Random Laplace Feature Maps for Semigroup Kernels on Histograms

Jiyan Yang
Stanford University
jiyan@stanford.edu

Vikas Sindhwani, Quanfu Fan, Haim Avron
IBM Research, NY
{vsindhw,qfan,haimav}@us.ibm.com

Michael W. Mahoney
University of California at Berkeley
mmahoney@stat.berkeley.edu

Abstract

With the goal of accelerating the training and testing complexity of nonlinear kernel methods, several recent papers have proposed explicit embeddings of the input data into low-dimensional feature spaces, where fast linear methods can instead be used to generate approximate solutions. Analogous to random Fourier feature maps to approximate shift-invariant kernels, such as the Gaussian kernel, on $\mathbb{R}^d$, we develop a new randomized technique called random Laplace features, to approximate a family of kernel functions adapted to the semigroup structure of $\mathbb{R}^d_+$. This is the natural algebraic structure on the set of histograms and other non-negative data representations. We provide theoretical results on the uniform convergence of random Laplace features. Empirical analyses on image classification and surveillance event detection tasks demonstrate the attractiveness of using random Laplace features relative to several other feature maps proposed in the literature.

1. Introduction

A wide spectrum of statistical learning problems in computer vision have been elegantly framed within the framework of kernel methods [19]. The algorithmic recipe in this framework is as follows. A kernel function, $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, is defined on an input domain of visual features, $\mathcal{X} \subset \mathbb{R}^d$. The kernel provides access to an implicit embedding of the input domain, $\Psi : \mathcal{X} \mapsto \mathcal{H}$, such that $k(\mathbf{x}, \mathbf{z}) = \langle \Psi(\mathbf{x}), \Psi(\mathbf{z}) \rangle_{\mathcal{H}}$, where $\mathcal{H}$ is a possibly infinite-dimensional inner product "feature space" (with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$). Regularized linear models in $\mathcal{H}$ then define non-parametric, non-linear models with respect to the original domain. Despite its conceptual simplicity, this idea is highly generalizable and versatile: it leads to nonlinear algorithms for supervised image classification and object detection, unsupervised visual feature extraction, image denoising, action recognition in videos, integration of multiple descriptors [10] , and many other tasks [14].

In the face of "big data" in computer vision [7, 20], the scalability of kernel methods, which is typically super-linear in the number of data points, is well-recognized as a valid concern. In recent years, approximations to kernel functions via explicit low-dimensional feature maps [18, 21, 15, 17, 12] have emerged as an appealing strategy to turn the complexity of learning nonlinear kernel methods back to that of training linear models, which typically scale linearly in the number of data points in a variety of settings such as regression, classification [13] and principal component analysis. Importantly, storage requirements and test-time prediction speed can also be dramatically improved. These feature maps provide a low-distortion embedding, $\hat{\Psi} : \mathcal{X} \mapsto \mathbb{R}^s$, such that,

$$k(\mathbf{x}, \mathbf{z}) \approx \langle \hat{\Psi}(\mathbf{x}), \hat{\Psi}(\mathbf{z}) \rangle_{\mathbb{R}^s} . \tag{1}$$

Shift-invariant kernels on $\mathbb{R}^d$, such as the Gaussian kernel, $k(\mathbf{x}, \mathbf{z}) = e^{-\frac{\|\mathbf{x}-\mathbf{z}\|_2^2}{2\sigma^2}}$, can be written as $k(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x} - \mathbf{z})$, for a positive-definite function $\phi : \mathbb{R}^d \mapsto \mathbb{R}$. A randomized construction of approximate feature maps for this class of kernels was recently suggested by Rahimi and Recht [18]. Their construction is based on a classical characterization of the family of positive-definite functions on $\mathbb{R}^d$ given by Bochner's Theorem [2].

Theorem 1 (Bochner's Theorem [2])**.** *A continuous shift-invariant kernel function $k(\mathbf{x}, \mathbf{z}) \equiv \phi(\mathbf{x} - \mathbf{z})$ on $\mathbb{R}^d$ is positive definite if and only if it is the Fourier transform of a unique finite non-negative measure on $\mathbb{R}^d$. That is, for any $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$,*

$$k(\mathbf{x}, \mathbf{z}) = \int_{\mathbb{R}^s} e^{-i(\mathbf{x}-\mathbf{z})^T \mathbf{w}} p(\mathbf{w}) d\mathbf{w} = \mathbb{E}_{\mathbf{w} \sim p}[e^{-i(\mathbf{x}-\mathbf{z})^T \mathbf{w}}] .$$

In the above, we assume, without loss of generality, that the non-negative measure is a probability measure with the

associated density p. Bochner's theorem establishes one-to-one correspondence between shift-invariant kernel functions and probability densities on $\mathbb{R}^d$, via the Fourier transform. For the Gaussian kernel with bandwidth σ, the associated density is again Gaussian with σ^{-2} times the identity as the covariance matrix. The above result immediately suggests a Monte Carlo approximation to the kernel function of the form,

$$k(\mathbf{x},\mathbf{z}) \approx \frac{1}{s}\sum_{j=1}^{s} e^{-i\mathbf{x}^T\mathbf{w}_j} e^{i\mathbf{z}^T\mathbf{w}_j} = \langle \bar{\Psi}(\mathbf{x}), \bar{\Psi}(\mathbf{z}) \rangle_{\mathbb{C}^s} , \quad (2)$$

where the points $\mathbf{w}_j$ are drawn from p, yielding a feature map of the form $\bar{\Psi}(\mathbf{x}) = \frac{1}{\sqrt{s}}[e^{-i\mathbf{x}^T\mathbf{w}_1} \dots e^{-i\mathbf{x}^T\mathbf{w}_s}]$ from which real-valued feature maps satisfying (1) can be derived. Following Rahimi and Recht [18], this map is referred to as *random Fourier feature map* due to the central role of the Fourier transform in characterizing shift-invariant kernels on $\mathbb{R}^d$.

Beyond shift-invariant kernels on $\mathbb{R}^d$, several recent papers have attempted to develop explicit low-dimensional feature maps to approximate specific kernels that excel in computer vision applications [21, 15, 17, 16]. These kernels are typically much better adapted to data representations in the form of finite probability distributions or normalized histograms, that common descriptors such as bag of visual words [4] and spatial pyramids [11] assume. Vedaldi and Zisserman [21] suggest approximate feature maps for the additive family of Intersection, Hellinger's, χ^2 and Jensen-Shannon kernels whose feature spaces respectively induce well-known divergence measures on finite probability distributions. Li et al. [15] suggest approximate feature maps for "skewed" multiplicative variants of the Intersection and χ^2 kernels, in an attempt to match the empirical performance of the exponentiated-χ^2 kernel, considered state of the art [3] for histogram descriptors. Table 1 catalogues these kernels and their associated approximate feature maps obtained through a randomized or deterministic sampling process.

