Journal of Machine Learning Research 27 (2026) 1-78 Submitted 9/24; Revised 1/26; Published 2/26

Asymptotics of Stochastic Gradient Descent with Dropout Regularization in Linear Models

Jiaqi Li JQLI@UCHICAGO.EDU
Department of Statistics
University of Chicago
Chicago, IL, USA

Johannes Schmidt-Hieber A.J.SCHMIDT-HIEBER@UTWENTE.NL
Department of Applied Mathematics
University of Twente
Enschede, Netherlands

Wei Biao Wu WBWU@UCHICAGO.EDU
Department of Statistics
University of Chicago
Chicago, IL, USA

Editor: Michael Mahoney

Abstract

This paper proposes an asymptotic theory for online inference of the stochastic gradient descent (SGD) iterates with dropout regularization in linear regression. Specifically, we establish the geometric-moment contraction (GMC) for constant step-size SGD dropout iterates to show the existence of a unique stationary distribution of the dropout recursive function. Based on the GMC property, we use the functional dependence measure to provide quenched central limit theorems (CLT) for the gradient descent iterates with dropout regularization. Moreover, we obtain CLTs for the Ruppert-Polyak averaged GD (AGD) and averaged SGD (ASGD) iterates with dropout. Based on these asymptotic normality results, we further introduce an online estimator for the long-run covariance matrix of ASGD dropout to facilitate inference in a recursive manner with efficiency in computational time and memory. The numerical experiments demonstrate that for large samples, the proposed confidence intervals for ASGD with dropout achieve the nominal coverage probability.

Keywords: stochastic gradient descent, dropout regularization, ℓ^2-regularization, online inference, quenched central limit theorems

1. Introduction

Dropout regularization is a popular method in deep learning (Hinton et al., 2012; Krizhevsky et al., 2012; Srivastava et al., 2014). During each training iteration, each hidden unit is randomly masked with probability $1 - p$. This ensures that a hidden unit cannot rely on the presence of another hidden unit. Dropout therefore provides an incentive for different units to act more independently and avoids co-adaptation, which means that different units do the same.

There is a rich literature contributing to the theoretical understanding of dropout regularization. As pointed out in Srivastava et al. (2014), the core idea of dropout is to artificially

introduce stochasticity to the training process, preventing the model from learning statistical noise in the data. Starting with the connection of dropout and ℓ^2-regularization that appeared already in the original dropout article Srivastava et al. (2014), numerous works investigated the statistical properties of dropout by marginalizing the loss functions over dropout noises and linking them with explicit regularization (Arora et al., 2021; Baldi and Sadowski, 2013; Cavazza et al., 2018; McAllester, 2013; Mianjy and Arora, 2019; Mianjy et al., 2018; Senen-Cerda and Sanders, 2022; Srivastava et al., 2014; Wager et al., 2013). The empirical study in Wei et al. (2020) concluded that adding dropout noise to gradient descent also introduces implicit effects, which cannot be characterized by connections between the gradients of marginalized loss functions and explicit regularizers. For the linear regression model and fixed learning rates, Clara et al. (2024) proved that the implicit effect of dropout adds noise to the iterates and that for a large class of design matrices, this implicit noise does not vanish in the limit.

Though the convergence theory of dropout in fixed design and full gradients has been widely investigated, an analysis of dropout with random design or sequential observations is still lacking, not to mention online statistical inference. To bridge this gap, we provide a theoretical framework for dropout applied to stochastic gradient descent (SGD). In particular, we establish the geometric-moment contraction (GMC) for the SGD dropout iterates for a range of constant learning rates α. We provide useful and sharp moment inequalities to prove the q-th moment convergence of SGD dropout for any $q > 1$.

Besides the convergence and error bounds of SGD dropout, statistical inference of SGD-based estimators is also gaining attention (Fang, 2019; Fang et al., 2019; Liang and Su, 2019; Su and Zhu, 2023; Zhong et al., 2024). Instead of focusing on point estimators using dropout regularization, we quantify the uncertainty of the estimates through their confidence intervals or confidence regions (Chen et al., 2020; Zhu et al., 2023). Nevertheless, it is challenging to derive asymptotic normality for SGD dropout or its variants, such as averaged SGD (ASGD) (Ruppert, 1988; Polyak and Juditsky, 1992). The reason is that the SGD iterates are dependent and the initialization makes the SGD iterates non-stationary. In this paper, we leverage the GMC property of SGD dropout and show quenched central limit theorems (CLT) for both SGD and ASGD dropout estimates. Additionally, we propose an online estimator for the long-run covariance matrix of ASGD dropout to facilitate the online inference.

Contributions. This study employs powerful techniques from time series analysis to derive a general asymptotic theory for the SGD iterates with dropout regularization. Specifically, the key contributions can be summarized as follows.

(1) We establish the geometric-moment contraction (GMC) of the non-stationary SGD dropout iterates, whose recursion can be viewed as a vector auto-regressive process (VAR). The possible range of learning rates that ensures GMC can be related to the condition number of the design matrix with dropout.

(2) The GMC property guarantees the existence of a unique stationary distribution of the SGD iterates with dropout, and leads to the L^q-convergence, the asymptotic normality, and the Gaussian approximation rate of the SGD dropout estimates and their Ruppert-Polyak averaged version.

(3) We derive a new moment inequality in Lemma 27, proving that for any two random vectors $\boldsymbol{x}, \boldsymbol{y}$ of the same length, the q-th moment $\mathbb{E}\|\boldsymbol{x}+\boldsymbol{y}\|_2^q$ can have a sharp bound in terms of $\mathbb{E}\|\boldsymbol{x}\|_2^q$, $\mathbb{E}\|\boldsymbol{y}\|_2^q$ and $\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{y})$, without the condition $\mathbb{E}[\boldsymbol{y} \mid \boldsymbol{x}] \overset{\text{a.s.}}{=} 0$ required in previous results (Rio, 2009). The derived moment inequality is also applicable to many other L^q-convergence problems in machine learning.

(4) An online statistical inference method is introduced to construct joint confidence intervals for averaged SGD dropout iterates. The coverage probability is shown to be asymptotically accurate in theory and simulation studies.

The rest of the paper is organized as follows. We introduce the dropout regularization in gradient descent in Section 2. Followed by Section 3, we establish the geometric-moment contraction for dropout in gradient descent and provide the asymptotic normality. In Section 4, we generalize the theory to stochastic gradient descent. In Section 5, we provide an online inference algorithm for the ASGD dropout with theoretical guarantees. Finally, we present simulation studies in Section 6. All the technical proofs are postponed to the Appendix.

1.1 Background

Dropout regularization. After its introduction by Hinton et al. (2012); Srivastava et al. (2014), dropout regularization was found to be closely related to ℓ^2-regularization in linear regression and generalized linear models. See also Baldi and Sadowski (2013); McAllester (2013). Wager et al. (2013) extended this connection to more general injected forms of noise, showing that dropout induces an ℓ^2-penalty after rescaling the data by the estimated inverse diagonal Fisher information. In neural networks with a single hidden layer, dropout noise marginalization leads to a nuclear norm regularization, as studied in matrix factorization (Cavazza et al., 2018), linear neural networks (Mianjy et al., 2018), deep linear neural networks (Mianjy and Arora, 2019) and shallow ReLU-activated networks (Arora et al., 2021). Moreover, Gal and Ghahramani (2016b) showed that dropout can be interpreted as a variational approximation to the posterior of a Bayesian neural network. Gal and Ghahramani (2016a) applied this new variational inference based dropout technique in recurrent neural networks (RNN) and long-short term memory (LSTM) models. Additional research has explored the impact of dropout on convolutional neural networks (Wu and Gu, 2015) and generalization properties via Rademacher complexity bounds (Arora et al., 2021; Gao and Zhou, 2016; Wan et al., 2013; Zhai and Wang, 2018). Dropout has been successfully applied in various domains, including image classification (Krizhevsky et al., 2012), handwriting recognition (Pham and Le, 2021) and heart sound classification (Kay and Agarwal, 2016).

Stochastic gradient descent. To learn from huge datasets, stochastic gradient descent (SGD) (Robbins and Monro, 1951; Kiefer and Wolfowitz, 1952) is a computationally attractive variant of the gradient descent method. While dropout and SGD have been studied separately, only little theory has been developed so far for SGD training with dropout regularization. Mianjy and Arora (2020) showed the necessary number of SGD iterations to achieve suboptimality in ReLU shallow neural networks for classification tasks, which is independent of the dropout probability due to a strict condition on data structures. Senen-

Cerda and Sanders (2023) extended this to more generic results without assuming any specific data structures, focusing instead on reaching stationarity in non-convex functions using dropout-like SGD. Furthermore, Senen-Cerda and Sanders (2022) analyzed the gradient flow of dropout in shallow linear networks and studied the asymptotic convergence rate of dropout by marginalizing the dropout noise in a shallow network. However, a theoretical convergence analysis or inference theory of SGD dropout iterates without marginalization has not been explored yet in the literature.

1.2 Notation

We denote column vectors in $\mathbb{R}^d$ by lowercase bold letters, that is, $\boldsymbol{x} := (x_1, \ldots, x_d)^\top$ and write $\|\boldsymbol{x}\|_2 := (\boldsymbol{x}^\top \boldsymbol{x})^{1/2}$ for the Euclidean norm. The expectation and covariance of random vectors are respectively denoted by $\mathbb{E}[\cdot]$ and $\mathrm{Cov}(\cdot)$. For two positive number sequences (a_n) and (b_n), we say $a_n = O(b_n)$ (resp. $a_n \asymp b_n$) if there exists $c > 0$ such that $a_n/b_n \leq c$ (resp. $1/c \leq a_n/b_n \leq c$) for all large n, and say $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. Let (x_n) and (y_n) be two sequences of random variables. Write $x_n = O_{\mathbb{P}}(y_n)$ if for $\forall \epsilon > 0$, there exists $c > 0$ such that $\mathbb{P}(|x_n/y_n| \leq c) > 1 - \epsilon$ for all large n, and say $x_n = o_{\mathbb{P}}(y_n)$ if $x_n/y_n \to 0$ in probability as $n \to \infty$.

We denote matrices by uppercase letters. The $d \times d$ identity matrix is symbolized by I_d. Given matrices A and B of compatible dimension, their matrix product is denoted by juxtaposition. Write $A^\top$ for the transpose of A and define $\mathbb{A} := A^\top A$. When A and B are of the same dimension, the Hadamard product $A \odot B$ is given by element-wise multiplication $(A \odot B)_{ij} = A_{ij} B_{ij}$. For any $A \in \mathbb{R}^{d \times d}$, let $\mathrm{Diag}(A) := I_d \odot A$ denote the diagonal matrix with the same main diagonal as A. Given $p \in (0, 1)$, define the matrices

$$\begin{aligned} \overline{A} &:= A - \mathrm{Diag}(A), \\ A_p &:= pA + (1-p)\mathrm{Diag}(A). \end{aligned}$$

In particular, $A_p = p\overline{A} + \mathrm{Diag}(A)$, so A_p results from re-scaling the off-diagonal entries of A by p. For a matrix A, the operator norm induced by the Euclidean norm $\|\cdot\|_2$ is the spectral norm and will always be written without sub-script, that is, $\|A\| := \|A\|_{\mathrm{op}}$.

2. Dropout Regularization

The stochasticity of dropout makes it challenging to analyze the asymptotic properties of dropout in stochastic gradient descent. To address the complex stochastic structure, we investigate in Section 2 the dropout regularization in gradient descent, and then consider stochastic gradient descent in Section 4.1.

Consider a linear regression model with fixed design matrix $X \in \mathbb{R}^{n \times d}$ and outcome $\boldsymbol{y} \in \mathbb{R}^n$, that is,

$$\boldsymbol{y} = X\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \tag{1}$$

with unknown regression vector $\boldsymbol{\beta}^* \in \mathbb{R}^d$, and random noise $\boldsymbol{\epsilon} \in \mathbb{R}^n$. The task is to recover $\boldsymbol{\beta}^*$ from the observed data $(\boldsymbol{y}, X)$. Moreover, we suppose that $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = I_n$. We highlight that the noise distribution of $\boldsymbol{\epsilon}$ is often explicitly modeled as multivariate normal, but this is not necessary for this analysis. We also assume that the design matrix X has no zero columns. Because of that we also say that model (1) is in reduced form. We

Notation	Definition	Equation	Index Range
$\tilde{\boldsymbol{\beta}}_k$	GD iterates with dropout	Eq. (2)	$k \in \mathbb{N}$
$\tilde{\boldsymbol{\beta}}_k^\circ$	stationary GD iterates with dropout	Eq. (14)	$k \in \mathbb{Z}$
$\tilde{\boldsymbol{\beta}}$	ℓ^2 regularizer in the GD setting with dropout	Eq. (3)	/
$\boldsymbol{\beta}_k^\dagger$	stationary affine GD iterates with dropout	Eq. (15)	$k \in \mathbb{Z}$
$\breve{\boldsymbol{\beta}}_k$	SGD iterates with dropout	Eq. (32)	$k \in \mathbb{N}$
$\breve{\boldsymbol{\beta}}_k^\circ$	stationary SGD iterates with dropout	Eq. (42)	$k \in \mathbb{Z}$
$\breve{\boldsymbol{\beta}}$	ℓ^2 regularizer in the SGD setting with dropout	Eq. (34)	/
$\bar{\boldsymbol{\beta}}_k^{\mathrm{gd}}$	AGD iterates with dropout	Eq. (24)	$k \in \mathbb{N}$
$\bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}$	ASGD iterates with dropout	Eq. (48)	$k \in \mathbb{N}$

Table 1: List of the sequences defined in the paper.

can always bring the model into reduced form, since zero columns and the corresponding regression coefficients have no effect on the outcome $\boldsymbol{y}$ and can thus be eliminated from the model.

We consider the least-squares criterion $\frac{1}{2}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2$ for the estimation of $\boldsymbol{\beta}^*$. For the minimization, we adopt a constant learning-rate gradient descent algorithm with random dropouts in each iteration. Following the seminal work on dropout by Srivastava et al. (2014), we call a $d \times d$ random diagonal matrix D a p-dropout matrix if its diagonal entries satisfy $D_{ii} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$, with some retaining probability $p \in (0, 1)$. On average, D has pd diagonal entries equal to 1 and $(1-p)d$ diagonal entries equal to 0. For simplicity, the dependence of D on p will only be stated if unclear from the context. For a sequence of independent and identically distributed (i.i.d.) dropout matrices D_k, $k = 1, 2, \ldots$, and some constant learning rate $\alpha > 0$, the k-th step gradient descent iterate with dropout takes the form

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_k(\alpha) &= \tilde{\boldsymbol{\beta}}_{k-1}(\alpha) - \alpha \nabla_{\tilde{\boldsymbol{\beta}}_{k-1}} \frac{1}{2}\big\|\boldsymbol{y} - XD_k\tilde{\boldsymbol{\beta}}_{k-1}(\alpha)\big\|_2^2 \\ &= \tilde{\boldsymbol{\beta}}_{k-1}(\alpha) + \alpha D_k X^\top\big(\boldsymbol{y} - XD_k\tilde{\boldsymbol{\beta}}_{k-1}(\alpha)\big), \end{aligned} \tag{2}$$

with $\tilde{\boldsymbol{\beta}}_0$ a given initial vector. Taking the expectation with respect to D_k, one obtains the gradient descent iterates with respect to the loss function $\mathbb{E}[\frac{1}{2}\|\boldsymbol{y} - XD\boldsymbol{\beta}\|_2^2 | \boldsymbol{y}, X]$, where the expectation is taken only over the stochasticity in the dropout matrix D. The minimizer of the loss is denoted by

$$\tilde{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \ \mathbb{E}\Big[\frac{1}{2}\big\|\boldsymbol{y} - XD\boldsymbol{\beta}\big\|_2^2 \ \Big|\ \boldsymbol{y}, X\Big]. \tag{3}$$

Throughout Section 3, we assume that the sample $(\boldsymbol{y}, X)$ is given, and all the expectations are taken only over dropout matrix D, i.e., $\mathbb{E}[\cdot] = \mathbb{E}_D[\cdot]$. In fact, the vector $\tilde{\boldsymbol{\beta}}$ has a closed form expression. To see this, we denote the Gram matrix by

$$\mathbb{X} = X^\top X$$

and recall

$$\overline{\mathbb{X}} = \mathbb{X} - \mathrm{Diag}(\mathbb{X}), \quad \mathbb{X}_p = p\mathbb{X} + (1-p)\mathrm{Diag}(\mathbb{X}). \tag{4}$$

Note that $D^2 = D$, $\mathrm{Diag}(\mathbb{X}) = \mathbb{X}_p - p\overline{\mathbb{X}}$, and that diagonal matrices always commute. Since the fixed design matrix X is assumed to be in reduced form with $\min_i \mathbb{X}_{ii} > 0$, one can show that solving the gradient for the minimizer $\tilde{\boldsymbol{\beta}}$ in (3) (Srivastava et al., 2014; Clara et al., 2024) leads to the closed form expression

$$\tilde{\boldsymbol{\beta}} = p\Big(p^2\mathbb{X} + p(1-p)\mathrm{Diag}(\mathbb{X})\Big)^{-1} X^\top \boldsymbol{y} = \mathbb{X}_p^{-1} X^\top \boldsymbol{y}. \tag{5}$$

If the columns of X are orthogonal, then $\mathbb{X}$ is a diagonal matrix, $\mathbb{X}_p = \mathbb{X}$ and $\tilde{\boldsymbol{\beta}}$ coincides with the classical least-squares estimator $\mathbb{X}^{-1}X^\top \boldsymbol{y}$. We refer to Section 4.1 for a counterpart of $\tilde{\boldsymbol{\beta}}$ using stochastic gradient. While previous studies have linked ℓ^2-regularization and GD with dropout by marginalizing the training loss over the dropout noise (Wager et al. (2013); McAllester (2013); Srivastava et al. (2014)), Theorem 7 of Clara et al. (2024) uncovers a persistent gap in their covariance structures under a constant learning rate. This proves that dropout moreover incurs an implicit effect beyond what an explicit ℓ^2-penalty can capture. Consequently, it is essential to investigate the GD dropout iterates $\tilde{\boldsymbol{\beta}}_k(\alpha)$.

A crucial argument in the analysis of the dropout iterate $\tilde{\boldsymbol{\beta}}_k$ is to rewrite the dropout update formula as

$$\tilde{\boldsymbol{\beta}}_k(\alpha) - \tilde{\boldsymbol{\beta}} = \underbrace{(I_d - \alpha D_k \mathbb{X} D_k)}_{=:A_k(\alpha)}(\tilde{\boldsymbol{\beta}}_{k-1}(\alpha) - \tilde{\boldsymbol{\beta}}) + \underbrace{\alpha D_k \overline{\mathbb{X}}(pI_d - D_k)\tilde{\boldsymbol{\beta}}}_{=:\boldsymbol{b}_k(\alpha)}, \tag{6}$$

see Section 4.1 in Clara et al. (2024).

3. Asymptotic Properties of Dropout in GD

To study the asymptotic properties of gradient descent with dropout, we first establish the geometric-moment contraction for the GD dropout sequence. Subsequently, we derive the quenched central limit theorems for both iterative dropout estimates and their Ruppert-Polyak averaged variants. Furthermore, we provide the quenched invariance principle for the Ruppert-Polyak averaged dropout with the optimal Gaussian approximation rate.

3.1 Geometric-Moment Contraction (GMC)

First, we extend the geometric-moment contraction in Wu and Shao (2004) to the cases where the inputs of iterated random functions are i.i.d. random matrices.

Definition 1 (Geometric-moment contraction) *For i.i.d. $d \times d$ random matrices Ψ_i, Ψ'_j, $i, j \in \mathbb{Z}$, consider a stationary causal process*

$$\boldsymbol{\theta}_k = g(\Psi_k, \ldots, \Psi_1, \Psi_0, \Psi_{-1}, \ldots), \quad k \in \mathbb{Z}, \tag{7}$$

for a measurable function $g(\cdot)$ such that the d-dimensional random vector $\boldsymbol{\theta}_k$ has a finite q-th moment $\mathbb{E}\|\boldsymbol{\theta}_k\|_2^q < \infty$, for some $q \geq 1$. We say that $\boldsymbol{\theta}_k$ is geometric-moment contracting if there exists some constant $r_q \in (0,1)$ such that

$$\big(\mathbb{E}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}'_k\|_2^q\big)^{1/q} = O(r_q^k), \quad \textit{for all } k = 1, 2, \ldots, \tag{8}$$

where $\boldsymbol{\theta}'_k = g(\Psi_k, \ldots, \Psi_1, \Psi'_0, \Psi'_{-1}, \ldots)$ *is a coupled version of* $\boldsymbol{\theta}_k$ *with* Ψ_i, $i \leq 0$, *replaced by i.i.d. copies* Ψ'_i.

In general, an iterated random function satisfies the geometric-moment contraction property under regularity conditions on convexity and stochastic Lipschitz continuity, see Section B.1 in the Appendix for details. Here, we focus on the contraction property with $\Psi_k = D_k$, the k-th dropout matrix. Setting

$$f_D(\boldsymbol{u}) := \boldsymbol{u} + \alpha D X^\top (\boldsymbol{y} - XD\boldsymbol{u}),$$

we can rewrite the recursion of the dropout gradient descent iterate $\tilde{\boldsymbol{\beta}}_k(\alpha)$ in (2) as

$$\tilde{\boldsymbol{\beta}}_k(\alpha) = \tilde{\boldsymbol{\beta}}_{k-1}(\alpha) + \alpha D_k X^\top \big(\boldsymbol{y} - XD_k\tilde{\boldsymbol{\beta}}_{k-1}(\alpha)\big) =: f_{D_k}\big(\tilde{\boldsymbol{\beta}}_{k-1}(\alpha)\big). \tag{9}$$

We shall show that, under quite general conditions on the constant learning rate $\alpha > 0$, this process satisfies the geometric-moment contraction in Definition 1, and converges weakly to a unique stationary distribution π_α on $\mathbb{R}^d$, that is, for any continuous function $h \in \mathcal{C}(\mathbb{R}^d)$ with $\|h\|_\infty < \infty$, $\mathbb{E}\big[h\big(\tilde{\boldsymbol{\beta}}_k(\alpha)\big)\big] \to \int h(\boldsymbol{u})\pi_\alpha(d\boldsymbol{u})$ as $k \to \infty$. We then write $\tilde{\boldsymbol{\beta}}_k(\alpha) \Rightarrow \pi_\alpha$. Set

$$r_{\alpha,q} := \left(\sup_{\boldsymbol{v}\in\mathbb{R}^d:\|\boldsymbol{v}\|_2=1} \mathbb{E}\Big\|\big(I_d - \alpha D_1 \mathbb{X} D_1\big)\boldsymbol{v}\Big\|_2^q\right)^{1/q}. \tag{10}$$

In particular, for $q = 2$, we can rewrite the squared norm and obtain $r^2_{\alpha,2} = \lambda_{\max}(\mathbb{E}(I_d - \alpha D_1\mathbb{X}D_1)^2)$ with $\lambda_{\max}(\cdot)$ the largest eigenvalue. The exponential stability of online algorithms with constant step size has also been studied by Durmus et al. (2021) in the context of linear stochastic approximation and by Samsonov et al. (2024) to establish convergence of the stationary law in Wasserstein distance for temporal-difference learning. Theorem 3 below complements these results by proving contraction in the Euclidean distance to show the existence and uniqueness of the stationary distribution for GD with dropout and constant learning rate.

Lemma 2 *If* $q > 1$ *and* $\alpha\|\mathbb{X}\| < 2$, *then,* $r_{\alpha,q} < 1$.

As we only assumed that the design matrix X has no zero column, $\mathbb{X} = X^\top X$ can be singular. The previous lemma shows that even for singular $\mathbb{X}$, contraction coefficient $r_{\alpha,q} < 1$ is possible. Without dropout, for any $\boldsymbol{v}$ in the kernel of $\mathbb{X}$, one has $\|(I_d - \alpha\mathbb{X})\boldsymbol{v}\|_2 = \|\boldsymbol{v}\|_2$, implying that $\sup_{\boldsymbol{v}\in\mathbb{R}^d:\|\boldsymbol{v}\|_2=1} \mathbb{E}\|\big(I_d - \alpha\mathbb{X}\big)\boldsymbol{v}\|_2^q \geq 1$.

Theorem 3 (Geometric-moment contraction of GD dropout) *Let* $q > 1$. *Choose a positive learning rate* α *satisfying* $\alpha\|\mathbb{X}\| < 2$. *For two dropout sequences* $\tilde{\boldsymbol{\beta}}_k(\alpha), \tilde{\boldsymbol{\beta}}'_k(\alpha)$, $k = 0, 1, \ldots$, *generated by the recursion (6) with the same dropout matrices but possibly different initial vectors* $\tilde{\boldsymbol{\beta}}_0$, $\tilde{\boldsymbol{\beta}}'_0$, *we have*

$$\big(\mathbb{E}\|\tilde{\boldsymbol{\beta}}_k(\alpha) - \tilde{\boldsymbol{\beta}}'_k(\alpha)\|_2^q\big)^{1/q} \leq r^k_{\alpha,q}\|\tilde{\boldsymbol{\beta}}_0 - \tilde{\boldsymbol{\beta}}'_0\|_2. \tag{11}$$

Moreover, there exists a unique stationary distribution π_α *which does not depend on the initialization* $\tilde{\boldsymbol{\beta}}_0$, *such that* $\tilde{\boldsymbol{\beta}}_k(\alpha) \Rightarrow \pi_\alpha$ *as* $k \to \infty$.

Regarding the condition $\alpha\|\mathbb{X}\| < 2$, numerical experiments in Section 6.1 show that, for a wide range of settings, choosing the learning rate α slightly larger than $2/\|\mathbb{X}\|$, the empirical version of $r_{\alpha,q}$ will be larger than 1, and consequently, the GMC property in (11) will no longer hold. As we only require $\alpha\|\mathbb{X}\| < 2$, the dimension d influences the GMC property in the previous theorem only through the structure of the Gram matrix $\mathbb{X} = X^\top X$.

As mentioned before, $r_{\alpha,2}^2 = \lambda_{\max}(\mathbb{E}(I_d - \alpha D_1 \mathbb{X} D_1)^2) < 1$. A special case of Theorem 3 is thus $\mathbb{E}\|\tilde{\boldsymbol{\beta}}_k(\alpha) - \tilde{\boldsymbol{\beta}}_k'(\alpha)\|_2^2 \le (\lambda_{\max}(\mathbb{E}(I_d - \alpha D_1 \mathbb{X} D_1)^2))^k \|\tilde{\boldsymbol{\beta}}_0 - \tilde{\boldsymbol{\beta}}_0'\|_2^2$. Theorem 3 indicates that although the GD dropout sequence $\{\tilde{\boldsymbol{\beta}}_k(\alpha)\}_{k\in\mathbb{N}}$ is non-stationary due to the initialization, it is asymptotically stationary and approaches the unique stationary distribution π_α at an exponential rate. Such geometric-moment contraction result is fundamental to establish a central limit theorem for the iterates.

Another consequence of Theorem 3 is that if $\boldsymbol{\beta}$ is drawn from the stationary distribution π_α and D is an independently sampled dropout matrix, then also $f_D(\boldsymbol{\beta}) \sim \pi_\alpha$. We refer to Corollary 23 in the Appendix for the detailed arguments. This result means that if the initialization $\tilde{\boldsymbol{\beta}}_0^\circ$ is sampled from the stationary distribution π_α, then, the marginal distribution of any of the GD dropout iterates $\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)$ will follow this stationary distribution as well.

We can also define the GD dropout iterates $\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)$ for negative integers k by considering i.i.d. dropout matrices D_k for all integers $k \in \mathbb{Z}$ and observing that the limit

$$\tilde{\boldsymbol{\beta}}_k^\circ(\alpha) := \lim_{m\to\infty} f_{D_k} \circ f_{D_{k-1}} \circ \cdots \circ f_{D_{k-m}}(\boldsymbol{\beta}) =: h_\alpha(D_k, D_{k-1}, \ldots), \tag{12}$$

exists almost surely and does not depend on $\boldsymbol{\beta}$ (see Corollary 23). Then, it follows that $\tilde{\boldsymbol{\beta}}_k^\circ(\alpha) = f_{D_k}\big(\tilde{\boldsymbol{\beta}}_{k-1}^\circ(\alpha)\big)$ also holds for negative integers and the geometric-moment contraction in Definition 1 is satisfied for $\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)$, that is,

$$\Big(\mathbb{E}\big\|h_\alpha(D_k, \ldots, D_1, D_0, D_{-1}, \ldots) - h_\alpha(D_k, \ldots, D_1, D_0', D_{-1}', \ldots)\big\|_2^q\Big)^{1/q} = O(r_{\alpha,q}^k), \tag{13}$$

for some $q \ge 1$, $r_{\alpha,q} \in (0,1)$ defined in (10), and i.i.d. dropout matrices D_k, D_ℓ', $k, \ell \in \mathbb{Z}$.

3.2 Iterative Dropout Schemes

Equation (6) rewrites the GD dropout iterates into $\tilde{\boldsymbol{\beta}}_k(\alpha) - \tilde{\boldsymbol{\beta}} = A_k(\alpha)(\tilde{\boldsymbol{\beta}}_{k-1}(\alpha) - \tilde{\boldsymbol{\beta}}) + \boldsymbol{b}_k(\alpha)$. If the initial vector $\tilde{\boldsymbol{\beta}}_0^\circ$ is sampled from the stationary distribution π_α, we also have

$$\tilde{\boldsymbol{\beta}}_k^\circ(\alpha) - \tilde{\boldsymbol{\beta}} = A_k(\alpha)(\tilde{\boldsymbol{\beta}}_{k-1}^\circ(\alpha) - \tilde{\boldsymbol{\beta}}) + \boldsymbol{b}_k(\alpha), \tag{14}$$

and for any $k = 0, 1, \ldots$, $\tilde{\boldsymbol{\beta}}_k^\circ(\alpha) \sim \pi_\alpha$. We can see that $\{\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ is a stationary vector autoregressive process (VAR) with random coefficients. While $(A_k(\alpha), \boldsymbol{b}_k(\alpha))$ are i.i.d., $A_k(\alpha)$ and $\boldsymbol{b}_k(\alpha)$ are dependent. This poses challenges to prove asymptotic normality of the dropout iterates. An intermediate recursion is obtained by replacing $A_k(\alpha) = I_d - \alpha D_k \mathbb{X} D_k$ by its expectation $\mathbb{E}[A_k(\alpha)] = I_d - \alpha p \mathbb{X}_p$. This gives the recursion

$$\boldsymbol{\beta}_k^\dagger(\alpha) - \tilde{\boldsymbol{\beta}} = (I_d - \alpha p \mathbb{X}_p)\big(\boldsymbol{\beta}_{k-1}^\dagger(\alpha) - \tilde{\boldsymbol{\beta}}\big) + \boldsymbol{b}_k(\alpha). \tag{15}$$

The condition $\alpha \in (0, 2/\|\mathbb{X}\|)$ implies the operator norm bound $\|I_d - \alpha p \mathbb{X}_p\| < 1$ (see Step 1 in the proof of Lemma 4). Thus GMC holds and the sequence $\{\boldsymbol{\beta}_k^\dagger(\alpha)\}_{k\in\mathbb{N}}$ converges to

a unique stationary distribution as $k \to \infty$. For brevity, throughout the paper, we shall assume that the initialization $\boldsymbol{\beta}_0^\dagger$ follows the stationary distribution, and therefore the affine sequence $\{\boldsymbol{\beta}_k^\dagger(\alpha)\}_{k\in\mathbb{N}}$ is stationary. The proof then derives the asymptotic normality for $\boldsymbol{\beta}_k^\dagger(\alpha)$, and shows that the difference between $\tilde{\boldsymbol{\beta}}_k(\alpha)$ and $\boldsymbol{\beta}_k^\dagger(\alpha)$ is negligible, in the sense that for $q \geq 2$, $(\mathbb{E}\|\tilde{\boldsymbol{\beta}}_k(\alpha) - \boldsymbol{\beta}_k^\dagger(\alpha)\|_2^q)^{1/q} = O\big(\alpha + r_{\alpha,q}^k \|\tilde{\boldsymbol{\beta}}_0 - \tilde{\boldsymbol{\beta}}_0^\circ\|_2\big)$, where the first part is due to the affine approximation in Lemma 4 and the second part results from the GMC property in Theorem 3.

Lemma 4 (Affine approximation) *If $\alpha \in (0, 2/\|\mathbb{X}\|)$, then the difference sequence $\boldsymbol{\delta}_k(\alpha) = \tilde{\boldsymbol{\beta}}_k^\circ(\alpha) - \boldsymbol{\beta}_k^\dagger(\alpha)$ satisfies $\mathbb{E}[\boldsymbol{\delta}_k(\alpha)] = 0$ and for any $q \geq 2$, $\max_k \big(\mathbb{E}\|\boldsymbol{\delta}_k(\alpha)\|_2^q\big)^{1/q} = O(\alpha)$.*

Lemma 5 (Moment convergence of iterative GD dropout) *Let $q \geq 2$. For the stationary GD dropout sequence $\{\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ defined in (6), if $\alpha \in (0, 2/\|\mathbb{X}\|)$, we have*

$$\max_k \big(\mathbb{E}\|\tilde{\boldsymbol{\beta}}_k^\circ(\alpha) - \tilde{\boldsymbol{\beta}}\|_2^q\big)^{1/q} = O(\sqrt{\alpha}). \tag{16}$$

If the Gram matrix $\mathbb{X}$ is ill-conditioned in high dimensions, then the allowed range for α is more restrictive. While a smaller α improves the moment convergence rate $O(\sqrt{\alpha})$ by Lemma 5, it causes the contraction constant $r_{\alpha,q}$ to approach 1, thereby requiring a larger number of iterations k for $\tilde{\boldsymbol{\beta}}_k(\alpha)$ to converge to the stationary distribution π_α. If $\alpha \downarrow 0$, the stationary distribution π_α converges to the normal distribution in a sense that is formally stated in the next result. Recall that $\pi_\alpha(\mathcal{A}) = \mathbb{P}(\tilde{\boldsymbol{\beta}}_1^\circ(\alpha) \in \mathcal{A})$ is the stationary distribution defined in (14).

Theorem 6 (Quenched CLT of iterative GD dropout) *Consider the iterative gradient descent dropout sequence $\{\tilde{\boldsymbol{\beta}}_k(\alpha)\}_{k\in\mathbb{N}}$ in (6) and the ℓ^2-regularized estimator $\tilde{\boldsymbol{\beta}}$ in (5). Assume that the constant learning rate α satisfies $\alpha \in (0, 2/\|\mathbb{X}\|)$, $\tilde{\boldsymbol{\beta}} \neq 0$, and suppose that for every $l = 1, \ldots, d$, there exists $m \neq l$ such that $\mathbb{X}_{lm} \neq 0$. For any set $\mathcal{B} = (-\infty, b_1] \times \ldots \times (-\infty, b_d]$ with real numbers $b_1, \ldots, b_d$,*

$$\pi_\alpha\big(\tilde{\boldsymbol{\beta}} + \sqrt{\alpha}\Xi^{1/2}(\alpha)\mathcal{B}\big) \to \mathbb{P}\big(\boldsymbol{z} \in \mathcal{B}\big), \quad \text{as } \alpha \to 0, \tag{17}$$

with $\boldsymbol{z} \sim \mathcal{N}(0, I_d)$ following the standard d-variate normal distribution, and

$$\Xi(\alpha) := \mathrm{Cov}\Big(\frac{\boldsymbol{\beta}_1^\dagger(\alpha) - \tilde{\boldsymbol{\beta}}}{\sqrt{\alpha}}\Big) = \frac{\mathbb{E}[(\boldsymbol{\beta}_1^\dagger(\alpha) - \tilde{\boldsymbol{\beta}})(\boldsymbol{\beta}_1^\dagger(\alpha) - \tilde{\boldsymbol{\beta}})^\top]}{\alpha}. \tag{18}$$

In (17), the centering term is $\tilde{\boldsymbol{\beta}}$ since one can show that the affine sequence $\boldsymbol{\beta}_k^\dagger(\alpha)$ satisfies $\mathbb{E}[\boldsymbol{\beta}_k^\dagger(\alpha) - \tilde{\boldsymbol{\beta}}] = 0$ (see the proof of Theorem 6), which along with Lemma 4 yields $\mathbb{E}[\tilde{\boldsymbol{\beta}}_1^\circ(\alpha) - \tilde{\boldsymbol{\beta}}] = 0$. The result also implies convergence on all hyper-rectangles $(a_1, b_1] \times \ldots \times (a_d, b_d]$. One can derive more explicit expressions of $\Xi(\alpha)$ and $\Xi(0) := \lim_{\alpha\downarrow 0} \Xi(\alpha)$. Reshaping a $d \times s$ matrix $U = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_s)$ with d-dimensional column vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_s$ into a ds-dimensional column vector gives $\mathrm{vec}(U) := (\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_s^\top)^\top$. Moreover, for any two matrices $A \in \mathbb{R}^{p\times q}$ and $B \in \mathbb{R}^{m\times n}$, the Kronecker product $A \otimes B$ is the $pm \times qn$ block matrix, with each block given by $(A \otimes B)_{ij} = A_{ij}B$. Following Theorem 1 in Pflug (1986),

and assuming that $\Xi(\alpha)$ is differentiable with respect to α, $\Xi(\alpha)$ becomes the solution of a classical Lyapunov equation

$$\Xi(\alpha)(p\mathbb{X}_p) + (p\mathbb{X}_p)\Xi(\alpha) = S,$$

that is,

$$\Xi(\alpha) = V_0 + \alpha B_p, \tag{19}$$

where the $d \times d$ matrices S, V_0 and B_p are respectively defined as

$$S = \frac{1}{\alpha^2}\mathrm{Cov}\big(\boldsymbol{b}_1(\alpha)\big) = \mathrm{Cov}\big(D_1\overline{\mathbb{X}}(pI_d - D_1)\tilde{\boldsymbol{\beta}}\big), \tag{20}$$

$$\mathrm{vec}(V_0) = (I_d \otimes p\mathbb{X}_p + p\mathbb{X}_p \otimes I_d)^{-1} \cdot \mathrm{vec}(S), \tag{21}$$

$$\mathrm{vec}(B_p) = (I_d \otimes p\mathbb{X}_p + p\mathbb{X}_p \otimes I_d)^{-1} \cdot \mathrm{vec}(p^2\mathbb{X}_p V_0 \mathbb{X}_p). \tag{22}$$

The matrix V_0 satisfies the equation $V_0 p\mathbb{X}_p + p\mathbb{X}_p V_0 = S$. By definition, the matrix S is independent of α and $\overline{\mathbb{X}} \neq 0$ since there exist non-zero diagonal and off-diagonal elements by assumptions in Theorem 6. Let $S_0 = \tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top$. By the proof of Theorem 6, we can express S in terms of p, X and $\tilde{\boldsymbol{\beta}}$ as follows,

$$\begin{aligned} S = {} & p^3(\overline{\mathbb{X}}S_0\overline{\mathbb{X}})_p - 2p\Big(p\overline{\mathbb{X}}_p(S_0\overline{\mathbb{X}})_p + p^2(1-p)\mathrm{Diag}(\overline{\mathbb{X}}S_0\overline{\mathbb{X}})\Big) \\ & + p\overline{\mathbb{X}}_p(S_0)_p\overline{\mathbb{X}}_p + p^2(1-p)\Big(\mathrm{Diag}(\overline{\mathbb{X}}(S_0)_p\overline{\mathbb{X}}) + 2\overline{\mathbb{X}}_p\mathrm{Diag}(\overline{S_0\mathbb{X}}) + (1-p)\overline{\mathbb{X}} \odot \overline{S_0}^\top \odot \overline{\mathbb{X}}\Big). \end{aligned} \tag{23}$$

One can see that, $\Xi(0) = \lim_{\alpha\downarrow 0}\Xi(\alpha) = V_0$, and in particular, for small p, $\mathrm{vec}(V_0)$ can be approximated by $(I_d \otimes \mathbb{X}_p + \mathbb{X}_p \otimes I_d)^{-1} \cdot \mathrm{vec}(\overline{\mathbb{X}}_p[\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top]_p\overline{\mathbb{X}}_p)$. By the expansion in (19), the CLT in (17) still holds with $\Xi(\alpha)$ therein replaced by $\Xi(0) = V_0$.

By Theorems 3 and 6, the GD dropout iterates $\tilde{\boldsymbol{\beta}}_k(\alpha)$ exhibit asymptotic normality via a two-step limit. First, for any fixed constant learning rate $\alpha > 0$, the GMC property (cf. Theorem 3) guarantees that the GD dropout iterates will first converge to the stationary distribution π_α as the iteration number k grows. Decreasing then the learning rate $\alpha \to 0$, π_α converges to the normal distribution (cf. Theorem 6). Similar two-step schemes are also explored in the literature; for example, Defazio et al. (2024) proposed a modern learning schedule which consists of an initial warm-up stage with relatively large learning rates followed by rapid learning rate annealing near the end of training.

3.3 Dropout with Ruppert-Polyak Averaging

To reduce the variance of the gradient descent iterates $\tilde{\boldsymbol{\beta}}_k(\alpha)$ introduced by the random dropout matrix D_k, we now consider the averaged GD dropout (AGD) iterate

$$\bar{\boldsymbol{\beta}}_k^{\mathrm{gd}}(\alpha) = \frac{1}{k}\sum_{i=1}^{k}\tilde{\boldsymbol{\beta}}_i(\alpha), \tag{24}$$

following the averaging scheme in Ruppert (1988); Polyak and Juditsky (1992). We derive the asymptotic normality of $\bar{\boldsymbol{\beta}}_k^{\mathrm{gd}}(\alpha)$ in the following theorem.

Theorem 7 (Quenched CLT of averaged GD dropout) *For a constant learning rate $\alpha \in (0, 2/\|\mathbb{X}\|)$ and any fixed initial vector $\tilde{\boldsymbol{\beta}}_0$, the averaged GD dropout sequence satisfies*

$$\sqrt{k}\big(\bar{\boldsymbol{\beta}}_k^{\mathrm{gd}}(\alpha) - \tilde{\boldsymbol{\beta}}\big) \Rightarrow \mathcal{N}\big(0, \Sigma(\alpha)\big), \quad \text{as } k \to \infty, \tag{25}$$

with $\Sigma(\alpha) = \sum_{i=-\infty}^{\infty} \mathrm{Cov}(\tilde{\boldsymbol{\beta}}_0^{\circ}(\alpha), \tilde{\boldsymbol{\beta}}_i^{\circ}(\alpha))$ the long-run covariance matrix of the stationary process $\tilde{\boldsymbol{\beta}}_i^{\circ}(\alpha) \sim \pi_\alpha$.

One can choose a few learning rates, say $\alpha_1, \ldots, \alpha_s$, and run gradient descent for each of these learning rates in parallel by computing $\tilde{\boldsymbol{\beta}}_k(\alpha_1), \ldots, \tilde{\boldsymbol{\beta}}_k(\alpha_s)$ for $k = 1, 2, \ldots$. An example is federated learning where data are distributed across different clients (Dean et al., 2012; Karimireddy et al., 2020; Zinkevich et al., 2010).

Corollary 8 (Quenched CLT of parallel averaged GD dropout) *Let $s \geq 1$. Consider constant learning rates $\alpha_1, \ldots, \alpha_s \in (0, 2/\|\mathbb{X}\|)$. Then, for any initial vectors $\tilde{\boldsymbol{\beta}}_0(\alpha_1)$, $\ldots, \tilde{\boldsymbol{\beta}}_0(\alpha_s)$,*

$$\sqrt{k} \cdot \mathrm{vec}\big(\bar{\boldsymbol{\beta}}_k^{\mathrm{gd}}(\alpha_1) - \tilde{\boldsymbol{\beta}}, \ldots, \bar{\boldsymbol{\beta}}_k^{\mathrm{gd}}(\alpha_s) - \tilde{\boldsymbol{\beta}}\big) \Rightarrow \mathcal{N}(0, \Sigma^{\mathrm{vec}}), \quad \text{as } k \to \infty, \tag{26}$$

with $\mathrm{vec}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_s) = (\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_s^\top)^\top \in \mathbb{R}^{ds}$ for d-dimensional vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_s$, and the long-run covariance matrix

$$\Sigma^{\mathrm{vec}} = \sum_{i=-\infty}^{\infty} \mathrm{Cov}\Big(\mathrm{vec}\big(\tilde{\boldsymbol{\beta}}_0^{\circ}(\alpha_1), \ldots, \tilde{\boldsymbol{\beta}}_0^{\circ}(\alpha_s)\big), \mathrm{vec}\big(\tilde{\boldsymbol{\beta}}_i^{\circ}(\alpha_1), \ldots, \tilde{\boldsymbol{\beta}}_i^{\circ}(\alpha_s)\big)\Big).$$

Next, we provide a stronger Gaussian approximation result, namely the rate for the Komlós–Major–Tusnády (KMT) approximation (Komlós et al., 1975, 1976; Berkes et al., 2014). In the quenched invariance principle below, we show that one can achieve the optimal Gaussian approximation rate for the averaged GD dropout process.

