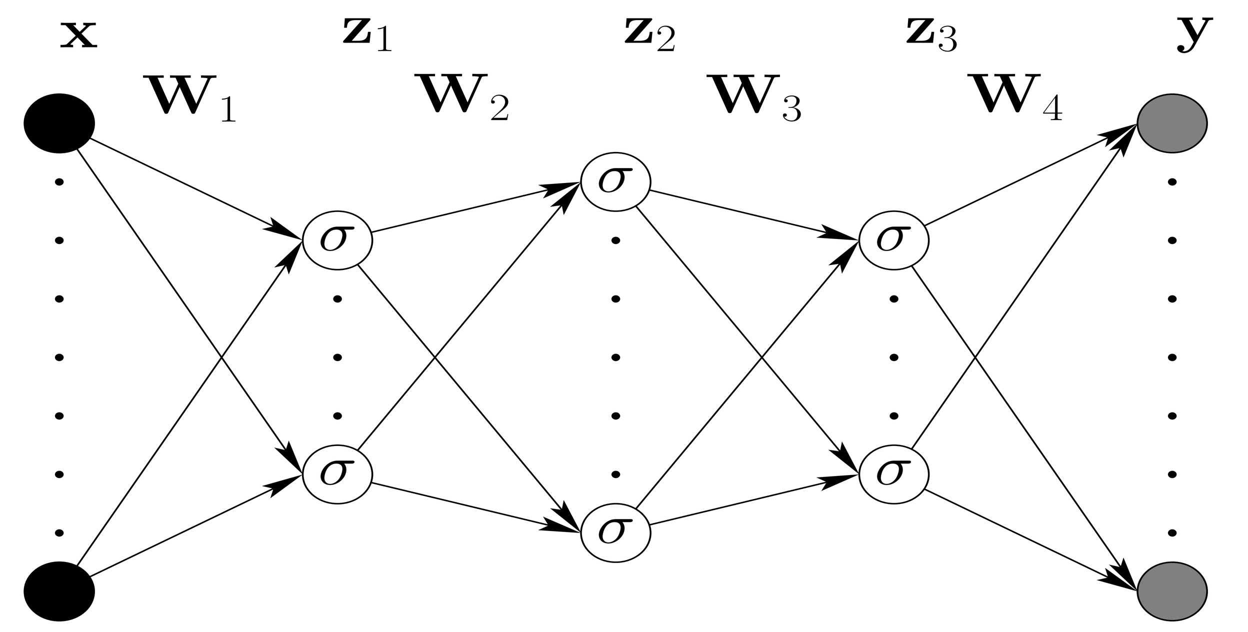
DISTRIBUTED OPTIMIZATION OF DEEPLY NESTED SYSTEMS

