Approximate guarantees for dictionary learning

Bhaskara, Tai '19

Geelon So
(agso@eng.ucsd.edu)

June 17, 2020

Proceedings of Machine Learning Research vol 99:1–18, 2019 32nd Annual Conference on Learning Theory

Approximate Guarantees for Dictionary Learning

Aditya Bhaskara BHASKARA@CS.UTAH.EDU
Wai Ming Tai* WMTAI@CS.UTAH.EDU
University of Utah, Salt Lake City, Utah, USA

Introduction

Dictionary learning (DL) recap

Recall the **dictionary learning problem**: given $X \in \mathbb{R}^{d\times n}$, find **dictionary** $A \in \mathbb{R}^{d\times m}$ and **sparse encoding** $Z \in \mathbb{R}^{m\times n}$ such that:

$$d\left\{\begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix}\right. \approx \underbrace{\begin{bmatrix} | & & | \\ A_1 & \cdots & A_m \\ | & & | \end{bmatrix}}_{m} \underbrace{\begin{bmatrix} | & & | \\ | & & | \\ z_1 & \cdots & z_n \\ | & & | \\ | & & | \end{bmatrix}}_{n}$$

where $\|A_j\|_2 = 1$ and $\|z_i\|_0 \leq k$.

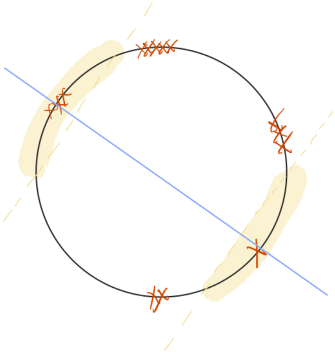
Dictionary learning objective

Subject to the sparsity constraint $\|z_i\|_0 \leq k$, the objective is to minimize the **total reconstruction error**:

$$\|X - AZ\|_F^2 = \sum_{i=1}^{n} \|x_i - Az_i\|_2^2.$$

Building intuition

Let $k = 1$. Assume $x_1, \ldots, x_n \in S^{d-1}$.



- **Sparse coding:** project each data point to maximally correlated dictionary atom.
- **Dictionary learning:** how to select dictionary atoms?

Building intuition: assumption

Assume that there exists $A_1^*, \ldots, A_m^* \in S^{d-1}$ such that for all x_i:

$$\max_{j \in [m]} \langle A_j^*, x_i \rangle^2 \geq \tau,$$

there always exists a highly-correlated dictionary atom.

Building intuition: search for highly-correlated atom

We could then iteratively **search** for directions A_t that maximizes total correlation over only highly-correlated data points:

$$\max_{A_t \in S^{d-1}} \sum_{x \in \mathcal{X}} \langle A_t, x \rangle^2 \cdot \mathbf{1}_{\langle A_t, x \rangle^2 \geq \tau},$$

where $\langle A_t, x \rangle^2$ is a correlation contribution that only counts when A_t and x have high correlation, $\mathbf{1}_{\langle A_t, x \rangle^2 \geq \tau}$.

τ-threshold correlation

Definition (τ-threshold correlation)

Let $v_1, \ldots, v_n \in \mathbb{R}^d$ *and* $w_1, \ldots, w_n \in \mathbb{R}$ *be nonnegative weights. Let* $u \in S^{d-1}$ *is a direction. The* τ-**threshold correlation** *between* u *and* $\{v_1, \ldots, v_n\}$ *is:*

$$\mathrm{TC}_\tau(u; v_1, \ldots, v_n; w_1, \ldots, w_n) = \sum_{i=1}^{n} w_i^2 \langle u, v_i \rangle^2 \cdot \mathbf{1}_{\langle u, v_i \rangle^2 \geq \tau}.$$

We'll drop the weights $w_1, \ldots, w_n$ when all the $w_i = 1$.

Building intuition: greedy algorithm

Initialize $\mathcal{X}_0 = \{x_1, \ldots, x_n\}$. While $\mathcal{X}_t \neq \varnothing$,

- Find new dictionary atom:

$$A_t \leftarrow \max_{u \in S^{d-1}} \mathrm{TC}_\tau(u; \mathcal{X}_t)$$

- Update points without representation:

$$\mathcal{X}_{t+1} \leftarrow \mathcal{X}_t \setminus \{x_i : \langle A_t, x_i \rangle^2 \geq \tau\}$$

τ-threshold correlation and reconstruction error

Notice that the reconstruction error depends on correlation:

$$\min_{z_i} \|x_i - A_t z_i\|^2 = \|x_i\|^2 - \langle A_t, x_i \rangle^2.$$

A_t will decrease reconstruction error by at least $\mathrm{TC}_\tau(A_t; \mathcal{X})$.

