Journal of Machine Learning Research 25 (2024) 1-48 Submitted 6/23; Revised 4/24; Published 7/24

Dropout Regularization Versus ℓ_2-Penalization in the Linear Model

Gabriel Clara G.CLARA@UTWENTE.NL
Sophie Langer S.LANGER@UTWENTE.NL
Johannes Schmidt-Hieber A.J.SCHMIDT-HIEBER@UTWENTE.NL
Faculty of Electrical Engineering, Mathematics, and Computer Science
University of Twente
7522 NB, Enschede, The Netherlands

Editor: Samory Kpotufe

Abstract

We investigate the statistical behavior of gradient descent iterates with dropout in the linear regression model. In particular, non-asymptotic bounds for the convergence of expectations and covariance matrices of the iterates are derived. The results shed more light on the widely cited connection between dropout and ℓ_2-regularization in the linear model. We indicate a more subtle relationship, owing to interactions between the gradient descent dynamics and the additional randomness induced by dropout. Further, we study a simplified variant of dropout which does not have a regularizing effect and converges to the least squares estimator.

Keywords: dropout, algorithmic regularization, gradient descent

1. Introduction

Dropout is a simple, yet effective, algorithmic regularization technique, intended to prevent neural networks from overfitting. First introduced in Srivastava et al. (2014), the method is implemented via random masking of neurons at training time. Specifically, during every gradient descent iteration, the output of each neuron is replaced by zero based on the outcome of an independently sampled $\mathrm{Ber}(p)$-distributed variable. This temporarily removes each neuron with a probability of $1-p$, see Figure 1 for an illustration. The method has demonstrated effectiveness in various applications, see for example Krizhevsky et al. (2012); Srivastava et al. (2014). On the theoretical side, dropout is often studied by exhibiting connections with explicit regularizers (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). Rather than analyzing the gradient descent iterates with dropout noise, these results consider the marginalized training loss with marginalization over the dropout noise. Within this framework, Srivastava et al. (2014) established a connection between dropout and weighted ℓ_2-penalization in the linear regression model. This connection is now cited in popular textbooks (Efron and Hastie, 2016; Goodfellow et al., 2016).

However, Wei et al. (2020) show empirically that injecting dropout noise in the gradient descent iterates also induces an implicit regularization effect that is not captured by the link

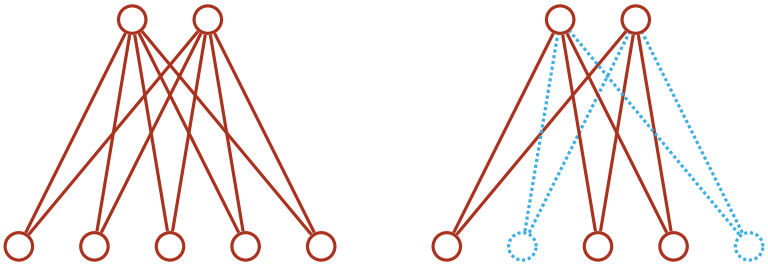

Figure 1: Regular neurons (left) with all connections active. Sample of the same neurons with dropout (right). The dashed connections are ignored during the current iteration of training.

between the marginalized loss and explicit regularization. This motivates our approach to directly derive the statistical properties of gradient descent iterates with dropout. We study the linear regression model due to mathematical tractability and because the minimizer of the explicit regularizer is unique and admits a closed-form expression. In line with the implicit regularization observed in Wei et al. (2020), our main result provides a theoretical bound quantifying the amount of randomness in the gradient descent scheme that is ignored by the previously considered minimizers of the marginalized training loss. More specifically, Theorem 5 shows that for a fixed learning rate there is additional randomness which fails to vanish in the limit, while Theorem 7 characterizes the gap between dropout and ℓ_2-penalization with respect to the learning rate, the dropout parameter p, the design matrix, and the distribution of the initial iterate. Theorem 8 shows that this gap disappears for the Ruppert-Polyak averages of the iterates.

To provide a clearer understanding of the interplay between gradient descent and variance, we also investigate a simplified variant of dropout featuring more straightforward interactions between the two. Applying the same analytical techniques to this simplified variant, Theorem 10 establishes convergence in quadratic mean to the conventional linear least-squares estimator. This analysis illustrates the sensitivity of gradient descent to small changes in the way noise is injected during training.

Many randomized optimization methods can be formulated as noisy gradient descent schemes. The developed strategy to treat gradient descent with dropout may be generalized to other settings. An example is the recent analysis of forward gradient descent in Bos and Schmidt-Hieber (2023).

The article is organized as follows. After discussing related results below, Section 2 contains preliminaries and introduces two different variants of dropout. Section 3 discusses some extensions of previous results for averaged dropout obtained by marginalizing over the dropout distribution in the linear model considered in Srivastava et al. (2014). Section 4 illustrates the main results on gradient descent with dropout in the linear model, and examines its statistical optimality. Section 5 contains further discussion and mentions a number of natural follow-up problems. All proofs are deferred to the Appendix.

1.1 Other Related Work

Considering linear regression and the marginalized training loss with marginalization over the dropout noise, the initial dropout article (Srivastava et al., 2014) already connects dropout with ℓ_2-regularization. This connection was also noted by Baldi and Sadowski (2013) and by McAllester (2013). As this argument is crucial in our own analysis, we will discuss it in more detail in Section 3.

Wager et al. (2013) extends the reasoning to generalized linear models and more general forms of injected noise. Employing a quadratic approximation to the loss function after marginalization over the injected noise, the authors exhibit an explicit regularizer. In case of dropout noise, this regularizer induces in first-order an ℓ_2-penalty after rescaling of the data by the estimated inverse of the diagonal Fisher information.

For two-layer models, marginalizing the dropout noise leads to a nuclear norm penalty on the product matrix, both in matrix factorization (Cavazza et al., 2018) and linear neural networks (Mianjy et al., 2018). The latter may be seen as a special case of a particular "ℓ_2-path regularizer", which appears in deep linear networks (Mianjy and Arora, 2019) and shallow ReLU-activated networks (Arora et al., 2021). Further, Arora et al. (2021) exhibit a data distribution-dependent regularizer in two-layer matrix sensing/completion problems. This regularizer collapses to a nuclear norm penalty for specific distributions.

Gal and Ghahramani (2016a) show that empirical risk minimization in deep neural networks with dropout may be recast as performing Bayesian variational inference to approximate the intractable posterior resulting from a deep Gaussian process prior. The Bayesian viewpoint also allows for the quantification of uncertainty. Gal and Ghahramani (2016b) further generalizes this technique to recurrent and long-short-term-memory (LSTM) networks. Wu and Gu (2015) analyze dropout applied to the max-pooling layers in convolutional neural networks. Wang and Manning (2013) present a Gaussian approximation to the gradient noise induced by dropout.

Generalization results for dropout training exist in various settings. Given bounds on the norms of weight vectors, Wan et al. (2013), Gao and Zhou (2016), and Zhai and Wang (2018) prove decreasing Rademacher complexity bounds as the dropout rate increases. Arora et al. (2021) bound the Rademacher complexity of shallow ReLU-activated networks with dropout. McAllester (2013) obtains a PAC-Bayes bound for dropout and illustrates a trade-off between large and small dropout probabilities for different terms in the bound.

Recently, Manita et al. (2022) demonstrated an universal approximation result in the vein of classic results (Cybenko, 1989; Hornik, 1991; Leshno et al., 1993), stating that any function in some generic semi-normed space that can be ε-approximated by a deterministic neural network may also be stochastically approximated in L^q-norm by a sufficiently large network with dropout.

Less is known about gradient descent training with dropout. Senen-Cerda and Sanders (2022) study the gradient flow associated with the explicit regularizer obtained by marginalizing the dropout noise in a shallow linear network. In particular, the flow converges exponentially fast within a neighborhood of a parameter vector satisfying a balancing condition. Mianjy and Arora (2020) study gradient descent with dropout on the logistic loss of a shallow ReLU-activated network in a binary classification task. Their main result includes an explicit rate for the misclassification error, assuming an overparametrized network oper-

ating in the so-called lazy regime, where the trained weights stay relatively close to their initializations, and two well-separated classes.

1.2 Notation

Column vectors $\mathbf{x} = (x_1, \dots, x_d)^\top$ are denoted by bold letters. We define $\mathbf{0} := (0, \dots, 0)^\top$, $\mathbf{1} := (1, \dots, 1)^\top$, and the Euclidean norm $\|\mathbf{x}\|_2 := \sqrt{\mathbf{x}^\top \mathbf{x}}$. The $d \times d$ identity matrix is symbolized by I_d, or simply I, when the dimension d is clear from context. For matrices A, B of the same dimension, $A \odot B$ denotes the Hadamard/entry-wise product $(A \odot B)_{ij} = A_{ij} B_{ij}$. We write $\mathrm{Diag}(A) := I \odot A$ for the diagonal matrix with the same main diagonal as A. Given $p \in (0, 1)$, we 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 rescaling the off-diagonal entries of A by p.

The smallest eigenvalue of a symmetric matrix A is denoted by $\lambda_{\min}(A)$. The operator norm of a linear operator $T : V \to W$ between normed linear spaces is given by $\|T\|_{\mathrm{op}} := \sup_{v \in V : \|v\|_V \le 1} \|Tv\|_W$. We write $\|\cdot\|$ for the spectral norm of matrices, which is the operator norm induced by $\|\cdot\|_2$. For symmetric matrices, the relation $A \ge B$ signifies $\mathbf{x}^\top (A - B)\mathbf{x} \ge 0$ for all non-zero vectors $\mathbf{x}$. The strict operator inequality $A > B$ is defined analogously.

2. Gradient Descent and Dropout

We consider a linear regression model with fixed $n \times d$ design matrix X and n outcomes $\mathbf{Y}$, so that

$$\mathbf{Y} = X\boldsymbol{\beta}_\star + \boldsymbol{\varepsilon}, \tag{1}$$

with unknown parameter $\boldsymbol{\beta}_\star$. We assume $\mathbb{E}\big[\boldsymbol{\varepsilon}\big] = \mathbf{0}$ and $\mathrm{Cov}\big(\boldsymbol{\varepsilon}\big) = I_n$. The task is to estimate $\boldsymbol{\beta}_\star$ from the observed data $(X, \mathbf{Y})$. As the Gram matrix $X^\top X$ appears throughout our analysis, we introduce the shorthand

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

Recovery of $\boldsymbol{\beta}_\star$ in the linear regression model (1) may be interpreted as training a neural network without intermediate hidden layers, see Figure 2. If X were to have a zero column, the corresponding regression coefficient would not affect the response vector $\mathbf{Y}$. Consequently, both the zero column and the regression coefficient may be eliminated from the linear regression model. Without zero columns, the model is said to be in *reduced form*.

The least squares criterion for the estimation of $\boldsymbol{\beta}_\star$ refers to the objective function $\boldsymbol{\beta} \mapsto \frac{1}{2}\|\mathbf{Y} - X\boldsymbol{\beta}\|_2^2$. Given a fixed learning rate $\alpha > 0$, performing gradient descent on the least squares objective leads to the iterative scheme

$$\widetilde{\boldsymbol{\beta}}^{\mathrm{gd}}_{k+1} = \widetilde{\boldsymbol{\beta}}^{\mathrm{gd}}_k - \alpha \nabla_{\widetilde{\boldsymbol{\beta}}^{\mathrm{gd}}_k} \frac{1}{2} \Big\| \mathbf{Y} - X\widetilde{\boldsymbol{\beta}}^{\mathrm{gd}}_k \Big\|_2^2 = \widetilde{\boldsymbol{\beta}}^{\mathrm{gd}}_k + \alpha X^\top \big(\mathbf{Y} - X\widetilde{\boldsymbol{\beta}}^{\mathrm{gd}}_k\big) \tag{2}$$

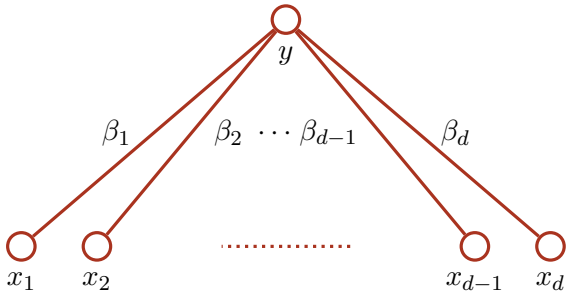

Figure 2: The linear regression model $y = \sum_{i=1}^{d} \beta_i x_i$, viewed as a neural network without hidden layers.

with $k = 0, 1, 2, \ldots$ and (possibly random) initialization $\widetilde{\boldsymbol{\beta}}_0^{\mathrm{gd}}$.

For standard gradient descent as defined in (2), the estimate is updated with the gradient of the full model. Dropout, as introduced in Srivastava et al. (2014), replaces the gradient of the full model with the gradient of a randomly reduced model during each iteration of training. To make this notion more precise, we call a random diagonal $d \times d$ matrix D a *p-dropout matrix*, or simply a *dropout matrix*, if its diagonal entries satisfy $D_{ii} \overset{i.i.d.}{\sim} \mathrm{Ber}(p)$ for some $p \in (0, 1)$. We note that the Bernoulli distribution may alternatively be parametrized with the failure probability $q := 1 - p$, but following Srivastava et al. (2014) we choose the success probability p.

On average, D has pd diagonal entries equal to 1 and $(1-p)d$ diagonal entries equal to 0. Given any vector $\boldsymbol{\beta}$, the coordinates of $D\boldsymbol{\beta}$ are randomly set to 0 with probability $1 - p$. For simplicity, the dependence of D on p will be omitted.

Now, let D_k, $k = 1, 2, \ldots$ be a sequence of i.i.d. dropout matrices, where D_k refers to the dropout matrix applied in the k^{th} iteration. Gradient descent with dropout takes the form

$$\widetilde{\boldsymbol{\beta}}_{k+1} = \widetilde{\boldsymbol{\beta}}_k - \alpha \nabla_{\widetilde{\boldsymbol{\beta}}_k} \frac{1}{2} \left\| \mathbf{Y} - X D_{k+1} \widetilde{\boldsymbol{\beta}}_k \right\|_2^2 = \widetilde{\boldsymbol{\beta}}_k + \alpha D_{k+1} X^\top \left(\mathbf{Y} - X D_{k+1} \widetilde{\boldsymbol{\beta}}_k \right) \tag{3}$$

with $k = 0, 1, 2, \ldots$ and (possibly random) initialization $\widetilde{\boldsymbol{\beta}}_0$. In contrast with (2), the gradient in (3) is taken on the model reduced by the action of multiplying $\widetilde{\boldsymbol{\beta}}_k$ with D_{k+1}. Alternatively, (3) may be interpreted as replacing the design matrix X with the reduced matrix XD_{k+1} during the $(k+1)^{\mathrm{th}}$ iteration. The columns of the reduced matrix are randomly deleted with a probability of $1 - p$. Observe that the dropout matrix appears inside the squared norm, making the gradient quadratic in D_{k+1}.

Dropout, as defined in (3), considers the full gradient of the reduced model, whereas another variant is obtained through reduction of the full gradient. The resulting iterative scheme takes the form

$$\widehat{\boldsymbol{\beta}}_{k+1} = \widehat{\boldsymbol{\beta}}_k - \alpha D_{k+1} \nabla_{\widehat{\boldsymbol{\beta}}_k} \frac{1}{2} \left\| \mathbf{Y} - X \widehat{\boldsymbol{\beta}}_k \right\|_2^2 = \widehat{\boldsymbol{\beta}}_k + \alpha D_{k+1} X^\top \left(\mathbf{Y} - X \widehat{\boldsymbol{\beta}}_k \right) \tag{4}$$

with $k = 0, 1, 2, \ldots$ and (possibly random) initialization $\widehat{\boldsymbol{\beta}}_0$. As opposed to $\widetilde{\boldsymbol{\beta}}_k$ defined above, the dropout matrix only occurs once in the updates, so we shall call this method

simplified dropout from here on. As we will illustrate, the quadratic dependence of $\widetilde{\boldsymbol{\beta}}_k$ on D_{k+1} creates various challenges, whereas the analysis of $\widehat{\boldsymbol{\beta}}_k$ is more straightforward.

Both versions (3) and (4) coincide when the Gram matrix $\mathbb{X} = X^\top X$ is diagonal, meaning when the columns of X are orthogonal. To see this, note that diagonal matrices commute, so $D_{k+1}^2 = D_{k+1}$ and hence $D_{k+1}\mathbb{X}D_{k+1} = D_{k+1}\mathbb{X}$.

We note that dropout need not require the complete removal of neurons. Each neuron may be multiplied by an arbitrary (not necessarily Bernoulli distributed) random variable. For instance, Srivastava et al. (2014) also report good performance for $\mathcal{N}(1,1)$-distributed diagonal entries of the dropout matrix. However, the Bernoulli variant seems well-motivated from a model averaging perspective. Srivastava et al. (2014) propose dropout with the explicit aim of approximating a Bayesian model averaging procedure over all possible combinations of connections in the network. The random removal of nodes during training is thought to prevent the neurons from co-adapting, recreating the model averaging effect. This is the main variant implemented in popular software libraries, such as *Caffe* (Jia et al., 2014), *TensorFlow* (Abadi et al., 2016), *Keras* (Chollet et al., 2015), and *PyTorch* (Paszke et al., 2019).

Numerous variations and extensions of dropout exist. Wan et al. (2013) show state-of-the-art results for networks with *DropConnect*, a generalization of dropout where connections are randomly dropped, instead of neurons. In the linear model, this coincides with standard dropout. Ba and Frey (2013) analyze the case of varying dropout probabilities, where the dropout probability for each neuron is computed using binary belief networks that share parameters with the underlying fully connected network. An adaptive procedure for the choice of dropout probabilities is presented in Kingma et al. (2015), while also giving a Bayesian justification for dropout.

For a comprehensive overview of established methods and cutting-edge variants, see Moradi et al. (2020) and Santos and Papa (2022).

3. Analysis of Averaged Dropout

Before presenting our main results on iterative dropout schemes, we further discuss some properties of the marginalized loss minimizer that was first analyzed by Srivastava et al. (2014). For the linear regression model (1), marginalizing the dropout noise leads to

$$\widetilde{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta}} \mathbb{E}\Big[\big\|\mathbf{Y} - XD\boldsymbol{\beta}\big\|_2^2 \mid \mathbf{Y}\Big]. \tag{5}$$

One may hope that the dropout gradient descent recursion for $\widetilde{\boldsymbol{\beta}}_k$ in (3) leads to a minimizer of (5), so that the marginalized loss minimizer may be studied as a surrogate for the behaviour of $\widetilde{\boldsymbol{\beta}}_k$ in the long run.

Intuitively, the gradient descent iterates with dropout represent a Monte-Carlo estimate of some deterministic algorithm (Wang and Manning, 2013). This can be motivated by separating the gradient descent update into a part without algorithmic randomness and a

centered noise term, meaning

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_{k+1} = \widetilde{\boldsymbol{\beta}}_k &- \frac{\alpha}{2}\mathbb{E}\Big[\nabla_{\widetilde{\boldsymbol{\beta}}_k}\Big\|\mathbf{Y} - XD_{k+1}\widetilde{\boldsymbol{\beta}}_k\Big\|_2^2 \mid \mathbf{Y}, \widetilde{\boldsymbol{\beta}}_k\Big] \\ &- \frac{\alpha}{2}\left(\nabla_{\widetilde{\boldsymbol{\beta}}_k}\Big\|\mathbf{Y} - XD_{k+1}\widetilde{\boldsymbol{\beta}}_k\Big\|_2^2 - \mathbb{E}\Big[\nabla_{\widetilde{\boldsymbol{\beta}}_k}\Big\|\mathbf{Y} - XD_{k+1}\widetilde{\boldsymbol{\beta}}_k\Big\|_2^2 \mid \mathbf{Y}, \widetilde{\boldsymbol{\beta}}_k\Big]\right).\end{aligned} \tag{6}$$

Notably, the stochastic terms form a martingale difference sequence with respect to $(\mathbf{Y}, \widetilde{\boldsymbol{\beta}}_k)$. It seems conceivable that the noise in (6) averages out; despite the random variables being neither independent, nor identically distributed, one may hope that a law of large numbers still holds, see Andrews (1988). In this case, after a sufficient number of gradient steps,

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_{k+1} &= \widetilde{\boldsymbol{\beta}}_0 - \frac{\alpha}{2}\sum_{\ell=1}^{k}\nabla_{\widetilde{\boldsymbol{\beta}}_\ell}\Big\|\mathbf{Y} - XD_{\ell+1}\widetilde{\boldsymbol{\beta}}_\ell\Big\|_2^2 \\ &\approx \widetilde{\boldsymbol{\beta}}_0 - \frac{\alpha}{2}\sum_{\ell=1}^{k}\mathbb{E}\Big[\nabla_{\widetilde{\boldsymbol{\beta}}_\ell}\Big\|\mathbf{Y} - XD_{\ell+1}\widetilde{\boldsymbol{\beta}}_\ell\Big\|_2^2 \mid \mathbf{Y}, \widetilde{\boldsymbol{\beta}}_k\Big].\end{aligned} \tag{7}$$

The latter sequence could plausibly converge to the marginalized loss minimizer $\widetilde{\boldsymbol{\beta}}$. While this motivates studying $\widetilde{\boldsymbol{\beta}}$, the main conclusion of our work is that this heuristic is not entirely correct and additional noise terms occur in the limit $k \to \infty$.

As the marginalized loss minimizer $\widetilde{\boldsymbol{\beta}}$ still plays a pivotal role in our analysis, we briefly recount and expand on some of the properties derived in Srivastava et al. (2014). Recall that $\mathbb{X} = X^\top X$, so we have

$$\big\|\mathbf{Y} - XD\boldsymbol{\beta}\big\|_2^2 = \|\mathbf{Y}\|_2^2 - 2\mathbf{Y}^\top XD\boldsymbol{\beta} + \boldsymbol{\beta}^\top D\mathbb{X}D\boldsymbol{\beta}.$$

Since D is diagonal, $\mathbb{E}[D] = pI_d$, and by Lemma 16(a), $\mathbb{E}[D\mathbb{X}D] = p^2\mathbb{X} + p(1-p)\mathrm{Diag}(\mathbb{X})$,

$$\begin{aligned}\mathbb{E}\Big[\big\|\mathbf{Y} - XD\boldsymbol{\beta}\big\|_2^2 \mid \mathbf{Y}\Big] &= \|\mathbf{Y}\|_2^2 - 2p\mathbf{Y}^\top X\boldsymbol{\beta} + p^2\boldsymbol{\beta}^\top\mathbb{X}\boldsymbol{\beta} + p(1-p)\boldsymbol{\beta}^\top\mathrm{Diag}(\mathbb{X})\boldsymbol{\beta} \\ &= \big\|\mathbf{Y} - pX\boldsymbol{\beta}\big\|_2^2 + p(1-p)\boldsymbol{\beta}^\top\mathrm{Diag}(\mathbb{X})\boldsymbol{\beta}.\end{aligned} \tag{8}$$

The right-hand side may be identified with a Tikhonov functional, or an ℓ_2-penalized least squares objective. Its gradient with respect to $\boldsymbol{\beta}$ is given by

$$\nabla_{\boldsymbol{\beta}}\mathbb{E}\big[\|\mathbf{Y} - XD\boldsymbol{\beta}\|_2^2 \mid \mathbf{Y}\big] = -2pX^\top\mathbf{Y} + 2\big(p^2\mathbb{X} + p(1-p)\mathrm{Diag}(\mathbb{X})\big)\boldsymbol{\beta}.$$

Recall from the discussion following Equation (1) that the model is assumed to be in reduced form, meaning $\min_i \mathbb{X}_{ii} > 0$. In turn,

$$p^2\mathbb{X} + p(1-p)\mathrm{Diag}(\mathbb{X}) \geq p(1-p)\mathrm{Diag}(\mathbb{X}) \geq p(1-p)\min_i \mathbb{X}_{ii} \cdot I_d$$

is bounded away from 0, making $p^2\mathbb{X} + p(1-p)\mathrm{Diag}(\mathbb{X})$ invertible. Solving the gradient for the minimizer $\widetilde{\boldsymbol{\beta}}$ now leads to

$$\widetilde{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^d}\mathbb{E}\Big[\big\|\mathbf{Y} - XD\boldsymbol{\beta}\big\|_2^2 \mid \mathbf{Y}\Big] = p\Big(p^2\mathbb{X} + p(1-p)\mathrm{Diag}(\mathbb{X})\Big)^{-1}X^\top\mathbf{Y} = \mathbb{X}_p^{-1}X^\top\mathbf{Y}, \tag{9}$$

where $\mathbb{X}_p := p\mathbb{X}+(1-p)\mathrm{Diag}(\mathbb{X})$. If the columns of X are orthogonal, then $\mathbb{X}$ is diagonal and hence $\mathbb{X}_p = \mathbb{X}$. In this case, $\widetilde{\boldsymbol{\beta}}$ matches the usual linear least squares estimator $\mathbb{X}^{-1}X^\top\mathbf{Y}$. Alternatively, $\widetilde{\boldsymbol{\beta}}$ minimizing the marginalized loss can also be deduced from the identity

$$\mathbb{E}\Big[\big\|\mathbf{Y} - XD\widehat{\boldsymbol{\beta}}\big\|_2^2 \mid \mathbf{Y}\Big] = \mathbb{E}\Big[\big\|\mathbf{Y} - XD\widetilde{\boldsymbol{\beta}}\big\|_2^2 \mid \mathbf{Y}\Big] + \mathbb{E}\Big[\big\|XD\big(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\big)\big\|_2^2 \mid \mathbf{Y}\Big], \tag{10}$$

which holds for all estimators $\widehat{\boldsymbol{\beta}}$. See Appendix A for a proof of (10). We now mention several other relevant properties of $\widetilde{\boldsymbol{\beta}}$.

