جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

ARTIFICIAL
INTELLIGENCE
INITIATIVE

Recent Advances in Optimization for Machine Learning*

Peter Richtárik

* By the OPTML Lab at KAUST

Optimization and Machine Learning Lab

Photo: February 2019

Research Scientists
Laurent Condat (from Grenoble)
Zhize Li (from Tsinghua)

Postdocs
Mher Safaryan (from Yerevan)
Adil Salim (from Télécom Paris)
Xun Qian (from Hong Kong)

PhD Students
Filip Hanzely (now Assistant Prof @ TTIC)
Konstantin Mishchenko (from ENS Paris-Saclay)
Alibek Sailanbayev (from Nazarbayev)
Samuel Horváth (from Comenius)
Elnur Gasanov (from MIPT)
Dmitry Kovalev (from MIPT)
Konstantin Burlachenko (from Huawei)
Slavomír Hanzely (from Comenius)
Lukang Sun (from Nanjing)

MS Students
Egor Shulgin (from MIPT)
Grigory Malinovsky (from MIPT)
Igor Sokolov (from MIPT)

Research Interns
Ilyas Fatkhullin (from Munich)
Rustem Islamov (from MIPT)
Bokun Wang (from UC Davis)
Eduard Gorbunov (from MIPT)
Ahmed Khaled (from Cairo)

Papers Since 2019

2021

[160] G. Malinovsky, A. Sailanbayev and P. Richtárik
Random reshuffling with variance reduction: new analysis and better rates

[159] A. Salim, L. Condat, D. Kovalev and P. Richtárik
An optimal algorithm for strongly convex minimization under affine constraints

[158] Zhen Shi, N. Loizou, P. Richtárik and M. Takáč
AI-SARAH: Adaptive and implicit stochastic recursive gradient methods

[157] D. Kovalev, E. Shulgin, P. Richtárik, A. Rogozin and A. Gasnikov
ADOM: Accelerated decentralized optimization method for time-varying networks
NSF-TRIPODS Workshop: Communication Efficient Distributed Optimization

[156] K. Mishchenko, B. Wang, D. Kovalev and P. Richtárik
IntSGD: Floatless compression of stochastic gradients

[155] E. Gorbunov, K. Burlachenko, Z. Li and P. Richtárik
MARINA: faster non-convex distributed learning with compression

[154] M. Safaryan, F. Hanzely and P. Richtárik
Smoothness matrices beat smoothness constants: better communication compression techniques for distributed optimization
ICLR 2021 (Workshop: Distributed and Private Machine Learning)
NSF-TRIPODS Workshop: Communication Efficient Distributed Optimization

[153] R. Islamov, X. Qian and P. Richtárik
Distributed second order methods with fast rates and compressed communication
NSF-TRIPODS Workshop: Communication Efficient Distributed Optimization

[152] K. Mishchenko, A. Khaled and P. Richtárik
Proximal and federated random reshuffling
NSF-TRIPODS Workshop: Communication Efficient Distributed Optimization

2020

[151] S. Horváth, A. Klein, P. Richtárik and C. Archambeau
Hyperparameter transfer learning with adaptive complexity
AISTATS 2021

[150] X. Qian, H. Dong, P. Richtárik and T. Zhang
Error compensated loopless SVRG for distributed optimization
NeurIPS 2020 (12th Annual Workshop on Optimization for ML)

[149] X. Qian, H. Dong, P. Richtárik and T. Zhang
Error compensated proximal SGD and RDA
NeurIPS 2020 (12th Annual Workshop on Optimization for ML)

[148] E. Gorbunov, F. Hanzely and P. Richtárik
Local SGD: unified theory and new efficient methods
AISTATS 2021

[147] D. Kovalev, A. Koloskova, M. Jaggi, P. Richtárik and S.U. Stich
A linearly convergent algorithm for decentralized optimization: sending less bits for free!
AISTATS 2021

[146] W. Chen, S. Horváth and P. Richtárik
Optimal client sampling for federated learning
NeurIPS 2020 (Workshop on Privacy Preserving Machine Learning)

[145] E. Gorbunov, D. Kovalev, D. Makarenko and P. Richtárik
Linearly converging error compensated SGD
NeurIPS 2020

[144] Alyazeed Albasyoni, M. Safaryan, L. Condat and P. Richtárik
Optimal gradient compression for distributed and federated learning
NeurIPS2020 (Scalability, Privacy and Security in Federated Learning)

[143] F. Hanzely, Slavomír Hanzely, S. Horváth and P. Richtárik
Lower bounds and optimal algorithms for personalized federated learning
NeurIPS 2020

[142] L. Condat, G. Malinovsky and P. Richtárik
Distributed proximal splitting algorithms with rates and acceleration
NeurIPS 2020 (12th Annual Workshop on Optimization for ML)

[141] R. M. Gower, M. Schmidt, F. Bach and P. Richtárik
Variance-reduced methods for machine learning
Proceedings of the IEEE, 2020

[140] X. Qian, P. Richtárik and T. Zhang
Error compensated distributed SGD can be accelerated
NeurIPS 2020 (12th Annual Workshop on Optimization for ML)

[139] A. S. Berahas, M. Jahani, P. Richtárik and M. Takáč
Quasi-Newton methods for deep learning: forget the past, just sample

[138] Z. Li, H. Bao, X. Zhang and P. Richtárik
PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization
NeurIPS 2020 (12th Annual Workshop on Optimization for ML)

[137] D. Kovalev, A. Salim and P. Richtárik
Optimal and practical algorithms for smooth and strongly convex decentralized optimization
NeurIPS 2020

[136] A. Khaled, O. Sebbouh, N. Loizou, R. M. Gower and P. Richtárik
Unified analysis of stochastic gradient methods for composite convex and smooth optimization

[135] S. Horváth and P. Richtárik
A better alternative to error feedback for communication-efficient distributed learning
ICLR 2021

[134] A. Salim and P. Richtárik
Primal dual interpretation of the proximal stochastic gradient Langevin algorithm
NeurIPS 2020

[133] Z. Li and P. Richtárik
A unified analysis of stochastic gradient methods for nonconvex federated optimization
NeurIPS 2020 (Scalability, Privacy and Security in Federated Learning)

[132] K. Mishchenko, A. Khaled and P. Richtárik
Random reshuffling: simple analysis with vast improvements
NeurIPS 2020

[131] M. Alfarra, S. Hanzely, A. Albasyoni, B. Ghanem and P. Richtárik
Adaptive learning of the optimal mini-batch size of SGD
NeurIPS 2020 (12th Annual Workshop on Optimization for ML)

[130] A. Salim, L. Condat, K. Mishchenko and P. Richtárik
Dualize, split, randomize: fast nonsmooth optimization algorithms
NeurIPS 2020 (12th Annual Workshop on Optimization for ML)

[129] A. N. Sahu, A. Dutta, A. Tiwari and P. Richtárik
On the convergence analysis of asynchronous SGD for solving consistent linear systems

