Bibliography

Allen-Zhu, Z. 2017. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, **18**(1), 8194–8244.

Attouch, H., Chbani, Z., Peypouquet, J., and Redont, P. 2018. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, **168**(1–2), 123–175.

Beck, A., and Teboulle, M. 2003. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, **31**, 167–175.

Beck, A., and Teboulle, M. 2009. A Fast iterative shrinkage-threshold algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, **2**(1), 183–202.

Beck, A., and Tetruashvili, L. 2013. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, **23**(4), 2037–2060.

Bertsekas, D. P. 1976. On the Goldstein-Levitin-Polyak gradient projection method. *IEEE Transactions on Automatic Control*, **AC-21**, 174–184.

Bertsekas, D. P. 1982. *Constrained Optimization and Lagrange Multiplier Methods*. New York: Academic Press.

Bertsekas, D. P. 1997. A new class of incremental gradient methods for least squares problems. *SIAM Journal on Optimization*, **7**(4), 913–926.

Bertsekas, D. P. 1999. *Nonlinear Programming*. Second edition. Belmont, MA: Athena Scientific.

Bertsekas, D. P. 2011. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. Pages 85–119 of: Sra, S., Nowozin, S., and Wright, S. J. (eds), *Optimization for Machine Learning*. NIPS Workshop Series. Cambridge, MA: MIT Press.

Bertsekas, D. P., and Tsitsiklis, J. N. 1989. *Parallel and Distributed Computation: Numerical Methods*. Englewood Cliffs, NJ: Prentice Hall.

Bertsekas, D. P., Nedić, A., and Ozdaglar, A. E. 2003. *Convex Analysis and Optimization*. Optimization and Computation Series. Belmont, MA: Athena Scientific.

Blatt, D., Hero, A. O., and Gauchman, H. 2007. A convergent incremental gradient method with a constant step size. *SIAM Journal on Optimization*, **18**(1), 29–51.

Bolte, J., and Pauwels, E. 2021. Conservative set valued fields, automatic differentiation, stochastic gradient methods, and deep learning. *Mathematical Programming*, **188**(1), 19–51.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. 1992. A training algorithm for optimal margin classifiers. Pages 144–152 of: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. Pittsburgh, PA: ACM Press.

Boyd, S., and Vandenberghe, L. 2003. *Convex Optimization*. Cambridge: Cambridge University Press.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. 2011. Distributed optimization and statistical learning via the alternating direction methods of multipliers. *Foundations and Trends in Machine Learning*, **3**(1), 1–122.

Bubeck, S., Lee, Y. T., and Singh, M. 2015. A geometric alternative to Nesterov's accelerated gradient descent. Technical Report arXiv:1506.08187. Microsoft Research.

Burachik, R. S., and Jeyakumar, V. 2005. A Simple closure condition for the normal cone intersection formula. *Transactions of the American Mathematical Society*, **133**(6), 1741–1748.

Burer, S., and Monteiro, R. D. C. 2003. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorizations. *Mathematical Programming, Series B*, **95**, 329–257.

Burke, J. V., and Engle, A. 2018. Line search methods for convex-composite optimization. Technical Report arXiv:1806.05218. Department of Mathematics, University of Washington.

Candès, E., and Recht, B. 2009. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, **9**, 717–772.

Chouzenoux, E., Pesquet, J.-C., and Repetti, A. 2016. A block coordinate variable metric forward-backward algorithm. *Journal of Global Optimization*, **66**, 457–485.

Conn, A. R., Gould, N. I. M., and Toint, P. L. 1992. *LANCELOT: A Fortran Package for Large-Scale Nonlinear Optimization*. Springer Series in Computational Mathematics, vol. 17. Heidelberg: Springer-Verlag.

Cortes, C., and Vapnik, V. N. 1995. Support-vector networks. *Machine Learning*, **20**, 273–297.

Danskin, J. M. 1967. *The Theory of Max-Min and Its Application to Weapons Allocation Problems*. Springer.

Davis, D., Drusvyatskiy, D., Kakade, S., and Lee, J. D. 2020. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, **20**(1), 119–154.

Defazio, A., Bach, F., and Lacoste-Julien, S. 2014. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. Pages 1646–1654 of: *Advances in Neural Information Processing Systems, November 2014, Montreal, Canada*.

