SnapStar Algorithm: a new way to ensemble Neural Networks

Sergey Zinchenko [1] **Dmitry Lishudi** [2]

Abstract

We propose a new neural network ensemble algorithm based on Audibert's empirical star algorithm and snapshot technique. We provide optimal theoretical minimax bound on the excess squared risk. Additionally, we empirically study this algorithm on regression and classification tasks and and show that it can be successfully applied to budget construct ensemble.

1. Introduction

Deep learning has been successfully applied to many types of problems and has reached the state-of-the-art performance. In many complex problems, such as the Imagenet competition (Deng et al., 2009), the best results are achieved by ensembles of neural networks, that is, it is often useful to combine the predictions of multiple neural networks to create a new one. As shown in work (Kawaguchi, 2016), the number of local minima grows exponentially with the number of parameters. And since modern neural network training methods are based on stochastic optimization, two identical architectures optimized with different initializations will probably converge to different solutions. Such a technique for obtaining neural networks with subsequent construction of an ensemble by majority voting or averaging is used, for example, in article (Caruana et al., 2004).

In addition to the fact that the class of deep neural networks has a huge number of local minima, it is also non-convex. It was shown in work (Lecué & Mendelson, 2009) that for the procedure of minimizing the empirical risk in a non-convex class of functions, the order of convergence is not optimal. J.-Y. Audibert proposed the star procedure method, which has optimal rate of convergence of excess squared risk (Audibert, 2007). Motivated by this observation and the huge success of ensembles of neural networks, we propose a modification of the star procedure that will combine the advantages of both methods.

We also take into account that training even a single neural network can be very expensive, and we propose an implementation of our algorithm that reuses the parameters of the trained models. In addition to this, we take into account that it is impossible to achieve a global minimum in the class of neural networks, and we consider the situation of imprecise minimization.

One can look at this procedure as a new way to train one large neural network with a block architecture, as well as a new way of aggregating models. In this work, we carry out a theoretical analysis of the behavior of the proposed algorithm for solving the regression problem with a class of sparse neural networks, and also check the operation of the algorithm in numerical experiments on classification and regression problems.

The *main results* of our work can be formulated as follows:

1. A multidimensional modification of the star procedure is proposed. We also offer budgetary implementation of this procedure.
2. We give an upper bound on the generalization error for the case of approximate empirical risk minimizers, which implies the optimality and stability against minimization errors of our algorithm.
3. We illustrate the efficiency of our approach with numerical experiments on real-world datasets.

2. Related work

2.1. Ensemble strategies

The main idea of the ensemble is to train several predictors and build a good metamodel on them. There are many techniques for its construction. We present some of them. A more detailed review can be found in the work (Ganaie et al., 2021).

Bagging The first of two stages is the generation of several samples with the same distribution as the training one. The next stage is training multiple models and aggregate their

[1]Novosibirsk State University, Novosibirsk, Russia [2]Higher School of Economics, Moscow, Russia. Correspondence to: Sergey Zinchenko <zinch.s.e@gmail.com>, Dmitry Lishudi <dlishudi@hse.ru>.

DyNN workshop at the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, 2022.

predictions. There are cases when the predictions of the constructed models are transferred to another model as new features (Kim et al., 2002). But still, most often, aggregation is performed either by majority voting or by averaging.

Boosting Another approach to construct ensembles is boosting. The idea is to build one strong model from several weak models by stepwise additive modeling. It was first applied to random trees to construct a so-called random forest. But it has also been applied to deep learning models as well. For example, in the task of recognizing facial expressions (Liu et al., 2014), or to improve the predictions of convolutional neural networks (Moghimi et al., 2016).

Snapshots The main problem in aggregation of deep learning models is the cost of training. Training even one modern model requires a lot of resources, and the ensemble needs a lot. A snapshot technique (Huang et al., 2017) and their modification (Garipov et al., 2018) have been created to combat this problem. In short, during the learning process, the step length in the gradient descent is cyclically changed. This allows a learner to get into various local minima (parameters of which are stored for subsequent aggregation) and, as a consequence, to build an ensemble using a computational budget comparable with the cost of training one model. Later, Zhang et al. tried combining this idea together with boosting in (Zhang et al., 2020).

Implicit ensembles In this approach a single model is trained in such a way as to behave like an ensemble. But it requires a much smaller computational budget for training. This is achieved due to the fact that in implicit ensembles the parameters of the models are shared, and their averaging is returned as predictions. For example, an implicit ensemble is the *DropOut* (Srivastava et al., 2014) method or the *DropConnect* (Wan et al., 2013) method. During training, each neuron or connection in the neural network has a chance to collapse, and after training, a neural network is returned, the elements of which are weighted by the probabilities of the presence of each element. A similar idea is implemented in the (Huang et al., 2016) *Stochastic depth* method for (He et al., 2016) residual neural networks. There, the residual blocks are randomly discarded during training, and the transformation goes only through a skip connection.

2.2. Star algorithm

Unlike the ensemble problem, the aggregation problem focuses on building a good predictor in a situation where there are already several ready-made models. The reader is referred to (Nemirovski, 2000; Tsybakov, 2003) for different types of aggregation. It is important to mention that, in contrast to the two-stage star procedure (Audibert, 2007), the usual empirical risk minimization procedure among the class of known predictors (or their convex hull) does not necessarily lead to the optimal rate of convergence (Lecué & Mendelson, 2009). This result was further developed in (Liang et al., 2015), where the authors extend the theoretical analysis of the star algorithm to the case of infinite classes using the offset Rademacher's complexity technique. It was also shown in the (Vijaykumar, 2021) that these results can be generalized to other loss functions. In particular, this means that the star procedure can be applied to more than just regression problems.

3. Theory

We have a $S_n = (\mathbf{X}_i, Y_i)_{i=1}^n$ sample of i.i.d. input-output pairs $(\mathbf{X}_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ distributed according to some unknown distribution $\mathcal{P}$. We also chose a certain family of solutions $\mathcal{F}$. Our goal is to build a new predictor $\widehat{f}$ minimizing the excess risk

$$\mathcal{E}(\widehat{g}) := \mathbb{E}(\widehat{g} - Y)^2 - \inf_{f \in \mathcal{F}} \mathbb{E}(f - Y)^2.$$

Let $\widehat{\mathbb{E}}$ denote the empirical expectation operator

$$\widehat{\mathbb{E}}(f) := \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{X}_i)$$

and call $\widehat{g} \in \mathcal{F}$ a Δ-*empirical risk minimizer in* $\mathcal{F}$ if the following inequality holds

$$\widehat{\mathbb{E}}(\widehat{g} - Y)^2 \leq \min_{f \in \mathcal{F}} \widehat{\mathbb{E}}(f - Y)^2 + \Delta.$$

