ANTICORRELATING MINIBATCH NOISE ACCELERATES ADAM

Anonymous authors
Paper under double-blind review

ABSTRACT

A minibatch gradient can be written as the full-dataset gradient plus a zero-mean sampling residual, which is temporally uncorrelated under independent sampling. This residual is central to stochastic optimization, but its temporal organization across updates remains largely overlooked. We introduce temporal antithesis, which makes consecutive residuals anticorrelated while preserving the minibatch one-step gradient's mean and covariance. We characterize analytically how temporal antithesis alters Adam's dynamics and show empirically that it substantially accelerates Adam per update, especially early in training. Because exact antithesis requires the full-dataset gradient, we propose superbatch constructions that approximate it using several ordinary minibatches. These constructions recover most of the early advantage of exact antithesis. We further observe the same early acceleration across convolutional networks, vision transformers, and a 124M-parameter GPT-2 model, as well as beyond Adam with SGD and Muon. Together, these findings establish temporal noise correlation as a consequential direction for understanding and designing stochastic optimizers.

1 INTRODUCTION

Modern neural networks owe much of their success to minibatch stochastic optimization. Beyond reducing computational cost, minibatch sampling introduces gradient noise that shapes optimization and generalization Smith et al. (2020); Zhu et al. (2019). In adaptive optimizers such as Adam Kingma & Ba (2017), this noise propagates through both the momentum and second-moment states, influencing the direction and coordinate-wise scale of subsequent updates Balles & Hennig (2018).

Most existing work on minibatch noise focuses on how its magnitude, geometry, and distribution affect optimization and generalization Wen et al. (2020). This perspective has motivated methods based on importance sampling and batch selection Katharopoulos & Fleuret (2018), as well as data-ordering strategies Gürbüzbalaban et al. (2019). A smaller body of work considers temporal dependence, either through correlated minibatches Kühn & Rosenow (2025) or deliberately correlated artificial perturbations Orvieto et al. (2022); Lucchi et al. (2022). This latter direction is related to antithetic sampling, a classical variance-reduction technique that has also been applied to stochastic gradient methods Hammersley & Morton (1956); Liu & Xu (2018). At the optimizer level, Positive-Negative Momentum alters the temporal aggregation of stochastic gradients by maintaining and combining separate momentum states Xie et al. (2021), while recent work studies how intrinsic minibatch noise shapes Adam's implicit bias Cattaneo & Shigida (2026). Despite these advances, how the temporal organization of intrinsic minibatch noise interacts with adaptive optimizer dynamics remains poorly understood.

We address this question by directly controlling the temporal dependence of the noise in minibatch gradients. Following the standard sampling-noise decomposition Wu et al. (2020), we write $g_B(\theta) = g_{\mathrm{tot}}(\theta) + \eta_B(\theta)$, where $g_{\mathrm{tot}}(\theta)$ is the gradient over the full training dataset and $\eta_B(\theta)$ is the residual induced by sampling minibatch B. Our central idea is to change only the temporal coupling of these residuals while preserving their one-step mean and covariance. We call this intervention *temporal antithesis*, in which consecutive optimizer updates receive approximately opposite residuals. By construction, each update retains the noise scale and geometry of ordinary minibatch training, while

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.

the noise accumulated across a pair is reduced. This controlled intervention allows us to analyze how temporal noise correlation propagates through Adam's internal states.

Temporal antithesis substantially accelerates Adam. We show that it reduces the spread of the parameters reached after a pair of updates while leaving the average two-step update nearly unchanged. The advantage is largest early in training and contracts as training proceeds. These findings identify temporal noise correlation as a direct lever for improving Adam's early optimization dynamics. Exact antithesis, however, requires the full-dataset gradient at every reflected update. To remove this dependency, we propose two superbatch constructions that approximate the desired temporal coupling from a group of ordinary minibatches, a superbatch, without explicitly computing the full-gradient. The *direct* construction keeps the first update of each pair an ordinary minibatch step and only builds the reflected second update from the superbatch, which makes it cheaper. The *dense* construction builds both updates from the same superbatch, which costs more but offers a more flexible basis on which future work can design richer temporal coupling patterns. Where exact antithesis is tractable, these constructions recover up to 95% of its early advantage. Across convolutional networks He et al. (2015), vision transformers Dosovitskiy et al. (2020), and a 124M-parameter GPT-2 model Radford et al. (2019), they also accelerate early Adam optimization. Additional experiments show that the acceleration persists with SGD with momentum Sutskever et al. (2013) and the Muon optimizer Jordan et al. (2024), further supporting temporal noise correlation as an overlooked yet consequential aspect of stochastic optimization. In brief, our contributions are as follows:

- **Temporal Antithesis Accelerates Early Optimization:** We propose to anticorrelate consecutive intrinsic minibatch residuals while preserving their one-step mean and covariance. Using the exact full-dataset gradient as the reflection center, we show that this coupling substantially accelerates Adam per update, especially early in training.
- **Characterizing Antithesis in Adam's Moments:** We show analytically that antithesis preserves Adam's expected moments while removing almost all first-moment fluctuations accumulated over a pair of updates. Second-moment fluctuations decrease only in signal-dominated coordinates. Empirically, it leaves the average parameters reached after two updates nearly unchanged, while reducing their spread.
- **Full-Gradient-Free Superbatch Constructions:** We propose two superbatch constructions that approximate temporal antithesis without computing the full-dataset gradient. Both retain the desired one-step gradient statistics and reproduce much of the early acceleration of exact antithesis.
- **Broad Transfer and a New Optimizer-Design Axis:** Temporal antithesis accelerates early optimization across convolutional networks, vision transformers, and GPT-2, and extends beyond Adam to SGD with momentum and Muon. These results establish temporal noise correlation as a new axis for optimizer design.

2 Related Work

Minibatch gradient noise. Prior work studies how the magnitude of minibatch noise affects optimization and generalization Smith et al. (2020), how its covariance interacts with the loss geometry Zhu et al. (2019); Wen et al. (2020), and whether its distribution matters beyond its first two moments Wu et al. (2020). The decomposition of a minibatch gradient into the full empirical gradient and a sampling residual is standard Wu et al. (2020). This perspective has also motivated artificial gradient noise with selected properties of minibatch noise Wu et al. (2020); Wen et al. (2020). Other approaches modify stochastic gradients through importance sampling or random reshuffling Katharopoulos & Fleuret (2018); Gürbüzbalaban et al. (2019). These methods change the sampling distribution or the order in which examples are processed. Our work instead controls the dependence between residuals presented at successive optimizer updates.

Temporal noise correlation. Temporal anticorrelations can arise naturally under epoch-based sampling without replacement Kühn & Rosenow (2025). Other work deliberately introduces temporal dependence into artificial perturbations. Anti-PGD (*Anticorrelated Perturbed Gradient Descent*) uses anticorrelated perturbations across consecutive gradient-descent updates Orvieto et al. (2022), while fractional-noise methods consider artificial perturbations with longer-range dependence Lucchi et al. (2022). Optimistic AMSGrad also uses information across updates, but ex-

ploits the predictability of consecutive minibatch gradients rather than controlling their sampling dependence Wang et al. (2021). Related variance-reduction approaches instead operate within a single update. Antithetic sampling is a classical technique Hammersley & Morton (1956) that has been applied to stochastic optimization by negatively correlating sample gradients within a minibatch Liu & Xu (2018). None of these approaches directly controls intrinsic minibatch residuals across parameter-changing updates. Anti-PGD explicitly identifies this setting as an open direction Orvieto et al. (2022); our work addresses this gap.

Minibatch noise in Adam. Adam propagates stochastic gradients through its first- and second-moment states Kingma & Ba (2017). Prior work interprets its updates through the mean and relative variance of stochastic gradients Balles & Hennig (2018), derives stochastic continuous-time approximations for its dynamics Malladi et al. (2022), and studies how minibatch noise affects its implicit bias Cattaneo & Shigida (2026). Positive-Negative Momentum combines two momentum streams to control effective stochastic-gradient noise, including in an Adam variant Xie et al. (2021). Our work instead intervenes directly on the cross-update coupling of intrinsic minibatch residuals while controlling their statistics at individual updates. This isolates how temporal coupling propagates jointly through Adam's first moment, second moment, and parameter trajectory.

3 METHOD

Here we define the temporal dependence of minibatch residuals and motivates its study. We then introduce temporal antithesis and analyze its effect on Adam. Finally, we present two practical superbatch constructions that approximate temporal antithesis without computing the full-dataset gradient.

3.1 MINIBATCH RESIDUALS ACROSS OPTIMIZER UPDATES

Recall the decomposition Wu et al. (2020) $g_B(\theta) = g_{\text{tot}}(\theta) + \eta_B(\theta)$ introduced in Section 1, where B is a minibatch of size b drawn uniformly from N training examples. At each update, the residual satisfies

$$\mathbb{E}_B[\eta_B(\theta)] = 0, \qquad \text{Cov}_B[\eta_B(\theta)] = \kappa_b \frac{\Sigma(\theta)}{b}, \tag{1}$$

where $\Sigma(\theta)$ is the covariance of the per-example gradients and κ_b equals 1 with replacement and $(N-b)/(N-1)$ without replacement (Appendix A.1). These marginal statistics, however, say nothing about how residuals across updates relate to one another, which is central to how optimizers with memory actually experience this noise. Writing $\eta_t = \eta_{B_t}(\theta_t)$ for the **residual** at update t, we investigate its temporal correlation through the joint law $p(\eta_t, \eta_{t+1} \mid \mathcal{F}_{t-1})$, where $\mathcal{F}_{t-1}$ contains the parameters, optimizer states, and data-stream state **before update** t. Two residual streams can share the marginals in Equation 1 while differing in their joint temporal law. With shuffled sampling without replacement, ordinary training already induces weak dependence between successive residuals Kühn & Rosenow (2025). To quantify this dependence, we use the local approximation

$$\theta_{t+1} \approx \theta_t, \tag{2}$$

which further assumes that the residual of a given minibatch, the full-dataset gradient, and the residual covariance change negligibly over one update (Appendix A.2). Under this approximation,

$$\text{Cov}\,(\eta_t, \eta_{t+1}) \approx -\frac{1}{M}\,\text{Cov}_B[\eta_B(\theta_t)], \tag{3}$$

where $M = N/b$ is the number of minibatches per epoch. Consecutive residuals therefore have normalized correlation $-1/M = -b/N$ in every direction. This weak anticorrelation is independent of the scale of $\Sigma(\theta)$ (see Appendix A.3). Kühn & Rosenow (2025) propagate the complete epoch-scale anticorrelation structure through local momentum-SGD dynamics and show that it reduces stationary parameter and loss fluctuations along sufficiently flat directions. However, this natural anticorrelation is spread across all M minibatches in an epoch. By contrast, Adam's first moment is an exponential average with decay rate β_1 that retains only about $1/(1-\beta_1)$ recent gradients, a window typically far shorter than one epoch. Most of the epoch's natural cancellation therefore falls outside the span Adam actually integrates over. This mismatch motivates strengthening the negative dependence between residuals and concentrating it on consecutive updates, where adaptive optimizers can actually exploit it.

