MEAN–FLUCTUATION DYNAMICS AT THE EDGE OF STABILITY

Anonymous authors
Paper under double-blind review

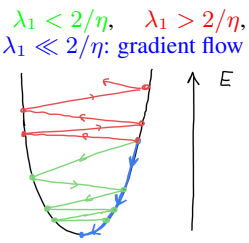
ABSTRACT

We study the dynamics of gradient descent in the Edge of Stability regime, where the learning rate is large enough to induce persistent oscillations in the trajectory, which has been linked to better generalization performances. We introduce the *mean–fluctuation* dynamics, a tractable continuous-time model coupling the window-averaged trajectory to its fluctuation covariance. Among our contributions, we rigorously derive this model from gradient descent in a sharp valley framework, characterize its stationary states and their linear stability, and establish precise connections with other effective dynamics. Numerical experiments illustrate these predictions and their finite-time limitations. We also study our model in the over-parametrized, fixed learning rate, regime of wide two-layer network, where we rigorously derive a kinetic equation describing weights and their fluctuations as a Wasserstein-2 gradient flow, for which we prove well-posedness, mean-field limit, and conditional convergence results.

1 INTRODUCTION

Many theoretical works on machine learning concern convergence guarantees of gradient descent (GD) based algorithms in the idealized regime of small learning rate. Such a regime is mathematically appealing because it reduces to the (already complex) study of gradient flows, for which we have a vast theory (Jordan et al., 1998; Otto, 2001; Ambrosio et al., 2008). However, practitioners usually operate in regimes where the learning rate is not arbitrarily small with respect to the other parameters (LeCun et al., 1998), a practice which has been associated with better generalization properties (Li et al., 2019; Lewkowycz et al., 2020). Yet, the implicit bias generated by such a procedure remains unclear.

The regime of optimization we are interested in is therefore the so-called *edge of stability* (EoS) regime (Cohen et al., 2021), in which the largest Hessian eigenvalues hover around the known stability threshold $2/\eta$, where η is the learning rate. As described by Damian et al. (2023), GD iterates dynamically stabilize around this threshold through the following mechanism: the sharpness first increases until it reaches the threshold $2/\eta$ ("progressive sharpening"), at which point GD jumps out of sharp wells to go back to flatter regions ("curvature-related drift"). Once in this region, the sharpness increases again as in the first stage.



Our goal is to derive a mathematical framework that is both amenable to analysis and quantitatively faithful of GD in this edge of stability regime, at least for a non-negligible timescale. Specifically, our model is a –rather simple yet rich– coupled ordinary differential equation (ODE) for window-averaged parameters and the covariance of their oscillations. We therefore call our model *mean–fluctuation* (MF) dynamics. Its distinguishing feature is an explicit covariance evolution, obtained from first principles, that allows several oscillating directions and unstable Hessian eigenvalues without imposing instantaneous covariance equilibration.

Edge of stability and implicit bias. Flat minima have long been associated with better generalization (Hochreiter & Schmidhuber, 1997), whether they are dynamically reached by some implicit bias (e.g. using small batches of data (Keskar et al., 2017)) or hard-coded in the training procedure (Foret

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

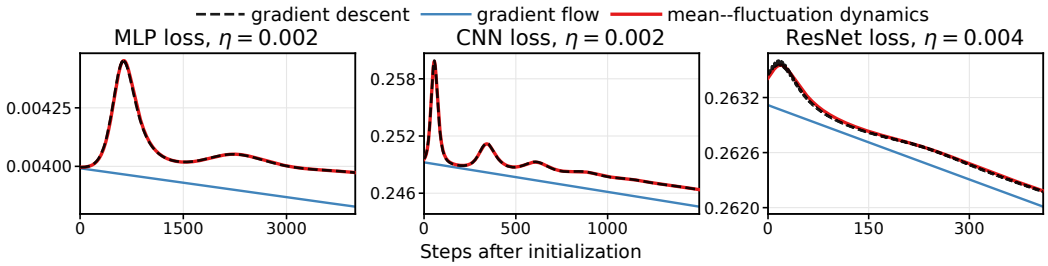


Figure 1: One EoS eigenvalue: our flow agrees with the GD iterates across architectures (Figure 14).

et al., 2021). Taking large learning rates is therefore believed to favor flatter regions (Lewkowycz et al., 2020) by driving the dynamics at the edge of stability (Cohen et al., 2021; Arora et al., 2022; Damian et al., 2023). Further analyses address less general situations in order to characterise as much as possible the edge of stability mechanism (Zhu et al., 2022; Chen & Bruna, 2023).

Continuous-time models. The idea of tracking the time-averaged position of parameters has been used in Cohen et al. (2025), where the covariance of the oscillations is importantly hard-coded (determined through a complementary problem at each time), which does not capture the transient dynamics of the unstable modes and the aforementioned stabilization mechanism, but is indeed a very good model for the coarse GD behavior over longer timescales of optimization.

Perhaps closer to our model, the Edge flow (Marion, 2026) and Rod flow (Regis & Chewi, 2026b;a) –of which we were initially unaware– were developed simultaneously as the first draft of this paper. We will detail the differences between all these flows in Section 2.2. Our model has the advantage of being mathematically analyzable while being faithful to gradient descent at the Edge of Stability.

Mean-field descriptions. Mean-field limits allow to describe the training dynamics of wide two-layer networks by transport equations for neuron distributions (Mei et al., 2018; Chizat & Bach, 2018; Rotskoff & Vanden-Eijnden, 2018; Mei et al., 2019). Yet, these models concern non-oscillating regimes of optimizations. We show that a similar overparametrized limit can be obtained for some oscillatory regimes (namely the "flat landscape limit"; Proposition A.14), yielding a partial differential equation on the space of averaged parameters and their oscillation covariance factors (see Section 4). Importantly, GD oscillations need not survive for long times as width increases in this regime.

Main contributions. We derive our new model from gradient descent over $O(\eta^{-1})$ iterations under sharp-valley assumptions (Theorem 2.1). We then classify stationary covariances and their linear stability in the locally convex regime (Theorem 2.3), and quantify the relation of our model to the Rod flow (Regis & Chewi, 2026b) near Edge of Stability (Proposition 2.4). Finally, for a bounded-kernel two-layer model, we obtain a kinetic PDE which we rigorously derive (Theorem 4.2) and analyse its Wasserstein gradient-flow structure and conditional long-time behavior (Theorem B.10). Experiments test the finite-time predictions and limitations of the mean–fluctuation model, including cases with several near-EoS Hessian modes, which are the trickiest to follow.

2 Mean–fluctuation dynamics

At the edge of stability, gradient descent iterates can be decomposed into a slow mean and fast oscillations. The *mean–fluctuation* (MF) dynamics couples the slow covariance of these fast oscillations to the mean trajectory.

2.1 Model derivation, stability analysis, and landscape asymptotics

We start from the gradient descent (GD) update rule $\tilde{\theta}_{t+1} = \tilde{\theta}_t - \eta \nabla E(\tilde{\theta}_t)$, with learning rate $\eta > 0$. Since we are interested in the slow mean and its coupling to the slow covariance of the fluctuations, we introduce an even window length $m \geq 2$, and define, for $t \geq m-1$,

$$\theta_t := \frac{1}{m}\sum_{j=0}^{m-1} \tilde{\theta}_{t-j}, \quad \Sigma_t := \frac{1}{m}\sum_{j=0}^{m-1} (\tilde{\theta}_{t-j} - \theta_t)(\tilde{\theta}_{t-j} - \theta_t)^T. \tag{1}$$

For any family $(a_j)_{j=0}^{m-1}$, write $\langle a \rangle_m := m^{-1} \sum_{j=0}^{m-1} a_j$ and define the fast oscillations by $\xi_{t,j} := \tilde{\theta}_{t-j} - \theta_t$, for $j \in \{0, \dots, m-1\}$. Every iterate in the window can then be written *exactly* as $\tilde{\theta}_{t-j} = \theta_t + \xi_{t,j}$, where $\langle \xi_t \rangle_m = 0$ and $\langle \xi_t \xi_t^T \rangle_m = \Sigma_t$. Since each $\tilde{\theta}_{t-j}$ advances to $\tilde{\theta}_{t+1-j}$ under GD, we have exactly

$$\theta_{t+1} = \theta_t - \eta \, \langle \nabla E(\theta_t + \xi_t) \rangle_m \,, \tag{2}$$

$$\xi_{t+1,j} = \xi_{t,j} - \eta \left(\nabla E(\theta_t + \xi_{t,j}) - \langle \nabla E(\theta_t + \xi_t) \rangle_m \right). \tag{3}$$

In particular, $\langle \xi_{t+1} \rangle_m = 0$ and $\langle \xi_{t+1} \xi_{t+1}^T \rangle_m = \Sigma_{t+1}$ exactly, with the same average over j. Taylor expanding in ξ as in Appendix A.2, we formally arrive at the *mean–fluctuation* (MF) dynamics:

$$\begin{cases} \dfrac{d\theta}{dt} = -\eta \nabla E(\theta) - \dfrac{1}{2} \eta \nabla H(\theta) : \Sigma, \qquad H(\theta) = \nabla^2 E(\theta), \\ \dfrac{d\Sigma}{dt} = -\eta (\Sigma H(\theta) + H(\theta) \Sigma) + \eta^2 H(\theta) \Sigma H(\theta), \end{cases} \tag{4}$$

where $A : B \equiv \mathrm{tr}(AB^T)$ is the Frobenius inner product (throughout the article, unsubscripted matrix norms will denote the Frobenius norm) and $(\nabla H : \Sigma)_i = \mathrm{tr}((\partial_i H)\Sigma)$.

To account for the contribution of fluctuations to the loss, we introduce the effective free energy $F(\theta, \Sigma) := E(\theta) + \frac{1}{2} H(\theta) : \Sigma$, corresponding to the second-order Taylor approximation of the window-averaged loss. The MF mean equation then reads $\dot{\theta} = -\eta \nabla_\theta F(\theta, \Sigma)$.

Under a more restricted "sharp-valley" framework (where sharp directions are orthogonal to the drift: see Berthier (2026) which uses this framework to formally derive the Central flow), we can actually prove that the mean–fluctuation dynamics indeed describes gradient descent on algorithmic timescales of order $t = O(\eta^{-1})$. To our knowledge, Theorem 2.1 is the first rigorous approximation of GD at the EoS by a continuous-time model that tracks both the central trajectory and the oscillation covariance.

Theorem 2.1 (Derivation of the mean–fluctuation model in a sharp valley)**.** *Let $a \in (0,1)$ denote the scale of $\xi_{t,j}$, and assume $\delta := a^2/\eta \in (0,1)$. Let $E_\eta \in C^4(\mathbb{R}^{d+r})$, with $\|D^3 E_\eta\|_{\infty,\mathrm{op}} + \|D^4 E_\eta\|_{\infty,\mathrm{op}} \le C/\eta$. Fix orthogonal coordinates $\theta = (\theta_1, \theta_2) \in \mathbb{R}^d \times \mathbb{R}^r$, and a constant $C_0 \ge 1$ independent of η, a. For $|\theta_1| \le \delta/C_0$, assume*

$$\nabla_2 E_\eta(\theta_1, 0) = 0, \ \left\| D_2^2 E_\eta(\theta_1, 0) - \frac{2}{\eta} \mathrm{Id}_r \right\|_{\mathrm{op}} \le C, \ |\nabla_1 E_\eta(\theta_1, 0)| \le C\delta, \ \|D_1^2 E_\eta(\theta_1, 0)\|_{\mathrm{op}} \le C.$$

Start GD with $|\tilde{\theta}_{1,0}| \le \delta/C_0$, writing $\tilde{\theta}_k = (\tilde{\theta}_{1,k}, \tilde{\theta}_{2,k})$. Assume $(\tilde{\theta}_{1,0} + \tilde{\theta}_{1,1})/2 = 0$ and $|(\tilde{\theta}_{2,0} + \tilde{\theta}_{2,1})/2| \le Ca^2$, with $|\tilde{\theta}_{2,0}| \le Ca$. Let (θ_k, Σ_k) be the windowed quantities defined in (1) *and let $(\theta(t), \Sigma(t))$ solve* (4)*, initialized at (θ_0, Σ_0). There exist $T_*, C_* > 0$, independent of η, a, m, such that for sufficiently small $m(\eta + \delta)$ both processes exist on the required interval and*

$$\sup_{0 \le t \le T_*/\eta} \left(\frac{|\theta_{\lfloor t \rfloor} - \theta(t)|}{\delta} + \frac{\|\Sigma_{\lfloor t \rfloor} - \Sigma(t)\|_F}{a^2} \right) \le C_* m(\eta + \delta). \tag{5}$$

In Appendix A.1, we comment on the scaling and assumptions considered here and prove the theorem. The key part of the proof is to control the remainders on the long timescale $t = O(\eta^{-1})$, which is done in Lemma A.4. For this, we exploit the stabilization mechanism of gradient descent at the Edge of Stability, which we could manage thanks to the sharp valley framework. In this regime of established oscillations, the fluctuation covariance is rank one at leading order (Equation (41)) as observed in our MLP runs (Appendix A.1, after Corollary A.7). By contrast, near the onset of EoS, the window covariance is dominated by the drift, and the full covariance of MF is what allows the small critical component to grow (Section 2.2 and Figure 5).

The advantage of our model (4) is that it both captures the Edge of Stability behavior and is amenable to the following mathematical analysis. Some of the statements and all proofs of our results are given in Appendix A.

Well-posedness, positivity, and dissipation (Proposition A.9). Under bounded second and third derivatives of the loss, the mean–fluctuation equations Equation (4) possess a unique, global solution. The covariance matrix Σ_t is symmetric positive semidefinite for all times, and the free energy $F(\theta, \Sigma) := E(\theta) + 1/2 H(\theta) : \Sigma$ is decreasing along the flow in stable regions.

Remark 2.2 (Rank of the fluctuation covariance). The MF dynamics does not preserve the rank of Σ. Indeed, the counter-example in Appendix A.5 starts from a rank-one covariance which becomes full rank immediately. Rank is preserved in what we call the "flat-landscape limit" (see Proposition A.14 and Section 4 below), but not by the coupled MF system in general.

Stability/instability of stationary states (Theorem 2.3). We also analyse the stability of stationary states of the MF dynamics. Specifically, a stationary state is *stable* if the spectrum of the linearized operator is contained in the closed left half-plane, *strictly stable* if it is contained in the open left half-plane, and *unstable* if it has an eigenvalue with positive real part.

Theorem 2.3 (Stationary covariance and EoS stability)**.** *Let $E \in C^4$ be locally convex near a stationary MF state (θ_*, Σ_*) with $\Sigma_* \geq 0$. Set $H_* = H(\theta_*)$, $V_0 = \ker H_*$ and $V_2 = \ker(2\mathrm{Id} - \eta H_*)$, with orthogonal projections P_0, P_2.*

If $\Sigma_* = 0$, then $\nabla E(\theta_*) = 0$; the state is stable exactly when $0 \leq \eta H_* \leq 2\mathrm{Id}$, and strictly stable exactly when $0 < \eta H_* < 2\mathrm{Id}$. For $\Sigma_* \neq 0$, we have the decomposition $\mathrm{Ran}\,\Sigma_* \subset V_0 \oplus V_2$, and the orthogonality condition $P_0 \Sigma_* P_2 = 0$. Thus the fluctuations in flat and EoS directions are decorrelated, and $(\theta_*, P_2 \Sigma_* P_2)$ is also stationary.

For a nonzero covariance supported in V_2, assume additionally that $\nabla^2_\theta F(\theta_*, \Sigma_*) \geq 0$. If all eigenvalues of $\eta H_*|_{V_2^\perp}$ lie in $[0, 2)$, the state is stable. An eigenvalue larger than 2 produces an exponentially growing covariance perturbation and hence instability.

Sketch of proof. The proof goes through the linearization of the stationary equations around (θ_*, Σ_*) and is split into three Lemmas. Lemma A.11 addresses the decomposition $\mathrm{Ran}\,\Sigma_*$ into flat and oscillating modes, which allows to reduce the analysis for $\Sigma_* \neq 0$ to purely oscillating cases, i.e. $\mathrm{Ran}\,\Sigma_* \subseteq V_2$. Lemma A.12 proves the no oscillating case $\Sigma_* = 0$. Finally, Lemma A.13 concerns the oscillating states $\Sigma_* \neq 0$, for which we introduce a Lyapunov function to show linear stability. □

In Appendix A.5, we make the remark that under a modified Polyak-Łojasiewicz inequality, the free energy $F(\theta, \Sigma) := E(\theta) + 1/2 H(\theta) : \Sigma$ converges at exponential rate.

Gradient-flow and flat-landscape limits (Proposition A.14). There are two interesting asymptotic regimes for the MF dynamics. The first one is the gradient-flow limit, where the learning rate η is small and the covariance is of order η (which is expected from scaling analysis, see Appendix C.1). The second one is the flat-landscape limit, where the energy is scaled by a small parameter ε. In both cases, the MF dynamics reduces to a simpler system of equations. In the rescaled time $\tau = \eta\varepsilon t$, the flat-landscape limit is the gradient flow of F for the product of the Euclidean metric on θ and the Bures–Wasserstein metric on covariance matrices (Malagò et al., 2018). Its covariance evolves as $\Sigma_\tau = P_\tau \Sigma_0 P_\tau^T$ with P_τ invertible, and hence preserves rank throughout its existence interval.

2.2 Comparison to other models

In this section, we compare our mean–fluctuation model to three other models: the Central flow (Cohen et al., 2025), the Edge flow (Marion, 2026) and the Rod flow (Regis & Chewi, 2026b). We focus on the Rod flow model, which we show is the only one –together with our model– to capture the *transient* fluctuation covariance dynamics when several Hessian eigenvalues are unstable. In the asymptotic validity regime of the mean–fluctuation model, we show that the covariance evolutions of Rod flow and our model are equivalent when restricted to rank-one Σ (Proposition 2.4). We also provide an explanation for the comparative robustness of the mean–fluctuation model to Rod flow's failure to grow oscillations when initialized at the onset at EoS.

Central flow (Cohen et al., 2025). The Central flow is a coarser description than our model. Cohen et al. (2025) decompose the discrete trajectory $\tilde{\theta}_k$ into a slow central path θ_k and fast oscillations $\delta\theta_k$. The mean path θ satisfies the same equation as for our mean–fluctuation model, while their covariance matrix Σ is hard-coded through the hypothesized relation $\mathrm{Ran}\;\Sigma \subseteq \mathrm{Ker}\,(\nabla^2 E(\theta) - 2/\eta\,\mathrm{Id})$ at each time, which assumes the instantaneous relaxation to an equilibrium value that balances the "progressive sharpening" with the "curvature-related drift". In particular, the Central flow is not designed to capture the transient, self-stabilizing, fluctuations of the loss and unstable

Hessian eigenvalues, contrary to our mean–fluctuation model. As shown in Figure 6, our model can nevertheless relax towards the Central flow on larger times of optimizations, which is consistent with the idea that Central flow is a coarser description of gradient descent.

Edge flow (Marion, 2026). The Edge flow model is another model for the Edge of Stability dynamics where the equations are coupling (θ, u, x), where u is the oscillation direction and x its amplitude. Contrary to the Central flow, Rod flow and our MF model, its derivation is not exactly based on first principles, but rather on the assumption that the oscillations are following the top eigenvector of the Hessian by a Rayleigh quotient gradient flow. In our experiments, Algorithm 1 of Marion (2026) seems unable to capture the Edge of Stability dynamics when several Hessian eigenvalues reach the stability threshold.

Rod flow (Regis & Chewi, 2026b;a). The derivation of Rod flow can be seen at first as the restriction of our model to the case of a window of size $m = 2$ (therefore restricting the derivation of their model to a rank-one covariance). In the numerical implementation of the Rod flow, the authors get rid of the constraint that Σ should be rank-one by instead *defining* the rod extent ξ as the top direction in the SVD of Σ. This allows them to obtain stronger numerical results, but it is not clear how to justify this from the point of view of the derivation of the model. In contrast, our model represents fluctuations in multiple directions from the start, but makes an additional closure assumption which makes it less precise than Rod flow in general, yet mathematically analyzable. The Rod flow equations are given by

$$\begin{aligned}\frac{d\theta}{dt} &= -\frac{\eta}{2}\big(\nabla E(\theta+\xi) + \nabla E(\theta-\xi)\big) - \frac{\eta^2}{8}\big(H(\theta+\xi) + H(\theta-\xi)\big)\big(\nabla E(\theta+\xi) + \nabla E(\theta-\xi)\big),\\ \frac{d\Sigma}{dt} &= \frac{\eta^2}{4}\big(\nabla E(\theta+\xi)\otimes\nabla E(\theta+\xi) + \nabla E(\theta-\xi)\otimes\nabla E(\theta-\xi)\big) - 2\Sigma,\end{aligned} \tag{6}$$

where ξ is obtained from the SVD of Σ. Arguably, our model is simpler to analyse than the Rod flow equations above, but this is because they are not exactly the right models to compare yet. Taylor expansion in ξ leads to the following *coarse Rod flow*:

$$\frac{d\theta}{dt} = -\eta\nabla E(\theta) - \frac{\eta}{2}\nabla H(\theta) : \xi\xi^T, \qquad \frac{d\Sigma}{dt} = \frac{\eta^2}{2}H(\theta)\xi\xi^T H(\theta) - 2\Sigma. \tag{7}$$

The following proposition quantifies this reduction and shows that the coarse Rod and MF vector fields, denoted by $\mathcal{V}_{\mathrm{R}}$ and $\mathcal{V}_{\mathrm{MF}}$, are equivalent for small fluctuations around the EoS for rank-one oscillations. The proof of this proposition is in Appendix A.9.

Proposition 2.4 (Coarse reduction of Rod flow and comparison with MF)**.** (i) Raw-to-coarse reduction. *Fix $T > 0$ and $\eta, a \in (0,1)$ such that $\delta := a^2/\eta \in (0,1)$. Let (θ_t, Σ_t) solve* (6) *on $[0, T/\eta]$, with $E = E_\eta$ of class C^5, satisfying $|\nabla E_\eta(\theta_t)| \le C\delta$, $C^{-1}a^2 \le |\xi_t|^2 \le Ca^2$, and $\eta\|D^j E_\eta\|_{\mathrm{op}} \le C$ for $2 \le j \le 5$, with $C \ge 1$ independent of η, a, t. For all sufficiently small η,*

$$\max\left\{\int_0^{T/\eta}\left|\frac{d\theta_t}{dt} - \mathcal{V}_{\mathrm{R}}^{\theta}(\theta_t,\Sigma_t)\right| dt,\quad \frac{1}{a^2}\int_0^{T/\eta}\left\|\frac{d\Sigma_t}{dt} - \mathcal{V}_{\mathrm{R}}^{\Sigma}(\theta_t,\Sigma_t)\right\|_F dt\right\} \le C_T\delta. \tag{8}$$

Here C_T depends only on C, T.

(ii) Comparison with MF. *At any common state (θ, Σ) with $\Sigma \ge 0$, put $D = \eta H(\theta) - 2\mathrm{Id}$. Then $\mathcal{V}_{\mathrm{MF}}^{\Sigma} - \mathcal{V}_{\mathrm{R}}^{\Sigma} = \frac{1}{2}D\Sigma D + \frac{\eta^2}{2}H(\theta)(\Sigma - \xi\xi^T)H(\theta)$. If P_Σ is the orthogonal projection onto $\operatorname{Ran}\Sigma$, then $\|DP_\Sigma\|_{\mathrm{op}} \le \varepsilon$ implies $\|\mathcal{V}_{\mathrm{MF}}^{\Sigma} - \mathcal{V}_{\mathrm{R}}^{\Sigma}\|_F \le \frac{\varepsilon^2}{2}\|\Sigma\|_F + \frac{(2+\varepsilon)^2}{2}\|\Sigma - \xi\xi^T\|_F$.*

The onset failure mode of Rod flow. When initialized substantially before the Edge of Stability regime, the Rod flow fails to stabilize at EoS and follows the gradient flow instead (Marion, 2026). In Appendix C.2, we reproduce this failure even starting at $\eta\lambda_1 \simeq 1.999$, whereas MF recovers the growing fluctuations. We provide an explanation which is that, at the onset of EoS, the mean–fluctuation dynamics can actually grow oscillations along the top Hessian eigenvector even when this covariance component is not dominant, contrary to Rod flow, which retains a rank-one flavor by defining the extent ξ from the top SVD direction of Σ.

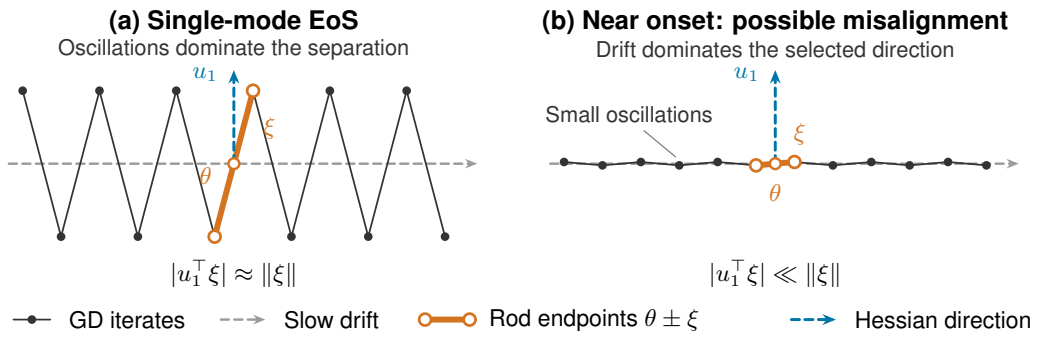


Figure 2: Rod's onset failure. The selected displacement is $\xi = \sqrt{s_1} v_1$, where (s_1, v_1) is the leading eigenpair of Σ. (a) At single-mode EoS, the oscillations are along the leading Hessian eigenvector u_1 (Marion, 2026, see also Fig. 2). (b) Near onset, the selected direction can instead be nearly orthogonal to u_1, even if the full covariance contains smaller fluctuations along u_1. MF can amplify these smaller fluctuations.

Raw counterpart of the MF dynamics: slow evolution hypothesis of the covariance. Finally, we compare MF to the following raw version of our model (dubbed *raw mean–fluctuation dynamics*):

$$\frac{d\theta}{dt} = -\eta \nabla E(\theta) - \frac{\eta}{2} \nabla H(\theta) : \Sigma_t, \quad \Sigma_{t+1} = (\mathrm{Id} - \eta H(\theta_t)) \Sigma_t (\mathrm{Id} - \eta H(\theta_t)). \tag{9}$$

This will numerically test the closure assumption together with the continuous approximation of the mean to establish the importance of the continuous approximation of the covariance evolution. In the numerical experiments, this model is actually better than Equation (4). Nevertheless, this is not necessarily the case: Equation (4) can actually be closer to Equation (9) than to gradient descent over long timescales. Indeed, for a common prescribed C^1 Hessian $H(t) \geq 0$, suppose a fixed invariant subspace V supports Σ_0, with $\|(\eta H - 2\mathrm{Id})|_V\|_{\mathrm{op}} \leq \varepsilon$ and $\|\dot{H}|_V\|_{\mathrm{op}} \leq L$. If $\eta \leq \varepsilon^2$, Proposition A.15 bounds the raw/coarse covariance gap by $O(\varepsilon)\|\Sigma_0\|$ over times $O(\varepsilon^{-1})$ as $\varepsilon \to 0$, while a toy example in Appendix A.3 shows that the relative covariance error against GD can already be of order one on this $O(\varepsilon^{-1})$ timescale.

3 Experiments

We validate our model, and compare it to the others, on three architectures (MLP, CNN, and ResNet) trained on a subset of CIFAR-10 (around 3,000–8,000 images). For each plot, the initialization is done at the onset of the first (or second) Hessian eigenvalue reaching the edge of stability. Every flow starts from the same stored GD window. We compare free energy F with GD loss, Hessian eigenvalues with those at the GD-window center, and the relative Euclidean center error $e_\theta(p) = \|\theta_p - \overline{\theta}_{\mathrm{GD}}\|_2 / \|\overline{\theta}_{\mathrm{GD}}\|_2$, where $\overline{\theta}_{\mathrm{GD}}$ is the window mean. Appendix D details the protocol.

The reference runs show close agreement when there is one unstable eigenvalue, for which we give the results in Appendix D.3 (see also Figure 1). MF and Rod have similar MLP errors when there is two unstable modes, while Rod is better for two-mode CNN and Resnet experiments. Raw MF has similar time-validity as Rod on the two-mode CNN experiment. In Appendix D.4, we plot the closure diagnostics for all these runs. Additional runs can be found in Appendix D.5. Since the plotted curves are small-rank numerical approximations, we proceed to a numerical validation in Appendix E.

Figure 5 supports our claim that MF can grow oscillations along the top Hessian eigenvector even when this covariance component is not dominant, while Rod fails to do so because of its remaining rank-one flavor. The full discussion is available in Appendix C.2.

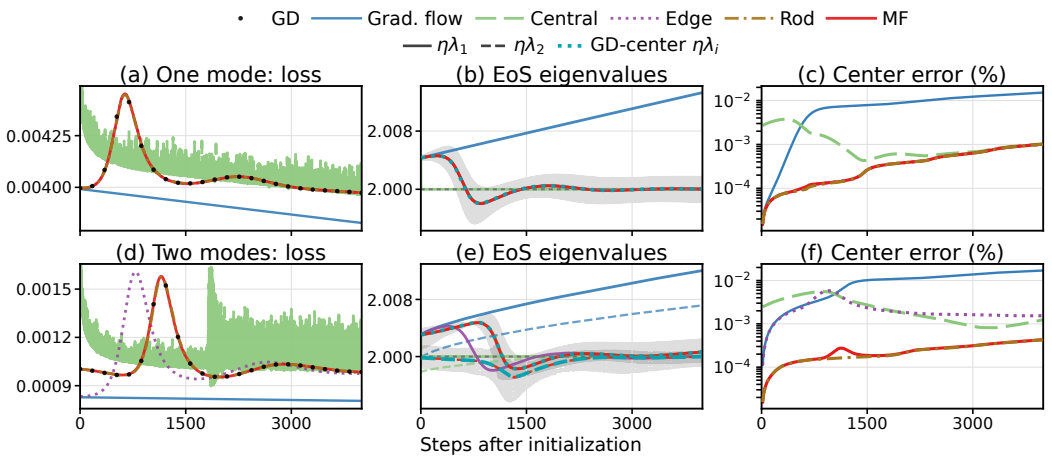


Figure 3: MLP on CIFAR-10, $\eta = 0.002$.

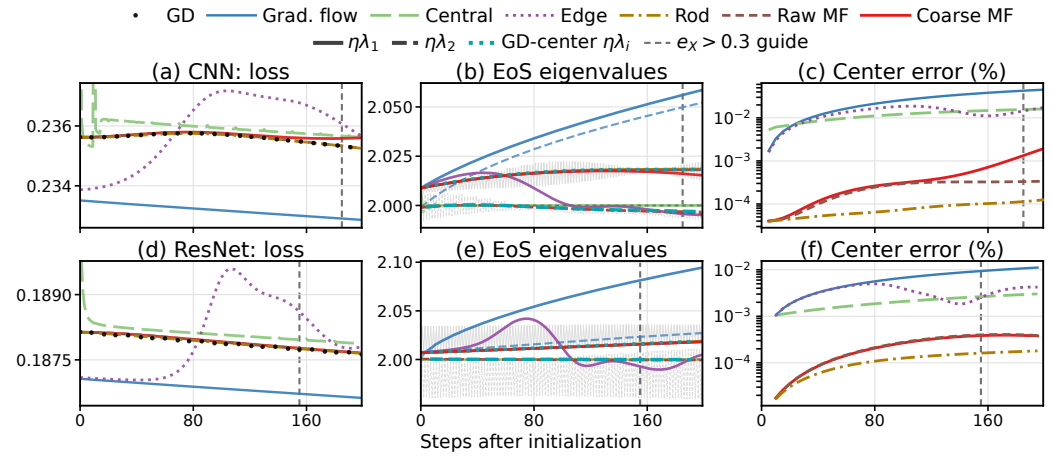


Figure 4: Two-mode CNN ($\eta = 0.002$) and ResNet ($\eta = 0.004$) on CIFAR-10. The vertical dashed line corresponds to accumulated covariance error above 0.3 (see Appendix D.4).

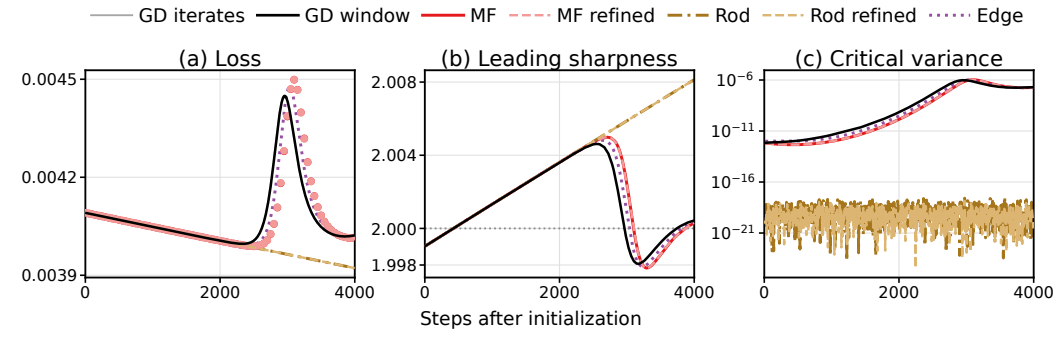


Figure 5: MLP initialized near $\eta\lambda_1 = 1.999$, $\eta = 0.002$. Writing u_1 for the leading GD-center Hessian eigenvector, the critical variance is defined as $|u_1^T \xi|^2$, where ξ is obtained from the SVD of Σ for Rod, and $u_1^T \Sigma u_1$ for Edge and MF. Full protocol: Appendix C.2.

In the longer CNN run, the covariance errors accumulate (Appendix D.4). Nevertheless, an interesting observation is that the MF and Central centers eventually become close to each other (Figure 6).

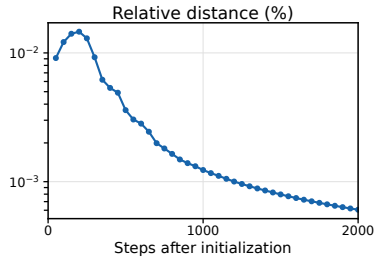


Figure 6: MF–Central relative distance for the two-mode CNN.

4 KINETIC THEORY FOR WIDE TWO-LAYER NETWORKS

Practical machine learning architectures are overparametrized, meaning that the number of its parameters are very large with respect to the number of training samples. We therefore study our model for wide two-layer networks in the mean-field scaling. Specifically, we consider N neurons with weights $(\theta_1, \dots, \theta_N) \in (\mathbb{R}^d)^N$ and the associated empirical measure $\rho_N := N^{-1} \sum_{i=1,\dots,N} \delta_{\theta_i}$. We apply our model to tasks of the form $E(\rho) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell \left(y, \int \phi(\theta, x) d\rho(\theta) \right) \right]$, where $\phi(\theta, x)$ denotes the output of a single neuron with weights θ and input x, and $\mathcal{D}$ is the data distribution. Rescaling time by a factor N/η, Equation (4) writes

$$\begin{cases} \dfrac{d\theta_i}{dt} = -N \nabla_{\theta_i} E(\rho_N) - \dfrac{1}{2} N \nabla_{\theta_i} \nabla^2 E(\rho_N) : \Sigma, \\ \dfrac{d\Sigma_{i,j}}{dt} = -N(\Sigma \nabla^2 E(\rho_N))_{i,j} - N(\nabla^2 E(\rho_N)\Sigma)_{i,j} + \eta N (\nabla^2 E(\rho_N) \Sigma \nabla^2 E(\rho_N))_{i,j}. \end{cases} \tag{10}$$

As we will show in Theorem 4.2, the quadratic term in the covariance equation is of order $O(N^{-1})$ when η is fixed. Therefore, this term disappears as $N \to \infty$, which is the flat landscape regime (Proposition A.14). The remaining part of the work is to show that, starting from $\Sigma^N = V^N (V^N)^T$ with a rank r independent of N, then, on every fixed time interval, the full system remains $O(N^{-1})$-close in block norm to the rank-r system obtained by omitting the quadratic term. The rows $v_i^N \in \mathbb{R}^{d\times r}$ of V^N then live in a fixed space, so their empirical law has an ordinary mean-field limit.

To fix ideas, we can write the equation in the case of a least square loss $\ell(r) = r^2/2$, in the teacher-student framework where $y = \int \phi(\theta, x)\, d\mu(\theta)$ for some target distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$. In the reproducible kernel Hilbert space scenario where we have a translation invariant kernel $\mathsf{g} : \mathbb{R}^d \to \mathbb{R}$ such that $\mathsf{g}(\theta - \theta') = \mathbb{E}_{\mathcal{D}}\big[\phi(\theta, x)\phi(\theta', x)\big]$, the risk functional now reads $E(\rho) = \frac{1}{2} \int \mathsf{g} * (\rho - \mu)\, d(\rho - \mu)$, which is called the Maximum Mean Discrepancy between ρ and μ (Arbel et al., 2019). Abusing notation by denoting $\mu \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^{d\times r})$ the measure $\mu(d\theta) \otimes \delta_0(dv)$, we rigorously obtain (see Theorem 4.2) the closed-form equation

$$\begin{cases} \dfrac{\partial f}{\partial t} - \operatorname{div}_\theta(f \nabla \mathsf{g} * (\rho - \mu)) - \dfrac{1}{2} \operatorname{div}_\theta(f \nabla_\theta(\nabla^2 \mathsf{g} : vv^T) * (f - \mu)) \\ - \operatorname{div}_v(f(\nabla^2 \mathsf{g}\, v) * (f - \mu)) = 0, \qquad \rho := \displaystyle\int f\, dv. \end{cases} \tag{11}$$

We could derive a more general PDE accounting for different losses ℓ, tasks E and data structures. From the point of view of analysis, this equation is the gradient flow for the Wasserstein-2 distance on the space $\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^{d\times r})$, associated with the macroscopic free energy

$$\mathcal{F}(f) := \frac{1}{2} \int \mathsf{g} * (\rho - \mu)\, d(\rho - \mu) + \frac{1}{4} \int (\nabla^2 \mathsf{g} : vv^T) * (f - \mu)\, d(f - \mu). \tag{12}$$

The fluctuation term has no sign, so $\mathcal{F}$ is generically unbounded from below. In Appendix B.1 we prove a uniform lower bound under a velocity second-moment constraint. After adding the uniform integrability needed, we also obtain minimizers and identify their small-variance limit.

Well-posedness and regularity (Theorem B.4). For even kernels $\mathsf{g} \in C_b^{3,1}$, target $\mu \in \mathcal{P}(\mathbb{R}^d)$ and initial data $f_0 \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^{d\times r})$ compactly supported in the v variable, the equation (11) has a unique and global solution, for which the free energy (12) decreases along the flow. It also propagates

the regularity of the initial data, provided regularity assumptions on the kernel (Proposition B.12). The key is to control fluctuations: we prove that f_t remains compactly supported in v along the flow.

