Multiple Kernel Learning: A Unifying Probabilistic Viewpoint

Hannes Nickisch   HANNES@NICKISCH.ORG
Max Planck Institute for Biological Cybernetics, Spemannstraße 38, 72076 Tübingen, Germany
Matthias Seeger   MATTHIAS.SEEGER@EPFL.CH
Ecole Polytechnique Fédérale de Lausanne, INJ 339, Station 14, 1015 Lausanne, Switzerland

Abstract

We present a probabilistic viewpoint to multiple kernel learning unifying well-known regularised risk approaches and recent advances in approximate Bayesian inference relaxations. The framework proposes a general objective function suitable for regression, robust regression and classification that is lower bound of the marginal likelihood and contains many regularised risk approaches as special cases. Furthermore, we derive an efficient and provably convergent optimisation algorithm.

Keywords: Multiple kernel learning, approximate Bayesian inference, double loop algorithms, Gaussian processes

1. Introduction

Nonparametric kernel methods, cornerstones of machine learning today, can be seen from different angles: as regularised risk minimisation in function spaces (Schölkopf and Smola, 2002), or as probabilistic Gaussian process methods (Rasmussen and Williams, 2006). In these techniques, the kernel (or equivalently covariance) function encodes interpolation characteristics from observed to unseen points, and two basic statistical problems have to be mastered. First, a latent function must be predicted which fits data well, yet is as smooth as possible given the fixed kernel. Second, the kernel function parameters have to be learned as well, to best support predictions which are of primary interest. While the first problem is simpler and has been addressed much more frequently so far, the central role of learning the covariance function is well acknowledged, and a substantial number of methods for "learning the kernel", "multiple kernel learning", or "evidence maximisation" are available now. However, much of this work has firmly been associated with one of the "camps" (referred to as *regularised risk* and *probabilistic* in the sequel) with surprisingly little crosstalk or acknowledgments of prior work across this boundary. In this paper, we clarify the relationship between major regularised risk and probabilistic kernel learning techniques precisely, pointing out advantages and pitfalls of either, as well as algorithmic similarities leading to novel powerful algorithms.

We develop a common analytical and algorithmical framework encompassing approaches from both camps and provide clear insights into the optimisation structure. Even though, most of the optimisation is non convex, we show how to operate a provably convergent "almost Newton" method nevertheless. Each step is not much more expensive than a gradient

based approach. Also, we do not require any foreign optimisation code to be available. Our framework unifies kernel learning for regression, robust regression and classification.

The paper is structured as follows: In section 2, we introduce the regularised risk and the probabilistic view of kernel learning. In increasing order of generality, we explain multiple kernel learning (MKL, section 2.1), maximum a posteriori estimation (MAP, section 2.2) and marginal likelihood maximisation (MLM, section 2.3). A taxonomy of the mutual relations between the approaches and important special cases is given in section 2.4. Section 3 introduces a general optimisation scheme and section 4 draws a conclusion.

2. Kernel Methods and Kernel Learning

Kernel-based algorithms come in many shapes, however, the primary goal is – based on training data $\{(\mathbf{x}_i, y_i) \mid i = 1..n\}$, $\mathbf{x}_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$ and a parametrised kernel function $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$ – to predict the output y_* for unseen inputs $\mathbf{x}_*$. Often, linear parametrisations $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^{M} \theta_m k_m(\mathbf{x}, \mathbf{x}')$ are used, where the k_m are fixed positive definite functions, and $\boldsymbol{\theta} \succeq \mathbf{0}$. Learning the kernel means finding $\boldsymbol{\theta}$ to best support this goal. In general, kernel methods employ a postulated latent function $u : \mathcal{X} \to \mathbb{R}$ whose smoothness is controlled via the function space squared norm $\|u(\cdot)\|_{k_{\boldsymbol{\theta}}}^2$. Most often, smoothness is traded against data fit, either enforced by a *loss function* $\ell(y_i, u(\mathbf{x}_i))$ or modeled by a *likelihood* $\mathbb{P}(y_i|u_i)$. Let us define kernel matrices $\mathbf{K}_{\boldsymbol{\theta}} := [k_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j)]_{ij}$, and $\mathbf{K}_m := [k_m(\mathbf{x}_i, \mathbf{x}_j)]_{ij}$ in $\mathbb{R}^{n \times n}$ and the vectors $\mathbf{y} := [y_i]_i \in \mathcal{Y}^n$, $\mathbf{u} := [u(\mathbf{x}_i)]_i \in \mathbb{R}^n$ collecting outputs and latent function values, respectively.

The *regularised risk* route to kernel prediction, which is followed by any support vector machine (SVM) or ridge regression technique, yields $\|u(\cdot)\|_{k_{\boldsymbol{\theta}}}^2 + \frac{C}{n} \sum_{i=1}^{n} \ell(y_i, u_i)$ as criterion, enforcing smoothness of $u(\cdot)$ as well as good data fit through the loss function $\frac{C}{n} \ell(y_i, u(\mathbf{x}_i))$. By the representer theorem, the minimiser can be written as $u(\cdot) = \sum_i \alpha_i k_{\boldsymbol{\theta}}(\cdot, \mathbf{x}_i)$, so that $\|u(\cdot)\|_{k_{\boldsymbol{\theta}}}^2 = \boldsymbol{\alpha}^\top \mathbf{K}_{\boldsymbol{\theta}} \boldsymbol{\alpha}$ (Schölkopf and Smola, 2002). As $\mathbf{u} = \mathbf{K}_{\boldsymbol{\theta}} \boldsymbol{\alpha}$, the regularised risk problem is given by

$$\min_{\mathbf{u}} \mathbf{u}^\top \mathbf{K}_{\boldsymbol{\theta}}^{-1} \mathbf{u} + \frac{C}{n} \sum_{i=1}^{n} \ell(y_i, u_i). \tag{1}$$

A *probabilistic* viewpoint of the same setting is based on the notion of a Gaussian process (GP) (Rasmussen and Williams, 2006): a Gaussian random function $u(\cdot)$ with mean function $\mathbb{E}[u(\mathbf{x})] = m(\mathbf{x}) \equiv 0$ and covariance function $\mathbb{V}[u(\mathbf{x}), u(\mathbf{x}')] = \mathbb{E}[u(\mathbf{x})u(\mathbf{x}')] = k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$. In practice, we only use finite-dimensional snapshots of the process $u(\cdot)$: for example, $\mathbb{P}(\mathbf{u}; \boldsymbol{\theta}) = \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{\boldsymbol{\theta}})$, a zero-mean joint Gaussian with covariance matrix $\mathbf{K}_{\boldsymbol{\theta}}$. We adopt this GP as prior distribution over $u(\cdot)$, estimating the latent function as maximum of the posterior process $\mathbb{P}(u(\cdot)|\mathbf{y}; \boldsymbol{\theta}) \propto \mathbb{P}(\mathbf{y}|\mathbf{u})\mathbb{P}(u(\cdot); \boldsymbol{\theta})$. Since the likelihood depends on $u(\cdot)$ only through the finite subset $\{u(\mathbf{x}_i)\}$, the posterior process has a finite-dimensional representation: $\mathbb{P}(u(\cdot)|\mathbf{y}, \mathbf{u}) = \mathbb{P}(u(\cdot)|\mathbf{u})$, so that $\mathbb{P}(u(\cdot)|\mathbf{y}; \boldsymbol{\theta}) = \int \mathbb{P}(u(\cdot)|\mathbf{u})\mathbb{P}(\mathbf{u}|\mathbf{y}; \boldsymbol{\theta})d\mathbf{u}$ is specified by the joint distribution $\mathbb{P}(\mathbf{u}|\mathbf{y}; \boldsymbol{\theta})$, a probabilistic equivalent of the representer theorem. Kernel prediction amounts to *maximum a posteriori* (MAP) estimation

