Interpretable unrolled dictionary learning networks

Bahareh Tolooshams

Demba Ba

Harvard John A. Paulson
School of Engineering
and Applied Sciences

DeepMath 2022

Deep learning theory

How do we learn?

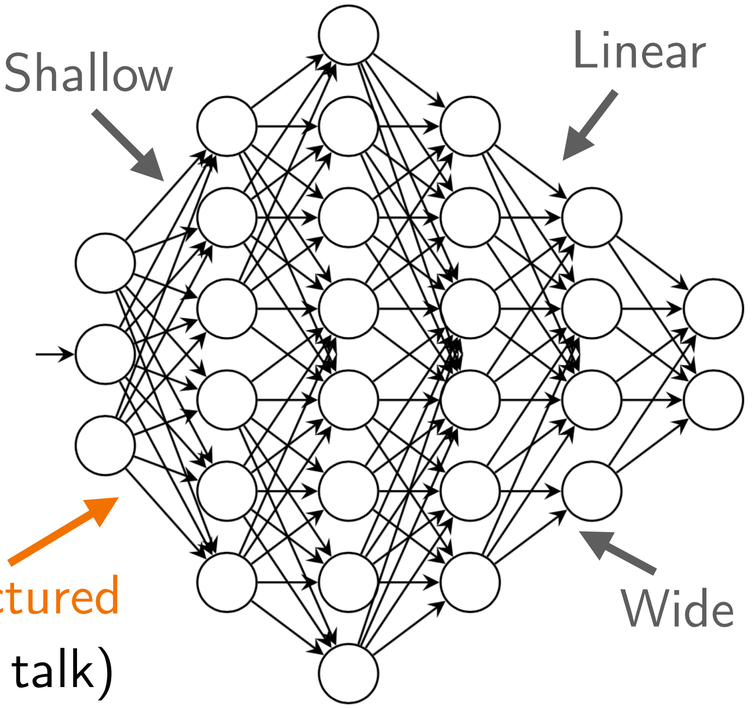
Loss landscape

Acceleration phenomenon

Neural collapse

Gradient dynamics

Convergence properties



Interpretability

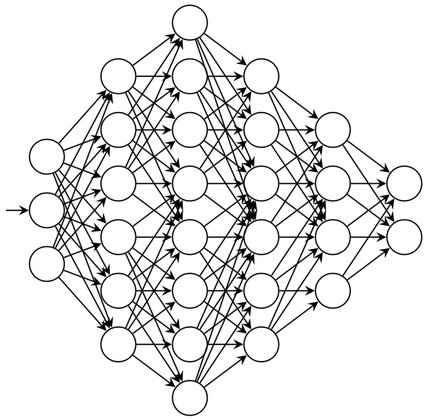
Decision making process?

Weights <—?—> Training set?

Inferred data <—?—> Training set?

Model-based deep learning

Deep learning



Statistical models

data representation

$$\mathbf{x} = g(\mathbf{z}), \qquad \mathbf{z} \sim p(\mathbf{z})$$

Sparse coding model (this talk)

Similarities: Use model as proxy for deep learning

Differences: Understand why deep learning is "better"

