OPT2025: 17th Annual Workshop on Optimization for Machine Learning

Stochastic Neural Tangent Kernel: Revisiting the NTK For SGD

Bhavesh Kumar BKUMAR2@UW.EDU
Dan Mikulincer DANMIKU@UW.EDU
University of Washington, Seattle

Abstract

Stochastic Gradient Descent (SGD) is a foundational algorithm for training neural networks, valued for its efficiency and generalization enabled by intrinsic stochasticity. Existing theoretical frameworks like the Neural Tangent Kernel (NTK) describe training dynamics in the infinite-width limit but omit the stochastic effects of minibatch sampling. This paper introduces the Stochastic Neural Tangent Kernel (SNTK), an extension of the NTK that incorporates adaptive, residual-weighted noise from SGD's minibatch-induced randomness. By modeling SGD as a continuous-time stochastic differential equation projected into function space, we rigorously characterize the noise structure and derive a kernel capturing the evolving stochastic dynamics. Our formulation unifies deterministic kernel regression and stochastic optimization, demonstrating how minibatch noise directs training toward flatter minima and greater robustness. Empirical evaluations on the two moons dataset show the SNTK's improved ability to represent SGD's functional behavior compared to classical NTK methods. The aim of this work is to address a gap in our understanding of SGD's function-space dynamics and provide a tool for further studies on optimization, generalization, and the interaction between noise and representation learning in wide neural networks.

1. Introduction

Stochastic Gradient Descent (SGD) is a cornerstone algorithm for training neural networks, valued for its scalability, efficiency in high-dimensional spaces, and intrinsic stochasticity that promotes superior generalization [5, 27, 31, 32, 35]. While the Neural Tangent Kernel (NTK) provides a powerful theoretical framework for analyzing training dynamics of wide networks [1, 15, 19], it is limited to deterministic, full-batch gradient descent and ignores SGD's minibatch-induced noise, which plays a crucial role in practical generalization and robustness [9, 22, 29]. In this work, we introduce the Stochastic Neural Tangent Kernel (SNTK) in the infinite-width regime, incorporating SGD's noise into the NTK framework. Our analysis characterizes SNTK's adaptive, residual-weighted noise properties, and we derive a formulation modeling idealized SGD trajectories in function space. This framework integrates stochastic optimization with NTK theory, showing how noise can guide training dynamics toward flatter minima, thus helping to improve generalization. The SNTK also opens the door for practical applications, such as distinguishing feature learning from lazy training regimes through the interaction of noise and kernel dynamics. Experiments in Section 4 validate these insights by comparing functional outputs of SGD, SNTK, and NTK.

2. Related Work

The Neural Tangent Kernel (NTK) is a linear model for infinitely wide neural networks, showing that gradient descent in the infinite-width limit corresponds to kernel regression with a fixed kernel [1, 7, 15]. This framework provides exact characterizations of convergence and generalization for deterministic full-batch methods, revealing that wide networks evolve linearly during training [8, 17, 19]. Extensions to convolutional architectures [1, 2, 28], asymptotic studies [9, 25, 26],

and finite-width corrections [11] have expanded its scope. However, NTK inherently assumes deterministic optimization and so is ill-suited to model SGD with minibatches. The added noise in such algorithms is a key factor in generalization, as the noise is known to promote flatter minima and robustness [6, 13, 18, 34].

This limitation has sparked interest in modeling SGD's noise separately, often through stochastic differential equations (SDEs) that capture scaling rules for adaptive optimizers [20, 21, 23] and the noise's role in implicit regularization [4, 24, 33, 35]. Empirical works have linked the network width to noise tolerance [10, 30], while kernel methods examine SGD learning curves [3, 9] and Gaussian process connections, emphasizing the noise's potential impact on universality [14, 16, 37]. Despite these insights into noise's benefits, no prior work unifies SGD's adaptive, state-dependent stochasticity with NTK's function-space dynamics.

3. The Stochastic Neural Tangent Kernel Framework

We introduce the Stochastic Neural Tangent Kernel (SNTK), a mathematical framework that extends the Neural Tangent Kernel, and characterizes SGD's dynamics in function space.

To derive the SNTK, we begin by modeling SGD in continuous time. Consider a neural network $f(\cdot;\theta_t)$ with parameters $\theta_t \in \mathbb{R}^P$ trained via SGD on dataset $\{(x_i, y_i)\}_{i=1}^N$ with mean squared error loss $L(\theta) = \frac{1}{2}\sum_{i=1}^N (f(x_i;\theta) - y_i)^2$. At each step, SGD samples a minibatch of size b and updates parameters based on the minibatch gradient (see Appendix A for further details and definitions). Following [23], in the infinite-width limit we model SGD as an idealized continuous-time stochastic differential equation (SDE). This SDE serves as the foundation of our SNTK derivation and takes the form:

$$d\theta_t = -\nabla L(\theta_t)dt + \sqrt{\eta\Sigma(\theta_t)}dW_t$$

In this equation, $\eta > 0$ is the learning rate, and W_t is a standard Brownian motion. The diffusion term captures the noise from minibatch sampling through the gradient noise covariance $\Sigma(\theta_t)$:

$$\Sigma(\theta) = \frac{1}{b}\sum_{i\in\mathcal{B}_t}(\nabla L_i(\theta) - \nabla L_{\mathcal{B}_t}(\theta))(\nabla L_i(\theta) - \nabla L_{\mathcal{B}_t}(\theta))^\top.$$

This setup ensures that the noise scales as $\sigma \sim 1/\sqrt{b}$, maintaining consistent SDE dynamics regardless of batch size.

With SGD modeled as an SDE in parameter space, we now derive the SNTK by projecting these dynamics into function space under the infinite-width limit as described in Assumption 1. Prior work [15] shows that, in this limit, deterministic full-batch gradient descent corresponds to kernel regression governed by the Neural Tangent Kernel:

Definition 1 (Neural Tangent Kernel) *The Neural Tangent Kernel (NTK) is defined as*

$$K(x, x') = \langle \nabla_\theta f(x;\theta_t), \nabla_\theta f(x';\theta_t)\rangle,$$

where $\nabla_\theta f(x;\theta_t)$ denotes the gradient of the network output at input x with respect to the parameters θ_t.

To incorporate the added noise, we project the parameter-space SDE into function space. Applying Itô's lemma to the network output $f_t(x) := f(x;\theta_t)$, and leveraging the infinite-width limit where the Hessian trace term is negligible (Lemma 5) gives:

$$df_t(x) = \nabla_\theta f(x;\theta_t)^\top d\theta_t.$$

Substituting this expression into the SDE and defining the residuals as $r_i(t) = f_t(x_i) - y_i$, we get:

$$df_t(x) = -\sum_{i=1}^{N} r_i(t) K(x, x_i) dt + \sqrt{\eta}\nabla_\theta f(x;\theta_t)^\top \sqrt{\Sigma(\theta_t)} dW_t.$$

The above equation describes the network's function evolution under SGD. The drift term corresponds to deterministic NTK regression, while the diffusion term introduces randomness from minibatch sampling, governed by the gradient-noise covariance $\Sigma(\theta_t)$. To understand this noise, we characterize $\Sigma(\theta_t)$ in Theorem 1. See the relevant assumptions about width and data separation in Appendix A. Assumption 1 ensures we operate in the infinite-width NTK regime where the kernel remains approximately constant during training. Assumption 2 provides data separation via the kernel, a standard condition in NTK convergence analysis [1, 8, 19] that is essential for bounding cross-terms in our noise analysis. This separation condition is empirically satisfied for common datasets including two moons (Section 4), and holds when training points are not excessively correlated in the kernel-induced feature space.

Theorem 1 (Minibatch Noise Structure) *Under Assumptions 1 and 2, the gradient noise covariance satisfies:*

$$\Sigma(\theta_t) = \frac{1}{b}\sum_{i\in\mathcal{B}_t} r_i^2(t)\nabla_\theta f(x_i;\theta_t)\nabla_\theta f(x_i;\theta_t)^\top + \mathcal{E}(\theta_t), \tag{1}$$

where $r_i(t) = f(x_i;\theta_t) - y_i$ are the residuals. The error term $\mathcal{E}(\theta_t)$ satisfies the function-space bound: for all training points $x_m, x_n \in \{x_1, \ldots, x_N\}$,

$$\left|\nabla_\theta f(x_m;\theta_t)^\top \mathcal{E}(\theta_t)\, \nabla_\theta f(x_n;\theta_t)\right| \le \frac{C_\delta}{b}\cdot\frac{1}{b}\sum_{i\in\mathcal{B}_t} r_i^2(t) + \mathcal{O}\left(n_{\min}^{-1/2}\right), \tag{2}$$

where $C_\delta = K_{\max}^2(1 + 2(1+\sqrt{\delta})^2)$ depends on both the kernel bound $K_{\max}$ and separation parameter δ. Here $n_{\min} := \min_{\ell\in\{1,\ldots,L-1\}} n_\ell$ denotes the minimum hidden layer width.

