Project Description

In this course project, you are asked to select one paper from the list and do a review on it.

You need to summarize the paper, particularly,

a) What is the problem the authors aimed to solve? Why it is important?
b) What are the approaches they took? What is the novelty?
c) What are their results/conclusion?
d) What are the constraints, i.e., imagine you are going to improve the result of this paper, what would you do?

- You need to form a group with a total of no more than 3 members and prepare a presentation with PPT slides on Dec. 27th and Dec. 27th: 15min presentation + 2min Q&A each. (Note: Every member of each group needs to participate in the presentation)

- You also need to submit a report (at most two pages) by Dec. 31st.

Please choose from the following list. If (more than) two group choose the same one, we will assign the paper to the one with an earlier timestamp.

Following are candidate paper list:

[1] Kuzmin, A., Van Baalen, M., Ren, Y., Nagel, M., Peters, J., & Blankevoort, T. (2022). "Fp8 quantization: The power of the exponent," *Advances in Neural Information Processing Systems*, 35, 14651-14662

[2] Blumenfeld, Yaniv, Itay Hubara, and Daniel Soudry. "Towards Cheaper Inference in Deep Networks with Lower Bit-Width Accumulators." In *Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ NeurIPS 2023)*. 2023.

[3] Avron, Haim, Petar Maymounkov, and Sivan Toledo. "Blendenpik: Supercharging LAPACK's least-squares solver." *SIAM Journal on Scientific Computing* 32.3 (2010): 1217-1236.

[4] Rokhlin, Vladimir, and Mark Tygert. "A fast randomized algorithm for overdetermined linear least-squares regression." *Proceedings of the National Academy of Sciences* 105.36 (2008): 13212-13217.

[5] Zouzias, Anastasios, and Nikolaos M. Freris. "Randomized extended Kaczmarz for solving least squares." *SIAM Journal on Matrix Analysis and Applications* 34.2 (2013): 773-793

[6] Sridhar, Srivatsan, Mert Pilanci, and Ayfer Özgür. "Lower bounds and a near-optimal shrinkage estimator for least squares using random projections." *IEEE Journal on Selected Areas in Information Theory* 1.3 (2020): 660-668.

[7] Lacotte, Jonathan, and Mert Pilanci. "Optimal randomized first-order methods for least-squares problems." *International Conference on Machine Learning*. PMLR, 2020.

[8] H. Avron, P. Maymounkov, and S. Toledo, "Blendenpik: Supercharging LAPACK's LeastSquares Solver," *SIAM Journal on Scientific Computing* 32 (January 2010), no. 3, 1217– 1236.

[9] Romanov, Elad. "On the Noise Sensitivity of the Randomized SVD." arXiv preprint arXiv:2305.17435 (2023).

