Modular Dimensionality Reduction ⋆

Henry W J Reeve[1], Tingting Mu[2], and Gavin Brown[2]

[1] University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom
[2] University of Manchester, Oxford Rd, Manchester, M13 9PL, United Kingdom

Abstract. We introduce an approach to modular dimensionality reduction, allowing efficient learning of multiple complementary representations of the same object. Modules are trained by optimising an unsupervised cost function which balances two competing goals: Maintaining the inner product structure within the original space, and encouraging structural diversity between complementary representations. We derive an efficient learning algorithm which outperforms gradient based approaches without the need to choose a learning rate. We also demonstrate an intriguing connection with Dropout. Empirical results demonstrate the efficacy of the method for image retrieval and classification.

Keywords: Ensemble learning · Dimensionality reduction · Dropout · Kernel principal components analysis.

1 Introduction

High dimensional data is a widespread challenge in machine learning applications, from computer vision through to bioinformatics and natural language processing. A natural solution is to find a structure-preserving mapping to a low dimensional space, many techniques for which can be found in the literature, such as kernel PCA, Isomap, LLE and Laplacian Eigenmaps [6, 23]. This paper provides a meta-level tool for modular dimensionality reduction, applicable to each of the aforementioned approaches.

We start from the observation that *multiple* abstractions of the same concept can be taken, and may provide complementary views on a task of interest. We therefore propose a *modular* approach to unsupervised dimensionality reduction, in which we learn a diverse collection of low-dimensional representations of the data. Once a modular representation is learned, each module may be used independently – with their respective predictions combined at test time. This procedure is naturally parallelisable in a distributed computing architecture; and, since each representation is low-dimensional, processing for each module is fast and efficient.

⋆ Acknowledgments: H. Reeve was supported by the EPSRC through the Centre for Doctoral Training Grant [EP/1038099/1]. G. Brown was supported by the EPSRC LAMBDA project [EP/N035127/1].

In the context of supervised learning, successful ensemble performance emanates from a fruitful trade-off between the accuracy of the individual members of the ensemble and the degree of diversity [15],[4]. We carry this insight across to the domain of unsupervised dimensionality reduction, by demonstrating the importance of diversity for a set of representation modules. We introduce an unsupervised loss function for training a *set* of dimensionality reduction modules, which balances two competing objectives. The first objective is for each module to preserve relational structure within the original feature space; the second is for modules to exhibit a *diversity* of relational structures.

The contributions of this paper are as follows:

1. An unsupervised loss function for modular dimensionality reduction.
2. A bespoke optimisation procedure which outperforms gradient based methods such as stochastic gradient descent in our setting.
3. A detailed empirical comparison with competitors.
4. An intriguing connection to the dropout algorithm from deep learning [13].

2 Background

We first review work on dimensionality reduction and ensemble learning.

2.1 Unsupervised Dimensionality Reduction

The canonical approach to unsupervised dimensionality reduction is PCA, and its kernelised generalisation, KPCA [21]. KPCA is a general approach which may be applied to a wide variety of application domains through an appropriate choice of kernel [19, 25]. Several manifold learning techniques have also been shown to be special cases of KPCA, with a data-dependent kernel [12].

Classically, KPCA has been viewed as the orthogonal projection which maximises the preserved variance [21]. We shall adopt an alternative perspective in which we view KPCA as a form of *unsupervised similarity learning*, whereby a mapping is chosen so that inner-products in the low dimensional space approximate the kernel. To make this precise we require some notation. We let X denote our original feature space and let $k : \mathsf{X} \quad \mathsf{X} \; ! \; \mathsf{R}$ denote a symmetric positive semi-definite kernel function. Take an unlabelled dataset $\mathsf{D} = \mathsf{f}\boldsymbol{x}_1, ..., \boldsymbol{x}_N\mathsf{g} \quad \mathsf{X}$. For simplicity, we assume throughout that k is centred with respect to D [21]. Let H_k denote the associated reproducing kernel Hilbert space (RKHS) of real-valued functions. For each $H \; 2 \; \mathsf{N}$, we let H_k^H denote the class of H-dimensional mappings $\varphi : \mathsf{X} \; ! \; \mathsf{R}^H$ with coordinate functions taken from the RKHS H_k. That is, for each $\varphi \; 2 \; \mathsf{H}_k^H$ there exists $\varphi^1, \quad , \varphi^H \; 2 \; \mathsf{H}_k$ such that for all $x \; 2 \; \mathsf{X}$, $\varphi(x) = \left(\varphi^h(x)\right)_{h=1}^H$.

Definition 1 (Inner product loss function). *Given an unsupervised data set D and a mapping $\varphi \; 2 \; \mathsf{H}_k^H$, the inner product loss is given by*

$$L_k\left(\varphi, \mathsf{D}\right) = \frac{1}{N^2} \sum_{i,j \quad N} \left(\mathsf{h}\varphi\left(\boldsymbol{x}_i\right), \varphi\left(\boldsymbol{x}_j\right)\mathsf{i} \quad k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\right)^2 .$$

We can interpret KPCA as a form of unsupervised similarity learning which minimises the inner product loss. Let $\xi : X \; ! \; H_k$ denote the canonical embedding given by $\xi(x)(y) = k(x, y)$.

Proposition 1. *The inner product loss* $L_k(\varphi, D)$ *is minimised by taking* φ *to be the member of* H_k^H *obtained by embedding* D *into* H_k *via* ξ *and projecting onto the top* H *kernel principal components.*

The proofs of all results within the main text are given in the appendices.

2.2 Ensembles and Diversity

Combining the outputs of multiple predictors often brings both statistical advantages, such as bias or variance reduction, and computational advantages, through parallelism. In order outperform an individual model, ensembles promote a level of diversity or disagreement between the predictions the constituent models [10, 15]. Whilst methods such as bagging and boosting encourage diversity through a manipulation of the training data, a more direct approach is the *Negative Correlation Learning* (NCL) algorithm of Liu and Yao [18] in which diversity is targeted explicitly.

Suppose we have have a supervised regression ensemble $H = fh_m g_{m=1}^M$ consisting of predictors h_m. In the previous section we used an unsupervised dataset $D = f\mathbf{x}_1, ..., \mathbf{x}_N g$. To distinguish this we use notation $T = f(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)g$ for a *supervised* dataset. We let $V(\,)$ denote the empirical variance of a finite sequence. The NCL algorithm can be understood in terms of the following *modular loss function.*

Definition 2 (Modular loss function). *The modular loss* E_λ *is defined by*

$$E_\lambda(H, T) = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} (h_m(\boldsymbol{x}_n) \quad y_n)^2 \quad \lambda \; \frac{1}{N} \sum_{n=1}^{N} V\left(fh_m(\boldsymbol{x}_n) g_{m=1}^M \right).$$

The modular loss function consists of two terms: A squared loss term which targets the average individual accuracy of the predictors h_m, combined with a diversity term which encourages disagreement between the predictors. The hyper-parameter λ controls the degree of emphasis placed on the diversity. This has the special property that when $\lambda = 1$, $E_\lambda(H, T)$ is exactly the squared loss for the ensemble predictor $\frac{1}{M} \sum_m h_m(\mathbf{x})$ from the target y.

The NCL algorithm is equivalent to stochastic gradient descent applied to the modular loss. This perspective differs from original formulation of the NCL algorithm first introduced by Liu and Yao which utilises a multiplicity of interacting cost functions [18]. However, the updates of the two formulations are equal up to a factor of $1/M$ applied to the learning rate.

3 The Modular Inner Product Loss

Our goal is to train a collection of M distinct but complementary representations of the data. With this goal in mind, we introduce the modular inner product loss which combines two contrasting objectives. On the one hand, we seek high quality representations which faithfully preserve the relational structure encoded by the kernel. On the other hand, we would like the relational structure encoded in our different representations to be diverse. Let $\mathsf{F}(H, M)$ denote the class of all M-tuples $\boldsymbol{\Phi} = \mathsf{f}\boldsymbol{\varphi}_m\mathsf{g}_{m=1}^{M}$ with each $\boldsymbol{\varphi}_m \; 2 \; \mathsf{H}_k^H$. Recall that $\mathsf{V}(\ )$ denotes the empirical variance.

Definition 3 (The modular inner product loss). *Suppose we have an unlabelled data set* $\mathsf{D} \quad \mathsf{X}$ *and a kernel* k*. Given* $\boldsymbol{\Phi} \; 2 \; \mathsf{F}(H, M)$*, the modular inner product loss is given by*

$$L_k^{\lambda}(\boldsymbol{\Phi}, \mathsf{D}) = \frac{1}{M}\sum_{m=1}^{M} L_k(\boldsymbol{\varphi}_m, \mathsf{D}) \quad \lambda \;\; \frac{1}{N^2}\sum_{1 \;\; i.j \;\; N} \mathsf{V}\left(\mathsf{f}\boldsymbol{\varphi}_m(\boldsymbol{x}_i) \;\; \boldsymbol{\varphi}_m(\boldsymbol{x}_j)\mathsf{g}_{m=1}^{M}\right). \tag{1}$$

The modular inner product loss is an analogue of the supervised modular loss function (Definition 2), with inner products between a pair of examples in a representation module replacing predictions for a single example, and the target replaced by an unsupervised inner product.

An equivalent reformulation of the modular inner product loss is as a convex combination between the average inner product loss of the individual modules and the inner product loss of a composite representation. Given $\boldsymbol{\Phi} \; 2 \; \mathsf{F}(H, M)$ we define $\overline{\boldsymbol{\Phi}} \; 2 \; (\mathsf{H}_k)^{H \;\; M}$ by $\overline{\boldsymbol{\Phi}}(\boldsymbol{x}) = \left(1/\sqrt{M}\right) \;\; \left[\boldsymbol{\varphi}_1(\boldsymbol{x})^T, \quad , \boldsymbol{\varphi}_M(\boldsymbol{x})^T\right]^T$. Proposition 2 is proved in Appendix 9.

Proposition 2. $L_k^{\lambda}(\boldsymbol{\Phi}, \mathsf{D}) = (1 \quad \lambda) \;\; \frac{1}{M}\sum_{m=1}^{M} L_k(\boldsymbol{\varphi}_m, \mathsf{D}) + \lambda \;\; L_k\left(\overline{\boldsymbol{\Phi}}, \mathsf{D}\right)$.

When $\lambda = 0$ the loss $L_k^{\lambda}(\boldsymbol{\Phi}, \mathsf{D})$ is minimised by taking each $\boldsymbol{\varphi}_m$ to be a projection onto the top H kernel principal components, whilst for $\lambda = 1$, $L_k^{\lambda}(\boldsymbol{\Phi}, \mathsf{D})$ is minimised by taking $\overline{\boldsymbol{\Phi}}$ to be the projection onto the top $M \;\; H$ kernel principal components. Hence, $L_k^{\lambda}(\boldsymbol{\Phi}, \mathsf{D})$ blends smoothly between training representation modules as individuals and targeting the composite representation.

4 Efficient Optimization

We now introduce the *module-by-module* (MBM) algorithm, which is a form of alternating optimisation designed to minimise the modular inner product loss without the need to choose a learning rate. Our objective is to minimise $L_{\lambda}(\boldsymbol{\Phi}, \mathsf{D})$ over $\boldsymbol{\Phi} \; 2 \; \mathsf{F}(H, M)$. We require an empirical kernel map.

Definition 4. *A rank* R *empirical kernel map is a function* $\boldsymbol{\psi} \; 2 \; \mathsf{H}_k^R$ *such that* $\boldsymbol{\psi}(\boldsymbol{x}_i)^T\psi(\boldsymbol{x}_j) = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ *for all pairs* $(\boldsymbol{x}_i, \boldsymbol{x}_j) \; 2 \; \mathsf{D}^2$.

One can always construct an empirical kernel map of rank N by taking $\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{D})^{\frac{1}{2}}[k(\boldsymbol{x},\boldsymbol{x}_1), \quad , k(\boldsymbol{x},\boldsymbol{x}_N)]^T$, where $\boldsymbol{K}(\boldsymbol{D}) = (k(\boldsymbol{x}_i,\boldsymbol{x}_j))_{ij}$ denotes the kernel gram matrix. Moreover, given a kernel k we can often obtain a low rank empirical kernel map ψ for a kernel $\tilde{k}$ which closely approximates k by employing a method such as random Fourier features [11] or the Nyström method [27]. By reasoning analogous to [20] we have the following useful proposition.

Proposition 3. *Given a rank R empirical kernel map ψ, the minimum for $L_k^{\lambda}(\boldsymbol{\Phi},\boldsymbol{D})$ is attained by $\boldsymbol{\Phi} = \boldsymbol{f}\boldsymbol{\varphi}_m\boldsymbol{g}_{m=1}^{M}$ with each $\boldsymbol{\varphi}_m$ of the form $\boldsymbol{\varphi}_m(\boldsymbol{x}) = \boldsymbol{W}_m \ \boldsymbol{\psi}(\boldsymbol{x})$ for some matrix $\boldsymbol{W}_m \ \boldsymbol{2} \ \mathsf{R}^{H \ R}$.*

Hence, our objective reduces to the following matrix optimisation problem: Minimise

$$C^{\lambda}(\boldsymbol{W},\boldsymbol{\Psi}) : = \sum_{m=1}^{M} \left|\left|\boldsymbol{F}_m^T\boldsymbol{F}_m \quad \boldsymbol{\Psi}^T\boldsymbol{\Psi}\right|\right|^2 \quad \lambda \sum_{m=1}^{M} \left|\left|\boldsymbol{F}_m^T\boldsymbol{F}_m \quad \frac{1}{M}\sum_{q=1}^{M}\boldsymbol{F}_q^T\boldsymbol{F}_q\right|\right|^2 \\ / \ L_k^{\lambda}(\boldsymbol{\Phi}_{\boldsymbol{\mathcal{W}}},\boldsymbol{D}),$$

where $\boldsymbol{\Psi} = [\boldsymbol{\psi}(\boldsymbol{x}_1), \quad , \boldsymbol{\psi}(\boldsymbol{x}_n)] \ \boldsymbol{2} \ \mathsf{R}^{R \ N}$, $\boldsymbol{F}_m = \boldsymbol{W}_m \ \boldsymbol{\Psi}$ and $\boldsymbol{\Phi}_{\boldsymbol{\mathcal{W}}} = \boldsymbol{f}\boldsymbol{W}_m \ \boldsymbol{\psi}\boldsymbol{g}_{m=1}^{M}$. We make use the concept of the rank-constrained approximate square root of a symmetric matrix.

Definition 5. *Define $RT_r : \mathsf{R}^{d \ d} \ ! \ \mathsf{R}^{r \ d}$ by*

$$RT_r(\boldsymbol{M}) = argmin_{\boldsymbol{F} \boldsymbol{2} \mathsf{R}^{r \ d}} \left\{ \left|\left|\boldsymbol{F}^T\boldsymbol{F} \quad \boldsymbol{M}\right|\right|^2 \right\}.$$

Dax has shown that the rank-constrained approximate square root $RT_r(\boldsymbol{M})$ of any $d \quad d$ symmetric matrix $\boldsymbol{M}$ (not necessarily positive semi-definite) may be computed via the singular value decomposition in $O(d^2 \ r)$ time and $O(d^2)$ space complexity [7]. The following proposition allows us to optimise the weights of a single module $\boldsymbol{\varphi}_m$, whilst leaving the remaining modules fixed.

Proposition 4. *Suppose we take $m \ \boldsymbol{2} \ \boldsymbol{f}1, \quad , M\boldsymbol{g}$, fix $\boldsymbol{W}_q$ for $q \ \boldsymbol{6}= m$, and let*

$$\boldsymbol{T}_m = \frac{M}{M \quad \lambda \ (M \quad 1)} \ \boldsymbol{\Psi}^T \left(\boldsymbol{I}_D \quad \frac{\lambda}{M} \ \sum_{q \boldsymbol{6}= m} \boldsymbol{W}_q^T\boldsymbol{W}_q \right) \boldsymbol{\Psi}.$$

Take $\boldsymbol{F}_m = RT_H(\boldsymbol{T}_m)$. Setting $\boldsymbol{W}_m = \boldsymbol{F}_m\boldsymbol{\Psi}^{\boldsymbol{y}}$ minimises $C^{\lambda}(\boldsymbol{W},\boldsymbol{\Psi})$ with respect to $\boldsymbol{W}_m$, under the constraint that $\boldsymbol{W}_q$ remains fixed for $q \ \boldsymbol{6}= m$, where $\boldsymbol{\Psi}^{\boldsymbol{y}}$ denotes the psuedo-inverse of $\boldsymbol{\Psi}$.

