High-Dimensional Limit of Stochastic Gradient Flow via Dynamical Mean-Field Theory

Sota Nishiyama SNISHIYAMA@G.ECC.U-TOKYO.AC.JP
Graduate School of Arts and Sciences, The University of Tokyo, Tokyo, Japan
RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan

Masaaki Imaizumi IMAIZUMI@G.ECC.U-TOKYO.AC.JP
Graduate School of Arts and Sciences, The University of Tokyo, Tokyo, Japan
RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan
Graduate School of Science, Kyoto University, Kyoto, Japan

Abstract

We study the high-dimensional dynamics of multi-pass stochastic gradient descent (SGD) with small batch sizes. Existing asymptotic descriptions of SGD primarily cover online learning, large-batch regimes, or linear models. We address this gap by analyzing a continuous-time stochastic differential equation, called *stochastic gradient flow* (SGF), which approximates small-batch multi-pass SGD. In the proportional limit $n, d \to \infty$ with $n/d \to \delta$, we derive a closed dynamical mean-field theory (DMFT) system and prove that it characterizes the asymptotic empirical distribution of the SGF parameters. The framework applies to nonlinear models including generalized linear models and two-layer neural networks, and recovers several existing high-dimensional descriptions, including noiseless gradient flow, online SGD, and high-dimensional linear regression.

1. Introduction

Stochastic gradient descent (SGD) [49] is the standard training algorithm for modern machine learning models. Its stochasticity can qualitatively change the training dynamics relative to noiseless gradient descent or gradient flow, affecting both optimization and the learned solution [34, 35]. Understanding these effects in high-dimensional regimes is therefore a central problem in the theory of learning dynamics.

A common approach is to derive low-dimensional equations that characterize the macroscopic behavior of high-dimensional training dynamics. For noiseless gradient descent and gradient flow, random matrix theory and dynamical mean-field theory (DMFT) have led to precise asymptotic descriptions [1, 11, 12, 15, 22, 41]. For SGD, however, existing frameworks remain restricted. Online or one-pass SGD has been extensively studied [8, 29, 51], but does not capture repeated reuse of a finite dataset. DMFT approaches for multi-pass SGD typically require batch sizes proportional to the dataset size [27, 41], while analyses of multi-pass small-batch SGD are largely limited to least-squares linear models through homogenized SGD and random matrix methods [44, 45]. Thus, a high-dimensional theory for multi-pass small-batch SGD in nonlinear models is still missing. A detailed review of related work is deferred to Appendix A.2.

In this work, we address this gap by studying a stochastic differential equation called *stochastic gradient flow* (SGF), which approximates multi-pass SGD with small batch sizes. We derive a DMFT

© S. Nishiyama & M. Imaizumi.

system that characterizes the high-dimensional limit of SGF and clarifies its relationship to several existing descriptions of SGD dynamics. Table 1 compares our framework with prior approaches.

Our contributions are as follows.

DMFT for stochastic gradient flow. We derive a closed system of low-dimensional stochastic processes that characterizes the empirical distribution of SGF parameters in the proportional asymptotic regime $n, d \to \infty$ with $n/d \to \delta$. We prove existence and uniqueness of the DMFT solution and show convergence of the empirical parameter and prediction distributions to this solution. The framework applies to a broad class of nonlinear models, including generalized linear models and two-layer neural networks.

Unification of existing analyses. We show that the DMFT equation for the SGF reduces to known descriptions in several limits: noiseless gradient flow when $\tau = 0$, online SGD dynamics in the infinite data limit $\delta \to \infty$, and the Volterra equations for high-dimensional linear regression [44, 45].

Concurrent with our work, Fan and Wang [23] derived a related DMFT characterization for high-dimensional SGD and SGF. Our work was developed independently and gives a complementary formulation; see the discussion in Appendix A.2.

Table 1: Comparison of frameworks for high-dimensional analysis of SGD.

	Online SGD [8]	**Previous DMFT** [41]	**HSGD** [45]	**Ours**
Multi-pass	✗	✓	✓	✓
Small batch sizes	✓	✗	✓	✓
Nonlinear models	✓	✓	✗	✓

2. Setup

Notations are summarized in Appendix A.1.

We study the high-dimensional dynamics of a *stochastic gradient flow* (SGF), a continuous-time approximation of small-batch multi-pass SGD. Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ be the data matrix, $\boldsymbol{z} \in \mathbb{R}^n$ an independent noise vector, and $\boldsymbol{\theta}^t \in \mathbb{R}^{d\times m}$ the parameter. For functions $h_t : \mathbb{R}^m \to \mathbb{R}^m$ and $\ell_t : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^m$ applied row-wise, define

$$\mathrm{d}\boldsymbol{\theta}^t = -\left(h_t(\boldsymbol{\theta}^t) + \frac{1}{\delta}\boldsymbol{X}^\mathsf{T}\ell_t(\boldsymbol{r}^t;\boldsymbol{z})\right)\mathrm{d}t + \sqrt{\frac{\tau}{\delta}}\sum_{i=1}^n \boldsymbol{x}_i\ell_t(r_i^t;z_i)^\mathsf{T}\,\mathrm{d}B_i^t\,, \quad \boldsymbol{r}^t = \boldsymbol{X}\boldsymbol{\theta}^t\,, \tag{1}$$

where $\boldsymbol{B}^t = (B_i^t)_{i\in[n]}$ is a Brownian motion in $\mathbb{R}^n$, $\delta > 0$ is the sample-to-dimension ratio $\delta := n/d$, and $\tau > 0$ is the *temperature* parameter controlling the noise magnitude. The SGF (1) can be seen as a continuous-time approximation of mini-batch SGD with learning rate η and batch size B in the regime $n, d \to \infty$, $B = o(n)$, where $\tau = \eta/B$. The time variable t corresponds to the number of SGD steps k via $t = \eta k/d$, so that $t = O(1)$ time corresponds to $k = O(d/\eta)$ steps. The noiseless case $\tau = 0$ reduces to the gradient flow dynamics studied in Celentano et al. [15].

The SGF (1) is obtained by matching the leading drift and covariance of one step of mini-batch SGD. The precise SGD recursion and moment calculation are deferred to Appendix A.3.

We analyze SGF as a tractable continuous model of small-batch SGD; the use of SDE approximations for SGD is standard in the literature and discussed in Appendix A.2.

We note that the case of planted models where the data label is generated by a planted signal $\boldsymbol{\theta}^*$ can be handled by augmenting the parameter $\boldsymbol{\theta}$ with $\boldsymbol{\theta}^*$; see Appendix E.1.

Assumptions. We assume independent centered sub-Gaussian covariates $\boldsymbol{X}$ with variance $1/d$, proportional asymptotics $n/d \to \delta \in (0, \infty)$, and empirical convergence of the initialization and noise. We further assume Lipschitz regularity of h_t, ℓ_t and their relevant derivatives, with polynomial growth of ℓ_t. The precise assumptions are stated in Appendix A.4 as Assumptions A.1 and A.2.

3. Main Result

We now state the DMFT characterization of the SGF (1). The DMFT equation describes the high-dimensional dynamics through low-dimensional effective processes $\theta^t, r^t \in \mathbb{R}^m$ together with correlation and response kernels $C_\theta, C_\ell, R_\theta, R_\ell : \mathbb{R}^2_{\geq 0} \to \mathbb{R}^{m\times m}$ and $\Gamma : \mathbb{R}_{\geq 0} \to \mathbb{R}^{m\times m}$. Informally, the system is

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\theta^t &= u^t - (h_t(\theta^t) + \Gamma(t)\theta^t) - \int_0^t R_\ell(t,s)\theta^s \,\mathrm{d}s\,, \quad u \sim \mathsf{GP}(0, C_\ell/\delta)\,,\\
r^t &= w^t - \frac{1}{\delta}\int_0^t R_\theta(t,s)\ell_s(r^s;z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\,, \quad w \sim \mathsf{GP}(0, C_\theta)\,,\\
C_\theta(t,t') &= \mathbb{E}[\theta^t\theta^{t'\mathsf{T}}]\,, \quad R_\theta(t,t') = \mathbb{E}\left[\frac{\partial\theta^t}{\partial u^{t'}}\right],\\
C_\ell(t,t') &= \mathbb{E}[\ell_t(r^t;z)(1+\sqrt{\tau\delta}\dot{B}^t)\ell_{t'}(r^{t'};z)^\mathsf{T}(1+\sqrt{\tau\delta}\dot{B}^{t'})]\,,\\
R_\ell(t,t') &= \mathbb{E}\left[\frac{\partial\ell_t(r^t;z)}{\partial w^{t'}}\right], \quad \Gamma(t) = \mathbb{E}[\nabla_r\ell_t(r^t;z)]\,,
\end{aligned}
\tag{2}
$$

with $R_\theta(t,t') = R_\ell(t,t') = 0$ for $t < t'$. The Brownian motion $B^t \in \mathbb{R}$ is one-dimensional and captures the stochasticity inherited from SGF. The equation is self-consistent: the law of (θ^t, r^t) determines the kernels, which in turn determine the law of (θ^t, r^t). The display above is informal because it uses functional derivatives and the white noise process $\dot{B}^t$; the rigorous formulation, replacing C_ℓ by a well-defined covariance object Σ_ℓ, is given in Appendix B.

When $\tau = 0$, the Brownian term vanishes and (2) reduces to the DMFT equation for noiseless gradient flow in Celentano et al. [15]. Thus, the present equation extends gradient-flow DMFT to stochastic dynamics.

Theorem 3.1 (Existence and uniqueness of the DMFT equation) *Suppose Assumptions A.1 and A.2 hold. Then there exists $T_* > 0$ such that, for every $T \in [0, T_*]$, the DMFT system admits a unique bounded fixed point $(C_\theta, \Sigma_\ell, R_\theta, R_\ell, \Gamma)$ on $[0, T]$. Moreover, the effective processes θ^t and r^t have continuous sample paths.*

If either $\tau = 0$ or $\nabla_r^2\ell_t(r;z) = 0$ for all t, r, z, then the result holds globally for all $T > 0$.

The global cases include noiseless gradient flow and SGF under quadratic losses, the latter covering least-squares linear regression. For general nonlinear ℓ_t with $\tau > 0$, the theorem gives local well-posedness; a quantitative lower bound on T_* and the proof are deferred to Appendix C.

Theorem 3.2 (DMFT characterization of SGF) *Suppose Assumptions A.1 and A.2 hold, and let T_* be as in Theorem 3.1. For any $T \in [0, T_*]$, $L \in \mathbb{N}$, and $0 \le t_1 < \cdots < t_L \le T$,*

$$\operatorname*{p-lim}_{n,d\to\infty} W_2\Big(\hat{\mathsf{P}}(\boldsymbol{\theta}^{t_1}, \ldots, \boldsymbol{\theta}^{t_L}), \mathsf{P}(\theta^{t_1}, \ldots, \theta^{t_L})\Big) = 0, \tag{3}$$

$$\operatorname*{p-lim}_{n,d\to\infty} W_2\Big(\hat{\mathsf{P}}(\boldsymbol{r}^{t_1}, \ldots, \boldsymbol{r}^{t_L}, \boldsymbol{z}), \mathsf{P}(r^{t_1}, \ldots, r^{t_L}, z)\Big) = 0. \tag{4}$$

Theorem 3.2 states that the empirical distribution of SGF parameters and predictions converges to the law of the effective DMFT process at any finite collection of times. Equivalently, empirical averages of test functions of the coordinates of $\boldsymbol{\theta}^t$ and $\boldsymbol{r}^t$ converge to expectations under the DMFT law. This extends the gradient-flow result of Celentano et al. [15] to stochastic gradient flow.

The proof follows the AMP-based strategy of Celentano et al. [15]: discretize the SGF, map the resulting iteration to an AMP recursion, identify its state evolution with the discretized DMFT system, and pass to the continuous-time limit. The stochastic setting requires an additional truncation argument for Brownian increments and Stein's lemma to derive the stochastic correction terms. Details are given in Appendix D.

4. Applications and Special Cases

We highlight two consequences of the DMFT characterization: its online SGD limit and its reduction to known Volterra equations in linear regression.

Infinite-data limit and online SGD. In the infinite-data limit $\delta \to \infty$, the response terms vanish and the DMFT equation simplifies to a low-dimensional SDE given in Equation (316) in Appendix E.2. This limit corresponds to online SGD: when n is much larger than the number of updates proportional to d, a constant-size mini-batch effectively consists of fresh samples. In Appendix E.2, we show that (316) recovers the online SGD characterization for linear regression derived in Wang et al. [56].

Linear regression. As a tractable special case, consider high-dimensional linear regression

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|_2^2, \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{z}, \tag{5}$$

with zero initialization. This corresponds to the planted model with $m = 1$, $h_t = 0$, and $\ell_t(r, r^*; z) = r - r^* - z$. Since ℓ_t is linear in r, Theorem 3.1 gives global well-posedness of the DMFT equation.

Let the training and test errors of the parameter $\boldsymbol{\theta}$ be given by

$$\mathscr{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta} - y_i)^2\,, \quad \mathscr{R}(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{x},y)}[(\boldsymbol{x}^\mathsf{T}\boldsymbol{\theta} - y)^2]\,. \tag{6}$$

Solving the DMFT equation yields a closed Volterra system for the limiting train and test errors:

$$\mathscr{L}(t) = \mathscr{L}_0(t) + \tau\int_0^t H_2(t-s)\mathscr{L}(s)\,\mathrm{d}s\,, \quad \mathscr{R}(t) = \mathscr{R}_0(t) + \tau\int_0^t H_1(t-s)\mathscr{L}(s)\,\mathrm{d}s\,, \tag{7}$$

where $H_i(t) = \int x^i e^{-2xt}\,\mathrm{d}\mu_{\mathrm{MP}}(x)$, and μ_{MP} is the Marchenko–Pastur law with aspect ratio δ. Here $\mathscr{L}_0$ and $\mathscr{R}_0$ are the noiseless limiting train and test errors; their explicit forms are given in Appendix E.3.

More precisely, for any finite collection of times $t_1, \dots, t_L$, we have the following convergence in probability as $n, d \to \infty$:

$$\operatorname*{p-lim}_{n,d\to\infty} \max_{l=1,\dots,L} |\mathscr{L}(\boldsymbol{\theta}^{t_l}) - \mathscr{L}(t_l)| = 0\,, \quad \operatorname*{p-lim}_{n,d\to\infty} \max_{l=1,\dots,L} |\mathscr{R}(\boldsymbol{\theta}^{t_l}) - \mathscr{R}(t_l)| = 0\,, \tag{8}$$

The scalar equation for $\mathscr{L}$ can be solved independently, after which $\mathscr{R}$ is obtained from the second equation in (7). These equations coincide, up to time rescaling, with the homogenized SGD Volterra equations of Paquette et al. [44, 45]. Thus, our DMFT framework recovers the known exact linear regression theory while extending naturally to nonlinear models.

We also note a close connection to the Volterra equations derived by Avidan et al. [6] for Langevin dynamics of infinitely wide neural networks with general data. In their setting, the mean training output satisfies a closed Volterra equation, while the mean test output is obtained through an integral transform of the training output. This structure is analogous to Equation (7), where the limiting training error satisfies a closed Volterra equation and the limiting test error is obtained from the training-error trajectory.

5. Numerical Simulations

We illustrate the DMFT prediction on a numerical simulation with a nonlinear model: logistic regression trained by multi-pass small-batch SGD. The data are sampled as $\boldsymbol{x}_i \sim \mathsf{N}(0, \boldsymbol{I}_d/d)$ with $d = 1024$, $n = 2048$ ($\delta = 2$), and labels $y_i = \operatorname{sign}(\boldsymbol{\theta}^{*\mathsf{T}}\boldsymbol{x}_i + z_i)$, $\boldsymbol{\theta}^* \sim \mathsf{N}(0, \boldsymbol{I}_d)$, $z_i \sim \mathsf{N}(0, 0.01)$. We train with the regularized logistic objective

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\theta})) + \frac{\lambda}{2d}\|\boldsymbol{\theta}\|_2^2, \qquad \lambda = 0.01. \tag{9}$$

SGD is run with batch size $B = 10$ and varying learning rates, corresponding to different temperatures $\tau = \eta/B$. We compare the average train and test zero-one errors over 10 independent trials with the numerical solution of the DMFT equation. Details of the time discretization used to solve the DMFT equation are deferred to Appendix F.

Figure 1 shows that the DMFT prediction closely tracks the SGD train and test error dynamics across temperatures. Although the logistic model used here is not formally covered by our assumptions because of the non-differentiability induced by $y = \operatorname{sign}(r^* + z)$, the agreement suggests that the DMFT equations remain predictive beyond the strictly covered setting.

6. Discussion

In summary, we derived a DMFT equation characterizing the high-dimensional limit of SGF, a continuous-time approximation of multi-pass small-batch SGD. The resulting theory covers nonlinear models, recovers known limits such as noiseless gradient flow and linear-regression Volterra equations, and gives accurate predictions in the logistic-regression experiment above. Future directions include proving sharper SGF–SGD approximation results, analyzing the long-time behavior of the DMFT equations, and applying the framework to more detailed nonlinear models.

A limitation of our work is that the rigorous high-dimensional characterization is for SGF, not for discrete-time mini-batch SGD itself. The SGF is obtained by matching the leading drift and covariance of one SGD step, and is expected to approximate SGD in the small-learning-rate regime.

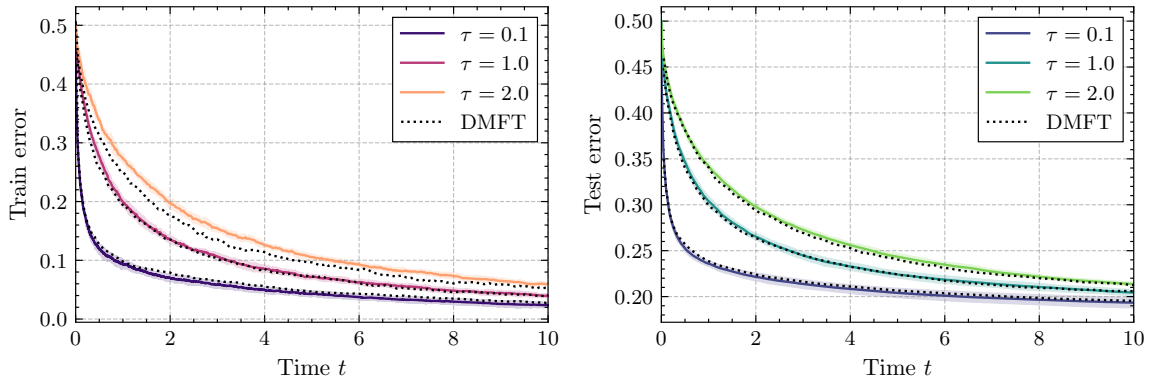

Figure 1: Train (left) and test (right) error dynamics for logistic regression at different temperatures $\tau = \eta/B$. Solid curves show the average over 10 SGD trials with $d = 1024$ and $n = 2048$; shaded regions show one standard deviation. Dotted curves show the DMFT predictions.

However, this approximation is not an exact equivalence in general: discrete SGD noise can differ qualitatively from a continuous-time diffusion approximation, and SGF and SGD may have different high-dimensional limits [23]. Thus, our results should be interpreted as an exact theory of the SGF approximation to small-batch multi-pass SGD, rather than as a complete exact theory of finite-step SGD. Nevertheless, the SGF approximation remains useful for understanding the anisotropic and state-dependent nature of the SGD noise, a viewpoint supported empirically by Li et al. [39]. Further study of SGF may reveal additional insights into SGD dynamics and its implicit biases.

References

[1] Madhu S. Advani, Andrew M. Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020. doi: 10.1016/j.neunet.2020.08.022.

[2] Elisabeth Agoritsas, Giulio Biroli, Pierfrancesco Urbani, and Francesco Zamponi. Out-of-equilibrium dynamical mean-field equations for the perceptron model. *Journal of Physics A: Mathematical and Theoretical*, 51(8):085002, 2018. doi: 10.1088/1751-8121/aaa68d.

[3] Alnur Ali, Edgar Dobriban, and Ryan Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 233–244, 2020.

[4] Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional & mean-field dynamics to dimensionless ODEs: a unifying approach to SGD in two-layers networks. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195, pages 1199–1227, 2023.

[5] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51, 2021.

[6] Yehonatan Avidan, Qianyi Li, and Haim Sompolinsky. Unified theoretical framework for wide neural network learning dynamics. *Physical Review E*, 111(4):045310, 2025. doi: 10.1103/PhysRevE.111.045310.

[7] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2): 764–785, 2011. doi: 10.1109/TIT.2010.2094817.

[8] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for SGD: effective dynamics and critical scaling. In *Advances in Neural Information Processing Systems*, volume 35, pages 25349–25362, 2022.

[9] Marc A. Berger and Victor J. Mizel. Volterra equations with Itô integrals—I. *Journal of Integral Equations*, 2(3):187–245, 1980.

[10] M Biehl and H Schwarze. Learning by on-line gradient descent. *Journal of Physics A: Mathematical and General*, 28(3):643–656, 1995. doi: 10.1088/0305-4470/28/3/018.

[11] Antoine Bodin and Nicolas Macris. Model, sample, and epoch-wise descents: exact solution of gradient flow in the random feature model. In *Advances in Neural Information Processing Systems*, volume 34, pages 21605–21617, 2021.

[12] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 4345–4382, 2024.

[13] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(8):084002, 2025. doi: 10.1088/1742-5468/adefb1.

[14] Michael Celentano, Andrea Montanari, and Yuchen Wu. The estimation error of general first order methods. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 1078–1141, 2020.

[15] Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021.

[16] Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10, 2018. doi: 10.1109/ITA.2018.8503224.

[17] Xiang Cheng, Dong Yin, Peter Bartlett, and Michael Jordan. Stochastic gradient and Langevin processes. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 1810–1819, 2020.

[18] Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: an ODE for SGD learning dynamics on GLMs and multi-index models. *Information and Inference: A Journal of the IMA*, 13(4):iaae028, 2024. doi: 10.1093/imaiai/iaae028.

[19] Andrea Crisanti, Heinz Horner, and H-J Sommers. The spherical p-spin interaction spin-glass model: the dynamics. *Zeitschrift für Physik B Condensed Matter*, 92(2):257–271, 1993.

[20] L. F. Cugliandolo and J. Kurchan. Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model. *Physical Review Letters*, 71(1):173–176, 1993. doi: 10.1103/PhysRevLett.71.173.

[21] Leticia F. Cugliandolo. Recent applications of dynamical mean-field methods. *Annual Review of Condensed Matter Physics*, 15(1):177–213, 2024. doi: 10.1146/annurev-conmatphys-040721-022848.

[22] Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborova, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: breaking the curse of information and leap exponents. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 9991–10016, 2024.

[23] Zhou Fan and Leda Wang. High-dimensional learning dynamics of multi-pass stochastic gradient descent in multi-index models. *arXiv preprint arXiv:2601.21093*, 2026.

[24] Zhou Fan, Justin Ko, Bruno Loureiro, Yue M. Lu, and Yandi Shen. Dynamical mean-field analysis of adaptive Langevin diffusions: propagation-of-chaos and convergence of the linear response. *arXiv preprint arXiv:2504.15556*, 2025.

[25] Zhou Fan, Justin Ko, Bruno Loureiro, Yue M. Lu, and Yandi Shen. Dynamical mean-field analysis of adaptive Langevin diffusions: replica-symmetric fixed point and empirical Bayes. *arXiv preprint arXiv:2504.15558*, 2025.

[26] Oliver Y. Feng, Ramji Venkataramanan, Cynthia Rush, and Richard J. Samworth. A unifying tutorial on approximate message passing. *Foundations and Trends in Machine Learning*, 15(4): 335–536, 2022. doi: 10.1561/2200000092.

[27] Cédric Gerbelot, Emanuele Troiani, Francesca Mignacco, Florent Krzakala, and Lenka Zdeborová. Rigorous dynamical mean-field theory for stochastic gradient descent methods. *SIAM Journal on Mathematics of Data Science*, 6(2):400–427, 2024. doi: 10.1137/23M1594388.

[28] Margalit Glasgow. SGD finds then tunes features in two-layer neural networks with near-optimal sample complexity: a case study in the XOR problem. In *International Conference on Learning Representations*, 2024.

[29] Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[30] Qiyang Han. Entrywise dynamics and universality of general first order methods. *The Annals of Statistics*, 53(4), 2025. doi: 10.1214/25-AOS2544.

[31] Qiyang Han and Masaaki Imaizumi. Precise gradient descent training dynamics for finite-width multi-layer neural networks. *arXiv preprint arXiv:2505.04898*, 2025.

[32] Yuma Ichikawa and Koji Hukushima. Learning dynamics in linear VAE: posterior collapse threshold, superfluous latent space pitfalls, and speedup with KL annealing. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pages 1936–1944, 2024.

[33] Yuma Ichikawa, Shuhei Kashiwamura, and Ayaka Sakata. High-dimensional learning dynamics of quantized models with straight-through estimator. *arXiv preprint arXiv:2510.10693*, 2025.

[34] Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in SGD. *arXiv preprint arXiv:1711.04623*, 2017.

[35] 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. In *International Conference on Learning Representations*, 2017.

[36] Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer Berlin Heidelberg, 1992. doi: 10.1007/978-3-662-12616-5.

[37] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 2101–2110, 2017.

[38] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms I: mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019.

[39] Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling sgd with stochastic differential equations (sdes). In *Advances in Neural Information Processing Systems*, volume 34, pages 12712–12725, 2021.

[40] 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.

[41] Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification. In *Advances in Neural Information Processing Systems*, volume 33, pages 9540–9550, 2020.

[42] Andrea Montanari and Pierfrancesco Urbani. Dynamical decoupling of generalization and overfitting in large two-layer networks. In *Advances in Neural Information Processing Systems*, 2025.

[43] Sota Nishiyama and Masaaki Imaizumi. Precise dynamics of diagonal linear networks: a unifying analysis by dynamical mean-field theory. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

[44] Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. SGD in the large: average-case analysis, asymptotics, and stepsize criticality. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pages 3548–3626, 2021.

[45] Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Homogenization of SGD in high-dimensions: exact dynamics and generalization properties. *Mathematical Programming*, 214(1-2):1–90, 2025. doi: 10.1007/s10107-024-02171-3.

[46] Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. In *Advances in Neural Information Processing Systems*, volume 34, pages 29218–29230, 2021.

[47] Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason D. Lee. Emergence and scaling laws in SGD learning of shallow neural networks. In *Advances in Neural Information Processing Systems*, 2025.

[48] P Riegler and M Biehl. On-line backpropagation in two-layered neural networks. *Journal of Physics A: Mathematical and General*, 28(20):L507–L513, 1995. doi: 10.1088/0305-4470/28/20/002.

[49] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[50] David Saad and Sara A. Solla. Exact solution for on-line learning in multilayer neural networks. *Physical Review Letters*, 74(21):4337–4340, 1995. doi: 10.1103/PhysRevLett.74.4337.

[51] David Saad and Sara A. Solla. On-line learning in soft committee machines. *Physical Review E*, 52(4):4225–4243, 1995. doi: 10.1103/PhysRevE.52.4225.

[52] Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Marvels and pitfalls of the Langevin algorithm in noisy high-dimensional inference. *Physical Review X*, 10(1):011057, 2020. doi: 10.1103/PhysRevX.10.011057.

[53] H. Sompolinsky and Annette Zippelius. Dynamic theory of the spin-glass phase. *Physical Review Letters*, 47(5):359–362, 1981. doi: 10.1103/PhysRevLett.47.359.

[54] H. Sompolinsky and Annette Zippelius. Relaxational dynamics of the Edwards-Anderson model and the mean-field theory of spin-glasses. *Physical Review B*, 25(11):6860–6875, 1982. doi: 10.1103/PhysRevB.25.6860.

[55] Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pages 23244–23255, 2022.

[56] Chuang Wang, Jonathan Mattingly, and Yue M. Lu. Scaling limit: exact and tractable analysis of online learning algorithms with applications to regularized regression and PCA. *arXiv preprint arXiv:1712.04332*, 2017.

[57] Chuang Wang, Hong Hu, and Yue Lu. A solvable high-dimensional model of GAN. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[58] Tianhao Wang, Xinyi Zhong, and Zhou Fan. Universality of approximate message passing algorithms and tensor networks. *The Annals of Applied Probability*, 34(4), 2024. doi: 10.1214/24-AAP2056.

Contents

1 Introduction **1**

2 Setup **2**

3 Main Result **3**

4 Applications and Special Cases **4**

5 Numerical Simulations **5**

6 Discussion **5**

A Preliminaries **12**
A.1 Notation . . . 12
A.2 Related Works . . . 12
A.3 Derivation of the SGF as an Approximation of SGD . . . 14
A.4 Formal Statement of the Assumptions . . . 15

B Definition of the DMFT Equation **15**
B.1 Rigorous Definition of the DMFT System . . . 16
B.2 Discretized DMFT System . . . 17
B.3 Correspondence to the Informal Definition . . . 18

C Proof of Theorem 3.1 **20**
C.1 Admissible Spaces $\mathcal{S}_\theta(T)$ and $\mathcal{S}_\ell(T)$ . . . 21
C.2 Equipping Metrics on $\mathcal{S}_\theta(T)$ and $\mathcal{S}_\ell(T)$ . . . 23
C.3 Proof of Lemma C.4 . . . 24
C.3.1 Construction of the Admissible Spaces . . . 25
C.3.2 $\mathcal{T}_{\theta\to\ell}$ maps $\mathcal{S}_\theta$ into $\mathcal{S}_\ell$. . . . 26
C.3.3 $\mathcal{T}_{\ell\to\theta}$ maps $\mathcal{S}_\ell$ into $\mathcal{S}_\theta$. . . . 29
C.3.4 A Rough Estimate of T_* . . . 30
C.4 Proof of Lemma C.6 . . . 31
C.5 Proof of Lemma C.7 . . . 35

D Proof of Theorem 3.2 **37**
D.1 Proof of Lemma D.1 . . . 39
D.2 Proof of Lemma D.2 . . . 41
D.2.1 Reduction to AMP . . . 41
D.2.2 Mapping the state evolution to DMFT . . . 44
D.3 Proof of Lemma D.3 . . . 46
D.3.1 Proof of Lemma D.6 . . . 48
D.3.2 Proof of Lemma D.7 . . . 53

E Details of Applications and Special Cases **53**
E.1 Planted Models . . . 53
E.1.1 Setup for Planted Models . . . 53
E.1.2 The DMFT Equation . . . 54
E.1.3 Proof of Corollary E.1 . . . 55
E.2 Infinite Data Limit . . . 57
E.2.1 Derivation of the Reduced DMFT Equation . . . 57
E.2.2 Example: Linear Regression . . . 58
E.3 Linear Regression . . . 58
E.3.1 Simplifying the DMFT Equations . . . 60
E.3.2 Solving the DMFT Equations . . . 64

F Details of Numerical Simulations **70**

G Discretization of SDEs in High Dimensions **71**

Appendix A. Preliminaries

A.1. Notation

$\boldsymbol{I}_d \in \mathbb{R}^{d\times d}$ denotes the $d \times d$ identity matrix. $\mathbf{1}_d \in \mathbb{R}^d$ denotes the all-ones vector $\mathbf{1}_d = (1, \dots, 1)^\mathsf{T}$. $\mathsf{GP}(0, Q)$ denotes a centered Gaussian process with covariance kernel Q. $\mathbb{I}(\cdot)$ denotes the indicator function that returns 1 if the argument is true and 0 otherwise. $\|\cdot\|_2$ denotes the ℓ_2 norm for vectors and the 2-operator norm for matrices. $\|\cdot\|_\mathrm{F}$ denotes the Frobenius norm for matrices. W_p denotes the p-Wasserstein distance between probability measures. $\mathrm{p\text{-}lim}$ denotes convergence in probability. $\mathsf{P}(X)$ denotes the law of a random variable X. $\mathsf{P}(X, Y)$ denotes the joint law of random variables X and Y. For a matrix $\boldsymbol{x} \in \mathbb{R}^{d\times m}$, we denote by $\hat{\mathsf{P}}(\boldsymbol{x})$ the empirical distribution of its rows, i.e., $\hat{\mathsf{P}}(\boldsymbol{x}) = \frac{1}{d}\sum_{i=1}^d \delta_{x_i}$. Similarly, $\hat{\mathsf{P}}(\boldsymbol{x}, \boldsymbol{y})$ denotes the empirical joint distribution of the rows of $\boldsymbol{x}$ and $\boldsymbol{y}$, i.e., $\hat{\mathsf{P}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{d}\sum_{i=1}^d \delta_{x_i, y_i}$. $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm (ψ_2-Orlicz norm).

A.2. Related Works

One-pass SGD. The study of online SGD using low-dimensional ODEs was pioneered in the statistical physics literature on learning in perceptrons and two-layer neural networks [10, 48, 50, 51]. They derived a closed system of low-dimensional ODEs for macroscopic parameters such as the correlation between the student and teacher weights and analyzed their dynamics, which closely approximates the original online SGD dynamics in high dimensions. The ODEs typically involve correction terms that account for the stochasticity of the dynamics, and analyzing these ODEs provides insights into how the noise affects the training dynamics.

Recently, these works have been put on a rigorous footing by Goldt et al. [29] using techniques developed in Wang et al. [56, 57]. Ben Arous et al. [8] extended these techniques to more general models and general scaling of learning rates. Analysis of online SGD in high dimensions using similar techniques has been applied to a wide range of models due to its versatility and simplicity [4, 5, 18, 28, 32, 33, 47, 55].

Multi-pass SGD. DMFT has recently gained attention as a powerful framework for analyzing high-dimensional random dynamics, including multi-pass GD and SGD, by reducing them to low-dimensional effective processes. DMFT was originally developed in spin glass theory [19, 20, 53, 54] and has been applied to analyzing various high-dimensional optimization dynamics [2, 21, 52].

In the context of SGD dynamics, Mignacco et al. [41] derived DMFT equations for multi-pass gradient flow and SGD in shallow neural networks, heuristically using statistical physics techniques. To avoid the problem of vanishing stochasticity in the continuous-time limit $\eta \to 0$, they considered a variant of SGD called *persistent SGD* to retain nontrivial noise in the continuous-time limit. Their analysis depends on the batch size growing proportionally to the number of samples. In contrast, we work with a stochastic gradient flow which approximates mini-batch SGD with small (sublinear) batch sizes compared to the number of samples, which is a common setting in practice.

There are several rigorous works that derived DMFT equations for GD/SGD. Celentano et al. [15] rigorously derived DMFT equations for gradient flow dynamics in shallow neural networks by using time discretization and mapping to approximate message passing [7, 26]. We build upon their proof technique to analyze SGD dynamics in this work. Gerbelot et al. [27] derived DMFT equations for discrete-time GD and SGD for shallow neural networks with batch sizes proportional to the number of samples and a constant number of updates. More recently, Fan et al. [24, 25] derived DMFT equations for Langevin dynamics of Bayesian linear regression. A closely related line of work is the study of *general first order methods* (GFOMs), which provides a framework for analyzing a broad class of iterative algorithms, including GD, using a low-dimensional recursion similar to DMFT [14, 30, 31].

DMFT equations have been used for analyzing long-time behavior of optimization dynamics and provided insights into deep learning phenomena such as scaling laws and timescale separation [12, 13, 22, 42, 43].

For linear models, there is a framework that analyzes multi-pass SGD with small batch sizes and proportionally many updates using continuous-time equations. Paquette et al. [44] derived a low-dimensional and continuous-time Volterra equation characterizing the training loss dynamics of SGD in high-dimensional linear regression models. Paquette et al. [45] extended this work and introduced an SDE called *homogenized SGD* (HSGD) as a high-dimensional equivalent of SGD dynamics in linear regression. The HSGD framework allows deriving equations for macroscopic quantities of SGD dynamics, such as training and test errors. Our work provides a similar framework for broader settings, including generalized linear models and shallow neural networks.

SDE approximations of SGD. The SDE approximation of SGD has been used extensively in the literature as a continuous model of discrete-time SGD [3, 16, 34, 39, 40, 46]. Tools from the Itô stochastic calculus can be used to analyze the dynamics of the SDE in detail, and this approach has led to fruitful insights into the behavior of SGD.

Although we do not prove that the high-dimensional behaviors of the SGD (10) and the SGF (1) match (and indeed they do not match exactly, as shown in the concurrent work [23]), we take the SGF (1) as an approximate continuous model of the SGD (10) and analyze its high-dimensional limiting behavior. It is shown that in the fixed-dimensional setting, when the learning rate is small ($\eta \to 0$), the dynamics of SGD is well approximated by that of the SGF (1) [17, 37, 38]. Hence, we expect that the SGF (1) approximates the SGD (10) well when τ is sufficiently small, which we empirically confirm by numerical experiments in Section 5.

Comparison with the concurrent work [23]. A concurrent work by Fan and Wang [23] also derives a DMFT equation characterizing the high-dimensional dynamics of SGD and SGF (which they call SME). Their DMFT equation for SGF is equivalent to ours, although defined differently; they define the response function R_ℓ as a linear operator, while we define it as a continuous function given by the expectation of a stochastic process. Another important difference is that they consider bounded ℓ_t, while we allow for unbounded, Lipschitz continuous ℓ_t. This allows us to rigorously apply our theory to linear regression settings which involve unbounded loss gradients.

A.3. Derivation of the SGF as an Approximation of SGD

Stochastic gradient descent. Fix $m \in \mathbb{N}$, a learning rate $\eta > 0$, a batch size $B \in \mathbb{N}$, a data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, a noise vector $\boldsymbol{z} \in \mathbb{R}^n$, the initial parameter $\boldsymbol{\theta}^0 \in \mathbb{R}^{d\times m}$, and functions $h\colon \mathbb{R}^m \times \mathbb{R}_{\geq 0} \to \mathbb{R}^m; (\theta, t) \mapsto h_t(\theta)$ and $\ell\colon \mathbb{R}^m \times \mathbb{R} \times \mathbb{R}_{\geq 0} \to \mathbb{R}^m; (r, z, t) \mapsto \ell_t(r; z)$. We consider the following stochastic process for $\hat{\boldsymbol{\theta}}^k \in \mathbb{R}^{d\times m}$ for $k = 0, 1, \ldots$, initialized with $\hat{\boldsymbol{\theta}}^0 = \boldsymbol{\theta}^0$:

$$\hat{\boldsymbol{\theta}}^{k+1} = \hat{\boldsymbol{\theta}}^k - \eta \cdot \left(\frac{1}{d} h_{t_k}(\hat{\boldsymbol{\theta}}^k) + \frac{1}{B} \sum_{i\in\mathcal{B}^k} \boldsymbol{x}_i \ell_{t_k}(\hat{r}_i^k; z_i)^\mathsf{T} \right), \quad \hat{\boldsymbol{r}}^k = \boldsymbol{X}\hat{\boldsymbol{\theta}}^k \in \mathbb{R}^{n\times m}, \tag{10}$$

where $t_k \coloneqq \eta k/d$ and $\boldsymbol{x}_i \in \mathbb{R}^d$ and $\hat{r}_i^k \in \mathbb{R}^m$ are the i-th row of $\boldsymbol{X}$ and $\hat{\boldsymbol{r}}^k$, respectively. Here, h_t is applied row-wise to $\hat{\boldsymbol{\theta}}^k$. In each update k, the mini-batch $\mathcal{B}^k$ is sampled uniformly at random from all subsets of $[n] \coloneqq \{1, 2, \ldots, n\}$ with size B.

This stochastic process includes the mini-batch stochastic gradient descent on the following training objective with the learning rate η and the batch size B:

$$\mathcal{L}(\boldsymbol{\theta}) \coloneqq \frac{1}{n} \sum_{i=1}^n L(\boldsymbol{\theta}^\mathsf{T}\boldsymbol{x}_i; z_i) + \frac{1}{d} \sum_{i=1}^d H(\theta_i), \tag{11}$$

where $\theta_i \in \mathbb{R}^m$ is the i-th row of $\boldsymbol{\theta} \in \mathbb{R}^{d\times m}$, $L\colon \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ is a loss function, and $H\colon \mathbb{R}^m \to \mathbb{R}$ is a regularization function. In this case, we have $h_t(\theta) = \nabla_\theta H(\theta)$ and $\ell_t(r; z) = \nabla_r L(r; z)$. This setting includes generalized linear models ($m = 1$) and two-layer neural networks with width m. In the following, we refer to the general process (10) as an 'SGD' for simplicity, although it may not correspond to the gradient descent on any objective function.

Derivation of SGF. To see that the SGF (1) approximates the SGD (10), consider the one-step increment of the SGD (10) given by $\Delta\hat{\boldsymbol{\theta}}^k = \hat{\boldsymbol{\theta}}^{k+1} - \hat{\boldsymbol{\theta}}^k$ and compute its first and second moments conditioned on $\hat{\boldsymbol{\theta}}^k$:

$$\mathbb{E}[\Delta\hat{\boldsymbol{\theta}}^k \mid \hat{\boldsymbol{\theta}}^k] = -\frac{\eta}{d}\left(h_{t_k}(\hat{\boldsymbol{\theta}}^k) + \frac{1}{\delta}\boldsymbol{X}^\mathsf{T}\ell_{t_k}(\hat{\boldsymbol{r}}^k; \boldsymbol{z}) \right), \tag{12}$$

$$\begin{aligned}\mathrm{Cov}(\Delta\hat{\boldsymbol{\theta}}^k \mid \hat{\boldsymbol{\theta}}^k) = {} & \frac{\eta^2(n-B)}{Bn(n-1)} \sum_{i=1}^n \boldsymbol{x}_i \ell_{t_k}(\hat{r}_i^k; z_i)^\mathsf{T} \otimes \boldsymbol{x}_i \ell_{t_k}(\hat{r}_i^k; z_i) \\ & + \frac{\eta^2(n-B)}{Bn^2(n-1)} \boldsymbol{X}^\mathsf{T}\ell_{t_k}(\boldsymbol{X}\hat{\boldsymbol{\theta}}^k; \boldsymbol{z}) \otimes \boldsymbol{X}^\mathsf{T}\ell_{t_k}(\boldsymbol{X}\hat{\boldsymbol{\theta}}^k; \boldsymbol{z}).\end{aligned} \tag{13}$$

Here, the outer product $\boldsymbol{A} \otimes \boldsymbol{B}$ for matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times m}$ is interpreted as $\mathrm{vec}(\boldsymbol{A})\,\mathrm{vec}(\boldsymbol{B})^\mathsf{T} \in \mathbb{R}^{dm\times dm}$, where vec is the vectorization operator. In the proportional high-dimensional limit where

$n, d \to \infty$ with $n/d \to \delta \in (0, \infty)$ and $B = o(n)$, the leading term of the covariance simplifies to

$$\mathrm{Cov}(\Delta\hat{\boldsymbol{\theta}}^k \mid \hat{\boldsymbol{\theta}}^k) = \frac{\eta}{d} \cdot \frac{\tau}{\delta} \sum_{i=1}^{n} \boldsymbol{x}_i \ell_{t_k}(\hat{r}_i^k; z_i)^\mathsf{T} \otimes \boldsymbol{x}_i \ell_{t_k}(\hat{r}_i^k; z_i) + (\text{sub-leading terms})\,. \tag{14}$$

Thus, the first and second moments of the increment $\Delta\hat{\boldsymbol{\theta}}^k$ match those of the increment of the SGF (1) up to the leading order in d over the time interval of length $\Delta t = \eta/d$.

A.4. Formal Statement of the Assumptions

Assumption A.1 (Data distribution)

- *The entries $\boldsymbol{X} = (x_{ij})_{i\in[n], j\in[d]}$ are independent, satisfying $\mathbb{E}\, x_{ij} = 0$, $\mathbb{E}\, x_{ij}^2 = 1/d$, and $\|x_{ij}\|_{\psi_2} \le C/\sqrt{d}$ for some constant $C > 0$.*
- *Proportional high-dimensional asymptotics: $n, d \to \infty,\ n/d \to \delta \in (0, \infty)$.*
- *$\boldsymbol{z} \in \mathbb{R}^n$ and $\boldsymbol{\theta}^0 \in \mathbb{R}^{d\times m}$ are independent of $\boldsymbol{X}$, and for all $p \ge 1$, their empirical distributions $\hat{\mathsf{P}}(\boldsymbol{\theta}^0)$ and $\hat{\mathsf{P}}(\boldsymbol{z})$ converge in p-Wasserstein distance to $\mathsf{P}(\theta^0)$ and $\mathsf{P}(z)$ respectively, almost surely as $d \to \infty$.*

The distribution of the data $\boldsymbol{X}$ is not restricted to the Gaussian distribution, and hence our analysis is *universal* with respect to the data distribution. The last condition on $\boldsymbol{z}$ and $\boldsymbol{\theta}^0$ is satisfied, for example, when their entries are i.i.d. samples from distributions with bounded moments of all orders.

The following assumption is used to guarantee that the SGF solution does not grow too fast and that the DMFT equation to be introduced is well-defined.

