STRUCTURED SAMPLING FOR STOCHASTIC DEPTH

Anonymous authors
Paper under double-blind review

ABSTRACT

Stochastic depth is a key technique for training residual networks, which randomly drops residual blocks during training to mitigate overfitting. Standard estimation of the expected gradient may result in high variance and unstable convergence due to averaging over i.i.d. masks. To address this issue, we formulate gradient estimation under stochastic depth as a Fourier projection problem over the Boolean cube. Building on this formulation, we develop two structured sampling designs that eliminate selected spectral components while preserving the original expectation: a coordinate-balanced design that removes first-order components at low sample cost, and an equiprobable stratification framework that cancels a user-specified set of modes exactly and guarantees covariance dominance. We derive explicit lower bounds on the variance reduction for both designs, and analyze their implications for stochastic gradient descent, including improved parameter-error decay and lower averaged training loss, under mild assumptions. Numerical experiments on CIFAR-10 and CIFAR-100 show that our designs substantially reduce gradient-estimation variance and slightly improving the accuracy and stability of training.

1 INTRODUCTION

Stochastic depth is a standard trick for training residual networks (ResNets) (Hayou & Ayed, 2021; Huang et al., 2016; Müller et al., 2026), which randomly drops residual blocks during training to reduce overfitting and improve generalization. The standard implementation is to minimize the expected loss $J(\theta) = \mathbb{E}_Z[f_\theta(Z)]$ and estimate the gradient $\mathbb{E}_Z[\nabla f_\theta(Z)]$, where Z is a stochastic mask and $f_\theta(Z)$ is the loss incurred by the network with mask Z.. In practice, this expectation is estimated by averaging over M independent and identically distributed (i.i.d.) samples of the random mask Z:

$$\nabla J(\theta) \approx \frac{1}{M}\sum_{i=1}^{M} \nabla f_\theta(Z_i), \qquad Z_i \overset{\text{i.i.d.}}{\sim} p(Z).$$

Although the standard i.i.d. estimator is unbiased, its variance can be substantial in the finite-sample regime, which in turn slows down convergence and can destabilize training.

Variance reduction has been extensively studied in stochastic optimization, *e.g.*, using control variates and their recursive variants (Johnson & Zhang, 2013; Nguyen et al., 2017), low-order moment approximations (Wang et al., 2013), or antithetic sampling with copulas for discrete variables (Dimitriev & Zhou, 2021a;b). However, these methods target the randomness of the *data*, whereas in stochastic depth the dominant source of variance is the *mask*. While structured dropout has been extensively studied (Li et al., 2023), its objects are paths, heads, or features (Larsson et al., 2017; Zhou et al., 2020; Ghiasi et al., 2018), rather than the mask over residual blocks.

Hence, we ask a sampling-design question:

> *Can we design a structured mask sampler that keeps the gradient estimator unbiased while reducing its variance?*

In this work, we view the estimation under stochastic depth as a Fourier projection problem on the Boolean cube O'Donnell (2014), where the mask takes its values. In the Fourier expansion, the expected gradient is its zeroth-order component, and the estimator variance is governed by the remaining components. This spectral perspective suggests a natural strategy: rather than i.i.d. sampling,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

we design a structured sampler that eliminates the specific Fourier components, thereby reducing the variance of the gradient estimate while preserving its expectation.

Building on this formulation, we study two structured sampling designs that share the same Boolean Fourier representation. Both are unbiased and eliminate the prescribed Fourier components. The two designs trade efficiency against exactness.

- *First-order spectral elimination*: a coordinate-balanced design that exactly removes first-order Fourier components at low sample cost, but mildly amplifies even high-order components. Instead of drawing each sampled mask independently, it fixes, for every layer, how many of the sampled masks may activate that layer.
- *Exact spectral elimination*: a design built on *equiprobable stratification* strategy. The framework rests on two ingredients: an equiprobable partition of the mask space into M strata, each sampled with equal probability, and a family of conditional distributions whose mixture recovers the original mask distribution. It cancels a specified set of Fourier modes exactly, at a sampling budget exponential in the linear rank of the target set.

Unlike classical stratified sampling (Neyman, 1934), which partitions a generic sampling space, our strata underlying exact elimination are constructed so that selected Fourier characters vanish exactly; the mixture condition preserves unbiasedness. An illustration of our designs is shown in Figure 1.

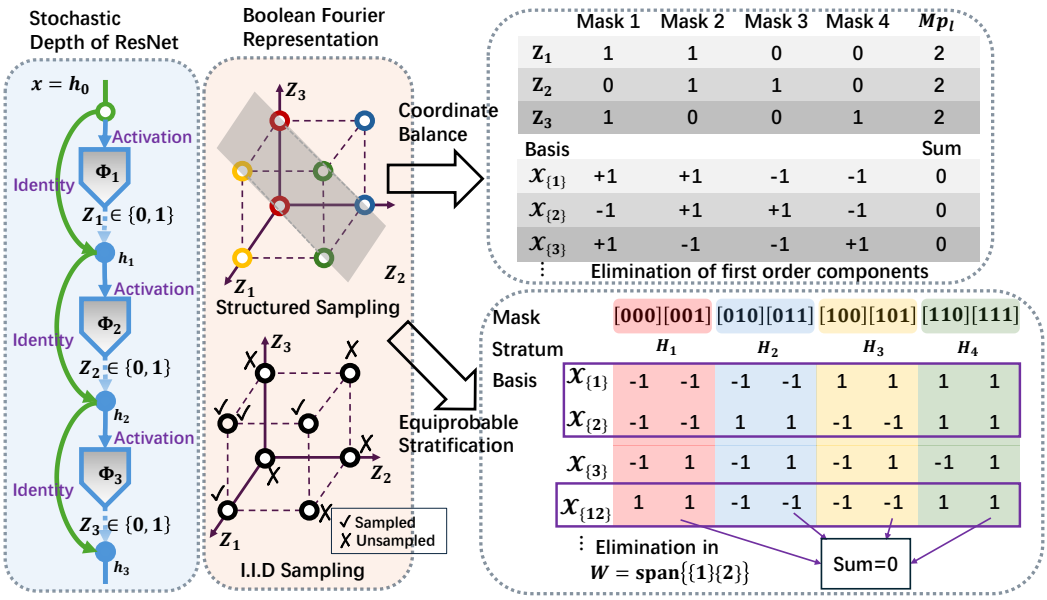


Figure 1: Boolean Fourier representation of stochastic depth in a three-block residual network. **Left:** the residual network with each block gated by a Bernoulli mask. **Middle:** the mask space $\{0,1\}^3$ as a Boolean cube. The upper cube illustrates the two structured sampling designs: the grey plane marks one sample set of the coordinate-balanced construction (Theorem 4.1), while the different colors indicate the distinct strata of the equiprobable stratification strategy (Theorem 4.2); the lower cube shows i.i.d. sampling. **Right:** the resulting masks and their first-order Fourier characters. The upper table shows that coordinate-balanced sampling gives every first-order character zero sample average; the lower table shows exact affine stratification with $W = \mathrm{span}\{\{1\},\{2\}\}$, where characters in $W \setminus \{0\}$ have zero stratified average while $\chi_{\{3\}}$ retains nonzero average.

Our main contributions are summarized as follows.

- **A spectral view of stochastic depth.** We formulate gradient estimation under stochastic depth as a Fourier projection problem on the Boolean cube, revealing that the estimator variance is governed by the Fourier components.

- **Two structured sampling designs.** We propose two designs that eliminate prescribed Fourier components of gradient estimate, differing in exactness and sample cost. We further establish lower bounds on the resulting variance reduction for both designs.
- **Training implications of variance reduction.** Under mild alignment and variance assumptions, we show that variance reduction improves a one-step expected parameter-error comparison from a common starting point and reduces the noise term in an averaged training-loss bound.

We present the main theoretical results in the main text and defer all proofs to the Appendix A.

2 Related Work

Stochastic Depth. Stochastic depth was introduced by (Huang et al., 2016) as a training procedure for ResNets that randomly drops layers to reduce overfitting and improve generalization. (Hayou & Ayed, 2021) showed that it acts as an explicit regularizer and derived principled guidelines for survival rates. Several variants extend it for efficiency or architectural combination, including Swapout (Singh et al., 2016), ShakeDrop (Yamada et al., 2019), and transfer-learning-oriented analyses (Candra et al., 2024). Stochastic depth has also been adopted in large language model pre-training (Elhoushi et al., 2026), and interpreted as Bayesian approximate inference for predictive uncertainty (Müller et al., 2026; Yao et al., 2024). However, these variants primarily target regularization, architectural combination, or Bayesian uncertainty; none explicitly addresses the variance of the training gradient induced by the stochastic depth mask.

