

References

- 310
311 Stefan Adams. Lecture notes in high-dimensional probability. [https://warwick.ac.uk/](https://warwick.ac.uk/fac/sci/math/people/staff/stefan_adams/high-dimensional_probability_ma3k0-notes.pdf)
312 [fac/sci/math/people/staff/stefan_adams/high-dimensional_probability_](https://warwick.ac.uk/fac/sci/math/people/staff/stefan_adams/high-dimensional_probability_ma3k0-notes.pdf)
313 [ma3k0-notes.pdf](https://warwick.ac.uk/fac/sci/math/people/staff/stefan_adams/high-dimensional_probability_ma3k0-notes.pdf), 2022.
- 314 Nicholas D. Alikakos, Peter W. Bates, and Xinfu Chen. Convergence of the Cahn-Hilliard equation
315 to the Hele-Shaw model. *Arch. Rational Mech. Anal.*, 128(2):165–205, 1994. ISSN 0003-9527.
316 doi: 10.1007/BF00375025. URL <http://dx.doi.org/10.1007/BF00375025>.
- 317 Annika Bach, Roberta Marziani, and Caterina Ida Zeppieri. γ -convergence and stochastic homogeni-
318 sation of singularly-perturbed elliptic functionals. *arXiv preprint arXiv:2102.09872*, 2021.
- 319 Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of*
320 *Machine Learning Research*, 18(1):629–681, 2017.
- 321 Heinz Bauer. Minimalstellen von funktionen und extremalpunkte. *Archiv der Mathematik*, 9(4):
322 389–393, 1958.
- 323 Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning
324 practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*,
325 116(32):15849–15854, 2019.
- 326 Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM*
327 *Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.
- 328 Kaushik Bhattacharya, Marta Lewicka, and Mathias Schäffner. Plates with incompatible prestrain.
329 *Archive for Rational Mechanics and Analysis*, 221(1):143–181, 2016.
- 330 Andrea Braides. Γ -convergence for beginners, volume 22 of *Oxford Lecture Series in Mathemat-*
331 *ics and its Applications*. Oxford University Press, Oxford, 2002. ISBN 0-19-850784-4. doi:
332 10.1093/acprof:oso/9780198507840.001.0001. URL [http://dx.doi.org/10.1093/acprof:](http://dx.doi.org/10.1093/acprof:oso/9780198507840.001.0001)
333 [oso/9780198507840.001.0001](http://dx.doi.org/10.1093/acprof:oso/9780198507840.001.0001).
- 334 Andrea Braides and Lev Truskinovsky. Asymptotic expansions by γ -convergence. *Continuum*
335 *Mechanics and Thermodynamics*, 20:21–62, 2008.
- 336 Lia Bronsard and Robert V. Kohn. On the slowness of phase boundary motion in one space dimension.
337 *Comm. Pure Appl. Math.*, 43(8):983–997, 1990. ISSN 0010-3640. doi: 10.1002/cpa.3160430804.
338 URL <http://dx.doi.org/10.1002/cpa.3160430804>.
- 339 Andrei Caragea, Philipp Petersen, and Felix Voigtlaender. Neural network approximation and estima-
340 tion of classifiers with classification boundary in a barron class. *arXiv preprint arXiv:2011.09363*,
341 2020.
- 342 Lenaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized
343 models using optimal transport. *Advances in neural information processing systems*, 31, 2018.
- 344 Lenaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks
345 trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.
- 346 Lenaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming.
347 *Advances in Neural Information Processing Systems*, 32, 2019.
- 348 Yaim Cooper. The loss landscape of overparameterized neural networks. *arXiv preprint*
349 *arXiv:1804.10200*, 2018.
- 350 Gianni Dal Maso. *An introduction to Γ -convergence*, volume 8. Springer Science & Business Media,
351 2012.
- 352 Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers.
353 *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.
- 354 Ennio De Giorgi and Tullio Franzoni. Su un tipo di convergenza variazionale. *Atti della Accademia*
355 *Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti*, 58(6):842–
356 850, 1975.

