

References

- Alacaoglu A, Cevher V, Wright SJ (2025) On the complexity of a simple primal-dual coordinate method. *Mathematical Programming*
- Amari S (1967) A theory of adaptive pattern classifier. *IEEE Transactions on Electronic Computers* EC-16(3):279–307
- An X, Zhu X, Gao Y, Xiao Y, Zhao Y, Feng Z, Wu L, Qin B, Zhang M, Zhang D, Fu Y (2021) Partial fc: Training 10 million identities on a single machine. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp 1445–1449
- Arjevani Y, Carmon Y, Duchi JC, Foster DJ, Srebro N, Woodworth B (2022) Lower bounds for non-convex stochastic optimization. *Math Program* 199(1–2):165–214, DOI 10.1007/s10107-022-01822-7, URL <https://doi.org/10.1007/s10107-022-01822-7>
- Ben-Tal A, Teboulle M (1986a) Expected utility, penalty functions, and duality in stochastic nonlinear programming. *Management Science* 32(11):1445–1466
- Ben-Tal A, Teboulle M (1986b) Expected utility, penalty functions, and duality in stochastic nonlinear programming. *Management Science* 32(11):1445–1466, DOI 10.1287/mnsc.32.11.1445, URL <https://doi.org/10.1287/mnsc.32.11.1445>
- Ben-Tal A, Teboulle M (2007) An old-new concept of convex risk measures: the optimized certainty equivalent. *Mathematical Finance* 17(3):449–476
- Ben-Tal A, El Ghaoui L, Nemirovski A (2009a) *Robust Optimization*. Princeton Series in Applied Mathematics, Princeton University Press
- Ben-Tal A, Ghaoui LE, Nemirovski A (2009b) *Robust Optimization*. Princeton Series in Applied Mathematics, Princeton University Press, Princeton, NJ
- Ben-Tal A, den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59(2):341–357, DOI 10.1287/mnsc.1120.1641, URL <https://doi.org/10.1287/mnsc.1120.1641>
- Bertsekas D (2005) Control of uncertain systems with a set-membership description of the uncertainty
- Bertsekas DP (2009) *Convex Optimization Theory*. Athena Scientific
- Bishop CM (2006) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg
- Bommasani R, Hudson DA, Adeli E, Altman RB, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunsell E, Brynjolfsson E, Buch S, Card D, Castellon R, Chatterji NS, Chen AS, Creel K, Davis JQ, Demszky D, Donahue C, Doumbouya M, Durmus E, Ermon S, Etchemendy J, Ethayarajh K, Fei-Fei L, Finn C, Gale T, Gillespie LE, Goel K, Goodman ND, Grossman S, Guha N, Hashimoto T, Henderson P, Hewitt J, Ho DE, Hong J, Hsu K, Huang J, Icard T, Jain S, Jurafsky D, Kalluri P, Karamcheti S, Keeling G, Khani F, Khattab O, Koh PW, Krass MS, Krishna R, Kudithipudi R, et al (2021) On the opportunities and risks of foundation models. *CoRR* abs/2108.07258, URL <https://arxiv.org/abs/2108.07258>, 2108.07258

-
- Boyd K, Eng KH, Page CD (2013) Area under the precision-recall curve: Point estimates and confidence intervals. In: Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2013, Springer, pp 451–466, DOI 10.1007/978-3-642-40994-3_29
- Boyd S, Vandenberghe L (2004) Convex Optimization. Cambridge University Press
- Bracken J, McGill JT (1973) Mathematical programs with optimization problems in the constraints. *Operations Research* 21:37–44
- Brown DB (2007) Large deviations bounds for estimating conditional value-at-risk. *Oper Res Lett* 35(6):722–730, DOI 10.1016/j.orl.2007.01.001, URL <https://doi.org/10.1016/j.orl.2007.01.001>
- Calafiore GC (2007) Ambiguous risk measures and optimal robust portfolios. *SIAM Journal on Optimization* 18(3):853–877, DOI 10.1137/050639379, URL <https://doi.org/10.1137/050639379>
- Cao K, Wei C, Gaidon A, Arechiga N, Ma T (2019) Learning imbalanced datasets with label-distribution-aware margin loss. In: Advances in Neural Information Processing Systems (NeurIPS), vol 32, pp 1567–1578
- Cao Z, Qin T, Liu TY, Tsai MF, Li H (2007) Learning to rank: From pairwise approach to listwise approach. In: Proceedings of the 24th International Conference on Machine Learning, pp 129–136
- Cauchy AL (1847) Méthode générale pour la résolution des systèmes d’équations simultanées. *Comptes Rendus de l’Académie des Sciences* 25:536, reprinted in *Œuvres complètes*, Série 1, Tome 10, pp. 399–402. Gallica digital document.
- Chang KW, Hsieh CJ, Lin CJ (2008) Coordinate descent method for large-scale l2-loss linear support vector machines. *Journal of Machine Learning Research* 9(45):1369–1398, URL <http://jmlr.org/papers/v9/chang08a.html>
- Chen L, Ma Y, Zhang J (2025a) Near-optimal nonconvex-strongly-convex bilevel optimization with fully first-order oracles. *Journal of Machine Learning Research* 26(109):1–56, URL <http://jmlr.org/papers/v26/23-1104.html>
- Chen T, Sun Y, Yin W (2021a) Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In: Beygelzimer A, Dauphin Y, Liang P, Vaughan JW (eds) Advances in Neural Information Processing Systems, URL <https://openreview.net/forum?id=r6cNUjs8cm0>
- Chen T, Sun Y, Yin W (2021b) Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing* 69:4937–4948, DOI 10.1109/TSP.2021.3092377
- Chen X, Wang B, Yang M, Lin Q, Yang T (2025b) Stochastic momentum methods for non-smooth non-convex finite-sum coupled compositional optimization. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems, URL <https://openreview.net/forum?id=kSgZiAAwDU>
- Chiang CK, Yang T, Lee CJ, Mahdavi M, Lu CJ, Jin R, Zhu S (2012) Online optimization with gradual variations. In: Mannor S, Srebro N, Williamson RC (eds) Proceedings of the 25th Annual Conference on Learning Theory, PMLR, Edinburgh, Scotland, Proceedings of Machine Learning Research, vol 23, pp 6.1–6.20, URL <https://proceedings.mlr.press/v23/chiang12.html>

