

A Limitations

linear stability does not explain noise injections.

Lastly, our analysis reveals a limitation of optimization dynamics in settings where the data exhibits high redundancy or strongly correlated features. In such cases, the linear stability criteria become less sensitive to solution complexity, as many parameter configurations yield similar activation patterns and induce comparable coherence measures. This results in a potential blind spot: the optimization algorithm may fail to distinguish between simple and complex solutions if the underlying data geometry does not sufficiently break symmetry. Recognizing this limitation offers a valuable direction for future theoretical work, particularly in understanding how optimization behavior is shaped not only by the loss landscape but also by the structure and diversity of the training data.

B Related Work

Flat vs. sharp minima and generalization. The connection between the geometry of minima and generalization in deep networks has been studied extensively. Hochreiter and Schmidhuber [1997] first argued that flat minima (regions in parameter space where the loss remains low) correspond to better generalization, while sharp minima might lead to worse generalization. Keskar et al. [2017] provided empirical evidence that large-batch SGD tends to find sharper minima than small-batch SGD, correlating with higher test error, bringing this idea to prominence. However, Dinh et al. [2017] pointed out that the sharpness of a minimum is not an invariant property (reparameterizations of the model can change the Hessian spectrum without affecting generalization), cautioning that one must carefully define “sharpness” (e.g., by normalizing for scale or using local subspace measures). Our work incorporates this perspective by focusing on a *relative* stability analysis: effectively, we look at sharpness in the context of the optimizer’s step size and algorithm, which is invariant to certain rescaling (for example, SAM’s notion of sharpness implicitly accounts for parameter scale through the perturbation magnitude).

Sharpness-Aware Minimization [Foret et al., 2021] and follow-up methods (e.g., adaptive SAM by Kwon et al., 2021, investigations in Chen and Flammarion, 2022) directly encode flatness into the training objective. By explicitly favoring flatter minima, SAM biases the training trajectory toward solutions that are less sensitive to perturbations in parameter space [Zhang et al., 2024, Chen et al., 2023a]. Empirically, SAM has demonstrated improved generalization across many tasks. The work of Andriushchenko et al. [2023] is particularly relevant to our findings: they show that SAM not only finds flatter minima but that the learned features (e.g., the covariance of layer activations) tend to be lower rank, suggesting the model focuses on a smaller set of principal components of the data. This aligns with our result that SAM bias can lead to simpler (more coherent) feature usage. There have also been studies connecting flatness to other measures like noise stability: Jiang et al. [2020] evaluate a variety of complexity measures (including some Hessian-based) to see which best predict generalization; they found that no single measure works universally, but a combination can. Our introduction of coherence could add a new dimension to such measures, since it incorporates data-dependent interactions.

Linear stability. Linear stability has gained increasing attention in recent machine learning research as a tool to characterize the local convergence or divergence behavior around minima. This framework enables a unified perspective that jointly considers the data distribution, loss landscape geometry, and optimization dynamics. Prior works such as Wu et al. [2018, 2022], Wu and Su [2023] leveraged linear stability to analyze how noise interacts with local minima and to derive convergence criteria based on the Frobenius norm of the Hessian and Ma and Ying [2021] use the framework of linear stability to study property of noise in terms of its higher order moment. More recently, Dexter et al. [2024] introduced a coherence-based measure that captures fine-grained alignment properties of the data through the Hessian. These lines of work provide valuable new perspectives on the interplay between data and optimization—perspectives that are difficult to obtain through classical optimization analysis alone—and offer a deeper understanding of local training dynamics.

Our work is closely related to [Wu et al., 2018, 2022, Dexter et al., 2024], but compared to prior works [Wu et al., 2018, 2022, Wu and Su, 2023], instead of assuming mean square loss, we have more general abstraction to include different kind of loss function. Compared to Dexter et al. [2024] who focused on the analysis of SGD, we take one step further and analyze random noise injected

SGD and SAM. Specifically, we investigate how SAM influences local optimization dynamics and how it interacts with the structure of the data. Furthermore, we provide an explicit analysis of two-layer ReLU networks, revealing connections between linear stability, neural activations, and solution stability. This helps elucidate the role of SAM in shaping both the geometry and generalization behavior of trained models. Finally, unlike these prior works, we also set up realization of the theory to a two layer neural network and discuss how the insight from analysis in linear stability can be transferred to the neural network and show how the pattern of activation in neural network can related to result of linear stability. Specifically, we explicitly construct a 2-layer neural network with several solutions of the same sharpness but different complexity (captured by the sparsity in the activation pattern), and show that SGD (and SAM even more aggressively) prefers simpler (sparser) solutions.

Stability and implicit bias in optimization. Our use of “stability” is in the sense of dynamical stability of fixed points for the parameter update. This differs from the notion of algorithmic stability in learning theory (e.g., Hardt et al., 2016), which concerns how sensitive the final model is to removal of a training example. Algorithmic stability yields generalization bounds but doesn’t directly explain which solution is picked. Nonetheless, both concepts are linked: an optimizer that always returns the same minimum despite small data perturbations might be one that has a strong attractor basin (stable solution).

A large body of work on implicit bias of gradient methods has focused on linear models or homogeneous models, proving that gradient descent converges to particular norm-minimizing solutions or maximum margin solutions [Soudry et al., 2018, Gunasekar et al., 2018]. For example, Soudry et al. [2018] show that for linearly separable data and logistic loss, SGD converges to the max- L_2 -margin classifier. This can be seen as a form of simplicity bias (since a max-margin separator in linear space is a simpler decision boundary than a complex wiggle that also separates the data). In deep networks, Lyu and Li [2019] extended this to deep homogeneous networks (showing convergence to margin maximization). These works explain *which* solution among the continuum of minimizers is chosen, in terms of margins or norms. Our work provides a complementary lens: rather than characterizing the final solution in closed-form, we explain it via the dynamics preferences (coherence and stability during training). Margin and flatness might be connected; indeed, a large margin classifier often corresponds to a broad basin in loss landscape. Exploring the link between coherence and margin could be interesting (perhaps high coherence solutions also align with large margin in classification tasks).

The notion of *simplicity bias* has been documented empirically by several works. Arpit et al. [2017] found that deep nets first fit the “easy” patterns (e.g., clean labels) before memorizing noisy data, indicating a bias towards simpler functions. Kalimeris et al. [2019] and Valle-Pérez et al. [2019] argued from an information/combinatorics perspective that, because there are exponentially more complex functions than simple ones, a random initialization plus SGD is more likely to land in a simple function that fits the data (if such exists). Shah et al. [2020] (Pitfalls of Simplicity Bias) constructed datasets with multiple features to quantify this bias and showed it can hurt robustness. Our results give a theoretical underpinning to these observations by linking them to the Hessian structure and training dynamics: effectively, the simple patterns correspond to directions in which many data points have aligned gradients, hence those get learned quickly and form a stable basis for the solution, whereas complex patterns do not align and either get learned later or not at all.

Recently, Morwani et al. [2023] provided a rigorous analysis of simplicity bias in one-hidden-layer ReLU networks (in the infinite width, lazy training regime). They defined simplicity in terms of the function depending on a low-dimensional projection of inputs and proved that indeed gradient descent finds such low-dimensional solutions under certain conditions. Their findings dovetail nicely with our coherence interpretation (low-dimensional projection usage implies high alignment among gradients of those inputs). While their analysis is specialized to a particular regime, ours aims to be more generally intuitive and spans beyond the NTK regime by considering the Hessian of the nonlinear model.

Another related concept is *Neural Collapse* [Papayan et al., 2020], which describes that at the final layer of a classifier, the class means and features tend to align in certain simple symmetric patterns. Neural collapse occurs in the late phase of training and indicates a sort of self-organization of features. This might be seen as a high-coherence structure in the last-layer gradients for examples of the same class. While our work did not directly address neural collapse, the idea that training dynamics lead to aligned and symmetric configurations is broadly consistent.

Data geometry and gradient alignment. The role of data distribution in learning dynamics has been explored under terms like *gradient confusion* [Sankararaman et al., 2020] and *gradient alignment*. When gradients of different examples are more aligned, training converges faster and perhaps finds simpler models. Sankararaman et al. [2020] demonstrated that increasing overparameterization can reduce gradient confusion (making gradients more aligned by virtue of more flexible models finding a common direction) up to a point, which speeds up convergence. Chatterjee [2020] studied how examples that are hard or easy influence learning; easy examples likely align well with the gradient direction. Our coherence matrix formalizes one aspect of gradient alignment (at a second-order level, but one could similarly define $G_{ij} = \nabla \ell_i^\top \nabla \ell_j$ for first-order gradients). In fact, one could incorporate first-order coherence in our analysis; we focused on Hessian since it directly ties to stability, but gradient dot products matter for the actual update direction in SGD. A high Hessian coherence usually also implies gradient coherence at w^* if w^* is a zero training error solution (gradients are zero at w^* , but consider nearby points or earlier in training).

In summary, our work synthesizes ideas from these threads: we put forth coherence as a data-dependent quantifier that influences stability of solutions, thereby linking the optimizer’s implicit bias to the geometry of data in parameter space. By doing so, we integrate perspectives from flat minima research, implicit bias theory, and empirical studies of feature learning. We hope this unification will spur further research in understanding and controlling the biases of gradient-based training in deep learning.

