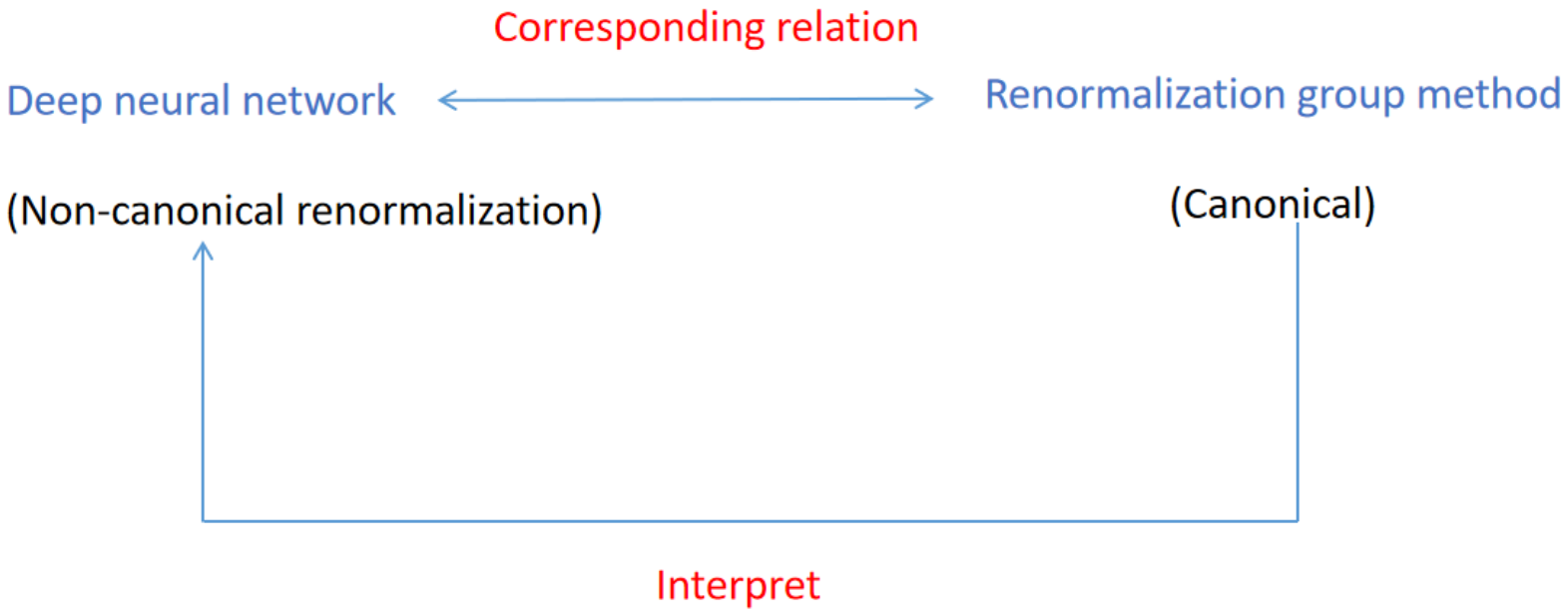
The Interpretability of Deep Neural Networks Based on Renormalization Group

Zigeng Xia
Peking University
Joint work with Fuzhou Gong

19th Workshop on Markov Processes and Related Topics
July 23rd, 2024

Framework



Data:

- One dimensional Ising model on finite lattice
- A class of continuous data from the exponential family

Framework

Backgrounds:

- Deep Neural Network (DNN)
  - Structure
  - Loss function and training algorithm
  - Interpretability
- Renormalization group method
  - Idea and conditions
  - Real space RG on Ising model

Deep neural network

- Deep Neural Network (DNN): The most widely used machine learning models today. Have shown excellent performance in various fields such as computer vision, natural language processing, speech recognition and synthesis, recommendation systems, finance, bioinformatics, etc.
- Characteristic: Strong feature extraction capabilities on large-scale datasets.

Deep neural network

- Fully connected neural network: Every two neurons are connected between two adjacent layers.

