Scalable Subsampling Inference for Deep Neural Networks

Kejin Wu [1] Dimitris N. Politis [1,2]

[1]Department of Mathematics, University of California, San Diego [2]Halıcıoğlu Data Science Institute, University of California, San Diego

Motivation

In the statistical view, there are two important factors participators need to consider when large models, e.g., DNN, are applied in practice:

- **The estimation uncertainty:** The estimation error rate of large models is unknown or in a sub-optimal rate;
- **The computational burden of training large models:** Training a large model with a huge sample size requires heavy computational resources.

We explore the possibility of **killing two birds with one stone**: (1) Improve the estimation error bound of large models; (2) Decrease the training time.

Intuition

- **Variance reduction:** To improve the convergence rate of an estimator, we can try to decrease its variance if its bias is acceptable. This is inspired by *bagging* method of Breiman (1996).
- **Build estimators on subsamples:** To relieve the computational burden, we can repeat the estimation with subsamples if the computation with the whole sample is infeasible. This approach shares a general *divide-and-conquer* idea originally proposed by Cormen et al. (1989).

Example

Suppose we need $O(n^{\varphi})$ operations to train one large model denoted by $\widehat{f}_n$. If we consider $q = O(n/b)$ number of estimations $\widehat{f}_{b,i}$ on the i-th subsample with size b for $i = 1, \ldots, q$, we can take $\overline{f}_{b,n,\mathrm{SS}} := \frac{1}{q}\sum_{i=1}^{q} \widehat{f}_{b,i}$ to approximate $\widehat{f}_n$.

Training time of $\overline{f}_{b,n,\mathrm{SS}}$:

Only

$$O\left(nb^{\varphi-1}\right) \tag{1}$$

operations are needed. For a DNN, the total number of operations to train a DNN is $O(n \cdot W \cdot E)$; here E represents the number of epochs and W is the number of parameters. If we take $b = n^{\beta}$ $(0 < \beta < 1)$, the ratio of number of operations (1) over $O(n^{\varphi})$ is

$$n^{-(\varphi-1)(1-\beta)},$$

the larger model, the larger φ and the more computational saving.

Variance of $\overline{f}_{b,n,\mathrm{SS}}$:

If all subsamples are non-overlapping,

$$\mathrm{Var}\left(\overline{f}_{b,n,\mathrm{SS}}\right) = \frac{1}{q}\mathrm{Var}\left(\widehat{f}_{b,1}\right), \tag{2}$$

assuming $\mathrm{Var}\left(\widehat{f}_{b,i}\right)$ are equal across i; $q = \lfloor n/b \rfloor$. When $\mathbb{E}\left(\overline{f}_{b,n,\mathrm{SS}}\right) = \mathbb{E}\left(\widehat{f}_{b,i}\right)$, result (2) implies the variance deduction, so the decrease of Mean Square Error.

Scalable Subsampling

Scalable subsampling is one type of non-stochastic *subsampling* technique proposed by Politis (2024).

Suppose that we observe the sample $\{U_1, \ldots, U_n\}$; U_i represents $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ which are predictors and response variables, respectively.

The scalable subsampling relies on $q = \lfloor (n-b)/h \rfloor + 1$ number of subsamples $B_1, \ldots, B_q$ where $B_j = \{U_{(j-1)h+1}, \ldots, U_{(j-1)h+b}\}$; h controls the amount of overlap (or separation) between B_j and B_{j+1}.

Tuning b and h can make scalable subsampling samples have different properties:

- if $h = 1$, the overlap is the maximum possible;
- if $h = 0.2b$, there is 80% overlap between B_j and B_{j+1};
- if $h = b$, there is no overlap between B_j and B_{j+1};
- if $h = 1.2b$, there is a block of about $0.2b$ data points that separates blocks.

Feasible Assumption

A crucial condition is that the bias of the large model estimator is relatively negligible compared to its variance. This is possible from several perspectives:

- The double-descent of the risk exists for over-parameterized estimator revealed by Belkin et al. (2019);
- A deeper DNN may possess a lower bias confirmed by Yang et al. (2020) with ResNet on some image datasets;
- The L^{∞} norm of a DNN on estimating a function f can be uniformly $O(W^{-2\xi/d})$ proved by Yarotsky and Zhevnerchuk (2020). ξ is the smoothness measure of f.

MSE Bound of DNN with Scalable Subsampling

Take the large model $\widehat{f}_n$ as a fully connected feedforward DNN with ReLU activation functions. Assume:

$$\mathbb{E}(\widehat{f}_n(x) - f(x)) = O(n^{-\Lambda/2})$$

uniformly for all x in its domain $\mathcal{X}$ and some constant $\Lambda > \frac{\xi}{\xi+d}$.

