1 Comments about reviews

1.1 General response

We sincerely thank the reviewers for their positive feedback and many useful suggestions.

All the reviewers point out that the paper is not easy to read, and some indicated the exposition of our loss expansion to be especially convoluted. We agree that the "Loss expansion" subsection in section 4 was excessively compressed. We have rewritten this subsection, emphasizing the intuition and general ideas. The technical derivations were moved to new Appendix D and now are written there with more detail. The loss expansion itself is formulated now as new Theorem 2. This has slightly shifted the numbering of formulas, theorems and sections in the paper; in our responses below we refer to the new numbers.

In addition, we have fixed a number of typos and clarified the exposition in several other places (in particular, the derivation of the convergence rate of the algorithm AM1 in section L).

1.2 Reviewer kpRS (8).

Thank you very much for your careful reading and such a positive evaluation of our work!

...the Loss expansion section with the introduction of propagators seemed too dense - and was lacking intuitions that would facilitate further reading of the paper..

Thank you for pointing out this weakness - we fully agree and have rewritten this piece.

Q1. *As mentioned, I wonder if there are any ways of improving the readability of the technical portions with further addition of intuitions and heuristics.*

Q2. *For example, the Loss expansion section with the introduction of propagators seemed too dense - and was lacking intuitions that would facilitate further reading of the paper, since those concepts were heavily relied on for further developments in the paper. Those intuitions were present in the earlier sections and were thus helpful for speeding up understanding of the paper*

1.3 Reviewer 1Va8 (8).

Thanks very much for your careful reading and a very positive evaluation of our work!

Q1. *One possible suggestion is that the paper develops the results initially in the as the shifts are not needed for the main contributions and then provide extensions, possibly in the appendix. The other is to hide some of the intermediate steps to provide space for more intuition for the different objects.*

Thank you for these suggestions - we seriously thought about them, especially removing the shifts. We didn't get to implement it, because of the lack of time and because we wanted to keep some discussion of shifts in the main text, but we plan to return to this issue in a future update. However, we have rewritten the excessively dense "loss expansion" section, moving intermediate steps in the appendix as you suggested (see our general response).

Q2. *One thing that threw me off is using capital letters for the propagators as these are scalars.*

Thank you for pointing this out. We now mention explicitly in the text that the propagators are scalar, to minimize possible confusion. There are several reasons why we use capital letters for the propagators: 1) they denote direct contribution to the averaged loss L_t which is also consistently denoted by a capital letter (e.g., in the noiseless case $L_t = \frac{1}{2}V_t$); 2) propagators U_t are extensively considered as entries of a circulant matrix $\mathbb{U}$ in various derivations in the appendix; 3) propagators are important objects, and most lowercase letters are already used for other purposes in the paper.

Q3. *Could such a framework be extended to non-IID batching, that is, batching with replacement (recent work [1,2])?*

We expect that yes, it could be extended in this way. Reading these papers, analysis of batching with replacement is certainly technically more involved than of batching without it. But at least for large training sets and small batches, we can expect both kinds of batching to be close, suggesting similar properties of our generalized SGD.

Q4. *Could you say anything about the generalization error using [3]?*

The paper [3] derives a precise asymptotic characterization of the test error under power-law spectrum. Yet, there seem to be several challenges left to obtain (and then analyze) such characterization for mini-batch gradient-based algorithms, including memory-M algorithms:

– The test error in [3] is derived for kernel ridge regression (which is especially convenient for tools from random matrix theory thanks to its connection to the resolvent), but not for gradient descent. The case of gradient descent is, however, addressed to some extent in some other papers, e.g. [4,5].

– Regarding the stochasticity of inputs, in [3] and other related works the training set is randomly drawn once before applying the learning algorithm. The SGD setting adds one more layer of stochasticity due to the random choice of mini-batches from the finite-size training set. The joint averaging over these two sources of randomness might be challenging.