[128] G. Malinovsky, D. Kovalev, E. Gasanov, L. Condat and P. Richtárik
From local SGD to local fixed point methods for federated learning
ICML 2020

[127] A. Beznosikov, S. Horváth, P. Richtárik and M. Safaryan
On biased compression for distributed learning
NeurIPS 2020 (Scalability, Privacy and Security in Federated Learning)

[126] Z. Li, D. Kovalev, X. Qian and P. Richtárik
Acceleration for compressed gradient descent in distributed and federated optimization
ICML 2020

[125] D. Kovalev, R. M. Gower, P. Richtárik and A. Rogozin
Fast linear convergence of randomized BFGS

[124] F. Hanzely, Nikita Doikov, P. Richtárik and Yurii Nesterov
Stochastic subspace cubic Newton method
ICML 2020

[123] M. Safaryan, E. Shulgin and P. 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, 2021

[122] F. Hanzely and P. Richtárik
Federated learning of a mixture of global and local models
NeurIPS 2020 (Scalability, Privacy and Security in Federated Learning)

[121] S. Horváth, L. Lei, P. Richtárik and M. I. Jordan
Adaptivity of stochastic gradient methods for nonconvex optimization
NeurIPS 2020 (12th Annual Workshop on Optimization for ML)

[120] F. Hanzely, D. Kovalev and P. Richtárik
Variance reduced coordinate descent with acceleration: new method with a surprising application to finite-sum problems
ICML 2020

[119] A. Khaled and P. Richtárik
Better theory for SGD in the nonconvex world

2019

[118] A. Khaled, K. Mishchenko and P. Richtárik
Tighter theory for local SGD on identical and heterogeneous data
AISTATS 2020

[117] S. Chraibi, A. Khaled, D. Kovalev, A. Salim, P. Richtárik and M. Takáč
Distributed fixed point methods with compressed iterates

[116] S. Horváth, C.-Y. Ho, Ľ. Horváth, A.N. Sahu, M. Canini and P. Richtárik
IntML: Natural compression for distributed deep learning
SOSP 2019 (Workshop on AI Systems)

[115] D. Kovalev, K. Mishchenko and P. Richtárik
Stochastic Newton and cubic Newton methods with simple local linear-quadratic rates
NeurIPS 2019 (Beyond First Order Methods in ML)

[114] A. Khaled, K. Mishchenko and P. Richtárik
Better communication complexity for local SGD
NeurIPS 2019 (Fed. Learning for Data Privacy and Confidentiality)

[113] A. Khaled and P. Richtárik
Gradient descent with compressed iterates
NeurIPS 2019 (Fed. Learning for Data Privacy and Confidentiality)

[112] A. Khaled, K. Mishchenko and P. Richtárik
First analysis of local GD on heterogeneous data
NeurIPS 2019 (Fed. Learning for Data Privacy and Confidentiality)

[111] J. Xiong, P. Richtárik and W. Heidrich
Stochastic convolutional sparse coding
Int. Symposium on Vision, Modeling and Visualization 2019

[110] X. Qian, Z. Qu and P. Richtárik
L-SVRG and L-Katyusha with arbitrary sampling
Journal of Machine Learning Research, 2021

[109] X. Qian, A. Sailanbayev, K. Mishchenko and P. Richtárik
MISO is making a comeback with better proofs and rates

[108] E. Gorbunov, A. Bibi, O. Sezer, E. H. Bergou and P. Richtárik
A stochastic derivative free optimization method with momentum
ICLR 2020

[107] M. Safaryan and P. Richtárik
On stochastic sign descent methods
NeurIPS 2020 (12th Annual Workshop on Optimization for ML)

[106] A. Salim, D. Kovalev and P. Richtárik
Stochastic proximal Langevin algorithm: potential splitting and nonasymptotic rates
NeurIPS 2019

[105] A. Dutta, E. H. Bergou, Y. Xiao, M. Canini and P. Richtárik
Direct nonlinear acceleration

[104] K. Mishchenko and P. Richtárik
A stochastic decoupling method for minimizing the sum of smooth and non-smooth functions

[103] K. Mishchenko, D. Kovalev, E. Shulgin, P. Richtárik and Y. Malitsky
Revisiting stochastic extragradient
AISTATS 2020

[102] F. Hanzely and P. Richtárik
One method to rule them all: variance reduction for data, parameters and many new methods

[101] E. Gorbunov, F. Hanzely and P. Richtárik
A unified theory of SGD: variance reduction, sampling, quantization and coordinate descent
AISTATS 2020

[100] S. Horváth, C.-Y. Ho, Ľ. Horváth, A. N. Sahu, M. Canini and P. Richtárik
Natural compression for distributed deep learning

[99] R. M. Gower, D. Kovalev, F. Lieder and P. Richtárik
RSN: Randomized Subspace Newton
NeurIPS 2019

[98] A. Dutta, F. Hanzely, J. Liang and P. Richtárik
Best pair formulation & accelerated scheme for non-convex principal component pursuit
IEEE Transactions on Signal Processing, 2020

[97] N. Loizou and P. Richtárik
Revisiting randomized gossip algorithms: general framework, convergence rates and novel block and accelerated protocols

[96] N. Loizou and P. Richtárik
Convergence analysis of inexact randomized iterative methods
SIAM Journal on Scientific Computing, 2020

[95] A. Sapio, M. Canini, C.-Y. Ho, J. Nelson, P. Kalnis, C. Kim, A. Krishnamurthy, M. Moshref, D. R. K. Ports and P. Richtárik
Scaling distributed machine learning with in-network aggregation
NSDI 2021

[94] S. Horváth, D. Kovalev, K. Mishchenko, P. Richtárik and S. Stich
Stochastic distributed learning with gradient quantization and variance reduction

[93] E. H. Bergou, E. Gorbunov and P. Richtárik
Stochastic three points method for unconstrained smooth minimization
SIAM Journal on Optimization, 2020

[92] A. Bibi, E. H. Bergou, O. Sener, B. Ghanem and P. Richtárik
A stochastic derivative-free optimization method with importance sampling
AAAI 2020

[91] K. Mishchenko, F. Hanzely and P. Richtárik
99% of distributed optimization is a waste of time: the issue and how to fix it
UAI 2020

[90] K. Mishchenko, E. Gorbunov, M. Takáč and P. Richtárik
Distributed learning with compressed gradient differences

[89] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin and P. Richtárik
SGD: general analysis and improved rates
ICML 2019

[88] D. Kovalev, S. Horváth and P. Richtárik
Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop
ALT 2020

[87] X. Qian, Z. Qu and P. Richtárik
SAGA with arbitrary sampling
ICML 2019

NeurIPS: Neural Inf. Process. Systems
ICML: Int. Conf. on Machine Learning
AISTATS: Artificial Intellig. & Statistics
ICLR: Int. Conf. on Learning Represent.
JMLR: J. Machine Learning Research
ALT: Algorithmic Learning Theory
UAI: Uncertainty in AI
AAAI: Conference on AI
SIAM, IEEE and IMA Journals
NSDI: USENIX Symp. on Networked Systems Design and Implementation
SOSP: Symp. Operating Syst. Principles

