Thank you for the reviews and suggestions; we will revise accordingly.

Reviewer 1
On the 14 references: Thank you for the suggestion, the period before the list was a typo and the list of references were examples of dimension reduction algorithms supporting the 1st sentence. We will condense to fewer citations.

On 10-Fold: Thank you for pointing out this potential for misunderstanding. For each of the 10-fold experiments, we trained W and the SVM classifier only on the training set while reporting the result only on the test set. The test set was never used during the training; training and testing data are strictly separated. We repeat this process for each fold of cross-validation. We will make sure to clarify the training/testing process in the camera ready version.

On additional experiments: There are 4 central claims of the paper. First, ISM can be generalized to many learning paradigms (supervised, semi-supervised, unsupervised). Second, ISM is significantly faster than existing optimization algorithms for IKDR. Third, ISM can generalize to multiple kernels with the appropriate Φ matrix. Lastly, the ISM family extends to conic combinations of kernels in the ISM family.

Like the first reviewer suggested, we conducted experiments with PCA and KPCA in a preliminary draft. Against PCA, IKDR significantly reduced the error rate while maintaining similar execution time across all datasets. Against KPCA, IKDR simultaneously reduced the error rate and the execution time. Although these results were interesting, when trimming the submission version we decided to focus the experiments on supporting the 4 central claims of the paper. We will add these results in our supplement on final revision.

Reviewer 2
On the algorithm seem straightforward: Our contribution focuses on the theoretical aspects of ISM for solving IKDR for a general class of kernels and our experiments are designed to support the claims. The ISM algorithm (which is an iterative eigendecomposition) itself is proposed previously by [19] and is not a part of our claim of contribution. We found the ISM algorithm simple and elegant, but its algorithmic simplicity hides the rigorous analysis required to guarantee its effectiveness. The sizable proofs provided in the appendix is a testament of the complexity involved to further extend the theoretical guarantees of ISM to other kernels for solving highly non-convex IKDR problems.

In addition to the theoretical contributions, another novelty of this work is the discovery of the Φ matrices *unique* to each ISM kernel. Indeed, every kernel within the ISM family possesses an alternative representation that is significantly smaller and yet contains all the necessary information. For IKDR applications, we discovered a matrix Φ that can replace kernels while significantly reducing its computational complexity. We consider discovering an alternative efficient representation to kernels of great interest for the IKDR community.

On experiments using toy data: All of the 5 datasets are real and not synthetically generated. They were carefully chosen for convenient comparison since many related work also used the same datasets. They were also carefully chosen to span different data types. The Flower and Face datasets are real images commonly used for alternative clustering. The cancer and wine datasets are used for many supervised and unsupervised domains while having both discrete and continuous features respectively. The MNIST handwritten digit dataset is a standard benchmark researchers use to compare classification performance. See citations 33-37.

Reviewer 3
On clarifying line 153: We apologize for the confusion. The language in this paragraph is commonly used in Spectral Clustering from Graph Theory. For a given matrix Ψ, its degree matrix is $D_\Psi = \text{Diag}(\Psi 1_n)$ where 1_n is a vector of 1s and Diag is a function that places the elements of a vector into the diagonal of a zero squared matrix. The Laplacian matrix is $\mathcal{L} = D_\Psi - \Psi$. The connection to graph theory is that Ψ is treated as a weighted adjacency matrix, and the Laplacian $\mathcal{L}$ shows up in many resulting formulas. We'll provide additional information to clarify the language.

On the intuition of the Φ matrix: Observe from Table 3 that Φ can be expressed as $X^T \Omega X$, where Ω is a positive semi-definite (PSD) matrix. Removing Ω, $X^T X$ is simply the covariance matrix. Applying Cholesky decomposition on Ω yields $(X^T L)(L^T X)$. Since the PSD Ω matrix is constructed from the kernel and label information, Ω is an interpretable way to scale the covariance matrix using the kernel and label information. We'll add this intuition.

On matrix A: Thank you for finding this oversight, we will include its definition within the paper.

On Definition 1: Given data $x_i \in \mathbb{R}^d$, $a(x_i, x_j)$ is a function defined as $a : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$. As a short hand, $\mathbf{a}$ is the image of a. From our analysis, a can be any function that allows a kernel function to be written in terms of $f(\beta)$ while remaining a valid kernel. The definition of $a(x_i, x_j)$ depends on the kernel shown in Table 1. b and $\mathbf{b}$ are defined the same way. We thank the reviewer for this point of confusion and we will address these definitions in the final release.

On Table 4: Thank you for noticing the minor typo. Instead of 47%, it is supposed to be 0.47. We will make this update accordingly. For the wine dataset, the bold text highlights the best performing results. Since 0.86 and 0.84 are all the best results from Gaussian and Polynomial, they were all highlighted.