Random Features for Sparse Signal Classification

Jen-Hao Rick Chang*, Aswin C. Sankaranarayanan, B. V. K. Vijaya Kumar
Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh, PA

Abstract

Random features is an approach for kernel-based inference on large datasets. In this paper, we derive performance guarantees for random features on signals, like images, that enjoy sparse representations and show that the number of random features required to achieve a desired approximation of the kernel similarity matrix can be significantly smaller for sparse signals. Based on this, we propose a scheme termed compressive random features that first obtains low-dimensional projections of a dataset and, subsequently, derives random features on the low-dimensional projections. This scheme provides significant improvements in signal dimensionality, computational time, and storage costs over traditional random features while enjoying similar theoretical guarantees for achieving inference performance. We support our claims by providing empirical results across many datasets.

1. Introduction

Random features [26,37,38,43,46] is an approach to perform kernel-based inference on very large datasets. In the traditional kernel approach, we need to construct the kernel-similarity matrix whose storage and computational time are quadratic in the size of the dataset; the quadratic dependence on the size of the dataset makes the approach infeasible for Big Data scenarios. Random features addresses this problem by explicitly constructing finite-dimensional random features from the data such that inner products between the random features approximate the kernel functions. Inference with random features achieves comparable performance as those of the kernel-based ones while enjoying the scalability of linear inference methods. Recently, it also achieves performance comparable to a convolutional neural network on datasets like the ImageNet [13].

We show that for signals enjoying sparse representations (either canonically or in a transform basis), the performance guarantees of random features can be significantly strengthened. Specifically, we prove that the dimension of random features required to approximate a stationary kernel function [42] dominantly depends on the signal sparsity instead of the ambient dimension. For images, whose ambient dimension is often far greater than their sparsity, our analysis greatly improves the theoretical bounds of random features.

We next show that both computational and storage costs of random features applied to sparse signals can be significantly improved by first performing a dimensionality reduction using random projection and subsequently, applying random features to the dimensionality-reduced signals. There are several advantages to this scheme. First, we show that the theoretical guarantees in approximating the original kernel function are similar to that of random features applied on sparse signals. This means that the additional dimensionality reduction step does not hinder our ability to approximate kernel functions. Second, the dimensionality reduction can be performed optically with compressive cameras [19]. In regimes where sensing is costly, (for example, short-wave infrared and midwave infrared), the use of compressive cameras enables sensing with low-resolution sensors [12, 32] with associated savings in cost of the camera. Third, in the context of compressive imaging, inference tasks such as classification and detection are often simpler than recovery [1, 31, 41] and hence, we can expect to use high compression ratios in the dimensionality reduction step. In our experiments, we are able to achieve $10 - 30\times$ compression with little loss in classification accuracy.

Contributions. In this paper, we propose a scheme called *compressive random features* that applies random features on compressive measurements of signals that enjoy sparse representations either canonically or in a transform basis (see Figure 1). Our contributions are three-fold:

- We prove that the number of random features required to accurately approximate the kernel function depends predominantly on the sparsity of the signals.
- We show that random features applied to dimensionality-reduced signals or equivalently, compressive measurements, are able to approximate isometric kernel functions of the original uncompressed data and provide analytical guarantees that bound the loss in performance.

*This work was supported by the ARO Grant W911NF-15-1-0126.

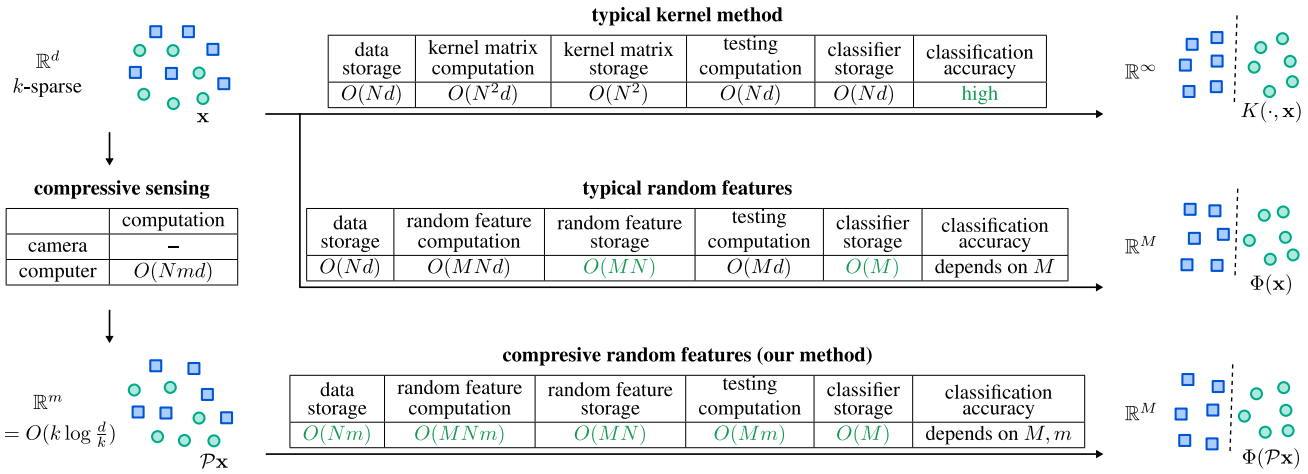


Figure 1: Overview of typical kernel methods, random features, and our compressive random features. N is the number of training samples, M is the dimension of the random features, d is the dimension of the original uncompressed data, and m is the dimension of the compressive measurements. In testing computations of typical and compressive random features schemes, we include the cost to construct random features and to apply classifiers for one test input.

- We also observe that our proposed scheme for compressive inference offers comparable classification performance, across many datasets, to similar approaches applied directly on the original data while providing reduced computational time and storage.

2. Related work

Notations. We use d as the dimension of original uncompressed signals, m as the dimension of compressive measurements, M as the dimension of random features, and N as the number of training samples. We use lowercase boldface letters to denote vectors and uppercase letters to denote matrices. We say a signal is k-sparse if it has at most k nonzero entries. All norms in this paper are ℓ_2-norm, denoted by $\|\cdot\|$. The element-wise complex conjugate of a vector $\mathbf{x}$ is written as $\overline{\mathbf{x}}$. We define the diameter and the radius of a set $\mathcal{X}$ as

$$\text{diam}(\mathcal{X}) = \max_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \|\mathbf{x}-\mathbf{y}\|,$$

$$\text{radius}(\mathcal{X}) = \max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|.$$

2.1. Random feature method

A hallmark of kernel-based inference is the development of kernel trick, which utilizes kernel function to efficiently evaluate similarity in infinitely high dimensional spaces; thereby, kernel-based inference is capable of approximating any decision boundary or function provided we have sufficient training samples [42]. Despite this attractive ability, kernel methods are prohibitive for large datasets because of their high storage and time complexities during both the training and testing phases. Specifically, with N training samples, kernel trick usually requires computing and storing a kernel matrix whose size is $N \times N$. Testing a single input requires evaluating kernel function between the input and a large portion of training samples [26].