Sparse coding model

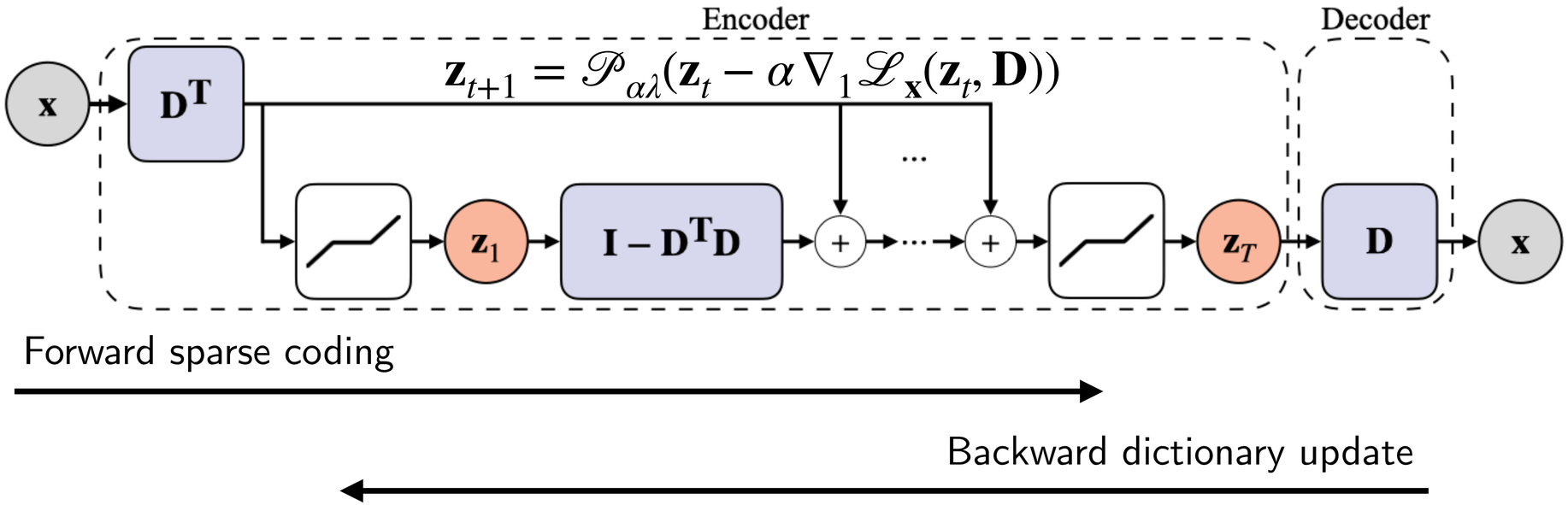
Model

data $\mathbf{x}$ = dictionary $\mathbf{D}^*$ sparse code $\mathbf{z}^*$

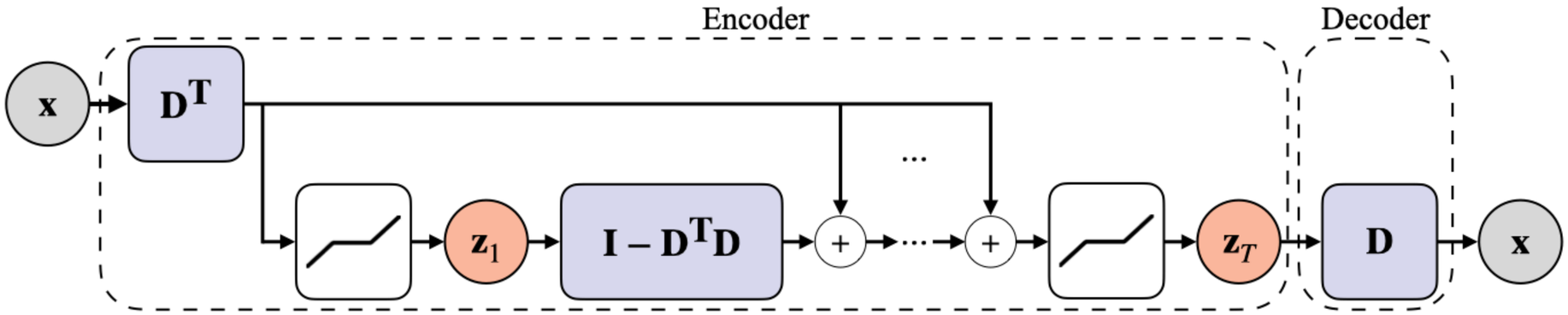
Optimization

data likelihood $\frac{1}{2}\|\mathbf{x} - \mathbf{D}\mathbf{z}\|_2^2$ + sparse prior $\lambda\|\mathbf{z}\|_1$

$$\min_{\mathbf{z},\mathbf{D}} \quad \mathscr{L}_{\mathbf{x}}(\mathbf{z}, \mathbf{D}) \quad + \quad h(\mathbf{z})$$



Forward sparse coding $\mathbf{x} = \mathbf{D}^* \mathbf{z}^*$



$$|\mathbf{z}_{t,(j)} - \mathbf{z}^*_{(j)}| \leq \mathcal{O}(\sqrt{s\|\mathbf{D}_j - \mathbf{D}^*_j\|_2} + e_{t,j} + \lambda)$$