- 357 Manfred Dobrowolski. *Angewandte Funktionalanalysis: Funktionalanalysis, Sobolev-Räume und*
358 *Elliptische Differentialgleichungen*. Springer-Verlag, 2010.
- 359 Patrick W Dondl, Matthias W Kurzke, and Stephan Wojtowytsch. The effect of forest dislocations on
360 the evolution of a phase-field model for plastic slip. *Archive for Rational Mechanics and Analysis*,
361 232(1):65–119, 2019.
- 362 Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes
363 over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.
- 364 Weinan E and Stephan Wojtowytsch. Representation formulas and pointwise properties for Barron
365 functions. *Calc. Var. Partial Differential Equations*, 61(46), 2020.
- 366 Weinan E and Stephan Wojtowytsch. Kolmogorov width decay and poor approximators in machine
367 learning: Shallow neural networks, random feature models and neural tangent kernels. *Research in*
368 *the Mathematical Sciences*, 8(1):1–28, 2021.
- 369 Weinan E and Stephan Wojtowytsch. On the emergence of simplex symmetry in the final and
370 penultimate layers of neural network classifiers. In *Mathematical and Scientific Machine Learning*,
371 pages 270–290. PMLR, 2022.
- 372 Weinan E, Chao Ma, and Lei Wu. A priori estimates of the population risk for two-layer neural
373 networks. *Communications in Mathematical Sciences*, 17(5):1407–1425, 2019a. doi: 10.4310/
374 cms.2019.v17.n5.a11. URL <https://doi.org/10.4310%2Fcms.2019.v17.n5.a11>.
- 375 Weinan E, Chao Ma, and Lei Wu. A comparative analysis of optimization and generalization
376 properties of two-layer neural network and random feature models under gradient descent dynamics.
377 *Sci. China Math*, 2019b.
- 378 Weinan E, Chao Ma, and Lei Wu. The barron space and the flow-induced function spaces for neural
379 network models. *arXiv:1906.08039 [cs.LG]*, 2019c.
- 380 Lawrence Craig Evans and Ronald F Gariepy. *Measure theory and fine properties of functions*. CRC
381 press, 2015.
- 382 Gero Friesecke, Richard D James, and Stefan Müller. Rigorous derivation of nonlinear plate theory
383 and geometric rigidity. *Comptes Rendus Mathématique*, 334(2):173–178, 2002a.
- 384 Gero Friesecke, Richard D James, and Stefan Müller. A theorem on geometric rigidity and the
385 derivation of nonlinear plate theory from three-dimensional elasticity. *Communications on Pure*
386 *and Applied Mathematics*, 55(11):1461–1506, 2002b.
- 387 Gero Friesecke, Richard D James, Maria Giovanna Mora, and Stefan Müller. Derivation of nonlinear
388 bending theory for shells from three-dimensional nonlinear elasticity by gamma-convergence.
389 *Comptes Rendus Mathématique*, 336(8):697–702, 2003.
- 390 Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural
391 networks. In *Proceedings of the thirteenth international conference on artificial intelligence and*
392 *statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.
- 393 Boris Hanin. Ridgeless interpolation with shallow relu networks in $1d$ is nearest neighbor curvature
394 extrapolation and provably generalizes on lipschitz functions. *arXiv preprint arXiv:2109.12960*,
395 2021.
- 396 Boris Hanin and David Rolnick. How to start training: The effect of initialization and architecture.
397 *Advances in Neural Information Processing Systems*, 31, 2018.
- 398 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing
399 human-level performance on imagenet classification. In *Proceedings of the IEEE international*
400 *conference on computer vision*, pages 1026–1034, 2015.
- 401 Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

- 402 Jean Honorio and Tommi Jaakkola. Tight Bounds for the Expected Risk of Linear Classifiers and
403 PAC-Bayes Finite-Sample Guarantees. In Samuel Kaski and Jukka Corander, editors, *Proceedings*
404 *of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of
405 *Proceedings of Machine Learning Research*, pages 384–392, Reykjavik, Iceland, 22–25 Apr 2014.
406 PMLR. URL <https://proceedings.mlr.press/v33/honorio14.html>.
- 407 Tom Ilmanen. Convergence of the Allen-Cahn equation to Brakke’s motion by mean curvature. *J.*
408 *Differential Geom.*, 38(2):417–461, 1993. ISSN 0022-040X. URL <http://projecteuclid.org/euclid.jdg/1214454300>.
- 409
- 410 Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*
411 *arXiv:1412.6980*, 2014.
- 412 Achim Klenke. *Wahrscheinlichkeitstheorie*, volume 1. Springer, 2006.
- 413 Marta Lewicka, Maria Giovanna Mora, and Mohammad Reza Pakzad. Shell theories arising as low
414 energy γ -limit of 3d nonlinear elasticity. *Annali della Scuola Normale Superiore di Pisa-Classe di*
415 *Scienze*, 9(2):253–295, 2010.
- 416 Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after sgd reaches zero loss?—a
417 mathematical framework. *arXiv preprint arXiv:2110.06914*, 2021.
- 418 Zhong Li, Chao Ma, and Lei Wu. Complexity measures for neural networks with general activation
419 functions using path-based norms. *arXiv preprint arXiv:2009.06132*, 2020.
- 420 Marco Loog, Tom Viering, Alexander Mey, Jesse H Krijthe, and David MJ Tax. A brief prehistory of
421 double descent. *Proceedings of the National Academy of Sciences*, 117(20):10625–10626, 2020.
- 422 Chao Ma, Lei Wu, et al. Machine learning from a continuous viewpoint, i. *Science China Mathematics*,
423 63(11):2233–2266, 2020.
- 424 Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-
425 layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671,
426 2018.
- 427 Luciano Modica. The gradient theory of phase transitions and the minimal interface criterion. *Arch*
428 *Ration Mech Anal*, 98(2):123–142, 1987.
- 429 Luciano Modica and Stefano Mortola. Un esempio di Γ -convergenza. *Boll. Un. Mat. Ital. B (5)*, 14
430 (1):285–299, 1977.
- 431 Luca Mugnai and Matthias Röger. Convergence of perturbed Allen-Cahn equations to forced mean
432 curvature flow. *Indiana Univ. Math. J.*, 60(1):41–75, 2011. ISSN 0022-2518. doi: 10.1512/iumj.
433 2011.60.3949. URL <http://dx.doi.org/10.1512/iumj.2011.60.3949>
- 434 Greg Ongie and Rebecca Willett. The role of linear layers in nonlinear interpolating networks. *arXiv*
435 *preprint arXiv:2202.00856*, 2022.
- 436 Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded
437 norm infinite width relu nets: The multivariate case. *arXiv preprint arXiv:1910.01635*, 2019.
- 438 Rahul Parhi and Robert D Nowak. Banach space representer theorems for neural networks and ridge
439 splines. *J. Mach. Learn. Res.*, 22(43):1–40, 2021.
- 440 Rahul Parhi and Robert D Nowak. What kinds of functions do deep neural networks learn? insights
441 from variational spline theory. *SIAM Journal on Mathematics of Data Science*, 4(2):464–489,
442 2022.
- 443 Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear
444 networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*,
445 34:29218–29230, 2021.
- 446 Grant M Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems:
447 Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *stat*,
448 1050:22, 2018.

