Proximal methods for hierarchical sparse coding

Jenatton, Mairal, Obozinski, Bach '11

Geelon So
(agso@eng.ucsd.edu)

May 20, 2020

Journal of Machine Learning Research 12 (2011) 2297-2334 Submitted 9/10; Revised 3/11; Published 7/11

Proximal Methods for Hierarchical Sparse Coding

Rodolphe Jenatton[*†] RODOLPHE.JENATTON@INRIA.FR
Julien Mairal[*†] JULIEN.MAIRAL@INRIA.FR
Guillaume Obozinski[†] GUILLAUME.OBOZINSKI@INRIA.FR
Francis Bach[†] FRANCIS.BACH@INRIA.FR
INRIA - WILLOW Project-Team
Laboratoire d'Informatique de l'Ecole Normale Supérieure (INRIA/ENS/CNRS UMR 8548)
23, avenue d'Italie
75214 Paris CEDEX 13, France

Dictionary learning s-sparse representations

Problem: given data points $y_1, \ldots, y_N \in \mathbb{R}^d$, find dictionary $\mathbf{D} = \begin{bmatrix} d_1 & \cdots & d_k \end{bmatrix}$ such that:

$$y_i \approx \mathbf{D} x_i$$

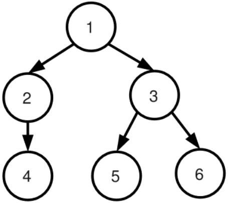
for $x_i \in \mathbb{R}^k$ and $\|x_i\|_0 \leq s$.

- Each y_i is approximately the linear combination of any s dictionary atoms.

Hierarchical sparse coding

Problem: we have **structured sparsity** assumptions in the form of a directed rooted tree where the nodes are dictionary atoms.

- if a representation uses the ith atom, then $\text{ancestors}(i)$ also count toward sparsity budget.



- valid 3-sparse representation: $\alpha_1 d_1 + \alpha_3 d_3 + \alpha_5 d_5$
- invalid 3-sparse representation: $\alpha_4 d_4 + \alpha_5 d_5 + \alpha_6 d_6$

Example: topic modeling

Let $y \in \mathbb{R}^d$ be a document:

- vocabulary of size d where **words** have **one-hot encoding**
- **documents** are represented by **normalized bag-of-words** (the ith word appears a $y(i)$ fraction of times in document)
- **structured sparsity** assumption: topics have subtopics have subtopics—topics correspond to a distribution over words

Example: topic modeling

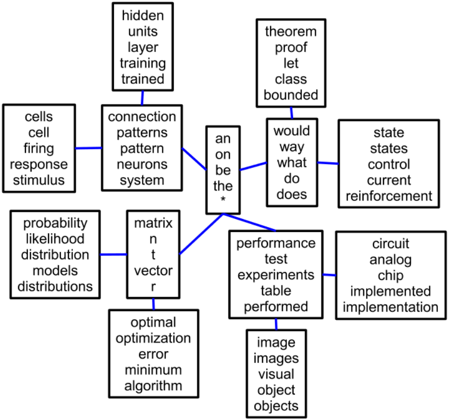


Figure 1: Example topic model generated by hierarchical sparse coding using 1714 NIPS proceedings papers (1988–1999); each node shows the top most common words.

Non-convex optimization problem

Given $y \in \mathbb{R}^d$ and dictionary $\mathbf{D}$, minimize:

$$\min_{\substack{x \in \mathbb{R}^k \\ \|x\|_0 \leq s}} \|y - \mathbf{D}x\|_2^2,$$

satisfying the constraint:

$$x(i) \neq 0$$

$$\Downarrow$$

$$\bigwedge_{j \in \text{ancestors}(i)} x(j) \neq 0.$$

Tree-structured groups

Given a directed graph $G = ([k], E)$, its associated **group** is the set $\mathcal{G}$ of subsets of $[k]$:

$$\mathcal{G} = \{\text{descendants}(i) : i \in [k]\}.$$

A group is **tree-structured** if $g, h \in \mathcal{G}$ such that $g \cap h \neq \varnothing$, then either $g \subset h$ or $h \subset g$.