Theorem 4.1 (Conditional convergence and rate)**.** *Let* $\mathrm{g} \in C^{3,1}(\mathbb{T}^d)$ *be even, with* $\widehat{\mathrm{g}}(k) \geq 0$ *for all* k *and* $\widehat{\mathrm{g}}(k) > 0$ *for* $k \neq 0$*, and let* $\mu \in \mathcal{P}(\mathbb{T}^d)$*. Let* f_t *solve* (11) *on* $\mathbb{T}^d \times \mathbb{R}^{d\times r}$ *from initial data* $f_0 \in \mathcal{P}_2$ *compactly supported in* v*, with positional marginal* ρ_t*. Assume* $\sup_{t\geq 0} \|\rho_t - \mu\|_{\dot{H}^{-1}(\rho_t)} \leq C_0$ *for some* $C_0 > 0$*, with the weighted dual norm defined in* (112)*. If* $M_2(t) := \int |v|^2 \,\mathrm{d}f_t \to 0$*, then* $W_2(f_t, \mu \otimes \delta_0) \to 0$*. Furthermore, if* $M_2(t) \leq A(1+t)^{-\alpha}$ *for some* $A, \alpha > 0$*, then*

$$E(\rho_t) \leq C(1+t)^{-\min\{1/2,\alpha/3\}},$$

where C *depends only on* $\mathrm{g}, C_0, A, \alpha$ *and* $E(\rho_0)$*.*

Sketch of proof. The proof is given in Theorem B.10 and Corollary B.11. The uniform bound in $\dot{H}^{-1}(\rho_t)$ was already assumed in Arbel et al. (2019) for classical flows. □

Theorem 4.2 (Kinetic limit of mean–fluctuation dynamics)**.** *Fix* $\eta > 0$ *and rank* r *independently of* N*. Let* $\mathrm{g} \in C_b^4(\mathbb{R}^d)$ *be even,* $\mu \in \mathcal{P}(\mathbb{R}^d)$*, and* $f_0 \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^{d\times r})$ *with* $|v| \leq M$ *on its support. Initialize the full MF system* (10) *with* $\Sigma^N(0) = V^N(0)V^N(0)^T$*, where the block rows* $v_i^{N,0}$ *of* $V^N(0)$ *satisfy* $|v_i^{N,0}| \leq M$*, and set* $f_N^0 = N^{-1}\sum_i \delta_{(\theta_i^{N,0}, v_i^{N,0})}$*.*
Let $(\bar{\theta}^N, V^N)$ *use the same mean equation with* $\Sigma = V^N(V^N)^T$ *and* $\partial_t V^N = -NH_N(\bar{\theta}^N)V^N$*, initialized from the same initial state. Set* $\bar{f}_N = N^{-1}\sum_i \delta_{(\bar{\theta}_i^N, v_i^N)}$*. If* f_t *solves* (11) *from* f_0*, then for every fixed* $T > 0$*, we have the mean-field-type result* $\sup_{t\leq T} W_1(\bar{f}_N(t), f_t) \leq C_T W_1(f_N^0, f_0)$*, and*

$$\sup_{t\leq T}\Big(\max_i |\theta_i^N - \bar{\theta}_i^N| + \|\Sigma^N - V^N(V^N)^T\|_{\max}\Big) \leq C_T/N,$$

where C_T *is independent of* N *and* $\|S\|_{\max} = \max_{i,j}\|S_{ij}\|$ *for the* $d \times d$ *covariance blocks. Consequently, the mean–fluctuation marginal law* $\rho_N^{\mathrm{full}} = N^{-1}\sum_i \delta_{\theta_i^N}$ *and the positional marginal* ρ_t *of* f_t *satisfy*

$$\sup_{t\leq T} W_1(\rho_N^{\mathrm{full}}(t), \rho_t) \leq C_T\big(W_1(f_N^0, f_0) + N^{-1}\big). \tag{13}$$

Sketch of proof. The essential step is to reduce to a fixed dimension space for V^N, proving that the quadratic Hessian term in (10) is $O(N^{-1})$. This is obtained by a stability argument: it replaces a $dN \times dN$ state by factors in an N-independent phase space. The proof is in Appendix B.5. □

5 Discussion

The mean–fluctuation dynamics we introduced is rich enough to capture the Edge of Stability retroaction mechanism between progressive sharpness and curvature-related drift (Figures 1, 3 and 4), while still being mathematically analyzable. We rigorously derived it from gradient descent (Theorem 2.1) and analyzed its stationary points and their linear stability, which can be related to the Edge of Stability threshold (Theorem 2.3). In the overparametrized limit, we rigorously derived the first kinetic description of wide two-layer models accounting for oscillations in the kinetic variable (Theorems 4.1 and 4.2). The experiments further identify where this covariance description follows GD and where it loses accuracy (Section 3 and Appendices C.2 and D.5).

Beyond sharp valleys. Our Theorem 2.1 assumes sharp directions orthogonal to the drift, as in the formal Central-flow derivation of Berthier (2026). Extending MF beyond this setting remains open: there is no general reason to discard higher-order covariance terms on timescales of order η^{-1}.

Hydrodynamic theory at the Edge of Stability. Our kinetic PDE describes fluctuations below the stability threshold. Deriving such a PDE in the Edge of Stability regime remains an open question.

Stochastic gradient descent. Extending our model to small-batch gradient descent would be a great benefit.

REPRODUCIBILITY STATEMENT

The anonymous supplement includes the solver code, architecture and data definitions, experiment launchers, saved configurations and initialization metadata, and numerical data and scripts for regenerating the figures without rerunning simulations. Initialization checkpoints are omitted to meet the upload size limit; preparation scripts implement the recorded initialization protocols, but regenerated starting states need not coincide with the historical ones. Experimental protocols are detailed in Appendix D. Solver settings and numerical validation procedures are detailed in Appendix E.

AI USE STATEMENT

LLMs (GPT-5.5, 5.6 Sol and 6.0 Astra) were used to write and run numerical code, to assist with literature search, to check proofs, and to draft parts of proofs (Theorem 2.1, Proposition 2.4, the case $\Sigma_\star \neq 0$ of Theorem 2.3, a first version of Proposition B.12, and Theorem 4.1) and of the text. The authors verified all statements, proofs and numerical results and take full responsibility for the content.

REFERENCES

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Birkhäuser, Basel, 2nd edition, 2008. doi: 10.1007/978-3-7643-8722-8. URL https://doi.org/10.1007/978-3-7643-8722-8.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 948–1024, 2022. URL https://proceedings.mlr.press/v162/arora22a.html.

Raphaël Berthier. The multiple timescales of gradient descent on the edge of stability: A perturbative derivation of the Central Flow. *arXiv preprint arXiv:2609.01034*, 2026. doi: 10.48550/arXiv.2609.01034. URL https://arxiv.org/abs/2609.01034.

W. Braun and K. Hepp. The Vlasov dynamics and its fluctuations in the $1/N$ limit of interacting classical particles. *Communications in Mathematical Physics*, 56(2):101–113, 1977. doi: 10.1007/BF01611497. URL https://doi.org/10.1007/BF01611497.

Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: A review of models, methods and applications. I. models and methods. *Kinetic and Related Models*, 15(6):895–1015, 2022. doi: 10.3934/krm.2022017. URL https://doi.org/10.3934/krm.2022017.

Lei Chen and Joan Bruna. Beyond the edge of stability via two-step gradient updates. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4330–4391. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/chen23b.html.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Lénaïc Chizat, Maria Colombo, Roberto Colombo, and Xavier Fernández-Real. Quantitative convergence of Wasserstein gradient flows of kernel mean discrepancies. *arXiv preprint arXiv:2603.01977*, 2026. URL https://arxiv.org/abs/2603.01977.

Antonin Chodron de Courcel and Matthew Rosenzweig. Wasserstein gradient flows for Coulomb discrepancies. *arXiv preprint arXiv:2607.12579*, 2026. URL https://arxiv.org/abs/2607.12579.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2103.00065.

Jeremy M Cohen, Alex Damian, Ameet Talwalkar, J Zico Kolter, and Jason D Lee. Understanding optimization in deep learning with central flows. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.24206.

Alex Damian, Eshaan Nichani, and Jason D Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *International Conference on Learning Representations*, 2023.

R. L. Dobrushin. Vlasov equations. *Functional Analysis and Its Applications*, 13(2):115–123, 1979. doi: 10.1007/BF01077243. URL https://doi.org/10.1007/BF01077243.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

François Golse. On the dynamics of large particle systems in the mean field limit. In Adrian Muntean, Jens Rademacher, and Antonios Zagaris (eds.), *Macroscopic and Large Scale Phenomena: Coarse Graining, Mean Field Limits and Ergodicity*, volume 3 of *Lecture Notes in Applied Mathematics and Mechanics*, pp. 1–144. Springer, Cham, 2016. doi: 10.1007/978-3-319-26883-5_1. URL https://doi.org/10.1007/978-3-319-26883-5_1.

Martin H. Gutknecht. Block Krylov space methods for linear systems with multiple right-hand sides: An introduction. In A. H. Siddiqi, I. S. Duff, and O. Christensen (eds.), *Modern Mathematical Models, Methods and Algorithms for Real World Systems*, pp. 420–447. Anamaya Publishers, New Delhi, 2007. URL https://people.math.ethz.ch/~mhg/pub/delhipap.pdf.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

Pierre-Emmanuel Jabin. A review of the mean field limits for Vlasov equations. *Kinetic and Related Models*, 7(4):661–711, 2014. doi: 10.3934/krm.2014.7.661. URL https://doi.org/10.3934/krm.2014.7.661.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998. doi: 10.1137/S0036141096303359. URL https://doi.org/10.1137/S0036141096303359.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1609.04836.

Yann LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. Efficient BackProp. In Genevieve B. Orr and Klaus-Robert Müller (eds.), *Neural Networks: Tricks of the Trade*, volume 1524 of *Lecture Notes in Computer Science*, pp. 9–50. Springer, 1998. doi: 10.1007/3-540-49430-8_2. URL https://leon.bottou.org/papers/lecun-98x.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020. URL https://arxiv.org/abs/2003.02218.

Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/bce9abf229ffd7e570818476ee5d7dde-Abstract.html.

Luigi Malagò, Luigi Montrucchio, and Giovanni Pistone. Wasserstein riemannian geometry of Gaussian densities. *Information Geometry*, 1(2):137–179, 2018. doi: 10.1007/s41884-018-0014-4. URL https://doi.org/10.1007/s41884-018-0014-4.

Pierre Marion. Edge Flow: A tractable and predictive continuous-time model for gradient descent at the edge of stability. *arXiv preprint arXiv:2606.18080*, 2026. doi: 10.48550/arXiv.2606.18080. URL https://arxiv.org/abs/2606.18080.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2388–2464, 2019. URL https://proceedings.mlr.press/v99/mei19a.html.

Felix Otto. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26(1–2):101–174, 2001. doi: 10.1081/PDE-100002243. URL https://doi.org/10.1081/PDE-100002243.

Eric Regis and Sinho Chewi. A rod flow model for adam at the edge of stability. *arXiv preprint arXiv:2605.06821*, 2026a. URL https://arxiv.org/abs/2605.06821.

Eric Regis and Sinho Chewi. Rod flow: A continuous-time model for gradient descent at the edge of stability. *arXiv preprint arXiv:2602.01480*, 2026b. URL https://arxiv.org/abs/2602.01480.

Matthew Rosenzweig, Dejan Slepčev, and Lihan Wang. Wasserstein gradient flows of Maximum Mean Discrepancy with energy kernels. *arXiv preprint arXiv:2608.01182*, 2026. URL https://arxiv.org/abs/2608.01182.

Grant Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Yousef Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, Philadelphia, 2nd edition, 2003. doi: 10.1137/1.9780898718003. URL https://doi.org/10.1137/1.9780898718003.

Gilbert Strang. On the construction and comparison of difference schemes. *SIAM Journal on Numerical Analysis*, 5(3):506–517, 1968. doi: 10.1137/0705041. URL https://doi.org/10.1137/0705041.

Xu Zhu, Zhaodong Wang, Xiao Wang, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. *arXiv preprint arXiv:2210.03294*, 2022.

CONTENTS

1 Introduction **1**

2 Mean–fluctuation dynamics **2**

2.1 Model derivation, stability analysis, and landscape asymptotics 2

2.2 Comparison to other models 4

3 Experiments **6**

4 Kinetic theory for wide two-layer networks **8**

5 Discussion **9**

A Proofs and conditional derivation of the mean–fluctuation dynamics **14**

A.1 Derivation of the mean–fluctuation dynamics 14

A.2 Formal derivation of the mean–fluctuation dynamics 25

A.3 Closure validity and a two-cycle counterexample 26

A.4 Well-posedness and positivity 26

A.5 Rank and quantitative dissipation 27

A.6 Stability of stationary states 28

A.7 Small-step and flat-landscape limits 33

A.8 Prescribed-Hessian raw-to-coarse approximation 35

A.9 Raw-to-coarse Rod reduction and comparison with MF 37

A.9.1 From raw Rod flow to coarse Rod flow 37

A.9.2 Comparison with MF 38

B Analysis of the kinetic equation **39**

B.1 Velocity moments and variational structure 39

B.2 Proof of the well-posedness theorem 41

B.2.1 Study of the flow map 42

B.2.2 Well-posedness of the equation 44

B.3 Conditional convergence to the target 46

B.4 Propagation of higher regularity 47

B.5 Proof of the mean-field limit 49

C Additional remarks **53**

C.1 Scaling of the fluctuations 53

C.2 On the onset failure of Rod flow 53

D Additional experimental details **56**

D.1 Experimental protocol and summary 56

D.2 Center error and oscillatory-state distance 57

D.3 One-mode CNN and ResNet experiments 59

D.4 Closure diagnostics for the six architecture/onset experiments 59

D.5 MLP sensitivity to seed and learning rate 63

D.6 Rod-flow solver resolution 66

D.7 Edge-flow timestep and direction-update controls 67

E Numerical solver of the mean–fluctuation dynamics **68**

E.1 Description of the solver 68

E.2 Numerical validation 71

A PROOFS AND CONDITIONAL DERIVATION OF THE MEAN–FLUCTUATION DYNAMICS

A.1 DERIVATION OF THE MEAN–FLUCTUATION DYNAMICS

In this section, we rigorously derive the mean–fluctuation dynamics from gradient descent, identifying its validity regime for the specific case of a sharp valley (Theorem 2.1). In a nutshell, the sharp-valley picture describes a landscape on orthogonal coordinates $\theta = (\theta_1, \theta_2) \in \mathbb{R}^d \times \mathbb{R}^r$ whose valley bottom $\theta_2 = 0$ satisfies $\nabla_2 E_\eta(\theta_1, 0) = 0$, while a small tangential gradient along θ_1 may remain. The curvature along the sharp directions θ_2 is $2/\eta + O(1)$, hence the name "sharp" valley as $\eta \to 0$.

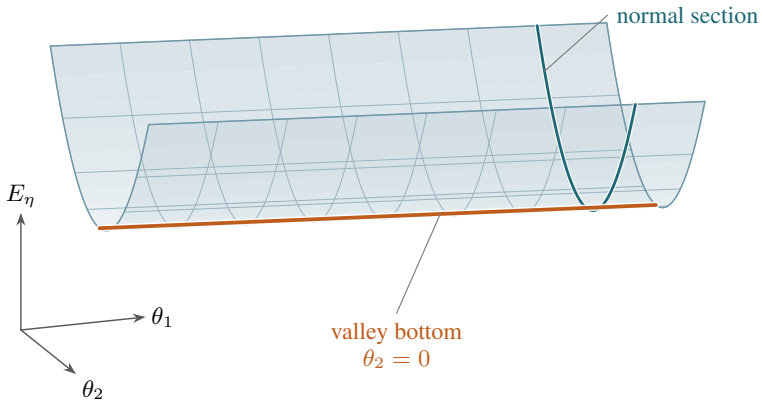


Figure 7: Local schematic of a sharp valley for $d = r = 1$. The highlighted bottom $\theta_2 = 0$ is stationary transversely but may have a tangential slope along θ_1. The transverse curvature is $2/\eta + O(1)$.

Structure of the section. We first make a scaling analysis which then motivates the assumptions needed to prove Theorem 2.1. After this, we start by decomposing the loss landscape into sharp H_η and tangential V_η components (see Lemma A.1). We then prove the main result in the restricted case where $m = 2$ and $V_\eta = 0$ (Proposition A.2), then in the full sharp-valley picture (Theorem A.6) for $m = 2$ and finally improve this to windows of larger sizes $m \geq 2$, therefore proving Theorem 2.1.

Scaling analysis. Before stating the assumptions and commenting on them, we make an informal scaling analysis to help understanding the picture.

The mean–fluctuation equations write, on the algorithmic timescale t:

$$\begin{aligned} \frac{d\theta}{dt} &= -\eta \nabla E_\eta(\theta) - \frac{1}{2}\eta D^3 E_\eta(\theta) : \Sigma, \\ \frac{d\Sigma}{dt} &= -(\eta H\Sigma + \eta \Sigma H) + \eta^2 H\Sigma H, \qquad H = D^2 E_\eta(\theta). \end{aligned} \tag{14}$$

First, we can factorize η in both the mean and covariance equation, so that it is natural to consider the slow timescale τ defined by $\tau = \eta t$. On this timescale, the mean–fluctuation dynamics rewrite

$$\begin{aligned} \frac{d\theta}{d\tau} &= -\nabla E_\eta(\theta) - \frac{1}{2} D^3 E_\eta(\theta) : \Sigma, \\ \frac{d\Sigma}{d\tau} &= -(H\Sigma + \Sigma H) + \eta H\Sigma H, \qquad H = D^2 E_\eta(\theta). \end{aligned} \tag{15}$$

Consider the mean equation. We shall consider a regime where both mean drifts equilibrate:

$$\nabla E_\eta \sim D^3 E_\eta : \Sigma,$$

which yields, if $\eta |D^3 E_\eta| = O(1)$ and denoting $\Sigma \sim a^2$,

$$|\nabla E| \sim \frac{a^2}{\eta} =: \delta.$$

Therefore, δ shall be thought of as the tangential displacement scale over slow times τ of order one. Consider the covariance equation. At the Edge of Stability, and if $\eta \to 0$, one must have $H \sim 2/\eta$, so that we have

$$-(H\Sigma + \Sigma H) \sim \frac{a^2}{\eta} = \delta,$$
$$\eta H\Sigma H \sim \frac{a^2}{\eta} \sim \delta.$$

However, their leading terms cancel. In the sharp subspace, writing $H = 2\mathrm{Id}/\eta + B$ with $\|B\| = O(1)$ gives

$$-(H\Sigma + \Sigma H) + \eta H\Sigma H = B\Sigma + \Sigma B + \eta B\Sigma B \sim a^2.$$

Thus, over $\tau = O(1)$, or $t = O(\eta^{-1})$ iterations, the mean θ can move by $O(\delta)$ while the covariance can change by $O(a^2)$, an amount comparable to its own size.

Normalizing these quantities as $\bar{\theta} = \theta/\delta$ and $\bar{\Sigma} = \Sigma/a^2$ therefore gives order-one bounds on the retained velocities in τ.

The scaling $H \sim 2/\eta$ is also the reason why one only assumes $\eta|D^j E_\eta| \le C$ for $j \ge 2$: uniform η-independent regularity of higher derivatives of E_η would more or less restrict the analysis to quadratic landscapes (derivatives of higher order being negligible with respect to the Hessian).

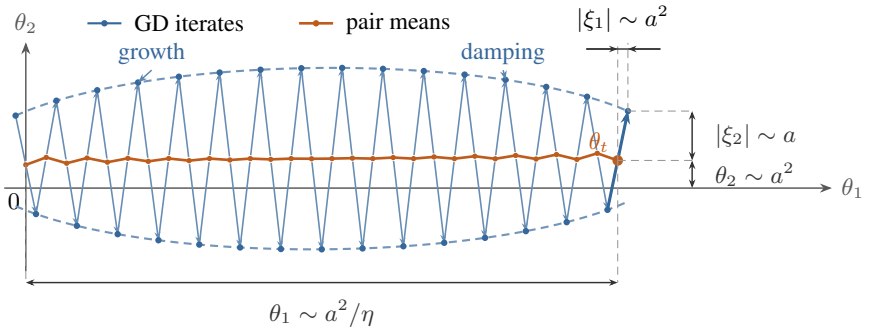


Figure 8: Schematic almost two-period GD oscillations for $m = 2$ and $V_\eta = 0$, with a prepared sharp pair mean. The orange points are consecutive-pair means; the highlighted pair has midpoint θ_t and half-separation ξ_t. The tangential origin is the initial pair mean. Labels indicate upper scales over $O(\eta^{-1})$ iterations, not lower bounds; the drawing is not to scale. The dashed envelope illustrates a possible $O(a)$ change in $|\xi_2|$: growth followed by damping as the sharp curvature decreases through $2/\eta$, for example when a and η are comparable.

Assumptions. Let $\eta \to 0$ and $\delta := a^2/\eta \to 0$, with no further restriction on a/η. Fix orthogonal coordinates $\theta = (\theta_1, \theta_2) \in \mathbb{R}^d \times \mathbb{R}^r$ and a constant $C_0 \ge 1$ independent of η, a. We assume $E_\eta \in C^4(\mathbb{R}^{d+r})$.

For $|\theta_1| \le \delta/C_0$, we assume

$$\nabla_2 E_\eta(\theta_1, 0) = 0, \qquad \left\| D_2^2 E_\eta(\theta_1, 0) - \frac{2}{\eta}\mathrm{Id}_r \right\|_{\mathrm{op}} \le C, \tag{16}$$

$$|\nabla_1 E_\eta(\theta_1, 0)| \le C\delta, \qquad \|D_1^2 E_\eta(\theta_1, 0)\|_{\mathrm{op}} \le C, \tag{17}$$

and, on the whole space, the derivative bound

$$\|D^3 E_\eta\|_{\infty,\mathrm{op}} + \|D^4 E_\eta\|_{\infty,\mathrm{op}} \le C/\eta. \tag{18}$$

Here $\|\cdot\|_{\infty,\mathrm{op}}$ is the supremum of the operator norm over $\mathbb{R}^{d+r}$. Dimensions and constants are independent of η, a and the window length m.

Start GD with $|\tilde{\theta}_{1,0}| \le \delta/C_0$ and $\tilde{\theta}_{k+1} = \tilde{\theta}_k - \eta \nabla E_\eta(\tilde{\theta}_k)$. After translating the tangential origin, assume the first pair satisfies

$$\frac{\tilde{\theta}_{1,0} + \tilde{\theta}_{1,1}}{2} = 0, \qquad \left| \frac{\tilde{\theta}_{2,0} + \tilde{\theta}_{2,1}}{2} \right| \le Ca^2, \qquad |\tilde{\theta}_{2,0}| \le Ca. \tag{19}$$

Let $m \ge 2$ be even, with $m(\eta + \delta)$ sufficiently small; for convergence we require $m(\eta + \delta) \to 0$. Define the actual GD window moments by

$$\theta_k = \frac{1}{m} \sum_{j=0}^{m-1} \tilde{\theta}_{k+j}, \qquad \Sigma_k = \frac{1}{m} \sum_{j=0}^{m-1} (\tilde{\theta}_{k+j} - \theta_k)(\tilde{\theta}_{k+j} - \theta_k)^T,$$

and initialize MF at (θ_0, Σ_0). For $m > 2$, this window mean need not equal the prepared first-pair mean in (19). Forward windows are the moving averages of Section 2.1, with the time origin shifted to the first complete window.

We use slow time $\tau = \eta t$, where t is iteration time. We write $(\theta(\tau), \Sigma(\tau))$ for MF in this time variable and Σ_{ij} for the covariance blocks in the coordinate split. The proof constructs a slow-time horizon $T_* > 0$ independent of η, a, m in which GD and MF asymptotically coincide.

Decomposition of the loss landscape. The following lemma proves a useful decomposition of the loss landscape of the form

$$E_\eta(\theta_1, \theta_2) = V_\eta(\theta_1) + \frac{1}{2}\theta_2^T H_\eta(\theta_1)\theta_2 + R_\eta(\theta_1, \theta_2).$$

Morally, V_η is responsible for the classical gradient-flow drift along the valley $\theta_2 = 0$, while H_η accounts for the orthogonal sharp directions that are driven at the Edge of Stability.

Therefore, the overall drift will be a combination of these two effects, while R_η only contributes to higher orders.

Lemma A.1 (Decomposition of the loss in orthogonal pieces)**.** *Under the landscape assumptions* (16)–(18) *and the regularity stated above, define*

$$V_\eta(\theta_1) = E_\eta(\theta_1, 0), \qquad H_\eta(\theta_1) = D_2^2 E_\eta(\theta_1, 0).$$

Then $V_\eta \in C^4$, $H_\eta \in C^2$, and, for $|\theta_1| \le \delta/C_0$,

$$E_\eta(\theta_1, \theta_2) = V_\eta(\theta_1) + \frac{1}{2}\theta_2^T H_\eta(\theta_1)\theta_2 + R_\eta(\theta_1, \theta_2), \tag{20}$$

where

$$R_\eta(\theta_1, \theta_2) = \frac{1}{2} \int_0^1 (1-s)^2 D_2^3 E_\eta(\theta_1, s\theta_2)[\theta_2, \theta_2, \theta_2] \, \mathrm{d}s. \tag{21}$$

On this region, the remainder is jointly C^2 and satisfies

$$\begin{gathered} R_\eta(\theta_1, 0) = \nabla_2 R_\eta(\theta_1, 0) = \nabla_2^2 R_\eta(\theta_1, 0) = 0, \\ |R_\eta(\theta_1, \theta_2)| \le C|\theta_2|^3/(6\eta). \end{gathered} \tag{22}$$

We furthermore have $|\nabla_1 V_\eta| \le C\delta$, $\|D_1^2 V_\eta\|_{\mathrm{op}} \le C$, $\|D_1^j V_\eta\|_{\mathrm{op}} \le C/\eta$ for $j = 3, 4$, $\|H_\eta - 2\mathrm{Id}_r/\eta\|_{\mathrm{op}} \le C$, and $\|D_1^j H_\eta\|_{\mathrm{op}} \le C/\eta$ for $j = 1, 2$.

Proof. Apply Taylor's formula with integral remainder to $s \mapsto E_\eta(\theta_1, s\theta_2)$ between 0 and 1. The constant and quadratic terms are V_η and $\theta_2^T H_\eta \theta_2/2$, while the linear term vanishes by assumption. This gives (21). The bound on $D^3 E_\eta$ and $\frac{1}{2}\int_0^1 (1-s)^2 \, \mathrm{d}s = 1/6$ give the bound on R_η. By their definitions, $V_\eta \in C^4$ and $H_\eta \in C^2$; their stated derivative bounds are restrictions of derivatives of E_η of order at most four. Hence $R_\eta = E_\eta - V_\eta - \theta_2^T H_\eta \theta_2/2$ is C^2; differentiating this identity with respect to θ_2 at $\theta_2 = 0$ gives the vanishing derivatives in (22). □

Two-point windows without tangential potential. We begin with $m = 2$ and $V_\eta = 0$. The aim is therefore to show the following proposition:

Proposition A.2 (Two-point window without tangential potential)**.** *Fix $C_0 \geq 1$ and $C \geq 1$, independently of η, a. Let $\eta, a \in (0,1)$ such that $\delta := a^2/\eta \in (0,1)$. The constant C_0 fixes the tangential neighborhood $|\theta_1| \leq \delta/C_0$; C is a common upper-bound constant for the loss derivatives and the initial data specified below. Assume $E_\eta \in C^4(\mathbb{R}^{d+r})$ satisfies* (16) *and* (18) *with this constant C. Let $H_\eta = D_2^2 E_\eta(\cdot, 0)$ as in Lemma A.1 and further assume that $V_\eta \equiv 0$.*

Starting from $|\tilde\theta_{1,0}| \leq \delta/C_0$, run $\tilde\theta_{t+1} = \tilde\theta_t - \eta \nabla E_\eta(\tilde\theta_t)$ and define

$$\theta_t = \frac{\tilde\theta_t + \tilde\theta_{t+1}}{2}, \qquad \xi_t = \frac{\tilde\theta_t - \tilde\theta_{t+1}}{2}, \qquad \Sigma_t = \xi_t \xi_t^T.$$

Assume $\theta_{1,0} = 0$, $|\theta_{2,0}| \leq Ca^2$, and $|\tilde\theta_{2,0}| \leq Ca$. Let $(\theta(\tau), \Sigma(\tau))$ solve the full MF equations (15) *in slow time $\tau = \eta t$, initialized at (θ_0, Σ_0), and set $\bar\theta = \theta/\delta$, $\bar\Sigma = \Sigma/a^2$ for both dynamics.*

There exist $T_* > 0$, $C_* < \infty$ and a positive threshold, all depending only on C_0, C, d, r, such that whenever $\eta + \delta$ is below this threshold, both dynamics exist on the corresponding time intervals and

$$\sup_{0 \leq \tau \leq T_*} \left(|\bar\theta_{\lfloor \tau/\eta \rfloor} - \bar\theta(\tau)| + \|\bar\Sigma_{\lfloor \tau/\eta \rfloor} - \bar\Sigma(\tau)\|_F \right) \leq C_*(\eta + \delta). \tag{23}$$

We decompose the proof of this proposition into three lemmas. We first bound the Taylor remainders between two iterates of (θ_t, Σ_t), which is Lemma A.3. Then, we control the components of (θ_t, Σ_t) that remain small on the long timescale η^{-1} (i.e. the mean along the sharp directions $\theta_{2,t}$ and the tangential blocks $\Sigma_{11,t}, \Sigma_{12,t}, \Sigma_{21,t}$) in Lemma A.4. This part crucially uses an established stabilization mechanism to propagate the smallness over long timescales of order η^{-1}. Finally, we rescale the retained quantities $(\theta_{1,t}, \Sigma_{22,t})$ and combine the remainder estimates to obtain a bound for the accumulated error on timescales η^{-1} (Lemma A.5).

We use a bootstrap argument. Increase C_0, if necessary, so that it exceeds the uniform bound on $\operatorname{tr} \bar\Sigma_{22,0}$ by a fixed positive margin. This enlargement depends only on the original C_0, C and only shrinks the neighborhood required in the assumptions. The prepared pair identity and Taylor expansion give $|\tilde\theta_{1,0}| = |\tilde\theta_{1,1}| = O(a^2) = o(\delta)$, so the initial iterates remain in this smaller neighborhood for small enough parameters. Keep C_0 fixed thereafter, independently of η, a. For $t = \lfloor \tau/\eta \rfloor$, consider the temporary bounds

$$\begin{aligned} \max\{|\bar\theta_1(\tau)|, |\bar\theta_{1,t}|\} < \frac{1}{2C_0}, \qquad & \max\{\operatorname{tr} \bar\Sigma_{22}(\tau), \operatorname{tr} \bar\Sigma_{22,t}\} < C_0, \\ \max_{j=0,1} |\tilde\theta_{1,t+j}| < \frac{\delta}{C_0}, \qquad & \max_{j=0,1} |\tilde\theta_{2,t+j}| < 2\sqrt{C_0}\, a, \\ |\theta_2(\tau)| < a, \qquad & \max\{\|\Sigma_{11}(\tau)\|_F, \|\Sigma_{12}(\tau)\|_F\} < a^2. \end{aligned} \tag{24}$$

Define their common stopping time in slow time by

$$T := \sup \left\{ s \in (0,1] : \begin{array}{l} \text{MF exists on } [0,s) \text{ and (24) holds} \\ \text{for every } 0 \leq \tau < s \end{array} \right\}. \tag{25}$$

Thus T is the first exit time, capped at 1, and may depend on η, a. The prepared initial data satisfy these bounds strictly for small enough parameters, so $T > 0$. The next three lemmas apply before T; since $T \leq 1$, their constants C_T can be chosen uniformly in η, a. After the lemmas, we prove $T > T_* > 0$ with T_* independent of the parameters, thereby closing the bootstrap.

Lemma A.3 (Taylor remainders for a GD pair)**.** *Under the assumptions of Proposition A.2, set $H_t = D^2 E_\eta(\theta_t)$ and $A_t = \mathrm{Id} - \eta H_t$. For every integer $t \geq 0$ with $\eta t < T$, where T is defined in* (25)*, the pair moments satisfy*

$$\theta_{t+1} = \theta_t - \eta \nabla E_\eta(\theta_t) - \frac{\eta}{2} D^3 E_\eta(\theta_t) : \Sigma_t + R_t^\theta, \tag{26}$$

$$\Sigma_{t+1} = A_t \Sigma_t A_t^T + R_t^\Sigma, \tag{27}$$

where

$$|R_{1,t}^\theta| + |R_{2,t}^\theta| \leq Ca^3, \tag{28}$$

and

$$\|R_{11,t}^\Sigma\|_F \leq Ca^5, \qquad \|R_{12,t}^\Sigma\|_F + \|R_{22,t}^\Sigma\|_F \leq Ca^4. \tag{29}$$

Proof. The exact pair identities give

$$R_t^\theta = -\frac{\eta}{2}\left[\nabla E_\eta(\theta_t+\xi_t) + \nabla E_\eta(\theta_t-\xi_t) - 2\nabla E_\eta(\theta_t) - D^3E_\eta(\theta_t) : \Sigma_t\right]. \tag{30}$$

Writing the odd gradient defect as $q_t = \frac{1}{2}[\nabla E_\eta(\theta_t+\xi_t) - \nabla E_\eta(\theta_t-\xi_t)] - H_t\xi_t$, we have $\xi_{t+1} = A_t\xi_t - \eta q_t$ and hence

$$R_t^\Sigma = -\eta\left[(A_t\xi_t)q_t^T + q_t(A_t\xi_t)^T\right] + \eta^2 q_t q_t^T. \tag{31}$$

The identity $\nabla_2 E_\eta(\theta_1, 0) = 0$ and Taylor's formula, applied directly to the gradient of E_η with respect to θ_2, give

$$\nabla_1 E_\eta(\theta_1,\theta_2) = \frac{1}{2}\nabla_1 H_\eta(\theta_1) : \theta_2\theta_2^T + O(|\theta_2|^3/\eta),$$
$$\nabla_2 E_\eta(\theta_1,\theta_2) = H_\eta(\theta_1)\theta_2 + O(|\theta_2|^2/\eta).$$

Since $\xi_t = \eta\nabla E_\eta(\tilde\theta_t)/2$, these estimates imply

$$|\xi_{1,t}| \le Ca^2, \qquad |\xi_{2,t}| \le Ca, \qquad |\theta_{2,t}| \le Ca(\eta+a). \tag{32}$$

For the last estimate, the leading sharp term cancels in $\theta_{2,t} = \tilde\theta_{2,t} - \eta\nabla_2 E_\eta(\tilde\theta_t)/2$. Taylor expansion of the even and odd gradient combinations, using (18), yields $|R_t^\theta| \le Ca^3$ and $|q_t| \le Ca^3/\eta$. At the pair mean, $H_{12,t} = O(a+\delta)$; together with (32), this gives $|(A_t\xi_t)_1| \le Ca^2$ and $|(A_t\xi_t)_2| \le Ca$. Substitution into (31) proves the blockwise covariance bounds. □

Lemma A.4 (Small sharp mean and tangential covariance blocks)**.** *For MF at $0 \le \tau < T$, and for GD pairs at $0 \le \eta t < T$, with T defined in* (25)*, we have*

$$|\theta_2| \le C_T a^2, \qquad \|\Sigma_{11}\|_F \le C_T a^4, \qquad \|\Sigma_{12}\|_F \le C_T a^3. \tag{33}$$

In particular, after normalization, $|\bar\theta_2| \le C_T\eta$, $\|\bar\Sigma_{11}\|_F \le C_T a^2$ and $\|\bar\Sigma_{12}\|_F \le C_T a$.

Proof. Gradient descent long-time estimates. The covariance bounds follow directly from $\Sigma_t = \xi_t\xi_t^T$ and (32). The same estimate gives $|\theta_{2,t}| \le Ca(\eta+a)$, which proves the sharp-mean bound when $a \ge \eta$. Suppose now that $a \le \eta$. We use the cancellation between two successive steps, rather than taking absolute values in each step. We first justify the expansion along the sharp directions used to identify the forcing. For fixed θ_1, $\nabla_2 E_\eta(\theta_1, 0) = 0$ and Taylor's formula give the exact identity

$$\nabla_2 E_\eta(\theta_1,\theta_2) = H_\eta(\theta_1)\theta_2 + \int_0^1 (1-s)D_2^3 E_\eta(\theta_1, s\theta_2) : (\theta_2\theta_2^T)\,\mathrm{d}s.$$

The segment $(\theta_1, s\theta_2)$, $s \in [0,1]$, keeps θ_1 fixed in the allowed neighborhood, and the derivative bounds hold along the entire segment. Hence

$$|\nabla_2 E_\eta(\theta_1,\theta_2) - H_\eta(\theta_1)\theta_2| \le \frac{C}{2\eta}|\theta_2|^2,$$
$$\|D_2^3 E_\eta(\theta_1,\theta_2) - D_2^3 E_\eta(\theta_1,0)\|_{\mathrm{op}} \le \frac{C}{\eta}|\theta_2|.$$

The first bound uses D^3E_η and the second uses D^4E_η. Thus the expansion requires a small displacement $|\theta_2|$ along the sharp directions. Set

$$M_t := \eta H_\eta(\theta_{1,t}) - \mathrm{Id}_r, \qquad \|M_t - \mathrm{Id}_r\|_{\mathrm{op}} \le C\eta,$$
$$F_t := -\frac{\eta}{2}D_2^3 E_\eta(\theta_{1,t}, 0) : \Sigma_{22,t}, \qquad |F_t| \le Ca^2.$$

Substituting these estimates into the sharp mean update (26) gives

$$\theta_{2,t+1} = -M_t\theta_{2,t} + F_t + O(a^3 + \eta^2 a^2).$$

Indeed, after multiplication by η, the nonlinear gradient contributes $O(|\theta_{2,t}|^2)$, and replacing the sharp cubic coefficient by its value at $(\theta_{1,t}, 0)$ costs $O(|\theta_{2,t}|a^2)$ because $\|\Sigma_{22,t}\|_F \le Ca^2$. The other covariance contractions are $O(a^3)$, using $\|\Sigma_{12,t}\|_F \le Ca^3$ and $\|\Sigma_{11,t}\|_F \le Ca^4$; the pair Taylor

defect is also $O(a^3)$. We use only the preliminary bound $|\theta_{2,t}| \le C\eta a$ from (32), not the desired $O(a^2)$ bound. The total error is therefore $O(\eta^2 a^2 + a^3) = O(\eta a^2)$.

The forcing varies slowly. Using only the preliminary bounds, Lemma A.3 gives

$$|\theta_{1,t+1} - \theta_{1,t}| \le Ca^2, \qquad \|\Sigma_{22,t+1} - \Sigma_{22,t}\|_F \le C\eta a^2.$$

For the second estimate, the leading sharp multiplier in A_t is $-\mathrm{Id}_r + O(\eta)$, so its two minus signs cancel in $A_t \Sigma_t A_t^T$; the covariance defect $O(a^4)$ is also $O(\eta a^2)$. Together with $\eta \|D_1 D_2^3 E_\eta\|_{\mathrm{op}} \le C$, these bounds imply $|F_{t+1} - F_t| \le C(a^4 + \eta a^2) \le C\eta a^2$. Composing two mean updates now makes the cancellation explicit:

$$\theta_{2,t+2} = M_{t+1} M_t \theta_{2,t} + (F_{t+1} - F_t) - (M_{t+1} - \mathrm{Id}_r) F_t + O(\eta a^2).$$

Consequently,

$$|\theta_{2,t+2}| \le (1 + C\eta)|\theta_{2,t}| + C\eta a^2.$$

The initial preparation controls $\theta_{2,0}$, and the one-step update then gives $|\theta_{2,1}| \le Ca^2$. For either parity $j \in \{0, 1\}$ and $\eta(2k + j) < T$, iteration yields

$$\begin{aligned} |\theta_{2,2k+j}| &\le (1 + C\eta)^k |\theta_{2,j}| + C\eta a^2 \sum_{\ell=0}^{k-1} (1 + C\eta)^\ell \\ &\le Ca^2 (1 + T) e^{CT}. \end{aligned}$$

The Grönwall factor is uniform in η, since $(1 + C\eta)^k \le e^{Ck\eta} \le e^{CT}$.