Boolean Fourier analysis. The Fourier (Walsh) expansion is a standard tool for analyzing functions on the Boolean cube (O'Donnell, 2014). In learning theory, (Linial et al., 1993) showed that AC^0 functions concentrate almost all of their Fourier mass on low-degree coefficients, enabling efficient learning. Similar spectral concentration underlies the $O(n^{\log\log n})$-time learning algorithm for DNF under the uniform distribution (Mansour, 1992). Recent work (Zhang, 2024) connects the Bohnenblust–Hille inequality with learning low-degree Boolean functions. Beyond learning theory, Boolean function theory has also been used to analyze the expressivity of graph neural networks (Pal, 2026). Our work instead views the gradient as a function of the mask, so that analyzing variance amounts to controlling specific Fourier components.

Structured Dropout. Structured dropout methods drop groups of units rather than individual neurons to address the limited effectiveness of standard dropout on structured data. (Ghiasi et al., 2018) drops contiguous regions in convolutional feature maps; (Larsson et al., 2017) drops entire paths in multi-branch architectures; (Zoph et al., 2018) linearly increases the drop rate during training; and (Zhou et al., 2020) drops entire attention heads in Transformers. Related variants include BatchDropBlock (Dai et al., 2019), which drops a shared region across the batch, and ProbDropBlock (Zhao et al., 2024), an adaptive structured dropout method. Structured dropout has also been interpreted through a Bayesian lens (Nguyen et al., 2021). Beyond vision and language, dropout-based regularization has been applied to recommendation (Volkovs et al., 2017). These methods operate on features, paths, or heads, whereas our work operates on the mask over residual blocks and specifically targets gradient variance in stochastic depth.

3 Formulation

3.1 Stochastic Depth and Residual Network Setup

Consider a ResNet with L residual blocks. Let $p_\ell \in (0,1)$ denote the survival probability of block ℓ, and let $Z = (Z_1, \dots, Z_L) \in \{0,1\}^L$ be a mask vector of independent Bernoulli variables $Z_\ell \sim \mathrm{Bernoulli}(p_\ell)$, distributed as $\mu_p = \bigotimes_{\ell=1}^{L} \mathrm{Bernoulli}(p_\ell)$.

The ResNet is defined recursively by

$$h_\ell = h_{\ell-1} + a_\ell Z_\ell \Phi_\ell(h_{\ell-1}; \theta_\ell), \qquad \ell = 1, \dots, L,$$

where $h_0 = x$ is the input, Φ_ℓ is the ℓ-th residual branch, a_ℓ is the residual scaling factor, and θ_ℓ collects the parameters of block ℓ; block ℓ is skipped when $Z_\ell = 0$. Let $\theta = (\theta_1, \dots, \theta_L)$ denote all parameters.

Given a loss function $\mathcal{L}$, define $f_\theta(Z) = \mathcal{L}(F_\theta(x; Z), y)$, where $F_\theta(x; Z) = h_L$ is the network output and y is the target. At parameter values where all masked losses are differentiable, define $g_\theta(Z) = \nabla_\theta f_\theta(Z)$. Since the mask space is finite, $\nabla_\theta J(\theta) = \mathbb{E}_{Z\sim\mu_p}[g_\theta(Z)]$. For networks with nonsmooth activations, we instead let $g_\theta(Z)$ denote the backpropagation direction under a fixed derivative convention at nondifferentiable points, and define $\bar{g}(\theta) = \mathbb{E}_{Z\sim\mu_p}[g_\theta(Z)]$. The sampling results below apply to this vector-valued function regardless of whether $\bar{g}(\theta)$ is a classical gradient. At differentiability points, $\bar{g}(\theta) = \nabla_\theta J(\theta)$. When θ is fixed, we write $g(Z) = g_\theta(Z)$ and $\mu = \mathbb{E}[g(Z)]$.

3.2 Boolean Fourier Decomposition

For each layer ℓ, define the centered and standardized basis function $\chi_\ell(z_\ell) = \frac{z_\ell - p_\ell}{\sqrt{p_\ell(1-p_\ell)}}$, and for a subset $S \subseteq [L]$ let $\chi_S(z) = \prod_{\ell\in S} \chi_\ell(z_\ell)$. The family $\{\chi_S\}_{S\subseteq[L]}$ is an orthonormal basis of $L^2(\mu_p)$:

$$\mathbb{E}_{Z\sim\mu_p}[\chi_S(Z)\chi_T(Z)] = \mathbf{1}\{S = T\}.$$

Any function $g : \{0,1\}^L \to \mathbb{R}^P$ admits the Fourier expansion $g(z) = \mu + \sum_{S\neq\varnothing} b_S\chi_S(z)$, where $\mu = \mathbb{E}_{Z\sim\mu_p}[g(Z)], b_S = \mathbb{E}_{Z\sim\mu_p}[g(Z)\chi_S(Z)] \in \mathbb{R}^P$. Define the energy of order k and the spectral tail above order t

$$E_k = \sum_{|S|=k} \|b_S\|^2 \quad \text{and} \quad R_t = \sum_{|S|>t} \|b_S\|^2 = \sum_{k>t} E_k$$

respectively, so that $R_0 = \sum_{S\neq\varnothing} \|b_S\|^2$ is the total variance. By Parseval's identity,

$$\mathrm{Var}(g(Z)) = \mathbb{E}\|g(Z) - \mathbb{E}g(Z)\|^2 = R_0,$$

and under i.i.d. sampling with M masks the variance of the sample mean is $\mathrm{Var}(\widehat{g}_{\mathrm{iid}}) = \frac{R_0}{M}$.

Throughout, the variance of a random vector is the trace of its covariance matrix:

$$\mathrm{Var}(Y) = \mathbb{E}\|Y - \mathbb{E}Y\|^2 = \mathrm{tr}(\mathrm{Cov}(Y)).$$

Relative variance-reduction statements are understood under $R_0 > 0$. If $R_0 = 0$, then g is constant on the mask space and all estimators considered below have zero variance.

3.3 Structured Sampling and Equiprobable Stratification

Definition 3.1 (Equiprobable stratification)**.** *An equiprobable stratification consists of a joint distribution of (Z, H) such that $Z \sim \mu_p, H \in \{1, \ldots, K\}, \Pr(H = h) = \frac{1}{K}$. Equivalently, its conditional distributions satisfy*

$$\frac{1}{K}\sum_{h=1}^{K} \Pr(Z = z \mid H = h) = \mu_p(z), \forall z \in \{0,1\}^L.$$

The variable H may be a deterministic or randomized function of Z, or may be defined through an auxiliary latent variable.

Balanced sampling procedure. Assume that $M = nK$ for some integer $n \geq 1$. For each $h \in \{1, \ldots, K\}$, draw

$$Z_{h,1}, \ldots, Z_{h,n} \overset{\text{i.i.d.}}{\sim} \mathcal{L}(Z \mid H = h),$$

independently across all strata, and return

$$\widehat{g} = \frac{1}{M}\sum_{h=1}^{K}\sum_{j=1}^{n} g(Z_{h,j}).$$

Thus, every stratum receives exactly n samples. The mixture constraint implies

$$\mathbb{E}[\widehat{g}] = \frac{1}{K}\sum_{h=1}^{K} \mathbb{E}[g(Z) \mid H = h] = \mu.$$

The fixed allocation is essential. Independently sampling a new uniform stratum index for every observation would instead recover ordinary i.i.d. sampling from μ_p.

Theorem 3.2 (Covariance dominance)**.** *Consider the balanced sampling procedure of Definition 3.1, with $M = nK$ and independent conditional draws. Let $\widehat{g}_{\mathrm{iid}}$ be the average of M independent samples of $g(Z)$ with $Z \sim \mu_p$. Then*

$$\mathrm{Cov}(\widehat{g}) = \frac{1}{M}\mathbb{E}_H[\mathrm{Cov}(g(Z) \mid H)]\,,$$

and

$$\mathrm{Cov}(\widehat{g}_{\mathrm{iid}}) - \mathrm{Cov}(\widehat{g}) = \frac{1}{M}\,\mathrm{Cov}(\mathbb{E}[g(Z) \mid H]) \succeq 0.$$

If $R_0 > 0$, then

$$\frac{\mathrm{Var}(\widehat{g})}{\mathrm{Var}(\widehat{g}_{\mathrm{iid}})} = 1 - \eta_H, \qquad \eta_H = \frac{\mathbb{E}_H\,\|\mathbb{E}[g(Z) - \mu \mid H]\|^2}{R_0} \in [0, 1].$$

4 Targeted Spectral Elimination

This section studies two forms of targeted spectral elimination. The first balances each mask coordinate across the sample set. It eliminates first-order Fourier components exactly, but its strict variance improvement requires a spectral condition. The second uses balanced conditional sampling over affine strata and satisfies the covariance-dominance guarantee of Theorem 3.2.

4.1 First-Order Spectral Elimination

Construction. Let M be the number of masks and suppose $Mp_\ell \in \mathbb{Z}$ for all ℓ. For each layer ℓ, independently and uniformly at random select Mp_ℓ positions out of M and set them to 1; set the remaining positions to 0. The resulting M masks form the sampling scheme.

