Straight-Through Meets Sparse Recovery: the Support Exploration Algorithm

Mimoun Mohamed [1, 2] François Malgouyres [3] Valentin Emiya [1] Caroline Chaux [4]

[1]Aix Marseille Université, CNRS, LIS, Marseille, France [2]Aix Marseille Université, CNRS, I2M, Marseille, France
[3]Université de Toulouse, CNRS, IMT, Toulouse, France [4]CNRS, IPAL, Singapour

Summary

- Need a better understanding of the **Straight-Through Estimator** (STE) initially proposed for quantization in neural networks [1, 2]
- Propose a **sparse support recovery** algorithm by deriving the STE
  1. **Enhanced exploration capability** beyond local minima
  2. **Superior performance** with **highly-coherent dictionaries** (spike **deconvolution**)
  3. **Theoretical guarantees** for sparse recovery
  4. Can be **warm-started** with state-of-the-art algorithms

Sparse support recovery

Goal

Recover $S^* = \mathrm{supp}(x^*)$ from

$$y = Ax^* + e \in \mathbb{R}^m$$

with $x^* \in \mathbb{R}^n$ s.t. $\|\boldsymbol{x^*}\|_0 \leq \boldsymbol{k}$ and $A \in \mathbb{R}^{m\times n}$

Optimization problem

$$\underset{x\in\mathbb{R}^n,\, \|\boldsymbol{x}\|_0\leq \boldsymbol{k}}{\mathrm{Minimize}}\ F(x) := \frac{1}{2}\|Ax-y\|_2^2$$

Problem reformulation with a sparsification operator H

$$\underset{\mathcal{X}\in\mathbb{R}^n}{\mathbf{Minimize}}\ \boldsymbol{F(H(\mathcal{X}))} \quad \text{with} \quad \boldsymbol{H(\mathcal{X})} \in \underset{\substack{x\in\mathbb{R}^n \\ \mathrm{supp}(x)\subseteq \mathrm{largest}_k(\mathcal{X})}}{\mathbf{argmin}} \frac{\mathbf{1}}{\mathbf{2}}\|\boldsymbol{Ax}-\boldsymbol{y}\|_2^2$$

Straight-Through Estimator for sparsification

Differentiate $\boldsymbol{F(x) = F(H(\mathcal{X}))}$ where $\boldsymbol{H}$ is non-differentiable ?

$$\mathcal{X} \longrightarrow \boxed{H} \xrightarrow{x} \boxed{F} \longrightarrow F(H(\mathcal{X}))$$

Straight-through estimator $\dfrac{\partial(F\circ H)}{\partial\mathcal{X}}(\mathcal{X}) = \dfrac{\partial F}{\partial x}(H(\mathcal{X}))\dfrac{\boldsymbol{\partial x}}{\boldsymbol{\partial \mathcal{X}}}(\mathcal{X}) \approx \dfrac{\partial F}{\partial x}(H(\mathcal{X}))$

Gradient update: $\boldsymbol{\mathcal{X}^{t+1} = \mathcal{X}^t - \eta\frac{\partial F}{\partial x}(H(\mathcal{X})) = \mathcal{X}^t - \eta A^T(Ax^t - y)}$

Support Exploration Algorithm (SEA)

Main idea: Support exploration variable $\mathcal{X}^t$ searches for S^*

Algorithm 1 SEA [3]

```
Initialize 𝒳^0
repeat
    S^t = largest_k(𝒳^t)
    x^t = argmin_{x∈R^n, supp(x)⊂S^t} ‖Ax − y‖_2^2
    𝒳^{t+1} = 𝒳^t − ηA^T(Ax^t − y)
until halting criterion is true
t_BEST = argmin_{t'∈⟦0,t⟧} ‖Ax^{t'} − y‖_2^2
return x^{t_BEST}
```

Algorithm 2 HTP [4]

```
Initialize x^0
repeat
    S^t = largest_k(x^t)
    x^t = argmin_{x∈R^n, supp(x)⊂S^t} ‖Ax − y‖_2^2
    x^{t+1} = x^t − ηA^T(Ax^t − y)
until halting criterion is true

return x^t
```

$\Rightarrow$ Explore sparse solutions

$\Rightarrow$ Stop in a local minimum

- Support exploration variable $\mathcal{X}^t$ updated with an STE update
- $\mathcal{X}^t$ is a sum of gradients of explored sparse approximates

References

[1] G. Hinton. Neural networks for machine learning. Coursera, video lectures, 2012. Lecture 15b.

