RANDOM PROJECTIONS THROUGH MULTIPLE OPTICAL SCATTERING: APPROXIMATING KERNELS AT THE SPEED OF LIGHT

A. Saade[1]*, F. Caltagirone*[1]*, I. Carron* [2]*, L. Daudet* [2,3,7]*, A. Drémeau* [4]*, S. Gigan* [5,6,7]*, F. Krzakala*[1,6,7]

[1] Laboratoire de Physique Statistique, CNRS UMR 8550 & École Normale Supérieure, Paris, France.
[2]Institut Langevin, ESPCI and CNRS UMR 7587, Paris, F-75005, France
[3]Paris Diderot University, Sorbonne Paris Cité, Paris, F-75013, France
[4] ENSTA Bretagne and Lab-STICC UMR 6285, F-29806 Brest, France
[5]Laboratoire Kastler Brossel, CNRS UMR 8552 & École Normale Supérieure,75005 Paris, France.
[6]Sorbonne Universités, Université Pierre et Marie Curie Paris 06, F-75005, Paris, France
[7] PSL Research University, 75005 Paris, France

ABSTRACT

Random projections have proven extremely useful in many signal processing and machine learning applications. However, they often require either to store a very large random matrix, or to use a different, structured matrix to reduce the computational and memory costs. Here, we overcome this difficulty by proposing an analog, optical device, that performs the random projections *literally* at the speed of light without having to store any matrix in memory. This is achieved using the physical properties of multiple coherent scattering of coherent light in random media. We use this device on a simple task of classification with a kernel machine, and we show that, on the MNIST database, the experimental results closely match the theoretical performance of the corresponding kernel. This framework can help make kernel methods practical for applications that have large training sets and/or require real-time prediction. We discuss possible extensions of the method in terms of a class of kernels, speed, memory consumption and different problems.

Index Terms— Kernel methods, random projections, machine learning, large-scale data processing, optical computing

1. INTRODUCTION

Random projections have proven useful in signal processing and machine learning in several ways. A first line of applications is concerned with dimensionality reduction [1, 2, 3, 4]. Given a dataset with a very large number of features, we look for a simple transformation of the data that reduces the number of features while approximately preserving the pairwise distances between data points. It turns out that linear random projection are suitable to this purpose (Johnson-Lindenstrauss lemma [5]). Similarly, a small number of random projections of sparse signals can carry sufficient information for their reconstruction, as shown in compressed sensing [6, 7].

Conversely, non-linear random projections allow the embedding of a dataset in a larger dimensional feature space. Notably, the data may become linearly separable in this larger space, while it was not in the native feature space. A linear model can then be used to fit or classify the data. An interesting class of methods relying on this embedding is that of so-called kernel machines [8], among which is the celebrated SVM. These offer the advantage, sometimes called *kernel trick*, of not requiring the explicit mapping, but only the inner products of all pairs of embedded data points, i.e. a function $k(x, y)$ where k is called the kernel, and x, y are data points in the native feature space. While removing the dependency on the dimension of the embedding (which is possibly infinite), this trick relies on a matrix storing all the values of k for all pairs of data points. This typically does not scale to large datasets.

Recently [9], a cure has been proposed to this problem that actually makes explicit a non-linear mapping of the data points such that the inner product of two transformed data points approximately equals their kernel evaluation. In this new feature space, the regression or classification task can be solved by a linear machine, which can be trained very quickly compared to non-linear ones [10]. Random projections have thus made some large-scale machine learning problems tractable, to the point that actually computing them - i.e. generating and storing a large random matrix, and computing matrix-vector products - has become one of the major bottlenecks in the above mentioned approaches [11, 12, 13].