- Chung KL (1954) On a Stochastic Approximation Method. *The Annals of Mathematical Statistics* 25(3):463–483, DOI 10.1214/aoms/1177728716, URL <https://doi.org/10.1214/aoms/1177728716>
- Cortes C, Mohri M (2003) AUC optimization vs. error rate minimization. *Advances in Neural Information Processing Systems* 16, URL https://proceedings.neurips.cc/paper_files/paper/2003/file/2518-auc-optimization-vs-error-rate-minimization.pdf
- Crammer K, Singer Y (2002) On the algorithmic implementation of multiclass kernel-based vector machines. *J Mach Learn Res* 2:265–292
- Csiszár I (1967) Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica* 2:299–318
- Cutkosky A, Orabona F (2019) Momentum-based variance reduction in non-convex SGD, Curran Associates Inc., Red Hook, NY, USA
- Dang CD, Lan G (2015) Stochastic block mirror descent methods for nonsmooth and stochastic optimization. *SIAM Journal on Optimization* 25(2):856–882
- Danskin J (1967) *The Theory of Max-min and Its Applications to Weapons Allocation Problems*. *Econometrics and operations research*, Springer, URL <https://books.google.ca/books?id=bvrfAQAACAAJ>
- Daskalakis C, Ilyas A, Syrgkanis V, Zeng H (2018) Training GANs with optimism. In: *International Conference on Learning Representations*, URL <https://openreview.net/forum?id=SJJySbbAZ>
- Daubechies I, Defrise M, Mol CD (2004) An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics* 57(11):1413–1457, DOI 10.1002/cpa.20042
- Davis D, Drusvyatskiy D (2019) Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization* 29(1):207–239, DOI 10.1137/18M1178244
- Dekel O, Singer Y (2005) Data-driven online to batch conversions. In: Weiss Y, Schölkopf B, Platt J (eds) *Advances in Neural Information Processing Systems*, MIT Press, vol 18, URL https://proceedings.neurips.cc/paper_files/paper/2005/file/4a5876b450b45371f6cfe5047ac8cd45-Paper.pdf
- Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research* 58(3):595–612
- Deleu T, Bengio Y (2021) Structured sparsity inducing adaptive optimizers for deep learning. *ArXiv abs/2102.03869*, URL <https://api.semanticscholar.org/CorpusID:231846689>
- Dodd LE, Pepe MS (2003) Partial AUC estimation and regression. *Biometrics* 59(3):614–623, DOI 10.1111/1541-0420.00071, URL <https://doi.org/10.1111/1541-0420.00071>
- Drusvyatskiy D, Paquette C (2019) Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming* 178:503–558

-
- Drusvyatskiy D, Ioffe AD, Lewis AS (2021) Nonsmooth optimization using taylor-like models: error bounds, convergence, and termination criteria. *Mathematical Programming* 185:357–383
- Duchi J, Singer Y (2009) Efficient online and batch learning using forward backward splitting. *J Mach Learn Res* 10:2899–2934
- Duchi J, Hazan E, Singer Y (2011) Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12:2121–2159
- Duchi JC, Ruan F (2017) Stochastic methods for composite and weakly convex optimization problems. *arXiv preprint arXiv:1703.08570* URL <https://arxiv.org/abs/1703.08570>, [1703.08570](https://arxiv.org/abs/1703.08570)
- Duchi JC, Ruan F (2018) Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization* 28(4):3229–3259, DOI 10.1137/17M1135086
- Duchi JC, Glynn PW, Namkoong H (2022) Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research* 47(2):882–910, DOI 10.1287/moor.2020.1085, URL <https://doi.org/10.1287/moor.2020.1085>
- Dupačová J (1966) On minimax solutions of stochastic linear programming problems. *Časopis pro pěstování matematiky* 091(4):423–430, URL <http://eudml.org/doc/20949>
- Ermoliev Y, Wets RJB (eds) (1988) *Numerical Techniques for Stochastic Optimization*, Springer Series in Computational Mathematics, vol 10. Springer-Verlag
- Ermoliev YM (1976) *Methods of Stochastic Programming*. Monographs in Optimization and Operations Research, Nauka, Moscow
- Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456):1348–1360, DOI 10.1198/016214501753382273
- Fang C, Li CJ, Lin Z, Zhang T (2018) Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol 31, URL https://proceedings.neurips.cc/paper_files/paper/2018/file/1543843a4723ed2ab08e18053ae6dc5b-Paper.pdf
- Fazel M, Hindi H, Boyd SP (2001) A rank minimization heuristic with application to minimum order system approximation. In: *2001 IEEE International Conference on Control Applications*, IEEE, pp 1347–1352
- Fletcher R (1982) A model algorithm for composite nondifferentiable optimization problems. *Mathematical Programming Study* 17:67–76
- Fletcher R, Watson GA (1980) First and second order conditions for a class of non-differentiable optimization problems. *Mathematical Programming* 18:291–307
- Frankel SP (1950) Convergence rates of iterative treatments of partial differential equations. *Mathematics of Computation* 4:65–75, URL <https://api.semanticscholar.org/CorpusID:119690385>

- Freund Y, Schapire RE (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J Comput Syst Sci* 55(1):119–139, DOI 10.1006/jcss.1997.1504
- Freund Y, Iyer R, Schapire RE, Singer Y (2003) An efficient boosting algorithm for combining preferences. *J Mach Learn Res* 4(null):933–969
- Ghadimi S, Lan G (2012) Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization* 22(4):1469–1492, DOI 10.1137/110848864
- Ghadimi S, Lan G (2013) Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization* 23(4):2341–2368, DOI 10.1137/120880811
- Ghadimi S, Ruszczyński A, Wang M (2020) A single timescale stochastic approximation method for nested stochastic optimization. *SIAM Journal on Optimization* 30(1):960–979, DOI 10.1137/18M1230542
- Ghosh A, Kumar H, Sastry PS (2017) Robust loss functions under label noise for deep neural networks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol 31
- Goldstein AA (1964) Convex programming in hilbert space. *Bulletin of the American Mathematical Society* 70(5):709–710
- Gong P, Zhang C, Lu Z, Huang J, Ye J (2013) A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In: Dasgupta S, McAllester D (eds) *Proceedings of the 30th International Conference on Machine Learning*, PMLR, Atlanta, Georgia, USA, *Proceedings of Machine Learning Research*, vol 28, pp 37–45, URL <https://proceedings.mlr.press/v28/gong13a.html>
- Green DM, Swets JA (1966) *Signal Detection Theory and Psychophysics*. John Wiley and Sons Inc., New York, NY
- Guo S, Hong I, Balmaseda V, Yu C, Qiu L, Liu X, Jiang H, Zhao T, Yang T (2025) Discriminative finetuning of generative large language models without reward models and human preference data. In: *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, OpenReview.net, URL <https://openreview.net/forum?id=1jutKQ5R8T>
- Guo Z, Hu Q, Zhang L, Yang T (2021a) Randomized stochastic variance-reduced methods for stochastic bilevel optimization. *CoRR abs/2105.02266*, URL <https://arxiv.org/abs/2105.02266>, 2105.02266
- Guo Z, Xu Y, Yin W, Jin R, Yang T (2021b) Unified convergence analysis for adaptive optimization with moving average estimator. *Mach Learn* 114(4), DOI 10.1007/s10994-024-06650-8, URL <https://doi.org/10.1007/s10994-024-06650-8>
- Guo Z, Yan Y, Yuan Z, Yang T (2023) Fast objective & duality gap convergence for non-convex strongly-concave min-max problems with PL condition. *J Mach Learn Res* 24:148:1–148:63, URL <https://jmlr.org/papers/v24/21-1471.html>
- Hadsell R, Chopra S, LeCun Y (2006) Dimensionality reduction by learning an invariant mapping. In: *Proceedings of the 2006 IEEE Computer Society Conference*

