Dictionary Learning via Projected Maximal Exploration

Boris Mailhé, Mark D. Plumbley
Queen Mary University of London
School of Electronic Engineering and Computer Science
Centre for Digital Music
Mile End Road, London E1 4NS, United Kingdom
firstname.name@eecs.qmul.ac.uk

***Abstract*—This work presents a geometrical analysis of the Large Step Gradient Descent (LGD) dictionary learning algorithm. LGD updates the atoms of the dictionary using a gradient step with a step size equal to twice the optimal step size.**

We show that the large step gradient descent can be understood as a maximal exploration step where one goes as far away as possible without increasing the the error. We also show that the LGD iteration is monotonic when the algorithm used for the sparse approximation step is close enough to orthogonal.

***Index Terms*—Dictionary learning, sparse representations, global optimization, projected gradient descent**

I. Dictionary learning

We consider the dictionary learning problem of minimizing the cost function:

$$f(\mathbf{\Phi}, \mathbf{X}) = \|\mathbf{S} - \mathbf{\Phi}\mathbf{X}\|_{FRO}^2 \tag{1}$$

under the constraints

$$\forall m \in [1, M],\ \|\boldsymbol{\varphi}_m\|_2 = 1 \tag{2}$$

$$\forall n \in [1, N],\ \|\mathbf{x}_n\|_0 \leq K \tag{3}$$

with $\boldsymbol{\varphi}$ an atom (or column) of $\mathbf{\Phi}$ and $\|\mathbf{x}_n\|_0$ the number of non-zero coefficients in the n^{th} column of the sparse coefficients $\mathbf{X}$. Common algorithms such as MOD [2] or K-SVD [3] solve this problem by alternating between the optimization of $\mathbf{X}$ (sparse approximation) and of $\mathbf{\Phi}$ (dictionary update). For the LGD algorithm, the dictionary update step is performed using a projected gradient descent iteration for each atom with a step size equal to twice the optimal step [1]. In this work we provide guarantees on the monotonicity of the LGD iteration and some insight on the reason why it outperforms MOD and K-SVD.

II. LGD

Numerical simulations showed that when the support of the sparse decomposition is known, a simple fixed step gradient descent is more likely to retrieve the best dictionary than either an optimal step gradient, MOD or K-SVD. Since the optimal step size for one iteration does not yield the best performance over the whole algorithm, LGD uses a step size that is twice the optimal one for each atom $\boldsymbol{\varphi}_n$ instead:

$$\boldsymbol{\varphi}_n^{(i+0.5)} = \boldsymbol{\varphi}_n^{(i)} + \frac{2}{\|\mathbf{x}^n\|_2^2}\mathbf{R}\mathbf{x}^{nT} \tag{4}$$

$$\boldsymbol{\varphi}_n^{(i+1)} \leftarrow \frac{\boldsymbol{\varphi}_n^{(i+0.5)}}{\left\|\boldsymbol{\varphi}_n^{(i+0.5)}\right\|_2} \tag{5}$$

with $\mathbf{R} = \mathbf{s} - \mathbf{\Phi}\mathbf{X}$ the residual and $\mathbf{x}^i$ the i^{th} row of $\mathbf{X}$.

III. Results

When updating one atom only, one iteration of optimal step size gradient descent descent finds the global minimum of the cost function f, and the level sets of that function are hyperspheres. In that case, the points $\boldsymbol{\varphi}_n^{(i)}$ and $\boldsymbol{\varphi}_n^{(i+0.5)}$ are diametrically opposed on the same level set. Instead of decreasing the error as much as possible, the LGD gradient step computes the point that is the furthest away from the starting point, under the constraint that the error does not increase. This explains why LGD is better at exploring the space than the optimal step size descent, but it does not explain why the error decreases.

It is in fact the renormalization step that decreases the error. One can prove that if $\left\|\boldsymbol{\varphi}_n^{(i+0.5)}\right\|_2 > 1$, then $f(\boldsymbol{\varphi}_n^{(i+1)}) < f(\boldsymbol{\varphi}_n^{(i)})$. If the algorithm used for the sparse decomposition is either an orthogonal algorithm or a thresholding algorithm, then that condition is always satisfied.

LGD is a deterministic, monotonic, parameter-free algorithm that performs global optimization by iterating a maximal exploration step and a projection step that reduces the error although it is not its primary goal. In our experiments, it succeeds at finding the best dictionary in 80% of the tries whereas the K-SVD success rate is only 18%.

This work was supported by the EPSRC Project EP/G007144/1 Machine Listening using Sparse Representations and by the EU FET-Open project FP7-ICT-225913-SMALL.

References

[1] B. Mailhé and M. D. Plumbley, "Dictionary Learning with Large Step Gradient Descent for Sparse Representations", *in Proc. LVA/ICA 2012*, pp. 231-238.

[2] K. Engan, S. O. Aase and J. Hakon Usoy, "Method of Optimal Directions for Frame Design", *in Proc. ICASSP 1999*, vol. 5, pp. 2443-2446.

[3] M. Aharon, M. Elad and A. Bruckstein , "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation", *in IEEE Trans. Signal Processing*, 2006, vol. 54, no. 11, pp. 4311-4322.

 Published in: Proceedings of the 2013 IEEE Global Conference on Signal and Information Processing (GlobalSIP), December 3-5, 2013, Austin, Texas, USA, p. 626. DOI:10.1109/GlobalSIP.2013.6736963