The starting point of this paper is the observation that the natural algebraic structure on the space of histograms and other non-negative descriptors, is that of an *abelian semigroup*.

Definition 2 (Abelian semigroup)**.** *A semigroup $(S, \circ)$ is a nonempty set S equipped with an associative composition $\circ$, i.e. for any $x, y, z \in S : x \circ (y \circ z) = (x \circ y) \circ z$ and a neutral/identity element e, i.e., for any $x \in S : x \circ e = x$. For an abelian semigroup, the composition is commutative, i.e., for any $x, y \in S : x \circ y = y \circ x$.*

In particular, $(\mathbb{R}^d_+, +)$ forms an abelian semigroup with $0 \in \mathbb{R}^d_+$ as the identity element. This basic definition is sufficient to introduce the concept of kernels on semigroups.

Definition 3 (Kernels on Abelian semigroups [1])**.** *A function $k : S \times S \mapsto \mathbb{R}$ is a positive definite kernel function on an abelian semigroup $(S, \circ)$ if $k(s,t) = \phi(s \circ t)$ where $\phi : S \mapsto \mathbb{R}$ is a positive definite function, i.e., for any $s_1 \dots s_n \in S$, and any real-valued scalars $c_1 \dots c_n$, the following holds[1]: $\sum_{i,j=1}^{n} c_i c_j \phi(s_i \circ s_j) \geq 0$.*

We now state our main contributions.

- We propose new randomized approximate feature maps for kernels based on the semigroup structure of $\mathbb{R}^d_+$. These *semigroup kernels* are characterized via extensions of Bochner's theorem [1] developed in the theory of harmonic analysis for general algebraic structures such as groups and semigroups. The Laplace transform assumes the role of the Fourier transform in our setting. Our proposed technique is therefore termed as *random Laplace features*, analogous to random Fourier features for shift-invariant kernels on $\mathbb{R}^d$.
- We provide theoretical analysis on the uniform convergence of random features associated with semigroup kernels. In particular, we show that with high probability, for all pairs of points drawn from a bounded input domain, semigroup kernels can be approximated to within ϵ error with $O(d\epsilon^{-2}\log(\epsilon^{-2}))$ random Laplace features.
- Random Laplace features provide approximations to a broad family of kernels on histograms, which includes two kernels that we call *Exponential-Semigroup* and *Reciprocal-Semigroup*. The definitions of these kernels and their associated approximate feature maps are provided in Table 1 as well. We provide a thorough empirical comparison of these feature maps on image classification and surveillance event detection tasks. Results show favorable accuracy-time tradeoffs from using our random Laplace feature maps.

2. Semigroup Kernels on Histograms

As per Definition 3, kernels respecting the algebraic structure of the semigroup $(R^d_+, +)$ can be written as $k(\mathbf{x},\mathbf{z}) = \phi(\mathbf{x}+\mathbf{z})$ where $\phi : \mathbb{R}^d_+ \mapsto \mathbb{R}$ is a positive definite function in the sense of satisfying $\sum_{i=1}^{n} c_i c_j \phi(\mathbf{x}_i + \mathbf{x}_j) \geq 0$ for any set of n non-negative vectors $\mathbf{x}_1 \dots \mathbf{x}_n$ and choice of real-valued scalars $c_1 \dots c_n$. The key observation is that positive-definite functions on R^d_+ are characterized by a theorem similar to Bochner's Theorem, in which the Laplace transform replaces the Fourier transform.

Theorem 4 (Berg et al.[1])**.** *A bounded continuous kernel function $k(\mathbf{x},\mathbf{z}) \equiv \phi(\mathbf{x}+\mathbf{z})$ on the Abelian semigroup $(\mathbb{R}^d_+, +)$ is positive definite if and only if it is the Laplace transform of a unique non-negative measure on $\mathbb{R}^d_+$. That*

[1]For semigroups with involution operator $*$, the condition is $\sum_{i,j=1}^{n} c_i c_j \phi(s_i^* \circ s_j)$; see Berg et al. [1].