-
- on Computer Vision and Pattern Recognition (CVPR'06), IEEE, vol 2, pp 1735–1742, DOI 10.1109/CVPR.2006.100
- Hamilton WL, Ying R, Leskovec J (2017) Inductive representation learning on large graphs. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, NIPS'17, p 1025–1035
- Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 143(1):29–36
- Hardt M, Recht B, Singer Y (2016) Train faster, generalize better: Stability of stochastic gradient descent. In: Proceedings of the 33rd International Conference on Machine Learning (ICML), PMLR, pp 1225–1234
- Hazan E, Kale S (2011) Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In: Kakade SM, von Luxburg U (eds) Proceedings of the 24th Annual Conference on Learning Theory, PMLR, Budapest, Hungary, Proceedings of Machine Learning Research, vol 19, pp 421–436, URL <https://proceedings.mlr.press/v19/hazan11a.html>
- Hazan E, Agarwal A, Kale S (2007) Logarithmic regret algorithms for online convex optimization. *Mach Learn* 69(2–3):169–192, DOI 10.1007/s10994-007-5016-8, URL <https://doi.org/10.1007/s10994-007-5016-8>
- Hinton G (2018) Neural networks for machine learning, lecture 6. Coursera online course
- Hong M, Wai HT, Wang Z, Yang Z (2020) A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization* 33(1):147–180
- Hsieh CJ, Chang KW, Lin CJ, Keerthi SS, Sundararajan S (2008) A dual coordinate descent method for large-scale linear svm. In: Proceedings of the 25th International Conference on Machine Learning, Association for Computing Machinery, New York, NY, USA, ICML '08, p 408–415, DOI 10.1145/1390156.1390208, URL <https://doi.org/10.1145/1390156.1390208>
- Hu Q, Qiu Z, Guo Z, Zhang L, Yang T (2023) Blockwise stochastic variance-reduced methods with parallel speedup for multi-block bilevel optimization. *CoRR* abs/2305.18730, DOI 10.48550/ARXIV.2305.18730, URL <https://doi.org/10.48550/arXiv.2305.18730>, 2305.18730
- Hu Q, Qi Q, Lu Z, Yang T (2024a) Single-loop stochastic algorithms for difference of max-structured weakly convex functions. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems, URL <https://openreview.net/forum?id=NhtBXSXKA>
- Hu Q, Qi Q, Lu Z, Yang T (2024b) Single-loop stochastic algorithms for difference of max-structured weakly convex functions. In: Globersons A, Mackey L, Belgrave D, Fan A, Paquet U, Tomczak JM, Zhang C (eds) Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, URL http://papers.nips.cc/paper_files/paper/2024/hash/67e79c8e9b11f068a7cafd79505175c0-Abstract-Conference.html

- Hu W, Niu G, Sato I, Sugiyama M (2018) Does distributionally robust supervised learning give robust classifiers? In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, PMLR, Proceedings of Machine Learning Research, vol 80, pp 2029–2037, URL <https://proceedings.mlr.press/v80/hu18a.html>
- Hu W, Li CJ, Lian X, Liu J, Yuan H (2019) Efficient smooth non-convex stochastic compositional optimization via stochastic recursive gradient descent. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, URL https://proceedings.neurips.cc/paper_files/paper/2019/file/21ce689121e39821d07d04faab328370-Paper.pdf
- Hu Y, Zhang S, Chen X, He N (2020) Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, NIPS '20
- Huang F, Gao S, Pei J, Huang H (2022) Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. J Mach Learn Res 23(1)
- Huo Z, Gu B, Liu J, Huang H (2018) Accelerated method for stochastic composition optimization with nonsmooth regularization. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI Press, AAAI'18/IAAI'18/EAAI'18
- Ilharco G, Wortsman M, Wightman R, Gordon C, Carlini N, Taori R, Dave A, Shankar V, Namkoong H, Miller J, Hajishirzi H, Farhadi A, Schmidt L (2021) Openclip. DOI 10.5281/zenodo.5143773, URL <https://doi.org/10.5281/zenodo.5143773>, if you use this software, please cite it as below.
- Ilse M, Tomczak J, Welling M (2018) Attention-based deep multiple instance learning. In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, PMLR, Proceedings of Machine Learning Research, vol 80, pp 2127–2136, URL <https://proceedings.mlr.press/v80/ilse18a.html>
- Iouditski A, Nesterov Y (2010) Primal-dual subgradient methods for minimizing uniformly convex functions. arXiv: Optimization and Control URL <https://api.semanticscholar.org/CorpusID:117741989>
- Järvelin K, Kekäläinen J (2000) Ir evaluation methods for retrieving highly relevant documents. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, USA, SIGIR '00, p 41–48, DOI 10.1145/345508.345545, URL <https://doi.org/10.1145/345508.345545>
- Ji K, Yang J, Liang Y (2020) Bilevel optimization: Convergence analysis and enhanced design. In: International Conference on Machine Learning, URL <https://api.semanticscholar.org/CorpusID:235825903>
- Jiang W, Li G, Wang Y, Zhang L, Yang T (2022) Multi-block-single-probe variance reduced estimator for coupled compositional optimization. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A (eds) Advances in Neural Informa-