The goal of random features [38] is to achieve a scalable implementation of kernel methods. Given a stationary kernel function $K(\mathbf{x},\mathbf{y}) = f(\mathbf{x}-\mathbf{y}) := f(\boldsymbol{\delta})$, its Fourier transform, $p(\boldsymbol{\omega})$, has only nonnegative entries [40] due to the positive definiteness of the kernel and hence, can be treated as a probabilistic density function. The inverse Fourier transform of the kernel function is given as

$$K(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d} p(\boldsymbol{\omega}) e^{j\boldsymbol{\omega}^\top(\mathbf{x}-\mathbf{y})} d\boldsymbol{\omega} = E_p[\phi_{\boldsymbol{\omega}}(\mathbf{x})\overline{\phi_{\boldsymbol{\omega}}(\mathbf{y})}], \quad (1)$$

where d is the dimension of the data, and $\phi_{\boldsymbol{\omega}}(\mathbf{x}) := e^{j\boldsymbol{\omega}^\top\mathbf{x}}$. The sample mean $\frac{1}{M}\sum_{i=1}^M \phi_{\boldsymbol{\omega}_i}(\mathbf{x})\overline{\phi_{\boldsymbol{\omega}_i}(\mathbf{y})}$ is thus an unbiased estimator of $K(\mathbf{x},\mathbf{y})$ when $\{\boldsymbol{\omega}_i\}$ are i.i.d. samples from p. Since $f(\boldsymbol{\delta})$ and $p(\boldsymbol{\omega})$ are real, we can reduce $\phi_{\boldsymbol{\omega}}(\mathbf{x})$ and define a real-valued random feature generating function, $\Phi : \mathbb{R}^d \to \mathbb{R}^M$, as

$$\Phi(\mathbf{x}) = \sqrt{\frac{2}{M}} \left[\cos(\boldsymbol{\omega}_1^\top\mathbf{x}+b), \cdots, \cos(\boldsymbol{\omega}_M^\top\mathbf{x}+b)\right]^\top, \quad (2)$$

where $\boldsymbol{\omega}_i$ is drawn from the distribution p and b is drawn uniformly from $[0, 2\pi]$. For the commonly used Gaussian kernel $K(\mathbf{x},\mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|^2/(2\sigma^2))$, $p(\boldsymbol{\omega}) = \mathcal{N}(0, \sigma^{-2}I_d)$, where I_d is the $d \times d$ identity matrix.

Rahimi and Recht [38] showed that the inner product of random features uniformly converges to $K(\mathbf{x},\mathbf{y})$ in probability. In particular, when training samples are from a compact set $\mathcal{X} \subset \mathbb{R}^d$, in order to have $\mathbb{P}\left(\sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}} |\langle\Phi(\mathbf{x}),\Phi(\mathbf{y})\rangle - K(\mathbf{x},\mathbf{y})| > \epsilon\right)$ less than a

constant q, the dimension of random features

$$M = O\left(\frac{d}{\epsilon^2}\log\frac{\sigma_p \operatorname{diam}(\mathcal{X})}{q\epsilon}\right), \tag{3}$$

where $\sigma_p^2 := E_p(\boldsymbol{\omega}^\top\boldsymbol{\omega})$ is the second moment of $p(\boldsymbol{\omega})$.

2.2. Compressive sensing and compressive inference

Compressive sensing [3] aims to sense a high-dimensional signal from a low-dimensional measurements. Specifically, any d-dimensional, k-sparse signal can be exactly recovered from its m-compressive measurements, provided $m = O(k\log\frac{d}{k})$.

One of the main results in CS is the restricted isometry property (RIP) [7, 8] which suggests that distances between sparse signals are approximately preserved by certain measurement matrices, including random projections and partial Fourier matrices [39]. A $m \times d$ matrix $\mathcal{P}$ satisfies RIP (of order $2k$) if for all k-sparse signals $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have $(1-\delta)\|\mathbf{x}-\mathbf{y}\|^2 \leq \|\mathcal{P}\mathbf{x}-\mathcal{P}\mathbf{y}\|^2 \leq (1+\delta)\|\mathbf{x}-\mathbf{y}\|^2$ with some $\delta \in (0,1)$. This means that all pairwise distances between k-sparse signals are approximately preserved after projected by $\mathcal{P}$. Sub-Gaussian random matrices and random orthoprojectors are known to satisfy RIP with high probability [9, 10]. To generate a $m \times d$ random orthoprojector, we first i.i.d. sample its entries from a zero-mean Gaussian or Bernoulli distribution. Then we run the Gram-Schmidt process row-wise (assuming its rows are linearly independent) and multiply the result by $\sqrt{d/m}$.

The approximate preservation of distances enables inference directly in the compressive domain. This idea — termed *compressive inference* — has resulted in many theoretical and practical algorithms for estimation [4, 25, 33, 41] and classification [1, 6, 11, 15, 16, 23, 28, 31, 36] without the need of an intermediate reconstruction step. This saves the computation required to recover the signals and thus, lowers the computational and memory requirements of the inference algorithm.

The goal of this paper is to provide theoretical guarantees for applying random features onto compressive measurements. We can, therefore, perform non-linear inference on compressive inference without sacrificing its benefits — low time and storage complexities.

3. Random features for sparse signals

The theoretical guarantee of random features provided in [38] is for generic datasets and does not exploit any model on the data. If we know that our signals enjoys sparse representations, either canonically or in some transform basis, can we tighten the bound required for approximating a kernel function? We address this question in this section.

The following theorem characterizes the performance of random features approximating stationary kernel functions for signals that enjoy sparse representations.