– If we want to characterize the generalization error under a label noise, this also needs to be added to the framework (see another question below for more details).

On a positive side, though, the SE approximation used in our framework holds exactly for Gaussian features used in [3].

[4] Yicheng Li, Weiye Gan, Zuoqiang Shi, and Qian Lin. Generalization error curves for analytic spectral algorithms under power-law decay, 2024. URL https://arxiv.org/ abs/2401.01599.

[5] M. Velikanov, M. Panov, and D. Yarotsky. Generalization error of spectral algorithms. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=3SJE1WLB4M.

Q5. *Could we add label noise or regularization?*

Both label noise and regularization could be introduced in our framework by generalizing some of its parts. Basically, we need to modify second moments update (12),(13), and then add new respective terms into propagator expansion (19).

In case of additive i.i.d. label noise with variance σ^2, second moments update acquires a term (one can use derivation in sec. C.1 as an example)

$$\frac{\sigma^2}{|B|}\mathbf{H} \otimes \begin{pmatrix} -\alpha_t \\ \mathbf{c}_t \end{pmatrix} \begin{pmatrix} -\alpha_t & \mathbf{c}_t \end{pmatrix}.$$

The case of L_2 regularization is a bit more involved with more terms changed/added. Since it pulls the parameters to 0, the optimum will shift to a new location $\mathbf{w}^*_{reg}$, modifying both the main evolution term A_t and the noise term P_t that make up the evolution operator F_t (see section D).

However, the biggest changes should be made to the loss propagator expansion. In its current version, we heavily rely on the multiplicative structure of the evolution operator, its form $F_t = A_t + P_t$, and P_t being rank-1. Adding extra terms of a different type to F_t, or introducing additive terms into the second moment's update as in the case of label noise, would require redesigning the loss expansion. This is an interesting direction for future research.

1.4 Reviewer LKb1 (3).

Thank you for your feedback!

Q1. *In general, the text is difficult to read....*

Q2 (Maksim). *I don't fully understand about the experimental part. Is it just not there?*

While the paper primarily focuses on theoretical analysis, we do provide experiments to validate the derived results, as presented in figures 1, 4, 5. All the settings are described in the dedicated section L of the appendix. In short, we believe that our current set of experiments covers the following points

– Figure 1 demonstrates the divergence of Jacobi accelerated HB that was mentioned in the introduction. Importantly, the divergence of Jacobi accelerated HB happens at any batch size, with larger batch size values just delaying the moment of divergence. This serves as a motivation to look into accelerated strategies, leading us to the formulation and analysis of Memory-M algorithms.

– Same figure 1 shows that the proposed AM1 achieves an accelerated power-low convergence rate, although not as high as that of Jacobi accelerated HB in the noiseless setting. Moreover, the AM1 algorithm is stable even at quite small batch sizes ($|B| = 1$ and $|B| = 10$ in the figure).

– Figure 1 also provides the noisy and noiseless (full-batch) loss curves in otherwise identical settings. This visually illustrates the effect of noise on the loss trajectory in both stable and diverging cases.

– In terms of scope, we consider either synthetic Gaussian data with the manually set power-law spectrum and MNIST classification with a shallow ReLU network. We describe the details of each setting in section M. These two settings serve different goals.

— The Gaussian setting is directly described by the developed theory and was expected to follow our theoretical predictions. Yet, the last step of designing a schedule for AM1 algorithm and obtaining its convergence rate is heuristic and non-rigorous, in contrast to the rest of the paper. Thus it was important for us to validate the intuition used in the heuristic derivation of AM1. Indeed, the predicted convergence rate, depicted by dashed lines, matches the experimental curves.

— The MNIST setting, although basic by modern deep learning standards, is non-convex and thus deviates from the quadratic optimization problem analyzed in the paper. Yet, we see that the main properties of AM1, stability and accelerated convergence, survive deviation from purely quadratic problems. This gives us some hope that AM1 algorithm could have broad practical applications in real-world optimization problems.

