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: 03.03.01 Прикладные математика и физика
Направленность (профиль) подготовки: Математическая физика, компьютерные технологии и математическое моделирование в экономике

РАСПРЕДЕЛЕННЫЕ МЕТОДЫ ВТОРОГО ПОРЯДКА С БЫСТРОЙ СКОРОСТЬЮ СХОДИМОСТИ И КОМПРЕССИЕЙ

(бакалаврская работа)

Студент:
Исламов Рустем Ильфакович

____
(подпись студента)

Научный руководитель:
Стрижов Вадим Викторович,
д-р физ.-мат. наук

____
(подпись научного руководителя)

Консультант *(при наличии)***:**
Питер Рихтарик

____
(подпись консультанта)

Москва 2021

Аннотация

Данная бакалаврская диссертация основана на статье «Distributed Second Order Methods with Fast Rates and Compressed Communication» [15] за авторством Рустема Исламова, Шуна Кяна и Питера Рихтарика.

Мы разработали несколько новых эффективных с точки зрения коммуникации методов второго порядка для распределенной оптимизации. Наш первый метод, NEWTON-STAR, является модификацией метода Ньютона, от которого он наследует свою локальную квадратичную сходимость. Кроме этого, NEWTON-STAR имеет ту же стоимость коммуницирования, что и градиентный спуск. Хотя этот метод непрактичен, поскольку опирается на использование неизвестных параметров, характеризующих Гессиан целевой функции в оптимуме, он служит отправной точкой для создания практического метода с доказанными теоретическими гарантиями сходимости. Мы разработали стратегию обучения неизвестных параметров, основанную на использовании случайной разреженности. Применение этой стратегии к NEWTON-STAR приводит к следующему методу, NEWTON-LEARN, для которого мы доказали локальные линейные и сверхлинейные скорости сходимости, не зависящие от числа обусловленности функции. Когда эти методы применимы, они имеют значительно более высокие скорости сходимости по сравнению с современными методами. Теоретические результаты подкреплены экспериментами на реальных наборах данных и показывают превосходство на несколько порядков по сравнению с классическими методами с точки зрения коммуницирования.

Abstract

This bachelor thesis is based on paper “Distributed Second Order Methods with Fast Rates and Compressed Communication” [15] written by Rustem Islamov, Xun Qian, and Peter Richtárik.

We develop new communication-efficient second-order method for distributed optimization. Our first method, NEWTON-STAR, is a variant of Newton's method from which it inherits its fast local quadratic rate. However, unlike Newton's method, NEWTON-STAR enjoys the same per iteration communication cost as gradient descent. While this method is impractical as it relies on the use of certain unknown parameters characterizing the Hessian of the objective function at the optimum, it serves as the starting point which enables us design practical variants thereof with strong theoretical guarantees. In particular, we design a stochastic sparsification strategy for learning the unknown parameters in an iterative fashion in a communication efficient manner. Applying this strategy to NEWTON-STAR leads to our next method, NEWTON-LEARN, for which we prove local linear and superlinear rates independent of the condition number. When applicable, this method can have dramatically superior convergence behavior when compared to state-of-the-art methods. Our results are supported with experimental results on real datasets, and show several orders of magnitude improvement on baseline and state-of-the-art methods in terms of communication complexity.

References

[1] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, pages 1709–1720, 2017.

[2] Amir Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB.* Society for Industrial and Applied Mathematics, USA, 2014. ISBN 1611973643.

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

[4] J. Bernstein, Y. X. Wang, K. Azizzadenesheli, and A. Anandkumar. SignSGD: Compressed optimisation for non-convex problems. *The 35th International Conference on Machine Learning*, pages 560–569, 2018.

[5] Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *arXiv:2002.12410*, 2020.

[6] Charles G Broyden. Quasi-newton methods and their application to function minimisation. *Mathematics of Computation*, 21(99):368–381, 1967.

[7] Rixon Crane and Fred Roosta. DINGO: Distributed Newton-type method for gradient-norm optimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 9498–9508. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/9718db12cae6be37f7349779007ee589-Paper.pdf`.

[8] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1646–1654. Curran Associates, Inc., 2014.

[9] Rodger Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–323, 1970.

[10] Avishek Ghosh, Raj Kumar Maity, Arya Mazumdar, and Kannan Ramchandran. Communication efficient distributed approximate Newton method. In *IEEE International Symposium on Information Theory (ISIT)*, 2020. doi: 10.1109/ISIT44484.2020.9174216.

[11] Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.

[12] Andreas Griewank. The modification of Newton's method for unconstrained optimization by bounding cubic terms. Technical report, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, 1981. Technical Report NA/12.

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

[14] 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.

[15] Rustem Islamov, Xun Qian, and Peter Richtárik. Distributed second order methods with fast rates and compressed communication. *arXiv preprint arXiv:2102.07158*, 2021.