Theorem 9 (Quenched invariance principle of averaged GD dropout) *Assume the constant learning rate satisfies $\alpha \in (0, 2/\|\mathbb{X}\|)$. Define a partial sum process $(S_i^{\circ}(\alpha))_{1 \leq i \leq t}$ for $t \in \mathbb{N}_+$ with*

$$S_i^{\circ}(\alpha) = \sum_{k=1}^{i} (\tilde{\boldsymbol{\beta}}_k^{\circ}(\alpha) - \tilde{\boldsymbol{\beta}}). \tag{27}$$

Then, there exists a (richer) probability space $(\Omega^\star, \mathcal{A}^\star, \mathbb{P}^\star)$ on which one can define d-dimensional random vectors $\tilde{\boldsymbol{\beta}}_k^\star$, the associated partial sum process $S_i^\star(\alpha) = \sum_{k=1}^{i} (\tilde{\boldsymbol{\beta}}_k^\star(\alpha) - \tilde{\boldsymbol{\beta}})$, and a Gaussian process $G_i^\star = \sum_{k=1}^{i} \boldsymbol{z}_k^\star$, with independent Gaussian random vectors $\boldsymbol{z}_k^\star \sim \mathcal{N}(0, I_d)$, such that $(S_i^{\circ})_{1 \leq i \leq t} \overset{\mathcal{D}}{=} (S_i^\star)_{1 \leq i \leq t}$ and

$$\max_{1 \leq i \leq t} \big\| S_i^\star - \Sigma^{1/2}(\alpha) G_i^\star \big\|_2 = o_{\mathbb{P}}(t^{1/q}), \quad \text{in } (\Omega^\star, \mathcal{A}^\star, \mathbb{P}^\star), \tag{28}$$

where $\Sigma(\alpha)$ is the long-run covariance matrix defined in Theorem 7. In addition, this approximation holds for all $(S_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha))_{1 \leq i \leq t}$ given any arbitrary initial vector $\tilde{\boldsymbol{\beta}}_0 \in \mathbb{R}^d$, where

$$S_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha) = \sum_{k=1}^{i} (\tilde{\boldsymbol{\beta}}_k(\alpha) - \tilde{\boldsymbol{\beta}}). \tag{29}$$

Theorem 9 shows that one can approximate the averaged GD dropout sequence by Brownian motions. Specifically, for any fixed initial vector $\tilde{\boldsymbol{\beta}}_0 \in \mathbb{R}^d$, the partial sum process converges in the Euclidean norm, uniformly over u,

$$\big\{t^{-1/2} S^{\tilde{\boldsymbol{\beta}}_0}_{\lfloor tu \rfloor}(\alpha),\ 0 \le u \le 1\big\} \Rightarrow \big\{\Sigma^{1/2}(\alpha)\mathbb{B}(u),\ 0 \le u \le 1\big\}, \tag{30}$$

where $\lfloor m \rfloor = \max\{i \in \mathbb{Z} : i \le m\}$, and $\mathbb{B}(u)$ is the standard d-dimensional Brownian motion, that is, it can be represented as a d-dimensional vector of independent standard Brownian motions. According to the arguments in Karmakar and Wu (2020), the KMT approximation rate $o_{\mathbb{P}}(t^{1/q})$ is optimal for fixed-dimension time series. Since we can view the GD dropout sequence as a VAR(1) process, the approximation rate in Theorem 9 is optimal for the partial sum process $(S_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha))_{1 \le i \le t}$.

4. Asymptotic Properties of Dropout in SGD

In the previous section, we considered a fixed design matrix and (full) gradient descent with dropout. Computing the gradient over the entire dataset can be computationally expensive, especially with large datasets. We now investigate stochastic gradient descent with dropout regularization.

4.1 Dropout Regularization in SGD

Consider i.i.d. covariate vectors $\boldsymbol{x}_k \in \mathbb{R}^d$, $k = 1, 2, \ldots$, from some distribution Π, and the realizations $y_k|\boldsymbol{x}_k$ from a linear regression model

$$y_k = \boldsymbol{x}_k^\top \boldsymbol{\beta}^* + \epsilon_k, \tag{31}$$

with unknown regression vector $\boldsymbol{\beta}^* \in \mathbb{R}^d$. We assume that the model is in reduced form, which here means that $\min_i(\mathbb{E}[\boldsymbol{x}_1 \boldsymbol{x}_1^\top])_{ii} > 0$. In addition, we assume that the i.i.d. random noises ϵ_k satisfy $\mathbb{E}[\epsilon_k] = 0$ and $\mathrm{Var}(\epsilon_k) = 1$. In this paper, we focus on the classical case where the SGD computes the gradient based on an individual observation $(y_k, \boldsymbol{x}_k)$ and a dropout matrix D_k. For constant learning rate α and initialization $\breve{\boldsymbol{\beta}}_0$, the k-th step SGD iterate with Bernoulli dropout is

$$\begin{aligned}\breve{\boldsymbol{\beta}}_k(\alpha) &= \breve{\boldsymbol{\beta}}_{k-1}(\alpha) - \alpha \nabla_{\breve{\boldsymbol{\beta}}_{k-1}} \frac{1}{2}\big(y_k - \boldsymbol{x}_k^\top D_k \breve{\boldsymbol{\beta}}_{k-1}(\alpha)\big)^2 \\ &= \breve{\boldsymbol{\beta}}_{k-1}(\alpha) + \alpha D_k \boldsymbol{x}_k \big(y_k - \boldsymbol{x}_k^\top D_k \breve{\boldsymbol{\beta}}_{k-1}(\alpha)\big).\end{aligned} \tag{32}$$

This is a sequential estimation, or online learning scheme, as computing $\breve{\boldsymbol{\beta}}_k(\alpha)$ from $\breve{\boldsymbol{\beta}}_{k-1}(\alpha)$ only requires the k-th sample $(y_k, \boldsymbol{x}_k)$ and the dropout matrix D_k. To study the contraction property of the SGD dropout iterates $\breve{\boldsymbol{\beta}}_k(\alpha)$, we express the recursion (32) by an iterated random function $\breve{f} : \mathbb{R}^{d\times d} \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}^d$ with

$$f_{D,(y,\boldsymbol{x})}(\boldsymbol{u}) = \boldsymbol{u} + \alpha D \boldsymbol{x}(y - \boldsymbol{x} D \boldsymbol{u}),$$

that is,

$$\begin{aligned}\breve{\boldsymbol{\beta}}_k(\alpha) &= \breve{\boldsymbol{\beta}}_{k-1}(\alpha) + \alpha D_k \boldsymbol{x}_k \big(y_k - \boldsymbol{x}_k^\top D_k \breve{\boldsymbol{\beta}}_{k-1}(\alpha)\big) \\ &=: \breve{f}_{D_k,(y_k,\boldsymbol{x}_k)}(\breve{\boldsymbol{\beta}}_{k-1}(\alpha)).\end{aligned} \tag{33}$$

We shall show that this iterated random function $\breve{f}$ is geometrically contracting under suitable conditions, and therefore, there exists a unique stationary distribution $\breve{\pi}_\alpha$ such that $\breve{\boldsymbol{\beta}}_k(\alpha) \Rightarrow \breve{\pi}_\alpha$, where $\Rightarrow$ denotes the convergence in distribution.

From now on, let $(y, \boldsymbol{x})$ be a sample with the same distribution as $(y_k, \boldsymbol{x}_k)$. By marginalizing over all randomness, we can view the SGD dropout in (32) as a minimizer of the ℓ^2-regularized least-squares loss

$$\breve{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \mathbb{E}_{(y,\boldsymbol{x})}\mathbb{E}_D\Big[\frac{1}{2}(y - \boldsymbol{x}^\top D\boldsymbol{\beta})^2\Big]. \tag{34}$$

Here, the expectation is take over both the random sample $(y, \boldsymbol{x})$ and the dropout matrix D. Throughout the rest of the paper, we shall write $\mathbb{E}[\cdot] = \mathbb{E}_{(y,\boldsymbol{x})}\mathbb{E}_D[\cdot]$ when no confusion should be caused.

Denote the $d \times d$ Gram matrix by $\mathbb{X}_k = \boldsymbol{x}_k\boldsymbol{x}_k^\top$, and define

$$\overline{\mathbb{X}}_k = \mathbb{X}_k - \mathrm{Diag}(\mathbb{X}_k), \quad \mathbb{X}_{k,p} = p\mathbb{X}_k + (1-p)\mathrm{Diag}(\mathbb{X}_k). \tag{35}$$

By Lemma 28 in the Appendix, we have a closed form solution for $\breve{\boldsymbol{\beta}}$ as follows

$$\breve{\boldsymbol{\beta}} = p\Big(p^2\mathbb{E}[\boldsymbol{x}_1\boldsymbol{x}_1^\top] + p(1-p)\mathrm{Diag}(\mathbb{E}[\boldsymbol{x}_1\boldsymbol{x}_1^\top])\Big)^{-1}\mathbb{E}[y_1\boldsymbol{x}_1] = (\mathbb{E}[\mathbb{X}_{1,p}])^{-1}\mathbb{E}[y_1\boldsymbol{x}_1],$$

and thus, we obtain the relationship $\mathbb{E}[\mathbb{X}_{1,p}]\breve{\boldsymbol{\beta}} = \mathbb{E}[y_1\boldsymbol{x}_1]$. To study the SGD with dropout, we now focus on the difference process $\{\breve{\boldsymbol{\beta}}_k(\alpha) - \breve{\boldsymbol{\beta}}\}_{k\in\mathbb{N}}$. As in the case of gradient descent, this process can be written in autoregressive form,

$$\breve{\boldsymbol{\beta}}_k(\alpha) - \breve{\boldsymbol{\beta}} = \underbrace{\big(I_d - \alpha D_k\mathbb{X}_k D_k\big)}_{=:\breve{A}_k(\alpha)}(\breve{\boldsymbol{\beta}}_{k-1}(\alpha) - \breve{\boldsymbol{\beta}}) + \underbrace{\alpha D_k\boldsymbol{x}_k(y_k - \boldsymbol{x}_k^\top D_k\breve{\boldsymbol{\beta}})}_{=:\breve{\boldsymbol{b}}_k(\alpha)}. \tag{36}$$

4.2 GMC of Dropout in SGD

Establishing the geometric-moment contraction (GMC) property to the stochastic gradient descent iterates with dropout is non-trivial as the randomness of $\breve{\boldsymbol{\beta}}_k(\alpha)$ not only comes from the dropout matrix D_k, but also the random sample $(y_k, \boldsymbol{x}_k)$. Recall that $\mathbb{X}_{k,p}$ is defined in (35) and by Lemma 21(ii), $\mathbb{E}[D\mathbb{X}_k D] = p\mathbb{E}[\mathbb{X}_{k,p}]$.

Lemma 10 (Learning-rate range in SGD dropout) *Assume that for some $q \geq 2$, the q-th moment $\mu_q(\boldsymbol{v}) = (\mathbb{E}\|D_k\mathbb{X}_k D_k\boldsymbol{v}\|_2^q)^{1/q} < \infty$ for all unit vectors $\boldsymbol{v} \in \mathbb{R}^d$. If the learning rate $\alpha > 0$ satisfies*

$$\frac{\alpha(q-1)}{2}\sup_{\boldsymbol{v}\in\mathbb{R}^d,\,\|\boldsymbol{v}\|_2=1}\frac{(1+\alpha\mu_q(\boldsymbol{v}))^{q-2}\mu_q(\boldsymbol{v})^2}{p\boldsymbol{v}^\top\mathbb{E}[\mathbb{X}_{k,p}]\boldsymbol{v}} < 1, \tag{37}$$

then, for a dropout matrix D,

$$\breve{r}_{\alpha,q}^q := \sup_{\boldsymbol{v}\in\mathbb{R}^d,\|\boldsymbol{v}\|_2=1}\mathbb{E}\big\|(I_d - \alpha D\mathbb{X}_k D)\boldsymbol{v}\big\|_2^q < 1. \tag{38}$$

This provides a sufficient condition for the learning rate α which ensures contraction of $I_d - \alpha D\mathbb{X}_k D$ for moments $q \geq 2$. This will lead to L^q-convergence of the SGD dropout iterates and determines the convergence rate in the Gaussian approximation in Theorem 17.

For the special and important case $q = 2$, the identities $\mu_2(\boldsymbol{v})^2 = \mathbb{E}\|D_k\mathbb{X}_k D_k\boldsymbol{v}\|_2^2$ and $\mathbb{E}(D_k\mathbb{X}_k D_k) = p\mathbb{E}[\mathbb{X}_{k,p}]$ imply that condition (37) can be rewritten into

$$\frac{\alpha}{2} \sup_{\boldsymbol{v}\in\mathbb{R}^d, \|\boldsymbol{v}\|_2=1} \frac{\mu_2(\boldsymbol{v})^2}{p\boldsymbol{v}^\top \mathbb{E}[\mathbb{X}_{k,p}]\boldsymbol{v}} < 1, \tag{39}$$

and

$$0 < \alpha < \inf_{\boldsymbol{v}\in\mathbb{R}^d, \|\boldsymbol{v}\|_2=1} \frac{2\boldsymbol{v}^\top \mathbb{E}(D_k\mathbb{X}_k D_k)\boldsymbol{v}}{\mathbb{E}\|D_k\mathbb{X}_k D_k\boldsymbol{v}\|_2^2}. \tag{40}$$

For $q = 2$, Lemma 29 in the Appendix states that the conclusion of the previous lemma is also implied by the condition $\mathbb{E}[2\mathbb{X}_k - \alpha\mathbb{X}_k^2] > 0$.

Remark 11 *The upper bound in condition* (40) *can be interpreted as a generalized eigenvalue. For a given pair of real symmetric matrices* (A, B) *and a given non-zero real vector* $\boldsymbol{u}$*, the generalized Rayleigh quotient is defined as*

$$R(A, B; \boldsymbol{u}) := \frac{\boldsymbol{u}^\top A\boldsymbol{u}}{\boldsymbol{u}^\top B\boldsymbol{u}}.$$

One can show that $\sup_{\boldsymbol{u}\neq 0} R(A, B; \boldsymbol{u}) = \max\{\lambda : \det(A - \lambda B) = 0\}$ *is the largest generalized eigenvalue of* (A, B)*. Recall* $\mathbb{X}_k = \boldsymbol{x}_k\boldsymbol{x}_k^\top$*, let* $\tilde{\boldsymbol{x}}_k = D_k\boldsymbol{x}_k$*, and define the matrices*

$$\Sigma_1 := \mathbb{E}(D_k\mathbb{X}_k D_k) = \mathbb{E}(\tilde{\boldsymbol{x}}_k\tilde{\boldsymbol{x}}_k^\top),\ \Sigma_2 := \mathbb{E}(D_k\mathbb{X}_k D_k\mathbb{X}_k D_k) = \mathbb{E}(\tilde{\boldsymbol{x}}_k\tilde{\boldsymbol{x}}_k^\top\tilde{\boldsymbol{x}}_k\tilde{\boldsymbol{x}}_k^\top) = \mathbb{E}(\tilde{\boldsymbol{x}}_k\tilde{\boldsymbol{x}}_k^\top)^2.$$

Letting $\boldsymbol{u} = \boldsymbol{v}$*,* $A = \Sigma_2$ *and* $B = \Sigma_1$*, condition* (40) *is equivalent to*

$$0 < \alpha < \inf_{\boldsymbol{v}\in\mathbb{R}^d, \|\boldsymbol{v}\|_2=1} \frac{2\boldsymbol{v}^\top\Sigma_1\boldsymbol{v}}{\boldsymbol{v}^\top\Sigma_2\boldsymbol{v}} = \frac{2}{\sup_{\boldsymbol{v}} R(\Sigma_2, \Sigma_1; \boldsymbol{v})}.$$

For the geometric-moment contraction for the SGD dropout sequence, we impose the following moment conditions.

Assumption 1 (Finite moment) *Assume that for some* $q \geq 2$*, the random noises* ϵ *and the random sample* $\boldsymbol{x}$ *in model (31) have finite* $2q$*-th moment* $\mathbb{E}[|\epsilon|^{2q}] + \|\boldsymbol{x}\|_2^{2q}] < \infty$.

Lemma 30 in the Appendix shows that this assumption ensures the finite q-th moment of the stochastic gradient in (32) evaluated at the true parameter $\boldsymbol{\beta}^*$ and the ℓ^2-minimizer $\breve{\boldsymbol{\beta}}$ in model (31), that is,

$$\Big(\mathbb{E}\Big\|\nabla_{\boldsymbol{\beta}^*}\frac{1}{2}\big(y - \boldsymbol{x}^\top D\boldsymbol{\beta}^*\big)^2\Big\|_2^q\Big)^{1/q} = \Big(\mathbb{E}\big\|D\boldsymbol{x}\big(y - \boldsymbol{x}^\top D\boldsymbol{\beta}^*\big)\big\|_2^q\Big)^{1/q} < \infty,$$

and $(\mathbb{E}\|\nabla_{\breve{\boldsymbol{\beta}}}\frac{1}{2}\big(y - \boldsymbol{x}^\top D\breve{\boldsymbol{\beta}}\big)^2\|_2^q)^{1/q} < \infty$. Assumption 1 moreover guarantees $(\mathbb{E}\|\nabla_{\boldsymbol{\beta}}\frac{1}{2}\big(y - \boldsymbol{x}^\top D\boldsymbol{\beta}\big)^2\|_2^q)^{1/q} < \infty$ for all $\boldsymbol{\beta} \in \mathbb{R}^d$.

Theorem 12 (Geometric-moment contraction of SGD dropout) *Let $q \geq 2$. Suppose that Assumption 1 holds and the learning rate α satisfies (37). For two dropout sequences $\breve{\boldsymbol{\beta}}_k(\alpha)$ and $\breve{\boldsymbol{\beta}}'_k(\alpha)$, $k = 0, 1, \ldots$, that are generated by the recursion (32) with the same sequence of dropout matrices $\{D_k\}_{k\in\mathbb{N}}$ but possibly different initializations $\breve{\boldsymbol{\beta}}_0$, $\breve{\boldsymbol{\beta}}'_0$, we have*

$$\Big(\mathbb{E}\big\|\breve{\boldsymbol{\beta}}_k(\alpha) - \breve{\boldsymbol{\beta}}'_k(\alpha)\big\|_2^q\Big)^{1/q} \leq \breve{r}_{\alpha,q}^k \|\breve{\boldsymbol{\beta}}_0 - \breve{\boldsymbol{\beta}}'_0\|_2, \quad \textit{for all } k = 1, 2, \ldots, \tag{41}$$

with

$$\breve{r}_{\alpha,q} = \Big(\sup_{\boldsymbol{v}\in\mathbb{R}^d:\, \|\boldsymbol{v}\|_2=1} \mathbb{E}\big\|\breve{A}_1(\alpha)\boldsymbol{v}\big\|_2^q\Big)^{1/q} < 1.$$

Moreover, for any initial vector $\breve{\boldsymbol{\beta}}_0 \in \mathbb{R}^d$, there exists a unique stationary distribution $\breve{\pi}_\alpha$ which does not depend on $\breve{\boldsymbol{\beta}}_0$, such that $\breve{\boldsymbol{\beta}}_k(\alpha) \Rightarrow \breve{\pi}_\alpha$ as $k \to \infty$.

By Theorem 12, initializing $\breve{\boldsymbol{\beta}}_0^\circ \sim \breve{\pi}_\alpha$ leads to the stationary SGD dropout sequence $\{\breve{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ by following the recursion

$$\breve{\boldsymbol{\beta}}_k^\circ(\alpha) - \breve{\boldsymbol{\beta}} = \breve{A}_k(\alpha)(\breve{\boldsymbol{\beta}}_{k-1}^\circ(\alpha) - \breve{\boldsymbol{\beta}}) + \breve{\boldsymbol{b}}_k(\alpha), \quad k = 1, 2, \ldots, \tag{42}$$

where the ℓ^2-regularized minimizer $\breve{\boldsymbol{\beta}}$ is defined in (34), and the random coefficients $\breve{A}_k(\alpha) = I_d - \alpha D_k \mathbb{X}_k D_k$ and $\breve{\boldsymbol{b}}_k(\alpha) = \alpha D_k \boldsymbol{x}_k(y_k - \boldsymbol{x}_k^\top D_k \breve{\boldsymbol{\beta}})$ are defined in (36). Furthermore, recall the iterated random function $\breve{f}_{D,(y,\boldsymbol{x})}(\boldsymbol{\beta})$ defined in (33). As a direct consequence of Theorem 12, we have

$$\breve{\boldsymbol{\beta}}_k^\circ(\alpha) = \breve{f}_{D_k,(y_k,\boldsymbol{x}_k)}(\breve{\boldsymbol{\beta}}_{k-1}^\circ(\alpha)), \tag{43}$$

which holds for all $k \in \mathbb{Z}$. To see the case with $k \leq 0$, we only need to notice that, for any $\boldsymbol{\beta} \in \mathbb{R}^d$, we have the limit

$$\breve{\boldsymbol{\beta}}_k^\circ(\alpha) := \lim_{m\to\infty} \breve{f}_{\boldsymbol{\xi}_k} \circ \cdots \breve{f}_{\boldsymbol{\xi}_{k-m}}(\boldsymbol{\beta}) =: \breve{h}_\alpha(\boldsymbol{\xi}_k, \boldsymbol{\xi}_{k-1}, \ldots), \tag{44}$$

where $\breve{h}_\alpha$ is a measurable function that depends on α, and we use $\boldsymbol{\xi}_k$ to denote all the new-coming random parts in the k-th iteration, that is,

$$\boldsymbol{\xi}_k = (D_k, (y_k, \boldsymbol{x}_k)), \quad k \in \mathbb{Z}. \tag{45}$$

For $k \leq 0$, $\boldsymbol{\xi}_k$ can be viewed as an i.i.d. copy of $\boldsymbol{\xi}_j$ for some $j \geq 1$. The limit $\breve{h}_\alpha(\boldsymbol{\xi}_k, \boldsymbol{\xi}_{k-1}, \ldots)$ exists almost surely and does not depend on $\boldsymbol{\beta}$. Therefore, the SGD dropout iteration $\breve{\boldsymbol{\beta}}_k^\circ(\alpha) = \breve{f}_{D_k,(y_k,\boldsymbol{x}_k)}(\breve{\boldsymbol{\beta}}_{k-1}^\circ(\alpha))$ in (43) holds for all $k \in \mathbb{Z}$.

As for GD with dropout, the parameter dimension d influences the contraction constant $\breve{r}_{\alpha,q}$ only indirectly through the matrix $\mathbb{E}[\boldsymbol{x}_k \boldsymbol{x}_k^\top]$. For i.i.d. Gaussian entries a more explicit expression is derived in the next lemma.

Lemma 13 (Dependence on d in SGD with dropout) *If the covariate vectors $\boldsymbol{x}_k$ are independently drawn from a standard multivariate Gaussian, then,*

$$\breve{r}_{\alpha,2}^2 = 1 - 2\alpha p + \alpha^2[(d-1)p^2 + 3p]. \tag{46}$$

Hence, $\alpha < 2/((d-1)p + 3)$ ensures $\breve{r}_{\alpha,2} < 1$.

4.3 Asymptotics of Dropout in SGD

In this section, we provide the asymptotics for the k-th iterate of SGD dropout and the Ruppert-Polyak averaged version.

Lemma 14 (Moment convergence of iterative SGD dropout) *Let $q \geq 2$ and suppose that Assumption 1 holds. For the stationary SGD dropout sequence $\{\breve{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ defined in (42) with learning rate α satisfying* (37)*, we have*

$$\max_k \big(\mathbb{E}\|\breve{\boldsymbol{\beta}}_k^\circ(\alpha) - \breve{\boldsymbol{\beta}}\|_2^q\big)^{1/q} = O(\sqrt{\alpha}). \tag{47}$$

Besides the stochastic order of the last iterate of SGD dropout $\breve{\boldsymbol{\beta}}_k(\alpha)$, we are also interested in the limiting distribution of the Ruppert-Polyak averaged SGD dropout, which can effectively reduce the variance and keep the online computing scheme. In particular, we define

$$\bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha) = \frac{1}{k}\sum_{i=1}^k \breve{\boldsymbol{\beta}}_i(\alpha). \tag{48}$$

Theorem 15 (Quenched CLT of averaged SGD dropout) *Under Assumption 1, if the learning rate α satisfies* (37)*, then,*

$$\sqrt{k}\big(\bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha) - \breve{\boldsymbol{\beta}}\big) \Rightarrow \mathcal{N}\big(0, \breve{\Sigma}(\alpha)\big), \quad \text{as } k\to\infty, \tag{49}$$

with $\breve{\Sigma}(\alpha) := \sum_{i=-\infty}^{\infty}\mathbb{E}[(\breve{\boldsymbol{\beta}}_0^\circ(\alpha) - \breve{\boldsymbol{\beta}})(\breve{\boldsymbol{\beta}}_i^\circ(\alpha) - \breve{\boldsymbol{\beta}})^\top]$ the long-run covariance matrix of the stationary process $\breve{\boldsymbol{\beta}}_i^\circ(\alpha) \sim \breve{\pi}_\alpha$.

The CLTs for fixed design (Theorem 7) and random design (Theorem 15) have different covariance matrices in the limit $k\to\infty$. In particular, the limiting covariance matrix of the averaged SGD iterates with dropout depends on the random design.

As discussed above Corollary 8, one can also choose different learning rates $\alpha_1,\dots,\alpha_s$ and then run the SGD dropout sequences $\breve{\boldsymbol{\beta}}_k(\alpha_1),\dots,\breve{\boldsymbol{\beta}}_k(\alpha_s)$ in parallel. For d-dimensional vectors $\boldsymbol{u}_1,\dots,\boldsymbol{u}_s$, recall that $\mathrm{vec}(\boldsymbol{u}_1,\dots,\boldsymbol{u}_s) := (\boldsymbol{u}_1^\top,\dots,\boldsymbol{u}_s^\top)^\top$ is the ds-dimensional concatenation.

Corollary 16 (Quenched CLT of parallel averaged SGD dropout) *Under Assumption 1, consider a sequence of constant learning rates $\alpha_1,\dots,\alpha_s$, for $s\geq 1$, satisfying the condition in (37). Then, for any initial vectors $\breve{\boldsymbol{\beta}}_0(\alpha_1),\dots,\breve{\boldsymbol{\beta}}_0(\alpha_s)$,*

$$\sqrt{k}\cdot\mathrm{vec}\Big(\bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha_1) - \breve{\boldsymbol{\beta}},\dots,\bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha_s) - \breve{\boldsymbol{\beta}}\Big) \Rightarrow \mathcal{N}\big(0, \breve{\Sigma}^{\mathrm{vec}}\big), \quad \text{as } k\to\infty, \tag{50}$$

with the long-run covariance matrix

$$\breve{\Sigma}^{\mathrm{vec}} = \sum_{i=-\infty}^{\infty}\mathrm{Cov}\big(\mathrm{vec}\big(\breve{\boldsymbol{\beta}}_0^\circ(\alpha_1),\dots,\breve{\boldsymbol{\beta}}_0^\circ(\alpha_s)\big), \mathrm{vec}\big(\breve{\boldsymbol{\beta}}_i^\circ(\alpha_1),\dots,\breve{\boldsymbol{\beta}}_i^\circ(\alpha_s)\big)\big).$$

Theorem 17 (Quenched invariance principle of averaged SGD dropout) *Let Assumption 1 hold for some $q > 2$ and consider the learning rate α satisfying (37). Define a partial sum process $(\breve{S}_i^\circ(\alpha))_{1\le i\le t}$ for $t \in \mathbb{N}_+$ with*

$$\breve{S}_i^\circ(\alpha) = \sum_{k=1}^{i} (\breve{\boldsymbol{\beta}}_k^\circ(\alpha) - \breve{\boldsymbol{\beta}}). \tag{51}$$

Then, there exists a (richer) probability space $(\breve{\Omega}^\star, \breve{\mathcal{A}}^\star, \breve{\mathbb{P}}^\star)$ on which one can define d-dimensional random vectors $\breve{\boldsymbol{\beta}}_k^\star$, the associated partial sum process $\breve{S}_i^\star(\alpha) = \sum_{k=1}^{i}(\breve{\boldsymbol{\beta}}_k^\star(\alpha) - \breve{\boldsymbol{\beta}})$, and a Gaussian process $\breve{G}_i^\star = \sum_{k=1}^{i} \breve{\boldsymbol{z}}_k^\star$, with independent Gaussian random vectors $\breve{\boldsymbol{z}}_k \sim \mathcal{N}(0, I_d)$, such that $(\breve{S}_i^\circ)_{1\le i\le t} \overset{\mathcal{D}}{=} (\breve{S}_i^\star)_{1\le i\le t}$ and

$$\max_{1\le i\le t} \left\| \breve{S}_i^\star - \breve{\Sigma}^{1/2}(\alpha)\breve{G}_i^\star \right\|_2 = o_{\mathbb{P}}(t^{1/q}), \quad \text{in } (\breve{\Omega}^\star, \breve{\mathcal{A}}^\star, \breve{\mathbb{P}}^\star), \tag{52}$$

where $\breve{\Sigma}(\alpha)$ is the long-run covariance matrix defined in Theorem 15. In addition, this approximation holds for all $(\breve{S}_i^{\breve{\boldsymbol{\beta}}_0}(\alpha))_{1\le i\le t}$ given any arbitrary initialization $\breve{\boldsymbol{\beta}}_0 \in \mathbb{R}^d$, where

$$\breve{S}_i^{\breve{\boldsymbol{\beta}}_0}(\alpha) = \sum_{k=1}^{i} (\breve{\boldsymbol{\beta}}_k(\alpha) - \breve{\boldsymbol{\beta}}). \tag{53}$$

5. Online Inference for SGD with Dropout

The long-run covariance matrix $\breve{\Sigma}(\alpha)$ of the averaged SGD dropouts is usually unknown and needs to be estimated. We now propose an online estimation method for $\breve{\Sigma}(\alpha)$, and establish theoretical guarantees.

The key idea is to adopt the non-overlapping batched means (NBM) method (Lahiri, 1999, 2003; Xiao and Wu, 2011), which resamples blocks of observations to estimate the long-run covariance of dependent data. Essentially, a sequences of non-overlapping blocks are pre-specified. When the block sizes are large enough, usually increasing as the the sample size grows, the block sums shall behave similar to independent observations and therefore can be used to estimate the long-run covariance. In this paper, to facilitate the online inference of the dependent SGD dropout iterates $\{\breve{\boldsymbol{\beta}}_k(\alpha)\}_{k\in\mathbb{N}}$, we shall extend the offline NBM estimators to online versions by only including the past SGD dropout iterates in each batch. The overlapped batch-means (OBM) methods are also investigated in literature; see for example Xiao and Wu (2011); Zhu et al. (2023). We shall only focus on the NBM estimates in this study given its simpler structure.

Let $\eta_1, \eta_2, \ldots$ be a strictly increasing integer-valued sequence satisfying $\eta_1 = 1$ and $\eta_{m+1} - \eta_m \to \infty$ as $m \to \infty$. For each m, we let B_m denote the block

$$B_m = \{\eta_m, \eta_m + 1, \ldots, \eta_{m+1} - 1\}. \tag{54}$$

For the k-th SGD dropout iteration, denote by $\psi(k)$ the largest index m such that $\eta_m \le k$. For any d-dimensional vector $\boldsymbol{v} = (v_1, \ldots, v_d)^\top$, the Kronecker product is the $d \times d$ matrix $\boldsymbol{v}^{\otimes 2} = (v_i v_j)_{i,j=1}^d$. Based on the non-overlapping blocks $\{B_m\}_{m\in\mathbb{N}}$, for the k-th iteration, we

can estimate the long-run covariance matrix $\breve{\Sigma}(\alpha)$ in Theorem 15 by

$$\hat{\Sigma}_k(\alpha) = \frac{1}{k}\sum_{m=1}^{\psi(k)-1}\Big(\sum_{i\in B_m}\big[\breve{\boldsymbol{\beta}}_i(\alpha) - \bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha)\big]\Big)^{\otimes 2} + \frac{1}{k}\Big(\sum_{i=\eta_{\psi(k)}}^{k}\big[\breve{\boldsymbol{\beta}}_i(\alpha) - \bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha)\big]\Big)^{\otimes 2}. \quad (55)$$

The estimator $\hat{\Sigma}_k(\alpha)$ is composed of two parts. The first part takes the sum within each block and then estimates the sample covariances of these centered block sums. The second part accounts for the remaining observations, which can be viewed as the estimated covariance of the tail block.

For the recursive computation of $\hat{\Sigma}_k(\alpha)$, we need to rewrite (55) such that, in the k-th iteration, we can update $\hat{\Sigma}_k(\alpha)$ based on the information from the $(k-1)$-th step and the latest iterate $\breve{\boldsymbol{\beta}}_k(\alpha)$. To this end, we denote the number of iterates included in the tail part (i.e., the second part) in (55) by

$$\delta_\eta(k) = k - \eta_{\psi(k)} + 1, \quad (56)$$

and define two partial sums

$$\mathcal{S}_m(\alpha) = \sum_{i\in B_m}\breve{\boldsymbol{\beta}}_i(\alpha) \quad \text{and} \quad \mathcal{R}_k(\alpha) = \sum_{i=\eta_{\psi(k)}}^{k}\breve{\boldsymbol{\beta}}_i(\alpha). \quad (57)$$

Then, we notice that the estimator $\hat{\Sigma}_k(\alpha)$ in (55) can be rewritten as follows,

$$\begin{aligned}\hat{\Sigma}_k(\alpha) = \frac{1}{k}\Bigg[&\Big(\sum_{m=1}^{\psi(k)-1}\mathcal{S}_m(\alpha)^{\otimes 2} + \mathcal{R}_k(\alpha)^{\otimes 2}\Big) + \Big(\sum_{m=1}^{\psi(k)-1}|B_m|^2 + |\delta_\eta(k)|^2\Big)\bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha)^{\otimes 2}\\ &- \Big(\sum_{m=1}^{\psi(k)-1}|B_m|\mathcal{S}_m(\alpha) + \delta_\eta(k)\mathcal{R}_k(\alpha)\Big)\bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha)^\top\\ &- \bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha)\Big(\sum_{m=1}^{\psi(k)-1}|B_m|\mathcal{S}_m(\alpha) + \delta_\eta(k)\mathcal{R}_k(\alpha)\Big)^\top\Bigg]\\ =:\ &\frac{1}{k}\Big[\mathcal{V}_k(\alpha) + K_k\bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha)^{\otimes 2} - H_k(\alpha)\bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha)^\top - \bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha)H_k(\alpha)^\top\Big]. \quad (58)\end{aligned}$$

As such, the estimation of $\breve{\Sigma}(\alpha)$ reduces to computing $\{\mathcal{V}_k(\alpha), K_k, H_k(\alpha), \bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha)\}$ recursively with respect to k. We provided the pseudo codes of the recursion in Algorithm 1. We shall further establish the convergence rate of the proposed online estimator $\hat{\Sigma}_k(\alpha)$ in Theorem 18.

The rational behind Algorithm 1 is as follows: if $k+1 < \eta_{\psi(k)+1}$, then the index $k+1$ still belongs to the block $B_{\psi(k)}$ and $\psi(k+1) = \psi(k)$. Also we have $\mathcal{R}_{k+1}(\alpha) = \mathcal{R}_k(\alpha) + \breve{\boldsymbol{\beta}}_{k+1}(\alpha)$ and $\delta_\eta(k+1) = \delta_\eta(k) + 1$. Consequently, $\{K_{k+1}, \mathcal{V}_{k+1}(\alpha), H_{k+1}(\alpha)\}$ can be recursively updated via

$$\begin{aligned}&K_{k+1} = K_k - |\delta_\eta(k)|^2 + |\delta_\eta(k+1)|^2,\\ &\mathcal{V}_{k+1}(\alpha) = \mathcal{V}_k(\alpha) - \mathcal{R}_k(\alpha)^{\otimes 2} + \mathcal{R}_{k+1}(\alpha)^{\otimes 2},\\ &H_{k+1}(\alpha) = H_k(\alpha) - \delta_\eta(k)\mathcal{R}_k(\alpha) + \delta_\eta(k+1)\mathcal{R}_{k+1}(\alpha).\end{aligned}$$

Algorithm 1: Online estimation of long-run covariance matrices of ASGD dropout

Data: Sequential random samples $(y_1, \boldsymbol{x}_1), \ldots, (y_k, \boldsymbol{x}_k)$; sequential dropout matrices $D_1, \ldots, D_k$; constant learning rate α; predefined sequences $\{\eta_m\}_{m\in\mathbb{N}}$

Result: ASGD dropout $\bar{\boldsymbol{\beta}}_{k+1}^{\mathrm{sgd}}(\alpha)$; estimated long-run covariance matrix $\tilde{\Sigma}_{k+1}(\alpha)$

Initialize $\breve{\boldsymbol{\beta}}_0(\alpha) = \bar{\boldsymbol{\beta}}_0^{\mathrm{sgd}}(\alpha) = \mathcal{R}_0(\alpha) \leftarrow 0$,
$\psi(0) \leftarrow 1$, $\delta_\eta(0) \leftarrow 1$, $K_0 = H_0(\alpha) \leftarrow 1$, $\mathcal{V}_0(\alpha) \leftarrow 0$

for $k = 0, 1, 2, 3, \ldots$ **do**

$\breve{\boldsymbol{\beta}}_{k+1}(\alpha) \leftarrow \breve{\boldsymbol{\beta}}_k(\alpha) + \alpha D_k \boldsymbol{x}_k \big(y_k - \boldsymbol{x}_k^\top D_k \breve{\boldsymbol{\beta}}_{k-1}(\alpha)\big)$; `/* SGD dropout */`

$\bar{\boldsymbol{\beta}}_{k+1}^{\mathrm{sgd}}(\alpha) \leftarrow \{k\bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha) + \breve{\boldsymbol{\beta}}_{k+1}(\alpha)\}/(k+1)$; `/* ASGD dropout */`

if $k + 1 < \eta_{\psi(k)+1}$ **then**

$\mathcal{R}_{k+1}(\alpha) \leftarrow \mathcal{R}_k(\alpha) + \breve{\boldsymbol{\beta}}_{k+1}(\alpha)$, $\delta_\eta(k+1) \leftarrow \delta_\eta(k) + 1$;
$K_{k+1} \leftarrow K_k - \delta_\eta^2(k) + \delta_\eta^2(k+1)$, $\psi(k+1) \leftarrow \psi(k)$;
$H_{k+1}(\alpha) \leftarrow H_k(\alpha) - \delta_\eta(k)\mathcal{R}_k(\alpha) + \delta_\eta(k+1)\mathcal{R}_{k+1}(\alpha)$;
$\mathcal{V}_{k+1}(\alpha) \leftarrow \mathcal{V}_k(\alpha) - \mathcal{R}_k(\alpha)^{\otimes 2} + \mathcal{R}_{k+1}(\alpha)^{\otimes 2}$;

else

$\mathcal{R}_{k+1}(\alpha) \leftarrow \breve{\boldsymbol{\beta}}_{k+1}(\alpha)$, $\delta_\eta(k+1) \leftarrow 1$;
$\psi(k+1) \leftarrow \psi(k) + 1$;
$K_{k+1} \leftarrow K_k + 1$, $H_{k+1}(\alpha) \leftarrow H_k(\alpha) + \mathcal{R}_{k+1}(\alpha)$;
$\mathcal{V}_{k+1}(\alpha) \leftarrow \mathcal{V}_k(\alpha) + \mathcal{R}_{k+1}(\alpha)^{\otimes 2}$;

end

$\hat{\Sigma}_{k+1}(\alpha) \leftarrow$
$\big[\mathcal{V}_{k+1}(\alpha) + K_{k+1}\bar{\boldsymbol{\beta}}_{k+1}^{\mathrm{sgd}}(\alpha)^{\otimes 2} - H_{k+1}(\alpha)\bar{\boldsymbol{\beta}}_{k+1}^{\mathrm{sgd}}(\alpha)^\top - \bar{\boldsymbol{\beta}}_{k+1}^{\mathrm{sgd}}(\alpha)H_{k+1}(\alpha)^\top\big]/(k+1)$;
`/* Estimated long-run covariance matrix */`

end

Otherwise, if $k + 1 = \eta_{\psi(k)}$, we have $\psi(k+1) = \psi(k) + 1$. Hence $\mathcal{R}_{k+1}(\alpha) = \breve{\boldsymbol{\beta}}_{k+1}(\alpha)$ and $\delta_\eta(k+1) = 1$. In this case, $\{K_{k+1}, \mathcal{V}_{k+1}(\alpha), H_{k+1}(\alpha)\}$ can be recursively updated as follows,

$$\begin{aligned} K_{k+1} &= K_k + 1, \\ \mathcal{V}_{k+1}(\alpha) &= \mathcal{V}_k(\alpha) + \mathcal{R}_{k+1}(\alpha)^{\otimes 2}, \\ H_{k+1}(\alpha) &= H_k(\alpha) + \mathcal{R}_{k+1}(\alpha). \end{aligned}$$

As such, given $\breve{\boldsymbol{\beta}}_1(\alpha), \ldots, \breve{\boldsymbol{\beta}}_k(\alpha)$, the estimator $\hat{\Sigma}_k(\alpha)$ for the long-run covariance matrix $\breve{\Sigma}(\alpha)$ can be updated online, requiring only $O(1)$ memory storage.

Theorem 18 (Precision of $\hat{\Sigma}_k(\alpha)$) *Let $\eta_m = \lfloor cm^\zeta \rfloor$ for some $c > 0$ and $\zeta > 1$. If the conditions of Theorem 12 are satisfied with $q \geq 4$, we have*

$$\mathbb{E}\|\hat{\Sigma}_k(\alpha) - \breve{\Sigma}(\alpha)\|_F \lesssim dk^{(1/\zeta - 1)\vee(-1/(2\zeta))},$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and the constants in $\lesssim$ are independent of k and d. Consequently, for any dimension d, if the number of iterations k satisfies

$$k \gtrsim d^{\frac{\zeta}{\zeta-1}\vee\frac{1}{2\zeta}},$$

then $\mathbb{E}\|\hat{\Sigma}_k(\alpha) - \breve{\Sigma}(\alpha)\|_F = o(1)$.

In particular, for $\zeta = 3/2$,

$$\mathbb{E}\|\hat{\Sigma}_k(\alpha) - \breve{\Sigma}(\alpha)\|_F \lesssim dk^{-1/3}. \tag{59}$$

This rate is optimal among long-run covariance estimators in the fixed-dimensional settings, even when comparing to offline estimation; see Xiao and Wu (2011) for details. This result indicates that for the dimension d, at least $n \gtrsim d^3$ training budget is needed to ensure consistency. The slow scaling of the algorithm in the dimension makes it, however, challenging to run the procedure for $d \gtrsim 1000$ requiring $\gtrsim 10^9$ training samples.

By the estimation procedure summarized in Algorithm 1, we can asymptotically estimate the long-run covariance matrix of the SGD dropout iterates $\bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha)$ for any arbitrarily fixed initial vector. For some given confidence level $\omega \in (0,1)$, in the k-th iteration the online confidence interval for each coordinate $\breve{\boldsymbol{\beta}}_j$, $j = 1, \ldots, d$, of the vector $\breve{\boldsymbol{\beta}}$ in (34) is

$$\mathrm{CI}_{\omega,k,j} := \Big[\bar{\boldsymbol{\beta}}_{k,j}^{\mathrm{sgd}}(\alpha) - z_{1-\omega/2}\sqrt{\hat{\sigma}_{k,jj}(\alpha)/k},\ \bar{\boldsymbol{\beta}}_{k,j}^{\mathrm{sgd}}(\alpha) + z_{1-\omega/2}\sqrt{\hat{\sigma}_{k,jj}(\alpha)/k}\Big], \tag{60}$$

with $z_{1-\omega/2}$ denoting the $(1-\omega/2)$-percentile of the standard normal distribution. Here, $\hat{\sigma}_{k,jj}(\alpha)$ is the j-th diagonal of the proposed online long-run covariance estimator $\hat{\Sigma}_k(\alpha)$ in (55), and $\bar{\boldsymbol{\beta}}_{k,j}^{\mathrm{sgd}}(\alpha)$ is the j-th coordinate of the averaged SGD dropout estimate $\bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha)$. Furthermore, the online joint confidence regions for the vector $\breve{\boldsymbol{\beta}}$ is

$$\mathrm{CI}_{\omega,k} := \Big\{\boldsymbol{\beta} \in \mathbb{R}^d :\ k\big(\bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha) - \boldsymbol{\beta}\big)^\top \hat{\Sigma}_k^{-1}(\alpha)\big(\bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha) - \boldsymbol{\beta}\big) \le \chi^2_{d,1-\omega/2}\Big\}, \tag{61}$$

where $\chi^2_{d,1-\omega/2}$ is the $(1-\omega/2)$-percentile of the χ^2_d distribution with d degrees of freedom.

Corollary 19 (Asymptotic coverage probability) *Suppose that Assumption 1 holds and the learning rate α satisfies (37). Given $\omega \in (0,1)$ and $\eta_m = \lfloor cm^\zeta \rfloor$ for some $c > 0$ and $\zeta > 1$, $\mathrm{CI}_{\omega,k,j}$ defined in (60), and $\mathrm{CI}_{\omega,k}$ defined in (61) are asymptotic $100(1-\omega)\%$ confidence intervals, that is, $\mathbb{P}\big(\breve{\boldsymbol{\beta}}_j \in \mathrm{CI}_{\omega,k,j}\big) \to 1-\omega$ for all $j = 1, \ldots, d$, and $\mathbb{P}\big(\breve{\boldsymbol{\beta}} \in \mathrm{CI}_{\omega,k}\big) \to 1-\omega$, as $k \to \infty$. More generally, for any d-dimensional unit-length vector $\boldsymbol{v}$ with $\|\boldsymbol{v}\|_2 = 1$, and $z_{1-\omega/2}$ the $(1-\omega/2)$-quantile of the standard normal distribution,*

$$\mathrm{CI}_{\omega,k}^{\mathrm{Proj}} := \Big[\boldsymbol{v}^\top \bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha) - z_{1-\omega/2}\sqrt{k^{-1}(\boldsymbol{v}^\top \hat{\Sigma}_k(\alpha)\boldsymbol{v})},\ \boldsymbol{v}^\top \bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha) + z_{1-\omega/2}\sqrt{k^{-1}(\boldsymbol{v}^\top \hat{\Sigma}_k(\alpha)\boldsymbol{v})}\Big] \tag{62}$$

is an asymptotic $100(1-\omega)\%$ confidence interval for the one-dimensional projection $\boldsymbol{v}^\top\breve{\boldsymbol{\beta}}$, that is, $\mathbb{P}\big(\boldsymbol{v}^\top\breve{\boldsymbol{\beta}} \in \mathrm{CI}_{\omega,k}^{\mathrm{Proj}}\big) \to 1-\omega$, as $k \to \infty$.

By the quenched CLT of the averaged SGD dropout sequence $\{\bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha)\}_{k\in\mathbb{N}}$ in Theorem 17 and the consistency of $\hat{\Sigma}_k(\alpha)$ in Theorem 18, we can apply Slutsky's theorem and obtain the results in Corollary 19. In Section 6, we shall validate the proposed online inference method by examining the estimation accuracy of the proposed online estimator $\hat{\Sigma}_k(\alpha)$ and the coverage probability of $\mathrm{CI}_{\omega,k}^{\mathrm{Proj}}$ under different settings.