We suggest the next two step procedure. In the first, calculate $\left\{\widehat{g}_i\right\}_{i=1}^d$ – different Δ_1-empirical risk minimizers in $\mathcal{F}$. And then look for a Δ_2-empirical risk minizer in the next set:

$$Star_d\big(\mathcal{F}, \widehat{g}_1, \dots \widehat{g}_d\big) := \Bigg\{ \sum_{i=1}^{d} \lambda_i \widehat{g}_i + \underbrace{\big(1 - \sum_{i=1}^{d} \lambda_i\big)}_{\lambda} f \;\Bigg|\; \lambda_i, \lambda \in [0,1];\ f \in \mathcal{F} \Bigg\}. \quad (1)$$

We will call the found function $\widehat{f} = \widehat{f}\,(\mathcal{F},\, d,\, \Delta_1, \Delta_2)$ as $Star_d$ *estimator*. The main result of our work is the proof that the proposed estimator has an optimal excess risk convergence rate in the case when $\mathcal{F}$ is a class of sparse fully connected neural networks $\mathcal{F}(L,\, \mathbf{p},\, s)$ (Schmidt-Hieber, 2020).

Let define the risk-minimizer in $\mathcal{F}$ and some sets:

$$Hull_d\big(\mathcal{F}\big) := \Bigg\{ \sum_{i=1}^{d} \lambda_i (g_i - f) \;\Bigg|\; \lambda_i \in [0,1];\ \sum_{i=1}^{d} \lambda_i \leq 1;\ f, g_1 \dots g_d \in \mathcal{F} \Bigg\}, \quad (2)$$

$$f^* := \arg\min_{f \in \mathcal{F}} \mathbb{E}(f(\mathbf{X}) - Y)^2, \tag{3}$$

$$\mathcal{H} := \mathcal{F} - f^* + Hull_d(\mathcal{F}). \tag{4}$$

Notice, that $Star_d$ estimator $\widehat{f}$ lies in $\mathcal{H} + f^*$. With the introduced notation, one of our main results is stated as follows.

Theorem 3.1. *Let $\widehat{f}$ is $Star_d$ estimator and $\mathcal{H}$ is the set defined in 3 for $\mathcal{F} = \mathcal{F}(L, \boldsymbol{p}, s)$. The following expectation bound on excess loss holds:*

$$\mathbb{E}\,\mathcal{E}(\hat{f}) \le C_{3.1}\left(\frac{d\,\log n}{n} + \Delta_1 + \Delta_2\right), \tag{5}$$

where $C_{3.1}$ depends only on the complexity of the class of neural networks $\mathcal{F}$.

In order to formulate an upper bound for the excess risk, performed with a high probability, we need to impose some constraints on the class of functions.

Definition 3.2 (Lower Isometry Bound)**.** Class $\mathcal{F}$ satisfies the lower isometry bound with some parameters $0 < \eta < 1$ and $0 < \delta < 1$ if

$$\mathbb{P}\left(\inf_{f \in \mathcal{F}\setminus\{0\}} \frac{1}{n}\sum_{i=1}^{n} \frac{f^2(\mathbf{X}_i)}{\mathbb{E}\,f^2} \ge 1 - \eta\right) \ge 1 - \delta$$

for all $n \ge n_0(\mathcal{F}, \delta, \eta)$, where $n_0(\mathcal{F}, \delta, \eta)$ depends on the complexity of the class of functions $\mathcal{F}$.

Theorem 3.3. *Let $\widehat{f}$ is $Star_d$ estimator and $\mathcal{H}$ is the set defined in 3 for $\mathcal{F} = \mathcal{F}(L, \boldsymbol{p}, s)$. Assume for $\mathcal{H}$ the lower isometry bound in Definition 3.2 holds with $\eta_{lib} = c_{A.2}/4$ and some $\delta_{lib} < 1$. Let $\xi_i = Y_i - f^*(\boldsymbol{X}_i)$. Then there exist constant $C_{3.3} = C_{3.3}(K, M, A, B)$ and absolute constants $c_{\tilde{A}.10}, c'_{A.10}, c_{A.10}$ such that*

$$\mathbb{P}\left(\mathcal{E}(\widehat{f}) > C_{3.3}\left[\frac{\log n/\delta}{n} + \Delta_1 + \Delta_2\right]\right) \le 4(\delta_{lib} + \delta)$$

as long as $n > \frac{16(1-c'_{A.10})^2 A}{c'^2_{A.10}} \vee n_0(\mathcal{H}, \delta_{lib}, c_{A.10}/4)$, where

$$A := \sup_{h \in \mathcal{H}} \frac{\mathbb{E}\,h^4}{(\mathbb{E}\,h^2)^2} \text{ and } B := \sup_{\boldsymbol{X}, Y} \mathbb{E}\,\xi^4,$$

$$K := \left(\sqrt{\sum_{i=1}^{n} \xi^2/n} + 2c_{\tilde{A}.10}\right),$$

$$M := \sup_{h \in \mathcal{H}\setminus\{0\}} \frac{\sum_{i=1}^{n} h(\boldsymbol{X}_i)^2 \xi_i^2}{c_{\tilde{A}.10} \sum_{i=1}^{n} h(\boldsymbol{X}_i)^2}.$$

That is, with some assumptions on the class of neural networks $\mathcal{F}$, we again obtained the order $\mathcal{O}\left(\frac{\log n}{n}\right)$ of convergence of the excess risk. Note that in the general case for an infinite class functions such an asymptotic rate with respect to the sample size n is *unimprovable* (Rakhlin et al., 2017).

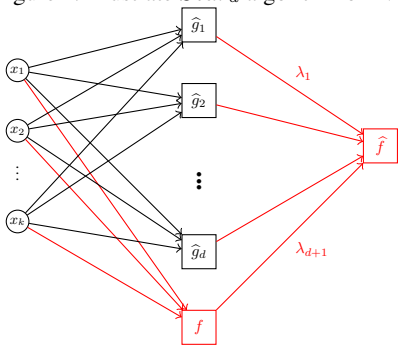
4. Realization

Algorithm 1 Star-d algorithm

Input: data S_n, parameters d,$epochs$,lr
for $i = 1$ **to** d **do**
  $\widehat{g}_i$ = $calculate\_erm(S_n, \mathcal{F}, epochs, lr)$
end for
$\widehat{f}$ = $calculate\_erm(S_n, Star_d(\mathcal{F}), epochs, lr)$
return $\widehat{f}$

The proposed $Star_d$ procedure can be represented by the following pseudocode (see Algorithm 1). The $calculate\_erm$ procedure is some optimization process that reduces the empirical risk. In practice it is impossible to search for a global empirical risk minimizer in the space of neural networks, which is why we introduced the concept of Δ-minimizers. As follows from our results, the more accurate the optimization is at each step of the algorithm, the higher the accuracy guarantee of the final predictor $\widehat{f}$.

