Sparse Multiple Kernel Learning: Support Identification via Mirror Stratifiability

Guillaume Garrigos
École Normale Supérieure, CNRS
Paris, France
guillaume.garrigos@ens.fr

Lorenzo Rosasco
Università degli Studi di Genova,
Massachusetts Institute of Technology,
Istituto Italiano di Tecnologia,
Genova, Italy
lrosasco@mit.edu

Silvia Villa
Dipartimento di Matematica
Politecnico di Milano
Milano, Italy
silvia.villa@polimi.it

Abstract—In statistical machine learning, kernel methods allow to consider infinite dimensional feature spaces with a computational cost that only depends on the number of observations. This is usually done by solving an optimization problem depending on a data term and a suitable regularizer. In this paper we consider feature maps which are the concatenation of a fixed, possibly large, set of simpler feature maps. The penalty is a sparsity inducing one, promoting solutions depending only on a small subset of the features. The group lasso problem is a sparsity inducing one, promoting solutions depending only on a fixed, possibly large, set of simpler feature maps. The penalty to consider infinite dimensional feature spaces with a computer-

I. INTRODUCTION

Kernel methods provide a practical and elegant way to derive nonparametric approaches to supervised learning [13]. In this context, the goal is to determine a function $f*: \mathbb{R}^d \rightarrow \mathbb{R}$ observing a finite training set of data points \{(x_1, y_1), \ldots, (x_m, y_m)\} \subset \mathbb{R}^d \times \mathbb{R}$. One of the most popular approaches to find estimates of $f*$ relies on the solution of an empirical minimization problem

$$
\min_{f \in \mathcal{F}} \lambda R(f) + \frac{1}{2m} \sum_{i=1}^{m} (f(x_i) - y_i)^2, \quad \lambda > 0,
$$

where $(\mathcal{F}, \langle \cdot , \cdot \rangle)$ is an appropriate Hilbert space of real valued functions on $\mathbb{R}^d$ and $R: \mathcal{F} \rightarrow \mathbb{R}$ a regularization functional. The choice of the space $\mathcal{F}$ and the regularization $R$ is crucial, since it allows to incorporate a priori information on the problem [15]. In kernel methods, $\mathcal{F}$ is realized through a (nonlinear) feature map $\Phi: \mathbb{R}^d \rightarrow H$, for a separable Hilbert space $H$, in the sense that

$$
(\forall f \in \mathcal{F})(\forall x \in \mathbb{R}^d) \quad f(x) = \langle w, \Phi(x) \rangle_H,
$$

for some $w \in H$. A positive definite kernel can then be defined by setting $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_H$. In this paper, we focus on a class of structured feature spaces, that can be written as direct sums of a finite number of spaces, involving thus different feature maps. This choice is known as multiple kernel learning [2], [9], [10].

Sparse Multiple Kernel Learning. We consider $H_1, \ldots, H_G$ a finite number of separable Hilbert spaces and, for every $i \in \{1, \ldots, G\}$, we let $\Phi_i: \mathbb{R}^d \rightarrow H_i$ be a map. Then, we consider the space of functions defined by the concatenated feature map

$$
\Phi: \mathbb{R}^d \rightarrow H_1 \times \ldots \times H_G, \quad \Phi = (\Phi_1, \ldots, \Phi_G).
$$

The supervised learning problem in $\mathcal{F}$ is called multiple kernel learning. In the context of multiple kernel learning, $f*$ is assumed to have a sparse representation in the feature space $H := H_1 \times \ldots \times H_G$, namely

$$
(\forall x \in \mathbb{R}^d) \quad f*(x) = \sum_{g=1}^{G} (w_g^*) \Phi_g(x)
$$

with $(w_g^*) = 0$ in $H_g$ for several $g$'s. To approximate the function $f*$, the idea is to devise a regularization term inducing sparsity in the sum representations [2], [9], [10]. This leads to the minimization problem with respect to $w = (w_1, \ldots, w_G) \in H$

$$
\arg\min_{w \in H} \lambda \sum_{g=1}^{G} \|w_g\|_{H_g} + \frac{1}{2m} \sum_{i=1}^{m} \sum_{g=1}^{G} (w_g \Phi_g(x_i))_{H_g} - y_i)^2.
$$