-
- tion Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, URL http://papers.nips.cc/paper_files/paper/2022/hash/d13ee73683fd5567e5c07634a25cd7b8-Abstract-Conference.html
- Jiang W, Qin J, Wu L, Chen C, Yang T, Zhang L (2023) Learning unnormalized statistical models via compositional optimization. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J (eds) International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, PMLR, Proceedings of Machine Learning Research, vol 202, pp 15105–15124, URL <https://proceedings.mlr.press/v202/jiang23g.html>
- Jin L, Ma J, Liu Z, Gromov A, Defazio A, Xiao L (2025) PARQ: Piecewise-affine regularized quantization. In: Forty-second International Conference on Machine Learning, URL <https://openreview.net/forum?id=8PCx0lwbIn>
- Johnson R, Zhang T (2013) Accelerating stochastic gradient descent using predictive variance reduction. In: Burges C, Bottou L, Welling M, Ghahramani Z, Weinberger K (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 26, URL https://proceedings.neurips.cc/paper_files/paper/2013/file/ac1dd209cbcc5e5d1c6e28598e8cbbe8-Paper.pdf
- Jordan K, Jin Y, Boza V, Jiacheng Y, Cesista F, Newhouse L, Bernstein J (2024) Muon: An optimizer for hidden layers in neural networks. URL <https://kellerjordan.github.io/posts/muon/>
- Juditsky A, Nemirovski A, Tauvel C (2011) Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems* 1(1):17–58
- Kar P, Narasimhan H, Jain P (2014) Online and stochastic gradient methods for non-decomposable loss functions. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, MIT Press, Cambridge, MA, USA, NIPS’14, p 694–702
- Karush W (1939) Minima of functions of several variables with inequalities as side constraints. M.sc. thesis, University of Chicago, Chicago, Illinois
- Khanduri P, Zeng S, Hong M, Wai HT, Wang Z, Yang Z (2021) A near-optimal algorithm for stochastic bilevel optimization via double-momentum. In: Proceedings of the 35th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, NIPS ’21
- Kiefer J, Wolfowitz J (1952) Stochastic Estimation of the Maximum of a Regression Function. *The Annals of Mathematical Statistics* 23(3):462 – 466, DOI 10.1214/aoms/1177729392, URL <https://doi.org/10.1214/aoms/1177729392>
- Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. CoRR abs/1412.6980, URL <https://api.semanticscholar.org/CorpusID:6628106>
- Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: Bengio Y, LeCun Y (eds) ICLR (Poster), URL <http://dblp.uni-trier.de/db/conf/iclr/iclr2015.html#KingmaB14>
- Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations, URL <https://openreview.net/forum?id=SJU4ayYgl>

- Kivinen J, Warmuth MK (1997) Exponentiated gradient versus gradient descent for linear predictors. *Inf Comput* 132(1):1–63, DOI 10.1006/inco.1996.2612, URL <https://doi.org/10.1006/inco.1996.2612>
- Koltchinskii V (2011) Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. *Ecole d’ Eté de Probabilités de Saint-Flour XXXVIII-2008*, Springer
- Korpelevich GM (1976) The extragradient method for finding saddle points and other problems. URL <https://api.semanticscholar.org/CorpusID:118602977>
- Kouvelis P, Yu G (1997) *Robust Discrete Optimization and Its Applications*, 1st edn. Springer, Boston, MA, DOI 10.1007/978-1-4757-2620-6
- Kuhn HW, Tucker AW (2014) *Nonlinear Programming*, Springer Basel, Basel, pp 247–258. DOI 10.1007/978-3-0348-0439-4_11, URL https://doi.org/10.1007/978-3-0348-0439-4_11
- Kwon J, Kwon D, Wright S, Nowak R (2023) A fully first-order method for stochastic bilevel optimization. In: *Proceedings of the 40th International Conference on Machine Learning*, JMLR.org, ICML’23
- Lacoste-Julien S, Schmidt M, Bach FR (2012) A simpler approach to obtaining an $o(1/t)$ convergence rate for the projected stochastic subgradient method. *CoRR abs/1212.2002*, URL <http://arxiv.org/abs/1212.2002>, 1212.2002
- Lan G (2012) An optimal method for stochastic composite optimization. *Math Program* 133(1–2):365–397, DOI 10.1007/s10107-010-0434-y, URL <https://doi.org/10.1007/s10107-010-0434-y>
- Lan G (2020) *First-order and Stochastic Optimization Methods for Machine Learning*, 1st edn. Springer Series in the Data Sciences, Springer International Publishing, Cham
- Lan G, Ouyang Y, Zhang Z (2023) Optimal and parameter-free gradient minimization methods for convex and nonconvex optimization. URL <https://api.semanticscholar.org/CorpusID:265506741>
- Lapin M, Hein M, Schiele B (2018) Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(7):1533–1554, DOI 10.1109/TPAMI.2017.2751607, URL <https://doi.org/10.1109/TPAMI.2017.2751607>
- Lee J, Park S, Shin J (2020) Learning bounds for risk-sensitive learning. *ArXiv abs/2006.08138*, URL <https://api.semanticscholar.org/CorpusID:219686788>
- Lei YX, Ying Y (2019) Fine-grained analysis of stability and generalization for stochastic gradient descent. In: *International Conference on Machine Learning (ICML)*, pp 5809–5819
- Lewis A, Wright S (2009) A proximal method for composite minimization. *Mathematical Programming* 158, DOI 10.1007/s10107-015-0943-9
- Li G, Yu W, Yao Y, Tong W, Liang Y, Lin Q, Yang T (2024) Model developmental safety: A safety-centric method and applications in vision-language models. *CoRR abs/2410.03955*, DOI 10.48550/ARXIV.2410.03955, URL <https://doi.org/10.48550/arXiv.2410.03955>, 2410.03955