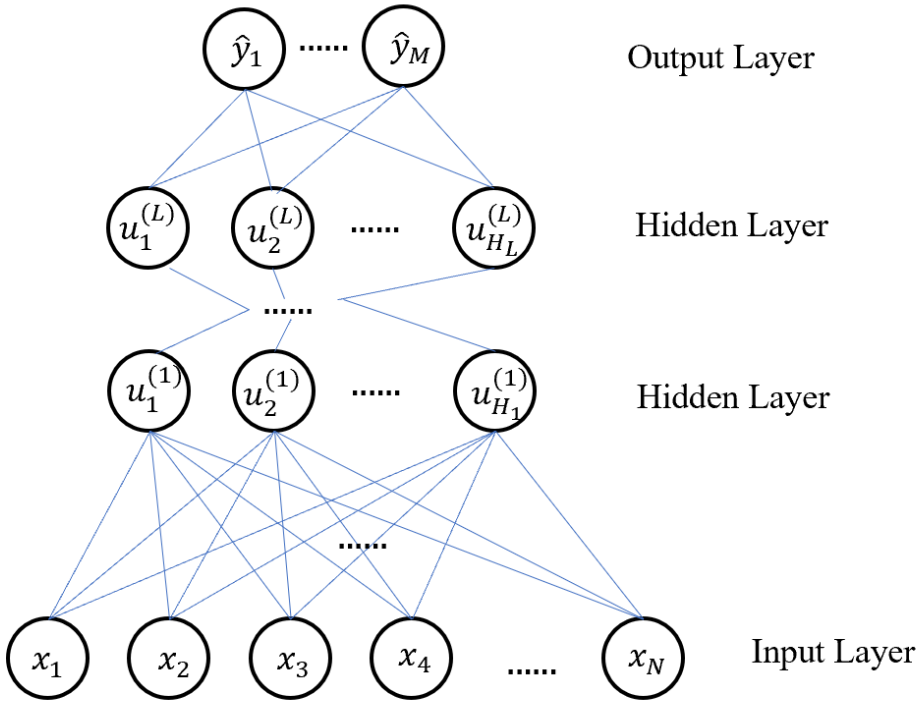


Figure: Structure of the fully connected neural network

Training process

- Loss function: F denotes neural network,

$$L(\mathbf{W}) = \frac{1}{2}||F(\mathbf{W}) - F(\mathbf{W}^*)||_2^2 + \lambda||\mathbf{W} - \mathbf{W}^*||_2^2.$$

  $\mathbf{W}^*$: the unique optimal point of parameters, $L(\mathbf{W}^*) = 0$.

- Back propagation: find the derivative of the loss function with respect to the network parameters.
- Stochastic optimization algorithm :Simulated annealing, Stochastic gradient descent (SGD) and its variants.

Simulated annealing

Continuation of simulated annealing (Stephan, Hoffman, Blei, 2017):

$$d\mathbf{W}^{(t)} = -\nabla L(\mathbf{W}^{(t)})\,dt + \sqrt{\eta_t}\,dB^{(t)},\ t{\geq}0.$$
$$\mathbf{W}^{(0)}{\sim}P_0(\mathbf{W}^{(0)}).$$

Mainly according to (Holley, Kusuoka, Stroock, 1989), we have:

Lemma 1

Assume that loss function $L(\mathbf{W}){\geq}0$, $\mathbf{W}^{(t)}, t{\in}[0,+\infty)$ satisfy the equation above, $L(\mathbf{W})$ has unique global minimum point $\mathbf{W}=\mathbf{W}^*$ and $L(\mathbf{W}^*)=0$. Then there exists function η_t of t, s.t. for $\delta>0$:

$$\mathbb{P}(L(\mathbf{W}^{(t)}){\geq}\delta){\leq}\epsilon(\delta,t).$$

Where $\epsilon(\delta,t){\to}0$ when $t{\to}\infty$.

Interpretability

- Definition: The capability of providing explanations for the model's decision outcomes and the internal mechanisms that lead to those decisions, aligning with human reasoning processes.
- Significance
  - (1) Safety
  - (2) Human trust in the model
- Limitations of current research: Mainly targeting a specific application problem or model structure.

Renormalization group method (RG)

- First proposed by Kadanoff in 1966, and then developed by Wilson.
- Phase transitions and critical phenomena
- Idea of RG
  - Microscopic variables $\Rightarrow$ macroscopic properties
  - Ising model: coarse graining procedure
    continuous models: scaling transformation
- Condition: The state function and its functional form with respect to characteristic parameters must remain unchanged: same physical system.

RG of one dimensional Ising model

Hamiltonian

$$\mathscr{H} = -J\sum_{i=1}^{N} x_i x_{i+1} - h\sum_{i=1}^{N} x_i.$$

$J > 0$: coupling constant, periodic boundary condition:

$$x_{N+1} \triangleq x_1 .$$

The probability distribution and the partition function:

$$\mathbb{P}(x_1, x_2, \cdots, x_N) = \frac{1}{\mathscr{Z}} e^{-\frac{1}{k_B T}\mathscr{H}} = \frac{1}{\mathscr{Z}} e^{\frac{J}{k_B T}\sum_{i=1}^{N} x_i x_{i+1} + \frac{h}{k_B T}\sum_{i=1}^{N} x_i}.$$

$$\mathscr{Z} = \sum_{\{x_i\}=\pm 1} e^{\frac{J}{k_B T}\sum_{i=1}^{N} x_i x_{i+1} + \frac{h}{k_B T}\sum_{i=1}^{N} x_i}.$$