Assumption A.2 (Function regularity) *There exists a constant $M > 0$ such that the following holds.*

- *$h_t(\theta)$ and its Jacobian $\mathrm{D}h = (\nabla_\theta h, \partial_t h)$ are Lipschitz continuous in t and θ, i.e., for $t_1, t_2 \ge 0$ and $\theta_1, \theta_2 \in \mathbb{R}^m$,*

$$\|h_{t_1}(\theta_1) - h_{t_2}(\theta_2)\|_2 + \|\mathrm{D}h_{t_1}(\theta_1) - \mathrm{D}h_{t_2}(\theta_2)\|_2 \le M(\|\theta_1 - \theta_2\|_2 + |t_1 - t_2|)\,. \tag{15}$$

- *$\ell_t(r; z)$, its Jacobian $\mathrm{D}\ell = (\nabla_r \ell, \partial_t \ell)$, and its Hessian $\mathrm{D}^2\ell$ are Lipschitz continuous in t and r for any $z \in \mathbb{R}$, i.e., for $t_1, t_2 \ge 0$ and $r_1, r_2 \in \mathbb{R}^m$,*

$$\|\ell_{t_1}(r_1; z) - \ell_{t_2}(r_2; z)\|_2 + \|\mathrm{D}\ell_{t_1}(r_1; z) - \mathrm{D}\ell_{t_2}(r_2; z)\|_2 + \|\mathrm{D}^2\ell_{t_1}(r_1; z) - \mathrm{D}^2\ell_{t_2}(r_2; z)\|_2 \le M(\|r_1 - r_2\|_2 + |t_1 - t_2|)\,. \tag{16}$$

- *$\ell_t(r; z)$ has polynomial growth for any $t \ge 0$, i.e., there exists some $p \ge 1$ such that $\|\ell_t(r; z)\|_2 \le M(1 + \|r\|_2 + |z|)^p$.*

Appendix B. Definition of the DMFT Equation

In this section, we provide a rigorous definition of the DMFT system $\mathfrak{S}$ introduced informally in Equation (2). The key idea is to define all objects through well-defined auxiliary stochastic processes, avoiding functional derivatives and the formal derivative of the Brownian motion. We also introduce a discretized DMFT equation for which the connection between the informal definition and the rigorous definition is more transparent. The discretized DMFT equation is also used as an intermediate step in the proof of Theorem 3.2 and as a numerical method for solving the DMFT equation (see Appendix F).

B.1. Rigorous Definition of the DMFT System

We rigorously define the DMFT system $\mathfrak{S}$ for functions $C_\theta, \Sigma_\ell, R_\theta, R_\ell \colon \mathbb{R}^2_{\geq 0} \to \mathbb{R}^{m\times m}$ and $\Gamma \colon \mathbb{R}_{\geq 0} \to \mathbb{R}^{m\times m}$ self-consistently as follows. First, given $\Sigma_\ell, R_\ell, \Gamma$, define stochastic processes $\{\theta^t \in \mathbb{R}^m\}_{t\geq 0}$ and $\{\rho_\theta^{t,t'} \in \mathbb{R}^{m\times m}\}_{t\geq t'\geq 0}$ by the following equations.

$$\theta^t = \theta^0 + U^t - \int_0^t \left(h_s(\theta^s) + \Gamma(s)\theta^s + \int_0^s R_\ell(s,s')\theta^{s'}\,\mathrm{d}s' \right) \mathrm{d}s\,, \quad U \sim \mathsf{GP}(0, \Sigma_\ell/\delta)\,, \tag{17}$$

$$\rho_\theta^{t,t'} = I_m - \int_{t'}^t \left((\nabla_\theta h_s(\theta^s) + \Gamma(s))\rho_\theta^{s,t'} + \int_{t'}^s R_\ell(s,s')\rho_\theta^{s',t'}\,\mathrm{d}s' \right) \mathrm{d}s\,. \tag{18}$$

Then, set C_θ, R_θ as

$$C_\theta(t,t') = \mathbb{E}[\theta^t \theta^{t'\mathsf{T}}]\,, \tag{19}$$

$$R_\theta(t,t') = \mathbb{E}[\rho_\theta^{t,t'}] \quad (t \geq t')\,, \tag{20}$$

and $R_\theta(t,t') = 0$ for $t < t'$. Here, the expectation is with respect to the randomness of $\theta^0 \sim \mathsf{P}(\theta^0)$ and the Gaussian process U.

Next, given C_θ, R_θ, define stochastic processes $\{r^t \in \mathbb{R}^m\}_{t\geq 0}$ and $\{\rho_\ell^{t,t'}, D_\ell^{t,t'} \in \mathbb{R}^{m\times m}\}_{t\geq t'\geq 0}$ by the following equations.

$$r^t = w^t - \frac{1}{\delta}\int_0^t R_\theta(t,s)\ell_s(r^s;z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\,, \quad w \sim \mathsf{GP}(0, C_\theta)\,, \tag{21}$$

$$\rho_\ell^{t,t'} = \nabla_r \ell_t(r^t;z)\rho_r^{t,t'}\,, \tag{22}$$

$$D_\ell^{t,t'} = \nabla_r \ell_t(r^t;z)\left(-\frac{1}{\delta}\int_{t'}^t R_\theta(t,s) D_\ell^{s,t'}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) \right) + \nabla_r^2 \ell_t(r^t;z)[D_r^{t,t'}]\cdot \rho_r^{t,t'}\,, \tag{23}$$

where B^t is a Brownian motion in $\mathbb{R}$, and we defined the auxiliary processes $\rho_r^{t,t'} \in \mathbb{R}^{m\times m}$ and $D_r^{t,t'} \in \mathbb{R}^m$ as

$$\rho_r^{t,t'} = -\frac{1}{\delta}\int_{t'}^t R_\theta(t,s)\rho_\ell^{s,t'}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) - \frac{1}{\delta}R_\theta(t,t')\nabla_r \ell_{t'}(r^{t'};z)\,, \tag{24}$$

$$D_r^{t,t'} = -\frac{1}{\delta}\int_{t'}^t R_\theta(t,s)\nabla_r \ell_s(r^s;z) D_r^{s,t'}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) - \sqrt{\frac{\tau}{\delta}} R_\theta(t,t')\ell_{t'}(r^{t'};z)\,, \tag{25}$$

and the notation $\nabla_r^2 \ell_t(r^t;z)[v]$ for $v \in \mathbb{R}^m$ in Equation (23) denotes the product

$$\sum_{i=1}^m \partial_{r_i} \nabla_r \ell_t(r^t;z) v_i \in \mathbb{R}^{m\times m}\,. \tag{26}$$

Then, set $\Sigma_\ell, R_\ell, \Gamma$ as

$$\Sigma_\ell(t,t') = \mathbb{E}[L^t L^{t'\mathsf{T}}]\,, \quad L^t := \int_0^t \ell_s(r^s;z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\,, \tag{27}$$

$$R_\ell(t,t') = \mathbb{E}[\rho_\ell^{t,t'}] + \sqrt{\tau\delta}\,\mathbb{E}[D_\ell^{t,t'}] \quad (t \geq t')\,, \tag{28}$$

$$\Gamma(t) = \mathbb{E}[\nabla_r \ell_t(r^t;z)]\,, \tag{29}$$

and $R_\ell(t,t') = 0$ for $t < t'$. Here, the expectation is with respect to the randomness of $z \sim \mathsf{P}(z)$, the Gaussian process w, and the Brownian motion B.

Then, the solution of the DMFT system $\mathfrak{S}$ is defined as a fixed point of the above two mappings.

Note that the above definition is written purely in terms of standard Itô integrals and is thus amenable to rigorous analysis.

B.2. Discretized DMFT System

We present a discretized version of the DMFT system defined above, in which the time variable t is discretized with step size $\gamma > 0$.

Let $t_i \coloneqq i\gamma$ for $i = 0, 1, 2, \dots$. We define the discretized DMFT system $\mathfrak{S}^\gamma$ by the following equations.

$$\theta_\gamma^{t_i} = \theta^0 + U_\gamma^{t_i} - \gamma \sum_{j=0}^{i-1} \left(h_{t_j}(\theta_\gamma^{t_j}) + \Gamma^\gamma(t_j)\theta_\gamma^{t_j} + \gamma \sum_{k=0}^{j-1} R_\ell^\gamma(t_j, t_k)\theta_\gamma^{t_k} \right), \tag{30a}$$

$$\rho_{\theta,\gamma}^{t_i,t_j} = I_m - \gamma \sum_{k=j+1}^{i-1} \left((\nabla_\theta h_{t_k}(\theta_\gamma^{t_k}) + \Gamma^\gamma(t_k))\rho_{\theta,\gamma}^{t_k,t_j} + \gamma \sum_{l=j+1}^{k-1} R_\ell^\gamma(t_k, t_l)\rho_{\theta,\gamma}^{t_l,t_j} \right), \tag{30b}$$

$$r_\gamma^{t_i} = w_\gamma^{t_i} - \frac{1}{\delta} \sum_{j=0}^{i-1} R_\theta^\gamma(t_i, t_j)\ell_{t_j}(r_\gamma^{t_j}; z)(\gamma + \sqrt{\tau\delta}(B^{t_{j+1}} - B^{t_j})), \tag{30c}$$

$$\rho_{\ell,\gamma}^{t_i,t_j} = \nabla_r \ell_{t_i}(r_\gamma^{t_i}; z)\rho_{r,\gamma}^{t_i,t_j}, \tag{30d}$$

$$\rho_{r,\gamma}^{t_i,t_j} = -\frac{1}{\delta} \sum_{k=j+1}^{i-1} R_\theta^\gamma(t_i, t_k)\rho_{\ell,\gamma}^{t_k,t_j}(\gamma + \sqrt{\tau\delta}(B^{t_{k+1}} - B^{t_k})) - \frac{1}{\delta} R_\theta^\gamma(t_i, t_j)\nabla_r \ell_{t_j}(r_\gamma^{t_j}; z), \tag{30e}$$

$$\begin{aligned} D_{\ell,\gamma}^{t_i,t_j} &= \nabla_r \ell_{t_i}(r_\gamma^{t_i}; z) \left(-\frac{1}{\delta} \sum_{k=j+1}^{i-1} R_\theta^\gamma(t_i, t_k) D_{\ell,\gamma}^{t_k,t_j}(\gamma + \sqrt{\tau\delta}(B^{t_{k+1}} - B^{t_k})) \right) \\ &\quad + \nabla_r^2 \ell_{t_i}(r_\gamma^{t_i}; z)[D_{r,\gamma}^{t_i,t_j}] \cdot \rho_{r,\gamma}^{t_i,t_j}, \end{aligned} \tag{30f}$$

$$\begin{aligned} D_{r,\gamma}^{t_i,t_j} &= -\frac{1}{\delta} \sum_{k=j+1}^{i-1} R_\theta^\gamma(t_i, t_k)\nabla_r \ell_{t_k}(r_\gamma^{t_k}; z) D_{r,\gamma}^{t_k,t_j}(\gamma + \sqrt{\tau\delta}(B^{t_{k+1}} - B^{t_k})) \\ &\quad - \sqrt{\frac{\tau}{\delta}} R_\theta^\gamma(t_i, t_j)\ell_{t_j}(r_\gamma^{t_j}; z), \end{aligned} \tag{30g}$$

where $(U_\gamma^{t_i}, w_\gamma^{t_i})_{i \geq 0}$ satisfies

$$\mathbb{E}[U_\gamma^{t_i} U_\gamma^{t_j \mathsf{T}}] = \Sigma_\ell^\gamma(t_i, t_j)/\delta, \quad \mathbb{E}[w_\gamma^{t_i} w_\gamma^{t_j \mathsf{T}}] = C_\theta^\gamma(t_i, t_j). \tag{31}$$

Then, set $C_\theta^\gamma, R_\theta^\gamma, \Sigma_\ell^\gamma, R_\ell^\gamma, \Gamma^\gamma$ as

$$\begin{gathered} C_\theta^\gamma(t_i, t_j) = \mathbb{E}[\theta_\gamma^{t_i}\theta_\gamma^{t_j \mathsf{T}}], \quad R_\theta^\gamma(t_i, t_j) = \mathbb{E}[\rho_{\theta,\gamma}^{t_i,t_j}], \\ \Sigma_\ell^\gamma(t_i, t_j) = \mathbb{E}[L_\gamma^{t_i} L_\gamma^{t_j \mathsf{T}}], \quad L_\gamma^{t_i} \coloneqq \sum_{k=0}^{i-1} \ell_{t_k}(r_\gamma^{t_k}; z)(\gamma + \sqrt{\tau\delta}(B^{t_{k+1}} - B^{t_k})), \\ R_\ell^\gamma(t_i, t_j) = \mathbb{E}[\rho_{\ell,\gamma}^{t_i,t_j}] + \sqrt{\tau\delta}\,\mathbb{E}[D_{\ell,\gamma}^{t_i,t_j}], \quad \Gamma^\gamma(t_i) = \mathbb{E}\big[\nabla_r \ell_{t_i}(r_\gamma^{t_i}; z)\big], \end{gathered} \tag{32}$$

where we set $R_\theta^\gamma(t_i, t_j) = R_\ell^\gamma(t_i, t_j) = 0$ for $i \le j$.

We will show that the solution of the discretized DMFT equation $\mathfrak{S}^\gamma$ converges to the unique solution of the continuous-time DMFT equation $\mathfrak{S}$ as $\gamma \to 0$ in Lemma D.3.

B.3. Correspondence to the Informal Definition

Once discretized, it is easy to see the correspondence between the rigorous definition of the DMFT system given above and the informal definition given in Equation (2). We distinguish the variables in the two definitions by writing bars over the variables in the informal definition, e.g., $\overline{\theta}^t$, $\overline{r}^t$, etc.

We first discretize the informal definition in Equation (2) with step size $\gamma > 0$ in the same manner as in the previous section. We obtain

$$
\begin{aligned}
\frac{\overline{\theta}^{t_{i+1}} - \overline{\theta}^{t_i}}{\gamma} &= \overline{u}^{t_i} - (h_{t_i}(\overline{\theta}^{t_i}) + \overline{\Gamma}(t_i)\overline{\theta}^{t_i}) - \gamma \sum_{j=0}^{i-1} \overline{R}_\ell(t_i, t_j)\overline{\theta}^{t_j}\,, \quad \overline{u} \sim \mathsf{GP}(0, \overline{C}_\ell/\delta)\,, \\
\overline{r}^{t_i} &= \overline{w}^{t_i} - \frac{1}{\delta}\sum_{j=1}^{i-1} \overline{R}_\theta(t_i, t_j)\ell_{t_j}(\overline{r}^{t_j}; z)(\gamma + \sqrt{\tau\delta}(\overline{B}^{t_{j+1}} - \overline{B}^{t_j}))\,, \quad \overline{w} \sim \mathsf{GP}(0, \overline{C}_\theta)\,, \\
\overline{C}_\theta(t_i, t_j) &= \mathbb{E}[\overline{\theta}^{t_i}\overline{\theta}^{t_j\mathsf{T}}]\,, \quad \overline{R}_\theta(t_i, t_j) = \frac{1}{\gamma}\mathbb{E}\left[\frac{\partial \overline{\theta}^{t_i}}{\partial \overline{u}^{t_j}}\right] \quad (i > j)\,, \\
\overline{C}_\ell(t_i, t_j) &= \mathbb{E}\left[\ell_{t_i}(\overline{r}^{t_i}; z)\left(1 + \sqrt{\tau\delta}\frac{\overline{B}^{t_{i+1}} - \overline{B}^{t_i}}{\gamma}\right)\ell_{t_j}(\overline{r}^{t_j}; z)^\mathsf{T}\left(1 + \sqrt{\tau\delta}\frac{\overline{B}^{t_{j+1}} - \overline{B}^{t_j}}{\gamma}\right)\right]\,, \\
\overline{R}_\ell(t_i, t_j) &= \frac{1}{\gamma}\mathbb{E}\left[\frac{\partial \ell_{t_i}(\overline{r}^{t_i}; z)}{\partial \overline{w}^{t_j}}\right]\,, \quad \overline{\Gamma}(t_i) = \mathbb{E}[\nabla_r \ell_{t_i}(\overline{r}^{t_i}; z)] \quad (i > j)\,.
\end{aligned}
\tag{33}
$$

Then, we transform the above equations to show their correspondence to the discretized DMFT system $\mathfrak{S}^\gamma$. The equations for $\overline{r}^{t_i}$, $\overline{C}_\theta$, and $\overline{\Gamma}$ directly correspond to definitions of $r_\gamma^{t_i}$, C_θ^γ, and Γ^γ in $\mathfrak{S}^\gamma$. Next, we show correspondence for $\overline{\theta}^{t_i}$ and $\overline{C}_\ell$. Summing the equation for $\overline{\theta}^{t_i}$ in (33) over i and multiplying by γ, we obtain

$$
\overline{\theta}^{t_i} - \overline{\theta}^0 = \gamma\sum_{j=0}^{i-1}\overline{u}^{t_j} - \gamma\sum_{j=0}^{i-1}\left(h_{t_j}(\overline{\theta}^{t_j}) + \overline{\Gamma}(t_j)\overline{\theta}^{t_j} + \gamma\sum_{k=0}^{j-1}\overline{R}_\ell(t_j, t_k)\overline{\theta}^{t_k}\right)\,, \quad \overline{u} \sim \mathsf{GP}(0, \overline{C}_\ell/\delta)\,.
\tag{34}
$$

Let $\overline{U}^{t_i} := \gamma\sum_{j=0}^{i-1}\overline{u}^{t_j}$. $\overline{U}^{t_i}$ is a Gaussian process with covariance given by

$$
\frac{1}{\delta}\overline{\Sigma}_\ell(t_i, t_j) := \mathbb{E}[\overline{U}^{t_i}\overline{U}^{t_j\mathsf{T}}] = \gamma^2\sum_{k=0}^{i-1}\sum_{l=0}^{j-1}\mathbb{E}[\overline{u}^{t_k}\overline{u}^{t_l\mathsf{T}}] = \frac{\gamma^2}{\delta}\sum_{k=0}^{i-1}\sum_{l=0}^{j-1}\overline{C}_\ell(t_k, t_l) = \frac{1}{\delta}\mathbb{E}[\overline{L}^{t_i}\overline{L}^{t_j\mathsf{T}}]\,,
\tag{35}
$$

where we set

$$
\overline{L}^{t_i} := \sum_{j=0}^{i-1}\ell_{t_j}(\overline{r}^{t_j}; z)(\gamma + \sqrt{\tau\delta}(\overline{B}^{t_{j+1}} - \overline{B}^{t_j}))\,.
\tag{36}
$$

Thus, $\overline{\theta}^{t_i}$ and $\overline{\Sigma}_\ell(t_i,t_j) = \gamma^2 \sum_{k=0}^{i-1}\sum_{l=0}^{j-1} \overline{C}_\ell(t_k,t_l)$ correspond to $\theta_\gamma^{t_i}$ and Σ_ℓ^γ in $\mathfrak{S}^\gamma$. Furthermore, differentiating $\overline{\theta}^{t_i}$ with respect to $\overline{u}^{t_j}$ $(j<i)$, we obtain

$$\frac{\partial \overline{\theta}^{t_i}}{\partial \overline{u}^{t_j}} = \gamma I_m - \sum_{k=j+1}^{i-1}\left((\nabla_\theta h_{t_k}(\overline{\theta}^{t_k}) + \overline{\Gamma}(t_k)) \frac{\partial \overline{\theta}^{t_k}}{\partial \overline{u}^{t_j}} + \gamma \sum_{l=j+1}^{k-1} \overline{R}_\ell(t_k,t_l) \frac{\partial \overline{\theta}^{t_l}}{\partial \overline{u}^{t_j}} \right). \tag{37}$$

This shows correspondence for $\gamma^{-1}\partial\overline{\theta}^{t_i}/\partial\overline{u}^{t_j}$ and $\rho_{\theta,\gamma}^{t_i,t_j}$ and thus $\overline{R}_\theta$ and R_θ^γ. Finally, we show correspondence for $\overline{R}_\ell$. Differentiating $\ell_{t_i}(\overline{r}^{t_i};z)$ with respect to $\overline{w}^{t_j}$ $(j<i)$, we obtain

$$\frac{\partial \ell_{t_i}(\overline{r}^{t_i};z)}{\partial \overline{w}^{t_j}} = \nabla_r \ell_{t_i}(\overline{r}^{t_i};z) \frac{\partial \overline{r}^{t_i}}{\partial \overline{w}^{t_j}}\,, \tag{38}$$

$$\begin{aligned}\frac{\partial \overline{r}^{t_i}}{\partial \overline{w}^{t_j}} &= -\frac{1}{\delta}\sum_{k=j+1}^{i-1} \overline{R}_\theta(t_i,t_k) \frac{\partial \ell_{t_k}(\overline{r}^{t_k};z)}{\partial \overline{w}^{t_j}} (\gamma + \sqrt{\tau\delta}(\overline{B}^{t_{k+1}} - \overline{B}^{t_k})) \\ &\quad - \frac{1}{\delta}\overline{R}_\theta(t_i,t_j)\nabla_r \ell_{t_j}(\overline{r}^{t_j};z)(\gamma + \sqrt{\tau\delta}(B^{t_{j+1}} - B^{t_j}))\,.\end{aligned} \tag{39}$$

Let $\overline{\rho}_\ell^{t_i,t_j}$ and $\overline{\rho}_r^{t_i,t_j}$ be the solution of the following equation:

$$\overline{\rho}_\ell^{t_i,t_j} = \nabla_r \ell_{t_i}(\overline{r}^{t_i};z)\overline{\rho}_r^{t_i,t_j}\,, \tag{40}$$

$$\overline{\rho}_r^{t_i,t_j} = -\frac{1}{\delta}\sum_{k=j+1}^{i-1} \overline{R}_\theta(t_i,t_k)\overline{\rho}_\ell^{t_k,t_j}(\gamma + \sqrt{\tau\delta}(B^{t_{k+1}} - B^{t_k})) - \frac{1}{\delta}\overline{R}_\theta(t_i,t_j)\nabla_r \ell_{t_j}(\overline{r}^{t_j};z)\,. \tag{41}$$

These equations correspond to definitions of $\rho_{\ell,\gamma}^{t_i,t_j}$ and $\rho_{r,\gamma}^{t_i,t_j}$. By the linearity of the above equations, we see that

$$\frac{\partial \ell_{t_i}(\overline{r}^{t_i};z)}{\partial \overline{w}^{t_j}} = \overline{\rho}_\ell^{t_i,t_j}(\gamma + \sqrt{\tau\delta}(B^{t_{j+1}} - B^{t_j}))\,. \tag{42}$$

Let $\overline{G}^j := (\overline{B}^{t_{j+1}} - \overline{B}^{t_j})/\sqrt{\gamma}$. Then, we have $\overline{G}^j \sim \mathsf{N}(0,1)$ i.i.d. for $j = 0,1,2,\ldots$. Taking the expectation of the above equation, we obtain

$$\overline{R}_\ell(t_i,t_j) = \frac{1}{\gamma}\,\mathbb{E}\left[\frac{\partial \ell_{t_i}(\overline{r}^{t_i};z)}{\partial \overline{w}^{t_j}}\right] = \mathbb{E}[\overline{\rho}_\ell^{t_i,t_j}] + \sqrt{\frac{\tau\delta}{\gamma}}\,\mathbb{E}[\overline{\rho}_\ell^{t_i,t_j}\overline{G}^j]\,. \tag{43}$$

By Stein's lemma (Gaussian integration by parts), we obtain

$$\mathbb{E}[\overline{\rho}_\ell^{t_i,t_j}\overline{G}^j] = \mathbb{E}\left[\frac{\partial \overline{\rho}_\ell^{t_i,t_j}}{\partial \overline{G}^j}\right]. \tag{44}$$

Differentiating $\overline{\rho}_\ell^{t_i,t_j}$ with respect to $\overline{G}^j$ and using independence of $\overline{r}^{t_k}$ and $\overline{G}^j$ for $k \le j$, we obtain

$$\frac{\partial \overline{\rho}_\ell^{t_i,t_j}}{\partial \overline{G}^j} = \nabla_r \ell_{t_i}(\overline{r}^{t_i};z)\frac{\partial \overline{\rho}_r^{t_i,t_j}}{\partial \overline{G}^j} + \nabla_r^2 \ell_{t_i}(\overline{r}^{t_i};z)\left[\frac{\partial \overline{r}^{t_i}}{\partial \overline{G}^j}\right]\cdot \overline{\rho}_r^{t_i,t_j}\,, \tag{45}$$

$$\frac{\partial \overline{\rho}_r^{t_i,t_j}}{\partial \overline{G}^j} = -\frac{1}{\delta} \sum_{k=j+1}^{i-1} \overline{R}_\theta(t_i,t_k) \frac{\partial \overline{\rho}_\ell^{t_k,t_j}}{\partial \overline{G}^j} (\gamma + \sqrt{\tau\delta\gamma}\overline{G}^k)\,, \tag{46}$$

$$\frac{\partial \overline{r}^{t_i}}{\partial \overline{G}^j} = -\frac{1}{\delta} \sum_{k=j+1}^{i-1} \overline{R}_\theta(t_i,t_k) \nabla_r \ell_{t_k}(\overline{r}^{t_k};z) \frac{\partial \overline{r}^{t_k}}{\partial \overline{G}^j} (\gamma + \sqrt{\tau\delta\gamma}\overline{G}^k) - \sqrt{\frac{\tau\gamma}{\delta}} \overline{R}_\theta(t_i,t_j) \ell_{t_j}(\overline{r}^{t_j};z)\,. \tag{47}$$

This shows that $\gamma^{-1/2}\partial\overline{\rho}_\ell^{t_i,t_j}/\partial\overline{G}^j$ and $\gamma^{-1/2}\partial\overline{r}^{t_i}/\partial\overline{G}^j$ correspond to $D_{\ell,\gamma}^{t_i,t_j}$ and $D_{r,\gamma}^{t_i,t_j}$. Thus, we have correspondence for $\overline{R}_\ell$ and R_ℓ^γ.

Summarizing, the informal definition corresponds to the formal definition by the following correspondence (informal definition on the left, rigorous definition on the right):

$$\begin{gathered}
\overline{\theta}^{t_i} \Leftrightarrow \theta_\gamma^{t_i}\,, \quad \frac{1}{\gamma}\frac{\partial \overline{\theta}^{t_i}}{\partial \overline{u}^{t_j}} \Leftrightarrow \rho_{\theta,\gamma}^{t_i,t_j}\,, \quad \overline{C}_\theta(t_i,t_j) \Leftrightarrow C_\theta^\gamma(t_i,t_j)\,, \quad \overline{R}_\theta(t_i,t_j) \Leftrightarrow R_\theta^\gamma(t_i,t_j)\,, \\
\overline{r}^{t_i} \Leftrightarrow r_\gamma^{t_i}\,, \quad \overline{\rho}_\ell^{t_i,t_j} \Leftrightarrow \rho_{\ell,\gamma}^{t_i,t_j}\,, \quad \frac{1}{\sqrt{\gamma}}\frac{\partial \overline{\rho}_\ell^{t_i,t_j}}{\partial \overline{G}^{t_j}} \Leftrightarrow D_{\ell,\gamma}^{t_i,t_j}\,, \quad \frac{1}{\gamma}\,\mathbb{E}\left[\frac{\partial \ell_{t_i}(\overline{r}^{t_i};z)}{\partial \overline{w}^{t_j}}\right] \Leftrightarrow \mathbb{E}[\rho_{\ell,\gamma}^{t_i,t_j}] + \sqrt{\tau\delta}\,\mathbb{E}[D_{\ell,\gamma}^{t_i,t_j}] \\
\gamma^2 \sum_{k=0}^{i-1}\sum_{l=0}^{j-1} \overline{C}_\ell(t_k,t_l) \Leftrightarrow \Sigma_\ell^\gamma(t_i,t_j)\,, \quad \overline{R}_\ell(t_i,t_j) \Leftrightarrow R_\ell^\gamma(t_i,t_j)\,, \quad \overline{\Gamma}(t_i) \Leftrightarrow \Gamma^\gamma(t_i)\,.
\end{gathered} \tag{48}$$

Although these two definitions are equivalent in discrete time, the informal definition does not have a well-defined continuous-time limit as $\gamma \to 0$, while the rigorous definition does. Thus, for theoretical purposes, we work with the rigorous definition. In numerics, however, we work with the informal definition after time discretization, as it leads to simpler numerical schemes (see Appendix F for details).

Appendix C. Proof of Theorem 3.1

We prove Theorem 3.1 using a contraction mapping argument similar to that of Celentano et al. [15, Theorem 1] and Fan et al. [24, Theorem 2.4]. It proceeds as follows.

1. For $T > 0$, we define *admissible spaces* $\mathcal{S}_\theta(T)$ and $\mathcal{S}_\ell(T)$ for the DMFT objects (C_θ, R_θ) and $(\Sigma_\ell, R_\ell, \Gamma)$, respectively. We define mappings $\mathcal{T}_{\theta\to\ell}\colon \mathcal{S}_\theta(T) \to \mathcal{S}_\ell(T)$ and $\mathcal{T}_{\ell\to\theta}\colon \mathcal{S}_\ell(T) \to \mathcal{S}_\theta(T)$ such that the fixed point of their composition $\mathcal{T} := \mathcal{T}_{\ell\to\theta} \circ \mathcal{T}_{\theta\to\ell}$ solves the DMFT system. We show that for sufficiently small $T > 0$, these mappings are well-defined.

2. Next, we construct a metric on the function spaces $\mathcal{S}_\theta(T)$ and $\mathcal{S}_\ell(T)$ such that $\mathcal{T}$ is a contraction.

3. Finally, we apply Banach's fixed point theorem to show the uniqueness and existence of the fixed point of $\mathcal{T}$.

C.1. Admissible Spaces $\mathcal{S}_\theta(T)$ and $\mathcal{S}_\ell(T)$

Since the following quantities are bounded by assumptions, we take $M > 0$ sufficiently large so that we have

$$\max\left\{\mathbb{E}\|\theta^0\|_2^2,\ \sup_{t\in[0,T]} h_t(0),\ \sup_{t\in[0,T]} \mathbb{E}\|\ell_t(0;z)\|_2^{2p}\right\} \le M\,, \tag{49}$$

for $p = 1, 2$.

For $T > 0$, we define admissible spaces $\mathcal{S}_\theta(T)$ and $\mathcal{S}_\ell(T)$ as follows.

Definition C.1 (Admissible space $\mathcal{S}_\theta(T)$) *Let $D \subset (0,T)$ be a finite set. Let $\mathcal{S}_\theta(T)$ be a set of function pairs (C_θ, R_θ) defined on $[0,T]^2$. We say that $\mathcal{S}_\theta(T)$ is* admissible *if there exist constants $\Phi_\theta, M_\theta > 0$ such that every $(C_\theta, R_\theta) \in \mathcal{S}_\theta(T)$ satisfies the following.*

- *C_θ is a covariance kernel (in particular, it satisfies $C_\theta(t,t') = C_\theta(t',t)^\mathsf{T}$) and satisfies $\|C_\theta(t,t)\|_2 \le \Phi_\theta$ for all $t \in [0,T]$ and $C_\theta(0,0) = \mathbb{E}[\theta^0\theta^{0\mathsf{T}}]$. Furthermore, $C_\theta(t,t')$ is uniformly continuous over $t, t' \in I$ for each maximal interval I of $[0,T] \setminus D$ and satisfies*

$$\|C_\theta(t,t) - 2C_\theta(t,t') + C_\theta(t',t')\|_2 \le M_\theta|t-t'|\,, \tag{50}$$

  for any $t, t' \in I$.

- *R_θ satisfies $R_\theta(t,t') = 0$ for $0 \le t < t' \le T$ and $\|R_\theta(t,t')\|_2 \le \Phi_\theta$ for $0 \le t' \le t \le T$. Furthermore, $R_\theta(t,t')$ is uniformly continuous over $t \in I$ and $t' \in I'$ for any two maximal intervals I, I' of $[0,T] \setminus D$.*

We define $\mathcal{S}_\theta^{\mathrm{cont}}(T)$ as the subset of $\mathcal{S}_\theta(T)$ with $D = \emptyset$ in the above definition.

Definition C.2 (Admissible space $\mathcal{S}_\ell(T)$) *Let $D \subset (0,T)$ be a finite set. Let $\mathcal{S}_\ell(T)$ be a set of function triples $(\Sigma_\ell, R_\ell, \Gamma)$ defined on $[0,T]^2$ and $[0,T]$. We say that $\mathcal{S}_\ell(T)$ is* admissible *if there exist constants $\Phi_\ell, M_\ell > 0$ such that every $(\Sigma_\ell, R_\ell, \Gamma) \in \mathcal{S}_\ell(T)$ satisfies the following.*

- *$\Sigma_\ell(t,s)$ is a covariance kernel (in particular, it satisfies $\Sigma_\ell(t,s) = \Sigma_\ell(s,t)^\mathsf{T}$) and satisfies $\|\Sigma_\ell(t,t)\|_2 \le \Phi_\ell$ for $t \in [0,T]$ and $\Sigma_\ell(0,0) = \mathbb{E}[\ell_0(r^0;z)\ell_0(r^0;z)^\mathsf{T}]$ for $r^0 \sim \mathsf{N}(0, \mathbb{E}[\theta^0\theta^{0\mathsf{T}}])$. Furthermore, $\Sigma_\ell(t,t')$ is uniformly continuous over $t, t' \in I$ for each maximal interval I of $[0,T] \setminus D$ and satisfies*

$$\|\Sigma_\ell(t,t) - 2\Sigma_\ell(t,t') + \Sigma_\ell(t',t')\|_2 \le M_\ell|t-t'|\,, \tag{51}$$

  for any $t, t' \in I$.

- *$R_\ell(t,t')$ satisfies $R_\ell(t,t') = 0$ for $0 \le t < t' \le T$ and $\|R_\ell(t,t')\|_2 \le \Phi_\ell$ for $0 \le t' \le t \le T$. Furthermore, $R_\ell(t,t')$ is uniformly continuous over $t \in I$ and $t' \in I'$ for any two maximal intervals I, I' of $[0,T] \setminus D$.*

- *$\Gamma(t)$ satisfies $\|\Gamma(t)\|_2 \le M$ for $t \in [0,T]$ and $\Gamma(0) = \mathbb{E}[\nabla_r \ell_0(r^0;z)]$ for $r^0 \sim \mathsf{N}(0, \mathbb{E}[\theta^0\theta^{0\mathsf{T}}])$. Furthermore, $\Gamma(t)$ is uniformly continuous over $t \in I$ for each maximal interval I of $[0,T] \setminus D$.*

We define $\mathcal{S}_\ell^{\mathrm{cont}}(T)$ as the subset of $\mathcal{S}_\ell(T)$ with $D = \emptyset$ in the above definition.

In the above definitions, we allow for discontinuities at a finite set of time points D to handle the discretized DMFT system later in the proof of Theorem 3.2 in Appendix D.

We now show that the stochastic processes are uniquely defined given functions in admissible spaces.

Lemma C.3 *Given an admissible space $\mathcal{S}_\theta(T)$ and any element $(C_\theta, R_\theta) \in \mathcal{S}_\theta(T)$, there exists a unique tuple of stochastic processes $\{r^t, \rho_\ell^{t,t'}, D_\ell^{t,t'}\}_{0\le t'\le t\le T}$ satisfying Equations (21) to (23). Furthermore, for any $(C_\theta, R_\theta) \in \mathcal{S}_\theta^{\mathrm{cont}}(T)$, the processes $\{r^t, \rho_\ell^{t,t'}, D_\ell^{t,t'}\}_{0\le t'\le t\le T}$ have continuous sample paths.*

Similarly, given an admissible space $\mathcal{S}_\ell(T)$ and any element $(\Sigma_\ell, R_\ell, \Gamma) \in \mathcal{S}_\ell(T)$, there exists a unique pair of stochastic processes $\{\theta^t, \rho_\theta^{t,t'}\}_{0\le t'\le t\le T}$ satisfying Equations (17) and (18). Furthermore, for any $(\Sigma_\ell, R_\ell, \Gamma) \in \mathcal{S}_\ell^{\mathrm{cont}}(T)$, the processes $\{\theta^t, \rho_\theta^{t,t'}\}_{0\le t'\le t\le T}$ have continuous sample paths.

Proof First, we show that r^t is uniquely defined. Let $\{w^t\}_{t\in[0,T]}$ be a centered Gaussian process with covariance kernel C_θ. Then, for any maximal interval I of $[0,T]\setminus D$ and any $t, t' \in I$, we have

$$\mathbb{E}\|w^t - w^{t'}\|_2^4 \le 3m^2\|C_\theta(t,t) - 2C_\theta(t,t') + C_\theta(t',t')\|_2^2 \le 3m^2 M_\theta^2 (t-t')^2 \,. \tag{52}$$

By the Kolmogorov continuity theorem, there exists a modification of w^t that is locally Hölder continuous on I. Then, for each maximal interval I, r^t follows a nonlinear Volterra stochastic integral equation of the second kind with a Lipschitz nonlinearity, a continuous kernel, and a continuous forcing term. By Berger and Mizel [9, Theorem 3.A], it has a unique continuous solution adapted to the filtration $\mathcal{F}_t^\ell$ generated by $(\boldsymbol{B}^s)_{s\le t}$. Applying this argument inductively over the maximal intervals of $[0,T]\setminus D$, we conclude that r^t is uniquely defined over $[0,T]$. The well-posedness of $\rho_\ell^{t,t'}$ and $D_\ell^{t,t'}$ can be shown similarly using the continuity of r^t.

Next, we show that θ^t is uniquely defined. Let $\{U^t\}_{t\in[0,T]}$ be a centered Gaussian process with covariance kernel Σ_ℓ/δ. Again, by the Kolmogorov continuity theorem, there exists a modification of U^t that is locally Hölder continuous on each I. From Equation (17), θ^t satisfies the following equation:

$$\theta^t = \theta^0 + U^t - \int_0^t \left(h_s(\theta^s) + \Gamma(s)\theta^s + \left(\int_s^t R_\ell(t',s)\,\mathrm{d}t'\right)\theta^s \right) \mathrm{d}s \,. \tag{53}$$

For each I, this is a nonlinear Volterra integral equation of the second kind with a continuous kernel and a continuous forcing term. Again, by Berger and Mizel [9, Theorem 3.A], it has a unique continuous solution. Applying this argument inductively over the maximal intervals of $[0,T]\setminus D$, we conclude that θ^t is uniquely defined over $[0,T]$. The well-posedness of $\rho_\theta^{t,t'}$ can be shown similarly using the continuity of θ^t. ■

Next, we define mappings between the admissible spaces.

First, we define the map $\mathcal{T}_{\theta\to\ell}\colon (C_\theta, R_\theta) \mapsto (\Sigma_\ell, R_\ell, \Gamma)$. Given $(C_\theta, R_\theta) \in \mathcal{S}_\theta$, take the unique processes $r^t, \rho_\ell^{t,t'}, D_\ell^{t,t'}$ satisfying Equations (21) to (23) whose existence is guaranteed by Lemma C.3. Then, we define $(\Sigma_\ell, R_\ell, \Gamma)$ by Equations (27) to (29).

Next, we define the map $\mathcal{T}_{\ell\to\theta}\colon (\Sigma_\ell, R_\ell, \Gamma) \mapsto (C_\theta, R_\theta)$. Given $(\Sigma_\ell, R_\ell, \Gamma) \in \mathcal{S}_\ell$, take the unique processes $\theta^t, \rho_\theta^{t,t'}$ satisfying Equations (17) and (18) whose existence is guaranteed by Lemma C.3. Then, we define (C_θ, R_θ) by Equations (19) and (20).

Finally, we define the composite map $\mathcal{T} = \mathcal{T}_{\ell\to\theta} \circ \mathcal{T}_{\theta\to\ell}$.

In the following lemma, we show that for sufficiently small $T > 0$, $\mathcal{T}_{\theta\to\ell}$ and $\mathcal{T}_{\ell\to\theta}$ map $\mathcal{S}_\theta(T)$ into $\mathcal{S}_\ell(T)$ and $\mathcal{S}_\ell(T)$ into $\mathcal{S}_\theta(T)$, respectively. We defer the proof to Appendix C.3.

Lemma C.4

1. *There exists some $T_* > 0$ such that, for any $0 < T \le T_*$, there exist admissible spaces $\mathcal{S}_\theta(T)$ and $\mathcal{S}_\ell(T)$ such that $\mathcal{T}_{\theta\to\ell}$ maps $\mathcal{S}_\theta(T)$ into $\mathcal{S}_\ell^{\mathrm{cont}}(T)$ and $\mathcal{T}_{\ell\to\theta}$ maps $\mathcal{S}_\ell(T)$ to $\mathcal{S}_\theta^{\mathrm{cont}}(T)$.*
2. *If either $\tau = 0$ or $\nabla^2 \ell_t(r; z) = 0$, T can be taken arbitrarily large (thus $T_* = \infty$).*

C.2. Equipping Metrics on $\mathcal{S}_\theta(T)$ and $\mathcal{S}_\ell(T)$

In the following, we fix $T > 0$ such that Lemma C.4 holds and fix admissible spaces $\mathcal{S}_\theta := \mathcal{S}_\theta(T)$ and $\mathcal{S}_\ell := \mathcal{S}_\ell(T)$.

We equip the spaces $\mathcal{S}_\theta$ and $\mathcal{S}_\ell$ with metrics. For a constant $\lambda > 0$, we define

$$\mathsf{dist}_\lambda(C_\theta^1, C_\theta^2) := \inf_{w_1\sim\mathsf{GP}(0,C_\theta^1), w_2\sim\mathsf{GP}(0,C_\theta^2)} \sup_{0\le t\le T} \mathrm{e}^{-\lambda t}\sqrt{\mathbb{E}\|w_1^t - w_2^t\|_2^2}\,, \tag{54a}$$

$$\mathsf{dist}_\lambda(\Sigma_\ell^1, \Sigma_\ell^2) := \inf_{U_1\sim\mathsf{GP}(0,\Sigma_\ell^1/\delta), U_2\sim\mathsf{GP}(0,\Sigma_\ell^2/\delta)} \sup_{0\le t\le T} \mathrm{e}^{-\lambda t}\sqrt{\mathbb{E}\|U_1^t - U_2^t\|_2^2}\,, \tag{54b}$$

$$\mathsf{dist}_\lambda(R_\theta^1, R_\theta^2) := \sup_{0\le s\le t\le T} \mathrm{e}^{-\lambda t}\|R_\theta^1(t,s) - R_\theta^2(t,s)\|\,, \tag{54c}$$

$$\mathsf{dist}_\lambda(R_\ell^1, R_\ell^2) := \sup_{0\le s\le t\le T} \mathrm{e}^{-\lambda t}\|R_\ell^1(t,s) - R_\ell^2(t,s)\|\,, \tag{54d}$$

$$\mathsf{dist}_\lambda(\Gamma^1, \Gamma^2) := \sup_{0\le t\le T} \mathrm{e}^{-\lambda t}\|\Gamma^1(t) - \Gamma^2(t)\|\,. \tag{54e}$$

In the first two definitions, the infima are taken over all couplings of the Gaussian processes with given marginal covariances. Finally, for $X^i = (C_\theta^i, R_\theta^i) \in \mathcal{S}_\theta$ and $Y^i = (\Sigma_\ell^i, R_\ell^i, \Gamma^i) \in \mathcal{S}_\ell$, we define the distances

$$\mathsf{dist}_\lambda(X^1, X^2) := \mathsf{dist}_\lambda(C_\theta^1, C_\theta^2) + \mathsf{dist}_\lambda(R_\theta^1, R_\theta^2)\,, \tag{55}$$

$$\mathsf{dist}_\lambda(Y^1, Y^2) := \sqrt{\lambda}\mathsf{dist}_\lambda(\Sigma_\ell^1, \Sigma_\ell^2) + \mathsf{dist}_\lambda(R_\ell^1, R_\ell^2) + \mathsf{dist}_\lambda(\Gamma^1, \Gamma^2)\,. \tag{56}$$

Notice the $\sqrt{\lambda}$ factor in front of $\mathsf{dist}_\lambda(\Sigma_\ell^1, \Sigma_\ell^2)$.

We show several properties of the metric spaces $(\mathcal{S}_\theta, \mathsf{dist}_\lambda)$ and $(\mathcal{S}_\ell, \mathsf{dist}_\lambda)$.

Lemma C.5 *The metric spaces $(\mathcal{S}_\theta, \mathsf{dist}_\lambda)$ and $(\mathcal{S}_\ell, \mathsf{dist}_\lambda)$ are complete.*

Proof For finite T, the distance dist_λ for R_θ, R_ℓ, and Γ are equivalent to L^∞ distance and hence complete. Completeness for C_θ and Σ_ℓ are shown in the proof of Fan et al. [24, Theorem 2.4(b)]. ■

Lemma C.6 *Let $X^i = (C_\theta^i, R_\theta^i) \in \mathcal{S}_\theta$ and $Y^i = \mathcal{T}_{\theta\to\ell}(X^i) = (\Sigma_\ell^i, R_\ell^i, \Gamma^i) \in \mathcal{S}_\ell$ for $i = 1, 2$. Then, there exists a constant $K > 0$ such that for any sufficiently large $\lambda > 0$ in the definitions of the metrics (54), we have*

$$\mathsf{dist}_\lambda(Y^1, Y^2) \le K \cdot \mathsf{dist}_\lambda(X^1, X^2)\,. \tag{57}$$

Lemma C.7 *Let $Y^i = (C^i_\ell, R^i_\ell, \Gamma^i) \in \mathcal{S}_\ell$ and $X^i = \mathcal{T}_{\ell\to\theta}(Y^i) = (C^i_\theta, R^i_\theta) \in \mathcal{S}_\theta$ for $i = 1, 2$. Then, for any $\varepsilon > 0$, for sufficiently large $\lambda > 0$, we have*

$$\mathsf{dist}_\lambda(X^1, X^2) \le \varepsilon \cdot \mathsf{dist}_\lambda(Y^1, Y^2)\,. \tag{58}$$

We defer the proof of the last two lemmas to Appendices C.4 and C.5.

