Entropy of Multilayer Generalized Linear Models: Proof of the Replica Formula with the Adaptive Interpolation Method

Jean Barbier, Clément Luneau, Nicolas Macris

Laboratoire de Théorie des Communications (LTHC) - EPFL

Problem presentation

L-layer Generalized Linear Model (GLM)

Multilayer model leveraged to simulate prototypical settings of deep supervised learning on synthetic datasets [1]

Input signal $\mathbf{X}^0 \in \mathbb{R}^{n_0}$ with $X_1^0, \dots, X_{n_0}^0 \overset{\text{i.i.d.}}{\sim} P_0$

Hidden layers Input feedforwarded through $L-1$ unobserved layers.

For $1 \leq \ell \leq L$, ℓ^{th} layer

❶ is fed with $\mathbf{X}^{\ell-1} \in \mathbb{R}^{n_{\ell-1}}$ ❷ forwards $\mathbf{X}^{\ell} \in \mathbb{R}^{n_\ell}$

$$\forall i \in \{1, \dots, n_\ell\} : X_i^\ell = \varphi_\ell\left(\left[\frac{\mathbf{W}_\ell \mathbf{X}^{\ell-1}}{\sqrt{n_{\ell-1}}}\right]_i, \mathbf{A}_{\ell,i}\right)$$

Layer characterized by

- *known* activation function $\varphi_\ell : \mathbb{R} \times \mathbb{R}^{k_\ell} \to \mathbb{R}$ with $k_\ell \in \mathbb{N}$
- *known* Gaussian matrix $\mathbf{W}_\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ with entries i.i.d. $\mathcal{N}(0,1)$
- *unknown* stochastic stream $\mathbf{A}_{\ell,1}, \dots, \mathbf{A}_{\ell,n_\ell} \in \mathbb{R}^{k_\ell} \overset{\text{i.i.d.}}{\sim} P_{A_\ell}$

Observations L^{th} layer output with AWGN $\mathbf{Z} \in \mathbb{R}^{n_L}$

$$\forall i \in \{1, \dots, n_L\} : Y_i = \varphi_L\left(\left[\frac{\mathbf{W}_L \mathbf{X}^{L-1}}{\sqrt{n_{L-1}}}\right]_i, \mathbf{A}_{L,i}\right) + \sqrt{\Delta}\, Z_i$$

Bayesian estimation

Hamiltonian

$$\mathcal{H}(\mathbf{x}, \mathbf{a}_1, \dots, \mathbf{a}_L; \mathbf{Y}, \mathbf{W}) = \frac{1}{2\Delta} \sum_{\mu=1}^{n_L} \left(Y_\mu - \varphi_L\left(\left[\frac{\mathbf{W}_L \mathbf{x}^{L-1}}{\sqrt{n_{L-1}}}\right]_\mu, \mathbf{a}_{L,\mu}\right)\right)^2$$

where

$$\forall \ell \in \{1, \dots, L-1\} : \mathbf{x}^\ell \equiv \mathbf{x}^\ell(\mathbf{x}, \mathbf{a}_1, \dots, \mathbf{a}_\ell) = \varphi_\ell\left(\frac{\mathbf{W}_\ell \mathbf{x}^{\ell-1}}{\sqrt{n_{\ell-1}}}, \mathbf{a}_\ell\right), \ \mathbf{x}^0 \equiv \mathbf{x}$$

Joint posterior distribution given *quenched* variables $\mathbf{Y}, \mathbf{W}$:

$$dP(\mathbf{x}, \mathbf{a}_1, \dots, \mathbf{a}_L | \mathbf{Y}, \mathbf{W}) = \frac{1}{\mathcal{Z}(\mathbf{Y}, \mathbf{W})} dP_0(\mathbf{x}) \prod_{\ell=1}^{L} dP_{A_\ell}(\mathbf{a}_\ell)\, e^{-\mathcal{H}(\mathbf{x}, \mathbf{a}_1, \dots, \mathbf{a}_L; \mathbf{Y}, \mathbf{W})}$$

Averaged free entropy

$$f_{n_0} = \mathbb{E}\left[\frac{\ln \mathcal{Z}(\mathbf{Y}, \mathbf{W})}{n_0}\right] = \frac{1}{n_0} \mathbb{E}\left[\ln \int dP_0(\mathbf{x}) \prod_{\ell=1}^{L} dP_{A_\ell}(\mathbf{a}_\ell)\, e^{-\mathcal{H}(\mathbf{x}, \mathbf{a}_1, \dots, \mathbf{a}_L; \mathbf{Y}, \mathbf{W})}\right]$$

