding input observations under the axillary map play the role of support vectors, which do not belong to the reference span $S^\circ := \mathrm{span}\{\overset{\circ}{x}_i\}_{i=1}^N$. This conclusion can be seen as an extension of the previous analysis in section 5.2 implying that there is no linear isomorphism between $\Phi_k(x)$

[2] In literature on SVM, support vectors are commonly defined as the subset of feature vectors laying on the margin and determining the optimal margin hyperplane of the SVM. In context of the current work, the term of support vectors means arbitrary feature vectors, corresponding to nonzero expansion coefficients of a hypothesis.

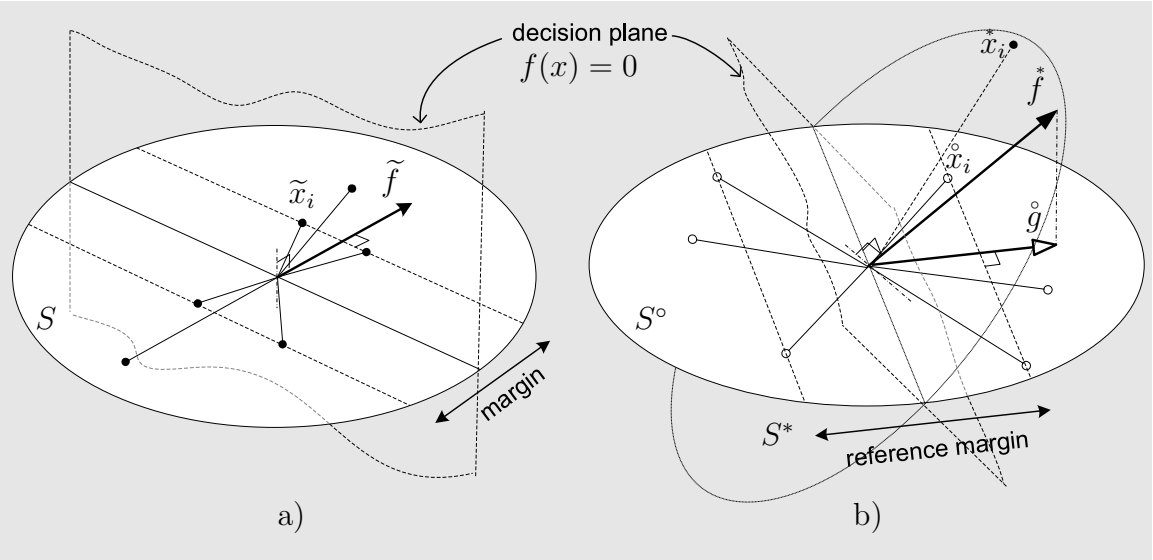

Figure 5.1: Schematic representation of the hypothesis f in the conventional (a) and reference (b) feature spaces.

and $\Phi_k^*(x)$. Otherwise, the spans of the corresponding RKHSs of k°, k and k^* must coincide simultaneously for all $k \in K$, which is generally not true.

The representation of the hypothesis f in the conventional and reference spaces is demonstrated schematically in Fig. 5.1. The auxiliary image $\overset{*}{f}$ is the normal vector of the decision hyperplane $\left\langle \overset{\circ}{x}, \overset{*}{f} \right\rangle = 0$ in the reference space, similarly to f in the RKHS. Hence, the hyperplanes

$$\begin{cases} \left\langle \overset{\circ}{x}, \overset{*}{f} \right\rangle = 1 \\ \left\langle \overset{\circ}{x}, \overset{*}{f} \right\rangle = -1 \end{cases} \tag{5.17}$$

also determine the margin associated with $\overset{*}{f}$. However, $\overset{*}{f}$ does not lie in the span S° of the training data, and the geometrical margin of its separation with $\overset{*}{f}$ (denoted as the *reference margin* in Fig. 5.1) differs from $2/\|\overset{*}{f}\|$.

Taking a closer look at the layout of vectors in the reference space, one can conclude that the reference margin (the geometrical margin, with which the training set is separated in the span of itself) is the distance between the intersections of the margin hyperplanes (5.17) with the span S°. Since the margin hyperplanes are parallel to the decision hyperplane with the normal vector $\overset{*}{f}$, it is straightforward to show that the distance between the intersections of (5.17) with S° is related to the orthogonal

projection $\mathring{g}$ of $\overset{*}{f}$ into S°. We shall refer $\mathring{g}$ as the reference image of f, which separates the training set with the reference margin $2/\|\mathring{g}\|$.

Since $\mathring{g} \in S^\circ$, $\mathring{g}$ admits the expansion

$$\mathring{g} = \sum_i \beta_i \mathring{x}_i. \tag{5.18}$$

Consequently, there exists a function g in the RKHS $\mathcal{H}_{k^\circ}$ of the reference kernel k°, such that

$$g(x) = \langle \mathring{x}, \mathring{g} \rangle = \sum_i \beta_i k^\circ(x, x_i).$$

In other words, the projection of $\overset{*}{f}$ into the reference span S° is the hypothesis associated with kernel k° and vector $\beta = (\beta_1, \beta_2, \ldots, \beta_N)^T$ of expansion coefficients.

In view of properties of the orthogonal projection, the following identity holds:

$$\left\langle u, \overset{*}{f} \right\rangle = \langle u, \mathring{g} \rangle, \quad \text{for all } u \in \operatorname{span}\{\mathring{x}_i\},$$

from where it immediately follows that

$$g(x_i) = f(x_i), \quad i = 1, \ldots, N. \tag{5.19}$$

The vector of expansion coefficients β can be found directly by solving system of linear equations (5.19) with respect to β_i, $i = 1, \ldots, N$. However, let us also show that $\mathring{g}$, being the projection of $\overset{*}{f}$, does not depend on the auxiliary map.

By definition of the orthogonal projection, $\mathring{g}$ is the minimizer of

$$\begin{aligned} \|\overset{*}{f} - \mathring{g}\|^2 &= \|\overset{*}{f}\|^2 - 2\left\langle \overset{*}{f}, \mathring{g} \right\rangle + \|\mathring{g}\|^2 \\ &= \|\overset{*}{f}\|^2 - 2\sum_{i=1}^N \sum_{j=1}^N \alpha_i \beta_j \langle \overset{*}{x}_i, \mathring{x}_j \rangle + \sum_{i=1}^N \sum_{j=1}^N \beta_i \beta_j \langle \mathring{x}_i, \mathring{x}_j \rangle, \end{aligned} \tag{5.20}$$

or

$$\begin{aligned} \|\overset{*}{f} - \mathring{g}\|^2 &= \|\overset{*}{f}\|^2 - 2\sum_{i=1}^N \sum_{j=1}^N \alpha_i \beta_j k(x_i, x_j) + \sum_{i=1}^N \sum_{j=1}^N \beta_i \beta_j k^\circ(x_i, x_j) \\ &= \|\overset{*}{f}\|^2 - 2\alpha^T G_k \beta + \beta^T G_{k^\circ} \beta, \end{aligned}$$

after substitutions $\langle \overset{*}{x}_i, \overset{\circ}{x}_j \rangle = k(x_i, x_j)$ and $\langle \overset{\circ}{x}_i, \overset{\circ}{x}_j \rangle = k^{\circ}(x_i, x_j)$, where G_k and G_{k° are the Gram matrices associated with kernels k and k°, respectively. Therefore, the solution of

$$\frac{\partial \|\overset{*}{f} - \overset{\circ}{g}\|^2}{\partial \beta} = 0,$$

results in the vector

$$\beta = (G_{k^\circ})^{-1} G_k \alpha \tag{5.21}$$

of expansion coefficients minimizing (5.20). As seen from (5.21), β and consequently g are independent on k^* and its associated auxiliary map Φ_k^*.

5.3.3 The concept of margin in a reference space: an extension is needed

The identity (5.19) means that the hypothesis g provides the same response to the training set as f , thus $R_{\text{emp}}[g] = R_{\text{emp}}[f]$. This in turn leads to the important conclusion that all hypotheses, whose auxiliary images are projected into the same reference image $\overset{\circ}{g}$, are equivalent with respect to the empirical risk. Therefore, specification of the complexity measure on the basis of the reference margin (that is $Q[f] = \|g\|_{k^\circ}^2$) leads to a degeneracy of the hypothesis space $\mathcal{H}_K$, since all hypotheses represented by a certain g become indistinguishable with respect to both objective functions, the empirical risk and complexity, simultaneously. Again, such a learning machine becomes inconsistent from the SRM point view.

It now can be shown that the above declined choice of the complexity measure $Q[f] = \|\overset{*}{f}\|^2$ based on the auxiliary image of f also causes a degeneracy of $\mathcal{H}_K$: the minimizer of $Q[f] = \|\overset{*}{f}\|^2$ within the subset of hypotheses, corresponding to a certain level of empirical risk (with a certain projection $\widetilde{g}$) is $\widetilde{g}$, since the latter is the smallest orthogonal projection into itself.

Finally, one may conclude that both margin-based approaches to the complexity measure ($Q[f] = \|\overset{*}{f}\|^2$ and $Q[f] = \|g\|_{k^\circ}^2$) in the reference space reduce representability of the hypothesis space $\mathcal{H}_K$ provoking degeneracies and, therefore, are unacceptable. Consequently, the concept of geometrical margin must be certainly extended with another more general property of a separation hyperplane.

5.4 Stability of separation hyperplanes

The idea of margin maximization is closely related to other concepts. For instance, the large margin classification is known to be robust to the observation noise and perturbation of hypothesis parameters (see e.g., [Scholkopf and Smola, 2001], ch. 7.2). Indeed, when the training set is correctly classified and separated at margin ϱ, all unseen patterns distributed within the $\varrho/2$-cover of the training set will be classified correctly as well. It means that the larger width of margin ϱ guarantees a certain classification accuracy at higher noise levels. Another robustness interpretation concerns with the stability of classification results to small perturbations of expansion coefficients. In this case, if the feature vectors are bounded in length and separated with a larger width of margin, the classification results do not change with small angular disturbances of the hyperplane's normal. Informally, this interpretation is also related to the principle of minimum description length (MDL) [Wallace and Boulton, 1968]: the larger the margin, the lower precision is needed to encode a separation hyperplane without suffering the classification results.

Both the stability interpretations denote the margin as a distance between separable feature vectors[3] to the separation hyperplane as a function of its orientation. Within the framework of feature equalization, the separation hyperplane (and its orientation) associated with the hypothesis f is denoted with the reference image $\mathring{g}$, as the projection of the auxiliary image $\overset{*}{f}$ of f into the span S° of reference features. As highlighted in the previous section 5.3.2, the hypotheses associated with the same reference image are indistinguishable from the point of view of reference margin. However, since the auxiliary images $\overset{*}{f}$ consist of different auxiliary support vectors, the same disturbance of their expansion parameters results in different disturbances of their projections into S°. Therefore, according to the MDL interpretation, the complexities of such hypotheses can be distinguished by amounts of information needed to encode their expansion parameters, maintaining the same precision of the separation hyperplanes in S°.

Intuitively, one can conclude that hypotheses associated with smooth kernels (and therefore weakly angled support vectors) require less information to be encoded with

[3]Similar conclusions apply to the soft-margin classification, where a certain number of misclassified training samples is allowed. For further details refer, e.g., [Cortes and Vapnik, 1995].

a given precision, than hypotheses, whose associated support vectors are almost orthogonal. This conclusion becomes clear, after viewing a separation hyperplane in the principal subspace of its associated support vectors: when support vectors are weakly angled, less principal components are needed to describe the linear system with a given precision.

5.4.1 Leave-one-out stability criterion

Let the separation hyperplane $\langle x, f\rangle = 0$ be given by the expansion

$$f := \sum_{j=1}^{N} \alpha_j x_j \tag{5.22}$$

of the N support vectors x_j, $j = 1, \ldots, N$. For simplicity of notation, it is assumed that x_j and f are already the vectors in some Hilbert space, whose specification is irrelevant in current context.

Now, let us consider the disturbance of the hyperplane vector f approximated by the linear combination of the reduced system of $N - 1$ support vectors, where the i-th support vector is excluded. Let the vector

$$g^{(i)} := \sum_{j=1, j\neq i}^{N} \beta_j^{(i)} x_j$$

stand for the normal of disturbed hyperplane, minimizing the squared error

$$e_i^2 = \|f - g^{(i)}\|^2. \tag{5.23}$$

Since an exclusion of the support vector means a certain information reduction, the resulting approximation error e_i reflects the consequent precision loss. Then, the sum

$$E(f) = \sum_{i=1}^{N} e_i^2 \tag{5.24}$$

can be a measure of information amount, needed for description of the hyperplane with certain precision. As a matter of fact, $E(f)$ can be interpreted as the leave-one-

out stability of f with respect to its support vectors. Note the leave-one-out error $E(f)$ is related to the hyperplane and is irrelevant to the training error.

Obviously, $g^{(i)}$ is the orthogonal projection of f into the span of support vectors without the i-th one. Introducing the reduced matrix

$$X^{(i)} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N)$$

of $N-1$ support vectors and the $(N-1) \times 1$ vector

$$\beta^{(i)} = (\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_N)^T$$

of the corresponding expansion coefficients of the projection

$$g^{(i)} = X^{(i)}\beta^{(i)},$$

one can show that the minimizer of the i-th squared distance

$$e_i^2 = \|f - X^{(i)}\beta^{(i)}\|^2 = f^T f - 2f^T X^{(i)}\beta^{(i)} + \beta^{(i)^T} X^{(i)^T} X^{(i)}\beta^{(i)}$$

corresponds to the solution of

$$\frac{\partial e_i^2}{\partial \beta^{(i)}} = 0$$

with respect to $\beta^{(i)}$. Hence,

$$\beta^{(i)} = \left(X^{(i)^T} X^{(i)}\right)^{-1} X^{(i)^T} f$$

and

$$\begin{aligned} e_i^2 &= f^T f - f^T X^{(i)} \left(X^{(i)^T} X^{(i)}\right)^{-1} X^{(i)^T} f \\ &= f^T f - f^T P^{(i)} f, \end{aligned} \tag{5.25}$$

where

$$P^{(i)} := X^{(i)} \left(X^{(i)^T} X^{(i)}\right)^{-1} X^{(i)^T}$$

is the orthogonal projector into the span of $X^{(i)}$.

At this point, one can demonstrate the relation of the stability measure to the

geometrical margin by the combination of (5.24) with (5.25), that yields

$$E(f) = N\|f\|^2 - \sum_{i=1}^{N} f^T P^{(i)} f. \tag{5.26}$$

As seen, the criterion $E(f)$ grows with the squared inverse $\|f\|^2$ of the geometrical margin but decreases with the growing projection lengths $f^T P^{(i)} f$. In other words, the more mutual information is contained in the system of support vectors, the more stable becomes f.

Introducing the reduced $(N-1) \times 1$ vector

$$\alpha^{(i)} = (\alpha_1, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_N)^T$$

of the expansion coefficients associated with f, one can rewrite (5.22) as the sum

$$f = X^{(i)} \alpha^{(i)} + x_i \alpha_i \tag{5.27}$$

and show that

$$f^T f = {\alpha^{(i)}}^T {X^{(i)}}^T X^{(i)} \alpha^{(i)} + 2\alpha_i x_i^T X^{(i)} \alpha^{(i)} + \alpha_i^2 x_i^T x_i. \tag{5.28}$$

Also, using the property $P^{(i)} X^{(i)} = X^{(i)}$ of the orthogonal projector, it can be shown that

$$\begin{aligned} f^T P^{(i)} f &= {\alpha^{(i)}}^T {X^{(i)}}^T P^{(i)} X^{(i)} \alpha^{(i)} + 2\alpha_i x_i^T P^{(i)} X^{(i)} \alpha^{(i)} + \alpha_i^2 x_i P^{(i)} x_i \\ &= {\alpha^{(i)}}^T {X^{(i)}}^T X^{(i)} \alpha^{(i)} + 2\alpha_i x_i^T X^{(i)} \alpha^{(i)} + \alpha_i^2 x_i P^{(i)} x_i. \end{aligned} \tag{5.29}$$

The combination of (5.28) and (5.29) with (5.25) yields

$$e_i^2 = \alpha_i^2 (x_i^T x_i - x_i^T P^{(i)} x_i). \tag{5.30}$$

It is noteworthy that (5.30) splits into the product of two independent terms: α_i^2 and $x_i^T x_i - x_i^T P^{(i)} x_i$, where the former represents the size of the expansion coefficients responsible for the margin and the latter represents self-similarity of the support vectors.

Next, consider the following lemma.

Lemma 5.4.1 *(Diagonal elements of the Gram matrix inverse): Let the matrix* $X = (x_1, \ldots, x_N)$ *contain* N *column vectors and the matrix* $X^{(i)}$ *contain* $N-1$ *column vectors from* X*, except the* i*-th one. Then, if the matrix* ${X^{(i)}}^T X^{(i)}$ *is invertible, the* i*-th diagonal element of the inverse of the Gram matrix* $G = X^T X$ *is*

$$d_i = (x_i^T x_i - x_i^T P^{(i)} x_i)^{-1},$$

where $P^{(i)} = X^{(i)} \left({X^{(i)}}^T X^{(i)} \right)^{-1} {X^{(i)}}^T$*. See Appendix B.2 for proof.*

Finally, the lemma 5.4.1 gives rise to the compact and computationally-friendly form

$$E(f) = \sum_{i=1}^{N} \frac{\alpha_i^2}{d_i} \tag{5.31}$$

of the leave-one-out stability criterion (5.24), where $(d_1, d_2, \ldots, d_N) = \mathrm{diag}(G^{-1})$ and $G = X^T X$ is the Gram matrix associated with the support vectors $X = (x_1, x_2, \ldots, x_N)$. Note that for better numerical stability it is convenient to calculate the elements d_i^{-1} directly from the singular value decomposition of X.

5.4.2 Stability-based reference complexity measure

Given the feature space image $\widetilde{f}$ of the hypothesis f and its associated Gram matrix G_k, one can express its complexity with the leave-one-out stability $E(\widetilde{f})$ (5.31) of its separation hyperplane. Such measure is closely related to the RKHS norm $\|f\|_k^2$, as seen from (5.26), and thus can be viewed as the extension of the traditional concept of margin.

Likewise the norm $\|f\|_k^2$, the magnitude of $E(\widetilde{f})$ is related to the metric in particular feature space associated with f and, thus, feature spaces must be equalized to ensure the comparability of criterion E over hypotheses from different RKHSs.

Recall the results of 5.3.2, where the separation hyperplane of the hypothesis f was given in the reference feature space by the vector

$$\mathring{g} = \sum_i \beta_i \mathring{x}_i, \tag{5.32}$$

which is an orthogonal projection of the auxiliary image

$$\overset{*}{f} = \sum_i \alpha_i \overset{*}{x}_i$$

into the reference span S° of training observations. Accordingly, stability of the separation hyperplane associated with f can be expressed in a comparable manner as $E(\mathring{g})$. However, in order to express stability of the original hypothesis f and properties of its associated kernel k, the measure $E(\mathring{g})$ must be calculated for $\mathring{g}$ in the form different to (5.32). In particular, the coefficients α_i and the auxiliary support vectors $\overset{*}{x}_i$, projected into S°, should be considered as the elements of the expansion $\mathring{g}$, instead of the coefficients β_i and vectors $\mathring{x}_i$ given by (5.32).

Let the hypothesis f given with N support vectors, whose corresponding reference feature vectors are columns of the matrix $X^\circ = (\mathring{x}_1, \mathring{x}_2, \ldots, \mathring{x}_N)$. Then, using the result (5.21), the expansion (5.32) can be written in the matrix form

$$\mathring{g} = X^\circ \beta = U\alpha,$$

where

$$U = X^\circ (G_{k^\circ})^{-1} G_k.$$

It can be seen that columns of the matrix U are orthogonal projections of the vectors $\overset{*}{x}_i$ into S°, same as the projection of their combination $\overset{*}{f}$ is $\mathring{g}$. Also, recall that $G_{k^\circ} = X^{\circ T} X^\circ$. Hence, the Gram matrix associated with the columns of U is

$$U^T U = G_k (G_{k^\circ})^{-1} G_k.$$

Finally, using (5.31), the *reference complexity measure* can be denoted on the basis of $E(\mathring{g})$ as

$$Q_{\text{ref}}[f] := \sum_{i=1}^{N} \frac{\alpha_i^2}{d_i}, \tag{5.33}$$

where

$$(d_1, d_2, \ldots, d_N) = \operatorname{diag}\big(G_k^{-1} G_{k^\circ} G_k^{-1}\big).$$

Note that there are different ways of putting Q_{ref} into practice. The first one is to consider N as a number of support vectors associated only with non-zero coefficients.

The second way is to consider the complete Gram matrix associated with N (distinct) training input observations. Since the Gram matrix remains fixed for all $f \in \mathcal{H}_k$, the complexity measure Q_{ref} is strictly convex on $\mathcal{H}_k$ and can be viewed as the coefficient-based stabilizer

$$Q_{\text{ref}}[f] = \alpha^T D \alpha,$$

where D is a diagonal matrix with the elements $d_1^{-1}, d_2^{-1}, \ldots, d_N^{-1}$ on the main diagonal.

It can be shown that when the number of actual support vectors is close to that of training observations, or when the feature vectors of $\mathcal{H}_k$ and $\mathcal{H}_{k^\circ}$ are uniformly spaced, both the ways of calculation of Q_{ref} provide similar results. However, the second approach seems preferable from practical considerations: it is enough to compute D, once for each $k \in K$, while Q_{ref} is convex on $\mathcal{H}_k$.

5.4.3 On a practical choice of the reference kernel

A choice of the reference kernel determines the reference map, and, consequently, influences to the ordering of the hypothesis space via the complexity measure Q_{ref} (5.33).

Even though the complexity measure is computable for arbitrary choice of G_{k° that, in fact, does not require the existence of $\overset{*}{f}$, the proper choice of the reference kernel involves deeper analysis of the reference space paradigm proposed in 5.3.1.

The results in A.4 show that the choice of the Dirac delta function for the reference kernel is universal for arbitrary family of kernels, since it ensures the existence of the corresponding auxiliary map, though the reference feature vectors in this case are infinitely large. Hence, the projection $\mathring{g}$ of $\overset{*}{f}$ into S° is infinitely small and, therefore, the value $Q_{\text{ref}}[f]$ is meaningless. Nevertheless, the results of A.4 can be adopted to the case of finite projection $\overset{*}{f}$ into S°. In particular, the choice of the reference kernel to be Kronecker delta function

$$k^\circ(x_i, x_j) = \delta_{ji} = \begin{cases} 1, & i = j \\ 0, & \text{otherwise} \end{cases}$$

ensures the existence of $\mathring{g}$ and also simplifies (5.33) with identity Gram matrix $G_{k^\circ} = I$. In this case, Mercer's eigenvalues associated with k°, are infinitely small causing the eigenvalues of k^* to be infinitely large. Hence, formally, the auxiliary kernel k^*

and image $\overset{*}{f}$ are infinite, however the projection of the latter into S° remains finite providing existence for the complexity measure Q_{ref}.

