Use Online Dictionary Learning to Get Parts-based Decomposion of Noisy Data

Daming Lu
Baidu Research
Sunnyvale, CA USA
ludaming@baidu.com

***Abstract*—Huge amount of data are generated every day. Extracting interpretable features from the data is becoming important. Meanwhile, dimension reduction and low rank approximation are also becoming important as people want to factorize big matrix into smaller ones, which are easier to handle. Sparse coding is such a technique that can factorize matrix into sparse linear combinations of basis elements. We found that through online dictionary learning, an efficient sparse coding algorithm, we could decompose large data matrix with noise into interpretable dictionary atoms. Such atoms are useful in reconstructing a denoised data matrix.**

Keywords—machine learning, sparse coding, online dictionary learning, dimension reduction

I. Introduction

Large amount of high dimensional data are generated every day, thanks to the prosperity of the Internet and big data technology. Due to the difficulty of processing high dimensional data, people intend to factorize or decompose large data matrices into smaller ones. The linear decomposition of a matrix into a few basis elements has been a hot research spot for a long time. At first, general purposed basis matrices were used to represent the large matrix. Later, using ad hoc matrix learned from specific input data produces better results. However, although many such decomposition methods could produce smaller matrices, these matrices are hardly interpretable, especially when the input data are noisy. Such popular methods include PCA, CUR, etc. We found that online dictionary learning (ODL) methods, introduced in [], could not only reduce the matrix dimension, but the atoms in the learned dictionary are interpretable as well. We applied this technology to two different set of noisy data and found the extracted atoms very close to the ground truth.

We compared our method with UoI NMF cluster and found our accuracies are at the same level. Meanwhile, online dictionary learning runs faster. The following sections are organized as below:

- We first review the core part of ODL
- We then introduce the application of ODL on our datasets
- We finally compared our results with NMF and discussed potential advantages and disadvantages.

At last, we discussed future improvements.

II. Preliminaries

A. Online Dictionary Learning

Online dictionary learning was first introduced in []. Assume we have a finite training dataset as $X = [x_1, \dots x_n]$ in $\mathbb{R}^{m \times n}$, we want to learn a dictionary $\mathbf{D}$ as a "good" representation of signal x. Normally the dimension m is relatively small compared to the total amount of data n. We want to have a $k \ll n$ such as we can only use a few elements (atoms) in $\mathbf{D}$ to represent signal x. Our aim is to optimize $\ell(x, \mathbf{D})$ as the ℓ_1 sparse coding problem:

$$\ell(\mathrm{x}, \mathbf{D}) \triangleq \min_{\alpha \in \mathrm{R}^k} \frac{1}{2} \|\mathrm{x} - \mathbf{D}\alpha\|_2^2 + \lambda \|\alpha\|_1$$

where λ is a regularization parameter. We followed the algorithm as mentioned in [] with a mini-batch extension in order to prevent polluting the initialization. [UoI] used a clustering process to extract true bases from noise, which is also a good idea.

III. Numeric experiments

In this section, we illustrate the application of online dictionary learning on two datasets, Swimmer dataset and MNIST 2-digit dataset. The details of these two datasets were introduced in [UoI]. We used SPAMS library as our backbone. After 1000 iterations, the performance as well as learned atoms are shown in Fig 1, side by side with UoI.

We can tell from the metrics that online dictionary learning is on the same the accuracy level as UoI. Moreover, since online dictionary learning does not have the clustering part as in UoI, it ran faster.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do

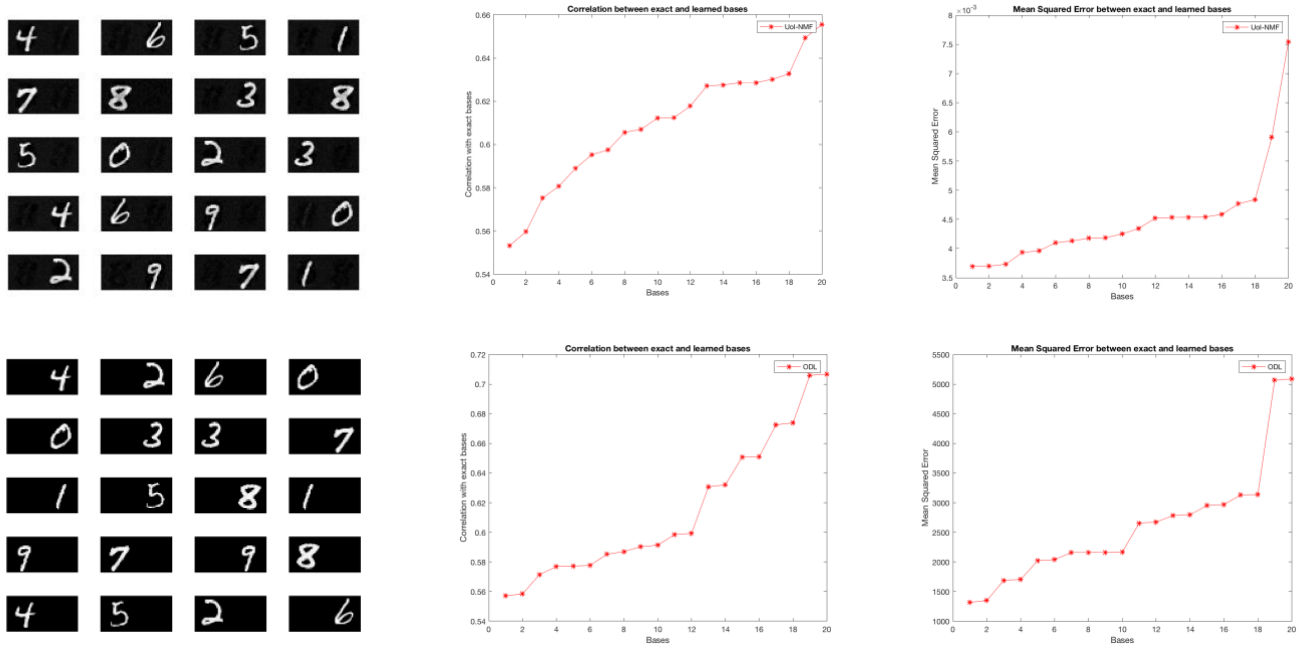

Figure 1: UoI-NMFcluster and Online Dictionary Learning for noisy MNIST two digits data. The top row includes learned bases, correlation between exact and learned bases and mean squared error between exact and learned bases, using *UoI-NMFcluster*. The bottom row includes the same metric while applying *Online Dictionary Learning*.

specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)", not "Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is . . ."

D. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within

et al.".

- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

IV. Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. Authors and Affiliations

The template is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

1) For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

2) For papers with less than six authors: To change the default, adjust the template as follows.

a) Selection: Highlight all author and affiliation lines.

b) Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

c) Deletion: Delete the author and affiliation lines for the extra authors.

B. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named "Heading 1", "Heading 2", "Heading 3", and "Heading 4" are prescribed.

C. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I. Table Type Styles

Table Head	Table Column Head		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a.] Sample of a Table footnote. (*Table footnote*)

Fig. 1. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

Acknowledgment *(Heading 5)*

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.