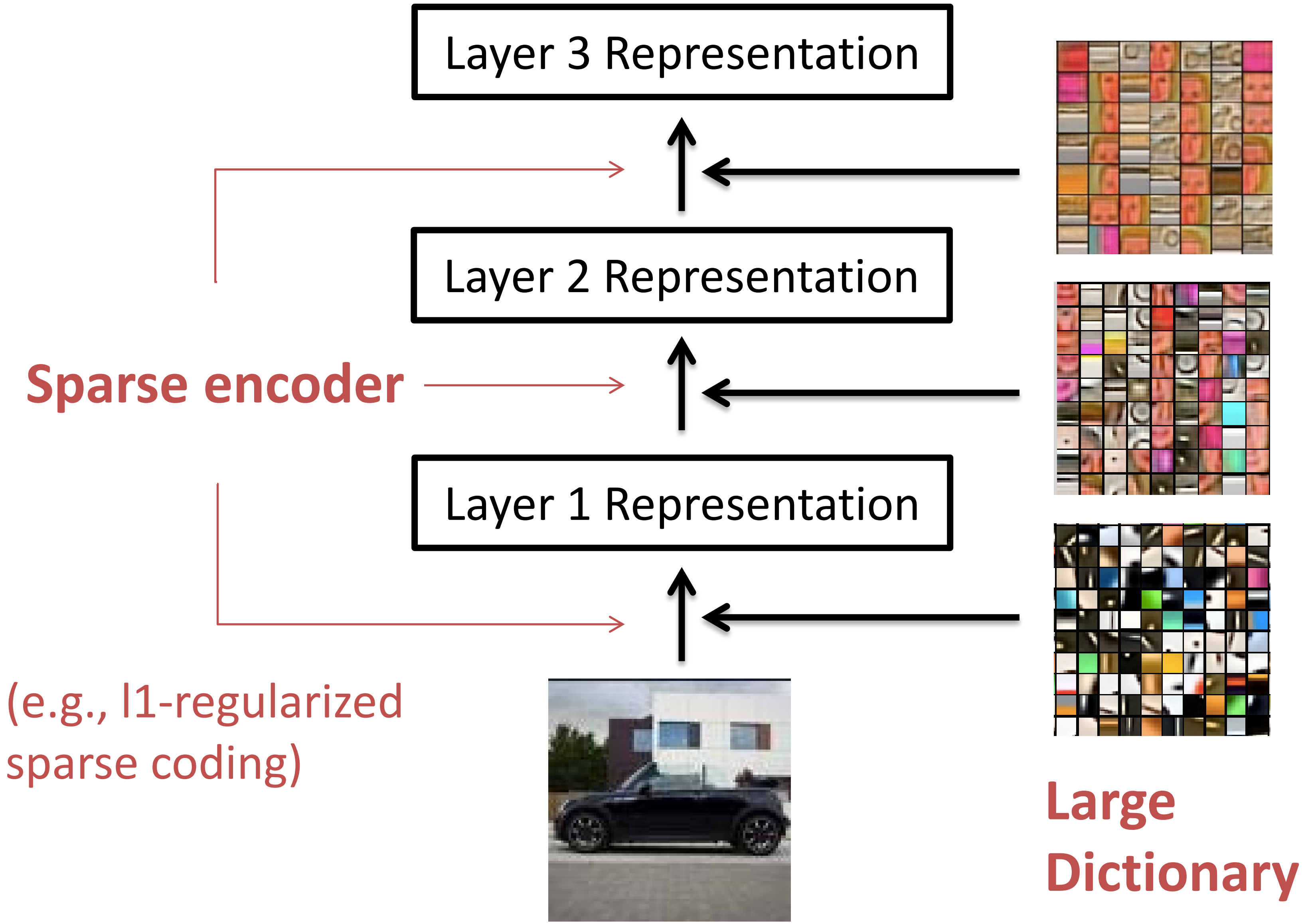
Stable and Efficient Representation Learning with Nonnegativity Constraints

Tsung-Han Lin and H.T. Kung

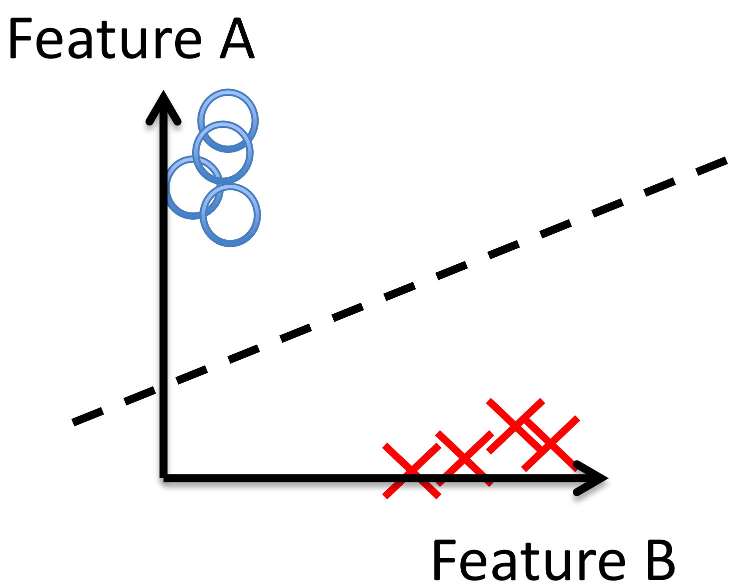
HARVARD
School of Engineering
and Applied Sciences