3.2 TEMPORAL ANTITHESIS

We propose *temporal antithesis*, a coupling in which consecutive Adam updates receive approximately opposite minibatch residuals. We present an ideal full-gradient construction, analyze its effect on Adam's moment states, and finally develop a practical superbatch approximation.

3.2.1 ANTITHETIC MECHANISM

Let $C(\theta) = \mathrm{Cov}_B[\eta_B(\theta)]$ denote the one-step residual covariance in Equation 1. Under the local approximation of Equation 2, $C(\theta_{2t+1}) \approx C(\theta_{2t})$, so both residuals of a pair can be taken at θ_{2t}. Let η and γ be independent zero-mean residuals with covariance $C(\theta_{2t})$, and define the independent and antithetic residual pairs construction as

$$(\eta_{2t}, \eta_{2t+1}) = \begin{cases} (\eta, \gamma)\,, & \textit{iid} \text{ (ordinary training)}, \\ (\eta, -\eta)\,, & \textit{antithetic}. \end{cases} \tag{4}$$

Both constructions satisfy $\mathbb{E}[\eta_{2t}] = \mathbb{E}[\eta_{2t+1}] = 0$ and $\mathrm{Cov}(\eta_{2t}) = \mathrm{Cov}(\eta_{2t+1}) = C(\theta_{2t})$, and differ only in their cross-update covariance $\mathrm{Cov}(\eta_{2t}, \eta_{2t+1})$ equal to 0 for the iid pair and $-C(\theta_{2t})$ for the antithetic pair. Ordinary shuffled training lies in between, with covariance $-C(\theta_{2t})/M$. Since the number of minibatches per epoch M is usually large, this coupling is negligible, and the iid pair provides a convenient model of ordinary training. Temporal antithesis therefore replaces this weak coupling with the strongest possible, $-C(\theta_{2t})$, without changing the one-step mean or covariance.

Application during training. In practice, each antithetic pair of updates uses a single minibatch B:

$$g_{2t} = g_B(\theta_{2t}) = g_{\mathrm{tot}}(\theta_{2t}) + \eta_B(\theta_{2t}), \tag{5}$$

$$g_{2t+1} = g_{\mathrm{tot}}(\theta_{2t+1}) - \eta_B(\theta_{2t+1}) = 2g_{\mathrm{tot}}(\theta_{2t+1}) - g_B(\theta_{2t+1}). \tag{6}$$

The ordinary gradient g_{2t} is applied before the same minibatch is reevaluated at θ_{2t+1} and its residual reflected around the new full-gradient center. Each antithetic pair consumes a single minibatch, so at a fixed number of updates exact antithesis draws half as many distinct minibatches as ordinary training. Computing g_{tot} at every reflected update is, however, impractical at scale. Section 3.2.3 proposes an alternative that does not require the full gradient.

3.2.2 EFFECT ON ADAM'S MOMENT STATES

Writing $g_t = g_{\mathrm{tot}}(\theta_t) + \eta_t$, Adam's two moments decompose exactly as $m_t = \bar{m}_t + \delta m_t$ and $v_t = \bar{v}_t + \delta v_t$, where each component follows the same exponential-average recursion as its moment. $\bar{m}_t$ averages the full gradients g_{tot} and δm_t the residuals η_t; $\bar{v}_t$ averages $g_{\mathrm{tot}}^2 + c_t$, with $c_t = \mathrm{diag}\, C(\theta_t)$ the residual variance, and δv_t the zero-mean terms $2g_{\mathrm{tot}} \odot \eta_t + \eta_t^2 - c_t$. Considering a given training history $\mathcal{F}_{2t-1}$ before a pair of updates made with either the *iid* or the *antithetic* construction, and under the local approximation of Equation 2, we can prove the following two propositions (see Appendices A.4.1 and A.4.2).

Proposition 1 (First moment)**.** *The iid and antithetic pairs give the same conditional expectation* $\mathbb{E}[m_{2t+1} \mid \mathcal{F}_{2t-1}]$*, and*

$$\mathrm{Cov}\big(m_{2t+1}^{\mathrm{anti}} \mid \mathcal{F}_{2t-1}\big) = \frac{(1-\beta_1)^2}{1+\beta_1^2}\, \mathrm{Cov}\big(m_{2t+1}^{\mathrm{iid}} \mid \mathcal{F}_{2t-1}\big)\,. \tag{7}$$

Proposition 2 (Second moment)**.** *The iid and antithetic pairs give the same conditional expectation* $\mathbb{E}[v_{2t+1} \mid \mathcal{F}_{2t-1}]$*. If, moreover, each residual coordinate has zero third moment, then for each coordinate* i*, with* $r_i = 4g_{\mathrm{tot},i}^2 c_{2t,i} / \mathrm{Var}(\eta_{2t,i}^2)$*,*

$$\frac{\mathrm{Var}\big(v_{2t+1,i}^{\mathrm{anti}} \mid \mathcal{F}_{2t-1}\big)}{\mathrm{Var}\big(v_{2t+1,i}^{\mathrm{iid}} \mid \mathcal{F}_{2t-1}\big)} = \frac{(1-\beta_2)^2 r_i + (1+\beta_2)^2}{(1+\beta_2^2)(1+r_i)}, \tag{8}$$

which is below one exactly when $r_i > 1$*.*

Hence, for $\beta_1 = 0.9$, antithesis removes 99.45% of the first-moment covariance accumulated over the pair, even though each update retains a full-size residual. Its effect on v is mixed: it stabilizes signal-dominated coordinates ($r_i > 1$) and destabilizes noise-dominated ones. Moreover, the first

moment averages only about $1/(1-\beta_1) \approx 10$ recent gradients for $\beta_1 = 0.9$, whereas the second averages about $1/(1-\beta_2) \approx 1000$ for $\beta_2 = 0.999$ and is nearly unchanged by the construction. The analysis therefore makes three predictions. First, antithesis should act mainly through the first moment. Second, since it leaves the expected displacement over a pair unchanged while reducing its dispersion, and dispersion raises the expected loss along curved directions (Appendix B.1), it should accelerate optimization. Third, this acceleration should be largest early in training, when the loss is far from its minimum and each reduction of dispersion translates into a large gain, and contract as both constructions approach the same loss level. We confirm these predictions experimentally in Section 4.2.

3.2.3 SUPERBATCH APPROXIMATIONS OF EXACT ANTITHESIS

Exact temporal antithesis requires the full-gradient center at every reflected update. We propose two superbatch constructions that avoid this expensive computation. Both use a superbatch of K independently drawn minibatches. The first estimates g_{tot} with the superbatch mean and requires $K+1$ minibatch-gradient evaluations per pair. The second uses dense signed contrasts over all K minibatches; it requires $2K$ evaluations per pair, but treats all minibatches symmetrically and permits more general temporal-correlation patterns.

Direct superbatch reflection. At update $2t$, we apply the ordinary gradient of a minibatch B_1, as in Equation 5. At update $2t+1$, we replace the full-gradient center in Equation 6 with the mean of K minibatch gradients evaluated at the new parameter state, including B_1:

$$g_{2t+1} = 2\bar{g}_K(\theta_{2t+1}) - g_{B_1}(\theta_{2t+1}), \qquad \bar{g}_K(\theta) = \frac{1}{K}\sum_{j=1}^{K} g_{B_j}(\theta), \tag{9}$$

where $B_2, \ldots, B_K$ are drawn independently from the same minibatch distribution. Under the local approximation, g_{2t+1} has the same marginal mean and covariance as an ordinary minibatch gradient, while the residuals across the pair have correlation

$$\rho_{\text{anti}} = -\frac{K-2}{K}. \tag{10}$$

The coupling strengthens with K. For $K = 2, 4$, and 8, the correlation is $0, -0.5$, and -0.75, respectively, and it approaches the exact antithetic value -1 as K grows. The proof is provided in Appendix A.5.1.

Dense signed-contrast construction. Direct reflection privileges B_1 as the first optimizer input. A symmetric alternative distributes both inputs across all K minibatches, reevaluating each minibatch at both parameter states. Let $z, w \in \mathbb{R}^K$ be contrast vectors satisfying $\sum_j z_j = \sum_j w_j = 0$ and $\frac{1}{K}\sum_j z_j^2 = \frac{1}{K}\sum_j w_j^2 = 1$. The two inputs are

$$g_{2t} = \bar{g}_K(\theta_{2t}) + \frac{\sqrt{K-1}}{K}\sum_{j=1}^{K} z_j\, g_{B_j}(\theta_{2t}), \qquad g_{2t+1} = \bar{g}_K(\theta_{2t+1}) + \frac{\sqrt{K-1}}{K}\sum_{j=1}^{K} w_j\, g_{B_j}(\theta_{2t+1}), \tag{11}$$

where $\bar{g}_K$ is defined in Equation 9. Because the contrast coefficients sum to zero, the full-gradient component cancels from the signed terms. Under the local approximation, each input again has the mean and covariance of an ordinary minibatch gradient, while the residuals across the pair have correlation

$$\rho = \frac{1 + (K-1)\langle z, w\rangle_K}{K}, \qquad \langle z, w\rangle_K = \frac{1}{K}\sum_{j=1}^{K} z_j w_j. \tag{12}$$

Choosing $w = -z$ recovers $\rho_{\text{anti}} = -(K-2)/K$, matching direct reflection, whereas $\langle z, w\rangle_K = -1/(K-1)$ gives exactly zero correlation. The latter provides a matched control with the same minibatches, computational cost, and one-step mean and covariance, differing only in its temporal coupling. The proof is provided in Appendix A.5.2.