RG of one dimensional Ising model

Renormalization transformation: coarse graining procedure

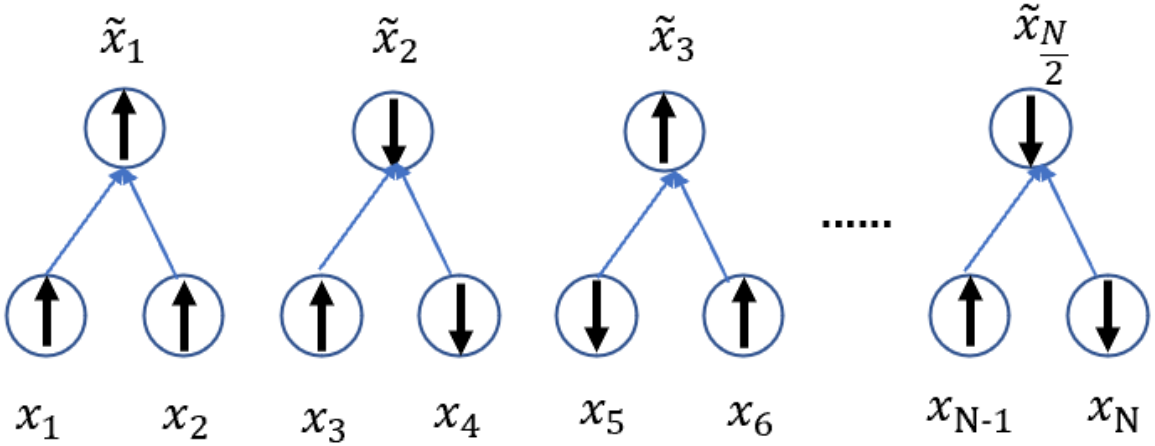


Figure: Coarse graining of Ising model

RG of one dimensional Ising model

For convenience, we consider $k_B T = 1$ and $h = 0$:

$$\begin{aligned} \mathbb{P}(x_1, x_2, \cdots, x_N) &= \frac{1}{Z(J)} e^{J\sum_{i=1}^{N} x_i x_{i+1}}, x_i \in \{\pm 1\}, i = 1, 2, \cdots, N, x_{N+1} \triangleq x_1, \\ Z(J) &= \sum_{\{x_i\}=\pm 1} e^{J\sum_{i=1}^{N} x_i x_{i+1}}. \end{aligned} \tag{1}$$

The renormalization transformation is (assume $N = 2^n$):

$$\tilde{x}_i = x_{2i}, i = 1, 2, \cdots, \frac{N}{2},$$

RG of one dimensional Ising model: canonical results

$$Z(J) = \sum_{\{x_i\}=\pm 1} e^{J\sum_{i=1}^{N} x_i x_{i+1}} = \sum_{\{\tilde{x}_j\}=\pm 1} e^{J_0 + \tilde{J}\sum_{j=1}^{\frac{N}{2}} \tilde{x}_i \tilde{x}_{i+1}}.$$

Where J_0 is a nonzero constant must be introduced, represents the non singular part of the Hamiltonian. Then

$$J_0 = \frac{N}{2}(\log(4\cosh(2J)), \tilde{J} = \frac{1}{2}\log(\cosh(2J))$$

$$J^* = \frac{1}{2}\log(\cosh(2J^*)). \tag{2}$$

From above equation J has two fixed point $J^* = 0$ and $J^* = \infty$, where $J^* = 0$ is the stable fixed point.

Research status

- The first notable work is that in (Mehta, Schwab, 2014) they constructs a mapping between the variational renormalization group and restricted Boltzmann machines (RBM).
  - Variational renormalization group: a method to calculate specifically the approximate RG transform according to the restrictions of the free energy.
  - RBM model:

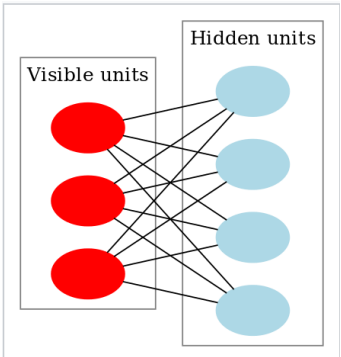


Figure: Restricted Boltzmann machine

Research status

Some other works based on (Mehta, Schwab, 2014):

- (Iso, Shiba, Yokoo, 2018) Trained an RBM model on Ising model and observed the temperature parameter of model converging toward a critical point, demonstrating self-similarity akin to RG method at the critical point.
- (Caticha, 2019) Considered the infinite-depth limit, approximately aligning neural networks with RG under continuous depth.
- (Koch, Koch, Cheng, 2020) Compared the similarity between a layer of RBM and a step of coarse graining in RG by computing the correlation function$< v_i h_j >$between the visible and hidden layers.
- (Erdmenger, Grosvenor, Jefferson, 2022) Compared the relative entropy in real space RG coarse graining and neural network forward propagation on Ising model.
- (Shukla, Thakur, 2022) (Lahoche, Samary, Tamaazousti, 2022) Compared RG with autoencoders, and principal component analysis (PCA).

Our main improvements

- (1): Conceptual illustration $\Rightarrow$ Theoretically rigorous mapping
- (2): The experimental analysis based on visual observation $\Rightarrow$ A theoretical proof of equivalence
- (3): RBM models $\Rightarrow$ general DNNs

Problem

- (1): How to construct the corresponding framework of DNN and RG.
  Input data and output $\Rightarrow$ physical systems
- (2): With the above corresponding relationship, how to prove the equivalence of DNN and RG on concrete data.