[10] J.A. Duersch and M. Gu, “Randomized QR with column pivoting,” *SIAM Journal on Scientific Computing* 39 (January 2017), no. 4, C263–C291.
[11] C. Musco, C. Musco, and A. Sidford. Stability of the lanczos method for matrix function approximation. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1605–1624. SIAM, 2018.
[12] Meier, Maike, et al. "Are sketch-and-precondition least squares solvers numerically stable?." *arXiv preprint arXiv:2302.07202* (2023).
[13] J. Demmel, L. Grigori, M. Hoemmen, and J. Langou, “Communication-optimal parallel and sequential QR and LU factorizations,” *SIAM Journal on Scientific Computing* 34 (2012), no. 1, A206–A239.
[14] T. Fukaya, R. Kannan, Y. Nakatsukasa, Y. Yamamoto, and Y. Yanagisawa, Shifted Cholesky QR for computing the QR factorization of ill-conditioned matrices, *SIAM Journal on Scientific Computing* 42 (2020), no. 1, A477–A503
[15] P.-G. Martinsson, G. Quintana-Ort´ı, N. Heavner, and R. van de Geijn, “Householder QR factorization with randomization for column pivoting (HQRRP),” *SIAM Journal on Scientific Computing 39* (January 2017), no. 2, C96–C115.
[16] Ubaru, Shashanka, Arya Mazumdar, and Yousef Saad. "Low rank approximation and decomposition of large matrices using error correcting codes." *IEEE Transactions on Information Theory* 63.9 (2017): 5544-5558.
[17] Hsu, Daniel, Sham M. Kakade, and Tong Zhang. "Random design analysis of ridge regression." *Conference on learning theory*. JMLR Workshop and Conference Proceedings, 2012
[18] Gu, Ming. "Subspace iteration randomization and singular value problems." *SIAM Journal on Scientific Computing* 37.3 (2015): A1139-A1173.
[19] Balabanov, Oleg, and Laura Grigori. "Randomized Gram--Schmidt Process with Application to GMRES." *SIAM Journal on Scientific Computing* 44.3 (2022): A1450-A1474.
[20] Coakley, E. S., Vladimir Rokhlin, and Mark Tygert. "A fast randomized algorithm for orthogonal projection." *SIAM Journal on Scientific Computing* 33.2 (2011): 849-868.
[21] Güttel, Stefan, and Marcel Schweitzer. "Randomized sketching for Krylov approximations of large-scale matrix functions." *SIAM Journal on Matrix Analysis and Applications* 44.3 (2023): 1073-1095.
[22] J. Lee, M. Simchowitz, M. Jordan, and B. Recht, ”Gradient descent converges to minimizers,” COLT, 2016.
[23] Andrychowicz, Marcin, et al. "Learning to learn by gradient descent by gradient descent." *Advances in neural information processing systems* 29 (2016).
[24] Byrd, Richard H., et al. "A stochastic quasi-Newton method for large-scale optimization." *SIAM Journal on Optimization* 26.2 (2016): 1008-1031.
[25] Allen-Zhu, Zeyuan. "How to make the gradients small stochastically: Even faster convex and nonconvex sgd." *Advances in Neural Information Processing Systems* 31 (2018).
[26] Panageas, Ioannis, Georgios Piliouras, and Xiao Wang. "First-order methods almost always avoid saddle points: The case of vanishing step-sizes." *Advances in Neural Information Processing Systems* 32 (2019).
[27] Du, Simon, et al. "Gradient descent finds global minima of deep neural networks." *International conference on machine learning*. PMLR, 2019.
[28] Duchi, John, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of machine learning research* 12.7 (2011).
[29] Su, Weijie, Stephen Boyd, and Emmanuel Candes. "A differential equation for modeling Nesterov’s accelerated gradient method: theory and insights." *Advances in neural information processing systems* 27 (2014).
[30] Arora, Sanjeev, Nadav Cohen, and Elad Hazan. "On the optimization of deep networks: Implicit acceleration by overparameterization." *International Conference on Machine Learning*. PMLR, 2018
[31] Bernstein, Jeremy, et al. "signSGD: Compressed optimisation for non-convex problems." *International Conference on Machine Learning*. PMLR, 2018.

[32] Woodworth, Blake, et al. "Is local SGD better than minibatch SGD?." *International Conference on Machine Learning*. PMLR, 2020.

[33] Stich, Sebastian U., Jean-Baptiste Cordonnier, and Martin Jaggi. "Sparsified SGD with memory." *Advances in Neural Information Processing Systems* 31 (2018)

[34] Zhou, Pan, et al. "Towards theoretically understanding why sgd generalizes better than adam in deep learning." *Advances in Neural Information Processing Systems* 33 (2020): 21285-21296.

[35] Roulet, Vincent, and Alexandre d'Aspremont. "Sharpness, restart and acceleration." *Advances in Neural Information Processing Systems* 30 (2017).

[36] Muehlebach, Michael, and Michael Jordan. "A dynamical systems perspective on Nesterov acceleration." *International Conference on Machine Learning*. PMLR, 2019.

[37] Jin, Chi, Praneeth Netrapalli, and Michael I. Jordan. "Accelerated gradient descent escapes saddle points faster than gradient descent." *Conference On Learning Theory*. PMLR, 2018.

[38] Attouch, Hedy, and Juan Peypouquet. "The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than 1/k^2." *SIAM Journal on Optimization* 26.3 (2016): 1824-1834.