5.5 Basic MOBJ implementation

Recall the general components of a MOBJ algorithm formulated in 3.3: the kernel family K; the functionals of empirical risk R_{emp}, complexity measure Q, and model selection criterion ζ on $\mathcal{H}_K$; and the procedure for finding Pareto-optimal elements $\mathcal{P}(\mathcal{H}_k, \phi)$, $\phi[f] = (R_{\text{emp}}[f], Q[f])$ on the RKHS of k. The latter performs convex optimization (assuming that both R_{emp} and Q are convex on $\mathcal{H}_k$), whose results serve for reconstruction of the Pareto set $\mathcal{P}(\mathcal{H}_K, \phi)$ of the whole problem by means of decomposition (3.10) (see 3.2.3 for details).

In particular, given the generic kernel algorithm

$$\text{KM}(Z_{\text{tr}}, k, \lambda) = \arg\min_{f \in \mathcal{H}_k} R_{\text{emp}}[f] + \lambda Q[f], \tag{5.34}$$

solving the corresponding regularization problem, one can find the elements of $\mathcal{P}(\mathcal{H}_k, \phi)$ for all $\lambda \geq 0$ and then reconstruct the complete Pareto set of the MOBJ problem using the following relation

$$\mathcal{P}(\mathcal{H}_K, \phi) = \mathcal{P}\left(\bigcup_{k \in K} \bigcup_{\lambda \in \mathbb{R}^+} \text{KM}(Z_{\text{tr}}, k, \lambda), \phi \right). \tag{5.35}$$

In practice, one aims for computationally efficient approximation of (5.35) with a finite set of solutions. This requires analysis of the particular multi-objective problem and development of special steps, similar to those in Chapter 4. Nevertheless, the reliability of the proposed multi-objective approach to margin maximization can be confirmed with the basic implementation of the MOBJ algorithm given below.

5.5.1 The MOBJ on a grid

In its basic form, the Pareto set (5.35) can be approximated with the set

$$\mathcal{P}\left(\bigcup_{\theta\in\Theta_{\text{grid}}} \text{KM}(Z_{\text{tr}},\theta),\phi\right) \tag{5.36}$$

of nondominated hypotheses, corresponding to the grid Θ_{grid} of kernel and regularization hyperparameters. Here, $\text{KM}(Z_{\text{tr}},\theta)$ is the alternative notation to (5.34), where the kernel and regularization parameters are specified by the element θ of the grid Θ_{grid}.

Combination of (5.36) with a general MOBJ procedure (3.7) provides the basic multi-objective grid search algorithm:

1. Given the training set Z_{tr} and the set Θ_{grid} of grid elements generate the set of hypothesis
$$F := \Big\{ f = \text{KM}(Z_{\text{tr}},\theta) \mid \theta \in \Theta_{\text{grid}} \Big\},$$
corresponding to the elements of the grid;

2. Determine the set $F_{\text{nd}} := \mathcal{P}(F,\phi)$ of nondominated hypothesis with respect to R_{emp} and specified complexity measure Q, (e.g., by means of a pairwise comparison of elements in F).

3. Find the final solution
$$f_{\text{mobj}} = \arg\min_{f\in F\text{nd}} \zeta[f]$$
with respect to the specified model selection criterion.

It is noteworthy that omitting the second step, the above algorithm reduces to the conventional grid search procedure

$$f_{\text{grid}} = \text{KM}(Z_{\text{tr}}, \arg\min_{\theta\in\Theta_{\text{grid}}} \zeta[\text{KM}(Z_{\text{tr}},\theta)]), \tag{5.37}$$

where the model selection criterion is evaluated for all elements of Θ_{grid}, instead of their subset associated with only nondominated hypotheses.

In order to complete the above implementation scheme, one has to specify the procedure $\mathrm{KM}(Z_{\mathrm{tr}}, \theta)$ for solution of the regularization problem for the given training set Z_{tr} with respect to empirical risk R_{emp}, complexity measure Q, and hyperparameters θ.

The both complexity measures Q_{norm} and Q_{ref} can be expressed in the form $\alpha^T Q_k \alpha$, where Q_k is a square matrix. Hence, choosing the squared error loss function one can build the procedure $\mathrm{KM}(Z_{\mathrm{tr}}, \theta)$ on the basis of the solution

$$\alpha = (H_k^T H_k + \lambda Q_k)^{-1} H_k^T Y$$

of the modified generalized regularization network (GRN), where H_k is the design matrix associated with k on the training set and Y is the training target vector (see 2.3.4 for details). However, this approach is not convenient due to the known drawbacks of GRNs. Also, since Q_{norm} and Q_{ref} are not common regularizers, the solutions achieved with the corresponding MOBJ algorithm will be difficult to compare with the existing techniques due to differences between underlying learning machines. Therefore, it is encouraging to adapt the proposed grid-based MOBJ scheme to a certain existing kernel algorithm.

5.5.2 Adaptation to SVM classifier

Recall the setting of the C-SVC classifier, given in 2.5.1 by the hinge loss function

$$l(x, y, f(x)) = \max(0, 1 - y \cdot f(x)). \tag{5.38}$$

The traditional C-SVC algorithm maximizes the geometrical margin of the separation hyperplane, minimizing the squared RKHS norm $\|f\|_k^2$ along with the penalty term $C \cdot R_{\mathrm{emp}}[f]$, where C is the regularization hyperparameter and $R_{\mathrm{emp}}[f]$ is calculated with respect to the loss function (5.38) for the training set Z_{tr}.

As highlighted in 5.2.2, for the particular case of the Gaussian RBF kernel $k(x, x') = \exp(-\gamma \|x - x'\|^2)$, the complexity measure Q_{norm} corresponds to the squared RKHS norm $\|f\|_k^2$. In this case, the results of the C-SVC algorithm are elements of the nondominated set $\mathcal{P}(\mathcal{H}_k, \phi)$, as supposedly made with $\mathrm{KM}(Z_{\mathrm{tr}}, \theta)$. However, in a

general case of kernel k, both the complexities $Q_{\text{norm}}[f] \neq \|f\|_k^2$ and $Q_{\text{ref}}[f] \neq \|f\|_k^2$ differ to $\|f\|_k^2$ minimized by the C-SVC. Hence, the results of the procedure $\text{KM}(Z_{\text{tr}}, \theta)$ based on the C-SVC are generally not Pareto-optimal.

However, the application of C-SVC in the MOBJ algorithm is possible under certain considerations. In particular, one can assume the hypothesis space of a learning machine to be limited by the subset of $\mathcal{H}_K$ corresponding to such hypotheses, whose hyperplanes are optimal in SV sense. In other words such a hypothesis space consists of all possible SVMs on a given kernel family K. Then, the nondominated set of C-SVC solutions can be considered Pareto-optimal.

Alternatively, viewing Q_{norm} and Q_{ref} as certain metrics on $\mathcal{H}_k$, one can consider them to be equivalent[4] to $\|\cdot\|_k^2$. Consequently, the nondominated set

$$\mathcal{P}\left(\bigcup_{k\in K} \mathcal{P}\big(\mathcal{H}_k, (R_{\text{emp}}, \|\cdot\|_k^2)\big), \phi\right)$$

can be viewed as an approximation to the sought $\mathcal{P}(\mathcal{H}_K, \phi)$.

Finally, both assumptions allow the standard implementation of C-SVC to be employed directly in place of $\text{KM}(Z_{\text{tr}}, \theta)$ of the grid-based MOBJ algorithm described in 5.5.1. Note that a similar approach provides adaptations to other kernel algorithms (e.g., ε-SVR for regression).

5.6 Experiment

The goal of current experimental study is to verify the reliability of supervised learning with the MOBJ algorithms, endowed with the proposed complexity measures Q_{norm} and Q_{ref}. In particular, the experiment is designed to confirm the central claim of current research, formulated in 2.6, that the exhaustive search within the space of hyperparameters is redundant and unnecessary, in contrast to the MOBJ algorithm, implementing the SRM principle.

The experiment plain consists of several classification benchmarks of the grid-based MOBJ algorithm 5.5, powered by the SVM classifier and endowed with the

[4]The term is used in the sense of equivalent norms, i.e. the norm $\|\cdot\|_a$ is equivalent to $\|\cdot\|_b$ on X if there exist positive constants c and d, such that $c\|x\|_b \leq \|x\|_a \leq d\|x\|_b$ for all $x \in X$.

complexity measures Q_{norm} and Q_{ref}, according to 5.5.2. The conclusions are expected to be drawn from the comparison between the results the MOBJ algorithm and the conventional grid search procedure (5.37) treated under exactly the same conditions.

5.6.1 Benchmark setup

Several classification benchmarks were selected for the tests. The data-sets, listed in Table 5.1 along with their short descriptions, are based on real-world data of diverse nature. Most of the data-sets are available from the UCI repository [Asuncion and Newman, 2007]. All data-sets have passed through the unified preprocessing steps: missing values were removed, binary or metric attributes (including the target) were normalized to the interval $[-1, 1]$, and categorical attributes were previously expanded onto corresponding binary vectors. The total number of samples (length) and the number of attributes after the preprocessing are shown in the corresponding columns of Table 5.1.

Table 5.1: List of the benchmark data-sets

Name	Alias	Length	Attributes
Iris (cl. 3 vs. 1&2)	*iris3*	150	4
Wine (cl. 2 vs. 1&3)	*wine2*	178	13
Sonar	*sonar*	208	60
Heart disease[1]	*heart*	270	13
Liver disorders (BUPA)	*liver*	345	6
Ionosphere	*ionosphere*	351	34
Vehicle silhouettes (cl. 1 vs. 2)	*vehicle12*	417	18
Vehicle silhouettes (cl. 3 vs. 4)	*vehicle34*	429	18
Credit approval	*credit*	653	43
Wisconsin breast cancer	*cancer*	699	9
Inidian diabetes	*pima*	768	8

The multi-class data-sets were reduced to binary classification by a combination of several classes together (the data-sets *iris3* and *wine2*), or by separation of the observations (the data-sets *vehicle12* and *vehicle34*). In particular, in *iris3* and *wine2* the most overlapping classes were selected to be classified against the others.

[1]The *heart* data-set were taken from the StatLog Project folder at UCI.

The vehicle silhouette data-set was split in two independent classification problems: *vehicle12* (Opel vs. Saab cars) and *vehicle34* (vans vs. buses).

Aiming for representability of scores, the benchmark technique of randomized cross-validation sampling was used from [Meyer, Leisch, and Hornik, 2003]. Similarly to the benchmark described in section 4.5.4, 10 random permutations were generated from each data-set. Then, each permutation was split into 10 non-overlapping training/test pairs (9/10 for training and 1/10 for test) and 100 different training/test cases were generated from each data-set. Accordingly, the scores of the particular algorithm for a given data-set were calculated on the basis of the test classification error rates, after training 100 times on different training sets.

Likewise [Meyer, Leisch, and Hornik, 2003], four types of scores were considered in the experiment for evaluation of the performance and stability of algorithms using two kinds of average/dispersion measures: mean/standard deviation (std.), median/interquartile range (iqr.). The classification error rates (%), averaged over 100 test cases, stand for performance score. The dispersion of mean classification error on each data-set permutation (cross-validation error) stands for the score of stability. Hence, the lower values of both scores indicate better results.

5.6.2 Configurations of the algorithms

The standard implementation of C-SVC was used from the LIBSVM [Chang and Lin, 2001] with two kernel classes: the Gaussian RBF

$$k(x, x') = \exp(-\gamma \|x - x'\|^2)$$

of variable bandwidth γ and the polynomial

$$k(x, x') = (\langle x, x \rangle + 1)^p$$

of integer[5] order p.

[5]The integer polynomial orders were used not to deal with complex-valued kernels, though, an extension of the proposed measures to this case is straightforward.

The two-dimensional rectangular grids were used to control the hyperparameters (C, γ) and (C', p) in configurations with Gaussian and polynomial kernels, respectively. The hyperparameter $C = C' \cdot 2^{-2p}$, inversely controlling the strength of regularization in the standard implementation of C-SVC, was substituted by C' for better fitting of the region of interest[6] into a rectangular grid in case of the polynomial kernel.

The values $p \in \{1, 2, \ldots, 10\}$ of the polynomial degree were used for all data-sets, whereas the values of the hyperparameters C, C', and γ were aligned on the exponential grids of the base 2 and multiplier step $2^{\frac{1}{2}}$. Due to the diverse nature and dimensions of the data-sets, the ranges of hyperparameters C, C', and γ were distinguished into several groups, as shown in Table 5.2. Such kind of division in groups were necessary to avoid computational burden and undesirable numerical issues of the SVM algorithm, which usually occur with the choices of hyperparameters outside a certain range.

Table 5.2: Selected ranges of hyperparameters

Parameter	Range ($\log_2$)	Data-sets
C'	[−6, 14]	*iris3*, *wine2*, *liver*
	[−10, 10]	*heart*, *iono*, *vehicle12*, *vehicle34*, *pima*
	[−12, 10]	*sonar*, *credit*, *cancer*
C	[−5, 25]	*iris3*, *wine2*, *sonar*, *heart*, *iono*
	[−8, 16]	*liver*, *vehicle12*, *vehicle34*, *credit*, *cancer*, *pima*
γ	[−28, 2]	*wine2*, *sonar*, *heart*, *iono*
	[−20, 5]	*iris*, *vehicle12*, *vehicle34*, *credit*, *cancer*, *pima*
	[−12, 8]	*liver*

The corresponding MOBJ algorithm was implemented in accordance with 5.5 for its two configurations: MOBJ-Q_{norm} and MOBJ-Q_{ref}, corresponding to the proposed complexity measures (5.9) and (5.33), respectively. In the experiment, the results of MOBJ-Q_{norm} and MOBJ-Q_{ref} are compared with the conventional grid search procedure (5.37) (GS) under exactly the same settings. The 5-fold CV model selection criterion was used in all configurations of the MOBJ algorithm, including the grid search GS.

[6]The region of hyperparameter space, corresponding to representative models generated with C-SVC. The hypotheses are mostly degenerated or strongly overfitted outside this region.

5.6.3 Benchmark results

Tables 5.3-5.4 show performance and stability scores achieved by MOBJ-Q_{norm}, MOBJ-Q_{ref}, and GS. The best scores are marked bold. The intensity plots of the magnitudes of complexity (Q_{norm} and Q_{ref}), training error (R_{emp}), and 5-fold CV criterion are available in Appendix C for one particular training case of each data-set, including the marked Pareto-optimal elements and final solutions.

As numerical results show, most differences between the scores are small. Thus, further statistical tests are necessary for making conclusions about the properties of the benchmarked algorithms.

5.6.4 Significance tests

Recent results on the methods of comparison of multiple algorithms on multiple data-sets are discussed in [Demšar, 2006]. In [Demšar, 2006], the correctness of application of the traditional variance-based statistical tests, such as Student's t-test, is argued for comparison of the classifiers' performances expressed by the cross-validation scores. Since in these sampling schemes the training/test cases are mutually dependent, the calculation of the correct number of degrees of freedom becomes the major issue. Moreover, there are no corresponding statistics to compare several classifiers on multiple data-sets. On the other hand, the results obtained from multiple data-sets are naturally independent. Hence, non-parametric rank tests can be applied to compare two or more algorithms. Such techniques, as demonstrated in [Demšar, 2006] for comparison of classifiers, turned to be state-of-the-art tools in machine learning community.

In particular, the well-known Friedman rank test [Friedman, 1937, 1940] can be applied to verify the significance of ranking of multiple algorithms. Given l algorithms, ranked on T data-sets with their average ranks R_j, $j = 1, \ldots, l$, one assumes the null-hypothesis, stating that all algorithms should be ranked equal. In this case, the statistic

$$F_F = \frac{(T-1)\chi_F^2}{T(l-1) - \chi_F^2}$$

Table 5.3: Scores of GS, MOBJ-Q_{norm} and MOBJ-Q_{ref} with the Gaussian RBF kernel on benchmark data-sets.

Data-set	Method	Performance		Stability	
		mean	median	std.	iqr.
iris3	GS	4.67	6.67	0.70	0.00
	MOBJ-Q_{norm}	4.73	6.67	**0.38**	0.00
	MOBJ-Q_{ref}	**4.00**	6.67	0.44	0.00
wine2	GS	**2.52**	**0.00**	0.76	0.56
	MOBJ-Q_{norm}	3.19	2.78	0.80	0.59
	MOBJ-Q_{ref}	2.96	**0.00**	**0.60**	**0.03**
sonar	GS	23.42	23.81	2.17	3.36
	MOBJ-Q_{norm}	**14.60**	**14.29**	**0.97**	**1.93**
	MOBJ-Q_{ref}	21.84	19.05	2.76	4.71
heart	GS	17.07	**14.82**	0.64	1.11
	MOBJ-Q_{norm}	18.41	18.52	0.87	1.11
	MOBJ-Q_{ref}	**17.04**	**14.82**	**0.60**	**0.74**
liver	GS	28.99	**28.57**	**0.72**	**0.62**
	MOBJ-Q_{norm}	35.48	35.29	1.52	2.34
	MOBJ-Q_{ref}	**28.90**	**28.57**	0.76	1.34
iono	GS	**5.25**	5.71	0.57	0.87
	MOBJ-Q_{norm}	**5.25**	5.71	**0.31**	**0.31**
	MOBJ-Q_{ref}	6.38	5.71	0.85	1.40
vehicle12	GS	25.10	**25.58**	**1.25**	2.08
	MOBJ-Q_{norm}	39.35	39.53	3.17	6.00
	MOBJ-Q_{ref}	**24.92**	**25.58**	1.35	**1.15**
vehicle34	GS	1.53	**0.00**	**0.20**	**0.24**
	MOBJ-Q_{norm}	1.63	2.38	0.25	0.47
	MOBJ-Q_{ref}	**1.49**	**0.00**	0.37	0.48
credit	GS	**13.63**	**13.85**	**0.01**	**0.01**
	MOBJ-Q_{norm}	13.95	14.50	0.34	0.32
	MOBJ-Q_{ref}	**13.63**	**13.85**	**0.01**	**0.01**
cancer	GS	3.62	3.59	**0.31**	0.57
	MOBJ-Q_{norm}	3.75	4.29	0.26	**0.28**
	MOBJ-Q_{ref}	**3.58**	**2.88**	0.32	0.43
pima	GS	23.40	23.38	0.71	0.78
	MOBJ-Q_{norm}	23.59	23.38	**0.35**	**0.41**
	MOBJ-Q_{ref}	**23.35**	23.38	0.69	0.78

Table 5.4: Scores of GS, MOBJ-Q_{norm} and MOBJ-Q_{ref} with the polynomial kernel on benchmark data-sets.

Data-set	Method	Performance		Stability	
		mean	median	std.	iqr.
iris3	GS	4.67	6.67	0.77	1.33
	MOBJ-Q_{norm}	4.40	6.67	0.84	**0.67**
	MOBJ-Q_{ref}	**3.80**	**0.00**	**0.45**	**0.67**
wine2	GS	**2.52**	**0.00**	**0.84**	**1.08**
	MOBJ-Q_{norm}	3.93	2.78	1.36	1.70
	MOBJ-Q_{ref}	**2.52**	**0.00**	**0.84**	**1.08**
sonar	GS	20.97	20.00	2.21	4.26
	MOBJ-Q_{norm}	21.09	20.00	1.50	2.40
	MOBJ-Q_{ref}	**18.35**	**19.05**	**1.37**	**2.00**
heart	GS	17.04	**14.82**	0.86	**0.74**
	MOBJ-Q_{norm}	23.89	22.22	1.90	2.59
	MOBJ-Q_{ref}	**16.56**	**14.82**	**0.55**	**0.74**
liver	GS	30.61	31.43	**1.11**	**0.92**
	MOBJ-Q_{norm}	31.86	31.89	1.77	2.67
	MOBJ-Q_{ref}	**29.92**	**29.41**	1.15	2.08
iono	GS	8.29	8.57	0.75	0.89
	MOBJ-Q_{norm}	12.79	11.43	0.81	1.17
	MOBJ-Q_{ref}	**7.84**	**7.02**	**0.43**	**0.83**
vehicle12	GS	27.14	27.91	1.10	1.43
	MOBJ-Q_{norm}	33.08	32.56	**1.00**	1.64
	MOBJ-Q_{ref}	**26.88**	**26.74**	1.24	**0.94**
vehicle34	GS	1.61	**0.00**	0.37	**0.47**
	MOBJ-Q_{norm}	3.26	2.44	0.81	1.20
	MOBJ-Q_{ref}	**1.39**	**0.00**	**0.29**	0.48
credit	GS	**13.66**	**13.85**	**0.10**	**0.01**
	MOBJ-Q_{norm}	26.68	26.15	1.46	1.97
	MOBJ-Q_{ref}	**13.66**	**13.85**	**0.10**	**0.01**
cancer	GS	**3.11**	2.86	**0.12**	**0.14**
	MOBJ-Q_{norm}	3.39	2.86	0.21	0.29
	MOBJ-Q_{ref}	3.36	2.86	**0.12**	0.15
pima	GS	23.05	23.38	**0.55**	0.62
	MOBJ-Q_{norm}	25.62	25.97	1.24	**0.39**
	MOBJ-Q_{ref}	**23.02**	**22.08**	0.64	0.67

is distributed according to the F-distribution with $l-1$ and $(l-1)(T-1)$ degrees of freedom, where

$$\chi_F^2 = \frac{12T}{l(l+1)} \left(\sum_j R_j^2 - \frac{l(l+1)^2}{4} \right).$$

In the current experiment $T = 11$ and $l = 3$, hence 2 and 20 are, respectively, the nominator and denominator degrees of freedom. According to the table of critical F_F distribution, the null-hypothesis is rejected at 95% confidence level (i.e., rank differences are significant) when $F_F > 3.493$.

The average ranks and their corresponding statistics χ_F^2 and F_F are shown in Tables 5.5 and 5.6 for all types of scores. The best average ranks are marked bold and the significant cases of $F_F > 3.493$ are underlined.