Unsupervised Representation Learning



Why Sparse Representations?

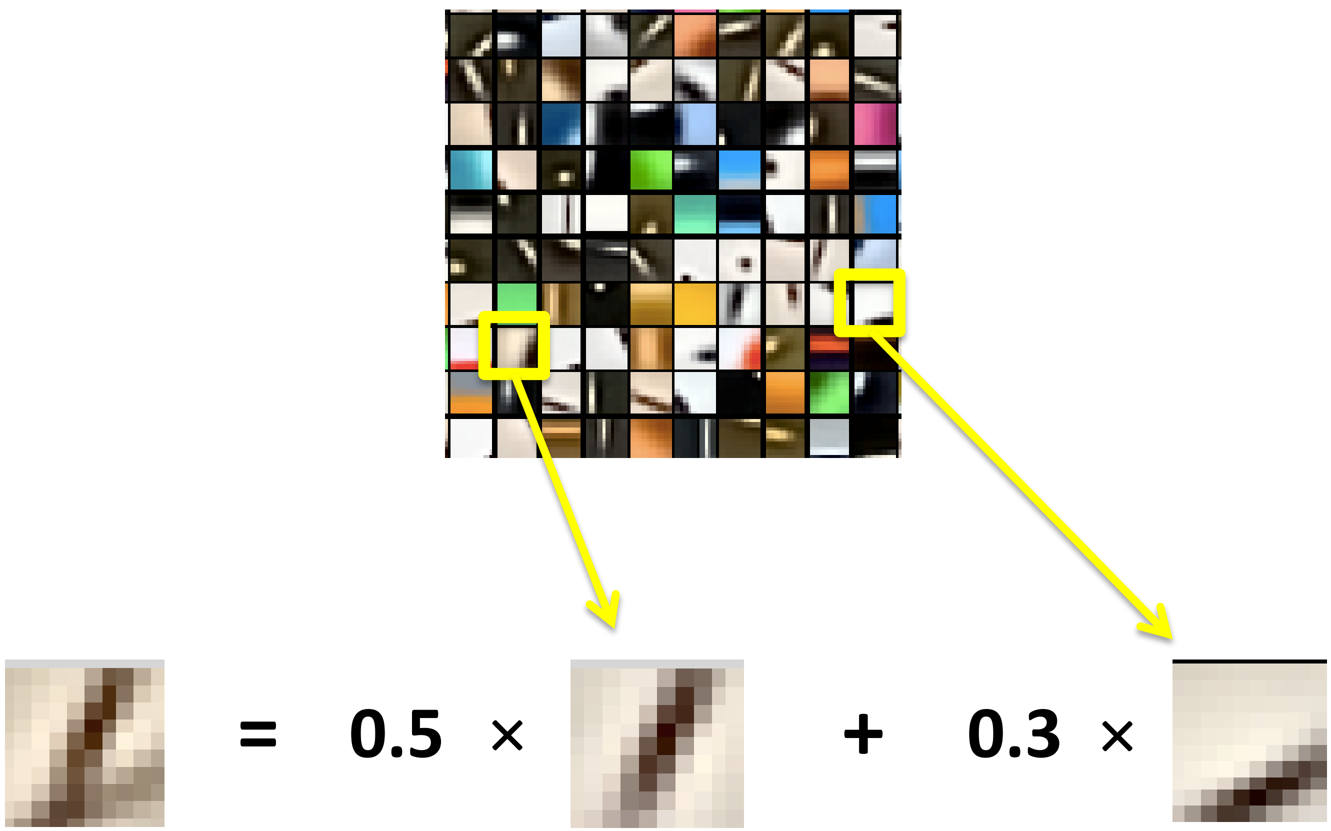
- Prior knowledge is better encoded into sparse representations
  - Data is explained by only a few underlying factors
  - Representations are more linearly separable



<u>Simplifies supervised classifier training</u>: sparse representations work well even when labeled samples are few

Computing Sparse Representations

Sparse approximation: $\min_{\mathbf{z}} \|\mathbf{x} - \mathbf{D}\mathbf{z}\|_2 \;\text{ s.t. }\; \|\mathbf{z}\|_0 \leq k$



