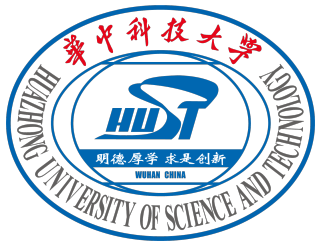
Deep Equilibrium Models are Almost Equivalent to Not-so-deep Explicit Models for High-dimensional Gaussian Mixtures

Zenan Ling [1] Longbo Li [1] Zhanbo Feng [2] Yixuan Zhang [3] Fengzhou [4]
Robert C. Qiu [1] Zhengyu Liao [†1]

[1]Huazhong University of Science and Technology [2]Shanghai Jiao Tong University
[3]Hangzhou Dianzi University [4]Renmin University of China
[†]Correspondence to: zhenyu_liao@hust.edu.cn

2024/07



ICML International Conference On Machine Learning

Deep Equilibrium Models

- **Explicit models:** $z^{(l+1)} = f_\theta^{(l)}(z^{(l)}; x)$, for $l = 0, 1, \cdots, L-1$

[1] Bai, S. *et al.* *Deep Equilibrium Models.* in *NeurIPS* (2019).

Deep Equilibrium Models

- **Explicit models:** $z^{(l+1)} = f_\theta^{(l)}(z^{(l)}; x)$, for $l = 0, 1, \cdots, L-1$
- **Implicit Deep Equilibrium Models (DEQs)**[1]**:** $z^* = f_\theta(z^*; x)$

[1] Bai, S. *et al.* *Deep Equilibrium Models.* in *NeurIPS* (2019).

Deep Equilibrium Models

- **Explicit models:** $z^{(l+1)} = f_\theta^{(l)}(z^{(l)}; x)$, for $l = 0, 1, \cdots, L-1$
- **Implicit Deep Equilibrium Models (DEQs)**[1]**:** $z^* = f_\theta(z^*; x)$
  - a typical "single-layer" implicit model
    $\Rightarrow$ an infinite-depth weight-tied model with an input injection

[1] Bai, S. *et al.* *Deep Equilibrium Models.* in *NeurIPS* (2019).

Deep Equilibrium Models

- **Explicit models:** $z^{(l+1)} = f_\theta^{(l)}(z^{(l)}; x)$, for $l = 0, 1, \cdots, L-1$
- **Implicit Deep Equilibrium Models (DEQs)**[1]**:** $z^* = f_\theta(z^*; x)$
  - a typical "single-layer" implicit model
    $\Rightarrow$ an infinite-depth weight-tied model with an input injection
  - Memory efficient: DEQs solve an equilibrium point directly and compute gradients with *implicit differentiation*.

[1] Bai, S. *et al.* *Deep Equilibrium Models.* in *NeurIPS* (2019).

Deep Equilibrium Models

- **Explicit models:** $z^{(l+1)} = f_\theta^{(l)}(z^{(l)}; x)$, for $l = 0, 1, \cdots, L-1$
- **Implicit Deep Equilibrium Models (DEQs)**[1]**:** $z^* = f_\theta(z^*; x)$
  - a typical "single-layer" implicit model
    $\Rightarrow$ an infinite-depth weight-tied model with an input injection
  - Memory efficient: DEQs solve an equilibrium point directly and compute gradients with *implicit differentiation*.
  - Universality of DEQs: any deep explicit NN can be reformulated as a "single-layer" DEQ.

[1] Bai, S. *et al.* *Deep Equilibrium Models.* in *NeurIPS* (2019).

ICML International Conference On Machine Learning

Deep Equilibrium Models

- **Explicit models:** $z^{(l+1)} = f_\theta^{(l)}(z^{(l)}; x)$, for $l = 0, 1, \cdots, L-1$
- **Implicit Deep Equilibrium Models (DEQs)**[1]**:** $z^* = f_\theta(z^*; x)$
  - a typical "single-layer" implicit model
    $\Rightarrow$ an infinite-depth weight-tied model with an input injection
  - Memory efficient: DEQs solve an equilibrium point directly and compute gradients with *implicit differentiation*.
  - Universality of DEQs: any deep explicit NN can be reformulated as a "single-layer" DEQ.
  - Remarkable success in various tasks, e.g., NPL, CV [2].

[1] Bai, S. *et al. Deep Equilibrium Models.* in *NeurIPS* (2019).

Deep Equilibrium Models

- **Explicit models:** $z^{(l+1)} = f_\theta^{(l)}(z^{(l)}; x)$, for $l = 0, 1, \cdots, L-1$
- **Implicit Deep Equilibrium Models (DEQs)**[1]**:** $z^* = f_\theta(z^*; x)$
  - a typical "single-layer" implicit model
    $\Rightarrow$ an infinite-depth weight-tied model with an input injection
  - Memory efficient: DEQs solve an equilibrium point directly and compute gradients with *implicit differentiation*.
  - Universality of DEQs: any deep explicit NN can be reformulated as a "single-layer" DEQ.
  - Remarkable success in various tasks, e.g., NPL, CV [2].
  - Significant computational overhead: a consequence of root-finding.