We discuss one dimensional Ising model on finite lattice and a class of models obey the exponential family distribution.

Fully connected neural network without hidden layers

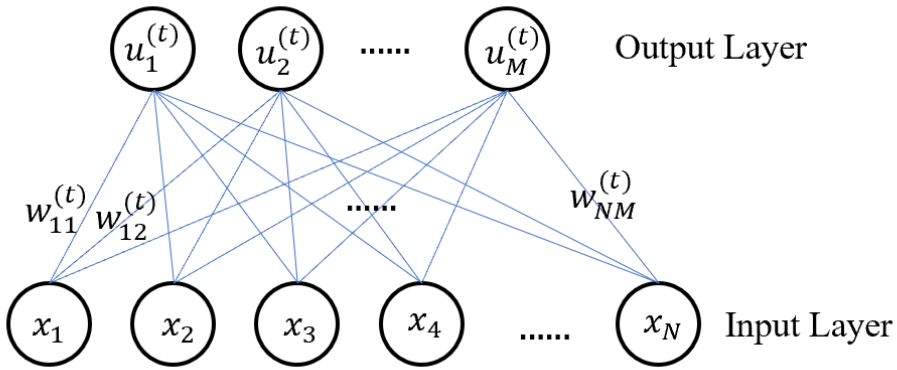


Figure: Fully connected neural network without hidden layers

Data and training algorithm

- Input data: One dimensional Ising model on finite lattice:

$$\mathbb{P}(x_1, x_2, \cdots, x_N) = \frac{1}{Z(J)} e^{J\sum_{i=1}^{N} x_i x_{i+1}}, x_i \in \{\pm 1\}, i = 1, 2, \cdots, N, x_{N+1} \triangleq x_1,$$

$$Z(J) = \sum_{\{x_i\}=\pm 1} e^{J\sum_{i=1}^{N} x_i x_{i+1}}.$$

- Loss function:

$$L(\mathbf{W}^{(t)}) = \frac{1}{2} \sum_{\{x_i\}=\pm 1} [\sum_{j=1}^{M} (f(x_1 w_{1j}^{(t)} + x_2 w_{2j}^{(t)} + \cdots + x_N w_{Nj}^{(t)} + b_j^{(t)})$$
$$- f(x_1 w_{1j}^{*} + x_2 w_{2j}^{*} + \cdots + x_N w_{Nj}^{*} + b_j^{*}))^2]$$
$$+ \lambda \sum_{i=1}^{N} \sum_{j=1}^{M} (w_{ij}^{(t)} - w_{ij}^{*})^2 + \lambda \sum_{j=1}^{M} (b_j^{(t)} - b_j^{*})^2.$$

- Training algorithm: Simulated annealing.

Method to construct the correspondence relation

Criteria:

- (1) State function: partition function, value remains unchanged.
- (2) Functional form of Hamiltonian with respect to the fundational part of the system and the coupling constant remains unchanged.

Equation: Let $y_j = f(x_1 w_{1j}^{(t)} + x_2 w_{2j}^{(t)} + \cdots + x_N w_{Nj}^{(t)} + b_j^{(t)}), j = 1, \cdots, M,$

$$Z(J) \equiv e^{Ng(t)} \mathbb{E}[\sum_{\{x_i\}=\pm 1} e^{J_t \sum_{j=1}^{M} y_j y_{j+1}}], \tag{3}$$

$$Z(J) = \sum_{\{x_i\}=\pm 1} e^{J \sum_{i=1}^{N} x_i x_{i+1}}.$$

$g(t)$ is a non-singular and non-stochastic function of time t with $g(0) = 0$.
Correspondence relation: $\{x_i\} \Leftrightarrow \{y_j\}$, $J \Leftrightarrow J_t$, $g(t) \Leftrightarrow J_0$.

Results

$$Z(J) \equiv e^{Ng(t)} \mathbb{E}[\sum_{\{x_i\}=\pm 1} e^{J_t \sum_{j=1}^{M} y_j y_{j+1}}], \tag{3}$$

Theorem 2 (Gong, X., 2024+; arXiv:2212.00005)

With respect to above network structure and training algorithm, η_t is chosen according to Lemma 1, then $\exists g(t) \in C^{\infty}$, s.t. when equation (3) holds for $t \in [0, \infty)$, we have $\lim_{t\to\infty} J_t = J^* = 0$. J^* denotes the stable fixed point of J in real space renormalization group of one-dimensional Ising model on finite lattice.