Computing Sparse Representations

Sparse approximation: $\min_{\mathbf{z}} \|\mathbf{x} - \mathbf{D}\mathbf{z}\|_2 \ \text{s.t.} \ \|\mathbf{z}\|_0 \le k$

- <u>L1 relaxation approach</u>: good classification accuracy, but computation is expensive
- <u>Greedy approach (e.g., orthogonal matching pursuit)</u>: fast, but yields suboptimal classification accuracy

CIFAR-10 classification with single-layer architecture

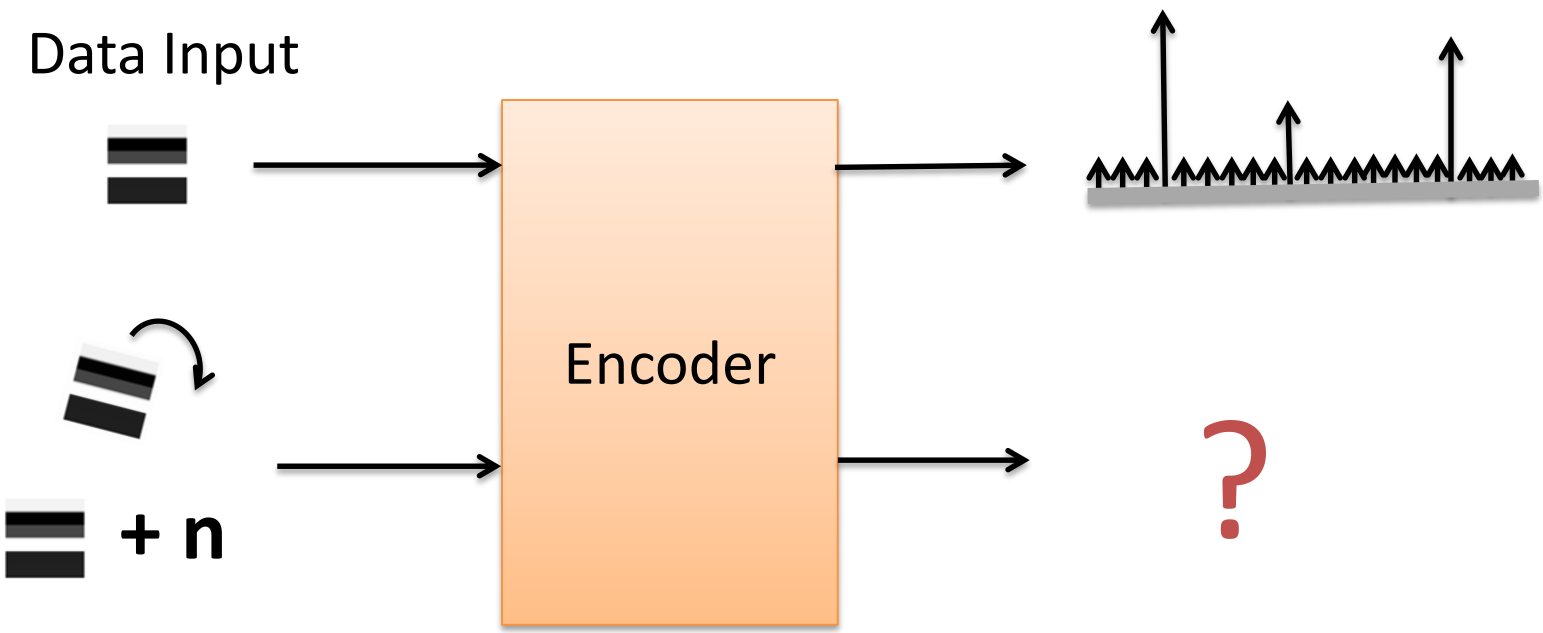
	L1-regularized	**OMP**
Classification accuracy (%)	78.7	76.0

[Coates 2011]

Major Findings

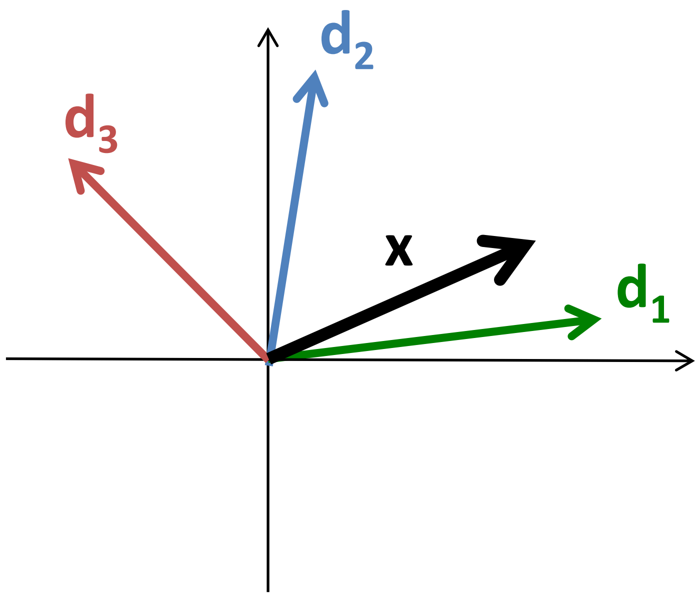
- Weak **stability** is the key to OMP's suboptimal performance
- By allowing only **additive features** (via nonnegativity constraints), classification with OMP delivers higher accuracy by large margins
- **Competitive** classification accuracy with deep neural networks

