Random Fourier Approximations for Skewed Multiplicative Histogram Kernels

Fuxin Li, Catalin Ionescu, Cristian Sminchisescu

Institut für Numerische Simulation, Universität Bonn

Abstract. Approximations based on random Fourier features have recently emerged as an efficient and elegant methodology for designing large-scale kernel machines [4]. By expressing the kernel as a Fourier expansion, features are generated based on a finite set of random basis projections with inner products that are Monte Carlo approximations to the original kernel. However, the original Fourier features are only applicable to translation-invariant kernels and are not suitable for histograms that are always non-negative. This paper extends the concept of translation-invariance and the random Fourier feature methodology to arbitrary, locally compact Abelian groups. Based on empirical observations drawn from the exponentiated χ^2 kernel, the state-of-the-art for histogram descriptors, we propose a new group called the *skewed-multiplicative group* and design translation-invariant kernels on it. Experiments show that the proposed kernels outperform other kernels that can be similarly approximated. In a semantic segmentation experiment on the PASCAL VOC 2009 dataset, the approximation allows us to train large-scale learning machines more than two orders of magnitude faster than previous nonlinear SVMs.

1 Introduction

In recent years, datasets containing large amounts of labeled data are increasingly common in learning problems, such as text classification [1], spam filtering [2] and visual object recognition [3]. It is however difficult to apply high-performance kernel methods to these tasks, as the constraint to operate with the kernel matrix makes such methods scale more than quadratically in the size of the dataset. A number of recent algorithms perform explicit feature transforms [4–6], so that nonlinear kernels can be approximated by linear kernels in the transformed space. This makes possible to use efficient linear methods that depend only linearly on the size of the training set [7, 8]. If the approximations are accurate, complex nonlinear functions can be learned using linear algorithms, thus allowing to solve large-scale learning problems efficiently.

Random Fourier approximations (RF) provides an elegant and efficient methodology to create explicit feature transforms. By applying Bochner's theorem, translation-invariant kernels are computed as inner products in the frequency domain (after a Fourier transform). Then m-dimensional feature vectors are created for examples so that their inner products are Monte Carlo approximations

of the original kernel. The method has the convergence rate of Monte Carlo: $O(m^{-\frac{1}{2}})$ independent of the input dimension. One usually needs only a few hundred dimensions to approximate the original kernel accurately.

Previously, RF were developed for translation-invariant kernels on $\mathbb{R}^n$. In this paper we study the applicability of RF for histogram features where it is known that kernels defined on $\mathbb{R}^n$ do not usually produce good results [9]. The best performing kernel to-date on histogram features [9] is the exponentiated χ^2 kernel [10]. However, this kernel cannot be approximated with RF. Our aim is to design a kernel that has similar performance, but fits within the RF framework.

We first extend the random Fourier feature methodology to translation-invariant kernels on general locally compact Abelian groups. It is hypothesized that two factors are important for the performance of the χ^2 kernel: the sensitivity to the scale of the features and the multiplicative decomposition as a product of components along each dimension, instead of a sum. Therefore we design a new group called the *skewed multiplicative group*, which has built-in sensitivity to feature scale. We propose multiplicative kernels on this group and apply the RF framework on it.

In experiments, we show that our designed kernels are easy to approximate, have better performance than other kernels usable within the RF framework, and offer a substantial speed-up over previous nonlinear SVM approaches.

2 Fourier Transform and Random Features on Groups

We use n to denote the number of training examples, d the input dimensionality and m the dimensionality of the extracted random features. $\mathcal{F}[f]$ denotes the Fourier transform of f, and $\mathrm{U}[a,b]$ is the uniform distribution on $[a,b]$. $\mathbf{E}_\mu[x]$ takes the expectation of x w.r.t to the measure μ.

2.1 Fourier Transform on Groups

Let $(G,+)$ be any locally compact abelian (LCA) group, with 0 the identity. There exists a non-negative regular measure m called the *Haar measure* of G, which is translation-invariant: $m(E+x) = m(E)$ for every $x \in G$ and every Borel set E in G. The Haar measure is provably unique up to a multiplicative positive constant, and is required in the *Haar integral*: $\int_G f(x)dm$, essentially a Lebesgue integral on the Haar measure [11].