Finally, we show that $\mathcal{T}$ is a contraction mapping under the above metrics, finishing the proof of Theorem 3.1. Take $T \in [0, T_*]$ where T_* is as in Lemma C.4, and take admissible spaces $\mathcal{S}_\theta := \mathcal{S}_\theta(T)$ and $\mathcal{S}_\ell := \mathcal{S}_\ell(T)$. By Lemmas C.6 and C.7, we can choose $\varepsilon < 1/K$ and λ sufficiently large such that $\mathcal{T} = \mathcal{T}_{\ell\to\theta} \circ \mathcal{T}_{\theta\to\ell}$ is a contraction mapping on the metric space $(\mathcal{S}_\theta, \mathsf{dist}_\lambda)$ which is complete by Lemma C.5. By the Banach fixed-point theorem, there exists a unique fixed point $(C_\theta, R_\theta) \in \mathcal{S}^{\mathrm{cont}}_\theta$ such that $\mathcal{T}(C_\theta, R_\theta) = (C_\theta, R_\theta)$. Thus, this (C_θ, R_θ) and $(\Sigma_\ell, R_\ell, \Gamma) = \mathcal{T}_{\theta\to\ell}(C_\theta, R_\theta) \in \mathcal{S}^{\mathrm{cont}}_\ell$ together form a unique pair of fixed points satisfying the DMFT equation $\mathfrak{S}$.

The continuity of the sample paths follows from Lemma C.3.

C.3. Proof of Lemma C.4

We will use the following bounds repeatedly in the proof.

Lemma C.8 *Let X_1 be a random variable and X^s_2, X^s_3 be stochastic processes in $\mathbb{R}^m$ adapted to the Brownian motion $(B^{s'})_{0\le s'<s}$. Then, for any $0 \le t' \le t$ and any integer $p \ge 1$, there exists a constant $C_p > 0$ such that we have*

$$\begin{aligned}
&\mathbb{E}\left\| X_1 + \int_{t'}^t X^s_2 \,\mathrm{d}s + \int_{t'}^t X^s_3 \,\mathrm{d}B^s \right\|_2^{2p} \\
&\quad \le 3^{2p-1}\left(\mathbb{E}\|X_1\|_2^{2p} + (t-t')^{2p-1} \int_{t'}^t \mathbb{E}\|X^s_2\|_2^{2p}\,\mathrm{d}s + C_p (t-t')^{p-1} \int_{t'}^t \mathbb{E}\|X^s_3\|_2^{2p}\,\mathrm{d}s \right).
\end{aligned} \tag{59}$$

Proof By Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{E}\left\| X_1 + \int_{t'}^t X^s_2 \,\mathrm{d}s + \int_{t'}^t X^s_3 \,\mathrm{d}B^s \right\|_2^{2p} \\
&\le 3^{2p-1}\left(\mathbb{E}\|X_1\|_2^{2p} + (t-t')^{2p-1} \int_{t'}^t \mathbb{E}\|X^s_2\|_2^{2p}\,\mathrm{d}s + \mathbb{E}\left\| \int_{t'}^t X^s_3 \,\mathrm{d}B^s \right\|_2^{2p} \right).
\end{aligned} \tag{60}$$

By the Burkholder–Davis–Gundy inequality and Jensen's inequality again, we have

$$\mathbb{E}\left\| \int_{t'}^t X^s_3 \,\mathrm{d}B^s \right\|_2^{2p} \le C_p\, \mathbb{E}\left(\int_{t'}^t \|X^s_3\|_2^2 \,\mathrm{d}s \right)^p \le C_p (t-t')^{p-1} \int_{t'}^t \mathbb{E}\|X^s_3\|_2^{2p}\,\mathrm{d}s\,. \tag{61}$$

This proves the claim. Note that we can set $C_1 = 1$ by the Itô isometry. ■

C.3.1. CONSTRUCTION OF THE ADMISSIBLE SPACES

Take constants $\Phi_{C_\theta} \geq 7(M + 4T^2M^2)$ and $\Phi_{R_\theta} \geq 2$. Define constants $\Phi_{C_\ell}, \Phi_{\Sigma_\ell}, \Phi_{R_\ell}, \Phi_i$ $(i = 1, \dots, 5)$, M_ℓ, M_θ, and $\overline{\lambda}$ as follows:

$$\Phi_{C_\ell} \coloneqq (2M + 6M^2 m \Phi_{C_\theta}) \exp\left(6M^2\left(\frac{T}{\delta^2} + \frac{\tau}{\delta}\right)\Phi_{R_\theta}^2 T\right), \tag{62}$$

$$\Phi_{\Sigma_\ell} \coloneqq (T + \tau\delta)\Phi_{C_\ell}, \tag{63}$$

$$\Phi_{R_\ell} \coloneqq \sqrt{2\Phi_1 + 2\tau\delta\Phi_5}, \tag{64}$$

$$\Phi_1 \coloneqq \frac{3M^4}{\delta^2}\Phi_{R_\theta}^2 \exp\left(3M^2\left(\frac{T}{\delta^2} + \frac{\tau}{\delta}\right)\Phi_{R_\theta}^2 T\right), \tag{65}$$

$$\Phi_2 \coloneqq \frac{27M^4\Phi_{R_\theta}^4}{\delta^4} \exp\left(27M^4\left(\frac{T^3}{\delta^4} + \frac{C_2\tau^2 T}{\delta^2}\right)\Phi_{R_\theta}^4 T\right), \tag{66}$$

$$\Phi_3 \coloneqq \left(8M + 648M^4 m^2 \Phi_{C_\theta}^2\right) \exp\left(216M^4\left(\frac{T^3}{\delta^4} + \frac{C_2\tau^2 T}{\delta^2}\right)\Phi_{R_\theta}^4 T\right), \tag{67}$$

$$\Phi_4 \coloneqq \frac{27\tau^2\Phi_{R_\theta}^4\Phi_3 \mathrm{e}^4}{\delta^2} \exp\left(27M^4\left(\frac{T^3}{\delta^4} + \frac{C_2\tau^2 T}{\delta^2}\right)\Phi_{R_\theta}^4 T\right), \tag{68}$$

$$\Phi_5 \coloneqq 2M^2\sqrt{\Phi_2\Phi_4}\mathrm{e}^2 \exp\left(4M^2\left(\frac{T}{\delta^2} + \frac{\tau}{\delta}\right)\Phi_{R_\theta}^2 T\right), \tag{69}$$

$$M_\ell \coloneqq 2(T + \tau\delta)\Phi_{C_\ell}\mathrm{e}^{2\overline{\lambda}T}, \tag{70}$$

$$M_\theta \coloneqq 2\left(\frac{mM_\ell}{\delta} + 2T\left(2M^2 + 3M^2\Phi_{C_\theta}\mathrm{e}^{2\overline{\lambda}T} + T^2\Phi_{R_\ell}^2\Phi_{C_\theta}\mathrm{e}^{4\overline{\lambda}T}\right)\right), \tag{71}$$

$$\overline{\lambda} \coloneqq \max\left\{2T(3M^2 + T\Phi_{R_\ell}^2), \frac{6m\Phi_{\Sigma_\ell}}{\delta M}, 2(2M + \Phi_{R_\ell}T)\right\}, \tag{72}$$

where $C_2 > 0$ is the constant in Lemma C.8 for $p = 2$.

For $T > 0$, we define the function spaces $\mathcal{S}_\theta \coloneqq \mathcal{S}_\theta(T)$ and $\mathcal{S}_\ell \coloneqq \mathcal{S}_\ell(T)$ as follows. We define $\mathcal{S}_\theta$ as the space of pairs of functions (C_θ, R_θ) satisfying the continuity conditions and initial conditions in Definition C.1 and the following bounds for $0 \leq t' \leq t \leq T$:

$$\|C_\theta(t,t)\|_2 \leq \Phi_{C_\theta}\mathrm{e}^{2\overline{\lambda}t}, \quad \|R_\theta(t,t')\|_2 \leq \Phi_{R_\theta}\mathrm{e}^{\overline{\lambda}(t-t')}. \tag{73}$$

Then, $\mathcal{S}_\theta$ is admissible with parameters $\Phi_\theta = \max\{\Phi_{C_\theta}\mathrm{e}^{2\overline{\lambda}T}, \Phi_{R_\theta}\mathrm{e}^{\overline{\lambda}T}\}$ and M_θ.

We define $\mathcal{S}_\ell$ as the space of triples of functions $(\Sigma_\ell, R_\ell, \Gamma)$ satisfying the continuity conditions and initial conditions in Definition C.2 and the following bounds for $0 \leq t' \leq t \leq T$:

$$\|\Sigma_\ell(t,t)\|_2 \leq \frac{\Phi_{\Sigma_\ell}\mathrm{e}^{2\overline{\lambda}t}}{\overline{\lambda}}, \quad \|R_\ell(t,t')\|_2 \leq \Phi_{R_\ell}\mathrm{e}^{\overline{\lambda}(t-t')}, \quad \|\Gamma(t)\|_2 \leq M. \tag{74}$$

Then, $\mathcal{S}_\ell$ is admissible with parameters $\Phi_\ell = \max\{\Phi_{\Sigma_\ell}\mathrm{e}^{2\overline{\lambda}T}/\overline{\lambda}, \Phi_{R_\ell}\mathrm{e}^{\overline{\lambda}T}\}$ and M_ℓ.

In the following, we show that

1. For sufficiently small T with $\overline{\lambda}T \leq 1$, the mappings $\mathcal{T}_{\theta\to\ell}$ and $\mathcal{T}_{\ell\to\theta}$ map $\mathcal{S}_\theta$ and $\mathcal{S}_\ell$ into each other, respectively.

2. If either $\tau = 0$ or $\nabla_r^2 \ell_t(r; z) = 0$, the above holds for any $T > 0$.

Note that it is possible to take $\overline{\lambda} T \leq 1$ since $\overline{\lambda}$ is monotonically increasing in T. Then, we can take T_* as the supremum of such T. We provide a rough estimate of $T_* \gtrsim \delta^2/(\tau^2 M^7 m^2)$ where $\gtrsim$ hides subleading terms in $M, m, \tau, 1/\delta$ and the constant factor in Appendix C.3.4.

The proof is almost identical for both cases; the only difference lies in bounding R_ℓ.

C.3.2. $\mathcal{T}_{\theta\to\ell}$ MAPS $\mathcal{S}_\theta$ INTO $\mathcal{S}_\ell$.

Condition for Σ_ℓ. We have by the assumptions and the triangle inequality that

$$\mathbb{E}\|\ell_t(r^t; z)\|_2^2 \leq \mathbb{E}\big(\|\ell_t(0; z)\|_2 + M\|r^t\|_2\big)^2 \leq 2\,\mathbb{E}\|\ell_t(0; z)\|_2^2 + 2M^2\,\mathbb{E}\|r^t\|_2^2 \leq 2M + 2M^2\,\mathbb{E}\|r^t\|_2^2\,. \tag{75}$$

We apply Lemma C.8 to r^t in Equation (21) with

$$\mathbb{E}\|X_1\|_2^2 \coloneqq \mathbb{E}\|w^t\|_2^2 = \operatorname{tr}(C_\theta(t,t)) \leq m\|C_\theta(t,t)\|_2 \leq m\Phi_{C_\theta}\mathrm{e}^{2\overline{\lambda}t}\,, \tag{76}$$

$$\mathbb{E}\|X_2^s\|_2^2 \coloneqq \mathbb{E}\left\|\frac{1}{\delta}R_\theta(t,s)\ell_s(r^s; z)\right\|_2^2 \leq \frac{1}{\delta^2}\Phi_{R_\theta}^2\mathrm{e}^{2\overline{\lambda}(t-s)}\,\mathbb{E}\|\ell_s(r^s; z)\|_2^2\,, \tag{77}$$

$$\mathbb{E}\|X_3^s\|_2^2 \coloneqq \mathbb{E}\left\|\sqrt{\frac{\tau}{\delta}}R_\theta(t,s)\ell_s(r^s; z)\right\|_2^2 \leq \frac{\tau}{\delta}\Phi_{R_\theta}^2\mathrm{e}^{2\overline{\lambda}(t-s)}\,\mathbb{E}\|\ell_s(r^s; z)\|_2^2\,, \tag{78}$$

to obtain

$$\mathbb{E}\|r^t\|_2^2 \leq 3\left(m\Phi_{C_\theta}\mathrm{e}^{2\overline{\lambda}t} + \left(\frac{T}{\delta^2} + \frac{\tau}{\delta}\right)\Phi_{R_\theta}^2\int_0^t \mathrm{e}^{2\overline{\lambda}(t-s)}\,\mathbb{E}\|\ell_s(r^s; z)\|_2^2\,\mathrm{d}s\right). \tag{79}$$

Thus, we have

$$\mathrm{e}^{-2\overline{\lambda}t}\,\mathbb{E}\|\ell_t(r^t; z)\|_2^2 \leq 2M + 6M^2 m\Phi_{C_\theta} + 6M^2\left(\frac{T}{\delta^2} + \frac{\tau}{\delta}\right)\Phi_{R_\theta}^2\int_0^t \mathrm{e}^{-2\overline{\lambda}s}\,\mathbb{E}\|\ell_s(r^s; z)\|_2^2\,\mathrm{d}s\,. \tag{80}$$

By Grönwall's inequality, we have

$$\mathbb{E}\|\ell_t(r^t; z)\|_2^2 \leq (2M + 6M^2 m\Phi_{C_\theta})\exp\left(6M^2\left(\frac{T}{\delta^2} + \frac{\tau}{\delta}\right)\Phi_{R_\theta}^2 T\right)\mathrm{e}^{2\overline{\lambda}t} = \Phi_{C_\ell}\mathrm{e}^{2\overline{\lambda}t}\,, \tag{81}$$

where we used the definition of Φ_{C_ℓ} in Equation (62).

Now, we check the condition $\|\Sigma_\ell(t,t)\|_2 \leq \Phi_{\Sigma_\ell}\mathrm{e}^{2\overline{\lambda}t}/\overline{\lambda}$. Using Lemma C.8, we have

$$\begin{aligned}\|\Sigma_\ell(t,t)\|_2 &= \|\mathbb{E}[L^t L^{t\mathsf{T}}]\|_2 \leq \mathbb{E}\|L^t\|_2^2 \leq 2(T+\tau\delta)\int_0^t \mathbb{E}\|\ell_s(r^s; z)\|_2^2\,\mathrm{d}s \\ &\leq 2(T+\tau\delta)\int_0^t \Phi_{C_\ell}\mathrm{e}^{2\overline{\lambda}s}\,\mathrm{d}s = \frac{(T+\tau\delta)\Phi_{C_\ell}}{\overline{\lambda}}(\mathrm{e}^{2\overline{\lambda}t} - 1) \leq \frac{\Phi_{\Sigma_\ell}}{\overline{\lambda}}\mathrm{e}^{2\overline{\lambda}t}\,,\end{aligned} \tag{82}$$

where we used the definition of Φ_{Σ_ℓ} in Equation (63).

Next, we check the condition (51). We have

$$\begin{aligned}&\|\Sigma_\ell(t,t) - 2\Sigma_\ell(t,t') + \Sigma_\ell(t',t')\|_2 = \|\mathbb{E}[(L^t - L^{t'})(L^t - L^{t'})^\mathsf{T}]\|_2 \leq \mathbb{E}\|L^t - L^{t'}\|_2^2 \\ &= \mathbb{E}\left\|\int_{t'}^t \ell_s(r^s; z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\right\|_2^2 \leq 2(T+\tau\delta)\int_{t'}^t \mathbb{E}\|\ell_s(r^s; z)\|_2^2\,\mathrm{d}s \\ &\leq 2(T+\tau\delta)\Phi_{C_\ell}\mathrm{e}^{2\overline{\lambda}T}|t-t'| \leq M_\ell|t-t'|\,.\end{aligned} \tag{83}$$

Condition for R_ℓ: case (1). We have

$$\|R_\ell(t,t')\|_2^2 \le 2\,\mathbb{E}\|\rho_\ell^{t,t'}\|_2^2 + 2\tau\delta\,\mathbb{E}\|D_\ell^{t,t'}\|_2^2\,. \tag{84}$$

First, we bound $\mathbb{E}\|\rho_\ell^{t,t'}\|_2^2$. By the Lipschitz continuity of $\ell_t(r;z)$ in r, we have $\|\nabla_r \ell_t(r^t;z)\|_2 \le M$ for all t. Therefore, we have

$$\mathbb{E}\|\rho_\ell^{t,t'}\|_2^2 = \mathbb{E}\|\nabla_r \ell_t(r^t;z)\rho_r^{t,t'}\|_2^2 \le M^2\,\mathbb{E}\|\rho_r^{t,t'}\|_2^2\,. \tag{85}$$

Applying Lemma C.8 to $\rho_r^{t,t'}$ in Equation (22), we have

$$\mathbb{E}\|\rho_r^{t,t'}\|_2^2 \le 3\left(\frac{M^2}{\delta^2}\Phi_{R_\theta}^2 \mathrm{e}^{2\overline{\lambda}(t-t')} + \left(\frac{T}{\delta^2}+\frac{\tau}{\delta}\right)\Phi_{R_\theta}^2\int_{t'}^t \mathrm{e}^{2\overline{\lambda}(t-s)}\,\mathbb{E}\|\rho_\ell^{s,t'}\|_2^2\,\mathrm{d}s\right). \tag{86}$$

Thus, we have

$$\mathrm{e}^{-2\overline{\lambda}(t-t')}\,\mathbb{E}\|\rho_\ell^{t,t'}\|_2^2 \le \frac{3M^4}{\delta^2}\Phi_{R_\theta}^2 + 3M^2\left(\frac{T}{\delta^2}+\frac{\tau}{\delta}\right)\Phi_{R_\theta}^2\int_{t'}^t \mathrm{e}^{-2\overline{\lambda}(s-t')}\,\mathbb{E}\|\rho_\ell^{s,t'}\|_2^2\,\mathrm{d}s\,. \tag{87}$$

By Grönwall's inequality, we have

$$\mathrm{e}^{-2\overline{\lambda}(t-t')}\,\mathbb{E}\|\rho_\ell^{t,t'}\|_2^2 \le \frac{3M^4}{\delta^2}\Phi_{R_\theta}^2 \exp\left(3M^2\left(\frac{T}{\delta^2}+\frac{\tau}{\delta}\right)\Phi_{R_\theta}^2 T\right) \le \Phi_1\,. \tag{88}$$

Next, we bound $\mathbb{E}\|D_\ell^{t,t'}\|_2^2$. We have

$$\begin{aligned}\mathbb{E}\|D_\ell^{t,t'}\|_2^2 &= 4M^2\left(\frac{T}{\delta^2}+\frac{\tau}{\delta}\right)\int_{t'}^t \|R_\theta(t,s)\|_2^2\,\mathbb{E}\|D_\ell^{s,t'}\|_2^2\,\mathrm{d}s + 2\,\mathbb{E}[\|\nabla_r^2\ell_t(r^t;z)\|_2^2\|D_r^{t,t'}\|_2^2\|\rho_r^{t,t'}\|_2^2] \\ &\le 4M^2\left(\frac{T}{\delta^2}+\frac{\tau}{\delta}\right)\Phi_{R_\theta}^2\int_{t'}^t \mathrm{e}^{2\overline{\lambda}(t-s)}\,\mathbb{E}\|D_\ell^{s,t'}\|_2^2\,\mathrm{d}s + 2M^2\sqrt{\mathbb{E}\|D_r^{t,t'}\|_2^4}\sqrt{\mathbb{E}\|\rho_r^{t,t'}\|_2^4}\,,\end{aligned} \tag{89}$$

where we used that $\|\nabla_r^2\ell_t(r^t;z)\|_2 \le M$ by assumption. We first bound $\mathbb{E}\|\rho_r^{t,t'}\|_2^4$. Applying Lemma C.8, we obtain

$$\begin{aligned}\mathbb{E}\|\rho_r^{t,t'}\|_2^4 &\le 27\Bigg(\left(\frac{T^3}{\delta^4}+\frac{C_2\tau^2 T}{\delta^2}\right)\int_{t'}^t \|R_\theta(t,s)\|_2^4\,\mathbb{E}[\|\nabla_r\ell_s(r^s;z)\|_2^4\|\rho_r^{s,t'}\|_2^4]\,\mathrm{d}s \\ &\qquad + \frac{1}{\delta^4}\|R_\theta(t,t')\|_2^4\,\mathbb{E}\|\nabla_r\ell_{t'}(r^{t'};z)\|_2^4\Bigg) \\ &\le 27\left(\left(\frac{T^3}{\delta^4}+\frac{C_2\tau^2 T}{\delta^2}\right)\Phi_{R_\theta}^4 M^4\int_{t'}^t \mathrm{e}^{4\overline{\lambda}(t-s)}\,\mathbb{E}\|\rho_r^{s,t'}\|_2^4\,\mathrm{d}s + \frac{M^4}{\delta^4}\Phi_{R_\theta}^4\mathrm{e}^{4\overline{\lambda}(t-t')}\right).\end{aligned} \tag{90}$$

By Grönwall's inequality, we have

$$\mathbb{E}\|\rho_r^{t,t'}\|_2^4 \le \frac{27M^4\Phi_{R_\theta}^4}{\delta^4}\exp\left(27M^4\left(\frac{T^3}{\delta^4}+\frac{C_2\tau^2 T}{\delta^2}\right)\Phi_{R_\theta}^4 T\right)\mathrm{e}^{4\overline{\lambda}(t-t')} \le \Phi_2\mathrm{e}^{4\overline{\lambda}(t-t')}\,. \tag{91}$$

We next bound $\mathbb{E}\|D_r^{t,t'}\|_2^4$. Applying Lemma C.8, we have

$$\begin{aligned}\mathbb{E}\|D_r^{t,t'}\|_2^4 &\le 27\bigg(\left(\frac{T^3}{\delta^4}+\frac{C_2\tau^2T}{\delta^2}\right)\int_{t'}^t \|R_\theta(t,s)\|_2^4\,\mathbb{E}[\|\nabla_r\ell_s(r^s;z)\|_2^4\|D_r^{s,t'}\|_2^4]\,\mathrm{d}s \\ &\qquad + \frac{\tau^2}{\delta^2}\|R_\theta(t,t')\|_2^4\,\mathbb{E}\|\ell_{t'}(r^{t'};z)\|_2^4\bigg) \\ &\le 27\bigg(\left(\frac{T^3}{\delta^4}+\frac{C_2\tau^2T}{\delta^2}\right)\Phi_{R_\theta}^4M^4\int_{t'}^t \mathrm{e}^{4\overline{\lambda}(t-s)}\,\mathbb{E}\|D_r^{s,t'}\|_2^4\,\mathrm{d}s + \frac{\tau^2}{\delta^2}\Phi_{R_\theta}^4\mathrm{e}^{4\overline{\lambda}(t-t')}\,\mathbb{E}\|\ell_{t'}(r^{t'};z)\|_2^4\bigg)\,.\end{aligned} \tag{92}$$

$\mathbb{E}\|\ell_{t'}(r^{t'};z)\|_2^4$ can be bounded as

$$\begin{aligned}&\mathbb{E}\|\ell_t(r^t;z)\|_2^4 \le 8\,\mathbb{E}\|\ell_t(0;z)\|_2^4 + 8M^4\,\mathbb{E}\|r^t\|_2^4 \\ &\le 8M + 8M^4\cdot 27\bigg(\mathbb{E}\|w^t\|_2^4 + \left(\frac{T^3}{\delta^4}+\frac{C_2\tau^2T}{\delta^2}\right)\int_0^t\|R_\theta(t,s)\|_2^4\,\mathbb{E}\|\ell_s(r^s;z)\|_2^4\,\mathrm{d}s\bigg) \\ &\le 8M + 216M^4\bigg(3m^2\Phi_{C_\theta}^2\mathrm{e}^{4\overline{\lambda}t} + \left(\frac{T^3}{\delta^4}+\frac{C_2\tau^2T}{\delta^2}\right)\Phi_{R_\theta}^4\int_0^t\mathrm{e}^{4\overline{\lambda}(t-s)}\,\mathbb{E}\|\ell_s(r^s;z)\|_2^4\,\mathrm{d}s\bigg)\,.\end{aligned} \tag{93}$$

By Grönwall's inequality, we have

$$\mathbb{E}\|\ell_t(r^t;z)\|_2^4 \le \left(8M + 648M^4m^2\Phi_{C_\theta}^2\right)\exp\bigg(216M^4\left(\frac{T^3}{\delta^4}+\frac{C_2\tau^2T}{\delta^2}\right)\Phi_{R_\theta}^4T\bigg)\mathrm{e}^{4\overline{\lambda}t} \le \Phi_3\mathrm{e}^{4\overline{\lambda}t}\,. \tag{94}$$

Taking T small enough such that $\overline{\lambda}T \le 1$, we have $\mathrm{e}^{4\overline{\lambda}t} \le \mathrm{e}^4$. Thus, we have

$$\mathbb{E}\|D_r^{t,t'}\|_2^4 \le 27\bigg(\left(\frac{T^3}{\delta^4}+\frac{C_2\tau^2T}{\delta^2}\right)\Phi_{R_\theta}^4M^4\int_{t'}^t\mathrm{e}^{4\overline{\lambda}(t-s)}\,\mathbb{E}\|D_r^{s,t'}\|_2^4\,\mathrm{d}s + \frac{\tau^2}{\delta^2}\Phi_{R_\theta}^4\Phi_3\mathrm{e}^4\mathrm{e}^{4\overline{\lambda}(t-t')}\bigg)\,. \tag{95}$$

By Grönwall's inequality, we have

$$\mathbb{E}\|D_r^{t,t'}\|_2^4 \le \frac{27\tau^2\Phi_{R_\theta}^4\Phi_3\mathrm{e}^4}{\delta^2}\exp\bigg(27M^4\left(\frac{T^3}{\delta^4}+\frac{C_2\tau^2T}{\delta^2}\right)\Phi_{R_\theta}^4T\bigg)\mathrm{e}^{4\overline{\lambda}(t-t')} \le \Phi_4\mathrm{e}^4\,. \tag{96}$$

Combining the above bounds, we have

$$\mathbb{E}\|D_\ell^{t,t'}\|_2^2 \le 4M^2\left(\frac{T}{\delta^2}+\frac{\tau}{\delta}\right)\Phi_{R_\theta}^2\int_{t'}^t\mathrm{e}^{2\overline{\lambda}(t-s)}\,\mathbb{E}\|D_\ell^{s,t'}\|_2^2\,\mathrm{d}s + 2M^2\sqrt{\Phi_4\mathrm{e}^4}\sqrt{\Phi_2\mathrm{e}^{4\overline{\lambda}(t-t')}}\,. \tag{97}$$

By Grönwall's inequality, we have

$$\mathbb{E}\|D_\ell^{t,t'}\|_2^2 \le 2M^2\sqrt{\Phi_2\Phi_4}\mathrm{e}^2\exp\bigg(4M^2\left(\frac{T}{\delta^2}+\frac{\tau}{\delta}\right)\Phi_{R_\theta}^2T\bigg)\mathrm{e}^{2\overline{\lambda}(t-t')} \le \Phi_5\mathrm{e}^{2\overline{\lambda}(t-t')}\,. \tag{98}$$

Thus, we have

$$\|R_\ell(t,t')\|_2^2 \le (2\Phi_1 + 2\tau\delta\Phi_5)\mathrm{e}^{2\overline{\lambda}(t-t')} = \Phi_{R_\ell}^2\mathrm{e}^{2\overline{\lambda}(t-t')}\,. \tag{99}$$

Condition for R_ℓ: case (2). When $\nabla_r^2 \ell_t(r;z) = 0$, we have $D_\ell^{t,t'} = 0$. Thus, when $\tau = 0$ or $\nabla^2 \ell_t(r;z) = 0$, we have

$$\|R_\ell(t,t')\|_2^2 \le \mathbb{E}\|\rho_\ell^{t,t'}\|_2^2 \le \Phi_1 \mathrm{e}^{-2\overline{\lambda}(t-t')} \le \Phi_{R_\ell}^2 \mathrm{e}^{2\overline{\lambda}(t-t')} . \tag{100}$$

This holds without taking T small.

Condition for Γ. We have

$$\|\Gamma(t)\|_2 \le \mathbb{E}\|\nabla_r \ell_t(r^t;z)\|_2 \le M . \tag{101}$$

C.3.3. $\mathcal{T}_{\ell\to\theta}$ MAPS $\mathcal{S}_\ell$ INTO $\mathcal{S}_\theta$.

Condition for C_θ. We have

$$\mathbb{E}\|U^t\|_2^2 = \frac{1}{\delta}\operatorname{tr}(\Sigma_\ell(t,t)) \le \frac{m}{\delta}\|\Sigma_\ell(t,t)\|_2 \le \frac{m\Phi_{\Sigma_\ell}}{\delta\overline{\lambda}}\mathrm{e}^{2\overline{\lambda}t} . \tag{102}$$

By the Lipschitz continuity of h, we have $\mathbb{E}\|h_t(\theta^t)\|_2^2 \le 2\|h_t(0)\|_2^2 + 2\,\mathbb{E}\|h_t(\theta^t) - h_t(0)\|_2^2 \le 2M^2 + 2M^2\,\mathbb{E}\|\theta^t\|_2^2$. Thus, we have

$$\begin{aligned}
&\mathrm{e}^{-2\overline{\lambda}t}\,\mathbb{E}\|\theta^t\|_2^2 \\
&\le 3\mathrm{e}^{-2\overline{\lambda}t}\Bigg(\mathbb{E}\|\theta^0\|_2^2 + \mathbb{E}\|U^t\|_2^2 \\
&\qquad + 2T\int_0^t \left(\mathbb{E}\|h_s(\theta^s)\|_2^2 + \|\Gamma(s)\|_2^2\,\mathbb{E}\|\theta^s\|_2^2 + T\int_0^s \|R_\ell(s,s')\|_2^2\,\mathbb{E}\|\theta^{s'}\|_2^2\,\mathrm{d}s'\right)\mathrm{d}s\Bigg) \\
&\le 3\Bigg(\mathrm{e}^{-2\overline{\lambda}t}(M + 4T^2M^2) + \frac{m\Phi_{\Sigma_\ell}}{\delta\overline{\lambda}} \\
&\qquad + 2T\int_0^t \mathrm{e}^{-2\overline{\lambda}(t-s)}\left(3M^2\mathrm{e}^{-2\overline{\lambda}s}\,\mathbb{E}\|\theta^s\|_2^2 + T\Phi_{R_\ell}^2\int_0^s \mathrm{e}^{-2\overline{\lambda}s'}\,\mathbb{E}\|\theta^{s'}\|_2^2\,\mathrm{d}s'\right)\mathrm{d}s\Bigg) \\
&\le 3\left(M + 4T^2M^2 + \frac{m\Phi_{\Sigma_\ell}}{\delta\overline{\lambda}} + 2T(3M^2 + T^2\Phi_{R_\ell}^2)\left(\int_0^t \mathrm{e}^{-2\overline{\lambda}(t-s)}\,\mathrm{d}s\right)\sup_{s\in[0,T]}\mathrm{e}^{-2\overline{\lambda}s}\,\mathbb{E}\|\theta^s\|_2^2\right) \\
&\le 3\left(M + 4T^2M^2 + \frac{m\Phi_{\Sigma_\ell}}{\delta\overline{\lambda}} + \frac{T(3M^2 + T^2\Phi_{R_\ell}^2)}{\overline{\lambda}}\sup_{s\in[0,T]}\mathrm{e}^{-2\overline{\lambda}s}\,\mathbb{E}\|\theta^s\|_2^2\right)
\end{aligned} \tag{103}$$

By $\overline{\lambda} \ge 2T(3M^2 + T^2\Phi_{R_\ell}^2)$, we have

$$\sup_{t\in[0,T]}\mathrm{e}^{-2\overline{\lambda}t}\,\mathbb{E}\|\theta^t\|_2^2 \le 6\left(M + 4T^2M^2 + \frac{m\Phi_{\Sigma_\ell}}{\delta\overline{\lambda}}\right). \tag{104}$$

By $\overline{\lambda} \ge 6m\Phi_{\Sigma_\ell}/(\delta(M + 4T^2M^2))$, the right-hand side is bounded by $7(M + 4T^2M^2) \le \Phi_{C_\theta}$. Therefore, we have

$$\|C_\theta(t,t)\|_2 \le \mathbb{E}\|\theta^t\|_2^2 \le \Phi_{C_\theta}\mathrm{e}^{2\overline{\lambda}t} . \tag{105}$$

Next, we check the condition (50). We have

$$
\begin{aligned}
&\|C_\theta(t,t) - 2C_\theta(t,t') + C_\theta(t',t')\|_2 = \|\mathbb{E}[(\theta^t - \theta^{t'})(\theta^t - \theta^{t'})^\mathsf{T}]\|_2 \le \mathbb{E}\|\theta^t - \theta^{t'}\|_2^2 \\
&\le \mathbb{E}\left\| U^t - U^{t'} - \int_{t'}^t \left(h_s(\theta^s) + \Gamma(s)\theta^s + \int_0^s R_\ell(s,s')\theta^{s'}\,\mathrm{d}s' \right)\mathrm{d}s \right\|_2^2 \\
&\le 2\left(\mathbb{E}\|U^t - U^{t'}\|_2^2 + 2(t-t')^2 \sup_{s\in[0,T]} \left\{ 2M^2 + 3M^2\,\mathbb{E}\|\theta^s\|_2^2 + T\int_0^s \|R_\ell(s,s')\|_2^2\,\mathbb{E}\|\theta^{s'}\|_2^2\,\mathrm{d}s' \right\} \right) \\
&\le 2\left(\frac{mM_\ell}{\delta}|t-t'| + 2(t-t')^2\Big(2M^2 + 3M^2\Phi_{C_\theta}\mathrm{e}^{2\overline{\lambda}T} + T^2\Phi_{R_\ell}^2\Phi_{C_\theta}\mathrm{e}^{4\overline{\lambda}T} \Big) \right) \\
&\le M_\theta|t-t'|\,,
\end{aligned}
\tag{106}
$$

where we used

$$
\mathbb{E}\|U^t - U^{t'}\|_2^2 \le \frac{m}{\delta}\|\Sigma_\ell(t,t) - 2\Sigma_\ell(t,t') + \Sigma_\ell(t',t')\|_2 \le \frac{mM_\ell}{\delta}|t-t'|\,. \tag{107}
$$

Condition for R_θ. By the Lipschitz continuity of h, we have $\|\nabla_\theta h_t(\theta)\|_2 \le M$ for all t and θ. Thus, we have

$$
\begin{aligned}
\|\rho_\theta^{t,t'}\|_2 &\le 1 + \int_{t'}^t \left((\|\nabla_\theta h_s(\theta^s)\|_2 + \|\Gamma(s)\|_2)\|\rho_\theta^{s,t'}\|_2 + \int_{t'}^s \|R_\ell(s,s')\|_2\|\rho_\theta^{s',t'}\|_2\,\mathrm{d}s' \right)\mathrm{d}s \\
&\le 1 + \int_{t'}^t \left(2M\|\rho_\theta^{s,t'}\|_2 + \Phi_{R_\ell}\int_{t'}^s \mathrm{e}^{\overline{\lambda}(s-s')}\|\rho_\theta^{s',t'}\|_2\,\mathrm{d}s' \right)\mathrm{d}s\,,
\end{aligned}
\tag{108}
$$

and thus

$$
\begin{aligned}
&\mathrm{e}^{-\overline{\lambda}(t-t')}\|\rho_\theta^{t,t'}\|_2 \\
&\le 1 + \int_{t'}^t \mathrm{e}^{-\overline{\lambda}(t-s)}\left(2M\mathrm{e}^{-\overline{\lambda}(s-t')}\|\rho_\theta^{s,t'}\|_2 + \Phi_{R_\ell}\int_{t'}^s \mathrm{e}^{\overline{\lambda}(s'-t')}\|\rho_\theta^{s',t'}\|_2\,\mathrm{d}s' \right)\mathrm{d}s \\
&\le 1 + (2M + \Phi_{R_\ell}T)\left(\int_{t'}^t \mathrm{e}^{-\overline{\lambda}(t-s)}\,\mathrm{d}s \right) \sup_{s\in[t',t]} \mathrm{e}^{-\overline{\lambda}(s-t')}\|\rho_\theta^{s,t'}\|_2 \\
&\le 1 + \frac{2M + \Phi_{R_\ell}T}{\overline{\lambda}} \sup_{s\in[t',t]} \mathrm{e}^{-\overline{\lambda}(s-t')}\|\rho_\theta^{s,t'}\|_2\,.
\end{aligned}
\tag{109}
$$

By $\overline{\lambda} \ge 2(2M + \Phi_{R_\ell}T)$, we have

$$
\|\rho_\theta^{t,t'}\|_2 \le 2\mathrm{e}^{\overline{\lambda}(t-t')}\,, \tag{110}
$$

and thus

$$
\|R_\theta(t,t')\|_2 \le \mathbb{E}\|\rho_\theta^{t,t'}\|_2 \le 2\mathrm{e}^{\overline{\lambda}(t-t')} \le \Phi_{R_\theta}\mathrm{e}^{\overline{\lambda}(t-t')}\,. \tag{111}
$$

C.3.4. A ROUGH ESTIMATE OF T_*

We derive a rough lower bound on T_* up to the leading dependencies on $M, m, \tau, 1/\delta$ and ignoring constant factors. Take $T \le \min\{1, (M^2(1/\delta^2 + \tau/\delta)\Phi_{R_\theta}^2)^{-1}\}$. Then, the exponents in the definitions of Φ_{C_ℓ} and Φ_i $(i = 1, \dots, 5)$ are all bounded by constants since

$$
M^2\left(\frac{T}{\delta^2} + \frac{\tau}{\delta} \right)\Phi_{R_\theta}^2 T \le 1\,, \quad M^4\left(\frac{T^2}{\delta^4} + \frac{C_2\tau^2}{\delta^2} \right)\Phi_{R_\theta}^4 T^2 \lesssim 1\,. \tag{112}
$$

Then, the Φ quantities can be bounded as

$$\Phi_{C_\theta} \lesssim M\,, \quad \Phi_{R_\theta} \lesssim 1\,, \quad \Phi_{C_\ell} \lesssim M^3 m\,, \quad \Phi_{\Sigma_\ell} \lesssim \tau M^3 m\,, \quad \Phi_1 \lesssim \frac{M^4}{\delta^2}\,, \quad \Phi_2 \lesssim \frac{M^4}{\delta^4}\,, \\ \Phi_3 \lesssim M^6 m^2\,, \quad \Phi_4 \lesssim \frac{\tau^2 M^6 m^2}{\delta^2}\,, \quad \Phi_5 \lesssim \frac{\tau M^7 m}{\delta^3}\,, \quad \Phi_{R_\ell} \le \frac{\tau M^{7/2} m^{1/2}}{\delta} \tag{113}$$

Then, $\overline{\lambda}$ is bounded as

$$\overline{\lambda} \lesssim (M^2 + \Phi_{R_\ell}^2) + \frac{m\Phi_{\Sigma_\ell}}{\delta M} + (M + \Phi_{R_\ell}) \lesssim \frac{\tau^2 M^7 m^2}{\delta^2}\,. \tag{114}$$

Thus, further taking $T \lesssim \delta^2/(\tau^2 M^7 m^2)$, we have $\overline{\lambda} T \lesssim 1$.

C.4. Proof of Lemma C.6

Set $\Phi \coloneqq \max\{\Phi_\theta, \Phi_\ell\}$. In the following, K denotes a positive constant that may depend on M, m, T, δ, τ, and Φ, but not on other variables, and may change from line to line.