In this study, we demonstrate experimentally that an optical-based hardware can be built which instantaneously provides a large number of "ideal" random projections of any input data. We show that this physical process mimics, for the goal of classification, the computation of a well-defined elliptic kernel. In other words, we believe that this optical setup can be seen as a generic digital data pre-processor, for many subsequent uses of the data. It has to be empha-

sized that, contrary to previous studies on *imaging* techniques based on multiple scattering (with a similar experimental setup [14, 15]), this random transformation does not have to be precisely determined through a lengthy calibration stage : the knowledge of its statistical properties is enough to guarantee its effectiveness. Although this process is fundamentally analog, we show here that the associated experimental noise does not substantially change the results, which are in very good agreement with computer simulations.

2. RANDOM PROJECTIONS AND KERNEL MACHINE LEARNING

Let us start by introducing the standard ridge regression problem, arguably the simplest kernel method for supervised classification. Throughout the section, we are given a training set composed of labeled data, represented by a matrix $\mathbf{U} \in \mathbb{R}^{n \times p}$ where n is the number of samples and p is the dimension of the data (number of features). Each data point $\mathbf{U}_i \in \mathbb{R}^p$ has a unique label $l_i \in [\![1, q]\!]$ which we encode in a matrix $\mathbf{Y} \in \mathbb{R}^{n \times q}$ with elements $\mathbf{Y}_{i,j} = \delta_{j,l_i}$. We are also given a test set of unlabeled data, represented by a matrix $\tilde{\mathbf{U}} \in \mathbb{R}^{\tilde{n} \times p}$ where $\tilde{n}$ is the number of unlabeled samples. Our goal is to estimate their label $\tilde{\mathbf{Y}} \in \mathbb{R}^{\tilde{n} \times q}$ using a classifier trained on the training set. Recall that the ridge regression problem reads

$$\underset{\beta \in \mathbb{R}^{p \times q}}{\operatorname{argmin}} ||\mathbf{U}\beta - \mathbf{Y}||_2^2 + \gamma ||\beta||_2^2 \tag{1}$$

where γ controls the trade-off between the estimation error and the regularization. Its closed-form solution is given by

$$\beta = (\mathbf{U}^T\mathbf{U} + \gamma \mathbf{I}_p)^{-1}\mathbf{U}^T\mathbf{Y} = \mathbf{U}^T(\mathbf{U}\mathbf{U}^T + \gamma \mathbf{I}_n)^{-1}\mathbf{Y} \tag{2}$$

where we have used a classical algebraic identity. Our prediction for the labels of the test data points is then

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{U}}\beta = \tilde{\mathbf{U}}(\mathbf{U}^T\mathbf{U} + \gamma \mathbf{I}_p)^{-1}\mathbf{U}^T\mathbf{Y} \tag{3}$$

$$= \tilde{\mathbf{U}}\mathbf{U}^T(\mathbf{U}\mathbf{U}^T + \gamma \mathbf{I}_n)^{-1}\mathbf{Y} \tag{4}$$

More precisely, since we are interested in classification, we will set the label of $\tilde{\mathbf{U}}_i$ to $\tilde{l}_i = \max_{j \in [\![1,q]\!]} \tilde{\mathbf{Y}}_{ij}$. Note that Eq. (4) only depends on inner products of the data points.

2.1. Kernel classification

Given a kernel k, we define the $n \times n$ (resp. $\tilde{n} \times n$) kernel matrix $\mathbf{K}$ (resp. $\tilde{\mathbf{K}}$) with elements

$$\mathbf{K}_{i,j} = k(\mathbf{U}_i, \mathbf{U}_j) \text{ and } \tilde{\mathbf{K}}_{i,j} = k(\tilde{\mathbf{U}}_i, \mathbf{U}_j) \tag{5}$$

In the kernel ridge regression problem, we simply replace the Euclidean inner product in feature space by this kernel. Because Eq. (4) only depends on these inner products, we do not have to explicit the mapping, and the solution of the kernel ridge regression is directly given by

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{K}}(\mathbf{K} + \gamma \mathbf{I}_n)^{-1}\mathbf{Y} \tag{6}$$