$$\max_{\mathbf{u}} \mathbb{P}(\mathbf{u}|\mathbf{y}; \boldsymbol{\theta}) \equiv \max_{\mathbf{u}} \mathbb{P}(\mathbf{u}; \boldsymbol{\theta})\mathbb{P}(\mathbf{y}|\mathbf{u}) \equiv \min_{\mathbf{u}} \mathbf{u}^\top \mathbf{K}_{\boldsymbol{\theta}}^{-1} \mathbf{u} - 2\ln \mathbb{P}(\mathbf{y}|\mathbf{u}) + \ln |\mathbf{K}_{\boldsymbol{\theta}}|, \tag{2}$$

ignoring an additive constant. Minimising equations (1) and (2) for any fixed kernel matrix $\mathbf{K}$ gives the same minimiser $\hat{\mathbf{u}}$ and prediction $u(\mathbf{x}_*) = \hat{\mathbf{u}}^\top \mathbf{K}_{\boldsymbol{\theta}}^{-1} [k_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_*)]_i$.

$\mathcal{Y}$	Loss function	$\ell(y_i, u_i)$	$\mathbb{P}(y_i\|u_i)$	Likelihood
$\{\pm 1\}$	SVM Hinge loss	$\max(0, 1 - y_i u_i)$	$\nexists$	
$\{\pm 1\}$	Log loss	$\ln(\exp(-y_i u_i) + 1)$	$1/(\exp(-\tau y_i u_i) + 1)$	Logistic
$\mathbb{R}$	SVM ϵ-insensitive loss	$\max(0, \|y_i - u_i\|/\epsilon - 1)$	$\nexists$	
$\mathbb{R}$	Quadratic loss	$(y_i - u_i)^2$	$\mathcal{N}(y_i\|u_i, \sigma^2)$	Gaussian
$\mathbb{R}$	Linear loss	$\|y_i - u_i\|$	$\mathcal{L}(y_i\|u_i, \tau)$	Laplace

Table 1: Relations between loss functions and likelihoods

The correspondence between likelihood and loss function bridges probabilistic and regularised risk techniques. More specifically, any likelihood $\mathbb{P}(\mathbf{y}|\mathbf{u})$ induces a loss function $\ell(\mathbf{y}, \mathbf{u})$ via

$$-2 \ln \mathbb{P}(\mathbf{y}|\mathbf{u}) = -2 \sum_i \ln \mathbb{P}(y_i|u_i) \rightsquigarrow \frac{C}{n} \sum_{i=1}^{n} \ell(y_i, u_i) = \ell(\mathbf{y}, \mathbf{u}),$$

however some loss functions cannot be interpreted as a negative log likelihood as shown in table (1) and as discussed for the SVM by Sollich (2000). If, the likelihood is a *log-concave* function of $\mathbf{u}$, it corresponds to a convex loss function (Boyd and Vandenberghe, 2002, Sect. 3.5.1). Common loss functions and likelihoods for classification $\mathcal{Y} = \{\pm 1\}$ and regression $\mathcal{Y} = \mathbb{R}$ are listed in table (1).

In the following, we discuss several approaches to learn the kernel parameters $\boldsymbol{\theta}$ and show how all of them can be understood as instances of or approximations to Bayesian evidence maximisation. Although the exposition MKL section 2.1 and MAP section 2.2 use a linear parametrisation $\boldsymbol{\theta} \mapsto \mathbf{K}_{\boldsymbol{\theta}} = \sum_{m=1}^{M} \theta_m \mathbf{K}_m$, much of the results in MLM 2.3 and all the aforementioned discussion are still applicable to non-linear parametrisations.

2.1 Multiple Kernel Learning

A widely adopted regularised risk principle, known as *multiple kernel learning* (MKL) (Christianini et al., 2001; Lanckriet et al., 2004; Bach et al., 2004), is to minimise equation (1) w.r.t. the kernel parameters $\boldsymbol{\theta}$ as well. One obvious caveat is that for any fixed $\mathbf{u}$, equation (1) becomes ever smaller as $\theta_m \to \infty$: it cannot per se play a meaningful statistical role. In order to prevent this, researchers constrain the domain of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and obtain

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \min_{\mathbf{u}} \mathbf{u}^{\top} \mathbf{K}_{\boldsymbol{\theta}}^{-1} \mathbf{u} + \ell(\mathbf{y}, \mathbf{u}),$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\theta} \succeq \mathbf{0}, \|\boldsymbol{\theta}\|_2 \leq 1\}$ or $\boldsymbol{\Theta} = \{\boldsymbol{\theta} \succeq \mathbf{0}, \mathbf{1}^{\top} \boldsymbol{\theta} \leq 1\}$ (Varma and Ray, 2007). Notably, these constraints are imposed independently of the statistical problem, the model and of the parametrization $\boldsymbol{\theta} \mapsto \mathbf{K}_{\boldsymbol{\theta}}$. The Lagrangian form of the MKL problem with parameter λ and a general p-norm unit ball constraint where $p \geq 1$ (Kloft et al., 2009) is given by

$$\min_{\boldsymbol{\theta} \succeq \mathbf{0}} \phi_{\text{MKL}}(\boldsymbol{\theta}), \quad \text{where } \phi_{\text{MKL}}(\boldsymbol{\theta}) := \min_{\mathbf{u}} \mathbf{u}^{\top} \mathbf{K}_{\boldsymbol{\theta}}^{-1} \mathbf{u} + \ell(\mathbf{y}, \mathbf{u}) + \underbrace{\lambda \cdot \mathbf{1}^{\top} \boldsymbol{\theta}^p}_{\rho(\boldsymbol{\theta})}, \ \lambda > 0. \quad (3)$$

Since, the *regulariser* $\rho(\boldsymbol{\theta})$ for the kernel parameter $\boldsymbol{\theta}$ is convex, the map $(\mathbf{u}, \mathbf{K}) \mapsto \mathbf{u}^{\top} \mathbf{K}^{-1} \mathbf{u}$ is jointly convex for $\mathbf{K} \succeq \mathbf{0}$ (Boyd and Vandenberghe, 2002) and the parametrisation $\boldsymbol{\theta} \mapsto \mathbf{K}_{\boldsymbol{\theta}}$ is linear, MKL is a jointly convex problem for $\boldsymbol{\theta} \succeq \mathbf{0}$ whenever the loss function
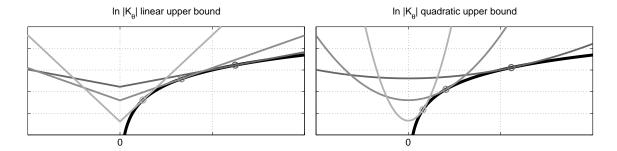
Figure 1: Convex upper bounds on (the concave non-decreasing) $\ln |\mathbf{K}_{\boldsymbol{\theta}}|$

By Fenchel duality, we can represent any concave non-decreasing function and hence the log determinant function by $\ln |\mathbf{K}_{\boldsymbol{\theta}}| = \min_{\boldsymbol{\lambda} \succeq \mathbf{0}} \boldsymbol{\lambda}^{\top} |\boldsymbol{\theta}|^p - g^*(\boldsymbol{\lambda})$. As a consequence, we obtain a piecewise polynomial upper bound for any particular value $\boldsymbol{\lambda}$.