Figure 1. Illustrate $Star_d$ algorithm on NN



Despite the fact that the second step of the star algorithm requires an optimization procedure over some complex set $Star_d$, this is fairly easy to implement in practice (see Figure 1). Suppose that we have fixed some architecture of estimator (black block), then in the first step we independently optimize the weights of the blocks $\widehat{g}_i$, freeze them and in the second step we add a new block f, connecting all of them by convex layer λ_i (red elements) and optimize them. This actually iterates over all possible simplices, optimizing the weights of the lower block, and all possible points within the simplex, optimizing the convex weights.

But training even one neural network is a rather complicated process, and in our algorithm it is required to train $d + 1$ predictors. To solve this problem, we propose to train the d models consequentially using the snapshot technique and at the last stage add convex coefficients and optimize $d + 1$-st block together with them (**Snap Star**). This does not contradict our theoretical result, since no conditions were imposed on obtaining the minimizer. This solution allows us to *reuse the obtained data* from previous models and *save*

computing resources.

5. Experiments

As competitors for numerical experiments, we chose 3 models: training in the classical way one large neural network of $d+1$ blocks (**Big NN**), learning $d+1$ blocks independently and averaging (**Ensemble**), learning blocks sequentially using the snapshot technique with subsequent averaging (**Snap Ensemble**). For the purposes of reproducing the results, the code and extended tables with results are publicly available at[1]. Adam was chosen as the optimizer.

BOSTON Task is to predict the value of real estate according to some characteristic (Harrison Jr & Rubinfeld, 1978). The ratio of training and test samples is equal to $7:3$. Standard scaler was used as preprocessing, batch size is 32. A small fully connected ReLu neural network of 4 layers was chosen as the architecture of the neural network, the number of neurons on the first layer is 128, then with the growth of the layer it decreases by 2 times, DropOut with parameter p and batch normalization are applied between the layers. Averaging over 5 runs.

Table 1. BOSTON ($epochs = 30$, $p = 0.1$, $lr = 0.01$)

NAME	D	MSE	MAE	R^2
SNAP STAR	5	**10.881±0.575**	**2.229**	**0.869**
SNAP ENSEMBLE	5	11.862±0.616	2.306	0.858
ENSEMBLE	5	12.568±0.878	2.399	0.849
BIG NN	5	12.068±0.860	2.411	0.855
SNAP STAR	4	**11.276±0.582**	**2.269**	**0.865**
SNAP ENSEMBLE	4	11.819±0.341	2.316	0.858
ENSEMBLE	4	12.059±0.614	2.365	0.855
BIG NN	4	12.556±0.904	2.383	0.849

FMNIST The second experiment was carried out on the Fashion Mnist dataset (Xiao et al., 2017), which consists of $70,000$ images (28×28 pixels). It is required to classify images by clothing classes. The ratio of training and test samples is equal to $6:1$. No scaler is used, batch size is 64. A simple convolutional network LeNet was chosen as a solution to this task. Averaging over 3 runs.

Table 2. FMNIST ($epochs = 5$, $lr = 0.001$)

NAME	D	ACCURACY	ENTROPY
SNAP STAR	3	**0.900±0.002**	**0.284±0.008**
SNAP ENSEMBLE	3	0.897±0.003	0.290±0.009
ENSEMBLE	3	0.887±0.001	0.310±0.005
BIG NN	3	0.890±0.010	0.299±0.022
SNAP STAR	2	**0.894±0.007**	**0.294±0.020**
SNAP ENSEMBLE	2	0.891±0.006	0.302±0.021
ENSEMBLE	2	0.886±0.004	0.313±0.008
BIG NN	2	0.892±0.003	0.304±0.007
SNAP STAR.	1	**0.891±0.002**	**0.299±0.006**
SNAP ENSEMBLE	1	0.889±0.001	0.304±0.007
ENSEMBLE	1	0.886±0.005	0.314±0.011
BIG NN	1	0.886±0.002	0.315±0.005

6. Discussion

The proposed algorithm performs well in the classification problem with cross-entropy loss, although this paper only presents a theoretical analysis for regression problem.

In fact, the star estimator we proposed is a multidimensional analogue of the Audibert's algorithm. It combines optimal orders as a solution to the aggregation problem of model selection, and at the same time behaves like an ensemble method. This decision can be viewed from 3 sides at once.

$Star_d$ algorithm as a new learning algorithm
It is worth noting that if we spend a fixed amount of computing resources B for each call to the optimization process $calculate\_erm$, then the total budget of our algorithm is about $(d+1) \cdot B$. But the surprising fact is that the result obtained is able to compete with other methods for training the final large neural network from $d+1$ blocks, although our theoretical analysis guarantees optimality only in comparison with the best single block architecture model. Thus, the procedure we proposed can be perceived as a new method for training neural networks with block architecture.

$Star_d$ algorithm as a new way of model aggregation
Also note that the predictors $\widehat{g}_i$ need not be trained in the first step. Then the $Star_d$ algorithm can be perceived as an algorithm for aggregating these models. It will consist of the following: a new predictive model f is added, a connecting layer, and the process of optimization by a parameter is started. At the same time, generally speaking, it is not necessary to have all blocks be of the same architecture. As intuition suggests, the main thing is that the expressive abilities of those classes of solutions to which the predictors given to us will relate should be approximately equal. Then it will be possible to formally consider the union of those decision classes to which each of the predictors belongs, and consider them as Δ-minimizers from the following class $\mathcal{F} = \bigcup_i \mathcal{F}_i$, where given predictors $\widehat{g}_i \in \mathcal{F}_i$ (which may be heterogeneous).

$Star_d$ algorithm as a new way to budget build ensemble
In combination with the snapshot technique, our algorithm shows good results with an extremely small number of epochs. Classical methods require more computational resources to achieve the same performance.

[1] https://github.com/mordiggian174/star-ensembling

7. Acknowledgments

We are grateful to Nikita Puchkin for essential comments and productive discussions, and also to Alexander Trushin for help with the design of the work. The article was prepared within the framework of the HSE University Basic Research Program.

References

Audibert, J.-Y. Progressive mixture rules are deviation suboptimal. *NeurIPS*, 2007.

Caruana, R., Niculescu-Mizil, A., Crew, G., and Ksikes, A. Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 18, 2004.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., and Suganthan, P. N. Ensemble deep learning: A review, 2021. URL https://arxiv.org/abs/2104.02395.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns, 2018. URL https://arxiv.org/abs/1802.10026.

Harrison Jr, D. and Rubinfeld, D. L. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K. Q. Deep networks with stochastic depth. In *European conference on computer vision*, pp. 646–661. Springer, 2016.