C Appendix – Experiments and Proofs

C.1 Illustrative example for (C, r) solution and calculation of the r and trace

Recall our construction for (C, r) -generalizing solutions. We design W_1 by an exhaustive enumeration of all possible feature constructions of size C . In other words, $\forall \{a_1, a_2, \dots, a_C\} \in \{0, 1\}^C$, let the j^{th} row of W_1 be $W_{1,j} = r[(-1)^{a_1}, (-1)^{a_2}, \dots, (-1)^{a_C}, 0, 0, \dots, 0]$, with $j = 1 + \sum 2^{i-1}a_i$. Similarly, let $b[j] = -r(C - 1)$. We set $W_2[j] = \frac{1}{r}(-1)^{a_1+a_2}$. For $k > C$, $W_{1,k} = 0$, $W_2[k] = 0$, $b[k] = 0$. The following is the W_1 with $d = 5$, $c = 3$, $r = 1$ and hidden layer with 10 neurons.

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ -1 & 1 & 1 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 \\ -1 & 1 & -1 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 \\ -1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (6)$$

The following is the W_2 with $d = 5$, $c = 3$, $r = 1$ and hidden layer with 10 neurons.

$$\begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \\ 1 \\ -1 \\ -1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad (7)$$

569 The following is the b with $d = 5$, $c = 3$, $r = 1$ and hidden layer with 10 neurons.

$$\begin{bmatrix} 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 0 \\ 0 \end{bmatrix} \quad (8)$$

570 For the following, we show calculation of relationship between trace of Hessian. We first show the
 571 trace of one sample and corresponding flatness. As we known in previous calculation that gradient
 572 can be as follows:

$$\nabla f_w(x_i) = \begin{bmatrix} \text{ReLU}(W_{1,1}x_i) \\ \dots \\ \text{ReLU}(W_{1,d_2}x_i) \\ W_{2,1}\mathbf{1}[W_{1,1}x_i > 0]x_i \\ \dots \\ W_{2,j}\mathbf{1}[W_{1,d_2}x_i > 0]x_i \\ W_{2,j}\mathbf{1}[W_{1,1}x_i > 0] \\ \dots \\ W_{2,j}\mathbf{1}[W_{1,d_2}x_i > 0] \end{bmatrix} \quad (9)$$

573 And the Hessian is $\nabla f_w(x_i)\nabla f_w(x_i)^T$. (for zero loss solution) Take the trace will be
 574 $\text{Tr}[\nabla f_w(x_i)\nabla f_w(x_i)^T] = \|\nabla f_w(x_i)\|^2$ Now, we have exactly one activation at a time due to the
 575 bias (b) that impose such restriction. Therefore, the $\|\nabla f_w(x_i)\|^2 = r^2 + \frac{1}{r^2}d + \frac{1}{r^2} = r^2 + \frac{1}{r^2}(d+1)$

576 C.2 Role of coherence measure in dynamics

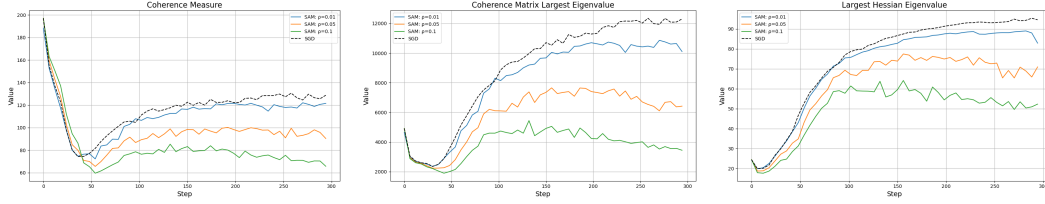


Figure 3: **2-layer ReLU network.** We found that the SAM method can impose strong regulation on the maximum eigenvalue elementwise, and this also reduce the strengthen of the largest eigenvalue of the coherence matrix. It means that the stability condition can be satisfied with smaller σ . From our experiments, we find that the sharpness of the solution impose strong regulation of the eigenvalue of the coherence matrix.

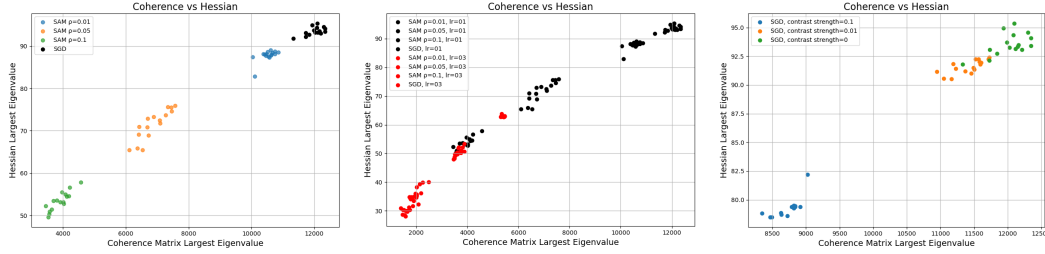


Figure 4: **2-layer ReLU network.** (Left) Comparison of SGD and SAM with different ρ . (Middle) We perform the same set of experiment with increased learning rate from 0.1 to 0.3. (Black to Red) (Right) SGD with different contrast loss strengthen (0.0, 0.1, 0.01). Through out the experiments, we find uniform shifting behavior for different algorithm with different strength but the relationship between $\max_i \lambda_{\max}(H_i)$ and $\lambda_{\max}(S)$ form strong regression line.

577 C.3 Experiment details - Local Linear stability in quadratic loss for different algorithms

578 (B, σ)

579 The experiments in this section serve to understand the local behavior of in terms of the linear
 580 stability. By studying the behavior near the local minimum, we aim to verify the correctness of our
 581 theory. We follow the same experiment set up to reproduce the plot in Dexter et al. [2024]. We first
 582 initialize $H_i = m e_1 e_1^T$ for all $i \in [\sigma]$ and $H_i = m e_{i-\sigma+1} e_{i-\sigma+1}^T$ otherwise. We use $m = \frac{2n}{\sigma}$ so that
 583 the sharpness of the minima ($\lambda_{\max}(H)$) is controlled to be 2. We set the learning rate to be smaller
 584 than 1 make sure diverging behavior arise due to noise not the sharpness. The loss function in the
 585 optimization is $l(w) = \frac{1}{n} \sum_{i=1}^n w H_i w$ and the gradient is $\nabla l(w) = \frac{2}{n} \sum_{i=1}^n H_i w$ that satisfies our
 586 theory setting. For all the experiment in this section, we set $n = 100$. For each set of parameters
 587 (B, η, σ) , we determine divergence or convergence by conducting 1000 steps update of the weight and
 588 calculate the norm of the weight. If the weight norm is 1000 times larger than original initialization,
 589 we classify it as diverging and vice versa. For each tuple, we perform the experiment 10 times. If
 590 the diverging behavior occurs more than half of the experiments set, we mark the specific tuple as
 591 diverging. The experiments involved in our work are done with CPU only.

592 C.4 Experiment details - Local Linear stability in mse loss for different algorithms in 2-layer

593 ReLU network.

594 We use the dimension of data $d = 100$ and the dimension of hidden layer is set to 50. Further, we
 595 use the batch size $B = 10$, the SAM $\rho = 0.01$, and the learning rate $\eta = 0.01$. We train for 50
 596 epochs and log the loss over epochs. All experiments comparing different algorithms are done with
 597 same initialization using the same random seed. The results are averaged over 5 runs.

598 **C.5 Experiment details - Global: The role of coherence in the training.**

599 To make the analysis more computationally tractable while tracking multiple quantities simultane-
600 ously, we reduce the model size: the input dimension is set to 15, the hidden layer size to 10, and the
601 number of training samples to 50. All other hyperparameters remain the same as in the Section 4.

602 C.6 Some identities and definition

603 We summarize the background and identities used through out the proof.

604 **Definition 3.** *The definition of Hessian and subset of Hessian where x_i is random variables with*
605 *Bernoulli distribution*

$$H_t = \frac{1}{B} \sum_{i=1}^n x_i H_i, \quad H = \frac{1}{n} \sum_{i=1}^n H_i \quad (10)$$

606 **Lemma C.1.** *Consider two matrix A, B with A being Positive semidefinite, then*

$$\lambda_{\max}(A) \text{Tr}[B] \geq \text{Tr}[AB] \geq \lambda_{\min}(A) \text{Tr}[B] \quad (11)$$

607 *The $\lambda_{\min}, \lambda_{\max}$ are smallest and largest eigenvalue of the matrix A .*

608 **Lemma C.2.** *Consider two matrix A, B, C , then*

$$\text{Tr}[ABC] = \text{Tr}[BCA] = \text{Tr}[CAB] \quad (12)$$

609 **Lemma C.3.** *l_1 - l_2 norm inequality: For any $x \in \mathbb{R}, \|x\|_2 \leq \|x\|_1 \leq \sqrt{d}\|x\|_2$*

610 **Lemma C.4. Binomial coefficient:** *For all $n, k \in \mathbb{N}$ such that $k \leq n$, the binomial coefficients*
611 *satisfy that*

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k} \quad (13)$$

612 **Lemma C.5.** *For any matrix $M \in \mathbb{R}^{n \times n}$, $\|M\|_F \leq \|M\|_{S_1} \leq \sqrt{n}\|M\|_F$, where $\|M\|_{S_p}$ is p*
613 *norm of the spectrum of M , and the inequality is obtain through l_1 - l_2 norm inequality.*

614 **Lemma C.6.** *For matrices $M_1 \dots M_k \in \mathbb{R}^{n \times n}$, $\text{Tr}[M_1 \dots M_k] \leq \|M_1 \dots M_k\|_{S_1}$ (see Bhatia [2013])*

