On Finite-Sample Identifiability of Contrastive Learning-Based Nonlinear Independent Component Analysis

Qi Lyu and Xiao Fu

School of EECS, Oregon State University

July 11, 2022

Independent Component Analysis (ICA)

- ICA learns statistically independent latent factors from data.
- ICA is widely applied.
  - Biomedical signal processing [Ziehe et al., 2000, Oveisi et al., 2012]
  - Speech separation [Comon and Jutten, 2010]
  - Causal discovery [Zhang and Hyvärinen, 2010, Monti et al., 2020]
  - Disentanglement [Locatello et al., 2020, Khemakhem et al., 2020]
  - Self-supervised Learning [Zimmermann et al., 2021]
  - ......

The "cocktail party problem". Source: https://dbcover.com/cocktail-party-effect-and-room-acoustics/

Nonlinear ICA (nICA) Model

- nICA assumes

$$\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{s}),$$

where $\boldsymbol{x} \in \mathbb{R}^M$ is the data, $\boldsymbol{s} \in \mathbb{R}^D$ are the D latent components.

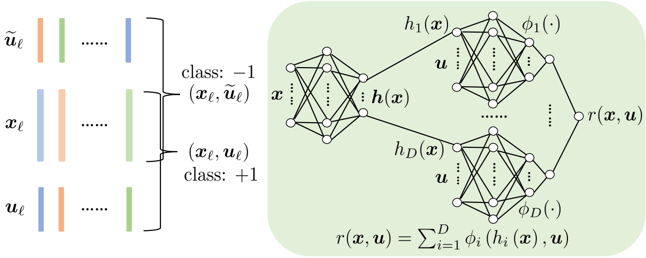
  - $\boldsymbol{g}(\cdot) : \mathbb{R}^D \to \mathbb{R}^M$ is a smooth and invertible ***unknown*** function.
  - $s_1, \ldots, s_D$ are statistically independent.
  - Challenge: nICA is not identifiable [Hyvärinen and Pajunen, 1999].
  - Solution: additional information is needed.
  - Works on model identification are developing [Hyvarinen and Morioka, 2016, Hyvarinen and Morioka, 2017, Hyvarinen et al., 2019, Khemakhem et al., 2020, Locatello et al., 2020, Gresele et al., 2020].

Contrastive Learning-Based nICA

- [Hyvarinen et al., 2019] assumes that an *auxiliary variable* $\boldsymbol{u}$ is observed.
- Given $\boldsymbol{u}$, $\boldsymbol{s}$ is conditionally independent, i.e.,

$$\log p(\boldsymbol{s}|\boldsymbol{u}) = \sum_{i=1}^{D} q_i(s_i, \boldsymbol{u}),$$

  where $q_i(\cdot, \cdot)$ is a continuous function.
- Goal: learn a logistic regression function $\boldsymbol{r}(\boldsymbol{x}, \boldsymbol{u})$ [Hyvarinen et al., 2019].



Identifiability Result

- Criterion: realize using the logistic loss

$$\min_{\phi,\boldsymbol{h}} \mathcal{L} = \min_{\phi,\boldsymbol{h}} \mathbb{E}_{\boldsymbol{z}} \left[\log(1+\exp[-dr(\boldsymbol{z})])\right],$$

where $d=+1$ for $\boldsymbol{z}=(\boldsymbol{x},\boldsymbol{u})$, $d=-1$ for $\boldsymbol{z}=(\boldsymbol{x},\widetilde{\boldsymbol{u}})$.

Theorem (Model Identifiability) [Hyvarinen et al., 2019] [Informal]

- Variability Assumption is satisfied (i.e., $\boldsymbol{u}$ is informative);
- $\boldsymbol{s}$ is conditionally independent given $\boldsymbol{u}$;
- With infinite data samples.

Then, $h^\star_{\pi(i)}(\boldsymbol{x}) = v_i^{-1}(s_i)$, for $i=1,\dots,D$, where $\{\pi(1),\dots,\pi(D)\}$ is a permutation of $\{1,\dots,D\}$

- A notable **gap**: in practice, we only have **finite samples**.

Challenges

- There is no sample complexity analysis on nICA.
  - [Arora et al., 2012] analyzed classic linear ICA.
  - [Lyu and Fu, 2021] assumed a structured (post-nonlinear) model.
  - [Lyu et al., 2022] considered a multiview mixture model.

- What are the **challenges** for analyzing nICA?
  - The optimal solution is based on sample size $N = \infty$;
  - Taking derivatives only holds on continuous open domain.
  - No unified metric: in supervised learning, one measures if $y \approx \boldsymbol{f}(\boldsymbol{x})$.

