Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа Прикладной Математики и Информатики
Кафедра интеллектуальных систем

Направление подготовки / специальность: 03.04.01 Прикладные математика и физика
Направленность (профиль) подготовки: Математическая физика, компьютерные технологии и математическое моделирование в экономике

КОМПРЕССИЯ КАШИНА ДЛЯ РАСПРЕДЕЛЁННОГО ОБУЧЕНИЯ

(магистерская диссертация)

Студент:
Шульгин Егор Владимирович

____
(подпись студента)

Научный руководитель:
Гасников Александр Владимирович,
д-р физ.-мат. наук, доц.

____
(подпись научного руководителя)

Консультант *(при наличии)*:
____
____
(подпись консультанта)

Москва 2021

Аннотация

Данная магистерская диссертация основана на статье «Uncertainty Principle for Communication Compression in Distributed and Federated Learning and the Search for an Optimal Compressor» [35] за авторством Мера Сафаряна, Егора Шульгина и Питера Рихтарика.

Для снижения высоких затрат на связь при распределенном и федеративном обучении, стали очень популярными различные схемы сжатия векторов, такие как квантизация и разрежение. При разработке метода сжатия необходимо передавать как можно меньше битов для минимизации затрат на ранд коммуникации, и в то же время вносимое искажение («дисперсия») в передаваемые сообщения должно быть как можно меньше для минимизации неблагоприятного эффекта сжатия на общее количество раундов коммуникации. Однако интуитивно эти две цели принципиально противоречат друг другу: чем большее сжатие мы допускаем, тем более искаженными становятся сообщения. Мы формализуем эту интуицию и доказываем принцип неопределенности для рандомизированных операторов сжатия, таким образом количественно оценивая это ограничение математически и показывая асимптотически точные нижние границы того, что может быть достигнуто с помощью сжатия коммуникации. Это мотивирует нас поставить задачу по поиску оптимального оператора сжатия. Делая первый шаг в этом направлении, мы рассматриваем метод несмещенного сжатия, основанный на векторном представлении Кашина, который мы называем сжатием Кашина. В отличие от всех ранее предложенных механизмов сжатия, компрессия Кашина имеет не зависящую от размерности оценку дисперсии.

Abstract

This master's thesis is based on paper «Uncertainty Principle for Communication Compression in Distributed and Federated Learning and the Search for an Optimal Compressor» [35] by Mher Safaryan, Egor Shulgin, Peter Richtarik.

In order to mitigate the high communication cost in distributed and federated learning, various vector compression schemes, such as quantization, sparsification and dithering, have become very popular. In designing a compression method, one aims to communicate as few bits as possible, which minimizes the cost per communication round, while at the same time attempting to impart as little distortion (variance) to the communicated messages as possible, which minimizes the adverse effect of the compression on the overall number of communication rounds. However, intuitively, these two goals are fundamentally in conflict: the more compression we allow, the more distorted the messages become. We formalize this intuition and prove an *uncertainty principle* for randomized compression operators, thus quantifying this limitation mathematically, and *effectively providing asymptotically tight lower bounds on what might be achievable with communication compression*. Motivated by these developments, we call for the search for the optimal compression operator. In an attempt to take a first step in this direction, we consider an unbiased compression method inspired by the Kashin representation of vectors, which we call *Kashin compression (KC)*. In contrast to all previously proposed compression mechanisms, KC enjoys a *dimension independent* variance bound for which we derive an explicit formula even in the regime when only a few bits need to be communicate per each vector entry.

Table of Contents

References

[1] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems 30*, pages 1709–1720, 2017.

[2] Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems*, pages 5977–5987, 2018.

[3] Ron Bekkerman, Mikhail Bilenko, and John Langford. *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press, 2011.

[4] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. 80:560–569, 2018.

[5] Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD with majority vote is communication efficient and fault tolerant. 2019.

[6] Sebastian Caldas, Jakub Konečný, H. Brendan McMahan, and Ameet Talwalkar. Expanding the reach of federated learning by reducing client resource requirements. *arXiv:1812.07210*, 2019.

[7] E. J. Candès and T. Tao. Decoding by linear programming. 51, 2005.

[8] E. J. Candès and T. Tao. Near-optimal signal recovery from random projections and universal encoding strategies. 52, 2006.

[9] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

[10] D Gabor. Theory of communication. *Journal of the Institute of Electrical Engineering*, 93:429–457, 1946.

[11] A. A. Giannopoulos and V. Milman. Concentration property on probability spaces. *Advances in Mathematics*, 156:77–106, 2000.

[12] W. M. Goodall. Television by pulse code modulation. *The Bell System Technical Journal*, 30(1):33–49, Jan 1951. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1951.tb01365.x.

[13] Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: training ImageNet in 1 hour. *CoRR*, abs/1706.02677, 2017.

