An iterative thresholding and K residual means algorithm for dictionary learning

Karin Schnass

Department of Mathematics
University of Innsbruck

karin.schnass@uibk.ac.at

Cambridge, July 6

reminder: sparsity & dictionaries

N vectors $y_n \in \mathbb{R}^d$
$Y = (y_1, \ldots, y_N)$
d, N large

reminder: sparsity & dictionaries

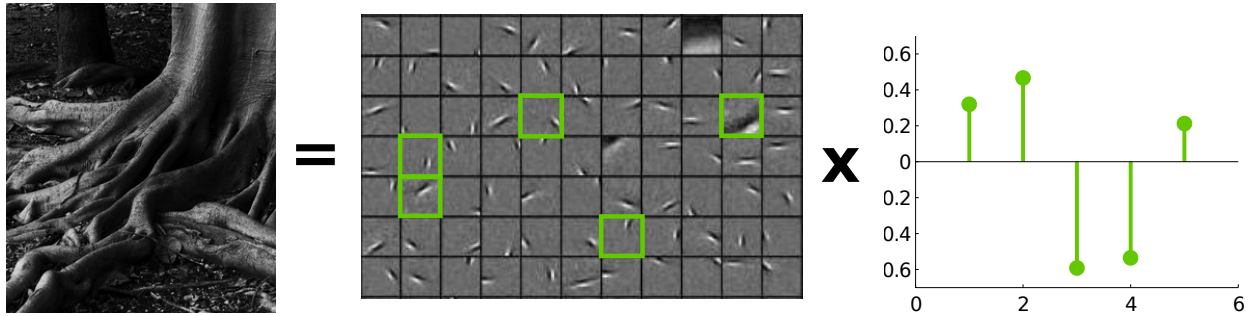
N vectors $y_n \in \mathbb{R}^d$
$Y = (y_1, \ldots, y_N)$
d, N large

A dictionary is a set of building blocks (atoms) so that every signal of our class can be constructed using only a small (sparse) number of blocks...

reminder: sparsity & dictionaries

N vectors $y_n \in \mathbb{R}^d$
$Y = (y_1, \ldots, y_N)$
d, N large

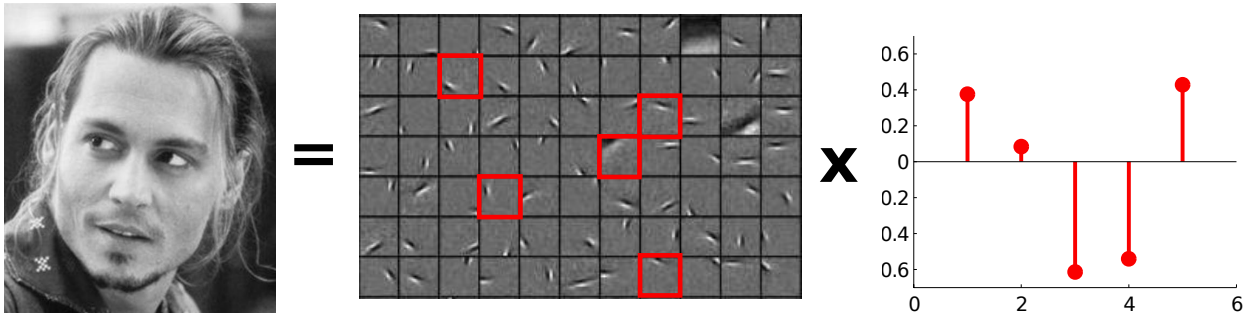
A dictionary is a set of building blocks (atoms) so that every signal of our class can be constructed using only a small (sparse) number of blocks...



reminder: sparsity & dictionaries

N vectors $y_n \in \mathbb{R}^d$
$Y = (y_1, \ldots, y_N)$
d, N large

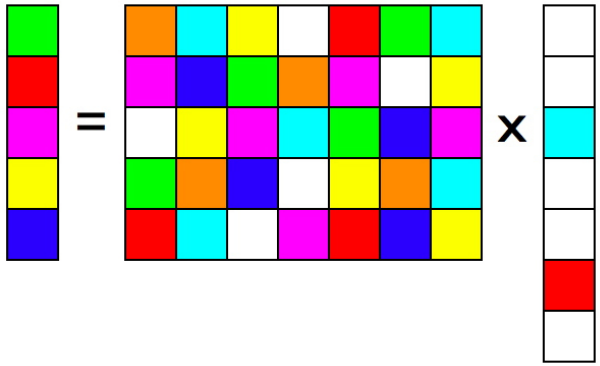
A dictionary is a set of building blocks (atoms) so that every signal of our class can be constructed using only a small (sparse) number of blocks...



