Benign Overfitting with Retrieval Augmented Models

Zhenisbek Assylbekov[1], Maxat Tezekbayev[1], Vassilina Nikoulina[2], and Matthias Gallé[*3]

[1]Department of Mathematics, Nazarbayev University
{maxat.tezekbayev, zhassylbekov}@nu.edu.kz
[2]NAVER LABS Europe
vassilina.nikoulina@naverlabs.com
[3]Cohere.ai
matthias@cohere.com

Abstract

Despite the fact that modern deep neural networks have the ability to memorize (almost) the entire training set they generalize well to unseen data, contradicting traditional learning theory. This phenomenon — dubbed *benign overfitting* — has been theoretically studied so far in simplified settings only. At the same time, ML practitioners (especially in NLP) figured out how to exploit this feature for more efficient training: retrieval-augmented models (e.g., kNN-LM, RETRO) explicitly store (part of) the training sample in the storage and thus try to (partially) remove a load of memorization from the neural network. In this paper we link these apparently separate threads of research, and propose several possible research directions regarding benign overfitting in retrieval-augmented models.

1 Introduction

In the classical learning theory (Valiant, 1984), generalizing ability and model complexity (Vapnik and Chervonenkis, 1974) are usually opposed to each other: the more complex the model,[1] the worse its generalizing ability on new data. This is well illustrated by typical curves of test and training errors as functions of the complexity of the model being trained. The training error tends to decrease whenever we increase the model complexity, that is, whenever we fit the data harder. With too much fitting, the model adapts itself too closely to the training data, and will not generalize well (i.e., have large test error).

However, modern machine learning models such as deep neural networks (DNNs) break this principle: they are usually complex enough to be able to memorize the entire training set, and nevertheless show excellent generalization ability. This phenomenon, called **benign overfitting**, was discovered empirically by Zhang et al. (2017) and has since attracted the attention of many minds in the field of machine learning, both experimentalists and theorists. We refer the reader to the survey of Bartlett et al. (2021) and Belkin (2021) for a more comprehensive overview of benign overfitting.

In Section 2 we link benign overfitting to two causes: 1) when learning from natural data with a so-called **long tail**, memorization of rare/atypical examples is inevitable, and this requires learning a complex model, 2) the gradient-based learner is biased towards **simplicity**: it prefers to fit the training set with the least complex models available. In Section 3, we discuss retrieval-augmented models, which complement a model with explicit memory thus allowing to reduce its complexity. This seems like a viable alternative to [complex models + simplicity bias] to achieve benign overfitting. Fig. 1 summarizes the argument made in Sections 2&3. In Section 4, we express

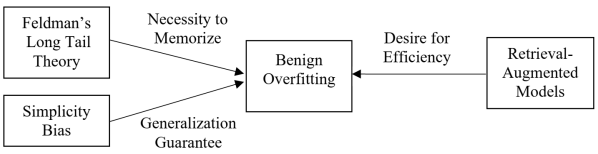


Figure 1: Established chain of reasoning.

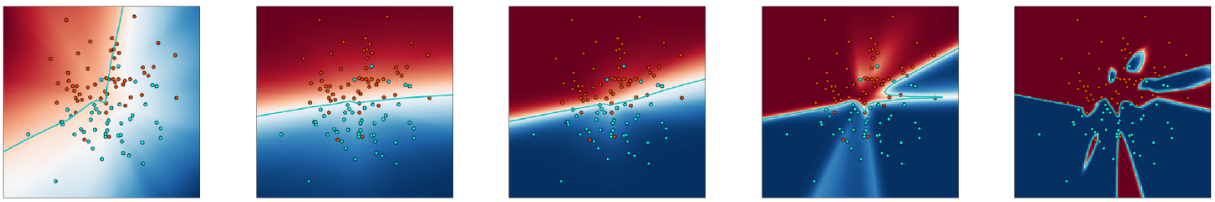
our concerns that the current theoretical studies of benign overfitting ignore the fundamental factor of long-tailedness of natural data, as well as RAMs. Thus, we suggest possible research directions that, in our opinion, may give us a better understanding of the benign overfitting phenomenon.

2 Benign Overfitting

Our departure point is the **long tail theory** of Feldman (2020), which considers learning from natural data (such as texts or images). The fact is that such data, as a rule, has a distribution with the so-called long tail, i.e. in such data, the proportion

[*]work done while at NAVER LABS Europe

[1]By *complexity* of a model we mean its ability to fit an arbitrary dataset.