Now we establish the character and (Pontryagin) dual group of G [11]. A complex function γ on G is called a *character* if $|\gamma(x)| = 1$ for all $x \in G$ and if

$$\gamma(x+y) = \gamma(x)\gamma(y), \forall x, y \in G \tag{1}$$

All complex $\gamma(x)$ with $|\gamma(x)| = 1$ can be represented as $\gamma(x) = e^{ig(x)}$. Therefore to make a unique character, only a real-valued $g(x)$ needs to be decided. The set of all continuous characters of G forms the *dual group* Γ, where addition is defined by $(\gamma_1 + \gamma_2)(x) = \gamma_1(x)\gamma_2(x)$. It follows that Γ is also an LCA group. To emphasize duality, we write $(x,\gamma) = \gamma(x)$.

For all f that are integrable on G, the function $\mathcal{F}$ defined on Γ by

$$\mathcal{F}[f](\gamma) = \frac{1}{2\pi} \int_G f(x)(-x, \gamma)dm \tag{2}$$

is called the *Fourier transform* of f.

The simplest example is $\mathbb{R}$, $\gamma_\eta(x) = e^{\eta x i}$, where η is an arbitrary number. Eq. (1) could easily be verified, and (2) becomes the conventional Fourier transform $\mathcal{F}[f](\gamma_\eta) = \frac{1}{2\pi} \int_{\mathbb{R}} f(x)e^{-\eta x i}dx$.

2.2 Random Features on Groups

Now we introduce Bochner's theorem which is the main result we need [11]:

Theorem 1. *A continuous function f on G is positive-definite if and only if there is a non-negative measure μ on Γ such that $f(x) = \int_\Gamma (x, \gamma)d\mu(\gamma)$.*

Usually one is able to verify if a translation-invariant kernel $k(x, y) = f(x-y)$ is positive-definite. For such kernels, we can use Bochner's theorem for the explicit feature transform [4]:

$$k(x - y) = \int_G (y - x, \gamma)d\mu(\gamma) = \mathbf{E}_\mu[\zeta_\gamma(x)\zeta_\gamma(y)^*], \tag{3}$$

where $\zeta_\gamma(x) = (-x, \gamma)$ and $*$ is the conjugate. To construct ζ_γ explicitly, note that $(-x, \gamma) = e^{-ig_\gamma(x)} = \cos(g_\gamma(x)) - i\sin(g_\gamma(x))$. Then, $k(x-y) = \mathbf{E}_\mu[\cos(g_\gamma(x) - g_\gamma(y))] + i\mathbf{E}_\mu[\sin(g_\gamma(x) - g_\gamma(y))]$. For the real kernels we work with, the imaginary part must be zero. Therefore we only need to approximate the real part $\mathbf{E}_\mu[\cos(g_\gamma(x) - g_\gamma(y))]$. Define

$$z_\gamma(x) = \cos(g_\gamma(x) + b), \tag{4}$$

where $b \sim \mathrm{U}[0, 2\pi]$. It follows that $\mathbf{E}_\mu[\cos(g_\gamma(x) - g_\gamma(y))] = \mathbf{E}_\mu[z_\gamma(x)z_\gamma(y)]$, thus (4) is the explicit transform we seek. To approximate the expectation $\mathbf{E}_\mu[\zeta_\gamma(x)\zeta_\gamma(y)^*]$, we sample from the distribution μ. In principle, the expectation can be approximated by linear functions on explicit features:

$$Z_x = \left[\cos(g_{\gamma_1}(x) + b_1), \cos(g_{\gamma_2}(x) + b_2), \ldots, \cos(g_{\gamma_k}(x) + b_k)\right] \tag{5}$$

Basically, the algorithm has the following steps: 1) Generate k random samples $\gamma_1, \ldots, \gamma_k$ from the distribution μ; 2) Compute Z_x as the RF feature for all training examples and use linear methods to perform the learning task. In practice, g_γ uniquely decides γ, and the group G defines the form of g_γ. In $\mathbb{R}^d$ for example, the form is $g_\gamma(x) = r_\gamma^{\mathrm{T}} x$, where r_γ is a real vector with the same length as x [4]. Therefore, sampling only needs to be done on r_γ. The distribution is decided by the Fourier transform of the kernel. For example, in the case of a Gaussian kernel, the distribution is still Gaussian. See [4] for details on other kernels.