reminder: sparsity & dictionaries

N vectors $y_n \in \mathbb{R}^d$
$Y = (y_1, \ldots, y_N)$
d, N large

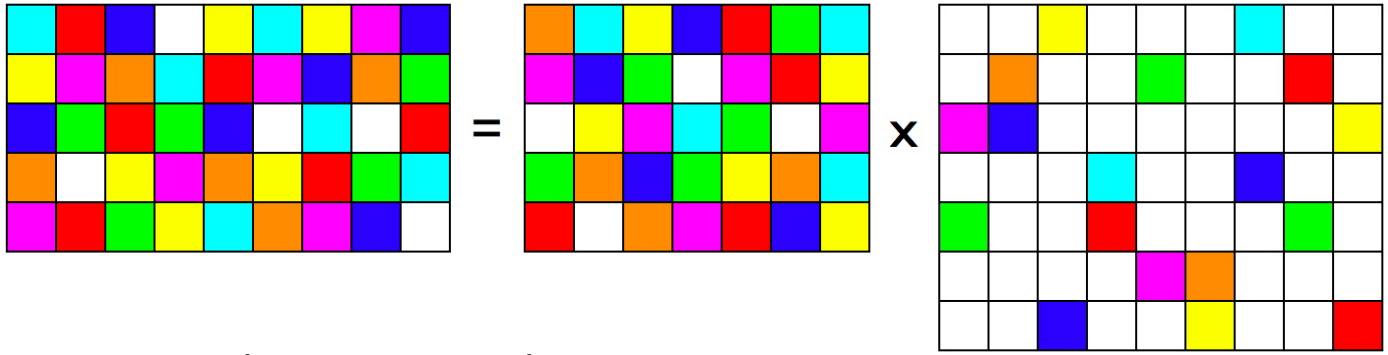
A dictionary is a set of building blocks (atoms) so that every signal of our class can be constructed using only a small (sparse) number of blocks...



reminder: sparsity & dictionaries

N vectors $y_n \in \mathbb{R}^d$
$Y = (y_1, \ldots, y_N)$
d, N large

A dictionary is a set of building blocks (atoms) so that every signal of our class can be constructed using only a small (sparse) number of blocks...



$K \ll N$ atoms (normalised)
$S \ll d$ atoms/signal

reminder: why do we care about sparsity?

A sparse representation of the data is the basis for

reminder: why do we care about sparsity?

A sparse representation of the data is the basis for

- efficient data processing,
  e.g denoising, compressed sensing, inpainting

Example: inpainting[a]

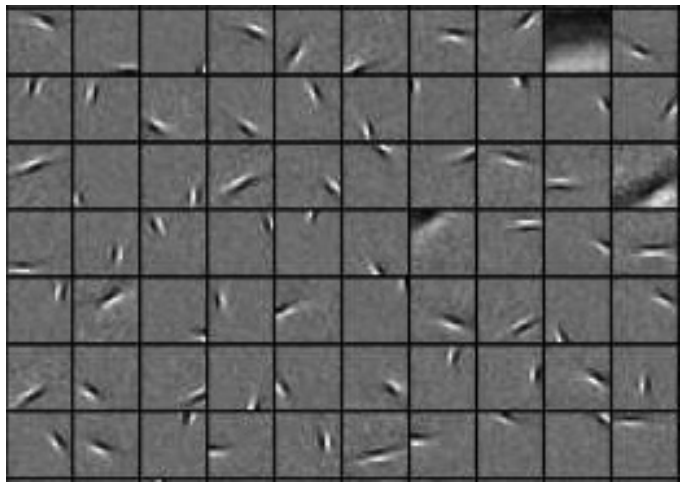
[a] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online learning for matrix factorization and sparse coding.

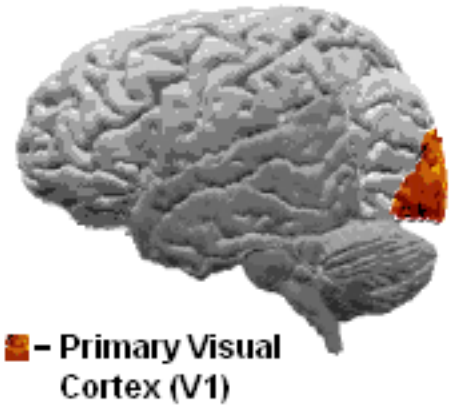
reminder: why do we care about sparsity?

A sparse representation of the data is the basis for

- efficient data processing,
  e.g denoising, compressed sensing, inpainting
- efficient data analysis,
  e.g source separation, anomaly detection, sparse components

Example: sparse components[a]





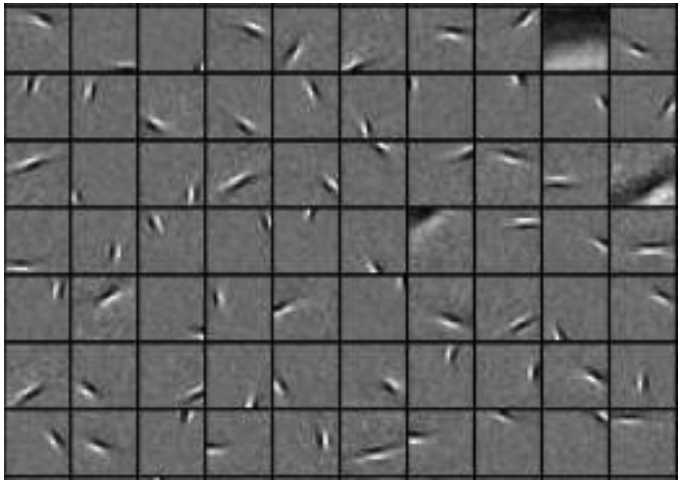
[a] D.J. Field, B.A. Olshausen, Emergence of simple-cell receptive field properties by learning a sparse code for natural images.

reminder: why do we care about sparsity?