Unbiasedness. Each mask has marginal distribution $\mathrm{Bernoulli}(p_\ell)$: for any fixed m,

$$\Pr(Z_\ell^{(m)} = 1) = Mp_\ell/M = p_\ell.$$

Independence across layers follows from the construction. Hence $\mathbb{E}[\widehat{g}] = \mu$.

First-order elimination. For every layer ℓ, $\frac{1}{M}\sum_{m=1}^{M} Z_\ell^{(m)} = p_\ell$. Then $\frac{1}{M}\sum_{m=1}^{M} \chi_\ell(Z^{(m)}) = 0$. All first-order Fourier components are exactly eliminated.

Theorem 4.1 (Variance of first-order spectral elimination)**.** *Assume $M \geq 2$ and $Mp_\ell \in \mathbb{Z}$ for every ℓ. For each layer ℓ, independently choose a uniformly random subset of Mp_ℓ positions among the M sample positions, and set precisely those entries to one. Let $\widehat{g}_{\mathrm{FO}}$ be the average of the resulting M masked gradients. Then $\mathbb{E}[\widehat{g}_{\mathrm{FO}}] = \mu$, and*

$$\mathrm{Cov}(\widehat{g}_{\mathrm{FO}}) = \frac{1}{M}\sum_{S \neq \varnothing}\left[1 + \frac{(-1)^{|S|}}{(M-1)^{|S|-1}}\right] b_S b_S^\top.$$

When does first-order elimination help? Writing the variance difference from Theorem 4.1 in terms of spectral orders,

$$\mathrm{Var}(\widehat{g}_{\mathrm{iid}}) - \mathrm{Var}(\widehat{g}_{\mathrm{FO}}) = \frac{1}{M}\Big[\underbrace{E_1}_{\text{first-order gain}} \underbrace{- \sum_{\substack{k \geq 2 \\ k \text{ even}}} \frac{E_k}{(M-1)^{k-1}}}_{\text{even-order cost}} + \underbrace{\sum_{\substack{k \geq 3 \\ k \text{ odd}}} \frac{E_k}{(M-1)^{k-1}}}_{\text{odd-order gain}}\Big],$$

which reveals a trade-off: first-order modes are eliminated exactly, even orders are mildly *amplified*, and odd orders $k \geq 3$ are mildly attenuated. Consequently, first-order elimination is highly effective when the first-order energy dominates the spectrum, as we verify empirically in Section 6.

A rigorous sufficient condition is $E_1 > \frac{R_1}{M-1}$, under which the relative variance reduction satisfies $\eta_{\mathrm{FO}} \geq \frac{E_1}{R_0} - \frac{R_1}{(M-1)R_0}$. The derivation is given in Appendix B.

4.2 Exact Spectral Elimination

We first present the construction for the uniform case $p_\ell = 1/2$, where the mask Z is itself uniform on the Boolean cube and linear-algebraic constructions apply directly; the general case is discussed at the end of the subsection.

Target and design. Assume $p_\ell = 1/2$ for every ℓ. Identify each subset $S \subseteq [L]$ with its indicator vector $s \in \mathbb{F}_2^L$. Let $\mathcal{T} \subseteq 2^{[L]} \setminus \{\varnothing\}$ be the collection of target frequencies, and choose a full-row-rank matrix $A \in \mathbb{F}_2^{r \times L}$ whose rows form a basis of $W = \mathrm{span}_{\mathbb{F}_2}(\mathcal{T})$. Thus $r = \dim_{\mathbb{F}_2}(W)$.

All finite-field operations below are performed on the bit vector $Z \in \{0,1\}^L = \mathbb{F}_2^L$. Define $H = AZ \in \mathbb{F}_2^r$. The $K = 2^r$ strata are the affine subspaces $C_h = \{z \in \mathbb{F}_2^L : Az = h\}, h \in \mathbb{F}_2^r$. Since A has full row rank, every stratum has size 2^{L-r}, and H is uniform on $\mathbb{F}_2^r$.

Sampling. Set $M = K = 2^r$. For each $h \in \mathbb{F}_2^r$, choose any particular solution z_h of $Az_h = h$, independently draw $v_h \sim \mathrm{Uniform}(\ker A)$, and set $Z_h = z_h + v_h$, where addition is over $\mathbb{F}_2$. Then Z_h is uniform on C_h. The estimator is $\widehat{g}_{\mathrm{exact}} = \frac{1}{M} \sum_{h \in \mathbb{F}_2^r} g(Z_h)$.

Unbiasedness. Unbiasedness follows from the equal-weight mixture: $\frac{1}{M} \sum_{h \in \mathbb{F}_2^r} \mathcal{L}(Z_h) = \mathrm{Uniform}(\mathbb{F}_2^L)$. Consequently, $\mathbb{E}[\widehat{g}_{\mathrm{exact}}] = \frac{1}{M} \sum_h \mathbb{E}[g(Z) \mid AZ = h] = \mu$.

Theorem 4.2 (Exact spectral elimination)**.** *Consider the construction above with $p_\ell = 1/2$, $W = \mathrm{row}_{\mathbb{F}_2}(A)$, and $M = 2^r$. For every nonzero frequency $S \in W$,*

$$\frac{1}{M} \sum_{h \in \mathbb{F}_2^r} \chi_S(Z_h) = 0, \quad \mathrm{Cov}(\widehat{g}_{\mathrm{iid}}) - \mathrm{Cov}(\widehat{g}_{\mathrm{exact}}) = \frac{1}{M} \sum_{S \in W \setminus \{0\}} b_S b_S^\top \succeq 0.$$

If $R_0 > 0$, the relative variance reduction is $\eta_{\mathrm{exact}} = \sum_{S \in W \setminus \{0\}} \|b_S\|^2 / R_0$.

Sampling budget. The construction eliminates every nonzero frequency in $W = \mathrm{span}_{\mathbb{F}_2}(\mathcal{T})$, not merely the frequencies explicitly listed in $\mathcal{T}$. Its sample budget is $M = 2^{\mathrm{rank}_{\mathbb{F}_2}(\mathcal{T})}$.

General survival probabilities. For general p_ℓ, the mask is no longer uniform, and the linear construction must be lifted through a latent uniform variable $X_\ell \in \{-1, 1\}$, with $\Pr(Z_\ell = 1 \mid X_\ell = x) = p_\ell + x m_\ell$, $m_\ell = \min(p_\ell, 1 - p_\ell)$. Exact cancellation on the mask cube is replaced by weighted spectral attenuation, while unbiasedness and covariance dominance are preserved:

$$\mathrm{Cov}(\widehat{g}_{\mathrm{iid}}) - \mathrm{Cov}(\widehat{g}_{\mathrm{exact}}) = \frac{1}{M} \sum_{S \in W \setminus \{0\}} \rho_S^2 \, b_S b_S^\top, \qquad \rho_S = \prod_{\ell \in S} \rho_\ell, \quad \rho_\ell = \sqrt{\tfrac{\min(p_\ell, 1-p_\ell)}{\max(p_\ell, 1-p_\ell)}}.$$

The factor ρ_S^2 decays with $|S|$, so higher-order modes benefit less from elimination; the construction remains unbiased and is never worse than i.i.d. sampling. The key technical step is the alignment $\mathbb{E}[\chi_S(Z) \mid X] = \rho_S \chi_S(X)$, which bridges the non-uniform mask distribution and the uniform cube. Full details are given in Appendix C.

5 Training Benefits of Variance Reduction

In this section, we write $g(\theta) = \mathbb{E}_Z[g_\theta(Z)]$. At differentiability points this is $\nabla J(\theta)$; otherwise it denotes the mean backpropagation direction under the convention fixed in Section 3. Let $\theta^\star$ be a reference point such that $\mathcal{L}(\theta) = J(\theta) - J(\theta^\star) \geq 0$ on the region considered. The following analysis uses the stated conditions on $g(\theta)$ directly and does not require differentiability of J.

5.1 Single-Step Reduction

Consider the stochastic update $\theta_{t+1} = \theta_t - \alpha_t \hat{g}_t$. Let $\mathcal{F}_t$ denote the filtration generated by the parameters and all information available before sampling at step t, and let $\hat{g}_t$ be conditionally unbiased, $\mathbb{E}[\hat{g}_t \mid \mathcal{F}_t] = g_t$. Write $\xi_t = \hat{g}_t - g_t$, so that $\mathbb{E}_t \xi_t = 0$, and let $v_t = \mathbb{E}_t \|\xi_t\|^2 < \infty$.

Suppose the sampling rule at step t is fixed conditionally on $\mathcal{F}_t$, and all estimation draws use fresh randomness. For balanced conditional stratification, Theorem 3.2 applies conditionally and gives

$v_t^{\mathrm{struct}} = v_t^{\mathrm{iid}} - \Delta_t, \Delta_t \geq 0$. For first-order coordinate-balanced sampling, the same notation can be used, but $\Delta_t \geq 0$ must be established separately, for example through the spectral criterion in Appendix B. A reduction in total variance for that design does not by itself imply covariance dominance in every direction.