(MKL)

The MKL approach has several possible motivations. First, it allows the use of richer hypotheses spaces, and therefore allows in principle for better solutions to be found. Indeed, we can think of it as a way to combine different feature maps/kernels. Moreover, multiple kernel learning can also be used as an alternative to model selection. By creating a concatenation of feature maps, model selection is replaced by
an aggregation procedure. Finally, multiple kernel learning is useful for data fusion. For example, different features/kernels could capture different characteristics of the data, and a better description is obtained by combining them. In the regime where we consider a large number of feature maps, the identification of the active features is crucial for good results. The special case of MKL, where each feature map is finite dimensional, is equivalent to the so called group lasso.

**Example I.1** (Group lasso). Partition $\mathbb{R}^d$ into non overlapping groups of variables in $\mathbb{R}^{d_i}$ for $i \in [1, \ldots, G]$. Set $H_i = \mathbb{R}^{d_i}$ and let $\Phi_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$ be the canonical projection. Then $H = \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_G}$, and problem (MKL) becomes

$$\arg\min_{w \in \mathbb{R}^d} \lambda \sum_{g=1}^G \|w_g\|_{\mathbb{R}^{d_i}} + \frac{1}{2m} \sum_{i=1}^m \left(\langle w, x_i \rangle - y_i \right)^2.$$  

(GL)

This is a well studied problem, and is called group lasso [16]. It is of interest when the vector $w$ has a structured sparsity pattern. More precisely, the considered penalty induces block-wise sparse vectors, where all the components belonging to a given group are zero.

**The Iterative Kernel Thresholding Algorithm.** From a computational point of view, as observed in [11], problem (MKL) can be solved using proximal splitting methods [3]: the objective function structure is amenable for the forward backward splitting algorithm, since it is the sum of a differentiable term (the square loss) with Lipschitz continuous gradient, and a convex continuous regularizer. Though this algorithm is defined on a possibly infinite-dimensional space $H$, a generalization of the representer theorem allows to compute its iterates as a finite linear combination of the features [11], as discussed next.

For every $g \in [1, \ldots, G]$, let

$$X_g : H_g \to \mathbb{R}^m, \quad w_g \mapsto (\langle \Phi_g(x_i), w_g \rangle)_{i=1}^m,$$

and

$$K_g := X_g X_g^* = (\langle \Phi(x_i), \Phi(x_{i'}) \rangle)_{i,i'}, \quad K = [K_1, \ldots, K_G].$$

Then, the Iterative Kernel Thresholding Algorithm can be written as

$$\begin{align*}
\alpha^0 & \in \mathbb{R}^{m \times G}, \quad \tau \in [0, 2\|K\|^{-1}], \\
\alpha^{n+1}_g & = \tilde{g}_{\lambda, \tau, g} (\alpha^n_g - \tau (K \alpha^n - y)), \\
w^{n+1}_g & = X_g^* \alpha^{n+1}_g,
\end{align*}$$

(IKTA)

where $\tilde{g}_{\lambda, \tau, g}$ denotes a group-thresholding operator defined by:

$$\tilde{g}_{\lambda, \tau, g}(\alpha_g) :=
\begin{cases}
0 & \text{if } \|X_g^* \alpha_g\|_{H_g} \leq \lambda, \\
\left(1 - \frac{\lambda}{\|X_g^* \alpha_g\|_{H_g}}\right) \alpha_g & \text{otherwise}.
\end{cases}$$

In the rest of the paper we show that the proposed algorithm produces sparse iterates, and more notably is able to identify the relevant features of the solutions of the regularized empirical risk functional if a suitable qualification condition holds. This is a crucial and desirable property in sparse regularization, since the exact solutions of the regularized problems will be never computed in closed form and the sparsity pattern gives key information on the sought model. In case the qualification condition does not hold, we show that the support of the iterates, though in general bigger than the ideal one, can still be controlled.