6. Numerical Experiments

In this section, we present the results of the numerical experiments to demonstrate the validity of the proposed online inference methodology. For a real-data application see Appendix H. The codes for reproducing all results and figures can be found online[1].

6.1 On the Range of the Learning Rate

The GD dropout iterates can be defined via the recursion (6), $\tilde{\boldsymbol{\beta}}_k(\alpha)-\tilde{\boldsymbol{\beta}} = A_k(\alpha)(\tilde{\boldsymbol{\beta}}_{k-1}(\alpha)-\tilde{\boldsymbol{\beta}}) + \boldsymbol{b}_k(\alpha)$, and the derived theory requires the learning rate α to satisfy $\alpha\|\mathbb{X}\| < 2$. Via a simulation study we show that this range is close to sharp to guarantee that the contraction constant

$$r^2_{\alpha,2} = \sup_{\boldsymbol{v}\in\mathbb{R}^d:\|\boldsymbol{v}\|_2=1} \mathbb{E}\big\|A_1(\alpha)\boldsymbol{v}\big\|_2^2 = \lambda_{\max}\big(\mathbb{E}[A_1^\top(\alpha)A_1(\alpha)]\big) < 1. \tag{63}$$

This then indicates that the condition $\alpha\|\mathbb{X}\| < 2$ in Lemma 2 is fairly sharp.

For the $n \times d$ full design matrix X, we independently generate each entry of X from the standard normal distribution. Since $\mathbb{X} = X^\top X$, the upper bound $2/\lambda_{\max}(\mathbb{X})$ of the learning rate α can be computed. Then, we independently generate $N = 500$ dropout matrices D_i, $i = 1, \ldots, N$ with retaining probability p. The simulation study evaluates the empirical contraction constant

$$\hat{r}^2_{\alpha,2} := \lambda_{\max}\Big(N^{-1}\sum_{i=1}^N A_i^\top(\alpha)A_i(\alpha)\Big), \tag{64}$$

for different training budget n, dimension d, retaining probability p, and learning rate α. Table 2 shows that even if the learning rate α exceeds the upper bound $2/\lambda_{\max}(\mathbb{X})$ by a

n, d	$2/\lambda_{\max}(\mathbb{X})$	p	α	$\hat{r}^2_{\alpha,2}$
100, 5	0.0151	0.9	0.0150	0.97
			0.0154	1.02
100, 50	0.0068	0.9	0.0067	0.93
			0.0072	1.01
		0.8	0.0068	0.90
			0.0075	1.02
100, 100	0.0052	0.9	0.0050	0.93
			0.0057	1.15
		0.5	0.0050	0.94
			0.0075	1.06

Table 2: Effects of the learning rate α on the geometric moment contraction of the GD iterates with dropout.

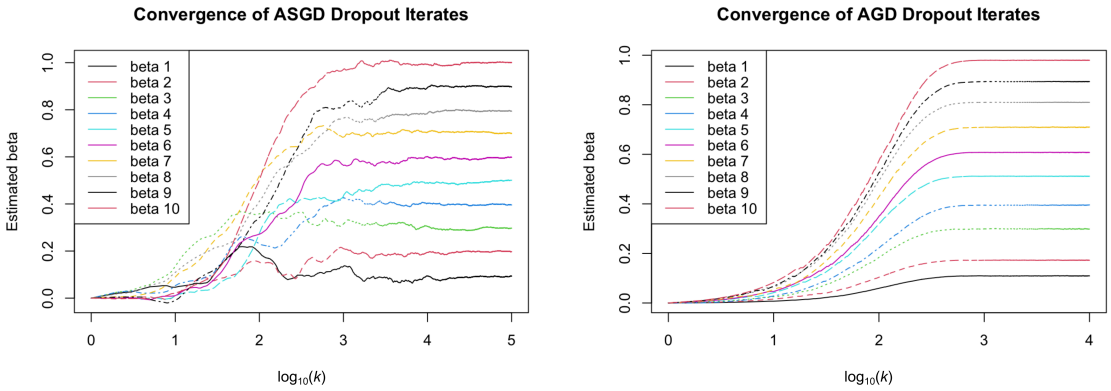
small margin, the contraction will not hold any more since $\hat{r}^2_{\alpha,2} > 1$. This indicates that the condition $\alpha\|\mathbb{X}\| < 2$ is close to sharp.

1. https://github.com/jiaqili97/Dropout_SGD

6.2 Estimation of Long-Run Covariance Matrix

In this section, we provide the simulation results of the proposed long-run covariance estimator $\hat{\Sigma}_k(\alpha)$ defined in (55), and its online version (58).

Figure 1 shows the convergence of the GD and SGD iterates with dropout. The coordinates of the true regression vector $\boldsymbol{\beta}^*$ are equidistantly spaced between 0 and 1. One can see that the initialization is quickly forgotten in both GD and SGD algorithms.



(a) AGD. (b) ASGD.

Figure 1: Convergence traces of AGD and ASGD iterates with dropout regularization based on a single run, with dimension $d = 10$ and initialization at zero. The coordinates of the true parameter $\boldsymbol{\beta}^*$ are equidistantly spaced between 0 and 1, the learning rate $\alpha = 0.01$, and the retaining probability $p = 0.9$. Each curve represents the convergence trace of one coordinate.

Figures 2–3 evaluate the performance of the online long-run covariance matrix estimator $\hat{\Sigma}_k(\alpha)$. Suppose that the training budget is n. The parameters in the blocks $B_m = \{\eta_m, \eta_{m+1}, \ldots, \eta_{m+1} - 1\}$, $m = 1, \ldots, M$, defined in (54), are chosen as $M = \lfloor \sqrt{n} \rfloor$ and $\eta_m = m^2$. In Figure 2, we can see that the long-run variances of each coordinate of the ASGD dropout iterates $\bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha)$ converges as the number of iterations k grows. Since there is no closed-form expression for the true long-run covariance matrix $\check{\Sigma}(\alpha)$ defined in Theorem 15, in Figure 3, we set the true long-run covariance matrix as the average of the 10^8-th iteration over 20 replications of the experiment, while we plot the first 10^7 iterations only. Figure 3(a) shows that the empirical convergence of the Frobenius norm $\mathbb{E}\|\hat{\Sigma}_k(\alpha) - \check{\Sigma}(\alpha)\|_F$ for dimension d between 100 and 1000 agrees with the theoretical convergence rate $\propto k^{-1/3}$ in (59). For $\boldsymbol{v}^\top = d^{-1/2}(1, \ldots, 1)^\top$, the length of the confidence interval for the one-dimensional projection $\boldsymbol{v}^\top \boldsymbol{\beta}^*$ is displayed in Figure 3(b). The same $\boldsymbol{v}$ is used for the simulation results summarized in Table 3. In the next section, we shall show that by using these estimated long-run variances, the online confidence intervals achieve asymptotically the nominal coverage probability.

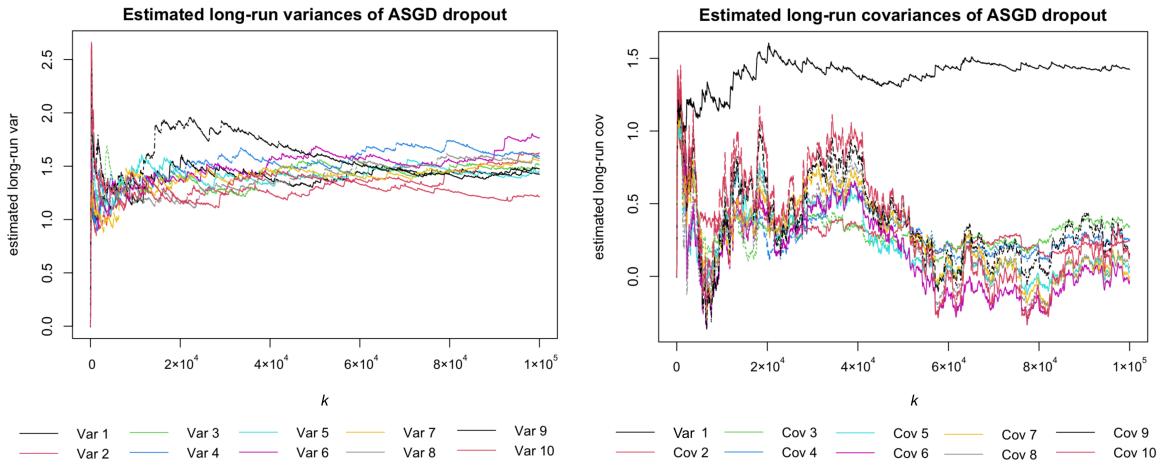


Figure 2: Under the same setting as in Figure 1, the left panel displays the estimated long-run variances of ASGD dropout iterates, that is, the diagonal entries of the estimated long-run covariance matrix $\hat{\Sigma}_k(\alpha)$; the right panel displays the first row of $\hat{\Sigma}_k(\alpha)$.

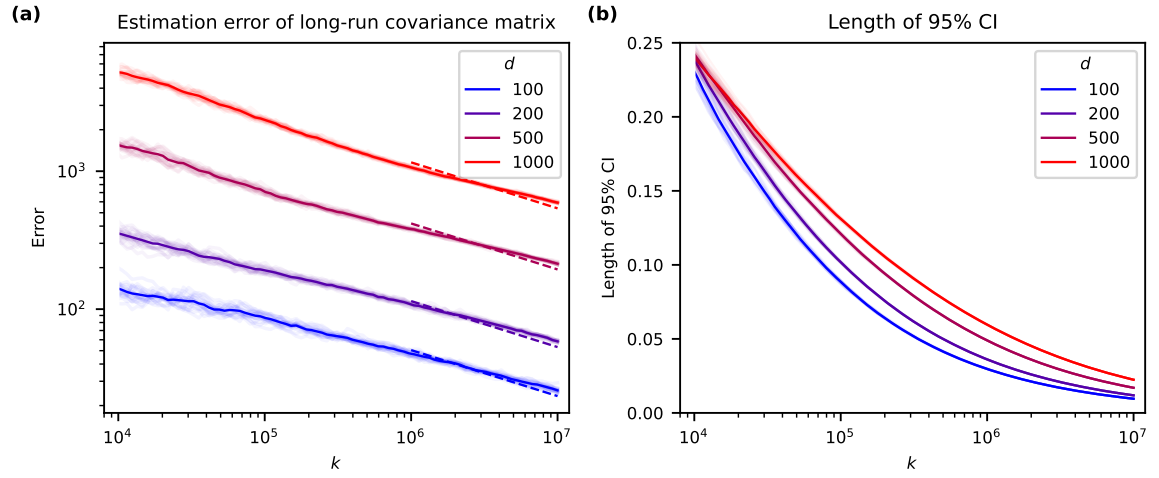


Figure 3: Performance of the long-run covariance matrix estimator $\hat{\Sigma}_k(\alpha)$ for learning rate $\alpha = 1/d$ and $p = 0.9$. (a) Empirical estimation errors $\mathbb{E}\|\hat{\Sigma}_k(\alpha) - \breve{\Sigma}(\alpha)\|_F$ averaged over 20 replications. The dashed lines show the theoretical decay rate $\propto k^{-1/3}$. The coordinates of $\boldsymbol{\beta}^*$ are equidistantly spaced between 0 and 1. (b) Length of the confidence intervals $\mathrm{CI}^{\mathrm{Proj}}_{\omega,k}$ in (62).

6.3 Online Confidence Intervals of ASGD Dropout Iterates

Recall the $100(1-\omega)\%$ online confidence interval $\mathrm{CI}_{\omega,k}^{\mathrm{Proj}}$ in (62) for the one-dimensional projection of the true parameter $\boldsymbol{\beta}^*$, i.e., $\boldsymbol{v}^\top\boldsymbol{\beta}^*$. For dimension $d \in \{50, 100, 200, 500\}$, a similar performance in convergence of the coverage probabilities is observed in Figure 4. In Table 3, we report the coverage probabilities of the confidence intervals $\mathrm{CI}_{\omega,k}^{\mathrm{Proj}}$ for dimension $d \in \{50, 100, 200, 500\}$, dropout probabilities $p \in \{0.5, 0.9\}$, and constant learning rates α ranging from 0.002 to 0.02. The results demonstrate that the online confidence intervals can achieve the target coverage probability as the number of iterations increases.

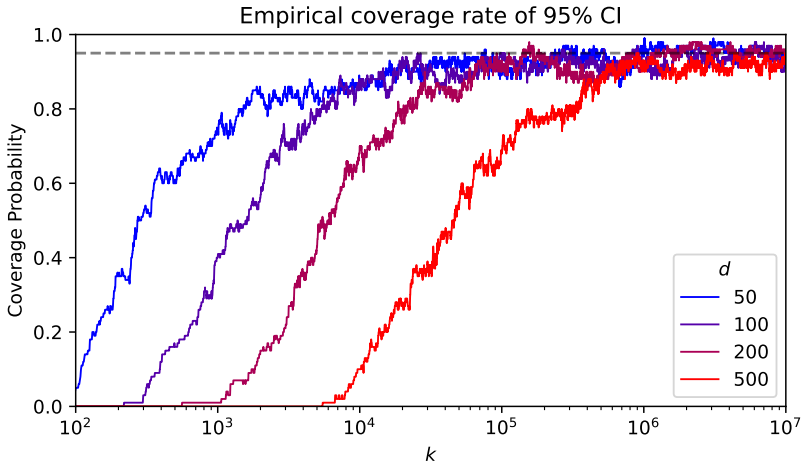


Figure 4: Coverage probabilities of 95% confidence intervals for one-dimensional projection of ASGD dropout iterates based on 100 independent repetitions, dropout probability $p = 0.9$, learning rate $\alpha = 1/d$, and coordinates of $\boldsymbol{\beta}^*$ equidistantly spaced between 0 and 1 and initialized at zero. The gray dashed line shows the nominal 0.95 coverage.

			$p = 0.9$		
d	$k = 10^5$	$k = 5 * 10^5$	$k = 10^6$	$k = 5 * 10^6$	$k = 10^7$
50	0.91 (0.0286)	0.95 (0.0218)	0.95 (0.0218)	0.95 (0.0218)	0.92 (0.0271)
100	0.94 (0.0237)	0.92 (0.0271)	0.91 (0.0286)	0.94 (0.0237)	0.92 (0.0271)
200	0.91 (0.0286)	0.89 (0.0313)	0.92 (0.0271)	0.93 (0.0255)	0.96 (0.0196)
500	0.68 (0.0466)	0.89 (0.0313)	0.92 (0.0271)	0.93 (0.0255)	0.95 (0.0218)
			$p = 0.5$		
d	$k = 10^5$	$k = 5 * 10^5$	$k = 10^6$	$k = 5 * 10^6$	$k = 10^7$
50	0.94 (0.0237)	0.92 (0.0271)	0.92 (0.0271)	0.95 (0.0218)	0.96 (0.0196)
100	0.94 (0.0237)	0.96 (0.0196)	0.93 (0.0255)	0.94 (0.0237)	0.96 (0.0196)
200	0.81 (0.0392)	0.91 (0.0286)	0.93 (0.0255)	0.95 (0.0218)	0.93 (0.0255)
500	0.55 (0.0497)	0.86 (0.0347)	0.90 (0.0300)	0.96 (0.0196)	0.96 (0.0196)

Table 3: Empirical coverage probabilities of 95% confidence intervals based on 100 independent repetitions (with standard errors in the brackets) and learning rate $\alpha = 1/d$.

Acknowledgments

We are grateful to the referees and the associate editor for constructive feedback that improved the work considerably. Moreover, the authors want to thank Haoxiong Yan for the support of large-scale computation, and thank Gabriel Clara for several helpful suggestions. Jiaqi Li's research is partially supported by the NSF (Grant NSF/DMS-2515926). Johannes Schmidt-Hieber has received funding from the Dutch Research Council (NWO) via the Vidi grant VI.Vidi.192.021. Wei Biao Wu's research is partially supported by the NSF (Grants NSF/DMS-2311249, NSF/DMS-2027723).

References

Raman Arora, Peter Bartlett, Poorya Mianjy, and Nathan Srebro. Dropout: Explicit Forms and Capacity Control. In *Proceedings of the 38th International Conference on Machine Learning*, pages 351–361. PMLR, 2021.

Pierre Baldi and Peter Sadowski. Understanding Dropout. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Roja Bandari, Sitaram Asur, and Bernardo Huberman. The pulse of news in social media: forecasting popularity. *Proceedings of the International AAAI Conference on Web and Social Media*, 6(1):26–33, August 2021.

István Berkes, Weidong Liu, and Wei Biao Wu. Komlós–Major–Tusnády approximation under dependence. *The Annals of Probability*, 42(2):794–817, 2014.

Andreas Brandt. The Stochastic Equation $Y_{n+1} = A_n Y_n + B_n$ with Stationary Coefficients. *Advances in Applied Probability*, 18(1):211–220, 1986.

Peter J Brockwell and Richard A Davis. *Time Series: Theory and Methods*. Springer, second edition edition, 1991.

Donald L. Burkholder. Sharp inequalities for martingales and stochastic integrals. In *Colloque Paul Lévy sur les processus stochastiques*, number 157-158 in Astérisque, pages 75–94. Société mathématique de France, 1988.

Jacopo Cavazza, Pietro Morerio, Benjamin Haeffele, Connor Lane, Vittorio Murino, and Rene Vidal. Dropout as a Low-Rank Regularizer for Matrix Factorization. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pages 435–444. PMLR, 2018.

Likai Chen and Wei Biao Wu. Stability and asymptotics for autoregressive processes. *Electronic Journal of Statistics*, 10(2):3723–3751, 2016.

Xi Chen, Jason D. Lee, Xin T. Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020.

Giovanni Luca Ciampaglia, Azadeh Nematzadeh, Filippo Menczer, and Alessandro Flammini. How algorithmic popularity bias hinders or promotes quality. *Scientific Reports*, 8 (1):15951, October 2018.

Gabriel Clara, Sophie Langer, and Johannes Schmidt-Hieber. Dropout regularization versus l2-penalization in the linear model. *Journal of Machine Learning Research*, 25(204):1–48, 2024.

Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc' aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, Quoc Le, and Andrew Ng. Large Scale Distributed Deep Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Aaron Defazio, Ashok Cutkosky, Harsh Mehta, and Konstantin Mishchenko. Optimal Linear Decay Learning Rate Schedules and Further Refinements. *arXiv preprint*, October 2024. arXiv:2310.07831.

Persi Diaconis and David Freedman. Iterated Random Functions. *SIAM Review*, 41(1): 45–76, 1999.

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, Kevin Scaman, and Hoi-To Wai. Tight High Probability Bounds for Linear Stochastic Approximation with Fixed Stepsize. In *Advances in Neural Information Processing Systems*, volume 34, pages 30063–30074. Curran Associates, Inc., 2021.

Yixin Fang. Scalable statistical inference for averaged implicit stochastic gradient descent. *Scandinavian Journal of Statistics*, 46(4):987–1002, 2019.

Yixin Fang, Jinfeng Xu, and Lei Yang. Online Bootstrap Confidence Intervals for the Stochastic Gradient Descent Estimator. *Journal of Machine Learning Research*, 19:1–21, 2019.

Kelwin Fernandes, Pedro Vinagre, Paulo Cortez, and Pedro Sernadela. Online News Popularity. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C5NS3V.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016a.

Yarin Gal and Zoubin Ghahramani. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016b.

Wei Gao and Zhi-Hua Zhou. Dropout Rademacher complexity of deep neural networks. *Science China Information Sciences*, 59(7):072104, 2016.

Geoffrey E. Hinton, Nitish Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint*, 2012. arXiv:1207.0580.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

Sayar Karmakar and Wei Biao Wu. Optimal Gaussian Approximation for Multiple Time Series. *Statistica Sinica*, 30(3):1399–1417, 2020.

Edmund Kay and Anurag Agarwal. DropConnected neural network trained with diverse features for classifying heart sounds. In *2016 Computing in Cardiology Conference (CinC)*, pages 617–620, 2016.

J. Kiefer and J. Wolfowitz. Stochastic Estimation of the Maximum of a Regression Function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

J. Komlós, P. Major, and G. Tusnády. An approximation of partial sums of independent RV'-s, and the sample DF. I. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 32(1):111–131, 1975.

J. Komlós, P. Major, and G. Tusnády. An approximation of partial sums of independent RV's, and the sample DF. II. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 34(1):33–58, 1976.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Soumendra N. Lahiri. Theoretical comparisons of block bootstrap methods. *The Annals of Statistics*, 27(1):386–404, 1999.

Soumendra N. Lahiri. *Resampling Methods for Dependent Data*. Springer Series in Statistics. Springer, New York, NY, 2003.

Tengyuan Liang and Weijie J. Su. Statistical Inference for the Population Landscape via Moment-Adjusted Stochastic Gradients. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(2):431–456, 2019.

David McAllester. A PAC-Bayesian tutorial with a dropout bound. *arXiv preprint*, 2013. arXiv:1307.2118.

Poorya Mianjy and Raman Arora. On Dropout and Nuclear Norm Regularization. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4575–4584. PMLR, 2019.

Poorya Mianjy and Raman Arora. On Convergence and Generalization of Dropout Training. In *Advances in Neural Information Processing Systems*, volume 33, pages 21151–21161. Curran Associates, Inc., 2020.

Poorya Mianjy, Raman Arora, and Rene Vidal. On the Implicit Bias of Dropout. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3540–3548. PMLR, 2018.

Georg Ch. Pflug. Stochastic Minimization with Constant Step-Size: Asymptotic Laws. *SIAM Journal on Control and Optimization*, 24(4):655–666, 1986.

Hieu Pham and Quoc Le. AutoDropout: Learning Dropout Patterns to Regularize Deep Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9351–9359, 2021.

B. T. Polyak and A. B. Juditsky. Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Emmanuel Rio. Moment inequalities for sums of dependent random variables under projective conditions. *Journal of Theoretical Probability*, 22(1):146–163, 2009.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

David Ruppert. Efficient Estimations from a Slowly Convergent Robbins-Monro Process, 1988. Technical report, Cornell University Operations Research and Industrial Engineering.

Sergey Samsonov, Daniil Tiapkin, Alexey Naumov, and Eric Moulines. Improved High-Probability Bounds for the Temporal Difference Learning Algorithm via Exponential Stability. In *Proceedings of Thirty Seventh Conference on Learning Theory*, pages 4511–4547. PMLR, June 2024.

Albert Senen-Cerda and Jaron Sanders. Asymptotic Convergence Rate of Dropout on Shallow Linear Neural Networks. *SIGMETRICS Perform. Eval. Rev.*, 50(1):105–106, 2022.

Albert Senen-Cerda and Jaron Sanders. Almost Sure Convergence of Dropout Algorithms for Neural Networks. *arXiv preprint*, 2023. arXiv:2002.02247.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Weijie Su and Yuancheng Zhu. HiGrad: Uncertainty Quantification for Online Learning and Stochastic Approximation. *Journal of Machine Learning Research*, 24:1–53, 2023.

Gabor Szabo and Bernardo A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88, August 2010.

Stefan Wager, Sida Wang, and Percy S Liang. Dropout Training as Adaptive Regularization. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of Neural Networks using DropConnect. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1058–1066. PMLR, 2013.

Colin Wei, Sham Kakade, and Tengyu Ma. The Implicit and Explicit Regularization Effects of Dropout. In *Proceedings of the 37th International Conference on Machine Learning*, pages 10181–10192. PMLR, 2020.

Haibing Wu and Xiaodong Gu. Towards dropout training for convolutional neural networks. *Neural Networks*, 71:1–10, 2015.

Wei Biao Wu. Nonlinear system theory: Another look at dependence. *PNAS*, 102(40): 14150–14154, 2005.

Wei Biao Wu. Asymptotic theory for stationary processes. *Statistics and Its Interface*, 4 (2):207–226, 2011.

Wei Biao Wu and Xiaofeng Shao. Limit Theorems for Iterated Random Functions. *Journal of Applied Probability*, 41(2):425–436, 2004.

Han Xiao and Wei Biao Wu. A Single-Pass Algorithm for Spectrum Estimation With Fast Convergence. *IEEE Transactions on Information Theory*, 57(7):4720–4731, 2011.

Ke Zhai and Huan Wang. Adaptive dropout with rademacher complexity regularization. In *International Conference on Learning Representations*, 2018.

Yanjie Zhong, Jiaqi Li, and Soumendra N. Lahiri. Probabilistic guarantees of stochastic recursive gradient in non-convex finite sum problems. In *Advances in Knowledge Discovery and Data Mining*, pages 142–154. Springer Nature Singapore, 2024.

Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online Covariance Matrix Estimation in Stochastic Gradient Descent. *Journal of the American Statistical Association*, 118(541):393–404, 2023.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized Stochastic Gradient Descent. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

Appendix A. Technical Lemmas

Lemma 20 (Burkholder, 1988; Rio, 2009) *Let $q > 1, q' = \min\{q, 2\}$, and define $M_T = \sum_{t=1}^{T} \xi_t$, where ξ_t are martingale differences with a finite q-th moment. Then*

$$\big(\mathbb{E}\|M_T\|_2^q\big)^{q'/q} \le K_q^{q'} \sum_{t=1}^{T} \big(\mathbb{E}\|\xi_T\|_2^q\big)^{q'/q}, \quad \text{where } K_q = \max\{(q-1)^{-1}, \sqrt{q-1}\}.$$

Lemma 21 (Clara et al., 2024) *For any matrices A and B in $\mathbb{R}^{d\times d}$, $p \in (0,1)$, and a diagonal matrix $D \in \mathbb{R}^{d\times d}$, the following results hold:*

(i) $\overline{AD} = \overline{A}D$, $\overline{DA} = D\overline{A}$, and $\overline{A}_p = p\overline{A} = \overline{A_p}$;

If in addition, the diagonal matrix D is random and independent of A and B, with the diagonal entries satisfying $D_{ii} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$, $1 \le i \le d$, then,

(ii) $\mathbb{E}[DAD] = pA_p$, where $A_p = pA + (1-p)\text{Diag}(A)$;

(iii) $\mathbb{E}[DADBD] = pA_pB_p + p^2(1-p)\text{Diag}(\overline{A}B)$, where $\overline{A} = A - \text{Diag}(A)$;

(iv) $\mathbb{E}[DADBDCD] = pA_pB_pC_p + p^2(1-p)\big[\text{Diag}(\overline{A}B_p\overline{C}) + A_p\text{Diag}(\overline{B}C) + \text{Diag}(A\overline{B})C_p + (1-p)A \odot \overline{B}^\top \odot C\big]$, where $\odot$ denotes the Hadamard product.

Lemma 22 (Properties of operator norm) *Let $A = (a_{ij})_{1\le i,j\le d}$ be a real $d\times d$ matrix. View A as a linear map $\mathbb{R}^d \mapsto \mathbb{R}^d$ and denote its operator norm by $\|A\|$.*

(i) (Inequalities for variants of A). $\|\text{Diag}(A)\| \le \|A\|$, $\|A_p\| \le \|A\|$, and if in addition, A is positive semi-definite, then also $\|\overline{A}\| \le \|A\|$;

(ii) (Frobenius norm). $\|A\| = \sup_{\boldsymbol{v}\in\mathbb{R}^d, \|\boldsymbol{v}\|_2=1} \|A\boldsymbol{v}\|_2 \le \|A\|_F$, where $\|A\|_F$ denotes the Frobenius norm, i.e., $\|A\|_F = \big(\sum_{i,j=1}^d |a_{ij}|^2\big)^{1/2}$;

(iii) (Largest magnitude of eigenvalues). For a symmetric $d\times d$ matrix A, we have $\max_{1\le i\le d} |\lambda_i(A)| = \|A\|$, where $\lambda_i(A)$ denotes the i-th largest eigenvalue of A. If in addition, A is positive semi-definite, then also $\lambda_{\max}(A) = \|A\|$.

Proof The inequalities in (i) follow directly from Lemma 19 in Clara et al. (2024). For (ii), we notice that for any unit vector $\boldsymbol{v} \in \mathbb{R}^d$, one can find a basis $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$ and write $\boldsymbol{v}$ into $\boldsymbol{v} = \sum_{j=1}^d c_j\boldsymbol{e}_j$, with $\boldsymbol{e}_j \in \mathbb{R}^d$, and the real coefficients c_j satisfying $\sum_{j=1}^d c_j^2 = 1$. Then, it follows from the orthogonality of $\boldsymbol{e}_j$ and the Cauchy-Schwarz inequality that

$$\begin{aligned}\|A\boldsymbol{v}\|_2^2 &= \Big\|\sum_{j=1}^d c_jA\boldsymbol{e}_j\Big\|_2^2 \le \Big(\sum_{j=1}^d |c_j|\|A\boldsymbol{e}_j\|_2\Big)^2 \le \Big(\sum_{j=1}^d |c_j|^2\Big)\Big(\sum_{j=1}^d \|A\boldsymbol{e}_j\|_2^2\Big)\\ &= \sum_{j=1}^d \|A\boldsymbol{e}_j\|_2^2 = \|A\|_F^2. \end{aligned} \tag{65}$$

Since this result holds for any unit vector $\boldsymbol{v} \in \mathbb{R}^d$, the desired result in (ii) is achieved.

(iii) is well-known. Nevertheless, we provide a proof here for completeness. For any eigenvalue $\lambda_i(A)$, denote its associated unit eigenvector by $\boldsymbol{v}$. Then, $\|A\boldsymbol{v}\|_2 = |\lambda_i(A)|\|\boldsymbol{v}\|_2 = |\lambda_i(A)|$, which further yields $|\lambda_i(A)| \le \sup_{\boldsymbol{v}\in\mathbb{R}^d, \|\boldsymbol{v}\|_2=1} \|A\boldsymbol{v}\|_2 = \|A\|$, uniformly over i. Hence, the inequality can be obtained. If in addition, A is symmetric, then A can be diagonalized by an orthogonal matrix Q and a diagonal matrix Λ such that $A = Q^\top\Lambda Q$.

Therefore, $\sup_{\boldsymbol{v}\in\mathbb{R}^d,\|\boldsymbol{v}\|_2=1}\|A\boldsymbol{v}\|_2 = \max_{1\le i\le d}|\lambda_i(A)|$, which completes the proof. ■

Appendix B. Proofs in Section 3.1

This section is devoted to the proofs of the geometric-moment contraction (GMC) for the dropout iterates with gradient descent (GD), i.e., $\{\tilde{\boldsymbol{\beta}}_k(\alpha)-\boldsymbol{\beta}\}_{k\in\mathbb{N}}$. We first extend the results in Wu and Shao (2004) to the cases where the inputs of iterated random functions are i.i.d. random matrices. Then, we present the proof for the sufficient condition of the GMC in terms of the constant learning rate α in Lemma 2, and showcase the GMC of $\{\tilde{\boldsymbol{\beta}}_k(\alpha)-\boldsymbol{\beta}\}_{k\in\mathbb{N}}$ in Theorem 3.

B.1 GMC – Random Matrix Version

Let $(\mathcal{Y},\rho)$ be a complete and separable metric space, endowed with its Borel sets $\mathbb{Y}$. Consider an iterated random function on the state space $\mathcal{Y}\subset\mathbb{R}^d$, for some fixed $d\ge 1$, with the form

$$\boldsymbol{y}_k = f(\boldsymbol{y}_{k-1},X_k) = f_{X_k}(\boldsymbol{y}_{k-1}),\quad k\in\mathbb{N}, \tag{66}$$

where $f(\boldsymbol{y},\cdot)$ is the $\boldsymbol{y}$-section of a jointly measurable function $f:\mathcal{Y}\times\mathcal{X}\mapsto\mathcal{Y}$; the random matrices X_k, $k\in\mathbb{N}$, take values in a second measurable space $\mathcal{X}\subset\mathbb{R}^{d\times d}$, and are independently distributed with identical marginal distribution H. The initial point $\boldsymbol{y}_0\in\mathcal{Y}$ is independent of all X_k.

We are interested in the sufficient conditions on $f_X(\boldsymbol{y})$ such that there is a unique stationary probability π on $\mathcal{Y}$ with $\boldsymbol{y}_k\Rightarrow\pi$ as $k\to\infty$. To this end, define a composite function

$$\boldsymbol{y}_k(\boldsymbol{y}) = f_{X_k}\circ f_{X_{k-1}}\circ\cdots\circ f_{X_1}(\boldsymbol{y}),\quad \text{for } \boldsymbol{y}\in\mathcal{Y}. \tag{67}$$

We say that $\boldsymbol{y}_k$ is *geometric-moment contracting* if for any two independent random vectors $\boldsymbol{y}\sim\pi$ and $\boldsymbol{y}'\sim\pi$ in $\mathcal{Y}$, there exist some $q>0$, $C_q>0$ and $r_q\in(0,1)$, such that for all $k\in\mathbb{N}$,

$$\mathbb{E}\big[\rho^q\big(\boldsymbol{y}_k(\boldsymbol{y}),\boldsymbol{y}_k(\boldsymbol{y}')\big)\big]\le C_q r_q^k. \tag{68}$$

Wu and Shao (2004) provided the sufficient conditions for (68) when X_k are random variables and $\boldsymbol{y}_k$ and $\boldsymbol{y}$ are one-dimensional. Their results can be directly extended to the random matrix version and we state them here for the completeness of this paper.

Assumption 2 (Finite moment) *Assume that there exists a fixed vector $\boldsymbol{y}^*\in\mathcal{Y}$ and some $q>0$ such that*

$$I(q,\boldsymbol{y}^*) := \mathbb{E}_{X\sim H}\big[\rho^q\big(\boldsymbol{y}^*,f_X(\boldsymbol{y}^*)\big)\big] = \int_{\mathcal{X}}\rho^q\big(\boldsymbol{y}^*,f_X(\boldsymbol{y}^*)\big)H(dX)<\infty.$$

Assumption 3 (Stochastic Lipschitz continuity) *Assume that there exists some $q>0$ and some $\boldsymbol{y}_0\in\mathcal{Y}$ such that*

$$L_q := \sup_{\boldsymbol{y}_0'\in\mathcal{Y},\,\boldsymbol{y}_0\neq\boldsymbol{y}_0'}\frac{\mathbb{E}_{X\sim H}\big[\rho^q\big(f_X(\boldsymbol{y}_0),f_X(\boldsymbol{y}_0')\big)\big]}{\rho^q(\boldsymbol{y}_0,\boldsymbol{y}_0')}<1,$$

where $L_q=L_q(\boldsymbol{y}_0)$ is a local Lipschitz constant.

Because of the affine form of the recursion in (36), the ergodic theory in Diaconis and Freedman (1999) also applies here for the SGD sequence with dropout.

Corollary 23 (GMC – random matrix version) *Suppose that Assumptions 2 and 3 hold. Define a backward iteration process*

$$\boldsymbol{z}_k(\boldsymbol{y}) = \boldsymbol{z}_{k-1}(f_{X_k}(\boldsymbol{y})) = f_{X_1} \circ f_{X_2} \circ \cdots \circ f_{X_k}(\boldsymbol{y}), \quad \textit{for } \boldsymbol{y} \in \mathcal{Y}. \tag{69}$$

Then, $\boldsymbol{z}_k(\boldsymbol{y}) \stackrel{\mathcal{D}}{=} \boldsymbol{y}_k(\boldsymbol{y})$*, and there exists a random vector* $\boldsymbol{z}_\infty \in \mathcal{Y}$ *such that for any* $\boldsymbol{y} \in \mathcal{Y}$*,*

$$\boldsymbol{z}_k(\boldsymbol{y}) \stackrel{a.s.}{\to} \boldsymbol{z}_\infty.$$

The limit $\boldsymbol{z}_\infty$ *is measurable with respect to the* σ*-algebra* $\sigma(X_1, X_2, \ldots)$ *and does not depend on* $\boldsymbol{y}$*. In addition,*

$$\mathbb{E}\big[\rho^q\big(\boldsymbol{z}_k(\boldsymbol{y}), \boldsymbol{z}_\infty\big)\big] \le C r_q^k, \quad \textit{for all } k \in \mathbb{N}, \tag{70}$$

where the constant $C > 0$ *only depends on* q *and* $\boldsymbol{y}^*$ *in Assumption 2,* L_q *in Assumption 3 and* $\boldsymbol{y}_0$*. Additionally, (68) holds.*

Before showing the detailed proof of Corollary 23, we first introduce a key technique for this proof in the following remark.

Remark 24 (Backward iteration) *We shall comment on the intuition for defining the backward iteration* $\boldsymbol{z}_k$ *in (69). Recall the i.i.d. random samples* $X_1, \ldots, X_n$*. Clearly, for any fixed initial point* $\boldsymbol{y}_0 \in \mathcal{Y}$*, for all* $k \in \mathbb{N}$*, we have the relations*

$$\begin{aligned} \boldsymbol{y}_{k+1}(\boldsymbol{y}_0) &= f_{X_{k+1}}\big(\boldsymbol{y}_k(\boldsymbol{y}_0)\big), \\ \boldsymbol{z}_{k+1}(\boldsymbol{y}_0) &= \boldsymbol{z}_k\big(f_{X_{k+1}}(\boldsymbol{y}_0)\big). \end{aligned}$$

To prove the existence of the limit for $\boldsymbol{y}_k = f_{X_k} \circ f_{X_{k-1}} \circ \cdots \circ f_{X_1}(\boldsymbol{y}_0)$*, we need to make use of the contracting property of the function* $f_X(\cdot)$ *stated in Assumption 3. However, we cannot directly apply it to the forward iteration, because by the Markov property, given the present position of the chain, the conditional distribution of the future does not depend on the past. This indicates*

$$\mathbb{E}\big[\rho^q\big(\boldsymbol{y}_{k+1}(\boldsymbol{y}^*), \boldsymbol{y}_k(\boldsymbol{y}^*)\big)\big] = \mathbb{E}\big[\mathbb{E}\big[\rho^q\big(f_{X_{k+1}}\big(\boldsymbol{y}_k(\boldsymbol{y}^*)\big), \boldsymbol{y}_k(\boldsymbol{y}^*)\big) \mid X_{k+1}\big]\big], \tag{71}$$

where the two parts inside of $\rho(\cdot,\cdot)$ *are operated by two different functions, which are* $f_{X_{k+1}}(\cdot)$ *and* $f_{X_k} \circ f_{X_{k-1}} \circ \cdots \circ f_{X_1}(\cdot)$ *respectively. In fact, as pointed out by Diaconis and Freedman (1999), the forward iteration* $\boldsymbol{y}_k$ *moves ergodically through* $\mathcal{Y}$*, which behaves quite differently from the backward iteration* $\boldsymbol{z}_k(\cdot) = f_{X_1} \circ f_{X_2} \circ \cdots \circ f_{X_k}(\cdot)$ *in (69), which does converge to a limit. To see this, we note that by Assumptions 2 and 3, there exists some* $\boldsymbol{y}^* \in \mathcal{Y}$ *such that*

$$\begin{aligned} \mathbb{E}\big[\rho^q\big(\boldsymbol{z}_{k+1}(\boldsymbol{y}^*), \boldsymbol{z}_k(\boldsymbol{y}^*)\big)\big] &= \mathbb{E}\big[\mathbb{E}\big[\rho^q\big(\boldsymbol{z}_k\big(f_{X_{k+1}}(\boldsymbol{y}^*)\big), \boldsymbol{z}_k(\boldsymbol{y}^*)\big) \mid X_{k+1}\big]\big] \\ &\le L_q^k \mathbb{E}\big[\rho^q\big(f_{X_{k+1}}(\boldsymbol{y}^*), \boldsymbol{y}^*\big)\big] \\ &= L_q^k I(q, \boldsymbol{y}^*), \end{aligned} \tag{72}$$

which is summable over k by Assumption 3. Since $\mathcal{Y}$ is a complete space, we can mimic the idea of a Cauchy sequence to prove the existence of the limit $\boldsymbol{z}_\infty$ and further show $\boldsymbol{z}_k \overset{a.s.}{\to} \boldsymbol{z}_\infty$, by applying the Borel-Cantelli lemma. Since $X_1, \ldots, X_n$ are i.i.d. and thus exchangeable, we have $\boldsymbol{z}_k(\boldsymbol{y}_0) \overset{\mathcal{D}}{=} \boldsymbol{y}_k(\boldsymbol{y}_0)$. Hence, we can show that $\boldsymbol{y}_k$ also converges to $\boldsymbol{z}_\infty$ in distribution.

Proof Let $q \in (0, 1]$ such that both Assumptions 2 and 3 hold. We will only show the desired results for this choice of q, since if Assumptions 2 and 3 are satisfied for some $q > 1$, then they are also valid for all $q \leq 1$ by Hölder's inequality (Wu and Shao, 2004). Recall the definition of integral $I(q, \boldsymbol{y}^*)$ in Assumption 2. Let $\boldsymbol{y}_0 \in \mathcal{Y}$ satisfy Assumption 3. Then,

$$\begin{aligned} I(q, \boldsymbol{y}_0) &= \mathbb{E}\big[\rho^q\big(\boldsymbol{y}_0, f_X(\boldsymbol{y}_0)\big)\big] \\ &\leq \mathbb{E}\big[\rho(\boldsymbol{y}_0, \boldsymbol{y}^*) + \rho\big(\boldsymbol{y}^*, f_X(\boldsymbol{y}^*)\big) + \rho\big(f_X(\boldsymbol{y}^*), f_X(\boldsymbol{y}_0)\big)\big]^q \\ &\leq \rho^q(\boldsymbol{y}_0, \boldsymbol{y}^*) + I(q, \boldsymbol{y}^*) + \mathbb{E}\big[\rho^q\big(f_X(\boldsymbol{y}^*), f_X(\boldsymbol{y}_0)\big)\big] \\ &\leq \rho^q(\boldsymbol{y}_0, \boldsymbol{y}^*) + I(q, \boldsymbol{y}^*) + L_q \rho^q(\boldsymbol{y}^*, \boldsymbol{y}_0) < \infty, \end{aligned}$$

where the first inequality follows from the triangle inequality, the second one is by Assumption 3 and Jensen's inequality, and the last one is due to Assumption 2. A similar argument as in (72) yields

$$\mathbb{E}\big[\rho^q\big(\boldsymbol{z}_{k+1}(\boldsymbol{y}_0), \boldsymbol{z}_k(\boldsymbol{y}_0)\big)\big] \leq L_q^k I(q, \boldsymbol{y}_0) =: \delta_k, \tag{73}$$

where $\delta_k = \delta_k(q, \boldsymbol{y}_0)$ solely depends on k, q, L_q and $\boldsymbol{y}_0$. By Markov's inequality, we have

$$\mathbb{P}\Big(\rho\big(\boldsymbol{z}_{k+1}(\boldsymbol{y}_0), \boldsymbol{z}_k(\boldsymbol{y}_0)\big) \geq \delta_k^{1/(2q)}\Big) \leq \delta_k^{1/2}. \tag{74}$$

Since $\sum_{k=1}^\infty \delta_k^{1/2} < \infty$, it follows from the first Borel-Cantelli lemma that

$$\mathbb{P}\Big(\rho\big(\boldsymbol{z}_{k+1}(\boldsymbol{y}_0), \boldsymbol{z}_k(\boldsymbol{y}_0)\big) \geq \delta_k^{1/(2q)} \text{ for infinitely many } k\Big) = 0. \tag{75}$$

Again, since $\delta_k^{1/2}$ is summable, $\boldsymbol{z}_k$ is a Cauchy sequence in space $\mathcal{Y}$, which together with the completeness of $\mathcal{Y}$ gives that almost surely, there exists a random vector $\boldsymbol{z}_\infty \in \mathcal{Y}$ such that

$$\boldsymbol{z}_k(\boldsymbol{y}_0) \overset{a.s.}{\to} \boldsymbol{z}_\infty, \quad \text{as } k \to \infty,$$

where $\boldsymbol{z}_\infty$ is $\sigma(X_1, X_2, \ldots)$-measurable. Let π be the probability distribution of $\boldsymbol{z}_\infty$.

Furthermore, it follows from the triangle inequality and Jensen's inequality that for any fixed $\boldsymbol{y}_0 \in \mathcal{Y}$,

$$\begin{aligned} \mathbb{E}\big[\rho^q\big(\boldsymbol{z}_k(\boldsymbol{y}_0), \boldsymbol{z}_\infty\big)\big] &\leq \mathbb{E}\Big[\sum_{l=0}^\infty \rho\big(\boldsymbol{z}_{k+1+l}(\boldsymbol{y}_0), \boldsymbol{z}_{k+l}(\boldsymbol{y}_0)\big)\Big]^q \\ &\leq \sum_{l=0}^\infty \mathbb{E}\big[\rho^q\big(\boldsymbol{z}_{k+1+l}(\boldsymbol{y}_0), \boldsymbol{z}_{k+l}(\boldsymbol{y}_0)\big)\big] \\ &\leq \delta_k/(1 - L_q), \end{aligned} \tag{76}$$

For any $\boldsymbol{y} \in \mathcal{Y}$, by Assumption 3 and triangle inequality,

$$\begin{aligned}\mathbb{E}\big[\rho^q\big(\boldsymbol{z}_k(\boldsymbol{y}), \boldsymbol{z}_\infty\big)\big] &\le \mathbb{E}\big[\rho^q\big(\boldsymbol{z}_k(\boldsymbol{y}), \boldsymbol{z}_k(\boldsymbol{y}_0)\big)\big] + \mathbb{E}\big[\rho^q\big(\boldsymbol{z}_k(\boldsymbol{y}_0), \boldsymbol{z}_\infty\big)\big] \\ &\le L_q^k \rho^q(\boldsymbol{y}_0, \boldsymbol{y}) + \delta_k/(1-L_q). \end{aligned} \tag{77}$$

Recall that $\delta_k = L_q^k I(q, \boldsymbol{y}_0)$. Let $C = I(q, \boldsymbol{y}_0)/(1-L_q) + \rho^q(\boldsymbol{y}_0, \boldsymbol{y})$ and we have shown result (70) with $r_q = L_q$. Since Cr_q^k in (70) is summable over k, it again follows from Borel-Cantelli lemma that for any $\boldsymbol{y} \in \mathcal{Y}$,

$$\boldsymbol{z}_k(\boldsymbol{y}) \overset{a.s.}{\to} \boldsymbol{z}_\infty, \quad \text{as } k \to \infty,$$

and therefore, the limit

$$\boldsymbol{v}_k(\boldsymbol{y}) = \lim_{m\to\infty} f_{X_{k+1}} \circ f_{X_{k+2}} \circ \cdots \circ f_{X_{k+m}}(\boldsymbol{y}) \tag{78}$$

exists almost surely.