Unfortunately, computing $\boldsymbol{F}_m$ via Proposition 4 is $O(N^2 \ H)$ which is intractable for large N. The following proposition enables us to reduce the complexity of this optimisation whenever we have access to an empirical kernel map of rank $R \quad N$.

Proposition 5. *Suppose that $\boldsymbol{\psi}$ is an empirical kernel map of rank R. Take $\tilde{\boldsymbol{\Psi}} = (RT_R(\boldsymbol{\Psi}\boldsymbol{\Psi}^T))^T \ 2 \ \mathsf{R}^{R \ R}$. For all $\mathsf{W} = \mathsf{f}\boldsymbol{W}_m\mathsf{g}_{m=1}^M$ with $\boldsymbol{W}_m \ 2 \ \mathsf{R}^{H \ R}$ we have $C_\lambda\left(\mathsf{W}, \tilde{\boldsymbol{\Psi}}\right) = C_\lambda\left(\mathsf{W}, \boldsymbol{\Psi}\right)$. Moreover, computing $\tilde{\boldsymbol{\Psi}}$ is $O\left(R^2 \ N\right)$ in time complexity and $O(R^2)$ in space complexity.*

Combining Propositions 3, 4 and 5 gives rise to the *module-by-module* algorithm (MBM, Algorithm 1), which is $O(NR^2 + EHR^2)$ in time and $O(NR)$ in space complexity, and has the advantage of reducing the modular inner product loss at every iteration until a critical point is reached.

Inputs: A data set $\mathcal{D} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$, a rank R empirical kernel map ψ, a number of modules M, a number of dimensions per module H, a diversity parameter λ and $\epsilon > 0$.
Compute $\boldsymbol{\Psi} = [\boldsymbol{\psi}(\boldsymbol{x}_1), \cdots, \boldsymbol{\psi}(\boldsymbol{x}_n)]$;
Update $\boldsymbol{\Psi} = (RT_\rho\left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T\right))^T$;
Randomly initialise $\boldsymbol{F}_m \in \mathsf{R}^{H \ R}$ for $m = 1, \cdots, M$;
Compute $\boldsymbol{Q} = \boldsymbol{\Psi}^T\boldsymbol{\Psi}$ and $\boldsymbol{S} = \sum_{m=1}^M \boldsymbol{F}_{\boldsymbol{m}}^T\boldsymbol{F}_m$;
Compute $c = ((1-\lambda) + \lambda/M + \epsilon) \ ^1$;
for $e = 1, \cdots, E$ **do**
  for $m = 1, \cdots, M$ **do**
    Compute $\boldsymbol{S} \ _m = \boldsymbol{S} - \boldsymbol{F}_m^T\boldsymbol{F}_m$;
    Compute $\boldsymbol{T} = c \cdot \left(\boldsymbol{Q} - (\lambda/M)\, \boldsymbol{S} \ _m + \epsilon \cdot \boldsymbol{F}_{\boldsymbol{m}}^T\boldsymbol{F}_m\right)$;
    Update $\boldsymbol{F}_m = RT_H\left(\boldsymbol{T}\right)$;
    Update $\boldsymbol{S} = \boldsymbol{S} \ _m + \boldsymbol{F}_m^T\boldsymbol{F}_m$;
  end
end
Compute $\boldsymbol{W}_m = \boldsymbol{F}_m\boldsymbol{\Psi}^{\mathsf{y}}$ for $m = 1, \cdots, M$;
Output: $\boldsymbol{\Phi} = \{\boldsymbol{W}_m \cdot \boldsymbol{\psi}\}_{m=1}^M$.

Algorithm 1: The module-by-module (MBM) algorithm.

The following theorem justifies the use of the MBM algorithm - it is guaranteed to reduce the modular inner product loss at every epoch until a critical point is reached.

Theorem 1. *Given $E \ 2 \ \mathsf{N}$, let $\boldsymbol{\Phi}^E \ 2 \ \mathsf{F}(H, M)$ denote the set obtained by training with Algorithm 1, for E epochs. Then for all $E \ 2 \ \mathsf{N}$, $L_k^\lambda\left(\boldsymbol{\Phi}^{E+1}, \mathsf{D}\right) < L_k^\lambda\left(\boldsymbol{\Phi}^E, \mathsf{D}\right)$, unless $\boldsymbol{\Phi}^E$ is a critical point of $L_k^\lambda\left(\boldsymbol{\Phi}, \mathsf{D}\right)$, in which case $L_k^\lambda\left(\boldsymbol{\Phi}^{E+1}, \mathsf{D}\right)$ $L_k^\lambda\left(\boldsymbol{\Phi}^E, \mathsf{D}\right)$.*

5 The Dropout Connection

In this section we introduce a surprising connection between the modular inner product loss and the dropout algorithm [13, 22]. Dropout is a state of the art approach to regularising deep neural networks in which a random collection of hidden neurons is "dropped out" at each stochastic gradient update. The dropout algorithm can be understood as implicitly minimising the expectation of a stochastic loss function based on predictions from a random subnetwork [22, 26]. There is a natural analogue of this, in our setting: to minimise the expectation of a stochastic variant of the inner product loss, based on inner products computed from a random subset of modules. We refer to this analogue as the drop-module (DM) algorithm. To be precise, given an ensemble of feature mappings $\boldsymbol{\Phi} = (\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M) \in \mathcal{F}(H, M)$ each binary vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_M) \in \{0,1\}^M$ corresponds to a 'noisy' representation $\boldsymbol{\Phi}_{\boldsymbol{\eta}}$ given by $\boldsymbol{\Phi}_{\boldsymbol{\eta}}(\boldsymbol{x}) = (1/\sqrt{M})(\eta_1 \boldsymbol{\varphi}_1(\boldsymbol{x}), \ldots, \eta_M \boldsymbol{\varphi}_M(\boldsymbol{x}))$.

Fix a probability $p \in [0,1]$ and let $B(p)$ denote the probability measure on $\{0,1\}$ with $\mathbb{E}_{B(p)}(\eta) = p$. Let $\boldsymbol{\Theta}$ denote the parameters of $\boldsymbol{\Phi}$. The DM algorithm proceeds by randomly sampling $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{D}$ and $\eta_m \sim B(p)$ and updating

$$\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta} - \alpha \frac{\partial}{\partial \boldsymbol{\Theta}} \left(\langle \boldsymbol{\Phi}_{\boldsymbol{\eta}}(\boldsymbol{x}_i), \boldsymbol{\Phi}_{\boldsymbol{\eta}}(\boldsymbol{x}_j) \rangle - k(\boldsymbol{x}_i, \boldsymbol{x}_j) \right)^2 .$$

The DM algorithm implicitly minimises the following stochastic loss function

$$L_{k,p}^{\text{drop}}(\boldsymbol{\Phi}, \mathcal{D}) = \mathbb{E}_{\boldsymbol{\eta} \sim B(p)^M} \left[L_k(\boldsymbol{\Phi}_{\boldsymbol{\eta}}, \mathcal{D}) \right].$$

Previously, Baldi et al. demonstrated that dropout may be understood as training an exponentially large ensemble with shared weights [1]. In our setting this corresponds to shared weight a ensemble of size 2^M with an ensemble member for each $\boldsymbol{\eta} \in \{0,1\}^M$. We demonstrate that the DM algorithm can be related to an ensemble of size M, trained via the modular inner product loss (see Definition 3). We emphasise that unlike the shared weight ensembles considered by Baldi et al. [1], here we consider ensembles with separate weights in which the interaction takes place purely via the diversity term in the modular inner product loss.

Theorem 2. *The drop-module inner product loss at p is equivalent to the modular inner product loss at $\lambda = Mp/(1+p(M-1))$. To be precise, for $\boldsymbol{\Phi}_0 \in \mathcal{F}(H, M)$ we have*

$$\left. \frac{\partial L_{k,p}^{\text{drop}}(\boldsymbol{\Phi}, \mathcal{D})}{\partial \boldsymbol{\Phi}} \right|_{\boldsymbol{\Phi} = \boldsymbol{\Phi}_0} = p \left. \frac{\partial L_k^{\lambda}(\boldsymbol{\Phi}, \mathcal{D})}{\partial \boldsymbol{\Phi}} \right|_{\boldsymbol{\Phi} = (p/\sqrt{p/\lambda})\, \boldsymbol{\Phi}_0} .$$

Theorem 2 implies that if we take $\lambda = Mp/(1 + p(M-1))$ then the minima of $L_k^{\lambda}(\boldsymbol{\Phi}, \mathcal{D})$ are equal to the minima of $L_{k,p}^{\text{drop}}(\boldsymbol{\Phi}, \mathcal{D})$, up to a constant scaling factor of $\sqrt{p/\lambda}$. In this sense, the minima for the two loss functions are *representationally equivalent*. The relationship between the diversity parameter λ in the MBM algorithm and the probability of keeping module p in the DM algorithm is illustrated in Figure 1.

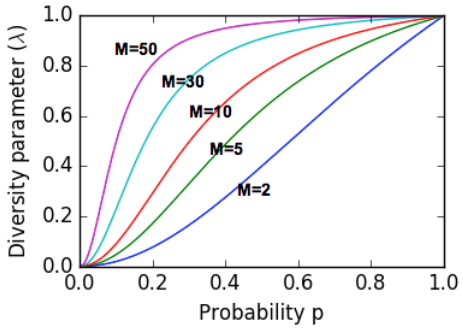

Fig. 1. The diversity parameter λ in MBM vs. the corresponding p in DM.

6 Experimental Results

In this section we first demonstrate the optimisation performance of the MBM algorithm before comparing our method for other natural approaches for training multiple kernelised representations. The data sets used in all experiments are described in Section 12.1.

6.1 Optimisation performance of the MBM algorithm

In this section we assess the MBM algorithm (Algorithm 1) in terms of its efficiency at optimising the modular inner product loss. We compare with two gradient based approaches. As a baseline we consider stochastic gradient descent (SGD) applied directly to the modular inner product loss. This is expected to perform poorly in our setting since the modular inner product loss sums over all pairs of examples. We also consider the state of the art Adam optimiser of Kingma et al. [14] (Adam). The Adam optimiser is applied in batch mode, first compressing the data by applying Proposition 5. We set $M = 10$, $H = 10$, $\lambda = 0.9$ and let k to be the Gaussian kernel with γ set using the the heuristic of Kwok and Tsang [16, Section 4]. In addition, we employ a rank $R = 1000$ Nyström approximation [27]. For SGD and Adam we we consider learning rates in the range $f10^{-6}, \ldots, 10^{2}g$. We evaluated the algorithms by training for one hour and evaluating the minimum value of the loss function attained during training and the convergence time - the time taken for the loss function to fall within 1% of its minimum. For SGD and Adam we report results corresponding to the learning rate which achieves the lowest minimum loss. The results are shown in Table 1. The SGD method was extremely slow and typically failed to converge within one hour. The compressed Adam method performed much better and typically converged within 30 minutes. However, the bespoke MBM algorithm achieved the same minimum loss in at most twice the speed on each of the data sets. The MBM algorithm also has the advantage of not requiring the user to set a learning rate.

Table 1. A comparison of the MBM algorithm with gradient based methods.

	Loss function minimum			Convergence time (seconds)		
Data set	SGD	Adam	MBM	SGD	Adam	MBM
Convex	110.0±21.2	**13.4±0.0**	**13.3±0.0**	3369.8±163.8	963.9±368.9	**334.7±151.2**
MNIST	427.6±16.3	**59.9±0.0**	**59.9±0.0**	3484.7±63.2	1609.1±69.2	**400.6±69.4**
NUS Wide	598.6±8.4	**73.1±0.0**	**73.1±0.0**	3376.3±185.2	1495.1±149.5	**422.7±47.7**
NVDR	110.0±18.5	**10.7±0.0**	**10.6±0.0**	1891.3±578.8	1157.2±227.0	**402.4±35.1**
Rectangles	29.4±0.4	**10.3±0.0**	**10.3±0.0**	1796.1±687.6	746.9±162.0	**256.2±36.5**

6.2 Image retrieval & classification performance of MBM modules

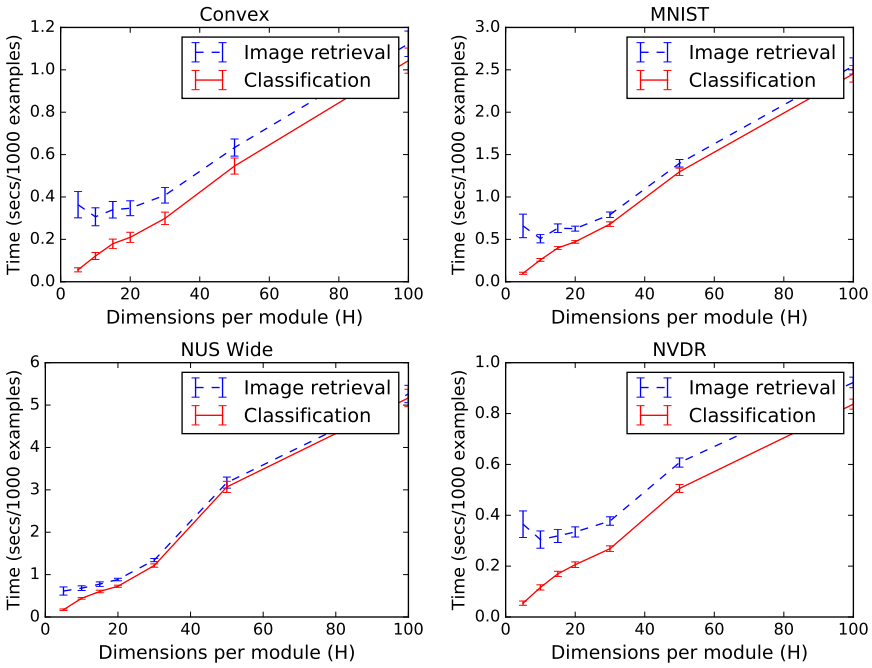

Fig. 2. Test time as vs. H. See Appendix 12.2 for discussion & other data sets.

We compare four unsupervised approaches to training multiple kernelised feature mappings:

Partition We compute the top HM KPCAs and randomly partition these into M sets of H, so that each mapping φ_m 2 H_k^H is a projection onto a disjoint subset of the top H M KPCAs.

Bootstrap Bagging [3] applied to KPCA. For each $m \in \{1, \ldots, M\}$ take a bootstrap sample $\tilde{D}_m$, of size N, and let $\varphi_m \in \mathsf{H}_k^H$ be the KPCA projection mapping onto H dimensions for $\tilde{D}_m$.

Random A kernelised variant of the widely used technique of random projections [2, 8]. For each $m = 1, \ldots, M$ we sample a random matrix $\boldsymbol{R}_m \in \mathsf{R}^{H \times D}$ from an $H \times D$ standard normal distribution, and normalise each row so that it has unit norm. In order to kernelise this technique the feature space for the random matrices is the output of the empirical kernel map ψ (see Section 4).

MBM Our proposed approach in which $\boldsymbol{\Phi}$ is trained to minimise the modular inner product loss via MBM algorithm. The diversity parameter λ is set based upon performance on a validation set. We shall consider $H \in \{5, 10, 15, 20, 30, 50, 100\}$ and take M so that $H \times M = 300$. We also compare with the following non-modular base line.

Monolithic A single mapping $\varphi \in \mathsf{H}_k^H$ - the projection onto the top 300 KPCAs.