A sparse representation of the data is the basis for

- efficient data processing,
  e.g denoising, compressed sensing, inpainting
- efficient data analysis,
  e.g source separation, anomaly detection, sparse components

In all examples:



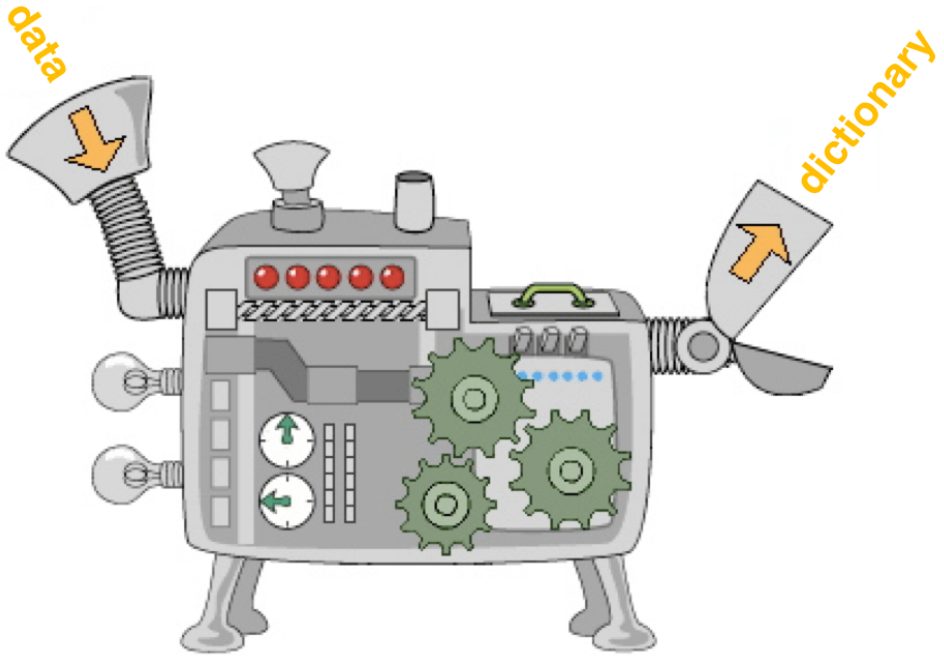
the sparser - the more efficient

reminder: dictionary learning

data: $Y = (y_1, \ldots, y_N)$
N vectors $y_n \in \mathbb{R}^d$
d, N large

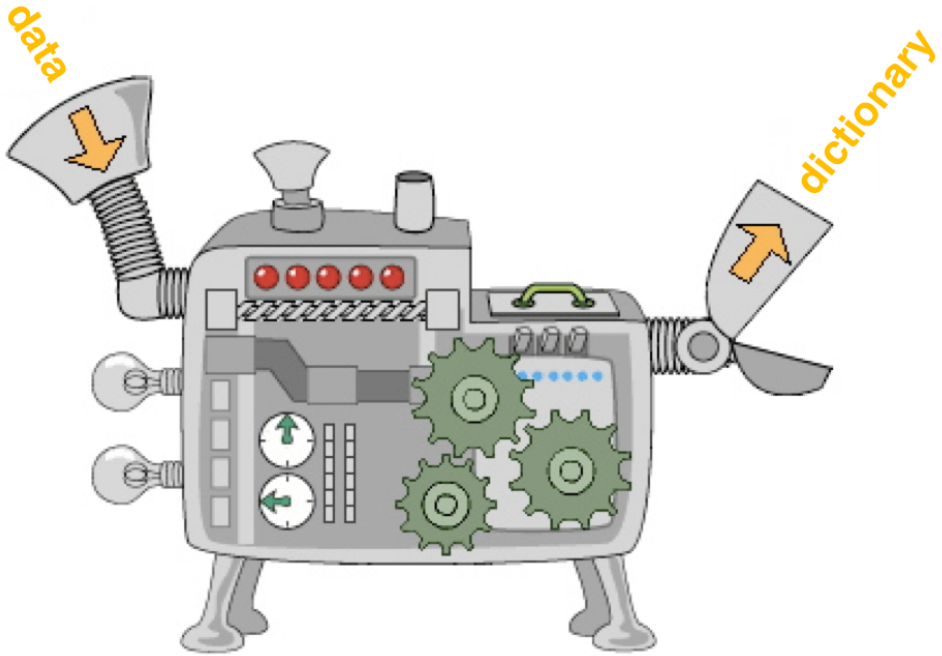
reminder: dictionary learning

data: $Y = (y_1, \ldots, y_N)$
N vectors $y_n \in \mathbb{R}^d$
d, N large



reminder: dictionary learning

data: $Y = (y_1, \ldots, y_N)$
N vectors $y_n \in \mathbb{R}^d$
d, N large

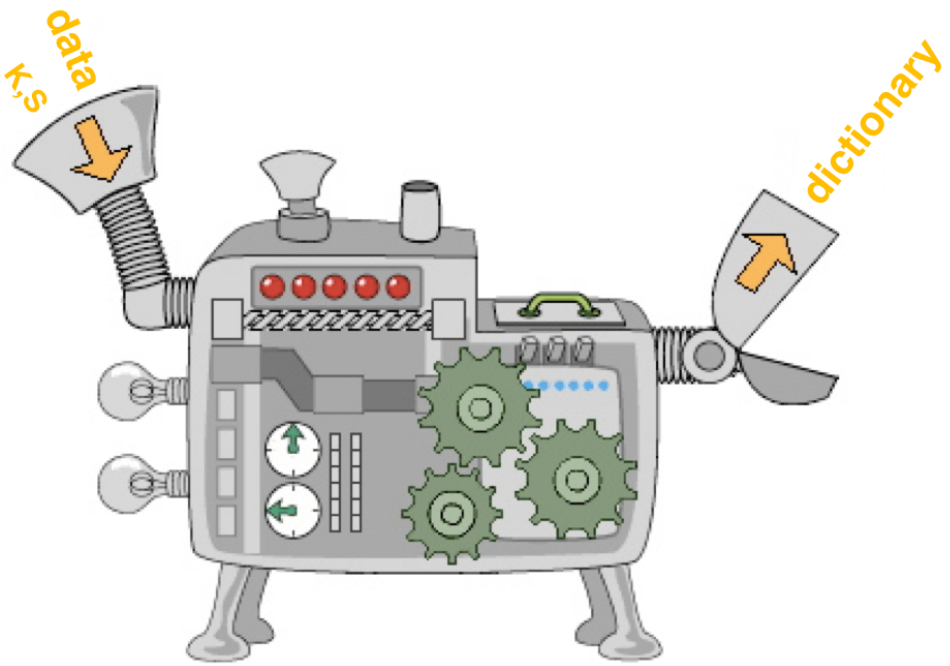


No need for

- intuition
- time (days vs. years)

reminder: dictionary learning

data: $Y = (y_1, \ldots, y_N)$
N vectors $y_n \in \mathbb{R}^d$
d, N large



No need for

- intuition
- time (days vs. years)

Another dictionary learning algorithm????

We have:

- data Y
- a model (Y is S-sparse in a $d \times K$ dictionary Φ)

Another dictionary learning algorithm????

We have:

- data Y
- a model (Y is S-sparse in a $d \times K$ dictionary Φ)

We want:

- an algorithm (fast, cheap)
- guarantees that the algorithm will find Φ.

Another dictionary learning algorithm????

We have:

- data Y