$\ell(\mathbf{y}, \mathbf{u})$ is convex. Furthermore, there are efficient algorithms to solve equation (3) for large models (Sonnenburg et al., 2006).

2.2 Joint MAP Estimation

Adopting a probabilistic MAP viewpoint, we can minimise equation (2) w.r.t. $\mathbf{u}$ and $\boldsymbol{\theta} \succeq \mathbf{0}$:

$$\min_{\boldsymbol{\theta} \succeq \mathbf{0}} \phi_{\mathrm{MAP}}(\boldsymbol{\theta}), \quad \text{where } \phi_{\mathrm{MAP}}(\boldsymbol{\theta}) := \min_{\mathbf{u}} \mathbf{u}^{\top} \mathbf{K}_{\boldsymbol{\theta}}^{-1} \mathbf{u} - 2 \ln \mathbb{P}(\mathbf{y}|\mathbf{u}) + \ln |\mathbf{K}_{\boldsymbol{\theta}}|. \quad (4)$$

While equation (3) and equation (4) share the "inner solution" $\hat{\mathbf{u}}$ for fixed $\mathbf{K}_{\boldsymbol{\theta}}$ – in case the loss $\ell(\mathbf{y}, \mathbf{u})$ corresponds to a likelihood $\mathbb{P}(\mathbf{y}|\mathbf{u})$ – they are different when it comes to optimising $\boldsymbol{\theta}$. The *joint MAP* problem is not in general jointly convex in $(\boldsymbol{\theta}, \mathbf{u})$, since $\boldsymbol{\theta} \mapsto \ln |\mathbf{K}_{\boldsymbol{\theta}}|$ is concave, see figure 2. However, it is always a well-posed statistical procedure, since $\ln |\mathbf{K}_{\boldsymbol{\theta}}| \to \infty$ as $\theta_m \to \infty$ for all m.

We show in the following, how the regularisers $\rho(\boldsymbol{\theta}) = \lambda \|\boldsymbol{\theta}\|_p^p$ of equation (3) can be related to the probabilistic term $f(\boldsymbol{\theta}) = \ln |\mathbf{K}_{\boldsymbol{\theta}}|$. In fact, the same reasoning can be applied to any concave non-decreasing function.

Since the function $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}) = \ln |\mathbf{K}_{\boldsymbol{\theta}}|$, $\boldsymbol{\theta} \succeq \mathbf{0}$ is jointly concave, we can represent it by $f(\boldsymbol{\theta}) = \min_{\boldsymbol{\lambda} \succeq \mathbf{0}} \boldsymbol{\lambda}^{\top} \boldsymbol{\theta} - f^*(\boldsymbol{\lambda})$ where $f^*(\boldsymbol{\lambda})$ denotes Fenchel dual of $f(\boldsymbol{\theta})$. Furthermore, the mapping $\boldsymbol{\vartheta} \mapsto \ln |\sum_{m=1}^{M} \sqrt[p]{\vartheta_m} \mathbf{K}_m| = f(\sqrt[p]{\boldsymbol{\vartheta}}) = g(\boldsymbol{\vartheta})$, $\boldsymbol{\vartheta} \succeq \mathbf{0}$ is jointly concave due to the composition rule (Boyd and Vandenberghe, 2002, §3.2.4), because $\boldsymbol{\vartheta} \mapsto \sqrt[p]{\boldsymbol{\vartheta}}$ is jointly concave and $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta})$ is non-decreasing in all components θ_m as all matrices $\mathbf{K}_m$ are positive (semi-)definite which guarantees that the eigenvalues of $\mathbf{K}_{\boldsymbol{\theta}}$ increase as θ_m increases. Thus we can – similarly to Zhang (2010) – represent $\ln |\mathbf{K}_{\boldsymbol{\theta}}|$ as

$$\ln |\mathbf{K}_{\boldsymbol{\theta}}| = f(\boldsymbol{\theta}) = g(\boldsymbol{\vartheta}) = \min_{\boldsymbol{\lambda} \succeq \mathbf{0}} \boldsymbol{\lambda}^{\top} \boldsymbol{\vartheta} - g^*(\boldsymbol{\lambda}) = \min_{\boldsymbol{\lambda} \succeq \mathbf{0}} \boldsymbol{\lambda}^{\top} |\boldsymbol{\theta}|^p - g^*(\boldsymbol{\lambda}).$$

Choosing a particular value $\boldsymbol{\lambda} = \lambda \cdot \mathbf{1}$, we obtain the bound $\ln |\mathbf{K}_{\boldsymbol{\theta}}| \leq \lambda \cdot \|\boldsymbol{\theta}\|_p^p - g^*(\lambda \cdot \mathbf{1})$. Figure 1 illustrates the bounds for $p = 1$ and $p = 2$. The bottom line is that one can interpret the regularisers $\rho(\boldsymbol{\theta}) = \lambda \|\boldsymbol{\theta}\|_p^p$ in equation (3) as corresponding to parametrised upper bounds to the $\ln |\mathbf{K}_{\boldsymbol{\theta}}|$ part in equation (4), hence $\phi_{\mathrm{MKL}}(\boldsymbol{\theta}) = \psi_{\mathrm{MAP}}(\boldsymbol{\theta}, \boldsymbol{\lambda} = \lambda \cdot \mathbf{1})$, where $\phi_{\mathrm{MAP}}(\boldsymbol{\theta}) = \min_{\boldsymbol{\lambda} \succeq \mathbf{0}} \psi_{\mathrm{MAP}}(\boldsymbol{\theta}, \boldsymbol{\lambda})$. Far from an ad hoc choice to keep $\boldsymbol{\theta}$ small, the

$\ln |\mathbf{K}_{\boldsymbol{\theta}}|$ term embodies the Occam's razor concept behind MAP estimation: overly large $\boldsymbol{\theta}$ are ruled out, since their explanation of the data $\mathbf{y}$ is extremely unlikely under the prior $\mathbb{P}(\mathbf{u}; \boldsymbol{\theta})$. The Occam's razor effect depends crucially on the proper normalization of the prior (MacKay, 1992). For example, the weighting parameter C of k ($k = C\tilde{k}$) can be learned by joint MAP: if $C = e^c$, then equation (4) is convex in c for any fixed $\mathbf{u}$. If kernel-regularised estimation equation (1) is interpreted as MAP estimation under a GP prior equation (2), the correct extension to kernel learning is joint MAP: the MKL criterion equation (3) lacks prior normalization, which renders MAP w.r.t. $\boldsymbol{\theta}$ meaningful in the first place. From a non-probabilistic viewpoint, the $\ln |\mathbf{K}_{\boldsymbol{\theta}}|$ term comes with a model and data dependent structure at least as complex as the rest of equation (3).