17 Spotlight (5 min) Talks

2021

[160] G. Malinovsky, A. Sailanbayev and P. Richtárik
Random reshuffling with variance reduction: new analysis and better rates

[159] A. Salim, L. Condat, D. Kovalev and P. Richtárik
An optimal algorithm for strongly convex minimization under affine constraints

[158] Zhen Shi, N. Loizou, P. Richtárik and M. Takáč
AI-SARAH: Adaptive and implicit stochastic recursive gradient methods

[157] D. Kovalev, E. Shulgin, P. Richtárik, A. Rogozin and A. Gasnikov
ADOM: Accelerated decentralized optimization method for time-varying networks
NSF-TRIPODS Workshop: Communication Efficient Distributed Optimization

[156] K. Mishchenko, B. Wang, D. Kovalev and P. Richtárik
IntSGD: Floatless compression of stochastic gradients

[155] E. Gorbunov, K. Burlachenko, Z. Li and P. Richtárik
MARINA: faster non-convex distributed learning with compression

[154] M. Safaryan, F. Hanzely and P. Richtárik
Smoothness matrices beat smoothness constants: better communication compression techniques for distributed optimization
ICLR 2021 (Workshop: Distributed and Private Machine Learning)
NSF-TRIPODS Workshop: Communication Efficient Distributed Optimization

[153] R. Islamov, X. Qian and P. Richtárik
Distributed second order methods with fast rates and compressed communication
NSF-TRIPODS Workshop: Communication Efficient Distributed Optimization

[152] K. Mishchenko, A. Khaled and P. Richtárik
Proximal and federated random reshuffling
NSF-TRIPODS Workshop: Communication Efficient Distributed Optimization

2020

[151] S. Horváth, A. Klein, P. Richtárik and C. Archambeau
Hyperparameter transfer learning with adaptive complexity
AISTATS 2021

[150] X. Qian, H. Dong, P. Richtárik and T. Zhang
Error compensated loopless SVRG for distributed optimization
NeurIPS 2020 (12th Annual Workshop on Optimization for ML)

[149] X. Qian, H. Dong, P. Richtárik and T. Zhang
Error compensated proximal SGD and RDA
NeurIPS 2020 (12th Annual Workshop on Optimization for ML)

[148] E. Gorbunov, F. Hanzely and P. Richtárik
Local SGD: unified theory and new efficient methods
AISTATS 2021

[147] D. Kovalev, A. Koloskova, M. Jaggi, P. Richtárik and S.U. Stich
A linearly convergent algorithm for decentralized optimization: sending less bits for free!
AISTATS 2021

[146] W. Chen, S. Horváth and P. Richtárik
Optimal client sampling for federated learning
NeurIPS 2020 (Workshop on Privacy Preserving Machine Learning)

[145] E. Gorbunov, D. Kovalev, D. Makarenko and P. Richtárik
Linearly converging error compensated SGD
NeurIPS 2020

[144] Alyazeed Albasyoni, M. Safaryan, L. Condat and P. Richtárik
Optimal gradient compression for distributed and federated learning
NeurIPS2020 (Scalability, Privacy and Security in Federated Learning)

[143] F. Hanzely, Slavomír Hanzely, S. Horváth and P. Richtárik
Lower bounds and optimal algorithms for personalized federated learning
NeurIPS 2020

[142] L. Condat, G. Malinovsky and P. Richtárik
Distributed proximal splitting algorithms with rates and acceleration
NeurIPS 2020 (12th Annual Workshop on Optimization for ML)

[141] R. M. Gower, M. Schmidt, F. Bach and P. Richtárik
Variance-reduced methods for machine learning
Proceedings of the IEEE, 2020

[140] X. Qian, P. Richtárik and T. Zhang
Error compensated distributed SGD can be accelerated
NeurIPS 2020 (12th Annual Workshop on Optimization for ML)

[139] A. S. Berahas, M. Jahani, P. Richtárik and M. Takáč
Quasi-Newton methods for deep learning: forget the past, just sample

[138] Z. Li, H. Bao, X. Zhang and P. Richtárik
PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization
NeurIPS 2020 (12th Annual Workshop on Optimization for ML)

[137] D. Kovalev, A. Salim and P. Richtárik
Optimal and practical algorithms for smooth and strongly convex decentralized optimization
NeurIPS 2020

[136] A. Khaled, O. Sebbouh, N. Loizou, R. M. Gower and P. Richtárik
Unified analysis of stochastic gradient methods for composite convex and smooth optimization

[135] S. Horváth and P. Richtárik
A better alternative to error feedback for communication-efficient distributed learning
ICLR 2021

[134] A. Salim and P. Richtárik
Primal dual interpretation of the proximal stochastic gradient Langevin algorithm
NeurIPS 2020

[133] Z. Li and P. Richtárik
A unified analysis of stochastic gradient methods for nonconvex federated optimization
NeurIPS 2020 (Scalability, Privacy and Security in Federated Learning)

[132] K. Mishchenko, A. Khaled and P. Richtárik
Random reshuffling: simple analysis with vast improvements
NeurIPS 2020

[131] M. Alfarra, S. Hanzely, A. Albasyoni, B. Ghanem and P. Richtárik
Adaptive learning of the optimal mini-batch size of SGD
NeurIPS 2020 (12th Annual Workshop on Optimization for ML)

[130] A. Salim, L. Condat, K. Mishchenko and P. Richtárik
Dualize, split, randomize: fast nonsmooth optimization algorithms
NeurIPS 2020 (12th Annual Workshop on Optimization for ML)

[129] A. N. Sahu, A. Dutta, A. Tiwari and P. Richtárik
On the convergence analysis of asynchronous SGD for solving consistent linear systems

[128] G. Malinovsky, D. Kovalev, E. Gasanov, L. Condat and P. Richtárik
From local SGD to local fixed point methods for federated learning
ICML 2020

[127] A. Beznosikov, S. Horváth, P. Richtárik and M. Safaryan
On biased compression for distributed learning
NeurIPS 2020 (Scalability, Privacy and Security in Federated Learning)

[126] Z. Li, D. Kovalev, X. Qian and P. Richtárik
Acceleration for compressed gradient descent in distributed and federated optimization
ICML 2020

[125] D. Kovalev, R. M. Gower, P. Richtárik and A. Rogozin
Fast linear convergence of randomized BFGS

[124] F. Hanzely, Nikita Doikov, P. Richtárik and Yurii Nesterov
Stochastic subspace cubic Newton method
ICML 2020

[123] M. Safaryan, E. Shulgin and P. 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, 2021

[122] F. Hanzely and P. Richtárik
Federated learning of a mixture of global and local models
NeurIPS 2020 (Scalability, Privacy and Security in Federated Learning)

