Towards Understanding Momentum Acceleration in River-Valley Loss Landscape

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

Momentum is a critical and ubiquitous component of modern optimizers, while the role of momentum remains unclear beyond restricted settings, especially in optimization for large-scale neural networks. Recent studies suggest that the highly non-convex loss landscape for large language models exhibits certain "river-valley" structure: a low-loss manifold (the river) bordered by sharp, high-loss directions (the valley), where the essential optimization progress is determined primarily by the progress made along river in the long run. Motivated by such a structure, in this work, we investigate the role of heavy-ball momentum in such an emerging setting. Specifically, we analyze gradient descent with heavy-ball momentum and show that compared to vanilla gradient descent, momentum can accelerate the progress along the river by enabling use of a substantially larger learning rate. In fact, momentum acts as a stabilizer in the presence of oscillations caused by an aggressive choice of learning rate, which the vanilla gradient descent cannot tolerate. We validate the insights with experiments on synthetic functions and language model training, offering practical guidance for tuning learning rate and momentum parameters.

1. Introduction

Momentum is a foundational component of modern optimizers and is widely used in training large neural networks. In practice, gradient descent variants with heavy-ball momentum [57], e.g., Adam [32] and AdamW [42], are standard choices of optimizer, and empirical studies underscore its critical role. Despite this ubiquity, however, the theoretical conditions under which momentum actually accelerates learning remain poorly understood. Classical theory can only guarantee the acceleration of heavy-ball momentum within restricted settings with quadratic-like properties via improving the exponent in linear convergence [31, 57]. However, the loss landscapes of neural networks are typically highly non-convex and linear convergence is rarely observed in practice. As a result, the conventional intuition – that momentum accelerates local linear convergence – may not fully explain its success in deep learning, calling for rethinking of the role of momentum in such complex loss landscapes. In particular, theoretically answering this question requires understanding momentum in the context of a tractable but still realistic and meaningful loss landscape beyond convex settings.

A recent work on optimization in language model pretraining has identified and proposed an intriguing "river-valley" structure of loss landscape [69]. From this viewpoint, the training loss surface can be depicted by a deep valley with a narrow, low-loss manifold at the bottom ("river") surrounded by orthogonal steep directions ("mountains"). Empirically, this picture of loss landscape is evidenced by the loss curves of the Warmup-Stable-Decay (WSD) learning rate scheduler [26] and the cosine learning rate scheduler [41]: WSD first uses a constant large learning rate, causing

faster descent along the river than cosine scheduler but with higher loss due to oscillations across valley walls. In the end, a sharp decay of the learning rate can suppress the oscillations, revealing the genuine optimization progress along the river by WSD and resulting in lower final loss than cosine scheduler. Wen et al. [69] also probe LLM pretraining loss landscape to showcase the existence of such a landscape. We illustrate such structure through a synthetic loss function in Figure 1. The black curve represents the river manifold. The red and the purple points represent GD with and w/o momentum respectively (the lighter the points, the larger the step).

Within such a landscape of "a deep valley with a river at its bottom", the long-term optimization progress is essentially governed by the motion along the river, making the main challenge how to move quickly along the low-loss river manifold while damping out the instability in the high-curvature valley directions. Motivated by the lack of theoretical underpinning of momentum in complex loss landscapes, in this paper, we theoretically analyze the dynamics of heavy-ball momentum gradient descent under the emerging setting of river-valley loss landscape. Within such a landscape, we ask the following fundamental questions:

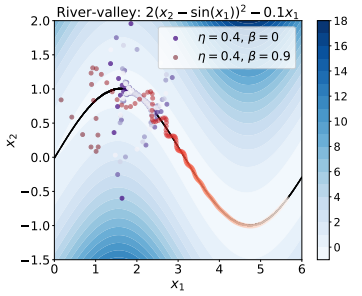

Figure 1: river-valley example.

What is the role of momentum? How does the momentum parameter interplay with learning rate? How do they jointly accelerate optimization along the river?

1.1. Our Contributions

Theoretically answering these questions requires a new analysis for GD with momentum in such a *non-convex* loss landscape, especially, how momentum shapes the interactions between the benign progress in the river and the oscillating updates in the valley directions. We consider GD iteration with heavy-ball momentum: given loss function L and initial parameter and momentum (w_0, m_0),

$$m_{k+1} = \beta \cdot m_k + (1-\beta) \cdot \nabla L(w_k), \qquad w_{k+1} = w_k - \eta \cdot m_{k+1}. \tag{1.1}$$

Mechanism of momentum acceleration in river-valley. We show that momentum enlarges the largest tolerable learning rate by the leading factor $(1+\beta)/(1-\beta)$, i.e., $\eta_{\max}^{\text{GD-M}} \approx (1+\beta)/(1-\beta)\eta_{\max}^{\text{GD}}$, even when the flat and sharp eigenspaces rotate along the non-convex river-valley landscape. Thus, momentum accelerates progress along the river mainly by stabilizing aggressive learning rates that vanilla GD cannot tolerate. We also point out that this acceleration is primarily achieved through stabilization rather than by substantially changing the river-direction speed.

Technical contributions for analyzing momentum. We prove that (1.1) can track the river without exploding under this enlarged learning-rate range; see Lemma B.1. The key technique is an induction argument that jointly controls the gradient and the momentum projections onto the river and valley directions as the eigenspaces rotate. As a by-product, we also improve the largest tolerable learning rate for vanilla GD in river-valley landscapes [69]; see Remark A.2.

Experiment on synthesis loss & language model training. We verify the theory on a synthetic river-valley loss and in GPT-Neo pretraining on TinyStories. Across SGD-Momentum and Adam, the largest stable learning rate consistently increases with the momentum parameter and induces lower training loss within same training iterations. Meanwhile, same (tolerable) learning rates with different momentum parameters lead to similar training loss, validating the theoretical prediction that the main acceleration of momentum comes from its stabilization effect under a larger learning rate.

2. Preliminaries

Notations. We denote the eigenvalues of $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ as $\{\lambda_k(\boldsymbol{A})\}_{k=1}^d$ with $\lambda_1(\boldsymbol{A}) \geq \cdots \geq \lambda_d(\boldsymbol{A})$. We denote $\{v_k(\boldsymbol{A})\}_{k=1}^d$ as the normalized eigenvectors of $\boldsymbol{A}$ with $v_k(\boldsymbol{A})$ corresponding to $\lambda_k(\boldsymbol{A})$. We denote $\|\boldsymbol{A}\|_{\mathrm{Op}}$ as the operator norm of $\boldsymbol{A}$. For the third derivative tensor, we define $\|\nabla^3 L(w)\|_{\mathrm{Op}} := \sup_{\|u\|_2=\|v\|_2=\|z\|_2=1} |\nabla^3 L(w)[u,v,z]|$, where $\nabla^3 L(w)[u,v,z]$ denotes the corresponding trilinear form. We denote $\lfloor t \rfloor$ as the largest integer smaller or equal to t. We denote $\mathcal{B}(w,r)$ as the Euclidean ball of radius r centered at w. For $\beta \in (0,1)$, we use $\mathcal{O}_\beta(\cdot)$ to hide factors polynomial in $1/(1-\beta)$.

2.1. River-Valley Loss Landscape: Setups and Key Assumptions

We use $w \in \mathbb{R}^d$ to denote the parameters. We denote $L : \mathbb{R}^d \to \mathbb{R}$ as the loss function and assume L is smooth. The concept of the river-valley loss landscape is as follows.

Assumption 2.1 (Existence of the manifold of river) *There exists a one-dimensional sub-manifold $\mathcal{M}$ of $\mathbb{R}^d$ s.t. for $\forall w \in \mathcal{M}$, $\nabla L(w)$ aligns with $v_d(\nabla^2 L(w))$, i.e., $\nabla L(w)/\|\nabla L(w)\|_2 = v_d(\nabla^2 L(w))$.*

Along the river, the gradient $\nabla L(w)$ aligns with the flattest direction at w, i.e., $v_d(\nabla^2 L(w))$, which we refer to as the river direction or the *flat direction*. All the other directions orthogonal to the river are dubbed as mountain directions or *sharp directions*, corresponding to the steep valley. Intuitively, the river captures the path with the lowest loss locally, while the hill components reflect the additional loss incurred by the deviations from the river.

Conceptually, the optimization process goes as follows: The parameter w starts from a neighborhood $\mathcal{U}$ of the river $\mathcal{M}$, and approaches the river $\mathcal{M}$ by decreasing the loss. Then the optimization proceeds by flowing downstream for further optimization progress, which in the long term, is determined primarily by the progress along the river. To rigorously characterize this process, we introduce technical assumptions following the prior work Wen et al. [69].

Assumption 2.2 (Existence and regularity of an open neighborhood around river) *There exist an open set $\mathcal{U}$ with $\mathcal{M} \subset \mathcal{U}$ and constants $\gamma_{\mathrm{flat}}, \gamma, \gamma_{\max}, g_{\max}, \kappa \geq 0$, s.t.:*

1. ***Diameter of** $\mathcal{U}$: For any $w \in \mathcal{M}$, $\mathcal{B}(w, 6g_{\max}/\gamma) \subset \mathcal{U}$, where $g_{\max}$ satisfies $\|\nabla L(w)\|_2 \leq g_{\max}$.*
2. ***Eigen-gap**: the constants $\gamma_{\mathrm{flat}}, \gamma, \gamma_{\max}$, κ satisfy that $\kappa\gamma \leq \gamma_{\mathrm{flat}} \leq \kappa^{1/2}\gamma_{\max}$, $\gamma \geq \kappa^{1/32}\gamma_{\max}$, and for any $w \in \mathcal{U}$, $\lambda_1(\nabla^2 L(w)) \leq \gamma_{\max}$, $\lambda_{d-1}(\nabla^2 L(w)) > \gamma + 4\gamma_{\mathrm{flat}}$, $|\lambda_d(\nabla^2 L(w))| < \gamma_{\mathrm{flat}}$.*
3. ***Slow spinning of flat direction and Hessian**: For any $w \in \mathcal{U}$, $\|\nabla(v_d(\nabla^2 L(w)))\|_{\mathrm{Op}} \leq \kappa\gamma/2g_{\max}$ and $\|\nabla^3 L(w)\|_{\mathrm{Op}} \leq \kappa\gamma^2/2g_{\max}$.*
4. ***Uniqueness of the river** $\mathcal{M}$: For any $w \in \mathcal{U} \setminus \mathcal{M}$, $\nabla L(w)$ does not align with $v_d(\nabla^2 L(w))$.*

The first one simply requires that the size of the neighborhood is not too small. The last one ensures the uniqueness of river in the neighborhood for the simplicity of analysis. The second and the third assumptions are key to the river-valley model, which we explain in detail. The *Eigen-gap* assumption posts conditions on the relative magnitude of the Hessian eigenvalues. It assumes a gap of γ between the sharpness of the river direction and the other sharp directions, which formalizes the picture of a flat river and steep mountains. Finally, the *Slow spinning* assumption limits the spinning speed of the river by imposing third-order conditions on the loss. Note that under the *Eigen-gap* assumption, the bound on $\|\nabla^3 L(w)\|_{\mathrm{Op}}$ implies the bound on the spinning speed of the flat direction. Throughout our analysis, both $\kappa \ll 1$ and $\gamma_{\mathrm{flat}}\gamma_{\max}^{-1} \ll 1$ are treated as *very small dimensionless constants*, guaranteeing the flatness as well as slow spinning of the river.

Definition 2.3 (River-valley landscape) *If the loss function L satisfies Assumption 2.1 and Assumption 2.2, we call the corresponding loss landscape a river-valley landscape.*

Reference flow. Next, to formalize how close the iterates are to the river and how fast they flow downstream, we introduce the concept of the *reference flow*. Specifically, the reference flow is a Riemannian gradient flow on the river, representing the dynamics of gradient flow on the loss when constrained to the river. We denote the projection matrix onto the tangent space of the manifold $\mathcal{M}$ at $w \in \mathcal{M}$ as $\boldsymbol{P}_{\mathcal{M}}(w)$. Given an initial point $x \in \mathcal{M}$, the Riemannian gradient flow is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = -\boldsymbol{P}_{\mathcal{M}}(x(t))\nabla L(x(t)), \quad x(0) = x. \tag{2.1}$$

Throughout the paper, $x(t)$ always refers to points on the river $\mathcal{M}$ or the reference flow, and t is the continuous time index of the reference flow (2.1).

Examples of river-valley loss landscape. We provide concrete examples to illustrate the landscape. The simplest one is a convex function $L(x_1, x_2) = \gamma \cdot x_1^2/2 - x_2$, where the river is just $x_1 = 0$ and it does not spin, i.e., $\kappa = 0$. Another example is a non-convex function $L(x_1, x_2) = (x_2 - \sin(x_1))^2 - \alpha \cdot x_1$ for some $\alpha \geq 0$. See Figure 1. The optimization progress for it is essentially determined by how fast the variable x_1 grows by tracking the river.

2.2. Gradient Descent with Heavy-Ball Momentum

Our theoretical analysis concerns vanilla GD and GD with heavy-ball momentum. Starting from some $w_0 \in \mathbb{R}^d$, the vanilla GD iterates as following,

$$w_{k+1} = w_k - \eta \cdot \nabla L(w_k), \tag{GD}$$

where η denotes the learning rate. Furthermore, for GD with heavy-ball momentum [57], when initialized at some $(w_0, m_0) \in \mathbb{R}^d \times \mathbb{R}^d$, it iterates as

$$m_{k+1} = \beta \cdot m_k + (1-\beta) \cdot \nabla L(w_k), \qquad w_{k+1} = w_k - \eta \cdot m_{k+1}. \tag{GD-M}$$

where $\beta \in [0, 1)$ is the momentum parameter. In particular, when $\beta \equiv 0$, GD-M reduces to vanilla GD. There is another version of GDM where the $1-\beta$ is absent in the update of momentum, but it is equivalent to the above formulation up to simple reparametrization.

3. Analysis of GD with Momentum in River-Valley

3.1. Main Theory: Momentum Acceleration in River-Valley Landscape

We investigate (GD-M) near the river by assuming that the initialization w_0 is already on the river $w_0 \in \mathcal{M}$. Our main theorem predicts how close and fast (GD-M) can track the river manifold. For ease of presentation, we state an informal version of the main theorem as follows.

Theorem 3.1 (GD-momentum in river-valley (informal version of Theorem G.1)) *Suppose that Assumptions 2.1 and 2.2 hold. For a momentum parameter $\beta \in (0, 1-\varepsilon)$, let $\eta < \eta_{\max}^{\text{GD-M}}$, where*

$$\eta_{\max}^{\text{GD-M}} \approx \left(\frac{1+\beta}{1-\beta} - \mathcal{O}_\beta\left(\kappa + \gamma_{\text{flat}}\gamma_{\max}^{-1}\right)\right) \cdot \frac{2}{\gamma_{\max}}. \tag{3.1}$$

Then for (GD-M) *with initialization* $w_0 \in \mathcal{M}$ *on the river and initial momentum* $m_0 = 0$*, there is a time index* T_0 *such that for any step* $k \geq \log(\beta\eta\gamma_{\text{flat}}/(1-\beta))/\log\beta$*, there exists another* $T(k)$ *such that the following two things hold:*

1. *GD-M stays close to the river:* $\|x(T_0 + T(k)) - w_k\|_2 = \mathcal{O}_\beta(\kappa^{1/2}/\gamma)$*;*
2. *The speed on the river is proportional to the learning rate:* $|T(k) - \eta \cdot k| = \mathcal{O}_\beta((\kappa + \eta\gamma_{\text{flat}}) \cdot \eta)$.

Here $\mathcal{O}_\beta(\cdot)$ *hides multiplicative factors depending on* β*.*

Here we omit details of the error terms to convey the essential message in a clean way. When $\beta = 0$, (GD-M) reduces to (GD), and Theorem 3.1 recovers Theorem A.1. We refer to Theorem G.1 for the complete statement of the above theorem. We outline the proof sketch of Theorem 3.1 in Section B, with all details provided in Appendix G. In the remainder of this section, we discuss the interpretation of the main theorem, as well as its predictions.

3.2. Theoretical Predictions by Theorem 3.1

For a constant $\beta \in (0, 1)$, e.g. $\beta = 0.9$ as the default choice used in practice, $\mathcal{O}_\beta(\cdot)$ is a constant. Then since κ and γ_{flat} are small, the error terms in Theorem 3.1 are negligible.

Heavy-ball momentum: acceleration via stabilization. As discussed in Section A, a key factor that determines the optimization progress along the river is the largest learning rate that the algorithm can tolerate. For (GD), recall from (A.1) that $\eta_{\max}^{\text{GD}} \approx 2/\gamma_{\max}$. In contrast, for (GD-M), in the regime of $\gamma_{\text{flat}} \ll \gamma_{\max}$ (very flat river) and $\kappa \ll 1$ (very slowly spinning river), it follows from (3.1) that the largest learning rate $\eta_{\max}^{\text{GD-M}}$ is larger than $\eta_{\max}^{\text{GD}}$ by a multiplicative factor of $(1+\beta)/(1-\beta)$. That is,

$$\eta_{\max}^{\text{GD-M}} \approx \frac{1+\beta}{1-\beta} \cdot \eta_{\max}^{\text{GD}}. \tag{3.2}$$

A larger learning rate η in-turn results in a faster speed along the river, because by Theorem 3.1, the pace of tracking the river is approximately proportional to η.

For flat & slow-spinning river, momentum itself doesn't alter the speed on river much. Another interesting prediction by Theorem 3.1 is that for a flat and slow spinning river ($\gamma_{\text{flat}}\gamma_{\max}^{-1} \ll 1$, $\kappa \ll 1$), the speed of tracking the river is not influenced much by the momentum parameter β itself. Indeed, the speed is approximately proportional to the learning rate and is similar to that of (GD). Given that, the benefit of momentum for acceleration of tracking the river is mainly brought by its stabilization effects under an aggressive choice of larger learning rates η. But still, if the landscape features a quickly spinning or relatively sharp river, the influences from the factors of $\gamma_{\text{flat}}\eta$ and κ are no longer negligible. In that case, $\mathcal{O}_\beta(\cdot)$ would become especially large when β is very close to 1, for which large momentum could hurt optimization due to failure of tracking the river closely. It is an interesting future work to investigate in that regime how momentum works and how the momentum parameter interplays with these geometric properties of the loss landscape in a quantitative manner. See numerical verification in Appendix C, and language model training experiments in Section J.

4. Conclusions

We prove that in river-valley landscape, momentum enables a significantly larger learning rate than vanilla GD by stabilizing oscillations across the sharp valley, thus accelerating progress along the flat, low-loss river manifold. Our theory is validated by synthetic experiments and language model training. Future directions include extending the analysis to adaptive optimization methods.

References

[1] Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Sgd with large step sizes learns sparse features. In *International Conference on Machine Learning*, pages 903–925. PMLR, 2023.

[2] Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pages 948–1024. PMLR, 2022.

[3] Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pages 483–513. PMLR, 2020.

[4] Raghu Bollapragada, Tyler Chen, and Rachel Ward. On the fast convergence of minibatch heavy ball momentum. *IMA Journal of Numerical Analysis*, page drae033, 2024.

[5] Yuhang Cai, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter Bartlett. Large stepsize gradient descent for non-homogeneous two-layer networks: Margin improvement and fast optimization. *Advances in Neural Information Processing Systems*, 37:71306–71351, 2024.

[6] Matias D Cattaneo, Jason Matthew Klusowski, and Boris Shigida. On the implicit bias of adam. In *International Conference on Machine Learning*, pages 5862–5906. PMLR, 2024.

[7] Yineng Chen, Zuchao Li, Lefei Zhang, Bo Du, and Hai Zhao. Bidirectional looking with a novel double exponential moving average to adaptive and non-adaptive momentum optimizers. In *International Conference on Machine Learning*, pages 4764–4803. PMLR, 2023.

[8] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on learning theory*, pages 1305–1338. PMLR, 2020.

[9] 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. *arXiv preprint arXiv:2103.00065*, 2021.

[10] Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers. In *Advances in Neural Information Processing Systems*, 2021.

[11] Anh Quang Dang, Reza Babanezhad Harikandeh, and Sharan Vaswani. Noise-adaptive (accelerated) stochastic heavy-ball momentum. In *OPT 2023: Optimization for Machine Learning*, 2023.

[12] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[13] Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Gradient descent with adaptive stepsize converges (nearly) linearly under fourth-order growth. *arXiv preprint arXiv:2409.19791*, 2024.

[14] Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023.

[15] C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. In *International Conference on Learning Representations*, 2017.

[16] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[17] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

[18] Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pages 310–315. IEEE, 2015.

[19] Avrajit Ghosh, He Lyu, Xitong Zhang, and Rongrong Wang. Implicit regularization in heavy-ball momentum accelerated stochastic gradient descent. In *The Eleventh International Conference on Learning Representations*, 2023.

[20] Igor Gitman, Hunter Lang, Pengchuan Zhang, and Lin Xiao. Understanding the role of momentum in stochastic gradient methods. *Advances in neural information processing systems*, 32, 2019.

[21] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[22] Xinran Gu, Kaifeng Lyu, Longbo Huang, and Sanjeev Arora. Why (and when) does local sgd generalize better than sgd? In *The Eleventh International Conference on Learning Representations*, 2023.

[23] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. *Advances in neural information processing systems*, 31, 2018.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[26] Shengding Hu, Yuge Tu, Xu Han, Ganqu Cui, Chaoqun He, Weilin Zhao, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Xinrong Zhang, Zhen Leng Thai, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, dahai li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=3X2L2TFr0f.

[27] Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference On Learning Theory*, pages 545–604. PMLR, 2018.

[28] Stanislaw Jastrzebski, Devansh Arpit, Oliver Astrand, Giancarlo B Kerg, Huan Wang, Caiming Xiong, Richard Socher, Kyunghyun Cho, and Krzysztof J Geras. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In *International Conference on Machine Learning*, pages 4772–4784. PMLR, 2021.

[29] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on learning theory*, pages 1772–1798. PMLR, 2019.

[30] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020.

[31] Sebastian Kassing and Simon Weissmann. Polyak's heavy ball method achieves accelerated local rate of convergence under polyak-lojasiewicz inequality. *arXiv preprint arXiv:2410.16849*, 2024.

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] Lingkai Kong and Molei Tao. Stochasticity of deterministic gradient descent: Large learning rate for multiscale objective function. *Advances in neural information processing systems*, 33: 2625–2638, 2020.

[34] Xiaoyu Li, Mingrui Liu, and Francesco Orabona. On the last iterate convergence of momentum methods. In *International Conference on Algorithmic Learning Theory*, pages 699–717. PMLR, 2022.

[35] Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after sgd reaches zero loss?–a mathematical framework. In *International Conference on Learning Representations*, 2022.

[36] Fangshuo Liao and Anastasios Kyrillidis. Provable accelerated convergence of nesterov's momentum for deep relu neural networks. In *International Conference on Algorithmic Learning Theory*, pages 732–784. PMLR, 2024.

[37] Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. Same pre-training loss, better downstream: Implicit bias matters for language models. In *International Conference on Machine Learning*, pages 22188–22214. PMLR, 2023.

[38] Xin Liu, Wei Tao, and Zhisong Pan. A convergence analysis of nesterov's accelerated gradient method in training deep linear neural networks. *Information Sciences*, 612:898–925, 2022.

[39] Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems*, 33:18261–18271, 2020.

[40] Nicolas Loizou and Peter Richtárik. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77(3):653–710, 2020.

[41] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

[42] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[43] Miao Lu, Beining Wu, Xiaodong Yang, and Difan Zou. Benign oscillation of stochastic gradient descent with large learning rate. In *The Twelfth International Conference on Learning Representations*, 2024.

[44] James Lucas, Shengyang Sun, Richard Zemel, and Roger Grosse. Aggregated momentum: Stability through passive damping. In *International Conference on Learning Representations*, 2019.

[45] Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. *Advances in Neural Information Processing Systems*, 35:34689–34708, 2022.

[46] Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon Shaolei Du, Jason D Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *The Twelfth International Conference on Learning Representations*, 2024.

[47] Chao Ma, Daniel Kunin, Lei Wu, and Lexing Ying. Beyond the quadratic approximation: the multiscale structure of neural network loss landscapes. *arXiv preprint arXiv:2204.11326*, 2022.

[48] Si Yi Meng, Antonio Orvieto, Daniel Yiming Cao, and Christopher De Sa. Gradient descent on logistic regression with non-separable data and large step sizes. *arXiv preprint arXiv:2406.05033*, 2024.

[49] Yurii Nesterov. A method for solving the convex programming problem with convergence rate o (1/k2). In *Dokl akad nauk Sssr*, volume 269, page 543, 1983.

[50] Matteo Pagliardini, Pierre Ablin, and David Grangier. The ademamix optimizer: Better, faster, older. *arXiv preprint arXiv:2409.03137*, 2024.

[51] Rui Pan, Haishan Ye, and Tong Zhang. Eigencurve: Optimal learning rate schedule for sgd on quadratic objectives with skewed hessian spectrums. In *International Conference on Learning Representations*, 2022.

[52] Rui Pan, Yuxing Liu, Xiaoyu Wang, and Tong Zhang. Accelerated convergence of stochastic heavy ball method under anisotropic gradient noise. In *The Twelfth International Conference on Learning Representations*, 2024.

[53] Yan Pan and Yuanzhi Li. Toward understanding why adam converges faster than sgd for transformers. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

[54] Hristo Papazov, Scott Pesme, and Nicolas Flammarion. Leveraging continuous time to understand momentum when training diagonal linear networks. In *International Conference on Artificial Intelligence and Statistics*, pages 3556–3564. PMLR, 2024.

[55] Courtney Paquette and Elliot Paquette. Dynamics of stochastic momentum methods on large-scale, quadratic models. *Advances in Neural Information Processing Systems*, 34:9229–9240, 2021.

[56] Maximilian Plattner. On sgd with momentum. *URL http://infoscience. epfl.ch/record/295398*, 2022.

[57] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[58] Damien Scieur and Fabian Pedregosa. Universal average-case optimality of polyak momentum. In *International conference on machine learning*, pages 8565–8572. PMLR, 2020.

[59] Othmane Sebbouh, Robert M Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on Learning Theory*, pages 3935–3971. PMLR, 2021.

[60] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19 (70):1–57, 2018.

[61] Jun-Kun Wang, Chi-Heng Lin, and Jacob D. Abernethy. Escaping saddle points faster with stochastic momentum. In *ICLR*, 2020. URL https://openreview.net/forum?id=rkeNfp4tPr.

[62] Jun-Kun Wang, Chi-Heng Lin, and Jacob D Abernethy. A modular analysis of provable acceleration via polyak's momentum: Training a wide relu network and a deep linear network. In *International Conference on Machine Learning*, pages 10816–10827. PMLR, 2021.

[63] Jun-Kun Wang, Chi-Heng Lin, Andre Wibisono, and Bin Hu. Provable acceleration of heavy ball beyond quadratics for a class of polyak-lojasiewicz functions when the non-convexity is averaged-out. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22839–22864. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wang22p.html.

[64] Runzhe Wang, Sadhika Malladi, Tianhao Wang, Kaifeng Lyu, and Zhiyuan Li. The marginal value of momentum for small learning rate SGD. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3JjJezzVkT.

[65] Yuqing Wang, Minshuo Chen, Tuo Zhao, and Molei Tao. Large learning rate tames homogeneity: Convergence and balancing effect. In *International Conference on Learning Representations*, 2022.

[66] Yuqing Wang, Zhenghao Xu, Tuo Zhao, and Molei Tao. Good regularity creates large learning rate implicit biases: edge of stability, balancing, and catapult. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023.

[67] Jingrong Wei and Long Chen. Accelerated over-relaxation heavy-ball methods with provable acceleration and global convergence. *arXiv preprint arXiv:2406.09772*, 2024.

[68] Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How sharpness-aware minimization minimizes sharpness? In *The eleventh international conference on learning representations*, 2023.

[69] Kaiyue Wen, Zhiyuan Li, Jason S. Wang, David Leo Wright Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape view. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=m51BgoqvbP.

[70] Jingfeng Wu, Difan Zou, Vladimir Braverman, and Quanquan Gu. Direction matters: On the implicit bias of stochastic gradient descent with moderate learning rate. In *International Conference on Learning Representations*, 2021.

[71] Jingfeng Wu, Vladimir Braverman, and Jason D Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. *Advances in Neural Information Processing Systems*, 36:74229–74256, 2023.

[72] Jingfeng Wu, Peter L Bartlett, Matus Telgarsky, and Bin Yu. Large stepsize gradient descent for logistic loss: Non-monotonicity of the loss improves optimization efficiency. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 5019–5073. PMLR, 2024.

[73] Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with sgd. *arXiv preprint arXiv:1802.08770*, 2018.

[74] Huaqing Xiong, Yuejie Chi, Bin Hu, and Wei Zhang. Analytical convergence regions of accelerated gradient descent in nonconvex optimization under regularity condition. *Automatica*, 113:108715, 2020.

[75] Yang Xu, Yihong Gu, and Cong Fang. The implicit bias of heterogeneity towards invariance: A study of multi-environment matrix sensing. *arXiv preprint arXiv:2403.01420*, 2024.

[76] Zhenghao Xu, Yuqing Wang, Tuo Zhao, Rachel Ward, and Molei Tao. Provable acceleration of nesterov's accelerated gradient for asymmetric matrix factorization and linear neural networks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[77] Yan Yan, Tianbao Yang, Zhe Li, Qihang Lin, and Yi Yang. A unified analysis of stochastic momentum methods for deep learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2955–2961, 2018.

[78] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

[79] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

[80] Ruiqi Zhang, Jingfeng Wu, Licong Lin, and Peter L Bartlett. Minimax optimal convergence of gradient descent in logistic regression via large and adaptive stepsizes. *arXiv preprint arXiv:2504.04105*, 2025.

[81] Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *The Eleventh International Conference on Learning Representations*, 2023.

Appendix A. Review: What Happens for Vanilla GD?

Before giving our results on (GD-M), we first review what happens for vanilla (GD) in the river-valley landscape. As we mentioned in the previously, GD starting from the neighborhood would (i) converge close to the river within distance of order $\mathcal{O}(\kappa/\gamma)$ (Theorem 3.1 of Wen et al. [69]), and then (ii) closely track the river afterwards. We recap the results for the second stage as follows.

Theorem A.1 (Informal, GD in river-valley landscape (improved version of Theorem 3.2 in Wen et al. 69)) *Suppose Assumptions 2.1 and 2.2 hold. Let η be a learning rate such that $\eta < \eta_{\max}^{\mathrm{GD}}$, where*

$$\eta_{\max}^{\mathrm{GD}} \approx \frac{2 - \mathcal{O}(\kappa + \gamma_{\mathrm{flat}}\gamma_{\max}^{-1})}{\gamma_{\max}}. \tag{A.1}$$

Then there exists a time index T_0 such that (GD) with initialization $w_0 \in \mathcal{M}$ on the river satisfies that for any step k, there exists $T(k)$ s.t. the following two things hold:

1. *GD stays close to the river: $\|x(T_0 + T(k)) - w_k\|_2 \le \mathcal{O}(\kappa/\gamma)$;*
2. *The speed on river is nearly proportional to η: $|T(k) - \eta \cdot k| \le \mathcal{O}(\kappa + \eta\gamma_{\mathrm{flat}}) \cdot \eta$.*

See formal version in Appendix I.1. By Theorem A.1, when GD closely tracks the river, the larger the learning rate (if tolerable), the faster the iterates move along the river, which means faster optimization progress. The theory partially demystifies the success and the non-trivial loss curve of WSD scheduler in practice [26]. Using a large constant learning rate induces a larger training loss before learning rate decay due to oscillations in the sharp directions, but after the learning rate decay, the loss reflects the true progress it has made (Theorem 3.5 of Wen et al. [69]).