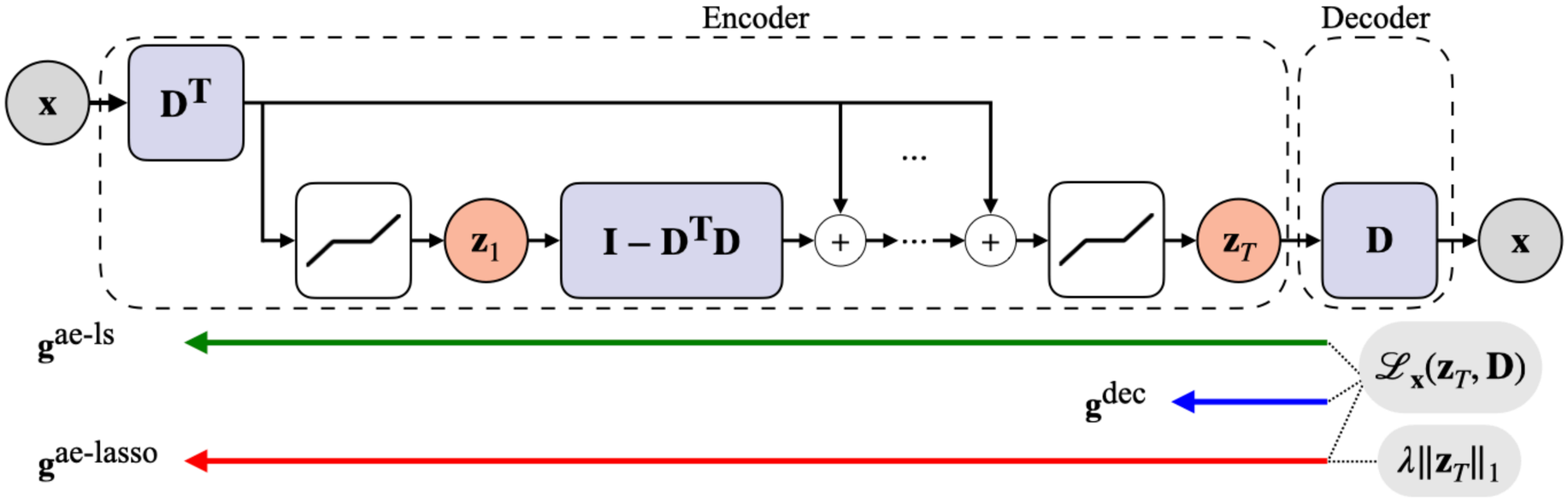
code sparsity

dictionary estimate

unrolling error
($e_{t,j} \to 0$ as $t \to \infty$)

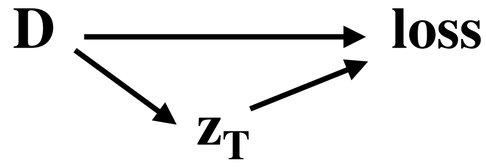
sparsity inducing
via architecture
(ReLU bias)

Gradient-based learning $\mathbf{D} \leftarrow \mathbf{D} - \eta \mathbf{g}_T$



direct gradient

$$\mathbf{g}_T^{\text{dec}} = \frac{d\,\text{loss}_\mathbf{x}(\mathbf{z}_T, \mathbf{D})}{d\mathbf{D}} = \nabla_2 \mathscr{L}_\mathbf{x}(\mathbf{z}_T, \mathbf{D})$$



