CRIK: CONVOLUTIONAL RESHAPING ON PERTURBATION INTERVALS PARTITIONED BY KEY POINTS TO ENHANCE GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalization is a primary criterion for evaluating deep neural networks. Research shows that sharp minima in the loss landscape often lead to poor generalization. Sharpness-aware minimization (SAM) employs worst-case perturbations around minima to enhance generalization. However, SAM overlooks important geometric information about the loss landscape within the perturbation interval, which can be reflected by the key points we proposed. To address this limitation, we partition the perturbation interval using the key points and employ convolutional operations to adaptively adjust each interval's contribution to generalization performance. Furthermore, our analysis shows the impact of interval numbers on generalization. Based on this analysis, we propose convolutional reshaping on perturbation intervals partitioned by key points (CRIK). We theoretically examine CRIK's convergence, derive its escape time from sharp minima, and demonstrate that it implicitly combines gradient and Hessian information adaptively. Simulations and experiments demonstrate that CRIK improves generalization performance.

1 INTRODUCTION

Deep neural networks (DNNs) are widely used across a range of fields (Frey et al. (2026); Shandirasegaran et al. (2026)). Generalization is crucial for deploying DNNs in real-world tasks. It is shown to depend on the geometric properties of minima in the loss landscape (Wang et al. (2026)). Specifically, flat minima tend to yield good generalization, whereas sharp minima usually lead to poor generalization (Vastola (2025); Xu et al. (2026); Lei et al. (2025)).

As for sharpness-aware minimization (SAM), it improves model generalization by seeking parameters in flat loss regions (Foret et al. (2021)). In more detail, SAM formulates optimization as a min-

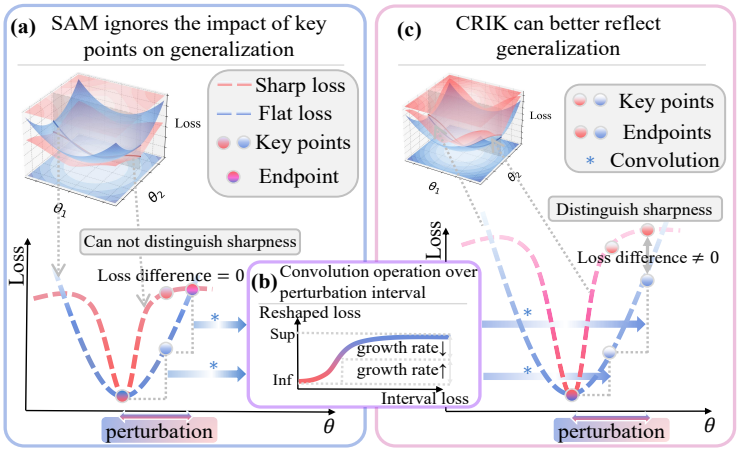


Figure 1: Framework of this paper.

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

imax problem. It identifies the worst-case perturbation within a perturbation interval of the current parameters, and then updates the model using the gradient at this perturbed point (Andriushchenko & Flammarion (2022); Chen et al. (2023)).

However, this updating procedure on the loss landscape only considers the two endpoints of the perturbation interval, i.e., the current and the worst-case point, while overlooking the key points that also govern the local geometric properties. Specifically, when the loss landscape changes nonlinearly within the perturbation interval, different geometric forms of the loss can create complex intersection patterns within that interval. As shown in Fig. 1(a), even in the simplified case where a sharp minimum and a flat minimum happen to have the same loss value at the endpoints, SAM still cannot tell them apart, making it unreliable for assessing generalization performance. To address this limitation, we propose convolutional reshaping on perturbation intervals partitioned by key points (CRIK). CRIK first identifies the key points within the perturbation interval that characterize its local geometry, partitions the interval into sub-intervals based on these points, and then applies a convolutional reshaping operation to each sub-interval, which is usually utilized in signal processing to capture the special characteristics of a signal (Parada-Mayorga et al. (2025)). This operation aggregates geometric information across the partitioned intervals, thereby amplifying the structural differences between sharp and flat minima. Consequently, CRIK captures additional landscape structure beyond the endpoints, enabling the model to escape sharp minima effectively, as shown in Fig. 1(b) and (c). To reduce the time overhead of CRIK, we employ a dynamic computation and reuse mechanism, which can be theoretically approximated as a delay stochastic differential equation (SDE). We also prove the convergence of CRIK and conduct simulations and experiments to demonstrate CRIK's effectiveness.

The key contributions of this work are summarized as follows:

- *A new understanding of SAM reflecting generalization*: By describing the necessity of key points existing within the perturbation interval, we demonstrate that SAM cannot reliably reflect generalization.
- *Multi-perspective analyses*: Applying diffusion theory and SDE, we demonstrate CRIK's superiority and discuss how the number of partitions affects the model's ability to escape sharp minima. We also derive the unique properties of the key points and discuss how CRIK can be approximated by delayed SDE. We show that CRIK is an adaptive combination of gradient and Hessian information. We also derive the generalization bound and prove the convergence.
- *A plug-and-play algorithm with computational overhead similar to SAM*: We introduce the key points and employ them to partition perturbation intervals, thereby obtaining CRIK.

2 Related Work

This section discusses the reshaping of nonlinear functions on the loss landscape, as well as SAM and its variants.

2.1 Reshaping of Nonlinear Functions

Some existing studies enhance model generalization by focusing on reshaping nonlinear functions. Drawing on this insight, a gradient activation function (GAF) that satisfies monotonicity, odd symmetry, and second-derivative-sign constraints is presented (Liu et al. (2023)). This function operates directly on gradients during training. Similarly, adaptive gradient activation (AGA) can be seamlessly embedded into the training phase of federated learning (Qin et al. (2025)). After clients compute raw gradients, these are subjected to an element-wise adaptive transformation via a tanh-based activation function. This design enables the optimizer to escape sharp minima and converge more readily toward flat loss regions (Chen et al. (Oct. 2024)).

2.2 SAM and Its Variants

SAM explicitly penalizes sharp minima and enhances generalization by minimizing the worst-case loss within a parameter neighborhood (Foret et al. (2021)). Unlike SAM, which focuses only on

the worst loss within the neighborhood, tilted SAM (TSAM) uses exponential tilt to perform soft reweighting across multiple high-loss local solutions (Li et al. (2025a)). To reduce the additional overhead of SAM, momentum SAM (MSAM) perturbs the parameters in the negative direction of the accumulated momentum vector to estimate sharpness without incurring significant additional computation or memory (Becker et al. (2025)).

3 METHOD

In this section, we discuss the construction of the CRIK algorithm and its generalization properties in detail.

3.1 ALGORITHM CONSTRUCTION

The optimization objective for SAM can be expressed as

$$\min_{\boldsymbol{\theta}} \left(\mathcal{L}\left(\boldsymbol{\theta}\right) + \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} \left(\mathcal{L}\left(\boldsymbol{\theta} + \boldsymbol{\epsilon}\right) - \mathcal{L}\left(\boldsymbol{\theta}\right) \right) \right), \tag{1}$$

where $\boldsymbol{\epsilon}$ denotes the perturbation, $\mathcal{L}$ denotes the loss function, $\boldsymbol{\theta}$ denotes the model parameters, and $||.||_2$ is the 2-norm. It can be written as

$$\mathcal{L}\left(\boldsymbol{\theta} + \boldsymbol{\epsilon}\right) - \mathcal{L}\left(\boldsymbol{\theta}\right) = \sum_{i=1}^{k} (\mathcal{L}\left(\boldsymbol{\theta} + \boldsymbol{\epsilon}_i\right) - \mathcal{L}\left(\boldsymbol{\theta} + \boldsymbol{\epsilon}_{i-1}\right)), \tag{2}$$

where $\boldsymbol{\epsilon}_0 = \mathbf{0}$, $\boldsymbol{\epsilon}_k = \boldsymbol{\epsilon}$, and $\boldsymbol{\epsilon}_i \in (\mathbf{0}, \boldsymbol{\epsilon})$. Equation (2) partitions the perturbation interval into k intervals. As for $\boldsymbol{\epsilon}_i$, we consider the interval $(\boldsymbol{\theta} + \boldsymbol{\epsilon}_{i-1}, \boldsymbol{\theta} + \boldsymbol{\epsilon}_i)$, and a Taylor expansion around $\boldsymbol{\theta} + \boldsymbol{\epsilon}_i'$ yields:

$$\begin{aligned} &\mathcal{L}\left(\boldsymbol{\theta} + \boldsymbol{\epsilon}_i\right) - \mathcal{L}\left(\boldsymbol{\theta} + \boldsymbol{\epsilon}_{i-1}\right) = \nabla \mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon}_i')^\top (\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1}) \\ &+ (\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1})^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon}_i')(\frac{\boldsymbol{\epsilon}_i + \boldsymbol{\epsilon}_{i-1}}{2} - \boldsymbol{\epsilon}_{\boldsymbol{i}}') + \mathcal{O}\left(\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{\boldsymbol{i}}'\|^2 + \|\boldsymbol{\epsilon}_{i-1} - \boldsymbol{\epsilon}_{\boldsymbol{i}}'\|^2\right), \end{aligned} \tag{3}$$

where $\boldsymbol{\epsilon}_i' \in (\boldsymbol{\epsilon}_{i-1}, \boldsymbol{\epsilon}_i)$; $\top$ is a transpose operation. Calculating the geometric properties of the perturbation interval is challenging, obtaining the key points based on these properties is also difficult (Zhang et al. (2023)). As the center point is often correlated with the geometric properties (Wu et al. (2024)), we iteratively bisect the interval $(\boldsymbol{\theta} + \boldsymbol{\epsilon}_{i-1}, \boldsymbol{\theta} + \boldsymbol{\epsilon}_i)$ that is largest and closest to $\boldsymbol{\theta}$ by inserting $\boldsymbol{\theta} + \frac{\boldsymbol{\epsilon}_{i-1} + \boldsymbol{\epsilon}_i}{2}$, yielding the following equation for $\boldsymbol{\epsilon}_i$:

$$\boldsymbol{\epsilon}_i = \frac{i + \max(i - 2k + 2^{\lfloor \log_2(2k-1) \rfloor}, 0)}{2^{\lfloor \log_2(2k-1) \rfloor}} \boldsymbol{\epsilon} := q_i \boldsymbol{\epsilon}, \tag{4}$$

where $i \in \{0, 1, \ldots, k\}$. The derivation is provided in Appendix A. Let $u_i = f(\mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon}_{i-1}))$, $v_i = \mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon}_i) - \mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon}_{i-1})$, and $\widetilde{v}_i = v_{k+1-i}$. The discrete convolution of u and $\widetilde{v}$ is

$$\omega\left(\boldsymbol{\theta}\right) = (u * \widetilde{v})[k] = \sum_{i=1}^{k} f\left(\mathcal{L}\left(\boldsymbol{\theta} + \boldsymbol{\epsilon}_{i-1}\right)\right) \left(\mathcal{L}\left(\boldsymbol{\theta} + \boldsymbol{\epsilon}_i\right) - \mathcal{L}\left(\boldsymbol{\theta} + \boldsymbol{\epsilon}_{i-1}\right)\right), \tag{5}$$

where f is an increasing function with an upper bound and positive values; $*$ is a convolution operation; u is an input signal; v is a convolution kernel (Parada-Mayorga et al. (2025)). Since $\tanh$ is a common activation function in DNNs and based on observations of $\tanh$ in (Liu et al. (2023)), we use a tanh-based function, i.e., $f(x) = \gamma_1 \tanh^{\gamma_3}(x) + \gamma_2$, where $\gamma_1 > 1$; $\gamma_2 \geq 1$; $\gamma_3 \geq 1$. Combining Eqs. (1) and (5) , and introducing λ yields

$$\hat{\boldsymbol{g}} = \nabla\left(\mathcal{L}\left(\boldsymbol{\theta}\right) + \lambda \omega\left(\boldsymbol{\theta}\right)\right) = \nabla(\mathcal{L}\left(\boldsymbol{\theta}\right)) + \lambda \nabla(\omega\left(\boldsymbol{\theta}\right)). \tag{6}$$

It can be rewritten as

$$\begin{aligned} \nabla \omega(\boldsymbol{\theta}) = \sum_{i=1}^{k} \Bigg(&\frac{\partial f(\mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon}_{i-1}))\left(\mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon}_i) - \mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon}_{i-1})\right)}{\partial \boldsymbol{\theta}} \\ &+ \frac{f(\mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon}_{i-1}))\partial\left(\mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon}_i) - \mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon}_{i-1})\right)}{\partial \boldsymbol{\theta}} \Bigg), \end{aligned} \tag{7}$$

where $\partial(\boldsymbol{\theta}+\boldsymbol{\epsilon_i})/\partial\boldsymbol{\theta}$ incurs a high computational overhead. Therefore, $\partial(\boldsymbol{\theta}+\boldsymbol{\epsilon_i})/\partial\boldsymbol{\theta} \approx \boldsymbol{I}$ is commonly adopted to simplify the derivations (Andriushchenko & Flammarion (2022)). Under this approximation, it can be written as

$$\begin{aligned}\nabla\omega\left(\boldsymbol{\theta}\right) \approx & \sum_{i=1}^{k}\left(f'(\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1}))\nabla\mathcal{L}\left(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1}\right)\ \left(\mathcal{L}\left(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i}\right)-\mathcal{L}\left(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1}\right)\right)\right) \\ & +\sum_{i=1}^{k}\left(f(\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1}))\left(\nabla\mathcal{L}\left(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i}\right)-\nabla\mathcal{L}\left(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1}\right)\right)\right).\end{aligned} \tag{8}$$

Algorithm 1 CRIK

Input: Initial parameters $\boldsymbol{\theta}_0$, the learning rate η, the perturbation size ρ, hyperparameters λ, k, and c_Δ, the total epoch E.
Initialize iteration counter $t \leftarrow 0$, epoch counter $e \leftarrow 0$, the reused gradient $\boldsymbol{g}'_t \leftarrow$ None.
while $\boldsymbol{\theta}_t$ not converge **do**
  Sample mini-batch $\mathcal{B}_t$ and compute $\boldsymbol{g}_t \leftarrow \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{B}_t}(\boldsymbol{\theta}_t)$
  Compute update interval $\Delta \leftarrow \text{round}\left(1 + c_\Delta\sqrt{1-(\text{e}/E)^2}\right)$
  Initialize perturbation: $\boldsymbol{\epsilon} = \rho\frac{\boldsymbol{g}_t}{\|\boldsymbol{g}_t\|_2}$
  Initialize $\omega(\boldsymbol{\theta}_t) \leftarrow 0$
  if $t \bmod \Delta = 0$ **then**
    for $i = 0$ **to** k **do**
      $\boldsymbol{\epsilon}_i \leftarrow \frac{i+\max(i-2k+2^{\lfloor\log_2(2k-1)\rfloor},0)}{2^{\lfloor\log_2(2k-1)\rfloor}}\boldsymbol{\epsilon}$
    end for
    for $i = 1$ **to** k **do**
      $\mathcal{L}_i \leftarrow \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{\theta}_t+\boldsymbol{\epsilon}_i),\ \mathcal{L}_{i-1} \leftarrow \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{\theta}_t+\boldsymbol{\epsilon}_{i-1})$
      $\omega(\boldsymbol{\theta}_t) \leftarrow \omega(\boldsymbol{\theta}_t) + f\left(\mathcal{L}_{i-1}\right)\left(\mathcal{L}_i - \mathcal{L}_{i-1}\right)$
    end for
    $\boldsymbol{g}'_t \leftarrow \nabla_{\boldsymbol{\theta}}\omega(\boldsymbol{\theta}_t)$
  end if
  $\hat{\boldsymbol{g}}_t \leftarrow \boldsymbol{g}_t + \lambda\boldsymbol{g}'_t$
  $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta\hat{\boldsymbol{g}}_t$
  $t \leftarrow t+1$
  if the current epoch is finished **then**
    $e \leftarrow e+1$
  end if
end while
Output: The optimized parameter $\boldsymbol{\theta}_t$.

3.2 Theoretical Analysis

We state the following properties of $\omega(\boldsymbol{\theta})$. The proofs of these lemmas and theorems are provided in Appendix C.