$\equiv$ **mutual information between input and output of the multilayer network**

Theorem: Replica formula for the free entropy

In the *high-dimensional regime*

$$n_0, \dots, n_L \to +\infty \text{ such that } \forall\, \ell \in \{0, \dots, L\} : \frac{n_\ell}{n_0} \to \tilde{\alpha}_\ell > 0$$

we have

$$\lim_{n_0 \to \infty} f_{n_0} = \sup_{q_{L-1} \in [0, \rho_{L-1}]} \inf_{r_{L-1} \geq 0} \dots \sup_{q_0 \in [0, \rho_0]} \inf_{r_0 \geq 0} f_{\mathrm{RS}}\left(\{q_\ell, r_\ell\}_{\ell=0}^{L-1}; \{\rho_\ell\}_{\ell=0}^{L-1}\right)$$

with

$$\begin{cases} \rho_0 = \mathbb{E}[X^2],\ X \sim P_0 \\ \rho_\ell = \mathbb{E}[\varphi_\ell^2(T_\ell, \mathbf{A}_\ell)],\ T_\ell \sim \mathcal{N}(0, \rho_{\ell-1}) \perp\!\!\!\perp \mathbf{A}_\ell \sim P_{A_\ell} \text{ for } \ell = 1, \dots, L-1 \end{cases}$$

Replica-symmetric potential

$$f_{\mathrm{RS}}\left(\{q_\ell, r_\ell\}_{\ell=0}^{L-1}; \{\rho_\ell\}_{\ell=0}^{L-1}\right) = \psi_{P_0}(r_0) + \sum_{\ell=1}^{L-1} \tilde{\alpha}_\ell \Psi_{\varphi_\ell}(q_{\ell-1}, r_\ell; \rho_{\ell-1}) + \tilde{\alpha}_L \Psi_{P_{\mathrm{out,L}}}(q_{L-1}; \rho_{L-1}) - \sum_{\ell=1}^{L} \tilde{\alpha}_{\ell-1} \frac{r_{\ell-1} q_{\ell-1}}{2}$$

- First obtained in [2]
- Per layer: 1 free entropy of a scalar problem to evaluate $\to$ computable integral!

$\psi_{P_0}(r) \leftarrow$ observation $Y = \sqrt{r}\, X + Z$ with $X \sim P_0$, $Z \sim \mathcal{N}(0,1)$

$\Psi_{\varphi_\ell}(q, r; \rho) \leftarrow$ observations V, $Y = \sqrt{r}\varphi_\ell(\sqrt{q}\, V + \sqrt{\rho - q}\, U, \mathbf{A}) + Z$ with $U, V, Z \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $\mathbf{A} \sim P_{A_\ell}$

$\Psi_{P_{\mathrm{out,L}}}(q; \rho) \leftarrow$ observations V, $Y = \varphi_L(\sqrt{q}\, V + \sqrt{\rho - q}\, U, \mathbf{A}) + \sqrt{\Delta}\, Z$ with $U, V, Z \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $\mathbf{A} \sim P_{A_L}$

Induction proof via adaptive interpolation

Base case Formula for 1-layer GLMs proved in [3]

Induction hypothesis Theorem proved for $(L-1)$-layer GLMs

Interpolation problems

Continuity of inference problems parametrized by $t \in [0,1]$. 2 kinds of observations:

$$\begin{cases} \mathbf{Y}_t &= \varphi_L(\mathbf{S}_t, \mathbf{A}_L) + \sqrt{\Delta}\, \mathbf{Z} \\ \mathbf{Y}'_t &= \sqrt{r\, t}\, \mathbf{X}^{L-1} + \mathbf{Z}' \end{cases}$$

with

$$\mathbf{S}_t = \sqrt{\frac{1-t}{n_{L-1}}} \mathbf{W}_L \mathbf{X}^{L-1} + \sqrt{\int_0^t q(v) dv}\, \mathbf{V} + \sqrt{\int_0^t \big(\rho_{L-1} - q(v)\big) dv}\, \mathbf{U}$$

- *Freely chosen interpolation function* $q : [0,1] \to [0, \rho_{L-1}]$
- $V_1, \dots, V_{n_L}, U_1, \dots, U_{n_L} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $\mathbf{V}$ **known** while $\mathbf{U}$ **unknown**
- $\mathbf{Z}' \in \mathbb{R}^{n_{L-1}}$ AWGN