3 The Skewed Multiplicative Group

3.1 Fourier Transform of the Skewed Multiplicative Group

We make use of a group operation that combines multiplication and addition, The inclusion of an additive part makes the group sensitive to scaling:

$$x \otimes y = (x + c)(y + c) - c \tag{6}$$

The group is defined on $(-c, \infty)$ with $c \geq 0$. Here $1 - c$ is the identity element, since $(1 - c) \otimes y = y$. Then x^{-1} could be solved from $x \otimes x^{-1} = 1 - c$, to obtain $x^{-1} = \frac{1}{x+c} - c$. Therefore, the translation-invariant kernel on this group is

$$k(x, y) = f(x \otimes y^{-1}) = f\left(\frac{x + c}{y + c} - c\right), \tag{7}$$

The Haar measure and the Fourier transform are given next.

Proposition 1. *On the skewed multiplicative group $((-c, \infty), \otimes)$, the following results hold:*
1) The Haar measure is given by $\mu(S) = \int_S \frac{1}{t+c} dt$.
2) The characters are $\gamma_\eta(x) = e^{\eta \log(x+c)i}$, with $\eta \in \mathbb{R}$.
3) The Fourier transform is given by $\mathcal{F}[f](\eta) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(e^x - c)e^{-\eta x i} dx$.

Proof. 1) Since $m([d \otimes a, d \otimes b]) = \int_{(d+c)(a+c)-c}^{(d+c)(b+c)-c} \frac{1}{t+c} dt = \log(d + c)(b + c) - \log(d + c)(a + c) = \log(b + c) - \log(a + c) = \int_a^b \frac{1}{t+c} dt = m([a, b])$, the measure is translation-invariant. Since the Haar measure is unique, we conclude that $\mu(S) = \int_S \frac{1}{t+c} dt$ is the Haar measure on the group.
2) We only need to verify (1): $\gamma_\eta(x \otimes y) = e^{\eta \log((x+c)(y+c))i}$
$= e^{\eta \log(x+c)i} e^{\eta \log(y+c)i} = \gamma_\eta(x)\gamma_\eta(y)$.
3) From (2), $\mathcal{F}[f](\eta) = \int_{-c}^{\infty} \frac{f(x)}{x+c} e^{-\eta \log(x+c)i} dx = \int_{-c}^{\infty} f(x) e^{-\eta \log(x+c)i} d(\log(x + c)) = \int_{-\infty}^{\infty} f(e^x - c)e^{-\eta x i} dx$.

When $c = 0$, we obtain the regular multiplicative group on $\mathbb{R}^+$, denoted as $(\mathbb{R}^+, \times)$. The identity on this group is 1. The translation-invariance property in this group is scale invariance, since translation-invariant kernels have $k(x, y) = f(x \times y^{-1}) = f(\frac{x}{y}) = f(\frac{d \times x}{d \times y})$. For this group, the Fourier transform is known to be $\mathcal{F}[f](e^x)$ in $\mathbb{R}$ [19] which is equivalent to Proposition 1.

3.2 Kernels

Only a few functions have explicit Fourier transforms. Here we consider two functions

$$f_1(x) = \frac{2}{\sqrt{x + c} + \sqrt{\frac{1}{x+c}}}, f_2(x) = \min\left(\sqrt{x + c}, \frac{1}{\sqrt{x + c}}\right) \tag{8}$$

which correspond to kernels that we refer as the *skewed* χ^2 and the *skewed intersection* kernels, respectively:

$$k_1(x,y) = \frac{2\sqrt{x+c}\sqrt{y+c}}{x+y+2c}, k_2(x,y) = \min\left(\sqrt{\frac{x+c}{y+c}}, \sqrt{\frac{y+c}{x+c}}\right) \tag{9}$$

From Proposition 1, the corresponding Fourier transforms can be computed. In this case, they are the hyperbolic secant and Cauchy distributions, respectively:

$$\mathcal{F}_1(\omega) = \operatorname{sech}(\pi\omega), \mathcal{F}_2(\omega) = \frac{2}{\pi(1+4\omega_i^2)} \tag{10}$$

The multidimensional kernels are defined as a product of one-dimensional kernels: $k(\mathbf{x},\mathbf{y}) = \prod_{i=1}^{d} k(x_i, y_i)$, where d is the dimensionality of the data. The multi-dimensional Fourier transform is just the product of the transform on each dimension, $\mathcal{F}(\omega) = \prod_{i=1}^{d} \mathcal{F}(\omega_i)$. In the case of $\mathcal{F}_1(\omega)$ and $\mathcal{F}_2(\omega)$, this just means that the Fourier transform of the kernel is a joint distribution on ω, where each dimension is independent of others.

