Quadratic polynomial kernel approximation with asymmetric embeddings

Remi Delogne and Laurent Jacques
INMA/ICTEAM, UCLouvain
{remi.delogne,laurent.jacques}@uclouvain.be

Abstract: Random embedding techniques, such as random Fourier features, are widely used to sketch initial data to a new, kernelised feature space. In this work, we leverage a specific property of random rank-one projection operators, the sign product embedding, to approximate a quadratic polynomial kernel using the scalar product of a pair asymmetric vector embeddings, with one taking only binary values. We demonstrate empirically that the approximated kernel compares favourably to the initial one on toy binary classification examples.

Keywords: Signal Processing, Kernel methods, Random Features

1 Introduction

In the realm of signal processing and machine learning, many applications require *sketching techniques*. These are employed to provide a new representation or approximation of data vectors while preserving their essential characteristics. Sketching methods prove particularly useful when dealing with large datasets or large objects that can make their processing too costly to perform, especially in a context with limited computational resources. While traditional sketching methods (such as linear random projections in compressive sensing) aim to reduce the number of dimensions, one can also *lift* data vectors into a higher dimensional space where tasks such as classification may be easier to perform.

We here consider the particular case of *quadratic* sketching, a variant of the classical linear version often used in compressive sensing. This particular sketching method relies on a *sketching operator* that uses rank-one projections (ROP) and sends a vector into a new *sketched* domain. In [1] and [2] we showed that estimating specific functions of a given signal (such as localized feature estimation) was possible having only access to the sketched signal, without reconstructing the original signal with costly optimisation methods.

In this paper, we show that the estimation process can in fact approximate a kernel function with corresponding asymmetric feature maps. This allows us to use kernels to classify data embedded in different ways.

2 Sketching and Sign Product Embedding (SPE)

In this section we briefly describe the sketching procedure considered. The studied sketching operator relies on rank-one projections (ROP) [3].

Given a vector $\boldsymbol{x} \in \mathbb{R}^n$ a single ROP measurement is defined as $(\boldsymbol{a}_i^\top \boldsymbol{x})^2$, for some random[1] vector $\boldsymbol{a}_i \in \mathbb{R}^n$. The quadratic operator is defined as a vector containing m ROP measurements of the vector to be observed: $\mathcal{A}^{\mathrm{v}} : \boldsymbol{x} \in \mathbb{R}^n \mapsto \mathcal{A}^{\mathrm{v}}(\boldsymbol{x}) := \big((\boldsymbol{a}_i^\top \boldsymbol{x})^2\big)_{i=1}^m \in \mathbb{R}_+^m$.

The quadratic nature of the ROP operator $\mathcal{A}^{\mathrm{v}}$ makes it linear with respect to the *lifted* signal $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$. We can indeed see that $\mathcal{A}^{\mathrm{v}}(\boldsymbol{x}) = \big(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle\big)_{i=1}^m$, where $\boldsymbol{A}_i = \boldsymbol{a}_i\boldsymbol{a}_i^\top$, and where the scalar product is the Frobenius inner product. As suggested in [4] we consider here a slightly different version of $\mathcal{A}^{\mathrm{v}}$—this operator presenting a certain ℓ_2-norm *bias* since $\mathbb{E}\|\mathcal{A}^{\mathrm{v}}(\boldsymbol{x})\|^2 \neq \|\boldsymbol{x}\|^4$. We thus instead consider this operator

$$\mathcal{B} : \boldsymbol{x} \in \mathbb{R}^n \mapsto \mathcal{B}(\boldsymbol{x}) = \big(\mathcal{A}^{\mathrm{v}}_{2i}(\boldsymbol{x}) - \mathcal{A}^{\mathrm{v}}_{2i+1}(\boldsymbol{x})\big)_{i=1}^m. \quad (1)$$

Interestingly, $\mathcal{B}$ satisfies this property [1, 2]: given a unit vector $\boldsymbol{y} \in \mathbb{R}^n$, $\kappa = \pi/4$, and a distortion $0 < \delta < 1$, provided that $m \geq C\delta^{-2}k\log(\frac{n}{k\delta})$, then, with probability exceeding $1 - C\exp(-c\delta^2 m)$, we have for all k-sparse signals $\boldsymbol{x} \in \Sigma_k := \{\boldsymbol{v} \in \mathbb{R}^n : |\mathrm{supp}(\boldsymbol{v})| \leq k\}$,

$$\Big|\tfrac{\kappa}{m}\langle \mathcal{B}(\boldsymbol{x}), \mathrm{sign}(\mathcal{B}(\boldsymbol{y}))\rangle - \langle \boldsymbol{x}, \boldsymbol{y}\rangle^2\Big| \leq \delta\|\boldsymbol{x}\|^2, \quad (2)$$

where sign is applied componentwise on vectors. We call (2) the *Sign Product Embedding* (SPE). In other words, with m, the number of quadratic measurements sufficiently large, one can approximate $\langle \boldsymbol{x}, \boldsymbol{y}\rangle^2$ up to a controlled distortion.