One key factor of Theorem A.1 is the upper bound on the learning rate, which ensures that the iterates remain close to the river without exploding due to oscillations between the valley. Intuitively, the bottleneck of using a large learning rate lies in *the curvature of sharp directions* and *the spinning of the river*, for which $\eta_{\max}^{\mathrm{GD}}$ is derived. Therefore, the dominant term in $\eta_{\max}^{\mathrm{GD}} \approx 2/\gamma_{\max}$, roughly the maximum learning rate for GD optimizing a quadratic with largest Hessian eigenvalue $\gamma_{\max}$.

Remark A.2 (Comparison with Wen et al. [69]) *Theorem A.1 improves the largest tolerable learning rate $\eta_{\max}^{\mathrm{GD}}$ compared to the original $\gamma/2\gamma_{\max}^2$ in Wen et al. [69]. The new proofs are in Appendix I.2. The improvement results from a tighter analysis of sharp direction dynamics.*

Given all the results so far, the natural questions are that, how does the inclusion of momentum in the optimization iteration effect the dynamics of tracking the river? and how does the additional momentum parameter interplays with the learning rate? We answer the question in Section 3.

Appendix B. Proof Outline, Key Challenges, and Technical Novelty

In this section we sketch the proof of Theorem 3.1. We first introduce several key concepts used in the proofs. Then we outline the main proofs and point out the key challenges. Then we walk through the key components of the proof, highlighting technical novelties.

Projections onto flat and sharp directions. Recall that $v_d(\nabla^2 L(w))$ is the river direction of the loss Hessian at w. We abbreviate it as $v_d(w)$. On the river, $v_d(w)$ aligns with $\nabla L(w)$. This motivates us to decompose the gradient and the momentum into the projections onto $v_d(w)$ and its orthogonal space. To this end, we define projection matrices $\boldsymbol{P}_{\mathrm{f}}(w) := v_d(w)v_d(w)^\top$ and $\boldsymbol{P}_{\mathrm{s}}(w) := \boldsymbol{I}_d - \boldsymbol{P}_{\mathrm{f}}(w)$.

Projection onto the river $\mathcal{M}$. To formalize the idea of how the iteration $\{w_k\}_{k\in\mathbb{N}}$ tracks the river, we introduce the projection onto the river of any point nearby. The definition of it is chosen as the limit of an ODE flow starting from the point to be projected. Formally, consider any $w \in \mathcal{U}$ (the nice set near the river $\mathcal{M}$, see Assumption 2.2). We define the ODE flow $\{\phi(w,t) : w \in \mathcal{U}, t \geq 0\}$:

$$\phi(w,0) = w, \quad \frac{\mathrm{d}}{\mathrm{d}t}\phi(w,t) = -\boldsymbol{P}_{\mathrm{s}}(\phi(w,t))\nabla L(\phi(w,t)), \quad t \geq 0.$$

As is shown in Lemma H.4, the ODE flow has a limit $\lim_{t\to+\infty}\phi(w,t) \in \mathcal{M}$, and we define it as $\Phi(w)$, i.e., the projection of w onto the river. Other important properties are listed in Lemma H.4.

Continuous time linear interpolation of trajectory. To help characterize how the discrete-time sequence $\{w_k\}_{k\in\mathbb{N}}$ tracks the continuous-time river $\{x(t)\}_{t\geq 0}$, we consider a linear interpolation. For any $k \in \mathbb{N}$ and $\tau \in [0,1]$, we let

$$w_{k,\tau} := \tau \cdot w_{k+1} + (1-\tau)\cdot w_k = w_k - \tau\eta \cdot m_{k+1}. \tag{B.1}$$

We then obtain a continuous-time process $\{w_{\lfloor t\rfloor, t-\lfloor t\rfloor}\}_{t\geq 0}$. In the proof, we iteratively show that $\{w_{\lfloor t\rfloor, t-\lfloor t\rfloor}\}_{t\geq 0} \subseteq \mathcal{U}$ to make sure that the projection to the river $\Phi(w_{\lfloor t\rfloor, t-\lfloor t\rfloor})$ is properly defined for any $t \geq 0$[1]. Thus, such an interpolated continuous-time sequence induces a trajectory on the river, i.e., $\{\Phi(w_{\lfloor t\rfloor, t-\lfloor t\rfloor})\}_{t\geq 0}$. We fix the initialization $x(T_0) = \Phi(w_{0,0})$ in the reference flow (2.1) and define $T(t)$ to be the time index associated with $\Phi(w_{\lfloor t\rfloor, t-\lfloor t\rfloor})$, i.e.,

$$x(T(t)) = \Phi\big(w_{\lfloor t\rfloor, t-\lfloor t\rfloor}\big). \tag{B.2}$$

B.1. Proof Outline and Key Challenges

Proof outline. With the above setups, we consider the following route of proof:

1. We first prove that for (GD-M) the gradient in the flat direction dominates the gradient and the momentum in the sharp directions, and the maximal learning rate is limited by the dynamics in sharp directions so that the iterates in the sharp directions do not explode.
2. Then, we are able to prove that (GD-M) always stays close to the river. In more specific, we show that w_k stays close to its projection onto the river $\Phi(w_k)$.
3. Finally, we prove that the time index $T(t)$ grows nearly proportional to η, finishing proof.

Key challenges. The essential challenge here is to establish the conditions under which the iteration in such a non-convex landscape does not explode, and how the momentum stabilizes the training process. This requires a delicate analysis of the dynamics of both the gradient and the momentum in the river direction and the sharp directions. But due to the river spinning, the projections of the gradient and the momentum onto these directions interfere with each other in a complicated manner, necessitating a careful induction argument to clearly track all of their dynamics.

B.2. Key Lemmas and Technical Novelties

Analysis of projections onto flat and sharp directions. We first prove that the dynamics in the flat direction dominate in the sense that the gradient and momentum mostly live in the flat direction. Also, the dynamics in the sharp direction determine the maximal tolerable learning rate such that the updates do not explode. This is the following lemma.

1. Concretely, $\mathcal{B}(w_{\lfloor t\rfloor, t-\lfloor t\rfloor}, 2g_{\max}/\gamma) \subseteq \mathcal{U}, \forall t \geq 0$ to ensure existence of $\Phi(w_{\lfloor t\rfloor, t-\lfloor t\rfloor})$, see Lemma H.4.

Lemma B.1 (Projections onto flat and sharp directions (informal and short version of Lemma G.2)) *Under the conditions and assumptions of Theorem 3.1, taking the step size η and momentum parameter β satisfying $\eta \le \eta_{\max}^{\text{GD-M}}$ as defined in Theorem 3.1, for any iteration $k \in \mathbb{N}$ and $\tau \in [0,1]$, it holds that*

$$\max\left\{\|\boldsymbol{P}_{\text{s}}(w_{k,\tau})\nabla L(w_{k,\tau})\|_2, \|\boldsymbol{P}_{\text{s}}(w_{k,\tau})m_{k+1}\|_2\right\} \le \mathcal{O}_\beta\big(\kappa^{1/2}\cdot\|\boldsymbol{P}_{\text{f}}(w_{k,\tau})\nabla L(w_{k,\tau})\|_2\big),$$

where κ characterizes the slow spinning speed of the river (item 3 of Assumption 2.2).

We refer the readers to Lemma G.2 in Appendix G.2 for the full version of Lemma B.1 as well as the detailed proofs. One of the main technical novelty here is a delicate induction argument to separately track the magnitude of the projections of the gradient and the momentum onto different directions in the river-valley landscape. Compared with the corresponding result for vanilla GD (Lemma I.2), momentum enlarges the maximum learning rate by a multiplicative factor of $(1+\beta)/(1-\beta)$. This is the key reason behind the larger learning rate blessed by momentum in river-valley.

GD with momentum tracks the river closely. Then we establish that the GD momentum trajectory can track the river closely by the following lemma. It demonstrates that the interpolated continuous time trajectory $\{w_{\lfloor t\rfloor, t-\lfloor t\rfloor}\}_{t\ge 0}$ has a well-defined projection onto the river and the points in the trajectory are not far away from their projected counterparts. See proofs in Appendix G.3.1.

Lemma B.2 (The trajectory tracks the river closely (informal and short version of Lemma G.10)) *Under the conditions and assumptions of Theorem 3.1, taking the learning rate η and momentum parameter β satisfying $\eta \le \eta_{\max}^{\text{GD-M}}$ as defined in Theorem 3.1, then for any iteration $k \in \mathbb{N}$ and $\tau \in [0,1]$ it holds that: (i) $\mathcal{B}(w_{k,\tau}, 2g_{\max}/\gamma) \subset \mathcal{U}$ and thus $\Phi(w_{k,\tau})$ exists; (ii) it holds that for any step $k \in \mathbb{N}$ and $\tau \in [0,1]$,*

$$\|w_{k,\tau} - \Phi(w_{k,\tau})\|_2 \le \mathcal{O}_\beta\left(\frac{\kappa^{1/2}\cdot\|\boldsymbol{P}_{\text{f}}(w_{k,\tau})\nabla L(w_{k,\tau})\|_2}{\gamma + 2\gamma_{\text{flat}}}\right).$$

Time index grows almost linearly w.r.t. learning rate. Finally, with previous results, we are able to show that the time index $\{T(t)\}_{t\ge 0}$ defined in (B.2) approximately grows linearly with t, where the linear coefficient approximately equals to the learning rate η. See proofs in Appendix G.3.2.

Lemma B.3 (Time index grows almost linearly with learning rate (informal and short version of Lemma G.11)) *Under the conditions and assumptions of Theorem 3.1, taking the learning rate η and momentum parameter β satisfying $\eta \le \eta_{\max}^{\text{GD-M}}$ as defined in Theorem 3.1, then for any iteration $k \ge \log(\beta\eta\gamma_{\text{flat}}/(1-\beta))/\log\beta$ and $\tau \in (0,1)$, by letting $t = \tau + k$, the derivative of $T(t)$ exists and satisfies $|\mathrm{d}T/\mathrm{d}t(t) - \eta| \le \epsilon(\beta)\cdot\eta$. Equivalently, this bound holds almost everywhere on each interpolation interval, with the integer breakpoints understood through one-sided derivatives. Here $\epsilon(\beta)$ is defined as $\epsilon(\beta) := \mathcal{O}_\beta(\kappa + \eta\gamma_{\text{flat}})$.*

Combining Lemma B.2 and Lemma B.3 gives Theorem 3.1.

Appendix C. Synthetic 2-D Non-Convex Experiment

Experiment: a 2-d non-convex loss. We first verify our theory in a synthetic non-convex loss function $L(x_1, x_2) = 2\cdot(x_2 - \sin(x_1))^2 - 0.1\cdot x_1$. We run (GD-M) with extensive choices of (β, η)

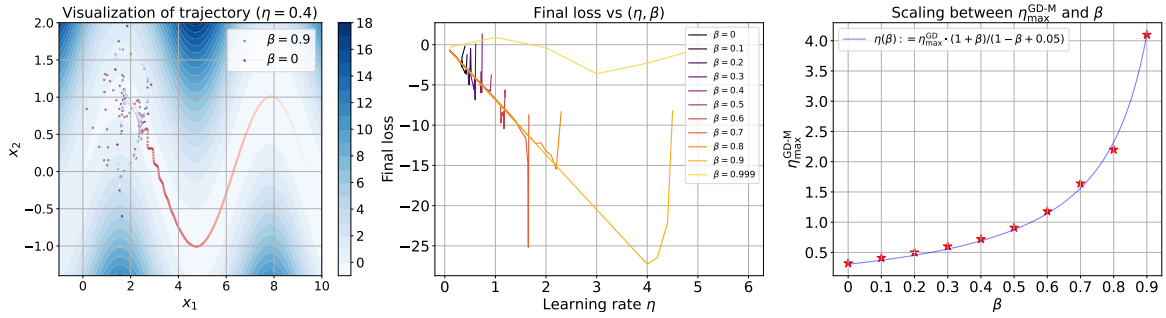

Figure 2: Results of experiment for $L(x_1, x_2) = 2 \cdot (x_2 - \sin(x_1))^2 - 0.1 \cdot x_1$. The left figure shows the trajectories of (GD) and (GD-M) under the same learning rate. The former diverges while the latter stabilizes the oscillation and tracks the river closely. The figure in the middle shows relation between the final loss and η under different β. The right figure shows the scaling of the largest tolerable learning rate w.r.t. β.

to validate theory. Results are in Figure 2. We start from $w_0 = (0, 1)$, a point near the river, and run (GD-M) for 1000 steps. In Figure 2 (left), we plot the trajectories of (GD-M) with $\beta = 0.9$ and $\beta = 0$ (corresponding to (GD)), where both $\eta = 0.4$. We can see (GD-M) with $\beta = 0.9$ manages to stabilize the trajectory and tracks the river smoothly. In contrast, GD can not tolerate such a learning rate and does not converge, bouncing between the valley. This matches the basic picture of our theory.

Furthermore, we demonstrate quantitative results. In Figure 2 (middle), we plot the final loss after 1000 steps under different (β, η) pairs. Here $\beta \in \{0, 0.1, 0.2, \ldots, 0.8, 0.9\}$, while the learning rate is taken from 0.1 to a threshold above which the iteration does not track the river closely and diverges. We have three key observations from the results:

1. For larger β, the largest tolerable LR $\eta_{\max}^{\text{GD-M}}$ is larger.
2. Given β, the final loss is nearly proportional to the LR.
3. Given a LR η, for those β such that η is tolerable, the choice of β does not significantly change the final loss.

All of these results demonstrate our theoretical predictions given by main theorem. Besides, we test $\beta = 0.999$ and it does not converge for any η, which echos the condition on β not being too close to 1 when the river spins (see previous discussions). The detailed constraints on β are in Theorem G.1.

Finally, we extract the numerically obtained $\eta_{\max}^{\text{GD-M}}$ and plot its relation with β. See Figure 2 (right). As shown, they nearly lie in the curve of $\eta(\beta) := \eta_{\max}^{\text{GD}} \cdot (1+\beta)/(1-\beta+0.05)$, matching theoretical prediction (3.2). The 0.05 factor is caused by the spinning and the curvature of the river.

Appendix D. Experiments on Language Model Training

Experimental setup. We train GPT-Neo [16] (adjusted to about 20 million parameters) on the TinyStories dataset [14], containing about 0.5 billion tokens. The training split contains about 2.2 million examples, and the validation split contains about 22000 examples. All experiments use batch size 32 and context window size 256. We conduct two groups of experiments: one with SGD-Momentum, which is closest to our theory, and one with Adam, which tests whether the same

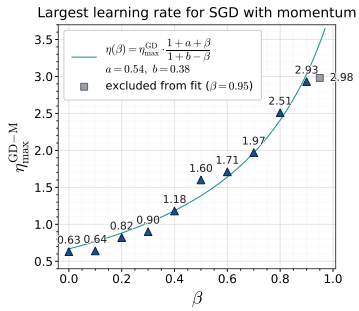

Figure 3: SGD-M: largest stable learning rate versus β.

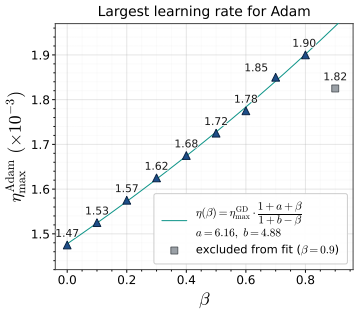

Figure 4: Adam: largest stable learning rate versus β_1.

η	β	Train	Val.
0.1	0	1.867	2.066
0.1	0.9	1.843	2.065
0.1	0.95	1.840	2.073
1	0	div.	div.
1	0.9	1.594	1.781
1	0.95	1.602	1.783

Table 1: Fixed-grid SGD-M final losses.

mechanism appears in practical language model training. For the Adam experiments, we train for 1 epoch, fix $\beta_2 = 0.95$, and vary the learning rate η and the first-order moment parameter β_1. For the SGD-M experiments, we use the same model, data processing, and training protocol, but replace Adam by stochastic gradients with heavy-ball momentum and vary the momentum parameter β. For both optimizers, we warm up the learning rate linearly for the first 100 steps, and for the last 15% of training steps, we decay the learning rate to 0.0001 to reduce the oscillations between the valley and reveal the underlying progress along the river, similar to the WSD scheduler [26, 69]. Additional experiment details and full grids are provided in Appendix J.

SGD-M: direct evidence for the learning-rate stabilization mechanism. We first test SGD-M, which removes Adam preconditioning and is closest to the optimizer analyzed in Theorem 3.1. For each momentum parameter β, we grid search the largest learning rate under which the training loss curve does not diverge. Figure 3 shows that the largest tolerable learning rate $\eta_{\max}^{\mathrm{SGD\text{-}M}}$ increases with β, matching the increasing trend predicted by (3.2). The fitted curve uses two additional constants, which soften the singularity near $\beta = 1$ and capture non-asymptotic effects from river spinning, river curvature, and stochastic gradient noise. Thus, we interpret this sweep as empirical support for the scaling trend rather than an exact realization of the idealized formula.

Table 1 gives a complementary fixed-grid comparison. With the large learning rate $\eta = 1$, training without momentum diverges, while training with $\beta = 0.9$ or $\beta = 0.95$ converges to low training and validation losses. In contrast, under the same stable learning rate $\eta = 0.1$, changing β from 0.9 to 0.95 only has a marginal effect on the final losses. This supports the message of our theory that momentum accelerates primarily by stabilizing larger learning rates, rather than by substantially changing the river-direction speed at a fixed learning rate.

Adam: the same stabilization mechanism appears in practical training. We next test whether the same qualitative behavior appears for Adam, a standard optimizer for language models.
We again first grid search the learning rate η under different momentum β_1 for the largest learning rate such that the training loss curve does not diverge. As shown in Figure 4, the largest tolerable learning rate under Adam increases as β_1 grows, which is consistent with the stabilization mechanism predicted by our theory. When β_1 approaches 1, we also observe a decrease in the largest tolerable learning rate, which is interpreted by the role of river curvature and spinning discussed in Section 3.2.

We then examine whether the larger stability window translates into better optimization and validation performance. Figure 5 plots the training and validation curves for $\eta \in \{0.0001, 0.0006, 0.0008\}$

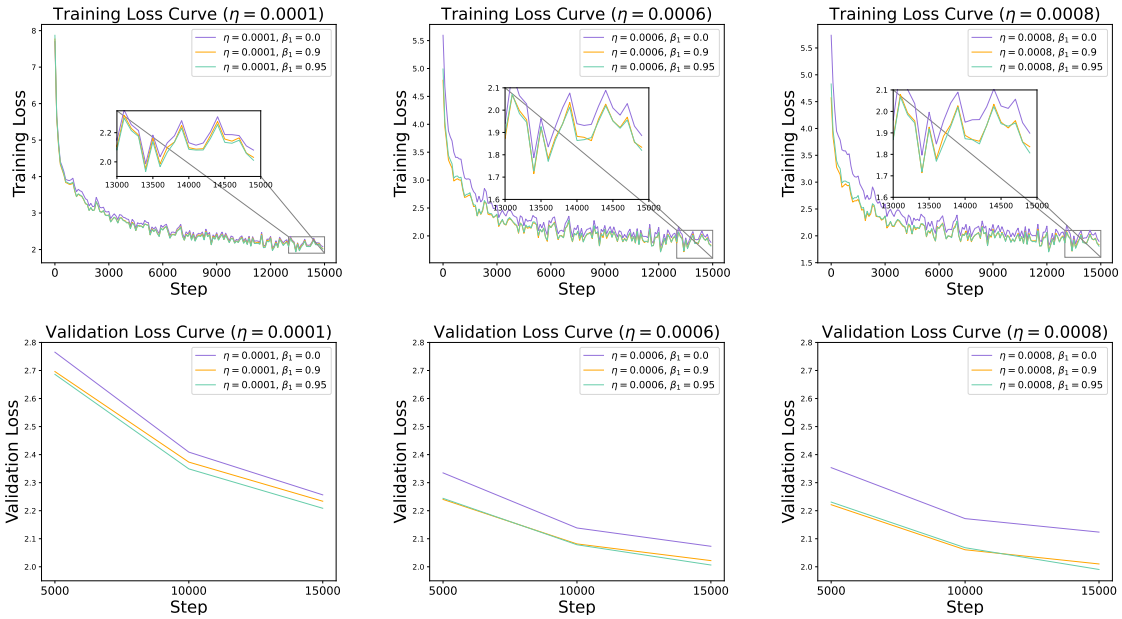

Figure 5: GPT-Neo training on TinyStories [14]. Top: training loss; bottom: validation loss. Columns correspond to $\eta \in \{0.0001, 0.0006, 0.0008\}$.

and $\beta_1 \in \{0, 0.9, 0.95\}$. For the small learning rate $\eta = 0.0001$, different choices of β_1 lead to similar loss curves. For the larger learning rates $\eta = 0.0006$ and $\eta = 0.0008$, Adam with $\beta_1 \in \{0.9, 0.95\}$ is more stable and reaches lower training and validation losses than Adam with $\beta_1 = 0$. Moreover, with momentum, using a larger learning rate gives lower training and validation losses than using the small learning rate. These observations again support the acceleration-by-stabilization explanation in practice: momentum improves training mainly by allowing a more aggressive stable learning rate.

Appendix E. Related Works

Theoretical analysis of momentum. Momentum acceleration in convex optimization dates back to Polyak's heavy-ball method [57]. Accelerated convergence of heavy-ball momentum (stochastic) gradient descent and its variants has been established in locally near-quadratic setups, including the convex case [4, 11, 18, 27, 34, 39, 40, 52, 55, 58, 59, 67] and the non-convex case with the PL condition [31, 62, 63], as well as non-convex convergence to stationary points [20, 39, 59, 61, 74, 77]. However, the understanding of momentum beyond such structured setups remains limited, especially for the non-convex loss landscapes emerging from training large neural networks. Wang et al. [62] analyze the convergence acceleration of heavy-ball momentum for wide ReLU neural networks and deep linear networks. Plattner [56] empirically show that momentum enlarges the tolerable learning rate in deep learning but cannot improve performance much. Wang et al. [64] demonstrate the marginal value brought by momentum in the small learning rate regime. Other lines of work study the implicit bias of heavy-ball momentum to understand its influence on generalization [6, 19, 54]. Chen et al. [7], Lucas et al. [44], Pagliardini et al. [50] consider modifications of heavy-ball momentum by mixing several exponential moving averages to better use past gradients with different patterns. Besides heavy-ball momentum, another recent line of work studies Nesterov's momentum [49]; see Liao and Kyrillidis [36], Liu et al. [38], Xu et al. [76]. This is a different approach, while we investigate heavy-ball momentum due to its wide use in modern optimizers. Our theory takes an abstract river-valley perspective of deep learning loss landscapes, providing insights for a wider range of problems exhibiting such a landscape. Instead of following a structured convergence-rate analysis, we customize the analysis of heavy-ball momentum in the river-valley loss landscape by rigorously showing how it accelerates via enabling stable tracking of the river in the presence of large learning rates.

Large learning rate training. The benefits of large learning rate training have been increasingly investigated in recent years [9, 24, 73]. Meng et al. [48], Wu et al. [72], Zhang et al. [80] prove accelerated optimization with large learning rates in logistic regression. Cai et al. [5] studies large learning rate acceleration in non-homogeneous two-layer neural networks. Wen et al. [69] demonstrates the benefits of large learning rate GD training in terms of tracking the low-loss river manifold in the river-valley landscape. A recent line of work on the "Edge of Stability" [2, 9, 28] demonstrates that large learning rate training can lead to flatter minima [1, 33, 43, 65, 66, 71, 75, 81], which is believed to be crucial to better generalization. However, existing analyses are restricted to setups without momentum. To fill this gap, we study GD with momentum and demonstrate that momentum enables stable training with larger learning rates than vanilla GD, giving new insights into modern optimizers.

Analysis of loss landscape. Understanding the interactions between the loss landscape and optimization dynamics, as well as generalization capabilities, is key to demystifying and further boosting the success of deep learning [15, 17]. The most relevant works to ours are Davis et al. [13], Wen et al. [69], Xing et al. [73], which all consider a "river-valley" like landscape of the loss function to understand the behavior of optimizers. Xing et al. [73] propose the conceptual picture where SGD locally bounces around the valley on top of a valley floor. Davis et al. [13] identify the existence of a manifold with vanishing gradient within the sharp directions of the Hessian for smooth loss functions that exhibit fourth-order growth near minimizers. All these works demonstrate the ubiquity of the river-valley like landscape. Pan et al. [51], Pan and Li [53], Zhang et al. [78, 79] study optimization

dynamics in landscapes with large sharpness in certain directions and heavy-tailed noise, all of which share a landscape similar to the "river-valley" one. Andriushchenko et al. [1], Blanc et al. [3], Gunasekar et al. [23], Jiang et al. [30], Li et al. [35], Liu et al. [37], Lu et al. [43], Lyu et al. [45], Ma et al. [47] examine the relationship between the loss landscape and generalization properties and demonstrate the belief that flatter minima result in better generalization. Finally, there is a line of work examining algorithmic implicit bias towards minima with different kinds of properties for different variants of (stochastic) gradient descent, including Arora et al. [2], Blanc et al. [3], Chizat and Bach [8], Damian et al. [10], Gu et al. [22], Ji and Telgarsky [29], Li et al. [35], Lu et al. [43], Lyu et al. [45, 46], Soudry et al. [60], Wen et al. [68], Wu et al. [70]. See also the references therein.

Comparison with Wen et al. [69]. The closest prior work to ours is Wen et al. [69], which introduces the river-valley loss landscape to explain WSD learning-rate schedules and analyzes how vanilla GD/SGD can make fast downstream progress along a one-dimensional river under large learning rates. Our work adopts this geometric framework as the starting point, but studies a different optimizer and a different dynamical question: how heavy-ball momentum (1.1) changes the stability threshold and the tracking dynamics near the river. This extension is not a direct consequence of the vanilla-GD analysis because the flat and sharp projections of both the gradient and the momentum interact with each other when the Hessian eigenspaces rotate along the trajectory. To handle this coupling, we develop a new induction argument for momentum dynamics and prove that the leading stability threshold is enlarged by the factor near $(1+\beta)/(1-\beta)$. Regarding the technical assumptions on river-valley, we differ from Wen et al. [69] in the following sense: since we focus on the stage-two, on-river regime rather than the initial approach-to-river phase, so the gradient-norm lower bound used for that first stage is not needed in our main tracking analysis. We also re-derive the vanilla-GD tracking result with an improved largest tolerable learning rate (Theorem A.1). The one-dimensional river model is inherited from Wen et al. [69], and extending the analysis to higher-dimensional river manifolds is an important future direction.

Appendix F. Further Discussions

Interpretation of river-valley loss landscape. Here we provide a more concrete realistic situation that induces the river-valley structure. In the standard next-token prediction loss, where the uncertainty of the next token can vary significantly, the variations in the uncertainty can shape this type of loss landscape. When the next token is predictable with high certainty, a larger learning rate can help the model learn faster. In contrast, when the next token is inherently ambiguous, e.g., given a phrase like "I am", the model must learn a well-calibrated probability distribution, which may require smaller updates. These fluctuations in uncertainty actually lead to the variations in the sharpness of the loss landscape, giving rise to the river-valley structure [69].

The stochastic setting. Beyond the deterministic setting in this paper, it is also interesting to consider the stochastic setting where gradients are corrupted by noise. However, it has been shown by Wang et al. [64] that when the learning rate is small for SGD (or when the batch size is small according to the linear scaling rule [21]), momentum has limited benefit. See Figure 1 of Wang et al. [64]. In contrast, when the batch size increases (more deterministic setting), the gap becomes significant. Therefore, it is more meaningful to study the benefit of momentum for large batch size, which is close to the deterministic setting we study here. We leave it for future work to systematically study SGD-momentum in river-valley.

Second-order ODE interpretation. We clarify why Theorem 3.1 compares GD-momentum with a first-order reference flow on the river, even though momentum methods are often interpreted through a second-order ODE. The key point is that the apparent mismatch is caused by both the parametrization of momentum and the very-flat-river regime. Eliminating m_k from (GD-M) gives the equivalent two-step recursion

$$w_{k+1} - 2w_k + w_{k-1} = -(1-\beta)(w_k - w_{k-1}) - \eta(1-\beta)\nabla L(w_k).$$

If w_k is formally viewed as samples of a continuous curve $y(s)$ with time spacing h, i.e., $w_k \approx y(kh)$, then the corresponding second-order ODE is

$$\ddot{y}(s) + a\dot{y}(s) + ab\nabla L(y(s)) = 0, \qquad a := \frac{1-\beta}{h}, \quad b := \frac{\eta}{h}.$$

This differs from the common unnormalized heavy-ball parametrization $w_{k+1} = w_k - \eta\nabla L(w_k) + \beta(w_k - w_{k-1})$, for which the gradient coefficient in the ODE is not multiplied by $1-\beta$. Our parametrization is the normalized first-moment update used in (GD-M); it is also the parametrization that matches the first-moment part of Adam and allows a controlled comparison with vanilla GD under the same learning-rate parameter η.

Now focus on the motion along the river, and denote the projected flat gradient by $g_{\mathrm{f}}(x) = \boldsymbol{P}_{\mathcal{M}}(x)\nabla L(x)$ for $x \in \mathcal{M}$. In the river-valley regime, the Hessian along the flat direction is small and the river spins slowly, quantified by γ_{flat} and κ. Therefore, over a short segment of the river, $g_{\mathrm{f}}(x)$ varies slowly and can be approximated by a local vector g_{loc}. Projecting the above ODE onto this local flat direction yields the local model

$$\ddot{y}_{\mathrm{f}}(s) + a\dot{y}_{\mathrm{f}}(s) + abg_{\mathrm{loc}} \approx 0.$$

Solving this linear equation gives

$$\dot{y}_{\mathrm{f}}(s) \approx \exp(-as)\big(\dot{y}_{\mathrm{f}}(0) + bg_{\mathrm{loc}}\big) - bg_{\mathrm{loc}}.$$

Thus, after the momentum burn-in period, the velocity satisfies $\dot{y}_{\mathrm{f}}(s) \approx -bg_{\mathrm{loc}}$. Taking the discrete-time parametrization $h = 1$, this gives the leading-order update direction $-\eta g_{\mathrm{loc}}$, which is the same first-order speed as vanilla GD with learning rate η. The lower bound on k in Theorem 3.1 is precisely used to remove the transient momentum term; in the discrete proof, this transient appears through the factor β^{k+1}.

This explains why the time-tracking statement in Theorem 3.1 takes the form $T(k) \approx \eta k$ and has no additional leading factor depending on β. Momentum does not directly multiply the river speed under the normalized parametrization (GD-M). Instead, its main effect is through stability in the sharp directions: it permits a much larger tolerable learning rate, whose leading improvement is the factor $(1+\beta)/(1-\beta)$. Once this larger η is chosen, the first-order tracking relation $T(k) \approx \eta k$ translates the enlarged stable learning rate into faster progress along the river.