The error bound (23) comprises two components: a batch-size-dependent term $\mathcal{O}(C_\delta/b)$ and a width-dependent term $\mathcal{O}(n_{\min}^{-1/2})$. In the infinite-width limit ($n_{\min}\to\infty$), the width-dependent component vanishes while the batch-size-dependent term persists, reflecting the fundamental discretization of minibatch sampling. Thus, $\mathcal{E}(\theta_t) = \mathcal{O}(1/b)$ in the infinite-width regime. This reveals that the noise covariance is dominated by the parameter gradients of minibatch samples, scaled by the squared residuals $r_i^2(t)$, amplifying perturbations in directions of poor predictions.

Building on this, we define the function-space noise process $dN_t(x) := \nabla_\theta f(x;\theta_t)^\top \sqrt{\Sigma(\theta_t)} dW_t$. Lemma 6 in the Appendix derives its infinitesimal covariance:

$$\mathbb{E}[dN_t(x) dN_t(x')] = \nabla_\theta f(x;\theta_t)^\top \Sigma(\theta_t) \nabla_\theta f(x';\theta_t)\, dt.$$

Substituting the leading term from Theorem 1 into this covariance yields our Stochastic Neural Tangent Kernel, a residual-weighted kernel that captures SGD's stochastic effects.

Definition 2 (Stochastic Neural Tangent Kernel) *The Stochastic Neural Tangent Kernel is defined as:*

$$K_S(x, x'; \{r_i\}) := \frac{1}{b} \sum_{i \in \mathcal{B}_t} r_i^2 K(x, x_i) K(x_i, x')$$

where b is the minibatch size, $\{r_i\}_{i \in \mathcal{B}_t}$ are the residuals of the current minibatch, and $K(\cdot, \cdot)$ is the Neural Tangent Kernel.

The SNTK exhibits several fundamental characteristics that distinguish it from classical kernel methods. It is state-dependent, explicitly incorporating current residuals r_i to enable adaptive behavior that evolves with the learning dynamics. Its quadratic scaling with prediction errors ensures that noise magnitude naturally diminishes as training progresses and accuracy improves. Moreover, as minibatch size b increases, stochastic perturbations become negligible, recovering the deterministic NTK regime in the full-batch limit.

These properties enable a precise characterization of SGD as kernel regression with adaptive noise, as formalized in our main result:

Theorem 2 (SNTK Regression Formulation) *Under the assumptions of Theorem 1, there exists a continuous process $N_t(x)$ such that*

$$df_t(x) = -\sum_{i=1}^{N} r_i(t) K(x, x_i) dt + \sqrt{\eta} dN_t(x)$$

where $N_t(x)$ is a continuous local martingale with quadratic covariation

$$\langle N(\cdot, x), N(\cdot, x') \rangle_t = \int_0^t K_S(x, x'; \{r_i(s)\}) ds$$

and K_S is the SNTK: $K_S(x, x'; \{r_i\}) = \frac{1}{b} \sum_{i \in \mathcal{B}_t} r_i^2 K(x, x_i) K(x_i, x')$.

This formulation unifies SGD's dynamics in function space, enabling the study of neural network training under stochastic optimization through a precise lens. The SNTK governs the evolution of network outputs by capturing how residual-weighted perturbations adapt to the model's current state, directing stronger updates toward regions of high uncertainty while preserving the underlying NTK structure. The deterministic drift component steadily advances predictions toward optimal alignment with training targets, complemented by stochastic elements that naturally attenuate as residuals diminish, thereby illuminating the enhanced generalization and efficiency conferred by minibatch training in wide neural networks.

4. Experiment Results

In this section, we experimentally validate our theory for the Stochastic Neural Tangent Kernel (SNTK) by examining how well SNTK captures the noisy behavior of real Stochastic Gradient Descent (SGD) training in wide neural networks, compared to the standard Neural Tangent Kernel (NTK).

We use the two moons dataset ($N = 500$ samples) and train a three-layer fully-connected network with 1,024 units per layer using mean squared error (MSE) loss. SGD is performed with batch size 8. For each method, we compute the corresponding function values after 100 epochs, using NTK as in Definition 1 and SNTK as in Definition 2, and visualize the decision boundaries.

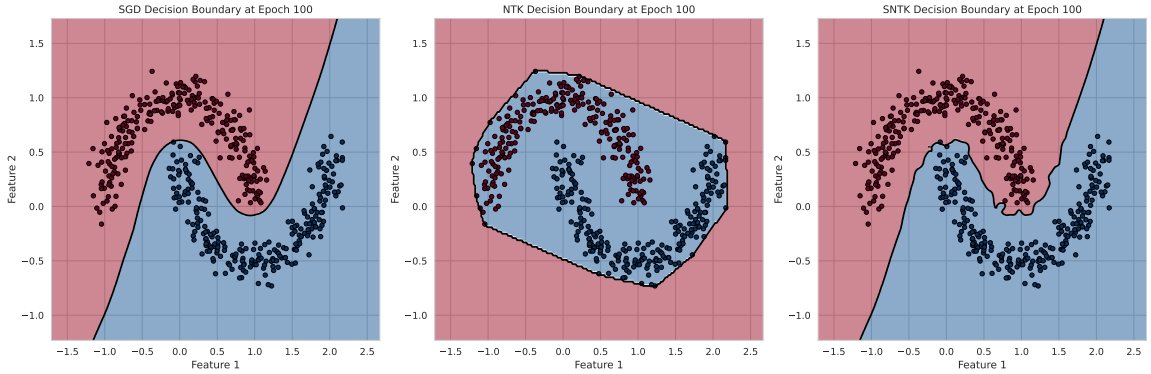

Figure 1: Decision boundaries after 100 training epochs for SGD (left), NTK (center), and SNTK (right) on the two moons dataset.

Figure 1 demonstrates SNTK's ability to more closely model SGD's evolution compared to NTK. The SNTK boundary resembles the SGD boundary but exhibits greater local adaptation and increased uncertainty in high-residual regions, consistent with Theorem 1. While the NTK produces a rigid, convex boundary typical of deterministic kernel regression, SNTK introduces adaptive stochasticity that evolves with the residuals, capturing the enhanced generalization and local feature learning induced by minibatch stochasticity. These experimental results demonstrate how SGD's intrinsic minibatch noise guides optimization towards flatter minima and improved generalization, and reinforce the SNTK's potential as a refined tool for understanding and modeling SGD dynamics in wide neural networks.

5. Conclusion

This paper introduces the Stochastic Neural Tangent Kernel (SNTK), a framework that extends the classical Neural Tangent Kernel to incorporate the stochastic dynamics of minibatch SGD in infinite-width neural networks. By explicitly modeling adaptive, residual-weighted noise in function space, the SNTK provides a refined lens to understand how SGD's intrinsic noise guides training towards better generalization and robustness.

The SNTK framework presents several promising avenues for future research. One potential direction is developing noise-aware optimization algorithms that exploit adaptive stochasticity, potentially enhancing training robustness and efficiency. Another important area is the investigation of scaling laws in large neural networks, which could provide guidance on effectively balancing batch size and learning rate for optimal performance.

References

[1] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[2] Alberto Bietti, Joan Bruna, Guillermo Sapiro, and Julien Mairal. Deep equals shallow for relu networks in kernel regimes. In *International Conference on Learning Representations (ICLR)*, 2021.

[3] Alberto Bietti, Joan Bruna, Clayton Sanford, and Laurent Sifre. Learning single-index models with shallow neural networks. *arXiv preprint arXiv:2210.15651*, 2022.

[4] Guy Blanc, Neha Gupta, Gregory Valiant Lee, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. *Conference on Learning Theory*, 2020.

[5] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 2018.

[6] Pratik Chaudhari, Adam Oberman, Stanley Osher, Stefano Soatto, and Guillaume Carlier. Deep relaxation: Partial differential equations for optimizing deep neural networks. *Research in the Mathematical Sciences*, 5(3):30, 2018.

[7] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 2019.

[8] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *International Conference on Learning Representations*, 2019.

[9] Ethan Dyer and Guy Gur-Ari. Asymptotics of wide networks from feynman diagrams. *arXiv preprint arXiv:2009.00613*, 2020.

[10] Markus Geiger, Tess Smidt, and Benjamin Kurt Miller. Disentangling the roles of curation, data-augmentation and the prior in the cold posterior effect. *Advances in Neural Information Processing Systems*, 2020.

[11] Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. In *International Conference on Learning Representations*, 2020.

[12] Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376:287–322, 2020.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 1997.

[14] Hong Hu and Yue M. Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 67(3):1932–1964, 2020.

[15] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pages 8571–8580, Montréal, Canada, 2018. Curran Associates, Inc.

[16] Arthur Jacot, Franck Ged, Franck Gabriel, Clément Hongler, and Francis Bach. Kernel alignment risk predicts trained linear model performance. *arXiv preprint arXiv:2006.09027*, 2020.

[17] Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *International Conference on Learning Representations*, 2019.

[18] 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.

[19] Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[20] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. *International Conference on Machine Learning*, 2017.

[21] Xiang Li, Zebang Shen, Liang Zhang, and Niao He. A hessian-aware stochastic differential equation for modelling sgd. *arXiv preprint arXiv:2405.18373*, 2024.

[22] Kangqiao Liu and Mikhail Belkin. Noise and fluctuation of finite learning rate stochastic gradient descent. *International Conference on Machine Learning*, 2021.

[23] Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and scaling rules for adaptive gradient algorithms. In *Advances in Neural Information Processing Systems*, volume 35, pages 4632–4648. Curran Associates, Inc., 2022.

