Independent Component Analysis (ICA)

Goal is to decompose a multivariate signal into independent non-Gaussian signals

Consider the model of identical signal strength

$$\boldsymbol{X} = \mathbf{A}\boldsymbol{Z}$$

where

- $\mathbf{A} \equiv (\mathbf{a}_1, \ldots, \mathbf{a}_d) \in \mathbb{R}^{d \times d}$ full-rank mixing matrix consisting of orthogonal components
- $\boldsymbol{Z} = (Z_1, \ldots, Z_d)^\top \in \mathbb{R}^d$ non-Gaussian data vector consisting of independent entries
- $Z_1, \ldots, Z_d$ share a fourth moment $\mu_4 \neq 3\mu_2^2$

- More accurately, find the projected directions $\mathbf{a}_1, \ldots, \mathbf{a}_d$ such that data projected onto these directions have maximal statistical independence
- How to actually maximize independence? Maximize the nonnormality

Independent Component Analysis (ICA)

Example: cocktail party problem

- Separate the mixed (sound) signal into sources
- Assumption: different sources are independent
- Question: is it possible to separate the mixed total signal into different sources?

The offline tensorial ICA procedure is as follows:

1. Whiten the data using the Singular Value Decomposition (SVD) to achieve identity covariance
2. Maximize the kurtosis (fourth moment subtracting off by 3) of the signal via the gradient ascent method, which is called the projection pursuit

Tensor Formulation of Independent Component Analysis (ICA)

Let $\mathbf{T} = \mathbb{E}(\boldsymbol{X}^{\otimes 4})$ be the cumulant tensor whose (i, j, k, l)-entry is $\mathbb{E}\,(X_i X_j X_k X_l)$

$$\mathbf{T}(\mathbf{u}, \mathbf{u}, \mathbf{u}, \mathbf{u}) \equiv \mathbb{E}(\mathbf{u}^\top \boldsymbol{X})^4 = 3 + (\mu_4 - 3) \sum_{i=1}^{d} (\mathbf{a}_i^\top \mathbf{u})^4$$

Finding $\mathbf{a}_i$'s can be cast into the solution to the following stochastic optimization problem (Comon, 1994; Frieze et al., 1996)

$$\mathbf{u}^* = \operatorname*{argmin}_{\|\mathbf{u}\|=1} -\mathrm{sign}(\mu_4 - 3) \cdot \mathbb{E}(\mathbf{u}^\top \boldsymbol{X})^4 = \operatorname*{argmin}_{\|\mathbf{u}\|=1} \sum_{i=1}^{d} -(\mathbf{a}_i^\top \mathbf{u})^4$$

Assumption

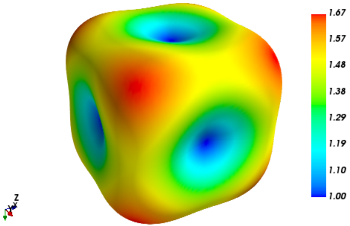
Let $\boldsymbol{X} = \mathbf{A}\boldsymbol{Z}$ where $\mathbf{A} \in \mathbb{R}^{d \times d}$ is the orthonormal mixing matrix, and $\boldsymbol{Z} \in \mathbb{R}^d$ is a random vector that has i.i.d. entries $Z_1, \ldots, Z_d$ satisfying

1. $Z_i, i = 1, \ldots, d$ are independent with identical jth-moment for $j = 1, 2, 4$, denoted as $\mu_j \equiv \mathbb{E} Z_i^j$
2. $\mu_1 = \mathbb{E} Z_i = 0$, $\mu_2 = \mathbb{E} Z_i^2 = 1$, $\mu_4 = \mathbb{E} Z_i^4 \neq 3$
3. Each Z_i has an Orlicz-ψ_2 norm bounded by $\sqrt{3/8}B$

Geometry Landscape of Tensorial ICA

Stationary points are of three types:

- $2d$ local minimizers: $\pm\mathbf{a}_i$ where $\mathbf{a}_i$'s are the column vectors of $\mathbf{A}$
- 2^d local maximizers: $\mathbf{Au}$ where $\mathbf{u} = d^{-1/2}(\pm 1, \cdots, \pm 1)^\top$
- Exponentially $3^d - 2^d - 2d - 1$ many saddle points



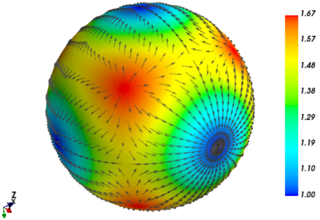


Figure 1: Landscape of Tensorial ICA objectives

Online Tensorial ICA

Online tensorial ICA updates:

$$\mathbf{u}^{(t)} = \Pi \left\{ \mathbf{u}^{(t-1)} + \eta \cdot \mathrm{sign}(\mu_4 - 3) \left(\mathbf{u}^{(t-1)\,\top} \boldsymbol{X}^{(t)} \right)^3 \boldsymbol{X}^{(t)} \right\} \qquad \text{(ICA)}$$