Finally, the local constant-gradient calculation above should not be interpreted as a global assumption. Theorem 3.1 compares the GD-momentum trajectory with the changing reference flow on the river,

$$\frac{\mathrm{d}}{\mathrm{d}T}x(T) = -\boldsymbol{P}_{\mathcal{M}}(x(T))\nabla L(x(T)).$$

The reference flow already captures the variation of $\boldsymbol{P}_{\mathcal{M}}(x(T))\nabla L(x(T))$ along the river. The proof also controls this variation explicitly; for instance, Lemma G.11 bounds the difference between the projected momentum direction and the tangent vector of the reference flow by terms of order κ and $\eta\gamma_{\mathrm{flat}}$. Along the reference flow, the loss satisfies

$$\frac{\mathrm{d}}{\mathrm{d}T}L(x(T)) = -\|\boldsymbol{P}_{\mathcal{M}}(x(T))\nabla L(x(T))\|_2^2.$$

Hence, if the loss is lower bounded, the projected flat gradient cannot remain bounded away from zero forever. The second-order ODE calculation is therefore only a local explanation of the post-burn-in behavior in a very flat segment of the river, while the theorem itself tracks the globally changing river flow under the stated assumptions.

Appendix G. Key Proofs of the Main Theorem

G.1. Formal Statement of Theorem 3.1

Theorem G.1 (GD-momentum in river valley loss landscape (formal version)) *Suppose that Assumptions 2.1 and 2.2 hold. Define the following functions of momentum parameter $\beta \in (0,1)$,*

$$\mathsf{A}(\beta) := \frac{1}{2}\cdot\left(\frac{8\sqrt{2}}{(1-\beta)^4}+\frac{22}{(1-\beta)^2}\right)+d\cdot\left(\frac{64\sqrt{2}}{(1-\beta)^4}+\frac{88}{(1-\beta)^2}\right)+\left(\frac{8}{(1-\beta)^3}+3\right)^{\frac{1}{2}},$$

$$\mathsf{B}(\beta) := 2\cdot\left(\frac{16}{(1-\beta)^3}+6\right)^{\frac{1}{2}}\cdot\left(\frac{10}{1-\beta}+1\right)\cdot\frac{2}{1-\beta},\quad \mathsf{C}(\beta) := 1+5\cdot\frac{1+\beta}{1-\beta},$$

$$\mathsf{D}_1(\beta) := 1+\mathsf{A}(\beta)+\mathsf{B}(\beta),\quad \mathsf{D}_2(\beta) := 4\big(1+6\mathsf{C}(\beta)\big)\cdot\frac{1+\beta}{1-\beta},$$

and we denote $\varsigma(\beta)\in\mathbb{R}$ as (for simplicity we sometimes omit the dependency on β when it makes no confusion)

$$\varsigma(\beta) := \mathsf{D}_1(\beta)\cdot\kappa^{3/16}+\mathsf{D}_2(\beta)\cdot\gamma_{\text{flat}}\gamma_{\max}^{-1}. \tag{G.1}$$

Now we take the learning rate η and momentum parameter β satisfying

$$\beta \le 1-\varsigma(\beta),\quad \eta\in\mathcal{I}(\beta), \tag{G.2}$$

where $\varsigma(\beta)$ is defined in (G.1), and the interval $\mathcal{I}(\beta)$ is defined as $\mathcal{I}(\beta) := \mathcal{I}_1\cap\mathcal{I}_2\cap\mathcal{I}_3\cap\mathcal{I}_4$ with[2]

$$\mathcal{I}_1 := \left(\frac{\kappa^{1/32}}{\gamma_{\max}},\ \frac{2\cdot(1+\beta)-\kappa^{1/16}}{1-\beta}\cdot\frac{1}{\gamma_{\max}}\right), \tag{G.3}$$

$$\mathcal{I}_2 := \left(0,\ \frac{1}{2(1+6\mathsf{C}(\beta))}\cdot\frac{1}{\gamma_{\text{flat}}}\right),\quad \mathcal{I}_3 := \left(0,\ \frac{1-\beta}{12\beta+2\beta(1+6\mathsf{C}(\beta))}\cdot\frac{1}{\gamma_{\text{flat}}}\right),$$

$$\mathcal{I}_4 := \left(\left(\mathsf{D}_1(\beta)\cdot\kappa^{5/32}+\mathsf{D}_2(\beta)\cdot\kappa^{15/32}\right)\cdot\frac{2}{\gamma_{\max}},\ \left(\frac{1+\beta}{1-\beta}-\frac{5\beta^4-2\beta^3+6\beta^2-2\beta+1}{2(1-\beta)^4}\cdot\varsigma-\mathcal{O}(\varsigma^2)\right)\cdot\frac{2}{\gamma_{\max}}\right).$$

Then there exists a time index T_0 such that the GD-momentum iteration (GD-M) with initialization $w_0\in\mathcal{M}$ on the river and initial momentum $m_0=0$ satisfies that for any step $k\ge\log(\beta\eta\gamma_{\text{flat}}/(1-\beta))/\log\beta$, there exists another $T(k)$ such that the following two things hold:

1. *GD-Momentum stays close to the river: $\|x(T_0+T(k))-w_k\|_2\le 6\mathsf{B}(\beta)\cdot\kappa^{1/2}\cdot g_{\max}/\gamma$;*
2. *The speed on the river is approximately proportional to the learning rate: $|T(k)-\eta\cdot k|\le\epsilon(\beta)\cdot\eta$, where*

$$\epsilon(\beta) := 9\kappa+\left(6\big(1+\mathsf{C}(\beta)\big)+8+\frac{100\beta}{1-\beta}\right)\cdot\eta\gamma_{\text{flat}}+o(\kappa+\eta\gamma_{\text{flat}}).$$

Proof [Proof of Theorem G.1] The proof of Theorem G.1 is outlined in Section B. The proof is decomposed into several key lemmas (Lemmas G.2, G.10, and G.11) whose proofs are provided in the subsequent sections. ■

2. The exact formula of the right endpoint of $\mathcal{I}_4$ is given in (G.41). Here we only present the asymptotic expansion for better readability.

About the range of learning rate and momentum parameter. Here we remark on the valid range of learning rate η and momentum β specified in (G.2) and (G.3). Firstly, the condition on momentum parameter β in (G.2) is mild since when both κ and $\gamma_{\rm flat}/\gamma_{\max}$ are considered sufficiently small in our theory, the right hand side of $\beta \le 1 - \varsigma(\beta)$ is close to 1. That being said, we need β not too close to 1 due to the existence of slow spinning and the small curvature of the river manifold.

Now given a fixed β satisfying the first inequality in (G.2), we discuss the non-emptiness of the interval $\mathcal{I}(\beta)$ defined in (G.3). The second and third interval constraints $\mathcal{I}_2$ and $\mathcal{I}_3$ are relatively mild since they only require the learning rate η to be not too large for the flat direction, and under $\gamma_{\rm flat} \ll \gamma_{\max}$, the right endpoint of both $\mathcal{I}_2$ and $\mathcal{I}_3$ are larger than the right endpoint of $\mathcal{I}_1$ and $\mathcal{I}_4$. So it suffices to consider $\mathcal{I}_1$ and $\mathcal{I}_4$. For the left endpoints, in the regime of $\kappa \ll 1$, both left endpoints of $\mathcal{I}_1$ and $\mathcal{I}_4$ are close to zero and are smaller than the right endpoints of $\mathcal{I}_1$ and $\mathcal{I}_4$. Therefore, we can conclude that the intersection $\mathcal{I}(\beta) = \mathcal{I}_1 \cap \mathcal{I}_2 \cap \mathcal{I}_3 \cap \mathcal{I}_4$ is non-empty, and the right endpoint of $\mathcal{I}(\beta)$ is captured by that of $\mathcal{I}_1$ and $\mathcal{I}_4$ which approximates $2(1+\beta)/(1-\beta) \cdot 1/\gamma_{\max}$ when both κ and $\gamma_{\rm flat}/\gamma_{\max}$ are sufficiently small.

G.2. Proofs for Projections onto Flat and Sharp Directions

G.2.1. MAIN RESULT AND PROOF OUTLINE

Lemma G.2 (Dominance of the flat direction) *Suppose that the momentum iteration (GD-M) starts from an initial point $w_0 \in \mathcal{M}$ on the river and an initial momentum $m_0 = 0$. Under Assumptions 2.1 and 2.2, taking the learning rate η and momentum parameter β satisfying the conditions in Theorem G.1, then for any iteration $k \in \mathbb{N}$ and $\tau \in [0,1]$, the following conclusions hold simultaneously,*

$$\begin{aligned}
&\|\boldsymbol{P}_{\rm f}(w_{k,\tau})\nabla L(w_{k,\tau})\|_2 \ge \big(1 - \underline{q}_k\big) \cdot \|\boldsymbol{P}_{\rm f}(w_k)\nabla L(w_k)\|_2, \\
&b_k \cdot \|\boldsymbol{P}_{\rm f}(w_k)\nabla L(w_k)\|_2 \ge \left\|(1-\beta)\cdot\sum_{j=0}^{k}\beta^j \cdot \boldsymbol{P}_{\rm f}(w_k)\nabla L(w_k) - \boldsymbol{P}_{\rm f}(w_k)m_{k+1}\right\|_2, \\
&\max\Big\{\|\boldsymbol{P}_{\rm s}(w_{k,\tau})\nabla L(w_{k,\tau})\|_2, \|\boldsymbol{P}_{\rm s}(w_{k,\tau})m_{k+1}\|_2\Big\} \le c \cdot \|\boldsymbol{P}_{\rm f}(w_{k,\tau})\nabla L(w_{k,\tau})\|_2,
\end{aligned}$$

where b_k, $\underline{q}_k$, and c are defined as following,

$$\begin{aligned}
c &:= 6\mathsf{B}(\beta)\cdot\kappa^{1/2}, \\
b_k &:= 12\beta\cdot\eta\gamma_{\rm flat}\cdot\sum_{j=0}^{k}\beta^j\cdot\Big(1+2\eta\gamma_{\rm flat}\cdot\big(1+6\mathsf{C}(\beta)\big)\Big)^j, \\
\underline{q}_k &:= 2\eta\gamma_{\rm flat}\cdot\Big(1+2\mathsf{C}(\beta)\cdot\big(1+4\eta\gamma_{\rm flat}\cdot\big(1+6\mathsf{C}(\beta)\big)\big)\Big)\cdot(1-\beta)\cdot\sum_{j=0}^{k}\beta^j,
\end{aligned}$$

and the constants $\mathsf{B}(\beta)$ and $\mathsf{C}(\beta)$ are defined in Theorem G.1.

Proof [Proof of Lemma G.2] The key idea to prove Lemma G.2 is by induction over k on the relationships between four quantities: gradient in the sharp/flat direction and momentum in the sharp/flat direction. We prove the result for $\tau = 1$, and the result for general $\tau \in [0,1]$ can be obtained in the same manner.

Induction objective. We are going to inductively prove the following augmented results: for each $k \in \mathbb{N}$, the following four induction conditions hold:

- Firstly, the sharp direction does not explode:

$$\begin{aligned}
&\left\| \boldsymbol{P}_k^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_{k+1}) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_{k+1}) \nabla L(w_{k+1}) \end{pmatrix} \right\|_2 \\
&\quad \le \left(1 - \overline{q}_k\right) \cdot \left(1 + \mathsf{A}(\beta) \cdot \kappa^{3/16}\right) \cdot \left\| \boldsymbol{P}_{k-1}^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_{k-1}^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2 \\
&\qquad + \mathsf{B}(\beta) \cdot \kappa^{15/16} \cdot \left\| \begin{pmatrix} \beta \cdot m_k \\ (1-\beta) \cdot \nabla L(w_k) \end{pmatrix} \right\|_2 .
\end{aligned} \tag{G.4}$$

- Secondly, the flat direction gradient keeps its magnitude:

$$\|\boldsymbol{P}_{\mathrm{f}}(w_{k+1}) \nabla L(w_{k+1})\|_2 \ge (1 - \underline{q}_k) \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_k) \nabla L(w_k)\|_2. \tag{G.5}$$

- Thirdly, the flat direction momentum resembles the moving average of flat direction gradients:

$$b_k \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_k) \nabla L(w_k)\|_2 \ge \left\| (1-\beta) \cdot \sum_{j=0}^{k} \beta^j \cdot \boldsymbol{P}_{\mathrm{f}}(w_k) \nabla L(w_k) - \boldsymbol{P}_{\mathrm{f}}(w_k) m_{k+1} \right\|_2 . \tag{G.6}$$

- Lastly, the flat direction gradient dominates the sharp direction gradient and momentum:

$$\begin{aligned}
&\left\| \boldsymbol{P}_k^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_{k+1}) m_{k+1} \\ \boldsymbol{P}_{\mathrm{s}}(w_{k+1}) \nabla L(w_{k,\tau}) \end{pmatrix} \right\|_2 \le c \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_{k+1}) \nabla L(w_{k+1})\|_2, \\
&\max\left\{ \|\boldsymbol{P}_{\mathrm{s}}(w_{k+1}) \nabla L(w_{k+1})\|_2, \|\boldsymbol{P}_{\mathrm{s}}(w_{k+1}) m_{k+1}\|_2 \right\} \le c \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_{k+1}) \nabla L(w_{k+1})\|_2.
\end{aligned} \tag{G.7}$$

Additionally, the induction condition (G.7) holds for a dummy step $k = -1$. Here $\mathsf{A}(\beta)$ and $\mathsf{B}(\beta)$ are defined in Theorem G.1. The matrices $\{\boldsymbol{P}_k, \boldsymbol{O}, \boldsymbol{V}_k\}_{k \ge 0}$ are constructed in Lemma G.4: $\boldsymbol{V}_k$ duplicates the eigenbasis of the sharp-direction Hessian $\boldsymbol{P}_{\mathrm{s}}(w_k) \nabla^2 L(w_k) \boldsymbol{P}_{\mathrm{s}}(w_k)$ in the momentum-gradient product space, $\boldsymbol{O}$ is a fixed orthogonal matrix that groups each sharp eigen-direction into a two-dimensional momentum-gradient block, and $\boldsymbol{P}_k$ is the block-diagonal Lyapunov weight obtained from these two-dimensional blocks. The sequences $\{b_k\}_{k \ge 0}$, $\{\overline{q}_k\}_{k \ge 0}$, and $\{\underline{q}_k\}_{k \ge 0}$ are positive sequences, and $c > 0$ is a constant, defined as following:

$$\begin{aligned}
c &:= 6\mathsf{B}^2(\beta) \cdot \kappa^{1/2}, \\
b_k &:= 12\beta \cdot \eta\gamma_{\mathrm{flat}} \cdot \sum_{j=0}^{k} \beta^j \cdot \Big(1 + 2\eta\gamma_{\mathrm{flat}} \cdot \big(1 + 6\mathsf{C}(\beta)\big)\Big)^j, \\
\underline{q}_k &:= 2\eta\gamma_{\mathrm{flat}} \cdot \Big(1 + 2\mathsf{C}(\beta) \cdot \big(1 + 4\eta\gamma_{\mathrm{flat}} \cdot \big(1 + 6\mathsf{C}(\beta)\big)\big)\Big) \cdot (1-\beta) \cdot \sum_{j=0}^{k} \beta^j, \\
\overline{q}_k &:= \left(\max_{i \in [d-1]} \frac{(2\beta^3 - 2\beta^2 + 3\beta - 1) \cdot (\eta\lambda_{k,i})^2 + 2(1-\beta^2) \cdot \eta\lambda_{k,i} + 2(1-\beta)^2(1+\beta)}{\eta\lambda_{k,i}(1-\beta)^2 \cdot (2(1+\beta) - (1-\beta)\eta\lambda_{k,i})} \right)^{-1}.
\end{aligned}$$

In the sufficiently small river-valley regime of Assumption 2.2, all scalar error terms carrying the small factors $\gamma_{\mathrm{flat}}/\gamma_{\max}$ and κ are taken to be below a fixed constant; in particular, $\underline{q}_k \le 1/2$ for all relevant k.

Correctness for step $k = 0$. Firstly we show that all of (G.4), (G.5), (G.6), and (G.7) are correct for step $k = 0$, and additionally (G.7) hold for a dummy step $k = -1$.

- The correctness of (G.4) at step $k = 0$ is given by Lemma G.4 for step $k = 0$.
- The correctness of (G.6) at step $k = 0$ is trivial since the right hand side of (G.6) is 0.
- The correctness of (G.7) at step $k = -1$ is also trivial since both sides are 0.
- The correctness of (G.5) at step $k = 0$ conditioning on the correctness of (G.6) at step $k = 0$ and (G.7) at step $k = -1$ is given by Lemma G.7.
- Finally the correctness of (G.7) at step $k = 0$ conditioning on the correctness of (G.4), (G.5), and (G.6) at step $k = 0$ and (G.7) at step $k = -1$ is given by Lemma G.8.

Main induction argument. Now suppose that all of (G.4), (G.5), (G.6), and (G.7) are correct up to some step $k - 1$ with $k \geq 1$. To prove the correctness for step k, we adopt the following strategy:

- Condition (G.4) can be directly proved correct for all k (proved by Lemma G.4). Then:
- First prove that Condition (G.6) is correct (with Lemma G.6).
- Then prove that Condition (G.5) is correct (with Lemma G.7).
- Finally prove that Condition (G.7) is correct (with Lemma G.8).

In the sequel, we show how to obtain (G.5), (G.6), and (G.7) for step k following the above strategy.

Proof of (G.6) *for step* k. By Lemma G.6, given the correctness of (G.5), (G.6), and (G.7) up to step $k - 1$, we have that

$$\left\| (1-\beta) \cdot \sum_{j=0}^{k} \beta^j \cdot \boldsymbol{P}_{\mathrm{f}}(w_k) \nabla L(w_k) - \boldsymbol{P}_{\mathrm{f}}(w_k) m_{k+1} \right\|_2 \leq \widetilde{b}_k \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_k) \nabla L(w_k)\|_2, \quad \text{where}$$

$$\widetilde{b}_k := \big(1 + 2\underline{q}_{k-1}\big) \cdot \left(\beta b_{k-1} + \beta(1-\beta) \sum_{j=0}^{k-1} \beta^j \underline{q}_{k-1} + \kappa\eta\gamma\beta \left(1 + (1-\beta) \sum_{j=0}^{k-1} \beta^j + b_{k-1} \right) \right).$$

Then it suffices to prove that the above $\widetilde{b}_k \leq b_k$. We first gather some useful properties needed in this proof (and later induction proofs) in the following.

Proposition G.3 *Under the same setups as Lemma G.2, we have the following results:*

1. *Under the condition that $\eta \in \mathcal{I}_1$, we have that*

$$1 + 2\eta\gamma_{\max} + \frac{\eta\gamma}{2} \leq 1 + 5 \cdot \frac{1+\beta}{1-\beta} := \mathsf{C}(\beta). \tag{G.8}$$

2. *If we denote $q(\beta)$ as following*

$$q(\beta) := 2\eta\gamma_{\mathrm{flat}} \cdot \big(1 + 6\mathsf{C}(\beta)\big),$$

then by $\eta \in \mathcal{I}_2$, we have that

$$\eta \leq \frac{1}{2(1+6\mathsf{C}(\beta))} \cdot \frac{1}{\gamma_{\mathrm{flat}}} \quad \Rightarrow \quad q(\beta) \leq 1. \tag{G.9}$$

3. *By $\eta \in \mathcal{I}_3$, we have that for any $k \in \mathbb{N}$,*

$$\eta \le \frac{1-\beta}{12\beta + 2\beta\cdot(1+6\mathsf{C}(\beta))}\cdot\frac{1}{\gamma_{\text{flat}}} \quad \Rightarrow \quad b_k \le (1-\beta)\sum_{j=0}^{k}\beta^j \le 1. \tag{G.10}$$

4. *By $\eta \in \mathcal{I}_1 \cap \mathcal{I}_2 \cap \mathcal{I}_3$, we have that for any $k \in \mathbb{N}$,*

$$\underline{q}_k \le q(\beta). \tag{G.11}$$

Proof [Proof of Proposition G.3] Most conclusions can be directly checked from the definitions. For the third conclusion, recall that $q(\beta) = 2\eta\gamma_{\text{flat}}(1+6\mathsf{C}(\beta))$. Then

$$\begin{aligned} b_k &= 12\beta\eta\gamma_{\text{flat}}\sum_{j=0}^{k}\beta^j\big(1+q(\beta)\big)^j \\ &\le 12\beta\eta\gamma_{\text{flat}}\cdot\frac{1-\beta}{1-\beta(1+q(\beta))}\sum_{j=0}^{k}\beta^j \\ &\le (1-\beta)\sum_{j=0}^{k}\beta^j, \end{aligned}$$

where the last step uses $\eta \in \mathcal{I}_3$. This proves the third conclusion and the remaining conclusions follow from the definitions. ■

Now we are ready to prove $\widetilde{b}_k \le b_k$. Using (G.9), (G.10), and (G.11), we have the following,

$$\begin{aligned} \widetilde{b}_k &= \big(1+2\underline{q}_{k-1}\big)\cdot\left(\beta b_{k-1} + \beta(1-\beta)\sum_{j=0}^{k-1}\beta^j\underline{q}_{k-1} + \kappa\eta\gamma\beta\left(1+(1-\beta)\sum_{j=0}^{k-1}\beta^j + b_{k-1}\right)\right) \\ &\le \big(1+2q(\beta)\big)\cdot\left(\beta b_{k-1} + \beta(1-\beta)\sum_{j=0}^{k-1}\beta^j q(\beta) + \kappa\eta\gamma\beta\left(1+(1-\beta)\sum_{j=0}^{k-1}\beta^j + b_{k-1}\right)\right) \\ &\le 12\beta\cdot\eta\gamma_{\text{flat}}\cdot\sum_{j=1}^{k}\beta^j\cdot\big(1+q(\beta)\big)^j + \beta\cdot q(\beta)\cdot\big(1+2q(\beta)\big) + \beta\cdot\eta\gamma_{\text{flat}}\cdot 3\cdot\big(1+2q(\beta)\big) \\ &\le 12\beta\cdot\eta\gamma_{\text{flat}}\cdot\sum_{j=1}^{k}\beta^j\cdot\big(1+q(\beta)\big)^j + 12\beta\cdot\eta\gamma_{\text{flat}} \\ &= 12\beta\cdot\eta\gamma_{\text{flat}}\cdot\sum_{j=0}^{k}\beta^j\cdot\big(1+q(\beta)\big)^j \\ &= b_k, \end{aligned}$$

where the first inequality uses (G.11), the second inequality combines (i) the definition of b_{k-1} and (ii) (G.10) to bound $b_{k-1} \le 1$, and (iii) item 2 of Assumption 2.2 to obtain $\kappa\gamma \lesssim \gamma_{\text{flat}}$, the third inequality uses (G.9) to bound $q(\beta) \le 1$ and obtain $\eta\gamma_{\text{flat}} \le 1$. This proves that $\widetilde{b}_k \le b_k$ and thus the induction condition (G.6) hold for step k.

Proof of (G.5) *for step* k. By Lemma G.7, given the correctness of (G.5), (G.6), and (G.7) up to step $k-1$ plus (G.6) for step k, we have the following,

$$\|\boldsymbol{P}_{\mathrm{f}}(w_{k+1})\nabla L(w_{k+1})\|_2 \geq \big(1-\widetilde{\underline{q}}_k\big)\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2, \quad \text{where}$$

$$\widetilde{\underline{q}}_k := \eta\gamma_{\mathrm{flat}}\cdot\left((1-\beta)\sum_{j=0}^{k}\beta^j + b_k\right)$$
$$+ 2\eta\gamma\kappa\big(1+2\eta\gamma_{\max}+\eta\gamma/2\big)\cdot\left((1-\beta)+\beta\left((1-\beta)\sum_{j=0}^{k-1}\beta^j + b_{k-1}\right)\big(1+2\underline{q}_{k-1}\big)\right).$$

Then it suffices to prove that $\widetilde{\underline{q}}_k \leq \underline{q}_k$. To this end, consider the following,

$$\widetilde{\underline{q}}_k \leq \eta\gamma_{\mathrm{flat}}\cdot\left((1-\beta)\cdot\sum_{j=0}^{k}\beta^j + b_k\right.$$
$$\left.+ 2\mathsf{C}(\beta)\cdot\left((1-\beta)+\beta\cdot\left((1-\beta)\cdot\sum_{j=0}^{k-1}\beta^j + b_{k-1}\right)\cdot\big(1+2q(\beta)\big)\right)\right), \quad \text{(G.12)}$$

where we apply (i) item 2 of Assumption 2.2 to obtain that $\kappa\gamma \leq \gamma_{\mathrm{flat}}$, (ii) inequality (G.8) to bound $1+\eta\gamma_{\max}+\eta\gamma/2$, and (iii) inequality (G.11) to bound $\underline{q}_{k-1} \leq q(\beta)$. To proceed, notice that by the definition of b_k,

$$(1-\beta)\cdot\sum_{j=0}^{k}\beta^j + b_k = (1-\beta)\cdot\sum_{j=0}^{k}\beta^j$$
$$+ 12\beta\cdot\eta\gamma_{\mathrm{flat}}\cdot\sum_{j=0}^{k}\beta^j\cdot\Big(1+2\eta\gamma_{\mathrm{flat}}\cdot\big(1+6\mathsf{C}(\beta)\big)\Big)^j. \quad \text{(G.13)}$$

The second term on the right hand side of the above equality can be bounded by

$$12\beta\cdot\eta\gamma_{\mathrm{flat}}\cdot\sum_{j=0}^{k}\beta^j\cdot\Big(1+2\eta\gamma_{\mathrm{flat}}\cdot\big(1+6\mathsf{C}(\beta)\big)\Big)^j$$
$$\leq 12\beta\cdot\eta\gamma_{\mathrm{flat}}\cdot\max_{k'\in\mathbb{N}}\left\{\frac{\sum_{j=0}^{k'}\beta^j\cdot(1+2\eta\gamma_{\mathrm{flat}}\cdot(1+6\mathsf{C}(\beta)))^j}{\sum_{j=0}^{k'}\beta^j}\right\}\cdot\sum_{j=0}^{k}\beta^j$$
$$= 12\beta\cdot\eta\gamma_{\mathrm{flat}}\cdot\frac{1-\beta}{1-\beta\cdot(1+2\eta\gamma_{\mathrm{flat}}\cdot(1+6\mathsf{C}(\beta)))}\cdot\sum_{j=0}^{k}\beta^j$$
$$\leq (1-\beta)\cdot\sum_{j=0}^{k}\beta^j, \quad \text{(G.14)}$$

where the last inequality uses $\eta\in\mathcal{I}_3$. Consequently, by (G.13) and (G.14), we have that

$$(1-\beta)\cdot\sum_{j=0}^{k}\beta^j + b_k \leq 2(1-\beta)\cdot\sum_{j=0}^{k}\beta^j. \quad \text{(G.15)}$$

Therefore, combining (G.12) and (G.15), we have that

$$\begin{aligned}\widetilde{q}_k &\le \eta\gamma_{\text{flat}} \cdot \left(2(1-\beta)\cdot\sum_{j=0}^{k}\beta^j + 2\mathsf{C}(\beta)\left((1-\beta)+\beta\cdot 2(1-\beta)\cdot\sum_{j=0}^{k-1}\beta^j\right)\cdot\big(1+2q(\beta)\big)\right)\\ &\le \eta\gamma_{\text{flat}} \cdot \left(2(1-\beta)\cdot\sum_{j=0}^{k}\beta^j + 4\mathsf{C}(\beta)\cdot(1-\beta)\cdot\sum_{j=0}^{k}\beta^j\cdot\big(1+2q(\beta)\big)\right)\\ &= 2\eta\gamma_{\text{flat}} \cdot \Big(1+2\mathsf{C}(\beta)\cdot\big(1+2q(\beta)\big)\Big)\cdot(1-\beta)\cdot\sum_{j=0}^{k}\beta^j\\ &= \underline{q}_k\end{aligned}$$

This proves $\widetilde{q}_k \le \underline{q}_k$, finishing the proof of induction condition (G.5) for step k.

Proof of (G.7) *for step* k. This follows by Lemma G.8 without extra proofs.

Thus, we finish the **Main induction argument**, completing the proof of Lemma G.2. ■

G.2.2. LEMMAS FOR INDUCTION ARGUMENT

Lemma G.4 (Analysis of the sharp directions) *Suppose that Assumptions 2.1 and 2.2 hold. Take the learning rate* η *and momentum parameter* β *satisfying*

$$\eta \in \mathcal{I}_1 := \left(\frac{\kappa^{1/32}}{\gamma_{\max}},\ \frac{2\cdot(1+\beta)-\kappa^{1/16}}{1-\beta}\cdot\frac{1}{\gamma_{\max}}\right),$$

then there exist orthogonal matrices $\boldsymbol{O}, \{\boldsymbol{V}_k\}_{k\ge 0} \subset \mathbb{R}^{2d\times 2d}$ *and positive definite matrices* $\{\boldsymbol{P}_k\}_{k\ge 0} \subset \mathbb{R}^{2d\times 2d}$ *such that for any* $k \in \mathbb{N}$ *and* $\tau \in [0,1]$*, it holds that*

$$\begin{aligned}&\left\|\boldsymbol{P}_k^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_k^\top\begin{pmatrix}\boldsymbol{P}_{\text{s}}(w_{k,\tau})m_k\\ \boldsymbol{P}_{\text{s}}(w_{k,\tau})\nabla L(w_{k+1})\end{pmatrix}\right\|_2\\ &\quad\le \big(1-\overline{q}_k\big)\cdot\big(1+\mathsf{A}(\beta)\cdot\kappa^{3/16}\big)\cdot\left\|\boldsymbol{P}_{k-1}^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_{k-1}^\top\begin{pmatrix}\boldsymbol{P}_{\text{s}}(w_k)m_k\\ \boldsymbol{P}_{\text{s}}(w_k)\nabla L(w_k)\end{pmatrix}\right\|_2\\ &\qquad+\mathsf{B}(\beta)\cdot\kappa^{15/16}\cdot\left\|\begin{pmatrix}\beta\cdot m_k\\ (1-\beta)\cdot\nabla L(w_k)\end{pmatrix}\right\|_2.\end{aligned}$$

Here $\overline{q}_k$, $\mathsf{A}(\beta)$ *and* $\mathsf{B}(\beta)$ *are defined as*

$$\overline{q}_k := \left(\max_{i\in[d-1]}\frac{(2\beta^3-2\beta^2+3\beta-1)\cdot(\eta\lambda_{k,i})^2+2(1-\beta^2)\cdot\eta\lambda_{k,i}+2(1-\beta)^2(1+\beta)}{\eta\lambda_{k,i}(1-\beta)^2\cdot(2(1+\beta)-(1-\beta)\eta\lambda_{k,i})}\right)^{-1},$$

$$\mathsf{A}(\beta) := \frac{1}{2}\cdot\left(\frac{8\sqrt{2}}{(1-\beta)^4}+\frac{22}{(1-\beta)^2}\right)+d\cdot\left(\frac{64\sqrt{2}}{(1-\beta)^4}+\frac{88}{(1-\beta)^2}\right)+\left(\frac{8}{(1-\beta)^3}+3\right)^{\frac{1}{2}},$$

$$\mathsf{B}(\beta) := 2\cdot\left(\frac{16}{(1-\beta)^3}+6\right)^{\frac{1}{2}}\cdot\left(\frac{10}{1-\beta}+1\right)\cdot\frac{2}{1-\beta},$$

and $\lambda_{k,1} \geq \lambda_{k,2} \geq \cdots \geq \lambda_{k,d-1}$ are the largest $d-1$ eigenvalues of $\nabla^2 L(w_k)$. The matrices $\boldsymbol{V}_k$, $\boldsymbol{O}$, and $\boldsymbol{P}_k$ are constructed explicitly in Step 1 of the proof: $\boldsymbol{V}_k$ diagonalizes the sharp Hessian, $\boldsymbol{O}$ regroups the momentum and gradient coordinates into independent two-dimensional blocks, and $\boldsymbol{P}_k$ is assembled from the Lyapunov solutions for those blocks.