Kernel	$k(\mathbf{x}, \mathbf{z})$	$\hat{\Psi}(\mathbf{x})$	Sampling	Ref
Gaussian	$e^{\frac{\|\mathbf{x}-\mathbf{z}\|_2^2}{2\sigma^2}}$	$\bigoplus_{j=1}^{s} \sqrt{\frac{1}{s}} e^{-i\mathbf{x}^T\mathbf{w}_j}$	$\frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{\sigma^2 w^2}{2}}$ (Normal)	[18]
Laplacian	$e^{\frac{\|\mathbf{x}-\mathbf{z}\|_1}{\sigma}}$	$\bigoplus_{j=1}^{s} \sqrt{\frac{1}{s}} e^{-\mathbf{x}^T\mathbf{w}_j}$	$\frac{\sigma}{\pi(1+\sigma^2 w^2)}$ (Cauchy)	[18]
Hellinger	$\sum_{j=1}^{d} \sqrt{x_j z_j}$	$\bigoplus_{j=1}^{d} \sqrt{x_j}$	–	[21]
χ^2	$2\sum_{j=1}^{d} \frac{x_j z_j}{x_j+z_j}$	$\bigoplus_{j=1,k}^{d} e^{iw_k \log x_j} \sqrt{x_j \operatorname{sech}(\pi w_k)}$	$w_k = kL, -r \le k \le r$	[21]
Intersection	$\sum_{j=1}^{s} \min(x_j, z_j)$	$\bigoplus_{j=1,k}^{d} e^{iw_k \log x_j} \sqrt{\frac{2x_j}{\pi(1+4w_k^2)}}$	$w_k = kL, -r \le k \le r$	[21]
Jensen-Shannon	$\sum_{j=1}^{d} h(x_j, z_j)$	$\bigoplus_{j=1,k}^{d} e^{iw_k \log x_j} \sqrt{\frac{2x_j \operatorname{sech}(\pi w_k)}{\log 4(1+4w_k^2)}}$	$w_k = kL, -r \le k \le r$	[21]
Exponentiated-χ^2	$e^{-\sigma^{-2}\sum_{j=1}^{d} \frac{(x_j-z_j)^2}{x_j+z_j}}$	$\bigoplus_{j=1}^{s} \sqrt{\frac{1}{s}} e^{-i\hat{\Psi}_{\chi^2}(\mathbf{x})^T\mathbf{w}_j}$	$\frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{\sigma^2 w^2}{2}}$ (Normal)	[21]
Jensen-Shannon	$\sum_{j=1}^{d} h(x_j, z_j)$	$\bigoplus_{j=1,k}^{d} e^{iw_k \log x_j} \sqrt{\frac{2x_j \operatorname{sech}(\pi w_k)}{\log 4(1+4w_k^2)}}$	$w_k = kL, -r \le k \le r$	[21]
Skewed-χ^2	$\prod_{j=1}^{d} \frac{2\sqrt{x_j+\epsilon}\sqrt{z_j+\epsilon}}{x_j+z_j+2\epsilon}$	$\bigoplus_{j=1}^{s} \sqrt{\frac{1}{s}} e^{-i\log(\mathbf{x}+\epsilon)^T\mathbf{w}_j}$	$\operatorname{sech}(\pi w)$ (Hyperbolic-secant)	[15]
Skewed-Interection	$\prod_{j=1}^{d} \min\left(\sqrt{\frac{x_j+\epsilon}{z_j+\epsilon}}, \sqrt{\frac{z_j+\epsilon}{x_j+\epsilon}}\right)$	$\bigoplus_{j=1}^{s} \sqrt{\frac{1}{s}} e^{-i\log(\mathbf{x}+\epsilon)^T\mathbf{w}_j}$	$\frac{2}{\pi(1+4w^2)}$ (Cauchy)	[15]
Exponential-Semigroup [1, 9]	$e^{-\beta\sum_{j=1}^{d}\sqrt{x_j+z_j}}$	$\bigoplus_{j=1}^{s} \sqrt{\frac{1}{s}} e^{-\mathbf{x}^T\mathbf{w}_j}$	$\frac{\beta}{2\sqrt{\pi}} w^{-\frac{3}{2}} e^{\frac{-\beta^2}{4w}}$ (Lévy)	**This Paper**
Reciprocal-Semigroup [1, 9]	$\prod_{j=1}^{d} \frac{\lambda}{x_j+z_j+\lambda}$	$\bigoplus_{j=1}^{s} \sqrt{\frac{1}{s}} e^{-\mathbf{x}^T\mathbf{w}_j}$	$\lambda e^{-\lambda w}$ (Exponential)	**This Paper**

Table 1. Summary of kernels and associated approximate feature maps. Above, $\bigoplus_{j=1}^{s} x_j = (x_1, \ldots, x_s), i = \sqrt{-1}, h(x,z) = \frac{x}{2}\log_2\frac{x+z}{2} + \frac{y}{2}\log_2\frac{x+z}{z}, \log(\mathbf{x}) = [\log(x_1)\ldots\log(x_d)]$. The feature map for Exponentiated-χ^2 is a composition of feature maps for the χ^2 and Gaussian kernels.

is, for any $\mathbf{x}, \mathbf{z} \in \mathbb{R}_+^d$,

$$k(\mathbf{x}, \mathbf{z}) = \int_{\mathbb{R}_+^d} e^{-(\mathbf{x}+\mathbf{z})^T\mathbf{w}} p(\mathbf{w}) d\mathbf{w} = \mathbb{E}_{\mathbf{w}\sim p} e^{-(\mathbf{x}+\mathbf{z})^T\mathbf{w}} .$$

This theorem should be contrasted with the Bochner's characterization for shift-invariant kernels on $\mathbb{R}^d$ (Theorem 1). As before, we assume without loss of generality that the non-negative measure above is a probability measure with associated density p. This result establishes one-to-one correspondence between semigroup kernels and probability densities on $\mathbb{R}^d$, via the Laplace transform. Exactly analogous to the random Fourier construction, we can now develop random Laplace feature maps via a Monte Carlo approximation,

$$k(\mathbf{x}, \mathbf{z}) \approx \frac{1}{s}\sum_{j=1}^{s} e^{-\mathbf{x}^T\mathbf{w}_j} e^{-\mathbf{z}^T\mathbf{w}_j} = \langle \hat{\Psi}(\mathbf{x}), \hat{\Psi}(\mathbf{z}) \rangle , \quad (3)$$

where the points $\mathbf{w}_j$ are drawn from p, yielding a feature map of the form

$$\hat{\Psi}(\mathbf{x}) = \frac{1}{\sqrt{s}} [e^{-\mathbf{x}^T\mathbf{w}_1} \ldots e^{-\mathbf{x}^T\mathbf{w}_s}] . \quad (4)$$

This simple algorithm is summarized in the Algorithm 1.

When the density p corresponds to Lévy or Exponential distributions, the Laplace transform provides the associated *Exponential-Semigroup* and *Reciprocal-Semigroup* kernels, respectively. The exact form of these kernels is given in Table 1. These kernels have also been studied in the context of injective Reproducing Kernel Hilbert Space (RKHS) embeddings of probability distributions on groups and semigroups [9, 6]. Through random Laplace features, one can expect to approximate these kernels well and deploy them for large-scale applications.

Algorithm 1 Random Laplace Features

Require: Characteristic kernel k on $(\mathbb{R}_+^d, +)$, size s.
Ensure: Feature map $\hat{\Psi}(\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}^s$.
1: Find p, the inverse Laplace transform of k.
2: Draw sequence $\mathbf{w}_1, \ldots, \mathbf{w}_s$ from p.
3: Set $\hat{\Psi}(\mathbf{x}) = \sqrt{\frac{1}{s}} \left[e^{-\mathbf{x}^T\mathbf{w}_1}, \ldots, e^{-\mathbf{x}^T\mathbf{w}_s}\right]$.

In the next section we bound the approximation error $|k(\mathbf{x}, \mathbf{z}) - \langle \hat{\Psi}(\mathbf{x}), \hat{\Psi}(\mathbf{z}) \rangle|$ for inputs $\mathbf{x}, \mathbf{z}$ drawn from a bounded domain in $\mathbb{R}_+^d$; we then study the empirical behaviour of the random Laplace feature map with respect to predictive tasks at hand, and benchmark its performance against several alternative approximate feature maps detailed in Table 1.

3. Uniform Convergence of Random Laplace Features

In this section, we assume that the density function can be written as $p(\mathbf{w}) = \prod_{j=1}^{d} q(w_j)$, where $q(\cdot)$ is a univariate density function. The kernel function can be written as $k(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x} + \mathbf{z})$. We assume that ϕ is a differentiable function in $\mathbb{R}_d^+/\{0\}$. Proofs for all the assertions are provided in Appendix A.