We consider here the following elliptic kernel:

$$k(\mathbf{U}_i, \mathbf{U}_j) = \frac{\sqrt{\mathbf{U}_i^T\mathbf{U}_i\,\mathbf{U}_j^T\mathbf{U}_j}}{2}\Bigg\{ -(\sin^2\theta)\mathcal{E}_K\left[\cos^2\theta\right] + 2\mathcal{E}_E\left[\cos^2\theta\right] + |\sin\theta|\left(2\mathcal{E}_E\left[-\frac{\cos^2\theta}{\sin^2\theta}\right] - \mathcal{E}_K\left[-\frac{\cos^2\theta}{\sin^2\theta}\right]\right)\Bigg\} \tag{7}$$

where $\mathcal{E}_K[.]$ and $\mathcal{E}_E[.]$ are the complete elliptic integrals of the first and second kind respectively, and θ the angle between $\mathbf{U}_i$ and $\mathbf{U}_j$. Despite its apparent complexity, this function essentially looks like a bell curve as a function of θ, and therefore quantifies the similarity between $\mathbf{U}_i$ and $\mathbf{U}_j$. This particular choice will be motivated experimentally in section 2.3.

To illustrate this method, we consider the MNIST dataset of handwritten digits [16]. This dataset is split into a training set of 60000 digits of size 28×28, and a test set of 10000 digits. Using the elliptic kernel, we achieve a classification error of 1.31%, to be compared with a 12% error for the purely linear ridge regression (4). While better results can be achieved using much more complex deep neural networks, kernel ridge regression achieves a reasonable error and is remarkable in its simplicity. The major drawback of this approach is the computational cost of inverting an $n \times n$ matrix. In "big data" problems, where n can be in the billions, it would not even be possible to store such a matrix.

2.2. Random projections

Following [9], we now compute a non-linear mapping and solve the linear ridge regression problem (1) in this new feature space of dimension N. Specifically, we consider mappings of the form

$$\mathbf{X}_{i,j} = \phi((\mathbf{W}\mathbf{U}_i)_j + \mathbf{b}_j) \qquad i \in [\![1, n]\!], j \in [\![1, N]\!] \tag{8}$$

where $\mathbf{b}$ is a bias, ϕ is a non-linear function and $\mathbf{W} \in \mathbb{R}^{N \times p}$ is a linear projection which we take to be random. The solution of the ridge regression problem is

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \gamma \mathbf{I}_n)^{-1}\mathbf{Y} = \tilde{\mathbf{X}}(\mathbf{X}^T\mathbf{X} + \gamma \mathbf{I}_N)^{-1}\mathbf{X}^T\mathbf{Y} \tag{9}$$

In this last formulation, we see that if the inner products of the data points in the new feature space approximate a kernel, and if this new feature space has a reasonable dimension N, then we can approximate a kernel ridge regression by inverting a smaller matrix ($\mathbf{X}^T\mathbf{X} \in \mathbb{R}^{N \times N}$) instead of the kernel matrix ($\mathbf{X}\mathbf{X}^T \in \mathbb{R}^{n \times n}$). Concretely, the experimental device described in the next section can perform (8) with $\mathbf{W}$ a random i.i.d complex matrix with Gaussian real and imaginary parts, ϕ the modulus function, and a vanishing bias $\mathbf{b}$. In synthetic experiments, if we use $N = 10000$ random projections and the linear ridge regression (9), we get an error of approximately 2% (see Fig. 1), having only to invert a 10000^2 matrix instead of a 60000^2 one. Even better: now we have lost the dependency on n^2.

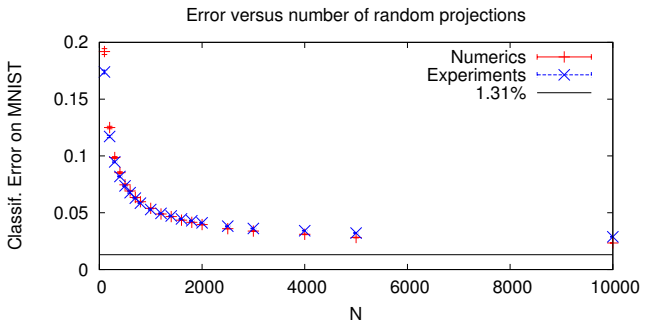

Fig. 1: Comparison between synthetic random projections and the experimental ones. The classification error is shown as a function of the dimension N of the feature space. As N increases, the error using synthetic random projections (red) approaches the asymptotic value using the true elliptic kernel (here 1.31%). Experimental results, made with our optical device, closely follow the synthetic ones though they tend to deviate from the synthetic data for larger N (see discussion).