In each case we take k to be the Gaussian kernel with the γ parameter set via the heuristic of Kwok and Tsang [16, Section 4]. For computational efficiency we employ a rank 1000 Nyström approximation [27] in each case. We shall consider two distinct tasks:

Image retrieval: We shall consider the modular low dimensional representation's capability for efficiently retrieving a set of κ close-by images. Let $\boldsymbol{\Phi} = \{\varphi_m\}_{m=1}^M \in \mathsf{F}(H, M)$ be a modular representation and D an unlabelled training set. Given a test point $\boldsymbol{x} \in \mathsf{X}$, for each module, we compute the set $I_{\kappa,n}^{\varphi_m}(\boldsymbol{x}) \subset D$ of κ-nearest neighbours of $\boldsymbol{x}$ based the distance $\|\varphi_m(\boldsymbol{x}_q) - \varphi_m(\boldsymbol{x})\|_2$. We then extract subset of size κ, $I_{\kappa,n}^{\boldsymbol{\Phi}}(\boldsymbol{x}) \subset \bigcup_{m=1}^M I_{\kappa,n}^{\varphi_m}(\boldsymbol{x})$ so that the elements $\boldsymbol{x}_q \in I_{\kappa,n}^{\boldsymbol{\Phi}}(\boldsymbol{x})$ minimise the average squared distance from the test point $\boldsymbol{x}$ over the low dimensional spaces ie. $(1/M) \sum_{m=1}^M \|\varphi_m(\boldsymbol{x}_q) - \varphi_m(\boldsymbol{x})\|_2^2$ is minimised. Let $I_{\kappa,n}(\boldsymbol{x})$ denote the set of κ nearest neighbours as computed in the original space X. To assess performance we compute the precision: the average value of $(1/\kappa) \times \#\left(I_{\kappa,n}^{\boldsymbol{\Phi}}(\boldsymbol{x}) \setminus I_{\kappa,n}(\boldsymbol{x})\right)$. This procedure is based upon the method of [24] and gives a quantitative assessment of the representation's ability to preserve structural information. The results of the image retrieval task for $\kappa = 10$, $H = 20$ and $M = 15$ are shown in Table 2. On each of the eight data sets the precision attained by the MBM method significantly exceeds the precision attained by the other modular methods: partition, bootstrap and random. Table 3 compares MBM with the monolithic method in which we simply compute the 10 nearest neighbours in the φ-projected space, where φ is the projection onto the top 300 KPCAs. For a relatively modest reduction in performance the MBM method obtains a significant speed up at test time. The speed up is due to the fact that each set of nearest neighbours $I_{\kappa,n}^{\varphi_m}(\boldsymbol{x})$ may be computed in parallel on a low-dimensional space (see Figure 2 and Appendix 12.2). Figure 3 demonstrates the precision as a function of the number of dimensions per module (H)

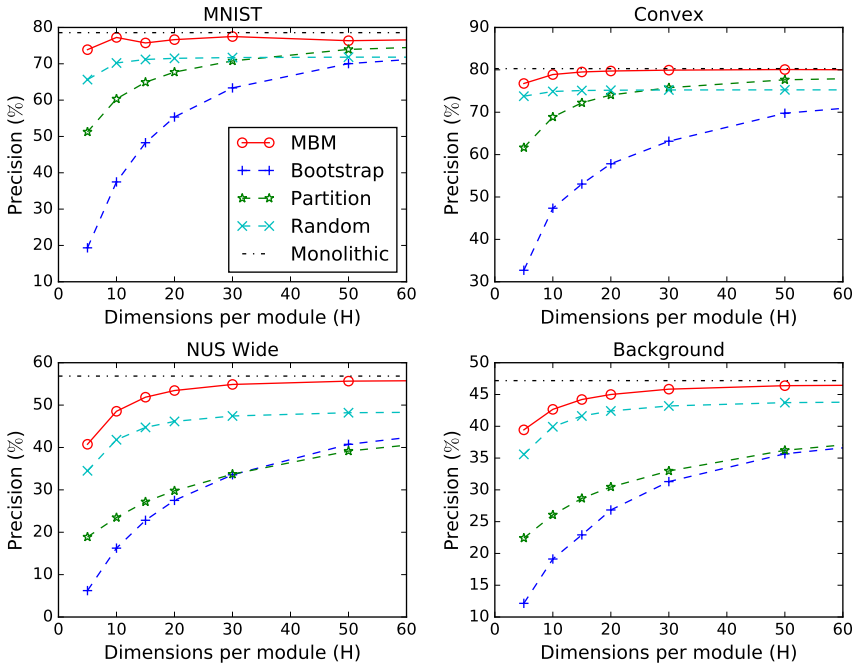

Fig. 3. Precision as a function of H (See Section 6.2).

with $\kappa = 10$ and $H \quad M = 300$. As H increases, the precision approaches the precision attained by the 300-dimensional Monolithic approach. The precision attained by the MBM approach typically exceeds that attained by the other modular approaches (Bootstrap, Partition, Random) across a range of values of H. Corresponding figures for other data sets are given in Appendix 12.3.

Classification We compare the methods in terms of their capacity for extracting multiple sets of features for use in a classification ensemble. Given a modular representation $\boldsymbol{\Phi} = f\varphi_m g_{m=1}^{M}\ 2\ F(H, M)$, for each m we train a classifier f_m based on the features extracted by φ_m. Given a test point $\boldsymbol{x}$ we combine the outputs of $f f_m\left(\varphi_m(\boldsymbol{x})\right) g_{m=1}^{M}$ by taking a modal average. Table 2 shows the classification accuracy for ensembles consisting of 15 5-nearest neighbour classifiers trained on 20-dimensional spaces. The MBM approach significantly outperforms the other approaches on five out of eight data sets, and performs comparably or better than the alternatives on every data set. Table 3 compares with the monolithic approach - a single 5-nearest neighbour classifier on 300 KPCAs. The MBM approach is both faster and more accurate than the monolithic method on all but one data set.

Table 2. A comparison of methods for modular dimensionality reduction. See Section 6.2 for details.

	Image retrieval precision (%)				Ensemble classification accuracy (%)			
Data set	Partition	Bootstrap	Random	MBM	Partition	Bootstrap	Random	MBM
NUS Wide	29.8±0.3	27.5±0.3	46.1±0.3	**53.4±0.3**	32.2±0.7	37.7±0.7	42.0±0.7	**43.7±0.7**
NVDR	68.0±0.5	48.1±0.5	73.4±0.5	**77.6±0.4**	**44.9±1.0**	39.6±1.0	**43.9±1.0**	**44.8±1.0**
MNIST	67.7±0.1	55.4±0.1	71.5±0.1	**76.6±0.1**	89.2±0.2	95.0±0.2	94.5±0.2	**95.8±0.1**
Background	30.5±0.2	26.8±0.2	42.4±0.2	**45.0±0.2**	41.0±0.4	47.3±0.4	53.8±0.4	**57.3±0.4**
Random	7.0±0.1	10.9±0.1	6.1±0.1	**17.2±0.1**	40.9±0.4	89.5±0.2	53.1±0.4	**90.3±0.2**
Rotations	59.6±0.2	54.3±0.1	68.4±0.1	**75.3±0.1**	72.0±0.3	85.5±0.3	83.8±0.3	**87.4±0.2**
Convex	74.1±0.1	57.8±0.2	75.2±0.1	**79.7±0.1**	55.7±0.4	**65.5±0.3**	60.3±0.4	**65.6±0.3**
Rectangles	63.3±0.1	44.6±0.1	46.7±0.1	**71.6±0.1**	93.0±0.2	95.5±0.2	93.9±0.2	**98.1±0.1**

Table 3. Comparing the MBM & the Monolithic approach (See Section 6.2).

	Image retrieval				Ensemble classification			
Data set	Monolithic	MBM Δ	Speed	λ	Monolithic	MBM Δ	Speed	λ
NUS Wide	56.8±0.3	-3.4±0.5	10.8×	0.999	40.6±0.7	+3.1±1.0	11.7×	0.990
NVDR	79.0±0.5	-1.4±0.6	7.7×	0.999	40.8±1.0	+3.9±1.4	12.6×	0.999
MNIST	78.6±0.1	-1.9±0.1	9.1×	0.990	95.7±0.1	+0.1±0.2	12.9×	0.900
Background	47.2±0.2	-2.2±0.2	5.8×	0.999	52.5±0.4	+4.9±0.5	8.1×	0.999
Random	23.1±0.1	-5.9±0.1	5.7×	0.950	83.9±0.3	+6.4±0.3	7.1×	0.500
Rotations	76.5±0.1	-1.2±0.1	7.9×	0.990	86.3±0.3	+1.0±0.4	10.9×	0.800
Convex	80.3±0.1	-0.6±0.1	7.2×	0.999	57.6±0.4	+8.0±0.5	23.0×	0.200
Rectangles	70.1±0.1	+1.5±0.2	2.4×	0.990	95.8±0.1	-2.3±0.2	10.7×	0.990

The two Monolothic columns show the image retrieval precision (%) and the classification accuracy (%) of the monolithic method. The MBM columns show the corresponding change in performance due to using the MBM method for each task. The Speed columns show the corresponding speed ups ie. the test time for the Monolithic method divided by the test time for the MBM method.

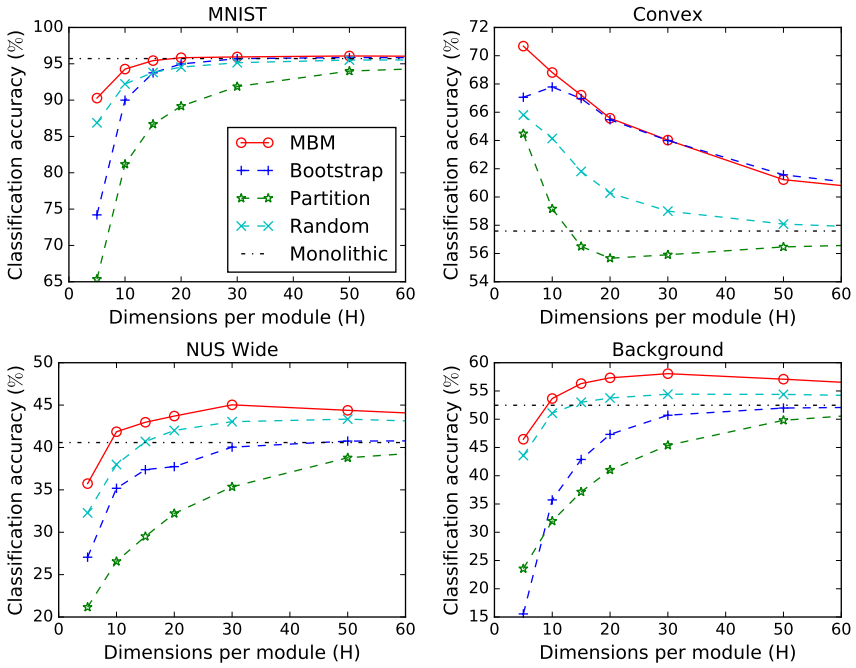

Fig. 4. Classification accuracy as a function of H (See Section 6.2).

The diversity parameter The diversity parameter λ in the MBM algorithm controls the level of emphasis placed upon encouraging a diversity of representations. We found that the optimal performance (both in terms of information retrieval and classification) was typically attained with λ just below 1, with performance declining sharply by taking $\lambda = 1$ (see Figure 5, cols 1& 2). It is interesting to observe that the dropout algorithm often performs well with p 0.5 and this corresponds a value of λ just below 1, when M is large (see Figure 1). However, whilst this pattern was observed on all data sets for image retrieval (see Appendix 12.6), for some data sets the best classification performance was attained by taking much lower values of λ (see Figure 5, col 3 and Appendix 12.6). Ultimately, the optimal value of λ is data dependent and must be set based on validation performance.

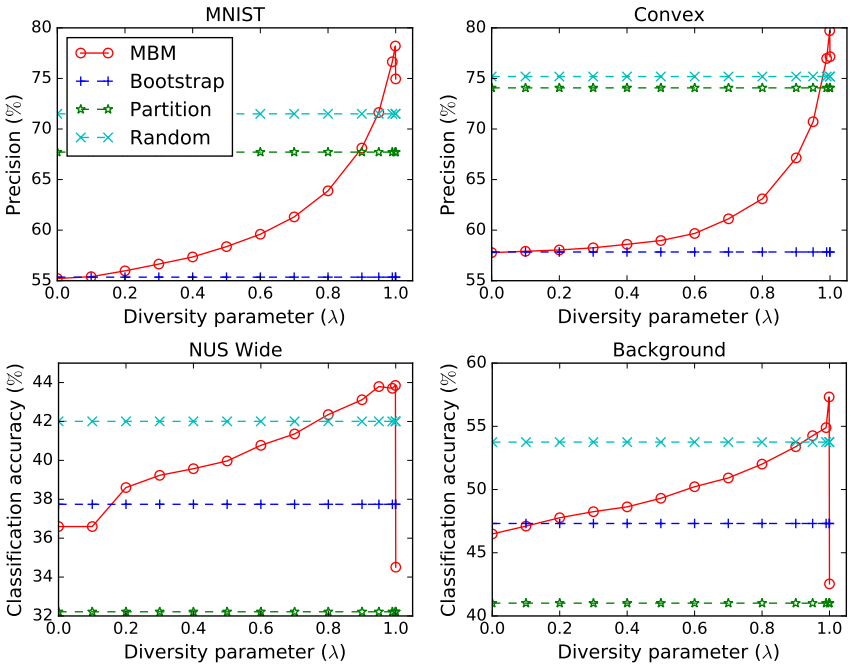

Fig. 5. Performance as a function of the diversity parameter (λ) (see Section 6.2).

7 Discussion

We have investigated a method for *modular* unsupervised dimensionality reduction. Our method is based upon the *modular inner product loss* (Definition 3), an adaptation of concepts from both negative correlation learning [4, 18] and kernel principal components analysis [21]. Whilst the modular loss could be optimised by gradient based methods we introduced a novel *module-by-module* algorithm, which converges at least twice as fast as a state of the art gradient based optimiser [14] without the need to tune the learning rate.

Modular representations have the potential to be applied on range of tasks. Empirical results on both image retrieval and classification tasks confirm that the MBM algorithm is superior to a range of competitors including random projections and bootstrapping, whilst providing a parallelisation advantage over "monolithic" dimensionality reduction. We also demonstrated an intriguing equivalency between our proposal and an analogue of the dropout algorithm - drop module, which deserves further attention.

In summary, this work has shown the potential of explicitly managing diversity in unsupervised representation learning.

References

1. Pierre Baldi and Peter J Sadowski. Understanding dropout. In *Advances in Neural Information Processing Systems*, pages 2814–2822, 2013.
2. Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250. ACM, 2001.
3. Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
4. Gavin Brown, Jeremy L Wyatt, and Peter Tiňo. Managing diversity in regression ensembles. *The Journal of Machine Learning Research*, 6:1621–1650, 2005.
5. Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, page 48. ACM, 2009.
6. John P Cunningham and Zoubin Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*, 16:2859–2900, 2015.
7. Achiya Dax. Low-rank positive approximants of symmetric matrices. *Advances in Linear Algebra & Matrix Theory*, 4(03):172, 2014.
8. Robert J Durrant and Ata Kabán. Random projections as regularizers: Learning a linear discriminant ensemble from fewer observations than dimensions.
9. Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.
10. Pascal Germain, Alexandre Lacasse, Francois Laviolette, Mario Marchand, and Jean-Francis Roy. Risk bounds for the majority vote: From a pac-bayesian analysis to a learning algorithm. *The Journal of Machine Learning Research*, 16(1):787–860, 2015.
11. Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.
12. Jihun Ham, Daniel D Lee, Sebastian Mika, and Bernhard Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the twenty- rst international conference on Machine learning*, page 47. ACM, 2004.
13. Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.
14. Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
15. Anders Krogh, Jesper Vedelsby, et al. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, pages 231–238, 1995.
16. James Tin-Yau Kwok and Ivor Wai-Hung Tsang. The pre-image problem in kernel methods. *Neural Networks, IEEE Transactions on*, 15(6):1517–1525, 2004.
17. Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra, and Yoshua Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on Machine learning*, pages 473–480. ACM, 2007.
18. Yong Liu and Xin Yao. Ensemble learning via negative correlation. *Neural Networks*, 12(10):1399–1404, 1999.

19. Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *The Journal of Machine Learning Research*, 2:419–444, 2002.
20. Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pages 416–426. Springer, 2001.
21. Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *Arti cial Neural NetworksICANN'97*, pages 583–588. Springer, 1997.
22. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
23. Dmitry Storcheus, Afshin Rostamizadeh, and Sanjiv Kumar. A survey of modern questions and challenges in feature extraction. In *Proceedings of The 1st International Workshop on Feature Extraction: Modern Questions and Challenges, NIPS*, pages 1–18, 2015.
24. Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11(Feb):451–490, 2010.
25. S Vichy N Vishwanathan, Nicol N Schraudolph, Risi Kondor, and Karsten M Borgwardt. Graph kernels. *The Journal of Machine Learning Research*, 11:1201–1242, 2010.
26. Sida I Wang and Christopher D Manning. Fast dropout training.
27. Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Proceedings of the 14th Annual Conference on Neural Information Processing Systems*, number EPFL-CONF-161322, pages 682–688, 2001.
28. Xiao Wu, Alexander G Hauptmann, and Chong-Wah Ngo. Practical elimination of near-duplicates from web video search. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 218–227. ACM, 2007.

Appendix for Modular Dimensionality Reduction

8 Proof of Proposition 1

In this section we prove Proposition 1. We will require some notation. Recall that $\xi : X \to \mathsf{H}_k$ denotes the canonical embedding given by $\xi(x)(y) = k(x,y)$.We also let $P_D^H : \mathsf{H}_k \to \mathsf{R}^H$ denote corresponding Hotelling transform, ie. the projection onto the top H kernel principal components embedded isometrically into R^H. Given a pair of Hilbert spaces $\mathsf{H}^1, \mathsf{H}^2$ with $a^i \in \mathsf{H}^i$ for $i \in \{1,2\}$ we define a bounded linear operator $a^2 \otimes a^1 : \mathsf{H}^1 \to \mathsf{H}^2$ by $(a^2 \otimes a^1)(z) = \langle a^1, z\rangle \, a^2$ for all $z \in \mathsf{H}^1$. Note that $(a^2 \otimes a^1)^* = a^2 \otimes a^1$, ie. given $z_i \in \mathsf{H}^i$ for $i \in \{1,2\}$ we have $\langle (a^2 \otimes a^1)(z_1), z_2\rangle = \langle z_1, (a^1 \otimes a^2)(z_2)\rangle = \langle a_1, z_1\rangle \langle a_2, z_2\rangle$. Finally, we let $\{e_1, \ldots, e_N\}$ denote the canonical orthonormal basis for R^N.

Proposition 1. *The inner product loss $L_k(\varphi, D)$ is minimised by taking φ to be the member of H_k^H obtained by embedding D into H_k via ξ and projecting onto the top H kernel principal components.*

Proof. We must show that the inner product loss $L_k(\varphi, D)$ is minimised over $\varphi \in X \to \mathsf{R}^H$ by taking $\varphi(x) = P_D^H(\xi(x))$. We shall define a data dependent operator $\boldsymbol{\Psi} : \mathsf{R}^N \to \mathsf{H}_k$ by $\boldsymbol{\Psi} = \sum_{n=1}^N \xi(x_n) \otimes e_n$. Clearly $\boldsymbol{\Psi}$ has rank at most N, so by the singular value theorem for finite rank operators there exists an orthonormal basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N\} \subset \mathsf{R}^N$, a set of orthonormal vectors $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N\} \subset \mathsf{H}_k$ and a set of singular values $\sigma_1, \ldots, \sigma_N \in \mathsf{R}^N$ with $|\sigma_1| \geq |\sigma_2| \geq \cdots \geq |\sigma_N|$ such that $\boldsymbol{\Psi} = \sum_{q=1}^N \sigma_q (\boldsymbol{u}_q \otimes \boldsymbol{v}_q)$. Now let $\boldsymbol{C}$ denote the covariance operator given by $\boldsymbol{C} = (1/N) \sum_{n=1}^N \xi(x_n) \otimes \xi(x_n)$, where we use the fact that k is centred with respect to D. Since $\{e_1, \ldots, e_N\}$ is orthonormal we have

$$
\begin{aligned}
N \boldsymbol{C} &= \sum_{n=1} \xi(x_n) \otimes \xi(x_n) = \left(\sum_{n=1} \xi(x_n) \otimes e_n\right)\left(\sum_{n=1} e_n \otimes \xi(x_n)\right) = \boldsymbol{\Psi}\boldsymbol{\Psi}^* \\
&= \left(\sum_{q=1}^N \sigma_q (\boldsymbol{u}_q \otimes \boldsymbol{v}_q)\right)\left(\sum_{\tilde{q}=1}^N \sigma_{\tilde{q}} (\boldsymbol{v}_{\tilde{q}} \otimes \boldsymbol{u}_{\tilde{q}})\right) = \sum_{q=1}^N \sigma_q^2 (\boldsymbol{u}_q \otimes \boldsymbol{u}_q)\,.
\end{aligned}
$$

Thus, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_H$ are the top H kernel principal components, ie. the eigenvectors of $\boldsymbol{C}$ with maximal eigenvalues. Consequently, $P_D^H = \sum_{h=1}^H \boldsymbol{e}_h \otimes \boldsymbol{u}_h$. Let $\boldsymbol{A} : \mathsf{H}_k \to \tilde{\mathsf{H}}$ be an arbitrary linear operator into a Hilbert space $\tilde{\mathsf{H}}$. We have $\boldsymbol{A}\boldsymbol{\Psi} = \sum_{n=1}^N (\boldsymbol{A}\xi(x_n)) \otimes e_n$, so $\boldsymbol{\Psi}^* \circ \boldsymbol{A}^* \circ \boldsymbol{A}\boldsymbol{\Psi}$ corresponds to an $N \times N$ matrix with entries

$$
\begin{aligned}
(\boldsymbol{\Psi}^* \circ \boldsymbol{A}^* \circ \boldsymbol{A}\boldsymbol{\Psi})_{n_1,n_2} &= e_{n_1}^T (\boldsymbol{\Psi}^* \circ \boldsymbol{A}^* \circ \boldsymbol{A}\boldsymbol{\Psi})\, e_{n_2} = \left\langle e_{n_1}, (\boldsymbol{A}\boldsymbol{\Psi}) \circ (\boldsymbol{A}\boldsymbol{\Psi})\, e_{n_2}\right\rangle \\
&= \langle (\boldsymbol{A}\boldsymbol{\Psi})\, e_{n_1}, (\boldsymbol{A}\boldsymbol{\Psi})\, e_{n_2}\rangle = \langle \boldsymbol{A}\xi(x_{n_1}), \boldsymbol{A}\xi(x_{n_1})\rangle\,.
\end{aligned}
$$

In particular, taking $\boldsymbol{A}$ to be the identity $\boldsymbol{I}_{N \times N}$ we have

$$
(\boldsymbol{\Psi}^* \boldsymbol{\Psi})_{n_1,n_2} = \langle \xi(x_{n_1}), \xi(x_{n_2})\rangle = k(x_{n_1}, x_{n_2})
$$

Hence, if let $\boldsymbol{K}$ denote the matrix with entries $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, we have

$$\boldsymbol{K} = \boldsymbol{\Psi}\ \boldsymbol{\Psi} = \left(\sum_{q=1}^{N} \sigma_q\ \ (\boldsymbol{v}_q \quad \boldsymbol{u}_q)\right)\left(\sum_{q=1}^{N} \sigma_q\ \ (\boldsymbol{u}_q \quad \boldsymbol{v}_q)\right) = \sum_{q=1}^{N} \sigma_q^2\ \ (\boldsymbol{v}_q \quad \boldsymbol{v}_q)\,.$$

Thus, by the Eckart-Young theorem [9] the rank H matrix $\boldsymbol{M}$ which minimises $\boldsymbol{kM} \quad \boldsymbol{K}\boldsymbol{k}_2$ (where $\boldsymbol{k}\ \ \boldsymbol{k}_2$ denotes the Frobenius norm) is $\boldsymbol{M}_H = \sum_{q=1}^{H} \sigma_q^2$ $(\boldsymbol{v}_q \quad \boldsymbol{v}_q)$. Given any matrix $\boldsymbol{\varphi} : \boldsymbol{X}\ \ ! \ \ \mathsf{R}^H$ we shall let $\boldsymbol{Z}(\boldsymbol{\varphi})$ denote the $H \quad N$ matrix with entries $\varphi_h(x_n)$ for $h\ \boldsymbol{2}\ \boldsymbol{f}1, \quad , H\boldsymbol{g}$ and $n\ \boldsymbol{2}\ \boldsymbol{f}1, \quad , N\boldsymbol{g}$. It follows that $\boldsymbol{Z}(\boldsymbol{\varphi})\ \ \boldsymbol{Z}(\boldsymbol{\varphi})$ is the $N \quad N$ matrices with entries $\boldsymbol{h}\boldsymbol{\varphi}(x_{n_1}), \boldsymbol{\varphi}(x_{n_2})\boldsymbol{i}$, so

$$\boldsymbol{k}\boldsymbol{Z}(\boldsymbol{\varphi})\ \ \boldsymbol{Z}(\boldsymbol{\varphi}) \quad \boldsymbol{K}\boldsymbol{k}_2 = \sum_{i,j \quad N} \left(\boldsymbol{h}\boldsymbol{\varphi}\left(\boldsymbol{x}_i\right), \boldsymbol{\varphi}\left(\boldsymbol{x}_j\right)\boldsymbol{i} \quad k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\right)^2 = N^2 \ \ L_k\left(\varphi, \boldsymbol{D}\right).$$

Given any function $\boldsymbol{\varphi} : \boldsymbol{X}\ \ ! \ \ \mathsf{R}^H$, $\boldsymbol{Z}(\boldsymbol{\varphi})\ \ \boldsymbol{Z}(\boldsymbol{\varphi})$ is of rank at most H, so if we can choose $\boldsymbol{\varphi}$ so that $\boldsymbol{Z}(\boldsymbol{\varphi})\ \ \boldsymbol{Z}(\boldsymbol{\varphi}) = \boldsymbol{M}_H$ then $L_k\left(\varphi, \boldsymbol{D}\right)$ will have attained its minimum. Now consider $\boldsymbol{\varphi}\ \boldsymbol{2}\ \mathsf{H}_k^H$ defined by $\boldsymbol{\varphi}(x) = \boldsymbol{P}_{\boldsymbol{D}}^H \boldsymbol{\xi}(x)$ for $x\ \boldsymbol{2}\ \boldsymbol{X}$. Given $n_1, n_2 \quad N$ we have

$$\left(\boldsymbol{Z}(\boldsymbol{\varphi})\ \ \boldsymbol{Z}(\boldsymbol{\varphi})\right)_{n_1,n_2} = \boldsymbol{h}\boldsymbol{\varphi}(x_{n_1}), \boldsymbol{\varphi}(x_{n_2})\boldsymbol{i} = \left\langle \boldsymbol{P}_{\boldsymbol{D}}^H \boldsymbol{\xi}(x_{n_1}), \boldsymbol{P}_{\boldsymbol{D}}^H \boldsymbol{\xi}(x_{n_2}) \right\rangle = \left(\boldsymbol{\Psi}\ \ \left(\boldsymbol{P}_{\boldsymbol{D}}^H\right)\ \ \boldsymbol{P}_{\boldsymbol{D}}^H \boldsymbol{\Psi}\right)_{n_1,n_2}.$$

Hence, we have

$$\begin{aligned}
\boldsymbol{Z}(\boldsymbol{\varphi})\ \ \boldsymbol{Z}(\boldsymbol{\varphi}) &= \boldsymbol{\Psi}\ \ \left(\boldsymbol{P}_{\boldsymbol{D}}^H\right)\ \ \boldsymbol{P}_{\boldsymbol{D}}^H \boldsymbol{\Psi} \\
&= \boldsymbol{\Psi}\ \ \left(\sum_{q=1}^{H} \boldsymbol{u}_q \quad e_q\right)\left(\sum_{q=1}^{H} e_q \quad \boldsymbol{u}_q\right)\boldsymbol{\Psi} = \boldsymbol{\Psi}\ \ \left(\sum_{q=1}^{H} \boldsymbol{u}_q \quad \boldsymbol{u}_q\right)\boldsymbol{\Psi} \\
&= \left(\sum_{q=1}^{N} \sigma_q\ \ (\boldsymbol{v}_q \quad \boldsymbol{u}_q)\right)\left(\sum_{q=1}^{H} \boldsymbol{u}_q \quad \boldsymbol{u}_q\right)\left(\sum_{q=1}^{N} \sigma_q\ \ (\boldsymbol{u}_q \quad \boldsymbol{v}_q)\right) \\
&= \sum_{q=1}^{H} \sigma_q^2\ \ (\boldsymbol{v}_q \quad \boldsymbol{v}_q) = \boldsymbol{M}_H.
\end{aligned}$$

Hence, taking $\boldsymbol{\varphi} = \boldsymbol{P}_{\boldsymbol{D}}^H \quad \boldsymbol{\xi}$ minimises the inner product loss $L_k\left(\varphi, \boldsymbol{D}\right)$.

9 Proof of Proposition 2

In this section we prove Proposition 2.

Proposition 2. $L_k^{\lambda}(\boldsymbol{\Phi}, \mathsf{D}) = (1 \quad \lambda) \;\; \frac{1}{M}\sum_{m=1}^{M} L_k(\boldsymbol{\varphi}_m, \mathsf{D}) + \lambda \;\; L_k\left(\overline{\boldsymbol{\Phi}}, \mathsf{D}\right)$.

Proof. It suffices to show that

$$L_k\left(\overline{\boldsymbol{\Phi}}, \mathsf{D}\right) = \frac{1}{M}\sum_{m=1}^{M} L_k(\boldsymbol{\varphi}_m, \mathsf{D}) \quad \frac{1}{N^2}\sum_{1 \;\; i.j \;\; N} \mathsf{V}\left(\mathsf{f}\varphi_m(\boldsymbol{x}_i) \;\; \varphi_m(\boldsymbol{x}_j)\mathsf{g}_{m=1}^{M}\right).$$

This may be seen by subtracting

$$(1 \quad \lambda) \;\; \frac{1}{M}\sum_{m=1}^{M} L_k(\boldsymbol{\varphi}_m, \mathsf{D}) + \lambda \;\; L_k\left(\overline{\boldsymbol{\Phi}}, \mathsf{D}\right).$$

from

$$\frac{1}{M}\sum_{m=1}^{M} L_k(\boldsymbol{\varphi}_m, \mathsf{D}) \quad \lambda \;\; \frac{1}{N^2}\sum_{1 \;\; i.j \;\; N} \mathsf{V}\left(\mathsf{f}\varphi_m(\boldsymbol{x}_i) \;\; \varphi_m(\boldsymbol{x}_j)\mathsf{g}_{m=1}^{M}\right),$$

Thus, it suffices to show that for any $\boldsymbol{x}_i, \boldsymbol{x}_j \; \mathsf{2} \; \mathsf{X}^2$,

$$\left(\mathsf{h}\overline{\boldsymbol{\Phi}}(\boldsymbol{x}_i), \overline{\boldsymbol{\Phi}}(\boldsymbol{x}_j)\mathsf{i} \quad k_c(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)^2 = \frac{1}{M}\sum_{m=1}^{M} \left(\mathsf{h}\boldsymbol{\varphi}_m(\boldsymbol{x}_i), \boldsymbol{\varphi}_m(\boldsymbol{x}_j)\mathsf{i} \quad k_c(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)^2 \mathsf{V}\left(\mathsf{f}\mathsf{h}\varphi_m(\boldsymbol{x}_i), \varphi_m(\boldsymbol{x}_j)\mathsf{i}\mathsf{g}_{m=1}^{M}\right).$$

Note that by construction $\mathsf{h}\overline{\boldsymbol{\Phi}}(\boldsymbol{x}_i), \overline{\boldsymbol{\Phi}}(\boldsymbol{x}_j)\mathsf{i} = (1/M) \;\; \sum_{m=1}^{M} \mathsf{h}\boldsymbol{\varphi}_m(\boldsymbol{x}_i), \boldsymbol{\varphi}_m(\boldsymbol{x}_j)\mathsf{i}$. Thus, the result follows from the observation that for any sequence $\mathsf{f}a_m\mathsf{g}_{m=1}^{M} \quad \mathsf{R}$ and $b \; \mathsf{2} \; \mathsf{R}$ if we let $\overline{a} = (1/M)\sum_{m=1}^{M} a_m$ then we have

$$(\overline{a} \quad b)^2 = \frac{1}{M}\sum_{m=1}^{M} (a_m \quad b)^2 \quad \frac{1}{M}\sum_{m=1}^{M} (a_m \quad \overline{a})^2.$$

This decomposition is known as the ambiguity-decomposition in the ensemble literature and was observed by Krogh and Vedelsby in the context of ensembles [15].