615 C.7 Proof for Random perturbation

616 **Theorem C.7.** *Give update rule (3),*

617 *1. The condition for divergence is the same as that for SGD [Dexter et al., 2024] as follows:*

$$\eta \geq \frac{\sigma}{\lambda_1} \left(\frac{n}{b} - 1 \right)^{-\frac{1}{2}}$$

618 *2. (Comparative Divergence Speed) Suppose $\text{Tr}[J^{2k}] \leq C_0 \alpha^k$ for some constants C_0 and*
619 *α_k , then the divergence rate of the random perturbation method is asymptotically within a*
620 *constant factor of that of standard SGD:*

$$\lim_{k \rightarrow \infty} \frac{E[\|w_k\|^2]_{\text{Random, lower bound}}}{E[\|w_k\|^2]_{\text{SGD, lower bound}}} = \mathcal{O}(1)$$

621 *3. Suppose the step size satisfies the convergence criterion established in prior stability anal-*
622 *yses (e.g., Dexter et al. [2024]). Then, under the random perturbation update (3), the*
623 *expected squared norm of the iterates remains bounded as $k \rightarrow \infty$:*

$$\lim_{k \rightarrow \infty} E[w_k^T w_k]_{\text{upper bound}} = \mathcal{O}(1)$$

624 *Proof.* Define $H = \frac{1}{n} \sum_{i=1}^n H_i$. Now consider k steps after, we can have expression for w_k as
625 following:

$$w_k = \hat{J}_k \dots \hat{J}_1 w_0 - \eta \sum_{t=1}^k \left(\prod_{t'=t+1}^k \hat{J}_{t'} \right) H_t \delta_t \quad (14)$$

626 We consider the dot product of w_k and take expectation over all random process in between:

$$\begin{aligned} E[w_k^T w_k] &= E[(\hat{J}_k \dots \hat{J}_1 w_0 - \eta \sum_{t=1}^k (\prod_{t'=t+1}^k \hat{J}_{t'}) H_t \delta_t)^T (\hat{J}_k \dots \hat{J}_1 w_0 - \eta \sum_{t=1}^k (\prod_{t'=t+1}^k \hat{J}_{t'}) H_t \delta_t)] \\ &= E[w_0^T \hat{J}_1 \dots \hat{J}_k \hat{J}_k \dots \hat{J}_1 w_0] + \eta^2 E[(\sum_{t=1}^k (\prod_{t'=t+1}^k \hat{J}_{t'}) H_t \delta_t)^T (\sum_{t=1}^k (\prod_{t'=t+1}^k \hat{J}_{t'}) H_t \delta_t)] \end{aligned} \quad (15)$$

627 We first consider the second term. Note that all the cross terms are eliminated as they are independent
 628 to each other:

$$\begin{aligned}
 \eta^2 E[(\sum_{t=1}^k (\prod_{t'=t+1}^k \hat{J}_{t'}) H_t \delta_t)^T (\sum_{t=1}^k (\prod_{t'=t+1}^k \hat{J}_{t'}) H_t \delta_t)] &= \eta^2 E[\sum_{i=1}^k \delta_i^T H_i (\hat{J}_{i+1} \dots \hat{J}_K^2 \dots \hat{J}_{i+1}) H_i \delta_i] \\
 &= \eta^2 E[\sum_{i=1}^k \text{Tr}((\hat{J}_{i+1} \dots \hat{J}_K^2 \dots \hat{J}_{i+1}) (H_i \delta_i \delta_i^T H_i))] \\
 &= \eta^2 \sigma_1^2 \sum_{i=1}^k E[\text{Tr}(\hat{J}_{i+1} \dots \hat{J}_K^2 \dots \hat{J}_{i+1})] E[H_i^2] \\
 &\geq \eta^2 \sigma_1^2 \lambda_{\min}^2 \sum_{i=1}^k E[\text{Tr}((\hat{J}_{i+1} \dots \hat{J}_K^2 \dots \hat{J}_{i+1}))]
 \end{aligned} \tag{16}$$

629 The term $\text{Tr}((J_{t+1} \dots J_K^2 \dots J_{t+1}))$ can be decomposed into the following according to Dexter et al.
 630 [2024].

$$\begin{aligned}
 \eta^2 \sigma_1^2 \lambda_{\min}^2(H) \sum_{t=1}^k E[\text{Tr}((J_{t+1} \dots J_K^2 \dots J_{t+1}))] &\geq \eta^2 \sigma_1^2 \lambda_{\min}^2(H) (\sum_{t=1}^k \text{Tr}[J^{2t} + \eta^{2t} (\frac{1}{Bn} - \frac{1}{n^2})^t \sum_{y_1 \dots y_t=1}^n H_{y_1} \dots H_{y_t}^2 \dots H_{y_t}]) \\
 &\geq \eta^2 \sigma_1^2 \lambda_{\min}^2(H) (\sum_{t=1}^k \text{Tr}[J^{2t}] + (\frac{\eta}{\sigma})^{2t} (\frac{n}{b} - 1)^t \lambda_{\max}(H)^{2t})
 \end{aligned} \tag{17}$$

631 The last term represent the growth of the perturbation over time step. Contrary to the original
 632 analysis, the dependency of the magnitude is a summation of geometric series. However, despite
 633 the summation dependency, the criterion for diverging is still the same as we only require that
 634 $(\frac{\eta}{\sigma})^2 (\frac{n}{b} - 1) \lambda_{\max}(H)^2$ to be larger than 1. i.e.,

$$\eta \geq \frac{\sigma}{\lambda_{\max}} (\frac{n}{b} - 1)^{-\frac{1}{2}} \tag{18}$$

635 Now, observe that the perturbation base method will not change the fundamental criterion for the
 636 diverging. However, it will change the speed of diverging. We first consider the summation.

$$\eta^2 \sigma_1^2 \lambda_{\min}^2(H) \sum_{t=1}^k (\frac{\eta}{\sigma})^{2t} (\frac{n}{b} - 1)^t \lambda_{\max}^{2t} = \eta^2 \sigma_1^2 \lambda_{\min}^2(H) \frac{(\frac{\eta}{\sigma} \lambda_{\max})^2 (\frac{n}{b} - 1) ((\frac{\eta}{\sigma} \lambda_{\max})^{2k} (\frac{n}{b} - 1)^k - 1)}{(\frac{\eta}{\sigma} \lambda_{\max})^2 (\frac{n}{b} - 1) - 1} \tag{19}$$

637 Now we impose assumption on the growth of the $\text{Tr}(J^{2k})$ by assuming that it grow with pattern
 638 $C_0 \alpha^k$ and calculate the sum of it.

$$\sum_{t=1}^k C_0 \alpha^t = \frac{C_0 \alpha (\alpha^k - 1)}{\alpha - 1} \tag{20}$$

639 Finally, we temporarily denote the term $(\frac{\eta}{\sigma} \lambda_{\max}(H))^2 (\frac{n}{b} - 1)$ to be r and the overall lower bound
 640 for the calculation will be:

$$E[w_k^T w_k] \geq C_0 \alpha^k + \frac{1}{nd^5} r^k + \eta^2 \sigma_1^2 \lambda_{\min}^2(H) (\frac{C_0 \alpha (\alpha^k - 1)}{\alpha - 1} + \frac{1}{nd^5} \frac{r(r^k - 1)}{r - 1}) \tag{21}$$

641 Now we set the $\sigma_1 = 1$ and compare with the original naive SGD and set that $\alpha \geq r$

$$\lim_{k \rightarrow \infty} \frac{E[w_k^T w_k]_{\text{Random, lower bound}}}{E[w_k^T w_k]_{\text{SGD, lower bound}}} = 1 + \eta^2 \sigma_1^2 \lambda_{\min}^2(H) \frac{\alpha}{\alpha - 1} \quad (22)$$

642 The escaping speed of the random perturbation base method is faster by a constant. Now we set that
643 $\alpha \leq r$

$$\lim_{k \rightarrow \infty} \frac{E[w_k^T w_k]_{\text{Random, lower bound}}}{E[w_k^T w_k]_{\text{SGD, lower bound}}} = 1 + \eta^2 \sigma_1^2 \lambda_{\min}^2 \frac{r}{r - 1} \quad (23)$$

644 No matter which term dominates, we will have constant faster escaping efficiency compared to the
645 original naive SGD.

