Grassmannian Optimization Drives Generalization in Overparameterized DNN

Changfeng C. Wang CHARLESCFWANG@GMAIL.COM
BDS | ApeiroM.ai, Boston, MA

Abstract

We present an overview of a geometric theory explaining why and how heavily overparameterized deep neural networks generalize despite being able to perfectly fit random labels. The key insight is that, contrary to the uniform hypothesis-class assumptions of classical statistical learning theory, deep learning admits an iso-loss–induced *fiber bundle structure* shaped jointly by the loss function, hypothesis class, and data distribution. Gradient-based optimization follows horizontal lifts across low-dimensional subspaces in the Grassmannian $\mathrm{Gr}(r, p)$, where $r \ll p$ is the rank of the Hessian at the optimum. The low-dimensional subspace is selected by random initialization near the origin and shaped by the data and the local trivialization structure. This yields: (i) a mechanistic explanation—effective complexity is r, not the ambient dimension p, because the $(p - r)$-dimensional fibers $F = \ker(H)$ are statistically inert; (ii) a unifying geometric framework for flat minima, PAC-Bayes, NTK, double descent, and implicit algorithmic regularization; and (iii) a closed-form finite-sample generalization gap equation together with a bias–variance decomposition (a theoretical scaling law) of the generalization error. Empirical evaluations of the gap equation achieve $> 90\%$ predictive accuracy, improving upon VC, PAC-Bayesian, and spectral bounds by orders of magnitude. The equation resolves the long-standing open problem of explaining generalization in gradient-trained overparameterized DNNs [12, 33]. The degeneracy of the Hessian post-training is thus a hallmark of generalization, rather than an empirical curiosity.

The framework provides a practical path for transforming current trial-and-error deep learning practice—especially for large models—into principled design and engineering. Practical translation to large-scale models requires computational innovations that we identify as key collaborative directions.[1]

1. Introduction

Deep neural networks (DNNs) are often massively overparameterized, with many more parameters than training examples. Overparameterization reshapes the optimization landscape: empirical loss surfaces exhibit large basins without spurious local minima and are locally well-behaved, enabling stochastic gradient descent (SGD) to converge reliably [13, 16]. Combined with universal approximation power [6, 10, 24, 31], such networks achieve state-of-the-art performance across diverse tasks.

The central puzzle is generalization: DNNs can perfectly memorize random labels [32], yet generalize well on structured data. This challenges the classical view that generalization is achieved by *externally* constraining capacity using VC dimension, norm constraints, or regularization [27]. It remains an open problem to explain *why and how* overparameterized DNNs trained with gradient-

1. This paper presents Part 1 of a broader framework. Part 2 (Optimization Dynamics) develops how hyperparameters map to generalization through the same geometry. Complete proofs appear in the full version [28].

based methods generalize; in particular, to obtain non-vacuous, data-dependent generalization guarantees in regimes where $p \gg n$ [12].

Classical theory treats the hypothesis class $\mathcal{F}$ as a uniform structure characterized by capacity measures (VC dimension, Rademacher complexity). We reveal that $\mathcal{F}$ has rich internal geometry: it decomposes into a fiber bundle (Θ, B, π, F). Around a minimizer θ^*, the Hessian $H(\theta^*)$ has rank $r \ll p$ and decomposes parameter space into an identifiable subspace $S(\theta^*) = \mathrm{range}(H)$ and an iso-loss fiber $F(\theta^*) = \ker(H)$. Random initialization near the origin favors low-dimension representation, and gradient-based algorithms naturally traverse a low-dimensional subspace around $S(\boldsymbol{\theta}^*) \subseteq \mathrm{Gr}(r, p)$ where the optimal fiber resides. This explains why externally imposed capacity control is often unnecessary—optimization itself imposes geometric constraints via confinement to data-determined subspaces.