Theorem 5.1 (Exact single-step reduction)**.** *Fix any reference point $\theta^\star$ and let $e_t = \theta_t - \theta^\star$. Then, conditionally on $\mathcal{F}_t$,*

$$\mathbb{E}_t\|\theta_{t+1} - \theta^\star\|^2 = \|e_t\|^2 - 2\alpha_t\langle g_t, e_t\rangle + \alpha_t^2\|g_t\|^2 + \alpha_t^2 v_t.$$

Consequently, if both estimators start from the same θ_t,

$$\mathbb{E}_t\|\theta_{t+1}^{\mathrm{iid}} - \theta^\star\|^2 - \mathbb{E}_t\|\theta_{t+1}^{\mathrm{struct}} - \theta^\star\|^2 = \alpha_t^2\Delta_t \geq 0.$$

5.2 Training Loss under a Pointwise Alignment Condition

Fix a reference point $\theta^\star$ and write $\mathcal{L}(\theta) = J(\theta) - J(\theta^\star) \geq 0$. We impose two conditions.

Assumption 5.2 (Directional alignment)**.** *There exist $c > 0$ such that, on the region visited by the iterates, $\langle g(\theta), \theta - \theta^\star\rangle \geq c\,\mathcal{L}(\theta)$.*

Assumption 5.3 (Bounded direction)**.** *There exists $B > 0$ such that, on the same region of Assumption 5.2, $\|g(\theta)\|^2 \leq B\,\mathcal{L}(\theta)$.*

Assume that, at every parameter value visited by the structured iterates, the corresponding i.i.d. mask estimator satisfies $v_t^{\mathrm{iid}} \leq V$ and the chosen sampling rule satisfies $v_t^{\mathrm{struct}} \leq (1-\eta_0)v_t^{\mathrm{iid}}, \eta_0 \in [0,1]$.

Theorem 5.4 (Averaged training loss bound)**.** *Suppose Assumptions 5.2 and 5.3 hold, the iterates remain in the region where they hold, and $v_t^{\mathrm{iid}} \leq V$ with $\eta_H \geq \eta_0$. Let $D_0 = \|\theta_0 - \theta^\star\|$ and choose a constant step size $0 < \alpha \leq c/B$. Then for every $T \geq 1$,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\,\mathcal{L}(\theta_t) \leq \frac{D_0^2}{\alpha c T} + \frac{\alpha}{cT}\sum_{t=0}^{T-1}\mathbb{E}\,v_t^{\mathrm{struct}}.$$

If $v_t^{\mathrm{struct}} \leq qV$ with $q = 1 - \eta_0$, then

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\,\mathcal{L}(\theta_t) \leq \frac{D_0^2}{\alpha c T} + \frac{\alpha q V}{c}.$$

Theorem 5.4 shows that the variance-reduction factor q enters the loss bound directly: in the noise-dominated regime, the random term scales by $q = 1 - \eta_0$. Assumptions 5.2 and 5.3 provide an alternative route to a training-loss bound without requiring a Lipschitz gradient or a Polyak–Lojasiewicz condition. We do not claim a general logical implication between these assumptions and those standard conditions. Appendix D gives sufficient conditions for particular squared-loss models.

Mask noise versus data noise. The preceding variance comparisons concern mask randomness at a fixed data batch. If a random batch B is sampled and the same batch is used for all M masks, define $\mu_B = \mathbb{E}[\widehat{g} \mid B]$. Then $\mathrm{Var}(\widehat{g}) = \mathrm{Var}_B(\mu_B) + \mathbb{E}_B[\mathrm{Var}(\widehat{g} \mid B)]$. A factor-$q$ reduction of conditional mask variance therefore gives $v_{\mathrm{total}}^{\mathrm{struct}} \leq v_{\mathrm{data}} + q\,v_{\mathrm{mask}}^{\mathrm{iid}}$, not in general $q\,v_{\mathrm{total}}^{\mathrm{iid}}$. For minibatch training, the bounds above must be applied with the total variance bound appropriate to this decomposition.

6 Experiments

We evaluate the structured sampling designs of Section 4.1 on image classification tasks to verify if the structured sampling can exactly reduce the variance and benefit the training. We focus on the coordinate-balanced design because the equiprobable stratification design requires a prescribed target mode set.

Data and architectures. For CIFAR-10 we use ResNet-18, ResNet-34, and ResNet-50; for CIFAR-100 we use ResNet-18. DropPath is applied to the residual branch of every block, giving 8, 16, and 16 DropPath blocks for the three architectures, respectively.

Training. Models are trained for 60 epochs with batch size 128 and standard SGD (learning rate 0.05, momentum 0.9, weight decay 5×10^{-4}, cosine annealing), with survival probability $p = 0.5$ and $M_{\text{train}} = 4$ masks per step.

Variance diagnostic and metric. At each checkpoint we estimate the gradient variance of each sampler on a fixed minibatch, using $N = 8$ masks and $R = 30$ repetitions, averaged over multiple seeds and reported as mean $\pm$ std. Since the absolute variance scales with the number of parameters, we report the *variance reduction factor*: i.i.d. variance divided by structured variance at the same checkpoint.

6.1 First-Order Elimination on CIFAR-10

Variance reduction. Table 1 reports the variance reduction factor of per-layer balanced sampling relative to i.i.d. sampling. The factor exceeds one at every architecture and checkpoint, with small across-seed standard deviation, confirming that the effect is robust and reproducible. It peaks early in training at around 3–4$\times$ and decreases to roughly 1.4–$1.9\times$ by the end. This is consistent with the spectral view: early in training the gradient noise is dominated by mask randomness, so eliminating first-order Fourier components yields the largest relative gain; as the model converges, the gradient becomes more stable and the relative benefit shrinks. The factor does not decrease monotonically with depth: ResNet-50, which has the same number of DropPath blocks as ResNet-34, exhibits a larger reduction factor throughout most of training, suggesting that the benefit depends on the residual block type rather than on depth alone.

Table 1: Variance reduction factor (i.i.d. variance divided by structured variance) of per-layer balanced sampling on CIFAR-10, reported as mean $\pm$ std over seeds. Values greater than one indicate variance reduction.

Epoch	ResNet-18	ResNet-34	ResNet-50
init	3.42 ± 0.03	2.70 ± 0.29	2.57 ± 0.35
ep10	4.01 ± 0.62	3.21 ± 0.74	3.17 ± 0.35
ep20	3.38 ± 0.08	2.43 ± 0.09	3.24 ± 0.03
ep30	2.38 ± 0.29	2.44 ± 0.17	3.01 ± 0.10
ep40	2.27 ± 0.20	2.05 ± 0.24	2.75 ± 0.30
ep50	1.68 ± 0.05	1.76 ± 0.09	2.54 ± 0.63
ep60	1.61 ± 0.18	1.38 ± 0.05	1.85 ± 0.10

Accuracy. Table 2 reports final and best accuracy, and the epoch at which each model reaches 85% accuracy. Balanced sampling matches or slightly improves final accuracy over i.i.d. sampling within one standard deviation, and consistently yields a smaller across-seed standard deviation, consistent with more stable training under lower gradient variance.

Table 2: Final and best accuracy on CIFAR-10, and the epoch at which each model reaches 85% accuracy, reported as mean $\pm$ std over seeds.

Architecture	Method	Final acc	Best acc	Epoch to 85%
ResNet-18	baseline	91.97 ± 0.13	92.02 ± 0.18	29.3 ± 0.5
ResNet-18	balanced	92.06 ± 0.03	92.08 ± 0.02	29.0 ± 0.8
ResNet-34	baseline	92.92 ± 0.24	92.96 ± 0.19	27.3 ± 0.5
ResNet-34	balanced	93.03 ± 0.12	93.07 ± 0.16	26.3 ± 0.5
ResNet-50	baseline	89.97 ± 1.18	90.03 ± 1.09	42.7 ± 3.7
ResNet-50	balanced	90.23 ± 0.54	90.30 ± 0.56	42.3 ± 1.2

First-order energy. Table 3 reports the first-order energy fraction E_1/R_0 on baseline checkpoints. It lies in 0.5–0.8 across architectures and checkpoints, indicating that the loss has a substantial but not dominant first-order dependence on the mask. This supports first-order elimination: the first-order component is large enough to make its removal worthwhile, but not so dominant that higher-order effects can be ignored.

Table 3: First-order energy fraction E_1/R_0 on baseline CIFAR-10 checkpoints, reported as mean $\pm$ std over seeds.