Theorem 1. *(Fourier random feature with k-sparse data) Let $\mathcal{X}$ be a compact set of k-sparse vectors in $\mathbb{R}^d$. Let the random features for a stationary kernel function, Φ, be defined as in (2) with $\sigma_p^2 = E_p[\boldsymbol{\omega}^\top\boldsymbol{\omega}]$ being the second moment of the Fourier transform of the kernel function. Given $\epsilon > 0$ and $q \in (0,1]$, there exists a constant $c_1 > 0$, such that, when*

$$M = c_1\frac{k}{\epsilon^2}\log\left(\frac{\sigma_p\, diam(\mathcal{X})}{q\epsilon}\frac{d}{k}\right), \tag{4}$$

the probability

$$\mathbb{P}\left(\sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}}\left|\langle\Phi(\mathbf{x}),\Phi(\mathbf{y})\rangle - K(\mathbf{x},\mathbf{y})\right| > \epsilon\right) \leq q.$$

The proof for Theorem 1 is provided in the appendix. As can be seen from the theorem, the dimension of random features depends predominantly on the sparsity of the signal, k, instead of its ambient dimensionality, d. Thereby, for sparse signals, the bound (4) greatly improves the original one shown in (3). We note that the factor $k\log\frac{d}{k}$ commonly appears in theoretical results of compressive sensing, *e.g.*, for constructing $m \times d$ random sub-Gaussian matrices satisfying RIP [2] and for ensuring stable recovery of sparse signals [17]. Since the approximation of kernel function with random features can be considered as constructing a low-dimensional embedding of the reproducing kernel Hilbert space associated with the kernel function [5], it is not surprising that $k\log\frac{d}{k}$ appears in our results.

The following corollary extends the above theorem to signals which are not canonically sparse but are sparse in some transform basis.

Corollary 1. *Suppose a stationary kernel function is also rotationally invariant, i.e, $f(B\boldsymbol{\delta}) = f(\boldsymbol{\delta})$ for any orthonormal basis B. Let $\mathcal{X}$ be a compact set in $\mathbb{R}^d$. Given an orthonormal basis Ψ, if for all $\mathbf{x} \in \mathcal{X}$, $\Psi\mathbf{x}$ is k-sparse, then Theorem 1 holds on $\mathcal{X}$.*

Examples of rotationally invariant stationary kernel functions include those depending only on the ℓ_2-norm of the signal, like the Gaussian kernel and the B-spline kernel [42]. Since images are often sparse in wavelet bases, this corollary allows us to apply random features on images with far-fewer features.

4. Compressive random features

We now consider the application of random features to compressive measurements. We term this scheme *compressive random features*. By performing inference directly with compressive random features, we bypass the computationally-expensive reconstruction step. For images, which are originally dense but sparse after transformation, our scheme effectively reduces computational and storage

costs and enjoys the low signal-acquisition cost provided by compressive cameras. These benefits make our scheme compelling in scenarios like Internet-of-things, where device cost, computation, and storage are of utmost concern.

Can we compute random features directly on the compressive measurements of sparse signals (either canonically or in a transform basis) without deteriorating its ability to approximate kernel functions? The following theorem addresses this question.

Theorem 2. *(Compressive random feature) Let $\mathcal{X}$ be a compact set of k-sparse vectors in $\mathbb{R}^d$. Let $\mathcal{P} : \mathbb{R}^d \to \mathbb{R}^m$ be a random orthoprojector constructed as described in Section 2.2. Let $\Phi : \mathbb{R}^m \to \mathbb{R}^M$ be the random features of an isometric kernel function, defined as $K(\mathbf{x}, \mathbf{y}) = f(\|\mathbf{x}-\mathbf{y}\|)$, with $\sigma_p^2 = E_p[\boldsymbol{\omega}^\top \boldsymbol{\omega}]$ being the second moment of its Fourier transform. Given $\epsilon > 0$ and $q \in (0, 1]$, there exist constants $c_1, c_2 > 0$, such that, when $m = c_1 \frac{k}{\epsilon^2} \log \frac{d}{k}$, $m \leq d$, and*

$$M = c_2 \frac{m}{\epsilon^2} \log \left(\frac{\sigma_p \ radius(\mathcal{X})}{q\epsilon} \frac{d}{k} \right), \tag{5}$$

the probability

$$\mathbb{P} \left(\sup_{\mathbf{x},\mathbf{y} \in \mathcal{X}} (|\langle \Phi(\mathcal{P}\mathbf{x}), \Phi(\mathcal{P}\mathbf{y}) \rangle - K(\mathbf{x}, \mathbf{y})|) > \epsilon \right) \leq q.$$

The proof is provided in the appendix. Comparing to the bound in Theorem 1, we can see that the effect of dimensionality reduction before constructing random features does not significantly impede its ability to approximate isometric kernel functions. By centering the data, we can reduce radius$(\mathcal{X})$ to diam$(\mathcal{X})$. Thereby, the required M only increases by an order of $\frac{1}{\epsilon^2} \log \frac{d}{k}$, but in return we gain the advantages of reduced device cost, computation, and storage (see Figure 1). In the context of compressive inference, this theorem provides a guarantee for applying random features directly on compressive measurements.

From our experiments in Section 5, we observe that it is possible to achieve a high compression ratio ($m \ll d$) and still obtain comparable accuracies as those of the typical random features. The reason may be that Theorem 2 requires all pairwise kernel function values are approximated; nevertheless, in classification scenarios, we are allowed to heavily compress the data as long as data points belonging to different classes do not collapse onto each other [1, 22, 24, 44, 45]. This enables us to use high compression ratios. We leave the analysis as a future work.

4.1. Analysis of storage and time complexity

We now analyze the storage and time complexity for compressive random features. As can be seen from Figure 1, if we use compressive cameras, we can get compressive measurements without actual computation, and the storage costs are $O(Nm)$. Since $m = O(k \log d)$, the saving of storage is large when $k \ll d$. For compressive random features, it costs $O(MNm)$ to construct and $O(MN)$ to store; in contrast, typical kernel methods require $O(N^2 d)$ computation and $O(N^2)$ storage for a kernel matrix. In the absence of compressive cameras where we obtain random sketches by computations, the total computational cost to construct compressive features is $O(Ndm+MNm)$, which is smaller than the cost to construct typical random features, $O(MNd)$, when m is small. We note that the accelerated random feature construction techniques [29] are also applicable to our compressive random features scheme.

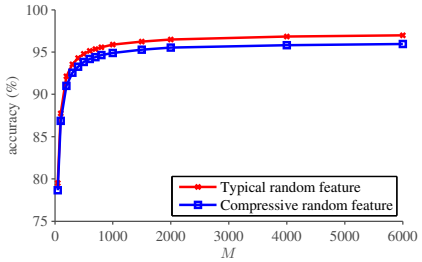
Testing time with our scheme are as follows. It takes $O(Mm)$ to construct compressive features and $O(M)$ to perform the inner product. To store a linear SVM, we only need to store the $M+1$ coefficients of the separating hyperplane. Instead, a typical kernel SVM requires the storage of all non-zero dual variables and their corresponding training samples. With large datasets, the number of non-zero dual variables usually grows linearly with N [26]. To test an image, typical kernel methods require $O(Nd)$ to evaluate the kernel function between the image and training samples. This makes kernel methods costly during the testing phase as well. In summary, with compressive random features, we can achieve nonlinear-class classification performance with improved storage and time complexity compared to both original kernel methods and typical random features.