The structure yields a data-dependent generalization gap equation driven by the gradient covariance and Hessian pseudoinverse, providing a closed-form solution to the gap estimation problem for overparameterized DNNs and a unified framework behind flat minima, implicit regularization, NTK theory, and double descent. Conceptually, the framework shifts the traditional philosophy of generalization via hypothesis-class capacity control by an *external designer* to an *optimization-driven* view in which generalization emerges from data-guided optimization that adapts to low-dimensional Grassmannians within the high capacity afforded by the global structure of the hypothesis class.

The framework provides a way to translate the "alchemy"-style trial-and-error approach into principled design and training of deep networks, leading to improved efficiency, performance, and economy, particularly for large-scale models (more details in [28]—a key direction for collaborative research).

2. Related Work

Generalization in DNNs has been studied from both algorithm-agnostic and algorithm-aware perspectives; see Kawaguchi et al. [12] for a survey of open problems.

Algorithm-agnostic approaches. Classical theory uses complexity measures (VC dimension, Rademacher complexity, covering numbers [3, 27]) or matrix norms [4, 20] to bound the generalization gap. In heavily overparameterized regimes these bounds typically scale with p and become vacuous [11]. PAC-Bayesian approaches [7, 18, 19] provide tighter, data-dependent bounds but rely on carefully chosen priors. Compression-based [1] and information-theoretic [30] approaches face similar limitations when $p \gg n$. Benign overfitting has been analyzed in linear models [2, 8], but extensions to deep nonlinear models are incomplete.

Algorithm-aware approaches. Work on implicit regularization shows GD/SGD can converge to margin-maximizing or minimum-norm solutions in simple settings [17, 22], while SGD noise can further aid generalization [26]. Double descent [5] reveals that test error can decrease again beyond the interpolation threshold, complicating classical capacity-based intuition. Neural Tangent Kernel (NTK) theory [**?**] analyzes infinitely wide networks via linearization at initialization, equivalent to holding the Hessian fixed. Empirical work has highlighted near-degenerate Hessians [21, 23, 25] and connections between flatness and generalization [9, 15, 16, 29]. Architecture-specific bounds [14] improve constants but still do not fully explain the role of optimization.

Summary. Existing approaches are either vacuous in modern overparameterized regimes or narrowly tailored to special cases. None provide non-vacuous, data-dependent generalization guarantees that explain *how* gradient-based algorithms achieve strong generalization when $p \gg n$, as high-

lighted by Zhang et al. [33]. Our fiber-bundle view is *both* geometric and data- and algorithm-aware: it shows that with proper configuration, GD/SGD is effectively constrained to an r-dimensional Grassmannian subspace determined by the Hessian, and derives a closed-form finite-sample gap equation scaling as r/n with tight experimental validation.

3. Preliminaries

Let $f(\mathbf{x}, \boldsymbol{\theta})$ denote a DNN with parameters $\boldsymbol{\theta} \in \mathbb{R}^p$ and inputs $\mathbf{x} \in \mathbb{R}^d$. Given i.i.d. training data $\mathcal{D} = \{(\mathbf{x}_i, y_i = T(x_i))\}_{i=1}^n$ drawn from distribution $P(\mathbf{x}, y)$ for some function $T(\cdot)$, define empirical and population risks $L_n(\boldsymbol{\theta}) := \mathbb{E}_{P_n}\ell(y, f(\mathbf{x}, \boldsymbol{\theta}))$ and $L(\boldsymbol{\theta}) := \mathbb{E}_P\ell(y, f(\mathbf{x}, \boldsymbol{\theta}))$, where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function. Let $\hat{\boldsymbol{\theta}} \in \operatorname{argmin}_{\boldsymbol{\theta}} L_n(\boldsymbol{\theta})$ and $\boldsymbol{\theta}^* \in \operatorname{argmin}_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ denote empirical and population minimizers. We study the generalization gap $L(\hat{\boldsymbol{\theta}}) - L_n(\hat{\boldsymbol{\theta}})$ in the overparameterized regime $p \gg n$, for $\hat{\boldsymbol{\theta}}$ obtained by gradient-based optimization.