**II. MAIN RESULTS**

As a preliminary result, we state the asymptotic convergence of IKTA towards a solution of (MKL).

**Proposition II.1.** Let $(w^n)_{n \in \mathbb{N}}$ be a sequence generated by IKTA. Then there exists $\bar{w} \in H$, a solution of (MKL), such that $\|w^n - \bar{w}\| \to 0$ as $n \to +\infty$.

In our first main result, we look at the exact support recovery property of IKTA. We recall that the (group) support of a vector $w \in H = H_1 \times \ldots \times H_G$ is given by

$$\text{supp}(w) := \{g \in [1 \ldots G] : w_g \neq 0\}.$$  

We show that, after a finite number of iterations, the algorithm identifies exactly the support of $\bar{w}$, provided the following qualification condition is verified:

$$\max_{g \notin \text{supp}(\bar{w})} \frac{1}{\lambda} \|X^*_g (X \bar{w} - y)\|_{H_g} < 1, \quad (QC)$$

where $X : H \to \mathbb{R}^m$ is given by $Xw = \sum_{g=1}^G X_g w_g$. This condition is also known as the "irrepresentable condition" [11, 17], or the "strict sparsity pattern" [5].

**Theorem II.2** (Exact support recovery). Let $(w^n)_{n \in \mathbb{N}}$ be a sequence generated by IKTA, converging to $\bar{w}$. If (QC) holds, then there exists $N \in \mathbb{N}$ such that

$$\forall n \geq N \quad \text{supp}(w^n) = \text{supp}(\bar{w}).$$

The condition (QC) is necessary for Theorem II.2, in the sense that the support of $\bar{w}$ cannot be identified in general after a finite number of iterations. Without (QC), it could be that the algorithm generates sequences identifying a strictly larger support. This is illustrated in the following example:

**Example II.3.** Let $G = 1, H_1 = \mathbb{R}, X_1 = 1, y_1 = 1, \lambda = 1$. Then the corresponding problem (MKL) has a unique solution, $\bar{w} = 0$, for which (QC) is clearly not satisfied. Take now $\alpha^0 = w^0 \in [0, +\infty]$, $\tau \in [0, 1]$, and consider the sequence $(w^n)_{n \in \mathbb{N}}$ generated by IKTA. It is not difficult to show that this sequence verifies $w^{n+1} = (1 - \tau) w^n$, and therefore will never satisfy $\text{supp}(w^n) = \text{supp}(\bar{w})$.

In our second main result, we tackle the case in which (QC) is not verified, and show that the support of the sequence can nevertheless be controlled. For this, one needs to look at the extended support of the solution, which corresponds to the set of active constraints of the dual problem of (MKL):

$$\text{esupp}(\bar{w}) := \{g \in [1, \ldots, G] : \|X_g^*(X \bar{w} - y)\|_{H_g} = \lambda\}.$$

**Theorem II.4** (Extended support control). Let $(w^n)_{n \in \mathbb{N}}$ be a sequence generated by IKTA, converging to $\bar{w}$. Then there exists $N \in \mathbb{N}$ such that

$$\forall n \geq N \quad \text{supp}(\bar{w}) \subset \text{supp}(w^n) \subset \text{esupp}(\bar{w}).$$
Relation with previous works: When the the groups have dimension 1 (i.e. $H_g \equiv \mathbb{R}$), the problem boils down to a simple $\ell^1$ regression problem. In that case, exact support recovery in presence of (QC) is obtained in [5], and extended support control in [7]. When the groups have finite dimension (i.e. $\dim H_g < +\infty$), exact support recovery under (QC) is known, see for instance [1], or [8], which contains results for the large class of partially smooth functions. Still in finite dimension, extended support control has been recently obtained in [6], where the authors introduce the class of mirror-stratifiable functions.