- a model (Y is S-sparse in a $d \times K$ dictionary Φ)

We want:

- an algorithm (fast, cheap)
- guarantees that the algorithm will find Φ.

Promising directions:

- Graph clustering algorithms (not so cheap)
- (Alternating) Optimisation (not so many guarantees)

Another dictionary learning algorithm????

We have:

- data Y
- a model (Y is S-sparse in a $d \times K$ dictionary Φ)

We want:

- an algorithm (fast, cheap)
- guarantees that the algorithm will find Φ.

Promising directions:

- Graph clustering algorithms (not so cheap)
- (Alternating) Optimisation (not so many guarantees)

first: Iterative Thresholding and K signal means (ITKsM)

Let's optimise:

$$\max_{\Psi \in \mathcal{D}} \sum_{n} \max_{|I|=S} \|\Psi_I^\star y_n\|_1 \tag{1}$$

first: Iterative Thresholding and K signal means (ITKsM)

Let's optimise:

$$\max_{\Psi \in \mathcal{D}} \sum_n \max_{|I|=S} \|\Psi_I^\star y_n\|_1 \tag{1}$$

Algorithm (ITKsM one iteration)

Given an input dictionary Ψ and N training signals y_n do:

- *For all n find $I^t_{\Psi,n} = \arg\max_{I:|I|=S} \|\Psi_I^\star y_n\|_1$.*
- *For all k calculate*

$$\bar{\psi}_k = \frac{1}{N} \sum_n y_n \cdot \text{sign}(\langle \psi_k, y_n \rangle) \cdot \chi(I^t_{\Psi,n}, k). \tag{2}$$

- *Output $\bar{\Psi} = (\bar{\psi}_1/\|\bar{\psi}_1\|_2, \ldots, \bar{\psi}_K/\|\bar{\psi}_K\|_2)$.*