Bound of $\mathsf{dist}_\lambda(\Sigma_\ell^1, \Sigma_\ell^2)$. Let $w_1 \sim \mathsf{GP}(0, C_\theta^1)$ and $w_2 \sim \mathsf{GP}(0, C_\theta^2)$ be Gaussian processes coupled such that

$$\sup_{t\in[0,T]} \mathrm{e}^{-\lambda t}\sqrt{\mathbb{E}\|w_1^t - w_2^t\|_2^2} \le 2 \cdot \mathsf{dist}_\lambda(C_\theta^1, C_\theta^2)\,. \tag{115}$$

For $i = 1, 2$, let r_i be the solution of

$$r_i^t = w_i^t - \frac{1}{\delta}\int_0^t R_\theta^i(t,s)\ell_s(r_i^s; z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\,. \tag{116}$$

Note that we use the same Brownian motion B^t for $i = 1, 2$. Applying Lemma C.8 to $r_1^t - r_2^t$ with

$$\mathbb{E}\|w_1^t - w_2^t\|_2^2 \le 4\mathrm{e}^{2\lambda t} \cdot \mathsf{dist}_\lambda(C_\theta^1, C_\theta^2)^2\,, \tag{117}$$

$$\begin{aligned}
&\mathbb{E}\|R_\theta^1(t,s)\ell_s(r_1^s; z) - R_\theta^2(t,s)\ell_s(r_2^s; z)\|_2^2 \\
&\quad \le 2\big(\|R_\theta^1(t,s) - R_\theta^2(t,s)\|_2^2\,\mathbb{E}\|\ell_s(r_1^s; z)\|_2^2 + \|R_\theta^2(t,s)\|_2^2\,\mathbb{E}\|\ell_s(r_1^s; z) - \ell_s(r_2^s; z)\|_2^2\big) \\
&\quad \le 2\Big(\Phi\mathrm{e}^{2\lambda t} \cdot \mathsf{dist}_\lambda(R_\theta^1, R_\theta^2)^2 + \Phi^2 M^2\,\mathbb{E}\|r_1^s - r_2^s\|_2^2\Big)\,,
\end{aligned} \tag{118}$$

we have

$$\begin{aligned}
\mathbb{E}\|r_1^t - r_2^t\|_2^2 \le 3\Bigg(&4\mathrm{e}^{2\lambda t} \cdot \mathsf{dist}_\lambda(C_\theta^1, C_\theta^2)^2 + 2\left(\frac{T}{\delta^2} + \frac{\tau}{\delta}\right)T\Phi\mathrm{e}^{2\lambda t} \cdot \mathsf{dist}_\lambda(R_\theta^1, R_\theta^2)^2 \\
&+ 2\left(\frac{T}{\delta^2} + \frac{\tau}{\delta}\right)\Phi^2 M^2 \int_0^t \mathbb{E}\|r_1^s - r_2^s\|_2^2\,\mathrm{d}s\Bigg)\,,
\end{aligned} \tag{119}$$

and thus

$$\mathrm{e}^{-2\lambda t}\,\mathbb{E}\|r_1^t - r_2^t\|_2^2 \le K \cdot \mathsf{dist}_\lambda(X^1, X^2)^2 + K\int_0^t \mathrm{e}^{-2\lambda s}\,\mathbb{E}\|r_1^s - r_2^s\|_2^2\,\mathrm{d}s\,. \tag{120}$$

By Grönwall's inequality, we have

$$\mathrm{e}^{-2\lambda t}\,\mathbb{E}\|r_1^t - r_2^t\|_2^2 \le K\mathrm{e}^{KT}\cdot \mathsf{dist}_\lambda(X^1,X^2)^2 \le K\cdot \mathsf{dist}_\lambda(X^1,X^2)^2\,. \tag{121}$$

For $i=1,2$, let

$$L_i^t := \int_0^t \ell_s(r_i^s;z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\,. \tag{122}$$

Then, we have

$$\begin{aligned}\mathrm{e}^{-2\lambda t}\,\mathbb{E}\|L_1^t - L_2^t\|_2^2 &\le 2(T+\tau\delta)\int_0^t \mathrm{e}^{-2\lambda(t-s)}\cdot \mathrm{e}^{-2\lambda s}\,\mathbb{E}\|\ell_s(r_1^s;z)-\ell_s(r_2^s;z)\|_2^2\,\mathrm{d}s\\ &\le \frac{(T+\tau\delta)M^2}{\lambda}\sup_{s\in[0,T]}\mathrm{e}^{-2\lambda s}\,\mathbb{E}\|r_1^s - r_2^s\|_2^2 \le \frac{K}{\lambda}\cdot \mathsf{dist}_\lambda(X^1,X^2)^2\,.\end{aligned} \tag{123}$$

Let $\{(U_1^t,U_2^t)\}_{t\in[0,T]}$ be a centered Gaussian process with covariance $\mathbb{E}\left[\begin{pmatrix}L_1^t\\L_2^t\end{pmatrix}\begin{pmatrix}L_1^t\\L_2^t\end{pmatrix}^\mathsf{T}\right]/\delta$. Since U_1 and U_2 have covariance kernels Σ_ℓ^1/δ and Σ_ℓ^2/δ respectively, we have

$$\mathsf{dist}_\lambda(\Sigma_\ell^1,\Sigma_\ell^2) \le \sup_{t\in[0,T]}\mathrm{e}^{-\lambda t}\sqrt{\mathbb{E}\|U_1^t-U_2^t\|_2^2} = \sup_{t\in[0,T]}\mathrm{e}^{-\lambda t}\sqrt{\mathbb{E}\|L_1^t-L_2^t\|_2^2/\delta} \le \sqrt{\frac{K}{\lambda}}\cdot \mathsf{dist}_\lambda(X^1,X^2)\,. \tag{124}$$

Bound of $\mathsf{dist}_\lambda(\Gamma^1,\Gamma^2)$**.** By Equation (121), we have

$$\begin{aligned}\mathsf{dist}_\lambda(\Gamma^1,\Gamma^2) &= \sup_{t\in[0,T]}\mathrm{e}^{-\lambda t}\|\Gamma^1(t)-\Gamma^2(t)\|_2 \le \sup_{t\in[0,T]}\mathrm{e}^{-\lambda t}\sqrt{\mathbb{E}\|\nabla_r\ell_t(r_1^t;z)-\nabla_r\ell_t(r_2^t;z)\|_2^2}\\ &\le M\sup_{t\in[0,T]}\mathrm{e}^{-\lambda t}\sqrt{\mathbb{E}\|r_1^t-r_2^t\|_2^2} \le K\cdot \mathsf{dist}_\lambda(X^1,X^2)\,.\end{aligned} \tag{125}$$

Bound of $\mathsf{dist}_\lambda(R_\ell^1,R_\ell^2)$**.** For $i=1,2$, let

$$\rho_{\ell,i}^{t,t'} = \nabla_r\ell_t(r_i^t;z)\rho_{r,i}^{t,t'}\,,\quad \rho_{r,i}^{t,t'} = -\frac{1}{\delta}\int_{t'}^t R_\theta^i(t,s)\rho_{\ell,i}^{s,t'}(\mathrm{d}s+\sqrt{\tau\delta}\,\mathrm{d}B^s) - \frac{1}{\delta}R_\theta^i(t,t')\nabla_r\ell_{t'}(r_i^{t'};z)\,. \tag{126}$$

Then, we have

$$\|R_\ell^1(t,t')-R_\ell^2(t,t')\|_2^2 \le 2\,\mathbb{E}\|\rho_{\ell,1}^{t,t'}-\rho_{\ell,2}^{t,t'}\|_2^2 + 2\tau\delta\,\mathbb{E}\|D_{\ell,1}^{t,t'}-D_{\ell,2}^{t,t'}\|_2^2\,. \tag{127}$$

We first bound $\mathbb{E}\|\rho_{\ell,1}^{t,t'}-\rho_{\ell,2}^{t,t'}\|_2^2$. We have

$$\begin{aligned}\mathbb{E}\|\rho_{\ell,1}^{t,t'}-\rho_{\ell,2}^{t,t'}\|_2^2 &= \mathbb{E}\|\nabla_r\ell_t(r_1^t;z)\rho_{r,1}^{t,t'}-\nabla_r\ell_t(r_2^t;z)\rho_{r,2}^{t,t'}\|_2^2\\ &\le 2\,\mathbb{E}\|(\nabla_r\ell_t(r_1^t;z)-\nabla_r\ell_t(r_2^t;z))\rho_{r,1}^{t,t'}\|_2^2 + 2\,\mathbb{E}\|\nabla_r\ell_t(r_2^t;z)(\rho_{r,1}^{t,t'}-\rho_{r,2}^{t,t'})\|_2^2\\ &\le 2M^2\sqrt{\mathbb{E}\|\rho_{r,1}^{t,t'}\|_2^4}\sqrt{\mathbb{E}\|r_1^t-r_2^t\|_2^4} + 2M^2\,\mathbb{E}\|\rho_{r,1}^{t,t'}-\rho_{r,2}^{t,t'}\|_2^2\,.\end{aligned} \tag{128}$$

First, we bound $\mathbb{E}\|\rho_{r,1}^{t,t'}\|_2^4$. Applying Lemma C.8 to $\rho_{r,1}^{t,t'}$, we have

$$\mathbb{E}\|\rho_{r,1}^{t,t'}\|_2^4 \le 27\left(\frac{\Phi^4 M^4}{\delta^4} + \left(\frac{T^3}{\delta^4} + \frac{3T\tau^2}{\delta^2}\right)\Phi^4 M^4 \int_{t'}^{t} \mathbb{E}\|\rho_{r,1}^{s,t'}\|_2^4 \,\mathrm{d}s\right) \le K + K\int_{t'}^{t} \mathbb{E}\|\rho_{r,1}^{s,t'}\|_2^4 \,, \tag{129}$$

By Grönwall's inequality, we have

$$\mathbb{E}\|\rho_{r,1}^{t,t'}\|_2^4 \le K\mathrm{e}^{KT} \le K\,. \tag{130}$$

Next, we bound $\mathbb{E}\|r_1^t - r_2^t\|_2^4$ by applying Lemma C.8 with

$$\mathbb{E}\|w_1^t - w_2^t\|_2^4 \le 3(\mathbb{E}\|w_1^t - w_2^t\|_2^2)^2 \le 12\mathrm{e}^{4\lambda t}\cdot \mathsf{dist}_\lambda(C_\theta^1, C_\theta^2)^4\,, \tag{131}$$

$$\begin{aligned}
&\mathbb{E}\|R_\theta^1(t,s)\ell_s(r_1^s;z) - R_\theta^2(t,s)\ell_s(r_2^s;z)\|_2^4 \\
&\le 8\big(\|R_\theta^1(t,s) - R_\theta^2(t,s)\|_2^4\,\mathbb{E}\|\ell_s(r_1^s;z)\|_2^4 + \|R_\theta^2(t,s)\|_2^4\,\mathbb{E}\|\ell_s(r_1^s;z) - \ell_s(r_2^s;z)\|_2^4\big) \\
&\le 8\Big(K\mathrm{e}^{4\lambda t}\cdot \mathsf{dist}_\lambda(R_\theta^1, R_\theta^2)^4 + \Phi^4 M^4\,\mathbb{E}\|r_1^s - r_2^s\|_2^4\Big)\,,
\end{aligned} \tag{132}$$

Here, we used that $\mathbb{E}\|\ell_t(r^t;z)\|_2^4$ is uniformly bounded by some constant $K > 0$ by a similar argument as the bound on $\mathbb{E}\|\rho_r^{t,t'}\|_2^4$. Then, we have

$$\mathbb{E}\|r_1^t - r_2^t\|_2^4 \le K\mathrm{e}^{4\lambda t}\cdot \mathsf{dist}_\lambda(X^1, X^2)^4 + K\int_0^t \mathbb{E}\|r_1^s - r_2^s\|_2^4 \,\mathrm{d}s\,. \tag{133}$$

By Grönwall's inequality, we have

$$\mathrm{e}^{-4\lambda t}\,\mathbb{E}\|r_1^t - r_2^t\|_2^4 \le K\mathrm{e}^{KT}\cdot \mathsf{dist}_\lambda(X^1, X^2)^4 \le K\cdot \mathsf{dist}_\lambda(X^1, X^2)^4\,. \tag{134}$$

Finally, we bound $\mathbb{E}\|\rho_{r,1}^{t,t'} - \rho_{r,2}^{t,t'}\|_2^2$. We apply Lemma C.8 to $\rho_{r,1}^{t,t'} - \rho_{r,2}^{t,t'}$ with

$$\begin{aligned}
&\mathbb{E}\|R_\theta^1(t,t')\nabla_r\ell_{t'}(r_1^{t'};z) - R_\theta^2(t,t')\nabla_r\ell_{t'}(r_2^{t'};z)\|_2^2 \\
&\le 2\Big(\|R_\theta^1(t,t') - R_\theta^2(t,t')\|_2^2\,\mathbb{E}\|\nabla_r\ell_{t'}(r_1^{t'};z)\|_2^2 + \|R_\theta^2(t,t')\|_2^2\,\mathbb{E}\|\nabla_r\ell_{t'}(r_1^{t'};z) - \nabla_r\ell_{t'}(r_2^{t'};z)\|_2^2\Big) \\
&\le 2\Big(M^2\mathrm{e}^{2\lambda t}\cdot \mathsf{dist}_\lambda(R_\theta^1, R_\theta^2)^2 + \Phi^2 M^2\,\mathbb{E}\|r_1^{t'} - r_2^{t'}\|_2^2\Big) \le K\mathrm{e}^{2\lambda t}\cdot \mathsf{dist}_\lambda(X^1, X^2)^2\,,
\end{aligned} \tag{135}$$

$$\begin{aligned}
&\mathbb{E}\|R_\theta^1(t,s)\rho_{\ell,1}^{s,t'} - R_\theta^2(t,s)\rho_{\ell,2}^{s,t'}\|_2^2 \\
&\le 2\Big(\|R_\theta^1(t,s) - R_\theta^2(t,s)\|_2^2\,\mathbb{E}\|\rho_{\ell,1}^{s,t'}\|_2^2 + \|R_\theta^2(t,s)\|_2^2\,\mathbb{E}\|\rho_{\ell,1}^{s,t'} - \rho_{\ell,2}^{s,t'}\|_2^2\Big) \\
&\le 2\Big(K\mathrm{e}^{2\lambda t}\cdot \mathsf{dist}_\lambda(R_\theta^1, R_\theta^2)^2 + \Phi^2\,\mathbb{E}\|\rho_{\ell,1}^{s,t'} - \rho_{\ell,2}^{s,t'}\|_2^2\Big)\,.
\end{aligned} \tag{136}$$

Here, we used that $\mathbb{E}\|\rho_{\ell,i}^{s,t'}\|_2^2$ is uniformly bounded by some constant $K > 0$ by a similar argument as the bound on $\mathbb{E}\|\rho_r^{t,t'}\|_2^2$. Then, we have

$$\mathbb{E}\|\rho_{r,1}^{t,t'} - \rho_{r,2}^{t,t'}\|_2^2 \le K\mathrm{e}^{2\lambda t}\cdot \mathsf{dist}_\lambda(X^1, X^2)^2 + K\int_{t'}^{t} \mathbb{E}\|\rho_{\ell,1}^{s,t'} - \rho_{\ell,2}^{s,t'}\|_2^2 \,\mathrm{d}s\,. \tag{137}$$

Plugging Equations (130), (134) and (137) into Equation (128), we have

$$\mathbb{E}\|\rho_{\ell,1}^{t,t'} - \rho_{\ell,2}^{t,t'}\|_2^2 \le K\mathrm{e}^{2\lambda t} \cdot \mathsf{dist}_\lambda(X^1, X^2)^2 + K\int_{t'}^t \mathbb{E}\|\rho_{\ell,1}^{s,t'} - \rho_{\ell,2}^{s,t'}\|_2^2 \,\mathrm{d}s\,. \tag{138}$$

Applying Grönwall's inequality, we have

$$\mathrm{e}^{-2\lambda t}\,\mathbb{E}\|\rho_{\ell,1}^{t,t'} - \rho_{\ell,2}^{t,t'}\|_2^2 \le K \cdot \mathsf{dist}_\lambda(X^1, X^2)^2\,. \tag{139}$$

Next, we bound $\mathbb{E}\|D_{\ell,1}^{t,t'} - D_{\ell,2}^{t,t'}\|_2^2$. We have

$$\begin{aligned}
&\|D_{\ell,1}^{t,t'} - D_{\ell,2}^{t,t'}\|_2 \\
&\le \frac{1}{\delta}\|\nabla_r \ell_t(r_1^t; z) - \nabla_r \ell_t(r_2^t; z)\|_2 \left\| \int_{t'}^t R_\theta^1(t,s) D_{\ell,1}^{s,t'} (\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) \right\|_2 \\
&\quad + \frac{1}{\delta}\|\nabla_r \ell_t(r_2^t; z)\|_2 \left\| \int_{t'}^t (R_\theta^1(t,s) D_{\ell,1}^{s,t'} - R_\theta^2(t,s) D_{\ell,2}^{s,t'}) (\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) \right\|_2 \\
&\quad + \|\nabla_r^2 \ell_t(r_1^t; z) - \nabla_r^2 \ell_t(r_2^t; z)\|_2 \|D_{r,1}^{t,t'} \rho_{r,1}^{t,t'}\|_2 + \|\nabla_r^2 \ell_t(r_2^t; z)\|_2 \|D_{r,1}^{t,t'} \rho_{r,1}^{t,t'} - D_{r,2}^{t,t'} \rho_{r,2}^{t,t'}\|_2 \\
&\le \frac{M}{\delta}\|r_1^t - r_2^t\|_2 \left\| \int_{t'}^t R_\theta^1(t,s) D_{\ell,1}^{s,t'} (\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) \right\|_2 \\
&\quad + \frac{M}{\delta} \left\| \int_{t'}^t (R_\theta^1(t,s) D_{\ell,1}^{s,t'} - R_\theta^2(t,s) D_{\ell,2}^{s,t'}) (\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) \right\|_2 \\
&\quad + M\|r_1^t - r_2^t\|_2 \|D_{r,1}^{t,t'}\|_2 \|\rho_{r,1}^{t,t'}\|_2 + M\|D_{r,1}^{t,t'} - D_{r,2}^{t,t'}\|_2 \|\rho_{r,1}^{t,t'}\|_2 + M\|D_{r,2}^{t,t'}\|_2 \|\rho_{r,1}^{t,t'} - \rho_{r,2}^{t,t'}\|_2\,.
\end{aligned} \tag{140}$$

Squaring and taking the expectation, we have

$$\begin{aligned}
&\mathbb{E}\|D_{\ell,1}^{t,t'} - D_{\ell,2}^{t,t'}\|_2^2 \\
&\le K\,\mathbb{E}\|r_1^t - r_2^t\|_2^2 \left\| \int_{t'}^t R_\theta^1(t,s) D_{\ell,1}^{s,t'} (\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) \right\|_2^2 \\
&\quad + K\,\mathbb{E} \left\| \int_{t'}^t (R_\theta^1(t,s) D_{\ell,1}^{s,t'} - R_\theta^2(t,s) D_{\ell,2}^{s,t'}) (\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) \right\|_2^2 \\
&\quad + K\,\mathbb{E}\|r_1^t - r_2^t\|_2^2 \|D_{r,1}^{t,t'}\|_2^2 \|\rho_{r,1}^{t,t'}\|_2^2 + K\,\mathbb{E}\|D_{r,1}^{t,t'} - D_{r,2}^{t,t'}\|_2^2 \|\rho_{r,1}^{t,t'}\|_2^2 + K\,\mathbb{E}\|D_{r,2}^{t,t'}\|_2^2 \|\rho_{r,1}^{t,t'} - \rho_{r,2}^{t,t'}\|_2^2 \\
&\le K\sqrt{\mathbb{E}\|r_1^t - r_2^t\|_2^4} \sqrt{\mathbb{E}\left\| \int_{t'}^t R_\theta^1(t,s) D_{\ell,1}^{s,t'} (\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) \right\|_2^4} \\
&\quad + K\int_{t'}^t \mathbb{E}\|R_\theta^1(t,s) D_{\ell,1}^{s,t'} - R_\theta^2(t,s) D_{\ell,2}^{s,t'}\|_2^2 \,\mathrm{d}s \\
&\quad + K\sqrt{\mathbb{E}\|r_1^t - r_2^t\|_2^4} \sqrt{\mathbb{E}\|D_{r,1}^{t,t'}\|_2^8} \sqrt{\mathbb{E}\|\rho_{r,1}^{t,t'}\|_2^8} + K\sqrt{\mathbb{E}\|D_{r,1}^{t,t'} - D_{r,2}^{t,t'}\|_2^4} \sqrt{\mathbb{E}\|\rho_{r,1}^{t,t'}\|_2^4} \\
&\quad + K\sqrt{\mathbb{E}\|D_{r,2}^{t,t'}\|_2^4} \sqrt{\mathbb{E}\|\rho_{r,1}^{t,t'} - \rho_{r,2}^{t,t'}\|_2^4} \\
&\le K\mathrm{e}^{2\lambda t} \cdot \mathsf{dist}_\lambda(X^1, X^2)^2 \cdot \left(\sqrt{\sup_{s\in[t',t]} \mathbb{E}\|D_{\ell,1}^{s,t'}\|_2^2} + \sqrt{\mathbb{E}\|D_{r,1}^{t,t'}\|_2^8} \sqrt{\mathbb{E}\|\rho_{r,1}^{t,t'}\|_2^8} \right)
\end{aligned}$$

$$+ K \int_{t'}^{t} \Big(\|R_\theta^1(t,s) - R_\theta^2(t,s)\|_2^2 \, \mathbb{E}\|D_{\ell,1}^{s,t'}\|_2^2 + \|R_\theta^2(t,s)\|_2^2 \, \mathbb{E}\|D_{\ell,1}^{s,t'} - D_{\ell,2}^{s,t'}\|_2^2 \Big) \, \mathrm{d}s$$
$$+ K \sqrt{\mathbb{E}\|D_{r,1}^{t,t'} - D_{r,2}^{t,t'}\|_2^4} \sqrt{\mathbb{E}\|\rho_{r,1}^{t,t'}\|_2^4} + K \sqrt{\mathbb{E}\|D_{r,2}^{t,t'}\|_2^4} \sqrt{\mathbb{E}\|\rho_{r,1}^{t,t'} - \rho_{r,2}^{t,t'}\|_2^4} \,. \tag{141}$$

We do not repeat all the details, but by a similar argument as the bounds on $\mathbb{E}\|\rho_{r,1}^{t,t'}\|_2^4$, we can show that $\mathbb{E}\|\rho_{r,i}^{t,t'}\|_2^p$, $\mathbb{E}\|D_{\ell,i}^{t,t'}\|_2^p$, $\mathbb{E}\|D_{r,i}^{t,t'}\|_2^p$, and $\mathbb{E}\|D_{\ell,i}^{t,t'}\|_2^4$ are uniformly bounded by some constant $K > 0$ for $p \geq 2$ and $i = 1, 2$. Moreover, by a similar argument as the bound on $\mathbb{E}\|\rho_{r,1}^{t,t'} - \rho_{r,2}^{t,t'}\|_2^2$, we can show that $\mathbb{E}\|D_{r,1}^{t,t'} - D_{r,2}^{t,t'}\|_2^4$ and $\mathbb{E}\|\rho_{r,1}^{t,t'} - \rho_{r,2}^{t,t'}\|_2^4$ are bounded by $K\mathrm{e}^{4\lambda t} \cdot \mathsf{dist}_\lambda(X^1, X^2)^4$. Thus, using these bounds and Grönwall's inequality, we have

$$\mathrm{e}^{-2\lambda t} \, \mathbb{E}\|D_{\ell,1}^{t,t'} - D_{\ell,2}^{t,t'}\|_2^2 \leq K \cdot \mathsf{dist}_\lambda(X^1, X^2)^2 \,. \tag{142}$$

Therefore, we have

$$\mathsf{dist}_\lambda(R_\ell^1, R_\ell^2) \leq K \cdot \mathsf{dist}_\lambda(X^1, X^2) \,. \tag{143}$$

Collecting the above bounds, we have

$$\mathsf{dist}_\lambda(Y^1, Y^2) = \sqrt{\lambda}\mathsf{dist}_\lambda(\Sigma_\ell^1, \Sigma_\ell^2) + \mathsf{dist}_\lambda(R_\ell^1, R_\ell^2) + \mathsf{dist}_\lambda(\Gamma^1, \Gamma^2) \leq K \cdot \mathsf{dist}_\lambda(X^1, X^2) \,. \tag{144}$$

C.5. Proof of Lemma C.7

Again, K denotes a positive constant that may depend on M, m, T, δ, τ, and Φ, but not on other variables, and may change from line to line.

Bound of $\mathsf{dist}_\lambda(C_\theta^1, C_\theta^2)$. Let $U_1 \sim \mathsf{GP}(0, \Sigma_\ell^1/\delta)$ and $U_2 \sim \mathsf{GP}(0, \Sigma_\ell^2/\delta)$ be Gaussian processes coupled such that

$$\sup_{t \in [0,T]} \mathrm{e}^{-\lambda t} \sqrt{\mathbb{E}\|U_1^t - U_2^t\|_2^2} \leq 2 \cdot \mathsf{dist}_\lambda(\Sigma_\ell^1, \Sigma_\ell^2) \,. \tag{145}$$

For $i = 1, 2$, let θ_i be the solution of

$$\theta_i^t = \theta^0 + U_i^t - \int_0^t \Big(h_s(\theta_i^s) + \Gamma^i(s)\theta_i^s + \int_0^s R_\ell^i(s,s')\theta_i^{s'} \, \mathrm{d}s' \Big) \, \mathrm{d}s \,. \tag{146}$$

Then, we have

$$\begin{aligned}
&\mathrm{e}^{-2\lambda t} \, \mathbb{E}\|\theta_1^t - \theta_2^t\|_2^2 \\
&\leq 6\mathrm{e}^{-2\lambda t} \Bigg(T \int_0^t \mathbb{E}\|h_s(\theta_1^s) - h_s(\theta_2^s)\|_2^2 \, \mathrm{d}s + T \int_0^t \|\Gamma^1(s)\|_2^2 \, \mathbb{E}\|\theta_1^s - \theta_2^s\|_2^2 \, \mathrm{d}s \\
&\quad + T^2 \int_0^t \int_0^s \|R_\ell^1(s,s')\|_2^2 \, \mathbb{E}\|\theta_1^{s'} - \theta_2^{s'}\|_2^2 \, \mathrm{d}s' \, \mathrm{d}s + \mathbb{E}\|U_1^t - U_2^t\|_2^2 \\
&\quad + T \int_0^t \|\Gamma^1(s) - \Gamma^2(s)\|_2^2 \, \mathbb{E}\|\theta_2^s\|_2^2 \, \mathrm{d}s + T^2 \int_0^t \int_0^s \|R_\ell^1(s,s') - R_\ell^2(s,s')\|_2^2 \, \mathbb{E}\|\theta_2^{s'}\|_2^2 \, \mathrm{d}s' \, \mathrm{d}s \Bigg)
\end{aligned}$$

$$
\begin{aligned}
&\leq 6\Bigg(2M^2T\int_0^t \mathrm{e}^{-2\lambda(t-s)}\cdot \mathrm{e}^{-2\lambda s}\,\mathbb{E}\|\theta_1^s-\theta_2^s\|_2^2\,\mathrm{d}s \\
&\quad + T^2\Phi^2\int_0^t \mathrm{e}^{-2\lambda(t-s)}\int_0^s \mathrm{e}^{-2\lambda s}\,\mathbb{E}\|\theta_1^{s'}-\theta_2^{s'}\|_2^2\,\mathrm{d}s'\,\mathrm{d}s + 4\cdot \mathsf{dist}_\lambda(\Sigma_\ell^1,\Sigma_\ell^2)^2 \\
&\quad + Tm\Phi\int_0^t \mathrm{e}^{-2\lambda(t-s)}\cdot \mathrm{e}^{-2\lambda s}\|\Gamma^1(s)-\Gamma^2(s)\|_2^2 \\
&\quad + T^2m\Phi\int_0^t \mathrm{e}^{-2\lambda(t-s)}\int_0^s \mathrm{e}^{-2\lambda s}\|R_\ell^1(s,s')-R_\ell^2(s,s')\|_2^2\,\mathrm{d}s'\,\mathrm{d}s\Bigg) \\
&\leq 6\Bigg(\frac{M^2T}{\lambda}\sup_{s\in[0,t]}\mathrm{e}^{-2\lambda s}\,\mathbb{E}\|\theta_1^s-\theta_2^s\|_2^2 + \frac{T^3\Phi^2}{2\lambda}\sup_{s\in[0,t]}\mathrm{e}^{-2\lambda s}\,\mathbb{E}\|\theta_1^s-\theta_2^s\|_2^2 \\
&\quad + \frac{4}{\lambda}\cdot\lambda\mathsf{dist}_\lambda(\Sigma_\ell^1,\Sigma_\ell^2)^2 + \frac{Tm\Phi}{2\lambda}\cdot\mathsf{dist}_\lambda(\Gamma^1,\Gamma^2)^2 + \frac{T^3m\Phi}{2\lambda}\cdot\mathsf{dist}_\lambda(R_\ell^1,R_\ell^2)^2\Bigg) \\
&\leq \frac{K}{\lambda}\sup_{s\in[0,t]}\mathrm{e}^{-2\lambda s}\,\mathbb{E}\|\theta_1^s-\theta_2^s\|_2^2 + \frac{K}{\lambda}\cdot\mathsf{dist}_\lambda(Y^1,Y^2)^2\,. \qquad (147)
\end{aligned}
$$

Since K is independent of λ, we can take $\lambda > 2K$ so that

$$
\sup_{t\in[0,T]}\mathrm{e}^{-2\lambda t}\,\mathbb{E}\|\theta_1^t-\theta_2^t\|_2^2 \leq \frac{2K}{\lambda}\cdot\mathsf{dist}_\lambda(Y^1,Y^2)^2\,. \qquad (148)
$$

Thus, for any $\varepsilon > 0$, we can take λ large enough such that

$$
\sup_{t\in[0,T]}\mathrm{e}^{-2\lambda t}\,\mathbb{E}\|\theta_1^t-\theta_2^t\|_2^2 \leq \varepsilon^2\cdot\mathsf{dist}_\lambda(Y^1,Y^2)^2\,. \qquad (149)
$$

Let $\{(w_1^t,w_2^t)\}_{t\in[0,T]}$ be a centered Gaussian process with covariance $\mathbb{E}\left[\begin{pmatrix}\theta_1^t\\ \theta_2^t\end{pmatrix}\begin{pmatrix}\theta_1^t\\ \theta_2^t\end{pmatrix}^\mathsf{T}\right]$. Since w_1 and w_2 have covariance kernels C_θ^1 and C_θ^2 respectively, we have

$$
\mathsf{dist}_\lambda(C_\theta^1,C_\theta^2) \leq \sup_{t\in[0,T]}\mathrm{e}^{-\lambda t}\sqrt{\mathbb{E}\|w_1^t-w_2^t\|_2^2} = \sup_{t\in[0,T]}\mathrm{e}^{-\lambda t}\sqrt{\mathbb{E}\|\theta_1^t-\theta_2^t\|_2^2} \leq \varepsilon\cdot\mathsf{dist}_\lambda(Y^1,Y^2)\,. \qquad (150)
$$

Bound of $\mathsf{dist}_\lambda(R_\theta^1,R_\theta^2)$**.** For $i=1,2$, let $\rho_{\theta,i}$ be the solution of

$$
\rho_{\theta,i}^{t,t'} = R_\theta^i(t,t') - \int_{t'}^t\big(\nabla_\theta h_s(\theta_i^s)+\Gamma^i(s)\big)\rho_{\theta,i}^{s,t'} + \int_{t'}^s R_\ell^i(s,s')\rho_{\theta,i}^{s',t'}\,\mathrm{d}s')\,\mathrm{d}s\,. \qquad (151)
$$

Thus, we have

$$
\begin{aligned}
&\mathrm{e}^{-\lambda t}\|\rho_{\theta,1}^{t,t'}-\rho_{\theta,2}^{t,t'}\|_2 \\
&\leq \mathrm{e}^{-\lambda t}\int_0^t(\|\nabla_\theta h_s(\theta_1^s)\|_2+\|\Gamma^1(s)\|_2)\|\rho_{\theta,1}^{s,t'}-\rho_{\theta,2}^{s,t'}\|_2\,\mathrm{d}s
\end{aligned}
$$

$$
\begin{aligned}
&+ \mathrm{e}^{-\lambda t} \int_0^t \big(\|\nabla_\theta h_s(\theta_1^s) - \nabla_\theta h_s(\theta_2^2)\|_2 + \|\Gamma^1(s) - \Gamma^2(s)\|_2\big) \|\rho_{\theta,2}^{s,t'}\| \,\mathrm{d}s \\
&+ \mathrm{e}^{-\lambda t} \int_0^t \int_{t'}^s \Big(\|R_\ell^1(s,s')\|_2 \|\rho_{\theta,1}^{s',t'} - \rho_{\theta,2}^{s',t'}\|_2 + \|R_\ell^1(s,s') - R_\ell^2(s,s')\|_2 \|\rho_{\theta,2}^{s',t'}\|_2\Big) \,\mathrm{d}s' \,\mathrm{d}s \\
\le{}& \int_0^t \mathrm{e}^{-\lambda(t-s)} \Big(2M\mathrm{e}^{-\lambda s} \|\rho_{\theta,1}^{s,t'} - \rho_{\theta,2}^{s,t'}\|_2 + \Phi \int_{t'}^s \mathrm{e}^{-\lambda s'} \|\rho_{\theta,1}^{s',t'} - \rho_{\theta,2}^{s',t'}\|_2 \,\mathrm{d}s'\Big) \,\mathrm{d}s \\
&+ \int_0^t \mathrm{e}^{-\lambda(t-s)} \Big(M\mathrm{e}^{-\lambda s} \|\theta_1^s - \theta_2^2\|_2 \|\rho_{\theta,2}^{s,t'}\|_2 + \mathrm{e}^{-\lambda s} \|\Gamma^1(s) - \Gamma^2(s)\|_2 \|\rho_{\theta,2}^{s',t'}\|_2\Big) \,\mathrm{d}s \\
&+ \int_0^t \mathrm{e}^{-\lambda(t-s)} \int_{t'}^s \mathrm{e}^{-\lambda s} \|R_\ell^1(s,s') - R_\ell^2(s,s')\|_2 \|\rho_{\theta,2}^{s',t'}\|_2 \,\mathrm{d}s' \,\mathrm{d}s\,. \qquad (152)
\end{aligned}
$$

By Equation (110), we have $\|\rho_{\theta,2}^{s',t'}\|_2 \le \Phi$. Thus, we have

$$
\begin{aligned}
&\mathrm{e}^{-\lambda t}\, \mathbb{E}\|\rho_{\theta,1}^{t,t'} - \rho_{\theta,2}^{t,t'}\|_2 \\
&\le \int_0^t \mathrm{e}^{-\lambda(t-s)} \Big(2M\mathrm{e}^{-\lambda s} \|\rho_{\theta,1}^{s,t'} - \rho_{\theta,2}^{s,t'}\|_2 + \Phi \int_{t'}^s \mathrm{e}^{-\lambda s'} \|\rho_{\theta,1}^{s',t'} - \rho_{\theta,2}^{s',t'}\|_2 \,\mathrm{d}s'\Big) \,\mathrm{d}s \\
&\quad + \Phi \int_0^t \mathrm{e}^{-\lambda(t-s)} \Big(M\mathrm{e}^{-\lambda s}\, \mathbb{E}\|\theta_1^s - \theta_2^2\|_2 + \mathrm{e}^{-\lambda s} \|\Gamma^1(s) - \Gamma^2(s)\|_2\Big) \,\mathrm{d}s \\
&\quad + \Phi \int_0^t \mathrm{e}^{-\lambda(t-s)} \int_{t'}^s \mathrm{e}^{-\lambda s} \|R_\ell^1(s,s') - R_\ell^2(s,s')\|_2 \,\mathrm{d}s' \,\mathrm{d}s \\
&\le \Big(\int_0^t \mathrm{e}^{-\lambda(t-s)} \,\mathrm{d}s\Big) \Bigg((2M + \Phi T) \sup_{0\le t'\le s\le t} \mathrm{e}^{-\lambda s} \|\rho_{\theta,1}^{s,t'} - \rho_{\theta,2}^{s,t'}\|_2 + \Phi M \sup_{0\le s\le t} \mathrm{e}^{-\lambda s} \sqrt{\mathbb{E}\|\theta_1^s - \theta_2^s\|_2} \\
&\quad + \Phi \sup_{0\le s\le t} \mathrm{e}^{-\lambda s} \|\Gamma^1(s) - \Gamma^2(s)\|_2 + \Phi T \sup_{0\le t'\le s'\le s\le t} \mathrm{e}^{-\lambda s} \|R_\ell^1(s,s') - R_\ell^2(s,s')\|_2\Bigg) \\
&\le \frac{K}{\lambda} \sup_{0\le t'\le s\le t} \mathrm{e}^{-\lambda s} \|\rho_{\theta,1}^{s,t'} - \rho_{\theta,2}^{s,t'}\|_2 + \frac{K}{\lambda} \cdot \mathsf{dist}_\lambda(Y^1, Y^2)\,. \qquad (153)
\end{aligned}
$$

Taking λ large enough such that $\lambda > 2K$, we have

$$
\sup_{0\le s\le t\le T} \mathrm{e}^{-\lambda t} \|R_\theta^1(t,s) - R_\theta^2(t,s)\|_2 \le \sup_{0\le s\le t\le T} \mathrm{e}^{-\lambda t}\, \mathbb{E}\|\rho_{\theta,1}^{t,s} - \rho_{\theta,2}^{t,s}\|_2 \le \frac{2K}{\lambda} \cdot \mathsf{dist}_\lambda(Y^1, Y^2)\,. \qquad (154)
$$

Thus, for any $\varepsilon > 0$, we can take λ large enough such that

$$
\mathsf{dist}_\lambda(R_\theta^1, R_\theta^2) \le \varepsilon \cdot \mathsf{dist}_\lambda(Y^1, Y^2)\,. \qquad (155)
$$

Collecting the above bounds, we have that for any $\varepsilon > 0$, we can take λ large enough such that

$$
\mathsf{dist}_\lambda(X^1, X^2) = \mathsf{dist}_\lambda(C_\theta^1, C_\theta^2) + \mathsf{dist}_\lambda(R_\theta^1, R_\theta^2) \le \varepsilon \cdot \mathsf{dist}_\lambda(Y^1, Y^2)\,. \qquad (156)
$$

Appendix D. Proof of Theorem 3.2

We follow the three-step strategy outlined in the main text, which we repeat here for clarity.

1. We discretize the SGF (1) with time step $\gamma > 0$ and analyze the discretization error (Lemma D.1).
2. We apply the AMP theory to characterize the asymptotic behavior of the discretized SGF using a low-dimensional state evolution recursion (Lemma D.2).
3. We take the continuous-time limit $\gamma \to 0$ and show that the state evolution converges to the unique solution of the DMFT equation $\mathfrak{S}$ (Lemma D.3).

We consider the discretization of the SGF (1) with time step $\gamma > 0$. Let $\lfloor t \rfloor := \max\{k\gamma : k\gamma \leq t, k \in \mathbb{N}\}$. We define $\boldsymbol{\theta}_\gamma^t$ and $\boldsymbol{r}_\gamma^t$ as the solution of the following equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\theta}_\gamma^t = -\left(h_{\lfloor t \rfloor}(\boldsymbol{\theta}_\gamma^{\lfloor t \rfloor}) + \frac{1}{\delta}\boldsymbol{X}^\mathsf{T}\ell_{\lfloor t \rfloor}(\boldsymbol{r}_\gamma^{\lfloor t \rfloor}; \boldsymbol{z})\right)\mathrm{d}t + \sqrt{\frac{\tau}{\delta}}\sum_{i=1}^n \boldsymbol{x}_i \ell_{\lfloor t \rfloor}(r_{\gamma,i}^{\lfloor t \rfloor}; z_i)^\mathsf{T}\,\mathrm{d}B_i^t\,, \tag{157}$$

and $\boldsymbol{r}_\gamma^t = \boldsymbol{X}\boldsymbol{\theta}_\gamma^t$ with initial condition $\boldsymbol{\theta}_\gamma^0 = \boldsymbol{\theta}^0$. At discrete time points $t_1, t_2, \dots$ where we define $t_k = k\gamma$ for $k \in \mathbb{N}$, $\boldsymbol{\theta}_\gamma^{t_k}$ satisfies the following recursion:

$$\boldsymbol{\theta}_\gamma^{t_{k+1}} = \boldsymbol{\theta}_\gamma^{t_k} - \gamma\left(h_{t_k}(\boldsymbol{\theta}_\gamma^{t_k}) + \frac{1}{\delta}\boldsymbol{X}^\mathsf{T}\ell_{t_k}(\boldsymbol{r}_\gamma^{t_k}; \boldsymbol{z})\right) + \sqrt{\frac{\tau}{\delta}}\sum_{i=1}^n \boldsymbol{x}_i \ell_{t_k}(r_{\gamma,i}^{t_k}; z_i)^\mathsf{T}(B_i^{t_{k+1}} - B_i^{t_k})\,. \tag{158}$$

First, we control the discretization error between the SGF (1) and its time-discretized version (158). We prove it in Appendix D.1.

Lemma D.1 *Under the assumptions of Theorem 3.2, for any $T > 0$, there exists a constant $C > 0$ such that we have, almost surely over the randomness of $\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{\theta}^0$,*

$$\limsup_{n,d\to\infty} \frac{1}{d}\mathbb{E}_{\boldsymbol{B}}\left[\sup_{0\leq t\leq T}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_\gamma^t\|_\mathrm{F}^2\right] \leq C\gamma\,, \quad \limsup_{n,d\to\infty} \frac{1}{n}\mathbb{E}_{\boldsymbol{B}}\left[\sup_{0\leq t\leq T}\|\boldsymbol{r}^t - \boldsymbol{r}_\gamma^t\|_\mathrm{F}^2\right] \leq C\gamma\,. \tag{159}$$

Furthermore, for any $L \in \mathbb{N}$ and $0 \leq t_1, \dots, t_L \leq T$, we have, almost surely,

$$\lim_{\gamma\to 0}\limsup_{n,d\to\infty}\mathbb{E}_{\boldsymbol{B}}\, W_2\Big(\hat{\mathsf{P}}(\boldsymbol{\theta}^{t_1}, \dots, \boldsymbol{\theta}^{t_L}), \hat{\mathsf{P}}(\boldsymbol{\theta}_\gamma^{t_1}, \dots, \boldsymbol{\theta}_\gamma^{t_L})\Big)^2 = 0\,, \tag{160}$$

$$\lim_{\gamma\to 0}\limsup_{n,d\to\infty}\mathbb{E}_{\boldsymbol{B}}\, W_2\Big(\hat{\mathsf{P}}(\boldsymbol{r}^{t_1}, \dots, \boldsymbol{r}^{t_L}, \boldsymbol{z}), \hat{\mathsf{P}}(\boldsymbol{r}_\gamma^{t_1}, \dots, \boldsymbol{r}_\gamma^{t_L}, \boldsymbol{z})\Big)^2 = 0\,. \tag{161}$$

Next, we relate the discretized SGF to the discretized DMFT equation $\mathfrak{S}^\gamma$ defined in Appendix B.2. The following lemma shows that the unique solution of $\mathfrak{S}^\gamma$ characterizes the asymptotic behavior of the discretized SGF (158). We prove it in Appendix D.2.

Lemma D.2 *Under the assumptions of Theorem 3.2, for any $T > 0, L \in \mathbb{N}$ and $0 \leq t_1 < \dots < t_L \leq T$, we have*

$$\operatorname*{p-lim}_{n,d\to\infty} W_2\Big(\hat{\mathsf{P}}(\boldsymbol{\theta}_\gamma^{t_1}, \dots, \boldsymbol{\theta}_\gamma^{t_L}), \mathsf{P}(\theta_\gamma^{t_1}, \dots, \theta_\gamma^{t_L})\Big) = 0\,, \tag{162}$$

$$\operatorname*{p-lim}_{n,d\to\infty} W_2\Big(\hat{\mathsf{P}}(\boldsymbol{r}_\gamma^{t_1}, \dots, \boldsymbol{r}_\gamma^{t_L}, \boldsymbol{z}), \mathsf{P}(r_\gamma^{t_1}, \dots, r_\gamma^{t_L}, z)\Big) = 0\,. \tag{163}$$

Finally, we establish the convergence of the discretized DMFT equation $\mathfrak{S}^\gamma$ to the original DMFT equation $\mathfrak{S}$ as $\gamma \to 0$. We prove it in Appendix D.3.

Lemma D.3 *Under the assumptions of Theorem 3.2, for any $T > 0, L \in \mathbb{N}$ and $t_1 < \cdots < t_L \in [0, T]$, we have*

$$\lim_{\gamma\to 0} W_2\big(\mathsf{P}(\theta_\gamma^{t_1}, \dots, \theta_\gamma^{t_L}), \mathsf{P}(\theta^{t_1}, \dots, \theta^{t_L})\big) = 0\,, \tag{164}$$

$$\lim_{\gamma\to 0} W_2\big(\mathsf{P}(r_\gamma^{t_1}, \dots, r_\gamma^{t_L}, z), \mathsf{P}(r^{t_1}, \dots, r^{t_L}, z)\big) = 0\,. \tag{165}$$

We are now ready to prove Theorem 3.2.

Proof [Proof of Theorem 3.2] We prove for $\boldsymbol{\theta}^t$; the proof for $\boldsymbol{r}^t$ is similar.

For any $0 \le t_1 < \cdots < t_L \le T$ and $\gamma > 0$, by the triangle inequality, we have

$$E_d := W_2\Big(\hat{\mathsf{P}}(\boldsymbol{\theta}^{t_1}, \dots, \boldsymbol{\theta}^{t_L}), \mathsf{P}(\theta^{t_1}, \dots, \theta^{t_L})\Big) \le E_{d,\gamma}^{(1)} + E_{d,\gamma}^{(2)} + E_\gamma^{(3)}\,, \tag{166}$$

where we defined

$$E_{d,\gamma}^{(1)} := W_2\Big(\hat{\mathsf{P}}(\boldsymbol{\theta}^{t_1}, \dots, \boldsymbol{\theta}^{t_L}), \hat{\mathsf{P}}(\boldsymbol{\theta}_\gamma^{t_1}, \dots, \boldsymbol{\theta}_\gamma^{t_L})\Big)\,, \tag{167}$$

$$E_{d,\gamma}^{(2)} := W_2\Big(\hat{\mathsf{P}}(\boldsymbol{\theta}_\gamma^{t_1}, \dots, \boldsymbol{\theta}_\gamma^{t_L}), \mathsf{P}(\theta_\gamma^{t_1}, \dots, \theta_\gamma^{t_L})\Big)\,, \tag{168}$$

$$E_\gamma^{(3)} := W_2\big(\mathsf{P}(\theta_\gamma^{t_1}, \dots, \theta_\gamma^{t_L}), \mathsf{P}(\theta^{t_1}, \dots, \theta^{t_L})\big)\,. \tag{169}$$

By the union bound, we have, for any $\varepsilon > 0$,

$$\mathbb{P}\{E_d \ge \varepsilon\} \le \mathbb{P}\{E_{d,\gamma}^{(1)} \ge \varepsilon/3\} + \mathbb{P}\{E_{d,\gamma}^{(2)} \ge \varepsilon/3\} + \mathbb{P}\{E_\gamma^{(3)} \ge \varepsilon/3\}\,. \tag{170}$$

Taking the limit $n, d \to \infty$ and applying Lemma D.2, the second term vanishes. Furthermore, by Markov's inequality, we have

$$\limsup_{n,d\to\infty} \mathbb{P}\{E_d \ge \varepsilon\} \le \frac{9}{\varepsilon^2} \limsup_{n,d\to\infty} \mathbb{E}[(E_{d,\gamma}^{(1)})^2] + \mathbb{P}\{E_\gamma^{(3)} \ge \varepsilon/3\}\,. \tag{171}$$

Since the left-hand side does not depend on γ, we can take the limit $\gamma \to 0$ and apply Lemmas D.1 and D.3 to obtain

$$\lim_{n,d\to\infty} \mathbb{P}\{E_d \ge \varepsilon\} = 0\,, \tag{172}$$

and thus $E_d \to 0$ in probability as $n, d \to \infty$. ■

D.1. Proof of Lemma D.1

In the following, C denotes a constant independent of n, d, γ, which may change from line to line.

We utilize the general results in Appendix G. We check that the SDE (1) satisfies Theorem G.1.

Drift term. For the drift coefficient $\boldsymbol{b}(t,\boldsymbol{\theta}) = h_t(\boldsymbol{\theta}) + \frac{1}{\delta}\boldsymbol{X}^\mathsf{T}\ell_t(\boldsymbol{X}\boldsymbol{\theta};\boldsymbol{z})$, we use the Lipschitz continuity of h and ℓ and that $\|\boldsymbol{X}\|_2 \le C$ and $\|\ell_t(0;\boldsymbol{z})\|_\mathrm{F} \le Cn$ almost surely for sufficiently large n, d by assumption to obtain

$$\begin{aligned}
\|\boldsymbol{b}(t,\theta)\|_\mathrm{F}^2 &\le 2\|h_t(\boldsymbol{\theta})\|_\mathrm{F}^2 + \frac{2}{\delta^2}\|\boldsymbol{X}\|_2^2\|\ell_t(\boldsymbol{X}\boldsymbol{\theta};\boldsymbol{z})\|_\mathrm{F}^2 \\
&\le C(|h_t(0)|d + \|\boldsymbol{\theta}\|_\mathrm{F}^2) + C(\|\ell_t(0;\boldsymbol{z})\|_\mathrm{F}^2 + \|\boldsymbol{X}\boldsymbol{\theta}\|_\mathrm{F}^2) \\
&\le C(d + \|\boldsymbol{\theta}\|_\mathrm{F}^2)\,, \qquad (173)\\
\|\boldsymbol{b}(t_1,\boldsymbol{\theta}_1) - \boldsymbol{b}(t_2,\boldsymbol{\theta}_2)\|_\mathrm{F}^2 &\le 2\|h_{t_1}(\boldsymbol{\theta}_1) - h_{t_2}(\boldsymbol{\theta}_2)\|_\mathrm{F}^2 + \frac{2}{\delta^2}\|\boldsymbol{X}\|^2\|\ell_{t_1}(\boldsymbol{X}\boldsymbol{\theta}_1;\boldsymbol{z}) - \ell_{t_2}(\boldsymbol{X}\boldsymbol{\theta}_2;\boldsymbol{z})\|_\mathrm{F}^2 \\
&\le C(d|t_1-t_2|^2 + \|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_\mathrm{F}^2) + C(n|t_1-t_2|^2 + \|\boldsymbol{X}(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)\|_\mathrm{F}^2) \\
&\le C(d|t_1-t_2|^2 + \|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_\mathrm{F}^2)\,. \qquad (174)
\end{aligned}$$

Diffusion term. For the diffusion coefficient $\boldsymbol{\sigma}_i(t,\boldsymbol{\theta}) = \sqrt{\tau/\delta}\boldsymbol{x}_i\ell_t(\boldsymbol{\theta}^\mathsf{T}\boldsymbol{x}_i;z_i)^\mathsf{T}$, we proceed similarly as above to obtain

$$\begin{aligned}
\sum_{i=1}^n\|\boldsymbol{\sigma}_i(t,\boldsymbol{\theta})\|_\mathrm{F}^2 &= \frac{\tau}{\delta}\sum_{i=1}^n\|\boldsymbol{x}_i\|_2^2\|\ell_t(\boldsymbol{\theta}^\mathsf{T}\boldsymbol{x}_i;z_i)\|_2^2 \le C(n + \|\boldsymbol{X}\boldsymbol{\theta}\|_\mathrm{F}^2) \\
&\le C(d + \|\boldsymbol{\theta}\|_\mathrm{F}^2)\,, \qquad (175)\\
\sum_{i=1}^n\|\boldsymbol{\sigma}_i(t_1,\boldsymbol{\theta}_1) - \boldsymbol{\sigma}_i(t_2,\boldsymbol{\theta}_2)\|_\mathrm{F}^2 &= \frac{\tau}{\delta}\sum_{i=1}^n\|\boldsymbol{x}_i\|_2^2\|\ell_{t_1}(\boldsymbol{\theta}_1^\mathsf{T}\boldsymbol{x}_i;z_i) - \ell_{t_2}(\boldsymbol{\theta}_2^\mathsf{T}\boldsymbol{x}_i;z_i)\|_2^2 \\
&\le C(d|t_1-t_2|^2 + \|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_\mathrm{F}^2)\,. \qquad (176)
\end{aligned}$$

Therefore, by Lemma G.4, we have that, for sufficiently large d,

$$\mathbb{E}\left[\sup_{0\le t\le T}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_\gamma^t\|_\mathrm{F}^2\right] \le C\gamma(d + \|\boldsymbol{\theta}^0\|_\mathrm{F}^2)\,. \qquad (177)$$

By Assumption A.1, $\|\boldsymbol{\theta}^0\|_\mathrm{F}^2 < Cd$ holds almost surely for sufficiently large d. Therefore, there exists a constant C independent of n, d, γ such that

$$\limsup_{n,d\to\infty}\frac{1}{d}\,\mathbb{E}_{\boldsymbol{B}}\left[\sup_{0\le t\le T}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_\gamma^t\|_\mathrm{F}^2\right] \le C\gamma\,. \qquad (178)$$

This shows the first claim. Furthermore, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{B}}\,W_2\Big(\hat{\mathsf{P}}(\boldsymbol{\theta}^{t_1},\dots,\boldsymbol{\theta}^{t_L}),\hat{\mathsf{P}}(\boldsymbol{\theta}_\gamma^{t_1},\dots,\boldsymbol{\theta}_\gamma^{t_L})\Big)^2 &\le \mathbb{E}_{\boldsymbol{B}}\left[\frac{1}{d}\sum_{l=1}^L\|\boldsymbol{\theta}^{t_l} - \boldsymbol{\theta}_\gamma^{t_l}\|_\mathrm{F}^2\right] \\
&\le \frac{L}{d}\,\mathbb{E}_{\boldsymbol{B}}\left[\sup_{0\le t\le T}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_\gamma^t\|_\mathrm{F}^2\right]. \qquad (179)
\end{aligned}$$

Taking the limit $n, d \to \infty$ followed by $\gamma \to 0$ shows the second claim. The claim for $\boldsymbol{r}^t$ follows from

$$\|\boldsymbol{r}^t - \boldsymbol{r}_\gamma^t\|_\mathrm{F} = \|\boldsymbol{X}(\boldsymbol{\theta}^t - \boldsymbol{\theta}_\gamma^t)\|_\mathrm{F} \le \|\boldsymbol{X}\|_2\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_\gamma^t\|_\mathrm{F}\,. \qquad (180)$$

D.2. Proof of Lemma D.2

D.2.1. REDUCTION TO AMP

For notational simplicity, we omit the subscript γ and denote $\boldsymbol{\theta}^t_\gamma, \boldsymbol{r}^t_\gamma$ by $\boldsymbol{\theta}^t, \boldsymbol{r}^t$. As we only work in discrete time, there is no risk of confusion.