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., and Weinberger, K. Q. Snapshot ensembles: Train 1, get m for free, 2017. URL https://arxiv.org/abs/1704.00109.

Kawaguchi, K. Deep learning without poor local minima. *Advances in neural information processing systems*, 29, 2016.

Kim, H.-C., Pang, S., Je, H.-M., Kim, D., and Bang, S.-Y. Support vector machine ensemble with bagging. In *International workshop on support vector machines*, pp. 397–408. Springer, 2002.

Lecué, G. and Mendelson, S. Aggregation via empirical risk minimization. *Probability theory and related fields*, 145(3):591–613, 2009.

Liang, T., Rakhlin, A., and Sridharan, K. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pp. 1260–1285. PMLR, 2015.

Liu, P., Han, S., Meng, Z., and Tong, Y. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1805–1812, 2014.

Moghimi, M., Belongie, S. J., Saberian, M. J., Yang, J., Vasconcelos, N., and Li, L.-J. Boosted convolutional neural networks. In *BMVC*, volume 5, pp. 6, 2016.

Nemirovski, A. Topics in non-parametric statistics. *Lectures on probability theory and statistics (Saint-Flour, 1998)*, 1738:85–277, 2000.

Rakhlin, A., Sridharan, K., and Tsybakov, A. B. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23 (2):789–824, 2017.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Tsybakov, A. B. Optimal rates of aggregation. In *COLT*, 2003.

Vijaykumar, S. Localization, convexity, and star aggregation. *Advances in Neural Information Processing Systems*, 34, 2021.

Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., and Fergus, R. Regularization of neural networks using dropconnect. In *International conference on machine learning*, pp. 1058–1066. PMLR, 2013.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Zhang, W., Jiang, J., Shao, Y., and Cui, B. Snapshot boosting: a fast ensemble framework for deep neural networks. *Science China Information Sciences*, 63(1):1–12, 2020.

A. Proofs

The main result of our work is the proof that the proposed estimator has an optimal excess risk convergence rate in the case when $\mathcal{F}$ is a class of fully connected neural networks. It is defined by the choice of the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ and the network architecture. We study neural network with activation function ReLu:

$$\sigma(x) := max(x, 0).$$

For $\mathbf{v} = (v_1, \ldots, v_r) \in \mathbb{R}^r$ define shifted activation function $\sigma_{\mathbf{v}} : \mathbb{R}^r \to \mathbb{R}^r$:

$$\sigma_{\mathbf{v}}(\mathbf{x}) := \big(\sigma(x_i - v_i)\big)_{i=1}^{r}.$$

The network architecture $(L, \mathbf{p})$ consists of a positive integer L called the *number of hidden layers or depth* and a *width vector* $\mathbf{p} = (p_0, \ldots, p_{L+1}) \in \mathbb{N}^{L+2}$. A neural network with network architecture $(L, \mathbf{p})$ is then any function of the form

$$f(\mathbf{x}) = W_L \sigma_{\mathbf{v}_L} W_{L-1} \sigma_{\mathbf{v}_{L-1}} \ldots W_1 \sigma_{\mathbf{v}_1} W_0 \mathbf{x}, \tag{6}$$

where W_j is a $p_{j+1} \times p_j$ matrix and $\mathbf{v}_i \in \mathbb{R}^{p_i}$ is a shift vector.

We will focus on the case when the model parameters satisfy some constraint. Denote $\|W_j\|_\infty$ the maximum-entry norm of W_j , $\|W_j\|_0$ the number of non-zero/active entries of W_j then the space of network functions with given network architecture and network parameters bounded by one is

$$\mathcal{F}(L, \mathbf{p}) := \left\{ f \text{of the form (6)} : \max_{j=0,\ldots,L} \|W_j\|_\infty \vee \|\mathbf{v}_j\|_\infty \leq 1 \right\}$$

and the s-sparse networks are given by

$$\mathcal{F}(L, \mathbf{p}, s) := \left\{ f \in \mathcal{F}(L, \mathbf{p}) : \sum_{j=0}^{L} \|W_j\|_0 + \|\mathbf{v}_j\|_0 \leq s \right\}.$$

Let's denote by $\mathcal{N}_\infty(\mathcal{F}, \varepsilon), \mathcal{N}_2(\mathcal{F}, \varepsilon)$ the size of the ε-net of $\mathcal{F}$ in the metric space L_∞ and L_2, respectively. Then from Lemma 5 in (Schmidt-Hieber, 2020) we have

$$\forall f \in \mathcal{F}(L, \mathbf{p}, s) : \|f\|_\infty \leq V(L+1) \tag{7}$$

and

$$\log \mathcal{N}_2(\mathcal{F}(L, \mathbf{p}, s), \delta) \leq \log \mathcal{N}_\infty(\mathcal{F}(L, \mathbf{p}, s), \delta) \leq (s+1) \log(2\delta^{-1}(L+1)V^2), \tag{8}$$

where

$$V := \prod_{l=0}^{L+1} (p_l + 1). \tag{9}$$

The combination of the following 2 Lemmas is a generalization of the geometric inequality proved by (Liang et al., 2015). In many respects the scheme of the proof is similar.

Lemma A.1. *(Geometric inequality for the exact $Star_d$ estimator in the second step)*
Let $\hat{g}_1 \ldots \hat{g}_d$ be Δ_1-empirical risk minimizers from the first step of the $Star_d$ procedure, $\widetilde{f}$ be the exact minimizer from the second step of the $Star_d$ procedure. Then, for $c_{A.1} = \frac{1}{18}$ the following inequality holds:

$$\widehat{\mathbb{E}}(h - Y)^2 - \widehat{\mathbb{E}}(\widetilde{f} - Y)^2 \geq c_{A.1}\, \widehat{\mathbb{E}}(\widetilde{f} - h)^2 - 2\Delta_1. \tag{10}$$

Proof. For any function f, g we denote the empirical ℓ_2 distance to be $\|f\|_n := \left[\widehat{\mathbb{E}}\, f^2\right]^{\frac{1}{2}}$, empirical product to be $\langle f, g \rangle_n := \widehat{\mathbb{E}}\,[fg]$ and the square of the empirical distance between $\mathcal{F}$ and Y as r_1. By definition of $Star_d$ estimator for some $\lambda \in [0; 1]$ we have:

$$\widetilde{f} = (1 - \lambda)\widehat{g} + \lambda f,$$

where $\widehat{g}$ lies in a convex hull of Δ_1-empirical risk minimizers $\{\widehat{g}_i\}_{i=1}^d$. Denote the balls centered at Y to be $\mathcal{B}_1 := \mathcal{B}(Y, \sqrt{r_1})$, $\mathcal{B}_1' := \mathcal{B}(Y, \|\widehat{g} - Y\|_n)$ and $\mathcal{B}_2 := \mathcal{B}(Y, \|\widetilde{f} - Y\|_n)$. The corresponding spheres will be called $\mathcal{S}_1, \mathcal{S}_1', \mathcal{S}_2$. We have $\mathcal{B}_2 \subseteq \mathcal{B}_1$ and $\mathcal{B}_2 \subseteq \mathcal{B}_1'$. Denote by $\mathcal{C}$ the conic hull of $\mathcal{B}_2$ with origin $\widehat{g}$ and define the spherical cap outside the cone $\mathcal{C}$ to be $\mathcal{S} = \mathcal{S}_1' \setminus \mathcal{C}$.