Table 1: Second update of each pair. $M = N/b$ is the number of minibatches in one AdamW epoch. Repeated batch and exact antithesis reuse the even-step minibatch at the odd step, so they see $M/2$ distinct minibatches per epoch.

Condition	Update	Batches
Repeated batch	$g_{B_{2t}}(\theta_{2t+1})$	$M/2$
IID reflection	$2g_{\text{tot}}(\theta_{2t+1}) - g_{B_{2t+1}}(\theta_{2t+1})$	M
AdamW	$g_{B_{2t+1}}(\theta_{2t+1})$	M
Exact antithesis	$2g_{\text{tot}}(\theta_{2t+1}) - g_{B_{2t}}(\theta_{2t+1})$	$M/2$

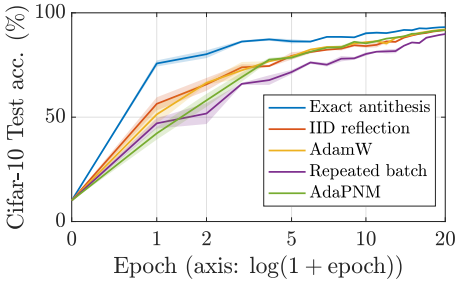

Figure 1: Test accuracy of ResNet-18 along the 20-epoch CIFAR-10 schedule. Epoch 0 is the untrained network. Mean ± s.e. over three seeds.

4 Experiments

We organize the experiments around four questions. First, we test whether temporal antithesis causally accelerates Adam, using exact reflections of real minibatch residuals on CIFAR-10 and comparing against controls matched in one-step gradient mean and covariance. Second, we examine why it accelerates Adam by testing the predictions of Section 3.2.2 using exact real residuals $\eta_B = g_B - g_{\text{tot}}$, which allow us to control the noise entering each moment state separately. Third, we ask whether the effect persists without the full gradient, evaluating both superbatch constructions for increasing K. Finally, we assess whether the superbatch constructions transfer to convolutional networks, vision transformers, and GPT-2, as well as to the SGD and Muon optimizers. All Adam experiments use AdamW (Loshchilov & Hutter, 2019), which adds decoupled weight decay to Adam; since weight decay does not enter the moment recursions, the analysis of Section 3.2.2 applies unchanged.

4.1 Temporal antithesis accelerates Adam

Using the exact intervention of Equations 5 and 6, we show that temporal antithesis substantially accelerates AdamW, especially early in training.

Setup. We use CIFAR-10 with $45{,}000$ training, $5{,}000$ validation, and $10{,}000$ test images. We train a ResNet-18 with BatchNorm using AdamW, batch size 128, weight decay 10^{-4}, random cropping and horizontal flipping, and a warmup–stable–decay schedule. The learning rate 1.5×10^{-3} was selected by sweeping ordinary AdamW, and the antithetic runs and all control use it without further tuning. All comparisons use three paired seeds. We also compare against AdaPNM (Xie et al., 2021), which enlarges the noise inside the momentum state by combining a positive and a negative momentum buffer, using one gradient per update. It is one of the few methods that act on minibatch noise across consecutive updates, and it requires no additional gradient evaluations, since it operates entirely within the optimizer. It thus provides a computationally inexpensive alternative from a different paradigm. Since it rescales the update, we sweep its learning rate over $\{7.5 \times 10^{-4}, 1.5 \times 10^{-3}, 3 \times 10^{-3}\}$ and report the best.

All methods follow the same pattern. At every even update, they apply an ordinary minibatch gradient and differ only in the subsequent update (Table 1). Exact antithesis reuses the same minibatch and flips the sign of its noise. To verify that the gain comes from this sign flip, we use two controls. IID reflection performs the same computation on a new minibatch, so the two noises are unrelated and do not cancel. Repeated batch reuses the same minibatch without flipping its noise, so the same noise is applied twice.

Early acceleration. Exact antithesis leads at every horizon, most strongly early in training (Table 2, Figure 1). After two epochs it reaches $\mathbf{88.07}\%$ test accuracy against 75.36% for AdamW. Accuracy orders exactly as the residual correlation ρ: repetition ($\rho \approx +1$) is worst, the two arms with near-zero correlation, AdamW and IID reflection, are indistinguishable, and antithesis ($\rho \approx -1$) is best. The gain therefore cannot come from computing the full gradient or from reusing a minibatch, since the

Table 2: CIFAR-10 test accuracy of exact antithesis and its controls (upper block) and of the two superbatch constructions at $K = 8$ with their zero-correlation controls (lower block). ρ is the correlation between the two residuals of a pair, under the local approximation with $M = 351$. Cost is the number of minibatch gradients per update relative to AdamW. Values are mean $\pm$ standard error over three seeds; Δ is the paired difference. The remaining values of K are shown in Figure 5.

			Short schedules		20-epoch schedule			
Method	ρ	Cost	2/2	5/5	5/20	10/20	15/20	20/20
Exact interventions								
Repeated batch	$+1$	$\times 1$	68.35 ± 0.32	80.16 ± 0.17	71.62 ± 1.01	80.27 ± 0.82	85.79 ± 0.25	89.83 ± 0.14
IID reflection	$+2.86\times10^{-3}$	$\times 176$	75.48 ± 0.30	85.44 ± 0.17	79.36 ± 1.77	84.07 ± 0.82	89.16 ± 0.49	91.73 ± 0.04
AdaPNM (Xie et al., 2021)	-2.85×10^{-3}	$\times 1$	73.48 ± 0.30	84.89 ± 0.09	78.51 ± 0.85	85.38 ± 0.35	89.97 ± 0.15	91.71 ± 0.04
AdamW	-2.85×10^{-3}	$\times 1$	75.36 ± 0.35	85.68 ± 0.06	79.05 ± 0.69	85.96 ± 0.06	89.42 ± 0.14	92.08 ± 0.11
Exact antithesis	$\mathbf{-1}$	$\times 176$	$\mathbf{88.07 \pm 0.13}$	$\mathbf{91.50 \pm 0.12}$	$\mathbf{86.35 \pm 1.00}$	$\mathbf{90.20 \pm 0.58}$	$\mathbf{91.66 \pm 0.49}$	$\mathbf{93.13 \pm 0.17}$
Δ(Antithesis $-$ AdamW)	–	–	$\mathbf{+12.71 \pm 0.29}$	$\mathbf{+5.81 \pm 0.16}$	$\mathbf{+7.30 \pm 1.32}$	$\mathbf{+4.24 \pm 0.65}$	$\mathbf{+2.24 \pm 0.35}$	$\mathbf{+1.05 \pm 0.26}$
Superbatch, $K = 8$								
Direct, $\rho = 0$ control	0	$\times 4.5$	75.73 ± 0.43	85.57 ± 0.06	80.83 ± 0.34	85.49 ± 0.29	89.72 ± 0.14	91.82 ± 0.08
Direct antithesis	-0.75	$\times 4.5$	$\mathbf{85.00 \pm 0.14}$	$\mathbf{90.49 \pm 0.13}$	$\mathbf{85.43 \pm 0.59}$	$\mathbf{86.96 \pm 1.12}$	$\mathbf{91.29 \pm 0.17}$	$\mathbf{92.89 \pm 0.10}$
Dense, $\rho = 0$ control	0	$\times 8$	75.73 ± 0.18	85.63 ± 0.17	77.13 ± 0.77	85.13 ± 0.81	88.95 ± 0.12	92.06 ± 0.05
Dense antithesis	-0.75	$\times 8$	$\mathbf{84.74 \pm 0.09}$	$\mathbf{90.53 \pm 0.05}$	$\mathbf{87.30 \pm 0.58}$	$\mathbf{89.20 \pm 0.23}$	$\mathbf{91.75 \pm 0.14}$	$\mathbf{93.24 \pm 0.07}$

controls do both without gaining. AdaPNM, which enlarges the noise inside the momentum at no extra cost, stays below AdamW. Under the 20-epoch schedule the advantage over AdamW narrows, from $+7.30$ points at epoch 5 to $+1.05$ at epoch 20, ending at $\mathbf{93.13}\%$ against 92.08%. The complete two- and five-epoch runs, each with its own learning-rate decay, show the same acceleration, so it is not an artifact of reading checkpoints in the middle of a longer schedule. We next analyse the source of this acceleration, testing the predictions of Section 3.2.2.

4.2 Understanding the acceleration

We now show that the gain comes from the negative correlation itself and test the predictions of Section 3.2.2. In particular, we trace the acceleration to reduced variance in Adam's first-moment state and the resulting reduction in parameter dispersion. Except if precised, experiments in this section use CIFAR-10 with ResNet-18 and a two-epoch training schedule.

Strength of the anticorrelation. We vary here the correlation of the pair from +1 to -1. Each pair starts with an ordinary update on a minibatch B. At the second update, we re-evaluate B at the current parameters, giving the residual $\eta_B = g_B - g_{\text{tot}}$, and draw an independent minibatch B' with residual $\gamma = g_{B'} - g_{\text{tot}}$. Let u be the component of γ orthogonal to η_B, rescaled to $\|\eta_B\|$. The second update uses $g_{\text{tot}} + \eta(\phi)$, with $\eta(\phi) = \cos\phi\,\eta_B + \sin\phi\,u$, a rotation of η_B in the plane spanned by η_B and u, so the pair correlation is exactly $\rho = \cos\phi$. Taking $\phi \in \{0°, 90°, 120°, 150°, 180°\}$ sweeps ρ from $+1$, where the same residual is repeated, to -1, which is antithesis. The gain over an independent minibatch grows monotonically as the pair becomes more opposed, from -7.48 points at $\rho = +1$ to $+11.28$ at $\rho = -1$, and is close to zero for the uncorrelated pair ($+0.66 \pm 0.13$ at $\rho = 0$; Figure 2). The gain is therefore set by the correlation of the pair.

Localization in Adam's moment states. Section 3.2.2 predicts that the acceleration comes from the first moment. We test this by feeding Adam's two states separate residuals, $m_t = \beta_1 m_{t-1} + (1-\beta_1)(g_{\text{tot}} + \eta_{m,t})$ and $v_t = \beta_2 v_{t-1} + (1-\beta_2)(g_{\text{tot}} + \eta_{v,t})^2$, each antithetic or iid; ordinary Adam corresponds to $\eta_{m,t} = \eta_{v,t}$ (see Table 3). Antithesis in m_t alone improves accuracy by $\mathbf{11.21 \pm 0.62}$ points over ordinary Adam, on all three seeds, and matches exact antithesis (87.95% vs. 88.07%), whereas antithesis in v_t alone has no detectable effect ($+0.34 \pm 0.36$). The acceleration is therefore carried entirely by the first moment.