1. $\Pi \mathbf{u} = \mathbf{u}/\|\mathbf{u}\|$ is projection operator onto the unit sphere
2. Uniform initialization
3. Extract one component at each time, can repeat the iteration for $\asymp d \log d$ times to find all tensor components
4. Improved temporal complexity $\mathcal{O}(Nd)$ and spatial complexity $\mathcal{O}(d)$
5. Recent line of nonconvex optimization literature, e.g. (Ge et al., 2015) injects artificial noises or special saddle-escaping treatments. In contrast, our analysis does not require such operations

Three-Stage Analysis and Diffusion Approximations

Uniform initialization is often close to a unstable stationary point (saddle point or local maximizer):

1. Initial Stage: escaping from unstable stationary points, characterized by SDE (unstable Ornstein-Uhlenbeck process), with traverse time $N_1^\eta \asymp d \cdot |\mu_4 - 3|^{-1} \cdot \eta^{-1} \log(\eta^{-1})$
2. Transient Stage: fast deterministic traverse period, characterized by an ODE, with traverse time $N_2^\eta \asymp d \cdot |\mu_4 - 3|^{-1} \cdot \eta^{-1}$
3. Fluctuation Stage: stable oscillation within a small basin around a local minimizer, characterized by SDE (Ornstein-Uhlenbeck process), with traverse time $N_3^\eta \asymp |\mu_4 - 3|^{-1} \cdot \eta^{-1} \log(\eta^{-1})$

Back to Discrete Time

Theorem 4 in (Li and Jordan, 2021)

Suppose the $\mathbf{u}^{(0)}$ is a warm initialization

$$\|\mathbf{u}^{(0)}\|_2 = 1 \quad \text{and} \quad \left|\tan\angle\left(\mathbf{u}^{(0)}, \mathbf{a}_i\right)\right| \leq \frac{1}{\sqrt{3}} \qquad \text{(warm)}$$

and the scaling condition that $T = \widetilde{\Omega}(d)$. Let the stepsize $\eta_2 \asymp \frac{\log T}{|\mu_4 - 3| T}$. Then with probability at least $1 - \epsilon$, the output of (ICA) satisfies

$$\left|\tan\angle\left(\mathbf{u}^{(T)}, \mathbf{a}_i\right)\right| \lesssim \frac{B^4}{|\mu_4 - 3|} \cdot \sqrt{\frac{d \log^2 T}{T}}$$

Theorem 7 in (Li and Jordan, 2021)

Suppose the $\mathbf{u}^{(0)}$ is uniformly sampled from the unit sphere, scaling condition $d \gtrsim \log \epsilon^{-1}$, $T = \widetilde{\Omega}(\epsilon^{-2} d^3)$. Let the stepsize $\eta_1 \asymp \frac{d \log T}{|\mu_4 - 3| T}$. Then with probability at least $1 - \epsilon$, the output of (ICA) satisfies

$$\left|\tan\angle\left(\mathbf{u}^{(T)}, \mathbf{a}_{\mathcal{I}}\right)\right| \lesssim \frac{B^4}{|\mu_4 - 3|} \cdot \sqrt{\frac{d^4 \log^2 T}{T}}$$

Two-Phase Training

Phase I (Theorem 7)

- Initialize $\mathbf{u}^{(0)}$ uniformly at random on unit sphere $\mathcal{D}_1$
- $\eta_1 \asymp \dfrac{d \log T}{T}$
- Update iteration $\mathbf{u}^{(t)}$ for $T/2$ iterates
- $\mathbf{u}^{(T/2)}$ satisfies the warm initialization condition (warm) under the scaling condition $T = \widetilde{\Omega}(d^4)$

Phase II (Theorem 4)

- Warm-initialize by $\mathbf{u}^{(T/2)}$
- $\eta_2 \asymp \dfrac{\log T}{|\mu_4 - 3| T}$
- Update $\mathbf{u}^{(t)}$ for $T/2$ iterates
- Achieves an error bound of $\widetilde{\mathcal{O}}\left(\sqrt{\dfrac{d}{T}}\right)$

Two-Phase Training

Combining the two phases: let the stepsize $\eta_1 \asymp \frac{d \log T}{|\mu_4 - 3| T}$ to obtain a warm initialization, then let $\eta_2 \asymp \frac{\log T}{|\mu_4 - 3| T}$ to achieve an $\widetilde{\mathcal{O}}\left(\sqrt{\frac{d}{T}}\right)$ error bound

Corollary 8 in (Li and Jordan, 2021)

Suppose the $\mathbf{u}^{(0)}$ is uniformly sampled from the unit sphere, scaling condition $d \gtrsim \log \epsilon^{-1}$, $T = \widetilde{\Omega}(\epsilon^{-2} d^4)$. Then with probability at least $1 - \epsilon$, the output of two-phase (ICA) satisfies

$$\left|\tan \angle \left(\mathbf{u}^{(T)}, \mathbf{a}_{\mathcal{I}}\right)\right| \lesssim \frac{B^4}{|\mu_4 - 3|} \cdot \sqrt{\frac{d \log^2 T}{T}}$$

