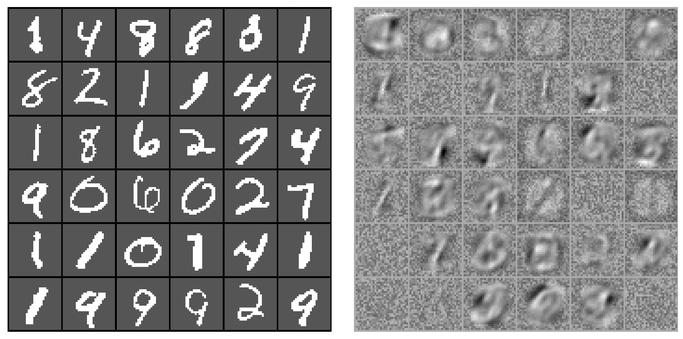
Enhanced Gradient for Learning Boltzmann Machines

Tapani.Raiko, KyungHyun.Cho, Alexander.Ilin @aalto.fi
Aalto University School of Science, Department of Information and Computer Science, Finland

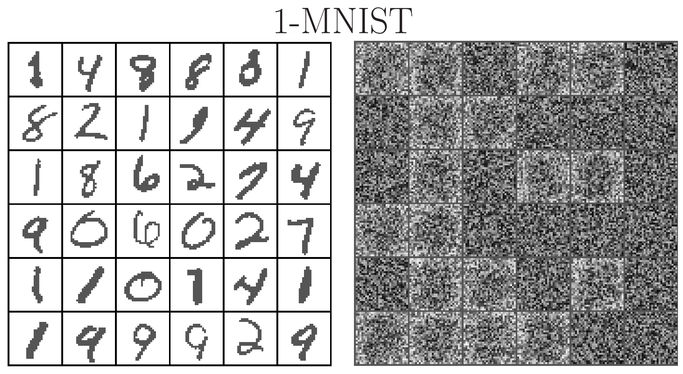
Problems of Traditional Learning Gradient

Sensitivity to Data Representation

MNIST

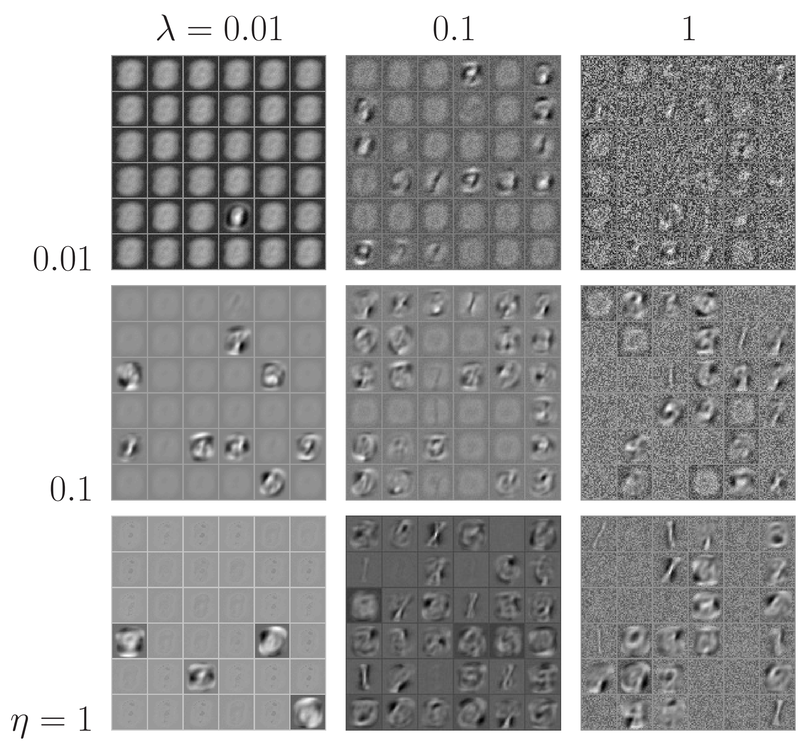


1-MNIST



- Dense data sets are more difficult to learn.
- Training BMs is not invariant to representation.

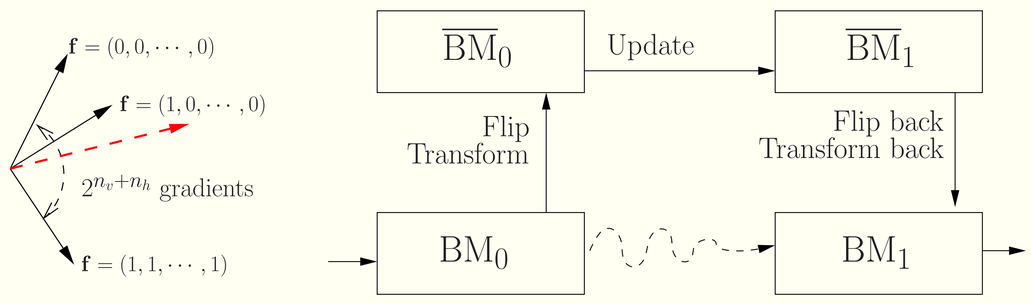
Sensitivity to Learning Parameters



- Learning parameters *greatly affect* learning:
  - Weight scale λ, learning rate η, momentum, weight decay, CDk

Enhanced Gradient

Bit-flipping Transformation



- Update: *Transform*, *update*, and *transform back*.
- $2^{n_v+n_h}$ well-founded ML updates exist.