-
- Li G, Lin M, Galanti T, Tu Z, Yang T (2025) DisCO: Reinforcing large reasoning models with discriminative constrained optimization. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems, URL <https://openreview.net/forum?id=zzUXS4f91r>
- Lin T, Jin C, Jordan M (2020) On gradient descent ascent for nonconvex-concave minimax problems. In: III HD, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning, PMLR, Proceedings of Machine Learning Research, vol 119, pp 6083–6093, URL <https://proceedings.mlr.press/v119/lin20a.html>
- Lions PL, Mercier B (1979) Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis* 16(6):964–979, DOI 10.1137/0716071
- Littlestone N (1988) Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Mach Learn* 2(4):285–318, DOI 10.1023/A:1022869011914, URL <https://doi.org/10.1023/A:1022869011914>
- Littlestone N, Warmuth MK (1994) The weighted majority algorithm. *Information and Computation* 108(2):212–261
- Liu B, Ye M, Wright S, Stone P, Liu Q (2022) Bome! bilevel optimization made easy: a simple first-order approach. In: Proceedings of the 36th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, NIPS ’22
- Liu M, Yuan Z, Ying Y, Yang T (2020) Stochastic AUC maximization with deep neural networks. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net, URL <https://openreview.net/forum?id=HJepXaVYDr>
- Liu R, Liu X, Zeng S, Zhang J, Zhang Y (2021) Value-function-based sequential minimization for bi-level optimization. *CoRR* abs/2110.04974, URL <https://arxiv.org/abs/2110.04974>, 2110.04974
- Luenberger D (1973) Introduction to Linear and Nonlinear Programming. Addison-Wesley Publishing Company, URL <https://books.google.com/books?id=1aCrPQAACAAJ>
- Luo ZQ, Tseng P (1992) On the convergence of the coordinate descent method for convex differentiable minimization. *J Optim Theory Appl* 72(1):7–35, DOI 10.1007/BF00939948, URL <https://doi.org/10.1007/BF00939948>
- Ma J, Xiao L (2025) Quantization through piecewise-affine regularization: Optimization and statistical guarantees. URL <https://arxiv.org/abs/2508.11112>, 2508.11112
- Mahdavi M, Jin R (2013) Mixedgrad: An $o(1/t)$ convergence rate algorithm for stochastic smooth optimization. In: Advances in Neural Information Processing Systems, vol 26, URL <https://proceedings.neurips.cc/paper/2013/file/f73b76ce8949fe29bf2a537cfa420e8f-Paper.pdf>
- Marcum JI (1947) A statistical theory of target detection by pulsed radar. Tech. Rep. RM-754, RAND Corporation, Santa Monica, CA, URL https://www.rand.org/pubs/research_memoranda/RM754.html

- Martinet B (1972) Détermination approchée d'un point fixe d'une application pseudo-contractante. cas de l'application prox. *Comptes Rendus de l'Académie des Sciences, Paris, Série A* 274:163–165
- Mindermann S, Brauner JM, Razzak MT, Sharma M, Kirsch A, Xu W, Hölting B, Gomez AN, Morisot A, Farquhar S, Gal Y (2022) Prioritized training on points that are learnable, worth learning, and not yet learnt. In: Chaudhuri K, Jegelka S, Song L, Szepesvari C, Niu G, Sabato S (eds) *Proceedings of the 39th International Conference on Machine Learning*, PMLR, *Proceedings of Machine Learning Research*, vol 162, pp 15630–15649, URL <https://proceedings.mlr.press/v162/mindermann22a.html>
- Mohri M, Rostamizadeh A, Talwalkar A (2018) *Foundations of Machine Learning*, 2nd edn. MIT Press
- Morgan W, Greiff W, Henderson J (2004) Direct maximization of average precision by hill-climbing, with a comparison to a maximum entropy approach. In: *Proceedings of HLT-NAACL 2004: Short Papers*, Association for Computational Linguistics, USA, HLT-NAACL-Short '04, p 93–96
- Namkoong H, Duchi JC (2017) Variance-based regularization with convex objectives. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA, NIPS'17, p 2975–2984
- Narasimhan H, Agarwal S (2013) A structural SVM based approach for optimizing partial auc. In: Dasgupta S, McAllester D (eds) *Proceedings of the 30th International Conference on Machine Learning*, PMLR, Atlanta, Georgia, USA, *Proceedings of Machine Learning Research*, vol 28, pp 516–524, URL <https://proceedings.mlr.press/v28/narasimhan13.html>
- Nemirovski A (2004) Prox-method with rate of convergence $o(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization* 15(1):229–251, DOI 10.1137/S1052623403425629
- Nemirovski A, Yudin D (1978) On cezari's convergence of the steepest descent method for approximating saddle point of convex-concave functions. *Soviet Mathematics Doklady* 19:341–362
- Nemirovski A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization* 19(4):1574–1609, DOI 10.1137/070704277
- Nemirovski AS, Yudin DB (1977) Information complexity of strongly convex optimization. *Ekonomika i Matematicheskie Metody* 13(3):550–559, translated into English in *MATEKON*
- Nemirovsky AS, Yudin DB (1983) *Problem Complexity and Method Efficiency in Optimization*, Wiley-Interscience Series in Discrete Mathematics, vol 15. John Wiley and Sons, New York
- Nesterov Y (1983) A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Proceedings of the USSR Academy of Sciences* 269:543–547, URL <https://api.semanticscholar.org/CorpusID:145918791>

-
- Nesterov Y (2004) *Introductory Lectures on Convex Programming: A Basic Course*. Springer
- Nesterov Y (2012) Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization* 22(2):341–362, DOI 10.1137/100802001
- Nguyen LM, Liu J, Scheinberg K, Takávc M (2017) Sarah: a novel method for machine learning problems using stochastic recursive gradient. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, JMLR.org, ICML’17, p 2613–2621
- Orabona F (2019) A modern introduction to online learning. CoRR abs/1912.13213, URL <http://arxiv.org/abs/1912.13213>, 1912.13213
- Ortega JM, Rheinboldt WC (1970) *Iterative solution of nonlinear equations in several variables*. Academic Press, New York
- Pazy GB (1979) Ergodic convergence to a zero of the sum of monotone operators in hilbert space. *Journal of Mathematical Analysis and Applications* 72:383–390
- Polyak B (1964) Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics* 4(5):1–17, URL <https://www.sciencedirect.com/science/article/pii/0041555364901375>
- Polyak BT, Juditsky AB (1992) Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization* 30(4):838–855
- Qi C, Su H, Mo K, Guibas LJ (2016) Pointnet: Deep learning on point sets for 3d classification and segmentation. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 77–85, URL <https://api.semanticscholar.org/CorpusID:5115938>
- Qi Q, Xu Y, Jin R, Yin W, Yang T (2020) Attentional-biased stochastic gradient descent. *Trans Mach Learn Res* 2023, URL <https://api.semanticscholar.org/CorpusID:255125618>
- Qi Q, Guo Z, Xu Y, Jin R, Yang T (2021a) An online method for a class of distributionally robust optimization with non-convex objectives. In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA, NIPS ’21
- Qi Q, Guo Z, Xu Y, Jin R, Yang T (2021b) An online method for A class of distributionally robust optimization with non-convex objectives. In: Ranzato M, Beygelzimer A, Dauphin YN, Liang P, Vaughan JW (eds) *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp 10067–10080
- Qi Q, Luo Y, Xu Z, Ji S, Yang T (2021c) Stochastic optimization of areas under precision-recall curves with provable convergence. In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA, NIPS ’21
- Qi Q, Lyu J, Chan K, Bai E, Yang T (2023) Stochastic constrained DRO with a complexity independent of sample size. *Trans Mach Learn Res* 2023, URL <https://openreview.net/forum?id=VpaXrBFYZ9>

