Online Convolutional Dictionary Learning

Jialin Liu[a] Cristina Garcia-Cardona[b]
Brendt Wohlberg[b] Wotao Yin[a]

[a]Department of Mathematics, UCLA
[b]Los Alamos National Laboratory

ICIP 2017, Beijing

Convolutional Sparse Coding

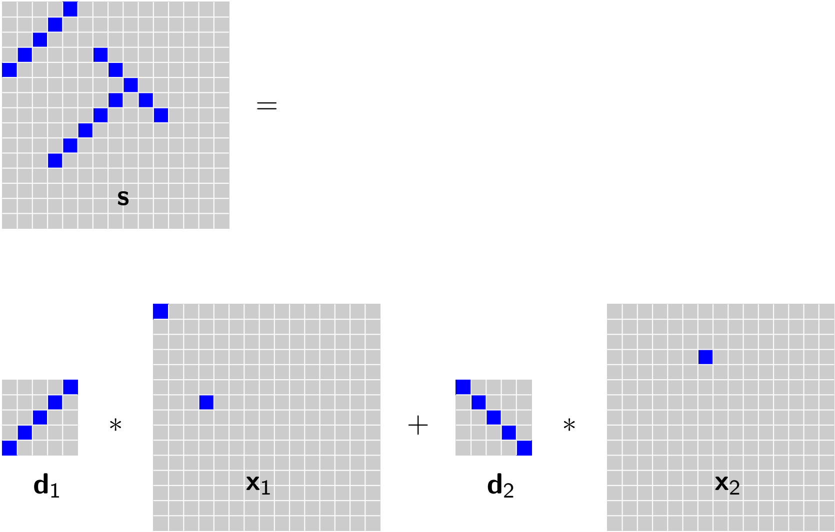
- Signal $\mathbf{s} \in \mathbb{R}^N$.
- Dictionary $\mathbf{d}$ and its kernels $\mathbf{d} = (\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_M)^T, \mathbf{d}_m \in \mathbb{R}^D$.
- Sparse coefficient maps $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M)^T, \mathbf{x}_m \in \mathbb{R}^N$.
- The model is

$$\mathbf{s} \approx \sum_{m=1}^{M} \mathbf{d}_m * \mathbf{x}_m.$$

- (Zeiler et al. 2010) Given $\mathbf{s}$ and $\mathbf{d}$, convolutional basis pursuit denoising (CBPDN):

$$\min_{\mathbf{x}} \ell(\mathbf{d}, \mathbf{x}; \mathbf{s}) = \min_{\{\mathbf{x}_m\}} \frac{1}{2} \left\| \sum_{m=1}^{M} \mathbf{d}_m * \mathbf{x}_m - \mathbf{s} \right\|_2^2 + \lambda \sum_{m=1}^{M} \|\mathbf{x}_m\|_1 .$$

An example of Convolutional Sparse Coding



Applications of CSC

- Image super-resolution (Gu et al. 2015)
- Trajectory Reconstruction (Zhu and Lucey 2015)
- Denoising (Wohlberg 2016)
- Image Decomposition (Zhang and Patel 2016)
- ...

Convolutional Dictionary Learning

- Given training signals $\{\mathbf{s}_k\}$,
  convolutional dictionary learning (CDL):

$$\min_{\mathbf{d}\in C, \{\mathbf{x}_k\}} \sum_{k=1}^{K} \ell(\mathbf{d}, \mathbf{x}_k; \mathbf{s}_k) \ .$$

- Conventional methods: batch learning.
  Alternative update $\mathbf{d}$ and $\{\mathbf{x}_k\}$.
- Single step complexity and memory usage[1]: $\mathcal{O}(KMN)$.
  Typical value: $K = 40, M = 64, N = 256 \times 256$.
  Total time: 15 hours ; memory: 7.5 GB.

[1] [Šorel and Šroubek 2016] and [Garcia-Cardona and Wohlberg 2017]

Surrogate Function Approach

- A statistic estimator:

$$\mathbf{d}^{(t)} = \arg\min_{\mathbf{d}\in C} \left\{ \min_{\mathbf{x}} \ell(\mathbf{d}, \mathbf{x}, \mathbf{s}^{(1)}) + \cdots + \min_{\mathbf{x}} \ell(\mathbf{d}, \mathbf{x}, \mathbf{s}^{(t)}) \right\} .$$

- An online estimator (Mairal et al. 2009):

$$\mathbf{x}^{(t)} = \arg\min_{\mathbf{x}} \ell(\mathbf{d}^{(t-1)}, \mathbf{x}; \mathbf{s}^{(t)}).$$

$$\mathbf{d}^{(t)} = \arg\min_{\mathbf{d}\in C} \left\{ \underbrace{\ell(\mathbf{d}, \mathbf{x}^{(1)}, \mathbf{s}^{(1)}) + \cdots + \ell(\mathbf{d}, \mathbf{x}^{(t)}, \mathbf{s}^{(t)})}_{\text{surrogate function } \mathcal{F}^{(t)}(\mathbf{d})} \right\} .$$

- $\mathcal{F}^{(t)}$ is quadratic on $\mathbf{d}$.
  Keeping Hessian matrix and a vector in memory.
  Constant computational cost.