First, $\widetilde{f} \in \mathcal{B}_2$ and it is a contact point of $\mathcal{C}$ and $\mathcal{S}_2$. Indeed, $\widetilde{f}$ is necessarily on a line segment between $\hat{g}$ and a point outside $\mathcal{B}_1$ that does not pass through the interior of $\mathcal{B}_2$ by optimality of $\widetilde{f}$. Let K be the set of all contact points of $\mathcal{C}$ and $\mathcal{S}_2$ – potential locations of $\widetilde{f}$.

Second, for any $h \in \mathcal{F}$, we have $\|h - Y\|_n \geq \sqrt{r_1}$ i.e. any $h \in \mathcal{F}$ is not in the interior of $\mathcal{B}_1$. Furthermore, let $\mathcal{C}'$ be bounded subset cone $\mathcal{C}$ cut at K. Thus $h \in (int\mathcal{C})^c \cap (\mathcal{B}_1)^c$ or $h \in \mathcal{T}$, where $\mathcal{T} := (int\mathcal{C}') \cap (\mathcal{B}_1)^c$.

For any $h \in \mathcal{F}$ consider the two dimensional plane $\mathcal{L}$ that passes through three points $\hat{g}$, Y, h, depicted in Figure 2. Observe that the left-hand side of the desired inequality (10) is constant as $\widetilde{f}$ ranges over K. The maximization of $\|h - f'\|_n^2$ over $f' \in K$ is achieved by $f' \in K \cap \mathcal{L}$. Hence, to prove the desired inequality, we can restrict our attention to the plane $\mathcal{L}$ and f'. Let $h_\perp$ be the projection of h onto the shell $L \cap S_1'$. By the geometry of the cone and triangle inequality we have:

$$\|f' - \widehat{g}\|_n \geq \frac{1}{2}\|\widehat{g} - h_\perp\|_n \geq \frac{1}{2}\left(\|f' - h_\perp\|_n - \|f' - \widehat{g}\|_n\right),$$

and, hence, $\|f' - \widehat{g}\|_n \geq \|f' - h_\perp\|_n/3$. By the Pythagorean theorem,

$$\|h_\perp - Y\|_n^2 - \|f' - Y\|_n^2 = \|\widehat{g} - Y\|_n^2 - \|f' - Y\|_n^2 = \|f' - \widehat{g}\|_n^2 \geq \frac{1}{9}\|f' - h_\perp\|_n^2.$$

We can now extend this claim to h. Indeed, due to the geometry of the projection $h \to h_\perp$ and the fact that $h \in (int\mathcal{C})^c \cap (int\mathcal{B}_1)^c$ or $h \in \mathcal{T}$ there are 2 possibilities:

a) $h \in (\mathcal{B}_1')^c$. Then $\langle h_\perp - Y, h_\perp - h\rangle_n \leq 0$;

b) $h \in \mathcal{B}_1'$. Then, since $h \in (\mathcal{B}_1)^c$, we have

$$\langle h_\perp - Y, h_\perp - h\rangle_n \leq \big(\|h - Y\| + \|h - h_\perp\|\big)\|h - h_\perp\| \leq \|h_\perp - Y\|_n^2 - \|h - Y\|_n^2 \leq \Delta_1.$$

In both cases, the following inequality is true

$$\begin{aligned}\|h - Y\|_n^2 - \|f' - Y\|_n^2 &= \|h_\perp - h\|_n^2 - 2\langle h_\perp - Y, h_\perp - h\rangle_n + (\|h_\perp - Y\|_n^2 - \|f' - Y\|_n^2) \\ &\geq \|h_\perp - h\|_n^2 - 2\Delta_1 + \frac{1}{9}\|f' - Y\|_n^2 \geq \frac{1}{18}\|f' - h\|_n^2 - 2\Delta_1.\end{aligned}$$

□

Lemma A.2 (Geometric Inequality for Δ-empirical minimizers)**.** *Let $\hat{g}_1 \dots \hat{g}_d$ be Δ_1-empirical risk minimizers from the first step of the $Star_d$ procedure, and $\widehat{f}$ be the Δ_2-empirical risk minimizer from the second step of the $Star_d$ procedure. Then, for any $h \in \mathcal{F}$ and $c_{A.2} = \frac{1}{36}$ the following inequality holds:*

$$\widehat{\mathbb{E}}(h - Y)^2 - \widehat{\mathbb{E}}(\widehat{f} - Y)^2 \geq c_{A.2}\widehat{E}(\widehat{f} - h)^2 - 2(1 + c_{A.2})[\Delta_1 + \Delta_2].$$

Proof. Since Lemma A.1 was actually proven for any $f \in K$, let f'' be the closest point to $\widehat{f}$ from K. For this f'' the inequality (10) holds. Similarly to Lemma A.1, there are 2 options: either $\widehat{f} \in (int\mathcal{C})^c$, or $\widehat{f} \in \mathcal{T}$.

a) Let $\widehat{f} \in (int\mathcal{C})^c$, then $\langle \widehat{f} - f'', f'' - Y\rangle \geq 0$. Since $\widehat{f}$ is Δ_2-empirical risk minimizer, we have $\|\widehat{f} - f''\|_n^2 + 2\langle \widehat{f} - f'', f'' - Y\rangle + \|f'' - Y\|_n^2 = \|\widehat{f} - Y\|_n^2 \leq \|f'' - Y\|_n^2 + \Delta_2$. It means, that $\|\widehat{f} - f''\|_n^2 \leq \Delta_2$.

b) Let $\widehat{f} \in \mathcal{T}$, then by the cosine theorem (as depicted on Figure 2, $\mathcal{L}$ is the two dimensional plane which passes through $\widehat{f}$, $\widehat{g}$, Y):

$$\|\widehat{f} - f''\|_n^2 = \|f'' - Y\|_n^2 + \|\widehat{f} - Y\|_n^2 - 2\|f'' - Y\|_n\|\widehat{f} - Y\|_n \cos(\varphi').$$