Conclusion

- $\lim_{t\to\infty} J_t$: macroscopic property of the output system in neural network
- J^*: macroscopic property of one dimensional Ising model on finite lattice, calculated by real space RG

$\lim_{t\to\infty} J_t = J^*$: two methods extract same macroscopic property

Remarks

- Generalize: stochastic gradient descent (SGD) algorithm
- Stable fixed point: finite system, no phase transition

Fully connected deep neural network

- Network structure: Multiple hidden layers.
- Loss function:

$$L(\mathbf{W},\mathbf{B}) = \sum_{\{x_i\}=\pm 1}[\sum_{j=1}^{M}(f(u_1^{(L)}w_{1j}^{(L+1)} + u_2^{(L)}w_{2j}^{(L+1)} + \cdots + u_{H_L}^{(L)}w_{H_Lj}^{(L+1)} + b_j^{(L+1)}) - f(u_1^{*(L)}w_{1j}^{*(L+1)} + u_2^{*(L)}w_{2j}^{*(L+1)} + \cdots + u_{H_L}^{*(L)}w_{H_Lj}^{*(L+1)} + b_j^{*(L+1)}))^2] + \lambda(\sum_{l=1}^{L+1}\sum_{i=1}^{H_{l-1}}\sum_{j=1}^{H_l}(w_{ij}^{(l)} - w_{ij}^{*(l)})^2 + \sum_{l=1}^{L+1}\sum_{i=1}^{H_l}(b_i^{(l)} - b_i^{*(l)})^2).$$

- Training algorithm: Simulated annealing ($\Theta^{(t)} \triangleq (\mathbf{W}^{(t)}, \mathbf{B}^{(t)})$),

$$d\Theta^{(t)} = -\nabla L(\Theta^{(t)})\,dt + \sqrt{\eta_t}\,dB^{(t)}.$$

Method to construct the correspondence relation

Output of the network (value in the last layer):

$$y_j = f(u_1^{(L)} w_{1j}^{(L+1)} + u_2^{(L)} w_{2j}^{(L+1)} + \cdots + u_{H_L}^{(L)} w_{H_L j}^{(L+1)} + b_j^{(L+1)}), j = 1, 2, \cdots, M.$$

Equation:

$$Z(J) \equiv e^{Ng(t)} \mathbb{E}[\sum_{\{x_i\}=\pm 1} e^{J_t \sum_{j=1}^{M} y_j y_{j+1}}]. \tag{4}$$

Boundary condition:

$$\begin{aligned} & f(u_1^{(L)} w_{1,M+1}^{(L+1)} + u_2^{(L)} w_{2,M+1}^{(L+1)} + \cdots + u_{H_L}^{(L)} w_{H_L,M+1}^{(L+1)} + b_{M+1}^{(L+1)}) \\ =& f(u_1^{(L)} w_{11}^{(L+1)} + u_2^{(L)} w_{21}^{(L+1)} + \cdots + u_{H_L}^{(L)} w_{H_L 1}^{(L+1)} + b_1^{(L+1)}). \end{aligned}$$

Results

$$Z(J) \equiv e^{Ng(t)} \mathbb{E}[\sum_{\{x_i\}=\pm 1} e^{J_t \sum_{j=1}^{M} y_j y_{j+1}}]. \tag{4}$$

Theorem 3 (Gong, X., 2024+; arXiv:2212.00005)

With respect to above network structure and training algorithm, η_t is chosen according to Lemma 1, then $\exists g(t) \in C^{\infty}$, such that when equation (4) holds for $t \in [0, \infty)$, we have $\lim_{t \to \infty} J_t = J^* = 0$. J^* denotes the fixed point of J in real space renormalization group of one-dimensional Ising model on finite lattice.

Continuous distribution: exponential family

Exponential family of distributions:

$$\mathbb{P}(\mathbf{x};\theta) = C(\theta)e^{\sum_{r=1}^{R}-\theta_r T_r(\mathbf{x})}h(\mathbf{x}), \mathbf{x} = (x_1, x_2, \cdots, x_N)\in\mathbb{R}^N$$

Here we choose:

$$T_r(\mathbf{x}) = (\sum_{j,k=1}^{N} a_{jk}x_j x_k + \sum_{j=1}^{N} b_j x_j)^r,$$

Where $a_{jk}\in\mathbb{R}, j, k = 1, \cdots, N$ and $b_j, j = 1, \cdots, N$ are constants, s.t. matrix $A = (a_{jk})$ symmetric and non negative definite. $h(\mathbf{x})\equiv 1$ and

$$C(\theta) = \frac{1}{\mathcal{Z}(\theta)}, \mathcal{Z}(\theta) = \int e^{\sum_{r=1}^{R}-\theta_r(\sum_{j,k=1}^{N} a_{jk}x_j x_k+\sum_{j=1}^{N} b_j x_j)^r} dx,$$

the distribution density of the data is:

$$\mathbb{P}(\mathbf{x};\theta) = \frac{1}{\mathcal{Z}(\theta)} e^{\sum_{r=1}^{R}-\theta_r(\sum_{j,k=1}^{N} a_{jk}x_j x_k+\sum_{j=1}^{N} b_j x_j)^r}.$$

Transformation of the data

Transform the distribution of data into a form of Landau-Ginzburg-Wilson Hamiltonian (Zinn-Justin, 2007).