Table 5.5: Friedman test of the algorithms with the RBF kernel

Algorithm	Performance ranks		Stability ranks	
	mean	median	std.	iqr.
GS	1.909	1.818	**1.864**	2.045
MOBJ-Q_{norm}	2.682	2.545	2.000	2.000
MOBJ-Q_{ref}	**1.409**	**1.636**	2.136	**1.955**
χ_F^2	9.045	5.091	0.409	5.091
F_F	<u>6.982</u>	3.011	0.189	0.021

Table 5.6: Friedman test of the algorithms with the polynomial kernel

Algorithm	Performance ranks		Stability ranks	
	mean	median	std.	iqr.
GS	1.909	1.909	1.773	**1.727**
MOBJ-Q_{norm}	2.909	2.818	2.727	2.545
MOBJ-Q_{ref}	**1.182**	**1.273**	**1.500**	**1.727**
χ^2	16.545	13.273	9.136	13.273
F_F	<u>30.333</u>	<u>15.208</u>	<u>7.102</u>	2.872

As seen, the average ranks corresponding to the mean performance score are detected significantly different for both the kernel cases. Also, the ranks associated with median performance and std. stability scores are significantly different in the experiments with the polynomial kernel.

Next, as suggested in [Demšar, 2006], the one-against-one comparison of particular algorithms can be done by means of the post-hoc Nemenyi test, namely, if the average

ranks of two classifiers differ by at least the critical value

$$CT = q\sqrt{\frac{l(l+1)}{6T}} \approx 0.99991,$$

the difference is significant with 95% confidence, where the constant [7] $q = 2.343$ corresponds to $l = 3$. Consequently, the differences of the average ranks in Tables 5.5and 5.6, combined together, lead to the following of conclusions:

- The performance of the MOBJ-Q_{ref} is significantly higher than MOBJ-Q_{norm}, but not significantly higher than that of GS;
- With the Gaussian RBF kernel, both the MOBJ-Q_{ref} and MOBJ-Q_{norm} are equivalently stable;
- With the polynomial kernel, MOBJ-Q_{ref} is more stable than MOBJ-Q_{norm} and the performance of MOBJ-Q_{norm} is significantly lower than that of GS;

5.6.5 Discussion

The statistical tests of the experiment results have confirmed theoretical expectations about the capabilities of the benchmarked MOBJ algorithms endowed with the proposed complexity measures. In particular, there were no significant differences in classification accuracy and stability between the MOBJ algorithm endowed with Q_{ref} and the conventional grid search, whereas the search area of the former (only nondominated elements of the grid) is significantly smaller than the complete grid. This in turn confirms a redundancy of the traditional grid search procedures.

As shown by the visualizations in Appendinx C, the Pareto-optimal pathes of the MOBJ-Q_{ref} algorithm tend to cross the regions of small CV error, passing close to the grid search solution. As the consequence, the MOBJ-Q_{ref} solutions likely coincide with the GS solutions, especially when the CV error surface is regular in the neighborhood of its minima (see e.g., the bottom-right plots in Figures C.4,C.8,C.13,

[7] The table of constants for calculation of critical values of Nemenyi test is available in [Demšar, 2006].

and C.14). However, when CV surfaces are irregular (e.g., Figures C.21,C.9) or their minima are weak (e.g., Figures C.17, and C.15), the GS solutions become sensitive to permutations of the data-sets and may be biased. In contrast, the Pareto-optimal pathes are irrelevant to the values model selection criterion, suffering less from its biasedness due to a smaller area of search, limited only to nondominated hypotheses. This explains a slightly better performance of MOBJ-Q_{ref} in comparison to GS, which significance, unfortunately, could not be detected from the available amount of data-sets.

Recall that the complexity measure $Q_{\mathrm{norm}}[f]$ is a squared RKHS norm $\|f\|_k^2$ in case of the Gaussian RBF kernel. Since $\|f\|_k^2$ is minimized by the SVM algorithm, its solutions are Pareto-optimal with respect to the empirical risk and complexity measure Q_{norm} on $\mathcal{H}_k$. In contrast, the Pareto-optimality of SVM solutions on $\mathcal{H}_k$ does not hold for the complexity measure Q_{ref}, as discussed above in section 5.5.2. Despite of that, the results of MOBJ-Q_{norm} demonstrated lower efficiency than MOBJ-Q_{ref}, in most cases. Moreover, with the polynomial kernel, whose feature space topology significantly changes with p, MOBJ-Q_{norm} demonstrated even worse performance and lower stability in comparison to MOBJ-Q_{ref}. This fact whiteness the drawbacks of the feature normalization approach, as predicted in 5.3.

The basic MOBJ algorithm proposed in 5.5.1 was sufficient to demonstrate the reliability of the developed multi-objective approach. However, such kind of grid-based scheme is computationally inefficient for most practical applications, especially when dimensionalities of grids (number of hyperparameters) are large. In spite of that, the multi-objective grid search implemented by MOBJ-Q_{ref} and MOBJ-Q_{norm} requires less computational time by approximately the factor of T, in comparison to the traditional grid search with the T-fold CV, since the number of nondominated elements is small. In particular, the experiments show that the number of nondominated grid elements is $\mathcal{O}(a)$, while the number of grid elements is $\mathcal{O}(a^h)$, where h is the number of hyperparameters. Hence, the traditional grid search with the T-fold CV requires $\mathcal{O}(T \cdot a^h)$ runs of the kernel algorithm, whereas the MOBJ algorithm is expected to spend only $\mathcal{O}(a^h) + \mathcal{O}(T \cdot a) = \mathcal{O}(a^h)$, which is T times faster than $\mathcal{O}(T \cdot a^h)$.

5.7 Summary

Both the feature normalization and feature equalization techniques are aimed to match different feature spaces. Whereas the former lies in a correction feature vectors' lengths by scaling with a certain metric, the latter is related to the underlying concept of reference spaces. In general, the application of feature equalization requires derivation of auxiliary kernels, whose closed forms may not exist for arbitrary settings. However, a special study in Appendix A provides several examples of derivation and demonstrates the existence of auxiliary maps and kernels.

The analysis of the reference space paradigm (see 5.3.2) revealed that the complexity measure of hypotheses in the reference space can not be expressed within the traditional concept of geometrical margin and, thus, its further extension has been required. The extension was therefore proposed on the basis of formalization of the widely-known robustness interpretation of the margin maximization principle. Specifically, the introduction of the leave-one-out stability criterion of separation hyperplanes, in combination with the developed feature equalization technique, led to a development of the so-called reference complexity measure.

The application of the proposed reference complexity measure in comparison with the normalized complexity measure involves significantly more efforts at both the computation (requires computation of the diagonal elements of the Gram matrix inverse) and optimization (minimization the regularization functional with the non-standard stabilizing term) sides. Nevertheless, its theoretical advantages were experimentally confirmed by the superior generalization performance and stability of the corresponding MOBJ algorithm. Also, the MOBJ algorithm employing, the reference complexity measure confirmed the redundancy of the exhaustive grid search under the same conditions, theoretically predicted from the SRM point view. This in turn demonstrates the reliability of the proposed multi-objective extension of the concept of margin maximization.

Chapter 6

Conclusions

Current research contributes to MOML with a novel multi-objective approach to supervised learning, incorporating advantages of the traditional (single-objective) learning concepts and multi-objective optimization. The proposed framework is built on the SRM principle, viewed as a generally non-convex bi-objective problem and addressed by decomposition into its convex elements in a certain deterministic algorithm. In contrast to evolutionary algorithms, commonly employed in MOML, the proposed algorithm efficiently approximates the Pareto set providing arbitrary precision within a guaranteed time, taking full advantage of convex programming. Also, unlike the common hyperparameter selection procedures, the proposed solution scheme extends the known learning machines to larger hypothesis spaces (e.g., associated with multiple kernels) in the SRM-consistent manner, reduces uncertainty and thereby improves generalization performance. The complexity measure, as a key element of the study, is treated from several perspectives, each one of which led to an independent branch of results.

The concept of smoothness applied to the hypothesis space of Gaussian RBF networks led to a development of the efficient multi-objective learning algorithm for rendering a wide spectrum of Pareto-optimal hypothesis of diverse structures. Moreover, the information criteria adopted for selection of Pareto-optimal models in combination with the proposed algorithm show high generalization performance. Although the series of experiments have already demonstrated the algorithm as a self-contained and ready-to-use tool, its further extensions and improvements are still possible. Among

them is the adaptation of the proposed complexity measure to other classes of radial-basis functions, inducing new modifications of the proposed algorithm. Also, significant improvements of the computational performance on long data-sets may be achieved with the adaptation of the algorithm to approximate large-scale LASSO procedures.

The indirect extension of the concept of margin maximization to the multi-kernel context required development of a number of additional elements. Among them are the so-called matching techniques of feature space normalization and equalization and the leave-one-out stability criterion of separation hyperplanes, playing a role of the extended measure of geometrical margin. Being combined together, the technique of equalization and stability criterion result in a new complexity measure, which extensive experimental analysis has confirmed its reliability as well as the theoretical advantages of the proposed multi-objective framework in whole. The proposed complexity measure is ready to be applied for arbitrary hypothesis spaces induced with positive definite kernels and, within the scope of developed framework, opens new perspectives for the construction of multi-objective kernel machines. Specifically, in a similar scheme of the above multi-objective algorithm for RBF networks, a new efficient multi-objective algorithm for SV machines may be elaborated by means of construction of regularization pathes (e.g., using the recent finding [Ong, Shao, and Yang, 2010]), in a similar manner to that of the MOBJ algorithm for RBF networks.

A special attention deserves the theoretical concept underlying the technique of feature equalization. The provided study demonstrates the existence and possibility of construction of a single Hilbert space, isomorphic to the union of arbitrary RKHSs, while maintaining the ability of functions to be evaluated via the dot product. This finding allows one to view a wider class of nonlinear functions from the new perspective of linear spaces, making a firm basis for the analysis of multi-kernel models and algorithms in a generalized context.

The provided study offers new insights in the MOML field by advanced concepts, methods, and algorithms ready for practical applications and theoretical evolution towards more sophisticated learning machines.

Appendix A

Auxiliary kernels

Although reference complexity measure developed in (5.33) was shown to be irrelevant to the auxiliary map, the idea of equivalent representation of hypotheses in a common feature space itself is fruitful from the theocratical point of view and may find its further applications. Aiming to demonstrate the existence of such representation, a framework for derivation of auxiliary maps and their associated kernels is provided in this chapter, along with several examples of derivation.

A.1 Basic considerations

Recall the abstract definitions of the reference Φ° and auxiliary ${\Phi^*}_k$ maps given in 5.3.1. In order to show their existence and find their associated kernels in computation-friendly closed forms, one first has to specify a particular reference space $\mathcal{H}^\circ$.

Obviously, $\mathcal{H}^\circ$ must embed the RKHS $\mathcal{H}_{k^\circ}$ of k°. However, $\mathcal{H}^\circ$ can not be equal to $\mathcal{H}_{k^\circ}$. Otherwise, only the trivial case of $k^* = k^\circ = k$ would be supported with such a choice due to the known uniqueness of RKHSs. In fact, the isomorphic representations of feature spaces induced by k^* and k° are simultaneously possible in ℓ_2 via Mercer's theorem.

Recall the definition (2.33) of the Mercer's feature map

$$\Phi(x) := \left(\sqrt{\lambda_j}\psi_j(x)\right)_j,$$

associated with the kernel k, where $(\psi_j)_j$ is the orthonormal eigenbasis of the integral operator (2.32) and $(\lambda_j) \in \ell_1$ is the non-negative sequence of the corresponding eigenvalues (see 2.4.2 for details).

Let $\Phi : \mathcal{X} \to \ell_2$ and $\Phi^\circ : \mathcal{X} \to \ell_2$ be the Mercer's maps associated with k and k°, respectively. Assuming an existence of the common eigenbasis $(\psi_j)_j$ for both the kernels, the identity (5.12) can be rewritten as

$$\begin{aligned} k(x, x') &= \left\langle \left(\sqrt{\lambda_j}\psi_j(x)\right)_j, \left(\sqrt{\lambda_j}\psi_j(x')\right)_j \right\rangle \\ &= \left\langle \left(\sqrt{\lambda_j^\circ}\psi_j(x)\right)_j, \left(\sqrt{\lambda_j^*}\psi_j(x')\right)_j \right\rangle, \end{aligned} \tag{A.1}$$

where $(\lambda_j)_j$, $\left(\lambda_j^\circ\right)_j$, and $\left(\lambda_j^*\right)_j$ are the eigenvalue sequences of the kernels k, k°, and k^*, respectively. Then, it is straightforward to show that the auxiliary map

$$\Phi_k^*(x) = \left(\sqrt{\lambda_j^*}\psi_j(x)\right)_j,$$

satisfying (5.12) is given by the eigenvalue sequence

$$\left(\lambda_j^*\right)_j = \left(\frac{\lambda_j^2}{\lambda_j^\circ}\right)_j \in \ell_1. \tag{A.2}$$

Note that eigenvalue sequences are non-negative. Thus, assuming their decreasing order, one can immediately conclude that $\left(\lambda_j^\circ\right)_j$ must decay slower than $\left(\lambda_j^2\right)_j$ to ensure the existence of (A.2) in ℓ_1. In other words, the feature space induced by k° must be sufficiently "rich" to handle the feature space induced by k.

The equation (A.2) establishes the relation between k and its associated auxiliary kernel k^* for the particular reference kernel k° in terms of their eigenvalues. Hence, one is already able to calculate the auxiliary feature map $\Phi_k^*(x)$ directly from (A.2), though its evaluation is useless, until a closed form of the auxiliary kernel

$$k^*(x, x') = \langle \Phi_k^*(x), \Phi_k^*(x') \rangle$$

exists and found. Aiming to derive the closed form of k^*, it is proposed to seek for k^* within the particular family of kernels K by assuming that:

- K is homogenous, i.e., all elements of K can be viewed within the same eigenbasis;
- $k^{\circ} \in K$ and $k \in K$;
- all eigenvalues of any element of K can be analytically drawn.

A.2 Convolution kernels

Consider a positive definite kernel $k(x, x') = \kappa(x - x')$ such that $\kappa : L^2(\mathbb{R}^n \to \mathbb{R})$. Such a translationally invariant kernel is also known as a convolution kernel, since the integral operator (2.32) associated with k is the convolution operator

$$(T_k f)(\cdot) = (\kappa * f)(\cdot). \tag{A.3}$$

Following the idea of derivation of the closed form of auxiliary kernel, one has to find the eigenspectrum of (A.3) first.

As known, the operator T_k must be compact in order to have a countable set of eigenvalues (discrete spectrum). However, many commonly-used convolution kernels have unbounded support, resulting in non-compact T_k (e.g., Gaussian RBF) and therefore its continuous spectrum. Even though it has been recently shown [Sun, 2005] that, under certain assumptions, Mercer's theorem holds on non-compact domains, one is interested in discrete eigenspectrum of (A.3) for making use of (A.2). Fortunately, the technique of kernel approximation on a periodic domain developed in [Williamson, Smola, and Scholkopf, 2001] can be used for finding a discrete eigenspectrum of kernels with an unbounded support. In particular, the technique relies on the $\tau-$periodic extension

$$k_\tau(x, x') := \kappa_\tau(x - x') = \sum_{z \in \mathbb{Z}} \kappa(x - x' + \tau z) \tag{A.4}$$

of the convolution kernel k. Assuming the existence of k_τ for a given k, it can be shown that k_τ approximates k arbitrary well on the hyperbox $\mathcal{X}_\tau = [-\frac{\tau}{2}, \frac{\tau}{2}]^n$. In other

words, for any $\epsilon > 0$ there exist such sufficiently large τ, that

$$\sup_{(x,x')\in\mathcal{X}_\tau^2} |k(x,x') - k_\tau(x,x')| \leq \epsilon$$

holds.

According to [Williamson, Smola, and Scholkopf, 2001], the integral operator T_{k_τ} is compact on $\mathcal{X}_\tau$. Hence, its eigenspectrum is discrete and can be immediately found from the Fourier series expansion of the eigenvalue equation

$$(k_\tau * \psi_j)(x) = \lambda_j \psi_j(x)$$

, namely, introducing the Fourier transform of κ on $\mathbb{R}^n$ as

$$F[\kappa](\omega) := (2\pi)^{-\frac{n}{2}} \int_{\mathbb{R}^n} e^{-i\cdot\langle\omega,x\rangle} \kappa(x)\partial x,$$

one can show that the j-th eigenvalue

$$\lambda_j = (2\pi)^{\frac{n}{2}} F[\kappa]\left(\frac{2\pi}{\tau} j\right), \quad j \in \mathbb{Z}^n \tag{A.5}$$

corresponds to both the eigenfunctions

$$\psi_j^+(x) = \tau^{-n} \cos\left(\frac{2\pi}{\tau}\langle j, x\rangle\right)$$

and

$$\psi_j^-(x) = \tau^{-n} \sin\left(\frac{2\pi}{\tau}\langle j, x\rangle\right),$$

whose L_2 norms are unit on $\mathcal{X}_\tau$. However, since the common eigenbasis is assumed for all kernels, the eigenfunctions themselves are irrelevant.

Now, is all ready for derivation of the auxiliary kernel. Consider the family of convolution kernels

$$K_{\text{conv}} := \left\{ k_{\beta,\sigma} \middle| k_{\beta,\sigma}(x,x') = \beta\kappa\left(\frac{x-x'}{\sigma}\right), \quad \beta,\sigma \in \mathbb{R}^+ \right\},$$

induced by scalings of κ and its argument. Here, β determines a simple scaling factor applied to a kernel, whereas σ determines a scaling of the input domain or, in other words, is the width of a kernel.

Let $k_{\beta^\circ,\sigma^\circ} \in K_{\text{conv}}$ be the reference kernel, then the auxiliary kernel corresponding to the arbitrary kernel $k_{\beta,\sigma} \in K_{\text{conv}}$ is also expected in K_{conv}. In other words, there must exist such kernel parameters $\beta^* > 0$ and $\sigma^* > 0$ that the eigenvalues associated with the kernels $k^* = k_{\beta^*,\sigma^*}$, $k_{\beta,\sigma}$, and $k_{\beta^\circ,\sigma^\circ}$ satisfy (A.2).

Assuming an existence of the corresponding $\tau-$periodic extensions (A.4) for all $k_{\beta,\sigma} \in K$, it is possible to rewrite the eigenvalues (A.5) of elements of K_{conv} in terms of the kernel parameters β and σ, and the Fourier transform $F[\kappa]$ as

$$\lambda_j = (2\pi)^{\frac{n}{2}} \beta \sigma^n F[\kappa]\left(\frac{2\pi\sigma}{\tau} j\right), \quad j \in \mathbb{Z}^n. \tag{A.6}$$

Then, the combination of (A.6) with the eigenvalue equation (A.2) leads to the functional equation with respect to $F[\kappa]$:

$$\beta^* \sigma^{*n} F[\kappa](\sigma^* \omega) = \frac{\beta^2 \sigma^{2n} F^2[\kappa](\sigma\omega)}{\beta^\circ \sigma^{\circ n} F[\kappa](\sigma^\circ \omega)}. \tag{A.7}$$

Note that (A.7) is independent on τ, hence the equation (A.7) applies directly for the elements of K as $\tau \to \infty$[1].

Therefore, the auxiliary kernel k^* exists in K and can be found, if the Fourier transform of κ is a root of (A.7) with the parameters β^* and σ^*. Fortunately, one can immediately show that the exponential functions are roots of (A.7). For instance, let

$$\kappa(x) = \exp\left(-\frac{1}{2}\|x\|^2\right)$$

then K_{conv} be a family of conventional Gaussian RBF kernels. Combination of the Fourier transform

$$F[\kappa](\omega) = \exp\left(-\frac{1}{2}\|\omega\|^2\right)$$

[1] Also, it is possible to show that the same equation (A.7) could be found without τ-periodic approximation, if the continuous form of Mercer's theorem existed.

with (A.7) yields the solution

$$\beta^* = \frac{\beta^2}{\beta^\circ}\left(\frac{\sigma^2}{\sigma^\circ\sqrt{2\sigma^2-\sigma^\circ}}\right)^n$$

and

$$\sigma^* = \sqrt{2\sigma^2 - \sigma^{\circ 2}}.$$

As seen the auxiliary kernel exists when $\sigma > \frac{\sigma^\circ}{2}$ in case of the Gaussian RBF kernels. This fact is also supported by the intuitive conclusion that a "capacity" of the reference feature space associated with a sharp kernel (large bandwidth or small σ°) is sufficient to handle smoother kernels (narrow bandwidth or large σ).

Assume that $\beta = 1$ and consider a special case of the reference kernel with infinitesimal width $\sigma^\circ \to 0$ and $\beta^\circ = \left(\sigma^\circ\sqrt{2\pi}\right)^{-n}$. In this case, the reference kernel is a Dirac delta function

$$k^\circ(x, x') = \delta(x - x')$$

and the auxiliary kernel corresponding to the Gaussian kernel

$$k(x, x') = \exp\left(-\frac{\|x_i - x'_j\|^2}{2\sigma^2}\right)$$

is

$$k^*(x, x') = (\sigma\sqrt{\pi})^n \exp\left(-\frac{\|x_i - x'_j\|^2}{4\sigma^2}\right). \tag{A.8}$$

Astonishingly, the reference feature vectors associated with the such k° are orthogonal and lie along the axis in ℓ_2. Also, the kernel k° is absolutely sharp, representing the "richest" feature space in which the learning capacity of any hypothesis is infinitely large.

A.3 Polynomial kernels

The above technique directly involves Mercer's theorem for derivation of feature maps, however this is not a unique way of representation kernel's feature map by series (see e.g., [Minh, Niyogi, and Yao, 2006]). For example, feature maps associated with the polynomial kernels (which are not only non-compact, but are also unbounded) are

convenient to be represented via the binomial expansion instead of derivation of the associated eigenspectrum. Hence, the proposed approach of finding of auxiliary kernel needs in a special adaptation to cover such a case.