Dem'yanov, V. F., and Rubinov, A. M. 1967. The minimization of a smooth convex functional on a convex set. *SIAM Journal on Control*, **5**(2), 280–294.

Dem'yanov, V. F., and Rubinov, A. M. 1970. *Approximate Methods in Optimization Problems*. Vol. 32. New York: Elsevier.

Drusvyatskiy, D., Fazel, M., and Roy, S. 2018. An optimal first order method based on optimal quadratic averaging. *SIAM Journal on Optimization*, **28**(1), 251–271.

Dunn, J. C. 1980. Convergence rates for conditional gradient sequences generated by implicit step length rules. *SIAM Journal on Control and Optimization*, **18**(5), 473–487.

Dunn, J. C. 1981. Global and asymptotic convergence rate estimates for a class of projected gradient processes. *SIAM Journal on Control and Optimization*, **19**(3), 368–400.

Eckstein, J., and Bertsekas, D. P. 1992. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, **55**, 293–318.

Eckstein, J., and Yao, W. 2015. Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives. *Pacific Journal of Optimization*, **11**(4), 619–644.

Fercoq, O., and Richtarik, P. 2015. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, **25**, 1997–2023.

Fletcher, R., and Reeves, C. M. 1964. Function minimization by conjugate gradients. *Computer Journal*, **7**, 149–154.

Frank, M., and Wolfe, P. 1956. An algorithm for Quadratic Programming. *Naval Research Logistics Quarterly*, **3**, 95–110.

Gabay, D., and Mercier, B. 1976. A dual algorithm for the solution of nonlinear variational problems via finite element approximations. *Computers and Mathematics with Applications*, **2**, 17–40.

Gelfand, I. 1941. Normierte ringe. *Recueil Mathématique [Matematicheskii Sbornik]*, **9**, 3–24.

Glowinski, R., and Marrocco, A. 1975. Sur l'approximation, par elements finis d'ordre un, en al resolution, par penalisation-dualité, d'une classe dre problems de Dirichlet non lineares. *Revue Francaise d'Automatique, Informatique, et Recherche Operationelle*, **9**, 41–76.

Goldstein, A. A. 1964. Convex programming in Hilbert space. *Bulletin of the American Mathematical Society*, **70**, 709–710.

Goldstein, A. A. 1974. On gradient projection. Pages 38–40 of: *Proceedings of the 12th Allerton Conference on Circuit and System Theory, Allerton Park, Illinois*.

Golub, G. H., and van Loan, C. F. 1996. *Matrix Computations*. Third edition. Baltimore: The Johns Hopkins University Press.

Griewank, A., and Walther, A. 2008. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Second edition. Frontiers in Applied Mathematics. Philadelphia, PA: SIAM.

Hestenes, M. R. 1969. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, **4**, 303–320.

Hestenes, M., and Steifel, E. 1952. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, **49**(6), 409–436.

Hu, B., Wright, S. J., and Lessard, L. 2018. Dissipativity theory for accelerating stochastic variance reduction: A unified analysis of SVRG and Katyusha using semidefinite programs. Pages 2038–2047 of: *International Conference on Machine Learning (ICML)*.

Jaggi, M. 2013. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. Pages 427–435 of: *International Conference on Machine Learning (ICML)*.

Jain, P., Netrapalli, P., Kakade, S. M., Kidambi, R., and Sidford, A. 2018. Accelerating stochastic gradient descent for least squares regression. Pages 545–604 of: *Conference on Learning Theory (COLT)*.

Johnson, R., and Zhang, T. 2013. Accelerating stochastic gradient descent using predictive variance reduction. Pages 315–323 of: *Advances in Neural Information Processing Systems*.

Kaczmarz, S. 1937. Angenäherte Auflösung von Systemen linearer Gleichungen. *Bulletin International de l'Académie Polonaise des Sciences et des Lettres. Classe des Sciences Mathématiques et Naturelles. Série A, Sciences Mathématiques*, **35**, 355–357.

Karimi, H., Nutini, J., and Schmidt, M. 2016. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. Pages 795–811 of: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer.

Kiwiel, K. C. 1990. Proximity control in bundle methods for convex nondifferentiable minimization. *Mathematical Programming*, **46**(1–3), 105–122.

Kurdyka, K. 1998. On gradients of functions definable in o-minimal structures. *Annales de l'Institut Fourier*, **48**, 769–783.