While the MKL objective, equation (3), enjoys the benefit of being convex in the (linear) kernel parameters $\boldsymbol{\theta}$, this does not hold true for joint MAP estimation, equation (4), in general. We illustrate the differences in figure 2. The function $\psi_{\text{MAP}}(\boldsymbol{\theta}, \mathbf{u})$ is a building block of the MAP objective $\phi_{\text{MAP}}(\boldsymbol{\theta}) = \min_{\mathbf{u}}[\psi_{\text{MAP}}(\boldsymbol{\theta}, \mathbf{u}) - 2 \ln \mathbb{P}(\mathbf{y}|\mathbf{u})]$, where

$$\psi_{\text{MAP}}(\boldsymbol{\theta}, \mathbf{u}) = \underbrace{\mathbf{u}^{\top}\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{u}}_{\psi_{\cup}(\boldsymbol{\theta}, \mathbf{u})} + \underbrace{\ln |\mathbf{K}_{\boldsymbol{\theta}}|}_{\psi_{\cap}(\boldsymbol{\theta})} \leq \psi_{\text{MKL}}(\boldsymbol{\theta}, \mathbf{u}) - g^{*}(\lambda \cdot \mathbf{1}), \quad \psi_{\text{MKL}}(\boldsymbol{\theta}, \mathbf{u}) = \mathbf{u}^{\top}\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{u} + \lambda \|\boldsymbol{\theta}\|_{p}^{p}.$$

More concretely, $\psi_{\text{MAP}}(\boldsymbol{\theta}, \mathbf{u})$ is a sum of a nonnegative, jointly convex function $\psi_{\cup}(\boldsymbol{\theta}, \mathbf{u})$ that is strictly decreasing in every component θ_m and a concave function $\psi_{\cap}(\boldsymbol{\theta})$ that is strictly increasing in every component θ_m. Both functions $\psi_{\cup}(\boldsymbol{\theta}, \mathbf{u})$ and $\psi_{\cap}(\boldsymbol{\theta})$ alone do not have a stationary point due to their monotonicity in θ_m. However, their sum can have (even multiple) stationary points as shown in figure 2 on the left left. We can show, that the map $\mathbf{K} \mapsto \mathbf{u}^{\top}\mathbf{K}^{-1}\mathbf{u} + \ln |\mathbf{K}|$ is *invex* i.e. every stationary point $\hat{\mathbf{K}}$ is a global minimum. Using the convexity of $\mathbf{A} \mapsto \mathbf{u}^{\top}\mathbf{A}\mathbf{u} - \ln |\mathbf{A}|$ (Boyd and Vandenberghe, 2002) and the fact that the derivative of $\mathbf{A} \mapsto \mathbf{A}^{-1}$ for $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{A} \succ \mathbf{0}$ has full rank n^2, we see by Mishra and Giorgi (2008, theorem 2.1) that $\mathbf{K} \mapsto \mathbf{u}^{\top}\mathbf{K}^{-1}\mathbf{u} + \ln |\mathbf{K}|$ is indeed invex.

Often, the MKL objective for the case $p = 1$ is motivated by the fact that the optimal solution $\boldsymbol{\theta}^{\star}$ is *sparse* (e.g. Sonnenburg et al., 2006), meaning that many components θ_m are zero. Figure 2 illustrates that $\phi_{\text{MAP}}(\boldsymbol{\theta})$ also yields sparse solutions; in fact it enforces even more sparsity. In MKL, $\phi_{\text{MAP}}(\boldsymbol{\theta})$ is simply relaxed to a convex objective $\phi_{\text{MKL}}(\boldsymbol{\theta})$ at the expense of having only a single less sparse solution.

Intuition for the Gaussian Case

We can gain further intuition about the criteria ϕ_{MKL} and ϕ_{MAP} by asking which *matrices* $\mathbf{K}$ minimise them. For simplicity, assume that $\mathbb{P}(\mathbf{y}|\mathbf{u}) = \mathcal{N}(\mathbf{y}|\mathbf{u}, \sigma^2 \mathbf{I})$ and $n/C = \sigma^2$, hence $\ell(\mathbf{y}, \mathbf{u}) = \frac{1}{\sigma^2} \|\mathbf{y} - \mathbf{u}\|_2^2$. The inner minimiser $\hat{\mathbf{u}}$ for both ϕ_{MKL} and ϕ_{MAP} is given by $\mathbf{K}_{\boldsymbol{\theta}}^{-1}\hat{\mathbf{u}} = (\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2 \mathbf{I})^{-1}\mathbf{y}$. With $\sigma^2 \to 0$, we find for joint MAP that $\frac{\partial}{\partial \mathbf{K}}\phi_{\text{MAP}} = \mathbf{0}$ results in $\hat{\mathbf{K}} = \mathbf{y}\mathbf{y}^{\top}$. While this "nonparametric" estimate requires smoothing to be useful in practice, closeness to $\mathbf{y}\mathbf{y}^{\top}$ is fundamental to covariance estimation and can be found in regularised risk kernel learning work (Christianini et al., 2001). On the other hand, for $\text{tr}(\mathbf{K}_m) = 1$ and hence $\rho(\boldsymbol{\theta}) = \lambda \text{tr}(\mathbf{K}_{\boldsymbol{\theta}}) = \lambda \|\boldsymbol{\theta}\|_1$, $\frac{\partial}{\partial \mathbf{K}}\phi_{\text{MKL}} = \mathbf{0}$ leads to $\hat{\mathbf{K}}^2 = \lambda^{-1}\mathbf{y}\mathbf{y}^{\top}$: an odd way of estimating covariance, not supported by any statistical literature we are aware of.
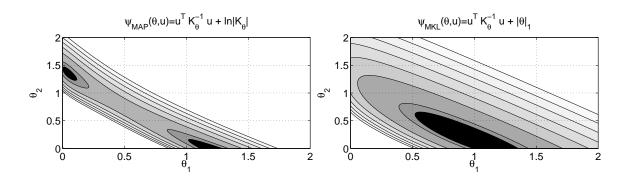
Figure 2: Convex and non-convex building blocks of the MKL and MAP objective function

2.3 Marginal Likelihood Maximisation

While the joint MAP criterion uses a properly normalised prior distribution, it is still not probabilistically consistent. Kernel learning amounts to finding a value $\hat{\boldsymbol{\theta}}$ of high data likelihood, no matter what the latent function $u(\cdot)$ is. The correct likelihood to be maximised is *marginal*: $\mathbb{P}(\mathbf{y}|\boldsymbol{\theta}) = \int \mathbb{P}(\mathbf{y}|\mathbf{u})\mathbb{P}(\mathbf{u}|\boldsymbol{\theta})\mathrm{d}\mathbf{u}$ ("max-sum"), while joint MAP employs the plug-in surrogate $\max_{\mathbf{u}} \mathbb{P}(\mathbf{y}|\mathbf{u})\mathbb{P}(\mathbf{u}|\boldsymbol{\theta})$ ("max-max"). *Marginal likelihood maximisation* (MLM) is also known as Bayesian estimation, and it underlies the EM algorithm or maximum likelihood learning of conditional random fields just as well: complexity is controlled (and overfitting avoided) by averaging over unobserved variables $\mathbf{u}$ (MacKay, 1992), rather than plugging in some point estimate $\hat{\mathbf{u}}$

$$\phi_{\mathrm{MLM}}(\boldsymbol{\theta}) := -2\ln \int \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{\boldsymbol{\theta}})\mathbb{P}(\mathbf{y}|\mathbf{u})\mathrm{d}\mathbf{u}. \tag{5}$$

The Gaussian Case

Before developing a general MLM approximation, we note an important analytically solvable exception: for Gaussian likelihood $\mathbb{P}(\mathbf{y}|\mathbf{u}) = \mathcal{N}(\mathbf{y}|\mathbf{u}, \sigma^2\mathbf{I})$, $\mathbb{P}(\mathbf{y}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_{\boldsymbol{\theta}} + \sigma^2\mathbf{I})$, and MLM becomes

$$\phi_{\mathrm{GAU}}(\boldsymbol{\theta}) := \mathbf{y}^{\top}(\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2\mathbf{I})^{-1}\mathbf{y} + \ln|\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2\mathbf{I}|. \tag{6}$$