To our knowledge, Theorems II.2 and II.4 are the first to treat the support recovery/control of an algorithm for solving a multiple kernel learning problem involving possibly infinite-dimensional kernels. Our theoretical results rely on an extension of [6] to the separable Hilbert space setting. In particular, we underline that these results build on the concept of stratification, and can thus be extended to the large class of mirror-stratifiable regularizers (see section IV-B).

III. Numerical experiments

In this section we illustrate our main results, by considering different variants of (MKL). For each experiment, we draw data points $(x_i)_{i=1}^{\text{data points}} \subset \mathbb{R}^p$ at random with i.i.d. entries from a zero-mean standard Gaussian distribution. The Hilbert space of functions we consider is either the space of linear functionals or the space generated by combinations of Gaussian kernels. Then, we take $\alpha_\ast \in \mathbb{R}^{m \times G}$ randomly among the vectors verifying $|\text{supp}(\alpha_\ast)| = s$, and take $y = K\alpha_\ast + \varepsilon$, where $\varepsilon$ is a zero-mean white Gaussian noise with standard deviation $10^{-2}$. Here $\text{supp}(\alpha_\ast)$ must be understood as the group-support of $\alpha_\ast \in \mathbb{R}^{m \times G} = \mathbb{R}^m \times \ldots \times \mathbb{R}^m$.

We consider two problems, for which we take the following parameters $(m, G, s, \lambda) = (50, 20, 5, 0.2)$. The first is the group-lasso setting, where we chose $p = 100$, $H_g \equiv \mathbb{R}^5$ and $\Phi_g : \mathbb{R}^{100} \rightarrow H_g$ to be the canonical projection on the $g$-th group of coordinates. The second is the Gaussian-Kernel setting, where we choose $p = 2$, and each space $H_g$ to be an infinite-dimensional RKHS associated to the Gaussian kernel

$$k(x, x') = e^{-\frac{||x-e_{x'}||^2}{2\sigma_g^2}},$$

with $(\sigma_g)_{g=1}^{G}$ being taken at random in $[10^{-1}, 10^1]$. For each problem, we draw 200 instances of $((x_i, y_i)_{i=1}^{\text{data points}}, \alpha_\ast)$, and solve the corresponding (MKL) problem with IKTA. The step size for the algorithm is $\tau = 0.8||K||^{-1}$, and we stop the algorithm after $N = 5000$ (resp. $N = 50000$) iterations for the group-lasso (resp. Gaussian-kernel).

Description of the results. In Figure 1, we show the distribution of $|\text{supp}(w^N)|$ over the 200 instances of the algorithm. We see that for the group-lasso, the method recovers most of the time a support of size 5-7, and that the probability of finding another support decays quickly. For the Gaussian-kernel, there are a significant number of instances where $|\text{supp}(w^N)| < 5$ and the probability of finding a smaller support decays slowly, while the probability of finding a large support decays as quickly as in the linear case. In Figure 2, we illustrate the evolution of the support along the iterations of IKTA.

Discussion on the results. An important point to observe is the fact that the algorithm often fails at recovering exactly the desired support. This means that qualification conditions which are often assumed in the literature to guarantee the stability of the support cannot be taken for granted in practice. In the light of Theorem II.4, the results of Figure 1 seem to suggest that IKTA often fails at recovering the support of $\tilde{w}$.

But it is important to keep in mind that another phenomenon enters into play here: $\tilde{w}$ and $w_\ast$ does not always share the same support, even for small $\lambda$. This is particularly clear when we look at the iterates whose support is smaller than the one of $w_\ast$: it would contradict Theorem II.4 if $\text{supp}(w_\ast) = \text{supp}(\tilde{w})$. This fact is well-known in the finite dimensional setting, when no qualification condition holds at $w_\ast$, see e.g. [6]. To have a finer understanding of our results, it would be interesting to have a result like [1, Theorem 2], or [6, Theorem 3], adapted to our infinite dimensional setting.