Mean–fluctuation mean in sharp directions. All derivatives are now in slow time τ. Covariance positivity and the stopping bounds give $\operatorname{tr} \Sigma \le Ca^2$, hence $|D_2 D^2 E_\eta(\theta) : \Sigma| \le Ca^2/\eta$. Taylor expansion of the gradient gives

$$\frac{\mathrm{d}\theta_2}{\mathrm{d}\tau} = -H_\eta(\theta_1)\theta_2 + O(|\theta_2|^2/\eta) - \frac{1}{2} D_2 D^2 E_\eta(\theta) : \Sigma.$$

Since $H_\eta \ge (2/\eta - C)\mathrm{Id}_r$, taking the inner product with $\theta_2/|\theta_2|$ gives

$$\begin{aligned} D_\tau^+ |\theta_2| &\le -\left(\frac{2}{\eta} - C\right) |\theta_2| + \frac{C}{\eta} |\theta_2|^2 + \frac{C}{\eta} a^2 \\ &\le -\frac{1}{\eta} |\theta_2| + \frac{C}{\eta} a^2. \end{aligned}$$

Here D_τ^+ is the upper right Dini derivative, so the inequality also holds at $\theta_2 = 0$. The last line uses $|\theta_2| < a$ and small a, η to absorb the quadratic term into the damping. The integrating-factor estimate is therefore

$$|\theta_2(\tau)| \le e^{-\tau/\eta} |\theta_2(0)| + Ca^2 (1 - e^{-\tau/\eta}) \le Ca^2.$$

Mean–fluctuation covariance blocks. Write $H = D^2 E_\eta(\theta)$ and H_{ij} for its blocks. Since $D_1^2 E_\eta(\theta_1, 0) = 0$ and differentiating $\nabla_2 E_\eta(\theta_1, 0) = 0$ twice in θ_1 gives $D_1^2 D_2 E_\eta(\theta_1, 0) = 0$, the fourth-derivative bound yields $D_1^2 E_\eta(\theta_1, \theta_2) = O(|\theta_2|^2/\eta)$. Expanding the other blocks and using the sharp-mean bound just proved, we obtain

$$H_{11} = O(a^2 \delta), \qquad H_{12} = O(\delta), \qquad H_{22} = 2\mathrm{Id}_r/\eta + O(1).$$

The stabilizing term in the mixed covariance equation is visible by writing that block explicitly:

$$\begin{aligned} \frac{\mathrm{d}\Sigma_{12}}{\mathrm{d}\tau} &= -\Sigma_{12} H_{22} + H_{11} \Sigma_{12} (\eta H_{22} - \mathrm{Id}_r) \\ &\quad + H_{12} \Sigma_{22} (\eta H_{22} - \mathrm{Id}_r) + (\eta H_{11} - \mathrm{Id}_d) \Sigma_{11} H_{12} + \eta H_{12} \Sigma_{21} H_{12}. \end{aligned}$$

Only the first term has a coefficient of order $1/\eta$; it contributes at most $-(2/\eta - C)\|\Sigma_{12}\|_F^2$ to the Frobenius inner product with Σ_{12}. The remaining terms, and the 11 block of the same covariance equation, give

$$\begin{aligned} D_\tau^+ \|\Sigma_{12}\|_F &\le -(2/\eta - C)\|\Sigma_{12}\|_F + C\delta(a^2 + \|\Sigma_{11}\|_F), \\ D_\tau^+ \|\Sigma_{11}\|_F &\le C\|\Sigma_{11}\|_F + C\delta \|\Sigma_{12}\|_F + C\eta\delta^2 a^2. \end{aligned}$$

In the second inequality, the last forcing term comes from $\eta H_{12}\Sigma_{22}H_{21}$; all other couplings to the sharp block contain Σ_{12} or Σ_{21}. For small η, $2/\eta - C \geq 1/\eta$. The stopping bound $\|\Sigma_{11}\|_F < a^2$ and the initial estimate $\|\Sigma_{12}(0)\|_F \leq Ca^3$ thus give

$$\begin{aligned}\|\Sigma_{12}(\tau)\|_F &\leq Ca^3 e^{-\tau/\eta} + C\delta a^2 \int_0^\tau e^{-(\tau-s)/\eta}\,\mathrm{d}s \\ &\leq Ca^3 e^{-\tau/\eta} + Ca^4,\end{aligned}$$

where $\eta\delta = a^2$. This proves the mixed-block bound. Substitution into the 11 inequality, followed by Grönwall, yields

$$\begin{aligned}\|\Sigma_{11}(\tau)\|_F &\leq e^{C\tau}\left[Ca^4 + C\delta a^3 \int_0^\tau e^{-s/\eta}\,\mathrm{d}s + C\tau\delta a^4 + C\tau\eta\delta^2 a^2\right] \\ &\leq Ce^{CT}\left[a^4 + a^5 + T\delta a^4\right] \leq C_T a^4.\end{aligned}$$

Together with the sharp-mean bound, this proves (33). □

Now that we have proved the propagation, over times $t = O(\eta^{-1})$, of the smallness of θ_2 and the covariance blocks $\Sigma_{12}, \Sigma_{21}, \Sigma_{11}$, we combine the remainders to keep only the relevant, non-vanishing contributions:

Lemma A.5 (Rescaled equations and error bounds)**.** *Let T be the stopping time* (25)*. For $0 \leq \eta t < T$, the GD pair moments satisfy*

$$\begin{aligned}\bar\theta_{1,t+1} &= \bar\theta_{1,t} - \frac{\eta^2}{2}\nabla_1 H_\eta : \bar\Sigma_{22,t} + \bar R_t^\theta, \\ \bar\Sigma_{22,t+1} &= \bar\Sigma_{22,t} - \eta\left[H_\eta\bar\Sigma_{22,t} + \bar\Sigma_{22,t}H_\eta\right] + \eta^2 H_\eta\bar\Sigma_{22,t}H_\eta + \bar R_t^\Sigma,\end{aligned} \tag{34}$$

where H_η and its derivatives are evaluated at $\delta\,\bar\theta_{1,t}$, and

$$\begin{aligned}&|\bar R_t^\theta| \leq C_T\eta a, \qquad \|\bar R_t^\Sigma\|_F \leq C_T\eta\delta, \\ &\sum_{0\leq t<T/\eta}\left(|\bar R_t^\theta| + \|\bar R_t^\Sigma\|_F\right) \leq C_T(a+\delta).\end{aligned} \tag{35}$$

For MF, in slow time $0 \leq \tau < T$,

$$\begin{aligned}\frac{\mathrm{d}\bar\theta_1}{\mathrm{d}\tau} &= -\frac{\eta}{2}\nabla_1 H_\eta : \bar\Sigma_{22} + O_T(a^2), \\ \frac{\mathrm{d}\bar\Sigma_{22}}{\mathrm{d}\tau} &= -H_\eta\bar\Sigma_{22} - \bar\Sigma_{22}H_\eta + \eta H_\eta\bar\Sigma_{22}H_\eta + O_T(\delta),\end{aligned} \tag{36}$$

where H_η and its derivatives are evaluated at $\delta\,\bar\theta_1(\tau)$. The leading drifts are uniformly bounded on $|\bar\theta_1| \leq 1/(2C_0)$, $\|\bar\Sigma_{22}\|_F \leq C_0$. On this region, we also have

$$\|D_1 H_\eta(\delta\bar\theta_1)\|_{\mathrm{op}} \leq \frac{C}{\eta + a}. \tag{37}$$

Proof. By Lemma A.4 and Taylor expansion along the sharp directions θ_2,

$$\begin{aligned}\eta\nabla_1 E_\eta(\theta_t) &= O_T(a^4), \\ \eta D_1 D^2 E_\eta(\theta_t) : \Sigma_t &= \eta\nabla_1 H_\eta(\theta_{1,t}) : \Sigma_{22,t} + O_T(a^4).\end{aligned}$$

Here $\nabla_2 E_\eta(\theta_1, 0) = 0$ and the fourth-derivative bound give $D_1^2 D_2 E_\eta = O(|\theta_2|/\eta)$ and $D_1 D_2^2 E_\eta - \nabla_1 H_\eta = O(|\theta_2|/\eta)$. For the covariance, the Hessian bounds $H_{22,t} - H_\eta(\theta_{1,t}) = O_T(\delta)$ and $H_{12,t} = O_T(\delta)$ show that replacing the sharp block of $A_t\Sigma_t A_t^T$ by $(\mathrm{Id}_r - \eta H_\eta)\Sigma_{22,t}(\mathrm{Id}_r - \eta H_\eta)$ also costs $O_T(\eta\delta a^2) = O_T(a^4)$. Adding the Taylor defects from Lemma A.3 thus gives physical errors $O_T(a^3)$ and $O_T(a^4)$. Division by δ and a^2 yields the pointwise bounds in (35); summation over fewer than $T/\eta + 1$ steps gives the accumulated bound $C_T(a+\delta)$.

For MF, the same discarded terms have size $O_T(a^4/\eta)$ per unit slow time, with no discrete Taylor defects. Normalization therefore gives the stated errors $O_T(a^2)$ and $O_T(\delta)$.

The mean drift is bounded by $C\eta\|D_1 H_\eta\|_{\mathrm{op}} \le C$. For the covariance drift, substituting $H_\eta = 2\mathrm{Id}_r/\eta + O(1)$ cancels the singular terms and gives the same bound. To prove (37), the assumed bound $\|D_1 H_\eta\|_{\mathrm{op}} \le C/\eta$ suffices when $a \le \eta$. When $a \ge \eta$, Taylor's theorem on tangential segments of length proportional to $a \le \delta$, inside $|\theta_1| \le \delta/C_0$, gives

$$a\|D_1 H_\eta\|_{\mathrm{op}} \le C(1 + a^2/\eta) = C(1+\delta),$$

using $\|H_\eta - 2\mathrm{Id}_r/\eta\|_{\mathrm{op}} \le C$ and $\|D_1^2 H_\eta\|_{\mathrm{op}} \le C/\eta$. Since $\eta + a \le 2a$ in this case, this proves (37). □

Proof of Proposition A.2. Local MF existence and covariance positivity follow as in Proposition A.9. We compare the two retained components at their respective error scales $\eta + a$ and $\eta + \delta$, both tending to zero:

$$e_t = \frac{|\bar\theta_{1,t} - \bar\theta_1(\eta t)|}{\eta + a} + \frac{\|\bar\Sigma_{22,t} - \bar\Sigma_{22}(\eta t)\|_F}{\eta + \delta}.$$

The leading vector field in (36) is uniformly Lipschitz after this normalization. Indeed, its two cross-derivative bounds follow from (37) and

$$\frac{\eta(\eta+\delta)}{(\eta+a)^2} = \frac{\eta^2 + a^2}{(\eta+a)^2} \le 1, \qquad \frac{\delta}{\eta+\delta} \le 1;$$

the derivatives in each component's own variable are bounded by $Ca^2\|D_1^2 H_\eta\|_{\mathrm{op}} \le C\delta$ and C, respectively. The GD remainders and integrated MF remainders contribute at most $C_T\eta$ per step in this error norm. Since the retained MF velocities are bounded, the quadrature error is at most $C_T\eta^2[(\eta+a)^{-1} + (\eta+\delta)^{-1}] \le C_T\eta$. Thus, for $\eta(t+1) < T$,

$$e_{t+1} \le (1 + C\eta)e_t + C_T\eta, \qquad e_0 = 0, \qquad \sup_{\eta t < T} e_t \le C_T,$$

where the last bound is discrete Grönwall, with constants independent of a/η. The two component errors are therefore $O_T(\eta + a)$ and $O_T(\eta + \delta)$. Adding the small components from Lemma A.4, the $O(\eta)$ variation between samples, and using $a = \sqrt{\eta\delta} \le (\eta+\delta)/2$ proves (23) on $[0, T)$.

Final bootstrap argument. To close the bootstrap, Lemma A.4 gives

$$|\theta_2| \le C_T a^2 < a, \qquad \|\Sigma_{11}\|_F \le C_T a^4 < a^2, \qquad \|\Sigma_{12}\|_F \le C_T a^3 < a^2,$$

strictly improving the MF bounds for small parameters. By Lemma A.5, the retained normalized variables move by at most $C\tau$. Choose $T_* \in (0,1)$ so that this motion preserves the fixed initial margins in the bounds on $\bar\theta_1$ and $\mathrm{tr}\,\bar\Sigma_{22}$. The constants are uniform since $T \le 1$, so T_* is independent of η, a. For the individual GD iterates, $\tilde\theta_t = \theta_t + \xi_t$ and $\tilde\theta_{t+1} = \theta_t - \xi_t$ give, for $j = 0, 1$,

$$|\tilde\theta_{1,t+j}| \le \frac{\delta}{2C_0} + Ca^2 < \frac{\delta}{C_0}, \qquad |\tilde\theta_{2,t+j}| \le \sqrt{C_0}\,a + C_T a^2 < 2\sqrt{C_0}\,a,$$

using $|\xi_{1,t}| \le Ca^2$ and $|\xi_{2,t}|^2 = \mathrm{tr}\,\Sigma_{22,t}$. MF continuity and the GD update from the last admissible pair therefore exclude a first exit at or before T_*, including the extra iterate needed for the final pair. Boundedness permits MF continuation, so $T > T_*$ and the comparison holds on $[0, T_*]$. □

Adding a tangential potential. A tangential potential adds the drift $-\eta\nabla_1 V_\eta$. We require it to be $O(a^2)$ per GD step, like the curvature-induced drift, and assume a bounded tangential Hessian to preserve stability.

Theorem A.6 (GD–MF comparison with a tangential potential)**.** *Keep the assumptions of Proposition A.2, except that $V_\eta := E_\eta(\cdot, 0)$ need not vanish and satisfies*

$$|\nabla_1 V_\eta| \le C\delta, \qquad \|D_1^2 V_\eta\|_{\mathrm{op}} \le C \qquad \textit{for } |\theta_1| \le \delta/C_0. \tag{38}$$

With the pair moments and common MF initial data computed from this energy, the conclusions of Proposition A.2 remain valid.

Proof. Only the tangential drift changes at leading order. Indeed, $\eta|\nabla_1 V_\eta| \le Ca^2$ preserves the fluctuation and Taylor bounds of Lemma A.3. The sharp-mean estimate is unchanged because V_η is independent of θ_2. For MF, the additional block $D_1^2 V_\eta = O(1)$ changes only the constants in the covariance inequalities of Lemma A.4. The rescaled GD mean therefore satisfies

$$\bar\theta_{1,t+1} = \bar\theta_{1,t} - \frac{\eta}{\delta}\nabla_1 V_\eta(\delta\bar\theta_{1,t}) - \frac{\eta^2}{2}\nabla_1 H_\eta(\delta\bar\theta_{1,t}) : \bar\Sigma_{22,t} + \bar R_t^\theta. \tag{39}$$

The covariance update and remainder bounds are unchanged. In the normalized MF mean equation, the additional covariance correction is $-D_1^3 V_\eta : \Sigma_{11}/(2\delta) = O_T(a^2)$, so it fits the remainder in Lemma A.5. The leading drift $-\nabla_1 V_\eta(\delta\bar\theta_1)/\delta$ is bounded and uniformly Lipschitz, since its derivative is $-D_1^2 V_\eta$. The Grönwall estimate and bootstrap of Proposition A.2 therefore apply with the same error bound and a uniform $T_* > 0$. □

Larger even windows. An even window is a union of adjacent GD pairs. Its covariance is the average of their covariances plus the covariance of their means. This identity extends the pair estimate without a new Taylor expansion.

Corollary A.7 (GD–MF comparison for even windows)**.** *Under the hypotheses of Theorem A.6, let $m \ge 2$ be even, possibly depending on η, a, and define*

$$\theta_t^{(m)} = \frac{1}{m}\sum_{j=0}^{m-1}\tilde\theta_{t+j}, \qquad \Sigma_t^{(m)} = \frac{1}{m}\sum_{j=0}^{m-1}(\tilde\theta_{t+j} - \theta_t^{(m)})(\tilde\theta_{t+j} - \theta_t^{(m)})^T.$$

Initialize full MF at these window moments, and normalize both processes by $\bar\theta = \theta/\delta$, $\bar\Sigma = \Sigma/a^2$. There exist $T_*, \varepsilon_*, C_* > 0$, independent of m, η, a, such that, whenever $m(\eta+\delta) \le \varepsilon_*$, both dynamics exist on the required intervals and

$$\sup_{0\le\tau\le T_*}\left(|\bar\theta^{(m)}_{\lfloor\tau/\eta\rfloor} - \bar\theta(\tau)| + \|\bar\Sigma^{(m)}_{\lfloor\tau/\eta\rfloor} - \bar\Sigma(\tau)\|_F\right) \le C_* m(\eta+\delta). \tag{40}$$

The preparation condition remains imposed on the first GD pair; the window length may grow provided $m(\eta+\delta) \to 0$.

Proof. Use superscript (2) for the pair moments. Partitioning the window into $m/2$ pairs gives the exact identities

$$\theta_t^{(m)} = \frac{2}{m}\sum_{j=0}^{m/2-1}\theta_{t+2j}^{(2)},$$

$$\Sigma_t^{(m)} = \frac{2}{m}\sum_{j=0}^{m/2-1}\left[\Sigma_{t+2j}^{(2)} + (\theta_{t+2j}^{(2)} - \theta_t^{(m)})(\theta_{t+2j}^{(2)} - \theta_t^{(m)})^T\right].$$

The tangential pair mean changes by $O(a^2)$ per step, the sharp mean stays $O_T(a^2)$, and the sharp covariance changes by $O_T(\eta a^2)$ per step. Substitution into these identities gives

$$|\theta_{2,t}^{(m)}| \le C_T a^2, \qquad \|\Sigma_{11,t}^{(m)}\|_F \le C_T m^2 a^4, \qquad \|\Sigma_{12,t}^{(m)}\|_F \le C_T(a^3 + ma^4),$$
$$|\bar\theta_{1,t}^{(m)} - \bar\theta_{1,t}^{(2)}| \le C_T m\eta, \qquad \|\bar\Sigma_{22,t}^{(m)} - \bar\Sigma_{22,t}^{(2)}\|_F \le C_T(m\eta + a^2),$$

with constants independent of m. The last a^2 term comes from the covariance of the sharp pair means, divided by a^2.

Since $2ma \le m(\eta+\delta)$, the larger initial tangential covariance is still small compared with a^2. For MF initialized at the window moments, the differential inequalities of Lemma A.4 give

$$|\theta_2(\tau)| \le C_T a^2, \qquad \|\Sigma_{12}(\tau)\|_F \le C_T(a^3 e^{-\tau/\eta} + a^4), \qquad \|\Sigma_{11}(\tau)\|_F \le C_T m^2 a^4.$$

The mixed block retains its damping on slow times of order η; the factor m^2 affects only the tangential covariance bound. Consequently, the normalized mean remainder becomes $O_T(m^2 a^2)$ (including $D_1^3 V_\eta : \Sigma_{11}/(2\delta)$), while the sharp-covariance remainder remains $O_T(\delta)$.

Compare this MF solution with the GD pair moments, using the normalized error e_t from the proof of Proposition A.2. The initial window–pair bounds give $e_0 \le Cm$. The leading drift has the same Lipschitz bound, and $m^2a^2 \ll 1$ preserves bounded velocities and the quadrature estimate. Grönwall therefore yields

$$\sup_t e_t \le C_T \left[m + \frac{m^2a^2}{\eta + a} \right] \le C_T m,$$

since $ma^2/(\eta + a) \le m\delta \le \varepsilon_*$. Returning to the window moments and including the small blocks gives (40), using $2a \le \eta + \delta$ and $m^2a^2 = (m\eta)(m\delta) \le \varepsilon_* m\eta$.

To close the bootstrap, note that $\|\Sigma_{11}\|_F/a^2 \le C_T(ma)^2 \ll 1$; the other small-block bounds also strictly improve the stopping bounds. Choose T_* below half the pair-comparison horizon, then ε_* small enough to accommodate the initial window's tangential shift and the extra slow-time interval $m\eta$ needed for the final window. This gives a common interval $[0, T_*]$. Returning to iteration time proves Theorem 2.1. □

Why the window must be even. An odd window leaves an uncancelled alternating component. For $E_\eta(\theta_2) = \theta_2^2/\eta$ and $\tilde{\theta}_{2,0} = a$, an odd window has mean $(-1)^t a/m$, whereas its window-initialized MF mean is $(a/m)e^{-2\tau/\eta}$. Thus at $\tau = \eta$ the normalized mean error equals $(1 + e^{-2})\eta/(ma)$. With $a = \eta$, this error is $(1 + e^{-2})/m$ and does not vanish for fixed odd m, although $\delta \to 0$.

Rank of established oscillations. The pair covariance $\Sigma_t^{(2)} = \xi_t \xi_t^\top$ has rank at most one. The window estimates above give

$$\left\| \bar{\Sigma}_t^{(m)} - \operatorname{diag}\left(0, \bar{\Sigma}_{22,t}^{(2)}\right) \right\|_F \le C_T\left(m\eta + a + m^2a^2 + ma^2\right) \le C\, m(\eta + \delta), \tag{41}$$

and Theorem 2.1 gives the same bound for MF. Thus, for $0 \le \tau \le T_*$, both $\bar{\Sigma}^{(m)}_{\lfloor \tau/\eta \rfloor}$ and $\bar{\Sigma}(\tau)$ stay within $C_* m(\eta + \delta)$ of a matrix of rank at most one. For established oscillations, $|\xi_0| \asymp a$ and $\|\bar{\Sigma}_0\|_F \asymp 1$, so the additional MF covariance directions are of lower order (Remark 2.2).

In the scaling limit of Corollary A.8, this factorization is exact. Write $\bar{\Sigma}^0_{22} = \zeta_0 \zeta_0^\top$ and solve $\dot{\zeta} = \bar{H}(\bar{\theta}_1)\zeta$ with $\zeta(0) = \zeta_0$. Differentiating $\zeta\zeta^\top$ gives the covariance equation in (43); hence $\bar{\Sigma}_{22}(\tau) = \zeta(\tau)\zeta(\tau)^\top$. Its rank remains one if $\zeta_0 \neq 0$, and zero otherwise.

Numerical check. In our MLP runs (seed 0, $\eta \in \{0.0025, 0.004\}$, one- and two-mode initializations), let $\widehat{X}_t = U_t^\top \widehat{\Sigma}_t U_t$ be the 16-iterate GD covariance projected on the three tracked leading Hessian directions. The fraction outside its principal direction, $(\lambda_2(\widehat{X}_t) + \lambda_3(\widehat{X}_t))/\operatorname{tr} \widehat{X}_t$, stays below 0.4% at every sample (Figure 9), even for two-mode initializations. This check does not measure fluctuations outside the tracked subspace.

Why MF still retains several directions. The rank-one limit concerns established oscillations, not their onset. In the onset experiment, drift dominates the initial window covariance and the critical Hessian direction carries little variance (Figure 2(b)). Truncating to the principal covariance direction can discard this small component. MF retains it, allowing it to grow and become dominant as oscillations develop (Appendix C.2); at fixed Hessian, its variance grows at rate $\eta\lambda_1(\eta\lambda_1 - 2)$.

Last but not least, this rank-one reduction only applies to the sharp valley picture we consider here, so retaining several directions might still be an advantage in general situations.

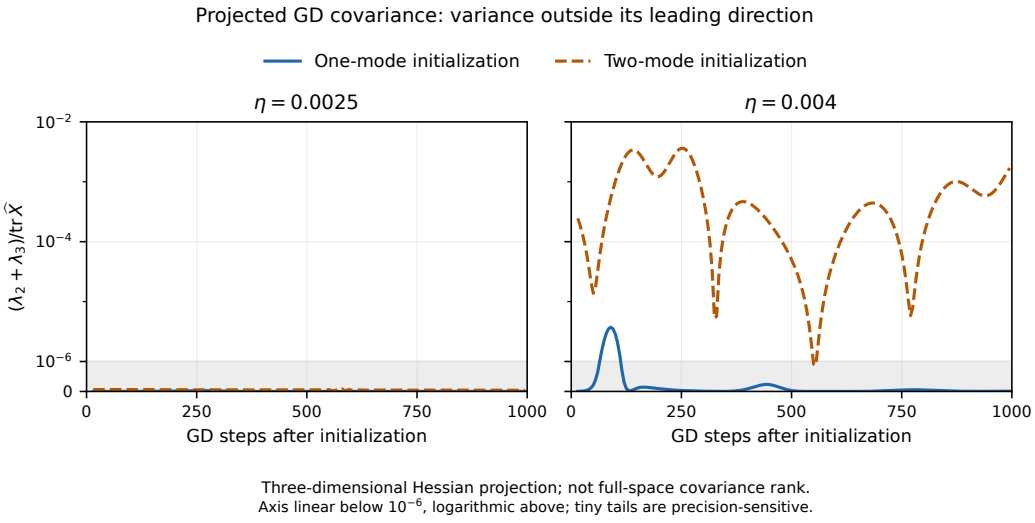


Figure 9: Fraction of the projected GD window covariance outside its principal direction (MLP, windows of 16 iterates, sampled every five steps). The vertical axis is linear below 10^{-6} (shaded) and logarithmic above; values in this range are at float32 precision.

A parameter-free scaling limit. The preceding comparison allows any relative scaling of a and η. The additional choice $a = \eta$ gives a limiting system in which the mean and covariance can remain coupled, provided the rescaled landscape has a limit.

Corollary A.8 (Coupled sharp-valley limit)**.** *Under the hypotheses of Corollary A.7, assume $a = \eta \to 0$, so that $\delta = \eta$, and let the even window length satisfy $m\eta \to 0$. For $|y| < 1/C_0$, define*

$$\bar{V}_\eta(y) = \frac{V_\eta(\eta y) - V_\eta(0)}{\eta^2}, \qquad \bar{H}_\eta(y) = H_\eta(\eta y) - \frac{2}{\eta}\mathrm{Id}_r. \tag{42}$$

Suppose $\bar{V}_\eta \to \bar{V}$ and $\bar{H}_\eta \to \bar{H}$ locally in C^1, and that the initial GD pair covariance satisfies $\Sigma^{(2)}_{22,0}/\eta^2 \to \bar{\Sigma}^0_{22}$. Then, on a common interval $[0, T_*]$ independent of η, m, the normalized GD window moments and their associated MF solution converge uniformly to

$$(\bar{\theta}_1, 0), \qquad \begin{pmatrix} 0 & 0 \\ 0 & \bar{\Sigma}_{22} \end{pmatrix},$$

where the GD moments are evaluated at $\lfloor \tau/\eta \rfloor$ and the normalization is $\bar{\theta} = \theta/\eta$, $\bar{\Sigma} = \Sigma/\eta^2$. The limiting components solve

$$\begin{aligned} \frac{\mathrm{d}\bar{\theta}_1}{\mathrm{d}\tau} &= -\nabla\bar{V}(\bar{\theta}_1) - \frac{1}{2}\nabla\bar{H}(\bar{\theta}_1) : \bar{\Sigma}_{22}, \\ \frac{\mathrm{d}\bar{\Sigma}_{22}}{\mathrm{d}\tau} &= \bar{H}(\bar{\theta}_1)\bar{\Sigma}_{22} + \bar{\Sigma}_{22}\bar{H}(\bar{\theta}_1), \\ \bar{\theta}_1(0) &= 0, \qquad \bar{\Sigma}_{22}(0) = \bar{\Sigma}^0_{22}. \end{aligned} \tag{43}$$

All derivatives of $\bar{V}, \bar{H}$ are with respect to the rescaled tangential coordinate y; in particular, $(\nabla\bar{H} : S)_j = \mathrm{tr}((\partial_j\bar{H})S)$.

Proof. First consider MF initialized at the GD pair moments. By Lemma A.5 and the potential correction in the proof of Theorem A.6, its retained equations are

$$\begin{aligned} \frac{\mathrm{d}\bar{\theta}_1}{\mathrm{d}\tau} &= -\nabla\bar{V}_\eta(\bar{\theta}_1) - \frac{1}{2}\nabla\bar{H}_\eta(\bar{\theta}_1) : \bar{\Sigma}_{22} + O_T(\eta), \\ \frac{\mathrm{d}\bar{\Sigma}_{22}}{\mathrm{d}\tau} &= \bar{H}_\eta\bar{\Sigma}_{22} + \bar{\Sigma}_{22}\bar{H}_\eta + \eta\bar{H}_\eta\bar{\Sigma}_{22}\bar{H}_\eta + O_T(\eta), \end{aligned}$$

where $\bar{H}_\eta$ is evaluated at $\bar{\theta}_1$. The first equation uses $\eta\nabla_1 H_\eta(\eta y) = \nabla\bar{H}_\eta(y)$. In the second, substituting $H_\eta = 2\mathrm{Id}_r/\eta + \bar{H}_\eta$ into $-H_\eta\bar{\Sigma}_{22} - \bar{\Sigma}_{22}H_\eta + \eta H_\eta\bar{\Sigma}_{22}H_\eta$ cancels the singular terms. Thus the Hessian-quadratic contribution is essential to the limit.

The retained vector fields have a uniform Lipschitz bound on the bounded region constructed by the bootstrap. Their coefficients converge there by assumption, and the remaining drift terms tend uniformly to zero. Writing the equations in integral form and applying Grönwall's lemma therefore gives convergence to (43), on a possibly smaller uniform interval. The small-component bounds in Lemma A.4 show that all other normalized blocks vanish. The pair GD moments have the same limit by Theorem A.6.

Finally, the window estimates in the proof of Corollary A.7 bound the difference between the normalized pair and even-window GD moments by $C_T m\eta$. That corollary gives the same order of error between the window moments and their associated MF solution. Since $m\eta \to 0$, both have the asserted limit and the same limiting initial data. □

Under the present derivative assumptions, the limiting curvature is necessarily affine. Indeed,

$$\|D_y^2\bar{H}_\eta\|_{\mathrm{op}} = \eta^2\|D_1^2 H_\eta(\eta y)\|_{\mathrm{op}} \le C\eta, \qquad \|D_y^4\bar{V}_\eta\|_{\mathrm{op}} \le C\eta.$$

Taylor's formula therefore shows that $\bar{H}$ is affine and $\bar{V}$ is a polynomial of degree at most three locally. An affine $\bar{H}$ still allows two-way coupling: the covariance drives the mean, which changes the curvature governing covariance growth when $\bar{H}$ is nonconstant. A nonlinear limiting curvature would require weaker bounds on the physical derivatives and a new remainder analysis.

A.2 Formal derivation of the mean–fluctuation dynamics

Let L_3, L_4 be the sup norm of third and fourth order derivatives of E. Taylor expansion in Equation (2) gives

$$\begin{aligned}\theta_{t+1} - \theta_t &= -\eta\left\langle \nabla E(\theta_t) + H(\theta_t)\xi_t + \frac{1}{2}\nabla H(\theta_t) : \xi_t\xi_t^T \right\rangle_m + R_t^\theta \\ &= -\eta\nabla E(\theta_t) - \frac{\eta}{2}\nabla H(\theta_t) : \Sigma_t + R_t^\theta, \end{aligned} \tag{44}$$

where we used $\langle\xi_t\rangle_m = 0$ and $\langle\xi_t\xi_t^T\rangle_m = \Sigma_t$, and

$$|R_t^\theta| \le \frac{\eta L_4}{6}\left\langle|\xi_t|^3\right\rangle_m.$$

We now turn to the fluctuation equation (3). Set $A_t = \mathrm{Id} - \eta H_t$ and

$$r_{t,j} = \nabla E(\theta_t + \xi_{t,j}) - \nabla E(\theta_t) - H(\theta_t)\xi_{t,j}, \qquad \bar{r}_t = \langle r_t\rangle_m.$$

Then $|r_{t,j}| \le L_3|\xi_{t,j}|^2/2$ and

$$\xi_{t+1,j} = A_t\xi_{t,j} - \eta(r_{t,j} - \bar{r}_t).$$

Taking the outer product and averaging over j, with $\langle\xi_t\rangle_m = 0$, yields exactly

$$\Sigma_{t+1} = A_t\Sigma_t A_t^T - \eta\left\langle A_t\xi_t r_t^T + r_t\xi_t^T A_t^T\right\rangle_m + \eta^2\left(\langle r_t r_t^T\rangle_m - \bar{r}_t\bar{r}_t^T\right). \tag{45}$$

Expanding $A_t\Sigma_t A_t^T$, we obtain

$$\Sigma_{t+1} - \Sigma_t = -\eta(H_t\Sigma_t + \Sigma_t H_t) + \eta^2 H_t\Sigma_t H_t + R_t^\Sigma, \tag{46}$$

with the Frobenius-norm bound

$$\|R_t^\Sigma\| \le \eta L_3\|A_t\|_{\mathrm{op}}\left\langle|\xi_t|^3\right\rangle_m + \frac{\eta^2 L_3^2}{4}\left\langle|\xi_t|^4\right\rangle_m. \tag{47}$$

Neglecting higher order terms, we have arrived at the coupled updates

$$\theta_{t+1} = \theta_t - \eta\nabla E(\theta_t) - \frac{\eta}{2}\nabla H(\theta_t) : \Sigma_t, \tag{48}$$

$$\Sigma_{t+1} = (\mathrm{Id} - \eta H(\theta_t))\Sigma_t(\mathrm{Id} - \eta H(\theta_t)). \tag{49}$$

Role of the window length. Notice that no $m \to \infty$ limit has been taken: the window should only average the fast phase so that the drift is slow between t and $t+1$. The value of m remains implicit in θ and Σ as well as the remainders.

Continuous-time flows. Neglecting higher order terms, and assuming that the window has been taken such that the evolution of θ is slow between t and $t+1$, one obtains the raw mean–fluctuation updates:

$$\frac{d\theta}{dt} = -\eta \nabla E(\theta) - \frac{\eta}{2} \nabla H(\theta) : \Sigma_t, \tag{50}$$

$$\Sigma_{t+1} = (\mathrm{Id} - \eta H(\theta_t)) \Sigma_t (\mathrm{Id} - \eta H(\theta_t)). \tag{51}$$

The diagnostics in Appendix D test these approximations on individual trajectories, while Proposition A.15 controls the raw-to-coarse covariance error along a prescribed Hessian path with a fixed invariant subspace.

A.3 Closure validity and a two-cycle counterexample

Small oscillations alone do not guarantee a long-time approximation by MF. The following example has a constant windowed GD covariance, whereas both MF and raw MF predict growing covariance.

Let $\eta = 1$ and $E(x) = (2+\varepsilon)x^2/2 - x^4/4$, with $0 < \varepsilon < 1$. The GD update is

$$x_{n+1} = T(x_n), \qquad T(x) = x - E'(x) = -(1+\varepsilon)x + x^3.$$

At $x = \pm\sqrt{\varepsilon}$, the identity $x^3 = \varepsilon x$ gives $T(x) = -x$. Thus, starting from $x_0 = \sqrt{\varepsilon}$, the exact trajectory is

$$x_n = (-1)^n \sqrt{\varepsilon}, \qquad x_{2k} = \sqrt{\varepsilon}, \quad x_{2k+1} = -\sqrt{\varepsilon}.$$

The two values are not fixed points: GD exchanges them at each step, while each parity subsequence remains constant. Since $T'(\pm\sqrt{\varepsilon}) = -1 + 2\varepsilon$, the derivative of the two-step map at either point is $(-1+2\varepsilon)^2 < 1$, so this cycle is locally attracting. For any even window length m, each window contains equally many positive and negative iterates; hence its mean and covariance are exactly

$$\theta_n = 0, \qquad \Sigma_n = \varepsilon.$$

Initialize MF and raw MF with these same values, $(\theta_0, \Sigma_0) = (0, \varepsilon)$. Their mean remains zero because $E'(0) = E'''(0) = 0$. The Hessian at this mean is $H(0) = 2 + \varepsilon$, so their covariance equations reduce to

$$\frac{d\Sigma_{\mathrm{MF}}}{dt} = (2+\varepsilon)\varepsilon \Sigma_{\mathrm{MF}}, \qquad \Sigma_{\mathrm{raw},n+1} = (1+\varepsilon)^2 \Sigma_{\mathrm{raw},n}.$$

Consequently,

$$\Sigma_{\mathrm{MF}}(t) = \varepsilon e^{(2+\varepsilon)\varepsilon t}, \qquad \Sigma_{\mathrm{raw},n} = \varepsilon (1+\varepsilon)^{2n}.$$

The discrepancy comes from the cubic term discarded by the fluctuation closure. On the exact orbit, this term cancels the excess amplification in the linearized update:

$$-(1+\varepsilon)x_n + x_n^3 = -(1+\varepsilon)x_n + \varepsilon x_n = -x_n.$$

Without this cancellation, raw MF multiplies the variance by $(1+\varepsilon)^2$ at each step, and MF predicts the corresponding exponential growth. Both relative covariance errors $|\Sigma - \varepsilon|/\varepsilon$ vanish as $\varepsilon \to 0$ over a fixed time interval. However, at $t = c/\varepsilon$ and $n = \lfloor c/\varepsilon \rfloor$, respectively, they tend to $e^{2c} - 1$ for every fixed $c > 0$. Thus the relative error becomes order one on the timescale ε^{-1}, despite the vanishing oscillation amplitude $\sqrt{\varepsilon}$. Nevertheless, their mutual gap satisfies $|\Sigma_{\mathrm{MF}}(n) - \Sigma_{\mathrm{raw},n}|/\varepsilon \to 0$ at $n = \lfloor c/\varepsilon \rfloor$: raw/coarse agreement does not imply relative covariance accuracy against GD.

A.4 Well-posedness and positivity

Proposition A.9 (Well-posedness, positivity, and dissipation)**.** *Let $E \in C_b^4(\mathbb{R}^d)$, $\eta > 0$, and $(\theta_0, \Sigma_0) \in \mathbb{R}^d \times \mathbb{R}^{d\times d}$. Then* (4) *has a unique global solution. If Σ_0 is symmetric positive semidefinite, so is Σ_t for every $t \geq 0$. Moreover, we have for all $t \geq 0$,*

$$\frac{d}{dt} F(\theta_t, \Sigma_t) = -\eta \|\nabla_\theta F(\theta_t, \Sigma_t)\|^2 - \frac{\eta}{2} \big[H(\theta_t)(2\mathrm{Id} - \eta H(\theta_t)) H(\theta_t) \big] : \Sigma_t. \tag{52}$$

In particular, the free energy is nonincreasing in the stable regime where $\lambda_{\max}(H(\theta_t)) \leq 2/\eta$.

Proof of Proposition A.9. The vector field is locally Lipschitz. Since the derivatives of E are bounded,

$$\|\dot{\Sigma}\| \le C\|\Sigma\|, \qquad |\dot{\theta}| \le C(1 + \|\Sigma\|).$$

Gronwall's inequality prevents finite-time blow-up and gives global well-posedness. Symmetry is preserved. If $\Sigma \ge 0$ and $v \in \ker \Sigma$, then

$$v^T \dot{\Sigma} v = \eta^2 (Hv)^T \Sigma (Hv) \ge 0.$$

The invariance criterion for closed convex sets therefore preserves the positive semidefinite cone.

It remains to differentiate the free energy. Since $\dot{\theta} = -\eta \nabla_\theta F$,

$$\begin{aligned} \frac{\mathrm{d}F}{\mathrm{d}t} &= \nabla_\theta F \cdot \dot{\theta} + \frac{1}{2} H : \dot{\Sigma} \\ &= -\eta \|\nabla_\theta F\|^2 - \frac{\eta}{2} \big[H(2\mathrm{Id} - \eta H) H \big] : \Sigma. \end{aligned}$$

Under the sharpness bound, $H(2\mathrm{Id} - \eta H)H$ is positive semidefinite; its Frobenius product with $\Sigma \ge 0$ is therefore nonnegative. This proves (52).

□

A.5 Rank and quantitative dissipation

Rank of the fluctuation covariance. The rank of Σ is generally not preserved along the MF flow: the following example starts from rank one and becomes full rank immediately. In contrast, the flat-landscape limit preserves rank throughout its existence interval (Proposition A.14).

Take $d = 2$, $\eta = 1$, and a smooth bounded function which agrees near the origin with

$$E(x, y) = xy + \frac{x^3}{6}.$$

With $\theta_0 = 0$ and $\Sigma_0 = \mathrm{diag}(1, 0)$, one has

$$H(0) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \nabla H(0) : \Sigma_0 = (1, 0).$$

Consequently $\dot{\theta}_0 = (-1/2, 0)$ and

$$\dot{\Sigma}_0 = -(H(0)\Sigma_0 + \Sigma_0 H(0)) + H(0)\Sigma_0 H(0) = \begin{pmatrix} 0 & -1 \\ -1 & 1 \end{pmatrix}.$$

Thus

$$\Sigma_t = \begin{pmatrix} 1 & -t \\ -t & t \end{pmatrix} + O(t^2), \qquad \det \Sigma_t = t + O(t^2) > 0,$$

and the covariance becomes full rank immediately.

Modified Polyak-Łojasiewicz inequality. Suppose that $\eta \lambda_{\max}(H) \le 2$, that $F \ge F_\star$ along the flow, and that we have the following *modified Polyak-Łojasiewicz inequality*:

$$\exists \mu > 0, \forall t \ge 0, \quad \|\nabla_\theta F(\theta_t, \Sigma_t)\|^2 \ge \frac{\mu}{\eta} \big(F(\theta_t, \Sigma_t) - F_\star \big).$$

Then, for all $t \ge 0$, we have the exponential convergence $F(\theta_t, \Sigma_t) - F_\star \le e^{-\mu t}(F(\theta_0, \Sigma_0) - F_\star)$.

The claim follows directly from (52):

$$\frac{\mathrm{d}}{\mathrm{d}t}(F - F_\star) \le -\eta \|\nabla_\theta F\|^2 \le -\mu (F - F_\star).$$

Gronwall's inequality gives the stated rate.

A.6 STABILITY OF STATIONARY STATES

In this section, we analyse the stability (or instability) of stationary states of the MF dynamics.

Roughly speaking, the stationary states can be classified into two categories: no oscillating states (i.e. $\Sigma_\star = 0$) which can be stable or unstable depending on the sharpness of the Hessian, and oscillating states (i.e. $\Sigma_\star \neq 0$). To each oscillating stationary state, we can associate a physically meaningful state for which the covariance is purely in the EoS subspace. The stability of these states is then determined by the sharpness of the Hessian on the EoS subspace.

On the reduction to purely EoS stationary states. There are two distinct statements behind the reduction to a pure EoS covariance. The first is an exact property of the MF stationary equations: Lemma A.11 shows that

$$(\theta_\star, \Sigma_{0,\star} + \Sigma_{2,\star}) \text{ stationary} \quad \Longrightarrow \quad (\theta_\star, \Sigma_{2,\star}) \text{ stationary}.$$

Thus every stationary MF state has a pure EoS representative with the same mean and the same EoS covariance block. However, this does not assert that either covariance is generated by a temporal window of GD iterates.