ITKsM is

- ridiculously cheap $O(dKN)$ (parallelisable, online version)
- robust to noise, not exact or low sparsity ($S = O(\mu^{-2})$)
- locally convergent (radius $1/\sqrt{\log K}$) for sparsity $S = O(\mu^{-2})$,
- needs only $O(K \log K \varepsilon^{-2})$ samples,

ITKsM is

- ridiculously cheap $O(dKN)$ (parallelisable, online version)
- robust to noise, not exact or low sparsity ($S = O(\mu^{-2})$)
- locally convergent (radius $1/\sqrt{\log K}$) for sparsity $S = O(\mu^{-2})$,
- needs only $O(K \log K \varepsilon^{-2})$ samples,
- but is not globally convergent

ITKsM is

- ridiculously cheap $O(dKN)$ (parallelisable, online version)
- robust to noise, not exact or low sparsity ($S = O(\mu^{-2})$)
- locally convergent (radius $1/\sqrt{\log K}$) for sparsity $S = O(\mu^{-2})$,
- needs only $O(K \log K \varepsilon^{-2})$ samples,
- but is not globally convergent

Algorithm (ITKrM one iteration)

Given an input dictionary Ψ and N training signals y_n do:

- *For all n find* $I^t_{\Psi,n} = \arg\max_{I:|I|=S} \|\Psi_I^\star y_n\|_1$.
- *For all k calculate*

$$\bar{\psi}_k = \sum_{n:k\in I^t_{\Psi,n}} \operatorname{sign}(\langle \psi_k, y_n \rangle) \cdot y_n.$$

- *Output* $\bar{\Psi} = (\bar{\psi}_1/\|\bar{\psi}_1\|_2, \ldots, \bar{\psi}_K/\|\bar{\psi}_K\|_2)$.

ITKsM is

- ridiculously cheap $O(dKN)$ (parallelisable, online version)
- robust to noise, not exact or low sparsity ($S = O(\mu^{-2})$)
- locally convergent (radius $1/\sqrt{\log K}$) for sparsity $S = O(\mu^{-2})$,
- needs only $O(K \log K \varepsilon^{-2})$ samples,
- but is not globally convergent

Algorithm (ITKrM one iteration)

Given an input dictionary Ψ and N training signals y_n do:

- *For all n find* $I^t_{\Psi,n} = \arg\max_{I:|I|=S} \|\Psi_I^\star y_n\|_1$.
- *For all k calculate*

$$\bar{\psi}_k = \sum_{n:k\in I^t_{\Psi,n}} \operatorname{sign}(\langle \psi_k, y_n\rangle) \cdot \left[\mathbb{I} - P(\Psi_{I^t_n}) + P(\psi_k)\right] y_n.$$

- *Output* $\bar{\Psi} = (\bar{\psi}_1/\|\bar{\psi}_1\|_2, \ldots, \bar{\psi}_K/\|\bar{\psi}_K\|_2)$.

first fine-tune the model

Signal model:

- take Φ with $\max_{i \neq j} |\langle \phi_i, \phi_j \rangle| = \mu < 1$.
- draw a positive, decaying, normed sequence c so that a.s.

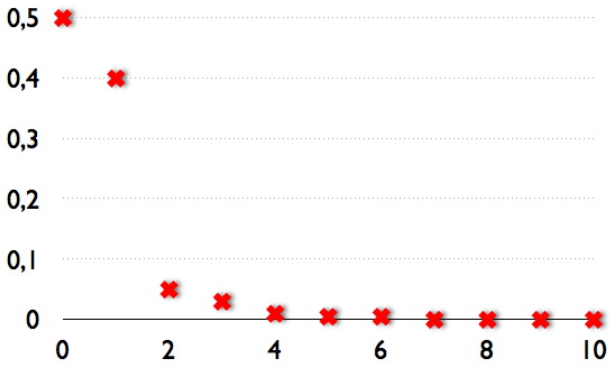
$$c(S) - c(S+1) > \beta_S \quad \text{and} \quad \frac{c(S) - c(S+1)}{c(1)} > \Delta_S.$$

- for a random permutation p, sign sequence σ and subgaussian noise r set

$$y = \frac{\Phi x_{c,p,\sigma} + r}{\sqrt{1 + \|r\|_2^2}}, \quad \text{where} \quad x_{c,p,\sigma}(k) = \sigma(k) c(p(k)). \qquad (3)$$

E.g.
$c_1 = \ldots = c_S = 1/\sqrt{S}$
$c_i = 0, i > S$

then $\beta_S = 1/\sqrt{S}$, $\Delta_S = 1$



a very detailed result

Theorem

Let Φ be a unit norm frame with frame constants $A \leq B$ and coherence μ and assume that the N training signals y_n are generated according to the signal model in (3) with coefficients that are S-sparse with absolute gap β_S and relative gap Δ_S. Assume further that $S \leq \frac{K}{98B}$ and $\varepsilon_\delta := K \exp\left(-\frac{1}{4741\mu^2 S}\right) \leq \frac{1}{24(B+1)}$.