Architecture	init	ep20	ep40	ep60
ResNet-18	0.63 ± 0.09	0.82 ± 0.13	0.70 ± 0.08	0.74 ± 0.06
ResNet-34	0.72 ± 0.03	0.67 ± 0.27	0.48 ± 0.05	0.57 ± 0.05
ResNet-50	0.61 ± 0.08	0.69 ± 0.25	0.52 ± 0.02	0.82 ± 0.03

6.2 Extension to CIFAR-100

We repeat the first-order elimination experiment on CIFAR-100, using ResNet-18.

Table 4: Variance reduction factor of per-layer balanced sampling on CIFAR-100 with ResNet-18, mean $\pm$ std over seeds. Values greater than one indicate variance reduction.

Epoch	init	ep10	ep20	ep30	ep40	ep50	ep60
i.i.d./ Balanced	3.07 ± 0.01	5.39 ± 0.45	3.99 ± 0.32	2.97 ± 0.09	2.65 ± 0.06	2.22 ± 0.17	2.00 ± 0.28

As on CIFAR-10, the reduction factor exceeds one at every checkpoint with small across-seed standard deviation, peaks early at about $5.4\times$, and decreases to roughly $2.0\times$ by the end. The higher early-training peak is consistent with the harder task producing noisier gradients and thus larger headroom for mask-variance reduction.

Table 5 reports final and best accuracy: balanced sampling matches i.i.d. within one standard deviation, so the variance reduction does not come at an accuracy cost.

Table 5: Final and best accuracy on CIFAR-100 with ResNet-18, mean $\pm$ std over seeds.

Method	Final acc	Best acc
baseline	67.75 ± 0.07	67.77 ± 0.05
balanced	67.90 ± 0.05	67.94 ± 0.00

The first-order energy fraction E_1/R_0 is 0.78–0.92 (Table 6), higher than on CIFAR-10, indicating more headroom for first-order elimination on the harder benchmark.

Table 6: First-order energy fraction E_1/R_0 on baseline CIFAR-100 checkpoints with ResNet-18, mean $\pm$ std over seeds.

init	ep20	ep40	ep60
0.92 ± 0.09	0.83 ± 0.08	0.84 ± 0.01	0.78 ± 0.03

7 Conclusion

We studied gradient estimation under stochastic depth through a Fourier lens on the Boolean cube, where variance reduction becomes the control of specific spectral components. This view yields two structured samplers: a coordinate-balanced design that removes first-order components at low sample cost, and an equiprobable stratification framework that cancels a user-specified set of modes exactly and guarantees covariance dominance. We derive explicit variance-reduction bounds and their implications for stochastic gradient descent. Our analysis concerns mask randomness at a fixed batch, and the stratification design assumes that the target mode set is known. Relaxing this assumption—selecting the target set without spectral prior knowledge—is one natural direction for future work. Another is to extend the framework to other structured-dropout settings, provided that the corresponding masks admit a tractable Fourier representation.

AI USE STATEMENT

We used generative AI tools to aid and polish the writing of this paper, and to assist in checking and revising our theoretical claims. All AI-assisted work has been reviewed by the authors: theoretical results and proofs were verified by hand, and all AI-generated or AI-edited text was checked for accuracy and potential plagiarism through a manual literature survey. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Calvin Linardy Candra, Anderies, Donghai Guan, Tjeng Wawan Cenggoro, and Bens Pardamean. Analysing potential of resnet for transfer learning with stochastic depth. In Yuan Tian, Tinghuai Ma, and Muhammad Khurram Khan (eds.), *Big Data and Security*, pp. 127–137, Singapore, 2024. Springer Nature Singapore. ISBN 978-981-97-4387-2.

Zuozhuo Dai, Mingqiang Chen, Xiaodong Gu, Siyu Zhu, and Ping Tan. Batch dropblock network for person re-identification and beyond. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3690–3700, 2019. doi: 10.1109/ICCV.2019.00379.

Alek Dimitriev and Mingyuan Zhou. Carms: Categorical-antithetic-reinforce multi-sample gradient estimator. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 13217–13229. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/6e16656a6ee1de7232164767ccfa7920-Paper.pdf.

Aleksandar Dimitriev and Mingyuan Zhou. Arms: Antithetic-reinforce-multi-sample gradient for binary variables. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2717–2727. PMLR, 18–24 Jul 2021b. URL https://proceedings.mlr.press/v139/dimitriev21a.html.

Mostafa Elhoushi, Alexander D. Pretko, Nolan Simran Dey, Bin Claire Zhang, Gavia Gray, Gurpreet Gosal, Abdulrahman Mahmoud, Shane Bergsma, and Joel Hestness. Don't drop dropout: Optimizing layer sparsity for efficient LLM training and inference. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=AEHn9ya4Re.

Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/7edcfb2d8f6a659ef4cd1e6c9b6d7079-Paper.pdf.

Soufiane Hayou and Fadhel Ayed. Regularization in resnet with stochastic depth. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 15464–15474. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/82ba9d6eee3f026be339bb287651c3d8-Paper.pdf.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 646–661, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, pp. 315–323, Red Hook, NY, USA, 2013. Curran Associates Inc.

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Fractalnet: Ultra-deep neural networks without residuals, 2017. URL https://arxiv.org/abs/1605.07648.

Yangkun Li, Weizhi Ma, Chong Chen, Min Zhang, Yiqun Liu, Shaoping Ma, and Yuekui Yang. A Survey on Dropout Methods and Experimental Verification in Recommendation . *IEEE Transactions on Knowledge & Data Engineering*, 35(07):6595–6615, July 2023. ISSN 1558-2191. doi: 10.1109/TKDE.2022.3187013. URL https://doi.ieeecomputersociety.org/10.1109/TKDE.2022.3187013.

Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, fourier transform, and learnability. *J. ACM*, 40(3):607–620, July 1993. ISSN 0004-5411. doi: 10.1145/174130.174138. URL https://doi.org/10.1145/174130.174138.

Yishay Mansour. An o(nlog log n) learning algorithm for dnf under the uniform distribution. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pp. 53–61, New York, NY, USA, 1992. Association for Computing Machinery. ISBN 089791497X. doi: 10.1145/130385.130391. URL https://doi.org/10.1145/130385.130391.

Adam T. Müller, Tobias Rögelein, and Nicolaj C. Stache. Monte carlo stochastic depth for uncertainty estimation in deep learning, 2026. URL https://arxiv.org/abs/2604.12719.

Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4): 558–625, 1934. ISSN 09528385. URL http://www.jstor.org/stable/2342192.

Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2613–2621. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/nguyen17b.html.

Son Nguyen, Duong Nguyen, Khai Nguyen, Khoat Than, Hung Bui, and Nhat Ho. Structured Dropout Variational Inference for Bayesian Neural Networks. In *Neural Information Processing Systems*, 2021. URL https://mlanthology.org/neurips/2021/nguyen2021neurips-structured/.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

Manjish Pal. A boolean function-theoretic framework for expressivity in gnns with applications to fair graph mining, 2026. URL https://arxiv.org/abs/2601.12751.

Saurabh Singh, Derek Hoiem, and David Forsyth. Swapout: Learning an ensemble of deep architectures. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/c51ce410c124a10e0db5e4b97fc2af39-Paper.pdf.

Maksims Volkovs, Guangwei Yu, and Tomi Poutanen. Dropoutnet: Addressing cold start in recommender systems. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/dbd22ba3bd0df8f385bdac3e9f8be207-Paper.pdf.

Chong Wang, Xi Chen, Alexander Smola, and Eric Xing. Variance reduction for stochastic gradient optimization. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/9766527f2b5d3e95d4a733fcfb77bd7e-Paper.pdf.

Yoshihiro Yamada, Masakazu Iwamura, Takuya Akiba, and Koichi Kise. Shakedrop regularization for deep residual learning. *IEEE Access*, 7:186126–186136, 2019. doi: 10.1109/ACCESS.2019.2960566.

Linghong Yao, Denis Hadjivelichkov, Andromachi Maria Delfaki, Yuanchang Liu, Brooks Paige, and Dimitrios Kanoulas. Balancing calibration and performance: Stochastic depth in segmentation bnns. In *35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024*. BMVA, 2024. URL https://papers.bmvc2024.org/0546.pdf.

Haonan Zhang. Three lectures on fourier analysis and learning theory, 2024. URL https://arxiv.org/abs/2409.10886.

Yiren Zhao, Oluwatomisin Dada, Robert Mullins, and Xitong Gao. Revisiting structured dropout. In Berrin Yanıkoğlu and Wray Buntine (eds.), *Proceedings of the 15th Asian Conference on Machine Learning*, volume 222 of *Proceedings of Machine Learning Research*, pp. 1699–1714. PMLR, 11–14 Nov 2024. URL https://proceedings.mlr.press/v222/zhao24a.html.

Wangchunshu Zhou, Tao Ge, Furu Wei, Ming Zhou, and Ke Xu. Scheduled DropHead: A regularization method for transformer models. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1971–1980, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.178. URL https://aclanthology.org/2020.findings-emnlp.178/.