In the skewed multiplicative group, the form of g_γ is $g_\gamma(x) = r_\gamma^{\mathrm{T}} \log(x+c)$. To apply the RF methodology, one would need to sample from (10), in order to obtain r_γ to compute the random features (5). We use the inverse transformation method: sampling uniformly from $\mathrm{U}[0,1]$ and transforming the samples by multiplying with the inverse CDF of the distribution.

4 Motivation for the skewed approximations

The exponentiated χ^2 kernel $k(x,y) = \exp(-\sum_i \frac{(x_i-y_i)^2}{x_i+y_i})$ has achieved the best performance to-date on histogram features for visual object detection and recognition [9]. However, for the multiplicative group of $\mathbb{R}^+$, we would need to compute the equivalent Fourier transform of $f(\exp(\frac{x}{y}))$ in $\mathbb{R}$. In the case of any exponentiated kernel, we might need to compute the Fourier transform of a function represented as $\exp(-\gamma g(\exp(\frac{x}{y})))$, for some $g(x)$. With two exponentials, it is difficult to find analytical forms for the transform.

Our motivation is to design a kernel within the RF framework that preserves some properties of the χ^2 kernel, while being at the same time tractable to approximate. To do this, we develop some intuition on why the χ^2 kernel works better than others. First, we conjecture that the exponentiated χ^2 kernel works well because *it adapts to different scales* in the input features. Secondly, we conjecture that its *multiplicative* properties might be an advantage over additive kernels. We will explain these two conjectures in the sequel.

The scale in a histogram feature is proportional to the number of occurrences of a random variable (its frequency). The χ^2 kernel is based on the Pearson χ^2 test, designed to favor variables that are observed more frequently. The gist is that higher frequencies are more stable finite-sample estimators of probabilities. Hence, kernel dimensions with higher frequency should be emphasized when two histograms are compared. The translation-invariant Gaussian kernel does not

Table 1. A list of kernels used in visual recognition. Previous work empirically showed that the exponentiated χ^2 kernel performs best among the kernels listed.

Name	$k(x,y)$	Group	Mult. or Add.	RF proved in
Gaussian	$\exp(-\gamma\|\|x-y\|\|^2)$	$(\mathbb{R}^d,+)$	Multiplicative	[4]
$1-\chi^2$	$\sum_i \left(1-\frac{(x_i-y_i)^2}{x_i+y_i}\right)$	$(\mathbb{R}_+^d,\times)$	Additive	[12]
Exponentiated χ^2	$\exp\left(-\gamma\sum_i \frac{(x_i-y_i)^2}{x_i+y_i}\right)$	N/A	Multiplicative	N/A
Intersection	$\sum_i \min(x_i,y_i)$	$(\mathbb{R}_+^d,\times)$	Additive	[12]
Linear Kernel	$\sum_i x_i y_i$	N/A	Additive	N/A
Skewed-χ^2	$\prod_i \frac{2\sqrt{x_i+c}\sqrt{y_i+c}}{x_i+y_i+2c}$	$((-c,\infty),\otimes)$	Multiplicative	This paper
Skewed-Intersection	$\prod_i \min\left(\sqrt{\frac{x_i+c}{y_i+c}},\sqrt{\frac{y_i+c}{x_i+c}}\right)$	$((-c,\infty),\otimes)$	Multiplicative	This paper

have this property. This may explain why the χ^2 kernel significantly outperforms the Gaussian in visual learning problems.

In Table 1, several other kernels that adapt to the scale of the features are shown. E. g., the $1-\chi^2$ kernel is based on exactly the same χ^2 statistic as the exponentiated one. We conjecture that one difference is important: the $1-\chi^2$ kernel and the other kernels are additive, i.e. the kernel value on multiple dimensions is a sum of the kernel value on each dimension. In contrast, the exponentiated kernel is multiplicative: $\exp\left(-\gamma\sum_i \frac{(x_i-y_i)^2}{x_i+y_i}\right)=\prod_i \exp\left(-\gamma\frac{(x_i-y_i)^2}{x_i+y_i}\right)$.