Solving subproblem

To compute $\mathcal{F}^{(t)}(\mathbf{d})$,

- Spacial domain:
  Flops: $\mathcal{O}(M^2D^2N)$; memory usage: $\mathcal{O}(M^2D^2)$.
- Frequency domain:
  Flops: $\mathcal{O}(M^2N)$; memory usage: $\mathcal{O}(M^2N)$.

To solve $\mathbf{d}^{(t)} \leftarrow \arg\min_{\mathbf{d}\in C} \mathcal{F}^{(t)}(\mathbf{d})$,

- Degraux et al. 2017 uses block-coordinate gradient descent.
  Flops: $\mathcal{O}(1/\epsilon)$.
- Wang et al. 2017 uses Augmented Lagrangian method + iterated Sherman-Morrison. Flops:$\mathcal{O}(1/\epsilon)$..
- Our work uses FISTA. Flops: $\mathcal{O}(1/\sqrt{\epsilon})$.

Frequency-domain FISTA

Frequency domain FISTA:

- Start with $\mathbf{g}^0 = \mathbf{g}^0_{\text{aux}} = \mathbf{d}^{(t-1)}$.
- Do

$$
\begin{aligned}
\hat{\mathbf{g}}^j_{\text{aux}} =& \text{FFT}(\mathbf{g}^j_{\text{aux}}) \\
\mathbf{g}^{j+1} =& \text{proj}_C\left(\text{IFFT}\left(\hat{\mathbf{g}}^j_{\text{aux}} - \eta \nabla \hat{\mathcal{F}}^{(t)}(\hat{\mathbf{g}}^j_{\text{aux}})\right)\right) . \\
\gamma^{j+1} =& \left(1 + \sqrt{1 + 4(\gamma^j)^2}\right)/2 , \\
\mathbf{g}^{j+1}_{\text{aux}} =& \mathbf{g}^{j+1} + \frac{\gamma^j - 1}{\gamma^{j+1}}(\mathbf{g}^{j+1} - \mathbf{g}^j) .
\end{aligned}
$$

- $\mathbf{d}^{(t)} \leftarrow$ the last $\mathbf{g}^j$.

Technique I - forgetting factor

Weighted loss function:

$$\mathbf{d}^{(t)} = \arg\min_{\mathbf{d}\in C} \left\{ \sum_{\tau=1}^{t} w^{\tau} \ell(\mathbf{d}, \mathbf{x}^{(\tau)}, \mathbf{s}^{(\tau)}) \right\},$$

where the weight is:

$$w^{\tau} = (\tau/t)^{p}, \quad p \geq 0.$$

Proposition (Weighted central limit theorem)

Suppose $Z_{\tau} \overset{i.i.d}{\sim} P_Z(z)$, *with a compact support, expectation* μ, *and variance* σ^2. *Define the approximation of* Z*:* $\hat{Z}^t \triangleq \frac{1}{\sum_{\tau=1}^{t} w^{\tau}} \sum_{\tau=1}^{t} w^{\tau} Z_{\tau}$. *Then, we have*

$$\sqrt{t}(\hat{Z}^t - \mu) \xrightarrow{d} N\left(0, \frac{p+1}{\sqrt{2p+1}}\sigma\right), \quad \text{as } t \to \infty.$$

Technique II - stopping of FISTA

$$\left\|\mathbf{d} - \mathrm{Proj}_C\left(\mathbf{d} - \eta\nabla\mathcal{F}^{(t)}(\mathbf{d})\right)\right\| \leq \tau_0/(1+\alpha t)\ .$$

Proposition (Convergence of FPR implies convergence of iterates)

Let $(\mathbf{d}^*)^{(t)}$ be the exact minimizer of the t^{th} subproblem:

$$(\mathbf{d}^*)^{(t)} = \arg\min_{\mathbf{d}\in C} \mathcal{F}^{(t)}(\mathbf{d})\ .$$

Let $\mathbf{d}^{(t)}$ be the solution obtained with the above stopping condition. Then, we have

$$\left\|\mathbf{d}^{(t)} - (\mathbf{d}^*)^{(t)}\right\| \leq \mathcal{O}\left(t^{-1}\right)\ .$$

With the two propositions, we prove the convergence of the whole algorithm.