Proof [Proof of Lemma G.4] By the iteration of GD momentum (GD-M), we have that

$$\boldsymbol{P}_{\mathrm{s}}(w_{k+1})m_{k+1} = \beta \cdot \boldsymbol{P}_{\mathrm{s}}(w_k)m_k + (1-\beta)\boldsymbol{P}_{\mathrm{s}}(w_k)\nabla L(w_k) + \big(\boldsymbol{P}_{\mathrm{s}}(w_{k+1}) - \boldsymbol{P}_{\mathrm{s}}(w_k)\big)m_{k+1},$$

and that

$$\begin{aligned}
&\boldsymbol{P}_{\mathrm{s}}(w_{k+1})\nabla L(w_{k+1}) \\
&\quad = \boldsymbol{P}_{\mathrm{s}}(w_k)\nabla L(w_k) - \eta \cdot \boldsymbol{P}_{\mathrm{s}}(w_k)\nabla^2 L(w_k)\boldsymbol{P}_{\mathrm{s}}(w_k)m_{k+1} \\
&\qquad - \eta \cdot \int_0^1 \Big(\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})\nabla^2 L(w_{k,\tau})\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau}) - \boldsymbol{P}_{\mathrm{s}}(w_k)\nabla^2 L(w_k)\boldsymbol{P}_{\mathrm{s}}(w_k)\Big)m_{k+1}\mathrm{d}\tau \\
&\qquad - \eta \cdot \int_0^1 \nabla \boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})[m_{k+1}]\nabla L(w_{k,\tau})\mathrm{d}\tau.
\end{aligned}$$

Therefore, we have the following iteration formula,

$$\begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_{k+1})m_{k+1} \\ \boldsymbol{P}_{\mathrm{s}}(w_{k+1})\nabla L(w_{k+1}) \end{pmatrix} = \boldsymbol{T}_k \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k)m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k)\nabla L(w_k) \end{pmatrix} + \boldsymbol{E}_k \tag{G.16}$$

where $\boldsymbol{T}_k$ and $\boldsymbol{E}_k$ are defined as following,

$$\begin{aligned}
\boldsymbol{T}_k &:= \begin{pmatrix} \beta \cdot \boldsymbol{P}_{\mathrm{s}}(w_k) & (1-\beta) \cdot \boldsymbol{P}_{\mathrm{s}}(w_k) \\ -\eta\beta \cdot \boldsymbol{P}_{\mathrm{s}}(w_k)\nabla^2 L(w_k)\boldsymbol{P}_{\mathrm{s}}(w_k) & \boldsymbol{P}_{\mathrm{s}}(w_k)\big(\boldsymbol{I}_d - \eta(1-\beta)\cdot\nabla^2 L(w_k)\big)\boldsymbol{P}_{\mathrm{s}}(w_k) \end{pmatrix}, \\
\boldsymbol{E}_k &:= \begin{pmatrix} E_{k,1} \\ E_{k,2} \end{pmatrix},
\end{aligned}$$

and $E_{k,1}$ and $E_{k,2}$ are defined as following respectively,

$$\begin{aligned}
E_{k,1} &:= \big(\boldsymbol{P}_{\mathrm{s}}(w_{k+1}) - \boldsymbol{P}_{\mathrm{s}}(w_k)\big)m_{k+1} \\
E_{k,2} &:= -\eta \cdot \int_0^1 \Big(\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})\nabla^2 L(w_{k,\tau})\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau}) - \boldsymbol{P}_{\mathrm{s}}(w_k)\nabla^2 L(w_k)\boldsymbol{P}_{\mathrm{s}}(w_k)\Big)m_{k+1}\mathrm{d}\tau \\
&\qquad - \eta \cdot \int_0^1 \nabla \boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})[m_{k+1}]\nabla L(w_{k,\tau})\mathrm{d}\tau.
\end{aligned}$$

Step 1: Analysis of matrix $\boldsymbol{T}_k$. We denote the eigenvalues of the matrix $\boldsymbol{P}_{\mathrm{s}}(w_k)\nabla^2 L(w_k)\boldsymbol{P}_{\mathrm{s}}(w_k)$ as $\lambda_{k,1} \geq \lambda_{k,2} \geq \cdots \geq \lambda_{k,d}$, and we denote the corresponding orthonormal eigenvectors as $\{v_{k,i}\}_{i=1}^d$. Specifically, we have that $\lambda_{k,d} = 0$ and that $v_{k,i} = v_i(\nabla^2 L(w_k))$ for any $i \in [d]$. Then we denote the orthogonal matrix $\boldsymbol{V}_k \in \mathbb{R}^{2d\times 2d}$ and a diagonal matrix $\boldsymbol{\Lambda}_k \in \mathbb{R}^{d\times d}$ as

$$\boldsymbol{V}_k := \begin{pmatrix} v_{k,1}, \cdots, v_{k,d} & \boldsymbol{0} \\ \boldsymbol{0} & v_{k,1}, \cdots, v_{k,d}, \end{pmatrix}, \quad \boldsymbol{\Lambda}_k := \mathrm{diag}\Big(\lambda_{k,1}, \cdots, \lambda_{k,d-1}, 0\Big)$$

which allows us to transform the matrix $\boldsymbol{T}_k$ to

$$\boldsymbol{T}_k = \boldsymbol{V}_k \begin{pmatrix} \beta \cdot \boldsymbol{I}_{1:d-1} & (1-\beta) \cdot \boldsymbol{I}_{1:d-1} \\ -\eta\beta \cdot \boldsymbol{\Lambda}_k & \boldsymbol{I}_{1:d-1} - \eta(1-\beta) \cdot \boldsymbol{\Lambda}_k \end{pmatrix} \boldsymbol{V}_k^\top.$$

Furthermore, we can find another constant orthogonal matrix $\boldsymbol{O} \in \mathbb{R}^{2d\times 2d}$ such that

$$\begin{aligned} \boldsymbol{T}_k &= \boldsymbol{V}_k \boldsymbol{O} \mathrm{diag}\Big(\boldsymbol{T}_{k,1}, \cdots, \boldsymbol{T}_{k,d-1}, \boldsymbol{0}_{2\times 2}\Big) \boldsymbol{O}^\top \boldsymbol{V}_k^\top, \quad \text{where} \\ \boldsymbol{T}_{k,i} &= \begin{pmatrix} \beta & (1-\beta) \\ -\beta \cdot \eta\lambda_{k,i} & 1-(1-\beta)\cdot\eta\lambda_{k,i} \end{pmatrix} \in \mathbb{R}^{2\times 2}, \quad \forall i \in [d-1]. \end{aligned}$$

Now for each block $\boldsymbol{T}_{k,i}$ with $i \in [d-1]$, consider the following Lyapunov equation,

$$\boldsymbol{T}_{k,i}^\top \boldsymbol{P}_{k,i} \boldsymbol{T}_{k,i} = \boldsymbol{P}_{k,i} - \boldsymbol{I}_2.$$

By $\eta \in \mathcal{I}_1$, especially using $\eta\gamma_{\max} < 2(1+\beta)/(1-\beta)$, the spectral radius of $\boldsymbol{T}_{k,i}$ satisfies that $\rho(\boldsymbol{T}_{k,i}) < 1$ for any $k \geq 0$ and $i \in [d-1]$ (see Proposition H.2). Therefore, the above Lyapunov equation admits a unique positive definite solution $\boldsymbol{P}_{k,i} \in \mathbb{R}^{2\times 2}$, given by

$$\boldsymbol{P}_{k,i} = \sum_{j=0}^{+\infty} \boldsymbol{T}_{k,i}^\top \boldsymbol{T}_{k,i} = \begin{pmatrix} \frac{2(1-\beta)+(2\beta^3-2\beta^2+3\beta-1)\eta\lambda_{k,i}}{(1-\beta)^2\cdot(2(1+\beta)-(1-\beta)\eta\lambda_{k,i})} & \frac{-\beta(2\beta-\eta\lambda_{k,i})}{(1-\beta)\cdot(2(1+\beta)-(1-\beta)\eta\lambda_{k,i})} \\ \frac{-\beta(2\beta-\eta\lambda_{k,i})}{(1-\beta)\cdot(2(1+\beta)-(1-\beta)\eta\lambda_{k,i})} & \frac{-2(-1+\beta^2-\beta\eta\lambda_{k,i})}{\eta\lambda_{k,i}(1-\beta)\cdot(2(1+\beta)-(1-\beta)\eta\lambda_{k,i})} \end{pmatrix} \succ \boldsymbol{I}_2.$$

Also, from the Lyapunov equation we can write $\boldsymbol{P}_{k,i}^{\frac{1}{2}} \boldsymbol{T}_{k,i} = \boldsymbol{Q}_{k,i} (\boldsymbol{P}_{k,i} - \boldsymbol{I}_2)^{\frac{1}{2}}$ for some othogonal matrix $\boldsymbol{Q}_{k,i}$. Now we combine these $\boldsymbol{P}_{k,i}$'s to form a block diagonal matrix $\boldsymbol{P}_k \in \mathbb{R}^{2d\times 2d}$ as

$$\boldsymbol{P}_k := \mathrm{diag}\Big(\boldsymbol{P}_{k,1}, \cdots, \boldsymbol{P}_{k,d-1}, \boldsymbol{I}_2\Big).$$

Step 2: Bounding the next step in the sharp direction: main analysis. With the analysis of matrix $\boldsymbol{T}_k$, we can continue the analysis for upper bounding the sharp direction. We have

$$\begin{aligned} \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_{k+1}) m_{k+1} \\ \boldsymbol{P}_{\mathrm{s}}(w_{k+1}) \nabla L(w_{k+1}) \end{pmatrix} &= \boldsymbol{T}_k \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} + \boldsymbol{E}_k \\ &= \boldsymbol{V}_k \boldsymbol{O} \mathrm{diag}\Big(\boldsymbol{T}_{k,1}, \cdots, \boldsymbol{T}_{k,d-1}, \boldsymbol{0}_{2\times 2}\Big) \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} + \boldsymbol{E}_k, \end{aligned}$$

and thus

$$\begin{aligned} &\boldsymbol{P}_k^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_{k+1}) m_{k+1} \\ \boldsymbol{P}_{\mathrm{s}}(w_{k+1}) \nabla L(w_{k+1}) \end{pmatrix} \\ &\quad = \boldsymbol{P}_k^{\frac{1}{2}} \mathrm{diag}\Big(\boldsymbol{T}_{k,1}, \cdots, \boldsymbol{T}_{k,d-1}, \boldsymbol{0}_{2\times 2}\Big) \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} + \boldsymbol{P}_k^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \boldsymbol{E}_k. \end{aligned}$$

Consider the $\|\cdot\|_2$-norm on both sides, we have

$$\left\| \boldsymbol{P}_k^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_{k+1}) m_{k+1} \\ \boldsymbol{P}_{\mathrm{s}}(w_{k+1}) \nabla L(w_{k+1}) \end{pmatrix} \right\|_2$$

$$\leq \left\| \boldsymbol{P}_k^{\frac{1}{2}} \operatorname{diag}\Big(\boldsymbol{T}_{k,1}, \cdots, \boldsymbol{T}_{k,d-1}, \boldsymbol{0}_{2\times 2}\Big) \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2 + \left\| \boldsymbol{P}_k^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \boldsymbol{E}_k \right\|_2 .$$

In **Step 2** we focus on the first term on the right-hand side. By the definition of $\boldsymbol{P}_k$, we have

$$\begin{aligned}
&\left\| \boldsymbol{P}_k^{\frac{1}{2}} \operatorname{diag}\Big(\boldsymbol{T}_{k,1}, \cdots, \boldsymbol{T}_{k,d-1}, \boldsymbol{0}_{2\times 2}\Big) \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2^2 \\
&\quad = \sum_{i=1}^{d-1} \left\| \boldsymbol{P}_{k,i}^{\frac{1}{2}} \boldsymbol{T}_{k,i} (\boldsymbol{V}_k \boldsymbol{O})_i^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2^2 \\
&\quad = \sum_{i=1}^{d} \left\| \boldsymbol{Q}_{k,i} (\boldsymbol{P}_{k,i} - \boldsymbol{I}_2)^{\frac{1}{2}} (\boldsymbol{V}_k \boldsymbol{O})_i^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2^2 \\
&\quad = \sum_{i=1}^{d} \left\| \boldsymbol{Q}_{k,i} (\boldsymbol{I}_2 - \boldsymbol{P}_{k,i}^{-1})^{\frac{1}{2}} \boldsymbol{P}_{k,i}^{\frac{1}{2}} (\boldsymbol{V}_k \boldsymbol{O})_i^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2^2 \\
&\quad \leq \max_{i\in[d-1]} \left\| \boldsymbol{I}_2 - \boldsymbol{P}_{k,i}^{-1} \right\|_{\mathrm{Op}} \cdot \sum_{i=1}^{d} \left\| \boldsymbol{P}_{k,i}^{\frac{1}{2}} (\boldsymbol{V}_k \boldsymbol{O})_i^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2^2 \\
&\quad = \max_{i\in[d-1]} \left\| \boldsymbol{I}_2 - \boldsymbol{P}_{k,i}^{-1} \right\|_{\mathrm{Op}} \cdot \left\| \boldsymbol{P}_k^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2^2 \\
&\quad = \max_{i\in[d-1]} \Big(1 - \left\| \boldsymbol{P}_{k,i} \right\|_{\mathrm{Op}}^{-1}\Big) \cdot \left\| \boldsymbol{P}_k^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2^2 .
\end{aligned} \tag{G.17}$$

In the following, we upper bound the right hand side of (G.17). On the one hand, we have that

$$\max_{i\in[d-1]} \Big(1 - \left\| \boldsymbol{P}_{k,i} \right\|_{\mathrm{Op}}^{-1}\Big) \leq 1 - \Big(\max_{i\in[d-1]} \operatorname{Tr}\big(\boldsymbol{P}_{k,i}\big)\Big)^{-1} := 1 - \overline{q}_k$$

where we define $\overline{q}_k := (\max_{i\in[d-1]} \operatorname{Tr}(\boldsymbol{P}_{k,i}))^{-1}$ and $\operatorname{Tr}(\boldsymbol{P}_{k,i})$ is given by

$$\begin{aligned}
\operatorname{Tr}(\boldsymbol{P}_{k,i}) &= \frac{2(1-\beta) + (2\beta^3 - 2\beta^2 + 3\beta - 1)\eta\lambda_{k,i}}{(1-\beta)^2 \cdot (2(1+\beta) - (1-\beta)\eta\lambda_{k,i})} - \frac{2(-1+\beta^2 - \beta\eta\lambda_{k,i})}{\eta\lambda_{k,i}(1-\beta)\cdot(2(1+\beta) - (1-\beta)\eta\lambda_{k,i})} \\
&= \frac{(2\beta^3 - 2\beta^2 + 3\beta - 1)\cdot(\eta\lambda_{k,i})^2 + 2(1-\beta^2)\cdot\eta\lambda_{k,i} + 2(1-\beta)^2(1+\beta)}{\eta\lambda_{k,i}(1-\beta)^2\cdot(2(1+\beta) - (1-\beta)\eta\lambda_{k,i})}.
\end{aligned}$$

On the other hand, we have that

$$\begin{aligned}
&\left\| \boldsymbol{P}_k^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2 \\
&\quad \leq \left\| (\boldsymbol{P}_k^{\frac{1}{2}} - \boldsymbol{P}_{k-1}^{\frac{1}{2}}) \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2 + \left\| \boldsymbol{P}_{k-1}^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2 .
\end{aligned} \tag{G.18}$$

For the first term on the right-hand side of (G.18), we have

$$\left\| (\boldsymbol{P}_k^{\frac{1}{2}} - \boldsymbol{P}_{k-1}^{\frac{1}{2}}) \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2$$

$$
\begin{aligned}
&\le \left\|\boldsymbol{P}_k^{\frac{1}{2}} - \boldsymbol{P}_{k-1}^{\frac{1}{2}}\right\|_{\mathrm{Op}} \cdot \left\| \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2 \\
&\le \frac{1}{2} \cdot \left\|\boldsymbol{P}_k - \boldsymbol{P}_{k-1}\right\|_{\mathrm{Op}} \cdot \left\| \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2 \\
&\le \frac{1}{2} \cdot \left\|\boldsymbol{P}_k - \boldsymbol{P}_{k-1}\right\|_{\mathrm{Op}} \cdot \left\| \boldsymbol{P}_{k-1}^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_{k-1}^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2,
\end{aligned}
$$

where in the second and the third inequalities we use the fact that $\boldsymbol{P}_k \succeq \boldsymbol{I}_{2d}$. Furthermore, the difference between $\boldsymbol{P}_k$ and $\boldsymbol{P}_{k-1}$ can be bounded as

$$
\boldsymbol{P}_{k,i} - \boldsymbol{P}_{k-1,i} = \frac{\eta(\lambda_{k,i} - \lambda_{k-1,i})}{D(\eta\lambda_{k,i}) D(\eta\lambda_{k-1,i})} \cdot \begin{pmatrix} \frac{4\beta^2(1+\beta^2)}{(1-\beta)^2} & \frac{2\beta(1+\beta^2)}{1-\beta} \\ \frac{2\beta(1+\beta^2)}{1-\beta} & \frac{2(\beta\eta^2\lambda_{k,i}\lambda_{k-1,i} + (1-\beta^2)\eta(\lambda_{k,i}+\lambda_{k-1,i}) - 2(1+\beta)^2)}{\eta\lambda_{k,i}\cdot\eta\lambda_{k-1,i}} \end{pmatrix}.
$$

Here for simplicity we denote $D(x) := 2(1+\beta) - (1-\beta)x$ for any $x \in \mathbb{R}$. This allows us to bound the operator norm of the difference as (see details in Proposition H.3),

$$
\begin{aligned}
&\left\|\boldsymbol{P}_k - \boldsymbol{P}_{k-1}\right\|_{\mathrm{Op}} \\
&\quad \le \max_{i\in[d-1]} \left\|\boldsymbol{P}_{k,i} - \boldsymbol{P}_{k-1,i}\right\|_{\mathrm{Op}} \\
&\quad \le \max_{i\in[d-1]} \left\|\boldsymbol{P}_{k,i} - \boldsymbol{P}_{k-1,i}\right\|_{\mathrm{F}} \\
&\quad \le \max_{i\in[d-1]} \frac{\eta\left|\lambda_{k,i} - \lambda_{k-1,i}\right|}{D(\eta\lambda_{k,i}) D(\eta\lambda_{k-1,i})} \\
&\qquad \cdot \sqrt{\frac{8\beta^2(1+\beta^2)^2(3\beta^2 - 2\beta + 1)}{(1-\beta)^4} + \frac{4(\beta\eta^2\lambda_{k,i}\lambda_{k-1,i} + (1-\beta^2)\eta(\lambda_{k,i}+\lambda_{k-1,i}) - 2(1+\beta)^2)^2}{(\eta\lambda_{k,i})^2(\eta\lambda_{k-1,i})^2}}.
\end{aligned}
$$

By $\eta \in \mathcal{I}_1$, we have that

$$
\begin{aligned}
\eta \le \frac{2(1+\beta) - \kappa^{1/16}}{(1-\beta)\cdot\gamma_{\max}} \le \frac{2(1+\beta) - \kappa^{1/16}}{(1-\beta)\cdot\lambda_{k,i}}, \quad \forall k \ge 0,\ i \in [d-1], \\
\eta \ge \frac{\kappa^{1/32}}{\gamma_{\max}} \ge \frac{\kappa^{1/32}}{\lambda_{k,i}} \cdot \frac{\gamma}{\gamma_{\max}} \ge \frac{\kappa^{1/16}}{\lambda_{k,i}}, \quad \forall k \ge 0,\ i \in [d-1],
\end{aligned}
$$

where we have used item 2 of Assumption 2.2 that $\gamma/\gamma_{\max} \ge \kappa^{1/32}$. Therefore, we can lower bound

$$
D(\eta\lambda_{k,i}) \ge \kappa^{1/16}, \quad \eta\lambda_{k,i} \ge \kappa^{1/16}, \quad \forall k \ge 0,\ i \in [d-1].
$$

Thus we further simplify the difference as

$$
\begin{aligned}
&\left\|\boldsymbol{P}_k - \boldsymbol{P}_{k-1}\right\|_{\mathrm{Op}} \\
&\quad \le \max_{i\in[d-1]} \frac{\eta\cdot|\lambda_{k,i} - \lambda_{k-1,i}|}{\kappa^{1/8}} \cdot \left(\frac{2\sqrt{2}\beta(1+\beta^2)\sqrt{3\beta^2 - 2\beta + 1}}{(1-\beta)^2} + 2\beta + \frac{4(1-\beta^2)}{\kappa^{1/16}} + \frac{4(1+\beta)^2}{\kappa^{1/8}} \right)
\end{aligned}
$$

$$\leq \max_{i\in[d-1]} \eta\cdot|\lambda_{k,i}-\lambda_{k-1,i}|\cdot\kappa^{-1/4}\cdot\left(\frac{2\sqrt{2}\beta(1+\beta^2)\sqrt{3\beta^2-2\beta+1}}{(1-\beta)^2}+2\beta+4(1-\beta^2)+4(1+\beta)^2\right)$$

$$\leq \max_{i\in[d-1]} \eta\cdot|\lambda_{k,i}-\lambda_{k-1,i}|\cdot\kappa^{-1/4}\cdot\left(\frac{8\sqrt{2}}{(1-\beta)^2}+22\right).$$

Now by Lemma H.11, we have the following,

$$\left|\lambda_{k,i}-\lambda_{k-1,i}\right|\leq\left\|\nabla^2 L(w_k)-\nabla^2 L(w_{k-1})\right\|_{\mathrm{Op}}=\left\|\eta\cdot\int_0^1\nabla^3 L(w_{k-1,\tau})[m_k]\mathrm{d}\tau\right\|_{\mathrm{Op}}\leq\frac{\eta\gamma^2\kappa}{2},$$

where we have applied item 3 of Assumption 2.2 and the fact that $\|m_k\|_2\leq g_{\max}$ in the last inequality. Then we can bound the difference between $\boldsymbol{P}_k$ and $\boldsymbol{P}_{k-1}$ as

$$\left\|\boldsymbol{P}_k-\boldsymbol{P}_{k-1}\right\|_{\mathrm{Op}}\leq\eta^2\gamma^2\cdot\left(\frac{4\sqrt{2}}{(1-\beta)^2}+11\right)\cdot\kappa^{3/4}.$$

Consequently, the first term on the right-hand side of (G.18) is bounded as

$$\begin{aligned}&\left\|(\boldsymbol{P}_k^{\frac{1}{2}}-\boldsymbol{P}_{k-1}^{\frac{1}{2}})\boldsymbol{O}^\top\boldsymbol{V}_k^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}(w_k)m_k\\ \boldsymbol{P}_{\mathrm{s}}(w_k)\nabla L(w_k)\end{pmatrix}\right\|_2\\ &\quad\leq\frac{1}{2}\eta^2\gamma^2\cdot\left(\frac{4\sqrt{2}}{(1-\beta)^2}+11\right)\cdot\kappa^{3/4}\cdot\left\|\boldsymbol{P}_{k-1}^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_{k-1}^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}(w_k)m_k\\ \boldsymbol{P}_{\mathrm{s}}(w_k)\nabla L(w_k)\end{pmatrix}\right\|_2.\end{aligned}$$

Then we handle the second term on the right-hand side of (G.18). We want to bound the difference between

$$\left\|\boldsymbol{P}_{k-1}^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_k^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|_2\quad\text{and}\quad\left\|\boldsymbol{P}_{k-1}^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_{k-1}^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|_2.$$

Consider the following twisted approach based on Lemma H.12 (Davis–Kahan $\sin\theta$ theorem). [2] consider a clustering mechanism to make sure that the eigenvalues of Hessian in different groups are at least η away, and consecutive eigenvalues in a group is no more than η away. In this manner, the gap condition in Lemma H.12 is satisfied because we can lower bound the gap between eigenvalues from two groups using $|\lambda_{k,i}-\lambda_{k-1,i}|\leq O(\eta)$ and that the groups at step $k-1$ is already η away. However, such a clustering mechanism is not friendly in our setting because the eigenvalues within a group could differ a lot (up to $d\cdot\eta$) and can make the weighting matrix $\boldsymbol{P}_{k-1}^{1/2}$ very heterogeneous within a cluster. To tackle this challenge, we observe that here we have already bounded the eigenvalues between time steps via $|\lambda_{k,i}-\lambda_{k-1,i}|\leq O(\eta\gamma^2\cdot\kappa)$ which is small. So we can actually narrow the grouping threshold to something related to κ. In this way, the difference between eigenvalues within a group can be controlled by κ, which means that within a group we reduce back to the setting of [2].

To make it precise, consider the following clustering of the eigenvalues of the Hessian. At time step k, we devide $[d]$ into $p(k-1)$ disjoint groups $S_1,\cdots,S_{p(k-1)}$ such that

$$\forall\ell,h\in[p(k-1)],\ell\neq h,\quad\text{and}\quad i\in S_\ell,j\in S_h,\quad|\lambda_{k-1,i}-\lambda_{k-1,j}|\geq2\eta\gamma^2\kappa^{1/2},$$

$$\forall \ell \in [p(k-1)], \quad \text{and} \quad i,j \in S_\ell, \quad |\lambda_{k-1,i} - \lambda_{k-1,j}| < 2\eta\gamma^2\kappa^{1/2}.$$

Then on the one hand, since $|\lambda_{k,i} - \lambda_{k-1,i}| \le \eta\gamma^2\kappa/2 \le \eta\gamma^2\kappa^{1/2}/2$, we can obtain that the Δ in Lemma H.12 is lower bounded by $2\eta\gamma^2\kappa^{1/2} - 2\cdot\eta\gamma^2\kappa^{1/2}/2 = \eta\gamma^2\kappa^{1/2}$, and thus by Lemma H.12 we have that for any $\ell \in [p(k)]$, it holds that

$$\left\|\widetilde{\boldsymbol{V}}_{k,S_\ell}\widetilde{\boldsymbol{V}}_{k,S_\ell}^\top - \widetilde{\boldsymbol{V}}_{k-1,S_\ell}\widetilde{\boldsymbol{V}}_{k-1,S_\ell}^\top\right\|_{\mathrm{Op}} \le \frac{\left\|\nabla^2 L\left(w_k\right) - \nabla^2 L\left(w_{k-1}\right)\right\|_{\mathrm{Op}}}{\Delta} \le \frac{\eta\gamma^2\kappa/2}{\eta\gamma^2\kappa^{1/2}} = \frac{1}{2}\kappa^{1/2}.$$

Here we denote $\widetilde{\boldsymbol{V}}_{k,S} := (v_{k,i})_{i\in S} \in \mathbb{R}^{d\times|S|}$ for any index set $S \subseteq [d]$. On the other hand, within a group S_ℓ the difference between any two eigenvalues is bounded by $2d\eta\gamma^2\cdot\kappa^{1/2}$. Denoting $\widehat{\boldsymbol{P}}_{k-1}$ as

$$\begin{aligned}
\widehat{\boldsymbol{P}}_{k-1} &:= \operatorname{diag}\Big(\widehat{\boldsymbol{P}}_{k-1,S_1}, \cdots, \widehat{\boldsymbol{P}}_{k-1,S_{p(k-1)}}\Big), \quad \text{where} \\
\widehat{\boldsymbol{P}}_{k-1,S_\ell} &:= \operatorname{diag}\Big(\underbrace{\boldsymbol{P}_{k-1,i_\ell}, \cdots, \boldsymbol{P}_{k-1,i_\ell}}_{|S_\ell|}\Big), \quad \forall \ell \in [p(k-1)],
\end{aligned}$$

where $i_\ell \in S_\ell$ denotes an arbitrary index in group S_ℓ, we can bound the difference between $\boldsymbol{P}_{k-1}$ and $\widehat{\boldsymbol{P}}_{k-1}$ as (see details in Proposition H.3),

$$\begin{aligned}
\left\|\boldsymbol{P}_{k-1} - \widehat{\boldsymbol{P}}_{k-1}\right\|_{\mathrm{Op}} &\le \max_{\ell\in[p(k-1)]}\max_{i,j\in S_\ell}\left\|\boldsymbol{P}_{k-1,i} - \boldsymbol{P}_{k-1,j}\right\|_{\mathrm{Op}} \\
&\le \max_{\ell\in[p(k-1)]}\max_{i,j\in S_\ell}\eta\cdot|\lambda_{k-1,i} - \lambda_{k-1,j}|\cdot\kappa^{-1/4}\cdot\left(\frac{8\sqrt{2}}{(1-\beta)^2} + 22\right) \\
&\le d\eta^2\gamma^2\cdot\left(\frac{16\sqrt{2}}{(1-\beta)^2} + 44\right)\cdot\kappa^{1/4}. \qquad \text{(G.19)}
\end{aligned}$$

Then we can proceed as follows. Consider that by (G.19) we have

$$\begin{aligned}
&\left\|\boldsymbol{P}_{k-1}^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_k^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|_2 - \left\|\boldsymbol{P}_{k-1}^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_{k-1}^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|_2 \\
&\le \left\|\widehat{\boldsymbol{P}}_{k-1}^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_k^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|_2 - \left\|\widehat{\boldsymbol{P}}_{k-1}^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_{k-1}^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|_2 \\
&\quad + d\eta^2\gamma^2\cdot\left(\frac{32\sqrt{2}}{(1-\beta)^2} + 88\right)\cdot\kappa^{1/4}\cdot\left\|\boldsymbol{P}_{k-1}^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_{k-1}^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|_2 \qquad \text{(G.20)}
\end{aligned}$$

Now for the first term on the right-hand side above, we can decompose the norm by

$$\begin{aligned}
&\left\|\widehat{\boldsymbol{P}}_{k-1}^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_k^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|_2 \\
&= \left(\sum_{\ell\in[p(k-1)]}\left\|\widehat{\boldsymbol{P}}_{k-1,S_\ell}^{\frac{1}{2}}\boldsymbol{O}_{S_\ell}^\top\boldsymbol{V}_{k,S_\ell}^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|_2^2\right)^{\frac{1}{2}} \\
&= \left(\sum_{\ell\in[p(k-1)]}\left\|\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|^2_{\boldsymbol{P}_{k-1,i_\ell}\otimes\widetilde{\boldsymbol{V}}_{k,S_\ell}\widetilde{\boldsymbol{V}}_{k,S_\ell}^\top}\right)^{\frac{1}{2}},
\end{aligned}$$

where $\boldsymbol{O}_{S_\ell}$ is an orthogonal matrix corresponding to the transformation of $\boldsymbol{O}$ on the subspace spanned by the eigenvectors in group S_ℓ. Similarly, we have that

$$\begin{aligned}
&\left\|\widehat{\boldsymbol{P}}_{k-1}^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_{k-1}^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|_2\\
&\quad=\left(\sum_{\ell\in[p(k-1)]}\left\|\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|^2_{\boldsymbol{P}_{k-1,i_\ell}\otimes\widetilde{\boldsymbol{V}}_{k-1,S_\ell}\widetilde{\boldsymbol{V}}_{k-1,S_\ell}^\top}\right)^{\frac{1}{2}}.
\end{aligned}$$