(a) Step 0 (b) Step 10 (c) Step 100 (d) Step 1000 (e) Step 10000

Figure 2: SGD training on a 3-layer, width-100 dense neural network. The blue line corresponds to the decision boundary of the neural network which becomes more "linear" in the initial stages before starting to overfit to the label noise. Source: Kalimeris et al. (2019). Reproduced with permission.

of rare/atypical examples is significant. An example of a long-tail distribution is the distribution of words in a text corpus, which can be approximated by the Zipf's law. Now recall that in a language modeling task the prediction is over the word given its context, and therefore the word types are classes, which means that the distribution over classes in such setting is essentially long-tailed. Roughly speaking, Feldman (2020) showed analytically that if the class distribution has a long tail (as in language modeling) or the distribution of subpopulations within classes has a long tail, then to achieve optimal performance, the learning algorithm needs to memorize rare/atypical examples from the training set. Unfortunately, simple models such as linear classifiers are *not* able to overfit (unless they are trained in an overparameterized mode). Thus, for achieving optimal performance one needs **more complex models**, such as DNNs, that *can* perfectly fit the training set. Therefore, Feldman's theory explains the need to use complex models for which we can observe the phenomenon of benign overfitting, but it remains unclear where this phenomenon itself comes from.

One of the ways to answering this question is the so-called **simplicity bias** of stochastic gradient descent (SGD), the default algorithm for training DNNs. As the experiments of Kalimeris et al. (2019) showed, although a neural net with 3 hidden layers can malignantly overfit a training set shown in Fig. 3 using a complex decision boundary (Fig. 4), this does not happen in practice. When initialized randomly, the SGD algorithm starts to fit using a simple, almost linear, classifier (Fig. 2). And only if there are some examples that do not fall under the simple classification pattern, the SGD uses the available capacity of the neural network to fit such rare/atypical examples. At the same time, the simple part of the model is preserved to a certain extent, which generalizes well to new data.

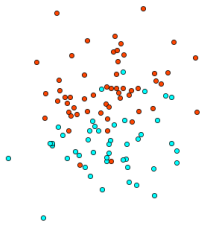


Figure 3: Training sample. Source: Kalimeris et al. (2019)

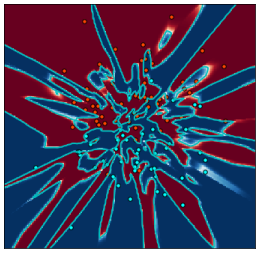


Figure 4: Malignant overfit with a 3-layer width-100 dense network. Source: Kalimeris et al. (2019).

Even though simplicity bias explains benign overfitting, where does this bias come from? There are empirical and theoretical studies on this. For example, Mingard et al. (2021) argue that SGD is essentially a Bayesian sampler that randomly selects the minimum of the loss function, and that in the space of functions representable by a neural network and consistent with the training sample, simple functions occupy a much larger volume than complex functions. At the same time, there are several theoretical works that prove the simplicity bias of the SGD, though in more simplified setups than those in which this bias is observed in practice. Here we present the seminal result of Soudry et al. (2018).

Theorem 1 (Soudry et al. (2018))**.** *For a linearly separable training sample and for a small enough step-size of the gradient descent, we have*

$$\frac{\mathbf{w}}{\|\mathbf{w}\|} = \lim_{t\to\infty} \frac{\mathbf{v}^{(t)}}{\|\mathbf{v}^{(t)}\|}$$

where $\mathbf{w}$ *is the direction of the max-margin linear classifier, and* $\mathbf{v}^{(t)}$ *is the direction of the linear classifier learnt by the gradient descent at step* t *when minimizing the logistic loss.*

Notice that the simplicity bias is characterized as margin maximization. This result was later extended to the cases of nonseparable data (Ji and Telgarsky, 2019) and one-hidden-layer neural networks (Lyu et al., 2021).

3 Retrieval-Augmented Models

Machine Learning practitioners came up with a viable alternative to [complex models + simplicity bias] to achieve benign overfitting with [not so complex models + retrieval mechanism]. In this approach, the neural network, instead of trying to memorize rare/atypical examples in the wilds of its parameters, explicitly writes them to the storage and then retrieves them on the inference when necessary.