IV. Proofs

For this section we need to introduce some notation. Given $\Omega \subset H$, $\text{cl} \Omega$, $\text{int} \Omega$ and $\text{qri} \Omega$ will denote respectively the (norm-)closure, interior and quasirelative interior [3, Definition 6.9] of $\Omega$. The closed unit ball, and the corresponding unit sphere of a Hilbert space $H$ are noted $B_{H_g}$ and $S_{H_g}$. The set of all proper lower semi-continuous convex functions on $H$ are denoted $\Gamma_0(H)$. For $R \in \Gamma_0(H)$, its Fenchel conjugate [3, Definition 13.1] is denoted $R^\ast$. Finally, for a vector in the product of Hilbert spaces $w \in H = H_1 \times \ldots \times H_G$, we will often use $||w||$ instead of $||w||_{H_g}$, if there is no ambiguity.
A. Proof of Proposition II.1

Let \( D : \mathbb{R}^{n \times G} \to H \) be the linear operator defined by \( D\alpha = (X^{g} \alpha_{g})_{g \in [1..G]} \). Also define \( y = (y; \ldots; y) \in \mathbb{R}^{n \times G} \), \( K = (K; \ldots; K) \in \mathbb{R}^{(m \times G)^{2}} \), and \( \hat{G}_{\lambda \tau} = (\hat{G}_{\lambda \tau}(\alpha_{g}))_{g} \). Then, IKTA can be rewritten as

\[
w^{n+1} = D \hat{G}_{\lambda \tau} (\alpha - \tau (K \alpha - y))
\]

It can also be verified that

\[
D \circ \hat{G}_{\lambda \tau} = \hat{G}_{\lambda \tau} \circ D, \quad D \mathbf{K} = X^{*} X \mathbf{D} \quad \text{and} \quad D \mathbf{y} = X^{*} y
\]

holds, where \( \hat{G}_{\lambda \tau} \) is the group thresholding operator:

\[
(\forall g \in [1..G]) \ (\hat{G}_{\lambda}(\alpha))_{g} := \begin{cases} 0 & \text{if } \| \alpha_{g} \| \leq \lambda, \\ \left( 1 - \frac{\lambda}{\| \alpha_{g} \|} \right) \alpha_{g} & \text{otherwise}. \end{cases}
\]

We see then that the iterates of IKTA verify:

\[
w^{n+1} = \hat{G}_{\lambda \tau} \circ D (\alpha - \tau (K \alpha - y)), \quad \hat{G}_{\lambda \tau} (w^{n} - \tau X^{*} (Xw^{n} - y)).
\]

We recognize here the iterates of the Forward-Backward algorithm applied to \( R + h \), where

\[
R(w) := \sum_{g=1}^{G} \| w_{g} \|, \quad h(w) := \| Xw - y \|^{2}/(2\lambda).
\]

Since \( R \) is coercive and \( h \) is bounded from below, we know that (MKL) admits solutions, so we deduce from [3, Corollary 28.9] that the sequence converges weakly towards a solution of (MKL). Since \((w^{n})_{n \in \mathbb{N}} \) belongs to \( \mathbb{N} \), which has finite dimension, this weak convergence actually turns out to be strong.

B. Identification for mirror-stratifiable functions

To prove Theorems II.2 and II.4, we adapt to our Hilbert setting the tools of stratification (of a set) and mirror-stratifiability (of a function) introduced in [6].