Finally, we notice that for any two independent random vectors $\boldsymbol{y} \sim \pi$ and $\boldsymbol{y}' \sim \pi$,

$$\begin{aligned}\mathbb{E}\big[\rho^q\big(\boldsymbol{y}_k(\boldsymbol{y}), \boldsymbol{y}_k(\boldsymbol{y}')\big)\big] &\le \mathbb{E}\big[\rho^q\big(\boldsymbol{y}_k(\boldsymbol{y}), \boldsymbol{y}_k(\boldsymbol{y}_0)\big)\big] + \mathbb{E}\big[\rho^q\big(\boldsymbol{y}_k(\boldsymbol{y}_0), \boldsymbol{y}_k(\boldsymbol{y}')\big)\big] \\ &= 2\mathbb{E}\big[\rho^q\big(\boldsymbol{z}_k(\boldsymbol{v}_k), \boldsymbol{z}_k(\boldsymbol{y}_0)\big)\big] \\ &= 2\mathbb{E}\big[\rho^q\big(\boldsymbol{z}_\infty, \boldsymbol{z}_k(\boldsymbol{y}_0)\big)\big] \le 2\delta_k/(1-L_q), \end{aligned} \tag{79}$$

where the first equation follows from the observation that $\boldsymbol{v}_k$ has the identical distribution as $\boldsymbol{z}_\infty = \boldsymbol{z}_k(\boldsymbol{v}_k(\boldsymbol{y})) \sim \pi$ and is independent of i.i.d. random matrices $X_1, \ldots, X_k$ because $\boldsymbol{v}_k$ as defined in (78) only depends on X_i for large $i \ge k+1$. The desired result in (68) has been achieved. ■

The recursion $\boldsymbol{y}_k = f(\boldsymbol{y}_{k-1}, X_k)$ is only defined for positive integers k. Nevertheless, Corollary 23 guarantees that for $k = 0, -1, \ldots$ the relation $\boldsymbol{y}_k = f(\boldsymbol{y}_{k-1}, X_k)$ also holds. See Remark 2 in Wu and Shao (2004) for a simple way to define $\boldsymbol{y}_k$ when $k = 0, -1, \ldots$ in the one-dimensional case. The vector versions can be similarly constructed.

B.2 Proof of Lemma 2

Proof Let D be a dropout matrix with the same distribution as D_1. Since $\mathbb{X} = X^\top X$ is positive semi-definite and by assumption $\alpha\|\mathbb{X}\| < 2$, we have $-I_d < I_d - \alpha D\mathbb{X}D \le I_d$ and consequently $\|I_d - \alpha D\mathbb{X}D\| \le 1$. Thus for a unit vector $\boldsymbol{v}$, $\|(I_d - \alpha D_k\mathbb{X}D_k)\boldsymbol{v}\|_2 \le 1$. This means that for $q \ge 2$, we can use $\|\cdot\|_2^q = \|\cdot\|_2^2\|\cdot\|_2^{q-2}$ to bound

$$\begin{aligned} r_{\alpha,q}^q &\le \sup_{\boldsymbol{v}\in\mathbb{R}^d:\|\boldsymbol{v}\|_2=1} \mathbb{E}\Big\|\big(I_d - \alpha D\mathbb{X}D\big)\boldsymbol{v}\Big\|_2^2 \\ &= \sup_{\boldsymbol{v}\in\mathbb{R}^d:\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^\top \mathbb{E}\Big[\big(I_d - \alpha D\mathbb{X}D\big)^2\Big]\boldsymbol{v} \\ &= \Big\|\mathbb{E}\Big[\big(I_d - \alpha D\mathbb{X}D\big)^2\Big]\Big\|. \end{aligned} \tag{80}$$

For a $d \times d$ and positive semi-definite matrix A, we have $A^2 \leq \|A\|A$. To see this, let $\boldsymbol{v}_j$ be the eigenvectors of A with corresponding eigenvalues λ_j. Any vector $\boldsymbol{w}$ can be written as $\boldsymbol{w} = \gamma_1 \boldsymbol{v}_1 + \ldots + \gamma_d \boldsymbol{v}_d$ with coefficients $\gamma_1, \ldots, \gamma_d$. Now $\boldsymbol{w}^\top A^2 \boldsymbol{w} = \gamma_1^2 \lambda_1^2 + \ldots + \gamma_d^2 \lambda_d^2 \leq (\max_j \lambda_j)(\gamma_1^2 \lambda_1 + \ldots + \gamma_d^2 \lambda_d) = \boldsymbol{w}^\top \|A\| A \boldsymbol{w}$. Since $\boldsymbol{w}$ was arbitrary, this proves $A^2 \leq \|A\|A$. Moreover, recall that D_k is a diagonal matrix with diagonal entries 0 and 1. Thus $D_k^2 = D_k \leq I_d$. Because $\mathbb{X}$ is positive semi-definite and by assumption $\Delta := 2 - \alpha\|\mathbb{X}\| > 0$, we have $\alpha^2 D_1 \mathbb{X} D_1^2 \mathbb{X} D_1 \leq \alpha^2 D_1 \mathbb{X}^2 D_1 \leq \alpha^2 D_1 \|\mathbb{X}\| \mathbb{X} D_1 \leq (2 - \Delta) \alpha D_1 \mathbb{X} D_1$. Thus,

$$(I_d - \alpha D_1 \mathbb{X} D_1)^2 = I_d - 2\alpha D_1 \mathbb{X} D_1 + \alpha^2 D_1 \mathbb{X} D_1^2 \mathbb{X} D_1 \leq I_d - \Delta \alpha D_1 \mathbb{X} D_1.$$

Taking expectation and using Lemma 21 (ii) yields $\mathbb{E}\big[(I_d - \alpha D_1 \mathbb{X} D_1)^2\big] \leq I_d - \Delta \alpha p \mathbb{X}_p$. The fact that $\mathbb{E}\big[(I_d - \alpha D_1 \mathbb{X} D_1)^2\big]$ is positive semi-definite implies that $\|\mathbb{E}\big[(I_d - \alpha D_1 \mathbb{X} D_1)^2\big]\|$ is bounded by the largest eigenvalue of $I_d - \Delta \alpha p \mathbb{X}_p$. By definition, $\mathbb{X}_p = p\mathbb{X} + (1 - p)\mathrm{Diag}(\mathbb{X}) \geq (1 - p) \min_j \mathbb{X}_{jj} I_d$. By assumption the design is in reduced form which implies that $\min_j \mathbb{X}_{jj} > 0$. This shows that $\mathbb{X}_p$ is positive definite and the largest eigenvalue of $I_d - \Delta \alpha p \mathbb{X}_p$ must be strictly smaller than 1. This implies $\|\mathbb{E}[(I_d - \alpha D_1 \mathbb{X} D_1)^2]\| < 1$. Combined with (80) this proves $r_{\alpha,q} < 1$.

If Lemma 2 holds for some $q \geq 2$, then by Hölder's inequality, it also holds for all $1 < q < 2$. To see this, consider a unit vector $\boldsymbol{v} \in \mathbb{R}^d$ and set $r(q) := \big(\mathbb{E}\|(I_d - \alpha D_1 \mathbb{X} D_1)\boldsymbol{v}\|_2^q\big)^{1/q}$. Then for any $1 < q' < q$, it follows from Hölder's inequality that

$$r(q')^{q'} = \mathbb{E}\Big\|\big(I_d - \alpha D_1 \mathbb{X} D_1\big)\boldsymbol{v}\Big\|_2^{q'} \leq \Big(\mathbb{E}\Big\|\big(I_d - \alpha D_1 \mathbb{X} D_1\big)\boldsymbol{v}\Big\|_2^{q}\Big)^{q'/q} = r(q)^{q'} < 1. \tag{81}$$

The desired result is obtained. ■

Throughout the rest of this paper, when there is no ambiguity, we omit the dependence on α, writing $\tilde{\boldsymbol{\beta}}_k$ instead of $\tilde{\boldsymbol{\beta}}_k(\alpha)$, $A_k = A_k(\alpha)$, and $\boldsymbol{b}_k = \boldsymbol{b}_k(\alpha)$.

B.3 Proof of Theorem 3

Proof Recall the recursive estimator $\tilde{\boldsymbol{\beta}}_k$ defined in (6). Write $A_k = A_k(\alpha) = I_d - \alpha D_k \mathbb{X} D_k$. We consider arbitrary d-dimensional initialization vectors $\tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{\beta}}'_0 \in \mathbb{R}^d$ and write $\tilde{\boldsymbol{\beta}}_k$ and $\tilde{\boldsymbol{\beta}}'_k$ for the respective iterates (sharing the same dropout matrices). Now $\tilde{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}'_k = A_k(\tilde{\boldsymbol{\beta}}_{k-1} - \tilde{\boldsymbol{\beta}}'_{k-1}) =: A_k \Delta_{k-1}$, with independent A_k and Δ_{k-1}. By Lemma 2, $r := \sup_{\boldsymbol{v} \in \mathbb{R}^d, \|\boldsymbol{v}\|_2 = 1} \big(\mathbb{E}\|A_k \boldsymbol{v}\|_2^q\big)^{1/q} < 1$, and thus, for any fixed vector $\boldsymbol{v}$, $\mathbb{E}\|A_k \boldsymbol{v}\|_2^q \leq r^q \|\boldsymbol{v}\|_2^q$. Due to the independence between A_k and Δ_{k-1}, it follows from the tower rule and the condition above that

$$\mathbb{E}\|\tilde{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}'_k\|_2^q = \mathbb{E}\|A_k \Delta_{k-1}\|_2^q = \mathbb{E}\big[\mathbb{E}[\|A_k \Delta_{k-1}\|_2^q \mid \Delta_{k-1}]\big] \leq \mathbb{E}\big[r^q \|\Delta_{k-1}\|_2^q\big] = r^q \mathbb{E}\|\Delta_{k-1}\|_2^q.$$

Since $A_k(\alpha)$ are i.i.d. random matrices induction on k yields the claimed geometric-moment contraction $\big(\mathbb{E}\|\tilde{\boldsymbol{\beta}}_k(\alpha) - \tilde{\boldsymbol{\beta}}'_k(\alpha)\|_2^q\big)^{1/q} \leq r_{\alpha,q}^k \|\tilde{\boldsymbol{\beta}}_0 - \tilde{\boldsymbol{\beta}}'_0\|_2$.

Finally, recall the quantity $I(q, \boldsymbol{y}^*)$ defined in Assumption 2. For the recursion of the sequence $\{\tilde{\boldsymbol{\beta}}_k(\alpha) - \tilde{\boldsymbol{\beta}}\}_{k \in \mathbb{N}}$ in (6), we have $I(q, 0) = \mathbb{E}\|\boldsymbol{b}_k(\alpha)\|_2^q < \infty$. To see this, note that

$$\|\boldsymbol{b}_k(\alpha)\|_2 = \alpha \|D_k \overline{\mathbb{X}}(pI_d - D_k)\tilde{\boldsymbol{\beta}}\|_2 \leq \alpha \|\overline{\mathbb{X}}\| \|\tilde{\boldsymbol{\beta}}\|_2,$$

where $\|\cdot\|$ denotes the operator norm, and the last inequality holds since $\|D_k\| < 1$, $\|pI_d - D_k\| \le \max\{p, 1-p\} \le 1$. By Corollary 23, the geometric-moment contraction proved above implies the existence of a unique stationary distribution $\tilde{\pi}_\alpha$ of the GD dropout sequence $\tilde{\boldsymbol{\beta}}_k(\alpha)$. This completes the proof. ■

Appendix C. Proofs in Section 3.2

C.1 Proof of Lemma 4

Proof Since $\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)$ is stationary and $\mathbb{E}[\boldsymbol{b}_k] = 0$ by (100), it follows that $\tilde{\boldsymbol{\beta}}_k^\circ(\alpha) - \tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_k^\dagger(\alpha) - \tilde{\boldsymbol{\beta}}$ both have zero mean, and thus $\mathbb{E}[\boldsymbol{\delta}_k(\alpha)] = 0$.

To prove the second claim, we first note that

$$\begin{aligned}\boldsymbol{\delta}_k(\alpha) &= (I_d - \alpha D_k \mathbb{X} D_k)(\tilde{\boldsymbol{\beta}}_{k-1}^\circ(\alpha) - \tilde{\boldsymbol{\beta}}) + \boldsymbol{b}_k(\alpha) - \big[(I_d - \alpha p \mathbb{X}_p)(\boldsymbol{\beta}_{k-1}^\dagger(\alpha) - \tilde{\boldsymbol{\beta}}) + \boldsymbol{b}_k(\alpha)\big] \\ &= (I_d - \alpha p \mathbb{X}_p)\boldsymbol{\delta}_{k-1} + \alpha(p\mathbb{X}_p - D_k \mathbb{X} D_k)(\tilde{\boldsymbol{\beta}}_{k-1}^\circ - \tilde{\boldsymbol{\beta}}) \end{aligned} \tag{82}$$

is a stationary sequence. By induction on k, we can write $\boldsymbol{\delta}_k(\alpha)$ into

$$\begin{aligned}\boldsymbol{\delta}_k(\alpha) &= \alpha\Big[(p\mathbb{X}_p - D_k\mathbb{X}D_k)(\tilde{\boldsymbol{\beta}}_{k-1}^\circ - \tilde{\boldsymbol{\beta}}) + \cdots \\ &\qquad + (p\mathbb{X}_p - D_1\mathbb{X}D_1)(I_d - \alpha p\mathbb{X}_p)^{k-1}(\tilde{\boldsymbol{\beta}}_0^\circ - \tilde{\boldsymbol{\beta}}) + \cdots\Big] \\ &= \alpha\sum_{i=1}^\infty (p\mathbb{X}_p - D_{k-i+1}\mathbb{X}D_{k-i+1})(I_d - \alpha p\mathbb{X}_p)^{i-1}(\tilde{\boldsymbol{\beta}}_{k-i}^\circ - \tilde{\boldsymbol{\beta}}) \\ &=: \alpha\sum_{i=1}^\infty \mathcal{M}_{k-i}(\alpha). \end{aligned} \tag{83}$$

For any $k \in \mathbb{N}$, $\{\mathcal{M}_{k-i}(\alpha)\}_{i\ge 1}$ is a sequence of martingale differences with respect to the filtration $\mathcal{F}_{k-i} = \sigma(\ldots, D_{k-i-1}, D_{k-i})$, since the dropout matrix D_k is independent of $\tilde{\boldsymbol{\beta}}_{k-1}^\circ$ and $\tilde{\boldsymbol{\beta}}$. Therefore, we can apply Burkholder's inequality in Lemma 20 to $\sum_{i=1}^\infty \mathcal{M}_{k-i}(\alpha)$, and obtain, for $q \ge 2$,

$$\begin{aligned}&\Big(\mathbb{E}\Big\|\sum_{i=1}^\infty \mathcal{M}_{k-i}(\alpha)\Big\|_2^q\Big)^{1/q} \\ &= \Big(\mathbb{E}\Big\|\sum_{i=1}^\infty (I_d - \alpha p\mathbb{X}_p)^{i-1}(p\mathbb{X}_p - D_{k-i+1}\mathbb{X}D_{k-i+1})(\tilde{\boldsymbol{\beta}}_{k-i}^\circ - \tilde{\boldsymbol{\beta}})\Big\|_2^q\Big)^{1/q} \\ &\lesssim \Big[\sum_{i=1}^\infty \big(\mathbb{E}\big\|(I_d - \alpha p\mathbb{X}_p)^{i-1}(p\mathbb{X}_p - D_{k-i+1}\mathbb{X}D_{k-i+1})(\tilde{\boldsymbol{\beta}}_{k-i}^\circ - \tilde{\boldsymbol{\beta}})\big\|_2^q\big)^{2/q}\Big]^{1/2} \\ &\le \Big[\sum_{i=1}^\infty \|I_d - \alpha p\mathbb{X}_p\|^{2(i-1)}\big(\mathbb{E}\big\|(p\mathbb{X}_p - D_{k-i+1}\mathbb{X}D_{k-i+1})(\tilde{\boldsymbol{\beta}}_{k-i}^\circ - \tilde{\boldsymbol{\beta}})\big\|_2^q\big)^{2/q}\Big]^{1/2}, \end{aligned}$$

where the constant in $\lesssim$ here and the rest of the proof only depends on q unless it is additionally specified.

We shall proceed the proof with two main steps. First, we show the bound $\|I_d-\alpha p\mathbb{X}_p\| < 1$ for the operator norm and thus $\sum_{i=1}^{\infty}\|I_d-\alpha p\mathbb{X}_p\|^{2(i-1)} < \infty$. Second, we provide a bound for $\mathbb{E}\big\|(p\mathbb{X}_p - D_k\mathbb{X}D_k)(\tilde{\boldsymbol{\beta}}^{\circ}_{k-1}-\tilde{\boldsymbol{\beta}})\big\|_2^q$ uniformly over k.

Step 1. Since $\alpha\|\mathbb{X}\| < 2$, it follows from Lemma 22 (i) that $\alpha\|\mathbb{X}_p\| \le \alpha\|\mathbb{X}\| < 2$. Moreover, the assumption that the design matrix X has no zero columns guarantees that all diagonal entries of $\mathbb{X}$ are positive and thus $\mathrm{Diag}(\mathbb{X}) > 0$. Together with $p < 1$, this lead to $\mathbb{X}_p = p\mathbb{X} + (1-p)\mathrm{Diag}(\mathbb{X}) \ge (1-p)\mathrm{Diag}(\mathbb{X}) > 0$. We thus have $-I_d < I_d - \alpha p\mathbb{X}_p < I_d$. Consequently, $\|I_d - \alpha p\mathbb{X}_p\| < 1$ and

$$\sum_{i=1}^{\infty}\|I_d-\alpha p\mathbb{X}_p\|^{2(i-1)} = \frac{1}{1-\|I_d-\alpha p\mathbb{X}_p\|^2} = O\big(\alpha^{-1}\big). \tag{84}$$

Step 2. Next, we shall bound the term $\mathbb{E}\big\|(p\mathbb{X}_p-D_i\mathbb{X}D_i)(\tilde{\boldsymbol{\beta}}^{\circ}_{i-1}-\tilde{\boldsymbol{\beta}})\big\|_2^q$. We first consider the case $q=2$. Denote $\mathbb{M}_i = p\mathbb{X}_p - D_i\mathbb{X}D_i$. Using that $\mathbb{E}[D_i\mathbb{X}D_i] = p\mathbb{X}_p$, we find $\mathbb{E}[\mathbb{M}_i] = 0$ and by the tower rule,

$$\begin{aligned}\mathbb{E}\big\|(p\mathbb{X}_p - D_i\mathbb{X}D_i)(\tilde{\boldsymbol{\beta}}^{\circ}_{i-1}-\tilde{\boldsymbol{\beta}})\big\|_2^2 &= \mathbb{E}[(\tilde{\boldsymbol{\beta}}^{\circ}_{i-1}-\tilde{\boldsymbol{\beta}})^{\top}\mathbb{M}_i^{\top}\mathbb{M}_i(\tilde{\boldsymbol{\beta}}^{\circ}_{i-1}-\tilde{\boldsymbol{\beta}})]\\ &= \mathbb{E}\big[\mathbb{E}\big[(\tilde{\boldsymbol{\beta}}^{\circ}_{i-1}-\tilde{\boldsymbol{\beta}})^{\top}\mathbb{M}_i^{\top}\mathbb{M}_i(\tilde{\boldsymbol{\beta}}^{\circ}_{i-1}-\tilde{\boldsymbol{\beta}}) \mid \mathcal{F}_{i-1}\big]\big]\\ &\le \|\mathbb{E}[\mathbb{M}_i^{\top}\mathbb{M}_i]\|\cdot\mathbb{E}\|\tilde{\boldsymbol{\beta}}^{\circ}_{i-1}-\tilde{\boldsymbol{\beta}}\|_2^2. \end{aligned}\tag{85}$$

By Lemma 5, we have $\mathbb{E}\|\tilde{\boldsymbol{\beta}}^{\circ}_{i-1}-\tilde{\boldsymbol{\beta}}\|_2^2 = O(\alpha)$. We only need to bound the operator norm $\|\mathbb{E}[\mathbb{M}_i^{\top}\mathbb{M}_i]\|$. To this end, we use again $\mathbb{E}[D_i\mathbb{X}D_i] = p\mathbb{X}_p$ and moreover $\mathbb{E}[D_i\mathbb{X}D_i\mathbb{X}D_i] = p\mathbb{X}_p^2 + p^2(1-p)\mathrm{Diag}(\overline{\mathbb{X}}\mathbb{X})$, which yields,

$$\begin{aligned}\mathbb{E}[\mathbb{M}_i^{\top}\mathbb{M}_i] &= \mathbb{E}[(p\mathbb{X}_p - D_i\mathbb{X}D_i)^{\top}(p\mathbb{X}_p - D_i\mathbb{X}D_i)]\\ &= p^2\mathbb{X}_p^2 - 2p^2\mathbb{X}_p^2 + p\mathbb{X}_p^2 + p^2(1-p)\mathrm{Diag}(\overline{\mathbb{X}}\mathbb{X})\\ &= p^2(1-p)\mathrm{Diag}(\overline{\mathbb{X}}\mathbb{X}).\end{aligned}\tag{86}$$

Recall that $\mathbb{X} = X^{\top}X$, where X is the fixed design matrix. Then, by Lemma 22 (i) and the sub-multiplicativity of the operator norm, we have $\|\mathrm{Diag}(\overline{\mathbb{X}}\mathbb{X})\| \le \|\overline{\mathbb{X}}\mathbb{X}\| \le \|\overline{\mathbb{X}}\|\|\mathbb{X}\| \le \|\mathbb{X}\|^2$. As a direct consequence, $\|\mathbb{E}[\mathbb{M}_i^{\top}\mathbb{M}_i]\| \le p^2(1-p)\|\mathbb{X}\|^2 < \infty$, which together with Lemma 5 and (84) gives

$$\mathbb{E}\|\boldsymbol{\delta}_k(\alpha)\|_2 = \Big(\mathbb{E}\Big\|\alpha\sum_{i=1}^{\infty}\mathcal{M}_{k-i}(\alpha)\Big\|_2^2\Big)^{1/2} \lesssim \alpha\Big(\sum_{i=1}^{\infty}\|I_d-\alpha p\mathbb{X}_p\|^{2(i-1)}\alpha\Big)^{1/2} = O(\alpha),$$

uniformly over k. For the case with $q > 2$, we can similarly apply the tower rule and obtain

$$\begin{aligned}&\mathbb{E}\big\|(p\mathbb{X}_p - D_i\mathbb{X}D_i)(\tilde{\boldsymbol{\beta}}^{\circ}_{i-1}-\tilde{\boldsymbol{\beta}})\big\|_2^q\\ &= \mathbb{E}\Big[\mathbb{E}\big[\|(p\mathbb{X}_p - D_i\mathbb{X}D_i)(\tilde{\boldsymbol{\beta}}^{\circ}_{i-1}-\tilde{\boldsymbol{\beta}})\|_2^q \mid \mathcal{F}_{i-1}\big]\Big]\\ &\le \sup_{\boldsymbol{v}\in\mathbb{R}^d,\|\boldsymbol{v}\|_2=1}\mathbb{E}\|(p\mathbb{X}_p - D_i\mathbb{X}D_i)\boldsymbol{v}\|_2^q\cdot\mathbb{E}\|\tilde{\boldsymbol{\beta}}^{\circ}_{i-1}-\tilde{\boldsymbol{\beta}}\|_2^q,\end{aligned}\tag{87}$$

where the last inequality can be achieved by writing $\tilde{\boldsymbol{\beta}}_{i-1}^{\circ} - \tilde{\boldsymbol{\beta}} = \|\tilde{\boldsymbol{\beta}}_{i-1}^{\circ} - \tilde{\boldsymbol{\beta}}\|_2 \boldsymbol{v}$. Here $\boldsymbol{v}$ is the unit vector $(\tilde{\boldsymbol{\beta}}_{i-1}^{\circ} - \tilde{\boldsymbol{\beta}})/\|\tilde{\boldsymbol{\beta}}_{i-1}^{\circ} - \tilde{\boldsymbol{\beta}}\|_2$ with $\|\boldsymbol{v}\|_2 = 1$. In addition, recall the Frobenius norm denoted by $\|\cdot\|_F$. It follows from Lemma 22 (i) and (ii) that

$$\begin{aligned}\sup_{\boldsymbol{v}\in\mathbb{R}^d, \|\boldsymbol{v}\|_2=1} \mathbb{E}\|(p\mathbb{X}_p - D_i\mathbb{X}D_i)\boldsymbol{v}\|_2^q &\le \mathbb{E}\|p\mathbb{X}_p - D_i\mathbb{X}D_i\|_F^q \\ &\lesssim \mathbb{E}\big(\|p\mathbb{X}_p\|_F^q + \|D_i\mathbb{X}D_i\|_F^q\big) \\ &\le (p^q+1)\|\mathbb{X}\|^q < \infty,\end{aligned}$$

where the constant in $\lesssim$ only depends on q. Combining this with the inequality (84), we obtain $(\mathbb{E}\|\boldsymbol{\delta}_k(\alpha)\|_2^q)^{1/q} = O(\alpha)$, completing the proof. ■

C.2 Proof of Lemma 5

Proof Recall that by applying induction on k to Equation (6), we can rewrite the GD dropout iterates $\tilde{\boldsymbol{\beta}}_k(\alpha)$ into

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_k(\alpha) - \tilde{\boldsymbol{\beta}} &= A_k(\alpha)(\tilde{\boldsymbol{\beta}}_{k-1}(\alpha) - \tilde{\boldsymbol{\beta}}) + \boldsymbol{b}_k(\alpha) \\ &= \sum_{i=0}^{k-1}\Big(\prod_{j=k-i+1}^{k} A_j(\alpha)\Big)\boldsymbol{b}_{k-i}(\alpha) + \Big(\prod_{j=1}^{k} A_j(\alpha)\Big)(\tilde{\boldsymbol{\beta}}_0(\alpha) - \tilde{\boldsymbol{\beta}}),\end{aligned}$$

where we set $\prod_{j=k+1}^{k} A_j(\alpha) = I_d$. Following Brandt (1986), since both A_k and $\boldsymbol{b}_k$ are i.i.d. random coefficients, the stationary solution $\{\tilde{\boldsymbol{\beta}}_k^{\circ}(\alpha) - \tilde{\boldsymbol{\beta}}\}_{k\in\mathbb{N}}$ of this recursion can be written into

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_k^{\circ}(\alpha) - \tilde{\boldsymbol{\beta}} &= A_k(\alpha)(\tilde{\boldsymbol{\beta}}_{k-1}^{\circ}(\alpha) - \tilde{\boldsymbol{\beta}}) + \boldsymbol{b}_k(\alpha) \\ &= \sum_{i=0}^{\infty}\Big(\prod_{j=k-i+1}^{k} A_j(\alpha)\Big)\boldsymbol{b}_{k-i}(\alpha) \\ &= \alpha\sum_{i=0}^{\infty}\Big[\prod_{j=k-i+1}^{k}(I_d - \alpha D_j\mathbb{X}D_j)\Big]D_{k-i}\overline{\mathbb{X}}(pI_d - D_{k-i})\tilde{\boldsymbol{\beta}} \\ &=: \alpha\sum_{i=0}^{\infty}\tilde{\mathcal{M}}_{i,k}(\alpha). \qquad (88)\end{aligned}$$

We observe that, for any $k \in \mathbb{N}$, $\{\tilde{\mathcal{M}}_{i,k}(\alpha)\}_{i\in\mathbb{N}}$ is a sequence of martingale differences with respect to the filtration $\mathcal{F}_{k-i} = \sigma(D_{k-i}, D_{k-i-1}, \ldots)$. Hence, it follows from Burkholder's inequality in Lemma 20 that, for $q \ge 2$,

$$\begin{aligned}\big(\mathbb{E}\|\tilde{\boldsymbol{\beta}}_k^{\circ}(\alpha) - \tilde{\boldsymbol{\beta}}\|_2^q\big)^{1/q} &= \alpha\Big(\mathbb{E}\Big\|\sum_{i=0}^{\infty}\tilde{\mathcal{M}}_{i,k}(\alpha)\Big\|_2^q\Big)^{1/q} \\ &\lesssim \alpha\Big(\sum_{i=0}^{\infty}\big(\mathbb{E}\|\tilde{\mathcal{M}}_{i,k}(\alpha)\|_2^q\big)^{2/q}\Big)^{1/2}, \qquad (89)\end{aligned}$$

where the constant in $\lesssim$ only depends on q. Recall H_k defined in (96), and we define a $d\times d$ matrix $B_{i,k}$ by

$$B_{i,k}=\Big[\prod_{j=k-i+1}^{k}(I_d-\alpha D_j\mathbb{X}D_j)\Big]D_{k-i}\overline{\mathbb{X}}(pI_d-D_{k-i})=\Big(\prod_{j=k-i+1}^{k}A_j\Big)H_{k-i}. \tag{90}$$

This random matrix is independent of $\tilde{\boldsymbol{\beta}}$. For $q=2$, by the tower rule, we have

$$\begin{aligned}\mathbb{E}\|\tilde{\mathcal{M}}_{i,k}(\alpha)\|_2^2&=\mathbb{E}\big[\mathbb{E}\big[\tilde{\boldsymbol{\beta}}^\top B_{i,k}^\top B_{i,k}\tilde{\boldsymbol{\beta}}\mid\mathcal{F}_k\big]\big]\\&=\mathbb{E}\big[\mathbb{E}\big[\mathrm{tr}(\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top B_{i,k}^\top B_{i,k})\mid\mathcal{F}_k\big]\big]\\&=\mathbb{E}\big[\mathrm{tr}\big(\mathbb{E}\big[\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top B_{i,k}^\top B_{i,k}\mid\mathcal{F}_k\big]\big)\big]\\&=\mathbb{E}\big[\mathrm{tr}\big(\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top B_{i,k}^\top B_{i,k}\big)\big]\\&=\mathrm{tr}\big(\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top\mathbb{E}[B_{i,k}^\top B_{i,k}]\big)\\&\le\|\mathbb{E}[B_{i,k}^\top B_{i,k}]\|\cdot\|\tilde{\boldsymbol{\beta}}\|_2^2.\end{aligned} \tag{91}$$

Following the similar arguments, we obtain for $q\ge 2$,

$$\mathbb{E}\|\tilde{\mathcal{M}}_{i,k}(\alpha)\|_2^q\le\sup_{\boldsymbol{v}\in\mathbb{R}^d,\|\boldsymbol{v}\|_2=1}\mathbb{E}\|B_{i,k}\boldsymbol{v}\|_2^q\cdot\|\tilde{\boldsymbol{\beta}}\|_2^q. \tag{92}$$

Moreover, we notice that by the tower rule

$$\begin{aligned}\|\mathbb{E}[B_{i,k}^\top B_{i,k}]\|&=\Big\|\mathbb{E}\Big[H_{k-i}^\top\Big(\prod_{j=k-i+1}^{k}A_j\Big)^\top\Big(\prod_{j=k-i+1}^{k}A_jH_{k-i}\Big)\Big]\Big\|\\&\le\|\mathbb{E}[H_{k-i}^\top A_k^\top A_kH_{k-i}]\|\cdot\Big\|\mathbb{E}\Big[\Big(\prod_{j=k-i+1}^{k-1}A_j\Big)^\top\Big(\prod_{j=k-i+1}^{k-1}A_j\Big)\Big]\Big\|.\end{aligned} \tag{93}$$

By a similar argument as Step 2 in the proof of Lemma 4, we obtain

$$\|\mathbb{E}[H_{k-i}^\top A_k^\top A_kH_{k-i}]\|\lesssim p^2\|\mathbb{X}\|^2<\infty, \tag{94}$$

where the constant in $\lesssim$ is independent of α. Further, recall that A_j are i.i.d. random matrices and $\|\mathbb{E}[A_1^\top A_1]\|\le 1-\alpha p\lambda_{\min}[X^\top(2I_d-\alpha\mathbb{X})X]$ by the proof of Lemma 2. When $\alpha\|\mathbb{X}\|<2$, it follows from the sub-multiplicativity of operator norm and the similar lines as the Step 1 in the proof of Lemma 4 that

$$\begin{aligned}&\sum_{i=0}^{\infty}\Big\|\mathbb{E}\Big[\Big(\prod_{j=k-i+1}^{k-1}A_j\Big)^\top\Big(\prod_{j=k-i+1}^{k-1}A_j\Big)\Big]\Big\|\\&=\sum_{i=0}^{\infty}\Big\|\prod_{j=k-i+1}^{k-1}\mathbb{E}[A_j^\top A_j]\Big\|\\&\le\sum_{i=2}^{\infty}\big\|\mathbb{E}[A_1^\top A_1]\big\|^{i-2}=O(1/\alpha).\end{aligned} \tag{95}$$

Therefore, $\sum_{i=0}^{\infty}\mathbb{E}\|\tilde{\mathcal{M}}_{i,k}(\alpha)\|_2^2 = O(1/\alpha)$, which yields $(\mathbb{E}\|\tilde{\boldsymbol{\beta}}_k^{\circ}(\alpha)-\tilde{\boldsymbol{\beta}}\|_2^2)^{1/2} = O(\sqrt{\alpha})$. By leveraging the inequality in (92) and the similar techniques adopted in the proof of Lemma 2 for the case with $q>2$, we obtain that for any $q\geq 2$, $\sum_{i=0}^{\infty}(\mathbb{E}\|\tilde{\mathcal{M}}_{i,k}(\alpha)\|_2^q)^{2/q} = O(1/\alpha)$. As a direct consequence, we obtain $(\mathbb{E}\|\tilde{\boldsymbol{\beta}}_k^{\circ}(\alpha)-\tilde{\boldsymbol{\beta}}\|_2^q)^{1/q} = O(\sqrt{\alpha})$, which completes the proof. ■

C.3 Proof of Theorem 6

Proof If we can establish the asymptotic normality for the affine sequence $\{\boldsymbol{\beta}_k^{\dagger}(\alpha)-\tilde{\boldsymbol{\beta}}\}_{k\in\mathbb{N}}$, then by applying Lemma 4 and Markov's inequality, we can prove the CLT for the stationary sequence $\{\tilde{\boldsymbol{\beta}}_k^{\circ}(\alpha)-\tilde{\boldsymbol{\beta}}\}_{k\in\mathbb{N}}$. Therefore, in this proof, we shall show the CLT for $\{\boldsymbol{\beta}_k^{\dagger}(\alpha)-\tilde{\boldsymbol{\beta}}\}_{k\in\mathbb{N}}$, that is,

$$\frac{\boldsymbol{\beta}_k^{\dagger}(\alpha)-\tilde{\boldsymbol{\beta}}}{\sqrt{\alpha}} \Rightarrow \mathcal{N}(0,\Xi(\alpha)), \quad \text{as } \alpha\to 0.$$

First, we recall the random vectors $\boldsymbol{b}_k(\alpha)$ in (15) and let

$$\boldsymbol{b}_k(\alpha) =: \alpha H_k\tilde{\boldsymbol{\beta}}, \quad \text{with} \quad H_k := D_k\overline{\mathbb{X}}(pI_d - D_k). \tag{96}$$

Then, since $\{\boldsymbol{\beta}_k^{\dagger}(\alpha)-\tilde{\boldsymbol{\beta}}\}_{k\in\mathbb{N}}$ is a stationary sequence and using induction on k, we can rewrite $\boldsymbol{\beta}_k^{\dagger}(\alpha)-\tilde{\boldsymbol{\beta}}$ into

$$\begin{aligned}\boldsymbol{\beta}_k^{\dagger}(\alpha)-\tilde{\boldsymbol{\beta}} &= \alpha\Big(I_dH_k + (I_d-\alpha p\mathbb{X}_p)H_{k-1} + \cdots + (I_d-\alpha p\mathbb{X}_p)^{k-1}H_1 + \cdots\Big)\tilde{\boldsymbol{\beta}}\\ &= \alpha\sum_{i=0}^{\infty}(I_d-\alpha p\mathbb{X}_p)^iH_{k-i}\tilde{\boldsymbol{\beta}}.\end{aligned} \tag{97}$$

Here, the random matrices H_k are i.i.d. and independent of $\boldsymbol{\beta}_{k-1}^{\dagger}$. For negative index k, H_k is independently drawn from the same distribution as H_1. In the second row of (97), we can write the recursion until $H_{-\infty}$ since in the definition of $\boldsymbol{\beta}_k^{\dagger}(\alpha)-\tilde{\boldsymbol{\beta}}$, we chose the initialization as the stationary random vector $\boldsymbol{\beta}_0^{\dagger} = \tilde{\boldsymbol{\beta}}_0^{\circ} \sim \pi_{\alpha}$. Thus, $\{\boldsymbol{\beta}_k^{\dagger}(\alpha)-\tilde{\boldsymbol{\beta}}\}_{k\in\mathbb{N}}$ is a stationary sequence. This together with the discussion at the end of Section B.1 in Appendix gives Eq. (97).

Again, since H_k are i.i.d. we shall apply the Lindeberg-Feller central limit theorem to the partial sum in (97). To this end, we first take the expectation on both sides of (97). Since the random matrices H_i are independent for all $i\in\mathbb{N}$, we obtain

$$\begin{aligned}\mathbb{E}[\boldsymbol{\beta}_k^{\dagger}(\alpha)-\tilde{\boldsymbol{\beta}}] &= \alpha\sum_{i=0}^{\infty}(I_d-\alpha p\mathbb{X}_p)^i\mathbb{E}[H_{k-i}\tilde{\boldsymbol{\beta}}]\\ &= \alpha\sum_{i=0}^{\infty}(I_d-\alpha p\mathbb{X}_p)^i\mathbb{E}[H_{k-i}]\tilde{\boldsymbol{\beta}} = 0.\end{aligned} \tag{98}$$

To see the last equality, we apply Lemma 21 (i) and (ii) and obtain $\mathbb{E}[D_k\overline{\mathbb{X}}D_k] = p\overline{\mathbb{X}}_p = p^2\overline{\mathbb{X}}$, which gives

$$\mathbb{E}[H_k] = \mathbb{E}[D_k\overline{\mathbb{X}}(pI_d-D_k)] = p^2\overline{\mathbb{X}} - p^2\overline{\mathbb{X}} = 0, \tag{99}$$

As a direct consequence, by (96), we have

$$\mathbb{E}[\boldsymbol{b}_k(\alpha)] = \alpha\mathbb{E}[H_k]\tilde{\boldsymbol{\beta}} = 0. \tag{100}$$

Next, we shall provide a closed form of the covariance matrix $\mathrm{Cov}(\boldsymbol{\beta}_k^\dagger(\alpha) - \tilde{\boldsymbol{\beta}})$. Notice that the random vectors $H_i\tilde{\boldsymbol{\beta}}$ are uncorrelated over different i, and $\mathbb{E}[H_i\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top H_i] = \mathbb{E}[H_1\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top H_1]$ due to the stationarity of the sequence $\{H_i\tilde{\boldsymbol{\beta}}\}_{i\in\mathbb{N}}$. Hence, by (97), we have

$$\begin{aligned}
V_\alpha &:= \mathrm{Cov}\Big(\alpha^{-1/2}(\boldsymbol{\beta}_k^\dagger(\alpha) - \tilde{\boldsymbol{\beta}})\Big) \\
&= \alpha^{-1}\mathbb{E}\big[(\boldsymbol{\beta}_k^\dagger(\alpha) - \tilde{\boldsymbol{\beta}})(\boldsymbol{\beta}_k^\dagger(\alpha) - \tilde{\boldsymbol{\beta}})^\top\big] \\
&= \alpha\sum_{i=0}^{\infty}(I_d - \alpha p\mathbb{X}_p)^i\mathbb{E}[H_{k-i}\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top H_{k-i}](I_d - \alpha p\mathbb{X}_p)^i \\
&=: \alpha\sum_{i=0}^{\infty}(I_d - \alpha p\mathbb{X}_p)^i S(I_d - \alpha p\mathbb{X}_p)^i,
\end{aligned} \tag{101}$$

with $d \times d$ matrix

$$S := \mathbb{E}[H_1\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top H_1] =: \mathbb{E}[H_1 S_0 H_1], \quad \text{where } S_0 = \tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top. \tag{102}$$

Furthermore, by Lemma 21 (i), one can show that $\overline{(\overline{\mathbb{X}})} = \overline{\mathbb{X}}$ and $\mathrm{Diag}(A\overline{\mathbb{X}}) = \mathrm{Diag}(\overline{A\mathbb{X}})$ for any matrix A. Then, by the definition of H_k in (96) and Lemma 21 (ii)–(iv), we can simplify $\mathbb{E}[H_1 S_0 H_1]$ as follows:

$$\begin{aligned}
&\mathbb{E}[H_1 S_0 H_1] \\
&= \mathbb{E}[D_k\overline{\mathbb{X}}(pI_d - D_k)S_0(pI_d - D_k)\overline{\mathbb{X}}D_k] \\
&= p^2\mathbb{E}[D_1\overline{\mathbb{X}}S_0\overline{\mathbb{X}}D_1] - p\mathbb{E}[D_1\overline{\mathbb{X}}D_1S_0\overline{\mathbb{X}}D_1] - p\mathbb{E}[D_1\overline{\mathbb{X}}S_0D_1\overline{\mathbb{X}}D_1] + \mathbb{E}[D_1\overline{\mathbb{X}}D_kS_0D_k\overline{\mathbb{X}}D_1] \\
&= p^3(\overline{\mathbb{X}}S_0\overline{\mathbb{X}})_p - 2p\Big(p\overline{\mathbb{X}}_p(S_0\overline{\mathbb{X}})_p + p^2(1-p)\mathrm{Diag}(\overline{\mathbb{X}}S_0\overline{\mathbb{X}})\Big) \\
&\quad + p\overline{\mathbb{X}}_p(S_0)_p\overline{\mathbb{X}}_p + p^2(1-p)\Big(\mathrm{Diag}(\overline{\mathbb{X}}(S_0)_p\overline{\mathbb{X}}) + 2\overline{\mathbb{X}}_p\mathrm{Diag}(\overline{S_0\mathbb{X}}) + (1-p)\overline{\mathbb{X}}\odot\overline{S_0}^\top\odot\overline{\mathbb{X}}\Big).
\end{aligned} \tag{103}$$

By (103), we obtain a closed form solution of S which is independent of α.

Now we are ready to solve the covariance matrix V_α in (101). We multiply the matrix $I_d - \alpha p\mathbb{X}_p$ to the left and right sides of (101) and obtain

$$(I_d - \alpha p\mathbb{X}_p)V_\alpha(I_d - \alpha p\mathbb{X}_p) = \alpha\sum_{i=0}^{\infty}(I_d - \alpha p\mathbb{X}_p)^i S(I_d - \alpha p\mathbb{X}_p)^i. \tag{104}$$

Taking the difference between V_α and $(I_d - \alpha p\mathbb{X}_p)V_\alpha(I_d - \alpha p\mathbb{X}_p)$ yields

$$V_\alpha - (I_d - \alpha p\mathbb{X}_p)V_\alpha(I_d - \alpha p\mathbb{X}_p) = \alpha S. \tag{105}$$

Denote the symmetric matrix $A_p = p\mathbb{X}_p$. By simplifying the equation above, for $\alpha > 0$, we have

$$V_\alpha A_p - A_pV_\alpha + \alpha A_pV_\alpha A_p = S. \tag{106}$$

Let $V_0 = \lim_{\alpha\to 0} V_\alpha$. As $\alpha \to 0$, the quadratic term $\alpha A_p V_\alpha A_p$ vanishes. Thus, we only need to solve the equation

$$S - V_0 A_p - A_p V_0 = 0, \tag{107}$$

to get the solution for

$$V_0 = \lim_{\alpha\to 0} V_\alpha.$$

Following Theorem 1 in Pflug (1986) and the subsequent Remark therein, we can get the closed form solution of V_0, that is,

$$\mathrm{vec}(V_0) = (I_d \otimes A_p + A_p \otimes I_d)^{-1} \cdot \mathrm{vec}(S), \tag{108}$$

where the $d^2 \times d^2$ matrix $I_d \otimes A_p + A_p \otimes I_d$ is invertible since the fixed design matrix X is assumed to be in a reduced form with no zero columns. For a small $\alpha > 0$, we shall provide a similar closed form solution for $V_\alpha = V_0 + \alpha B_p$. Specifically, we need to get the closed form of the matrix B_p by solving a similar equation:

$$A_p V_0 A_p - B_p A_p - A_p B_p = 0, \tag{109}$$

which gives

$$\mathrm{vec}(B_p) = (I_d \otimes A_p + A_p \otimes I_d)^{-1} \times \mathrm{vec}(A_p V_0 A_p). \tag{110}$$

The deterministic matrices V_0, A_p and B_p are all independent of α. By inserting the results of V_0 and B_p into $V_\alpha = V_0 + \alpha B_p$, we obtain

$$\Xi(\alpha) = V_\alpha = V_0 + \alpha B_p,$$

which holds uniformly over k due to the stationarity of $\{\boldsymbol{\beta}_k^\dagger(\alpha) - \tilde{\boldsymbol{\beta}}\}_{k\in\mathbb{N}}$.

Finally, by applying the Lindeberg-Feller central limit theorem to the partial sum in (97), we establish the asymptotic normality of $\{\boldsymbol{\beta}_k^\dagger(\alpha) - \tilde{\boldsymbol{\beta}}\}_{k\in\mathbb{N}}$ and complete the proof. ■

Appendix D. Proofs in Section 3.3

We first outline the main techniques for establishing the asymptotic normality of the averaged GD dropout sequence $\{\bar{\boldsymbol{\beta}}_k^{\mathrm{gd}}(\alpha)\}_{k\in\mathbb{N}}$ defined in (24).