Table 3: Antithesis applied to each moment state separately.

Condition	m_t	v_t	Accuracy (%)
Ordinary Adam	*iid*	*iid*	76.73 ± 0.43
Antithetic m_t only	*anti*	*iid*	$\mathbf{87.95 \pm 0.47}$
Antithetic v_t only	*iid*	*anti*	77.07 ± 0.07
Exact *antithetic*	*anti*	*anti*	$\mathbf{88.07 \pm 0.50}$

Momentum variance reduction. Having localized the effect in m, we ask whether the variance reduction predicted in Section 3.2.2 is the channel through which it acts. As shown in Proposition 1, antithesis suppresses a fraction $1 - (1-\beta_1)^2/(1+\beta_1^2)$ of the pair-induced first-moment covariance. Thus, varying β_1 directly tests whether the gain follows the predicted suppression (see

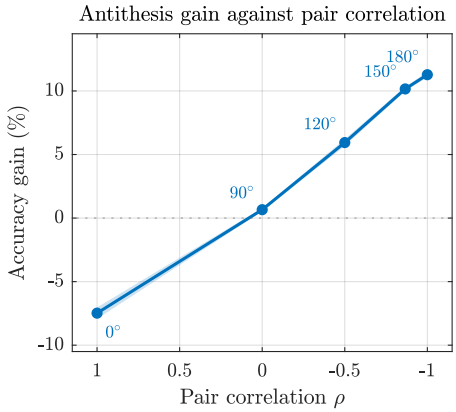

Figure 2: **Strength of the anticorrelation.** Test-accuracy gain over an independent minibatch as a function of the pair correlation $\rho = \cos\phi$, from repetition (0°) to antithesis (180°) (two-epoch CIFAR-10; mean ± s.e. over three seeds).

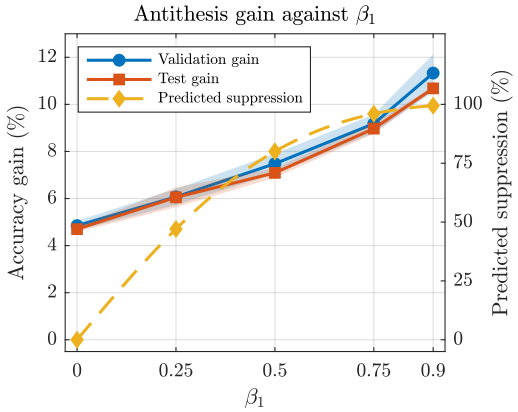

Figure 3: **Antithesis gain against β_1.** Accuracy gain over an independent minibatch (left axis) and predicted first-moment suppression (right axis) (two-epoch CIFAR-10; mean ± s.e. over three seeds).

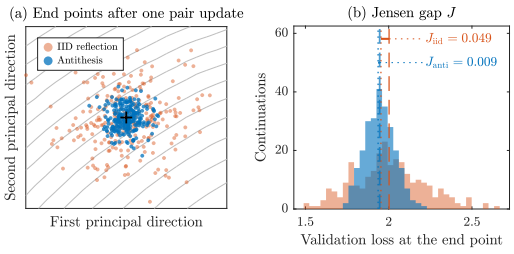

Figure 4: Reduced dispersion (one seed). (a) End points on the two leading *iid* directions. (b) End-point losses; J is the gap between the dotted and dashed lines.

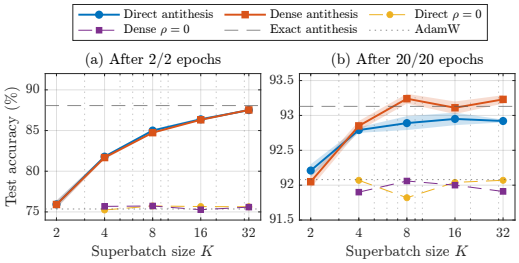

Figure 5: Superbatch antithesis against K after (a) two and (b) 20 epochs, with zero-correlation controls, exact antithesis (dashed) and AdamW (dotted); mean ± s.e. over three seeds.

Figure 3). The validation gain grows from $+4.9$ points at $\beta_1 = 0$ to $+11.3$ at $\beta_1 = 0.9$, increasing at every step of β_1, so the predicted suppression has Spearman correlation 1.0 with the gain. The first-moment variance reduction is therefore an important channel. At $\beta_1 = 0$, however, antithesis still improves validation accuracy by 4.9 points despite no predicted variance reduction, indicating additional effects not captured by the local approximation, such as the parameter change between the two gradients (Appendix A.4.1).

Reduced dispersion. We next examine how the variance reduction in the first moment affects the parameters reached (the end points) after a pair of updates. From a checkpoint taken after one epoch of AdamW, we run 256 independent two-update continuations with each construction (*iid* and *antithetic*), for three seeds (see Figure 4a). We compare the loss averaged over the end points with the loss at their average end point. The difference, $J = \mathbb{E}[L_{\text{val}}(\theta_{t+2})] - L_{\text{val}}(\mathbb{E}[\theta_{t+2}])$, is the Jensen gap: it is zero when all continuations end at the same point, and grows with their spread along curved directions, since $J \approx \frac{1}{2}\operatorname{tr}(H\,\mathrm{Cov}(\theta_{t+2}))$ (Appendix B.1). In units of 10^{-2} validation loss, the average end points of the two constructions have the same loss (difference -0.60 ± 0.56), but antithesis reduces J from 4.48 ± 0.28 to $\mathbf{0.84 \pm 0.07}$, an $\mathbf{81}\%$ reduction. Its lower average loss (-4.23 ± 0.72) therefore comes almost entirely from reduced dispersion (Figure 4b). Repeated over every pair of updates, this lower loss per pair is what accelerates training.

4.3 Superbatch antithesis reproduces the acceleration

Exact antithesis accelerates Adam, but at $\times 176$ the gradient cost of AdamW, since every reflected update needs g_{tot}. We now evaluate the two superbatch constructions of Section 3.2.3, which build

Table 4: Dense $K = 8$ antithesis applied at every update for the whole schedule, against the same optimizer without antithesis. The first row is the reference setting; in the other rows, bold marks what changes with respect to it. The four gain columns give the paired test-accuracy gain of antithesis after 10%, 25%, 50% and 100% of the training schedule, in green when positive and red when negative (for GPT, change in validation loss). The last column gives the final test accuracy, without $\rightarrow$ with antithesis. All values are mean $\pm$ standard error over three seeds.

				Gain after % of the training schedule				
Model	Dataset	Optimizer	Schedule	10%	25%	50%	100%	Final acc. (%)
ResNet-18	CIFAR-10	AdamW	20 ep. (7,040 upd.)	$+14.01 \pm 0.63$	$+8.25 \pm 1.23$	$+3.24 \pm 0.28$	$+1.16 \pm 0.04$	$92.08 \pm 0.11 \rightarrow 93.24 \pm 0.07$
ResNet-18	CIFAR-10	**SGD**	20 ep. (7,040 upd.)	$+15.39 \pm 0.91$	$+9.10 \pm 1.95$	$+4.55 \pm 1.35$	$+1.49 \pm 0.09$	$92.80 \pm 0.10 \rightarrow 94.29 \pm 0.05$
ResNet-18	CIFAR-10	**Muon**	20 ep. (7,040 upd.)	$+7.87 \pm 0.87$	$+4.38 \pm 0.29$	$+2.18 \pm 0.37$	$+0.47 \pm 0.14$	$93.89 \pm 0.06 \rightarrow 94.37 \pm 0.10$
ResNet-18	**CIFAR-100**	AdamW	20 ep. (7,040 upd.)	$+16.69 \pm 1.16$	$+11.75 \pm 1.65$	$+5.00 \pm 1.38$	$+0.30 \pm 0.49$	$70.78 \pm 0.22 \rightarrow 71.08 \pm 0.32$
ResNet-18	**TinyImageNet**	AdamW	10 ep. (7,030 upd.)	$+15.53 \pm 0.99$	$+19.22 \pm 0.83$	$+13.77 \pm 0.57$	$+1.49 \pm 0.15$	$56.65 \pm 0.16 \rightarrow 58.15 \pm 0.19$
ViT-Tiny	**CIFAR-100**	AdamW	30 ep. (10,530 upd.)	$+8.84 \pm 0.22$	$+11.96 \pm 0.51$	$+1.44 \pm 0.49$	-2.69 ± 0.15	$53.94 \pm 0.11 \rightarrow 51.25 \pm 0.19$
GPT-124M	**FineWeb**	AdamW	5,722 upd. (1.5B tok.)	-0.176 ± 0.011	-0.194 ± 0.001	-0.178 ± 0.001	-0.112 ± 0.001	$3.480 \pm 0.001 \rightarrow 3.368 \pm 0.001$ (loss)

the reflection from K ordinary minibatches, together with their zero-correlation controls, which use the same minibatches at identical cost. Table 2 (lower block) reports both constructions at $K = 8$, Figure 5 the sweep over $K \in \{2, 4, 8, 16, 32\}$, and Appendix B.2 the full numerical results.

Both constructions recover most of the acceleration at a small fraction of the cost. At $K = 8$, the direct construction reaches 85.00% after two epochs and the dense one 84.74%, against 88.07% for exact antithesis and 75.36% for AdamW, at $\times 4.5$ and $\times 8$ rather than $\times 176$; by the end of the 20-epoch schedule they reach 92.89% and 93.24%, matching exact antithesis (93.13%) within error. The gain grows with K, while the zero-correlation controls stay at the AdamW level for every K (Figure 5).

4.4 The effect transfers across architectures, modalities, and optimizers

The effect generalizes beyond our CIFAR-10 setup. We apply the dense construction with $K = 8$ at every update for the whole schedule and compare it with the same optimizer without antithesis (Table 4). Starting from ResNet-18 He et al. (2015) on CIFAR-10 with AdamW, we change the optimizer, to SGD with momentum and Muon Jordan et al. (2024); the dataset, to CIFAR-100 and TinyImageNet; the architecture, to ViT-Tiny Dosovitskiy et al. (2020); and the modality, to a 124M-parameter GPT-2 model Radford et al. (2019) trained on FineWeb Penedo et al. (2024).