To fix ideas, consider the *frozen homogeneous linear model*

$$\xi_{k+1} = A_\star \xi_k, \qquad A_\star = \mathrm{Id} - \eta H_\star. \tag{53}$$

For $z_k = P_0 \xi_k$ and $w_k = P_2 \xi_k$, the spectral identities $P_0 A_\star = P_0$ and $P_2 A_\star = -P_2$ give

$$z_{k+1} = z_k, \qquad w_{k+1} = -w_k.$$

The constant flat component has zero centered temporal covariance, whereas the EoS component is period two. This motivates the modeling restriction

$$\Sigma_{0,\star} = 0, \qquad \text{equivalently} \qquad \mathrm{Ran}\, \Sigma_\star \subset V_2, \tag{54}$$

but does not prove it for temporal windows of nonlinear GD. In particular, it is a selection principle, not a realizability theorem.

We prove Theorem 2.3, stated in the main text. We use $\mathrm{Sym}(V)$ for the self-adjoint maps on a subspace V.

The proof is split into three lemmas. The first Lemma A.11 addresses the covariance decomposition alluded to above, while the two remaining Lemma A.12 and Lemma A.13 address zero-covariance and pure EoS states, respectively.

We first state again the main Theorem of this section:

Theorem A.10 (Stability in the locally convex regime)**.** *Let $(\theta_\star, \Sigma_\star)$ be stationary for* (4)*, with $\Sigma_\star \geq 0$, and suppose that E is locally convex near $\theta_\star$. Set*

$$H_\star = H(\theta_\star), \qquad V_0 = \ker H_\star, \qquad V_2 = \ker(2\mathrm{Id} - \eta H_\star),$$

and let P_0, P_2 be the corresponding orthogonal projections. We write $\mathrm{Sym}(V)$ for the self-adjoint linear maps from V to itself.

If $\Sigma_\star = 0$, then $\nabla E(\theta_\star) = 0$. The state is stable if and only if $0 \leq \eta H_\star \leq 2\mathrm{Id}$, and it is strictly stable if and only if $0 < \eta H_\star < 2\mathrm{Id}$. If $\lambda_{\max}(\eta H_\star) > 2$, it has an exponentially growing covariance perturbation.

Suppose now that $\Sigma_\star \neq 0$. Then

$$\mathrm{Ran}\, \Sigma_\star \subset \ker H_\star \oplus \ker(2\mathrm{Id} - \eta H_\star), \tag{55}$$

and the fluctuations in the flat and EoS subspaces are decorrelated, in the sense that

$$u^T \Sigma_\star w = 0 \qquad \textit{for every } u \in \ker H_\star \textit{ and } w \in \ker(2\mathrm{Id} - \eta H_\star).$$

Consequently,

$$\Sigma_\star = \Sigma_{0,\star} + \Sigma_{2,\star}, \qquad \Sigma_{0,\star} = P_0 \Sigma_\star P_0, \qquad \Sigma_{2,\star} = P_2 \Sigma_\star P_2.$$

Moreover, $(\theta_\star, \Sigma_{2,\star})$ is also stationary. Thus every stationary state can be assigned a pure EoS representative with the same $\theta_\star$, and the stability analysis below focuses on these representatives.

Finally, consider a nonzero pure EoS representative and assume that

$$0 \neq \Sigma_\star \geq 0, \qquad \operatorname{Ran} \Sigma_\star \subset V_2, \qquad H_\star^{\text{eff}} := \nabla_\theta^2 F(\theta_\star, \Sigma_\star) = H_\star + \frac{1}{2}\nabla^2 H(\theta_\star) : \Sigma_\star \geq 0.$$

If every eigenvalue of $\eta H_\star$ on $V_2^\perp$ lies in $[0, 2)$, then the state is stable. Conversely, an eigenvalue of $\eta H_\star|_{V_2^\perp}$ larger than 2 gives an exponentially growing covariance perturbation and hence instability.

Lemma A.11 (Stationary covariance decomposition)**.** *Let $(\theta_\star, \Sigma_\star)$ be stationary for* (4)*, with $\Sigma_\star \geq 0$, and suppose that E is locally convex near $\theta_\star$. Set*

$$H_\star = H(\theta_\star), \qquad V_0 = \ker H_\star, \qquad V_2 = \ker(2\mathrm{Id} - \eta H_\star),$$

and let P_0, P_2 be the corresponding orthogonal projections. Then

$$\operatorname{Ran} \Sigma_\star \subset V_0 \oplus V_2, \qquad u^T \Sigma_\star w = 0 \quad (u \in V_0,\ w \in V_2).$$

Consequently,

$$\Sigma_\star = \Sigma_{0,\star} + \Sigma_{2,\star}, \qquad \Sigma_{0,\star} = P_0 \Sigma_\star P_0, \qquad \Sigma_{2,\star} = P_2 \Sigma_\star P_2.$$

Moreover, $(\theta_\star, \Sigma_{2,\star})$ is stationary.

Proof of Lemma A.11. Local convexity gives $H_\star = H(\theta_\star) \geq 0$. Set $A_\star = \mathrm{Id} - \eta H_\star$. The covariance equation factors as

$$\dot{\Sigma} = (\mathrm{Id} - \eta H)\Sigma(\mathrm{Id} - \eta H) - \Sigma, \tag{56}$$

so covariance stationarity is equivalent to

$$A_\star \Sigma_\star A_\star = \Sigma_\star. \tag{57}$$

We first prove the support inclusion. Let U be an orthogonal matrix such that

$$U^T H_\star U = D = \operatorname{diag}(\lambda_1, \ldots, \lambda_d).$$

We distinguish the covariance from its matrix in this eigenbasis by writing

$$\widetilde{\Sigma}_\star = U^T \Sigma_\star U, \qquad \widetilde{A}_\star = U^T A_\star U = \operatorname{diag}(a_1, \ldots, a_d), \qquad a_i = 1 - \eta\lambda_i.$$

Multiplying (57) on the left by U^T and on the right by U gives

$$\widetilde{A}_\star \widetilde{\Sigma}_\star \widetilde{A}_\star = \widetilde{\Sigma}_\star.$$

Therefore, entry by entry,

$$(a_i a_j - 1)(\widetilde{\Sigma}_\star)_{ij} = 0, \qquad 1 \leq i, j \leq d. \tag{58}$$

Taking $i = j$ yields

$$(a_i^2 - 1)(\widetilde{\Sigma}_\star)_{ii} = 0.$$

Since $\lambda_i \geq 0$, the coefficient vanishes precisely when $\lambda_i \in \{0, 2/\eta\}$. Otherwise $(\widetilde{\Sigma}_\star)_{ii} = 0$. The matrix $\widetilde{\Sigma}_\star$ is positive semidefinite. Hence each of its 2×2 principal submatrices is positive semidefinite and has nonnegative determinant:

$$\det \begin{pmatrix} (\widetilde{\Sigma}_\star)_{ii} & (\widetilde{\Sigma}_\star)_{ij} \\ (\widetilde{\Sigma}_\star)_{ij} & (\widetilde{\Sigma}_\star)_{jj} \end{pmatrix} \geq 0.$$

Equivalently,

$$|(\widetilde{\Sigma}_\star)_{ij}|^2 \leq (\widetilde{\Sigma}_\star)_{ii}(\widetilde{\Sigma}_\star)_{jj}.$$

Thus $(\widetilde{\Sigma}_\star)_{ii} = 0$ implies $(\widetilde{\Sigma}_\star)_{ij} = 0$ for every j; by symmetry, the entire i-th row and column vanish. Consequently, if

$$V = \operatorname{span}\{e_i : \lambda_i \in \{0, 2/\eta\}\},$$

then $\operatorname{Ran} \widetilde{\Sigma}_\star \subset V$.

We now return explicitly to the original coordinates. Since $\Sigma_\star = U\widetilde{\Sigma}_\star U^T$, for every $x \in \mathbb{R}^d$,

$$\Sigma_\star x = U\widetilde{\Sigma}_\star(U^T x) \in UV.$$

Moreover, the columns Ue_i are the eigenvectors of $H_\star$ with eigenvalues λ_i. Therefore

$$UV = \operatorname{span}\{Ue_i : \lambda_i = 0 \text{ or } \lambda_i = 2/\eta\} = \ker H_\star \oplus \ker(2\mathrm{Id} - \eta H_\star).$$

Combining the last two displays proves the support inclusion. Finally, if $\lambda_i = 0$ and $\lambda_j = 2/\eta$, then $a_i a_j = -1$, so (58) gives $(\widetilde{\Sigma}_\star)_{ij} = 0$. The corresponding transposed entry also vanishes. Hence, for every $u \in \ker H_\star$ and $w \in \ker(2\mathrm{Id} - \eta H_\star)$, expansion in the eigenbasis U gives

$$u^T \Sigma_\star w = 0.$$

When $\Sigma_\star = \mathbb{E}[\xi\xi^T]$, this identity reads

$$\mathbb{E}\big[(u^T\xi)(w^T\xi)\big] = 0.$$

Thus the fluctuations in the flat and EoS subspaces are decorrelated.

We now show that the flat block can be removed without changing the stationary mean. Let P_0 and P_2 be the orthogonal projections onto

$$V_0 = \ker H_\star, \qquad V_2 = \ker(2\mathrm{Id} - \eta H_\star),$$

respectively. The support and decorrelation properties proved above yield

$$\Sigma_\star = \Sigma_{0,\star} + \Sigma_{2,\star}, \qquad \Sigma_{0,\star} = P_0\Sigma_\star P_0, \qquad \Sigma_{2,\star} = P_2\Sigma_\star P_2. \tag{59}$$

Fix $u \in V_0$. Local convexity implies $g_u(\theta) := u^T H(\theta) u \geq 0$ near $\theta_\star$, while $g_u(\theta_\star) = 0$. Therefore $\theta_\star$ is a local minimizer of g_u and

$$\nabla g_u(\theta_\star) = \nabla^3 E(\theta_\star)[u, u] = 0. \tag{60}$$

Take a spectral decomposition $\Sigma_{0,\star} = \sum_r s_r u_r u_r^T$, with $s_r \geq 0$ and $u_r \in V_0$. Then

$$H_\star \Sigma_{0,\star} = \Sigma_{0,\star} H_\star = 0, \qquad \nabla H(\theta_\star) : \Sigma_{0,\star} = \sum_r s_r \nabla g_{u_r}(\theta_\star) = 0. \tag{61}$$

Since $A_\star = \mathrm{Id}$ on V_0, the covariance identity also gives

$$A_\star \Sigma_{0,\star} A_\star = \Sigma_{0,\star}.$$

Subtracting these flat contributions from the stationary equations for $(\theta_\star, \Sigma_\star)$ gives

$$\nabla E(\theta_\star) + \frac{1}{2}\nabla H(\theta_\star) : \Sigma_{2,\star} = 0,$$
$$A_\star \Sigma_{2,\star} A_\star = A_\star \Sigma_\star A_\star - A_\star \Sigma_{0,\star} A_\star = \Sigma_\star - \Sigma_{0,\star} = \Sigma_{2,\star}.$$

Thus $(\theta_\star, \Sigma_{2,\star})$ is stationary: removing the flat block changes neither the stationary mean nor the EoS block. □

For the two stability lemmas below, we record the linearization at an arbitrary stationary state $(\theta_\star, \Sigma_\star)$. Set

$$H_\star = H(\theta_\star), \qquad A_\star = \mathrm{Id} - \eta H_\star, \qquad H_\star^{\mathrm{eff}} = H_\star + \frac{1}{2} D^2 H(\theta_\star) : \Sigma_\star = \nabla_\theta^2 F(\theta_\star, \Sigma_\star).$$

If x and $S = S^T$ are perturbations of $\theta_\star$ and $\Sigma_\star$, respectively, differentiation of (4) gives

$$\dot{x} = -\eta\left(H_\star^{\mathrm{eff}} x + \frac{1}{2}\nabla H(\theta_\star) : S\right), \tag{62}$$

$$\dot{S} = A_\star S A_\star - S - \eta\big(DH(\theta_\star)[x]\Sigma_\star A_\star + A_\star \Sigma_\star DH(\theta_\star)[x]\big). \tag{63}$$

Lemma A.12 (Zero-covariance states)**.** *Let $(\theta_\star, 0)$ be stationary for* (4)*, and suppose that E is locally convex near $\theta_\star$. Set $H_\star = H(\theta_\star)$. Then $\nabla E(\theta_\star) = 0$. The state is stable if and only if $0 \leq \eta H_\star \leq 2\mathrm{Id}$, and it is strictly stable if and only if $0 < \eta H_\star < 2\mathrm{Id}$. If $\lambda_{\max}(\eta H_\star) > 2$, it has an eigenvalue with positive real part.*

Proof of Lemma A.12. The stationary mean equation gives $\nabla E(\theta_\star) = 0$, while local convexity gives $H_\star \geq 0$. Since $\Sigma_\star = 0$, one has $H_\star^{\text{eff}} = H_\star$. Choose an orthogonal matrix U whose columns $u_i = Ue_i$ form an eigenbasis of $H_\star$:

$$H_\star u_i = \lambda_i u_i, \qquad A_\star u_i = (1 - \eta\lambda_i)u_i.$$

Set

$$\widetilde{x} = U^T x, \qquad \widetilde{S} = U^T S U, \qquad D = U^T H_\star U = \operatorname{diag}(\lambda_i), \qquad \widetilde{A}_\star = U^T A_\star U = \operatorname{diag}(1-\eta\lambda_i).$$

Thus the vector equation is written in the eigenbasis by multiplying it on the left by U^T, whereas the covariance equation is conjugated by multiplying it on the left by U^T and on the right by U. Since $S = U\widetilde{S}U^T$, this gives

$$\dot{\widetilde{x}} = -\eta D\widetilde{x} - \frac{\eta}{2}U^T\big(\nabla H(\theta_\star) : (U\widetilde{S}U^T)\big),$$
$$\dot{\widetilde{S}} = \widetilde{A}_\star\widetilde{S}\widetilde{A}_\star - \widetilde{S}.$$

The covariance equation is diagonal entry by entry:

$$\dot{\widetilde{S}}_{ij} = \big((1-\eta\lambda_i)(1-\eta\lambda_j) - 1\big)\widetilde{S}_{ij}.$$

Consequently, the mean direction u_i has rate $-\eta\lambda_i$, while the symmetric covariance direction $u_iu_j^T + u_ju_i^T = U(e_ie_j^T + e_je_i^T)U^T$ has rate

$$(1-\eta\lambda_i)(1-\eta\lambda_j) - 1. \tag{64}$$

These are continuous-time MF rates, not the discrete GD multipliers $1-\eta\lambda_i$. The covariance equation does not depend on x, whereas the covariance perturbation may force the mean through $\nabla H(\theta_\star) : S$. The full linearized system is therefore block triangular, so this forcing does not change its eigenvalues. Since $\lambda_i \geq 0$, all these spectral values are nonpositive exactly when $0 \leq \eta H_\star \leq 2\mathrm{Id}$, and they are all strictly negative exactly when $0 < \eta H_\star < 2\mathrm{Id}$. If some $\eta\lambda_i > 2$, the diagonal covariance perturbation $u_iu_i^T$ has the positive rate

$$(1-\eta\lambda_i)^2 - 1 = \eta\lambda_i(\eta\lambda_i - 2) > 0.$$

This gives the complete linear classification of zero-covariance states. □

Lemma A.13 (Pure EoS states)**.** *Let $(\theta_\star, \Sigma_\star)$ be stationary for* (4)*, and suppose that E is locally convex near $\theta_\star$. Set*

$$H_\star = H(\theta_\star), \qquad V_2 = \ker(2\mathrm{Id} - \eta H_\star), \qquad H_\star^{\text{eff}} = \nabla_\theta^2 F(\theta_\star, \Sigma_\star).$$

Assume that

$$0 \neq \Sigma_\star \geq 0, \qquad \operatorname{Ran}\Sigma_\star \subset V_2, \qquad H_\star^{\text{eff}} \geq 0.$$

If every eigenvalue of $\eta H_\star|_{V_2^\perp}$ belongs to $[0, 2)$, then the state is stable. Conversely, if this restriction has an eigenvalue larger than 2, the state is unstable.

Proof of Lemma A.13. The proof uses coordinates adapted simultaneously to the Hessian and the stationary covariance. This does not diagonalize the whole linearization: the third derivative of E still couples its mean and covariance parts. It does, however, make the sign of every spectral value accessible through one weighted eigenmode identity.

Let P and Q be the orthogonal projections onto V_2 and $V_2^\perp$. Since $A_\star := \mathrm{Id} - \eta H_\star$ equals $-\mathrm{Id}$ on V_2, every covariance perturbation $S = S^T$ decomposes as

$$S = B + Z + Z^T + C, \qquad B = PSP, \quad Z = PSQ, \quad C = QSQ.$$

Projecting (63) onto these blocks gives

$$\dot{x} = -\eta\left(H_\star^{\text{eff}}x + \frac{1}{2}\nabla H(\theta_\star) : (B + Z + Z^T + C)\right), \tag{65}$$

$$\dot{B} = \eta\big(PDH(\theta_\star)[x]P\Sigma_\star + \Sigma_\star PDH(\theta_\star)[x]P\big), \tag{66}$$

$$\dot{Z} = -Z(2\mathrm{Id} - \eta H_\star|_{V_2^\perp}) + \eta\Sigma_\star PDH(\theta_\star)[x]Q, \tag{67}$$

$$\dot{C} = A_\perp C A_\perp - C, \qquad A_\perp = A_\star|_{V_2^\perp}. \tag{68}$$

Assume first that the eigenvalues of $\eta H_\star|_{V_2^\perp}$ belong to $[0,2)$. Since $H_\star = (2/\eta)\mathrm{Id}$ on V_2, we may choose an orthonormal basis (e_i) of V_2 that diagonalizes $\Sigma_\star$:

$$\Sigma_\star e_i = \sigma_i e_i, \qquad \sigma_i \geq 0.$$

Choose also an orthonormal eigenbasis (u_α) of $V_2^\perp$ such that

$$H_\star u_\alpha = \lambda_\alpha u_\alpha, \qquad \delta_\alpha := 2 - \eta\lambda_\alpha > 0.$$

For $G = DH(\theta_\star)[x]$, write

$$B_{ij} = e_i^T B e_j, \qquad Z_{i\alpha} = e_i^T Z u_\alpha, \qquad G_{ij} = e_i^T G e_j, \qquad G_{i\alpha} = e_i^T G u_\alpha.$$

Equations (66)– (68) then read

$$\dot{B}_{ij} = \eta(\sigma_i + \sigma_j)G_{ij}, \tag{69}$$

$$\dot{Z}_{i\alpha} = -\delta_\alpha Z_{i\alpha} + \eta\sigma_i G_{i\alpha}, \tag{70}$$

$$\dot{C}_{\alpha\beta} = \big((1-\eta\lambda_\alpha)(1-\eta\lambda_\beta) - 1\big)C_{\alpha\beta}. \tag{71}$$

In particular, the C-variables form an autonomous subsystem. Their rates are nonpositive because $|1 - \eta\lambda_\alpha| \leq 1$.

Full rank on the EoS subspace. Suppose first that $\operatorname{Ran}\Sigma_\star = V_2$, so every σ_i is positive. The full linearized system is block triangular, with the autonomous C-subsystem on the diagonal. It therefore remains to study the subsystem obtained by setting $C = 0$.

Let $r \in \mathbb{C}$ be an eigenvalue of this subsystem, with nonzero complex eigenvector (x, B, Z). Replace the time derivatives in (65), (69), and (70) by multiplication by r. Take the real part after multiplying these equations, respectively, by

$$x^*, \qquad \frac{\overline{B_{ij}}}{2(\sigma_i + \sigma_j)}, \qquad \frac{\overline{Z_{i\alpha}}}{\sigma_i},$$

and summing over the matrix indices. The identity

$$DH(\theta_\star)[x] : Y = x \cdot \big(\nabla H(\theta_\star) : Y\big)$$

and the symmetry of $DH(\theta_\star)[x]$ give

$$\Re \sum_{i,j} G_{ij}\overline{B_{ij}} = \Re\big(x^*(\nabla H(\theta_\star) : B)\big),$$

$$2\Re \sum_{i,\alpha} G_{i\alpha}\overline{Z_{i\alpha}} = \Re\big(x^*(\nabla H(\theta_\star) : (Z + Z^T))\big).$$

Thus the B- and Z-couplings cancel those in the x-equation, and we obtain

$$\Re(r)\left(|x|^2 + \frac{1}{2}\sum_{i,j}\frac{|B_{ij}|^2}{\sigma_i + \sigma_j} + \sum_{i,\alpha}\frac{|Z_{i\alpha}|^2}{\sigma_i}\right)$$

$$= -\eta x^* H_\star^{\mathrm{eff}} x - \sum_{i,\alpha}\frac{\delta_\alpha}{\sigma_i}|Z_{i\alpha}|^2 \leq 0. \tag{72}$$

The factor multiplying $\Re(r)$ is positive because $(x, B, Z) \neq 0$. The right-hand side is nonpositive because $H_\star^{\mathrm{eff}} \geq 0$ and $\delta_\alpha > 0$. Hence $\Re(r) \leq 0$, which proves stability in the full-rank case.

Rank-deficient covariance. Suppose now that some σ_i vanish, and set

$$I_+ = \{i : \sigma_i > 0\}, \qquad I_0 = \{i : \sigma_i = 0\}.$$

The coordinates excluded by the denominators in (72) already have explicit nonpositive rates. Indeed, (69)– (70) give

$$\dot{B}_{ij} = 0 \quad (i,j \in I_0), \qquad \dot{Z}_{i\alpha} = -\delta_\alpha Z_{i\alpha} \quad (i \in I_0).$$

Together with the C-variables, these coordinates form an autonomous block. They may force the remaining variables through the x-equation, but the remaining variables do not force them; hence the full system is block triangular and their rates remain part of its spectrum.

Set this autonomous block to zero. The remaining coordinates satisfy $\sigma_i + \sigma_j > 0$ for B_{ij} and $i \in I_+$ for $Z_{i\alpha}$. Repeating the preceding eigenmode calculation, now with the sums restricted to these indices, gives

$$\begin{aligned}\Re(r)\left(|x|^2 + \frac{1}{2}\sum_{\sigma_i+\sigma_j>0}\frac{|B_{ij}|^2}{\sigma_i+\sigma_j} + \sum_{i\in I_+,\alpha}\frac{|Z_{i\alpha}|^2}{\sigma_i}\right)\\ = -\eta x^* H_\star^{\mathrm{eff}} x - \sum_{i\in I_+,\alpha}\frac{\delta_\alpha}{\sigma_i}|Z_{i\alpha}|^2 \le 0.\end{aligned}$$

Thus every rate of the remaining block also has nonpositive real part. This proves stability in the rank-deficient case.

Finally, suppose that $H_\star u = \lambda u$ for some $u \in V_2^\perp$ with $\eta\lambda > 2$. Equation (68) shows that $C = uu^T$ has rate

$$(1-\eta\lambda)^2 - 1 = \eta\lambda(\eta\lambda - 2) > 0.$$

This rate belongs to the autonomous C-block and hence to the spectrum of the full block-triangular linearization. The stationary state is therefore unstable. □

Proof of Theorem 2.3. The covariance decomposition is Lemma A.11. The zero-covariance case is Lemma A.12. The nonzero pure EoS case is Lemma A.13. □

A.7 Small-step and flat-landscape limits

Proposition A.14 (Gradient-flow and flat-landscape limits)**.** *Let $E \in C^4(\mathbb{R}^d)$.*

1. *Fix $(\theta_0, \bar\Sigma_0)$ and, for each $\eta > 0$, let $(\theta^\eta, \Sigma^\eta)$ denote the MF solution in the rescaled time $\tau = \eta t$, with $\theta^\eta(0) = \theta_0$ and $\Sigma^\eta(0) = \eta\bar\Sigma_0$. Let θ solve*

$$\frac{\mathrm{d}\theta}{\mathrm{d}\tau} = -\nabla E(\theta), \qquad \theta(0) = \theta_0.$$

For every T smaller than the maximal existence time of θ, the rescaled MF solution exists on $[0,T]$ for all sufficiently small η, and there is a constant C_T such that

$$\sup_{0\le\tau\le T}|\theta^\eta(\tau) - \theta(\tau)| \le C_T\eta.$$

2. *Fix $\eta > 0$ and (θ_0, Σ_0). For each $\varepsilon > 0$, let $(\theta^\varepsilon, \Sigma^\varepsilon)$ denote the MF solution associated with the energy εE, in the rescaled time $\tau = \eta\varepsilon t$, and with initial data (θ_0, Σ_0). Let (θ, Σ) solve*

$$\begin{aligned}\frac{\mathrm{d}\theta}{\mathrm{d}\tau} &= -\nabla E(\theta) - \frac{1}{2}\nabla H(\theta) : \Sigma,\\ \frac{\mathrm{d}\Sigma}{\mathrm{d}\tau} &= -(H(\theta)\Sigma + \Sigma H(\theta)),\\ (\theta,\Sigma)|_{\tau=0} &= (\theta_0, \Sigma_0).\end{aligned} \tag{73}$$

For every T smaller than the maximal existence time of (θ, Σ), the rescaled MF solution exists on $[0,T]$ for all sufficiently small ε, and there is a constant C_T such that

$$\sup_{0\le\tau\le T}(|\theta^\varepsilon(\tau) - \theta(\tau)| + \|\Sigma^\varepsilon(\tau) - \Sigma(\tau)\|) \le C_T\varepsilon.$$

Moreover, $\operatorname{rank}\Sigma(\tau) = \operatorname{rank}\Sigma_0$ *throughout the maximal existence interval of the limiting solution.*

Proof of Proposition A.14. We treat the two limits separately.

Gradient-flow limit. Fix $T > 0$, let θ solve

$$\frac{\mathrm{d}\theta}{\mathrm{d}\tau} = -\nabla E(\theta), \qquad \theta(0) = \theta_0,$$

on $[0, T]$, and set

$$\bar{\Sigma}^\eta = \eta^{-1}\Sigma^\eta.$$

Since $\Sigma_0^\eta = \eta\bar{\Sigma}_0$, the rescaled equations are

$$\frac{\mathrm{d}\theta^\eta}{\mathrm{d}\tau} = -\nabla E(\theta^\eta) - \frac{\eta}{2}\nabla H(\theta^\eta) : \bar{\Sigma}^\eta, \tag{74}$$

$$\frac{\mathrm{d}\bar{\Sigma}^\eta}{\mathrm{d}\tau} = -\big(\bar{\Sigma}^\eta H(\theta^\eta) + H(\theta^\eta)\bar{\Sigma}^\eta\big) + \eta H(\theta^\eta)\bar{\Sigma}^\eta H(\theta^\eta), \qquad \bar{\Sigma}^\eta(0) = \bar{\Sigma}_0. \tag{75}$$

Only local bounds are needed. Let

$$R = 1 + \max_{0\le\tau\le T} |\theta(\tau)|, \qquad K = \{y \in \mathbb{R}^d : |y| \le R\}.$$

The limiting trajectory lies at distance at least one from the boundary of the compact set K. Let L be a Lipschitz constant for ∇E on K, and choose finite constants M_0, M_1 such that

$$\|H(y)\|_{\mathrm{op}} \le M_0, \qquad |\nabla H(y) : S| \le M_1\|S\| \quad (y \in K).$$

Let T_η be the first time at which θ^η leaves K, capped at T and at the maximal existence time of the rescaled MF solution. On $[0, T_\eta]$, (75) and $0 < \eta \le 1$ give

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\|\bar{\Sigma}^\eta\| \le (2M_0 + M_0^2)\|\bar{\Sigma}^\eta\|.$$

Hence, by Gronwall's inequality,

$$\sup_{0\le\tau\le T_\eta} \|\bar{\Sigma}^\eta(\tau)\| \le \|\bar{\Sigma}_0\| e^{(2M_0+M_0^2)T} =: M_T. \tag{76}$$

Subtracting the gradient-flow equation from (74) and using the local bounds on K yields, for $0 \le \tau \le T_\eta$,

$$|\theta^\eta(\tau) - \theta(\tau)| \le L\int_0^\tau |\theta^\eta(s) - \theta(s)|\,\mathrm{d}s + \frac{\eta}{2}M_1 M_T \tau.$$

A second application of Gronwall's inequality gives

$$\sup_{0\le\tau\le T_\eta} |\theta^\eta(\tau) - \theta(\tau)| \le \frac{\eta}{2}M_1 M_T T e^{LT} =: C_T\eta. \tag{77}$$

For η small enough that $C_T\eta < 1$, this estimate prevents θ^η from reaching the boundary of K before time T. The local continuation theorem for ODEs then gives $T_\eta = T$, and (77) proves that $\theta^\eta \to \theta$ uniformly on $[0, T]$. All constants used above depend only on the compact neighborhood K of the limiting trajectory.

Flat-landscape limit and rank preservation. Let $(\theta^\varepsilon, \Sigma^\varepsilon)$ denote the rescaled MF solution associated with the energy εE. Its equations are

$$\frac{\mathrm{d}\theta^\varepsilon}{\mathrm{d}\tau} = -\nabla E(\theta^\varepsilon) - \frac{1}{2}\nabla H(\theta^\varepsilon) : \Sigma^\varepsilon, \tag{78}$$

$$\frac{\mathrm{d}\Sigma^\varepsilon}{\mathrm{d}\tau} = -\big(\Sigma^\varepsilon H(\theta^\varepsilon) + H(\theta^\varepsilon)\Sigma^\varepsilon\big) + \eta\varepsilon H(\theta^\varepsilon)\Sigma^\varepsilon H(\theta^\varepsilon). \tag{79}$$

Fix $T > 0$ and let $z = (\theta, \Sigma)$ solve (73) on $[0, T]$ with the same initial data. On the product space, use the norm

$$|(y, S)|_* = |y| + \|S\|,$$

and define the compact ball

$$R = 1 + \max_{0\le\tau\le T} |z(\tau)|_*, \qquad \mathcal{K} = \{(y, S) : |(y, S)|_* \le R\}.$$

The limiting vector field

$$\mathcal{W}_0(y,S) = \left(-\nabla E(y) - \frac{1}{2}\nabla H(y) : S, -\big(SH(y) + H(y)S\big)\right)$$

is Lipschitz on $\mathcal{K}$; denote a Lipschitz constant by L_T. This uses only local regularity: on the projection of $\mathcal{K}$ onto $\mathbb{R}^d$, the derivatives of E involved in $\mathcal{W}_0$ are bounded. In particular, local boundedness of ∇H makes H locally Lipschitz, while local boundedness of $\nabla^2 H$ makes ∇H locally Lipschitz. On the same compact set, there is a constant M_T^{flat} such that

$$\eta\|H(y)SH(y)\| \le M_T^{\mathrm{flat}} \qquad ((y,S)\in\mathcal{K}).$$

Let T_ε be the first time at which $z^\varepsilon = (\theta^\varepsilon, \Sigma^\varepsilon)$ leaves $\mathcal{K}$, capped at T and at the maximal existence time of the rescaled MF solution. Equations (78)– (79) and (73) give, on $[0, T_\varepsilon]$,

$$|z^\varepsilon(\tau) - z(\tau)|_* \le L_T \int_0^\tau |z^\varepsilon(s) - z(s)|_* \,\mathrm{d}s + \varepsilon M_T^{\mathrm{flat}}\tau.$$

Therefore

$$\sup_{0\le\tau\le T_\varepsilon} |z^\varepsilon(\tau) - z(\tau)|_* \le \varepsilon M_T^{\mathrm{flat}} T e^{L_T T}. \tag{80}$$

For ε small enough, the right-hand side is smaller than one, so z^ε cannot reach the boundary of $\mathcal{K}$ before time T. The local continuation theorem then gives $T_\varepsilon = T$, and (80) proves uniform convergence on $[0,T]$ using only constants on a compact neighborhood of the limiting trajectory.

It remains to prove rank preservation for the limiting covariance. If P and Q solve

$$\frac{\mathrm{d}P}{\mathrm{d}\tau} = -H(\theta(\tau))P, \qquad P(0) = \mathrm{Id},$$
$$\frac{\mathrm{d}Q}{\mathrm{d}\tau} = QH(\theta(\tau)), \qquad Q(0) = \mathrm{Id},$$

then

$$\frac{\mathrm{d}}{\mathrm{d}\tau}(QP) = QH(\theta)P - QH(\theta)P = 0.$$

Thus $Q(\tau)P(\tau) = \mathrm{Id}$, so $P(\tau)$ is invertible. Define

$$\widetilde{\Sigma}(\tau) = P(\tau)\Sigma_0 P(\tau)^T.$$

Since $H(\theta)$ is symmetric, differentiation gives

$$\begin{aligned}\frac{\mathrm{d}\widetilde{\Sigma}}{\mathrm{d}\tau} &= -H(\theta)P\Sigma_0 P^T - P\Sigma_0 P^T H(\theta)\\ &= -H(\theta)\widetilde{\Sigma} - \widetilde{\Sigma}H(\theta), \qquad \widetilde{\Sigma}(0) = \Sigma_0.\end{aligned}$$

Hence $\widetilde{\Sigma}$ and Σ solve the same linear matrix ODE with the same initial condition, and uniqueness yields $\Sigma(\tau) = P(\tau)\Sigma_0 P(\tau)^T$. Multiplication on the left and right by invertible matrices preserves rank, so $\operatorname{rank}\Sigma(\tau) = \operatorname{rank}\Sigma_0$ throughout the maximal existence interval.

□

A.8 Prescribed-Hessian raw-to-coarse approximation

In this section, we provide a minimal argument for the validity of the coarse covariance equation, assuming that the Hessian is prescribed and varies at a bounded rate. This compares the two covariance equations, not either equation with GD.

Proposition A.15 (Prescribed-Hessian validity of the coarse covariance equation)**.** *Fix $T, L > 0$, $\eta > 0$, and $0 < \varepsilon \le 1/4$. Let $V \subset \mathbb{R}^d$ be a linear subspace and let $H \in C^1([0, T/\varepsilon]; \mathbb{S}_+^d)$, where $\mathbb{S}_+^d$ denotes the symmetric positive semidefinite matrices. Assume that $H(t)V \subset V$ for every t, that the active Hessian varies at bounded speed,*

$$\sup_{0\le t\le T/\varepsilon} \| \dot{H}(t)\Big|_V \|_{\mathrm{op}} \le L,$$

and that

$$\sup_{0\le t\le T/\varepsilon} \| \left(\eta H(t) - 2\mathrm{Id}\right)|_V \|_{\mathrm{op}} \le \varepsilon. \tag{81}$$

Consider the raw covariance recursion

$$\Sigma_{k+1}^{\mathrm{raw}} = (\mathrm{Id} - \eta H(k))\Sigma_k^{\mathrm{raw}}(\mathrm{Id} - \eta H(k)), \tag{82}$$

and its coarse counterpart

$$\frac{\mathrm{d}\Sigma}{\mathrm{d}t} = -\eta(\Sigma H(t) + H(t)\Sigma) + \eta^2 H(t)\Sigma H(t), \tag{83}$$

from the same $\Sigma_0 \ge 0$, with $\operatorname{Ran}\Sigma_0 \subset V$. Set $\Sigma^{\mathrm{raw}}(t) = \Sigma_k^{\mathrm{raw}}$ for $k \le t < k+1$. There is a constant $C_{T,L}$, independent of η, ε, H, and Σ_0, such that

$$\sup_{0\le t\le T/\varepsilon} \|\Sigma^{\mathrm{raw}}(t) - \Sigma(t)\| \le C_{T,L}\left(\varepsilon + \frac{\eta}{\varepsilon}\right)\|\Sigma_0\|. \tag{84}$$

In particular, if $\eta \le \varepsilon^2$, the error is at most $C_{T,L}\varepsilon\|\Sigma_0\|$ on $O(\varepsilon^{-1})$ algorithmic times.

Proof of Proposition A.15. Let P be the orthogonal projection onto V. Since $H(t)$ is symmetric and $H(t)V \subset V$, the orthogonal complement $V^\perp$ is also invariant under $H(t)$: for $v \in V$ and $w \in V^\perp$,

$$v \cdot H(t)w = H(t)v \cdot w = 0.$$

Consequently, $PH(t) = H(t)P$. Moreover, $\operatorname{Ran}\Sigma_0 \subset V$ and the symmetry of Σ_0 give $\Sigma_0 = P\Sigma_0 P$. The raw update then shows by induction that

$$\Sigma_k^{\mathrm{raw}} = P\Sigma_k^{\mathrm{raw}}P \quad \text{for every } k.$$

The coarse vector field has the same property: if $S = PSP$, then every term in (83) is supported in V. Uniqueness of the linear matrix ODE therefore gives $\Sigma(t) = P\Sigma(t)P$. It is thus enough to work on V; the behavior of $H(t)$ on $V^\perp$ never enters the two dynamics.

On V, set $D(t) = \left(\eta H(t) - 2\mathrm{Id}\right)|_V$. For a matrix $S : V \to V$, define

$$\mathcal{L}_t(S) = D(t)S + SD(t) + D(t)SD(t).$$

Expanding the raw map and the coarse vector field gives, respectively,

$$(\mathrm{Id} - \eta H(k))S(\mathrm{Id} - \eta H(k)) = S + \mathcal{L}_k(S), \tag{85}$$

$$-\eta(SH(t) + H(t)S) + \eta^2 H(t)SH(t) = \mathcal{L}_t(S). \tag{86}$$

Thus the raw recursion is the explicit Euler discretization, with unit step, of the non-autonomous coarse equation.

The tube assumption (81) implies

$$\|\mathcal{L}_t(S)\| \le 3\varepsilon\|S\|, \tag{87}$$

$$\|(\mathcal{L}_t - \mathcal{L}_s)(S)\| \le 3L\eta|t-s|\|S\|. \tag{88}$$

Indeed, the first estimate follows by bounding the three terms in $\mathcal{L}_t$. For the second, expand

$$D(t)SD(t) - D(s)SD(s) = (D(t) - D(s))SD(t) + D(s)S(D(t) - D(s))$$

and use $\|D(t) - D(s)\|_{\mathrm{op}} \le L\eta|t-s|$.

Before comparing the two dynamics, we bound their size. From (87) and Gronwall's inequality,

$$\|\Sigma(t)\| \le e^{3\varepsilon t}\|\Sigma_0\|.$$

Likewise, (85) gives

$$\|\Sigma_{k+1}^{\mathrm{raw}}\| \le (1 + 3\varepsilon)\|\Sigma_k^{\mathrm{raw}}\|.$$

Both covariances are therefore bounded by $C_T\|\Sigma_0\|$ for $t, k \le T/\varepsilon$. In particular, if $k \le t < k+1$, then

$$\|\Sigma(t) - \Sigma(k)\| \le C_T\varepsilon(t-k)\|\Sigma_0\|. \tag{89}$$

The exact coarse evolution over one step is

$$\Sigma(k+1) = \Sigma(k) + \mathcal{L}_k(\Sigma(k)) + r_k,$$

where

$$r_k = \int_k^{k+1} \big(\mathcal{L}_t(\Sigma(t)) - \mathcal{L}_k(\Sigma(k))\big)\,\mathrm{d}t.$$

Using (87) for the change of Σ, (88) for the change of H, and (89), we obtain

$$\|r_k\| \le C_{T,L}(\varepsilon^2 + \eta)\|\Sigma_0\|. \tag{90}$$

Subtracting this identity from the raw update yields

$$\|\Sigma^{\mathrm{raw}}_{k+1} - \Sigma(k+1)\| \le (1+3\varepsilon)\|\Sigma^{\mathrm{raw}}_k - \Sigma(k)\| + C_{T,L}(\varepsilon^2+\eta)\|\Sigma_0\|.$$

The initial error vanishes. Discrete Gronwall, for $k \le T/\varepsilon$, gives

$$\|\Sigma^{\mathrm{raw}}_k - \Sigma(k)\| \le C_{T,L}\left(\varepsilon + \frac{\eta}{\varepsilon}\right)\|\Sigma_0\|.$$

Together with (89), this proves (84) for the piecewise-constant interpolation of the raw covariance. Since the embedding of matrices supported in V into $\mathbb{R}^{d\times d}$ preserves the Frobenius norm, the same estimate holds in the ambient space. □

A.9 Raw-to-coarse Rod reduction and comparison with MF

We prove the two parts of Proposition 2.4 in turn. Throughout this subsection, $\|\cdot\|$ denotes the Frobenius norm and $\|\cdot\|_{\mathrm{op}}$ the operator norm.

