Unsupervised Learning by Discriminating Data from Artificial Noise

Michael Gutmann
University of Helsinki
michael.gutmann@helsinki.fi

Aapo Hyvärinen
University of Helsinki
aapo.hyvarinen@helsinki.fi

Discriminative learning & logistic regression

- Logistic regression can be used to obtain a classifier which discriminates between the data sets $\mathbf{X} = (\mathbf{x}(1), \ldots, \mathbf{x}(T))$ and $\mathbf{Y} = (\mathbf{y}(1), \ldots, \mathbf{y}(T))$.
- Logistic regression uses the model

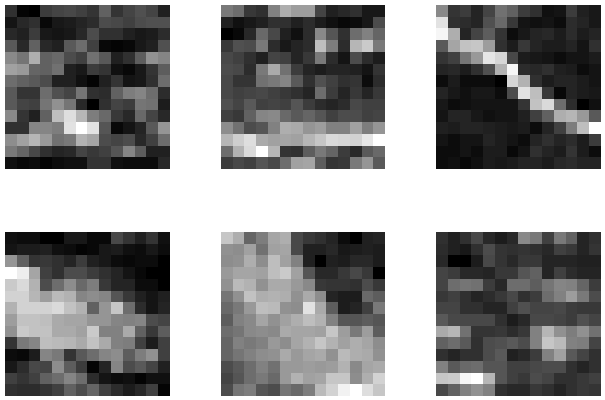
$$P(\mathbf{u} \in \mathbf{X}; \theta) = \frac{1}{1 + \exp(-G(\mathbf{u}; \theta))} \quad (1)$$

$$P(\mathbf{u} \in \mathbf{Y}; \theta) = 1 - P(\mathbf{u} \in \mathbf{X}; \theta), \quad (2)$$

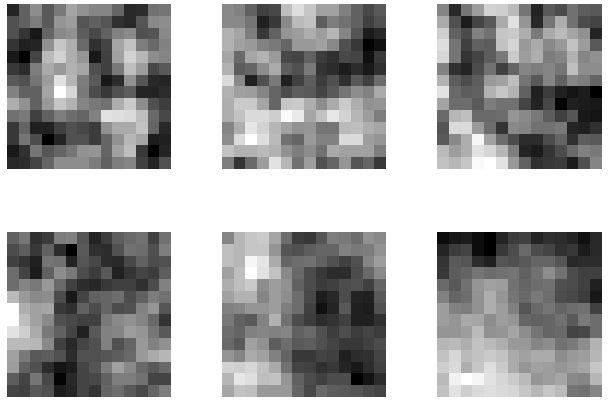
  where $G(\mathbf{u}; \theta)$ is a function parameterized by θ.
- For $G(\mathbf{u}; \theta) > 0$, $P(\mathbf{u} \in \mathbf{X}; \theta) > 0.5$ and the input $\mathbf{u}$ is classified to belong to $\mathbf{X}$.
- For a linear classifier: $G(\mathbf{u}; \theta) = w_0 + \mathbf{w}^T \mathbf{u}$. Parameters θ are $\{w_0, \mathbf{w}\}$.

Learning by comparison

- To succeed in the discrimination task, the classifier must learn differences between the two data sets $\mathbf{X}$ and $\mathbf{Y}$.
- Assume we know the properties of $\mathbf{Y}$. From the learned difference between $\mathbf{X}$ and $\mathbf{Y}$, we can thus deduce properties of $\mathbf{X}$.
- This can be formalized using estimation theory.



$\mathbf{X}$: natural images



$\mathbf{Y}$: noise images

Noise-contrastive estimation

- Observe data $\mathbf{X} = (\mathbf{x}(1), \ldots, \mathbf{x}(T))$ with *unknown* pdf p_d
- Generate noise $\mathbf{Y} = (\mathbf{y}(1), \ldots, \mathbf{y}(T))$ with *known* pdf p_n
- Define a parameterized function $f(\mathbf{u}; \theta)$, which models the data log-density $\log p_d(\mathbf{u})$
- Use logistic regression with the nonlinearity

$$G(\mathbf{u}; \theta) = f(\mathbf{u}; \theta) - \log p_n(\mathbf{u}) \tag{3}$$

- Conditional likelihood leads to the objective function

$$J(\theta) = \sum_t \log\left[h(\mathbf{x}(t); \theta)\right] + \log\left[1 - h(\mathbf{y}(t); \theta)\right]$$

$$\text{where } h(\mathbf{u}; \theta) = \frac{1}{1 + \exp[-G(\mathbf{u}; \theta)]} \tag{4}$$

- The estimator is defined as $\hat{\theta} = \text{argmax} J(\theta)$
- (Link to classification: $P(\mathbf{u} \in \mathbf{X}; \theta) = h(\mathbf{u}; \theta)$)

Properties of the estimator

- Assume the parametric model $f(\mathbf{u};\theta)$ can approximate any function. Then, the maximum of objective J is attained when

$$f(\mathbf{u};\theta) = \log p_d(\mathbf{u}), \tag{5}$$

  where $p_d(\mathbf{u})$ is the pdf of the observed data.
- For data generated according to the model, i.e.

$$\log p_d(\mathbf{u}) = \log p_m(\mathbf{u};\theta^*), \tag{6}$$

  we can show that the estimator is *statistically consistent*.
- No normalization constraint such as $\int \exp(f(\mathbf{u};\theta))d\mathbf{u} = 1$ is needed in the optimization of J. We thus don't need to know the partition function!
  $\Rightarrow$ Estimation of unnormalized models is possible.
  (This is not possible with maximum likelihood estimation)

Unsupervised learning by supervised learning

A central message of this talk:

Supervised learning leads to unsupervised estimation of a (unnormalized) probabilistic model given by the log-density $f(\mathbf{u}; \theta)$.