Recall the observation $\boldsymbol{y}$ in model (1) and the dropout matrix D. For the GD dropout $\{\tilde{\boldsymbol{\beta}}_k(\alpha)\}_{k\in\mathbb{N}}$ in (2), by Theorem 3, we can define a centering term as follows,

$$\tilde{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha) = \lim_{k\to\infty} \mathbb{E}_D[\tilde{\boldsymbol{\beta}}_k(\alpha)] = \mathbb{E}_D[\tilde{\boldsymbol{\beta}}_1^\circ(\alpha)], \tag{111}$$

where $\tilde{\boldsymbol{\beta}}_1^\circ(\alpha)$ follows the stationary distribution π_α as stated in (14). According to Lemma 1 in Clara et al. (2024), we note that $\mathbb{E}_D[\tilde{\boldsymbol{\beta}}_k(\alpha) - \tilde{\boldsymbol{\beta}}] \neq 0$ but $\|\mathbb{E}_D[\tilde{\boldsymbol{\beta}}_k(\alpha) - \tilde{\boldsymbol{\beta}}]\|_2 \to 0$ if $\alpha p \|\mathbb{X}\| < 1$ with a geometric rate as $k \to \infty$. Therefore, we shall first show the central limit theorems for the partial sum of $\{\tilde{\boldsymbol{\beta}}_k(\alpha) - \tilde{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)\}_{k\in\mathbb{N}}$ and then for the one of $\{\tilde{\boldsymbol{\beta}}_k(\alpha) - \tilde{\boldsymbol{\beta}}\}_{k\in\mathbb{N}}$.

Next, we take a closer look at the partial sum of $\{\tilde{\boldsymbol{\beta}}_k(\alpha) - \tilde{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)\}_{k\in\mathbb{N}}$. The iterative function f defined in (9) allows us to write $\tilde{\boldsymbol{\beta}}_k(\alpha) = f_{D_k}(\tilde{\boldsymbol{\beta}}_{k-1}(\alpha))$ for all $k \in \mathbb{N}$. Similarly,

for the initialization $\tilde{\boldsymbol{\beta}}_0^\circ(\alpha)$ that follows the unique stationary distribution π_α in Theorem 3, we can write the stationary GD dropout sequence $\{\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ into

$$\tilde{\boldsymbol{\beta}}_k^\circ(\alpha) = f_{D_k}(\tilde{\boldsymbol{\beta}}_{k-1}^\circ(\alpha)), \quad k \in \mathbb{N}. \tag{112}$$

Recall $\tilde{\boldsymbol{\beta}}_{\text{mean}}(\alpha)$ defined in (111). Then, we can recursively rewrite $\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)$ using the iterative function f and obtain the partial sum

$$\begin{aligned}\tilde{S}_t^\circ(\alpha) &:= \sum_{k=1}^{t}[\tilde{\boldsymbol{\beta}}_k^\circ(\alpha) - \tilde{\boldsymbol{\beta}}_{\text{mean}}(\alpha)] \\ &= \big\{f_{D_1}(\tilde{\boldsymbol{\beta}}_0^\circ(\alpha)) - \mathbb{E}\big[f_{D_1}(\tilde{\boldsymbol{\beta}}_0^\circ(\alpha))\big]\big\} + \big\{f_{D_2}\circ f_{D_1}(\tilde{\boldsymbol{\beta}}_0^\circ(\alpha)) - \mathbb{E}\big[f_{D_2}\circ f_{D_1}(\tilde{\boldsymbol{\beta}}_0^\circ(\alpha))\big]\big\} \\ &\quad + \cdots + \big\{f_{D_t}\circ\cdots\circ f_{D_1}(\tilde{\boldsymbol{\beta}}_0^\circ(\alpha)) - \mathbb{E}\big[f_{D_t}\circ\cdots\circ f_{D_1}(\tilde{\boldsymbol{\beta}}_0^\circ(\alpha))\big]\big\}. \end{aligned} \tag{113}$$

Primarily, we aim to (i) prove the central limit theorem for the partial sum $t^{-1/2}\tilde{S}_t^\circ(\alpha)$, and (ii) prove the invariance principle for the partial sum process $(\tilde{S}_i^\circ(\alpha))_{1\le i\le t}$. To this end, we borrow the idea of *functional dependence measure* in Wu (2005), which was further investigated in Wu (2011) to establish the asymptotic normality for sequences with *short-range dependence* (see (120) for the definition). We shall show that the GD dropout sequence $\{\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ that satisfies the geometric-moment contraction (as proved in Theorem 3) satisfies such short-range dependence condition.

Finally, we shall complete the proofs of the quenched central limit theorems by showing that, for any given constant learning rate $\alpha > 0$ satisfying the conditions in Theorem 7, and any initialization $\tilde{\boldsymbol{\beta}}_0 \in \mathbb{R}^d$, the partial sum

$$\tilde{S}_t^{\tilde{\boldsymbol{\beta}}_0}(\alpha) := \sum_{k=1}^{t}[\tilde{\boldsymbol{\beta}}_k(\alpha) - \tilde{\boldsymbol{\beta}}_{\text{mean}}(\alpha)] \tag{114}$$

converges to the stationary partial sum process $\tilde{S}_t^\circ(\alpha)$, in the sense that $t^{-1/2}\big(\mathbb{E}\|\tilde{S}_t^{\tilde{\boldsymbol{\beta}}_0}(\alpha) - \tilde{S}_t^\circ(\alpha)\|_2^q\big)^{1/q} = o(1)$ as $t\to\infty$.

D.1 Functional Dependence Measure

Before proceeding to the proofs of Theorems 7 and 9, we first provide the detailed form of the functional dependence measure in Wu (2005) for the iterated random functions with i.i.d. random matrices as inputs. This will serve as the foundational pillar to build the asymptotic normality of averaged GD dropout iterates.

First, for any random vector $\boldsymbol{\zeta} \in \mathbb{R}^d$ satisfying $\mathbb{E}\|\boldsymbol{\zeta}\|_2 < \infty$, define projection operators

$$\mathcal{P}_k[\boldsymbol{\zeta}] = \mathbb{E}[\boldsymbol{\zeta} \mid \mathcal{F}_k] - \mathbb{E}[\boldsymbol{\zeta} \mid \mathcal{F}_{k-1}], \quad k \in \mathbb{Z}, \tag{115}$$

where we recall the filtration $\mathcal{F}_i = \sigma(D_i, D_{i-1}, \ldots)$ with i.i.d. dropout matrices D_i, $i \in \mathbb{Z}$. By Theorem 3 and (12), there exists a measurable function $h_\alpha(\cdot)$ such that the stationary GD dropout sequence $\{\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ can be written as the following causal process

$$\tilde{\boldsymbol{\beta}}_k^\circ(\alpha) = h_\alpha(D_k, D_{k-1}, \ldots) = h_\alpha(\mathcal{F}_k). \tag{116}$$

Define a coupled version of filtration $\mathcal{F}_i$ as $\mathcal{F}_{i,\{j\}} = \sigma(D_i, \ldots, D_{j+1}, D'_j, D_{j-1}, \ldots)$. In addition, $\mathcal{F}_{i,\{j\}} = \mathcal{F}_i$ if $j > i$. For $q > 1$, define the *functional dependence measure* of $\tilde{\boldsymbol{\beta}}^\circ_k(\alpha)$ as

$$\theta_{k,q}(\alpha) = \big(\mathbb{E}\|\tilde{\boldsymbol{\beta}}^\circ_k(\alpha) - \tilde{\boldsymbol{\beta}}^\circ_{k,\{0\}}(\alpha)\|_2^q\big)^{1/q}, \quad \text{where } \tilde{\boldsymbol{\beta}}^\circ_{k,\{0\}}(\alpha) = h_\alpha(\mathcal{F}_{k,\{0\}}). \tag{117}$$

The above quantity can be interpreted as the dependence of $\tilde{\boldsymbol{\beta}}^\circ_k(\alpha)$ on D_0 (see the discussion below Theorem 3 for the meaning of $\tilde{\boldsymbol{\beta}}^\circ_k$ with $k \le 0$), and $\tilde{\boldsymbol{\beta}}^\circ_{k,\{0\}}(\alpha)$ is a coupled version of $\tilde{\boldsymbol{\beta}}^\circ_k(\alpha)$ with D_0 in the latter replaced by its i.i.d. copy D'_0. If $\tilde{\boldsymbol{\beta}}^\circ_k(\alpha)$ does not functionally depend on D_0, then $\theta_{k,q}(\alpha) = 0$.

Furthermore, if $\sum_{k=0}^\infty \theta_{k,q}(\alpha) < \infty$, we define the tail of the *cumulative dependence measure* as

$$\Theta_{m,q}(\alpha) = \sum_{k=m}^\infty \theta_{k,q}(\alpha), \quad m \in \mathbb{N}. \tag{118}$$

This can be interpreted as the cumulative dependence of $\{\tilde{\boldsymbol{\beta}}^\circ_k(\alpha)\}_{k\ge m}$ on D_0, or equivalently, the cumulative dependence of $\tilde{\boldsymbol{\beta}}^\circ_0(\alpha)$ on D_j, $j \ge m$. The functional dependence measure in (117) and its cumulative variant in (118) are easy to work with and they can directly reflect the underlying data-generating mechanism of the iterative function $\tilde{\boldsymbol{\beta}}^\circ_k(\alpha) = f_{D_k}(\tilde{\boldsymbol{\beta}}^\circ_{k-1}(\alpha))$.

Specifically, for all $q \ge 2$, Theorem 1 in Wu (2005) pointed out a useful inequality for the functional dependence measure as follows,

$$\sum_{k=0}^\infty \big(\mathbb{E}\|\mathcal{P}_0[\tilde{\boldsymbol{\beta}}^\circ_k(\alpha)]\|_2^q\big)^{1/q} \le \sum_{k=0}^\infty \theta_{k,q}(\alpha) = \Theta_{0,q}(\alpha). \tag{119}$$

In particular, for some given learning rate $\alpha > 0$, we say the sequence $\{\tilde{\boldsymbol{\beta}}^\circ_k(\alpha)\}_{k\in\mathbb{N}}$ satisfies the short-range dependence condition if

$$\Theta_{0,q}(\alpha) < \infty, \quad \text{for some } q \ge 2. \tag{120}$$

This dependence assumption has been widely adopted in the literature; see for example the invariance principle in Wu (2011); Berkes et al. (2014); Karmakar and Wu (2020). If condition (120) fails, then $\tilde{\boldsymbol{\beta}}^\circ_k(\alpha)$ can be long-range dependent, and the partial sum (resp. partial sum processes) behave no longer like Gaussian random vectors (resp. Brownian motions).

Here, we introduce Theorem 3 in Wu (2011) and Theorem 2 in Karmakar and Wu (2020), which are the fundamental tools for the proofs of Theorems 7 and 9, respectively.

Lemma 25 (Asymptotic normality (Wu, 2011)) *Consider a sequence of stationary mean-zero random variables $x_k = g(\epsilon_k, \epsilon_{k-1}, \ldots) \in \mathbb{R}$, for $k = 1, \ldots, n$, where ϵ_k's are i.i.d. random variables, and $g(\cdot)$ is a measurable function such that each x_k is a proper random variable. Recall the projection operator $\mathcal{P}_0[\cdot]$ defined in* (115)*. Let $\theta_q(i) = (\mathbb{E}\|\mathcal{P}_0[x_i]\|_2^q)^{1/q}$, $q > 1$. Assume $\mathbb{E}[x_i] = 0$ and*

$$\Theta_q := \sum_{i=0}^\infty \theta_q(i) < \infty. \tag{121}$$

Let $S_n = \sum_{k=1}^n x_k$ and define the process $S_t = S_{\lfloor t\rfloor} + (t - \lfloor t\rfloor)x_{\lfloor t\rfloor+1}$, for $t \ge 0$ and the floor function $\lfloor t\rfloor = \max\{k \in \mathbb{Z} : k \le t\}$. Then,

(i) we have the moment inequality

$$(\mathbb{E}\|S_n\|_2^q)^{1/q} \leq \begin{cases} (q-1)^{1/2} n^{1/2} \Theta_q, & q > 2, \\ (q-1)^{-1} n^{1/q} \Theta_q, & 1 < q \leq 2. \end{cases} \tag{122}$$

(ii) If moreover (121) holds with $q = 2$, the invariance principle

$$\Big\{ \frac{1}{\sqrt{n}} S_{nu},\, 0 \leq u \leq 1 \Big\} \Rightarrow \{\sigma \mathbb{B}(u),\, 0 \leq u \leq 1\}, \tag{123}$$

holds with long-run variance $\sigma^2 = \mathbb{E}\| \sum_{i=0}^{\infty} \mathcal{P}_0[x_i]\|_2^2$.

Lemma 26 (Gaussian approximation (Karmakar and Wu, 2020)) *Suppose that we have a sequence of nonstationary mean-zero random vectors $\boldsymbol{x}_k = g_k(\epsilon_k, \epsilon_{k-1}, \ldots) \in \mathbb{R}^d$, for $k = 1, \ldots, n$, where the ϵ_k's are i.i.d. random variables, and $g_k(\cdot)$ is a measurable function such that each $\boldsymbol{x}_k$ is a proper random vector. Let $S_j = \sum_{k=1}^{j} \boldsymbol{x}_k$. Assume the following conditions hold for some $q > 2$:*

(i) The series $(\|\boldsymbol{x}_k\|_2^q)_{k \geq 1}$ is uniformly integrable: $\sup_{k \geq 1} \mathbb{E}\big[\|\boldsymbol{x}_k\|_2^q \mathbf{1}_{\|\boldsymbol{x}_k\|_2 \geq u}\big] \to 0$ as $u \to \infty$,

(ii) The eigenvalues of covariance matrices of increment processes are lower-bounded, that is, there exists $\lambda_* > 0$ and $l_* \in \mathbb{N}$, such that for all $t \geq 1$, $l \geq l_*$,

$$\lambda_{\min}\big(\mathrm{Cov}(S_{t+l} - S_t)\big) \geq \lambda_* l;$$

(iii) There exist constants $\chi > \chi_0$ and $\kappa > 0$, where

$$\chi_0 = \frac{q^2 - 4 + (q-2)\sqrt{q^2 + 20q + 4}}{8q},$$

such that the tail cumulative dependence measure

$$\Theta_{m,q}(\alpha) = \sum_{k=m}^{\infty} \theta_{k,q}(\alpha) = O\big\{ m^{-\chi} \big(\log(m) \big)^{-\kappa} \big\}. \tag{124}$$

Then, for all $q > 2$, there exists a probability space $(\Omega^\star, \mathcal{A}^\star, \mathbb{P}^\star)$ on which we can define random vectors $\boldsymbol{x}_k^\star$, with the partial sum process $S_i^\star = \sum_{k=1}^{i} \boldsymbol{x}_k^\star$ and a Gaussian process $G_i^\star = \sum_{k=1}^{i} \boldsymbol{z}_k^\star$. Here $\boldsymbol{z}_k^\star$ is a mean-zero independent Gaussian vector, such that $(S_i^\star)_{1 \leq i \leq n} \overset{\mathcal{D}}{=} (S_i)_{1 \leq i \leq n}$ and

$$\max_{i \leq n} |S_i^\star - G_i^\star| = o_{\mathbb{P}}(n^{1/q}) \quad \text{in } (\Omega^\star, \mathcal{A}^\star, \mathbb{P}^\star).$$

We notice that condition (ii) in Lemma 26 on the non-singularity is required when the sequence $\{\boldsymbol{x}_k\}_{k \in \mathbb{N}}$ is non-stationary. However, if the function $g_k(\cdot) \equiv g(\cdot)$, that is, the sequence $\{\boldsymbol{x}_k\}_{k \in \mathbb{N}}$ is stationary, then the covariance matrix of the increments is allowed to be singular. To see this, consider a stationary partial sum $S_l = (S_{l,1}, \ldots, S_{l,d})^\top$ with a singular covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ and assume $\mathrm{rank}(\Sigma) = d - 1$. Then, there exists

a unit vector $\boldsymbol{v} \in \mathbb{R}^d$ such that $\Sigma \boldsymbol{v} = 0$, which indicates that $S_{l,1}$ can be written into a linear combination of $S_{l,2}, \ldots, S_{l,d}$, and the covariance matrix of this linear combination is non-singular. Hence, condition (ii) in Lemma 26 is not required for stationary processes.

In addition, the original Theorem 3 in Wu (2011) and Theorem 2 in Karmakar and Wu (2020) considered a simple case where the i.i.d. inputs ϵ_k are one-dimensional. These two theorems still hold even if the inputs are i.i.d. random matrices such as the dropout matrices D_k in our case. In fact, as long as the inputs are i.i.d. elements, the functional dependence measure can be similarly computed as the one in one-dimensional case. The essence is that the short-range dependence condition (120) is satisfied using an appropriate norm (e.g., L^2-norm for vectors, operator norm for matrices) by the output x_k. For example, Wu and Shao (2004) considered iterated random functions on a general metric space, and Chen and Wu (2016) assumed the ϵ_i's to be i.i.d. random elements to derive asymptotics for x_k. We will verify this short-range dependence condition on the GD dropout vector estimates $\{\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ in the proof of Theorem 7.

D.2 Proof of Theorem 7

Proof We verify the short-range dependence condition for the stationary GD dropout sequence $\{\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$.

First, consider two different initial vectors $\tilde{\boldsymbol{\beta}}_0^\circ$, $\tilde{\boldsymbol{\beta}}_0^{\circ'} \in \mathbb{R}^d$ following the unique stationary distribution π_α in Theorem 3. Denote the two GD dropout sequences by $\{\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ and $\{\tilde{\boldsymbol{\beta}}_k^{\circ'}(\alpha)\}_{k\in\mathbb{N}}$ accordingly. By the geometric-moment contraction in Theorem 3, for all $q \geq 2$, we have

$$\sup_{\tilde{\boldsymbol{\beta}}_0^\circ, \tilde{\boldsymbol{\beta}}_0^{\circ'} \in \mathbb{R}^d, \tilde{\boldsymbol{\beta}}_0^\circ \neq \tilde{\boldsymbol{\beta}}_0^{\circ'}} \frac{\big(\mathbb{E}\|\tilde{\boldsymbol{\beta}}_k^\circ(\alpha) - \tilde{\boldsymbol{\beta}}_k^{\circ'}(\alpha)\|_2^q\big)^{1/q}}{\|\tilde{\boldsymbol{\beta}}_0^\circ - \tilde{\boldsymbol{\beta}}_0^{\circ'}\|_2} \leq r_{\alpha,q}^k, \quad k \in \mathbb{N}, \tag{125}$$

for some constant $r_{\alpha,q} \in (0,1)$. Equivalently, it can be rewritten in terms of the iterative function f defined in (9) and $h_\alpha(\cdot)$ defined in (12). That is, for all $\tilde{\boldsymbol{\beta}}_0^\circ$, $\tilde{\boldsymbol{\beta}}_0^{\circ'} \in \mathbb{R}^d$, such that $\tilde{\boldsymbol{\beta}}_0^\circ \neq \tilde{\boldsymbol{\beta}}_0^{\circ'}$, we have

$$\begin{aligned}
&\big(\mathbb{E}\|f_{D_k} \circ \cdots \circ f_{D_1}(\tilde{\boldsymbol{\beta}}_0^\circ) - f_{D_k} \circ \cdots \circ f_{D_1}(\tilde{\boldsymbol{\beta}}_0^{\circ'})\|_2^q\big)^{1/q} \\
&= \big(\mathbb{E}\|h_\alpha(D_k, \ldots, D_1, D_0, D_{-1}, \ldots) - h_\alpha(D_k, \ldots, D_1, D_0', D_{-1}', \ldots)\|_2^q\big)^{1/q} \\
&= \big(\mathbb{E}\|h_\alpha(\mathcal{F}_k) - h_\alpha(\mathcal{F}_{k,\{0,-1,\ldots\}})\|_2^q\big)^{1/q} \\
&\leq c_q r_{\alpha,q}^k,
\end{aligned} \tag{126}$$

where we recall the filtration $\mathcal{F}_{i,\{j\}} = \sigma(D_i, \ldots, D_{j+1}, D_j', D_{j-1}, \ldots)$, and $c_q > 0$ is some constant independent of k. Moreover, since $h_\alpha(\mathcal{F}_k)$ is stationary over k and D_i and D_j' are i.i.d. random matrices, for all $i, j \in \mathbb{Z}$, it follows that

$$\begin{aligned}
&\mathbb{E}\|h_\alpha(\mathcal{F}_{k,\{0\}}) - h_\alpha(\mathcal{F}_{k,\{0,-1,\ldots\}})\|_2^q \\
&= \mathbb{E}\|h_\alpha(\mathcal{F}_k) - h_\alpha(\mathcal{F}_{k,\{-1,-2\ldots\}})\|_2^q \\
&= \mathbb{E}\|h_\alpha(\mathcal{F}_{k+1}) - h_\alpha(\mathcal{F}_{k+1,\{0,-1,\ldots\}})\|_2^q \leq c_q' r_{\alpha,q}^k,
\end{aligned} \tag{127}$$

where the constant $c_q' > 0$ is also independent of k. Hence, by (126) and (127), we can bound the functional dependence measure defined in (117) as follows

$$\begin{aligned}\theta_{k,q}(\alpha) &= \big(\mathbb{E}\|h_\alpha(\mathcal{F}_k) - h_\alpha(\mathcal{F}_{k,\{0\}})\|_2^q\big)^{1/q} \\ &\le \big(\mathbb{E}\|h_\alpha(\mathcal{F}_k) - h_\alpha(\mathcal{F}_{k,\{0,-1,\dots\}})\|_2^q\big)^{1/q} + \big(\mathbb{E}\|h_\alpha(\mathcal{F}_{k,\{0,-1,\dots\}}) - h_\alpha(\mathcal{F}_{k,\{0\}})\|_2^q\big)^{1/q} \\ &\le (c_q + c_q')r_{\alpha,q}^k. \end{aligned} \tag{128}$$

As a direct result, we have finite cumulative dependence measure defined in (118), i.e.,

$$\Theta_{m,q}(\alpha) = \sum_{k=m}^{\infty} \theta_{k,q}(\alpha) = O(r_{\alpha,q}^m) < \infty. \tag{129}$$

Therefore, for the constant learning rate $\alpha > 0$ satisfying the assumptions in Theorem 3, the stationary GD dropout sequence $\{\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ meets the short-range dependence requirement in (120). Consequently, the condition (121) in Lemma 25 is satisfied, which along with the Cramér-Wold device yields the central limit theorem for $\tilde{S}_t^\circ(\alpha)$ defined in (113), that is,

$$t^{-1/2}\tilde{S}_t^\circ(\alpha) \Rightarrow \mathcal{N}(0, \Sigma(\alpha)), \tag{130}$$

where the long-run covariance matrix $\Sigma(\alpha)$ is defined in Theorem 7.

Next, we bound the difference between $\tilde{S}_t^\circ(\alpha)$ and $\tilde{S}_t^{\tilde{\boldsymbol{\beta}}_0}(\alpha)$ for any arbitrarily fixed $\tilde{\boldsymbol{\beta}}_0 \in \mathbb{R}^d$ in the q-th moment, for all $q \ge 2$. For the constant learning rate $\alpha > 0$ satisfying $\alpha\|\mathbb{X}\| < 2$, applying Theorem 3 yields

$$\begin{aligned}&\Big(\mathbb{E}\big\|\tilde{S}_t^\circ(\alpha) - \tilde{S}_t^{\tilde{\boldsymbol{\beta}}_0}(\alpha)\big\|_2^q\Big)^{1/q} \\ &= \Big(\mathbb{E}\big\|\big[f_{D_1}(\tilde{\boldsymbol{\beta}}_0^\circ(\alpha)) + f_{D_2}\circ f_{D_1}(\tilde{\boldsymbol{\beta}}_0^\circ(\alpha)) + \cdots + f_{D_t}\circ\cdots\circ f_{D_1}(\tilde{\boldsymbol{\beta}}_0^\circ(\alpha))\big] \\ &\quad - \big[f_{D_1}(\tilde{\boldsymbol{\beta}}_0(\alpha)) + f_{D_2}\circ f_{D_1}(\tilde{\boldsymbol{\beta}}_0(\alpha)) + \cdots + f_{D_t}\circ\cdots\circ f_{D_1}(\tilde{\boldsymbol{\beta}}_0(\alpha))\big]\big\|_2^q\Big)^{1/q} \\ &\le \Big(\sum_{k=1}^t r_{\alpha,q}^k\Big)\|\tilde{\boldsymbol{\beta}}_0^\circ - \tilde{\boldsymbol{\beta}}_0\|_2. \end{aligned} \tag{131}$$

Since the contraction constant $r_{\alpha,q} \in (0,1)$, we can derive the limit for the sum of the geometric series $\{r_{\alpha,q}^k\}_{k=1}^t$ as follows

$$\lim_{t\to\infty}\sum_{k=1}^t r_{\alpha,q}^k = \lim_{t\to\infty}\frac{r_{\alpha,q}\big(1 - r_{\alpha,q}^t\big)}{1 - r_{\alpha,q}} = \frac{r_{\alpha,q}}{1 - r_{\alpha,q}}. \tag{132}$$

This, together with (131) gives

$$\big(\mathbb{E}\big\|\tilde{S}_t^\circ(\alpha) - \tilde{S}_t^{\tilde{\boldsymbol{\beta}}_0}(\alpha)\big\|_2^q\big)^{1/q} = O(1) = o(\sqrt{t}), \tag{133}$$

which yields the quenched central limit theorem for the partial sum $\tilde{S}_t^{\tilde{\boldsymbol{\beta}}_0}(\alpha)$ defined in (114), that is, for any fixed initial point $\tilde{\boldsymbol{\beta}}_0 \in \mathbb{R}^d$,

$$t^{-1/2}\tilde{S}_t^{\tilde{\boldsymbol{\beta}}_0}(\alpha) \Rightarrow \mathcal{N}(0, \Sigma(\alpha)). \tag{134}$$

Finally, we shall show that $\|\sum_{k=1}^{t}\mathbb{E}[\tilde{\boldsymbol{\beta}}_k(\alpha)-\tilde{\boldsymbol{\beta}}]\|_2 = o(\sqrt{t})$. To see this, we note that given two independently chosen initial vectors $\tilde{\boldsymbol{\beta}}_0$ and $\tilde{\boldsymbol{\beta}}_0^\circ$, where $\tilde{\boldsymbol{\beta}}_0^\circ$ follows the stationary distribution π_α while $\tilde{\boldsymbol{\beta}}_0$ is an arbitrary initial point in $\mathbb{R}^d$, it follows from the triangle inequality that

$$\begin{aligned}\Big\|\sum_{k=1}^{t}\mathbb{E}[\tilde{\boldsymbol{\beta}}_k(\alpha)-\tilde{\boldsymbol{\beta}}]\Big\|_2 &= \Big\|\mathbb{E}\Big[\sum_{k=1}^{t}\big(\tilde{\boldsymbol{\beta}}_k(\alpha)-\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)+\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)-\tilde{\boldsymbol{\beta}}\big)\Big]\Big\|_2 \\ &\le \Big\|\mathbb{E}\Big[\sum_{k=1}^{t}\big(\tilde{\boldsymbol{\beta}}_k(\alpha)-\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)\big)\Big]\Big\|_2 + \Big\|\mathbb{E}\Big[\sum_{k=1}^{t}\big(\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)-\tilde{\boldsymbol{\beta}}\big)\Big]\Big\|_2 \\ &=: \mathrm{I}_1+\mathrm{I}_2. \end{aligned} \tag{135}$$

We first show $\mathrm{I}_2 = 0$. Recall the representation of $\{\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)-\tilde{\boldsymbol{\beta}}\}_{k\in\mathbb{N}}$ in (14). Since $\mathbb{E}[A_k(\alpha)] = \mathbb{E}[I_d-\alpha D_k\mathbb{X}D_k] = I_d-\alpha p\mathbb{X}_p$ and $\mathbb{E}[\boldsymbol{b}_k(\alpha)]=0$ by (100), it follows that

$$\mathbb{E}[\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)-\tilde{\boldsymbol{\beta}}] = (I_d-\alpha p\mathbb{X}_p)\mathbb{E}[\tilde{\boldsymbol{\beta}}_{k-1}^\circ(\alpha)-\tilde{\boldsymbol{\beta}}]. \tag{136}$$

Thus, due to the stationarity of $\{\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ and the non-singularity of $\mathbb{X}_p$, we obtain that uniformly over $k\in\mathbb{N}$,

$$\mathbb{E}[\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)-\tilde{\boldsymbol{\beta}}] = 0. \tag{137}$$

As a direct consequence,

$$\mathrm{I}_2 = \Big\|\mathbb{E}\Big[\sum_{k=1}^{t}\big(\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)-\tilde{\boldsymbol{\beta}}\big)\Big]\Big\|_2 = \Big\|\sum_{k=1}^{t}\mathbb{E}[\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)-\tilde{\boldsymbol{\beta}}]\Big\|_2 = 0. \tag{138}$$

In addition, for the part I_1, it follows from Jensen's inequality and (131) that

$$\begin{aligned}\mathrm{I}_1 &= \Big\|\mathbb{E}\Big[\sum_{k=1}^{t}\big(\tilde{\boldsymbol{\beta}}_k(\alpha)-\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)\big)\Big]\Big\|_2 \\ &\le \Big(\mathbb{E}\Big\|\sum_{k=1}^{t}\big(\tilde{\boldsymbol{\beta}}_k(\alpha)-\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)\big)\Big\|_2^2\Big)^{1/2} \\ &\le \Big(\sum_{k=1}^{t} r_{\alpha,2}^k\Big)\|\tilde{\boldsymbol{\beta}}_0-\tilde{\boldsymbol{\beta}}_0^\circ\|_2. \end{aligned} \tag{139}$$

By inserting the results of parts I_1 and I_2 back to (135), we obtain

$$\Big\|\sum_{k=1}^{t}\mathbb{E}[\tilde{\boldsymbol{\beta}}_k(\alpha)-\tilde{\boldsymbol{\beta}}]\Big\|_2 \le \Big(\sum_{k=1}^{t} r_{\alpha,2}^k\Big)\|\tilde{\boldsymbol{\beta}}_0-\tilde{\boldsymbol{\beta}}_0^\circ\|_2. \tag{140}$$

which remains bounded as $t\to\infty$ when $\alpha\|\mathbb{X}\|<2$ by Theorem 3. This completes the proof. ■

D.3 Proof of Corollary 8

Proof Recall the stationary GD dropout sequence $\{\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ which follows the unique stationary distribution π_α. Since this sequence satisfies the short-range dependence condition as stated in (120), it follows from Lemma 25 and the Cramér-Wold device that any fixed linear combination of the coordinates of $\tilde{S}_t^\circ(\alpha)$ in (113) converges to the corresponding linear combination of normal vectors in distribution. Then, the CLT for the averaged GD dropout with multiple learning rates holds by applying the Cramér-Wold device again, that is,

$$t^{-1/2}\mathrm{vec}\big(\tilde{S}_t^\circ(\alpha_1),\ldots,\tilde{S}_t^\circ(\alpha_s)\big)\Rightarrow\mathcal{N}(0,\Sigma^{\mathrm{vec}}). \tag{141}$$

Then, following the similar arguments as in the proof of Theorem 7, we obtain the quenched CLT for $\mathrm{vec}\big(\bar{\boldsymbol{\beta}}_k^{\mathrm{gd}}(\alpha_1)-\tilde{\boldsymbol{\beta}},\ldots,\bar{\boldsymbol{\beta}}_k^{\mathrm{gd}}(\alpha_s)-\tilde{\boldsymbol{\beta}}\big)$. We omit the details here. ■

D.4 Proof of Theorem 9

Proof Recall the stationary GD dropout sequence $\{\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ in (14), where $\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)$ follows the stationary distribution π_α for all $k\in\mathbb{N}$. Also, recall the centering term $\tilde{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)=\mathbb{E}[\tilde{\boldsymbol{\beta}}_1^\circ(\alpha)]$ as defined in (111). By (137), we have $\mathbb{E}[\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)-\tilde{\boldsymbol{\beta}}]=0$ uniformly over $k\in\mathbb{N}$. Hence,

$$\big(\mathbb{E}\|\tilde{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)-\tilde{\boldsymbol{\beta}}\|_2^q\big)^{1/q}=\big(\mathbb{E}\|\mathbb{E}[\tilde{\boldsymbol{\beta}}_1^\circ(\alpha)-\tilde{\boldsymbol{\beta}}]\|_2^q\big)^{1/q}=0. \tag{142}$$

This, along with Lemma 5 gives, for $q>2$,

$$\big(\mathbb{E}\|\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)-\tilde{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)\|_2^q\big)^{1/q}\le\big(\mathbb{E}\|\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)-\tilde{\boldsymbol{\beta}}\|_2^q\big)^{1/q}+\big(\mathbb{E}\|\tilde{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)-\tilde{\boldsymbol{\beta}}\|_2^q\big)^{1/q}=O(\sqrt{\alpha}). \tag{143}$$

Moreover, we notice that by Markov's inequality and (143), for any $u\in\mathbb{R}$ and $\delta>0$, we have

$$\begin{aligned}
&\sup_{k\ge1}\mathbb{E}\big[\|\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)-\tilde{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)\|_2^q\mathbf{1}_{\|\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)-\tilde{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)\|_2\ge u}\big]\\
&\le\sup_{k\ge1}\mathbb{E}\Big[\|\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)-\tilde{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)\|_2^q\cdot\|\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)-\tilde{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)\|_2^\delta/u^\delta\Big]\\
&=\sup_{k\ge1}\mathbb{E}\big[\|\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)-\tilde{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)\|_2^{q+\delta}\big]/u^\delta\\
&=O\big\{\alpha^{(q+\delta)/2}/u^\delta\big\},
\end{aligned} \tag{144}$$

which converges to 0 as $u\to\infty$. Therefore, condition (i) in Lemma 26 is satisfied. Since $\{\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ is stationary, following the arguments below Lemma 26, condition (ii) is not required. Regarding condition (iii), for the constant learning rate $\alpha>0$ satisfying $\alpha\|\mathbb{X}\|<2$, it follows from (128) that the functional dependence measure $\theta_{k,q}(\alpha)\le c\cdot r_{\alpha,q}^k$, for all $q>2$ and $k\in\mathbb{N}$, where the constant $c>0$ is independent of k. Consequently, there exists a

constant $\kappa > 0$ such that the tail cumulative dependence measure of $\{\tilde{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ can be bounded by

$$\Theta_{m,q}(\alpha) = \sum_{k=m}^{\infty} \theta_{k,q}(\alpha) = O\big\{m^{-\chi}\big(\log(m)\big)^{-\kappa}\big\}, \tag{145}$$

where $\chi > 0$ is some constant that can be taken to be arbitrarily large. Then, the condition (iii) in Lemma 26 is satisfied.

Thus, we can obtain the invariance principle for the stationary partial sum process $(\tilde{S}_i^\circ(\alpha))_{1\le i\le t}$ defined in (113). That is, there exists a (richer) probability space $(\tilde{\Omega}^\star, \tilde{\mathcal{A}}^\star, \tilde{\mathbb{P}}^\star)$ on which we can define random vectors $\tilde{\boldsymbol{\beta}}_k^\star$'s with the partial sum process $\tilde{S}_i^\star = \sum_{k=1}^i(\tilde{\boldsymbol{\beta}}_k^\star - \tilde{\boldsymbol{\beta}}_{\text{mean}})$, and a Gaussian process $\tilde{G}_i^\star = \sum_{k=1}^i \tilde{\boldsymbol{z}}_k^\star$, where $\tilde{\boldsymbol{z}}_k^\star$'s are independent Gaussian random vectors in $\mathbb{R}^d$ following $\mathcal{N}(0, I_d)$, such that

$$(\tilde{S}_i^\star)_{1\le i\le t} \overset{\mathcal{D}}{=} (\tilde{S}_i^\circ)_{1\le i\le t}, \tag{146}$$

and

$$\max_{1\le i\le t} \big\|\tilde{S}_i^\star - \Sigma^{1/2}(\alpha)\tilde{G}_i^\star\big\|_2 = o_{\mathbb{P}}(t^{1/q}), \quad \text{in } (\tilde{\Omega}^\star, \tilde{\mathcal{A}}^\star, \tilde{\mathbb{P}}^\star), \tag{147}$$

where the long-run covariance matrix $\Sigma(\alpha)$ is defined in Theorem 7.

Next, recall the partial sum $\tilde{S}_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha) = \sum_{k=1}^i[\tilde{\boldsymbol{\beta}}_k(\alpha) - \tilde{\boldsymbol{\beta}}_{\text{mean}}(\alpha)]$ as defined in (114), given an arbitrarily fixed initial point $\tilde{\boldsymbol{\beta}}_0 \in \mathbb{R}^d$. It follows from the triangle inequality that

$$\begin{aligned}
&\Big(\mathbb{E}\Big[\max_{1\le i\le t}\big\|\tilde{S}_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha) - \Sigma^{1/2}(\alpha)\tilde{G}_i^\star\big\|_2\Big]^q\Big)^{1/q} \\
&= \Big(\mathbb{E}\Big[\max_{1\le i\le t}\big\|\tilde{S}_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha) - \tilde{S}_i^\circ(\alpha) + \tilde{S}_i^\circ(\alpha) - \Sigma^{1/2}(\alpha)\tilde{G}_i^\star\big\|_2\Big]^q\Big)^{1/q} \\
&\le \Big(\mathbb{E}\Big[\max_{1\le i\le t}\big\|\tilde{S}_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha) - \tilde{S}_i^\circ(\alpha)\big\|_2 + \max_{1\le i\le t}\big\|\tilde{S}_i^\circ(\alpha) - \Sigma^{1/2}(\alpha)\tilde{G}_i^\star\big\|_2\Big]^q\Big)^{1/q} \\
&\le \Big(\mathbb{E}\Big[\max_{1\le i\le t}\big\|\tilde{S}_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha) - \tilde{S}_i^\circ(\alpha)\big\|_2\Big]^q\Big)^{1/q} + \Big(\mathbb{E}\Big[\max_{1\le i\le t}\big\|\tilde{S}_i^\circ(\alpha) - \Sigma^{1/2}(\alpha)\tilde{G}_i^\star\big\|_2\Big]^q\Big)^{1/q}.
\end{aligned} \tag{148}$$

Therefore, to show the invariance principle for $\big(\tilde{S}_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha)\big)_{1\le i\le t}$, it suffices to bound the difference part $\max_{1\le i\le t}\|\tilde{S}_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha) - \tilde{S}_i^\circ(\alpha)\|_2$ in terms of the q-th moment. To this end, recall the iterative function $f_D(\boldsymbol{\beta}) = \boldsymbol{\beta} + \alpha DX^\top(\boldsymbol{y} - XD\boldsymbol{\beta})$ in (9) that rewrites the GD dropout recursion (2). We note that

$$\begin{aligned}
&\max_{1\le i\le t}\big\|\tilde{S}_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha) - \tilde{S}_i^\circ(\alpha)\big\|_2 \\
&= \max_{1\le i\le t}\Big\|\big[f_{D_1}(\tilde{\boldsymbol{\beta}}_0) - f_{D_1}(\tilde{\boldsymbol{\beta}}_0^\circ)\big] + \cdots + \big[f_{D_i}\circ\cdots\circ f_{D_1}(\tilde{\boldsymbol{\beta}}_0) - f_{D_i}\circ\cdots\circ f_{D_1}(\tilde{\boldsymbol{\beta}}_0^\circ)\big]\Big\|_2 \\
&\le \max_{1\le i\le t}\Big(\big\|f_{D_1}(\tilde{\boldsymbol{\beta}}_0) - f_{D_1}(\tilde{\boldsymbol{\beta}}_0^\circ)\big\|_2 + \cdots + \big\|f_{D_i}\circ\cdots\circ f_{D_1}(\tilde{\boldsymbol{\beta}}_0) - f_{D_i}\circ\cdots\circ f_{D_1}(\tilde{\boldsymbol{\beta}}_0^\circ)\big\|_2\Big) \\
&= \big\|f_{D_1}(\tilde{\boldsymbol{\beta}}_0) - f_{D_1}(\tilde{\boldsymbol{\beta}}_0^\circ)\big\|_2 + \cdots + \big\|f_{D_t}\circ\cdots\circ f_{D_1}(\tilde{\boldsymbol{\beta}}_0) - f_{D_t}\circ\cdots\circ f_{D_1}(\tilde{\boldsymbol{\beta}}_0^\circ)\big\|_2.
\end{aligned} \tag{149}$$

This, along with the triangle inequality and Theorem 3 yields

$$\begin{aligned}
&\Big(\mathbb{E}\Big[\max_{1\le i\le t}\big\|\tilde{S}_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha)-\tilde{S}_i^{\circ}(\alpha)\big\|_2\Big]^q\Big)^{1/q}\\
&\le \big(\mathbb{E}\big\|f_{D_1}(\tilde{\boldsymbol{\beta}}_0)-f_{D_1}(\tilde{\boldsymbol{\beta}}_0^{\circ})\big\|_2^q\big)^{1/q}+\cdots+\big(\mathbb{E}\big\|f_{D_t}\circ\cdots\circ f_{D_1}(\tilde{\boldsymbol{\beta}}_0)-f_{D_t}\circ\cdots\circ f_{D_1}(\tilde{\boldsymbol{\beta}}_0^{\circ})\big\|_2^q\big)^{1/q}\\
&\le \frac{r_{\alpha,q}(1-r_{\alpha,q}^t)}{1-r_{\alpha,q}}\|\tilde{\boldsymbol{\beta}}_0-\tilde{\boldsymbol{\beta}}_0^{\circ}\|_2=o(t^{1/q}). \qquad (150)
\end{aligned}$$

We insert this result back into (149), which together with (148) gives the invariance principle for the partial sum $\tilde{S}_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha)=\sum_{k=1}^i[\tilde{\boldsymbol{\beta}}_k(\alpha)-\tilde{\boldsymbol{\beta}}_{\text{mean}}(\alpha)]$.

Finally, let the partial sum $S_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha)=\sum_{k=1}^i[\tilde{\boldsymbol{\beta}}_k(\alpha)-\tilde{\boldsymbol{\beta}}]$ be as defined in Theorem 9. We shall bound the difference between $S_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha)$ and $\tilde{S}_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha)$. Since $\tilde{\boldsymbol{\beta}}_{\text{mean}}(\alpha)=\mathbb{E}_D[\tilde{\boldsymbol{\beta}}_1^{\circ}(\alpha)]$ as defined in (111) and $\tilde{\boldsymbol{\beta}}_{\text{mean}}(\alpha)-\tilde{\boldsymbol{\beta}}=\mathbb{E}_D[\tilde{\boldsymbol{\beta}}_1^{\circ}(\alpha)-\tilde{\boldsymbol{\beta}}]=0$ by (137), it follows that

$$\Big(\mathbb{E}\Big[\max_{1\le i\le t}\big\|\tilde{S}_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha)-S_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha)\big\|_2\Big]^q\Big)^{1/q}=\Big(\mathbb{E}\Big[\max_{1\le i\le t}\big\|\sum_{k=1}^i\tilde{\boldsymbol{\beta}}_{\text{mean}}(\alpha)-\tilde{\boldsymbol{\beta}}\big\|_2\Big]^q\Big)^{1/q}=0. \qquad (151)$$

Combining this with the invariance principle for $(\tilde{S}_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha))_{1\le i\le t}$, we obtain the same approximation rate $o_{\mathbb{P}}(t^{1/q})$ for the partial sum process $(S_i^{\tilde{\boldsymbol{\beta}}_0}(\alpha))_{1\le i\le t}$. This completes the proof. ■

Appendix E. Proofs in Section 4.2

Recall the SGD dropout sequence $\{\breve{\boldsymbol{\beta}}_k(\alpha)\}_{k\in\mathbb{N}}$ and the random coefficient $\breve{A}_k(\alpha)$ in (36). To prove that $\sup_{\boldsymbol{v}\in\mathbb{R}^d,\|\boldsymbol{v}\|_2=1}\mathbb{E}\|\breve{A}_k(\alpha)\boldsymbol{v}\|_2^q<1$ is a sufficient condition for the geometric-moment contraction (GMC) of the SGD dropout sequence, we first introduce two useful moment inequalities in Lemma 27.

E.1 Proof of Lemma 10

Lemma 27 (Moment inequality) *Let $q\ge 2$. For any two random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^d$ with fixed $d\ge 1$, the following inequalities holds:*
(i) $\mathbb{E}\big|\|\boldsymbol{x}+\boldsymbol{y}\|_2^q-\|\boldsymbol{x}\|_2^q-q\|\boldsymbol{x}\|_2^{q-2}\boldsymbol{x}^\top\boldsymbol{y}\big|\le\mathbb{E}\big(\|\boldsymbol{x}\|_2+\|\boldsymbol{y}\|_2\big)^q-\mathbb{E}\|\boldsymbol{x}\|_2^q-q\mathbb{E}(\|\boldsymbol{x}\|_2^{q-1}\|\boldsymbol{y}\|_2)$.
(ii) $\mathbb{E}\big|\|\boldsymbol{x}+\boldsymbol{y}\|_2^q-\|\boldsymbol{x}\|_2^q-q\|\boldsymbol{x}\|_2^{q-2}\boldsymbol{x}^\top\boldsymbol{y}\big|\le\big[(\mathbb{E}\|\boldsymbol{x}\|_2^q)^{1/q}+(\mathbb{E}\|\boldsymbol{y}\|_2^q)^{1/q}\big]^q$
$-\mathbb{E}\|\boldsymbol{x}\|_2^q-q(\mathbb{E}\|\boldsymbol{x}\|_2^q)^{(q-1)/q}(\mathbb{E}\|\boldsymbol{y}\|_2^q)^{1/q}$.

Lemma 27(i) immediately follows if we can prove the inequality

$$\Big|\|\boldsymbol{x}+\boldsymbol{y}\|_2^q-\|\boldsymbol{x}\|_2^q-q\|\boldsymbol{x}\|_2^{q-2}\boldsymbol{x}^\top\boldsymbol{y}\Big|\le\big(\|\boldsymbol{x}\|_2+\|\boldsymbol{y}\|_2\big)^q-\|\boldsymbol{x}\|_2^q-q(\|\boldsymbol{x}\|_2^{q-1}\|\boldsymbol{y}\|_2), \qquad (152)$$

which is of independent interest. The right hand side of the inequality in Lemma 27(ii) only depends on expectations of either one of the random vectors $\boldsymbol{x}$ or $\boldsymbol{y}$. This makes the inequality particularly useful if $\boldsymbol{x}$ and $\boldsymbol{y}$ are dependent. Lemma 27(i) is more favorable in cases where $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent, or if one vector is deterministic.