- Qiu S, Yang Z, Wei X, Ye J, Wang Z (2020) Single-timescale stochastic nonconvex-concave optimization for smooth nonlinear td learning. ArXiv abs/2008.10103, URL <https://api.semanticscholar.org/CorpusID:221266692>
- Qiu Z, Hu Q, Zhong Y, Zhang L, Yang T (2022) Large-scale stochastic optimization of NDCG surrogates for deep learning with provable convergence. In: Chaudhuri K, Jegelka S, Song L, Szepesvári C, Niu G, Sabato S (eds) International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, PMLR, Proceedings of Machine Learning Research, vol 162, pp 18122–18152, URL <https://proceedings.mlr.press/v162/qiu22a.html>
- Qiu Z, Hu Q, Yuan Z, Zhou D, Zhang L, Yang T (2023) Not all semantics are created equal: Contrastive self-supervised learning with automatic temperature individualization. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J (eds) International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, PMLR, Proceedings of Machine Learning Research, vol 202, pp 28389–28421, URL <https://proceedings.mlr.press/v202/qiu23a.html>
- Qiu Z, Guo S, Xu M, Zhao T, Zhang L, Yang T (2024) To cool or not to cool? temperature network meets large foundation models via DRO. In: Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024, OpenReview.net, URL <https://openreview.net/forum?id=YWuSLBkf0w>
- Rafique H, Liu M, Lin Q, Yang T (2018) Weakly-convex-concave min-max optimization: provable algorithms and applications in machine learning. Optimization Methods and Software 37:1087 – 1121, URL <https://api.semanticscholar.org/CorpusID:233790522>
- Rafique H, Liu M, Lin Q, Yang T (2022) Weakly-convex-concave min-max optimization: provable algorithms and applications in machine learning. Optim Methods Softw 37(3):1087–1121, DOI 10.1080/10556788.2021.1895152, URL <https://doi.org/10.1080/10556788.2021.1895152>
- Rakhlin A, Sridharan K (2013) Optimization, learning, and games with predictable sequences. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, Curran Associates Inc., Red Hook, NY, USA, NIPS’13, p 3066–3074
- Rakhlin A, Shamir O, Sridharan K (2012) Making gradient descent optimal for strongly convex stochastic optimization. In: Proceedings of the 29th International Conference on International Conference on Machine Learning, Omnipress, Madison, WI, USA, ICML’12, p 1571–1578
- Recht B, Wright SJ (2025) Optimization for Modern Data Analysis. Cambridge University Press, this is a hypothetical example; details may vary for actual publications.
- Robbins H, Monro S (1951) A stochastic approximation method. Annals of Mathematical Statistics 22:400–407
- Rockafellar RT (1970a) Convex Analysis. Princeton University Press
- Rockafellar RT (1970b) Convex analysis. Princeton Mathematical Series, Princeton University Press, Princeton, N. J.

-
- Rockafellar RT (1976) Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization* 14(5):877–898, DOI 10.1137/0314056
- Rosenblatt F (1962) *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington, D.C.
- Rustem B, Howe M (2002) *Algorithms for Worst-Case Design and Applications to Risk Management*. Princeton University Press, Princeton, NJ
- Sagawa S, Koh PW, Hashimoto TB, Liang P (2019) Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *CoRR* abs/1911.08731, URL <http://arxiv.org/abs/1911.08731>, 1911.08731
- Scarf H (1958) A min-max solution of an inventory problem. In: *Studies in the Mathematical Theory of Inventory and Production*, Stanford University Press, pp 201–209
- Schmidt M, Le Roux N, Bach F (2017) Minimizing finite sums with the stochastic average gradient. *Math Program* 162(1–2):83–112, DOI 10.1007/s10107-016-1030-6, URL <https://doi.org/10.1007/s10107-016-1030-6>
- Shalev-Shwartz S, Ben-David S (2014) *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press
- Shalev-Shwartz S, Wexler Y (2016) Minimizing the maximal loss: How and why? *CoRR* abs/1602.01690, URL <http://arxiv.org/abs/1602.01690>, 1602.01690
- Shalev-Shwartz S, Singer A, Srebro N (2007) Pegasos: Primal estimated sub-gradient solver for svm. In: *Proceedings of the 24th International Conference on Machine Learning*, pp 807–814
- Shamir O, Zhang T (2013) Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In: Dasgupta S, McAllester D (eds) *Proceedings of the 30th International Conference on Machine Learning*, PMLR, Atlanta, Georgia, USA, *Proceedings of Machine Learning Research*, vol 28, pp 71–79, URL <https://proceedings.mlr.press/v28/shamir13.html>
- Shapiro A, Kleywegt AJ (2002) Minimax analysis of stochastic problems. *Optimization Methods and Software* 17(3):523–542
- Shen H, Chen T (2023) On penalty-based bilevel gradient descent method. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J (eds) *Proceedings of the 40th International Conference on Machine Learning*, PMLR, *Proceedings of Machine Learning Research*, vol 202, pp 30992–31015, URL <https://proceedings.mlr.press/v202/shen23c.html>
- Sohn K (2016) Improved deep metric learning with multi-class n-pair loss objective. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA, NIPS’16, p 1857–1865
- Spackman KA (1989) Signal detection theory: valuable tools for evaluating inductive learning. In: *Proceedings of the Sixth International Workshop on Machine*

- Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p 160–163
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)* 58:267–288
- Tieleman T, Hinton G (2012) Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* 4(2):26–31
- Tseng P (1990) Dual ascent methods for problems with strictly convex costs and linear constraints: A unified approach. *SIAM Journal on Control and Optimization* 28(1):214–242
- Tseng P (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. *J Optim Theory Appl* 109(3):475–494, DOI 10.1023/A:1017501703105, URL <https://doi.org/10.1023/A:1017501703105>
- Tseng P, Bertsekas DP (1987) Relaxation methods for problems with strictly convex separable costs and linear constraints. *Math Program* 38(3):303–321
- Verrelst H, Moreau Y, Vandewalle J, Timmerman D (1998) Use of a multi-layer perceptron to predict malignancy in ovarian tumors. In: *Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems 10*, MIT Press, Cambridge, MA, USA, NIPS '97, p 978–984
- Vogel R, Bellet A, Cl'emencon S (2020) Learning fair scoring functions: Bipartite ranking under roc-based fairness constraints. In: *International Conference on Artificial Intelligence and Statistics*, URL <https://api.semanticscholar.org/CorpusID:224899598>
- Wang B, Yang T (2022) Finite-sum coupled compositional stochastic optimization: Theory and applications. In: Chaudhuri K, Jegelka S, Song L, Szepesvári C, Niu G, Sabato S (eds) *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, PMLR, *Proceedings of Machine Learning Research*, vol 162, pp 23292–23317, URL <https://proceedings.mlr.press/v162/wang22ak.html>
- Wang B, Yang T (2023) A near-optimal single-loop stochastic algorithm for convex finite-sum coupled compositional optimization. In: *International Conference on Machine Learning*, URL <https://api.semanticscholar.org/CorpusID:265658854>
- Wang B, Lei Y, Ying Y, Yang T (2025) On discriminative probabilistic modeling for self-supervised representation learning. In: *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, OpenReview.net, URL <https://openreview.net/forum?id=s15HrqCqbr>
- Wang G, Yang M, Zhang L, Yang T (2022) Momentum accelerates the convergence of stochastic AUPRC maximization. In: Camps-Valls G, Ruiz FJR, Valera I (eds) *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, PMLR, *Proceedings of Machine Learning Research*, vol 151, pp 3753–3771, URL <https://proceedings.mlr.press/v151/wang22b.html>
- Wang M, Fang EX, Liu H (2017a) Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Math Pro-*