646 Now, we consider the convergence behavior, we first note that there exists ϵ and C such that
647 $E[\hat{J}_k \dots \hat{J}_1^2 \dots \hat{J}_k] \leq C((1 - \epsilon)^2 + \epsilon)^k$. Here, we temporarily denote $((1 - \epsilon)^2 + \epsilon)$ to be r . We apply
648 this identity to the above equation and we will get the following:

$$\begin{aligned} E[w_k^T w_k] &\leq C r^k + \eta^2 \lambda_{\max} \sum_{t=1}^k C r^t \\ &= C r^k + \eta^2 \lambda_{\max} \frac{r(1 - r^k)}{1 - r} \end{aligned} \quad (24)$$

649 We consider long term behavior (i.e., $k \rightarrow \infty$)

$$\lim_{k \rightarrow \infty} E[w_k^T w_k] \leq \eta^2 \lambda_{\max} \frac{r}{1 - r} \quad (25)$$

650 We observe that there exist residual terms relating to the perturbation itself and this fits our intuition
651 that the random perturbation method will usually hover around the minimum as the noise injected
652 can lead to less accurate estimation of the gradient direction. \square

653 **C.8 Proof for divergence theorem**

654 **Theorem C.8.** *For update rules as following:*

$$W_{t+1} = (I - \eta H_t (I + \frac{\rho}{\alpha} H)) W_t \quad (26)$$

655 Define $\hat{J}_t = (I - \eta H_t (I + \frac{\rho}{\alpha} H))$, then

656

657 **1. There exist M_k such that**

$$E[\hat{J}_k^T \dots \hat{J}_1^T \hat{J}_1 \dots \hat{J}_k] \succeq M_k \quad (27)$$

658 with

$$M_k = J^{2k} + \eta^{2k} \left(\frac{1}{Bn} - \frac{1}{n^2} \right)^k \sum_{y_1 \dots y_k=1}^n (I + \frac{\rho}{\alpha} H) H_{y_k} \dots (I + \frac{\rho}{\alpha} H) H_{y_1}^2 (I + \frac{\rho}{\alpha} H) \dots H_{y_k} (I + \frac{\rho}{\alpha} H) \quad (28)$$

659 **2. The Trace of M_k can be lower bounded by the following:**

$$\text{Tr}[M_k] \geq \eta^{2k} \left(\frac{1}{Bn} - \frac{1}{n^2} \right)^k (1 + \frac{\rho}{\alpha} \lambda_{\min}(H))^{2k} \text{Tr} \left[\sum_{y_1 \dots y_k=1}^n H_{y_k} \dots H_{y_1}^2 \dots H_{y_k} \right] \quad (29)$$

660 **3. The Trace of M_k is lower bounded through σ :**

$$\text{Tr}[M_k] \geq \eta^{2k} \left(\frac{n}{B} - 1 \right)^k (1 + \frac{\rho}{\alpha} \lambda_{\min}(H))^{2k} \frac{1}{\sigma^{2k}} \frac{1}{nd^5} \lambda_{\max}(H)^{2k} \quad (30)$$

661 **4. The diverging criterion for SAM under linear stability is:**

$$\lambda_{\max}(H) \geq \frac{\sigma}{\eta} \left(\frac{n}{B} - 1 \right)^{-\frac{1}{2}} \left(1 + \frac{\rho}{\alpha} \lambda_{\min}(H) \right)^{-1} \quad (31)$$

662 *Proof.* We prove by induction as follows.

663 **Base case: $k=1$**

$$\begin{aligned} E[\hat{J}_1^T \hat{J}_1] &= E[(I - \eta H_t (I + \frac{\rho}{\alpha} H))^T (I - \eta H_t (I + \frac{\rho}{\alpha} H))] \\ &= E[I - 2\eta H_t - \frac{\eta \rho}{\alpha} (H_t H + H H_t) + \eta^2 H_t^2 + \frac{\eta^2 \rho}{\alpha} (H_t^2 H + H H_t^2) + \frac{\eta^2 \rho^2}{\alpha^2} H H_t^2 H] \\ &= I - 2\eta H - 2\frac{\eta \rho}{\alpha} H^2 + \eta^2 E[H_t^2] + \frac{\eta^2 \rho}{\alpha} (E[H_t^2] H + H E[H_t^2]) + \frac{\eta^2 \rho^2}{\alpha^2} H E[H_t^2] H \end{aligned} \quad (32)$$

664 We know that

$$E[H_t^2] = H^2 + \left(\frac{1}{Bn} - \frac{1}{n^2} \right) \sum_{i=1} H_i^2 \quad (33)$$

665 and we will have

$$\begin{aligned} E[\hat{J}_1^T \hat{J}_1] &= J^2 + \eta^2 \left(\frac{1}{Bn} - \frac{1}{n^2} \right) (I + \frac{\rho}{\alpha} H) \left(\sum_{i=1} H_i^2 \right) (I + \frac{\rho}{\alpha} H) \\ &= M_1 \end{aligned} \quad (34)$$

666 **Induction case: k-1 to k**

$$\begin{aligned}
& E[\hat{J}_k^T \dots \hat{J}_1^T \hat{J}_1 \dots \hat{J}_k] \succeq E[\hat{J}_k^T M_{k-1} \hat{J}_k] \\
& = E[(I - \eta H_k - \frac{\eta \rho}{\alpha} H_k H)^T M_{k-1} (I - \eta H_k - \frac{\eta \rho}{\alpha} H_k H)] \\
& = E[M_{k-1} - \eta M_{k-1} H_k - \frac{\eta \rho}{\alpha} M_{k-1} H_k H - \eta H_k M_{k-1} + \eta^2 H_k M_{k-1} H_k + \frac{\eta^2 \rho}{\alpha} H_k M_{k-1} H_k H - \\
& \quad \frac{\eta \rho}{\alpha} H H_k M_{k-1} + \frac{\eta^2 \rho}{\alpha} H H_k M_{k-1} H_k + \frac{\eta^2 \rho^2}{\alpha^2} H H_k M_{k-1} H_k H] \\
& = J M_{k-1} J + \eta^2 (\frac{1}{nB} - \frac{1}{n^2}) (I + \frac{\rho}{\alpha} H) (\sum_i H_i M_{k-1} H_i) (I + \frac{\rho}{\alpha} H)
\end{aligned} \tag{35}$$

667 Now, we substitute the expression of M_{k-1} into the expression and we will have the follows:

$$\begin{aligned}
& E[\hat{J}_k^T \dots \hat{J}_1^T \hat{J}_1 \dots \hat{J}_k] \succeq J M_{k-1} J + \eta^2 (\frac{1}{nB} - \frac{1}{n^2}) (I + \frac{\rho}{\alpha} H) (\sum_i H_i M_{k-1} H_i) (I + \frac{\rho}{\alpha} H) \\
& = J [J^{2(k-1)} + \eta^{2(k-1)} (\frac{1}{Bn} - \frac{1}{n^2})^{(k-1)} \sum_{y_1 \dots y_k=1}^n (I + \frac{\rho}{\alpha} H) H_{y_{k-1}} \dots (I + \frac{\rho}{\alpha} H) H_{y_1}^2 (I + \frac{\rho}{\alpha} H) \dots H_{y_{k-1}} (I + \frac{\rho}{\alpha} H)] J + \\
& \quad \eta^2 (\frac{1}{nB} - \frac{1}{n^2}) (I + \frac{\rho}{\alpha} H) (\sum_i H_i [J^{2(k-1)} + \\
& \quad \eta^{2(k-1)} (\frac{1}{Bn} - \frac{1}{n^2})^{(k-1)} \sum_{y_1 \dots y_k=1}^n (I + \frac{\rho}{\alpha} H) H_{y_{k-1}} \dots (I + \frac{\rho}{\alpha} H) H_{y_1}^2 (I + \frac{\rho}{\alpha} H) \dots H_{y_{k-1}} (I + \frac{\rho}{\alpha} H)] H_i) (I + \frac{\rho}{\alpha} H) \\
& \succeq J^{2k} + \eta^{2k} (\frac{1}{Bn} - \frac{1}{n^2})^k \sum_{y_1 \dots y_k=1}^n (I + \frac{\rho}{\alpha} H) H_{y_k} \dots (I + \frac{\rho}{\alpha} H) H_{y_1}^2 (I + \frac{\rho}{\alpha} H) \dots H_{y_k} (I + \frac{\rho}{\alpha} H) \\
& = M_k
\end{aligned} \tag{36}$$

668 Now, we wish to analyze the trace of the matrix M_k . For simplicity, we analyze the latter term of
669 the expression.

$$\text{Tr}[M_k] = \text{Tr}[\eta^{2k} (\frac{1}{Bn} - \frac{1}{n^2})^k \sum_{y_1 \dots y_k=1}^n (I + \frac{\rho}{\alpha} H) H_{y_k} \dots (I + \frac{\rho}{\alpha} H) H_{y_1}^2 (I + \frac{\rho}{\alpha} H) \dots H_{y_k} (I + \frac{\rho}{\alpha} H)] \tag{37}$$

670 We first focus on specific term in the summation.

$$\begin{aligned}
& \text{Tr}[(I + \frac{\rho}{\alpha} H) H_{y_k} \dots (I + \frac{\rho}{\alpha} H) H_{y_1}^2 (I + \frac{\rho}{\alpha} H) \dots H_{y_k} (I + \frac{\rho}{\alpha} H)] = \\
& \text{Tr}[(I + \frac{\rho}{\alpha} H)^2 H_{y_k} (I + \frac{\rho}{\alpha} H) H_{y_{k-1}} \dots (I + \frac{\rho}{\alpha} H) H_{y_1}^2 (I + \frac{\rho}{\alpha} H) \dots H_{y_{k-1}} (I + \frac{\rho}{\alpha} H) H_{y_k}] \\
& \geq (1 + \frac{\rho}{\alpha} \lambda_{\min}(H))^2 \text{Tr}[H_{y_k} \dots (I + \frac{\rho}{\alpha} H) H_{y_1}^2 (I + \frac{\rho}{\alpha} H) \dots H_{y_k}] \\
& = (1 + \frac{\rho}{\alpha} \lambda_{\min}(H))^3 \text{Tr}[H_{y_{k-1}} \dots (I + \frac{\rho}{\alpha} H) H_{y_1}^2 (I + \frac{\rho}{\alpha} H) \dots H_{y_{k-1}} (I + \frac{\rho}{\alpha} H) H_{y_k}^2] \\
& \geq (1 + \frac{\rho}{\alpha} \lambda_{\min}(H))^4 \text{Tr}[H_{y_{k-1}} H_{y_k}^2 H_{y_{k-1}} (I + \frac{\rho}{\alpha} H) \dots (I + \frac{\rho}{\alpha} H) H_{y_1}^2 (I + \frac{\rho}{\alpha} H) \dots (I + \frac{\rho}{\alpha} H)]
\end{aligned} \tag{38}$$