– Finally, in figures 4 and 5 we provide more experiments by changing various parameters of both Gaussian and MNIST settings. These experiments show that the conclusions from figure 1 are robust.

Q3 (Maksim,Dmitry?). *Can authors convince me that their work fits the spirit of this conference? This question is not obvious to me at this point.....*

We are trying to solve a problem -acceleration of SGD - that is relatively hard and does not seem to have a simple intuitive solution. To this end, we develop and present an original theoretical approach that includes many new elements (general memory-M algorithms, a propagator expansion of the loss, generalized effective learning rate, stable memory-1 algorithms with unbounded effective learning rate ...) and eventually reveals a solution. We believe our work to advance the ML science and thus fit the spirit of ICLR quite well.

Many of the works directly connected to our research and cited in our publication list have also been published either at ICLR or similar conferences like NeurIPS, ICML, COLT. This further confirms that our work fits the spirit of ICLR well.

1.5 Reviewer eHN7 (6).

Thank you for your careful reading and positive evaluation of our work!

W1. *The paper seemed quite convoluted*

That's true; we have rewritten one part that was especially convoluted, please see our general response.

W2. *The empirical validation of the results remains quite limited*

We agree that in our empirical validations we covered a limited set of scenarios. We have validated that the heuristically derived AM1 algorithm actually gives the predicted convergence rate while being stable in numerical simulations, and in the MNIST experiments, we showed that AM1 still works in non-convex setting without exact power-law spectrum and on real-world data. See also a detailed summary of our experiments in the response to the reviewer LKb1.

A comprehensive test of AM1 in a variety of modern deep-learning scenarios would be an interesting direction for future research. Yet, we are convinced that the provided set of experiments fits well the current manuscript. We would like to keep the focus of this submission on theoretical analysis and derivation of the algorithm, limiting experiments to validation of main theoretical claims and providing proof-of-concept evidence that the algorithm can work beyond the purely quadratic setting considered in the paper.

W3. *It is seems like the optimal rate depends on the right choice of hyperparameters, such as* α_0*,* q_0 *and* δ*. However it is unclear to me how to select or estimate them in practice. This limits its application to real-world problems.*

Thank you for this remark! We tried to address this question in the paper, although not very directly.

– Although theorems 6 and 7 provide specific choice of $\delta, q_0, \alpha_{\text{eff}}$ hyperparameters, it is natural to assume that the precise values will change if we deviate from the setting of quadratic models with exact power-law spectrum. Yet, we believe that the region of hyperparamters $\delta^{-1} \gg \alpha_{\text{eff}} \gg q_0 = const$, identified in theorem 7, will be more robust to the change of the setting (non-power law spectrum, non-quadratic loss, etc). See the remark in the penultimate paragraph on page 9.

– Another possible difficulty for practical application is formulation of the algorithm in terms of variables $\delta, q_0, \alpha_{\text{eff}}$. These variables appear to be the most convenient parametrization for our theoretical analysis, but not necessarily for the practical implementation of the algorithm. We address this in the beginning of section M, where we rewrite the AM1 update rule as a modification of HB (a.k.a SGD in most deep learning libraries). Specifically, we can just add can an extra "kick" term - updating parameters directly with the gradients of the current iteration. Denoting the learning rate of the new kick term as α_1, and the learning rate of the classical momentum term as α_2, the AM1 schedule becomes $(\delta_t, \alpha_{1,t}, \alpha_{2,t}) \sim (t^{-\overline{\delta}}, const, t^{-\overline{\delta}+\overline{\alpha}})$.

