Learning Sparse Distributions using Iterative Hard Thresholding

Jacky Y. Zhang[1] Rajiv Khanna[2] Anastasios Kyrillidis[3] Oluwasanmi Koyejo[1]

[1]University of Illinois at Urbana-Champaign [2]University of California at Berkeley [3]Rice University

Scan Me for Full Paper

Motivation

Goal: Find a sparse distribution that optimizes a given loss functional $F[\cdot]$.

Problem

$$\min_{p} \quad F[p] \qquad \text{s.t.} \quad p \in \mathcal{D}_k,$$

where $\mathcal{D}_k$ is the set of all sparse distributions.

Example: find priors for sparse structure, where

$$F[p] = KL(p\|q)$$

Background

Sparse Distribution: A k-sparse distribution is a distribution where any sample drawn has the same k-sparse support, where the k-sparse support is arbitrary.

Formally, denote the set of distributions on an n-dimensional domain $\mathcal{X}$ as:

$$\mathcal{P} = \left\{ p(\cdot) : \mathcal{X} \to \mathbb{R}_+ \middle| \sum_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x}) = 1 \right\} .$$

The set of domain restricted densities, denoted by , is the set of probability density functions with support $\mathcal{S} \subset [n]$, i.e.,

$$\mathcal{P}_{\mathcal{S}} = \{ q(\cdot) \in \mathcal{P} \mid \forall \text{supp}(\boldsymbol{x}) \not\subseteq \mathcal{S} : q(\boldsymbol{x}) = 0 \}$$

The distribution sparsity is defined as the union of all possible k-sparse support domain restricted densities, thus a **non-convex** function space:

$$\mathcal{D}_k = \cup_{|\mathcal{S}| \leq k} \mathcal{P}_{\mathcal{S}}.$$

Problem Setting: We consider discrete densities on an n-dimensional integer lattice, with totally m^n positions:

$$\mathcal{X} = \{ \boldsymbol{x} \in \mathbb{Z}^n \mid \forall i \in [n], 0 \leq x_i \leq m-1 \} .$$

Hardness of the Projection

Projection: The sparse distribution l_2 projection is

Problem

$$\Pi_{\mathcal{D}_k}(p(\cdot)) := \underset{q(\cdot) \in \mathcal{D}_k}{\arg\min} \|q(\cdot) - p(\cdot)\|_2^2.$$

The combinatorial nature of $\mathcal{D}_k$ makes the projection hard ($O(n^k)$ to solve exactly by enumeration). We prove the following two theoretical results:

Theorems

*1. The sparse distribution l_2 projection is **NP-hard***

*2. **No** deterministic algorithm exists that approximate it **in polynomial time***

Example of a generic 2-dimensional distribution with $m = 3$ projecting to $\mathcal{D}_1$.

Greedy Sparse Projection

Although the projection is hard in general, we find that a simple greedy heuristic is good when using in our main algorithm, i.e., Distribution IHT.

Algorithm

$\mathcal{S} := \emptyset$
while $|\mathcal{S}| < k$ **do**
  $j \in \arg\min_{i \in [n] \setminus \mathcal{S}} \{ \min_{p \in \mathcal{P}_{\mathcal{S} \cup i}} \|p(\cdot) - q(\cdot)\|_2^2 \}$
  $\mathcal{S} := \mathcal{S} \cup j$
end while
return $\arg\min_{p \in \mathcal{P}_{\mathcal{S}}} \|p(\cdot) - q(\cdot)\|_2^2$

Theorem

*3. The Greedy finds the **optimal** projection when used in Distribution IHT, and certain conditions are satisfied*

Distribution IHT

Algorithm

$t \leftarrow 0$
while $t < T$ **do**
  $q_{t+1}(\cdot) = p_t(\cdot) - \mu \frac{\delta F}{\delta p_t}(\cdot)$
  $p_{t+1}(\cdot) = \Pi_{\mathcal{D}_k}(q_{t+1})$
end while
return $p_T(\cdot)$

With some regular assumptions: strong convexity/smoothness and Lipschitz continuity of $F[\cdot]$, and the greedy procedure solves the projection approximately, we can prove the convergence of distribution IHT.