5. Experiments

We conducted experiments on 5 datasets to examine the classification performance of linear SVMs using our compressive random features. We compared the performance against six methods, whose legends are as follows:

- **Original linear:** Linear SVM trained directly with original uncompressed data.
- **Original kernel:** Typical kernel-based SVM trained directly with original uncompressed data.
- **Compressive linear:** Linear SVM trained directly on compressive measurements, a technique commonly used in prior work [6, 28, 36].
- **Compressive kernel:** Typical kernel-based SVM trained directly with compressive measurements.
- **Typical random features:** Linear SVM trained with random features applied to the original data.
- **Compressive random features:** Linear SVM trained with our compressive random features.

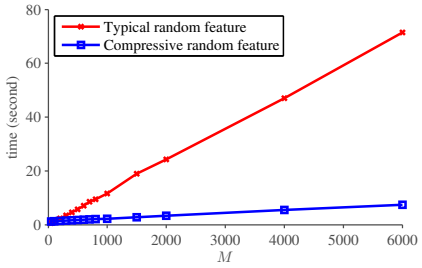
Among the last three methods, the compressive kernel approach is expected to achieve highest accuracies, since its kernel function is computed exactly. Further, in spite of undergoing both dimensionality reduction and kernel function



(a) $m = 100$ ($\frac{m}{d} = 0.13$).

	Typical random feature	Compressive random feature (Ours)			
		$m = 50$ ($\frac{m}{d} = 0.06$)	$m = 100$ ($\frac{m}{d} = 0.13$)	$m = 200$ ($\frac{m}{d} = 0.26$)	$m = 300$ ($\frac{m}{d} = 0.38$)
$M = 100$	87.75	85.14	86.87	87.45	87.66
$M = 500$	94.78	91.92	93.83	94.45	94.67
$M = 1000$	95.89	93.03	94.89	95.56	95.65
$M = 2000$	96.47	93.73	95.51	96.15	96.27
$M = 4000$	96.82	94.12	95.79	96.45	96.61
$M = 6000$	96.96	94.08	95.95	96.61	96.76
Compressive linear	–	85.63	89.85	91.86	92.45
Compressive kernel	–	93.22	95.34	96.36	96.31
Original linear	92.82				
Original kernel	96.63				

(b) Accuracy (%) under various settings.



(c) $m = 100$ ($\frac{m}{d} = 0.13$).

	Typical random feature	Compressive random feature (Ours)			
		$m = 50$ ($\frac{m}{d} = 0.06$)	$m = 100$ ($\frac{m}{d} = 0.13$)	$m = 200$ ($\frac{m}{d} = 0.26$)	$m = 300$ ($\frac{m}{d} = 0.38$)
$M = 100$	1.26	0.71	1.29	2.43	3.65
$M = 500$	5.74	1.05	1.72	3.04	4.40
$M = 1000$	11.64	1.48	2.24	3.82	5.38
$M = 2000$	24.22	2.37	3.33	5.29	7.17
$M = 4000$	46.97	4.06	5.43	8.00	10.88
$M = 6000$	71.46	5.84	7.41	11.02	14.64

(d) Random feature construction time (in seconds). For compressive random features, the random projection time is included. The experiments were conducted on Intel Xeon CPU E5-2620 2.0GHz with 16GB memory. Note that in our experiments, typical and compressive random features have similar SVM training and testing time.

Figure 2: MNIST results.

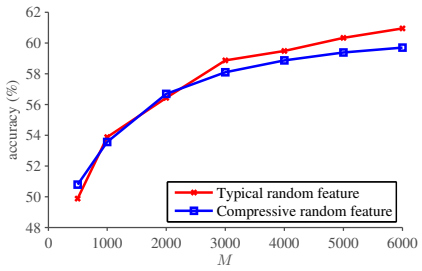
approximation, the proposed compressive random features is expected to achieve accuracy that is comparable to typical random features, especially when the dimension of compressive measurements, m, is large enough. We also expect to achieve accuracies comparable to the kernel-based SVM when the dimension of random features, M, is large enough so to precisely approximate the kernel function.

In all experiments, the SVMs are directly trained with pixel values or their compressive measurements (although our scheme also supports sparse features, like features learned by convolutional neural networks). Due to memory issues, in some instances, we downsampled images. We use the Gaussian kernel function $K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/(2\sigma^2))$ in all experiments, with σ kept the same across different methods. We used C-SVM in LIBLINEAR [21] with $C = 1$. Finally, all results are averages over 20 trials.

We briefly introduce the 5 datasets used for validation.

- **MNIST** [30] contains $60,000$ training images and $10,000$ test images. The 28×28 gray-scale images contain digits from 0 to 9. We set $\sigma = 10$. The results are shown in Figure 2.
- **8 scene categories dataset** [35] contains 256×256 RGB images of 8 different scenes, like mountain views, streets, highways, and coast, . . . , *etc*. There are 2688 images, and we randomly split them into 2150 training images and 538 test images. We resized images into 32×32. We set $\sigma = 8$. The results are shown in Figure 3.
- **INRIA person dataset** [14] contains 128×64 RGB images. Each positive image contains a standing person, and the negative images do not. There are 8506 training images and 2482 test images. We resized the images to 32×16. We set $\sigma = 5$. The results are shown in Figure 4.
- **CIFAR-10** [27] contains 32×32 RGB images of 10 different objects, like airplanes and horses. Each class has 5000 training images and 1000 test images. We set $\sigma = 18$. The results are shown in Figure 5.
- **Street view house numbers dataset** [34] contains 32×32 RGB images. It contains images with different digits taken from house numbers in Google Street View images. It has 73257 training images and 26032 test images. We set $\sigma = 13$. The results are shown in Figure 6.

Observations. Across all experiments, compressive random features has a performance that is comparable to typical random features and outperforms compressive linear SVMs even under high compression ratio ($\frac{m}{d}$=0.07). Further, as shown in Figure 2d, working with dimensionality-reduced compressive measurements effectively reduces the

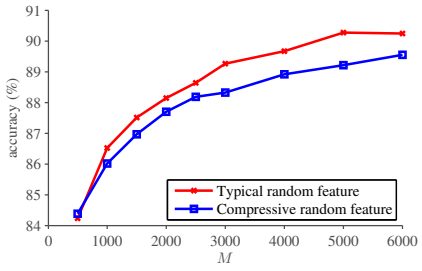


(a) $m = 400$ ($\frac{m}{d} = 0.13$).