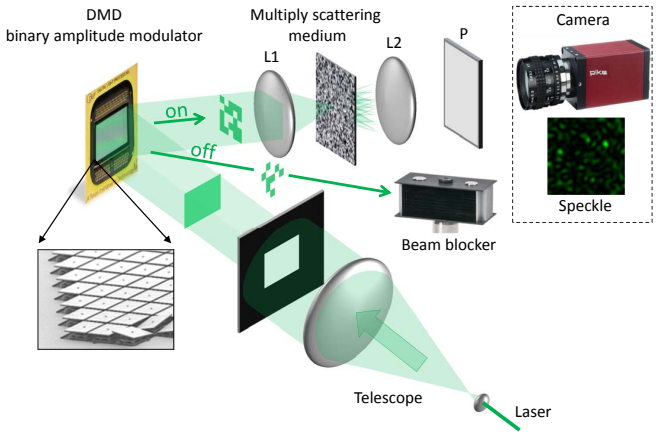

Fig. 2: Experimental setup (from [15]). A monochromatic laser at 532nm is expanded by a telescope, then illuminates a digital micromirror device (DMD), able to spatially encode digital information on the light beam by amplitude modulation, as described in section 3.2. The light beam carrying the signal is then focused on a random medium by means of a lens. Here, the medium is a thick (several tens of microns) layer of TiO_2 (Titanium Dioxide) nanoparticles (white paint pigments deposited on a microscope glass slide). The transmitted light is collected on the far side by a second lens, passes through a polarizer, and is measured by a standard monochrome CCD camera.

2.3. Random complex projections and the elliptic kernel

In the limit where the number of projections N grows to infinity, due to the concentration of the measure, it is possible to show that the inner product between the projected data points $\mathbf{X}_i \in \mathbb{R}^N$ tends to a kernel function that depends only on the $\mathbf{U}_i$ [17]. With the choice of $\mathbf{W}$ and ϕ corresponding to our experiment, a tedious but simple computation shows that the limiting kernel is precisely the elliptic kernel of equation (7). To approximate this kernel, we can therefore use our experimental device to perform the non-linear projections, and solve the ridge regression problem (9). This concentration phenomenon is shown on Fig. 1 where we plot the classification error on the MNIST database when the number of random projections is increased (red points). As the number of random projections grows, the classification error using synthetic random projections approaches the asymptotic value using the true elliptic kernel (1.31%). Empirically, we find that these corrections are very well fitted by a power law in $N^{-2/3}$.

3. EXPERIMENTAL APPARATUS

The approach exposed in the previous section still requires to store, and multiply by, a potentially huge random matrix, and to apply the modulus function. We now move to the discussion of our experimental apparatus which will make this procedure a trivial, instantaneous one.

3.1. Principle of optical analog random projections

At the basis of the analog random projections, we exploit the ability of a heterogeneous material, such as paper, paint, or any white translucid material, to scatter light impinging on them in a very complex way. Due to their extremely high complexity, their behavior for light scattering is considered "random", and this kind of medium are often called "random medium". This term is abusive since scattering is a linear, deterministic, and reproducible phenomenon. However, the unpredictable nature of the process makes it effectively a random process. This is why these materials are called "opaque", since all information on the incoming light is seemingly lost (but only irremediably mixed) during the propagation. As an example, consider a cube of edge size 100 μm. It comprises $\approx 10^7$ paint nanoparticles, whose positions and shape would have to be exactly known in order to predict its effect on light. Propagation through such a layer can be seen as a random walk because of frequent scattering with the nanoparticles. The characteristic step is the transport mean free path, typically of a few μm, so light would explore the whole volume and endure on average tens of thousands of such steps before exiting on the other side, in a few picoseconds.