– Hyperparameter choice. One can parameterized the whole schedule as $(\delta_t, \alpha_{1,t}, \alpha_{2,t}) = (c_\delta t^{-\overline{\delta}}, c_{\alpha_1}, c_{\alpha_2} t^{-\overline{\delta}+\overline{\alpha}})$ leading to 5 tunable hyperparameters. A quick starting point could be (i) setting reasonable values for constants $c_\delta, c_{\alpha_1}, c_{\alpha_2}$ (ii) picking $\overline{\delta}$ close to its theoretically predicted value $\overline{\delta} = 1$ to avoid some edge effects, e.g. $\overline{\delta} = 0.9$ (iii) doing a sweep on value of $\overline{\alpha}$, since its theoretical optimal prediction $\overline{\alpha} = \overline{\delta}(1 - \frac{1}{\nu})$ requires knowledge of the spectral exponent ν which is typically unknown for real-world problems. We roughly followed this strategy in our experiments, as shown on figures 4,5.

– Finally, we note that there is freedom in choosing schedule exponents $\overline{\delta}, \overline{\alpha}$, removing the need to precisely estimate spectrum exponent ν. From our results in sec. 6, the algorithm would be stable as soon as $\overline{\alpha} \leq \overline{\delta}(1 - \frac{1}{\nu})$, and acceleration is determined by the value of $\overline{\alpha}$. Thus, a less aggressive choice than theoretically optimal $\overline{\delta} = 1, \overline{\alpha} = 1 - \frac{1}{\nu}$ would still provide accelerated convergence while being stable.

Q1. *I do not fully understand the image on the right of Figure 2. What are* $t_1, \dots, t_4$ *supposed to denote or perhaps more generally what does it mean that a signal/noise propagator is "long" or "short"?*

The values t_k are the iteration numbers ("time values") from loss expansion (19). By "long"/"short" we simply mean a propagator V_t or U_t with a large or small t. More precisely, when considering the propagator expansion for L_t, "long" means length $t - o(t)$ and "short" means length $o(t)$. By inspecting the proof of Theorem 2, only those configurations of propagator factors that include one long propagator an several short ones (as shown in Figure 2) make the leading contribution to the loss expansion (19) in the signal- and noise-dominated convergence regimes. This is the key intuition behind the proofs of the loss asymptotics for these regimes in Theorem 2.

We are happy to discuss this question as it relates to the propagator expansion central to our approach. First, the times $t_1, \dots, t_4$ could be understood formally, as the iterations over which the summation in (19) is performed. Recall that the loss is given by the product of single iteration evolution operators F_t, which consists of diagonal term A_t and rank-1 term P_t

$$F_t F_{t-1} \dots F_2 F_1 = (A_t + P_t)(A_{t-1} + P_{t-1}) \dots (A_2 + P_2)(A_1 + P_1).$$

Then, the times $t_1, \dots, t_4$ correspond to time steps at which P term was chosen when expanding the parentheses. The length of the propagator equals to the amount of A operators chosen subsequently, before choosing P or terminating the dynamics with loss computation.

Now, we can try to build an intuitive picture. Picking P_t term at iteration t corresponds to injecting the noise, created due to mini-batch sampling on this iteration, into parameter moments $\mathbf{M}_{t+1}$. Picking A_t corresponds to ignoring the noise (at least for $\tau_2 = 0$) and shrinking $\mathbf{M}_{t+1}$ as per convergence of the algorithm, given by $S_{\lambda,t}$. Then, having a long V propagator in the signal-dominated phase means that the convergence is bottlenecked by decaying the initial parameter displacement in $\mathbf{M}_0$. Having a long U propagator means the convergence is bottlenecked by decaying the noise created at the beginning of the optimization process.

TODO: interpretation of t_m? U_t, V_t are scalar

Version 2.