Lemma 3.1 *Let $\mathcal{L}:\mathbb{R}^n \to \mathbb{R}$ be a C^3 function and $H(\boldsymbol{\theta}) = \nabla^2\mathcal{L}(\boldsymbol{\theta})$ denote its Hessian. Consider the hyper-rectangle C_i with opposite vertices $(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1})$ and $(\boldsymbol{\theta}+\boldsymbol{\epsilon}_i)$, where the subscript i indexes the partition interval. Define the averaged Hessian as $\bar{H}_i(\boldsymbol{\theta}) = \frac{1}{|C_i|}\int_{C_i} H(\boldsymbol{\phi})\,d\boldsymbol{\phi}$, where $|C_i|$ denotes the volume of C_i. Let $\boldsymbol{\theta}_i^{\text{c}} = \boldsymbol{\theta} + \frac{\boldsymbol{\epsilon}_{i-1}+\boldsymbol{\epsilon}_i}{2}$ denote the geometric center of C_i, and let $\Lambda_{\max}$ and $\Lambda_{\min}$ denote the largest and smallest eigenvalues, respectively. Then, as $\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1}\|_2 \to 0$, we have $\left\|H(\boldsymbol{\theta}_i^{\text{c}}) - \bar{H}_i(\boldsymbol{\theta})\right\|_{\text{F}} = \mathcal{O}(\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1}\|_2^2)$, $\left|\Lambda_{\max}(H(\boldsymbol{\theta}_i^{\text{c}})) - \Lambda_{\max}(\bar{H}_i(\boldsymbol{\theta}))\right| = \mathcal{O}(\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1}\|_2^2)$, and $\left|\Lambda_{\min}(H(\boldsymbol{\theta}_i^{\text{c}})) - \Lambda_{\min}(\bar{H}_i(\boldsymbol{\theta}))\right| = \mathcal{O}(\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1}\|_2^2)$. Furthermore, for any other point $\boldsymbol{\theta}' = \boldsymbol{\theta} + \boldsymbol{\epsilon}'$ in C_i with $\boldsymbol{\epsilon}' \neq \frac{\boldsymbol{\epsilon}_{i-1}+\boldsymbol{\epsilon}_i}{2}$ and $\nabla^3\mathcal{L}(\boldsymbol{\theta}_i^{\text{c}}) \neq \mathbf{0}$, we have $\left\|H(\boldsymbol{\theta}') - \bar{H}_i(\boldsymbol{\theta})\right\|_{\text{F}} = \mathcal{O}(\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1}\|_2)$, $\left|\Lambda_{\max}(H(\boldsymbol{\theta}')) - \Lambda_{\max}(\bar{H}_i(\boldsymbol{\theta}))\right| = \mathcal{O}(\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1}\|_2)$, and $\left|\Lambda_{\min}(H(\boldsymbol{\theta}')) - \Lambda_{\min}(\bar{H}_i(\boldsymbol{\theta}))\right| = \mathcal{O}(\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1}\|_2)$.*

This lemma illustrates that the center points as the key points are better than the other points.

Lemma 3.2 *When $k \to \infty$, we have*

$$\lim_{k\to\infty} \omega(\boldsymbol{\theta}) = \int_0^1 f\Big(\mathcal{L}\big(\boldsymbol{\theta}+q\boldsymbol{\epsilon}\big)\Big)\Big(\nabla\mathcal{L}(\boldsymbol{\theta}+q\boldsymbol{\epsilon})\cdot\boldsymbol{\epsilon}\Big)\,dq, \tag{9}$$

where dq denotes taking the Stieltjes differential. $\lim_{k\to\infty}\omega(\boldsymbol{\theta})$ can be expressed to reflect that the contribution of a single key point progressively diminishes, as $\omega(\boldsymbol{\theta})$ tends toward a limiting value. Consequently, we treat k as a finite constant rather than an unbounded parameter, ensuring different key points have different contributions.

Lemma 3.3 *According to Eq. (5), we have*

$$\begin{aligned}
&\sum_{i=2}^{k} f(\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1}))(\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_i)-\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1})) \\
=&\underbrace{\sum_{i=2}^{k} f(\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1}))((\boldsymbol{\epsilon}_i-\boldsymbol{\epsilon}_{i-1})^\top\nabla^2\mathcal{L}(\boldsymbol{\xi}_i')(\boldsymbol{\epsilon}_i-\boldsymbol{\epsilon}_{i-1}))}_{\textit{Directional-curvature terms}} \\
&+\underbrace{\sum_{i=1}^{k-1} f(\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_i))\left(\nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\xi}_i'')^\top(\boldsymbol{\epsilon}_i-\boldsymbol{\epsilon}_{i-1})\right)}_{\textit{Directional-gradient terms}},
\end{aligned} \tag{10}$$

where $\boldsymbol{\xi}_i' \in (\boldsymbol{\epsilon}_{i-1}, \boldsymbol{\epsilon}_i)$; $\boldsymbol{\xi}_i'' \in (\boldsymbol{\epsilon}_{i-1}, \boldsymbol{\epsilon}_i)$. This lemma indicates that Eq. (5) implicitly constitutes an adaptive combination of gradient and Hessian information.

Lemma 3.4 *When $\mathcal{L}$ is γ'-Lipschitz smooth, we have*

$$\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}) \le \mathcal{L}(\boldsymbol{\theta}) + |\omega(\boldsymbol{\theta})| + \gamma'\rho(\gamma_1+\gamma_2-1). \tag{11}$$

This Lemma indicates that Eq. (5) yields a boundary that differs from SAM and reflects generalization performance. Due to the high computational overhead of $\omega(\boldsymbol{\theta})$, we calculate $\omega(\boldsymbol{\theta})$ at specified intervals, and reuse the previously computed gradients for the remaining iterations, as shown in Algorithm 1. The average number of single-step gradient computations for this dynamically updated interval is derived in Appendix B, which shows that the update interval approaches 1 in the later stages of training and incurs a computational overhead similar to SAM when c_Δ takes specific values. To show the impact of the reused gradient and update interval Δ, we use a stochastic differential equation (SDE) to analyze CRIK.

Lemma 3.5 *The piecewise-defined delayed SDE derived from CRIK is*

$$d\boldsymbol{\Theta}_t = b(\boldsymbol{\Theta}_t, \boldsymbol{\Theta}_{t_0}, r(t))dt + \sqrt{\eta}\,\sigma(\boldsymbol{\Theta}_t, r(t))dW_t, \tag{12}$$

where $\boldsymbol{\Theta}_t$ is the parameter of SDE; $r(t) = t - \lfloor t/\Delta \rfloor\Delta \in [0,\Delta); t \in [j\Delta, (j+1)\Delta); j \in \{0,1,\ldots\}$. We define $\boldsymbol{\Theta}_{t_0} = \boldsymbol{\Theta}_{j\Delta}$, then we have

$$\begin{aligned}
&b(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{t_0}, 0) = -\boldsymbol{g}(\boldsymbol{\Theta}) - \lambda\rho C_f(\boldsymbol{\Theta})\frac{H(\boldsymbol{\Theta})\boldsymbol{g}(\boldsymbol{\Theta})}{\|\boldsymbol{g}(\boldsymbol{\Theta})\|} - \frac{\lambda\rho^2}{2}\nabla\mathcal{M}(\boldsymbol{\Theta}) \\
&- \lambda\rho f'(\mathcal{L}(\boldsymbol{\Theta}))\|\boldsymbol{g}(\boldsymbol{\Theta})\|\boldsymbol{g}(\boldsymbol{\Theta}) + \mathcal{O}(\lambda\rho^3) + \mathcal{O}(\frac{\lambda\rho^2}{k}), \\
&\sigma(\boldsymbol{\Theta},0)\sigma(\boldsymbol{\Theta},0)^\top = \Sigma_{\text{SGD}}(\boldsymbol{\Theta}) + \mathcal{O}(\frac{\lambda\rho}{\sqrt{\beta}}) + \mathcal{O}(\lambda^2\rho^2) + \mathcal{O}(\frac{\lambda\rho^2}{k}), \\
&b(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{t_0}, r) = -\boldsymbol{g}(\boldsymbol{\Theta}) - \lambda\nabla\omega(\boldsymbol{\Theta}_{t_0}) + \mathcal{O}(\frac{\lambda\rho^2}{k}), \quad r \neq 0, \\
&\sigma(\boldsymbol{\Theta}, r)\sigma(\boldsymbol{\Theta}, r)^\top = \Sigma_{\text{SGD}}(\boldsymbol{\Theta}) + \lambda^2\Sigma_{\text{switch}}(\boldsymbol{\Theta}_{t_0}), \quad r \neq 0,
\end{aligned} \tag{13}$$

where $\boldsymbol{g}(\boldsymbol{\Theta}) = \nabla\mathcal{L}(\boldsymbol{\Theta})$; $H(\boldsymbol{\Theta}) = \nabla^2\mathcal{L}(\boldsymbol{\Theta})$; $C_f(\boldsymbol{\Theta}) = f\big(\mathcal{L}(\boldsymbol{\Theta})\big) + \|\boldsymbol{g}(\boldsymbol{\Theta})\| f'\big(\mathcal{L}(\boldsymbol{\Theta})\big)$; $\mathcal{M}(\boldsymbol{\Theta}) = f\big(\mathcal{L}(\boldsymbol{\Theta})\big)\boldsymbol{g}(\boldsymbol{\Theta})^\top H\boldsymbol{g}(\boldsymbol{\Theta})/\|\boldsymbol{g}(\boldsymbol{\Theta})\|^2 + f'\big(\mathcal{L}(\boldsymbol{\Theta})\big)\|\boldsymbol{g}(\boldsymbol{\Theta})\|^2$. $\Sigma_{\text{SGD}}(\boldsymbol{\Theta})$ is the mini-batch gradient covariance; $\Sigma_{\text{switch}}(\boldsymbol{\Theta}_{t_0}) = \mathcal{O}(\rho)$ is the switching-step covariance evaluated at $\boldsymbol{\Theta}_{t_0}$. $\mathcal{O}(\lambda\rho^2/k)$ arises from the binary partition approximation error, where k is the number of intervals. We assume $\lambda = \mathcal{O}(1)$ and $\beta = \mathcal{O}(1)$ are constants independent of η.

Lemma 3.6 *When $\rho = \mathcal{O}(\sqrt{\eta})$; mini-batch size β and $\Delta = \mathcal{O}(1)$ are fixed; $k = \mathcal{O}(1/\eta)$ as $\eta \to 0$, the constructed piecewise-defined delayed SDE is a weak approximation of CRIK, in the sense that for any test function $\phi \in C_b^4(\mathbb{R}^n)$, there exists a constant $C(T, \phi)$ such that*

$$\big|\mathbb{E}[\phi(\boldsymbol{\theta}_N)] - \mathbb{E}[\phi(\boldsymbol{\Theta}_{N\eta})]\big| \le C(T, \phi) \cdot \sqrt{\eta}, \tag{14}$$

where $N\eta \le T$.

These lemmas show that CRIK's reuse mechanism can be approximated in a distributional sense by a delayed piecewise SDE. The drift term in the switching step includes curvature-aware adjustments that actively penalize sharp directions; the reuse step reduces to a deterministic descent based on reused gradients. At the same time, as long as the delayed SDE converges to a stationary point, this strategy does not affect the overall convergence order.

Assumption 3.7 (Lipschitz Smoothness) *The gradient $\nabla\mathcal{L}(\boldsymbol{\theta})$ is γ-Lipschitz smooth, and $\mathcal{L}$ is γ'-Lipschitz smooth.*

Assumption 3.8 (Boundedness) *The loss function is bounded: $|\mathcal{L}(\boldsymbol{\theta})| \le m$.*

Lemma 3.9 *According to these Assumptions, we have $|\omega(\boldsymbol{\theta})| \le m_\omega, \quad \|\nabla\omega(\boldsymbol{\theta})\| \le h$, and $\nabla\omega(\boldsymbol{\theta})$ is μ-Lipschitz smooth, where m_ω, h, and μ depend on γ, γ', γ_1, γ_2, γ_3, k, ρ, and m.*

Theorem 3.10 (Convergence of CRIK) *Under these Assumptions, with the learning rate $\eta_t = \eta_0/\sqrt{t}$, CRIK achieves the following convergence rate:*

$$\frac{1}{t'}\sum_{t=1}^{t'}\mathbb{E}\left[\left\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_t)\right\|^2\right] \le \frac{c_1 + c_2(1+\log t')}{\sqrt{t'}}, \tag{15}$$

where $c_1 = 2(m + \lambda m_\omega)/\eta_0$ and $c_2 = (\gamma + \lambda\mu)\eta_0(\|\boldsymbol{g}\|^2 + \lambda^2\|\boldsymbol{h}\|^2)$.

When $\omega(\boldsymbol{\theta})$ is computed at every iteration, CRIK minimizes $\hat{\mathcal{L}}(\boldsymbol{\theta}_t) = \mathcal{L}(\boldsymbol{\theta}_t) + \lambda\omega(\boldsymbol{\theta}_t)$ and $\hat{\boldsymbol{g}}_t = \boldsymbol{g}_t + \lambda\nabla\omega(\boldsymbol{\theta}_t)$. When computing $\omega(\boldsymbol{\theta})$ at each iteration, we present the generalization bound for CRIK.

Theorem 3.11 (Generalization Bound for CRIK) *Suppose the per-sample loss function $\mathcal{L}$ is differentiable and bounded by m. Then with probability at least $1-\delta$ over the training set generated from the distribution:*

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\epsilon}'\sim\mathcal{N}(0,\sigma'^2)}\left[\mathcal{L}'(\boldsymbol{\theta}+\boldsymbol{\epsilon}')\right] \\
&\le \mathcal{L}(\boldsymbol{\theta}) + |\omega(\boldsymbol{\theta})| + \gamma'\rho(\gamma_1+\gamma_2-1) + \frac{m}{\sqrt{s}} + \\
&\sqrt{\frac{\frac{n\log\left(1+\frac{\|\boldsymbol{\theta}\|_2^2(\sqrt{n}+\sqrt{\log s})^2}{n\rho^2}\right)+1}{4} + \log\frac{s}{\delta} + 2\log(6s+3n)}{s-1}},
\end{aligned} \tag{16}$$

where $\sigma' = \rho/(\sqrt{n}+\sqrt{\log s})$; $\mathcal{L}'$ denotes the test loss; s is the total number of samples; m is the upper bound of the loss; n is the dimension of $\boldsymbol{\theta}$.

Based on Kramers' stochastic escape theory, we obtain the Fokker-Planck equation in a high-dimensional phase space of parameters (Xie et al. (2022)). Under this framework, the problem of escaping sharp minima during optimization is transformed into a Kramers escape problem, and the mean escape time is then obtained.

Assumption 3.12 (Second Order Taylor Approximation) *The loss landscape around the parameter $\boldsymbol{\theta}_0$ can be written as $\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta_0}) + \boldsymbol{g}(\boldsymbol{\theta_0})^\top(\boldsymbol{\theta} - \boldsymbol{\theta_0}) + 1/2(\boldsymbol{\theta} - \boldsymbol{\theta_0})^\top H(\boldsymbol{\theta_0})(\boldsymbol{\theta} - \boldsymbol{\theta_0})$, where $\boldsymbol{g}(\boldsymbol{\theta_0})$ is the gradient and $H(\boldsymbol{\theta_0})$ is the Hessian matrix.*

Assumption 3.13 (Quasi-Equilibrium Approximation) *The system is in a quasi-equilibrium state near the minimum, and the parameter distribution obeys the Boltzmann distribution.*

Assumption 3.14 (Low Temperature Approximation) *The escape dynamics are mainly governed by the barrier's shape and height, rather than by thermal noise; i.e., gradient noise is negligible during training.*

Theorem 3.15 *Assume that only one most probable escape path exists from a sharp minimum $\boldsymbol{a}$ to the region outside its basin of attraction, passing through a saddle point $\boldsymbol{b}$ along an escape direction $\boldsymbol{e}$. Under these Assumptions, if CRIK governs optimization dynamics and computes $\omega(\boldsymbol{\theta})$ at each iteration, the mean escape time τ for CRIK to escape from such a sharp minimum is*

$$\tau = \frac{2\pi}{|H_{be}|} \exp\left(\frac{2\beta\Delta\hat{\mathcal{L}}}{\eta}\left(\frac{p}{H_{ae}} + \frac{1-p}{|H_{be}|}\right)\right), \tag{17}$$

where $\Delta\hat{\mathcal{L}} = \Delta\mathcal{L} - \rho^2\lambda/2(f(\mathcal{L}(\boldsymbol{a}))\lambda_a - f(\mathcal{L}(\boldsymbol{b}))\lambda_b)$; $p \in (0,1)$ is a path-dependent parameter; λ_a is a dominant eigenvalue at the sharp minimum $\boldsymbol{a}$; λ_b is a dominant eigenvalue at the saddle point $\boldsymbol{b}$; $\Delta\mathcal{L}$ is the barrier height $\mathcal{L}(\boldsymbol{b}) - \mathcal{L}(\boldsymbol{a})$; β is the batch size. A detailed analysis is presented in Appendix D.

3.3 Simulations

To demonstrate generalization performance of the optimizers, we use CRIK and SAM to find minima on two-dimensional loss landscapes f_1 and f_2:

$$\begin{aligned} f_1(X,Y) = &-\log\left(0.9\exp\left(-\frac{(X-1.5)^2+(Y-2.5)^2}{1.8^2}\right)\right. \\ &\left.+0.2\exp\left(-\frac{(X+1)^2+(Y+2.5)^2}{0.5^2}\right)\right), \\ f_2(X,Y) = &\frac{15\log(\sin(47\sin X)+1.2)+\sin(10X)\sin(6Y)}{100} \\ &+\frac{30\log\left(\sin\left(47\sin Y\right)+1.2\right)+\cos(10X)\cos(6Y)}{100}. \end{aligned}$$

As shown in Fig. 2, CRIK escapes from sharp minima more effectively and converges to flatter minima than SAM. By escaping sharp minima, CRIK demonstrates better generalization.