Enhanced Gradient

- *Weighted sum* of all updates:

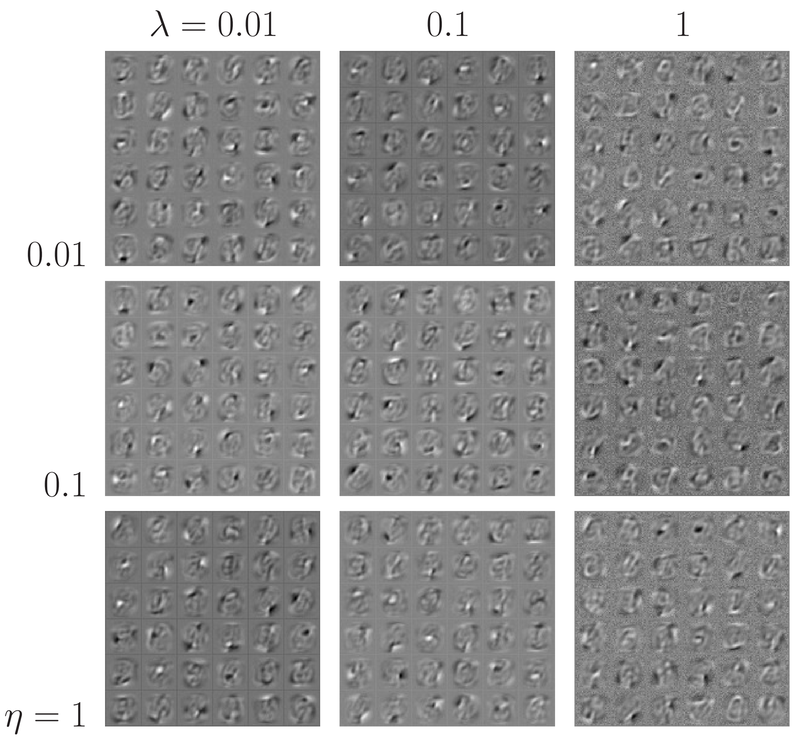
$$\tilde{\nabla} w_{ij} = \mathrm{Cov}_{\mathrm{d}}\left(x_i, x_j\right) - \mathrm{Cov}_{\mathrm{m}}\left(x_i, x_j\right)$$
$$\tilde{\nabla} b_i = \langle x_i \rangle_{\mathrm{d}} - \langle x_i \rangle_{\mathrm{m}} - \sum_j \langle x_j \rangle_{\mathrm{dm}} \tilde{\nabla} w_{ij}$$

- Weight for each gradient prefers *sparse representation*:

$$\prod_{k=1}^{n_v+n_h} \langle x_k \rangle_{\mathrm{dm}}^{f_k} \left(1 - \langle x_k \rangle_{\mathrm{dm}}\right)^{1-f_k}$$

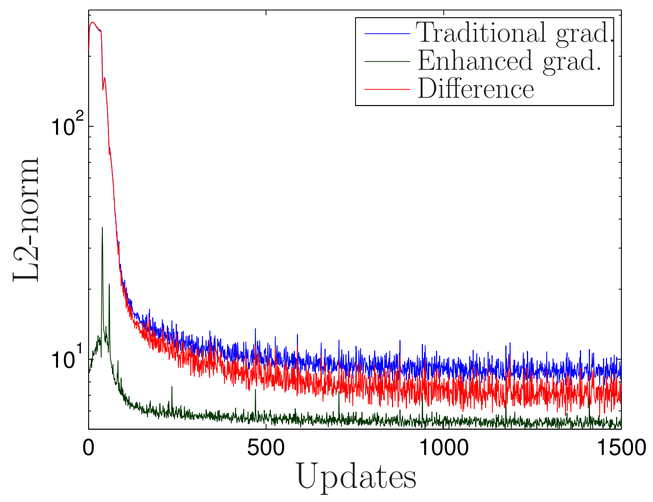
Improvements by Enhanced Gradient

Robustness to Learning Parameters

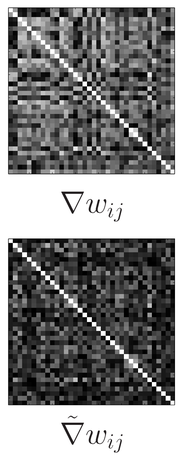


- Enhanced Gradient + Adaptive Learning Rate
- Reasonable filters regardless of initializations and learning rates.

Better Gradient



Cosine angles



Experimental Result: Caltech 101 Silhouettes

Hidden neurons	Log-probability PT	CD	(M)	Accuracy (%) PT	CD	(M)
500	-127.40	-280.91	**-125**	71.56	68.48	65.8
1000	-129.69	-190.80		**72.61**	70.39	
2000	-131.19	-166.72		71.82	71.39	

- Enhanced gradient works simply "*out of the box*".

(M) Marlin et al. 2010