Moreover, we argue that a multiplicative kernel is more sensitive to large deviations between x and y in one or a few dimensions. Assuming $\chi^2(x_i,y_i)\leq\chi^2(x_u,y_u)$ for all i, we have $\exp(-\gamma\chi^2(x,y))\leq\exp(-\gamma\chi^2(x_u,y_u))$. Therefore, one extremely noisy dimension may negatively impact the exponentiated kernel severely. Otherwise said, to make $k(x,y)$ large (i.e., x,y similar), the two histograms must be similar in almost all dimensions. For an additive kernel, this effect is much less obvious. $k(x,y)$ is high if x and y match on some important bins, but not necessarily all.

Why is matching all bins important? Intuitively, in localization tasks, under relatively weak models, the number of negative object hypotheses one must go over is usually huge. Therefore, if the similarity between two object hypotheses is large when they are only partially matched, there might be simply too many hypotheses with good similarity to the ground truth. In such circumstances the false positive rate may increase significantly.

5 Related Work

RF belongs to the class of methods that replace the kernel with a low-rank approximation. In [13, 14], the authors proposed incomplete Cholesky decomposition methods that compute a low-rank approximation to the kernel matrix while simultaneously solving the SVM optimization. These methods are computationally powerful but to predict new data, kernel values still have to be computed between all test and training examples, which is slow for large-scale problems. Alternatively, one can use Nyström methods [15] to subsample the

training set and operate on a reduced kernel matrix. However the convergence rate of this approximation is slow, $(O(m^{-\frac{1}{4}}))$ [16], where m is the number of samples used.

In computer vision, the exponentiated χ^2 kernel was known to be both the best-performing and the most expensive to compute. A cheaper variant is the histogram intersection kernel [17], for which a computational trick for fast testing is available [18]. However, training time remains a severe problem in this approach since the speedup does not apply. Therefore, many systems directly use linear kernels. Vedaldi et al. proposed a 3-step approach starting with 2 fast linear filtering steps, followed by a non-linear SVM on the exponentiated χ^2 kernel [9]. Bo and Sminchisescu proposed EMK to learn low-dimensional kernel approximations and showed comparable performances with RF for the Gaussian kernel [6].

The work of [12] complements ours, in that it also seeks a low-dimensional linear approximation based on the Fourier theory. However, their development is based on the result of [19], which only applies to scale-invariant kernels in $\mathbb{R}^+$. To adapt to scale, one has to use a kernel that is *additive*, so that the scale of the data $\sqrt{x}$ is multiplied to the kernel on each dimension. Using this approach one could approximate the $1-\chi^2$ and the intersection kernels (Table 1). However, the technique does not immediately extend to the important case of multiplicative kernels. When one has null components in some dimensions, multiplying by $\sqrt{x}$ sets the entire kernel to 0. Although one may palliate such effects e.g., by multiplying with $\exp(-x)$ instead of $\sqrt{x}$, it may be difficult to identify the form of the kernel after such transformations.

6 Experiments

We conduct experiments in a semantic image segmentation task within the PASCAL VOC 2009 Challenge, widely acknowledged as one of the most difficult benchmarks for the problem [20]. In this task, we need to both recognize objects in an image correctly, and generate pixel-wise segmentations for these objects. Ground truth segments of objects paired with their category labels are available for training. A recent approach that achieves state-of-the-art results train a scoring function for each class on many putative figure-ground segmentation hypotheses, obtained using a parametric min-cut method [21]. This creates a large-scale learning task even if the original image database has moderate size: with 90 segments in each image, training on 5000 images creates a learning problem with $450,000$ training examples.

