Convergence Rates of Variational Inference in Sparse Deep Learning

Badr-Eddine Chérief-Abdellatif
CREST - ENSAE - Institut Polytechnique de Paris

International Conference on Machine Learning 2020

Motivation

- Generalization properties of DL are not well understood.
- Recent works addressed the estimation of smooth functions in a nonparametric regression framework using sparse DNNs.
- From a Bayesian point of view, the posterior distribution using some sparsity-inducing prior concentrates at the near-minimax rate when estimating Hölder smooth functions.
- Variational inference is wisely used in practice to compute the exact posterior.

Question

Do Bayesian neural networks retain the same properties when using **variational inference** ?

Contributions of this work

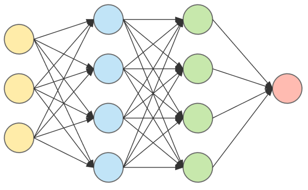
Answer

Yes ! If the variational family is well chosen.

- Nonasymptotic generalization bound based on PAC-Bayes theory ensuring the consistency of approximations of Bayesian DNNs, along with rates of convergence.
- Posterior concentration at near-minimax rates for a wise choice of the architecture when estimating Hölder smooth functions.
- Extension of the oracle inequality when the optimization algorithm incurs error as measured by its effect on the ELBO.
- Selection of the network architecture using the ELBO criterion.

Nonparametric regression

- $X_i \sim \mathcal{U}([-1,1]^d)$,
- $Y_i = f_0(X_i) + \zeta_i$,
- $\zeta_i \sim \mathcal{N}(0, \sigma^2)$.



Deep neural networks

- Depth $L \geq 3$, width $D \geq d$, sparsity $S \leq T$ where T is the total number of connections.
- Parameter $\theta = \{(A_1, b_1), ..., (A_L, b_L)\}$.
- $f_\theta(x) = A_L \rho(A_{L-1} ... \rho(A_1 x + b_1) + ... + b_{L-1}) + b_L$.

Bayesian approach

Spike-and-Slab prior π

- First, we select uniformly at random a number S of active connections, and denote $\gamma_t = 1$ if connection t is active and $\gamma_t = 0$ otherwise.
- Then, for $t = 1, ..., T$: $\begin{cases} \theta_t | \gamma_t = 1 \sim \mathcal{N}(0,1) \\ \theta_t | \gamma_t = 0 \sim \delta_{\{0\}} \end{cases}$

V. Rockova, N. Polson. Posterior Concentration for Sparse Deep Learning. *NeurIPS*, 2018.

Bayesian approach

Spike-and-Slab prior π

- First, we select uniformly at random a number S of active connections, and denote $\gamma_t = 1$ if connection t is active and $\gamma_t = 0$ otherwise.
- Then, for $t = 1, ..., T$: $\begin{cases} \theta_t | \gamma_t = 1 \sim \mathcal{N}(0,1) \\ \theta_t | \gamma_t = 0 \sim \delta_{\{0\}} \end{cases}$

V. Rockova, N. Polson. Posterior Concentration for Sparse Deep Learning. *NeurIPS*, 2018.

Tempered posterior distribution ($0 < \alpha < 1$)

By tempering Bayes' rule using a temperature parameter α, we get the tempered posterior distribution :

$$\pi_{n,\alpha}(d\theta) \propto \exp\left(- \frac{\alpha}{2\sigma^2} \sum_{i=1}^{n} (Y_i - f_\theta(X_i))^2 \right) \pi(d\theta).$$

Sparse variational inference

Idea of variational inference

Choose a family $\mathcal{F}_{S,L,D}$ of probability distributions over θ and approximate $\pi_{n,\alpha}$ by a distribution in $\mathcal{F}_{S,L,D}$:

$$\tilde{\pi}_{n,\alpha} = \arg \min_{q \in \mathcal{F}_{S,L,D}} KL(q, \pi_{n,\alpha}).$$

Sparse variational inference

Idea of variational inference

Choose a family $\mathcal{F}_{S,L,D}$ of probability distributions over θ and approximate $\pi_{n,\alpha}$ by a distribution in $\mathcal{F}_{S,L,D}$:

$$\tilde{\pi}_{n,\alpha} = \arg \min_{q \in \mathcal{F}_{S,L,D}} KL(q, \pi_{n,\alpha}).$$

Spike-and-Slab variational family $\mathcal{F}_{S,L,D}$

- Hyperparameters of the family : discrete distribution q_γ on the set of T-dimensional 0-1 vector γ with S nonzero entries, mean-variance pairs $\{(m_t, s_t^2)\}_{t=1,\dots,T}$.
- First, we select $\gamma \sim q_\gamma$.
- Then, for $t = 1, \dots, T$: $\begin{cases} \theta_t | \gamma_t = 1 \sim \mathcal{N}(m_t, s_t^2) \\ \theta_t | \gamma_t = 0 \sim \delta_{\{0\}} \end{cases}$

Generalization error bound

Generalization error

$$R(\tilde{\pi}_{n,\alpha}) = \mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \tilde{\pi}_{n,\alpha}(d\theta)\right]$$

Generalization error bound

Generalization error

$$R(\tilde{\pi}_{n,\alpha}) = \mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \tilde{\pi}_{n,\alpha}(d\theta)\right]$$

Theorem

$$R(\tilde{\pi}_{n,\alpha}) \leq \frac{2}{1-\alpha} \inf_{\|\theta^*\|_\infty \leq B} \|f_{\theta^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1 + \frac{\sigma^2}{\alpha}\right) r_n$$

where the rate of convergence r_n is of order $\frac{LS}{n}\log(BD)$.