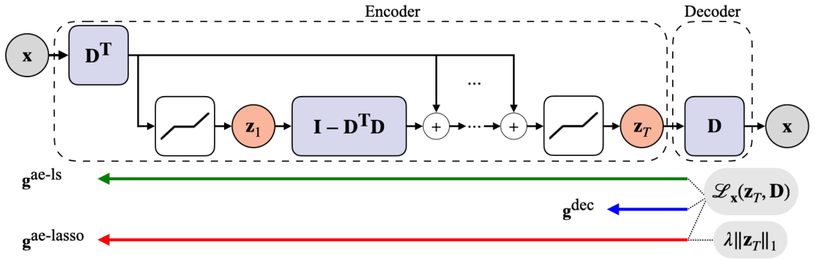
response gradient

$$\mathbf{g}_T^{\text{ae-lasso}} = \frac{d\,\text{loss}_\mathbf{x}(\mathbf{z}_T(\mathbf{D}), \mathbf{D})}{d\mathbf{D}} = \nabla_2 \mathscr{L}_\mathbf{x}(\mathbf{z}_T, \mathbf{D}) + \frac{\partial \mathbf{z}_T}{\partial \mathbf{D}} (\nabla_1 \mathscr{L}_\mathbf{x}(\mathbf{z}_T, \mathbf{D}) + \partial h(\mathbf{z}_T))$$

$$\mathbf{g}_T^{\text{ae-ls}} = \frac{d\,\text{loss}_\mathbf{x}(\mathbf{z}_T(\mathbf{D}), \mathbf{D})}{d\mathbf{D}} = \nabla_2 \mathscr{L}_\mathbf{x}(\mathbf{z}_T, \mathbf{D}) + \frac{\partial \mathbf{z}_T}{\partial \mathbf{D}} \nabla_1 \mathscr{L}_\mathbf{x}(\mathbf{z}_T, \mathbf{D})$$

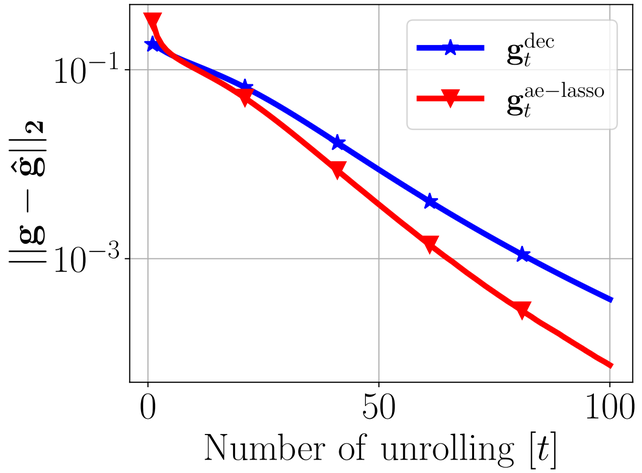
Effect of backpropagation

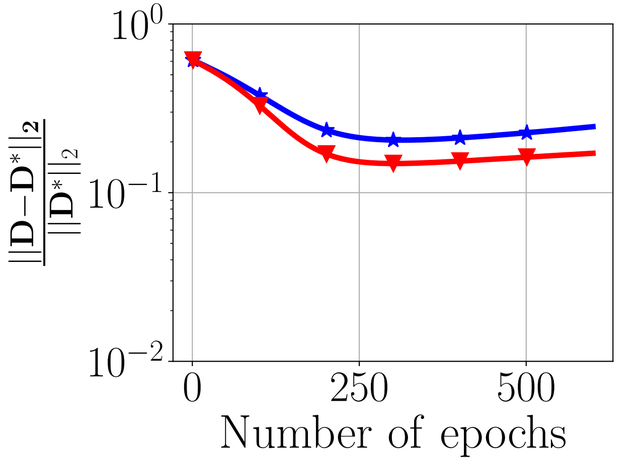
Best local direction $\hat{\mathbf{g}}$
(unroll for ∞)



$$\|\mathbf{g}_t^{\text{dec}} - \hat{\mathbf{g}}\|_2 \leq \mathcal{O}(\rho^t)$$

$$\|\mathbf{g}_t^{\text{ae-lasso}} - \hat{\mathbf{g}}\|_2 \leq \mathcal{O}(t\rho^{2t})$$





Dictionary learning with $\mathbf{g}_T^{\text{dec}}$

$$\|\mathbf{D}_j^{(l+1)} - \mathbf{D}_j^*\|_2^2 \leq (1-\psi)\|\mathbf{D}_j^{(l)} - \mathbf{D}_j^*\|_2^2 + \epsilon_\lambda^{(l)}$$