τ-threshold correlation for dictionary learning

τ-threshold correlation problem

Problem (τ-threshold correlation)

Let $v_1, \ldots, v_n \in S^{d-1}$ *and* $w_1, \ldots, w_n \in \mathbb{R}_{\geq 0}$. *The* τ-**threshold correlation** *problem is the optimization problem:*

$$\max_{u \in S^{d-1}} \mathrm{TC}_\tau(u; v_1, \ldots, v_n; w_1, \ldots, w_n).$$

(τ, α, β)-threshold correlation problem

Let $v_1, \ldots, v_n$ and $w_1, \ldots, w_n$ as before. Let OPT be the optimal value of the τ-threshold problem.

Problem ((τ, α, β)-threshold correlation)

*If $\alpha, \beta \in [0, 1]$, then an algorithm solving the (τ, α, β)-***threshold correlation** *problem returns a vector $u \in S^{d-1}$ such that:*

$$\mathrm{TC}_{\boldsymbol{\alpha}\cdot\tau}(u; v_1, \ldots, v_n; w_1, \ldots, w_n) \geq \beta \cdot \mathrm{OPT}.$$

That is, the threshold is reduced by a factor $1 - \alpha$ and the objective by a factor $1 - \beta$.

Claim: (τ, α, β)-TC implies approximate DL

In the following, let X and A^*, Z^* satisfy:

$$\|X - A^* Z^*\|_F^2 \leq \gamma^* \|X\|_F^2.$$

Further, let A^* be incoherent.

Theorem (Informal)

If there is an efficient algorithm solving (τ, α, β)-TC, then there is an efficient algorithm that obtains $X \approx AZ$:

- *with reconstruction error $\|X - AZ\|_F^2 \leq (\gamma^* + \varepsilon)\|X\|_F^2$*
- *with dictionary size $M = O(m/\beta\varepsilon)$*
- *with sparsity $K = O(k/\alpha\varepsilon^2)$.*

Expected reconstruction error

Notice that the condition $\|X - A^*Z^*\|_F^2 \leq \gamma^*\|X\|_F^2$ is:

$$\gamma^* \geq \frac{1}{\|X\|_F^2}\sum_{i=1}^{n}\|x_i - A^*z_i^*\|_2^2$$
$$= \frac{1}{\sum_{i=1}^{n} w_i^2}\sum_{i=1}^{n} w_i^2\|v_i - A^*z_i'\|_2^2,$$

where $w_i = \|x_i\|_2$ and $v_i = x_i/w_i$ and $z_i' = z_i^*/w_i$.

Greedy algorithm for dictionary learning: idea

Initialize $\mathcal{X} = \{x_1, \ldots, x_n\}$. While reconstruction error large:

- Find new dictionary atom: if u satisfies

$$\mathrm{TC}_{\alpha \cdot \tau}(u; \mathcal{X}) \geq \beta \cdot \mathrm{OPT},$$

  then $A_t \leftarrow u$

- Replace highly-correlated points with residual:
  - if $\langle A_t, x_i \rangle^2 \geq \alpha \cdot \tau$,

$$x_i \leftarrow x_i - \Pi_{A_t} x_i,$$

    where Π_S is the projection onto the subspace S.[1]

[1] Let Π_u be the projection onto the subspace $\mathrm{span}(u)$.

Greedy algorithm

Algorithm *GreedyPursuit*

Input: $X \in \mathbb{R}^{d \times n}$, sparsity k, dictionary size m, norm bound Λ, approximation quality ε

Initialize: $w_i \leftarrow \|x_i\|$, $v_i \leftarrow x_i/\|x_i\|$ for $i \in [n]$, $\tau \leftarrow \frac{\varepsilon^2}{k\Lambda}$, $M = mk$

1. **for** rounds $t = 1$ **to** M
2. **do** obtain solution u to (τ, α, β)-threshold correlation problem:

$$\mathrm{TC}_{\alpha \cdot \tau}(u; v_1, \ldots, v_n, w_1, \ldots, w_n) \geq \beta \cdot \mathrm{OPT}$$

and set $A_t \leftarrow u$

3. **for** $i = 1$ **to** n
4. **if** $\langle A_t, v_i \rangle^2 \geq \alpha\tau$
5. **then** $Z_{ti} \leftarrow w_i \cdot \langle A_t, v_i \rangle$ and update $v_i \leftarrow v_i - \Pi_{A_t} v_i$ with residual
6. **else** $Z_{ti} \leftarrow 0$
7. **return** $A = [A_1, \ldots, A_M] \in \mathbb{R}^{d \times M}$ and $Z \in \mathbb{R}^{M \times n}$

Potential issue with algorithm: high correlation?