Assumptions

We work under standard regularity conditions ensuring differentiability and controlled curvature:

1. **Smoothness:** $\ell(y, f)$ is C^2 in f, and $f(\mathbf{x}, \boldsymbol{\theta})$ is twice differentiable in $\boldsymbol{\theta}$ almost everywhere on Θ.
2. **Lipschitz gradients:** $\nabla_{\boldsymbol{\theta}}\ell$ and $\nabla_{\boldsymbol{\theta}} f$ are Lipschitz continuous on Θ.
3. **Boundedness:** $\Theta \subset \mathbb{R}^p$ is compact, inputs $\mathbf{x}$ are bounded, and y has finite variance.

These conditions (together with standard dominated-convergence assumptions) ensure that $L_n(\boldsymbol{\theta})$ and $L(\boldsymbol{\theta})$ are twice differentiable almost everywhere with well-controlled Hessians.

4. Fiber Bundle Structure and Grassmannian Geometry

Define the hypothesis class $\mathcal{F} = \{f(\cdot; \theta) : \theta \in \Theta\}$. The loss $L(\boldsymbol{\theta})$ defines iso-loss manifolds $\mathcal{M}(l) := \{\boldsymbol{\theta} \in \Theta : L(\boldsymbol{\theta}) = l\}$ for $l \in [0, C]$. Define loss-based equivalence $\boldsymbol{\theta} \sim \boldsymbol{\theta}'$ iff $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}')$. The quotient $B := \mathcal{F}/\sim$ is the base space of loss-equivalence classes, and the canonical projection $\pi : \Theta \to B$ maps $\boldsymbol{\theta}$ to its class. Fibers $F_b := \pi^{-1}(b)$ are iso-loss manifolds, yielding a fiber bundle (Θ, B, π, F) over B with typical fiber F. Under our main assumptions 3, the fiber bundle structure is Morse–Bott.

At a minimizer $\boldsymbol{\theta}^*$, the Hessian $H(\boldsymbol{\theta}^*) = \nabla^2 L(\boldsymbol{\theta}^*)$ has rank $r \ll p$ and splits parameter space as $\boldsymbol{\theta} - \boldsymbol{\theta}^* = \delta_S + \delta_F$, where $\delta_S \in S(\boldsymbol{\theta}^*) := \operatorname{range}(H(\boldsymbol{\theta}^*))$ and $\delta_F \in F(\boldsymbol{\theta}^*) := \ker(H(\boldsymbol{\theta}^*))$. A Taylor expansion gives $L(\boldsymbol{\theta}^* + \delta) = L(\boldsymbol{\theta}^*) + \frac{1}{2}\delta_S^\top H(\boldsymbol{\theta}^*)\delta_S + o(\|\delta\|^2)$, so only δ_S affects loss at second order; fiber directions δ_F are iso-loss. Each $S(\boldsymbol{\theta}^*)$ corresponds to a point in the Grassmannian $\operatorname{Gr}(r, p)$, and we define the *generalized Grassmannian* $\mathcal{G}_H(r, p) := \bigcup_{\boldsymbol{\theta}^*} S(\boldsymbol{\theta}^*) \subseteq \operatorname{Gr}(r, p)$, the union of all r-dimensional critical subspaces.

Gradient-based algorithms are (approximately) orthogonal to fiber directions. Near $\boldsymbol{\theta}^*$, they move along directions in $S(\boldsymbol{\theta}^*)$ because $\nabla L(\boldsymbol{\theta}) \in S(\boldsymbol{\theta}^*)$ (up to second order), while motion in $F(\boldsymbol{\theta}^*)$ is neutral for the loss. Thus optimization follows *horizontal lifts* across the Grassmannian base, never exploring most of the $(p - r)$-dimensional fiber directions, even though they exist in parameter space.