[14] V. Havin and B. Jöricke. *The Uncertainty Principle in Harmonic Analysis*. Springer-Verlag, 1994.

[15] Werner Heisenberg. Über den anschaulichen inhalt der quantentheoretischen kinematik und mechanik. *Zeitschrift für Physik*, 43(3–4):172–198, 1927.

[16] Samuel Horváth, Chen-Yu Ho, L'udovít Horváth, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural compression for distributed deep learning. *arXiv:1905.10988*, 2019.

[17] Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*, 2019.

[18] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for on-device federated learning. *ArXiv*, abs/1910.06378, 2019.

[19] Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 3252–3261, 2019.

[20] Boris S. Kashin. Diameters of some finite-dimensional sets and classes of smooth functions. *Jour. Izv. Akad. Nauk SSSR Ser. Mat.*, 41(2):334–351, 1977. URL http://mi.mathnet.ru/izv1805.

[21] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local SGD on identical and heterogeneous data. *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, 2020.

[22] Sarit Khirirat, Hamid Reza Feyzmahdavian, and Mikael Johansson. Distributed learning with compressed gradients. 2018.

[23] Martin Kochol. Constructive approximation of a ball by polytopes. *Mathematica Slovaca*, 44(1):99–105, 1994. ISSN 0139-9918. URL https://eudml.org/doc/34376.

[24] Martin Kochol. A note on approximation of a ball by polytopes. *Discrete Optimization*, 1(2):229 – 231, 2004. ISSN 1572-5286. doi: https://doi.org/10.1016/j.disopt.2004.07.003. URL http://www.sciencedirect.com/science/article/pii/S1572528604000295.

[25] Jakub Konečný and Peter Richtárik. Randomized distributed mean estimation: accuracy vs communication. *Frontiers in Applied Mathematics and Statistics*, 4(62):1–11, 2018.

[26] Jakub Konečný, H. Brendan McMahan, Felix Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: strategies for improving communication efficiency. *NIPS Private Multi-Party Machine Learning Workshop*, 2016.

[27] Michel Ledoux. *The Concentration of Measure Phenomenon*. American Mathematical Society, 2001.

[28] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: challenges, methods, and future directions. *arXiv preprint arXiv:1908.07873*, 2019.

[29] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J. Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. 2018.

[30] Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. signSGD via zeroth-order oracle. *International Conference on Learning Representations*, 2019.

[31] Yurii Lyubarskii and Roman Vershynin. Uncertainty principles and vector quantization. *IEEE Trans. Inf. Theor.*, 56(7):3491–3501, July 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2048458. URL http://dx.doi.org/10.1109/TIT.2010.2048458.

[32] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[33] L. Roberts. Picture coding using pseudo-random noise. *IRE Transactions on Information Theory*, 8(2):145–154, February 1962. ISSN 0096-1000. doi: 10.1109/TIT.1962.1057702.

[34] Mher Safaryan and Peter Richtárik. On stochastic sign descent methods. *arXiv preprint arXiv:1905.12938*, 2019.

[35] Mher Safaryan, Egor Shulgin, and Peter Richtárik. Uncertainty principle for communication compression in distributed and federated learning and the search for an optimal compressor. *Information and Inference: A Journal of the IMA*, 04 2021. ISSN 2049-8772. doi: 10.1093/imaiai/iaab006. URL `https://doi.org/10.1093/imaiai/iaab006`. iaab006.

[36] Jürgen Schmidhuber. Deep learning in neural networks: An overview. In *Neural networks*, volume 61, page 85–117, 2015.

[37] Sebastian U. Stich. Local SGD converges fast and communicates little. *arXiv:1805.09767*, 2018.

[38] Hanlin Tang, Chen Yu, Xiangru Lian, Tong Zhang, and Ji Liu. `DoubleSqueeze`: Parallel stochastic gradient descent with double-pass error-compensated compression. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6155–6165, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/tang19d.html`.

[39] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *PMLR*, pages 1195–1204, 2019.

[40] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

[41] Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical low-rank gradient compression for distributed optimization. *CoRR*, abs/1905.13727, 2019. URL `http://arxiv.org/abs/1905.13727`.

[42] Hongyi Wang, Scott Sievert, Shengchao Liu, Zachary Charles, Dimitris Papailiopoulos, and Stephen Wright. Atomo: Communication-efficient learning via atomic sparsification. *Advances in Neural Information Processing Systems*, 2018.

[43] Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, pages 1306–1316, 2018.

[44] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in Neural Information Processing Systems*, pages 1509–1519, 2017.

[45] Hantian Zhang, Jerry Li, Kaan Kara, Dan Alistarh, Ji Liu, and Ce Zhang. ZipML: Training linear models with end-to-end low precision, and a little bit of deep learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, page 4035–4043, 2017.