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8697–8710, 2018. doi: 10.1109/CVPR.2018.00907.

A Proof

A.1 Proof of Theorem 3.2

Proof. Write

$$\Sigma_h = \mathrm{Cov}(g(Z) \mid H = h), \qquad m_h = \mathbb{E}[g(Z) \mid H = h].$$

The conditional samples are independent, and $M = nK$. Therefore,

$$\begin{aligned}\mathrm{Cov}(\widehat{g}) &= \frac{1}{M^2} \sum_{h=1}^{K} \sum_{j=1}^{n} \Sigma_h \\ &= \frac{nK}{M^2} \left(\frac{1}{K} \sum_{h=1}^{K} \Sigma_h \right) \\ &= \frac{1}{M} \mathbb{E}_H[\mathrm{Cov}(g(Z) \mid H)].\end{aligned}$$

By the law of total covariance,

$$\mathrm{Cov}(g(Z)) = \mathbb{E}_H[\mathrm{Cov}(g(Z) \mid H)] + \mathrm{Cov}(\mathbb{E}[g(Z) \mid H]).$$

Since

$$\mathrm{Cov}(\widehat{g}_{\mathrm{iid}}) = \frac{1}{M} \mathrm{Cov}(g(Z)),$$

subtracting proves the covariance identity. Taking traces gives

$$\mathrm{Var}(\widehat{g}_{\mathrm{iid}}) - \mathrm{Var}(\widehat{g}) = \frac{1}{M} \mathbb{E}_H \left\| \mathbb{E}[g(Z) - \mu \mid H] \right\|^2.$$

Dividing by R_0/M proves the relative variance formula. □

A.2 Proof of Theorem 4.1

We first record a second-moment computation for the first-order characters.

Lemma A.1 (Second-moment of first-order characters)**.** *For $m \neq n$,*

$$\mathbb{E}[\chi_\ell(Z^{(m)}) \chi_\ell(Z^{(n)})] = -\frac{1}{M-1};$$

for $m = n$,

$$\mathbb{E}[\chi_\ell(Z^{(m)})^2] = 1.$$

Proof. Let $K_\ell = Mp_\ell$. The vector $(Z_\ell^{(1)}, \dots, Z_\ell^{(M)})$ is uniform over all binary vectors with exactly K_ℓ ones. For $m \neq n$,

$$\mathbb{E}[Z_\ell^{(m)} Z_\ell^{(n)}] = \frac{K_\ell(K_\ell - 1)}{M(M-1)}.$$

Substituting $K_\ell = Mp_\ell$ and simplifying gives

$$\mathbb{E}[\chi_\ell(Z^{(m)})\chi_\ell(Z^{(n)})] = \frac{\mathbb{E}[Z_\ell^{(m)} Z_\ell^{(n)}] - p_\ell^2}{p_\ell(1-p_\ell)} = -\frac{1}{M-1}.$$

The diagonal case is immediate from standardization. □

Proof of Theorem 4.1. Write

$$\widehat{g}_{\mathrm{FO}} - \mu = \frac{1}{M}\sum_{m=1}^{M}\sum_{S\neq\varnothing} b_S \chi_S(Z^{(m)}).$$

By orthogonality across distinct subsets and independence across layers, cross terms between different S vanish. For a fixed S with $|S| = k$,

$$\mathbb{E}\left[\left(\frac{1}{M}\sum_{m=1}^{M}\chi_S(Z^{(m)})\right)^2\right] = \frac{1}{M^2}\sum_{m,n}\mathbb{E}[\chi_S(Z^{(m)})\chi_S(Z^{(n)})].$$

By independence across layers and Lemma A.1,

$$\mathbb{E}[\chi_S(Z^{(m)})\chi_S(Z^{(n)})] = \left(-\frac{1}{M-1}\right)^k \quad (m \neq n),$$

and the diagonal terms equal 1. There are M diagonal terms and $M(M-1)$ off-diagonal terms, so

$$\mathbb{E}\left[\left(\frac{1}{M}\sum_{m=1}^{M}\chi_S(Z^{(m)})\right)^2\right] = \frac{1}{M}\left[1 + (M-1)\left(-\frac{1}{M-1}\right)^k\right] = \frac{1}{M}\left[1 + \frac{(-1)^k}{(M-1)^{k-1}}\right].$$

The same calculation applied to the covariance matrix gives

$$\mathrm{Cov}(\widehat{g}_{\mathrm{FO}}) = \sum_{S\neq\varnothing} b_S b_S^\top \mathbb{E}\left[\left(\frac{1}{M}\sum_{m=1}^{M}\chi_S(Z^{(m)})\right)^2\right].$$

Substituting the computed second moments proves the covariance formula. Taking traces and grouping by $|S|$ proves the variance formula. □

A.3 Proof of Theorem 4.2

Proof. We identify each subset $S \subseteq [L]$ with its indicator vector $s \in \mathbb{F}_2^L$, and write $S \in W$ when $s \in W$. For $p_\ell = 1/2$,

$$\chi_S(z) = \prod_{\ell\in S}(2z_\ell - 1) = (-1)^{|S|}(-1)^{s^\top z},$$

where $s^\top z$ is computed over $\mathbb{F}_2$.

First, we compute the conditional mean of each character. If $S \in W = \mathrm{row}(A)$, there is a unique $t_S \in \mathbb{F}_2^r$ such that $A^\top t_S = s$. For every $z \in C_h$,

$$\chi_S(z) = (-1)^{|S|}(-1)^{t_S^\top h} =: \psi_S(h).$$

Thus χ_S is constant within each stratum.

If $S \notin W$, the identity

$$W = (\ker A)^\perp$$

implies that there exists $v \in \ker A$ with $s^\top v = 1$. Translation by v preserves C_h and changes the sign of χ_S, so

$$\mathbb{E}[\chi_S(Z) \mid H = h] = 0.$$

Consequently,

$$\mathbb{E}[\chi_S(Z) \mid H = h] = \begin{cases} \psi_S(h), & S \in W, \\ 0, & S \notin W. \end{cases}$$

It is important to distinguish this conditional-mean identity from exact cancellation in the stratified sample average. Define

$$U_S = \frac{1}{M} \sum_{h \in \mathbb{F}_2^r} \chi_S(Z_h).$$

For every nonzero $S \in W$, we have $t_S \neq 0$ and hence

$$U_S = \frac{1}{M} \sum_{h} \psi_S(h) = 0 \qquad \text{almost surely.}$$

Thus the nonzero frequencies in W are canceled exactly by averaging over all strata.

In contrast, for $S \notin W$, each $\chi_S(Z_h)$ has mean zero and variance one. Independence across strata gives

$$\mathbb{E}[U_S] = 0, \qquad \mathrm{Var}(U_S) = \frac{1}{M}.$$

Thus a zero conditional mean outside W does not imply exact cancellation of the sampled character.

To compute the covariance, the conditional-mean formula gives

$$\mathbb{E}[g(Z) \mid H] = \mu + \sum_{S \in W \setminus \{0\}} b_S \psi_S(H).$$

The functions ψ_S, for nonzero $S \in W$, have mean zero and are mutually orthonormal under uniform H. Therefore,

$$\mathrm{Cov}(\mathbb{E}[g(Z) \mid H]) = \sum_{S \in W \setminus \{0\}} b_S b_S^\top.$$

By Theorem 3.2,

$$\mathrm{Cov}(\widehat{g}_{\mathrm{iid}}) - \mathrm{Cov}(\widehat{g}_{\mathrm{exact}}) = \frac{1}{M} \sum_{S \in W \setminus \{0\}} b_S b_S^\top.$$

Since Fourier orthogonality gives

$$\mathrm{Cov}(\widehat{g}_{\mathrm{iid}}) = \frac{1}{M} \sum_{S \neq \varnothing} b_S b_S^\top,$$

we obtain

$$\mathrm{Cov}(\widehat{g}_{\mathrm{exact}}) = \frac{1}{M} \sum_{\substack{S \neq \varnothing \\ S \notin W}} b_S b_S^\top.$$

Taking traces proves the variance formula. When $R_0 > 0$, dividing the eliminated variance by R_0/M proves the relative variance-reduction formula. □

A.4 Proof of Theorem 5.1

Proof. Conditionally on $\mathcal{F}_t$, both e_t and g_t are fixed and $\mathbb{E}_t[\xi_t] = 0$. Therefore,

$$\begin{aligned} \mathbb{E}_t\|e_t - \alpha_t g_t - \alpha_t \xi_t\|^2 &= \|e_t - \alpha_t g_t\|^2 + \alpha_t^2 \mathbb{E}_t\|\xi_t\|^2 \\ &= \|e_t\|^2 - 2\alpha_t \langle g_t, e_t \rangle + \alpha_t^2 \|g_t\|^2 + \alpha_t^2 v_t. \end{aligned}$$