Antithesis speeds up early training in every setting, leading the same optimizer without antithesis by 4 to 19 accuracy points after a quarter of the schedule, and by 0.19 in validation loss for GPT. The gain is largest with SGD with momentum and persists with Muon (whose standard recipe still trains biases, normalization parameters, and the classifier with AdamW). By the end of training the advantage narrows on image tasks but it remains positive in four settings, is within noise for ResNet-18 on CIFAR-100, and reverses on ViT-Tiny, which fits its training set within about ten epochs. On GPT the gain remains large at the end of training (-0.112 in validation loss), but our 1.5B-token schedule is short for a 124M-parameter model, and the gain may narrow with longer training as it does on image tasks.

5 Conclusion

We studied the temporal organization of minibatch noise. Temporal antithesis presents opposite residuals to consecutive updates while preserving their one-step statistics, and this change alone substantially accelerates Adam, most strongly early in training. It acts mainly through Adam's first moment, removing most of the noise a pair of updates accumulates there and reducing the dispersion of the parameters. Superbatch constructions built from a few ordinary minibatches recover most of the effect without the full gradient, and the acceleration transfers across architectures, datasets, optimizers, and language modeling. These constructions still evaluate several minibatches per update, so our gains are per update rather than per unit of compute. Our core contribution, however, is to establish the temporal structure of noise as a meaningful axis of stochastic optimization, distinct from batch size and learning rate. Estimating g_{tot} was the most direct way to demonstrate the effect, not an intrinsic requirement of temporal antithesis. Reducing this cost is therefore a well-posed direction that the magnitude and consistency of our gains make worth pursuing.

AI USE STATEMENT

Generative AI tools, were used to help edit the manuscript for language, grammar, and clarity. They were also used to help implement and run experiments, including writing training and analysis code, launching and monitoring runs, aggregating results, and producing figures. The authors reviewed all AI-assisted code and verified every reported number against the logged experimental outputs. The scientific conclusions of this work are the authors' own, and the authors take full responsibility for all of its content, including text, claims, code, experiments, and any artifacts produced with AI assistance.

REFERENCES

Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 404–413. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/balles18a.html.

Matias D. Cattaneo and Boris Shigida. The effect of mini-batch noise on the implicit bias of adam, 2026. URL https://arxiv.org/abs/2602.01642.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929.

M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186(1-2):49–84, October 2019. ISSN 1436-4646. doi: 10.1007/s10107-019-01440-w. URL http://dx.doi.org/10.1007/s10107-019-01440-w.

J. M. Hammersley and K. W. Morton. A new monte carlo technique: antithetic variates. *Mathematical Proceedings of the Cambridge Philosophical Society*, 52(3):449–475, 1956. doi: 10.1017/S0305004100031455.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. https://kellerjordan.github.io/posts/muon/, 2024.

Angelos Katharopoulos and Francois Fleuret. Not all samples are created equal: Deep learning with importance sampling. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2525–2534. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/katharopoulos18a.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Marcel Kühn and Bernd Rosenow. Anti-correlated noise in epoch-based stochastic gradient descent: implications for weight variances in flat directions. *Machine Learning: Science and Technology*, 6(4):045003, October 2025. ISSN 2632-2153. doi: 10.1088/2632-2153/ae0ab2. URL http://dx.doi.org/10.1088/2632-2153/ae0ab2.

Jingchang Liu and Linli Xu. Accelerating stochastic gradient descent using antithetic sampling, 2018. URL https://arxiv.org/abs/1810.03124.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Aurelien Lucchi, Frank Proske, Antonio Orvieto, Francis Bach, and Hans Kersting. On the theoretical properties of noise correlation in stochastic optimization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 14261–14273. Curran Associates, Inc., 2022. doi: 10.52202/068431-1037. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/5bed8703db85ab27dc32f6a42f8fbdb6-Paper-Conference.pdf.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the sdes and scaling rules for adaptive gradient algorithms. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 7697–7711. Curran Associates, Inc., 2022. doi: 10.52202/068431-0559. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/32ac710102f0620d0f28d5d05a44fe08-Paper-Conference.pdf.

Antonio Orvieto, Hans Kersting, Frank Proske, Francis Bach, and Aurelien Lucchi. Anticorrelated noise injection for improved generalization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17094–17116. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/orvieto22a.html.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024. URL https://arxiv.org/abs/2406.17557.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Accessed: 2024-11-15.

Samuel Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic gradient descent. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9058–9067. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/smith20a.html.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/sutskever13.html.

Jun-Kun Wang, Xiaoyun Li, Belhal Karimi, and Ping Li. An optimistic acceleration of amsgrad for nonconvex optimization. In Vineeth N. Balasubramanian and Ivor Tsang (eds.), *Proceedings of The 13th Asian Conference on Machine Learning*, volume 157 of *Proceedings of Machine Learning Research*, pp. 422–437. PMLR, 17–19 Nov 2021. URL https://proceedings.mlr.press/v157/wang21c.html.

Yeming Wen, Kevin Luk, Maxime Gazeau, Guodong Zhang, Harris Chan, and Jimmy Ba. An empirical study of stochastic gradient descent with structured covariance noise. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3621–3631. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/wen20a.html.

Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as SGD. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10367–10376. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/wu20c.html.

Zeke Xie, Li Yuan, Zhanxing Zhu, and Masashi Sugiyama. Positive-negative momentum: Manipulating stochastic gradient noise to improve generalization. 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. 11448–11458. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/xie21h.html.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7654–7663. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/zhu19e.html.

A MINIBATCH RESIDUAL STATISTICS

A.1 MINIBATCH-RESIDUAL MOMENTS

The main text represents a minibatch gradient as the full-dataset gradient plus a sampling residual. This section defines these quantities precisely and derives the residual mean and covariance under uniform sampling, both with and without replacement. These results specify the one-update statistics that are controlled by our temporal interventions.

Let $g_i(\theta)$ denote the gradient of training example i, and let $B = (i_1, \dots, i_b)$ be a minibatch of size b. The full-training and minibatch gradients are

$$g_{\text{tot}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} g_i(\theta), \qquad g_B(\theta) = \frac{1}{b}\sum_{j=1}^{b} g_{i_j}(\theta). \tag{13}$$

The minibatch residual is then defined as

$$\eta_B(\theta) = g_B(\theta) - g_{\text{tot}}(\theta). \tag{14}$$

We use the empirical per-example gradient covariance

$$\Sigma(\theta) = \frac{1}{N}\sum_{i=1}^{N} \left(g_i(\theta) - g_{\text{tot}}(\theta)\right)\left(g_i(\theta) - g_{\text{tot}}(\theta)\right)^\top. \tag{15}$$

Mean of the residual. Under either uniform sampling with replacement or uniform sampling without replacement, each sampled index has a uniform marginal distribution. Hence,

$$\mathbb{E}_B[g_{i_j}(\theta)] = g_{\text{tot}}(\theta). \tag{16}$$

It follows that

$$\mathbb{E}_B[\eta_B(\theta)] = \frac{1}{b}\sum_{j=1}^{b} \mathbb{E}_B[g_{i_j}(\theta)] - g_{\text{tot}}(\theta) \tag{17}$$

$$= 0. \tag{18}$$

Covariance with replacement. When the indices $i_1, \dots, i_b$ are sampled independently,

$$\text{Cov}_B[\eta_B(\theta)] = \text{Cov}_B\left[\frac{1}{b}\sum_{j=1}^{b} g_{i_j}(\theta)\right] \tag{19}$$

$$= \frac{1}{b^2}\sum_{j=1}^{b} \text{Cov}_B[g_{i_j}(\theta)] \tag{20}$$

$$= \frac{\Sigma(\theta)}{b}. \tag{21}$$

Covariance without replacement. When the minibatch contains b distinct examples sampled uniformly without replacement, two different sampled positions satisfy

$$\text{Cov}_B\left(g_{i_j}(\theta), g_{i_k}(\theta)\right) = -\frac{\Sigma(\theta)}{N-1}, \qquad j \neq k. \tag{22}$$

Consequently,

$$\text{Cov}_B[\eta_B(\theta)] = \frac{1}{b^2}\left[b\Sigma(\theta) - \frac{b(b-1)}{N-1}\Sigma(\theta)\right]$$

$$= \frac{N-b}{N-1}\frac{\Sigma(\theta)}{b}. \tag{23}$$

Therefore, under either sampling scheme,

$$\text{Cov}_B[\eta_B(\theta)] = \kappa_b \frac{\Sigma(\theta)}{b}, \qquad \kappa_b = \begin{cases} 1, & \text{with replacement,} \\ \dfrac{N-b}{N-1}, & \text{without replacement.} \end{cases} \tag{24}$$

A.2 Sampling and state transport in temporal residual covariance

The preceding section characterizes the distribution of a minibatch residual at a single parameter state. During training, consecutive residuals are evaluated at different states because the first stochastic update changes the parameters before the second gradient is computed. Consequently, the observed cross-update covariance combines dependence introduced by the sampling law with deformation of the residual between parameter states.

This section separates these two contributions exactly and bounds the state-transport correction under a local Lipschitz condition. This identifies when the same-state sampling covariance used in the main text provides an accurate local approximation, without assuming that the parameters remain fixed during training.

For two consecutive minibatches B_t and B_{t+1}, define the state-transport correction

$$\Delta^{\mathrm{tr}}\eta_{t+1} = \eta_{B_{t+1}}(\theta_{t+1}) - \eta_{B_{t+1}}(\theta_t). \tag{25}$$

Adding and subtracting $\eta_{B_{t+1}}(\theta_t)$ gives the exact decomposition

$$\begin{aligned}&\mathrm{Cov}\big(\eta_{B_t}(\theta_t), \eta_{B_{t+1}}(\theta_{t+1})\big)\\ &\quad= \underbrace{\mathrm{Cov}\big(\eta_{B_t}(\theta_t), \eta_{B_{t+1}}(\theta_t)\big)}_{\text{sampling contribution}} + \underbrace{\mathrm{Cov}\big(\eta_{B_t}(\theta_t), \Delta^{\mathrm{tr}}\eta_{t+1}\big)}_{\text{state-transport correction}}.\end{aligned} \tag{26}$$

State-transport correction. Assume that every minibatch residual is locally L_η-Lipschitz over the region traversed by one optimizer update:

$$\|\eta_B(\theta') - \eta_B(\theta)\| \le L_\eta \|\theta' - \theta\|. \tag{27}$$

Equation 25 then implies

$$\left\|\Delta^{\mathrm{tr}}\eta_{t+1}\right\| \le L_\eta \|\theta_{t+1} - \theta_t\|. \tag{28}$$

Using Cauchy–Schwarz,

$$\begin{aligned}&\left\|\mathrm{Cov}\big(\eta_{B_t}(\theta_t), \Delta^{\mathrm{tr}}\eta_{t+1}\big)\right\|_F\\ &\quad\le L_\eta \sqrt{\mathbb{E}\Big[\|\eta_{B_t}(\theta_t) - \mathbb{E}[\eta_{B_t}(\theta_t)]\|^2\Big]}\sqrt{\mathbb{E}\Big[\|\theta_{t+1} - \theta_t\|^2\Big]}.\end{aligned} \tag{29}$$

The state-transport correction is therefore controlled by the magnitude of the optimizer update and the local sensitivity of the residual geometry. When this correction is small relative to the sampling contribution, the first term in Equation 26 dominates. Its value under random reshuffling is derived in Appendix A.3, yielding Equation 3.