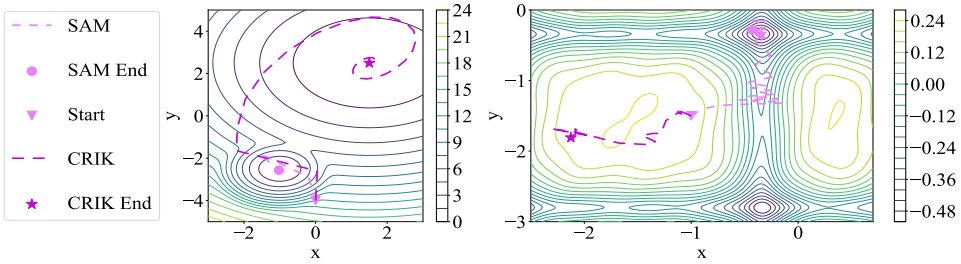


Figure 2: Optimization trajectories on f_1 and f_2.

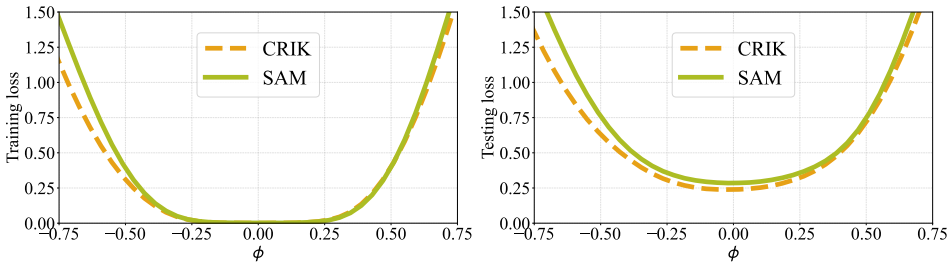
3.4 Experiments on Image Classification

The CIFAR-10, CIFAR-100, and ImageNet datasets are widely used in computer vision and pattern recognition (Wang et al. (2026)). To demonstrate the advantages over other optimizers, we compare

Table 1: Image classification accuracies (Mean ± Std) using ResNet18.

CIFAR-10		CIFAR-100	
Method	Test Accuracy (%)	Method	Test Accuracy (%)
SAM	$95.88_{\pm 0.06}$	SAM	$78.31_{\pm 0.09}$
LPF	$95.19_{\pm 0.11}$	GAF	$77.50_{\pm 0.27}$
FSGDM	$95.81_{\pm 0.04}$	AGA	$78.13_{\pm 0.33}$
MSAM	$95.26_{\pm 0.15}$	TSAM	$78.70_{\pm 0.29}$
TSAM	$95.97_{\pm 0.08}$	Adam	$72.86_{\pm 0.06}$
CRIK	$95.91_{\pm 0.03}$	CRIK	$78.50_{\pm 0.16}$
CRIK + MSAM	$95.63_{\pm 0.17}$	CRIK + TSAM	$\mathbf{79.22}_{\pm 0.21}$
CRIK + TSAM	$\mathbf{96.03}_{\pm 0.09}$	CRIK + Adam	$72.95_{\pm 0.21}$

CRIK with SAM and several recent variants, including MSAM (Becker et al. (2025)), TSAM (Li et al. (2025a)), GAF (Liu et al. (2023)), AGA (Qin et al. (2025)), LLR (Chen et al. (Oct. 2024)), low-pass filter (LPF) (Bisla et al. (2022)), frequency stochastic gradient descent with momentum (FSGDM) (Li et al. (2025b)), adaptive moment estimation (Adam) (Zhang et al. (2025)), and relativistic adaptive gradient descent (RAD) (Lyu et al. (2025)). We evaluate optimizers on ResNet18 (Vastola (2025)), ResNet50 (Izzo et al. (2026)), and RepVGG-A1 (Lei et al. (2025)). On CIFAR-10 and CIFAR-100, all models are trained for 200 epochs with a mini-batch size of 128, and each experiment is repeated three times. The initial learning rates for SGD-based and Adam-based optimizers are 0.1 and 1e-3, respectively. All other hyperparameters for each optimizer are kept at their original published values. As shown in Tables 1 and 6, CRIK exhibits higher accuracy. Following the method in (Li et al. (2018)), we further evaluate the sharpness of the loss landscape using VGG9 on the CIFAR-10 dataset. As illustrated in Fig. 3, CRIK consistently converges toward flat minima. More experiments and sensitivity of hyperparameters are presented in Appendix E.



(a) (b)

Figure 3: The loss landscape is visualized. The loss is defined as $\mathcal{L}(\boldsymbol{\theta}^* + \phi \boldsymbol{l})$, where ϕ is a scalar that perturbs the optimal parameters ϕ by scaling the random vector l taken from a Gaussian distribution.

3.5 Experiments on Self-Supervised Learning

EB-JEPA constitutes a joint embedding prediction architecture (JEPA) in its fully trained form (Terver et al. (2026)). Its image-focused submodule, Image-JEPA, is used for self-supervised pre-training on raw image inputs and excels at extracting feature representations robust to viewpoint changes. Table 2 presents two distinct regularization strategies designed to avoid collapsed latent representations: variance-invariance-covariance regularization (VICReg) and slice-information Gaussian

Table 2: Linear probing accuracy (%) on CIFAR-10 With ResNet18.

Model	SIGReg	VICReg
SAM	89.49	89.91
TSAM	89.23	90.35
CRIK	89.51	90.34
CRIK + TSAM	**89.72**	**90.62**

regularization (SIGReg). Empirical results show that CRIK achieves a higher linear probing accuracy after 300 training epochs.

3.6 Experiments on Language Modeling

WikiText-103, a widely used large corpus for long-sequence language modeling with more than 100 million tokens, is selected as the experimental dataset (Merity et al. (2017)). Pythia-160M is a 160M-parameter autoregressive Transformer model from the Pythia suite, pretrained on the Pile for language modeling. The Pythia-160M network is trained with preset hyperparameter configurations: a batch size of 8, a constant learning rate of 1e-4, and 20 epochs. The curves in Fig. 4 show that CRIK achieves lower perplexity values and effectively mitigates overfitting during training compared with vanilla Adam, thereby confirming its stronger generalization performance.

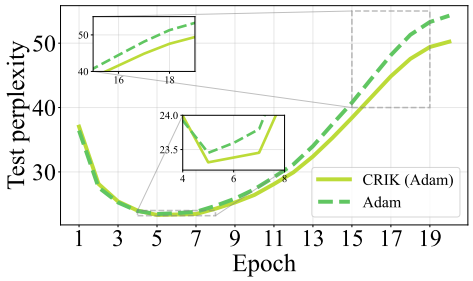


Figure 4: Perplexity on Pythia-160M.

3.7 Experiments on Time Series

This medium-sized time-series benchmark is derived from a traffic dataset containing approximately 15.12 million individual data samples. We adopt the four transformer-based architectures for our empirical tests, namely Informer(Li et al. (2021)), Transformer (Vaswani et al. (2017)), Flowformer (Wu et al. (2022)), and Flashformer (Dao et al. (2022)). Mean absolute error (MAE) is adopted as a core evaluation metric to quantify the predictive capability of all the candidate models, and we apply CRIK to Adam and RAD to train them. The outcomes summarized in Table 3 demonstrate that CRIK outperforms the conventional baseline algorithms.

Table 3: MAE on Traffic.

	Transformer	Informer	Flowformer	Flashformer
Adam	0.4907	0.5185	0.4933	0.4940
RAD	0.5271	0.5291	0.5150	0.5201
CRIK + Adam	**0.4804**	**0.5106**	**0.4757**	**0.4887**
CRIK + RAD	0.4935	0.5204	0.4990	0.4933
Best increment	6.37%	1.64%	3.57%	5.15%

4 Conclusion

This work has proposed CRIK, an optimizer that enhances generalization performance by partitioning perturbation intervals around the key points and employing convolutional reshaping. SAM has ignored the impact of the key points within the perturbation intervals on generalization. In contrast, CRIK partitions the intervals using the key points and amplifies the differences between the intervals via convolutional operations, enabling the model to escape from sharp minima more effectively. Theoretical analyses have shown that CRIK converges. We have derived the escape time from sharp minima and generalization bounds, demonstrating that CRIK constitutes an adaptive combination of the gradient and Hessian information. Experiments have shown that CRIK has a strong ability to escape sharp minima. Although the lemma and some experiments have shown the advantages of using center points as key points, whether geometrically adaptable points can exhibit better generalization performance remains an important issue. Future work will focus on the impact of different types of key points on generalization.

AI USE STATEMENT

In this study, we strictly follow the ICLR 2027 author policy and responsibly use generative artificial intelligence tools to assist with this research work. Specifically, we use major language models such as GPT and Claude for initial construction and debugging of the code framework, syntax optimization and expression polishing of the paper language, and retrieval and organization of relevant literature. All AI-generated content, including code snippets, text expressions, and literature references, underwent item-by-item review, manual verification, and necessary modifications by all authors to ensure technical accuracy and scientific reliability. We take full responsibility for all aspects of the work, including AI-assisted generation, and have ensured the authenticity, completeness, and reproducibility of the research results through internal cross-validation.

REPRODUCIBILITY STATEMENT

Section 3 defines the construction of the CRIK algorithm, the strategy for partitioning perturbation intervals, the convolution reshaping operation, and the theoretical analysis of convergence and generalization bounds. Sections 3.3 to 3.7 provide detailed baseline models for tasks such as 2D loss landscapes, image classification, self-supervised learning, language modeling, and time-series prediction. The appendix provides a detailed derivation of key points, key hyperparameter settings, detailed mathematical proofs, and additional experimental results. The simulated trajectory, algorithm pseudocode, experimental environment configuration, and AI-assisted usage statement in the manuscript support internal traceability. This manuscript is currently undergoing double-blind review and does not claim to have undergone public anonymous data release or independent reproduction.

REFERENCES

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *International Conference on Machine Learning*, pp. 639–668, 2022.

Marlon Becker, Frederick Altrock, and Benjamin Risse. Momentum-SAM: Sharpness aware minimization without computational overhead. In *Advances in Neural Information Processing Systems*, pp. 55336–55366, 2025.

Devansh Bisla, Jing Wang, and Anna Choromanska. Low-pass filtering SGD for recovering flat optima in the deep learning optimization landscape. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 8299–8339, 2022.

Liangming Chen, Long Jin, and Mingsheng Shang. Efficient loss landscape reshaping for convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7):12 535–12549, Oct. 2024.

Zixiang Chen, Junkai Zhang, Yiwen Kou, Xiangning Chen, Cho-Jui Hsieh, and Quanquan Gu. Why does sharpness-aware minimization generalize better than SGD? In *Advances in Neural Information Processing Systems*, volume 36, pp. 72325–72376, 2023.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/67d57c32e20fd0a7a302cb81d36e40d5-Abstract-Conference.html.

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

Matthias Frey, Jingge Zhu, and Michael Gastpar. Block-sample MAC-bayes generalization bounds. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=pelxC06W8U.

Zachary Izzo, Eshaan Nichani, and Jason D. Lee. Quantitative bounds for length generalization in transformers. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=TLSUIyBIfs.

Shiye Lei, Fengxiang He, Haowen Chen, and Dacheng Tao. Attentive learning facilitates generalization of neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2): 3329–3342, 2025.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, pp. 1–11, 2018.

Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *arXiv preprint*, 2021. URL https://arxiv.org/abs/2012.07436.

Tian Li, Tianyi Zhou, and Jeffrey A. Bilmes. Tilted sharpness-aware minimization. In *International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=bwfiFv8KG2.

Xianliang Li, Jun Luo, Zhiwei Zheng, Hanxiao Wang, Li Luo, Lingkun Wen, Linlong Wu, and Sheng Xu. On the performance analysis of momentum method: A frequency domain perspective. In *International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=tznvtmSEiN.

Mei Liu, Liangming Chen, Xiaohao Du, Long Jin, and Mingsheng Shang. Activated gradients for deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 34(4): 2156–2168, 2023. doi: 10.1109/TNNLS.2021.3106044.

Yao Lyu, Xiangteng Zhang, Shengbo Eben Li, Jiangtao Duan, Letian Tao, Qi Xu, Lei He, and Keqiang Li. Conformal symplectic optimization for stable reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):11049–11063, 2025.

S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Alejandro Parada-Mayorga, Leopoldo Agorio, Alejandro Ribeiro, and Juan Bazerque. Convolutional filtering with rkhs algebras. *IEEE Transactions on Signal Processing*, 73:2353–2367, 2025.

Chuan Qin, Mei Liu, and Long Jin. Adaptive gradient activation for federated learning on consumer electronic devices. *IEEE Transactions on Consumer Electronics*, 71(2):5489–5500, 2025. doi: 10.1109/TCE.2025.3539274.

Mugunthan Shandirasegaran, Hongkang Li, Songyang Zhang, Meng Wang, and Shuai Zhang. A theoretical analysis of mamba's training dynamics: Filtering relevant features for generalization in state space models. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=hvpKqEYJjj.

Basile Terver, Randall Balestriero, Megi Dervishi, David Fan, Quentin Garrido, Tushar Nagarajan, Koustuv Sinha, Wancong Zhang, Michael Rabbat, Yann LeCun, and Amir Bar. A lightweight library for energy-based joint-embedding predictive architectures. *arXiv preprint arXiv:2602.03604*, 2026. URL https://arxiv.org/abs/2602.03604.

John Vastola. Generalization through variance: how noise shapes inductive biases in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7lUdo8Vuqa.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Siqi Wang, Aoming Liu, and Bryan A. Plummer. Noise-aware generalization: Robustness to in-domain noise and out-of-domain generalization. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=wb83wO41QT.

Haixu Wu, Jialong Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Flowformer: Linearizing transformers with conservation flows. In *International Conference on Machine Learning*, 2022. URL https://proceedings.mlr.press/v162/wu22m.html.

Tao Wu, Tie Luo, and Donald C. Wunsch II. CR-SAM: Curvature regularized sharpness-aware minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6144–6152, 2024.

Zeke Xie, Xinrui Wang, Huishuai Zhang, Issei Sato, and Masashi Sugiyama. Adaptive inertia: Disentangling the effects of adaptive learning rate and momentum. In *International Conference on Machine Learning*, pp. 24430–24459, 2022.

Zhipeng Xu, Zilong Wang, XINYANG JIANG, Dongsheng Li, De Cheng, and Nannan Wang. Reasoning-driven multimodal LLM for domain generalization. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=psJiUopUt7.

Xingxuan Zhang, Renzhe Xu, Han Yu, Hao Zou, and Peng Cui. Gradient norm aware minimization seeks first-order flatness and improves generalization. In *Proc. Comput. Vis. Pattern Recognit.*, pp. 20247–20257, 2023.

Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik P. Kingma, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=iBExhaU3Lc.

Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha C. Dvornek, Sekhar Tatikonda, James Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=edONMAnhLu-.

A Derivation of Key Points

Continuously inserting $\boldsymbol{\theta}+(\boldsymbol{\epsilon}_{i-1}+\boldsymbol{\epsilon}_i)/2$ into the largest interval in $(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1}, \boldsymbol{\theta}+\boldsymbol{\epsilon}_i)$ that is closest to $\boldsymbol{\theta}$ can be understood as a binary tree whose nodes' values are intervals, by performing node split operations according to level-order traversal, as shown in Fig. 5.

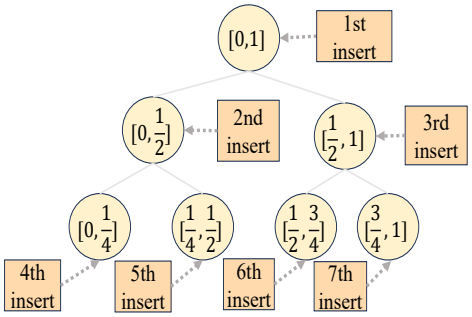


Figure 5: Binary tree whose nodes' values are intervals.

When we use k intervals, we obtain $2k-1$ nodes. Since a level-order traversal of the binary tree is performed, each insertion will only form a complete binary tree. According to the properties of complete binary trees, the depth of the complete binary tree is calculated as $d=\log_2(2k-1)+1$; the

deepest and shallowest leaf nodes differ by at most tone level; the interval of leaf nodes on the same level differs by $1/2^{depth-1}$; the number of leaf nodes on the deepest level is $(2k-1)-2^{d-1}$; the nodes connected to the deepest leaf nodes are $((2k-1)-2^{d-1})/2$. Based on the above conclusions, when $i \leq (2k-1)-2^{d-1}$, the interval has an endpoint value of $i/2^{d-1}$. When $i > (2k-1)-2^{d-1}$, the right endpoint of the interval is $(i-((2k-1)-2^{d-1})/2)/2^{d-2}$. Based on these viewpoints, we can get $(i+\max(i-2k+2^{\lfloor \log_2(2k-1) \rfloor}, 0))/2^{\lfloor \log_2(2k-1) \rfloor}$.