-
- gram 161(1–2):419–449, DOI 10.1007/s10107-016-1017-3, URL <https://doi.org/10.1007/s10107-016-1017-3>
- Wang M, Liu J, Fang EX (2017b) Accelerating stochastic composition optimization. *Journal of Machine Learning Research* 18(105):1–23, URL <http://jmlr.org/papers/v18/16-504.html>
- Warga J (1963) Minimizing certain convex functions. *Journal of the Society for Industrial and Applied Mathematics* 11(3):588–593
- Wei X, Ye F, Yonay O, Chen X, Sun B, Tao D, Yang T (2024) Fastclip: A suite of optimization techniques to accelerate CLIP training with limited resources. CoRR abs/2407.01445, DOI 10.48550/ARXIV.2407.01445, URL <https://doi.org/10.48550/arXiv.2407.01445>, 2407.01445
- Wei X, Lin MC, Ye F, Song F, Cao L, Thai MT, Yang T (2025) Model steering: Learning with a reference model improves generalization bounds and scaling laws. In: Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13–19, 2025, OpenReview.net, URL <https://openreview.net/forum?id=QC4dfob0LQ>
- Wei X, Zhou L, Wang B, Lin CJ, Yang T (2026) A geometry-aware efficient algorithm for compositional entropic risk minimization. arXiv
- Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10(9):207–244, URL <http://jmlr.org/papers/v10/weinberger09a.html>
- Widrow B, Hoff ME (1960) Adaptive switching circuits. IRE WESCON Convention Record 4:96–104
- Xu Y, Lin Q, Yang T (2017) Stochastic convex optimization: Faster local growth implies faster global convergence. In: Precup D, Teh YW (eds) *Proceedings of the 34th International Conference on Machine Learning*, PMLR, *Proceedings of Machine Learning Research*, vol 70, pp 3821–3830, URL <https://proceedings.mlr.press/v70/xu17a.html>
- Xu Y, Jin R, Yang T (2019a) Non-asymptotic analysis of stochastic methods for non-smooth non-convex regularized problems. In: Wallach H, Larochelle H, Beygelzimer A, d’Alché-Buc F, Fox E, Garnett R (eds) *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol 32
- Xu Y, Zhu S, Yang S, Zhang C, Jin R, Yang T (2019b) Learning with non-convex truncated losses by SGD. In: Globerson A, Silva R (eds) *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI 2019, Tel Aviv, Israel, July 22–25, 2019, AUAI Press, *Proceedings of Machine Learning Research*, vol 115, pp 701–711, URL <http://proceedings.mlr.press/v115/xu20b.html>
- Yan L, Dodier R, Mozer MC, Wolniewicz R (2003) Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic. In: *Proceedings of the 20th International Conference on Machine Learning (ICML)*, AAAI Press, pp 848–855
- Yan Y, Xu Y, Lin Q, Liu W, Yang T (2020a) Optimal epoch stochastic gradient descent ascent methods for min-max optimization. arXiv: Optimization and Control URL <https://api.semanticscholar.org/CorpusID:226148475>

- Yan Y, Xu Y, Lin Q, Liu W, Yang T (2020b) Optimal epoch stochastic gradient descent ascent methods for min-max optimization. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, URL <https://proceedings.neurips.cc/paper/2020/hash/3f8b2a81da929223ae025fcec26dde0d-Abstract.html>
- Yang F, Koyejo S (2020) On the consistency of top-k surrogate losses. In: International Conference on Machine Learning (ICML), PMLR, pp 10727–10735
- Yang H, Lu K, Lyu X, Hu F (2019) Two-way partial AUC and its properties. *Statistical Methods in Medical Research* 28(1):184–195, DOI 10.1177/0962280217718866, URL <https://doi.org/10.1177/0962280217718866>
- Yang J, Ji K, Liang Y (2021) Provably faster algorithms for bilevel optimization. In: Proceedings of the 35th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, NIPS '21
- Yang L, Jin R (2006) Distance metric learning: A comprehensive survey. Tech. Rep. 2, Department of Computer Science and Engineering, Michigan State University
- Yang M, Li G, Hu Q, Lin Q, Yang T (2025) Single-loop algorithms for stochastic non-convex optimization with weakly-convex constraints. *CoRR* abs/2504.15243, DOI 10.48550/ARXIV.2504.15243, URL <https://doi.org/10.48550/arXiv.2504.15243>, 2504.15243
- Yang T (2022) Algorithmic foundation of empirical x-risk minimization. *arXiv preprint arXiv:220600439*
- Yang T, Ying Y (2022) AUC maximization in the era of big data and ai: A survey. *ACM Comput Surv* 55(8), DOI 10.1145/3554729, URL <https://doi.org/10.1145/3554729>
- Yang T, Ying Y (2023) AUC maximization in the era of big data and AI: A survey. *ACM Comput Surv* 55(8):172:1–172:37, DOI 10.1145/3554729, URL <https://doi.org/10.1145/3554729>
- Yang T, Lin Q, Li Z (2016) Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. URL <https://arxiv.org/abs/1604.03257>, 1604.03257
- Ye J, Zhu D, Zhu Q (1997) Exact penalization and necessary optimality conditions for generalized bilevel programming problems. *SIAM Journal on Optimization* 7(2):481–507
- Ying Y, Wen L, Lyu S (2016a) Stochastic online AUC maximization. In: Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 29, pp 451–459
- Ying Y, Wen L, Lyu S (2016b) Stochastic online auc maximization. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, NIPS'16, p 451–459
- Yu H, Wang L, Wang B, Liu M, Yang T, Ji S (2022) Graphfm: Improving large-scale GNN training via feature momentum. In: Chaudhuri K, Jegelka S, Song L, Szepesvári C, Niu G, Sabato S (eds) International Conference on Machine

