We thank the reviewers for their time, effort, and helpful feedback. We address individual comments below.

Reviewer 1: *Time homogeneity.* The *training* loss in all of our examples can be written in the form of lines 17-18. When each data batch ξ is sampled uniformly with replacement, the time homogeneity follows from the form of the update dynamics, since α and β in Eqn. (1) are constant while we collect samples and do statistical tests. The fact that the iterates of SGD form a homogeneous Markov chain is also used, for example, by Bach and Moulines (2013) and Dieuleveut et al. (2017). We will include these additional references and add a proof to the appendix. While we sample batches *without* replacement in the experiments, such practice is common in deep learning and is arguably a small gap.

Statistics in loss space. Consider equation (6) when the assumptions of Section 2.1 apply, i.e., $F(x)$ is a quadratic function and the additive noise to the gradient is independent of x. Then the left hand side of (6) is $\mathbf{E}_\pi[x^T Ax]$, the mean value of the loss at stationarity, but the right-hand side is $\frac{\alpha}{2}\mathbf{tr}(\Sigma)$, the (scaled) trace of the noise covariance. In this case, we test whether the mean loss has converged to a constant, but we *also* have a different estimator for that constant. This test should be more sample-efficient than one that compares the mean loss to itself if the other estimator converges quickly. Our test can be considered as a more general version of testing whether the loss has reached a constant value.

Hyperparameter δ. In our numerical experiments, we found that $\delta = 0.02$ worked well on all the examples we studied, and Appendix A contains a study of the sensitivity of the algorithm to changes of δ around this value. Intuitively, because the term that δ is multiplying is sensitive to the scale of the statistics, it is reasonable to expect roughly the same value of δ to work on different problems. In some sense, any method for testing stationarity must include a "slack term."

Reviewer 2: *Results with (9).* In addition to Figure 4, results for tuning using equation (9) are provided in Figures 12 and 13 in Appendix B. In short, (9) performs similarly to (10) on average, but has (potentially much) higher variance.

Sensitivity to significance parameter. The middle rows of Figures 6, 8, and 10 in Appendix A show the sensitivity of Algorithm 2 to changes in γ around the default value of 0.2 on CIFAR-10, ImageNet, and MNIST, respectively.

Hand-tuning Adam. In the CIFAR-10 and ImageNet experiments, we used a hand-tuned "warmup phase" for Adam. The learning rates are not plotted here because they changed per parameter after the warmup phase. In the RNN example, the global learning rate of Adam is dropped based on the validation loss, and is simply missing from the bottom-right panel of Figure 2. We will add this global learning rate curve upon revision. It is similar to the one for SGM. Wilson et al. (2017, Section 4.2) observed that step-wise decay of Adam's global learning rate did not improve their results on CIFAR-10, so we only tuned the warmup phase for our image experiments.

SASA for Adam. Unfortunately, unlike SGM with fixed values of α and β, the dynamics of Adam depend heavily on time. Adam converges to a stationary point rather than to a stationary distribution with nonzero variance. This makes the SASA approach of testing for stationarity inapplicable to Adam without significant modification.

Figure 4. The five curves plotted are the performance across five independent runs. The y axes are equalized throughout the figure, and in the (1,2) panel only one of the five curves is on the same scale as the others because of the variance of testing with (9). This is described in the main text, and we will update the Figure 4 caption to match the main text.

Reviewer 3: *Statistical testing.* The null hypothesis is that $|\mathbf{E}_\pi[\langle x, g\rangle] - \frac{\alpha}{2}\frac{1+\beta}{1-\beta}\mathbf{E}_\pi[\langle d, d\rangle]| \geq \Delta$; the alternative is that $|\mathbf{E}_\pi[\langle x, g\rangle] - \frac{\alpha}{2}\frac{1+\beta}{1-\beta}\mathbf{E}_\pi[\langle d, d\rangle]| < \Delta$. That is, the alternative is that the equation (6) holds up to a slack of Δ. This is known as *equivalence testing* (Streiner, 2003). We have a relative threshold $\Delta = \delta\frac{\alpha}{2}\frac{1+\beta}{1-\beta}\mathbf{E}_\pi[\langle d, d\rangle]$ with the hyperparameter δ to make the threshold adaptive to the scale of the statistics. We can clarify the presentation of the test by including a brief overview of equivalence testing and by more explicitly stating the hypotheses. Intuitively, the null is "not stationary" and the alternative is "stationary," but with the caveats that we can only test equation (6) up to a slack term Δ, and that equation (6) is merely necessary for stationarity, not sufficient.

Interpreting Yaida's relation. We gave some intuition on the condition in our "loss space" response to Reviewer 1 for quadratic F, which we can add after (6). By "general functions F" we mean any function of the form given in Section 1 such that the other assumptions (SGM$\to$stationary distribution) apply. Understanding what these assumptions imply about F seems to be quite challenging in the general nonconvex case, but the convergence of the statistics in Figures 5, 7, and 9 of the appendix suggests that they can hold in practice even for complicated, nonsmooth functions.

Biased estimators. It is unclear if the precise test used in this paper works when the gradient estimator is biased. Passing from Yaida's original formula (lines 121-122) to equation (6) requires unbiasedness. However, the general procedure of testing for stationarity still applies—the bias simply must be accounted for. We are pursuing some follow-up work to find more general stationary relations, but unbiased estimators remain the most common type in practice.

Small datasets. Note that the sample size N is adaptive, but the *test frequency* M may need to be small for a small dataset. Approach (10) shines compared to (9) when there is large noise (due e.g. to a small M). When the variance of the statistics is high, not accounting for it can cause huge variance in the learning rate schedule, as in Figure 4.