Technique III - image splitting

- Memory cost $\mathcal{O}(M^2N)$ is still large. To reduce N:

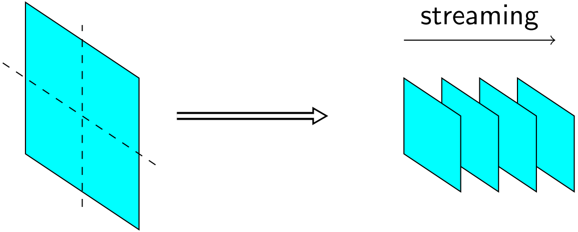


Figure: An example: $N = 256 \times 256 \to \tilde{N} = 128 \times 128$

- Boundary issue:
  $\tilde{N}$ should be at least twice D in each dimension.
  For 2D images, $\tilde{N} \geq 2^2 D$.
- In our experiment, we take $D = 12 \times 12$, $\tilde{N} = 64 \times 64$.

Online Algorithm II - Frequency-domain SGD

- Recall the CDL problem:

$$\min_{\mathbf{d}\in C} \mathbb{E}_{\mathbf{s}} \Big\{ \overbrace{\min_{\mathbf{x}} \ell(\mathbf{d}, \mathbf{x}; \mathbf{s})}^{f(\mathbf{d};\mathbf{s})} \Big\}.$$

- Projected Stochastic Gradient Descent (SGD):

$$\mathbf{d}^{(t)} = \mathrm{Proj}_C\left(\mathbf{d}^{(t-1)} - \eta^{(t)} \nabla f(\mathbf{d}^{(t-1)}; \mathbf{s}^{(t)})\right) .$$

- Frequency domain SGD:

$$\hat{\mathbf{d}}^{(t)} = \mathrm{Proj}_C\left(\mathrm{IFFT}\left(\hat{\mathbf{d}}^{(t-1)} - \eta^{(t)} \nabla \hat{f}(\hat{\mathbf{d}}^{(t-1)}; \hat{\mathbf{s}}^{(t)})\right)\right).$$

Learning from incomplete images

- Masked CDL:

$$\min_{\mathbf{d}\in C} \mathbb{E}_{\mathbf{s}}[f_{\text{mask}}(\mathbf{d};\mathbf{s})] \ ,$$

where f_{mask} is

$$f_{\text{mask}}(\mathbf{d};\mathbf{s}) \triangleq \min_{\{\mathbf{x}_m\}} \frac{1}{2}\left\| W \odot \left(\sum_{m=1}^{M} \mathbf{d}_m * \mathbf{x}_m - \mathbf{s}\right)\right\|_2^2 + \lambda \sum_{m=1}^{M} \|\mathbf{x}_m\|_1 \ .$$

- W is a *masking matrix*, usually $\{0,1\}$-valued. Masking unknown or unreliable pixels.
- Online algorithm for masked CDL:

$$\mathbf{d}^{(t)} = \text{Proj}_{C_{\text{PN}}}\left(\text{IFFT}\left(\hat{\mathbf{d}}^{(t-1)} - \eta^{(t)}\nabla \hat{f}_{\text{mask}}(\hat{\mathbf{d}}^{(t-1)};\hat{\mathbf{s}}^{(t)})\right)\right).$$

Numerical Results

- Platform: MATLAB R2016a; 2 Intel Xeon(R) X5650 CPUs @ 2.67GHz.
- Dictionary size: $12 \times 12 \times 64$
- Signal size: 256×256.
- Dataset: MIRFlickr25k. (Huiskes et al. 2010)
  40 training images and 20 testing images.

Comparison: Convergence Speed

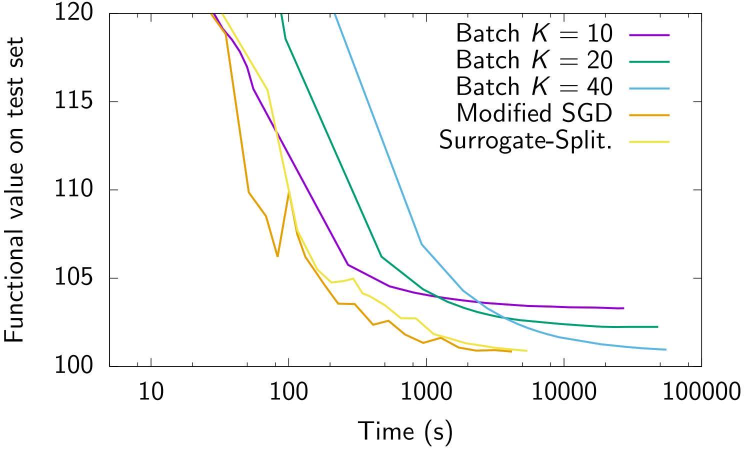


Figure: Convergence speed comparison on the clean data set.