Lang, S. 1983. *Real Analysis*. Second edition. Reading, MA: Addison-Wesley.

Le Roux, N., Schmidt, M., and Bach, F. 2012. A stochastic gradient method with an exponential convergence _rate for finite training sets. *Advances in Neural Information Processing Systems*, **25**, 2663–2671.

Lee, C.-P., and Wright, S. J. 2018. Random permutations fix a worst case for cyclic coordinate descent. *IMA Journal of Numerical Analysis*, **39**, 1246–1275.

Lee, Y. T., and Sidford, A. 2013. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. Pages 147–156 of: *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE.

Lemaréchal, C. 1975. An extension of Davidon methods to non differentiable problems. Pages 95–109 of: *Nondifferentiable Optimization*. Springer.

Lemaréchal, C., Nemirovskii, A., and Nesterov, Y. 1995. New variants of bundle methods. *Mathematical Programming*, **69**(1–3), 111–147.

Lessard, L., Recht, B., and Packard, A. 2016. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, **26**(1), 57–95.

Levitin, E. S., and Polyak, B. T. 1966. Constrained minimization problems. *USSR Journal of Computational Mathematics and Mathematical Physics*, **6**, 1–50.

Li, X., Zhao, T., Arora, R., Liu, H., and Hong, M. 2018. On Faster convergence of cyclic block coordinate descent-type methods for strongly convex minimization. *Journal of Machine Learning Research*, **18**, 1–24.

Liu, J., and Wright, S. J. 2015. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM Journal on Optimization*, **25**(1), 351–376.

Liu, J., Wright, S. J., Ré, C., Bittorf, V., and Sridhar, S. 2015. An asynchronous parallel stochastic coordinate descent algorithm. *Journal of Machine Learning Research*, **16**, 285–322.

Łojasiewicz, S. 1963. Une propriété topologique des sous-ensembles analytiques réels. *Les Équations aus Dérivées Partielles*, **117**, 87–89.

Lu, Z., and Xiao, L. 2015. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming, Series A*, **152**, 615–642.

Luo, Z.-Q., Sturm, J. F., and Zhang, S. 2000. Conic convex programming and self-dual embedding. *Optimization Methods and Software*, **14**, 169–218.

Maddison, C. J., Paulin, D., Teh, Y. W., O'Donoghue, B., and Doucet, A. 2018. Hamiltonian descent methods. arXiv preprint arXiv:1809.05042.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. 2009. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, **19**(4), 1574–1609.

Nesterov, Y. 1983. A method for unconstrained convex problem with the rate of convergence $O(1/k^2)$. *Doklady AN SSSR*, **269**, 543–547.

Nesterov, Y. 2004. *Introductory Lectures on Convex Optimization: A Basic Course*. Boston: Kluwer Academic Publishers.

Nesterov, Y. 2012. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, **22**(January), 341–362.

Nesterov, Y. 2015. Universal gradient methods for convex optimization problems. *Mathematical Programming*, **152**(1–2), 381–404.

Nesterov, Y., and Nemirovskii, A. S. 1994. *Interior Point Polynomial Methods in Convex Programming*. Philadelphia, PA: SIAM.

Nesterov, Y., and Stich, S. U. 2017. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, **27**(1), 110–123.

Nocedal, J., and Wright, S. J. 2006. *Numerical Optimization*. Second edition. New York: Springer.

Parikh, N., and Boyd, S. 2013. Proximal algorithms. *Foundations and Trends in Optimization*, **1**(3), 123–231.

Polyak, B. T. 1963. Gradient methods for minimizing functionals (in Russian). *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 643–653.

Polyak, B. T. 1964. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, **4**, 1–17.

Powell, M. J. D. 1969. A method for nonlinear constraints in minimization problems. Pages 283–298 of: Fletcher, R. (ed), *Optimization*. New York: Academic Press.

Rao, C. V., Wright, S. J., and Rawlings, J. B. 1998. Application of interior-point methods to model predictive control. *Journal of Optimization Theory and Applications*, **99**, 723–757.

Recht, B., Fazel, M., and Parrilo, P. 2010. Guaranteed Minimum-rank solutions to linear matrix equations via nuclear norm minimization. *SIAM Review*, **52**(3), 471–501.