B Derivation of Update Interval in Algorithm

Since a function is needed that has a large value early in training and a small value later in training to calculate the update interval, the elliptic function satisfies these requirements:

$$g(x) = 1 + c_\Delta \sqrt{1-x^2}, \tag{18}$$

where $g(1) = 1$ indicates that CRIK gradually switches in the later stages of training; $\mathbb{E}[k/g(x)] = 1$ indicates that the average number of perturbations per step is approximately 1, which means CRIK has computational overhead similar to SAM. When there are many training epochs, we have

$$\mathbb{E}[\frac{k}{g(x)}] = \int_0^1 \frac{k}{g(x)} \mathrm{d}x = 1. \tag{19}$$

We have

$$\begin{aligned} &\int_0^1 \frac{k}{g(x)} \mathrm{d}x \\ =&\int_0^1 \frac{k}{1+c_\Delta \sqrt{1-x^2}} \mathrm{d}x \\ =&k \cdot \frac{\frac{\pi \sqrt{c_\Delta - 1}\sqrt{c_\Delta + 1}}{2} + \ln\left(\frac{\left|\sqrt{c_\Delta - 1} - \sqrt{c_\Delta + 1}\right|}{\left|\sqrt{c_\Delta - 1} + \sqrt{c_\Delta + 1}\right|}\right)}{c_\Delta \sqrt{c_\Delta - 1}\sqrt{c_\Delta + 1}} \\ =&1. \end{aligned} \tag{20}$$

This equation has no closed-form solution. Therefore, we use the bisection method to approximate the solution to the equation. As shown in Table 4, we can obtain a table of b values for specific values of k.

Table 4: The solution of equation.

	c_Δ value	Absolute error
$k=3$	2.76061899294622	$7.72e^{-11}$
$k=4$	4.21460501267140	$6.15e^{-11}$
$k=5$	5.69027561416121	$4.25e^{-11}$
$k=6$	7.18053623322681	$5.36e^{-12}$
$k=7$	8.68133953386824	$9.60e^{-11}$
$k=8$	10.19014402408835	$2.36e^{-11}$
$k=9$	11.70524292464979	$2.84e^{-11}$
$k=10$	13.22543156907276	$9.40e^{-11}$

C Detailed Proof

C.1 Proof of Lemma 3.1

Proof. Let $\boldsymbol{\delta}_i = \boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1}$. Throughout this proof, the subscript $j \in \{1, \ldots, n\}$ indexes the coordinate dimension, and the subscripts $k, l \in \{1, \ldots, n\}$ enumerate the row and column entries

of the Hessian matrix, respectively. Any point in C_i can be written as $\boldsymbol{\phi} = \boldsymbol{\theta} + \boldsymbol{\epsilon}_{i-1} + \boldsymbol{\delta}$ with $\delta_j \in [0, \delta_{i,j}]$ for all j. Shifting to the center $\boldsymbol{\theta}_i^{\mathrm{c}}$, set $\boldsymbol{\delta}' = \boldsymbol{\delta} - \frac{\boldsymbol{\delta}_i}{2}$, so that $\delta'_j \in [-\frac{\delta_{i,j}}{2}, \frac{\delta_{i,j}}{2}]$ and $\boldsymbol{\phi} = \boldsymbol{\theta}_i^{\mathrm{c}} + \boldsymbol{\delta}'$. For any entry H_{kl}, Taylor expansion at $\boldsymbol{\theta}_i^{\mathrm{c}}$ gives

$$H_{kl}(\boldsymbol{\theta}_i^{\mathrm{c}} + \boldsymbol{\delta}') = H_{kl}(\boldsymbol{\theta}_i^{\mathrm{c}}) + \sum_{j=1}^{n} \partial_j H_{kl}(\boldsymbol{\theta}_i^{\mathrm{c}})\, \delta'_j + \mathcal{O}(\|\boldsymbol{\delta}'\|_2^2). \tag{21}$$

Averaging over C_i and using the symmetry $\int_{-\delta_{i,j}/2}^{\delta_{i,j}/2} \delta'_j \, \mathrm{d}\delta'_j = 0$, we obtain

$$\bar{H}_{i,kl}(\boldsymbol{\theta}) = H_{kl}(\boldsymbol{\theta}_i^{\mathrm{c}}) + \mathcal{O}(\|\boldsymbol{\delta}_i\|_2^2), \tag{22}$$

which implies $\|H(\boldsymbol{\theta}_i^{\mathrm{c}}) - \bar{H}_i(\boldsymbol{\theta})\|_{\mathrm{F}} = \mathcal{O}(\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1}\|_2^2)$. For any other point $\boldsymbol{\theta}' = \boldsymbol{\theta} + \boldsymbol{\epsilon}'$ in C_i, write $\boldsymbol{\epsilon}' = \frac{\boldsymbol{\epsilon}_{i-1} + \boldsymbol{\epsilon}_i}{2} + \boldsymbol{\delta}_0$ with $\boldsymbol{\delta}_0 \neq \mathbf{0}$ and $\|\boldsymbol{\delta}_0\|_2 = \mathcal{O}(\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1}\|_2)$. Then

$$H(\boldsymbol{\theta}') - \bar{H}_i(\boldsymbol{\theta}) = \nabla H(\boldsymbol{\theta}_i^{\mathrm{c}})[\boldsymbol{\delta}_0] + \mathcal{O}(\|\boldsymbol{\delta}_i\|_2^2), \tag{23}$$

where $(\nabla H(\boldsymbol{\theta}_i^{\mathrm{c}})[\boldsymbol{v}])_{kl} = \sum_{j=1}^{n} \partial_j H_{kl}(\boldsymbol{\theta}_i^{\mathrm{c}}) v_j$. Since $\nabla^3 \mathcal{L}(\boldsymbol{\theta}_i^{\mathrm{c}}) \neq \mathbf{0}$, the linear term does not vanish in general, giving $\|H(\boldsymbol{\theta}') - \bar{H}_i(\boldsymbol{\theta})\|_{\mathrm{F}} = \mathcal{O}(\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1}\|_2)$. For the eigenvalues, Weyl's inequality gives

$$\begin{aligned} |\Lambda_{\max}(H(\boldsymbol{\theta}_i^{\mathrm{c}})) - \Lambda_{\max}(\bar{H}_i(\boldsymbol{\theta}))| &\leq \|H(\boldsymbol{\theta}_i^{\mathrm{c}}) - \bar{H}_i(\boldsymbol{\theta})\|_2 = \mathcal{O}(\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1}\|_2^2), \\ |\Lambda_{\min}(H(\boldsymbol{\theta}_i^{\mathrm{c}})) - \Lambda_{\min}(\bar{H}_i(\boldsymbol{\theta}))| &\leq \|H(\boldsymbol{\theta}_i^{\mathrm{c}}) - \bar{H}_i(\boldsymbol{\theta})\|_2 = \mathcal{O}(\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1}\|_2^2). \end{aligned} \tag{24}$$

For the other point $\boldsymbol{\theta}'$, let $\mathbf{u}_i$ be the unit eigenvector of $\bar{H}_i(\boldsymbol{\theta})$ corresponding to $\Lambda_{\max}(\bar{H}_i(\boldsymbol{\theta}))$. By the Rayleigh quotient,

$$\Lambda_{\max}(H(\boldsymbol{\theta}')) \geq \mathbf{u}_i^{\top} H(\boldsymbol{\theta}') \mathbf{u}_i = \Lambda_{\max}(\bar{H}_i(\boldsymbol{\theta})) + \mathbf{u}_i^{\top} \big(H(\boldsymbol{\theta}') - \bar{H}_i(\boldsymbol{\theta})\big) \mathbf{u}_i. \tag{25}$$

Because $H(\boldsymbol{\theta}') - \bar{H}_i(\boldsymbol{\theta}) = \mathcal{O}(\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1}\|_2)$ and the third derivative is non-zero, the perturbation term is $\mathcal{O}(\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1}\|_2)$ in general. The same argument applied to $-\mathcal{L}$ yields the lower bound for $\Lambda_{\min}$. Hence the eigenvalue deviations at $\boldsymbol{\theta}'$ are $\mathcal{O}(\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1}\|_2)$. The proof is completed. ■

C.2 Proof of Lemma 3.2

Proof. When $\mathcal{L}$ is γ'-Lipschitz smooth, we have

$$|\mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon_i}) - \mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon_{i-1}})| \leq \gamma' ||(q_i - q_{i-1})\boldsymbol{\epsilon}||_2 = \gamma'(q_i - q_{i-1})||\boldsymbol{\epsilon}||_2, \tag{26}$$

where $q_i = (i + \max(i - 2k + 2^{\lfloor \log_2(2k-1) \rfloor}, 0))/2^{\lfloor \log_2(2k-1) \rfloor}$. Then, we have:

$$\sum_{i=1}^{k} |\mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon_i}) - \mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon_{i-1}})| \leq \sum_{i=1}^{k} \gamma'(q_i - q_{i-1})||\boldsymbol{\epsilon}||_2 = \gamma'||\boldsymbol{\epsilon}||_2 < +\infty. \tag{27}$$

so $\mathcal{L}\big(\boldsymbol{\theta} + q\boldsymbol{\epsilon}\big)$ is bounded in $[0, 1]$. Because f is continuous in $(-\infty, +\infty)$ and $\mathcal{L}\big(\boldsymbol{\theta} + q\boldsymbol{\epsilon}\big)$is bounded variation in $[0, 1]$, then the Riemann-Stieltjes integral $\int_0^1 f\big(\mathcal{L}\big(\boldsymbol{\theta} + q\boldsymbol{\epsilon}\big)\big)\, \mathrm{d}_q \mathcal{L}\big(\boldsymbol{\theta} + q\boldsymbol{\epsilon}\big)$ exists; moreover, for any sequence of partitions q_k with $q_k \to 0$, the Riemann-Stieltjes sums with any choice of sample points converge to $\int_0^1 f\big(\mathcal{L}\big(\boldsymbol{\theta} + q\boldsymbol{\epsilon}\big)\big)\, \mathrm{d}_q \mathcal{L}\big(\boldsymbol{\theta} + q\boldsymbol{\epsilon}\big)$. The proof is thus completed. ■

C.3 Proof of Lemma 3.3

Proof. Using Taylor expansion, we have

$$\begin{aligned}
&\sum_{i=2}^{k} f(\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1}))(\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_i)-\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1})) \\
=&\sum_{i=2}^{k} f\left(\mathcal{L}\left(\boldsymbol{\theta}+\epsilon_{i-1}\right)\right)\left(\mathcal{L}\left(\boldsymbol{\theta}+\epsilon_i\right)+\mathcal{L}\left(\boldsymbol{\theta}+\epsilon_{i-2}\right)-2\mathcal{L}\left(\boldsymbol{\theta}+\epsilon_{i-1}\right)\right) \\
&+\sum_{i=1}^{k-1} f\left(\mathcal{L}\left(\boldsymbol{\theta}+\epsilon_i\right)\right)\left(\mathcal{L}\left(\boldsymbol{\theta}+\epsilon_i\right)-\mathcal{L}\left(\boldsymbol{\theta}+\epsilon_{i-1}\right)\right) \\
=&\sum_{i=2}^{k} f(\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1}))((\boldsymbol{\epsilon}_i-\boldsymbol{\epsilon}_{i-1})^\top\nabla^2\mathcal{L}(\boldsymbol{\xi}_i')(\boldsymbol{\epsilon}_i-\boldsymbol{\epsilon}_{i-1})) \\
&+\sum_{i=1}^{k-1} f(\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_i))\left(\nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\xi}_i'')^\top(\boldsymbol{\epsilon}_i-\boldsymbol{\epsilon}_{i-1})\right),
\end{aligned} \tag{28}$$

where $\boldsymbol{\xi}_i' \in (\boldsymbol{\epsilon}_{i-1}, \boldsymbol{\epsilon}_i)$ and $\boldsymbol{\xi}_i'' \in (\boldsymbol{\epsilon}_{i-1}, \boldsymbol{\epsilon}_i)$. The proof is thus completed. ■

C.4 Proof of Lemma 3.4

Proof. For $\|\boldsymbol{\epsilon}'\| \le \rho$, by the γ'-Lipschitz continuity of $\mathcal{L}$:

$$|\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}')-\mathcal{L}(\boldsymbol{\theta})| \le \gamma'\rho. \tag{29}$$

Meanwhile, by the definition of ω and the triangle inequality:

$$\begin{aligned}
|\omega(\boldsymbol{\theta})| &= \left|\sum_{i=1}^{k} f(\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1}))\Delta\mathcal{L}_i\right| \\
&\ge \left|\sum_{i=1}^{k}\Delta\mathcal{L}_i\right| - \sum_{i=1}^{k}|f(\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1}))-1|\cdot|\Delta\mathcal{L}_i|,
\end{aligned} \tag{30}$$

where $\Delta\mathcal{L}_i = \mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_i)-\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1})$. Since $f(\mathcal{L}_{i-1}) \le \gamma_1+\gamma_2$ and $\sum_{i=1}^{k}|\Delta\mathcal{L}_i| \le \gamma'\rho$, we have

$$|\omega(\boldsymbol{\theta})| \ge |\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon})-\mathcal{L}(\boldsymbol{\theta})| - (\gamma_1+\gamma_2-1)\cdot\gamma'\rho. \tag{31}$$

Therefore, we have

$$|\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon})-\mathcal{L}(\boldsymbol{\theta})| \le |\omega(\boldsymbol{\theta})| + (\gamma_1+\gamma_2-1)\gamma'\rho. \tag{32}$$

It can be rewritten as

$$\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}') \le \mathcal{L}(\boldsymbol{\theta}) + |\omega(\boldsymbol{\theta})| + (\gamma_1+\gamma_2-1)\gamma'\rho. \tag{33}$$

The proof is thus completed. ■

C.5 Proof of Lemma 3.5

Proof. We consider the empirical risk minimization problem $\mathcal{L}(\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^n$ denotes the parameter vector. When $\|\boldsymbol{g}(\boldsymbol{\theta})\| > 0$, we perform a Taylor expansion at $\boldsymbol{\theta}$ of $\mathcal{L}_i = \mathcal{L}(\boldsymbol{\theta}+q_i\boldsymbol{\epsilon})$, with $\boldsymbol{\epsilon} = \rho\boldsymbol{g}(\boldsymbol{\theta})/\|\boldsymbol{g}(\boldsymbol{\theta})\|$:

$$\mathcal{L}_i = \mathcal{L}(\boldsymbol{\theta}) + q_i\rho\|\boldsymbol{g}(\boldsymbol{\theta})\| + \frac{q_i^2\rho^2}{2}\frac{\boldsymbol{g}(\boldsymbol{\theta})^\top H\boldsymbol{g}(\boldsymbol{\theta})}{\|\boldsymbol{g}(\boldsymbol{\theta})\|^2} + \mathcal{O}(\rho^3). \tag{34}$$

Since $\boldsymbol{g}(\boldsymbol{\theta})^\top\boldsymbol{\epsilon} = \rho\|\boldsymbol{g}(\boldsymbol{\theta})\|$, we obtain

$$\begin{aligned}
f(\mathcal{L}_{i-1}) =& f(\mathcal{L}(\boldsymbol{\theta})) + f'(\mathcal{L}(\boldsymbol{\theta}))\left(q_{i-1}\rho\|\boldsymbol{g}(\boldsymbol{\theta})\| + \frac{q_{i-1}^2\rho^2}{2}\frac{\boldsymbol{g}(\boldsymbol{\theta})^\top H\boldsymbol{g}(\boldsymbol{\theta})}{\|\boldsymbol{g}(\boldsymbol{\theta})\|^2}\right) \\
&+ \frac{1}{2}f''(\mathcal{L}(\boldsymbol{\theta}))q_{i-1}^2\rho^2\|\boldsymbol{g}(\boldsymbol{\theta})\|^2 + \mathcal{O}(\rho^3).
\end{aligned} \tag{35}$$

Therefore, we have

$$\begin{aligned}\mathcal{L}_i - \mathcal{L}_{i-1} =&(q_i - q_{i-1})\rho\|\boldsymbol{g}(\boldsymbol{\theta})\| + \frac{q_i^2 - q_{i-1}^2}{2}\rho^2\frac{\boldsymbol{g}(\boldsymbol{\theta})^\top H\boldsymbol{g}(\boldsymbol{\theta})}{\|\boldsymbol{g}(\boldsymbol{\theta})\|^2} \\ &+ \mathcal{O}(\rho^3).\end{aligned} \tag{36}$$