[2] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio. Binarized neural networks. *NeurIPS*, 2016.

[3] **M. Mohamed and F. Malgouyres and V. Emiya and C. Chaux. Straight-Through Meets Sparse Recovery: the Support Exploration Algorithm.** *ICML*, 2024.

[4] S. Foucart. Hard thresholding pursuit: An algorithm for compressive sensing. *SIAM J. Numer. Anal.*, 2011.

[5] E. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 2005.

Theorem - Recovery with RIP assumption

Upper bound on the number of iterations for sparse recovery

Assume A satisfies the $(2k+1)$-RIP [5] and $\|A_i\|_2 = 1$. If x^* satisfies

$$\alpha_k^{RIP}\|x^*\|_2 + \gamma_k^{RIP}\|e\|_2 < \frac{\min_{i\in S^*}|x_i^*|}{2k}$$

then for all $\mathcal{X}^0$, η, there exists $t_s \leq T_{RIP}$ such that $S^* \subseteq S^{t_s}$, where

$$T_{RIP} = \frac{2k\frac{\|\mathcal{X}^0\|_\infty}{\eta} + (k+1)\min_{i\in S^*}|x_i^*|}{\min_{i\in S^*}|x_i^*| - 2k\left(\alpha_k^{RIP}\|x^*\|_2 + \gamma_k^{RIP}\|e\|_2\right)}$$

Exact support recovery

If moreover, x^* is such that $\min_{i\in S^*}|x_i^*| > \frac{2}{\sqrt{1-\delta_{2k}}}\|e\|_2$, and SEA performs more than T_{RIP} iterations, then $S^* \subseteq S^{t_{BEST}}$ and $\|x^{t_{BEST}} - x^*\|_2 \leq \frac{2}{\sqrt{1-\delta_k}}\|e\|_2$