Stability of Representations



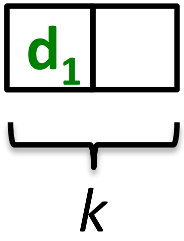
Orthogonal Matching Pursuit (OMP)

Select k atoms from a dictionary D that minimize $|x-Dz|$



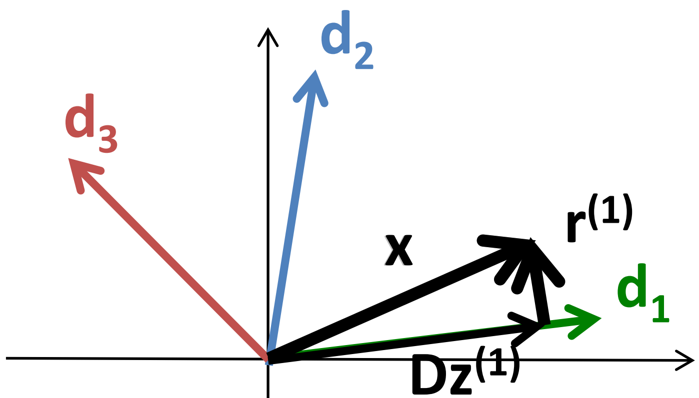
Select the atom that has the largest **correlation** with the residual

Support set

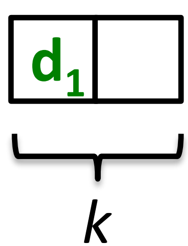


Orthogonal Matching Pursuit (OMP)

Select k atoms from a dictionary D that minimize $|x-Dz|$



Support set



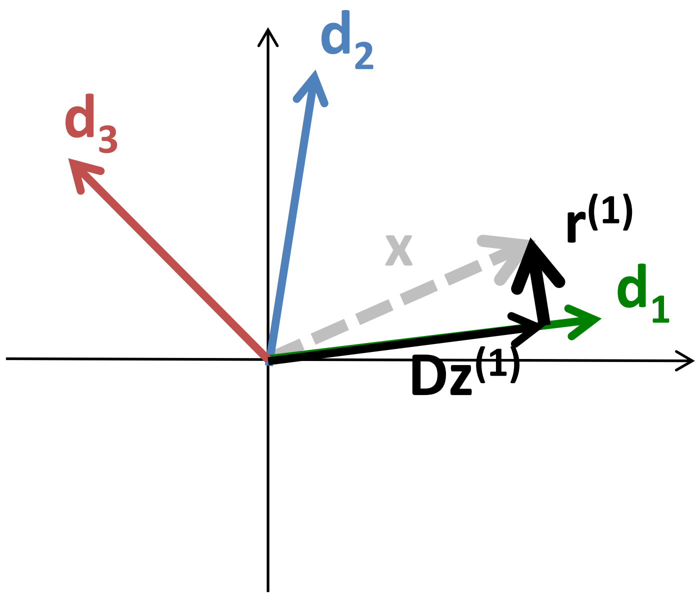
Select the atom that has the largest **correlation** with the residual

Estimate the coefficients of the selected atoms by **least squares**

Update the **residual** using current estimate

Orthogonal Matching Pursuit (OMP)

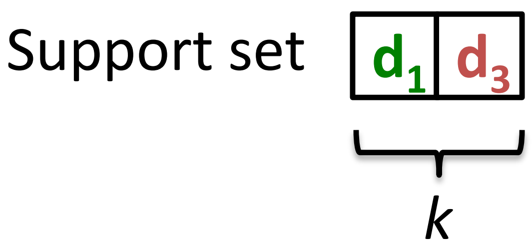
Select k atoms from a dictionary D that minimize $|x-Dz|$