Calibration: Srivastava et al. (2014) recommend multiplying $\widetilde{\boldsymbol{\beta}}$ by p, which may be motivated as follows: Since $\mathbf{Y} = X\boldsymbol{\beta}_\star + \boldsymbol{\varepsilon}$, a small squared error $\big\|\mathbf{Y} - pX\widetilde{\boldsymbol{\beta}}\big\|_2^2$ in (8) implies $\boldsymbol{\beta}_\star \approx p\widetilde{\boldsymbol{\beta}}$. Moreover, multiplying $\widetilde{\boldsymbol{\beta}}$ by p leads to $p\widetilde{\boldsymbol{\beta}} = \big(\mathbb{X} + (1/p-1)\mathrm{Diag}(\mathbb{X})\big)^{-1}X^\top\mathbf{Y}$ which may be identified with the minimizer of the objective function

$$\boldsymbol{\beta} \mapsto \big\|\mathbf{Y} - X\boldsymbol{\beta}\big\|_2^2 + (p^{-1}-1)\boldsymbol{\beta}^\top\mathrm{Diag}(\mathbb{X})\boldsymbol{\beta} = \mathbb{E}\big[\|\mathbf{Y} - Xp^{-1}D\boldsymbol{\beta}\|_2^2 \mid \mathbf{Y}\big].$$

This recasts $p\widetilde{\boldsymbol{\beta}}$ as resulting from a weighted form of ridge regression. Comparing the objective function to the original marginalized loss $\mathbb{E}\big[\|\mathbf{Y} - XD\boldsymbol{\beta}\|_2^2 \mid \mathbf{Y}\big]$, the rescaling replaces D with the normalized dropout matrix $p^{-1}D$, which has the identity matrix as its expected value. In popular machine learning software, the sampled dropout matrices are usually rescaled by p^{-1} (Abadi et al., 2016; Chollet et al., 2015; Jia et al., 2014; Paszke et al., 2019).

In some settings multiplication by p may worsen $\widetilde{\boldsymbol{\beta}}$ as a statistical estimator. As an example, consider the case $n = d$ with $X = nI_n$ a multiple of the identity matrix. Now, (9) turns into $\widetilde{\boldsymbol{\beta}} = n^{-1}\mathbf{Y} = \boldsymbol{\beta}_\star + n^{-1}\boldsymbol{\varepsilon}$. If the noise vector $\boldsymbol{\varepsilon}$ consists of independent standard normal random variables, then $\widetilde{\boldsymbol{\beta}}$ has mean squared error $\mathbb{E}\big[\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star\|_2^2\big] = n^{-1}$. In contrast, $\mathbb{E}\big[\|p\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star\|_2^2\big] = (1-p)\|\boldsymbol{\beta}_\star\|_2^2 + p^2n^{-1}$, so $\widetilde{\boldsymbol{\beta}}$ converges to $\boldsymbol{\beta}_\star$ at the rate n^{-1} while $p\widetilde{\boldsymbol{\beta}}$ cannot be consistent as $n \to \infty$, unless $\boldsymbol{\beta}_\star = \mathbf{0}$.

The correct rescaling may also depend on the parameter dimension d and the spectrum of $\mathbb{X}$. Suppose that all columns of X have the same Euclidean norm, so that $\mathrm{Diag}(\mathbb{X}) = \mathbb{X}_{11}\cdot I_d$. Let $X = \sum_{\ell=1}^{\mathrm{rank}(X)} \sigma_\ell \mathbf{v}_\ell \mathbf{w}_\ell^\top$ denote a singular value decomposition of X, with singular values $\sigma_1, \ldots, \sigma_{\mathrm{rank}(X)}$. Now, $\widetilde{\boldsymbol{\beta}}$ satisfies

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}} &= \sum_{\ell=1}^{\mathrm{rank}(X)} \frac{1}{p\sigma_\ell^2 + (1-p)\mathbb{X}_{11}}\big(\sigma_\ell \mathbf{w}_\ell \mathbf{v}_\ell^\top\big)\mathbf{Y} \\
\mathbb{E}\big[X\widetilde{\boldsymbol{\beta}}\big] &= \sum_{\ell=1}^{\mathrm{rank}(X)} \frac{1}{p + (1-p)\mathbb{X}_{11}/\sigma_\ell^2}\big(\sigma_\ell \mathbf{v}_\ell \mathbf{w}_\ell^\top\big)\boldsymbol{\beta}_\star.
\end{aligned} \tag{11}$$

For a proof of these identities, see Appendix A. To get an unbiased estimator for $X\boldsymbol{\beta}_\star = \sum_{\ell=1}^{\mathrm{rank}(X)} \sigma_\ell \mathbf{v}_\ell \mathbf{w}_\ell^\top \boldsymbol{\beta}_\star$, we must undo the effect of the spectral multipliers $1/(p+(1-p)\mathbb{X}_{11}/\sigma_\ell^2)$ which take values in the interval $[0, 1/p]$. Consequently, the proper rescaling depends on the eigenspace. Multiplication of the estimator by p addresses the case where the singular values σ_ℓ are large. In particular, if $X = \sigma\mathbf{v}\mathbf{w}^\top$ with $\sigma = \sqrt{nd}$, $\mathbf{v} = (n^{-1/2}, \ldots, n^{-1/2})^\top$, and $\mathbf{w} = (d^{-1/2}, \ldots, d^{-1/2})^\top$, then X is the $n \times d$ matrix with all entries equaling 1. Now

$\mathbb{X}_{11} = n$ and so $\mathbb{X}_{11}/\sigma^2 = d^{-1}$, meaning the correct scaling factor depends explicitly on the parameter dimension d and converges to the dropout probability p as $d \to \infty$.

Invariance properties: The minimizer $\widetilde{\boldsymbol{\beta}} = \mathbb{X}_p^{-1} X^\top \mathbf{Y}$ is scale invariant in the sense that $\mathbf{Y}$ and X may be replaced with $\gamma\mathbf{Y}$ and γX for some arbitrary $\gamma \neq 0$, without changing $\widetilde{\boldsymbol{\beta}}$. This does not hold for the gradient descent iterates (3) and (4), since rescaling by γ changes the learning rate from α to $\alpha\gamma^2$. Moreover, $\widetilde{\boldsymbol{\beta}}$ as well as the gradient descent iterates $\widetilde{\boldsymbol{\beta}}_k$ in (3) and $\widehat{\boldsymbol{\beta}}_k$ in (4) are invariant under replacement of $\mathbf{Y}$ and X by $Q\mathbf{Y}$ and QX for any orthogonal $n \times n$ matrix Q. See Helmbold and Long (2017) for further results on scale-invariance of dropout in deep networks.

Overparametrization: Dropout has been successfully applied in the overparametrized regime, see for example Krizhevsky et al. (2012). For the overparametrized linear regression model, the data-misfit term in (8) suggests that $pX\widetilde{\boldsymbol{\beta}} = X\big(\mathbb{X} + (p^{-1} - 1)\mathrm{Diag}(\mathbb{X})\big)^{-1} X^\top \mathbf{Y}$ should be close to the data vector $\mathbf{Y}$. However,

$$\Big\| X\big(\mathbb{X} + (p^{-1} - 1)\mathrm{Diag}(\mathbb{X})\big)^{-1} X^\top \Big\| < 1. \tag{12}$$

See Appendix A for a proof. Hence, $pX\widetilde{\boldsymbol{\beta}}$ also shrinks $\mathbf{Y}$ towards zero in the overparametrized regime and does not interpolate the data. The variational formulation $\widetilde{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \big\|\mathbf{Y} - pX\boldsymbol{\beta}\big\|_2^2 + p(1-p)\boldsymbol{\beta}^\top \mathrm{Diag}(\mathbb{X})\boldsymbol{\beta}$ reveals that $\widetilde{\boldsymbol{\beta}}$ is a minimum-norm solution in the sense

$$\widetilde{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta}: X\boldsymbol{\beta} = X\widetilde{\boldsymbol{\beta}}} \boldsymbol{\beta}^\top \mathrm{Diag}(\mathbb{X})\boldsymbol{\beta},$$

which explains the induced shrinkage.

4. Analysis of Iterative Dropout Schemes

In the linear model, gradient descent with a small but fixed learning rate, as in (2), leads to exponential convergence in the number of iterations. Accordingly, we analyze the iterative dropout schemes (3) and (4) for fixed learning rate α and only briefly discuss the algebraically less tractable case of decaying learning rates.

4.1 Convergence of Dropout

We proceed by assessing convergence of the iterative dropout scheme (3), as well as some of its statistical properties. Recall that gradient descent with dropout takes the form

$$\widetilde{\boldsymbol{\beta}}_k = \widetilde{\boldsymbol{\beta}}_{k-1} + \alpha D_k X^\top \big(\mathbf{Y} - X D_k \widetilde{\boldsymbol{\beta}}_{k-1}\big) = \big(I - \alpha D_k \mathbb{X} D_k\big)\widetilde{\boldsymbol{\beta}}_{k-1} + \alpha D_k X^\top \mathbf{Y}. \tag{13}$$

As alluded to in the beginning of Section 3, the gradient descent iterates should be related to the minimizer $\widetilde{\boldsymbol{\beta}}$ of (5). It then seems natural to study the difference $\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}$, with $\widetilde{\boldsymbol{\beta}}$ as an "anchoring point". Comparing $\widetilde{\boldsymbol{\beta}}_k$ and $\widetilde{\boldsymbol{\beta}}$ demands an explicit analysis, without marginalization of the dropout noise.

To start, we rewrite the updating formula (13) in terms of $\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}$. Using $D_k^2 = D_k$, $\mathrm{Diag}(\mathbb{X}) = \mathbb{X}_p - p\overline{\mathbb{X}}$, and that diagonal matrices always commute, we obtain $D_k \mathbb{X} D_k =$

$D_k\overline{\mathbb{X}}D_k + D_k\mathrm{Diag}(\mathbb{X}) = D_k\overline{\mathbb{X}}D_k + D_k\mathbb{X}_p - pD_k\overline{\mathbb{X}}$. As defined in (9), $\mathbb{X}_p\widetilde{\boldsymbol{\beta}} = X^\top\mathbf{Y}$ and thus

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}} &= \big(I - \alpha D_k\mathbb{X}D_k\big)\big(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}}\big) + \alpha D_k\overline{\mathbb{X}}\big(pI - D_k\big)\widetilde{\boldsymbol{\beta}} \\ &= \big(I - \alpha D_k\mathbb{X}_p\big)\big(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}}\big) + \alpha D_k\overline{\mathbb{X}}\big(pI - D_k\big)\widetilde{\boldsymbol{\beta}}_{k-1}.\end{aligned} \tag{14}$$

In both representations, the second term is centered and uncorrelated for different values of k. Vanishing of the mean follows from the independence of D_k and $\big(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}_{k-1}\big)$, combined with $\mathbb{E}\big[D_k\overline{\mathbb{X}}(pI - D_k)\big] = 0$, the latter being shown in (29). If $k > \ell$, independence of D_k and $\big(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}_{k-1}, \widetilde{\boldsymbol{\beta}}_{\ell-1}\big)$, as well as $\mathbb{E}\big[D_k\overline{\mathbb{X}}(pI - D_k)\big] = 0$, imply $\mathrm{Cov}\big(D_k\overline{\mathbb{X}}(pI - D_k)\widetilde{\boldsymbol{\beta}}, D_\ell\overline{\mathbb{X}}(pI - D_\ell)\widetilde{\boldsymbol{\beta}}\big) = 0$ and $\mathrm{Cov}\big(D_k\overline{\mathbb{X}}(pI - D_k)\widetilde{\boldsymbol{\beta}}_{k-1}, D_\ell\overline{\mathbb{X}}(pI - D_\ell)\widetilde{\boldsymbol{\beta}}_{\ell-1}\big) = 0$, which proves uncorrelatedness.

Defining $Z_k := \widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}$, $G_k := I - \alpha D_k\mathbb{X}D_k$, and $\boldsymbol{\xi}_k := \alpha D_k\overline{\mathbb{X}}\big(pI - D_k\big)\widetilde{\boldsymbol{\beta}}$, the first representation in (14) may be identified with a lag one vector autoregressive (VAR) process

$$Z_k = G_kZ_{k-1} + \boldsymbol{\xi}_k \tag{15}$$

with i.i.d. random coefficients G_k and noise/innovation process $\boldsymbol{\xi}_k$. As just shown, $\mathbb{E}[\boldsymbol{\xi}_k] = 0$ and $\mathrm{Cov}(\boldsymbol{\xi}_k, \boldsymbol{\xi}_\ell) = 0$ whenever $k \neq \ell$, so the noise process is centered and serially uncorrelated. The random coefficients G_k and $\boldsymbol{\xi}_k$ are, however, dependent. While most authors do not allow for random coefficients G_k in VAR processes, such processes are special cases of a random autoregressive process (RAR) (Regis et al., 2022).

In the VAR literature, identifiability and estimation of the random coefficients G_k is considered (Nicholls and Quinn, 1982; Regis et al., 2022). In contrast, we aim to obtain bounds for the convergence of $\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big]$ and $\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big)$. Difficulties arise from the involved structure and coupled randomness of G_k and $\boldsymbol{\xi}_k$. Estimation of coefficients under dependence of G_k and $\boldsymbol{\xi}_k$ is treated in Hill and Peng (2014).

For a sufficiently small learning rate α, the random matrices $I - \alpha D_k\mathbb{X}D_k$ and $I - \alpha D_k\mathbb{X}_p$ in both representations in (14) are contractive maps in expectation. By Lemma 16(a), their expected values coincide since

$$\mathbb{E}\big[I - \alpha D_k\mathbb{X}D_k\big] = \mathbb{E}\big[I - \alpha D_k\mathbb{X}_p\big] = I - \alpha p\mathbb{X}_p.$$

For the subsequent analysis, we impose the following mild conditions that, among other things, establish contractivity of $I - \alpha p\mathbb{X}_p$ as a linear map.

Assumption 1 *The learning rate α and the dropout probability p are chosen such that $\alpha p\|\mathbb{X}\| < 1$, the initialization $\widetilde{\boldsymbol{\beta}}_0$ is a square integrable random vector that is independent of the data $\mathbf{Y}$ and the model is in reduced form, meaning that X does not have zero columns.*

For gradient descent without dropout and fixed learning rate, as defined in (2), $\alpha\|\mathbb{X}\| < 1$ guarantees converge of the scheme in expectation. We will see shortly that dropout essentially replaces the expected learning rate with αp, which motivates the condition $\alpha p\|\mathbb{X}\| < 1$.

As a straightforward consequence of the definitions, we are now able to show that $\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}$ vanishes in expectation at a geometric rate. For a proof of this as well as subsequent results, see Appendix B.

Lemma 1 (Convergence of Expectation) *Given Assumption 1,* $\|I - \alpha p \mathbb{X}_p\| \leq 1 - \alpha p(1-p)\min_i \mathbb{X}_{ii} < 1$ *and for any* $k = 0, 1, \ldots,$

$$\left\|\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big]\right\|_2 \leq \|I - \alpha p \mathbb{X}_p\|^k \left\|\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_0 - \widetilde{\boldsymbol{\beta}}\big]\right\|_2.$$

Before turning to the analysis of the covariance structure, we highlight a property of the sequence $\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k \mid \mathbf{Y}\big]$. As mentioned, these conditional expectations may be viewed as gradient descent iterates generated by the marginalized objective $\frac{1}{2}\mathbb{E}\big[\|\mathbf{Y} - XD\boldsymbol{\beta}\|_2^2 \mid \mathbf{Y}\big]$ that gives rise to $\widetilde{\boldsymbol{\beta}}$. Indeed, combining (13) with $\mathbb{E}[D_{k+1}] = pI_d$, $\mathbb{E}\big[D_{k+1}\mathbb{X}D_{k+1}\big] = p\mathbb{X}_p$ from Lemma 16(a), and (8) yields

$$\begin{aligned}\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_{k+1} \mid \mathbf{Y}\big] &= \mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k \mid \mathbf{Y}\big] + \alpha p X^\top \mathbf{Y} - \alpha p \mathbb{X}_p \mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k \mid \mathbf{Y}\big] \\ &= \mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k \mid \mathbf{Y}\big] - \frac{\alpha}{2}\nabla_{\mathbb{E}[\widetilde{\boldsymbol{\beta}}_k|\mathbf{Y}]}\mathbb{E}\Big[\big\|\mathbf{Y} - XD\mathbb{E}[\widetilde{\boldsymbol{\beta}}_k \mid \mathbf{Y}]\big\|_2^2 \mid \mathbf{Y}\Big].\end{aligned}$$

This establishes a connection between the dropout iterates and the averaged analysis of the previous section. However, the relationship between the (unconditional) covariance matrices $\mathrm{Cov}(\widetilde{\boldsymbol{\beta}}_k)$ and the added noise remains unclear. A new dropout matrix is sampled for each iteration, whereas $\widetilde{\boldsymbol{\beta}}$ results from minimization only after applying the conditional expectation $\mathbb{E}[\,\cdot \mid \mathbf{Y}]$ to the randomized objective function. Hence, we may expect that $\widetilde{\boldsymbol{\beta}}$ features smaller variance than the iterates as the latter also depend on the noise added via dropout.

As a first result for the covariance analysis, we establish an extension of the Gauss-Markov theorem stating that the covariance matrix of a linear estimator lower-bounds the covariance matrix of an affine estimator, provided that both estimators have the same asymptotic mean. Moreover, the covariance matrix of their difference characterizes the gap. We believe that a similar result may already be known, but we are not aware of any reference, so a full proof is provided in Appendix B for completeness.

Theorem 2 *In the linear regression model* (1)*, consider estimators* $\widetilde{\boldsymbol{\beta}}_A = AX^\top \mathbf{Y}$ *and* $\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} = B\mathbf{Y} + \mathbf{a}$*, with* $B \in \mathbb{R}^{d\times n}$ *and* $\mathbf{a} \in \mathbb{R}^d$ *(possibly) random, but independent of* $\mathbf{Y}$*, and* $A \in \mathbb{R}^{d\times d}$ *deterministic. Then,*

$$\Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}}\big) - \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_A\big) - \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A\big)\Big\| \leq 4\|A\| \sup_{\boldsymbol{\beta}_\star : \|\boldsymbol{\beta}_\star\|_2 \leq 1} \big\|\mathbb{E}_{\boldsymbol{\beta}_\star}\big[\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A\big]\big\|_2,$$

where $\mathbb{E}_{\boldsymbol{\beta}_\star}$ *denotes the expectation with respect to* $\boldsymbol{\beta}_\star$ *being the true regression vector in the linear regression model* (1)*.*

Since $\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}}\big) - \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_A\big) - \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A\big) = \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A, \widetilde{\boldsymbol{\beta}}_A\big) + \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_A, \widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A\big)$, Theorem 2 may be interpreted as follows: if the estimators $\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}}$ and $\widetilde{\boldsymbol{\beta}}_A$ are nearly the same in expectation, then $\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A$ and $\widetilde{\boldsymbol{\beta}}_A$ must be nearly uncorrelated. In turn, $\widetilde{\boldsymbol{\beta}}_k$ may be decomposed into $\widetilde{\boldsymbol{\beta}}_A$ and (nearly) orthogonal noise $\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A$, so that $\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}}\big) \approx \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_A\big) + \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A\big)$ is lower bounded by $\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_A\big)$. Therefore, the covariance matrix $\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A\big)$ quantifies the gap in the bound.

Taking $A := \mathbb{X}^{-1}$ and considering linear estimators with $\mathbf{a} = \mathbf{0}$ recovers the usual Gauss-Markov theorem, stating that $\mathbb{X}^{-1}X^\top\mathbf{Y}$ is the best linear unbiased estimator (BLUE) for

the linear model. Applying the generalized Gauss-Markov theorem with $A = (\mathbb{X} + \Gamma)^{-1}$, where Γ is a positive definite matrix, we obtain the following statement about ℓ_2-penalized estimators.

Corollary 3 *The minimizer $\widetilde{\boldsymbol{\beta}}_\Gamma := (\mathbb{X} + \Gamma)^{-1} X^\top \mathbf{Y}$ of the ℓ_2-penalized functional $\|\mathbf{Y} - X\boldsymbol{\beta}\|_2^2 + \boldsymbol{\beta}^\top \Gamma \boldsymbol{\beta}$ has the smallest covariance matrix among all affine estimators with the same expectation as $\widetilde{\boldsymbol{\beta}}_\Gamma$.*

We now return to our analysis of the covariance structure induced by dropout. If $A := \mathbb{X}_p^{-1}$, then $\widetilde{\boldsymbol{\beta}} = \mathbb{X}_p^{-1} X^\top \mathbf{Y} = A X^\top \mathbf{Y} = \widetilde{\boldsymbol{\beta}}_A$ in Theorem 2. Further, the dropout iterates may be rewritten as affine estimators $\widetilde{\boldsymbol{\beta}}_k = B_k \mathbf{Y} + \mathbf{a}_k$ with

$$B_k := \sum_{j=1}^{k-1} \left(\prod_{\ell=0}^{k-j-1} \left(I - \alpha D_{k-\ell} \mathbb{X} D_{k-\ell} \right) \right) \alpha D_j X^\top + \alpha D_k X^\top$$
$$\mathbf{a}_k := \left(\prod_{\ell=0}^{k-1} \left(I - \alpha D_{k-\ell} \mathbb{X} D_{k-\ell} \right) \right) \widetilde{\boldsymbol{\beta}}_0.$$

By construction, $(B_k, \mathbf{a}_k)$ and $\mathbf{Y}$ are independent, so Theorem 2 applies. As shown in Lemma 1, $\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big]$ vanishes exponentially fast, so we conclude that $\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k\big)$ is asymptotically lower-bounded by $\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}\big)$. Further, the covariance structure of $\widetilde{\boldsymbol{\beta}}$ is optimal in the sense of Corollary 3.

We proceed by studying $\mathrm{Cov}(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}})$, with the aim of quantifying the gap between the covariance matrices. To this end, we exhibit a particular recurrence for the second moments $\mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}})(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}})^\top\big]$. Recall that $\odot$ denotes the Hadamard product, $B_p = pB + (1-p)\mathrm{Diag}(B)$, and $\overline{B} = B - \mathrm{Diag}(B)$.

Lemma 4 (Second Moment - Recursive Formula) *Under Assumption 1, for all positive integers k*

$$\begin{aligned}
&\left\| \mathbb{E}\Big[\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big) \big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big)^\top \Big] - S\left(\mathbb{E}\Big[\big(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}}\big) \big(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}}\big)^\top \Big] \right) \right\| \\
&\leq 6 \big\| I - \alpha p \mathbb{X}_p \big\|^{k-1} \Big\| \mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_0 - \widetilde{\boldsymbol{\beta}}) \widetilde{\boldsymbol{\beta}}^\top \big] \Big\|,
\end{aligned}$$

where $S : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ denotes the affine operator

$$\begin{aligned}
S(A) = {} & \big(I - \alpha p \mathbb{X}_p\big) A \big(I - \alpha p \mathbb{X}_p\big) \\
& + \alpha^2 p(1-p) \mathrm{Diag}\big(\mathbb{X}_p A \mathbb{X}_p\big) + \alpha^2 p^2 (1-p)^2 \mathbb{X} \odot \overline{A + \mathbb{E}\big[\widetilde{\boldsymbol{\beta}} \widetilde{\boldsymbol{\beta}}^\top\big]} \odot \mathbb{X} \\
& + \alpha^2 p^2 (1-p) \left(\Big(\overline{\mathbb{X}} \mathrm{Diag}\Big(A + \mathbb{E}\big[\widetilde{\boldsymbol{\beta}} \widetilde{\boldsymbol{\beta}}^\top\big] \Big) \overline{\mathbb{X}} \Big)_p + \overline{\mathbb{X}} \mathrm{Diag}\big(\mathbb{X}_p A\big) + \mathrm{Diag}\big(\mathbb{X}_p A\big) \overline{\mathbb{X}} \right).
\end{aligned}$$

Intuitively, the lemma states that the second moment of $\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}$ evolves as an affine dynamical system, up to some exponentially decaying remainder. This may be associated

with the implicit regularization of the dropout noise, as illustrated empirically in Wei et al. (2020).

Mathematically, the result may be motivated via the representation of the dropout iterates as a random autoregressive process $Z_k = G_k Z_{k-1} + \boldsymbol{\xi}_k$ in (15). Writing out $Z_k Z_k^\top = G_k Z_{k-1} Z_{k-1}^\top G_k + \boldsymbol{\xi}_k \boldsymbol{\xi}_k^\top + G_k Z_{k-1} \boldsymbol{\xi}_k^\top + \boldsymbol{\xi}_k Z_{k-1}^\top G_k$ and comparing with the proof of the lemma, we see that the remainder term, denoted by ρ_{k-1} in the proof, coincides with the expected value of the cross terms $G_k Z_{k-1} \boldsymbol{\xi}_k^\top + \boldsymbol{\xi}_k Z_{k-1}^\top G_k$. Moreover, the operator S is obtained by computing

$$S(A) = \mathbb{E}\Big[G_k A G_k + \boldsymbol{\xi}_k \boldsymbol{\xi}_k^\top\Big].$$

As $(G_k, \boldsymbol{\xi}_k)$ are i.i.d., S does not depend on k. Moreover, independence of G_k and Z_{k-1} implies

$$\begin{aligned}
\mathbb{E}\Big[Z_k Z_k^\top\Big] &= \mathbb{E}\Big[G_k Z_{k-1} Z_{k-1}^\top G_k + \boldsymbol{\xi}_k \boldsymbol{\xi}_k^\top\Big] + \rho_{k-1} \\
&= \mathbb{E}\Big[G_k \mathbb{E}\big[Z_{k-1} Z_{k-1}^\top\big] G_k + \boldsymbol{\xi}_k \boldsymbol{\xi}_k^\top\Big] + \rho_{k-1} \\
&= S\Big(\mathbb{E}\big[Z_{k-1} Z_{k-1}^\top\big]\Big) + \rho_{k-1}.
\end{aligned}$$

Inserting the definition $Z_k = \widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}$ results in the statement of the lemma. The random vector $\boldsymbol{\xi}_k$ depends on $\widetilde{\boldsymbol{\beta}}$ and by Theorem 2 the correlation between $Z_k = \widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ decreases as $k \to \infty$. This leads to the exponentially decaying bound for the remainder term ρ_{k-1}.

The previous lemma entails equality between $\mathbb{E}\big[Z_k Z_k^\top\big]$ and $S^k\big(\mathbb{E}[Z_0 Z_0^\top]\big)$, up to the remainder terms. The latter may be computed further by decomposing the affine operator S into its intercept and linear part

$$S_0 := S(0) = \mathbb{E}\big[\boldsymbol{\xi}_k \boldsymbol{\xi}_k^\top\big] \qquad \text{and} \qquad S_{\text{lin}}(A) := S(A) - S_0 = \mathbb{E}\big[G_k A G_k\big]. \tag{16}$$

If S_{lin} were to have operator norm less than one, then the Neumann series for $(\mathrm{id} - S_{\text{lin}})^{-1}$ (see Lemma 18) gives

$$S^k(A) = \sum_{j=0}^{k-1} S_{\text{lin}}^j(S_0) + S_{\text{lin}}^k(A) \to \sum_{j=0}^{\infty} S_{\text{lin}}^j(S_0) = \big(\mathrm{id} - S_{\text{lin}}\big)^{-1} S_0,$$

with id the identity operator on $d \times d$ matrices. Surprisingly, the operator "forgets" A in the sense that the limit does not depend on A anymore. The argument shows that $\mathbb{E}\big[Z_k Z_k^\top\big] = \mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}})(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}})^\top\big]$ should behave like $\big(\mathrm{id} - S_{\text{lin}}\big)^{-1} S_0$ in the first order. The next result makes this precise, taking into account the remainder terms and approximation errors.