Gaussian deconvolution with $x^*_{|S^*} \sim \mathbb{U}([-2,-1]\cup[1,2])$

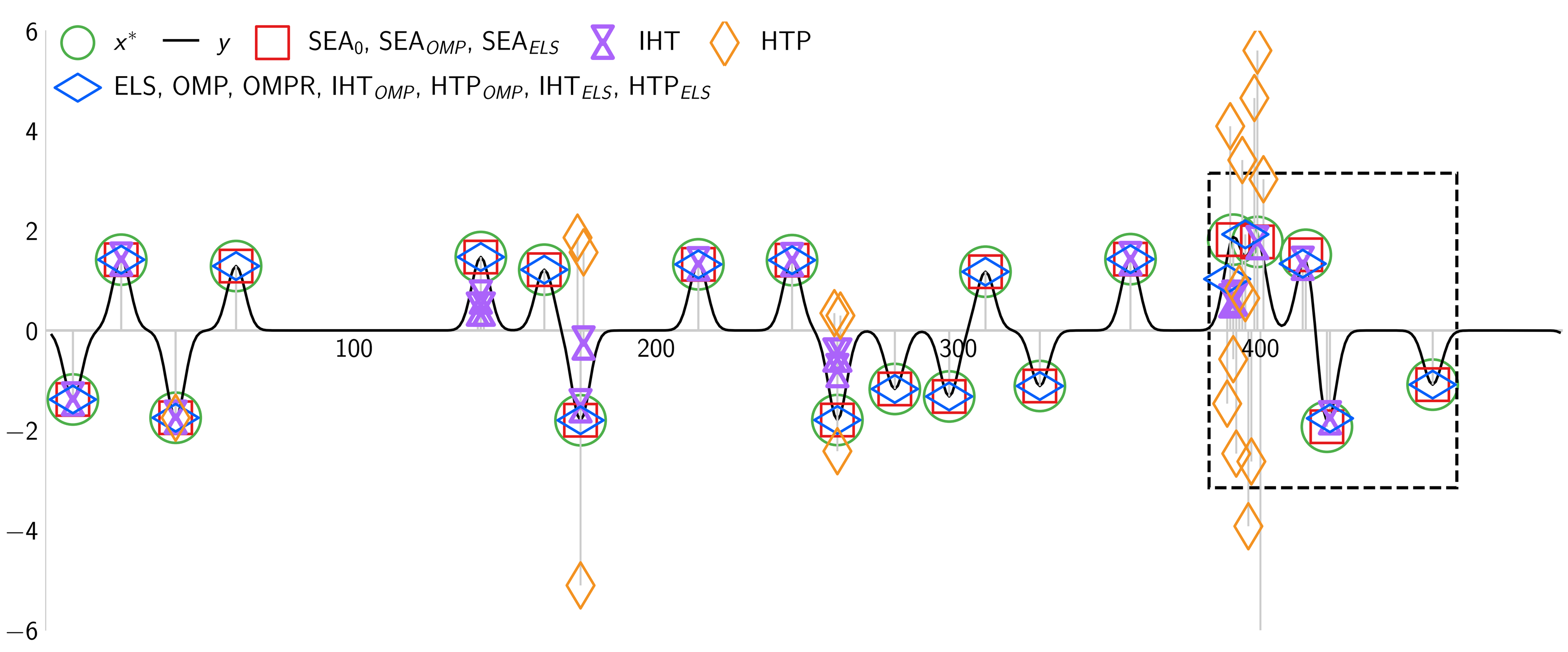

Figure 1. x^* and y with the solutions provided by the algorithms when $k = 20$

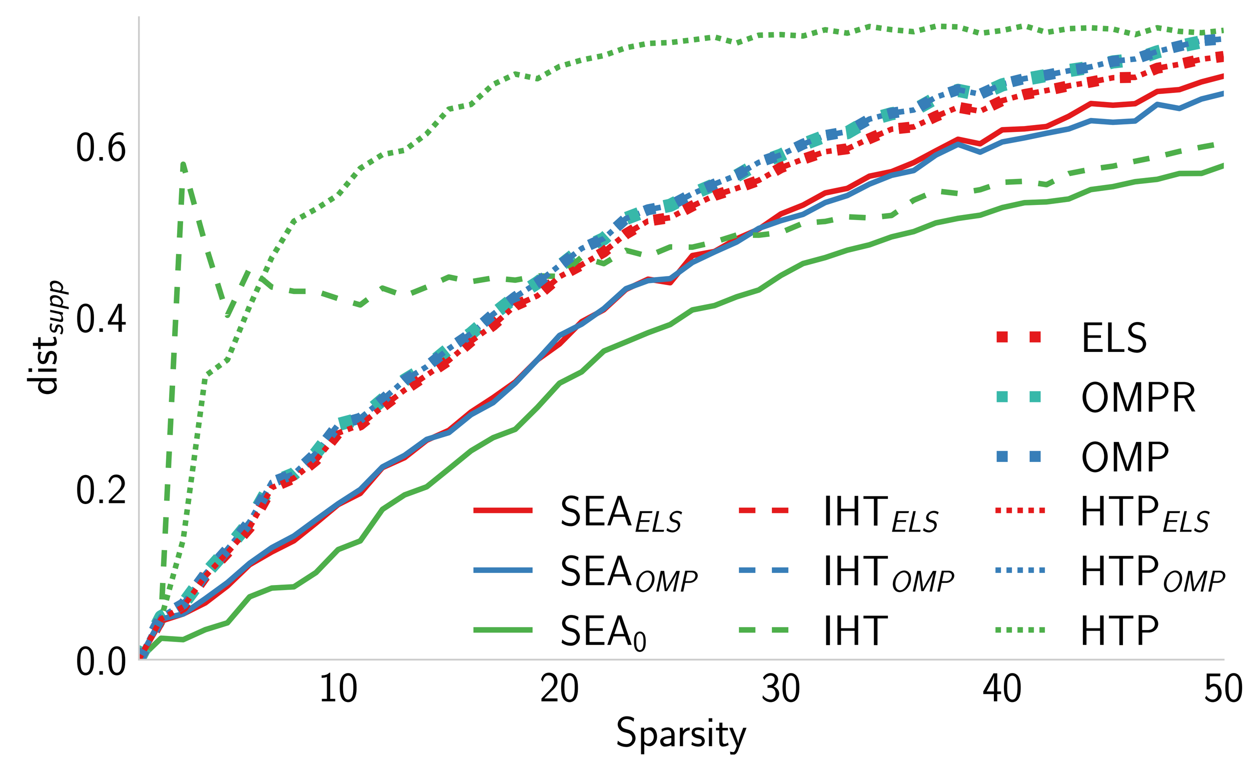

Figure 2. Average support distance $\mathrm{supp}_{\mathrm{dist}}(x) = \frac{k-|S^*\cap\,\mathrm{supp}(x)|}{k}$ between S^* and the