Proof [Lemma 27(i)] We can assume that $\|\boldsymbol{x}\|_2 > 0$ and $\|\boldsymbol{y}\|_2 > 0$ as otherwise, the inequality holds trivially. It is moreover sufficient to assume $\boldsymbol{x} = w\boldsymbol{e}_1$, for a positive number w and $\boldsymbol{e}_1 \in \mathbb{R}^d$ a unit vector. Then, we can find two numbers u, v such that

$$\boldsymbol{y} = u \cdot (w\boldsymbol{e}_1) + v\boldsymbol{e}_2, \tag{153}$$

where $\boldsymbol{e}_2 \in \mathbb{R}^d$ is a unit vector orthogonal to $\boldsymbol{e}_1$. Let $r = \|\boldsymbol{y}\|_2 = \sqrt{(uw)^2 + v^2} > 0$. We note that $\boldsymbol{x}^\top \boldsymbol{y} = uw^2$ and $\|\boldsymbol{x} + \boldsymbol{y}\|_2^2 = (1+u)^2 w^2 + v^2$, which gives

$$\begin{aligned}
&\|\boldsymbol{x} + \boldsymbol{y}\|_2^q - \|\boldsymbol{x}\|_2^q - q\|\boldsymbol{x}\|_2^{q-2}\boldsymbol{x}^\top \boldsymbol{y} \\
&= \big[(1+u)^2 w^2 + v^2\big]^{q/2} - w^q - qw^{q-2}uw^2 \\
&= (w^2 + 2uw^2 + r^2)^{q/2} - w^q - quw^q.
\end{aligned} \tag{154}$$

Since $r = \sqrt{(uw)^2 + v^2}$, we can rewrite $uw = r\delta$ for some scalar δ with $\delta \in [-\delta^*, 1]$, where $\delta^* = (w^2 + r^2)/(2wr)$, i.e., $w^2 - 2wr\delta^* + r^2 = 0$. Here, $|\delta|$ can be viewed as the projection length of $\boldsymbol{y}$ on the direction of $\boldsymbol{x}$, and the end point δ^* falls in $[0, 1]$. Then, (154) can be rewritten into

$$\varphi(\delta) := (w^2 + 2wr\delta + r^2)^{q/2} - w^q - qw^{q-1}r\delta. \tag{155}$$

Recall that $w = \|\boldsymbol{x}\|_2 > 0$ and $r = \|\boldsymbol{y}\|_2 > 0$. The first order derivative of $\varphi(\delta)$ is

$$\begin{aligned}
\varphi'(\delta) &= \frac{q}{2} 2wr(w^2 + 2wr\delta + r^2)^{(q/2)-1} - qw^{q-1}r \\
&= qwr\big[(w^2 + 2wr\delta + r^2)^{(q/2)-1} - w^{q-2}\big] \\
&= qw^{q-1}r\Big[\Big(1 + \frac{2r\delta}{w} + \frac{r^2}{w^2}\Big)^{(q/2)-1} - 1\Big].
\end{aligned} \tag{156}$$

This indicates that, for $q \geq 2$, $\varphi'(\delta) \leq 0$ when $\delta \in [-\delta^*, r/(2w)]$, and $\varphi'(\delta) > 0$ when $\delta \in (-r/(2w), 1]$. In particular, by Bernoulli's inequality, we can observe that

$$\begin{aligned}
\varphi(-\delta^*) &= -w^q + qw^{q-1}r\delta^* = \big(q/2 - 1\big)w^q + qw^{q-2}r^2 > 0, \\
\varphi(-r/(2w)) &= -qw^{q-2}r^2/2 < 0, \\
\varphi(1) &= (w + r)^q - w^q - qw^{q-1}r > 0.
\end{aligned} \tag{157}$$

Moreover, regarding $\varphi(-\delta^*)$ on $\delta^* \in [0, 1]$, we consider a new function $\tilde{\varphi}(s) = -w^q - qw^{q-1}rs$, which is decreasing on $s \in [-1, 0]$. Note that $\tilde{\varphi}(-1) = -w^q + qw^{q-1}r$. Thus, by comparison, we have $-\varphi(-r/(2w)) < \varphi(-\delta^*) \leq \tilde{\varphi}(-1) < \varphi(1)$. As a direct result, we obtain

$$\sup_{|\delta| \leq 1} |\varphi(\delta)| = \max\{\varphi(-\delta^*), -\varphi(-r/(2w)), \varphi(1)\} = \varphi(1). \tag{158}$$

By inserting $w = \|\boldsymbol{x}\|_2$ and $r = \|\boldsymbol{y}\|_2$ back to $\varphi(\delta)$, we obtain, for any $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^d$,

$$\Big|\|\boldsymbol{x} + \boldsymbol{y}\|_2^q - \|\boldsymbol{x}\|_2^q - q\|\boldsymbol{x}\|_2^{q-2}\boldsymbol{x}^\top \boldsymbol{y}\Big| \leq (\|\boldsymbol{x}\|_2 + \|\boldsymbol{y}\|_2)^q - \|\boldsymbol{x}\|_2^q - q\|\boldsymbol{x}\|_2^{q-1}\|\boldsymbol{y}\|_2. \tag{159}$$

The desired result holds by taking the expectation on the both sides. ■

Proof [Lemma 27(ii)] First, we define a function $\phi(t) = \|\boldsymbol{x} + t\boldsymbol{y}\|_2^q$ on $t \in [0, \infty)$. The first and second order derivatives of $\phi(t)$ are as follows

$$\begin{aligned}
\phi'(t) = \frac{d}{dt}\phi(t) &= q\|\boldsymbol{x} + t\boldsymbol{y}\|_2^{q-2}\big(t\|\boldsymbol{y}\|_2^2 + \boldsymbol{x}^\top\boldsymbol{y}\big), \\
\phi''(t) = \frac{d^2}{dt^2}\phi(t) &= q(q-2)\|\boldsymbol{x} + t\boldsymbol{y}\|_2^{q-4}\big(t\|\boldsymbol{y}\|_2^2 + \boldsymbol{x}^\top\boldsymbol{y}\big)^2 + q\|\boldsymbol{x} + t\boldsymbol{y}\|_2^{q-2}\|\boldsymbol{y}\|_2^2 \\
&= q(q-2)\|\boldsymbol{x} + t\boldsymbol{y}\|_2^{q-4}\big[(\boldsymbol{x} + t\boldsymbol{y})^\top\boldsymbol{y}\big]^2 + q\|\boldsymbol{x} + t\boldsymbol{y}\|_2^{q-2}\|\boldsymbol{y}\|_2^2 \\
&\le q(q-1)\|\boldsymbol{x} + t\boldsymbol{y}\|_2^{q-2}\|\boldsymbol{y}\|_2^2, \qquad (160)
\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. In the previous inequality, equality holds when both random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are scalars. Note that $\phi(1) = \|\boldsymbol{x} + \boldsymbol{y}\|_2^q$, $\phi(0) = \|\boldsymbol{x}\|_2^q$, and $\phi'(0) = q\|\boldsymbol{x}\|_2^{q-2}\boldsymbol{x}^\top\boldsymbol{y}$. Since

$$\begin{aligned}
\phi(1) - \phi(0) - \phi'(0) &= \int_{t=0}^1 \phi'(t)dt - \phi'(0) \\
&= \int_{t=0}^1 \Big(\int_{s=0}^t \phi''(s)ds + \phi'(0)\Big)dt - \phi'(0) \\
&= \int_{t=0}^1\int_{s=0}^t \phi''(s)dsdt, \qquad (161)
\end{aligned}$$

it follows from the upper bound of $\phi''(s)$ in (160) that

$$\|\boldsymbol{x} + \boldsymbol{y}\|_2^q - \|\boldsymbol{x}\|_2^q - q\|\boldsymbol{x}\|_2^{q-2}\boldsymbol{x}^\top\boldsymbol{y} \le q(q-1)\int_{t=0}^1\int_{s=0}^t \|\boldsymbol{x} + s\boldsymbol{y}\|_2^{q-2}\|\boldsymbol{y}\|_2^2 dsdt. \qquad (162)$$

Taking the expectation yields

$$\mathbb{E}\Big(\|\boldsymbol{x} + \boldsymbol{y}\|_2^q - \|\boldsymbol{x}\|_2^q - q\|\boldsymbol{x}\|_2^{q-2}\boldsymbol{x}^\top\boldsymbol{y}\Big) \le q(q-1)\int_{t=0}^1\int_{s=0}^t \mathbb{E}\big(\|\boldsymbol{x} + s\boldsymbol{y}\|_2^{q-2}\|\boldsymbol{y}\|_2^2\big)dsdt. \qquad (163)$$

Furthermore, by Hölder's inequality and the triangle inequality, we obtain

$$\begin{aligned}
\mathbb{E}\big(\|\boldsymbol{x} + s\boldsymbol{y}\|_2^{q-2}\|\boldsymbol{y}\|_2^2\big) &\le \big(\mathbb{E}\|\boldsymbol{x} + s\boldsymbol{y}\|_2^q\big)^{(q-2)/q}\big(\mathbb{E}\|\boldsymbol{y}\|_2^q\big)^{2/q} \\
&\le \Big[\big(\mathbb{E}\|\boldsymbol{x}\|_2^q\big)^{1/q} + s\big(\mathbb{E}\|\boldsymbol{y}\|_2^q\big)^{1/q}\Big]^{q-2}\big(\mathbb{E}\|\boldsymbol{y}\|_2^q\big)^{2/q}, \qquad (164)
\end{aligned}$$

which together with (163) gives

$$\begin{aligned}
&\mathbb{E}\Big(\|\boldsymbol{x} + \boldsymbol{y}\|_2^q - \|\boldsymbol{x}\|_2^q - q\|\boldsymbol{x}\|_2^{q-2}\boldsymbol{x}^\top\boldsymbol{y}\Big) \\
&\le q(q-1)\int_{t=0}^1\int_{s=0}^t \Big[\big(\mathbb{E}\|\boldsymbol{x}\|_2^q\big)^{1/q} + s\big(\mathbb{E}\|\boldsymbol{y}\|_2^q\big)^{1/q}\Big]^{q-2}\big(\mathbb{E}\|\boldsymbol{y}\|_2^q\big)^{2/q}dsdt \\
&= q\int_{t=0}^1 \Big\{\Big[\big(\mathbb{E}\|\boldsymbol{x}\|_2^q\big)^{1/q} + t\big(\mathbb{E}\|\boldsymbol{y}\|_2^q\big)^{1/q}\Big]^{q-1} - \big(\mathbb{E}\|\boldsymbol{x}\|_2^q\big)^{(q-1)/q}\Big\}\big(\mathbb{E}\|\boldsymbol{y}\|_2^q\big)^{1/q}dt \\
&= \Big[\big(\mathbb{E}\|\boldsymbol{x}\|_2^q\big)^{1/q} + \big(\mathbb{E}\|\boldsymbol{y}\|_2^q\big)^{1/q}\Big]^q - \mathbb{E}\|\boldsymbol{x}\|_2^q - q\big(\mathbb{E}\|\boldsymbol{x}\|_2^q\big)^{(q-1)/q}\big(\mathbb{E}\|\boldsymbol{y}\|_2^q\big)^{1/q}. \qquad (165)
\end{aligned}$$

In addition, recall that by the proof of Lemma 27(i), we have

$$\begin{aligned}\Big|\|\boldsymbol{x}+\boldsymbol{y}\|_2^q - \|\boldsymbol{x}\|_2^q - q\|\boldsymbol{x}\|_2^{q-2}\boldsymbol{x}^\top\boldsymbol{y}\Big| &\le (\|\boldsymbol{x}\|_2+\|\boldsymbol{y}\|_2)^q - \|\boldsymbol{x}\|_2^q - q\|\boldsymbol{x}\|_2^{q-1}\|\boldsymbol{y}\|_2 \\ &= q(q-1)\int_{t=0}^1\int_{s=0}^t (\|\boldsymbol{x}\|_2+s\|\boldsymbol{y}\|_2)^{q-2}\|\boldsymbol{y}\|_2^2 ds dt.\end{aligned}$$

Taking expectation on both sides, we obtain

$$\mathbb{E}\Big|\|\boldsymbol{x}+\boldsymbol{y}\|_2^q - \|\boldsymbol{x}\|_2^q - q\|\boldsymbol{x}\|_2^{q-2}\boldsymbol{x}^\top\boldsymbol{y}\Big| \le q(q-1)\int_0^1\int_0^t \mathbb{E}\big[(\|\boldsymbol{x}\|_2+s\|\boldsymbol{y}\|_2)^{q-2}\|\boldsymbol{y}\|_2^2\big]\, ds dt. \tag{166}$$

It follows from Hölder's inequality and the triangle inequality that

$$\begin{aligned}\mathbb{E}\big[(\|\boldsymbol{x}\|_2+s\|\boldsymbol{y}\|_2)^{q-2}\|\boldsymbol{y}\|_2^2\big] &\le \big[\mathbb{E}\big(\|\boldsymbol{x}\|_2+s\|\boldsymbol{y}\|_2\big)^q\big]^{(q-2)/q}\big(\mathbb{E}\|\boldsymbol{y}\|_2^q\big)^{2/q} \\ &\le \Big[\big(\mathbb{E}\|\boldsymbol{x}\|_2^q\big)^{1/q} + s\big(\mathbb{E}\|\boldsymbol{y}\|_2^q\big)^{1/q}\Big]^{q-2}\big(\mathbb{E}\|\boldsymbol{y}\|_2^q\big)^{2/q}. \end{aligned}\tag{167}$$

Evaluating the double integral as in (165) yields

$$\begin{aligned}&\mathbb{E}\Big|\|\boldsymbol{x}+\boldsymbol{y}\|_2^q - \|\boldsymbol{x}\|_2^q - q\|\boldsymbol{x}\|_2^{q-2}\boldsymbol{x}^\top\boldsymbol{y}\Big| \\ &\le \Big[\big(\mathbb{E}\|\boldsymbol{x}\|_2^q\big)^{1/q} + \big(\mathbb{E}\|\boldsymbol{y}\|_2^q\big)^{1/q}\Big]^q - \mathbb{E}\|\boldsymbol{x}\|_2^q - q\big(\mathbb{E}\|\boldsymbol{x}\|_2^q\big)^{(q-1)/q}\big(\mathbb{E}\|\boldsymbol{y}\|_2^q\big)^{1/q}.\end{aligned}$$

This completes the proof. ■

Proof [Lemma 10] Since a dropout matrix D_k is a diagonal matrix with values 0 and 1 on the diagonal, $D_k^2 = D_k$ and

$$\begin{aligned}\big\|(I_d-\alpha D_k\mathbb{X}_k D_k)\boldsymbol{v}\big\|_2^2 &= \boldsymbol{v}^\top\big(I_d - 2\alpha D_k\mathbb{X}_k D_k + \alpha^2 D_k\mathbb{X}_k D_k\mathbb{X}_k D_k\big)\boldsymbol{v} \\ &= 1 - 2\alpha\boldsymbol{v}^\top D_k\mathbb{X}_k D_k\boldsymbol{v} + \alpha^2\boldsymbol{v}^\top D_k\mathbb{X}_k D_k^2\mathbb{X}_k D_k\boldsymbol{v} \\ &= 1 - \alpha\boldsymbol{v}^\top\big[2D_k\mathbb{X}_k D_k - \alpha D_k\mathbb{X}_k D_k\mathbb{X}_k D_k\big]\boldsymbol{v} \\ &=: 1 - \alpha\boldsymbol{v}^\top M_k\boldsymbol{v},\end{aligned}\tag{168}$$

with

$$M_k(\alpha) = 2D_k\mathbb{X}_k D_k - \alpha D_k\mathbb{X}_k D_k\mathbb{X}_k D_k. \tag{169}$$

Recall the condition on the learning rate α in (40), it follows that $\mathbb{E}[M_k]$ is positive definite (p.d.), which further implies the lower bound $\mathbb{E}(\alpha\boldsymbol{v}^\top M_k\boldsymbol{v}) \ge \alpha\lambda_{\min}(\mathbb{E}[M_k]) > 0$, that holds uniformly over all unit vectors $\boldsymbol{v}$. As a direct consequence,

$$\sup_{\boldsymbol{v}\in\mathbb{R}^d,\|\boldsymbol{v}\|_2=1}\mathbb{E}\big\|(I_d-\alpha D_k\mathbb{X}_k D_k)\boldsymbol{v}\big\|_2^2 \le \sup_{\boldsymbol{v}\in\mathbb{R}^d,\|\boldsymbol{v}\|_2=1}\big(1-\mathbb{E}(\alpha\boldsymbol{v}^\top M_k\boldsymbol{v})\big) < 1, \tag{170}$$

proving the result in the case $q = 2$.

Next, we shall show that for all $q > 2$, we also have

$$\sup_{\boldsymbol{v}\in\mathbb{R}^d,\|\boldsymbol{v}\|_2=1}\mathbb{E}\|(I_d-\alpha D_k\mathbb{X}_k D_k)\boldsymbol{v}\|_2^q < 1.$$

In this case, the techniques in the proof of Lemma 2 cannot be directly applied due to the randomness of $\mathbb{X}_k$, and we need to leverage the moment inequalities in Lemma 27 instead. Specifically, let $\boldsymbol{x}$ and $\boldsymbol{y}$ in Lemma 27 be

$$\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{v}) = \boldsymbol{v}, \quad \boldsymbol{y} = \boldsymbol{y}(\boldsymbol{v}) = D_k \mathbb{X}_k D_k \boldsymbol{v}, \tag{171}$$

respectively, for $\boldsymbol{v}$ a deterministic d-dimensional unit vector. It remains to show that for any $q > 2$, $\mathbb{E}\|\boldsymbol{x} - \alpha \boldsymbol{y}\|_2^q < 1$ holds for any arbitrary unit vector $\boldsymbol{v}$. By Lemma 27,

$$\mathbb{E}\|\boldsymbol{x} - \alpha \boldsymbol{y}\|_2^q - \|\boldsymbol{x}\|_2^q - q\|\boldsymbol{x}\|_2^{q-2}\mathbb{E}(-\boldsymbol{x}^\top \alpha \boldsymbol{y}) \leq \mathbb{E}\big(\|\boldsymbol{x}\|_2 + \|\alpha \boldsymbol{y}\|_2\big)^q - \|\boldsymbol{x}\|_2^q - q\|\boldsymbol{x}\|_2^{q-1}\mathbb{E}\|\alpha \boldsymbol{y}\|_2, \tag{172}$$

which along with $\|\boldsymbol{x}\|_2 = \|\boldsymbol{v}\|_2 = 1$ further yields,

$$\mathbb{E}\|\boldsymbol{x} - \alpha \boldsymbol{y}\|_2^q - 1 + q\alpha\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{y}) \leq \mathbb{E}\big(1 + \alpha\|\boldsymbol{y}\|_2\big)^q - 1 - q\alpha\mathbb{E}\|\boldsymbol{y}\|_2. \tag{173}$$

Therefore, to prove $\sup_{\boldsymbol{v}\in\mathbb{R}^d, \|\boldsymbol{v}\|_2=1} \mathbb{E}\|\boldsymbol{x} - \alpha \boldsymbol{y}\|_2^q < 1$, it suffices to show

$$\mathbb{E}\big(1 + \alpha\|\boldsymbol{y}\|_2\big)^q - 1 - q\alpha\mathbb{E}\|\boldsymbol{y}\|_2 < q\alpha\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{y}). \tag{174}$$

By applying Lemma 27 again, we have

$$\mathbb{E}\big(1 + \alpha\|\boldsymbol{y}\|_2\big)^q - 1 - q\alpha\mathbb{E}\|\boldsymbol{y}\|_2 \leq \big(1 + \alpha(\mathbb{E}\|\boldsymbol{y}\|_2^q)^{1/q}\big)^q - 1 - q\alpha(\mathbb{E}\|\boldsymbol{y}\|_2^q)^{1/q}. \tag{175}$$

Thus, we only need to show that for any d-dimensional vector $\boldsymbol{v}$,

$$\big(1 + \alpha(\mathbb{E}\|\boldsymbol{y}\|_2^q)^{1/q}\big)^q - 1 - q\alpha(\mathbb{E}\|\boldsymbol{y}\|_2^q)^{1/q} < q\alpha\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{y}). \tag{176}$$

Recall the definitions of $\boldsymbol{x}$ and $\boldsymbol{y}$ in (171). By Lemma 21 (i), it follows that $\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{y}) = p\boldsymbol{v}^\top \mathbb{E}[\mathbb{X}_{k,p}]\boldsymbol{v}$, where $\mathbb{X}_{k,p} = p\mathbb{X}_k + (1-p)\mathrm{Diag}(\mathbb{X}_k)$. With $\mu_q = \mu_q(\boldsymbol{v}) = (\mathbb{E}\|\boldsymbol{y}(\boldsymbol{v})\|_2^q)^{1/q} = (\mathbb{E}\|D_k\mathbb{X}_k D_k \boldsymbol{v}\|_2^q)^{1/q}$, it suffices to show that

$$\sup_{\boldsymbol{v}\in\mathbb{R}^d, \|\boldsymbol{v}\|_2=1} \frac{\big(1 + \alpha\mu_q\big)^q - 1 - q\alpha\mu_q}{p\boldsymbol{v}^\top \mathbb{E}[\mathbb{X}_{k,p}]\boldsymbol{v}} < q\alpha. \tag{177}$$

Let $q > 2$. Consider a function $f : \mathbb{R}_+ \mapsto \mathbb{R}$ with $f(t) = (1+t)^q - 1 - qt$, which is strictly increasing on $\mathbb{R}_+$. Note that $f''(t) = q(q-1)(1+t)^{q-2}$. Since $q > 2$, it follows that $f(t) = \int_{s=0}^t \int_{u=0}^s q(q-1)(1+u)^{q-2} du ds$. Therefore,

$$f(t) = (1+t)^q - 1 - qt \leq \frac{q(q-1)}{2}(1+t)^{q-2}t^2. \tag{178}$$

Thus, the condition (177) is satisfied if

$$\frac{q(q-1)}{2}\alpha^2 \sup_{\boldsymbol{v}\in\mathbb{R}^d, \|\boldsymbol{v}\|_2=1} \frac{\big(1 + \alpha\mu_q\big)^{q-2}\mu_q^2}{p\boldsymbol{v}^\top \mathbb{E}[\mathbb{X}_{k,p}]\boldsymbol{v}} < q\alpha. \tag{179}$$

As this is true by assumption the proof is complete. ■

E.2 Proof of Lemma 13

Proof For any fixed unit vector $\boldsymbol{u} \in \mathbb{R}^d$, since $\breve{A}_k(\alpha) = I_d - \alpha D_k \mathbb{X}_k D_k$ as defined in (36) with $\mathbb{X}_k = \boldsymbol{x}_k \boldsymbol{x}_k^\top$, it follows that

$$\begin{aligned}\|\breve{A}_k(\alpha)\boldsymbol{v}\|_2^2 &= \|(I_d - \alpha D_k \boldsymbol{x}_k \boldsymbol{x}_k^\top D_k)\boldsymbol{v}\|_2^2 \\ &= \|\boldsymbol{v}\|_2^2 - 2\alpha \boldsymbol{v}^\top D_k \boldsymbol{x}_k \boldsymbol{x}_k^\top D_k \boldsymbol{v} + \alpha^2 \boldsymbol{v}^\top D_k \boldsymbol{x}_k \boldsymbol{x}_k^\top D_k^2 \boldsymbol{x}_k \boldsymbol{x}_k^\top D_k \boldsymbol{v} \\ &= 1 - 2\alpha(\boldsymbol{x}_k^\top D_k \boldsymbol{v})^2 + \alpha^2 \|D_k \boldsymbol{x}_k\|_2^2 (\boldsymbol{x}_k^\top D_k \boldsymbol{v})^2. \end{aligned} \tag{180}$$

Write $\boldsymbol{v} = (v_1, \ldots, v_d)^\top$ and $\boldsymbol{x}_k = (x_{k1}, \ldots, x_{kd})^\top$. Let $D_{k,jj}$ be the j-th diagonal element in D_k. Since $\boldsymbol{x}_k \sim \mathcal{N}(0, I_d)$ and D_k is independent of $\boldsymbol{x}_k$, it follows that

$$\mathbb{E}(\boldsymbol{x}_k^\top D_k \boldsymbol{v})^2 = \sum_{j=1}^d v_j^2 \mathbb{E}[D_{k,jj} x_{kj}^2] = \sum_{j=1}^d v_j^2 p = p. \tag{181}$$

Moreover,

$$\begin{aligned}\|D_k \boldsymbol{x}_k\|_2^2 (\boldsymbol{x}_k^\top D_k \boldsymbol{v})^2 &= \Big(\sum_{i=1}^d D_{k,ii} x_{ki}^2 \Big) \Big(\sum_{j,l=1}^d D_{k,jj} D_{k,ll} v_j v_l x_{kj} x_{kl} \Big) \\ &= \sum_{i,j,l=1}^d D_{k,ii} D_{k,jj} D_{k,ll} v_j v_l x_{ki}^2 x_{kj} x_{kl}. \end{aligned} \tag{182}$$

For $j \neq l$, $\mathbb{E}[x_{ki}^2 x_{kj} x_{kl}] = 0$. For $j = l$ and $i \neq j$, $\mathbb{E}[x_{ki}^2 x_{kj}^2] = 1$. For $j = l$ and $i = j$, $\mathbb{E}[x_{ki}^4] = 3$. Therefore, by the independence of the diagonal elements in D_k, we have

$$\begin{aligned}\mathbb{E}\big[\|D_k \boldsymbol{x}_k\|_2^2 (\boldsymbol{x}_k^\top D_k \boldsymbol{v})^2\big] &= \sum_{j=1}^d v_j^2 \mathbb{E}\Big(3 D_{k,jj} + \sum_{i \neq j} D_{k,ii} D_{k,jj} \Big) \\ &= \sum_{j=1}^d v_j^2 \big(3p + p^2(d-1) \big) \\ &= 3p + p^2(d-1). \end{aligned} \tag{183}$$

Finally, we obtain

$$\mathbb{E}\|\breve{A}_k(\alpha)\boldsymbol{v}\|_2^2 = 1 - 2\alpha p + \alpha^2 [3p + (d-1)p^2],$$

for all unit vectors $\boldsymbol{v} \in \mathbb{R}^d$. This completes the proof. ■

E.3 Proof of Theorem 12

Proof Let the random coefficient matrix $\breve{A}_k(\alpha) = I_d - \alpha D_k \mathbb{X}_k D_k$ be as defined in (36). We write $\breve{A}_k(\alpha) = \breve{A}_k$ exchangeably in this proof.

First, we study the case with $q = 2$. Consider two SGD dropout sequences $\{\breve{\boldsymbol{\beta}}_k(\alpha)\}_{k\in\mathbb{N}}$ and $\{\breve{\boldsymbol{\beta}}'_k(\alpha)\}_{k\in\mathbb{N}}$, given two arbitrarily fixed initial vectors $\breve{\boldsymbol{\beta}}_0$, $\breve{\boldsymbol{\beta}}'_0$. Let $\breve{\boldsymbol{\delta}} = \breve{\boldsymbol{\beta}}_0 - \breve{\boldsymbol{\beta}}'_0$. Then, it follows from the tower rule that

$$\begin{aligned}
\mathbb{E}\|\breve{\boldsymbol{\beta}}_k(\alpha) - \breve{\boldsymbol{\beta}}'_k(\alpha)\|_2^2 &= \mathbb{E}\big[\breve{\boldsymbol{\delta}}^\top \breve{A}_1^\top \cdots \breve{A}_k^\top \breve{A}_k \cdots \breve{A}_1 \breve{\boldsymbol{\delta}}\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[\breve{\boldsymbol{\delta}}^\top \breve{A}_1^\top \cdots \breve{A}_k^\top \breve{A}_k \cdots \breve{A}_1 \breve{\boldsymbol{\delta}} \mid \breve{A}_1, \ldots, \breve{A}_{k-1}\big]\big] \\
&= \mathbb{E}\big[\breve{\boldsymbol{\delta}}^\top \breve{A}_1^\top \cdots \breve{A}_{k-1}^\top \mathbb{E}(\breve{A}_k^\top \breve{A}_k) \breve{A}_{k-1} \cdots \breve{A}_1 \breve{\boldsymbol{\delta}}\big] \\
&\leq \|\mathbb{E}(\breve{A}_k^\top \breve{A}_k)\| \cdot \mathbb{E}\big[\breve{\boldsymbol{\delta}}^\top \breve{A}_1^\top \cdots \breve{A}_{k-1}^\top \breve{A}_{k-1} \cdots \breve{A}_1 \breve{\boldsymbol{\delta}}\big] \\
&\leq \prod_{i=1}^{k} \|\mathbb{E}(\breve{A}_i^\top \breve{A}_i)\| \cdot \|\breve{\boldsymbol{\delta}}\|_2^2. \qquad (184)
\end{aligned}$$

Recall that for the constant learning rate $\alpha > 0$ satisfying the conditions in Lemma 10, we have $\|\mathbb{E}(\breve{A}_i^\top \breve{A}_i)\| < 1$ uniformly over $i \in \mathbb{N}$. Thus, $\prod_{i=1}^{k} \|\mathbb{E}(\breve{A}_i^\top \breve{A}_i)\| < 1$. Since the dropout matrices D_k's are i.i.d. and are independent of the i.i.d. observations $\boldsymbol{x}_k$'s, it follows that $\prod_{i=1}^{k} \|\mathbb{E}(\breve{A}_i^\top \breve{A}_i)\| = \|\mathbb{E}(\breve{A}_1^\top \breve{A}_1)\|^k < 1$. This gives the desired result for the case with $q = 2$.

For $q > 2$, we note that

$$\begin{aligned}
\mathbb{E}\|\breve{\boldsymbol{\beta}}_k(\alpha) - \breve{\boldsymbol{\beta}}'_k(\alpha)\|_2^q = \mathbb{E}\|\breve{A}_k \cdots \breve{A}_1 \breve{\boldsymbol{\delta}}\|_2^q &= \mathbb{E}\big(\|\breve{A}_k \cdots \breve{A}_1 \breve{\boldsymbol{\delta}}\|_2^2\big)^{q/2} \\
&= \mathbb{E}\big(\breve{\boldsymbol{\delta}}^\top \breve{A}_1^\top \cdots \breve{A}_k^\top \breve{A}_k \cdots \breve{A}_1 \breve{\boldsymbol{\delta}}\big)^{q/2}. \qquad (185)
\end{aligned}$$

Similarly, it follows from the tower rule that

$$\begin{aligned}
\mathbb{E}\big(\breve{\boldsymbol{\delta}}^\top \breve{A}_1^\top \cdots \breve{A}_k^\top \breve{A}_k \cdots \breve{A}_1 \breve{\boldsymbol{\delta}}\big)^{q/2} &= \mathbb{E}\big[\mathbb{E}\big[\big(\breve{\boldsymbol{\delta}}^\top \breve{A}_1^\top \cdots \breve{A}_k^\top \breve{A}_k \cdots \breve{A}_1 \breve{\boldsymbol{\delta}}\big)^{q/2} \mid \breve{A}_1, \ldots, \breve{A}_{k-1}\big]\big] \\
&\leq \sup_{\boldsymbol{v}\in\mathbb{R}^d, \|\boldsymbol{v}\|_2=1} \mathbb{E}\|\breve{A}_k \boldsymbol{v}\|_2^q \cdot \mathbb{E}\big(\breve{\boldsymbol{\delta}}^\top \breve{A}_1^\top \cdots \breve{A}_{k-1}^\top \breve{A}_{k-1} \cdots \breve{A}_1 \breve{\boldsymbol{\delta}}\big)^{q/2} \\
&\leq \prod_{i=1}^{k} \sup_{\boldsymbol{v}\in\mathbb{R}^d, \|\boldsymbol{v}\|_2=1} \mathbb{E}\|\breve{A}_i \boldsymbol{v}\|_2^q \cdot \|\breve{\boldsymbol{\delta}}\|_2^q. \qquad (186)
\end{aligned}$$

Since $\sup_{\boldsymbol{v}\in\mathbb{R}^d, \|\boldsymbol{v}\|_2=1} \mathbb{E}\|\breve{A}_i \boldsymbol{v}\|_2^q < 1$ holds uniformly over $i \in \mathbb{N}$, we obtain the geometric-moment contraction in (41) for $q > 2$.

Finally, by Corollary 23, the geometric-moment contraction in (41) implies the existence of a unique stationary distribution $\breve{\pi}_\alpha$ of the SGD dropout $\{\breve{\boldsymbol{\beta}}_k(\alpha)\}_{k\in\mathbb{N}}$. This completes the proof. ■

E.4 Proofs of Lemmas 28–30

Lemma 28 (Closed-form solution of the ℓ^2 minimizer) *Assume that model (31) is in reduced form, i.e.,* $\min_i \mathbb{E}[\boldsymbol{x}_1 \boldsymbol{x}_1^\top]_{ii} > 0$. *Then, for the minimizer of the ℓ^2-regularized least-squares loss* $\breve{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \mathbb{E}\big[(y - \boldsymbol{x}^\top D\boldsymbol{\beta})^2/2\big]$ *as defined in (34), we have the closed form solution*

$$\breve{\boldsymbol{\beta}} = (\mathbb{E}[\mathbb{X}_{1,p}])^{-1} \mathbb{E}[y_1 \boldsymbol{x}_1].$$

Proof Recall the $d \times d$ Gram matrix $\mathbb{X}_k = \boldsymbol{x}_k \boldsymbol{x}_k^\top$ and

$$\overline{\mathbb{X}}_k = \mathbb{X}_k - \mathrm{Diag}(\mathbb{X}_k), \quad \mathbb{X}_{k,p} = p\mathbb{X}_k + (1-p)\mathrm{Diag}(\mathbb{X}_k).$$

To obtain the closed form solution, we first compute the gradient of the ℓ^2-regularized least-squares loss as follows,

$$\begin{aligned}
&(y - \boldsymbol{x}^\top \boldsymbol{\beta})^2 = y^2 - 2y\boldsymbol{x}^\top D\boldsymbol{\beta} + \boldsymbol{\beta}^\top D(\boldsymbol{x}\boldsymbol{x}^\top)D\boldsymbol{\beta},\\
&\mathbb{E}_D[(y - \boldsymbol{x}^\top \boldsymbol{\beta})^2] = y^2 - 2py\boldsymbol{x}^\top \boldsymbol{\beta} + p^2 \boldsymbol{\beta}^\top (\boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{\beta} + p(1-p)\boldsymbol{\beta}^\top \mathrm{Diag}(\boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{\beta},\\
&\mathbb{E}_{(y,\boldsymbol{x})}\mathbb{E}_D[(y - \boldsymbol{x}^\top \boldsymbol{\beta})^2] = \mathbb{E}[y^2] - 2p\mathbb{E}[y\boldsymbol{x}^\top]\boldsymbol{\beta} + p^2 \boldsymbol{\beta}^\top \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{\beta} + p(1-p)\boldsymbol{\beta}^\top \mathrm{Diag}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top])\boldsymbol{\beta},\\
&\nabla_{\boldsymbol{\beta}}\mathbb{E}_{(y,\boldsymbol{x})}\mathbb{E}_D[(y - \boldsymbol{x}^\top \boldsymbol{\beta})^2] = -2p\mathbb{E}[y\boldsymbol{x}] + 2\Big(p^2\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] + p(1-p)\mathrm{Diag}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top])\Big)\boldsymbol{\beta}.
\end{aligned}$$

Recall that the i.i.d. random noise ϵ_k is independent of the i.i.d. random covariates $\boldsymbol{x}_k$. Since model (31) is assumed to be in a reduced form, i.e., $\min_i \mathbb{E}[\boldsymbol{x}_1 \boldsymbol{x}_1^\top]_{ii} > 0$, the closed form solution of $\breve{\boldsymbol{\beta}}$ is

$$\breve{\boldsymbol{\beta}} = p\Big(p^2\mathbb{E}[\boldsymbol{x}_1\boldsymbol{x}_1^\top] + p(1-p)\mathrm{Diag}(\mathbb{E}[\boldsymbol{x}_1\boldsymbol{x}_1^\top])\Big)^{-1}\mathbb{E}[y_1\boldsymbol{x}_1] = (\mathbb{E}[\mathbb{X}_{1,p}])^{-1}\mathbb{E}[y_1\boldsymbol{x}_1].$$

This completes the proof. ■

Recall that for any $d \times d$ matrix A, $A_p := pA + (1-p)\mathrm{Diag}(A)$.

Lemma 29 *If the $d \times d$ matrix $\mathbb{E}[2\mathbb{X}_k - \alpha\mathbb{X}_k^2]_p$ is positive definite, then the condition on the learning rate α in (37) holds for $q = 2$.*

Proof By rewriting the condition (37) with $q = 2$, for all the unit vector $\boldsymbol{v} \in \mathbb{R}^d$, $\|\boldsymbol{v}\|_2 = 1$, we aim to show

$$0 < \alpha < \frac{2\boldsymbol{v}^\top \mathbb{E}(D_k\mathbb{X}_k D_k)\boldsymbol{v}}{\mathbb{E}\|D_k\mathbb{X}_k D_k\boldsymbol{v}\|_2^2}. \tag{187}$$

Since $D_k^2 = D_k \le I_d$, it follows from Lemma 21(ii) that

$$\begin{aligned}
&2\boldsymbol{v}^\top \mathbb{E}(D_k\mathbb{X}_k D_k)\boldsymbol{v} - \alpha\boldsymbol{v}^\top \mathbb{E}[D_k\mathbb{X}_k D_k\mathbb{X}_k D_k]\boldsymbol{v}\\
&\ge 2\boldsymbol{v}^\top \mathbb{E}(D_k\mathbb{X}_k D_k)\boldsymbol{v} - \alpha\boldsymbol{v}^\top \mathbb{E}[D_k\mathbb{X}_k^2 D_k]\boldsymbol{v}\\
&= \boldsymbol{v}^\top \mathbb{E}[D_k(2\mathbb{X}_k - \alpha\mathbb{X}_k^2)D_k]\boldsymbol{v}\\
&= p\boldsymbol{v}^\top \mathbb{E}[2\mathbb{X}_k - \alpha\mathbb{X}_k^2]_p\boldsymbol{v} > 0.
\end{aligned}$$

As the unit vector $\boldsymbol{v} \in \mathbb{R}^d$ was arbitrary, condition (37) holds for $q = 2$. ■

Lemma 30 (ℓ^2-minimizer $\breve{\boldsymbol{\beta}}$ and true parameter $\boldsymbol{\beta}^*$) *Assume that $\mathbb{E}[|\epsilon|^{2q}] + \|\boldsymbol{x}\|_2^{2q}] < \infty$. Then, the q-th moment of the gradient in (32) exists at the true parameter $\boldsymbol{\beta}^*$ in model (31), for some $q \ge 2$, that is,*

$$\Big(\mathbb{E}\Big\|\nabla_{\boldsymbol{\beta}^*}\frac{1}{2}\big(y - \boldsymbol{x}^\top D\boldsymbol{\beta}^*\big)^2\Big\|_2^q\Big)^{1/q} = \Big(\mathbb{E}\big\|D\boldsymbol{x}\big(y - \boldsymbol{x}^\top D\boldsymbol{\beta}^*\big)\big\|_2^q\Big)^{1/q} < \infty,$$

which further implies the finite q-th moment of the stochastic gradient at the ℓ^2-minimizer $\breve{\boldsymbol{\beta}}$ defined in (34), that is

$$\Big(\mathbb{E}\Big\|\nabla_{\breve{\boldsymbol{\beta}}}\frac{1}{2}\big(y-\boldsymbol{x}^\top D\breve{\boldsymbol{\beta}}\big)^2\Big\|_2^q\Big)^{1/q}=\Big(\mathbb{E}\big\|D\boldsymbol{x}\big(y-\boldsymbol{x}^\top D\breve{\boldsymbol{\beta}}\big)\big\|_2^q\Big)^{1/q}<\infty,$$

Proof First, it follows from the triangle inequality that

$$\begin{aligned}&\Big(\mathbb{E}\big\|D\boldsymbol{x}\big(y-\boldsymbol{x}^\top D\boldsymbol{\beta}^*\big)\big\|_2^q\Big)^{1/q}\\&=\Big(\mathbb{E}\big\|D\boldsymbol{x}\big(\boldsymbol{x}^\top\boldsymbol{\beta}^*+\epsilon-\boldsymbol{x}^\top D\boldsymbol{\beta}^*\big)\big\|_2^q\Big)^{1/q}\\&\le\big(\mathbb{E}\big\|D\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}^*\big\|_2^q\big)^{1/q}+\big(\mathbb{E}\big\|D\boldsymbol{x}\epsilon\big\|_2^q\big)^{1/q}+\big(\mathbb{E}\big\|D\boldsymbol{x}\boldsymbol{x}^\top D\boldsymbol{\beta}^*\big\|_2^q\big)^{1/q}.\end{aligned}\tag{188}$$

By Assumption 1, since the dimension of $\boldsymbol{\beta}^*$ is fixed, we have

$$\big(\mathbb{E}\big\|D\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}^*\big\|_2^q\big)^{1/q}\le\big(\mathbb{E}\|\boldsymbol{x}\|_2^{2q}\big)^{1/q}\|\boldsymbol{\beta}^*\|_2<\infty.$$

Due the independence between $\boldsymbol{x}$ and ϵ, Assumption 1 gives

$$\big(\mathbb{E}\big\|D\boldsymbol{x}\epsilon\big\|_2^q\big)^{1/q}\le(\mathbb{E}\|\boldsymbol{x}\|_2^q)^{1/q}(\mathbb{E}\|\epsilon\|_2^q)^{1/q}<\infty.$$

Moreover, we obtain

$$\big(\mathbb{E}\big\|D\boldsymbol{x}\boldsymbol{x}^\top D\boldsymbol{\beta}^*\big\|_2^q\big)^{1/q}=\big(\mathbb{E}\big\|D\boldsymbol{x}\big\|_2^{2q}\big)^{1/q}\|\boldsymbol{\beta}^*\|_2\le\big(\mathbb{E}\big\|\boldsymbol{x}\big\|_2^{2q}\big)^{1/q}\|\boldsymbol{\beta}^*\|_2<\infty.\tag{189}$$

Inserting the inequalities into (188), we obtain the finite q-th moment at the true parameter $\boldsymbol{\beta}^*$. Replacing $\boldsymbol{\beta}^*$ in (189) by any vector $\boldsymbol{\beta}\in\mathbb{R}^d$ satisfying $\|\boldsymbol{\beta}\|_2<\infty$, the result still holds.