Theorem 5 (Second Moment - Limit Formula) *In addition to Assumption 1 suppose* $\alpha < \frac{\lambda_{\min}(\mathbb{X}_p)}{3\|\mathbb{X}\|^2}$*, then, for any* $k = 1, 2, \ldots$

$$\Big\| \mathbb{E}\Big[\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big)\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big)^\top\Big] - \big(\mathrm{id} - S_{\text{lin}}\big)^{-1} S_0 \Big\| \leq Ck \big\| I - \alpha p \mathbb{X}_p \big\|^{k-1}$$

and

$$\Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big) - \big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1} S_0\Big\| \leq Ck\big\|I - \alpha p \mathbb{X}_p\big\|^{k-1}$$

with constant C given by

$$C := \Big\|\mathbb{E}\Big[\big(\widetilde{\boldsymbol{\beta}}_0 - \widetilde{\boldsymbol{\beta}}\big)\big(\widetilde{\boldsymbol{\beta}}_0 - \widetilde{\boldsymbol{\beta}}\big)^\top\Big] - \big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1} S_0\Big\| + 6\Big\|\mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_0 - \widetilde{\boldsymbol{\beta}})\widetilde{\boldsymbol{\beta}}^\top\big]\Big\| + \Big\|\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_0 - \widetilde{\boldsymbol{\beta}}\big]\Big\|_2^2.$$

In short, $\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big)$ converges exponentially fast to the limit $(\mathrm{id} - S_{\mathrm{lin}})^{-1} S_0$. Combining the generalized Gauss-Markov Theorem 2 with Theorem 5 also establishes

$$\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k\big) \to \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}\big) + (\mathrm{id} - S_{\mathrm{lin}})^{-1} S_0, \qquad \text{as } k \to \infty,$$

with exponential rate of convergence. Recall the intuition gained from Theorem 2 that $\widetilde{\boldsymbol{\beta}}_k$ may be decomposed into a sum of $\widetilde{\boldsymbol{\beta}}$ and (approximately) orthogonal centered noise. We now conclude that up to exponentially decaying terms, the covariance matrix of this orthogonal noise must be given by $(\mathrm{id} - S_{\mathrm{lin}})^{-1} S_0$, which fully describes the (asymptotic) gap between the covariance matrices of $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}_k$.

Taking the trace and noting that $\big|\mathrm{Tr}(A)\big| \leq d\|A\|$, we obtain a bound for the convergence of $\widetilde{\boldsymbol{\beta}}_k$ with respect to the squared Euclidean loss,

$$\left|\mathbb{E}\Big[\big\|\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big\|_2^2\Big] - \mathrm{Tr}\Big(\big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1} S_0\Big)\right| \leq Cdk\big\|I - \alpha p \mathbb{X}_p\big\|^{k-1}. \tag{17}$$

Since $(\mathrm{id} - S_{\mathrm{lin}})^{-1} S_0$ is a $d \times d$ matrix, the term $\mathrm{Tr}\big((\mathrm{id} - S_{\mathrm{lin}})^{-1} S_0\big)$ describing the asymptotic discrepancy between $\widetilde{\boldsymbol{\beta}}_k$ and $\widetilde{\boldsymbol{\beta}}$ can be large in high dimensions d, even if the spectral norm of $(\mathrm{id} - S_{\mathrm{lin}})^{-1} S_0$ is small. Since $\mathrm{id} - S_{\mathrm{lin}}$ is a positive definite operator, the matrix $(\mathrm{id} - S_{\mathrm{lin}})^{-1} S_0$ is zero if, and only if, S_0 is zero. By (16), $S_0 = \mathbb{E}[\boldsymbol{\xi}_k \boldsymbol{\xi}_k^\top]$. The explicit form $\boldsymbol{\xi}_k = \alpha D_k \overline{\mathbb{X}}\big(pI - D_k\big)\widetilde{\boldsymbol{\beta}}$, shows that $\boldsymbol{\xi}_k = 0$ and $S_0 = 0$ provided that $\overline{\mathbb{X}} = 0$, meaning whenever $\mathbb{X}$ is diagonal. To give a more precise quantification, we show that the operator norm of $(\mathrm{id} - S_{\mathrm{lin}})^{-1} S_0$ is of order $\alpha p/(1-p)^2$.

Lemma 6 *In addition to Assumption 1 suppose $\alpha < \frac{\lambda_{\min}(\mathbb{X}_p)}{3\|\mathbb{X}\|^2}$, then, for any $k = 1, 2, \ldots$*

$$\Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k\big) - \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}\big)\Big\| \leq \frac{k\|I - \alpha p \mathbb{X}_p\|^{k-1} C' + \alpha p C''}{(1-p)^2}$$

and

$$\Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k\big) - \mathrm{Diag}(\mathbb{X})^{-1}\mathbb{X}\mathrm{Diag}(\mathbb{X})^{-1}\Big\| \leq \frac{k\|I - \alpha p \mathbb{X}_p\|^{k-1} C' + p(1+\alpha) C''}{(1-p)^2},$$

where C' and C'' are constants that are independent of (α, p, k).

The first bound describes the interplay between αp and k. Making αp small will decrease the second term in the bound, but conversely requires a larger number of iterations k for the first term to decay.

In the second bound, $\mathrm{Diag}(\mathbb{X})^{-1}\mathbb{X}\mathrm{Diag}(\mathbb{X})^{-1}$ matches the covariance matrix $\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}\big)$ of the marginalized loss minimizer $\widetilde{\boldsymbol{\beta}}$ up to a term of order p. Consequently, the covariance structures induced by dropout and ℓ_2-regularization approximately coincide for sufficiently small p. However, in this regime we have $\mathbb{X}_p = p\mathbb{X} + (1-p)\mathrm{Diag}(\mathbb{X}) \approx \mathrm{Diag}(\mathbb{X})$, and $\widetilde{\boldsymbol{\beta}} = \mathbb{X}_p^{-1}X^\top\mathbf{Y} \approx \mathrm{Diag}(\mathbb{X})^{-1}X^\top\mathbf{Y}$ becomes extremely biased whenever the Gram matrix $\mathbb{X}$ is not diagonal.

Theorem 2 already establishes $\mathrm{Cov}(\widetilde{\boldsymbol{\beta}})$ as the optimal covariance among all affine estimators that are asymptotically unbiased for $\widetilde{\boldsymbol{\beta}}$. To conclude our study of the gap between $\mathrm{Cov}(\widetilde{\boldsymbol{\beta}}_k)$ and $\mathrm{Cov}(\widetilde{\boldsymbol{\beta}})$, we provide a lower-bound.

Theorem 7 (Sub-Optimality of Variance) *In addition to the assumptions of Theorem 5, suppose for every $\ell = 1, \dots, d$ there exists $m \neq \ell$ such that $\mathbb{X}_{\ell m} \neq 0$, then*

$$\lim_{k\to\infty} \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k\big) - \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}\big) \geq \frac{\alpha p(1-p)^2\lambda_{\min}(\mathbb{X})}{2\|\mathbb{X}\|^3} \min_{i\neq j:\mathbb{X}_{ij}\neq 0} \mathbb{X}_{ij}^2 \cdot I_d.$$

The lower-bound is positive whenever $\mathbb{X}$ is invertible. In general, Theorem 7 entails asymptotic statistical sub-optimality of the gradient descent iterates $\widetilde{\boldsymbol{\beta}}_k$ for a large class of design matrices. Moreover, the result does not require any further assumptions on the tuning parameters α and p, other than α being sufficiently small.

To summarize, compared with the marginalized loss minimizer $\widetilde{\boldsymbol{\beta}}$, the covariance matrix of the gradient descent iterates with dropout may be larger. The difference may be significant, especially if the data dimension d is large. Proving the results requires a refined second moment analysis, based on explicit computation of the dynamics of $\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}$. Simple heuristics such as (7) do not fully reveal the properties of the underlying dynamics.

4.1.1 Ruppert-Polyak Averaging

To reduce the gradient noise induced by dropout, one may consider the running average over the gradient descent iterates. This technique is also known as Ruppert-Polyak averaging (Ruppert, 1988; Polyak, 1990) The k^{th} Ruppert-Polyak average of the gradient descent iterates is given by

$$\widetilde{\boldsymbol{\beta}}_k^{\mathrm{rp}} := \frac{1}{k}\sum_{\ell=1}^{k} \widetilde{\boldsymbol{\beta}}_\ell.$$

Averages of this type are well-studied in the stochastic approximation literature, see Polyak and Juditsky (1992); Györfi and Walk (1996) for results on linear regression and Zhu et al. (2021); Dereich and Kassing (2023) for stochastic gradient descent. The next theorem illustrates convergence of $\widetilde{\boldsymbol{\beta}}_k^{\mathrm{rp}}$ towards $\widetilde{\boldsymbol{\beta}}$.

Theorem 8 *In addition to Assumption 1, suppose $\alpha < \frac{\lambda_{\min}(\mathbb{X}_p)}{3\|\mathbb{X}\|^2}$, then, for any $k = 1, 2, \dots$*

$$\Big\|\mathbb{E}\Big[\big(\widetilde{\boldsymbol{\beta}}_k^{\mathrm{rp}} - \widetilde{\boldsymbol{\beta}}\big)\big(\widetilde{\boldsymbol{\beta}}_k^{\mathrm{rp}} - \widetilde{\boldsymbol{\beta}}\big)^\top\Big]\Big\| \leq \frac{2\|\mathbb{X}\|^2 \cdot \big\|\mathbb{E}[X^\top\mathbf{Y}\mathbf{Y}^\top X]\big\|}{k(1-p)(\min_i \mathbb{X}_{ii})^4} + \frac{2C}{k^2(\alpha p(1-p)\min_i \mathbb{X}_{ii})^3},$$

where C is the constant from Theorem 5.

The first term in the upper bound is independent of α and decays at the rate k^{-1}, whereas the second term scales with $(\alpha p)^{-3}k^{-2}$. Accordingly, for small αp, the second term will dominate initially, until k grows sufficiently large.

Since the right hand side eventually tends to zero, the theorem implies convergence of the Ruppert-Polyak averaged iterates to the marginalized loss minimizer $\widetilde{\boldsymbol{\beta}}$, so the link between dropout and ℓ_2-regularization persists at the variance level. The averaging comes at the price of a slower convergence rate k^{-1} of the remainder terms, as opposed to the exponentially fast convergence in Theorem 5. As in (17), the bound can be converted into a convergence rate for $\mathbb{E}\big[\|\widetilde{\boldsymbol{\beta}}_k^{\mathrm{rp}} - \widetilde{\boldsymbol{\beta}}\|_2^2\big]$ by taking the trace,

$$\mathbb{E}\Big[\big\|\widetilde{\boldsymbol{\beta}}_k^{\mathrm{rp}} - \widetilde{\boldsymbol{\beta}}\big\|_2^2\Big] \leq d\left(\frac{2\|\mathbb{X}\|^2 \cdot \big\|\mathbb{E}[X^\top \mathbf{Y}\mathbf{Y}^\top X]\big\|}{k(1-p)(\min_i \mathbb{X}_{ii})^4} + \frac{2C}{k^2(\alpha p(1-p)\min_i \mathbb{X}_{ii})^3}\right).$$

4.2 Convergence of Simplified Dropout

To further illustrate how dropout and gradient descent are coupled, we will now study the simplified dropout iterates

$$\widehat{\boldsymbol{\beta}}_k = \widehat{\boldsymbol{\beta}}_{k-1} + \alpha D_k X^\top \big(\mathbf{Y} - X\widehat{\boldsymbol{\beta}}_{k-1}\big), \tag{18}$$

as defined in (4). While the original dropout reduces the model before taking the gradient, this version takes the gradient first and applies dropout afterwards. As shown in Section 2, both versions coincide if the Gram matrix $\mathbb{X}$ is diagonal. Recall from the discussion preceding Lemma 6 that for diagonal $\mathbb{X}$, $\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k\big)$ converges to the optimal covariance matrix. This suggests that for the simplified dropout, no additional randomness in the limit $k \to \infty$ occurs.

The least squares objective $\boldsymbol{\beta} \mapsto \|\mathbf{Y} - X\boldsymbol{\beta}\|_2^2$ always admits a minimizer, with any minimizer $\widehat{\boldsymbol{\beta}}$ necessarily solving the so-called normal equations $X^\top \mathbf{Y} = \mathbb{X}\widehat{\boldsymbol{\beta}}$. Provided $\mathbb{X}$ is invertible, the least-squares estimator $\widehat{\boldsymbol{\beta}} = \mathbb{X}^{-1}X^\top \mathbf{Y}$ gives the unique solution. We will not assume invertibility for all results below, so we let $\widehat{\boldsymbol{\beta}}$ denote any solution of the normal equations, unless specified otherwise. In turn, (18) may be rewritten as

$$\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}} = \big(I - \alpha D_k \mathbb{X}\big)\big(\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}\big), \tag{19}$$

which is simpler than the analogous representation of $\widetilde{\boldsymbol{\beta}}_k$ as a VAR process in (15).

As a first result, we will show that the expectation of the simplified dropout iterates $\widehat{\boldsymbol{\beta}}_k$ converges to the mean of the unregularized least squares estimator $\widehat{\boldsymbol{\beta}}$, provided that $\mathbb{X}$ is invertible. Indeed, using (19), independence of D_k and $\big(\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}\big)$, and $\mathbb{E}\big[D_k\big] = pI$, observe that

$$\mathbb{E}\big[\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big] = \mathbb{E}\Big[\big(I - \alpha D_k \mathbb{X}\big)\big(\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}\big)\Big] = \big(I - \alpha p \mathbb{X}\big)\mathbb{E}\big[\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}\big]. \tag{20}$$

Induction on k now shows $\mathbb{E}\big[\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big] = \big(I - \alpha p \mathbb{X}\big)^k \mathbb{E}\big[\widehat{\boldsymbol{\beta}}_0 - \widehat{\boldsymbol{\beta}}\big]$ and so

$$\Big\|\mathbb{E}\big[\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big]\Big\|_2 = \Big\|\big(I - \alpha p \mathbb{X}\big)^k \mathbb{E}\big[\widehat{\boldsymbol{\beta}}_0 - \widehat{\boldsymbol{\beta}}\big]\Big\|_2 \leq \big\|I - \alpha p \mathbb{X}\big\|^k \Big\|\mathbb{E}\big[\widehat{\boldsymbol{\beta}}_0 - \widehat{\boldsymbol{\beta}}\big]\Big\|_2.$$

Assuming $\alpha p\|\mathbb{X}\| < 1$, invertibility of $\mathbb{X}$ implies $\|I - \alpha p\mathbb{X}\| < 1$. Consequently, the convergence is exponential in the number of iterations.

Invertibility of $\mathbb{X}$ may be avoided if the initialization $\widehat{\boldsymbol{\beta}}_0$ lies in the orthogonal complement of the kernel of $\mathbb{X}$ and $\widehat{\boldsymbol{\beta}}$ is the $\|\cdot\|_2$-minimal solution to the normal equations. We can then argue that $(I - \alpha p\mathbb{X})^{k-1}\mathbb{E}[\widehat{\boldsymbol{\beta}}_0 - \widehat{\boldsymbol{\beta}}]$ always stays in a linear subspace on which $(I - \alpha p\mathbb{X})$ still acts as a contraction.

We continue with our study of $\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}$ by employing the same techniques as in the previous section to analyze the second moment. The linear operator

$$T(A) := \big(I - \alpha p\mathbb{X}\big)A\big(I - \alpha p\mathbb{X}\big) + \alpha^2 p(1-p)\mathrm{Diag}\big(\mathbb{X}A\mathbb{X}\big), \tag{21}$$

defined on $d \times d$ matrices, takes over the role of the affine operator S encountered in Lemma 13. In particular, the second moments $A_k := \mathbb{E}\big[(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}})^\top\big]$ evolve as the linear dynamical system

$$A_k = T\big(A_{k-1}\big), \qquad k = 1, 2, \ldots \tag{22}$$

without remainder terms. To see this, observe via (19) the identity $(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}})^\top = (I - \alpha D_k\mathbb{X})(\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}})^\top(I - \alpha\mathbb{X}D_k)$. Taking the expectation on both sides, conditioning on D_k, and recalling that D_k is independent of $(\widehat{\boldsymbol{\beta}}_k, \widehat{\boldsymbol{\beta}})$ gives $A_k = \mathbb{E}\big[(I - \alpha D_k\mathbb{X})A_{k-1}(I - \alpha\mathbb{X}D_k)\big]$. We have $\mathbb{E}[D_k] = pI_d$ and by Lemma 16, $\mathbb{E}\big[D_k\mathbb{X}A_{k-1}\mathbb{X}D_k\big] = p(\mathbb{X}A_{k-1}\mathbb{X})_p = p^2\mathbb{X}A_{k-1}\mathbb{X} + p(1-p)\mathrm{Diag}(\mathbb{X}A_{k-1}\mathbb{X})$. Together with the definition of $T(A)$, this proves (22).

Further results are based on analyzing the recursion in (22). It turns out that convergence of $\widehat{\boldsymbol{\beta}}_k$ to $\widehat{\boldsymbol{\beta}}$ in second mean requires a non-singular Gram matrix $\mathbb{X}$.

Lemma 9 *Suppose the initialization $\widehat{\boldsymbol{\beta}}_0$ is independent of all other sources of randomness and the number of parameters satisfies $d \geq 2$, then there exists a singular $\mathbb{X}$, such that for any positive integer k, $\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_k\big) \geq \mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big) + \mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}\big)$ and*

$$\Big\|\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big)\Big\| \geq \alpha^2 p(1-p).$$

For invertible $\mathbb{X}$, we can apply Theorem 2 to show that $\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}\big) = \mathbb{X}^{-1}$ is the optimal covariance matrix for the sequence of affine estimators $\widehat{\boldsymbol{\beta}}_k$. The simplified dropout iterates actually achieve the optimal variance when $\mathbb{X}$ is invertible, which stands in contrast with the situation in Theorem 7.

Theorem 10 *Suppose $\mathbb{X}$ is invertible, $\alpha \leq \min\Big\{\frac{1}{p\|\mathbb{X}\|}, \frac{\lambda_{\min}(\mathbb{X})}{\|\mathbb{X}\|^2}\Big\}$, and let $\widehat{\boldsymbol{\beta}}_0$ be square-integrable, then, for any $k = 1, 2, \ldots$*

$$\Big\|\mathbb{E}\Big[\big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big)\big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big)^\top\Big]\Big\| \leq \big(1 - \alpha p\lambda_{\min}(\mathbb{X})\big)^k\Big\|\mathbb{E}\Big[\big(\widehat{\boldsymbol{\beta}}_0 - \widehat{\boldsymbol{\beta}}\big)\big(\widehat{\boldsymbol{\beta}}_0 - \widehat{\boldsymbol{\beta}}\big)^\top\Big]\Big\|.$$

Intuitively, the result holds due to the operator T in (21) being linear, as opposed to affine like in the case of Lemma 4. Choosing α sufficiently small ensures that T acts as a contraction, meaning $T^k(A) \to 0$ for any matrix A. Hence, linearity of T serves as an algebraic expression of the simplified dynamics. As in (17), we may take the trace to obtain the bound

$$\mathbb{E}\Big[\big\|\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big\|_2^2\Big] \leq d\big(1 - \alpha p\lambda_{\min}(\mathbb{X})\big)^k\Big\|\mathbb{E}\Big[\big(\widehat{\boldsymbol{\beta}}_0 - \widehat{\boldsymbol{\beta}}\big)\big(\widehat{\boldsymbol{\beta}}_0 - \widehat{\boldsymbol{\beta}}\big)^\top\Big]\Big\|.$$

5. Discussion and Outlook

Our main contributions may be summarized as follows: We studied dropout in the linear regression model, but unlike previous results, we explicitly analyzed the gradient descent dynamics with new dropout noise being sampled in each iteration. This allows us to characterize the limiting variance of the gradient descent iterates exactly (Theorem 5), which sheds light on the covariance structure induced via dropout. Our main tool in the analysis is a particular recursion (Lemma 4), which may be exhibited by "anchoring" the gradient descent iterates around the marginalized loss minimizer $\widetilde{\boldsymbol{\beta}}$. To further understand the interaction between noise and gradient descent dynamics, we analyze the running average of the process (Theorem 8) and a simplified version of dropout (Theorem 10).

We view our analysis of the linear model as a fundamental first step towards understanding the dynamics of gradient descent with dropout. Analyzing the linear model has been a fruitful approach to study other phenomena in deep learning, such as overfitting (Tsigler and Bartlett, 2023), sharpness of local minima (Bartlett et al., 2023), and in-context learning (Zhang et al., 2024). We conclude by proposing some natural directions for future work.

Random minibatch sampling: For yet another way of incorporating dropout, we may compute the gradient based on a random subset of the data (mini batches). In this case, the updating formula satisfies

$$\overline{\boldsymbol{\beta}}_{k+1} = \overline{\boldsymbol{\beta}}_k - \alpha \nabla_{\overline{\boldsymbol{\beta}}_k} \frac{1}{2} \Big\| D_{k+1}\big(\mathbf{Y} - X\overline{\boldsymbol{\beta}}_k\big) \Big\|_2^2 = \overline{\boldsymbol{\beta}}_k + \alpha X^\top D_{k+1}\big(\mathbf{Y} - X\overline{\boldsymbol{\beta}}_k\big). \tag{23}$$

The dropout matrices are now of dimension $n \times n$ and select a random subset of the data points in every iteration. This version of dropout also relates to randomly weighted least squares and resampling methods (Dümbgen et al., 2013). The update formula may be written in the form $\overline{\boldsymbol{\beta}}_{k+1} - \widehat{\boldsymbol{\beta}} = (I - \alpha X^\top D_{k+1} X)(\overline{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}) + \alpha X^\top D_{k+1}(\mathbf{Y} - X\widehat{\boldsymbol{\beta}})$ with $\widehat{\boldsymbol{\beta}}$ solving the normal equations $X^\top \mathbf{Y} = \mathbb{X}\widehat{\boldsymbol{\beta}}$. Similarly to the corresponding reformulation (14) of the original dropout scheme, this defines a vector autoregressive process with random coefficients and lag one.

Learning rates: The proof ideas may be generalized to a sequence of iteration-dependent learning rates α_k. We expect this to come at the cost of more involved formulas. Specifically, the operator S in Lemma 4 will depend on the iteration number through α_k, so the limit in Theorem 5 cannot be expressed as $(\mathrm{id} - S_{\mathrm{lin}})^{-1} S_0$ anymore.

Random design and stochastic gradients: We considered a fixed design matrix X and (full) gradient descent, whereas in machine learning practice inputs are typically assumed to be random and parameters are updated via stochastic gradient descent (SGD). The recent works Bos and Schmidt-Hieber (2023); Schmidt-Hieber and Koolen (2023) derive convergence rates for SGD considering linear regression and another form of noisy gradient descent. We believe that parts of theses analyses carry over to dropout.

Generic dropout distributions: As already mentioned, Srivastava et al. (2014) carry out simulations with dropout where $D_{ii} \overset{i.i.d.}{\sim} \mathcal{N}(1,1)$. Gaussian dropout distributions are currently supported, or easily implemented, in major software libraries (Abadi et al., 2016; Chollet et al., 2015; Jia et al., 2014; Paszke et al., 2019). Analyzing a generic dropout distribution with mean μ and variance σ may also paint a clearer picture of how the dropout noise interacts with the gradient descent dynamics. For the linear regression model, results

that marginalize over the dropout noise generalize to arbitrary dropout distribution. In particular, (9) turns into

$$\begin{aligned}\widetilde{\boldsymbol{\beta}} :=& \arg\min_{\boldsymbol{\beta}} \mathbb{E}\Big[\big\|\mathbf{Y} - XD\boldsymbol{\beta}\big\|_2^2 \mid \mathbf{Y}\Big] \\ =& \arg\min_{\boldsymbol{\beta}} \big\|\mathbf{Y} - \mu X\boldsymbol{\beta}\big\|_2^2 + \sigma^2\boldsymbol{\beta}^\top \mathrm{Diag}(\mathbb{X})\boldsymbol{\beta} \\ =& \mu\big(\mu^2\mathbb{X} + \sigma^2\mathrm{Diag}(\mathbb{X})\big)^{-1}X^\top\mathbf{Y}.\end{aligned}$$

If $D_{ii} \overset{i.i.d.}{\sim} \mathcal{N}(1,1)$, then $\widetilde{\boldsymbol{\beta}} = \big(\mathbb{X} + \mathrm{Diag}(\mathbb{X})\big)^{-1}X^\top\mathbf{Y}$.

In contrast, treatment of the corresponding iterative dropout scheme seems more involved. The analysis of dropout with Bernoulli distributions relies in parts on the projection property $D^2 = D$. Without it, additional terms occur in the moments in Lemma 16, which is required for the computation of the covariance matrix. For example, the formula $\mathbb{E}\big[DADBD\big] = pA_pB_p + p^2(1-p)\mathrm{Diag}(\overline{A}B)$ turns into

$$\begin{aligned}\mathbb{E}\big[DADBD\big] =& \frac{1}{\mu_1}\big(\mu_1^2 A + \sigma^2\mathrm{Diag}(A)\big)\big(\mu_1^2 B + \sigma^2\mathrm{Diag}(B)\big) + \sigma^2\mu_1\mathrm{Diag}(\overline{A}\,\overline{B}) \\ &+ \left(\mu_3 - \frac{\mu_2^2}{\mu_1}\right)\mathrm{Diag}(A)\mathrm{Diag}(B),\end{aligned}$$

where μ_r denotes the r^{th} moment of the dropout distribution and σ^2 its variance. For the Bernoulli distribution, all moments equal p, so $\mu_3 - \mu_2^2/\mu_1 = 0$ and the last term disappears. Similarly, more terms will appear in the fourth moment of D, making the expression for the operator corresponding to S in Lemma 4 more complicated.

Inducing robustness via dropout: Among the possible ways of explaining the data, dropout should, by design, favor an explanation that is robust against setting a random subset of the parameters to zero. Mianjy et al. (2018) indicate that dropout in two-layer linear networks tends to equalize the norms of different weight vectors.