A.3 Residual autocorrelation under random reshuffling

This section derives the adjacent residual covariance used in Equation 3. We consider a fixed empirical population of $N = Mb$ examples, partitioned uniformly at random into M disjoint minibatches of size b. The calculation is performed at a fixed parameter state θ; its application to consecutive parameter states uses the local approximation in Equation 2.

Let $B_1, \dots, B_M$ be the minibatches of one randomly shuffled epoch and let

$$\eta_j = g_{B_j}(\theta) - g_{\mathrm{tot}}(\theta). \tag{30}$$

Because every example appears exactly once during the epoch,

$$\sum_{j=1}^{M} g_{B_j}(\theta) = M g_{\mathrm{tot}}(\theta), \tag{31}$$

and therefore

$$\sum_{j=1}^{M} \eta_j = 0. \tag{32}$$

The residuals are exchangeable under a uniformly random partition. Define their common marginal covariance as

$$C_b(\theta) = \mathrm{Cov}(\eta_j). \tag{33}$$

For two distinct positions $j \neq k$, exchangeability implies that their cross-covariance has a common value $D_b(\theta)$. Taking the covariance of η_j with the zero sum in Equation 32 gives

$$0 = \mathrm{Cov}\left(\eta_j, \sum_{k=1}^{M} \eta_k\right) \tag{34}$$

$$= C_b(\theta) + (M-1)D_b(\theta). \tag{35}$$

Hence,

$$\mathrm{Cov}(\eta_j, \eta_k) = -\frac{C_b(\theta)}{M-1}, \qquad j \neq k. \tag{36}$$

Now consider a stationary stream of independently reshuffled epochs. Two consecutive updates belong to the same epoch with probability $(M-1)/M$. With probability $1/M$, they cross an epoch boundary and belong to independent random partitions, giving zero cross-covariance. Therefore,

$$\mathrm{Cov}(\eta_t, \eta_{t+1}) = \frac{M-1}{M}\left(-\frac{C_b(\theta)}{M-1}\right) \tag{37}$$

$$= -\frac{C_b(\theta)}{M}. \tag{38}$$

Since $C_b(\theta) = \mathrm{Cov}_B[\eta_B(\theta)]$, the normalized adjacent correlation is

$$\rho_1 = -\frac{1}{M} = -\frac{b}{N}. \tag{39}$$

More generally, two updates separated by h positions belong to the same epoch with probability $(M - |h|)/M$. The fixed-state autocovariance is therefore

$$\mathrm{Cov}(\eta_t, \eta_{t+h}) = \begin{cases} C_b(\theta), & h = 0, \\ -\dfrac{M-|h|}{M(M-1)} C_b(\theta), & 1 \leq |h| < M, \\ 0, & |h| \geq M. \end{cases} \tag{40}$$

Random reshuffling thus distributes negative dependence across an epoch, with the residuals cancelling exactly when accumulated over all M minibatches.

During optimization, consecutive residuals are evaluated at different parameter states. Under $\eta_B(\theta_{t+1}) \approx \eta_B(\theta_t)$, the fixed-state result gives

$$\mathrm{Cov}\left(\eta_{B_t}(\theta_t), \eta_{B_{t+1}}(\theta_{t+1})\right) \approx -\frac{1}{M}\,\mathrm{Cov}_B[\eta_B(\theta_t)], \tag{41}$$

which is Equation 3. The state-transport correction neglected by this approximation is discussed in Appendix A.2.

This derivation isolates randomness due to index sampling. If training also uses independently sampled data augmentation, augmentation introduces an additional stochastic component that need not satisfy the epoch-wise zero-sum identity and can weaken the normalized anticorrelation.

A.4 Temporal residual correlation in Adam's moment states

This section derives the statements of Section 3.2.2. Throughout, we condition on $\mathcal{F}_{2t-1}$, the history immediately before update $2t$, and use the local approximation of Section 3.1, so that both updates of the pair see the full gradient $g = g_{\mathrm{tot}}(\theta_{2t})$ and residuals with covariance $C = C(\theta_{2t})$. We write $c = \mathrm{diag}\, C$; all products and squares of vectors are coordinate-wise.

A.4.1 FIRST MOMENT

Unrolling the recursion over a pair. With $g_s = g + \eta_s$ for $s \in \{2t, 2t+1\}$, applying the first-moment recursion twice gives

$$\begin{aligned} m_{2t} &= \beta_1 m_{2t-1} + (1-\beta_1)(g+\eta_{2t}), \\ m_{2t+1} &= \beta_1 m_{2t} + (1-\beta_1)(g+\eta_{2t+1}) \\ &= \beta_1^2 m_{2t-1} + (1-\beta_1)(1+\beta_1)\, g + (1-\beta_1)\left(\beta_1\eta_{2t} + \eta_{2t+1}\right). \end{aligned} \tag{42}$$

Substituting $m_{2t-1} = \bar{m}_{2t-1} + \delta m_{2t-1}$ and separating the terms driven by g from those driven by the residuals yields

$$\bar{m}_{2t+1} = \beta_1^2 \bar{m}_{2t-1} + (1-\beta_1)(1+\beta_1)\, g, \qquad \delta m_{2t+1} = \beta_1^2 \delta m_{2t-1} + (1-\beta_1)\left(\beta_1\eta_{2t} + \eta_{2t+1}\right), \tag{43}$$

Given $\mathcal{F}_{2t-1}$, the quantities $\bar{m}_{2t-1}$, δm_{2t-1}, and g are fixed, and the new residuals have zero mean. Hence $\mathbb{E}[m_{2t+1} \mid \mathcal{F}_{2t-1}] = \bar{m}_{2t+1} + \beta_1^2 \delta m_{2t-1}$ for any temporal coupling.

Conditional covariance. Only the term $\xi = \beta_1\eta_{2t} + \eta_{2t+1}$ is random given $\mathcal{F}_{2t-1}$, so $\mathrm{Cov}(m_{2t+1} \mid \mathcal{F}_{2t-1}) = (1-\beta_1)^2 \mathrm{Cov}(\xi)$. By bilinearity,

$$\mathrm{Cov}(\xi) = \beta_1^2 \mathrm{Cov}(\eta_{2t}) + \mathrm{Cov}(\eta_{2t+1}) + \beta_1 \left[\mathrm{Cov}(\eta_{2t}, \eta_{2t+1}) + \mathrm{Cov}(\eta_{2t+1}, \eta_{2t})\right]. \tag{44}$$

Both residuals have covariance C. If $\mathrm{Cov}(\eta_{2t}, \eta_{2t+1}) = \rho\, C$, the symmetry of C gives

$$\mathrm{Cov}(m_{2t+1} \mid \mathcal{F}_{2t-1}) = (1-\beta_1)^2 \left(1 + \beta_1^2 + 2\beta_1\rho\right) C. \tag{45}$$

The iid pair has $\rho = 0$, giving $(1-\beta_1)^2(1+\beta_1^2)C$. The antithetic pair has $\rho = -1$, and since $1 + \beta_1^2 - 2\beta_1 = (1-\beta_1)^2$, it gives $(1-\beta_1)^4 C$. Their ratio is $(1-\beta_1)^2/(1+\beta_1^2)$, Equation 7. The same formula covers the other couplings in the paper: random reshuffling locally has $\rho \approx -1/M$ (Equation 3), and the superbatch constructions have $\rho = -(K-2)/K$.

Longer horizons. Unrolling the recursion from $\delta m_0 = 0$ gives $\delta m_t = (1-\beta_1)\sum_{k=1}^{t} \beta_1^{t-k}\eta_k$, so

$$\mathrm{Cov}(\delta m_t) = (1-\beta_1)^2 \sum_{i=1}^{t}\sum_{j=1}^{t} \beta_1^{2t-i-j}\, \mathrm{Cov}(\eta_i, \eta_j). \tag{46}$$

When the symmetrized cross-covariances $\mathrm{Cov}(\eta_i, \eta_j) + \mathrm{Cov}(\eta_j, \eta_i)$ are negative semidefinite, as under random reshuffling, their positive Adam weights reduce the first-moment covariance. Adjacent correlations receive the largest weights, which is why antithesis concentrates the cancellation on consecutive updates.

Validity of the local approximation. Without the local approximation, the second update sees $g_{\mathrm{tot}}(\theta_{2t+1})$ rather than g. Writing $\theta_{2t+1} = \theta_{2t} - \alpha u_{2t}$ with $u_{2t} \approx D^{-1} m_{2t}$ and $D = \mathrm{diag}(\sqrt{v_{2t}} + \epsilon)$, and neglecting the dependence of v_{2t} on η_{2t}, which carries the small weight $1-\beta_2$, a first-order expansion gives

$$g_{\mathrm{tot}}(\theta_{2t+1}) \approx g_{\mathrm{tot}}(\theta_{2t}) - \alpha H D^{-1} m_{2t} = (\text{terms fixed by } \mathcal{F}_{2t-1}) - \alpha(1-\beta_1) H D^{-1}\eta_{2t}, \tag{47}$$

where H is the Hessian of the training loss at θ_{2t}. The residual-dependent part of m_{2t+1} therefore becomes $(1-\beta_1)\left[(\beta_1 I - (1-\beta_1)A)\eta_{2t} + \eta_{2t+1}\right]$, with $A = \alpha H D^{-1}$. For the antithetic pair, the coefficient of η_{2t} becomes

$$(\beta_1 - 1)I - (1-\beta_1)A = -(1-\beta_1)\left(I + A\right). \tag{48}$$

The closed-form ratio thus holds to leading order when $\alpha\|HD^{-1}\| \ll 1$. Along directions of large preconditioned curvature, the intervening parameter update partly restores the cancelled residual. This correction is one route by which antithesis can still act when the first moment has no memory, as observed at $\beta_1 = 0$ in Section 4.2. The transport of the residual itself between the two states adds a further correction of the same order, bounded in Appendix A.2.