[1] Bai, S. *et al.* *Deep Equilibrium Models.* in *NeurIPS* (2019).

ICML International Conference On Machine Learning

Deep Equilibrium Models

- **Explicit models:** $z^{(l+1)} = f_\theta^{(l)}(z^{(l)}; x)$, for $l = 0, 1, \cdots, L-1$
- **Implicit Deep Equilibrium Models (DEQs)**[1]**:** $z^* = f_\theta(z^*; x)$
  - a typical "single-layer" implicit model
    $\Rightarrow$ an infinite-depth weight-tied model with an input injection
  - Memory efficient: DEQs solve an equilibrium point directly and compute gradients with *implicit differentiation*.
  - Universality of DEQs: any deep explicit NN can be reformulated as a "single-layer" DEQ.
  - Remarkable success in various tasks, e.g., NPL, CV [2].
  - Significant computational overhead: a consequence of root-finding.
- **Limited theoretical understanding of DEQs:** the connections and differences between implicit DEQs and explicit models

[1] Bai, S. *et al. Deep Equilibrium Models.* in *NeurIPS* (2019).

Deep Equilibrium Models

- **Explicit models:** $z^{(l+1)} = f_\theta^{(l)}(z^{(l)}; x)$, for $l = 0, 1, \cdots, L-1$
- **Implicit Deep Equilibrium Models (DEQs)**[1]**:** $z^* = f_\theta(z^*; x)$
  - a typical "single-layer" implicit model
    $\Rightarrow$ an infinite-depth weight-tied model with an input injection
  - Memory efficient: DEQs solve an equilibrium point directly and compute gradients with *implicit differentiation*.
  - Universality of DEQs: any deep explicit NN can be reformulated as a "single-layer" DEQ.
  - Remarkable success in various tasks, e.g., NPL, CV [2].
  - Significant computational overhead: a consequence of root-finding.
- **Limited theoretical understanding of DEQs:** the connections and differences between implicit DEQs and explicit models
  - whether general DEQs have advantages over explicit networks, or
  - whether an equivalent explicit NN exists for a given implicit DEQ.

[1] Bai, S. *et al.* *Deep Equilibrium Models.* in *NeurIPS* (2019).

DEQ Models

Vanilla DEQs

Let $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n] \in \mathbb{R}^{p \times n}$ denote the input data, consider a vanilla DEQ with output $f(\boldsymbol{x}_i)$ given by

$$f(\boldsymbol{x}_i) = \boldsymbol{a}^\top \boldsymbol{z}_i^*, \tag{1}$$

where $\boldsymbol{a} \in \mathbb{R}^m$ and $\boldsymbol{z}_i^{(*)} \triangleq \lim_{l \to \infty} \boldsymbol{z}_i^{(l)} \in \mathbb{R}^m$ with

$$\boldsymbol{z}_i^{(l)} = \frac{1}{\sqrt{m}} \phi\left(\sigma_a \boldsymbol{A} \boldsymbol{z}_i^{(l-1)} + \sigma_b \boldsymbol{B} \boldsymbol{x}_i\right) \in \mathbb{R}^m, \text{ for } l \geq 1, \tag{2}$$

for some appropriate initialization $\boldsymbol{z}_i^{(0)}$. Here, $\boldsymbol{A} \in \mathbb{R}^{m \times m}$ and $\boldsymbol{B} \in \mathbb{R}^{m \times p}$ are the DEQ weight parameters, $\sigma_a, \sigma_b \in \mathbb{R}$ are constants, and ϕ is an element-wise activation. Note that $\boldsymbol{z}_i^*$ can also be determined as the equilibrium point of

$$\boldsymbol{z}_i^* = \frac{1}{\sqrt{m}} \phi\left(\sigma_a \boldsymbol{A} \boldsymbol{z}_i^* + \sigma_b \boldsymbol{B} \boldsymbol{x}_i\right). \tag{3}$$

Weight and Activation

Assumption

(1) ***Initialization:*** *$\boldsymbol{a} \in \mathbb{R}^m$, $\boldsymbol{A} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{B} \in \mathbb{R}^{m\times p}$ are initialized with i.i.d. entries of zero mean, unit variance, and finite fourth-order moment;*

(2) ***Weak differentiablility:*** *ϕ is centered and L_1-Lipschitz, and $\max_{k\in\{0,1,2,3,4\}} |\mathbb{E}[\phi^{(k)}(\xi)]| < \infty$;*

(3) ***Variance parameter:*** *$\sigma_a^2 < 1/(4L_1^2)$ and $\sigma_a^2 < 2/(\mathbb{E}[(\phi^2(\tau\xi))''])$ for $\tau > 0$ and $\xi \sim \mathcal{N}(0,1)$.*

Remark

(i) hold for commonly-used initialization;

(ii) hold for commonly-used smooth, *e.g.*, Tanh, and piecewise linear activations, *e.g.*, ReLU and Leaky ReLU;

(iii) guarantee the existence and uniqueness of the equilibrium point.