Let us consider the family of generalized polynomial kernels

$$K_{\text{poly}} := \left\{ k_\lambda : \mathcal{X}^2 \to \mathbb{R} \mid k_\lambda(x, x') = \sum_{j=0}^{p} \lambda_j \langle x, x' \rangle^j , \quad \right\}, \tag{A.9}$$

on $\mathcal{X} = \mathbb{R}^n$, where $\lambda = (\lambda_j)_{j=0}^p$ is the sequence non-negative coefficients, $p \in \mathbb{N}$. Indeed, the term $\langle x, x' \rangle^j$ is the polynomial kernel itself, whose features map consists of all j-th order monomials (see e.g., [Hofmann, Schöolkopf, and Smola, 2008], ch. 2.2.4) . Moreover, it can be shown that

$$\langle x, x' \rangle^j = \langle P_j(x), P_j(x') \rangle$$

where

$$P_j(x) = (\psi_{j,1}(x), \psi_{j,2}(x), \ldots, \psi_{j,n^j}(x))^T$$

is the $n^j \times 1$-vector of all j-th order products of components of x. Therefore, almost any $k_\lambda \neq 0$ is a positive definite kernel whose feature map can be written as the finite sequence

$$\Phi_\lambda(x) = \left(\left(\sqrt{\lambda_j} \psi_{j,i}(x) \right)_{i=1}^{n^j} \right)_{j=0}^{p} . \tag{A.10}$$

Unlike the general form (A.9), the closed form

$$k_{\gamma,c,p}(x, x') = (\gamma \langle x, x' \rangle + c)^p \tag{A.11}$$

of a polynomial kernel is common on practice, where $\gamma > 0$, $c \geq 0$, and $p \in \mathbb{N}$ are the kernel parameters. Using the binomial expansion, it can be shown that $k_{\gamma,c,p} \in K_{\text{poly}}$ is given by the coefficients

$$\lambda_j = \binom{p}{j} \gamma^j c^{p-j}, \; j = 0 \ldots p. \tag{A.12}$$

Even though Φ_λ is not a Mercer's map, the relation (A.2) still can be used with a slight abuse of notation. Specifically, given the reference kernel $k_{\lambda^\circ} \in K$ and the

arbitrary kernel $k_\lambda \in K$, elements of the coefficient sequence λ^* corresponding to the auxiliary kernel $k_{\lambda^*} \in K$ can be calculated using the algebraical form of (A.2). It is straightforward to show that the resulting sequence of coefficients λ^* satisfies the identity (A.1) (with the monomials $\psi_{j,i}$ in place of eigenfunctions) and therefore Φ_{λ^*} is the auxiliary map satisfying (5.12).

Hence, the combination of (A.12) with (A.2) yields the elements

$$\lambda^*_j = \frac{\binom{p}{j}^2 \gamma^{2j} c^{2p-2j}}{\lambda^\circ_j}$$

of the coefficient sequence of the auxiliary kernel k_{λ^*} corresponding to $k_{\gamma,c,p}$ with respect to the reference kernel k_{λ°. Unfortunately, it is due to binomial coefficients the auxiliary kernel k_{λ^*} does not admit the closed form (A.11), even when the reference k_{λ° is also given by (A.11). However, since p is finite and usually small, it is enough to assume the reference kernel k_{λ° with the sufficiently large ($p^\circ \geq p$) sequence of unit coefficients $\left(\lambda^\circ_j\right)_{j=0}^{p^\circ} = 1$, giving rise to the computable form of auxiliary kernel:

$$k_{\lambda^*}(x, x') = \sum_{j=0}^{p} \binom{p}{j}^2 \gamma^{2j} c^{2(p-j)} \langle x, x' \rangle^j . \tag{A.13}$$

Despite of the lack of elegance and requirement of numerical evaluations when compared to (A.8), the auxiliary kernel (A.13) demonstrates the possibility of an extension of the developed approach to different feature spaces expressed not via the Mercer's theorem.

A.4 Universal reference map

The idea of using a Dirac delta function as the choice of the reference kernel can be extended to arbitrary classes of Mercer's kernels. In particular, given the Mercer's feature map

$$\Phi(x) := \left(\sqrt{\lambda_j}\psi_j(x)\right)_j$$

with an arbitrary orthonormal eigenbasis $(\psi_j(x))_j$, it is straightforward to show that a reference kernel associated with the sequence of unit eigenvalues is

$$k^\circ(x,x') = \sum_j \psi_j(x)\psi_j(x') = \delta(x-x') = \begin{cases} \infty, x = x' \\ 0, \text{otherwise} \end{cases}$$

Next, assuming that eigenvalues of the reference map are unit, the relation (A.2) provides the auxiliary map

$$\Phi_k^*(x) := (\lambda_j \psi_j(x))_j \,,$$

whose associated kernel k^* exists, *iff* the sequence $\left(\lambda_j^2\right)_j$ of squared eigenvalues is in ℓ_1. Since $(\lambda_j)_j \in \ell_1$ holds by definition, the series $\sum_j \lambda_j^2$ converge and therefore $\left(\lambda_j^2\right)_j \in \ell_1$ holds as well.

Consequently, for an arbitrary Mercer's kernel k there exist a corresponding auxiliary kernel k^*, associated with the reference space of a Dirac delta function. Therefore, such a choice of reference space is universal.

A.5 Summary

The developed technique demonstrates an existence of auxiliary maps and possibility of their analysis and application via the kernels in closed forms. Further development of the approach may be evolve into new learning techniques for multiple kernels and the construction of corresponding learning algorithms within the concept of reference feature spaces.

Appendix B

Proofs of some lemmas

B.1 Matrix inversion lemma: particular case

Lemma B.1.1 *Let symmetric $(N \times N)$-matrix X given with the block form*

$$X = \begin{bmatrix} A & b \\ b^T & c \end{bmatrix},$$

where A is the $(N-1)\times(N-1)$ invertible matrix, b is the $(N-1)\times 1$ column vector and c is a scalar. Then the inverse X^{-1} exists and admits the form

$$X^{-1} = \begin{bmatrix} A^{-1} + \frac{1}{r}A^{-1}bb^TA^{-1} & -\frac{1}{r}A^{-1}b \\ -\frac{1}{r}b^TA^{-1} & \frac{1}{r} \end{bmatrix},$$

if the Schur complement $r = c - b^TA^{-1}b$ of the block A is non-zero.

Proof Let us write the lower-diagonal-upper factorization of X

$$X = \begin{bmatrix} A & b \\ b^T & c \end{bmatrix} = LDU = \begin{bmatrix} I & 0 \\ b^TA^{-1} & 1 \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & r \end{bmatrix} \begin{bmatrix} I & A^{-1}b \\ 0 & 1 \end{bmatrix},$$

from where the inverse of X can be found as

$$X^{-1} = U^{-1}D^{-1}L^{-1} = \begin{bmatrix} I & -A^{-1}b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} A^{-1} & 0 \\ 0 & \frac{1}{r} \end{bmatrix} \begin{bmatrix} I & 0 \\ -b^T A^{-1} & 1 \end{bmatrix}$$
$$= \begin{bmatrix} A^{-1} + \frac{1}{r}A^{-1}bb^T A^{-1} & -\frac{1}{r}A^{-1}b \\ -\frac{1}{r}b^T A^{-1} & \frac{1}{r} \end{bmatrix}. \quad \blacksquare$$

B.2 Proof of lemma 5.4.1 (Diagonal elements of the Gram matrix inverse)

Proof Assume that X is the $n \times N$-matrix, where $1 \leq n \leq \infty$ and introduce the right-hand circular column shift operator

$$S := \begin{bmatrix} 0 & 1 \\ I_{N-1} & 0 \end{bmatrix},$$

where I_{N-1} is the $(N-1) \times (N-1)$ identity matrix. It is straightforward to show $SS^T = S^T S = I$ and thus $S^{-1} = S^T$. One can verify, that the operator S shifts the columns of the operand matrix to the left as follows:

$$XS = (x_2, \ldots, x_N, x_1).$$

Consequently, one can show that

$$XS^i = (x_{i+1}, \ldots, x_N, x_1, x_2, \ldots, x_i) \tag{B.1}$$

where $S^i = S \cdots S$ is the i-times product and $S^N = S^0 = I$ is the identity. Also, the property $(S^i)^{-1} = S^{N-i}$ is straightforward to confirm.

Since the squared length $x_i^T P^{(i)} x_i$ of the orthogonal projection of x_i into the span of the rest $N-1$ is irrelevant to the column order of $X^{(i)}$, let us denote $X^{(i)}$ to be the corresponding block of XS^i. Then, one can rewrite (B.1) in the block form

$$XS^i = \begin{bmatrix} X^{(i)} & x_i \end{bmatrix}. \tag{B.2}$$

Now, consider the shifted Gram matrix

$$G' = (XS^i)^T XS^i = S^{N-i}X^T XS^i = S^{N-i}GS^i$$

and its inverse

$$(G')^{-1} = ((XS^i)^T XS^i)^{-1} = S^{N-i}(X^T X)^{-1}S^i = S^{N-i}G^{-1}S^i.$$

One can see that G' can be obtained by the circular diagonal shift of G (i.e., column-wise and row-wise) and the same conclusion applies to their inverses. Hence, one can show that

$$\begin{aligned}\operatorname{diag}\big((G')^{-1}\big) &= \operatorname{diag}\big(S^{N-i}G^{-1}S^i\big) \\ &= \operatorname{diag}\big(G^{-1}\big)S^i = (d_{i+1}, \ldots, d_N, d_1, d_2, \ldots, d_i).\end{aligned}$$

In other words, the i-th diagonal element d_i of G^{-1} is the last diagonal element of $(G')^{-1}$.

Next, using the particular case of the matrix inversion lemma B.1.1 one can show that the last diagonal element of $(G')^{-1}$ is the inverse of the Schur complement

$$r = x_i^T x_i - x_i^T X^{(i)}({X^{(i)}}^T X^{(i)})^{-1}{X^{(i)}}^T x_i = x_i^T x_i - x_i^T P^{(i)} x_i$$

of the the block form

$$G' = ((XS^i)^T XS^i) = \begin{bmatrix} {X^{(i)}}^T X^{(i)} & {X^{(i)}}^T x_i \\ x_i^T X^{(i)} & x_i^T x_i \end{bmatrix}.$$

Hence, $d_i = \frac{1}{r}$. ■

Appendix C

Visualizations of the experiment results from Chapter 5

In this appendix, the empirical risk, complexity measures, and model selection criterion are visualized for a single benchmark case of each data-set of the experiment described in section 5.6.

Each data-set is represented here by four intensity plots in Figures C.1–C.22, displaying the values of complexity measures Q_{norm} (top-left) and Q_{ref} (top-right), the values of the empirical risk R_{emp} (bottom-left), and the values of the 5-fold CV error (bottom-right) for the corresponding elements of hyperparameter grids. In addition, the grid elements corresponding to nondominated hypotheses generated by MOBJ-Q_{norm} and MOBJ-Q_{ref} are marked by crosses ($\times$) and dots ($\cdot$), respectively. The minima of the 5-fold CV model selection criterion within the nondominated subsets of MOBJ-Q_{norm} and MOBJ-Q_{ref} (which are their corresponding final solutions) are marked by squares ($\square$) and circles ($\circ$), while the global minima of the 5-fold CV error (finale solutions of the GS) are marked by "diamonds" ($\lozenge$). Thereby, the Pareto-optimal pathes of MOBJ-Q_{norm} and MOBJ-Q_{ref} are visualized in spaces of hyperparameters.

Note that the information from the model selection criterion (bottom-right plot) remains "invisible" to MOBJ-Q_{norm} and MOBJ-Q_{ref}, until the corresponding nondominated sets are found. For a detailed description of the experiment settings and analysis of the results, refer to section 5.6.

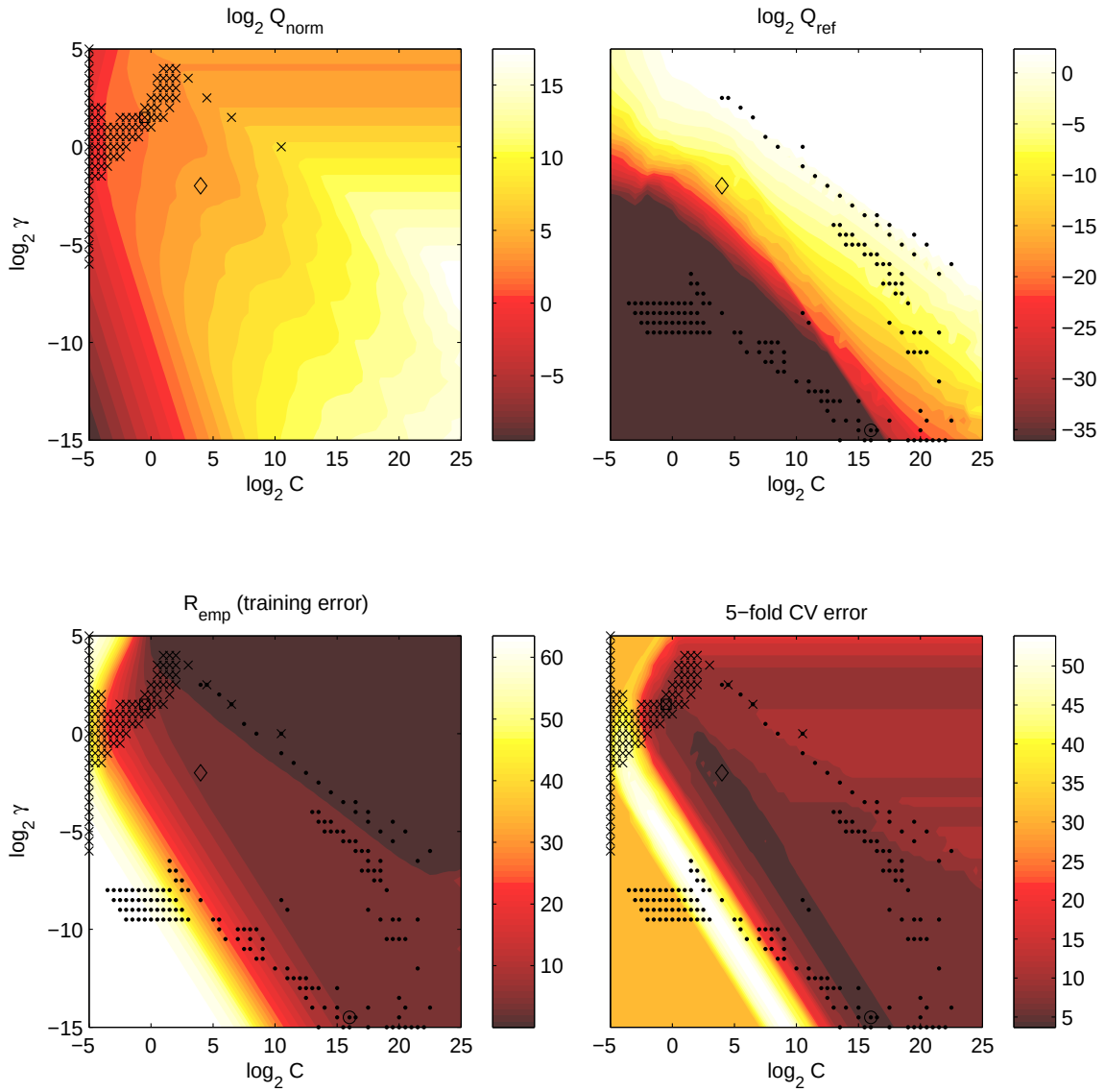

Figure C.1: Visualizations of the experiment results from Chapter 5 for *iris3* data-set and the Gaussian RBF kernel.

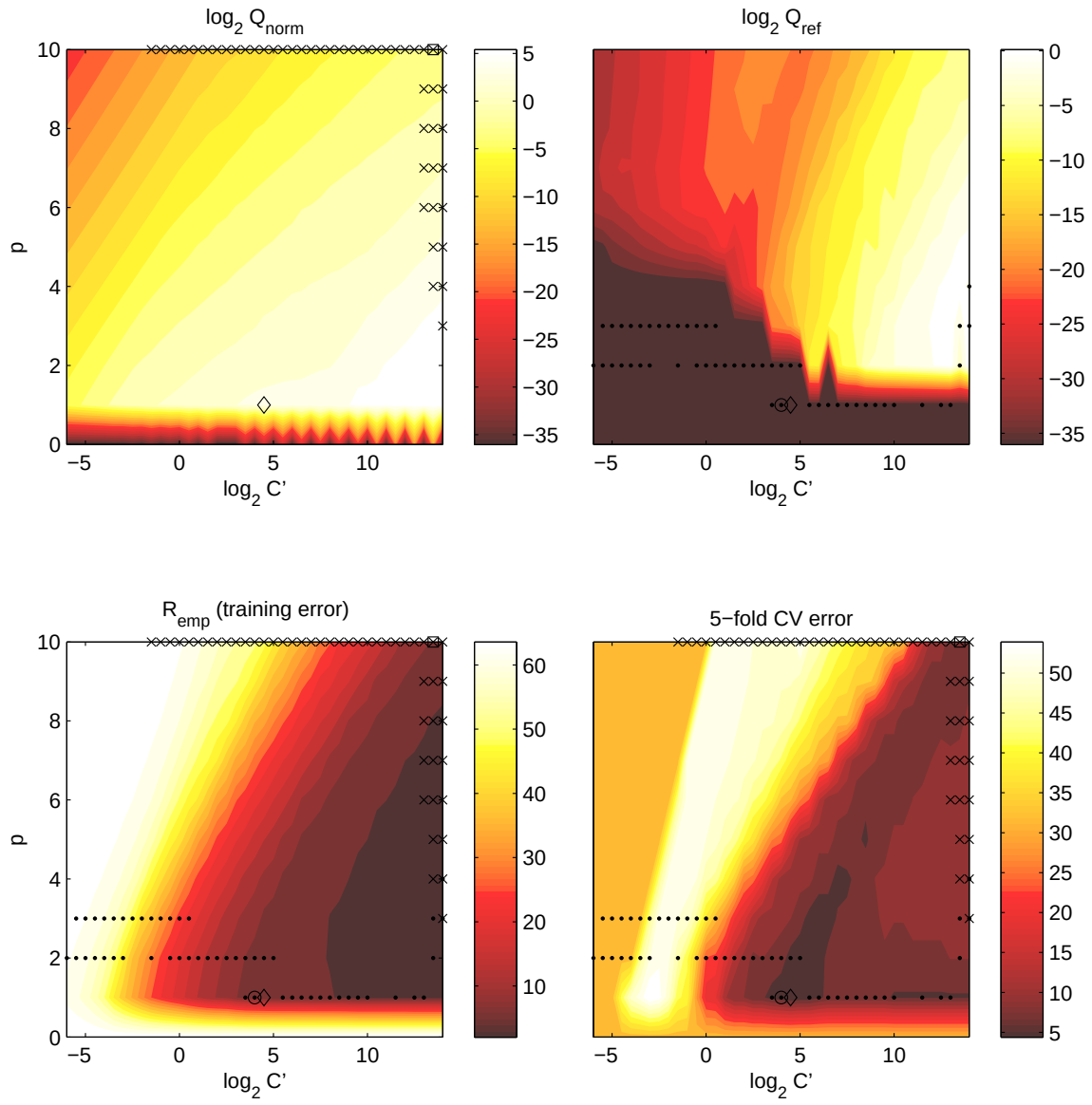

Figure C.2: Visualizations of the experiment results from Chapter 5 for *iris3* data-set and the polynomial kernel.

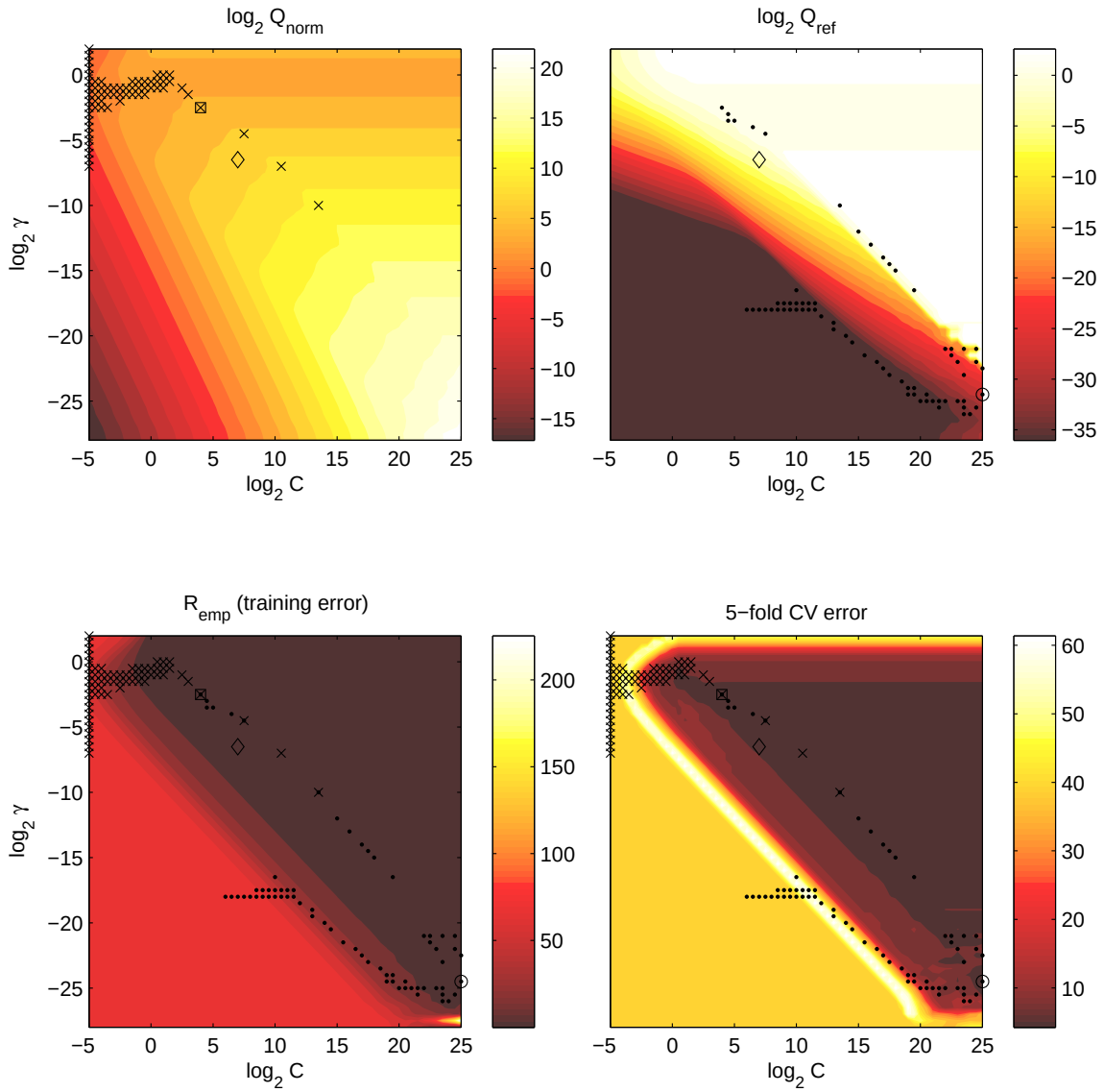

Figure C.3: Visualizations of the experiment results from Chapter 5 for *wine2* dataset and the Gaussian RBF kernel.