Finite Sample Analysis

- How do we approach the problem?
- **Step1**: Logistic regression, learn $r(\boldsymbol{x}, \boldsymbol{u}) = \sum_{i=1}^{D} \phi_i \left(h_i \left(\boldsymbol{x}\right), \boldsymbol{u}\right)$.
  - $N = \infty$: after convergence [Hyvarinen et al., 2019, Goodfellow et al., 2014],

$$\underbrace{\sum_{i=1}^{D} \phi_i^{\star} \left(h_i^{\star} \left(\boldsymbol{x}\right), \boldsymbol{u}\right)}_{\widehat{r}^{\star}(\boldsymbol{x},\boldsymbol{u})} = \underbrace{\log p(\boldsymbol{x}|\boldsymbol{u}) - \log p(\boldsymbol{x})}_{r^{\star}(\boldsymbol{x},\boldsymbol{u})}. \tag{1}$$

  - $N \neq \infty$: we derive that

$$\mathbb{E}[|\widehat{r}^{\star}(\boldsymbol{x}, \boldsymbol{u}) - r^{\star}(\boldsymbol{x}, \boldsymbol{u})|^2] \leq \varepsilon,$$

    where ε depends on modeling error, function learner and sample size.

Finite Sample Analysis

- **Step2**: Characterize the unobserved data point.
  - $N = \infty$: the equation holds **everywhere** (i.e., $r^\star(\boldsymbol{x}, \boldsymbol{u}) = \widehat{r}^\star(\boldsymbol{x}, \boldsymbol{u})$)

$$\sum_{i=1}^{D} q_i(v_i(\boldsymbol{y}_\ell), \boldsymbol{u}_\ell) - \log p_s(\boldsymbol{v}(\boldsymbol{y}_\ell)) = \sum_{i=1}^{D} \phi_i\left([\boldsymbol{y}_\ell]_i, \boldsymbol{u}_\ell\right), \ \forall \ (\boldsymbol{x}_\ell, \boldsymbol{u}_\ell)$$

    where $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x}), \quad \boldsymbol{v}(\boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{h}^{-1}(\boldsymbol{y})) = \boldsymbol{s}$.
  - $N \neq \infty$: for each **unobserved** $(\boldsymbol{x}_\ell, \boldsymbol{u}_\ell)$, characterize the distance

$$\varepsilon_\ell = \left(\sum_{i=1}^{D} q_i(v_i(\boldsymbol{y}_\ell), \boldsymbol{u}_\ell) - \log p_s(\boldsymbol{v}(\boldsymbol{y}_\ell)) - \sum_{i=1}^{D} \phi_i\left([\boldsymbol{y}_\ell]_i, \boldsymbol{u}_\ell\right) \right)^2,$$

    with $\mathbb{E}_{\mathcal{D}}[\varepsilon_\ell] \leq \varepsilon$.
- **Step3**: Compute the derivatives.
  - $N = \infty$: taking derivative gives $\boldsymbol{\gamma}_{jk} = \left[\frac{\partial^2 v_1(\boldsymbol{y})}{\partial y_j \partial y_k}, \cdots, \frac{\partial^2 v_D(\boldsymbol{y})}{\partial y_j \partial y_k} \right]^\top = 0$
  - $N \neq \infty$: numerically estimating the cross-derivatives gives

$$\mathbb{E}_{\mathcal{D}}\left[\|\widehat{\boldsymbol{\gamma}}_{jk}\|_2^2 \right] \leq \text{ certain bound.}$$

Sample Complexity Result

Theorem (Sample Complexity) [Informal]

- Assume the problem is solved with N i.i.d. samples $\{\boldsymbol{z}_\ell\}_{\ell=1}^{N}$;
- the learned $\boldsymbol{h}$ is invertible;
- the 4th-order derivative of $\widehat{r}^\star(\boldsymbol{z}) - r^\star(\boldsymbol{z})$ is bounded.

Then, we have the following bound with probability of at least $1-\delta$,

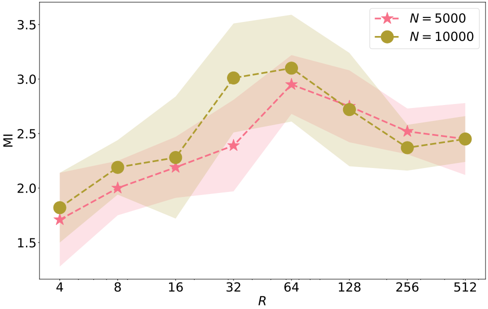
$$\mathbb{E}_{\mathcal{D}}\left[\|\widehat{\gamma}_{jk}\|_2^2\right] \leq O\left(\frac{D(1+e^{\alpha})}{e^{\alpha/2}}\left(\mathfrak{R}_N + \nu + \alpha\sqrt{\frac{\ln(1/\delta)}{N}}\right)^{1/2}\right),$$

where α is a bound of $|\boldsymbol{r}(\boldsymbol{z})|$, $\mathfrak{R}_N$ is Rademacher complexity.