[24] Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

[25] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 2018.

[26] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. The mean-field limit of the neural tangent kernel. *Advances in Neural Information Processing Systems*, 2019.

[27] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media, 2003.

[28] Toan Q Nguyen and Julian Salazar. Transformers without tears: Improving the normalization of self-attention. *arXiv preprint arXiv:1910.05895*, 2019.

[29] Atsushi Nitanda and Taiji Suzuki. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. *International Conference on Learning Representations*, 2021.

[30] Daniel S. Park, Jascha Sohl-Dickstein, Quoc V. Le, and Samuel L. Smith. The effect of network width on stochastic gradient descent and generalization: an empirical study. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pages 5042–5051. PMLR, 2019.

[31] Boris T. Polyak. *Introduction to Optimization*. Optimization Software, 1987.

[32] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 1951.

[33] Antonio Sclocchi, Mario Geiger, and Matthieu Wyart. Dissecting the effects of sgd noise in distinct regimes of deep learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 30600–30616. PMLR, 2023.

[34] Samuel L. Smith and Quoc V. Le. A bayesian perspective on generalization and stochastic gradient descent. *arXiv preprint arXiv:1710.06451*, 2017.

[35] Samuel L Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic gradient descent. *International Conference on Machine Learning*, 2020.

[36] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z. URL https://doi.org/10.1007/s10208-011-9099-z.

[37] Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2020.

Appendix A. Notation and Assumptions

A.1. Notation

DATASET

- N: Number of training samples
- Training data: $\{(x_i, y_i)\}_{i=1}^N$ where $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$
- Minibatch: $\mathcal{B}_t \subseteq \{1, \ldots, N\}$ of size b sampled uniformly at random

NETWORK ARCHITECTURE

- Depth: L (fixed)
- Widths: $(n_0, n_1, \ldots, n_{L-1}, n_L)$ where
  - $n_0 = d$ (input dimension)
  - $n_L = 1$ (output dimension, scalar regression)
  - n_ℓ for $\ell = 1, \ldots, L-1$ are hidden layer widths
  - $n_{\min} := \min_{\ell \in \{1,\ldots,L-1\}} n_\ell$ denotes the minimum hidden layer width.
- Weights: $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ for $\ell = 1, \ldots, L$
  - Individual weight: $W_{ij}^{(\ell)}$ connects neuron j in layer $\ell - 1$ to neuron i in layer ℓ
- Biases: $b^{(\ell)} \in \mathbb{R}^{n_\ell}$ for $\ell = 1, \ldots, L$
  - Individual bias: $b_i^{(\ell)}$ for neuron i in layer ℓ
- Parameter vector: $\theta = \{W^{(\ell)}, b^{(\ell)}\}_{\ell=1}^L \in \mathbb{R}^P$ where

$$P = \sum_{\ell=1}^{L} n_\ell n_{\ell-1} + \sum_{\ell=1}^{L} n_\ell$$

  is the total number of parameters

FORWARD PROPAGATION

For input $x \in \mathbb{R}^d$:

1. **Input layer** ($\ell = 0$):
$$a_j^{(0)}(x) = x_j, \quad j = 1, \ldots, d$$
2. **Hidden and output layers** ($\ell = 1, \ldots, L$):
   - **Pre-activation:**
$$h_i^{(\ell)}(x) = \frac{1}{\sqrt{n_{\ell-1}}} \sum_{j=1}^{n_{\ell-1}} W_{ij}^{(\ell)} a_j^{(\ell-1)}(x) + b_i^{(\ell)}$$

- **Activation:**
$$a_i^{(\ell)}(x) = \phi(h_i^{(\ell)}(x)), \quad i = 1, \ldots, n_\ell$$

3. **Network output:**
$$f(x;\theta) = h_1^{(L)}(x) \in \mathbb{R}$$

BACKWARD PROPAGATION

For input x, define backward signals $\delta_i^{(\ell)}(x) := \frac{\partial f(x;\theta)}{\partial h_i^{(\ell)}(x)}$:

- **Output layer** ($\ell = L$):
$$\delta_1^{(L)}(x) = 1$$

- **Hidden layers** ($\ell = L-1, \ldots, 1$):
$$\delta_j^{(\ell)}(x) = \phi'(h_j^{(\ell)}(x)) \cdot \frac{1}{\sqrt{n_\ell}} \sum_{i=1}^{n_{\ell+1}} W_{ij}^{(\ell+1)} \delta_i^{(\ell+1)}(x)$$

  for $j = 1, \ldots, n_\ell$

PARAMETER GRADIENTS

For input x:
$$\frac{\partial f(x;\theta)}{\partial W_{ij}^{(\ell)}} = \delta_i^{(\ell+1)}(x) \cdot a_j^{(\ell-1)}(x), \qquad \frac{\partial f(x;\theta)}{\partial b_i^{(\ell)}} = \delta_i^{(\ell+1)}(x)$$

Vectorized: $\nabla_\theta f(x;\theta) \in \mathbb{R}^P$ stacks all parameter gradients.

LOSS AND RESIDUALS

- Individual loss: $\mathcal{L}_i(\theta) = \frac{1}{2}(f(x_i;\theta) - y_i)^2$
- Full loss: $\mathcal{L}(\theta) = \sum_{i=1}^N \mathcal{L}_i(\theta)$
- Residual: $r_i(t) = f(x_i;\theta_t) - y_i$
- Individual gradient: $\nabla \mathcal{L}_i(\theta) = r_i(t) \nabla_\theta f(x_i;\theta)$

TRAINING DYNAMICS

- Learning rate: $\eta > 0$
- Time: $t \in [0, \infty)$
- Parameters evolve as: $\theta_t \in \mathbb{R}^P$
- Network output: $f_t(x) := f(x;\theta_t)$

SECOND-ORDER DERIVATIVES

- Hessian: $H(x;\theta) = \nabla_\theta^2 f(x;\theta) \in \mathbb{R}^{P \times P}$

A.2. Assumptions

Assumption 1 (Infinite-Width Neural Network) *Consider an L-layer fully connected neural network with widths $(d, n_1, n_2, \ldots, n_{L-1}, 1)$ satisfying:*

(A1.1) Width scaling: $n_\ell \geq C \log(N)$ *for all* $\ell \in \{1, \ldots, L-1\}$ *and some universal constant* $C > 0$.

(A1.2) Weight initialization: $W_{ij}^{(\ell)} \sim \mathcal{N}(0, \sigma_w^2/n_{\ell-1})$ *i.i.d. with* $\sigma_w^2 = 2$, *and* $b_i^{(\ell)} \sim \mathcal{N}(0, \sigma_b^2)$ *i.i.d. with* $\sigma_b^2 = 1$.

(A1.3) Activation function: $\phi : \mathbb{R} \to \mathbb{R}$ *is twice continuously differentiable with bounded derivatives:* $|\phi'(z)| \leq C_1$ *and* $|\phi''(z)| \leq C_2$ *for all* $z \in \mathbb{R}$ *and universal constants* $C_1, C_2 > 0$.

(A1.4) Input boundedness: *All inputs satisfy* $\|x\|_2 \leq R$ *for some constant* $R > 0$.

(A1.5) Target boundedness: *All targets satisfy* $|y_i| \leq M$ *for some constant* $M > 0$.

(A1.6) Infinite-width limit: *All asymptotic results hold in the limit* $\min_{\ell \in \{1,\ldots,L-1\}} n_\ell \to \infty$, *with depth* L *and other constants remaining fixed.*

Notation: Throughout, $\mathcal{O}_{\mathbb{P}}(\cdot)$ denotes the order in probability in the infinite-width limit, and $o_{\mathbb{P}}(\cdot)$ denotes convergence to zero in probability.

Assumption 2 (Data Separation) *The training data $\{x_1, \ldots, x_N\}$ is δ-separated with respect to the Neural Tangent Kernel: there exists $\delta \in [0, 1)$ such that for all distinct pairs $i \neq j$,*

$$\frac{|K(x_i, x_j)|^2}{K(x_i, x_i)K(x_j, x_j)} \leq \delta. \tag{3}$$

Appendix B. Technical Theorems & Lemmas

Lemma 1 (Residual Boundedness) *Under Assumptions 1 (A1.1)–(A1.6), the NTK regime implies that for any finite training time $T > 0$, the residuals satisfy*

$$|r_i(t)| = \mathcal{O}_{\mathbb{P}}(1)$$

uniformly over $i \in \{1, \ldots, N\}$ and $t \in [0, T]$.