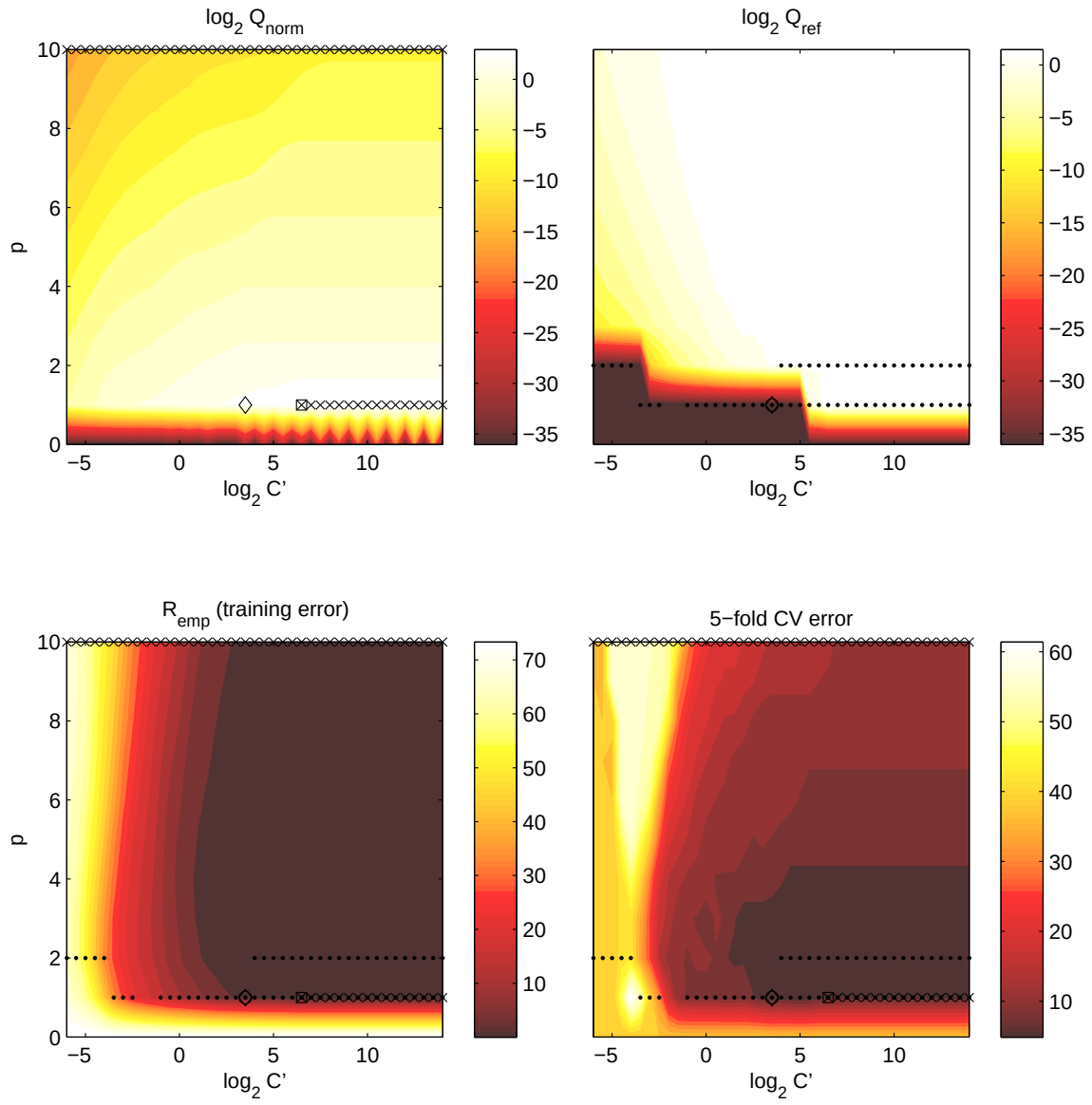

Figure C.4: Visualizations of the experiment results from Chapter 5 for *wine2* dataset and the polynomial kernel.

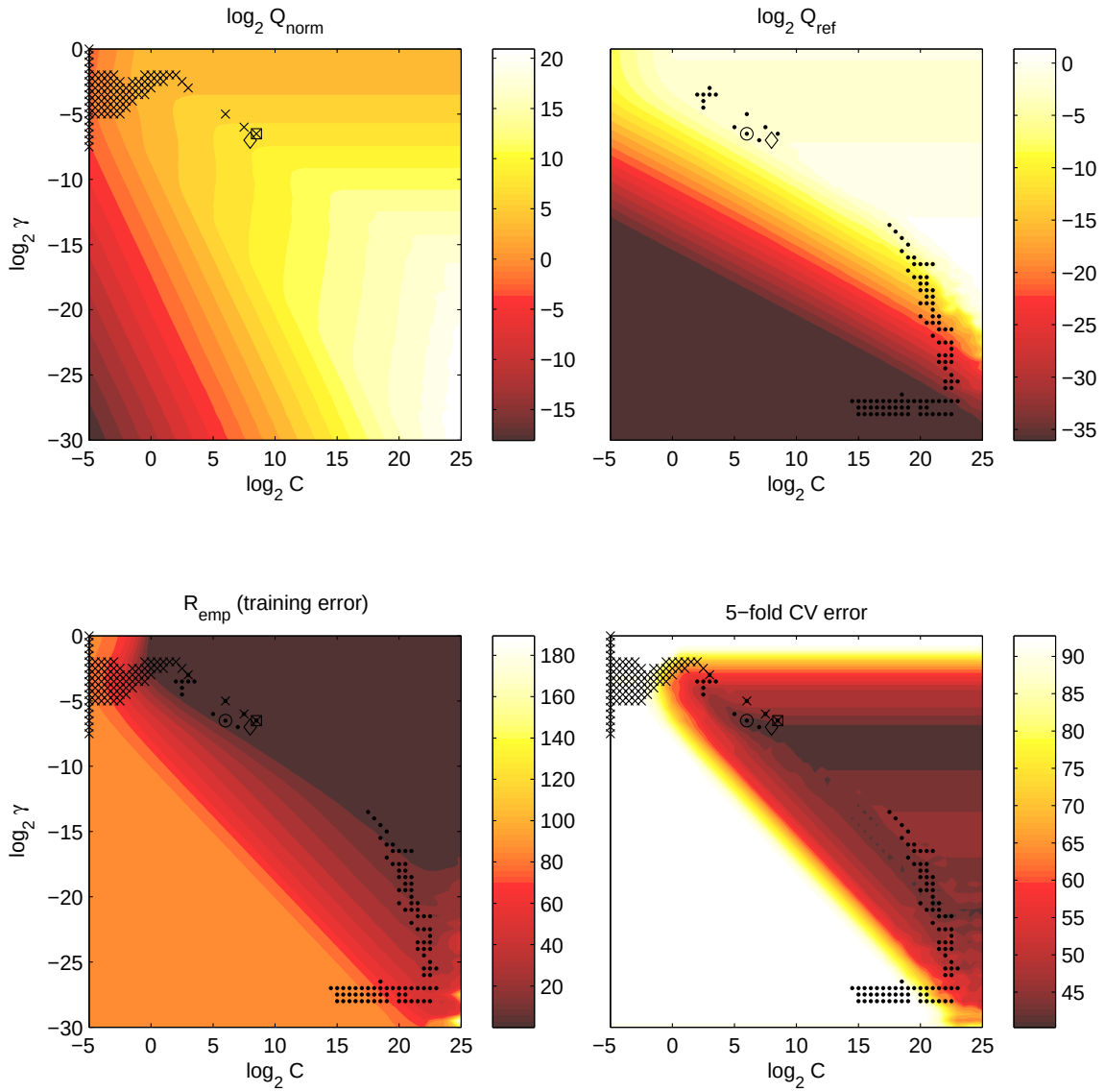

Figure C.5: Visualizations of the experiment results from Chapter 5 for *sonar* data-set and the Gaussian RBF kernel.

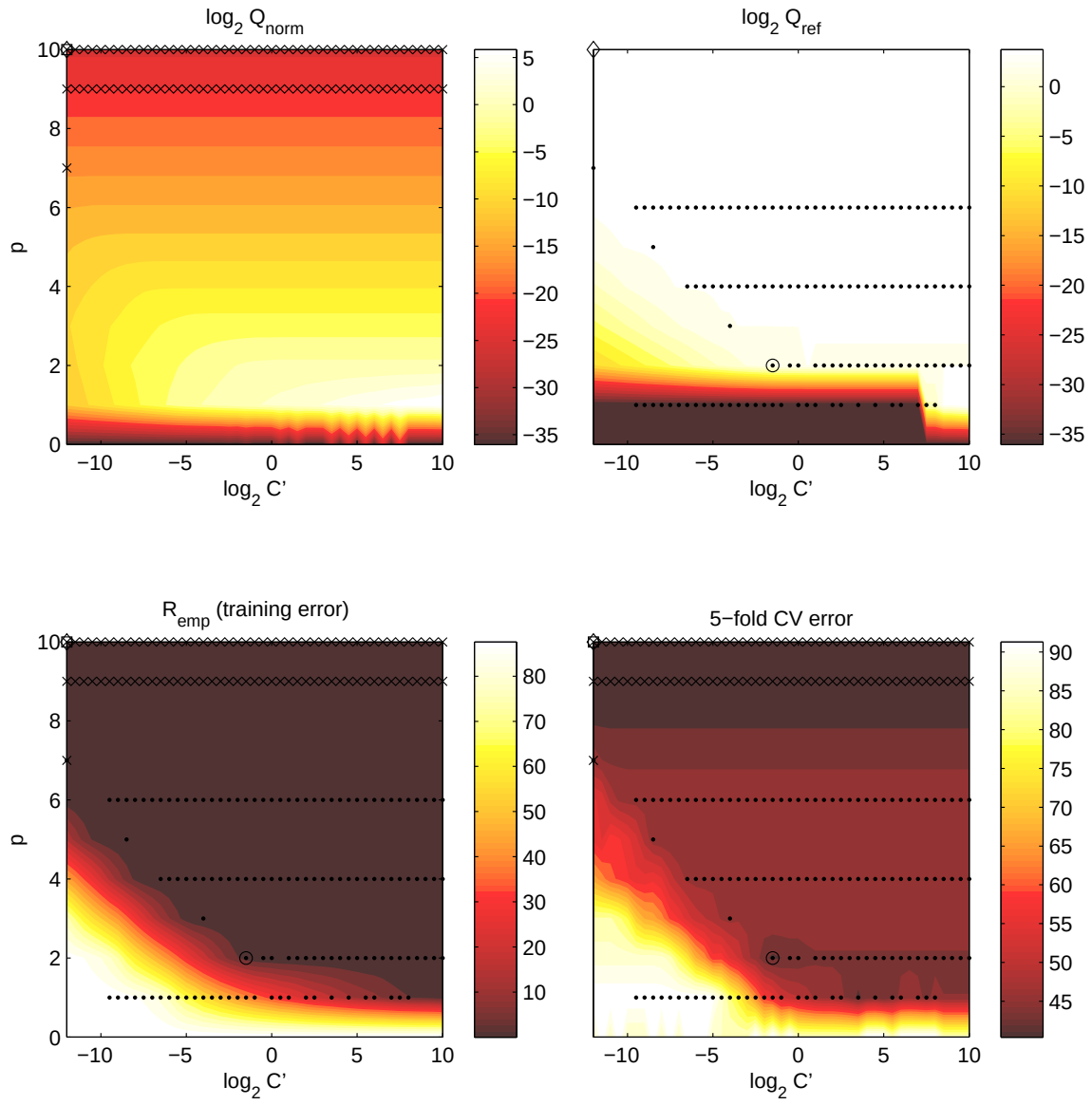

Figure C.6: Visualizations of the experiment results from Chapter 5 for *sonar* data-set and the polynomial kernel.

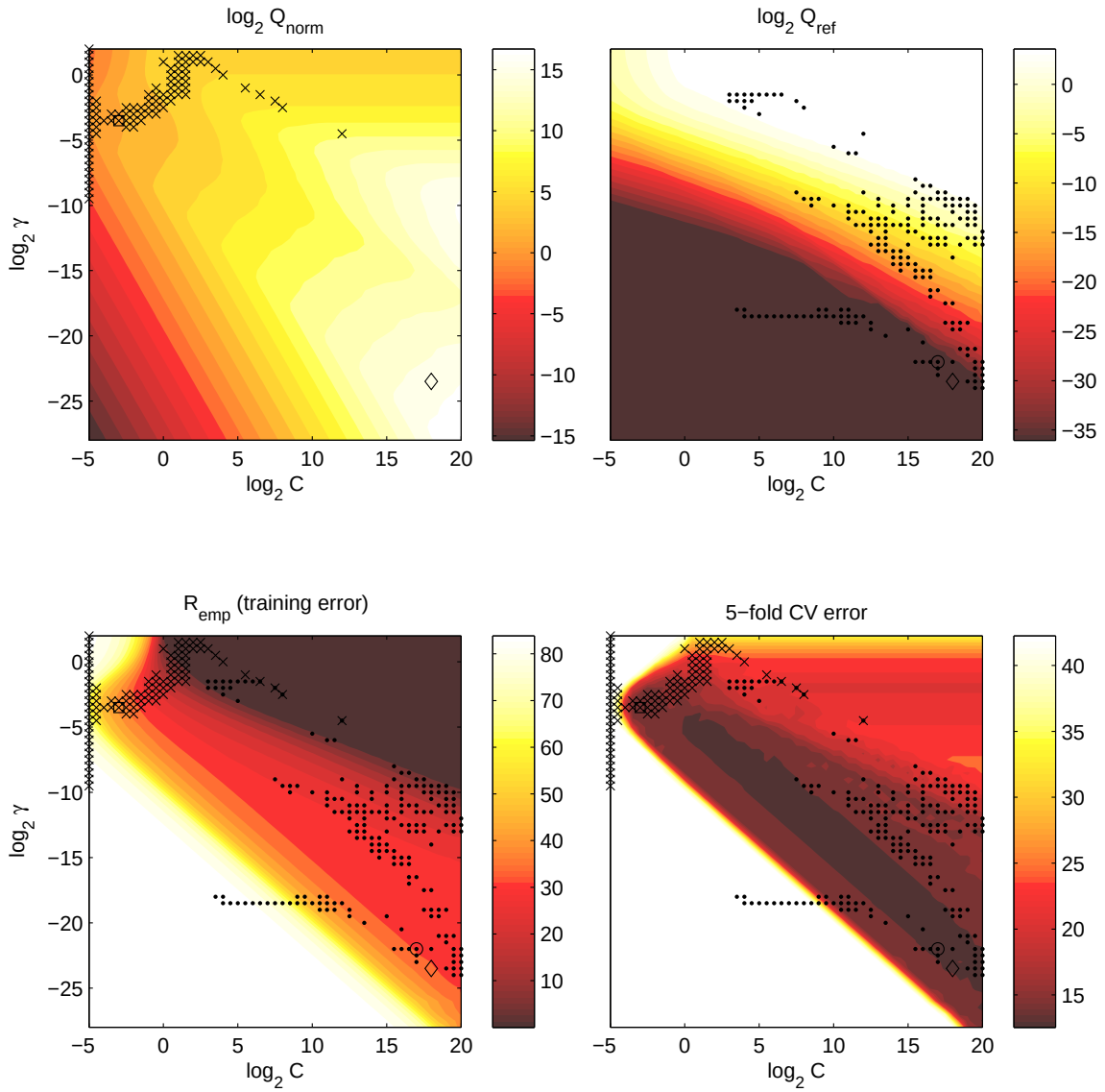

Figure C.7: Visualizations of the experiment results from Chapter 5 for *heart* data-set and the Gaussian RBF kernel.

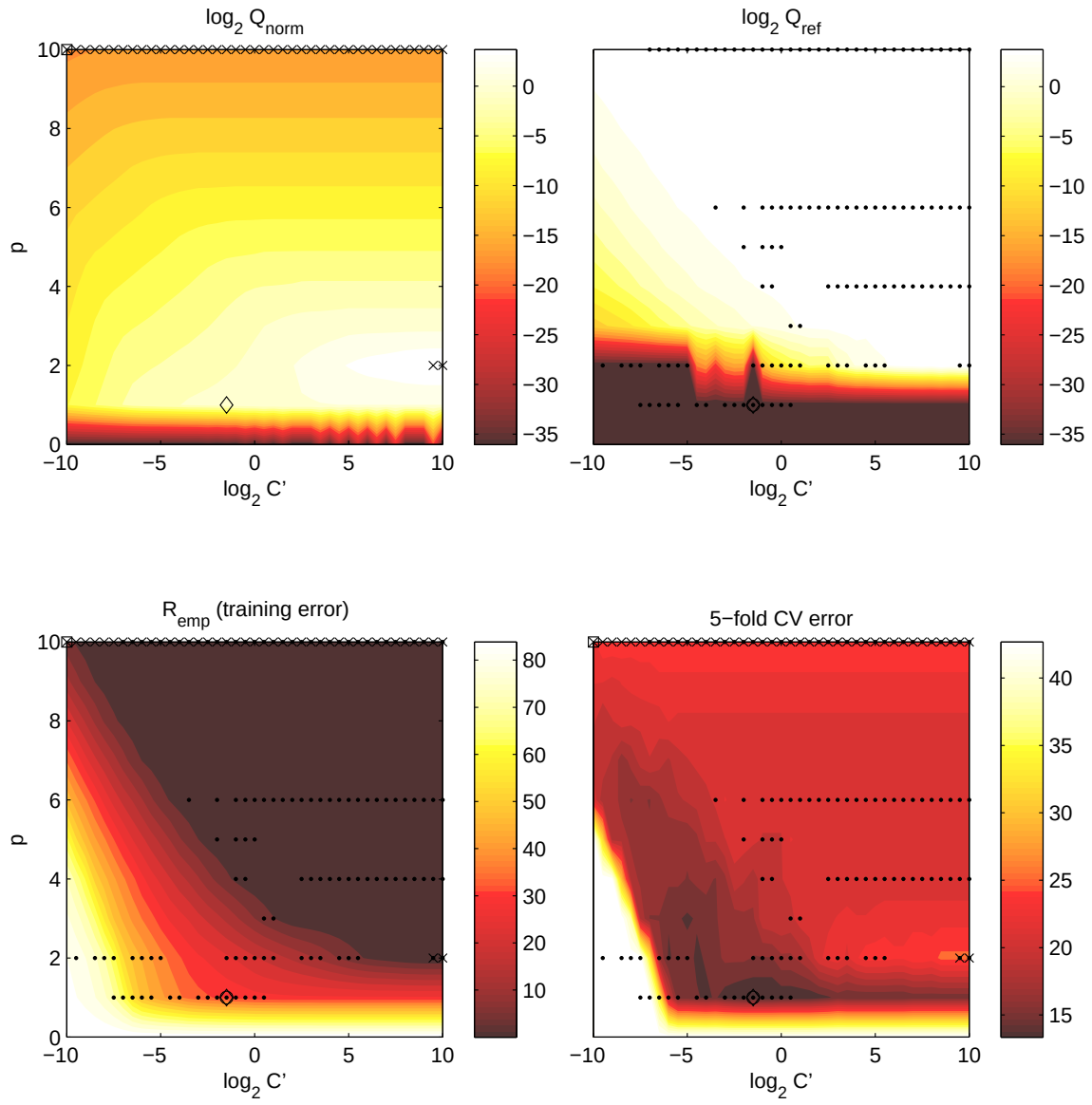

Figure C.8: Visualizations of the experiment results from Chapter 5 for *heart* data-set and the polynomial kernel.

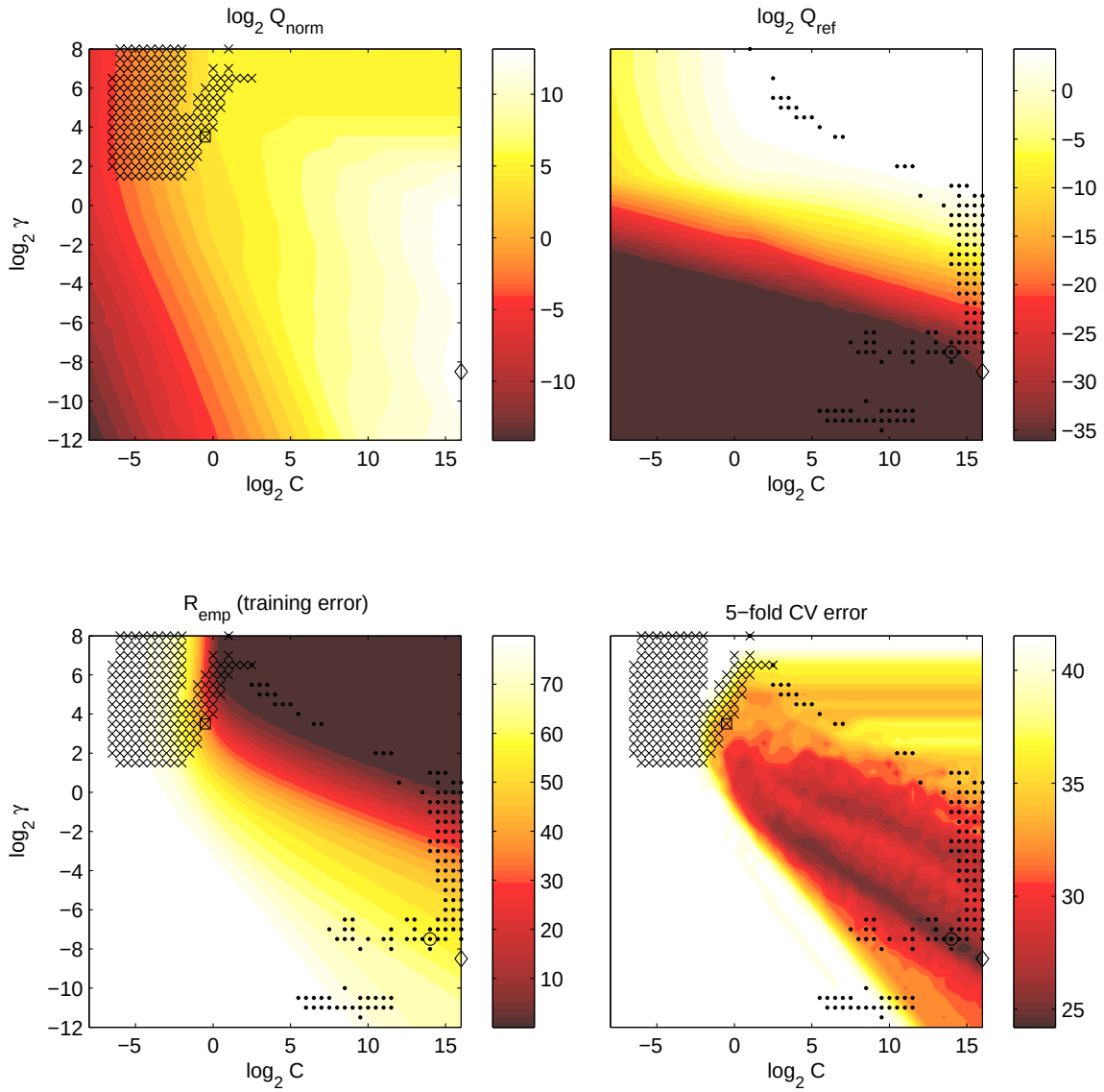

Figure C.9: Visualizations of the experiment results from Chapter 5 for *liver* data-set and the Gaussian RBF kernel.