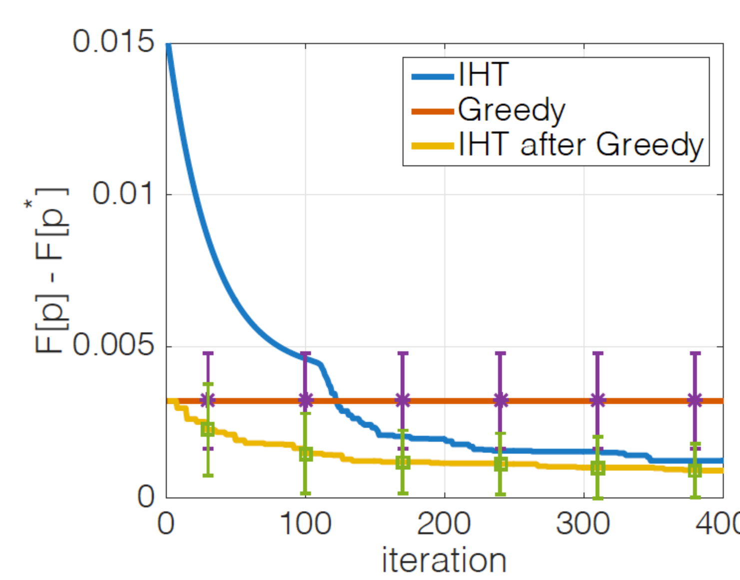
Theorem

4. If the conditions are satisfied, we have

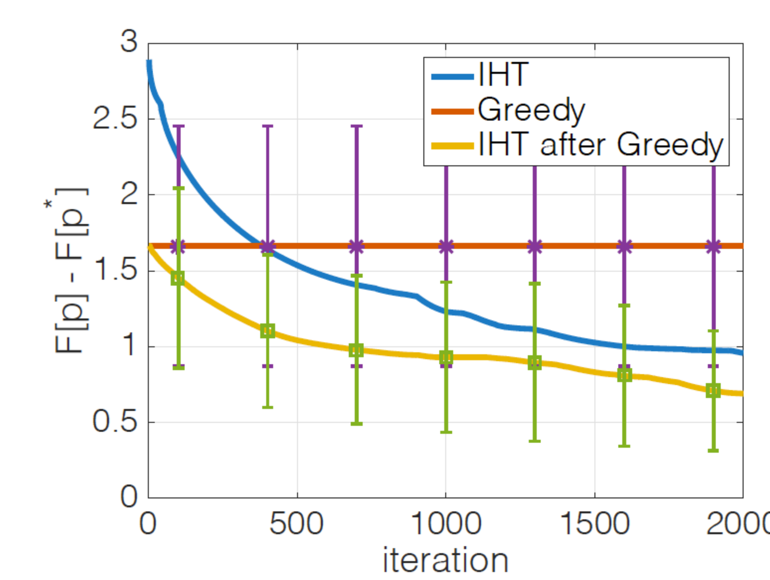
$$F[p_T(\cdot)] \leq OPT + c + \epsilon$$

after iterations $T \geq O(\log \frac{1}{\epsilon})$.

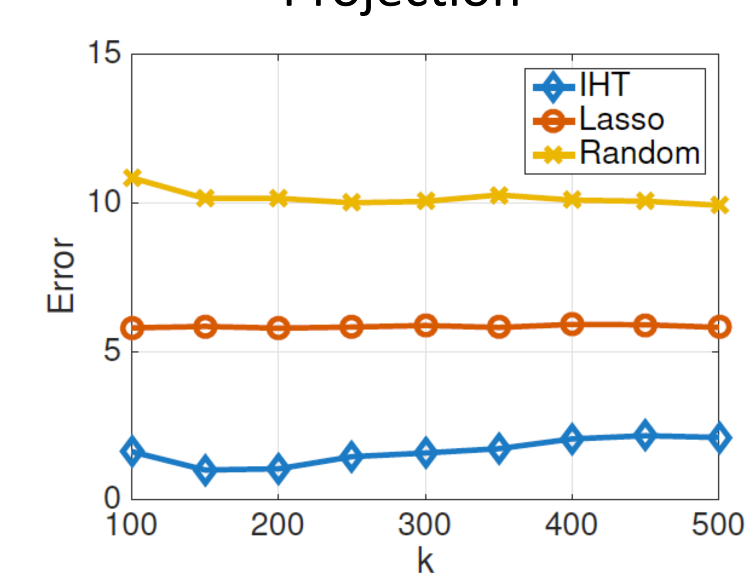
Experiments



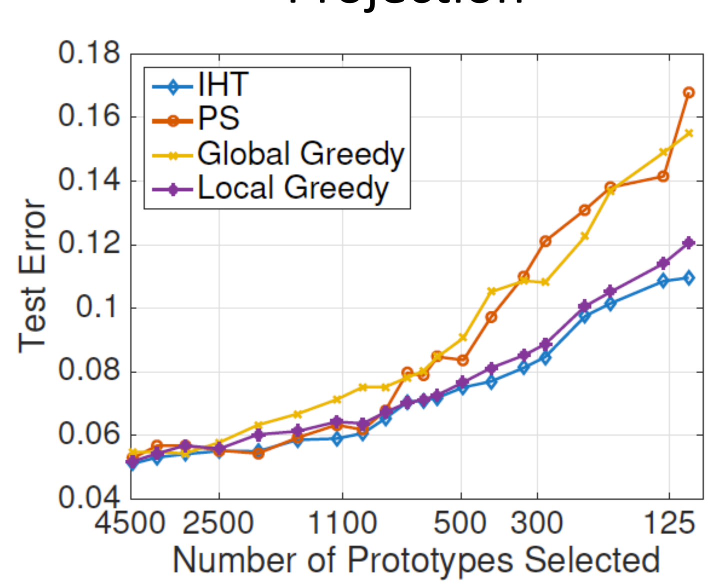
(a) Sparse Distribution l_2 Projection



(b) Sparse Distribution KL Projection



(c) Distribution Compression



(d) Dataset Compression