Fix a target error $\bar{\varepsilon} \geq 8\varepsilon_{\mu,\rho}$, with

$$\varepsilon_{\mu,\rho} = \frac{8K^2\sqrt{B+1}}{C_r\gamma_{1,S}} \exp\left(\frac{-\beta_S^2}{98\max\{\mu^2,\rho^2\}}\right), \tag{4}$$

compare (??), and assume that $\bar{\varepsilon} \leq 1 - \gamma_{2,S} + d\rho^2$.
If for the input dictionary Ψ we have

$$d(\Psi,\Phi) \leq \frac{\Delta_S}{\sqrt{98B}\left(\frac{1}{4} + \sqrt{\log\left(\frac{2544K^2(B+1)}{\Delta_S C_r \gamma_{1,S}}\right)}\right)} \quad \text{and} \quad d(\Psi,\Phi) \leq \frac{1}{32\sqrt{S}}, \tag{5}$$

then after $12\lceil\log(\bar{\varepsilon}^{-1})\rceil$ iterations the output dictionary $\tilde{\Psi}$ of ITKrM both in its batch and online version satisfies $d(\bar{\Psi},\Phi) \leq \bar{\varepsilon}$ except with probability

$$60\lceil\log(\bar{\varepsilon}^{-1})\rceil K \exp\left(\frac{-C_r^2\gamma_{1,S}^2 N\bar{\varepsilon}^2}{576K\max\{S,B+1\}\left(\bar{\varepsilon}+1-\gamma_{2,S}+d\rho^2\right)}\right). \tag{6}$$

and its understandable version

Theorem

Assume the number of training samples N scales as $O(K \log K \varepsilon^{-2})$. If $S \leq O(\frac{1}{\ell \mu^2 \log K})$ then with high probability for any starting dictionary Ψ within distance $O(1/\sqrt{S})$ to the generating dictionary Φ, i.e.,

$$\max_k \|\phi_k - \psi_k\|_2 \leq O(1/\sqrt{S}),$$

after $O(\log(\varepsilon^{-1}))$ iterations of ITKM the distance of the output dictionary $\bar{\Psi}$ to the generating dictionary will be smaller than

$$\max_k \|\phi_k - \bar{\psi}_k\|_2 \leq \max\left\{\varepsilon, O\left(K^{2-\ell}\right)\right\}. \tag{7}$$

If the signals are noiseless and exactly S-sparse with $S \leq O(\mu^{-1})$, the right hand side above reduces to ε and the number of necessary training samples reduces to $O(K \log K \varepsilon^{-1})$.

and its understandable version

Theorem

Assume the number of training samples N scales as $O(K \log K \varepsilon^{-2})$. If $S \leq O(\frac{1}{\ell \mu^2 \log K})$ then with high probability for any starting dictionary Ψ within distance $O(1/\sqrt{S})$ to the generating dictionary Φ, i.e.,

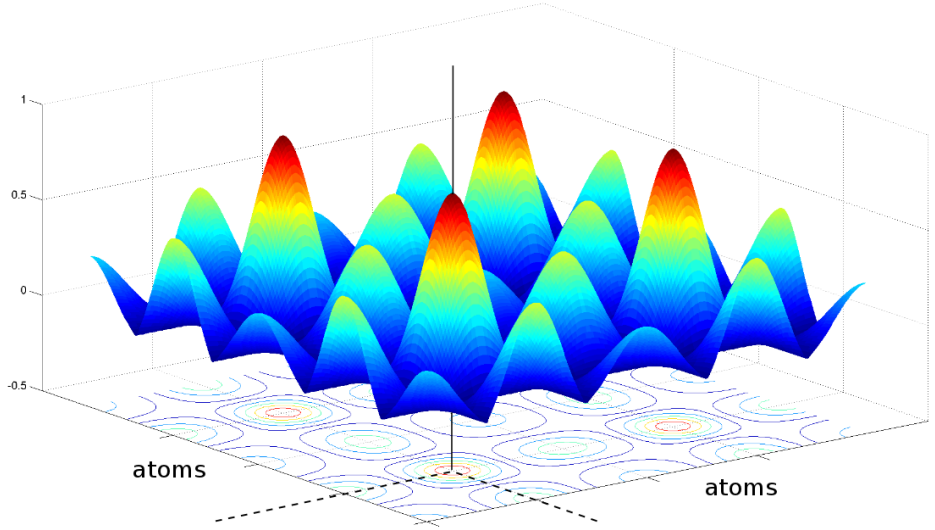
$$\max_k \|\phi_k - \psi_k\|_2 \leq O(1/\sqrt{S}),$$

after $O(\log(\varepsilon^{-1}))$ iterations of ITKM the distance of the output dictionary $\bar{\Psi}$ to the generating dictionary will be smaller than

$$\max_k \|\phi_k - \bar{\psi}_k\|_2 \leq \max\left\{\varepsilon, O\left(K^{2-\ell}\right)\right\}. \tag{7}$$

If the signals are noiseless and exactly S-sparse with $S \leq O(\mu^{-1})$, the right hand side above reduces to ε and the number of necessary training samples reduces to $O(K \log K \varepsilon^{-1})$.