According to the definition of q_i, we have

$$\sum_{i=1}^{k}(q_i - q_{i-1}) = 1, \quad \sum_{i=1}^{k}(q_i^2 - q_{i-1}^2) = 1, \sum_{i=1}^{k} q_{i-1}(q_i - q_{i-1}) = \frac{1}{2} + \mathcal{O}(\frac{1}{k}) \approx \frac{1}{2}. \tag{37}$$

Substituting Eq. (37) into the definition of $\omega(\boldsymbol{\theta})$, we obtain:

$$\begin{aligned}\omega(\boldsymbol{\theta}) =&f\big(\mathcal{L}(\boldsymbol{\theta})\big)\rho\|\boldsymbol{g}(\boldsymbol{\theta})\| + \frac{\rho^2}{2}\left[f\big(\mathcal{L}(\boldsymbol{\theta})\big)\frac{\boldsymbol{g}(\boldsymbol{\theta})^\top H\boldsymbol{g}(\boldsymbol{\theta})}{\|\boldsymbol{g}(\boldsymbol{\theta})\|^2} + f'\big(\mathcal{L}(\boldsymbol{\theta})\big)\|\boldsymbol{g}(\boldsymbol{\theta})\|^2\right] \\ &+ \mathcal{O}(\rho^3) + \mathcal{O}(\frac{\rho^2}{k}).\end{aligned} \tag{38}$$

By taking the gradient with respect to $\boldsymbol{\theta}$, we have

$$\begin{aligned}\nabla\omega =&\rho\nabla\left[f\big(\mathcal{L}(\boldsymbol{\theta})\big)\|\boldsymbol{g}(\boldsymbol{\theta})\|\right] + \frac{\rho^2}{2}\nabla\left[f\big(\mathcal{L}(\boldsymbol{\theta})\big)\frac{\boldsymbol{g}(\boldsymbol{\theta})^\top H\boldsymbol{g}(\boldsymbol{\theta})}{\|\boldsymbol{g}(\boldsymbol{\theta})\|^2} + f'\big(\mathcal{L}(\boldsymbol{\theta})\big)\|\boldsymbol{g}(\boldsymbol{\theta})\|^2\right] \\ &+ \mathcal{O}(\rho^3) + \mathcal{O}(\frac{\rho^2}{k}).\end{aligned} \tag{39}$$

Computing the gradient term, we obtain

$$\nabla\left[f\big(\mathcal{L}(\boldsymbol{\theta})\big)\|\boldsymbol{g}(\boldsymbol{\theta})\|\right] = f'\big(\mathcal{L}(\boldsymbol{\theta})\big)\|\boldsymbol{g}(\boldsymbol{\theta})\|\boldsymbol{g}(\boldsymbol{\theta}) + f\big(\mathcal{L}(\boldsymbol{\theta})\big)\frac{H\boldsymbol{g}(\boldsymbol{\theta})}{\|\boldsymbol{g}(\boldsymbol{\theta})\|}. \tag{40}$$

Therefore, we have

$$\nabla\omega =\rho f'\big(\mathcal{L}(\boldsymbol{\theta})\big)\|\boldsymbol{g}(\boldsymbol{\theta})\|\boldsymbol{g}(\boldsymbol{\theta}) + \rho C_f(\boldsymbol{\theta})\frac{H\boldsymbol{g}(\boldsymbol{\theta})}{\|\boldsymbol{g}(\boldsymbol{\theta})\|} + \frac{\rho^2}{2}\nabla\mathcal{M}(\boldsymbol{\theta}) + \mathcal{O}(\rho^3) + \mathcal{O}(\frac{\rho^2}{k}), \tag{41}$$

where $C_f(\boldsymbol{\theta}) = f\big(\mathcal{L}(\boldsymbol{\theta})\big) + \|\boldsymbol{g}(\boldsymbol{\theta})\|f'\big(\mathcal{L}(\boldsymbol{\theta})\big)$, and $\mathcal{M}(\boldsymbol{\theta}) = f\big(\mathcal{L}(\boldsymbol{\theta})\big)\boldsymbol{g}(\boldsymbol{\theta})^\top H\boldsymbol{g}(\boldsymbol{\theta})/\|\boldsymbol{g}(\boldsymbol{\theta})\|^2 + f'\big(\mathcal{L}(\boldsymbol{\theta})\big)\|\boldsymbol{g}(\boldsymbol{\theta})\|^2$. At switching steps, $\hat{\boldsymbol{g}}(\boldsymbol{\theta}) = \boldsymbol{g}(\boldsymbol{\theta}) + \lambda\nabla\omega(\boldsymbol{\theta})$. Therefore, we have

$$\begin{aligned}\hat{\boldsymbol{g}}(\boldsymbol{\theta}) =&\boldsymbol{g}(\boldsymbol{\theta}) + \lambda\rho f'\big(\mathcal{L}(\boldsymbol{\theta})\big)\|\boldsymbol{g}(\boldsymbol{\theta})\|\boldsymbol{g}(\boldsymbol{\theta}) + \lambda\rho C_f(\boldsymbol{\theta})\frac{H\boldsymbol{g}(\boldsymbol{\theta})}{\|\boldsymbol{g}(\boldsymbol{\theta})\|} \\ &+ \frac{\lambda\rho^2}{2}\nabla\mathcal{M}(\boldsymbol{\theta}) + \mathcal{O}(\lambda\rho^3) + \mathcal{O}(\frac{\lambda\rho^2}{k}).\end{aligned} \tag{42}$$

It gives the drift term:

$$\begin{aligned}b(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{t_0}, 0) =& - \boldsymbol{g}(\boldsymbol{\Theta}) - \lambda\rho f'(\mathcal{L}(\boldsymbol{\Theta}))\|\boldsymbol{g}(\boldsymbol{\Theta})\|\boldsymbol{g}(\boldsymbol{\Theta}) - \lambda\rho C_f(\boldsymbol{\Theta})\frac{H\boldsymbol{g}(\boldsymbol{\Theta})}{\|\boldsymbol{g}(\boldsymbol{\Theta})\|} - \frac{\lambda\rho^2}{2}\nabla\mathcal{M}(\boldsymbol{\Theta}) \\ &+ \mathcal{O}(\lambda\rho^3) + \mathcal{O}(\frac{\lambda\rho^2}{k}).\end{aligned} \tag{43}$$

We consider the mini-batch setting. Let $\mathcal{B}$ be a mini-batch of size β. The empirical gradient is $\boldsymbol{g}_\mathcal{B}(\boldsymbol{\theta}) = \boldsymbol{g}(\boldsymbol{\theta}) + \boldsymbol{\zeta}(\boldsymbol{\theta})$, where $\boldsymbol{\zeta}$ is zero-mean noise with covariance $\Sigma_{\text{SGD}}(\boldsymbol{\theta})$. At the switching steps, $\nabla\omega_\mathcal{B}(\boldsymbol{\theta})$ depends on $\boldsymbol{g}_\mathcal{B}(\boldsymbol{\theta})$ and $H_\mathcal{B}$. Through linearizing $\nabla\omega_\mathcal{B}$ around $\boldsymbol{g}(\boldsymbol{\theta})$ and $H(\boldsymbol{\theta})$, we have

$$\nabla\omega_\mathcal{B}(\boldsymbol{\theta}) = \nabla\omega(\boldsymbol{\theta}) + \mathcal{O}(\|\boldsymbol{g}_\mathcal{B} - \boldsymbol{g}\| + \|H_\mathcal{B} - H\|). \tag{44}$$

Since $\|\boldsymbol{g}_\mathcal{B} - \boldsymbol{g}\| = \mathcal{O}(1/\sqrt{\beta})$ and $\|H_\mathcal{B} - H\| = \mathcal{O}(1/\sqrt{\beta})$, the covariance of the switching-step gradient $\hat{\boldsymbol{g}} = \boldsymbol{g}_\mathcal{B} + \lambda\nabla\omega_\mathcal{B}$ is

$$\Sigma_{\text{switch}}(\boldsymbol{\theta}) = \Sigma_{\text{SGD}}(\boldsymbol{\theta}) + \mathcal{O}(\frac{\lambda\rho}{\sqrt{\beta}}) + \mathcal{O}(\lambda^2\rho^2) + \mathcal{O}(\frac{\lambda\rho^2}{k}), \tag{45}$$

where $\mathcal{O}(\lambda\rho/\sqrt{\beta})$ represents the cross-covariance term $2\lambda\,\mathrm{Cov}(\boldsymbol{g}_{\mathcal{B}}, \nabla\omega_{\mathcal{B}})$. At non-switching steps, the cached gradient $\nabla\omega(\boldsymbol{\theta}_{t_0})$ is computed using the mini-batch $\mathcal{B}_0$ from the switching step. Since $\mathcal{B}$ and $\mathcal{B}_0$ are independent, the conditional covariance of $\hat{\boldsymbol{g}}(\boldsymbol{\theta}) = \boldsymbol{g}_{\mathcal{B}}(\boldsymbol{\theta}) + \lambda\nabla\omega_{\mathcal{B}_0}(\boldsymbol{\theta}_{t_0})$ gives

$$\Sigma_{\text{non-switch}}(\boldsymbol{\theta}_t) = \Sigma_{\text{SGD}}(\boldsymbol{\theta}_t) + \lambda^2\Sigma_{\text{switch}}(\boldsymbol{\theta}_{t_0}), \tag{46}$$

where $\Sigma_{\text{switch}}(\boldsymbol{\theta}_{t_0}) = \mathcal{O}(\rho)$ is the covariance of $\nabla\omega$ at the switching steps. When $\rho = \mathcal{O}(\sqrt{\eta})$, this term is $\mathcal{O}(\sqrt{\eta}) \to 0$ as $\eta \to 0$. So we construct the piecewise-defined delayed SDE:

$$\mathrm{d}\boldsymbol{\Theta}_t = b(\boldsymbol{\Theta}_t, \boldsymbol{\Theta}_{t_0}, r(t))\mathrm{d}t + \sqrt{\eta}\sigma(\boldsymbol{\Theta}_t, r(t))\mathrm{d}W_t, \tag{47}$$

where $r(t) = t - \lfloor t/\Delta \rfloor\Delta$, and for $t \in [j\Delta, (j+1)\Delta)$ with $j = 0, 1, 2, \ldots$. We define $\boldsymbol{\Theta}_{t_0} = \boldsymbol{\Theta}_{j\Delta}$. This SDE is well-posed. On each open interval $(j\Delta, (j+1)\Delta)$, the delay term is fixed, reducing it to a standard SDE with Lipschitz coefficients, whose solution exists uniquely. At the switching points $t = j\Delta$, the coefficients may be discontinuous, but since these points form a discrete set of measure zero, the solution can be extended continuously across them. When $r(t) = 0$, i.e., switching instants $t = j\Delta$, we have

$$\begin{aligned} b(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{t_0}, 0) =& -\boldsymbol{g}(\boldsymbol{\Theta}) - \lambda\rho f'(\mathcal{L}(\boldsymbol{\Theta}))\|\boldsymbol{g}(\boldsymbol{\Theta})\|\boldsymbol{g}(\boldsymbol{\Theta}) - \lambda\rho C_f(\boldsymbol{\Theta})\frac{H(\boldsymbol{\Theta})\boldsymbol{g}(\boldsymbol{\Theta})}{\|\boldsymbol{g}(\boldsymbol{\Theta})\|} - \frac{\lambda\rho^2}{2}\nabla\mathcal{M}(\boldsymbol{\Theta}) \\ &+ \mathcal{O}(\lambda\rho^3) + \mathcal{O}(\lambda\rho^2/k), \\ \sigma(\boldsymbol{\Theta}, 0)\sigma(\boldsymbol{\Theta}, 0)^\top =& \Sigma_{\text{SGD}}(\boldsymbol{\Theta}) + \mathcal{O}(\frac{\lambda\rho}{\sqrt{\beta}}) + \mathcal{O}(\lambda^2\rho^2) + \mathcal{O}(\frac{\lambda\rho^2}{k}), \end{aligned} \tag{48}$$

when $r(t) \neq 0$, i.e., non-switching instants, we have

$$\begin{aligned} b(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{t_0}, r) =& -\boldsymbol{g}(\boldsymbol{\Theta}) - \lambda\nabla\omega(\boldsymbol{\Theta}_{t_0}) + \mathcal{O}(\frac{\lambda\rho^2}{k}), \\ \sigma(\boldsymbol{\Theta}, r)\sigma(\boldsymbol{\Theta}, r)^\top =& \Sigma_{\text{SGD}}(\boldsymbol{\Theta}) + \lambda^2\Sigma_{\text{switch}}(\boldsymbol{\Theta}_{t_0}), \end{aligned} \tag{49}$$

where $\boldsymbol{\Theta}_{t_0} = \boldsymbol{\Theta}_{j\Delta}$ with $t \in [j\Delta, (j+1)\Delta)$. This completes the proof. ■

C.6 Proof of Lemma 3.6

Proof. We adopt the standard semigroup expansion method. Define the discrete-step Markov operator $\mathcal{P}_\eta$:

$$(\mathcal{P}_\eta\phi)(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{B}}\big[\phi\big(\boldsymbol{\theta} - \eta\hat{\boldsymbol{g}}_{\mathcal{B}}(\boldsymbol{\theta})\big)\big]. \tag{50}$$

Since the randomness of CRIK is time-homogeneous within each switching cycle, $\{\mathcal{P}_\eta^N\}_{N\geq 0}$ forms a discrete semigroup on each cycle. The solution of the piecewise-defined delayed SDE corresponds to an evolution system $\{\mathcal{P}_{s,t}\}_{0\leq s\leq t}$ (which reduces to a continuous semigroup on each interval $[j\Delta, (j+1)\Delta)$), whose generator is given by

$$\mathcal{L}_t\phi = \big\langle b(\cdot, \cdot, r(t)), \nabla\phi\big\rangle + \frac{\eta}{2}\Sigma(\cdot, r(t)) : \nabla^2\phi, \tag{51}$$

where $t \neq j\Delta$. At switching steps $t = j\Delta$, the generator changes discontinuously. At a specific switching step, i.e., $r(0) = 0$, we perform Taylor expansion on ϕ:

$$\phi(\boldsymbol{\theta}_1) = \phi(\boldsymbol{\theta}_0) + \big\langle\nabla\phi(\tilde{\boldsymbol{\theta}}_0), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\big\rangle + \frac{1}{2}\big\langle\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0, \nabla^2\phi(\tilde{\boldsymbol{\theta}}_0)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)\big\rangle, \tag{52}$$

where $\tilde{\boldsymbol{\theta}}_0 = \gamma\boldsymbol{\theta}_1 + (1-\gamma)\boldsymbol{\theta}_0$, $\gamma \in (0, 1)$. Taking the conditional expectation given $\boldsymbol{\theta}_0$, we have

$$\mathbb{E}\big[\phi(\boldsymbol{\theta}_1) \,\big|\, \boldsymbol{\theta}_0\big] = \phi(\boldsymbol{\theta}_0) - \eta\big\langle\nabla\phi(\boldsymbol{\theta}_0), \boldsymbol{g}(\boldsymbol{\theta}_0) + \lambda\nabla\omega(\boldsymbol{\theta}_0)\big\rangle + \frac{\eta}{2}\nabla^2\phi(\boldsymbol{\theta}_0) : \Sigma_{\text{switch}}(\boldsymbol{\theta}_0) + \mathcal{R}_1, \tag{53}$$

where $\mathcal{R}_1 = \mathcal{O}(\eta^{3/2}) + \mathcal{O}(\lambda\eta^2) + \mathcal{O}(\lambda\eta\rho^2/k)$ with $\|\mathcal{R}_1\| \leq C(\eta^{3/2} + \lambda\eta^2 + \lambda\eta\rho^2/k)\|\phi\|_{C^3}$. $\mathcal{O}(\lambda\eta^2)$ arises from the second-order remainder of $\nabla\omega$ because of $\nabla\omega = \mathcal{O}(\rho) = \mathcal{O}(\sqrt{\eta})$. At

non-switching steps, the cached gradient $\nabla\omega(\boldsymbol{\theta}_{t_0})$ is used. Since $\mathcal{B}$ (current mini-batch) and $\mathcal{B}_0$ (switching mini-batch) are independent, the conditional expectation given the history $\mathcal{F}_t$ is:

$$\begin{aligned}\mathbb{E}_{\mathcal{B}}\big[\phi(\boldsymbol{\theta}_{t+1})\,\big|\,\mathcal{F}_t\big] &= \phi(\boldsymbol{\theta}_t) - \eta\big\langle\nabla\phi(\boldsymbol{\theta}_t), \boldsymbol{g}(\boldsymbol{\theta}_t) + \lambda\nabla\omega(\boldsymbol{\theta}_{t_0})\big\rangle \\ &\quad + \frac{\eta}{2}\nabla^2\phi(\boldsymbol{\theta}_t) : \big(\Sigma_{\text{SGD}}(\boldsymbol{\theta}_t) + \lambda^2\Sigma_{\text{switch}}(\boldsymbol{\theta}_{t_0})\big) + \mathcal{R}_2,\end{aligned} \tag{54}$$

where $\mathcal{R}_2 = \mathcal{O}(\eta^{3/2}) + \mathcal{O}(\lambda\eta^2) + \mathcal{O}(\lambda\eta\rho^2/k)$. Let $\mathcal{P}^{\text{SDE}}_{t,t+\eta}$ denote the one-step evolution operator of the SDE from t to $t+\eta$:

$$\mathcal{P}^{\text{SDE}}_{t,t+\eta}\phi(\boldsymbol{\Theta}_t) = \mathbb{E}\big[\phi(\boldsymbol{\Theta}_{t+\eta})\,\big|\,\boldsymbol{\Theta}_t\big]. \tag{55}$$