To study the robustness properties of dropout, one may suggest analysis of loss functions measuring prediction of the response vector if each estimated regression coefficients is deleted with probability p. Given an estimator $\widehat{\boldsymbol{\beta}}$, a natural choice would be the loss

$$L\big(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}_\star\big) := \mathbb{E}\Big[\big\|X(D\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star)\big\|_2^2 \mid \mathbf{Y}\Big] = \big(p\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star\big)^\top\mathbb{X}\big(p\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star\big) + p(1-p)\widehat{\boldsymbol{\beta}}^\top\mathrm{Diag}\big(\mathbb{X}\big)\widehat{\boldsymbol{\beta}},$$

with D a new draw of the dropout matrix, independent of all other randomness. This loss depends on the unknown true regression vector $\boldsymbol{\beta}_\star$. Since $\mathbb{E}[\mathbf{Y}] = X\boldsymbol{\beta}_\star$, an empirical version of the loss may replace $X\boldsymbol{\beta}_\star$ with $\mathbf{Y}$, considering $\mathbb{E}\big[\|XD\widehat{\boldsymbol{\beta}} - \mathbf{Y}\|_2^2 \mid \mathbf{Y}\big]$. As shown in (9), $\widetilde{\boldsymbol{\beta}} = \mathbb{X}_p^{-1}X^\top\mathbf{Y}$ minimizes this loss function. This suggests that $\widetilde{\boldsymbol{\beta}}$ may possess some optimality properties for the loss $L(\,\cdot\,, \boldsymbol{\beta}_\star)$ defined above.

Shallow networks with linear activation function: Multi-layer neural networks do not admit unique minimizers. In comparison with the linear regression model, this poses a major challenge for the analysis of dropout. Mianjy et al. (2018) consider shallow linear networks of the form $f(\mathbf{x}) = UV^\top\mathbf{x}$ with $U = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$ an $n \times m$ matrix and $V = (\mathbf{v}_1, \ldots, \mathbf{v}_m)$ a $d \times m$ matrix. Suppose D is an $m \times m$ dropout matrix. Assuming

the random design vector $\mathbf{X}$ satisfies $\mathbb{E}\big[\mathbf{X}\mathbf{X}^\top\big] = I_d$, and marginalizing over dropout noise applied to the columns of U (or equivalently to the rows of V) leads to an ℓ_2-penalty via

$$\begin{aligned}
&\mathbb{E}\Big[\big\|\mathbf{Y} - p^{-1}UDV^\top\mathbf{X}\big\|_2^2 \mid \mathbf{Y}\Big] \\
&= \mathbb{E}\Big[\big\|\mathbf{Y} - UV^\top\mathbf{X}\big\|_2^2 \mid \mathbf{Y}\Big] + \frac{1-p}{p}\sum_{i=1}^m \|\mathbf{u}_i\|_2^2\|\mathbf{v}_i\|_2^2 \\
&= \mathbb{E}\Big[\big\|\mathbf{Y} - UV^\top\mathbf{X}\big\|_2^2 \mid \mathbf{Y}\Big] + \frac{1-p}{p}\mathrm{Tr}\Big(\mathrm{Diag}(U^\top U)\mathrm{Diag}(V^\top V)\Big).
\end{aligned} \tag{24}$$

As an extension of our approach, it seems natural to investigate whether gradient descent with dropout will converge to the same minimizer or involve additional terms in the variance. In contrast with linear regression, the marginalized loss (24) is non-convex and does not admit a unique minimizer. Hence, we cannot simply center the gradient descent iterates around a specific closed-form estimator, as in Section 4. To extend our techniques we may expect to replace the centering estimator with the gradient descent iterates for the marginalized loss function, demanding a more careful analysis.

Senen-Cerda and Sanders (2022) study the gradient flow associated with (24) and exhibit exponential convergence of U and V towards a minimizer. Extending the existing result on gradient flows to gradient descent is, however, non-trivial, see for example the gradient descent version of Theorem 3.1 in Bah et al. (2022) provided in Theorem 2.4 of Nguegnang et al. (2021).

To be more precise, suppose $\mathbf{Y} = W_\star\mathbf{X} + \boldsymbol{\varepsilon}$, where $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ are independent random vectors, so the task reduces to learning a factorization $W_\star \approx UV^\top$ based on noisy evaluation of $W_\star\mathbf{X}$. Consider the randomized loss

$$L\big(U,V\big) \mapsto \frac{1}{2}\Big\|\mathbf{Y} - p^{-1}UDV^\top\mathbf{X}\Big\|_2^2,$$

with respective gradients

$$\begin{aligned}
\nabla_U L\big(U,V\big) &= -\big(\mathbf{Y} - p^{-1}UDV^\top\mathbf{X}\big)\mathbf{X}^\top V p^{-1}D \\
\nabla_{V^\top} L\big(U,V\big) &= -p^{-1}DU^\top\big(\mathbf{Y} - Up^{-1}DV^\top\mathbf{X}\big)\mathbf{X}^\top.
\end{aligned}$$

Given observations $(\mathbf{Y}_k, \mathbf{X}_k) \overset{i.i.d.}{\sim} (\mathbf{Y}, \mathbf{X})$ and independent dropout matrices $D_k \overset{i.i.d.}{\sim} D$, the factorized structure leads to two coupled dynamical systems $U_{k+1} = U_k - \alpha\nabla_{U_k}L(U_k, V_k)$ and $V_{k+1}^\top = V_k^\top - \alpha\nabla_{V_k^\top}L(U_k, V_k)$, which are linked through the appearance of V_k in $\nabla_{U_k}L(U_k, V_k)$ and U_k in $\nabla_{V_k}L(U_k, V_k)$. Due to non-convexity of the underlying marginalized objective (24), the resulting dynamics should be sensitive to initialization. Suppose $U_k = P_k(U_0, V_0)$ and $V_k = Q_k(U_0, V_0)$ are given as random matrix polynomials (P_k, Q_k) in (U_0, V_0), meaning finite sums of expressions like $A_1X_{i_1}A_2X_{i_2}\cdots A_nX_{i_n}A_{n+1}$, where $X_{i_j} \in \{U_0, U_0^\top, V_0, V_0^\top\}$ and the A_j are random coefficient matrices. Now, the gradient descent recursions lead to

$$\begin{aligned}
U_{k+1} &= P_k(U_0,V_0) + \alpha\Big(\mathbf{Y}_k - P_k(U_0,V_0)p^{-1}D_kQ_k^\top(U_0,V_0)\mathbf{X}_k\Big)\mathbf{X}_k^\top Q_k(U_0,V_0)p^{-1}D_k \\
&=: P_{k+1}(U_0,V_0), \\
V_{k+1}^\top &= Q_k^\top(U_0,V_0) + \alpha p^{-1}D_kP_k^\top(U_0,V_0)\Big(\mathbf{Y}_k - P_k(U_0,V_0)p^{-1}D_kQ_k^\top(U_0,V_0)\mathbf{X}_k\Big)\mathbf{X}_k^\top \\
&=: Q_{k+1}^\top(U_0,V_0),
\end{aligned} \tag{25}$$

so (U_{k+1}, V_{k+1}) is also a polynomial in (U_0, V_0) with random coefficients. A difficulty in analyzing this recursion is that the degree of the polynomial increases exponentially fast. Indeed, since P_{k+1} includes the term $p^{-2}P_kD_kQ_k^\top\mathbf{X}_k\mathbf{X}_k^\top Q_kD_k$, the degree of P_{k+1} is the degree of P_k plus twice the degree of Q_k. During each gradient descent step, additional randomness is introduced via the newly sampled dropout matrix D_k and the training data $(\mathbf{Y}_k, \mathbf{X}_k)$. Accordingly, the coefficients of P_{k+1} and Q_{k+1} fluctuate around the coefficients of $\mathbb{E}[P_{k+1} \mid U_k, V_k]$ and $\mathbb{E}[Q_{k+1} \mid U_k, V_k]$. A principled analysis of the resulting dynamics requires careful accounting of how these fluctuations propagate through the iterations. Senen-Cerda and Sanders (2022) show that the gradient flow trajectories and minimizers of (24) satisfy specific symmetries, so one should hope to reduce the algebraic complexity of the problem by finding analogous symmetries in the stochastic recursions (25).

Alternatively, one may consider layer-wise training of the weight matrices to break the dependence between U_k and V_k. Given $K_1 > 0$, suppose we keep U_0 fixed while taking K_1 gradient steps

$$V_{k+1}^\top = V_k^\top + \alpha p^{-1}D_kU_0^\top\big(\mathbf{Y} - U_0p^{-1}D_kV_k^\top\mathbf{X}_k\big)\mathbf{X}_k^\top$$

followed by $K_2 > 0$ gradient steps of the form

$$U_{k+1} = U_k + \alpha\big(\mathbf{Y} - U_kp^{-1}D_kV_{K_1}^\top\mathbf{X}_k\big)\mathbf{X}_k^\top V_{K_1}p^{-1}D_k.$$

In each phase, the gradient descent recursion solves a linear regression problem similar to to our analysis of the linear model. We leave the details for future work.

Acknowledgments

The authors thank the editor and three anonymous referees for their valuable time and effort; their comments improved the article tremendously.

This publication is part of the project *Statistical foundation for multilayer neural networks* (project number VI.Vidi.192.021 of the Vidi ENW programme) financed by the Dutch Research Council (NWO).

Appendix A. Proofs for Section 3

A.1 Proof of Equation (10)

Recall the definition $\widetilde{\boldsymbol{\beta}} = \mathbb{X}_p^{-1}X^\top\mathbf{Y}$. By Lemma 16(a), $\mathbb{E}[D\mathbb{X}D] = p\mathbb{X}_p$ and so

$$\mathbb{E}\Big[DX^\top\big(\mathbf{Y} - XD\widetilde{\boldsymbol{\beta}}\big) \mid \mathbf{Y}\Big] = pX^\top\mathbf{Y} - p\mathbb{X}_p\widetilde{\boldsymbol{\beta}} = 0. \tag{26}$$

Note the identity $\mathbf{Y} = XD\widetilde{\boldsymbol{\beta}} + (\mathbf{Y} - XD\widetilde{\boldsymbol{\beta}})$. Hence, if $\widehat{\boldsymbol{\beta}}$ denotes any estimator, $XD\widehat{\boldsymbol{\beta}} - \mathbf{Y} = XD(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}) - (\mathbf{Y} - XD\widetilde{\boldsymbol{\beta}})$. By (26), $\mathbb{E}\big[(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})^\top DX^\top(\mathbf{Y} - XD\widetilde{\boldsymbol{\beta}}) \mid \mathbf{Y}\big] = 0$ and

$$\begin{aligned}\mathbb{E}\Big[\big\|XD\widehat{\boldsymbol{\beta}} - \mathbf{Y}\big\|_2^2 \mid \mathbf{Y}\Big] &= \mathbb{E}\Big[\big\|XD\widetilde{\boldsymbol{\beta}} - \mathbf{Y}\big\|_2^2 \mid \mathbf{Y}\Big] + \mathbb{E}\Big[\big\|XD\big(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\big)\big\|_2^2 \mid \mathbf{Y}\Big] \\ &\quad - 2\mathbb{E}\Big[\big(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\big)^\top DX^\top\big(\mathbf{Y} - XD\widetilde{\boldsymbol{\beta}}\big) \mid \mathbf{Y}\Big] \\ &= \mathbb{E}\Big[\big\|XD\widetilde{\boldsymbol{\beta}} - \mathbf{Y}\big\|_2^2 \mid \mathbf{Y}\Big] + \mathbb{E}\Big[\big\|XD\big(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\big)\big\|_2^2 \mid \mathbf{Y}\Big],\end{aligned}$$

which is the claimed expression. ■

A.2 Proof of Equation (11)

Let $r := \operatorname{rank}(X)$ and consider a singular value decomposition $X = \sum_{\ell=1}^{r} \sigma_\ell \mathbf{v}_\ell \mathbf{w}_\ell^\top$. The left-singular vectors $\mathbf{w}_\ell$, $\ell = 1, \dots, d$ are orthonormal, meaning $\mathbb{X} = \sum_{\ell=1}^{r} \sigma_\ell^2 \mathbf{w}_\ell \mathbf{w}_\ell^\top$ and since $\operatorname{Diag}(\mathbb{X}) = \mathbb{X}_{11} \cdot I_d$,

$$\mathbb{X}_p = p\left(\sum_{\ell=1}^{r} \sigma_\ell^2 \mathbf{w}_\ell \mathbf{w}_\ell^\top\right) + (1-p)\mathbb{X}_{11} I_d.$$

Each left-singular vector $\mathbf{w}_\ell$ is an eigenvector of $\mathbb{X}_p$, with associated eigenvalue $p\sigma_\ell^2 + (1-p)\mathbb{X}_{11}$. If $r < d$, suppose $\mathbf{w}_m$, $m = r+1, \dots, d$ complete the orthonormal basis, then $\mathbb{X}_p \mathbf{w}_m = (1-p)\mathbb{X}_{11}\mathbf{w}_m$. Consequently, $\mathbb{X}_p^{-1}$ admits $\mathbf{w}_\ell$ as an eigenvector for every $\ell = 1, \dots, d$ and

$$\mathbb{X}_p^{-1} = \sum_{\ell=1}^{r} \frac{1}{p\sigma_\ell^2 + (1-p)\mathbb{X}_{11}} \mathbf{w}_\ell \mathbf{w}_\ell^\top + \sum_{m=r+1}^{d} \frac{1}{(1-p)\mathbb{X}_{11}} \mathbf{w}_m \mathbf{w}_m^\top. \tag{27}$$

By definition, $X^\top \mathbf{Y} = \sum_{\ell=1}^{r} \sigma_\ell \mathbf{w}_\ell \mathbf{v}_\ell^\top \mathbf{Y}$ and $\mathbf{Y} = \sum_{\ell=1}^{r} \sigma_\ell \mathbf{v}_\ell \mathbf{w}_\ell^\top \boldsymbol{\beta}_\star + \boldsymbol{\varepsilon}$. Combining these facts now leads to

$$\widetilde{\boldsymbol{\beta}} = \sum_{\ell=1}^{r} \frac{\sigma_\ell}{p\sigma_\ell^2 + (1-p)\mathbb{X}_{11}} \mathbf{w}_\ell \mathbf{v}_\ell^\top \mathbf{Y}$$

and

$$\begin{aligned}
\mathbb{E}\big[X\widetilde{\boldsymbol{\beta}}\big] &= \sum_{\ell=1}^{r} \frac{\sigma_\ell^2}{p\sigma_\ell^2 + (1-p)\mathbb{X}_{11}} \mathbf{v}_\ell \mathbf{v}_\ell^\top \mathbb{E}\big[\mathbf{Y}\big] \\
&= \sum_{\ell=1}^{r} \frac{1}{p + (1-p)\mathbb{X}_{11}/\sigma_\ell^2} \mathbf{v}_\ell \mathbf{v}_\ell^\top X\boldsymbol{\beta}_\star \\
&= \sum_{\ell=1}^{r} \frac{1}{p + (1-p)\mathbb{X}_{11}/\sigma_\ell^2} \sigma_\ell \mathbf{v}_\ell \mathbf{w}_\ell^\top \boldsymbol{\beta}_\star.
\end{aligned}$$

■

A.3 Proof of Equation (12)

Set $A := X\big(\mathbb{X} + (p^{-1}-1)\operatorname{Diag}(\mathbb{X})\big)^{-1} X^\top$ and consider $\mathbf{w} = \mathbf{u} + \mathbf{v}$ with $\mathbf{u}^\top \mathbf{v} = 0$ and $X^\top \mathbf{v} = \mathbf{0}$. Observe that

$$A^2 = A - X\big(\mathbb{X} + (p^{-1}-1)\operatorname{Diag}(\mathbb{X})\big)^{-1}(p^{-1}-1)\operatorname{Diag}(\mathbb{X})\big(\mathbb{X} + (p^{-1}-1)\operatorname{Diag}(\mathbb{X})\big)^{-1} X^\top$$

and thus $\mathbf{w}^\top A^2 \mathbf{w} = \mathbf{u}^\top A^2 \mathbf{u} \le \mathbf{u} A \mathbf{u} = \mathbf{w}^\top A \mathbf{w}$. If $\mathbf{u}$ is the zero vector, $0 = \mathbf{w}^\top A^2 \mathbf{w} = \mathbf{w}^\top A \mathbf{w}$. Otherwise, $\operatorname{Diag}(\mathbb{X}) > 0$ implies $\big(\mathbb{X} + (p^{-1}-1)\operatorname{Diag}(\mathbb{X})\big)^{-1}(p^{-1}-1)\operatorname{Diag}(\mathbb{X})\big(\mathbb{X} +$

$(p^{-1}-1)\mathrm{Diag}(\mathbb{X})\big)$ being positive definite, so we have the strict inequality $\mathbf{w}^\top A^2\mathbf{w} < \mathbf{w}^\top A\mathbf{w}$ whenever $\mathbf{u} \neq 0$.

Now, suppose that A has an eigenvector $\mathbf{w}$ with corresponding eigenvalue $\lambda \geq 1$, which implies $\mathbf{u} \neq \mathbf{0}$. Then, we have $\mathbf{w}^\top A^2\mathbf{w} = \lambda^2 \geq \lambda = \mathbf{w}^\top A\mathbf{w}$. Equality only holds if $\mathbf{w}^\top A\mathbf{w} = 1$. This contradicts the strict inequality $\mathbf{w}^\top A^2\mathbf{w} < \mathbf{w}^\top A\mathbf{w}$ so all eigenvalues have to be strictly smaller than one. ■

Appendix B. Proofs for Section 4

B.1 Proof of Lemma 1

To show that $\big\|I - \alpha p\mathbb{X}_p\big\| \leq 1 - \alpha p(1-p)\min_i \mathbb{X}_{ii} < 1$, note that $\|\mathbb{X}_p\| \leq \|\mathbb{X}\|$ by Lemma 19 and recall $\alpha p\|\mathbb{X}\| < 1$ from Assumption 1. Hence,

$$\big\|I - \alpha p\mathbb{X}_p\big\| = 1 - \alpha p\lambda_{\min}(\mathbb{X}_p). \tag{28}$$

For any vector $\mathbf{v}$ with $\|\mathbf{v}\|_2 = 1$ and any two $d \times d$ positive semi-definite matrices A and B, $\mathbf{v}^\top(A+B)\mathbf{v} = \mathbf{v}^\top A\mathbf{v} + \mathbf{v}^\top B\mathbf{v} \geq \lambda_{\min}(A) + \lambda_{\min}(B)$. Hence, $\lambda_{\min}(A+B) \geq \lambda_{\min}(A) + \lambda_{\min}(B)$ and $\lambda_{\min}(\mathbb{X}_p) \geq (1-p)\lambda_{\min}\big(\mathrm{Diag}(\mathbb{X})\big) = (1-p)\min_i \mathbb{X}_{ii}$. By Assumption 1, the design matrix X has no zero columns, guaranteeing $\min_i \mathbb{X}_{ii} > 0$. Combined with (28), we now obtain $\big\|I - \alpha p\mathbb{X}_p\big\| \leq 1 - \alpha p(1-p)\min_i \mathbb{X}_{ii} < 1$.

To prove the bound on the expectation, recall from (14) that $\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}$ equals $\big(I - \alpha D_k\mathbb{X}_p\big)\big(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}}\big) + \alpha D_k\overline{\mathbb{X}}\big(pI - D_k\big)\widetilde{\boldsymbol{\beta}}_{k-1}$. Lemma 16(a) shows that $\mathbb{E}[DAD] = pA_p$ and Lemma 15(b) gives $\overline{A}_p = p\overline{A}$. In turn, we have $\mathbb{E}\big[D_k\overline{\mathbb{X}}D_k\big] = p\overline{\mathbb{X}}_p = p^2\overline{\mathbb{X}}$ and

$$\mathbb{E}\big[D_k\overline{\mathbb{X}}(pI - D_k)\big] = p^2\overline{\mathbb{X}} - p^2\overline{\mathbb{X}} = 0. \tag{29}$$

Conditioning on all randomness except D_k now implies

$$\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}} \mid \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}_{k-1}\big] = \big(I - \alpha p\mathbb{X}_p\big)\big(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}}\big). \tag{30}$$

By the tower rule $\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big] = \big(I - \alpha p\mathbb{X}_p\big)\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}}\big]$, so induction on k gives

$$\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big] = \big(I - \alpha p\mathbb{X}_p\big)^k\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_0 - \widetilde{\boldsymbol{\beta}}\big].$$

Sub-multiplicativity of the spectral norm implies $\big\|(I - \alpha p\mathbb{X}_p)^k\big\| \leq \|I - \alpha p\mathbb{X}_p\|^k$, proving that $\big\|\mathbb{E}[\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}]\big\|_2 \leq \big\|I - \alpha p\mathbb{X}_p\big\|^k\big\|\mathbb{E}[\widetilde{\boldsymbol{\beta}}_0 - \widetilde{\boldsymbol{\beta}}]\big\|_2$. ■

B.2 Proof of Theorem 2

We have $\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}}\big) = \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_A + (\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A)\big) = \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_A\big) + \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A\big) + \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_A, \widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A\big) + \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A, \widetilde{\boldsymbol{\beta}}_A\big)$, so the triangle inequality implies

$$\Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}}\big) - \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_A\big) - \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A\big)\Big\| \leq 2\Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A, \widetilde{\boldsymbol{\beta}}_A\big)\Big\|. \tag{31}$$

Write $B' := B - AX^\top$, then $\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A = B'\mathbf{Y} + \mathbf{a}$. When conditioned on $\mathbf{Y}$, the estimator $\widetilde{\boldsymbol{\beta}}_A = AX^\top\mathbf{Y}$ is deterministic. Hence, the law of total covariance yields

$$\begin{aligned}\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A, \widetilde{\boldsymbol{\beta}}_A\big) &= \mathbb{E}\Big[\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A, \widetilde{\boldsymbol{\beta}}_A \mid \mathbf{Y}\big)\Big] + \mathrm{Cov}\Big(\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A \mid \mathbf{Y}\big], \mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_A \mid \mathbf{Y}\big]\Big) \\ &= 0 + \mathrm{Cov}\big(\mathbb{E}[B']\mathbf{Y} + \mathbb{E}[\mathbf{a}], AX^\top\mathbf{Y}\big) = \mathrm{Cov}\big(\mathbb{E}[B']\mathbf{Y}, AX^\top\mathbf{Y}\big).\end{aligned}$$

Further, $\mathrm{Cov}(\mathbf{Y}) = I$ implies $\mathrm{Cov}\big(\mathbb{E}[B']\mathbf{Y}, AX^\top\mathbf{Y}\big) = \mathbb{E}[B']\mathrm{Cov}(\mathbf{Y})XA^\top = \mathbb{E}[B']XA^\top$. Using $\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A = B'\mathbf{Y} + \mathbf{a}$, note that $\mathbb{E}[B']X\boldsymbol{\beta}_\star = \mathbb{E}_{\boldsymbol{\beta}_\star}\big[\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A\big] - \mathbb{E}_{\mathbf{0}}\big[\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A\big]$ with $\mathbf{0} = (0, \dots, 0)^\top$. Combining these identities, sub-multiplicativity of the spectral norm, and the triangle inequality leads to

$$\begin{aligned}
\Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A, \widetilde{\boldsymbol{\beta}}_A\big)\Big\| &\le \|A\| \cdot \big\|\mathbb{E}[B']X\big\| \\
&= \|A\| \sup_{\boldsymbol{\beta}_\star : \|\boldsymbol{\beta}_\star\|_2 \le 1} \big\|\mathbb{E}[B']X\boldsymbol{\beta}_\star\big\|_2 \\
&\le \|A\| \sup_{\boldsymbol{\beta}_\star : \|\boldsymbol{\beta}_\star\|_2 \le 1} \Big(\Big\|\mathbb{E}_{\boldsymbol{\beta}_\star}\big[\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A\big]\Big\|_2 + \Big\|\mathbb{E}_{\mathbf{0}}\big[\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A\big]\Big\|_2\Big) \\
&\le 2\|A\| \sup_{\boldsymbol{\beta}_\star : \|\boldsymbol{\beta}_\star\|_2 \le 1} \Big\|\mathbb{E}_{\boldsymbol{\beta}_\star}\big[\widetilde{\boldsymbol{\beta}}_{\mathrm{aff}} - \widetilde{\boldsymbol{\beta}}_A\big]\Big\|_2.
\end{aligned}$$

Together with (31), this proves the result. ■

B.3 Proof of Lemma 4

Given a random vector U and a random element V, observe that $\mathbb{E}\big[\mathrm{Cov}(U \mid V)\big] = \mathbb{E}\big[UU^\top\big] - \mathbb{E}\big[\mathbb{E}[U \mid V]\mathbb{E}[U \mid V]^\top\big]$. Inserting $U = \widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}$ and $V = \big(\widetilde{\boldsymbol{\beta}}_{k-1}, \widetilde{\boldsymbol{\beta}}\big)$, as well as defining $A_k := \mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}})(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}})^\top\big]$, leads to

$$A_k = \mathbb{E}\Big[\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}} \mid \widetilde{\boldsymbol{\beta}}_{k-1}, \widetilde{\boldsymbol{\beta}}\big]\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}} \mid \widetilde{\boldsymbol{\beta}}_{k-1}, \widetilde{\boldsymbol{\beta}}\big]^\top\Big] + \mathbb{E}\Big[\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}} \mid \widetilde{\boldsymbol{\beta}}_{k-1}, \widetilde{\boldsymbol{\beta}}\big)\Big] \tag{32}$$

Recall $\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}} \mid \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}_{k-1}\big] = (I - \alpha p\mathbb{X}_p)\big(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}}\big)$ from (30), and so

$$\mathbb{E}\Big[\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}} \mid \widetilde{\boldsymbol{\beta}}_{k-1}, \widetilde{\boldsymbol{\beta}}\big]\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}} \mid \widetilde{\boldsymbol{\beta}}_{k-1}, \widetilde{\boldsymbol{\beta}}\big]^\top\Big] = \big(I - \alpha p\mathbb{X}_p\big)A_{k-1}\big(I - \alpha p\mathbb{X}_p\big), \tag{33}$$

where $A_{k-1} := \mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}})(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}})^\top\big]$.