10 Proof of Theorem 1 & Propositions 3, 4, 5

In this section we prove Theorem 1 which justifies the procedure given by Algorithm 1 (the MBM algorithm), along with several supporting propositions. Throughout this section $jj\ jj$ denotes the Frobenius norm. Recall that we assume the existence of an empirical kernel map $\psi\ 2\ \mathsf{H}_k^H$.

Proposition 3. *Given a rank R empirical kernel map ψ, the minimum for $L_k^\lambda(\boldsymbol{\Phi}, \mathsf{D})$ is attained by $\boldsymbol{\Phi} = f\varphi_m g_{m=1}^M$ with each φ_m of the form $\varphi_m(\boldsymbol{x}) = \boldsymbol{W}_m\ \ \psi(\boldsymbol{x})$ for some matrix $\boldsymbol{W}_m\ 2\ \mathsf{R}^{H\ \ R}$.*

Recall that $\xi : X\ \ !\ \ \mathsf{H}_k$ denote the canonical embedding given by $\xi(x)(y) = k(x, y)$.

Lemma 1. *Given $f\ 2\ \mathsf{H}_k$ there exists $\boldsymbol{a}\ 2\ \mathsf{R}^R$ such that $\tilde{f} = \boldsymbol{a}^T\psi$ satisfies $\tilde{f}(x_n) = f(x_n)$ for $n = 1, \quad , N$.*

Proof. Write $f = \sum_{q=1}^N \beta_q\ \ \xi(\boldsymbol{x}_q) + g$ where $g\ 2\ \mathsf{H}_k$ is orthogonal to $f\xi(\boldsymbol{x}_n)g_{n=1}^N$. By orthogonality, combined with the reproducing property, for each $n = 1, \quad , N$ we have $g(\boldsymbol{x}_n) = hg, \xi(\boldsymbol{x}_n)i = 0$. Hence, $f(\boldsymbol{x}_n) = \sum_{q=1}^Q \beta_q \xi(\boldsymbol{x}_q)(\boldsymbol{x}_n)$ for $n = 1, \quad , N$. Now let $\boldsymbol{a} := \sum_{q=1}^N \beta_q \psi(\boldsymbol{x}_q)$ and define $\tilde{f} := \boldsymbol{a}^T\psi$. Then, for each $n = 1, \quad , N$ we have

$$\begin{aligned}\tilde{f}(\boldsymbol{x}_n) = \boldsymbol{a}^T\psi(\boldsymbol{x}_n) &= \sum_{q=1}^N \beta_q \psi(\boldsymbol{x}_q)^T\psi(\boldsymbol{x}_n) \\ &= \sum_{q=1}^N \beta_q k(\boldsymbol{x}_q, \boldsymbol{x}_n) = \sum_{q=1}^N \beta_q\ \ \xi(\boldsymbol{x}_q)(\boldsymbol{x}_n) = f(\boldsymbol{x}_n).\end{aligned}$$

Proof (Proof of Proposition 3). Take $\boldsymbol{\Phi} = f\varphi_m g_{m=1}^M\ 2\ F(H, M)$. By Lemma 1, for each $m = 1, \quad , M$, there exists matrices $\boldsymbol{W}_m\ 2\ \mathsf{H}\quad \mathsf{R}$ such that $\tilde{\varphi}_m := \boldsymbol{W}_m\ \ \psi_m$ satisfies $\tilde{\varphi}_m(\boldsymbol{x}_n) = \varphi_m(\boldsymbol{x}_n)$ for each $n = 1, \quad , N$. It follows that $\tilde{\boldsymbol{\Phi}} = f\tilde{\varphi}_m g_{m=1}^M$ satisfies $\tilde{\boldsymbol{\Phi}}(x_n) = \boldsymbol{\Phi}(\boldsymbol{x}_n)$ for all $\boldsymbol{x}_n\ 2\ \mathsf{D}$, so $L_k^\lambda\left(\tilde{\boldsymbol{\Phi}}, \mathsf{D}\right) = L_k^\lambda\left(\tilde{\Phi}, \mathsf{D}\right)$.

This leads to the objective of minimising

$$C^\lambda(\boldsymbol{W}, \boldsymbol{\Psi}) := \sum_{m=1}^M \left|\left|\boldsymbol{F}_m^T\boldsymbol{F}_m \quad \boldsymbol{\Psi}^T\boldsymbol{\Psi}\right|\right|^2 \quad \lambda\ \sum_{m=1}^M \left|\left|\boldsymbol{F}_m^T\boldsymbol{F}_m \quad \frac{1}{M}\sum_{q=1}^M \boldsymbol{F}_q^T\boldsymbol{F}_q\right|\right|^2\ \ /\ \ L_k^\lambda(\boldsymbol{\Phi}_{\boldsymbol{\mathcal{W}}}, \mathsf{D}),$$

where $\boldsymbol{\Psi} = [\psi(\boldsymbol{x}_1), \quad , \psi(\boldsymbol{x}_n)]\ 2\ \mathsf{R}^{R\ \ N}$, $\boldsymbol{F}_m = \boldsymbol{W}_m\ \boldsymbol{\Psi}$ and $\boldsymbol{\Phi}_{\boldsymbol{\mathcal{W}}} = f\boldsymbol{W}_m\ \ \psi g_{m=1}^M$. Recall Definition 5.

Definition 5. *Define $RT_r : \mathsf{R}^{d\ \ d}\ \ !\ \ \mathsf{R}^{r\ \ d}$ by*

$$RT_r(\boldsymbol{M}) = argmin_{\boldsymbol{F}\,2\,\mathsf{R}^{r\ \ d}} \left\{\left|\left|\boldsymbol{F}^T\boldsymbol{F} \quad \boldsymbol{M}\right|\right|^2\right\}.$$

Dax has shown that when $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ is symmetric, $RT_r(\boldsymbol{M}) \in \mathbb{R}^{r \times d}$ may be computed via the singular value decomposition for any $r \leq D$ [7]. We may take an eigen-decomposition of $\boldsymbol{M} = \sum_{i=1}^{d} \gamma_i \boldsymbol{u}_i^T \boldsymbol{u}_i$, where $\gamma_i \in \mathbb{R}$ are eigen-values and $\boldsymbol{u}_i$ are unit eigen-vectors of $\boldsymbol{M}$. We can compute the approximate root as follows,

$$RT_r(\boldsymbol{M}) = \operatorname{diag}\left(\sqrt{\max\{\gamma_1, 0\}}, \ldots, \sqrt{\max\{\gamma_r, 0\}}\right) [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r]^T. \quad (2)$$

Note that in the special case in which $\boldsymbol{M}$ is positive semi-definite ie. $\gamma_i \geq 0$ for all i, this result is the classical Eckart and Young Theorem [9].

We first prove Proposition 5.

Proposition 5. *Suppose that ψ is an empirical kernel map of rank R. Take $\tilde{\boldsymbol{\Psi}} = (RT_R(\boldsymbol{\Psi}\boldsymbol{\Psi}^T))^T \in \mathbb{R}^{R \times R}$. For all $\mathsf{W} = \{\boldsymbol{W}_m\}_{m=1}^M$ with $\boldsymbol{W}_m \in \mathbb{R}^{H \times R}$ we have $C_\lambda\left(\mathsf{W}, \tilde{\boldsymbol{\Psi}}\right) = C_\lambda(\mathsf{W}, \boldsymbol{\Psi})$. Moreover, computing $\tilde{\boldsymbol{\Psi}}$ is $O\left(R^2 N\right)$ in time complexity and $O(R^2)$ in space complexity.*

Proof (Proof of Proposition 5). Take a singular value decomposition of $\boldsymbol{\Psi}$, $\boldsymbol{\Psi} = \sum_{l=1}^{\rho} \sigma_l \boldsymbol{u}_l \boldsymbol{v}_l^T$, let $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_\rho] \in \mathbb{R}^{D \times \rho}$, $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_\rho) \in \mathbb{R}^{\rho \times \rho}$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_\rho] \in \mathbb{R}^{N \times \rho}$. Hence, $\boldsymbol{\Psi} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$. We define $\tilde{\boldsymbol{\Psi}} \in \mathbb{R}^{D \times \rho}$ by $\tilde{\boldsymbol{\Psi}} = \boldsymbol{U}\boldsymbol{\Sigma}$. Note that $\tilde{\boldsymbol{\Psi}}$ may computed using the singular value decomposition in $O(R^2 N)$ time and $O(R^2)$ space complexity. Note also that since the columns of $\boldsymbol{V}$ are orthonormal and $\|\cdot\|$ is the the Frobenius norm, for any matrix $\boldsymbol{A} \in \rho \times \rho$ we have $\left\|\boldsymbol{V}\boldsymbol{A}\boldsymbol{V}^T\right\| = \|\boldsymbol{A}\|$. Hence, we have

$$\begin{aligned}
C^\lambda\left(\mathsf{W}, \tilde{\boldsymbol{\Psi}}\right) &= (1-\lambda)\ \frac{1}{M}\sum_{m=1}^{M}\left\|\left(\boldsymbol{W}_m\tilde{\boldsymbol{\Psi}}\right)^T\left(\boldsymbol{W}_m\tilde{\boldsymbol{\Psi}}\right) - \tilde{\boldsymbol{\Psi}}^T\tilde{\boldsymbol{\Psi}}\right\|^2 \\
&\quad + \lambda\ \left\|\frac{1}{M}\sum_{m=1}^{M}\left(\boldsymbol{W}_m\tilde{\boldsymbol{\Psi}}\right)^T\left(\boldsymbol{W}_m\tilde{\boldsymbol{\Psi}}\right) - \tilde{\boldsymbol{\Psi}}^T\tilde{\boldsymbol{\Psi}}\right\|^2 \\
&= (1-\lambda)\ \frac{1}{M}\sum_{m=1}^{M}\left\|\boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{W}_m^T\boldsymbol{W}_m\boldsymbol{U}\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Sigma}\right\|^2 \\
&\quad + \lambda\ \left\|\boldsymbol{\Sigma}\boldsymbol{U}^T\left(\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{W}_m^T\boldsymbol{W}m\right)\boldsymbol{U}\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Sigma}\right\|^2 \\
&= (1-\lambda)\ \frac{1}{M}\sum_{m=1}^{M}\left\|\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{W}_m^T\boldsymbol{W}_m\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T - \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\right\|^2 \\
&\quad + \lambda\ \left\|\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T\left(\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{W}_m^T\boldsymbol{W}m\right)\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T - \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\right\|^2 = C^\lambda(\mathsf{W}, \boldsymbol{\Psi}).
\end{aligned}$$

We now prove Proposition 4.

Proposition 4. *Suppose we take $m \in \{1, \ldots, M\}$, fix $\boldsymbol{W}_q$ for $q \neq m$, and let*

$$\boldsymbol{T}_m = \frac{M}{M - \lambda (M - 1)} \boldsymbol{\Psi}^T \left(\boldsymbol{I}_D - \frac{\lambda}{M} \sum_{q \neq m} \boldsymbol{W}_q^T \boldsymbol{W}_q \right) \boldsymbol{\Psi}.$$

Take $\boldsymbol{F}_m = RT_H(\boldsymbol{T}_m)$. Setting $\boldsymbol{W}_m = \boldsymbol{F}_m \boldsymbol{\Psi}^{\dagger}$ minimises $C^{\lambda}(\mathcal{W}, \boldsymbol{\Psi})$ with respect to $\boldsymbol{W}_m$, under the constraint that $\boldsymbol{W}_q$ remains fixed for $q \neq m$, where $\boldsymbol{\Psi}^{\dagger}$ denotes the psuedo-inverse of $\boldsymbol{\Psi}$.

Proposition 4 is a special case of the more general Proposition 10 which will be useful in proving Theorem 1.

Suppose we take $m \in \{1, \ldots, M\}$ and fix $\boldsymbol{W}_q$ for $q \neq m$ and take $\epsilon > 0$. Fix a matrix $\boldsymbol{W}_m^{old} \in \mathbb{R}^{H \times D}$ and let $\boldsymbol{\Theta} = (\boldsymbol{W}_m^{old} \boldsymbol{\Psi})^T (\boldsymbol{W}_m^{old} \boldsymbol{\Psi})$. Take $c = ((1 - \lambda) + \lambda / M + \epsilon)^{-1}$. Define $\boldsymbol{S}_m = \sum_{q \neq m} \boldsymbol{F}_q^T \boldsymbol{F}_q$ where $\boldsymbol{F}_q = \boldsymbol{W}_q \boldsymbol{\Psi}$, and $\boldsymbol{Q} = \boldsymbol{\Psi}^T \boldsymbol{\Psi}$,

$$\boldsymbol{T} = c \left(\boldsymbol{Q} - \frac{\lambda}{M} \boldsymbol{S}_m + \epsilon \boldsymbol{\Theta} \right).$$

Take $\boldsymbol{F}_m = RT_H(\boldsymbol{T})$. Then taking $\boldsymbol{W}_m = \boldsymbol{F}_m \boldsymbol{\Psi}^{\dagger}$ minimises

$$C^{\lambda}(\mathcal{W}, \boldsymbol{\Psi}) + \frac{\epsilon}{M} \left|\left| (\boldsymbol{W}_m \boldsymbol{\Psi})^T (\boldsymbol{W}_m \boldsymbol{\Psi}) - \boldsymbol{\Theta} \right|\right|^2,$$

with respect to $\boldsymbol{W}_m$, under the constraint that $\boldsymbol{W}_q$ remains fixed for $q \neq m$.

Proof (Proof of Proposition 10). We note that with $\boldsymbol{F}_q = \boldsymbol{W}_q \boldsymbol{\Psi}$ for all $q = 1, \ldots, M$ we have

$$\begin{aligned} &C^{\lambda}(\mathcal{W}, \boldsymbol{\Psi}) + \frac{\epsilon}{M} \left|\left| (\boldsymbol{W}_m \boldsymbol{\Psi})^T (\boldsymbol{W}_m \boldsymbol{\Psi}) - \boldsymbol{\Theta} \right|\right|^2 \\ &\quad = (1 - \lambda) \frac{1}{M} \sum_{q=1}^{M} \left|\left| \boldsymbol{F}_q^T \boldsymbol{F}_q - \boldsymbol{Q} \right|\right|^2 + \lambda \left|\left| \frac{1}{M} \sum_{q=1}^{M} \boldsymbol{F}_q^T \boldsymbol{F}_q - \boldsymbol{Q} \right|\right|^2 + \frac{\epsilon}{M} \left|\left| \boldsymbol{F}_m^T \boldsymbol{F}_m - \boldsymbol{\Theta} \right|\right|^2. \end{aligned}$$

By removing terms independent of $\boldsymbol{F}_m$ and multiplying by M we see that this is equivalent to minimising

$$(1 - \lambda) \left|\left| \boldsymbol{F}_m^T \boldsymbol{F}_m - \boldsymbol{Q} \right|\right|^2 + \frac{\lambda}{M} \left|\left| \boldsymbol{F}_m^T \boldsymbol{F}_m - (M \boldsymbol{Q} - \boldsymbol{S}_m) \right|\right|^2 + \epsilon \left|\left| \boldsymbol{F}_m^T \boldsymbol{F}_m - \boldsymbol{\Theta} \right|\right|^2.$$

Moreover, this expression is equal up to a constant term (independent of $\boldsymbol{F}_m$) to the following,

$$\left((1 - \lambda) + \frac{\lambda}{M} + \epsilon \right) \left|\left| \boldsymbol{F}_m^T \boldsymbol{F}_m \right|\right|^2 - 2 \left|\left| \boldsymbol{F}_m^T \boldsymbol{F}_m \left(\boldsymbol{Q} - \frac{\lambda}{M} \boldsymbol{S}_m + \epsilon \boldsymbol{\Theta} \right) \right|\right|^2.$$

In addition, by adjusting by another constant and multiplying through by $c = \left((1 - \lambda) + \frac{\lambda}{M} + \epsilon \right)^{-1}$ we see that this is equivalent to minimising $\left|\left| \boldsymbol{F}_m^T \boldsymbol{F}_m - \boldsymbol{T} \right|\right|^2$

where $\boldsymbol{T} = c\ \left(\boldsymbol{Q} \quad \frac{\lambda}{M}\ \ \boldsymbol{S}\ \ _m + \epsilon\ \ \boldsymbol{\Theta}\right)$. Thus, we minimise $\boldsymbol{F}_m\ 2\ \mathsf{R}^{H\ \ N}$ by taking $\boldsymbol{F}_m = RT_H(\boldsymbol{T})$. Now take $\boldsymbol{W}_m = \boldsymbol{F}_m \boldsymbol{\Psi}^{\mathsf{y}}$. It suffices to show that $\boldsymbol{W}_m \boldsymbol{\Psi} = \boldsymbol{F}_m$.