671 Here, we use the lemma C.1. By continue pruning, we will have the following:

$$\text{Tr}[(I + \frac{\rho}{\alpha}H)H_{y_k} \dots (I + \frac{\rho}{\alpha}H)H_{y_1}^2(I + \frac{\rho}{\alpha}H) \dots H_{y_k}(I + \frac{\rho}{\alpha}H)] \geq (1 + \frac{\rho}{\alpha}\lambda_{\min}(H))^{2k} \text{Tr}[H_{y_k} \dots H_{y_1}^2 \dots H_{y_k}] \quad (39)$$

672 We apply this to the summation in M_k and we will get

$$\text{Tr}[M_k] \geq \eta^{2k} (\frac{1}{Bn} - \frac{1}{n^2})^k (1 + \frac{\rho}{\alpha}\lambda_{\min}(H))^{2k} \text{Tr}[\sum_{y_1 \dots y_k=1}^n H_{y_k} \dots H_{y_1}^2 \dots H_{y_k}] \quad (40)$$

673 Now, we connect the M_k with the coherence measure in the following:

$$\begin{aligned} \text{Tr}[M_k] &\geq \eta^{2k} (\frac{1}{Bn} - \frac{1}{n^2})^k (1 + \frac{\rho}{\alpha}\lambda_{\min}(H))^{2k} \sum_{y_1 \dots y_k=1}^n \text{Tr}[H_{y_k} \dots H_{y_1}^2 \dots H_{y_k}] \\ &= \eta^{2k} (\frac{1}{Bn} - \frac{1}{n^2})^k (1 + \frac{\rho}{\alpha}\lambda_{\min}(H))^{2k} \sum_{y_1 \dots y_k=1}^n \|H_{y_k} \dots H_{y_1}\|_F^2 \\ &\geq \eta^{2k} (\frac{1}{Bn} - \frac{1}{n^2})^k (1 + \frac{\rho}{\alpha}\lambda_{\min}(H))^{2k} \sum_{y_1 \dots y_k=1}^n \frac{1}{d} \|H_{y_k} \dots H_{y_1}\|_{S_1}^2 \\ &\geq \eta^{2k} (\frac{1}{Bn} - \frac{1}{n^2})^k (1 + \frac{\rho}{\alpha}\lambda_{\min}(H))^{2k} \sum_{y_1 \dots y_k=1}^n \frac{1}{d} \text{Tr}[H_{y_k} \dots H_{y_1}]^2 \\ &\geq \eta^{2k} (\frac{1}{Bn} - \frac{1}{n^2})^k (1 + \frac{\rho}{\alpha}\lambda_{\min}(H))^{2k} \sum_{y=1}^n \frac{1}{d} \text{Tr}[H_y^k]^2 \\ &\geq \eta^{2k} (\frac{1}{Bn} - \frac{1}{n^2})^k (1 + \frac{\rho}{\alpha}\lambda_{\min}(H))^{2k} \frac{1}{nd} (\sum_{y=1}^n \text{Tr}[H_y^k])^2 \end{aligned} \quad (41)$$

674 for the above we use lemma C.5 and we know the following from Dexter et al. [2024]:

$$\frac{n^k}{d^2 \sigma^k} \text{Tr}[H^k] \leq \sum_{y=1}^n \text{Tr}[H_y^k] \quad (42)$$

675 Finally, we can have the following:

$$\begin{aligned} \text{Tr}[M_k] &\geq \eta^{2k} (\frac{1}{Bn} - \frac{1}{n^2})^k (1 + \frac{\rho}{\alpha}\lambda_{\min}(H))^{2k} \frac{1}{nd} (\frac{n^k}{d^2 \sigma^k})^2 (\text{Tr}[H^k])^2 \\ &= \eta^{2k} (\frac{1}{Bn} - \frac{1}{n^2})^k (1 + \frac{\rho}{\alpha}\lambda_{\min}(H))^{2k} \frac{1}{nd} (\frac{n^k}{d^2 \sigma^k})^2 (\text{Tr}[H^k])^2 \\ &= \eta^{2k} (\frac{n}{B} - 1)^k (1 + \frac{\rho}{\alpha}\lambda_{\min}(H))^{2k} \frac{1}{\sigma^{2k}} \frac{1}{nd^5} (\text{Tr}[H^k])^2 \\ &\geq \eta^{2k} (\frac{n}{B} - 1)^k (1 + \frac{\rho}{\alpha}\lambda_{\min}(H))^{2k} \frac{1}{\sigma^{2k}} \frac{1}{nd^5} \lambda_{\max}(H)^{2k} \end{aligned} \quad (43)$$

676 We then will have the following condition for diverging:

$$\lambda_{\max}(H) \geq \frac{\sigma}{\eta} (\frac{n}{B} - 1)^{-\frac{1}{2}} (1 + \frac{\rho}{\alpha}\lambda_{\min}(H))^{-1} \quad (44)$$

677 □

678 **C.9 Proof for convergence theorem**

679 **Theorem C.9.** *For update rules as following:*

$$W_{t+1} = (I - \eta H_t(I + \frac{\rho}{\alpha} H))W_t \quad (45)$$

680 **1. There exist N_r such that**

$$E[\hat{J}_k^T \dots \hat{J}_1^T \hat{J}_1 \dots \hat{J}_k] \preceq \sum_{r=0}^k (1 - \epsilon)^{2(k-r)} \binom{k}{r} N_r \quad (46)$$

681 *and*

$$N_k = \eta^{2k} \left(\frac{1}{nB} - \frac{1}{n^2} \right)^k \sum_{y_1, \dots, y_r=1}^n (I + \frac{\rho}{\alpha} H) H_{y_k} \dots (I + \frac{\rho}{\alpha} H) H_{y_1}^2 (I + \frac{\rho}{\alpha} H) \dots H_{y_k} (I + \frac{\rho}{\alpha} H) \quad (47)$$

682 **2. The N_r can be upper bounded as following**

$$\text{Tr}[N_r] \leq \eta^{2k} \left(\frac{1}{B} - \frac{1}{n} \right)^k d^{3k+\frac{1}{2}} n^{4k} \frac{\lambda_{\max}(H_{SAM})^{4k}}{\sigma_{SAM}^{2k}} \quad (48)$$

683 **3. Suppose there exist $\epsilon \in (0, 1)$ and we will have converging criterion such that**

$$\begin{aligned} \frac{\epsilon}{\eta} &\leq \lambda_i + \frac{\rho}{\alpha} \lambda_i^2 \leq \frac{2 - \epsilon}{\eta} \quad \forall i \in [d] \quad \text{and} \\ \lim_{k \rightarrow \infty} \frac{1}{\epsilon^k} \eta^{2k} \left(\frac{1}{nB} - \frac{1}{n^2} \right)^k \sum_{y_1, y_2, \dots, y_k=1}^n (I + \frac{\rho}{\alpha} H) H_{y_k} \dots (I + \frac{\rho}{\alpha} H) H_{y_1}^2 (I + \frac{\rho}{\alpha} H) \dots H_{y_k} (I + \frac{\rho}{\alpha} H) &= 0 \end{aligned} \quad (49)$$

684 *then we will have that $\lim_{k \rightarrow \infty} E[\hat{J}_k^T \dots \hat{J}_1^T \hat{J}_1 \dots \hat{J}_k] = 0$*

685 *Proof.* We first define N_r as follows:

$$N_k = \eta^{2k} \left(\frac{1}{nB} - \frac{1}{n^2} \right)^k \sum_{y_1, \dots, y_r=1}^n (I + \frac{\rho}{\alpha} H) H_{y_k} \dots (I + \frac{\rho}{\alpha} H) H_{y_1}^2 (I + \frac{\rho}{\alpha} H) \dots H_{y_k} (I + \frac{\rho}{\alpha} H) \quad (50)$$

686 We define $N_0 = I$

687 we want to prove the following:

$$E[\hat{J}_k^T \dots \hat{J}_1^T \hat{J}_1 \dots \hat{J}_k] \preceq \sum_{r=0}^k (1 - \epsilon)^{2(k-r)} \binom{k}{r} N_r \quad (51)$$

688 **Base case: $k=1$**

$$\begin{aligned} E[\hat{J}_1^T \hat{J}_1] &= E[(I - \eta \hat{H} - \frac{\eta \rho}{\alpha} \hat{H} H)^T (I - \eta \hat{H} - \frac{\eta \rho}{\alpha} \hat{H} H)] \\ &= J^2 + \eta^2 \left(\frac{1}{nB} - \frac{1}{n^2} \right) \sum_i (I + \frac{\rho}{\alpha} H) H_i^2 (I + \frac{\rho}{\alpha} H) \\ &\preceq (1 - \epsilon)^2 N_0 + N_1 \end{aligned} \quad (52)$$

689 The first condition $(1 - \epsilon)^2 I$ is achieved when the condition of the assumption is satisfied. We
690 demonstrate why that is the case

$$J = I - \eta H(I + \frac{\rho}{\alpha} H) \quad (53)$$

691 and observe the following

$$-(1 - \epsilon)^2 I \preceq J^2 \preceq (1 - \epsilon)^2 I \quad (54)$$

692 First, we focus on the

$$J^2 \preceq (1 - \epsilon)^2 I \quad (55)$$

693 By replacing the definition of J into the equation, we will reach

$$(I - \eta H(I + \frac{\rho}{\alpha} H))^2 \preceq (1 - \epsilon)^2 I \quad (56)$$