Evaluating the conditional covariance $\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}} \mid \widetilde{\boldsymbol{\beta}}_{k-1}, \widetilde{\boldsymbol{\beta}}\big)$ is the more challenging part, requiring moments up to fourth order in D_k, see Lemma 17. Recall that

$$\begin{aligned}
S(A) = {} & \big(I - \alpha p\mathbb{X}_p\big)A\big(I - \alpha p\mathbb{X}_p\big) \\
& + \alpha^2 p(1-p)\mathrm{Diag}\big(\mathbb{X}_p A\mathbb{X}_p\big) + \alpha^2 p^2(1-p)^2\mathbb{X} \odot \overline{A + \mathbb{E}\big[\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^\top\big]} \odot \mathbb{X} \\
& + \alpha^2 p^2(1-p)\bigg(\Big(\overline{\mathbb{X}}\mathrm{Diag}\Big(A + \mathbb{E}\big[\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^\top\big]\Big)\overline{\mathbb{X}}\Big)_p + \overline{\mathbb{X}}\mathrm{Diag}\big(\mathbb{X}_p A\big) + \mathrm{Diag}\big(\mathbb{X}_p A\big)\overline{\mathbb{X}}\bigg).
\end{aligned}$$

Lemma 11 *For every positive integer k,*

$$\mathbb{E}\Big[\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}} \mid \widetilde{\boldsymbol{\beta}}_{k-1}, \widetilde{\boldsymbol{\beta}}\big)\Big] = S(A_{k-1}) - \big(I - \alpha p\mathbb{X}_p\big)A_{k-1}\big(I - \alpha p\mathbb{X}_p\big) + \rho_{k-1},$$

with remainder ρ_{k-1} vanishing at the rate

$$\big\|\rho_{k-1}\big\| \le 6\big\|I - \alpha p\mathbb{X}_p\big\|^{k-1}\Big\|\mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_0 - \widetilde{\boldsymbol{\beta}})\widetilde{\boldsymbol{\beta}}^\top\big]\Big\|.$$

Proof Recall from (14) that $\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}} = \big(I - \alpha D_k \mathbb{X}_p\big)\big(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}}\big) + \alpha D_k \overline{\mathbb{X}}\big(pI - D_k\big)\widetilde{\boldsymbol{\beta}}_{k-1}$. The covariance is invariant under deterministic shifts and sign flips, so

$$\mathrm{Cov}\Big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}} \mid \widetilde{\boldsymbol{\beta}}_{k-1}, \widetilde{\boldsymbol{\beta}}\Big) = \alpha^2 \mathrm{Cov}\Big(D_k \mathbb{X}_p\big(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}}\big) + D_k \overline{\mathbb{X}}\big(D_k - pI\big)\widetilde{\boldsymbol{\beta}}_{k-1} \mid \widetilde{\boldsymbol{\beta}}_{k-1}, \widetilde{\boldsymbol{\beta}}\Big).$$

Applying Lemma 17 with $\mathbf{u} := \mathbb{X}_p\big(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}}\big)$, $\overline{A} := \overline{\mathbb{X}}$, and $\mathbf{v} := \widetilde{\boldsymbol{\beta}}_{k-1}$, we find

$$\begin{aligned}\frac{\mathrm{Cov}(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}} \mid \widetilde{\boldsymbol{\beta}}_{k-1}, \widetilde{\boldsymbol{\beta}})}{\alpha^2 p(1-p)} &= \mathrm{Diag}\Big(\mathbb{X}_p\big(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}}\big)\big(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}}\big)^\top \mathbb{X}_p\Big) \\ &\quad + p\overline{\mathbb{X}}\mathrm{Diag}\Big(\mathbb{X}_p\big(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}}\big)\widetilde{\boldsymbol{\beta}}_{k-1}^\top\Big) \\ &\quad + p\mathrm{Diag}\Big(\widetilde{\boldsymbol{\beta}}_{k-1}\big(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}}\big)^\top \mathbb{X}_p\Big)\overline{\mathbb{X}} \\ &\quad + p\Big(\overline{\mathbb{X}}\mathrm{Diag}\big(\widetilde{\boldsymbol{\beta}}_{k-1}\widetilde{\boldsymbol{\beta}}_{k-1}^\top\big)\overline{\mathbb{X}}\Big)_p \\ &\quad + p(1-p)\mathbb{X} \odot \overline{\widetilde{\boldsymbol{\beta}}_{k-1}\widetilde{\boldsymbol{\beta}}_{k-1}^\top} \odot \mathbb{X}.\end{aligned} \tag{34}$$

Set $B_{k-1} := \mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}})\widetilde{\boldsymbol{\beta}}^\top\big]$ and recall $A_{k-1} = \mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}})(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}})^\top\big]$. Note the identities $\mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}})\widetilde{\boldsymbol{\beta}}_{k-1}^\top\big] = A_{k-1} + B_{k-1}$ and $\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_{k-1}\widetilde{\boldsymbol{\beta}}_{k-1}^\top\big] = A_{k-1} + \mathbb{E}\big[\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^\top\big] + B_{k-1} + B_{k-1}^\top$. Taking the expectation of (34), multiplying both sides with $\alpha^2 p(1-p)$, and using the definition of $S(A)$ proves the claimed expression for $\mathbb{E}\big[\mathrm{Cov}(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}} \mid \widetilde{\boldsymbol{\beta}}_{k-1}, \widetilde{\boldsymbol{\beta}})\big]$ with remainder term

$$\begin{aligned}\rho_{k-1} = \alpha^2 p(1-p)\Big(&p\overline{\mathbb{X}}\mathrm{Diag}\big(\mathbb{X}_p B_{k-1}\big) + p\mathrm{Diag}\big(B_{k-1}^\top \mathbb{X}_p\big)\overline{\mathbb{X}} + p\Big(\overline{\mathbb{X}}\mathrm{Diag}\big(B_{k-1} + B_{k-1}^\top\big)\overline{\mathbb{X}}\Big)_p \\ &+ p(1-p)\mathbb{X} \odot \overline{\big(B_{k-1} + B_{k-1}^\top\big)} \odot \mathbb{X}\Big).\end{aligned} \tag{35}$$

For any $d \times d$ matrices A and B, Lemma 19 provides the inequalities $\big\|\mathrm{Diag}(A)\big\| \le \|A\|$, $\|A_p\| \le \|A\|$, and $\|A \odot B\| \le \|A\| \cdot \|B\|$. If A is moreover positive semi-definite, then also $\big\|\overline{A}\big\| \le \|A\|$. Combined with the sub-multiplicativity of the spectral norm, this implies

$$\begin{aligned}\big\|\overline{\mathbb{X}}\mathrm{Diag}(\mathbb{X}_p B_{k-1})\big\| &\le \big\|\overline{\mathbb{X}}\big\| \cdot \big\|\mathrm{Diag}(\mathbb{X}_p B_{k-1})\big\| \\ &\le \|\mathbb{X}\| \cdot \|\mathbb{X}_p\| \cdot \big\|B_{k-1}\big\| \le \|\mathbb{X}\|^2 \cdot \big\|B_{k-1}\big\| \\ \Big\|\Big(\overline{\mathbb{X}}\mathrm{Diag}\big(B_{k-1} + B_{k-1}^\top\big)\overline{\mathbb{X}}\Big)_p\Big\| &\le \big\|\overline{\mathbb{X}}\mathrm{Diag}\big(B_{k-1} + B_{k-1}^\top\big)\overline{\mathbb{X}}\big\| \\ &\le \big\|\overline{\mathbb{X}}\big\|^2 \cdot \big\|B_{k-1} + B_{k-1}^\top\big\| \le 2\|\mathbb{X}\|^2 \cdot \big\|B_{k-1}\big\| \\ \Big\|\mathbb{X} \odot \overline{\big(B_{k-1} + B_{k-1}^\top\big)} \odot \mathbb{X}\Big\| &\le 2\|\mathbb{X}\|^2 \cdot \big\|B_{k-1}\big\|.\end{aligned}$$

By Assumption 1 also $\alpha p\|\mathbb{X}\| < 1$, so combining the upper-bounds with (35) leads to

$$\big\|\rho_{k-1}\big\| \le 6(\alpha p)^2 \|\mathbb{X}\|^2 \cdot \big\|B_{k-1}\big\| \le 6\big\|B_{k-1}\big\|. \tag{36}$$

The argument is to be completed by bounding $\big\|B_{k-1}\big\|$. Using (14), we have $\big(\widetilde{\boldsymbol{\beta}}_{k-1} - \widetilde{\boldsymbol{\beta}}\big)\widetilde{\boldsymbol{\beta}}^\top = \big(I - \alpha D_{k-1}\mathbb{X}_p\big)\big(\widetilde{\boldsymbol{\beta}}_{k-2} - \widetilde{\boldsymbol{\beta}}\big)\widetilde{\boldsymbol{\beta}}^\top + \alpha D_{k-1}\overline{\mathbb{X}}\big(pI - D_{k-1}\big)\widetilde{\boldsymbol{\beta}}_{k-1}\widetilde{\boldsymbol{\beta}}^\top$. In (29), it was

shown that $\mathbb{E}\big[D_{k-1}\overline{\mathbb{X}}(pI - D_{k-1})\big] = 0$. Recalling that D_{k-1} is independent of $\big(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}_{k-2}\big)$ and $\mathbb{E}[D_{k-1}] = pI_d$, we obtain $B_{k-1} = \big(I - \alpha p\mathbb{X}_p\big)B_{k-2}$. By induction on k, $B_{k-1} = \big(I - \alpha p\mathbb{X}_p\big)^{k-1}\mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_0 - \widetilde{\boldsymbol{\beta}})\widetilde{\boldsymbol{\beta}}^\top\big]$. Using sub-multiplicativity of the spectral norm,

$$\big\|B_{k-1}\big\| \le \|I - \alpha p\mathbb{X}_p\|^{k-1}\Big\|\mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_0 - \widetilde{\boldsymbol{\beta}})\widetilde{\boldsymbol{\beta}}^\top\big]\Big\|.$$

Together with (36) this finishes the proof. ■

Combining Lemma 11 with (32) and (33) leads to $\big\|A_k - S(A_{k-1})\big\| = \big\|\rho_{k-1}\big\|$ with remainder ρ_{k-1} as above. This completes the proof of Lemma 4. ■

B.4 Proof of Theorem 5

Let $S : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ be the affine operator introduced in Lemma 4 and recall the definitions $S_0 := S(0)$ and $S_{\text{lin}}(A) := S(A) - S_0$. First, the operator norm of S_{lin} will be analyzed.

Lemma 12 *The linear operator S_{lin} satisfies $\|S_{\text{lin}}\|_{\text{op}} \le \big\|I - \alpha p\mathbb{X}_p\big\| < 1$, provided that*

$$\alpha < \min\left\{\frac{1}{p\|\mathbb{X}\|}, \frac{\lambda_{\min}(\mathbb{X}_p)}{3\|\mathbb{X}\|^2}\right\},$$

where $\lambda_{\min}(\mathbb{X}_p)$ denotes the smallest eigenvalue of $\mathbb{X}_p$.

Proof Let A be a $d \times d$ matrix. Applying the triangle inequality, Lemma 19, and sub-multiplicativity of the spectral norm,

$$\begin{aligned}\big\|S_{\text{lin}}(A)\big\| &\le \big\|I - \alpha p\mathbb{X}_p\big\|^2\|A\| + \Big(\alpha^2 p(1-p) + 3\alpha^2p^2(1-p) + \alpha^2p^2(1-p)^2\Big)\|\mathbb{X}\|^2\|A\| \\ &\le \Big(\big\|I - \alpha p\mathbb{X}_p\big\|^2 + 2\alpha^2 p\|\mathbb{X}\|^2\Big)\|A\|,\end{aligned}$$

where the second inequality follows from $p(1-p) \le 1/4$.

As shown in (28), $\big\|I - \alpha p\mathbb{X}_p\big\| = 1 - \alpha p\lambda_{\min}(\mathbb{X}_p)$. Lemma 19 now implies $\big(1 - \alpha p\lambda_{\min}(\mathbb{X}_p)\big)^2 = 1 - 2\alpha p\lambda_{\min}(\mathbb{X}_p) + \alpha^2p^2\lambda_{\min}(\mathbb{X}_p) \le 1 - 2\alpha p\lambda_{\min}(\mathbb{X}_p) + \alpha^2 p\|\mathbb{X}\|^2$, so that

$$\big\|S_{\text{lin}}(A)\big\| \le \big(1 - 2\alpha p\lambda_{\min}(\mathbb{X}_p) + 3\alpha^2 p\|\mathbb{X}\|^2\big)\|A\|.$$

If $\alpha < \lambda_{\min}(\mathbb{X}_p)/\big(3\|\mathbb{X}\|^2\big)$, then also $3\alpha^2p\|\mathbb{X}\|^2 \le 3\alpha p\lambda_{\min}(\mathbb{X}_p)$, so that in turn $\|S_{\text{lin}}\|_{\text{op}} \le \big\|I - \alpha p\mathbb{X}_p\big\|$. The constraint $\alpha < 1/\big(p\|\mathbb{X}\|\big)$ now enforces $\alpha p\|\mathbb{X}\| < 1$, which implies $\big\|I - \alpha p\mathbb{X}_p\big\| < 1$. ■

As before, set $A_k := \mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}})(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}})^\top\big]$ for each $k \ge 0$ and let $\rho_k := A_{k+1} - S(A_k)$. Using induction on k, we now prove

$$A_k = S_{\text{lin}}^k(A_0) + \sum_{\ell=0}^{k-1} S_{\text{lin}}^\ell\big(S_0 + \rho_{k-1-\ell}\big). \tag{37}$$

Taking $k = 1$, $A_1 = S(A_0)+\rho_0 = S_{\mathrm{lin}}(A_0)+S_0+\rho_0$, so the claimed identity holds. Assuming the identity is true for $k-1$, the recursion $A_k = S(A_{k-1}) + \rho_{k-1}$ leads to

$$\begin{aligned}
A_k &= S\left(S_{\mathrm{lin}}^{k-1}(A_0) + \sum_{\ell=0}^{k-2} S_{\mathrm{lin}}^{\ell}\Big(S_0 + \rho_{k-2-\ell}\Big)\right) + \rho_{k-1} \\
&= S_{\mathrm{lin}}^{k}(A_0) + S_{\mathrm{lin}}\left(\sum_{\ell=0}^{k-2} S_{\mathrm{lin}}^{\ell}\Big(S_0 + \rho_{k-2-\ell}\Big)\right) + S_0 + \rho_{k-1} \\
&= S_{\mathrm{lin}}^{k}(A_0) + \sum_{\ell=0}^{k-1} S_{\mathrm{lin}}^{\ell}\Big(S_0 + \rho_{k-1-\ell}\Big),
\end{aligned}$$

thereby establishing the induction step and proving (37).

Assuming $\|S_{\mathrm{lin}}\|_{\mathrm{op}} \le \|I - \alpha p \mathbb{X}_p\| < 1$, we move on to show the bound

$$\Big\|A_k - \big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1} S_0\Big\| \le \big\|I - \alpha p \mathbb{X}_p\big\|^k \Big\|A_0 - \big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1} S_0\Big\| + C_0 k \|I - \alpha p \mathbb{X}_p\|^{k-1}$$

with id the identity operator on $\mathbb{R}^{d\times d}$, and $C_0 := 6\big\|\mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_0 - \widetilde{\boldsymbol{\beta}})\widetilde{\boldsymbol{\beta}}^\top\big]\big\|$. By linearity, $\sum_{\ell=0}^{k-1} S_{\mathrm{lin}}^{\ell}\big(S_0 + \rho_{k-1-\ell}\big) = \sum_{\ell=0}^{k-1} S_{\mathrm{lin}}^{\ell}(S_0) + \sum_{m=0}^{k-1} S_{\mathrm{lin}}^{m}(\rho_{k-1-m})$. Since $\|S_{\mathrm{lin}}\|_{\mathrm{op}} < 1$, Lemma 18 asserts that $\big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1} = \sum_{\ell=0}^{\infty} S_{\mathrm{lin}}^{\ell}$ and

$$\begin{aligned}
\sum_{\ell=0}^{k-1} S_{\mathrm{lin}}^{\ell}(S_0) &= \big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1} S_0 + \left(\sum_{\ell=0}^{k-1} S_{\mathrm{lin}}^{\ell} - \big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1}\right) S_0 \\
&= \big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1} S_0 - \sum_{\ell=k}^{\infty} S_{\mathrm{lin}}^{\ell}(S_0) \\
&= \big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1} S_0 - S_{\mathrm{lin}}^{k}\Big(\big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1} S_0\Big). \qquad (38)
\end{aligned}$$

Lemma 4 ensures $\|\rho_{k-1-m}\| \le C_0\big\|I - \alpha p \mathbb{X}_p\big\|^{k-1-m}$ for all $m \le k-1$. Moreover, $\|S_{\mathrm{lin}}\|_{\mathrm{op}} \le \big\|I - \alpha p \mathbb{X}_p\big\|$ and hence $\big\|S_{\mathrm{lin}}^{m}(\rho_{k-1-m})\big\| \le C_0\big\|I - \alpha p \mathbb{X}_p\big\|^{k-1}$ for every $m = 0, 1, \ldots$, so the triangle inequality implies

$$\left\|\sum_{m=0}^{k-1} S_{\mathrm{lin}}^{m}(\rho_{k-1-m})\right\| \le C_0 k \big\|I - \alpha p \mathbb{X}_p\big\|^{k-1}. \qquad (39)$$

Combining (37) and (38), as well as applying the triangle inequality and the bound (39), leads to the first bound asserted in Theorem 5,

$$\begin{aligned}
\Big\|A_k - \big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1} S_0\Big\| &\le \Big\|S_{\mathrm{lin}}^{k}\Big(A_0 - \big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1} S_0\Big)\Big\| + \left\|\sum_{m=0}^{k-1} S_{\mathrm{lin}}^{m}(\rho_{k-1-m})\right\| \\
&\le \big\|I - \alpha p \mathbb{X}_p\big\|^k \Big\|A_0 - \big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1} S_0\Big\| + C_0 k \big\|I - \alpha p \mathbb{X}_p\big\|^{k-1}. \qquad (40)
\end{aligned}$$

To show the corresponding bound for the variance, observe that $\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big) = A_k - \mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big]\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big]^\top$. Lemma 1 and (70) imply

$$\begin{aligned}\Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big) - A_k\Big\| &= \Big\|\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big]\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big]^\top\Big\| \\ &\le \Big\|\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big]\Big\|_2^2 \\ &\le \big\|I - \alpha p \mathbb{X}_p\big\|^{k-1}\Big\|\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}_0 - \widetilde{\boldsymbol{\beta}}\big]\Big\|_2^2.\end{aligned}$$

Together with (40) and the triangle inequality, this proves the second bound asserted in Theorem 5. ∎

B.5 Proof of Lemma 6

Applying Theorem 2, Lemma 1, and the triangle inequality,

$$\Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k\big) - \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}\big)\Big\| \le \Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big)\Big\| + 4\big\|\mathbb{X}_p^{-1}\big\|\big\|I - \alpha p \mathbb{X}_p\big\|^k \sup_{\boldsymbol{\beta}_\star : \|\boldsymbol{\beta}_\star\|_2 \le 1}\Big\|\mathbb{E}_{\boldsymbol{\beta}_\star}\big[\widetilde{\boldsymbol{\beta}}_0 - \widetilde{\boldsymbol{\beta}}\big]\Big\|_2. \tag{41}$$

Lemma 19 implies $\mathbb{X}_p \ge (1-p)\mathrm{Diag}(\mathbb{X})$, so that $\big\|\mathbb{X}_p^{-1}\big\| = \lambda_{\min}(\mathbb{X}_p)^{-1} \le \big((1-p)\min_i \mathbb{X}_{ii}\big)^{-1}$. Next, $\mathbb{E}_{\boldsymbol{\beta}_\star}\big[\widetilde{\boldsymbol{\beta}}\big] = \mathbb{X}_p^{-1}\mathbb{X}\boldsymbol{\beta}_\star$ entails equality between $\sup_{\boldsymbol{\beta}_\star:\|\boldsymbol{\beta}_\star\|_2\le 1}\big\|\mathbb{E}_{\boldsymbol{\beta}_\star}[\widetilde{\boldsymbol{\beta}}]\big\|_2$ and $\big\|\mathbb{X}_p^{-1}\mathbb{X}\big\| \le \big((1-p)\min_i \mathbb{X}_{ii}\big)^{-1}\|\mathbb{X}\|$. The second term on the right-hand side of (41) is then bounded by $C_1\big\|I - \alpha p\mathbb{X}_p\big\|^k/(1-p)^2$, for some constant C_1 independent of (α, p, k).

To prove the first claim of the lemma, it now suffices to show

$$\Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big)\Big\| \le \frac{1}{(1-p)^2}\Big(k\big\|I - \alpha p\mathbb{X}_p\big\|^{k-1}C_2 + \alpha p C_3\Big), \tag{42}$$

where C_2 and C_3 are constants independent of (α, p, k). As $\big\|\mathbb{X}_p^{-1}\big\| \le \big((1-p)\min_i \mathbb{X}_{ii}\big)^{-1}$, the constant C in Theorem 5 satisfies $C \le C_4/(1-p)^2 + \big\|(\mathrm{id} - S_{\mathrm{lin}})^{-1}S_0\big\|$, with C_4 depending only on the distribution of $\big(\mathbf{Y}, \widetilde{\boldsymbol{\beta}}_0, X\big)$. Consequently, Theorem 5 and the triangle inequality imply

$$\begin{aligned}\Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big)\Big\| &\le \Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big) - \big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1}S_0\Big\| + \Big\|\big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1}S_0\Big\| \\ &\le k\big\|I - \alpha p\mathbb{X}_p\big\|^{k-1}\left(\frac{C_4}{(1-p)^2} + \Big\|\big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1}S_0\Big\|\right) + \Big\|\big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1}S_0\Big\|.\end{aligned} \tag{43}$$

Consider a bounded linear operator G on $\mathbb{R}^{d\times d}$ satisfying $\|G\|_{\mathrm{op}} < 1$. For an arbitrary $d\times d$ matrix A, Lemma 18 asserts $(\mathrm{id} - G)^{-1}A = \sum_{\ell=0}^\infty G^\ell(A)$ and therefore $\big\|(\mathrm{id}-G)^{-1}A\big\| \le \sum_{\ell=0}^\infty \|G\|_{\mathrm{op}}^\ell \cdot \|A\| = \big(1 - \|G\|_{\mathrm{op}}\big)^{-1}\|A\|$. Theorem 2 states that $\|S_{\mathrm{lin}}\|_{\mathrm{op}} \le \big\|I - \alpha p\mathbb{X}_p\big\|$. As shown following (28), $\big\|I - \alpha p\mathbb{X}_p\big\| \le 1 - \alpha p(1-p)\min_i \mathbb{X}_{ii}$. Therefore,

$$\Big\|\big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1}S_0\Big\| \le \big(1 - \|S_{\mathrm{lin}}\|_{\mathrm{op}}\big)^{-1}\|S_0\| \le \big(\alpha p(1-p)\min_i \mathbb{X}_{ii}\big)^{-1}\|S_0\|.$$

Taking $A = 0$ in Lemma 4, $S_0 = \alpha^2 p^2 (1-p)\big(\overline{\mathbb{X}}\mathrm{Diag}\big(\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^\top\big]\big)\overline{\mathbb{X}}\big)_p + \alpha^2 p^2 (1-p)^2 \mathbb{X} \odot \overline{\mathbb{E}[\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^\top]} \odot \mathbb{X}$. Using Lemma 19 and $\big\|\mathbb{X}_p^{-1}\big\| \leq \big((1-p)\min_i \mathbb{X}_{ii}\big)^{-1}$,

$$\begin{aligned}\|S_0\| &\leq \alpha^2 p^2 (1-p)\Big(\Big\|\overline{\mathbb{X}}\mathrm{Diag}\big(\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^\top\big]\big)\overline{\mathbb{X}}\Big\| + \Big\|\mathbb{X} \odot \overline{\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^\top\big]} \odot \mathbb{X}\Big\|\Big) \\ &\leq \frac{(\alpha p \|\mathbb{X}\|)^2 \big\|\mathbb{E}[X^\top \mathbf{Y}\mathbf{Y}^\top X]\big\|}{(1-p)(\min_i \mathbb{X}_{ii})^2}\end{aligned}$$

proving that

$$\big\|(\mathrm{id} - S_{\mathrm{lin}})^{-1} S_0\big\| \leq \alpha p (1-p)^{-2} (\min_i \mathbb{X}_{ii})^{-3} \|\mathbb{X}\|^2 \cdot \big\|\mathbb{E}[X^\top \mathbf{Y}\mathbf{Y}^\top X]\big\|. \tag{44}$$

Note that $\alpha p \|\mathbb{X}\|^2 \leq \|\mathbb{X}\|$ by Assumption 1. Applying these bounds in (43) leads to

$$\begin{aligned}\Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big)\Big\| \leq \frac{k\|I - \alpha p \mathbb{X}_p\|^{k-1}}{(1-p)^2}\left(C_4 + \frac{\|\mathbb{X}\|\big\|\mathbb{E}[X^\top \mathbf{Y}\mathbf{Y}^\top X]\big\|}{(\min_i \mathbb{X}_{ii})^3}\right) \\ + \frac{\alpha p \|\mathbb{X}\|^2 \big\|\mathbb{E}[X^\top \mathbf{Y}\mathbf{Y}^\top X]\big\|}{(1-p)^2 (\min_i \mathbb{X}_{ii})^3},\end{aligned}$$

which proves (42). Combined with (41), this proves the first claim of the lemma since

$$\Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k\big) - \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}\big)\Big\| \leq \frac{k\|I - \alpha p \mathbb{X}_p\|^{k-1}(C_1 + C_2) + \alpha p C_3}{(1-p)^2}. \tag{45}$$

To start proving the second claim, recall that $\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}\big) = \mathbb{X}_p^{-1}\mathbb{X}\mathbb{X}_p^{-1}$. Hence, the triangle inequality leads to

$$\begin{aligned}&\Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k\big) - \mathrm{Diag}(\mathbb{X})^{-1}\mathbb{X}\mathrm{Diag}(\mathbb{X})^{-1}\Big\| \\ &\leq \Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k\big) - \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}\big)\Big\| + \Big\|\mathrm{Diag}(\mathbb{X})^{-1}\mathbb{X}\mathrm{Diag}(\mathbb{X})^{-1} - \mathbb{X}_p^{-1}\mathbb{X}\mathbb{X}_p^{-1}\Big\|.\end{aligned}$$

Let A, B, and C be square matrices of the same dimension, with A and B invertible. Observe the identity $A^{-1}CA^{-1} - B^{-1}CB^{-1} = A^{-1}(B-A)B^{-1}CA^{-1} + B^{-1}CA^{-1}(B-A)B^{-1}$, so sub-multiplicativity implies $\big\|A^{-1}CA^{-1} - B^{-1}CB^{-1}\big\| \leq 2\max\big\{\|A^{-1}\|, \|B^{-1}\|\big\}\|A^{-1}\| \cdot \|B^{-1}\| \cdot \|A - B\| \cdot \|C\|$. Using $\big\|\mathbb{X}_p^{-1}\big\| \leq \big((1-p)\min_i \mathbb{X}_{ii}\big)^{-1}$, and inserting $A = \mathrm{Diag}(\mathbb{X})$, $B = \mathbb{X}_p = A + p\overline{\mathbb{X}}$, and $C = \mathbb{X}$ results in

$$\Big\|\mathrm{Diag}(\mathbb{X})^{-1}\mathbb{X}\mathrm{Diag}(\mathbb{X})^{-1} - \mathbb{X}_p^{-1}\mathbb{X}\mathbb{X}_p^{-1}\Big\| \leq \frac{pC_5}{(1-p)^2}$$

with C_5 independent of (α, p, k). Combined with (45), this results in

$$\Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k\big) - \mathrm{Diag}(\mathbb{X})^{-1}\mathbb{X}\mathrm{Diag}(\mathbb{X})^{-1}\Big\| \leq \frac{k\|I - \alpha p \mathbb{X}_p\|^{k-1}(C_1 + C_2) + \alpha p C_3 + pC_5}{(1-p)^2},$$

which proves the second claim of the lemma by enlarging C'', if necessary. ■

B.6 Proof of Theorem 7

Start by noting that $\lambda_{\min}(A) = \inf_{\mathbf{v}:\|\mathbf{v}\|=1} \mathbf{v}^\top A\mathbf{v}$ for symmetric matrices, see Horn and Johnson (2013), Theorem 4.2.6. Using super-additivity of infima, observe the lower bound

$$\begin{aligned}
&\liminf_{k\to\infty} \inf_{\mathbf{v}:\|\mathbf{v}\|=1} \mathbf{v}^\top \Big(\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k\big) - \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}\big)\Big)\mathbf{v} \\
&\geq \liminf_{k\to\infty} \left(\lambda_{\min}\Big(\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big)\Big) - \sup_{\mathbf{v}:\|\mathbf{v}\|=1} \Big|\mathbf{v}^\top \Big(\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k\big) - \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}\big) - \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big)\Big)\mathbf{v}\Big|\right) \\
&\geq \liminf_{k\to\infty} \lambda_{\min}\Big(\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big)\Big) - \limsup_{k\to\infty} \Big\|\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k\big) - \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}\big) - \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big)\Big\|. \qquad (46)
\end{aligned}$$

Combining Lemma 1 and Theorem 2, the limit superior in (46) vanishes. Further, $\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big)$ converges to $\big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1} S_0$ by Theorem 5, so it suffices to analyze the latter matrix.