- Directed trees and forests yield tree-structured groups.
- There exists a (non-unique) total order $g \preceq h$ extending the usual subset ordering on $\mathcal{G}$ if it is tree-structured.

Tree-structured groups

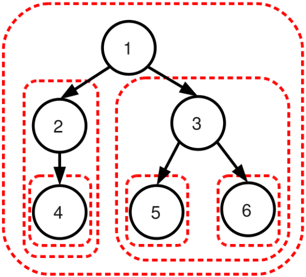


Figure 2: Groups of a tree.

Hierarchical sparsity-inducing norm

Let $\mathcal{G}$ be a group. Define $\Omega : \mathbb{R}^k \to \mathbb{R}$ by:

$$\Omega(x) = \sum_{g \in \mathcal{G}} \omega_g \|\mathbf{\Pi}_g \alpha\|,$$

where $\mathbf{\Pi}_g : \mathbb{R}^k \to \mathbb{R}^{|g|}$ projects onto the coordinates in g and $\omega_g \geq 0$ are positive weights.

- $\| \cdot \|$ is generally the ℓ_2 or ℓ_∞ norm.
- Analysis shows that solution will satisfy $\mathbf{\Pi}_g x = 0$ for some $g \in \mathcal{G}$, which means some subtrees are set to zero. [ZRY2009]

Convex optimization problem

The hierarchical sparse coding problem:

$$\min_{\mathbf{X} \in \mathbb{R}^{k \times N}} \sum_{i=1}^{N} \|y_i - \mathbf{D}x_i\|_2^2 + \lambda \Omega(x_i).$$

Proximal methods

Proximal methods

Let f be convex, continuously differentiable with L-Lipschitz gradient. If the current estimate of the minimzer is x_t, the **proximal problem** is:

$$x_{t+1} = \underset{x \in \mathbb{R}^k}{\arg\min} \, f(x_t) + (x - x_t)^\top \nabla f(x_t) + \lambda \Omega(x) + \frac{L}{2} \|x - x_t\|_2^2.$$

- By strong convexity, minimizer of proximal problem is unique.
- Achieves optimal first-order convergence rates.

Proximal methods

It is equivalent to solve the following problem:

$$x_{t+1} = \underset{x \in \mathbb{R}^k}{\arg\min} \frac{1}{2} \left\| x - \left(x_t - \frac{1}{L} \nabla f(x_t) \right) \right\|_2^2 + \frac{\lambda}{L} \Omega(x).$$

Proximal methods

Definition

The **proximal operator** *associated with the regularization term $\lambda\Omega$ is a function $\mathrm{Proc}_{\lambda\Omega}$ that maps $u \in \mathbb{R}^k$ to the unique solution:*

$$\mathrm{Proc}_{\lambda\Omega}(u) = \min_{v \in \mathbb{R}^k} \frac{1}{2}\|u - v\|_2^2 + \lambda\Omega(v).$$

- The proximal operator often has closed form.

One-pass convergence

Theorem

Let $\mathcal{G} = \{g_1, \ldots, g_m\}$ be ordered, so that $g_1 \preceq \cdots \preceq g_m$. If $\| \cdot \|$ is the ℓ_2 or ℓ_∞-norm, then:

$$\mathrm{Proc}_{\lambda\Omega} = \mathrm{Proc}_{\lambda\omega_{g_m}\| \cdot \|} \circ \cdots \circ \mathrm{Proc}_{\lambda\omega_{g_1}\| \cdot \|}.$$

- Proof makes use of conic duality, enabling block coordinate ascent algorithm in the dual.
- This does not hold for other ℓ_p-norms.

References

[JMOB2011] R. Jenatton, J. Mairal, G. Obozinski, F. Bach. "Proximal methods for hierarchical sparse coding." *JMLR*, 2011.

[ZRY2009] P. Zhao, G. Rocha, and B. Yu. "The composite absolute penalties family for grouped and hierarchical variable selection." *Annals of Statistics*, 2009.