[121] S. Horváth, L. Lei, P. Richtárik and M. I. Jordan
Adaptivity of stochastic gradient methods for nonconvex optimization
NeurIPS 2020 (12th Annual Workshop on Optimization for ML)

[120] F. Hanzely, D. Kovalev and P. Richtárik
Variance reduced coordinate descent with acceleration: new method with a surprising application to finite-sum problems
ICML 2020

[119] A. Khaled and P. Richtárik
Better theory for SGD in the nonconvex world

2019

[118] A. Khaled, K. Mishchenko and P. Richtárik
Tighter theory for local SGD on identical and heterogeneous data
AISTATS 2020

[117] S. Chraibi, A. Khaled, D. Kovalev, A. Salim, P. Richtárik and M. Takáč
Distributed fixed point methods with compressed iterates

[116] S. Horváth, C.-Y. Ho, Ľ. Horváth, A.N. Sahu, M. Canini and P. Richtárik
IntML: Natural compression for distributed deep learning
SOSP 2019 (Workshop on AI Systems)

[115] D. Kovalev, K. Mishchenko and P. Richtárik
Stochastic Newton and cubic Newton methods with simple local linear-quadratic rates
NeurIPS 2019 (Beyond First Order Methods in ML)

[114] A. Khaled, K. Mishchenko and P. Richtárik
Better communication complexity for local SGD
NeurIPS 2019 (Fed. Learning for Data Privacy and Confidentiality)

[113] A. Khaled and P. Richtárik
Gradient descent with compressed iterates
NeurIPS 2019 (Fed. Learning for Data Privacy and Confidentiality)

[112] A. Khaled, K. Mishchenko and P. Richtárik
First analysis of local GD on heterogeneous data
NeurIPS 2019 (Fed. Learning for Data Privacy and Confidentiality)

[111] J. Xiong, P. Richtárik and W. Heidrich
Stochastic convolutional sparse coding
Int. Symposium on Vision, Modeling and Visualization 2019

[110] X. Qian, Z. Qu and P. Richtárik
L-SVRG and L-Katyusha with arbitrary sampling
Journal of Machine Learning Research, 2021

[109] X. Qian, A. Sailanbayev, K. Mishchenko and P. Richtárik
MISO is making a comeback with better proofs and rates

[108] E. Gorbunov, A. Bibi, O. Sezer, E. H. Bergou and P. Richtárik
A stochastic derivative free optimization method with momentum
ICLR 2020

[107] M. Safaryan and P. Richtárik
On stochastic sign descent methods
NeurIPS 2020 (12th Annual Workshop on Optimization for ML)

[106] A. Salim, D. Kovalev and P. Richtárik
Stochastic proximal Langevin algorithm: potential splitting and nonasymptotic rates
NeurIPS 2019

[105] A. Dutta, E. H. Bergou, Y. Xiao, M. Canini and P. Richtárik
Direct nonlinear acceleration

[104] K. Mishchenko and P. Richtárik
A stochastic decoupling method for minimizing the sum of smooth and non-smooth functions

[103] K. Mishchenko, D. Kovalev, E. Shulgin, P. Richtárik and Y. Malitsky
Revisiting stochastic extragradient
AISTATS 2020

[102] F. Hanzely and P. Richtárik
One method to rule them all: variance reduction for data, parameters and many new methods

[101] E. Gorbunov, F. Hanzely and P. Richtárik
A unified theory of SGD: variance reduction, sampling, quantization and coordinate descent
AISTATS 2020

[100] S. Horváth, C.-Y. Ho, Ľ. Horváth, A. N. Sahu, M. Canini and P. Richtárik
Natural compression for distributed deep learning

[99] R. M. Gower, D. Kovalev, F. Lieder and P. Richtárik
RSN: Randomized Subspace Newton
NeurIPS 2019

[98] A. Dutta, F. Hanzely, J. Liang and P. Richtárik
Best pair formulation & accelerated scheme for non-convex principal component pursuit
IEEE Transactions on Signal Processing, 2020

[97] N. Loizou and P. Richtárik
Revisiting randomized gossip algorithms: general framework, convergence rates and novel block and accelerated protocols

[96] N. Loizou and P. Richtárik
Convergence analysis of inexact randomized iterative methods
SIAM Journal on Scientific Computing, 2020

[95] A. Sapio, M. Canini, C.-Y. Ho, J. Nelson, P. Kalnis, C. Kim, A. Krishnamurthy, M. Moshref, D. R. K. Ports and P. Richtárik
Scaling distributed machine learning with in-network aggregation
NSDI 2021

[94] S. Horváth, D. Kovalev, K. Mishchenko, P. Richtárik and S. Stich
Stochastic distributed learning with gradient quantization and variance reduction

[93] E. H. Bergou, E. Gorbunov and P. Richtárik
Stochastic three points method for unconstrained smooth minimization
SIAM Journal on Optimization, 2020

[92] A. Bibi, E. H. Bergou, O. Sener, B. Ghanem and P. Richtárik
A stochastic derivative-free optimization method with importance sampling
AAAI 2020

[91] K. Mishchenko, F. Hanzely and P. Richtárik
99% of distributed optimization is a waste of time: the issue and how to fix it
UAI 2020

[90] K. Mishchenko, E. Gorbunov, M. Takáč and P. Richtárik
Distributed learning with compressed gradient differences

[89] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin and P. Richtárik
SGD: general analysis and improved rates
ICML 2019

[88] D. Kovalev, S. Horváth and P. Richtárik
Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop
ALT 2020

[87] X. Qian, Z. Qu and P. Richtárik
SAGA with arbitrary sampling
ICML 2019

+ Samuel Horváth:
Federated Learning under Heterogeneous Clients

+ Adil Salim:
Complexity Analysis of Stein Variational Gradient Descent

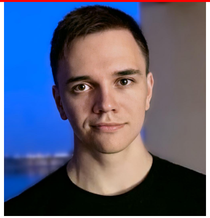
17 Spotlight (5 min) Talks


Konstantin Burlachenko
CS PhD Student


Samuel Horváth
STAT PhD Student


Ahmed Khaled
Intern (Cairo) -> PhD @ Caltech


Filip Hanzely
Assistant Professor @ TTIC


Konstantin Mishchenko
CS PhD Student


Laurent Condat
VCC Research Scientist


Dmitry Kovalev
CS PhD Student


Eduard Gorbunov
Intern (MIPT)


Egor Shulgin
CS MS/PhD Student


Bokun Wang
Intern (UC Davis)


Grigory Malinovsky
AMCS MS/PhD Student


Adil Salim
Postdoc -> Berkeley

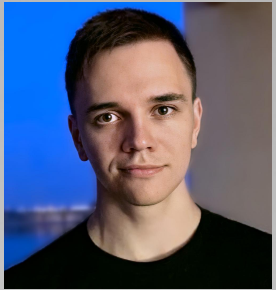

Mher Safaryan
Postdoc


Xun Qian
VCC/ECRC Postdoc


Rustem Islamov
Intern (MIPT) -> PhD @ Paris


Slavomír Hanzely
AMCS PhD Student


Zhize Li
VCC/ECRC Research Scientist



Konstantin Mishchenko
(KAUST PhD Student)