To prove this consider the singular value decomposition $\boldsymbol{\Psi} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ where $\boldsymbol{\Psi} = \sum_{l=1}^{\rho} \sigma_l \boldsymbol{u}_l \boldsymbol{v}_l^T$, with orthonormal vectors $\boldsymbol{U} = [\boldsymbol{u}_1, \quad , \boldsymbol{u}_\rho]\ 2\ \mathsf{R}^{D\ \ \rho}$, $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \quad , \sigma_\rho)\ 2\ \mathsf{R}^{\rho\ \ \rho}$ and orthonormal vectors $\boldsymbol{V} = [\boldsymbol{v}_1, \quad , \boldsymbol{v}_\rho]\ 2\ \mathsf{R}^{N\ \ \rho}$. Here ρ denotes the rank of $\boldsymbol{\Psi}$ and for all $l = 1, \quad , \rho$ we have $\sigma_l > 0$. Thus, the psuedo-inverse is given by $\boldsymbol{\Psi}^{\mathsf{y}} = \boldsymbol{V}\boldsymbol{\Sigma}\ \ ^1\boldsymbol{U}^T$.

Given any vector $\boldsymbol{z}\ 2\ \mathsf{R}^N$ which lies in the orthogonal complement to the span of $\mathsf{f}\boldsymbol{v}_1, \quad , \boldsymbol{v}_\rho\mathsf{g}$ (equivalently, in the null space of $\boldsymbol{V}^T$) we have $\boldsymbol{\Psi}\boldsymbol{z} = \boldsymbol{0}$. Noting that $\boldsymbol{T}$ may be written

$$\boldsymbol{T} = c\ \ \boldsymbol{\Psi}^T \left(\boldsymbol{I}_D \quad \frac{\lambda}{M} \sum_{q \neq m} \boldsymbol{W}_q^T \boldsymbol{W}_q + \epsilon (\boldsymbol{W}_m^{old})^T (\boldsymbol{W}_m^{old}) \right) \boldsymbol{\Psi}.$$

Hence for z in the null space of $\boldsymbol{V}^T$ we have $\boldsymbol{T}\boldsymbol{z} = 0$. By equation 2 this implies $\boldsymbol{F}_m \boldsymbol{z} = 0$. Note also that $\boldsymbol{\Psi}^{\mathsf{y}}\boldsymbol{\Psi} = \boldsymbol{V}\boldsymbol{V}^T$, so for $\boldsymbol{z}$ in the null space of $\boldsymbol{V}^T$ we have $\boldsymbol{\Psi}^{\mathsf{y}}\boldsymbol{\Psi}\boldsymbol{z} = 0$ whereas for $\boldsymbol{w}$ in the linear span of $\mathsf{f}\boldsymbol{v}_1, \quad , \boldsymbol{v}_\rho\mathsf{g}$ we have $\boldsymbol{\Psi}^{\mathsf{y}}\boldsymbol{\Psi}\boldsymbol{w} = \boldsymbol{w}$. Hence, if we take $\boldsymbol{W}_m = \boldsymbol{F}_m\boldsymbol{\Psi}^{\mathsf{y}}$, then for $\boldsymbol{z}$ in the null space of $\boldsymbol{V}^T$ we have

$$\boldsymbol{F}_m \boldsymbol{z} = 0 = \boldsymbol{F}_m \boldsymbol{\Psi}^{\mathsf{y}} \boldsymbol{\Psi} \boldsymbol{z}.$$

On the other hand, if $\boldsymbol{w}$ is in the linear span of $\mathsf{f}\boldsymbol{v}_1, \quad , \boldsymbol{v}_\rho\mathsf{g}$ then we have $(\boldsymbol{\Psi}^{\mathsf{y}}\boldsymbol{\Psi})\boldsymbol{w} = \boldsymbol{w}$, so

$$\boldsymbol{F}_m \boldsymbol{w} = \boldsymbol{F}_m \boldsymbol{\Psi}^{\mathsf{y}} \boldsymbol{\Psi} \boldsymbol{w}.$$

Thus, by taking an orthogonal decomposition we see if we take $\boldsymbol{W}_m = \boldsymbol{F}_m \boldsymbol{\Psi}^{\mathsf{y}}$ then indeed

$$\boldsymbol{W}_m \boldsymbol{\Psi} = (\boldsymbol{F}_m \boldsymbol{\Psi}^{\mathsf{y}}) \boldsymbol{\Psi} = \boldsymbol{F}_m.$$

This completes the proof of Proposition 10 which implies Proposition 4.

We now combine Propositions 5 and 10 to prove Theorem 1.
First we recall Algorithm 1 and the statement of Theorem 1.

Theorem 1. *Given $E\ 2\ \mathsf{N}$, let $\boldsymbol{\Phi}^E\ 2\ \mathsf{F}(H, M)$ denote the set obtained by training with Algorithm 1, for E epochs. Then for all $E\ 2\ \mathsf{N}$, $L_k^{\lambda}\left(\boldsymbol{\Phi}^{E+1}, \mathsf{D}\right) < L_k^{\lambda}\left(\boldsymbol{\Phi}^E, \mathsf{D}\right)$, unless $\boldsymbol{\Phi}^E$ is a critical point of $L_k^{\lambda}(\boldsymbol{\Phi}, \mathsf{D})$, in which case $L_k^{\lambda}\left(\boldsymbol{\Phi}^{E+1}, \mathsf{D}\right)$ $L_k^{\lambda}\left(\boldsymbol{\Phi}^E, \mathsf{D}\right)$.*

Proof (Proof of Theorem 1). Firstly, by Proposition 5, the replacement of $\boldsymbol{\Psi}$ with $\left(RT_\rho(\boldsymbol{\Psi}\boldsymbol{\Psi}^T)\right)^T$ at the start of Algorithm 1, while computationally essential, has no effect upon the loss function. Secondly, by Proposition 10, each update in the inner loop of Algorithm 1, minimises

$$f(\boldsymbol{W}_m) = N^2 \quad L_k^{\lambda}(\boldsymbol{\Phi}, \mathsf{D}) + \frac{\epsilon}{M} \left|\left|(\boldsymbol{W}_m \boldsymbol{\Psi})^T (\boldsymbol{W}_m \boldsymbol{\Psi}) \quad \boldsymbol{\Theta}\right|\right|^2,$$

Inputs: A data set $\mathcal{D} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$, a rank R empirical kernel map ψ, a number of modules M, a number of dimensions per module H, a diversity parameter λ and $\epsilon > 0$.
Compute $\boldsymbol{\Psi} = [\boldsymbol{\psi}(\boldsymbol{x}_1), \cdots, \boldsymbol{\psi}(\boldsymbol{x}_n)]$;
Update $\boldsymbol{\Psi} = (RT_\rho\left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T\right))^T$;
Randomly initialise $\boldsymbol{F}_m \in \mathsf{R}^{H \quad R}$ for $m = 1, \cdots, M$;
Compute $\boldsymbol{Q} = \boldsymbol{\Psi}^T\boldsymbol{\Psi}$ and $\boldsymbol{S} = \sum_{m=1}^{M} \boldsymbol{F}_{\boldsymbol{m}}^T\boldsymbol{F}_m$;
Compute $c = ((1-\lambda) + \lambda/M + \epsilon)^{\quad 1}$;
for $e = 1, \cdots, E$ **do**
  for $m = 1, \cdots, M$ **do**
    Compute $\boldsymbol{S}_{\quad m} = \boldsymbol{S} - \boldsymbol{F}_m^T\boldsymbol{F}_m$;
    Compute $\boldsymbol{T} = c \cdot \left(\boldsymbol{Q} - (\lambda/M)\,\boldsymbol{S}_{\quad m} + \epsilon \cdot \boldsymbol{F}_{\boldsymbol{m}}^T\boldsymbol{F}_m\right)$;
    Update $\boldsymbol{F}_m = RT_H\left(\boldsymbol{T}\right)$;
    Update $\boldsymbol{S} = \boldsymbol{S}_{\quad m} + \boldsymbol{F}_m^T\boldsymbol{F}_m$;
  end
end
Compute $\boldsymbol{W}_m = \boldsymbol{F}_m\boldsymbol{\Psi}^y$ for $m = 1, \cdots, M$;
Output: $\boldsymbol{\Phi} = \{\boldsymbol{W}_m \cdot \boldsymbol{\psi}\}_{m=1}^{M}$.

Algorithm 2: The module-by-module (MBM) algorithm.

where $\boldsymbol{\Theta} = (\boldsymbol{W}_m^{old}\boldsymbol{\Psi})^T(\boldsymbol{W}_m^{old}\boldsymbol{\Psi})$. In particular, we have $f(\boldsymbol{W}_m) \quad f(\boldsymbol{W}_m^{old})$. Thus, by the definition of $\boldsymbol{\Theta}$ we see that replacing $\boldsymbol{W}_m^{old}$ cannot increase $L_k^\lambda\left(\boldsymbol{\Phi}, \boldsymbol{D}\right)$, as $\left|\left|(\boldsymbol{W}_m\boldsymbol{\Psi})^T(\boldsymbol{W}_m\boldsymbol{\Psi}) \quad \boldsymbol{\Theta}\right|\right|^2$ is minimised by $\boldsymbol{W}_m = \boldsymbol{W}_m^{old}$. Hence, we have $L_k^\lambda\left(\boldsymbol{\Phi}^{E+1}, \boldsymbol{D}\right) \quad L_k^\lambda\left(\boldsymbol{\Phi}^E, \boldsymbol{D}\right)$ for all E *2* N.

Now suppose $L_k^\lambda\left(\boldsymbol{\Phi}^{E+1}, \boldsymbol{D}\right) = L_k^\lambda\left(\boldsymbol{\Phi}^E, \boldsymbol{D}\right)$. By Proposition 10, the combined cost $L_k^\lambda\left(\boldsymbol{\Phi}, \boldsymbol{D}\right) + \frac{\epsilon}{M}\left|\left|(\boldsymbol{W}_m\boldsymbol{\Psi})^T(\boldsymbol{W}_m\boldsymbol{\Psi}) \quad \boldsymbol{\Theta}\right|\right|^2$ is minimised at each stage, so we must have $(\boldsymbol{W}_m^{E+1}\boldsymbol{\Psi})^T(\boldsymbol{W}_m^{E+1}\boldsymbol{\Psi}) = (\boldsymbol{W}_m^{E}\boldsymbol{\Psi})^T(\boldsymbol{W}_m^{E}\boldsymbol{\Psi})$ for every m (otherwise taking $\boldsymbol{W}_m = \boldsymbol{W}_m^E = \boldsymbol{W}_m^{old}$ would lead to a lower value of $L_k^\lambda\left(\boldsymbol{\Phi}, \boldsymbol{D}\right) + \frac{\epsilon}{M}\left|\left|(\boldsymbol{W}_m\boldsymbol{\Psi})^T(\boldsymbol{W}_m\boldsymbol{\Psi}) \quad \boldsymbol{\Theta}\right|\right|^2$). Hence, we see that $\boldsymbol{W}_m = \boldsymbol{W}_m^E$ is a critical point for

$$f(\boldsymbol{W}_m) = L_k^\lambda\left(\boldsymbol{\Phi}, \boldsymbol{D}\right) + \frac{\epsilon}{M}\left|\left|(\boldsymbol{W}_m\boldsymbol{\Psi})^T(\boldsymbol{W}_m\boldsymbol{\Psi}) \quad \boldsymbol{\Theta}\right|\right|^2,$$

Since $\boldsymbol{W}_m^E$ is clearly a critical point for the map $\boldsymbol{W}_m$ *7* $\left|\left|(\boldsymbol{W}_m\boldsymbol{\Psi})^T(\boldsymbol{W}_m\boldsymbol{\Psi}) \quad \boldsymbol{\Theta}\right|\right|^2$, it follows that $\boldsymbol{W}_m = \boldsymbol{W}_m^E$ is also a critical point for $L_k^\lambda\left(\boldsymbol{\Phi}, \boldsymbol{D}\right)$. Since this holds for each $m = 1, \quad , M$ we see that $\mathsf{W}^E = \mathsf{f}\boldsymbol{W}_m^E\mathsf{g}_{m=1}$ is a critical point for $L_k^\lambda\left(\boldsymbol{\Phi}, \boldsymbol{D}\right)$. Equivalently, $\boldsymbol{\Phi}^E$ is a critical point for $L_k^\lambda\left(\boldsymbol{\Phi}, \boldsymbol{D}\right)$.

11 Proof of Theorem 2

In this section we prove Theorem 2 which connects the modular loss $L_k^{\lambda}(\boldsymbol{\Phi}, \mathsf{D})$, defined in Section 3, with the drop-module loss $L_{k,p}^{drop}(\boldsymbol{\Phi}, \mathsf{D})$ defined in Section 5.