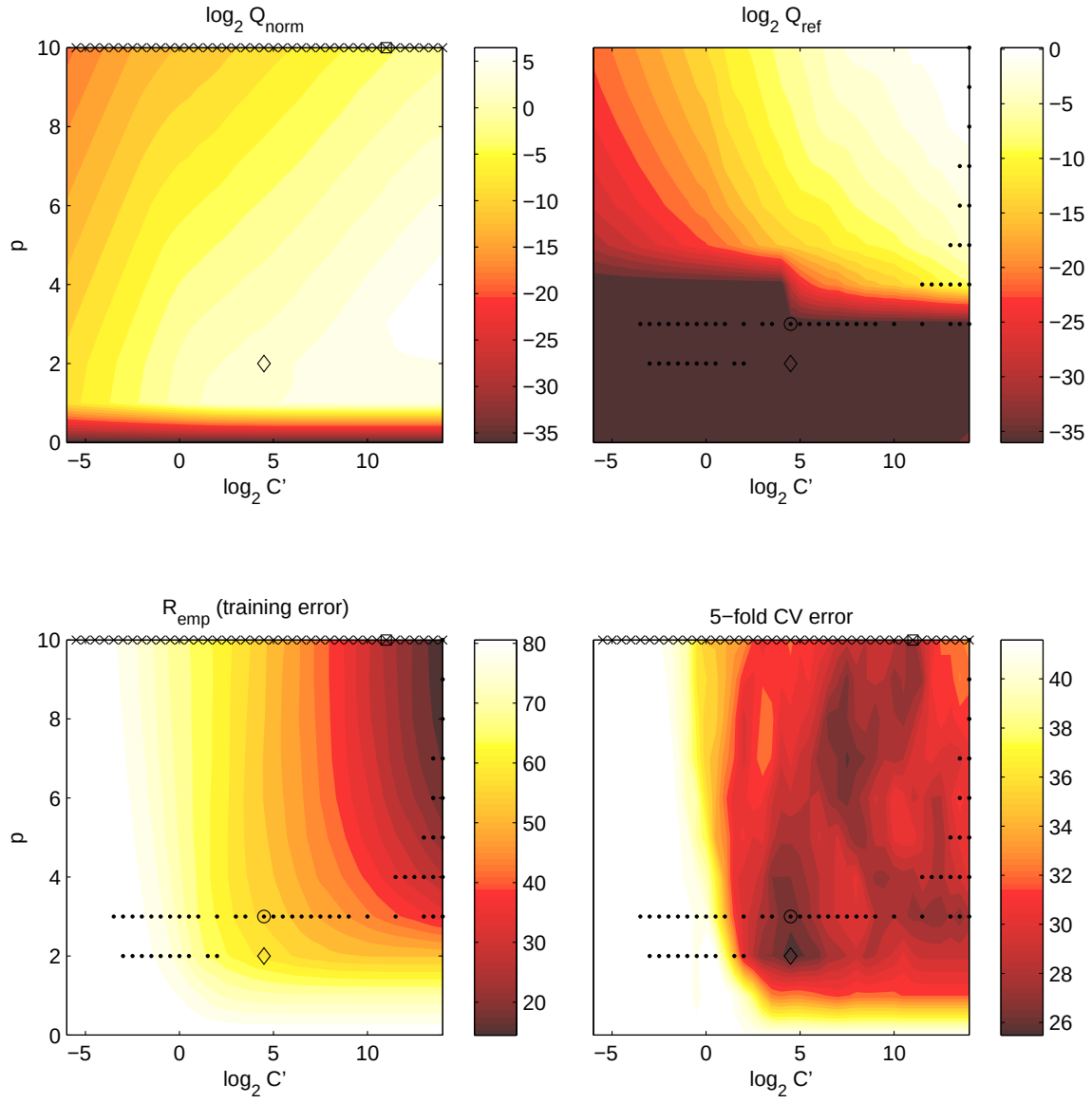

Figure C.10: Visualizations of the experiment results from Chapter 5 for *liver* data-set and the polynomial kernel.

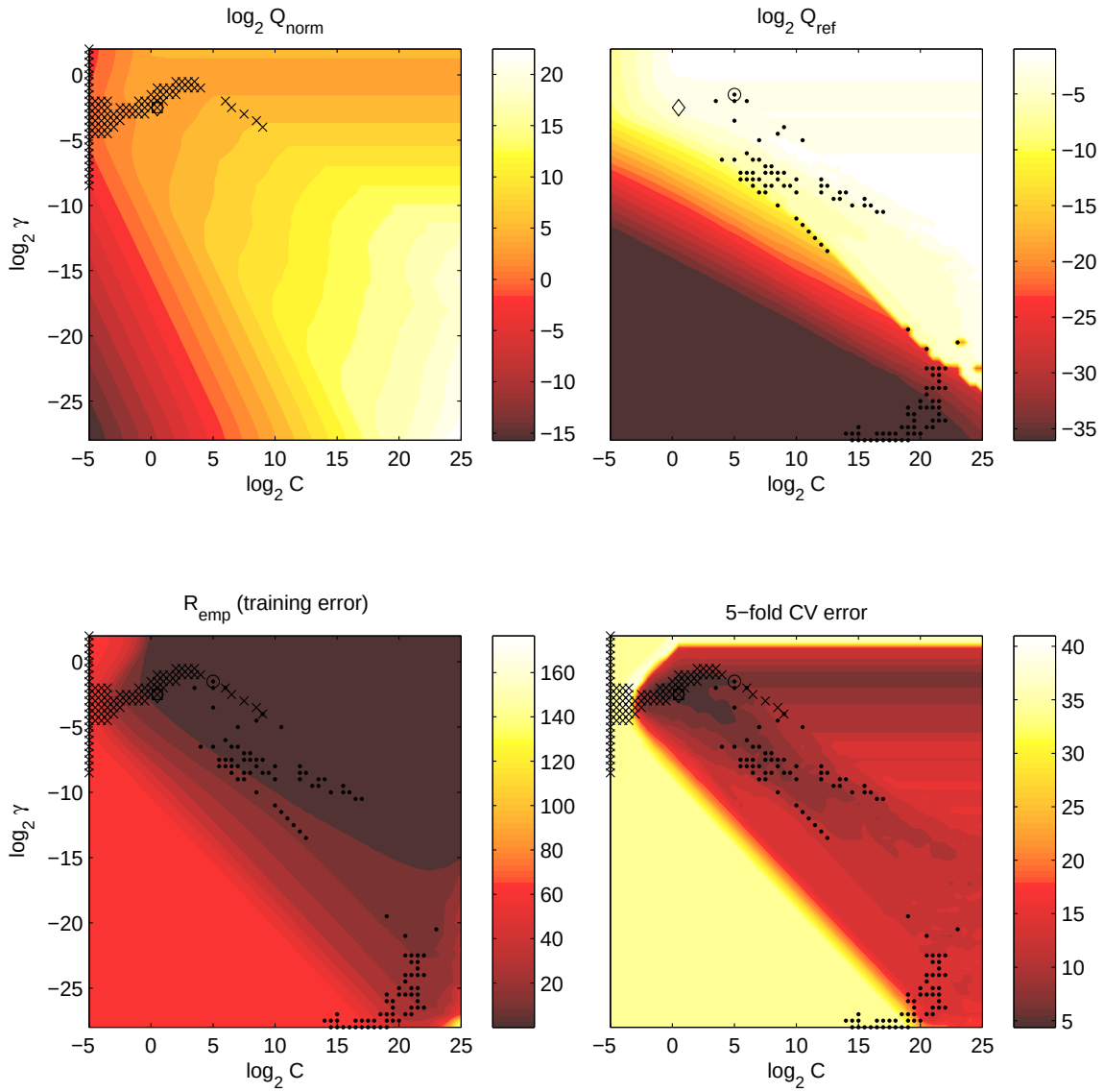

Figure C.11: Visualizations of the experiment results from Chapter 5 for *iono* data-set and the Gaussian RBF kernel.

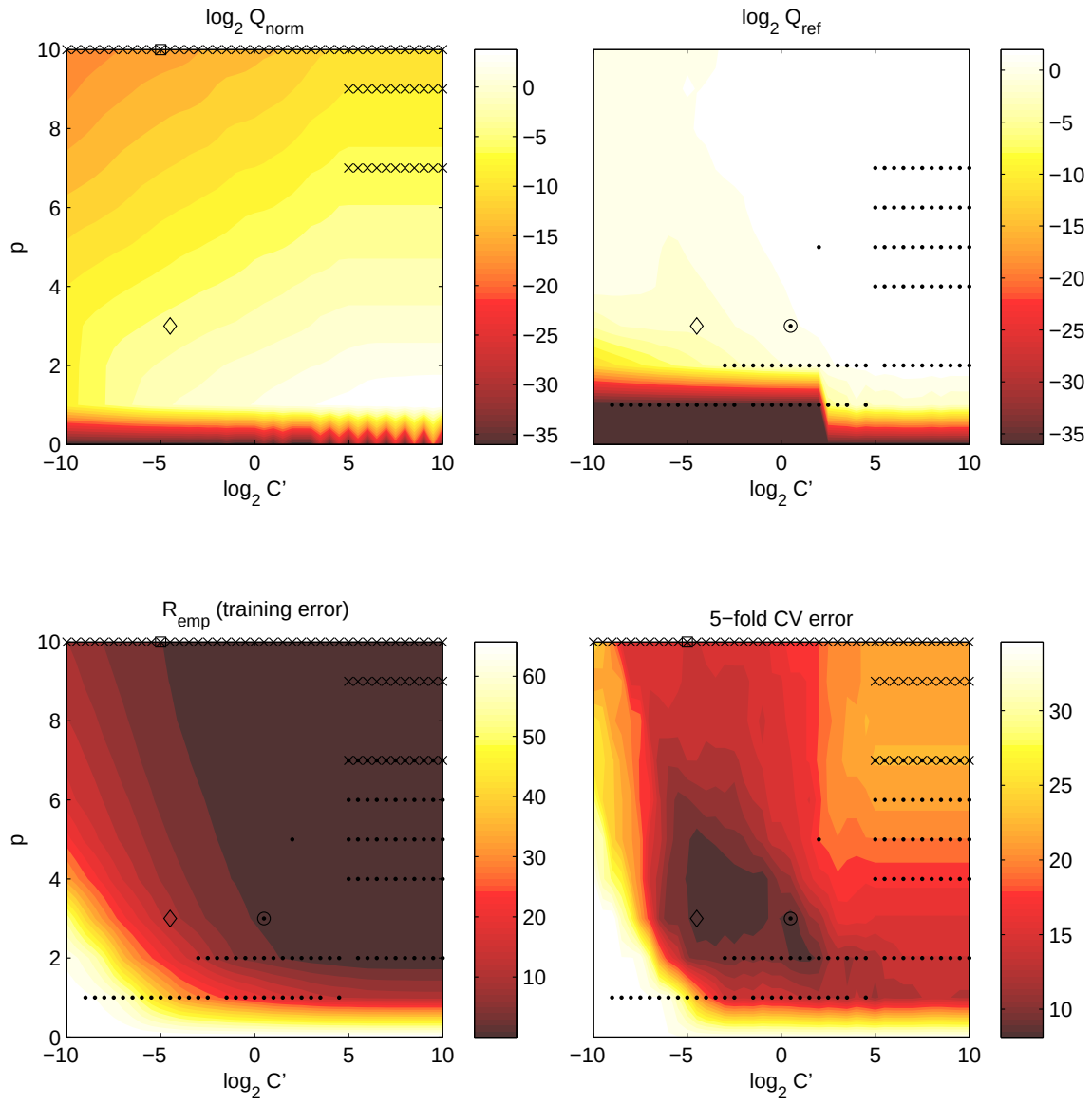

Figure C.12: Visualizations of the experiment results from Chapter 5 for *iono* data-set and the polynomial kernel.

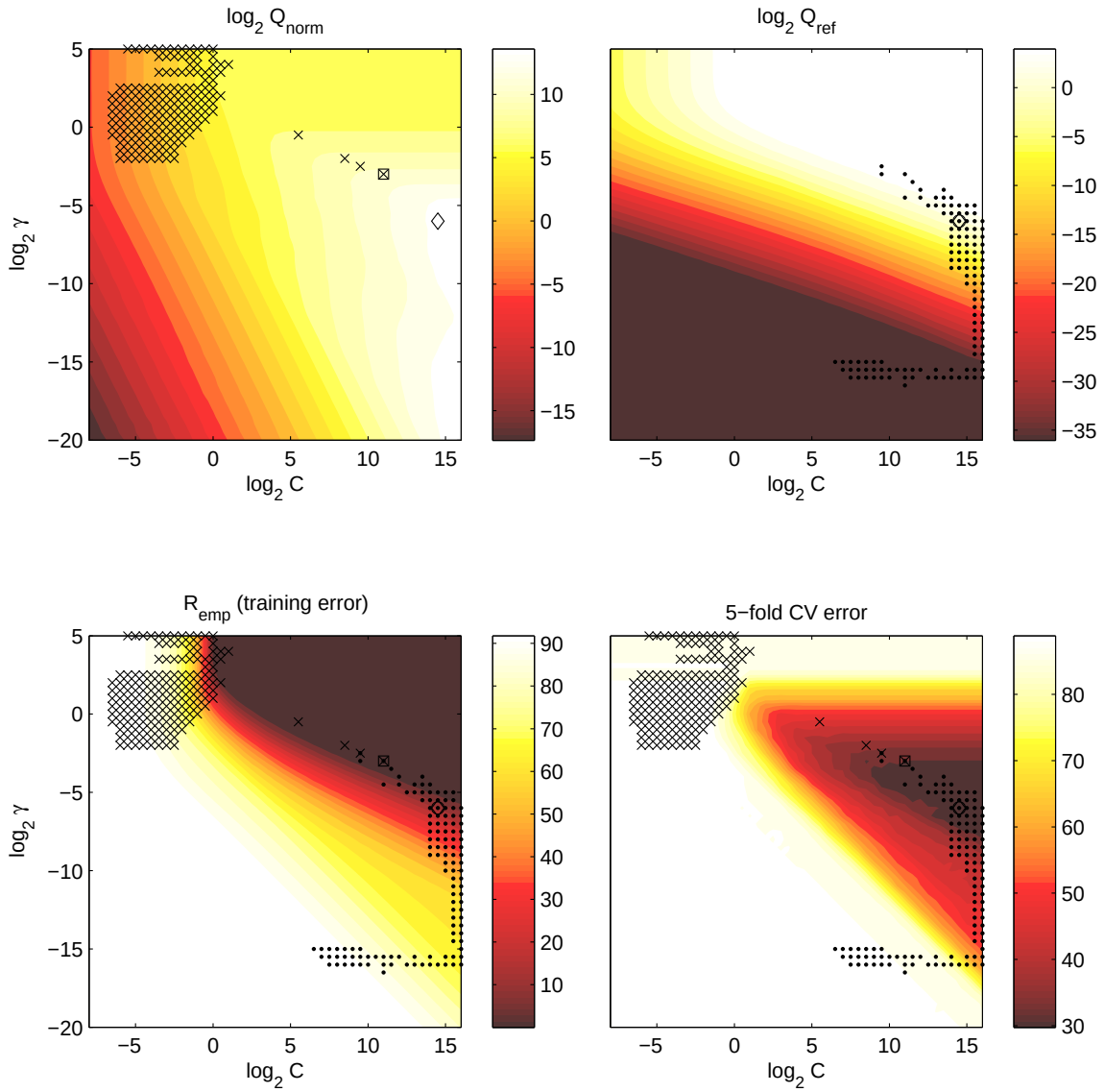

Figure C.13: Visualizations of the experiment results from Chapter 5 for *vehicle12* data-set and the Gaussian RBF kernel.

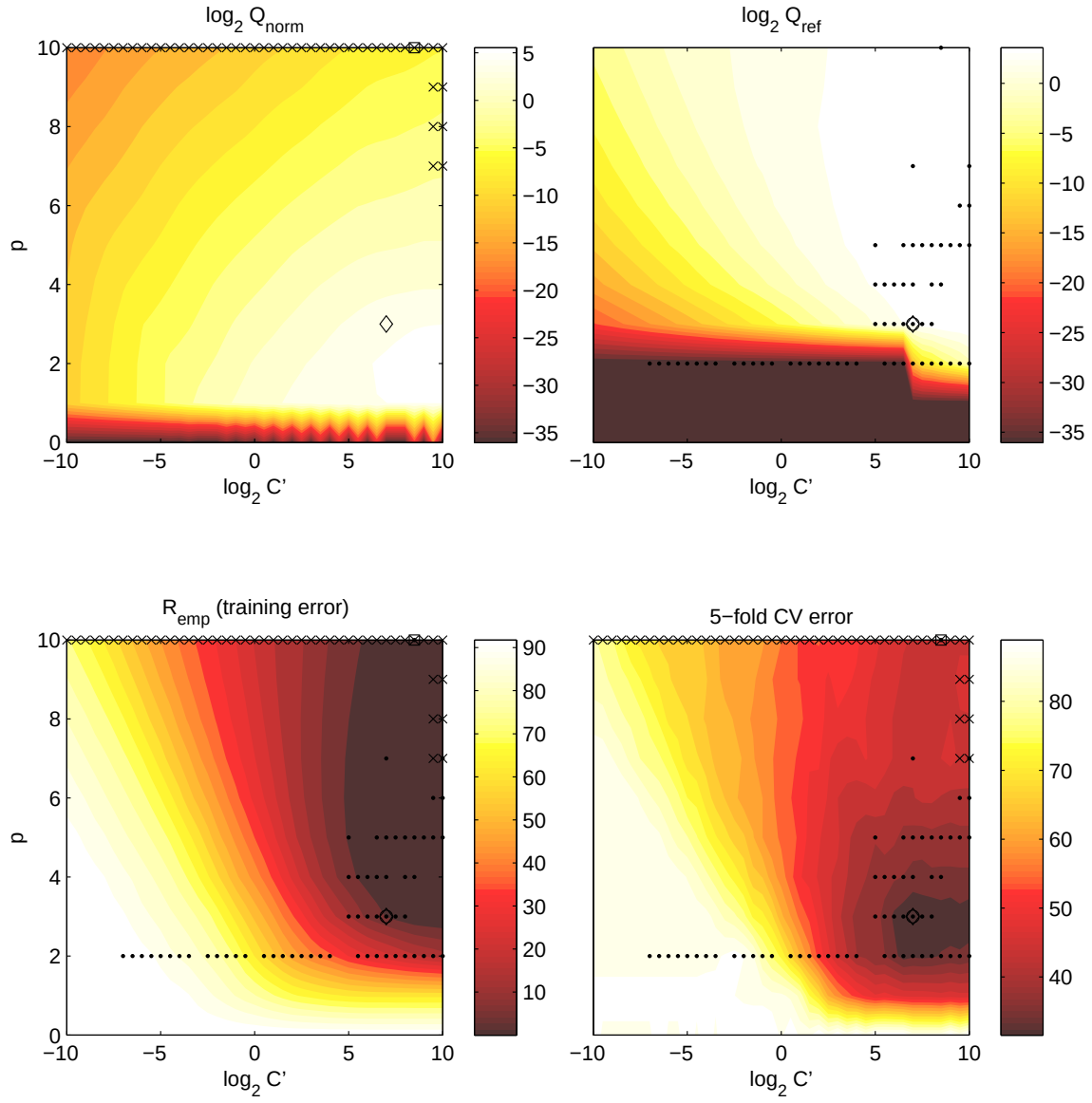

Figure C.14: Visualizations of the experiment results from Chapter 5 for *vehicle12* data-set and the polynomial kernel.

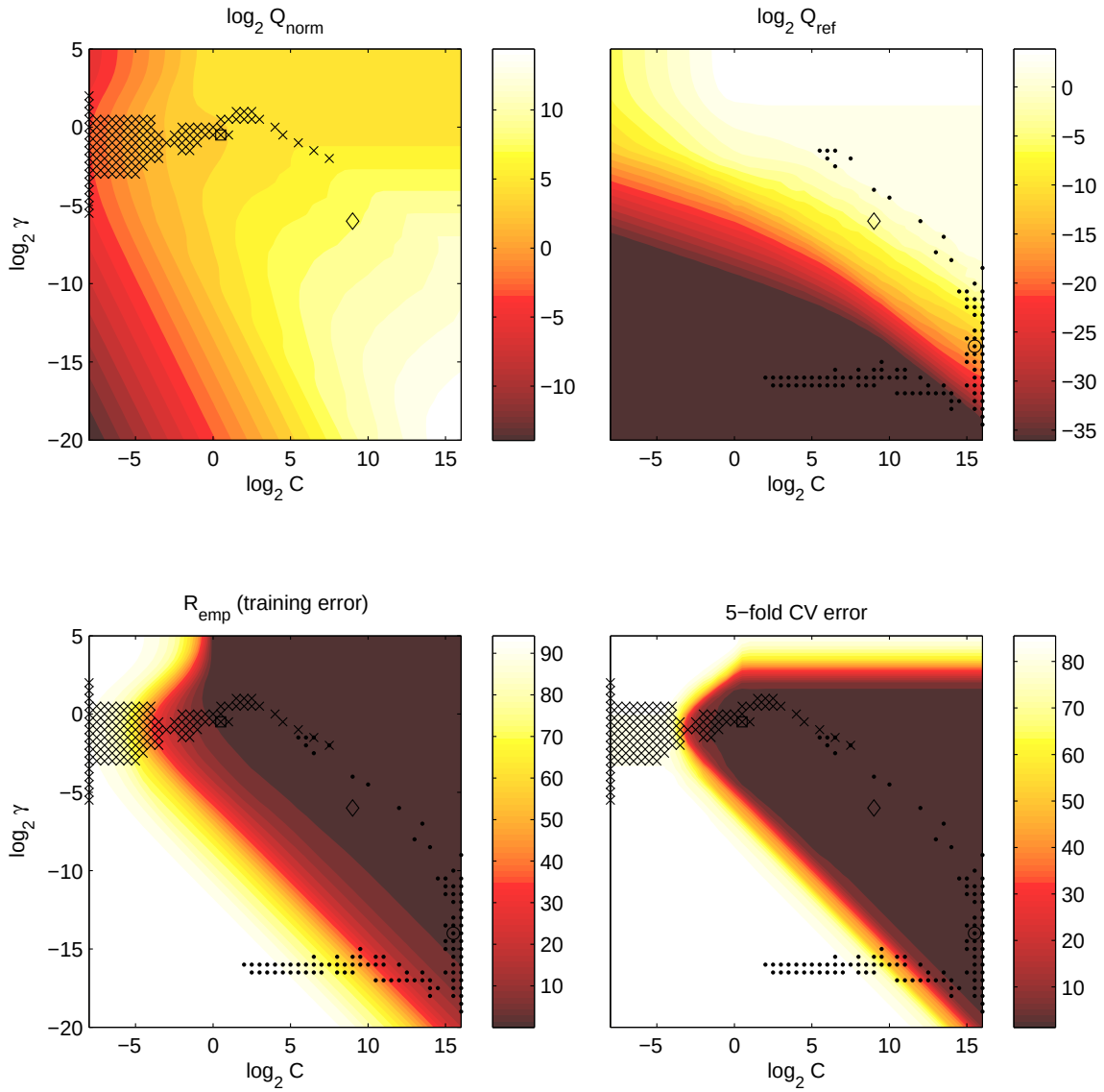

Figure C.15: Visualizations of the experiment results from Chapter 5 for *vehicle34* data-set and the Gaussian RBF kernel.

