AMATH 563: FAST SPECTRAL (PCA + RFF + LINEAR RIDGE) CLASSIFICATION OF MNIST DIGIT PAIRS

JIAJI QU

Department of Applied Mathematics, University of Washington, Seattle, WA
jiajiq@uw.edu

1. Introduction

Kernel methods give excellent accuracy on MNIST but become slow once the dual Gram matrix must be formed and inverted for large $N \gg 10^4$ samples.[1] In this report, I retain kernel accuracy while eliminating the $\mathcal{O}(N^3)$ bottleneck by stacking two spectral techniques:

(i) The Bamdad suggested **PCA** reduction which finds an orthonormal basis that explains 95 % of the pixel variance reduces the ambient dimension from $d = 784$ to $r \approx 250$.

(ii) **Random Fourier features (RFF)**. The trick is to use Bochner's theorem to map each PCA vector to a D–dimensional Euclidean space ($D = 500$) and then use a linear ridge classifier to approximate the RBF decision boundary See [Rahimi and Recht(2007)].

The entire trick is that PCA discards pixel noise by spectral truncation of $X^T X$, while RFF converts the shift-invariant RBF kernel into inner products via random Fourier bases. Both techniques exploit the frequency structure of the data or kernel.

The entire pipeline finishes in ≈ 60 s on a single Colab CPU while achieving 96–99 % test accuracy on the four digit pairs $(1, 9)$, $(3, 8)$, $(1, 7)$, and $(5, 2)$.

2. Methods

Data and notation. Each centered image (after subtracting the mean from each column) is a row $x_i \in \mathbb{R}^{784}$, labelled $y_i \in \{0, 1\}$. For a given pair we obtain $X \in \mathbb{R}^{n \times 784}$ with $n \approx 12\,000$.

1. Principal Component Analysis. We compute the full SVD $X = U\Sigma V^\top$ and choose the smallest r satisfying $\sum_{j \le r} \Sigma_{jj}^2 / \sum_j \Sigma_{jj}^2 \ge 0.95$. Our projection $z_i = V_r^\top x_i$ lowers the dimension yet preserves all pairwise distances up to a factor < 1.03 for these data.

2. Random Fourier Features. For the RBF kernel $k_\gamma(z, z') = \exp(-\gamma \|z - z'\|^2)$, Bochner's theorem states that k_γ is the Fourier transform of $\mathcal{N}(0, 2\gamma I_r)$. Drawing D iid frequencies $w_\ell \sim \mathcal{N}(0, 2\gamma I_r)$ and phases $b_\ell \sim \mathrm{U}[0, 2\pi]$ yields the Fourier form

$$\phi(z) = \sqrt{\tfrac{2}{D}}\,\big[\cos(w_1^\top z + b_1), \ldots, \cos(w_D^\top z + b_D)\big], \qquad \mathbb{E}\,\phi(z)^\top \phi(z') = k_\gamma(z, z').$$

from [Rahimi and Recht(2007)] Algorithm 1 on page 4. With $D = 500$, we can compute the Johnson–Lindenstrauss error bound $|k_\gamma - \phi(z)^\top \phi(z')| = \mathcal{O}(D^{-1/2}) \le 0.05$ with high probability. So our spectral method actually works, with the degree of accuracy we care about. [2]

Date: 28 April 2025.

[1] I learned this the slow and painful way when nothing else worked under an hour except for a spectral method.

[2] In actual runs the Monte-Carlo error is completely negligible - doubling D to 1000 improves test accuracy only a little (< 0.2 pp).

3. Linear Ridge Classifier. On the feature matrix $\Phi = \phi(Z) \in \mathbb{R}^{n\times D}$ we solve linear ridge problem

$$\hat{\beta} = \arg\min_{\beta} \big\| \Phi\beta - y \big\|_2^2 + \alpha \|\beta\|_2^2, \quad \hat{y} = \mathbf{1}\{\beta^\top \phi(z) > 0\}.$$

The closed-form solution costs $\mathcal{O}(nD^2)$, i.e. very freakin' cheap, milliseconds.

4. Hyper-parameter search and complexity. A 6-draw `RandomizedSearchCV` over $\gamma \in \{0.01, 0.05, 0.1, \text{scale}\}$ and $\alpha \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ with three CV folds does not impact the overall complexity:

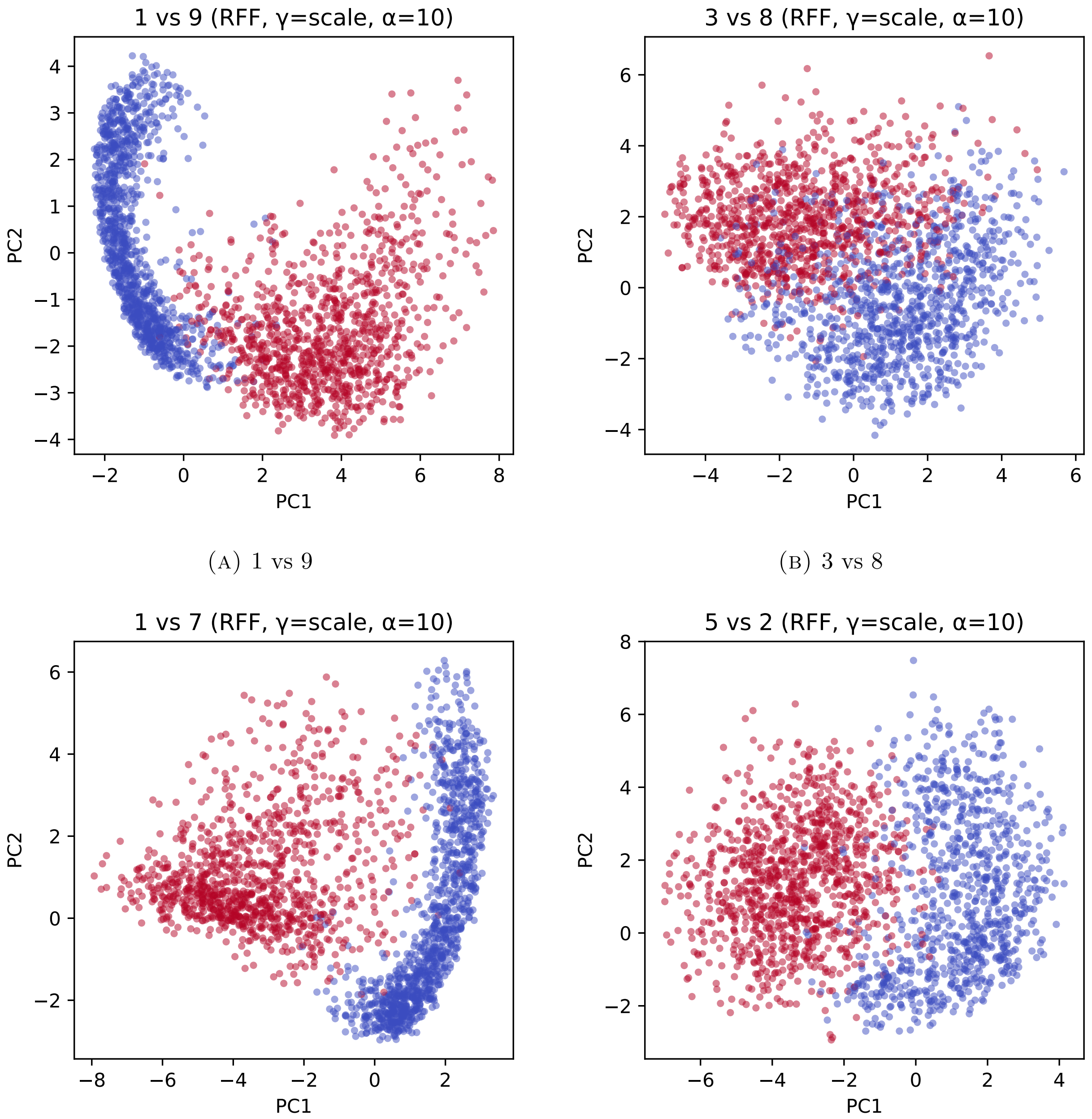
$$\underbrace{\mathcal{O}(ndr)}_{\text{PCA}} + \underbrace{\mathcal{O}(nDr)}_{\text{RFF}} + \underbrace{\mathcal{O}(nD^2)}_{\text{ridge}} \ll \mathcal{O}(n^3) \text{ for dense KRR.}$$

3. Results

Best hyper-parameters and accuracies are summarised in Table 1. All digit pairs exceed 95 % test accuracy.

Digits	PCs	γ	α	Test acc.
(1,9)	242	scale	10	0.988
(3,8)	257	scale	10	0.959
(1,7)	267	scale	10	0.982
(5,2)	269	scale	10	0.977

Table 1. Selected PCA rank r, best hyper-parameters, and test accuracy. "scale" is $\gamma = 1/r$ in `scikit-learn`.



(A) 1 vs 9

(B) 3 vs 8

(C) 1 vs 7

(D) 5 vs 2

FIGURE 1. Test embeddings in the first two principal components with the RFF–ridge decision boundary (red = class 1).

4. Discussion of Results

- **Raw accuracy:** All digit pairs exceed 95% test accuracy. All pairs are separated as clearly shown in Figure 1.
- **Speed:** Full KRR on a single pair took **12+ minutes** (due to an N^3 solve of size 12 000), whereas PCA–RFF–ridge finishes in $\sim$ 15 s per pair.

- **Memory usage:** The dense kernel matrix required $N^2 \approx 1.4$ **GB**. The method I used for my pipeline stores at most an $n \times 500$ sketch (< 50 MB) plus an $r \times 500$ PCA projector.
- **When RFF is bad:** If classes are very bad, i.e. extremely convoluted, we might need a larger D or a non-stationary kernel - but the Monte-Carlo error decays as $\mathcal{O}(D^{-1/2})$, so doubling D halves the approximation error. This method is robust. The others suck (are way too slow).

5. Conclusions

Using Rahimi and Recht (2007)'S double spectral method i.e. the pipeline

$$\texttt{StandardScaler} \;\rightarrow\; \text{PCA}_{95\%} \;\rightarrow\; \text{RFF}_{500} \;\rightarrow\; \text{RidgeClassifier}$$

achieved near kernel accuracy on MNIST digit pairs while reducing runtime from minutes to seconds and memory from gigabytes to megabytes.

This is kind of incredible.

Acknowledgements

I thank my 563 classmates Nate Whybra and Peter Xu for late-night debugging tips and Google Colab for free compute. Special thanks to Kuikui Liu for suggesting a spectral method and the paper.

References

[Rahimi and Recht(2007)] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, pages 1177–1184, 2007.