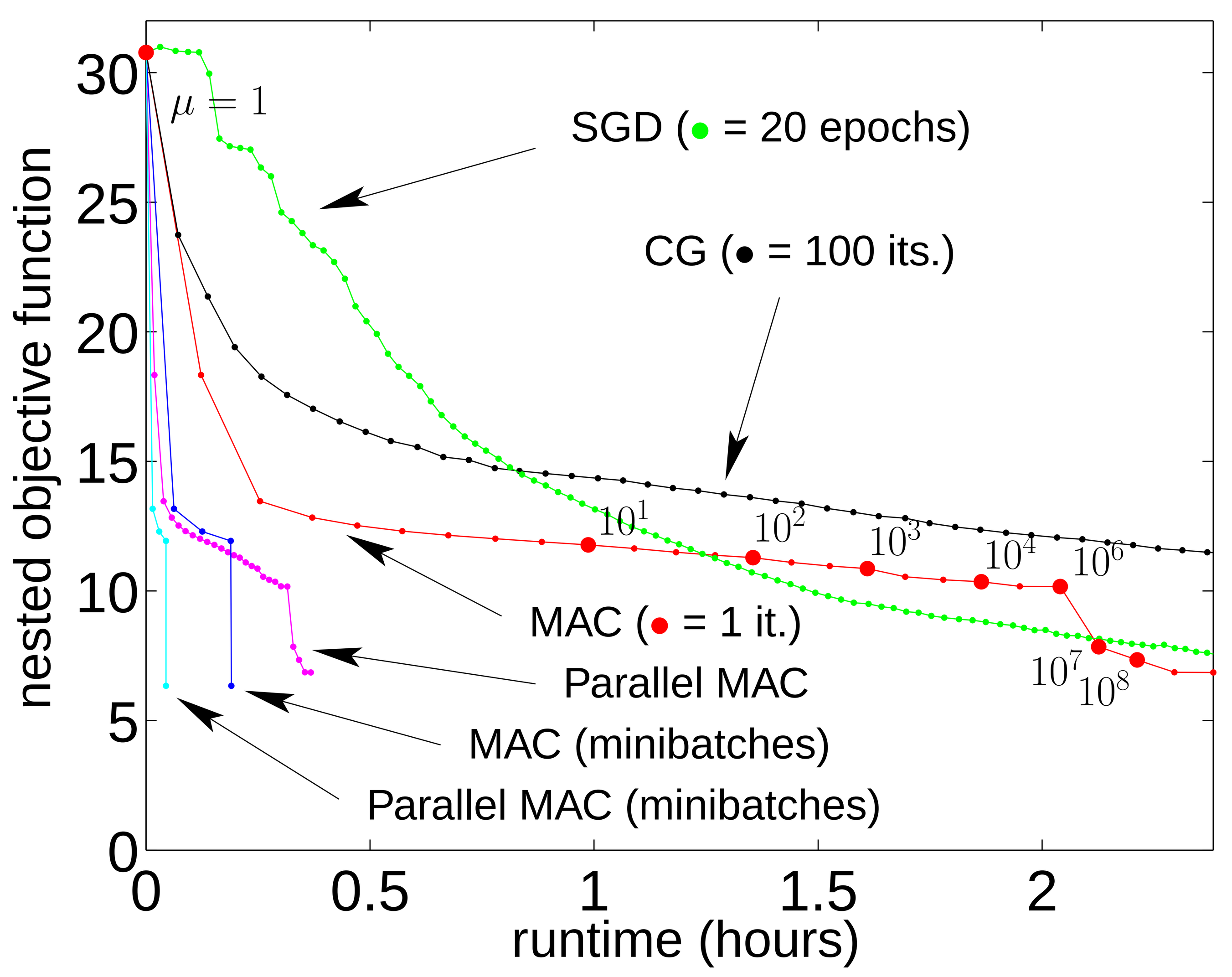
Miguel Á. Carreira-Perpiñán and **Weiran Wang**
EECS, University of California, Merced

1 Abstract

- Many architectures share the fundamental design principle of constructing a **deeply nested mapping from inputs to outputs**.
- Learning these architectures is challenging because nesting (i.e., function composition) produces inherently **nonconvex** functions.
- Backprop suffers from **vanishing gradients** and is hard to parallelize, is only applicable if the mappings are differentiable with respect to the parameters, and needs careful tuning of learning rates.
- Selecting the best architecture, for example the number of units in each layer of a deep net, or the number of filterbanks in a speech front-end processing, requires a **combinatorial search**.
- We describe a general optimization strategy called **method of auxiliary coordinates (MAC)**. It has **provable convergence**, is **easy to implement** reusing existing algorithms for single layers, can be **parallelized trivially and massively**, applies even when parameter **derivatives are not available** or not desirable, can perform some **model selection on the fly**, and is competitive with state-of-the-art nonlinear optimizers even in the serial computation setting, often providing **reasonable models within a few iterations**.



Net with $K=3$ hidden layers ($\mathbf{W}_k$: weights, $\mathbf{z}_k$: auxiliary coordinates).



Deep autoencoder ($K=5$) with USPS handwritten digit images.

2 The method of auxiliary coordinates (MAC)

A typical objective function to learn a deep net with K hidden layers:

$$\min_{\mathbf{W}} \ \frac{1}{2}\sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{f}(\mathbf{x}_n;\mathbf{W})\|^2 \qquad (1)$$

$$\mathbf{f}(\mathbf{x};\mathbf{W}) = \mathbf{f}_{K+1}(\dots\mathbf{f}_2(\mathbf{f}_1(\mathbf{x};\mathbf{W}_1);\mathbf{W}_2)\dots;\mathbf{W}_{K+1})$$

where each layer function has the form $\mathbf{f}_k(\mathbf{x};\mathbf{W}_k) = \sigma(\mathbf{W}_k\mathbf{x})$.