Richtarik, P., and Takac, M. 2014. Iteration complexity of a randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming, Series A*, **144**(1), 1–38.

Richtarik, P., and Takac, M. 2016a. Distributed coordinate descent method for learning with big data. *Journal of Machine Learning Research*, **17**, 1–25.

Richtarik, P., and Takac, M. 2016b. Parallel coordinate descent methods for big data optimization. *Mathematical Programming, Series A*, **156**, 433–484.

Robbins, H., and Monro, S. 1951. A stochastic approximation method. *Annals of Mathematical Statistics*, **22**(3), 400–407.

Rockafellar, R. T. 1970. *Convex Analysis*. Princeton, NJ: Princeton University Press.

Rockafellar, R. T. 1973. The multiplier method of Hestenes and Powell applied to convex programming. *Journal of Optimization Theory and Applications*, **12**(6), 555–562.

Rockafellar, R. T. 1976a. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, **1**, 97–116.

Rockafellar, R. T. 1976b. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, **14**, 877–898.

Rosenblatt, F. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**(6), 386.

Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. 2011. Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical Programming*, **127**(1), 3–30.

Shi, B., Du, S. S., Jordan, M. I., and Su, W. J. 2018. Understanding the acceleration phenomenon via high-resolution differential equations. arXiv preprint arXiv:1810.08907.

Sion, M. 1958. On general minimax theorems. *Pacific Journal of Mathematics*, **8**(1), 171–176.

Stellato, B., Banjac, G., Goulart, P., Bemporad, A., and Boyd, S. 2020. OSQP: An operator splitting solver for quadratic programs. *Mathematical Programming Computation*, **12**(4), 637–672.

Strohmer, T., and Vershynin, R. 2009. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, **15**(2), 262.

Su, W., Boyd, S., and Candès, E. 2014. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. Pages 2510–2518 of: *Advances in Neural Information Processing Systems*.

Sun, R., and Hong, M. 2015. Improved iteration complexity bounds of cyclic block coordinate descent for convex problems. Pages 1306–1314 of: *Advances in Neural Information Processing Systems*.

Teo, C. H., Vishwanathan, S. V. N., Smola, A., and Le, Q. V. 2010. Bundle methods for regularized risk minimization. *Journal of Machine Learning Research*, **11**(1), 311–365.

Tibshirani, R. 1996. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society B*, **58**, 267–288.

Todd, M. J. 2001. Semidefinite optimization. *Acta Numerica*, **10**, 515–560.

Tseng, P., and Yun, S. 2010. A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training. *Computational Optimization and Applications*, **47**(2), 179–206.

Vandenberghe, L. 2016. *Slides for EE236C: Optimization Methods for Large-Scale Systems*.

Vandenberghe, L., and Boyd, S. 1996. Semidefinite programming. *SIAM Review*, **38**, 49–95.

Vapnik, V. 1992. Principles of risk minimization for learning theory. Pages 831–838 of: *Advances in Neural Information Processing Systems*.

Vapnik, V. 2013. *The Nature of Statistical Learning Theory*. Berlin: Springer Science & Business Media.

Wibisono, A., Wilson, A. C., and Jordan, M. I. 2016. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, **113**(47), E7351–E7358.

Wolfe, P. 1975. A method of conjugate subgradients for minimizing nondifferentiable functions. Pages 145–173 of: *Nondifferentiable Optimization*. Springer.

Wright, S. J. 1997. *Primal-Dual Interior-Point Methods*. Philadelphia, PA: SIAM.

Wright, S. J. 2012. Accelerated block-coordinate relaxation for regularized optimization. *SIAM Journal on Optimization*, **22**(1), 159–186.

Wright, S. J. 2018. Optimization algorithms for data analysis. Pages 49–97 of: Mahoney, M., Duchi, J. C., and Gilbert, A. (eds), *The Mathematics of Data*. IAS/Park City Mathematics Series, vol. 25. AMS.

Wright, S. J., and Lee, C.-P. 2020. Analyzing random permutations for cyclic coordinate descent. *Mathematics of Computation*, **89**, 2217–2248.

Wright, S. J., Nowak, R. D., and Figueiredo, M. A. T. 2009. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, **57**(August), 2479–2493.

Zhang, T. 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. Page 116 of: *Proceedings of the Twenty-First International Conference on Machine Learning*.