The bound is better than the best SGD analysis in nonconvex optimization

Warm Initialization Analysis: Sketch

We simplify our problem by rotate the iteration $\{\mathbf{v}^{(t)}\}_{t\geq 0}$ such that $\mathbf{a}_i^\top \mathbf{u}^{(t)} = \mathbf{e}_1^\top \mathbf{v}^{(t)}$

$$\mathbf{v}^{(t)} \equiv \mathbf{P}\mathbf{A}^\top \mathbf{u}^{(t)}$$

This rotation ensures that $\pm\mathbf{e}_1$ is the closest independent components pair at initialization and (with high probability) at convergence. We study instead the "tangent" at coordinate k:

$$U_k^{(t)} \equiv \frac{v_k^{(t)}}{v_1^{(t)}}$$

Warm Regions and Warm-Auxilliary Regions

$$\mathcal{D}_{\text{warm}} = \left\{\mathbf{v} : \left|\tan\angle(\mathbf{v}, \mathbf{e}_1)\right| \leq \frac{1}{\sqrt{3}}\right\}, \quad \mathcal{D}_{\text{warm-aux}} = \left\{\mathbf{v} : \left|\tan\angle(\mathbf{v}, \mathbf{e}_1)\right| \leq \frac{1}{\sqrt{2}}\right\}$$

$$\mathcal{T}_x \equiv \inf\left\{t \geq 1 : \mathbf{v}^{(t)} \in \mathcal{D}_{\text{warm-aux}}^c\right\}$$

Warm Initialization Analysis: Sketch

Lemma 5 in (Li and Jordan, 2021)

With high probability, we have

$$\sup_{t \le T \wedge \mathcal{T}_x} \left| U_k^{(t)} - U_k^{(0)} \prod_{s=0}^{t-1} \left[1 - \eta |\mu_4 - 3| \left((v_1^{(s)})^2 - (v_k^{(s)})^2 \right) \right] \right| \le \widetilde{\mathcal{O}}(\eta T^{1/2})$$

This indicates that with high probability, the dynamics of $U_k^{(t)}$ is tightly controlled within a deterministic vessel whose center converges to zero at least exponentially fast

Lemma 3 in (Li and Jordan, 2021)

Under certain scaling condition on η that $\frac{B^8}{|\mu_4 - 3|} \cdot d \cdot LOG^9 \lesssim \eta^{-1}$, with high probability we have for the vessel

$$\left| \tan \angle \left(\mathbf{u}^{(t)}, \mathbf{a}_i \right) \right| \lesssim \underbrace{\left| \tan \angle \left(\mathbf{u}^{(0)}, \mathbf{a}_i \right) \right| \left(1 - \frac{\eta}{3} |\mu_4 - 3| \right)^t}_{\text{exponential mixing}} + \underbrace{\frac{B^4}{|\mu_4 - 3|^{1/2}} \cdot \sqrt{d\eta}}_{\text{noise term}}$$

Uniform Initialization Analysis: Sketch

Similarly as in the warm initialization analysis, we have the following convergence result for the uniform initialization analysis:

Lemma 6 in (Li and Jordan, 2021)

Under certain scaling conditions that $\frac{B^8}{|\mu_4-3|}\cdot d^2 \lesssim \epsilon^2\eta^{-1}$, with high probability at least $1-\epsilon$ we have for an applicable range of t

$$\left|\tan\angle\left(\mathbf{u}^{(t)},\mathbf{a}_i\right)\right| \lesssim \underbrace{\sqrt{\frac{|\mu_4-3|}{B^8}\cdot d\eta^{-1}}\cdot\left(1-\frac{\eta}{2d}|\mu_4-3|\right)^t}_{\text{exponential mixing}} + \underbrace{\frac{B^4}{|\mu_4-3|^{1/2}}\cdot\sqrt{d^3\eta}}_{\text{noise term}}$$

1. Under scaling condition $T=\widetilde{\Omega}(d^3)$, we choose

$$\eta_1 \asymp \frac{d\log T}{|\mu_4-3|T}$$

   to establish our Theorem 7 in (**Li and Jordan**, 2021).
2. Our analysis consists of the "cotangent" and "tangent" parts, as well as an initialization lemma

Conclusion and Future Directions

- Online tensorial ICA algorithm achieves a $\widetilde{O}\left(\sqrt{\frac{d}{T}}\right)$-convergence rate

- Dynamics-based approach outperforms the best existing analyses of online stochastic approximation for tensorial ICA estimation
- Requires no noise-injection steps or specially-designed loops for saddle-point avoidance

- Our dynamics-based analysis can potentially generalize to a broader class of statistical estimation problems that can be cast as nonconvex stochastic optimization problems

- Examples include phase-retrieval, dictionary learning, matrix completion, subspace PCA, sparse models, training deep neural networks, higher-order tensor decomposition

Future Directions

- Further improvements of the convergence rate and scaling conditions or justification of the impossibility (or minimax optimality) of such rates
- Analyzing the mini-batch stochastic approximation algorithm as well as the non-identical kurtosis case for ICA
- Generalizing our analysis of the dynamics of stochastic online algorithms to the nonorthogonal tensor decomposition and over-parameterized cases