Next, we show that the finite q-th moment of the stochastic gradient at $\boldsymbol{\beta}^*$ can also imply the finite q-th moment at $\breve{\boldsymbol{\beta}}$. Note that

$$\begin{aligned}&\Big(\mathbb{E}\big\|D_1\boldsymbol{x}_1\big(y_1-\boldsymbol{x}_1^\top D_1\breve{\boldsymbol{\beta}}\big)\big\|_2^q\Big)^{1/q}\\&\le\Big(\mathbb{E}\big\|D_1\boldsymbol{x}_1\big(y_1-\boldsymbol{x}_1^\top D_1\boldsymbol{\beta}^*\big)\big\|_2^q\Big)^{1/q}+\Big(\mathbb{E}\big\|D_1\mathbb{X}_1D_1(\breve{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\big\|_2^q\Big)^{1/q}.\end{aligned}\tag{190}$$

We only need to show that the second term is bounded. Since $\mathbb{X}_{1,p}=p\mathbb{X}_1+(1-p)\mathrm{Diag}(\mathbb{X}_1)$, $\overline{\mathbb{X}}_1=\mathbb{X}_1-\mathrm{Diag}(\mathbb{X}_1)$, and $\breve{\boldsymbol{\beta}}=(\mathbb{E}[\mathbb{X}_{1,p}])^{-1}\mathbb{E}[y_1\boldsymbol{x}_1]$, it follows that

$$\begin{aligned}\mathbb{E}\big\|D_1\mathbb{X}_1D_1(\breve{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\big\|_2^q&=\mathbb{E}\big\|D_1\mathbb{X}_1D_1\big((\mathbb{E}[\mathbb{X}_{1,p}])^{-1}\mathbb{E}[y_1\boldsymbol{x}_1]-\boldsymbol{\beta}^*\big)\big\|_2^q\\&=\mathbb{E}\big\|D_1\mathbb{X}_1D_1\big((\mathbb{E}[\mathbb{X}_{1,p}])^{-1}\mathbb{E}[(\boldsymbol{x}_1^\top\boldsymbol{\beta}^*+\epsilon_1)\boldsymbol{x}_1]-\boldsymbol{\beta}^*\big)\big\|_2^q\\&=\mathbb{E}\big\|D_1\mathbb{X}_1D_1\big((\mathbb{E}[\mathbb{X}_{1,p}])^{-1}\mathbb{E}[\mathbb{X}_1]\boldsymbol{\beta}^*-\boldsymbol{\beta}^*\big)\big\|_2^q\\&=(1-p)^q\mathbb{E}\big\|D_1\mathbb{X}_1D_1(\mathbb{E}[\mathbb{X}_{1,p}])^{-1}\mathbb{E}[\overline{\mathbb{X}}_1]\boldsymbol{\beta}^*\big\|_2^q\\&\le(1-p)^q\big\|(\mathbb{E}[\mathbb{X}_{1,p}])^{-1}\mathbb{E}[\overline{\mathbb{X}}_1]\big\|^q\sup_{\boldsymbol{v}\in\mathbb{R}^d,\|\boldsymbol{v}\|_2=1}\mathbb{E}\big\|D_1\mathbb{X}_1D_1\boldsymbol{v}\big\|_2^q\cdot\|\boldsymbol{\beta}^*\|_2^q.\end{aligned}\tag{191}$$

The sub-multiplicativity of the operator norm yields

$$\big\|(\mathbb{E}[\mathbb{X}_{1,p}])^{-1}\mathbb{E}[\overline{\mathbb{X}}_1]\big\| \le \frac{\lambda_{\max}(\mathbb{E}[\overline{\mathbb{X}}_1])}{\lambda_{\min}(\mathbb{E}[\mathbb{X}_{1,p}])} < \infty, \tag{192}$$

since $\lambda_{\min}(\mathbb{E}[\mathbb{X}_{1,p}]) \ge (1-p)\lambda_{\min}(\mathbb{E}[\mathrm{Diag}(\mathbb{X}_1)]) = (1-p)\min_i \mathbb{E}[\mathbb{X}_1]_{ii} > 0$, and $\lambda_{\max}(\mathbb{E}[\overline{\mathbb{X}}_1]) \le \lambda_{\max}(\mathbb{E}[\mathbb{X}_1]) < \infty$. Moreover, $\|\boldsymbol{\beta}^*\|_2 < \infty$. As we assume $\sup_{\boldsymbol{v}\in\mathbb{R}^d,\|\boldsymbol{v}\|_2=1}\mathbb{E}\|D_1\mathbb{X}_1D_1\boldsymbol{v}\|_2^q < \infty$ in Lemma 10, also (191) is bounded. ■

Appendix F. Proofs in Section 4.3

F.1 Proof of Lemma 14

Proof The recursion in (36) is $\breve{\boldsymbol{\beta}}_k(\alpha) - \breve{\boldsymbol{\beta}} = \breve{A}_k(\alpha)(\breve{\boldsymbol{\beta}}_{k-1}(\alpha) - \breve{\boldsymbol{\beta}}) + \breve{\boldsymbol{b}}_k(\alpha)$, with random matrix $\breve{A}_k(\alpha) = I_d - \alpha D_k\mathbb{X}_kD_k$, and random vector $\breve{\boldsymbol{b}}_k(\alpha) = \alpha D_k\boldsymbol{x}_k(y_k - \boldsymbol{x}_k^\top D_k\breve{\boldsymbol{\beta}})$. Recall $\breve{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d}\mathbb{E}[(y-\boldsymbol{x}^\top D\boldsymbol{\beta})^2/2]$ in (34), where the expectation is taken over both $(y,\boldsymbol{x})$ and D. By Lemma 21 (ii),

$$\begin{aligned}
\mathbb{E}_{(y,\boldsymbol{x})}\mathbb{E}_D[\breve{\boldsymbol{b}}_k(\alpha)] &= \mathbb{E}_{(y,\boldsymbol{x})}\mathbb{E}_D[\alpha D_k\boldsymbol{x}_k(y_k - \boldsymbol{x}_k^\top D_k\breve{\boldsymbol{\beta}})] \\
&= \mathbb{E}_{(y,\boldsymbol{x})}[\alpha p I_d y_k\boldsymbol{x}_k - \alpha p\mathbb{X}_{k,p}\breve{\boldsymbol{\beta}}] \\
&= \alpha p\mathbb{E}[\mathbb{X}_{1,p}]\breve{\boldsymbol{\beta}} - \alpha p\mathbb{E}[\mathbb{X}_{1,p}]\breve{\boldsymbol{\beta}} \\
&= 0.
\end{aligned} \tag{193}$$

Similar to (88), we can rewrite the stationary SGD dropout sequence $\breve{\boldsymbol{\beta}}_k^\circ(\alpha)$ into

$$\begin{aligned}
\breve{\boldsymbol{\beta}}_k^\circ(\alpha) - \breve{\boldsymbol{\beta}} &= \sum_{i=0}^{\infty}\Big(\prod_{j=k-i+1}^{k}\breve{A}_j(\alpha)\Big)\breve{\boldsymbol{b}}_{k-i}(\alpha) \\
&= \alpha\sum_{i=0}^{\infty}\Big(\prod_{j=k-i+1}^{k}(I_d - \alpha D_j\mathbb{X}_jD_j)\Big)D_{k-i}\boldsymbol{x}_{k-i}(y_{k-i} - \boldsymbol{x}_{k-i}^\top D_{k-i}\breve{\boldsymbol{\beta}}) \\
&=: \alpha\sum_{i=0}^{\infty}\breve{\mathcal{M}}_{i,k}(\alpha).
\end{aligned} \tag{194}$$

Recall the filtration $\breve{\mathcal{F}}_i = \sigma(\boldsymbol{\xi}_i, \boldsymbol{\xi}_{i-1}, \ldots)$ in (205) for $i \in \mathbb{Z}$, where $\boldsymbol{\xi}_i = (y_i, \boldsymbol{x}_i, D_i)$. Notice that $\mathbb{E}[\breve{\boldsymbol{b}}_k(\alpha)] = 0$ by (193), and therefore we have

$$\begin{aligned}
&\mathbb{E}[\breve{\mathcal{M}}_{i,k}(\alpha) \mid \breve{\mathcal{F}}_{k-i+1}] \\
&= \mathbb{E}\Big[\prod_{j=k-i+1}^{k}(I_d - \alpha D_j\mathbb{X}_jD_j)\;\Big|\;\breve{\mathcal{F}}_{k-i+1}\Big]\mathbb{E}\Big[D_{k-i}\boldsymbol{x}_{k-i}(y_{k-i} - \boldsymbol{x}_{k-i}^\top D_{k-i}\breve{\boldsymbol{\beta}})\Big] \\
&= \mathbb{E}\Big[\prod_{j=k-i+1}^{k}(I_d - \alpha D_j\mathbb{X}_jD_j)\;\Big|\;\breve{\mathcal{F}}_{k-i+1}\Big]\cdot\mathbb{E}[\breve{\boldsymbol{b}}_{k-i}(\alpha)] \\
&= 0.
\end{aligned} \tag{195}$$

Hence, for any $k \in \mathbb{N}$, $\{\breve{\mathcal{M}}_{i,k}(\alpha)\}_{i\in\mathbb{N}}$ is a sequence of martingale differences with respect to the filtration $\breve{\mathcal{F}}_{k-i}$. Let $t = k - i$. By applying Burkholder's inequality in Lemma 20, we have,

$$\begin{aligned}\big(\mathbb{E}\|\breve{\boldsymbol{\beta}}_k^\circ(\alpha) - \breve{\boldsymbol{\beta}}\|_2^q\big)^{1/q} &= \alpha\Big(\mathbb{E}\Big\|\sum_{t=-\infty}^{k} \breve{\mathcal{M}}_{k-t,k}(\alpha)\Big\|_2^q\Big)^{1/q} \\ &\lesssim \alpha\Big(\sum_{t=-\infty}^{k}\big(\mathbb{E}\|\breve{\mathcal{M}}_{k-t,k}(\alpha)\|_2^q\big)^{2/q}\Big)^{1/2}, \end{aligned} \tag{196}$$

where the constant in $\lesssim$ only depends on q. Denote the vector $\breve{\boldsymbol{s}}_t = D_t\boldsymbol{x}_t(y_t - \boldsymbol{x}_t^\top D_t\breve{\boldsymbol{\beta}})$ and the matrix product $\breve{A}_{(t+1):k} = \breve{A}_{(t+1):k}(\alpha) = \prod_{j=t+1}^k \breve{A}_j(\alpha)$ for simplicity. Then, we can write

$$\breve{\boldsymbol{b}}_t(\alpha) = \alpha\breve{\boldsymbol{s}}_t \quad \text{and} \quad \breve{\mathcal{M}}_{k-t,k}(\alpha) = \breve{A}_{(t+1):k}(\alpha)\breve{\boldsymbol{s}}_t. \tag{197}$$

For the case with $q = 2$, notice that $\breve{A}_{(t+1):k}$ is independent of $\breve{\boldsymbol{s}}_t$, and by the tower rule, we have

$$\begin{aligned}\mathbb{E}\|\breve{\mathcal{M}}_{k-t,k}(\alpha)\|_2^2 &= \mathbb{E}\|\breve{A}_{(t+1):k}\breve{\boldsymbol{s}}_t\|_2^2 \\ &= \mathbb{E}\big[\mathbb{E}\big[\breve{\boldsymbol{s}}_t^\top\breve{A}_{(t+1):k}^\top\breve{A}_{(t+1):k}\breve{\boldsymbol{s}}_t \mid \breve{\mathcal{F}}_t\big]\big] \\ &= \mathbb{E}\big[\mathbb{E}\big[\mathrm{tr}(\breve{\boldsymbol{s}}_t\breve{\boldsymbol{s}}_t^\top\breve{A}_{(t+1):k}^\top\breve{A}_{(t+1):k}) \mid \breve{\mathcal{F}}_t\big]\big] \\ &= \mathbb{E}\big[\mathrm{tr}\big(\mathbb{E}[\breve{\boldsymbol{s}}_t\breve{\boldsymbol{s}}_t^\top\breve{A}_{(t+1):k}^\top\breve{A}_{(t+1):k} \mid \breve{\mathcal{F}}_t]\big)\big] \\ &= \mathbb{E}\big[\mathrm{tr}\big(\mathbb{E}[\breve{\boldsymbol{s}}_t\breve{\boldsymbol{s}}_t^\top]\breve{A}_{(t+1):k}^\top\breve{A}_{(t+1):k}\big)\big] \\ &= \mathrm{tr}\big(\mathbb{E}[\breve{\boldsymbol{s}}_t\breve{\boldsymbol{s}}_t^\top]\mathbb{E}[\breve{A}_{(t+1):k}^\top\breve{A}_{(t+1):k}]\big) \\ &\le \big\|\mathbb{E}[\breve{A}_{(t+1):k}^\top\breve{A}_{(t+1):k}]\big\| \cdot \mathbb{E}\|\breve{\boldsymbol{s}}_t\|_2^2. \end{aligned} \tag{198}$$

Next, we shall bound the parts $\big\|\mathbb{E}[\breve{A}_{(t+1):k}^\top\breve{A}_{(t+1):k}]\big\|$ and $\mathbb{E}\|\breve{\boldsymbol{s}}_t\|_2^2$ separately. First, recall $\breve{A}_j(\alpha) = I_d - \alpha D_j\mathbb{X}_j D_j$ in (36), which are i.i.d. over j. By the tower rule with the induction over $j = t+1, t+2, \ldots, k$, we have

$$\begin{aligned}\big\|\mathbb{E}[\breve{A}_{(t+1):k}^\top\breve{A}_{(t+1):k}]\big\| &= \big\|\mathbb{E}\big[\mathbb{E}[\breve{A}_{(t+1):k}^\top\breve{A}_{(t+1):k} \mid \breve{A}_{t+1}, \breve{A}_{t+2}, \ldots, \breve{A}_{k-1}]\big]\big\| \\ &\le \big\|\mathbb{E}[\breve{A}_k^\top(\alpha)\breve{A}_k(\alpha)]\big\| \cdot \big\|\mathbb{E}[\breve{A}_{(t+1):(k-1)}^\top\breve{A}_{(t+1):(k-1)}]\big\| \\ &\le \prod_{j=t+1}^{k}\big\|\mathbb{E}[\breve{A}_j^\top(\alpha)\breve{A}_j(\alpha)]\big\| \\ &= \big\|\mathbb{E}[\breve{A}_1^\top(\alpha)\breve{A}_1(\alpha)]\big\|^{k-t}. \end{aligned} \tag{199}$$

Moreover, recall the random matrix $M_k(\alpha) = 2D_k\mathbb{X}_kD_k - \alpha D_k\mathbb{X}_kD_k\mathbb{X}_kD_k$ as defined in (169). For any unit vector $\boldsymbol{v} \in \mathbb{R}^d$, by (170) in the proof of Lemma 10, we have

$$\begin{aligned}\boldsymbol{v}^\top\mathbb{E}[\breve{A}_1^\top(\alpha)\breve{A}_1(\alpha)]\boldsymbol{v} &\le 1 - \alpha\boldsymbol{v}^\top\mathbb{E}[M_1(\alpha)]\boldsymbol{v} \\ &\le 1 - \alpha\lambda_{\min}\big(\mathbb{E}[M_1(\alpha)]\big) < 1, \end{aligned} \tag{200}$$

as the constant learning rate α satisfies condition (37). In fact, condition (37) also implies that $\mathbb{E}[M_k(\alpha)]$ is positive definite for each $k \in \mathbb{N}$, which can be seen by (40).

We shall show that the term $\mathbb{E}\|\breve{\boldsymbol{s}}_k\|_2^2 = \mathbb{E}\|D_k\boldsymbol{x}_k(y_k - \boldsymbol{x}_k^\top D_k\breve{\boldsymbol{\beta}})\|_2^2$ remains bounded as $k \to \infty$. In Assumption 1, we have assumed that the stochastic gradient in the SGD dropout recursion $\nabla_{\boldsymbol{\beta}}(y - \boldsymbol{x}^\top D\boldsymbol{\beta})^2/2 = D\boldsymbol{x}(y - \boldsymbol{x}^\top D\boldsymbol{\beta})$, has finite q-th moment when $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ for some $q \geq 2$. By Lemma 30, Assumption 1 also implies the bounded q-th moment when $\boldsymbol{\beta} = \breve{\boldsymbol{\beta}}$. As a direct consequence, $\mathbb{E}\|\breve{\boldsymbol{s}}_k\|_2^2$ is bounded as $k \to \infty$. This, along with (196), (198) and (200), yields

$$\begin{aligned}
\big(\mathbb{E}\|\breve{\boldsymbol{\beta}}_k^\circ(\alpha) - \breve{\boldsymbol{\beta}}\|_2^2\big)^{1/2} &\lesssim \alpha\Big(\sum_{t=-\infty}^{k} \mathbb{E}\|\breve{\mathcal{M}}_{k-t,k}(\alpha)\|_2^2\Big)^{1/2} \\
&\leq \alpha\Big(\sum_{t=-\infty}^{k} \big\|\mathbb{E}[\breve{A}_{(t+1):k}^\top \breve{A}_{(t+1):k}]\big\| \cdot \mathbb{E}\|\breve{\boldsymbol{s}}_t\|_2^2\Big)^{1/2} \\
&\lesssim \alpha\Big(\sum_{t=-\infty}^{k} (1 - \alpha\lambda_*)^{k-t}\Big)^{1/2} \\
&= \alpha\Big(\sum_{i=0}^{\infty} (1 - \alpha\lambda_*)^{i}\Big)^{1/2} \\
&= O(\sqrt{\alpha}),
\end{aligned} \tag{201}$$

where the last equation holds since $\sum_{i=0}^{\infty}(1 - \alpha\lambda_*)^i = \frac{1}{1-(1-\alpha\lambda_*)} = O(1/\alpha)$. Here, the constants in $\lesssim$ are independent of k and α, and $\lambda_* = \lambda_{\min}\big(\mathbb{E}[M_1(\alpha)]\big)$ is bounded away from zero since $\mathbb{E}[M_1(\alpha)] = \mathbb{E}[2D_1\mathbb{X}_1D_1 - \alpha D_1\mathbb{X}_1D_1\mathbb{X}_1D_1]$ is positive definite by condition (40).

For the case $q > 2$, following similar arguments as in (198), we obtain

$$\begin{aligned}
\mathbb{E}\|\breve{\mathcal{M}}_{k-t,k}(\alpha)\|_2^q &= \mathbb{E}(\|\breve{A}_{(t+1):k}\breve{\boldsymbol{s}}_t\|_2^2)^{q/2} \\
&= \mathbb{E}\big(\breve{\boldsymbol{s}}_t^\top \breve{A}_{(t+1):k}^\top \breve{A}_{(t+1):k}\breve{\boldsymbol{s}}_t\big)^{q/2} \\
&\leq \sup_{\boldsymbol{v}\in\mathbb{R}^d, \|\boldsymbol{v}\|_2=1} \mathbb{E}\|\breve{A}_{(t+1):k}\boldsymbol{v}\|_2^q \cdot \mathbb{E}\|\breve{\boldsymbol{s}}_t\|_2^q \\
&\leq \big(\sup_{\boldsymbol{v}\in\mathbb{R}^d, \|\boldsymbol{v}\|_2=1} \mathbb{E}\|\breve{A}_1\boldsymbol{v}\|_2^q\big)^{k-t} \cdot \mathbb{E}\|\breve{\boldsymbol{s}}_t\|_2^q.
\end{aligned} \tag{202}$$

With $\mu_q(\boldsymbol{v}) = (\mathbb{E}\|D_1\mathbb{X}_1D_1\boldsymbol{v}\|_2^q)^{1/q} < \infty$ as defined in Lemma 10 and the Equations (173) and (175) in the proof of Lemma 10, we have

$$\mathbb{E}\|\breve{A}_1\boldsymbol{v}\|_2^q \leq (1 + \alpha\mu_q(\boldsymbol{v}))^q - q\alpha\mu_q(\boldsymbol{v}) - q\alpha p\boldsymbol{v}^\top\mathbb{E}[\mathbb{X}_{1,p}]\boldsymbol{v}. \tag{203}$$

This, together with Taylor expansion around $\alpha = 0$ and the inequalities (196) and (202) gives finally $\max_k \big(\mathbb{E}\|\breve{\boldsymbol{\beta}}_k^\circ(\alpha) - \breve{\boldsymbol{\beta}}\|_2^q\big)^{1/q} = O(\sqrt{\alpha})$. ■

F.2 Proof of Theorem 15

The proofs of the quenched CLT and the invariance principle for averaged SGD dropout follow similar arguments as the ones for averaged GD dropout. The key differences lie in the functional dependence measures, because for SGD settings, both the dropout matrix D_k and the sequential observation $\boldsymbol{x}_k$ are random. We shall first introduce some necessary definitions and then proceed with the rigorous proofs.

Recall the generic dropout matrix $D \in \mathbb{R}^{d\times d}$ and random sample $(y, \boldsymbol{x}) \in \mathbb{R}\times\mathbb{R}^d$. For the SGD dropout sequence $\{\breve{\boldsymbol{\beta}}_k(\alpha)\}_{k\in\mathbb{N}}$, we define the centering term

$$\breve{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha) = \lim_{k\to\infty} \mathbb{E}[\breve{\boldsymbol{\beta}}_k(\alpha)] = \mathbb{E}[\breve{\boldsymbol{\beta}}_1^\circ(\alpha)], \tag{204}$$

where the expectation is taken over both $(y, \boldsymbol{x})$ and D, and $\breve{\boldsymbol{\beta}}_1^\circ(\alpha)$ defined in (44) follows the unique stationary distribution $\breve{\pi}_\alpha$. We first use Lemma 25 to prove the CLT for the partial sum of the stationary sequence $\{\breve{\boldsymbol{\beta}}_k^\circ(\alpha) - \breve{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)\}_{k\in\mathbb{N}}$, and then apply the geometric-moment contraction in Theorem 12 to show the quenched CLT for the partial sum of the non-stationary one $\{\breve{\boldsymbol{\beta}}_k(\alpha) - \breve{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)\}_{k\in\mathbb{N}}$. Finally, we extend the quenched CLT to the partial sum of $\{\breve{\boldsymbol{\beta}}_k(\alpha) - \breve{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)\}_{k\in\mathbb{N}}$ by providing the upper bound of $\breve{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha) - \breve{\boldsymbol{\beta}}$ in terms of the q-th moment for some $q \geq 2$.

Similar to Section D.1, we introduce the *functional dependence measure* in Wu (2005) for the stationary SGD dropout sequence $\{\breve{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$. However, the randomness in $\breve{\boldsymbol{\beta}}_k^\circ(\alpha)$ is induced from both the dropout matrix D_k and the random sample $(y_k, \boldsymbol{x}_k)$. Therefore, we define a new filtration $\breve{\mathcal{F}}_k$ by

$$\breve{\mathcal{F}}_k = \sigma(\boldsymbol{\xi}_k, \boldsymbol{\xi}_{k-1}, \ldots), \quad k \in \mathbb{Z}, \tag{205}$$

where the i.i.d. random elements $\boldsymbol{\xi}_k = (D_k, (y_k, \boldsymbol{x}_k))$, $k \in \mathbb{Z}$, are defined in (45). For any random vector $\boldsymbol{\zeta} \in \mathbb{R}^d$ satisfying $\mathbb{E}\|\boldsymbol{\zeta}\|_2 < \infty$, define projection operators

$$\breve{\mathcal{P}}_k[\boldsymbol{\zeta}] = \mathbb{E}[\boldsymbol{\zeta} \mid \breve{\mathcal{F}}_k] - \mathbb{E}[\boldsymbol{\zeta} \mid \breve{\mathcal{F}}_{k-1}], \quad k \in \mathbb{Z}. \tag{206}$$

By Theorem 12 and (44), there exists a measurable function $\breve{h}_\alpha(\cdot)$ such that the stationary SGD dropout sequence $\{\breve{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ can be represented by a causal process

$$\breve{\boldsymbol{\beta}}_k^\circ(\alpha) = \breve{h}_\alpha(\boldsymbol{\xi}_k, \boldsymbol{\xi}_{k-1}, \ldots) = \breve{h}_\alpha(\breve{\mathcal{F}}_k). \tag{207}$$

We denote the coupled version of $\breve{\mathcal{F}}_i$ by

$$\breve{\mathcal{F}}_{i,\{j\}} = \sigma(\boldsymbol{\xi}_i, \ldots, \boldsymbol{\xi}_{j+1}, \boldsymbol{\xi}_j', \boldsymbol{\xi}_{j-1} \ldots), \tag{208}$$

and let $\breve{\mathcal{F}}_{i,\{j\}} = \breve{\mathcal{F}}_i$ if $j > i$, where $\boldsymbol{\xi}_j'$ is an i.i.d. copy of $\boldsymbol{\xi}_i$. For $q > 1$, define the *functional dependence measure* of $\breve{\boldsymbol{\beta}}_k(\alpha)$ as

$$\breve{\theta}_{k,q}(\alpha) = \big(\mathbb{E}\|\breve{\boldsymbol{\beta}}_k^\circ(\alpha) - \breve{\boldsymbol{\beta}}_{k,\{0\}}^\circ(\alpha)\|_2^q\big)^{1/q}, \quad \text{where } \breve{\boldsymbol{\beta}}_{k,\{0\}}^\circ(\alpha) = \breve{h}_\alpha(\breve{\mathcal{F}}_{k,\{0\}}). \tag{209}$$

In addition, if $\sum_{k=0}^\infty \breve{\theta}_{k,q}(\alpha) < \infty$, we define the tail of *cumulative dependence measure* as

$$\breve{\Theta}_{m,q}(\alpha) = \sum_{k=m}^\infty \breve{\theta}_{k,q}(\alpha), \quad m \in \mathbb{N}. \tag{210}$$

Both $\breve{\theta}_{k,q}(\alpha)$ and $\breve{\Theta}_{k,q}(\alpha)$ are useful to study the dependence structure of the stationary SGD dropout iteration $\breve{\boldsymbol{\beta}}_k^\circ(\alpha) = f_{D_k,(y_k,\boldsymbol{x}_k)}(\breve{\boldsymbol{\beta}}_{k-1}^\circ(\alpha))$. To apply Lemma 25, we only need to show that the stationary SGD dropout sequence $\{\breve{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ is short-range dependent in the sense that $\breve{\Theta}_{0,q}(\alpha) < \infty$, for some $q \geq 2$.

Proof [Theorem 15] Consider two initial vectors $\breve{\boldsymbol{\beta}}_0^\circ, \breve{\boldsymbol{\beta}}_0^{\circ\prime}$ following the stationary distribution $\breve{\pi}_\alpha$ defined in Theorem 12. By the recursion in (36), we obtain two stationary SGD dropout sequences $\{\breve{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ and $\{\breve{\boldsymbol{\beta}}_k^{\circ\prime}(\alpha)\}_{k\in\mathbb{N}}$. It follows from Theorem 12 that for all $q \geq 2$,

$$\sup_{\breve{\boldsymbol{\beta}}_0^\circ, \breve{\boldsymbol{\beta}}_0^{\circ\prime} \in \mathbb{R}^d, \breve{\boldsymbol{\beta}}_0^\circ \neq \breve{\boldsymbol{\beta}}_0^{\circ\prime}} \frac{\big(\mathbb{E}\|\breve{\boldsymbol{\beta}}_k^\circ(\alpha) - \breve{\boldsymbol{\beta}}_k^{\circ\prime}(\alpha)\|_2^q\big)^{1/q}}{\|\breve{\boldsymbol{\beta}}_0^\circ - \breve{\boldsymbol{\beta}}_0^{\circ\prime}\|_2} \leq \breve{r}_{\alpha,q}^k, \quad k \in \mathbb{N}, \tag{211}$$

with $\breve{r}_{\alpha,q} = \big(\sup_{\boldsymbol{v}\in\mathbb{R}^d:\,\|\boldsymbol{v}\|_2=1} \mathbb{E}\|\breve{A}_1(\alpha)\boldsymbol{v}\|_2^q\big)^{1/q}$ as defined in (38) and random matrix $\breve{A}_1(\alpha) = I_d - \alpha D_1 \mathbb{X}_1 D_1$. When the constant learning rate $\alpha > 0$ satisfies the condition in (37), we have $\breve{r}_{\alpha,q} \in (0,1)$ as shown in Theorem 12. Recall the coupled filtration $\breve{\mathcal{F}}_{i,\{j\}} = \sigma(\boldsymbol{\xi}_i, \ldots, \boldsymbol{\xi}_{j+1}, \boldsymbol{\xi}_j', \boldsymbol{\xi}_{j-1} \ldots)$ as defined in (208). Then, (44) and (211) show that

$$\begin{aligned}
&\big(\mathbb{E}\|\breve{f}_{\boldsymbol{\xi}_k} \circ \cdots \circ \breve{f}_{\boldsymbol{\xi}_1}(\breve{\boldsymbol{\beta}}_0^\circ) - \breve{f}_{\boldsymbol{\xi}_k} \circ \cdots \circ \breve{f}_{\boldsymbol{\xi}_1}(\breve{\boldsymbol{\beta}}_0^{\circ\prime})\|_2^q\big)^{1/q} \\
&= \big(\mathbb{E}\|\breve{h}_\alpha(\boldsymbol{\xi}_k, \ldots, \boldsymbol{\xi}_1, \boldsymbol{\xi}_0, \boldsymbol{\xi}_{-1}, \ldots) - \breve{h}_\alpha(\boldsymbol{\xi}_k, \ldots, \boldsymbol{\xi}_1, \boldsymbol{\xi}_0', \boldsymbol{\xi}_{-1}', \ldots)\|_2^q\big)^{1/q} \\
&= \big(\mathbb{E}\|\breve{h}_\alpha(\breve{\mathcal{F}}_k) - \breve{h}_\alpha(\breve{\mathcal{F}}_{k,\{0,-1,\ldots\}})\|_2^q\big)^{1/q} \\
&\leq \breve{c}_q \breve{r}_{\alpha,q}^k,
\end{aligned} \tag{212}$$

for some constant $\breve{c}_q > 0$ that is independent of k. Following a similar argument in (127), for all $q \geq 2$ and $k \in \mathbb{N}$, we can bound the functional dependence measure $\breve{\theta}_{k,q}(\alpha)$ in (209) as follows

$$\begin{aligned}
\breve{\theta}_{k,q}(\alpha) &= \big(\mathbb{E}\|\breve{h}_\alpha(\breve{\mathcal{F}}_k) - \breve{h}_\alpha(\breve{\mathcal{F}}_{k,\{0\}})\|_2^q\big)^{1/q} \\
&\leq \big(\mathbb{E}\|\breve{h}_\alpha(\breve{\mathcal{F}}_k) - \breve{h}_\alpha(\breve{\mathcal{F}}_{k,\{0,-1,\ldots\}})\|_2^q\big)^{1/q} + \big(\mathbb{E}\|\breve{h}_\alpha(\breve{\mathcal{F}}_{k,\{0,-1,\ldots\}}) - \breve{h}_\alpha(\breve{\mathcal{F}}_{k,\{0\}})\|_2^q\big)^{1/q} \\
&\leq \breve{c}_q' \breve{r}_{\alpha,q}^k,
\end{aligned} \tag{213}$$

for some constant $\breve{c}_q' > 0$ that is independent of k. Consequently, the cumulative dependence measure $\breve{\Theta}_{m,q}(\alpha)$ in (210) is also bounded for all $q \geq 2$ and $m \in \mathbb{N}$, that is,

$$\breve{\Theta}_{m,q}(\alpha) = \sum_{k=m}^\infty \breve{\theta}_{k,q}(\alpha) = O(\breve{r}_{\alpha,q}^m) < \infty. \tag{214}$$

The inequality in (119) derived by Wu (2005) holds for a general class of functional dependence measures, as long as the inputs of the functional system (i.e., the measurable function $\breve{h}(\boldsymbol{\xi}_k, \boldsymbol{\xi}_{k-1}, \ldots)$) are i.i.d. elements. Thus, we can apply (119) to the projection operator $\breve{\mathcal{P}}_k[\cdot]$ in (206) and obtain

$$\sum_{k=0}^\infty \big(\mathbb{E}\|\breve{\mathcal{P}}_0[\breve{\boldsymbol{\beta}}_k^\circ(\alpha)]\|_2^q\big)^{1/q} \leq \sum_{k=0}^\infty \breve{\theta}_{k,q}(\alpha) = \breve{\Theta}_{0,q}(\alpha) < \infty, \tag{215}$$

implying the short-range dependence of the stationary SGD dropout sequence $\{\breve{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$. Then, it follows from Lemma 25 that

$$t^{-1/2}\sum_{k=1}^{t}\big(\breve{\boldsymbol{\beta}}_k^\circ(\alpha)-\breve{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)\big)\Rightarrow\mathcal{N}(0,\breve{\Sigma}(\alpha)),\tag{216}$$

where the long-run covariance matrix $\breve{\Sigma}(\alpha)$ is defined in Theorem 15. Following similar arguments as in (131)–(133), for any initial vector $\breve{\boldsymbol{\beta}}_0\in\mathbb{R}^d$, we can leverage the geometric-moment contraction in Theorem 12 and achieve the quenched CLT for the corresponding SGD dropout sequence $\{\breve{\boldsymbol{\beta}}_k(\alpha)\}_{k\in\mathbb{N}}$, that is,

$$t^{-1/2}\sum_{k=1}^{t}\big(\breve{\boldsymbol{\beta}}_k(\alpha)-\breve{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)\big)\Rightarrow\mathcal{N}(0,\breve{\Sigma}(\alpha)).\tag{217}$$

Recall the ℓ^2-minimizer $\breve{\boldsymbol{\beta}}$ in (34) and the centering term $\breve{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)=\lim_{k\to\infty}\mathbb{E}(\breve{\boldsymbol{\beta}}_k(\alpha))=\mathbb{E}(\breve{\boldsymbol{\beta}}_1^\circ(\alpha))$ in (204). We shall prove $\|\sum_{k=1}^{t}\mathbb{E}[\breve{\boldsymbol{\beta}}_k(\alpha)-\breve{\boldsymbol{\beta}}]\|_2=o(\sqrt{t})$. For any two initial vectors $\breve{\boldsymbol{\beta}}_0$ and $\breve{\boldsymbol{\beta}}_0^\circ$, where $\breve{\boldsymbol{\beta}}_0^\circ$ follows the stationary distribution $\breve{\pi}_\alpha$ in Theorem 12, while $\breve{\boldsymbol{\beta}}_0$ is an arbitrary initial vector in $\mathbb{R}^d$, it follows from the triangle inequality that

$$\begin{aligned}\Big\|\sum_{k=1}^{t}\mathbb{E}[\breve{\boldsymbol{\beta}}_k(\alpha)-\breve{\boldsymbol{\beta}}]\Big\|_2&=\Big\|\mathbb{E}\Big[\sum_{k=1}^{t}\big(\breve{\boldsymbol{\beta}}_k(\alpha)-\breve{\boldsymbol{\beta}}_k^\circ(\alpha)+\breve{\boldsymbol{\beta}}_k^\circ(\alpha)-\breve{\boldsymbol{\beta}}\big)\Big]\Big\|_2\\&\le\Big\|\mathbb{E}\Big[\sum_{k=1}^{t}\big(\breve{\boldsymbol{\beta}}_k(\alpha)-\breve{\boldsymbol{\beta}}_k^\circ(\alpha)\big)\Big]\Big\|_2+\Big\|\mathbb{E}\Big[\sum_{k=1}^{t}\big(\breve{\boldsymbol{\beta}}_k^\circ(\alpha)-\breve{\boldsymbol{\beta}}\big)\Big]\Big\|_2\\&=:\breve{\mathrm{I}}_1+\breve{\mathrm{I}}_2.\end{aligned}\tag{218}$$

For the part $\breve{\mathrm{I}}_1$, Jensen's inequality and a similar argument as for (131) yield

$$\begin{aligned}\breve{\mathrm{I}}_1&=\Big\|\mathbb{E}\Big[\sum_{k=1}^{t}\big(\breve{\boldsymbol{\beta}}_k(\alpha)-\breve{\boldsymbol{\beta}}_k^\circ(\alpha)\big)\Big]\Big\|_2\\&\le\Big(\mathbb{E}\Big\|\sum_{k=1}^{t}\big(\breve{\boldsymbol{\beta}}_k(\alpha)-\breve{\boldsymbol{\beta}}_k^\circ(\alpha)\big)\Big\|_2^2\Big)^{1/2}\\&\le\Big(\sum_{k=1}^{t}\breve{r}_{\alpha,2}^{k}\Big)\|\breve{\boldsymbol{\beta}}_0-\breve{\boldsymbol{\beta}}_0^\circ\|_2.\end{aligned}\tag{219}$$

For the part $\breve{\mathrm{I}}_2$, we recall the random matrix $\breve{A}_1(\alpha)=I_d-\alpha D_1\mathbb{X}_1D_1$ in (36) with $\mathbb{X}_1=\boldsymbol{x}_1\boldsymbol{x}_1^\top$. Recall the notation $\mathbb{X}_{1,p}=p\mathbb{X}_1+(1-p)\mathrm{Diag}(\mathbb{X}_1)$ in (35). Notice that by Lemma 21 (ii), we have $\mathbb{E}_{(y,\boldsymbol{x})}\mathbb{E}_D[\breve{A}_1(\alpha)]=\mathbb{E}_{(y,\boldsymbol{x})}\mathbb{E}_D[I_d-\alpha D_1\mathbb{X}_1D_1]=\mathbb{E}_{(y,\boldsymbol{x})}[I_d-\alpha p\mathbb{X}_{1,p}]=I_d-\alpha p\mathbb{E}[\mathbb{X}_{1,p}]$, which along with $\mathbb{E}[\breve{\boldsymbol{b}}_k]=0$ in (193) gives $\mathbb{E}[\breve{\boldsymbol{\beta}}_k^\circ(\alpha)-\breve{\boldsymbol{\beta}}]=\big(I_d-\alpha p\mathbb{E}[\mathbb{X}_{1,p}]\big)\mathbb{E}[\breve{\boldsymbol{\beta}}_{k-1}^\circ(\alpha)-\breve{\boldsymbol{\beta}}]$. Since $\{\breve{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ is stationary and $\mathbb{E}[\mathbb{X}_{1,p}]$ is non-singular by the reduced-form condition $\min_i(\mathbb{E}[\boldsymbol{x}_1\boldsymbol{x}_1^\top])_{ii}>0$ imposed on model (31), it follows that

$$\mathbb{E}[\breve{\boldsymbol{\beta}}_k^\circ(\alpha)-\breve{\boldsymbol{\beta}}]=0,\quad\text{for all } k\in\mathbb{N}.\tag{220}$$

This further yields

$$\breve{\mathrm{I}}_2 = \Big\| \mathbb{E}\Big[\sum_{k=1}^{t} \big(\breve{\boldsymbol{\beta}}_k^{\circ}(\alpha) - \breve{\boldsymbol{\beta}} \big) \Big] \Big\|_2 = \Big\| \sum_{k=1}^{t} \mathbb{E}[\breve{\boldsymbol{\beta}}_k^{\circ}(\alpha) - \breve{\boldsymbol{\beta}}] \Big\|_2 = 0. \tag{221}$$

By applying the results for $\breve{\mathrm{I}}_1$ and $\breve{\mathrm{I}}_2$ to (218), we obtain

$$\Big\| \sum_{k=1}^{t} \mathbb{E}[\breve{\boldsymbol{\beta}}_k(\alpha) - \breve{\boldsymbol{\beta}}] \Big\|_2 \leq \Big(\sum_{k=1}^{t} \breve{r}_{\alpha,2}^{k} \Big) \big\| \breve{\boldsymbol{\beta}}_0 - \breve{\boldsymbol{\beta}}_0^{\circ} \big\|_2. \tag{222}$$

When the constant learning rate $\alpha > 0$ satisfies the condition in (37), $\breve{r}_{\alpha,2} \in (0,1)$ by Theorem 12, and hence (222) remains bounded as $t \to \infty$. By this result and (217), the desired quenched CLT for the partial sum $\sum_{k=1}^{t} (\breve{\boldsymbol{\beta}}_k^{\circ}(\alpha) - \breve{\boldsymbol{\beta}})$ follows. ■

Proof The proof of Corollary 16 applies the Cramér-Wold device to Theorem 15 and can be derived in the same way as the proof of Corollary 8. We omit the details here. ■

F.3 Proof of Theorem 17

Proof Consider a stationary SGD dropout sequence $\{\breve{\boldsymbol{\beta}}_k^{\circ}(\alpha)\}_{k\in\mathbb{N}}$ following the stationary distribution $\breve{\pi}_\alpha$ in Theorem 12. Define the mean-zero stationary partial sum

$$\breve{S}_i^{\circ}(\alpha) = \sum_{k=1}^{i} \big(\breve{\boldsymbol{\beta}}_k^{\circ}(\alpha) - \breve{\boldsymbol{\beta}}_{\mathrm{mean}} \big), \quad i \in \mathbb{N}. \tag{223}$$

Recall that $\breve{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha) = \mathbb{E}[\breve{\boldsymbol{\beta}}_1^{\circ}(\alpha)]$ as defined in (204). Due to the stationarity, we have $\mathbb{E}[\breve{\boldsymbol{\beta}}_k^{\circ}(\alpha) - \breve{\boldsymbol{\beta}}] = 0$ for all $k \in \mathbb{N}$ by (220). This gives

$$\big(\mathbb{E}\|\breve{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha) - \breve{\boldsymbol{\beta}}\|_2^q \big)^{1/q} = \big(\mathbb{E}\|\mathbb{E}[\breve{\boldsymbol{\beta}}_1^{\circ}(\alpha) - \breve{\boldsymbol{\beta}}]\|_2^q \big)^{1/q} = 0. \tag{224}$$

Since we supposed in Theorem 17 that Assumption 1 holds for some $q > 2$, it follows from Lemma 14 that, for all $q > 2$,

$$\big(\mathbb{E}\|\breve{\boldsymbol{\beta}}_k^{\circ}(\alpha) - \breve{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)\|_2^q \big)^{1/q} \leq \big(\mathbb{E}\|\breve{\boldsymbol{\beta}}_k^{\circ}(\alpha) - \breve{\boldsymbol{\beta}}\|_2^q \big)^{1/q} + \big(\mathbb{E}\|\breve{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha) - \breve{\boldsymbol{\beta}}\|_2^q \big)^{1/q} = O(\sqrt{\alpha}). \tag{225}$$

Following a similar argument as for (144), we can show that $\breve{\boldsymbol{\beta}}_k^{\circ}(\alpha) - \breve{\boldsymbol{\beta}}_{\mathrm{mean}}(\alpha)$ satisfies condition (i) on the uniform integrability in Lemma 26. Due to the stationarity of $\{\breve{\boldsymbol{\beta}}_k^{\circ}(\alpha)\}_{k\in\mathbb{N}}$, condition (ii) in Lemma 26 is not required (see the discussion below Lemma 26 for details). Condition (iii) is also satisfied, since when the constant learning rate $\alpha > 0$ satisfies condition (37), the stationary SGD dropout sequence $\{\breve{\boldsymbol{\beta}}_k^{\circ}(\alpha)\}_{k\in\mathbb{N}}$ is shown to be short-range dependent by (215), i.e., the tail of cumulative dependence measure $\breve{\Theta}_{m,q}(\alpha) = \sum_{k=m}^{\infty} \breve{\theta}_{k,q}(\alpha) < \infty$.

Hence, by Lemma 26, there exists a (richer) probability space $(\breve{\Omega}^\star, \breve{\mathcal{A}}^\star, \breve{\mathbb{P}}^\star)$ on which we can define random vectors $\breve{\boldsymbol{\beta}}_k^\star$'s with the partial sum process $\breve{S}_i^\star = \sum_{k=1}^i (\breve{\boldsymbol{\beta}}_k^\star - \breve{\boldsymbol{\beta}}_{\text{mean}})$, and a Gaussian process $\breve{G}_i^\star = \sum_{k=1}^i \breve{\boldsymbol{z}}_k^\star$, where $\breve{\boldsymbol{z}}_k^\star$'s are independent Gaussian random vectors in $\mathbb{R}^d$ following $\mathcal{N}(0, I_d)$, such that

$$(\breve{S}_i^\star)_{1\le i\le t} \overset{\mathcal{D}}{=} (\breve{S}_i^\circ)_{1\le i\le t}, \tag{226}$$

and

$$\max_{1\le i\le t} \left\| \breve{S}_i^\star - \breve{\Sigma}^{1/2}(\alpha)\breve{G}_i^\star \right\|_2 = o_{\mathbb{P}}(t^{1/q}), \quad \text{in } (\breve{\Omega}^\star, \breve{\mathcal{A}}^\star, \breve{\mathbb{P}}^\star), \tag{227}$$

where the long-run covariance matrix $\breve{\Sigma}(\alpha)$ is defined in Theorem 15.

Following similar arguments as for (148)–(150), we can leverage the geometric-moment contraction in Theorem 12 to show the same Gaussian approximation rate, i.e., $o_{\mathbb{P}}(t^{1/q})$, for the partial sum sequence $\left(\sum_{k=1}^i (\breve{\boldsymbol{\beta}}_k(\alpha) - \breve{\boldsymbol{\beta}}_{\text{mean}}(\alpha))\right)_{1\le i\le t}$, for any arbitrarily fixed initial vector $\breve{\boldsymbol{\beta}}_0 \in \mathbb{R}^d$. Finally, recall the partial sum process $\breve{S}_i^{\breve{\boldsymbol{\beta}}_0}(\alpha) = \sum_{k=1}^i (\breve{\boldsymbol{\beta}}_k(\alpha) - \breve{\boldsymbol{\beta}})$. By a similar argument in (151), the desired Gaussian approximation result for the partial sum process $(\breve{S}_i^{\breve{\boldsymbol{\beta}}_0}(\alpha))_{1\le i\le t}$ follows. ∎

Appendix G. Proofs in Section 5

G.1 Proof of Theorem 18

Lemma 31 *For any $d \times d$ symmetric matrix $S = (S_{ij})$, we have $\mathbb{E}\|S\|_F \le \sqrt{\text{tr}\mathbb{E}(S^2)} \le d \max_{i,j} \left(\mathbb{E}[S_{ij}^2]\right)^{1/2}$.*

Proof For a symmetric matrix S, $\text{tr}(S^2) = \|S\|_F^2 = \sum_{i,j} S_{ij}^2$. Since $\sqrt{(\cdot)}$ is a concave function, by Jensen's inequality,

$$\mathbb{E}\|S\|_F = \mathbb{E}\sqrt{\text{tr}(S^2)} \le \sqrt{\text{tr}[\mathbb{E}(S^2)]} = \sqrt{\sum_{i,j=1}^d \mathbb{E}[S_{ij}^2]} \le d \max_{i,j} \left(\mathbb{E}[S_{ij}^2]\right)^{1/2}.$$

This completes the proof. ∎

Proof Recall the SGD dropout sequence $\{\breve{\boldsymbol{\beta}}_k(\alpha)\}_{k\in\mathbb{N}}$ in (36), the long-run covariance matrix $\breve{\Sigma}(\alpha)$ of the averaged SGD dropout iterates in Theorem 15, and the online estimator $\hat{\Sigma}_k(\alpha)$ in (55). When there is no ambiguity, we omit the dependence on α, e.g., $\hat{\Sigma}_k = \hat{\Sigma}_k(\alpha)$ and $\breve{\Sigma} = \breve{\Sigma}(\alpha)$. We shall bound $\mathbb{E}\|\hat{\Sigma}_k - \breve{\Sigma}\|_F$.

Recall the stationary SGD dropout sequence $\{\breve{\boldsymbol{\beta}}_k^\circ(\alpha)\}_{k\in\mathbb{N}}$ in (42), which follows the stationary distribution $\breve{\pi}_\alpha$ in Theorem 12. For simplicity, we define $V_k(\alpha) = k\hat{\Sigma}_k(\alpha)$. By

Equation (58), we can write $V_k(\alpha)$ into

$$V_k(\alpha) = \Big(\sum_{m=1}^{\psi(k)-1} \mathcal{S}_m(\alpha)^{\otimes 2} + \mathcal{R}_k(\alpha)^{\otimes 2} \Big) + \Big(\sum_{m=1}^{\psi(k)-1} |B_m|^2 + |\delta_\eta(k)|^2 \Big) \bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha)^{\otimes 2} - \Big(\sum_{m=1}^{\psi(k)-1} |B_m| \mathcal{S}_m(\alpha) + \delta_\eta(k) \mathcal{R}_k(\alpha) \Big) \bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha)^\top - \bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}(\alpha) \Big(\sum_{m=1}^{\psi(k)-1} |B_m| \mathcal{S}_m(\alpha) + \delta_\eta(k) \mathcal{R}_k(\alpha) \Big)^\top. \tag{228}$$

Recall the partial sums $\mathcal{S}_m(\alpha) = \sum_{k \in B_m} \breve{\boldsymbol{\beta}}_k(\alpha)$ and $\mathcal{R}_k(\alpha) = \sum_{i=\eta_{\psi(k)}}^{k} \breve{\boldsymbol{\beta}}_i(\alpha)$ in (57). We similarly define

$$V_k^\circ(\alpha) = \sum_{m=1}^{\psi(k)-1} \Big(\sum_{i \in \mathcal{B}_m} \breve{\boldsymbol{\beta}}_i^\circ(\alpha) \Big)^{\otimes 2} + \Big(\sum_{i=\eta_{\psi(k)}}^{k} \breve{\boldsymbol{\beta}}_i^\circ(\alpha) \Big)^{\otimes 2} =: \sum_{m=1}^{\psi(k)-1} \mathcal{S}_m^\circ(\alpha)^{\otimes 2} + \mathcal{R}_k^\circ(\alpha)^{\otimes 2}. \tag{229}$$

By the triangle inequality, we have

$$k \mathbb{E} \|\hat{\Sigma}_k - \breve{\Sigma}\|_F = \mathbb{E} \|V_k - k\breve{\Sigma}\|_F \le \mathbb{E} \|V_k - V_k^\circ\|_F + \mathbb{E} \|V_k^\circ - k\breve{\Sigma}\|_F. \tag{230}$$

We shall bound these two terms separately.