The meaning of times $t_1, \dots, t_m$ is central to our propagator expansion. We tried to make this part clearer in the revised version and also added a dedicated section D in the appendix. In the proof of new Theorem 2 (loss expansion), the loss at time T appears by considering the second moments update at each step as the sum of a diagonal operator A_t (mainly, the noiseless part of the evolution) and a rank 1 operator P_t (coming from the $\mathbf{H}\,\mathrm{Tr}[\mathbf{HC}]$ term of the minibatch noise variance in eq. (15)):

$$L_T = \mathrm{Tr}[\mathbf{H}\mathbf{C}_T] = \mathrm{Tr}\left[\langle \left(\begin{smallmatrix}1 & 0\\ 0 & 0\end{smallmatrix}\right), (A_{T-1} + P_{T-1})(A_{T-2} + P_{T-2}) \dots (A_1 + P_1)(A_0 + P_0)\mathbf{M}_0 \rangle\right].$$

Then, the times $t_1, \dots, t_m$ correspond to those iterations where we chose P_t when expanding the parentheses above. Taking into account all the ways to expand the parentheses translates into the sum over all propagator configurations in eq. (19).

By "long"/"short" we simply mean a propagator V_t or U_t with a large or small t. More precisely, when considering the propagator expansion for L_T, "long" means length $T - o(T)$ and "short" means length $o(T)$. By inspecting the proof of Theorem 3, only those configurations of propagator factors that include one long propagator an several short ones (as shown in Figure 3) make the leading contribution to the loss expansion (19) in the signal- and noise-dominated convergence regimes. This is the key intuition behind the proofs of the loss asymptotics for these regimes in Theorem 3.

Q2. *I have another question regarding the stability of the maps*

No, $\lambda_{\max}$ does not need to be small for A_λ to be stable throughout $0 < \lambda < \lambda_{\max}$. Theorem 3 (4 after revision) characterizes stability of A_λ only for small $\lambda > 0$ and does not say anything about larger λ, where the statement $\mu_{A,\lambda} = 1 - 2\alpha_{\text{eff}}\lambda + O(\lambda^2)$ becomes uninformative.

For example, as can be seen in Figure 3, both in Heavy Ball and our memory-1 methods, at some critical value λ the leading eigenvalue $\mu_{A,\lambda}$ collides with the second eigenvalue, after which they both start moving along a circle in the complex plane. In the examples shown in Figure 3, all the eigenvalues μ_λ of $A_\lambda, 0 < \lambda \le \lambda_{\max}$, stay within the unit circle and so A_λ are stable for all $0 < \lambda \le \lambda_{\max}$. However, the regime associated with the eigenvalues on the circle is not described by Theorem 3 (4).

At the same time, in the case of memory-1 algorithms, we do give a complete characterization of stability for all $0 < \lambda \le \lambda_{\max}$ (and in particular covering all regimes) in Theorem 6, part 1. This characterization is necessarily significantly more complicated than the condition $\alpha_{\text{eff}} > 0$ from Theorem 3 (4).

Q3. *What are the values of ξ_v, ξ_U in Eq. (21) and (22)?*

Theorem 2 is a combinatorial result that assumes that the propagators are asymptotically power-law, and derives the loss asymptotics based only on the propagator expansion (19) of the loss. So, in Theorem 2 the values ξ_v, ξ_U can be any numbers such that $\xi_U > 1, \xi_V > 0$. But then, in Theorem 5 we show that for the particular propagators given by Eqs. (17),(18) we have $\xi_U = 2 - 1/\nu, \xi_V = \zeta$. We separate the two theorems because we feel that such structure of the exposition clarifies the origin of the loss asymptotics $L_t \sim t^{-\zeta}$ and $L_t \sim t^{1/\nu - 2}$ in the signal- and noise-dominated regimes.

Q4. *Have the authors also considered non-quadratic settings and how does it perform here?*

This is a good question. Empirically, we observe our algorithm to work on slightly non-quadratic problems: note that in our MNIST experiments (see Figure 1, right, and Figure 5) the network is finite and so the loss is non-quadratic. We observe both acceleration and stability to be present in these experiments.

Theoretically, it is a good future research direction, and there are some interesting challenges. In particular, our work relies heavily on spectral source and capacity conditions: a reasonable extension to non-quadratic problems would need to propose their non-quadratic generalizations while demonstrating that these can be fulfilled in realistic problems.