-
- Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, PMLR, Proceedings of Machine Learning Research, vol 162, pp 25684–25701, URL <https://proceedings.mlr.press/v162/yu22g.html>
- Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(1):49–67
- Yuan Z, Yan Y, Sonka M, Yang T (2021) Large-scale robust deep AUC maximization: A new surrogate loss and empirical studies on medical image classification. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, IEEE, pp 3020–3029, DOI 10.1109/ICCV48922.2021.00303, URL <https://doi.org/10.1109/ICCV48922.2021.00303>
- Yuan Z, Guo Z, Chawla N, Yang T (2022a) Compositional training for end-to-end deep AUC maximization. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=gPvB4pdu_Z
- Yuan Z, Guo Z, Chawla NV, Yang T (2022b) Compositional training for end-to-end deep AUC maximization. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net, URL https://openreview.net/forum?id=gPvB4pdu_Z
- Yuan Z, Wu Y, Qiu Z, Du X, Zhang L, Zhou D, Yang T (2022c) Provable stochastic optimization for global contrastive learning: Small batch does not harm performance. In: Chaudhuri K, Jegelka S, Song L, Szepesvári C, Niu G, Sabato S (eds) International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, PMLR, Proceedings of Machine Learning Research, vol 162, pp 25760–25782, URL <https://proceedings.mlr.press/v162/yuan22b.html>
- Yuan Z, Zhu D, Qiu Z, Li G, Wang X, Yang T (2023a) Libauc: A deep learning library for x-risk optimization. In: Singh AK, Sun Y, Akoglu L, Gunopulos D, Yan X, Kumar R, Ozcan F, Ye J (eds) Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023, ACM, pp 5487–5499, DOI 10.1145/3580305.3599861, URL <https://doi.org/10.1145/3580305.3599861>
- Yuan Z, Zhu D, Qiu ZH, Li G, Wang X, Yang T (2023b) Libauc: A deep learning library for x-risk optimization. In: 29th SIGKDD Conference on Knowledge Discovery and Data Mining
- Zaheer M, Kottur S, Ravanbakhsh S, Póczos B, Salakhutdinov R, Smola AJ (2017) Deep sets. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, NIPS’17, p 3394–3404
- Zhang CH (2010) Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 38(2):894–942
- Zhang J, Xiao L (2019) A stochastic composite gradient method with incremental variance reduction. In: Wallach H, Larochelle H, Beygelzimer A, d’Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, URL

- https://proceedings.neurips.cc/paper_files/paper/2019/file/a68259547f3d25ab3c0a5c0adb4e3498-Paper.pdf
- Zhang L, Mahdavi M, Jin R (2013) Linear convergence with condition number independent access of full gradients. In: Burges C, Bottou L, Welling M, Ghahramani Z, Weinberger K (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 26, URL https://proceedings.neurips.cc/paper_files/paper/2013/file/37f0e884fbad9667e38940169d0a3c95-Paper.pdf
- Zhang T (2004a) Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: Greiner R, Schuurmans D (eds) Proceedings of the Twenty-First International Conference on Machine Learning, ICML 2004, ACM, pp 919–926
- Zhang T (2004b) Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research* 5:1225–1251
- Zhang T (2013) Multi-stage convex relaxation for feature selection. *Bernoulli* 19(5B):2277–2293
- Zhang X, Aybat NS, Gürbüzbalaban M (2021) Robust accelerated primal-dual methods for computing saddle points. *SIAM J Optim* 34:1097–1130, URL <https://api.semanticscholar.org/CorpusID:244709301>
- Zhang X, Aybat N, Gurbuzbalaban M (2022) SAPD+: An accelerated stochastic method for nonconvex-concave minimax problems. In: Oh AH, Agarwal A, Belgrave D, Cho K (eds) Advances in Neural Information Processing Systems, URL <https://openreview.net/forum?id=GiUpEVQmNx8>
- Zhang Z, Lan G (2024) Optimal methods for convex nested stochastic composite optimization: Optimal methods for convex nested... *Math Program* 212(1):1–48, DOI 10.1007/s10107-024-02090-3, URL <https://doi.org/10.1007/s10107-024-02090-3>
- Zhou L, Wang B, Thai MT, Yang T (2025) Stochastic primal-dual double block-coordinate for two- way partial AUC maximization. *Transactions on Machine Learning Research* URL <https://openreview.net/forum?id=M3kibBFP4q>
- Zhu D, Li G, Wang B, Wu X, Yang T (2022a) When AUC meets DRO: optimizing partial AUC for deep learning with non-convex convergence guarantee. In: Chaudhuri K, Jegelka S, Song L, Szepesvári C, Niu G, Sabato S (eds) International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, PMLR, Proceedings of Machine Learning Research, vol 162, pp 27548–27573, URL <https://proceedings.mlr.press/v162/zhu22g.html>
- Zhu D, Li G, Wang B, Wu X, Yang T (2022b) When auc meets dro: Optimizing partial auc for deep learning with non-convex convergence guarantee. *ArXiv abs/2203.00176*, URL <https://api.semanticscholar.org/CorpusID:247187969>
- Zhu D, Wu X, Yang T (2022c) Benchmarking deep AUROC optimization: Loss functions and algorithmic choices. *CoRR abs/2203.14177*, DOI 10.48550/ARXIV.2203.14177, URL <https://doi.org/10.48550/arXiv.2203.14177>, 2203.14177

-
- Zhu D, Wang B, Chen Z, Wang Y, Sonka M, Wu X, Yang T (2023a) Provable multi-instance deep AUC maximization with stochastic pooling. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J (eds) International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, PMLR, Proceedings of Machine Learning Research, vol 202, pp 43205–43227, URL <https://proceedings.mlr.press/v202/zhu23l.html>
- Zhu D, Ying Y, Yang T (2023b) Label distributionally robust losses for multi-class classification: Consistency, robustness and adaptivity. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J (eds) International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, PMLR, Proceedings of Machine Learning Research, vol 202, pp 43289–43325, URL <https://proceedings.mlr.press/v202/zhu23o.html>
- Zhu L, Gürbüzbalaban M, Ruszczyński A (2023c) Distributionally robust learning with weakly convex losses: Convergence rates and finite-sample guarantees. URL <https://arxiv.org/abs/2301.06619>, 2301.06619
- Zinkevich M (2003) Online convex programming and generalized infinitesimal gradient ascent. In: Proceedings of the Twentieth International Conference on International Conference on Machine Learning, AAAI Press, ICML'03, p 928–935
- Zou H, Hastie T (2003) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301–320