- 449 W. Rudin. *Functional Analysis*. International series in pure and applied mathematics. McGraw-Hill,
450 1991. ISBN 9780070542365.
- 451 Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural net-
452 works. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018,*
453 *Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine*
454 *Learning Research*, pages 4430–4438. PMLR, 2018a. URL [http://proceedings.mlr.press/
455 v80/safran18a.html](http://proceedings.mlr.press/v80/safran18a.html).
- 456 Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks.
457 In *International conference on machine learning*, pages 4433–4441. PMLR, 2018b.
- 458 Etienne Sandier and Sylvia Serfaty. Gamma-convergence of gradient flows with applications to
459 Ginzburg-Landau. *Communications on Pure and Applied Mathematics: A Journal Issued by the*
460 *Courant Institute of Mathematical Sciences*, 57(12):1627–1672, 2004.
- 461 Sylvia Serfaty. Gamma-convergence of gradient flows on Hilbert and metric spaces and applications.
462 *Discrete Contin. Dyn. Syst*, 31(4):1427–1451, 2011.
- 463 Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to*
464 *algorithms*. Cambridge university press, 2014.
- 465 Jonathan W Siegel and Jinchao Xu. Approximation rates for neural networks with general activation
466 functions. *Neural Networks*, 128:313–321, 2020.
- 467 Jonathan W Siegel and Jinchao Xu. Characterization of the variation spaces corresponding to shallow
468 neural networks. *arXiv preprint arXiv:2106.15002*, 2021.
- 469 Jonathan W Siegel and Jinchao Xu. Sharp bounds on the approximation rates, metric entropy, and
470 n-widths of shallow neural networks. *Foundations of Computational Mathematics*, pages 1–57,
471 2022.
- 472 Jonathan W Siegel and Jinchao Xu. Characterization of the variation spaces corresponding to shallow
473 neural networks. *Constructive Approximation*, pages 1–24, 2023.
- 474 Justin Sirignano and Konstantinos Spiliopoulos. Scaling limit of neural networks with the xavier
475 initialization and convergence to a global minimum. *arXiv preprint arXiv:1907.04108*, 2019.
- 476 Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of
477 large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020a.
- 478 Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central
479 limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020b.
- 480 Luca Venturi, Afonso S Bandeira, and Joan Bruna. Spurious valleys in two-layer neural network
481 optimization landscapes. *arXiv preprint arXiv:1802.06384*, 2018.
- 482 Stephan Wojtowytsch. On the convergence of gradient descent training for two-layer relu-networks
483 in the mean field regime. *arXiv preprint arXiv:2005.13530*, 2020.
- 484 Stephan Wojtowytsch. Optimal bump functions for shallow relu networks: Weight decay, depth
485 separation and the curse of dimensionality. *arXiv preprint arXiv:2209.01173*, 2022.
- 486 Lei Wu, Mingze Wang, and Weijie Su. The alignment property of sgd noise and how it helps select flat
487 minima: A stability analysis. *Advances in Neural Information Processing Systems*, 35:4680–4693,
488 2022.
- 489 Yaoqing Yang, Liam Hodgkinson, Ryan Theisen, Joe Zou, Joseph E Gonzalez, Kannan Ramchandran,
490 and Michael W Mahoney. Taxonomizing local versus global structure in neural network loss
491 landscapes. *Advances in Neural Information Processing Systems*, 34:18722–18733, 2021.
- 492 Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, et al. Towards theoretically
493 understanding why sgd generalizes better than adam in deep learning. *Advances in Neural*
494 *Information Processing Systems*, 33:21285–21296, 2020.