American Journal of Data Science and Machine Learning
ISSN: - applied

Global Convergence, Stochastic Approximation, and Optimization Landscapes in Overparameterized Deep Learning: A Unified Theoretical Analysis of Gradient Based Methods

[1] Prof. Martina L. Rossi
[1] Department of Mathematics and Computer Science, University of Zurich, Switzerland

Received: 11th Nov 2025 | Received Revised Version: 22th Nov 2025 | Accepted: 29th Nov 2025 | Published: 16th Dec 2025

Volume 01 Issue 02 2025 | Crossref DOI: 10.64917/ajdsml/V01I02-003

Abstract

The rapid expansion of deep learning has placed gradient based optimization methods at the center of modern machine learning theory and practice. Despite their apparent simplicity, algorithms such as gradient descent, stochastic gradient descent, momentum variants, proximal methods, and adaptive schemes demonstrate remarkable empirical performance even in highly nonconvex and overparameterized regimes. This article develops a comprehensive and unified theoretical framework for understanding convergence, stability, and generalization of gradient based optimization methods in convex, weakly convex, and nonconvex settings, with particular emphasis on deep neural networks. Drawing exclusively on foundational and contemporary research in stochastic approximation, incremental gradient methods, mean field theory, neural tangent kernels, and Polyak Lojasiewicz geometry, this work synthesizes classical optimization principles with modern overparameterized learning theory.

We begin by revisiting deterministic gradient descent under convex and Polyak Lojasiewicz conditions, establishing its convergence properties and complexity guarantees. We then extend the analysis to stochastic approximation frameworks rooted in the Robbins Monro paradigm, examining almost sure convergence and finite time convergence rates under diminishing and constant step sizes. The interplay between variance, minibatching, and interpolation is explored to explain the surprising efficiency of stochastic gradient descent in large scale machine learning. Accelerated and momentum based methods are analyzed in both convex and nonconvex contexts, with special attention to variance reduction techniques and adaptive step size strategies.

A central contribution of this article is the integration of mean field limits and neural tangent kernel perspectives with classical stochastic approximation theory. We demonstrate how overparameterization reshapes the optimization landscape, producing regimes in which gradient descent enjoys global convergence guarantees. The role of lazy training, optimal transport formulations, and gradient flow approximations is examined in depth. Furthermore, we connect Lojasiewicz gradient inequalities to generalization behavior, illustrating how optimization dynamics influence statistical performance.

Through extensive theoretical elaboration, we reveal that many seemingly disparate results share a common geometric and probabilistic structure. This unified view clarifies the mechanisms underlying large minibatch training, structured nonconvex objectives, and composite optimization. We conclude with a detailed discussion of limitations, open theoretical questions, and promising directions for bridging optimization and generalization in deep learning.

Keywords: Gradient descent, stochastic gradient descent, overparameterization, Polyak Lojasiewicz condition, mean field theory, neural tangent kernel, stochastic approximation.

Cite This Article: Prof. Martina L. Rossi. 2025. Global Convergence, Stochastic Approximation, and Optimization Landscapes in Overparameterized Deep Learning: A Unified Theoretical Analysis of Gradient Based Methods. American Journal of Data Science and Machine Learning 1, 02, 13-18. https://doi.org/10.64917/ajdsml/V01I02-003

1. Introduction

The success of modern machine learning, and deep learning in particular, rests fundamentally on the effectiveness of gradient based optimization methods. Algorithms such as gradient descent and stochastic gradient descent have become the primary tools for training neural networks with millions or even billions of parameters. From image recognition tasks involving large datasets to language models and generative architectures, the practical dominance of these methods raises profound theoretical questions. Why do simple first order methods converge reliably in highly nonconvex landscapes? How do stochastic approximations interact with overparameterization? Under what geometric conditions can one guarantee global convergence? And how do optimization dynamics influence generalization?