2 General response.

We thank the reviewers for useful comments. Following the feedback, we have substantially revised the paper. The main changes are:

- We have generally revised the text for better clarity and accuracy:
    1. We completely reworked the introduction, excluding the controversial statements and also providing a better motivation for our approach and explaining its novelty.
    2. We restructured the main section of the paper so that each section corresponds to a single stage in our framework. In particular, we put all the results which require the power-law assumption into a single section. (Our stability analysis does not require this assumption.)
    3. We moved some of the technical details and derivations into the appendix, and added to the main text more theorems and discussions of the results instead.
- We added new experiments with pretrained ResNet-18 and MobileNet-V2 applied to CIFAR10 and showed that our SE approximation and power-law ansatz work well in these cases too (See section E.5, Figures 9 and 11).
- To make our theory more convincing, we have added several new theoretical results and strengthened or formalized previous results:
    1. Added a new estimate of the impact of non-spectral details (Sections 3 and D.1);
    2. Added an argument that SE approximation with $\tau_1 = \tau_2 = 1$ results if pointwise losses are statistically independent of the eigencomponents (Sections 3 and D.5);
    3. Added Proposition 4, rigorously establishing the necessity and tightness of the simplified convergence condition in terms of the effective learning rate;
    4. Moved Proposition 5 on the effect of momenta into the main text and generalized its "signal-dominated" part to all $0 \le \tau \le 1, 0 < \gamma \le 1$.

3 response to reviewer KUvw

Thank you for your feedback! We believe that we have improved writing in the revised version of the paper. Please also refer to our general comment. Regarding your specific points:

1. We have tried to remove everywhere ambiguous notation (such as $O(N)$) and generally ensure that the meaning of all formulas is clear. We decided not to include a separate section with notation because the paper has much different notation in different sections, and so we find it more efficient to introduce notation along the way.
2. (a) We have completely rewritten the introduction and removed all controversial claims.

    (b) We have tried to restructure the paper to clarify our hierarchy of assumptions. This hierarchy is, in fact, rather simple: 1) Initial linear model; 2) Our SE approximation; 3) Power-law spectral assumptions; 4) Minor auxiliary assumptions such as the one relating individual losses to their cumulative sums (Eq. (28)).

    Different results in our paper require different levels of assumptions. Our first results (Section 3) provide various justifications for the SE approximation from several fundamental perspectives. Next results (Sections 4 and 5) use only SE approximation to solve the SGD dynamics and analyze its stability. Later results (Section 6) provide detailed analysis of various subtle effects and naturally require more detailed (power-law) assumptions. At the same time, we show experimentally that all the assumptions are reasonable and the results agree well with experiment. In the revision we added experiments with CIFAR10 and ResNet/MobileNet, and also tried to better clarify this agreement.