A.9.1 From raw Rod flow to coarse Rod flow

Detailed statement of part (i). Fix $T > 0$, $C \ge 1$, and $\eta, a \in (0,1)$ with $\delta = a^2/\eta \in (0,1)$. Let (θ_t, Σ_t) solve (6) on $[0, T/\eta]$, and use the same principal extent ξ_t when evaluating the coarse field $\mathcal{V}_{\mathrm{R}}$. It is enough that E_η be C^5 on a neighborhood of the segments $\{\theta_t + s\xi_t : |s| \le 1\}$ and that, uniformly in t,

$$|\nabla E_\eta(\theta_t)| \le C\delta, \qquad C^{-1}a^2 \le |\xi_t|^2 \le Ca^2, \qquad \|D^j E_\eta\|_{\mathrm{op}} \le \frac{C}{\eta} \quad (2 \le j \le 5). \tag{91}$$

Here C is independent of η, a, t, and each derivative norm is the supremum of the multilinear operator norm on the corresponding segment. The two conclusions of Proposition 2.4(i) are

$$\int_0^{T/\eta} \left| \frac{d\theta_t}{dt} - \mathcal{V}^\theta_{\mathrm{R}}(\theta_t, \Sigma_t) \right| dt \le C_T\delta, \tag{92}$$

$$\frac{1}{a^2}\int_0^{T/\eta} \left\| \frac{d\Sigma_t}{dt} - \mathcal{V}^\Sigma_{\mathrm{R}}(\theta_t, \Sigma_t) \right\|_F dt \le C_T\delta, \tag{93}$$

where C_T depends only on C, T. The first estimate is an absolute mean defect; the second divides the covariance defect by its natural scale a^2.

Proof of Proposition 2.4, part (i). Taylor expanding in ξ the equations (6), the odd powers cancel between the two endpoints, and we obtain

$$\frac{d\theta}{dt} = -\eta\nabla E(\theta) - \frac{\eta}{2}\nabla H(\theta) : \xi\xi^T + R^\theta, \tag{94}$$

$$\frac{d\Sigma}{dt} = \frac{\eta^2}{2}\left(\nabla E(\theta)\nabla E(\theta)^T + H(\theta)\xi\xi^T H(\theta)\right) - 2\Sigma + R^\Sigma. \tag{95}$$

We include all terms proportional to η^2 in the mean equation in R^θ. In the covariance equation, R^Σ includes the products of $\nabla E(\theta)$ with the quadratic Taylor correction $\frac{1}{2}\nabla H(\theta) : \xi\xi^T$. Explicitly,

$$|R^\theta| \le \frac{\eta}{24}\|D^5E\|_{\mathrm{op}}|\xi|^4 + \frac{\eta^2}{2}\|D^2E\|_{\mathrm{op}}\left(|\nabla E(\theta)| + \frac{1}{2}\|D^3E\|_{\mathrm{op}}|\xi|^2\right), \tag{96}$$

$$\begin{aligned}\|R^\Sigma\|_F \le{}& \frac{\eta^2}{2}\|D^3E\|_{\mathrm{op}}|\nabla E(\theta)|\,|\xi|^2 \\ &+ \eta^2\left(\frac{1}{8}\|D^3E\|^2_{\mathrm{op}} + \frac{1}{6}\|D^2E\|_{\mathrm{op}}\|D^4E\|_{\mathrm{op}}\right)|\xi|^4.\end{aligned} \tag{97}$$

For the first bound, the symmetric gradient average has Taylor remainder at most $\|D^5E\|_{\mathrm{op}}|\xi|^4/24$, since its linear and cubic terms cancel. The averaged Hessian is bounded by $\|D^2E\|_{\mathrm{op}}$, and the averaged gradient by $|\nabla E(\theta)| + \|D^3E\|_{\mathrm{op}}|\xi|^2/2$. For the second bound, decompose the endpoint gradients into their even and odd parts. Their outer-product average is the sum of the outer products of these two parts. The even part differs from $\nabla E(\theta)$ by at most $\|D^3E\|_{\mathrm{op}}|\xi|^2/2$. The odd part differs from $H(\theta)\xi$ by at most $\|D^4E\|_{\mathrm{op}}|\xi|^3/6$, and both have norm at most $\|D^2E\|_{\mathrm{op}}|\xi|$. Expanding these outer products gives the stated covariance bound.

By (95), the integrand in (93) is the norm of $\frac{\eta^2}{2}\nabla E(\theta_t)\nabla E(\theta_t)^T + R_t^\Sigma$. The gradient contribution is bounded by $\frac{\eta^2}{2}|\nabla E(\theta_t)|^2 \lesssim \eta^2\delta^2 = a^4$. Substituting (91) into (97) gives

$$\|R_t^\Sigma\|_F \lesssim \eta|\nabla E(\theta_t)|\,|\xi_t|^2 + |\xi_t|^4 \lesssim \eta\delta a^2 + a^4 = 2a^4,$$

with constants depending only on C. Integration over an interval of length T/η gives a bound $C_T a^4/\eta = C_T a^2\delta$; dividing by the covariance scale a^2 proves (93). Similarly, (96) yields

$$|R_t^\theta| \lesssim |\xi_t|^4 + \eta|\nabla E(\theta_t)| + |\xi_t|^2 \lesssim a^4 + \eta\delta + a^2 \lesssim a^2.$$

Since $\frac{d\theta_t}{dt} - \mathcal{V}_{\mathrm{R}}^\theta(\theta_t, \Sigma_t) = R_t^\theta$, integration over $[0, T/\eta]$ gives $C_T a^2/\eta = C_T\delta$ and proves (92). □

Part (i) of Proposition 2.4 controls the accumulated defect along the raw trajectory. Controlling the distance between independently evolved raw and coarse solutions would require a stability estimate on $[0, T/\eta]$. In particular, the absolute mean bound is not smaller than the available $O(\delta)$ bound on the mean displacement. The effect of retaining only $\xi\xi^T$, rather than Σ, is quantified in the comparison with MF below.

A.9.2 Comparison with MF

Proof of Proposition 2.4, part (ii). Write $D := \eta H(\theta) - 2\mathrm{Id}$. The covariance fields write

$$\begin{aligned}\mathcal{V}_{\mathrm{MF}}^\Sigma &= -\eta(\Sigma H + H\Sigma) + \eta^2 H\Sigma H,\\ \mathcal{V}_{\mathrm{R}}^\Sigma &= \frac{1}{2}\eta^2 H\xi\xi^T H - 2\Sigma.\end{aligned} \tag{98}$$

Subtracting the two identities, we obtain

$$\begin{aligned}\mathcal{V}_{\mathrm{MF}}^\Sigma - \mathcal{V}_{\mathrm{R}}^\Sigma &= -\eta(\Sigma H + H\Sigma) + 2\Sigma + \frac{1}{2}\eta^2 H\Sigma H\\ &\quad + \frac{\eta^2}{2}H(\Sigma - \xi\xi^T)H\\ &= \frac{1}{2}D\Sigma D + \frac{\eta^2}{2}H(\Sigma - \xi\xi^T)H.\end{aligned} \tag{99}$$

For $\Sigma \geq 0$, the factorization

$$D\Sigma D = (D\Sigma^{1/2})(D\Sigma^{1/2})^T$$

and the inequality $\|AA^T\| \leq \|A\|^2$ also give

$$\|\mathcal{V}_{\mathrm{MF}}^\Sigma - \mathcal{V}_{\mathrm{R}}^\Sigma\| \leq \frac{1}{2}\|D\Sigma^{1/2}\|^2 + \frac{\eta^2}{2}\|H(\Sigma - \xi\xi^T)H\|. \tag{100}$$

Let $P = P_\Sigma$ be the orthogonal projection onto $\operatorname{Ran}\Sigma$, and assume $\|DP\|_{\mathrm{op}} \leq \varepsilon$. Since $\Sigma = P\Sigma P$ and $\xi \in \operatorname{Ran}\Sigma$,

$$D\Sigma = (DP)\Sigma, \qquad \Sigma D = \Sigma(PD), \qquad D\Sigma D = (DP)\Sigma(PD).$$

The matrix $\Sigma - \xi\xi^T$ is also supported in $\operatorname{Ran}\Sigma$, so

$$H(\Sigma - \xi\xi^T)H = (HP)(\Sigma - \xi\xi^T)(PH), \qquad \|\eta HP\|_{\mathrm{op}} = \|(D + 2\mathrm{Id})P\|_{\mathrm{op}} \leq 2 + \varepsilon.$$

Symmetry of D gives $\|PD\|_{\mathrm{op}} = \|DP\|_{\mathrm{op}}$. Applying $\|A\Sigma C\| \leq \|A\|_{\mathrm{op}}\|\Sigma\|\|C\|_{\mathrm{op}}$ to these products therefore yields

$$\|\mathcal{V}_{\mathrm{MF}}^\Sigma - \mathcal{V}_{\mathrm{R}}^\Sigma\| \leq \frac{1}{2}\varepsilon^2\|\Sigma\| + \frac{(2+\varepsilon)^2}{2}\|\Sigma - \xi\xi^T\|.$$

For the individual fields, expanding $\eta H = 2\mathrm{Id} + D$ gives

$$\mathcal{V}_{\mathrm{MF}}^{\Sigma} = D\Sigma + \Sigma D + D\Sigma D,$$
$$\mathcal{V}_{\mathrm{R}}^{\Sigma} = D\xi\xi^T + \xi\xi^T D + \tfrac{1}{2} D\xi\xi^T D - 2(\Sigma - \xi\xi^T).$$

Since $\|\xi\xi^T\| \le \|\Sigma\|$, we obtain

$$\|\mathcal{V}_{\mathrm{MF}}^{\Sigma}\| \le (2\varepsilon + \varepsilon^2)\|\Sigma\|,$$
$$\|\mathcal{V}_{\mathrm{R}}^{\Sigma}\| \le (2\varepsilon + \tfrac{1}{2}\varepsilon^2)\|\Sigma\| + 2\|\Sigma - \xi\xi^T\|. \tag{101}$$

□

B Analysis of the kinetic equation

B.1 Velocity moments and variational structure

We work on $\mathbb{T}^d \times \mathbb{R}^{d\times r}$ and use the Fourier convention $\widehat{\nabla h}(k) = ik\widehat{h}(k)$. Throughout this subsection, $\mathsf{g} \in C^2(\mathbb{T}^d)$ is even and $\widehat{\mathsf{g}}(k) \ge 0$. For a probability measure f with positional marginal ρ, set

$$M_p(f) = \int |v|^p \,\mathrm{d}f, \qquad \mathcal{E}(\rho) = \frac{1}{2}\int \mathsf{g} * (\rho - \mu)\,\mathrm{d}(\rho - \mu).$$

The free energy is unbounded when velocities are unrestricted, but a bound on their second moment restores a uniform lower bound.

Proposition B.1 (Free energy with and without a velocity bound)**.** *Suppose that* g *is nonconstant. Then*

$$\inf_{f\in\mathcal{P}_2(\mathbb{T}^d\times\mathbb{R}^{d\times r})} \mathcal{F}(f) = -\infty. \tag{102}$$

Moreover, every $f \in \mathcal{P}_2$ *satisfies*

$$\mathcal{F}(f) \ge \mathcal{E}(\rho) - \|\nabla^2\mathsf{g}\|_{L^\infty} M_2(f). \tag{103}$$

Consequently, for every $\varepsilon > 0$*,*

$$-\|\nabla^2\mathsf{g}\|_{L^\infty}\varepsilon^2 \le \inf_{\substack{f\in\mathcal{P}_2 \\ M_2(f)\le\varepsilon^2}} \mathcal{F}(f) \le 0. \tag{104}$$

Proof. We first prove unboundedness. If $\mathsf{g} * \mu$ is nonconstant, there are $\theta \in \mathbb{T}^d$ and $w \in \mathbb{R}^{d\times r}$ such that $\nabla^2(\mathsf{g}*\mu)(\theta) : ww^T > 0$. Indeed, otherwise $\nabla^2(\mathsf{g}*\mu) \le 0$ everywhere, so the periodic function $\mathsf{g} * \mu$ would be concave and hence constant. Direct substitution of $f_R = \delta_{(\theta,Rw)}$ in (12) gives

$$\mathcal{F}(f_R) = \mathcal{E}(\delta_\theta) - \frac{R^2}{2}\nabla^2(\mathsf{g}*\mu)(\theta) : ww^T \longrightarrow -\infty.$$

It remains to consider the case where $\mathsf{g} * \mu$ is constant. Fix θ and take

$$f_R = \frac{1}{2}\delta_{(\theta,Rw)} + \frac{1}{2}\delta_{(\theta,-Rw)}.$$

Direct substitution now gives

$$\mathcal{F}(f_R) = \mathcal{E}(\delta_\theta) + \frac{R^2}{2}\nabla^2\mathsf{g}(0) : ww^T.$$

Because $\nabla^2\mathsf{g}(0) = -\sum_{k\neq 0}\widehat{\mathsf{g}}(k)k\otimes k$ and g is nonconstant, w can be chosen so that the last term is negative. This proves (102) in both cases.

We next prove the lower bound. Expanding $(v - v')(v - v')^T$ in (12) yields

$$\mathcal{F}(f) - \mathcal{E}(\rho) = \frac{1}{2}\int [\nabla^2\mathsf{g} * (\rho - \mu)](\theta) : vv^T \,\mathrm{d}f - \frac{1}{2}\iint v : \nabla^2\mathsf{g}(\theta - \theta')v' \,\mathrm{d}f\,\mathrm{d}f'.$$

Since $\widehat{\nabla^2 \mathsf{g}}(k) = -k \otimes k\, \widehat{\mathsf{g}}(k)$, Fourier expansion applied columnwise gives

$$-\iint v : \nabla^2 \mathsf{g}(\theta - \theta') v' \,\mathrm{d}f \,\mathrm{d}f' = \sum_{k \neq 0} \widehat{\mathsf{g}}(k) \left| k^T \int e^{-ik\cdot\theta} v \,\mathrm{d}f(\theta, v) \right|^2 \geq 0.$$

Moreover,

$$\|\nabla^2 \mathsf{g} * (\rho - \mu)\|_{L^\infty} \leq \|\nabla^2 \mathsf{g}\|_{L^\infty} \|\rho - \mu\|_{\mathrm{TV}} \leq 2\|\nabla^2 \mathsf{g}\|_{L^\infty}.$$

Hence

$$\frac{1}{2} \int [\nabla^2 \mathsf{g} * (\rho - \mu)](\theta) : vv^T \,\mathrm{d}f \geq -\|\nabla^2 \mathsf{g}\|_{L^\infty} M_2(f),$$

which proves (103). If $M_2(f) \leq \varepsilon^2$, this gives the lower bound in (104); the upper bound follows from taking $f = \mu \otimes \delta_0$, whose free energy vanishes. □

A natural question is whether the free energy attains its minimum under velocity constraints, which we now address.

Proposition B.2 (Existence under uniform velocity integrability)**.** *Fix $p > 2$ and $L > 0$, and let*

$$\mathcal{B}_L^p := \left\{ f \in \mathcal{P}_p(\mathbb{T}^d \times \mathbb{R}^{d\times r}) : M_p(f) \leq L \right\}.$$

Then $\mathcal{F}$ attains its minimum on $\mathcal{B}_L^p$.

Proof. For every $f \in \mathcal{B}_L^p$, Hölder's inequality gives $M_2(f) \leq M_p(f)^{2/p} \leq L^{2/p}$, so (103) bounds $\mathcal{F}$ from below on $\mathcal{B}_L^p$. Let (f_j) be a minimizing sequence. Since the θ-space is compact and $M_p(f_j) \leq L$, Markov's inequality makes (f_j) tight. Prokhorov's theorem therefore gives, up to a subsequence, $f_j \rightharpoonup f$ narrowly as $j \to \infty$. The moment constraint passes to the limit by lower semicontinuity.

It remains to pass to the limit in the energy, which is quadratic in velocity. For $z = (\theta, v)$ and $z' = (\theta', v')$, set

$$K(z, z') = \mathsf{g}(\theta - \theta') + \frac{1}{2} \nabla^2 \mathsf{g}(\theta - \theta') : (v - v')(v - v')^T.$$

With the target understood as $\mu \otimes \delta_0$, (12) reads

$$\mathcal{F}(f_j) = \frac{1}{2} \iint K(z, z') \,\mathrm{d}(f_j - \mu)(z) \,\mathrm{d}(f_j - \mu)(z').$$

Let $\chi \in C_c(\mathbb{R}^{d\times r}; [0, 1])$ satisfy $\chi = 1$ on B_1 and $\chi = 0$ outside B_2, and set

$$\chi_R(v) = \chi(v/R), \qquad K_R(z, z') = \chi_R(v)\chi_R(v')K(z, z').$$

The kernel K_R is bounded and continuous. Moreover, $f_j \rightharpoonup f$ implies convergence of $f_j \otimes f_j$, $f_j \otimes \mu$, and $\mu \otimes f_j$ to the corresponding products with f. Consequently,

$$\frac{1}{2} \iint K_R \,\mathrm{d}(f_j - \mu) \,\mathrm{d}(f_j - \mu) \longrightarrow \frac{1}{2} \iint K_R \,\mathrm{d}(f - \mu) \,\mathrm{d}(f - \mu) \qquad (j \to \infty). \tag{105}$$

It remains to remove the cutoff uniformly in j. For $R \geq 1$,

$$\int_{\{|v|>R\}} \mathrm{d}f_j \leq LR^{-p}, \qquad \int_{\{|v|>R\}} |v|^2 \,\mathrm{d}f_j \leq LR^{2-p}, \qquad M_2(f_j) \leq L^{2/p}.$$

The middle estimate is precisely the uniform integrability of $|v|^2$. Since

$$|K(z, z')| \leq C(1 + |v|^2 + |v'|^2), \qquad |f_j - \mu| \leq f_j + \mu,$$

and $1 - \chi_R(v)\chi_R(v')$ vanishes unless $|v| > R$ or $|v'| > R$, it is enough to estimate either of these two symmetric regions. Since the lifted target has velocity zero,

$$\begin{aligned} &\iint_{\{|v|>R\}} (1 + |v|^2 + |v'|^2) \,\mathrm{d}(f_j + \mu)(z) \,\mathrm{d}(f_j + \mu)(z') \\ &\quad = 2 \int_{\{|v|>R\}} (1 + |v|^2) \,\mathrm{d}f_j + M_2(f_j) \int_{\{|v|>R\}} \mathrm{d}f_j \\ &\quad \leq 2L\big(R^{-p} + R^{2-p}\big) + L^{1+2/p} R^{-p}. \end{aligned}$$

The same estimate holds with v and v' exchanged. We conclude that

$$\left|\mathcal{F}(f_j) - \frac{1}{2}\iint K_R \,\mathrm{d}(f_j - \mu)\,\mathrm{d}(f_j - \mu)\right| \tag{106}$$

$$\leq CL\big(1 + L^{2/p}\big)R^{2-p}. \tag{107}$$

Here C depends only on $\|\mathsf{g}\|_{L^\infty} + \|\nabla^2\mathsf{g}\|_{L^\infty}$ and not on j or R. Since $p > 2$, the right-hand side tends to zero as $R \to \infty$. The limit measure also satisfies $M_p(f) \leq L$, so the same estimate holds with f in place of f_j. We may therefore first let $j \to \infty$ in (105) and then let $R \to \infty$ in (107). This proves $\mathcal{F}(f_j) \to \mathcal{F}(f)$, and f is a minimizer. □

Finally, we prove the convergence of minimizers to the zero-velocity limit. We actually prove a slightly stronger statement, where we need not consider only minimizers, but any sequence of measures with vanishing second moment and nonpositive free energy.

Proposition B.3 (Small-variance limit)**.** *Assume that $\widehat{\mathsf{g}}(k) > 0$ for every $k \neq 0$. Let $\varepsilon_n \to 0$ and let $f_n \in \mathcal{P}_2$ satisfy*

$$M_2(f_n) \leq \varepsilon_n^2, \qquad \mathcal{F}(f_n) \leq 0.$$

Then

$$W_2(f_n, \mu \otimes \delta_0) \longrightarrow 0. \tag{108}$$

In particular, f_n converges narrowly to $\mu \otimes \delta_0$.

Moreover, fix any $p > 2$ and arbitrary $L_n > 0$. The conclusion applies when f_n minimizes $\mathcal{F}$ on

$$\{f \in \mathcal{B}^p_{L_n} : M_2(f) \leq \varepsilon_n^2\}.$$

Such minimizers exist by Proposition B.2, and $\mu \otimes \delta_0$ is an admissible competitor.

Proof. Write ρ_n for the positional marginal of f_n. By (103),

$$0 \geq \mathcal{F}(f_n) \geq \mathcal{E}(\rho_n) - \|\nabla^2\mathsf{g}\|_{L^\infty}\varepsilon_n^2.$$

Therefore

$$\mathcal{E}(\rho_n) \leq \|\nabla^2\mathsf{g}\|_{L^\infty}\varepsilon_n^2 \longrightarrow 0.$$

Since every nonzero Fourier coefficient of g is positive, this implies $\widehat{\rho}_n(k) \to \widehat{\mu}(k)$ for every k. Trigonometric polynomials are dense in $C(\mathbb{T}^d)$, hence $\rho_n \rightharpoonup \mu$. Compactness of the torus upgrades this to $W_2(\rho_n, \mu) \to 0$.

Finally, coupling each (θ, v) with $(\theta, 0)$ gives

$$W_2(f_n, \rho_n \otimes \delta_0)^2 \leq M_2(f_n) \leq \varepsilon_n^2.$$

The triangle inequality proves (108). □

B.2 Proof of the well-posedness theorem

Notations. We identify the target with $\mu \otimes \delta_0$ and use the characteristic flow $Z = (\Theta, V)$ defined in (110); the solution is represented by (109). We write $\mathcal{V} = (\mathcal{V}_\theta, \mathcal{V}_v)$ for the characteristic velocity and $M_k(t) := M_k(f_t)$ for its velocity moments. The products vv^T and VV^T are $d \times d$ matrices, obtained by summing the outer products of the columns. Derivatives in $v \in \mathbb{R}^{d\times r}$ are Fréchet derivatives, with operator norms induced by the Frobenius norm.

Formally, the equation (11) is equivalent to

$$f(t) = Z(t, \cdot;\, f)_\# f_0, \tag{109}$$

where $Z(t, \theta, v) = (\Theta(t, \theta, v), V(t, \theta, v))$ is the flow map

$$\begin{cases} \dfrac{d\Theta}{ds} = \mathcal{V}_\theta(s, \Theta, V) := -\nabla\mathsf{g} * (\rho - \mu)(s, \Theta) - \dfrac{1}{2}\nabla_\theta(\nabla^2\mathsf{g} : VV^T) * (f - \mu)(s, \Theta, V), \\ \dfrac{dV}{ds} = \mathcal{V}_v(s, \Theta, V) := -(\nabla^2\mathsf{g}\, V) * (f - \mu)(s, \Theta, V). \end{cases} \tag{110}$$

Theorem B.4 (Well-posedness and regularity of the MMD kinetic equation)**.** *Let* $\mathrm{g} \in C_b^{3,1}$ *be even,* $f_0 \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^{d\times r})$ *compactly supported in the* v *variable and* $\mu \in \mathcal{P}(\mathbb{R}^d)$*.*

For all $T > 0$*, there exists a unique, compactly supported in* v*, solution* $f \in C([0,T], \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^{d\times r}))$ *to* (109)*, which furthermore satisfies*

$$\frac{d}{dt}\mathcal{F}(f) + \int_{\mathbb{R}^d\times\mathbb{R}^{d\times r}} \left(\left| \nabla_\theta \frac{\delta \mathcal{F}}{\delta f} \right|^2 + \left| \nabla_v \frac{\delta \mathcal{F}}{\delta f} \right|^2 \right) df \le 0. \tag{111}$$

Finally, if g *is smooth enough, the regularity of* f_0 *is propagated in time.*

B.2.1 Study of the flow map

In this section, we study the flow (110) and establish its regularity properties. Specifically, the goal is to prove the following two lemmas.

Lemma B.5 (Regularity of the velocity fields)**.** *Let* $n \ge 0$*,* $f \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^{d\times r})$*,* $\mu \in \mathcal{P}_2(\mathbb{R}^d)$*, and* $\mathrm{g} \in C_b^{n+2,1}(\mathbb{R}^d)$*. For every* $0 \le a \le n$ *and* $b \ge 0$*, there is a constant* $C_{a,b}$ *such that*

$$\sup_{\substack{|v|\le R \\ \theta\in\mathbb{R}^d}} \big(\|\nabla_\theta^a \nabla_v^b \mathcal{V}_\theta\| + \|\nabla_\theta^{a+1}\nabla_v^b \mathcal{V}_v\| \big) \le C_{a,b}\big(1 + R^2 + M_2(f)\big).$$

The constant depends only on the indicated derivatives of g*.*

Lemma B.6 (Stability of the velocity fields)**.** *Let* $R > 0$ *and* $|v| \le R$*. Assume* $\mathrm{g} \in C_b^{3,1}(\mathbb{R}^d)$*. For any* $f_1, f_2 \in \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^{d\times r})$*, we have*

$$\begin{aligned} &|\mathcal{V}_\theta(t,\theta,v;f_1) - \mathcal{V}_\theta(t,\theta,v;f_2)| \\ &\quad \le C\Big[1 + R + R^2 + (1+R)\big(M_2(f_1)^{\frac12} + M_2(f_2)^{\frac12}\big) + M_4(f_1)^{\frac14} M_4(f_2)^{\frac14}\Big] W_2(f_1,f_2), \end{aligned}$$

and

$$|\mathcal{V}_v(t,\theta,v;f_1) - \mathcal{V}_v(t,\theta,v;f_2)| \le C\big(1 + R + M_2(f_1)^{\frac12} + M_2(f_2)^{\frac12}\big) W_2(f_1,f_2),$$

where the constant $C > 0$ *depends on* $\|\mathrm{g}\|_{C^{3,1}}$*.*

Remark B.7. Once these two lemmas are established, the Cauchy–Lipschitz theorem gives existence and uniqueness of the flow in (110). The flow representation (109) and its stability with respect to f then give the corresponding well-posedness theory for (11). It remains only to control the velocity moments needed in Lemmas B.5 and B.6; this is done in Proposition B.8.

Proof of Lemma B.5. We fix some $R > 0$ and $|v| \le R$. We have from Hölder's inequality:

$$\|\nabla^k \nabla \mathrm{g} * (\rho - \mu)\|_{L^\infty} \le 2\|\nabla^{k+1}\mathrm{g}\|_{L^\infty}.$$

Moreover, expanding the convolution gives

$$\begin{aligned} \|\nabla_\theta^k(\nabla^2\mathrm{g} : vv^T) * (f-\mu)\| &\le \|\nabla^{k+2}\mathrm{g}\|_{L^\infty}\big(2R^2 + 2RM_1(f) + M_2(f)\big) \\ &\le \|\nabla^{k+2}\mathrm{g}\|_{L^\infty}\big(3R^2 + 2M_2(f)\big), \end{aligned}$$

where we have used $ab \le a^2/2 + b^2/2$ to obtain the last line. Now, we differentiate with respect to v; for any matrix increment $h \in \mathbb{R}^{d\times r}$,

$$\begin{aligned} \nabla_v\big[(\nabla^2\mathrm{g} : vv^T) * (f-\mu)\big] &= 2(\nabla^2\mathrm{g}\, v) * (f-\mu), \\ \nabla_v^2\big[(\nabla^2\mathrm{g} : vv^T) * (f-\mu)\big]h &= 2\big[\nabla^2\mathrm{g} * (\rho-\mu)\big]h. \end{aligned}$$

The higher order derivatives in v vanish identically. Therefore,

$$\begin{aligned} &\|\nabla_v\nabla_\theta^k(\nabla^2\mathrm{g} : vv^T) * (f-\mu)\| \\ &\le 4R\|\nabla^{k+2}\mathrm{g}\|_{L^\infty} + 2\|\nabla^{k+2}\mathrm{g}\|_{L^\infty} \int_{\mathbb{R}^d\times\mathbb{R}^{d\times r}} |v| f(\theta,v)\, d\theta dv, \end{aligned}$$

and

$$\|\nabla_v^2\nabla_\theta^k(\nabla^2\mathsf{g} : vv^T) * (f-\mu)\|_{L^\infty} \le 4\|\nabla^{k+2}\mathsf{g}\|_{L^\infty}.$$

Let us now focus on the regularity of $\mathcal{V}_v$. We have

$$\begin{aligned}\|\nabla_\theta^k(\nabla^2\mathsf{g}\, v) * (f-\mu)\| &\le 2R\|\nabla^{k+2}\mathsf{g}\|_{L^\infty} + \|\nabla^{k+2}\mathsf{g}\|_{L^\infty}\int_{\mathbb{R}^d\times\mathbb{R}^{d\times r}} |v| f(\theta,v)\, d\theta dv \\ &\le \frac{5}{2}R\|\nabla^{k+2}\mathsf{g}\|_{L^\infty} + \frac{1}{2}\|\nabla^{k+2}\mathsf{g}\|_{L^\infty}\int_{\mathbb{R}^d\times\mathbb{R}^{d\times r}} |v|^2 f(\theta,v)\, d\theta dv.\end{aligned}$$

Moreover, differentiating in v in the matrix direction h gives

$$\nabla_v\big[(\nabla^2\mathsf{g}\, v) * (f-\mu)\big]h = \big[\nabla^2\mathsf{g} * (\rho-\mu)\big]h.$$

Therefore,

$$\|\nabla_v\nabla_\theta^k(\nabla^2\mathsf{g}\, v) * (f-\mu)\|_{L^\infty} \le 2\|\nabla^{k+2}\mathsf{g}\|_{L^\infty}.$$

Again, all the higher order terms in v identically vanish. □

Proof of Lemma B.6. We have

$$\mathcal{V}_\theta(t,\theta,v;f_1) - \mathcal{V}_\theta(t,\theta,v;f_2) = -\nabla\mathsf{g} * (\rho_1-\rho_2) - \frac{1}{2}\nabla_\theta(\nabla^2\mathsf{g} : vv^T) * (f_1-f_2).$$

and

$$\mathcal{V}_v(t,\theta,v;f_1) - \mathcal{V}_v(t,\theta,v;f_2) = -(\nabla^2\mathsf{g}\, v) * (f_1-f_2).$$

For the first term of $\mathcal{V}_\theta$, we have from Kantorovich-Rubinstein duality:

$$|\nabla\mathsf{g} * (\rho_1-\rho_2)| \le \|\nabla^2\mathsf{g}\|_{L^\infty} W_1(\rho_1,\rho_2) \le \|\nabla^2\mathsf{g}\|_{L^\infty} W_2(f_1,f_2).$$

For the oscillatory part, let π be a coupling between f_1 and f_2 for the W_2 distance. We have

$$\begin{aligned}&\nabla_\theta(\nabla^2\mathsf{g} : vv^T) * (f_1-f_2) = \int \nabla^3\mathsf{g}(\theta-\theta') : (v-v')(v-v')^T\, d(f_1-f_2)(\theta',v') \\ &= \int \big[\nabla^3\mathsf{g}(\theta-\theta_1') : (v-v_1')(v-v_1')^T - \nabla^3\mathsf{g}(\theta-\theta_2') : (v-v_2')(v-v_2')^T\big]\, d\pi(Z_1',Z_2') \\ &= \int \big[\nabla^3\mathsf{g}(\theta-\theta_1') : (v_2'-v_1')(v-v_1')^T + \nabla^3\mathsf{g}(\theta-\theta_1') : (v-v_2')(v-v_1')^T \\ &- \nabla^3\mathsf{g}(\theta-\theta_2') : (v-v_2')(v-v_1')^T - \nabla^3\mathsf{g}(\theta-\theta_2') : (v-v_2')(v_1'-v_2')^T\big]\, d\pi(Z_1',Z_2') \\ &= \int \nabla^3\mathsf{g}(\theta-\theta_1') : (v_2'-v_1')(v-v_1')^T\, d\pi(Z_1',Z_2') \\ &+ \int \big[\nabla^3\mathsf{g}(\theta-\theta_1') - \nabla^3\mathsf{g}(\theta-\theta_2')\big] : (v-v_2')(v-v_1')^T\, d\pi(Z_1',Z_2') \\ &- \int \nabla^3\mathsf{g}(\theta-\theta_2') : (v-v_2')(v_1'-v_2')^T\, d\pi(Z_1',Z_2').\end{aligned}$$

The first integral is bounded by

$$\begin{aligned}&\|\nabla^3\mathsf{g}\|_{L^\infty}\int \big(|v|+|v_1'|\big)|v_2'-v_1'|\, d\pi(Z_1',Z_2') \\ &\le \|\nabla^3\mathsf{g}\|_{L^\infty}\bigg(\int (|v|+|v_1'|)^2\, d\pi(Z_1',Z_2')\bigg)^{\frac{1}{2}}\bigg(\int |v_2'-v_1'|^2\, d\pi(Z_1',Z_2')\bigg)^{\frac{1}{2}}.\end{aligned}$$

The second one is bounded by

$$\begin{aligned}&\|\nabla^4\mathsf{g}\|_{L^\infty}\int |\theta_1'-\theta_2'||v-v_2'||v-v_1'|\, d\pi(Z_1',Z_2') \\ &\le \|\nabla^4\mathsf{g}\|_{L^\infty}\bigg(\int |v-v_2'|^2|v-v_1'|^2\, d\pi(Z_1',Z_2')\bigg)^{\frac{1}{2}}\bigg(\int |\theta_1'-\theta_2'|^2\, d\pi(Z_1',Z_2')\bigg)^{\frac{1}{2}}.\end{aligned}$$

We then use $|v| \le R$ to obtain

$$\begin{aligned}
&\|\nabla^4 \mathsf{g}\|_{L^\infty} \left(\int |v - v_2'|^2 |v - v_1'|^2 \, d\pi \right)^{\frac{1}{2}} \left(\int |\theta_1' - \theta_2'|^2 \, d\pi \right)^{\frac{1}{2}} \\
&\quad \le 4 \|\nabla^4 \mathsf{g}\|_{L^\infty} \left(R^4 + R^2 \big(M_2(f_1) + M_2(f_2) \big) + M_4(f_1)^{\frac{1}{2}} M_4(f_2)^{\frac{1}{2}} \right)^{\frac{1}{2}} \\
&\qquad \times \left(\int |\theta_1' - \theta_2'|^2 \, d\pi \right)^{\frac{1}{2}}.
\end{aligned}$$

Finally, the last term is bounded similarly to the first one. Taking the infimum over the couplings π, we obtain the stability result for $\mathcal{V}_\theta$.

For $\mathcal{V}_v$, we have

$$\begin{aligned}
&\mathcal{V}_v(t, \theta, v; f_1) - \mathcal{V}_v(t, \theta, v; f_2) \\
&= \int \left[\nabla^2 \mathsf{g}(\theta - \theta_1)(v - v_1) - \nabla^2 \mathsf{g}(\theta - \theta_2)(v - v_2) \right] d\pi(Z_1, Z_2) \\
&= \int \left[(\nabla^2 \mathsf{g}(\theta - \theta_1) - \nabla^2 \mathsf{g}(\theta - \theta_2))(v - v_1) + \nabla^2 \mathsf{g}(\theta - \theta_2)(v_2 - v_1) \right] d\pi(Z_1, Z_2).
\end{aligned}$$

We therefore have

$$\begin{aligned}
&|\mathcal{V}_v(t, \theta, v; f_1) - \mathcal{V}_v(t, \theta, v; f_2)| \\
&\le \|\nabla^3 \mathsf{g}\|_{L^\infty} \int |\theta_1 - \theta_2| |v - v_1| \, d\pi(Z_1, Z_2) + \|\nabla^2 \mathsf{g}\|_{L^\infty} \int |v_1 - v_2| \, d\pi(Z_1, Z_2) \\
&\le \|\nabla^3 \mathsf{g}\|_{L^\infty} \left(\int |\theta_1 - \theta_2|^2 \, d\pi(Z_1, Z_2) \right)^{\frac{1}{2}} (2R^2 + 2M_2(f_1))^{\frac{1}{2}} \\
&+ \|\nabla^2 \mathsf{g}\|_{L^\infty} \left(\int |v_1 - v_2|^2 \, d\pi(Z_1, Z_2) \right)^{\frac{1}{2}}.
\end{aligned}$$

This allows us to conclude by considering the infimum over couplings. □

B.2.2 Well-posedness of the equation

We now establish the estimates needed to globalize the characteristic flow. The only point is to control the velocity moments appearing in Lemmas B.5 and B.6.

Proposition B.8 (A priori estimates)**.** *Let $\mathsf{g} \in C_b^{2,1}(\mathbb{R}^d)$ be even. Consider $f_0 \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^{d \times r})$ and $\mu \in \mathcal{P}(\mathbb{R}^d)$ such that f_0 is compactly supported in the v variable. Let $f \in C([0,T], \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^{d \times r}))$ be a locally Lipschitz solution to* (11) *that is compactly supported in the v variable for all $t \in [0,T]$. Then the following properties hold:*

i. Conservation of mass: $\int \mathrm{d} f_t = 1$ *for every* $t \in [0,T]$.

ii. Propagation of moments: *for every $k \ge 1$ and $t \in [0,T]$,*

$$M_k(t) \le M_k(0) \exp\big(3k \|\nabla^2 \mathsf{g}\|_{L^\infty} t \big).$$

iii. Finite propagation speed in velocity: *if* $\operatorname{supp}_v f_0 \subset B_R(0)$*, then*

$$\operatorname{supp}_v f_t \subset B_{(R + Ct) e^{Ct}}(0), \qquad 0 \le t \le T,$$

where C depends only on $\|\nabla^2 \mathsf{g}\|_{L^\infty}$ and $M_1(0)$.

iv. Free-energy dissipation: *the free energy defined in* (12) *is nonincreasing and satisfies* (111)*.*

Proof. The conservation of mass follows from the divergence form of (11). To establish the propagation of velocity moments, let $p \ge 1$ and differentiate $M_p(t)$ with respect to time:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} M_p(t) &= -p \int |v|^{p-2} v : \big[(\nabla^2 \mathsf{g} \, v) * (f_t - \mu) \big](\theta, v) \, \mathrm{d} f_t(\theta, v) \\
&\le p \|\nabla^2 \mathsf{g}\|_{L^\infty} \left(M_{p-1}(t) M_1(t) + 2 M_p(t) \right).
\end{aligned}$$

Using Jensen's inequality to obtain $M_{p-1}M_1 \le M_p$, we have $\frac{\mathrm{d}}{\mathrm{d}t}M_p(t) \le 3p\|\nabla^2 \mathsf{g}\|_{L^\infty} M_p(t)$, and the result follows by Grönwall's lemma.

The finite speed of propagation in the velocity variable is obtained by considering a characteristic curve $(\Theta(s), V(s))$, whose velocity in v satisfies:

$$\begin{aligned}\frac{\mathrm{d}V}{\mathrm{d}s}(s) &= -(\nabla^2 \mathsf{g} V) * (f-\mu)(s,\Theta,V) \\ &= -\nabla^2 \mathsf{g} * (\rho-\mu)(s,\Theta)V(s) + \int_{\mathbb{R}^d\times\mathbb{R}^{d\times r}} \nabla^2 \mathsf{g}(\Theta-\Theta')V' f(s,\Theta',V')\, d\Theta' dV'.\end{aligned}$$

Since M_1 is bounded on $[0,T]$, we have $|\frac{\mathrm{d}V}{\mathrm{d}s}(s)| \le C|V(s)| + C$, where C depends on $\|\nabla^2 \mathsf{g}\|_\infty$ and the first moment. Grönwall's lemma then implies $|V(t)| \le (|V(0)| + Ct)e^{Ct}$, so that the support remains compact for all $t \in [0,T]$.

