A diffusion strategy for distributed dictionary learning

Pierre Chainais, Cédric Richard

A diffusion strategy for distributed dictionary learning

Pierre Chainais[1], Cédric Richard[2].

[1]LAGIS UMR CNRS 8219 - Ecole Centrale Lille, INRIA-Lille Nord Europe, SequeL, France
pierre.chainais@ec-lille.fr.

[2]Laboratoire Lagrange UMR CNRS 7293, University of Nice Sophia-Antipolis, France
cedric.richard@unice.fr

***Abstract*— We consider the problem of a set of nodes which is required to collectively learn a common dictionary from noisy measurements. This distributed dictionary learning approach may be useful in several contexts including sensor networks. Diffusion cooperation schemes have been proposed to estimate a consensus solution to distributed linear regression. This work proposes a diffusion-based adaptive dictionary learning strategy. Each node receives measurements which may be shared or not with its neighbors. All nodes cooperate with their neighbors by sharing their local dictionary to estimate a common representation. In a diffusion approach, the resulting algorithm corresponds to a distributed alternate optimization. Beyond dictionary learning, this strategy could be adapted to many matrix factorization problems in various settings. We illustrate its efficiency on some numerical experiments, including the difficult problem of blind hyperspectral images unmixing.**

1 Introduction

In a variety of contexts, huge amounts of high dimensional data are recorded from multiple sensors. When sensor networks are considered, it is desirable that computations be distributed over the network rather than centralized in some fusion unit. Indeed, centralizing all measurements lacks robustness - a failure of the central node is fatal - and scalability due to the needed energy and communication resources. In distributed computing, every node communicates with its neighbors only and processing is carried out by every node in the network. Another important remark is that relevant information from the data usually lives in a space of much reduced dimension compared to the physical space. The extraction of this relevant information calls for the identification of some adapted sparse representation of the data. Learning an adaptive *sparse representation* of the data using a *redundant dictionary* is useful for many tasks such as storing, transmitting or analyzing the data to understand its content, see [1] for an up-to-date review. Furthermore, the problem of dictionary learning belongs to the more general family of matrix factorization problems that appears in a host of applications. We study the problem of dictionary learning distributed over a sensor network in a setting where a set of nodes is required to collectively learn an adaptive sparse representation from independent observations. We consider the situation where a set of connected nodes records data from observations of the same kind of physical system: each observation is assumed to be described by a sparse representation using a common dictionary over all sensors. For instance, a set of cameras observe the same kind of scenes or a set of microphones records the same kind of sound environment.

The dictionary learning and the matrix factorization problems are connected to the linear regression problem. Indeed, the classical approach based on alternate minimization on the coefficients $\mathbf{X}$ and the dictionary $\mathbf{D}$ solves two linear regression problems knowing respectively $\mathbf{D}$ or $\mathbf{X}$. Several recent works have proposed efficient solutions to the problem of least mean square (LMS) distributed linear regression, see [2] and references therein. The main idea is to use a so-called *diffusion strategy*: each node n carries out its own estimation $\mathbf{D}_n$ of the same underlying linear regression vector $\mathbf{D}$ but can communicate with its neighbors as well. The information provided to some node by its neighbors is taken into account according to weights interpreted as diffusion coefficients. Under some mild conditions, the performance of such an approach in terms of mean squared error is similar to that of a centralized approach [3]. Let $\mathbf{D}_c$ the centralized estimate which uses all the observations at once. It can be shown that the error $\mathbf{E}\|\mathbf{D}_n - \mathbf{D}\|_2$ of the distributed estimate is of the same order as $\mathbf{E}\|\mathbf{D}_c - \mathbf{D}\|_2$: diffusion networks match the performance of the centralized solution.

Our work [4] gives strong indication that the classical dictionary learning technique based on block coordinate descent on the dictionary $\mathbf{D}$ and the coefficients $\mathbf{X}$ can be adapted to the distributed framework by adapting the diffusion strategy mentionned above. Our numerical experiments also strongly support this idea. Note that solving this type of matrix factorization problems is really at stake since it corresponds to many inverse problems: denoising, adaptive compression, recommendation systems... A distributed approach is highly desirable both for use in sensor networks and for parallelization of numerically expensive learning algorithms. In a second step, we may consider the more general situation where observations may also be shared between connected nodes.