Proof By the mean value theorem, for some $\bar{\theta}$ on the line segment between θ_0 and θ_t:

$$f(x_i; \theta_t) = f(x_i; \theta_0) + \nabla_\theta f(x_i; \bar{\theta})^\top (\theta_t - \theta_0).$$

From the NTK regime (A1.7), $\|\theta_t - \theta_0\| = \mathcal{O}_{\mathbb{P}}(n_{\min}^{-1/2})$. By NTK convergence [15], the gradient norms satisfy $\|\nabla_\theta f(x_i; \bar{\theta})\|^2 = K(x_i, x_i) = \Theta_{\mathbb{P}}(1)$ uniformly, with fluctuations $\mathcal{O}_{\mathbb{P}}(n_{\min}^{-1/2})$. Therefore:

$$|f(x_i; \theta_t) - f(x_i; \theta_0)| \leq \|\nabla_\theta f(x_i; \bar{\theta})\| \cdot \|\theta_t - \theta_0\| = \mathcal{O}_{\mathbb{P}}(1) \cdot \mathcal{O}_{\mathbb{P}}(n_{\min}^{-1/2}) = \mathcal{O}_{\mathbb{P}}(n_{\min}^{-1/2}).$$

At initialization, with the specified random initialization (A1.2), the network output converges in distribution to a Gaussian process [15] with bounded second moment, giving $f(x_i;\theta_0) = \mathcal{O}_{\mathbb{P}}(1)$. Thus:

$$f(x_i;\theta_t) = f(x_i;\theta_0) + \mathcal{O}_{\mathbb{P}}(n_{\min}^{-1/2}) = \mathcal{O}_{\mathbb{P}}(1).$$

Combined with $|y_i| \leq M$ from (A1.5), we obtain $|r_i(t)| = |f(x_i;\theta_t) - y_i| = \mathcal{O}_{\mathbb{P}}(1)$ uniformly. ■

Lemma 2 (Gradient Decorrelation) *Under Assumptions 1 and 2, for distinct training inputs x_i, x_j with $i \neq j$:*

$$|\langle \nabla_\theta f(x_i;\theta_t), \nabla_\theta f(x_j;\theta_t)\rangle| \leq \sqrt{\delta} \cdot \|\nabla_\theta f(x_i;\theta_t)\| \, \|\nabla_\theta f(x_j;\theta_t)\| + \mathcal{O}\left(n_{\min}^{-1/2}\right). \tag{4}$$

Equivalently, the normalized kernel satisfies $|K(x_i,x_j)| \leq \sqrt{\delta} \cdot \sqrt{K(x_i,x_i)K(x_j,x_j)} + \mathcal{O}(n_{\min}^{-1/2})$.

Proof By definition of the Neural Tangent Kernel:

$$K(x_i,x_j) = \langle \nabla_\theta f(x_i;\theta_t), \nabla_\theta f(x_j;\theta_t)\rangle.$$

Under Assumption 1, by [15, Theorem 1], the NTK converges to a deterministic limit with fluctuations of order $\mathcal{O}_{\mathbb{P}}((n_{\min})^{-1/2})$. Taking square roots of the separation condition (3) yields:

$$|K(x_i,x_j)| \leq \sqrt{\delta} \cdot \sqrt{K(x_i,x_i)K(x_j,x_j)} + \mathcal{O}\left(n_{\min}^{-1/2}\right).$$

Since $K(x_i,x_i) = \|\nabla_\theta f(x_i)\|^2$, this gives (4). ■

Lemma 3 (Cross-Term Cancellation) *Under Assumptions 1 and 2, the cross-correlation term*

$$R_{\text{cross}} := -\frac{1}{b^2} \sum_{\substack{i,j \in \mathcal{B}_t \\ i \neq j}} r_i r_j \nabla_\theta f(x_i;\theta_t) \nabla_\theta f(x_j;\theta_t)^\top$$

satisfies: for all training points $x_m, x_n \in \{x_1, \ldots, x_N\}$,

$$\left| \nabla_\theta f(x_m;\theta_t)^\top R_{\text{cross}} \, \nabla_\theta f(x_n;\theta_t) \right| \leq \frac{C_\delta}{b} \cdot \left| \frac{1}{b} \sum_{i \in \mathcal{B}_t} r_i^2(t) K(x_m,x_i) K(x_i,x_n) \right|, \tag{5}$$

where $C_\delta = 2(1+\sqrt{\delta})^2$ depends only on the separation parameter δ from Assumption 2.

Proof Computing the function-space projection at training points x_m, x_n yields

$$\nabla_\theta f(x_m)^\top R_{\text{cross}} \, \nabla_\theta f(x_n) = -\frac{1}{b^2} \sum_{\substack{i,j \in \mathcal{B}_t \\ i \neq j}} r_i r_j K(x_m,x_i) K(x_j,x_n). \tag{6}$$

Using the identity $\sum_{i\neq j} a_i b_j = (\sum_i a_i)(\sum_j b_j) - \sum_i a_i b_i$, we decompose this as

$$\sum_{\substack{i,j\in\mathcal{B}_t\\ i\neq j}} r_i r_j K(x_m,x_i)K(x_j,x_n) = \underbrace{\left(\sum_{i\in\mathcal{B}_t} r_i K(x_m,x_i)\right)}_{=:\alpha}\underbrace{\left(\sum_{j\in\mathcal{B}_t} r_j K(x_j,x_n)\right)}_{=:\beta} - \underbrace{\sum_{i\in\mathcal{B}_t} r_i^2 K(x_m,x_i)K(x_i,x_n)}_{=:D}, \tag{7}$$

where D/b is precisely the diagonal contribution appearing in the SNTK. Taking absolute values gives

$$\left|\nabla_\theta f(x_m)^\top R_{\text{cross}}\, \nabla_\theta f(x_n)\right| = \frac{1}{b^2}\left|\alpha\beta - D\right| \le \frac{|\alpha\beta|}{b^2} + \frac{|D|}{b^2}. \tag{8}$$

We now bound $|\alpha|$. Decomposing into diagonal and off-diagonal contributions,

$$\alpha = r_m K(x_m,x_m) + \sum_{\substack{i\in\mathcal{B}_t\\ i\neq m}} r_i K(x_m,x_i). \tag{9}$$

For the off-diagonal sum, applying Cauchy-Schwarz yields

$$\left|\sum_{\substack{i\in\mathcal{B}_t\\ i\neq m}} r_i K(x_m,x_i)\right|^2 \le \left(\sum_{\substack{i\in\mathcal{B}_t\\ i\neq m}} r_i^2\right)\left(\sum_{\substack{i\in\mathcal{B}_t\\ i\neq m}} K(x_m,x_i)^2\right). \tag{10}$$

By Assumption 2, for $i \neq m$ we have $K(x_m,x_i)^2 \le \delta \cdot K(x_m,x_m)K(x_i,x_i)$. Summing over the off-diagonal indices,

$$\sum_{\substack{i\in\mathcal{B}_t\\ i\neq m}} K(x_m,x_i)^2 \le \delta \cdot K(x_m,x_m)\sum_{\substack{i\in\mathcal{B}_t\\ i\neq m}} K(x_i,x_i) \le \delta \cdot K(x_m,x_m)\cdot b\bar{K}, \tag{11}$$

where $\bar{K} := \frac{1}{b}\sum_{i\in\mathcal{B}_t} K(x_i,x_i)$ is the average diagonal kernel value. Define $\bar{r}^2 := \frac{1}{b}\sum_{i\in\mathcal{B}_t} r_i^2$ as the average squared residual. Substituting (11) into (10),

$$\left|\sum_{\substack{i\in\mathcal{B}_t\\ i\neq m}} r_i K(x_m,x_i)\right|^2 \le b\bar{r}^2 \cdot \delta K(x_m,x_m)\cdot b\bar{K} = \delta b^2 \bar{r}^2 K(x_m,x_m)\bar{K}. \tag{12}$$

Taking square roots and applying the triangle inequality to (9),

$$|\alpha| \le |r_m| K(x_m,x_m) + \sqrt{\delta}\cdot b\cdot \bar{r}\sqrt{K(x_m,x_m)\bar{K}}. \tag{13}$$

Under Assumption 1, by [15, Theorem 1], the diagonal kernel values satisfy $K(x_i,x_i) = \Theta(1)$ uniformly, so there exist constants $0 < K_{\min} \le K_{\max} < \infty$ such that $K_{\min} \le K(x_i,x_i) \le K_{\max}$ for all i. In particular, $\bar{K} \le K_{\max}$. Since $r_m^2 \le \sum_{i\in\mathcal{B}_t} r_i^2 = b\bar{r}^2$, we have $|r_m| \le \sqrt{b}\bar{r}$, and thus (13) becomes

$$|\alpha| \le K_{\max}\sqrt{b}\bar{r} + \sqrt{\delta}\cdot b\cdot \bar{r}K_{\max} = K_{\max} b\bar{r}\left(\frac{1}{\sqrt{b}} + \sqrt{\delta}\right). \tag{14}$$

By an identical argument,

$$|\beta| \leq K_{\max} b\bar{r} \left(\frac{1}{\sqrt{b}} + \sqrt{\delta} \right). \tag{15}$$

For the diagonal term D, we use the bound $|K(x_m, x_i)| \leq K_{\max}$ and $|K(x_i, x_n)| \leq K_{\max}$ to obtain

$$|D| \leq K_{\max}^2 \sum_{i \in \mathcal{B}_t} r_i^2 = K_{\max}^2 b\bar{r}^2. \tag{16}$$

Combining (14) and (15),

$$|\alpha\beta| \leq K_{\max}^2 b^2 \bar{r}^2 \left(\frac{1}{\sqrt{b}} + \sqrt{\delta} \right)^2. \tag{17}$$

Therefore,

$$\frac{|\alpha\beta|}{b^2} \leq K_{\max}^2 \bar{r}^2 \left(\frac{1}{\sqrt{b}} + \sqrt{\delta} \right)^2. \tag{18}$$