	Typical random feature	Compressive random feature (Ours)				
		$m = 100$ ($\frac{m}{d} = 0.03$)	$m = 200$ ($\frac{m}{d} = 0.07$)	$m = 400$ ($\frac{m}{d} = 0.13$)	$m = 500$ ($\frac{m}{d} = 0.16$)	$m = 750$ ($\frac{m}{d} = 0.24$)
$M = 1000$	53.87	52.19	53.60	53.55	53.56	53.86
$M = 2000$	56.42	54.66	56.14	56.68	56.64	56.83
$M = 3000$	58.86	56.25	56.59	58.09	58.32	58.56
$M = 4000$	59.47	56.42	57.64	58.87	58.89	59.38
$M = 5000$	60.33	57.40	58.89	59.38	59.10	60.09
$M = 6000$	60.95	57.64	59.21	59.69	60.25	60.42
Compressive linear	–	43.00	40.53	36.63	36.26	35.91
Compressive kernel	–	57.93	59.69	60.49	60.62	60.99
Original linear	36.43					
Original kernel	61.34					

(b) Accuracy (%) under various settings.

Figure 3: 8 scene categories dataset results.



(a) $m = 200$ ($\frac{m}{d} = 0.13$).

	Typical random feature	Compressive random feature (Ours)				
		$m = 50$ ($\frac{m}{d} = 0.03$)	$m = 100$ ($\frac{m}{d} = 0.07$)	$m = 200$ ($\frac{m}{d} = 0.13$)	$m = 300$ ($\frac{m}{d} = 0.20$)	$m = 400$ ($\frac{m}{d} = 0.26$)
$M = 500$	84.23	82.87	83.62	84.38	84.36	84.17
$M = 1000$	86.52	84.75	85.66	86.01	86.30	86.21
$M = 2000$	88.15	85.65	87.23	87.70	87.85	87.86
$M = 4000$	89.67	86.63	88.44	88.91	89.50	89.42
$M = 5000$	90.27	87.19	88.42	89.21	89.68	89.86
$M = 6000$	90.24	87.11	88.89	89.55	89.76	90.15
Compressive linear	–	81.93	84.36	86.06	87.18	87.96
Compressive kernel	–	87.99	89.92	90.93	91.27	91.43
Original linear	89.20					
Original kernel	91.90					

(b) Accuracy (%) under various settings.

Figure 4: INRIA person dataset results.

time to construct random features. In some datasets, we observe that when the dimension of random feature M is small, compressive linear SVMs are able to achieve better accuracies than both compressive random features and typical random features. This could be due to poor approximations of the kernel similarities at small values of M. As expected, when using larger values of M, both random feature methods achieve higher accuracies.

Looking at the results of CIFAR-10 and the street view house numbers datasets, all these methods still have room for improvement compared with state-of-the-art methods like convolutional neural networks (CNN). This gap in performance can be attributed, in part, to our reliance on pixel-values as the underlying features.

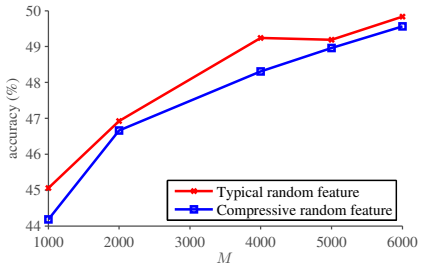
6. Conclusion and discussion

We propose compressive random features, a framework for kernel-based inference on compressive measurements that enjoys low acquisition, computation, and storage costs, along with theoretical guarantees on its ability to approximate kernel similarities. In the context of compressive inference, we introduced a novel method to perform scalable nonlinear inference. Thereby, for many applications, our scheme provides an effective solution that provides a trade-off between inference performance and design considerations like cost, computation, and storage. Finally, we note that even though we focused on sparse signals and stationary kernel functions, we conjecture that similar results can be derived for low-dimensional smooth manifolds and for dot-product kernels.

Comparison to the Nyström method. The Nyström method [18] is another popular method for large-scale kernel-based inference. By first obtaining a low-rank approximation of the kernel matrix, the Nyström method obtains eigen-decomposition of the low-rank matrix and generates features using the eigenvectors. Since this process involves learning from data, the Nyström method can achieve better performance by exploiting structures specific to the dataset. However, the dependency on training data also makes the Nyström method less flexible than random features, whose feature construction function can be designed independent to the overall training-testing process. In this context, we can view the results in this paper as a potential approach to incorporate more knowledge of the signal into random features without having to perform learning.

A. Proofs

As discussed in Section 2.2, random orthoprojectors satisfy RIP with high probability. We state the following theorem which will be utilized to prove our theorem.

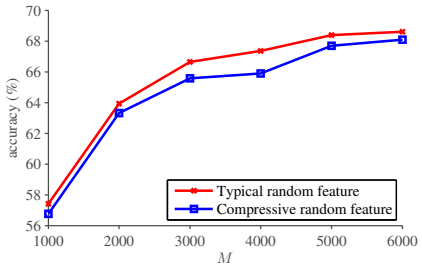


(a) $m = 400$ ($\frac{m}{d} = 0.13$).

	Typical random feature	Compressive random feature (Ours)				
		$m = 100$ ($\frac{m}{d} = 0.03$)	$m = 200$ ($\frac{m}{d} = 0.07$)	$m = 400$ ($\frac{m}{d} = 0.13$)	$m = 600$ ($\frac{m}{d} = 0.20$)	$m = 800$ ($\frac{m}{d} = 0.26$)
$M = 1000$	45.06	41.66	43.56	44.19	44.31	44.22
$M = 2000$	46.93	44.36	45.95	46.66	47.15	47.22
$M = 4000$	49.24	45.65	47.48	48.31	48.81	49.01
$M = 5000$	49.19	45.96	48.22	48.96	49.08	49.20
$M = 6000$	49.84	46.06	48.43	49.56	49.76	49.68
Compressive linear	–	25.60	25.66	32.45	34.23	35.45
Compressive kernel	–	45.27	47.42	48.51	48.93	49.20
Original linear	36.96					
Original kernel	49.56					

(b) Accuracy (%) under various settings.

Figure 5: CIFAR-10 dataset results.



(a) $m = 400$ ($\frac{m}{d} = 0.13$).

	Typical random feature	Compressive random feature (Ours)				
		$m = 100$ ($\frac{m}{d} = 0.03$)	$m = 200$ ($\frac{m}{d} = 0.07$)	$m = 400$ ($\frac{m}{d} = 0.13$)	$m = 500$ ($\frac{m}{d} = 0.16$)	$m = 750$ ($\frac{m}{d} = 0.24$)
$M = 1000$	57.42	57.20	57.15	56.76	57.25	57.66
$M = 2000$	63.93	61.05	63.85	63.32	63.21	64.49
$M = 3000$	66.64	63.72	65.29	65.57	65.25	66.20
$M = 4000$	67.36	63.58	66.30	65.89	66.83	67.39
$M = 5000$	68.38	65.10	67.43	67.69	67.89	67.52
$M = 6000$	68.61	65.63	67.05	68.09	68.73	68.10
Compressive linear	–	11.26	14.46	15.93	16.73	12.84
Compressive kernel	–	53.89	56.80	57.78	57.19	57.59
Original linear	19.16					
Original kernel	57.79					

(b) Accuracy (%) under various settings.