For the next step, the matrix $S_0 := S(0)$ in Theorem 5 will be lower-bounded. Taking $A = 0$ in Lemma 4 and exchanging the expectation with the Diag operator results in

$$\begin{aligned}
S_0 &= \alpha^2 p^2 (1-p) \left(\overline{\mathbb{X}}\mathrm{Diag}\Big(\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^\top\big]\Big)\overline{\mathbb{X}}\right)_p + \alpha^2 p^2 (1-p)^2 \mathbb{X} \odot \overline{\mathbb{E}\big[\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^\top\big]} \odot \mathbb{X} \\
&= \alpha^2 p^2 (1-p) \mathbb{E}\left[p\overline{\mathbb{X}}\mathrm{Diag}\big(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^\top\big)\overline{\mathbb{X}} + (1-p)\left(\mathrm{Diag}\Big(\overline{\mathbb{X}}\mathrm{Diag}\big(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^\top\big)\overline{\mathbb{X}}\Big) + \mathbb{X} \odot \overline{\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^\top} \odot \mathbb{X}\right)\right]. \qquad (47)
\end{aligned}$$

The first matrix in (47) is always positive semi-definite and we will now lower bound the matrix $B := \mathrm{Diag}\big(\overline{\mathbb{X}}\mathrm{Diag}(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^\top)\overline{\mathbb{X}}\big) + \mathbb{X} \odot \overline{\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^\top} \odot \mathbb{X}$. Given distinct $i, j = 1, \ldots, d$, symmetry of $\mathbb{X}$ implies

$$\begin{aligned}
\Big(\overline{\mathbb{X}}\mathrm{Diag}\big(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^\top\big)\overline{\mathbb{X}}\Big)_{ii} &= \sum_{k=1}^d \overline{\mathbb{X}}_{ik}\mathrm{Diag}\big(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^\top\big)_{kk}\overline{\mathbb{X}}_{ki} = \sum_{k=1}^d \mathbb{1}_{\{k\neq i\}}\mathbb{X}_{ik}^2\widetilde{\beta}_k^2, \\
\Big(\mathbb{X} \odot \overline{\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^\top} \odot \mathbb{X}\Big)_{ij} &= \mathbb{X}_{ij}^2\widetilde{\beta}_i\widetilde{\beta}_j.
\end{aligned}$$

In turn, for any unit-length vector $\mathbf{v}$,

$$\begin{aligned}
\mathbf{v}^\top\mathbb{E}[B]\mathbf{v} &= \mathbb{E}\left[\sum_{i=1}^d\sum_{k=1}^d \mathbb{1}_{\{k\neq i\}} v_i^2\mathbb{X}_{ik}^2\widetilde{\beta}_k^2 + \sum_{\ell=1}^d\sum_{m=1}^d \Big(\mathbb{1}_{\{\ell\neq m\}} v_\ell\mathbb{X}_{\ell m}\widetilde{\beta}_\ell\Big)\Big(\mathbb{1}_{\{\ell\neq m\}}\widetilde{\beta}_m\mathbb{X}_{\ell m}v_m\Big)\right] \\
&= \sum_{\ell=1}^d\sum_{m=1}^d \mathbb{1}_{\{\ell\neq m\}}\mathbb{X}_{\ell m}^2\mathbb{E}\Big[v_\ell^2\widetilde{\beta}_m^2 + v_\ell\widetilde{\beta}_\ell v_m\widetilde{\beta}_m\Big] \\
&= \frac{1}{2}\sum_{\ell=1}^d\sum_{m=1}^d \mathbb{1}_{\{\ell\neq m\}}\mathbb{X}_{\ell m}^2\mathbb{E}\Big[\big(v_\ell\widetilde{\beta}_m + v_m\widetilde{\beta}_\ell\big)^2\Big], \qquad (48)
\end{aligned}$$

where the last equality follows by noting that each square $(v_\ell\widetilde{\beta}_m + v_m\widetilde{\beta}_\ell)^2$ appears twice in (48) since the expression is symmetric in (ℓ, m). Every summand in (48) is non-negative.

If $v_\ell \neq 0$, then there exists $m(\ell) \neq \ell$ such that $\mathbb{X}_{\ell m} \neq 0$. Write $\mathbf{w}(\ell)$ for the vector with entries

$$w_i(\ell) = \begin{cases} v_\ell & \text{if } i = m(\ell), \\ v_{m(\ell)} & \text{if } i = \ell, \\ 0 & \text{otherwise.} \end{cases}$$

By construction, $\mathbb{E}\big[(v_\ell\widetilde{\beta}_{m(\ell)} + v_{m(\ell)}\widetilde{\beta}_\ell)^2\big] = \mathbf{w}(\ell)^\top \mathbb{E}\big[\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^\top\big]\mathbf{w}(\ell) \geq \mathbf{w}(\ell)^\top \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}\big)\mathbf{w}(\ell)$. Recall that $\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}\big) = \mathbb{X}_p^{-1}\mathbb{X}\mathbb{X}_p^{-1}$ and note that $\lambda_{\min}\big(\mathbb{X}_p^{-1}\mathbb{X}\mathbb{X}_p^{-1}\big) \geq \lambda_{\min}(\mathbb{X})/\|\mathbb{X}_p\|^2$. Together with $\big\|\mathbf{w}(\ell)\big\|_2^2 \geq v_\ell^2$ and $\sum_{\ell=1}^d v_\ell^2 = \|\mathbf{v}\|_2^2 = 1$, (48) now satisfies

$$\begin{aligned} &\frac{1}{2}\sum_{\ell=1}^d \sum_{m=1}^d \mathbb{1}_{\{\ell\neq m\}}\mathbb{X}_{\ell m}^2 \mathbb{E}\Big[\big(v_\ell\widetilde{\beta}_m + v_m\widetilde{\beta}_\ell\big)^2\Big] \\ &\geq \frac{1}{2}\sum_{\ell=1}^d \mathbb{1}_{\{v_\ell\neq 0\}}\mathbb{X}_{\ell m(\ell)}^2 \mathbf{w}(\ell)^\top \mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}\big)\mathbf{w}(\ell) \\ &\geq \frac{\lambda_{\min}(\mathbb{X})}{2\|\mathbb{X}_p\|^2}\sum_{\ell=1}^d \mathbb{1}_{\{v_\ell\neq 0\}}\big\|\mathbf{w}(\ell)\big\|_2^2 \min_{m:\mathbb{X}_{\ell m}\neq 0}\mathbb{X}_{\ell m}^2 \\ &\geq \frac{\lambda_{\min}(\mathbb{X})}{2\|\mathbb{X}_p\|^2}\min_{i\neq j:\mathbb{X}_{ij}\neq 0}\mathbb{X}_{ij}^2. \end{aligned}$$

As $\lambda_{\min}(S_0) \geq \alpha^2p^2(1-p)^2\lambda_{\min}(B)$, this proves the matrix inequality

$$S_0 \geq \frac{\alpha^2p^2(1-p)^2\lambda_{\min}(\mathbb{X})}{2\|\mathbb{X}_p\|^2}\min_{i\neq j:\mathbb{X}_{ij}\neq 0}\mathbb{X}_{ij}^2 \cdot I_d. \tag{49}$$

Next, let $\boldsymbol{\xi}$ be a centered random vector with covariance matrix M and suppose D is a $d\times d$ dropout matrix, independent of $\boldsymbol{\xi}$. Conditioning on $\boldsymbol{\xi}$, the law of total variance states

$$\begin{aligned} &\mathrm{Cov}\Big(\big(I-\alpha D\mathbb{X}_p\big)\boldsymbol{\xi} + \alpha D\overline{\mathbb{X}}\big(pI-D\big)\boldsymbol{\xi}\Big) \\ &= \mathrm{Cov}\Big(\mathbb{E}\big[(I-\alpha D\mathbb{X}_p)\boldsymbol{\xi} + D\overline{\mathbb{X}}(pI-D)\boldsymbol{\xi} \mid \boldsymbol{\xi}\big]\Big) \qquad (50) \\ &\quad + \mathbb{E}\Big[\mathrm{Cov}\big((I-\alpha D\mathbb{X}_p)\boldsymbol{\xi} + \alpha D\overline{\mathbb{X}}(pI-D)\boldsymbol{\xi} \mid \boldsymbol{\xi}\big)\Big] \\ &= \big(I-\alpha p\mathbb{X}_p\big)M\big(I-\alpha p\mathbb{X}_p\big) + \alpha^2\mathbb{E}\Big[\mathrm{Cov}\Big(D\mathbb{X}_p\boldsymbol{\xi} + D\overline{\mathbb{X}}\big(D-pI\big)\boldsymbol{\xi} \mid \boldsymbol{\xi}\Big)\Big]. \qquad (51) \end{aligned}$$

Applying Lemma 17 with $A := \mathbb{X}$, $\mathbf{u} := \mathbb{X}_p\boldsymbol{\xi}$, and $\mathbf{v} := \boldsymbol{\xi}$ now shows that $S_{\mathrm{lin}}(M) = \mathrm{Cov}\big((I-\alpha D\mathbb{X}_p)\boldsymbol{\xi} + \alpha D\overline{\mathbb{X}}(pI-D)\boldsymbol{\xi}\big)$. The second term in (51) is always positive semi-definite, proving that $S_{\mathrm{lin}}(M) \geq \lambda_{\min}(I-\alpha p\mathbb{X}_p)^2\lambda_{\min}(M)\cdot I_d$. As $(\mathrm{id}-S_{\mathrm{lin}})^{-1} = \sum_{\ell=0}^\infty S_{\mathrm{lin}}^\ell$ and $\lambda_{\min}\big(I-\alpha p\mathbb{X}_p\big) = 1-\alpha p\|\mathbb{X}_p\|$, this implies

$$\begin{aligned} \big(\mathrm{id}-S_{\mathrm{lin}}\big)^{-1}M &\geq \left(\lambda_{\min}(M)\sum_{\ell=0}^\infty \lambda_{\min}\big(I-\alpha p\mathbb{X}_p\big)^{2\ell}\right)\cdot I_d \\ &= \frac{\lambda_{\min}(M)}{2\alpha p\|\mathbb{X}_p\| - (\alpha p)^2\|\mathbb{X}_p\|^2}\cdot I_d \geq \frac{\lambda_{\min}(M)}{\alpha p\|\mathbb{X}_p\|}\cdot I_d. \end{aligned}$$

Lemma 19 moreover gives $\|\mathbb{X}_p\| \leq \|\mathbb{X}\|$. Together with the lower-bound (49) for $\lambda_{\min}(S_0)$, this proves the result. ∎

B.7 Proof of Theorem 8

As in Section 4.1, write $\widetilde{\boldsymbol{\beta}}_k^{\mathrm{rp}} := k^{-1}\sum_{j=1}^k \widetilde{\boldsymbol{\beta}}_j$ for the running average of the iterates and define

$$A_k^{\mathrm{rp}} := \mathbb{E}\Big[\big(\widetilde{\boldsymbol{\beta}}_k^{\mathrm{rp}} - \widetilde{\boldsymbol{\beta}}\big)\big(\widetilde{\boldsymbol{\beta}}_k^{\mathrm{rp}} - \widetilde{\boldsymbol{\beta}}\big)^\top\Big] = \frac{1}{k^2}\sum_{j,\ell=1}^k \mathbb{E}\Big[\big(\widetilde{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{\beta}}\big)\big(\widetilde{\boldsymbol{\beta}}_\ell - \widetilde{\boldsymbol{\beta}}\big)^\top\Big]. \tag{52}$$

Suppose $j > \ell$ and take $r = 0, \dots, j-\ell$. Using induction on r, we now prove that $\mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{\beta}})(\widetilde{\boldsymbol{\beta}}_\ell - \widetilde{\boldsymbol{\beta}})^\top\big] = (I - \alpha p\mathbb{X}_p)^r \mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_{j-r} - \widetilde{\boldsymbol{\beta}})(\widetilde{\boldsymbol{\beta}}_\ell - \widetilde{\boldsymbol{\beta}})^\top\big]$. The identity always holds when $r = 0$. Next, suppose the identity holds for some $r - 1 < j - \ell$. Taking $k = j + 1 - r$ in (14), $\widetilde{\boldsymbol{\beta}}_{j+1-r} - \widetilde{\boldsymbol{\beta}} = \big(I - \alpha D_{j+1-r}\mathbb{X}_p\big)\big(\widetilde{\boldsymbol{\beta}}_{j-r} - \widetilde{\boldsymbol{\beta}}\big) + \alpha D_{j+1-r}\overline{\mathbb{X}}\big(pI - D_{j+1-r}\big)\widetilde{\boldsymbol{\beta}}_{j-r}$. Since $j - r \geq \ell$, D_{j+1-r} is by assumption independent of $\big(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}_{j-r}, \widetilde{\boldsymbol{\beta}}_\ell\big)$. Recall from (29) that $\mathbb{E}\big[D_{j+1-r}\overline{\mathbb{X}}(pI - D_{j+1-r})\big] = 0$. Conditioning on $\big(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}_{j-r}, \widetilde{\boldsymbol{\beta}}_\ell\big)$ and applying tower rule now gives

$$\begin{aligned}
\mathbb{E}\Big[\big(\widetilde{\boldsymbol{\beta}}_{j+1-r} - \widetilde{\boldsymbol{\beta}}\big)\big(\widetilde{\boldsymbol{\beta}}_\ell - \widetilde{\boldsymbol{\beta}}\big)^\top\Big] &= \mathbb{E}\Big[I - \alpha D_{j+1-r}\mathbb{X}_p\Big]\mathbb{E}\Big[\big(\widetilde{\boldsymbol{\beta}}_{j-r} - \widetilde{\boldsymbol{\beta}}\big)\big(\widetilde{\boldsymbol{\beta}}_\ell - \widetilde{\boldsymbol{\beta}}\big)^\top\Big] \\
&\quad + \alpha\mathbb{E}\Big[D_{j+1-r}\overline{\mathbb{X}}\big(pI - D_{j+1-r}\big)\Big]\mathbb{E}\Big[\widetilde{\boldsymbol{\beta}}_{j-r}\big(\widetilde{\boldsymbol{\beta}}_\ell - \widetilde{\boldsymbol{\beta}}\big)^\top\Big] \\
&= \big(I - \alpha p\mathbb{X}_p\big)\mathbb{E}\Big[\big(\widetilde{\boldsymbol{\beta}}_{j-r} - \widetilde{\boldsymbol{\beta}}\big)\big(\widetilde{\boldsymbol{\beta}}_\ell - \widetilde{\boldsymbol{\beta}}\big)^\top\Big].
\end{aligned}$$

Together with the induction hypothesis, this proves the desired equality

$$\begin{aligned}
\mathbb{E}\Big[\big(\widetilde{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{\beta}}\big)\big(\widetilde{\boldsymbol{\beta}}_\ell - \widetilde{\boldsymbol{\beta}}\big)^\top\Big] &= \big(I - \alpha p\mathbb{X}_p\big)^{r-1}\mathbb{E}\Big[\big(\widetilde{\boldsymbol{\beta}}_{j+1-r} - \widetilde{\boldsymbol{\beta}}\big)\big(\widetilde{\boldsymbol{\beta}}_\ell - \widetilde{\boldsymbol{\beta}}\big)^\top\Big] \\
&= \big(I - \alpha p\mathbb{X}_p\big)^{r}\mathbb{E}\Big[\big(\widetilde{\boldsymbol{\beta}}_{j-r} - \widetilde{\boldsymbol{\beta}}\big)\big(\widetilde{\boldsymbol{\beta}}_\ell - \widetilde{\boldsymbol{\beta}}\big)^\top\Big].
\end{aligned}$$

For $j < \ell$, transposing and flipping the roles of j and ℓ also shows that $\mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{\beta}})(\widetilde{\boldsymbol{\beta}}_\ell - \widetilde{\boldsymbol{\beta}})^\top\big] = \mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{\beta}})(\widetilde{\boldsymbol{\beta}}_{\ell-r} - \widetilde{\boldsymbol{\beta}})^\top\big](I - \alpha p\mathbb{X}_p)^r$ with $r = 0, \dots, \ell - j$.

Defining $A_\ell := \mathbb{E}\big[(\widetilde{\boldsymbol{\beta}}_\ell - \widetilde{\boldsymbol{\beta}})(\widetilde{\boldsymbol{\beta}}_\ell - \widetilde{\boldsymbol{\beta}})^\top\big]$ and taking $r = |j - \ell|$, (52) may now be rewritten as

$$A_k^{\mathrm{rp}} = \frac{1}{k^2}\sum_{j=1}^k \left(\sum_{\ell=0}^j \big(I - \alpha p\mathbb{X}_p\big)^{j-\ell} A_\ell + \sum_{\ell=j+1}^\infty A_j\big(I - \alpha p\mathbb{X}_p\big)^{\ell-j}\right). \tag{53}$$

Set $\gamma := \|I - \alpha p\mathbb{X}_p\|$, then $\gamma \leq 1 - \alpha p(1-p)\min_i \mathbb{X}_{ii} < 1$ by Lemma 1. Note also that $\sum_{r=0}^j \gamma^r \leq \sum_{r=0}^\infty \gamma^r = (1-\gamma)^{-1}$. Using the triangle inequality and sub-multiplicativity of the spectral norm, (53) then satisfies

$$\begin{aligned}
\big\|A_k^{\mathrm{rp}}\big\| \leq \frac{1}{k^2}\sum_{j=1}^k \left(\sum_{\ell=0}^j \gamma^{j-\ell}\big\|A_\ell\big\| + \sum_{\ell=j+1}^\infty \big\|A_j\big\|\gamma^{\ell-j}\right) &\leq \frac{2}{k^2}\sum_{\ell=1}^k \big\|A_\ell\big\| \sum_{r=0}^\infty \gamma^r \\
&= \frac{2}{k^2(1-\gamma)}\sum_{\ell=1}^k \big\|A_\ell\big\|.
\end{aligned}$$

As shown in Theorem 5, $\|A_\ell\| \le \big\|(\mathrm{id}-S_{\mathrm{lin}})^{-1}S_0\big\|+C\ell\gamma^{\ell-1}$ for some constant C. Observing that $\sum_{\ell=1}^{\infty}\ell\gamma^{\ell-1} = \partial_\gamma\sum_{\ell=1}^{\infty}\gamma^\ell = \partial_\gamma\big((1-\gamma)^{-1}-1\big) = (1-\gamma)^{-2}$, this implies

$$\big\|A_k^{\mathrm{rp}}\big\| \le \frac{2}{k(1-\gamma)}\big\|(\mathrm{id}-S_{\mathrm{lin}})^{-1}S_0\big\| + \frac{2C}{k^2(1-\gamma)^3}.$$

To complete the proof note that $\gamma \le 1-\alpha p(1-p)\min_i \mathbb{X}_{ii}$ may be rewritten as $(1-\gamma)^{-1} \le \big(\alpha p(1-p)\min_i \mathbb{X}_{ii}\big)^{-1}$ and $\big\|(\mathrm{id}-S_{\mathrm{lin}})^{-1}S_0\big\| \le \alpha p(1-p)^{-2}(\min_i \mathbb{X}_{ii})^{-3}\|\mathbb{X}\|^2\cdot\big\|\mathbb{E}[X^\top\mathbf{Y}\mathbf{Y}^\top X]\big\|$ by (44). ∎

B.8 Proof of Lemma 9

Recall the definition $T(A) := \big(I-\alpha p\mathbb{X}\big)A\big(I-\alpha p\mathbb{X}\big) + \alpha^2 p(1-p)\mathrm{Diag}\big(\mathbb{X}A\mathbb{X}\big)$.

Lemma 13 *For every $k = 1, 2, \ldots$*

$$\begin{aligned}\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big) &= \big(I-\alpha p\mathbb{X}\big)\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}\big)\big(I-\alpha p\mathbb{X}\big)\\ &\qquad + \alpha^2 p(1-p)\mathrm{Diag}\Big(\mathbb{X}\mathbb{E}\Big[\big(\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}\big)\big(\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}\big)^\top\Big]\mathbb{X}\Big)\\ &\ge T\big(\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}\big)\big),\end{aligned}$$

with equality if $\widehat{\boldsymbol{\beta}}_0 = \mathbb{E}\big[\widehat{\boldsymbol{\beta}}\big]$ almost surely.

Proof Recall the definition $A_k := \mathbb{E}\big[(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}})^\top\big]$, so that $\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big) = A_k - \mathbb{E}\big[\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big]\mathbb{E}\big[\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big]^\top$. As shown in (22), $A_k = T(A_{k-1})$ and hence

$$\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}\big) = T\big(A_{k-1}\big) - \mathbb{E}\big[\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big]\mathbb{E}\big[\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big]^\top$$

By definition, $T(A_{k-1}) = \big(I-\alpha p\mathbb{X}\big)A_{k-1}\big(I-\alpha p\mathbb{X}\big) + \alpha^2 p(1-p)\mathrm{Diag}\big(\mathbb{X}A_{k-1}\mathbb{X}\big)$. Recall from (20) that $\mathbb{E}\big[\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big] = \big(I-\alpha p\mathbb{X}\big)\mathbb{E}\big[\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}\big]$, so $\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}\big) = A_{k-1} - \mathbb{E}\big[\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}\big]\mathbb{E}\big[\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}\big]^\top$ implies

$$\big(I-\alpha p\mathbb{X}\big)A_{k-1}\big(I-\alpha p\mathbb{X}\big) = \big(I-\alpha p\mathbb{X}\big)\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}\big)\big(I-\alpha p\mathbb{X}\big) + \mathbb{E}[\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}]\mathbb{E}[\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}]^\top.$$

Together, these identities prove the first claim.

The lower bound follows from $\mathbb{X}\mathbb{E}\big[\widehat{\boldsymbol{\beta}}_{k-1}-\widehat{\boldsymbol{\beta}}\big]\mathbb{E}\big[\widehat{\boldsymbol{\beta}}_{k-1}-\widehat{\boldsymbol{\beta}}\big]^\top\mathbb{X}$ being positive semi-definite. A positive semi-definite matrix has non-negative diagonal entries, meaning $\mathrm{Diag}\big(\mathbb{X}\mathbb{E}[\widehat{\boldsymbol{\beta}}_{k-1}-\widehat{\boldsymbol{\beta}}]\mathbb{E}[\widehat{\boldsymbol{\beta}}_{k-1}-\widehat{\boldsymbol{\beta}}]^\top\mathbb{X}\big)$ is also positive semi-definite. Next, note that $A_{k-1} = \mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}\big) + \mathbb{E}\big[\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}\big]\mathbb{E}\big[\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}\big]^\top$ and in turn

$$\begin{aligned}\mathrm{Diag}\big(\mathbb{X}A_{k-1}\mathbb{X}\big) &= \mathrm{Diag}\big(\mathbb{X}\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}})\mathbb{X}\big) + \mathrm{Diag}\big(\mathbb{X}\mathbb{E}[\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}]\mathbb{E}[\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}]^\top\mathbb{X}\big)\\ &\ge \mathrm{Diag}\big(\mathbb{X}\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}})\mathbb{X}\big).\end{aligned} \tag{54}$$

Together with the first part of the lemma and the definition of T, the lower-bound follows.

Lastly, if $\widehat{\boldsymbol{\beta}}_0 = \mathbb{E}[\widehat{\boldsymbol{\beta}}]$ almost surely, then (20) implies $\mathbb{E}\big[\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big] = \big(I-\alpha p\mathbb{X}\big)^k\mathbb{E}\big[\widehat{\boldsymbol{\beta}}_0 - \widehat{\boldsymbol{\beta}}\big] = 0$, so equality holds in (54). ∎

Consider arbitrary positive semi-definite matrices $A \geq B$, then $\mathbf{w}^\top(A-B)\mathbf{w} \geq 0$ for all vectors $\mathbf{w}$. Given any vector $\mathbf{v}$, this implies

$$\begin{aligned}\mathbf{v}^\top T(A-B)\mathbf{v} &= \mathbf{v}^\top\big(I-\alpha p\mathbb{X}\big)\big(A-B\big)\big(I-\alpha p\mathbb{X}\big)\mathbf{v} + \alpha^2 p(1-p)\sum_{\ell=1}^d v_\ell^2 \mathbf{e}_\ell^\top \mathbb{X}\big(A-B\big)\mathbb{X}\mathbf{e}_\ell \\ &\geq 0\end{aligned}$$

with $\mathbf{e}_\ell$ the ℓ^{th} standard basis vector. Accordingly, T is operator monotone with respect to the ordering of positive semi-definite matrices, in the sense that $T(A) \geq T(B)$ whenever $A \geq B$. Using induction on k, Lemma 13 may now be rewritten as

$$\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big) \geq T^k\Big(\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_0 - \widehat{\boldsymbol{\beta}}\big)\Big). \tag{55}$$

To complete the proof, the right-hand side of (55) will be analyzed for a suitable choice of $\mathbb{X}$.