The following is a general result characterizing the error in approximating the kernel function using random Laplace features (Algorithm 1). The proof follows a similar strategy to the one used by Rahimi and Recht [18].

Theorem 5. *Let $\mathcal{M}$ be the set consisting of all the points in $\mathbb{R}^d$ satisfying $\|\mathbf{x}\|_2 \leq R$ and $x_i \geq r \geq 0$, $i = 1, \ldots, d$. Then, provided that $L_{q,r} \equiv \mathbb{E}_{w\sim q}[e^{-2wr}w^2] < \infty$ and $L_{q,r}R > \epsilon$, then for the mapping $\hat{\Psi}$ defined in Algorithm 1, we have*

$$\mathbb{P}\left[\sup_{\mathbf{x},\mathbf{z}\in\mathcal{M}} |\langle\hat{\Psi}(\mathbf{x}), \hat{\Psi}(\mathbf{z})\rangle - k(\mathbf{x},\mathbf{z})| \geq \epsilon\right] \leq 2^6 \left(\frac{dR^2 L_{q,r}}{\epsilon^2}\right) \exp\left(-\frac{s\epsilon^2}{d+2}\right), \quad (5)$$

Furthermore,

$$\sup_{\mathbf{x},\mathbf{z}\in\mathcal{M}} |\langle\hat{\Psi}(\mathbf{x}), \hat{\Psi}(\mathbf{z})\rangle - k(\mathbf{x},\mathbf{z})| < \epsilon$$

with any constant probability when $s = \Omega\left(\frac{d}{\epsilon^2}\log\frac{R^2 L}{\epsilon^2}\right)$.

Note that the quantity $L_{q,r}$ depends on the specific choice of kernel. We now give explicit expression for $L_{q,r}$ for q corresponding to two popular semigroup kernels.

The following lemma gives an explicit expression for the Exponential-Semigroup kernel.

Lemma 6. *Let $\beta > 0$, and let $q(w) = \frac{\beta}{2\sqrt{\pi}} w^{-3/2} e^{-\beta^2/4w}$ (this corresponds to the kernel $k(\mathbf{x},\mathbf{z}) = e^{-\beta\sum_{i=1}^d \sqrt{x_i+z_i}}$). For $r > 0$ we have $L_{q,r} = \frac{\beta}{4}\frac{\sqrt{2r}\beta+1}{(2r)^{\frac{3}{2}}e^{\sqrt{2r}\beta}}$.*

In the above we require all the coordinates of $\mathbf{x}$ and $\mathbf{z}$ to be positive. Furthermore, if β is fixed, $L_{q,r}$ will go to infinity as r approaches zero. To get an approximate feature map with finite error bound for the Exponential-Semigroup kernel even when some coordinates of $\mathbf{x}$ or $\mathbf{z}$ are zero, it is natural to consider building a feature map on perturbed dataset, i.e., on $\mathbf{x}+\delta$ and $\mathbf{z}+\delta$ for some small δ (the addition here denotes a component-wise addition of the scalar).

Let $\mathbf{x}' = \mathbf{x} + \delta$ and $\mathbf{z}' = \mathbf{z} + \delta$. For any approximate kernel $c(\cdot,\cdot)$, by the triangle inequality we have,

$$|k(\mathbf{x},\mathbf{z}) - c(\mathbf{x}',\mathbf{z}')| \leq |k(\mathbf{x},\mathbf{z}) - k(\mathbf{x}',\mathbf{z}')| + |k(\mathbf{x}',\mathbf{z}') - c(\mathbf{x}',\mathbf{z}')|. \quad (6)$$

So, if the kernel function is sufficiently smooth and $c(\mathbf{x}+\delta, \mathbf{z}+\delta)$ approximates $k(\mathbf{x}+\delta, \mathbf{z}+\delta)$ well, then $c(\mathbf{x}+\delta, \mathbf{z}+\delta)$ will approximate $k(\mathbf{x},\mathbf{z})$ well. In particular, for the Exponential-Semigroup kernel we have the following lemma.

Lemma 7. *Let $\mathcal{M}$ be the set consisting of all the points in $\mathbb{R}^d$ satisfying $\|\mathbf{x}\|_2 \leq R$ and $x_i \geq 0$, $i = 1, \ldots, d$. Let $\beta > 0$, and let $q(w) = \frac{\beta}{2\sqrt{\pi}} w^{-3/2} e^{-\beta^2/4w}$ (this corresponds to the kernel $k(\mathbf{x},\mathbf{z}) = e^{-\beta\sum_{i=1}^d \sqrt{x_i+z_i}}$). For $\delta = \frac{\epsilon^2}{4d^2}$ we have*

$$\mathbb{P}\left[\sup_{\mathbf{x},\mathbf{z}\in\mathcal{M}} |k(\mathbf{x},\mathbf{z}) - \langle\hat{\Psi}(\mathbf{x}+\delta), \hat{\Psi}(\mathbf{z}+\delta)\rangle| \geq \epsilon\right] \leq 1 - 2^6 \left(\frac{d(R+\delta)^2 L_{q,\delta}}{\epsilon^2/4}\right) \exp\left(-\frac{s\epsilon^2/4}{d+2}\right), \quad (7)$$

where $L_{q,\delta} = \frac{\beta}{4}\frac{\sqrt{2\delta}\beta+1}{(2\delta)^{\frac{3}{2}}e^{\sqrt{2\delta}\beta}}$.

The following lemma gives an explicit expression for the Reciprocal-Semigroup kernel.

Lemma 8. *Let $\lambda > 0$, $r \geq 0$, and let $q(w) = \lambda e^{-\lambda w}$ (this corresponds to the kernel $k(\mathbf{x},\mathbf{z}) = \prod_{i=1}^d \left(\frac{\lambda}{x_i+z_i+\lambda}\right)$). We have, $L_{q,r} = \frac{2\lambda}{(\lambda+2r)^3}$.*

4. Empirical Analysis

The number of random features controls the kernel approximation quality and the computational cost of solving a downstream task such as classification. Several empirical questions are of interest: For the same number of random features, how well do random Laplace features perform relative to other alternative feature maps on a given predictive task? How well is the underlying exact semigroup kernel approximated? Do histogram-based kernels outperform common shift-invariant kernels on $\mathbb{R}^d$ for problems of interest?

In the results reported in this section, we use the kernel name to denote the associated feature map. For example, by "Exp-Semigroup", we mean the use of random Laplace features to approximate the Exponential-Semigroup kernel. We report experiments on an image classification task (Caltech-101 [8]) and a surveillance event detection task (TRECVID SED).