When light is coherent, it gives rise to interferences. The complex interference pattern arising from the multiple scattering process is called "speckle". It is characteristic not only of the heterogeneous material, but also of the "shape" of the input light. In essence, the propagation of light through a random medium can be modeled as $y = \mathbf{H}x$ where y (resp. x)

are vector amplitudes between a set of spatial modes at the output (resp. input) of the medium, and where $\mathbf{H}$ is the so-called transmission matrix of the medium, which has been shown to be very close to Gaussian i.i.d matrices [18]. By "interrogating" the medium with an appropriate set of input illuminations, which can be conveniently done using a spatial light modulator and a laser, and measuring the resulting "answer" by means of a CCD or CMOS camera for instance, one would therefore record the resulting intensity $|y|^2$ (see Fig. 2). For a stable medium, such as a paint layer for instance, the transmission matrix $\mathbf{H}$ is stable and the medium can therefore provide a convenient platform for random projections, without the need to determine $\mathbf{H}$.

3.2. Encoding the data on the DMD

The DMD we use is a 1920×1080 array of micro-mirrors. Each micro-mirror encodes a binary value (lit or not). Levels of grey can be conventionally coded in time by lighting up a micro-mirror for a fraction of the exposure time of the camera. This approach however reduces the frequency at which data can be processed through the apparatus. We therefore resorted to spatial encoding. We encoded each pixel of an image on a 4×4 array of micro-mirrors in which the number of lit micro-mirrors reflects the level of grey of the pixel. This allowed us to encode 17 grey levels.

Each digit in the MNIST dataset can be seen as a 28^2 array of integers between 0 and 255. We first quantized the grey levels between 0 and 16. We then encoded each of these quantized pixels as a 4×4 array of micro-mirrors. This results in a 112^2 array of binary variables. To rescale the resulting image to 1920×1080, which is the size of the DMD, we need to first add a border of zeros, so that the image becomes of size 120^2. Finally, we can rescale the image by a factor of 16×9. The result is a 1920×1080 binary image, encoding each digit over 17 levels of grey. This image can then be projected on the DMD. It is in principle possible to encode more levels of grey by encoding each pixel on a larger array of micro-mirrors. The procedure described to encode the data on the DMD is not specific to images and generalizes to any kind of data which can be written as a numerical vector.

After sending the data through the disordered medium, we acquire a snapshot of the resulting random projection using a standard camera. To reduce the correlations between neighboring pixels of the output, we start by acquiring a 400^2 pixel area, which we then reduce to 100^2 pixels where each of these pixels is the average of a 4^2 patch in the original snapshot.

4. RESULTS

Our results are summarized in Fig. 1 where we compare the efficiency in classification using synthetic random projection (in red) and the random projection obtained within the experimental set-up (in blue). We find that the two curves agree remarkably well, thus validating our procedure to generate analog random projections for classification tasks.

We note, however, that when the number of projections is increased to large values, i.e. $N \gtrsim 2000$, the optical experiment starts to deviate from the synthetic one. We checked numerically that adding even large noise to the synthetic projection did not change the results of the classification, which is therefore robust to experimental noise. We thus believe that this deviation is due to residual correlations between pixels of the output image. The number of independent pixels can be tuned in several ways, for instance by ensuring that the CCD pixel on the output camera matches the size of a speckle grain, or by tuning the thickness of the random medium itself to increase the number of independent modes [18]. We are therefore confident that this deviation can be decreased with further development of the experimental apparatus.