[16] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pages 315–323, 2013.

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

[18] Sarit Khirirat, Hamid Reza Feyzmahdavian, and Mikael Johansson. Distributed learning with compressed gradients. In *arXiv preprint arXiv:1806.06573*, 2018.

[19] Dmitry Kovalev, Konstanting Mishchenko, and Peter Richtárik. Stochastic Newton and cubic Newton methods with simple local linear-quadratic rates. In *NeurIPS Beyond First Order Methods Workshop*, 2019.

[20] Dmitry Kovalev, Robert M. Gower, Peter Richtárik, and Alexander Rogozin. Fast linear convergence of randomized BFGS. *arXiv:2002.11337*, 2020.

[21] Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtárik. Acceleration for compressed gradient descent in distributed and federated optimization. In *International Conference on Machine Learning*, 2020.

[22] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. *arXiv preprint arXiv:1506.02186*, 2015.

[23] Nicolas Loizou and Peter Richtárik. Linearly convergent stochastic heavy ball method for minimizing generalization error. In *NIPS Workshop on Optimization for Machine Learning*, 2017.

[24] Nicolas Loizou and Peter Richtárik. Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. *arXiv:1712.09677*, 2017.

[25] Chenxin Ma, Jakub Konečný, Martin Jaggi, Virginia Smith, Michael I. Jordan, Peter Richtárik, and Martin Takáč. Distributed optimization with arbitrary local solvers. *Optimization Methods and Software*, 32(4):813–848, 2017.

[26] Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6702–6712. PMLR, 13–18 Jul 2019.

[27] Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019.

[28] Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysiswith vast improvements. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

[29] Deanna Needell, Nathan Srebro, and Rachel Ward. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. *Mathematical Programming*, 155(1–2):549–573, 2015.

[30] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[31] Yurii Nesterov and Boris T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[32] Boris Polyak and Andrey Tremba. New versions of Newton method: step-size choice, convergence domain and under-determined equations. *arXiv preprint arXiv:1703.07810*, 2019.

[33] Boris Polyak and Andrey Tremba. New versions of newton method: step-size choice, convergence domain and under-determined equations. *Optimization Methods and Software*, 35(6):1272–1303, 2020.

[34] Josepho Raphson. Analysis aequationum universalis seu ad aequationes algebraicas resolvendas methodus generalis, & expedita, ex nova infinitarum serierum methodo, deducta ac demonstrata. *Oxford: Richard Davis*, 1697.

[35] Sashank J. Reddi, Jakub Konečný, Peter Richtárik, Barnabás Póczos, and Alex Smola. AIDE: fast and communication efficient distributed optimization. *arXiv:1608.06879*, 2016.

[36] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[37] M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Math. Program.*, 162(1-2):83–112, 2017.

[38] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu. 1-bit stochastic gradient descent and its application to data- parallel distributed training of speech DNNs. *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[39] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: from theory to algorithms*. Cambridge University Press, 2014.

[40] Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate Newton-type method. In *Proceedings of the 31st International Conference on Machine Learning, PMLR*, volume 32, pages 1000–1008, 2014.

[41] David F Shanno. Conditioning of quasi-Newton methods for function minimization. *Mathematics of computation*, 24(111):647–656, 1970.

[42] S. U. Stich and S. P. Karimireddy. The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication. *arXiv: 1909.05350*, 2019.

[43] Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2020.

[44] Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems*, pages 4447–4458, 2018.

[45] Martin Takáč, Avleen Bijral, Peter Richtárik, and Nathan Srebro. Mini-batch primal and dual methods for SVMs. In *30th International Conference on Machine Learning*, pages 537–552, 2013.

[46] H. Tang, X. Lian, T. Zhang, and J. Liu. DoubleSqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6155–6165, 2019.

[47] John Wallis. A treatise of algebra, both historical and practical. *Philosophical Transactions of the Royal Society of London*, 15(173):1095–1106, 1685. doi: 10.1098/rstl.1685.0053. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rstl.1685.0053`.

[48] Shusen Wang, Fred Roosta, Peng Xu, and Michael W Mahoney. Giant: Globally improved approximate newton method for distributed optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 2332–2342. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/file/dabd8d2ce74e782c65a973ef76fd540b-Paper.pdf`.

[49] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, and H. Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. *Advances in Neural Information Processing Systems*, pages 1509–1519, 2017.

[50] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

[51] Jiaqi Zhang, Keyou You, and Tamer Başar. Distributed adaptive Newton methods with globally superlinear convergence. *arXiv preprint arXiv:2002.07378*, 2020.

[52] Yuchen Zhang and Lin Xiao. DiSCO: Distributed optimization for self-concordant empirical loss. In *Proceedings of the 32nd International Conference on Machine Learning, PMLR*, volume 37, pages 362–370, 2015.

[53] Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling. *The 32nd International Conference on Machine Learning*, 37:1–9, 2015.