A.4.2 SECOND MOMENT

Unrolling the recursion over a pair. Applying the second-moment recursion twice gives

$$v_{2t+1} = \beta_2^2 v_{2t-1} + (1-\beta_2)\left(\beta_2 g_{2t}^2 + g_{2t+1}^2\right). \tag{49}$$

Each squared gradient decomposes into its expectation and a zero-mean part,

$$g_s^2 = (g + \eta_s)^2 = \left(g^2 + c\right) + \left(2g\,\eta_s + \eta_s^2 - c\right). \tag{50}$$

As for the first moment, we split $v_t = \bar{v}_t + \delta v_t$ into the part driven by $g^2 + c$ and the part driven by the zero-mean terms. Substituting $v_{2t-1} = \bar{v}_{2t-1} + \delta v_{2t-1}$ and separating the two parts yields

$$\bar{v}_{2t+1} = \beta_2^2 \bar{v}_{2t-1} + (1-\beta_2)(1+\beta_2)\left(g^2 + c\right), \qquad \delta v_{2t+1} = \beta_2^2 \delta v_{2t-1} + (1-\beta_2)X, \tag{51}$$

with

$$X = 2g\left(\beta_2 \eta_{2t} + \eta_{2t+1}\right) + \beta_2\left(\eta_{2t}^2 - c\right) + \left(\eta_{2t+1}^2 - c\right). \tag{52}$$

Expectation. The deterministic component depends on the residuals only through c. Antithesis preserves the one-step covariance, so c, and hence $\bar{v}_{2t+1}$, is the same for both constructions. Since $\mathbb{E}[\eta_s] = 0$ and $\mathbb{E}[\eta_s^2] = c$, we have $\mathbb{E}[X] = 0$, so both constructions have the same conditional expected second moment. Given $\mathcal{F}_{2t-1}$, only X is random, so for each coordinate $\mathrm{Var}(v_{2t+1,i} \mid \mathcal{F}_{2t-1}) = (1-\beta_2)^2 \,\mathrm{Var}(X_i)$.

A variance identity. Fix a coordinate, drop its index, and let η have $\mathbb{E}[\eta] = 0$, $\mathbb{E}[\eta^2] = c$, $\mu_3 = \mathbb{E}[\eta^3]$, and $V_4 = \mathrm{Var}(\eta^2) = \mathbb{E}[\eta^4] - c^2$. For constants a and b,

$$\mathrm{Var}\left(a\eta + b(\eta^2 - c)\right) = a^2 c + b^2 V_4 + 2ab\,\mu_3, \tag{53}$$

because $\mathrm{Cov}(\eta, \eta^2 - c) = \mathbb{E}[\eta^3] = \mu_3$.

Independent pair. When η_{2t} and η_{2t+1} are independent copies of η, X is the sum of two independent terms of the form in Equation 53, with $(a, b) = (2g\beta_2, \beta_2)$ and $(a, b) = (2g, 1)$. Adding their variances gives

$$\mathrm{Var}\left(X^{\mathrm{iid}}\right) = (1+\beta_2^2)\left(4g^2 c + V_4 + 4g\mu_3\right). \tag{54}$$

Antithetic pair. When $\eta_{2t+1} = -\eta_{2t} = -\eta$, the linear terms combine to $2g(\beta_2 - 1)\eta$, whereas the squared terms coincide, so $X^{\mathrm{anti}} = -2g(1-\beta_2)\eta + (1+\beta_2)(\eta^2 - c)$. Equation 53 with $(a, b) = (-2g(1-\beta_2), 1+\beta_2)$ gives

$$\mathrm{Var}\left(X^{\mathrm{anti}}\right) = 4g^2 c\,(1-\beta_2)^2 + (1+\beta_2)^2 V_4 - 4g\,(1-\beta_2^2)\,\mu_3. \tag{55}$$

Antithesis thus nearly cancels the linear term, while the two squared residuals add coherently.

Residuals with zero third moment. When $\mu_3 = 0$, dividing Equation 55 by Equation 54 and writing $r = 4g^2 c / V_4$ gives Equation 8. The difference between its numerator and denominator is

$$(1-\beta_2)^2 r + (1+\beta_2)^2 - (1+\beta_2^2)(1+r) = 2\beta_2(1-r), \tag{56}$$

so the ratio is below one exactly when $r > 1$. It tends to $(1+\beta_2)^2/(1+\beta_2^2)$ as $r \to 0$ and to $(1-\beta_2)^2/(1+\beta_2^2)$ as $r \to \infty$. For Gaussian residuals, $\mathbb{E}[\eta^4] = 3c^2$, so $V_4 = 2c^2$, $r = 2g^2/c$, and the condition $r > 1$ becomes $|g| > \sqrt{c/2}$.

Size of the third moment. With replacement, a residual coordinate is the average of b independent per-example deviations with variance σ^2 and third central moment κ_3. Variances and third central moments both add over independent terms, so the average has variance σ^2/b and third moment κ_3/b^2, hence skewness

$$\frac{\kappa_3/b^2}{(\sigma^2/b)^{3/2}} = \frac{1}{\sqrt{b}}\,\frac{\kappa_3}{\sigma^3}. \tag{57}$$

Sampling without replacement changes this only through the finite-population correction. The symmetric case therefore describes minibatch residuals approximately rather than exactly; Equations 54 and 55 hold without this assumption.

A.4.3 Complete Adam direction

Adam applies the bias corrections $\widehat{m}_t = m_t/(1-\beta_1^t)$ and $\widehat{v}_t = v_t/(1-\beta_2^t)$. These are deterministic scalings shared by both constructions, so they leave all the ratios above unchanged. The update direction $u_t = \widehat{m}_t/(\sqrt{\widehat{v}_t} + \epsilon)$ is, however, a nonlinear function of both states, so its behavior cannot be inferred from either state alone. Section 4.2 therefore separates the contributions of m, v, and their coupling through controlled interventions, and finds that the first moment carries the effect.

A.5 SUPERBATCH CONSTRUCTIONS

This section derives the statements of Section 3.2.3. Under the local approximation, we evaluate all gradients of a pair at a common state and write $g_{B_j} = g + \eta_j$ for $j \in \{1, \dots, K\}$, with $g = g_{\mathrm{tot}}$ and residuals η_j that are independent, zero-mean, and have covariance C. Sampling the K minibatches without replacement is treated in Appendix A.5.4.

A.5.1 DIRECT SUPERBATCH REFLECTION

Residual of the reflected input. Since $\bar{g}_K = g + \frac{1}{K}\sum_{j=1}^{K} \eta_j$, the second input of Equation 9 is

$$2\bar{g}_K - g_{B_1} = g + \frac{2}{K}\sum_{j=1}^{K} \eta_j - \eta_1 = g - \frac{K-2}{K}\,\eta_1 + \frac{2}{K}\sum_{j=2}^{K} \eta_j, \tag{58}$$

while the first input $g_{B_1} = g + \eta_1$ carries the residual η_1.

Marginal statistics. The reflected input has mean g. By independence of the residuals, its covariance is

$$\left(\frac{K-2}{K}\right)^2 C + \frac{4(K-1)}{K^2}\,C = \frac{(K-2)^2 + 4(K-1)}{K^2}\,C = C, \tag{59}$$

that of an ordinary minibatch gradient.

Correlation. Only the term in η_1 is correlated with the first input, so $\mathrm{Cov}(g_{B_1}, 2\bar{g}_K - g_{B_1}) = -\frac{K-2}{K}\,C$. Since both inputs have covariance C, their correlation is $\rho_{\mathrm{anti}} = -(K-2)/K$, Equation 10.

Distance to exact reflection. Exact reflection would present the residual $-\eta_1$ at the second update. The direct construction deviates from it by

$$\Delta_K = \frac{2}{K}\sum_{j=1}^{K} \eta_j, \qquad \mathbb{E}[\Delta_K] = 0, \qquad \mathrm{Cov}(\Delta_K) = \frac{4}{K}\,C, \tag{60}$$

so it converges to exact reflection in mean square as K increases.

Cost. The first input requires only B_1, and the second requires all K minibatches at the new state, for $K + 1$ evaluations per pair.

A.5.2 DENSE SIGNED-CONTRAST CONSTRUCTION

Residual of a signed-contrast input. For $z \in \mathbb{R}^K$ with $\sum_j z_j = 0$ and $\frac{1}{K}\sum_j z_j^2 = 1$, define $q(z) = \bar{g}_K + \frac{\sqrt{K-1}}{K}\sum_{j=1}^{K} z_j\, g_{B_j}$. Since $\sum_j z_j\, g = 0$,

$$q(z) = g + \frac{1}{K}\sum_{j=1}^{K} a_j \eta_j, \qquad a_j = 1 + \sqrt{K-1}\, z_j, \tag{61}$$

with $\sum_j a_j = K$ and $\sum_j a_j^2 = K + 2\sqrt{K-1}\sum_j z_j + (K-1)\sum_j z_j^2 = K^2$.

Marginal statistics. Hence $\mathbb{E}[q(z)] = g$ and $\mathrm{Cov}(q(z)) = \frac{1}{K^2}\sum_j a_j^2\, C = C$.

Correlation. For two contrasts z and w, with $b_j = 1 + \sqrt{K-1}\, w_j$ and $\sum_j a_j b_j = K + (K-1)\sum_j z_j w_j$,

$$\mathrm{Cov}(q(z), q(w)) = \frac{1}{K^2}\sum_{j=1}^{K} a_j b_j\, C = \frac{1 + (K-1)\langle z, w\rangle_K}{K}\,C, \tag{62}$$

which gives Equation 12. Taking $w = -z$ gives $\langle z, w\rangle_K = -1$ and $\rho_{\mathrm{anti}} = -(K-2)/K$.