- (1) First through the transformation $\tilde{x}_j = x_j + f_j, j = 1, \cdots, N$, change the Hamiltonian into a polynomial with only even terms using undetermined coefficients.

$$\sum_{r=1}^{R} \theta_r (\sum_{j,k=1}^{N} a_{jk} x_j x_k)^r.$$

- (2) Transform the quadratic form into a diagonal form.

$$\sum_{j,k=1}^{N} a_{jk} x_j x_k = \sum_{i=1}^{N} \mu_i y_i^2.$$

- (3) Transform the Hamiltonian into translation-invariant form through dilation:

$$\mathscr{H} = \sum_{r=1}^{R} \theta_r (\sum_{j,k=1}^{N} a_{jk} x_j x_k)^r = \sum_{r=1}^{R} \theta_r (\sum_{i=1}^{N} y_i^2)^r.$$

RG on continuous field

Continuous field $\sigma(x)$ defined on $\mathbb{R}^N$, Hamiltonian of the system:

$$\mathscr{H}(\sigma) = \mathscr{H}(\sigma(x)), x \in \mathbb{R}^N.$$

The probability of a configuration σ satisfies

$$\mathbb{P}(\sigma) \propto e^{-\mathscr{H}(\sigma)}.$$

The partition function of the system is:

$$\mathscr{Z}(H) = \int e^{-\mathscr{H}(\sigma)}[d\sigma],$$

(where $[d\sigma]$ is only a formal sign since the uniform measure does not exist on the field space), the $n-$points correlation function is:

$$\mathscr{G}^{(n)}(\sigma(x_1), \sigma(x_2), \cdots, \sigma(x_N)) \triangleq \frac{1}{\mathscr{Z}} \int \sigma(x_1)\sigma(x_2)\cdots\sigma(x_N) e^{-\mathscr{H}(\sigma)}[d\sigma].$$

RG on continuous field

RG transformation: dilation parameter λ, $\mathscr{H}_\lambda(\sigma)$ is the Hamiltonian under transformation.

$$\mathscr{H}_{\lambda=1}(\sigma) = \mathscr{H}(\sigma).$$

Criterion: renormalization group equation (RGE) based on correlation function.

$$\begin{aligned}&\mathscr{G}_\lambda^{(n)}(\sigma(x_1), \sigma(x_2), \cdots, \sigma(x_N)) - Z^{-\frac{n}{2}}(\lambda)\mathscr{G}^{(n)}(\sigma(\lambda x_1), \sigma(\lambda x_2), \cdots, \sigma(\lambda x_N)) \\ &= \mathscr{R}_\lambda^{(n)}(\sigma(x_1), \sigma(x_2), \cdots, \sigma(x_N)).\end{aligned}$$

Where

$$\mathscr{G}_\lambda^{(n)}(\sigma(x_1), \sigma(x_2), \cdots, \sigma(x_N)) \triangleq \frac{1}{\mathscr{Z}_\lambda}\int \sigma(x_1)\sigma(x_2)\cdots\sigma(x_N)e^{-\mathscr{H}_\lambda(\sigma)}[d\sigma].$$

- $Z(\lambda)$: function of λ (choice of $Z(\lambda)$ is a difficult point).
- $\mathscr{R}_\lambda^{(n)}$: Schwartz function of λ.

RG on continuous field

The form of RG transformation:

$$\mathscr{H}_{\lambda}(\sigma(x)) = \mathscr{H}_{\lambda=1}(Z(\lambda)^{\frac{1}{2}}\sigma(\frac{x}{\lambda})).$$

When $\lambda \to \infty$, if $\mathscr{H}_{\lambda}(\sigma)$ has a limit $\mathscr{H}^{*}(\sigma)$: the fixed-point Hamiltonian, its characteristic parameters : the fixed point of corresponding characteristic parameters in the system.

Compatibility with the real space RG

Proposition (Cui, Gong, 2022)

For the Ising model without external field, using coarse graining as the RG transformation, then for one step, the dilation parameter $\lambda = 2$, the RGE with respect to the 2-points correlation function is (here choose $Z(\lambda) = 1$):

$$\mathscr{G}_{\lambda}^{(2)}(x_1, x_2) - \mathscr{G}(\lambda x_1, \lambda x_2) = \mathscr{R}_{\lambda}(x_1, x_2). \tag{5}$$

when the renormalization process goes on, let $\lambda = 2^n$, $n \to \infty$ then the remainder $\mathscr{R}_{\lambda}(x_1, x_2)$ tends to 0 exponentially.

RG of the input data

We define a general form of the Hamiltonian:

$$\tilde{\mathscr{H}} :\Sigma\times\mathbb{R}^R\times\mathbb{N}\to\mathbb{R}$$
$$\tilde{\mathscr{H}}(\sigma,\theta,N)=\int_{\mathbb{R}^N}\sum_{r=1}^{R}\theta_r(\sum_{i=1}^{N}\sigma_i(x)^2)^r dx.$$

Where the space of field Σ is the subspace of functions $\mathbb{R}^N\to\mathbb{R}^N$.