4.1. Caltech-101

For this dataset, we evaluate how well inner products in the Euclidean space induced by random Laplace features approximates the true semigroup kernel; we also compare random Laplace features with other approximate feature maps on the predictive problem of classifying the 102 (101 + background) classes of the Caltech-101 benchmark dataset [8]. Our data preparation follows the one used by Vedaldi and Zisserman [21]. In particular, our results are competitive with the state of the art on this dataset for methods that use a single but strong image feature (multi-scale dense SIFT). We use the *phow_caltech* function of VLFeat[2] which rescales images to have a largest side of 480 pixels; dense SIFT features are extracted every four pixels at four scales and quantized into a 200 visual words dictionary estimated using k-means. Each image is described by a 4200-dimensional histogram of visual words with $1\times1, 2\times2$ and 4×4 spatial subdivisions.

Quality of Gram matrix approximation

Given data points $\{\mathbf{x}_i\}_{i=1}^n$, the Gram matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ is defined as $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. Suppose $\mathbf{Z} \in \mathbb{R}^{n\times s}$ is the data matrix in the induced feature space, with the i-th row

[2] http://www.vlfeat.org

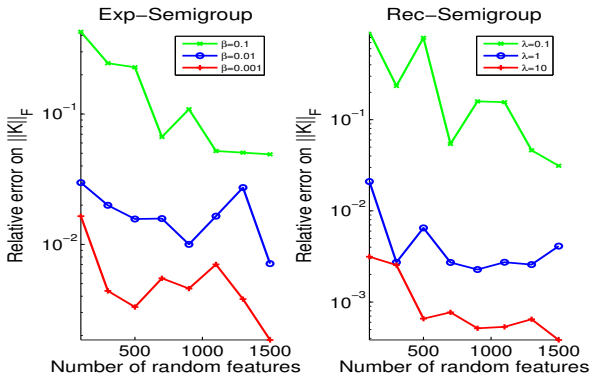

Figure 1. Gram matrix approximation relative error on Caltech-101 by using random Laplace features with s ranging from 100 to 1500. The two plots correspond to Exponential-Semigroup and Reciprocal-Semigroup kernel, respectively (see Table 1). Each curve corresponds to one value of β or λ. For the fixed pair of parameter, ten independent trials are executed and the mean is reported.

$\mathbf{Z}^{(i)} = \hat{\Psi}(\mathbf{x}_i)$ where $\hat{\Psi}(\mathbf{x}_i)$ is the random Laplace feature of $\mathbf{x}_i$ generated from Algorithm 1. We will evaluate relative error in terms of Frobenius norm, $\|\mathbf{K} - \mathbf{Z}\mathbf{Z}^T\|_F / \|\mathbf{K}\|_F$, as a function of s, as this provides an overall measure of approximation quality over the entire set of points. We consider both Exp-Semigroup and Rec-Semigroup kernels.

In Figure 1, we show the relative error of random Laplace feature with the number of random features s ranging from 100 to 1500 on the full Caltech-101 dataset comprising of $n = 3060$ samples. We can see that as s grows, random Laplace feature converges to the exact kernel quickly, meaning the relative approximation error goes to zero fast in practice. As expected, the rate of convergence depends on the choice of the kernel parameters.

Classification performance

Next, we compare SVM classification accuracies on Caltech-101 by using approximate feature maps associated with different kernels as described in Table 1. We use the usual training-test splitting protocol with 15-images per class in the training set and an equal number in the test set. SVM parameters are tuned with cross-validation. Figure 2 reports mean test set accuracy as a function of the number of random features s. We include the original input features (i.e., using linear kernel) as a baseline. Among the semigroup kernels, the Exponential-Semigroup significantly outperforms the reciprocal semigroup whose performance is below baseline levels, and hence results for the latter are ommited.

Among the seven feature maps compared in Figure 2, the random Laplace features for the Exponential-Semigroup ($\beta = 0.01$) consistently yield the highest accuracy, outperforming the Exponentiated-χ^2 kernel which is widely considered state of the art on histogram descriptors.

The Gaussian kernel does not improve over the linear kernel baseline, while kernels such as the skewed-Intersection kernel [15] perform significantly better, confirming the need to design kernels better adapted to histogram-like data.

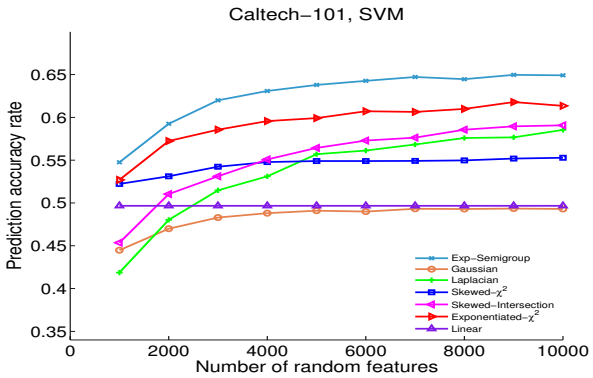

Figure 2. Prediction accuracy on Caltech-101 by using random feature maps associated to different kernels with s ranging from 1000 to 10000. The results are generated by using SVM. For a feature map and an s, five independent trials are executed and the mean is reported. For comparisons with χ^2, Intersection and Jensen-Shannon, see Table 2.

The computation time for generating random features and solving the resulting classification problem is near-identical for all feature maps shown in Figure 2, except for the Exponentiated-χ^2. While Exponentiated-χ^2 yields the second highest accuracy, its running time is six times higher than the rest, since it generates a much higher dimensional intermediate χ^2 feature map, which is then composed with the feature map for the Gaussian kernel. For the feature maps of the homogeneous additive kernels [21] which include χ^2, the number of random features that can be generated is of the form $(2r+1)d$ where d is the dimension of the original feature and r is a parameter. For high-dimensional input spaces, the resulting feature maps can be very high-dimensional and hence costly for downstream processing. For the Exponentiated-χ^2 feature map, we used $r = 3$ which corresponds to 29400 intermediate χ^2 features. The comparison with homogeneous additive kernels [21] is reported in Table 2 for $s = 12600, 21000, 29400$ corresponding to $r = 1, 2, 3$. Again, the proposed random Laplace feature maps for the Exponential-Semigroup kernel are significantly better.