- In the $k = 1$ case, we assumed for all x_i, there exists:

$$\langle A_j, x_i \rangle^2 \geq \tau.$$

- The analogous assumption for $k = 2$ would be for all x_i, there exists $u \in \operatorname{span}(A_j, A_{j'})$ where:

$$\langle u, x_i \rangle^2 \geq \tau.$$

- But this does not imply that either:

$$\langle A_j, x_i \rangle^2 \geq \tau \quad \text{or} \quad \langle A_{j'}, x_i \rangle^2 \geq \tau.$$

- So, it is not obvious that maximizing τ-TC will be useful for choosing dictionary atoms.
- Fix: incoherence assumption (actually, something weaker).

Potential issue with algorithm: residuals?

- Suppose that initially, there are directions that are highly-correlated with many of the x_i's.
- After some of the x_i's have been replaced by their residuals,

  $$x_i \leftarrow x_i - \Pi_S x_i,$$

  where $S = \mathrm{span}(A_{i_1}, \ldots, A_{i_t})$, are there still directions highly-correlated with many of the x_i's?
- Yes, as long as the residuals are not too small.

(τ, α, β)-TC implies approximate DL

In the following, let $X \in \mathbb{R}^{d \times n}$ such that there exists $A^* \in \mathbb{R}^{d \times m}$ and $Z^* \in \mathbb{R}^{m \times n}$ satisfying:

(a) the columns of A^* are unit vectors

(b) the columns of Z^* are k-sparse and satisfy $\|z_i^*\|^2 \leq \Lambda \|x_i\|^2$

(c) we have $\|X - A^* Z^*\|_F^2 \leq \gamma^* \|X\|_F^2$.

Theorem

Let $\varepsilon > 0$ be an accuracy parameter. If there is an efficient algorithm solving the (τ, α, β)-TC problem, then there is an efficient algorithm that outputs $A \in \mathbb{R}^{d \times M}$ and $Z \in \mathbb{R}^{M \times n}$ with:

- $\|X - AZ\|_F^2 \leq (\gamma^* + \varepsilon) \|X\|_F^2$,
- *the size of the dictionary is $M = O\left(m\Lambda / \beta\varepsilon\right)$,*
- *the z_i's are K-sparse with $K = O(k\Lambda / \alpha\varepsilon^2)$.*

Norm bound to fix first potential issue

Condition (b) states $\|z_i^*\|^2 \leq \Lambda\|x_i\|^2$. In the exact case, $X = A^*Z^*$, this means:

$$\|z_i^*\|^2 \leq \Lambda\|A^*z_i^*\|^2.$$

That is, A^* doesn't shrink z_i^* much: if two columns contribute to the sparse representation, then their contributions can't significantly cancel each other out.

- Incoherence assumption on A^* implies norm bound.

Lemma to fix second potential issue

Lemma (Residuals preserve correlations)

Let $v, A_1^*, \ldots, A_k^* \in S^{d-1}$ *and* $v = \sum_i \alpha_i A_i^* + y$*. For any subspace* $S \subset \mathbb{R}^d$*, let* $v' = v - \Pi_S v$ *be the residual. Then:*

$$\sum_{i=1}^{k} \langle A_i^*, v' \rangle^2 \geq \frac{(\|v'\|^2 - \|y\|^2)_+^2}{4\left(\sum_i \alpha_i^2\right)},$$

where $(r)_+ = \max\{0, r\}$ *for* $r \in \mathbb{R}$*.*

- **Read:** v has optimal representation $\sum_i \alpha_i A_i^*$ and error y.
- As long as residual still large: $\|v'\|^2 - \|y\|^2 \geq \varepsilon^2$,
- then the residual will be highly-correlated with at least one A_i^*,