Applying the Itô-Taylor expansion to SDE at t, we obtain

$$\begin{aligned}\mathcal{P}^{\text{SDE}}_{t,t+\eta}\phi(\boldsymbol{\Theta}_t) =& \phi(\boldsymbol{\Theta}_t) + \eta\mathcal{L}_t\phi(\boldsymbol{\Theta}_t) + \mathcal{O}(\eta^{\frac{3}{2}})\|\phi\|_{C^4} + \mathcal{O}(\lambda\eta^{\frac{3}{2}})\|\phi\|_{C^4} + \mathcal{O}(\lambda\eta^2)\|\phi\|_{C^4} \\ &+ \mathcal{O}(\frac{\lambda\eta\rho^2}{k})\|\phi\|_{C^4},\end{aligned} \tag{56}$$

where $\mathcal{O}(\lambda\eta^{3/2})$ arises from the delay discrepancy $\|\boldsymbol{\Theta}_{t_0} - \boldsymbol{\Theta}_t\| = \mathcal{O}(\sqrt{\eta})$ and Lipschitz continuity of $\nabla\omega$, which holds when the process is near a stationary region where $\|\boldsymbol{g}(\boldsymbol{\Theta})\| = \mathcal{O}(\sqrt{\eta})$, or when $\Delta = \mathcal{O}(\sqrt{\eta})$; the $\mathcal{O}(\lambda\eta^2)$ arises from the second-order remainder of the Taylor expansion of $\nabla\omega$. Since both operators share the same drift and diffusion coefficients up to the specified accuracy, for each t we have

$$\|(\mathcal{P}_\eta - \mathcal{P}^{\text{SDE}}_{t,t+\eta})\phi\|_\infty \le C\eta^{\frac{3}{2}}\|\phi\|_{C^4} + C\lambda\eta^2\|\phi\|_{C^4} + C\lambda\eta\rho^2/k\|\phi\|_{C^4}, \tag{57}$$

where $\mathcal{O}(\lambda\eta^{3/2})$ is absorbed into the leading $\mathcal{O}(\eta^{3/2})$ because of $\lambda = \mathcal{O}(1)$. We establish the global error using the semigroup decomposition. Since the coefficient is time-homogeneous on each $[j\Delta, (j+1)\Delta)$, we apply the argument within each cycle. For a complete cycle starting at $s = j\Delta$, let $M = \Delta/\eta$ be the number of steps in one cycle. Then we have

$$\mathcal{P}^M_\eta\phi(\boldsymbol{\Theta}_{j\Delta}) - \mathcal{P}_{j\Delta,(j+1)\Delta}\phi(\boldsymbol{\Theta}_{j\Delta}) = \sum_{l=1}^{M}\mathcal{P}^{M-l}_\eta \circ \Big(\mathcal{P}_\eta - \mathcal{P}^{\text{SDE}}_{j\Delta+(l-1)\eta,j\Delta+l\eta}\Big) \circ \mathcal{P}_{j\Delta,j\Delta+(l-1)\eta}\phi(\boldsymbol{\Theta}_{j\Delta}), \tag{58}$$

where Eq. (58) follows from the semigroup property $\mathcal{P}_{j\Delta,j\Delta+l\eta} = \mathcal{P}_{j\Delta+(l-1)\eta,j\Delta+l\eta} \circ \mathcal{P}_{j\Delta,j\Delta+(l-1)\eta}$ on the interval. Taking expectation conditioned on $\boldsymbol{\Theta}_{j\Delta+(l-1)\eta}$ and using the Markov property together with the contraction property of Markov operators, i.e., $\|\mathcal{P}\|_1 \le 1$ in the L^1 sense, we obtain

$$\begin{aligned}&\big|\mathcal{P}^M_\eta\phi(\boldsymbol{\Theta}_{j\Delta}) - \mathcal{P}_{j\Delta,(j+1)\Delta}\phi(\boldsymbol{\Theta}_{j\Delta})\big| \\ &\le \sum_{l=1}^{M}\mathbb{E}\left[\left|\Big(\mathcal{P}_\eta - \mathcal{P}^{\text{SDE}}_{j\Delta+(l-1)\eta,j\Delta+l\eta}\Big) \circ \mathcal{P}_{j\Delta,j\Delta+(l-1)\eta}\phi(\boldsymbol{\Theta}_{j\Delta})\right|\right] \\ &\le \sum_{l=1}^{M}\left(C\eta^{\frac{3}{2}}\|\phi\|_{C^4} + C\lambda\eta^2\|\phi\|_{C^4} + \frac{C\lambda\eta\rho^2}{k}\|\phi\|_{C^4}\right) \\ &= \Delta\cdot C\eta^{\frac{1}{2}}\|\phi\|_{C^4} + \Delta\cdot C\lambda\eta\|\phi\|_{C^4} + \Delta\cdot C\lambda\frac{\rho^2}{k}\|\phi\|_{C^4},\end{aligned} \tag{59}$$

where $M = \Delta/\eta$. At switching steps $t = j\Delta$, the generator changes discontinuously. However, since $j\Delta$ forms a set of measure zero in continuous time, and the number of switching steps in $[0, T]$ is $J = \lfloor T/\Delta\rfloor$, the contribution from each switching step is bounded by the single-step error $\mathcal{O}(\eta^{3/2})$. Summing over all J cycles, with $T = N\eta$ and $N = \lfloor T/\eta\rfloor$, the total error satisfies

$$\begin{aligned}&\big|\mathbb{E}[\phi(\boldsymbol{\theta}_N)] - \mathbb{E}[\phi(\boldsymbol{\Theta}_{N\eta})]\big| \\ &\le J\left(\Delta\cdot C\eta^{\frac{1}{2}}\|\phi\|_{C^4} + \Delta\cdot C\lambda\eta\|\phi\|_{C^4} + \Delta\cdot C\lambda\rho^2/k\|\phi\|_{C^4}\right) \quad + J\cdot\mathcal{O}(\eta^{\frac{3}{2}})\|\phi\|_{C^4} \\ &\le CT\eta^{\frac{1}{2}}\|\phi\|_{C^4} + CT\lambda\eta\|\phi\|_{C^4} + CT\lambda\rho^2/k\|\phi\|_{C^4} + \mathcal{O}(T\eta^{\frac{1}{2}}).\end{aligned} \tag{60}$$

When $\rho = \mathcal{O}(\sqrt{\eta})$, $\lambda = \mathcal{O}(1)$, and $k = \mathcal{O}(1/\eta)$, we have

$$\begin{aligned}&\big|\mathbb{E}[\phi(\boldsymbol{\theta}_N)] - \mathbb{E}[\phi(\boldsymbol{\Theta}_{N\eta})]\big| \\ &\le CT\eta^{\frac{1}{2}}\|\phi\|_{C^4} + CT\lambda\eta\|\phi\|_{C^4} + CT\lambda\eta\|\phi\|_{C^4} = \mathcal{O}(\sqrt{\eta}),\end{aligned} \tag{61}$$

where $CT\lambda\rho^2/k = CT\lambda\eta/(1/\eta) = CT\lambda\eta^2 = \mathcal{O}(\eta^2)$. Thus the bound is $\mathcal{O}(\sqrt{\eta})$. If $\rho = \mathcal{O}(\eta)$, $\mathcal{O}(T\lambda\rho^2/k)$ becomes $\mathcal{O}(T\lambda\eta^3)$, but $\mathcal{O}(T\eta^{1/2})$ persists because the delayed SDE structure causes the Euler-Maruyama discretization to be only $\mathcal{O}(\sqrt{\eta})$ accurate when Δ is fixed, due to the discontinuity of the coefficients at each switching step. Achieving $\mathcal{O}(\eta)$ accuracy would require higher-order discretization schemes or a different scaling regime for Δ. This completes the proof. ■

C.7 Proof of Lemma 3.9

Proof. When $f(x) = \gamma_1 \tanh^{\gamma_3}(x) + \gamma_2$, where $\gamma_1 > 1$, $\gamma_3 > 1$, and $x > 0$, we have $|f(x)| \le \gamma_1 + \gamma_2$; $|f'(x)| \le \gamma_1\gamma_3 \cdot \frac{2}{\gamma_3+1}\left(\frac{\gamma_3-1}{\gamma_3+1}\right)^{\frac{\gamma_3-1}{2}}$; $|f''(x)| \le \gamma_1\gamma_3(\gamma_3+1)$. For brevity, denote

$$\begin{aligned}L_0 &:= \gamma_1 + \gamma_2, \\ L_1 &:= \gamma_1\gamma_3\frac{2}{\gamma_3+1}\Big(\frac{\gamma_3-1}{\gamma_3+1}\Big)^{\frac{\gamma_3-1}{2}}, \\ L_2 &:= \gamma_1\gamma_3(\gamma_3+1).\end{aligned} \tag{62}$$

Since $|\mathcal{L}(\boldsymbol{\theta})| \le m$ and each perturbation satisfies $\|\boldsymbol{\epsilon}_i\| \le \rho$, we have:

$$|\omega(\boldsymbol{\theta})| \le \sum_{i=1}^{k} L_0 \cdot 2m = 2kL_0m =: m_\omega. \tag{63}$$

Using $\|\nabla\mathcal{L}\| \le \gamma'$ and $\|\nabla\mathcal{L}(\boldsymbol{\theta}_1) - \nabla\mathcal{L}(\boldsymbol{\theta}_2)\| \le \gamma\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$, and noting $\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_{i-1}\| \le 2\rho$, we obtain

$$\|\nabla\omega(\boldsymbol{\theta})\| \le \sum_{i=1}^{k}\Big(L_1\gamma' \cdot 2m + L_0 \cdot 2\gamma\rho\Big) = 2kmL_1\gamma' + 2kL_0\gamma\rho =: h. \tag{64}$$

For $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, by the definition of $\nabla\omega$ and the triangle inequality, we have

$$\begin{aligned}&\|\nabla\omega(\boldsymbol{\theta}_1) - \nabla\omega(\boldsymbol{\theta}_2)\| \\ &\le \sum_{i=1}^{k}\Big[\big\|f'(\mathcal{L}(\boldsymbol{\theta}_1+\boldsymbol{\epsilon}_{i-1}))\nabla\mathcal{L}(\boldsymbol{\theta}_1+\boldsymbol{\epsilon}_{i-1})(\mathcal{L}(\boldsymbol{\theta}_1+\boldsymbol{\epsilon}_i) - \mathcal{L}(\boldsymbol{\theta}_1+\boldsymbol{\epsilon}_{i-1})) \\ &\quad - f'(\mathcal{L}(\boldsymbol{\theta}_2+\boldsymbol{\epsilon}_{i-1}))\nabla\mathcal{L}(\boldsymbol{\theta}_2+\boldsymbol{\epsilon}_{i-1})(\mathcal{L}(\boldsymbol{\theta}_2+\boldsymbol{\epsilon}_i) - \mathcal{L}(\boldsymbol{\theta}_2+\boldsymbol{\epsilon}_{i-1}))\big\| \\ &\quad + \big\|f(\mathcal{L}(\boldsymbol{\theta}_1+\boldsymbol{\epsilon}_{i-1}))(\nabla\mathcal{L}(\boldsymbol{\theta}_1+\boldsymbol{\epsilon}_i) - \nabla\mathcal{L}(\boldsymbol{\theta}_1+\boldsymbol{\epsilon}_{i-1})) \\ &\qquad - f(\mathcal{L}(\boldsymbol{\theta}_2+\boldsymbol{\epsilon}_{i-1}))(\nabla\mathcal{L}(\boldsymbol{\theta}_2+\boldsymbol{\epsilon}_i) - \nabla\mathcal{L}(\boldsymbol{\theta}_2+\boldsymbol{\epsilon}_{i-1}))\big\|\Big].\end{aligned} \tag{65}$$

For the first part, write $\alpha_1(\boldsymbol{\theta}) = f'(\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1}))$, $\alpha_2(\boldsymbol{\theta}) = \nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1})$, $\alpha_3(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_i) - \mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1})$. Using $|\alpha_1| \le L_1$, $\|\alpha_2\| \le \gamma'$, $|\alpha_3| \le 2m$, $\|\alpha_2(\boldsymbol{\theta}_1) - \alpha_2(\boldsymbol{\theta}_2)\| \le \gamma\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$, $|\alpha_3(\boldsymbol{\theta}_1) - \alpha_3(\boldsymbol{\theta}_2)| \le 2\gamma'\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$, and $|\alpha_1(\boldsymbol{\theta}_1) - \alpha_1(\boldsymbol{\theta}_2)| \le L_2\gamma'\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$, we have

$$\begin{aligned}&\big\|\alpha_1(\boldsymbol{\theta}_1)\alpha_2(\boldsymbol{\theta}_1)\alpha_3(\boldsymbol{\theta}_1) - \alpha_1(\boldsymbol{\theta}_2)\alpha_2(\boldsymbol{\theta}_2)\alpha_3(\boldsymbol{\theta}_2)\big\| \\ &\le |\alpha_1(\boldsymbol{\theta}_1)|\|\alpha_2(\boldsymbol{\theta}_1)\||\alpha_3(\boldsymbol{\theta}_1) - \alpha_3(\boldsymbol{\theta}_2)| \\ &\quad + |\alpha_1(\boldsymbol{\theta}_1)|\|\alpha_2(\boldsymbol{\theta}_1) - \alpha_2(\boldsymbol{\theta}_2)\||\alpha_3(\boldsymbol{\theta}_2)| \\ &\quad + |\alpha_1(\boldsymbol{\theta}_1) - \alpha_1(\boldsymbol{\theta}_2)|\|\alpha_2(\boldsymbol{\theta}_2)\||\alpha_3(\boldsymbol{\theta}_2)| \\ &\le \big(2L_1\gamma'^2 + 2mL_1\gamma + 2mL_2\gamma'^2\big)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.\end{aligned} \tag{66}$$

For the second part, write $p_1(\boldsymbol{\theta}) = f(\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1}))$, $p_2(\boldsymbol{\theta}) = \nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_i) - \nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_{i-1})$. Using $|p_1| \le L_0$, $\|p_2(\boldsymbol{\theta}_1) - p_2(\boldsymbol{\theta}_2)\| \le 2\gamma\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$, $|p_1(\boldsymbol{\theta}_1) - p_1(\boldsymbol{\theta}_2)| \le L_1\gamma'\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$, and

$\|p_2(\boldsymbol{\theta}_2)\| \leq 2\gamma\rho$, we have

$$\begin{aligned}&\big\|p_1(\boldsymbol{\theta}_1)p_2(\boldsymbol{\theta}_1) - p_1(\boldsymbol{\theta}_2)p_2(\boldsymbol{\theta}_2)\big\| \\ \leq\ & |p_1(\boldsymbol{\theta}_1)|\|p_2(\boldsymbol{\theta}_1) - p_2(\boldsymbol{\theta}_2)\| + |p_1(\boldsymbol{\theta}_1) - p_1(\boldsymbol{\theta}_2)|\|p_2(\boldsymbol{\theta}_2)\| \\ \leq\ & \big(2L_0\gamma + 2\rho\gamma L_1\gamma'\big)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.\end{aligned} \tag{67}$$

Combining the above bounds yields

$$\|\nabla\omega(\boldsymbol{\theta}_1) - \nabla\omega(\boldsymbol{\theta}_2)\| \leq \mu\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|, \tag{68}$$

where $\mu = k\Big(2mL_2\gamma'^2 + 2mL_1\gamma + 2L_1\gamma'^2 + 2L_0\gamma + 2\rho\gamma L_1\gamma'\Big)$. The proof is thus completed. ■

C.8 Proof of Theorem 3.10

Proof. We first establish that $\hat{\mathcal{L}}$ inherits the Lipschitz smoothness from $\mathcal{L}$ and ω. For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, we have

$$\begin{aligned}&\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_1) - \nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_2)\| \\ =&\|\nabla\mathcal{L}(\boldsymbol{\theta}_1) - \nabla\mathcal{L}(\boldsymbol{\theta}_2) + \lambda(\nabla\omega(\boldsymbol{\theta}_1) - \nabla\omega(\boldsymbol{\theta}_2))\| \\ \leq&\|\nabla\mathcal{L}(\boldsymbol{\theta}_1) - \nabla\mathcal{L}(\boldsymbol{\theta}_2)\| + \lambda\|\nabla\omega(\boldsymbol{\theta}_1) - \nabla\omega(\boldsymbol{\theta}_2)\| \\ \leq&\gamma\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \lambda\mu\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \\ =&(\gamma + \lambda\mu)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.\end{aligned} \tag{69}$$