To express the bound in terms of the diagonal contribution, note that from (16) we have $|D|/b \leq K_{\max}^2 \bar{r}^2$. Taking the ratio of (18) to $|D|/b$,

$$\frac{|\alpha\beta|/b^2}{|D|/b} \leq \frac{\left(\frac{1}{\sqrt{b}} + \sqrt{\delta} \right)^2}{1} = \left(\frac{1}{\sqrt{b}} + \sqrt{\delta} \right)^2. \tag{19}$$

For $b \geq 1$, we have $1/\sqrt{b} \leq 1$, so

$$\left(\frac{1}{\sqrt{b}} + \sqrt{\delta} \right)^2 \leq (1 + \sqrt{\delta})^2. \tag{20}$$

Returning to (8) and substituting the bounds,

$$\begin{aligned}
\left| \nabla_\theta f(x_m)^\top R_{\text{cross}} \, \nabla_\theta f(x_n) \right| &\leq \frac{|\alpha\beta|}{b^2} + \frac{|D|}{b^2} \\
&\leq (1 + \sqrt{\delta})^2 \cdot \frac{|D|}{b} + \frac{|D|}{b^2} \\
&\leq \left((1 + \sqrt{\delta})^2 + 1 \right) \cdot \frac{1}{b} \cdot \frac{|D|}{b} \\
&\leq \frac{2(1 + \sqrt{\delta})^2}{b} \cdot \frac{1}{b} \left| \sum_{i \in \mathcal{B}_t} r_i^2 K(x_m, x_i) K(x_i, x_n) \right|,
\end{aligned} \tag{21}$$

$$\left| \nabla_\theta f(x_m; \theta_t)^\top R_{\text{cross}} \, \nabla_\theta f(x_n; \theta_t) \right| \leq \frac{C_\delta}{b} \cdot \left| \frac{1}{b} \sum_{i \in \mathcal{B}_t} r_i^2(t) K(x_m, x_i) K(x_i, x_n) \right|$$

where in the last line we used $(1 + \sqrt{\delta})^2 + 1 \leq 2(1 + \sqrt{\delta})^2$ for $\delta \in [0, 1)$. Setting $C_\delta = 2(1 + \sqrt{\delta})^2$ completes the proof. ■

Theorem 1 (Minibatch Noise Structure) *Under Assumptions 1 and 2, the gradient noise covariance satisfies:*

$$\Sigma(\theta_t) = \frac{1}{b}\sum_{i\in\mathcal{B}_t} r_i^2(t)\nabla_\theta f(x_i;\theta_t)\nabla_\theta f(x_i;\theta_t)^\top + \mathcal{E}(\theta_t), \tag{22}$$

where $r_i(t) = f(x_i;\theta_t) - y_i$ are the residuals. The error term $\mathcal{E}(\theta_t)$ satisfies the function-space bound: for all training points $x_m, x_n \in \{x_1,\dots,x_N\}$,

$$\left|\nabla_\theta f(x_m;\theta_t)^\top \mathcal{E}(\theta_t)\,\nabla_\theta f(x_n;\theta_t)\right| \le \frac{C_\delta}{b}\cdot\frac{1}{b}\sum_{i\in\mathcal{B}_t} r_i^2(t) + \mathcal{O}\left(n_{\min}^{-1/2}\right), \tag{23}$$

where $C_\delta = K_{\max}^2(1+2(1+\sqrt{\delta})^2)$ depends on both the kernel bound $K_{\max}$ and separation parameter δ. Here $n_{\min} := \min_{\ell\in\{1,\dots,L-1\}} n_\ell$ denotes the minimum hidden layer width.

Proof By definition, as described in [23], the gradient noise covariance is:

$$\Sigma(\theta_t) = \frac{1}{b}\sum_{i\in\mathcal{B}_t}(\nabla\mathcal{L}_i(\theta_t) - \nabla\mathcal{L}_{\mathcal{B}_t}(\theta_t))(\nabla\mathcal{L}_i(\theta_t) - \nabla\mathcal{L}_{\mathcal{B}_t}(\theta_t))^\top.$$

For mean squared error loss, we have $\nabla\mathcal{L}_i(\theta) = r_i\nabla_\theta f(x_i;\theta)$ and $\nabla\mathcal{L}_{\mathcal{B}_t}(\theta) = \frac{1}{b}\sum_{j\in\mathcal{B}_t} r_j\nabla_\theta f(x_j;\theta)$. Expanding:

$$\nabla\mathcal{L}_i - \nabla\mathcal{L}_{\mathcal{B}_t} = r_i\nabla_\theta f(x_i;\theta_t) - \frac{1}{b}\sum_{j\in\mathcal{B}_t} r_j\nabla_\theta f(x_j;\theta_t).$$

Computing the outer product and summing over the minibatch:

$$\begin{aligned}\Sigma(\theta_t) = \frac{1}{b}\sum_{i\in\mathcal{B}_t} r_i^2\nabla_\theta f(x_i;\theta_t)\nabla_\theta f(x_i;\theta_t)^\top \\ - \frac{1}{b^2}\left(\sum_{i\in\mathcal{B}_t} r_i\nabla_\theta f(x_i;\theta_t)\right)\left(\sum_{j\in\mathcal{B}_t} r_j\nabla_\theta f(x_j;\theta_t)\right)^\top.\end{aligned}$$

Expanding the second term and separating diagonal ($i=j$) from off-diagonal ($i\neq j$) contributions:

$$\begin{aligned}\left(\sum_{i\in\mathcal{B}_t} r_i\nabla_\theta f(x_i)\right)\left(\sum_{j\in\mathcal{B}_t} r_j\nabla_\theta f(x_j)\right)^\top &= \sum_{i\in\mathcal{B}_t} r_i^2\nabla_\theta f(x_i)\nabla_\theta f(x_i)^\top \\ &\quad + \sum_{\substack{i,j\in\mathcal{B}_t\\ i\neq j}} r_i r_j\nabla_\theta f(x_i)\nabla_\theta f(x_j)^\top.\end{aligned}$$

Substituting back yields:

$$\Sigma(\theta_t) = \frac{b-1}{b^2}\sum_{i\in\mathcal{B}_t} r_i^2\nabla_\theta f(x_i;\theta_t)\nabla_\theta f(x_i;\theta_t)^\top + R_{\text{cross}}, \tag{24}$$

where

$$R_{\text{cross}} := -\frac{1}{b^2} \sum_{\substack{i,j \in \mathcal{B}_t \\ i \neq j}} r_i r_j \nabla_\theta f(x_i; \theta_t) \nabla_\theta f(x_j; \theta_t)^\top.$$

We now bound the error incurred by approximating $\Sigma(\theta_t)$ with the leading-order term

$$\frac{1}{b} \sum_{i \in \mathcal{B}_t} r_i^2 \nabla_\theta f(x_i) \nabla_\theta f(x_i)^\top.$$

Define

$$\mathcal{E}(\theta_t) := \Sigma(\theta_t) - \frac{1}{b} \sum_{i \in \mathcal{B}_t} r_i^2(t) \nabla_\theta f(x_i; \theta_t) \nabla_\theta f(x_i; \theta_t)^\top.$$

From (24), this error decomposes as

$$\mathcal{E}(\theta_t) = -\frac{1}{b^2} \sum_{i \in \mathcal{B}_t} r_i^2 \nabla_\theta f(x_i; \theta_t) \nabla_\theta f(x_i; \theta_t)^\top + R_{\text{cross}}, \tag{25}$$

where the first term arises from the coefficient correction $\frac{b-1}{b^2} - \frac{1}{b} = -\frac{1}{b^2}$.

For any training points x_m, x_n, projecting (25) into function space and applying the triangle inequality:

$$\begin{aligned} \left| \nabla_\theta f(x_m)^\top \mathcal{E}(\theta_t) \nabla_\theta f(x_n) \right| \leq \frac{1}{b^2} \sum_{i \in \mathcal{B}_t} r_i^2 \left| K(x_m, x_i) K(x_i, x_n) \right| \\ + \left| \nabla_\theta f(x_m)^\top R_{\text{cross}} \nabla_\theta f(x_n) \right|. \end{aligned} \tag{26}$$

For the first term in (26), we bound the absolute value of the kernel product. Since $K(x_i, x_i) = \Theta(1)$ uniformly by NTK convergence, there exist constants $0 < K_{\min} \leq K_{\max} < \infty$ such that $K_{\min} \leq K(x_i, x_i) \leq K_{\max}$ for all i. Moreover, by Lemma 2, for $m \neq i$ we have $|K(x_m, x_i)| \leq \sqrt{\delta}\sqrt{K(x_m, x_m) K(x_i, x_i)} + \mathcal{O}(n_{\min}^{-1/2}) \leq \sqrt{\delta} K_{\max} + \mathcal{O}(n_{\min}^{-1/2})$. Therefore, using the triangle inequality $|AB| \leq |A| \cdot |B|$:

$$\begin{aligned} \frac{1}{b^2} \sum_{i \in \mathcal{B}_t} r_i^2 \left| K(x_m, x_i) K(x_i, x_n) \right| &\leq \frac{1}{b^2} \sum_{i \in \mathcal{B}_t} r_i^2 |K(x_m, x_i)| \cdot |K(x_i, x_n)| \\ &\leq \frac{1}{b^2} \sum_{i \in \mathcal{B}_t} r_i^2 \cdot K_{\max}^2 \\ &= \frac{K_{\max}^2}{b} \cdot \frac{1}{b} \sum_{i \in \mathcal{B}_t} r_i^2. \end{aligned} \tag{27}$$

Setting $\kappa := K_{\max}^2$, we obtain:

$$\frac{1}{b^2} \sum_{i \in \mathcal{B}_t} r_i^2 \left| K(x_m, x_i) K(x_i, x_n) \right| \leq \frac{\kappa}{b} \cdot \frac{1}{b} \sum_{i \in \mathcal{B}_t} r_i^2 + \mathcal{O}(n_{\min}^{-1/2}). \tag{28}$$

For the second term in (26), by Lemma 3 under Assumptions 1 and 2:

$$\left|\nabla_\theta f(x_m)^\top R_{\text{cross}} \nabla_\theta f(x_n)\right| \leq \frac{C_\delta}{b} \cdot \left|\frac{1}{b}\sum_{i\in\mathcal{B}_t} r_i^2(t) K(x_m, x_i) K(x_i, x_n)\right|, \tag{29}$$

where $C_\delta = 2(1+\sqrt{\delta})^2$.