Figure 6: Street view house numbers dataset results.

Theorem 3. *Let $\mathcal{P} : \mathcal{X} \to \mathbb{R}^m$ be a $m \times d$ random orthoprojector constructed as described in Section 2.2. If $\mathcal{X}$ is the set of k-sparse vectors in $\mathbb{R}^d$ and $m \in \left[\frac{\log(2)+k\log(12/\delta)+k\log(ed/k)}{\delta^2/16-\delta^3/48}, d\right]$, then for $\delta \in (0,1)$ we have*

$$\mathbb{P}\left\{\forall \mathbf{x} \in \mathcal{X}, (1-\delta)\|\mathbf{x}\|^2 \leq \|\mathcal{P}\mathbf{x}\|^2 \leq (1+\delta)\|\mathbf{x}\|^2\right\} \geq 1 - 2\left(\frac{12}{\delta}\right)^k \left(\frac{ed}{k}\right)^k \exp\left(-\left(\frac{\delta^2}{16} - \frac{\delta^3}{48}\right)m\right). \tag{6}$$

The proof of the theorem is a simple extension of the results in [2].

Proof of Theorem 1. The difference of two k-sparse vectors is at most $2k$-sparse. Let $z := 2k$. For any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, $\mathbf{x}-\mathbf{y}$ belongs to one of the $\binom{d}{z}$ z-dimensional subspaces, $\mathcal{M}_1, \ldots, \mathcal{M}_{\binom{d}{z}}$. Each $\mathcal{M}_j$, $j=1,\ldots,\binom{d}{z}$, is compact and has diameter at most twice $\operatorname{diam}(\mathcal{X})$. Thus, we can construct a ϵ-net in each of $\mathcal{M}_j$, $j=1,\ldots,\binom{d}{z}$, and $\{\mathbf{x}-\mathbf{y}|\forall \mathbf{x},\mathbf{y} \in \mathcal{X}\} \subseteq \cup_{j=1}^{\binom{d}{z}} \mathcal{M}_j$. Each net will have at most $T = (2\operatorname{diam}(\mathcal{X})/r)^z$ balls of radius r [20, Chapter 5]. Denote Δ as $(\mathbf{x},\mathbf{y})$ and $\mathcal{M}$ as $\{\Delta|\forall \mathbf{x},\mathbf{y} \in \mathcal{X}\}$. Let $\Delta_{i,j}, i = 1,\ldots T$ be the i-th center in the ϵ-net of $\mathcal{M}_j$. Define $f : \mathcal{M} \to \mathbb{R}$, $f(\Delta) := \langle \Phi(\mathbf{x}), \Phi(\mathbf{y})\rangle - K(\mathbf{x},\mathbf{y})$ and let L_f be the Lipschitz constant of f. By limiting L_f and making sure that all $f(\Delta_{i,j})$ are small, we can provide a bound to the overall approximation error:

$$\begin{aligned}&\mathbb{P}\left(\sup_{\Delta\in\mathcal{M}} |f(\Delta)| > \epsilon\right) \\ &\leq \mathbb{P}\left(L_f > \frac{\epsilon}{2r}\right) + \mathbb{P}\left(\cup_{j=1}^{\binom{d}{z}} \cup_{i=1}^{T} \{|f(\Delta_{i,j})| > \frac{\epsilon}{2}\}\right).\end{aligned} \tag{7}$$

Let $\Delta^* = \arg\max_{\Delta\in\mathcal{M}} \|\bigtriangledown f(\Delta)\|$. By expanding $\Phi(\mathbf{x}) = e^{jW\mathbf{x}}$, where i-th row of W contains $\boldsymbol{\omega}_i$, we have

$$\begin{aligned}&E[L_f^2] = E\|\nabla f(\Delta^*)\|^2 = E\|\nabla(\langle\Phi(\Delta^*), \Phi(\mathbf{0})\rangle)\|^2 \\ &-E\|\nabla K(\Delta^*, \mathbf{0})\|^2 \leq E\|\nabla(\langle\Phi(\Delta^*), \Phi(\mathbf{0})\rangle)\|^2 \leq E\|\boldsymbol{\omega}\|^2 = \sigma_p^2,\end{aligned}$$

as [38], we use Markov's inequality and get

$$\mathbb{P}\left(L_f \geq \frac{\epsilon}{2r}\right) = \mathbb{P}\left(L_f^2 \geq \left(\frac{\epsilon}{2r}\right)^2\right) \leq \frac{E[L_f^2]}{\left(\frac{\epsilon}{2r}\right)^2} \leq \left(\frac{2r\sigma_p}{\epsilon}\right)^2.$$

Using a union bound and Hoeffding's inequality, we have

$$\mathbb{P}\left(\cup_{j=1}^{\binom{d}{z}} \cup_{i=1}^{T} \{|f(\Delta_{i,j})| > \frac{\epsilon}{2}\}\right) \leq 2\binom{d}{z} T \exp\left(\frac{-M\epsilon^2}{8}\right).$$

Because $\binom{d}{z} \leq \left(\frac{ed}{z}\right)^z$, we can bound (7) by

$$\begin{aligned}&\mathbb{P}\left(\sup_{\Delta\in\mathcal{M}} |f(\Delta)| > \epsilon\right) \\ &\leq \left(\frac{2r\sigma_p}{\epsilon}\right)^2 + 2\left(\frac{ed}{z}\right)^z \left(\frac{2\operatorname{diam}(\mathcal{X})}{r}\right)^z \exp\left(\frac{-M\epsilon^2}{8}\right) \\ &= 2\left[\left(\frac{2ed\operatorname{diam}(\mathcal{X})}{z}\right)^z \exp\left(\frac{-M\epsilon^2}{8}\right)\right] r^{-z} + \left(\frac{2\sigma_p}{\epsilon}\right)^2 r^2 \\ &:= 2\alpha r^{-z} + \beta r^2.\end{aligned}$$

Minimizing the right hand side w.r.t. r results in $r = (\frac{\alpha z}{\beta})^{\frac{1}{z+2}}$. After substituting r, the right hand side becomes $\alpha^{\frac{2}{z+2}} \beta^{\frac{z}{z+2}} \left(2z^{\frac{-z}{z+2}} + z^{\frac{2}{z+2}}\right)$, and we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\Delta \in \mathcal{M}} |f(\Delta)| > \epsilon\right) \\
&\leq \left(\frac{ed\sigma_p \operatorname{diam}(\mathcal{X})}{z\epsilon}\right)^{\frac{2z}{z+2}} \exp\left(\frac{-M\epsilon^2}{4(z+2)}\right)\left(2z^{\frac{-z}{z+2}} + z^{\frac{2}{z+2}}\right) \\
&\leq 3\left(\frac{2ed\sigma_p \operatorname{diam}(\mathcal{X})}{z\epsilon}\right)^{\frac{2z}{z+2}} \exp\left(\frac{-M\epsilon^2}{4(z+2)}\right).
\end{aligned}$$