Let $\boldsymbol{G}^k = (\boldsymbol{B}^{t_{k+1}} - \boldsymbol{B}^{t_k})/\sqrt{\gamma} \sim \mathsf{N}(0, \boldsymbol{I}_n)$. Then the recursion (158) can be rewritten as

$$\boldsymbol{\theta}^{t_{k+1}} = \boldsymbol{\theta}^{t_k} - \gamma h_{t_k}(\boldsymbol{\theta}^{t_k}) - \frac{\gamma}{\delta}\boldsymbol{X}^\mathsf{T}((\mathbf{1}_n + \sqrt{\tau\delta/\gamma}\boldsymbol{G}^k) \odot \ell_{t_k}(\boldsymbol{r}^{t_k}; \boldsymbol{z}))\,, \quad \boldsymbol{r}^{t_k} = \boldsymbol{X}\boldsymbol{\theta}^{t_k}\,. \tag{181}$$

Let $M > 0$ be a constant and let $[\cdot]_M \colon x \mapsto \max\{-M, \min\{x, M\}\}$ be the clipping function. We clip the Gaussian vector in (181) entry-wise as

$$\check{\boldsymbol{\theta}}^{t_{k+1}} = \check{\boldsymbol{\theta}}^{t_k} - \gamma h_{t_k}(\check{\boldsymbol{\theta}}^{t_k}) - \frac{\gamma}{\delta}\boldsymbol{X}^\mathsf{T}((\mathbf{1}_n + \sqrt{\tau\delta/\gamma}[\boldsymbol{G}^k]_M) \odot \ell_{t_k}(\check{\boldsymbol{r}}^{t_k}; \boldsymbol{z}))\,, \quad \check{\boldsymbol{r}}^{t_k} = \boldsymbol{X}\check{\boldsymbol{\theta}}^{t_k}\,. \tag{182}$$

We first control the difference between (181) and (182).

Lemma D.4 *We have, almost surely over the randomness of $\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{\theta}^0$,*

$$\lim_{M\to\infty} \limsup_{n,d\to\infty} \frac{1}{d}\,\mathbb{E}_{\boldsymbol{B}}\left[\max_{0\le k\le T/\gamma} \|\boldsymbol{\theta}^{t_k} - \check{\boldsymbol{\theta}}^{t_k}\|_\mathrm{F}^2\right] = 0\,, \tag{183}$$

and thus

$$\lim_{M\to\infty} \limsup_{n,d\to\infty} \mathbb{E}_{\boldsymbol{B}}\, W_2\Big(\hat{\mathsf{P}}(\boldsymbol{\theta}^{t_1}, \dots, \boldsymbol{\theta}^{t_L}), \hat{\mathsf{P}}(\check{\boldsymbol{\theta}}^{t_1}, \dots, \check{\boldsymbol{\theta}}^{t_L})\Big)^2 = 0\,, \tag{184}$$

$$\lim_{M\to\infty} \limsup_{n,d\to\infty} \mathbb{E}_{\boldsymbol{B}}\, W_2\Big(\hat{\mathsf{P}}(\boldsymbol{r}^{t_1}, \dots, \boldsymbol{r}^{t_L}, \boldsymbol{z}), \hat{\mathsf{P}}(\check{\boldsymbol{r}}^{t_1}, \dots, \check{\boldsymbol{r}}^{t_L}, \boldsymbol{z})\Big)^2 = 0\,, \tag{185}$$

Proof We have

$$\begin{aligned}
&\|\boldsymbol{\theta}^{t_{k+1}} - \check{\boldsymbol{\theta}}^{t_{k+1}}\|_\mathrm{F} \\
&\le \|\boldsymbol{\theta}^{t_k} - \check{\boldsymbol{\theta}}^{t_k}\|_\mathrm{F} + \gamma\|h_{t_k}(\boldsymbol{\theta}^{t_k}) - h_{t_k}(\check{\boldsymbol{\theta}}^{t_k})\|_\mathrm{F} \\
&\qquad + \frac{\gamma}{\delta}\Big\|\boldsymbol{X}^\mathsf{T}\Big((\mathbf{1}_n + \sqrt{\tau\delta/\gamma}\boldsymbol{G}^k) \odot \ell_{t_k}(\boldsymbol{r}^{t_k}; \boldsymbol{z}) - (\mathbf{1}_n + \sqrt{\tau\delta/\gamma}[\boldsymbol{G}^k]_M) \odot \ell_{t_k}(\check{\boldsymbol{r}}^{t_k}; \boldsymbol{z})\Big)\Big\|_\mathrm{F} \\
&\le C\|\boldsymbol{\theta}^{t_k} - \check{\boldsymbol{\theta}}^{t_k}\|_\mathrm{F} + C\|(\boldsymbol{G}^k - [\boldsymbol{G}^k]_M) \odot \ell_{t_k}(\boldsymbol{r}^{t_k}; z)\|_\mathrm{F} + C\|[\boldsymbol{G}^k]_M \odot (\ell_{t_k}(\boldsymbol{r}^{t_k}; \boldsymbol{z}) - \ell_{t_k}(\check{\boldsymbol{r}}^{t_k}; \boldsymbol{z}))\|_\mathrm{F}\,.
\end{aligned} \tag{186}$$

Since $\boldsymbol{G}^k$ is independent of $\boldsymbol{r}^{t_k}$ and $\check{\boldsymbol{r}}^{t_k}$, we have

$$\begin{aligned}
&\mathbb{E}\|(\boldsymbol{G}^k - [\boldsymbol{G}^k]_M) \odot \ell_{t_k}(\boldsymbol{r}^{t_k}; z)\|_\mathrm{F}^2 = \sum_{i=1}^n \mathbb{E}[(G_i^k - [G_i^k]_M)^2\|\ell_{t_k}(r_i^{t_k}; z_i)\|_2^2] \\
&= f(M)\sum_{i=1}^n \mathbb{E}[\|\ell_{t_k}(\boldsymbol{r}_i^{t_k}; z_i)\|_2^2] = f(M)\,\mathbb{E}\|\ell_{t_k}(\boldsymbol{r}_i^{t_k}; \boldsymbol{z})\|_\mathrm{F}^2 \le Cf(M)\,\mathbb{E}\|\boldsymbol{\theta}^{t_k}\|_\mathrm{F}^2\,, \\
&\mathbb{E}\|[\boldsymbol{G}^k]_M \odot (\ell_{t_k}(\boldsymbol{r}^{t_k}; \boldsymbol{z}) - \ell_{t_k}(\check{\boldsymbol{r}}^{t_k}; \boldsymbol{z}))\|_\mathrm{F}^2 = \sum_{i=1}^n \mathbb{E}[[G_i^k]_M^2\|\ell_{t_k}(\boldsymbol{r}_i^{t_k}; z_i) - \ell_{t_k}(\check{\boldsymbol{r}}_i^{t_k}; z_i)\|_2^2]
\end{aligned} \tag{187}$$

$$= \mathbb{E}[[G]_M^2] \sum_{i=1}^{n} \mathbb{E}[\|\ell_{t_k}(\boldsymbol{r}_i^{t_k}; z_i) - \ell_{t_k}(\check{\boldsymbol{r}}_i^{t_k}; z_i)\|_2^2] \le C\, \mathbb{E}[G^2]\, \mathbb{E}\|\boldsymbol{r}^{t_k} - \check{\boldsymbol{r}}^{t_k}\|_{\mathrm{F}}^2 \le C\, \mathbb{E}\|\boldsymbol{\theta}^{t_k} - \check{\boldsymbol{\theta}}^{t_k}\|_{\mathrm{F}}^2 \,, \tag{188}$$

where $f(M) = \mathbb{E}[(G - [G]_M)^2]$ for $G \sim \mathsf{N}(0,1)$. Therefore, we have

$$\frac{1}{d}\, \mathbb{E}\|\boldsymbol{\theta}^{t_{k+1}} - \check{\boldsymbol{\theta}}^{t_{k+1}}\|_{\mathrm{F}}^2 \le \frac{C}{d}\, \mathbb{E}\|\boldsymbol{\theta}^{t_k} - \check{\boldsymbol{\theta}}^{t_k}\|_{\mathrm{F}}^2 + C f(M) \cdot \frac{1}{d}\, \mathbb{E}\|\boldsymbol{\theta}^{t_k}\|_{\mathrm{F}}^2 \,. \tag{189}$$

Iterating this inequality yields

$$\frac{1}{d}\, \mathbb{E}\|\boldsymbol{\theta}^{t_k} - \check{\boldsymbol{\theta}}^{t_k}\|_{\mathrm{F}}^2 \le C f(M) \cdot \frac{1}{d} \sup_{k \le T/\gamma} \mathbb{E}\|\boldsymbol{\theta}^{t_k}\|_{\mathrm{F}}^2 \,. \tag{190}$$

As $n, d \to \infty$, $\mathbb{E}\|\boldsymbol{\theta}^{t_k}\|_{\mathrm{F}}^2/d$ is uniformly bounded in t_k almost surely for large d by Lemma G.2. As $M \to \infty$, we have $f(M) \to 0$ by the dominated convergence theorem, and the first claim follows. Convergence of the 2-Wasserstein distances follows from the first claim by the same argument as in the proof of Lemma D.1. ■

Consider the following AMP iteration. Given sequences of functions $f_i \colon \mathbb{R}^{(i+1)m+i+2} \to \mathbb{R}^m$ and $g_i \colon \mathbb{R}^{im+m} \to \mathbb{R}^m$ $(i \ge 0)$ that are Lipschitz in the first $(i+1)m$ and im arguments, respectively, we generate sequences of matrices $\boldsymbol{a}^{i+1} \in \mathbb{R}^{d \times m}$ and $\boldsymbol{b}^i \in \mathbb{R}^{n \times m}$ $(i \ge 0)$ as follows.

$$\boldsymbol{a}^{i+1} = -\frac{1}{\delta} \boldsymbol{X}^{\mathsf{T}} f_i(\boldsymbol{b}^0, \dots, \boldsymbol{b}^i; \boldsymbol{z}, [\boldsymbol{G}^0]_M, \dots, [\boldsymbol{G}^i]_M) + \sum_{j=0}^{i} g_j(\boldsymbol{a}^1, \dots, \boldsymbol{a}^j; \boldsymbol{\theta}^0) \xi_{i,j}^{\mathsf{T}} \,, \tag{191}$$

$$\boldsymbol{b}^i = \boldsymbol{X} g_i(\boldsymbol{a}^1, \dots, \boldsymbol{a}^i; \boldsymbol{\theta}^0) + \frac{1}{\delta} \sum_{j=0}^{i-1} f_j(\boldsymbol{b}^0, \dots, \boldsymbol{b}^j; \boldsymbol{z}, [\boldsymbol{G}^0]_M, \dots, [\boldsymbol{G}^j]_M) \zeta_{i,j}^{\mathsf{T}} \,, \tag{192}$$

with initial values $g_0(\boldsymbol{\theta}^0) = \boldsymbol{\theta}^0, \boldsymbol{b}^0 = \boldsymbol{X}\boldsymbol{\theta}^0$. Here, f_i and g_i are applied row-wise, and $\{\xi_{i,j}\}_{0 \le j \le i}$, $\{\zeta_{i,j}\}_{0 \le j \le i-1} \subset \mathbb{R}^{m \times m}$ are defined as follows. We define a sequence of centered Gaussian random variables $\{\overline{u}^{i+1}, \overline{w}^i\}_{i \ge 0}$ recursively as

$$\mathbb{E}[\overline{w}^i (\overline{w}^j)^{\mathsf{T}}] = \mathbb{E}[g_i(\overline{u}^1, \dots, \overline{u}^i; \theta^0) g_j(\overline{u}^1, \dots, \overline{u}^j; \theta^0)^{\mathsf{T}}] \,, \tag{193}$$

$$\mathbb{E}[\overline{u}^{i+1} (\overline{u}^{j+1})^{\mathsf{T}}] = \frac{1}{\delta}\, \mathbb{E}[f_i(\overline{w}^0, \dots, \overline{w}^i; z, [\overline{G}^0]_M, \dots, [\overline{G}^i]_M) f_j(\overline{w}^0, \dots, \overline{w}^j; z, [\overline{G}^0]_M, \dots, [\overline{G}^j]_M)^{\mathsf{T}}] \,, \tag{194}$$

for $0 \le j \le i$, and set $\zeta_{i,j}, \xi_{i,j}$ as

$$\zeta_{i,j} = \mathbb{E}\left[\frac{\partial}{\partial \overline{u}^{j+1}} g_i(\overline{u}^1, \dots, \overline{u}^i; \theta^0)\right] , \qquad 0 \le j \le i-1 \,, \tag{195}$$

$$\xi_{i,j} = \mathbb{E}\left[\frac{\partial}{\partial \overline{w}^j} f_i(\overline{w}^0, \dots, \overline{w}^i; z, [\overline{G}^0]_M, \dots, [\overline{G}^i]_M)\right] , \qquad 0 \le j \le i \,, \tag{196}$$

where the expectations are taken over $\overline{u}^i$, $\overline{w}^i$, $\theta^0 \sim \mathsf{P}(\theta^0)$, $z \sim \mathsf{P}(z)$, and $\overline{G}^i \sim \mathsf{N}(0,1)$.

This AMP iteration can be mapped to the recursion (181) by considering the specific choice of f_i and g_i as follows.

$$g_i(\boldsymbol{a}^1, \dots, \boldsymbol{a}^i; \boldsymbol{\theta}^0) = \check{\boldsymbol{\theta}}^{t_i}\,, \tag{197}$$

$$f_i(\boldsymbol{b}^0, \dots, \boldsymbol{b}^i; \boldsymbol{z}, [\boldsymbol{G}^0]_M, \dots, [\boldsymbol{G}^i]_M) = (\mathbf{1}_n + \sqrt{\tau\delta/\gamma}[\boldsymbol{G}^i]_M) \odot \ell_{t_i}(\boldsymbol{r}^{t_i}; \boldsymbol{z})\,. \tag{198}$$

We show that g_i is indeed a function of $\boldsymbol{a}^1, \dots, \boldsymbol{a}^i, \boldsymbol{\theta}^0$ and Lipschitz in $\boldsymbol{a}^j$, and that f_i is a function of $\boldsymbol{b}^0, \dots, \boldsymbol{b}^i, \boldsymbol{z}, [\boldsymbol{G}^0]_M, \dots, [\boldsymbol{G}^i]_M$ and Lipschitz in $\boldsymbol{b}^j$. It can be shown by the Lipschitz continuity of ℓ, boundedness of $[G]_M$, and induction over i as follows.

$$\begin{aligned}
\check{\boldsymbol{\theta}}^{t_{i+1}} &= \check{\boldsymbol{\theta}}^{t_i} - \gamma h_{t_i}(\check{\boldsymbol{\theta}}^{t_i}) - \frac{\gamma}{\delta}\boldsymbol{X}^\mathsf{T} f_i(\boldsymbol{b}^0, \dots, \boldsymbol{b}^i; \boldsymbol{z}, [\boldsymbol{G}^0]_M, \dots, [\boldsymbol{G}^i]_M) \\
&= \check{\boldsymbol{\theta}}^{t_i} - \gamma h_{t_i}(\check{\boldsymbol{\theta}}^{t_i}) + \gamma\left(\boldsymbol{a}^{i+1} - \sum_{j=0}^{i} g_j(\boldsymbol{a}^1, \dots, \boldsymbol{a}^j; \boldsymbol{\theta}^0)\xi_{i,j}^\mathsf{T}\right), \qquad (199)\\
\boldsymbol{r}^{t_i} &= \boldsymbol{X}\boldsymbol{\theta}^{t_i} = \boldsymbol{X} g_i(\boldsymbol{a}^1, \dots, \boldsymbol{a}^i; \boldsymbol{\theta}^0) \\
&= \boldsymbol{b}^i - \frac{1}{\delta}\sum_{j=0}^{i-1} f_j(\boldsymbol{b}^0, \dots, \boldsymbol{b}^j; \boldsymbol{z}, [\boldsymbol{G}^0]_M, \dots, [\boldsymbol{G}^j]_M)\zeta_{i,j}^\mathsf{T}\,. \qquad (200)
\end{aligned}$$

By Wang et al. [58, Theorem 2.21], for any second order pseudo-Lipschitz functions $\psi\colon \mathbb{R}^{im+m} \to \mathbb{R}$ and $\widetilde{\psi}\colon \mathbb{R}^{(i+1)m+i+2} \to \mathbb{R}$, we have almost surely

$$\lim_{n,d\to\infty} \frac{1}{d}\sum_{j=1}^{d} \psi(a_j^1, \dots, a_j^i; \theta_j^0) = \mathbb{E}[\psi(\overline{u}^1, \dots, \overline{u}^i; \theta^0)]\,, \tag{201}$$

$$\lim_{n,d\to\infty} \frac{1}{n}\sum_{j=1}^{n} \widetilde{\psi}(b_j^0, \dots, b_j^i; z_j, G_j^0, \dots, G_j^i) = \mathbb{E}[\widetilde{\psi}(\overline{w}^0, \dots, \overline{w}^i; z, \overline{G}^0, \dots, \overline{G}^i)]\,. \tag{202}$$

Since $\check{\boldsymbol{\theta}}^{t_i}$ is a Lipschitz function of $\boldsymbol{a}^1, \dots, \boldsymbol{a}^i, \boldsymbol{\theta}^0$, we can take a Lipschitz function h_θ such that $\check{\boldsymbol{\theta}}^{t_i} = h_\theta(\boldsymbol{a}^1, \dots, \boldsymbol{a}^i; \boldsymbol{\theta}^0)$ and define $\overline{\theta}^i := h_\theta(\overline{u}^1, \dots, \overline{u}^i; \theta^0)$. Similarly, we can take a Lipschitz function h_r such that $\boldsymbol{r}^{t_i} = h_r(\boldsymbol{b}^0, \dots, \boldsymbol{b}^i; \boldsymbol{z}, [\boldsymbol{G}^0]_M, \dots, [\boldsymbol{G}^i]_M)$ and define $\overline{r}^i := h_r(\overline{w}^0, \dots, \overline{w}^i; z, [\overline{G}^0]_M, \dots, [\overline{G}^i]_M)$. Considering the composition of $\psi, \widetilde{\psi}$ with h_θ, h_r, we have almost surely

$$\lim_{n,d\to\infty} \frac{1}{d}\sum_{j=1}^{d} \psi(\check{\theta}_j^{t_1}, \dots, \check{\theta}_j^{t_i}; \theta_j^0) = \mathbb{E}[\psi(\overline{\theta}^1, \dots, \overline{\theta}^i; \theta^0)]\,, \tag{203}$$

$$\lim_{n,d\to\infty} \frac{1}{n}\sum_{j=1}^{n} \widetilde{\psi}(r_j^0, \dots, r_j^{t_i}; z_j, [G_j^0]_M, \dots, [G_j^i]_M) = \mathbb{E}[\widetilde{\psi}(\overline{r}^0, \dots, \overline{r}^i; z, [\overline{G}^0]_M, \dots, [\overline{G}^i]_M)]\,. \tag{204}$$

As $M \to \infty$, by the dominated convergence theorem, we have $W_2([G]_M, G) \to 0$ for $G \sim \mathsf{N}(0,1)$. Combining this with the above and Lemma D.4, we have

$$\operatorname*{p-lim}_{n,d\to\infty} W_2\Big(\hat{\mathsf{P}}(\boldsymbol{\theta}^{t_0}, \dots, \boldsymbol{\theta}^{t_i}), \mathsf{P}(\theta^0, \overline{\theta}^1, \dots, \overline{\theta}^i)\Big) = 0\,, \tag{205}$$

$$\operatorname*{p-lim}_{n,d\to\infty} W_2\Big(\hat{\mathsf{P}}(\boldsymbol{r}^{t_0}, \dots, \boldsymbol{r}^{t_i}, \boldsymbol{z}), \mathsf{P}(\overline{r}^0, \dots, \overline{r}^i; z, \overline{G}^0, \dots, \overline{G}^i)\Big) = 0\,. \tag{206}$$

D.2.2. MAPPING THE STATE EVOLUTION TO DMFT

It remains to show that the state evolution process $(\overline{\theta}^i, \overline{r}^i)_{i\geq 0}$ defined above satisfies the discretized DMFT equations (30)–(32).

By Equations (193) and (194), we have

$$\begin{aligned}\mathbb{E}[\overline{w}^i(\overline{w}^j)^\mathsf{T}] &= \mathbb{E}[\overline{\theta}^i(\overline{\theta}^j)^\mathsf{T}]\,,\\ \mathbb{E}[\overline{u}^{i+1}(\overline{u}^{j+1})^\mathsf{T}] &= \frac{1}{\delta}\,\mathbb{E}[(1+\sqrt{\tau\delta/\gamma}\overline{G}^i)\ell_{t_i}(\overline{r}^i;z)(1+\sqrt{\tau\delta/\gamma}\overline{G}^j)\ell_{t_j}(\overline{r}^j;z)^\mathsf{T}]\,.\end{aligned} \tag{207}$$

Define $\overline{U}^i$ and $\overline{L}^i$ as

$$\overline{U}^i := \gamma\sum_{j=1}^{i}\overline{u}^i\,,\quad \overline{L}^i := \gamma\sum_{j=0}^{i-1}(1+\sqrt{\tau\delta/\gamma}\overline{G}^j)\ell_{t_j}(\overline{r}^j;z)\,. \tag{208}$$

Then, we have

$$\mathbb{E}[\overline{U}^i(\overline{U}^j)^\mathsf{T}] = \frac{1}{\delta}\,\mathbb{E}[\overline{L}^i(\overline{L}^j)^\mathsf{T}]\,. \tag{209}$$

By Equation (199), $\overline{\theta}^i$ follows the following recursion.

$$\begin{aligned}\overline{\theta}^{i+1} &= \overline{\theta}^i - \gamma h_{t_i}(\overline{\theta}^i) + \gamma\left(\overline{u}^{i+1} - \sum_{j=0}^{i}\xi_{i,j}\overline{\theta}^j\right)\\ &= \overline{\theta}^i + \gamma\left(\overline{u}^{i+1} - h_{t_i}(\overline{\theta}^i) - \xi_{i,i}\overline{\theta}^i - \sum_{j=0}^{i-1}\xi_{i,j}\overline{\theta}^j\right)\,.\end{aligned} \tag{210}$$

Thus,

$$\begin{aligned}\overline{\theta}^i = \overline{\theta}^0 + \sum_{j=0}^{i-1}(\overline{\theta}^{j+1}-\overline{\theta}^j) &= \overline{\theta}^0 + \gamma\sum_{j=0}^{i-1}\overline{u}^{j+1} - \gamma\sum_{j=0}^{i-1}\left(h_{t_j}(\overline{\theta}^j) + \xi_{j,j}\overline{\theta}^j + \sum_{k=0}^{j-1}\xi_{j,k}\overline{\theta}^k\right)\\ &= \overline{\theta}^0 + \overline{U}^i - \gamma\sum_{j=0}^{i-1}\left(h_{t_j}(\overline{\theta}^j) + \xi_{j,j}\overline{\theta}^j + \sum_{k=0}^{j-1}\xi_{j,k}\overline{\theta}^k\right)\,.\end{aligned} \tag{211}$$

Furthermore, $\partial\overline{\theta}^i/\partial\overline{u}^{j+1}$ satisfies

$$\frac{\partial\overline{\theta}^i}{\partial\overline{u}^{j+1}} = \gamma I_m - \gamma\sum_{k=j+1}^{i-1}\left((\nabla_\theta h_{t_k}(\overline{\theta}^k) + \xi_{k,k})\frac{\partial\overline{\theta}^k}{\partial\overline{u}^{j+1}} + \sum_{l=j+1}^{k-1}\xi_{k,l}\frac{\partial\overline{\theta}^l}{\partial\overline{u}^{j+1}}\right)\,. \tag{212}$$

By Equation (200), $\overline{r}^i$ follows the following recursions.

$$\overline{r}^i = \overline{w}^i - \frac{1}{\delta}\sum_{j=0}^{i-1}\zeta_{i,j}\ell_{t_j}(\overline{r}^j;z)(1+\sqrt{\tau\delta/\gamma}\overline{G}^j)\,. \tag{213}$$

Furthermore, $\partial \ell_{t_i}(\overline{r}^i; z)/\partial \overline{w}^j$ satisfies

$$\frac{\partial \ell_{t_i}(\overline{r}^i; z)}{\partial \overline{w}^j} = \nabla_r \ell_{t_i}(\overline{r}^i; z) \frac{\partial \overline{r}^i}{\partial \overline{w}^j}\,, \tag{214}$$

$$\frac{\partial \overline{r}^i}{\partial \overline{w}^j} = -\frac{1}{\delta} \sum_{k=j+1}^{i-1} \zeta_{i,k} \frac{\partial \ell_{t_k}(\overline{r}^k; z)}{\partial \overline{w}^j} (1 + \sqrt{\tau\delta/\gamma}\overline{G}^k) - \frac{1}{\delta} \zeta_{i,j} \nabla_r \ell_{t_j}(\overline{r}^j; z)(1 + \sqrt{\tau\delta/\gamma}\overline{G}^j)\,. \tag{215}$$

Let $\{\overline{\rho}_\ell^{i,j}\}_{0 \le j < i}$ be the stochastic process satisfying

$$\overline{\rho}_\ell^{i,j} = \nabla_r \ell_{t_i}(\overline{r}^i; z) \overline{\rho}_r^{i,j}\,, \quad \overline{\rho}_r^{i,j} = -\frac{1}{\delta} \sum_{k=j+1}^{i-1} \zeta_{i,k} \overline{\rho}_\ell^{k,j} (1 + \sqrt{\tau\delta/\gamma}\overline{G}^k) - \frac{1}{\delta} \zeta_{i,j} \nabla_r \ell_{t_j}(\overline{r}^j; z)\,. \tag{216}$$

Using the linearity of Equation (214), we have

$$\frac{\partial \ell_{t_i}(\overline{r}^i; z)}{\partial \overline{w}^j} = \overline{\rho}_\ell^{i,j} (1 + \sqrt{\tau\delta/\gamma}\overline{G}^j)\,. \tag{217}$$

Then, $\xi_{i,j}$ satisfies

$$\xi_{i,j} = \mathbb{E}\left[\frac{\partial \ell_{t_i}(\overline{r}^i; z)}{\partial \overline{w}^j}\right] = \mathbb{E}\left[\overline{\rho}_\ell^{i,j}(1 + \sqrt{\tau\delta/\gamma}\overline{G}^j)\right] = \mathbb{E}[\overline{\rho}_\ell^{i,j}] + \sqrt{\frac{\tau\delta}{\gamma}}\, \mathbb{E}\left[\frac{\partial \overline{\rho}_\ell^{i,j}}{\partial \overline{G}^j}\right]\,, \tag{218}$$

where we used Stein's lemma (Gaussian integration by parts) in the last equality. By Equation (216), $\partial \overline{\rho}_\ell^{i,j}/\partial \overline{G}^j$ satisfies

$$\begin{aligned}\frac{\partial \overline{\rho}_\ell^{i,j}}{\partial \overline{G}^j} &= \nabla_r \ell_{t_i}(\overline{r}^i; z) \frac{\partial \overline{\rho}_r^{i,j}}{\partial \overline{G}^j} + \frac{\partial \nabla_r \ell_{t_i}(\overline{r}^i; z)}{\partial \overline{G}^j} \overline{\rho}_r^{i,j} \\ &= \nabla_r \ell_{t_i}(\overline{r}^i; z) \left(-\frac{1}{\delta} \sum_{k=j+1}^{i-1} \zeta_{i,k} \frac{\partial \overline{\rho}_\ell^{k,j}}{\partial \overline{G}^j} (1 + \sqrt{\tau\delta/\gamma}\overline{G}^k) \right) + \nabla_r^2 \ell_{t_i}(\overline{r}^i; z) \left[\frac{\partial \overline{r}^i}{\partial \overline{G}^j}\right] \overline{\rho}_r^{i,j}\,,\end{aligned} \tag{219}$$

where $\partial \overline{r}^i/\partial \overline{G}^j$ satisfies

$$\frac{\partial \overline{r}^i}{\partial \overline{G}^j} = -\frac{1}{\delta} \sum_{k=j+1}^{i-1} \zeta_{i,k} \nabla_r \ell_{t_k}(\overline{r}^k; z) \frac{\partial \overline{r}^k}{\partial \overline{G}^j} (1 + \sqrt{\tau\delta/\gamma}\overline{G}^k) - \sqrt{\frac{\tau}{\delta\gamma}} \zeta_{i,j} \ell_{t_j}(\overline{r}^j; z)\,. \tag{220}$$

These state evolution recursions exactly correspond to the discrete DMFT equations $\mathfrak{S}^\gamma$ shown in Equations (30) and (32) by the following mappings:

$$\theta_\gamma^{t_i} \overset{\mathrm{d}}{=} \overline{\theta}^i\,, \qquad i \ge 0\,, \qquad \text{(Compare (30a) with (211))}$$

$$\rho_{\theta,\gamma}^{t_i,t_j} \overset{\mathrm{d}}{=} \frac{1}{\gamma} \frac{\partial \overline{\theta}^i}{\partial \overline{u}^{j+1}}\,, \qquad 0 \le j < i\,, \qquad \text{(Compare (30b) with (212))}$$

$$r_\gamma^{t_i} \overset{\mathrm{d}}{=} \overline{r}^i\,, \qquad i \ge 0\,, \qquad \text{(Compare (30c) with (213))}$$

$$\rho_{\ell,\gamma}^{t_i,t_j} \overset{\mathrm{d}}{=} \frac{1}{\gamma}\overline{\rho}_\ell^{i,j}\,, \qquad 0 \le j < i\,, \qquad \text{(Compare (30d), (30e) with (216))}$$

$$D_{\ell,\gamma}^{t_i,t_j} \overset{\mathrm{d}}{=} \frac{1}{\sqrt{\gamma}}\frac{\partial \overline{\rho}_\ell^{i,j}}{\partial \overline{G}^j}\,, \qquad 0 \le j < i\,, \qquad \text{(Compare (30f), (30g) with (219), (220))}$$

$$U_\gamma^{t_i} \overset{\mathrm{d}}{=} \overline{U}^i\,, \qquad i \ge 0\,, \tag{221}$$

$$w_\gamma^{t_i} \overset{\mathrm{d}}{=} \overline{w}^i\,, \qquad i \ge 0\,, \tag{222}$$

$$B^{t_{i+1}} - B^{t_i} \overset{\mathrm{d}}{=} \sqrt{\gamma}\overline{G}^i\,, \qquad i \ge 0\,, \tag{223}$$

and

$$C_\theta^\gamma(t_i,t_j) = \mathbb{E}[\overline{\theta}^i(\overline{\theta}^j)^\mathsf{T}]\,, \qquad i,j \ge 0\,, \tag{224}$$

$$\Sigma_\ell^\gamma(t_i,t_j) = \mathbb{E}[\overline{L}^i(\overline{L}^j)^\mathsf{T}]\,, \qquad i,j \ge 0\,, \tag{225}$$

$$R_\theta^\gamma(t_i,t_j) = \zeta_{i,j}/\gamma\,, \qquad 0 \le j < i\,, \tag{226}$$

$$R_\ell^\gamma(t_i,t_j) = \xi_{i,j}/\gamma\,, \qquad 0 \le j < i\,, \tag{227}$$

$$\Gamma^\gamma(t_i) = \xi_{i,i}\,, \qquad 0 \le i\,. \tag{228}$$

D.3. Proof of Lemma D.3

We first embed the discretized DMFT equation $\mathfrak{S}^\gamma$ defined for discrete time knots $t_k = k\gamma$ ($k \ge 0$) into continuous time $t \in [0,T]$ in a piecewise constant manner. We define the stochastic processes $\{\theta_\gamma^t, r_\gamma^t\}_{t\in[0,T]}$ and $\{\rho_{\theta,\gamma}^{t,t'}, \rho_{\ell,\gamma}^{t,t'}, D_{\ell,\gamma}^{t,t'}\}_{t\ge t'\ge 0}$ by the following equations.

$$\theta_\gamma^t = \theta^0 + U_\gamma^t - \int_0^{\lfloor t\rfloor}\big(h_{\lfloor s\rfloor}(\theta_\gamma^s) + \Gamma^\gamma(s)\theta_\gamma^s\big) + \int_0^{\lfloor s\rfloor} R_\ell^\gamma(s,s')\theta_\gamma^{s'}\,\mathrm{d}s')\,\mathrm{d}s\,, \tag{229}$$

$$\rho_{\theta,\gamma}^{t,t'} = I_m - \mathbb{I}(\lceil t'\rceil \le \lfloor t\rfloor)\int_{\lceil t'\rceil}^{\lfloor t\rfloor}\left((\nabla_\theta h_{\lfloor s\rfloor}(\theta_\gamma^s) + \Gamma^\gamma(s))\rho_{\theta,\gamma}^{s,t'} + \int_{\lceil t'\rceil}^{\lfloor s\rfloor} R_\ell^\gamma(s,s')\rho_{\theta,\gamma}^{s',t'}\,\mathrm{d}s'\right)\mathrm{d}s\,, \tag{230}$$

$$r_\gamma^t = w_\gamma^t - \frac{1}{\delta}\int_0^{\lfloor t\rfloor} R_\theta^\gamma(t,s)\ell_{\lfloor s\rfloor}(r_\gamma^s;z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\,, \tag{231}$$

$$\rho_{\ell,\gamma}^{t,t'} = \nabla_r\ell_{\lfloor t\rfloor}(r_\gamma^t;z)\rho_{r,\gamma}^{t,t'}\,, \tag{232}$$

$$\rho_{r,\gamma}^{t,t'} = -\frac{1}{\delta}\mathbb{I}(\lceil t'\rceil \le \lfloor t\rfloor)\int_{\lceil t'\rceil}^{\lfloor t\rfloor} R_\theta^\gamma(t,s)\rho_{\ell,\gamma}^{s,t'}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) - \frac{1}{\delta}R_\theta^\gamma(t,t')\nabla_r\ell_{\lfloor t'\rfloor}(r_\gamma^{t'};z)\,, \tag{233}$$

$$\begin{aligned} D_{\ell,\gamma}^{t,t'} &= \nabla_r\ell_{\lfloor t\rfloor}(r_\gamma^t;z)\left(-\frac{1}{\delta}\mathbb{I}(\lceil t'\rceil \le \lfloor t\rfloor)\int_{\lceil t'\rceil}^{\lfloor t\rfloor} R_\theta^\gamma(t,s)D_{\ell,\gamma}^{s,t'}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\right) \\ &\quad + \nabla_r^2\ell_{\lfloor t\rfloor}(r_\gamma^t;z)[D_{r,\gamma}^{t,t'}]\rho_{r,\gamma}^{t,t'}\,, \end{aligned} \tag{234}$$

$$\begin{aligned} D_{r,\gamma}^{t,t'} &= -\frac{1}{\delta}\mathbb{I}(\lceil t'\rceil \le \lfloor t\rfloor)\int_{\lceil t'\rceil}^{\lfloor t\rfloor} R_\theta^\gamma(t,s)\nabla_r\ell_{\lfloor s\rfloor}(r_\gamma^s;z)D_{r,\gamma}^{s,t'}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) \\ &\quad - \sqrt{\frac{\tau}{\delta}}R_\theta^\gamma(t,t')\ell_{\lfloor t'\rfloor}(r_\gamma^{t'};z)\,, \end{aligned} \tag{235}$$

where U_γ^t, w_γ^t are centered Gaussian processes with covariance kernels $\Sigma_\ell^\gamma/\delta$ and C_θ^γ respectively. Then, set $C_\theta^\gamma, R_\theta^\gamma, \Sigma_\ell^\gamma, R_\ell^\gamma, \Gamma^\gamma$ as

$$\begin{aligned} C_\theta^\gamma(t,t') &= \mathbb{E}[\theta_\gamma^t \theta_\gamma^{t'\mathsf{T}}]\,, \quad R_\theta^\gamma(t,t') = \mathbb{E}[\rho_{\theta,\gamma}^{t,t'}]\,, \\ \Sigma_\ell^\gamma(t,t') &= \mathbb{E}[L_\gamma^t L_\gamma^{t'\mathsf{T}}]\,, \quad L_\gamma^t := \int_0^{\lfloor t \rfloor} \ell_{\lfloor s \rfloor}(r_\gamma^s; z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\,, \\ R_\ell^\gamma(t,t') &= \mathbb{E}[\rho_{\ell,\gamma}^{t,t'}] + \sqrt{\tau\delta}\,\mathbb{E}[D_{\ell,\gamma}^{t,t'}]\,, \quad \Gamma^\gamma(t) = \mathbb{E}\big[\nabla_r \ell_{\lfloor t \rfloor}(r_\gamma^t; z)\big]\,, \end{aligned} \tag{236}$$

where we set $R_\theta^\gamma(t,t') = R_\ell^\gamma(t,t') = 0$ for $\lfloor t \rfloor < \lceil t' \rceil$.

The above equation agrees with the discretized DMFT equation $\mathfrak{S}^\gamma$ at discrete time points $t = t_i = i\gamma$ $(i \geq 0)$, and has a unique solution since it is piecewise constant in each interval $[t_i, t_{i+1})$.

We then define mappings $\mathcal{T}_{\theta\to\ell}^\gamma \colon (C_\theta, R_\theta) \mapsto (\Sigma_\ell, R_\ell, \Gamma)$ and $\mathcal{T}_{\ell\to\theta}^\gamma \colon (\Sigma_\ell, R_\ell, \Gamma) \mapsto (C_\theta, R_\theta)$ similarly to $\mathcal{T}_{\theta\to\ell}$ and $\mathcal{T}_{\ell\to\theta}$ in Appendix C but with the DMFT equation $\mathfrak{S}$ replaced by its discretized version $\mathfrak{S}^\gamma$. We also define their composition $\mathcal{T}^\gamma := \mathcal{T}_{\ell\to\theta}^\gamma \circ \mathcal{T}_{\theta\to\ell}^\gamma$.

Since the solution to the discretized DMFT equation $\mathfrak{S}^\gamma$ is determined by the values at discrete time points $\{t_i = i\gamma : i \geq 0\}$, the solution exists uniquely by induction. Let $X^\gamma = (C_\theta^\gamma, R_\theta^\gamma)$ and $Y^\gamma = (\Sigma_\ell^\gamma, R_\ell^\gamma, \Gamma^\gamma)$ be the solution to $\mathfrak{S}^\gamma$. Then, X^γ is the unique fixed point of $\mathcal{T}^\gamma$, i.e., $\mathcal{T}^\gamma(X^\gamma) = X^\gamma$.

We show that the unique solutions X^γ and Y^γ belong to some admissible spaces defined in Appendix C.1.

Lemma D.5 *There exist admissible spaces $\mathcal{S}_\theta$ and $\mathcal{S}_\ell$ such that $X^\gamma \in \mathcal{S}_\theta$ and $Y^\gamma \in \mathcal{S}_\ell$.*

Proof Since $C_\theta^\gamma, \Sigma_\ell^\gamma, R_\theta^\gamma, R_\ell^\gamma$, and Γ^γ are piecewise constant, the continuity conditions are automatically satisfied. Since the solutions are bounded, we can take the spaces $\mathcal{S}_\theta$ and $\mathcal{S}_\ell$ large enough so that the boundedness conditions are also satisfied. ■

Let $X = (C_\theta, R_\theta)$ be the unique fixed point of $\mathcal{T}$ shown in Theorem 3.1. We control their distance as follows.

$$\mathsf{dist}_\lambda(X, X^\gamma) = \mathsf{dist}_\lambda(\mathcal{T}(X), \mathcal{T}^\gamma(X^\gamma)) \leq \mathsf{dist}_\lambda(\mathcal{T}(X), \mathcal{T}(X^\gamma)) + \mathsf{dist}_\lambda(\mathcal{T}(X^\gamma), \mathcal{T}^\gamma(X^\gamma))\,. \tag{237}$$

As in the proof of Theorem 3.1, we can take λ large enough so that $\mathcal{T}$ is a contraction and the first term is bounded by $(1/2)\mathsf{dist}_\lambda(X, X^\gamma)$. Then, we have

$$\mathsf{dist}_\lambda(X, X^\gamma) \leq 2 \cdot \mathsf{dist}_\lambda(\mathcal{T}(X^\gamma), \mathcal{T}^\gamma(X^\gamma))\,. \tag{238}$$

The next lemma bounds the right-hand side. We prove it in Appendix D.3.1.

Lemma D.6 *For any $\lambda > 0$, there exists a constant $K > 0$ independent of γ such that*

$$\mathsf{dist}_\lambda(\mathcal{T}(X^\gamma), \mathcal{T}^\gamma(X^\gamma)) \leq K\sqrt{\gamma}\,. \tag{239}$$

Therefore, for sufficiently large λ, we have

$$\mathsf{dist}_\lambda(X, X^\gamma) \leq K\sqrt{\gamma}\,, \quad \mathsf{dist}_\lambda(Y, Y^\gamma) \leq K\sqrt{\gamma}\,. \tag{240}$$

Finally, we couple θ^t and θ_γ^t so that they are close. We prove it in Appendix D.3.2.

Lemma D.7 *There exist a constant $K > 0$ independent of γ and a coupling of the processes θ^t and θ^t_γ such that*

$$\sup_{0\le t\le T} \sqrt{\mathbb{E}\|\theta^t - \theta^t_\gamma\|_2^2} \le K(\sqrt{\gamma} + \mathsf{dist}_\lambda(Y, Y^\gamma))\,, \tag{241}$$

$$\sup_{0\le t\le T} \sqrt{\mathbb{E}\|r^t - r^t_\gamma\|_2^2} \le K(\sqrt{\gamma} + \mathsf{dist}_\lambda(X, X^\gamma))\,. \tag{242}$$

By the above Lemmas, we have

$$W_2\big(\mathsf{P}(\theta^{t_1}, \dots, \theta^{t_L}), \mathsf{P}(\theta^{t_1}_\gamma, \dots, \theta^{t_L}_\gamma)\big) \le \sqrt{\sum_{i=1}^L \mathbb{E}\|\theta^{t_i} - \theta^{t_i}_\gamma\|_2^2} \le \sqrt{KL}\cdot\sqrt{\gamma}\,, \tag{243}$$

$$W_2\big(\mathsf{P}(r^{t_1}, \dots, r^{t_L}, z), \mathsf{P}(r^{t_1}_\gamma, \dots, r^{t_L}_\gamma, z)\big) \le \sqrt{\sum_{i=1}^L \mathbb{E}\|r^{t_i} - r^{t_i}_\gamma\|_2^2} \le \sqrt{KL}\cdot\sqrt{\gamma}\,. \tag{244}$$

Sending $\gamma \to 0$ completes the proof of Lemma D.3.