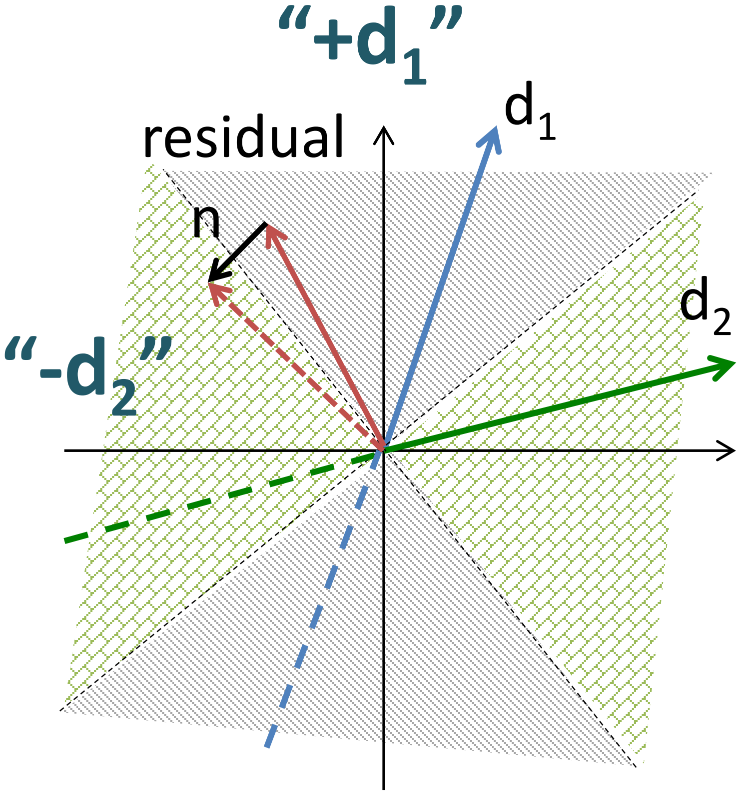


Select the atom that has the largest **correlation** with the residual

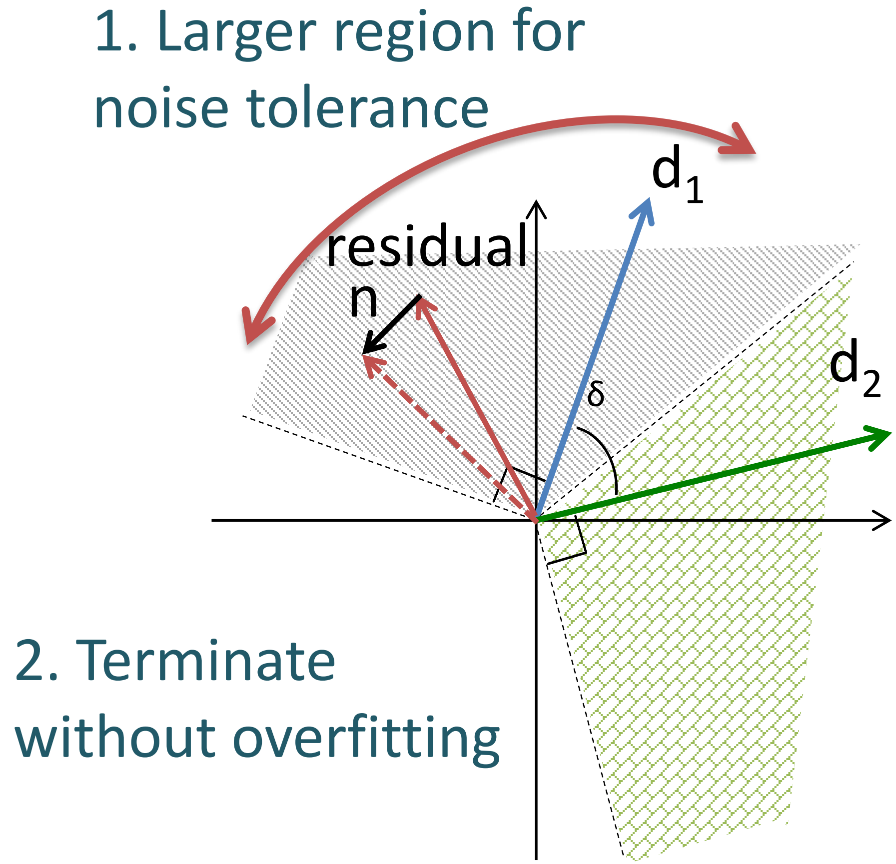
Estimate the coefficients of the selected atoms by **least squares**