694 By removing the square term,

$$(I - \eta H(I + \frac{\rho}{\alpha} H)) \preceq (1 - \epsilon) I \quad (57)$$

695 Rearrange will give

$$\epsilon I \preceq \eta H(I + \frac{\rho}{\alpha} H) \quad (58)$$

696 By decomposing each eigenvalue direction, we can have that

$$\frac{\epsilon}{\eta} \leq \lambda_i + \frac{\rho}{\alpha} \lambda_i^2 \quad (59)$$

697 We perform the same operation on the other direction and we will have

$$\lambda_i + \frac{\rho}{\alpha} \lambda_i^2 \leq \frac{2 - \epsilon}{\eta} \quad (60)$$

698 The ϵ in our analysis represent the deterministic term in each step as we use it to upper bound J^2 .
 699 Compared to SGD, the ϵ can be larger as the term $I - \eta H$ is larger than $I - \eta H - \eta \frac{\rho}{\alpha} H^2$ in each
 700 direction. The deterministic part of update process can shrink faster compared to the SGD. The
 701 analysis implicitly incorporate it through ϵ . For the other part, N_1 represent the randomness in the
 702 operation which can be directly checked that if we increase the batch size to n , we will have the
 703 term vanished. The format of the N_r rely on how different sample align with each other and this
 704 form origin of the noise. Compared to the tradition analysis, we can have more subtle observation
 705 of the noise through this specific form as we no longer need to assume the structure of the noise or
 706 we can assume the how each sample align with each other to see the final form of the noise and this
 707 can reveal more insight of the relationship between sample for further analysis. Now, we proceed to
 708 induction step.

709 **Induction case: k-1**

$$\begin{aligned} E[\hat{J}_k^T \dots \hat{J}_1^T \hat{J}_1 \dots \hat{J}_k] &\preceq E[\hat{J}_k^T (\sum_{r=0}^{k-1} (1 - \epsilon)^{2(k-1-r)} \binom{k-1}{r} N_r) \hat{J}_k] \\ &= J^T (\sum_{r=0}^{k-1} (1 - \epsilon)^{2(k-1-r)} \binom{k-1}{r} N_r) J + \eta^2 (\frac{1}{nb} - \frac{1}{n^2}) \sum_i (I + \frac{\rho}{\alpha} H) H_i (\sum_{r=0}^{k-1} (1 - \epsilon)^{2(k-1-r)} \binom{k-1}{r} N_r) H_i (I + \frac{\rho}{\alpha} H) \\ &\preceq (1 - \epsilon)^2 \sum_{r=0}^{k-1} (1 - \epsilon)^{2(k-1-r)} \binom{k-1}{r} N_r + \eta^2 (\frac{1}{nb} - \frac{1}{n^2}) \sum_i \sum_{r=0}^{k-1} (1 - \epsilon)^{2(k-1-r)} \binom{k-1}{r} (I + \frac{\rho}{\alpha} H) H_i N_r H_i (I + \frac{\rho}{\alpha} H) \\ &= \sum_{r=0}^{k-1} (1 - \epsilon)^{2(k-r)} \binom{k-1}{r} N_r + \eta^2 (\frac{1}{nb} - \frac{1}{n^2}) \sum_{r=0}^{k-1} (1 - \epsilon)^{2(k-1-r)} \binom{k-1}{r} \sum_i (I + \frac{\rho}{\alpha} H) H_i N_r H_i (I + \frac{\rho}{\alpha} H) \\ &= \sum_{r=0}^{k-1} (1 - \epsilon)^{2(k-r)} \binom{k-1}{r} N_r + \sum_{r=0}^{k-1} (1 - \epsilon)^{2(k-1-r)} \binom{k-1}{r} N_{r+1} \end{aligned} \quad (61)$$

710 By reordering the term, we will have the following using lemma C.4:

$$\begin{aligned}
& \sum_{r=0}^{k-1} (1-\epsilon)^{2(k-r)} \binom{k-1}{r} N_r + \sum_{r=1}^k (1-\epsilon)^{2(k-r)} \binom{k-1}{r-1} N_r \\
&= (1-\epsilon)^{2k} N_0 + \sum_{r=0}^{k-1} ((1-\epsilon)^{2(k-r)} \binom{k-1}{r} + (1-\epsilon)^{2(k-r)} \binom{k-1}{r-1}) N_r + N_k \\
&= \sum_{r=0}^k (1-\epsilon)^{2(k-r)} \binom{k}{r} N_r
\end{aligned} \tag{62}$$

711 Similarly, as we require that the $\text{Tr}[N_r]$ term to be smaller than ϵ^r , we can upper bound the term with
712 constant C such that $\frac{1}{\epsilon^r} \text{Tr}[N_r] \leq C$, and therefore,

$$\begin{aligned}
\sum_{r=0}^k (1-\epsilon)^{2(k-r)} \binom{k}{r} \text{Tr}[N_r] &\leq C \sum_{r=0}^k \binom{k}{r} (1-\epsilon)^{2(k-r)} \epsilon^r \\
&= C((1-\epsilon)^2 + \epsilon)^k
\end{aligned} \tag{63}$$

713 For the last step, as we ask the $\text{Tr}[N_r]$ term to be smaller than ϵ , we can further bound it. Despite we
714 can upper bound it through ϵ , we can still analysis its magnitude. If it is smaller, it will also converge
715 faster. The distinction between SAM and SGD is that SAM has extra multiplication $(I + \frac{\rho}{\alpha} H)$ and
716 this result from the operation with the gradient ascent intermediate step. We can find that this can
717 potentially make the process unstable as it amplify the noise through out the process and this fit in
718 to our general believe that SAM can make the optimization process less stable but still converge fast
719 due to the shrink of the deterministic term. Note that unlike the tradition analysis that require step
720 size to be $\eta \leq \frac{2}{\lambda_{\max}(H)}$, we ask for different criterion for converging. This is due to the fact that
721 we analysis the origin of noise which comes from alignment of samples and the traditional analysis
722 focus more on the deterministic part which directly involve eigenvalue of Hessian and usually the
723 analysis specify the noise with covariance matrix instead. \square

724 To answer the relationship between ϵ and the N_r . We first consider the following inequility

$$\begin{aligned}
\lambda_1(S)^k &\leq \text{Tr}(S^k) \\
&= \sum_{y_1 \dots y_k=1}^n \|H_{y_1}^{\frac{1}{2}} H_{y_k}^{\frac{1}{2}}\|_F \dots \|H_{y_2}^{\frac{1}{2}} H_{y_1}^{\frac{1}{2}}\|_F \\
&= \sum_{y_1 \dots y_k=1}^n \text{Tr}(H_{y_1} H_{y_k}) \dots \text{Tr}(H_{y_2} H_{y_1}) \\
&= n^{2k} (\text{Tr}(H^2))^k \\
&\leq n^{2k} d^k \lambda_{\max}(H)^{2k}
\end{aligned} \tag{64}$$

725 Now, we defined a new form of coherence matrix and Hessian to accomadate the SAM algorithm as
726 following:

$$S_{\text{SAM}_{ij}} = \sqrt{\text{Tr}((I + \frac{\rho}{\alpha} H) H_i (I + \frac{\rho}{\alpha} H) H_j)} \tag{65}$$

727

$$H_{\text{SAM}_{ij}} = (I + \frac{\rho}{\alpha} H) \sum_{i=1}^n H_i \tag{66}$$

728 Now, we go back to the N_r

$$\begin{aligned}
\text{Tr}(N_r) &= \eta^{2k} \left(\frac{1}{nB} - \frac{1}{n^2} \right)^k \sum_{y_1 \dots y_k=1}^n \text{Tr} \left(\left(I + \frac{\rho}{\alpha} H \right) H_{y_k} \dots \left(I + \frac{\rho}{\alpha} H \right) H_{y_1}^2 \left(I + \frac{\rho}{\alpha} H \right) \dots H_{y_k} \left(I + \frac{\rho}{\alpha} H \right) \right) \\
&= \eta^{2k} \left(\frac{1}{nB} - \frac{1}{n^2} \right)^k \sum_{y_1 \dots y_k=1}^n \left\| \left(I + \frac{\rho}{\alpha} H \right) H_{y_k} \dots \left(I + \frac{\rho}{\alpha} H \right) H_{y_1} \right\|_F^2 \\
&\leq \eta^{2k} \left(\frac{1}{nB} - \frac{1}{n^2} \right)^k \sqrt{d} \sum_{y_1 \dots y_k=1}^n \left\| \left(I + \frac{\rho}{\alpha} H \right) H_{y_k} \right\|_F^2 \dots \left\| \left(I + \frac{\rho}{\alpha} H \right) H_{y_1} \right\|_F^2 \\
&\leq \eta^{2k} \left(\frac{1}{B} - \frac{1}{n} \right)^k \sqrt{d} \max_{i=1 \dots n} \left\| \left(I + \frac{\rho}{\alpha} H \right) H_i \right\|_F^{2k} \\
&\leq \eta^{2k} \left(\frac{1}{B} - \frac{1}{n} \right)^k \sqrt{d} \max_{i=1 \dots n} d^k (\lambda_{\max}((I + \frac{\rho}{\alpha} H) H_i))^{2k} \\
&\leq \eta^{2k} \left(\frac{1}{B} - \frac{1}{n} \right)^k d^{3k + \frac{1}{2}} n^{4k} \frac{\max_{i=1 \dots n} (\lambda_{\max}((I + \frac{\rho}{\alpha} H) H_i))^{2k}}{\lambda_{\max}(S_{\text{SAM}})^{2k}} \lambda_{\max}(H_{\text{SAM}})^{4k} \\
&\leq \eta^{2k} \left(\frac{1}{B} - \frac{1}{n} \right)^k d^{3k + \frac{1}{2}} n^{4k} \frac{\lambda_{\max}(H_{\text{SAM}})^{4k}}{\sigma_{\text{SAM}}^{2k}}
\end{aligned} \tag{67}$$

In our analysis, through new definition of the coherence matrix and Hessian, we find that SAM is performing optimization on the loss surface that is amplified by $I + \frac{\rho}{\alpha} H$. The loss surface is sharper as it give larger eigenvalue in each direction. If we compared about the ratio $\frac{\lambda_{\max}(H)^4}{\sigma^2}$ and $\frac{\lambda_{\max}(H_{\text{SAM}})^4}{\sigma_{\text{SAM}}^2}$, we can see question about which one is larger or smaller will need more information about the exact coherence matrix to determine. They can be the same or different depending on the relationship between samples. However, for both of the method, if the solution give larger coherence measure, they both converge faster for the specific solution and vice versa.