Classical optimization theory provides strong answers in convex settings. Foundational treatments such as those of Polyak and Nesterov describe deterministic gradient descent as a method with well characterized convergence rates under smoothness and strong convexity assumptions (Polyak, 1987; Nesterov, 2003). Extensions to incremental and proximal methods broaden this theory to composite and structured problems (Bertsekas, 2012). However, deep neural networks are neither convex nor low dimensional. Their loss surfaces exhibit nonconvexity, saddle points, and flat directions. Traditional guarantees therefore appear insufficient.

The stochastic nature of modern training further complicates analysis. The introduction of stochastic approximation by Robbins and Monro formalized the idea of replacing exact gradients with noisy observations (Robbins and Monro, 1951). Subsequent developments by Chung and by Kushner and Yin refined convergence theory for recursive stochastic algorithms (Chung, 1954; Kushner and Yin, 2003). These works established almost sure convergence under diminishing step sizes and appropriate noise conditions. Yet they were not originally designed for the massive overparameterized regimes of deep learning.

Recent advances have revisited gradient based methods in large scale machine learning. Bottou, Curtis, and Nocedal provide a comprehensive review of optimization methods tailored for large datasets, emphasizing stochastic and variance reduced techniques (Bottou et al., 2018). Reddi et al. extend stochastic variance reduction to nonconvex problems, offering improved convergence guarantees (Reddi et al., 2016). Ghadimi and Lan analyze accelerated gradient methods in nonconvex settings, revealing nuanced trade offs between speed and stability (Ghadimi and Lan, 2016). These works collectively suggest that nonconvexity does not preclude meaningful convergence guarantees.

The phenomenon of overparameterization introduces an additional layer of structure. Empirical evidence shows that neural networks with more parameters than training samples can achieve near zero training error and often generalize well. Theoretical investigations have revealed surprising geometric properties in such regimes. Jacot, Gabriel, and Hongler introduce the neural tangent kernel framework, demonstrating that infinitely wide networks trained by gradient descent behave like kernel methods (Jacot et al., 2018). Chizat and Bach show that gradient descent can converge globally in overparameterized models via optimal transport formulations (Chizat and Bach, 2018). Mean field analyses further describe the loss landscape of two layer networks, revealing the absence of spurious local minima under certain conditions (Mei et al., 2018; Soltanolkotabi et al., 2018).

At the same time, alternative perspectives challenge simplistic interpretations. Chizat, Oyallon, and Bach describe lazy training, a regime in which parameters move minimally and the model behaves linearly (Chizat et al., 2019). Liu, Zhu, and Belkin examine loss landscapes in nonlinear overparameterized systems, uncovering complex geometric structures (Liu et al., 2022). These results suggest that the interplay between optimization and model structure is subtle and multifaceted.

Another line of research focuses on geometric conditions weaker than strong convexity. The Polyak Lojasiewicz condition guarantees linear convergence without requiring convexity (Karimi et al., 2016). Extensions of the Lojasiewicz gradient inequality connect optimization dynamics to generalization properties (Liu et al., 2022). Jentzen and Riekert investigate the existence of global minima and convergence of gradient descent in deep neural networks, revealing conditions under which convergence is guaranteed (Jentzen and Riekert, 2021).

Despite this wealth of research, a unified narrative remains elusive. Convex analysis, stochastic approximation, mean field theory, neural tangent kernels, and nonconvex geometry are often treated as separate domains. The objective of this article is to integrate these strands into a cohesive theoretical framework. By systematically examining deterministic and stochastic gradient methods across convex and nonconvex regimes, and by situating overparameterization within classical approximation theory, we aim to clarify the structural reasons for convergence and stability in deep learning.

The core problem addressed in this work is the following. Given a potentially nonconvex objective function arising from empirical risk minimization in an overparameterized neural network, under what conditions and through which mechanisms does gradient based optimization converge to a global or near global minimum? Furthermore, how do stochastic approximations, minibatching, variance reduction, and adaptive step sizes influence both convergence rates and generalization?