To combine these bounds, we use the triangle inequality. Since

$$\left|\sum_{i\in\mathcal{B}_t} r_i^2 K(x_m, x_i) K(x_i, x_n)\right| \leq \sum_{i\in\mathcal{B}_t} r_i^2 |K(x_m, x_i)||K(x_i, x_n)| \leq K_{\max}^2 \sum_{i\in\mathcal{B}_t} r_i^2,$$

we have

$$\frac{C_\delta}{b} \cdot \left|\frac{1}{b}\sum_{i\in\mathcal{B}_t} r_i^2 K(x_m, x_i) K(x_i, x_n)\right| \leq \frac{C_\delta K_{\max}^2}{b} \cdot \frac{1}{b}\sum_{i\in\mathcal{B}_t} r_i^2.$$

Combining (26), (28), and (29):

$$\begin{aligned}\left|\nabla_\theta f(x_m)^\top \mathcal{E}(\theta_t) \nabla_\theta f(x_n)\right| &\leq \frac{\kappa}{b} \cdot \frac{1}{b}\sum_{i\in\mathcal{B}_t} r_i^2 + \frac{C_\delta K_{\max}^2}{b} \cdot \frac{1}{b}\sum_{i\in\mathcal{B}_t} r_i^2(t) + \mathcal{O}(n_{\min}^{-1/2}) \\ &= \frac{\kappa + C_\delta K_{\max}^2}{b} \cdot \frac{1}{b}\sum_{i\in\mathcal{B}_t} r_i^2(t) + \mathcal{O}(n_{\min}^{-1/2}).\end{aligned}$$

Redefining $C_\delta \leftarrow \kappa + C_\delta K_{\max}^2 = K_{\max}^2(1 + 2(1+\sqrt{\delta})^2)$ yields (23), where C_δ depends on both the kernel bound $K_{\max}$ and the separation parameter δ, but remains bounded independently of network width as $n_{\min} \to \infty$. ■

Lemma 4 (Trace Bound for Noise Error) *Under the assumptions of Theorem 1, the error term $\mathcal{E}(\theta_t)$ from equation (22) satisfies:*

$$\text{Tr}(\mathcal{E}(\theta_t)) = \mathcal{O}_{\mathbb{P}}(1). \tag{30}$$

Proof From the decomposition in (25), we have:

$$\mathcal{E}(\theta_t) = -\frac{1}{b^2}\sum_{i\in\mathcal{B}_t} r_i^2 \nabla_\theta f(x_i;\theta_t) \nabla_\theta f(x_i;\theta_t)^\top + R_{\text{cross}},$$

where $R_{\text{cross}} = -\frac{1}{b^2}\sum_{\substack{i,j\in\mathcal{B}_t \\ i\neq j}} r_i r_j \nabla_\theta f(x_i;\theta_t) \nabla_\theta f(x_j;\theta_t)^\top$.

Taking the trace:

$$\begin{aligned}\text{Tr}(\mathcal{E}(\theta_t)) &= -\frac{1}{b^2}\sum_{i\in\mathcal{B}_t} r_i^2 \|\nabla_\theta f(x_i;\theta_t)\|^2 + \text{Tr}(R_{\text{cross}}) \\ &= -\frac{1}{b^2}\sum_{i\in\mathcal{B}_t} r_i^2 K(x_i, x_i) + \text{Tr}(R_{\text{cross}}).\end{aligned}$$

Since $K(x_i, x_i) = \Theta(1)$ uniformly by NTK convergence [1], there exists $K_{\max} < \infty$ such that $K(x_i, x_i) \leq K_{\max}$ for all i. By Lemma 1, $|r_i| = \mathcal{O}_{\mathbb{P}}(1)$. Therefore:

$$\begin{aligned}\left| -\frac{1}{b^2} \sum_{i \in \mathcal{B}_t} r_i^2 K(x_i, x_i) \right| &\leq \frac{K_{\max}}{b^2} \sum_{i \in \mathcal{B}_t} \mathcal{O}_{\mathbb{P}}(1) \\ &= \frac{K_{\max}}{b^2} \cdot b \cdot \mathcal{O}_{\mathbb{P}}(1) \\ &= \frac{K_{\max}}{b} \cdot \mathcal{O}_{\mathbb{P}}(1) = \mathcal{O}_{\mathbb{P}}(1).\end{aligned}$$

Computing the trace of R_{cross}:

$$\begin{aligned}\operatorname{Tr}(R_{\text{cross}}) &= -\frac{1}{b^2} \sum_{\substack{i,j \in \mathcal{B}_t \\ i \neq j}} r_i r_j \langle \nabla_\theta f(x_i; \theta_t), \nabla_\theta f(x_j; \theta_t) \rangle \\ &= -\frac{1}{b^2} \sum_{\substack{i,j \in \mathcal{B}_t \\ i \neq j}} r_i r_j K(x_i, x_j).\end{aligned}$$

By Lemma 2 and Assumption 2, for $i \neq j$:

$$\begin{aligned}|K(x_i, x_j)| &\leq \sqrt{\delta} \sqrt{K(x_i, x_i) K(x_j, x_j)} + \mathcal{O}(n_{\min}^{-1/2}) \\ &\leq \sqrt{\delta} K_{\max} + \mathcal{O}(n_{\min}^{-1/2}).\end{aligned}$$

Therefore:

$$\begin{aligned}|\operatorname{Tr}(R_{\text{cross}})| &\leq \frac{1}{b^2} \sum_{\substack{i,j \in \mathcal{B}_t \\ i \neq j}} |r_i||r_j| \cdot \left(\sqrt{\delta} K_{\max} + \mathcal{O}(n_{\min}^{-1/2}) \right) \\ &\leq \frac{1}{b^2} \cdot b(b-1) \cdot \mathcal{O}_{\mathbb{P}}(1) \cdot \sqrt{\delta} K_{\max} \\ &= \left(1 - \frac{1}{b} \right) \cdot \mathcal{O}_{\mathbb{P}}(1) \cdot \sqrt{\delta} K_{\max} \\ &= \mathcal{O}_{\mathbb{P}}(1).\end{aligned}$$

Combining both terms yields $\operatorname{Tr}(\mathcal{E}(\theta_t)) = \mathcal{O}_{\mathbb{P}}(1)$. ■

Lemma 5 (Negligible Hessian Trace in Infinite Width) *Under Assumption 1, let b be the mini-batch size, $\eta > 0$ the learning rate, and $\Sigma(\theta_t)$ the gradient covariance matrix. For any fixed $t \geq 0$ and bounded input x, in the infinite-width limit, we have*

$$\left| \operatorname{Tr}\left[\nabla_\theta^2 f(x; \theta_t) \cdot \eta \Sigma(\theta_t) \right] \right| = \mathcal{O}_{\mathbb{P}}\left(\eta n_{\min}^{-1/2} \right). \tag{31}$$

In particular, when applying Itô's lemma to $f(x; \theta_t)$, this Hessian trace term is $o_{\mathbb{P}}(dt)$ as $n_{\min} \to \infty$ with η fixed, making it negligible compared to the first-order drift term $|\nabla_\theta f(x; \theta_t)^\top \nabla L(\theta_t)|\, dt = \Theta_{\mathbb{P}}(1)\, dt$.

Proof Since $\Sigma(\theta_t)$ is symmetric positive semidefinite, we have for $H = \nabla^2_\theta f(x;\theta_t)$:

$$|\mathrm{Tr}(H \cdot \eta\Sigma)| \le \eta \|H\|_{\mathrm{op}} \cdot \mathrm{Tr}(\Sigma). \tag{32}$$

We bound $\|H\|_{\mathrm{op}}$ and $\mathrm{Tr}(\Sigma)$ separately.