Even if the primary purpose is classification, the Gaussian likelihood is used for its analytical simplicity (Kapoor et al., 2009). Only for the Gaussian case, joint MAP and MLM have an analytically closed form. From the product formula of Gaussians (Brookes, 2005, §5.1)

$$\mathbb{Q}(\mathbf{u}) := \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{\boldsymbol{\theta}})\mathcal{N}(\mathbf{y}|\mathbf{u}, \boldsymbol{\Gamma}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_{\boldsymbol{\theta}} + \boldsymbol{\Gamma})\mathcal{N}(\mathbf{u}|\mathbf{m}, \mathbf{V}),$$

where $\mathbf{V} = (\mathbf{K}_{\boldsymbol{\theta}}^{-1} + \boldsymbol{\Gamma}^{-1})^{-1}$ and $\mathbf{m} = \mathbf{V}\boldsymbol{\Gamma}^{-1}\mathbf{y}$ we can deduce that

$$-2\ln \int \mathbb{Q}(\mathbf{u})\mathrm{d}\mathbf{u} = \ln|\mathbf{K}_{\boldsymbol{\theta}}^{-1} + \boldsymbol{\Gamma}^{-1}| + \min_{\mathbf{u}}[-2\ln \mathbb{Q}(\mathbf{u})] - n\ln|2\pi|. \tag{7}$$

Using $\sigma^2\mathbf{I} = \boldsymbol{\Gamma}$ and $\min_{\mathbf{u}}[-2\ln \mathbb{Q}(\mathbf{u})] = -2\ln \mathbb{Q}(\mathbf{m})$, we see that by

$$\phi_{\mathrm{MAP/GAU}}(\boldsymbol{\theta}) :\overset{c}{=} \phi_{\mathrm{GAU}}(\boldsymbol{\theta}) - \ln|\mathbf{K}_{\boldsymbol{\theta}}^{-1} + \sigma^{-2}\mathbf{I}| \overset{c}{=} \mathbf{y}^{\top}(\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2\mathbf{I})^{-1}\mathbf{y} + \ln|\mathbf{K}_{\boldsymbol{\theta}}| \tag{8}$$

MLM and MAP are very similar for the Gaussian case.

The "ridge regression" approximation is also used together with p-norm constraints instead of the $\ln|\mathbf{K}_{\boldsymbol{\theta}}|$ term (Cortes et al., 2009)

$$\phi_{\mathrm{RR}}(\boldsymbol{\theta}) := \mathbf{y}^{\top}(\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2\mathbf{I})^{-1}\mathbf{y} + \lambda\|\boldsymbol{\theta}\|_p^p. \tag{9}$$

Unfortunately, most GP methods to date work with a Gaussian likelihood for simplicity, a restriction which often proves short-sighted. Gaussian-linear models come with unrealistic properties, and benefits of MLM over joint MAP cannot be realised.

Kernel parameter learning has been an integral part of probabilistic GP methods from the very beginning. Williams and Rasmussen (1996) proposed MLM for Gaussian noise equation 6, fifteen years ago. They treated sums of exponential and linear kernels as well as learning lengthscales (ARD), predating recent proposals such as "products of kernels" (Varma and Babu, 2009).

THE GENERAL CASE

In general, joint MAP always has the analytical form equation 4, while $\mathbb{P}(\mathbf{y}|\boldsymbol{\theta})$ can only be approximated. For non-Gaussian $\mathbb{P}(\mathbf{y}|\mathbf{u})$, numerous approximate inference methods have been proposed, specifically motivated by learning kernel parameters via MLM. The simplest such method is Laplace's approximation, applied to GP binary and multi-way classification by Williams and Barber (1998): starting with convex joint MAP, $\ln\mathbb{P}(\mathbf{y},\mathbf{u})$ is expanded to second order around the posterior mode $\hat{\mathbf{u}}$. More recent approximations can be much more accurate, yet come with non-convex problems and less robust algorithms (Nickisch and Rasmussen, 2008). In this paper, we concentrate on the variational lower bound relaxation (VB) by Jaakkola and Jordan (2000), which is convex for log-concave likelihoods $\mathbb{P}(\mathbf{y}|\mathbf{u})$ (Nickisch and Seeger, 2009), providing a novel simple and efficient algorithm. While our VB approximation to MLM is more expensive to run than joint MAP for non-Gaussian likelihood (even using Laplace's approximation), the implementation complexity of our VB algorithm is comparable to what is required in the Gaussian noise case equation 6.

More, specifically, we exploit that super-Gaussian of likelihoods $\mathbb{P}(y_i|u_i)$ can be lower bounded by scaled Gaussians $\mathcal{N}_{\gamma_i}$ of any width γ_i:

$$\mathbb{P}(y_i|u_i) = \max_{\gamma_i>0}\mathcal{N}_{\gamma_i} = \max_{\gamma_i>0}\exp\left(\beta_i u_i - \frac{u_i^2}{2\gamma_i} - \frac{1}{2}h_i(\gamma_i)\right),$$

where $\beta_i \propto y_i$ are constants, and $h_i(\cdot)$ is convex (Nickisch and Seeger, 2009) whenever the likelihood $\mathbb{P}(y_i|u_i)$ is log-concave. If the posterior distribution is $\mathbb{P}(\mathbf{u}|\mathbf{y}) = Z^{-1}\mathbb{P}(\mathbf{y}|\mathbf{u})\mathbb{P}(\mathbf{u})$, then $\ln Z \geq Ce^{-\psi_{\mathrm{VB}}(\boldsymbol{\theta},\boldsymbol{\gamma})/2}$ by plugging in these bounds, where C is a constant and

$$\phi_{\mathrm{VB}}(\boldsymbol{\theta}) := \min_{\boldsymbol{\gamma}\succ\mathbf{0}}\psi_{\mathrm{VB}}(\boldsymbol{\theta},\boldsymbol{\gamma}), \quad \psi_{\mathrm{VB}}(\boldsymbol{\theta},\boldsymbol{\gamma}) := h(\boldsymbol{\gamma}) - 2\ln\int\mathcal{N}(\mathbf{u}|\mathbf{0},\mathbf{K}_{\boldsymbol{\theta}})e^{\mathbf{u}^{\top}(\boldsymbol{\beta}-\frac{1}{2}\boldsymbol{\Gamma}^{-1}\mathbf{u})}\mathrm{d}\mathbf{u}, \tag{10}$$

$h(\boldsymbol{\gamma}) := \sum_i h_i(\gamma_i)$, $\boldsymbol{\Gamma} := \mathrm{dg}(\boldsymbol{\gamma})$. The variational relaxation[1] amounts to maximising the lower bound, which means that $\mathbb{P}(\mathbf{u}|\mathbf{y})$ is fitted by the *Gaussian* approximation $\mathbb{Q}(\mathbf{u}|\mathbf{y};\boldsymbol{\gamma})$ with covariance matrix $\mathbf{V} = (\mathbf{K}_{\boldsymbol{\theta}}^{-1}+\boldsymbol{\Gamma}^{-1})^{-1}$ (Nickisch and Seeger, 2009). Alternatively, we can interpret $\psi_{\mathrm{VB}}(\boldsymbol{\theta},\boldsymbol{\gamma})$ as an upper bound on the Kullback-Leibler divergence $\mathrm{KL}(\mathbb{Q}(\mathbf{u}|\mathbf{y};\boldsymbol{\gamma})||\mathbb{P}(\mathbf{u}|\mathbf{y}))$

---

1. Generalisations to other super-Gaussian potentials (log-concave or not) or models including linear couplings and mixed potentials are given by Nickisch and Seeger (2009).