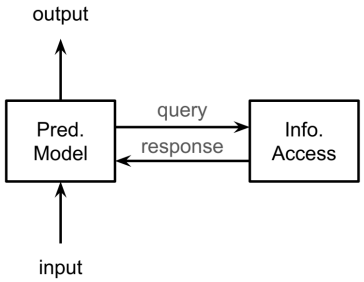


Figure 5: RAM overview. Source: Zamani et al. (2022)

Retrieval-augmented models (RAMs) refer to models composed of two coupled components (Fig. 5): one model that makes predictions by communicating with another model mediating access to a repository of information or knowledge. Zamani et al. (2022) define RAM as $f_\theta(x; R_\omega)$. The model f_θ parameterized by θ is called the prediction model and R_ω denotes the information access model parameterized by ω. Thus, to produce $\hat{y}$, the prediction model can interface with the information access model. R_ω includes a collection or repository C that is available—through an information access model—to the prediction model. C reflects a large set of parameters available to the model that can be leveraged ad hoc.

RAMs implementations for the language modeling problem turned out to be particularly successful, since natural language text is long-tailed data and, accordingly, memorizing part of the training set is inevitable for optimal generalization to new data. For example, REALM of Guu et al. (2020) combines a masked language model with a differentiable retriever, which allows the model to retrieve and attend over documents from a large corpus such as Wikipedia, used during pre-training, fine-tuning and inference. The effectiveness of REALM was demonstrated by fine-tuning on an open-domain question answering task. Khandelwal et al. (2020) introduced kNN-LM, where a retrieval mechanism is used to find the nearest neighbor tokens given the prefix as query. kNN-LM linearly interpolates the predicted distribution for the next token using distance information from the retrieval mechanism. This idea has also been extended to machine translation (Khandelwal et al., 2021). It was shown that retrieval augmentation improves domain adaptation by using a domain-specific datastore for retrieval. The recently introduced RETRO model (Borgeaud et al., 2022) combines a frozen BERT (Devlin et al., 2019) retriever, a differentiable encoder and a chunked cross-attention mechanism to predict tokens based on a 2 trillion token database. RETRO obtains comparable performance to GPT-3 (Brown et al., 2020), despite using 25$\times$ fewer parameters. After fine-tuning, RETRO performance translates to downstream knowledge-intensive tasks such as question answering.

4 Proposed Research Directions

Currently learning theorists prove benign overfitting as an *implication* of margin maximization and light-tailedness of data distributions within classes under assumptions like those from Theorem 1. For example, Chatterji and Long (2021) showed that an over-parameterized max-margin linear classifier trained on a linearly separable-with-noise data can perfectly fit the training sample, yet generalize nearly optimally. A similar result was shown by Shamir (2022), and extensions to neural networks with one hidden dense layer and one hidden convolutional layer were recently given by Frei et al. (2022) and Cao et al. (2022) respectively.

4.1 Questions on Long-Tailedness

We are mainly concerned with the assumptions made in these works: the setup of binary classification with light-tailed distributions within classes is very different from what Feldman (2020) suggested in his long tail theory. Recall that the key point in Feldman's theory is the huge number of classes with long-tailed frequency distribution over classes (or the huge number of long-tailed subpopulations within few classes). According to Feldman (2020), in this case, memorization of some training examples is *necessary* for optimal performance

on test data. However, for those setups where benign overfitting is now being proven, there are non-overfitting models that generalize just as well. Accordingly, it is not clear why benign overfitting is needed in the first place. Thus, one of the research directions can be as follows.

Question 1. *Develop a mathematical framework for the analysis of gradient-based learning algorithms that aligns with Feldman (2020)'s long tail theory. Prove the simplicity bias of SGD and benign overfitting of overparametrized neural networks within such a framework. Show that overparameterization is necessary to achieve optimal performance.*

The recent work by Bubeck and Sellke (2021) is a first possible step towards solving Task 1. In it, the authors show that when learning parameterized classes, overparametrization is necessary for *smooth* interpolation of the training set. However, it remains unclear whether smooth interpolation gives the optimal generalizing ability, i.e. whether there is benign overfitting. Moreover, the problem of regression, not classification, is considered, and therefore it is not entirely clear how Feldman's assumptions about the nature of the data can be integrated into this framework.

4.2 Questions on RAMs

Another concern with the existing analyzes of the benign overfitting is that too much emphasis is placed on pure neural networks, while ignoring RAMs as more efficient way of achieving benign overfitting. Empirically, RAMs have lower generalization error than baseline models without retrieval. At the same time, RAMs memorize (a compressed version of) the training sample in the storage. This brings us to the next

Question 2. *Study benign overfitting for RAMs within the framework established in Question 1. Study their generalization error bounds and computational complexity compared to models without retrieval.*

We emphasize that the study of benign overfitting and the generalizing ability of retrieval-augmented learning should be carried out under the assumptions of Feldman (2020)'s long tail theory. We suspect that the empirical success of RAMs is precisely based on this property, which is inherent in natural data. We hypothesize that in simplified setups (for example, in the case of light-tailed distribution of data), the gain of RAMs compared to models without a retrieval will be minimal or even zero. This conjecture is based on the fact that RAMs are more efficient when there is a need for memorization, which in turn naturally arises from the long tail assumption in Feldman's theory.