Relation to the direct reflection. The contrast $z^{(1)} = \left(\sqrt{K-1}, -\frac{1}{\sqrt{K-1}}, \dots, -\frac{1}{\sqrt{K-1}}\right)$ satisfies both constraints and gives coefficients a_j/K equal to 1 for $j = 1$ and 0 otherwise, so $q(z^{(1)}) = g_{B_1}$.

Its opposite gives coefficients $-(K-2)/K$ and $2/K$, so $q(-z^{(1)}) = 2\bar{g}_K - g_{B_1}$. The direct reflection is therefore the dense construction with contrast $z^{(1)}$, consistent with Appendix A.5.1.

Cost. The dense construction evaluates all K minibatches at both states, for $2K$ evaluations per pair.

A.5.3 Zero-correlation control

Let $u \in \mathbb{R}^K$ satisfy $\sum_j u_j = 0$, $\langle u, z\rangle_K = 0$, and $\frac{1}{K}\sum_j u_j^2 = 1$, which exists for $K \geq 3$, and set

$$w_0 = -\frac{1}{K-1}\, z + \sqrt{1 - \frac{1}{(K-1)^2}}\; u. \tag{63}$$

Then $\sum_j (w_0)_j = 0$, $\frac{1}{K}\sum_j (w_0)_j^2 = 1$, and $\langle z, w_0\rangle_K = -1/(K-1)$, so by Appendix A.5.2 the cross-update covariance vanishes. The control shares the first input, the minibatches, the gradient-evaluation cost, and the one-step mean and covariance of the antithetic construction, and differs from it only in the temporal coupling. The direct control uses $z = z^{(1)}$. For $K = 2$, the antithetic construction itself has zero correlation.

A.5.4 Sampling without replacement

If the K minibatches are drawn without replacement from the same epoch, distinct residuals have covariance $-C/(M-1)$ (Equation 36). With the coefficients of Equation 61, which also cover the direct construction, the input covariance becomes

$$\frac{1}{K^2}\left[\sum_j a_j^2 - \frac{(\sum_j a_j)^2 - \sum_j a_j^2}{M-1}\right] C = C, \tag{64}$$

unchanged because $(\sum_j a_j)^2 = \sum_j a_j^2 = K^2$. The correlation between the two inputs becomes

$$\rho - \frac{1-\rho}{M-1}, \tag{65}$$

a correction of order $1/M$. For the zero-correlation control, this recovers the natural adjacent correlation $-1/(M-1)$ of random reshuffling.

B Additional experimental results

B.1 The Jensen gap

Let θ denote the end point of a pair of updates from a common state, with mean $\mu = \mathbb{E}[\theta]$, and let H be the Hessian of the loss L at μ. Expanding L to second order around μ,

$$L(\theta) \approx L(\mu) + \nabla L(\mu)^\top(\theta - \mu) + \tfrac{1}{2}(\theta-\mu)^\top H(\theta-\mu). \tag{66}$$

Taking the expectation, the linear term vanishes because $\mathbb{E}[\theta - \mu] = 0$, and $\mathbb{E}[(\theta-\mu)^\top H(\theta-\mu)] = \operatorname{tr}\big(H\,\mathbb{E}[(\theta-\mu)(\theta-\mu)^\top]\big) = \operatorname{tr}(H\,\mathrm{Cov}(\theta))$. Hence

$$J = \mathbb{E}[L(\theta)] - L(\mathbb{E}[\theta]) \approx \tfrac{1}{2}\operatorname{tr}(H\,\mathrm{Cov}(\theta))\,. \tag{67}$$

The Jensen gap is thus the spread of the end points weighted by the curvature in each direction. It is positive along directions where the loss is locally convex, and reducing the spread without moving the mean lowers the expected loss.

Figure 6 illustrates this. After a pair of updates from a common state, the end points scatter around a mean that is nearly the same for the *iid* and *antithetic* constructions, but antithesis scatters much less (panel a). Along a curved direction, the loss averaged over the scattered end points, $\mathbb{E}[L(\theta)]$, lies above the loss at their average, $L(\mathbb{E}[\theta])$ (panel b), and the gap shrinks with the scatter. The figure uses a synthetic quadratic loss and Gaussian end points; only the ratio of the two scatters is taken from the experiment, matching the measured reduction of J on CIFAR-10.

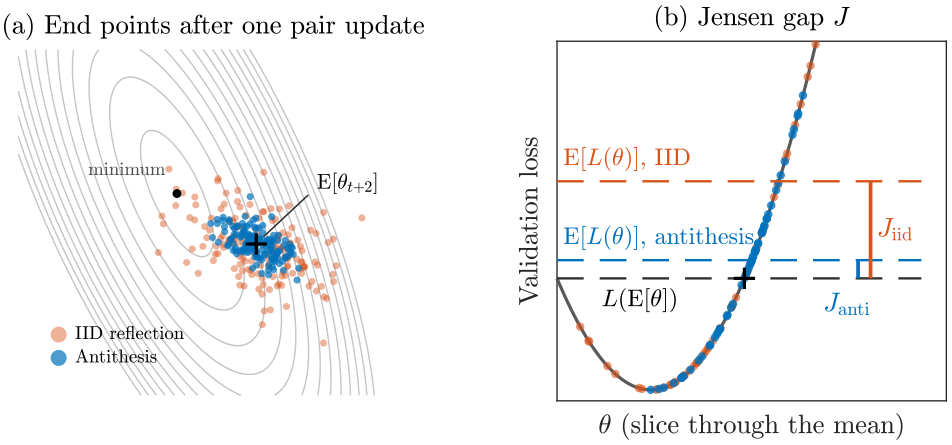

Figure 6: Illustration of the Jensen gap (synthetic). (a) End points after one pair update for *iid* reflection and antithesis, around a common mean, on the contours of a quadratic loss. (b) Along a one-dimensional slice, the average loss $\mathbb{E}[L(\theta)]$ exceeds the loss at the average point $L(\mathbb{E}[\theta])$ by the Jensen gap J, which is smaller for antithesis.

B.2 Superbatch results for all K

Table 5 reports the test accuracy of both superbatch constructions and their zero-correlation controls for every superbatch size, with AdamW and exact antithesis as references.

Table 5: CIFAR-10 test accuracy of the direct and dense superbatch constructions and their zero-correlation controls. Cost is the number of minibatch gradients per update relative to AdamW. Each control matches its antithetic construction in minibatches, one-step mean and covariance, optimizer steps and cost. Mean ± standard error over three seeds.

			Independent schedules		20-epoch schedule			
Construction	Method	Cost	2/2	5/5	5/20	10/20	15/20	20/20
–	AdamW	×1	75.36 ± 0.35	85.68 ± 0.06	79.05 ± 0.69	85.96 ± 0.06	89.42 ± 0.14	92.08 ± 0.11
Direct	$K = 4$, $\rho = 0$	×2.5	75.24 ± 0.36	85.49 ± 0.02	80.59 ± 0.52	85.21 ± 0.59	89.70 ± 0.14	92.07 ± 0.13
	$K = 4$, antithesis	×2.5	**81.79 ± 0.15**	**89.01 ± 0.12**	**84.49 ± 1.08**	**88.16 ± 0.61**	**90.87 ± 0.32**	**92.79 ± 0.04**
	$K = 8$, $\rho = 0$	×4.5	75.73 ± 0.43	85.57 ± 0.06	80.83 ± 0.34	85.49 ± 0.29	89.72 ± 0.14	91.82 ± 0.08
	$K = 8$, antithesis	×4.5	**85.00 ± 0.14**	**90.49 ± 0.13**	**85.43 ± 0.59**	**86.96 ± 1.12**	**91.29 ± 0.17**	**92.89 ± 0.10**
	$K = 16$, $\rho = 0$	×8.5	75.66 ± 0.16	85.74 ± 0.13	80.48 ± 0.83	85.29 ± 0.53	89.39 ± 0.24	92.04 ± 0.11
	$K = 16$, antithesis	×8.5	**86.38 ± 0.18**	**91.22 ± 0.02**	**85.60 ± 1.44**	**88.58 ± 0.61**	**91.60 ± 0.28**	**92.95 ± 0.09**
	$K = 32$, $\rho = 0$	×16.5	75.63 ± 0.59	85.66 ± 0.13	79.50 ± 1.09	84.74 ± 0.14	89.40 ± 0.27	92.07 ± 0.04
	$K = 32$, antithesis	×16.5	**87.51 ± 0.08**	**91.61 ± 0.03**	**86.34 ± 0.85**	**88.52 ± 1.16**	**90.89 ± 0.37**	**92.92 ± 0.02**
Dense	$K = 4$, $\rho = 0$	×4	75.68 ± 0.17	85.54 ± 0.11	76.54 ± 3.32	84.99 ± 0.65	89.47 ± 0.28	91.90 ± 0.08
	$K = 4$, antithesis	×4	**81.68 ± 0.21**	**88.95 ± 0.05**	**83.59 ± 0.98**	**87.60 ± 0.63**	**91.01 ± 0.41**	**92.85 ± 0.07**
	$K = 8$, $\rho = 0$	×8	75.73 ± 0.18	85.63 ± 0.17	77.13 ± 0.77	85.13 ± 0.81	88.95 ± 0.12	92.06 ± 0.05
	$K = 8$, antithesis	×8	**84.74 ± 0.09**	**90.53 ± 0.05**	**87.30 ± 0.58**	**89.20 ± 0.23**	**91.75 ± 0.14**	**93.24 ± 0.07**
	$K = 16$, $\rho = 0$	×16	75.27 ± 0.37	85.79 ± 0.20	78.12 ± 1.26	84.65 ± 0.75	89.34 ± 0.09	92.00 ± 0.09
	$K = 16$, antithesis	×16	**86.32 ± 0.02**	**91.14 ± 0.09**	**88.03 ± 0.31**	**90.51 ± 0.55**	**91.83 ± 0.13**	**93.11 ± 0.09**
	$K = 32$, $\rho = 0$	×32	75.59 ± 0.42	85.86 ± 0.27	77.96 ± 0.54	85.26 ± 0.37	88.89 ± 0.06	91.91 ± 0.16
	$K = 32$, antithesis	×32	**87.53 ± 0.15**	**91.64 ± 0.02**	**89.20 ± 0.35**	**91.21 ± 0.11**	**92.02 ± 0.17**	**93.23 ± 0.06**
–	Exact antithesis	×176	**88.07 ± 0.13**	**91.50 ± 0.12**	**86.35 ± 1.00**	**90.20 ± 0.58**	**91.66 ± 0.49**	**93.13 ± 0.17**