The last inequality holds because $\left(2z^{\frac{-z}{z+2}} + z^{\frac{2}{z+2}}\right) \leq 3$ for all $z \geq 1$. Setting an upper bound for the right hand side and solving for M will prove the theorem. □

Proof of Corollary 1. We use the following property of multi-variant Fourier transform:

$$\mathcal{F}(f(B\mathbf{x})) = \frac{1}{\det(B)}(\mathcal{F}f)(B^{-T}\boldsymbol{\omega}).$$

Since B is a orthonormal basis and f is rotational invariant, we have

$$p(\boldsymbol{\omega}) = \mathcal{F}(f(\mathbf{x})) = \mathcal{F}(f(B\mathbf{x})) = (\mathcal{F}f)(B^{-T}\boldsymbol{\omega}) = p(B\boldsymbol{\omega}).$$

Therefore, p is rotational invariant. Let $\boldsymbol{\alpha} = B^\top \mathbf{x}$ for all $\mathbf{x} \in \mathcal{X}$. We have $\boldsymbol{\omega}^\top \mathbf{x} = \boldsymbol{\omega}^\top B\boldsymbol{\alpha} = (B^\top\boldsymbol{\omega})^\top\boldsymbol{\alpha} := \mathbf{z}^\top\boldsymbol{\alpha}$. Since p is rotational invariant and $\boldsymbol{\omega} \sim p$, $\mathbf{z} \sim p$. So Theorem 1 can be applied to $\boldsymbol{\alpha}$, which is k-sparse. □

The sketch to prove Theorem 2 is as follows. In order to uniformly bound the approximation of kernel function with compressive features, we simply need to uniformly bound the errors caused by compressive sensing and random feature approximation separately.

Proof of Theorem 2. Let $f(\|\mathbf{x}-\mathbf{y}\|) := K(\mathbf{x},\mathbf{y})$, $\forall \mathbf{x},\mathbf{y}$. By triangular inequality, we have

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}}\left(\left|\langle\Phi(\mathcal{P}\mathbf{x}),\Phi(\mathcal{P}\mathbf{y})\rangle - K(\mathbf{x},\mathbf{y})\right|\right) > \epsilon\right\} \\
&\leq \mathbb{P}\left\{\sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}}\left(\left|\langle\Phi(\mathcal{P}\mathbf{x}),\Phi(\mathcal{P}\mathbf{y})\rangle - f(\|\mathcal{P}\mathbf{x}-\mathcal{P}\mathbf{y}\|)\right|\right.\right. \\
&\quad \left.\left. + \left|f(\|\mathcal{P}\mathbf{x}-\mathcal{P}\mathbf{y}\|) - f(\|\mathbf{x}-\mathbf{y}\|)\right|\right) > \epsilon\right\} \\
&\leq \mathbb{P}\left\{\sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}}\left(\left|\langle\Phi(\mathcal{P}\mathbf{x}),\Phi(\mathcal{P}\mathbf{y})\rangle - f(\|\mathcal{P}\mathbf{x}-\mathcal{P}\mathbf{y}\|)\right|\right) > \frac{\epsilon}{2}\right\} \\
&\quad + \mathbb{P}\left\{\sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}}\left(\left|f(\|\mathcal{P}\mathbf{x}-\mathcal{P}\mathbf{y}\|) - f(\|\mathbf{x}-\mathbf{y}\|)\right|\right) > \frac{\epsilon}{2}\right\}
\end{aligned} \tag{8}$$

Let D_P be the diameter of $\mathcal{P}\mathcal{X}$ and $D = 2\,\text{radius}(\mathcal{X})$. Using the result in [38], we can bound the first term:

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\mathbf{x},\mathbf{y}\in\mathcal{P}\mathcal{X}}\left(\left|\langle\Phi(\mathbf{x}),\Phi(\mathbf{y})\rangle - f(\|\mathbf{x}-\mathbf{y}\|)\right|\right) > \frac{\epsilon}{2}\right\} \\
&\leq 2^{10}\left(\frac{\sigma_p D_P}{\epsilon}\right)^2 \exp\left(\frac{-M\epsilon^2}{16(m+2)}\right) \\
&\leq 2^{10}\frac{d}{m}\left(\frac{\sigma_p D}{\epsilon}\right)^2 \exp\left(\frac{-M\epsilon^2}{16(m+2)}\right).
\end{aligned}$$

The last inequality holds because by the construction of random orthoprojector $\|\mathcal{P}\mathbf{x}\| \leq \sqrt{\frac{d}{m}}\|\mathbf{x}\|$ for all $\mathbf{x}$. The second term can be bounded as follows. Since f is differentiable and continuous and $\mathcal{X}$ is compact, the derivative of f attains its maximum and minimum in $\mathcal{X}$. Therefore, the kernel function is Lipschitz continuous in $\mathcal{X}$. Let L be the Lipschitz constant. By Theorem 3, we have $\forall \mathbf{x},\mathbf{y} \in \mathcal{X}$, $|\|\mathcal{P}\mathbf{x}-\mathcal{P}\mathbf{y}\| - \|\mathbf{x}-\mathbf{y}\|| \leq \delta\|\mathbf{x}-\mathbf{y}\| \leq \delta D$ with probability at least $1 - 2\exp(-c_1\delta^2 m)$. By the definition of Lipschitz continuous we have $|f(\|\mathcal{P}\mathbf{x}-\mathcal{P}\mathbf{y}\|) - f(\|\mathbf{x}-\mathbf{y}\|)| \leq \delta DL$. Thus, by setting $\delta = \frac{\epsilon}{2DL}$, we bound the second term using (6). Therefore, we have the following result:

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}}\left(\left|\langle\Phi(\mathcal{P}\mathbf{x}),\Phi(\mathcal{P}\mathbf{y})\rangle - K(\mathbf{x},\mathbf{y})\right|\right) > \epsilon\right\} \\
&\leq 2^{10}\frac{d}{m}\left(\frac{\sigma_p D}{\epsilon}\right)^2 \exp\left(\frac{-M\epsilon^2}{16(m+2)}\right) \\
&\quad + 2\left(\frac{24DL}{\epsilon}\right)^k \left(\frac{ed}{k}\right)^k \exp\left[-\left(\frac{\epsilon^2}{64D^2L^2} - \frac{\epsilon^3}{384D^3L^3}\right)m\right].
\end{aligned} \tag{9}$$