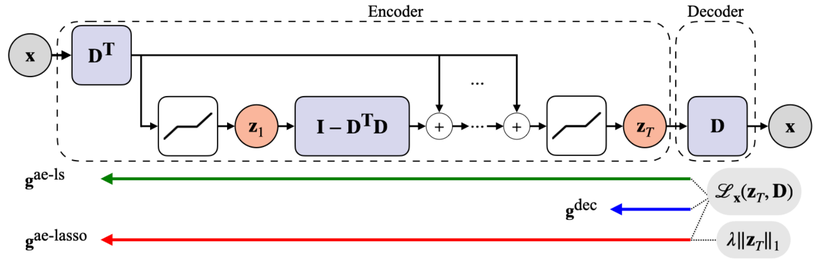
forward bias

Forward bias is propagated into the backward dictionary estimation

How to reduce this bias in the backward pass *without* changing the forward pass?

Reducing estimation bias

Desired global direction $\mathbf{g}^*$

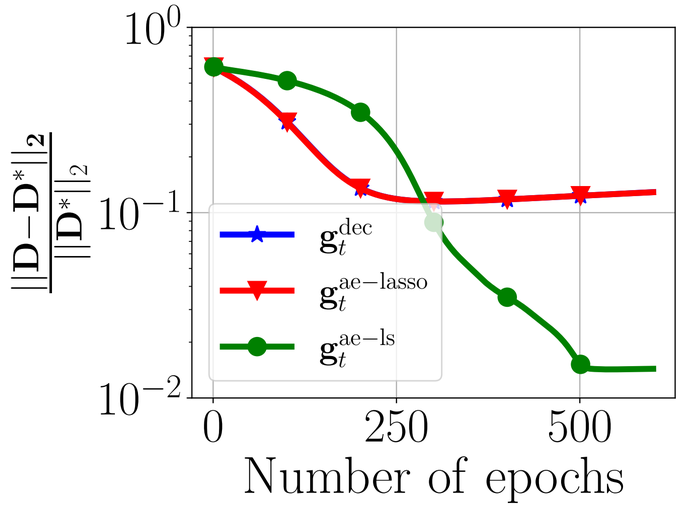


forward bias

$\mathcal{O}(\lambda\sqrt{s})$ lasso loss

$$\|\mathbf{g}_T^{\text{ae-lasso}} - \mathbf{g}^*\|_2 \leq \mathcal{O}(\|\mathbf{D} - \mathbf{D}^*\|_2 + \delta^* + C_{lasso})$$

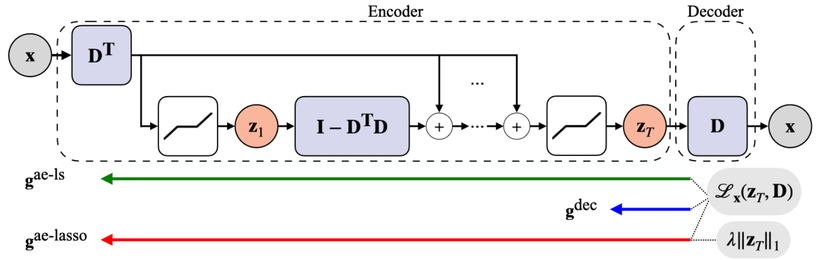
$$\|\mathbf{g}_T^{\text{ae-ls}} - \mathbf{g}^*\|_2 \leq \mathcal{O}(\|\mathbf{D} - \mathbf{D}^*\|_2 + \delta^*)$$



METHOD	PSNR [dB]			
λ	0.08	0.12	0.16	0.2
$\boldsymbol{g}_t^{\text{dec}}$	24.21 (0.12)	24.93 (0.14)	25.25 (0.06)	24.88 (0.00)
$\boldsymbol{g}_t^{\text{ae-ls}}$	24.79 (0.03)	25.43 (0.03)	**25.63** (0.04)	25.46 (0.05)