C.10 Proof for theorem 3.4

Proof. We know that the $\nabla f_w(x_i)$ can be written as following:

$$\nabla f_w(x_i) = \begin{bmatrix} \text{ReLU}(W_{1,1}x_i) \\ \dots \\ \text{ReLU}(W_{1,d_2}x_i) \\ W_{2,1}\mathbf{1}[W_{1,1}x_i > 0]x_i \\ \dots \\ W_{2,j}\mathbf{1}[W_{1,d_2}x_i > 0]x_i \\ W_{2,j}\mathbf{1}[W_{1,1}x_i > 0] \\ \dots \\ W_{2,j}\mathbf{1}[W_{1,d_2}x_i > 0] \end{bmatrix} \tag{68}$$

where the gradient is taken with respect to parameter (W_2, W_1, b) in sequence. For each element in the coherence matrix, we will have

$$\begin{aligned}
S_{i,j} &= \|H_i^{\frac{1}{2}} H_j^{\frac{1}{2}}\|_F = \sqrt{\text{Tr}(H_j^{\frac{1}{2}} H_i^{\frac{1}{2}} H_i^{\frac{1}{2}} H_j^{\frac{1}{2}})} = \sqrt{\text{Tr}(H_i H_j)} = \sqrt{(\nabla f_w^T(x_i) \nabla f_w(x_j))^2} \\
&= |\nabla f_w^T(x_i) \nabla f_w(x_j)|
\end{aligned} \tag{69}$$

As the activation of the samples are orthogonal to each other in memorizing solution. The orthogonal in activation will also give the gradient orthogonal property and therefore,

$$\text{Tr}(H_i H_j) = (\nabla f_i \nabla f_j)^2 = 0 \tag{70}$$

Therefore, the coherence matrix is diagonal in the setting. The corresponding coherence measure is small compared to other solution and we can conclude that the memorizing solution is relatively hard to find during optimization process as seen in the prior work with coherence measure. The reverse is also true. If the coherence matrix is diagonal, then the solution is memorizing solution. As if we have two data activation overlap, the gradient product will not be zero. \square

747 **C.11 Proof for theorem 3.5**

748 *Proof.* Suppose we draw a dataset with size n uniformly at random. The coherence matrix becomes
 749 block diagonal matrix with eigenvalue being $2(d+1)^{\frac{1}{2}} \max_{i=1 \dots 2^C} |S_i|$ where S_i is the set with data
 750 matched to specific feature extracted by $W_{1,i}$ and we know the following:

$$S_{i,j} = |\nabla f_w^T(x_i) \nabla f_w(x_j)| = 2(d+1)^{\frac{1}{2}}. \quad (71)$$

751 We estimate the following:

$$P(\max_{i=1 \dots 2^C} |S_i| \geq nu + n\epsilon). \quad (72)$$

752 We can find that by union bound

$$P(\max_{i=1 \dots 2^C} |S_i| \geq nu + n\epsilon) \leq \sum_{i=1}^{2^C} P(|S_i| \geq nu + n\epsilon) \quad (73)$$

753 Let $u = \frac{1}{2^C}$ and $\epsilon = \sqrt{\frac{C + \log \frac{1}{\delta}}{2n}}$. Also, we know that by $|S_i|$ is a sum of independent variables X_{ik}
 754 that fall into the category and we can formulate through chernoff bound:

$$P(|S_i| \geq nu + n\epsilon) = P\left(\frac{1}{n} \sum_{j=1}^n X_j \geq u + \epsilon\right) \leq \exp(-2\epsilon n^2) \leq \exp(-(C + \log \frac{1}{\delta})) = e^{-C} \delta \quad (74)$$

755 Therefore,

$$\begin{aligned} P(\max_{i=1 \dots 2^C} |S_i| \geq nu + n\epsilon) &\leq \sum_{i=1}^{2^C} P(|S_i| \geq nu + n\epsilon) \\ &\leq \sum_{i=1}^{2^C} e^{-C} \delta \\ &\leq \delta \end{aligned} \quad (75)$$

756

□

757 **C.12 Proof for theorem 3.6**

758 *Proof.* For the generalizing solution, we can analyze the element in the coherence matrix. $\text{Tr}((I +$
 759 $\frac{\rho}{\alpha} H) H_i (I + \frac{\rho}{\alpha} H) H_j)$. We can see that if two samples do not share the same activation, the specific
 760 element will be zero. We consider average value for the element that are within the same cluster. As
 761 they are in the same cluster, the $H_i = H_j = H_S$

$$\begin{aligned} E[(\sum_k X_k) \sqrt{\text{Tr}[(I + \frac{\rho}{\alpha} H) H_i (I + \frac{\rho}{\alpha} H) H_j]}] &= E[(\sum_k X_k) \sqrt{\text{Tr}[H_i H_j + \frac{\rho}{\alpha} H H_i H_j + \frac{\rho}{\alpha} H_j H H_i + \frac{\rho^2}{\alpha^2} H H_i H H_j]}] \\ &= E[(\sum_k X_k) \sqrt{\text{Tr}[H_S^2] + \frac{2\rho}{\alpha} \text{Tr}[H_S^3] \sum_{k=1}^n X_k + \frac{\rho^2}{\alpha^2} \text{Tr}[H_S^4] \sum_{kk'} X_k X_{k'}}] \\ &= E[(\sum_k X_k) \text{Tr}[H_S] \sqrt{1 + \frac{2\rho}{\alpha} \text{Tr}[H_S] \sum_{k=1}^n X_k + \frac{\rho^2}{\alpha^2} \text{Tr}[H_S^2] \sum_{kk'} X_k X_{k'}}] \end{aligned} \quad (76)$$

762 where the $\sum_k X_k$ are random variables indicating the sample inside the specific cluster or not. The
 763 other term is the strengthen of coherence elementwise. We can observe that this is a convex function
 764 in terms of the random variables and therefore we can lower bound it by taking the expectation first
 765 in each random variable:

$$\begin{aligned}
& E[(\sum_k X_k) \sqrt{\text{Tr}[(I + \frac{\rho}{\alpha} H) H_i (I + \frac{\rho}{\alpha} H) H_j]}] \\
& \geq E[\sum_k X_k] \text{Tr}[H_S] \sqrt{1 + \frac{1}{n} \frac{2\rho}{\alpha} \text{Tr}[H_S] E[\sum_{k=1}^n X_k] + \frac{1}{n^2} \frac{\rho^2}{\alpha^2} \text{Tr}[H_S^2] E[\sum_{kk'} X_k X_{k'}]}] \quad (77)
\end{aligned}$$

766 The key lies in the term $E[\sum_{kk'} X_k X_{k'}]$ which is not simply the multiplication of the two individual
767 probability and we can find that it is $\frac{1}{2^{2c}} + \frac{1}{n}(\frac{1}{2^c} - \frac{1}{2^{2c}})$ and we will have the following:

$$\begin{aligned}
& E[(\sum_k X_k) \sqrt{\text{Tr}[(I + \frac{\rho}{\alpha} H) H_i (I + \frac{\rho}{\alpha} H) H_j]}] \\
& \geq \frac{n}{2^c} 2(d+1)^{\frac{1}{2}} \sqrt{1 + \frac{2\rho}{\alpha} \frac{1}{2^c} 2(d+1)^{\frac{1}{2}} + \frac{\rho^2}{\alpha^2} (\frac{1}{2^{2c}} + \frac{1}{n}(\frac{1}{2^c} - \frac{1}{2^{2c}})) 4(d+1)} \quad (78) \\
& = \frac{n}{2^c} 2(d+1)^{\frac{1}{2}} \sqrt{(1 + \frac{\rho}{\alpha} \frac{2(d+1)^{\frac{1}{2}}}{2^c})^2 + \frac{\rho^2}{\alpha^2} (\frac{1}{n}(\frac{1}{2^c} - \frac{1}{2^{2c}})) 4(d+1)}
\end{aligned}$$

768 Now, the even stronger dependency of the number of features used can also be translated to the
769 probability statement. With probability $1 - \delta$, the eigenvalue of the coherence matrix is upper
770 bounded by $\mathcal{O}(\frac{n}{2^c} (d+1)^{\frac{1}{2}} \sqrt{(1 + \frac{\rho}{\alpha} \frac{2(d+1)^{\frac{1}{2}}}{2^c})^2 + \frac{\rho^2}{\alpha^2} (\frac{1}{n}(\frac{1}{2^c} - \frac{1}{2^{2c}})) 4(d+1)})$ using the same method
771 as in appendix C.12.