First, for the term $\mathbb{E}\|V_k^\circ - k\breve{\Sigma}\|_F$, we shall use the results in Xiao and Wu (2011). To this end, we need to verify the assumptions on the weak dependence of SGD dropout iterates $\{\breve{\boldsymbol{\beta}}_k\}_{k \in \mathbb{N}}$ in (36) and the growing sizes of blocks $\{B_m\}_{m \in \mathbb{N}}$ in (54). Denote the elements of $d \times d$ matrices V_k° and $\breve{\Sigma}$ respectively by

$$V_k^\circ =: (v_{ij,k}^\circ)_{1 \le i,j \le d} \quad \text{and} \quad \breve{\Sigma} =: (\breve{\sigma}_{ij})_{1 \le i,j \le d}. \tag{231}$$

Moreover, we write the stationary SGD dropout iterate $\breve{\boldsymbol{\beta}}_k^\circ$ in (42) and the ℓ^2-minimizer $\breve{\boldsymbol{\beta}}$ in (34) respectively into

$$\breve{\boldsymbol{\beta}}_k^\circ = (\breve{\beta}_{k1}^\circ, \ldots, \breve{\beta}_{kd}^\circ)^\top \quad \text{and} \quad \breve{\boldsymbol{\beta}} = (\breve{\beta}_1, \ldots, \breve{\beta}_d)^\top. \tag{232}$$

By the short-range dependence of $\{\breve{\boldsymbol{\beta}}_k^\circ\}_{k \in \mathbb{N}}$ in (215), it can be shown that for any $1 \le i, j \le d$, $\mathrm{Cov}(\breve{\beta}_{ki}^\circ, \breve{\beta}_{0j}^\circ) \le c_{ij} \rho_{ij}^k$ for some constants $c_{ij} > 0$ and $0 \le \rho_{ij} < 1$. For the non-overlapping blocks $\{B_m\}_{m \in \mathbb{N}}$, since the positive integers $\{\eta_m\}_{m \in \mathbb{N}}$ satisfy $\eta_{m+1} - \eta_m \to \infty$, it follows that $\eta_{m+1}/\eta_m \to 1$ as $m \to \infty$, and therefore,

$$\sum_{m=1}^{M} (\eta_{m+1} - \eta_m)^2 \asymp \eta_{M+1}(\eta_{M+1} - \eta_M). \tag{233}$$

Then, by Theorems 1(i) and 2(iii) in Xiao and Wu (2011), we obtain, for each $1 \le j \le d$,

$$\begin{aligned} \mathbb{E}(v_{jj,k}^\circ - k\breve{\sigma}_{jj})^2 &\le \Big\{ \big[\mathbb{E}(v_{jj,k}^\circ - \mathbb{E}v_{jj,k}^\circ)^2 \big]^{1/2} + \big[(\mathbb{E}v_{jj,k}^\circ - k\breve{\sigma}_{jj})^2 \big]^{1/2} \Big\}^2 \\ &\lesssim k^{(2/\zeta) \vee (2 - 1/\zeta)}. \end{aligned} \tag{234}$$

For $\mathbb{E}(v^\circ_{ij,k} - k\breve{\sigma}_{ij})^2$ with $i \neq j$, the same rate as in (234) holds. To see this, define two new sequences $\{\beta^+_k\}_{k\in\mathbb{N}}$ and $\{\beta^-_k\}_{k\in\mathbb{N}}$ with $\beta^+_k = (\breve{\beta}^\circ_{ki} + \breve{\beta}^\circ_{0j}) - (\breve{\beta}_i + \breve{\beta}_j)$, and $\beta^-_k = (\breve{\beta}^\circ_{ki} - \breve{\beta}^\circ_{0j}) - (\breve{\beta}_i - \breve{\beta}_j)$. Notice that

$$\begin{aligned}
\breve{\sigma}_{ij} &= \sum_{k=-\infty}^{\infty} \mathbb{E}(\breve{\beta}^\circ_{ki} - \breve{\beta}_i)(\breve{\beta}^\circ_{0j} - \breve{\beta}_j) \\
&= \sum_{k=-\infty}^{\infty} \mathbb{E}\frac{[(\breve{\beta}^\circ_{ki} - \breve{\beta}_i) + (\breve{\beta}^\circ_{0j} - \breve{\beta}_j)]^2 - [(\breve{\beta}^\circ_{ki} - \breve{\beta}_i) - (\breve{\beta}^\circ_{0j} - \breve{\beta}_j)]^2}{4} \\
&= \frac{1}{4}\sum_{k=-\infty}^{\infty} \mathbb{E}(\beta^+_k)^2 - \frac{1}{4}\sum_{k=-\infty}^{\infty} \mathbb{E}(\beta^-_k)^2,
\end{aligned} \tag{235}$$

which can be viewed as the long-run variances of the sequences $\{\beta^+_k\}_{k\in\mathbb{N}}$ and $\{\beta^-_k\}_{k\in\mathbb{N}}$ as indicated by the last line. A similar decomposition can be applied to $v^\circ_{ij,k}$. Since the results in Xiao and Wu (2011) hold for any linear combination of weak-dependent sequences, again by the short-range dependence of $\{\breve{\boldsymbol{\beta}}^\circ_k\}_{k\in\mathbb{N}}$ in (215), we have

$$\mathbb{E}(v^\circ_{ij,k} - k\breve{\sigma}_{ij})^2 \lesssim k^{(2/\zeta)\vee(2-1/\zeta)}, \quad \text{for all } 1 \leq i, j \leq d. \tag{236}$$

For dimension $d \geq 1$, it follows from Lemma 31 that

$$\mathbb{E}\|V^\circ_k - k\breve{\Sigma}\|_F \leq d \max_{1\leq i,j\leq d} \sqrt{\mathbb{E}(v^\circ_{ij,k} - k\breve{\sigma}_{ij})^2} \lesssim dk^{(1/\zeta)\vee(1-1/(2\zeta))}, \tag{237}$$

where the constants in $\lesssim$ are independent of k and d.

Next, we bound $\mathbb{E}\|V_k - V^\circ_k\|_F$. Similar to (237), by Lemma 31, we have

$$\mathbb{E}\|V_k - V^\circ_k\|_F \leq d \max_{1\leq i,j\leq d} \sqrt{\mathbb{E}(v_{ij,k} - v^\circ_{ij,k})^2}. \tag{238}$$

Thus, we only need to show the bound for the one-dimensional case. Now, consider V_k, V°_k, $\mathcal{V}_k$, H_k and $\bar{\boldsymbol{\beta}}^{\mathrm{sgd}}_k$ as scalars. Note that $\mathbb{E}\|\cdot\|_F = (\mathbb{E}[\cdot]^2)^{1/2}$ for $d = 1$. By the decomposition in (58) and applying Jensen's inequality, we have

$$(\mathbb{E}[V_k - V^\circ_k]^2)^{1/2} \leq (\mathbb{E}[\mathcal{V}_k - V^\circ_k]^2)^{1/2} + 2(\mathbb{E}[H_k\bar{\boldsymbol{\beta}}^{\mathrm{sgd}}_k]^2)^{1/2} + K_n(\mathbb{E}[\bar{\boldsymbol{\beta}}^{\mathrm{sgd}}_k]^4)^{1/2}. \tag{239}$$

We shall bound the three term respectively. First, recall the contraction constant $\breve{r}_{\alpha,q}$ defined in (38). By the GMC in Theorem 12 and Hölder's inequality, it follows that

$$\begin{aligned}
&(\mathbb{E}[\mathcal{V}_k - V^\circ_k]^2)^{1/2} \\
&\leq \sum_{m=1}^{\psi(k)-1} \big(\mathbb{E}[\mathcal{S}_m - \mathcal{S}^\circ_m]^4\big)^{1/4}\big(\mathbb{E}[\mathcal{S}_m + \mathcal{S}^\circ_m]^4\big)^{1/4} + \big(\mathbb{E}[\mathcal{R}_k - \mathcal{R}^\circ_k]^4\big)^{1/4}\big(\mathbb{E}[\mathcal{R}_k + \mathcal{R}^\circ_k]^4\big)^{1/4} \\
&\lesssim \sum_{m=1}^{\psi(k)-1} (\breve{r}_{\alpha,q})^{\eta_m}(\eta_{m+1} - \eta_m)^{1/2} + (\breve{r}_{\alpha,q})^{\eta_{\psi(k)}}(k - \eta_{\psi(k)} + 1)^{1/2} \\
&= O(1),
\end{aligned} \tag{240}$$

where the constants in $\lesssim$ and $O(\cdot)$ are independent of k. Similarly, since the dimension d is fixed and $\mathbb{E}[\breve{\boldsymbol{\beta}}_k^\circ] = \breve{\boldsymbol{\beta}}$ by (220) with the ℓ^2-minimizer $\breve{\boldsymbol{\beta}}$ defined in (34), we can show that

$$(\mathbb{E}[\bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}]^4)^{1/2} \lesssim \Big(\mathbb{E}\Big[\frac{1}{k}\sum_{i=1}^k \breve{\boldsymbol{\beta}}_i^\circ\Big]^4\Big)^{1/2} + \Big(\mathbb{E}\Big[\frac{1}{k}\sum_{i=1}^k (\breve{\boldsymbol{\beta}}_i - \breve{\boldsymbol{\beta}}_i^\circ)\Big]^4\Big)^{1/2} \asymp \frac{1}{k}, \tag{241}$$

and $(\mathbb{E}[H_k]^4)^{1/2} \lesssim k(k-\eta_{\psi(k)})^2$. Combining these results with the fact that $K_k \asymp k(k-\eta_{\psi(k)})$ yields

$$K_k(\mathbb{E}[\bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}]^4)^{1/2} \asymp k - \eta_{\psi(k)} \quad \text{and} \quad (\mathbb{E}[H_k \bar{\boldsymbol{\beta}}_k^{\mathrm{sgd}}]^2)^{1/2} \lesssim k - \eta_{\psi(k)}. \tag{242}$$

Inserting all these expressions back into (239), we obtain

$$(\mathbb{E}[V_k - V_k^\circ]^2)^{1/2} \lesssim k - \eta_{\psi(k)} \asymp k^{1-(1/\zeta)}. \tag{243}$$

Consequently, by (238), for the multi-dimensional case, it follows from (238) that

$$\mathbb{E}\|V_k - V_k^\circ\|_F \le d \max_{1\le i,j\le d} \sqrt{\mathbb{E}(v_{ij,k} - v_{ij,k}^\circ)^2} \lesssim dk^{1-1/(2\zeta)}.$$

Since $\zeta > 1$, compared to the rate of $\mathbb{E}\|V_k^\circ - k\breve{\Sigma}\|_F \lesssim dk^{(1/\zeta)\vee(1-1/(2\zeta))}$ in (237), the latter dominates. This, along with (230) gives the desired result $\mathbb{E}\|\hat{\Sigma}_k(\alpha) - \breve{\Sigma}(\alpha)\|_F = O(dk^{(1/\zeta)\vee(1-1/(2\zeta))})$.

■

G.2 Long-Run Covariances for Gaussian Random Samples

Consider i.i.d. covariate vectors $\boldsymbol{x}_k \sim \mathcal{N}(0, I_d)$, $k = 1, 2, \ldots$, and the responses $y_k \mid \boldsymbol{x}_k$ from the linear regression model

$$y_k = \boldsymbol{x}_k^\top \boldsymbol{\beta}^* + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, 1). \tag{244}$$

For the SGD iterates with dropout

$$\breve{\boldsymbol{\beta}}_k(\alpha) - \breve{\boldsymbol{\beta}} = \underbrace{\big(I_d - \alpha D_k \mathbb{X}_k D_k\big)}_{=:\breve{A}_k(\alpha)}(\breve{\boldsymbol{\beta}}_{k-1}(\alpha) - \breve{\boldsymbol{\beta}}) + \underbrace{\alpha D_k \boldsymbol{x}_k (y_k - \boldsymbol{x}_k^\top D_k \breve{\boldsymbol{\beta}})}_{=:\breve{\boldsymbol{b}}_k(\alpha)},$$

defined in (36), and the minimizer of the ℓ^2-regularized least-squares loss

$$\breve{\boldsymbol{\beta}} = p\Big(p^2\mathbb{E}[\boldsymbol{x}_1\boldsymbol{x}_1^\top] + p(1-p)\mathrm{Diag}(\mathbb{E}[\boldsymbol{x}_1\boldsymbol{x}_1^\top])\Big)^{-1}\mathbb{E}[y_1\boldsymbol{x}_1] = (\mathbb{E}[\mathbb{X}_{1,p}])^{-1}\mathbb{E}[y_1\boldsymbol{x}_1],$$

defined in (34), we have

$$\breve{\boldsymbol{\beta}} = \boldsymbol{\beta}^*. \tag{245}$$

This result holds since $\mathbb{E}[\mathbb{X}_{k,p}] = \mathbb{E}[D_k\boldsymbol{x}_k\boldsymbol{x}_k^\top D_k] = pI_d = \mathbb{E}[\mathbb{X}_k]$. Recall the long-run covariance matrix

$$\breve{\Sigma}(\alpha) := \sum_{i=-\infty}^{\infty} \mathbb{E}[(\breve{\boldsymbol{\beta}}_0^\circ(\alpha) - \breve{\boldsymbol{\beta}})(\breve{\boldsymbol{\beta}}_i^\circ(\alpha) - \breve{\boldsymbol{\beta}})^\top],$$

defined in the quenched CLT in Theorem 15, where $\breve{\boldsymbol{\beta}}_i^\circ(\alpha) \sim \breve{\pi}_\alpha$ denotes the stationary SGD process with dropout. We can derive a closed form expression of the limit

$$\breve{\Sigma}^* := \lim_{\alpha \to 0} \breve{\Sigma}(\alpha). \tag{246}$$

Lemma 32 (Long-run covariance matrices for Gaussian random samples) *For the model* (244) *with Gaussian random samples, suppose that conditions in Theorem 18 are satisfied. Then,* $\breve{\Sigma}^*$ *is diagonal, with the* i*-th diagonal element*

$$\breve{\Sigma}_{ii}(\alpha) \to \breve{\Sigma}_{ii}^*, \quad \text{as } \alpha \to 0, \tag{247}$$

for all $i = 1, \ldots, d$*, where we write* $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_d^*)^\top$ *and*

$$\breve{\Sigma}_{ii}^* = \frac{1}{p} + \frac{1-p}{p} \sum_{j \neq i} \beta_j^{*2}. \tag{248}$$

In particular, $\breve{\Sigma}^* = I_d$ *if* $p = 1$.

Proof The proof consists of two steps. In **Step 1**, we construct an affine sequence with long-run covariance matrix $\breve{\Sigma}$ and prove that with i.i.d. Gaussian random samples $(y_k, \boldsymbol{x}_k)$, $\breve{\Sigma}^\dagger$ is diagonal, and the i-th diagonal element $\breve{\Sigma}_{ii}^\dagger(\alpha) = \Sigma_{ii}^*$ for each i. Then in **Step 2**, we bound the error of this affine approximation by showing that $\|\breve{\Sigma}(\alpha) - \breve{\Sigma}^\dagger\| \to 0$ as $\alpha \to 0$.

Step 1. Define the affine approximation to the SGD dropout sequence $\{\breve{\boldsymbol{\beta}}_k(\alpha)\}_{k \in \mathbb{N}}$ as $\{\breve{\boldsymbol{\beta}}_k^\dagger(\alpha)\}_{k \in \mathbb{N}}$, which follows the recursion

$$\breve{\boldsymbol{\beta}}_k^\dagger(\alpha) - \breve{\boldsymbol{\beta}} = (I_d - \alpha H_p)(\breve{\boldsymbol{\beta}}_{k-1}^\dagger(\alpha) - \breve{\boldsymbol{\beta}}) + \breve{\boldsymbol{b}}_k(\alpha), \tag{249}$$

where

$$H_p = \mathbb{E}[D_1 \mathbb{X}_1 D_1], \quad \breve{\boldsymbol{b}}_k(\alpha) = \alpha D_k \boldsymbol{x}_k (y_k - \boldsymbol{x}_k^\top D_k \breve{\boldsymbol{\beta}}). \tag{250}$$

Recall the contraction constant $\breve{r}_{\alpha,q}$ defined in (38). Let $q = 2$. Since $\|\cdot\|_2^2$ is convex, it follows from Jensen's inequality that

$$\mathbb{E}\|(I_d - \alpha D \mathbb{X}_1 D)\boldsymbol{v}\|_2^2 \geq \big\|\mathbb{E}\big[(I_d - \alpha D \mathbb{X}_1 D)\boldsymbol{v}\big]\big\|_2^2 = \|(I_d - \alpha H_p)\boldsymbol{v}\|_2^2. \tag{251}$$

Since the matrix $I - \alpha H_p$ is symmetric, we therefore obtain

$$\breve{r}_{\alpha,2}^2 = \sup_{\|\boldsymbol{v}\|_2 = 1} \mathbb{E}\|(I_d - \alpha D \mathbb{X}_1 D)\boldsymbol{v}\|_2^2 \geq \sup_{\|\boldsymbol{v}\|_2 = 1} \|(I_d - \alpha H_p)\boldsymbol{v}\|_2^2 = \|I_d - \alpha H_p\|^2. \tag{252}$$

By Theorem 12, since the constant learning rate α satisfies condition (37), we have $\|I_d - \alpha H_p\| \leq \breve{r}_{\alpha,2} < 1$. Hence, GMC also holds for the affine sequence $\{\breve{\boldsymbol{\beta}}_k^\dagger(\alpha)\}_{k \in \mathbb{N}}$ and $\breve{\boldsymbol{\beta}}_k^\dagger(\alpha)$ converges to a unique stationary distribution as $k \to \infty$. Throughout this proof, we assume that the initialization $\breve{\boldsymbol{\beta}}_0^\dagger$ follows this stationary distribution, and thus the sequence $\{\breve{\boldsymbol{\beta}}_k^\dagger(\alpha)\}_{k \in \mathbb{N}}$ is stationary.

Since $\{\breve{\boldsymbol{\beta}}_k^\dagger(\alpha)\}_{k\in\mathbb{N}}$ is a stationary sequence, define its long-run covariance matrix as

$$\breve{\Sigma}^\dagger(\alpha) := \sum_{i=-\infty}^{\infty} \mathbb{E}[(\breve{\boldsymbol{\beta}}_0^\dagger(\alpha) - \breve{\boldsymbol{\beta}})(\breve{\boldsymbol{\beta}}_i^\dagger(\alpha) - \breve{\boldsymbol{\beta}})^\top].$$

We denote the i-th diagonal element of $\breve{\Sigma}^\dagger(\alpha)$ by $\breve{\Sigma}_{ii}(\alpha)$, $i = 1, \ldots, d$. Next, we show that

$$\breve{\Sigma}_{ii}^\dagger(\alpha) = \breve{\Sigma}_{ii}^*, \quad \text{for all } i. \tag{253}$$

Recall that $\breve{\boldsymbol{b}}_k(\alpha)$ are i.i.d. and by (193), $\mathbb{E}[\breve{\boldsymbol{b}}_k(\alpha)] = 0$. Thus, recursion (249) yields a VAR(1) process. By iterating (249), we obtain

$$\breve{\boldsymbol{\beta}}_k^\dagger(\alpha) - \breve{\boldsymbol{\beta}} = \sum_{j=0}^{\infty} (I_d - \alpha H_p)^j \breve{\boldsymbol{b}}_{k-j}(\alpha), \tag{254}$$

which converges since $\|I_d - \alpha H_p\| < 1$. Define the covariance matrix

$$Q_p(\alpha) := \mathbb{E}[\breve{\boldsymbol{b}}_1(\alpha)\breve{\boldsymbol{b}}_1(\alpha)^\top]. \tag{255}$$

Since $\breve{\boldsymbol{b}}_{k-j}$ are independent over j, it follows that

$$\begin{aligned}
\breve{\Sigma}^\dagger(\alpha) &= \sum_{j=0}^{\infty} (I_d - \alpha H_p)^j Q_p(\alpha)[(I_d - \alpha H_p)^j]^\top \\
&= \big[I_d - (I_d - \alpha H_p)\big]^{-1} Q_p(\alpha) \big[I_d - (I_d - \alpha H_p)\big]^{-1} \\
&= (\alpha H_p)^{-1} Q_p(\alpha) (\alpha H_p)^{-1},
\end{aligned} \tag{256}$$

where the sandwich form in the second equation follows from the the spectral density of $\breve{\boldsymbol{\beta}}_k^\dagger(\alpha) - \breve{\boldsymbol{\beta}}$ at frequency 0 (see Brockwell and Davis (1991), Chapter 11).

Now we compute H_p and $Q_p(\alpha)$ for i.i.d. Gaussian random samples. Since $\boldsymbol{x}_k \sim \mathcal{N}(0, I_d)$,

$$H_p = \mathbb{E}[D_1 \mathbb{X}_1 D_1] = p I_d, \tag{257}$$

which yields

$$\breve{\Sigma}^\dagger(\alpha) = (\alpha p)^{-2} Q_p(\alpha). \tag{258}$$

Moreover, by (245), $\breve{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$, and thus $y_k = \boldsymbol{x}_k^\top \breve{\boldsymbol{\beta}} + \epsilon_k$ for all $k \in \mathbb{N}_+$, which leads to

$$\begin{aligned}
Q_p(\alpha) &= \mathbb{E}[\breve{\boldsymbol{b}}_1(\alpha)\breve{\boldsymbol{b}}_1(\alpha)^\top] \\
&= \alpha^2 \mathbb{E}\big([D_1 \boldsymbol{x}_1 (y_1 - \boldsymbol{x}_1^\top D_1 \breve{\boldsymbol{\beta}})][D_1 \boldsymbol{x}_1 (y_1 - \boldsymbol{x}_1^\top D_1 \breve{\boldsymbol{\beta}})]^\top\big) \\
&= \alpha^2 \mathbb{E}\big[D_1 \mathbb{X}_1 D_1 (\epsilon_1 + \boldsymbol{x}_1^\top (I_d - D_1)\breve{\boldsymbol{\beta}})^2\big].
\end{aligned} \tag{259}$$

Expanding the square gives

$$(\epsilon_1 + \boldsymbol{x}_1^\top (I_d - D_1)\breve{\boldsymbol{\beta}})^2 = (\boldsymbol{x}_1^\top (I_d - D_1)\breve{\boldsymbol{\beta}})^2 + 2\epsilon_1 \boldsymbol{x}_1^\top (I_d - D_1)\breve{\boldsymbol{\beta}} + \epsilon_1^2.$$

We consider the three terms separately. Since ϵ_k is independent of $\boldsymbol{x}_k$ and D_k,

$$\mathbb{E}_{(y,\boldsymbol{x})}\big[D_1\mathbb{X}_1D_1\epsilon_1\boldsymbol{x}_1^\top(I_d-D_1)\breve{\boldsymbol{\beta}}\mid D_1\big]=0. \tag{260}$$

By $\mathbb{E}[\epsilon_k^2]=1$ and $D_1^2=D_1$, we have

$$\mathbb{E}_{(y,\boldsymbol{x})}\big[D_1\mathbb{X}_1D_1\epsilon_1^2\mid D_1\big]=D_1. \tag{261}$$

In addition, since $\boldsymbol{x}_k\sim\mathcal{N}(0,I_d)$,

$$\begin{aligned}
&\mathbb{E}_{(y,\boldsymbol{x})}[D_1\mathbb{X}_1D_1(\boldsymbol{x}_1^\top(I_d-D_1)\breve{\boldsymbol{\beta}})^2\mid D_1]\\
&=D_1\mathbb{E}_{(y,\boldsymbol{x})}\big[\|(I_d-D_1)\breve{\boldsymbol{\beta}}\|_2^2I_d+2(I_d-D_1)\breve{\boldsymbol{\beta}}\breve{\boldsymbol{\beta}}^\top(I_d-D_1)\big]D_1\\
&=D_1\mathbb{E}_{(y,\boldsymbol{x})}\big[\|(I_d-D_1)\breve{\boldsymbol{\beta}}\|_2^2I_d\big],
\end{aligned} \tag{262}$$

where the last equation holds due to $D_1(I_d-D_1)=0$. Combining all the three equations above, we obtain

$$\mathbb{E}_{(y,\boldsymbol{x})}[\breve{\boldsymbol{b}}_1(\alpha)\breve{\boldsymbol{b}}_1(\alpha)^\top\mid D_1]=\alpha^2\big(1+\|(I_d-D_1)\breve{\boldsymbol{\beta}}\|_2^2\big)D_1. \tag{263}$$

Taking expectation over D gives the diagonal matrix $Q_p(\alpha)$ with entries

$$Q_{p,ii}(\alpha)=\alpha^2\Big(p+p(1-p)\sum_{j\neq i}\breve{\beta}_j^2\Big), \tag{264}$$

and $Q_{p,ij}(\alpha)=0$ for $i\neq j$. Inserting this back into (258) yields, for all i,

$$\breve{\Sigma}_{ii}^\dagger(\alpha)=\frac{1}{p}+\frac{1-p}{p}\sum_{j\neq i}\beta_j^{*2}=\breve{\Sigma}_{ii}^*, \tag{265}$$

which is independent of α. Write $\breve{\Sigma}^\dagger(\alpha)=\breve{\Sigma}$.

Step 2. Finally, we complete the proof by showing that

$$\|\breve{\Sigma}(\alpha)-\breve{\Sigma}^\dagger\|_F\to 0,\quad \text{as } \alpha\to 0. \tag{266}$$

To this end, consider the difference

$$\breve{\boldsymbol{\delta}}_k(\alpha)=\breve{\boldsymbol{\beta}}_k^\circ(\alpha)-\breve{\boldsymbol{\beta}}_k^\dagger(\alpha)=\breve{A}_k(\alpha)\breve{\boldsymbol{\delta}}_{k-1}(\alpha)-\alpha(D_k\mathbb{X}_kD_k-H_p)[\breve{\boldsymbol{\beta}}_{k-1}^\dagger(\alpha)-\breve{\boldsymbol{\beta}}]. \tag{267}$$

Since both initializations $\breve{\boldsymbol{\beta}}_0^\circ$ and $\breve{\boldsymbol{\beta}}_0^\dagger$ follow the stationary distributions, respectively, the difference sequence $\{\breve{\boldsymbol{\delta}}_k(\alpha)\}_{k\in\mathbb{N}}$ is stationary. By Lemma 10 and the independence between $(D_k,\boldsymbol{x}_k)$ and $\breve{\boldsymbol{\beta}}_{k-1}^\dagger(\alpha)$, we obtain, for any $\eta>0$,

$$\begin{aligned}
\mathbb{E}\|\breve{\boldsymbol{\delta}}_k(\alpha)\|_2^2&\le(1+\eta)\mathbb{E}\|\breve{A}_k(\alpha)\breve{\boldsymbol{\delta}}_{k-1}(\alpha)\|_2^2+\Big(1+\frac{1}{\eta}\Big)\alpha^2\mathbb{E}\|(D_k\mathbb{X}_kD_k-H_p)[\breve{\boldsymbol{\beta}}_{k-1}^\dagger(\alpha)-\breve{\boldsymbol{\beta}}]\|_2^2\\
&\le(1+\eta)\breve{r}_{\alpha,2}^2\mathbb{E}\|\breve{\boldsymbol{\delta}}_{k-1}(\alpha)\|_2^2\\
&\quad+\Big(1+\frac{1}{\eta}\Big)\alpha^2\sup_{\boldsymbol{v}\in\mathbb{R}^d,\,\|\boldsymbol{v}\|_2=1}\mathbb{E}\|(D_k\mathbb{X}_kD_k-H_p)\boldsymbol{v}\|_2^2\cdot\mathbb{E}\|\breve{\boldsymbol{\beta}}_{k-1}^\dagger(\alpha)-\breve{\boldsymbol{\beta}}\|_2^2.
\end{aligned} \tag{268}$$

By Assumption 2, there exists some finite constant $c > 0$ such that

$$\sup_{\boldsymbol{v}\in\mathbb{R}^d,\, \|\boldsymbol{v}\|_2=1} \mathbb{E}\|(D_k\mathbb{X}_k D_k - H_p)\boldsymbol{v}\|_2^2 \le c. \tag{269}$$

Moreover, by similar arguments as in the proof of Lemma 14, one can show that

$$\mathbb{E}\|\breve{\boldsymbol{\beta}}_{k-1}^{\dagger}(\alpha) - \breve{\boldsymbol{\beta}}\|_2^2 = O(\alpha). \tag{270}$$

Since $\breve{r}_{\alpha,2} < 1$, by choosing η such that $(1+\eta)\breve{r}_{\alpha,2}^2 < 1$, it follows from the stationarity of $\{\breve{\boldsymbol{\delta}}_k(\alpha)\}_{k\in\mathbb{N}}$ that

$$\mathbb{E}\|\breve{\boldsymbol{\delta}}_k(\alpha)\|_2^2 = O(\alpha^3). \tag{271}$$

For brevity, denote

$$\boldsymbol{u}_k = \boldsymbol{u}_k(\alpha) = \breve{\boldsymbol{\beta}}_k^{\circ}(\alpha) - \breve{\boldsymbol{\beta}}, \quad \boldsymbol{u}_k^{\dagger} = \boldsymbol{u}_k^{\dagger}(\alpha) = \breve{\boldsymbol{\beta}}_k^{\dagger}(\alpha) - \breve{\boldsymbol{\beta}}. \tag{272}$$

For any lag $h \in \mathbb{Z}$, the difference of the two covariance matrices can be decomposed into

$$\mathbb{E}[\boldsymbol{u}_0\boldsymbol{u}_h^{\top}] - \mathbb{E}[\boldsymbol{u}_0^{\dagger}\boldsymbol{u}_h^{\dagger\top}] = \mathbb{E}[\breve{\boldsymbol{\delta}}_0\boldsymbol{u}_h^{\top}] + \mathbb{E}[\boldsymbol{u}_0^{\dagger}\breve{\boldsymbol{\delta}}_h^{\top}]. \tag{273}$$

We take the Frobenius norm and it follows from Cauchy-Schwarz inequality, Lemma 14 and expression (271) that

$$\|\mathbb{E}[\breve{\boldsymbol{\delta}}_0\boldsymbol{u}_h^{\top}]\|_F \le (\mathbb{E}\|\breve{\boldsymbol{\delta}}_0\|_2^2)^{1/2}(\mathbb{E}\|\boldsymbol{u}_h\|_2^2)^{1/2} = O(\alpha^{3/2})\cdot O(\alpha^{1/2}) = O(\alpha^2). \tag{274}$$

Similarly, $\|\mathbb{E}[\boldsymbol{u}_0^{\dagger}\breve{\boldsymbol{\delta}}_h^{\top}]\|_F = O(\alpha^2)$, which together with the expression above yields

$$\|\mathbb{E}[\boldsymbol{u}_0\boldsymbol{u}_h^{\top}] - \mathbb{E}[\boldsymbol{u}_0^{\dagger}\boldsymbol{u}_h^{\dagger\top}]\|_F = O(\alpha^2). \tag{275}$$

For any truncation level $H \in \mathbb{N}$,

$$\begin{aligned}
\|\breve{\Sigma}(\alpha) - \breve{\Sigma}^{\dagger}\|_F &= \Big\| \sum_{h=-\infty}^{\infty} \big(\mathbb{E}[\boldsymbol{u}_0\boldsymbol{u}_h^{\top}] - \mathbb{E}[\boldsymbol{u}_0^{\dagger}\boldsymbol{u}_h^{\dagger\top}]\big)\Big\|_F \\
&\le \Big\| \sum_{|h|\le H} \big(\mathbb{E}[\boldsymbol{u}_0\boldsymbol{u}_h^{\top}] - \mathbb{E}[\boldsymbol{u}_0^{\dagger}\boldsymbol{u}_h^{\dagger\top}]\big)\Big\|_F + \Big\| \sum_{|h|> H} \big(\mathbb{E}[\boldsymbol{u}_0\boldsymbol{u}_h^{\top}] - \mathbb{E}[\boldsymbol{u}_0^{\dagger}\boldsymbol{u}_h^{\dagger\top}]\big)\Big\|_F \\
&\le O(H\alpha^2) + \sum_{|h|>H} \|\mathbb{E}[\boldsymbol{u}_0\boldsymbol{u}_h^{\top}]\|_F + \sum_{|h|>H} \|\mathbb{E}[\boldsymbol{u}_0^{\dagger}\boldsymbol{u}_h^{\dagger\top}]\|_F.
\end{aligned} \tag{276}$$

Recall the functional dependence measure in (209). By GMC of $\boldsymbol{u}_h$, the coupled distance also contracts, which gives

$$\breve{\theta}_{k,2}(\alpha) \le \breve{r}_{\alpha,2}^k \breve{\theta}_{0,2}(\alpha) \le 2\breve{r}_{\alpha,2}^k (\mathbb{E}\|\boldsymbol{u}_0\|_2^2)^{1/2} = 2\breve{r}_{\alpha,2}^k O(\sqrt{\alpha}). \tag{277}$$

Recall the projection operator $\breve{\mathcal{P}}_k[\cdot]$ in (206). By the martingale decomposition and the orthogonality of projections, for $h > 0$,

$$\begin{aligned}
\mathbb{E}[\boldsymbol{u}_0\boldsymbol{u}_h^{\top}] &= \mathbb{E}\Big[\Big(\sum_{i=0}^{\infty} \breve{\mathcal{P}}_{-i}[\boldsymbol{u}_0]\Big)\Big(\sum_{j=0}^{\infty} \breve{\mathcal{P}}_{h-j}[\boldsymbol{u}_h]\Big)^{\top}\Big] \\
&= \sum_{i=0}^{\infty} \mathbb{E}\Big[\breve{\mathcal{P}}_{-i}[\boldsymbol{u}_0]\Big(\breve{\mathcal{P}}_{-i}[\boldsymbol{u}_h]\Big)^{\top}\Big].
\end{aligned} \tag{278}$$

By the triangle inequality, Cauchy-Schwarz inequality, and Theorem 1 in Wu (2005),

$$\begin{aligned}\|\mathbb{E}[\boldsymbol{u}_0\boldsymbol{u}_h^\top]\|_F &\le \sum_{i=0}^{\infty}\Big\|\mathbb{E}\Big[\breve{\mathcal{P}}_{-i}[\boldsymbol{u}_0]\Big(\breve{\mathcal{P}}_{-i}[\boldsymbol{u}_h]\Big)^\top\Big]\Big\|_F \\ &\le \sum_{i=0}^{\infty}\big[\|\mathbb{E}\big(\breve{\mathcal{P}}_{-i}[\boldsymbol{u}_0]\big)\|_2^2\big]^{1/2}\big[\|\mathbb{E}\big(\breve{\mathcal{P}}_{-i}[\boldsymbol{u}_h]\big)\|_2^2\big]^{1/2} \\ &\le \sum_{j=0}^{\infty}\breve{\theta}_{j,2}(\alpha)\breve{\theta}_{h+j,2}(\alpha). \end{aligned} \tag{279}$$

As a direct consequence,

$$\sum_{h>H}\|\mathbb{E}[\boldsymbol{u}_0\boldsymbol{u}_h^\top]\|_F \le \sum_{h>H}\sum_{j=0}^{\infty}\breve{\theta}_{j,2}(\alpha)\breve{\theta}_{h+j,2}(\alpha) = O(\alpha\breve{r}_{\alpha,2}^H). \tag{280}$$

The bounds for the parts $\sum_{h<-H}\|\mathbb{E}[\boldsymbol{u}_0\boldsymbol{u}_h^\top]\|_F$ and $\sum_{|h|>H}\|\mathbb{E}[\boldsymbol{u}_0^\dagger\boldsymbol{u}_h^{\dagger\top}]\|_F$ can follow similar arguments. Choose $H = \lceil c''\log(1/\alpha)\rceil$ with some constant $c'' > 1/|\log(\breve{r}_{\alpha,2})|$. Then, we have $\alpha\breve{r}_{\alpha,2}^H \le \alpha^{c''|\log(r_{\alpha,2})|+1} = o(\alpha)$ for the tail part in (276), and $H\alpha^2 = c''\alpha^2\log(1/\alpha)$. Both converge to 0 as $\alpha\to 0$. This completes the proof. ■

Appendix H. Real-Data Application

The *Online News Popularity* dataset from UCI ML Repository (Fernandes et al., 2015) summarizes a set of features about articles published by *Mashable* (`www.mashable.com`) in a period of two years. This dataset contains 39,797 samples and 58 features. We apply the efficient ASGD-dropout algorithm to predict the number of *shares* in social networks (popularity) and use our proposed online inference methodology to test the significance of features towards prediction.

To standardize the dataset, we center all the predictors and scale them to unit variance. The response is transformed as $\log(1+\text{shares})$ and centered. This ensures comparable gradient magnitudes across features and avoids a few heavy-tailed article outliers dominating the updates. We adopt $p = 0.9$ and the dropout rate is $1-p = 0.1$. Dropout regularization is not strictly necessary for classical linear regression, but it demonstrates how the ASGD framework can incorporate randomized feature masking, especially when scaling to large-dimensional data. The online nature of the proposed inference methodology means we never store the full design matrix or compute costly matrix inversions, useful for streaming or memory-constrained environments.

Figures 5 and 6 plot the ASGD-dropout trajectories over $n = 39,797$ iterations, showing rapid stabilization of estimated regression coefficients. By estimating the long-run variances of these coefficients and constructing marginal 95% CIs using expression (60), at the last iteration, any feature whose CI does not cover zero can be deemed statistically significant. Prominent positive predictors include *number of videos*, *number of links*, *publication at weekends* and *global text subjectivity*. Primary negative predictors include *mid-week publication* and *average length of the words in the content*. Interestingly, for the five Latent Dirichlet

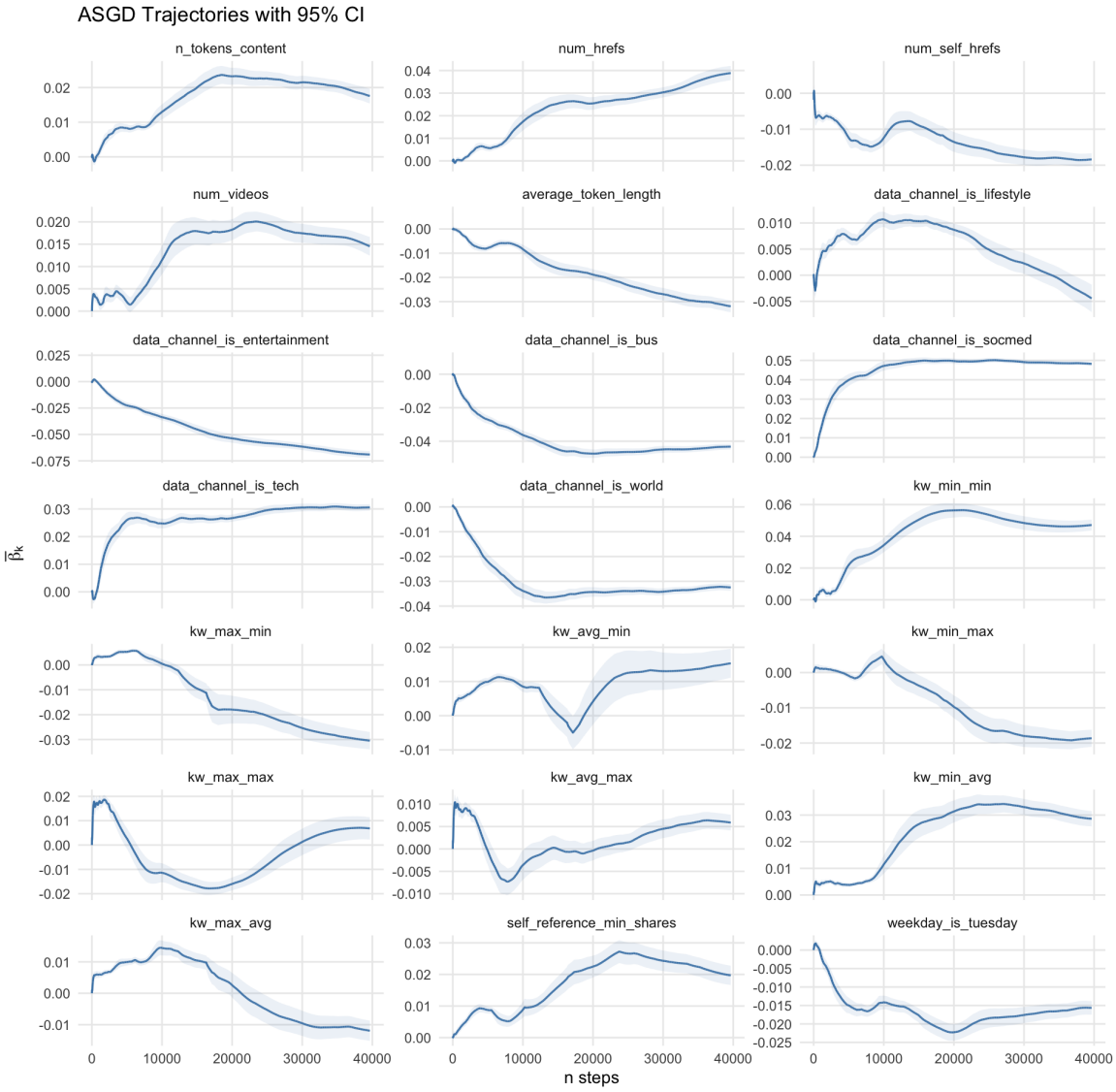


Figure 5: Convergence trace of 95% CI of ASGD iterates with dropout regularization for significant features. Each subplot represents the convergence trace of one feature.

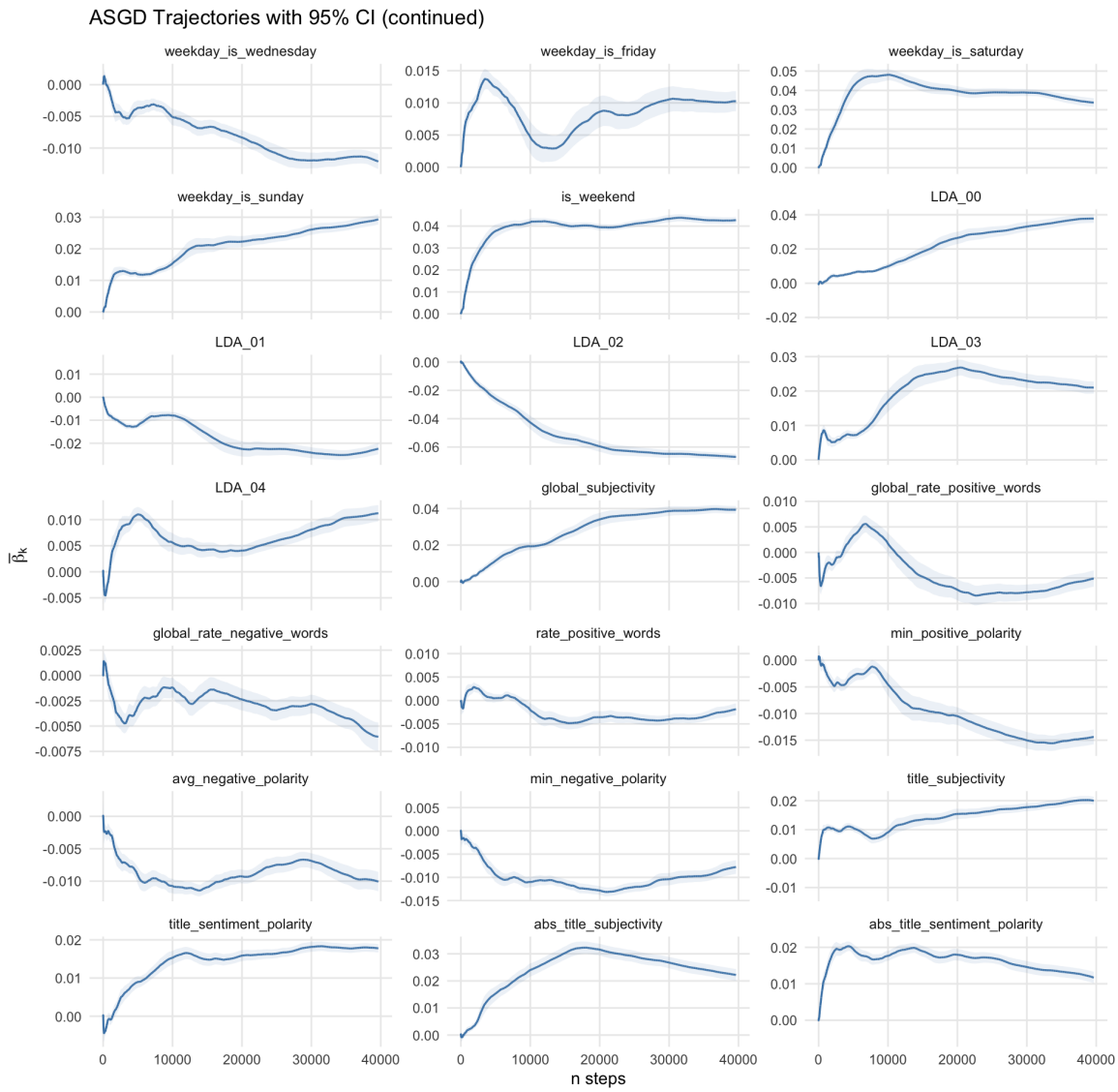


Figure 6: (Continued) Convergence trace of 95% CI of ASGD iterates with dropout regularization for significant features. Each subplot represents the convergence trace of one feature.

Allocation (LDA) topics, LDA topics 0, 3 and 4 have positive coefficients whose 95% CIs lie entirely above zero, while LDA topics 1 and 2 have coefficients whose CIs lie entirely below zero. We refer to Figures 5 and 6 for the detailed significant features, where we plot the trajectory of ASGD-dropout (solid line) with its 95% CI ribbon for the significant features.

Our findings of significance features echo the literature. For example, Ciampaglia et al. (2018) found that certain latent topics (e.g. technology, social media) boost sharing while others (e.g. politics, hard news) suppress it, matching our positive significance for certain LDA components and negative for others. Szabo and Huberman (2010) documented a "weekend effect" in which articles published on weekends tend to garner more attention, consistent with the positive coefficient on weekend publication. Moreover, Bandari et al. (2021) demonstrated that multimedia richness (links and video counts) are among the most correlated variables with news popularity.