Thus, $\hat{\mathcal{L}}$ is $(\gamma + \lambda\mu)$-Lipschitz smooth, and we have the quadratic upper bound:

$$\hat{\mathcal{L}}(\boldsymbol{\theta}_{t+1}) \leq \hat{\mathcal{L}}(\boldsymbol{\theta}_t) + \nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_t)^\top(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) + \frac{\gamma + \lambda\mu}{2}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|^2. \tag{70}$$

Substituting the update rule $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t\hat{\boldsymbol{g}}_t$:

$$\hat{\mathcal{L}}(\boldsymbol{\theta}_{t+1}) \leq \hat{\mathcal{L}}(\boldsymbol{\theta}_t) - \eta_t\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_t)^\top\hat{\boldsymbol{g}}_t + \frac{(\gamma + \lambda\mu)\eta_t^2}{2}\|\hat{\boldsymbol{g}}_t\|^2. \tag{71}$$

Taking expectation conditioned on $\boldsymbol{\theta}_t$ and noting that $\mathbb{E}[\hat{\boldsymbol{g}}_t \mid \boldsymbol{\theta}_t] = \nabla\mathcal{L}(\boldsymbol{\theta}_t) + \lambda\nabla\omega(\boldsymbol{\theta}_t) = \nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_t)$:

$$\mathbb{E}[\hat{\mathcal{L}}(\boldsymbol{\theta}_{t+1}) \mid \boldsymbol{\theta}_t] \leq \hat{\mathcal{L}}(\boldsymbol{\theta}_t) - \eta_t\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_t)\|^2 + \frac{(\gamma + \lambda\mu)\eta_t^2}{2}\mathbb{E}[\|\hat{\boldsymbol{g}}_t\|^2 \mid \boldsymbol{\theta}_t]. \tag{72}$$

We have

$$\begin{aligned}\mathbb{E}[\|\hat{\boldsymbol{g}}_t\|^2 \mid \boldsymbol{\theta}_t] &= \mathbb{E}[\|\boldsymbol{g}_t + \lambda\nabla\omega(\boldsymbol{\theta}_t)\|^2 \mid \boldsymbol{\theta}_t] \\ &\leq 2\mathbb{E}[\|\boldsymbol{g}_t\|^2 \mid \boldsymbol{\theta}_t] + 2\lambda^2\|\nabla\omega(\boldsymbol{\theta}_t)\|^2 \\ &\leq 2(\|\boldsymbol{g}\|^2 + \lambda^2h^2).\end{aligned} \tag{73}$$

Substituting back and taking the full expectation:

$$\eta_t\mathbb{E}[\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_t)\|^2] \leq \mathbb{E}[\hat{\mathcal{L}}(\boldsymbol{\theta}_t)] - \mathbb{E}[\hat{\mathcal{L}}(\boldsymbol{\theta}_{t+1})] + (\gamma + \lambda\mu)\eta_t^2(\|\boldsymbol{g}\|^2 + \lambda^2h^2). \tag{74}$$

Summing over $t = 1$ to t', we have

$$\sum_{t=1}^{t'}\eta_t\mathbb{E}[\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_t)\|^2] \leq \mathbb{E}[\hat{\mathcal{L}}(\boldsymbol{\theta}_1)] - \mathbb{E}[\hat{\mathcal{L}}(\boldsymbol{\theta}_{t'+1})] + (\gamma + \lambda\mu)(\|\boldsymbol{g}\|^2 + \lambda^2h^2)\sum_{t=1}^{t'}\eta_t^2. \tag{75}$$

By using $|\hat{\mathcal{L}}(\boldsymbol{\theta})| = |\mathcal{L}(\boldsymbol{\theta}) + \lambda\omega(\boldsymbol{\theta})| \leq m + \lambda m_\omega$. Thus:

$$\mathbb{E}[\hat{\mathcal{L}}(\boldsymbol{\theta}_1)] - \mathbb{E}[\hat{\mathcal{L}}(\boldsymbol{\theta}_{t'+1})] \leq 2(m + \lambda m_\omega). \tag{76}$$

Using $\eta_t = \eta_0/\sqrt{t}$ and the bounds $\sum_{t=1}^{t'} 1/\sqrt{t} \geq \sqrt{t'}$ and $\sum_{t=1}^{t'} 1/t \leq 1 + \log t'$:

$$\begin{aligned}&\frac{\eta_0}{\sqrt{t'}}\sum_{t=1}^{t'}\mathbb{E}[\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_t)\|^2] \\ \leq&2(m + \lambda m_\omega) + (\gamma + \lambda\mu)\eta_0^2(\|\boldsymbol{g}\|^2 + \lambda^2h^2)(1 + \log t').\end{aligned} \tag{77}$$

Multiplying both sides by $\sqrt{t'}/\eta_0$ and dividing by t' yields

$$\frac{1}{t'}\sum_{t=1}^{t'}\mathbb{E}\left[\left\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_t)\right\|^2\right] \leq \frac{c_1 + c_2(1 + \log t')}{\sqrt{t'}}, \tag{78}$$

where $c_1 = 2(m + \lambda m_\omega)/\eta_0$ and $c_2 = (\gamma + \lambda\mu)\eta_0(\|\boldsymbol{g}\|^2 + \lambda^2h^2)$. This completes the proof. ■

C.9 Proof of Theorem 3.11

Proof. Following the proof in Zhuang et al. (2022), we obtain that with probability at least $1-\delta$:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}'\sim\mathcal{N}(0,\sigma'^2)}\left[\mathcal{L}'(\boldsymbol{\theta}+\boldsymbol{\epsilon}')\right] \leq \mathbb{E}_{\boldsymbol{\epsilon}'\sim\mathcal{N}(0,\sigma'^2)}\left[\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}')\right] \\ + \sqrt{\frac{\frac{1}{4}n\log\left(1+\frac{\|\boldsymbol{\theta}\|_2^2}{n\sigma'^2}\right)+\frac{1}{4}+\log\frac{s}{\delta}+2\log(6s+3n)}{s-1}}.\end{aligned} \tag{79}$$

Since $\boldsymbol{\epsilon}' \sim \mathcal{N}(0,\sigma'^2)$, $\|\boldsymbol{\epsilon}'\|^2/\sigma'^2$ has a chi-square distribution. Therefore, according to Foret et al. (2021), we have that for any $t>0$,

$$\mathbb{P}\left(\frac{\|\boldsymbol{\epsilon}'\|^2}{\sigma'^2}-n \geq 2\sqrt{nt}+2t\right) \leq \exp(-t). \tag{80}$$

By letting $t=\frac{1}{2}\log s$, we obtain with probability at least $1-1/\sqrt{s}$:

$$\begin{aligned}\|\boldsymbol{\epsilon}'\|^2 &\leq \sigma'^2\left(n+\sqrt{2n\log s}+\log s\right) \\ &\leq \sigma'^2\left(\sqrt{n}+\sqrt{\log s}\right)^2=\rho^2.\end{aligned} \tag{81}$$

Therefore, we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\epsilon}'\sim\mathcal{N}(0,\sigma'^2)}\left[\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}')\right] \\ =&\mathbb{E}_{\boldsymbol{\epsilon}'\sim\mathcal{N}(0,\sigma'^2)}\Big[\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}') \mid \|\boldsymbol{\epsilon}'\|\leq\rho\Big]\mathbb{P}\left(\|\boldsymbol{\epsilon}'\|\leq\rho\right) \\ &+\mathbb{E}_{\boldsymbol{\epsilon}'\sim\mathcal{N}(0,\sigma'^2)}\Big[\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}') \mid \|\boldsymbol{\epsilon}'\|>\rho\Big]\mathbb{P}\left(\|\boldsymbol{\epsilon}'\|>\rho\right) \\ \leq&\mathbb{E}_{\boldsymbol{\epsilon}'\sim\mathcal{N}(0,\sigma'^2)}\Big[\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}') \mid \|\boldsymbol{\epsilon}'\|\leq\rho\Big]+\frac{m}{\sqrt{s}}.\end{aligned} \tag{82}$$

The claim follows from the Lemma, Eqs. 16, and 82. This completes the proof. ■

C.10 Proof of Theorem 3.15

Proof. Based on Xie et al. (2022), the covariance of the noise is

$$C(\boldsymbol{\theta}) \approx \frac{H(\boldsymbol{\theta})}{\beta}. \tag{83}$$

The diffusion tensor is then

$$D(\boldsymbol{\theta}) = \frac{\eta}{2}C(\boldsymbol{\theta}) \approx \frac{\eta}{2\beta}H(\boldsymbol{\theta}). \tag{84}$$

Furthermore, the stochastic differential equation takes the form

$$\mathrm{d}\boldsymbol{\theta}_t = -\nabla\hat{\mathcal{L}}\left(\boldsymbol{\theta}_t\right)\mathrm{d}t + \sqrt{2D\left(\boldsymbol{\theta}_t\right)}\mathrm{d}W_t. \tag{85}$$

Under these Assumptions, a second-order Taylor approximation of the density function around the sharp minimum is

$$\begin{aligned}&P\left(\boldsymbol{\theta}\in V_{\boldsymbol{a}}\right) \\ =&\int_{\boldsymbol{\theta}\in V_{\boldsymbol{a}}} P(\boldsymbol{\theta})\mathrm{d}V \\ =&P(\boldsymbol{a})\int_{\boldsymbol{\theta}\in V_{\boldsymbol{a}}}\exp\left[-\frac{(\boldsymbol{\theta}-\boldsymbol{a})^\top}{2}\left(D_{\boldsymbol{a}}^{-\frac{1}{2}}H_{\boldsymbol{a}}D_{\boldsymbol{a}}^{-\frac{1}{2}}\right)(\boldsymbol{\theta}-\boldsymbol{a})\right]\mathrm{d}V \\ =&P(\boldsymbol{a})\frac{(2\pi)^{\frac{n}{2}}}{\det\left(D_{\boldsymbol{a}}^{-1}H_{\boldsymbol{a}}\right)^{\frac{1}{2}}},\end{aligned} \tag{86}$$

where n is the dimension of $\boldsymbol{\theta}$. Utilizing the formulation for probability current and flux, and based on these Assumptions, we determine the flux exiting through the boundary between $\boldsymbol{a}$ and $\boldsymbol{b}$:

$$\begin{aligned}
&\int_{S_{\boldsymbol{b}}} J \mathrm{d}S \\
&= J_{1d} \int_{S_{\boldsymbol{b}}} \exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{b})^\top \left[D_{\boldsymbol{b}}^{-\frac{1}{2}} H_{\boldsymbol{b}} D_{\boldsymbol{b}}^{-\frac{1}{2}}\right]^{\perp e} (\boldsymbol{\theta}-\boldsymbol{b})\right] \mathrm{d}S \\
&= J_{1d} \frac{(2\pi)^{\frac{n-1}{2}}}{\left(\prod_{i\neq e}\left(D_{\boldsymbol{b}_i}^{-1} H_{\boldsymbol{b}_i}\right)\right)^{\frac{1}{2}}},
\end{aligned} \tag{87}$$

where $[\cdot]^{\perp \boldsymbol{e}}$ denotes directions perpendicular to the escape path $\boldsymbol{e}$. Consequently:

$$\chi = \frac{\sqrt{\frac{\det\left(H_{\boldsymbol{a}} D_{\boldsymbol{a}}^{-1}\right)}{-\det\left(H_{\boldsymbol{b}} D_{\boldsymbol{b}}^{-1}\right)}}}{2\pi} |H_{\boldsymbol{be}}| \exp\left(-\frac{p\Delta\hat{\mathcal{L}}}{T_{\boldsymbol{a}}} - \frac{(1-p)\Delta\hat{\mathcal{L}}}{T_{\boldsymbol{b}}}\right), \tag{88}$$

where $T_{\boldsymbol{a}}$ and $T_{\boldsymbol{b}}$ denote the eigenvalues of $H_{\boldsymbol{a}}^{-1} D_{\boldsymbol{a}}$ and $H_{\boldsymbol{b}}^{-1} D_{\boldsymbol{b}}$, respectively, along the escape direction. It follows that $D_{\boldsymbol{a}} = \frac{\gamma\eta}{2\beta} H_{\boldsymbol{a}}$ and $D_{\boldsymbol{b}} = \frac{\gamma\eta}{2\beta} [H_{\boldsymbol{b}}]^+$. Since D must be positive semidefinite, $H_{\boldsymbol{b}} = U_{\boldsymbol{b}}^\top \operatorname{diag}\left(H_{\boldsymbol{b}_1}, \ldots, H_{\boldsymbol{b}_{n-1}}, H_{\boldsymbol{be}}\right) U_{\boldsymbol{b}}$ is substituted with its positive semidefinite equivalent $[H_{\boldsymbol{b}}]^+ = U_{\boldsymbol{b}}^\top \operatorname{diag}\left(H_{\boldsymbol{b}_1}, \ldots, H_{\boldsymbol{b}_{n-1}}, |H_{\boldsymbol{be}}|\right) U_{\boldsymbol{b}}$. Therefore, the mean escape time τ for CRIK to escape from such a sharp minimum is

$$\tau = \frac{1}{\chi} = \frac{2\pi}{|H_{\boldsymbol{be}}|} \exp\left[\frac{2\beta\Delta\hat{\mathcal{L}}}{\eta}\left(\frac{p}{H_{\boldsymbol{ae}}} + \frac{(1-p)}{|H_{\boldsymbol{be}}|}\right)\right]. \tag{89}$$

To calculate $\hat{\mathcal{L}}$, we do a Taylor expansion of $\mathcal{L}$ by using $\nabla\mathcal{L}(a) = 0$:

$$\mathcal{L}(\boldsymbol{a} + \boldsymbol{\epsilon}_i) = \mathcal{L}(\boldsymbol{a}) + \frac{1}{2} q_i^2 \rho^2 \Lambda_a + O(\rho^3). \tag{90}$$

where Λ_a is dominant eigenvalue at the sharp minimum $\boldsymbol{a}$. So we calculate loss differences:

$$\mathcal{L}(\boldsymbol{a} + \boldsymbol{\epsilon}_i) - \mathcal{L}(\boldsymbol{a} + \boldsymbol{\epsilon}_{i-1}) = \frac{\rho^2 \Lambda_a}{2}\left(q_i^2 - q_{i-1}^2\right) + O(\rho^3). \tag{91}$$

We rewrite $f(\mathcal{L}(\boldsymbol{a} + \boldsymbol{\epsilon}_i))$ as

$$f(\mathcal{L}(\boldsymbol{a} + \boldsymbol{\epsilon}_{i-1})) = f\big(\mathcal{L}(\boldsymbol{a}) + O(\rho^2)\big) = f(\mathcal{L}(\boldsymbol{a})) + O(\rho^2). \tag{92}$$

Then we have

$$\begin{aligned}
&f(\mathcal{L}(\boldsymbol{a} + \boldsymbol{\epsilon}_{i-1}))\big(\mathcal{L}(\boldsymbol{a} + \boldsymbol{\epsilon}_i) - \mathcal{L}(\boldsymbol{a} + \boldsymbol{\epsilon}_{i-1})\big) \\
=&f(\mathcal{L}(\boldsymbol{a})) \cdot \frac{\rho^2 \Lambda_a}{2}\left(q_i^2 - q_{i-1}^2\right) + O(\rho^3).
\end{aligned} \tag{93}$$

It can be summed up as

$$\omega(\boldsymbol{a}) = \frac{f(\mathcal{L}(\boldsymbol{a}))\rho^2 \Lambda_a}{2} + O(\rho^3). \tag{94}$$

Similarly, $\omega(\boldsymbol{b})$ can be calculated as

$$\omega(\boldsymbol{b}) = \frac{f(\mathcal{L}(\boldsymbol{b}))\rho^2 \Lambda_b}{2} + O(\rho^3). \tag{95}$$

So $\Delta\hat{\mathcal{L}}$ can be written as

$$\begin{aligned}
\Delta\hat{\mathcal{L}} &= \mathcal{L}(\boldsymbol{b}) + \lambda\omega(\boldsymbol{b}) - \mathcal{L}(\boldsymbol{a}) - \lambda\omega(\boldsymbol{a}) \\
&= \Delta\mathcal{L} - \frac{\rho^2\lambda}{2}(f(\mathcal{L}(\boldsymbol{a}))\Lambda_a - f(\mathcal{L}(\boldsymbol{b}))\Lambda_b).
\end{aligned} \tag{96}$$

Furthermore, the mean escape time τ_{SAM} for SAM to escape from such a sharp minimum is

$$\tau_{\text{SAM}} = \frac{2\pi}{|H_{\boldsymbol{be}}|} \exp\left(\frac{2\beta(\Delta\mathcal{L} - \frac{\rho^2}{2}(\Lambda_a - \Lambda_b))}{\eta}\left(\frac{p}{H_{\boldsymbol{ae}}} + \frac{1-p}{|H_{\boldsymbol{be}}|}\right)\right). \tag{97}$$