When m is large enough, the second term is less than the first term, the right hand side of (9) is less than $2^{11}\frac{d}{m}\left(\frac{\sigma_p D}{\epsilon}\right)^2 \exp\left(\frac{-M\epsilon^2}{16(m+2)}\right)$. To make it less than a constant probability, we need

$$M = O\left(\frac{m}{\epsilon^2}\log\frac{d\sigma_p D}{m\epsilon}\right).$$

□

References

[1] A. S. Bandeira, D. G. Mixon, and B. Recht. Compressive classification and the rare eclipse problem. *arXiv:1404.3203*, 2014. 1, 3, 4

[2] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008. 3, 7

[3] R. G. Baraniuk. Compressive sensing. *IEEE Signal Proc. Mag.*, 24(4):118–121, 2007. 3

[4] Z. Ben-Haim and Y. C. Eldar. The cramér-rao bound for estimating a sparse parameter vector. *IEEE Trans. Signal Proc.*, 58(6):3384–3389, 2010. 3

[5] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011. 3

[6] R. Calderbank, S. Jafarpour, and R. Schapire. Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. Technical report, 2009. 3, 4

[7] E. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Info. Theory*, 51(12):4203–4215, 2005. 3

[8] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9–10):589–592, 2008. 3

[9] E. J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure and Applied Mathematics*, 59(8):1207–1223, 2006. 3

[10] E. J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Info. Theory*, 52(12):5406–5425, 2006. 3

[11] V. Cevher, A. Sankaranarayanan, M. F. Duarte, D. Reddy, R. G. Baraniuk, and R. Chellappa. Compressive sensing for background subtraction. In *ECCV*, 2008. 3

[12] H. Chen, M. S. Asif, A. C. Sankaranarayanan, and A. Veeraraghavan. FPA-CS: Focal plane array-based compressive imaging in short-wave infrared. In *CVPR*, 2015. 1

[13] B. Dai, B. Xie, N. He, Y. Liang, A. Raj, M.-F. F. Balcan, and L. Song. Scalable kernel methods via doubly stochastic gradients. In *NIPS*, 2014. 1

[14] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 5

[15] M. Davenport, P. T. Boufounos, M. B. Wakin, R. G. Baraniuk, et al. Signal processing with compressive measurements. *IEEE J. Selected Topics in Signal Processing*, 4(2):445–460, 2010. 3

[16] M. A. Davenport, M. F. Duarte, M. B. Wakin, J. N. Laska, D. Takhar, K. F. Kelly, and R. G. Baraniuk. The smashed filter for compressive classification and target recognition. In *Electronic Imaging*, 2007. 3

[17] K. Do Ba, P. Indyk, E. Price, and D. P. Woodruff. Lower bounds for sparse recovery. In *ACM-SIAM Symp. Discrete Algorithms (SODA)*, 2010. 3

[18] P. Drineas and M. W. Mahoney. On the Nyström method for approximating a gram matrix for improved kernel-based learning. *JMLR*, 6:2153–2175, 2005. 6

[19] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. E. Kelly, R. G. Baraniuk, et al. Single-pixel imaging via compressive sampling. *IEEE Signal Proc. Mag.*, 25(2):83, 2008. 1

[20] Y. C. Eldar and G. Kutyniok. *Compressed sensing: theory and applications*. Cambridge University Press, 2012. 7

[21] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *JMLR*, 9:1871–1874, 2008. 5

[22] A. Globerson and S. T. Roweis. Metric learning by collapsing classes. In *NIPS*, 2005. 4

[23] J. Haupt, R. Castro, R. Nowak, G. Fudge, and A. Yeh. Compressive sampling for signal classification. In *IEEE Asilomar Conf. Signals, Systems and Computers*, pages 1430–1434, 2006. 3

[24] C. Hegde, A. C. Sankaranarayanan, W. Yin, and R. Baraniuk. NuMax: A convex approach for learning near-isometric linear embeddings. *IEEE Trans. Signal Proc.*, 63(22):6109–6121, 2015. 4

[25] C. Hegde, M. Wakin, and R. Baraniuk. Random projections for manifold learning. In *NIPS*, 2008. 3

[26] P. Kar and H. Karnick. Random feature maps for dot product kernels. In *AISTATS*, 2012. 1, 2, 4

[27] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images, 2009. 5

[28] K. Kulkarni and P. Turaga. Reconstruction-free action inference from compressive imagers. *PAMI*, 38(4):772–784, 2015. 3, 4

[29] Q. Le, T. Sarlós, and A. Smola. Fastfood-approximating kernel expansions in loglinear time. In *ICML*, 2013. 4

[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5

[31] S. Lohit, K. Kulkarni, P. Turaga, J. Wang, and A. Sankaranarayanan. Reconstruction-free inference on compressive measurements. In *CVPR*, 2015. 1, 3

[32] A. Mahalanobis, R. Shilling, R. Murphy, and R. Muise. Recent results of medium wave infrared compressive sensing. *Applied optics*, 53(34):8060–8070, 2014. 1

[33] O. Maillard and R. Munos. Compressed least-squares regression. In *NIPS*, 2009. 3

[34] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. 5

[35] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 5

[36] S. Paul, C. Boutsidis, M. Magdon-Ismail, and P. Drineas. Random projections for support vector machines. In *AISTATS*, 2013. 3, 4

[37] N. Pham and R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *ACM Intl. Conf. Knowledge Discovery and Data Mining (SIGKDD)*, 2013. 1

[38] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NIPS*, 2007. 1, 2, 3, 7, 8

[39] M. Rudelson and R. Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Comm. Pure and Applied Mathematics*, 61(8):1025–1045, 2008. 3

[40] W. Rudin. *Fourier analysis on groups*. Number 12. John Wiley & Sons, 1990. 2

[41] A. C. Sankaranarayanan, P. K. Turaga, R. Chellappa, and R. G. Baraniuk. Compressive acquisition of linear dynamical systems. *SIAM Journal on Imaging Sciences*, 6(4):2109–2133, 2013. 1, 3

[42] B. Scholkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001. 1, 2, 3

[43] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *PAMI*, 34(3):480–492, 2012. 1

[44] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10:207–244, 2009. 4

[45] E. P. Xing, M. I. Jordan, S. Russell, and A. Y. Ng. Distance metric learning with application to clustering with side-information. In *NIPS*, 2002. 4

[46] J. Yang, V. Sindhwani, Q. Fan, H. Avron, and M. Mahoney. Random laplace feature maps for semigroup kernels on histograms. In *CVPR*, 2014. 1