Ahmed Khaled
(KAUST Intern -> Caltech PhD)

NEURAL INFORMATION PROCESSING SYSTEMS

Random Reshuffling: Simple Analysis with Vast Improvements

Konstantin Mishchenko
KAUST
Thuwal, Saudi Arabia

Ahmed Khaled
Cairo University
Giza, Egypt

Peter Richtárik
KAUST
Thuwal, Saudi Arabia

Abstract

Random Reshuffling (RR) is an algorithm for minimizing finite-sum functions that utilizes iterative gradient descent steps in conjunction with data reshuffling. Often contrasted with its sibling Stochastic Gradient Descent (SGD), RR is usually faster in practice and enjoys significant popularity in convex and non-convex optimization. The convergence rate of RR has attracted substantial attention recently and, for strongly convex and smooth functions, it was shown to converge faster than SGD if 1) the stepsize is small, 2) the gradients are bounded, and 3) the number of epochs is large. We remove these 3 assumptions, improve the dependence on the condition number from κ^2 to κ (resp. from κ to $\sqrt{\kappa}$) and, in addition, show that RR has a different type of variance. We argue through theory and experiments that the new variance type gives an additional justification of the superior performance of RR. To go beyond strong convexity, we present several results for non-strongly convex and non-convex objectives. We show that in all cases, our theory improves upon existing literature. Finally, we prove fast convergence of the Shuffle-Once (SO) algorithm, which shuffles the data only once, at the beginning of the optimization process. Our theory for strongly convex objectives tightly matches the known lower bounds for both RR and SO and substantiates the common practical heuristic of shuffling once or only a few times. As a byproduct of our analysis, we also get new results for the Incremental Gradient algorithm (IG), which does not shuffle the data at all.

1 Introduction

We study the finite-sum minimization problem

$$\min_{x\in\mathbb{R}^d}\left[f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)\right], \qquad (1)$$

where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is differentiable and smooth, and are particularly interested in the big data machine learning setting where the number of functions n is large. Thanks to their scalability and low memory requirements, first-order methods are especially popular in this setting (Bottou et al., 2018). *Stochastic* first-order algorithms in particular have attracted a lot of attention in the machine learning community and are often used in combination with various practical heuristics. Explaining these heuristics may lead to further development of stable and efficient training algorithms. In this work, we aim at better and sharper theoretical explanation of one intriguingly simple but notoriously elusive heuristic: *data permutation/shuffling*.

1.1 Data permutation

In particular, the goal of our paper is to obtain deeper theoretical understanding of methods for solving (1) which rely on random or deterministic *permutation/shuffling* of the data $\{1, 2, \dots, n\}$ and

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Part I

RANDOM RESHUFFLING: SIMPLE ANALYSIS WITH VAST IMPROVEMENTS

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of Science and Technology

ARTIFICIAL INTELLIGENCE INITIATIVE

Peter Richtárik

Random Reshuffling: Simple Analysis With Vast Improvements

Problem: Find Model Which Minimizes Prediction Loss on Training Data

Training data $\{1, 2, \ldots, n\}$

Loss of model *x* on *i*th data

$$\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

d features/parameters representing a ML model

$f(x) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$

Theorem (Strongly Convex Case)

Model trained after *t* data passes

learning rate $\gamma \leq \frac{1}{L}$

A new notion: "shuffling variance"

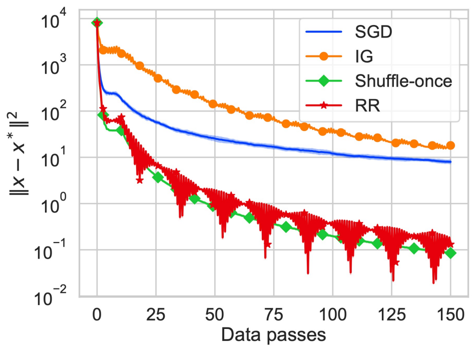
$$\mathbb{E}\left[\left\|x^{tn} - x^\star\right\|^2\right] \leq (1 - \gamma\mu)^{tn} \left\|x^0 - x^\star\right\|^2 + \frac{2\gamma\sigma^2_{\text{Shuffle}}}{\mu}$$

solution

Strong convexity parameter

- **Dramatically new proof technique**
- Better dependence on *n* and condition number
- New notion: **shuffling variance**
- Variant Shuffle-Once: **tightly matches lower bound** of Safran and Shamir (2020)

Rajput et al (2020): ***"Current theoretical bounds [for Shuffle-Once] are insufficient to explain this phenomenon, and a new theoretical breakthrough may be required to tackle it."***



Algorithm: How to Choose the Next Training Data Point to Learn From?

1. Sampling **With** Replacement, aka **Stochastic Gradient Descent (SGD)**

$$x^{k+1} = x^k - \gamma \nabla f_{i^k}(x^k)$$

Example for $n = 5$: $\{i^1, i^2, i^3, i^4, i^5\} = \{3, 2, 2, 1, 3\}$

- **unbiased** gradient estimator $\mathbb{E}\left[\nabla f_{i^k}(x^k) \mid x^k\right] = \nabla f(x^k)$
- thousands of papers since 1950s
- **well understood**

2. Sampling **Without** Replacement, aka **Random Reshuffling (RR)**

$$x^{k+1} = x^k - \gamma \nabla f_{\pi^k}(x^k)$$

Example for $n = 5$: $\{\pi^1, \pi^2, \pi^3, \pi^4, \pi^5\} = \{4, 3, 1, 2, 5\}$

- **biased** gradient estimator $\mathbb{E}\left[\nabla f_{\pi^k}(x^k) \mid x^k\right] \neq \nabla f(x^k)$
- a handful of papers only!
- **not understood**
- **default in deep learning software**

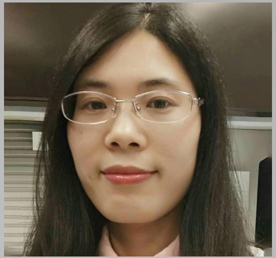
Our Theoretical Rates Significantly Improve on SOTA (in strongly convex, convex and also nonconvex regimes)