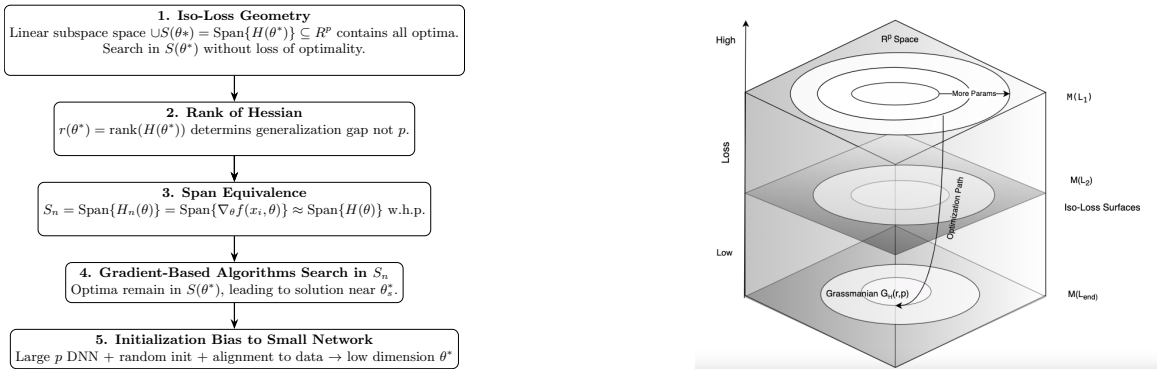

Figure 1: **Left:** How strong generalization emerges from optimization: gradient-based algorithms find generalizing minima by following horizontal lifts across space in low-dimensional Grassmannians affine to $S(\boldsymbol{\theta}^*) \in \mathrm{Gr}(r, p)$, while fibers $F(\boldsymbol{\theta}^*) = \ker(H)$ remain inert toward convergence. **Right:** The fiber bundle structure (Θ, B, π, F) of deep learning: iso-loss manifolds $\mathcal{M}(l)$ form fibers over the base B. The Hessian decomposes parameter space into identifiable directions $S(\boldsymbol{\theta}^*)$ (curved surface) and iso-loss directions $F(\boldsymbol{\theta}^*)$ (vertical lines), and optimization traverses fibers toward an optimal fiber.

Overview of the Geometric Theory

Combining this geometry with tools from empirical process theory and high-dimensional probability yields the following picture (formal statements are given in [28]):

- **Subspace sufficiency:** Any minimizer $\boldsymbol{\theta}^*$ has an equivalent minimizer lying in a low-dimensional $S(\boldsymbol{\theta}^*) \in \mathcal{G}_H(r, p)$, reducing effective search to exponentially smaller subspaces determined by the data distribution.
- **Horizontal confinement:** Following early exploration, GD/SGD dynamics toward the interpolation phase are confined to $S(\boldsymbol{\theta}^*)$: gradients live in $S(\boldsymbol{\theta}^*)$ and optimization follows horizontal lifts across the Grassmannian base rather than exploring fiber directions $F(\boldsymbol{\theta}^*)$.
- **Data-dependent generalization:** The generalization gap depends only on gradient covariance and Hessian structure restricted to $S(\boldsymbol{\theta}^*)$, with effective dimension $r = \mathrm{rank}(H(\boldsymbol{\theta}^*))$ rather than p.
- **Low-dimensional bias:** Isotropic initialization (e.g., He or Xavier) induces an ε^r volume concentration toward low-rank fibers in $\mathcal{G}_H(r, p)$, explaining why optimization reliably finds good minima despite enormous ambient dimension. This bias toward low dimension is a geometric consequence via the Morse–Bott tube theorem.

Taken together, these points formalize an optimization-driven capacity-control mechanism: effective complexity is set by the Hessian rank r and the geometry of $\mathcal{G}_H(r, p)$, rather than by the ambient parameter count p or other global complexity measures of $\mathcal{F}$.

5. Main Theoretical Result

Let $\Sigma_n(\boldsymbol{\theta})$ denote the empirical gradient covariance, $H_n(\boldsymbol{\theta})$ the empirical Hessian, and $H_n^+(\boldsymbol{\theta})$ its Moore–Penrose pseudoinverse. We write $Z_n := L(\boldsymbol{\theta}^*) - L_n(\boldsymbol{\theta}^*)$ for the centered loss fluctuation.