Bounding $\|H\|_{\mathrm{op}}$ via backward signal propagation. Define the backward Jacobian at layer ℓ:

$$J^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell+1}}, \quad J^{(\ell)}_{ji} = \phi'(h^{(\ell)}_j(x)) \cdot \frac{1}{\sqrt{n_\ell}} W^{(\ell+1)}_{ji}. \tag{33}$$

The backward signals satisfy the recursion $\delta^{(\ell)} = J^{(\ell)}\delta^{(\ell+1)}$ with $\delta^{(L)} = (1)$.

Since $W^{(\ell+1)}_{ji} \sim \mathcal{N}(0, 2/n_\ell)$ independently and $|\phi'| \le C_1$, each entry of $J^{(\ell)}$ is sub-Gaussian with parameter $O(C_1/\sqrt{n_\ell})$. Applying matrix Bernstein inequality [36] with variance proxy $\sigma^2 = O(C_1^2 n_{\ell+1}/n_\ell)$ and row bound $K = O(C_1\sqrt{n_{\ell+1}\log n_{\ell+1}}/\sqrt{n_\ell})$ yields:

$$\|J^{(\ell)}\|_{\mathrm{op}} = \mathcal{O}_{\mathbb{P}}\left(C_1\sqrt{\frac{n_{\ell+1}}{n_\ell}} + C_1\sqrt{\log n_{\ell+1}}\right) = \mathcal{O}_{\mathbb{P}}(C_1)$$

with probability $1-\exp(-\Omega(\log n_{\ell+1}))$, where the second equality uses balanced widths $n_{\ell+1}/n_\ell = O(1)$ from Assumption 1.

Starting from $\|\delta^{(L)}\|_2 = 1$ and applying the operator norm bound inductively:

$$\|\delta^{(\ell)}\|_2 \le \prod_{k=\ell}^{L-1} \|J^{(k)}\|_{\mathrm{op}} = \mathcal{O}_{\mathbb{P}}(C_1^{L-\ell}). \tag{34}$$

This bound on products of random matrices in deep networks is consistent with the precise asymptotic analysis in [12].

Block decomposition of the Hessian. Partition H into layer-pair blocks $H_{\ell,\ell'} \in \mathbb{R}^{p_\ell \times p_{\ell'}}$ where $p_\ell = n_\ell n_{\ell-1} + n_\ell$ is the number of parameters in layer ℓ, and entries are $[H_{\ell,\ell'}]_{ij} = \partial^2 f/\partial\theta_{\ell,i}\partial\theta_{\ell',j}$. Then:

$$\|H\|_{\mathrm{op}} \le \sum_{\ell,\ell'=1}^{L} \|H_{\ell,\ell'}\|_{\mathrm{op}}. \tag{35}$$

Cross-layer blocks ($\ell < \ell'$). For a representative weight-weight entry, the product rule gives:

$$\frac{\partial^2 f}{\partial W^{(\ell)}_{ab}\partial W^{(\ell')}_{cd}} = a^{(\ell'-1)}_d \frac{\partial \delta^{(\ell')}_c}{\partial W^{(\ell)}_{ab}} + \delta^{(\ell')}_c \frac{\partial a^{(\ell'-1)}_d}{\partial W^{(\ell)}_{ab}}. \tag{36}$$

The first term chains backward from layer ℓ' to ℓ through the recursion $\delta^{(k)} = J^{(k)}\delta^{(k+1)}$. Differentiating this recursion introduces one factor of ϕ'' (bounded by C_2) when differentiating through ϕ'. Combined with (34) and the $1/\sqrt{n_k}$ scaling per layer transition:

$$\left|a^{(\ell'-1)}_d \frac{\partial \delta^{(\ell')}_c}{\partial W^{(\ell)}_{ab}}\right| = \mathcal{O}_{\mathbb{P}}\left(C_1^{L-\ell} C_2 (\min_k n_k)^{-(\ell'-\ell)/2}\right).$$

The second term in (36) chains forward from ℓ to $\ell' - 1$ through successive forward Jacobians, each contributing $\mathcal{O}_{\mathbb{P}}(C_1)$:

$$\left| \delta_c^{(\ell')} \frac{\partial a_d^{(\ell'-1)}}{\partial W_{ab}^{(\ell)}} \right| = \mathcal{O}_{\mathbb{P}}(C_1^{L-\ell'} \cdot C_1^{\ell'-1-\ell}) = \mathcal{O}_{\mathbb{P}}(C_1^{L-\ell-1}).$$

Thus $\|H_{\ell,\ell'}\|_{\text{op}} = \mathcal{O}_{\mathbb{P}}(C_1^L C_2 n_{\min}^{-(\ell'-\ell)/2})$ where $n_{\min} = \min_k n_k$. Summing over $L(L-1)/2$ pairs with minimum gap $\ell' - \ell = 1$:

$$\sum_{\ell<\ell'} \|H_{\ell,\ell'}\|_{\text{op}} = \mathcal{O}_{\mathbb{P}}(L^2 C_1^L C_2 n_{\min}^{-1/2}). \tag{37}$$

Same-layer blocks ($\ell = \ell'$). For off-diagonal entries with $i \neq k$ or $j \neq l$:

$$\frac{\partial^2 f}{\partial W_{ij}^{(\ell)} \partial W_{kl}^{(\ell)}} = \frac{\partial^2 f}{\partial h_i^{(\ell)} \partial h_k^{(\ell)}} a_j^{(\ell-1)} a_l^{(\ell-1)} + \delta_i^{(\ell+1)} \delta_{ik} \frac{\partial a_j^{(\ell-1)}}{\partial W_{kl}^{(\ell)}}.$$

Since activations are applied elementwise, $h_i^{(\ell)}$ and $h_k^{(\ell)}$ affect disjoint sets of downstream activations when $i \neq k$, giving $\partial^2 f / \partial h_i^{(\ell)} \partial h_k^{(\ell)} = 0$. The Kronecker delta $\delta_{ik} = 0$ eliminates the second term. Similar reasoning applies to weight-bias and bias-bias entries, so all off-diagonal entries vanish.

For diagonal entries:

$$\frac{\partial^2 f}{\partial (W_{ij}^{(\ell)})^2} = a_j^{(\ell-1)} \frac{\partial}{\partial h_i^{(\ell)}} \left(\delta_i^{(\ell+1)} a_j^{(\ell-1)} \right) = [a_j^{(\ell-1)}]^2 \left[\frac{\partial^2 f}{\partial (h_i^{(\ell)})^2} [\phi'(h_i^{(\ell)})]^2 + \delta_i^{(\ell)} \phi''(h_i^{(\ell)}) \right].$$

Differentiating the backward recursion yields $|\partial^2 f / \partial (h_i^{(\ell)})^2| = \mathcal{O}_{\mathbb{P}}(C_1^{L-\ell})$. Since pre-activations concentrate via Gaussian process limits [15], $|a_j^{(\ell-1)}| = \mathcal{O}_{\mathbb{P}}(\sqrt{\log n})$. Combined with $|\phi'| \leq C_1$, $|\phi''| \leq C_2$, and $\log n = o(\sqrt{n_{\min}})$:

$$\left| \frac{\partial^2 f}{\partial (W_{ij}^{(\ell)})^2} \right| = \mathcal{O}_{\mathbb{P}}(C_1^{L-\ell+2} C_2 \log n) = \mathcal{O}_{\mathbb{P}}(1).$$

The spectral norm of each diagonal block equals its maximum entry, so:

$$\sum_{\ell=1}^{L} \|H_{\ell,\ell}\|_{\text{op}} = \mathcal{O}_{\mathbb{P}}(L) = \mathcal{O}_{\mathbb{P}}(1). \tag{38}$$

Combining (35), (37), and (38):

$$\|H\|_{\text{op}} = \mathcal{O}_{\mathbb{P}}(L^2 C_1^L C_2 n_{\min}^{-1/2}). \tag{39}$$

Bounding $\text{Tr}(\Sigma)$. From Theorem 1:

$$\Sigma(\theta_t) = \frac{1}{b} \sum_{i=1}^{b} r_i^2 \nabla_\theta f(x_i) \nabla_\theta f(x_i)^\top + \mathcal{E}(\theta_t).$$

Define $A = \sum_{i=1}^{b} r_i^2 \nabla_\theta f(x_i) \nabla_\theta f(x_i)^\top$. Then:

$$\mathrm{Tr}(\Sigma) = \frac{1}{b}\,\mathrm{Tr}(A) + \mathrm{Tr}(\mathcal{E}(\theta_t)).$$

For the main term:

$$\mathrm{Tr}(A) = \sum_{i=1}^{b} r_i^2 \|\nabla_\theta f(x_i)\|_2^2 = \sum_{i=1}^{b} r_i^2 K(x_i, x_i).$$

By NTK convergence [1], $K(x_i, x_i) = \Theta_{\mathbb{P}}(1)$ uniformly in i. From Lemma 1, $|r_i| = \mathcal{O}_{\mathbb{P}}(1)$. Thus:

$$\frac{1}{b}\,\mathrm{Tr}(A) = \frac{1}{b} \sum_{i=1}^{b} \mathcal{O}_{\mathbb{P}}(1) = \mathcal{O}_{\mathbb{P}}(1).$$