Assumptions N.L.(1)	Assumptions U.V.(2)	μ-Strongly Convex	Non-Strongly Convex	Non-Convex	Citation
✓	✓	$\kappa^2 n + \frac{\kappa n \sigma_*}{\mu\sqrt{\varepsilon}}$	–	–	Ying et al. (2019)
✗	✗	$\kappa^2 n + \frac{\kappa\sqrt{n}G}{\mu\sqrt{\varepsilon}}$	$\frac{LD^2}{\varepsilon} + \frac{G^2D^2}{\varepsilon^2}$ (3)	–	Nagaraj et al. (2019)
✗	✗	–	–	$\frac{Ln}{\varepsilon^2} + \frac{LnG}{\varepsilon^3}$	Nguyen et al. (2020)
✓	✓	$\frac{\kappa^2 n}{\sqrt{\mu\varepsilon}} + \frac{\kappa^2 n\sigma_*}{\mu\sqrt{\varepsilon}}$ (4)	–	–	Nguyen et al. (2020)
✗	✗	$\frac{\kappa\alpha}{\varepsilon^{1/\alpha}} + \frac{\kappa\sqrt{n}G\alpha^{3/2}}{\mu\sqrt{\varepsilon}}$ (5)	–	–	Ahn and Sra (2020)
✓	✓	$\kappa n + \frac{\sqrt{n}}{\sqrt{\mu\varepsilon}} + \frac{\kappa\sqrt{n}G_0}{\mu\sqrt{\varepsilon}}$ (6)	–	–	Ahn et al. (2020)
✓	✓	$\kappa + \frac{\sqrt{\kappa n}\sigma_*}{\mu\sqrt{\varepsilon}}$ (7) $\kappa n + \frac{\sqrt{\kappa n}\sigma_*}{\mu\sqrt{\varepsilon}}$	$\frac{Ln}{\varepsilon} + \frac{\sqrt{Ln}\sigma_*}{\varepsilon^{3/2}}$	$\frac{Ln}{\varepsilon^2} + \frac{L\sqrt{n}(B+\sqrt{A})}{\varepsilon^3}$	This work

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology
ARTIFICIAL INTELLIGENCE INITIATIVE

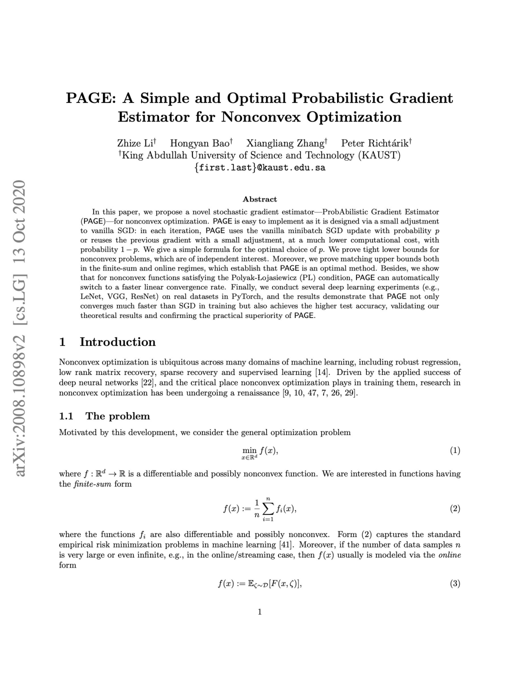
Peter Richtárik

Zhize Li
(KAUST Research Scientist)



Hongyan Bao
(KAUST PhD Student)

Xiangliang Zhang
(KAUST Associate Professor)



arXiv:2008.10898v2 [cs.LG] 13 Oct 2020

PAGE: A Simple and Optimal Probabilistic Gradient Estimator for Nonconvex Optimization

Zhize Li[†] Hongyan Bao[†] Xiangliang Zhang[†] Peter Richtárik[†]
[†]King Abdullah University of Science and Technology (KAUST)
{first.last}@kaust.edu.sa

Abstract

In this paper, we propose a novel stochastic gradient estimator—ProbAbilistic Gradient Estimator (PAGE)—for nonconvex optimization. PAGE is easy to implement as it is designed via a small adjustment to vanilla SGD: in each iteration, PAGE uses the vanilla minibatch SGD update with probability p or reuses the previous gradient with a small adjustment, at a much lower computational cost, with probability $1-p$. We give a simple formula for the optimal choice of p. We prove tight lower bounds for nonconvex problems, which are of independent interest. Moreover, we prove matching upper bounds both in the finite-sum and online regimes, which establish that PAGE is an optimal method. Besides, we show that for nonconvex functions satisfying the Polyak-Łojasiewicz (PL) condition, PAGE can automatically switch to a faster linear convergence rate. Finally, we conduct several deep learning experiments (e.g., LeNet, VGG, ResNet) on real datasets in PyTorch, and the results demonstrate that PAGE not only converges much faster than SGD in training but also achieves the higher test accuracy, validating our theoretical results and confirming the practical superiority of PAGE.

1 Introduction

Nonconvex optimization is ubiquitous across many domains of machine learning, including robust regression, low rank matrix recovery, sparse recovery and supervised learning [14]. Driven by the applied success of deep neural networks [22], and the critical place nonconvex optimization plays in training them, research in nonconvex optimization has been undergoing a renaissance [9, 10, 47, 7, 26, 29].

1.1 The problem

Motivated by this development, we consider the general optimization problem

$$\min_{x\in\mathbb{R}^d} f(x), \tag{1}$$

where $f:\mathbb{R}^d \to \mathbb{R}$ is a differentiable and possibly nonconvex function. We are interested in functions having the *finite-sum* form

$$f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{2}$$

where the functions f_i are also differentiable and possibly nonconvex. Form (2) captures the standard empirical risk minimization problems in machine learning [41]. Moreover, if the number of data samples n is very large or even infinite, e.g., in the online/streaming case, then $f(x)$ usually is modeled via the *online* form

$$f(x) := \mathbb{E}_{\zeta\sim\mathcal{D}}[F(x,\zeta)], \tag{3}$$

1

Part II

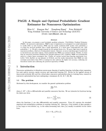
PAGE: A SIMPLE AND OPTIMAL PROBABILISTIC GRADIENT ESTIMATOR FOR NONCONVEX OPTIMIZATION

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

ARTIFICIAL
INTELLIGENCE
INITIATIVE

Peter Richtárik

PAGE: A Simple and Optimal Probabilistic Gradient Estimator for Nonconvex Optimization



Problem: Train a ML Model on 1 Machine Using Minimal # of Data Samples

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

Training data $\{1,2,\dots,n\}$

Loss of model x on i^{th} data

d features/parameters representing a ML model

$$f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

PAGE

$$x^{k+1} = x^k - \gamma g^k$$

Learning rate

Stochastic gradient estimator want: $g^k \approx \nabla f(x^k)$

Full gradient (requires 1 pass over all data)

Probability of computing the full gradient (should be small)

$$g^{k+1} = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x^{k+1}) & \text{with probability } p \\ g^k + \frac{1}{|S^k|}\sum_{i\in S^k} \left(\nabla f_i(x^{k+1}) - \nabla f_i(x^k)\right) & \text{with probability } 1-p \end{cases}$$