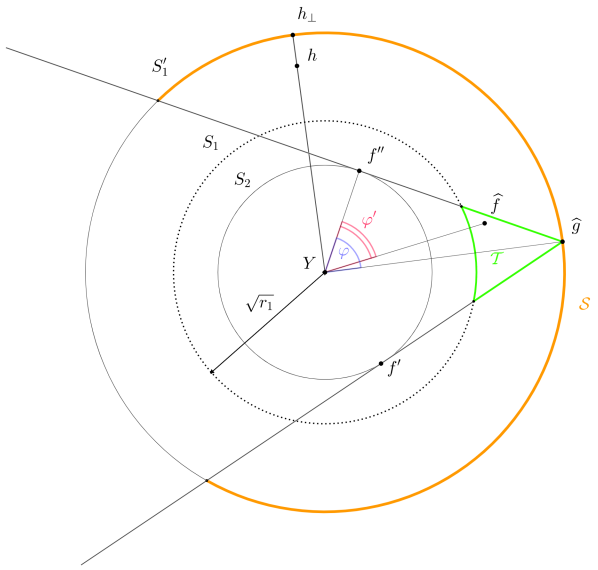


Figure 2. The cut surface $\mathcal{L}$

But $\cos(\varphi') \geq \cos(\varphi) = \frac{\|f''-Y\|_n}{\|\widehat{g}-Y\|_n}$ and $\|\widehat{f}-Y\|_n^2 \geq r_1$. Then we have:

$$\|\widehat{f}-f''\|_n^2 \leq \Delta_2 + 2\|f''-Y\|_n^2\left(1-\frac{\|\widehat{f}-Y\|_n}{\|\widehat{g}-Y\|_n}\right)$$
$$\leq \Delta_2 + 2\frac{\|f''-Y\|_n^2}{\|\widehat{g}-Y\|_n}\left(\frac{\|\widehat{g}-Y\|_n^2-\|\widehat{f}-Y\|_n^2}{\|\widehat{g}-Y\|_n+\|\widehat{f}-Y\|_n}\right) \leq \Delta_1+\Delta_2.$$

Lemma A.1 states:

$$\|h-Y\|_n^2 \geq \|f''-Y\|_n^2 + c_{A.1}\|f''-h\|_n^2 - 2\Delta_1.$$

By using the triangle inequality and the convexity of the quadratic function, we can get the following bound

$$\frac{c_{A.1}}{2}\|\widehat{f}-h\|_n^2 \leq c_{A.1}\left(\|\widehat{f}-f''\|_n^2+\|f''-h\|_n^2\right) \leq c_{A.1}[\Delta_2+\Delta_1]+c_{A.1}\|f''-h\|_n^2.$$

Combining everything together, we get the required result for the constant $c_{A.2} = \frac{c_{A.1}}{2} = \frac{1}{36}$:

$$\widehat{\mathbb{E}}(h-Y)^2 - \widehat{\mathbb{E}}(\widehat{f}-Y)^2 \geq c_{A.2}\cdot\widehat{\mathbb{E}}(\widehat{f}-h)^2 - 2(1+c_{A.2})[\Delta_1+\Delta_2].$$

□

For convenience, we introduce a Δ-excess risk

$$\mathcal{E}_\Delta(\widehat{g}) := \mathbb{E}(\widehat{g}-Y)^2 - \inf_{f\in\mathcal{F}}\mathbb{E}(f-Y)^2 - 2(1+c_{A.2})[\Delta_1+\Delta_2],$$

then the following 2 statements are the direct consequences of the corresponding statements from the article (Liang et al., 2015). The only difference is that in our case the geometric inequality has terms on the right side with minimization errors Δ_1, Δ_2. Also our definition of the set $\mathcal{H}$ is different, but all that was needed from it was the property that $\widehat{f}$ lies in $\mathcal{H}+f^*$. For brevity, we will not repeat the proofs, but only indicate the numbers of the corresponding results in the titles of the assertions. We will also proceed for statements the proofs for which we slightly modify or use without changes.

Corollary A.3 (Corollary 3)**.** *Conditioned on the data* $\{(\boldsymbol{X}_i, Y_i) : 1 \leq i \leq n\}$*, we have a deterministic upper bound for the* $Star_d$ *estimator:*

$$\mathcal{E}_\Delta(\widehat{f}) \leq (\widehat{\mathbb{E}}-\mathbb{E})[2(f^*-Y)(f^*-\widehat{f})] + \mathbb{E}(f^*-\widehat{f})^2 - (1+c_{A.2})\cdot\widehat{\mathbb{E}}(f^*-\widehat{f})^2.$$

Theorem A.4 (Theorem 4)**.** *The following expectation bound on excess loss of the $Star_d$ estimator holds:*

$$\mathbb{E}\,\mathcal{E}_\Delta(\widehat{f}) \le (2F' + F(2+c_{A.2})/2)\cdot \mathbb{E}_\sigma \sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^n 2\sigma_i h(\boldsymbol{X}_i) - c_{A.4}h(\boldsymbol{X}_i)^2\right\},$$

where $\sigma_1,\dots\sigma_n$ are independent Rademacher random variables, $c_{A.4} = \min\left\{\frac{c_{A.2}}{4F'}, \frac{c_{A.2}}{4F(2+c_{A.2})}\right\}$, $F = \sup_{f\in\mathcal{F}}|f|_\infty$ and $F' = \sup_{\mathcal{F}}|Y-f|_\infty$ almost surely.

Theorem A.5 (Theorem 7)**.** *Assume the lower isometry bound in Definition 3.2 holds with $\eta_{lib} = c_{A.2}/4$ and some $\delta_{lib} < 1$ and $\mathcal{H}$ is the set defined in 3. Let $\xi_i = Y_i - f^*(\boldsymbol{X}_i)$. Define*

$$A := \sup_{h\in\mathcal{H}}\frac{\mathbb{E}\,h^4}{(\mathbb{E}\,h^2)^2} \text{ and } B := \sup_{\boldsymbol{X},Y}\mathbb{E}\,\xi^4.$$

Then there exist two absolute constants $c'_{A.5}, c_{\tilde{A}.5} > 0$ (which only depend on $c_{A.2}$), such that

$$\mathbb{P}\left(\mathcal{E}_\Delta(\widehat{f}) > 4u\right) \le 4\delta_{lib} + 4\,\mathbb{P}\left(\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \sigma_i\xi_i h(\boldsymbol{X}_i) - c_{\tilde{A}.5}h(\boldsymbol{X}_i)^2 > u\right)$$

for any

$$u > \frac{32\sqrt{AB}}{c'_{A.5}}\frac{1}{n}$$

as long as $n > \frac{16(1-c'_{A.5})^2 A}{c'^2_{A.5}} \vee n_0(\mathcal{H}, \delta_{lib}, c_{A.2}/4)$.