KERNEL	s=12600	s=21000	s=29400
EXP-SEMIGROUP	0.6536	0.6627	0.6643
χ^2	0.6510	0.6497	0.6471
INTERSECTION	0.6399	0.6392	0.635
JENSEN-SHANNON	0.6510	0.6477	0.6477

Table 2. Caltech 101 SVM accuracies: Comparison against χ^2, Intersection and Jensen-Shannon can only be done for $s = (2r+1)d$. Here, $r = 1, 2, 3$. For a feature map and an s, five independent trials are executed and the mean is reported.

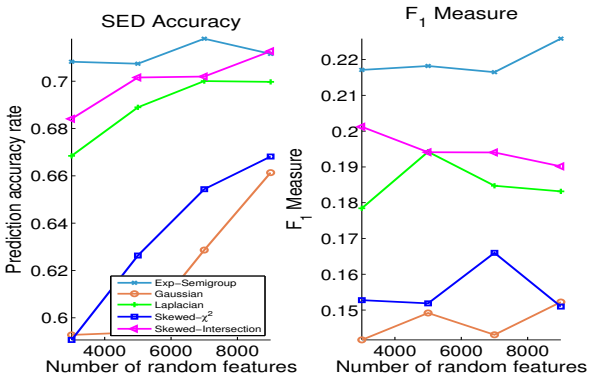

Figure 3. Prediction accuracy and F_1 score on SED by using random feature maps associated to different kernels with s ranging from 3000 to 9000. The two subplots are results of Prediction accuracy and F_1 score respectively. For a feature map and an s, five independent trials are executed and the mean is reported.

4.2. Surveillance Event Detection (SED)

There are seven target events in the TRECVED Surveillance Event Detection (SED) dataset, i.e., *CellToEar*, *Embrace*, *ObjectPut*, *Pointing*, *PeopleMeet*, *PeopleSplitUp* and *PersonRuns*. The dataset was captured in five locations at a busy airport. Many confounding issues exist in this dataset such as high activity levels, camera view changes, large variances in how events play out (e.g., "PeopleMeet") and small objects carrying predictive signals (i.e., "CellToEar"). The development set consists of 100 hours of video and the evaluation set has an additional 50 hours of data. The annotations of the dataset only include temporal extents and event labels, and no localization information is provided for events. We used the development set in our experiments, and divided it into two equal part for validation.

The training size of SED is 7024, containing approximately even number of event and non-event samples. The test size is 22437, containing 5381 events instances. By event sample, we mean an observation coming from any of the seven events described above. We use bag of visual words on motion-SIFT features resulting in final dimensionality of $d = 24000$.

Due to the high dimensionality of the SED dataset, we do not generate the homogeneous additive and Exponentiated-χ^2 kernels in this case. We report the results for a regularized least squares classification model (SVMs perform similarly) with parameters tuned using cross-validation. The test accuracies and F_1 scores are shown in Figure 3. As before, random Laplace feature maps for the exponential semigroup kernel ($\beta = 0.001$) outperform other feature maps in this task.

5. Conclusions

Our empirical results strongly suggest that the proposed random Laplace features for semigroup kernels are a valuable addition to the library of approximate feature maps proposed in the literature for scaling up kernel methods. These semigroup kernels are particularly well-suited to data representations in the form of non-negative attributes and histograms. The scalability of this approach can be further improved via design of specialized parallel solvers [12] to handle a larger number of random features, while replacing Monte Carlo approximations with more efficient numerical integration techniques [22]. We plan to investigate a broader family of semigroup kernels on $\mathbb{R}^d_+$ and benchmark their performance across several applications.

A. Technical Details

In this section we provide the proofs of Theorem 5, Lemma 6, Lemma 7 and Lemma 8.

A.1. Proof of Theorem 5

Let $c(\sigma) = \frac{1}{s}\sum_{i=1}^{s} g_i(\sigma)$ where $g_i(\sigma) = e^{-\sigma^T \mathbf{w}_i}$. It is easy to see that for any $\mathbf{x} + \mathbf{z} = \sigma$ we have $c(\sigma) = \langle \hat{\Psi}(\mathbf{x}), \hat{\Psi}(\mathbf{z}) \rangle$. The function k is additive-invariant, so we can abuse notation and write $k(\sigma) = k(\mathbf{x}, \mathbf{z})$ for $\mathbf{x} + \mathbf{z} = \sigma$. Now, let us denote $f(\sigma) = c(\sigma) - k(\sigma)$.

Let $\mathcal{M}_\sigma = \{\mathbf{x} + \mathbf{z} \mid \mathbf{x}, \mathbf{z} \in \mathcal{M}\}$. It is easy to see that $\mathcal{M}_\sigma$ is a subset of $\mathcal{T} = \{\sigma \mid \|\sigma\| \leq 2R$ and $\sigma_i \geq r, i = 1, \ldots, d\}$. Our goal is to show that with probability specified in (5),

$$|f(\sigma)| < \epsilon, \quad \forall \sigma \in \mathcal{T} \,. \tag{8}$$

Since $\mathrm{diam}(\mathcal{T}) \leq 4R$ and it is closed in $\mathbb{R}^d$, and hence it is compact. We can construct an ϵ-net over $\mathcal{T}$ using less than $J = (4\mathrm{diam}(\mathcal{T})/\gamma)^d$ balls with radius γ [5]. Denote the anchors of the net by $\{\eta_i\}_{i=1}^J$. We bound $|f(\sigma)|$ uniformly (with high probability) by showing that the following two claims hold with high probability:

- For $i = 1, \ldots, J$, $|f(\eta_i)| < \epsilon/2$;
- For $\forall \sigma \in \mathcal{M}_\sigma$, $\|\nabla f(\sigma)\| < \epsilon/2\gamma$.

These are sufficient, since for any $\sigma \in \mathcal{T}$, let η_i be the nearest anchor to σ. It satisfies $\|\sigma - \eta_i\| < \gamma$, so,

$$\begin{aligned} \|f(\sigma)\| &= \|f(\eta_i) + f(\sigma) - f(\eta_i)\| \\ &\leq \|f(\eta_i)\| + \|f(\sigma) - f(\eta_i)\| \\ &= \|f(\eta_i)\| + \|\nabla f(\xi)^T(\sigma - \eta_i)\| \\ &\leq \|f(\eta_i)\| + \|\nabla f(\xi)\| \cdot \|\sigma - \eta_i\| \\ &\leq \epsilon/2 + \gamma \cdot \epsilon/2\gamma = \epsilon \end{aligned} \tag{9}$$

for some ξ on the line connecting σ and η_i (note that $\mathcal{T}$ is convex, so the line is contained in it). The second equality uses the mean-value theorem.