Lemma 14 *Suppose $\widehat{\boldsymbol{\beta}}_0$ is independent of all other sources of randomness. Consider the linear regression model with a single observation $n = 1$, number of parameters $d \geq 2$, and design matrix $X = \mathbf{1}^\top$. Then, $\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_k\big) \geq \mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big) + \mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}\big)$ and for any d-dimensional vector $\mathbf{v}$ satisfying $\mathbf{v}^\top\mathbf{1} = 0$ and every $k = 1, 2, \ldots$*

$$\mathbf{v}^\top \mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big)\mathbf{v}^\top \geq \alpha^2 p(1-p)\|\mathbf{v}\|_2^2.$$

Proof By definition, $\mathbb{X} = \mathbf{1}\mathbf{1}^\top$ is the $d \times d$-matrix with all entries equal to one. Consequently, $\mathbb{X}^k = d^{k-1}\mathbb{X}$ for all $k \geq 1$ and $\mathbb{X}X^\top = \mathbf{1}\mathbf{1}^\top\mathbf{1} = d\mathbf{1} = dX^\top$, so $\widehat{\boldsymbol{\beta}} := d^{-1}X^\top\mathbf{Y}$ satisfies the normal equations $X^\top\mathbf{Y} = \mathbb{X}\widehat{\boldsymbol{\beta}}$.

To prove $\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_k\big) \geq \mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big) + \mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}\big)$, note that $\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}\big) \geq \mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big) + \mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}\big)$ whenever $\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}\big) \geq 0$. By conditioning on $\big(\widehat{\boldsymbol{\beta}}_k, \widehat{\boldsymbol{\beta}}\big)$, the identity $\mathbb{E}\big[\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big] = \big(I - \alpha p\mathbb{X}\big)\mathbb{E}\big[\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}\big]$ was shown in (20). The same argument also proves $\mathbb{E}\big[(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}}^\top\big] = \big(I - \alpha p\mathbb{X}\big)\mathbb{E}\big[(\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}}^\top\big]$. Induction on k and the assumed independence between $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}$ now lead to

$$\begin{aligned}\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}\big) &= \mathbb{E}\big[(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}}^\top\big] - \mathbb{E}\big[\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big]\mathbb{E}\big[\widehat{\boldsymbol{\beta}}^\top\big] \\ &= \big(I - \alpha p\mathbb{X}\big)\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}\big) \\ &= \big(I - \alpha p\mathbb{X}\big)^k\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}_0 - \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}\big) \\ &= \big(I - \alpha p\mathbb{X}\big)^k\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}\big).\end{aligned} \tag{56}$$

Next, note that $\mathrm{Cov}\big(\widetilde{\boldsymbol{\beta}}\big) = d^{-2}\mathbb{X}$ and $\big(I - \alpha p\mathbb{X}\big)\mathbb{X} = \big(1 - \alpha pd\big)\mathbb{X}$. As $\alpha p\|\mathbb{X}\| = \alpha pd < 1$, (56) satisfies

$$\big(I - \alpha p\mathbb{X}\big)^k\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}\big) = \frac{1}{d^2}\big(I - \alpha p\mathbb{X}\big)^k\mathbb{X} = \frac{1}{d^2}\big(1 - \alpha pd\big)^k\mathbb{X} \geq 0$$

which proves the first claim.

To prove the second claim, we first show that there are real sequences $\{\nu_k\}_k$ and $\{\lambda_k\}_k$, not depending on the distribution of $\widehat{\boldsymbol{\beta}}_0$, such that

$$\operatorname{Cov}\big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big) \geq \nu_k I_d + \tfrac{\lambda_k}{d}\mathbb{X} \tag{57}$$

for all $k \geq 0$, with equality if $\widehat{\boldsymbol{\beta}}_0 = \mathbb{E}\big[\widehat{\boldsymbol{\beta}}\big]$ almost surely. When $k = 0$, independence between $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}$, as well as $\operatorname{Cov}\big(\widehat{\boldsymbol{\beta}}\big) = d^{-2}\mathbb{X}$, imply $\operatorname{Cov}\big(\widehat{\boldsymbol{\beta}}_0 - \widehat{\boldsymbol{\beta}}\big) = \operatorname{Cov}\big(\widehat{\boldsymbol{\beta}}_0\big) + \operatorname{Cov}\big(\widehat{\boldsymbol{\beta}}\big) \geq d^{-2}\mathbb{X}$. Moreover, equality holds whenever $\widehat{\boldsymbol{\beta}}_0$ is deterministic.

For the sake of induction suppose the claim is true for some $k-1$. Lemma 13 and operator monotonicity of T then imply

$$\operatorname{Cov}\big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big) \geq T\Big(\operatorname{Cov}\big(\widehat{\boldsymbol{\beta}}_{k-1} - \widehat{\boldsymbol{\beta}}\big)\Big) \geq T\Big(\nu_{k-1} I_d + \tfrac{\lambda_{k-1}}{d}\mathbb{X}\Big).$$

In case $\widehat{\boldsymbol{\beta}} = \mathbb{E}\big[\widehat{\boldsymbol{\beta}}\big]$ almost surely, Lemma 13 and the induction hypothesis assert equality in the previous display. Recall $\mathbb{X}^\ell = d^{\ell-1}\mathbb{X}$, so $\big(I - \alpha p\mathbb{X}\big)^2 = I + \big((\alpha p)^2 d - 2\alpha p\big)\mathbb{X}$ as well as $\big(I - \alpha p\mathbb{X}\big)\mathbb{X}\big(I - \alpha p\mathbb{X}\big) = (1 - \alpha p d)^2\mathbb{X}$. Note also that $\operatorname{Diag}(\mathbb{X}) = I_d$. Setting $c := \alpha p d$, expanding the definition $T(A) = \big(I - \alpha p\mathbb{X}\big)A\big(I - \alpha p\mathbb{X}\big) + \alpha^2 p(1-p)\operatorname{Diag}\big(\mathbb{X}A\mathbb{X}\big)$ now results in

$$\begin{aligned}
&T\left(\nu_{k-1} I_d + \frac{\lambda_{k-1}}{d}\mathbb{X}\right) \\
&= \Big(\nu_{k-1} + \alpha^2 p(1-p) d(\nu_{k-1} + \lambda_{k-1})\Big) \cdot I_d + \left(\nu_{k-1}\Big((\alpha p)^2 d - 2\alpha p\Big) + \frac{\lambda_{k-1}(1 - \alpha p d)^2}{d}\right) \cdot \mathbb{X} \\
&= \Big(\nu_{k-1} + \alpha^2 p(1-p) d(\nu_{k-1} + \lambda_{k-1})\Big) \cdot I_d - \frac{\nu_{k-1} - (1 - \alpha p d)^2(\nu_{k-1} + \lambda_{k-1})}{d} \cdot \mathbb{X}
\end{aligned} \tag{58}$$

This establishes the induction step and thereby proves (57) for all $k \geq 0$.

As (ν_k, λ_k) do not depend on the distribution of $\widehat{\boldsymbol{\beta}}_0$, taking $\widehat{\boldsymbol{\beta}}_0 = \mathbb{E}\big[\widehat{\boldsymbol{\beta}}\big]$ shows that $\nu_k + \lambda_k \geq 0$ for all $k \geq 0$ since

$$0 \leq \mathbf{1}^\top \operatorname{Cov}\big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big)\mathbf{1} = \mathbf{1}^\top\Big(\nu_k I_d + \tfrac{\lambda_k}{d}\mathbb{X}\Big)\mathbf{1} = d\big(\nu_k + \lambda_k\big).$$

Consequently, (58) implies $\nu_k = \nu_{k-1} + \alpha^2 p(1-p) d(\nu_{k-1} + \lambda_{k-1}) \geq \nu_{k-1}$, proving that $\{\nu_k\}_k$ is non-decreasing in k.

Lastly, we show that $\nu_1 \geq \alpha^2 p(1-p)$. To this end, recall that $\mathbb{X}^3 = d^2\mathbb{X}$ and $\operatorname{Diag}(\mathbb{X}) = I$. As T is operator monotone and $T(A) \geq \alpha^2 p(1-p)\operatorname{Diag}\big(\mathbb{X}A\mathbb{X}\big)$, independence of $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}$ results in

$$\begin{aligned}
\operatorname{Cov}\big(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}\big) &\geq T\Big(\operatorname{Cov}\big(\widehat{\boldsymbol{\beta}}_0 - \widehat{\boldsymbol{\beta}}\big)\Big) \geq T\Big(\operatorname{Cov}\big(\widehat{\boldsymbol{\beta}}\big)\Big) \geq \alpha^2 p(1-p)\operatorname{Diag}\big(\mathbb{X}d^{-2}\mathbb{X}\mathbb{X}\big) \\
&= \frac{\alpha^2 p(1-p)}{d^2}\operatorname{Diag}\big(\mathbb{X}^3\big) = \alpha^2 p(1-p)\operatorname{Diag}(\mathbb{X}) = \alpha^2 p(1-p) I_d,
\end{aligned}$$

so $\nu_1 \geq \alpha^2 p(1-p)$.

To complete the proof, observe that $\mathbf{v}^\top\mathbf{1} = 0$ implies $\mathbb{X}\mathbf{v} = 0$. Accordingly,

$$\mathbf{v}^\top \operatorname{Cov}\big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}\big)\mathbf{v}^\top \geq \mathbf{v}^\top\Big(\nu_k I_d + \tfrac{\lambda_k}{d}\mathbb{X}\Big)\mathbf{v}^\top \geq \nu_k\|\mathbf{v}\|_2^2 \geq \nu_1\|\mathbf{v}\|_2^2 \geq \alpha^2 p(1-p)\|\mathbf{v}\|_2^2$$

which yields the second claim of the lemma. ∎

B.9 Proof of Theorem 10

Recall that $T(A) := \big(I - \alpha p \mathbb{X}\big) A \big(I - \alpha p \mathbb{X}\big) + \alpha^2 p(1-p) \mathrm{Diag}\big(\mathbb{X} A \mathbb{X}\big)$, as defined in (21). If $A_k := \mathbb{E}\big[(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}})^\top\big]$, then $A_k = T\big(A_{k-1}\big)$ by (22).

For an arbitrary $d \times d$ matrix A, the triangle inequality, Lemma 19, and submultiplicativity of the spectral norm imply

$$\big\|T(A)\big\| \leq \Big(\big\|I - \alpha p \mathbb{X}\big\|^2 + \alpha^2 p(1-p)\|\mathbb{X}\|^2\Big)\|A\|. \tag{59}$$

As $\mathbb{X}$ is positive definite, $\alpha p \|\mathbb{X}\| < 1$ implies $\big\|I - \alpha p \mathbb{X}\big\| = 1 - \alpha p \lambda_{\min}(\mathbb{X})$. If $\alpha \leq \frac{\lambda_{\min}(\mathbb{X})}{\|\mathbb{X}\|^2}$, then

$$\begin{aligned}
\big\|I - \alpha p \mathbb{X}\big\|^2 + \alpha^2 p(1-p)\|\mathbb{X}\|^2 &= 1 - 2\alpha p \lambda_{\min}(\mathbb{X}) + \alpha^2\big(p^2 \lambda_{\min}(\mathbb{X})^2 + p(1-p)\|\mathbb{X}\|^2\big) \\
&\leq \big(1 - 2\alpha p \lambda_{\min}(\mathbb{X})\big) + \alpha^2 p \|\mathbb{X}\|^2 \\
&\leq 1 - \alpha p \lambda_{\min}(\mathbb{X}).
\end{aligned}$$

Together with (59) this leads to $\|A_k\| \leq \big(1 - \alpha p \lambda_{\min}(\mathbb{X})\big)\|A_{k-1}\|$. By induction on k, $\|A_k\| \leq \big(1 - \alpha p \lambda_{\min}(\mathbb{X})\big)^k \|A_0\|$, completing the proof. ■

Appendix C. Higher Moments of Dropout Matrices

Deriving concise closed-form expressions for third and fourth order expectations of the dropout matrices presents one of the main technical challenges encountered in Section B.

All matrices in this section will be of dimension $d \times d$ and all vectors of length d. Moreover, D always denotes a random diagonal matrix such that $D_{ii} \overset{i.i.d.}{\sim} \mathrm{Ber}(p)$ for all $i = 1, \ldots, d$. The diagonal entries of D are elements of $\{0, 1\}$, meaning $D = D^k$ for all positive powers k.

Given a matrix A and $p \in (0, 1)$, recall the definitions

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

The first lemma contains some simple identities.

Lemma 15 *For arbitrary matrices A and B, $p \in (0, 1)$, and a diagonal matrix F,*

(a) $\overline{AF} = \overline{A}F$ *and* $\overline{FA} = F\overline{A}$

(b) $\overline{A}_p = p\overline{A} = \overline{A_p}$

(c) $\mathrm{Diag}\big(\overline{A}B\big) = \mathrm{Diag}\big(A\overline{B}\big)$.

Proof

(a) By definition, $(\overline{A}F)_{ij} = F_{jj} \mathbb{1}_{\{i \neq j\}} A_{ij}$ for all $i, j \in \{1, \ldots, d\}$, which equals $\overline{AF}$. The second equality then follows by transposition.

(b) Clearly, $\mathrm{Diag}(\overline{A}) = 0$ and in turn $\overline{A}_p = p\overline{A} + (1-p)\mathrm{Diag}(\overline{A}) = p\overline{A}$. On the other hand, $\mathrm{Diag}(A_p) = \mathrm{Diag}(A)$ and hence $\overline{A_p} = pA + (1-p)\mathrm{Diag}(A) - \mathrm{Diag}(A) = pA - p\mathrm{Diag}(A)$ equals $p\overline{A}$ as well.

(c) Observe that $\mathrm{Diag}\big(\overline{A}B\big)$ equals $\mathrm{Diag}\big(AB\big) - \mathrm{Diag}\big(\mathrm{Diag}(A)B\big)$. As $\mathrm{Diag}\big(\mathrm{Diag}(A)B\big) = \mathrm{Diag}(A)\mathrm{Diag}(B) = \mathrm{Diag}\big(A\mathrm{Diag}(B)\big)$, the claim follows.

■

With these basic properties at hand, higher moments involving the dropout matrix D may be computed by carefully accounting for equalities between the involved indices.

Lemma 16 *Given arbitrary matrices A, B, and C, the following hold:*

(a) $\mathbb{E}\big[DAD\big] = pA_p$

(b) $\mathbb{E}\big[DADBD\big] = pA_pB_p + p^2(1-p)\mathrm{Diag}(\overline{A}B)$

(c) $\mathbb{E}\big[DADBDCD\big] = pA_pB_pC_p + p^2(1-p)\Big(\mathrm{Diag}\big(\overline{A}B_p\overline{C}\big) + A_p\mathrm{Diag}(\overline{B}C) + \mathrm{Diag}(A\overline{B})C_p + (1-p)A \odot \overline{B}^\top \odot C\Big)$

Proof

(a) Recall that $D = D^2$ and hence $\mathbb{E}[D] = \mathbb{E}[D^2] = pI$, meaning $\mathbb{E}\big[D\mathrm{Diag}(A)D\big] = \mathrm{Diag}(A)\mathbb{E}[D] = p\mathrm{Diag}(A)$. On the other hand, $(D\overline{A}D)_{ij} = D_{ii}D_{jj}A_{ij}\mathbb{1}_{\{i\neq j\}}$ implies $\mathbb{E}\big[D\overline{A}D\big] = p^2\overline{A}$ due to independence of D_{ii} and D_{jj}. Combining both identities,

$$\begin{aligned}\mathbb{E}\big[DAD\big] &= \mathbb{E}\Big[D\overline{A}D + D\mathrm{Diag}(A)D\Big]\\ &= p^2\overline{A} + p\mathrm{Diag}(A)\\ &= p\Big(pA - p\mathrm{Diag}(A) + \mathrm{Diag}(A)\Big) = pA_p.\end{aligned}$$

(b) First, note that $D = D^2$ and commutativity of diagonal matrices imply

$$\begin{aligned}DADBD &= D\overline{A}DBD + D\mathrm{Diag}(A)DBD\\ &= D\overline{\overline{A}DB}D + \mathrm{Diag}(A)DBD + D\mathrm{Diag}\big(\overline{A}DB\big)D\\ &= D\overline{\overline{A}DB}D + \mathrm{Diag}(A)DBD + \mathrm{Diag}(\overline{A}DBD).\end{aligned} \tag{60}$$

Applying Lemma 15(a) twice, $D\overline{\overline{A}DB}D = \overline{D\overline{A}DBD}$ has no non-zero diagonal entries. Moreover, taking $i, j \in \{1, \ldots, d\}$ distinct,

$$\Big(D\overline{\overline{A}DB}D\Big)_{ij} = D_{ii}D_{jj}\Big(\overline{\overline{A}DB}\Big)_{ij} = D_{ii}D_{jj}\sum_{k=1}^{d} A_{ik}\mathbb{1}_{\{i\neq k\}}D_{kk}B_{kj},$$

so that both $i \neq j$ and $i \neq k$. Therefore,

$$\begin{aligned}\mathbb{E}\Big[D\overline{\overline{A}DB}D\Big] &= \mathbb{E}[D]\mathbb{E}\Big[\overline{\overline{A}DB}D\Big] \\ &= p\mathbb{E}\big[\overline{A}DBD\big] - p\mathrm{Diag}\big(\overline{A}\mathbb{E}[DBD]\big) \\ &= pA\mathbb{E}[DBD] - p\mathrm{Diag}(A)\mathbb{E}[DBD] - p\mathrm{Diag}\big(\overline{A}\mathbb{E}[DBD]\big).\end{aligned}$$

Reinserting this expression into the expectation of (60) and applying Part (i) of the lemma now results in the claimed identity

$$\begin{aligned}\mathbb{E}\big[DADBD\big] &= pA\mathbb{E}[DBD] + (1-p)\mathrm{Diag}(A)\mathbb{E}[DBD] + (1-p)\mathrm{Diag}\big(\overline{A}\mathbb{E}[DBD]\big) \\ &= p\big(pA + (1-p)\mathrm{Diag}(A)\big)B_p + p(1-p)\mathrm{Diag}\big(\overline{A}B_p\big) \\ &= pA_pB_p + p(1-p)\mathrm{Diag}\big(\overline{A}B_p\big) \\ &= pA_pB_p + p^2(1-p)\mathrm{Diag}\big(\overline{A}B\big),\end{aligned}$$

where $\mathrm{Diag}(\overline{A}\mathrm{Diag}(B)) = 0$ by Lemma 15(a).

(c) Following a similar strategy as in Part (ii), observe that

$$\begin{aligned}DADBDCD &= DAD\overline{B}DCD + DAD\mathrm{Diag}(B)DCD \\ &= DAD\overline{B}D\overline{C}D + DAD\mathrm{Diag}(B)CD + DAD\overline{B}D\mathrm{Diag}(C)D \\ &= D\overline{\overline{ADB}}D\overline{C}D + DAD\mathrm{Diag}(B)CD + DAD\overline{B}\mathrm{Diag}(C)D \\ &\quad + D\mathrm{Diag}(AD\overline{B})\overline{C}D.\end{aligned} \tag{61}$$

By construction of the latter matrix,

$$\begin{aligned}\big(D\overline{\overline{ADB}}D\overline{C}D\big)_{ij} &= D_{ii}D_{jj}\sum_{k=1}^{d}\big(\overline{\overline{ADB}}\big)_{ik}\big(D\overline{C}\big)_{kj} \\ &= D_{ii}D_{jj}\sum_{k=1}^{d}\mathbb{1}_{\{k\neq j\}}D_{kk}C_{kj}\sum_{\ell=1}^{d}\mathbb{1}_{\{i\neq k\}}A_{i\ell}D_{\ell\ell}\mathbb{1}_{\{\ell\neq k\}}B_{\ell k},\end{aligned}$$

meaning k is always distinct from the other indices. The k index corresponds to the third D-matrix from the left, so this proves $\mathbb{E}\big[D\overline{\overline{ADB}}D\overline{C}D\big] = p\mathbb{E}\big[D\overline{\overline{ADB}}\,\overline{C}D\big]$. Reversing the overlines of the latter expression in order, note that

$$\begin{aligned}D\overline{\overline{ADB}}\,\overline{C}D = DADBCD &- D\mathrm{Diag}(AD\overline{B})\overline{C}D - DAD\overline{B}\mathrm{Diag}(C)D \\ &- DAD\mathrm{Diag}(B)CD.\end{aligned}$$

Note that the subtracted terms match those added to $D\overline{AD\overline{B}}D\overline{C}D$ in (61) exactly. In turn, these identities prove

$$\begin{aligned}
&\mathbb{E}\big[DADBDCD\big] \\
&= \mathbb{E}\Big[D\overline{AD\overline{B}}D\overline{C}D + DAD\mathrm{Diag}(B)CD + DAD\overline{B}\mathrm{Diag}(C)D + D\mathrm{Diag}(AD\overline{B})\overline{C}D\Big] \\
&= p\mathbb{E}\Big[D\overline{AD\overline{B}}\,\overline{C}D\Big] + \mathbb{E}\Big[DAD\mathrm{Diag}(B)CD + DAD\overline{B}\mathrm{Diag}(C)D + D\mathrm{Diag}(AD\overline{B})\overline{C}D\Big] \\
&= p\mathbb{E}\Big[DADBCD\Big] + (1-p)\mathbb{E}\Big[DAD\mathrm{Diag}(B)CD\Big] \\
&\qquad + (1-p)\mathbb{E}\Big[DAD\overline{B}\mathrm{Diag}(C)D\Big] + (1-p)\mathbb{E}\Big[D\mathrm{Diag}(AD\overline{B})\overline{C}D\Big] \\
&= \mathbb{E}\big[DADB_pCD\big] + (1-p)\Big(\mathbb{E}\big[DAD\overline{B}\mathrm{Diag}(C)D\big] + \mathbb{E}\big[D\mathrm{Diag}(AD\overline{B})\overline{C}D\big]\Big). \qquad (62)
\end{aligned}$$

The first and second term in the last equality may be computed via Part (ii) of the lemma, whereas the third term remains to be treated.

By definition, the diagonal entries of $\overline{B}$ and $\overline{C}$ are all zero, so $\big(D\mathrm{Diag}(AD\overline{B})\overline{C}D\big)_{ii}$ equals 0 for all $i \in \{1,\dots,d\}$. Moreover, taking $i \neq j$ implies

$$\begin{aligned}
\big(D\mathrm{Diag}(AD\overline{B})\overline{C}D\big)_{ij} &= D_{ii}\big(AD\overline{B}\big)_{ii}C_{ij}D_{jj} \\
&= D_{ii}C_{ij}D_{jj}\sum_{k=1}^{d} A_{ik}D_{kk}B_{ki}\mathbb{1}_{\{k\neq i\}}.
\end{aligned}$$

On the set $\{i \neq j\} \cap \{i \neq k\}$, the entry D_{ii} is independent of D_{jj} and D_{kk}. Consequently,

$$\begin{aligned}
\mathbb{E}\Big[\big(D\mathrm{Diag}(AD\overline{B})\overline{C}D\big)_{ij}\Big] &= \mathbb{E}[D_{ii}]\sum_{k=1}^{d} A_{ik}\overline{B}_{ki}C_{ij}\mathbb{E}[D_{jj}D_{kk}] \\
&= \sum_{k=1}^{d} A_{ik}\overline{B}_{ki}C_{ij}\big(p^3 + p^2(1-p)\mathbb{1}_{\{k=j\}}\big) \\
&= p^3\sum_{k=1}^{d} A_{ik}\overline{B}_{ki}C_{ij} + p^2(1-p)A_{ij}\overline{B}_{ji}C_{ij}.
\end{aligned}$$

In matrix form, the previous equation reads

$$\mathbb{E}\Big[D\mathrm{Diag}\big(AD\overline{B}\big)\overline{C}D\Big] = p^3\mathrm{Diag}(A\overline{B})\overline{C} + p^2(1-p)A\odot\overline{B}^{\top}\odot C$$

where $\odot$ denotes the Hadamard product.

Reinserting the computed expressions into (62), as well as noting that $\big(\overline{B}\mathrm{Diag}(C)\big)_p = \overline{B}_p\mathrm{Diag}(C)$ and $\mathrm{Diag}\big(\overline{A}\,\overline{B}\mathrm{Diag}(C)\big) = \mathrm{Diag}(\overline{A}\,\overline{B})\mathrm{Diag}(C)$ yields

$$\begin{aligned}
\mathbb{E}\big[DADBDCD\big] = {}& pA_p\big(B_pC\big)_p + p^2(1-p)\mathrm{Diag}\big(\overline{A}B_pC\big) \\
&+ p(1-p)A_p\overline{B}_p\mathrm{Diag}(C) + p^2(1-p)^2\mathrm{Diag}(\overline{A}\,\overline{B})\mathrm{Diag}(C) \qquad (63) \\
&+ p^3(1-p)\mathrm{Diag}(A\overline{B})\overline{C} + p^2(1-p)^2A\odot\overline{B}^{\top}\odot C.
\end{aligned}$$

Next, using the identity $\overline{B}_p = B_p - \mathrm{Diag}(B)$ of Lemma 15(b), we combine the first and third terms of the latter display into

$$\begin{aligned} pA_p\big(B_pC\big)_p + p(1-p)A_p\overline{B}_p\mathrm{Diag}(C) &= pA_pB_p\big(pC + (1-p)\mathrm{Diag}(C)\big) \\ &\quad + p(1-p)A_p\mathrm{Diag}\big(B_pC\big) \\ &\quad - p(1-p)A_p\mathrm{Diag}(B)\mathrm{Diag}(C) \\ &= pA_pB_pC_p + p^2(1-p)A_p\mathrm{Diag}\big(\overline{B}C\big). \end{aligned}$$

Regarding the second term of (63), observe that

$$\begin{aligned} \mathrm{Diag}\big(\overline{A}B_pC\big) &= \mathrm{Diag}\big(\overline{A}B_p\overline{C}\big) + \mathrm{Diag}\big(\overline{A}B_p\big)\mathrm{Diag}(C) \\ &= \mathrm{Diag}\big(\overline{A}B_p\overline{C}\big) + p\mathrm{Diag}(\overline{A}B)\mathrm{Diag}(C), \end{aligned}$$

where the second equality follows from $\mathrm{Diag}\big(\overline{A}\mathrm{Diag}(B)\big) = 0$. Lastly, $\mathrm{Diag}(\overline{A}\,\overline{B}) = \mathrm{Diag}(\overline{A}B) = \mathrm{Diag}(A\overline{B})$ by Lemma 15(c), so the fourth and fifth term of (63) combine into

$$\begin{aligned} &p\mathrm{Diag}(A\overline{B})\overline{C} + (1-p)\mathrm{Diag}(\overline{A}\,\overline{B})\mathrm{Diag}(C) \\ &= \mathrm{Diag}(A\overline{B})\big(p\overline{C} + \mathrm{Diag}(C)\big) - p\mathrm{Diag}(A\overline{B})\mathrm{Diag}(C) \\ &= \mathrm{Diag}(A\overline{B})C_p - p\mathrm{Diag}(A\overline{B})\mathrm{Diag}(C), \end{aligned}$$

where the common factor $p^2(1-p)$ is omitted in the display.