Theorem: Under other appropriate conditions, let $b = h = n^{\beta}$; $\beta = \frac{1}{1+\Lambda-\frac{\xi}{\xi+d}}$. Then, with probability at least $(1 - \exp(-n^{\frac{d}{\xi+d}}\log^6 n))^q$:

$$\left\|\overline{f}_{b,n,\mathrm{SS}} - f\right\|^2_{L_2(X)} \leq n^{\frac{-\Lambda}{\Lambda+\frac{d}{\xi+d}}}\mathcal{L}(n); \tag{3}$$

where $\mathcal{L}(n)$ is a slowly varying function involving a constant and all $\log(n)$ terms.

Remark: The order of MSE in Eq. (3) is larger than the optimal and practically achievable MSE order without applying the scalable subsampling technique.

Simulation with DNN

To perform simulations, we consider models:

- Model-1: $Y_i = X_{i,1}^2 + \sin(X_{i,2} + X_{i,3}) + \epsilon_i$, where $X_i \sim N(0, I_3)$; $\epsilon_i \sim N(0,1)$;
- Model-2: $Y_i = X_{i,1}^2 + \sin(X_{i,2} + X_{i,3}) + \exp(-|X_{i,4} + X_{i,5}|) + \epsilon_i$, where $X_i \sim N(0, I_5)$; $\epsilon_i \sim N(0,1)$.

To be consistent with folk wisdom, we build $\widehat{f}_{b,i}$ with a relatively large depth to decrease the bias but also guarantee that the DNN estimator is in the under-parameterized region. We also consider other 5 DNN estimators trained with the whole sample:

(1) A DNN possesses the same depth and width as $\widehat{f}_{b,i}$, namely "S-DNN";
(2) A DNN possesses the same depth as $\widehat{f}_{b,i}$, but a larger width so that its size is close to the sample size, namely "DNN-deep-1";
(3) A DNN possesses the same depth as $\widehat{f}_{b,i}$, but a larger width so that its size is close to half of the sample size, namely "DNN-deep-2";
(4) A DNN possesses only one hidden layer, but a larger width so that its size is close to the sample size, namely "DNN-wide-1";
(5) A DNN possesses only one hidden layer, but a larger width so that its size is close to half of the sample size, namely "DNN-wide-2".

Table 1. Average MSE/MSPE and Training Time of different DNN models over 200 replications.

Estimator	SS-DNN	S-DNN	DNN-deep-1	DNN-deep-2	DNN-wide-1	DNN-wide-2
Model-1, $n = 10^4$						
Width	[15,15,15]	[15,15,15]	[65,65,65]	[45,45,45]	[2000]	[1000]
MSE	0.0296	0.0536	0.0533	0.0522	0.0426	0.0431
MSPE	0.0310	0.0564	0.0572	0.0570	0.0453	0.0449
Training Time (secs)	353	379	561	468	483	363
Model-2, $n = 10^4$						
Width	[15,15,15]	[15,15,15]	[65,65,65]	[45,45,45]	[2000]	[1000]
MSE	0.0757	0.0830	0.1076	0.0980	0.0729	0.0728
MSPE	0.0790	0.0875	0.1114	0.1045	0.0754	0.0749
Training Time (secs)	359	376	560	471	551	394
Model-2, $n = 2 \cdot 10^4$						
Width	[20,20,20]	[20,20,20]	[95,95,95]	[65,65,65]	[2800]	[1400]
MSE	0.0490	0.0653	0.0686	0.0675	0.0635	0.0635
MSPE	0.0502	0.0670	0.0692	0.0689	0.0623	0.0626
Training Time (secs)	748	775	1684	1198	1549	998

Empirical MSE: $\frac{1}{n}\sum_{i=1}^{n}(\widetilde{f}(x_i) - f(x_i))^2$; Empirical MSPE: $\frac{1}{N}\sum_{i=1}^{N}(\widetilde{f}(x_{0,i}) - f(x_{0,i}))^2$;

$\widetilde{f}$ represents different estimators; f is the true regression function; $\{x_i, y_i\}_{i=1}^n$ are observations of training data $\{X_i, Y_i\}_{i=1}^n$; $\{x_{0,i}, y_{0,i}\}_{i=1}^N$ are test data; $N = 2 \cdot 10^5$.

Future Work

- The estimation of exactly non-asymptotic bias and variance orders;
- The training of LLM and other large models with the scalable subsampling idea.

References

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24:123–140.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (1989). *Introduction to algorithms*. MIT press.

Politis, D. N. (2024). Scalable subsampling: computation, aggregation and inference. *Biometrika*, 111(1):347–354.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Yang, Z., Yu, Y., You, C., Steinhardt, J., and Ma, Y. (2020). Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pages 10767–10777. PMLR.

Yarotsky, D. and Zhevnerchuk, A. (2020). The phase diagram of approximation rates for deep neural networks. *Advances in neural information processing systems*, 33:13005–13015.