To show the first claim, it is sufficient to show $|f(\sigma)| < \epsilon$ holds with high probability for any (fixed) σ. The claim then follows using a union bound. For any $\sigma \in \mathcal{T}$, we have

$$f(\sigma) = c(\sigma) - k(\sigma) = \frac{1}{s}\sum_{i=1}^{s} (g_i(\sigma) - k(\sigma)) \,. \tag{10}$$

From Theorem 4, it is not hard to show that $\mathbb{E}_{\mathbf{w}_i}[g_i(\sigma)] = k(\sigma)$, for $i = 1, \ldots, s$. Also, $|g_i(\sigma)| \leq 1$. By Hoeffding inequality, we have

$$\mathbb{P}\left\{\left|\frac{1}{s}\sum_{i=1}^{s} g_i(\sigma) - k(\sigma)\right| > \epsilon\right\} \leq 2e^{-2s\epsilon^2} . \quad (11)$$

By a union bound, the first claim holds with probability at least $1 - 2Je^{-s\epsilon^2/2}$.

To show the second claim, we need to bound the Lipschitz constant of f uniformly. Let $\lambda = \arg\max_{\sigma\in\mathcal{T}} \|\nabla f(\sigma)\|$. Since $\mathbb{E}[c(\lambda)] = k(\lambda)$, we have $\mathbb{E}[\nabla c(\lambda)] = \nabla k(\lambda)$, as

$$\begin{aligned}\frac{\partial k(\lambda)}{\partial \lambda_i} &= \frac{\partial}{\partial \lambda_i}\int_{\mathbb{R}_+^d} e^{-\lambda^T\mathbf{w}} p(\mathbf{w}) d\mathbf{w} \\ &= \int_{\mathbb{R}_+^d} w_i e^{-\lambda^T\mathbf{w}} p(\mathbf{w}) d\mathbf{w} \\ &= \mathbb{E}[w_i e^{-\lambda^T\mathbf{w}}] = \mathbb{E}\left[\frac{\partial c(\lambda)}{\partial \lambda_i}\right] . \quad (12)\end{aligned}$$

The interchange between the integral and derivative is allowed since the functions $\mathbf{w}, \lambda \mapsto e^{-\lambda^T\mathbf{w}} p(\mathbf{w})$ and $\mathbf{w}, \lambda \mapsto w_i e^{-\lambda^T\mathbf{w}} p(\mathbf{w})$ are both continuous on λ and $\mathbf{w}$.

Taking the expectation on $\|\nabla f(\lambda)\|^2 = \|\nabla c(\lambda) - \nabla k(\lambda)\|^2$, we have

$$\begin{aligned}\mathbb{E}[\|\nabla f(\lambda)\|^2] &= \mathbb{E}[\|\nabla c(\lambda)\|^2 - 2\nabla c(\lambda)^T \nabla f(\lambda) + \|\nabla k(\lambda)\|^2] \\ &= \mathbb{E}[\|\nabla c(\lambda)\|^2] - \|\nabla k(\lambda)\|^2 \\ &\leq \mathbb{E}[\|\nabla c(\lambda)\|^2] \\ &= \mathbb{E}\left[\sum_{j=1}^{d}\left(\frac{\partial c(\lambda)}{\partial \lambda_j}\right)^2\right] . \quad (13)\end{aligned}$$

Recall that $c(\lambda) = \frac{1}{s}\sum_{i=1}^{s} g_i(\lambda)$, so $\frac{\partial c(\lambda)}{\partial \lambda_j} = \frac{1}{s}\sum_{i=1}^{s}\frac{\partial g_i(\lambda)}{\partial \lambda_i}$. As $g_i(\lambda) = e^{-\lambda^T\mathbf{w}_i}$, we have

$$\frac{\partial g_i(\lambda)}{\partial \lambda_j} = -e^{-\lambda^T\mathbf{w}_i} w_{ij} . \quad (14)$$

Hence, continuing (13), we have

$$\begin{aligned}\mathbb{E}[\|\nabla f(\lambda)\|^2] &\leq \frac{1}{s^2}\mathbb{E}\left[\sum_{j=1}^{d}\left(\sum_{i=1}^{s} e^{-\lambda^T\mathbf{w}_i} w_{ij}\right)^2\right] \\ &\leq \frac{1}{s^2}\mathbb{E}\left[\sum_{j=1}^{d}\left(\sum_{i=1}^{s} e^{-w_{ij}\lambda_j} w_{ij}\right)^2\right] \\ &\leq \frac{1}{s^2}\mathbb{E}\left[\sum_{j=1}^{d}\left(\sum_{i=1}^{s} e^{-w_{ij} r} w_{ij}\right)^2\right] \\ &= \frac{1}{s^2}\sum_{j=1}^{d}\mathbb{E}\left[\left(\sum_{i=1}^{s} e^{-w_{ij} r} w_{ij}\right)^2\right] . \quad (15)\end{aligned}$$

In the above inequalities we used the facts that w_{ij}, λ_j are positive and $\lambda_j \geq r$.

By our assumption, w_{ij} are i.i.d. variables with density q. Hence the expectations in the summand above are identical. Let ν be a random variable with density function q. We have

$$\mathbb{E}[\|\nabla f(\lambda)\|^2] \leq \frac{d}{s^2}\mathbb{E}\left[\left(\sum_{i=1}^{s} e^{-\nu_i r}\nu_i\right)^2\right] . \quad (16)$$

Expanding the last inequality and using Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}[\|\nabla f(\lambda)\|^2] &= \frac{d}{s^2}\left(s\mathbb{E}[(e^{-\nu r}\nu)^2] + (s^2 - s)(\mathbb{E}[e^{-\nu r}\nu])^2\right) \\ &\leq d \cdot \mathbb{E}[e^{-2\nu r}\nu^2] \\ &= dL_{q,r} . \quad (17)\end{aligned}$$

By Markov's inequality, we have

$$\mathbb{P}\{\|\nabla f(\lambda)\|^2 > \epsilon^2/4\gamma^2\} \leq 4\gamma^2 L_{q,r}/\epsilon^2 . \quad (18)$$

Overall, with probability at least $1 - 2\left(\frac{16R}{\gamma}\right)^d e^{-s\epsilon^2/2} - 4\gamma^2 L_{q,r}/\epsilon^2$, the two events will hold simultaneously. By setting $\gamma = (16R)^{\frac{d}{d+2}}(2e^{-s\epsilon^2/2}\gamma^2 L_{q,r})^{\frac{1}{d+2}}$, and since $L_{q,r}R > \epsilon$, the success probability is at least

$$1 - 2^6\left(\frac{R^2 L_{q,r}}{\epsilon^2}\right)^2 e^{-\frac{s\epsilon^2}{d+2}} . \quad (19)$$

The second part of the theorem follows by fixing the failure probability and solving for s.