For the error term, by Lemma 4:

$$\mathrm{Tr}(\mathcal{E}(\theta_t)) = \mathcal{O}_{\mathbb{P}}(1).$$

Therefore:

$$\begin{aligned} \mathrm{Tr}(\Sigma) &= \frac{1}{b}\,\mathrm{Tr}(A) + \mathrm{Tr}(\mathcal{E}(\theta_t)) \\ &= \mathcal{O}_{\mathbb{P}}(1) + \mathcal{O}_{\mathbb{P}}(1) = \mathcal{O}_{\mathbb{P}}(1). \end{aligned} \tag{40}$$

Combining (32), (39), and (40):

$$|\mathrm{Tr}(H\eta\Sigma)| \le \eta \cdot \mathcal{O}_{\mathbb{P}}(L^2 C_1^{2L} C_2 n_{\min}^{-1/2}) \cdot \mathcal{O}_{\mathbb{P}}(1) = \mathcal{O}_{\mathbb{P}}(\eta n_{\min}^{-1/2}),$$

where constants depending on fixed depth L and activation bounds C_1, C_2 are absorbed. For negligibility relative to the drift term $|\nabla_\theta f^\top d\theta_t| = \mathcal{O}_{\mathbb{P}}(b^{-1/2})dt$, we require $\eta n_{\min}^{-1/2} = o(b^{-1/2})$, giving the condition $\eta = o(n_{\min}^{1/2}/\sqrt{b})$. ■

Lemma 6 (Martingale Covariance Structure) *The martingale process $N_t(x)$ has covariance:*

$$\mathbb{E}[dN_t(x) dN_t(x')] = K_S(x, x'; \{r_i\})dt$$

Proof By definition, the martingale process is:

$$dN_t(x) = \frac{1}{\sqrt{\eta}} \nabla_\theta f(x; \theta_t)^\top \sqrt{\eta \Sigma(\theta_t)}\, dW_t$$

For two inputs x and x', the covariance is:

$$\begin{aligned} \mathbb{E}[dN_t(x) dN_t(x')] &= \mathbb{E}\left[\frac{1}{\sqrt{\eta}} \nabla_\theta f(x; \theta_t)^\top \sqrt{\eta \Sigma(\theta_t)}\, dW_t \times \frac{1}{\sqrt{\eta}} \nabla_\theta f(x'; \theta_t)^\top \sqrt{\eta \Sigma(\theta_t)}\, dW_t \right] \\ &= \frac{1}{\eta} \nabla_\theta f(x; \theta_t)^\top \eta \Sigma(\theta_t) \nabla_\theta f(x'; \theta_t) \mathbb{E}[dW_t dW_t^\top] \\ &= \nabla_\theta f(x; \theta_t)^\top \Sigma(\theta_t) \nabla_\theta f(x'; \theta_t) dt \end{aligned}$$

where we used $\mathbb{E}[dW_t dW_t^\top] = I dt$ for standard Brownian motion. Substituting the leading-order diagonal term from Theorem 1:

$$\begin{aligned}
\mathbb{E}[dN_t(x)dN_t(x')] &= \nabla_\theta f(x;\theta_t)^\top \left(\frac{1}{b}\sum_{i=1}^{b} r_i^2(t) \nabla_\theta f(x_i;\theta_t) \nabla_\theta f(x_i;\theta_t)^\top \right) \nabla_\theta f(x';\theta_t) dt \\
&= \frac{1}{b}\sum_{i=1}^{b} r_i^2(t) \left(\nabla_\theta f(x;\theta_t)^\top \nabla_\theta f(x_i;\theta_t) \right) \left(\nabla_\theta f(x_i;\theta_t)^\top \nabla_\theta f(x';\theta_t) \right) dt \\
&= \frac{1}{b}\sum_{i=1}^{b} r_i^2(t) K(x,x_i) K(x_i,x') dt \\
&= K_S(x,x';\{r_i\}) dt
\end{aligned}$$

■

Theorem 2 (SNTK Regression Formulation) *Under the assumptions of Theorem 1, there exists a continuous process $N_t(x)$ such that*

$$df_t(x) = -\sum_{i=1}^{N} r_i(t) K(x,x_i) dt + \sqrt{\eta} dN_t(x)$$

where $N_t(x)$ is a continuous martingale with quadratic covariation

$$\langle N(\cdot,x), N(\cdot,x') \rangle_t = \int_0^t K_S(x,x';\{r_i(s)\}) ds$$

and K_S is the SNTK: $K_S(x,x';\{r_i\}) = \frac{1}{b}\sum_{i\in\mathcal{B}_t} r_i^2 K(x,x_i)K(x_i,x')$.

Proof Starting from the SGD dynamics:

$$d\theta_t = -\nabla L(\theta_t) dt + \sqrt{\eta \Sigma(\theta_t)} dW_t$$

Applying Itô's lemma to $f_t(x) = f(x;\theta_t)$:

$$df_t(x) = \nabla_\theta f(x;\theta_t)^\top d\theta_t + \frac{1}{2}\mathrm{Tr}\left[\nabla_\theta^2 f(x;\theta_t) \cdot d\theta_t d\theta_t^\top \right]$$

Since $d\theta_t d\theta_t^\top = \eta \Sigma(\theta_t) dt$, by Lemma 5, the Hessian term vanishes in the infinite-width limit:

$$df_t(x) = \nabla_\theta f(x;\theta_t)^\top d\theta_t + \mathcal{O}((n_{\min})^{-1/2})$$

Substituting the SGD dynamics:

$$\begin{aligned}
df_t(x) &= \nabla_\theta f(x;\theta_t)^\top \left(-\nabla L(\theta_t) dt + \sqrt{\eta\Sigma(\theta_t)} dW_t \right) + \mathcal{O}((n_{\min})^{-1/2}) \\
&= -\nabla_\theta f(x;\theta_t)^\top \nabla L(\theta_t) dt + \nabla_\theta f(x;\theta_t)^\top \sqrt{\eta\Sigma(\theta_t)} dW_t + \mathcal{O}((n_{\min})^{-1/2})
\end{aligned}$$

Hereafter, we focus on the leading-order terms and omit the $\mathcal{O}((n_{\min})^{-1/2})$ corrections, which vanish in the infinite-width limit and do not affect the martingale structure of the stochastic component.

For the drift term, using $\nabla L(\theta_t) = \sum_{i=1}^{N} r_i(t) \nabla_\theta f(x_i; \theta_t)$:

$$-\nabla_\theta f(x; \theta_t)^\top \nabla L(\theta_t) = -\sum_{i=1}^{N} r_i(t) K(x, x_i)$$

For the stochastic term, consider the process:

$$\tilde{N}_t(x) = \frac{1}{\sqrt{\eta}} \int_0^t \nabla_\theta f(x; \theta_s)^\top \sqrt{\eta \Sigma(\theta_s)} dW_s$$

This is a continuous local martingale with differential $d\tilde{N}_t(x) = \frac{1}{\sqrt{\eta}} \nabla_\theta f(x; \theta_t)^\top \sqrt{\eta \Sigma(\theta_t)} dW_t$. To compute its covariance structure, we use the fact that for standard Brownian motion $\mathbb{E}[dW_t dW_t^\top] = I dt$:

$$\begin{aligned}
\mathbb{E}[d\tilde{N}_t(x) d\tilde{N}_t(x')] &= \mathbb{E}\left[\frac{1}{\sqrt{\eta}} \nabla_\theta f(x; \theta_t)^\top \sqrt{\eta \Sigma(\theta_t)} dW_t \times \frac{1}{\sqrt{\eta}} \nabla_\theta f(x'; \theta_t)^\top \sqrt{\eta \Sigma(\theta_t)} dW_t \right] \\
&= \frac{1}{\eta} \nabla_\theta f(x; \theta_t)^\top \eta \Sigma(\theta_t) \mathbb{E}[dW_t dW_t^\top] \sqrt{\eta \Sigma(\theta_t)}^\top \nabla_\theta f(x'; \theta_t) \\
&= \frac{1}{\eta} \nabla_\theta f(x; \theta_t)^\top \eta \Sigma(\theta_t) \cdot I \cdot \eta \Sigma(\theta_t) \nabla_\theta f(x'; \theta_t) dt \\
&= \nabla_\theta f(x; \theta_t)^\top \Sigma(\theta_t) \nabla_\theta f(x'; \theta_t) dt
\end{aligned}$$

Using the leading-order term from Theorem 1, this becomes:

$$\begin{aligned}
\mathbb{E}[d\tilde{N}_t(x) d\tilde{N}_t(x')] &= \nabla_\theta f(x; \theta_t)^\top \left(\frac{1}{b} \sum_{i=1}^{n} r_i^2(t) \nabla_\theta f(x_i; \theta_t) \nabla_\theta f(x_i; \theta_t)^\top \right) \nabla_\theta f(x'; \theta_t) dt \\
&= \frac{1}{b} \sum_{i=1}^{n} r_i^2(t) K(x, x_i) K(x_i, x') dt \\
&= K_S(x, x'; \{r_i(t)\}) dt
\end{aligned}$$

Setting $N_t(x) = \tilde{N}_t(x)$, we obtain:

$$df_t(x) = -\sum_{i=1}^{N} r_i(t) K(x, x_i) dt + \sqrt{\eta} dN_t(x)$$

The quadratic covariation follows by integration:

$$\langle N(\cdot, x), N(\cdot, x') \rangle_t = \int_0^t K_S(x, x'; \{r_i(s)\}) ds$$

■