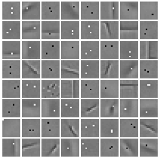
Comparison: Memory Usage

Scheme	Memory (MB)
Batch ($K = 10$)	1959.58
Batch ($K = 20$)	3887.08
Batch ($K = 40$)	7742.08
Surrogate-Split	158.11
Modified SGD	154.84

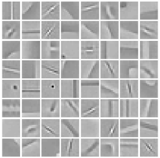
Table: Memory Usage Comparison in Megabytes.

Learning from noisy images

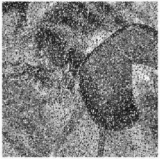
(a) One of the training images. (10% positions noised)



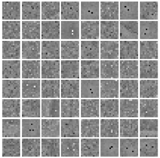
(b) Results by SGD: some valid features.



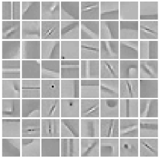
(c) Results by masked SGD: clean features learned.



(d) One of the training images. (30% positions noised)



(e) Results by SGD: almost no valid features.



(f) Results by masked SGD: clean features learned.

Comparison with batch methods

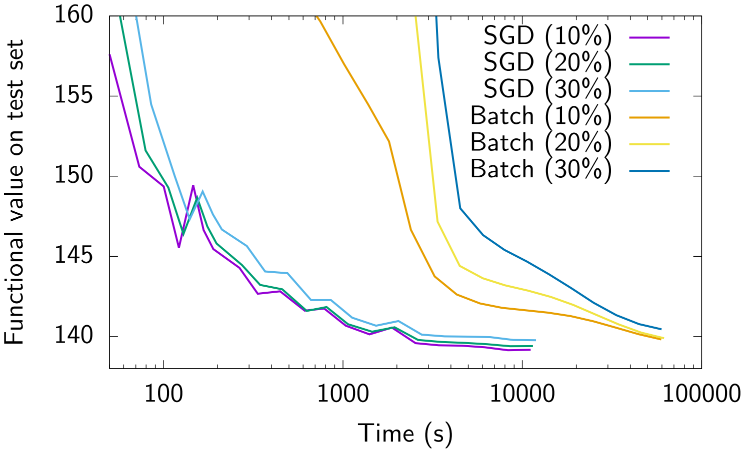


Figure: Comparison on masked CDL problem.

Conclusions

- We have proposed two efficient online convolutional dictionary learning methods. Both of them have theoretical convergence guarantee and show good performance on both time and memory usage.
- Frequency SGD shows better performance in time and memory usage, and requires fewer parameters to tune.
- Frequency SGD can be extended to masked CDL, which learns dictionaries from imcomplete images.
- See `arXiv:1709.00106` for details.
- Implementations of all of these algorithms will be made available as part of the SPORCO software library `http://purl.org/brendt/software/sporco`

References I

- Degraux, Kevin, Ulugbek S Kamilov, Petros T Boufounos, and Dehong Liu (2017). "Online Convolutional Dictionary Learning for Multimodal Imaging". In: *arXiv preprint arXiv:1706.04256*.
- Garcia-Cardona, Cristina and Brendt Wohlberg (2017). "Subproblem coupling in convolutional dictionary learning". In: *Proceedings of IEEE International Conference on Image Processing (ICIP)*.
- Gu, Shuhang et al. (2015). "Convolutional sparse coding for image super-resolution". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1823–1831.
- Huiskes, Mark J, Bart Thomee, and Michael S Lew (2010). "New trends and ideas in visual concept detection: the MIR flickr retrieval evaluation initiative". In: *Proceedings of the international conference on Multimedia information retrieval*. ACM, pp. 527–536.
- Mairal, Julien, Francis Bach, Jean Ponce, and Guillermo Sapiro (2009). "Online dictionary learning for sparse coding". In: *Proceedings of the 26th annual international conference on machine learning*. ACM, pp. 689–696.

References II

- Šorel, Michal and Filip Šroubek (2016). "Fast convolutional sparse coding using matrix inversion lemma". In: *Digital Signal Processing* 55, pp. 44–51.
- Wang, Yaqing, Quanming Yao, James T Kwok, and Lionel M Ni (2017). "Online convolutional sparse coding". In: *arXiv preprint arXiv:1706.06972*.
- Wohlberg, Brendt (2016). "Convolutional sparse representations as an image model for impulse noise restoration". In: *Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), 2016 IEEE 12th*. IEEE, pp. 1–5.
- Zeiler, Matthew D, Dilip Krishnan, Graham W Taylor, and Rob Fergus (2010). "Deconvolutional networks". In: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, pp. 2528–2535.
- Zhang, He and Vishal M Patel (2016). "Convolutional Sparse Coding-based Image Decomposition." In: *BMVC*.
- Zhu, Yingying and Simon Lucey (2015). "Convolutional sparse coding for trajectory reconstruction". In: *IEEE transactions on pattern analysis and machine intelligence* 37.3, pp. 529–540.

Thanks for listening !