D.3.1. PROOF OF LEMMA D.6

Let $\overline{Y}^\gamma = (\overline{\Sigma}^\gamma_\ell, \overline{R}^\gamma_\ell, \overline{\Gamma}^\gamma) = \mathcal{T}_{\theta\to\ell}(X^\gamma)$ and $Y^\gamma = (\Sigma^\gamma_\ell, R^\gamma_\ell, \Gamma^\gamma) = \mathcal{T}^\gamma_{\theta\to\ell}(X^\gamma)$. Also, let $\overline{X}^\gamma = (\overline{C}^\gamma_\theta, \overline{R}^\gamma_\theta) = \mathcal{T}_{\ell\to\theta}(\overline{Y}^\gamma)$.

In the following, K denotes a positive constant independent of γ whose value may change from line to line. Note that K can depend on λ since we fix λ and send $\gamma \to 0$.

The proof proceeds as follows.

1. We show that $\mathsf{dist}_\lambda(\overline{Y}^\gamma, Y^\gamma) \le K\sqrt{\gamma}$ for some constant $K > 0$.
2. We show that $\mathsf{dist}_\lambda(\overline{X}^\gamma, X^\gamma) \le K(\sqrt{\gamma} + \mathsf{dist}_\lambda(\overline{Y}^\gamma, Y^\gamma))$ for some constant $K > 0$.

Together they prove the Lemma D.6.

Bound of $\mathsf{dist}_\lambda(\overline{\Sigma}^\gamma_\ell, \Sigma^\gamma_\ell)$ Let $w_\gamma \sim \mathsf{GP}(0, C^\gamma_\theta)$. Let $\overline{r}^t_\gamma$ and r^t_γ be the solutions to the following equations:

$$\overline{r}^t_\gamma = w^t_\gamma - \frac{1}{\delta}\int_0^t R^\gamma_\theta(t,s)\ell_s(\overline{r}^s_\gamma; z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\,, \tag{245}$$

$$r^t_\gamma = w^t_\gamma - \frac{1}{\delta}\int_0^{\lfloor t\rfloor} R^\gamma_\theta(t,s)\ell_{\lfloor s\rfloor}(r^s_\gamma; z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\,. \tag{246}$$

We have

$$\begin{aligned}\overline{r}^t_\gamma - r^t_\gamma = &-\frac{1}{\delta}\int_0^{\lfloor t\rfloor} R^\gamma_\theta(t,s)(\ell_s(\overline{r}^s_\gamma; z) - \ell_{\lfloor s\rfloor}(r^s_\gamma; z))(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\\ &-\frac{1}{\delta}\int_{\lfloor t\rfloor}^t R^\gamma_\theta(t,s)\ell_s(\overline{r}^s_\gamma; z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\,,\end{aligned} \tag{247}$$

and thus

$$\begin{aligned}
\mathbb{E}\|\overline{r}_\gamma^t - r_\gamma^t\|_2^2 &\le 2\left(\frac{T}{\delta^2}+\frac{\tau}{\delta}\right)\int_0^{\lfloor t\rfloor}\|R_\theta^\gamma(t,s)\|_2^2\,\mathbb{E}\|\ell_s(\overline{r}_\gamma^s;z)-\ell_{\lfloor s\rfloor}(r_\gamma^s;z)\|_2^2\,\mathrm{d}s \\
&\quad + 2\left(\frac{T}{\delta^2}+\frac{\tau}{\delta}\right)\int_{\lfloor t\rfloor}^{t}\|R_\theta^\gamma(t,s)\|_2^2\,\mathbb{E}\|\ell_s(\overline{r}_\gamma^s;z)\|_2^2\,\mathrm{d}s \\
&\le 2\left(\frac{T}{\delta^2}+\frac{\tau}{\delta}\right)\Phi^2M^2\int_0^{\lfloor t\rfloor}\mathbb{E}\big[(|s-\lfloor s\rfloor|+\|\overline{r}_\gamma^s-r_\gamma^s\|_2)^2\big]\,\mathrm{d}s + 2\left(\frac{T}{\delta^2}+\frac{\tau}{\delta}\right)\Phi^3(t-\lfloor t\rfloor) \\
&\le 2\left(\frac{T}{\delta^2}+\frac{\tau}{\delta}\right)\Phi^2M^2\int_0^{\lfloor t\rfloor}(2\gamma^2+2\,\mathbb{E}\|\overline{r}_\gamma^s-r_\gamma^s\|_2^2)\,\mathrm{d}s + 2\left(\frac{T}{\delta^2}+\frac{\tau}{\delta}\right)\Phi^3\gamma \\
&\le K\gamma + K\int_0^t\mathbb{E}\|\overline{r}_\gamma^s-r_\gamma^s\|_2^2\,\mathrm{d}s\,. \qquad (248)
\end{aligned}$$

By Grönwall's inequality, we have

$$\sup_{t\in[0,T]}\mathbb{E}\|\overline{r}_\gamma^t-r_\gamma^t\|_2^2 \le K\gamma\mathrm{e}^{KT}\le K\gamma\,. \qquad (249)$$

Let

$$\overline{L}_\gamma^t := \int_0^t \ell_s(\overline{r}_\gamma^s;z)(\mathrm{d}s+\sqrt{\tau\delta}\,\mathrm{d}B^s)\,,\quad L_\gamma^t := \int_0^{\lfloor t\rfloor}\ell_{\lfloor s\rfloor}(r_\gamma^s;z)(\mathrm{d}s+\sqrt{\tau\delta}\,\mathrm{d}B^s)\,. \qquad (250)$$

Then, we have

$$\begin{aligned}
\mathbb{E}\|\overline{L}_\gamma^t-L_\gamma^t\|_2^2 &\le 2(T+\tau\delta)\int_0^{\lfloor t\rfloor}\mathbb{E}\|\ell_s(\overline{r}_\gamma^s;z)-\ell_{\lfloor s\rfloor}(r_\gamma^s;z)\|_2^2\,\mathrm{d}s + 2(T+\tau\delta)\int_{\lfloor t\rfloor}^t\mathbb{E}\|\ell_s(\overline{r}_\gamma^s;z)\|_2^2\,\mathrm{d}s \\
&\le 2(T+\tau\delta)M^2\int_0^{\lfloor t\rfloor}\mathbb{E}\big[(|s-\lfloor s\rfloor|+\|\overline{r}_\gamma^s-r_\gamma^s\|_2)^2\big]\,\mathrm{d}s + 2(T+\tau\delta)\Phi(t-\lfloor t\rfloor) \\
&\le K\gamma\,. \qquad (251)
\end{aligned}$$

Let $\{(\overline{U}_\gamma^t, U_\gamma^t)\}_{t\in[0,T]}$ be a centered Gaussian process with covariance $\mathbb{E}\left[\begin{pmatrix}\overline{L}_\gamma^t\\ L_\gamma^t\end{pmatrix}\begin{pmatrix}\overline{L}_\gamma^t\\ L_\gamma^t\end{pmatrix}^\mathsf{T}\right]/\delta$. Since $\overline{U}_\gamma$ and U_γ have covariance kernels $\overline{\Sigma}_\ell^\gamma/\delta$ and $\Sigma_\ell^\gamma/\delta$ respectively, we have

$$\mathsf{dist}_\lambda(\overline{\Sigma}_\ell^\gamma,\Sigma_\ell^\gamma) \le \sup_{t\in[0,T]}\mathrm{e}^{-\lambda t}\sqrt{\mathbb{E}\|\overline{U}_\gamma^t-U_\gamma^t\|_2^2} = \sup_{t\in[0,T]}\mathrm{e}^{-\lambda t}\sqrt{\mathbb{E}\|\overline{L}_\gamma^t-L_\gamma^t\|_2^2/\delta} \le K\sqrt{\gamma}\,. \qquad (252)$$

Bound of $\mathsf{dist}_\lambda(\overline{\Gamma}^\gamma,\Gamma^\gamma)$**.** By Equation (249), we have

$$\begin{aligned}
\mathsf{dist}_\lambda(\overline{\Gamma}^\gamma,\Gamma^\gamma) &= \sup_{t\in[0,T]}\mathrm{e}^{-\lambda t}\|\overline{\Gamma}^\gamma(t)-\Gamma^\gamma(t)\|_2 \le \sup_{t\in[0,T]}\sqrt{\mathbb{E}\|\nabla_r\ell_t(\overline{r}_\gamma^t;z)-\nabla_r\ell_{\lfloor t\rfloor}(r_\gamma^t;z)\|_2^2} \\
&\le M\sup_{t\in[0,T]}\sqrt{2(t-\lfloor t\rfloor)^2+2\,\mathbb{E}\|\overline{r}_\gamma^t-r_\gamma^t\|_2^2} \le K\sqrt{\gamma}\,. \qquad (253)
\end{aligned}$$

Bound of $\mathsf{dist}_\lambda(R^1_\ell, R^2_\ell)$**.** Let $\overline{\rho}^{t,t'}_{r,\gamma}$ and $\rho^{t,t'}_{r,\gamma}$ be the stochastic processes satisfying

$$\overline{\rho}^{t,t'}_{r,\gamma} = -\frac{1}{\delta}\int_{t'}^{t} R^\gamma_\theta(t,s)\overline{\rho}^{s,t'}_{\ell,\gamma}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) - \frac{1}{\delta}R^\gamma_\theta(t,t')\nabla_r \ell_{t'}(\overline{r}^{t'}_\gamma; z)\,, \tag{254}$$

$$\rho^{t,t'}_{r,\gamma} = -\frac{1}{\delta}\mathbb{I}(\lceil t'\rceil \le \lfloor t\rfloor)\int_{\lceil t'\rceil}^{\lfloor t\rfloor} R^\gamma_\theta(t,s)\rho^{s,t'}_{\ell,\gamma}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) - \frac{1}{\delta}R^\gamma_\theta(t,t')\nabla_r \ell_{\lfloor t'\rfloor}(r^{t'}_\gamma; z)\,. \tag{255}$$

Let $\overline{\rho}^{t,t'}_{\ell,\gamma} = \nabla_r \ell_t(\overline{r}^t_\gamma; z)\overline{\rho}^{t,t'}_{r,\gamma}$ and $\rho^{t,t'}_{\ell,\gamma} = \nabla_r \ell_{\lfloor t\rfloor}(r^t_\gamma; z)\rho^{t,t'}_{r,\gamma}$. Similarly, define $\overline{D}^{t,t'}_{\ell,\gamma}$ and $D^{t,t'}_{\ell,\gamma}$ for processes defined by $\mathcal{T}_{\theta\to\ell}$ and $\mathcal{T}^\gamma_{\theta\to\ell}$, respectively. Then, we have

$$\|\overline{R}^\gamma_\ell(t,t') - R^\gamma_\ell(t,t')\|_2^2 \le \mathbb{E}\|\overline{\rho}^{t,t'}_{\ell,\gamma} - \rho^{t,t'}_{\ell,\gamma}\|_2^2 + \tau\delta\,\mathbb{E}\|\overline{D}^{t,t'}_{\ell,\gamma} - D^{t,t'}_{\ell,\gamma}\|_2^2\,. \tag{256}$$

We first bound $\mathbb{E}\|\overline{\rho}^{t,t'}_{\ell,\gamma} - \rho^{t,t'}_{\ell,\gamma}\|_2^2$. We have

$$\begin{aligned}
\mathbb{E}\|\overline{\rho}^{t,t'}_{\ell,\gamma} - \rho^{t,t'}_{\ell,\gamma}\|_2^2 &= \mathbb{E}\|\nabla_r \ell_t(\overline{r}^t_\gamma; z)\overline{\rho}^{t,t'}_{r,\gamma} - \nabla_r \ell_{\lfloor t\rfloor}(r^t_\gamma; z)\rho^{t,t'}_{r,\gamma}\|_2^2 \\
&\le 2\sqrt{\mathbb{E}\|\overline{\rho}^{t,t'}_{r,\gamma}\|_2^4}\sqrt{\mathbb{E}\|\nabla_r \ell_t(\overline{r}^t_\gamma; z) - \nabla_r \ell_{\lfloor t\rfloor}(r^t_\gamma; z)\|_2^4} + 2M^2\,\mathbb{E}\|\overline{\rho}^{t,t'}_{r,\gamma} - \rho^{t,t'}_{r,\gamma}\|_2^2 \\
&\le K\sqrt{\gamma^4 + \mathbb{E}\|\overline{r}^t_\gamma - r^t_\gamma\|_2^4} + K\,\mathbb{E}\|\overline{\rho}^{t,t'}_{r,\gamma} - \rho^{t,t'}_{r,\gamma}\|_2^2\,.
\end{aligned} \tag{257}$$

First, we bound $\mathbb{E}\|\overline{r}^t_\gamma - r^t_\gamma\|_2^4$. We have

$$\begin{aligned}
&\mathbb{E}\|\overline{r}^t_\gamma - r^t_\gamma\|_2^4 \\
&\le K\int_0^{\lfloor t\rfloor}\|R^\gamma_\theta(t,s)\|_2^4\,\mathbb{E}\|\ell_s(\overline{r}^s_\gamma; z) - \ell_{\lfloor s\rfloor}(r^s_\gamma; z)\|_2^4\,\mathrm{d}s + K(t - \lfloor t\rfloor)\int_{\lfloor t\rfloor}^t \|R^\gamma_\theta(t,s)\|_2^4\,\mathbb{E}\|\ell_s(\overline{r}^s_\gamma; z)\|_2^4\,\mathrm{d}s \\
&\le K\int_0^{\lfloor t\rfloor}(\gamma^4 + \mathbb{E}\|\overline{r}^s_\gamma - r^s_\gamma\|_2^4)\,\mathrm{d}s + K\gamma^2 \le K\gamma^2 + K\int_0^t \mathbb{E}\|\overline{r}^s_\gamma - r^s_\gamma\|_2^4\,\mathrm{d}s\,.
\end{aligned} \tag{258}$$

By Grönwall's inequality, we have

$$\sup_{t\in[0,T]}\mathbb{E}\|r^t_1 - r^t_2\|_2^4 \le K\gamma^2\,. \tag{259}$$

Next, we bound $\mathbb{E}\|\overline{\rho}^{t,t'}_{r,\gamma} - \rho^{t,t'}_{r,\gamma}\|_2^2$. We have

$$\begin{aligned}
\|\rho^{t,t'}_{r,\gamma} - \overline{\rho}^{t,t'}_{r,\gamma}\|_2 \le{}& \frac{1}{\delta}\left\|\int_{t'}^{\lceil t'\rceil} R^\gamma_\theta(t,s)\overline{\rho}^{s,t'}_{\ell,\gamma}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\right\|_2 + \frac{1}{\delta}\left\|\int_{\lfloor t\rfloor}^t R^\gamma_\theta(t,s)\overline{\rho}^{s,t'}_{\ell,\gamma}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\right\|_2 \\
&+ \frac{1}{\delta}\left\|\int_{\lceil t'\rceil}^{\lfloor t\rfloor} R^\gamma_\theta(t,s)(\rho^{s,t'}_{\ell,\gamma} - \overline{\rho}^{s,t'}_{\ell,\gamma})(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\right\|_2 \\
&+ \frac{1}{\delta}\|R^\gamma_\theta(t,t')(\nabla_r \ell_t(\overline{r}^{t'}_\gamma; z) - \nabla_r \ell_{\lfloor t'\rfloor}(r^{t'}_\gamma; z))\|_2\,,
\end{aligned} \tag{260}$$

Thus, we have

$$\mathbb{E}\|\rho^{t,t'}_{r,\gamma} - \overline{\rho}^{t,t'}_{r,\gamma}\|_2^2 \le K\int_{t'}^{\lceil t'\rceil}\|R^\gamma_\theta(t,s)\|_2^2\,\mathbb{E}\|\overline{\rho}^{s,t'}_{\ell,\gamma}\|_2^2\,\mathrm{d}s + K\int_{\lfloor t\rfloor}^t \|R^\gamma_\theta(t,s)\|_2^2\,\mathbb{E}\|\overline{\rho}^{s,t'}_{\ell,\gamma}\|_2^2\,\mathrm{d}s$$

$$
\begin{aligned}
&+ K \int_{\lceil t' \rceil}^{\lfloor t \rfloor} \|R_\theta^\gamma(t,s)\|_2^2 \, \mathbb{E}\|\rho_{\ell,\gamma}^{s,t'} - \overline{\rho}_{\ell,\gamma}^{s,t'}\|_2^2 \, \mathrm{d}s \\
&+ K \|R_\theta^\gamma(t,t')\|_2^2 \, \mathbb{E}\|\nabla_r \ell_t(\overline{r}_\gamma^{t'}; z) - \nabla_r \ell_{\lfloor t' \rfloor}(r_\gamma^{t'}; z)\|_2^2 \\
\le\ & K\gamma + K \int_{t'}^{t} \mathbb{E}\|\rho_{\ell,\gamma}^{s,t'} - \overline{\rho}_{\ell,\gamma}^{s,t'}\|_2^2 \, \mathrm{d}s \,. \qquad (261)
\end{aligned}
$$

By Grönwall's inequality, we have

$$
\sup_{0 \le t' \le t \le T} \mathbb{E}\|\rho_{r,\gamma}^{t,t'} - \overline{\rho}_{r,\gamma}^{t,t'}\|_2^2 \le K\gamma \,. \qquad (262)
$$

Next, we bound $\mathbb{E}\|\overline{D}_{\ell,\gamma}^{t,t'} - D_{\ell,\gamma}^{t,t'}\|_2^2$ similarly. We do not repeat the details, but following the same argument, we have

$$
\mathbb{E}\|\overline{D}_{\ell,\gamma}^{t,t'} - D_{\ell,\gamma}^{t,t'}\|_2^2 \le K\gamma \,. \qquad (263)
$$

Therefore, we have

$$
\|\overline{R}_\ell^\gamma(t,t') - R_\ell^\gamma(t,t')\|_2^2 \le K\gamma \,. \qquad (264)
$$

By Equations (252), (253) and (264), we have

$$
\mathsf{dist}_\lambda(\overline{Y}^\gamma, Y^\gamma) \le K\sqrt{\gamma} \,. \qquad (265)
$$

Bound of $\mathsf{dist}_\lambda(\overline{C}_\theta^\gamma, C_\theta^\gamma)$**.** Let $\overline{U}_\gamma \sim \mathsf{GP}(0, \overline{\Sigma}_\ell^\gamma/\delta)$ and $U_\gamma \sim \mathsf{GP}(0, \Sigma_\ell^\gamma/\delta)$ be Gaussian processes coupled such that

$$
\sup_{t \in [0,T]} \mathrm{e}^{-\lambda t} \sqrt{\mathbb{E}\|\overline{U}_\gamma^t - U_\gamma^t\|_2^2} \le 2 \cdot \mathsf{dist}_\lambda(\overline{\Sigma}_\ell^\gamma, \Sigma_\ell^\gamma) \,. \qquad (266)
$$

Let $\overline{\theta}_\gamma^t$ and θ_γ^t be the solution of

$$
\overline{\theta}_\gamma^t = \theta^0 + \overline{U}_\gamma^t - \int_0^t \left(h_s(\overline{\theta}_\gamma^s) + \overline{\Gamma}^\gamma(s)\overline{\theta}_\gamma^s + \int_0^s \overline{R}_\ell^\gamma(s,s')\overline{\theta}_\gamma^{s'} \, \mathrm{d}s' \right) \mathrm{d}s \,, \qquad (267)
$$

$$
\theta_\gamma^t = \theta^0 + U_\gamma^t - \int_0^{\lfloor t \rfloor} \left(h_{\lfloor s \rfloor}(\theta_\gamma^s) + \Gamma^\gamma(s)\theta_\gamma^s + \int_0^{\lfloor s \rfloor} R_\ell^\gamma(s,s')\theta_\gamma^{s'} \, \mathrm{d}s' \right) \mathrm{d}s \,. \qquad (268)
$$

Then, we have

$$
\begin{aligned}
\mathbb{E}\|\overline{\theta}_\gamma^t - \theta_\gamma^t\|_2^2 \le\ & K \left(\gamma + \mathbb{E}\|\overline{U}_\gamma^t - U_\gamma^t\|_2^2 + \sup_{t \in [0,T]} \|\overline{\Gamma}^\gamma(t) - \Gamma^\gamma(t)\|_2^2 + \sup_{0 \le t' \le t \le T} \|\overline{R}_\ell^\gamma(t,t') - R_\ell^\gamma(t,t')\|_2^2 \right) \\
&+ K \int_0^t \mathbb{E}\|\overline{\theta}_\gamma^s - \theta_\gamma^s\|_2^2 \, \mathrm{d}s \\
\le\ & K\big(\gamma + \mathsf{dist}_\lambda(\overline{Y}^\gamma, Y^\gamma)^2\big) + K \int_0^t \mathbb{E}\|\overline{\theta}_\gamma^s - \theta_\gamma^s\|_2^2 \, \mathrm{d}s \,. \qquad (269)
\end{aligned}
$$

By Grönwall's inequality, we have

$$\sup_{t\in[0,T]} \mathbb{E}\|\overline{\theta}^t_\gamma - \theta^t_\gamma\|_2^2 \le K\big(\gamma + \mathsf{dist}_\lambda(\overline{Y}^\gamma, Y^\gamma)^2\big)\,. \tag{270}$$

Let $\{(\overline{w}^t_\gamma, w^t_\gamma)\}_{t\in[0,T]}$ be a centered Gaussian process with covariance $\mathbb{E}\left[\begin{pmatrix}\overline{\theta}^t_\gamma\\ \theta^t_\gamma\end{pmatrix}\begin{pmatrix}\overline{\theta}^t_\gamma\\ \theta^t_\gamma\end{pmatrix}^\mathsf{T}\right]$. Since $\overline{w}_\gamma$ and w_γ have covariance kernels $\overline{C}^\gamma_\theta$ and C^γ_θ respectively, we have

$$\mathsf{dist}_\lambda(\overline{C}^\gamma_\theta, C^\gamma_\theta) \le \sup_{t\in[0,T]} \mathrm{e}^{-\lambda t}\sqrt{\mathbb{E}\|\overline{w}^t_\gamma - w^t_\gamma\|_2^2} = \sup_{t\in[0,T]} \mathrm{e}^{-\lambda t}\sqrt{\mathbb{E}\|\overline{\theta}^t_\gamma - \theta^t_\gamma\|_2^2} \le K(\sqrt{\gamma} + \mathsf{dist}_\lambda(\overline{Y}^\gamma, Y^\gamma))\,. \tag{271}$$

Bound of $\mathsf{dist}_\lambda(R^1_\theta, R^2_\theta)$**.** Let $\overline{\rho}^{t,t'}_{\theta,\gamma}$ and $\rho^{t,t'}_{\theta,\gamma}$ be the stochastic processes satisfying

$$\overline{\rho}^{t,t'}_{\theta,\gamma} = I_m - \int_{t'}^t \left((\nabla_\theta h_s(\overline{\theta}^s_\gamma) + \overline{\Gamma}^\gamma(s))\overline{\rho}^{s,t'}_{\theta,\gamma} + \int_{t'}^s \overline{R}^\gamma_\ell(s,s')\overline{\rho}^{s',t'}_{\theta,\gamma}\,\mathrm{d}s'\right)\mathrm{d}s\,, \tag{272}$$

$$\rho^{t,t'}_{\theta,\gamma} = I_m - \mathbb{I}(\lceil t'\rceil \le \lfloor t\rfloor)\int_{\lceil t'\rceil}^{\lfloor t\rfloor}\left((\nabla_\theta h_{\lfloor s\rfloor}(\theta^s_\gamma) + \Gamma^\gamma(s))\rho^{s,t'}_{\theta,\gamma} + \int_{\lceil t'\rceil}^{\lfloor s\rfloor} R^\gamma_\ell(s,s')\rho^{s',t'}_{\theta,\gamma}\,\mathrm{d}s'\right)\mathrm{d}s\,, \tag{273}$$

and set $\overline{R}^\gamma_\theta(t,t') = \mathbb{E}[\overline{\rho}^{t,t'}_{\theta,\gamma}]$ and $R^\gamma_\theta(t,t') = \mathbb{E}[\rho^{t,t'}_{\theta,\gamma}]$. We have

$$\begin{aligned}\|\overline{\rho}^{t,t'}_{\theta,\gamma} - \rho^{t,t'}_{\theta,\gamma}\|_2 &\le K\left(\gamma + \sup_{t\in[0,T]}\|\overline{\Gamma}^\gamma(t) - \Gamma^\gamma(t)\|_2 + \sup_{0\le t'\le t\le T}\|\overline{R}^\gamma_\ell(t,t') - R^\gamma_\ell(t,t')\|_2\right)\\ &\quad + K\int_{t'}^t \|\overline{\rho}^{s,t'}_{\theta,\gamma} - \rho^{s,t'}_{\theta,\gamma}\|_2\,\mathrm{d}s\\ &\le K(\gamma + \mathsf{dist}_\lambda(\overline{Y}^\gamma, Y^\gamma)) + K\int_{t'}^t \|\overline{\rho}^{s,t'}_{\theta,\gamma} - \rho^{s,t'}_{\theta,\gamma}\|_2\,\mathrm{d}s\,.\end{aligned} \tag{274}$$

By Grönwall's inequality, we have

$$\sup_{0\le t'\le t\le T}\|\overline{\rho}^{t,t'}_{\theta,\gamma} - \rho^{t,t'}_{\theta,\gamma}\|_2 \le K(\gamma + \mathsf{dist}_\lambda(\overline{Y}^\gamma, Y^\gamma))\,. \tag{275}$$

Thus, we have

$$\begin{aligned}\mathsf{dist}_\lambda(\overline{R}^\gamma_\theta, R^\gamma_\theta) \le \sup_{0\le t'\le t\le T}\|\overline{R}^\gamma_\theta(t,t') - R^\gamma_\theta(t,t')\|_2 &\le \sup_{0\le t'\le t\le T}\mathbb{E}\|\overline{\rho}^{t,t'}_{\theta,\gamma} - \rho^{t,t'}_{\theta,\gamma}\|_2\\ &\le K(\gamma + \mathsf{dist}_\lambda(\overline{Y}^\gamma, Y^\gamma))\,.\end{aligned} \tag{276}$$

By Equations (271) and (276), we have

$$\mathsf{dist}_\lambda(\overline{X}^\gamma, X^\gamma) \le K(\sqrt{\gamma} + \mathsf{dist}_\lambda(\overline{Y}^\gamma, Y^\gamma))\,. \tag{277}$$

D.3.2. PROOF OF LEMMA D.7

Following the same calculation that led to Equation (249), we can show that there exists a coupling of the processes r^t and r^t_γ such that

$$\sup_{t\in[0,T]} \sqrt{\mathbb{E}\|r^t - r^t_\gamma\|_2^2} \le K(\sqrt{\gamma} + \mathsf{dist}_\lambda(X, X^\gamma))\,. \tag{278}$$

Following the same calculation that led to Equation (270), we can show that there exists a coupling of the processes θ^t and θ^t_γ such that

$$\sup_{t\in[0,T]} \sqrt{\mathbb{E}\|\theta^t - \theta^t_\gamma\|_2^2} \le K(\sqrt{\gamma} + \mathsf{dist}_\lambda(Y, Y^\gamma))\,. \tag{279}$$

Appendix E. Details of Applications and Special Cases

E.1. Planted Models

E.1.1. SETUP FOR PLANTED MODELS

The setting of supervised learning with (noisy) target labels $y_i = f_*(\boldsymbol{\theta}^{*\mathsf{T}}\boldsymbol{x}_i) + z_i$ for some target parameter $\boldsymbol{\theta}^* \in \mathbb{R}^{d\times m}$ and target function $f_*\colon \mathbb{R}^m \to \mathbb{R}$ can be handled within our framework. Consider the following stochastic process with a planted signal $\boldsymbol{\theta}^* \in \mathbb{R}^{d\times m}$:

$$\hat{\boldsymbol{\theta}}^{k+1} = \hat{\boldsymbol{\theta}}^k - \eta \cdot \left(\frac{1}{d} h_{t_k}(\hat{\boldsymbol{\theta}}^k) + \frac{1}{B}\sum_{i\in\mathcal{B}^k} \boldsymbol{x}_i \ell_{t_k}(\hat{r}_i^k, r_i^*; z_i)^\mathsf{T} \right), \tag{280}$$

with $\hat{\boldsymbol{r}}^k = \boldsymbol{X}\hat{\boldsymbol{\theta}}^k$ and $\boldsymbol{r}^* = \boldsymbol{X}\boldsymbol{\theta}^*$. We consider the SGF with the planted signal that approximates the above SGD dynamics:

$$\mathrm{d}\boldsymbol{\theta}^t = -\left(h_t(\boldsymbol{\theta}^t) + \frac{1}{\delta}\boldsymbol{X}^\mathsf{T}\ell_t(\boldsymbol{r}^t, \boldsymbol{r}^*; \boldsymbol{z}) \right)\mathrm{d}t + \sqrt{\frac{\tau}{\delta}} \sum_{i=1}^n \boldsymbol{x}_i \ell_t(r_i^t, r_i^*; z_i)^\mathsf{T}\,\mathrm{d}B_i^t\,, \tag{281}$$

with $\boldsymbol{r}^t = \boldsymbol{X}\boldsymbol{\theta}^t$. This can be mapped to our general SGF setup (1) by concatenating the parameters $\boldsymbol{\theta}^t$ and $\boldsymbol{\theta}^*$ as follows.

$$\begin{aligned}\mathrm{d}(\boldsymbol{\theta}^t, \boldsymbol{\theta}^*) = &-\Big(\big(h_t(\boldsymbol{\theta}^t), 0\big) + \frac{1}{\delta}\boldsymbol{X}^\mathsf{T}\big(\ell_t(\boldsymbol{X}(\boldsymbol{\theta}^t, \boldsymbol{\theta}^*); \boldsymbol{z}), 0\big) \Big)\,\mathrm{d}t \\ &+ \sqrt{\frac{\tau}{\delta}} \sum_{i=1}^n \boldsymbol{x}_i \Big(\ell_t((\boldsymbol{\theta}^t, \boldsymbol{\theta}^*)^\mathsf{T}\boldsymbol{x}_i; z_i)^\mathsf{T}, 0 \Big)\,\mathrm{d}B_i^t\,.\end{aligned} \tag{282}$$

Here, $\boldsymbol{\theta}^*$ is constant over time. By applying Theorem 3.1 and Theorem 3.2 to the concatenated parameter of shape $\mathbb{R}^{d\times 2m}$, we obtain a DMFT characterization of the SGF with a planted signal (281) as a corollary.

Corollary E.1 (DMFT characterization of SGF with a planted signal) *Suppose Assumptions A.1 and A.2 hold. Take $T_* > 0$ in Theorem 3.1. Furthermore, assume that $\boldsymbol{\theta}^* \in \mathbb{R}^{d\times m}$ is independent of*

$\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{\theta}^0$ and for all $p \geq 1$, its empirical distribution converge in p-Wasserstein distance to $\mathsf{P}(\theta^*)$ almost surely as $d \to \infty$. Then, for any $T \in [0, T_*]$, $L \in \mathbb{N}$, and $0 \leq t_1 < \cdots < t_L \leq T$, we have

$$\operatorname*{p-lim}_{n,d\to\infty} W_2\Big(\hat{\mathsf{P}}(\boldsymbol{\theta}^{t_1}, \ldots, \boldsymbol{\theta}^{t_L}, \boldsymbol{\theta}^*), \mathsf{P}(\theta^{t_1}, \ldots, \theta^{t_L}, \theta^*)\Big)^2 = 0\,, \tag{283}$$

$$\operatorname*{p-lim}_{n,d\to\infty} W_2\Big(\hat{\mathsf{P}}(\boldsymbol{r}^{t_1}, \ldots, \boldsymbol{r}^{t_L}, \boldsymbol{r}^*, \boldsymbol{z}), \mathsf{P}(r^{t_1}, \ldots, r^{t_L}, r^*, z)\Big)^2 = 0\,, \tag{284}$$

where $(\theta^t, \theta^*, r^t, r^*, z)$ is the unique solution of the DMFT equation given in Appendix E.1.2.

The proof appears in Appendix E.1.3. This result is an extension of Celentano et al. [15, Corollary 4.1] to the SGF dynamics.

E.1.2. THE DMFT EQUATION

We state the DMFT equation $\mathfrak{S}^*$ for planted model (281) for functions $R_\ell, C_\theta \colon (\mathbb{R}_{\geq 0} \cup \{*\})^2 \to \mathbb{R}^{m\times m}$, $R_\theta, \Sigma_\ell \colon \mathbb{R}^2_{\geq 0} \to \mathbb{R}^{m\times m}$, and $\Gamma \colon \mathbb{R}_{\geq 0} \to \mathbb{R}^{m\times m}$ self-consistently as follows. First, given $\Sigma_\ell, R_\ell, \Gamma$, define stochastic processes $\{\theta^t \in \mathbb{R}^m\}_{t\geq 0}$ and $\{\rho_\theta^{t,t'} \in \mathbb{R}^{m\times m}\}_{t\geq t'\geq 0}$ by the following equations.

$$\theta^t = \theta^0 + U^t - \int_0^t \Big((h_s(\theta^s) + \Gamma(s)\theta^s) + \int_0^s R_\ell(s,s')\theta^{s'}\,\mathrm{d}s' + R_\ell(t,*)\theta^* \Big)\,\mathrm{d}s\,, \tag{285}$$

$$\rho_\theta^{t,t'} = I_m - \int_{t'}^t \Big((\nabla_\theta h_s(\theta^s) + \Gamma(s))\rho_\theta^{s,t'} + \int_{t'}^s R_\ell(s,s')\rho_\theta^{s',t'}\,\mathrm{d}s' \Big)\,\mathrm{d}s\,, \tag{286}$$

where $U \sim \mathsf{GP}(0, \Sigma_\ell/\delta)$. Then, set C_θ, R_θ as

$$C_\theta(t,t') = \mathbb{E}[\theta^t \theta^{t'\mathsf{T}}] \quad (s,t \in [0,T] \cup \{*\})\,, \quad R_\theta(t,t') = \mathbb{E}[\rho_\theta^{t,t'}] \quad (t \geq t')\,, \tag{287}$$

and $R_\theta(t,t') = 0$ for $t < t'$.

Next, given C_θ, R_θ, define stochastic processes $\{r^t \in \mathbb{R}^m\}_{t\geq 0}$, $r^* \in \mathbb{R}^m$, and $\{\rho_\ell^{t,t'}, \rho_{\ell*}^{t,t'}, D_\ell^{t,t'}, D_{\ell*}^{t,t'} \in \mathbb{R}^{m\times m}\}_{t\geq t'\geq 0}$ by the following equations.

$$r^t = w^t - \frac{1}{\delta}\int_0^t R_\theta(t,s)\ell_s(r^s, r^*; z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\,, \quad r^* = w^*\,, \qquad w \sim \mathsf{GP}(0, C_\theta)\,, \tag{288}$$

$$\rho_\ell^{t,t'} = \nabla_r \ell_t(r^t, r^*; z)\rho_r^{t,t'}\,, \tag{289}$$

$$\rho_{\ell*}^{t,t'} = \nabla_r \ell_t(r^t, r^*; z)\rho_{r*}^{t,t'}\,, \tag{290}$$

$$D_\ell^{t,t'} = \nabla_r \ell_t(r^t, r^*; z)\Big(-\frac{1}{\delta}\int_{t'}^t R_\theta(t,s)D_\ell^{s,t'}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\Big) + \nabla_r^2 \ell_t(r^t, r^*; z)[D_r^{t,t'}]\rho_r^{t,t'}\,, \tag{291}$$

$$D_{\ell*}^{t,t'} = \nabla_r \ell_t(r^t, r^*; z)\Big(-\frac{1}{\delta}\int_{t'}^t R_\theta(t,s)D_{\ell*}^{s,t'}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\Big) + \nabla_r^2 \ell_t(r^t, r^*; z)[D_r^{t,t'}]\rho_{r*}^{t,t'}\,, \tag{292}$$

where B^t is a Brownian motion in $\mathbb{R}$, and we defined the auxiliary processes $\rho_r^{t,t'}$, $\rho_{r*}^{t,t'} \in \mathbb{R}^{m\times m}$, and $D_r^{t,t'} \in \mathbb{R}^m$ as

$$\rho_r^{t,t'} = -\frac{1}{\delta}\int_{t'}^{t} R_\theta(t,s)\rho_\ell^{s,t'}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) - \frac{1}{\delta}R_\theta(t,t')\nabla_r \ell_{t'}(r^{t'}, r^*; z)\,, \tag{293}$$

$$\rho_{r*}^{t,t'} = -\frac{1}{\delta}\int_{t'}^{t} R_\theta(t,s)\rho_{\ell*}^{s,t'}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) - \frac{1}{\delta}R_\theta(t,t')\nabla_{r^*} \ell_{t'}(r^{t'}, r^*; z)\,, \tag{294}$$

$$D_r^{t,t'} = -\frac{1}{\delta}\int_{t'}^{t} R_\theta(t,s)\nabla_r \ell_s(r^s, r^*; z) D_r^{s,t'}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) - \sqrt{\frac{\tau}{\delta}}R_\theta(t,t')\ell_{t'}(r^{t'}, r^*; z)\,. \tag{295}$$

Then, set $\Sigma_\ell, R_\ell, \Gamma$ as

$$\Sigma_\ell(t,t') = \mathbb{E}[L^t L^{t'\mathsf{T}}]\,, \quad L^t := \int_0^t \ell_s(r^s, r^*; z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\,, \tag{296}$$

$$R_\ell(t,t') = \mathbb{E}[\rho_\ell^{t,t'}] + \sqrt{\tau\delta}\,\mathbb{E}[D_\ell^{t,t'}] \quad (t \geq t')\,, \tag{297}$$

$$R_\ell(t,*) = \mathbb{E}[\nabla_{r^*}\ell_t(r^t, r^*; z)] + \int_0^t \mathbb{E}[\rho_{\ell*}^{t,t'} + \sqrt{\tau\delta} D_{\ell*}^{t,t'}]\,\mathrm{d}t' \quad (t \geq 0)\,, \tag{298}$$

$$\Gamma(t) = \mathbb{E}[\nabla_r \ell_t(r^t, r^*; z)]\,, \tag{299}$$

and $R_\ell(t,t') = 0$ for $t < t'$.

Then, the solution of the DMFT system $\mathfrak{S}^*$ is defined as a fixed point of the above two mappings. In the informal notation, the DMFT equation for the planted model can be written as

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\theta^t &= u^t - (h_t(\theta^t) + \Gamma(t)\theta^t) - \int_0^t R_\ell(t,s)\theta^s\,\mathrm{d}s - R_\ell(t,*)\theta^*\,, \quad u \sim \mathsf{GP}(0, C_\ell/\delta)\,,\\
r^t &= w^t - \frac{1}{\delta}\int_0^t R_\theta(t,s)\ell_s(r^s, r^*; z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\,, \quad r^* = w^*\,, \quad w \sim \mathsf{GP}(0, C_\theta)\,,\\
C_\theta(t,t') &= \mathbb{E}[\theta^t\theta^{t'\mathsf{T}}] \quad (t,t' \in [0,T]\cup\{*\})\,, \quad R_\theta(t,t') = \mathbb{E}\left[\frac{\partial\theta^t}{\partial u^{t'}}\right]\,,\\
C_\ell(t,t') &= \mathbb{E}[\ell_t(r^t, r^*; z)(1 + \sqrt{\tau\delta}\dot{B}^t)\ell_{t'}(r^{t'}, r^*; z)^\mathsf{T}(1 + \sqrt{\tau\delta}\dot{B}^{t'})]\,,\\
R_\ell(t,t') &= \mathbb{E}\left[\frac{\partial \ell_t(r^t, r^*; z)}{\partial w^{t'}}\right]\,, \quad R_\ell(t,*) = \mathbb{E}\left[\frac{\partial \ell_t(r^t, r^*; z)}{\partial w^*}\right]\,, \quad \Gamma(t) = \mathbb{E}[\nabla_r \ell_t(r^t, r^*; z)]\,.
\end{aligned} \tag{300}$$

E.1.3. PROOF OF COROLLARY E.1

We transform the DMFT equation $\mathfrak{S}$ applied to the planted model (281) to the DMFT equation $\mathfrak{S}^*$ defined above. We distinguish the variables in $\mathfrak{S}$ by adding a bar over them, e.g., $\overline{\theta}^t$. The variables in $\mathfrak{S}$ have dimensions $2m$. Identifying components that are trivially zero, we see that the solution to $\mathfrak{S}$

is of the form

$$\overline{\theta}^t = \begin{pmatrix} \overline{\theta}_1^t \\ \theta^* \end{pmatrix}, \quad \overline{U}^t = \begin{pmatrix} \overline{U}_1^t \\ 0 \end{pmatrix}, \quad \overline{\rho}_\theta^{t,t'} = \begin{pmatrix} \overline{\rho}_{\theta,1}^{t,t'} & - \\ 0 & I_m \end{pmatrix},$$

$$\overline{C}_\theta(t,t') = \begin{pmatrix} \overline{C}_\theta^{11}(t,t') & \overline{C}_\theta^{12}(t,t') \\ \overline{C}_\theta^{12}(t,t')^\mathsf{T} & \overline{C}_\theta^{22}(t,t') \end{pmatrix}, \quad \overline{R}_\theta(t,t') = \begin{pmatrix} \overline{R}_\theta^1(t,t') & - \\ 0 & I_m \end{pmatrix},$$

$$\overline{r}^t = \begin{pmatrix} \overline{r}_1^t \\ \overline{r}_2^t \end{pmatrix}, \quad \overline{w}^t = \begin{pmatrix} \overline{w}_1^t \\ \overline{w}_2^t \end{pmatrix}, \quad \overline{\rho}_\ell^{t,t'} = \begin{pmatrix} \overline{\rho}_{\ell,1}^{t,t'} & \overline{\rho}_{\ell,2}^{t,t'} \\ 0 & 0 \end{pmatrix}, \quad \overline{D}_\ell^{t,t'} = \begin{pmatrix} \overline{D}_{\ell,1}^{t,t'} & \overline{D}_{\ell,2}^{t,t'} \\ 0 & 0 \end{pmatrix},$$

$$\overline{\rho}_r^{t,t'} = \begin{pmatrix} \overline{\rho}_{r,1}^{t,t'} & \overline{\rho}_{r,2}^{t,t'} \\ 0 & 0 \end{pmatrix}, \quad \overline{D}_r^{t,t'} = \begin{pmatrix} \overline{D}_{r,1}^{t,t'} \\ 0 \end{pmatrix},$$

$$\overline{\Sigma}_\ell(t,t') = \begin{pmatrix} \overline{\Sigma}_\ell^1(t,t') & 0 \\ 0 & 0 \end{pmatrix}, \quad \overline{R}_\ell(t,t') = \begin{pmatrix} \overline{R}_\ell^1(t,t') & \overline{R}_\ell^2(t,t') \\ 0 & 0 \end{pmatrix}, \quad \overline{\Gamma}(t) = \begin{pmatrix} \overline{\Gamma}_1(t) & \overline{\Gamma}_2(t) \\ 0 & 0 \end{pmatrix}. \tag{301}$$