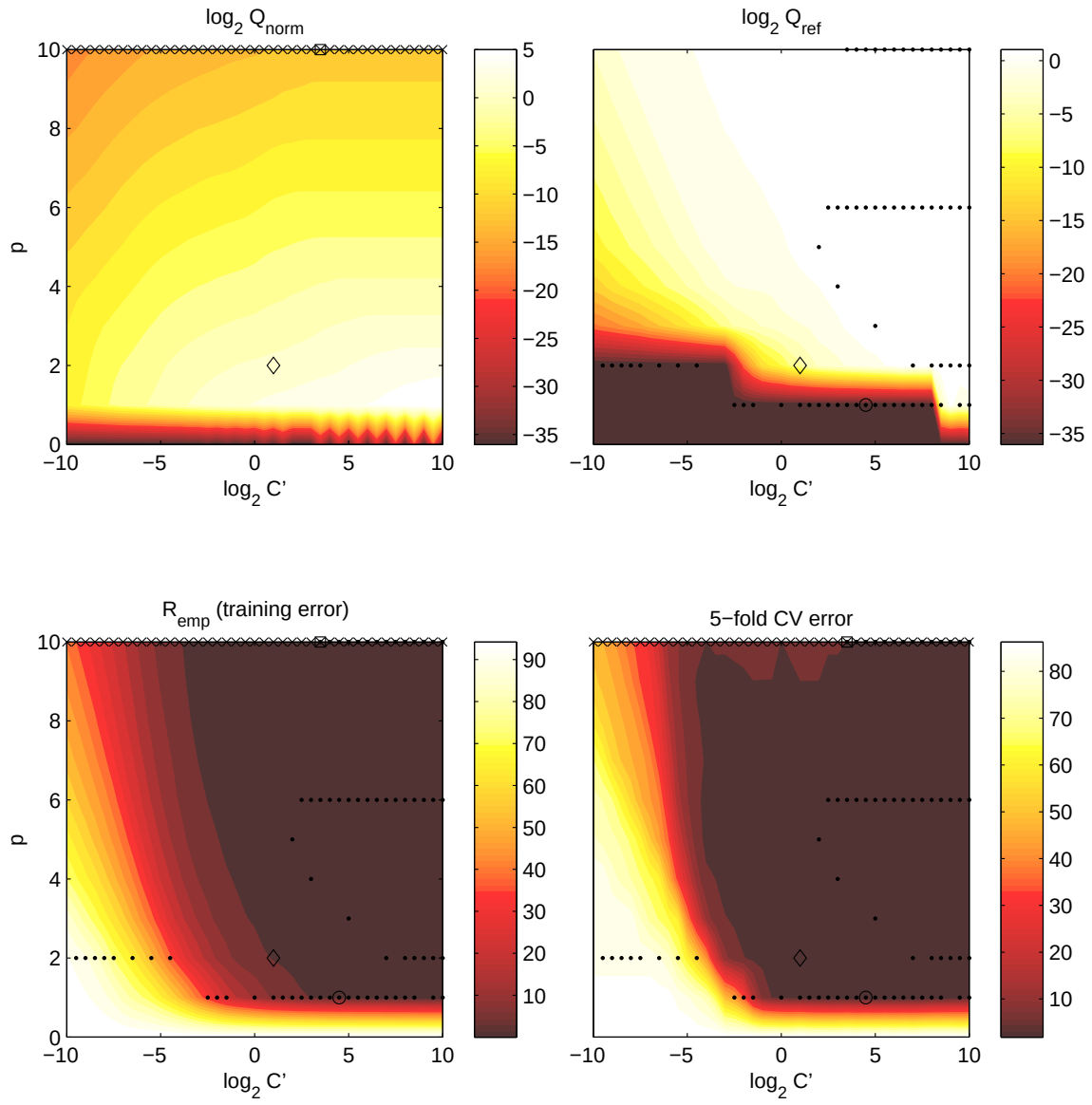

Figure C.16: Visualizations of the experiment results from Chapter 5 for *vehicle34* data-set and the polynomial kernel.

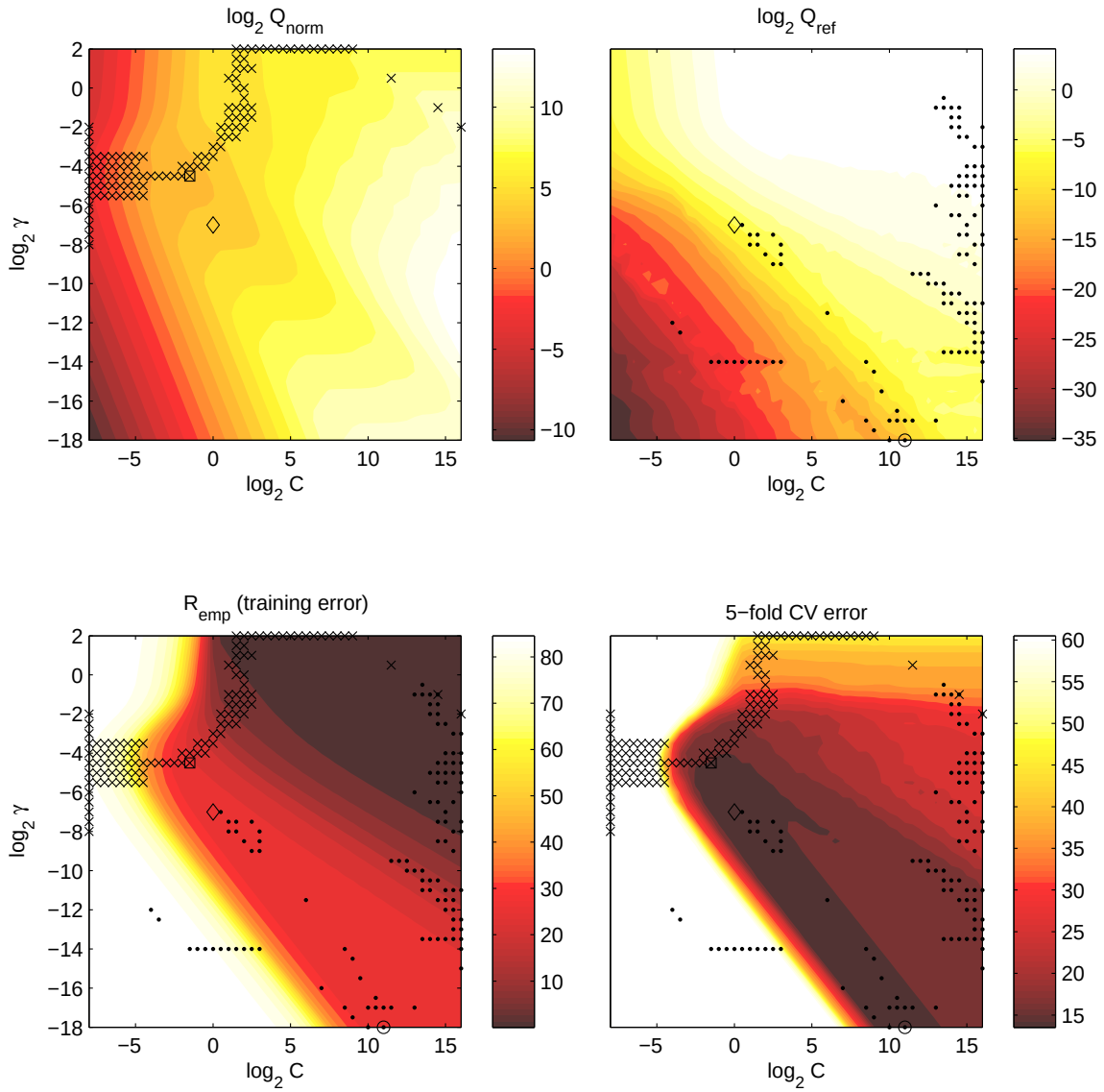

Figure C.17: Visualizations of the experiment results from Chapter 5 for *credit* dataset and the Gaussian RBF kernel.

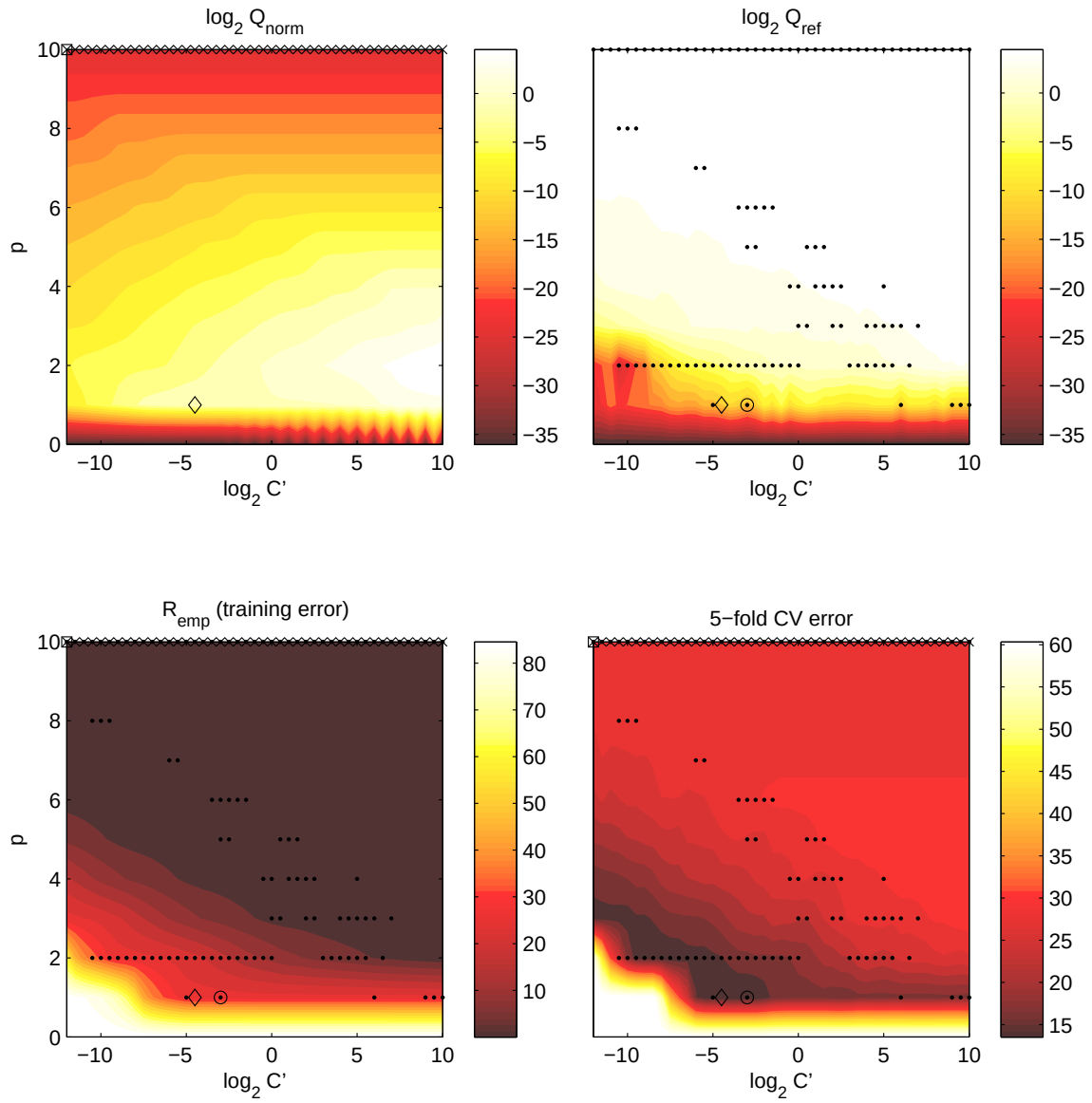

Figure C.18: Visualizations of the experiment results from Chapter 5 for *credit* dataset and the polynomial kernel.

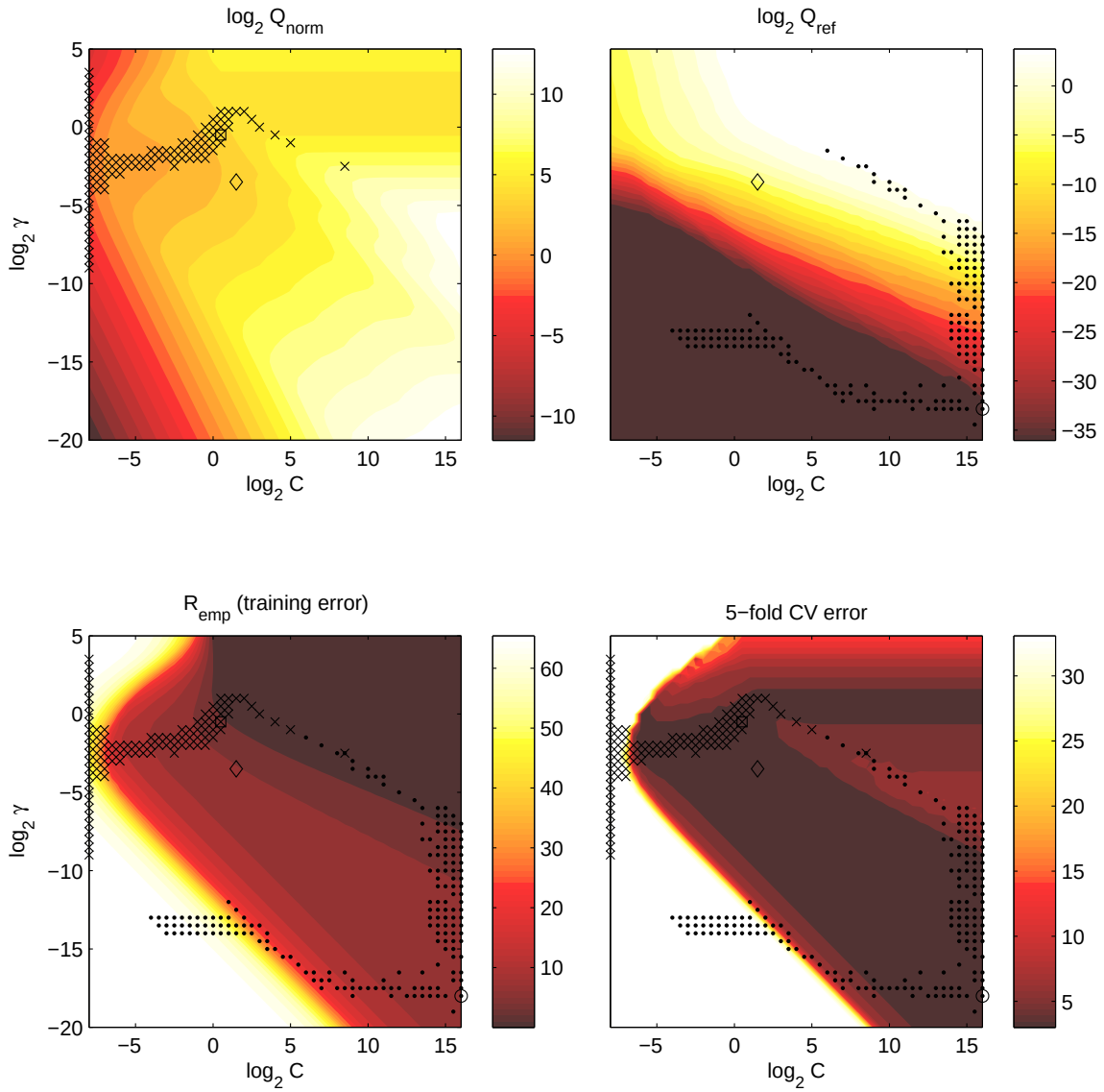

Figure C.19: Visualizations of the experiment results from Chapter 5 for *cancer* dataset and the Gaussian RBF kernel.

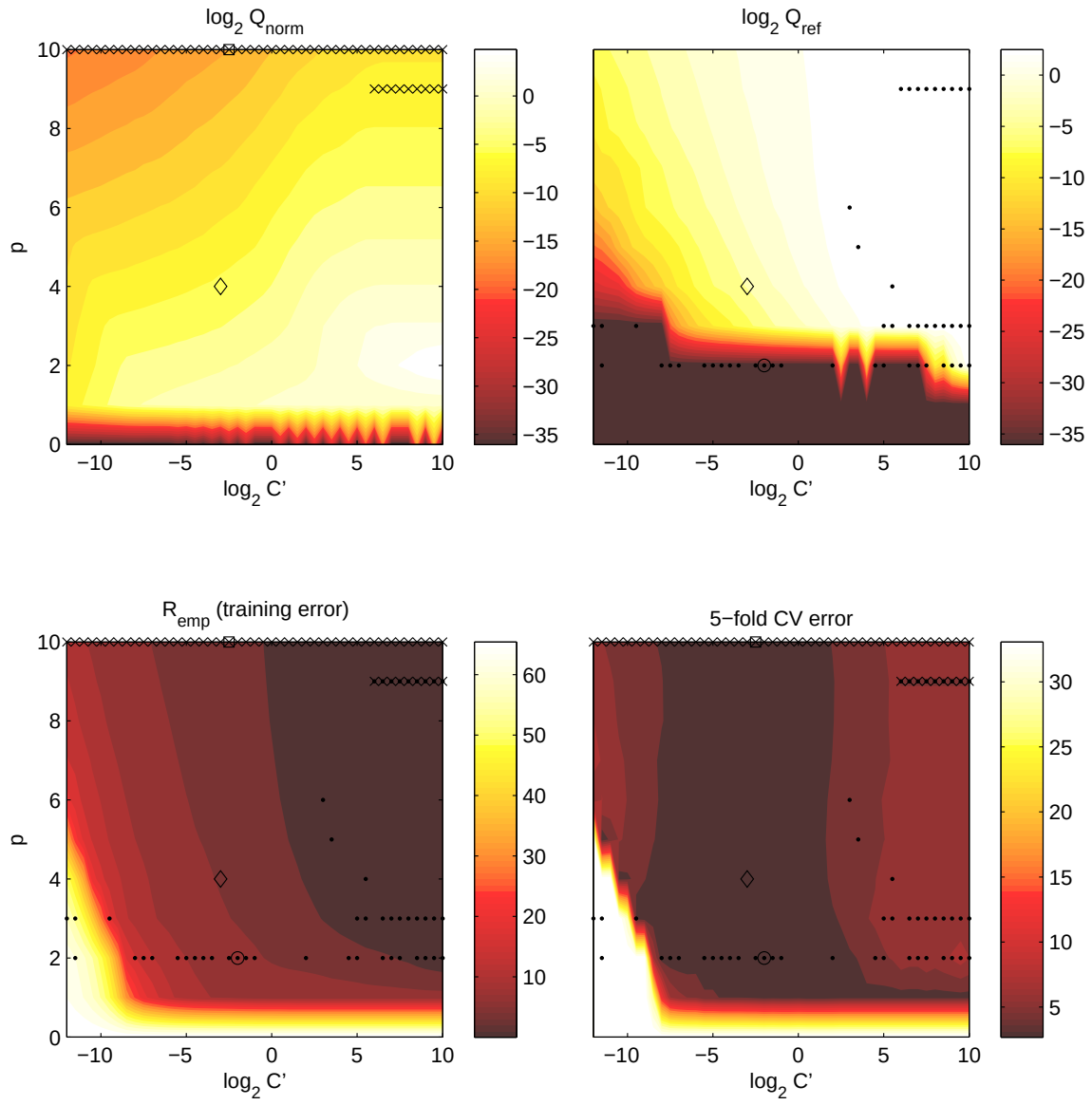

Figure C.20: Visualizations of the experiment results from Chapter 5 for *cancer* dataset and the polynomial kernel.