A gap in the literature lies in the absence of a comprehensive synthesis that traces a continuous line from Robbins Monro stochastic approximation to neural tangent kernel limits and Polyak Lojasiewicz geometry. Many analyses focus on specialized regimes or specific algorithmic variants. This article instead proposes a unified perspective that emphasizes common structural principles: smoothness, interpolation, variance control, geometric regularity, and infinite width limits.

In the sections that follow, we develop this perspective in detail. We first revisit deterministic gradient descent in convex and PL regimes, then extend to stochastic approximation and nonconvex convergence. We incorporate variance reduction and momentum methods, examine mean field and kernel limits, and explore the role of geometric inequalities in linking optimization to generalization. The resulting synthesis provides both conceptual clarity and a roadmap for future theoretical development.

2. Methodology

The methodological approach of this article is entirely theoretical and synthetic. Rather than introducing new empirical experiments or computational simulations, the analysis proceeds by systematically integrating results from established research on optimization and stochastic approximation. The guiding principle is structural unification. Each class of methods is examined through the lens of its convergence properties, geometric assumptions, and stochastic behavior, and then placed within a broader conceptual framework.

We begin with deterministic gradient descent under smooth convexity assumptions. Classical analyses show that if an objective function is smooth and strongly convex, gradient descent converges linearly to the unique minimizer (Nesterov, 2003; Polyak, 1987). This linear rate arises from a contraction property induced by curvature. However, strong convexity is often absent in machine learning problems. The Polyak Lojasiewicz condition relaxes convexity while preserving linear convergence. Under this condition, the squared gradient norm dominates the function suboptimality, ensuring that descent steps produce proportional reductions in objective value (Karimi et al., 2016).

The methodological insight here is to reinterpret the PL condition not merely as a technical assumption but as a geometric property of the loss landscape. It implies that the gradient field consistently points toward a region of minimal value, even if the surface is not globally convex. This interpretation becomes crucial when analyzing overparameterized networks, where global convexity is absent but certain interpolation properties induce PL like behavior locally or globally.

Next, we incorporate incremental and proximal methods, drawing on the survey of Bertsekas (2012). Incremental gradient methods decompose the objective into finite sums, updating parameters using one component at a time. This framework naturally aligns with empirical risk minimization. Proximal methods extend gradient descent to composite objectives with nonsmooth components. The unifying methodological device is to express these updates as approximate gradient flows in discrete time, highlighting how step size and smoothness govern stability.

Stochastic approximation theory provides the probabilistic backbone of the analysis. The Robbins Monro algorithm formalizes iterative updates driven by noisy gradient estimates (Robbins and Monro, 1951). Convergence proofs rely on diminishing step sizes and martingale difference noise conditions. Chung and later Kushner and Yin develop refined results on almost sure convergence and asymptotic normality (Chung, 1954; Kushner and Yin, 2003). In our synthesis, stochastic gradient descent is interpreted as a special case of stochastic approximation applied to finite sum objectives.

The methodological challenge lies in bridging classical asymptotic results with finite time convergence guarantees

relevant for modern machine learning. Bottou et al. emphasize non asymptotic complexity bounds and practical step size strategies (Bottou et al., 2018). Gower et al. provide general analyses of SGD, deriving improved rates under interpolation and structured noise conditions (Gower et al., 2019). Stich presents a unified optimal analysis of stochastic gradient methods, clarifying the role of smoothness and variance (Stich, 2019). By synthesizing these works, we construct a finite time probabilistic framework that captures both convergence speed and variance effects.

Variance reduction techniques such as those studied by Reddi et al. and Lei et al. aim to mitigate stochastic noise by periodically incorporating full gradient information (Reddi et al., 2016; Lei et al., 2017). The methodological integration here consists of interpreting variance reduction as modifying the effective noise covariance in the stochastic approximation recursion. This perspective allows direct comparison with classical convergence conditions.