The following theorem summarizes our main result; a detailed version with proofs appears in the supplement.

Theorem *[Generalization gap scales with Hessian rank r] Under the assumptions in Section 3, let $r = \operatorname{rank}(H(\boldsymbol{\theta}^*))$ and let $\lambda_r > 0$ be the smallest positive eigenvalue of $H(\boldsymbol{\theta}^*)$. Suppose $\|H(\boldsymbol{\theta})\| \leq M$ in a neighborhood of $\boldsymbol{\theta}^*$ and let C_L be a Lipschitz constant for ∇L. Define $C_R := MC_L/\lambda_r^2$ and $Z_n := L(\boldsymbol{\theta}^*) - L_n(\boldsymbol{\theta}^*)$. Then, for sufficiently large n, with probability at least $1-\delta$,*

$$L(\hat{\boldsymbol{\theta}}) = L_n(\hat{\boldsymbol{\theta}}) + Z_n + \frac{1}{n} tr\big(\Sigma_n(\boldsymbol{\theta}^*) H_n(\boldsymbol{\theta}^*)^+\big) + C_R\Big(\frac{r\log n}{n}\Big)^{3/2}. \tag{1}$$

Moreover, Z_n concentrates and satisfies

$$P\left(L(\hat{\boldsymbol{\theta}}) \leq L_n(\hat{\boldsymbol{\theta}}) + M\sqrt{\frac{\ln(4/\delta)}{2n}} + \frac{1}{n} tr\big(\Sigma_n(\boldsymbol{\theta}^*) H_n(\boldsymbol{\theta}^*)^+\big) + C_R\Big(\frac{r\log n}{n}\Big)^{3/2}\right) \geq 1-\delta, \tag{2}$$

and

$$\sqrt{n}\, Z_n \xrightarrow{d} \mathcal{N}(0, \sigma_l^2), \quad \operatorname{Var}(Z_n) = \frac{\sigma_l^2}{n}, \quad \sigma_l^2 = \operatorname{Var}\big[\ell(y, f(x, \boldsymbol{\theta}^*))\big]. \tag{3}$$

In particular, the loss admits the bias–variance-type expansion

$$L(\hat{\boldsymbol{\theta}}) = L(\boldsymbol{\theta}^*) + \frac{1}{2n} tr\big(\Sigma_n(\boldsymbol{\theta}^*) H_n(\boldsymbol{\theta}^*)^+\big) + C_R\Big(\frac{r\log n}{n}\Big)^{3/2}, \tag{4}$$

and the parameters satisfy

$$\sqrt{n}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\big) \xrightarrow{d} \mathcal{N}\big(0,\, H(\boldsymbol{\theta}^*)^+ S H(\boldsymbol{\theta}^*)^+\big), \quad S = Cov\big(\nabla_{\boldsymbol{\theta}} \ell(y, f(x, \boldsymbol{\theta}^*))\big). \tag{5}$$

Remark *[Interpretation] The dominant term in the generalization gap is $\frac{1}{n} tr(\Sigma_n H_n^+)$, which depends only on the r-dimensional identifiable subspace $S(\boldsymbol{\theta}^*)$; fiber directions in $F(\boldsymbol{\theta}^*) = \ker(H)$ are projected out by H_n^+. For L_2 loss, this trace reduces to $\sigma^2 r/n$, making explicit the r/n scaling and connecting directly to benign overfitting results in linear models. The $(p-r)$ flat directions create large fibers but do not contribute to the gap. This makes precise the heuristic link between flat minima and generalization and explains why classical bounds scaling with p are vacuous when $p \gg n$ but $r \ll n$.*

6. Experiments

We empirically validate the gap equation and its r/n scaling on teacher–student setups (FNNs and Transformers) under varying architectures and loss functions (cross-entropy and L_2), including strongly overparameterized regimes with $p \gg n$.