Lemma A.6 (Lemma 15)**.** *The offset Rademacher complexity for $\mathcal{H}$ is bounded as:*

$$\mathbb{E}_\sigma \sup_{\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^n 2\sigma_i\xi_i h(\boldsymbol{X}_i) - Ch(\boldsymbol{X}_i)^2\right\} \le K(C)\varepsilon + M(C)\cdot\frac{\log\mathcal{N}_2(\mathcal{H},\varepsilon)}{n}$$

and with probability at least $1-\delta$

$$\sup_{\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^n 2\sigma_i\xi_i h(\boldsymbol{X}_i) - Ch(\boldsymbol{X}_i)^2\right\} \le K(C)\varepsilon + M(C)\cdot\frac{\log\mathcal{N}_2(\mathcal{H},\varepsilon) + \log\frac{1}{\delta}}{n},$$

where

$$K(C) := 2\left(\sqrt{\sum_{i=1}^n \xi^2/n} + C\right),\ \ M(C) := \sup_{h\in\mathcal{H}\setminus\{0\}} 4\frac{\sum_{i=1}^n h(\boldsymbol{X}_i)^2\xi_i^2}{C\sum_{i=1}^n h(\boldsymbol{X}_i)^2}. \tag{11}$$

Proof. Let $N_2(\mathcal{H},\varepsilon)$ be the ε-net of the $\mathcal{H}$ of size at most $\mathcal{N}_2(\mathcal{H},\varepsilon)$ and $v[h]$ be the closest point from this net for function $h\in\mathcal{H}$, i.e. $\|h - v[h]\|_2 \le \varepsilon$. By using the inequality $v[h]_i^2 \le 2\Big(h_i^2 + (v[h]_i - h_i)^2\Big)$, we can get next upper bound:

$$\begin{aligned}
&\left\{\frac{1}{n}\sum_{i=1}^n 2\sigma_i\xi_i h(\mathbf{X}_i) - Ch(\mathbf{X}_i)^2\right\} \\
&\qquad\le \left\{\frac{1}{n}\sum_{i=1}^n 2\sigma_i\xi_i(h(\mathbf{X}_i) - v[h](\mathbf{X}_i)) + C\Big(v[h]^2(\mathbf{X}_i)/2 - h^2(\mathbf{X}_i)\Big)\right\} \\
&\qquad\qquad + \frac{1}{n}\sup_{v\in N_2(\mathcal{H},\varepsilon)}\left\{\sum_{i=1}^n 2\sigma_i\xi_i v(\mathbf{X}_i) - \frac{C}{2}v(\mathbf{X}_i)^2\right\} \\
&\qquad\qquad\le 2\varepsilon\left(\sqrt{\sum_{i=1}^n \xi_i^2/n} + C\right) + \frac{1}{n}\sup_{v\in N_2(\mathcal{H},\varepsilon)}\left\{\sum_{i=1}^n 2\sigma_i\xi_i v(\mathbf{X}_i) - \frac{C}{2}v(\mathbf{X}_i)^2\right\}.
\end{aligned}$$

The right summarand is supremum over set of cardinality not more than $\mathcal{N}_2(\mathcal{H},\varepsilon)$. By using Lemma A.11, we acquire the expected estimates. □

We have now obtained, using the offset Rademacher complexity technique, the upper bound on excess risk in terms of the coverage size of the set $\mathcal{H}$. To get the desired result, we need to obtain an upper bound on the size of the cover $\mathcal{H}$ in terms of the size of the cover $\mathcal{F}$.

Lemma A.7. *For any scale $\varepsilon > 0$, the covering number of $\mathcal{F} \subseteq V(L+1) \cdot \mathcal{B}_2$ (where $\mathcal{B}_2$ is a sphere of radius one in space with norm $\|\cdot\|_n$) and that of $\mathcal{H}$ are bounded in the sense:*

$$\log \mathcal{N}_2(\mathcal{F}, \varepsilon) \leq \log \mathcal{N}_2(\mathcal{H}, \varepsilon) \leq (d+2)\left[\log \mathcal{N}_2\left(\mathcal{F}, \frac{\varepsilon}{3(d+1)}\right) + \log \frac{6(d+1)V(L+1)}{\varepsilon}\right].$$

Proof. If we define as $N(\mathcal{F}, \varepsilon)$ the ε-net cardinality no more then $\mathcal{N}(\mathcal{F}, \varepsilon)$, then the following is true: $N(\mathcal{F}_1, \varepsilon_1) + N(\mathcal{F}_1, \varepsilon_2)$ is $(\varepsilon_1 + \varepsilon_2)$-net for $\mathcal{F}_1 + \mathcal{F}_2$. Hence, $\mathcal{N}(\mathcal{F}_1 + \mathcal{F}_2, \varepsilon_1 + \varepsilon_2) \leq \mathcal{N}(\mathcal{F}_1, \varepsilon_1) \cdot \mathcal{N}(\mathcal{F}_2, \varepsilon_2)$. With this we can obtain the following upper bound

$$\mathcal{N}_2(\mathcal{H}, \varepsilon) \leq \mathcal{N}_2(\mathcal{F} + Hull_d, \varepsilon) \leq \mathcal{N}_2\left(\mathcal{F}, \frac{\varepsilon}{3}\right) \cdot \mathcal{N}_2\left(Hull_d, \frac{2\varepsilon}{3}\right).$$

But since $Hull_d$ is the sum of $d+1$ functions from $\mathcal{F}$ with coefficients in $[-1; 1]$, by the inequaility (7), we can cover this with a net of size no more than

$$\left[\mathcal{N}_2\left(\mathcal{F}, \frac{\varepsilon}{3(d+1)}\right) \cdot \frac{6(d+1)V(L+1)}{\varepsilon}\right]^{d+1}.$$

□

Note that to obtain the required orders, we only need coverage with $\varepsilon = 1/n$.

Corollary A.8. *Let $\mathcal{H}$ defined in 3 for $\mathcal{F} = \mathcal{F}(L, \boldsymbol{p}, s)$, then for V defined in 9 holds*

$$\log \mathcal{N}_2\left(\mathcal{H}, \frac{1}{n}\right) \leq c_{A.8} d\, s \log\big(V L\, n\, d\big),$$

where $c_{A.8}$ is an indepedent constant.

Proof. By lemma A.7 and inequality 8, we have

$$\begin{aligned}\log \mathcal{N}_2(\mathcal{H}, 1/n) &\leq (d+2)\left[\log \mathcal{N}_2\left(\mathcal{F}(L, \mathbf{p}, s), \frac{1}{3n(d+1)}\right) + \log 6n(d+1)V(L+1)\right] \\ &\leq (d+2)\left[(s+1)\log\big(2V^2(L+1)(3n(d+1))\big) + \log\left(6n(d+1)V(L+1)\right)\right].\end{aligned}$$

□

We are now fully prepared to prove the two main results.