For two estimators evaluated at the same θ_t and having the same conditional mean, subtracting the two identities gives

$$\mathbb{E}_t\|\theta_{t+1}^{\mathrm{iid}} - \theta^\star\|^2 - \mathbb{E}_t\|\theta_{t+1}^{\mathrm{struct}} - \theta^\star\|^2 = \alpha_t^2 \Delta_t.$$

□

A.5 Proof of Theorem 5.4

Proof. By Theorem 5.1 and Assumption 5.2,

$$\mathbb{E}_t\|e_{t+1}\|^2 \le \|e_t\|^2 - (2\alpha c - \alpha^2 B)\mathcal{L}(\theta_t) + \alpha^2 v_t^{\mathrm{struct}},$$

where we used $\|g_t\|^2 \le B\mathcal{L}(\theta_t)$ and $\langle g_t, e_t\rangle \ge c\mathcal{L}(\theta_t)$. For $\alpha \le c/B$, $2\alpha c - \alpha^2 B \ge \alpha c$, hence

$$\mathbb{E}_t\|e_{t+1}\|^2 \le \|e_t\|^2 - \alpha c\,\mathcal{L}(\theta_t) + \alpha^2 v_t^{\mathrm{struct}}.$$

Summing over $t = 0, \dots, T-1$, dropping the nonnegative terminal term, and dividing by αcT yields the first bound; the second follows from $v_t^{\mathrm{struct}} \le qV$. □

B Advantage Criterion for First-Order Elimination

We formalize the spectral condition under which first-order elimination outperforms i.i.d. sampling, and derive the corresponding relative variance reduction.

Proposition B.1 (Advantage criterion)**.** *Assume $M \ge 2$ and $Mp_\ell \in \mathbb{Z}$ for every ℓ. Let $\widehat{g}_{\mathrm{FO}}$ be the first-order elimination estimator and $\widehat{g}_{\mathrm{iid}}$ the i.i.d. estimator with the same number M of masks. Then*

$$\mathrm{Var}(\widehat{g}_{\mathrm{iid}}) - \mathrm{Var}(\widehat{g}_{\mathrm{FO}}) = \frac{1}{M}\left[E_1 - \sum_{\substack{k\ge 2\\ k\ \mathit{even}}} \frac{E_k}{(M-1)^{k-1}} + \sum_{\substack{k\ge 3\\ k\ \mathit{odd}}} \frac{E_k}{(M-1)^{k-1}}\right], \tag{1}$$

and consequently

$$\mathrm{Var}(\widehat{g}_{\mathrm{iid}}) - \mathrm{Var}(\widehat{g}_{\mathrm{FO}}) \ge \frac{1}{M}\left[E_1 - \frac{R_1}{M-1}\right], \qquad R_1 = \sum_{k\ge 2} E_k. \tag{2}$$

In particular, if

$$E_1 > \frac{R_1}{M-1}, \tag{3}$$

then first-order elimination strictly outperforms i.i.d. sampling, and the relative variance reduction satisfies

$$\eta \;\ge\; \frac{E_1}{R_0} - \frac{R_1}{(M-1)R_0}. \tag{4}$$

Moreover, condition eq. (3) *is implied by the mixed-difference condition*

$$\mathcal{D}_1 > \frac{2M-1}{M-1}\,\mathcal{D}_2. \tag{5}$$

When $L \ge 3$, the implication is strict in general: there exist spectra for which the first-order energy criterion holds but the mixed-difference criterion fails.

Proof. By Theorem 4.1, the variance of first-order elimination is

$$\mathrm{Var}(\widehat{g}_{\mathrm{FO}}) = \frac{1}{M}\sum_{k\ge 1}\left[1 + \frac{(-1)^k}{(M-1)^{k-1}}\right]E_k,$$

while $\mathrm{Var}(\widehat{g}_{\mathrm{iid}}) = R_0/M$. Subtracting,

$$\mathrm{Var}(\widehat{g}_{\mathrm{iid}}) - \mathrm{Var}(\widehat{g}_{\mathrm{FO}}) = \frac{1}{M}\sum_{k\ge 1}\left[1 - 1 - \frac{(-1)^k}{(M-1)^{k-1}}\right]E_k = -\frac{1}{M}\sum_{k\ge 1}\frac{(-1)^k}{(M-1)^{k-1}}E_k.$$

For $k = 1$ the summand is

$$-\frac{-1}{1}E_1 = E_1,$$

and for $k \ge 2$ it equals

$$-\frac{(-1)^k}{(M-1)^{k-1}}E_k,$$

which is

$$-\frac{E_k}{(M-1)^{k-1}}$$

for even k and

$$+\frac{E_k}{(M-1)^{k-1}}$$

for odd k. This yields eq. (1). Dropping the nonnegative odd-k sums gives eq. (2), so eq. (3) guarantees a strictly positive difference. The relative reduction follows from dividing by $\mathrm{Var}(\widehat{g}_{\mathrm{iid}}) = R_0/M$, giving

$$\eta = M(\mathrm{Var}(\widehat{g}_{\mathrm{iid}}) - \mathrm{Var}(\widehat{g}_{\mathrm{FO}}))/R_0 \geq (E_1 - R_1/(M-1))/R_0,$$

which is eq. (4).

It remains to show that eq. (5) implies eq. (3). Recall $\mathcal{D}_s = \sum_{k\geq s} \binom{k}{s} E_k$, so that

$$\mathcal{D}_1 = \sum_{k\geq 1} kE_k, \qquad \mathcal{D}_2 = \sum_{k\geq 2} \binom{k}{2} E_k.$$

The inequality $\mathcal{D}_1 > \frac{2M-1}{M-1}\mathcal{D}_2$ is equivalent to

$$E_1 > \sum_{k\geq 2} \left[\frac{2M-1}{M-1}\binom{k}{2} - k\right] E_k = \sum_{k\geq 2} f(k)\, E_k, \qquad f(k) := \frac{2M-1}{M-1}\binom{k}{2} - k.$$

We claim $f(k) \geq \frac{1}{M-1}$ for every $k \geq 2$. For $k = 2$,

$$f(2) = \frac{2M-1}{M-1} \cdot 1 - 2 = \frac{1}{M-1},$$

with equality. For $k \geq 3$, since $\binom{k}{2} = \frac{k(k-1)}{2}$ and $\frac{k-1}{2} \geq 1$, we have $\binom{k}{2} \geq k$, and therefore

$$f(k) = \frac{2M-1}{M-1}\binom{k}{2} - k \geq \frac{2M-1}{M-1}\, k - k = \left(\frac{2M-1}{M-1} - 1\right) k = \frac{M}{M-1}\, k \geq \frac{1}{M-1},$$

where the last step uses $M \geq 1$ and $k \geq 3$. Hence

$$E_1 > \sum_{k\geq 2} f(k) E_k \geq \frac{1}{M-1} \sum_{k\geq 2} E_k = \frac{R_1}{M-1},$$

which is eq. (3).

Finally, we show that the implication is strict. Consider a spectrum supported on $k \in \{1, 3\}$, i.e., $E_k = 0$ for $k \notin \{1, 3\}$, with $E_1, E_3 > 0$. Then eq. (3) reads $E_1 > E_3/(M-1)$, whereas eq. (5) reads

$$E_1 + 3E_3 > \frac{2M-1}{M-1} \cdot 3E_3 \iff E_1 > \frac{3M}{M-1} E_3.$$

Since $\frac{3M}{M-1} > \frac{1}{M-1}$, the condition eq. (5) is strictly stronger: there exist E_1, E_3 with $\frac{E_3}{M-1} < E_1 \leq \frac{3M}{M-1} E_3$, for which eq. (3) holds but eq. (5) fails. Hence eq. (5) is sufficient but not necessary, as claimed. □

C General Survival Probabilities for Theorem 4.2

We present the construction and analysis for general $p_\ell \in (0, 1)$, extending the uniform case $p_\ell = 1/2$ of Section 4.2. The uniform case is recovered as $\rho_\ell = 1$.

Proposition C.1 (Weighted spectral attenuation)**.** *Let $p_\ell \in (0, 1)$, and define*

$$m_\ell = \min\{p_\ell, 1 - p_\ell\}, \qquad \rho_\ell = \frac{m_\ell}{\sqrt{p_\ell(1-p_\ell)}} = \sqrt{\frac{\min\{p_\ell, 1-p_\ell\}}{\max\{p_\ell, 1-p_\ell\}}}, \qquad \rho_S = \prod_{\ell \in S} \rho_\ell.$$

Let $A \in \mathbb{F}_2^{r\times L}$ have full row rank, let $W = \mathrm{row}_{\mathbb{F}_2}(A)$, and set $M = K = 2^r$.