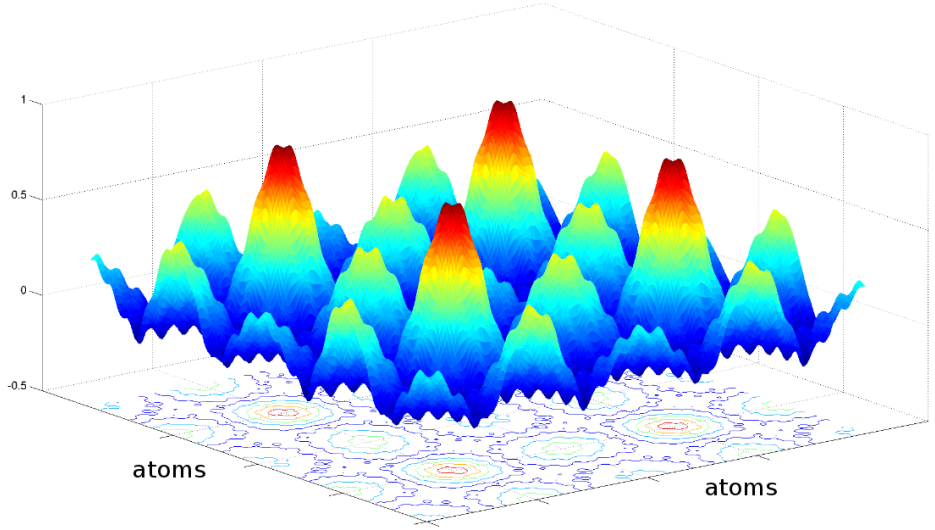
plus some graphical explanations



Answers:

- In expectation there is a local maximum of (1) at/near the generating dictionary.

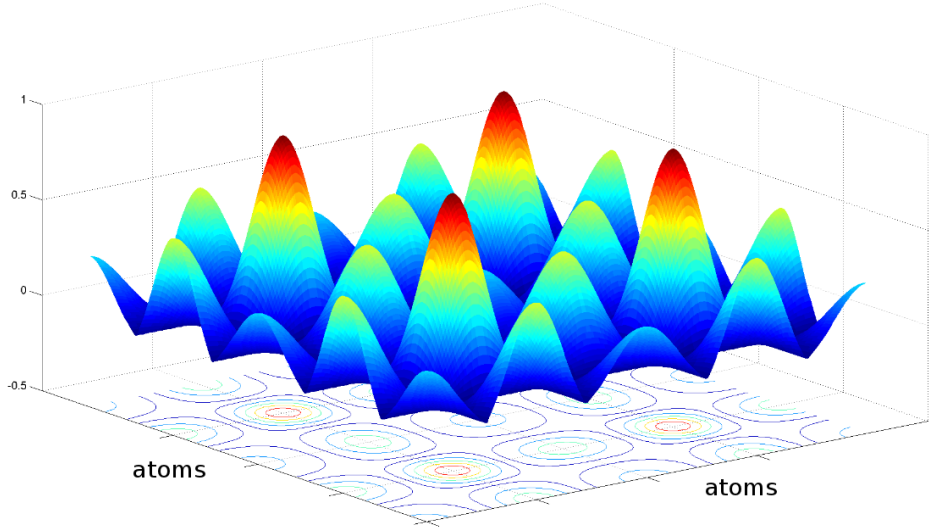
plus some graphical explanations



Answers:

- In expectation there is a local maximum of (1) at/near the generating dictionary.
- For final accuracy ε the ITKMs needs $K \log K \varepsilon^{-2}$ training signals, in the ideal case ITKrM only $K \log K \varepsilon^{-1}$.

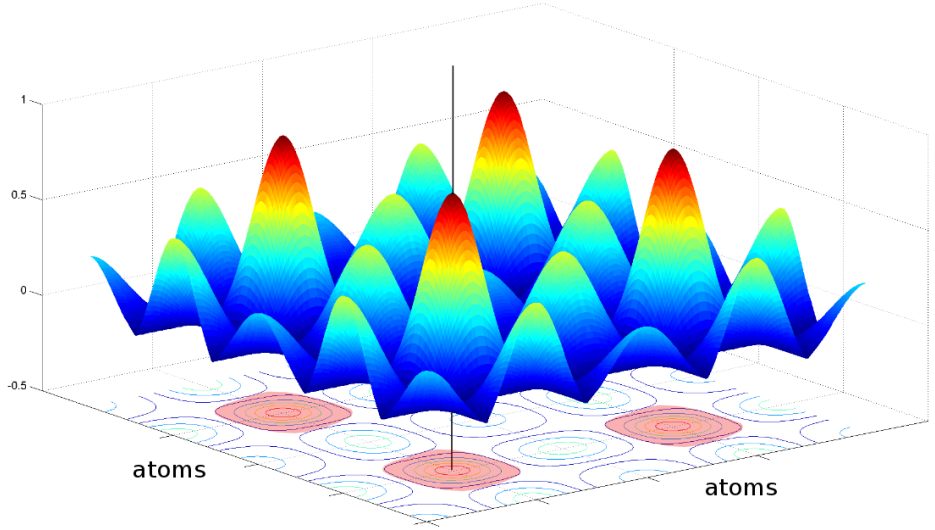
plus some graphical explanations



Answers:

- In expectation there is a local maximum of (1) at/near the generating dictionary.
- For final accuracy ε the ITKMs needs $K \log K \varepsilon^{-2}$ training signals, in the ideal case ITKrM only $K \log K \varepsilon^{-1}$.

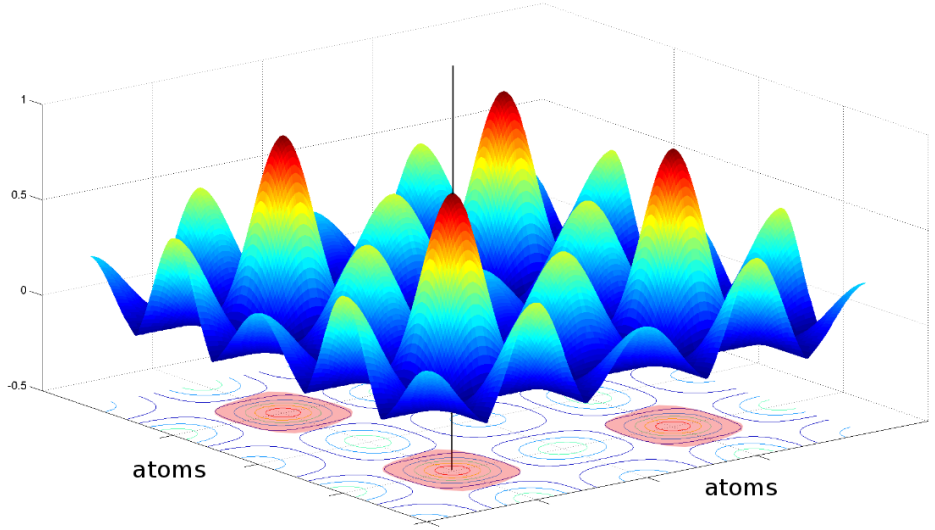
plus some graphical explanations



Answers:

- In expectation there is a local maximum of (1) at/near the generating dictionary.
- For final accuracy ε the ITKMs needs $K \log K \varepsilon^{-2}$ training signals, in the ideal case ITKrM only $K \log K \varepsilon^{-1}$.
- The convergence radius of ITKsM resp. ITKrM is of size at least $1/\sqrt{\log K}$ resp. $1/\sqrt{S}$.

plus some graphical explanations



Answers:

- In expectation there is a local maximum of (1) at/near the generating dictionary.
- For final accuracy ε the ITKMs needs $K \log K \varepsilon^{-2}$ training signals, in the ideal case ITKrM only $K \log K \varepsilon^{-1}$.
- The convergence radius of ITKsM resp. ITKrM is of size at least $1/\sqrt{\log K}$ resp. $1/\sqrt{S}$.
- Experimentally ITKrM shows global convergence properties.

proofs

proofs

and other gory details can be found in

- *Convergence radius and sample complexity of ITKM algorithms for dictionary learning*, arXiv:1503.07027
- *Identification of overcomplete dictionaries*, to appear Journal of Machine Learning Research, (arXiv: 1401.6354).

background and complementary literature


R. Rubinstein, A. Bruckstein, and M. Elad.
Dictionaries for sparse representation modeling.
Proceedings of the IEEE, 98(6):1045–1057, 2010.


K. Schnass.
A personal introduction to theoretical dictionary learning.
Internationale Mathematische Nachrichten, 228:5–15, 2015.


S. Arora, R. Ge, T. Ma, and A. Moitra.
Simple, efficient, and neural algorithms for sparse coding.
arXiv:1503.00778, 2015.


J. Sun, Q. Qu, and J. Wright.
Complete dictionary recovery over the sphere.
arXiv:1504.06785, 2015.

background and complementary literature


R. Rubinstein, A. Bruckstein, and M. Elad.
Dictionaries for sparse representation modeling.
Proceedings of the IEEE, 98(6):1045–1057, 2010.


K. Schnass.
A personal introduction to theoretical dictionary learning.
Internationale Mathematische Nachrichten, 228:5–15, 2015.


S. Arora, R. Ge, T. Ma, and A. Moitra.
Simple, efficient, and neural algorithms for sparse coding.
arXiv:1503.00778, 2015.


J. Sun, Q. Qu, and J. Wright.
Complete dictionary recovery over the sphere.
arXiv:1504.06785, 2015.

Today at 17.40!

the last slide

Questions

Comments

Thanks for your attention!!

an offer you cannot refuse