We say that \( \mathcal{M} = \{ M_{i} \}_{\iota \in \iota} \subset 2^{H} \) is a stratification of a set \( \Omega \subset H \), if it is a finite partition of \( \Omega \) verifying, for any pair of strata \( M \) and \( M' \) in \( \mathcal{M} \):

\[
M \cap \text{cl} M' \neq \emptyset \Rightarrow M \subset \text{cl} M'.
\]

Since it is a partition, we note \( M_{w} \in \mathcal{M} \) the unique strata containing a vector \( w \in \Omega \). It is important to note that the property (3) endows \( \mathcal{M} \) with a partial order \( \leq \), given by:

\[
M \leq M' \Leftrightarrow M \subset \text{cl} M'.
\]

We are now in position to define the mirror-stratifiability of a convex function \( R \in \Gamma_{0}(H) \). For this, we will need to associate to \( R \) a correspondence operator between subsets of \( \text{dom} \partial R \) and \( \text{dom} \partial R^{*} \):

\[
(\forall S \subset \text{dom} \partial R) \quad J_{R}(S) := \bigcup \{ \text{ qr } \partial R(w) \}_{w \in S}
\]

Definition IV.1. Let \( R \in \Gamma_{0}(H) \), and let \( \mathcal{M} \) (resp. \( \mathcal{M}^{*} \)) be a stratification of \( \text{dom} \partial R \) (resp. \( \text{dom} \partial R^{*} \)). We say that \( R \) is mirror-stratifiable (w.r.t. \( \mathcal{M} \) and \( \mathcal{M}^{*} \)) if \( J_{R} \) realizes a bijection between \( \mathcal{M} \) and \( \mathcal{M}^{*} \) (with inverse \( J_{R^{-1}} \)), and is decreasing:

\[
(\forall (M, M') \in \mathcal{M}^{2}) \quad M \leq M' \Leftrightarrow J_{R}(M') \leq J_{R}(M).
\]

Our main result for mirror-stratifiable functions controls the stability of their stratas up to perturbations on the graph of \( \partial R \):

Proposition IV.2. Let \( R \in \Gamma_{0}(H) \) be mirror-stratifiable w.r.t. \( (\mathcal{M}, \mathcal{M}^{*}) \), and consider a converging sequence on the graph of \( \partial R \):

\[
\eta_{n} \in \partial R(w^{n}), \quad w^{n} \xrightarrow{n \to \infty} \bar{w}, \quad \eta^{n} \xrightarrow{n \to \infty} \bar{\eta}.
\]

Then there exists \( N \in \mathbb{N} \) such that

\[
(\forall n \geq N) \quad M_{\bar{w}} \leq M_{w^{n}} \leq J_{R^{*}}(M_{\bar{\eta}}).
\]

Proof. Reasoning exactly as in [6, Proposition 1], we obtain the existence of some \( \delta > 0 \) such that

\[
\max \{ \| w - \bar{w} \|, \| \eta - \bar{\eta} \| \} \leq \delta \Rightarrow M_{\bar{w}} \leq M_{w^{n}} \text{ and } M_{\bar{\eta}} \leq M_{w^{n}}.
\]

From now, we consider that \( n \) is large enough so that \( \max \{ \| w^{n} - \bar{w} \|, \| \eta^{n} - \bar{\eta} \| \} \leq \delta \) holds. Thus, the first inequality \( M_{\bar{w}} \leq M_{w^{n}} \) holds. Using [3, Fact 6.14.ii], we see that

\[
\eta^{n} \in \partial R(w^{n}) = \text{cl qr } \partial R(w^{n}) = \text{cl qr } J_{R}(\{ w^{n} \}) \subset \text{cl qr } J_{R}(M_{w^{n}}),
\]

or, equivalently, \( M_{\bar{\eta}}^{*} \leq J_{R}(M_{w^{n}}) \). Considering that \( M_{\bar{\eta}}^{*} \leq M_{w^{n}}^{*} \), we deduce from the transitivity of \( \leq \) that \( M_{\bar{\eta}}^{*} \leq J_{R}(M_{w^{n}}) \). Since we assume \( J_{R} \) to be bijective and decreasing, the latter is equivalent to \( M_{w^{n}} \leq J_{R^{*}}(M_{\bar{\eta}}^{*}) \), which ends the proof.