Thus we can bound the difference between the two terms by

$$\begin{aligned}
&\left\|\widehat{\boldsymbol{P}}_{k-1}^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_{k}^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|_2-\left\|\widehat{\boldsymbol{P}}_{k-1}^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_{k-1}^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|_2\\
&\quad\le\left(\sum_{\ell\in[p(k-1)]}\left\|\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|^2_{\boldsymbol{P}_{k-1,i_\ell}\otimes\left(\widetilde{\boldsymbol{V}}_{k,S_\ell}\widetilde{\boldsymbol{V}}_{k,S_\ell}^\top-\widetilde{\boldsymbol{V}}_{k-1,S_\ell}\widetilde{\boldsymbol{V}}_{k-1,S_\ell}^\top\right)}\right)^{\frac{1}{2}}\\
&\quad\le\left(\sum_{\ell\in[p(k-1)]}\left\|\boldsymbol{P}_{k-1,i_\ell}\right\|_{\mathrm{Op}}\cdot\left\|\widetilde{\boldsymbol{V}}_{k,S_\ell}\widetilde{\boldsymbol{V}}_{k,S_\ell}^\top-\widetilde{\boldsymbol{V}}_{k-1,S_\ell}\widetilde{\boldsymbol{V}}_{k-1,S_\ell}^\top\right\|_{\mathrm{Op}}\right.\\
&\qquad\left.\cdot\left\|\boldsymbol{P}_{k-1,S_\ell}^{\frac{1}{2}}\boldsymbol{O}_{S_\ell}^\top\boldsymbol{V}_{k-1,S_\ell}^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|_2^2\right)^{\frac{1}{2}}.
\end{aligned}$$

Specifically we have that for any $\ell\in p(k-1)$,

$$\begin{aligned}
\left\|\boldsymbol{P}_{k-1,i_\ell}\right\|_{\mathrm{Op}}&\le\frac{(2\beta^3-2\beta^2+3\beta-1)\cdot(\eta\lambda_{k,i})^2+2(1-\beta)^2\cdot\eta\lambda_{k,i}+2(1-\beta)^2(1+\beta)}{\eta\lambda_{k,i}(1-\beta)^2\cdot(2(1+\beta)-(1-\beta)\eta\lambda_{k,i})}\\
&\le\frac{4\eta\lambda_{k,i_\ell}}{(1-\beta)^2D(\eta\lambda_{k,i_\ell})}+\frac{2}{D(\eta\lambda_{k,i_\ell})}+\frac{2(1+\beta)}{\eta\lambda_{k,i_\ell}D(\eta\lambda_{k,i_\ell})}\\
&\le\frac{8(1+\beta)}{(1-\beta)^3\kappa^{1/16}}+\frac{2}{\kappa^{1/16}}+\frac{2(1+\beta)}{\kappa^{1/8}}\\
&\le\left(\frac{16}{(1-\beta)^3}+6\right)\kappa^{-1/8}.
\end{aligned}$$

Also, we have already shown that the difference between the sub-projection matrices is bounded by $\kappa^{1/2}/2$. Therefore, we can conclude that

$$\begin{aligned}
&\left\|\widehat{\boldsymbol{P}}_{k-1}^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_{k}^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|_2-\left\|\widehat{\boldsymbol{P}}_{k-1}^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_{k-1}^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|_2\\
&\quad\le\left(\frac{8}{(1-\beta)^3}+3\right)^{\frac{1}{2}}\cdot\kappa^{3/16}\cdot\left\|\boldsymbol{P}_{k-1}^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_{k-1}^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|_2
\end{aligned}$$

Combing with (G.20), we have that the second term on the right-hand side of (G.18) can be bounded by

$$\left\|\boldsymbol{P}_{k-1}^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_{k}^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}\left(w_k\right)m_k\\ \boldsymbol{P}_{\mathrm{s}}\left(w_k\right)\nabla L\left(w_k\right)\end{pmatrix}\right\|_2$$

$$\leq \left(1 + d\eta^2\gamma^2 \cdot \left(\frac{32\sqrt{2}}{(1-\beta)^2} + 88\right) \cdot \kappa^{1/4} + \left(\frac{8}{(1-\beta)^3} + 3\right)^{\frac{1}{2}} \cdot \kappa^{3/16}\right) \cdot \left\| \boldsymbol{P}_{k-1}^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_{k-1}^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k)\, m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k)\, \nabla L(w_k) \end{pmatrix} \right\|_2 .$$

Now we can conclude that (G.18) can be bounded by

$$\begin{aligned}
&\left\| \boldsymbol{P}_k^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k)\, m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k)\, \nabla L(w_k) \end{pmatrix} \right\|_2 \\
&\quad \leq \left(1 + \frac{1}{2}\eta^2\gamma^2 \cdot \left(\frac{4\sqrt{2}}{(1-\beta)^2} + 11\right) \cdot \kappa^{3/4} + d\eta^2\gamma^2 \cdot \left(\frac{32\sqrt{2}}{(1-\beta)^2} + 88\right) \cdot \kappa^{1/4} + \left(\frac{8}{(1-\beta)^3} + 3\right)^{\frac{1}{2}} \cdot \kappa^{3/16}\right) \\
&\qquad \cdot \left\| \boldsymbol{P}_{k-1}^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_{k-1}^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k)\, m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k)\, \nabla L(w_k) \end{pmatrix} \right\|_2 \\
&\quad \leq \left(1 + \frac{1}{2} \cdot \left(\frac{8\sqrt{2}}{(1-\beta)^4} + \frac{22}{(1-\beta)^2}\right) \cdot \kappa^{3/4} + d \cdot \left(\frac{64\sqrt{2}}{(1-\beta)^4} + \frac{88}{(1-\beta)^2}\right) \cdot \kappa^{1/4} + \left(\frac{8}{(1-\beta)^3} + 3\right)^{\frac{1}{2}} \cdot \kappa^{3/16}\right) \\
&\qquad \cdot \left\| \boldsymbol{P}_{k-1}^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_{k-1}^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k)\, m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k)\, \nabla L(w_k) \end{pmatrix} \right\|_2 \\
&\quad \leq \left(1 + \mathsf{A}(\beta) \cdot \kappa^{3/16}\right) \cdot \left\| \boldsymbol{P}_{k-1}^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_{k-1}^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k)\, m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k)\, \nabla L(w_k) \end{pmatrix} \right\|_2 ,
\end{aligned}$$

where we denote $\mathsf{A}(\beta)$ as

$$\mathsf{A}(\beta) := \frac{1}{2} \cdot \left(\frac{8\sqrt{2}}{(1-\beta)^4} + \frac{22}{(1-\beta)^2}\right) + d \cdot \left(\frac{64\sqrt{2}}{(1-\beta)^4} + \frac{88}{(1-\beta)^2}\right) + \left(\frac{8}{(1-\beta)^3} + 3\right)^{\frac{1}{2}} .$$

Thus we conclude **Step 2** by the following,

$$\begin{aligned}
&\left\| \boldsymbol{P}_k^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_{k+1}) m_{k+1} \\ \boldsymbol{P}_{\mathrm{s}}(w_{k+1}) \nabla L(w_{k+1}) \end{pmatrix} \right\|_2 \qquad \text{(G.21)} \\
&\quad \leq (1 - \overline{q}_k) \cdot \left(1 + \mathsf{A}(\beta) \cdot \kappa^{3/16}\right) \cdot \left\| \boldsymbol{P}_{k-1}^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_{k-1}^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k)\, m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k)\, \nabla L(w_k) \end{pmatrix} \right\|_2 + \left\| \boldsymbol{P}_k^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \boldsymbol{E}_k \right\|_2 .
\end{aligned}$$

Step 3: Bounding the next step in the sharp direction: remaining terms. It then remains to control the remaining terms due to the spinning of the river, i.e., $\|\boldsymbol{P}_k^{1/2} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \boldsymbol{E}_k\|_2$. To this end, consider

$$\left\| \boldsymbol{P}_k^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \boldsymbol{E}_k \right\|_2 \leq \left\| \boldsymbol{P}_k^{\frac{1}{2}} \right\|_{\mathrm{Op}} \cdot \|\boldsymbol{E}_k\|_2 .$$

For the first term $\|\boldsymbol{P}_k^{1/2}\|_{\mathrm{Op}}$, as we have shown in **Step 2**, it holds that

$$\|\boldsymbol{P}_k\|_{\mathrm{Op}} \leq \left(\frac{16}{(1-\beta)^3} + 6\right)^{\frac{1}{2}} \cdot \kappa^{-1/16} . \qquad \text{(G.22)}$$

For the second term $\|\boldsymbol{E}_k\|_2$, by its definition in (G.16), we have that

$$\|\boldsymbol{E}_k\|_2 \le \|E_{k,1}\|_2 + \|E_{k,2}\|_2,$$

where we can bound $\|E_{k,1}\|_2$ and $\|E_{k,2}\|_2$ respectively. Firstly, the term $\|E_{k,1}\|_2$ can be bounded by

$$\|E_{k,1}\|_2 \le \big\|\big(\boldsymbol{P}_{\mathrm{s}}(w_{k+1}) - \boldsymbol{P}_{\mathrm{s}}(w_k)\big)m_{k+1}\big\|_2 \le \kappa\eta\gamma \cdot \big(\beta \cdot \|m_k\|_2 + (1-\beta)\cdot\|\nabla L(w_k)\|_2\big), \tag{G.23}$$

where in the second inequality we use Lemma H.6. Secondly, the term $\|E_{k,2}\|_2$ can be bounded by

$$\begin{aligned}\|E_{k,2}\|_2 \le &\left\|\eta\cdot\int_0^1 \Big(\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})\nabla^2 L(w_{k,\tau})\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau}) - \boldsymbol{P}_{\mathrm{s}}(w_k)\nabla^2 L(w_k)\boldsymbol{P}_{\mathrm{s}}(w_k)\Big)m_{k+1}\mathrm{d}\tau\right\|_2 \\ &+ \left\|\eta\cdot\int_0^1 \nabla\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})[m_{k+1}]\nabla L(w_{k,\tau})\mathrm{d}\tau\right\|_2,\end{aligned} \tag{G.24}$$

We further have the following bounds for the right hand side of the above,

$$\begin{aligned}&\left\|\eta\cdot\int_0^1 \Big(\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})\nabla^2 L(w_{k,\tau})\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau}) - \boldsymbol{P}_{\mathrm{s}}(w_k)\nabla^2 L(w_k)\boldsymbol{P}_{\mathrm{s}}(w_k)\Big)m_{k+1}\mathrm{d}\tau\right\|_2 \\ &\le \left\|\eta\cdot\int_0^1 \big(\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau}) - \boldsymbol{P}_{\mathrm{s}}(w_k)\big)\nabla^2 L(w_{k,\tau})\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})m_{k+1}\mathrm{d}\tau\right\|_2 \\ &\quad + \left\|\eta\cdot\int_0^1 \boldsymbol{P}_{\mathrm{s}}(w_k)\big(\nabla^2 L(w_{k,\tau}) - \nabla^2 L(w_k)\big)\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})m_{k+1}\mathrm{d}\tau\right\|_2 \\ &\quad + \left\|\eta\cdot\int_0^1 \boldsymbol{P}_{\mathrm{s}}(w_k)\nabla^2 L(w_k)\big(\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau}) - \boldsymbol{P}_{\mathrm{s}}(w_k)\big)m_{k+1}\mathrm{d}\tau\right\|_2 \\ &\le 2\eta\cdot\eta\gamma\kappa\cdot\gamma_{\max}\cdot\|m_{k+1}\|_2 + \eta\cdot\frac{1}{2}\eta\gamma^2\kappa\cdot\|m_{k+1}\|_2 \\ &\le \eta\gamma\kappa\cdot\big(2\eta\gamma_{\max} + \eta\gamma/2\big)\cdot\big(\beta\cdot\|m_k\|_2 + (1-\beta)\cdot\|\nabla L(w_k)\|_2\big),\end{aligned} \tag{G.25}$$

where the second inequality uses Lemma H.6 together with 2. and 3. of Assumption 2.2. Also,

$$\left\|\eta\cdot\int_0^1 \nabla\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})[m_{k+1}]\nabla L(w_{k,\tau})\mathrm{d}\tau\right\|_2 \le \eta\gamma\kappa\cdot\big(\beta\cdot\|m_k\|_2 + (1-\beta)\cdot\|\nabla L(w_k)\|_2\big), \tag{G.26}$$

where we apply Lemma H.5. Consequently, by (G.22), (G.23), (G.24), (G.25), and (G.26), we can conclude that

$$\begin{aligned}\left\|\boldsymbol{P}_k^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_k^\top\boldsymbol{E}_k\right\|_2 &\le \left(\frac{16}{(1-\beta)^3}+6\right)^{\frac{1}{2}}\eta\gamma\cdot\big(2+\eta\gamma_{\max}+\eta\gamma/2\big)\cdot\kappa^{15/16}\cdot\big(\beta\|m_k\|_2+(1-\beta)\|\nabla L(w_k)\|_2\big) \\ &\le \mathsf{B}(\beta)\cdot\kappa^{15/16}\cdot\left\|\begin{pmatrix}\beta\cdot m_k\\(1-\beta)\cdot\nabla L(w_k)\end{pmatrix}\right\|_2,\end{aligned} \tag{G.27}$$

where we define $\mathsf{B}(\beta)$ as

$$\mathsf{B}(\beta) := 2\cdot\left(\frac{16}{(1-\beta)^3}+6\right)^{\frac{1}{2}}\cdot\left(\frac{10}{1-\beta}+1\right)\cdot\frac{2}{1-\beta}.$$

Step 4: Concluding the proof. Finally, combining (G.21) and (G.27) , we can conclude that

$$\begin{aligned}
&\left\| \boldsymbol{P}_k^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_{k+1}) m_{k+1} \\ \boldsymbol{P}_{\mathrm{s}}(w_{k+1}) \nabla L(w_{k+1}) \end{pmatrix} \right\|_2 \\
&\quad \le \big(1 - \overline{q}_k\big) \cdot \big(1 + \mathsf{A}(\beta) \cdot \kappa^{3/16}\big) \cdot \left\| \boldsymbol{P}_{k-1}^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_{k-1}^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2 \\
&\qquad + \mathsf{B}(\beta) \cdot \kappa^{15/16} \cdot \left\| \begin{pmatrix} \beta \cdot m_k \\ (1-\beta) \cdot \nabla L(w_k) \end{pmatrix} \right\|_2
\end{aligned}$$

This completes the proof of Lemma G.4. ■

We recall again the river-valley regime specified by Assumption 2.2: the factors carrying $\gamma_{\mathrm{flat}}/\gamma_{\max}$ and κ are regarded as small enough so that the scalar induction errors are below a fixed constant. In particular, we use $\underline{q}_k \le 1/2$ for all relevant k, which follows from $\underline{q}_k = \mathcal{O}(\eta\gamma_{\mathrm{flat}}/(1-\beta))$ and that $\gamma_{\mathrm{flat}}/\gamma_{\max} \le \kappa^{1/2}$.

Lemma G.5 (One-step variation of the flat-direction gradient) *Suppose that Assumptions 2.1 and 2.2 hold. Suppose that the momentum iteration (GD-M) starts from an initial point $w_0 \in \mathcal{M}$ on the river and an initial momentum $m_0 = 0$. Take the learning rate η and momentum parameter β satisfying $\eta \in \mathcal{I}_1 \cap \mathcal{I}_2 \cap \mathcal{I}_3$. Suppose that the induction conditions (G.5), (G.6), and (G.7) hold up to step $k-1$ for some $k \ge 1$. Then*

$$\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k) - \boldsymbol{P}_{\mathrm{f}}(w_{k-1})\nabla L(w_{k-1})\|_2 \le \underline{q}_{k-1}\big(1 + 2\underline{q}_{k-1}\big) \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2.$$

Proof [Proof of Lemma G.5] Denote $G_\ell := \boldsymbol{P}_{\mathrm{f}}(w_\ell)\nabla L(w_\ell)$ for each ℓ. For the boundary case $k = 1$, we use the convention that $\sum_{j=0}^{-1} \beta^j = 0$, $b_{-1} = 0$, and $\underline{q}_{-1} = 0$. We first bound the perturbation from G_{k-1} to G_k in terms of G_{k-1}. By Taylor's formula along the segment $\{w_{k-1,\tau}\}_{\tau\in[0,1]}$,

$$\begin{aligned}
G_k - G_{k-1} = &-\eta \cdot \boldsymbol{P}_{\mathrm{f}}(w_{k-1})\nabla^2 L(w_{k-1})\boldsymbol{P}_{\mathrm{f}}(w_{k-1}) m_k \\
&- \eta \cdot \int_0^1 \Big(\boldsymbol{P}_{\mathrm{f}}(w_{k-1,\tau})\nabla^2 L(w_{k-1,\tau})\boldsymbol{P}_{\mathrm{f}}(w_{k-1,\tau}) \\
&\qquad - \boldsymbol{P}_{\mathrm{f}}(w_{k-1})\nabla^2 L(w_{k-1})\boldsymbol{P}_{\mathrm{f}}(w_{k-1}) \Big) m_k \mathrm{d}\tau \\
&- \eta \cdot \int_0^1 \nabla \boldsymbol{P}_{\mathrm{f}}(w_{k-1,\tau})[m_k]\nabla L(w_{k-1,\tau})\mathrm{d}\tau.
\end{aligned}$$

The first term on the right-hand side is controlled by induction condition (G.6) at step $k-1$:

$$\begin{aligned}
&\eta \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_{k-1})\nabla^2 L(w_{k-1})\boldsymbol{P}_{\mathrm{f}}(w_{k-1}) m_k\|_2 \\
&\quad \le \eta\gamma_{\mathrm{flat}} \cdot \left((1-\beta)\sum_{j=0}^{k-1} \beta^j + b_{k-1} \right) \cdot \|G_{k-1}\|_2.
\end{aligned} \tag{G.28}$$

For the two integral terms, the same estimates as in (G.25) and Lemma H.5, with the index shifted from k to $k-1$, give

$$\left\| \eta \cdot \int_0^1 \Big(\boldsymbol{P}_{\mathrm{f}}(w_{k-1,\tau})\nabla^2 L(w_{k-1,\tau})\boldsymbol{P}_{\mathrm{f}}(w_{k-1,\tau}) - \boldsymbol{P}_{\mathrm{f}}(w_{k-1})\nabla^2 L(w_{k-1})\boldsymbol{P}_{\mathrm{f}}(w_{k-1}) \Big) m_k \mathrm{d}\tau \right\|_2$$

$$\leq \eta\gamma\kappa \cdot \big(2\eta\gamma_{\max} + \eta\gamma/2\big) \cdot \big(\beta\|m_{k-1}\|_2 + (1-\beta)\|\nabla L(w_{k-1})\|_2\big), \tag{G.29}$$

$$\left\|\eta \cdot \int_0^1 \nabla \boldsymbol{P}_{\mathrm{f}}(w_{k-1,\tau})[m_k]\nabla L(w_{k-1,\tau})\mathrm{d}\tau\right\|_2$$

$$\leq \eta\gamma\kappa \cdot \big(\beta\|m_{k-1}\|_2 + (1-\beta)\|\nabla L(w_{k-1})\|_2\big). \tag{G.30}$$

By induction conditions (G.5), (G.6), and (G.7) up to step $k-1$, together with Lemma G.9, the factor multiplying the last two displays satisfies

$$\beta\|m_{k-1}\|_2 + (1-\beta)\|\nabla L(w_{k-1})\|_2$$

$$\leq 2\left((1-\beta) + \beta\left((1-\beta)\sum_{j=0}^{k-2}\beta^j + b_{k-2}\right)\big(1 + 2\underline{q}_{k-2}\big)\right) \cdot \|G_{k-1}\|_2, \tag{G.31}$$

where the boundary case $k=1$ follows from $m_0 = 0$ and the initialization $w_0 \in \mathcal{M}$. Combining (G.28), (G.29), (G.30), and (G.31), we obtain

$$\|G_k - G_{k-1}\|_2 \leq \widehat{q}_{k-1} \cdot \|G_{k-1}\|_2, \tag{G.32}$$

where

$$\widehat{q}_{k-1} := \eta\gamma_{\mathrm{flat}} \cdot \left((1-\beta)\sum_{j=0}^{k-1}\beta^j + b_{k-1}\right)$$
$$+ 2\eta\gamma\kappa \cdot \big(1 + 2\eta\gamma_{\max} + \eta\gamma/2\big)$$
$$\cdot \left((1-\beta) + \beta\left((1-\beta)\sum_{j=0}^{k-2}\beta^j + b_{k-2}\right)\big(1 + 2\underline{q}_{k-2}\big)\right).$$

It remains to compare $\widehat{q}_{k-1}$ with $\underline{q}_{k-1}$. By item 2 of Assumption 2.2 and Proposition G.3, $\kappa\gamma \leq \gamma_{\mathrm{flat}}$, $1 + 2\eta\gamma_{\max} + \eta\gamma/2 \leq \mathsf{C}(\beta)$, and $\underline{q}_{k-2} \leq q(\beta)$. Moreover, (G.10) gives $b_t \leq (1-\beta)\sum_{j=0}^{t}\beta^j$ for every $t \geq 0$. Therefore,

$$(1-\beta) + \beta\left((1-\beta)\sum_{j=0}^{k-2}\beta^j + b_{k-2}\right)\big(1 + 2\underline{q}_{k-2}\big)$$
$$\leq 2(1-\beta)\sum_{j=0}^{k-1}\beta^j\big(1 + 2q(\beta)\big).$$

Hence,

$$\widehat{q}_{k-1} \leq \eta\gamma_{\mathrm{flat}} \cdot \left(2(1-\beta)\sum_{j=0}^{k-1}\beta^j + 2\mathsf{C}(\beta)\cdot 2(1-\beta)\sum_{j=0}^{k-1}\beta^j \cdot \big(1 + 2q(\beta)\big)\right)$$

$$= 2\eta\gamma_{\mathrm{flat}} \cdot \Big(1 + 2\mathsf{C}(\beta)\big(1 + 2q(\beta)\big)\Big) \cdot (1-\beta)\sum_{j=0}^{k-1}\beta^j$$

$$= \underline{q}_{k-1}. \tag{G.33}$$

Combining (G.32) and (G.33) gives

$$\|G_k - G_{k-1}\|_2 \le \underline{q}_{k-1} \cdot \|G_{k-1}\|_2.$$

Finally, induction condition (G.5) at step $k-1$ and Lemma G.9 imply

$$\|G_{k-1}\|_2 \le \big(1 + 2\underline{q}_{k-1}\big) \cdot \|G_k\|_2.$$

The desired conclusion follows by substituting the last display into the previous one. ■

Lemma G.6 (Induction argument 1) *Suppose that Assumptions 2.1 and 2.2 hold. Suppose that the momentum iteration (GD-M) starts from an initial point $w_0 \in \mathcal{M}$ on the river and an initial momentum $m_0 = 0$. Take the learning rate η and momentum parameter β satisfying $\eta \in \mathcal{I}_1 \cap \mathcal{I}_2 \cap \mathcal{I}_3$. Also, suppose that the induction conditions (G.5), (G.6), and (G.7) hold up to step $k-1$ for some $k \ge 1$, and that (G.7) holds for the dummy step $k = -1$. Then for step k,*

$$\left\|(1-\beta)\cdot\sum_{j=0}^{k}\beta^j\cdot\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k) - \boldsymbol{P}_{\mathrm{f}}(w_k)m_{k+1}\right\|_2 \le \widetilde{b}_k\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2, \quad \textit{where}$$

$$\widetilde{b}_k := \big(1+2\underline{q}_{k-1}\big)\cdot\left(\beta b_{k-1} + \beta(1-\beta)\sum_{j=0}^{k-1}\beta^j\underline{q}_{k-1} + \kappa\eta\gamma\beta\left(1+(1-\beta)\sum_{j=0}^{k-1}\beta^j + b_{k-1}\right)\right).$$

Proof [Proof of Lemma G.6] By definition, we have the following decomposition of the left hand side,

$$\begin{aligned}
&\left\|(1-\beta)\cdot\sum_{j=0}^{k}\beta^j\cdot\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k) - \boldsymbol{P}_{\mathrm{f}}(w_k)m_{k+1}\right\|_2 \\
&\quad= \left\|(1-\beta)\cdot\sum_{j=1}^{k}\beta^j\cdot\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k) - \beta\cdot\boldsymbol{P}_{\mathrm{f}}(w_k)m_k\right\|_2 \\
&\quad\le \left\|\beta\cdot\left((1-\beta)\cdot\sum_{j=0}^{k-1}\beta^j\cdot\boldsymbol{P}_{\mathrm{f}}(w_{k-1})\nabla L(w_{k-1}) - \boldsymbol{P}_{\mathrm{f}}(w_{k-1})m_k\right)\right\|_2 \\
&\qquad+ \left\|\beta(1-\beta)\cdot\sum_{j=0}^{k-1}\beta^j\cdot\Big(\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k) - \boldsymbol{P}_{\mathrm{f}}(w_{k-1})\nabla L(w_{k-1})\Big)\right\|_2 \\
&\qquad+ \left\|\beta\cdot\Big(\boldsymbol{P}_{\mathrm{f}}(w_{k-1}) - \boldsymbol{P}_{\mathrm{f}}(w_k)\Big)m_k\right\|_2.
\end{aligned}$$

Now using induction condition (G.6) at step $k-1$ and Lemma G.5, we can derive that

$$\left\|(1-\beta)\cdot\sum_{j=0}^{k}\beta^j\cdot\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k) - \boldsymbol{P}_{\mathrm{f}}(w_k)m_{k+1}\right\|_2$$

$$\leq \left(\beta \cdot b_{k-1} + \beta(1-\beta) \cdot \sum_{j=0}^{k-1} \beta^j \cdot \underline{q}_{k-1}\right) \cdot (1 + 2\underline{q}_{k-1}) \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2$$
$$+ \beta \cdot \left\| \Big(\boldsymbol{P}_{\mathrm{f}}(w_{k-1}) - \boldsymbol{P}_{\mathrm{f}}(w_k)\Big) m_k \right\|_2,$$

where we have also applied Lemma G.9. Furthermore, the last term on the right hand side above is bounded by the following,

$$\left\| \Big(\boldsymbol{P}_{\mathrm{f}}(w_{k-1}) - \boldsymbol{P}_{\mathrm{f}}(w_k)\Big) m_k \right\|_2 \leq \kappa\eta\gamma \cdot \|m_k\|_2$$
$$\leq \kappa\eta\gamma \cdot \|\boldsymbol{P}_{\mathrm{s}}(w_{k-1})m_k\|_2 + \kappa\eta\gamma \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_{k-1})m_k\|_2,$$

where the first inequality uses Lemma H.6. On the one hand, by the induction conditions (G.5) and (G.6) at step $k-1$, applying Lemma G.9, we have that

$$\|\boldsymbol{P}_{\mathrm{f}}(w_{k-1})m_k\|_2 \leq \left((1-\beta) \cdot \sum_{j=0}^{k-1} \beta^j + b_{k-1}\right) \cdot (1 + 2\underline{q}_{k-1}) \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2.$$

On the other hand, by the induction condition (G.7) at step $k-2$ and condition (G.5) at step $k-1$,

$$\|\boldsymbol{P}_{\mathrm{s}}(w_{k-1})m_k\|_2 \leq \|\boldsymbol{P}_{\mathrm{s}}(w_{k-1})m_{k-1}\|_2 + \|\boldsymbol{P}_{\mathrm{s}}(w_{k-1})\nabla L(w_{k-1})\|_2$$
$$\leq 2c \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_{k-1})\nabla L(w_{k-1})\|_2$$
$$\leq (1 + 2\underline{q}_{k-1}) \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2,$$

where the last inequality applies Lemma G.9 and the condition that $c \leq 1/2$. Therefore, by combining all the above inequalities, we can conclude the proof of Lemma G.6. ■

Lemma G.7 (Induction argument 2) *Suppose that Assumptions 2.1 and 2.2 hold. Also, suppose that the induction conditions (G.5), (G.6), and (G.7) hold up to step $k-1$, plus that the induction condition (G.6) hold for step k, where $k \geq 1$, then it holds that for such a step k and any $\tau \in [0,1]$,*

$$\|\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})\nabla L(w_{k,\tau})\|_2 \geq \big(1 - \widetilde{\underline{q}}_k\big) \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2, \quad \textit{where}$$
$$\widetilde{\underline{q}}_k := \eta\gamma_{\mathrm{flat}} \cdot \left((1-\beta)\sum_{j=0}^{k} \beta^j + b_k\right)$$
$$+ 2\eta\gamma\kappa\big(1 + 2\eta\gamma_{\max} + \eta\gamma/2\big) \cdot \left((1-\beta) + \beta\left((1-\beta)\sum_{j=0}^{k-1}\beta^j + b_{k-1}\right)(1 + 2\underline{q}_{k-1})\right).$$

Moreover, the above conclusion also holds for step $k = 0$ supposing that the induction condition (G.7) holds for step $k = -1$ and that the induction condition (G.6) holds for step $k = 0$.