Theorem A.9. *Let $\widehat{f}$ be a $Star_d$ estimator and $\mathcal{H}$ be the set defined in 3 for $\mathcal{F} = \mathcal{F}(L, \boldsymbol{p}, s)$. The following expectation bound on excess loss holds:*

$$\mathbb{E}\,\mathcal{E}_\Delta(\hat{f}) \leq 2(F' + V(L+1)) \cdot \left[\frac{K(C)}{n} + M(C) \cdot \frac{c_{A.8} d\, s \log\left(V L\, n\, d\right)}{n}\right],$$

where $K(C),\ M(C)$ defined in (11) for constants

$$C = \min\left\{\frac{c_{A.2}}{4F'}, \frac{c_{A.2}}{4V(L+1)(2+c_{A.2})}\right\},\ \ F' = \sup_{\mathcal{F}} |Y - f|_\infty.$$

Proof. By using Theorem A.4 and inequality 7 we have

$$\mathbb{E}\,\mathcal{E}_\Delta(\hat{f}) \leq (2F' + V(L+1)(2+c_{A.2})/2) \cdot \mathbb{E}_\sigma \sup_{h \in \mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n} 2\sigma_i h(\mathbf{X}_i) - C h(\mathbf{X}_i)^2\right\},$$

where $C = \min\left\{\frac{c_{A.2}}{4F'}, \frac{c_{A.2}}{4V(L+1)(2+c_{A.2})}\right\}$, $F' = \sup_{\mathcal{F}} |Y - f|_\infty$ almost surely.

By using Lemma A.6 and corollary A.8 we get desired result

$$\mathbb{E}_\sigma \sup_{\mathcal{H}} \left\{ \frac{1}{n}\sum_{i=1}^n 2\sigma_i\xi_i h(\mathbf{X}_i) - Ch(\mathbf{X}_i)^2 \right\} \le \frac{K(C)}{n} + M(C)\cdot \frac{c_{A.8} d\, s \log\left(VL\, n\, d\right)}{n}.$$

□

Theorem A.10. *Let $\widehat{f}$ be a $Star_d$ estimator and let $\mathcal{H}$ be the set defined in 3 for $\mathcal{F} = \mathcal{F}(L, \boldsymbol{p}, s)$. Assume for $\mathcal{H}$ the lower isometry bound in Definition 3.2 holds with $\eta_{lib} = c_{A.2}/4$ and some $\delta_{lib} < 1$. Let $\xi_i = Y_i - f^*(\boldsymbol{X}_i)$. Define*

$$A := \sup_{h\in\mathcal{H}} \frac{\mathbb{E}\, h^4}{(\mathbb{E}\, h^2)^2} \text{ and } B := \sup_{\boldsymbol{X},Y} \mathbb{E}\, \xi^4.$$

Then there exist 3 absolute constants $c'_{A.10}, c_{\tilde{A}.10}, c_{A.10} > 0$ (which only depend on $c_{A.2}$), such that

$$\mathbb{P}\left(\mathcal{E}_\Delta(\widehat{f}) > 4D\right) \le 4(\delta_{lib} + \delta)$$

as long as $n > \frac{16(1-c'_{A.10})^2 A}{c'^2_{A.10}} \vee n_0(\mathcal{H}, \delta_{lib}, c_{A.10}/4)$, where

$$K := \left(\sqrt{\sum_{i=1}^n \xi^2/n} + 2c_{\tilde{A}.10}\right), \quad M := \sup_{h\in\mathcal{H}\setminus\{0\}} \frac{\sum_{i=1}^n h(\boldsymbol{X}_i)^2\xi_i^2}{c_{\tilde{A}.10}\sum_{i=1}^n h(\boldsymbol{X}_i)^2},$$

$$D := \max\left(\frac{K}{n} + M\cdot\frac{c_{A.8} d\, s \log\left(VL\, n\, d\right) + \log\frac{1}{\delta}}{n}, \frac{32\sqrt{AB}}{c'_{A.10}}\frac{1}{n}\right)$$

and $c_{A.8}$ is an independent constant.

Proof. By using Theorem A.5 for any $u > \frac{32\sqrt{AB}}{c'_{A.5}}\frac{1}{n}$ we have

$$\mathbb{P}\left(\mathcal{E}_\Delta(\widehat{f}) > 4u\right) \le 4\delta_{lib} + 4\,\mathbb{P}\left(\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \sigma_i\xi_i h(\mathbf{X}_i) - c_{\tilde{A}.5} h(\mathbf{X}_i)^2 > u\right)$$

as long as $n > \frac{16(1-c'_{A.5})^2 A}{c'^2_{A.5}} \vee n_0(\mathcal{H}, \delta_{lib}, c_{A.2}/4)$.

By using Lemmas A.6 and A.8 we have with probability no more than δ for any $C > 0$:

$$\sup_{\mathcal{H}} \left\{ \frac{1}{n}\sum_{i=1}^n \sigma_i\xi_i h(\mathbf{X}_i) - \frac{C}{2}h(\mathbf{X}_i)^2 \right\} \ge \frac{K(C)}{2}\varepsilon + \frac{M(C)}{2}\cdot\frac{\log\mathcal{N}_2(\mathcal{H},\varepsilon) + \log\frac{1}{\delta}}{n},$$

where $K(C),\ M(C)$ are defined in (11). Combining this inequality for $C = 2c_{\tilde{A}.10} = 2c_{\tilde{A}.5}$ and $c'_{A.10} = c'_{A.5}, c_{A.10} = c_{A.2}$ we get the required result. □

Lemma A.11 (Lemma 9)**.** *Let $V \subset \mathbb{R}^n$ be a finite set, $|V| = N$. Then, for any $C > 0$:*

$$\mathbb{E}_\sigma \max_{v\in V}\left[\frac{1}{n}\sum_{i=1}^n \sigma_i\xi_i v(\boldsymbol{X}_i) - Cv(\boldsymbol{X}_i)^2\right] \le M\frac{\log N}{n}.$$

For any $\delta > 0$:

$$\mathbb{P}_\sigma\left(\max_{v\in V}\left[\frac{1}{n}\sum_{i=1}^n \sigma_i\xi_i v(\boldsymbol{X}_i) - Cv(\boldsymbol{X}_i)^2\right] > M\frac{\log N + \log\frac{1}{\delta}}{n}\right) \le \delta,$$

where

$$M := \sup_{v\in V\setminus\{0\}} \frac{\sum_{i=1}^n v(\boldsymbol{X}_i)^2\xi_i^2}{2C\sum_{i=1}^n v(\boldsymbol{X}_i)^2}.$$