Theorem 2. *The drop-module inner product loss at p is equivalent to the modular inner product loss at $\lambda = Mp/(1+p(M \quad 1))$. To be precise, for $\boldsymbol{\Phi}_0$ 2 F(H, M) we have*

$$\left.\frac{\partial L_{k,p}^{drop}(\boldsymbol{\Phi}, \mathsf{D})}{\partial \boldsymbol{\Phi}}\right|_{\boldsymbol{\Phi}=\boldsymbol{\Phi}_0} = p \quad \left.\frac{\partial L_k^{\lambda}(\boldsymbol{\Phi}, \mathsf{D})}{\partial \boldsymbol{\Phi}}\right|_{\boldsymbol{\Phi}=(p_{\overline{p/\lambda)}}\,\boldsymbol{\Phi}_0}.$$

Proof (Proof of Theorem 2). Recall from Section 5 that

$$L_{k,p}^{\mathrm{drop}}(\boldsymbol{\Phi}, \mathsf{D}) = \mathsf{E}\left\{L_k(\boldsymbol{\Phi}_{\boldsymbol{\eta}}, \mathsf{D}) : \boldsymbol{\eta} \quad B(p)^M\right\},$$

where

$$\begin{aligned} L_k(\boldsymbol{\Phi}_{\boldsymbol{\eta}}, \mathsf{D}) &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(h\boldsymbol{\Phi}_{\boldsymbol{\eta}}(\boldsymbol{x}_i), \boldsymbol{\Phi}_{\boldsymbol{\eta}}(\boldsymbol{x}_j)i \quad k(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)^2 \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{1}{M}\sum_{m=1}^{M}\eta_m \quad h\boldsymbol{\varphi}_m(\boldsymbol{x}_i), \boldsymbol{\varphi}_m(\boldsymbol{x}_j)i \quad k(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)^2. \end{aligned}$$

Now fix q 2 f1, , Mg and take a pair $\boldsymbol{x}_i, \boldsymbol{x}_j$ 2 D. Then,

$$\begin{aligned} &\frac{\partial}{\partial \boldsymbol{\varphi}_q(\boldsymbol{x}_i)}\left(\frac{1}{M}\sum_{m=1}^{M}\eta_m \quad h\boldsymbol{\varphi}_m(\boldsymbol{x}_i), \boldsymbol{\varphi}_m(\boldsymbol{x}_j)i \quad k(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)^2 \\ &= \frac{\partial h\boldsymbol{\varphi}_q(\boldsymbol{x}_i), \boldsymbol{\varphi}_q(\boldsymbol{x}_j)i}{\partial \boldsymbol{\varphi}_q(\boldsymbol{x}_i)} \quad \frac{\partial}{\partial h\boldsymbol{\varphi}_q(\boldsymbol{x}_i), \boldsymbol{\varphi}_q(\boldsymbol{x}_j)i}\left(\frac{1}{M}\sum_{m=1}^{M}\eta_m \quad h\boldsymbol{\varphi}_m(\boldsymbol{x}_i), \boldsymbol{\varphi}_m(\boldsymbol{x}_j)i \quad k(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)^2 \\ &= \boldsymbol{\varphi}_q(\boldsymbol{x}_j) \quad \frac{2\eta_q}{M}\left(\frac{1}{M}\sum_{m=1}^{M}\eta_m \quad h\boldsymbol{\varphi}_m(\boldsymbol{x}_i), \boldsymbol{\varphi}_m(\boldsymbol{x}_j)i \quad k(\boldsymbol{x}_i, \boldsymbol{x}_j)\right). \end{aligned}$$

It follows that

$$\begin{aligned} &\frac{\partial}{\partial \boldsymbol{\varphi}_q(\boldsymbol{x}_i)}\mathsf{E}\left[\left(\frac{1}{M}\sum_{m=1}^{M}\eta_m \quad h\boldsymbol{\varphi}_m(\boldsymbol{x}_i), \boldsymbol{\varphi}_m(\boldsymbol{x}_j)i \quad k(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)^2\right] \\ &= \boldsymbol{\varphi}_q(\boldsymbol{x}_j) \quad \frac{2p}{M}\left(\frac{1}{M}\sum_{m=1}^{M}p \quad h\boldsymbol{\varphi}_m(\boldsymbol{x}_i), \boldsymbol{\varphi}_m(\boldsymbol{x}_j)i + \frac{1}{M} \quad (1 \quad p) \quad h\boldsymbol{\varphi}_q(\boldsymbol{x}_i), \boldsymbol{\varphi}_q(\boldsymbol{x}_j)i \quad k(\boldsymbol{x}_i, \boldsymbol{x}_j)\right). \end{aligned}$$

Hence, from the definition of $L_{k,p}^{\text{drop}}(\boldsymbol{\Phi}, \mathsf{D})$, for each $\boldsymbol{x}_i \in \mathsf{D}$ and $q \in \{1, \ldots, M\}$ we have

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\varphi}_q(\boldsymbol{x}_i)} \left(L_{k,p}^{\text{drop}}(\boldsymbol{\Phi}, \mathsf{D}) \right) \\
&\quad = \frac{1}{N^2} \boldsymbol{\varphi}_q(\boldsymbol{x}_i) \left[\frac{2p}{M} \left(\frac{1}{M} \sum_{m=1}^{M} p \left[\langle \boldsymbol{\varphi}_m(\boldsymbol{x}_i), \boldsymbol{\varphi}_m(\boldsymbol{x}_i) \rangle + \frac{1}{M} \right] (1 - p) \left[\langle \boldsymbol{\varphi}_q(\boldsymbol{x}_i), \boldsymbol{\varphi}_q(\boldsymbol{x}_i) \rangle - k(\boldsymbol{x}_i, \boldsymbol{x}_i) \right] \right) \right. \\
&\qquad \left. + \frac{1}{N^2} \sum_{j=1}^{N} \boldsymbol{\varphi}_q(\boldsymbol{x}_j) \left[\frac{4p}{M} \left(\frac{1}{M} \sum_{m=1}^{M} p \left[\langle \boldsymbol{\varphi}_m(\boldsymbol{x}_i), \boldsymbol{\varphi}_m(\boldsymbol{x}_j) \rangle + \frac{1}{M} \right] (1 - p) \left[\langle \boldsymbol{\varphi}_q(\boldsymbol{x}_i), \boldsymbol{\varphi}_q(\boldsymbol{x}_j) \rangle - k(\boldsymbol{x}_i, \boldsymbol{x}_j) \right] \right) \right].
\end{aligned}$$

On the other hand,

$$\begin{aligned}
L_k^{\lambda}(\boldsymbol{\Phi}, \mathsf{D}) &= (1 - \lambda) \left[\frac{1}{M} \sum_{m=1}^{M} L_k(\boldsymbol{\varphi}_m, \mathsf{D}) + \lambda \left[L_k\left(\overline{\Phi}, \mathsf{D}\right) \right] \right. \\
&= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left((1 - \lambda) \left[\frac{1}{M} \sum_{m=1}^{M} \left(\langle \boldsymbol{\varphi}_m(\boldsymbol{x}_i), \boldsymbol{\varphi}_m(\boldsymbol{x}_j) \rangle - k(\boldsymbol{x}_i, \boldsymbol{x}_j) \right)^2 \right] \right. \\
&\quad \left. \left. + \lambda \left[\left(\frac{1}{M} \sum_{m=1}^{M} \langle \boldsymbol{\varphi}_m(\boldsymbol{x}_i), \boldsymbol{\varphi}_m(\boldsymbol{x}_j) \rangle - k(\boldsymbol{x}_i, \boldsymbol{x}_j) \right)^2 \right] \right).
\end{aligned}$$

Hence for each $\boldsymbol{x}_i \in \mathsf{D}$ and $q \in \{1, \ldots, M\}$ we have

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\varphi}_q(\boldsymbol{x}_i)} \left(L_k^{\lambda}(\boldsymbol{\Phi}, \mathsf{D}) \right) \\
&\quad = \frac{1}{N^2} \boldsymbol{\varphi}_q(\boldsymbol{x}_i) \left[\frac{2}{M} \left(\frac{1}{M} \sum_{m=1}^{M} \lambda \left[\langle \boldsymbol{\varphi}_m(\boldsymbol{x}_i), \boldsymbol{\varphi}_m(\boldsymbol{x}_i) \rangle + (1 - \lambda) \left[\langle \boldsymbol{\varphi}_q(\boldsymbol{x}_i), \boldsymbol{\varphi}_q(\boldsymbol{x}_i) \rangle - k(\boldsymbol{x}_i, \boldsymbol{x}_i) \right] \right) \right. \right. \\
&\qquad \left. \left. + \frac{1}{N^2} \sum_{j=1}^{N} \boldsymbol{\varphi}_q(\boldsymbol{x}_j) \left[\frac{4}{M} \left(\frac{1}{M} \sum_{m=1}^{M} \lambda \left[\langle \boldsymbol{\varphi}_m(\boldsymbol{x}_i), \boldsymbol{\varphi}_m(\boldsymbol{x}_j) \rangle + (1 - \lambda) \left[\langle \boldsymbol{\varphi}_q(\boldsymbol{x}_i), \boldsymbol{\varphi}_q(\boldsymbol{x}_j) \rangle - k(\boldsymbol{x}_i, \boldsymbol{x}_j) \right] \right) \right. \right. \right].
\end{aligned}$$

Thus, with $\lambda = Mp/(1 + p(M - 1))$ and $\boldsymbol{\Phi}_0 = \left\{ \boldsymbol{\varphi}_m^0 \right\}_{m=1}^{M}$ we have

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\varphi}_q(\boldsymbol{x}_i)} \left(L_k^{\lambda}(\boldsymbol{\Phi}, \mathsf{D}) \right) \Big|_{\boldsymbol{\Phi} = (\sqrt{p/\lambda})\, \boldsymbol{\Phi}_0} \\
&\quad = \frac{1}{N^2} \boldsymbol{\varphi}_q^0(\boldsymbol{x}_i) \left[\frac{2}{M} \left(\frac{1}{M} \sum_{m=1}^{M} p \left[\langle \boldsymbol{\varphi}_m^0(\boldsymbol{x}_i), \boldsymbol{\varphi}_m^0(\boldsymbol{x}_i) \rangle + \frac{1}{M} (1 - p) \left[\langle \boldsymbol{\varphi}_q^0(\boldsymbol{x}_i), \boldsymbol{\varphi}_q^0(\boldsymbol{x}_i) \rangle - k(\boldsymbol{x}_i, \boldsymbol{x}_i) \right] \right) \right. \right. \\
&\qquad \left. \left. + \frac{1}{N^2} \sum_{j=1}^{N} \boldsymbol{\varphi}_q^0(\boldsymbol{x}_j) \left[\frac{4}{M} \left(\frac{1}{M} \sum_{m=1}^{M} p \left[\langle \boldsymbol{\varphi}_m^0(\boldsymbol{x}_i), \boldsymbol{\varphi}_m^0(\boldsymbol{x}_j) \rangle + \frac{1}{M} \right] (1 - p) \left[\langle \boldsymbol{\varphi}_q^0(\boldsymbol{x}_i), \boldsymbol{\varphi}_q^0(\boldsymbol{x}_j) \rangle - k(\boldsymbol{x}_i, \boldsymbol{x}_j) \right] \right) \right. \right].
\end{aligned}$$

Combining this with the above computation of $\partial\left(L_{k,p}^{\mathrm{drop}}(\boldsymbol{\Phi}, \mathsf{D})\right)/\partial\boldsymbol{\varphi}_q(\boldsymbol{x}_i)$ gives

$$\left.\frac{\partial\left(L_{k,p}^{\mathrm{drop}}(\boldsymbol{\Phi}, \mathsf{D})\right)}{\partial\boldsymbol{\varphi}_q(\boldsymbol{x}_i)}\right|_{\boldsymbol{\Phi}=\boldsymbol{\Phi}_0} = p \left.\frac{\partial\left(L_k^{\lambda}(\boldsymbol{\Phi}, \mathsf{D})\right)}{\partial\boldsymbol{\varphi}_q(\boldsymbol{x}_i)}\right|_{\boldsymbol{\Phi}=(\sqrt{p/\lambda})\,\boldsymbol{\Phi}_0}.$$

Since this holds for all $\boldsymbol{x}_i \in \mathsf{D}$ and $q \in \{1, \ldots, M\}$, this completes the proof of the theorem.

12 Experimental results

12.1 The data sets

We used the following eight data sets in our experiments.

Background The MNIST + Background images data set from Larochelle et al. [17]:
10 classes, 784 features per example, 12000 training examples, 50000 test examples.

Convex The Convex data set from Larochelle et al. [17]:
2 classes, 784 features per example, 8000 training examples, 50000 test examples.

MNIST The original MNIST data set as used in Larochelle et al. [17]:
10 classes, 784 features per example, 12000 training examples, 50000 test examples.

NUS Wide The NUS-Wide-Object data set from Chua et al. [5] with all extracted features concatenated: 31 classes, 639 features per example, 17928 training examples, 12072 test examples.

NVDR The NVDR data set from Wu et al. [28] with all extracted features concatenated:
24 classes, 418 features per example, 6438 training examples, 6429 test examples.

Random The MNIST + Random background data set from Larochelle et al. [17]:
10 classes, 784 features per example, 12000 training examples, 50000 test examples.

Rectangles The Rotated MNIST digits data set from Larochelle et al. [17]:
2 classes, 784 features per example, 1200 training examples, 50000 test examples.

Rotations The Rotated MNIST digits data set from Larochelle et al. [17]:
10 classes, 784 features per example, 12000 training examples, 50000 test examples.

12.2 Test time and the number of dimensions per module (H)

The following figures show the test time for both image retrieval and classification as a function of the number of dimensions per module H. Note that $H \quad M = 300$ throughout, so M is inversely proportional to H. Test time for classification increases with H as this corresponds to the number of dimensions per module, and the predictions for each module may be made in parallel. The aggregation of predictions corresponds to taking the most common label from a list, which

adds little relative overhead. In contrast, the test time for image retrieval is a non-monotonic function of the number of dimensions per module - it typically falls for very small H, before rising again as H grows. Recall from Section 6.2 that the computation of nearest neighbours for image retrieval consists of two steps. First a collection κ of nearest neighbours $I_{\kappa,n}^{\varphi_m}(\boldsymbol{x})$ is computed in each H-dimensional space separately. Second, the sets $I_{\kappa,n}^{\varphi_m}(\boldsymbol{x})$ are combined and a subset $I_{\kappa,n}^{\boldsymbol{\Phi}}(\boldsymbol{x}) \quad \bigcup_{m=1}^{M} I_{\kappa,n}^{\varphi_m}(\boldsymbol{x})$ of size κ is extracted so as to minimise the average squared distance $(1/M) \ \sum_{m=1}^{M} k\varphi_m(\boldsymbol{x}_q) \quad \varphi_m(\boldsymbol{x})k_2^2$. The first step is easily distributable over the M modules. Hence, the time taken is monotonoically increasing in H. The second stage is not straightforwardly distributable and the size of the search space $\bigcup_{m=1}^{M} I_{\kappa,n}^{\varphi_m}(\boldsymbol{x})$ will typically grow with M. This adds an additional overhead which accounts for fact that image retrieval test time can typically be seen to fall with H when H is small (ie. when M is large). Note that the size of this additional overhead is bounded by $M \ \kappa$, where as the time taken by the first stage, consisting of multiple nearest neighbour searches done in parallel, increases with the size of the training set. Hence, the effect of the additional overhead for small values of H is at its most dramatic when the size of the training set is small, such as for the Rectangles data set, where the training set consists of just 1200 images.

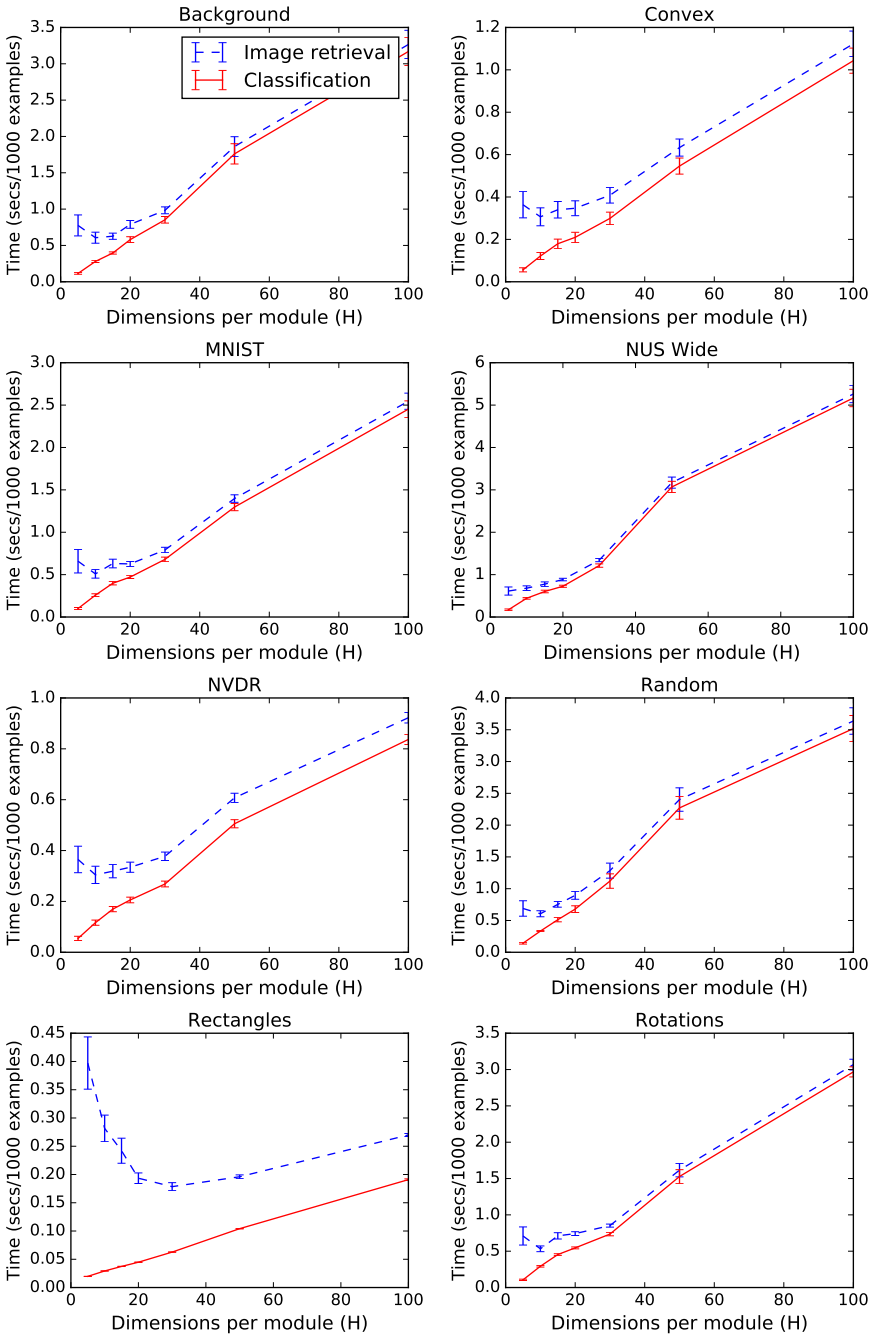

Fig. 6. Classification accuracy as a function of H (See Section 6.2).