GMM Data

Assumption (GMM Data)

For $\boldsymbol{x}_i \in \mathcal{C}_a$, $\sqrt{p}\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{C}_a), a \in [K]$. *For* n, p *both large that (i)* $p = \Theta(n)$ *and* $n_a = \Theta(n)$; *(ii)* $\|\boldsymbol{\mu}_a\| = \mathcal{O}(1)$; *(iii) for* $\boldsymbol{C}^\circ \equiv \sum_{a=1}^K \frac{n_a}{n}\boldsymbol{C}_a$ *and* $\boldsymbol{C}_a^\circ \equiv \boldsymbol{C}_a - \boldsymbol{C}^\circ$, *we have* $\|\boldsymbol{C}_a\| = \mathcal{O}(1)$, $\operatorname{tr}\boldsymbol{C}_a^\circ = \mathcal{O}(\sqrt{p})$ *and* $\operatorname{tr}(\boldsymbol{C}_a\boldsymbol{C}_b) = \mathcal{O}(p)$.

Remark

(i) Assumptions of GMM are non-trivial, widely studied in LDA, SVM, and NNs;

(ii) GMM is a universal approximator, can approximate any distribution to an arbitrary error;

(iii) For large p and n, data generated from generative models, *e.g.*, GANs, behaves as GMM [4].

High-dimensional Statistics

- **Label information:** $\boldsymbol{J} \equiv [\boldsymbol{j}_1, \cdots, \boldsymbol{j}_K] \in \mathbb{R}^{n\times K}, \quad [\boldsymbol{j}_a]_i = 1_{\boldsymbol{x}_i \in \mathcal{C}_a}$;
- **Second-order data fluctuation vector:** $\boldsymbol{\psi} \equiv \{\|\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i]\|^2 - \mathbb{E}[\|\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i]\|^2]\}_{i=1}^n \in \mathbb{R}^n$;
- **Second-order GMM statistics:** $\boldsymbol{T} = \{\operatorname{tr}\boldsymbol{C}_a\boldsymbol{C}_b/p\}_{a,b=1}^K \in \mathbb{R}^{K\times K}$, $\boldsymbol{t} = \{\operatorname{tr}\boldsymbol{C}_a^\circ/\sqrt{p}\} \in \mathbb{R}^K$;
- **A fixed point:** $\tau_0 \equiv \sqrt{\operatorname{tr}\boldsymbol{C}^\circ/p}$, and τ_* be the fixed point to the equation

$$\tau_* = \sqrt{\sigma_a^2 \mathbb{E}\left[\phi^2(\tau_*\xi)\right] + \sigma_b^2\tau_0^2}, \quad \xi \sim \mathcal{N}(0,1).$$

Implicit CK and NTK

Conjugate Kernels (CKs) and Neural Tangent Kernels (NTKs): an analytical assessment of the convergence and generalization properties of wide NNs.

Implicit CK and NTK

Conjugate Kernels (CKs) and Neural Tangent Kernels (NTKs): an analytical assessment of the convergence and generalization properties of wide NNs.

Implicit-CK [3]

$\boldsymbol{G}^* = \lim_{l\to\infty} \boldsymbol{G}^{(l)}$, where $\boldsymbol{G}_{ij}^{(l)} = \mathbb{E}_{(\mathsf{u}_l,\mathsf{v}_l)}[\phi(\mathsf{u}_l)\phi(\mathsf{v}_l)]$ with $(\mathsf{u}_l, \mathsf{v}_l) \sim \mathcal{N}\left(0, \begin{bmatrix} \boldsymbol{\Lambda}_{ii}^{(l)} & \boldsymbol{\Lambda}_{ij}^{(l)} \\ \boldsymbol{\Lambda}_{ji}^{(l)} & \boldsymbol{\Lambda}_{jj}^{(l)} \end{bmatrix}\right)$ and $\boldsymbol{\Lambda}_{ij}^{(l)} = \sigma_a^2 \boldsymbol{G}_{ij}^{(l-1)} + \sigma_b^2 \boldsymbol{x}_i^\top \boldsymbol{x}_j$, for $l \geq 1$, and $\boldsymbol{G}_{ij}^{(0)} = (\boldsymbol{z}_i^{(0)})^\top \boldsymbol{z}_j^{(0)}$.

Implicit CK and NTK

Conjugate Kernels (CKs) and Neural Tangent Kernels (NTKs): an analytical assessment of the convergence and generalization properties of wide NNs.