Input data:

$$\mathbb{P}(x)=\frac{1}{\mathscr{Z}(\theta)}e^{-\sum_{r=1}^{R}\theta_r(\sum_{i=1}^{N}x_i^2)^r}.$$

Corresponding field:

$$\sigma :\mathbb{R}^N\to\mathbb{R}^N$$
$$\sigma_i(x)=x_i,\ i=1,2,\cdots,N.$$

RG of the input data

RG transformation:

$$x_i \mapsto Z^{\frac{1}{2}}(\lambda) \cdot \frac{x_i}{\lambda}, i = 1, 2, \cdots, N. \tag{6}$$

Transformation of characteristic parameters θ_r are:

$$\theta_r \mapsto \theta_r \cdot Z(\lambda)^r \cdot \frac{1}{\lambda^{2r}} \triangleq \theta_r^{\lambda}, r = 1, 2, \cdots, R$$

When $\lambda \to \infty$, the limit of the characteristic parameters are the fixed points:

$$\theta_r^* \triangleq \theta_r \cdot \lim_{\lambda \to \infty} Z(\lambda)^r \cdot \frac{1}{\lambda^{2r}}$$

RG of the input data

RGE:

$$\frac{1}{\mathcal{Z}_\lambda}\int (x^{(1)})^{\vec{\kappa}_1}(x^{(2)})^{\vec{\kappa}_2}\cdots(x^{(n)})^{\vec{\kappa}_n} e^{-\sum_{r=1}^{R}\theta_r(Z(\lambda)\frac{1}{\lambda^2})^r(\sum_{i=1}^{N}x_i^2)^r}\,dx_1\,dx_2\cdots dx_N$$

$$- Z^{-\frac{n}{2}}(\lambda)\frac{1}{\mathcal{Z}}\int \lambda^{\sum_{l=1}^{n}|\vec{\kappa}_l|}\cdot(x^{(1)})^{\vec{\kappa}_1}(x^{(2)})^{\vec{\kappa}_2}\cdots(x^{(n)})^{\vec{\kappa}_n} e^{-\sum_{r=1}^{R}\theta_r(\sum_{i=1}^{N}x_i^2)^r}\,dx_1\,dx_2\cdots dx_N$$

$$= \mathscr{R}_\lambda^{(n)}(x^{(1)}, x^{(2)}, \cdots, x^{(n)}).$$

$\lambda\to\infty$, condition: $\mathcal{Z}_\lambda = \mathcal{Z}$,

$$\int (x^{(1)})^{\vec{\kappa}_1}(x^{(2)})^{\vec{\kappa}_2}\cdots(x^{(n)})^{\vec{\kappa}_n} e^{-\sum_{r=1}^{R}\theta_r^*(\sum_{i=1}^{N}x_i^2)^r}\,dx_1\,dx_2\cdots dx_N$$

$$- \lim_{\lambda\to\infty}(Z^{-\frac{n}{2}}(\lambda)\lambda^{\sum_{l=1}^{n}|\vec{\kappa}_l|})\int\cdot(x^{(1)})^{\vec{\kappa}_1}(x^{(2)})^{\vec{\kappa}_2}\cdots(x^{(n)})^{\vec{\kappa}_n} e^{-\sum_{r=1}^{R}\theta_r(\sum_{i=1}^{N}x_i^2)^r}\,dx_1\,dx_2\cdots dx_N$$

$$= 0.$$

Assume: non trivial parameter fixed points

Forward propagation as a RG transformation

General form of the Hamiltonian:

$$\tilde{\mathscr{H}} : \Sigma \times \mathbb{R}^R \times \mathbb{N} \to \mathbb{R}$$

$$\tilde{\mathscr{H}}(\sigma, \theta, N) = \int_{\mathbb{R}^N} \sum_{r=1}^{R} \theta_r \Big(\sum_{i=1}^{N} \sigma_i(x)^2 \Big)^r dx.$$

For the out put of neural network,

$$\phi : \mathbb{R}^N \to \mathbb{R}^M$$

$$\phi_j(x) = F_j(x; \mathbf{W}), j = 1, 2, \cdots, M.$$

Where $F(x; \mathbf{W})$ is the network. e.g. the neural network without hidden layers:

$$F_j(x; \mathbf{W}) = f(x_1 w_{1j} + x_2 w_{2j} + \cdots + x_N w_{Nj} + b_j), \tag{7}$$

$$\tilde{\mathscr{H}}(\phi, \theta, M) = \int_{\mathbb{R}^N} \sum_{r=1}^{R} \tilde{\theta}_r^t \Big(\sum_{j=1}^{M} F_j(x; \mathbf{W}t)^2 \Big)^r dx,$$

partition function:

$$\mathscr{Z}_t = \mathbb{E}_{\mathbf{W}^t} \Big[\int e^{Ng(t) - \int_{\mathbb{R}^N} \sum_{r=1}^{R} \tilde{\theta}_r^t (\sum_{j=1}^{M} F_j(x; \mathbf{W}^t)^2)^r dx} [d\phi] \Big].$$