Proof [Proof of Lemma G.7] Suppose that the induction conditions (G.5), (G.6), and (G.7) hold up to step $k-1$, plus that the induction condition (G.6) hold for step k, where $k \geq 1$. Then for such a step k,

$$\boldsymbol{P}_{\mathrm{f}}(w_{k+1})\nabla L(w_{k+1})$$

$$
\begin{aligned}
&= \boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k) - \eta\cdot\int_0^1 \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})\nabla^2 L(w_{k,\tau})\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})m_{k+1}\mathrm{d}\tau \\
&\quad - \eta\cdot\int_0^1 \nabla\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})[m_{k+1}]\nabla L(w_{k,\tau})\mathrm{d}\tau \\
&= \boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k) - \eta\cdot\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla^2 L(w_k)\boldsymbol{P}_{\mathrm{f}}(w_k)m_{k+1} \\
&\quad - \eta\cdot\int_0^1 \Big(\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})\nabla^2 L(w_{k,\tau})\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau}) - \boldsymbol{P}_{\mathrm{f}}(w_k)\nabla^2 L(w_k)\boldsymbol{P}_{\mathrm{f}}(w_k)\Big)m_{k+1}\mathrm{d}\tau \\
&\quad - \eta\cdot\int_0^1 \nabla\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})[m_{k+1}]\nabla L(w_{k,\tau})\mathrm{d}\tau \\
&= \boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k) - \eta\cdot(1-\beta)\cdot\sum_{j=0}^{k}\beta^j\cdot\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla^2 L(w_k)\boldsymbol{P}_{\mathrm{f}}(w_k)\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k) \\
&\quad + \eta\cdot\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla^2 L(w_k)\boldsymbol{P}_{\mathrm{f}}(w_k)\left((1-\beta)\cdot\sum_{j=0}^{k}\beta^j\cdot\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k) - \boldsymbol{P}_{\mathrm{f}}(w_k)m_{k+1}\right) \\
&\quad - \eta\cdot\int_0^1 \Big(\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})\nabla^2 L(w_{k,\tau})\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau}) - \boldsymbol{P}_{\mathrm{f}}(w_k)\nabla^2 L(w_k)\boldsymbol{P}_{\mathrm{f}}(w_k)\Big)m_{k+1}\mathrm{d}\tau \\
&\quad - \eta\cdot\int_0^1 \nabla\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})[m_{k+1}]\nabla L(w_{k,\tau})\mathrm{d}\tau,
\end{aligned}
$$

Therefore, we can lower bound $\|\boldsymbol{P}_{\mathrm{f}}(w_{k+1})\nabla(w_{k+1})\|_2$ as

$$
\begin{aligned}
&\|\boldsymbol{P}_{\mathrm{f}}(w_{k+1})\nabla(w_{k+1})\|_2 \\
&\quad\geq \left(1-\eta\gamma_{\mathrm{flat}}\cdot(1-\beta)\cdot\sum_{j=0}^{k}\beta^j - \eta\gamma_{\mathrm{flat}}\cdot b_k\right)\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla(w_k)\|_2 \\
&\qquad - \left\|\eta\cdot\int_0^1 \Big(\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})\nabla^2 L(w_{k,\tau})\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau}) - \boldsymbol{P}_{\mathrm{f}}(w_k)\nabla^2 L(w_k)\boldsymbol{P}_{\mathrm{f}}(w_k)\Big)m_{k+1}\mathrm{d}\tau\right\|_2 \\
&\qquad - \left\|\eta\cdot\int_0^1 \nabla\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})[m_{k+1}]\nabla L(w_{k,\tau})\mathrm{d}\tau\right\|_2,
\end{aligned}
$$

where we have applied item 2 of Assumption 2.2 and the induction condition (G.6) at step k. Moreover, for the last two terms on the right hand above, we have the following upper bounds,

$$
\begin{aligned}
&\left\|\eta\cdot\int_0^1 \Big(\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})\nabla^2 L(w_{k,\tau})\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau}) - \boldsymbol{P}_{\mathrm{f}}(w_k)\nabla^2 L(w_k)\boldsymbol{P}_{\mathrm{f}}(w_k)\Big)m_{k+1}\mathrm{d}\tau\right\|_2 \qquad \text{(G.34)} \\
&\quad\leq \eta\gamma\kappa\cdot\big(2\eta\gamma_{\max}+\eta\gamma/2\big)\cdot\big(\beta\cdot\|m_k\|_2 + (1-\beta)\cdot\|\nabla L(w_k)\|_2\big) \\
&\quad\leq \eta\gamma\kappa\cdot\big(2\eta\gamma_{\max}+\eta\gamma/2\big)\cdot\Big(\beta\cdot\big(\|\boldsymbol{P}_{\mathrm{f}}(w_{k-1})m_k\|_2 + \|\boldsymbol{P}_{\mathrm{s}}(w_{k-1})m_k\|_2\big) \\
&\qquad + (1-\beta)\cdot\big(\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2 + \|\boldsymbol{P}_{\mathrm{s}}(w_k)\nabla L(w_k)\|_2\big)\Big) \\
&\quad\leq \eta\gamma\kappa\cdot(1+c)\big(2\eta\gamma_{\max}+\eta\gamma/2\big)\Big(\beta\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_{k-1})m_k\|_2 + (1-\beta)\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2\Big) \\
&\quad\leq 2\eta\gamma\kappa\cdot\big(2\eta\gamma_{\max}+\eta\gamma/2\big)
\end{aligned}
$$

$$\cdot \left((1-\beta) + \beta \cdot \left((1-\beta) \cdot \sum_{j=0}^{k-1} \beta^j + b_{k-1} \right) \cdot (1 + 2\underline{q}_{k-1}) \right) \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_k) \nabla L(w_k)\|_2,$$

where the first inequality follows from the same arguments as in (G.25), the third inequality applies induction condition (G.7) at step $k-1$ and step $k-2$, and the last inequality uses induction conditions (G.5) and (G.6) at step $k-1$, and that $c \leq 1$. Similarly, we have that

$$\begin{aligned}
&\left\| \eta \cdot \int_0^1 \nabla \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})[m_{k+1}] \nabla L(w_{k,\tau}) \mathrm{d}\tau \right\|_2 \\
&\quad \leq \eta\gamma\kappa \cdot \Big(\beta \cdot \|m_k\|_2 + (1-\beta) \cdot \|\nabla L(w_k)\|_2 \Big) \\
&\quad \leq 2\eta\gamma\kappa \cdot \left((1-\beta) + \beta \cdot \left((1-\beta) \sum_{j=0}^{k-1} \beta^j + b_{k-1} \right) (1 + 2\underline{q}_{k-1}) \right) \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_k) \nabla L(w_k)\|_2,
\end{aligned}$$

where the first inequality uses Lemma H.5, and the second inequality applies the same argument for deriving the above (G.34). Consequently, we can conclude that

$$\begin{aligned}
&\|\boldsymbol{P}_{\mathrm{f}}(w_{k+1}) \nabla (w_{k+1})\|_2 \geq \big(1 - \widetilde{\underline{q}}_k\big) \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_k) \nabla L(w_k)\|_2, \quad \text{where} \\
&\widetilde{\underline{q}}_k := \eta\gamma_{\mathrm{flat}} \cdot \left((1-\beta) \sum_{j=0}^{k} \beta^j + b_k \right) \\
&\qquad + 2\eta\gamma\kappa\big(1 + 2\eta\gamma_{\max} + \eta\gamma/2\big) \cdot \left((1-\beta) + \beta \left((1-\beta) \sum_{j=0}^{k-1} \beta^j + b_{k-1} \right) (1 + 2\underline{q}_{k-1}) \right).
\end{aligned}$$

Finally, for the case of step $k = 0$ given the correctness of induction conditions (G.6) at step $k = 0$ and (G.7) at step $k = -1$, the conclusion can be proved similarly by additionally using the fact that $m_0 = 0$. This completes the proof of Lemma G.7. ■

Lemma G.8 (Induction argument 3) *Suppose that the momentum iteration (GD-M) starts from an initial point $w_0 \in \mathcal{M}$ on the river and an initial momentum $m_0 = 0$. Suppose that Assumptions 2.1 and 2.2 hold. Also, suppose that the induction conditions (G.4), (G.5), and (G.6) hold up to step k, plus that the induction condition (G.7) hold up to step $k-1$, where $k \geq 0$. Take the learning rate η and the momentum parameter β satisfying $\eta \in \mathcal{I}_1 \cap \mathcal{I}_2 \cap \mathcal{I}_3 \cap \mathcal{I}_4 \cap \mathcal{I}_5$ (see definitions in Theorem G.1). Then it holds that for such a step step k and any $\tau \in [0,1]$,*

$$\left\| \boldsymbol{P}_k^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_{k,\tau}) m_{k+1} \\ \boldsymbol{P}_{\mathrm{s}}(w_{k,\tau}) \nabla L(w_{k,\tau}) \end{pmatrix} \right\|_2 \leq 3\mathsf{B}(\beta) \cdot \kappa^{1/2} \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_{k+1}) \nabla L(w_{k+1})\|_2,$$

and consequently it holds that

$$\begin{aligned}
&\max \Big\{ \|\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau}) \nabla L(w_{k,\tau})\|_2, \|\boldsymbol{P}_{\mathrm{s}}(w_{k+1}) m_{k+1}\|_2 \Big\} \\
&\qquad \leq 3\mathsf{B}(\beta) \cdot \kappa^{1/2} \cdot \kappa \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_{k+1}) \nabla L(w_{k+1})\|_2.
\end{aligned}$$

Here the matrices $\{\boldsymbol{P}_k, \boldsymbol{O}, \boldsymbol{V}_k\}_{k \geq 0} \subset \mathbb{R}^{2d \times 2d}$ are defined in Lemma G.4.

Proof [Proof of Lemma G.8] By the induction conditions (G.4) and (G.5) for step k, we have that

$$\begin{aligned}
&\left\| \boldsymbol{P}_k^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_{k+1}) m_{k+1} \\ \boldsymbol{P}_{\mathrm{s}}(w_{k+1}) \nabla L(w_{k+1}) \end{pmatrix} \right\|_2 \\
&\quad \le \big(1 - \overline{q}_k\big) \cdot \big(1 + \mathsf{A}(\beta) \cdot \kappa^{3/16}\big) \cdot \left\| \boldsymbol{P}_{k-1}^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_{k-1}^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2 \\
&\qquad + \mathsf{B}(\beta) \cdot \kappa^{15/16} \cdot \left\| \begin{pmatrix} \beta \cdot m_k \\ (1-\beta) \cdot \nabla L(w_k) \end{pmatrix} \right\|_2,
\end{aligned} \tag{G.35}$$

and that

$$\|\boldsymbol{P}_{\mathrm{f}}(w_{k+1}) \nabla L(w_{k+1})\|_2 \ge \Big(1 - 2\eta\gamma_{\mathrm{flat}} \cdot \big(1 + 6\mathsf{C}(\beta)\big)\Big) \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_k) \nabla L(w_k)\|_2. \tag{G.36}$$

Note that here we also use $\eta \in \mathcal{I}_1 \cap \mathcal{I}_2 \cap \mathcal{I}_3$ to apply the following bounds (see Proposition G.3) to obtain (G.36) from (G.5),

$$1 + 2\eta\gamma_{\max} + \frac{\eta\gamma}{2} \le \mathsf{C}(\beta), \quad \underline{q}_k \le 2\eta\gamma_{\mathrm{flat}} \cdot \big(1 + 6\mathsf{C}(\beta)\big).$$

By combining (G.35) and (G.36), we have that, for any coefficient $c \in [0, 1]$,

$$\begin{aligned}
&\left\| \boldsymbol{P}_k^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_k^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_{k+1}) m_{k+1} \\ \boldsymbol{P}_{\mathrm{s}}(w_{k+1}) \nabla L(w_{k+1}) \end{pmatrix} \right\|_2 - c \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_{k+1}) \nabla L(w_{k+1})\|_2 \\
&\quad \le (1 - \overline{q}_k) \cdot \big(1 + \mathsf{A}(\beta) \cdot \kappa^{3/16}\big) \cdot \left\| \boldsymbol{P}_{k-1}^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_{k-1}^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2 \\
&\qquad - c \cdot \Big(1 - 2\eta\gamma_{\mathrm{flat}} \cdot \big(1 + 6\mathsf{C}(\beta)\big)\Big) \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_k) \nabla L(w_k)\|_2 \\
&\qquad + \mathsf{B}(\beta) \cdot \kappa^{15/16} \cdot \left\| \begin{pmatrix} \beta \cdot m_k \\ (1-\beta) \cdot \nabla L(w_k) \end{pmatrix} \right\|_2.
\end{aligned} \tag{G.37}$$

Notice that the last term above can be bounded by the following,

$$\left\| \begin{pmatrix} \beta \cdot m_k \\ (1-\beta) \cdot \nabla L(w_k) \end{pmatrix} \right\|_2 \le \left\| \begin{pmatrix} \beta \cdot \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ (1-\beta) \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2 + \left\| \begin{pmatrix} \beta \cdot \boldsymbol{P}_{\mathrm{f}}(w_k) m_k \\ (1-\beta) \boldsymbol{P}_{\mathrm{f}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2.$$

On the one hand, for the first term above, we have that

$$\left\| \begin{pmatrix} \beta \cdot \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ (1-\beta) \cdot \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2 \le \left\| \boldsymbol{P}_{k-1}^{\frac{1}{2}} \boldsymbol{O}^\top \boldsymbol{V}_{k-1}^\top \begin{pmatrix} \boldsymbol{P}_{\mathrm{s}}(w_k) m_k \\ \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \end{pmatrix} \right\|_2.$$

On the other hand, for the second term, we have,

$$\begin{aligned}
&\left\|\begin{pmatrix}\beta\cdot\boldsymbol{P}_{\mathrm{f}}(w_k)m_k\\(1-\beta)\cdot\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\end{pmatrix}\right\|_2\\
&\quad\le\|\beta\cdot\boldsymbol{P}_{\mathrm{f}}(w_k)m_k\|_2+\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2\\
&\quad\le\left\|\beta\cdot\boldsymbol{P}_{\mathrm{f}}(w_k)m_k-(1-\beta)\cdot\sum_{j=1}^{k}\beta^j\cdot\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\right\|_2\\
&\qquad+\left\|(1-\beta)\cdot\sum_{j=1}^{k}\beta^j\cdot\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\right\|_2+\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2\\
&\quad\le\big(b_k+2\big)\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2\\
&\quad\le 3\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2,
\end{aligned}$$

where the third inequality uses induction condition (G.5) at step k and that $b_k\le 1$ due to $\eta\in\mathcal{I}_3$ (see (G.10) in Proposition G.3). Consequently, we obtain that

$$\left\|\begin{pmatrix}\beta\cdot m_k\\(1-\beta)\cdot\nabla L(w_k)\end{pmatrix}\right\|_2\le\left\|\boldsymbol{P}_{k-1}^{\frac12}\boldsymbol{O}^\top\boldsymbol{V}_{k-1}^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}(w_k)m_k\\\boldsymbol{P}_{\mathrm{s}}(w_k)\nabla L(w_k)\end{pmatrix}\right\|_2+3\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2 \tag{G.38}$$

Combining (G.38) with (G.37), we can conclude that

$$\begin{aligned}
&\left\|\boldsymbol{P}_k^{\frac12}\boldsymbol{O}^\top\boldsymbol{V}_k^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})m_{k+1}\\\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})\nabla L(w_{k,\tau})\end{pmatrix}\right\|_2-c\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_{k+1})\nabla L(w_{k+1})\|_2\\
&\quad\le\left(\big(1-\overline{q}_k\big)\cdot\Big(1+\mathsf{A}(\beta)\cdot\kappa^{3/16}\Big)+\mathsf{B}(\beta)\cdot\kappa^{15/16}\right)\cdot\left\|\boldsymbol{P}_{k-1}^{\frac12}\boldsymbol{O}^\top\boldsymbol{V}_{k-1}^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}(w_k)m_k\\\boldsymbol{P}_{\mathrm{s}}(w_k)\nabla L(w_k)\end{pmatrix}\right\|_2\\
&\qquad-\left(c\cdot\Big(1-2\eta\gamma_{\mathrm{flat}}\cdot\big(1+6\mathsf{C}(\beta)\big)\Big)-3\mathsf{B}(\beta)\cdot\kappa^{15/16}\right)\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2.
\end{aligned}$$

Now we let the right hand side of the above inequality be equal to

$$\begin{aligned}
&\left(\big(1-\overline{q}_k\big)\cdot\Big(1+\mathsf{A}(\beta)\cdot\kappa^{3/16}\Big)+\mathsf{B}(\beta)\cdot\kappa^{15/16}\right)\\
&\qquad\cdot\left(\left\|\boldsymbol{P}_{k-1}^{\frac12}\boldsymbol{O}^\top\boldsymbol{V}_{k-1}^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}(w_k)m_k\\\boldsymbol{P}_{\mathrm{s}}(w_k)\nabla L(w_k)\end{pmatrix}\right\|_2-c\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2\right),
\end{aligned}$$

which is equivalent to letting

$$c\cdot\left(\big(1-\overline{q}_k\big)\cdot\Big(1+\mathsf{A}(\beta)\cdot\kappa^{3/16}\Big)+\mathsf{B}(\beta)\cdot\kappa^{15/16}\right)=c\cdot\Big(1-2\eta\gamma_{\mathrm{flat}}\cdot\big(1+6\mathsf{C}(\beta)\big)\Big)-3\mathsf{B}(\beta)\cdot\kappa^{15/16},$$

or equivalently,

$$\left(\overline{q}_k\cdot\Big(1+\mathsf{A}(\beta)\cdot\kappa^{3/16}\Big)-\mathsf{A}(\beta)\cdot\kappa^{3/16}-\mathsf{B}(\beta)\cdot\eta\gamma\cdot\kappa^{15/16}-2\eta\gamma_{\mathrm{flat}}\cdot\big(1+6\mathsf{C}(\beta)\big)\right)\cdot c=3\mathsf{B}(\beta)\cdot\kappa^{15/16}.$$

To this end, we let

$$\overline{q}_k \cdot \Big(1 + \mathsf{A}(\beta) \cdot \kappa^{3/16}\Big) - \mathsf{A}(\beta) \cdot \kappa^{3/16} - \mathsf{B}(\beta) \cdot \eta\gamma \cdot \kappa^{15/16} - 2\eta\gamma_{\text{flat}} \cdot \big(1 + 6\mathsf{C}(\beta)\big) \geq \kappa^{7/16}. \tag{G.39}$$

By the definition of $\overline{q}_k$, a sufficient condition for (G.39) to hold is letting

$$\max_{i \in [d-1]} \mathrm{Tr}(\boldsymbol{P}_{k,i}) \leq \frac{1}{\mathsf{D}_1(\beta) \cdot \kappa^{3/16} + \mathsf{D}_2(\beta) \cdot \gamma_{\text{flat}}\gamma_{\max}^{-1}} := \frac{1}{\varsigma}. \tag{G.40}$$

Here we denote $\mathsf{D}_1(\beta)$ and $\mathsf{D}_2(\beta)$ respectively as

$$\begin{aligned}
\mathsf{D}_1(\beta) &:= 1 + \mathsf{A}(\beta) + \mathsf{B}(\beta), \\
\mathsf{D}_2(\beta) &:= 4 \cdot \big(1 + 6\mathsf{C}(\beta)\big) \cdot \frac{1+\beta}{1-\beta}.
\end{aligned}$$

Solving the above inequality (G.40) is equivalent to the following inequality on η,

$$\frac{(2\beta^3 - 2\beta^2 + 3\beta - 1) \cdot (\eta\lambda_{k,i})^2 + 2(1-\beta^2) \cdot \eta\lambda_{k,i} + 2(1-\beta)^2(1+\beta)}{\eta\lambda_{k,i}(1-\beta)^2 \cdot (2(1+\beta) - (1-\beta)\eta\lambda_{k,i})} \leq \frac{1}{\varsigma}, \forall i \in [d-1].$$

This actually can be simplified to the following $d-1$ quadratic inequalities on η,

$$\begin{aligned}
\mathsf{P}(\eta\lambda_{k,i}) &= c_2 \cdot (\eta\lambda_{k,i})^2 + c_1 \cdot \eta\lambda_{k,i} + c_0, \quad \text{where } i \in [d-1], \\
c_2 &= a(1-\beta)^3 + \big(2\beta^3 - 2\beta^2 + 3\beta - 1\big), \\
c_1 &= 2(1+\beta)(1-\beta)\big(1 - a(1-\beta)\big), \\
c_0 &= 2(1+\beta)(1-\beta)^2.
\end{aligned}$$

Given the condition on β that $\beta \leq 1 - \varsigma = 1 - \mathsf{D}_1(\beta) \cdot \kappa^{3/16} - \mathsf{D}_2(\beta) \cdot \gamma_{\text{flat}}\gamma_{\max}^{-1}$, solving the above inequalities gives

$$\frac{(1-\beta)((1+\beta)(a(1-\beta)-1) - \sqrt{(1+\beta)\,R(\beta)})}{a(1-\beta)^3 + (2\beta^3 - 2\beta^2 + 3\beta - 1)} \leq \eta\lambda_{k,i} \leq \frac{(1-\beta)((1+\beta)(a(1-\beta)-1) + \sqrt{(1+\beta)\,R(\beta)})}{a(1-\beta)^3 + (2\beta^3 - 2\beta^2 + 3\beta - 1)},$$

where we let $a := 1/\varsigma$ and $R(\beta)$ is defined as

$$R(\beta) := (a^2 + 2a - 4) \cdot \beta^3 - (a^2 + 4a - 4) \cdot \beta^2 + (-a^2 + 6a - 5) \cdot \beta + (a^2 - 4a + 3).$$

This effectively requires that

$$\begin{aligned}
\eta\lambda_{k,1} &\leq \frac{(1-\beta)((1+\beta)(a(1-\beta)-1) + \sqrt{(1+\beta)\,R(\beta)})}{a(1-\beta)^3 + (2\beta^3 - 2\beta^2 + 3\beta - 1)}, \quad \text{and} \\
\eta\lambda_{k,d-1} &\geq \frac{(1-\beta)((1+\beta)(a(1-\beta)-1) - \sqrt{(1+\beta)\,R(\beta)})}{a(1-\beta)^3 + (2\beta^3 - 2\beta^2 + 3\beta - 1)}.
\end{aligned}$$

For the first inequality above, since $\lambda_{k,1} \le \gamma_{\max}$, we obtain that η need to satisfy

$$\begin{aligned}\eta &\le \frac{(1-\beta)((1+\beta)(a(1-\beta)-1)+\sqrt{(1+\beta)\,R(\beta)})}{a(1-\beta)^3+(2\beta^3-2\beta^2+3\beta-1)}\cdot\frac{1}{\gamma_{\max}}\\ &= \Bigg(\underbrace{\frac{2\cdot(1+\beta)}{1-\beta}}_{\text{main term}} - \underbrace{\frac{5\beta^4-2\beta^3+6\beta^2-2\beta+1}{(1-\beta)^4}\cdot\varsigma-\mathcal{O}(\varsigma^2)}_{\text{small terms caused by river spinning}}\Bigg)\cdot\frac{1}{\gamma_{\max}}. \end{aligned} \tag{G.41}$$

For the second inequality above, since $\lambda_{k,d-1} \ge \gamma$, we obtain that η need to satisfy

$$\eta \ge \frac{(1-\beta)((1+\beta)(a(1-\beta)-1)+\sqrt{(1+\beta)\,R(\beta)})}{a(1-\beta)^3+(2\beta^3-2\beta^2+3\beta-1)}\cdot\frac{1}{\gamma}.$$

By Assumption 2.2 that $\gamma/\gamma_{\max} \ge \kappa^{1/32}$ and that $\gamma_{\text{flat}}/\gamma_{\max} \le \kappa^{1/2}$, an upper bound of the right hand side of above inequality is

$$\begin{aligned}&\frac{(1-\beta)((1+\beta)(a(1-\beta)-1)+\sqrt{(1+\beta)\,R(\beta)})}{a(1-\beta)^3+(2\beta^3-2\beta^2+3\beta-1)}\cdot\frac{1}{\gamma}\\ &\le \frac{(1-\beta)((1+\beta)(a(1-\beta)-1)+\sqrt{(1+\beta)\,R(\beta)})}{a(1-\beta)^3+(2\beta^3-2\beta^2+3\beta-1)}\cdot\frac{\kappa^{-1/32}}{\gamma_{\max}}\\ &= \left(1+\frac{\beta^2+3}{2(1-\beta^2)}\cdot\varsigma+\mathcal{O}(\varsigma^2)\right)\cdot\varsigma\cdot\frac{\kappa^{-1/32}}{\gamma_{\max}}\\ &= \left(1+\frac{\beta^2+3}{2(1-\beta^2)}\cdot\varsigma+\mathcal{O}(\varsigma^2)\right)\cdot\left(\mathsf{D}_1(\beta)\cdot\kappa^{5/32}+\mathsf{D}_2(\beta)\cdot\frac{\gamma_{\text{flat}}}{\gamma_{\max}}\cdot\frac{1}{\kappa^{1/32}}\right)\cdot\frac{1}{\gamma_{\max}}\\ &\le \left(\mathsf{D}_1(\beta)\cdot\kappa^{5/32}+\mathsf{D}_2(\beta)\cdot\kappa^{15/32}\right)\cdot\frac{2}{\gamma_{\max}},\end{aligned}$$

This is the last interval $\mathcal{I}_5$. By $\eta \in \mathcal{I}_5$, we have

$$\left(\mathsf{D}_1(\beta)\cdot\kappa^{5/32}+\mathsf{D}_2(\beta)\cdot\kappa^{15/32}\right)\cdot\frac{2}{\gamma_{\max}} \le \eta \le \left(\frac{1+\beta}{1-\beta}-\frac{5\beta^4-2\beta^3+6\beta^2-2\beta+1}{2(1-\beta)^4}\cdot\varsigma-\mathcal{O}(\varsigma^2)\right)\cdot\frac{2}{\gamma_{\max}}$$

Thus (G.39) holds, and we can further obtain that

$$c \le 3\mathsf{B}(\beta)\cdot\kappa^{1/2},$$

By induction and initial condition, we can conclude that

$$\left\|\boldsymbol{P}_k^{\frac{1}{2}}\boldsymbol{O}^\top\boldsymbol{V}_k^\top\begin{pmatrix}\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})m_{k+1}\\ \boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})\nabla L(w_{k,\tau})\end{pmatrix}\right\|_2 \le 3\mathsf{B}(\beta)\cdot\kappa^{1/2}\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_{k+1})\nabla L(w_{k+1})\|_2,$$

This completes the proof of Lemma G.8. ■

Lemma G.9 (Auxiliary inequality 1) *Suppose that induction condition (G.5) hold for some $\underline{q}_k \le 1/2$ at step k, then it holds that*

$$\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2 \le (1+2\underline{q}_k)\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_{k+1})\nabla L(w_{k+1})\|_2.$$

Proof [Proof of Lemma G.9] This follows from the fact that $1/(1-x) \le 1+2x$ for $0 \le x \le 1/2$. ■

G.3. Remaining Proofs in the Outline (Section B)

G.3.1. PROOF OF LEMMA B.2

We first give the formal version of Lemma B.2 in the following.

Lemma G.10 (The trajectory tracks the river closely) *Under the conditions and assumptions of Theorem G.1, taking the learning rate η and momentum parameter $\beta \leq 0.99$ satisfying $\eta \in \mathcal{I}(\beta)$ defined in Theorem G.1, then for any iteration $k \in \mathbb{N}$ and $\tau \in [0,1]$ it holds that: (i) $\mathcal{B}(w_{k,\tau}, 2g_{\max}/\gamma) \subset \mathcal{U}$ and thus $\Phi(w_{k,\tau})$ exists; (ii) it holds that for any step $k \in \mathbb{N}$ and $\tau \in [0,1]$,*

$$\|w_{k,\tau} - \Phi(w_{k,\tau})\|_2 \leq \frac{6\mathsf{B}(\beta) \cdot \kappa^{1/2} \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})\nabla L(w_{k,\tau})\|_2}{\gamma + 2\gamma_{\mathrm{flat}}}.$$

Proof [Proof of Lemma G.10] We prove Lemma G.10 here by the following two steps.

Step 1: The GD trajectory has a well-defined projection onto the river. We first prove that the interpolated GD trajectory $\{w_{\lfloor t \rfloor, t-\lfloor t \rfloor}\}_{t \geq 0}$ can be projected onto the river $\mathcal{M}$ by induction. According to Lemma H.4, it suffices to check that any point $w \in \{w_{\lfloor t \rfloor, t-\lfloor t \rfloor}\}_{t \geq 0}$ satisfies $\mathcal{B}(w, 2g_{\max}/\gamma) \subset \mathcal{U}$.

For the initial point $w_0 \in \mathcal{M}$, it holds directly due to item 1 of Assumption 2.2. Then assume that $\{w_{\lfloor t \rfloor, t-\lfloor t \rfloor}\}_{0 \leq t \leq k}$ for some $k \in \mathbb{N}$ satisfies the desired property, which validates the projection of w_k onto $\mathcal{M}$. For step $k+1$, we have that

$$\begin{aligned}\|w_{k+1} - \Phi(w_k)\| &\leq \|w_{k+1} - w_k\|_2 + \|w_k - \Phi(w_k)\|_2 = \eta \cdot \|m_{k+1}\|_2 + \|w_k - \Phi(w_k)\|_2 \\ &\leq \frac{1+\beta}{1-\beta} \cdot \frac{2}{\gamma_{\max}} \cdot g_{\max} + \frac{2g_{\max}}{\gamma} \leq \frac{400 g_{\max}}{\gamma_{\max}} + \frac{2g_{\max}}{\gamma} \leq \frac{4g_{\max}}{\gamma}\end{aligned}$$

where the second inequality $\eta \in \mathcal{I}_1$ and item 2 of Lemma H.4. Thus we have, $\mathcal{B}(w_{k+1}, 2g_{\max}/\gamma) \subset \mathcal{B}(\Phi(w_k), 6g_{\max}/\gamma) \subset \mathcal{U}$ because of item 1 of Assumption 2.2 and $\Phi(w_k) \in \mathcal{M}$. This extends the desired property from $t \in [0,k]$ to $t \in [0, k+1]$ (notice that the interpolation is a linear interpolation).

Step 2: Control the distance between w_k and $\Phi(w_k)$. By item 2 of Lemma H.4, we can obtain that for any $k \in \mathbb{N}$ and $\tau \in [0,1]$,

$$\|w_{k,\tau} - \Phi(w_{k,\tau})\|_2 \leq \frac{2\|\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})\nabla L(w_{k,\tau})\|_2}{\gamma + 2\gamma_{\mathrm{flat}}}. \tag{G.42}$$

This shows that the distance between $w_{k,\tau}$ and its projection $\Phi(w_{k,\tau})$ is bounded by the norm of the gradient in the sharp direction at $w_{k,\tau}$. With this, we now further invoke Lemma G.2 to show that the gradient is almost fully living in the flat direction, meaning that the norm of the gradient in the sharp direction is dominated by its norm in the flat direction. More specifically, with $\eta \in \mathcal{I}(\beta)$, it holds that

$$\|\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})\nabla L(w_{k,\tau})\|_2 \leq 3\mathsf{B}(\beta) \cdot \kappa^{1/2} \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})\nabla L(w_{k,\tau})\|_2, \tag{G.43}$$

for any $k \in \mathbb{N}$. With (G.42) and (G.43), we finally obtain that

$$\|w_{k,\tau} - \Phi(w_{k,\tau})\|_2 \leq \frac{6\mathsf{B}(\beta) \cdot \kappa^{1/2} \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})\nabla L(w_{k,\tau})\|_2}{\gamma + 2\gamma_{\mathrm{flat}}},$$

for any $k \in \mathbb{N}$. Combing Step 1 and Step 2, we can conclude the proof of Lemma B.2. ■

G.3.2. PROOF OF LEMMA B.3

We first give the formal version of Lemma B.3 in the following.

Lemma G.11 (Time index grows almost linearly with learning rate) *Under the conditions and assumptions of Theorem G.1, taking the learning rate η and the momentum parameter $\beta \le 0.99$ satisfying $\eta \in \mathcal{I}(\beta)$ defined in Theorem G.1, then for any iteration $k \ge \log(\beta\eta\gamma_{\text{flat}}/(1-\beta))/\log\beta$ and any $\tau \in (0,1)$, by letting $t = \tau + k$, the derivative of $T(t)$ exists and satisfies $|\mathrm{d}T/\mathrm{d}t(t) - \eta| \le \epsilon(\beta)\cdot\eta$. Equivalently, this derivative bound holds almost everywhere on each admissible interpolation interval $(k, k+1)$. At the integer breakpoints, $T(t)$ is continuous, and the corresponding one-sided derivative bounds follow by taking limits from the adjacent open intervals whenever those intervals are admissible. Here $\epsilon(\beta)$ is defined as*

$$\epsilon(\beta) := 9\kappa + \left(6\big(1+6\mathsf{C}(\beta)\big) + 4 + \frac{100\beta}{1-\beta}\right)\cdot\eta\gamma_{\text{flat}} + o(\kappa + \eta\gamma_{\text{flat}}).$$

Proof [Proof of Lemma G.11]

We prove the lemma by the following steps.

Step 1: Differentiate two sides of (B.2). For each open interpolation interval $t \in (k, k+1)$, the path $w_{\lfloor t\rfloor, t-\lfloor t\rfloor} = w_{k,t-k}$ is affine in t, and thus $\Phi(w_{\lfloor t\rfloor, t-\lfloor t\rfloor})$ is continuously differentiable as a function of t. At an integer breakpoint $t \in \mathbb{N}$, this projected path is continuous, but its derivative need not exist because the affine segment changes. Therefore, all derivative statements below are made on the open intervals $(k, k+1)$, or equivalently almost everywhere in t. The corresponding one-sided statements at breakpoints follow by taking the limits $\tau \downarrow 0$ or $\tau \uparrow 1$ from the neighboring open intervals. By the implicit function theorem, we know that $T(t)$ is also a continuously differentiable function of t whenever $t \notin \mathbb{N}$. Therefore, for any $t = k + \tau$ with $k \in \mathbb{N}$ and $\tau \in (0,1)$, we can differentiate Equation (B.2) and apply the chain rule to obtain that

$$\frac{\mathrm{d}}{\mathrm{d}T}x(T(t))\cdot\frac{\mathrm{d}T}{\mathrm{d}t}(t) = \nabla\Phi(w_{k,\tau})\frac{\mathrm{d}}{\mathrm{d}t}w_{\lfloor t\rfloor,t-\lfloor t\rfloor} = -\eta\cdot\nabla\Phi(w_{k,\tau})m_{k+1}, \tag{G.44}$$

where the last equality is due to (B.1), and $\nabla\Phi(\cdot) \in \mathbb{R}^{d\times d}$ is the Jacobian matrix of $\Phi(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^d$. With (G.44), in order to prove Lemma G.11 which controls the magnitude of $\mathrm{d}T/\mathrm{d}t(t)$, it suffices to compare the two vectors $\mathrm{d}x/\mathrm{d}T(T(t))$ and $-\nabla\Phi(w_{k,\tau})m_{k+1}$. For simplicity, we denote

$$u(t) := \frac{\mathrm{d}}{\mathrm{d}T}x(T(t)),$$

which is the tangent vector of the manifold under the parametrization (2.1) $\{x(t)\}_{t\ge 0}$ at time $T(t)$.