Using these identities, Equation (63) now turns into

$$\begin{aligned} \mathbb{E}\big[DADBDCD\big] &= pA_pB_pC_p + p^2(1-p)^2A \odot \overline{B}^\top \odot C \\ &\quad + p^2(1-p)\Big(\mathrm{Diag}\big(\overline{A}B_pC\big) - p\mathrm{Diag}(A\overline{B})\mathrm{Diag}(C)\Big) \\ &\quad + p^2(1-p)\Big(A_p\mathrm{Diag}\big(\overline{B}C\big) + \mathrm{Diag}(A\overline{B})C_p\Big). \end{aligned}$$

Noting that $\mathrm{Diag}\big(\overline{A}B_pC\big) - p\mathrm{Diag}(A\overline{B})\mathrm{Diag}(C)$ equals $\mathrm{Diag}\big(\overline{A}B_p\overline{C}\big)$ finishes the proof.

■

In principle, any computations involving higher moments of D may be accomplished with the proof strategy of Lemma 16. A particular covariance matrix is needed in Section B, which will be given in the next lemma.

Lemma 17 *Given a symmetric matrix A, as well as vectors $\mathbf{u}$ and $\mathbf{v}$,*

$$\begin{aligned} \frac{\mathrm{Cov}\big(D\mathbf{u} + D\overline{A}(D-pI)\mathbf{v}\big)}{p(1-p)} &= \mathrm{Diag}(\mathbf{u}\mathbf{u}^\top) + p\overline{A}\mathrm{Diag}(\mathbf{v}\mathbf{u}^\top) + p\mathrm{Diag}(\mathbf{u}\mathbf{v}^\top)\overline{A} \\ &\quad + p\big(\overline{A}\mathrm{Diag}(\mathbf{v}\mathbf{v}^\top)\overline{A}\big)_p + p(1-p)A \odot \overline{\mathbf{v}\mathbf{v}^\top} \odot A. \end{aligned}$$

Proof The covariance of the sum is given by

$$\begin{aligned}\mathrm{Cov}\big(D\mathbf{u} + D\overline{A}(D-pI)\mathbf{v}\big) &= \mathrm{Cov}\big(D\mathbf{u}\big) + \mathrm{Cov}\big(D\overline{A}(D-pI)\mathbf{v}\big) + \mathrm{Cov}\big(D\mathbf{u}, D\overline{A}(D-pI)\mathbf{v}\big) \\ &\quad + \mathrm{Cov}\big(D\overline{A}(D-pI)\mathbf{v}, D\mathbf{u}\big) \\ &=: T_1 + T_2 + T_3 + T_4 \end{aligned} \tag{64}$$

where each of the latter terms will be treated separately. To this end, set $B_{\mathbf{u}} := \mathbf{u}\mathbf{u}^\top$, $B_{\mathbf{v}} := \mathbf{v}\mathbf{v}^\top$, $B_{\mathbf{u},\mathbf{v}} := \mathbf{u}\mathbf{v}^\top$ and $B_{\mathbf{v},\mathbf{u}} := \mathbf{v}\mathbf{u}^\top$.

First, recall that $\mathbb{E}\big[DB_{\mathbf{u}}D\big] = p(B_{\mathbf{u}})_p$ by Lemma 16(a) and $\mathbb{E}[D] = p$, so the definition of covariance implies

$$\begin{aligned} T_1 &= \mathbb{E}\big[DB_{\mathbf{u}}D\big] - \mathbb{E}[D]B_{\mathbf{u}}\mathbb{E}[D] = p^2B_{\mathbf{u}} + p(1-p)\mathrm{Diag}(B_{\mathbf{u}}) - p^2B_{\mathbf{u}} \\ &= p(1-p)\mathrm{Diag}(B_{\mathbf{u}}). \end{aligned} \tag{65}$$

Moving on to T_4, observe that

$$T_4 = \mathrm{Cov}\big(D\overline{A}(D-pI)\mathbf{v}, D\mathbf{u}\big) = \mathbb{E}\big[D\overline{A}(D-pI)B_{\mathbf{v},\mathbf{u}}D\big] - \mathbb{E}\big[D\overline{A}(D-pI)\mathbf{v}\big]\mathbb{E}\big[D\mathbf{u}\big]^\top.$$

By the same argument as in (29),

$$\mathbb{E}\big[D\overline{A}(D-pI)\mathbf{v}\big] = \mathbb{E}\big[D\overline{A}D\big]\mathbf{v} - p\mathbb{E}\big[D\overline{A}\big]\mathbf{v} = p^2\overline{A}\mathbf{v} - p^2\overline{A}\mathbf{v} = 0 \tag{66}$$

so that in turn

$$T_4 = \mathbb{E}\big[D\overline{A}DB_{\mathbf{v},\mathbf{u}}D\big] - p\mathbb{E}\big[D\overline{A}B_{\mathbf{v},\mathbf{u}}D\big].$$

Recall $\overline{A}_p = p\overline{A}$ from Lemma 15(b). Applying Lemma 16(b) for the first term in the previous display and Lemma 16(a) for the second term now leads to

$$\begin{aligned} T_4 &= p^2\overline{A}(B_{\mathbf{v},\mathbf{u}})_p + p^2(1-p)\mathrm{Diag}\big(\overline{A}B_{\mathbf{v},\mathbf{u}}\big) - p^2\big(\overline{A}B_{\mathbf{v},\mathbf{u}}\big)_p \\ &= p^3\overline{A}B_{\mathbf{v},\mathbf{u}} + p^2(1-p)\overline{A}\mathrm{Diag}(B_{\mathbf{v},\mathbf{u}}) + p^2(1-p)\mathrm{Diag}\big(\overline{A}B_{\mathbf{v},\mathbf{u}}\big) - p^3\overline{A}B_{\mathbf{v},\mathbf{u}} \\ &\quad - p^2(1-p)\mathrm{Diag}\big(\overline{A}B_{\mathbf{v},\mathbf{u}}\big) \\ &= p^2(1-p)\overline{A}\mathrm{Diag}(B_{\mathbf{v},\mathbf{u}}). \end{aligned} \tag{67}$$

Using a completely analogous argument, the reflected term T_3 satisfies

$$T_3 = p^2(1-p)\mathrm{Diag}\big(B_{\mathbf{u},\mathbf{v}}\big)\overline{A}. \tag{68}$$

The last term T_2 necessitates another decomposition into four sub-problems. First, recall from (66) that $\mathbb{E}\big[D\overline{A}(D-pI)\mathbf{v}\big]$ vanishes, which leads to

$$\begin{aligned} T_2 &= \mathrm{Cov}\big(D\overline{A}(D-pI)\mathbf{v}\big) \\ &= \mathbb{E}\big[D\overline{A}(D-pI)B_{\mathbf{v}}(D-pI)\overline{A}D\big] \\ &= \mathbb{E}\big[D\overline{A}DB_{\mathbf{v}}D\overline{A}D\big] - p\mathbb{E}\big[D\overline{A}B(D-pI)\overline{A}D\big] - p\mathbb{E}\big[D\overline{A}(D-pI)B_{\mathbf{v}}\overline{A}D\big] \\ &\quad + p^2\mathbb{E}\big[D\overline{A}B_{\mathbf{v}}\overline{A}D\big] \\ &=: T_{2,1} - T_{2,2} - T_{2,3} - T_{2,4} \end{aligned} \tag{69}$$

where the last term is negative since $-pB_{\mathbf{v}}(D-pI)-p(D-pI)B_{\mathbf{v}} = -pB_{\mathbf{v}}D-pDB_{\mathbf{v}}+2p^2B_{\mathbf{v}}$. Recall once more the identity $\overline{A}_p = p\overline{A}$ from Lemma 15(b) and apply Lemma 16(c), to rewrite $T_{2,1}$ as

$$\begin{aligned}
T_{2,1} &= \mathbb{E}\big[D\overline{A}DB_{\mathbf{v}}D\overline{A}D\big] \\
&= p^3\overline{A}(B_{\mathbf{v}})_p\overline{A} + p^2(1-p)\Big(\mathrm{Diag}\big(\overline{A}(B_{\mathbf{v}})_p\overline{A}\big) + p\overline{A}\mathrm{Diag}\big(\overline{B_{\mathbf{v}}}A\big) + p\mathrm{Diag}\big(A\overline{B_{\mathbf{v}}}\big)\overline{A}\Big) \\
&\quad + p^2(1-p)^2A\odot\overline{B_{\mathbf{v}}}\odot A.
\end{aligned}$$

As for $T_{2,2}$, start by noting that

$$\begin{aligned}
T_{2,2} &= p\mathbb{E}\big[D\overline{A}B_{\mathbf{v}}(D-pI)\overline{A}D\big] \\
&= p\mathbb{E}\big[D\overline{A}B_{\mathbf{v}}D\overline{A}D\big] - p^2\mathbb{E}\big[D\overline{A}B_{\mathbf{v}}\overline{A}D\big] \\
&= p^3\big(\overline{A}B_{\mathbf{v}}\big)_p\overline{A} + p^3(1-p)\mathrm{Diag}\big(\overline{\overline{A}B_{\mathbf{v}}\overline{A}}\big) - p^3\big(\overline{A}B_{\mathbf{v}}\overline{A}\big)_p,
\end{aligned}$$

where Lemma 16(a) computes the second expectation and Lemma 16(b) the first expectation. To progress, note first the identities

$$\begin{aligned}
\big(\overline{A}B_{\mathbf{v}}\big)_p\overline{A} - \big(\overline{A}B_{\mathbf{v}}\overline{A}\big)_p &= (1-p)\Big(\mathrm{Diag}\big(\overline{A}B_{\mathbf{v}}\big)\overline{A} - \mathrm{Diag}\big(\overline{A}B_{\mathbf{v}}\overline{A}\big)\Big) \\
\mathrm{Diag}\big(\overline{\overline{A}B_{\mathbf{v}}\overline{A}}\big) &= \mathrm{Diag}\big(\overline{A}B_{\mathbf{v}}\overline{A}\big)
\end{aligned}$$

so that

$$T_{2,2} = p^3(1-p)\Big(\mathrm{Diag}\big(\overline{A}B_{\mathbf{v}}\big)\overline{A} - \mathrm{Diag}\big(\overline{A}B_{\mathbf{v}}\overline{A}\big) + \mathrm{Diag}\big(\overline{A}B_{\mathbf{v}}\overline{A}\big)\Big) = p^3(1-p)\mathrm{Diag}\big(\overline{A}B_{\mathbf{v}}\big)\overline{A}.$$

By symmetry, the reflected term $T_{2,3}$ then also satisfies $T_{2,3} = p^3(1-p)\overline{A}\mathrm{Diag}\big(B_{\mathbf{v}}\overline{A}\big)$. Lastly, applying Lemma 16(a) to $T_{2,4}$ results in $T_{2,4} = p^3\big(\overline{A}B_{\mathbf{v}}\overline{A}\big)_p$. To finish the treatment of T_2, inserting the computed expressions for $T_{2,1}, T_{2,2}, T_{2,3}$, and $T_{2,4}$ into (69) and combining like terms now leads to

$$\begin{aligned}
T_2 &= p^3\overline{A}(B_{\mathbf{v}})_p\overline{A} + p^2(1-p)^2A\odot\overline{B_{\mathbf{v}}}\odot A \\
&\quad + p^2(1-p)\Big(\mathrm{Diag}\big(\overline{A}(B_{\mathbf{v}})_p\overline{A}\big) + p\overline{A}\mathrm{Diag}\big(\overline{B_{\mathbf{v}}}A\big) + \mathrm{Diag}\big(A\overline{B_{\mathbf{v}}}\big)p\overline{A}\Big) \\
&\quad - p^3(1-p)\mathrm{Diag}\big(\overline{A}B_{\mathbf{v}}\big)\overline{A} - p^3(1-p)\overline{A}\mathrm{Diag}\big(B_{\mathbf{v}}\overline{A}\big) + p^3\big(\overline{A}B_{\mathbf{v}}\overline{A}\big)_p \\
&= p^3\Big(\overline{A}(B_{\mathbf{v}})_p\overline{A} - \big(\overline{A}B_{\mathbf{v}}\overline{A}\big)_p\Big) + p^2(1-p)\mathrm{Diag}\big(\overline{A}(B_{\mathbf{v}})_p\overline{A}\big) + p^2(1-p)^2A\odot\overline{B_{\mathbf{v}}}\odot A \\
&= p^3(1-p)\overline{A}\mathrm{Diag}(B_{\mathbf{v}})\overline{A} + p^2(1-p)^2\mathrm{Diag}\big(\overline{A}\mathrm{Diag}(B_{\mathbf{v}})\overline{A}\big)p^2(1-p)^2A\odot\overline{B_{\mathbf{v}}}\odot A \\
&= p^2(1-p)\big(\overline{A}\mathrm{Diag}(B_{\mathbf{v}})\overline{A}\big)_p + p^2(1-p)^2A\odot\overline{B_{\mathbf{v}}}\odot A.
\end{aligned}$$

To conclude the proof, insert this expression for T_2 into (64), together with T_1 as in (65), T_3 as in (68), and T_4 as in (67) to obtain the desired identity

$$\begin{aligned}
\mathrm{Cov}\big(D\mathbf{u} + D\overline{A}(D-pI)\mathbf{v}\big) &= p(1-p)\mathrm{Diag}(B_{\mathbf{u}}) + p^2(1-p)\big(\overline{A}\mathrm{Diag}(B_{\mathbf{v}})\overline{A}\big)_p \\
&\quad + p^2(1-p)^2A\odot\overline{B}_{\mathbf{v}}\odot A \\
&\quad + p^2(1-p)\big(\mathrm{Diag}(B_{\mathbf{u},\mathbf{v}})\overline{A} + \overline{A}\mathrm{Diag}(B_{\mathbf{v},\mathbf{u}})\big).
\end{aligned}$$

■

Appendix D. Auxiliary Results

Below we collect identities and definitions referenced in other sections.

Neumann series: Let V denote a real or complex Banach space with norm $\|\cdot\|$. Recall that the operator norm of a linear operator λ on V is given by $\|\lambda\|_{\text{op}} = \sup_{\mathbf{v}\in V:\|\mathbf{v}\|\leq 1}\|\lambda(\mathbf{v})\|$.

Lemma 18 (Helemskii (2006), Proposition 5.3.4) *Suppose $\lambda : V \to V$ is bounded, linear, and satisfies $\|\mathrm{id} - \lambda\|_{\text{op}} < 1$, then λ is invertible and $\lambda^{-1} = \sum_{i=0}^{\infty}\big(\mathrm{id} - \lambda\big)^i$.*

Bounds on singular values: Recall that $\|\cdot\|_2$ denotes the Euclidean norm on $\mathbb{R}^d$ and $\|\cdot\|$ the spectral norm on $\mathbb{R}^{d\times d}$, which is given by the largest singular value $\sigma_{\max}(\cdot)$. The spectral norm is sub-multiplicative in the sense that $\|AB\| \leq \|A\|\|B\|$. The spectral norm of a vector $\mathbf{v} \in \mathbb{R}^d$, viewed as a linear functional on $\mathbb{R}^d$, is given by $\|\mathbf{v}\| = \|\mathbf{v}\|_2$, proving that

$$\big\|\mathbf{v}\mathbf{w}^\top\big\| \leq \|\mathbf{v}\|_2\|\mathbf{w}\|_2 \tag{70}$$

for any vectors $\mathbf{v}$ and $\mathbf{w}$ of the same length.

Recall the definitions $\overline{A} = A - \mathrm{Diag}(A)$ and $A_p = pA + (1-p)\mathrm{Diag}(A)$ with $p \in (0,1)$.

Lemma 19 *Given $d\times d$ matrices A and B, the inequalities $\big\|\mathrm{Diag}(A)\big\| \leq \|A\|$, $\|A_p\| \leq \|A\|$, and $\|A \odot B\| \leq \|A\| \cdot \|B\|$ hold. Moreover, if A is symmetric and positive semi-definite, then also $\big\|\overline{A}\big\| \leq \|A\|$.*

Proof For any matrix A, the maximal singular value $\sigma_{\max}(A)$ can be computed from the variational formulation $\sigma_{\max}(A) = \max_{\mathbf{v}\in\mathbb{R}^d\setminus\{\mathbf{0}\}}\|A\mathbf{v}\|_2/\|\mathbf{v}\|_2$, see Horn and Johnson (2013), Theorem 4.2.6.

Let $\mathbf{e}_i$ denote the i^{th} standard basis vector. The variational formulation of the maximal singular value implies $\|\mathrm{Diag}(A)\|^2 = \max_i A_{ii}^2$ which is bounded by $\max_i \sum_{k=1}^d A_{ki}^2 = \max_i \big(A^\top A\big)_{ii} = \max_i \mathbf{e}_i^\top A^\top A\mathbf{e}_i$. The latter is further bounded by $\big\|A^\top A\big\|$, proving the first statement. The second inequality follows from the first since

$$\|A_p\| \leq p\|A\| + (1-p)\big\|\mathrm{Diag}(A)\big\| \leq p\|A\| + (1-p)\|A\| = \|A\|.$$

For the inequality concerning the Hadamard product, see Theorem 5.5.7 of Horn and Johnson (1991).

For the last inequality, note that semi-definiteness entails $\min_i \mathrm{Diag}(A)_{ii} \geq 0$. Fixing $\mathbf{v} \in \mathbb{R}^d$, this ensures $\mathbf{v}^\top\overline{A}\mathbf{v} \leq \mathbf{v}^\top A\mathbf{v}$, which completes the proof. ■

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: A system for Large-Scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation*, pages 265–283. USENIX Association, 2016. ISBN 978-1-931971-33-1.

Donald W. K. Andrews. Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory*, 4(3):458–467, 1988. ISSN 0266-4666.

Raman Arora, Peter Bartlett, Poorya Mianjy, and Nathan Srebro. Dropout: Explicit forms and capacity control. In *38th International Conference on Machine Learning*, pages 351–361. Proceedings of Machine Learning Research, 2021.

Jimmy Ba and Brendan Frey. Adaptive dropout for training deep neural networks. In *Advances in Neural Information Processing Systems 26*, pages 3084–3092. Curran Associates, Inc., 2013. ISBN 978-1-632660-24-4.

Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 11(1):307–353, 2022. ISSN 2049-8772.

Pierre Baldi and Peter J. Sadowski. Understanding dropout. In *Advances in Neural Information Processing Systems 26*, pages 2814–2822. Curran Associates, Inc., 2013. ISBN 978-1-632660-24-4.

Peter L. Bartlett, Philip M. Long, and Olivier Bousquet. The dynamics of sharpness-aware minimization: Bouncing across ravines and drifting towards wide minima. *Journal of Machine Learning Research*, 24(316):1–36, 2023. ISSN 1533-7928.

Thijs Bos and Johannes Schmidt-Hieber. Convergence guarantees for forward gradient descent in the linear regression model. arXiv:2309.15001 [math.ST], 2023.

Jacopo Cavazza, Pietro Morerio, Benjamin Haeffele, Connor Lane, Vittorio Murino, and Rene Vidal. Dropout as a low-rank regularizer for matrix factorization. In *21st International Conference on Artificial Intelligence and Statistics*, pages 435–444. Proceedings of Machine Learning Research, 2018.

François Chollet et al. Keras. `https://keras.io`, 2015.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989. ISSN 0932-4194.

Steffen Dereich and Sebastian Kassing. Central limit theorems for stochastic gradient descent with averaging for stable manifolds. *Electronic Journal of Probability*, 28:1–48, 2023. ISSN 1083-6489.

Lutz Dümbgen, Richard J. Samworth, and Dominic Schuhmacher. Stochastic search for semiparametric linear regression models. In *From Probability to Statistics and Back: High-Dimensional Models and Processes. A Festschrift in Honor of Jon A. Wellner*, pages 78–90. Institute of Mathematical Statistics, 2013. ISBN 978-0-940600-83-6.

Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference. Algorithms, Evidence, and Data Science*, volume 5 of *Institute of Mathematical Statistics Monographs*. Cambridge University Press, 2016. ISBN 978-1-107-14989-2; 978-1-316-57653-3.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *33rd International Conference on International Conference on Machine Learning*, pages 1050–1059. Proceedings of Machine Learning Research, 2016a.

Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems 29*, pages 1027–1035. Curran Associates, Inc., 2016b. ISBN 978-1-510838-81-9.

Wei Gao and Zhi-Hua Zhou. Dropout Rademacher complexity of deep neural networks. *Science China Information Sciences*, 59(7):2104:1–12, 2016. ISSN 1869-1919.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. Adaptive Computation and Machine Learning. MIT Press, 2016. ISBN 978-0-262-03561-3; 978-0-262-33743-4.

László Györfi and Harro Walk. On the averaged stochastic approximation for linear regression. *SIAM Journal on Control and Optimization*, 34(1):31–61, 1996. ISSN 0363-0129.

Aleksandr Ya. Helemskii. *Lectures and Exercises on Functional Analysis*, volume 233 of *Translations of Mathematical Monographs*. American Mathematical Society, 2006. ISBN 0-8218-4098-3.

David P. Helmbold and Philip M. Long. Surprising properties of dropout in deep networks. In *Conference on Learning Theory*, pages 1123–1146. Proceedings of Machine Learning Research, 2017.

Jonathan Hill and Liang Peng. Unified interval estimation for random coefficient autoregressive models. *Journal of Time Series Analysis*, 35(3):282–297, 2014. ISSN 0143-9782.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991. ISBN 0-521-30587-X.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2013. ISBN 978-0-521-54823-6; 978-0-521-83940-2.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991. ISSN 0893-6080.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *22nd ACM International Conference on Multimedia*, pages 675–678. Association for Computing Machinery, 2014. ISBN 978-1-4503-3063-3.

Diederik P. Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems 28*, pages 2575–2583. Curran Associates, Inc., 2015. ISBN 978-1-510825-02-4.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. ISBN 978-1-627480-03-1.

Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993. ISSN 0893-6080.

Oxana A. Manita, Mark A. Peletier, Jacobus W. Portegies, Jaron Sanders, and Albert Senen-Cerda. Universal approximation in dropout neural networks. *Journal of Machine Learning Research*, 23(19):1–46, 2022. ISSN 1533-7928.

David McAllester. A PAC-Bayesian tutorial with a dropout bound. arXiv:1307.2118 [cs.LG], 2013.

Poorya Mianjy and Raman Arora. On dropout and nuclear norm regularization. In *36th International Conference on Machine Learning*, pages 4575–4584. Proceedings of Machine Learning Research, 2019.

Poorya Mianjy and Raman Arora. On convergence and generalization of dropout training. In *Advances in Neural Information Processing Systems 33*, pages 21151–21161. Curran Associates, Inc., 2020. ISBN 978-1-713829-54-6.

Poorya Mianjy, Raman Arora, and Rene Vidal. On the implicit bias of dropout. In *35th International Conference on Machine Learning*, pages 3540–3548. Proceedings of Machine Learning Research, 2018.

Reza Moradi, Reza Berangi, and Behrouz Minaei. A survey of regularization strategies for deep models. *Artificial Intelligence Review*, 53(6):3947–3986, 2020. ISSN 0269–2821.

Gabin Maxime Nguegnang, Holger Rauhut, and Ulrich Terstiege. Convergence of gradient descent for learning linear neural networks. arXiv:2108.02040 [cs.LG], 2021.

Des F. Nicholls and Barry G. Quinn. *Random Coefficient Autoregressive Models: An Introduction*, volume 11 of *Lecture Notes in Statistics*. Springer New York, 1982. ISBN 978-0-387-90766-6.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019. ISBN 978-1-713807-93-3.

Boris Polyak. New method of stochastic approximation type. *Avtomatica i Telemekhanika*, 7:98–107, 1990. ISSN 0005–2310.

Boris Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. ISSN 0363-0129.

Marta Regis, Paulo Serra, and Edwin R. van den Heuvel. Random autoregressive models: a structured overview. *Econometric Reviews*, 41(2):207–230, 2022. ISSN 0747-4938.

David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical Report 781. Cornell University Operations Research and Industrial Engineering, 1988.

Claudio Filipi Gonçalves Dos Santos and João Paulo Papa. Avoiding overfitting: A survey on regularization methods for convolutional neural networks. *ACM Computing Surveys*, 54(10s):213:1–25, 2022. ISSN 0360-0300.

Johannes Schmidt-Hieber and Wouter M. Koolen. Hebbian learning inspired estimation of the linear regression parameters from queries. arXiv:2311.03483 [math.ST], 2023.

Albert Senen-Cerda and Jaron Sanders. Asymptotic convergence rate of dropout on shallow linear neural networks. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(2):32:1–53, 2022. ISSN 2476-1249.

Nitish Srivastava, Geoffrey 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. ISSN 1533-7928.

Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023. ISSN 1533-7928.

Stefan Wager, Sida Wang, and Percy S Liang. Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems 26*, pages 351–359. Curran Associates, Inc., 2013. ISBN 978-1-632660-24-4.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *30th International Conference on Machine Learning*, pages 1058–1066. Proceedings of Machine Learning Research, 2013.

Sida Wang and Christopher Manning. Fast dropout training. In *30th International Conference on Machine Learning*, pages 118–126. Proceedings of Machine Learning Research, 2013.

Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. In *37th International Conference on Machine Learning*, pages 10181–10192. Proceedings of Machine Learning Research, 2020.

Haibing Wu and Xiaodong Gu. Towards dropout training for convolutional neural networks. *Neural Networks*, 71:1–10, 2015. ISSN 0893-6080.

Ke Zhai and Huan Wang. Adaptive dropout with Rademacher complexity regularization. In *6th International Conference on Learning Representations*, 2018.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024. ISSN 1533-7928.

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, 2021. ISSN 0162-1459.