Dynamic Momentum Recalibration in Online Gradient Learning Appendix

A. Bias-Variance Decomposition (Section 2 in main paper)

Definition A.1. The unified momentum update rule is defined as:

$$m_t = \beta m_{t-1} + u g_t, \quad \theta_t = \theta_{t-1} - \alpha m_t, \tag{1}$$

where $\beta \in [0, 1)$ denotes the momentum (decay) coefficient, and $u \geq 1 - \beta$ is a scaling parameter controlling the contribution of the current gradient. The stochastic gradient is given by $g_t = \nabla f_t(\theta_t) + \epsilon_t, \epsilon_t \sim \mathcal{N}(0, \sigma^2 I)$, where $f_t(\theta) = f(\theta; \xi_t)$ denotes the stochastic objective at iteration t with ξ_t sampled from the data distribution $\mathcal{D}$. Specific cases include:

- $u = 1 - \beta$: Exponential Moving Average (EMA),
- $u = 1$: Classical Momentum (CM) [28, 33].

Assumption A.2. We make the following assumptions on the smoothness and stochasticity of the objective function f:

1. **Lipschitz continuity:** There exists a constant $L > 0$ such that, for any θ and ϕ, $\|\nabla f(\theta) - \nabla f(\phi)\| \leq L\|\theta - \phi\|$.
2. **Bounded gradients:** There exists a constant $G > 0$ such that, for all t, $\|\nabla f(\theta_t)\| \leq G$.
3. **Bounded gradient noise:** The stochastic gradient noise ϵ_t is temporally uncorrelated with zero mean, i.e., $\mathbb{E}[\epsilon_t] = \mathbf{0}$, for all t. Furthermore, the variance of the stochastic gradients is uniformly bounded, meaning there exists a constant $\sigma > 0$ such that $\mathbb{E}\big[\|g_t - \nabla f(\theta_t)\|^2\big] \leq \sigma^2$.

Lemma A.3 (Bias-Variance Decomposition)**.** *Let the stochastic gradient be given by $g_t = \nabla f(\theta_t) + \epsilon_t, \epsilon_t \sim \mathcal{N}(0, \sigma^2 I)$, where $\nabla f(\theta_t)$ denotes the true gradient and ϵ_t represents zero-mean Gaussian noise. For any gradient estimator $\hat{g}_t = \mathcal{A}(g_1, \ldots, g_t)$ produced by an arbitrary algorithm $\mathcal{A}$, the mean squared error (MSE) satisfies the bias-variance decomposition:*

$$\mathbb{E}\big[\|\hat{g}_t - \nabla f(\theta_t)\|^2\big] = \underbrace{\|\mathbb{E}[\hat{g}_t] - \nabla f(\theta_t)\|^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\big[\|\hat{g}_t - \mathbb{E}[\hat{g}_t]\|^2\big]}_{\text{Variance}}. \tag{2}$$

Proof.

$$\begin{aligned}
\mathbb{E}\big[\|\hat{g}_t - \nabla f(\theta_t)\|^2\big] &= \mathbb{E}\big[\|\hat{g}_t - \mathbb{E}[\hat{g}_t] + \mathbb{E}[\hat{g}_t] - \nabla f(\theta_t)\|^2\big] \\
&= \mathbb{E}\big[\|\hat{g}_t - \mathbb{E}[\hat{g}_t]\|^2\big] + \|\mathbb{E}[\hat{g}_t] - \nabla f(\theta_t)\|^2 + 2\,\mathbb{E}[\langle \hat{g}_t - \mathbb{E}[\hat{g}_t],\, \mathbb{E}[\hat{g}_t] - \nabla f(\theta_t)\rangle] \\
&= \mathrm{Var}(\hat{g}_t) + \mathrm{Bias}^2(\hat{g}_t) + 2\underbrace{\langle \mathbb{E}[\hat{g}_t - \mathbb{E}[\hat{g}_t]],\, \mathbb{E}[\hat{g}_t] - \nabla f(\theta_t)\rangle}_{=0} \\
&= \mathrm{Var}(\hat{g}_t) + \mathrm{Bias}^2(\hat{g}_t).
\end{aligned} \tag{3}$$

■

Lemma A.4. *Refer to the SDEs of vanilla SGD [32], the Definition A.1 with learning rate α can be represented in continuous time as the stochastic differential equation (SDE):*

$$\begin{cases} dm(t) = [-(1-\beta)m(t) + u\nabla f(\theta(t))]\,dt + u\sigma\,dW(t), \\ d\theta(t) = -\alpha\, m(t)\,dt, \end{cases} \tag{4}$$

where $m(t)$ is the momentum, $\theta(t)$ is the parameter, $\beta \in [0, 1)$ is the momentum coefficient, $u \geq 1 - \beta$ is the gradient scaling factor, $\alpha > 0$ is the learning rate, $\nabla f(\theta(t))$ is the gradient of the objective function, σ is the noise standard deviation, and $W(t)$ is a standard d-dimensional Wiener process. This approximation holds when the learning rate α is sufficiently small.