Since g is even, the interaction kernel in (12) is symmetric under exchanging the two phase-space variables. Hence $\mathcal{V} = -\nabla_{\theta,v}(\delta\mathcal{F}/\delta f)$. Finally, differentiating the free energy along the equation gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(f_t) = &-\int \left|\nabla \mathsf{g} * (\rho_t - \mu) + \frac{1}{2}\nabla_\theta(\nabla^2 \mathsf{g} : vv^T) * (f_t - \mu)\right|^2 \mathrm{d}f_t \\ &- \int |(\nabla^2 \mathsf{g}\, v) * (f_t - \mu)|^2 \,\mathrm{d}f_t.\end{aligned}$$

Since both terms are non-positive, $\mathcal{F}$ is non-increasing. □

Proposition B.9 (Uniqueness)**.** *Let $f_0 \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^{d\times r})$ be a probability measure with compact support in the v variable. There exists at most one solution $f \in C([0,T], \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^{d\times r}))$ to* (109) *which maintains compact support in v for all $t \in [0,T]$.*

Proof. Let f_1, f_2 be two such solutions starting from f_0. Let Z_1, Z_2 be their respective characteristic flows defined by (110). Since $f_i(t) = Z_i(t,\cdot)_\# f_0$, the measure $\pi_t := (Z_1(t,\cdot), Z_2(t,\cdot))_\# f_0$ is a valid coupling between $f_1(t)$ and $f_2(t)$. By definition of the Wasserstein distance as an infimum over couplings, we have

$$W_2^2(f_1(t), f_2(t)) \le \int_{\mathbb{R}^d\times\mathbb{R}^{d\times r}} |Z_1(t,z) - Z_2(t,z)|^2 df_0(z).$$

where $z = (\theta, v)$. The difference in trajectories satisfies

$$\begin{aligned}\frac{d}{dt}|Z_1 - Z_2| &\le |\mathcal{V}(t, Z_1; f_1) - \mathcal{V}(t, Z_2; f_2)| \\ &\le |\mathcal{V}(t, Z_1; f_1) - \mathcal{V}(t, Z_1; f_2)| + |\mathcal{V}(t, Z_1; f_2) - \mathcal{V}(t, Z_2; f_2)|.\end{aligned}$$

By the compact support assumption, there exists $R > 0$ such that $\mathrm{supp}_v\, f_1(t) \cup \mathrm{supp}_v\, f_2(t) \subset B_R$ for all $t \in [0,T]$. Using Lemma B.6, the first term is bounded by $C_R W_2(f_1, f_2)$, while the second term is bounded by $C_R|Z_1 - Z_2|$ thanks to Lemma B.5 Combining these, we have

$$\begin{aligned}\frac{d}{dt}\int |Z_1 - Z_2|^2\, df_0 &= 2\int (Z_1 - Z_2)\cdot \frac{d}{dt}(Z_1 - Z_2)\, df_0 \\ &\le 2C_R \int |Z_1 - Z_2|(|Z_1 - Z_2| + W_2(f_1, f_2))\, df_0 \\ &\le 4C_R \int |Z_1 - Z_2|^2\, df_0 + C_R W_2^2(f_1, f_2).\end{aligned}$$

where we used $ab \le a^2/2 + b^2/2$. We conclude by Grönwall's lemma that $W_2(f_1(t), f_2(t)) = 0$ for all $t \in [0,T]$, which implies uniqueness. □

B.3 Conditional convergence to the target

For classical MMD gradient flow, Arbel et al. (2019, Proposition 7) prove global convergence under a uniform bound on a weighted negative Sobolev norm. Kernel-specific convergence results have also been obtained for Sobolev discrepancies (Chizat et al., 2026), while Chodron de Courcel & Rosenzweig (2026) and Rosenzweig et al. (2026) study convergence and rigidity of Lagrangian critical points for Coulomb and energy kernels. Here we work with a smooth kernel and quantify how kinetic fluctuations perturb the MMD convergence argument.

For probability measures ρ, μ on $\mathbb{T}^d$, write

$$\mathrm{MMD}_{\mathsf{g}}^2(\rho,\mu) := \int \mathsf{g} * (\rho-\mu)\,\mathrm{d}(\rho-\mu),$$

and define the weighted negative Sobolev norm

$$\|\rho-\mu\|_{\dot{H}^{-1}(\rho)} := \sup_{\substack{\varphi\in C^1(\mathbb{T}^d)\\ \int|\nabla\varphi|^2\,\mathrm{d}\rho\le 1}} \left|\int \varphi\,\mathrm{d}(\rho-\mu)\right|. \tag{112}$$

Theorem B.10 (Conditional convergence of the kinetic flow)**.** *Let $\mathsf{g}\in C^{3,1}(\mathbb{T}^d)$ be even, with $\widehat{\mathsf{g}}(k)\ge 0$ for all k and $\widehat{\mathsf{g}}(k)>0$ for $k\neq 0$, and let $\mu\in\mathcal{P}(\mathbb{T}^d)$. Let f_t be the characteristic solution of (11) on $\mathbb{T}^d\times\mathbb{R}^{d\times r}$, from initial data $f_0\in\mathcal{P}_2(\mathbb{T}^d\times\mathbb{R}^{d\times r})$ compactly supported in v, and denote its positional marginal by ρ_t. Set*

$$M_2(t)=\int |v|^2\,df_t.$$

Suppose that, for some $C_0>0$,

$$\sup_{t\ge 0}\|\rho_t-\mu\|_{\dot{H}^{-1}(\rho_t)}\le C_0. \tag{113}$$

Then

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{MMD}_{\mathsf{g}}^2(\rho_t,\mu)\le -\frac{2}{C_0^2}\big[\mathrm{MMD}_{\mathsf{g}}^2(\rho_t,\mu)\big]^2+2C_{\mathsf{g}}M_2(t)\sqrt{\mathrm{MMD}_{\mathsf{g}}^2(\rho_t,\mu)}, \tag{114}$$

where $C_{\mathsf{g}}=\frac{3}{2}\|\nabla^3\mathsf{g}\|_\infty\sqrt{-\Delta\mathsf{g}(0)}$. In particular, if $M_2(t)\to 0$, then

$$W_2(f_t,\mu\otimes\delta_0)\longrightarrow 0. \tag{115}$$

No time-integrability assumption on M_2 is required.

Proof. We first estimate the correction to the classical MMD dissipation. Write

$$\phi_t=\mathsf{g}*(\rho_t-\mu),\qquad b_t(\theta,v)=\frac{1}{2}\int\nabla^3\mathsf{g}(\theta-\theta'):(v-v')(v-v')^T\,\mathrm{d}(f_t-\mu)(\theta',v').$$

In this formula, μ is lifted to $\mu\otimes\delta_0$ and $\mathcal{V}_\theta=-\nabla\phi_t-b_t$. Since $\iint|v-v'|^2\,\mathrm{d}f_t\,\mathrm{d}f_t'\le 2M_2(t)$,

$$\int|b_t|\,\mathrm{d}f_t\le\frac{1}{2}\|\nabla^3\mathsf{g}\|_\infty\left(\iint|v-v'|^2\,\mathrm{d}f_t\,\mathrm{d}f_t'+M_2(t)\right)\le\frac{3}{2}\|\nabla^3\mathsf{g}\|_\infty M_2(t).$$

Differentiating the squared MMD along the characteristics gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{MMD}_{\mathsf{g}}^2(\rho_t,\mu)=-2J_t-2\int\nabla\phi_t\cdot b_t\,\mathrm{d}f_t,\qquad J_t:=\int|\nabla\phi_t|^2\,\mathrm{d}\rho_t. \tag{116}$$

By homogeneity in the test function, (113) gives

$$\mathrm{MMD}_{\mathsf{g}}^2(\rho_t,\mu)=\int\phi_t\,\mathrm{d}(\rho_t-\mu)\le C_0\sqrt{J_t},\qquad J_t\ge\frac{1}{C_0^2}\big[\mathrm{MMD}_{\mathsf{g}}^2(\rho_t,\mu)\big]^2.$$

To bound the remaining term, Fourier expansion and Cauchy–Schwarz yield

$$\|\nabla\phi_t\|_\infty^2\le\left(\sum_{k\neq 0}|k|^2\widehat{\mathsf{g}}(k)\right)\left(\sum_{k\neq 0}\widehat{\mathsf{g}}(k)|\widehat{\rho_t-\mu}(k)|^2\right)=-\Delta\mathsf{g}(0)\ \mathrm{MMD}_{\mathsf{g}}^2(\rho_t,\mu).$$

Combining these estimates with (116) proves (114).

Suppose now that $M_2(t) \to 0$. For any $\varepsilon > 0$, choose T so that $C_{\mathrm{g}} M_2(t) \le \varepsilon^{3/2}/(2C_0^2)$ for $t \ge T$. Whenever $\mathrm{MMD}_{\mathrm{g}}^2(\rho_t, \mu) \ge \varepsilon$, (114) then gives

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{MMD}_{\mathrm{g}}^2(\rho_t, \mu) \le -\frac{1}{C_0^2} \big[\mathrm{MMD}_{\mathrm{g}}^2(\rho_t, \mu)\big]^2 \le -\frac{\varepsilon^2}{C_0^2}.$$

Thus the squared MMD enters $[0, \varepsilon]$ in finite time and cannot subsequently cross ε upwards. Letting $\varepsilon \to 0$ proves $\mathrm{MMD}_{\mathrm{g}}^2(\rho_t, \mu) \to 0$.

Finally, compactness of the torus gives a narrowly convergent subsequence of any sequence (ρ_{t_n}) with $t_n \to \infty$. The Fourier formula for $\mathrm{MMD}_{\mathrm{g}}^2(\rho_t, \mu)$ and strict positivity of $\widehat{\mathrm{g}}(k)$ imply that every such limit has the same Fourier coefficients as μ, and hence equals μ. Thus $\rho_t \to \mu$ narrowly and in W_2. For the product metric,

$$W_2(f_t, \mu \otimes \delta_0)^2 = M_2(t) + W_2(\rho_t, \mu)^2,$$

because the target fluctuation is always zero and an optimal positional coupling can be lifted using the conditional law of v given θ. This proves (115). □

Corollary B.11 (Rates from fluctuation decay)**.** *Under the assumptions of Theorem B.10, suppose $M_2(t) \le A(1+t)^{-\alpha}$ for some $A, \alpha > 0$. Set $q = \min\{1, 2\alpha/3\}$. Then, for a constant C depending only on $\mathrm{g}, C_0, A, \alpha$ and $\mathrm{MMD}_{\mathrm{g}}^2(\rho_0, \mu)$,*

$$\mathrm{MMD}_{\mathrm{g}}^2(\rho_t, \mu) \le C(1+t)^{-q}. \tag{117}$$

Proof. Choose $C > \mathrm{MMD}_{\mathrm{g}}^2(\rho_0, \mu)$ large enough that $2C^2/C_0^2 - 2C_{\mathrm{g}} A\sqrt{C} \ge qC$. Since $3q/2 \le \alpha$ and $q \le 1$, the positive function $y_t = C(1+t)^{-q}$ satisfies

$$\frac{\mathrm{d}y_t}{\mathrm{d}t} \ge -\frac{2}{C_0^2} y_t^2 + 2C_{\mathrm{g}} A (1+t)^{-\alpha} \sqrt{y_t}.$$

Comparison with (114) gives $\mathrm{MMD}_{\mathrm{g}}^2(\rho_t, \mu) \le y_t$. Since $y_t > 0$, the right-hand side is locally Lipschitz at any possible contact point. □

This is a squared-MMD rate, not a rate in L^p norms of densities. When $\alpha \ge 3/2$, the estimate recovers the classical t^{-1} squared-MMD bound; no optimality of these exponents is asserted. The dual bound (113) and the decay of M_2 are assumptions on the trajectory, not consequences of well-posedness.

B.4 Propagation of higher regularity

We show that regularity in the full variable $z = (\theta, v)$, identifying $\mathbb{R}^{d \times r}$ with $\mathbb{R}^{dr}$, is propagated by the dynamic.

A multiindex $\alpha = (\alpha_1, \dots, \alpha_{d+dr}) \in \mathbb{N}_0^{d+dr}$ records the number of derivatives in each coordinate:

$$|\alpha| = \sum_{j=1}^{d+dr} \alpha_j, \qquad D_z^\alpha = \partial_{z_1}^{\alpha_1} \cdots \partial_{z_{d+dr}}^{\alpha_{d+dr}}.$$

Here $W^{s,q} = W^{s,q}(\mathbb{R}^d \times \mathbb{R}^{d \times r})$ includes all weak derivatives of total order at most s, including mixed derivatives, and we use the norm $\|h\|_{W^{s,q}} = \sum_{|\alpha| \le s} \|D_z^\alpha h\|_{L^q}$.

Proposition B.12 (Propagation of joint regularity)**.** *Let $s \ge 0$ be an integer, $q \in [1, \infty]$, $\mu \in \mathcal{P}(\mathbb{R}^d)$, and let $\mathrm{g} \in C_b^{s+4}(\mathbb{R}^d)$ be even. Suppose that $f_0 \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^{d \times r})$ has a density in $W^{s,q}$ supported in $\{|v| \le M\}$. Then the characteristic solution of* (109) *has a density satisfying*

$$\|f_t\|_{W^{s,q}} \le e^{C_T t} \|f_0\|_{W^{s,q}}, \qquad 0 \le t \le T,$$

for every $T < \infty$. The constant C_T depends only on T, M, s, q, d, r and $\|\mathrm{g}\|_{C_b^{s+4}}$.

Proof. Fix $T > 0$ and consider the transport velocity $\mathcal{V}(t, z; f_t) = (\mathcal{V}_\theta, \mathcal{V}_v)(t, z; f_t)$ defined by the right-hand sides of (110). By Proposition B.8, there is R_T such that $\operatorname{supp}_v f_t \subset B_{R_T}$ for $0 \le t \le T$. The θ-derivatives of this velocity act on the convolution kernels, not on f_t, and its dependence on v is at most quadratic. Since its expression uses at most three derivatives of g, the assumption $\mathrm{g} \in C_b^{s+4}$ bounds all its derivatives in $z = (\theta, v)$ through total order $s + 1$ on $\mathbb{R}^d \times B_{R_T+1}$. Choose a smooth cutoff $\chi(v)$ equal to one on B_{R_T} and zero outside B_{R_T+1}, and set

$$b_t(z) = \chi(v)\mathcal{V}(t, z; f_t).$$

This cutoff velocity agrees with the original transport velocity $\mathcal{V}(t, z; f_t)$ on the support of f_t and satisfies

$$\sup_{0\le t\le T} \|b_t\|_{W^{s+1,\infty}} \le C_T.$$

Let Φ_t be the characteristic flow generated by b_t. It is a C^{s+1} diffeomorphism; the derivatives of Φ_t and Φ_t^{-1} of orders $1, \dots, s + 1$ are bounded on $[0, T]$, and their Jacobian determinants are bounded away from zero. The characteristic formula

$$f_t = (f_0 \circ \Phi_t^{-1})\, |\det D_z \Phi_t^{-1}|$$

and the Sobolev chain and product rules therefore give $f_t \in W^{s,q}$. We now estimate this norm. For $|\alpha| \le s$, differentiating $\partial_t f + b_t \cdot \nabla_z f + (\operatorname{div}_z b_t) f = 0$ in the sense of distributions gives

$$\partial_t D_z^\alpha f + b_t \cdot \nabla_z D_z^\alpha f + (\operatorname{div}_z b_t) D_z^\alpha f = R_\alpha,$$

$$R_\alpha = -\sum_{0<\beta\le\alpha} \binom{\alpha}{\beta} \left[(D_z^\beta b_t) \cdot \nabla_z D_z^{\alpha-\beta} f + D_z^\beta (\operatorname{div}_z b_t) D_z^{\alpha-\beta} f \right],$$

where $\beta \le \alpha$ is understood coordinatewise and $\binom{\alpha}{\beta} = \prod_j \binom{\alpha_j}{\beta_j}$. In the first term, the derivative of f has order $|\alpha| - |\beta| + 1 \le |\alpha|$; in the second, at most $s + 1$ derivatives fall on b_t. Consequently,

$$\|R_\alpha\|_{L^q} \le C_T \|f\|_{W^{s,q}}.$$

For $1 < q < \infty$, computing the L^q norm of $g = D_z^\alpha f$ along the characteristics of b_t gives

$$\begin{aligned} \frac{1}{q}\frac{\mathrm{d}}{\mathrm{d}t}\|g\|_{L^q}^q &= -\left(1 - \frac{1}{q}\right) \int (\operatorname{div}_z b_t) |g|^q \,\mathrm{d}z + \int R_\alpha |g|^{q-2} g \,\mathrm{d}z \\ &\le C_T \|g\|_{L^q}^q + \|R_\alpha\|_{L^q} \|g\|_{L^q}^{q-1}. \end{aligned}$$

Dividing by $\|g\|_{L^q}^{q-1}$, with a regularization when this norm vanishes, gives $\frac{\mathrm{d}}{\mathrm{d}t}\|g\|_{L^q} \le C_T \|g\|_{L^q} + \|R_\alpha\|_{L^q}$. For $q = 1$ and $q = \infty$, the same inequality follows by taking the corresponding norm in the characteristic formula for g. Summing over $|\alpha| \le s$ and applying Grönwall's lemma proves the estimate. □

Initial regularity in θ alone is not enough to propagate. We record an explicit example where an initial jump in v can make $\partial_\theta f_t$ singular.

Take $d = r = 1$, $\mathrm{g}(\theta) = \cos\theta$, $\mu = \delta_0$, and

$$f_0(\theta, v) = \frac{h(\theta)}{2a} \mathbf{1}_{[-a,a]}(v), \qquad a > 0,$$

where h is a smooth, even, compactly supported probability density, with $h(\pi/2) > 0$. Every θ-derivative of f_0 belongs to every L^q, but f_0 has jumps at $v = \pm a$. Since $\int_{-a}^{a} v' \,\mathrm{d}v' = 0$ and $\int \sin\theta' \, h(\theta') \,\mathrm{d}\theta' = 0$, substitution into (110) gives

$$\begin{aligned} \mathcal{V}_v(0, \theta, v) &= v \left(\int \cos(\theta - \theta') h(\theta') \,\mathrm{d}\theta' - \cos\theta \right) \\ &= -c\, v \cos\theta, \qquad c = 1 - \int \cos\theta' \, h(\theta') \,\mathrm{d}\theta' > 0. \end{aligned}$$

The jump initially at $v = a$ is carried by the characteristics; write its position at small times as $v = a_t(\theta)$. Since $a_0(\theta) = a$ is constant, its initial velocity is $\partial_t a_t(\theta)|_{t=0} = \mathcal{V}_v(0, \theta, a) = -a\, c \cos\theta$. Expanding a_t and its θ-derivative in time gives

$$a_t(\theta) = a\big(1 - c\, t \cos\theta\big) + o(t), \qquad \partial_\theta a_t(\pi/2) = a\, c\, t + o(t) > 0$$

for sufficiently small $t > 0$. Thus the jump location, initially independent of θ, now varies with θ.

Near this jump and $\theta = \pi/2$, the transported density has the form $f_t(\theta, v) = A_t(\theta, v)\mathbf{1}_{\{v<a_t(\theta)\}}$, where A_t is smooth and positive. Differentiating in θ, one obtains:

$$\partial_\theta f_t = (\partial_\theta A_t)\mathbf{1}_{\{v<a_t(\theta)\}} + A_t(\theta, a_t(\theta))\,\partial_\theta a_t(\theta)\,\delta\big(v - a_t(\theta)\big).$$

The second term is singular in v, and its coefficient is nonzero near $\theta = \pi/2$. Hence, for every sufficiently small $t > 0$, the first weak derivative $\partial_\theta f_t$ is not an L^q function for any $1 \le q \le \infty$, although all θ-derivatives of f_0 are in L^q.

B.5 Proof of the mean-field limit

In this section, we prove Theorem 4.2.

The limiting passage on a fixed phase space follows the classical Braun–Hepp–Dobrushin argument (Braun & Hepp, 1977; Dobrushin, 1979). For more recent accounts on mean-field limits, we refer to Chaintron & Diez (2022); Golse (2016); Jabin (2014). The non-standard point has to do with the dimension of the fluctuation covariance: the full Σ is a $dN \times dN$ matrix, whereas the kinetic equation is posed on the N-independent phase space $\mathbb{R}^d \times \mathbb{R}^{d\times r}$. We overcome this obstruction by introducing an intermediate rank-r particle system and proving that the full dynamics stays close to it. This first step is essentially a stability estimate for two finite-dimensional ODEs. Once it is established, the classical mean-field argument applies directly.

We first state again the theorem:

Theorem B.13 (Validity of the kinetic limit)**.** *Fix $\eta > 0$ and r independently of N, let $\mathrm{g} \in C_b^4(\mathbb{R}^d)$ be even, and let $\mu \in \mathcal{P}(\mathbb{R}^d)$. Let $f_0 \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^{d\times r})$ be supported in $\{\|v\| \le M\}$. For each N, choose $z_i^{N,0} = (\theta_i^{N,0}, v_i^{N,0})$ with $\max_i \|v_i^{N,0}\| \le M$ and set*

$$f_N^0 = \frac{1}{N}\sum_{i=1}^N \delta_{z_i^{N,0}}, \qquad \Sigma^N(0) = V^N(0)V^N(0)^T,$$

where the i-th block row of $V^N(0)$ is $v_i^{N,0}$. For a matrix S with $d \times d$ blocks, write $\|S\|_{\max} = \max_{i,j}\|S_{ij}\|$.

Let (θ^N, Σ^N) solve the full system (10)*. Let $(\bar\theta^N, V^N)$ solve the rank-r system obtained by deleting the quadratic covariance term, with the same initial positions and factor, and define*

$$\bar f_N(t) = \frac{1}{N}\sum_{i=1}^N \delta_{(\bar\theta_i^N(t), v_i^N(t))}.$$

Finally, let f_t be the solution of (109) *with initial law f_0, and let ρ_t be its θ-marginal. Then, for every $T > 0$, there is a constant C_T, independent of N, such that*

$$\sup_{0\le t\le T} W_1(\bar f_N(t), f_t) \le C_T W_1(f_N^0, f_0), \tag{118}$$

$$\sup_{0\le t\le T}\Big(\max_i |\theta_i^N(t) - \bar\theta_i^N(t)| + \|\Sigma^N(t) - V^N(t)V^N(t)^T\|_{\max}\Big) \le \frac{C_T}{N}. \tag{119}$$

Consequently, for $\rho_N^{\mathrm{full}}(t) = N^{-1}\sum_i \delta_{\theta_i^N(t)}$,

$$\sup_{0\le t\le T} W_1(\rho_N^{\mathrm{full}}(t), \rho_t) \le C_T\left(W_1(f_N^0, f_0) + \frac{1}{N}\right). \tag{120}$$

In particular, if $(z_i^0)_{i\ge1}$ is an i.i.d. sequence with law f_0 and $z_i^{N,0} = z_i^0$ for $i \le N$, then all three conclusions hold almost surely with right-hand sides tending to zero.

Full and fixed-rank particle systems. For $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N)$, set

$$\rho_N = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i}, \qquad E_N(\boldsymbol{\theta}) = \frac{1}{2}\int \mathrm{g} * (\rho_N - \mu)\,\mathrm{d}(\rho_N - \mu).$$

We suppress the superscript N on all particle variables. In the time scale of (10), the full system is

$$\begin{cases} \dfrac{\mathrm{d}\theta_i}{\mathrm{d}t} = -N\nabla_{\theta_i} E_N(\boldsymbol{\theta}) - \dfrac{N}{2}\nabla_{\theta_i}\nabla^2 E_N(\boldsymbol{\theta}) : \Sigma, & 1 \le i \le N, \\ \dfrac{\mathrm{d}\Sigma_{ij}}{\mathrm{d}t} = -N\displaystyle\sum_{k=1}^{N}\left(\nabla^2_{\theta_i\theta_k} E_N(\boldsymbol{\theta})\Sigma_{kj} + \Sigma_{ik}\nabla^2_{\theta_k\theta_j} E_N(\boldsymbol{\theta})\right) \\ \qquad\qquad +\eta N\displaystyle\sum_{k,l=1}^{N}\nabla^2_{\theta_i\theta_k} E_N(\boldsymbol{\theta})\Sigma_{kl}\nabla^2_{\theta_l\theta_j} E_N(\boldsymbol{\theta}), \quad 1 \le i,j \le N. \end{cases} \tag{121}$$

Starting from $\Sigma(0) = V(0)V(0)^T$, where the i-th block row of V is $v_i \in \mathbb{R}^{d\times r}$, we compare it with the fixed-rank system

$$\begin{cases} \dfrac{\mathrm{d}\bar\theta_i}{\mathrm{d}t} = -N\nabla_{\theta_i} E_N(\bar{\boldsymbol{\theta}}) - \dfrac{N}{2}\nabla_{\theta_i}\nabla^2 E_N(\bar{\boldsymbol{\theta}}) : VV^T, & 1 \le i \le N, \\ \dfrac{\mathrm{d}v_i}{\mathrm{d}t} = -N\displaystyle\sum_{j=1}^{N}\nabla^2_{\theta_i\theta_j} E_N(\bar{\boldsymbol{\theta}})v_j, & 1 \le i \le N. \end{cases} \tag{122}$$

Writing $Q_{ij} = v_i v_j^T$, the second equation implies

$$\frac{\mathrm{d}Q_{ij}}{\mathrm{d}t} = -N\sum_{k=1}^{N}\left(\nabla^2_{\theta_i\theta_k} E_N(\bar{\boldsymbol{\theta}})Q_{kj} + Q_{ik}\nabla^2_{\theta_k\theta_j} E_N(\bar{\boldsymbol{\theta}})\right), \qquad \operatorname{rank} Q(t) \le r.$$

Thus the intermediate dynamics evolves in a fixed phase space. The next proposition contains the non-classical part of the proof. It states that every (i,j) block of the full covariance remains $O(N^{-1})$-close to the corresponding block of the rank-r matrix Q, although the quadratic term in (121) need not preserve algebraic rank.

Proposition B.14 (Stability of the fixed-rank reduction)**.** *Under the assumptions of Theorem 4.2, suppose that* (121) *and* (122) *start from the same positions, from* $\Sigma(0) = V(0)V(0)^T$*, and from* $\max_i \|v_i(0)\| \le M$*. Then, for every* $T > 0$*,*

$$\sup_{0\le t\le T}\left(\max_i |\theta_i(t) - \bar\theta_i(t)| + \|\Sigma(t) - V(t)V(t)^T\|_{\max}\right) \le \frac{C_T}{N}.$$

Proof. For a block matrix, write $\|S\|_{\max} = \max_{i,j}\|S_{ij}\|$. The rescaled Hessian writes, blockwise, as

$$N\nabla^2_{\theta_i\theta_j} E_N(\boldsymbol{\theta}) = \delta_{ij}\nabla^2\mathrm{g} * (\rho_N - \mu)(\theta_i) - \frac{1}{N}\nabla^2\mathrm{g}(\theta_i - \theta_j). \tag{123}$$

Hence, uniformly in N,

$$\max_i \sum_{j=1}^{N}\left\|N\nabla^2_{\theta_i\theta_j} E_N(\boldsymbol{\theta})\right\|_{\mathrm{op}} \le C,$$

$$\max_i \sum_{j=1}^{N}\left\|N\nabla^2_{\theta_i\theta_j} E_N(\boldsymbol{\theta}) - N\nabla^2_{\theta_i\theta_j} E_N(\bar{\boldsymbol{\theta}})\right\|_{\mathrm{op}} \le C\max_i |\theta_i - \bar\theta_i|.$$

Bounded fourth derivatives of g similarly show that, whenever $\|S\|_{\max} + \|\bar S\|_{\max} \le K$, the maximal difference between the right-hand sides of the position equations evaluated at $(\boldsymbol{\theta}, S)$ and $(\bar{\boldsymbol{\theta}}, \bar S)$ is bounded by

$$C_K\left(\max_i |\theta_i - \bar\theta_i| + \|S - \bar S\|_{\max}\right).$$

The constants are independent of N: each off-diagonal block carries a factor N^{-1}, while every diagonal term is a normalized sum.

By symmetry, the same bounds hold when one sums over i with j fixed. They therefore control multiplication by $N\nabla^2 E_N$ on either side in the $\|\cdot\|_{\max}$ norm. Gronwall's inequality yields, on $[0,T]$,

$$\max_i \|v_i(t)\| \le Me^{Ct}, \qquad \|\Sigma(t)\|_{\max} + \|Q(t)\|_{\max} \le C_T.$$

Set $R = \Sigma - Q$ and

$$u(t) = \max_i |\theta_i(t) - \bar{\theta}_i(t)|, \qquad r(t) = \max_{i,j} \|R_{ij}(t)\|.$$

For any i, the integral form of the two position equations, their common initial value, and the Lipschitz estimate above give

$$|\theta_i(t) - \bar{\theta}_i(t)| \le C_T \int_0^t \big(u(s) + r(s)\big)\, \mathrm{d}s.$$

On the other hand, subtraction of the equations for Σ and Q gives

$$\begin{aligned} \dot{R} = & -N\nabla^2 E_N(\boldsymbol{\theta})R - NR\nabla^2 E_N(\boldsymbol{\theta}) \\ & - N\big(\nabla^2 E_N(\boldsymbol{\theta}) - \nabla^2 E_N(\bar{\boldsymbol{\theta}})\big)Q \\ & - NQ\big(\nabla^2 E_N(\boldsymbol{\theta}) - \nabla^2 E_N(\bar{\boldsymbol{\theta}})\big) \\ & + \eta N\nabla^2 E_N(\boldsymbol{\theta})\Sigma\nabla^2 E_N(\boldsymbol{\theta}). \end{aligned}$$

All terms except the last one are Lipschitz in the current error; the last is $O(N^{-1})$ in block norm. Since $R(0) = 0$, for any i, j,

$$\|R_{ij}(t)\| \le C_T \int_0^t \big(u(s) + r(s)\big)\, \mathrm{d}s + \frac{C_T t}{N}.$$

Taking the maximum first over i, then over (i, j), and using $t \le T$, we obtain

$$u(t) + r(t) \le C_T \int_0^t \big(u(s) + r(s)\big)\, \mathrm{d}s + \frac{C_T}{N}.$$

Gronwall's inequality proves (119). □

Mean-field limit. It remains to pass to the limit in the fixed-rank system. This is the classical Dobrushin argument, which we give for completeness.

Proof of Theorem 4.2. Set $z = (\theta, v)$ and identify the target with $\widehat{\mu} = \mu \otimes \delta_0$. The interaction kernel on the product space is

$$K(z, z') = \mathsf{g}(\theta - \theta') + \frac{1}{2}\nabla^2 \mathsf{g}(\theta - \theta') : (v - v')(v - v')^T.$$

For a probability measure h, write

$$\mathcal{V}[h](z) = -\int \nabla_z K(z, z')\, \mathrm{d}(h - \widehat{\mu})(z').$$

If $z_i(t) = (\bar{\theta}_i(t), v_i(t))$ and

$$\bar{f}_N(t) = \frac{1}{N}\sum_{i=1}^N \delta_{z_i(t)},$$

then direct substitution in the preceding formula shows that (122) is exactly

$$\frac{\mathrm{d}z_i}{\mathrm{d}t} = \mathcal{V}[\bar{f}_N](z_i), \qquad 1 \le i \le N.$$

Consequently, if Z_N denotes the characteristic flow generated by $\mathcal{V}[\bar{f}_N]$, then

$$\bar{f}_N(t) = Z_N(t, \cdot)_\# f_N^0.$$

Likewise, if Z is the characteristic flow generated by $\mathcal{V}[f]$, then $f_t = Z(t, \cdot)_\# f_0$.

We first record the Lipschitz estimates used below. The bounds on the v_i obtained in the proof of Proposition B.14, together with Proposition B.8, give $R_T > 0$, independent of N, such that

$$\operatorname{supp}_v \bar{f}_N(t) \cup \operatorname{supp}_v f_t \subset B_{R_T}, \qquad 0 \le t \le T.$$

Since $\mathsf{g} \in C_b^4$, differentiation of K gives, for $z, \widetilde{z}, z', \widetilde{z}' \in \mathbb{R}^d \times B_{R_T}$,

$$|\nabla_z K(z, z') - \nabla_z K(\widetilde{z}, z')| \le L_T |z - \widetilde{z}|, \tag{124}$$

$$|\nabla_z K(z, z') - \nabla_z K(z, \widetilde{z}')| \le L_T |z' - \widetilde{z}'|. \tag{125}$$

Indeed, the θ-component of $\nabla_z K$ contains derivatives of g up to order three multiplied by at most two factors $v - v'$, while its v-component is linear in $v - v'$. One further derivative therefore uses at most four derivatives of g, and all velocity factors are bounded by $2R_T$.

It follows from (124) that, for every probability measure h supported in $v \in B_{R_T}$,

$$|\mathcal{V}[h](z) - \mathcal{V}[h](\widetilde{z})| \le 2L_T |z - \widetilde{z}|. \tag{126}$$

The factor two accounts for the total variation of $h - \widehat{\mu}$. Moreover, if h_1, h_2 have such support and $\pi \in \Pi(h_1, h_2)$, then the target terms cancel and

$$\begin{aligned} |\mathcal{V}[h_1](z) - \mathcal{V}[h_2](z)| &\le \int |\nabla_z K(z, z') - \nabla_z K(z, \widetilde{z}')| \, \mathrm{d}\pi(z', \widetilde{z}') \\ &\le L_T \int |z' - \widetilde{z}'| \, \mathrm{d}\pi(z', \widetilde{z}'). \end{aligned} \tag{127}$$

We now construct the coupling. Let $\pi_0 \in \Pi(f_N^0, f_0)$ be arbitrary and transport its two coordinates by the two flows:

$$\pi_t = \big((z, \widetilde{z}) \mapsto (Z_N(t, z), Z(t, \widetilde{z}))\big)_{\#} \pi_0.$$

Then $\pi_t \in \Pi(\bar{f}_N(t), f_t)$. Define its transportation cost by

$$D(t) = \int |Z_N(t, z) - Z(t, \widetilde{z})| \, \mathrm{d}\pi_0(z, \widetilde{z}).$$

In particular,

$$W_1(\bar{f}_N(t), f_t) \le D(t). \tag{128}$$

For fixed $(z, \widetilde{z})$, the integral form of the characteristic equations gives

$$\begin{aligned} |Z_N(t, z) - Z(t, \widetilde{z})| &\le |z - \widetilde{z}| \\ &\quad + \int_0^t |\mathcal{V}[\bar{f}_N(s)](Z_N(s, z)) - \mathcal{V}[f_s](Z(s, \widetilde{z}))| \, \mathrm{d}s. \end{aligned}$$

Split the last difference by inserting $\mathcal{V}[\bar{f}_N(s)](Z(s, \widetilde{z}))$. The spatial estimate (126) controls the first part by $2L_T |Z_N(s, z) - Z(s, \widetilde{z})|$. For the second part, use (127) with the coupling π_s to obtain the bound $L_T D(s)$. Integrating with respect to π_0 therefore yields

$$D(t) \le D(0) + 3L_T \int_0^t D(s) \, \mathrm{d}s.$$

Gronwall's inequality gives

$$D(t) \le e^{3L_T t} D(0) = e^{3L_T t} \int |z - \widetilde{z}| \, \mathrm{d}\pi_0(z, \widetilde{z}).$$

Combining this estimate with (128) and finally taking the infimum over $\pi_0 \in \Pi(f_N^0, f_0)$ proves Dobrushin's estimate

$$\sup_{0 \le t \le T} W_1(\bar{f}_N(t), f_t) \le e^{3L_T T} W_1(f_N^0, f_0),$$

which is (118).

The stability estimate of Proposition B.14 gives (119). Finally, coupling the full and fixed-rank particles with the same index and projecting onto the θ coordinate gives

$$\begin{aligned} W_1(\rho_N^{\mathrm{full}}(t), \rho_t) &\le \frac{1}{N} \sum_{i=1}^N |\theta_i(t) - \bar{\theta}_i(t)| + W_1\big((\mathsf{p}_\theta)_{\#} \bar{f}_N(t), \rho_t\big) \\ &\le \max_i |\theta_i(t) - \bar{\theta}_i(t)| + W_1(\bar{f}_N(t), f_t). \end{aligned}$$

This proves (13). For independent initial samples with law f_0, the empirical measures f_N^0 converge almost surely to f_0 in W_1 because $f_0 \in \mathcal{P}_2$. $\square$

C ADDITIONAL REMARKS

C.1 SCALING OF THE FLUCTUATIONS

A natural question concerns the scaling of the fluctuation covariance with respect to the parameters, in particular the learning rate η. We argue here that the natural scaling is, at the edge of stability, $\mathrm{tr}(\Sigma) \approx \eta$, as is the case in the gradient-flow limit Proposition A.14.

Indeed, the natural quantity we monitor, as we have seen, is

$$F(\theta, \Sigma) := E(\theta) + \frac{1}{2} H(\theta) : \Sigma.$$

At the edge of stability, we have $H : \Sigma \approx 2\,\mathrm{tr}(\Sigma)/\eta$, so that to equilibrate the energy part and the fluctuating part, one has

$$\boxed{\mathrm{tr}(\Sigma) \approx \eta E.}$$

C.2 ON THE ONSET FAILURE OF ROD FLOW

As observed by Marion (2026), the Rod flow is unable to settle at the edge of stability when initialized at the onset of the threshold: oscillations are not developing, as for the gradient flow. In this section, we show that because the mean–fluctuation dynamics retains covariance in all directions, it can be more robust than Rod flow to this failure mode.

First, we show that this failure mode is robust to numerical refinement, then assess the impact of flooring the covariance amplitude (a procedure used in Edge flow) and finally show that the mean–fluctuation dynamics does not always suffer from the same failure mode as Rod flow. The explanation of the latter is that the evolution of MF retains the full covariance matrix, which can amplify initially small oscillations aligned with the top Hessian eigenvector. This is crucial, because at the onset of EoS, the top eigenvector of Σ (which is the one retained in Rod flow) can be nearly orthogonal to the top Hessian direction.

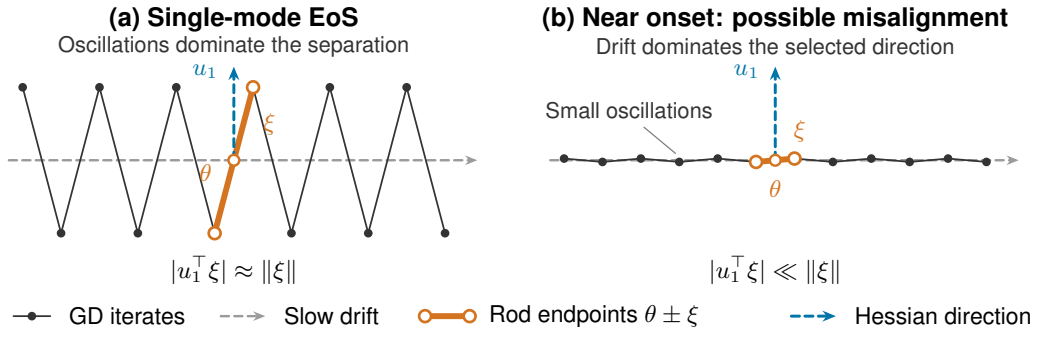


Figure 10: Rod's onset failure. The selected displacement is $\xi = \sqrt{s_1} v_1$, where (s_1, v_1) is the leading eigenpair of Σ. (a) At single-mode EoS, the oscillations are along the leading Hessian eigenvector u_1 (Marion, 2026, see also Fig. 2). (b) Near onset, the selected direction can instead be nearly orthogonal to u_1, even if the full covariance contains smaller fluctuations along u_1. MF can amplify these smaller fluctuations.

Onset comparison and numerical refinement. We first check that this failure mode persists under numerical refinement. To this aim, we initialize from the same saved GD window at $\eta\lambda_1 \simeq 1.999$ and evolve for 4000 steps. MF uses[1] $(r, p, K) = (3, 2, 4)$ as a baseline and $(4, 3, 8)$ for the refined run; Rod uses auxiliary rank three with $K = 4$ and rank four with $K = 8$. The initial MF covariance and Rod auxiliary matrix Σ are fixed and have rank three in both runs. Neither MF nor Rod has an

[1]We refer to Appendix E for the definition of these numerical parameters for MF. K is the number of Euler substep per gradient descent step.

amplitude floor; Edge retains its prescribed floor and uses $K = D = 4$ in both runs.[2] In Figure 5, MF recovers the growth of the GD fluctuations whereas Rod does not, at both resolutions. The maximum relative full-covariance differences over the 81 common checkpoints are 1.86% for MF and 0.0506% for Rod's endpoint covariance, using the normalization of (148). The replayed GD states agree exactly.

Persistent amplitude floors. After checking that this behavior persists under numerical refinement, we wonder about the role of a hypothetical floor for the amplitude of the fluctuations. Indeed, at each step of Algorithm 1 (Marion, 2026), the amplitude x of oscillations is floored by $\max(a, x)$, which may explain why Edge flow is not concerned by this failure mode.

We repeat the baseline Rod run with RMS floors $a \in \{10^{-16}, 10^{-12}, 10^{-8}, 10^{-6}, 10^{-4}\}$, using the same cached GD window, initial center, retained basis and 4000-step horizon.