Minibatch $S^k \subset \{1,2,\dots,n\}$ Random subset

PAGE is a **biased estimator** of the gradient:

$$\mathbb{E}\left[g^k \mid x^k, x^{k-1}\right] \neq \nabla f(x^k)$$

Comparison to Existing Results

Table 1: Gradient complexity for finding $\widehat{x}$ satisfying $\mathbb{E}\|\nabla f(\widehat{x})\| \le \epsilon$ in nonconvex problems

Problem	Assumption	Algorithm or Lower Bound	Gradient complexity
Finite-sum (2)	Asp. 2	GD [34]	$O(\frac{n}{\epsilon^2})$
Finite-sum (2)	Asp. 2	SVRG [3, 40], SCSG [24], SVRG+ [27]	$O(n+\frac{n^{2/3}}{\epsilon^2})$
Finite-sum (2)	Asp. 2	SNVRG [47], Geom-SARAH [13]	$\widetilde{O}\left(n+\frac{\sqrt{n}}{\epsilon^2}\right)$
Finite-sum (2)	Asp. 2	SPIDER [7], SpiderBoost [43], SARAH [37], SSRGD [26]	$O(n+\frac{\sqrt{n}}{\epsilon^2})$
Finite-sum (2)	Asp. 2	PAGE (this paper)	$O(n+\frac{\sqrt{n}}{\epsilon^2})$
Finite-sum (2)	Asp. 2	Lower bound [7]	$\Omega(\frac{\sqrt{n}}{\epsilon^2})$ if $n \le O(\frac{1}{\epsilon^4})$
Finite-sum (2)	Asp. 2	Lower bound (this paper)	$\Omega(n+\frac{\sqrt{n}}{\epsilon^2})$
Finite-sum (2)	Asp. 2 and 3	PAGE (this paper)	$O\left((n+\sqrt{n}\kappa)\log\frac{1}{\epsilon}\right)$ [1]
Online (3) [2]	Asp. 1 and 2	SGD [10, 16, 29]	$O(\frac{\sigma^2}{\epsilon^4})$
Online (3)	Asp. 1 and 2	SCSG [24], SVRG+ [27]	$O(b+\frac{b^{2/3}}{\epsilon^2})$
Online (3)	Asp. 1 and 2	SNVRG [47], Geom-SARAH [13]	$\widetilde{O}\left(b+\frac{\sqrt{b}}{\epsilon^2}\right)$
Online (3)	Asp. 1 and 2	SPIDER [7], SpiderBoost [43], SARAH [37], SSRGD [26]	$O(b+\frac{\sqrt{b}}{\epsilon^2})$
Online (3)	Asp. 1 and 2	PAGE (this paper)	$O(b+\frac{\sqrt{b}}{\epsilon^2})$ [3]
Online (3)	Asp. 1 and 2	Lower bound (this paper)	$\Omega(b+\frac{\sqrt{b}}{\epsilon^2})$
Online (3)	Asp. 1, 2 and 3	PAGE (this paper)	$O\left((b+\sqrt{b}\kappa)\log\frac{1}{\epsilon}\right)$

Assumptions:

1. f_i can be nonconvex
2. f is lower bounded
3. f_i is "smooth"

$$\|\nabla f_i(x) - \nabla f_i(y)\| \le L_i\|x-y\| \quad \forall x,y\in\mathbb{R}^d$$

Goal: Find random vector $\widehat{x}$ such that

$$\mathbb{E}\left[\|\nabla f(\widehat{x})\|^2\right] \le \varepsilon^2$$

Theorem

PAGE solves the problem using

$$\mathcal{O}\left(n + \frac{\sqrt{n}}{\varepsilon^2}\right)$$ data samples

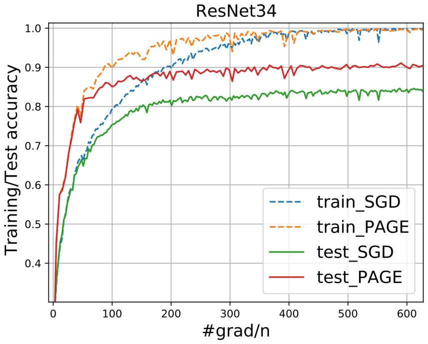
= # of stochastic gradient evaluations

with either of these two parameter choices:

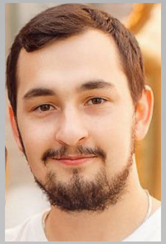
A. $|S^k| = 1$ and $p = \frac{1}{1+n}$

B. $|S^k| = \sqrt{n}$ and $p = \frac{1}{1+\sqrt{n}}$

We prove that PAGE is "optimal" (= in a precise mathematical sense, this is the best gradient-type method for solving smooth nonconvex problems)



جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology
ARTIFICIAL INTELLIGENCE INITIATIVE
Peter Richtárik



Eduard Gorbunov
(KAUST Intern)

Konstantin Burlachenko
(KAUST PhD Student)

Zhize Li
(KAUST Research Scientist)

arXiv:2102.07845v1 [cs.LG] 15 Feb 2021

MARINA: Faster Non-Convex Distributed Learning with Compression

Eduard Gorbunov[1,2,3] Konstantin Burlachenko[3] Zhize Li[3] Peter Richtárik[3]

[1] Moscow Institute of Physics and Technology, Russia
[2] Institute for Information Transmission Problems RAS, Russia
[3] King Abdullah University of Science and Technology, Kingdom of Saudi Arabia

Abstract

We develop and analyze MARINA: a new communication efficient method for non-convex distributed learning over heterogeneous datasets. MARINA employs a novel communication compression strategy based on the compression of gradient differences which is reminiscent of but different from the strategy employed in the DIANA method of Mishchenko et al (2019). Unlike virtually all competing distributed first-order methods, including DIANA, ours is based on a carefully designed *biased* gradient estimator, which is the key to its superior theoretical and practical performance. To the best of our knowledge, the communication complexity bounds we prove for MARINA are strictly superior to those of all previous first order methods. Further, we develop and analyze two variants of MARINA: VR-MARINA and PP-MARINA. The first method is designed for the case when the local loss functions owned by clients are either of a finite sum or of an expectation form, and the second method allows for partial participation of clients – a feature important in federated learning. All our methods are superior to previous state-of-the-art methods in terms of the oracle/communication complexity. Finally, we provide convergence analysis of all methods for problems satisfying the Polyak-Łojasiewicz condition.