We test a number of kernels on the VOC 2009 dataset. training on the VOC `train` and test on the `validation`, which has approximately 750 images and 1600 objects each. Using the methodology in [21, 22], we select up to 90 putative segments in each image for training and testing. Altogether, there are $62,747$ training segments and $60,522$ test segments. Four types of descriptors are used: a bag of words of dense gray-level SIFT, and three pyramid HOGs, as in [22]. The total number of dimensions is 3270. The final kernel is a weighted sum of

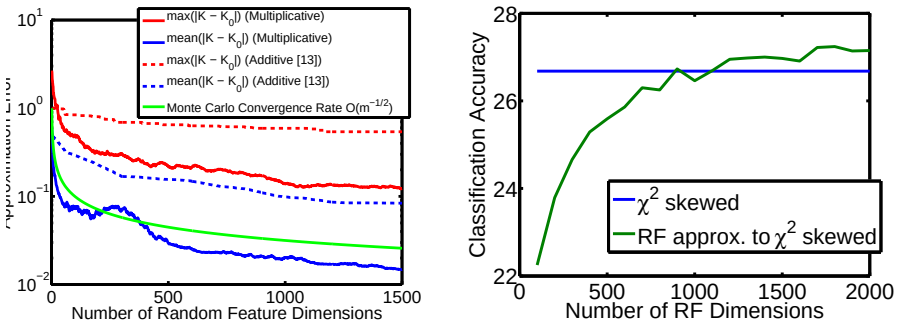

Fig. 1. (*left*) Approximation quality when a linear kernel $K = z_\mu z_\mu^T$ is used to estimate the original kernel matrix K_0. Both the L_∞ error (maximal error) and the average L_1 error are shown. It could be seen that the convergence rate of the multiplicative kernel is consistent with the theoretical $O(m^{-1/2})$ rate for Monte Carlo methods. The rate of the additive kernel is dependent on the input dimension, hence much slower. *(right)* The accuracy as a function of the number of dimensions needed to approximate the kernel.

kernels on each individual descriptor. The kernel parameters are estimated using the approach in [22]. Other parameters, such as c in the skewed kernel and the regularization parameter are chosen by cross-validation on the training set.

In the first experiment we test the quality of the RF approximation for the skewed-χ^2 kernel. Computing the full kernel matrix would require a prohibitive amount of memory. Therefore, testing is done on a 3202×3202 kernel matrix by selecting only the ground truth segment and the best-overlapping putative segment for each object. We plot the result for one HOG descriptor with 1700 dimensions (fig. 1(a)). Notice that the convergence rate of the RF is quite consistent with the theoretical $O(m^{-1/2})$. We also compare with the approximation of the additive χ^2 kernel given in [12]. It can be clearly seen that a skewed multiplicative kernel needs fewer dimensions for good approximation accuracy.

Speed of Training and Testing: Next we compare the speed of the RF approach with a previous nonlinear SVM regression approach [22]. For RF features, we use 2000 dimensions for each type of descriptor, for a total of 8000 dimensions. For RF features on additive kernels, 3 dimensions are used for each input dimension, to make the dimensionality of the RF feature comparable to our multiplicative ones. The results are obtained on an 8-core Pentium Xeon 3.0GHz computer. Since no fast linear SVM regression algorithms are available, we use ridge regression in conjunction with RF features.

Training and testing times for different methods are given in Table 2. One could see that RF offer a substantial speed-up over previous approaches, and is able to scale to much larger datasets[1].

[1] The code for the nonlinear χ^2 kernel is more heavily optimized (using Christoph Lampert's SIMD-optimized χ^2 code) than the skewed kernels, hence Table 2 should not be used to compare speeds among the nonlinear versions of those kernels.

Table 2. Running times (in seconds) for nonlinear and linear approaches. The nonlinear and linear RF histogram intersection [18, 12] has fast testing time, but is slower than the skewed kernels due to higher dimensionality.

Kernel Name	Nonlinear		Linear by Random Fourier Features		
	training	testing	Feature Generation	Training	Testing
Exponentiated χ^2	20647.82	34027.86	N/A	N/A	N/A
Skewed χ^2	70548.20	102277.78	519.22	914.70	57.39
Histogram Intersection	30082.08	742.36	3716.07	1498.05	69.91
Skewed Intersection	53235.17	79821.94	505.37	913.87	56.81

Table 3. Segment classification accuracies (for the best segments in our pool, as determined by ground truth data) for both original non-linear kernels and their approximation using RF.