772 The additional higher order interacting term give strong additional bias toward solution with lower
773 C , by observation, we can see that the term become more significant when the data dimension (also
774 model dimension) becomes higher and cannot be neglect for modern deep learning scenario in terms
775 of overparameter region. Also, to check the correctness of our result, we find that by replacing ρ to
776 be zero, we can recover back to the case for SGD exactly.

777 To calculate the eigenvalue of $\max_i \lambda_{\max}(H_i)$, we will need to calculate the average number of the
778 data that align with each other as follows:

$$\begin{aligned}
\max_i \lambda_{\max}((I + \frac{\rho}{\alpha} H) H_i) &= \max_i \lambda_{\max}((I + \frac{\rho}{n\alpha} \sum_{i=1}^n \nabla f_w(x_i) \nabla f_w(x_i)^T) \nabla f_w(x_i) \nabla f_w(x_i)^T) \\
&= \max_i \|\nabla f_w(x_i)\|^2 + \frac{\rho}{\alpha} \|\nabla f_w(x_i)\|^4 \frac{1}{2^c} \\
&= 2(d+1)^{\frac{1}{2}} (1 + \frac{\rho}{\alpha} \frac{1}{2^c} 2(d+1)^{\frac{1}{2}}) \quad (79)
\end{aligned}$$

779

□

780 **C.13 Proof for theorem 3.3**

781 *Proof.* We follow the construction of prior work Dexter et al. [2024] and focus on the term $E \text{Tr}[\hat{J}^T \hat{J}]$
782 where $\hat{J} = I - \eta H_t - \frac{\eta \rho}{\alpha} H_t H$ (Note that $H_t = \sum_i x_i H_i$, where x_i is Bernoulli with probability
783 $\frac{B}{n}$ being 1) with the probability of sampling each sample being independent Bernoulli distribution.
784 The construction of set $\{H_i\}_{i \in [n]}$ is such that $H_i = m e_1 e_1^T$, $\forall i \in [\sigma]$ and $H_i = 0$ otherwise, and
785 $m = \frac{\lambda_1 n}{\sigma}$ so that $\lambda_{\max}(H) = \frac{\sigma}{n} m = \lambda_1$. Note that $\lambda_{\max}(S) = m\sigma$ and $\max_i \lambda_{\max}(H_i) = m$
786 and the coherence measure is exactly σ . Also under this construction, $E[\text{Tr}[\hat{J}_k^T \dots \hat{J}_1^T \hat{J}_1 \dots \hat{J}_k]] =$
787 $\text{Tr}[E[(\hat{J}_1^T \hat{J}_1)^{2k}]]$ as all matrix involved are commuting with i.i.d sampled.

788 We have following:

$$\begin{aligned} E[\hat{J}^T \hat{J}] &= E[(I - \eta H_t - \frac{\eta \rho}{\alpha} H H_t)(I - \eta H_t - \frac{\eta \rho}{\alpha} H_t H)] \\ &= E[I - 2\eta H_t - \frac{\eta \rho}{\alpha} (H H_t + H_t H) + \eta H_t^2 + \frac{\eta^2 \rho}{\alpha} (H_t^2 H + H H_t^2) + \frac{\eta^2 \rho^2}{\alpha^2} H H_t^2 H] \\ &= I - 2\eta H_t - 2\frac{\eta \rho}{\alpha} H^2 + \eta H_t^2 + 2\frac{\eta^2 \rho}{\alpha} H^3 + \frac{\eta^2 \rho^2}{\alpha^2} H^4 + \eta^2 (\frac{1}{Bn} - \frac{1}{n^2}) \sum_i (I + \frac{\rho}{\alpha} H) H_i^2 (I + \frac{\rho}{\alpha} H) \end{aligned} \quad (80)$$

789 Now, we calculate $e_1^T E[\hat{J}^T \hat{J}] e_1$ (e_1 is the only direction that involve interaction of different samples)
790 will give us

$$e_1^T E[\hat{J}^T \hat{J}] e_1 = 1 - 2\eta \lambda_1 - 2\frac{\eta \rho}{\alpha} \lambda_1^2 + \eta \lambda_1^2 + 2\frac{\eta^2 \rho}{\alpha} \lambda_1^3 + \frac{\eta^2 \rho^2}{\alpha^2} \lambda_1^4 + \eta^2 (\frac{1}{Bn} - \frac{1}{n^2}) (1 + \frac{\rho}{\alpha} \lambda_1)^2 \frac{n^2 \lambda_1^2}{\sigma} \quad (81)$$

791 We need the term to be smaller than 1 to avoid growing infinitely

$$1 - 2\eta \lambda_1 - 2\frac{\eta \rho}{\alpha} \lambda_1^2 + \eta \lambda_1^2 + 2\frac{\eta^2 \rho}{\alpha} \lambda_1^3 + \frac{\eta^2 \rho^2}{\alpha^2} \lambda_1^4 + \eta^2 (\frac{1}{Bn} - \frac{1}{n^2}) (1 + \frac{\rho}{\alpha} \lambda_1)^2 \frac{n^2 \lambda_1^2}{\sigma} \leq 1 \quad (82)$$

792 We can rearrange and obtain the following:

$$-2\eta \lambda_1 (1 + \frac{\rho}{\alpha} \lambda_1) + \eta^2 \lambda_1^2 (1 + \frac{\rho}{\alpha} \lambda_1)^2 + \eta^2 (\frac{n}{B} - 1) (1 + \frac{\rho}{\alpha} \lambda_1)^2 \frac{\lambda_1^2}{\sigma} \leq 0 \quad (83)$$

793 We find that we can divide the equation on both side by $\eta \lambda_1 (1 + \frac{\rho}{\alpha} \lambda_1)$ and have

$$-2 + \eta \lambda_1 (1 + \frac{\rho}{\alpha} \lambda_1) + \eta (\frac{n}{B} - 1) (1 + \frac{\rho}{\alpha} \lambda_1) \frac{\lambda_1}{\sigma} \leq 0 \quad (84)$$

794 Now, we can rearrange and have the following:

$$\eta \lambda_1 (1 + \frac{\rho}{\alpha} \lambda_1) (\frac{\frac{n}{B} - 1}{\sigma} + 1) \leq 2 \quad (85)$$

795

$$\frac{\eta \lambda_1}{\sigma} (1 + \frac{\rho}{\alpha} \lambda_1) (\frac{n}{B} - 1 + \sigma) \leq 2 \quad (86)$$

796 Finally, we will have

$$\lambda_1 (1 + \frac{\rho}{\alpha} \lambda_1) \leq \frac{2\sigma}{\eta} (\frac{n}{B} - 1 + \sigma)^{-1} \quad (87)$$

797 The additional term $(1 + \frac{\rho}{\alpha} \lambda_1)$ result from the SAM modified surface gives a more restricted learning
798 rate choice compared to the SGD. We can also check the result by setting $\rho = 0$ and will find that it
799 reduce to the original SGD criterion. \square

800 **C.14 Theorem from Dexter et al. [2024]**

801 **Theorem 1** Let $\{\hat{J}_i\}_{i \in \mathbb{N}}$ be a sequence of i.i.d copies of \hat{J} defined in linearized SGD. Let $\{H_i\}_{i \in [n]}$
 802 have coherence measure σ . If

$$\lambda_{\max}(H) \geq \frac{2}{\eta} \text{ or } \lambda_{\max}(H) \geq \frac{\sigma}{\eta} \left(\frac{n}{B} - 1\right)^{-\frac{1}{2}}, \text{ then } \lim_{k \rightarrow \infty} E\|\hat{J}_k \dots \hat{J}_1\| = \infty \quad (88)$$

803 **Theorem 2** For every choice of $\lambda_{\max} > 0, n \in \mathbb{N}, B \in [n], \eta > 0$ and $\sigma \in [n]$, that satisfies:

$$\lambda_{\max} < \frac{2\sigma}{\eta} \left(\sigma + \frac{n}{B} - 1\right)^{-1} \quad (89)$$

804 There exists a set of PSD matrices $\{H_i\}_{i \in [n]}$ such that $\lambda_{\max}(H) = \lambda_{\max}$ and $\lim_{k \rightarrow \infty} E\|\hat{J}_k \dots \hat{J}_1\| <$
 805 n .

806 **Lemma 4.1** Let \hat{J}_i be independent Jacobians of SGD dynamics,

807 (1) If

$$\lambda_{\max} \geq \frac{2}{\eta} \text{ or } \lim_{k \rightarrow \infty} \left(\frac{\eta^2}{nB} - \frac{\eta^2}{n^2}\right) \sum_{y_1 \dots y_k=1}^n \|H_{y_k} \dots H_{y_1}\|_F = \infty \quad (90)$$

808 then $\lim_{k \rightarrow \infty} E\|\hat{J}_k \dots \hat{J}_1\|_F^2 = \infty$

809 (2) If, for some $\epsilon \in (0, 1)$,

$$\frac{\epsilon}{\eta} < \lambda_i(H) < \frac{2-\epsilon}{\eta} \quad \forall i \in [d] \text{ and } \frac{1}{\epsilon^k} \lim_{k \rightarrow \infty} \left(\frac{\eta^2}{nB} - \frac{\eta^2}{n^2}\right) \sum_{y_1 \dots y_k=1}^n \|H_{y_k} \dots H_{y_1}\|_F = 0 \quad (91)$$

810 then $\lim_{k \rightarrow \infty} E\|\hat{J}_k \dots \hat{J}_1\|_F^2 = 0$