When $\lambda\gamma_1 > 1$ and $\Lambda_a > \frac{\lambda(\gamma_1+\gamma_2)-1}{\lambda\gamma_1-1}\Lambda_b$, f and λ reduce the barrier. Therefore, the mean escape time for CRIK is less than that for SAM, i.e., $\tau < \tau_{\text{SAM}}$. It suggests that CRIK has a stronger ability to escape from sharp minima than SAM. In a low-order Taylor expansion, k does not explicitly appear in the result. Nevertheless, a high-order Taylor expansion shows that under temperature control influenced by saddle points and sharp minima, as k increases, the effect of escaping sharp minima gradually strengthens. However, taking k to infinity is not realistic, so we choose a relatively large k. At the same time, compared to uniform partitioning, binary partitioning is more sensitive to k, which allows for a better balance between the computational overhead and the efficiency of escaping sharp minima. This completes the proof. ■

D Detailed Analysis of Theorem 3.15

D.1 SAM's Ability to Escape Sharp Minima

Similar to the derivation in the main text, we can get

$$\tau = \frac{1}{\chi} = \frac{2\pi}{|H_{\boldsymbol{be}}|}\exp\left[\frac{2\beta\Delta\hat{\mathcal{L}}}{\eta}\left(\frac{p}{H_{\boldsymbol{ae}}} + \frac{(1-p)}{|H_{\boldsymbol{be}}|}\right)\right]. \tag{98}$$

To calculate $\Delta\hat{\mathcal{L}}$, we do a Taylor expansion of $\mathcal{L}$ (note that $\nabla\mathcal{L}(a) = 0$):

$$\begin{aligned}\mathcal{L}(\boldsymbol{a}+\boldsymbol{\epsilon}) &= \mathcal{L}(\boldsymbol{a}) + \frac{1}{2}\rho^2\Lambda_a + \mathcal{O}(\rho^3),\\ \mathcal{L}(\boldsymbol{b}+\boldsymbol{\epsilon}) &= \mathcal{L}(\boldsymbol{b}) + \frac{1}{2}\rho^2\Lambda_b + \mathcal{O}(\rho^3),\end{aligned} \tag{99}$$

where Λ_a is dominant eigenvalue at the sharp minimum $\boldsymbol{a}$, Λ_b is dominant eigenvalue at $\boldsymbol{b}$. So $\Delta\hat{\mathcal{L}}$ can be written as

$$\Delta\hat{\mathcal{L}} = \Delta\mathcal{L} - \frac{\rho^2}{2}(\Lambda_a - \Lambda_b). \tag{100}$$

Furthermore, the mean escape time τ_{SAM} for SAM to escape from such a sharp minimum is

$$\tau_{\text{SAM}} = \frac{2\pi}{|H_{\boldsymbol{be}}|}\exp\left(\frac{2\beta(\Delta\mathcal{L} - \frac{\rho^2}{2}(\Lambda_a - \Lambda_b))}{\eta}\left(\frac{s}{H_{\boldsymbol{ae}}} + \frac{1-s}{|H_{\boldsymbol{be}}|}\right)\right). \tag{101}$$

D.2 CRIK's Ability to Escape Sharp Minima

Similar to the derivation in the main text, we can get

$$\tau = \frac{1}{\chi} = \frac{2\pi}{|H_{\boldsymbol{be}}|}\exp\left[\frac{2\beta\Delta\hat{\mathcal{L}}}{\eta}\left(\frac{p}{H_{\boldsymbol{ae}}} + \frac{(1-p)}{|H_{\boldsymbol{be}}|}\right)\right]. \tag{102}$$

To calculate $\hat{\mathcal{L}}$, we do a Taylor expansion of $\mathcal{L}$ when $\nabla\mathcal{L}(a) = 0$:

$$\mathcal{L}(\boldsymbol{a}+\boldsymbol{\epsilon}_i) = \mathcal{L}(\boldsymbol{a}) + \frac{1}{2}q_i^2\rho^2\Lambda_a + \mathcal{O}(\rho^3). \tag{103}$$

where Λ_a is dominant eigenvalue at the sharp minimum $\boldsymbol{a}$. So we calculate loss differences:

$$\mathcal{L}(\boldsymbol{a}+\boldsymbol{\epsilon}_i) - \mathcal{L}(\boldsymbol{a}+\boldsymbol{\epsilon}_{i-1}) = \frac{\rho^2\Lambda_a}{2}\left(q_i^2 - q_{i-1}^2\right) + \mathcal{O}(\rho^3). \tag{104}$$

We rewrite $f(\mathcal{L}(\boldsymbol{a}+\boldsymbol{\epsilon}_i))$ as

$$\begin{aligned}f(\mathcal{L}(\boldsymbol{a}+\boldsymbol{\epsilon}_i)) &= f\big(\mathcal{L}(a) + \mathcal{O}(\rho^2)\big)\\ &= f(\mathcal{L}(\boldsymbol{a})) + f'(\mathcal{L}(\boldsymbol{a}))\frac{\rho^2\Lambda_a}{2}q_{i-1}^2 + \mathcal{O}(\rho^3).\end{aligned} \tag{105}$$

Then we have

$$\begin{aligned}
&f(\mathcal{L}(\boldsymbol{a}+\boldsymbol{\epsilon}_{i-1}))(\mathcal{L}(\boldsymbol{a}+\boldsymbol{\epsilon}_i)-\mathcal{L}(\boldsymbol{a}+\boldsymbol{\epsilon}_{i-1})) \\
=&\left(f(\mathcal{L}(\boldsymbol{a}))+\frac{f'(\mathcal{L}(\boldsymbol{a}))q_{i-1}^2\rho^2\Lambda_a}{2}+\mathcal{O}(\rho^4)\right)\left(\frac{\rho^2\Lambda_a}{2}(q_i^2-q_{i-1}^2)+\mathcal{O}(\rho^3)\right) \\
=&\frac{f(\mathcal{L}(\boldsymbol{a}))\rho^2\Lambda_a}{2}(q_i^2-q_{i-1}^2)+\frac{f'(\mathcal{L}(\boldsymbol{a})\rho^3\Lambda_a^2 q_{i-1}^2(q_i^2-q_{i-1}^2)}{4}+\mathcal{O}(\rho^4).
\end{aligned} \tag{106}$$

It can be summed up as

$$\omega(\boldsymbol{a})=\frac{f(\mathcal{L}(\boldsymbol{a}))\rho^2\Lambda_a}{2}+\frac{f'(\mathcal{L}(\boldsymbol{a}))\rho^3\Lambda_a^2}{4}\sum_{i=1}^{k}q_{i-1}^2(q_i^2-q_{i-1}^2). \tag{107}$$

Similarly, $\omega(\boldsymbol{b})$ can be calculated as

$$\omega(\boldsymbol{b})=\frac{f(\mathcal{L}(\boldsymbol{b}))\rho^2\Lambda_b}{2}+\frac{f'(\mathcal{L}(\boldsymbol{b}))\rho^3\Lambda_b^2}{4}\sum_{i=1}^{k}q_{i-1}^2(q_i^2-q_{i-1}^2). \tag{108}$$

So $\Delta\hat{\mathcal{L}}$ can be written as

$$\begin{aligned}
\hat{\mathcal{L}}&=\mathcal{L}(\boldsymbol{b})+\lambda w(\boldsymbol{b})-\mathcal{L}(\boldsymbol{a})-\lambda w(\boldsymbol{a}) \\
&=\Delta\mathcal{L}-\frac{\rho^2\lambda}{2}(f(\mathcal{L}(\boldsymbol{a}))\Lambda_a-f(\mathcal{L}(\boldsymbol{b}))\Lambda_b) \\
&-(\frac{f'(\mathcal{L}(\boldsymbol{a}))\rho^3\Lambda_a^2-f'(\mathcal{L}(\boldsymbol{b}))\rho^3\Lambda_b^2}{4})\sum_{i=1}^{k}q_{i-1}^2(q_i^2-q_{i-1}^2).
\end{aligned} \tag{109}$$

$S(k)=\sum_{i=1}^{k}q_{i-1}^2(q_i^2-q_{i-1}^2)$ can be written as

$$\begin{aligned}
S(k)=\frac{1}{k'^4}\Bigg(&\frac{(2k-k')^2(2k-k'+1)^2}{2}-\frac{5(2k-k')(2k-k'+1)(4k-2k'+1)}{6} \\
&+2(2k-k')(2k-k'+1)-(2k-k')+4(k'-k)(2k-k')^2(2k-k'+1) \\
&+4(2k-k')(6k-3k'+2)(k'-k)(k'-k-1)+8(2k-k'+\frac{1}{3}) \\
&(k'-k)(k'-k-1)(2k'-2k-1)+8(k'-k)^2(k'-k-1)^2\Bigg),
\end{aligned} \tag{110}$$

where $k'=2^{\lfloor\log_2(2k-1)\rfloor}$.

As shown in Fig. 5, $S(k)$ rises quickly at the beginning and then stays almost constant, while $\Delta S(k)$ only shows small fluctuations early on and then approaches zero. Overall, it shows a steady and stable evolution. This suggests that the effect of generalization has an upper limit and converges quickly at the beginning, so a relatively small k is suitable.

D.3 DIFFERENT q_i

To show the superiority of the binary partition, we use a uniform partition $q_i=i/k$, $\sum_{i=1}^{k}q_{i-1}^2(q_i^2-q_{i-1}^2)=(3k^3-4k^2+1)/(6k^3)$. It is observable that within the low value regime of k (the first half of the interval for small k) in Fig. 6, $\Delta S_2(k)$ derived from the binary partition exhibits higher peak magnitudes and substantially faster growth rates relative to $\Delta S_1(k)$ under uniform partitioning. This demonstrates that the proposed binary partition facilitates a more rapid ascent of the objective function over the low k interval, yielding superior short-term optimization performance.

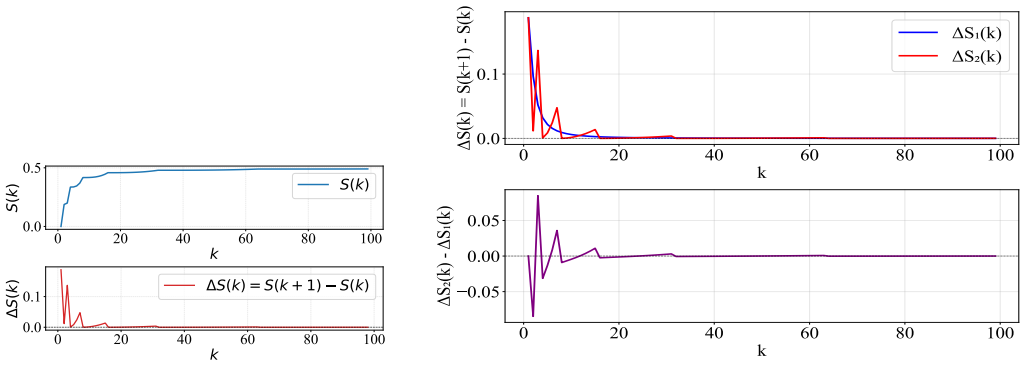


Figure 6: The effect of k on $\Delta S(k)$ and $S(k)$.

E Hyperparameter Settings and Additional Experiments

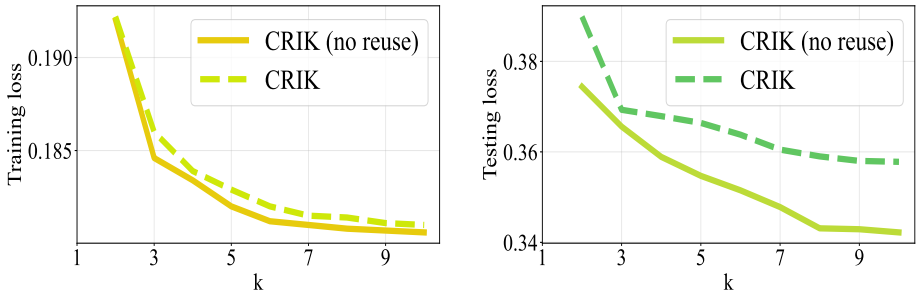
All experiments are conducted on Ubuntu 22.04 LTS using an NVIDIA GeForce RTX 4500 Ada GPU and an NVIDIA GeForce RTX 2080 Ti. The learning rates are set to 0.1 and 1e-3. For WikiText-103, the number of epochs is set to 20. The optimizer's detailed hyperparameter settings are the same as those used in the image classification experiment. We train for 200 epchos on the CIFAR-10 and CIFAR-100 datasets. The hyperparameter λ is set to 0.1; k is set to 5; γ_1, γ_2, and γ_3 are set to 1.5.

To illustrate the impact of the update interval and the total number of partitions k, we train ResNet18 on CIFAR-10. As shown in Fig. 7, the additional loss incurred by CRIK is acceptable, and the rate of decrease in loss slows as k increases.

To investigate the effects of the dynamic update interval in Algorithm 1, the convolution function f, and the key point calculation scheme, we conduct experiments with ResNet12 on CIFAR-100. Specifically, CRIK with $f = \arctan$ uses an arctan-based function for f; CRIK with $f =$ sigmoid uses a sigmoid-based function; CRIK with $q_i \sim \mathcal{N}(0, 1)$ replaces the coefficient generated by Eq. (5) with a truncated Gaussian distribution in $[0, 1]$; CRIK with $q_i \sim U(0, 1)$ replaces the coefficient generated by Eq. (5) with a uniform distribution in $[0, 1]$; and CRIK with $\Delta = \frac{1}{k}$ fixes Δ in Algorithm 1 to a constant. The results are shown in Table 5, indicating that CRIK achieves high accuracy with a minimal increase in computational overhead.

Table 5: Accuracy under different methods on CIFAR-100.

Method	Accuracy (%)	Increment (%)	Time (s)
SAM	$76.03_{\pm 0.11}$	-	4203
CRIK	$\mathbf{77.21}_{\pm 0.13}$	+1.18	4255
CRIK with $f = \arctan$	$76.47_{\pm 0.17}$	+0.44	4653
CRIK with $f =$ sigmoid	$76.72_{\pm 0.10}$	+0.69	4347
CRIK with $q_i \sim \mathcal{N}(0, 1)$	$76.44_{\pm 0.45}$	+0.41	4274
CRIK with $q_i \sim U(0, 1)$	$76.58_{\pm 0.39}$	+0.55	4303
CRIK with $\Delta = \frac{1}{k}$	$76.63_{\pm 0.15}$	+0.60	4579



(a) (b)

Figure 7: The impact of the number of partitions k on loss.

Table 6: Image classification accuracies (Mean ± Std) on CIFAR-100.

Model	Method	Test Accuracy (%)	Increment (%)
ResNet18	SAM	$78.31_{\pm 0.09}$	-
	GAF	$77.50_{\pm 0.27}$	-0.81
	AGA	$78.13_{\pm 0.33}$	-0.18
	TSAM	$78.70_{\pm 0.29}$	+0.39
	Adam	$72.86_{\pm 0.06}$	-
	RAD	$73.91_{\pm 0.17}$	+1.05
	CRIK	$78.50_{\pm 0.16}$	+0.19
	CRIK + TSAM	$\mathbf{79.22_{\pm 0.21}}$	+0.52
	CRIK + Adam	$72.95_{\pm 0.21}$	+0.09
	CRIK + RAD	$74.94_{\pm 0.19}$	+1.03
ResNet50	SAM	$78.45_{\pm 0.14}$	-
	MSAM	$77.13_{\pm 0.08}$	-1.32
	TSAM	$78.76_{\pm 0.23}$	+0.31
	LLR	$77.98_{\pm 0.39}$	-0.47
	RAD	$73.99_{\pm 0.19}$	-
	CRIK	$\mathbf{79.16}_{\pm 0.12}$	+0.71
	CRIK + MSAM	$77.57_{\pm 0.27}$	+0.44
	CRIK + TSAM	$78.82_{\pm 0.26}$	+0.06
	CRIK + RAD	$75.03_{\pm 0.14}$	+1.04
RepVGG-A1	SAM	$78.37_{\pm 0.12}$	-
	AGA	$78.34_{\pm 0.18}$	-0.03
	TSAM	$78.46_{\pm 0.16}$	+0.09
	Adam	$72.58_{\pm 0.11}$	-
	RAD	$73.74_{\pm 0.20}$	+1.16
	CRIK	$\mathbf{79.28}_{\pm 0.10}$	+0.91
	CRIK + TSAM	$79.22_{\pm 0.16}$	+0.76
	CRIK + Adam	$72.95_{\pm 0.27}$	+0.37
	CRIK + RAD	$74.44_{\pm 0.25}$	+0.70

Table 7: Test accuracy on ImageNet.

Model	Method	Test Accuracy(%)	Increment(%)
ResNet18	SAM	70.24	-
	CRIK	70.57	+0.33
ResNet50	SAM	76.42	-
	CRIK	76.74	+0.32