Here, we indicated by $-$ the irrelevant variables. These variables satisfy

$$\begin{aligned} \overline{\theta}_1^t = \theta^0 + \overline{U}_1^t - \int_0^t \Big(h_s(\overline{\theta}_1^s) + \overline{\Gamma}_1(s)\overline{\theta}_1^s + \int_0^s \overline{R}_\ell^1(s,s')\overline{\theta}_1^{s'}\,\mathrm{d}s' \Big)\,\mathrm{d}s \\ - \int_0^t \Big(\overline{\Gamma}_2(s) + \int_0^s \overline{R}_\ell^2(s,s')\,\mathrm{d}s' \Big)\theta^*\,\mathrm{d}s\,, \quad \overline{U}_1 \sim \mathsf{GP}(0, \overline{\Sigma}_\ell^{11}/\delta)\,, \end{aligned} \tag{302}$$

$$\overline{\rho}_{\theta,1}^{t,t'} = I_m - \int_{t'}^t \Big((\nabla_\theta h_s(\overline{\theta}_1^s) + \overline{\Gamma}_1(s))\overline{\rho}_{\theta,1}^{s,t'} + \int_{t'}^s \overline{R}_\ell^1(s,s')\overline{\rho}_{\theta,1}^{s',t'}\,\mathrm{d}s' \Big)\,\mathrm{d}s\,, \tag{303}$$

$$\overline{C}_\theta^{11}(t,t') = \mathbb{E}[\overline{\theta}_1^t \overline{\theta}_1^{t'\mathsf{T}}]\,, \quad \overline{C}_\theta^{12}(t,t') = \mathbb{E}[\overline{\theta}_1^t \theta^{*\mathsf{T}}]\,, \quad \overline{C}_\theta^{22}(t,t') = \mathbb{E}[\theta^*\theta^{*\mathsf{T}}]\,, \quad \overline{R}_\theta^1(t,t') = \mathbb{E}[\overline{\rho}_{\theta,1}^{t,t'}]\,, \tag{304}$$

$$\overline{r}_1^t = \overline{w}_1^t - \frac{1}{\delta}\int_0^t \overline{R}_\theta^1(t,s)\ell_s(\overline{r}_1^s, \overline{r}_2^s; z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\,, \quad \overline{r}_2^t = \overline{w}_2^t\,, \quad \begin{pmatrix} \overline{w}_1 \\ \overline{w}_2 \end{pmatrix} \sim \mathsf{GP}(0, \overline{C}_\theta)\,, \tag{305}$$

$$\overline{\rho}_{\ell,1}^{t,t'} = \nabla_r \ell_t(\overline{r}_1^t, \overline{r}_2^t; z)\overline{\rho}_{r,1}^{t,t'}\,, \quad \overline{\rho}_{\ell,2}^{t,t'} = \nabla_r \ell_t(\overline{r}_1^t, \overline{r}_2^t; z)\overline{\rho}_{r,2}^{t,t'}\,, \tag{306}$$

$$\overline{D}_{\ell,1}^{t,t'} = \nabla_r \ell_t(\overline{r}_1^t, \overline{r}_2^t; z)\Big(-\frac{1}{\delta}\int_{t'}^t \overline{R}_\theta^1(t,s)\overline{D}_{\ell,1}^{s,t'}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) \Big) + \nabla_{rr}^2 \ell_t(\overline{r}_1^t, \overline{r}_2^t; z)[\overline{D}_{r,1}^{t,t'}]\overline{\rho}_{r,1}^{t,t'}\,, \tag{307}$$

$$\overline{D}_{\ell,2}^{t,t'} = \nabla_r \ell_t(\overline{r}_1^t, \overline{r}_2^t; z)\Big(-\frac{1}{\delta}\int_{t'}^t \overline{R}_\theta^1(t,s)\overline{D}_{\ell,2}^{s,t'}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) \Big) + \nabla_{rr}^2 \ell_t(\overline{r}_1^t, \overline{r}_2^t; z)[\overline{D}_{r,1}^{t,t'}]\overline{\rho}_{r,2}^{t,t'}\,, \tag{308}$$

$$\overline{\rho}_{r,1}^{t,t'} = -\frac{1}{\delta}\int_{t'}^t \overline{R}_\theta^1(t,s)\overline{\rho}_{\ell,1}^{s,t'}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) - \frac{1}{\delta}\overline{R}_\theta^1(t,t')\nabla_r \ell_{t'}(\overline{r}_1^{t'}, \overline{r}_2^{t'}; z)\,, \tag{309}$$

$$\overline{\rho}_{r,2}^{t,t'} = -\frac{1}{\delta}\int_{t'}^t \overline{R}_\theta^1(t,s)\overline{\rho}_{\ell,2}^{s,t'}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) - \frac{1}{\delta}\overline{R}_\theta^1(t,t')\nabla_w \ell_{t'}(\overline{r}_1^{t'}, \overline{r}_2^{t'}; z)\,, \tag{310}$$

$$\overline{D}_{r,1}^{t,t'} = -\frac{1}{\delta}\int_{t'}^t \overline{R}_\theta^1(t,s)\nabla_r \ell_s(\overline{r}_1^s, \overline{r}_2^s; z)\overline{D}_{r,1}^{s,t'}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) - \sqrt{\frac{\tau}{\delta}}\overline{R}_\theta^1(t,t')\ell_{t'}(\overline{r}_1^{t'}, \overline{r}_2^{t'}; z)\,, \tag{311}$$

$$\overline{\Sigma}^1_\ell(t,t') = \mathbb{E}[\overline{L}^t_1 \overline{L}^{t'\mathsf{T}}_1]\,, \quad \overline{L}^t_1 := \int_0^t \ell_s(\overline{r}^s_1, \overline{r}^s_2; z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\,, \tag{312}$$

$$\overline{R}^1_\ell(t,t') = \mathbb{E}[\overline{\rho}^{t,t'}_{\ell,1}] + \sqrt{\tau\delta}\,\mathbb{E}[\overline{D}^{t,t'}_{\ell,1}]\,, \quad \overline{R}^2_\ell(t,t') = \mathbb{E}[\overline{\rho}^{t,t'}_{\ell,2}] + \sqrt{\tau\delta}\,\mathbb{E}[\overline{D}^{t,t'}_{\ell,2}]\,, \tag{313}$$

$$\overline{\Gamma}_1(t) = \mathbb{E}[\nabla_r \ell_t(\overline{r}^t_1, \overline{r}^t_2; z)]\,, \quad \overline{\Gamma}_2(t) = \mathbb{E}[\nabla_{r^*} \ell_t(\overline{r}^t_1, \overline{r}^t_2; z)]\,. \tag{314}$$

Since $\overline{C}^{22}_\theta(t,t')$ is constant, $\overline{r}^t_2 = \overline{w}^t_2$ is constant. Then, $\mathfrak{S}$ reduces to $\mathfrak{S}^*$ by identifying

$$\begin{gathered}
\overline{\theta}^t = \theta^t\,, \quad \overline{U}^t_1 = U^t\,, \quad \overline{\rho}^{t,t'}_{\theta,1} = \rho^{t,t'}_\theta\,, \\
\overline{C}^{11}_\theta(t,t') = C_\theta(t,t')\,, \quad \overline{C}^{12}_\theta(t,t') = C_\theta(t,*)\,, \quad \overline{C}^{22}_\theta(t,t') = C_\theta(*,*)\,, \quad \overline{R}^1_\theta(t,t') = R_\theta(t,t')\,, \\
\overline{r}^t_1 = r^t\,, \quad \overline{w}^t_1 = w^t\,, \quad \overline{r}^t_2 = \overline{w}^t_2 = w^*\,, \quad \overline{\rho}^{t,t'}_{\ell,1} = \rho^{t,t'}_\ell\,, \overline{\rho}^{t,t'}_{\ell,2} = \rho^{t,t'}_{\ell*}\,, \quad \overline{D}^{t,t'}_{\ell,1} = D^{t,t'}_\ell\,, \overline{D}^{t,t'}_{\ell,2} = D^{t,t'}_{\ell*}\,, \\
\overline{\rho}^{t,t'}_{r,1} = \rho^{t,t'}_r\,, \quad \overline{\rho}^{t,t'}_{r,2} = \rho^{t,t'}_{r*}\,, \quad \overline{D}^{t,t'}_{r,1} = D^{t,t'}_r\,, \\
\overline{\Sigma}^1_\ell(t,t') = \Sigma_\ell(t,t')\,, \quad \overline{R}^1_\ell(t,t') = R_\ell(t,t')\,, \quad \overline{\Gamma}_1(t) = \Gamma(t)\,, \quad \overline{\Gamma}_2(t) + \int_0^t \overline{R}^2_\ell(t,t')\,\mathrm{d}t' = R_\ell(t,*)\,.
\end{gathered} \tag{315}$$

E.2. Infinite Data Limit

E.2.1. DERIVATION OF THE REDUCED DMFT EQUATION

We show that in the infinite data limit $\delta \to \infty$, the DMFT equations reduce to the following simple form.

$$\begin{gathered}
\mathrm{d}\theta^t = -(h_t(\theta^t) + \Gamma(t)\theta^t)\,\mathrm{d}t + \sqrt{\tau C_\ell(t)}\,\mathrm{d}W^t\,, \quad C_\theta(t,t') = \mathbb{E}[\theta^t \theta^{t'\mathsf{T}}]\,, \\
C_\ell(t) = \mathbb{E}[\ell_t(w^t; z)\ell_t(w^t; z)^\mathsf{T}]\,, \quad \Gamma(t) = \mathbb{E}[\nabla_r \ell_t(w^t; z)]\,, \quad w \sim \mathsf{GP}(0, C_\theta)\,,
\end{gathered} \tag{316}$$

where W^t is a Brownian motion in $\mathbb{R}^m$.

It is easy to see that as $\delta \to \infty$, $R_\ell(t,t') \to 0$ and hence $R_\theta(t,t') \to I_m$. Then, the equations for θ^t and r^t reduce to

$$\theta^t = \theta^0 + U^t - \int_0^t (h_s(\theta^s) + \Gamma(s)\theta^s)\,\mathrm{d}s\,, \quad U \sim \mathsf{GP}(0, \widetilde{\Sigma}_\ell)\,, \quad r^t = w^t\,, \quad w \sim \mathsf{GP}(0, C_\theta)\,, \tag{317}$$

where we introduced the rescaled variable $\widetilde{\Sigma}_\ell := \Sigma_\ell/\delta$. In addition, the equation for $\widetilde{\Sigma}_\ell$ reduces to

$$\widetilde{\Sigma}_\ell(t,t') = \tau \int_0^{t\wedge t'} \mathbb{E}[\ell_s(w^s; z)\ell_s(w^s; z)^\mathsf{T}]\,\mathrm{d}s\,. \tag{318}$$

Defining $C_\ell(t) = \mathbb{E}[\ell_t(w^t; z)\ell_t(w^t; z)^\mathsf{T}]$, $\widetilde{\Sigma}_\ell$ is a covariance kernel of an integrated Brownian motion:

$$U^t = \int_0^t \sqrt{\tau C_\ell(s)}\,\mathrm{d}W^s\,, \tag{319}$$

where W^t is a standard Brownian motion in $\mathbb{R}^m$. Thus, θ^t follows the following SDE:

$$\mathrm{d}\theta^t = -(h_t(\theta^t) + \Gamma(t)\theta^t)\,\mathrm{d}t + \sqrt{\tau C_\ell(t)}\,\mathrm{d}W^t\,. \tag{320}$$

These equations recover Equation (316).

For planted models of the form in Equation (281), the DMFT equation in the infinite data limit (316) becomes

$$\begin{gathered}\mathrm{d}\theta^t = -\big(h_t(\theta^t) + \Gamma(t)\theta^t + \Gamma^*(t)\theta^*\big)\,\mathrm{d}t + \sqrt{\tau C_\ell(t)}\,\mathrm{d}W^t\,,\\ C_\theta(t,t') = \mathbb{E}[\theta^t\theta^{t'\mathsf{T}}] \quad (t,t' \in [0,T]\cup\{*\})\,, \quad C_\ell(t) = \mathbb{E}[\ell_t(w^t,w^*;z)\ell_t(w^t,w^*;z)^\mathsf{T}]\,,\\ \Gamma(t) = \mathbb{E}[\nabla_w \ell_t(w^t,w^*;z)]\,, \quad \Gamma^*(t) = \mathbb{E}[\nabla_{w^*}\ell_t(w^t,w^*;z)]\,, \quad w \sim \mathsf{GP}(0,C_\theta)\,.\end{gathered} \tag{321}$$

E.2.2. EXAMPLE: LINEAR REGRESSION

Consider the linear regression setting described in Section 4. This corresponds to the choice $m = 1$, $h_t = 0$, and $\ell_t(r, r^*; z) = r - r^* - z$. Then, Equation (321) reduces to

$$\begin{gathered}\mathrm{d}\theta^t = -(\theta^t - \theta^*)\,\mathrm{d}t + \sqrt{\tau C_\ell(t)}\,\mathrm{d}W^t\,,\\ C_\theta(t,t') = \mathbb{E}[\theta^t\theta^{t'}] \quad (t,t' \in [0,T]\cup\{*\})\,, \quad C_\ell(t) = \mathbb{E}[(w^t - w^* - z)^2]\,, \quad w \sim \mathsf{GP}(0,C_\theta)\,.\end{gathered} \tag{322}$$

Using the Fokker–Planck equation, the density of θ^t given by the above SDE is given by the following partial differential equation (PDE):

$$\partial_t \mu(t,\theta) = \partial_\theta((\theta - \theta^*)\mu(t,\theta)) + \frac{\tau C_\ell(t)}{2}\partial_\theta^2 \mu(t,\theta)\,. \tag{323}$$

This equation coincides with the PDE derived in Wang et al. [56].

Furthermore, the training dynamics for test errors $\mathscr{L}(t) = \mathbb{E}[(\theta - \theta^*)^2] + \sigma^2$ can be obtained in closed form, where $\sigma^2 := \mathbb{E}[z^2]$. Let $\rho^2 := \mathbb{E}[(\theta^*)^2]$. Then, we have $C_\ell(t) = C_\theta(t,t) - 2C_\theta(t,*) + \rho^2 + \sigma^2$. By using Itô's lemma on $C_\theta(t,*) = \mathbb{E}[\theta^t\theta^*]$ and $C_\theta(t,t) = \mathbb{E}[(\theta^t)^2]$, we obtain the following system of ODEs for $C_\theta(t,*)$ and $C_\theta(t,t)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}C_\theta(t,*) = -C_\theta(t,*) + \rho^2\,, \tag{324}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}C_\theta(t,t) = -2C_\theta(t,t) + 2C_\theta(t,*) + \tau(C_\theta(t,t) - 2C_\theta(t,*) + \rho^2 + \sigma^2)\,. \tag{325}$$

These are linear ODEs and can be solved in closed form as follows (assuming zero initialization, i.e., $\theta^0 = C_\theta(0,0) = C_\theta(0,*) = 0$):

$$C_\theta(t,*) = \rho^2(1 - \mathrm{e}^{-t})\,, \quad C_\theta(t,t) = \rho^2(1 - 2\mathrm{e}^{-t} + \mathrm{e}^{-(2-\tau)t}) + \frac{\tau\sigma^2}{2-\tau}(1 - \mathrm{e}^{-(2-\tau)t})\,. \tag{326}$$

E.3. Linear Regression

In this section, we derive the DMFT equation for the SGF for the ridge regression (regularized linear regression) problem. The loss for the ridge regression problem is defined as

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|_2^2 + \frac{\lambda}{2d}\|\boldsymbol{\theta}\|_2^2\,, \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{z}\,, \tag{327}$$

with regularization parameter $\lambda \geq 0$. The SGF for this problem is given by

$$\mathrm{d}\boldsymbol{\theta}^t = -\left(\lambda\boldsymbol{\theta}^t + \frac{1}{\delta}\boldsymbol{X}^\mathsf{T}(\boldsymbol{r}^t - \boldsymbol{r}^* - \boldsymbol{z})\right)\mathrm{d}t + \sqrt{\frac{\tau}{\delta}}\sum_{i=1}^n \boldsymbol{x}_i(r_i^t - r_i^* - z_i)\,\mathrm{d}B_i^t\,. \tag{328}$$

where $\boldsymbol{r}^t = \boldsymbol{X}\boldsymbol{\theta}^t$ and $\boldsymbol{r}^* = \boldsymbol{X}\boldsymbol{\theta}^*$. For simplicity, we consider the case of zero initialization $\boldsymbol{\theta}^0 = 0$. This is a special case of the SGF for planted models (281) with $m = 1$, $h_t(\theta) = \lambda\theta$, and $\ell_t(r, r^*; z) = r - r^* - z$.

We are particularly interested in the training and test errors of the parameter $\boldsymbol{\theta}^t$ given by

$$\mathscr{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta} - y_i)^2\,, \quad \mathscr{R}(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{x},y)}[(\boldsymbol{x}^\mathsf{T}\boldsymbol{\theta} - y)^2] = \frac{1}{d}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 + \sigma^2\,. \tag{329}$$

Applying Corollary E.1, we obtain a DMFT characterization of the SGF for ridge regression (328). Note that since we have $\nabla_r^2 \ell_t(r, r^*; z) = 0$, the DMFT equation is guaranteed to have a unique solution for all $T > 0$ by Theorem 3.1. Solving the DMFT equation, we can characterize the training and test errors of SGD for linear regression by a simple set of linear Volterra equations.

Proposition E.2 (DMFT characterization of SGF for ridge regression) *Assume that the noise $\boldsymbol{z} \in \mathbb{R}^n$ and the target parameter $\boldsymbol{\theta}^* \in \mathbb{R}^d$ satisfy the same assumptions as in Corollary E.1. Let $\rho^2 \coloneqq \mathbb{E}[(\theta^*)^2]$ and $\sigma^2 \coloneqq \mathbb{E}[z^2]$. Let μ_{MP} be the Marchenko–Pastur distribution with parameter δ given by*

$$\mu_{\mathrm{MP}}(x) = \frac{\delta\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{2\pi x} + (1-\delta)\delta(x)\mathbb{I}(\delta < 1)\,, \quad x \in [\lambda_-, \lambda_+]\,, \tag{330}$$

where $\lambda_\pm = (1 \pm 1/\sqrt{\delta})^2$ and δ is the Dirac delta function (an upright δ is used to distinguish it from $\delta = n/d$).

Then, for any $0 \le t_1, \dots, t_L < \infty$, we have

$$\operatorname*{p-lim}_{n,d\to\infty} \max_{l=1,\dots,L} |\mathscr{L}(\boldsymbol{\theta}^{t_l}) - \mathscr{L}(t_l)| = 0\,, \quad \operatorname*{p-lim}_{n,d\to\infty} \max_{l=1,\dots,L} |\mathscr{R}(\boldsymbol{\theta}^{t_l}) - \mathscr{R}(t_l)| = 0\,, \tag{331}$$

where $\mathscr{L}(t)$ and $\mathscr{R}(t)$ solve the following system of linear Volterra equations:

$$\mathscr{L}(t) = \mathscr{L}_0(t) + \tau\int_0^t H_2(t-s)\mathscr{L}(s)\,\mathrm{d}s\,, \quad \mathscr{R}(t) = \mathscr{R}_0(t) + \tau\int_0^t H_1(t-s)\mathscr{L}(s)\,\mathrm{d}s\,, \tag{332}$$

where $H_i(t) \coloneqq \int x^i \mathrm{e}^{-2(x+\lambda)t}\,\mathrm{d}\mu_{\mathrm{MP}}(x)$, and $\mathscr{L}_0(t)$ and $\mathscr{R}_0(t)$ are the asymptotic train and test errors for the noiseless case $\tau = 0$, which are given by

$$\mathscr{L}_0(t) = \rho^2\int \frac{x(\lambda + x\mathrm{e}^{-(x+\lambda)t})^2}{(x+\lambda)^2}\,\mathrm{d}\mu_{\mathrm{MP}}(x) + \frac{\sigma^2}{\delta}\int \frac{(\lambda + x\mathrm{e}^{-(x+\lambda)t})^2}{(x+\lambda)^2}\,\mathrm{d}\mu_{\mathrm{MP}}(x) + \frac{\delta-1}{\delta}\sigma^2\,, \tag{333}$$

$$\mathscr{R}_0(t) = \rho^2\int \frac{(\lambda + x\mathrm{e}^{-(x+\lambda)t})^2}{(x+\lambda)^2}\,\mathrm{d}\mu_{\mathrm{MP}}(x) + \frac{\sigma^2}{\delta}\int \frac{x}{(x+\lambda)^2}(1 - \mathrm{e}^{-(x+\lambda)t})^2\,\mathrm{d}\mu_{\mathrm{MP}}(x) + \sigma^2\,. \tag{334}$$

Setting $\lambda = 0$, we obtain the DMFT characterization in the main text (Section 4).

The equation for $\mathscr{L}$ in (332) is a scalar linear Volterra integral equation, which can be solved numerically efficiently. Once we obtain $\mathscr{L}$, we can compute $\mathscr{R}$ using the second equation in (332).

The equations (332) are equivalent to those derived in Paquette et al. [44, 45] using the theory of homogenized SGD and random matrix theory (up to time rescaling to match their settings). Our result provides an alternative derivation of these equations using the DMFT framework (although our framework does not directly apply to SGD rigorously). For further analysis of these equations, such as their exact solutions and long-time behaviors, see Paquette et al. [44, 45].

E.3.1. SIMPLIFYING THE DMFT EQUATIONS

We can derive the DMFT equation for ridge regression by specializing the DMFT equation $\mathfrak{S}^*$ for planted models in Appendix E.1.2. Since the loss is quadratic, i.e., $\partial_r \ell_t(r, r^*; z) = 1$, $\partial_{r^*} \ell_t(r, r^*; z) = -1$, and $\partial^2 \ell_t(r, r^*; z) = 0$, the DMFT equation simplifies significantly. We have $\Gamma(t) = 1$, $D_\ell^{t,t'} = D_{\ell*}^{t,t'} = 0$, $\rho_r^{t,t'} = -\rho_{r*}^{t,t'}$, and $\rho_\ell^{t,t'} = -\rho_{\ell*}^{t,t'}$. The DMFT equation for ridge regression reduces to

$$\theta^t = U^t - \int_0^t \left((1+\lambda)\theta^s - \theta^* + \int_0^s R_\ell(s,s')(\theta^{s'} - \theta^*)\,\mathrm{d}s' \right) \mathrm{d}s\,, \quad U \sim \mathsf{GP}(0, \Sigma_\ell/\delta)\,, \tag{335a}$$

$$\rho_\theta^{t,t'} = 1 - \int_{t'}^t \left((1+\lambda)\rho_\theta^{s,t'} + \int_{t'}^s R_\ell(s,s')\rho_\theta^{s',t'}\,\mathrm{d}s' \right) \mathrm{d}s\,, \tag{335b}$$

$$C_\theta(t,t') = \mathbb{E}[\theta^t \theta^{t'}] \quad (s,t \in [0,T] \cup \{*\})\,, \quad R_\theta(t,t') = \mathbb{E}[\rho_\theta^{t,t'}] \quad (t \geq t')\,, \tag{335c}$$

$$r^t = w^t - \frac{1}{\delta} \int_0^t R_\theta(t,s)(r^s - r^* - z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\,, \quad w \sim \mathsf{GP}(0, C_\theta)\,, \tag{335d}$$

$$\rho_\ell^{t,t'} = -\frac{1}{\delta} \int_{t'}^t R_\theta(t,s)\rho_\ell^{s,t'}(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s) - \frac{1}{\delta} R_\theta(t,t')\,, \tag{335e}$$

$$\Sigma_\ell(t,t') = \mathbb{E}[L^t L^{t'}]\,, \quad L^t := \int_0^t (r^s - r^* - z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\,, \quad R_\ell(t,t') = \mathbb{E}[\rho_\ell^{t,t'}] \quad (t \geq t')\,. \tag{335f}$$

Next, we eliminate the stochastic processes to close the system in terms of the correlation and response functions. To this end, we discretize time with step size γ as in Appendix B.2, reduce the equations, and then take the continuous-time limit $\gamma \to 0$. This operation can be justified along the lines of the proof of Theorem 3.2.

Let $t_i = i\gamma$ for $i = 0, 1, \ldots, \lfloor T/\gamma \rfloor$. The discretized version of Equation (335) is given by

$$\theta^{t_i} = \gamma \sum_{j=0}^{i-1} \left(u^{t_i} - (1+\lambda)\theta^{t_j} + \theta^* - \gamma \sum_{k=0}^{j-1} R_\ell(t_j, t_k)(\theta^{t_k} - \theta^*) \right)\,, \quad U \sim \mathsf{GP}(0, C_\ell/\delta)\,, \tag{336a}$$

$$C_\theta(t_i,t_j) = \mathbb{E}[\theta^{t_i}\theta^{t_j}]\,, \quad C_\theta(t_i,*) = \mathbb{E}[\theta^{t_i}\theta^*]\,, \quad R_\theta(t_i,t_j) = \gamma^{-1}\,\mathbb{E}\left[\frac{\partial \theta^{t_i}}{\partial u^{t_j}}\right] \quad (i > j)\,, \tag{336b}$$

$$r^{t_i} = w^{t_i} - \frac{1}{\delta} \sum_{j=0}^{i-1} R_\theta(t_i,t_j)(r^{t_j} - r^* - z)(\gamma + \sqrt{\tau\delta}(B^{t_{j+1}} - B^{t_j}))\,, \quad w \sim \mathsf{GP}(0, C_\theta)\,, \tag{336c}$$

$$C_\ell(t_i,t_j) = \mathbb{E}\left[(r^{t_i} - r^* - z)\left(1 + \sqrt{\tau\delta}\frac{B^{t_{i+1}} - B^{t_i}}{\gamma}\right)(r^{t_j} - r^* - z)\left(1 + \sqrt{\tau\delta}\frac{B^{t_{j+1}} - B^{t_j}}{\gamma}\right) \right]\,, \tag{336d}$$

$$R_\ell(t_i,t_j) = \gamma^{-1}\,\mathbb{E}\left[\frac{\partial r^{t_i}}{\partial w^{t_j}}\right] \quad (i > j)\,. \tag{336e}$$

Let $K = \lfloor T/\gamma \rfloor + 1$ and let $\boldsymbol{C}_\theta, \boldsymbol{R}_\theta, \boldsymbol{C}_\ell, \boldsymbol{R}_\ell \in \mathbb{R}^{K\times K}$ be the matrices with entries $(\boldsymbol{C}_\theta)_{ij} = C_\theta(t_i, t_j)$, $(\boldsymbol{R}_\theta)_{ij} = R_\theta(t_i, t_j)$, $(\boldsymbol{C}_\ell)_{ij} = C_\ell(t_i, t_j)$, and $(\boldsymbol{R}_\ell)_{ij} = R_\ell(t_i, t_j)$ for $i, j = 0, 1, \ldots, K-1$. Let $\boldsymbol{c}^* \in \mathbb{R}^K$ be the vector with entries $c_i^* = C_\theta(t_i, *)$ for $i = 0, 1, \ldots, K-1$. Let $\boldsymbol{\theta}, \boldsymbol{r}, \boldsymbol{u}, \boldsymbol{w}, \boldsymbol{G} \in \mathbb{R}^K$ be the vectors with entries $\theta^{t_i}, r^{t_i}, u^{t_i}, w^{t_i}, G^{t_i} := (B^{t_{i+1}} - B^{t_i})/\sqrt{\gamma}$ for $i = 0, 1, \ldots, K-1$.

Equation for R_θ. Differentiating Equation (336a) with respect to u^{t_j} for $j < i$ and taking the expectation, we have

$$R_\theta(t_i, t_j) = 1 - \gamma \sum_{k=j}^{i-1} \left((1+\lambda) R_\theta(t_k, t_j) + \gamma \sum_{l=0}^{k-1} R_\ell(t_k, t_l) R_\theta(t_l, t_j) \right). \tag{337}$$

Let $\boldsymbol{T} := \mathbb{R}^{K\times K}$ be the lower-triangular matrix with entries $T_{ij} = 1$ for $i > j$ and $T_{ij} = 0$ for $i \geq j$. Equation (337) can be written in matrix form as

$$\boldsymbol{R}_\theta = \boldsymbol{T} - \gamma \boldsymbol{T}((1+\lambda)\boldsymbol{R}_\theta + \gamma \boldsymbol{R}_\ell \boldsymbol{R}_\theta). \tag{338}$$

Solving for $\boldsymbol{R}_\theta$, we obtain

$$\boldsymbol{R}_\theta = (\boldsymbol{I} + \gamma \boldsymbol{T}((1+\lambda)\boldsymbol{I} + \gamma \boldsymbol{R}_\ell))^{-1} \boldsymbol{T}. \tag{339}$$

Equation for C_θ. Multiplying Equation (336a) by θ^* and taking the expectation, we have

$$C_\theta(t_i, *) = -\gamma \sum_{j=0}^{i-1} \left((1+\lambda) C_\theta(t_j, *) - \rho^2 + \gamma \sum_{k=0}^{j-1} R_\ell(t_j, t_k)(C_\theta(t_k, *) - \rho^2) \right). \tag{340}$$

In matrix form, this can be written as

$$\boldsymbol{c}^* = -\gamma \boldsymbol{T}((1+\lambda)\boldsymbol{c}^* - \rho^2 \boldsymbol{1} + \gamma \boldsymbol{R}_\ell(\boldsymbol{c}^* - \rho^2 \boldsymbol{1})). \tag{341}$$

Solving for $\boldsymbol{c}^*$, we obtain

$$\boldsymbol{c}^* = \rho^2 \gamma (\boldsymbol{I} + \gamma \boldsymbol{T}((1+\lambda)\boldsymbol{I} + \gamma \boldsymbol{R}_\ell))^{-1} \boldsymbol{T}(\boldsymbol{I} + \gamma \boldsymbol{R}_\ell)\boldsymbol{1} = \rho^2 \gamma \boldsymbol{R}_\theta (\boldsymbol{I} + \gamma \boldsymbol{R}_\ell)\boldsymbol{1}. \tag{342}$$

Multiplying Equation (336a) by θ^{t_j} and taking the expectation, we have

$$\begin{aligned} &C_\theta(t_i, t_j) \\ &= \gamma \sum_{k=0}^{i-1} \left(\mathbb{E}[u^{t_k} \theta^{t_j}] - (1+\lambda) C_\theta(t_k, t_j) + C_\theta(t_j, *) - \gamma \sum_{l=0}^{j-1} R_\ell(t_k, t_l)(C_\theta(t_l, t_j) - C_\theta(t_j, *)) \right). \end{aligned} \tag{343}$$

By Stein's lemma (the Gaussian integration by parts), we have

$$\mathbb{E}[u^{t_k} \theta^{t_j}] = \sum_{l=0}^{j-1} \mathrm{Cov}(u^{t_k}, u^{t_l}) \, \mathbb{E}\left[\frac{\partial \theta^{t_j}}{\partial u^{t_l}} \right] = \frac{\gamma}{\delta} \sum_{l=0}^{j-1} C_\ell(t_k, t_l) R_\theta(t_j, t_l). \tag{344}$$

Therefore, Equation (343) can be written in matrix form as

$$\boldsymbol{C}_\theta = \gamma \boldsymbol{T}\Big(\frac{\gamma}{\delta}\boldsymbol{C}_\ell \boldsymbol{R}_\theta^\mathsf{T} - (1+\lambda)\boldsymbol{C}_\theta + \mathbf{1}\boldsymbol{c}^{*\mathsf{T}} - \gamma \boldsymbol{R}_\ell(\boldsymbol{C}_\theta - \mathbf{1}\boldsymbol{c}^{*\mathsf{T}})\Big)\,. \tag{345}$$

Solving for $\boldsymbol{C}_\theta$, we obtain

$$\begin{aligned}\boldsymbol{C}_\theta &= \gamma(\boldsymbol{I} + \gamma\boldsymbol{T}((1+\lambda)\boldsymbol{I} + \gamma\boldsymbol{R}_\ell))^{-1}\boldsymbol{T}(\gamma\boldsymbol{C}_\ell\boldsymbol{R}_\theta^\mathsf{T} + (\boldsymbol{I}+\gamma\boldsymbol{R}_\ell)\mathbf{1}\boldsymbol{c}^{*\mathsf{T}}) \\ &= \frac{\gamma^2}{\delta}\boldsymbol{R}_\theta\boldsymbol{C}_\ell\boldsymbol{R}_\theta^\mathsf{T} + \frac{1}{\rho^2}\boldsymbol{c}^*\boldsymbol{c}^{*\mathsf{T}}\,. \end{aligned}\tag{346}$$

Let $C_\theta^*(t_i,t_j) := \mathbb{E}[(\theta^{t_i}-\theta^*)(\theta^{t_j}-\theta^*)] = C_\theta(t_i,t_j) - C_\theta(t_i,*) - C_\theta(t_j,*) + \rho^2$. Let $\boldsymbol{C}_\theta^* \in \mathbb{R}^{K\times K}$ be the matrix with entries $(\boldsymbol{C}_\theta^*)_{ij} = C_\theta^*(t_i,t_j)$ for $i,j = 0,1,\dots,K-1$. From Equation (346), we have

$$\boldsymbol{C}_\theta^* = \boldsymbol{C}_\theta - \boldsymbol{c}^*\mathbf{1}^\mathsf{T} - \mathbf{1}\boldsymbol{c}^{*\mathsf{T}} + \rho^2\mathbf{1}\mathbf{1}^\mathsf{T} = \frac{\gamma^2}{\delta}\boldsymbol{R}_\theta\boldsymbol{C}_\ell\boldsymbol{R}_\theta^\mathsf{T} + \frac{1}{\rho^2}(\boldsymbol{c}^* - \rho^2\mathbf{1})(\boldsymbol{c}^* - \rho^2\mathbf{1})^\mathsf{T}\,. \tag{347}$$

Equation for R_ℓ. Differentiating Equation (336c) with respect to w^{t_j} for $j < i$ and taking the expectation, we have

$$R_\ell(t_i,t_j) = -\frac{1}{\delta}\sum_{k=j}^{i-1} R_\theta(t_i,t_k)R_\ell(t_k,t_j) - \frac{1}{\delta}R_\theta(t_i,t_j)\,. \tag{348}$$

This can be written in matrix form as

$$\boldsymbol{R}_\ell = -\frac{1}{\delta}\boldsymbol{R}_\theta\boldsymbol{R}_\ell - \frac{1}{\delta}\boldsymbol{R}_\theta\,. \tag{349}$$

Solving for $\boldsymbol{R}_\ell$, we obtain

$$\boldsymbol{R}_\ell = -(\delta\boldsymbol{I} + \gamma\boldsymbol{R}_\theta)^{-1}\boldsymbol{R}_\theta\,. \tag{350}$$

Equation for C_ℓ. Since $G^{t_i} \sim \mathsf{N}(0,1)$ are i.i.d. standard normal variables, we have

$$\begin{aligned}C_\ell(t_i,t_j) &= \mathbb{E}[(r^{t_i}-r^*-z)(r^{t_j}-r^*-z)] + + \frac{\tau\delta}{\gamma}\,\mathbb{E}[(r^{t_i}-r^*-z)^2]\mathbb{I}(i=j) \\ &\quad + \sqrt{\frac{\tau\delta}{\gamma}}\big(\mathbb{E}[(r^{t_i}-r^*-z)(r_{t_j}-r^*-z)G^{t_j}] + \mathbb{E}[(r^{t_i}-r^*-z)(r^{t_j}-r^*-z)G^{t_i}]\big) \end{aligned}\tag{351}$$

$$= L(t_i,t_j) + \frac{\tau\delta}{\gamma}L(t_i,t_i)\mathbb{I}(i=j) + \sqrt{\tau\delta}(M(t_i,t_j) + M(t_j,t_i))\,, \tag{352}$$

where we defined

$$L(t_i,t_j) = \mathbb{E}[(r^{t_i}-r^*-z)(r^{t_j}-r^*-z)]\,, \tag{353}$$

$$M(t_i,t_j) = \gamma^{-1/2}\,\mathbb{E}[(r^{t_i}-r^*-z)(r^{t_j}-r^*-z)G^{t_j}] = \gamma^{-1/2}\,\mathbb{E}\left[\frac{\partial r^{t_i}}{\partial G^{t_j}}(r^{t_j}-r^*-z)\right]. \tag{354}$$

In the definition of $M(t_i, t_j)$, we used Stein's lemma.

We first derive a closed equation for $M(t_i, t_j)$. Differentiating Equation (336c) with respect to G^{t_j}, multiplying by $\gamma^{-1/2}(r^{t_j} - r^* - z)$, and taking the expectation, we have

$$M(t_i, t_j) = -\frac{\gamma}{\delta} \sum_{k=j}^{i-1} R_\theta(t_i, t_k) M(t_k, t_j) - \sqrt{\frac{\tau}{\delta}} R_\theta(t_i, t_j) L(t_j, t_j) \,. \tag{355}$$

In matrix form, this can be written as

$$\boldsymbol{M} = -\frac{\gamma}{\delta} \boldsymbol{R}_\theta \boldsymbol{M} - \sqrt{\frac{\tau}{\delta}} \boldsymbol{R}_\theta \operatorname{diag}(\boldsymbol{L}) \,, \tag{356}$$

where $\boldsymbol{M}, \boldsymbol{L} \in \mathbb{R}^{K \times K}$ are the matrices with entries $M_{ij} = M(t_i, t_j)$ and $L_{ij} = L(t_i, t_j)$, and $\operatorname{diag}(\boldsymbol{L})$ is the diagonal matrix with diagonal entries equal to those of $\boldsymbol{L}$. Solving for $\boldsymbol{M}$, we obtain

$$\boldsymbol{M} = -\sqrt{\tau\delta}(\delta \boldsymbol{I} + \gamma \boldsymbol{R}_\theta)^{-1} \boldsymbol{R}_\theta \operatorname{diag}(\boldsymbol{L}) = \sqrt{\tau\delta} \boldsymbol{R}_\ell \operatorname{diag}(\boldsymbol{L}) \,. \tag{357}$$

Next, we derive a closed equation for $L(t_i, t_j)$. Multiplying Equation (336c) by $r^{t_j} - r^* - z$ and taking the expectation, we have

$$L(t_i, t_j) = \mathbb{E}[(w^{t_i} - r^* - z)(r^{t_j} - r^* - z)] - \frac{\gamma}{\delta} \sum_{k=0}^{i-1} R_\theta(t_i, t_k)(L(t_k, t_j) + \sqrt{\tau\delta} M(t_j, t_k)) \,. \tag{358}$$

By Stein's lemma, we have

$$\begin{aligned}
&\mathbb{E}[(w^{t_i} - r^* - z)(r^{t_j} - r^* - z)] \\
&= \sum_{k=0}^{j} \operatorname{Cov}(w^{t_i} - r^* - z, w^{t_k}) \, \mathbb{E}\left[\frac{\partial r^{t_j}}{\partial w^{t_k}}\right] + \operatorname{Cov}(w^{t_i} - r^* - z, r^* + z)\left(\mathbb{E}\left[\frac{\partial r^{t_j}}{\partial (r^* + z)}\right] - 1\right) \\
&= \gamma \sum_{k=0}^{j-1} (C_\theta(t_i, t_k) - C_\theta(t_k, *)) R_\ell(t_j, t_k) + C_\theta(t_i, t_j) - C_\theta(t_j, *) \\
&\qquad - (C_\theta(t_i, *) - \rho^2 - \sigma^2)\left(\gamma \sum_{k=0}^{j-1} R_\ell(t_j, t_k) + 1\right) \\
&= C_\theta(t_i, t_j) - C_\theta(t_i, *) - C_\theta(t_j, *) + \rho^2 + \sigma^2 \\
&\qquad + \gamma \sum_{k=0}^{j-1} (C_\theta(t_i, t_k) - C_\theta(t_i, *) - C_\theta(t_k, *) + \rho^2 + \sigma^2) R_\ell(t_j, t_k) \\
&= C_\theta^*(t_i, t_j) + \sigma^2 + \gamma \sum_{k=0}^{j-1} (C_\theta^*(t_i, t_k) + \sigma^2) R_\ell(t_j, t_k) \,.
\end{aligned} \tag{359}$$

Therefore, Equation (358) can be written in matrix form as

$$\boldsymbol{L} = (\boldsymbol{C}_\theta^* + \sigma^2 \boldsymbol{1}\boldsymbol{1}^\mathsf{T})(\boldsymbol{I} + \gamma \boldsymbol{R}_\ell^\mathsf{T}) - \frac{\gamma}{\delta} \boldsymbol{R}_\theta (\boldsymbol{L} + \sqrt{\tau\delta} \boldsymbol{M}^\mathsf{T}) \,. \tag{360}$$

Solving for $\boldsymbol{L}$, we obtain

$$\begin{aligned}\boldsymbol{L} &= \delta(\delta\boldsymbol{I}+\gamma\boldsymbol{R}_\theta)^{-1}(\boldsymbol{C}_\theta^*+\sigma^2\boldsymbol{1}\boldsymbol{1}^\mathsf{T})(\boldsymbol{I}+\gamma\boldsymbol{R}_\ell^\mathsf{T}) - \sqrt{\tau\delta}\gamma(\delta\boldsymbol{I}+\gamma\boldsymbol{R}_\theta)^{-1}\boldsymbol{R}_\theta\boldsymbol{M}^\mathsf{T} \\ &= (\boldsymbol{I}+\gamma\boldsymbol{R}_\ell)(\boldsymbol{C}_\theta^*+\sigma^2\boldsymbol{1}\boldsymbol{1}^\mathsf{T})(\boldsymbol{I}+\gamma\boldsymbol{R}_\ell^\mathsf{T}) + \tau\delta\gamma\boldsymbol{R}_\ell\operatorname{diag}(\boldsymbol{L})\boldsymbol{R}_\ell^\mathsf{T}\,, \end{aligned} \tag{361}$$

where we used that

$$\begin{aligned}\boldsymbol{I}+\gamma\boldsymbol{R}_\ell = \boldsymbol{I}-\gamma(\delta\boldsymbol{I}+\gamma\boldsymbol{R}_\theta)^{-1}\boldsymbol{R}_\theta &= \boldsymbol{I}-(\delta\boldsymbol{I}+\gamma\boldsymbol{R}_\theta)^{-1}(\delta\boldsymbol{I}+\gamma\boldsymbol{R}_\theta-\delta\boldsymbol{I}) \\ &= \delta(\delta\boldsymbol{I}+\gamma\boldsymbol{R}_\theta)^{-1}\,. \end{aligned} \tag{362}$$

Continuous-time limit. Taking the continuous-time limit $\gamma\to 0$ in Equations (337), (342), and (347), we obtain

$$R_\theta(t,t') = 1-\int_{t'}^t\left((1+\lambda)R_\theta(s,t')+\int_{t'}^s R_\ell(s,s')R_\theta(s',t')\,\mathrm{d}s'\right)\mathrm{d}s\,, \tag{363}$$

$$C_\theta(t,*) = \rho^2\int_0^t R_\theta(t,s)\left(1+\int_0^s R_\ell(s,s')\,\mathrm{d}s'\right)\mathrm{d}s\,, \tag{364}$$

$$C_\theta^*(t,t') = \frac{1}{\delta}\int_0^t\int_0^{t'} R_\theta(t,s)R_\theta(t',s')C_\ell(s,s')\,\mathrm{d}s'\,\mathrm{d}s + \frac{1}{\rho^2}(C_\theta(t,*)-\rho^2)(C_\theta(t',*)-\rho^2)\,. \tag{365}$$

Taking the continuous-time limit $\gamma\to 0$ in Equations (348), (352), and (361), we obtain

$$R_\ell(t,t') = -\frac{1}{\delta}\int_{t'}^t R_\theta(t,s)R_\ell(s,t')\,\mathrm{d}s - \frac{1}{\delta}R_\theta(t,t')\,, \tag{366}$$

$$C_\ell(t,t') = L(t,t')+\tau\delta(R_\ell(t',t)L(t',t')+R_\ell(t,t')L(t,t)+L(t,t)\delta(t-t'))\,, \tag{367}$$

$$\begin{aligned}L(t,t') = &\int_0^t\int_0^{t'}(\delta(t-s)+R_\ell(t,s))(\delta(t'-s')+R_\ell(t',s'))(C_\theta^*(s,s')+\sigma^2)\,\mathrm{d}s'\,\mathrm{d}s \\ &+\tau\delta\int_0^{t\wedge t'} R_\ell(t,s)R_\ell(t',s)L(s,s)\,\mathrm{d}s\,. \end{aligned} \tag{368}$$

Here, $\delta(\cdot)$ is the Dirac delta function.

E.3.2. SOLVING THE DMFT EQUATIONS

We now solve the DMFT equations derived in the previous section for ridge regression.

Lemma E.3 *Define the function $K_i(t)$ for $i\geq 0$ as*

$$K_i(t) := \int x^i \mathrm{e}^{-(x+\lambda)t}\,\mathrm{d}\mu_{\mathrm{MP}}(x)\,, \tag{369}$$

where μ_{MP} is the Marchenko–Pastur law*, whose density is given by Equation* (330)*.*

The solution of the DMFT equations for ridge regression (363)–(368) *is given by*

$$R_\theta(t,t') = K_0(t-t')\,,\quad R_\ell(t,t') = -\frac{1}{\delta}K_1(t-t')\,, \tag{370}$$

$$C_\theta^*(t,t') = C_{\theta 0}^*(t,t') + \tau \int_0^{t\wedge t'} K_1(t+t'-2s) L(s,s)\,\mathrm{d}s\,, \tag{371}$$

$$L(t,t') = L_0(t,t') + \tau \int_0^{t\wedge t'} K_2(t+t'-2s) L(s,s)\,\mathrm{d}s\,, \tag{372}$$

where $C_{\theta 0}^*$ and L_0 are given by

$$\begin{aligned} C_{\theta 0}^*(t,t') &= \rho^2 \int \frac{(\lambda + x\mathrm{e}^{-(x+\lambda)t})(\lambda + x\mathrm{e}^{-(x+\lambda)t'})}{(x+\lambda)^2}\,\mathrm{d}\mu_{\mathrm{MP}}(x) \\ &\quad + \frac{\sigma^2}{\delta} \int \frac{x}{(x+\lambda)^2}(1-\mathrm{e}^{-(x+\lambda)t})(1-\mathrm{e}^{-(x+\lambda)t'})\,\mathrm{d}\mu_{\mathrm{MP}}(x)\,, \end{aligned} \tag{373}$$

$$\begin{aligned} L_0(t,t') &= \rho^2 \int x\frac{(\lambda + x\mathrm{e}^{-(x+\lambda)t})(\lambda + x\mathrm{e}^{-(x+\lambda)t'})}{(x+\lambda)^2}\,\mathrm{d}\mu_{\mathrm{MP}}(x) \\ &\quad + \frac{\sigma^2}{\delta} \int \frac{(\lambda + x\mathrm{e}^{-(x+\lambda)t})(\lambda + x\mathrm{e}^{-(x+\lambda)t'})}{(x+\lambda)^2}\,\mathrm{d}\mu_{\mathrm{MP}}(x) + \frac{\delta-1}{\delta}\sigma^2\,. \end{aligned} \tag{374}$$

Before proving Lemma E.3, we summarize the necessary background on the *Laplace transform*, which is a useful technique for analyzing linear differential equations and will be used extensively in the proof. Given a function $f\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$, its Laplace transform $\mathcal{L}[f] = \overline{f}$ is defined as

$$\overline{f}(p) \coloneqq \int_0^\infty f(t)\mathrm{e}^{-pt}\,\mathrm{d}t\,, \tag{375}$$

for $p \in \mathbb{C}$ with sufficiently large real part for the integral to be convergent.

We state several of its basic properties used in the proof.

- *Linearity:* For $f\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ and $a, b \in \mathbb{R}$, we have $\mathcal{L}[af+b] = a\mathcal{L}[f] + b$.
- *Laplace transforms of derivatives, integrals, and convolutions:* For $f, g\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$, we have

$$\mathcal{L}\big[f'(t)\big](p) = p\overline{f}(p) - f(0)\,, \tag{376}$$

$$\mathcal{L}\left[\int_0^t f(s)\,\mathrm{d}s\right](p) = \frac{\overline{f}(p)}{p}\,, \tag{377}$$

$$\mathcal{L}\left[\int_0^t f(t-s)g(s)\,\mathrm{d}s\right](p) = \overline{f}(p)\overline{g}(p)\,. \tag{378}$$

- *Laplace transform of the Dirac delta function:* We have $\mathcal{L}[\delta(t)](p) = 1$, where $\delta(t)$ is the Dirac delta function.