Update the **residual** using current estimate





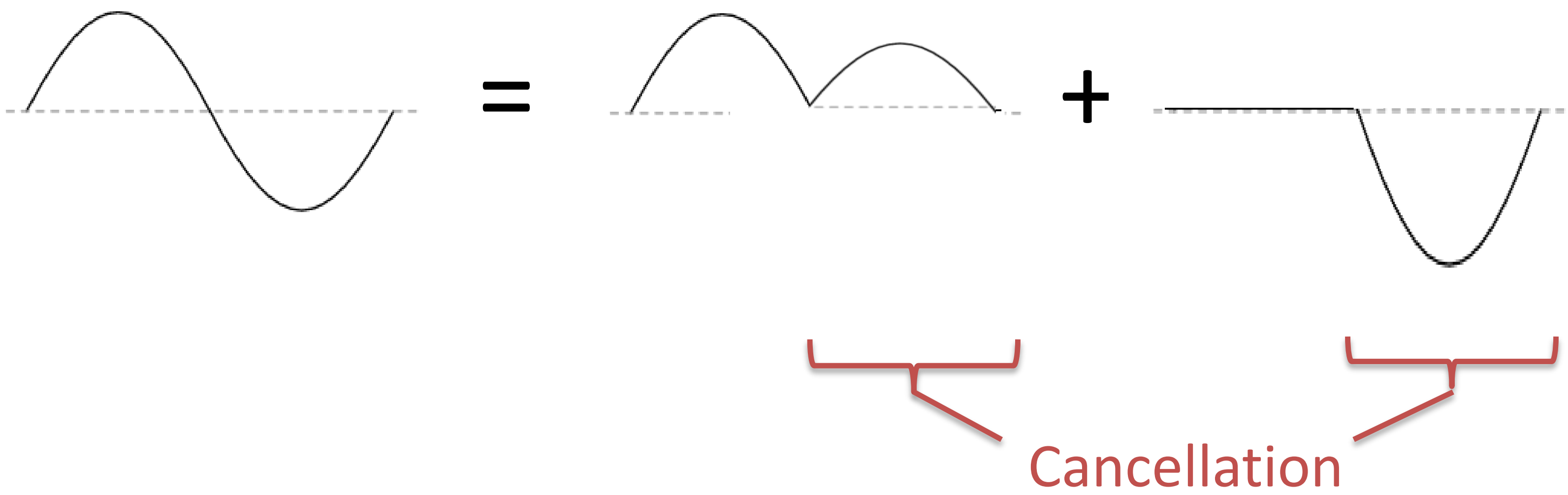
OMP



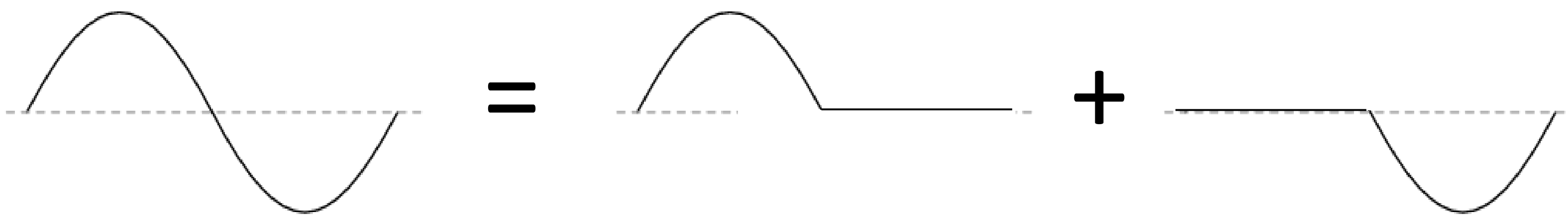
Nonnegative OMP

Use only additive features by constraining the atoms and coefficients to be nonnegative

Allowing Only Additive Features

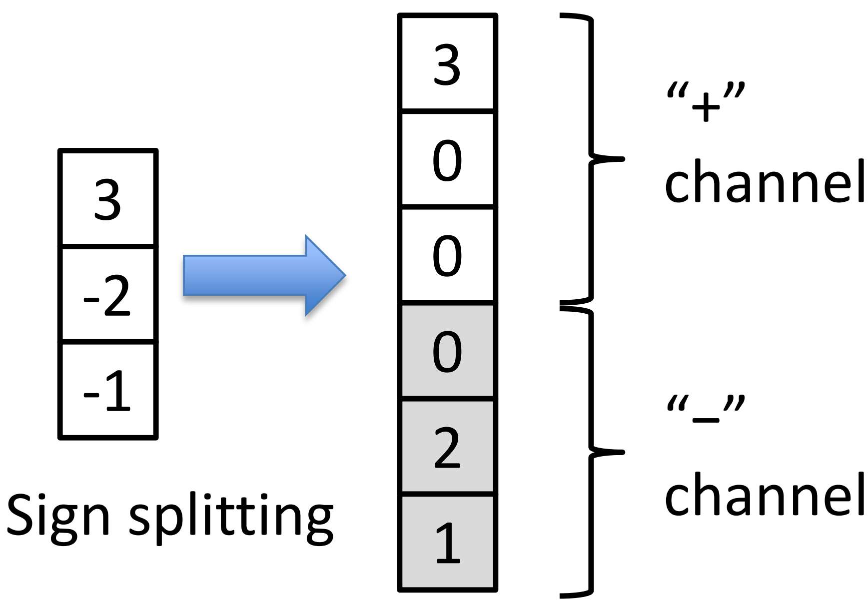


Allowing Only Additive Features



Enforce nonnegativity to eliminate cancellation

<u>On input</u>:



<u>On dictionary</u>:

- Any nonnegative sparse coding algorithms
- We use spherical K-means

<u>On representation</u>:

- Encode with nonnegative OMP (NOMP)

Evaluate the Stability of Representations

Feature dictionary learned from image datasets

Grating A  → Representation A

Rotate by some small angle δ

Encode by OMP/NOMP

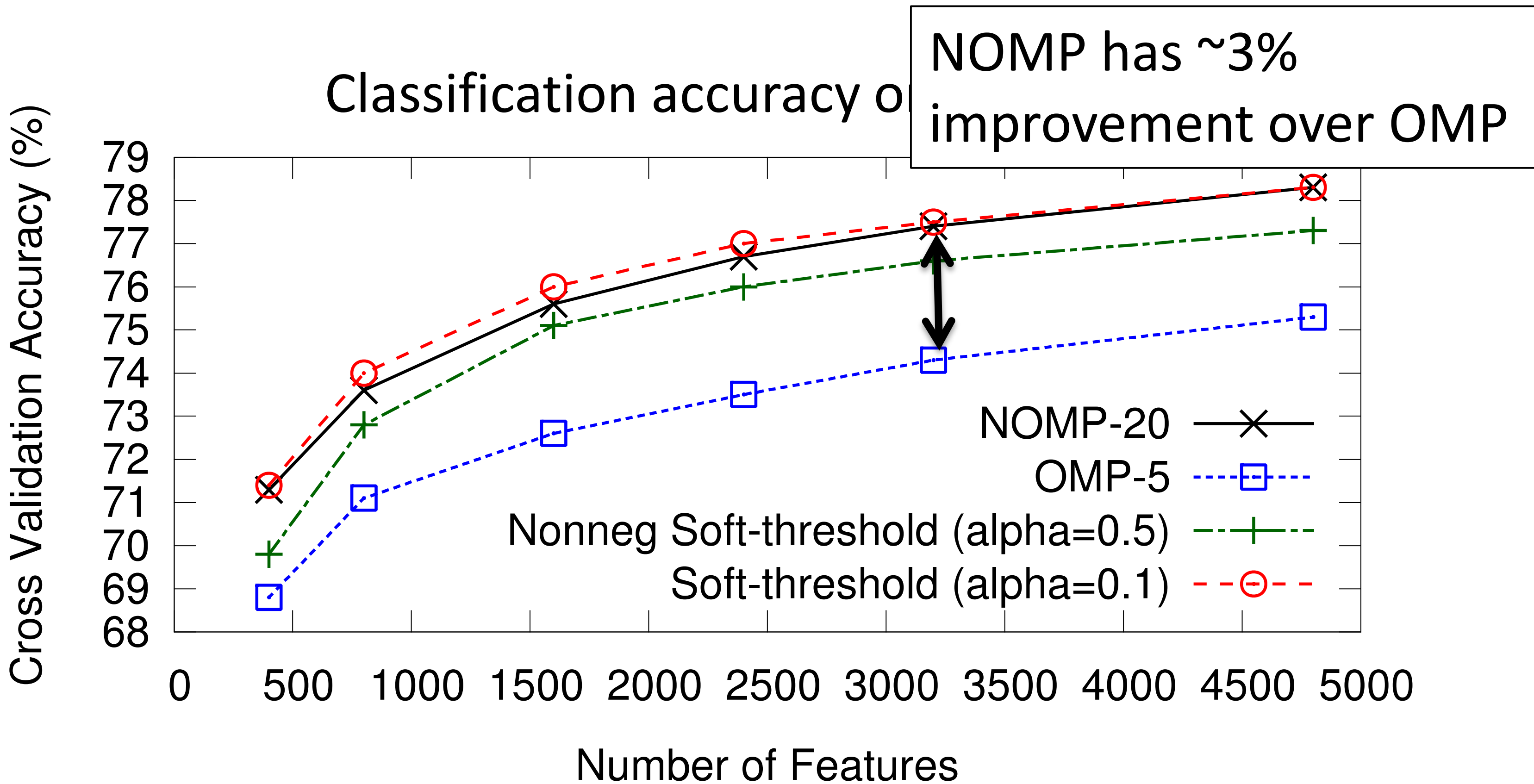
Measure change by their correlation

Grating B  → Representation B

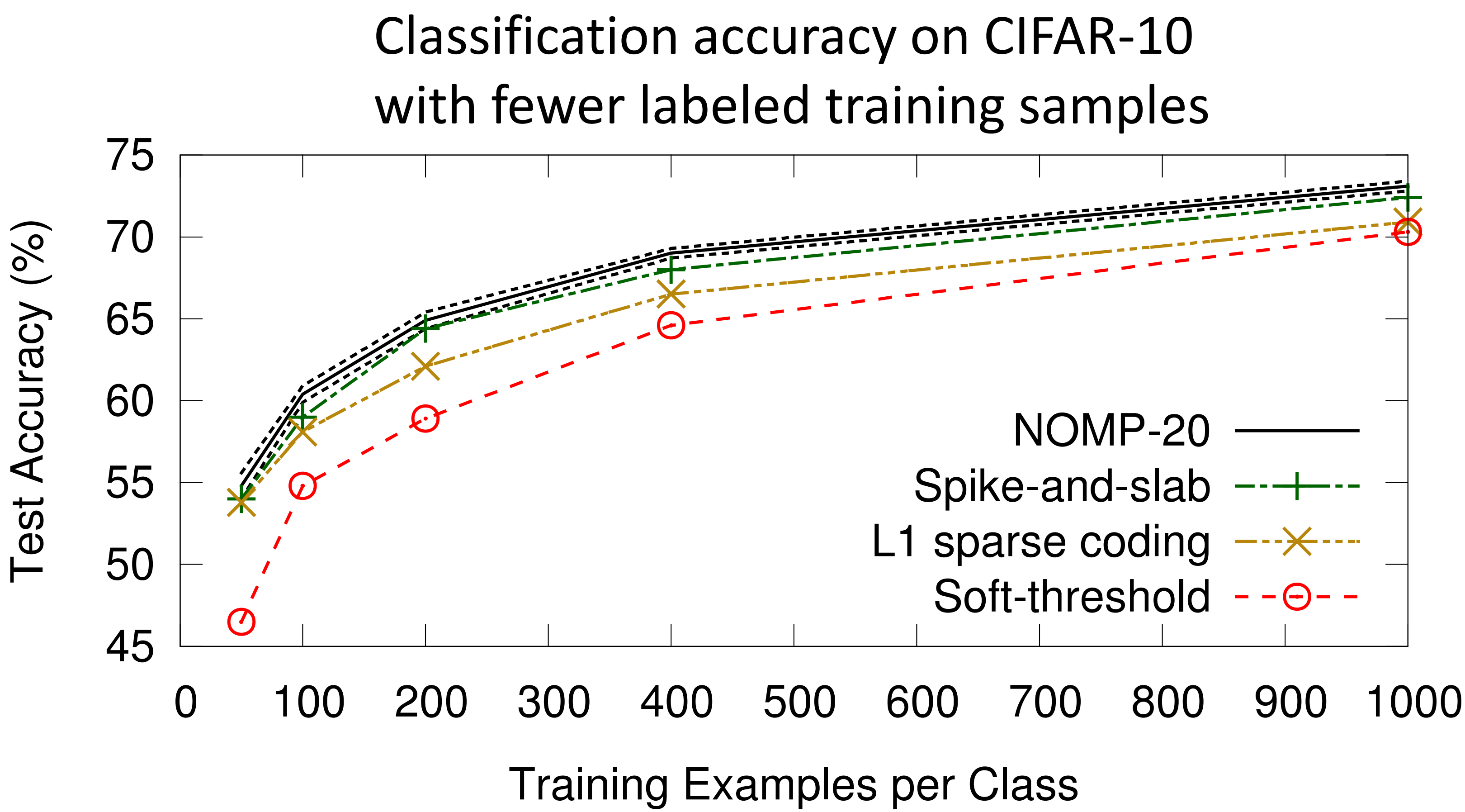
Correlation between representation A and B

Encoder	Rotation angle δ				
	0	0.01π	0.02π	0.03π	0.04π
OMP	1	0.71	0.54	0.43	0.34
NOMP	**1**	**0.92**	**0.80**	**0.68**	**0.57**

Classification: NOMP vs OMP



NOMP Outperforms When Fewer Labeled Samples Are Available



STL-10: 10 classes, 100 labeled samples/class, 96x96 images

airplane, bird, car, cat, deer, dog, horse, monkey, ship, truck

64.5%	→	**67.9%**
Hierarchical matching pursuit (2012)		This work

CIFAR-100: 100 classes, 500 labeled samples/class, 32x32 images

aquatic mammals, fish, flowers, food containers, fruit and vegetables, household electrical devices, household furniture, insects, large carnivores, large man-made outdoor things, large natural outdoor scenes, large omnivores and herbivores, medium-sized mammals, non-insect invertebrates, people, reptiles, small mammals, trees, vehicles

61.4%	**60.1%**
Maxout network (2013)	This work

Conclusion

- Greedy sparse encoder is useful, giving a scalable unsupervised representation learning pipeline that attains state-of-the-art classification performance
- Proper choice of encoder is critical: the stability of encoder is a key to the quality of representations