This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2487966, IEEE Transactions on Pattern Analysis and Machine Intelligence

Dictionary learning for sparse coding: Algorithms and convergence analysis

Chenglong Bao, Hui Ji, Yuhui Quan and Zuowei Shen

Abstract—In recent years, sparse coding has been widely used in many applications ranging from image processing to pattern recognition. Most existing sparse coding based applications require solving a class of challenging non-smooth and non-convex optimization problems. Despite the fact that many numerical methods have been developed for solving these problems, it remains an open problem to find a numerical method which is not only empirically fast, but also has mathematically guaranteed strong convergence. In this paper, we propose an alternating iteration scheme for solving such problems. A rigorous convergence analysis shows that the proposed method satisfies the global convergence property: the whole sequence of iterates is convergent and converges to a critical point. Besides the theoretical soundness, the practical benefit of the proposed method is validated in applications including image restoration and recognition. Experiments show that the proposed method achieves similar results with less computation when compared to widely used methods such as K-SVD.

Index Terms—dictionary learning, sparse coding, non-convex optimization, convergence analysis

1 Introduction

Sparse coding aims to construct succinct representations of input data, i.e. a linear combination of only a few atoms of the dictionary learned from the data itself. Sparse coding techniques have been widely used in applications, e.g. image processing, audio processing, visual recognition, clustering and machine learning [1]. Given a set of signals $Y := \{y_1, y_2, \dots, y_p\}$, sparse coding aims at finding a dictionary $D := \{d_1, d_2, \dots, d_m\}$ such that each signal $y \in Y$ can be well-approximated by a linear combination of $\{d_j\}_{j=1}^m$, i.e., $y = \sum_{\ell=1}^m c_\ell d_\ell$, and most coefficients c_ℓs are zero or close to zero. Sparse coding can be typically formulated as the following optimization problem:

$$\min_{D,\{c_i\}_{i=1}^p} \sum_{i=1}^p \frac{1}{2}\|y_i - Dc_i\|^2 + \lambda\|c_i\|_0, \qquad (1)$$

subject to $\|d_i\| = 1, 1 \le i \le m$. The dictionary dimension m is usually larger than the signal dimension n.

1.1 Overview of the problem

The problem (1) is a non-convex problem whose non-convexity comes from two sources: the sparsity-promoting ℓ_0-norm, and the bi-linearity between the dictionary D and codes $\{c_i\}_{i=1}^p$ in the fidelity term. Most sparse coding based applications adopt an alternating iteration scheme: for $k = 1, 2, \dots,$

(a) *sparse approximation*: update codes $\{c_i\}_{i=1}^p$ via solving (1) with the dictionary fixed from the previous iteration, i.e. $D := D^k$.

(b) *dictionary refinement*: update the dictionary D via solving (1) with codes fixed from the previous iteration, i.e. $c_i := c_i^{k+1}$ for $i = 1, \dots, p$.

Thus, each iteration requires solving two non-convex sub-problems (a) and (b).

The sub-problem (a) is an NP-hard problem [2], and thus only a sub-optimal solution can be found in polynomial time. Existing algorithms for solving (a) either use greedy strategies to obtain a local minimizer (e.g. orthogonal matching pursuit (OMP) [3]), or replace the ℓ_0-norm by its convex relaxation, the ℓ_1-norm, to provide an approximate solution (e.g. [4], [5], [6], [7]).

The sub-problem (b) is also a non-convex problem due to the existence of norm equality constraints on atoms $\{d_i\}_{i=1}^m$. Furthermore, some additional non-convex constraints on D are used for better performance in various applications, e.g. compressed sensing and visual recognition. One such constraint is an upper bound on the *mutual coherence* $\mu(D) = \max_{i\neq j} |\langle d_i, d_j\rangle|$ of the dictionary, which measures the correlation of atoms. A model often seen in visual recognition (see e.g. [8], [9], [10]) is defined as follows,

$$\min_{D,C} \|Y - DC\|^2 + \lambda\|C\|_0 + \frac{\mu}{2}\|D^\top D - I\|^2, \qquad (2)$$

subject to $\|d_i\| = 1, 1 \le i \le m$. Due to the additional term $\|D^\top D - I\|^2$, the problem (2) is harder than (1).

1.2 Motivations and our contributions

Despite the wide use of sparse coding techniques, the study of algorithms for solving (1) and its variants with rigorous convergence analysis has been scant in the literature. The most popular algorithm for solving the constrained version of (1) is the K-SVD

• *C. Bao, H. Ji, Y. Quan and Z. Shen are with the Department of Mathematics, National University of Singapore, Singapore, 119076.*
E-mail: matbc,matjh,matquan,matzuows@nus.edu.sg