support of the solutions provided by the algorithms over 200 run: $\mu(A) = 0.97$, $\sigma = 3$

Phase transition diagram with $A, x^*_{|S^*} \sim \mathcal{N}(0,1)$

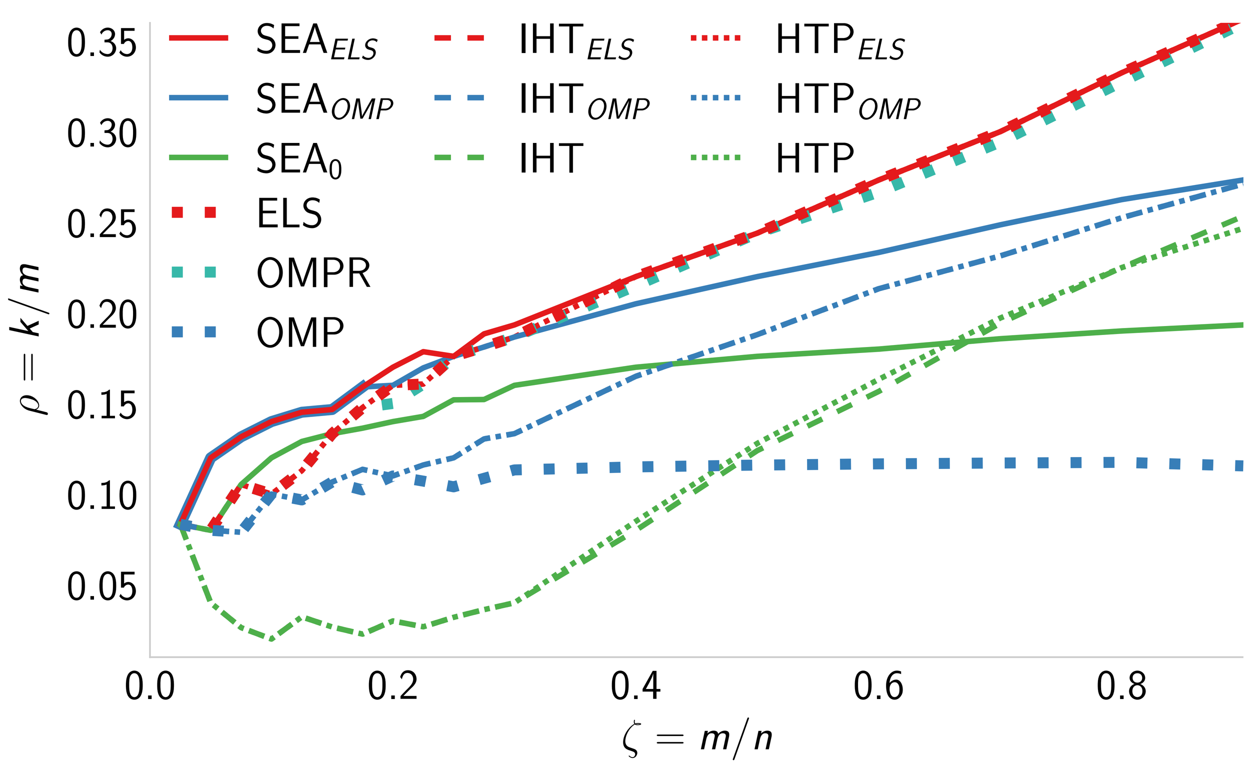

Figure 3. Empirical support recovery phase transition curves. Problems below each curve are solved by the algorithms with a success rate larger than 95% over 1000 runs