3. We explicitly defined λ_{crit} in Proposition 4 in the new version. We also made definition of the other terms more explicit (e.g. C_{signal} and C_{noise}).
4. (a) We added a new Proposition 7 and a remark in the main text clarifying a potential effect of non-spectral details. Roughly speaking, we show that, in general, non-spectral details do not allow us to predict the mean loss under SGD.
   (b) At the same time, our main point is that in many cases we can still neglect non-spectral details (thus obtaining the SE approximation (7)). In the revised version we give an extended list of 5 theoretical and empirical arguments supporting this claim.
5. This is an interesting question. Actually, the role of (8) in our paper was not to approximate (5), but rather to consider a related (more general) family of noise terms and use it to gain a better understanding of the SE approximation. We slightly changed the wording in the paper to make it more clear. We admit that (8) is more of a mathematical device; we don't claim it to describe any real learning dynamics and at present don't know of a natural interpretation of the cross-terms.
6. As we mention after theorem 2, the loss is actually described by the sum of two terms resulting from the two terms in the expansion of $\frac{d}{dz}\widetilde{L}(z)$ (Eq. (27) in the revised version). These two loss terms can be approximated as $C_{\text{signal}}t^{-\zeta}$ and $C_{\text{noise}}t^{1/\nu-2}$, so that the full loss $L_t \approx C_{\text{signal}}t^{-\zeta} + C_{\text{noise}}t^{1/\nu-2}$. However, at large t one of the terms (depending on the sign of $\zeta - 2 + 1/\nu$) dominates the other, so we can write $L_t \approx \max(C_{\text{signal}}t^{-\zeta}, C_{\text{noise}}t^{1/\nu-2})$. The transition between phases occurs when the dominant term changes.
7. (a) We have completely rewritten the section on positive vs. negative momenta in the main text. We believe that our analysis was actually solid, but the exposition was somewhat convoluted. We have now stated precise Proposition 5 describing the effect, and moved technical details to the appendix.
   (b) Moreover, we have generalized the statement for the "signal-dominated" regime: we show now that $\beta > 0$ are beneficial in this regime for all parameters $0 < \tau \leq 1, 0 \leq \gamma \leq 1$. We believe that this confirms the generality of the effect. As for the improvement from $\beta < 0$ in the "noise-dominated" regime: our goal was rather to show that this can happen at least in some cases, and can be analytically predicted – which we have achieved.

4 response to reviewer LoFN

Thank you for your feedback! We have revised the paper and believe to have improved the writing in various aspects, please refer to our general comment for more details. Some changes addressing your concerns are:

1. "*The introduction largely overstates the importance of least squares.*"

   We believe that the revised introduction does not overstate the importance of least square. We additionally motivate the use of least squares in the context of deep learning by the observation that networks can often be linearized in the late stages of training (Fort et al., 2020).

2. "*...developing better the discussion and highlighting better the main contributions would improve the reading. For a non-expert, it is not very clear what are the challenges and the progress made the manuscript with regard to previous literature.*"

   - We believe that the revised introduction and discussion clearly list multiple advances made by our work.
   - In the Related Work section (A) we provide a detailed and comprehensive discussion of our work in comparison to previous research.

Regarding your questions:

- **Q1.** This is an interesting question.

  1. On the one hand, we definitely see cases when non-Gaussian ($\tau \neq -1$) SE approximation is more appropriate than the Gaussian one. This can be seen theoretically for translation-invariant models or models with pointwise losses independent of feature eigencomponents (the first two items in our current list of evidence for SE in section 3). We also observe empirically that non-Gaussian values $\tau \geq 0$ are more suitable to at least some tasks (this is especially well seen in Figure 7 where we perform averaging to better estimate the mean losses).
  2. On the other hand, we don't see any fundamental difference between the Gaussian and non-Gaussian cases. The exponents in our power-law loss asymptotics do not depend on τ; only the leading coefficients depend on it. The phase structure and other general effects that we discuss, such as the dynamic phase transitions and the effects of momenta, also only include τ through some coefficients but not in any fundamental way (at least in the case $\tau > 0$ that we limit ourselves to for technical reasons).

- **Q2.**

  1. We have substantially revised the paper and removed all controversial claims.
  2. On the other hand, we have added new experiments with ResNet/MobileNet on CIFAR10 that agree well with our theory. We have also tried to better explain what exactly agrees with what (e.g. in Figure 1). We believe that our paper now has a sufficient amount of empirical evidence.

- **Q3.** That was indeed a wrong statement, now removed – thank you for pointing this out.
- **Q4.**

  1. Thank you for bringing that paper to our attention. The setting in that paper is actually very different from ours. We work with a fixed Hessian that is a compact operator with a tail of discrete eigenvalues converging to 0, and a non-random initial condition for SGD. We don't perform any limits involving growing numbers of degrees of freedom and rescaling of the learning rate. It's not obvious how to relate the Marchenko-Pastur distribution to our particular spectral power laws on which our momentum statements depend very sensitively (presumably only the MP with the critical aspect ratio would be suitable for that). So the differences between our settings are so significant that we admit that we find it not easy to establish a correspondence between our results.
  2. In our setting, the benefit of momentum is determined not by the presence of spectral gap, but rather by how fast the target expansion coefficients converge to 0 (see e.g. our Figure 4, in which faster convergence corresponds to larger $\varkappa$ and lower optimal β). One can probably say that the strongly convex case corresponds to eigencoefficients vanishing for very small eigenvalues – in this respect our conclusions roughly agree.