Unsupervised learning can thus be performed by supervised learning!

Example: estimation of an ICA model

- Assume data $\mathbf{X}$ follows the ICA model

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{7}$$

- The sources s_i are i.i.d Laplacian, the mixing matrix $\mathbf{A}$ invertible, and dim($\mathbf{x}$)=dim($\mathbf{s}$)=4.
- The pdf of the data (p_d) and the model (p_m) are

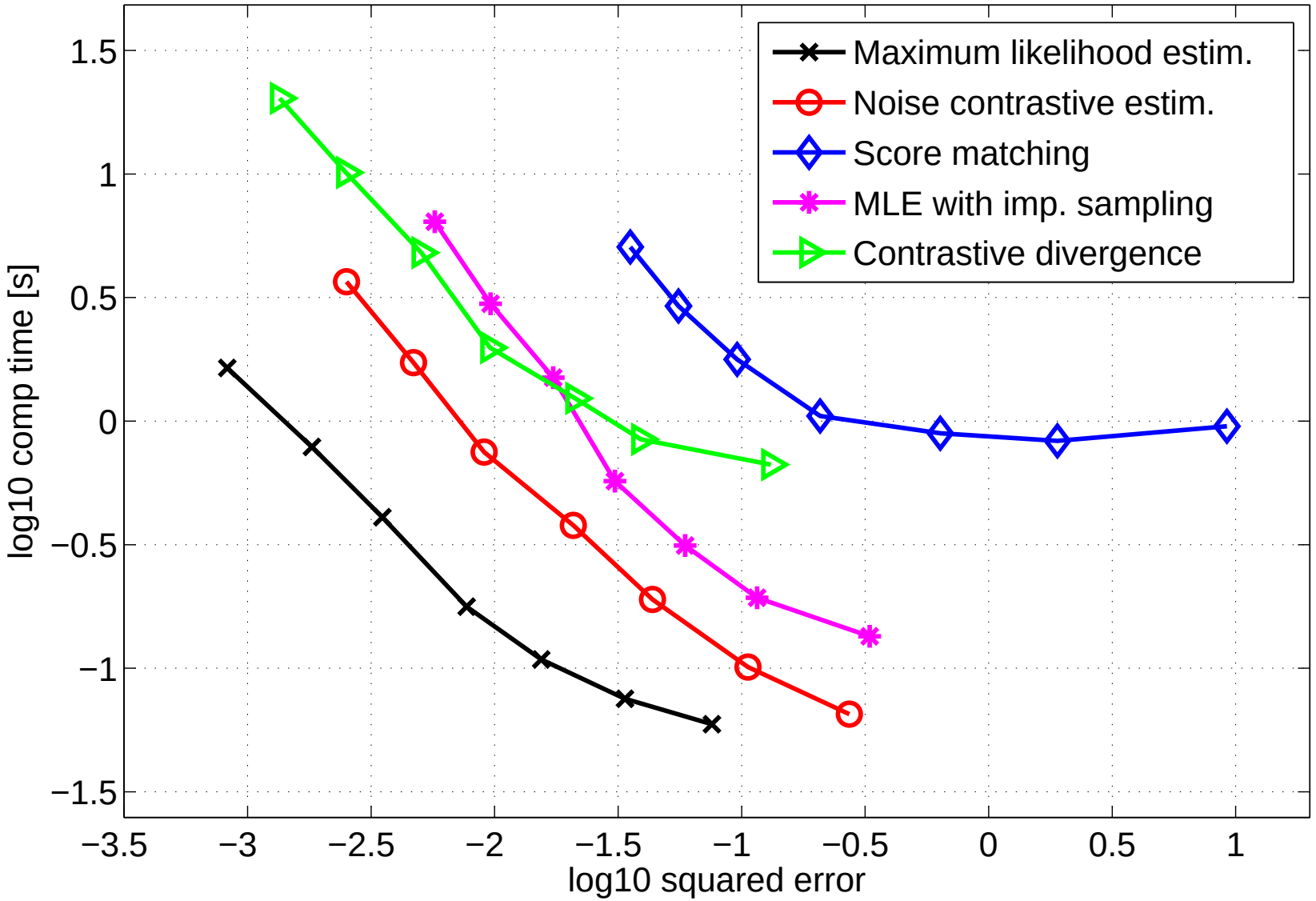
$$\log p_d(\mathbf{x}) = -\sum_{i=1}^{4} \sqrt{2}\,|\mathbf{b}_i^* \mathbf{x}| + (\ln|\det \mathbf{B}^*| - \ln 4) \tag{8}$$

$$\log p_m(\mathbf{x};\theta) = -\sum_{i=1}^{4} \sqrt{2}|\mathbf{b}_i \mathbf{x}| + c, \tag{9}$$

  where $\mathbf{B}^* = \mathbf{A}^{-1}$. The parameters $\theta \in \mathbb{R}^{17}$ are the row vectors $\mathbf{b}_i$ and c.

- Contrastive noise $\mathbf{Y}$: data with the same covariance as $\mathbf{x}$.

Estimation error versus computation time



Summary

1. By learning to discriminate between observed data and artificially generated data (noise), you can learn properties of the observed data (“learning by comparison”).
2. This principle provides a consistent estimator of a (possibly unnormalized) parametric model for the observed data.
3. It offers (for an ICA model) the best trade-off between computational and statistical efficiency. It proved highly useful in the estimation of models for natural images (results not shown).
4. It shows that unsupervised learning can be performed by supervised learning.
   $\Rightarrow$ Useful for deep learning with energy-based models?!