The upper bound on the generalization error is composed of the approximation error of f_0 (i.e. the bias) and the estimation error r_n (i.e. the variance).

Some remarks

Main idea of the proof : PAC-Bayes theory + (extended) prior mass condition

$$\pi\left(\theta / \|f_\theta - f_{\theta^*}\|^2 \leq r_n\right) \geq e^{-nr_n}.$$

P. Alquier, J. Ridgway. Concentration of Tempered Posteriors and of their Variational Approximations. *to appear in The Annals of Statistics*, 2017.

We recover exactly the rate of convergence of the empirical risk minimizer for DNNs which is obtained using different proof techniques (by computing the r_n-local covering entropy).

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *to appear in The Annals of Statistics*, 2017.

T. Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces : optimal rate and curse of dimensionality *ICLR*, 2019.

Theorem

Assume that f_0 is β-Hölder smooth with $0 < \beta < d$, that the activation function is ReLU, and that

$$L \asymp \log n,$$

$$D \asymp n^{\frac{d}{2\beta+d}} / \log n,$$

$$S \asymp n^{\frac{d}{2\beta+d}}.$$

Then the variational approximation $\tilde{\pi}_{n,\alpha}$ concentrates at the (near)-minimax rate $r_n = n^{\frac{-2\beta}{2\beta+d}}$ in the sense that :

$$\tilde{\pi}_{n,\alpha}\left(\theta \in \Theta_{S,L,D} \,/\, \|f_\theta - f_0\|_2^2 > M_n \cdot n^{\frac{-2\beta}{2\beta+d}} \cdot \log^2 n\right) \xrightarrow[n\to+\infty]{} 0$$

in probability as $n \to +\infty$ for any $M_n \to +\infty$.

Effect of an optimization error

VI alternatively maximizes the ELBO :

$$\text{ELBO}(q) = -\frac{\alpha}{2\sigma^2}\sum_{i=1}^{n}\int (Y_i - f_\theta(X_i))^2 q(d\theta) - KL(q, \pi_{n,\alpha}).$$

When considering an algorithm $(\tilde{\pi}_{n,\alpha}^{(j)})_j$ for computing the ideal approximation $\tilde{\pi}_{n,\alpha}$, there is an additional term in the generalization error bound :

Effect of an optimization error

VI alternatively maximizes the ELBO :

$$\text{ELBO}(q) = -\frac{\alpha}{2\sigma^2}\sum_{i=1}^{n}\int (Y_i - f_\theta(X_i))^2 q(d\theta) - KL(q, \pi_{n,\alpha}).$$

When considering an algorithm $(\tilde{\pi}_{n,\alpha}^{(j)})_j$ for computing the ideal approximation $\tilde{\pi}_{n,\alpha}$, there is an additional term in the generalization error bound :

Theorem

$$R(\tilde{\pi}_{n,\alpha}^{(j)}) \leq \frac{2}{1-\alpha}\inf \|f_{\theta^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1 + \frac{\sigma^2}{\alpha}\right) r_n + \frac{\mathbb{E}[\Delta_{n,j}]}{n}$$

where $\Delta_{n,j}$ is the difference between the maximum value of the ELBO and the value of the ELBO at the j^{th} iteration of the algorithm.

How to select the network architecture ?

- The choice of the architecture of the neural network is crucial and can lead to faster convergence and better approximation.
- Contrary to the approach of Rockova & Polson (2018) which is fully Bayesian and treats S, L, D as random variables, we formulate the architecture design of DNNs as a model selection problem.
- Choose among a number M_S (resp. M_L, M_D) of possible values of the sparsity (resp. depth, width).

Strategy : ELBO maximization criterion

$$\hat{S}, \hat{L}, \hat{D} = \arg\max_{S,L,D} \mathrm{ELBO}(\tilde{\pi}_{n,\alpha}^{S,L,D}).$$

B.-E. Chérief-Abdellatif. Consistency of ELBO maximization for model selection. *AABI*, 2019.

Adaptivity of ELBO maximization

Theorem

For any value of S, L, D,

$$R(\tilde{\pi}_{n,\alpha}^{\hat{S},\hat{L},\hat{D}}) \leq \frac{2}{1-\alpha} \inf_{\theta^*_{S,L,D}} \|f_{\theta^*_{S,L,D}} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1 + \frac{\sigma^2}{\alpha}\right) r_n^{S,L,D} + \frac{2\sigma^2}{\alpha(1-\alpha)} \cdot \frac{\log M_S + \log M_L + \log M_D}{n}.$$

- As soon as T is not exponentially larger than n, then we adaptively achieve the lowest generalization error among all architectures.
- Typically, it leads to (near-)minimax rates for Hölder smooth functions and selects the optimal architecture even without the knowledge of the smoothness parameter β (which was previously required).

Thank you !