Momentum and acceleration are treated through the lens of dynamical systems. Ghadimi and Lan analyze accelerated methods in nonconvex optimization, establishing convergence to stationary points under appropriate step sizes (Ghadimi and Lan, 2016). Cutkosky and Orabona study momentum based variance reduction, demonstrating improvements in nonconvex SGD (Cutkosky and Orabona, 2019). Rather than treating these methods as distinct algorithms, we analyze them as second order discretizations of gradient flow with memory, emphasizing their influence on stability and noise amplification.

A central methodological component involves mean field and neural tangent kernel analyses. Jacot et al. show that in the infinite width limit, training dynamics converge to linearized kernel regression governed by the neural tangent kernel (Jacot et al., 2018). Mei et al. and Soltanolkotabi et al. analyze the landscape of shallow networks in mean field regimes, identifying conditions under which global minima are reachable (Mei et al., 2018; Soltanolkotabi et al., 2018). Chizat and Bach employ optimal transport arguments to prove global convergence of gradient descent in overparameterized models (Chizat and Bach, 2018).

Methodologically, these results are unified by interpreting parameter distributions as probability measures evolving under gradient flow. In infinite width limits, the empirical distribution of parameters approximates a deterministic measure valued differential equation. This measure perspective reveals convexity in function space even when parameter space is nonconvex.

Finally, we incorporate Lojasiewicz gradient inequalities and their connection to generalization. Liu et al. demonstrate how the Lojasiewicz property yields convergence rates and links to generalization bounds (Liu et al., 2022). By interpreting optimization trajectories through geometric inequalities, we connect dynamic convergence behavior to statistical performance.

Throughout the methodology, no formulas are presented explicitly. Instead, convergence rates, inequalities, and probabilistic conditions are described in detailed conceptual terms. The objective is not to reproduce proofs verbatim but to integrate their logical structures into a unified explanatory narrative.

3. Results

The theoretical synthesis developed in this article yields several major results regarding gradient based optimization in machine learning.

First, deterministic gradient descent exhibits linear convergence under the Polyak Lojasiewicz condition even in the absence of convexity. This result extends classical convex guarantees and clarifies why certain overparameterized networks converge rapidly once they enter interpolation regimes (Karimi et al., 2016). The presence of zero training error often implies local PL behavior around global minima, ensuring geometric decay of training loss.

Second, stochastic gradient descent inherits convergence guarantees from stochastic approximation theory. Under diminishing step sizes, almost sure convergence to stationary points is established for broad classes of nonconvex functions (Mertikopoulos et al., 2020; Kushner and Yin, 2003). Finite time analyses demonstrate that expected gradient norms decrease at rates inversely proportional to the number of iterations, with constants depending on noise variance (Bottou et al., 2018; Stich, 2019). These results collectively explain why SGD remains stable despite stochastic fluctuations.

Third, variance reduction techniques significantly improve convergence in finite sum settings. By reducing gradient noise, methods analyzed by Reddi et al. and Lei et al. achieve faster decay of expected stationarity measures (Reddi et al., 2016; Lei et al., 2017). This confirms that controlling variance rather than eliminating stochasticity is key to efficiency.

Fourth, large minibatch training can achieve comparable accuracy to small batch SGD when learning rates are scaled

appropriately, as demonstrated in large scale image training (Goyal et al., 2017). Theoretical insights from Gower et al. show that interpolation properties allow larger step sizes without sacrificing convergence (Gower et al., 2021). These findings align with the observation that overparameterized models exhibit reduced gradient variance near minima.

Fifth, mean field and neural tangent kernel analyses reveal that in infinite width limits, gradient descent dynamics converge to globally optimal solutions under mild assumptions (Jacot et al., 2018; Mei et al., 2018; Pham and Nguyen, 2021). The landscape effectively becomes convex in function space. Chizat and Bach further demonstrate global convergence using optimal transport arguments (Chizat and Bach, 2018). This provides a structural explanation for the empirical absence of poor local minima in wide networks.