(Nickisch, 2010, §2.5.9), a measure for the dissimilarity between the exact posterior $\mathbb{P}(\mathbf{u}|\mathbf{y})$ and the parametrised Gaussian approximation $\mathbb{Q}(\mathbf{u}|\mathbf{y}; \boldsymbol{\gamma})$.

Finally, note that by equation (7), $\psi_{\mathrm{VB}}(\boldsymbol{\theta}, \boldsymbol{\gamma})$ can also be written as

$$\psi_{\mathrm{VB}}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \ln|\mathbf{K}_{\boldsymbol{\theta}}^{-1} + \boldsymbol{\Gamma}^{-1}| + h(\boldsymbol{\gamma}) + \min_{\mathbf{u}} R(\mathbf{u}, \boldsymbol{\theta}, \boldsymbol{\gamma}) + \ln|\mathbf{K}_{\boldsymbol{\theta}}|, \quad (11)$$

where $R(\mathbf{u}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \mathbf{u}^{\top}(\mathbf{K}_{\boldsymbol{\theta}}^{-1} + \boldsymbol{\Gamma}^{-1})\mathbf{u} - 2\boldsymbol{\beta}^{\top}\mathbf{u}$. Using the concavity of $\boldsymbol{\gamma}^{-1} \mapsto \ln|\mathbf{K}_{\boldsymbol{\theta}}^{-1} + \boldsymbol{\Gamma}^{-1}|$ and Fenchel duality $\ln|\mathbf{K}_{\boldsymbol{\theta}}^{-1} + \boldsymbol{\Gamma}^{-1}| = \min_{\mathbf{z} \succ \mathbf{0}} \mathbf{z}^{\top} \boldsymbol{\gamma}^{-1} - g_{\boldsymbol{\theta}}^{*}(\mathbf{z}) = \hat{\mathbf{z}}_{\boldsymbol{\theta}}^{\top} \boldsymbol{\gamma}^{-1} - g_{\boldsymbol{\theta}}^{*}(\hat{\mathbf{z}}_{\boldsymbol{\theta}})$, with the optimal value $\hat{\mathbf{z}}_{\boldsymbol{\theta}} = \mathrm{dg}(\mathbf{V})$, we can reformulate $\psi_{\mathrm{VB}}(\boldsymbol{\theta}, \boldsymbol{\gamma})$ as

$$\psi_{\mathrm{VB}}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \min_{\mathbf{z} \succ \mathbf{0}}[\mathbf{z}^{\top} \boldsymbol{\gamma}^{-1} - g_{\boldsymbol{\theta}}^{*}(\mathbf{z})] + h(\boldsymbol{\gamma}) + \min_{\mathbf{u}} R(\mathbf{u}, \boldsymbol{\theta}, \boldsymbol{\gamma}) + \ln|\mathbf{K}_{\boldsymbol{\theta}}|,$$

which allows to perform the minimisation w.r.t. $\boldsymbol{\gamma}$ in closed form (Nickisch, 2010, §3.5.6):

$$\phi_{\mathrm{VB}}(\boldsymbol{\theta}) = \min_{\mathbf{z} \succ \mathbf{0}} \psi_{\mathrm{VB}}(\boldsymbol{\theta}, \mathbf{z}), \quad \psi_{\mathrm{VB}}(\boldsymbol{\theta}, \mathbf{z}) = \min_{\mathbf{u}} \mathbf{u}^{\top} \mathbf{K}_{\boldsymbol{\theta}}^{-1} \mathbf{u} + \tilde{\ell}_{\mathbf{z}}(\mathbf{y}, \mathbf{u}) - g_{\boldsymbol{\theta}}^{*}(\mathbf{z}) + \ln|\mathbf{K}_{\boldsymbol{\theta}}|, \quad (12)$$

where $\tilde{\ell}_{\mathbf{z}}(\mathbf{y}, \mathbf{u}) := 2\boldsymbol{\beta}^{\top}(\mathbf{v} - \mathbf{u}) - 2\ln\mathbb{P}(\mathbf{y}|\mathbf{v})$ and finally $\mathbf{v} = \mathrm{sign}(\mathbf{u}) \odot \sqrt{\mathbf{u}^2 + \mathbf{z}}$. Note that for $\mathbf{z} = \mathbf{0}$, we exactly recover joint MAP estimation, equation (4), as $\mathbf{z} = \mathbf{0}$ implies $\mathbf{u} = \mathbf{v}$ and $\tilde{\ell}_{\mathbf{z}}(\mathbf{y}, \mathbf{u}) = \ell(\mathbf{y}, \mathbf{u})$. For fixed $\boldsymbol{\theta}$, the optimal value $\hat{\mathbf{z}}_{\boldsymbol{\theta}} = \mathrm{dg}(\mathbf{V})$ corresponds to the marginal variances of the Gaussian approximation $\mathbb{Q}(\mathbf{u}|\mathbf{y}; \boldsymbol{\gamma})$: Variational inference corresponds to variance-smoothed joint MAP estimation (Nickisch, 2010) with a loss function $\tilde{\ell}(\mathbf{y}, \mathbf{u}, \boldsymbol{\theta})$ that explicitly depends on the kernel parameters $\boldsymbol{\theta}$. We have two equivalent representations of the loss $\tilde{\ell}(\mathbf{y}, \mathbf{u}, \boldsymbol{\theta})$ that directly follow from equations (11) and (12):

$$
\begin{aligned}
\tilde{\ell}(\mathbf{y}, \mathbf{u}, \boldsymbol{\theta}) &= \min_{\boldsymbol{\gamma} \succ \mathbf{0}}[\ln|\mathbf{K}_{\boldsymbol{\theta}}^{-1} + \boldsymbol{\Gamma}^{-1}| + h(\boldsymbol{\gamma}) + \mathbf{u}^{\top}\boldsymbol{\Gamma}^{-1}\mathbf{u} - 2\boldsymbol{\beta}^{\top}\mathbf{u}], \text{ and} \\
\tilde{\ell}(\mathbf{y}, \mathbf{u}, \boldsymbol{\theta}) &= \min_{\mathbf{z} \succ \mathbf{0}}[2\boldsymbol{\beta}^{\top}(\mathbf{v} - \mathbf{u}) - 2\ln\mathbb{P}(\mathbf{y}|\mathbf{v}) - g_{\boldsymbol{\theta}}^{*}(\mathbf{z})], \quad \mathbf{v} = \mathrm{sign}(\mathbf{u}) \odot \sqrt{\mathbf{u}^2 + \mathbf{z}}.
\end{aligned}
$$

Our VB problem is $\min_{\boldsymbol{\theta} \succeq \mathbf{0}, \boldsymbol{\gamma} \succ \mathbf{0}} \psi_{\mathrm{VB}}(\boldsymbol{\theta}, \boldsymbol{\gamma})$ or equivalently $\min_{\boldsymbol{\theta} \succeq \mathbf{0}, \mathbf{z} \succ \mathbf{0}} \psi_{\mathrm{VB}}(\boldsymbol{\theta}, \mathbf{z})$. The inner variables here are $\boldsymbol{\gamma}$ and $\mathbf{z}$, in addition to $\mathbf{u}$ in joint MAP. There are further similarities: since $\psi_{\mathrm{VB}}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = -2\ln\int e^{-R(\mathbf{u}, \boldsymbol{\gamma}, \boldsymbol{\theta})}\mathrm{d}\mathbf{u} + h(\boldsymbol{\gamma}) + \ln|2\pi\mathbf{K}_{\boldsymbol{\theta}}|$, $(\boldsymbol{\gamma}, \boldsymbol{\theta}) \mapsto \psi_{\mathrm{VB}} - \ln|\mathbf{K}_{\boldsymbol{\theta}}|$ is jointly convex for $\boldsymbol{\gamma} \succ \mathbf{0}$, $\boldsymbol{\theta} \succeq \mathbf{0}$, by the joint convexity of $(\mathbf{u}, \boldsymbol{\gamma}, \boldsymbol{\theta}) \mapsto R$ and Prékopa's theorem (Boyd and Vandenberghe, 2002, §3.5.2). Joint MAP and VB share the same convexity structure. In contrast, approximating $\mathbb{P}(\mathbf{y}|\boldsymbol{\theta})$ by other techniques like Expectation Propagation (Minka, 2001) or general Variational Bayes (Opper and Archambeau, 2009) does not even constitute convex problems for fixed $\boldsymbol{\theta}$.