1. **Introduce one auxiliary variable per data point/hidden unit** and define the following **equality-constrained optimization** problem:

$$\min_{\mathbf{W},\mathbf{Z}} \ \frac{1}{2}\sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{f}_{K+1}(\mathbf{z}_{K,n};\mathbf{W}_{K+1})\|^2 \qquad (2)$$

$$\text{s.t.} \ \left\{\begin{array}{l} \mathbf{z}_{K,n} = \mathbf{f}_K(\mathbf{z}_{K-1,n};\mathbf{W}_K) \\ \dots \\ \mathbf{z}_{1,n} = \mathbf{f}_1(\mathbf{x}_n;\mathbf{W}_1) \end{array}\right\} \ n=1,\dots,N.$$

   Each $\mathbf{z}_{k,n}$ can be seen as the coordinates of $\mathbf{x}_n$ in an intermediate feature space, or as the hidden unit activations for $\mathbf{x}_n$.

2. Apply the **quadratic-penalty method** (or aug. Lagrangian). Optimize the following function over $(\mathbf{W},\mathbf{Z})$ for fixed $\mu > 0$ and drive $\mu \to \infty$:

$$\min_{\mathbf{W},\mathbf{Z}} \ \frac{1}{2}\sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{f}_{K+1}(\mathbf{z}_{K,n};\mathbf{W}_{K+1})\|^2 + \frac{\mu}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \|\mathbf{z}_{k,n} - \mathbf{f}_k(\mathbf{z}_{k-1,n};\mathbf{W}_k)\|^2.$$

   This defines a continuous path $(\mathbf{W}^*(\mu),\mathbf{Z}^*(\mu))$ which converges to a minimum of the constrained problem (2) and original problem (1).

3. Apply **alternating optimization** over $\mathbf{W}$ and $\mathbf{Z}$:
   - **$\mathbf{W}$-step** Minimizing over $\mathbf{W}$ for fixed $\mathbf{Z}$ results in a **separate minimization over the weights of each hidden unit**—each a single-layer, single-unit problem that is solved with existing algorithms.
   - **$\mathbf{Z}$-step** Minimizing over $\mathbf{Z}$ for fixed $\mathbf{W}$ **separates over the coordinates** $\mathbf{z}_n$ for each data point $n=1,\dots,N$:

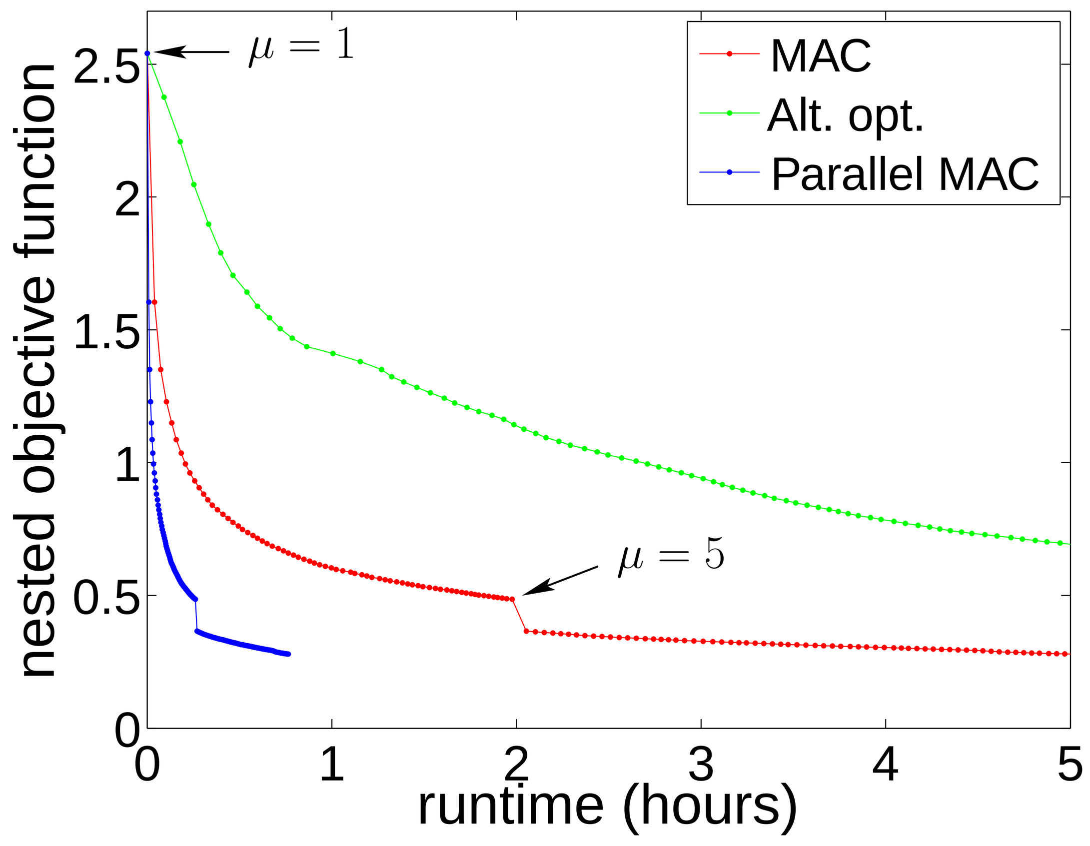
$$\min_{\mathbf{z}} \ \frac{1}{2}\|\mathbf{y} - \mathbf{f}_{K+1}(\mathbf{z}_K)\|^2 + \frac{\mu}{2}\sum_{k=1}^{K}\|\mathbf{z}_k - \mathbf{f}_k(\mathbf{z}_{k-1})\|^2,$$

     another nonlinear least squares problem that can be solved using the derivatives w.r.t. $\mathbf{z}$ of the single-layer functions $\mathbf{f}_1,\dots,\mathbf{f}_{K+1}$.

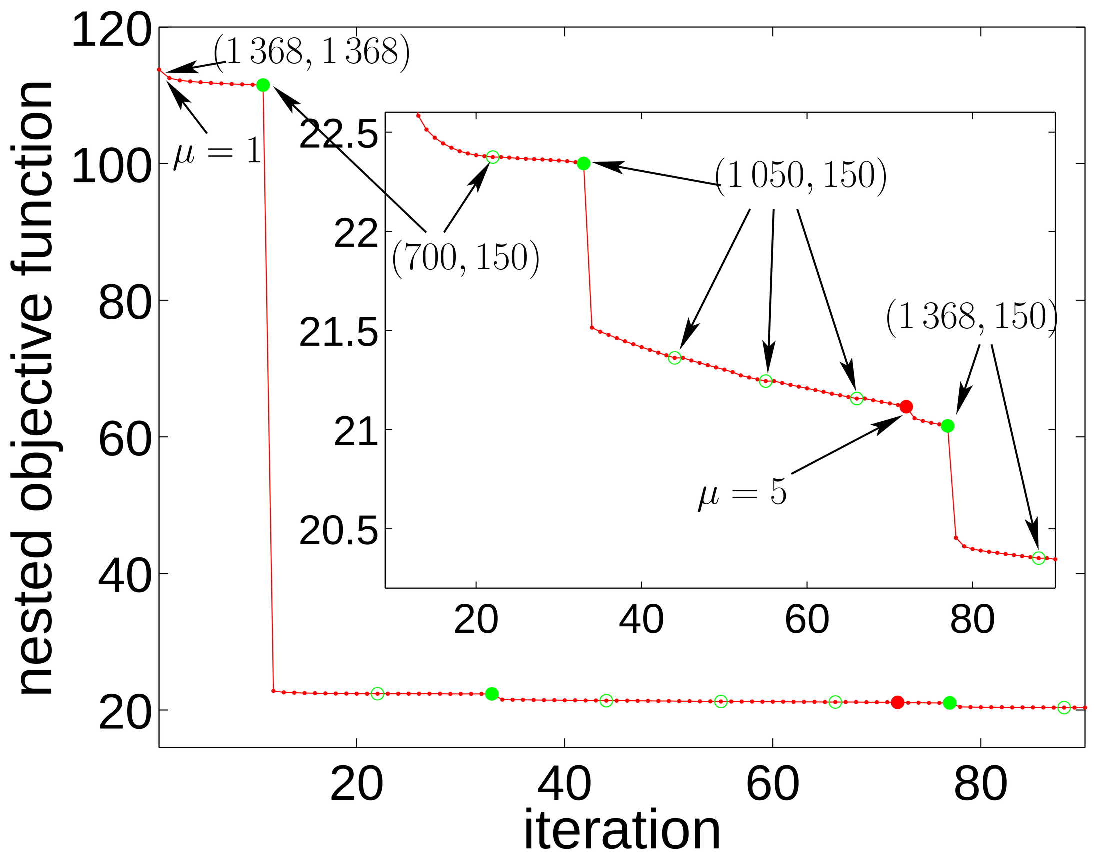
After stopping, we can apply a fast **post-processing step** to reduce the objective, achieve feasibility and eliminate the auxiliary coordinates.