Sixth, the Lojasiewicz gradient inequality provides a unifying geometric principle connecting optimization rates and generalization. When this inequality holds, gradient descent trajectories exhibit predictable decay behavior, which in turn bounds the complexity of the learned model (Liu et al., 2022). This bridges optimization and statistical learning theory.

Collectively, these results demonstrate that gradient based methods succeed not by accident but due to deep geometric and probabilistic structures inherent in modern machine learning models.

4. Discussion

The integration of classical optimization, stochastic approximation, and overparameterized neural network theory reveals a coherent conceptual landscape. At its core lies the principle that geometry and variance control determine convergence behavior. Strong convexity guarantees contraction. The Polyak Lojasiewicz condition generalizes this contraction without requiring convexity. Interpolation properties reduce stochastic variance. Infinite width limits induce convexity in function space.

One important limitation of existing theory is its reliance on asymptotic regimes, such as infinite width or diminishing step sizes. Real networks are finite and trained with constant step sizes. Bridging this gap remains an open challenge. Additionally, while global convergence has been established for shallow networks under certain assumptions (Soltanolkotabi et al., 2018; Pham and Nguyen, 2021), deep multilayer architectures pose additional complexity.

Another limitation concerns generalization. While optimization dynamics influence statistical performance, a complete theory linking stochastic gradient noise, implicit bias, and generalization remains incomplete. Path normalized optimization such as Path SGD suggests that geometry induced by parameterization affects implicit regularization (Neyshabur et al., 2015). Further integration of such ideas with stochastic approximation theory could yield deeper insights.

Future research directions include refining finite width analyses, exploring non Euclidean geometries, and developing unified frameworks that incorporate adaptive methods such as Adagrad, whose sharp convergence over nonconvex landscapes has been established (Ward et al., 2020). Understanding how adaptive step sizes interact with overparameterization and PL geometry is especially promising.

5. Conclusion

This article has developed a comprehensive theoretical synthesis of gradient based optimization in convex, nonconvex, and overparameterized regimes. By integrating classical deterministic gradient descent, stochastic approximation, variance reduction, acceleration, mean field limits, and geometric inequalities, we have shown that modern deep learning optimization is grounded in robust mathematical principles. Convergence arises from a combination of smoothness, interpolation, variance control, and geometric regularity. Overparameterization, far from being a mere empirical artifact, reshapes the loss landscape in ways that favor global convergence. The unified perspective presented here provides a foundation for future advances in optimization theory and its application to deep learning.

References

1. Bertsekas, D. P. Incremental gradient, subgradient, and proximal methods for convex optimization. In Optimization for Machine Learning, 85 to 115, 2012.
2. Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large scale machine learning. SIAM Review, 60(2):223 to 311, 2018.
3. Chatterjee, S. Convergence of gradient descent for deep neural networks. arXiv:2203.16462, 2022.
4. Chizat, L., and Bach, F. On the global convergence of gradient descent for over parameterized models using optimal transport. Advances in Neural Information Processing Systems, 31:3036 to 3046, 2018.
5. Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. Advances in Neural

Information Processing Systems, 32:2914 to 2924, 2019.