Contents

Part III

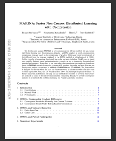
MARINA: FASTER NON-CONVEX DISTRIBUTED LEARNING WITH (COMMUNICATION) COMPRESSION

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of Science and Technology

ARTIFICIAL INTELLIGENCE INITIATIVE

Peter Richtárik

MARINA: Faster Non-convex Distributed Learning with (Communication) Compression



Problem: Train a ML Model on n Machines Using Minimal # of Bits Communciated by the n Workers to the Master

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

- # of machines
- Loss of model x on data stored on machine i
- d features/parameters representing a ML model

$$f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

MARINA

$$x^{k+1} = x^k - \gamma g^k$$

- Learning rate (γ)
- Stochastic gradient estimator (g^k). want: $g^k \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} g_i^k \approx \nabla f(x^k)$

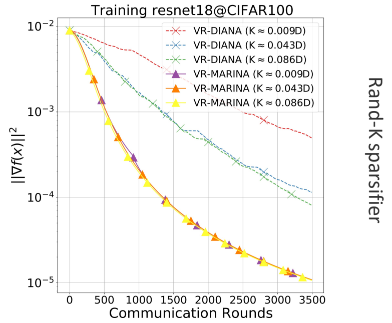
$$g_i^{k+1} = \begin{cases} \nabla f_i(x^{k+1}) & \text{with probability } p \\ g^k + \mathcal{Q}_i^k\left(\nabla f_i(x^{k+1}) - \nabla f_i(x^k)\right) & \text{with probability } 1-p \end{cases}$$

- Uncompressed gradient: $\nabla f_i(x^{k+1})$
- Probability of sending an uncompressed vector (should be small): p
- Random **compression operator** $\mathcal{Q}_i^k$: $\mathbb{E}\left[\mathcal{Q}_i^k(g)\right] = g \qquad \mathbb{E}\left[\|\mathcal{Q}_i^k(g)\|^2\right] \leq (\omega+1)\|g\|^2$

MARINA uses a **biased estimator** of the gradient:

$$\mathbb{E}\left[g^k \mid x^k, x^{k-1}\right] \neq \nabla f(x^k)$$

Comparison to Previous SOTA



Assumptions:

1. f_i can be nonconvex
2. f is lower bounded
3. f_i is "smooth"

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i\|x-y\| \quad \forall x,y\in\mathbb{R}^d$$

Goal: Find random vector $\widehat{x}$ such that

$$\mathbb{E}\left[\|\nabla f(\widehat{x})\|^2\right] \leq \varepsilon^2$$

Theorem (simplified)

If $\mathcal{Q}_i^k$ is the Rand-1 sparsifier, and $p = \frac{1}{d+1}$, then MARINA solves the problem using

$$\mathcal{O}\left(\frac{1+d/\sqrt{n}}{\varepsilon^2}\right)$$

communicated bits / machine

Previous SOTA:
Mishchenko et al 2019; Horváth et al 2019; Li & R. 2020

Gradient Descent

$$\mathcal{O}\left(\frac{d}{\varepsilon^2}\right)$$

DIANA

$$\mathcal{O}\left(\frac{1+d^{3/2}/\sqrt{n}}{\varepsilon^2}\right)$$

Comparison to Existing Results: MARINA is SOTA

Setup	Method	Citation	Communication Complexity	Oracle Complexity
(1)	DIANA	[Mishchenko et al., 2019] [Horváth et al., 2019] [Li and Richtárik, 2020]	$\frac{1+(1+\omega)\sqrt{\omega/n}}{\varepsilon^2}$	$\frac{1+(1+\omega)\sqrt{\omega/n}}{\varepsilon^2}$
	FedCOMGATE (1)	[Haddadpour et al., 2020]	$\frac{1+\omega}{\varepsilon^2}$	$\frac{1+\omega}{n\varepsilon^4}$
	FedSTEPH, $r=n$	[Das et al., 2020]	$\frac{1+\omega/n}{\varepsilon^4}$	$\frac{1+\omega/n}{\varepsilon^4}$
	MARINA (Alg. 1)	Thm. 2.1 & Cor. 2.1 (NEW)	$\frac{1+\omega/\sqrt{n}}{\varepsilon^2}$	$\frac{1+\omega/\sqrt{n}}{\varepsilon^2}$
(1)+(5)	DIANA	[Li and Richtárik, 2020]	$\frac{1+(1+\omega)\sqrt{\omega/n}}{\varepsilon^2} + \frac{1+\omega}{n\varepsilon^4}$	$\frac{1+(1+\omega)\sqrt{\omega/n}}{\varepsilon^2} + \frac{1+\omega}{n\varepsilon^4}$
	VR-DIANA	[Horváth et al., 2019]	$\frac{\left(m^{2/3}+\omega\right)\sqrt{1+\omega/n}}{\varepsilon^2}$	$\frac{\left(m^{2/3}+\omega\right)\sqrt{1+\omega/n}}{\varepsilon^2}$
	VR-MARINA (Alg. 2) $b'=1$ (2)	Thm. 3.1 & Cor. 3.1 (NEW)	$\frac{1+\max\left\{\omega,\sqrt{(1+\omega)m}\right\}/\sqrt{n}}{\varepsilon^2}$	$\frac{1+\max\left\{\omega,\sqrt{(1+\omega)m}\right\}/\sqrt{n}}{\varepsilon^2}$
(1)+(6)	DIANA (3)	[Mishchenko et al., 2019] [Li and Richtárik, 2020]	$\frac{1+(1+\omega)\sqrt{\omega/n}}{\varepsilon^2} + \frac{1+\omega}{n\varepsilon^4}$	$\frac{1+(1+\omega)\sqrt{\omega/n}}{\varepsilon^2} + \frac{1+\omega}{n\varepsilon^4}$
	FedCOMGATE (3)	[Haddadpour et al., 2020]	$\frac{1+\omega}{\varepsilon^2}$	$\frac{1+\omega}{n\varepsilon^4}$
	VR-MARINA (Alg. 2) $b'=1$	Thm. 3.2 & Cor. 3.2 (NEW)	$\frac{1+\omega/\sqrt{n}}{\varepsilon^2} + \frac{\sqrt{1+\omega}}{n\varepsilon^3}$	$\frac{1+\omega/\sqrt{n}}{\varepsilon^2} + \frac{\sqrt{1+\omega}}{n\varepsilon^3}$
	VR-MARINA (Alg. 2) $b'=\Theta\left(\frac{1}{n\varepsilon^2}\right)$	Thm. 3.2 & Cor. 3.2 (NEW)	$\frac{1+\omega/\sqrt{n}}{\varepsilon^2}$	$\frac{1+\omega/\sqrt{n}}{n\varepsilon^4} + \frac{1+\omega}{n\varepsilon^3}$
PP, (1)	FedSTEPH	[Das et al., 2020]	$\frac{1+\omega/n}{r\varepsilon^4} + \frac{(1+\omega)(n-r)}{r(n-1)\varepsilon^4}$	$\frac{1+\omega/n}{r\varepsilon^4} + \frac{(1+\omega)(n-r)}{r(n-1)\varepsilon^4}$
	PP-MARINA (Alg. 4)	Thm. 4.1 & Cor. 4.1 (NEW)	$\frac{1+(1+\omega)\sqrt{n}/r}{\varepsilon^2}$	$\frac{1+(1+\omega)\sqrt{n}/r}{\varepsilon^2}$

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology
ARTIFICIAL INTELLIGENCE INITIATIVE
Peter Richtárik

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology
ARTIFICIAL
INTELLIGENCE
INITIATIVE