3 Model selection

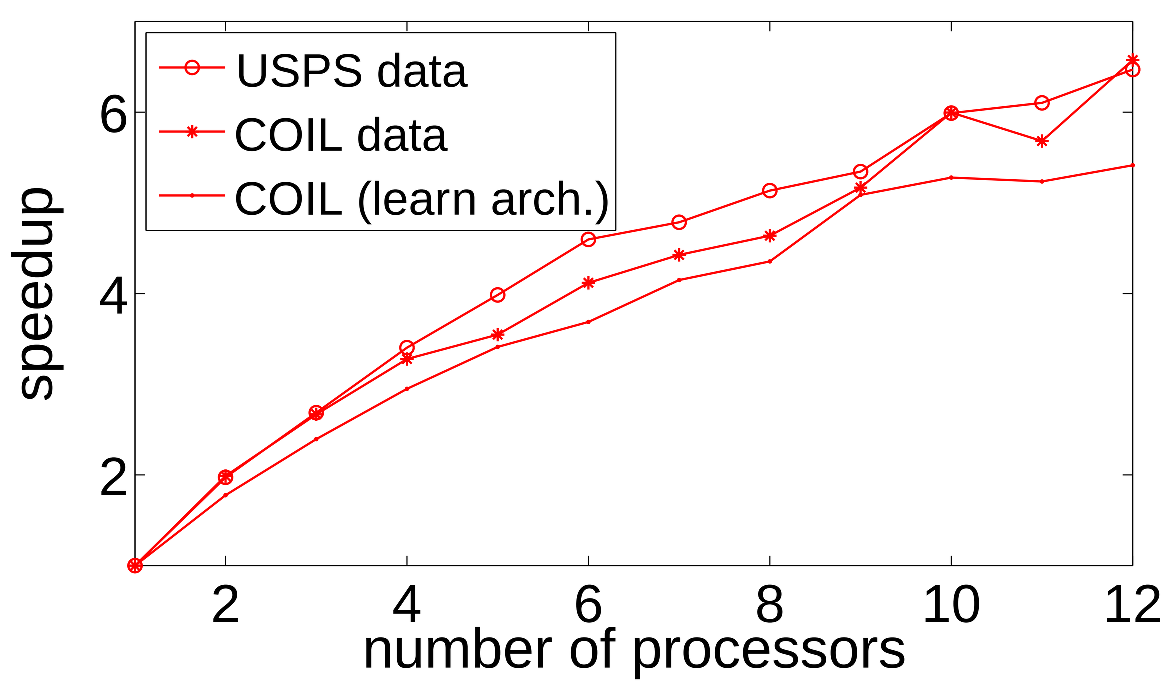
- Model selection may be achieved "on the fly" by having the $\mathbf{W}$-step do model selection **separately for each layer** (e.g., with criteria like BIC, AIC or minimum description length, or cross-validation).
- Instead of testing M^K deep nets, with MAC we can test only MK single-layer nets (in parallel) at each model-selection iteration.



RBF autoencoder with COIL–20 images.



Learning the architecture of RBF autoencoder.



Parallel processing speedup of MAC.