$$\langle A_i^*, v' \rangle^2 \geq \frac{\varepsilon^2}{4k\Lambda}.$$

Individual correlation to collective correlation

- The previous lemma shows that if the reconstruction error of x_i is large, then it is highly-correlated with some A_j^*.
- But is there an A_j^* highly correlated to many x_i's that still have large reconstruction error?
- Suppose at time t of Algorithm *GreedyPursuit*, we have matrices $A^{(t)}$ and $Z^{(t)}$ satisfying:

  $$\|X - A^{(t)}Z^{(t)}\|_F^2 = \gamma^{(t)}\|X\|_F^2 \geq (\gamma^* + \varepsilon)\|X\|_F^2,$$

  then a significant number of residuals still have large reconstruction error.
- Each residual is highly correlated to some A_j^*; a large fraction of reconstruction error is shared across these m directions.
- At least one of the A_j^* account for at least $\frac{1}{m}$-fraction.

Collective correlation

Lemma (Existence of highly-correlated dictionary atom)

Let $A^{(t)}$, $Z^{(t)}$ *and* $\gamma^{(t)}$ *as above. Let* $x_i^{(t)} = x_i - A^{(t)} z_i^{(t)}$ *be the residual of* x_i *at time* t. *There is a* $j \in [m]$ *and* $R \subset [n]$ *so that:*

$$\langle A_j^*, x_i^{(t)} \rangle^2 \geq \frac{\varepsilon^2}{16k\Lambda} \|x_i\|^2$$

for all $i \in R$ *and*

$$\sum_{i \in R} \langle A_j^*, x_i^{(t)} \rangle^2 \geq \frac{(\gamma^{(t)} - \gamma^*)^2}{16m\Lambda} \|X\|_F^2.$$

- **Read:** there is an A_j^* highly correlated to x_i for $i \in R \subset [n]$,
- and adding A_j^* can significantly decrease reconstruction error.

Apply (τ, α, β)-TC

Corollary

If we have an algorithm to solve (τ, α, β)-threshold correlation, where $\tau = \varepsilon^2/16k\Lambda$, then at time t of Algorithm GreedyPursuit, then we can construct a vector A_t so for some subset $R' \subset [m]$,

$$\langle A_t, x_i^{(t)} \rangle^2 \geq \frac{\alpha \cdot \varepsilon^2}{16k\Lambda} \|x_i\|^2$$

for all $i \in R'$ and

$$\sum_{i \in R'} \langle A_t, x_i^{(t)} \rangle^2 \geq \frac{\beta \cdot (\gamma^{(t)} - \gamma^*)^2}{16m\Lambda} \|X\|_F^2.$$

Corollary implies Theorem

Proof sketch of Theorem.

- The first result of the corollary: every time we update z_i with a nonzero entry, we decrease the approximation error for x_i by

$$(\alpha\varepsilon^2/16k\Lambda) \cdot \|x_i\|^2.$$

  The final sparsity of z_i is at most $K = O(k\Lambda/\alpha\varepsilon^2)$.

- As for the reconstruction error, algebra applied to the second result of the corollary shows that

$$\gamma^{(t)} - \gamma^* \leq 16m\Lambda/\beta t.$$

  To obtain approximation error $(\gamma^* + \varepsilon)\|X\|_F^2$, need at most $M = O(m\Lambda/\beta\varepsilon)$ steps.

□

Efficient algorithm for (τ, α, β)-TC

Existence of efficient algorithm

Theorem (Efficient solution to (τ, α, β)-TC)

Let $\tau \in (0, 1)$. There is a polynomial time algorithm that solves $(\tau, \tau/4, \tau^2/32)$-threshold correlation.

Algorithm

Suppose $v_1, \dots, v_n \in S^{d-1}$ and $w_1, \dots, w_n \in \mathbb{R}_{\geq 0}$ and u maximizes τ-TC. WLOG, let first q vectors be highly-correlated:

$$\langle u, v_1 \rangle^2, \dots, \langle u, v_q \rangle^2 \geq \tau.$$

- These v_i's are contained in spherical cap around $\pm u$.
- In fact, one of the v_ℓ's for $\ell \in [q]$ correlated with many others:

$$\mathrm{TC}_{\tau \cdot \tau/4}(v_\ell; v_1, \dots, v_q; w_1, \dots w_q) \geq \frac{\tau^2}{32} \sum_{i=1}^{q} w_i^2 \langle u, v_i \rangle^2.$$

- Algorithm: iterate through $v_1, \dots, v_n$ and compute the $\tau^2/4$-TC with the whole set. Return v_ℓ satisfying above.