Kernel	Accuracy		Kernel	Accuracy	
	Nonlinear	Fourier Approx.		Nonlinear	Fourier Approx.
Gaussian	21.31%	24.71%	Exponentiated χ^2	29.54%	N/A
$1-\chi^2$	20.63%	23.75%	Skewed χ^2	26.68%	27.16%
Intersection	22.08%	23.65%	Skewed Intersection	26.34%	26.73%

Results on Different Kernels: Having established that random features offer a substantial speed-up, the question is how good the prediction accuracy of the proposed skewed kernels is. In Table 3 we compare the classification accuracy on all the segments and skip the post-processing step in [22]. Usually this result correlates linearly to the VOC criteria. For the skewed χ^2 kernel, we plot the performance against the number of RF dimensions in fig. 1 (b). One can see that approximations based on random Fourier features can even improve performance of the original kernel. This might be caused by the difference in learning algorithms used (squared loss vs. hinge loss) or the fact that the RF function class is richer than the kernel method: the kernel can be represented by the inner product on RF, but some other functions may also be represented by weighted inner products on RF. Our skewed χ^2 kernel outperforms all the other kernels, but there is still a 2% performance lag with respect to the exponentiated kernel.

7 Conclusion

In this paper, we extend the random Fourier feature methodology to locally compact abelian groups, where kernels on histogram features are considered. Based on empirical observations on the exponentiated χ^2 kernel, we propose a new group on which we build kernels that are not scale-invariant, yet can be approximated linearly using random Fourier features. The experiments show that our kernels are much faster to compute than many nonlinear kernels, and outperform kernels for which approximations are previously known. However, the performance of the proposed kernels is still inferior to that of the exponentiated χ^2 kernel. Designing better kernels to close the gap is an interesting avenue for future work.

Acknowledgments This work was supported, in part, by the European Commission, under a Marie Curie Excellence Grant MCEXT-025481.

References

1. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: Rcv1: A new benchmark collection for text categorization research. JMLR **5** (2004) 361–397
2. Attenberg, J., Dasgupta, A., Langford, J., Smola, A., Weinberger, K.: Feature hashing for large scale multitask learning. In: ICML. (2009)
3. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: A database and web-based tool for image annotation. IJCV **77**(1-3) (2008) 157–173
4. Rahimi, A., Recht, B.: Random features for large-scale kernel machines. In: NIPS. (2007)
5. Shi, Q., Patterson, J., Dror, G., Langford, J., Smola, A., Strehl, A., Vishwanathan, V.: Hash kernels. In: AISTATS. (2009)
6. Bo, L., Sminchisescu, C.: Efficient match kernels between sets of features for visual recognition. In: NIPS. (2009)
7. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. Journal of Machine Learning Research (2008) 1871–1874
8. Shalev-Shwartz, S., Singer, Y., Srebro, N.: Pegasos: Primal estimated sub-gradient solver for svm. In: ICML. (2007)
9. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: ICCV. (2009)
10. Chapelle, O., Haffner, P., Vapnik, V.: Support vector machines for histogram-based image classification. IEEE Transactions on Neural Networks **10** (1999)
11. Rudin, W.: Fourier Analysis on Groups. (1962)
12. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. In: CVPR. (2010)
13. Fine, S., Scheinberg, K.: Efficient svm training using low-rank kernel representation. JMLR **2** (2001) 243–264
14. Bach, F., Jordan, M.I.: Predictive low-rank decomposition for kernel methods. In: ICML. (2005)
15. Williams, C.K.I., Seeger, M.: Using the nyström method to speed up kernel machines. In: NIPS. (2001)
16. Drineas, P., Mahoney, M.: On the nyström method for approximating a gram matrix for improved kernel-based learning. JMLR **6** (2005) 2153–2175
17. Grauman, K., Darrell, T.: The pyramid match kernel: Efficient learning with sets of features. JMLR **8** (2007) 725–760
18. Maji, S., Berg, A.C., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: CVPR. (2008)
19. Hein, M., Bousquet, O.: Hilbertian metrics and positive definite kernels on probability measures. In: AISTATS. (2005)
20. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results. http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html
21. Carreira, J., Sminchisescu, C.: Constrained parametric min cuts for automatic object segmentation. In: CVPR. (2010)
22. Li, F., Carreira, J., Sminchisescu, C.: Object recognition as ranking holistic figure-ground hypotheses. In: CVPR. (2010)