5 response to reviewer NWpJ

Thank you for the positive evaluation of our work! We are especially happy that you appreciated our explicit convergence condition. This is indeed one of our main contributions, as it shows in a simple way the fundamental difference between mini-batch SGD (where effective learning rate is bounded) and full-batch GD (where it is unbounded).

6 Further comments for reviewers

Dear reviewers, we noticed that there are a few features of our results which are not highlighted or commented in the paper. Thus, we decided to post such comments here

- **Experimental confirmation.** One of the main validations of SE approximation is the comparison with experiments on realistic data (MNIST,CIFAR10). In this regard, we believe that comparison of convergence regions (Fig. 2 for MNIST and Fig. 11 for CIFAR10) is the most illustrative comparison, as we observed a fundamental difference SGD and GD in terms of convergence condition. For GD the effective learning rate α_{eff} is unbounded, which enables asymptotic acceleration of GD by letting $\beta \to 1$. In contrast, for SGD the α_{eff} is bounded, which prevents unlimited asymptotic acceleration. On Figs. 2 and 11 we see not only qualitative correspondence: restriction on α_{eff} is present for both SE dynamics and basic training of linearized model, but also quantitative correspondence: actual convergence boundary essentially coincides with theoretical condition $\widetilde{U}(1) < 1$.

- **Value of obtained spectral expressions.** It is possible to criticize our approach as being non-rigorous, as main body of the work is devoted to analysis of SGD under SE approximation. However, we believe that SE approximation may be somewhat exhaustive, as it is in a certain sense (Theorem 1) contains all the cases where observables (train loss, critical learning rates, ...) can be expressed in terms of only spectral distributions (hence the name "Spectrally Expressible (SE)"). This makes our obtained spectral expressions (e.g. for generating functions $\widetilde{U}(z), \widetilde{V}(z)$) exhaustive in the same sense. As (to the best of our knowledge) our obtained spectral expressions are new, we believe that there is a value in just writing them, irrespective of the method or assumptions used for derivation. Let us make an analogy with spectral expression for test risk in kernel ridge regression (see e.g. Bordelon 2020). A significant series of works (Jacot20,Canatar20,Hastie19,Wei22,Simon21, and more) was built around this spectral expression, where it was either derived under various settings/methods/assumptions, or analyzed to explain different phenomena. Retrospectively, such attention puts a lot of value into the spectral expression itself.

- **Power-law assumption.** A portion of our results was obtained under an assumption of power-law spectral distributions. While it may look restrictive (even taking into account that many practical problems approximately have power-law distributions), let us present an alternative perspective on this assumption. First, note that spectral expressions (including ours and e.g. ridge regression risk estimates Bordelon 2020) usually operate sums of the type $\sum_k g(\lambda_k, c_k)$ where $g(\cdot,\cdot)$ is some elementary function. This makes spectral expressions particularly convenient for theoretical analysis. But spectral expressions especially shine when they are coupled with macroscopic distribution laws for λ_k, c_k - this allows to easily evaluate the sums $\sum_k g(\lambda_k, c_k)$ and obtain closed form expressions for the quantities of interest (e.g. our loss asymptotic Eq. (28)). Now, we can say that the alternative perspective on our power-law assumption is to showcase how our spectral expressions can be coupled with specific distribution laws. With this illustration, one can use other distribution laws with our spectral expressions and obtain some explicit results in similar way.

References