2 Problem formulation

Many nodes, one dictionary. Consider N nodes over some region. In the following, boldfaced letters denote column vectors, and capital letters denote matrices. The node n takes q_n measurements $\mathbf{y}_n(i)$, $1 \leq i \leq q_n$ from some physical system. All the observations are assumed to originate from independent realizations $\mathbf{s}_n(i)$ of the same underlying stochastic source process $\mathbf{s}$. Each measurement is a noisy measurement

$$\mathbf{y}_n(i) = \mathbf{s}_n(i) + \mathbf{z}_n(i) \tag{1}$$

where $\mathbf{z}$ denotes the usual i.i.d. Gaussian noise with covariance matrix $\Sigma_n = \sigma_n^2 \mathbb{I}$. Our purpose is to learn a common redundant dictionary $\mathbf{D}$ which carries the characteristic properties of the data. This dictionary must yield a sparse representation of $\mathbf{s}$ so that:

$$\forall n, \quad \mathbf{y}_n(i) = \underbrace{\mathbf{D}\mathbf{x}_n(i)}_{\mathbf{s}_n(i)} + \mathbf{z}_n(i) \tag{2}$$

where $\mathbf{x}_n(i)$ features the coefficients $x_{nk}(i)$ associated to the contribution of atom $\mathbf{d}_k$, the k-th column in the dictionary ma-

trix $\mathbf{D}$, to $\mathbf{s}_n(i)$. The sparsity of $\mathbf{x}_n(i)$ means that only few components of $\mathbf{x}_n(i)$ are non zero.

We consider the situation where a unique dictionary $\mathbf{D}$ generates the observations at all nodes. On the contrary, observations will first not be shared between nodes (this is one potential generalization). Our purpose is to learn this dictionary in a distributed manner thanks to in-network computing only. As a consequence, each node will locally estimate a local dictionary $\mathbf{D}_n$ thanks to i) its observations $\mathbf{y}_n$ and ii) communication with its neighbors. The neighborhood of node n will be denoted by $\mathcal{N}_n$, including node n itself.

Dictionary learning. Various approaches to dictionary learning have been proposed [1]. Usually, in the centralized setting, the q observations are denoted by $\mathbf{y}(i) \in \mathbb{R}^p$ and grouped in a matrix $\mathbf{Y} = [\mathbf{y}(1), ..., \mathbf{y}(q)]$. As a consequence, $\mathbf{Y} \in \mathbb{R}^{p \times q}$. The dictionary (associated to some linear transform) is denoted by $\mathbf{D} \in \mathbb{R}^{p \times K}$: each column is one atom $\mathbf{d}_k$ of the dictionary. The coefficients associated to observations are $\mathbf{X} = [\mathbf{x}(1), ..., \mathbf{x}(q)]$. We will consider learning methods based on block coordinate descent or alternate optimization on $\mathbf{D}$ and $\mathbf{X}$ with a sparsity constraint on $\mathbf{X}$ [5, 1, 6]. The data is represented as the sum of a linear combination of atoms and a noise term $\mathbf{Z} \in \mathbb{R}^{p \times q}$:

$$\mathbf{Y} = \mathbf{D}\mathbf{X} + \mathbf{Z} \tag{3}$$

In the most usual setting featuring white Gaussian noise, one wants to solve:

$$(\mathbf{D}, \mathbf{X}) = \text{argmin}_{(\mathbf{D},\mathbf{X})} \quad \frac{1}{2}||\mathbf{Y} - \mathbf{D}\mathbf{X}||_2^2 + \lambda||\mathbf{X}||_1 \tag{4}$$

Under some mild conditions, this problem is known to provide a solution to the underlying L0-penalized problem [7].

3 Distributed alternate optimization for dictionary learning

Algorithm. The Adapt Then Combine diffusion strategy [2] for distributed estimation originates the following approach to distributed alternate optimization for dictionary learning. Diffusion is ensured by the communication between nodes sharing their dictionary estimate with neighbors in $\mathcal{N}_n$. Observations are taken simultaneously at each node so that a whole data matrix $\mathbf{Y}_n$ is assumed to be available at node n. Here index i stands for iterations. The case where data arrive sequentially at each node can also be dealt with at the price of a natural adaptation of the present approach. Each node must estimate both its local dictionary $\mathbf{D}_n$ and the coefficients $\mathbf{X}_n$ which describe observations $\mathbf{Y}_n = \mathbf{D}_n\mathbf{X}_n + \mathbf{Z}_n$. At each iteration i, only the local dictionary estimates $\mathbf{D}_{n,i}$ are assumed to be shared between neighbors, not observations. In summary, sparse representations are computed locally. Then each node updates its dictionary as a function of its local observations $\mathbf{Y}_n$ (Adapt step) and its neighbors' dictionaries (Combine step). Based on known results for the ATC strategy in its usual setting, we expect that Algorithm 1 below converges to an accurate estimate of the common underlying dictionary $\mathbf{D}$. Various choices can be considered for $\mathbf{A}$ such as some a priori fixed matrix $\mathbf{A}$ or with the relative degree variance (ν_ℓ = degree of node ℓ):

$$a_{\ell,n} = \frac{\nu_\ell \sigma_\ell^2}{\sum_{m \in \mathcal{N}_n} \nu_m \sigma_m^2} \tag{5}$$