We also utilize a two-dimensional version of the Laplace transform, which is defined for $g\colon \mathbb{R}_{\geq 0}^2 \to \mathbb{R}$ as

$$\overline{g}(p,q) \coloneqq \int_0^\infty \int_0^\infty g(t,t')\mathrm{e}^{-pt}\mathrm{e}^{-qt'}\,\mathrm{d}t\,\mathrm{d}t'\,, \tag{379}$$

for $p, q \in \mathbb{C}$ with sufficiently large real parts for the integral to be convergent. Its properties are similar to the one-dimensional case.

Proof We proceed as follows. First, we solve the equation for the noiseless case $\tau = 0$ in the frequency domain using the Laplace transform. Next, we perform the inverse Laplace transform to obtain the time-domain solution for $\tau = 0$. Along the way, we use ideas and techniques from the random matrix theory. Finally, we solve the full equations for $\tau > 0$.

Solving in the frequency domain for $\tau = 0$. Note that the equations for R_θ, R_ℓ, and $C_\theta(\cdot, *)$ do not depend on τ and are thus the same for $\tau = 0$ and $\tau > 0$. Let $C^*_{\theta 0}$ and L_0 be the solutions of Equations (365) and (368) for $\tau = 0$.

Since the equations for $R_\theta(t, t')$ (363) and $R_\ell(t, t')$ (366) depend on time only through the time difference $t - t'$, they are time-translation invariant, i.e., $R_\theta(t, t') = R_\theta(t - t')$ and $R_\ell(t, t') = R_\ell(t - t')$. Therefore, they satisfy the following one-dimensional integral equations:

$$R_\theta(t) = 1 - \int_0^t \left((1+\lambda) R_\theta(s) + \int_0^s R_\ell(s - s') R_\theta(s') \, \mathrm{d}s' \right) \mathrm{d}s \,, \tag{380}$$

$$R_\ell(t) = -\frac{1}{\delta} R_\theta(t) - \frac{1}{\delta} \int_0^t R_\theta(t - s) R_\ell(s) \, \mathrm{d}s \,. \tag{381}$$

Taking the Laplace transforms of these equations, we have

$$\overline{R}_\theta(p) = \frac{1}{p} \big(1 - (1+\lambda) \overline{R}_\theta(p) - \overline{R}_\ell(p) \overline{R}_\theta(p) \big) \,, \tag{382}$$

$$\overline{R}_\ell(p) = -\frac{1}{\delta} \overline{R}_\theta(p) - \frac{1}{\delta} \overline{R}_\theta(p) \overline{R}_\ell(p) \,. \tag{383}$$

Rearranging, we obtain

$$\overline{R}_\theta(p) = -\frac{\delta(1 + p + \lambda) - 1 - \sqrt{(\delta(1 + p + \lambda) - 1)^2 + 4\delta(p + \lambda)}}{2(p + \lambda)} \,, \tag{384}$$

$$\overline{R}_\ell(p) = -\frac{\overline{R}_\theta(p)}{\delta + \overline{R}_\theta(p)} = \frac{(p + \lambda) \overline{R}_\theta(p) - 1}{\delta} \,. \tag{385}$$

Taking the Laplace transforms of the correlation functions in Equations (364), (365), and (368), we have

$$\overline{C}_\theta(p, *) = \rho^2 \overline{R}_\theta(p) \cdot \frac{1 + \overline{R}_\ell(p)}{p} \,, \tag{386}$$

$$\overline{C}^*_{\theta 0}(p, q) = \frac{1}{\delta} \overline{R}_\theta(p) \overline{R}_\theta(q) \overline{L}_0(p, q) + \frac{1}{\rho^2} \left(\overline{C}_\theta(p, *) - \frac{\rho^2}{p} \right) \left(\overline{C}_\theta(q, *) - \frac{\rho^2}{q} \right) , \tag{387}$$

$$\overline{L}_0(p, q) = (1 + \overline{R}_\ell(p))(1 + \overline{R}_\ell(q)) \left(\overline{C}^*_{\theta 0}(p, q) + \frac{\sigma^2}{pq} \right) . \tag{388}$$

Simplifying these equations, we obtain

$$\overline{C}^*_{\theta 0}(p, q) = \frac{\rho^2 (p + \lambda)(q + \lambda)}{pq} \frac{\overline{R}_\theta(p) \overline{R}_\theta(q)}{1 - \delta \overline{R}_\ell(p) \overline{R}_\ell(q)} + \frac{\sigma^2}{pq} \frac{\delta \overline{R}_\ell(p) \overline{R}_\ell(q)}{1 - \delta \overline{R}_\ell(p) \overline{R}_\ell(q)} \,, \tag{389}$$

$$\begin{aligned} \overline{L}_0(p, q) = {} & \frac{\rho^2 \delta (p + \lambda)(q + \lambda)}{pq} \frac{\delta \overline{R}_\ell(p) \overline{R}_\ell(q)}{1 - \delta \overline{R}_\ell(p) \overline{R}_\ell(q)} \\ & + \frac{\sigma^2 (p + \lambda)(q + \lambda)}{\delta pq} \frac{\overline{R}_\theta(p) \overline{R}_\theta(q)}{1 - \delta \overline{R}_\ell(p) \overline{R}_\ell(q)} + \frac{\delta - 1}{\delta} \frac{\sigma^2}{pq} \,. \end{aligned} \tag{390}$$

Using the relation

$$1 - \delta \overline{R}_\ell(p) \overline{R}_\ell(q) = \frac{(q - p) \overline{R}_\theta(p) \overline{R}_\theta(q)}{\overline{R}_\theta(p) - \overline{R}_\theta(q)} \,, \tag{391}$$

we have

$$\frac{\overline{R}_\theta(p)\overline{R}_\theta(q)}{1-\delta\overline{R}_\ell(p)\overline{R}_\ell(q)} = \frac{\overline{R}_\theta(p)-\overline{R}_\theta(q)}{q-p}\,, \qquad \frac{\delta\overline{R}_\ell(p)\overline{R}_\ell(q)}{1-\delta\overline{R}_\ell(p)\overline{R}_\ell(q)} = \frac{\overline{R}_\ell(q)-\overline{R}_\ell(p)}{q-p}\,. \tag{392}$$

Thus, we obtain

$$\overline{C}^*_{\theta 0}(p,q) = \frac{\rho^2(p+\lambda)(q+\lambda)}{pq}\frac{\overline{R}_\theta(p)-\overline{R}_\theta(q)}{q-p} + \frac{\sigma^2}{pq}\frac{\overline{R}_\ell(q)-\overline{R}_\ell(p)}{q-p}\,, \tag{393}$$

$$\overline{L}_0(p,q) = \frac{\rho^2\delta(p+\lambda)(q+\lambda)}{pq}\frac{\overline{R}_\ell(q)-\overline{R}_\ell(p)}{q-p} + \frac{\sigma^2(p+\lambda)(q+\lambda)}{\delta pq}\frac{\overline{R}_\theta(p)-\overline{R}_\theta(q)}{q-p} + \frac{\delta-1}{\delta}\frac{\sigma^2}{pq}\,. \tag{394}$$

Solving in the time domain for $\tau = 0$. We now perform the inverse Laplace transform to obtain the time-domain solution. Before proceeding, we introduce *Stieltjes transform*. The Stieltjes transform $S : \mathbb{C}\setminus I \to \mathbb{C}$ of a (signed) measure μ on an interval $I \subseteq \mathbb{R}$ is defined as follows:

$$S(z) = \int_I \frac{1}{x-z}\,\mathrm{d}\mu(x)\,. \tag{395}$$

We check that the time-domain solution stated in Lemma E.3 has the same Laplace transform as the frequency-domain solution obtained above. First, we take the Laplace transform of R_θ to obtain

$$\begin{aligned}\int_0^\infty R_\theta(t)\mathrm{e}^{-pt}\,\mathrm{d}t &= \int_0^\infty\left(\int \mathrm{e}^{-(x+\lambda+p)t}\,\mathrm{d}\mu_{\mathrm{MP}}(x)\right)\mathrm{d}t = \int\left(\int_0^\infty \mathrm{e}^{-(x+\lambda+p)t}\,\mathrm{d}t\right)\mathrm{d}\mu_{\mathrm{MP}}(x)\\ &= \int \frac{1}{x+\lambda+p}\,\mathrm{d}\mu_{\mathrm{MP}}(x) = S_{\mathrm{MP}}(-(p+\lambda))\,.\end{aligned} \tag{396}$$

Here, $S_{\mathrm{MP}}(z)$ is the Stieltjes transform of the Marchenko–Pastur law which is given by

$$S_{\mathrm{MP}}(z) = \frac{\delta(1-z)-1-\sqrt{(\delta(1-z)-1)^2-4\delta z}}{2z}\,. \tag{397}$$

Setting $z = -(p+\lambda)$, we obtain $\overline{R}_\theta(p)$ given in Equation (384).

Next, we check R_ℓ.

$$\int_0^\infty R_\ell(t)\mathrm{e}^{-pt}\,\mathrm{d}t = -\frac{1}{\delta}\int \frac{x}{x+p+\lambda}\,\mathrm{d}\mu_{\mathrm{MP}}(x) = -\frac{1}{\delta}(-(p+\lambda)S_{\mathrm{MP}}(-(p+\lambda))+1)\,, \tag{398}$$

which is equal to $\overline{R}_\ell(p)$ in Equation (385).

Finally, we check $C^*_{\theta 0}$ and L_0. Define $F_1(t,t'), F_2(t,t')$ as follows:

$$F_1(t,t') = K_0(t+t')\,, \quad F_2(t,t') = \frac{1}{\delta}K_1(t+t')\,. \tag{399}$$

The Laplace transforms of these functions are

$$\int_0^\infty\int_0^\infty F_1(t,t')\mathrm{e}^{-pt-qt'}\,\mathrm{d}t\,\mathrm{d}t' = \int \frac{1}{(x+p+\lambda)(x+q+\lambda)}\,\mathrm{d}\mu_{\mathrm{MP}}(x) = \frac{\overline{R}_\theta(p)-\overline{R}_\theta(q)}{q-p}\,, \tag{400}$$

$$\int_0^\infty \int_0^\infty F_2(t,t')\mathrm{e}^{-pt-qt'}\,\mathrm{d}t\,\mathrm{d}t' = \frac{1}{\delta}\int \frac{x}{(x+p+\lambda)(x+q+\lambda)}\,\mathrm{d}\mu_{\mathrm{MP}}(x) = \frac{\overline{R}_\ell(q)-\overline{R}_\ell(p)}{q-p}\,. \tag{401}$$

Furthermore, we have

$$\begin{aligned}
\mathcal{L}^{-1}\left[\frac{(p+\lambda)(q+\lambda)}{pq}\frac{\overline{R}_\theta(p)-\overline{R}_\theta(q)}{q-p}\right] &= \mathcal{L}^{-1}\left[\frac{(p+\lambda)(q+\lambda)}{pq}\overline{F}_1(p,q)\right] \\
&= \int_0^t\int_0^{t'} (\delta(t-s)+\lambda)(\delta(t'-s')+\lambda)K_0(s+s')\,\mathrm{d}s'\,\mathrm{d}t' \\
&= \int \frac{(\lambda+x\mathrm{e}^{-(x+\lambda)t})(\lambda+x\mathrm{e}^{-(x+\lambda)t'})}{(x+\lambda)^2}\,\mathrm{d}\mu_{\mathrm{MP}}(x)\,, \qquad (402)\\
\mathcal{L}^{-1}\left[\frac{1}{pq}\frac{\overline{R}_\ell(q)-\overline{R}_\ell(p)}{q-p}\right] &= \mathcal{L}^{-1}\left[\frac{\overline{F}_2(p,q)}{pq}\right] = \frac{1}{\delta}\int_0^t\int_0^{t'} K_1(s,s')\,\mathrm{d}s'\,\mathrm{d}s \\
&= \frac{1}{\delta}\int \frac{x}{(x+\lambda)^2}(1-\mathrm{e}^{-(x+\lambda)t})(1-\mathrm{e}^{-(x+\lambda)t'})\,\mathrm{d}\mu_{\mathrm{MP}}(x)\,, \qquad (403)\\
\mathcal{L}^{-1}\left[\frac{(p+\lambda)(q+\lambda)}{pq}\frac{\overline{R}_\ell(q)-\overline{R}_\ell(p)}{q-p}\right] &= \mathcal{L}^{-1}\left[\frac{(p+\lambda)(q+\lambda)}{pq}\frac{\overline{F}_2(p,q)}{pq}\right] \\
&= \frac{1}{\delta}\int_0^t\int_0^{t'} (\delta(t-s)+\lambda)(\delta(t'-s')+\lambda)K_1(s,s')\,\mathrm{d}s'\,\mathrm{d}t' \\
&= \frac{1}{\delta}\int x\frac{(\lambda+x\mathrm{e}^{-(x+\lambda)t})(\lambda+x\mathrm{e}^{-(x+\lambda)s})}{(x+\lambda)^2}\,\mathrm{d}\mu_{\mathrm{MP}}(x)\,. \qquad (404)
\end{aligned}$$

Thus, by Equations (393) and (394), we have the desired expressions for $C^*_{\theta 0}$ and L_0.

Solving for $\tau > 0$. Equations for $\Delta C^*_\theta(t,t') \coloneqq C^*_\theta(t,t') - C^*_{\theta 0}(t,t')$ and $\Delta L(t,t') \coloneqq L(t,t') - L_0(t,t')$ are given by

$$\begin{aligned}
&\Delta C^*_\theta(t,t') \\
&= \frac{1}{\delta}\int_0^t\int_0^{t'} R_\theta(t-s)R_\theta(t'-s')\Delta L(s,s')\,\mathrm{d}s'\,\mathrm{d}s \\
&+ \tau\int_0^t\int_0^{t'} R_\theta(t-s)R_\theta(t'-s)(R_\ell(s-s')L(s',s') + R_\ell(s'-s)L(s,s) + L(s,s')\delta(s-s'))\,\mathrm{d}s'\,\mathrm{d}s\,,
\end{aligned} \tag{405}$$

and

$$\begin{aligned}
\Delta L(t,t') = \frac{1}{\delta}\int_0^t\int_0^{t'} &(\delta(t-s)+R_\ell(t-s))(\delta(t'-s')+R_\ell(t'-s'))\Delta C^*_\theta(s,s')\,\mathrm{d}s'\,\mathrm{d}s \\
&+ \tau\int_0^t\int_0^{t'} R_\ell(t-s)R_\ell(t'-s)L(s,s)\,\mathrm{d}s\,. \qquad (406)
\end{aligned}$$

Define $L'(t,t') \coloneqq L(t,t)\delta(t-t')$. Taking the Laplace transform of these equations, we obtain

$$\Delta\overline{C}^*_\theta(p,q) = \frac{1}{\delta}\overline{R}_\theta(p)\overline{R}_\theta(q)\Delta\overline{L}(p,q) + \tau\overline{R}_\theta(p)\overline{R}_\theta(q)(1+\overline{R}_\ell(p)+\overline{R}_\ell(q))\overline{L}'(p,q)\,, \tag{407}$$

$$\Delta\overline{L}(p,q) = (1+\overline{R}_\ell(p))(1+\overline{R}_\ell(q))\Delta\overline{C}^*_\theta(p,q) + \tau\delta\overline{R}_\ell(p)\overline{R}_\ell(q)\overline{L}'(p,q)\,. \tag{408}$$

Further simplification gives

$$\Delta\overline{C}^*_\theta(p,q) = \frac{\delta^2\overline{R}_\ell(p)\overline{R}_\ell(q)}{1-\delta\overline{R}_\ell(p)\overline{R}_\ell(q)}\tau\overline{L}'(p,q) = \frac{\delta(\overline{R}_\ell(q)-\overline{R}_\ell(p))}{q-p}\tau\overline{L}'(p,q)\,, \tag{409}$$

$$\begin{aligned}\Delta\overline{L}(p,q) &= \frac{\delta(\delta(\overline{R}_\ell(p)+\overline{R}_\ell(q))+\delta+1)R_\ell(p)R_\ell(q)}{1-\delta\overline{R}_\ell(p)\overline{R}_\ell(q)}\tau\overline{L}'(p,q)\\ &= \frac{(\delta R_\ell(q)^2+(\delta+1)R_\ell(q)) - (\delta R_\ell(p)^2+(\delta+1)R_\ell(p))}{q-p}\tau\overline{L}'(p,q)\,.\end{aligned} \tag{410}$$

We perform the inverse Laplace transform. For $\Delta\overline{C}^*_\theta(p,q)$, we have

$$\Delta C^*_\theta(t,t') = \tau\delta\int_0^t\int_0^{t'} F_2(t-s,t'-s')L'(s,s')\,\mathrm{d}s'\,\mathrm{d}s = \tau\int_0^{t\wedge t'} K_1(t+t'-2s)L(s,s)\,\mathrm{d}s\,, \tag{411}$$

and we obtain Equation (371). For $\Delta\overline{L}(p,q)$, we use that

$$\begin{aligned}\int_0^\infty\int_0^\infty K_2(t+t')\mathrm{e}^{-pt-qt'}\,\mathrm{d}t'\,\mathrm{d}t &= \int\frac{x^2}{(x+\lambda+p)(x+\lambda+q)}\,\mathrm{d}\mu_{\mathrm{MP}}(x)\\ &= \frac{(\delta R_\ell(q)^2+(\delta+1)R_\ell(q)) - (\delta R_\ell(p)^2+(\delta+1)R_\ell(p))}{q-p}\,,\end{aligned} \tag{412}$$

and proceeding similarly, we obtain Equation (372).

■

Finally, we prove Proposition E.2. The asymptotic train and test errors can be expressed in terms of the DMFT solution as

$$\mathscr{L}(\boldsymbol{\theta}^t) = \frac{1}{n}\sum_{i=1}^n(\boldsymbol{x}_i^\top\boldsymbol{\theta}^t - y_i)^2 \to \mathbb{E}[(r^t - r^* - z)^2] = L(t,t)\,, \tag{413}$$

$$\mathscr{R}(\boldsymbol{\theta}^t) = \frac{1}{d}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|_2^2 + \sigma^2 \to \mathbb{E}[(\theta^t-\theta^*)^2] + \sigma^2 = C^*_\theta(t,t) + \sigma^2\,. \tag{414}$$

By Lemma E.3, the asymptotic train and test errors $\mathscr{L}(t) \coloneqq L(t,t)$ and $\mathscr{R}(t) \coloneqq C^*_\theta(t,t)+\sigma^2$ satisfy

$$\mathscr{L}(t) = \mathscr{L}_0(t) + \tau\int_0^t K_2(2(t-s))\mathscr{L}(s)\,\mathrm{d}s\,,\quad \mathscr{R}(t) = \mathscr{R}_0(t) + \tau\int_0^t K_1(2(t-s))\mathscr{L}(s)\,\mathrm{d}s\,, \tag{415}$$

where $\mathscr{L}_0(t) \coloneqq L_0(t,t)$ and $\mathscr{R}_0(t) \coloneqq C^*_{\theta 0}(t,t) + \sigma^2$ are the asymptotic train and test errors for the noiseless case $\tau = 0$, which can be written explicitly as

$$\mathscr{L}_0(t) = \rho^2 \int \frac{x(\lambda + x\mathrm{e}^{-(x+\lambda)t})^2}{(x+\lambda)^2}\,\mathrm{d}\mu_{\mathrm{MP}}(x) + \frac{\sigma^2}{\delta}\int \frac{(\lambda + x\mathrm{e}^{-(x+\lambda)t})^2}{(x+\lambda)^2}\,\mathrm{d}\mu_{\mathrm{MP}}(x) + \frac{\delta-1}{\delta}\sigma^2\,, \tag{416}$$

$$\mathscr{R}_0(t) = \rho^2 \int \frac{(\lambda + x\mathrm{e}^{-(x+\lambda)t})^2}{(x+\lambda)^2}\,\mathrm{d}\mu_{\mathrm{MP}}(x) + \frac{\sigma^2}{\delta}\int \frac{x}{(x+\lambda)^2}(1-\mathrm{e}^{-(x+\lambda)t})^2\,\mathrm{d}\mu_{\mathrm{MP}}(x) + \sigma^2\,. \tag{417}$$

Setting $H_i(t) \coloneqq K_i(2t)$, we obtain Equations (332)–(334). This concludes the proof of Proposition E.2.

Appendix F. Details of Numerical Simulations

We numerically solve the discretized DMFT equation given in Appendix B.2 using Monte Carlo sampling. Instead of directly working with the system $\mathfrak{S}^\gamma$ given in Appendix B.2, we work with the equivalent system (33), as it is simpler to implement. We solve Equation (33) by iterating the following steps until convergence:

0. Start with a random guess of the DMFT solution (C_θ, R_θ).
1. Given the current estimate of (C_θ, R_θ), sample M instances of the stochastic processes r^{t_i} and $\partial \ell_{t_i}(r^{t_i}; z)/\partial w^{t_j}$, and compute the functions (C_ℓ, R_ℓ, Γ) by averaging over the samples.
2. Given the current estimate of (C_ℓ, R_ℓ, Γ), sample M instances of the stochastic processes θ^{t_i} and $\partial\theta^{t_i}/\partial u^{t_j}$, and compute the functions $\widetilde{C}_\theta$ and $\widetilde{R}_\theta$ by averaging over the samples.
3. Update the DMFT solution as $(C_\theta, R_\theta) \leftarrow (1-\alpha)(C_\theta, R_\theta) + \alpha(\widetilde{C}_\theta, \widetilde{R}_\theta)$ where $\alpha \in (0,1]$ is a damping factor.

In our experiments, we set the number of samples to $M = 8000$, the damping factor to $\alpha = 0.8$, and the time step to $\gamma = 0.05$. We observe that the above iteration converges in around 10 iterations for the settings considered in this paper.

For the logistic regression setting with $\ell_t(r, r^*; z) = -y/(1+\exp(yr))$ where $y = \mathrm{sign}(r^* + z)$, there is an issue with computing the function $R_\ell(t_i, *) = \mathbb{E}[\partial \ell_{t_i}(r^{t_i}, r^*; z)/\partial r^*]$ due to the non-differentiability of ℓ with respect to r^* at $r^* = -z$. However, we can avoid the differentiation by r^* entirely by (heuristically) using Stein's lemma:

$$\begin{aligned}\mathbb{E}[r^*\ell_{t_i}(r^{t_i}, r^*; z)] &= \sum_{j=0}^{i-1} \mathrm{Cov}(r^*, w^{t_j})\,\mathbb{E}\left[\frac{\partial \ell_{t_i}(r^{t_i}, r^*; z)}{\partial w^{t_j}}\right] + \mathrm{Cov}(r^*, r^*)\,\mathbb{E}\left[\frac{\partial \ell_{t_i}(r^{t_i}, r^*; z)}{\partial r^*}\right] \\ &= \gamma \sum_{j=0}^{i-1} C_\ell(t_j, *) R_\ell(t_i, t_j) + \rho^2 R_\ell(t_i, *)\,,\end{aligned} \tag{418}$$

and computing $R_\ell(t_i, *)$ as

$$R_\ell(t_i, *) = \frac{1}{\rho^2}\left(\mathbb{E}[r^*\ell_{t_i}(r^{t_i}, r^*; z)] - \gamma \sum_{j=0}^{i-1} R_\ell(t_i, t_j) C_\ell(t_j, *)\right)\,, \tag{419}$$

and calculating the expectation using Monte Carlo integration.

Appendix G. Discretization of SDEs in High Dimensions

In this section, we analyze the discretization error of a general stochastic differential equation in high dimensions. The results presented here are used in the proof of Lemma D.1.

Let $T > 0$, $\boldsymbol{b}\colon [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$, and $\boldsymbol{\sigma}\colon [0,T] \times \mathbb{R}^d \to \mathbb{R}^{d\times n}$. Consider the following SDE in $\mathbb{R}^d$:

$$\mathrm{d}\boldsymbol{\theta}^t = \boldsymbol{b}(t,\boldsymbol{\theta}^t)\,\mathrm{d}t + \boldsymbol{\sigma}(t,\boldsymbol{\theta}^t)\,\mathrm{d}\boldsymbol{W}^t\,, \tag{420}$$

with given initial condition $\boldsymbol{\theta}^0 \in \mathbb{R}^d$. Here, $\boldsymbol{W}^t \in \mathbb{R}^n$ is a Brownian motion.

We discretize the SDE using the Euler–Maruyama method with step size $\gamma > 0$ as follows:

$$\boldsymbol{\theta}_\gamma^{t_{k+1}} = \boldsymbol{\theta}_\gamma^{t_k} + \boldsymbol{b}(t_k,\boldsymbol{\theta}_\gamma^{t_k})\gamma + \boldsymbol{\sigma}(t_k,\boldsymbol{\theta}_\gamma^{t_k})(\boldsymbol{W}^{t_{k+1}} - \boldsymbol{W}^{t_k})\,, \quad \boldsymbol{\theta}_\gamma^0 = \boldsymbol{\theta}^0\,, \tag{421}$$

where $t_k := k\gamma$.

There is a standard result [36, Theorem 10.2.2] for bounding the difference between $\boldsymbol{\theta}^t$ and $\boldsymbol{\theta}_\gamma^t$. However, they treat n, d as constants, and the dependence of the bound on n, d is not obvious. Here, we show a bound that explicitly tracks the dependence on n, d.

Assumption G.1 *There exists a constant $L > 0$ independent of n and d such that the following hold.*

1. *(Proportional asymptotics): $1/L \le n/d \le L$.*
2. *(Lipschitz continuity): $\|\boldsymbol{b}(t_1,\boldsymbol{x}_1) - \boldsymbol{b}(t_2,\boldsymbol{x}_2)\|_2^2 + \|\boldsymbol{\sigma}(t_1,\boldsymbol{x}_1) - \boldsymbol{\sigma}(t_2,\boldsymbol{x}_2)\|_{\mathrm{F}}^2 \le L(d(t_1-t_2)^2 + \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2)$ for $t_1, t_2 \in [0,T]$, $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$.*
3. *(Linear growth): $\|\boldsymbol{b}(t,\boldsymbol{x})\|_2^2 + \|\boldsymbol{\sigma}(t,\boldsymbol{x})\|_{\mathrm{F}}^2 \le L(d + dt^2 + \|\boldsymbol{x}\|_2^2)$ for $t \in [0,T]$, $\boldsymbol{x} \in \mathbb{R}^d$.*

Lemma G.2 (Norm bound on the discretized SDE iterates) *Under Assumption G.1 and $\gamma < 1$, there exists a constant $C := C(L) > 0$ independent of n, d, T, γ such that the following holds.*

$$\mathbb{E}\left[\max_{0\le k\le \lfloor T/\gamma\rfloor} \|\boldsymbol{\theta}_\gamma^{t_k}\|_2^2\right] \le \mathrm{e}^{CT}(Td + \|\boldsymbol{\theta}^0\|_2^2)\,. \tag{422}$$

Proof For notational simplicity, we denote $K := \lfloor T/\gamma\rfloor$, $\boldsymbol{\theta}^k := \boldsymbol{\theta}_\gamma^{t_k}$, $\boldsymbol{b}^k := \boldsymbol{b}(t_k,\boldsymbol{\theta}_\gamma^{t_k})$, $\boldsymbol{\sigma}^k := \boldsymbol{\sigma}(t_k,\boldsymbol{\theta}_\gamma^{t_k})$, and $\boldsymbol{\xi}^k := \boldsymbol{W}^{t_{k+1}} - \boldsymbol{W}^{t_k}$. From Equation (421), we have

$$\|\boldsymbol{\theta}^{k+1}\|_2^2 = \|\boldsymbol{\theta}^k\|_2^2 + 2\gamma\boldsymbol{\theta}^{k\mathsf{T}}\boldsymbol{b}^k + \gamma^2\|\boldsymbol{b}^k\|_2^2 + \|\boldsymbol{\sigma}^k\boldsymbol{\xi}^k\|_2^2 + 2(\boldsymbol{\theta}^k + \gamma\boldsymbol{b}^k)^\mathsf{T}\boldsymbol{\sigma}^k\boldsymbol{\xi}^k\,. \tag{423}$$

Summing over k steps, we have

$$\|\boldsymbol{\theta}^k\|_2^2 = \|\boldsymbol{\theta}^0\|_2^2 + \underbrace{\sum_{j=0}^{k-1}\Big(2\gamma\boldsymbol{\theta}^{j\mathsf{T}}\boldsymbol{b}^j + \gamma^2\|\boldsymbol{b}^j\|_2^2 + \|\boldsymbol{\sigma}^j\boldsymbol{\xi}^j\|_2^2\Big)}_{=:A^k} + \underbrace{\sum_{j=0}^{k-1} 2(\boldsymbol{\theta}^j + \gamma\boldsymbol{b}^j)^\mathsf{T}\boldsymbol{\sigma}^j\boldsymbol{\xi}^j}_{=:M^k}\,. \tag{424}$$

We first bound A^k. We have

$$\max_{0\le k\le K}|A^k| \le \sum_{j=0}^{K-1}\Big(2\gamma|\boldsymbol{\theta}^{j\mathsf{T}}\boldsymbol{b}^j| + \gamma^2\|\boldsymbol{b}^j\|_2^2 + \|\boldsymbol{\sigma}^j\boldsymbol{\xi}^j\|_2^2\Big)\,. \tag{425}$$

By the linear growth condition, we have

$$2\gamma|\boldsymbol{\theta}^{j\mathsf{T}}\boldsymbol{b}^j| + \gamma^2\|\boldsymbol{b}^j\|_2^2 \le 2\gamma\|\boldsymbol{\theta}^j\|_2^2 + (2\gamma+\gamma^2)\|\boldsymbol{b}^j\|_2^2 \le C\gamma((1+T^2)d + \|\boldsymbol{\theta}^j\|_2^2)\,. \tag{426}$$

Using the covariance of $\boldsymbol{\xi}^j \sim \mathsf{N}(0,\gamma\boldsymbol{I})$ and the linear growth condition, we have

$$\mathbb{E}[\|\boldsymbol{\sigma}^j\boldsymbol{\xi}^j\|_2^2 \mid \mathcal{F}^j] = \gamma\|\boldsymbol{\sigma}^j\|_{\mathrm{F}}^2 \le C\gamma((1+T^2)d + \|\boldsymbol{\theta}^j\|_2^2)\,, \tag{427}$$

where $\mathcal{F}^j$ is the filtration generated by $\{\boldsymbol{\xi}^0,\dots,\boldsymbol{\xi}^{j-1}\}$. Thus, taking the expectation, we have

$$\mathbb{E}\Big[\max_{0\le k\le K}|A^k|\Big] \le C\gamma\sum_{j=0}^{K-1}((1+T^2)d + \mathbb{E}\|\boldsymbol{\theta}^j\|_2^2)\,. \tag{428}$$

Next, we bound M^k. Notice that M^k is a martingale with respect to the filtration $\mathcal{F}^k$ since $\mathbb{E}[\boldsymbol{\xi}^j \mid \mathcal{F}^j] = 0$. Thus, by the Burkholder–Davis–Gundy inequality, we have

$$\begin{aligned}
\mathbb{E}\Big[\max_{0\le k\le K}|M^k|\Big] &\le C\,\mathbb{E}\left(\sum_{j=0}^{K-1}\mathbb{E}\Big[\Big(2(\boldsymbol{\theta}^j+\gamma\boldsymbol{b}^j)^{\mathsf{T}}\boldsymbol{\sigma}^j\boldsymbol{\xi}^j\Big)^2 \mid \mathcal{F}^j\Big]\right)^{1/2}\\
&= C\,\mathbb{E}\left(\sum_{j=0}^{K-1}4\gamma\|(\boldsymbol{\theta}^j+\gamma\boldsymbol{b}^j)^{\mathsf{T}}\boldsymbol{\sigma}^j\|_2^2\right)^{1/2}\\
&= C\,\mathbb{E}\left(\gamma\sum_{j=0}^{K-1}\|\boldsymbol{\theta}^j+\gamma\boldsymbol{b}^j\|_2^2\|\boldsymbol{\sigma}^j\|_{\mathrm{F}}^2\right)^{1/2}\\
&\le C\,\mathbb{E}\left[\left(\max_{0\le k\le K}\|\boldsymbol{\theta}^k+\gamma\boldsymbol{b}^k\|_2^2\right)^{1/2}\left(\gamma\sum_{j=0}^{K-1}\|\boldsymbol{\sigma}^j\|_{\mathrm{F}}^2\right)^{1/2}\right]\\
&\le C\,\mathbb{E}\left[\left(\gamma^2(1+T^2)d + \max_{0\le k\le K}\|\boldsymbol{\theta}^k\|_2^2\right)^{1/2}\left(\gamma\sum_{j=0}^{K-1}((1+T^2)d+\|\boldsymbol{\theta}^j\|_2^2)\right)^{1/2}\right],
\end{aligned} \tag{429}$$

where we used the linear growth condition in the last line. By Young's inequality, we have

$$\mathbb{E}\Big[\max_{0\le k\le K}|M^k|\Big] \le \frac{1}{2}\left(\gamma^2(1+T^2)d + \mathbb{E}\Big[\max_{0\le k\le K}\|\boldsymbol{\theta}^k\|_2^2\Big]\right) + C\gamma\sum_{j=0}^{K-1}((1+T^2)d + \mathbb{E}\|\boldsymbol{\theta}^j\|_2^2)\,. \tag{430}$$

Combining Equations (428) and (430), we have

$$\mathbb{E}\left[\max_{0\le k\le K}\|\boldsymbol{\theta}^k\|_2^2\right] \le \|\boldsymbol{\theta}^0\|_2^2 + C\gamma^2(1+T^2)d + C\gamma\sum_{j=0}^{K-1}((1+T^2)d + \mathbb{E}\|\boldsymbol{\theta}^j\|_2^2) + \frac{1}{2}\,\mathbb{E}\left[\max_{0\le k\le K}\|\boldsymbol{\theta}^k\|_2^2\right]. \tag{431}$$

Rearranging the terms, we have

$$\mathbb{E}\left[\max_{0\le k\le K}\|\boldsymbol{\theta}^k\|_2^2\right] \le C(T(1+T^2)d + \|\boldsymbol{\theta}^0\|_2^2) + C\gamma\sum_{j=0}^{K-1}\mathbb{E}\left[\max_{0\le j\le k}\|\boldsymbol{\theta}^j\|_2^2\right]. \tag{432}$$

The bound (422) follows by applying Grönwall's inequality and absorbing $1+T^2$ into the exponential factor e^{CT}. ■

Remark G.3 *Lemma G.2 holds verbatim to the case where the Gaussian increments $\boldsymbol{W}^{t_{k+1}} - \boldsymbol{W}^{t_k}$ are replaced by independent random vectors $\boldsymbol{\xi}^k \in \mathbb{R}^n$ with $\mathbb{E}[\boldsymbol{\xi}^k] = 0$ and $\mathbb{E}[\boldsymbol{\xi}^k\boldsymbol{\xi}^{k\mathsf{T}}] = \gamma\boldsymbol{I}_n$, as the proof only uses up to the second moment of the increments.*

Lemma G.4 (Strong approximation of SDE) *Under Assumption G.1 and $\gamma < 1$, there exists a constant $C \coloneqq C(L) > 0$ that does not depend on n, d, T, γ such that the following holds.*

$$\mathbb{E}\left[\sup_{0\le t\le T}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_\gamma^t\|_2^2\right] \le \mathrm{e}^{CT}\gamma(Td + \|\boldsymbol{\theta}^0\|_2^2)\,. \tag{433}$$

Proof Equation (420) can be written as

$$\boldsymbol{\theta}^t = \boldsymbol{\theta}^0 + \int_0^t \boldsymbol{b}(s,\boldsymbol{\theta}^s)\,\mathrm{d}s + \int_0^t \boldsymbol{\sigma}(s,\boldsymbol{\theta}^s)\,\mathrm{d}\boldsymbol{W}^s\,. \tag{434}$$

Let $\lfloor t\rfloor \coloneqq \max\{k\gamma \mid k\gamma \le t, k\in\mathbb{N}\}$ and consider the following stochastic process that embeds Equation (421) into continuous time.

$$\boldsymbol{\theta}_\gamma^t = \boldsymbol{\theta}^0 + \int_0^t \boldsymbol{b}(\lfloor s\rfloor,\boldsymbol{\theta}_\gamma^{\lfloor s\rfloor})\,\mathrm{d}s + \int_0^t \boldsymbol{\sigma}(\lfloor s\rfloor,\boldsymbol{\theta}_\gamma^{\lfloor s\rfloor})\,\mathrm{d}\boldsymbol{W}^s\,. \tag{435}$$

Using Itô's lemma on $\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_\gamma^t\|_2^2$, we have

$$\begin{aligned}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_\gamma^t\|_2^2 = 2\int_0^t (\boldsymbol{\theta}^s - \boldsymbol{\theta}_\gamma^s)^\mathsf{T}(\boldsymbol{b}(s,\boldsymbol{\theta}^s) - \boldsymbol{b}(\lfloor s\rfloor,\boldsymbol{\theta}_\gamma^{\lfloor s\rfloor}))\,\mathrm{d}s \\ + \int_0^t \|\boldsymbol{\sigma}(s,\boldsymbol{\theta}^s) - \boldsymbol{\sigma}(\lfloor s\rfloor,\boldsymbol{\theta}_\gamma^{\lfloor s\rfloor})\|_\mathrm{F}^2\,\mathrm{d}s + M^t\,,\end{aligned} \tag{436}$$

$$M^t \coloneqq 2\int_0^t (\boldsymbol{\theta}^s - \boldsymbol{\theta}_\gamma^s)^\mathsf{T}(\boldsymbol{\sigma}(s,\boldsymbol{\theta}^s) - \boldsymbol{\sigma}(\lfloor s\rfloor,\boldsymbol{\theta}_\gamma^{\lfloor s\rfloor}))\,\mathrm{d}\boldsymbol{W}^s\,. \tag{437}$$

By the Lipschitz assumption, we have

$$2(\boldsymbol{\theta}^s - \boldsymbol{\theta}_\gamma^s)^\mathsf{T}(\boldsymbol{b}(s,\boldsymbol{\theta}^s) - \boldsymbol{b}(\lfloor s\rfloor,\boldsymbol{\theta}_\gamma^{\lfloor s\rfloor})) \le \|\boldsymbol{\theta}^s - \boldsymbol{\theta}_\gamma^s\|_2^2 + \|\boldsymbol{b}(s,\boldsymbol{\theta}^s) - \boldsymbol{b}(\lfloor s\rfloor,\boldsymbol{\theta}_\gamma^{\lfloor s\rfloor})\|_2^2$$

$$\leq C(\gamma^2 d + \|\boldsymbol{\theta}^s - \boldsymbol{\theta}_\gamma^{\lfloor s \rfloor}\|_2^2)\,, \tag{438}$$

$$\|\boldsymbol{\sigma}(s, \boldsymbol{\theta}^s) - \boldsymbol{\sigma}(\lfloor s \rfloor, \boldsymbol{\theta}_\gamma^{\lfloor s \rfloor})\|_{\mathrm{F}}^2 \leq C(\gamma^2 d + \|\boldsymbol{\theta}^s - \boldsymbol{\theta}_\gamma^{\lfloor s \rfloor}\|_2^2)\,. \tag{439}$$

Therefore, we have

$$\mathbb{E}\left[\sup_{0\leq t\leq T} \|\boldsymbol{\theta}^t - \boldsymbol{\theta}_\gamma^t\|_2^2\right] \leq C\int_0^T (\gamma^2 d + \mathbb{E}\|\boldsymbol{\theta}^s - \boldsymbol{\theta}_\gamma^{\lfloor s \rfloor}\|_2^2)\,\mathrm{d}s + \mathbb{E}\left[\sup_{0\leq t\leq T} |M^t|\right]. \tag{440}$$

We next bound the martingale term M^t. By the Burkholder–Davis–Gundy inequality,

$$\begin{aligned}
\mathbb{E}\left[\sup_{0\leq t\leq T} |M^t|\right] &\leq C\,\mathbb{E}\left[\left(4\int_0^T \|(\boldsymbol{\theta}^s - \boldsymbol{\theta}_\gamma^s)^\mathsf{T}(\boldsymbol{\sigma}(s, \boldsymbol{\theta}^s) - \boldsymbol{\sigma}(\lfloor s \rfloor, \boldsymbol{\theta}_\gamma^{\lfloor s \rfloor}))\|_2^2\,\mathrm{d}s\right)^{1/2}\right] \\
&\leq C\,\mathbb{E}\left[\left(\int_0^T \|\boldsymbol{\theta}^s - \boldsymbol{\theta}_\gamma^s\|_2^2 \|\boldsymbol{\sigma}(s, \boldsymbol{\theta}^s) - \boldsymbol{\sigma}(\lfloor s \rfloor, \boldsymbol{\theta}_\gamma^{\lfloor s \rfloor})\|_{\mathrm{F}}^2\,\mathrm{d}s\right)^{1/2}\right] \\
&\leq C\,\mathbb{E}\left[\left(\sup_{0\leq t\leq T} \|\boldsymbol{\theta}^t - \boldsymbol{\theta}_\gamma^t\|_2^2\right)^{1/2}\left(\int_0^T (\gamma^2 d + \|\boldsymbol{\theta}^s - \boldsymbol{\theta}_\gamma^{\lfloor s \rfloor}\|_2^2)\,\mathrm{d}s\right)^{1/2}\right] \\
&\leq \frac{1}{2}\,\mathbb{E}\left[\sup_{0\leq t\leq T} \|\boldsymbol{\theta}^t - \boldsymbol{\theta}_\gamma^t\|_2^2\right] + C\int_0^T (\gamma^2 d + \mathbb{E}\|\boldsymbol{\theta}^s - \boldsymbol{\theta}_\gamma^{\lfloor s \rfloor}\|_2^2)\,\mathrm{d}s\,.
\end{aligned} \tag{441}$$

In the last line, we used Young's inequality. Therefore, we have

$$\mathbb{E}\left[\sup_{0\leq t\leq T} \|\boldsymbol{\theta}^t - \boldsymbol{\theta}_\gamma^t\|_2^2\right] \leq C\int_0^T (\gamma^2 d + \mathbb{E}\|\boldsymbol{\theta}^s - \boldsymbol{\theta}_\gamma^{\lfloor s \rfloor}\|_2^2)\,\mathrm{d}s + \frac{1}{2}\,\mathbb{E}\left[\sup_{0\leq t\leq T} \|\boldsymbol{\theta}^t - \boldsymbol{\theta}_\gamma^t\|_2^2\right]. \tag{442}$$

By rearranging the terms, we have

$$\mathbb{E}\left[\sup_{0\leq t\leq T} \|\boldsymbol{\theta}^t - \boldsymbol{\theta}_\gamma^t\|_2^2\right] \leq C\gamma^2 T d + C\int_0^T \mathbb{E}\|\boldsymbol{\theta}^s - \boldsymbol{\theta}_\gamma^{\lfloor s \rfloor}\|_2^2\,\mathrm{d}s\,. \tag{443}$$

Finally, we have $\|\boldsymbol{\theta}^s - \boldsymbol{\theta}_\gamma^{\lfloor s \rfloor}\|_2^2 \leq 2\|\boldsymbol{\theta}^s - \boldsymbol{\theta}_\gamma^s\|_2^2 + 2\|\boldsymbol{\theta}_\gamma^s - \boldsymbol{\theta}_\gamma^{\lfloor s \rfloor}\|_2^2$ and

$$\boldsymbol{\theta}_\gamma^s - \boldsymbol{\theta}_\gamma^{\lfloor s \rfloor} = \boldsymbol{b}(\lfloor s \rfloor, \boldsymbol{\theta}_\gamma^{\lfloor s \rfloor})(s - \lfloor s \rfloor) + \boldsymbol{\sigma}(\lfloor s \rfloor, \boldsymbol{\theta}_\gamma^{\lfloor s \rfloor})(\boldsymbol{W}^s - \boldsymbol{W}^{\lfloor s \rfloor})\,. \tag{444}$$

By the linear growth condition and Lemma G.2, we have

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\theta}_\gamma^s - \boldsymbol{\theta}_\gamma^{\lfloor s \rfloor}\|_2^2 &\leq C\gamma^2\,\mathbb{E}\|\boldsymbol{b}(\lfloor s \rfloor, \boldsymbol{\theta}_\gamma^{\lfloor s \rfloor})\|_2^2 + C\gamma\,\mathbb{E}\|\boldsymbol{\sigma}(\lfloor s \rfloor, \boldsymbol{\theta}_\gamma^{\lfloor s \rfloor})\|_{\mathrm{F}}^2 \\
&\leq C\gamma(d + \mathbb{E}\|\boldsymbol{\theta}_\gamma^{\lfloor s \rfloor}\|_2^2) \leq \mathrm{e}^{CT}\gamma(d + \|\boldsymbol{\theta}^0\|_2^2)\,.
\end{aligned} \tag{445}$$

Combining the above bounds, we have

$$\mathbb{E}\left[\sup_{0\leq t\leq T} \|\boldsymbol{\theta}^t - \boldsymbol{\theta}_\gamma^t\|_2^2\right] \leq \mathrm{e}^{CT}\gamma(Td + \|\boldsymbol{\theta}^0\|_2^2) + C\int_0^T \mathbb{E}\left[\sup_{0\leq s\leq t} \|\boldsymbol{\theta}^s - \boldsymbol{\theta}_\gamma^s\|_2^2\right]\mathrm{d}t\,. \tag{446}$$

Applying Grönwall's inequality, we obtain the bound (433). ■