Step 2: Compare the two vectors $u(t)$ and $-\nabla\Phi(w_{k,\tau})m_{k+1}$. Consider the decomposition:

$$\begin{aligned}&\|u(t) - (-\nabla\Phi(w_{k,\tau})m_{k+1})\|_2\\ &\quad\le \underbrace{\|u(t) - (-\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})m_{k+1})\|_2}_{\text{Term (i)}} + \underbrace{\|\nabla\Phi(w_{k,\tau})m_{k+1} - \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})m_{k+1}\|_2}_{\text{Term (ii)}}.\end{aligned} \tag{G.45}$$

The approach is to show that along the (GD-M) trajectory the right hand side of (G.45) is bounded by $\mathcal{O}(\kappa\cdot\|u(t)\|_2)$, which means that the compared two vectors are actually quite similar. We approach Term (i) and Term (ii) in (G.45) respectively in the following.

Step 2.1: Bound Term (i) in (G.45). Bounding this term is nearly about to say that the momentum in the flat direction is similar to the tangent vector of the projection of $w_{k,\tau}$ on the river. To put it formal, consider the following decomposition,

$$\begin{aligned}\|u(t) - (-\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})m_{k+1})\|_2 \le{}& \|u(t) - (-\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})\nabla L(w_{k,\tau}))\|_2 \\ &+ \|\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})\nabla L(w_{k,\tau}) - \boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2 \\ &+ \|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k) - \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})m_{k+1}\|_2. \end{aligned} \tag{G.46}$$

The first term in (G.46) is how similar the gradient in the flat direction at $w_{k,\tau}$ is to the tangent vector of the river at $\Phi(w_{k,\tau})$ (recall that $u(t) = \mathrm{d}x/\mathrm{d}\mathrm{T}(x(T(t)))$, $x(T(t)) = \Phi(w_{k,\tau})$). The second term in (G.46) is the difference of the gradient in the flat direction between time k and $k+\tau$. The last term in (G.46) characterizes the difference between the momentum projected onto the flat direction and the gradient in the flat direction. We handle these terms respectively in the following.

For the first term, we invoke Lemma H.9 with $\mathsf{F}(\beta) := 6\mathsf{B}(\beta)$ and $\iota := \kappa^{1/2}$, which shows that given the conclusion of Lemma G.10,

$$\|u(t) - (-\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})\nabla L(w_{k,\tau}))\|_2 \le \left(4 + \frac{36\mathsf{B}^2(\beta)\cdot(1+4\kappa)}{1 - 6\mathsf{B}(\beta)\cdot\kappa^{1/2}}\right)\cdot\kappa\cdot\|u(t)\|_2, \tag{G.47}$$

For the second term in (G.46), we invoke the first conclusion of Lemma G.2, which shows that

$$\|\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})\nabla L(w_{k,\tau}) - \boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2 \le 6\eta\gamma_{\mathrm{flat}}\cdot\big(1+6\mathsf{C}(\beta)\big)\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})\nabla L(w_{k,\tau})\|_2,$$

which, combined with (G.47), further gives that

$$\begin{aligned}&\|\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})\nabla L(w_{k,\tau}) - \boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2 \\ &\quad\le 6\eta\gamma_{\mathrm{flat}}\cdot\big(1+6\mathsf{C}(\beta)\big)\cdot\left(1 + \left(4 + \frac{36\mathsf{B}^2(\beta)\cdot(1+4\kappa)}{1 - 6\mathsf{B}(\beta)\cdot\kappa^{1/2}}\right)\cdot\kappa\right)\cdot\|u(t)\|_2. \end{aligned} \tag{G.48}$$

Finally, for the last term in (G.46), we have the following upper bound,

$$\begin{aligned}&\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k) - \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})m_{k+1}\|_2 \\ &\quad\le \|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k) - \boldsymbol{P}_{\mathrm{f}}(w_k)m_{k+1}\|_2 + \|\boldsymbol{P}_{\mathrm{f}}(w_k)m_{k+1} - \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})m_{k+1}\|_2. \end{aligned} \tag{G.49}$$

For the first term on the right hand side above, we have that

$$\begin{aligned}&\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k) - \boldsymbol{P}_{\mathrm{f}}(w_k)m_{k+1}\|_2 \\ &\quad\le \left\|(1-\beta)\cdot\sum_{j=0}^{k}\beta^j\cdot\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k) - \boldsymbol{P}_{\mathrm{f}}(w_k)m_{k+1}\right\|_2 + \beta^{k+1}\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2 \\ &\quad\le \left(\frac{24\beta}{1-\beta}\cdot\eta\gamma_{\mathrm{flat}} + \beta^{k+1}\right)\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2 \\ &\quad\le \frac{25\beta}{1-\beta}\cdot\eta\gamma_{\mathrm{flat}}\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2, \end{aligned} \tag{G.50}$$

where the second inequality uses the second conclusion of Lemma G.2 and $\eta \in \mathcal{I}_3$, and the last inequality uses $k \ge \log(\beta\eta\gamma_{\mathrm{flat}}/(1-\beta))/\log\beta$. For the second term on the right hand side of

(G.49), we have

$$\begin{aligned}\|\boldsymbol{P}_{\mathrm{f}}(w_k)m_{k+1} - \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})m_{k+1}\|_2 &\le \eta\gamma\kappa\cdot\|m_{k+1}\|_2 \\ &\le \eta\gamma\kappa\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)m_{k+1}\|_2 + \eta\gamma\kappa\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2 \\ &\le \eta\gamma\kappa\cdot\left(2+\frac{25\beta}{1-\beta}\cdot\eta\gamma_{\mathrm{flat}}\right)\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2 \\ &\le \eta\gamma_{\mathrm{flat}}\cdot\left(2+\frac{25\beta}{1-\beta}\cdot\eta\gamma_{\mathrm{flat}}\right)\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2,\end{aligned} \tag{G.51}$$

where the first inequality applies Lemma H.10, the second inequality applies the last conclusion of Lemma G.2, the third inequality uses (G.50), and the last inequality uses item 2 of Assumption 2.2. With (G.50) and (G.51), we can upper bound the last term in (G.46) as

$$\begin{aligned}&\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k) - \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})m_{k+1}\|_2 \\ &\quad\le\left(2+\frac{50\beta}{1-\beta}\right)\cdot\eta\gamma_{\mathrm{flat}}\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2 \\ &\quad\le\left(2+\frac{50\beta}{1-\beta}\right)\cdot\eta\gamma_{\mathrm{flat}} \\ &\qquad\cdot\Big(1+6\eta\gamma_{\mathrm{flat}}\big(1+6\mathsf{C}(\beta)\big)\Big)\cdot\left(1+\left(4+\frac{36\mathsf{B}^2(\beta)\cdot(1+4\kappa)}{1-\mathsf{B}(\beta)\cdot\kappa^{1/2}}\right)\kappa\right)\cdot\|u(t)\|_2 \\ &\quad\le\left(4+\frac{100\beta}{1-\beta}\right)\cdot\eta\gamma_{\mathrm{flat}}\cdot\left(1+\left(4+\frac{36\mathsf{B}^2(\beta)\cdot(1+4\kappa)}{1-\mathsf{B}(\beta)\cdot\kappa^{1/2}}\right)\kappa\right)\cdot\|u(t)\|_2.\end{aligned} \tag{G.52}$$

Here the first inequality is by (G.50) and (G.51), the second inequality is by (G.47) and (G.48), and the last inequality is by $\eta\in\mathcal{I}_2$ (see the second conclusion of Proposition G.3). Consequently, combining (G.47), (G.48), and (G.52), we conclude that the right hand side of (G.46) can be upper bounded by

$$\begin{aligned}&\|u(t) - (-\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})m_{k+1})\|_2 \\ &\quad\le\left(4\kappa+\left(6\big(1+6\mathsf{C}(\beta)\big)+4+\frac{100\beta}{1-\beta}\right)\cdot\eta\gamma_{\mathrm{flat}}+o(\kappa+\eta\gamma_{\mathrm{flat}})\right)\cdot\|u(t)\|_2.\end{aligned} \tag{G.53}$$

Here $o(\cdot)$ notation hides terms with multiplicative factor $\eta\gamma_{\mathrm{flat}}\kappa$ or higher orders.

Step 2.2: Bound Term (ii) in (G.45). Bounding this term means to prove that the Jacobian of the projection onto the river is similar to the projection onto the flat direction. This is shown through Lemma H.10, by which we have that

$$\begin{aligned}\|\nabla\Phi(w_{k,\tau})m_{k+1} - \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})m_{k+1}\|_2 &\le 5\kappa\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})m_{k+1}\|_2 \\ &\le 5\kappa\cdot\big(\|\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})m_{k+1} - (-u(t))\|_2 + \|-u(t)\|_2\big) \\ &\le\Big(5\kappa+o(\kappa+\eta\gamma_{\mathrm{flat}})\Big)\cdot\|u(t)\|_2,\end{aligned}$$

where the first inequality is by Lemma H.10 and the last inequality is by Step 2.1, i.e., (G.53).

Step 3: Summary up. In conclusion, combining Step 1 and Step 2, we can arrive at

$$\left|\frac{\mathrm{d}T}{\mathrm{d}t}(t) - \eta\right| \leq \epsilon(\beta) \cdot \eta,$$

where the coefficient $\epsilon(\beta)$ is defined as following,

$$\epsilon(\beta) := 9\kappa + \left(6\big(1 + 6\mathsf{C}(\beta)\big) + 4 + \frac{100\beta}{1-\beta}\right) \cdot \eta\gamma_{\mathrm{flat}} + o(\kappa + \eta\gamma_{\mathrm{flat}}).$$

This completes the proof of Lemma B.3. ■

Appendix H. Technical Lemmas

H.1. Analysis for Eigen-Decomposition

In this section, we study the eigen-decomposition of matrix in the form of

$$\boldsymbol{T} := \begin{pmatrix} \beta & 1-\beta \\ -\beta\cdot\eta\lambda & 1-(1-\beta)\cdot\eta\lambda \end{pmatrix} \in \mathbb{R}^{2\times 2},$$

where $\eta, \lambda \geq 0$ and $\beta \in (0,1)$. The characteristic polynomial $\psi_{\boldsymbol{T}}(\alpha)$ of $\boldsymbol{T}$ is given by

$$\psi_{\boldsymbol{T}}(\alpha) = \det(\alpha\cdot\boldsymbol{I}_2 - \boldsymbol{T}) = \alpha^2 - \big((1+\beta)-(1-\beta)\cdot\eta\lambda\big)\cdot\alpha + \beta. \tag{H.1}$$

The discriminant of $\psi_{\boldsymbol{T}}$ is then given by

$$\Delta_{\boldsymbol{T}} = (1-\beta)^2\cdot(1-\eta\lambda)^2 - 4\beta(1-\beta)\cdot\eta\lambda = (1-\beta)[(\eta\lambda-1)^2 - \beta(1+\eta\lambda)^2].$$

Therefore, the two eigenvalues of $\boldsymbol{T}$ is calculated as

$$\alpha_{\boldsymbol{T},\pm} = \frac{(1+\beta)-(1-\beta)\cdot\eta\lambda \pm \sqrt{(1-\beta)^2\cdot(1-\eta\lambda)^2 - 4\beta(1-\beta)\cdot\eta\lambda}}{2} \in \mathbb{C}$$

Lemma H.1 (Schur unit-disk stability test for a quadratic) *A real-coefficient quadratic*

$$z^2 + a_1 z + a_0$$

*has both roots strictly inside the unit disk ($|z|<1$) **iff** the Jury (Schur) inequalities hold:*

$$|a_0| < 1, \quad 1 + a_1 + a_0 > 0, \quad 1 - a_1 + a_0 > 0.$$

Proposition H.2 (Spectral radius of matrix $\boldsymbol{T}$) *Regarding the matrix $\boldsymbol{T}$, the following holds:*

$$\eta\lambda < \frac{2(1+\beta)}{1-\beta} \quad \Longrightarrow \quad \rho(\boldsymbol{T}) < 1.$$

Moreover by the discrete-time Lyapunov stability criterion:

$$\rho(\boldsymbol{T}) < 1 \quad \Longleftrightarrow \quad \exists \boldsymbol{P} \succ 0 \ s.t. \ \boldsymbol{T}^\top \boldsymbol{P}\boldsymbol{T} - \boldsymbol{P} = -\boldsymbol{I} \prec 0.$$

Proof [Proof of Proposition H.2] We check the Schur unit-disk stability test Lemma H.1 for the characteristic polynomial (H.1):

$$\begin{aligned} |a_0| &= |\beta| < 1, \\ 1 + a_1 + a_0 &= 1 - (1+\beta-(1-\beta)\eta\lambda) + \beta > 0 \Leftrightarrow (1-\beta)\eta\lambda > 0, \\ 1 - a_1 + a_0 &= 1 + (1+\beta-(1-\beta)\eta\lambda) + \beta > 0 \Leftrightarrow \eta\lambda < \frac{2(1+\beta)}{1-\beta}. \end{aligned}$$

This proves Proposition H.2. ■

Proposition H.3 (Sensitivity of Lyapunov equation solution $\boldsymbol{P}$) *Given $\boldsymbol{P}_1, \boldsymbol{P}_2$ with shared parameter η and β and different λ_1, λ_2. Denote*

$$D(\lambda) := 2(1+\beta) - (1-\beta)\eta\lambda$$
$$C(\lambda_1, \lambda_2) := -\frac{2}{\lambda_1\lambda_2\eta^2} \cdot \left(2(1+\beta)^2 + \left(\beta^2 - 1\right)(\lambda_1 + \lambda_2)\eta - \beta\lambda_1\lambda_2\eta^2\right).$$

Then the difference operator norm can be bounded by:

$$\|\boldsymbol{P}_1 - \boldsymbol{P}_2\|_{\mathrm{Op}} \le \frac{|\lambda_1 - \lambda_2|\eta}{|D(\lambda_1)D(\lambda_2)|} \cdot \sqrt{\frac{8\beta^2\left(\beta^2+1\right)^2}{(1-\beta)^2}\left(\frac{2\beta^2}{(1-\beta)^2} + 1\right) + C^2(\lambda_1, \lambda_2)}.$$

Proof [Proof of Proposition H.3] By definition, we have

$$\boldsymbol{P}_1 - \boldsymbol{P}_2 = \frac{(\lambda_1 - \lambda_2)\eta}{D(\lambda_1)D(\lambda_2)} \cdot \boldsymbol{M}(\lambda_1, \lambda_2),$$
$$\boldsymbol{M}(\lambda_1, \lambda_2) := \begin{pmatrix} \frac{4\beta^2(\beta^2+1)}{(\beta-1)^2} & -\frac{2\beta(\beta^2+1)}{1-\beta} \\ \frac{2\beta(\beta^2+1)}{1-\beta} & C(\lambda_1, \lambda_2) \end{pmatrix},$$
$$C(\lambda_1, \lambda_2) := -\frac{2}{\lambda_1\lambda_2\eta^2} \cdot \left(2(1+\beta)^2 + \left(\beta^2 - 1\right)(\lambda_1 + \lambda_2)\eta - \beta\lambda_1\lambda_2\eta^2\right).$$

Therefore we have the following upper bound,

$$\begin{aligned}\|\boldsymbol{P}_1 - \boldsymbol{P}_2\|_{\mathrm{Op}} &\le \frac{|\lambda_1 - \lambda_2|\eta}{|D(\lambda_1)D(\lambda_2)|} \cdot \|\boldsymbol{M}(\lambda_1, \lambda_2)\|_F \\ &= \frac{|\lambda_1 - \lambda_2|\eta}{|D(\lambda_1)D(\lambda_2)|} \cdot \sqrt{\frac{8\beta^2\left(\beta^2+1\right)^2}{(1-\beta)^2} \cdot \left(\frac{2\beta^2}{(1-\beta)^2} + 1\right) + C^2(\lambda_1, \lambda_2)}.\end{aligned}$$

This completes the proof of Proposition H.3. ∎

H.2. Basics of the River

Recall that given $w \in \mathcal{U}$, the projection ODE flow is defined as

$$\phi(w, 0) = w, \quad \frac{\mathrm{d}}{\mathrm{d}t}\phi(w, t) = -\boldsymbol{P}_{\mathrm{s}}(\phi(w, t))\nabla L(\phi(w, t)), \quad t \ge 0. \tag{H.2}$$

Lemma H.4 (Existence of projection onto the river & further properties) *Under Assumptions 2.1 and 2.2, for any w satisfying $\mathcal{B}(w, 2g_{\max}/\gamma) \subset \mathcal{U}$, it holds that*

1. *$\Phi(w) := \lim_{t\to\infty} \phi(w, t)$ exists and $\Phi(w) \in \mathcal{M}$;*
2. *it holds that*
$$\|w - \Phi(w)\|_2 \le \frac{2\|\boldsymbol{P}_{\mathrm{s}}(w)\nabla L(w)\|_2}{\gamma + 2\gamma_{\mathrm{flat}}};$$
3. *the movement along the ODE flow (H.2) decays exponentially, that is,*
$$\|\boldsymbol{P}_{\mathrm{s}}(\phi(w, t))\nabla L(\phi(w, t))\|_2^2 \le \exp(-\gamma t/2) \cdot \|\boldsymbol{P}_{\mathrm{s}}(w)\nabla L(w)\|_2;$$
4. *finally, the Jacobian $\nabla\Phi(w)$ is well defined.*

Proof [Proof of Lemma H.4] See Lemmas C.4 and C.5 in [69] for a proof of Lemma H.4. ∎

H.3. Analysis of the River and the Projections

Lemma H.5 (River spinning) *Under Assumptions 2.1 and 2.2, it holds that for any vector v and weight w,*

$$\|\nabla \boldsymbol{P}_{\mathrm{s}}(w)[v]\|_{\mathrm{Op}} \leq \frac{\gamma\kappa}{g_{\max}} \cdot \|v\|_2.$$

The same conclusion also holds for the flat direction projection $\boldsymbol{P}_{\mathrm{f}}$.

Proof [Proof of Lemma H.5] Please refer to Lemma C.2 in [69] for a proof of Lemma H.5. ■

Lemma H.6 (Change of projection matrix) *Under Assumptions 2.1 and 2.2, for any $k \in \mathbb{N}$ and $\tau, \tau' \in [0,1]$, it holds that*

$$\|\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau}) - \boldsymbol{P}_{\mathrm{s}}(w_{k,\tau'})\|_{\mathrm{Op}} \leq \eta\gamma\kappa.$$

The same conclusion also holds for the flat direction projection $\boldsymbol{P}_{\mathrm{f}}$.

Proof [Proof of Lemma H.6] This is a direct corollary of Lemma H.5. ■

Lemma H.7 (Tangent direction of the river) *Under Assumptions 2.1 and 2.2, for any point $w \in \mathcal{M}$ on the river, it holds that*

$$\|\boldsymbol{P}_{\mathcal{M}}(w)\nabla L(w) - \nabla L(w)\|_2 \leq 4\kappa \cdot \|\boldsymbol{P}_{\mathcal{M}}(w)\nabla L(w)\|_2.$$

Proof [Proof of Lemma H.7] Please see Lemma C.10 in [69] for a proof of Lemma H.7. ■

Lemma H.8 (Auxiliary inequality 2) *Under Assumptions 2.1 and 2.2, for any w such that $\mathcal{B}(w, 2g_{\max}/\gamma) \subset \mathcal{U}$, it holds that*

$$\|\boldsymbol{P}_{\mathrm{f}}(w)\nabla L(w) - \nabla L(\Phi(w))\| \leq (\gamma_{\mathrm{flat}} + \gamma\kappa) \cdot \|w - \Phi(w)\|_2.$$

Proof [Proof of Lemma H.8] Please see Lemma C.11 in [69] for a proof of Lemma H.8. ■

Lemma H.9 (Tangent vector and gradient in the flat direction) *Under Assumptions 2.1 and 2.2, suppose that w with $\mathcal{B}(w, 2g_{\max}/\gamma) \subset \mathcal{U}$ satisfies that*

$$\|w - \Phi(w)\|_2 \leq \frac{\mathsf{F}(\beta) \cdot \iota \cdot \|\boldsymbol{P}_{\mathrm{f}}(w)\nabla L(w)\|_2}{\gamma + 2\gamma_{\mathrm{flat}}},$$

for some function $\mathsf{F}(\beta)$ and $\iota > 0$, then it holds that

$$\left\| \frac{\mathrm{d}}{\mathrm{d}T} x(T(w)) + \boldsymbol{P}_{\mathrm{f}}(w)\nabla L(w) \right\|_2 \leq \left(4\kappa + \frac{\mathsf{F}^2(\beta) \cdot \iota^2 \cdot (1+4\kappa)}{1 - \mathsf{F}(\beta) \cdot \iota} \right) \cdot \left\| \frac{\mathrm{d}}{\mathrm{d}T} x(T(w)) \right\|_2,$$

where $x(T(w)) = \Phi(w)$.

Proof [Proof of Lemma H.9] According to Assumption 2.1, the tangent vector can be alternatively represented as

$$\frac{\mathrm{d}}{\mathrm{d}T}x(t) = -\boldsymbol{P}_{\mathcal{M}}(x(t))\nabla(x(t)). \tag{H.3}$$

By Lemma H.7, we have that

$$\begin{aligned}&\|\boldsymbol{P}_{\mathcal{M}}(x(T(w)))\nabla L(x(T(w))) - \nabla L(x(T(w)))\|_2\\ &\quad\le 4\kappa\cdot\|\boldsymbol{P}_{\mathcal{M}}(x(T(w)))\nabla L(x(T(w)))\|_2.\end{aligned} \tag{H.4}$$

By combining (H.3) and (H.4), we have that

$$\begin{aligned}&\left\|\frac{\mathrm{d}}{\mathrm{d}T}x(T(w)) + \boldsymbol{P}_{\mathrm{f}}(w)\nabla L(w)\right\|_2\\ &\quad\le 4\kappa\cdot\left\|\frac{\mathrm{d}}{\mathrm{d}T}x(T(w))\right\|_2 + \|\boldsymbol{P}_{\mathrm{f}}(w)\nabla L(w) - \nabla L(\Phi(w))\|_2.\end{aligned} \tag{H.5}$$

Now invoking Lemma H.8, we have that

$$\begin{aligned}\|\boldsymbol{P}_{\mathrm{f}}(w)\nabla L(w) - \nabla L(\Phi(w))\| &\le (\gamma_{\mathrm{flat}} + \gamma\kappa)\cdot\|w - \Phi(w)\|_2\\ &\le \mathsf{F}(\beta)\cdot\iota\cdot\|\boldsymbol{P}_{\mathrm{f}}(w)\nabla(w)\|_2,\end{aligned} \tag{H.6}$$

which further gives that

$$\begin{aligned}\|\boldsymbol{P}_{\mathrm{f}}(w)\nabla(w)\|_2 &\le \frac{\mathsf{F}(\beta)\cdot\iota}{1-\mathsf{F}(\beta)\cdot\iota}\cdot\|\nabla L(\Phi(w))\|_2\\ &\le \frac{\mathsf{F}(\beta)\cdot\iota\cdot(1+4\kappa)}{1-\mathsf{F}(\beta)\cdot\iota}\cdot\left\|\frac{\mathrm{d}}{\mathrm{d}T}x(T(w))\right\|_2.\end{aligned} \tag{H.7}$$

Consequently, with (H.5), (H.6), and (H.7), we have that

$$\left\|\frac{\mathrm{d}}{\mathrm{d}T}x(T(w)) + \boldsymbol{P}_{\mathrm{f}}(w)\nabla L(w)\right\|_2 \le \left(4\kappa + \frac{\mathsf{F}^2(\beta)\cdot\iota^2\cdot(1+4\kappa)}{1-\mathsf{F}(\beta)\cdot\iota}\right)\cdot\left\|\frac{\mathrm{d}}{\mathrm{d}T}x(T(w))\right\|_2.$$

This completes the proof of Lemma H.9. ■

Lemma H.10 (Jacobian of projection to the river) *Under Assumptions 2.1 and 2.2, it holds that for any w such that $\mathcal{B}(w, 2g_{\max}/\gamma) \subset \mathcal{U}$ and any direction u,*

$$\|\nabla\Phi(w)u - \boldsymbol{P}_{\mathrm{f}}(w)u\|_2 \le 5\kappa\cdot\|\boldsymbol{P}_{\mathrm{f}}(w)u\|_2.$$

Proof [Proof of Lemma H.10] See Lemmas C.8 and C.9 in [69] for a proof of Lemma H.10. ■

H.4. Matrix Inequalities

Lemma H.11 (Eigenvalue perturbation of symmetric matrices (Corollary 4.3.15 in [25])) *Let $\mathbf{\Sigma}$ and $\widehat{\mathbf{\Sigma}} \in \mathbb{R}^{d\times d}$ be two symmetric matrices with real eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$ and $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_d$ respectively. Then for any $i \in [d]$, it holds that*

$$\left|\lambda_i - \widehat{\lambda}_i\right| \leq \left\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\right\|_{\mathrm{Op}}.$$

Lemma H.12 (Davis–Kahan $\sin(\theta)$ theorem [12]) *Let $\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ and $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_p$ respectively. Fix $1 \leq r \leq s \leq p$, let $d := s - r + 1$, and let*

$$\boldsymbol{V} = (v_r, v_{r+1}, \ldots, v_s) \in \mathbb{R}^{p\times d}, \qquad \widehat{\boldsymbol{V}} = (\widehat{v}_r, \widehat{v}_{r+1}, \ldots, \widehat{v}_s) \in \mathbb{R}^{p\times d}$$

have orthonormal columns satisfying

$$\mathbf{\Sigma} v_j = \lambda_j v_j, \qquad \widehat{\mathbf{\Sigma}}\,\widehat{v}_j = \widehat{\lambda}_j\,\widehat{v}_j, \qquad j = r, r+1, \ldots, s.$$

Define

$$\Delta := \min\Big\{\max\{0, \lambda_s - \widehat{\lambda}_{s+1}\},\ \max\{0, \widehat{\lambda}_{r-1} - \lambda_r\}\Big\},$$

where $\widehat{\lambda}_0 := +\infty$ and $\widehat{\lambda}_{p+1} := -\infty$. Then for any unitary invariant norm $\|\cdot\|_*$,

$$\Delta \cdot \left\|\sin\Theta(\boldsymbol{V}, \widehat{\boldsymbol{V}})\right\|_* \ \leq\ \left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_*.$$

Here $\Theta(\boldsymbol{V}, \widehat{\boldsymbol{V}}) \in \mathbb{R}^{d\times d}$ is diagonal with

$$\Theta(\boldsymbol{V}, \widehat{\boldsymbol{V}})_{j,j} = \arccos(\sigma_j), \quad j \in [d],$$

and $\Theta(\boldsymbol{V}, \widehat{\boldsymbol{V}})_{i,j} = 0$ for $i \neq j$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ are the singular values of $\widehat{\boldsymbol{V}}^\top \boldsymbol{V}$. The matrix $\sin\Theta(\boldsymbol{V}, \widehat{\boldsymbol{V}})$ is defined entrywise by $[\sin\Theta]_{i,j} = \sin(\Theta_{i,j})$.

Appendix I. Proofs for Improved Analysis of Vanilla GD

I.1. Formal Statement of Theorem A.1

Theorem I.1 (GD in river-valley loss landscape (improved version of Theorem 3.2 in Wen et al. 69)) *Suppose Assumptions 2.1 and 2.2 hold. Let η be a learning rate such that $\eta < \eta_{\max}^{\mathrm{GD}}$, where*

$$\eta_{\max}^{\mathrm{GD}} = \frac{1.9 - 2\gamma_{\mathrm{flat}}\gamma_{\max}^{-1} - 12\kappa}{\gamma_{\max}}.$$

Then there exists a time index T_0 such that iteration (GD) with initialization $w_0 \in \mathcal{M}$ on the river satisfies that for any step k, there exists another $T(k)$ s.t. the following two things hold:

1. *GD stays close to the river: $\|x(T_0 + T(k)) - w_k\|_2 \le \mathcal{O}(\kappa \cdot g_{\max}/\gamma)$;*
2. *The speed on river is nearly proportional to η: $|T(k) - \eta \cdot k| \le \epsilon \cdot \eta$, with $\epsilon := \mathcal{O}(\kappa + \eta\gamma_{\mathrm{flat}})$.*

I.2. Proofs for Improved Analysis

The key to the improved analysis of the largest tolerable learning rate of GD compared with [69] is from the following lemma, which establishes a tighter condition on the learning rate such the dynamics on the sharp directions do not explode. Given the following lemma, Theorem I.1 follows from a similar proof to Theorem G.1 with $\beta = 0$.

Lemma I.2 (Gradient norm in the flat direction dominates) *Under Assumptions 2.1 and 2.2, with learning rate $\eta \le (1.9 - 2\gamma_{\mathrm{flat}}\gamma_{\max}^{-1} - 12\kappa)/\gamma_{\max}$, it holds that for any $k \in \mathbb{N}$ and $\tau \in [0, 1]$,*

$$\|\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})\nabla L(w_{k,\tau})\|_2 \le 120\kappa \cdot \|\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})\nabla L(w_{k,\tau})\|_2,$$

Proof [Proof of Lemma I.2] We present the proof for $\tau = 1$, but the proof holds for general $\tau \in [0, 1]$. To facilitate presentation, we break the proof into three steps.