A.2. Proof of Lemma 6

By the definition of $L_{q,r}$, we have

$$\begin{aligned}L_{q,r} &= \mathbb{E}[e^{-2wr}w^2] \\ &= \frac{\beta}{2\sqrt{\pi}}\int_0^\infty e^{-2wr} w^2 w^{-\frac{3}{2}} e^{-\beta^2/4w} dw \\ &= \frac{\beta}{2\sqrt{\pi}}\int_0^\infty e^{-2wr-\beta^2/4w} w^{\frac{1}{2}} dw \\ &= \frac{\beta}{2\sqrt{\pi}} c^{\frac{3}{2}}\int_0^\infty e^{-(2rc)(w+1/w)} w^{\frac{1}{2}} dw \\ &= \frac{\beta}{4}\frac{\sqrt{2r}\beta + 1}{(2r)^{\frac{3}{2}} e^{\sqrt{2r}\beta}} . \quad (20)\end{aligned}$$

Above, $c = \frac{\beta}{2\sqrt{2\sigma_i}}$ and we use the fact that

$$\begin{aligned}&\int e^{-a(x+1/x)} x^{1/2} dx \\ &= -\frac{\sqrt{\pi}e^{-2a}}{4a^{3/2}}\left((2a+1)erf(\sqrt{a}(1/\sqrt{x} - \sqrt{x})\right) - \\ &\frac{\sqrt{\pi}e^{-2a}}{4a^{3/2}}\left((2a-1)e^{4a}erf(\sqrt{a}(1/\sqrt{x} + \sqrt{x}))\right) - \\ &\frac{\sqrt{x}e^{-a(x+1/x)}}{a} + C . \quad (21)\end{aligned}$$

A.3. Proof of Lemma 7

Let $\phi(\sigma) = e^{-\beta \sum_{i=1}^{d} \sqrt{\sigma_i}}$. It is easy to verify that for $\rho = \epsilon^2/d^2$, we have $|\phi(\sigma + \rho) - \phi(\sigma)| < \epsilon$ for $\sigma \in \mathbb{R}^d_+$. Hence, for $\delta = \frac{\epsilon^2}{4d^2}$ we have $|k(\mathbf{x}, \mathbf{z}) - k(\mathbf{x}+\delta, \mathbf{z}+\delta)| < \epsilon/2$. Now, applying Lemma 5 to the set $\mathcal{M} + \delta$ shows that with the specified probability, $|k(\mathbf{x}+\delta, \mathbf{z}+\delta) - c(\mathbf{x}+\delta, \mathbf{z}+\delta)| < \epsilon/2$. The bound now follows from (6).

A.4. Proof of Lemma 8

By the definition of $L_{q,r}$, we have

$$\begin{aligned} L_{q,r} &= \mathbb{E}[e^{-2wr} w^2] \\ &= \lambda \int_0^\infty e^{-2wr} e^{-\lambda w} w^2 dw \\ &= \lambda \int_0^\infty e^{-(2r+\lambda)w} w^2 dw \\ &= \frac{2\lambda}{(\lambda + 2r)^3} \,. \end{aligned} \tag{22}$$

In the above, we use the fact that

$$\int e^{ax} x^2 dx = \frac{e^{ax}(a^2x^2 - 2ax + 2)}{a^3} + C \,. \tag{23}$$

References

[1] C. Berg, J. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups: Theory of Positive-definite and related functions*. Springer-Verlag, 1984. 2, 3

[2] S. Bochner. Monotone funktionen, Stieltjes integrale und harmonische analyse. *Math. Ann.*, 108:378–410, 1933. 1

[3] O. Chapelle. Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10:1055 – 1064, 1999. 2

[4] G. Csurka, C. Dance, L. Dan, J. Williamowski, and C. Bray. Visual categorization with bags of keypoints. In *European Conference on Computer Vision*, 2004. 2

[5] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Soc.*, 39:1–49, 2001. 6

[6] S. Danafar, A. Gretton, and J. Schmidhuber. Characteristic kernels on structured domains excel in robotics and human action recognition. *Machine Learning and Knowledge Discovery in Databases*, pages 264–279, 2010. 3

[7] W. Deng, R. Dong, L. Socher, K. Li, and L. Fei-Fei. Imagenet: A large-scale heirarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 1

[8] L. Fei-Fei, R. Fergus, and P. Perona. A Bayesian approach to unsupervised one-shot learning of object categories. In *International Conference on Computer Vision*, 2003. 4

[9] K. Fukumiuzu, B. Sriperumbudur, A. Gretton, and B. Scholkopf. Characteristic kernels on groups and semigroups. In *Neural Information Processing Systems*, 2008. 3

[10] P. V. Gehler and S. Nowozin. On feature combination methods for multiclass object classification. In *International Conference on Computer Vision*, 2009. 1

[11] K. Grauman and T. Darrel. The pyramid match kernel: Discriminative classification with sets of image features. In *International Conference on Computer Vision*, 2005. 2

[12] P. Huang, H. Avron, T. Sainath, V. Sindhwani, and B. Ramabhadran. Kernel methods match deep neural networks on timit. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2014. 1, 6

[13] T. Joachims. Training linear svms in linear time. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2006. 1

[14] C. H. Lampert. *Kernel Methods in Computer Vision*. Foundations and Trends in Computer Graphics and Vision, 2009. 1

[15] F. Li, C. Ionescu, and C. Sminchisescu. Random Fourier approximations for skewed multiplicative histogram kernels. *Pattern Recognition*, 6376:262–271, 2010. 1, 2, 3, 5

[16] F. Li, G. Lebanon, and C. Sminchisescu. Chebyshev approximation to the histogram χ^2 kernel. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 2

[17] S. Maji and A. Berg. Max-margin additive classifiers for detection. In *International Conference on Computer Vision*, 2009. 1, 2

[18] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Neural Information Processing Systems*, 2007. 1, 2, 3

[19] B. Schoelkopf and A. Smola, editors. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002. 1

[20] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008. 1

[21] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3(34):480–492, 2012. 1, 2, 3, 4, 5

[22] J. Yang, V. Sindhwani, H. Avron, and M. Mahoney. Quasi-monte carlo feature maps for shift-invariant kernels. In *International Conference on Machine Learning*, 2014. 6