Write Rod's auxiliary matrix as $\Sigma = \sum_{i=1}^{r} s_i v_i v_i^T$, with orthonormal v_i and $s_1 \geq \cdots \geq s_r$. At initialization and after every covariance substep, the floor replaces

$$s_i \leftarrow \max\{s_i, a^2\}, \qquad i = 1, \dots, r.$$

The two endpoints remain $\theta \pm \xi$, with $\xi = \sqrt{s_1}\, v_1$: their covariance is $\xi\xi^T$, not the full auxiliary matrix Σ. Along the leading Hessian eigenvector u_1 at the GD-window mean, the plotted endpoint variance is therefore

$$(u_1^T \xi)^2 = s_1 |u_1^T v_1|^2,$$

which need not exceed a^2 if ξ and u_1 are nearly orthogonal. It also leaves directions outside the retained subspace unchanged. Thus $a = 10^{-6}$ matches Edge's RMS floor magnitude, but not its rule for choosing the fluctuation direction.

Figure 11 separates two outcomes. For every tested $a \leq 10^{-6}$, Rod's loss and sharpness curves overlap with the unfloored run, and the critical endpoint variance remains negligible. Even at $a = 10^{-6}$, the principal RMS amplitude stays above $4.5\,10^{-6}$: the floor raises the smaller extent eigenvalues, rather than directly increasing $\|\xi\|$. At the last diagnostic sample, $|u_1^T v_1| \simeq 2.4\,10^{-5}$ and $(u_1^T \xi)^2 \simeq 1.2\,10^{-20}$, despite the nonzero extent.

For $a = 10^{-4}$, substantial critical variance does develop, and Rod's sharpness returns close to 2. At the last diagnostic sample, its critical variance $(u_1^T \xi)^2$ is $2.60\,10^{-7}$, compared with $1.99\,10^{-7}$ for GD. However, the variance grows and the loss peaks much earlier than in GD. This stronger floor also changes the initial extent: all three retained eigenvalues become a^2, so the largest is initially multiple.

The critical direction of oscillations is discarded by Rod while amplified by MF. The experiment of Figure 5 shows that the top Hessian direction u_1 and Rod's retained direction ξ are nearly orthogonal at the onset of EoS (see also Figure 2).

This may be the reason why Rod flow fails to develop oscillations, as explained by Marion (2026). On the opposite, MF can still increase the small oscillations aligned with u_1, because all directions of $\operatorname{Ran}\Sigma$ are retained for its evolution. We also note that Edge flow does not suffer from this failure mode because the oscillation direction is *defined* to align with the top Hessian direction through a Rayleigh-quotient flow.

To fix ideas, freeze the Hessian H and let (λ_1, u_1) denotes its top eigenpair: $Hu_1 = \lambda_1 u_1$ with $\|u_1\| = 1$. Consider that

$$\Sigma = s_{\text{drift}} v v^T + s_{\text{crit}} u_1 u_1^T, \quad \|v\| = 1, \quad v \perp u_1, \quad s_{\text{drift}} > s_{\text{crit}} > 0.$$

Rod flow selects $\xi = \sqrt{s_{\text{drift}}}\, v$. Therefore, linearizing the evolution around this initial condition gives

$$\begin{aligned} \left.\frac{ds_{\text{crit}}}{dt}\right|_{\text{Rod}} &\approx \frac{\eta^2}{2} |u_1^T \nabla E(\theta)|^2 - 2 s_{\text{crit}}, \\ &\approx -2 s_{\text{crit}}, \end{aligned}$$

assuming $\eta^2 |u_1^T \nabla E(\theta)|^2 \ll 4 s_{\text{crit}}$. As raised by Marion (2026), the critical variance s_{crit} therefore cannot grow under these assumptions.

[2] See Appendix D.7 for the numerical parameters of Edge flow.

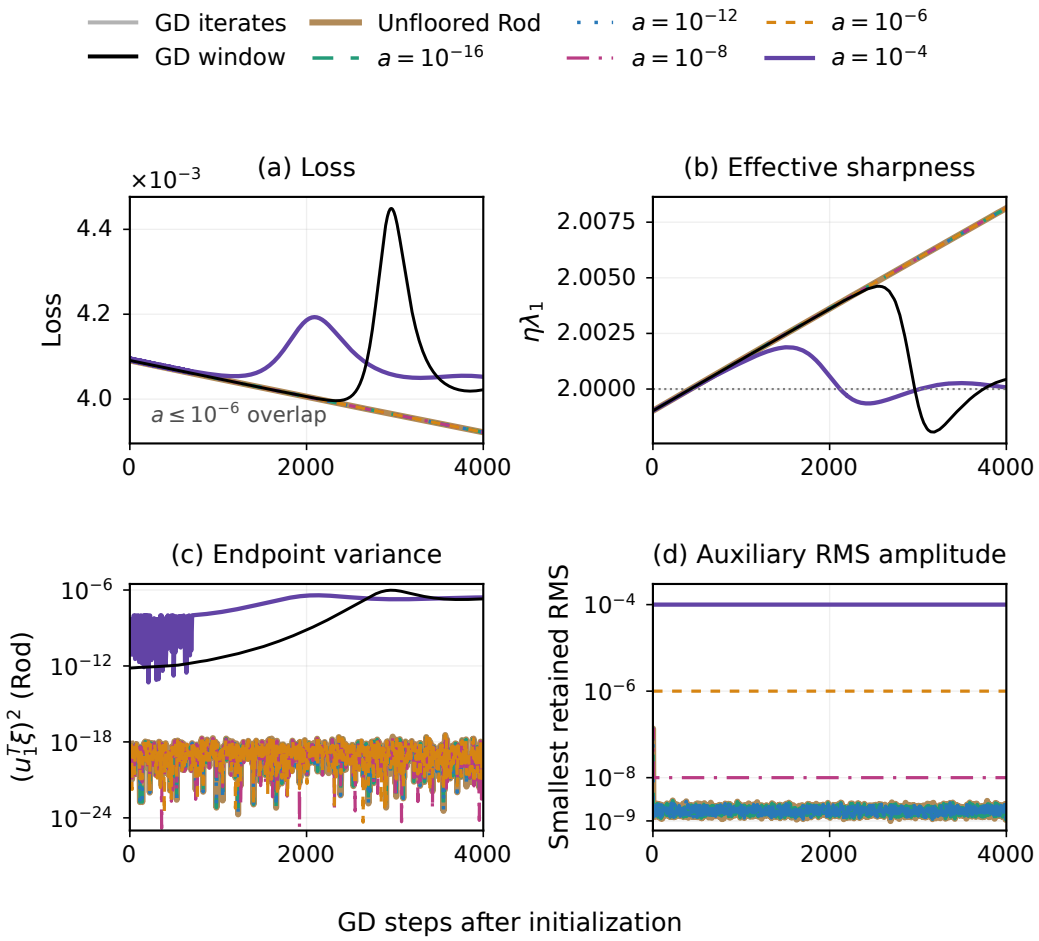


Figure 11: Floored Rod flows at $\eta = 0.002$, from the same cached GD window as Figure 5, with Σ of rank three, $K = 4$ and 4000 steps. The floor acts at initialization and throughout the evolution. The tested floors up to 10^{-6} leave the onset failure unchanged; 10^{-4} produces critical fluctuations, but earlier than in GD.

On the opposite, the mean–fluctuation evolution gives

$$\left.\frac{ds_{\text{crit}}}{dt}\right|_{\text{MF}} \approx -2\eta\lambda_1 s_{\text{crit}} + \eta^2\lambda_1^2 s_{\text{crit}} = \eta\lambda_1(-2 + \eta\lambda_1)s_{\text{crit}}.$$

There is therefore no need for a rotation of v to allow for growing oscillations in s_{crit}: as soon as $\eta\lambda_1 > 2$, the critical variance grows exponentially.

MF can therefore amplify critical fluctuations before they become dominant; the linearized Rod forcing does not amplify them when its selected direction is orthogonal to u_1. We finally note that this is a frozen-Hessian explanation: the directions can rotate along the nonlinear dynamics, so that we have not provided the full mechanism.

We do not claim that MF is totally immune to this failure mode: for fixed H, an initially zero critical variance remains zero, and strong subcritical damping can delay its recovery.

D Additional experimental details

D.1 Experimental protocol and summary

All experiments use full-batch GD on subsets of CIFAR-10, mean-squared loss, and no data augmentation. The MLP, CNN and ResNet have widths 256, 32 and 16, respectively, with GELU activations; the ResNet has three stages of three residual blocks. Gradients may be accumulated over smaller batches for memory efficiency, without stochastic sampling. The six main comparisons use seed 0 and the settings in Table 1.

For these six comparisons, coarse MF uses covariance rank cap three, Krylov depth two and four substeps per GD step (the solver's parameters are described in Appendix E). Rod uses auxiliary rank three and four substeps; Central uses four substeps, and Edge uses Algorithm 1 of Marion (2026) with $K = D = 4$. The initial MF covariance and Rod auxiliary extent have rank three for MLP, one/two for CNN, and three/two for ResNet in the one-/two-mode cases. No amplitude floor is added to MF or Rod; Edge retains its prescribed amplitude rule. Arithmetic precision and solver-refinement checks are detailed in Appendix E, including the full displayed horizons of the two-mode comparisons.

For reproducibility, the accompanying code provides Python scripts that prepare initial states and launch the simulations. Saved configuration files record the parameter choices. Initialization checkpoints are not included in the supplement; when rerunning a comparison, the same prepared state must be used for all flows and numerical resolutions. A separate record links each plotted curve to its simulation outputs and source-code version.

Experiment	Modes	Seeds	n	η	m	r	ℓ	K	Iterates	Figures
MLP	1	1	8000	.002	16	3	2	4	4000	1, 3
MLP	2	1	8000	.002	16	3	2	4	4000	3
CNN	1	1	8000	.002	4	3	2	4	1500	1, 14
CNN	2	1	8000	.002	4	3	2	4	2000	4, 18, 6
ResNet	1	1	3000	.004	8	3	2	4	411	1, 14
ResNet	2	1	3000	.004	8	3	2	4	200	4
MLP seeds	1	10	8000	.002	16	3	2	4	4000	19, 20
MLP seeds	2	9	8000	.002	16	3	2	4	4000	19, 20
MLP rates	1	1	8000	.0025	16	3	2	4	1000	21
MLP rates	1	1	8000	.004	16	3	2	4	1000	21
MLP rates	2	1	8000	.0025	16	3	2	4	1000	21
MLP rates	2	1	8000	.004	16	3	2	4	1000	21
MLP onset	1	1	8000	.002	16	3	2	4	4000	5
CNN sweep	1	1	8000	.002	4	1	–	$2, 4, 8, 16$	200	22

Table 1: Summary of the experiments. The Modes column counts near-EoS Hessian modes at initialization, independently of the retained covariance rank r. Here n is the sample count, m the averaging-window length, ℓ the MF Krylov depth, and K the MF/Rod substeps per GD step (Rod only in the sweep). Seeds counts completed trajectories. The seed cohorts attempt seeds $0, \ldots, 9$; two-mode seed 4 does not reach its initialization gate and is not replaced. All other rows use seed 0. Iterates gives the retained baseline comparison horizon per seed. Matched coarse-MF refinements for the two-mode comparisons are listed separately in Table 6. The MLP onset row gives the unfloored baseline; Figure 5 also shows its rank-four, eight-substep refinement, with MF Krylov depth three.

The one-mode comparisons start after the first instability, when $\frac{1}{2}H : \Sigma/E$ reaches approximately 10^{-3}, with H and E evaluated at the right endpoint of the stored GD window and Σ its empirical covariance. The two-mode comparisons continue the underlying GD trajectory until the second Hessian mode reaches EoS. Within each comparison, all flows share the stored initialization window. The separate MLP onset control from Appendix C.2 starts just below EoS, at $\eta\lambda_1 \simeq 1.999$, without the additional fluctuation-size condition; its baseline and refined MF/Rod runs have no amplitude floor. The CNN sweep reuses the one-mode CNN initialization.

GD Hessian eigenpairs are computed by warm-started LOBPCG every five steps for the MLP, every 100 for the one-mode CNN/ResNet, and every ten for their two-mode runs. Independent GD-window diagnostics are saved every five steps in all six cases, with both GD-center eigenvalues in every two-

mode comparison. The covariance-closure tests in Appendix D.4 compare the covariance measured from GD with predictions obtained by evolving the covariance equations along the GD-window means. The predicted covariance and Hessian coefficients are updated every GD step; the Hessian basis used for projection is refreshed every five steps. The CNN Rod sweep evaluates sharpness every 20 steps. Instantaneous GD sharpness and sharpness at the window mean are distinct observables; flow sharpness is evaluated at the flow center.

Figure 1 reuses the three one-mode runs, while Figures 6 and 18 reuse the two-mode CNN trajectory. Some figures combine flows computed in separate runs, each with a GD reference restarted from the same saved state. Each flow is compared with the GD reference computed in its own run, so we need to check reproducibility between GD runs. The repeated GD loss records are identical for MLP; for the long CNN, their maximum relative loss difference from the displayed GD references (absolute error normalized by the reference loss) is 0.00497%; the corresponding projected empirical GD covariance traces differ by at most 0.232%. For Figure 6, GD, MF and Central were computed together in one run, with the two-mode CNN settings in Table 1. Both flows use the same GD reference, and their mutual distance is computed directly from their saved parameter vectors.

D.2 Center error and oscillatory-state distance

The relative center error used in the main figures,

$$e_\theta(p) = \frac{\|\theta_p - \overline{\theta}_{\mathrm{GD}}\|_2}{\|\overline{\theta}_{\mathrm{GD}}\|_2}, \tag{129}$$

compares the predicted mean with the GD-window mean in the full parameter space. It does not compare their fluctuations. Moreover, its denominator is the total parameter norm, which can be much larger than the oscillation amplitude. A small value of e_θ therefore does not by itself establish accurate fluctuation dynamics.

We complement it with a comparison of means and covariances in the leading Hessian directions. At each sampled time, let U contain the leading one or two orthonormal Hessian eigenvectors at $\overline{\theta}_{\mathrm{GD}}$, according to the one- or two-mode comparison. For a flow p with mean θ_p and covariance Σ_p, set

$$m_p = U^T(\theta_p - \overline{\theta}_{\mathrm{GD}}), \qquad X_p = U^T \Sigma_p U.$$

For Rod, Σ_p is the covariance of its two endpoint positions, not its auxiliary covariance. GD is represented by $m_{\mathrm{GD}} = 0$ and $X_{\mathrm{GD}} = U^T \Sigma_{\mathrm{GD}} U$, where Σ_{GD} is the empirical covariance of the iterates in the same window. All states in a comparison use the same basis and GD-window mean.

To compare positive semidefinite covariance matrices, we use the Bures distance

$$d_{\mathrm{B}}^2(X, Y) = \operatorname{tr} X + \operatorname{tr} Y - 2 \operatorname{tr}\left[(X^{1/2} Y X^{1/2})^{1/2}\right],$$

where matrix square roots are positive semidefinite. Our normalized state discrepancy is

$$\mathcal{D}^2(p, q) = \frac{2\big(\|m_p - m_q\|^2 + d_{\mathrm{B}}^2(X_p, X_q)\big)}{\|m_p\|^2 + \operatorname{tr} X_p + \|m_q\|^2 + \operatorname{tr} X_q}. \tag{130}$$

For a positive denominator, this compares the mean and covariance differences with the average second moment of the two projected states about the GD-window mean, rather than with the full parameter norm. It does not detect errors orthogonal to the selected Hessian subspace, so it complements rather than replaces e_θ.

Figure 12 uses the same trajectories and sampled times as Figures 3 and 4. At the last two-mode MLP sample, MF and Rod have nearly identical center errors, about 0.00043%, but their discrepancies $\mathcal{D}$ from GD are 0.0263 and 0.00113, respectively. Thus agreement of the means alone does not characterize agreement of the mean–covariance states.

MF–Central comparison. For the shared 2000-step CNN run in Figure 6, the center-to-center curve shows $\|\theta_{\mathrm{MF}} - \theta_{\mathrm{Central}}\|_2 / \|\theta_{\mathrm{Central}}\|_2$ at the 41 saved checkpoints. To complement this comparison, we evaluate (130) in the same two GD-center Hessian directions for both flows. The discrepancy $\mathcal{D}(\mathrm{MF}, \mathrm{Central})$ decreases from 0.457 at step 195 to 0.00117 at step 1995, while their final discrepancies from GD are both about 0.697 (Figure 13).

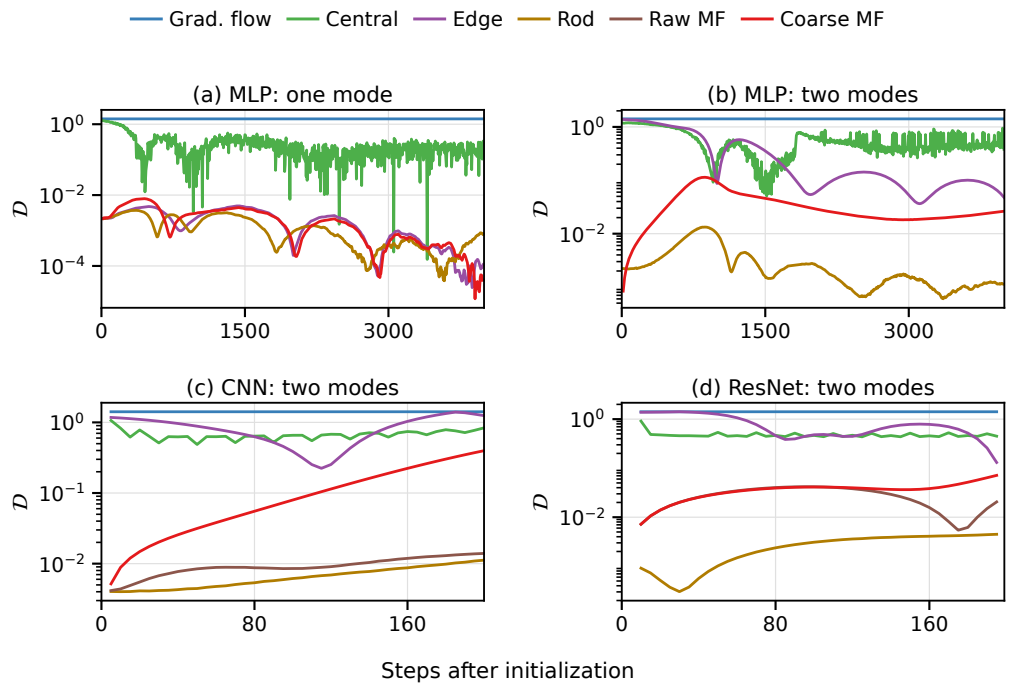


Figure 12: Normalized mean–covariance discrepancy $\mathcal{D}(p, \mathrm{GD})$ for the same runs as the main figures. Top: one- and two-mode MLP. Bottom: two-mode CNN and ResNet. Each comparison uses the leading one or two Hessian directions at the GD-window mean. The trajectories, horizons and sampled times are unchanged. Gradient flow has zero covariance, giving $\mathcal{D} = \sqrt{2}$.

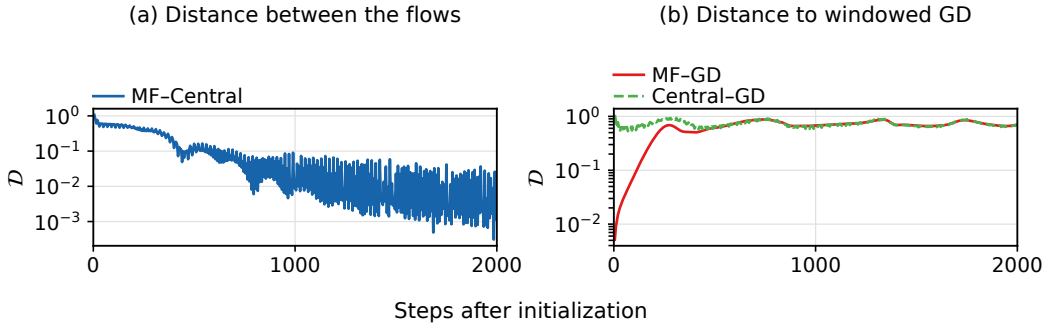


Figure 13: Mean–covariance comparison for the same 2000-step CNN run as Figure 6, using $\mathcal{D}$ from (130). Left: discrepancy between MF and Central flow. Right: discrepancy of each flow from windowed GD. All comparisons use the same two Hessian directions at the GD-window mean and all 399 recorded diagnostic samples.

The constant value for gradient flow. If $X_p = 0$, then $d_{\mathrm{B}}^2(X_p, X_{\mathrm{GD}}) = \operatorname{tr} X_{\mathrm{GD}}$. Since $m_{\mathrm{GD}} = 0$, (130) gives

$$\mathcal{D}^2(p, \mathrm{GD}) = \frac{2\big(\|m_p\|^2 + \operatorname{tr} X_{\mathrm{GD}}\big)}{\|m_p\|^2 + \operatorname{tr} X_{\mathrm{GD}}} = 2,$$

whenever the denominator is positive, in particular when $X_{\mathrm{GD}} \neq 0$. This applies to gradient flow, whose assigned covariance is zero.

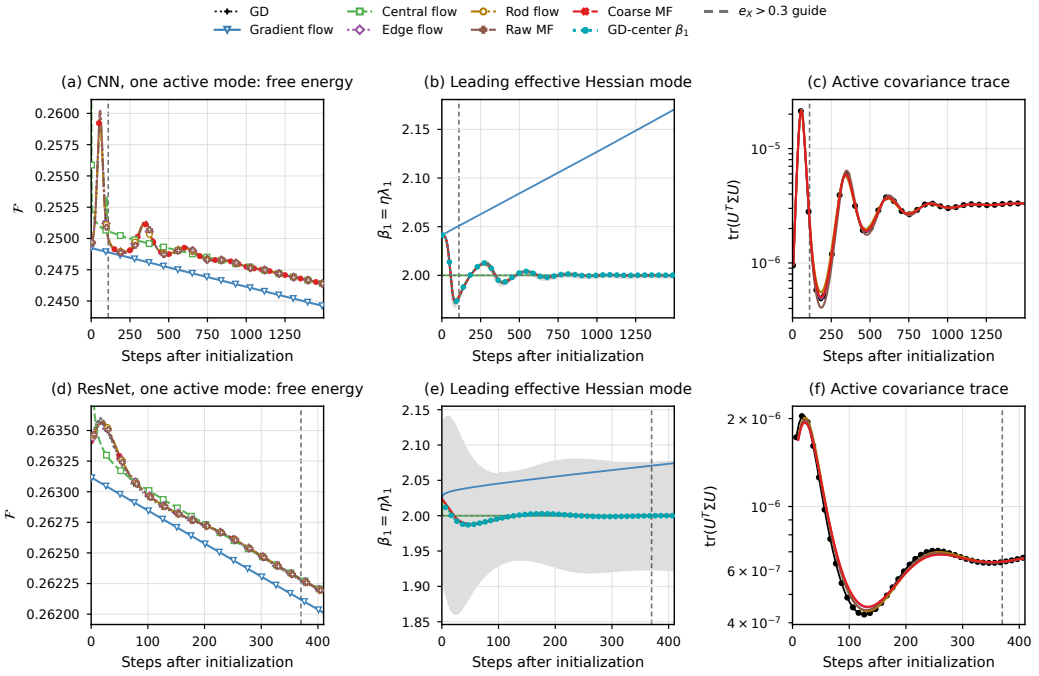


Figure 14: Additional one-mode architecture experiments. Top: CNN with $\eta = 0.002$. Bottom: ResNet with $\eta = 0.004$. Columns show free energy, the leading effective Hessian eigenvalue, and the variance $u_1^T \Sigma u_1$, where u_1 is the leading Hessian eigenvector at the GD-window mean. For the Rod curve, Σ is the covariance of its two endpoint positions. Instantaneous GD sharpness is drawn lightly; cyan markers show sharpness at the window center, and black markers show the empirical GD-window variance along u_1. The vertical guides mark the first sampled forced-covariance error above 0.3.

D.3 One-mode CNN and ResNet experiments

Figure 14 complements the two-mode comparisons with one-mode CNN and ResNet runs. The available trajectories contain 1500 and 411 iterates, respectively. Both use the baseline of Table 1.

In both one-mode runs, the coarse-MF free energy and sharpness follow the GD loss and the sharpness at the GD-window mean, respectively. Denoting u_1 the leading Hessian eigenvector at the GD-window mean, the predicted variance $u_1^T \Sigma_{\mathrm{MF}} u_1$ also follows GD accordingly.

D.4 Closure diagnostics for the six architecture/onset experiments

In this section, we numerically test our closure hypothesis, especially for the covariance evolution. Specifically, we wonder if the covariance evolution, when fed with the empirical GD Hessian, reproduces the GD fluctuations. Similarly, we run the mean equation when fed with the empirical GD covariance (projected on the low dimensional unstable modes).

We report these diagnostics for the six one- and two-mode MLP, CNN and ResNet runs in Table 1.

Closure diagnostics procedure. Let $\tilde{\theta}_t$ denote the GD iterates. The mean and covariance of the window of m consecutive iterates ending at step t are

$$\overline{\theta}_{\mathrm{GD},t} = \frac{1}{m} \sum_{j=0}^{m-1} \tilde{\theta}_{t-j},$$

$$\widehat{\Sigma}_t = \frac{1}{m} \sum_{j=0}^{m-1} (\tilde{\theta}_{t-j} - \overline{\theta}_{\mathrm{GD},t})(\tilde{\theta}_{t-j} - \overline{\theta}_{\mathrm{GD},t})^T.$$

These empirical quantities are measured from GD, not predicted by MF; the window lengths are given in Table 1. Denoting $U_t = (u_1(t), \dots, u_q(t))$ to have orthonormal columns approximating the eigenvectors associated with the q largest eigenvalues of $H(\overline{\theta}_{\mathrm{GD},t})$, we will use the empirical projected quantities

$$\widehat{X}_t = U_t^T \widehat{\Sigma}_t U_t, \qquad B_t = U_t^T H(\overline{\theta}_{\mathrm{GD},t}) U_t.$$

All six runs use the same diagnostic dimension $q = 3$. Note that the basis is constructed from GD alone. A Hessian eigensolve refreshes the tracked subspace every five steps; between solves, we diagonalize the current Hessian within that subspace.

Accumulated covariance errors. We evolve two covariance controls along the prescribed GD-window means which we now describe. The continuous control starts from $\widehat{X}_t$ and integrates

$$\frac{dX}{ds} = -\eta(B_t X + X B_t) + \eta^2 B_t X B_t, \qquad s \in [t, t+1],$$

exactly with B_t held fixed over the step. Before the next step, the covariance is transferred to the updated basis by $X \leftarrow (U_{t+1}^T U_t) X (U_t^T U_{t+1})$. The discrete control starts from a rank-truncated factor Q of the same initial empirical covariance and evolves

$$Q_{t+1} = (\mathrm{Id} - \eta H(\overline{\theta}_{\mathrm{GD},t})) Q_t$$

in the full parameter space. Its initial factor rank is three for MLP and one-mode ResNet, one for one-mode CNN, and two for two-mode CNN/ResNet. We compare its projection $U_t^T Q_t Q_t^T U_t$ with $\widehat{X}_t$. These controls differ in covariance representation, initialization and basis transfers, so their comparison does not isolate the effect of replacing discrete evolution by continuous time.

Neither control is reset to the empirical covariance at later windows so that the resulting errors accumulate over time. For either predicted covariance X_{pred}, Figure 15 shows

$$e_X = \frac{\|\widehat{X} - X_{\mathrm{pred}}\|_F}{\max\{\|\widehat{X}\|_F, \epsilon_X\}}.$$

To avoid division by nearly zero values, we set the fixed floor ϵ_X to one thousandth of the median covariance norm $\|\widehat{X}_t\|_F$, with a minimum of 10^{-12}.

The continuous crossings of $e_X = 0.3$ define the vertical guides in Figures 4 and 14.

Accumulated mean errors. We test the mean equation similarly, using the projected empirical covariance. To this aim, we evolve a forced mean θ^{f}, initialized from the GD-window mean. Over each GD interval, it solves

$$\frac{d\theta^{\mathrm{f}}}{ds} = -\eta \nabla E(\theta^{\mathrm{f}}) - \frac{\eta}{2} \nabla H(\theta^{\mathrm{f}}) : (U_t \widehat{X}_t U_t^T), \qquad s \in [t, t+1]. \tag{131}$$

Only the empirical covariance is held fixed over the interval. The covariance is refreshed at every GD step, but the forced mean is never reset, so that we measure accumulated mean errors. We use four Euler substeps per GD step, float64 accumulation of the mean updates, and float32 network and derivative evaluations.

The accumulated error is

$$e_\theta^{\mathrm{forced}}(t) = \frac{\|\theta^{\mathrm{f}}(t) - \overline{\theta}_{\mathrm{GD},t}\|_2}{\max\{\|\overline{\theta}_{\mathrm{GD},t}\|_2, 10^{-12}\}}. \tag{132}$$

Figure 16 reports this control for all six runs, using their original initialization caches: 4000 steps for each MLP run, $1500/2000$ steps for one-/two-mode CNN, and $411/200$ for ResNet. The sampled relative error stays below 1.02×10^{-5} throughout. In the two-mode CNN and ResNet runs, its maxima are 3.85×10^{-6} and 1.73×10^{-6}, respectively, even though the corresponding covariance controls develop large errors.

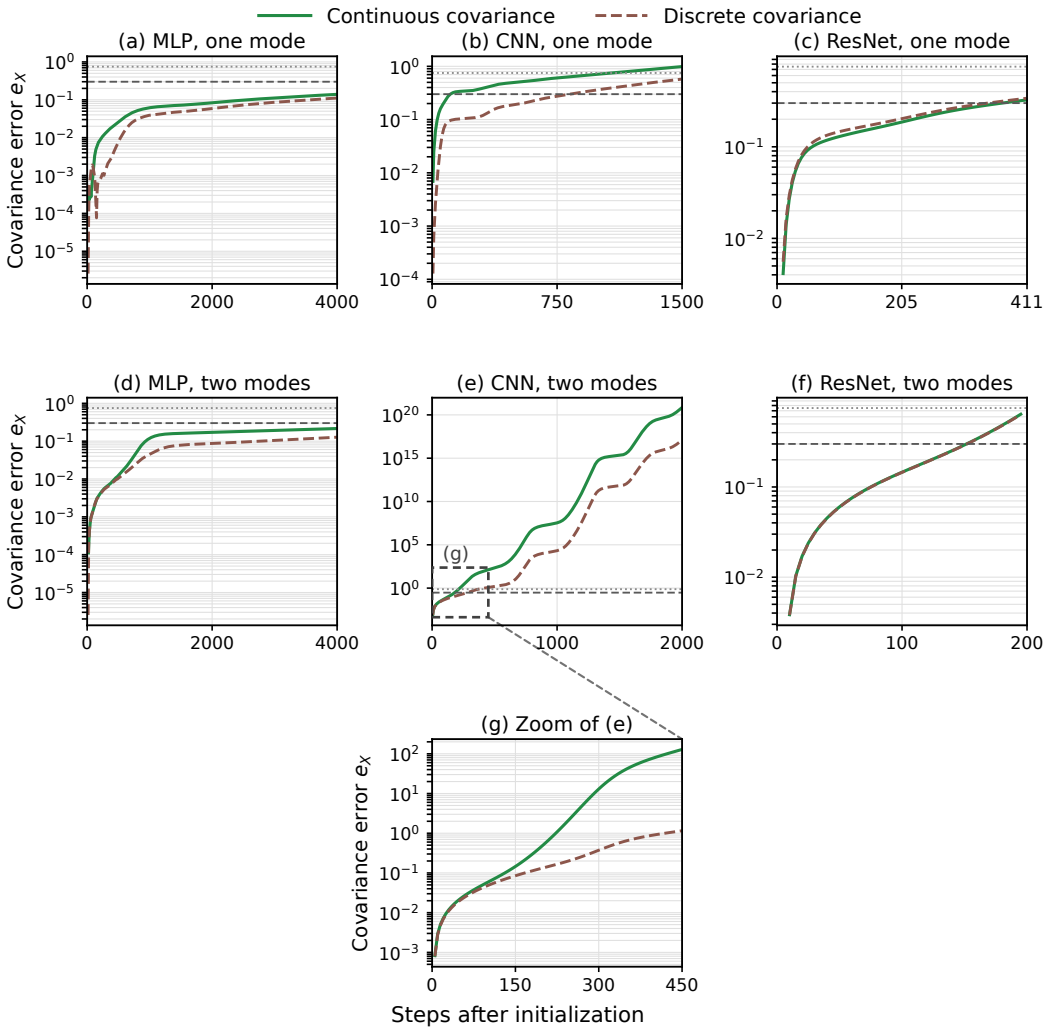


Figure 15: Accumulated covariance errors e_X along the prescribed GD-window means. Panels (a–c): one-mode initializations; (d–f): two-mode initializations. Columns: MLP, CNN and ResNet. Solid green: continuous MF covariance equation in the $q = 3$ diagnostic subspace. Dashed brown: discrete linearized covariance evolution in the full parameter space, projected only for comparison. Both controls are initialized once, then evolved without resetting to GD. Values are recorded every five steps; horizontal lines in (a–f) mark $e_X = 0.30, 0.75$.

Local mismatches of the mean and covariance equations. Beyond accumulated errors, the recorded diagnostics compare backward window increments with the current mean and covariance fields. Writing $\widehat{\theta}_t = \overline{\theta}_{\mathrm{GD},t}$, the MF predictions are

$$G_{\theta,t} = -\eta \nabla E(\widehat{\theta}_t) - \frac{\eta}{2} \nabla H(\widehat{\theta}_t) : (U_t \widehat{X}_t U_t^T),$$

$$G_{X,t} = -\eta (B_t \widehat{X}_t + \widehat{X}_t B_t) + \eta^2 B_t \widehat{X}_t B_t.$$

We compare them with the changes over one GD step,

$$\delta_t \widehat{\theta} = \widehat{\theta}_t - \widehat{\theta}_{t-1}, \qquad \delta_t \widehat{X} = U_t^T (\widehat{\Sigma}_t - \widehat{\Sigma}_{t-1}) U_t.$$

Both empirical covariances are projected onto the same current basis before subtraction, so basis rotations alone do not contribute to the residual. For $z \in \{\theta, X\}$, define

$$R_z(t) = \frac{\|\delta_t \widehat{z} - G_{z,t}\|}{\max\{\|\delta_t \widehat{z}\|, \|G_{z,t}\|, \epsilon_z\}}, \tag{133}$$

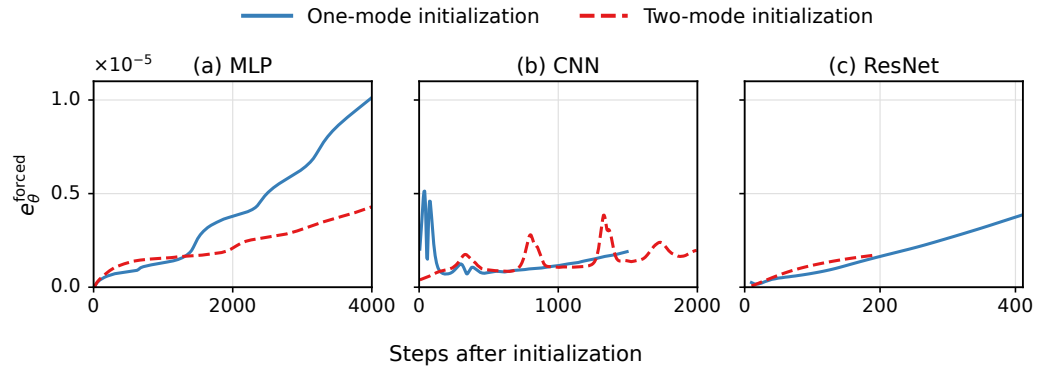


Figure 16: Accumulated mean errors (132), with projected empirical GD covariance forcing and $q = 3$. Curves end at their respective run horizons; the two-mode ResNet run covers 200 steps, compared with 411 for the one-mode run.

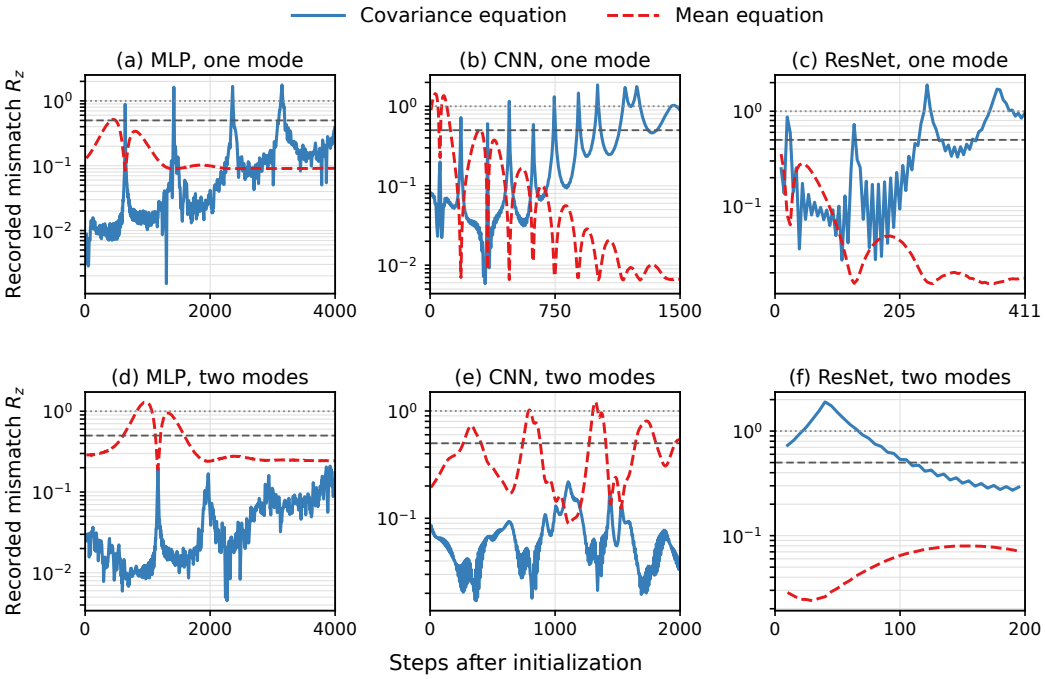


Figure 17: Local residual (133) for the same six runs and $q = 3$ diagnostic spaces, arranged as panels (a–f) of Figure 15. Solid blue: covariance equation. Dashed red: mean equation. Horizontal lines mark $R_z = 0.50, 1.00$.

using the Euclidean norm for the mean, the Frobenius norm for the covariance, and the same floor rule applied to $\max\{\|\delta_t \widehat{z}\|, \|G_{z,t}\|\}$. Figure 17 records these residuals every five steps, but each difference uses adjacent windows, not windows five steps apart. The mean residual is measured in the full parameter space, while the covariance is projected.

A more natural comparison would have been to record the defect with respect to the *forward* changes $\delta_t \widehat{\theta} = \widehat{\theta}_{t+1} - \widehat{\theta}_t$. However, we did not log such quantities, and expect this change to be negligible.

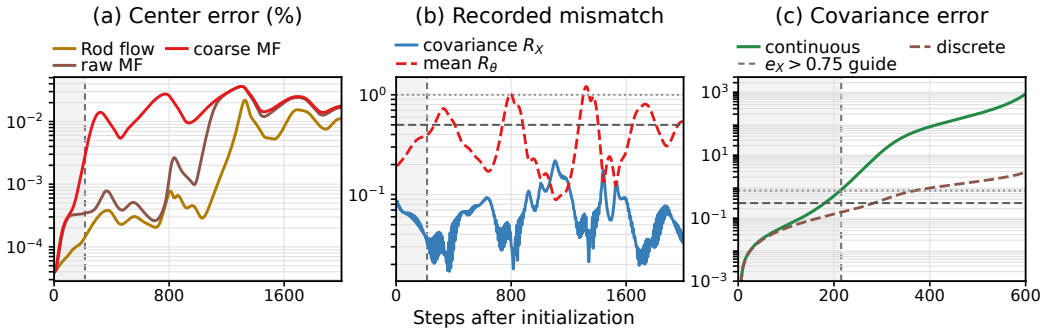


Figure 18: Two-mode CNN: (a) relative center errors of the autonomous flows with respect to gradient descent; (b) local residual R_X, R_θ; (c) accumulated covariance errors for the continuous and discrete controls, over the first 600 steps. Vertical guides mark the first sampled crossing of $e_X = 0.75$ by the continuous control; horizontal guides show the reference levels specified above.

D.5 MLP SENSITIVITY TO SEED AND LEARNING RATE

We vary the random seed at $\eta = 0.002$ and the learning rate at seed 0, using both one- and two-mode initializations. The MLP width, sample count and averaging window remain 256, 8000 and $m = 16$. For each seed and rate, GD is continued to the corresponding first- or second-eigenvalue threshold, following Table 1; all compared flows then use the same sampled GD window. The horizons are 4000 steps for the seed cohort and 1000 for the learning-rate comparisons. MF denotes the coarse dynamics, with baseline $(r, p, K) = (3, 2, 4)$. The other solver settings are unchanged. The available refinement checks are summarized in Table 6.