Free entropy of time-t interpolation problem: $f_{n_0}(t) = \frac{1}{n_0} \mathbb{E}\left[\ln \mathcal{Z}_t(\mathbf{Y}_t, \mathbf{Y}'_t, \mathbf{W}, \mathbf{V})\right]$

❶ Known formula at t=1:

$\mathbf{Y}_{t=1}$, $\mathbf{Y}'_{t=1}$ independent observations & $\mathbf{Y}_{t=1}$ observations of n_L independent scalar channels & $\mathbf{Y}'_{t=1}$ noisy observation of the last layer of a $(L-1)$-layer GLM

$\Rightarrow f_{n_0}(1)$ in high-dimensional regime given by **induction hypothesis**

❷ Problem of interest at t=0:

$$f_{n_0}(0) = f_{n_0} - \frac{n_{L-1}}{2n_0}$$

❸ Going from t=1 to t=0: Fundamental Theorem of Analysis

$$f_{n_0} = f_{n_0}(1) + \frac{n_{L-1}}{2n_0} - \int_0^1 \frac{df_{n_0}(t)}{dt} dt$$

Choosing the interpolation function

Goal: Cancelling remainder $R(t)$ in derivative

$$\frac{df_{n_0}(t)}{dt} = \frac{n_{L-1}}{n_0}\left(\frac{rq(t)}{2} - \frac{r\rho_{L-1}}{2}\right) + R(t) + \underbrace{o_{n_0}(1)}_{\text{vanishes uniformly in t}}$$

$R(t)$ satifies $\left|\int_0^1 dt\, R(t)\right| \lesssim \sqrt{\int_0^1 dt\, \mathbb{E}\left[\left\langle \big(Q_{L-1} - q(t)\big)^2 \right\rangle_t\right]}$

- $\langle - \rangle_t$ Gibbs measure associated to interpolating Hamiltonian $\mathcal{H}_t(\mathbf{x}, \mathbf{a}_1, \dots, \mathbf{a}_{L-1}, \mathbf{u}; \mathbf{Y}_t, \mathbf{Y}'_t, \mathbf{W}, \mathbf{V})$
- Requires concentration of overlap $Q_{L-1} = \frac{1}{n_{L-1}} \left(\mathbf{x}^{L-1}\right)^\intercal \cdot \mathbf{X}^{L-1}$

Natural choice

$q_{n_0}^{(r)}$ solution to differential equation $q(t) = \mathbb{E}\left[\langle Q_{L-1} \rangle_t\right]$

Sufficient condition for overlap concentration to $\mathbb{E}\left[\langle Q_{L-1} \rangle_t\right]$ - See [4]

$$\forall t \in [0,1] : \mathbb{E}\left[\left(\frac{\ln \mathcal{Z}_t}{n_0} - \mathbb{E}\left[\frac{\ln \mathcal{Z}_t}{n_0}\right]\right)^2\right] \leq \frac{C\left(\{\varphi_\ell, \tilde{\alpha}_\ell\}_{\ell=1}^L, P_0\right)}{n_0}$$

Concentration

- Proved for 1-layer, 2-layer & 3-layer GLMs [1, 3]
- Conjectured for $L > 3$

After canceling the remainder

$$f_{n_0} = \sup_{q_{L-2} \in [0, \rho_{L-2}]} \inf_{r_{L-2} \geq 0} \dots \sup_{q_0 \in [0, \rho_0]} \inf_{r_0 \geq 0} f_{\mathrm{RS}}\left(\{q_\ell, r_\ell\}_{\ell=0}^{L-2}, \int_0^1 q_{n_0}^{(r)}(v) dv, r; \{\rho_\ell\}_{\ell=0}^{L-1}\right) + o_{n_0}(1)$$

Last equation in the limit $n_0 \to +\infty \Rightarrow$ **Replica Formula (Theorem)**

References

[1] M. Gabrié, A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, and L. Zdeborová, "Entropy and mutual information in models of deep neural networks," *Arxiv e-print 1805.09785*, 2018.

[2] A. Manoel, F. Krzakala, M. Mézard, and L. Zdeborová, "Multi-layer generalized linear estimation," *ArXiv e-print 1701.06981*, 2017.

[3] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová, "Phase transitions, optimal errors and optimality of message-passing in generalized linear models," *ArXiv e-print 1708.03395*, Aug. 2017.

[4] J. Barbier and N. Macris, "The adaptive interpolation method: A simple scheme to prove replica formulas in bayesian inference," *Arxiv e-print 1705.02780*, 2017.