Implicit-CK [3]

$\boldsymbol{G}^* = \lim_{l\to\infty} \boldsymbol{G}^{(l)}$, where $\boldsymbol{G}_{ij}^{(l)} = \mathbb{E}_{(\mathsf{u}_l,\mathsf{v}_l)}[\phi(\mathsf{u}_l)\phi(\mathsf{v}_l)]$ with $(\mathsf{u}_l, \mathsf{v}_l) \sim \mathcal{N}\left(0, \begin{bmatrix} \boldsymbol{\Lambda}_{ii}^{(l)} & \boldsymbol{\Lambda}_{ij}^{(l)} \\ \boldsymbol{\Lambda}_{ji}^{(l)} & \boldsymbol{\Lambda}_{jj}^{(l)} \end{bmatrix}\right)$ and $\boldsymbol{\Lambda}_{ij}^{(l)} = \sigma_a^2 \boldsymbol{G}_{ij}^{(l-1)} + \sigma_b^2 \boldsymbol{x}_i^\top \boldsymbol{x}_j$, for $l \geq 1$, and $\boldsymbol{G}_{ij}^{(0)} = (\boldsymbol{z}_i^{(0)})^\top \boldsymbol{z}_j^{(0)}$.

Implicit-NTK

$\boldsymbol{K}^* = \lim_{l\to\infty} \boldsymbol{K}^{(l)}$, where $\boldsymbol{K}_{ij}^{(l)} = \sum_{h=1}^{l+1} \left(\boldsymbol{G}_{ij}^{(h-1)} \prod_{h'=h}^{l+1} \dot{\boldsymbol{G}}_{ij}^{(h')} \right)$, with $\dot{\boldsymbol{G}}_{ij}^{(l)} = \sigma_a^2 \mathbb{E}_{(\mathsf{u}_l,\mathsf{v}_l)}[\phi'(\mathsf{u}_l)\phi'(\mathsf{v}_l)]$, so that $\boldsymbol{K}_{ij}^* \equiv \boldsymbol{G}_{ij}^*/(1 - \dot{\boldsymbol{G}}_{ij}^*)$.

High-dimensional Equivalents

Theorem (High-dimensional approximation of Implicit-CKs [5])

Let Assumptions 1 and 2 hold, and let the activation ϕ be centered s.t., $\mathbb{E}[\phi(\tau_* \xi)] = 0$ for $\xi \sim \mathcal{N}(0,1)$. It holds that $\left\| \boldsymbol{G}^* - \overline{\boldsymbol{G}} \right\| = \mathcal{O}\left(n^{-1/2}\right)$ where

$$\overline{\boldsymbol{G}} \equiv \alpha_{*,1} \boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{V} \boldsymbol{C}_* \boldsymbol{V}^\top + (\gamma_*{}^2 - \tau_0^2 \alpha_{*,1}) \boldsymbol{I}_n,$$

with $\boldsymbol{V} = \left[\boldsymbol{J}/\sqrt{p},\ \boldsymbol{\psi} \right]$ and $\boldsymbol{C}_* = \begin{bmatrix} \alpha_{*,2} \boldsymbol{t}\boldsymbol{t}^\top + \alpha_{*,3} \boldsymbol{T} & \alpha_{*,2} \boldsymbol{t} \\ \alpha_{*,2} \boldsymbol{t}^\top & \alpha_{*,2} \end{bmatrix}$. Non-negative scalars $\gamma_*, \alpha_{*,1}, \alpha_{*,2}, \alpha_{*,3} \geq 0$ are defined, for $\xi \sim \mathcal{N}(0,1)$, as

$$\begin{aligned} \gamma_* &= \sqrt{\mathbb{E}[\phi^2(\tau_* \xi)]}, \quad \alpha_{*,1} = \frac{\sigma_b^2 \mathbb{E}[\phi'(\tau_* \xi)]^2}{1 - \sigma_a^2 \mathbb{E}[\phi'(\tau_* \xi)]^2}, \\ \alpha_{*,2} &= \frac{\mathbb{E}[\phi''(\tau_* \xi)]^2}{4(1 - \sigma_a^2 \mathbb{E}[\phi'(\tau_* \xi)]^2)} \alpha_{*,4}^2, \ \alpha_{*,3} = \frac{\mathbb{E}[\phi''(\tau_* \xi)]^2 (\sigma_a^2 \alpha_{*,1} + \sigma_b^2)^2}{2(1 - \sigma_a^2 \mathbb{E}[\phi'(\tau_* \xi)]^2)} \end{aligned} \tag{4}$$

with $\alpha_{*,4} = (1 - \frac{\sigma_a^2}{2} \mathbb{E}[(\phi^2(\tau_* \xi))''])^{-1} \sigma_b^2$.

High-dimensional Equivalents

Theorem (High-dimensional approximation of Implicit-NTKs)