2.4 Summary and Taxonomy

In the last paragraphs, we have detailed how a variety of kernel learning approaches can be obtained from Bayesian marginal likelihood maximisation in a sequence of nested upper bounding steps. Table 2 nicely illustrates how many kernel learning objectives are related to each other – either by upper bounds or by Gaussianity assumptions. We can clearly see, that $\phi_{\mathrm{VB}}(\boldsymbol{\theta})$ – as an upper bound to the negative log marginal likelihood – can be seen as the mother function. For a special case, $\mathbf{z} = \mathbf{0}$, we obtain joint maximum a posteriori

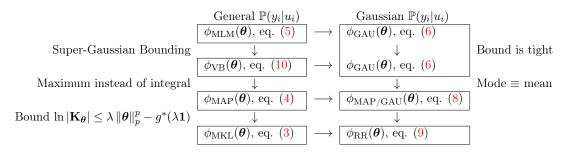| Name	Objective function
Marginal Likelihood Maximisation	$\phi_{\mathrm{MLM}}(\boldsymbol{\theta}) = -2\ln\left[\int \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{\boldsymbol{\theta}})\mathbb{P}(\mathbf{y}|\mathbf{u})\mathrm{d}\mathbf{u}\right]$
Variational Bounds	$\phi_{\mathrm{VB}}(\boldsymbol{\theta}) = \min_{\boldsymbol{\gamma}\succ\mathbf{0}} \psi_{\mathrm{VB}}(\boldsymbol{\theta}, \boldsymbol{\gamma}) \geq \phi_{\mathrm{MLM}}(\boldsymbol{\theta})$ by $\mathbb{P}(y_i|u_i) \geq \mathcal{N}_{\gamma_i}$
Maximum A Posteriori	$\phi_{\mathrm{MAP}}(\boldsymbol{\theta}) = -2\ln\left[\max_{\mathbf{u}} \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{\boldsymbol{\theta}})\mathbb{P}(\mathbf{y}|\mathbf{u})\right] = \psi_{\mathrm{VB}}(\boldsymbol{\theta}, \mathbf{z}=\mathbf{0})$
Multiple Kernel Learning	$\phi_{\mathrm{MKL}}(\boldsymbol{\theta}) = \phi_{\mathrm{MAP}}(\boldsymbol{\theta}) + \lambda\,\|\boldsymbol{\theta}\|_p^p - \ln|\mathbf{K}_{\boldsymbol{\theta}}| = \psi_{\mathrm{MAP}}(\boldsymbol{\theta}, \boldsymbol{\lambda}=\lambda\cdot\mathbf{1})$

Table 2: Taxonomy of kernel learning objective functions

The upper table visualises the relationship between several kernel learning objective functions for arbitrary likelihood/loss functions: Marginal likelihood maximisation (MLM) can be bounded by variational bounds (VB) and maximum a posteriori estimation (MAP) is a special case $\mathbf{z} = \mathbf{0}$ thereof. Finally multiple kernel learning (MKL) can be understood as an upper bound to the MAP estimation objective $\boldsymbol{\lambda} = \lambda\cdot\mathbf{1}$. The lower table complements the upper table by also covering the analytically important Gaussian case.

estimation, where the loss functions does not depend on the kernel parameters. Going further, a particular instance $\boldsymbol{\lambda} = \lambda\cdot\mathbf{1}$ yields the widely use multiple kernel learning objective that becomes convex in the kernel parameters $\boldsymbol{\theta}$. In the following, we will concentrate on the optimisation and computational similarities between the approaches.

3. Algorithms

In this section, we derive a simple, provably convergent and efficient algorithm for MKL, joint MAP and VB. We use the Lagrangian form of equation (3) and $\ell(\mathbf{y}, \mathbf{u}) := -2\ln\mathbb{P}(\mathbf{y}|\mathbf{u})$:

$$\psi_{\mathrm{MKL}}(\boldsymbol{\theta}, \mathbf{u}) = \mathbf{u}^\top\mathbf{K}^{-1}\mathbf{u} + \ell(\mathbf{y}, \mathbf{u}) + \lambda\cdot\mathbf{1}^\top\boldsymbol{\theta}, \ \lambda > 0,$$
$$\psi_{\mathrm{MAP}}(\boldsymbol{\theta}, \mathbf{u}) = \mathbf{u}^\top\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{u} + \ell(\mathbf{y}, \mathbf{u}) + \ln|\mathbf{K}_{\boldsymbol{\theta}}|, \quad \text{and}$$
$$\psi_{\mathrm{VB}}(\boldsymbol{\theta}, \mathbf{u}) = \mathbf{u}^\top\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{u} + \min_{\mathbf{z}\succ\mathbf{0}}\left[\ell(\mathbf{y}, \mathbf{v}) + 2\boldsymbol{\beta}^\top(\mathbf{v}-\mathbf{u}) - g_{\boldsymbol{\theta}}^*(\mathbf{z})\right] + \ln|\mathbf{K}_{\boldsymbol{\theta}}|,$$

$$\text{where } \mathbf{v} = \mathrm{sign}(\mathbf{u}) \odot \sqrt{\mathbf{u}^2 + \mathbf{z}}.$$

Many previous algorithms use alternating minimization, which is easy to implement but tends to converge slowly. Both ϕ_{VB} and ϕ_{MAP} are jointly convex up to the concave $\boldsymbol{\theta} \mapsto \ln|\mathbf{K}_{\boldsymbol{\theta}}|$ part. Since $\ln|\mathbf{K}_{\boldsymbol{\theta}}| = \min_{\boldsymbol{\lambda}\succ\mathbf{0}} \boldsymbol{\lambda}^\top\boldsymbol{\theta} - f^*(\boldsymbol{\lambda})$ (Legendre duality, Boyd and Vandenberghe, 2002), joint MAP becomes $\min_{\boldsymbol{\lambda}\succ\mathbf{0}, \boldsymbol{\theta}\succeq\mathbf{0}, \mathbf{u}} \phi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}, \mathbf{u})$ with $\phi_{\boldsymbol{\lambda}} := \mathbf{u}^\top\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{u} + \ell(\mathbf{y}, \mathbf{u}) + \boldsymbol{\lambda}^\top\boldsymbol{\theta} - f^*(\boldsymbol{\lambda})$ which is jointly convex in $(\boldsymbol{\theta}, \mathbf{u})$. Algorithm 1 iterates between refits of $\boldsymbol{\lambda}$ and joint Newton updates of $(\boldsymbol{\theta}, \mathbf{u})$.