[39] T. Sarlos, Improved approximation algorithms for large matrices via random projections, Proceedings of the 47th annual *IEEE Symposium on Foundations of Computer Science (FOCS),* 2006, pp. 143–152.

[40] Allen-Zhu, Zeyuan, David Simchi-Levi, and Xinshang Wang. "The lingering of gradients: how to reuse gradients over time." *Advances in Neural Information Processing Systems* 31 (2018).

[41] Hoffer, Elad, Itay Hubara, and Daniel Soudry. "Train longer, generalize better: closing the generalization gap in large batch training of neural networks." *Advances in neural information processing systems* 30 (2017).

[42] Zaheer, Manzil, et al. "Adaptive methods for nonconvex optimization." *Advances in neural information processing systems* 31 (2018).

[43] Sleijpen, Gerard LG, and Henk A. Van der Vorst. "A Jacobi--Davidson iteration method for linear eigenvalue problems." *SIAM review* 42.2 (2000): 267-293.

[44] Grimes, Roger G., John G. Lewis, and Horst D. Simon. "A shifted block Lanczos algorithm for solving sparse symmetric generalized eigenproblems." *SIAM Journal on Matrix Analysis and Applications* 15.1 (1994): 228-272.

[45] Ghorbani, Behrooz, Shankar Krishnan, and Ying Xiao. "An investigation into neural net optimization via hessian eigenvalue density." *International Conference on Machine Learning*. PMLR, 2019.

[46] Nash, Stephen G. "Newton-type minimization via the Lanczos method." *SIAM Journal on Numerical Analysis* 21.4 (1984): 770-788.

[47] Y. Chen, Y. Chi, J. Fan, C. Ma, “Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval”, *Mathematical Programming*, vol. 176, no. 1-2, pp. 5-37, July 2019.

[48] Cutkosky, Ashok, and Francesco Orabona. "Momentum-based variance reduction in non-convex sgd." *Advances in neural information processing systems* 32 (2019).

[49] Goldfarb, Donald, Yi Ren, and Achraf Bahamou. "Practical quasi-newton methods for training deep neural networks." *Advances in Neural Information Processing Systems* 33 (2020): 2386-2396.

[50] Pilanci, Mert, and Martin J. Wainwright. "Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence." *SIAM Journal on Optimization* 27.1 (2017): 205-245.

[51] Bollapragada, Raghu, et al. "A progressive batching L-BFGS method for machine learning." *International Conference on Machine Learning*. PMLR, 2018.

[52] Agarwal, Naman, Brian Bullins, and Elad Hazan. "Second-order stochastic optimization for machine learning in linear time." *The Journal of Machine Learning Research* 18.1 (2017): 4148-4187.

[53] Gower, Robert, Donald Goldfarb, and Peter Richtárik. "Stochastic block BFGS: Squeezing more curvature out of data." *International Conference on Machine Learning*. PMLR, 2016.
[54] Erdogdu, Murat A., and Andrea Montanari. "Convergence rates of sub-sampled Newton methods." *Advances in Neural Information Processing Systems* 28 (2015).
[55] Lacotte, Jonathan, Yifei Wang, and Mert Pilanci. "Adaptive newton sketch: Linear-time optimization with quadratic convergence and effective hessian dimensionality." *International Conference on Machine Learning*. PMLR, 2021.
[56] Berahas, Albert S., Jorge Nocedal, and Martin Takác. "A multi-batch L-BFGS method for machine learning." *Advances in Neural Information Processing Systems* 29 (2016).
[57] Castera, Camille, et al. "An Inertial Newton Algorithm for Deep Learning." *J. Mach. Learn. Res.* 22 (2021): 134-1.
[58] Schraudolph, Nicol N., Jin Yu, and Simon Günter. "A stochastic quasi-Newton method for online convex optimization." *Artificial intelligence and statistics*. PMLR, 2007.
[59] Bajovic, Dragana, et al. "Newton-like method with diagonal correction for distributed optimization." *SIAM Journal on Optimization* 27.2 (2017): 1171-1203.
[60] Wang, Yu, Wotao Yin, and Jinshan Zeng. "Global convergence of ADMM in nonconvex nonsmooth optimization." *Journal of Scientific Computing* 78.1 (2019): 29-63.