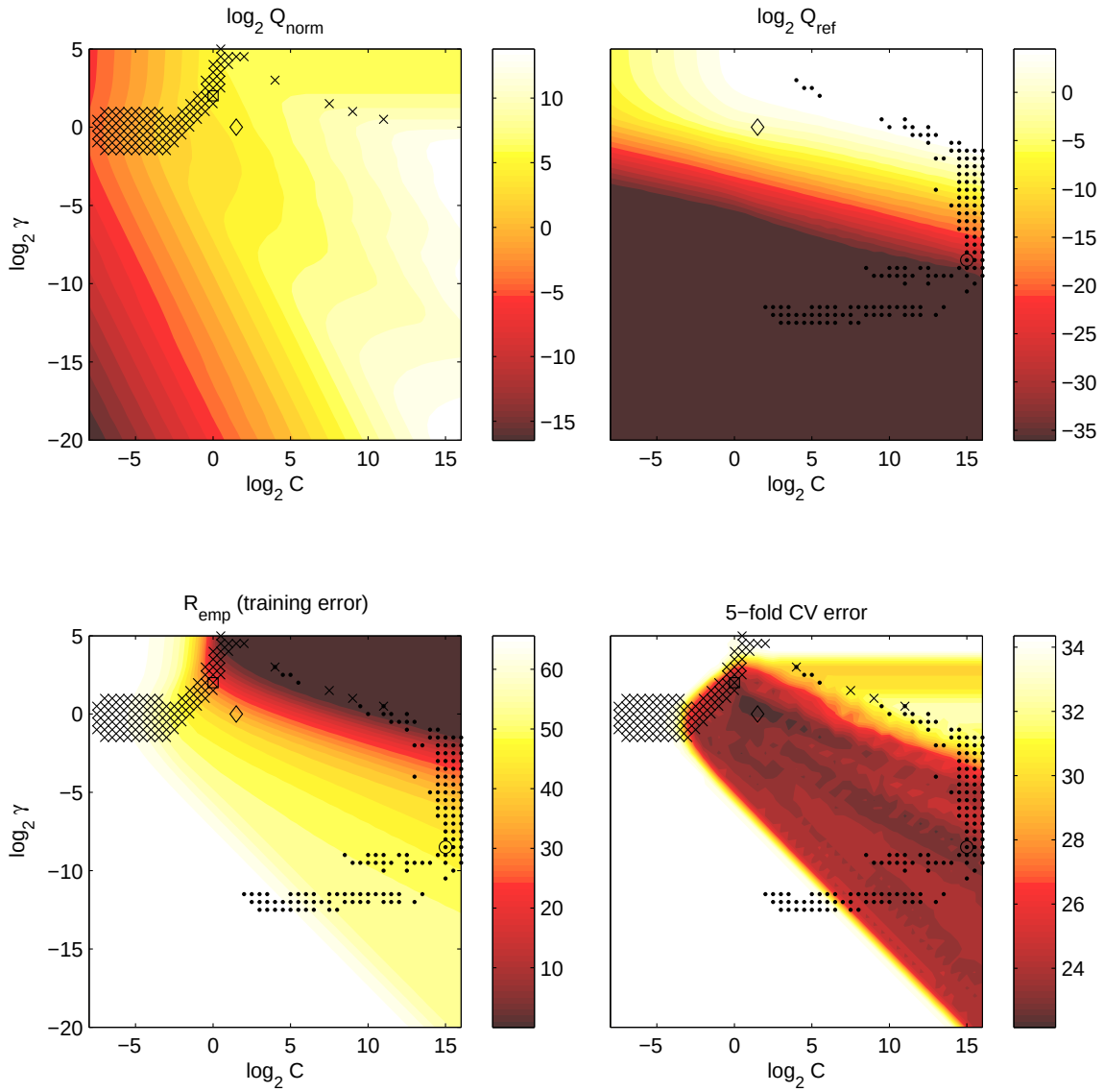

Figure C.21: Visualizations of the experiment results from Chapter 5 for *pima* dataset and the Gaussian RBF kernel.

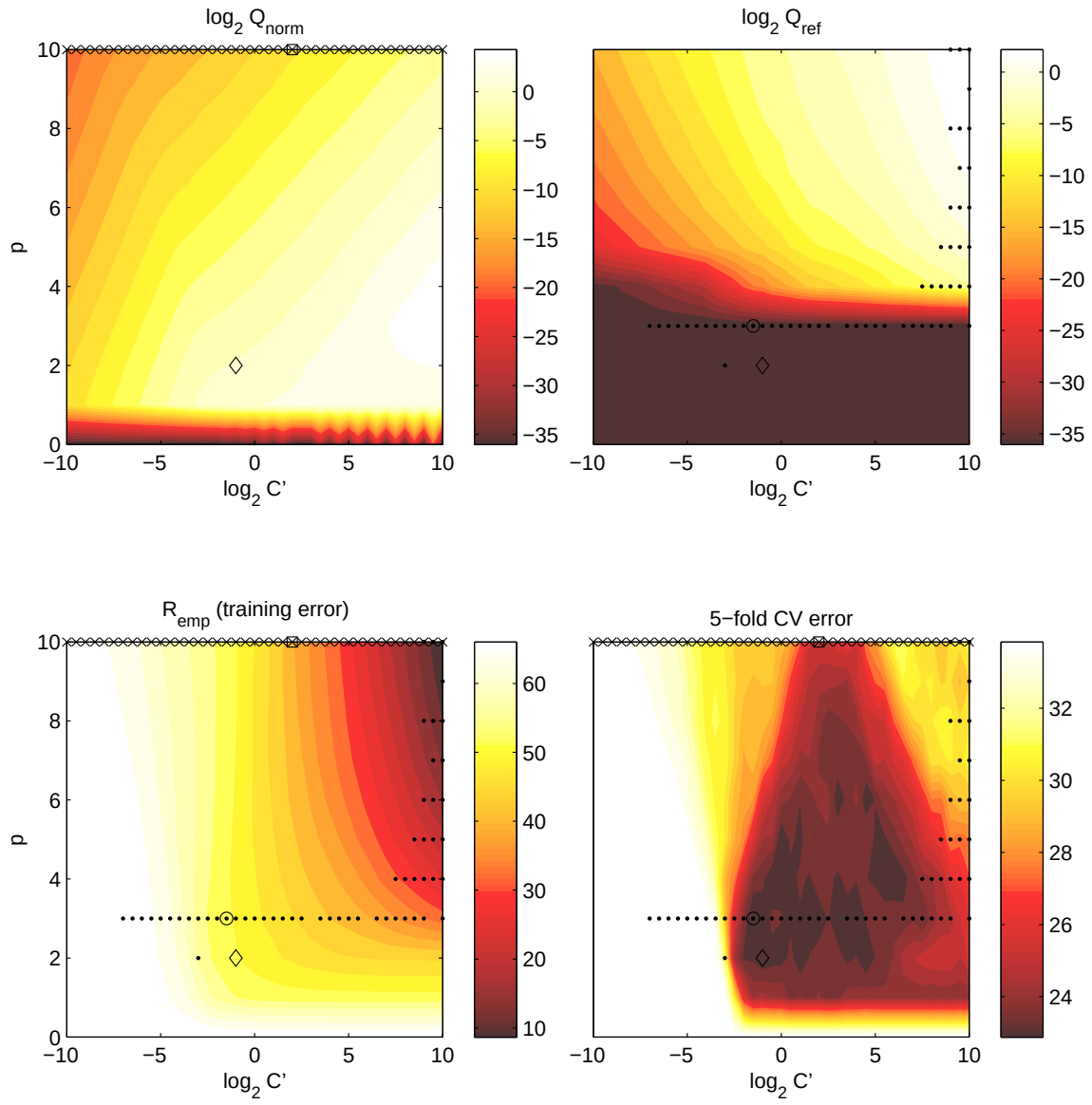

Figure C.22: Visualizations of the experiment results from Chapter 5 for *pima* dataset and the polynomial kernel.

Bibliography

R. A. Adams and J. J. Fournier. *Sobolev spaces*. Academic press, New York, second edition, 2003.

H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

N. Alon, S. Cesa-Bianchi, S. Ben-david, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 4:615–631, 1997.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

A. Asuncion and D. Newman. UCI machine learning repository, 2007. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 6, New York, NY, USA, 2004. ACM.

P. L. Bartlett. For valid generalization the size of the weights is more important than the size of the network. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, page 134. The MIT Press, 1997.

V. Bevilacqua, G. Mastronardi, F. Menolascina, P. Pannarale, and A. Pedone. A novel multi-objective genetic algorithm approach to artificial neural network topology optimisation: The breast cancer classification problem. In *International Joint Conferenct on Neural Networks (IJCNN'06), 2006*, pages 1958–1965, Vancouver, 2006.

J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

Y. Bodyanskiy, I. **Kokshenev**, Y. Gorshkov, and V. Kolodyazhniy. Outlier resistant recursive fuzzy clustering algorithms. In B. Reusch, editor, *Computational Intelligence, Theory and Applications (Series: Advances in Soft Computing)*, volume 38, pages 647–652. Springer-Verlag, Berlin Heidelberg, 2006.

Y. Bodyanskiy, Y. Gorshkov, I. **Kokshenev**, and V. Kolodyazhniy. *Evolving intelligent systems: Methodology and applications*, chapter Evolving Fuzzy Classification of Non-Stationary Time Series, pages 301–313. John Wiley, New York, 2010. ISBN 978-0470287194.

O. Bousquet and A. Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, 2002. ISSN 1532-4435.

K. P. Burnham and R. D. Anderson. *Model selection and inference: a practical information theoretic approach*. Springer-Verlag, New York, 1998.

C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

V. Chankong and Y. Haimes. *Multiobjective Decision Making: Theory and Methodology*. Elsevier Science, New York, 1983.

O. Chapelle and V. Vapnik. Model selection for support vector machines. In *Advances in Neural Information Processing Systems*, pages 230–236. MIT Press, 1999.

O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Mach. Learn.*, 46(1-3):131–159, 2002. ISSN 0885-6125. doi: http://dx.doi.org/10.1023/A:1012450327387.

S. Chen, C. F. N. Cowan, and P. M. Grant. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Trans. on Neural Networks*, 2:302 – 309, 1991.

S. Chen, E. S. Chng, and K. Alkadhimi. Regularized orthogonal least squares algorithm for constructing radial basis function networks. *International Journal of Control*, 64(5):829–837, 1996.

S. Chiu. Fuzzy model identification based on cluster estimation. *Journal of Intelligent & Fuzzy Systems*, 2(3):209 – 219, 1994.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

M. A. Costa and A. P. Braga. Optimization of neural networks with multi-objective lasso algorithm. In *International Joint Conferenct on Neural Networks (IJCNN'06), 2006*, pages 6344–6350, Vancouver, 2006.

M. A. Costa, A. Braga, B. R. Menezes, R. A. Teixeira, and G. G. Parma. Training neural networks with a multi-objective sliding mode control algorith. *Neurocomputing*, 51:467–473, 2003.

T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *Electronic Computers, IEEE Transactions on EC*, 14(3):326–334, 1965.

I. Das and J. E. Dennis. A closer look at drawbacks of minimizing weighted sums of objectives for pareto set generation in multicriteria optimization problems. *Structural and Multidisciplinary Optimization*, 14(1):63–69, 1997.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32: 407–499, 2004.

M. Ehrgott. *Multicriteria optimization*. Springer Berlin Heidelberg, 2005.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, (1):1–50, 2000.

E. Fix and J. L. Hodges. Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical report, U.S. Air Force, School of Aviation Medicine, 1951.

C. M. Fonseca and P. J. Fleming. An overview of evolutionary algorithms in multiobjective optimization. *Evolutionary Computation*, 3(1):1–16, 1995.

S. Forrest, editor. *Genetic algorithms for multiobjective optimization: Formulation, discussion and generalization*, 1993.

M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32:675–701, 1937.

M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics*, (11):86–92, 1940.

S. Geman, E. Bienenstock, and R. Doursat. Neural Network and the Bias/Variance Dilemma. *Neural Computation*, 4:1–58, 1992.

A. Geoffrion. Proper efficiency and the theory of vector maximization. *Journal of Mathematical Analysis and Applications*, 22:618–630, 1968.

F. Girosi, M. Jones, and T. Poggio. Priors stabilizers and basis functions: From regularization to radial, tensor and additive splines. Technical report, Cambridge, MA, USA, 1993.

F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7(2):219–269, 1995.

G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2), 1979.

J. Gonzalez, I. Rojas, J. Ortega, H., F. J. Fernandez, and A. F. Diaz. Multiobjective evolutionary optimization of the size, shape, and position parameters of radial basis function networks for function approximation. *IEEE Trans. on Neural Networks*, 14(6):1478–1495, 2003.

I. Guyon, A. Saffari, G. Dror, and G. Cawley. Model selection: Beyond the bayesian/frequentist divide. *Journal of Machine Learning Research*, 11:61–87, 2010.

Y. Haimes, L. Lasdon, and D. Wismer. On a bicriterion formulation of the problems of integrated system identification and system optimization. *IEEE Transactions on Systems, Man, and Cybernetics*, 1:296–297, 1971.

T. Hanne. Global multiobjective optimization using evolutionary algorithms. *Journal of Heuristics*.

P. C. Hansen and D. P. O'Leary. The use of the l-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput.*, 14(6):1487–1503, 1993. ISSN 1064-8275. doi: http://dx.doi.org/10.1137/0914086.

T. Hatanaka and N. K. K. Uosaki. Multi-objective structure selection for radial basis function networks based on genetic algorithm. *Evolutionary Computation, 2003. CEC '03. The 2003 Congress on*, 2:1095–1100, Dec. 2003. doi: 10.1109/CEC.2003.1299790.

S. Haykin. *Neural Networks. A Comprehensive Foundation*. Prentice Hall, Inc., Upper Saddle River, N.J., 1999.

T. Hofmann, B. Schöolkopf, and A. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 96(9):1171–1220, 2008.

J. Jahn. *Vector Optimization: Theory, Applications, and Extensions*. Springer, 2004.

J.-S. R. Jang and C.-T. Sun. Functional equivalence between radial basis function networks and fuzzy inference systems. *IEEE Trans. on Neural Networks*, 4(1):156–159, 1993.

J.-S. R. Jang, C.-T. Sun, and E. Mizutani. *Neuro-Fuzzy and Soft Computing: a computational approach to learning and machine intelligence*. Prentice-Hall, Inc., Upper Saddle River, NJ, 1997.

Y. Jin, editor. *Multi-Objective Machine Learning (Series: Studies in Computational Intelligence)*, volume 16. Heidelberg: Springer Verlag, 2006.

Y. Jin and B. Sendhoff. Pareto-based multi-objective machine learning: An overview and case studies. *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 38 (3):397–415, 2008.

Y. Jin, T. Okabe, and B. Sendhoff. Neural network regularization and ensembling using multi-objective evolutionary algorithms. *Evolutionary Computation, CEC2004*, 1:1–8, 2004.

S. Karlin. *Mathematical Methods and Theory in Games, Programming, and Economics, Vol. 1: Matrix Games, Programming, and Mathematical Economics.* Addison-Wesley, 1959.

S. Keerthi. Efficient tuning of svm hyperparameters using radius/margin bound and iterative algorithms. *Neural Networks, IEEE Transactions on*, 13(5):1225–1229, Sep 2002.

I. **Kokshenev** and A. P. Braga. Complexity bounds of radial basis functions and multi-objective learning. In *ESANN*, pages 73–78, 2007.

I. **Kokshenev** and A. P. Braga. A multi-objective approach to RBF network learning. *Neurocomputing*, 71(7-9):1203–1209, 2008a.

I. **Kokshenev** and A. P. Braga. A multi-objective learning algorithm for rbf neural network. In *10th Brazilian Symposium on Neural Networks, SBRN 2008*, pages 9–14, 2008b.

I. **Kokshenev** and A. P. Braga. An efficient multi-objective learning algorithm for RBF neural network. *Neurocomputing*, 2010. accepted for publication.

N. Kondo, T. Hatanaka, and K. Uosaki. Pattern classification via multi-objective evolutionary rbf networks ensemble. In *SICE-ICASE, 2006. International Joint Conference*, pages 137–142, 2006. doi: 10.1109/SICE.2006.315388.

H. König. *Eigenvalue Distribution of Compact Operators*. Basel, Switzerland: Birkhäuser, 1986.

K. Lang and M. Witbrock. Learning to tell two spirals apart. In *Proc. of the Connectionist Models Summer School*. Morgan Kaufmann, 1988.

W. A. Light. Some aspects on radial basis function approximation. In *Approximation Theory, Spline Functions and Applications*, pages 163–190, 1992.

G. Liu. and V. Kadirkamanathan. Learning with multi-objective criteria. In *Artificial Neural Networks, Fourth International Conference on*, pages 53–58, 1995.

G. Liu, J. Yang, and J. Whidborne. *Multiobjective optimisation and control*. Research Studies press, Baldock, Hertfordshire, England, 2003.

D. Lowe. Adaptive radial basis function nonliniarities, and the problem of generalisation. In *Proc. of the First IEE International Conference on Adaptive Signal Processing*, pages 171–175, 1989.

A. Luntz and V. Brailovsky. On estimation of characters obtained in statistical procedure of recognition (in russian). *Technicheskaya Kibernetica*, (3), 1969.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, Berkeley, University of California Press, 1967.

W. McCulloch and W. Pitts. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, (5):115–133, 1943.

J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, A 209:415–446, 1909.

D. Meyer, F. Leisch, and K. Hornik. The support vector machine under test. *Neurocomputing*, 55 (1-2):169–186, 2003.

C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *J. of Machine Learning Research*, 6:1099–1125, 2005.

C. A. Michelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2:11 –22, 1986.

H. Q. Minh, P. Niyogi, and Y. Yao. Mercer's theorem, feature maps, and smoothing. In *COLT*, volume 4005 of *Lecture Notes in Computer Science*, pages 154–168. Springer, 2006.

H. Minkowski. *Geometrie der Zahlen*. Chelsea, reprint, 1953.

J. Moody and C. J. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, (1):281 – 294, 1989.

E. A. Nadaraya. On estimating regression. *Theory Probab. Appl.*, pages 141–142, 1964.

R. M. Neal. *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics No. 118. New York, 1996.

P. Niyogi and F. Girosi. On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions. *Neural Computation*, 8:819–842, 1996.

C. J. Ong, S. Y. Shao, and J. B. Yang. An improved algorithm for the solution of the regularization path of support vector machine. *IEEE transactions on neural networks*, 21:451–462, 2010.

C. S. Ong, A. J. Smola, and R. C. Williamson. Learning the kernel with hyperkernels. *J. Mach. Learn. Res.*, 6:1043–1071, 2005. ISSN 1532-4435.

M. J. L. Orr. Regularised centre recruitment in radial basis function networks. *Centre for Cognitive Science, Edinburgh University*, 59, 1993.

M. J. L. Orr. Recent advances in radial basis function networks. Technical report, Technical Report www.ed.ac.uk/ mjo/papers/recad.ps, Institute for Adaptive and Neural Computation, 1999.

V. Pareto. *Manual d'économie politique (in French)*. F. Rouge, Lausanne, 1896.

J. Park and I. W. Sandberg. Universal approximation using radial-basis-function network. In *Neural Computation*, volume 3, pages 246–257, 1991.

M. Y. Park and T. Hastie. L1-regularization path algorithm for generalized linear models. *J. Royal Statistical Society*, 69(1):659–677, 2007.

E. Parzen. On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, (33):1065–1076, 1962.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 6(2):559–572, 1901.

T. Poggio and F. Girosi. Networks for approximation and learning. In *Proc. of the IEEE*, volume 78, pages 1481–1497, 1990.

M. J. D. Powell. Radial basis functions for multivariable interpolation: A review. In *IMA Conference on Algorithms for the Approximation of Functions and Data*, pages 143–167, 1985.

L. Prechelt. Proben1: A set of neural network benchmark problems and benchmarking rules. Technical Report 21/94, 1994.

J. R. Quinlan. Induction of decision trees. In *Machine Learning*, pages 81–106, 1986.

B. Scholkopf. *Advances in Kernel Methods: Support Vector Learning*. MIT Press, 1999.

B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

B. Scholkopf, R. Herbrich, A. J. Smola, and R. C. Williamson. A generalized representer theorem. technical report 2000-81, neurocolt, 2000. published in proceedings colt'2001. Technical report, 2001.

G. Schwarz. Estimating the dimension of a model. *Ann. Statistics*, 6:461–464, 1978.

C. E. Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

A. J. Shepherd. *Second-Order Methods for Neural Networks*. Springer-Verlag, London, 1997.

E. Snelson and Z. Ghahramani. Local and global sparse gaussian process approximations. *Journal of Machine Learning Research*, (2):524–531, 2007.

H. Sun. Mercer theorem for rkhs on noncompact sets. *J. Complex.*, 21(3):337–349, 2005. ISSN 0885-064X. doi: http://dx.doi.org/10.1016/j.jco.2004.09.002.

K. Tan, T. Lee, and E. Khor. Evolutionary algorithms for multi-objective optimization: Performance assessments and comparisons. *Artificial Intelligence Review*, 17(4):253–290, 2002.

R. A. Teixeira, A. P. Braga, R. H. C. Takahashi, and R. R. Saldanha. Improving generalization of MLPs with multi-objetive optimization. *Neurocomputing*, 35:189–194, 2000.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Statistical Society*, 58(1): 267–288, 1996.

A. N. Tikhonov. On the stability of inverse problems. *Dokl. Akad. Nauk SSSR*, 39(5):195–198, 1943.

A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet. Math. Dokl.*, 4:1035–1038, 1963.

V. Vapnik and O. Chapelle. Bounds on error expectation for support vector machines. *Neural Comput.*, 12(9):2013–2036, 2000. ISSN 0899-7667.

V. Vapnik and A. Y. Chervonenkis. The necessary and sufficient conditions for the consistency of the method of empirical risk minimization (in russian). *Yearbook of the Academy of Sciences of the USSR on Recognition, Classification and Forecasting*, (2):217–249, 1989.

V. N. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, Berlin, 1982.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of event to their probabilities. *Soviet. Math. Dokl.*, 9, 1968.

C. S. Wallace and D. M. Boulton. An information measure for classification. *Computer Jrnl*, 2(11): 185–194, 1968.

L. X. Wang. Fuzzy systems are universal approximators. In *Proc. 1-st IEEE Conf. on Fuzzy Systems*, pages 1163–1169, San Diego, 1992.

B. Widrow and S. D. Stearns. *Adaptive Signal Processing*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1985.

R. Williamson, A. Smola, and B. Scholkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *Information Theory, IEEE Transactions on*, 47(6):2516–2532, Sep 2001. ISSN 0018-9448. doi: 10.1109/18.945262.

H. Xu, C. Caramanis, and S. Mannor. Sparse algorithms are not stable: A no-free-lunch theorem. In *Forty-Sixth Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1299–1303, 2008.

L. Xu, A. Krzyzak, and A. Yuille. On radial basis function nets and kernel regression: Statistical consistency, convergence rates, and receptive field size. *Neural Networks*, 7(4):609 – 628, 1994. ISSN 0893-6080.

R. Yager and D. Filev. Generation of fuzzy rules by mountain clustering. *Journal of Intelligent & Fuzzy Systems*, 2(3):209 – 219, 1994.

G. G. Yen. Multi-objective evolutionary algorithm for radial basis function neural network design. In Y. Jin, editor, *Multi-Objective Machine Learning*, volume 16 of *Studies in Computational Intelligence*, pages 221–239. Springer, 2006.

H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the lasso. *Ann. of Statistics*, 35(5):2173–2192, 2007.