The Newton direction costs $O(n^3 + M\,n^2)$, with n the number of data points and M the number of base kernels. All algorithms discussed in this paper require $O(n^3)$ time, apart

---

Algorithm 1 Double loop algorithm for joint MAP, MKL and VB.

Require: Criterion $\psi_{\#}(\boldsymbol{\theta}, \mathbf{u}) = \tilde{\psi}_{\#}(\boldsymbol{\theta}, \mathbf{u}) + \ln|\mathbf{K}_{\boldsymbol{\theta}}|$ to minimise for $(\mathbf{u}, \boldsymbol{\theta}) \in \mathbb{R}^n \times \mathbb{R}_+^M$.

  repeat

    Newton $\min_{\mathbf{u}} \psi_{\#}$ for fixed $\boldsymbol{\theta}$ (optional; few steps).

    Refit upper bound: $\boldsymbol{\lambda} \leftarrow \nabla_{\boldsymbol{\theta}} \ln|\mathbf{K}_{\boldsymbol{\theta}}| = [\mathrm{tr}(\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{K}_1), .., \mathrm{tr}(\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{K}_M)]^{\top}$.

    Compute joint Newton search direction $\mathbf{d}$ for $\psi_{\boldsymbol{\lambda}} := \tilde{\psi}_{\#} + \boldsymbol{\lambda}^{\top}\boldsymbol{\theta}$:

        $\nabla^2_{[\boldsymbol{\theta};\mathbf{u}]}\psi_{\boldsymbol{\lambda}}\mathbf{d} = -\nabla_{[\boldsymbol{\theta};\mathbf{u}]}\psi_{\boldsymbol{\lambda}}$.

    Linesearch: Minimise $\psi_{\#}(\alpha)$ i.e. $\psi_{\#}(\boldsymbol{\theta}, \mathbf{u})$ along $[\boldsymbol{\theta}; \mathbf{u}] + \alpha\mathbf{d}$, $\alpha > 0$.

  until Outer loop converged

---

from the requirement to store the base matrices $\mathbf{K}_m$. The convergence proof hinges on the fact that ϕ and $\phi_{\boldsymbol{\lambda}}$ are tangentially equal (Nickisch and Seeger, 2009). Equivalently, the algorithm can be understood as Newton's method, yet dropping the part of the Hessian corresponding to the $\ln|\mathbf{K}|$ term (note that $\nabla_{(\mathbf{u},\boldsymbol{\theta})}\phi_{\boldsymbol{\lambda}} = \nabla_{(\mathbf{u},\boldsymbol{\theta})}\phi$ for the Newton direction computation). Exact Newton for MKL.

In practice, we use $\mathbf{K}_{\boldsymbol{\theta}} = \sum_m \theta_m \mathbf{K}_m + \varepsilon\mathbf{I}$, $\varepsilon = 10^{-8}$ to avoid numerical problems when computing $\boldsymbol{\lambda}$ and $\ln|\mathbf{K}_{\boldsymbol{\theta}}|$. We also have to enforce $\boldsymbol{\theta} \succeq \mathbf{0}$ in algorithm 1, which is done by the barrier method (Boyd and Vandenberghe, 2002). We minimise $t\phi + \mathbf{1}^{\top}(\ln\boldsymbol{\theta})$ instead of ϕ, increasing $t > 0$ every few outer loop iterations.

A variant algorithm 1 can be used to solve VB in a different parametrisation ($\boldsymbol{\gamma} \succ \mathbf{0}$ replaces $\mathbf{u}$), which has the same convexity structure as joint MAP. Transforming equation (10) similarly to equation (6), we obtain

$$\phi_{\mathrm{VB}}(\boldsymbol{\theta}) = \min_{\boldsymbol{\gamma} \succ \mathbf{0}} \ln|\mathbf{C}| - \ln|\boldsymbol{\Gamma}| + \boldsymbol{\beta}^{\top}\boldsymbol{\Gamma}\mathbf{C}^{-1}\boldsymbol{\Gamma}\boldsymbol{\beta} - \boldsymbol{\beta}^{\top}\boldsymbol{\Gamma}\boldsymbol{\beta} + h(\boldsymbol{\gamma}) \tag{13}$$

with $\mathbf{C} := \mathbf{K}_{\boldsymbol{\theta}} + \boldsymbol{\Gamma}$, computed using the Cholesky factorisation $\mathbf{C} = \mathbf{L}\mathbf{L}^{\top}$. They cost $O(M n^3)$ to compute, which is more expensive than for joint MAP or MKL. Note that the cost $O(M n^3)$ is not specific to our particular relaxation or algorithm e.g. the Laplace MLM approximation (Williams and Barber, 1998), solved using gradients w.r.t. $\boldsymbol{\theta}$ only, comes with the same complexity.

4. Conclusion

We presented a unifying probabilistic viewpoint to multiple kernel learning that derives regularised risk approaches as special cases of approximate Bayesian inference. We provided an efficient and provably convergent optimisation algorithm suitable for regression, robust regression and classification.

Our taxonomy of multiple kernel learning approaches connected many previously only loosely related ideas and provided insights into the common structure of the respective optimisation problems. Finally, we proposed an algorithm to solve the latter efficiently.

References

Francis Bach, Gert Lanckriet, and Michael Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *ICML*, 2004.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2002.

Mike Brookes. The matrix reference manual, 2005. URL http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro.html.

Nello Christianini, John Shawe-Taylor, André Elisseeff, and Jaz Kandola. On kernel-target alignment. In *NIPS*, 2001.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. L2 regularization for learning kernels. In *UAI*, 2009.

Tomi Jaakkola and Michael Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37, 2000.

Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Gaussian processes for object categorization. *IJCV*, 2009. doi: 10.1007/s11263-009-0268-3.

Marius Kloft, Ulf Brefeld, Sören Sonnenburg, Pavel Laskov, Klaus-Robert Müller, and Alexander Zien. Efficient and accurate lp-norm multiple kernel learning. In *NIPS*, 2009.

Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michal I. Jordan. Learning the kernel matrix with semidefinite programming. *JMLR*, 5:27–72, 2004.

Davic MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

Tom Minka. Expectation propagation for approximate Bayesian inference. In *Uncertainty in AI*, 2001.

Shashi Kant Mishra and Giorgio Giorgi. *Invexity and optimization*. Springer, 2008.

Hannes Nickisch. *Bayesian Inference and Experimental Design for Large Generalised Linear Models*. PhD thesis, TU Berlin, 2010.

Hannes Nickisch and Carl Edward Rasmussen. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078, 2008.

Hannes Nickisch and Matthias Seeger. Convex variational Bayesian inference for large scale generalized linear models. In *ICML*, 2009.

Manfred Opper and Cédric Archambeau. The variational Gaussian approximation revisited. *Neural Computation*, 21(3):786–792, 2009.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Bernhard Schölkopf and Alex Smola. *Learning with Kernels*. MIT Press, 1st edition, 2002.

Peter Sollich. Probabilistic methods for support vector machines. In *NIPS*, pages 349–355, 2000.

Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *JMLR*, 7:1531–1565, 2006.

Manik Varma and Bodla Rakesh Babu. More generality in efficient multiple kernel learning. In *ICML*, 2009.

Manik Varma and Debajyoti Ray. Learning the discriminative power-invariance trade-off. In *ICCV*, 2007.

Christopher K. I. Williams and David Barber. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.

Christopher K. I. Williams and Carl Edward Rasmussen. Gaussian processes for regression. In *Advances in Neural Information Processing Systems 8*, 1996.

Tong Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11:1081–1107, 2010.