Under the same settings and notations of Theorem 1, it holds that $\left\|\boldsymbol{K}^* - \overline{\boldsymbol{K}}\right\| = \mathcal{O}\left(n^{-1/2}\right)$ *where*

$$\overline{\boldsymbol{K}} \equiv \beta_{*,1}\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{V}\boldsymbol{D}_*\boldsymbol{V}^\top + (\kappa_*{}^2 - \tau_0^2\beta_{*,1})\boldsymbol{I}_n,$$

with $\boldsymbol{V}$ *defined in Theorem 1,* $\boldsymbol{D}_* = \begin{bmatrix} \beta_{*,2}\boldsymbol{t}\boldsymbol{t}^\top + \beta_{*,3}\boldsymbol{T} & \beta_{*,2}\boldsymbol{t} \\ \beta_{*,2}\boldsymbol{t}^\top & \beta_{*,2} \end{bmatrix}$, *as well as non-negative scalars* $\kappa_*, \beta_{*,1}, \beta_{*,2}, \beta_{*,3} \geq 0$ *defined as*

$$\kappa_* = \frac{\tau_*}{\sqrt{1 - \sigma_a^2\mathbb{E}[\phi'(\tau_*\xi)^2]}},\ \beta_{*,1} = \frac{\alpha_{*,1}}{1 - \sigma_a^2\mathbb{E}[\phi'(\tau_*\xi)]^2},$$
$$\beta_{*,2} = \frac{\alpha_{*,2}}{1 - \sigma_a^2\mathbb{E}[\phi'(\tau_*\xi)]^2},\ \beta_{*,3} = \frac{\alpha_{*,3} + \beta_{*,1}(\sigma_a^2\mathbb{E}[\phi''(\tau_*\xi)]^2 + \sigma_b^2)\alpha_{*,1}}{1 - \sigma_a^2\mathbb{E}[\phi'(\tau_*\xi)]^2},$$

for $\xi \sim \mathcal{N}(0,1)$.

High-dimensional Equivalents

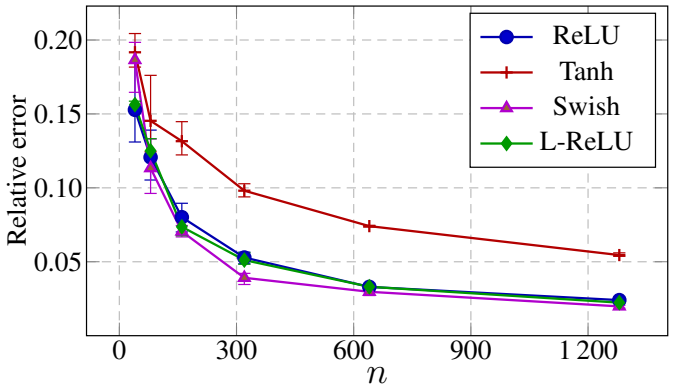


Figure: Evolution of relative spectral norm error $\|\boldsymbol{G}^* - \overline{\boldsymbol{G}}\|/\|\boldsymbol{G}^*\|$ *w.r.t.* sample size n, for DEQs with different activations and $\sigma_a^2 = 0.2$, on two-class GMM, $p/n = 0.8$, $\boldsymbol{\mu}_a = [\boldsymbol{0}_{8(a-1)}; 8; \boldsymbol{0}_{p-8a+7}]$, and $\boldsymbol{C}_a = (1 + 8(a-1)/\sqrt{p})\boldsymbol{I}_p, a \in \{1, 2\}$.

Remark

As, the "equivalent" Implicit-CK and NTK matrices, $\overline{\boldsymbol{G}}$ and $\overline{\boldsymbol{K}}$,

(1) depend on the input GMM data ($\boldsymbol{X}$), their class structure ($\boldsymbol{J}$) and higher-order statistics ($\boldsymbol{t}$ and $\boldsymbol{T}$), *explicitly* ; and

(2) are *independent* of the distribution of the weight matrices $\boldsymbol{A}$ and $\boldsymbol{B}$; and

(3) depend on σ_a^2, σ_b^2, and ϕ *only* via four scalars $\alpha_{*,1-3}, \gamma_*$, and $\beta_{*,1-3}, \kappa_*$, *explicitly*.

High-dimensional Equivalents

Theorem (High-dimensional approximation of Explicit-CKs [6])

Consider a fully-connected NN model $\boldsymbol{x}_i^{(l)} = \frac{1}{\sqrt{m_l}}\sigma_l(\boldsymbol{W}_l\boldsymbol{x}_i^{(l-1)})$, *for* $l = 1, \cdots, L$, *with GMM input. For the corresponding Explicit-CK matrix* $\boldsymbol{\Sigma}^{(l)}$, *it holds that* $\|\boldsymbol{\Sigma}^{(l)} - \overline{\boldsymbol{\Sigma}}^{(l)}\| = \mathcal{O}\big(n^{-1/2}\big)$ *where*

$$\overline{\boldsymbol{\Sigma}}^{(l)} = \widetilde{\alpha}_{l,1}\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{V}\widetilde{\boldsymbol{C}}_l\boldsymbol{V}^\top + (\widetilde{\tau}_l^2 - \tau_0^2\widetilde{\alpha}_{l,1})\boldsymbol{I}_n,$$