6. Chung, K. L. On a stochastic approximation method. Annals of Mathematical Statistics, 25(3):463 to 483, 1954.
7. Cutkosky, A., and Orabona, F. Momentum based variance reduction in non convex SGD. Advances in Neural Information Processing Systems, 32:15157 to 15166, 2019.
8. De Sa, C., Kale, S., Lee, J. D., Sekhari, A., and Sridharan, K. From gradient flow on population loss to learning with stochastic gradient descent. Advances in Neural Information Processing Systems, 35:30963 to 30976, 2022.
9. Duchi, J. C., and Ruan, F. Stochastic methods for composite and weakly convex optimization problems. SIAM Journal on Optimization, 28(4):3229 to 3259, 2018.
10. Fehrman, B., Gess, B., and Jentzen, A. Convergence rates for the stochastic gradient descent method for non convex objective functions. Journal of Machine Learning Research, 21(1):5354 to 5401, 2020.
11. Ghadimi, S., and Lan, G. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Mathematical Programming, 156(1 to 2):59 to 99, 2016.
12. Gower, R., Sebbouh, O., and Loizou, N. SGD for structured nonconvex functions: Learning rates, minibatching and interpolation. International Conference on Artificial Intelligence and Statistics, PMLR, 1315 to 1323, 2021.
13. Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtarik, P. SGD: General analysis and improved rates. International Conference on Machine Learning, PMLR, 5200 to 5209, 2019.
14. Goyal, P., Dollar, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch SGD: Training imagenet in 1 hour. arXiv:1706.02677, 2017.
15. Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. Advances in Neural Information Processing Systems, 31:8571 to 8580, 2018.
16. Jentzen, A., and Riekert, A. On the existence of global minima and convergence analyses for gradient descent methods in the training of deep neural networks. arXiv:2112.09684, 2021.
17. Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal gradient methods under the Polyak Lojasiewicz condition. Machine Learning and Knowledge Discovery in Databases, 9851:795 to 811, 2016.
18. Kushner, H. J., and Yin, G. Stochastic Approximation and Recursive Algorithms and Applications. Springer, 2003.
19. Lei, L., Ju, C., Chen, J., and Jordan, M. I. Non convex finite sum optimization via SCSG methods. Advances in Neural Information Processing Systems, 30:2349 to 2359, 2017.
20. Liu, C., Zhu, L., and Belkin, M. Loss landscapes and optimization in over parameterized non linear systems and neural networks. Applied and Computational Harmonic Analysis, 59:85 to 116, 2022.
21. Liu, F., Yang, H., Hayou, S., and Li, Q. From optimization dynamics to generalization bounds via Lojasiewicz gradient inequality. Transactions on Machine Learning Research, 2022.
22. Mei, S., Montanari, A., and Nguyen, P. M. A mean field view of the landscape of two layer neural networks. Proceedings of the National Academy of Sciences USA, 115(33):E7665 to E7671, 2018.
23. Mertikopoulos, P., Hallak, N., Kavis, A., and Cevher, V. On the almost sure convergence of stochastic gradient descent in non convex problems. Advances in Neural Information Processing Systems, 33:1117 to 1128, 2020.
24. Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. SIAM Journal on Optimization, 19(4):1574 to 1609, 2009.
25. Nesterov, Y. Introductory Lectures on Convex Optimization: A Basic Course. Springer, 2003.
26. Neyshabur, B., Salakhutdinov, R. R., and Srebro, N. Path SGD: Path normalized optimization in deep neural networks. Advances in Neural Information Processing Systems, 28:2422 to 2430, 2015.
27. Pham, H. T., and Nguyen, P. M. Global convergence of three layer neural networks in the mean field regime. International Conference on Learning Representations, 2021.
28. Polyak, B. T. Introduction to Optimization. Optimization Software Publications Division, 1987.
29. Reddi, S. J., Hefny, A., Sra, S., Poczos, B., and Smola, A. Stochastic variance reduction for nonconvex optimization. International Conference on Machine Learning, PMLR, 314 to 323, 2016.
30. Robbins, H., and Monro, S. A stochastic approximation method. Annals of Mathematical Statistics, 22(3):400 to 407, 1951.
31. Soltanolkotabi, M., Javanmard, A., and Lee, J. D.

Theoretical insights into the optimization landscape of over parameterized shallow neural networks. IEEE Transactions on Information Theory, 65(2):742 to 769, 2018.

32. Stich, S. U. Unified optimal analysis of the stochastic gradient method. arXiv:1907.04232, 2019.
33. Ward, R., Wu, X., and Bottou, L. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. Journal of Machine Learning Research, 21(1):9047 to 9076, 2020.