- $\mathfrak{R}_N$ grows when DNN is more complex and decreases when N grows
- ν: the expressiveness of the DNN; $\nu = 0$ when DNN is universal.
- **Implication**: Use an expressive DNN, not an overly complex one.

Experiment Results

- We follow the settings in [Hyvarinen et al., 2019].
- s_i is the product of a Gaussian and a Laplacian variable.
- $\boldsymbol{u}$ corresponds to different time frames.
- $\boldsymbol{g}(\cdot)$ is neural network with leaky ReLU.
- $\boldsymbol{h}(\cdot)$, $\phi_i(\cdot)$ are modeled with 3-hidden-layer network with R neurons.
- Metric: mutual information between s_i and $h_j(\boldsymbol{x})$.



- There is a trade-off in terms of expressiveness of $\boldsymbol{h}(\cdot)$ (i.e., R).

Conclusion

- We propose the first framework for sample complexity of nICA.
- The framework is a nontrivial integration of
  - statistical learning theory;
  - numerical differentiation;
  - problem-specific design of success metric.
- It is also applicable to other nonlinear mixture learning problems.

Arora, S., Ge, R., Moitra, A., and Sachdeva, S. (2012).
Provable ICA with unknown Gaussian noise, with implications for Gaussian mixtures and autoencoders.
In *Advances in Neural Information Processing Systems*, volume 25.

Comon, P. and Jutten, C. (2010).
Handbook of Blind Source Separation.
Elsevier.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014).
Generative adversarial nets.
In *Proceedings of NIPS 2014*, pages 2672–2680.

Gresele, L., Rubenstein, P. K., Mehrjou, A., Locatello, F., and Schölkopf, B. (2020).
The incomplete Rosetta stone problem: Identifiability results for multi-view nonlinear ICA.
In *Proceedings of UAI 2020*, pages 217–227.

Hyvarinen, A. and Morioka, H. (2016).
Unsupervised feature extraction by time-contrastive learning and nonlinear ica.
In *Advances in Neural Information Processing Systems*, volume 29.

Hyvarinen, A. and Morioka, H. (2017).
Nonlinear ICA of temporally dependent stationary sources.
In *International Conference on Artificial Intelligence and Statistics*, volume 54, pages 460–469.

Hyvärinen, A. and Pajunen, P. (1999).
Nonlinear Independent Component Analysis: Existence and uniqueness results.
Neural Networks, 12(3):429–439.

Hyvarinen, A., Sasaki, H., and Turner, R. (2019).
Nonlinear ICA using auxiliary variables and generalized contrastive learning.
In *International Conference on Artificial Intelligence and Statistics*, pages 859–868.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020).
Variational autoencoders and nonlinear ICA: A unifying framework.
In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR.

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. (2020).
Weakly-supervised disentanglement without compromises.
In *International Conference on Machine Learning*, pages 6348–6359. PMLR.

Lyu, Q. and Fu, X. (2021).
Identifiability-guaranteed simplex-structured post-nonlinear mixture learning via autoencoder.
IEEE Transactions on Signal Processing, 69:4921–4936.

Lyu, Q., Fu, X., Wang, W., and Lu, S. (2022).
Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective.
In *International Conference on Learning Representations*.

Monti, R. P., Zhang, K., and Hyvärinen, A. (2020).
Causal discovery with general non-linear relationships using non-linear ICA.
In *Uncertainty in Artificial Intelligence*, pages 186–195. PMLR.

Oveisi, F., Oveisi, S., Efranian, A., and Patras, I. (2012).
Nonlinear Independent Component Analysis for EEG-Based Brain-Computer Interface Systems.
Independent Component Analysis for Audio and Biosignal Applications, Edited by Ganesh R. Naik, page 165.

Zhang, K. and Hyvärinen, A. (2010).
Distinguishing causes from effects using nonlinear acyclic causal models.
In *Causality: Objectives and Assessment*, pages 157–164. PMLR.

Ziehe, A., Muller, K. ., Nolte, G., Mackert, B. ., and Curio, G. (2000).
Artifact reduction in magnetoneurography based on time-delayed second-order correlations.
IEEE Trans. Biomedical Eng., 47(1):75–87.

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. (2021).
Contrastive learning inverts the data generating process.
In *International Conference on Machine Learning*, pages 12979–12990. PMLR.