with $\boldsymbol{V}$ *defined in Theorem 1,* $\widetilde{\boldsymbol{C}}_l = \begin{bmatrix} \widetilde{\alpha}_{l,2}\boldsymbol{t}\boldsymbol{t}^\top + \widetilde{\alpha}_{l,3}\boldsymbol{T} & \widetilde{\alpha}_{l,2}\boldsymbol{t} \\ \widetilde{\alpha}_{l,2}\boldsymbol{t}^\top & \widetilde{\alpha}_{l,2} \end{bmatrix}$. *Non-negative scalars* $\widetilde{\alpha}_{l,1}, \widetilde{\alpha}_{l,2}, \widetilde{\alpha}_{l,3}$ *defined recursively as* $\widetilde{\alpha}_{0,1} = \widetilde{\alpha}_{0,4} = 1$, $\widetilde{\alpha}_{0,2} = \widetilde{\alpha}_{0,3} = 0$, *and*

$$\widetilde{\alpha}_{l,1} = \mathbb{E}[\sigma_l'(\widetilde{\tau}_{l-1}\xi)]^2\widetilde{\alpha}_{l-1,1},\ \widetilde{\alpha}_{l,2} = \mathbb{E}[\sigma_l'(\widetilde{\tau}_{l-1}\xi)]^2\widetilde{\alpha}_{l-1,2} + \frac{1}{4}\mathbb{E}[\sigma_l''(\widetilde{\tau}_{l-1}\xi)]^2\widetilde{\alpha}_{l-1,4}^2,$$
$$\widetilde{\alpha}_{l,3} = \mathbb{E}[\sigma_l'(\widetilde{\tau}_{l-1}\xi)]^2\widetilde{\alpha}_{l-1,3} + \frac{1}{2}\mathbb{E}[\sigma_l''(\widetilde{\tau}_{l-1}\xi)]^2\widetilde{\alpha}_{l-1,1}^2,$$

with $\widetilde{\alpha}_{l,4} = \mathbb{E}[(\sigma_l^2(\widetilde{\tau}_{l-1}\xi))'']\widetilde{\alpha}_{l-1,4}$, *for* $\xi \sim \mathcal{N}(0,1)$.

Observation

The high-dimensional approximation $\overline{G}$ of the Implicit-CK takes a consistent form with that ($\overline{\boldsymbol{\Sigma}}^{(l)}$) of the Explicit-CK.

High-dimensional Equivalence between Implicit and Explicit NNs

Observation

The high-dimensional approximation $\overline{\boldsymbol{G}}$ of the Implicit-CK takes a consistent form with that ($\overline{\boldsymbol{\Sigma}}^{(l)}$) of the Explicit-CK.

Key idea

Given a DEQ, design activations of an L-layer explicit NN *s.t.* its Explicit-CK $\boldsymbol{\Sigma}^{(L)}$ shares the same coefficients as the Implicit-CK $\boldsymbol{G}^*$, i.e.,
$\widetilde{\tau}_L = \gamma_*,\ \widetilde{\alpha}_{L,i} = \alpha_{*,i},\ i \in \{1, 2, 3\}$.

Observation

The high-dimensional approximation $\overline{\boldsymbol{G}}$ of the Implicit-CK takes a consistent form with that ($\overline{\boldsymbol{\Sigma}}^{(l)}$) of the Explicit-CK.

Key idea

Given a DEQ, design activations of an L-layer explicit NN *s.t.* its Explicit-CK $\boldsymbol{\Sigma}^{(L)}$ shares the same coefficients as the Implicit-CK $\boldsymbol{G}^*$, i.e., $\widetilde{\tau}_L = \gamma_*,\ \widetilde{\alpha}_{L,i} = \alpha_{*,i},\ i \in \{1,2,3\}$.

Implicit- versus Explicit-CK

It follows from Theorem 3 that, for the single-hidden-layer ENN, one *must* have $\widetilde{\alpha}_{1,2} = \frac{1}{2}\widetilde{\alpha}_{1,3}$. On the contrast, $\alpha_{*,2} = \frac{1}{2}\alpha_{*,3}$ does *not* necessarily hold for *all* DEQs. As such, for a given DEQ,

- if $\alpha_{*,2} = \frac{1}{2}\alpha_{*,3}$, a *single-hidden-layer* ENN suffices to match the given DEQ;
- if $\alpha_{*,2} \neq \frac{1}{2}\alpha_{*,3}$, an ENN with *(at least) two hidden layers* is required.

High-dimensional Equivalence between Implicit and Explicit NNs

Observation

The high-dimensional approximation $\overline{\boldsymbol{G}}$ of the Implicit-CK takes a consistent form with that ($\overline{\boldsymbol{\Sigma}}^{(l)}$) of the Explicit-CK.

Key idea

Given a DEQ, design activations of an L-layer explicit NN *s.t.* its Explicit-CK $\boldsymbol{\Sigma}^{(L)}$ shares the same coefficients as the Implicit-CK $\boldsymbol{G}^*$, i.e., $\widetilde{\tau}_L = \gamma_*,\ \widetilde{\alpha}_{L,i} = \alpha_{*,i},\ i \in \{1,2,3\}$.