C. Proof of Theorem II.4

Let \( R \) and \( h \) be as in (2), and let us define

\[
\mathcal{M} := \left\{ \prod_{g=1}^{G} M_{g} : (\forall g \in [1..G]) \ M_{g} \in \{ H_{g} \setminus \{0\}, \{0\} \} \right\},
\]

\[
\mathcal{M}^{*} := \left\{ \prod_{g=1}^{G} M_{g}^{*} : (\forall g \in [1..G]) \ M_{g}^{*} \in \{ \text{S}_{H_{g}}, \text{int B}_{H_{g}} \} \right\}.
\]

It is a simple exercise to verify that \( \mathcal{M} \) is a stratification of \( \text{dom} \partial R = H \), and that \( \mathcal{M}^{*} \) is a stratification of \( \text{dom} \partial R^{*} = \prod_{g} \text{B}_{H_{g}} \). Then, note that for all \( w \in H \) and \( \eta \in \prod_{g} \text{B}_{H_{g}} \):

\[
\partial R(w) = \prod_{g} M_{g}^{*}, \quad \text{with } M_{g}^{*} = \begin{cases} \frac{w_{g}}{\| w_{g} \|} & \text{if } w_{g} \neq 0 \\ \text{int B}_{H_{g}} & \text{else} \end{cases},
\]

\[
\partial R^{*}(\eta) = \prod_{g} M_{g}, \quad \text{with } M_{g} = \begin{cases} [0, +\infty] & \text{if } \eta_{g} \in \text{S}_{H_{g}} \\ \{0\} & \text{else} \end{cases}
\]

Using [4, Proposition 2.5], it is easy to deduce that \( J_{R} \) induces a bijection between \( \mathcal{M} \) and \( \mathcal{M}^{*} \), by putting in relation \( H_{g} \setminus \{0\} \) with \( \text{S}_{H_{g}} \), and \( \{0\} \) with \( \text{int B}_{H_{g}} \). It can be shown by the same arguments that \( J_{R} \) is decreasing for \( \leq \), meaning that \( R \) is mirror-stratifiable w.r.t. \( (\mathcal{M}, \mathcal{M}^{*}) \).

Now, according to the proof of Proposition II.1, the sequence \((w^{n})_{n \in \mathbb{N}} \) verifies

\[
w^{n+1} = \hat{G}_{\lambda \tau}(w^{n} - \tau \lambda \nabla h(w^{n})),
\]
which can be equivalently rewritten as
\[ \eta^n := \frac{w^n - w^{n+1}}{\lambda_T} - \nabla h(w^n) \in \partial R(w^{n+1}). \]

We know from Proposition II.1 that \( w^{n+1} \) converges strongly to \( \bar{w} \), so we can deduce from the continuity of \( \nabla h \) that \( \eta^n \) converges to \( \bar{\eta} := -\nabla h(\bar{w}) \). Applying Proposition IV.2, we deduce that for \( n \) large enough,
\[ M_{\bar{w}} \preceq M_{w^n} \preceq J_{R*}(M^*_\eta). \]

From the definition of \( \mathcal{M} \), we see that \( M_{\bar{w}} \preceq M_{w^n} \) is equivalent to \( \text{supp}(\bar{w}) \subset \text{supp}(w^n) \). Also, we have from the definition of \( J_{R*} \) and \( M^* \) that
\[ J_{R*}(M^*_\eta) = \{ w \in H : w_g \neq 0 \text{ if } \|\bar{\eta}_g\| = 1, w_g = 0 \text{ else} \}. \]

According to the definition of \( \text{esupp}(\bar{w}) \), we deduce that \( M_{w^n} \preceq J_{R*}(M^*_\eta) \) is equivalent to \( \text{supp}(w^n) \subset \text{esupp}(\bar{w}) \), which concludes the proof.

**D. Proof of Theorem II.2**

Theorem II.2 follows directly from Theorem II.4, by observing that (QC) is equivalent to \( \text{supp}(\bar{w}) = \text{esupp}(\bar{w}) \).

**REFERENCES**