Setup. Synthetic data are generated by a compact teacher network and split into train/validation/test sets. Overparameterized student networks of varying depth and width are trained with SGD+momentum (100 epochs, 5 random seeds per configuration). For each trained model we estimate $\operatorname{tr}(\Sigma_n H_n^+)$ and evaluate several practical approximations:

1. **Trace:** direct or Hutchinson/Lanczos-based estimation of $\frac{1}{n}\operatorname{tr}(\Sigma_n H_n^+)$;
2. **Sig-rank:** $\sigma_n^2 \frac{r}{n}$ for L_2 loss, where σ_n^2 is the empirical noise variance;
3. **Gnorm:** a gradient-based Monte Carlo estimator;

4. **Gnorm_rp:** a random-projection variant that approximates both curvature and gradient covariance.

We compare these with VC, spectral, and PAC-Bayesian bounds and summarize performance using Value $= \big((L_{\text{train}} - L_{\text{test}}) - \widehat{\text{Gap}}\big)/L_{\text{test}}$, so values near zero indicate tight prediction of the generalization gap.

Results. Table 1 reports min/max/average/variance of this value across configurations. The trace-based and Gnorm-based estimators nearly close the gap (within 10% error), and Gnorm_rp consistently yields the tightest approximation ($> 90\%$ of test error explained), dramatically outperforming VC, spectral, and PAC-Bayesian bounds.

Table 1: Empirical validation with cross-entropy (CE) and L_2 losses. Values closer to zero indicate tighter prediction of the generalization gap.

Statistic	Sig-rank	Trace	Gnorm	Gnorm_rp	VC	Spectral	PacBayesian
				CE Dataset			
Min	-0.0481	-0.1457	0.0842	-0.0100	0.6068	4.3914×10^4	1.2555
Max	-0.0044	-0.0088	0.1290	-0.0025	1.0707	6.3540×10^5	1.3476
Average	-0.0305	-0.0479	0.1121	-0.0055	0.9122	2.2692×10^5	1.2861
Variance	0.0003	0.0035	0.0003	1.0×10^{-5}	0.0369	6.9198×10^{10}	0.0016
				L_2 Dataset			
Min	0.0441	-3.8936	0.2405	-0.0241	2.2815	5.2081×10^5	1.8414
Max	0.1904	0.0023	0.3623	-0.0053	4.0249	1.0310×10^6	2.2957
Average	0.1162	-1.0391	0.2937	-0.0137	3.3365	7.4040×10^5	2.0832
Variance	0.0032	2.7275	0.0027	6.8×10^{-5}	0.4226	3.6694×10^{10}	0.0395

Notably, in regimes where $p \gg n$ but $r \ll n$, classical bounds scaling with p are vacuous, while our $O(r/n)$ theory remains accurate and predictive, consistent with the fiber-bundle explanation that only the r-dimensional base $S(\boldsymbol{\theta}^*)$ contributes to the generalization gap and the $(p-r)$-dimensional fibers are statistically inert.

7. Conclusion

We presented a geometric theory of generalization in overparameterized DNNs based on fiber bundle structure. Generalization emerges because gradient-based optimization is constrained to low-dimensional Grassmannian bases determined by the Hessian. This yields a closed-form gap equation 1, along with a theoretical scaling law, both scaling with r/n rather than p/n, achieving $> 90\%$ predictive accuracy and orders-of-magnitude improvement over classical bounds. The framework provides mechanistic understanding of implicit regularization, unifies flat minima/NTK/double descent phenomena, and addresses the open problem of explaining generalization in gradient-trained overparameterized networks [12, 33].

Our analysis relies on standard smoothness and boundedness assumptions that may be relaxed in future work. Part 2 of this framework extends the geometric viewpoint to the full dynamics of learning, relating initialization, learning schedules, and hyperparameter choices directly to generalization through the same fiber bundle and Grassmannian structure. Taken together, these results help move DNN practice from "alchemy" toward a principled science of deep learning.