Initialize $\mathbf{D}_{n,0}$, $\forall n$ (random subset of K observations $y_n(i)$).
Given a matrix $\mathbf{A}$ satisfying $\mathbf{1}^T\mathbf{A} = \mathbf{1}^T$, $i = 0$,

Repeat until convergence of $(\mathbf{D}_{n,i}, \mathbf{X}_{n,i})_{n=1:N}$

For each node n **repeat:**

1) Optimization w.r.t. $\mathbf{X}_{n,i}$ (sparse coding):
Given the dictionary $\mathbf{D}_{n,i}$, the coefficients $\mathbf{X}_{n,i}$ are estimated using a sparse coding method (Basis Pursuit, OMP, FOCUSS,...)

2) Optimization w.r.t. $\mathbf{D}_{n,i}$ (dictionary) e.g. by gradient descent:

$$\begin{cases} \boldsymbol{\psi}_{n,i+1} &= \mathbf{D}_{n,i} + \mu_n^D(\mathbf{Y}_n - \mathbf{D}_{n,i}\mathbf{X}_{n,i})\mathbf{X}_{n,i}^T \\ \mathbf{D}_{n,i+1} &= \sum_{\ell \in \mathcal{N}_k} a_{\ell,n}^D \boldsymbol{\psi}_{\ell,i} \text{ (diffusion)} \end{cases}$$

and $\forall 1 \leq k \leq K, \mathbf{d}_k \leftarrow \frac{\mathbf{d}_k}{\|\mathbf{d}_k\|_2}$ (normalization)

$\boldsymbol{\psi}_{n,i+1}$ can also be updated by MOD or K-SVD,...at node n.
EndFor (n)
$i \leftarrow i + 1$
EndRepeat

Numerical experiments. We present some numerical experiments to illustrate the relevance and efficiency of our approach. For instance we show the results obtained on a dataset built from a redundant random dictionary of 48 atoms of dimension 16 corresponding to image patches of size 4×4. Each data $y_n(i)$ is the linear combination of 3 atoms with i.i.d. coefficients uniformly distributed over $[-0.5, 0.5]$; various Gaussian noise levels are considered. We show that a set of 4 nodes in a symmetrically connected network consistently learn the same dictionary of 4×4 patches with good accuracy (45 atoms of the initial dictionary of 48 atoms are recovered with $\langle \mathbf{d}_j, \mathbf{d}_j^{(o)} \rangle \geq 0.99$).

We will also show the results of an application to the problem of blind unmixing of hyperspectral images. In this application, the network is simply made of connections between pixels which are spatially close or which carry similar spectral information. The graph underlying (hyper-)pixels makes them collaborate to learn spectral endmembers.

4 Conclusion

We present an original algorithm which solves the problem of distributed dictionary learning over a sensor network. This is made possible thanks to a diffusion strategy which permits local communication between neighbors. Connected nodes exchange their local dictionaries estimated from disjoint subsets of data. This algorithm adapts usual dictionary learning techniques for sparse representation to the context of in-network computing. This approach to the general problem of distributed matrix factorization paves the way towards many prospects and applications. Moreover, as far as computational complexity is concerned, distributed parallel implementations are a potentially interesting alternative to online learning techniques [8]. We may even consider a dynamical context where observations arrive over time so that the dictionary would also be learnt dynamically.

References

[1] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27 –38, march 2011.

[2] F. Cattivelli and A. Sayed, "Diffusion lms strategies for distributed estimation," *Signal Processing, IEEE Transactions on*, vol. 58, no. 3, pp. 1035 –1048, march 2010.

[3] X. Zhao and A. H. Sayed, "Performance limits for distributed estimation over lms adaptive networks," *IEEE Transactions on Signal Processing*, vol. 60, no. 10, pp. 5107–5124, 2012.

[4] P. Chainais and C. Richard, "Learning a common dictionary over a sensor network," in *Proc. of IEEE CAMSAP 2013*, 2013.

[5] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of Optimization Theory and Applications*, vol. 109, pp. 475–494, 2001. [Online]. Available: http://dx.doi.org/10.1023/A%3A1017501703105

[6] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311 –4322, nov. 2006.

[7] J.-L. Starck, F. Murtagh, and J. Fadili, *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity*. Cambridge University Press, 2010.

[8] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Mar. 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1756006.1756008