Seed cohort. We attempt seeds $0, \dots, 9$ at $\eta = 0.002$. All ten one-mode runs and nine two-mode runs complete. Two-mode seed 4 does not reach the second-eigenvalue threshold within 500000 additional GD steps and is not replaced. All completed runs are retained. The largest sampled next GD eigenvalue is 1.9422 in the one-mode cohort and 1.9264 in the two-mode cohort; these checks exclude an additional unstable mode only at the sampled times.

Figure 19 compares the mean loss/free energy and leading eigenvalues of GD, MF and Rod. Each seed is aligned at its own initialization. We compute the relative energy error $e_F = |F - \overline{E}_{\mathrm{GD}}|/|\overline{E}_{\mathrm{GD}}|$ against that seed's window-averaged GD loss before taking means or medians. Figure 20 shows the individual errors as well as both summaries, and Table 2 reports cohort medians for MF.

At the last common sample, the one-mode median energy errors for MF, raw MF and Rod are $0.00277\%, 0.0138\%, 0.0102\%$, respectively. The corresponding two-mode medians are $0.151\%, 0.0581\%, 0.0313\%$, whereas the means reach $3.27\%, 5.02\%, 0.242\%$.

Learning-rate sensitivity. Figure 21 compares seed 0 at $\eta = 0.0025, 0.004$, including Edge, Central and gradient flow. For one-mode initializations, the sampled MF energy error stays below 0.224% and 0.73%, respectively. For two-mode initializations, these bounds become 0.068% and 5.42%. The next GD-window eigenvalue remains below 2 at the sampled times: its maxima are below $1.803, 1.839$ for one mode and $1.978, 1.994$ for two modes. The deterioration at the larger rate therefore does not require an additional unstable mode in the recorded spectrum.

Dependence on the observable. Small center errors need not imply accurate fluctuations. For the two-mode $\eta = 0.004$ run, the final relative center error e_θ is 0.0228%, whereas the mean–covariance discrepancy $\mathcal{D}$ is 0.760 (Appendix D.2). The former divides by the full parameter norm; the latter compares means and covariances on the oscillation scale in the tracked Hessian directions.

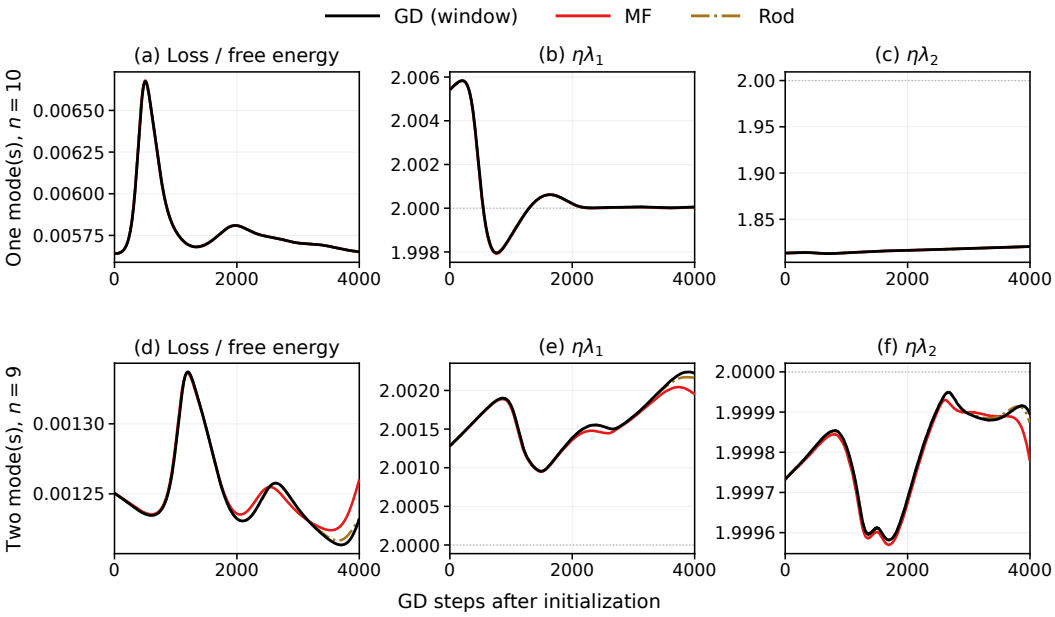


Figure 19: MLP seed cohort at $\eta = 0.002$: one-mode (10 seeds, top) and two-mode (9 seeds, bottom) initializations. Columns show averaged loss/free energy and the first two effective Hessian eigenvalues for GD, MF and Rod.

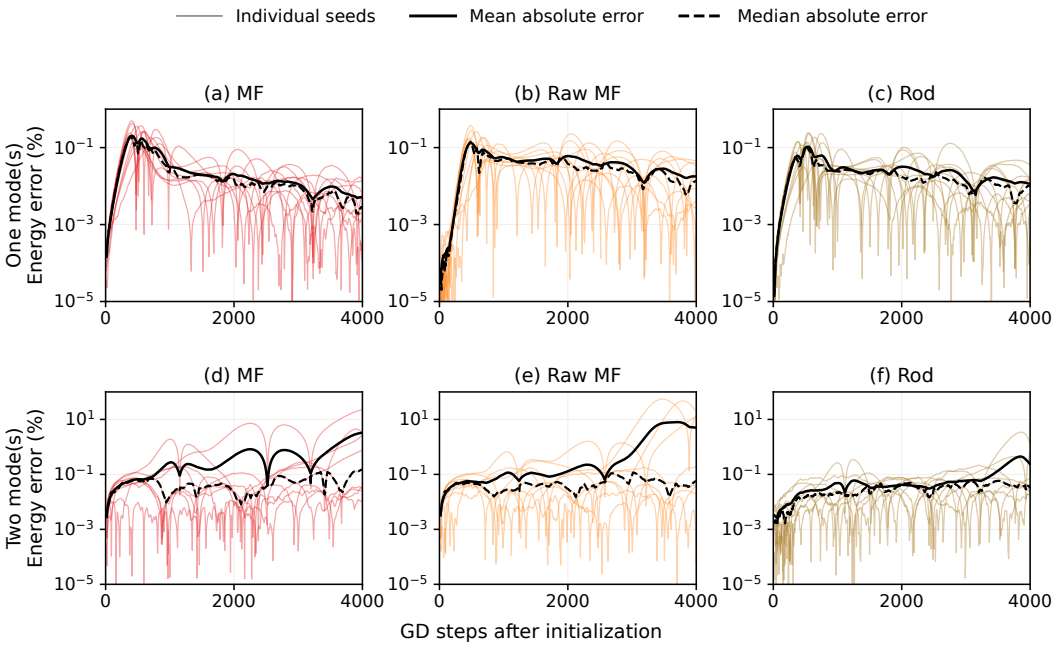


Figure 20: Per-seed absolute relative energy errors (thin colored lines), with their mean (solid black) and median (dashed black). Errors are computed against each seed's window-averaged GD loss before aggregation. Two-mode seed 4 has no comparison trajectory because its initialization gate was not reached.

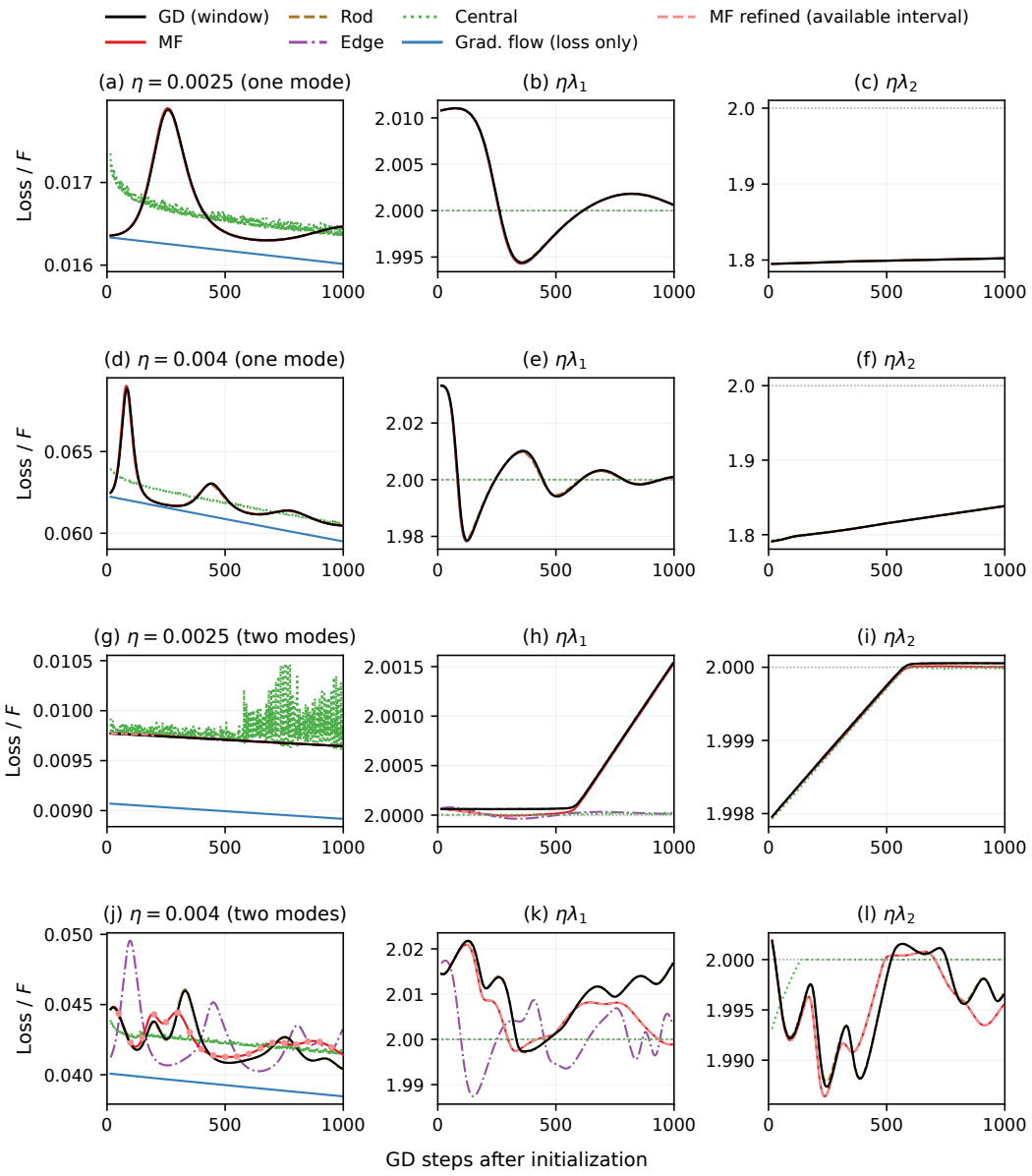


Figure 21: MLP learning-rate sensitivity, seed 0. One-mode initializations at $\eta = 0.0025, 0.004$ (first two rows), then two-mode initializations at the same rates (last two rows), each over 1000 steps. Left: GD loss (gray), its window average (black), and flow free energies. Middle/right: first/second effective Hessian eigenvalues at the flow centers and the GD-window center (black). For Edge, only the proxy $\eta\|Hu\|$ is available, shown in the middle column. Gradient flow appears only in the loss panels. Light dashed red shows the refinement over its available horizon: 200 steps in the third row and 1000 in the bottom row, where refined energies are available only at checkpoints and are shown as red markers. The first two rows have no matched full-horizon refinement.

Modes	η	nb. seeds	Steps	$\max e_F$ (%)	Final e_F (%)	Final e_θ (%)	Final $\mathcal{D}$
1	.002	10	4000	.249	.00277	.000511	.000337
2	.002	9	4000	.757	.151	.00182	.0263
1	.0025	1	1000	.224	.00431	.000133	.000845
1	.004	1	1000	.729	.00603	.000201	.00200
2	.0025	1	1000	.0680	.00672	.000274	.00190
2	.004	1	1000	5.41	2.51	.0228	.760

Table 2: Accuracy of the baseline MF trajectories: free-energy error e_F, relative center error e_θ (129), and mean–covariance discrepancy $\mathcal{D}$ (130). Modes refers to the initialization Hessian spectrum; nb. seeds is the number of completed trajectories. The first two rows give medians across seeds: each error is first computed within its own trajectory, and $\max e_F$ is maximized in time before taking the median. The remaining rows use seed 0 only.

D.6 Rod-flow solver resolution

The Rod solver evolves an auxiliary covariance matrix Σ, whose leading eigenpair determines the endpoints $\theta \pm \xi$. With K substeps per GD step, its forward-Euler update before rank truncation is

$$\Sigma^+ = \left(1 - \frac{2}{K}\right)\Sigma + \frac{\eta^2}{4K}\sum_{s=\pm 1} \nabla E(\theta + s\xi)\nabla E(\theta + s\xi)^T. \tag{134}$$

For the isolated decay $\dot{\Sigma} = -2\Sigma$, the exact multiplier over one substep is $e^{-2/K}$. At $K = 2$, the Euler multiplier instead vanishes; the endpoint-gradient term still depends on the previous state. A timestep check is therefore needed before interpreting a Rod–GD discrepancy as a limitation of the model.

One-mode timestep sweep. We use the one-mode CNN initialization of Table 1, fix the auxiliary rank to one and compare $K = 2, 4, 8, 16$ over 200 steps (Figure 22). This control uses single-precision arithmetic and changes only K. Taking the $K = 16$ trajectory as reference, the maximum relative differences in free energy and effective sharpness are 4.97% and 2.18% for $K = 2$, but 0.0454% and 0.0385% for $K = 4$. Each difference is divided by the reference value at the same recorded time. The largest relative covariance-eigenvalue difference for $K = 4$ is 2.99%. Thus the scalar observables are substantially less sensitive to further refinement beyond $K = 4$, although this rank-one control does not validate the higher-rank baseline by itself.

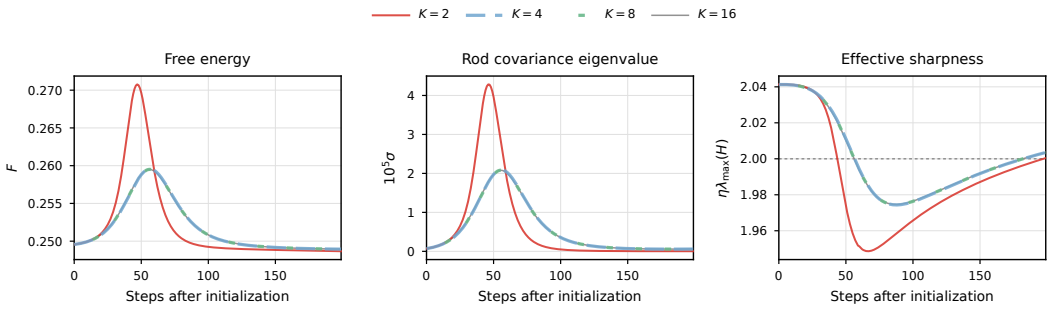


Figure 22: Rod timestep refinement from the one-mode CNN initialization, with auxiliary rank one and 200 steps after initialization. Left to right: free energy F, the retained covariance eigenvalue σ, and effective sharpness $\eta\lambda_1(H(\theta))$ at the Rod center. Here K is the number of Euler substeps per GD step. Light, interleaved curves distinguish the nearly overlapping $K = 4, 8, 16$ trajectories. The horizontal line in the right panel marks EoS.

Two-mode baseline refinement. We next compare auxiliary rank three with $K = 4$ against rank four with $K = 8$, using the two-mode CNN initialization over 200 steps. Both runs start from the same mean and rank-two auxiliary matrix, so increasing the rank cap does not change the initial state.

Both use double-precision mean accumulation and small eigensolves, with single-precision network evaluations, and neither applies an amplitude floor.

At step 200, the relative Frobenius difference between the endpoint covariances $\xi\xi^T$ is 0.140%, normalized by the larger covariance norm. The center difference is 0.00768% of the larger displacement from the common initial center, as in (147). The full auxiliary matrices also differ by 0.140%. These comparisons support the rank-three, $K = 4$ Rod baseline on this 200-step CNN interval; they do not establish its accuracy over the 2000-step extension or for other architectures.

D.7 Edge-flow timestep and direction-update controls

For Edge flow, increasing the number of substeps changes more than the integration step size. Algorithm 1 of Marion (2026) advances the center and amplitude with step $\rho = \eta/K$, but replaces the oscillation direction by $Hu/\|Hu\|$ at every substep, where $H = \nabla^2 E(\theta)$ is evaluated at the Edge center. This direction update has no factor ρ. Denoting its frequency per GD step by D, the published algorithm has $D = K$. We therefore test integration resolution and direction-update frequency separately.

All controls use the one- or two-mode CNN initializations of Table 1 and cover 200 GD steps. Within each case, the initial center, direction and amplitude are identical. Mean increments are accumulated in double precision; network and Hessian–vector evaluations remain single precision. The amplitude floor 10^{-6} is inactive throughout. At the five-step diagnostic times, the second GD-center effective eigenvalue lies in $[1.788, 1.815]$ for the one-mode case and $[1.9969, 2.0004]$ for the two-mode case.

Separating the two update frequencies. With Algorithm 1, changing $K = 4$ to 8 has little effect on the direction in the one-mode case, but produces a 61.74° angle in the two-mode case (Table 3). To isolate timestep sensitivity, we fix $D = 4$ and vary $K = 4, 8, 16$. At the start of each direction-update interval of length $\eta/4$, we compute $Hu/\|Hu\|$. We then advance the center and amplitude for $K/4$ substeps with the old direction held fixed, and apply the computed direction at the interval end. At $K = 4$, this reproduces Algorithm 1 exactly.

With $D = 4$ fixed, the covariance and center differences approximately halve when the pair $K = 4, 8$ is replaced by $K = 8, 16$, at every saved nonzero checkpoint. In contrast, changing only $D = 4$ to 8 at fixed $K = 8$ gives a 105.34% covariance difference and a 61.78° direction angle at step 200. The large two-mode sensitivity of Algorithm 1 is therefore predominantly due to the direction updates in this test, rather than the center and amplitude time discretization.

Integrating the direction equation. Fixing D still leaves discrete direction updates. As a separate check, we replace the power iteration used in Algorithm 1 (Marion, 2026) by a normalized Euler step for Marion's continuous direction equation:

$$\widetilde{u} = u + 2\rho\big(Hu - (u^T Hu)u\big), \qquad u^+ = \widetilde{u}/\|\widetilde{u}\|, \qquad \rho = \eta/K.$$

Only this update changes: the center and amplitude equations, including the sharpness estimate $\|Hu\|$ used for the amplitude, are retained. At step 200, the $K = 8, 16$ covariance difference is 0.0865%, and the $K = 16$ result differs from Algorithm 1 with $K = 4$ by 0.150% (Table 3).

These controls support retaining Algorithm 1 with $K = D = 4$ as the Edge comparator on the tested CNN interval. They neither establish long-horizon convergence nor validate the approximation $\lambda_1(H) \simeq \|Hu\|$. In particular, sensitivity to D alone does not demonstrate a failure of the continuous Edge equations.

CNN case	Comparison	Covariance gap (%)	Angle (°)
One-mode	Algorithm 1: $K = 4, 8$	2.49	0.0027
Two-mode	Algorithm 1: $K = 4, 8$	105.34	61.74
Two-mode	Algorithm 1: $K = 8, 16$	120.68	74.00
Two-mode	Fixed $D = 4$: $K = 4, 8$	0.181	0.0402
Two-mode	Fixed $D = 4$: $K = 8, 16$	0.0902	0.0201
Two-mode	Direction ODE: $K = 8, 16$	0.0865	0.0283
Two-mode	Direction ODE $K = 16$ vs. Algorithm 1 $K = 4$	0.150	0.0358

Table 3: Edge resolution controls at step 200, from identical initial states within each CNN case. For the two compared runs a, b, the covariance gap is $\|\Sigma_a - \Sigma_b\|_{\mathrm{F}} / \max\{\|\Sigma_a\|_{\mathrm{F}}, \|\Sigma_b\|_{\mathrm{F}}\}$, with $\Sigma_i = x_i^2 u_i u_i^T$; it can exceed 100%. The angle $\arccos|u_a^T u_b|$ compares normalized oscillation directions independently of their signs. These are differences between numerical solutions, not errors relative to GD.

E Numerical solver of the mean–fluctuation dynamics

For a network with d parameters, storing the covariance in (4) would require d^2 numbers. This is not feasible for the networks considered here. Our solver instead keeps a low-rank factor in parameter space and uses a small Krylov space to advance it. Neither the Hessian nor the covariance is ever formed as a dense $d \times d$ matrix.

E.1 Description of the solver

A covariance factor rather than a covariance matrix. We write

$$\Sigma \simeq QQ^T, \qquad Q = (q_1, \dots, q_r) \in \mathbb{R}^{d \times r}, \qquad r \ll d. \tag{135}$$

The columns of Q carry the oscillation directions and amplitudes. The initial factor is obtained from the truncated SVD of the centered GD window. Its initialization rank r_{init} is specified separately from the evolution rank cap r: in a resolution comparison, we keep the initial covariance fixed while increasing the number of directions the solver may retain. More precisely, if $\xi_1, \dots, \xi_m$ are the centered iterates and

$$\Xi = \frac{1}{\sqrt{m}}(\xi_1, \dots, \xi_m) \in \mathbb{R}^{d \times m}, \tag{136}$$

then $Q_0 Q_0^T$ is the rank-r_{init} spectral truncation of $\widehat{\Sigma}_0 = \Xi\Xi^T$. Computing the thin SVD of the m stored oscillations gives Q_0, and $\widehat{\Sigma}_0$ is never assembled.

The sign $\simeq$ in (135) is important because the full MF equation need not preserve rank (Remark 2.2), so that fixing r is an additional numerical approximation. What makes this approximation useful is that the empirical EoS covariance may concentrate along a few directions, although this does not establish convergence to the full MF equation. Indeed, if $P_\perp$ projects orthogonally to $\operatorname{Ran}\Sigma$, then

$$P_\perp \dot{\Sigma} P_\perp = \eta^2 P_\perp H \Sigma H P_\perp,$$

so covariance can grow outside its current range. The solver below captures new directions by Krylov enrichment, advances the projected covariance by a frozen-Hessian exponential, and then truncates to rank at most r. Numerical refinement therefore varies the rank cap r, enrichment depth p and number K of time substeps.

Krylov update of the covariance. Fix the mean during one covariance substep and write $H = H(\theta)$. Even when $\Sigma = QQ^T$ has small rank, the terms $H\Sigma$ and $H\Sigma H$ immediately probe the new directions HQ. We retain these directions without forming H by building

$$\mathcal{K}_p(Q, H) = \operatorname{span}\{Q, HQ, \dots, H^p Q\}. \tag{137}$$

Let the columns of $V \in \mathbb{R}^{d \times s}$ form an orthonormal basis of this space and define the small matrices

$$B = V^T H V, \qquad S = V^T Q Q^T V. \tag{138}$$

For fixed V, whose construction is described next, the projected covariance equation is

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\eta(BS + SB) + \eta^2 BSB. \tag{139}$$

This small equation can be integrated exactly while H is frozen, which we do in float64 precision (because this step adds numerical values of particularly different scales). Since B is symmetric, write

$$B = R\,\mathrm{diag}(\lambda_1, \dots, \lambda_s)R^T, \qquad Y = R^T S R,$$

where the columns of R are orthonormal eigenvectors of B and λ_a are its eigenvalues. In these coordinates, the covariance entries evolve independently. A substep of length Δt gives

$$Y^+_{ab} = \exp\big(\Delta t[-\eta(\lambda_a + \lambda_b) + \eta^2\lambda_a\lambda_b]\big)\, Y_{ab}. \tag{140}$$

Here and below, Y^+, S^+ and Q^+ denote the updated covariance variables after this substep. The rate of the variance along an eigenvector of the projected Hessian B is $(\eta\lambda_a)(\eta\lambda_a - 2)$, which makes explicit the change from decay to growth at the EoS threshold. We then return to the original Krylov coordinates:

$$S^+ = RY^+R^T.$$

To truncate the covariance rank, we perform a second eigendecomposition, $S^+ = W\Lambda W^T$. The columns of W are orthonormal eigenvectors of S^+, and Λ is diagonal with its covariance eigenvalues. We retain at most the r largest nonnegative eigenvalues in the diagonal matrix Λ_r, and collect their eigenvectors in W_r. The updated factor is

$$Q^+ = VW_r\Lambda_r^{1/2}. \tag{141}$$

Thus diagonalizing B serves to integrate the projected equation, whereas diagonalizing S^+ selects the covariance directions retained by the solver. The reconstructed covariance $Q^+(Q^+)^T$ is positive semidefinite.

In the implementation, covariance eigenvalues smaller than 10^{-10} times the largest one are also dropped when no amplitude floor is imposed. Thus r is an upper bound on the retained rank, not a prescribed rank at every step. Before rank truncation, $VS(t)V^T$ solves the full covariance equation with frozen H if H maps $\mathcal{K}_p$ into itself. Otherwise, directions outside $\mathcal{K}_p$ are discarded and the exponential update is exact only for the projected equation (139).

Construction of the orthonormal Krylov basis. There are two issues with the construction of the Krylov basis V. First, at EoS, we have roughly $\lambda_1(H) \approx 2/\eta$ and therefore H^2Q could be about 10^6 times larger than Q when $\eta = 0.002$. A naive SVD cutoff on $[Q, HQ, \dots, H^pQ]$ can thus discard directions carried by Q.

Replacing these blocks by $[Q, (\eta H)Q, \dots, (\eta H)^pQ]$ gives the same Krylov space and is a better idea. Nevertheless, the generated columns can still be almost linearly dependent, which makes basis extraction sensitive to roundoffs.

Our implementation constructs V incrementally by the *Arnoldi* algorithm applied to the scaled, frozen Hessian $q \mapsto \eta H(\theta)q$ (Saad, 2003; Gutknecht, 2007), which we describe now. First, the initial V_0 is obtained by QR factorization, i.e. V_0R_0 denotes the matrix formed by normalized columns of Q while $V_0^TV_0 = \mathrm{Id}$ and R_0 is upper triangular. Then, given $V = [V_0, \dots, V_j]$, we compute $Z = \eta H(\theta)V_j$ through Hessian–vector products, applying the scaled Hessian to each column of the newest block V_j. We remove its components along already existing basis vectors by applying

$$Z \leftarrow Z - V(V^TZ)$$

twice to reduce loss of orthogonality from roundoff. A thin SVD of this residual identifies the new directions: we retain its left singular vectors whose singular values exceed $10^{-5}\|\eta HV_j\|_{\mathrm{F}}$. This explicit cutoff uses the norm of the current HVP block *before* orthogonalization, not the norm of a concatenation of differently scaled powers. This cutoff selects only new columns to add to V; it does not remove any existing column, including those of V_0. If no singular value passes the cutoff, we stop enlarging the basis and use the current V for the covariance update. Otherwise, we apply the above projection twice more to the selected left singular vectors to remove roundoff components along V. A QR factorization then makes these new columns mutually orthonormal. They form the next block V_{j+1}, and we enlarge the basis by $V \leftarrow [V, V_{j+1}]$.

The depth p limits the number of enrichment rounds, so that the temporary basis has at most $(p+1)r$ columns, whereas at most r covariance directions are retained in (141). Normalizing the starting columns changes only the basis construction: $S = V^T Q Q^T V$ still uses the original Q, so no covariance amplitude is rescaled or floored.

Mean equation and free energy. The same factor makes the remaining high-dimensional contractions matrix-free. Namely,

$$\nabla H(\theta) : \Sigma \simeq \sum_{a=1}^{r} D^3 E(\theta)[q_a, q_a, \cdot\,], \tag{142}$$

$$F(\theta, \Sigma) \simeq E(\theta) + \frac{1}{2} \sum_{a=1}^{r} q_a^T H(\theta) q_a. \tag{143}$$

Automatic differentiation evaluates each third-order directional derivative and Hessian–vector product separately. Thus we form neither $D^3 E$, H, nor Σ. The coarse-MF mean correction uses the full-space factor Q, not its projection onto the tracked Hessian eigenvectors.

Energy observables in the figures. In the main-text architecture comparisons, coarse-MF energy curves use full-factor logs $F(\theta, QQ^T)$ for MLP and the projected logs $F(\theta, UXU^T)$ for CNN and ResNet, where U is the tracked Hessian basis and $X = U^T Q Q^T U$. This concerns only the displayed observable: the coarse-MF mean drift uses QQ^T in all cases. Furthermore, for these CNN and ResNet coarse-MF runs, the difference between the two energies is below 0.008% of $|F(\theta, QQ^T)|$ at the saved checkpoints. Across the architecture, seed, learning-rate and onset experiments, the relative difference is below 0.13% at the available logged or checkpointed states, so that the conclusions of the energy plots are unchanged.

Time stepping, controls, and cost. One unit of algorithmic time corresponds to one GD iteration. We divide it into K substeps and use a symmetric splitting : a half covariance update with the current Hessian, an explicit Euler mean update, then a half covariance update with the new Hessian. Despite the symmetric ordering, the Euler mean step makes the coupled scheme first order. For comparison, Strang splitting with exact subflows is second order (Strang, 1968).

For the neural-network flow runs, the mean increments are accumulated in a double-precision copy of θ, which is carried across substeps and saved in checkpoints. The gradient and HVP evaluations use a rounded single-precision copy. This prevents small increments from being lost each time they are added to a much larger parameter value, without converting all network computations to double precision. In particular, factors, Krylov bases and HVPs remain single precision.

The raw-MF control uses instead the full-space linearized-GD map

$$Q^+ = (\mathrm{Id} - \eta H) Q, \qquad \Sigma^+ = (\mathrm{Id} - \eta H) \Sigma (\mathrm{Id} - \eta H), \tag{144}$$

once per algorithmic step. Contrary to the coarse MF model, this map preserves rank when $\mathrm{Id} - \eta H$ is injective on $\mathrm{span}\{Q\}$.

The covariance state requires $O(dr)$ memory. At Krylov depth p, its temporary workspace has at most $(p+1)r$ ambient vectors and a dense matrix of size at most $(p+1)r$. The dominant operations are therefore a small number of Hessian–vector products and one third-order directional derivative per factor column; all dense linear algebra is low-dimensional.

Projection and truncation diagnostics. At each covariance substep we record

$$\rho_H = \frac{\|HV - VB\|_{\mathrm{F}}}{\|HV\|_{\mathrm{F}}}. \tag{145}$$

This measures how far HV lies outside the Krylov space, and vanishes when $H(\mathrm{Ran}\, V) \subseteq \mathrm{Ran}\, V$. We set $\rho_H = 0$ if $HV = 0$.

To measure the covariance discarded by truncation, let I_r index the retained eigenvalues of S^+ and put $\mu_a = \max\{\lambda_a(S^+), 0\}$. Before any amplitude floor, we record

$$d_{\mathrm{tr}} = \sum_{a \notin I_r} \mu_a. \tag{146}$$

This is the trace of the discarded positive covariance part. In the dense checks below, we report the sum of d_{tr} and the largest ρ_H over all covariance substeps of the run.

These diagnostics help assess the covariance representation. We complement them with comparisons of the full mean and covariance under refinement of (r, p, K) to assess numerical convergence.

E.2 Numerical validation

We first test the covariance solver against exact solutions for quadratic losses, then the coupled mean–covariance solver against a high-accuracy dense numerical reference for a nonlinear loss. These low-dimensional benchmarks use float64 throughout. Then, for neural networks, we instead compare MF trajectories at different resolutions of (r, p, K) from the same initial state. Agreement under these refinements supports the tested numerical settings: neither check should be confused with agreement between MF and GD, which tests the model approximation.

Dense reference checks. The first two checks use quadratic losses $E(w) = \frac{1}{2} w^T H w$ and initial mean $\theta_0 = 0$. Thus H is constant and the mean stays zero: specifying H, η and Q_0 suffices to determine the covariance evolution. For $\eta = 1$, $H = \mathrm{diag}(2, 1)$ and $Q_0 = (1, 1)^T$, the exact covariance is

$$\Sigma(t) = \begin{pmatrix} 1 & e^{-t} \\ e^{-t} & e^{-t} \end{pmatrix}, \qquad \det \Sigma(t) = e^{-t} - e^{-2t} > 0 \quad (t > 0).$$

At $t = 1$, the solver's covariance with $r = 2, p = 1$ agrees with this solution to relative Frobenius error below 10^{-11}. At $r = 1, p = 1$, that error is 0.182 for $K = 8$ and 0.186 for $K = 64$: time refinement does not cure rank bias. At $p = 0, r = 2, K = 8$, the error is 0.470 despite zero discarded trace. Here the missing enrichment is detected by $\max \rho_H = 0.316$, whereas this residual is below 10^{-14} in the $p = 1$ tests.

A second quadratic-loss test checks invariance under rescaling the loss and inversely rescaling the learning rate. In dimension six we set

$$E_\eta(w) = \frac{1}{2\eta} w^T A w, \qquad A = \mathrm{diag}(2.02, 1.97, 1.5, 1.1, 0.7, 0.3),$$

so that $H = A/\eta$ and $\eta H = A$ stays fixed. We use the same rank-two initial factor, $r = 6$, $p = 2$ and substep length $1/8$. Changing η from 1 to 0.002, while rescaling the loss as above, leaves the exact covariance update unchanged. At both scalings, the Arnoldi construction retains six directions and the relative covariance error is below 2×10^{-15}. Thus the solver reproduces the same exact update to floating-point accuracy.

The third check tests the coupled dynamics, where the Hessian changes along the mean trajectory. We compare the solver's covariance and mean-update with a dense numerical integration of the full coupled ODE in dimension six, with $\eta = 1$ and the nonlinear loss

$$E(w) = \tfrac{1}{2} w^T H_0 w + \frac{0.08}{6} \sum_i (Mw)_i^3 + \frac{0.02}{4} \sum_i w_i^4, \qquad H_0 = \mathrm{diag}(2.02, 1.8, 1.1, 0.7, 0.4, 0.1).$$

Here M is orthogonal and the initial covariance has rank two; the generator seed 17, matrix and initial state are saved with the benchmark. The dense reference uses DOP853 with relative tolerance 10^{-12} and absolute tolerance 10^{-14}; a second integration at $10^{-10}, 10^{-12}$ changes the final state by less than 3×10^{-13} in maximum norm. Table 4 separates time refinement from rank and Krylov-depth effects. These small-system checks validate the numerical building blocks.

The first three rows of Table 4 keep the full rank $r = 6$ and depth $p = 3$ fixed: doubling K approximately halves both errors, as expected for first-order time stepping. At fixed $K = 16$, reducing the rank or taking $p = 0$ produces a much larger covariance error. The diagnostics distinguish these two limitations: rank truncation discards covariance despite a negligible ρ_H, whereas $p = 0$ has zero discarded trace but a large ρ_H. However, $p = 1$ already gives a covariance error close to that at $p = 3$, despite $\max \rho_H = 0.117$. This residual measures subspace invariance without weighting directions by their covariance amplitudes, so that its value need not track the trajectory error. We therefore assess accuracy through the mean and covariance comparisons, using the local diagnostics to identify sources of error.

r	p	K	Relative mean error	Relative covariance error	$\sum d_{\mathrm{tr}}$	$\max \rho_H$
6	3	4	$2.14\,10^{-2}$	$1.38\,10^{-4}$	0	$< 2\,10^{-15}$
6	3	8	$1.06\,10^{-2}$	$6.79\,10^{-5}$	0	$< 2\,10^{-15}$
6	3	16	$5.25\,10^{-3}$	$3.37\,10^{-5}$	0	$< 2\,10^{-15}$
2	3	16	$5.14\,10^{-3}$	$4.68\,10^{-1}$	$4.61\,10^{-3}$	$< 2\,10^{-15}$
4	3	16	$5.25\,10^{-3}$	$1.34\,10^{-2}$	$1.02\,10^{-4}$	$< 2\,10^{-15}$
6	0	16	$5.42\,10^{-3}$	$6.58\,10^{-1}$	0	0.549
6	1	16	$5.25\,10^{-3}$	$4.37\,10^{-5}$	0	0.117

Table 4: Dense coupled MF check at $t = 2$. Errors use the dense-reference Euclidean and Frobenius norms as denominators. Discarded trace is summed over the whole run. The last column is the largest ρ_H over all covariance substeps. All rows start from exactly the same mean and covariance.

Neural-network refinement protocol. Within each comparison, the loss, dataset, seed, learning rate, GD initialization window and initial state (θ_0, Σ_0) are fixed. In particular, changing the solver rank r does not change the initial covariance rank. The runs use the same mixed-precision arithmetic described in Appendix E.1. We vary r, p and K separately in the short CNN controls, then compare the baseline $(3, 2, 4)$ with the combined refinement $(4, 3, 8)$ over the common intervals listed in Table 6.

At every common checkpoint, write $(\theta^{\mathrm{b}}, \Sigma^{\mathrm{b}})$ and $(\theta^{\mathrm{f}}, \Sigma^{\mathrm{f}})$ for the two MF states (baseline and refined), with $\Sigma^a = Q^a (Q^a)^T$ for $a \in \{\mathrm{b}, \mathrm{f}\}$. We measure

$$\delta_\theta(t) = \frac{\|\theta^{\mathrm{b}}(t) - \theta^{\mathrm{f}}(t)\|_2}{\max\{\|\theta^{\mathrm{b}}(t) - \theta_0\|_2, \|\theta^{\mathrm{f}}(t) - \theta_0\|_2\}}, \tag{147}$$

$$\delta_\Sigma(t) = \frac{\|\Sigma^{\mathrm{b}}(t) - \Sigma^{\mathrm{f}}(t)\|_{\mathrm{F}}}{\max\{\|\Sigma^{\mathrm{b}}(t)\|_{\mathrm{F}}, \|\Sigma^{\mathrm{f}}(t)\|_{\mathrm{F}}\}}. \tag{148}$$

We set $\delta_\theta = 0$ if both means equal their common initial value θ_0, and $\delta_\Sigma = 0$ if both covariances vanish. The mean gap is therefore relative to the displacement since initialization. The covariance comparison uses the full factors Q, evaluated in a shared QR basis without assembling $d \times d$ matrices so that we compare QQ^T, not the factors themselves or only their projections onto near-EoS Hessian modes.

The tables report $100\delta_\theta$ and $100\delta_\Sigma$, either at the last checkpoint or maximized over the common checkpoints, as specified in each caption.

Separate rank, depth and timestep controls. We vary each resolution parameter separately in the two-mode CNN over 200 steps, with width 32, $n = 8000$, $\eta = 0.002$, $m = 4$, seed 0 and the same rank-two initial covariance. The first four comparisons in Table 5 change only the indicated parameter from $(r, p, K) = (3, 2, 4)$. Rank and depth have little effect; the mean gaps approximately halve when K doubles.

Comparison	Covariance gap (%)	Mean gap (%)
$r = 3$ versus 4	0.0195	0.00139
$p = 2$ versus 3	0.0146	0.00134
$K = 2$ versus 4	0.0602	0.00738
$K = 4$ versus 8	0.0255	0.00366
$(3, 2, 4)$ versus $(4, 3, 8)$	0.0302	0.00357

Table 5: Separate CNN resolution controls. Mean and covariance gaps use (147)–(148) at checkpoint 200.

Matched refinement checks. Table 6 collects the combined $(3, 2, 4)$ versus $(4, 3, 8)$ checks: the full reference CNN, MLP and ResNet horizons, plus the seed-1 check and the two-mode learning-rate comparisons of Figure 21. The refinement covers 1000 steps at $\eta = 0.004$ and only the first 200 at $\eta = 0.0025$. The initial covariance rank is fixed within each pair.

Network	η	Seed	Matched steps	r_{init}	Max. mean gap (%)	Max. covariance gap (%)
CNN	.002	0	2000	2	.00752	.0540
MLP	.002	0	4000	3	.00728	.111
ResNet	.004	0	200	2	.0245	.126
MLP	.002	1	4000	3	.00583	.113
MLP	.004	0	1000	3	.00911	.0696
MLP	.0025	0	200	3	.00391	.0273

Table 6: Coarse-MF refinement checks for selected two-mode experiments. Entries are maxima of $100\delta_\theta$ and $100\delta_\Sigma$ over checkpoints spaced 50 steps apart. The baseline and refinement share their initial state. Matched steps gives the common comparison horizon, which may be shorter than the baseline horizon in Table 1.

Across the matched pairs, maximum relative mean and covariance gaps remain below 0.025% and 0.127%. These checks support the listed coarse-MF resolutions and horizons; they are not exact-solution error bounds and do not validate other flows or initializations. The accompanying code and metadata retain the launchers, configurations and full checkpoint comparisons.