References

[1] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang. Stronger generalization bounds for deep nets via a compression approach. *International Conference on Machine Learning*, pages 254–263, 2018.

[2] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020. doi: 10.1073/pnas.1907378117.

[3] P.L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[4] P.L. Bartlett, D.J. Foster, and M. Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30:6240–6249, 2017.

[5] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116.

[6] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989.

[7] G.K. Dziugaite and D.M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks. *International Conference on Machine Learning*, pages 1256–1265, 2017.

[8] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 48(4):1836–1863, 2020.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997. URL https://www.bioinf.jku.at/publications/older/3304.pdf.

[10] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2): 251–257, 1991.

[11] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 12 2019. URL https://arxiv.org/abs/1912.02178.

[12] K. Kawaguchi, Y. Bengio, and L. Kaelbling. Generalization in deep learning. page 112–148, 2022. doi: 10.1017/9781009025096.003. URL http://dx.doi.org/10.1017/9781009025096.003.

[13] Kenji Kawaguchi. Deep learning without poor local minima. *Advances in Neural Information Processing Systems*, 29:586–594, 2016.

[14] Kenji Kawaguchi, Prateek Kothari, Praneeth Netrapalli, Sujay Sanghavi, and Yuki Sugiyama. Generalization bounds for deep neural networks: A review and new results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9732–9748, 2023.

[15] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016. URL https://arxiv.org/abs/1609.04836.

[16] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in Neural Information Processing Systems*, 31:6389–6399, 2018.

[17] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019. URL https://arxiv.org/abs/1906.05890.

[18] D.A. McAllester. Pac-bayesian model averaging. *Conference on Computational Learning Theory*, pages 164–170, 1999.

[19] B. Neyshabur, S. Bhojanapalli, and N. Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *International Conference on Learning Representations*, 2018.

[20] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. *Conference on Learning Theory*, 65:1376–1401, 2017.

[21] Vardan Papyan. The full spectrum of deepnet hessians at scale. *arXiv:1811.07062*, 2018. URL https://arxiv.org/abs/1811.07062.

[22] Tomaso Poggio, Qianli Liao, and Andrzej Banburski. Complexity control by gradient descent in deep networks. *Nature Communications*, 11(1):1027, 2020. doi: 10.1038/s41467-020-14663-9.

[23] Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Léon Bottou. Empirical analysis of the hessian of deep networks. *arXiv:1706.04454*, 2017. URL https://arxiv.org/abs/1706.04454.

[24] Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of relu neural networks in sobolev spaces. *arXiv preprint arXiv:2012.07808*, 2020. Demonstrates near-optimal approximation rates for deep ReLU networks with Sobolev $W^{k,2}$ target functions and L_2-loss. Rate: $L(\theta^*) \leq C \cdot N^{-\frac{2k}{2k+d}} \cdot (\log N)^c$.

[25] Bernhard Schölkopf Sidak Pal Singh, Thomas Hofmann. The hessian perspective into the nature of convolutional neural networks. *https://arxiv.org/abs/2305.09088*, 2023.

[26] Samuel L. Smith and Quoc V. Le. A bayesian perspective on generalization and stochastic gradient descent. *arXiv preprint arXiv:1710.06451*, 2017. URL https://arxiv.org/abs/1710.06451.

[27] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[28] Changfeng Wang. The fiber bundle structure of deep learning. *Preprint*, 2025.

[29] Lei Wu, Zhanxing Zhu, and Weinan E. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv:1706.10239*, 2017. URL https://arxiv.org/abs/1706.10239.

[30] A. Xu and M. Raginsky. Information-theoretic analysis of generalization capability. *International Symposium on Information Theory*, pages 1356–1360, 2017.

[31] Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94: 103–114, 2017. Establishes approximation rates for deep ReLU networks with Sobolev $W^{k,\infty}$ target functions and L_∞-loss. Rate: $L(\theta^*) \leq C \cdot N^{-\frac{k}{d}}$.

[32] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[33] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3), 2021. URL https://arxiv.org/abs/2103.10438.