Implicit- versus Explicit-CK

It follows from Theorem 3 that, for the single-hidden-layer ENN, one *must* have $\widetilde{\alpha}_{1,2} = \frac{1}{2}\widetilde{\alpha}_{1,3}$. On the contrast, $\alpha_{*,2} = \frac{1}{2}\alpha_{*,3}$ does *not* necessarily hold for *all* DEQs. As such, for a given DEQ,

- if $\alpha_{*,2} = \frac{1}{2}\alpha_{*,3}$, a *single-hidden-layer* ENN suffices to match the given DEQ;
- if $\alpha_{*,2} \neq \frac{1}{2}\alpha_{*,3}$, an ENN with *(at least) two hidden layers* is required.

Remark. Results for NTKs can be similarly obtained with our Theorem 2.

Designing Equivalent Explicit NNs via CK matching

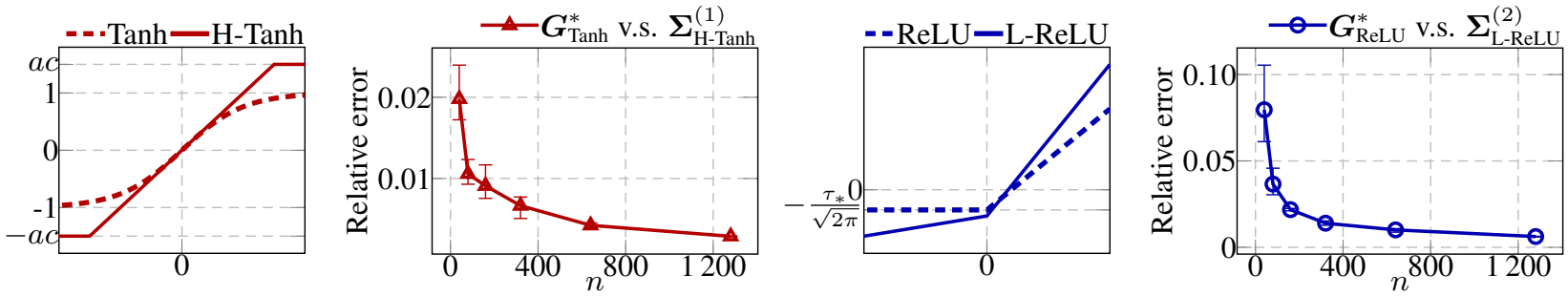


Figure: CKs of implicit DEQs and explicit NNs are close, if the activations of ENNs are constructed according to our Examples.

Example

- For a given Tanh-DEQ, use a *single-hidden-layer* H-Tanh-ENN with

$$\sigma_{\text{H-Tanh}}(x) \equiv ax \cdot 1_{-c \le x \le c} + ac \cdot (1_{x \ge c} - 1_{x \le -c}),$$

  s.t. $\|\boldsymbol{G}^*_{\text{Tanh}} - \boldsymbol{\Sigma}^{(1)}_{\text{H-Tanh}}\| = \mathcal{O}(n^{-1/2})$.

- For a given ReLU-DEQ, use a *two-hidden-layer* L-ReLU-ENN with

$$\sigma^{(l)}_{\text{L-ReLU}}(x) \equiv \max(a_l x, b_l x) - \frac{a_l - b_l}{\sqrt{2\pi}}\widetilde{\tau}_l,\ l = 1, 2,$$

  s.t. $\|\boldsymbol{G}^*_{\text{ReLU}} - \boldsymbol{\Sigma}^{(2)}_{\text{L-ReLU}}\| = \mathcal{O}(n^{-1/2})$.