12.3 Image retreival precision and the number of dimensions per module (H)

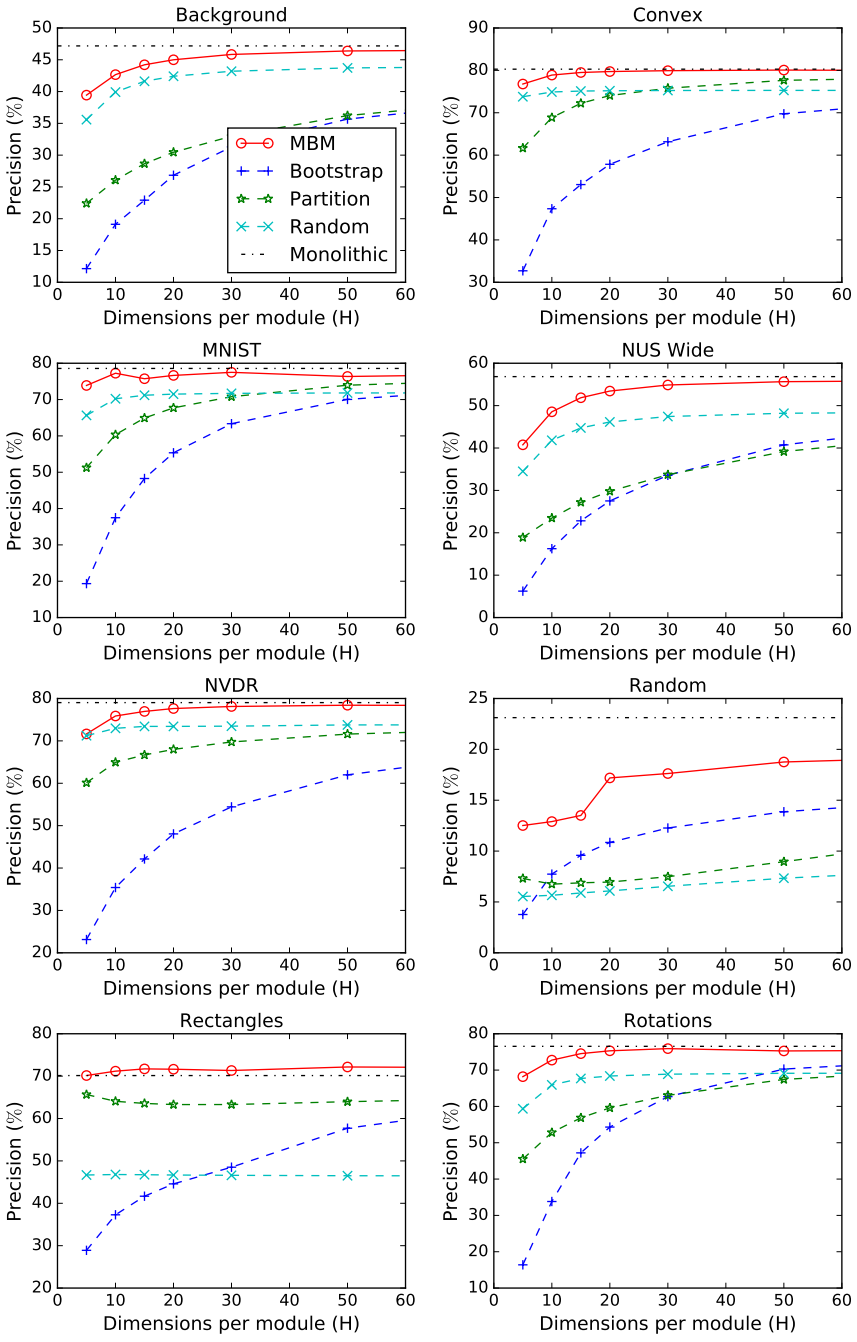

Fig. 7. Precision as a function of H (See Section 6.2).

12.4 Image retreival precision and the diversity parameter

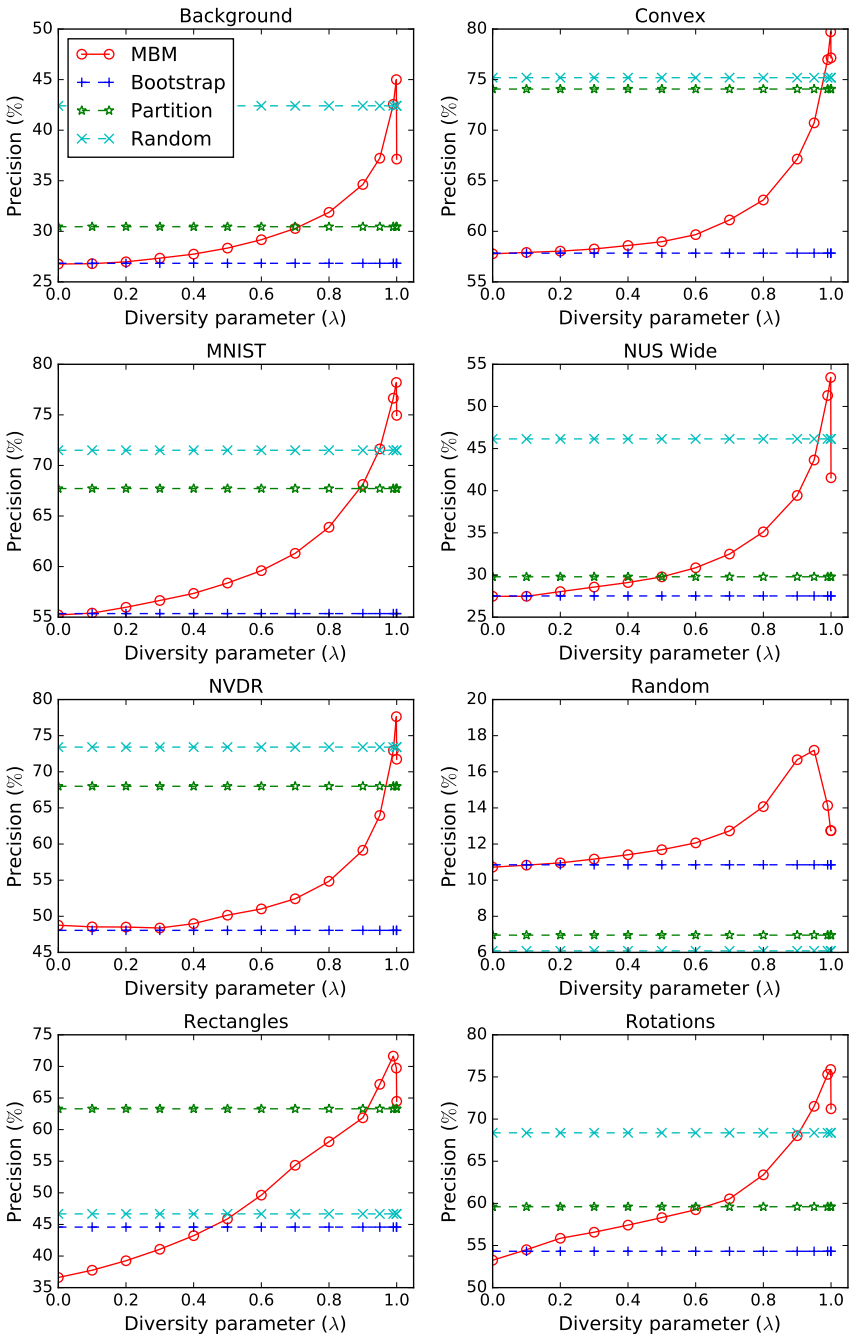

Fig. 8. Precision as a function of the diversity parameter λ (See Section 6.2).

12.5 Classification accuracy and the number of dimensions per module

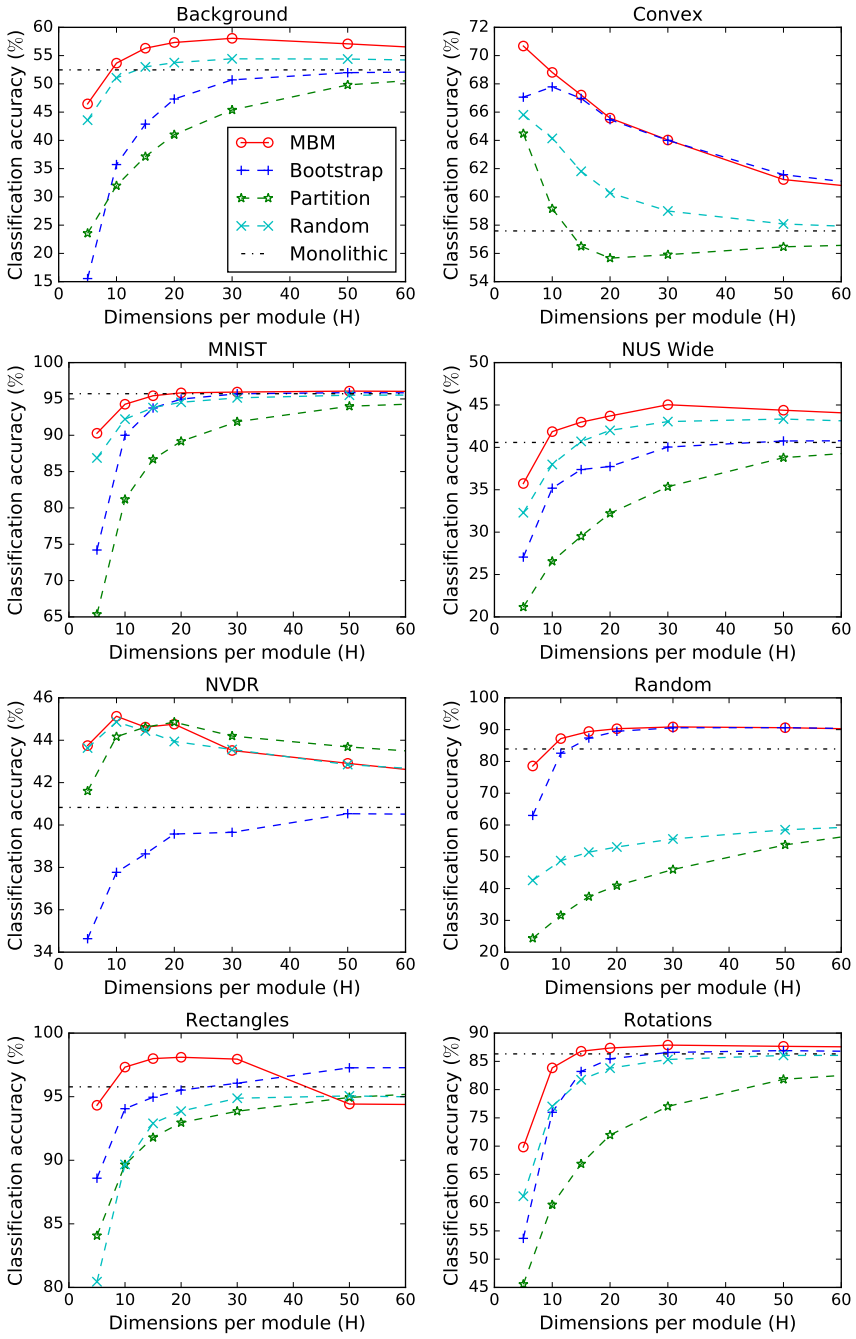

Fig. 9. Classification accuracy as a function of H (See Section 6.2).

12.6 Classification accuracy and the diversity parameter

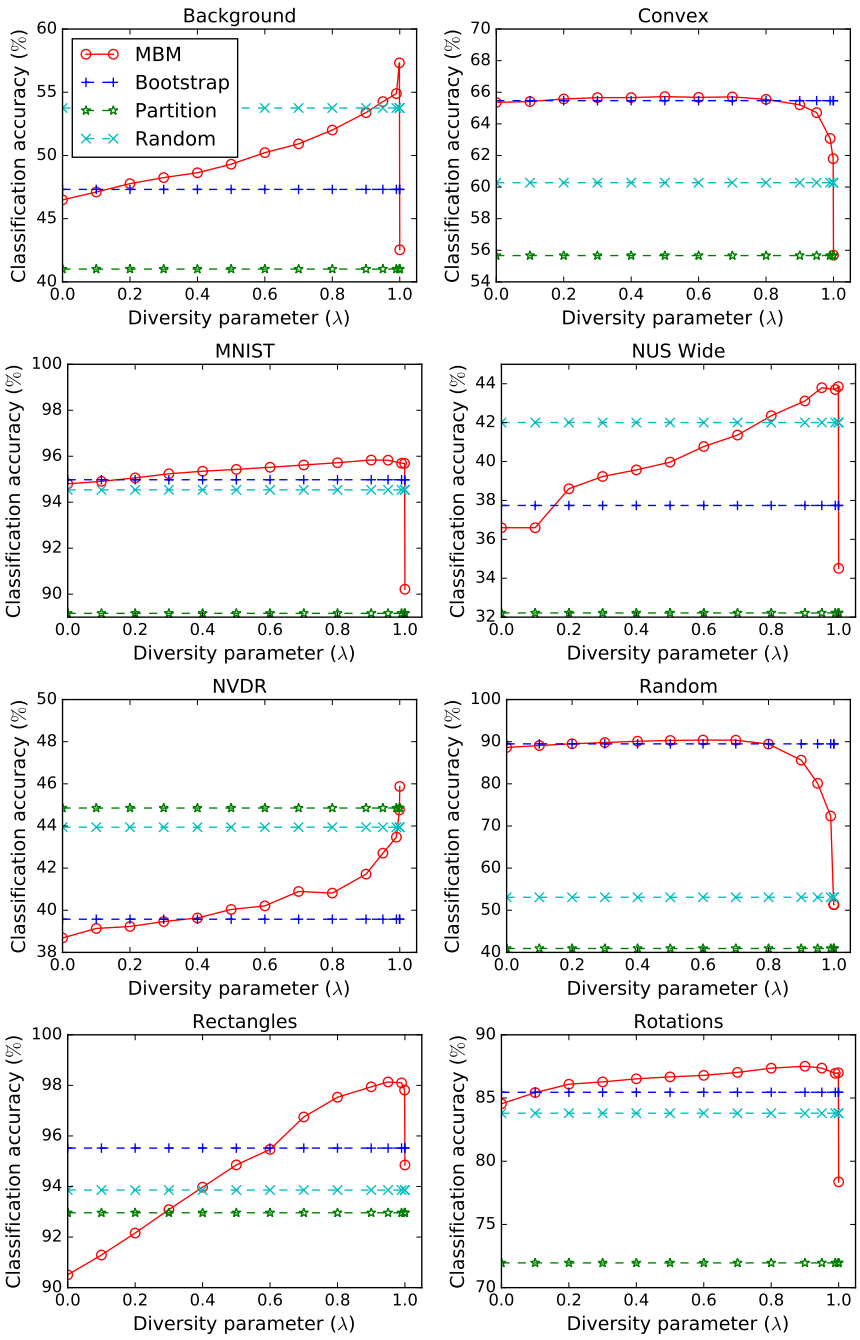

Fig. 10. Classification accuracy as a function of H (See Section 6.2).