For each $h \in \mathbb{F}_2^r$, choose $x_h \in \mathbb{F}_2^L$ satisfying $Ax_h = h$, draw

$$v_h \sim \mathrm{Uniform}(\ker A),$$

and define

$$X_h = 2(x_h + v_h) - \mathbf{1},$$

where $x_h + v_h$ is computed over $\mathbb{F}_2$ and then viewed as a bit vector. Given X_h, independently across coordinates generate Z_h by

$$\Pr((Z_h)_\ell = 1 \mid X_h) = p_\ell + m_\ell (X_h)_\ell.$$

All random draws associated with different strata are independent.

Then

$$\frac{1}{M}\sum_h \mathcal{L}(Z_h) = \mu_p, \qquad \mathbb{E}[\widehat{g}_{\mathrm{exact}}] = \mu,$$

where

$$\widehat{g}_{\mathrm{exact}} = \frac{1}{M}\sum_h g(Z_h).$$

Moreover,

$$\mathrm{Cov}(\widehat{g}_{\mathrm{iid}}) - \mathrm{Cov}(\widehat{g}_{\mathrm{exact}}) = \frac{1}{M}\sum_{S\in W\setminus\{0\}} \rho_S^2 b_S b_S^\top \succeq 0.$$

In particular,

$$\mathrm{Var}(\widehat{g}_{\mathrm{exact}}) = \frac{1}{M}\left(R_0 - \sum_{S\in W\setminus\{0\}} \rho_S^2 \|b_S\|^2\right).$$

Proof. Consider the reference joint distribution defined by independent uniform signs $X_1, \dots, X_L$ and the conditionally independent channel

$$\Pr(Z_\ell = 1 \mid X_\ell) = p_\ell + m_\ell X_\ell.$$

Under this reference distribution,

$$\Pr(Z_\ell = 1) = p_\ell,$$

and the pairs (X_ℓ, Z_ℓ) are independent across ℓ. Thus $Z \sim \mu_p$.

Define

$$H = A\left(\frac{X + \mathbf{1}}{2}\right) \in \mathbb{F}_2^r.$$

Then H is uniform, and the construction in the proposition samples independently from $\mathcal{L}(Z \mid H = h)$ for every h. The equal-weight mixture of these conditional distributions is therefore μ_p, proving unbiasedness.

For the Fourier alignment, write

$$\psi_S(X) = \prod_{\ell\in S} X_\ell.$$

For a single coordinate, using $\mathbb{E}[Z_\ell \mid X_\ell] = p_\ell + m_\ell X_\ell$, we obtain

$$\begin{aligned}\mathbb{E}[\chi_\ell(Z_\ell) \mid X_\ell] &= \frac{\mathbb{E}[Z_\ell \mid X_\ell] - p_\ell}{\sqrt{p_\ell(1-p_\ell)}} \\ &= \frac{m_\ell}{\sqrt{p_\ell(1-p_\ell)}} X_\ell = \rho_\ell X_\ell.\end{aligned}$$

Conditional independence across coordinates implies

$$\mathbb{E}[\chi_S(Z) \mid X] = \rho_S \psi_S(X).$$

Consequently,

$$\mathbb{E}[g(Z) \mid X] = \mu + \sum_{S \neq \varnothing} \rho_S b_S \psi_S(X).$$

Applying the affine-stratum character calculation from Theorem 4.2 on the latent bit cube gives

$$\mathrm{Cov}(\mathbb{E}[g(Z) \mid H]) = \sum_{S \in W \setminus \{0\}} \rho_S^2 b_S b_S^\top.$$

Theorem 3.2 now yields the covariance identity. Taking traces gives the variance formula. □

Interpretation of the weights. For nonuniform survival probabilities, the construction attenuates rather than necessarily cancels the selected mask Fourier modes. Each additional coordinate contributes a factor $\rho_\ell^2 \leq 1$:

$$S \subseteq T \quad \Longrightarrow \quad \rho_T^2 \leq \rho_S^2.$$

For heterogeneous survival probabilities, cardinality alone does not order the weights of arbitrary subsets. If

$$\rho_{\max} = \max_\ell \rho_\ell < 1,$$

then the uniform bound

$$\rho_S^2 \leq \rho_{\max}^{2|S|}$$

gives exponential attenuation of the guaranteed gain with degree.

D Certifying Assumptions 5.2 and 5.3

The following are sufficient conditions for the particular squared-loss objectives stated below. They do not automatically verify Assumptions 5.2 and 5.3 for every stochastic-depth objective.

In the realizable model

$$J(\theta) = \frac{1}{2}\mathbb{E}_z \|f_\theta(z) - f_{\theta^\star}(z)\|^2,$$

the target is generated by a common reference parameter. If z includes both data and a stochastic-depth mask, applying this model to the original supervised squared loss requires

$$F_{\theta^\star}(x; Z) = y \qquad \text{almost surely.}$$

This common-realizability condition is an additional assumption.

Assumptions 5.2 and 5.3 can be verified in two ways.

(i) Relative linearization error. Let $e = \theta - \theta^\star$, $r_\theta(z) = f_\theta(z) - f_{\theta^\star}(z)$, and

$$d_\theta(z) = J_\theta(z)e - r_\theta(z).$$

Assume, throughout the region under consideration, that for some $\kappa \in [0, 1)$,

$$\mathbb{E}_z \|d_\theta(z)\|^2 \leq \kappa^2 \mathbb{E}_z \|r_\theta(z)\|^2, \qquad \|J_\theta(z)\|_{\mathrm{op}} \leq K \quad \text{almost surely.}$$

For $J(\theta) = \frac{1}{2}\mathbb{E}_z\|r_\theta(z)\|^2$ and $g(\theta) = \mathbb{E}_z[J_\theta(z)^\top r_\theta(z)]$, Cauchy–Schwarz gives

$$\begin{aligned} \langle g(\theta), e\rangle &= \mathbb{E}_z\|r_\theta(z)\|^2 + \mathbb{E}_z\langle r_\theta(z), d_\theta(z)\rangle \\ &\geq (1-\kappa)\mathbb{E}_z\|r_\theta(z)\|^2 \\ &= 2(1-\kappa)J(\theta), \end{aligned}$$

while

$$\|g(\theta)\|^2 \leq K^2 \mathbb{E}_z\|r_\theta(z)\|^2 = 2K^2 J(\theta).$$

Hence Assumptions 5.2 and 5.3 hold with $c = 2(1-\kappa)$ and $B = 2K^2$.

(ii) Single ReLU neuron. Consider

$$F(w) = \frac{1}{2}\mathbb{E}_x\big[\sigma(w^\top x) - \sigma(w_\star^\top x)\big]^2, \qquad \sigma(a) = \max\{a, 0\},$$

with $\|x\| \le R_x$ almost surely and the derivative convention $\sigma'(a) = \mathbf{1}_{\{a>0\}}$. Suppose throughout the region of interest that

$$\mathbb{E}\big[xx^\top \mathbf{1}_{\{w^\top x>0,\ w_\star^\top x>0\}}\big] \succeq \lambda I, \qquad \lambda > 0.$$

Then Assumptions 5.2 and 5.3 hold with

$$c = \frac{2\lambda}{R_x^2}, \qquad B = 2R_x^2.$$

Proof. Write

$$e = w - w_\star, \qquad r(x) = \sigma(w^\top x) - \sigma(w_\star^\top x).$$

The backpropagation direction is

$$g(w) = \mathbb{E}\big[r(x)\mathbf{1}_{\{w^\top x>0\}} x\big].$$

If both preactivations are positive, the contribution to $\langle g(w), e\rangle$ is $(e^\top x)^2$. If $w^\top x > 0$ but $w_\star^\top x \le 0$, the contribution is

$$(w^\top x)\big(w^\top x - w_\star^\top x\big) \ge 0.$$

If $w^\top x \le 0$, it is zero. Therefore,

$$\langle g(w), e\rangle \ge \mathbb{E}\big[(e^\top x)^2 \mathbf{1}_{\{w^\top x>0,\ w_\star^\top x>0\}}\big] \ge \lambda \|e\|^2.$$

The Lipschitz property of ReLU gives

$$F(w) \le \frac{1}{2}\mathbb{E}(e^\top x)^2 \le \frac{R_x^2}{2}\|e\|^2.$$

Combining these inequalities yields

$$\langle g(w), e\rangle \ge \frac{2\lambda}{R_x^2} F(w).$$

Finally, Jensen's inequality gives

$$\begin{aligned}\|g(w)\|^2 &\le \mathbb{E}\big[r(x)^2 \mathbf{1}_{\{w^\top x>0\}} \|x\|^2\big] \\ &\le R_x^2 \mathbb{E} r(x)^2 \\ &= 2R_x^2 F(w).\end{aligned}$$

This proves the stated constants. □

Iteration complexity. Under the assumptions of Theorem 5.4, reaching an averaged excess loss at most ε requires

$$T = O\left(\frac{BD_0^2}{c^2\varepsilon} + \frac{D_0^2 qV}{c^2\varepsilon^2}\right),$$

where $q = 1 - \eta_0$ is the residual variance fraction of Theorem 5.4. In the noise-dominated regime the second term is scaled by q.