After building a mathematical framework and analyzing RAMs, it would be desirable to apply the gained knowledge to improve their performance and/or interpretability. In this regard, we note that in RAMs, the prediction model is usually a deep neural net and as such it still has enough capacity to partially memorize the training sample. If we already have an explicit memory in the form of the information access model can we decrease memorization capacity of the prediction model? This leads us to the following

Question 3. *How can we explicitly control the memorization capacity of the prediction model in RAMs and force it to focus on generalization while shifting memorization to the information access model?*

One way to do this experimentally is to penalize the prediction model when it tries to memorize examples from the training set. In this case, the penalty term should be based on the degree to which the neural network memorizes training examples.

A successful solution of Question 3 will give a control over memorization in prediction part of RAMs. Potentially such control could have applications such as integrating domain-specific (say medical) knowledge into a general model (say neural machine translation, NMT), isolating a generic language-independent part from a multilingual NMT, etc.

Moreover such control would make it possible to ask the following question: can we control RAM's runtime at inference by controlling its complexity and storage size as in the work of Latifi et al. (2022)?

5 Conclusion

As far as we know, our work is the first attempt to establish the chain of reasoning illustrated in Fig. 1. Moreover, Questions 1–3 we propose are novel and relevant. Their solution will provide a better understanding of the mechanics of retrieval-augmented models, and potentially motivate learning theorists to shift their focus towards more efficient approaches for benign overfitting, which ML practitioners currently use in applied problems.

Limitations

We identify two main limitations of our work:

1. This is an opinion paper. As such, some of the hypotheses have not yet been validated. For example, after Question 2, we argue that the advantage of RAMs over non-retrieval models will be minimal or even zero on light-tailed data. However, this assertion has not yet been supported by empirical evidence, because RAMs are usually trained and evaluated on natural data with a long tail, and as far as we know, they have not yet been evaluated on data with a light tail.

2. Due to page limitations, we were unable to provide a more comprehensive overview of the work on benign overfitting, simplicity bias, long-tail theory, and RAMs. However, we believe we have been able to touch on the most influential work in each of these areas.

References

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. 2021. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201.

Mikhail Belkin. 2021. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sébastien Bubeck and Mark Sellke. 2021. A universal law of robustness via isoperimetry. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 28811–28822.

Yuan Cao, Zixiang Chen, Mikhail Belkin, and Quanquan Gu. 2022. Benign overfitting in two-layer convolutional neural networks. *arXiv preprint arXiv:2202.06526*.

Niladri S. Chatterji and Philip M. Long. 2021. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *J. Mach. Learn. Res.*, 22:129:1–129:30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Vitaly Feldman. 2020. Does learning require memorization? a short tale about a long tail. In *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 954–959. ACM.

Spencer Frei, Niladri S. Chatterji, and Peter L. Bartlett. 2022. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 2668–2703. PMLR.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Ziwei Ji and Matus Telgarsky. 2019. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 1772–1798. PMLR.

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin L. Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. 2019. SGD on neural networks learns functions of increasing complexity. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3491–3501.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Salar Latifi, Saurav Muralidharan, and Michael Garland. 2022. Efficient sparsely activated transformers. *CoRR*, abs/2208.14580.

Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. 2021. Gradient descent on two-layer nets: Margin maximization and simplicity bias. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 12978–12991.

Chris Mingard, Guillermo Valle Pérez, Joar Skalse, and Ard A. Louis. 2021. Is SGD a bayesian sampler? well, almost. *J. Mach. Learn. Res.*, 22:79:1–79:64.

Ohad Shamir. 2022. The implicit bias of benign overfitting. In *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 448–478. PMLR.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. 2018. The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.*, 19:70:1–70:57.

Leslie G Valiant. 1984. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

Vladimir Vapnik and Alexey Chervonenkis. 1974. Theory of pattern recognition.

Hamed Zamani, Fernando Diaz, Mostafa Dehghani, Donald Metzler, and Michael Bendersky. 2022. Retrieval-enhanced machine learning. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2875–2886. ACM.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.