Experiments

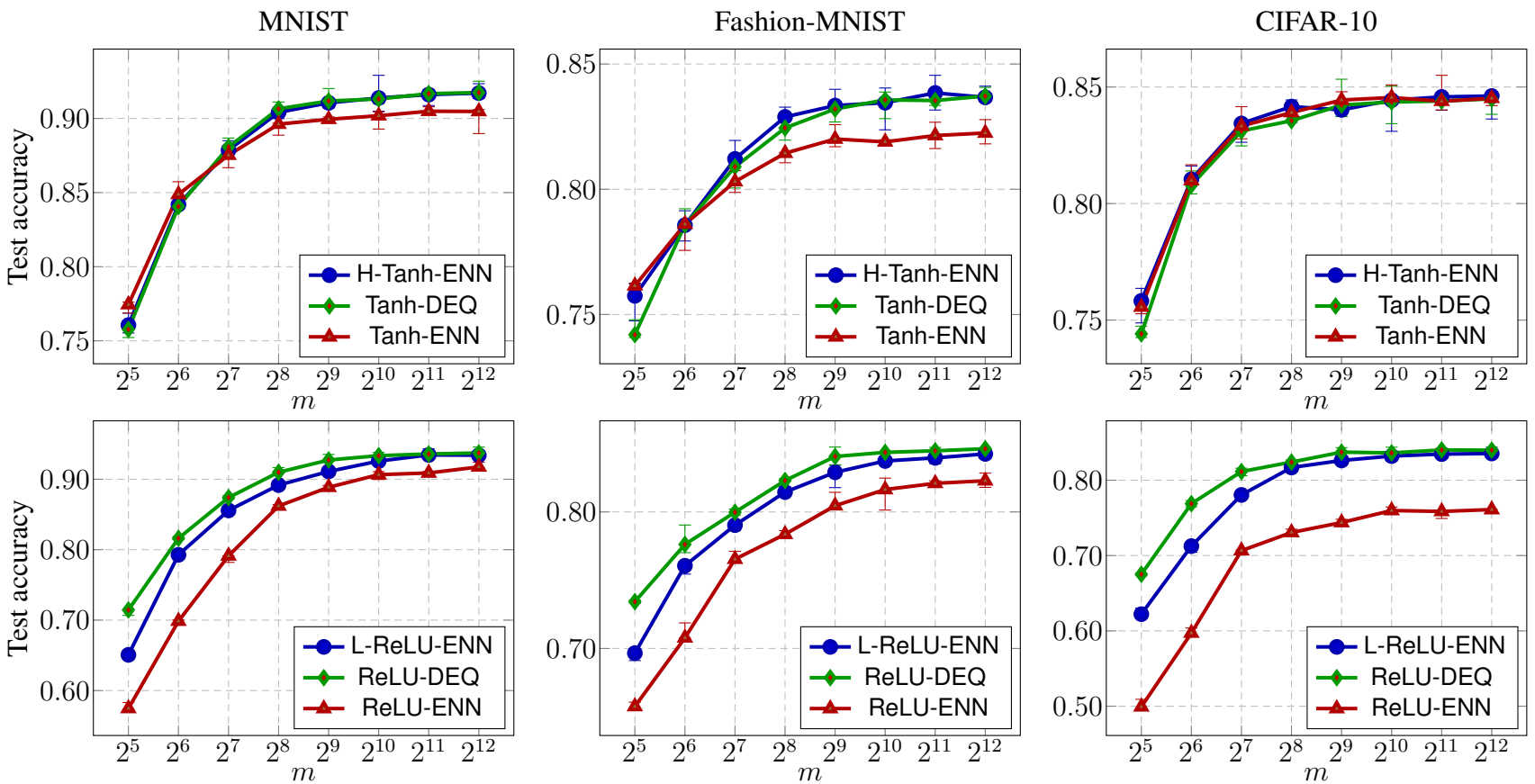


Figure: Classification accuracies of implicit DEQs and explicit models trained with SGD.

Take-away

Take away message:

- for GMM input data, **random matrix theory (RMT)** allows for precise characterization of (the CKs and NTKs of) random explicit and implicit NNs;

Take-away

Take away message:

- for GMM input data, **random matrix theory (RMT)** allows for precise characterization of (the CKs and NTKs of) random explicit and implicit NNs;
- explicit connections between implicit and explicit NNs: high-dimensional "equivalence", making implicit NNs explicit and significantly reducing the computational overhead;

Take-away

Take away message:

- for GMM input data, **random matrix theory (RMT)** allows for precise characterization of (the CKs and NTKs of) random explicit and implicit NNs;
- explicit connections between implicit and explicit NNs: high-dimensional "equivalence", making implicit NNs explicit and significantly reducing the computational overhead;
- future work: beyond the "lazy" CK/NTK regime $\Rightarrow$ feature learning;

ICML International Conference On Machine Learning

Take-away

Take away message:

- for GMM input data, **random matrix theory (RMT)** allows for precise characterization of (the CKs and NTKs of) random explicit and implicit NNs;
- explicit connections between implicit and explicit NNs: high-dimensional "equivalence", making implicit NNs explicit and significantly reducing the computational overhead;
- future work: beyond the "lazy" CK/NTK regime $\Rightarrow$ feature learning;
- future work: extends to another typical implicit models, Neural ODEs, e.g., diffusion models, providing theoretical understanding and accelerating the sampling process.

Reference and Acknowledgement

Reference:

1. Bai, S., Kolter, J. Z. & Koltun, V. *Deep Equilibrium Models.* in *NeurIPS* (2019).
2. Xie, X., Wang, Q., Ling, Z., Liu, G. & Lin, Z. Optimization Induced Equilibrium Networks: An Explicit Optimization Perspective for Understanding Equilibrium Models. *TPAMI* (2022).
3. Ling, Z., Xie, X., Wang, Q., Zhang, Z. & Lin, Z. *Global convergence of over-parameterized deep equilibrium models.* in *AISTATS* (2023).
4. Seddik, M. E. A., Louart, C., Tamaazousti, M. & Couillet, R. *Random matrix theory proves that deep learning representations of gan-data behave as gaussian mixtures.* in *ICML* (2020).
5. Ling, Z. *et al.* *Deep Equilibrium Models are Almost Equivalent to Not-so-deep Explicit Models for High-dimensional Gaussian Mixtures.* in *ICML* (2024).
6. Gu, L., Du, Y., Yuan, Z., Qiu, R. & Liao, Z. *" Lossless" Compression of Deep Neural Networks: A High-dimensional Neural Tangent Kernel Approach.* in *NeurIPS* (2022).

Thank you! Q & A?