Forward propagation as a RG transformation

Rewrite:

$$\mathscr{Z}_t = \mathbb{E}_{\mathbf{W}^t}\Big[\int_{\Sigma} e^{Ng(t)-\int\sum_{r=2}^{R}\tilde{\theta}_r^t(\sum_{j=1}^{M}\phi_j(x,\mathbf{W}t)^2)^r dx} d\mathcal{N}(0,\frac{1}{\tilde{\theta}_1^t})\Big].$$

Non-canonical RG transformation:

$$\tilde{\theta}_r^t = \theta_r \cdot \tilde{Z(t)}^r \cdot \frac{1}{t^{2r}}.$$

RGE of n-points correlation function:

$$\mathbb{E}_{\mathbf{W}^t}\Big[\int \phi(x^{(1)},\mathbf{W}^t)^{\vec{\kappa}_1}\phi(x^{(2)},\mathbf{W}^t)^{\vec{\kappa}_2}\cdots\phi(x^{(n)},\mathbf{W}^t)^{\vec{\kappa}_n} e^{Ng(t)-\int\sum_{r=2}^{R}\tilde{\theta}_r^t(\sum_{j=1}^{M}\phi_j(x,\mathbf{W}^t)^2)^r dx} d\mathcal{N}(0,\frac{1}{\tilde{\theta}_1^t})\Big]$$
$$-\int \phi(x^{(1)},\mathbf{W}^1)^{\vec{\kappa}_1}\phi(x^{(2)},\mathbf{W}^1)^{\vec{\kappa}_2}\cdots\phi(x^{(n)},\mathbf{W}^1)^{\vec{\kappa}_n} e^{Ng(1)-\int\sum_{r=2}^{R}\theta_r(\sum_{j=1}^{M}\phi_j(x,\mathbf{W}^1)^2)^r dx} d\mathcal{N}(0,\frac{1}{\theta_1})$$
$$=\mathscr{R}_t^{(n)}(x^{(1)},x^{(2)},\cdots,x^{(n)}).$$

Method to construct the correspondence relation and result

Theorem 4 (Gong, X., 2024+)

Assuming characteristic parameters $\theta_r > 0, r = 1, 2, \cdots, R$ *have fixed points* $\theta_r^* \geq 0, r = 1, 2, \cdots, R$ *under the RG method based on scaling transformation. The structure of the neural network is the fully connected network, training algorithm is simulated annealing and* $\eta(t)$ *is chosen through Lemma 1. Then* $\exists g(t) \in C^{\infty}[1, +\infty)$ *s.t.*

$$\mathscr{Z}_t = \mathscr{Z},$$

RGE for output holds true, and

$$\tilde{\theta}_r^* \triangleq \lim_{t \to \infty} \tilde{\theta}_r^t = \theta_r^*. r = 1, 2, \cdots, R.$$

Conclusion

As an exploration direction in the theoretical study of interpretability of deep neural networks, constructing a corresponding relationship between them and renormalization group theory: intuitive, theoretically valid.

- One dimensional Ising model on finite lattice
- A class of data in exponential family

Future works

More general form of Hamiltonian:

- Analysis of the stability of the fixed point:

$$\mathscr{H}(x;\theta^{\lambda}) = \mathscr{H}(x;\theta^{*}) + \Delta\theta\cdot(\mathscr{H}_{\theta}^{'}(x;\theta^{*})) + O(\Delta\theta^{2}).$$

  Guess the fixed point of $\mathscr{H} = \sum_{r=1}^{R}\theta_r(\sum_{i=1}^{N}x_i^2)^r$: Gaussian

- Functional renormalization group method

Thanks!

Remarks

Stochastic gradient descent (SGD) algorithm:

$$d\mathbf{W}^{(t)} = -\nabla \hat{L}(\mathbf{W}^{(t)})\,dt + \sqrt{\frac{\eta_t}{S}}\sigma(\mathbf{W}^{(t)})\,dB^{(t)},\ t \geq 0,$$

$$\sigma(\mathbf{W}^{(t)})\sigma(\mathbf{W}^{(t)})^T = \Sigma(\mathbf{W}^{(t)})$$

$$= \frac{1}{2^N}\sum_{\{x_i\}=\pm 1}[(\nabla \hat{l}(\mathbf{W}^{(t)}, \{x_i\}) - \nabla \hat{L}(\mathbf{W}^{(t)}))(\nabla \hat{l}(\mathbf{W}^{(t)}, \{x_i\}) - \nabla \hat{L}(\mathbf{W}^{(t)}))^T],$$

$$\hat{l}(\mathbf{W}^{(t)}, \{x_i\}) = [\sum_{j=1}^{M}(f(x_1 w_{1j}^{(t)} + x_2 w_{2j}^{(t)} + \cdots + x_N w_{Nj}^{(t)}) - f(x_1 w_{1j}^{*} + x_2 w_{2j}^{*} + \cdots + x_N w_{Nj}^{*}))^2],\ \mathbf{W}^{(0)}$$