5. CONCLUSION AND PERSPECTIVES

This study shows that large dimensional (here 10^4 - 10^6) random projections of digital data can be obtained almost instantaneously with an optical device, that can then be used for practical machine learning applications such as (but not limited to) kernel classification. The speed of "computation" is only limited by how fast one can modulate light at the input (off-the-shelf DMDs can handle images with millions of pixels, at rates of 20 kHz or above), and measure the corresponding scattered light. Note also that although optical sensors natively measure only the intensity (squared modulus) of the field, and not the complex-valued field itself, this is precisely what is needed for the kernel computations presented here, akin to the non-linear transform of each layer of a neural network. In the cases where the linear projections are needed, e.g. for randomized linear algebra, the experimental setup can be modified to include an interferometric arm. We can then apply any non-linear function using e.g. FPGAs, widening the range of kernels we can approximate. It is also possible to combine several DMDs together to process larger signals.

More generally, we believe that this experiment is a first proof-of-concept toward a new generation of analog optical-based co-processors, able to complement existing silicon-based chips for the processing of very large datasets, with potential benefits in terms of data throughput and energy consumption.

Acknowledgments

We thank Gilles Wainrib for illuminating discussions. This research has received funding from the European Research Council under the EU's 7^{th} Framework Programme (FP/2007-2013/ERC Grant Agreement 307087-SPARCS and 278025-COMEDIA) ; and from LABEX WIFI under references ANR-10-LABX-24 and ANR-10-IDEX-0001-02-PSL*.

6. REFERENCES

[1] Ella Bingham and Heikki Mannila, "Random projection in dimensionality reduction: applications to image and text data," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 245–250.

[2] Dmitriy Fradkin and David Madigan, "Experiments with random projections for machine learning," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 517–522.

[3] Dimitris Achlioptas, "Database-friendly random projections: Johnson-lindenstrauss with binary coins," *Journal of computer and System Sciences*, vol. 66, no. 4, pp. 671–687, 2003.

[4] Chinmay Hegde, Michael Wakin, and Richard Baraniuk, "Random projections for manifold learning," in *Advances in neural information processing systems*, 2008, pp. 641–648.

[5] William B Johnson and Joram Lindenstrauss, "Extensions of lipschitz mappings into a hilbert space," *Contemporary mathematics*, vol. 26, no. 189-206, pp. 1, 1984.

[6] Emmanuel J Candes, Justin K Romberg, and Terence Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on pure and applied mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[7] David L Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, 2006.

[8] John Shawe-Taylor and Nello Cristianini, *Kernel methods for pattern analysis*, Cambridge university press, 2004.

[9] Ali Rahimi and Benjamin Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2007, pp. 1177–1184.

[10] Thorsten Joachims, "Training linear svms in linear time," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 217–226.

[11] Nir Ailon and Edo Liberty, "Fast dimension reduction using rademacher series on dual bch codes," *Discrete & Computational Geometry*, vol. 42, no. 4, pp. 615–630, 2009.

[12] Quoc Le, Tamás Sarlós, and Alex Smola, "Fastfood-approximating kernel expansions in loglinear time," in *Proceedings of the international conference on machine learning*, 2013.

[13] Xiangrui Meng and Michael W Mahoney, "Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression," in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM, 2013, pp. 91–100.

[14] Antoine Liutkus, David Martina, Sébastien Popoff, Gilles Chardon, Ori Katz, Geoffroy Lerosey, Sylvain Gigan, Laurent Daudet, and Igor Carron, "Imaging with nature: Compressive imaging using a multiply scattering medium," *Scientific reports*, vol. 4, 2014.

[15] Angélique Drémeau, Antoine Liutkus, David Martina, Ori Katz, Christophe Schülke, Florent Krzakala, Sylvain Gigan, and Laurent Daudet, "Reference-less measurement of the transmission matrix of a highly scattering material using a dmd and phase retrieval techniques," *Optics express*, vol. 23, no. 9, pp. 11898–11911, 2015.

[16] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[17] Christopher KI Williams, "Computation with infinite neural networks," *Neural Computation*, vol. 10, no. 5, pp. 1203–1216, 1998.

[18] SM Popoff, G Lerosey, R Carminati, M Fink, AC Boccara, and S Gigan, "Measuring the transmission matrix in optics: an approach to the study and control of light propagation in disordered media," *Physical review letters*, vol. 104, no. 10, pp. 100601, 2010.