Step 1: upper bounding $\|\boldsymbol{P}_{\mathrm{s}}(w_{k+1})\nabla L(w_{k+1})\|_2$. By the fundamental theorem of calculus,

$$\begin{aligned}
&\boldsymbol{P}_{\mathrm{s}}(w_{k+1})\nabla L(w_{k+1}) - \boldsymbol{P}_{\mathrm{s}}(w_k)\nabla L(w_k) \\
&\quad = -\eta \cdot \int_0^1 \nabla \boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})[\nabla L(w_k)]\nabla L(w_{k,\tau}) + \boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})\nabla^2 L(w_{k,\tau})\nabla L(w_k)\mathrm{d}\tau \\
&\quad = -\eta \cdot \int_0^1 \nabla \boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})[\nabla L(w_k)]\nabla L(w_{k,\tau}) + \boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})\nabla^2 L(w_{k,\tau})\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})\nabla L(w_k)\mathrm{d}\tau \\
&\quad = -\eta \cdot \int_0^1 \boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})\nabla^2 L(w_{k,\tau})\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})\boldsymbol{P}_{\mathrm{s}}(w_k)\nabla L(w_k)\mathrm{d}\tau + \mathrm{Err}_1^{\mathrm{s}} + \mathrm{Err}_2^{\mathrm{s}}. \qquad \text{(I.1)}
\end{aligned}$$

where the first equality is from the fundamental theorem of calculus, the second equality uses the fact that $\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})\nabla^2 L(w_{k,\tau})\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau}) = \boldsymbol{0}$. The terms Err_1 and Err_2 are given respectively by

$$\begin{aligned}
\mathrm{Err}_1^{\mathrm{s}} &= -\eta \cdot \int_0^1 \boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})\nabla^2 L(w_{k,\tau})\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\mathrm{d}\tau, \\
\mathrm{Err}_2^{\mathrm{s}} &= -\eta \cdot \int_0^1 \nabla \boldsymbol{P}_{\mathrm{s}}(w_{k,\tau})[\nabla L(w_k)]\nabla L(w_{k,\tau})\mathrm{d}\tau.
\end{aligned}$$

We now upper bound the norms of Err_1 and $\mathrm{Err}_2^{\mathrm{s}}$ respectively. For $\mathrm{Err}_1^{\mathrm{s}}$, we have

$$\begin{aligned}
\|\mathrm{Err}_1^{\mathrm{s}}\|_2 &= \eta \cdot \left\| \int_0^1 \boldsymbol{P}_{\mathrm{s}}(w_{k,\tau}) \nabla^2 L(w_{k,\tau}) \boldsymbol{P}_{\mathrm{s}}(w_{k,\tau}) \boldsymbol{P}_{\mathrm{f}}(w_k) \nabla L(w_k) \mathrm{d}\tau \right\|_2 \\
&= \eta \cdot \left\| \int_0^1 \boldsymbol{P}_{\mathrm{s}}(w_{k,\tau}) \nabla^2 L(w_{k,\tau}) \Big(\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau}) - \boldsymbol{P}_{\mathrm{s}}(w_k) \Big) \boldsymbol{P}_{\mathrm{f}}(w_k) \nabla L(w_k) \mathrm{d}\tau \right\|_2 \\
&\le \eta \cdot \int_0^1 \|\nabla^2 L(w_{k,\tau})\|_2 \cdot \|\boldsymbol{P}_{\mathrm{s}}(w_{k,\tau}) - \boldsymbol{P}_{\mathrm{s}}(w_k)\|_2 \cdot \|\nabla L(w_k)\|_2 \mathrm{d}\tau \\
&\le \eta^2 \gamma_{\max} \gamma \cdot \kappa \cdot \|\nabla L(w_k)\|_2, \qquad \text{(I.2)}
\end{aligned}$$

where the last inequality uses Lemma H.10 and Assumption 2.2 (2). For Err_2, we have

$$\|\mathrm{Err}_2^{\mathrm{s}}\|_2 \le \eta\gamma \cdot \kappa \cdot \|\nabla L(w_k)\|_2, \qquad \text{(I.3)}$$

where we use Lemma H.5 and item 3 of Assumption 2.2. By (I.1), (I.2), and (I.3), we have

$$\begin{aligned}
&\|\boldsymbol{P}_{\mathrm{s}}(w_{k+1}) \nabla L(w_{k+1})\|_2 \qquad \text{(I.4)} \\
&\quad \le \left\| \left(\boldsymbol{I}_d - \eta \cdot \int_0^1 \boldsymbol{P}_{\mathrm{s}}(w_{k,\tau}) \nabla^2 L(w_{k,\tau}) \boldsymbol{P}_{\mathrm{s}}(w_{k,\tau}) \mathrm{d}\tau \right) \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \right\|_2 \\
&\qquad + \|\mathrm{Err}_1^{\mathrm{s}}\|_2 + \|\mathrm{Err}_2^{\mathrm{s}}\|_2 \\
&\quad \le \max\Big\{ \big|1 - \eta\gamma_{\max}\big|, \big|1 - \eta(\gamma + 4\gamma_{\mathrm{flat}})\big| \Big\} \cdot \|\boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k)\|_2 \\
&\qquad + \eta\gamma\big(1 + \eta\gamma_{\max}\big) \cdot \kappa \cdot \|\nabla L(w_k)\|_2.
\end{aligned}$$

This upper bounds the gradient norm in the sharp direction.

Step 2: lower bounding $\|\boldsymbol{P}_{\mathrm{f}}(w_{k+1}) \nabla L(w_{k+1})\|_2$. We use a similar way. Consider

$$\begin{aligned}
&\boldsymbol{P}_{\mathrm{f}}(w_{k+1}) \nabla L(w_{k+1}) - \boldsymbol{P}_{\mathrm{f}}(w_k) \nabla L(w_k) \\
&\quad = -\eta \cdot \int_0^1 \nabla \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau}) [\nabla L(w_k)] \nabla L(w_{k,\tau}) + \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau}) \nabla^2 L(w_{k,\tau}) \nabla L(w_k) \mathrm{d}\tau \\
&\quad = -\eta \cdot \int_0^1 \nabla \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau}) [\nabla L(w_k)] \nabla L(w_{k,\tau}) + \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau}) \nabla^2 L(w_{k,\tau}) \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau}) \nabla L(w_k) \mathrm{d}\tau \\
&\quad = -\eta \cdot \int_0^1 \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau}) \nabla^2 L(w_{k,\tau}) \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau}) \boldsymbol{P}_{\mathrm{f}}(w_k) \nabla L(w_k) \mathrm{d}\tau + \mathrm{Err}_1^{\mathrm{f}} + \mathrm{Err}_2^{\mathrm{f}}, \qquad \text{(I.5)}
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{Err}_1^{\mathrm{f}} &= -\eta \cdot \int_0^1 \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau}) \nabla^2 L(w_{k,\tau}) \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau}) \boldsymbol{P}_{\mathrm{s}}(w_k) \nabla L(w_k) \mathrm{d}\tau, \\
\mathrm{Err}_2^{\mathrm{f}} &= -\eta \cdot \int_0^1 \nabla \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau}) [\nabla L(w_k)] \nabla L(w_{k,\tau}) \mathrm{d}\tau.
\end{aligned}$$

Similar to (I.2) and (I.3), we can show that

$$\|\mathrm{Err}_1^{\mathrm{f}}\|_2 \le \eta^2 \gamma_{\max} \gamma \cdot \kappa \cdot \|\nabla L(w_k)\|_2, \quad \|\mathrm{Err}_2^{\mathrm{f}}\|_2 \le \eta\gamma \cdot \kappa \cdot \|\nabla L(w_k)\|_2. \qquad \text{(I.6)}$$

Thus combining (I.5) and (I.6), we obtain that

$$\begin{aligned}
&\|\boldsymbol{P}_{\mathrm{f}}(w_{k+1})\nabla L(w_{k+1})\|_2 \\
&\quad\geq \left\|\left(\boldsymbol{I}_d - \eta\cdot\int_0^1 \boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})\nabla^2 L(w_{k,\tau})\boldsymbol{P}_{\mathrm{f}}(w_{k,\tau})\mathrm{d}\tau\right)\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\right\|_2 - \|\mathrm{Err}_1^{\mathrm{f}}\|_2 - \|\mathrm{Err}_2^{\mathrm{f}}\|_2 \\
&\quad\geq (1-\eta\gamma_{\mathrm{flat}})\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2 - \eta\gamma\big(1+\eta\gamma_{\max}\big)\cdot\kappa\cdot\|\nabla L(w_k)\|_2. \qquad \text{(I.7)}
\end{aligned}$$

This lower bounds the gradient norm in the flat direction.

Step 3: summary up. Now with (I.4) and (I.7), we have that, for any coefficient $\alpha\in[0,1]$,

$$\begin{aligned}
&\|\boldsymbol{P}_{\mathrm{s}}(w_{k+1})\nabla L(w_{k+1})\|_2 - \alpha\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_{k+1})\nabla L(w_{k+1})\|_2 \\
&\quad\leq \max\Big\{\big|1-\eta\gamma_{\max}\big|, \big|1-\eta(\gamma+4\gamma_{\mathrm{flat}})\big|\Big\}\cdot\|\boldsymbol{P}_{\mathrm{s}}(w_k)\nabla L(w_k)\|_2 \\
&\qquad - \alpha\cdot(1-\eta\gamma_{\mathrm{flat}})\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2 + 2\eta\gamma\big(1+\eta\gamma_{\max}\big)\cdot\kappa\cdot\|\nabla L(w_k)\|_2. \qquad \text{(I.8)}
\end{aligned}$$

By the fact that

$$\|\nabla L(w_k)\|_2 \leq \|\boldsymbol{P}_{\mathrm{s}}(w_k)\nabla L(w_k)\|_2 + \|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2,$$

we can further upper bound the right hand side of (I.8) by

$$\begin{aligned}
&\|\boldsymbol{P}_{\mathrm{s}}(w_{k+1})\nabla L(w_{k+1})\|_2 - \alpha\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_{k+1})\nabla L(w_{k+1})\|_2 \\
&\quad\leq \left(2\eta\gamma\big(1+\eta\gamma_{\max}\big)\kappa + \max\Big\{\big|1-\eta\gamma_{\max}\big|, \big|1-\eta(\gamma+4\gamma_{\mathrm{flat}})\big|\Big\}\right)\cdot\|\boldsymbol{P}_{\mathrm{s}}(w_k)\nabla L(w_k)\|_2 \\
&\qquad - \Big(\alpha\cdot(1-\eta\gamma_{\mathrm{flat}}) - 2\eta\gamma\big(1+\eta\gamma_{\max}\big)\kappa\Big)\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2. \qquad \text{(I.9)}
\end{aligned}$$

Now we let the right hand side of (I.9) satisfying the equation

$$\begin{aligned}
\text{R.H.S. of (I.9)} = &\left(2\eta\gamma\big(1+\eta\gamma_{\max}\big)\kappa + \max\Big\{\big|1-\eta\gamma_{\max}\big|, \big|1-\eta(\gamma+4\gamma_{\mathrm{flat}})\big|\Big\}\right) \\
&\cdot\Big(\|\boldsymbol{P}_{\mathrm{s}}(w_k)\nabla L(w_k)\|_2 - \alpha\cdot\|\boldsymbol{P}_{\mathrm{f}}(w_k)\nabla L(w_k)\|_2\Big). \qquad \text{(I.10)}
\end{aligned}$$

and finding for sufficient conditions on η to bound α. A sufficient condition for (I.10) is

$$\begin{aligned}
&\alpha\cdot\left(2\eta\gamma\big(1+\eta\gamma_{\max}\big)\kappa + \max\Big\{\big|1-\eta\gamma_{\max}\big|, \big|1-\eta(\gamma+4\gamma_{\mathrm{flat}})\big|\Big\}\right) \\
&\quad= \alpha\cdot(1-\eta\gamma_{\mathrm{flat}}) - 2\eta\gamma\big(1+\eta\gamma_{\max}\big)\kappa,
\end{aligned}$$

or equivalently,

$$\begin{aligned}
&\left(1-\eta\gamma_{\mathrm{flat}} - 2\eta\gamma\big(1+\eta\gamma_{\max}\big)\kappa - \max\Big\{\big|1-\eta\gamma_{\max}\big|, \big|1-\eta(\gamma+4\gamma_{\mathrm{flat}})\big|\Big\}\right)\cdot\alpha \\
&\quad= 2\eta\gamma(1+\eta\gamma_{\max})\kappa.
\end{aligned}$$

Let's now narrow our focus on the regime $\eta < 2/\gamma_{\max}$ (other wise in the sharp direction the iteration explores). Consider two different sub-cases. Firstly, for $0 < \eta \le 2/(\gamma_{\max} + \gamma + 4\gamma_{\text{flat}})$, the above equation reduces to

$$\Big(\eta\gamma + 3\eta\gamma_{\text{flat}} - 2\eta\gamma\big(1+\eta\gamma_{\max}\big)\kappa\Big)\cdot\alpha = 2\eta\gamma(1+\eta\gamma_{\max})\kappa.$$

This solves α as

$$\alpha = \frac{2\gamma(1+\eta\gamma_{\max})}{\gamma + 3\gamma_{\text{flat}} - 2\gamma(1+\eta\gamma_{\max})\kappa}\cdot\kappa \le \frac{6\gamma}{\gamma - 6\gamma\kappa}\cdot\kappa \le 7\kappa$$

Secondly, for $\eta > 2/(\gamma_{\max} + \gamma + 4\gamma_{\text{flat}})$, the equation reduces to

$$\Big(2 - \eta\gamma_{\text{flat}} - 2\eta\gamma\big(1+\eta\gamma_{\max}\big)\kappa - \eta\gamma_{\max}\Big)\cdot\alpha = 2\eta\gamma(1+\eta\gamma_{\max})\kappa.$$

Letting η further satisfy

$$\eta < \frac{1.9 - 2\gamma_{\text{flat}}\gamma_{\max}^{-1} - 12\kappa}{\gamma_{\max}}, \tag{I.11}$$

then we can solve α as

$$\alpha = \frac{2\eta\gamma(1+\eta\gamma_{\max})}{2 - \eta\gamma_{\text{flat}} - 2\eta\gamma\big(1+\eta\gamma_{\max}\big)\kappa - \eta\gamma_{\max}}\cdot\kappa,$$

and under (I.11)

$$\alpha \le \frac{12}{2 - 2\gamma_{\text{flat}}\gamma_{\max}^{-1} - 12\kappa - 1.9 + 2\gamma_{\text{flat}}\gamma_{\max}^{-1} + 12\kappa}\cdot\kappa \le 120\kappa.$$

Thus in conclusion, for learning rate η satisfying (I.11), we can guarantee that

$$\begin{aligned}
&\|\boldsymbol{P}_{\text{s}}(w_{k+1})\nabla L(w_{k+1})\|_2 - \alpha\cdot\|\boldsymbol{P}_{\text{f}}(w_{k+1})\nabla L(w_{k+1})\|_2 \\
&\quad\le \left(2\eta\gamma\big(1+\eta\gamma_{\max}\big)\kappa + \max\Big\{\big|1-\eta\gamma_{\max}\big|, \big|1-\eta(\gamma+4\gamma_{\text{flat}})\big|\Big\}\right) \\
&\qquad\cdot\Big(\|\boldsymbol{P}_{\text{s}}(w_k)\nabla L(w_k)\|_2 - \alpha\cdot\|\boldsymbol{P}_{\text{f}}(w_k)\nabla L(w_k)\|_2\Big),
\end{aligned}$$

for some $\alpha \le 120\kappa$. Iterating the above inequality and use the fact that

$$\|\boldsymbol{P}_{\text{s}}(w_0)\nabla L(w_0)\|_2 - \alpha\cdot\|\boldsymbol{P}_{\text{f}}(w_0)\nabla L(w_0)\|_2 = -\alpha\cdot\|\boldsymbol{P}_{\text{f}}(w_0)\nabla L(w_0)\|_2 \le 0,$$

we can conclude the proof. ■

Appendix J. More Results and Details of Experiments

J.1. Further Experiment Details

The training set of the TinyStories dataset is of about 2.2 million instances, and the validation set is of about 22000 instances. The training loss curves are plotted every 100 steps, showing the training loss of that specific training batch. The evaluation loss curves are plotted every 5000 steps, showing the validation loss calculated on the entire validation set. All of the experiments are trained using a single NVIDIA H100 (80G PCIe) GPU.

J.2. More Experiment Results

We conduct more experiments to continue the study of Section D via more different choices of the learning rate η and the momentum β_1. Specifically, we take $\eta \in \{0.0001, 0.0002, 0.0004, 0.0006, 0.0008\}$ and $\beta_1 \in \{0, 0.85, 0.9, 0.95\}$, and plot the training loss curves and the validation loss curves.

J.2.1. FIX THE LEARNING RATE AND CHANGE THE MOMENTUM

We first fix the learning rate η and compare the loss curves under different choices of momentum β_1, see Figures 6 to 10. The results matches observations in Section D that momentum enables more stable and faster training under larger learning rates. The effect is more significant for larger learning rates.

J.2.2. FIX THE MOMENTUM AND CHANGE THE LEARNING RATE

We then fix the momentum β_1 and compare the loss curves under different choices of learning rate η, see Figures 11 to 14. The results show two trends, under the selected β_1: (i) for larger momentum, the training and validation loss curves are more stable (especially for the validation loss), demonstrating the role of stabilizer of momentum as predicted by the theory; (ii) for the same momentum β_1, during the early phase of the training, the larger the learning rate, the lower the training and validation loss, which matches our prediction of the speed on river by theory.

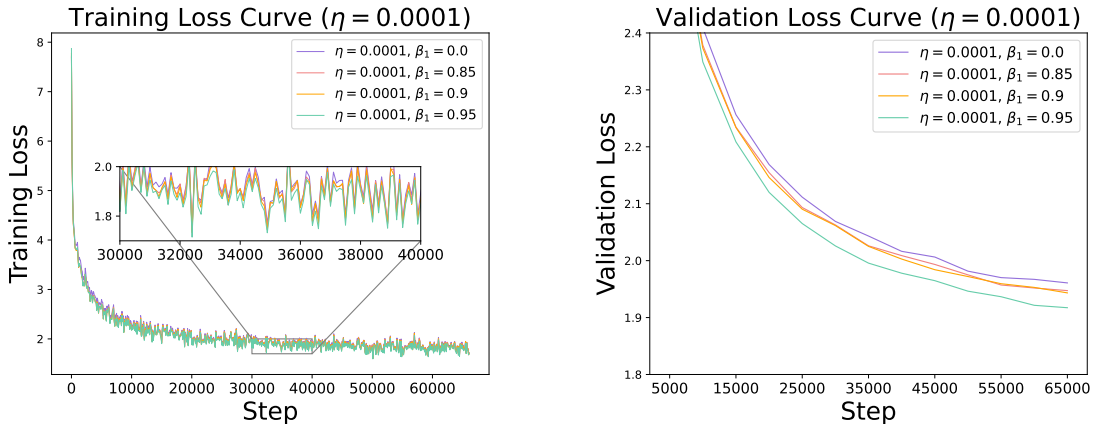

Figure 6: Training loss and validation loss under $\eta = 0.0001$ and $\beta_1 \in \{0, 0.85, 0.9, 0.95\}$.

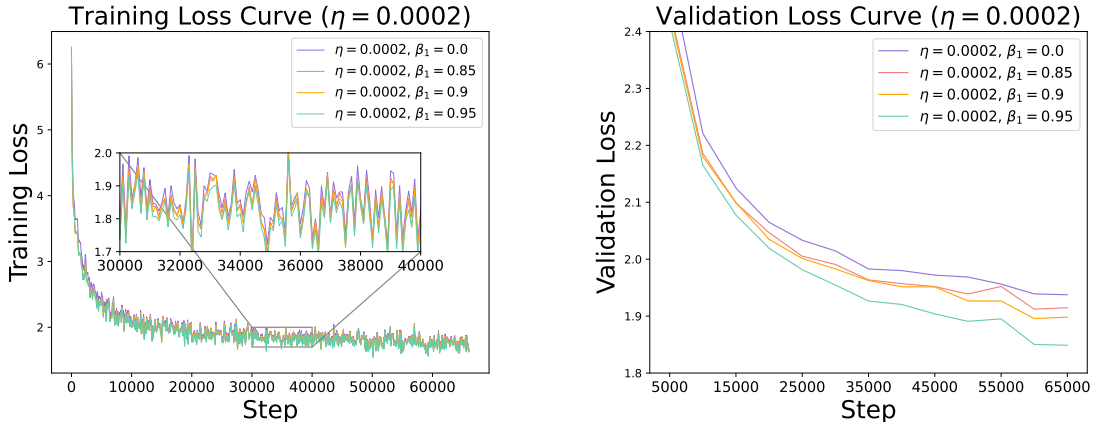

Figure 7: Training loss and validation loss under $\eta = 0.0002$ and $\beta_1 \in \{0, 0.85, 0.9, 0.95\}$.

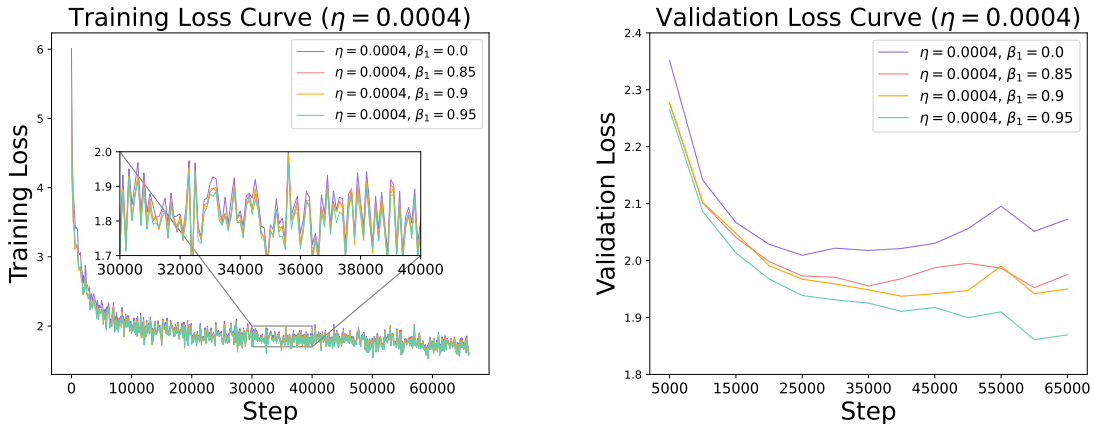

Figure 8: Training loss and validation loss under $\eta = 0.0004$ and $\beta_1 \in \{0, 0.85, 0.9, 0.95\}$.

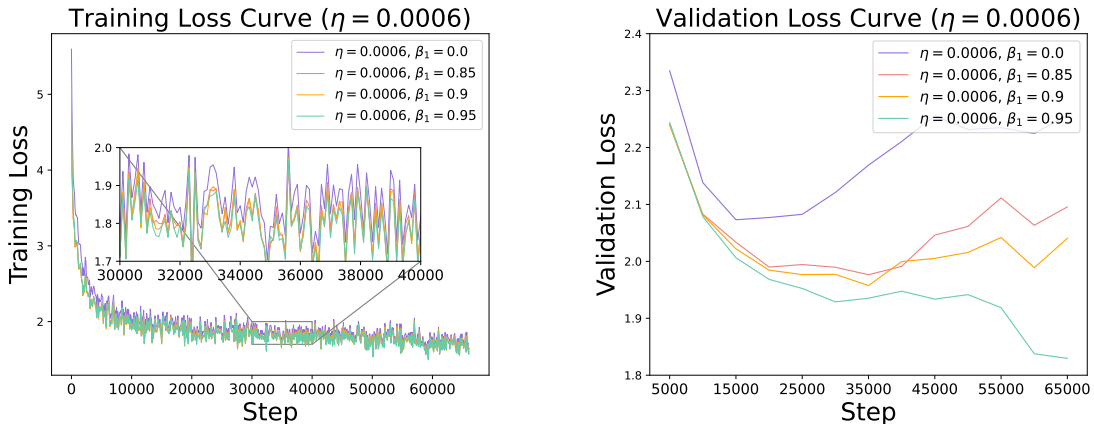

Figure 9: Training loss and validation loss under $\eta = 0.0006$ and $\beta_1 \in \{0, 0.85, 0.9, 0.95\}$.

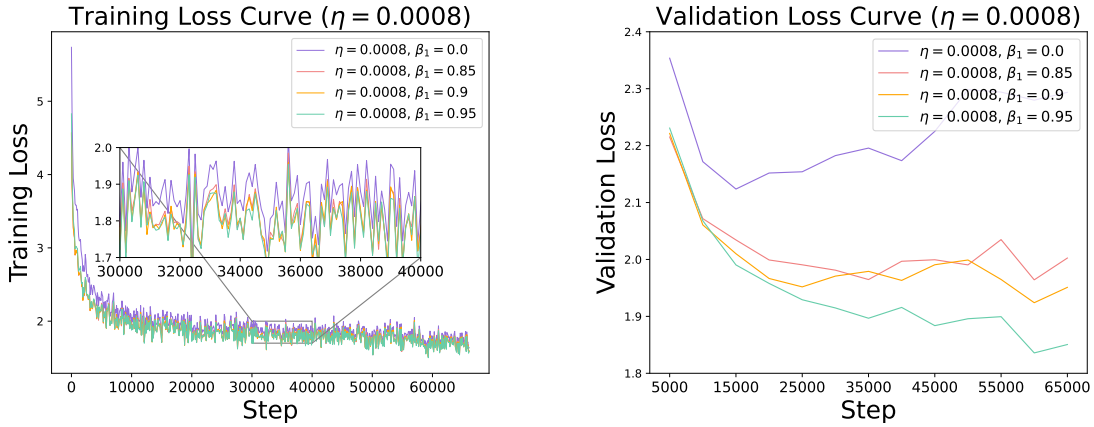

Figure 10: Training loss and validation loss under $\eta = 0.0008$ and $\beta_1 \in \{0, 0.85, 0.9, 0.95\}$.

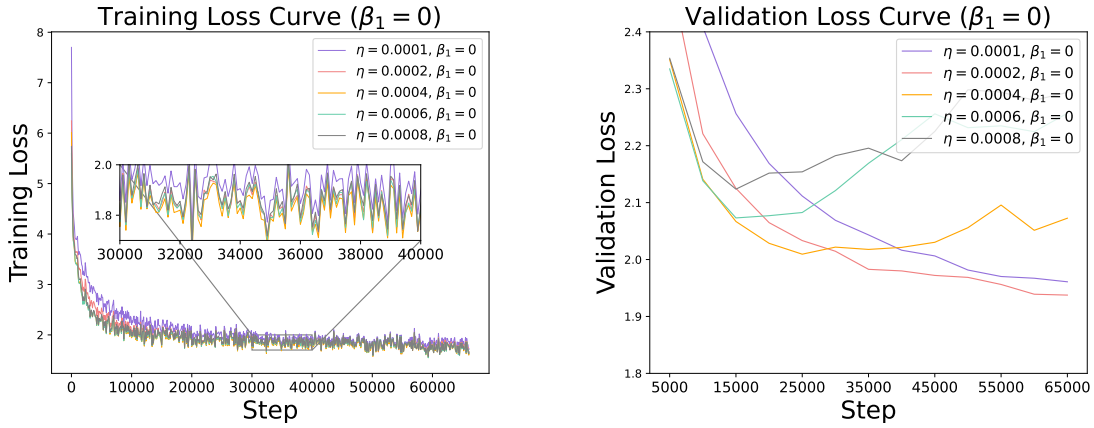

Figure 11: Training and validation loss of $\eta \in \{0.0001, 0.0002, 0.0004, 0.0006, 0.0008\}$, $\beta_1 = 0$.

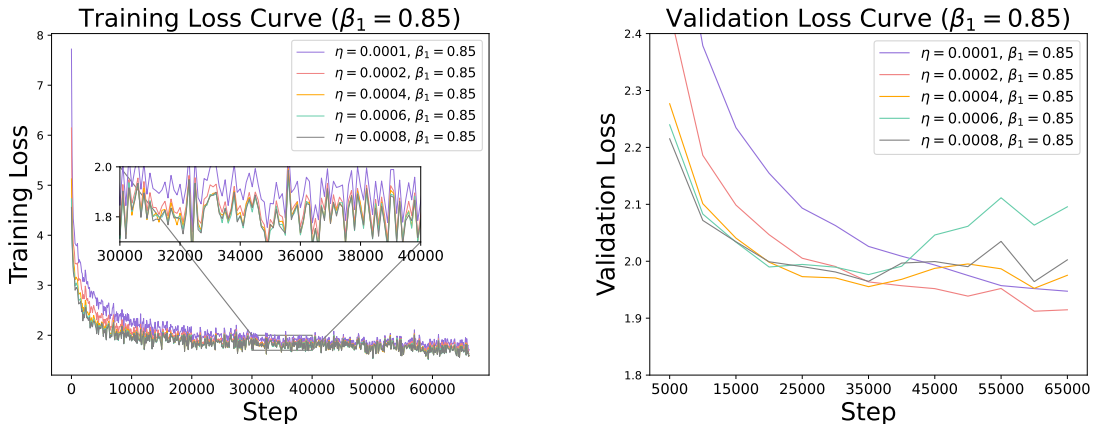

Figure 12: Training and validation loss of $\eta \in \{0.0001, 0.0002, 0.0004, 0.0006, 0.0008\}$, $\beta_1 = 0.85$.

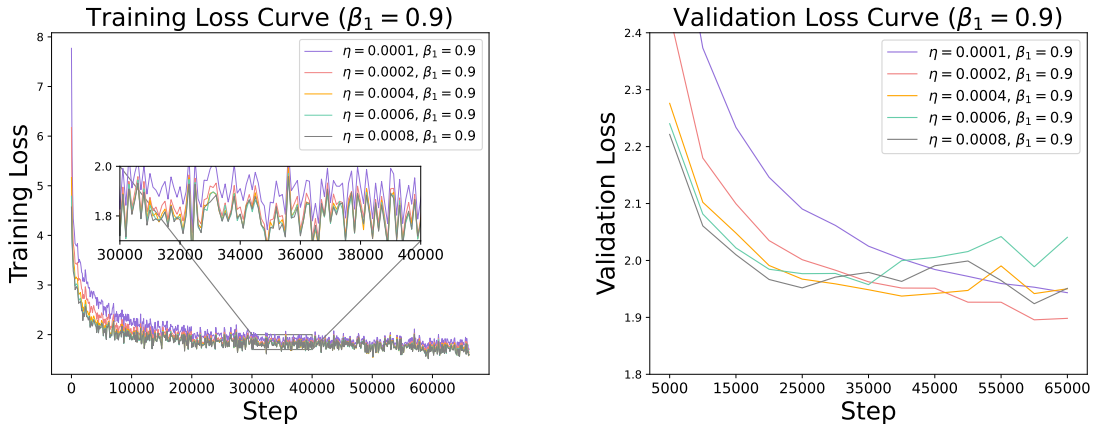

Figure 13: Training and validation loss of $\eta \in \{0.0001, 0.0002, 0.0004, 0.0006, 0.0008\}$, $\beta_1 = 0.9$.

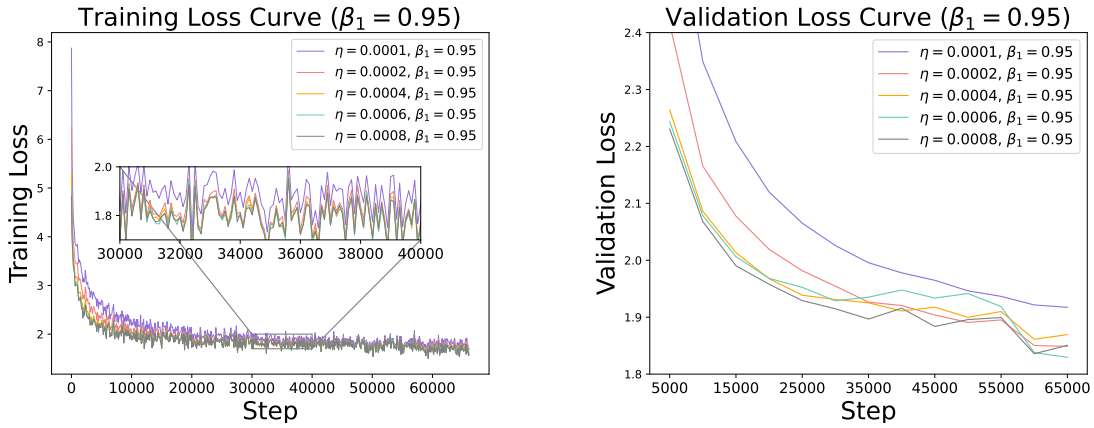

Figure 14: Training and validation loss of $\eta \in \{0.0001, 0.0002, 0.0004, 0.0006, 0.0008\}$, $\beta_1 = 0.95$.