3 Kernel Representation

The trained eye of a machine learner cannot help but notice the strong resemblance between the left-hand-

[1] Here we restrict to Gaussian random vectors $\boldsymbol{a}_i \sim \mathcal{N}(0, 1)$.

side of equation (2) and some type of kernel approximation. Indeed recall that kernel methods seek to replicate an inner product of a high (possible infinitely) dimensional Hilbert space $\mathcal{H}$ with a kernel function, via the kernel trick. Instances of a dataset in their original space (say $\mathbb{R}^n$) are mapped to the high dimensional space via an embedding $\varphi : \mathbb{R}^n \to \mathcal{H}$ such that for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, $k(\boldsymbol{x}, \boldsymbol{y}) = \langle \varphi(\boldsymbol{x}), \varphi(\boldsymbol{y}) \rangle_{\mathcal{H}}$. With mapped instances now belonging to $\mathcal{H}$, non-linear classifiers in $\mathbb{R}^n$ can be reached with linear ones in $\mathcal{H}$. In practice such embeddings are never explicitly constructed thanks to the kernel trick, however one can resort to *random features* to get an explicit approximation of $\mathcal{H}$ [5]. This yields

$$k(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{H}} \approx z(\boldsymbol{x})^\top z(\boldsymbol{y}), \tag{3}$$

where $z : \mathbb{R}^n \to \mathbb{R}^m$ is a random feature operator. This approximation can lower the computational complexity while inferring the classes of new data [5].

Going back to equation (2), we can set $k(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2$, and build a corresponding random feature embedding similar to the one described in equation (3). The specifics of equation (2) do however introduce a little twist to the general framework as we need two *distinct* random embeddings: $\phi : \mathbb{R}^n \to \mathbb{R}^m : \boldsymbol{x} \mapsto \mathcal{B}(\boldsymbol{x})$ and $\psi : \mathbb{R}^n \to \mathbb{R}^m : \boldsymbol{x} \mapsto \text{sign}(\mathcal{B}(\boldsymbol{x}))$. This follows the methodology used in [6] where the scalar product of asymmetric random periodic embeddings (incorporating quantisation for instance) yields approximate kernel functions. We propose to show experimentally that we can use these two embeddings ϕ and ψ for classification purposes.

4 Experiments

Let us first of all test the kernel approximation directly. We started by creating an artificial binary dataset of 10.000 points in $\mathbb{R}^{100}$ labelled 0 or 1, with only two informative features. We then trained a classifier using $k(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2$. This first classifier averaged an 87% accuracy over 5-fold cross-validation. We sketched the data using $\mathcal{B}$ and trained three different classifiers. The first approximated with $\langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle$, the second with $\langle \phi(\boldsymbol{x}), \psi(\boldsymbol{y}) \rangle$ and the third with $\langle \psi(\boldsymbol{x}), \psi(\boldsymbol{y}) \rangle$. The results are displayed in figure 1. We can see that the first (non-binarised) kernel unsurprisingly performs best (on average 4% better than the second) and that the second (asymmetric) kernel is on par with the third (fully binarised) kernel (on average 0.4% better).

For the second experiment, recall that given a classifier, a new instance $\boldsymbol{x}'$ is classified using the quantity $\sum_i y_i \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}') + b$, with y_i the class of $\boldsymbol{x}_i$ and α_i and b determined by the training. Using the asymmetric approximation of the kernel function we can train a classifier using sketched training data and evaluate a new instance binarised. This situation arises in setups where a sensor transmits observations to a server where

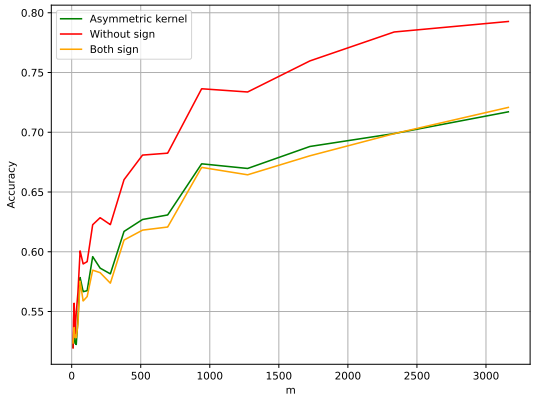

Figure 1: Accuracy of classifiers as a function of m

they will be classified on a pre-trained model, but the transmission of data must remain low and the sensor thus resorts to binarising its data. In a similar way if the server has memory requirements it may operate on binarised data and classify non-binarised new instances. We tested both these setups with the same artificial data as in the previous experiment. Using $m = 5000$ we obtained an accuracy of 74.2% and 79.5% respectively. Notice that thanks to the binarisation step, the sketched vectors with $m = 5000$ only require 5000 bits of memory.

5 Conclusion and further research

We empirically showed that the result given in [1] allows us to approximate an asymmetric kernel that is capable of performing classification. This result remains strictly empirical and needs some theoretical improvements to quantify more precisely the link between m and the approximation error of the kernel.

Furthermore the classification step described in the experiments can actually be seen as acting on labelled rank-one matrices due to the nature of ϕ. Since ϕ is linear with respect to rank-one matrices, this could potentially reduce the complexity of training algorithms on datasets of rank-one matrices.

References

[1] R. Delogne, V. Schellekens, L. Daudet L. Jacques, "Signal Processing with Optical Quadratic Random Sketches", *ICASSP*, 2023.

[2] R. Delogne, V. Shellekens, and L. Jacques, "ROP inception: signal estimation with quadratic random sketching", *ESANN*, 2022.

[3] T. T. Cai and A. Zhang, "Rop: Matrix recovery via rank-one projections," *The Annals of Statistics*, vol. 43, no. 1, pp. 102–138, 2015.

[4] Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and stable covariance estimation from quadratic sampling via convex programming", *IEEE Transactions on Information Theory*, 2015.

[5] A. Rahimi, B. Recht "Random features for large-scale kernel machines", *Advances in Neural Information Processing Systems*, 1177–1184, 2007.

[6] V. Schellekens, L. Jacques, "Breaking the waves: asymmetric random periodic features for low-bitrate kernel machines", *Information and Inference: A Journal of the IMA*, 2022.