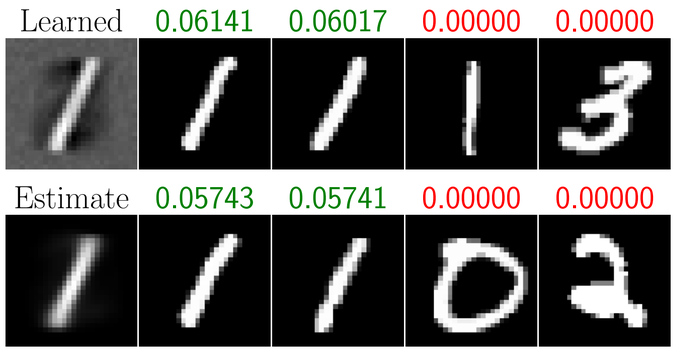
Interpretability

(Model) $\mathbf{x} = \mathbf{D}\mathbf{z}, \quad \mathbf{z}$ is sparse

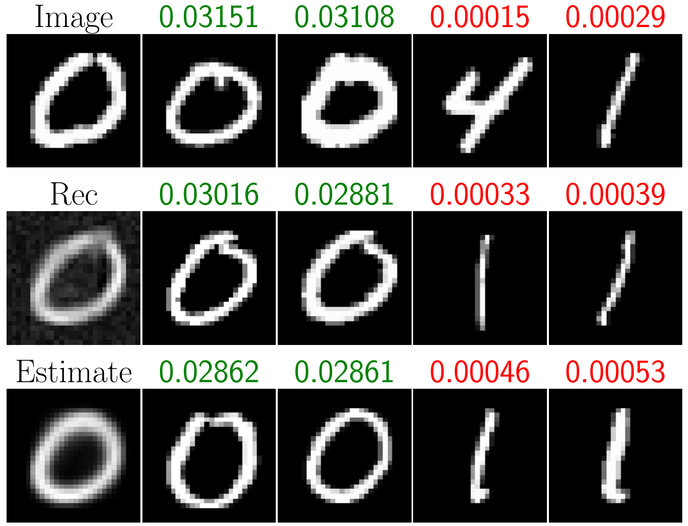


Learned dictionary atoms

$$\mathbf{D}_j^{\text{learned}} = \sum_{k=1}^{n} \alpha_k \mathbf{x}^k$$



$$\hat{\mathbf{x}}^{\text{test}} = \sum_{k=1}^{n} \beta_k \mathbf{x}^k$$



Published in Transactions on Machine Learning Research (08/2022)

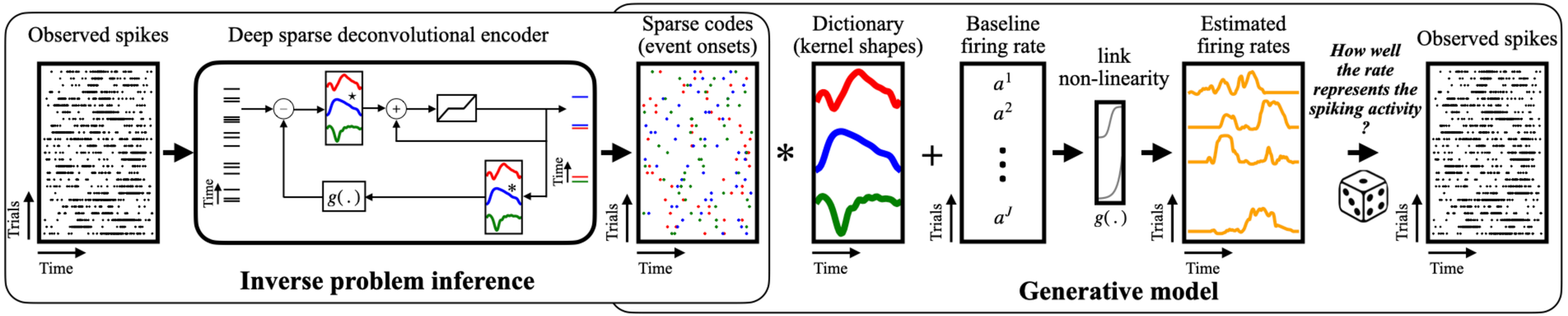
Stable and Interpretable Unrolled Dictionary Learning

Bahareh Tolooshams *btolooshams@seas.harvard.edu*
Demba Ba *demba@seas.harvard.edu*
School of Engineering and Applied Sciences
Harvard University

Reviewed on OpenReview: *https://openreview.net/forum?id=e3S0Bl2R08*

From theory to applications

Interpretable deep learning for computational neuroscience






Thank you!

Bahareh Tolooshams

btolooshams@seas.harvard.edu

https://btolooshams.github.io