Proof. Start with the discrete momentum update rule from Definition A.1:

$$m_{n+1} = \beta m_n + u g(t_n), \quad \theta_{n+1} = \theta_n - \alpha m_{n+1}, \tag{5}$$

where $m_n = m(t_n)$ is the momentum, $\theta_n = \theta(t_n)$ is the parameter, $g(t_n) = \nabla f(\theta_n) + \epsilon_n$ with $\epsilon_n \sim \mathcal{N}(0, \sigma^2 I)$, and $\alpha > 0$ is the learning rate.

Rewrite the momentum update in terms of the increment:

$$\begin{aligned} m_{n+1} - m_n &= \beta m_n + u g(t_n) - m_n \\ &= -(1-\beta) m_n + u g(t_n) \\ &= -(1-\beta) m_n + u \nabla f(\theta_n) + u \epsilon_n. \end{aligned} \tag{6}$$

For the parameter:

$$\theta_{n+1} - \theta_n = -\alpha m_{n+1}. \tag{7}$$

To model this as a continuous-time SDE, assume the learning rate α is sufficiently small, controlling the step size of the discrete updates. Define $t_n = n$ to index discrete iterations, each corresponding to a unit time step $dt = 1$. The learning rate α controls the update magnitude but does not rescale time.

For the momentum, interpret the increment as the rate of change over one iteration:

$$m_{n+1} - m_n \approx [-(1-\beta) m_n + u \nabla f(\theta_n)]\, dt + u \epsilon_n, \quad dt = 1. \tag{8}$$

For small constant step sizes α, this yields the drift:

$$dm(t) = [-(1-\beta) m(t) + u \nabla f(\theta(t))]\, dt. \tag{9}$$

For the stochastic part, assume $\epsilon_n = \sigma Z_n,\ Z_n \sim \mathcal{N}(0, I)$, so that the stochastic increment $u\epsilon_n$ has variance $u^2\sigma^2$ per iteration, matching the Brownian term $u\sigma\, dW(t)$ under the step-time scaling $dt = 1$.

Therefore, in the SDE limit, the momentum dynamics can be written as:

$$dm(t) = [-(1-\beta) m(t) + u \nabla f(\theta(t))]\, dt + u\sigma\, dW(t). \tag{10}$$

For the parameter update:

$$\theta_{n+1} - \theta_n = -\alpha m_{n+1} \approx -\alpha m(t) \cdot (\text{time step}), \tag{11}$$

where the time step is implicitly dt in the continuous limit, yielding:

$$d\theta(t) = -\alpha m(t)\, dt. \tag{12}$$

Combining both, when α is small, the discrete updates approximate:

$$\begin{cases} dm(t) = [-(1-\beta) m(t) + u \nabla f(\theta(t))]\, dt + u\sigma\, dW(t), \\ d\theta(t) = -\alpha m(t)\, dt. \end{cases} \tag{13}$$

This SDE captures the dynamics of the momentum and parameter updates, with α as the learning rate driving the continuous approximation. ■

Remark A.5 (Step-time scaling). Our continuous-time formulation adopts the *step-time* scaling of Mandt et al. [32]. An alternative is the *slow-time* scaling $t = n\alpha$, often used in stochastic modified equations [19]. In that regime, one typically sets $1 - \beta = \Theta(\alpha)$, and the diffusion term scales with $\sqrt{\alpha}$. We do not adopt this scaling here, since doing so would modify both the drift and diffusion coefficients, as well as the form of $d\theta$.

Lemma A.6. *Under Lemma A.4, the solution to the stochastic differential equation (SDE), with initial conditions $m(0) = 0$ and $\theta(0) = \theta_0$, is given by:*

$$\begin{cases} m(t) = u \displaystyle\int_0^t e^{-(1-\beta)(t-s)} \nabla f(\theta(s))\, ds + u\sigma \int_0^t e^{-(1-\beta)(t-s)}\, dW(s), \\ \theta(t) = \theta_0 - \alpha \displaystyle\int_0^t m(s)\, ds, \end{cases} \tag{14}$$

where $W(t)$ is a standard Wiener process, and the integrals represent the stochastic evolution driven by the gradient $\nabla f(\theta(t))$ and noise.

Proof. We solve the coupled stochastic differential equation (SDE) system step-by-step:

$$\begin{cases} dm(t) = [-(1-\beta)m(t) + u\nabla f(\theta(t))]\,dt + u\sigma\,dW(t), \\ d\theta(t) = -\alpha\, m(t)\,dt, \end{cases} \tag{15}$$

with initial conditions $m(0) = 0$ and $\theta(0) = \theta_0$.

The $\theta(t)$ dynamics have drift only (no explicit diffusion term), but $\theta(t)$ is still random because $m(t)$ is random. Integrate:

$$\begin{aligned} d\theta(t) &= -\alpha\, m(t)\,dt, \\ \theta(t) - \theta(0) &= -\alpha \int_0^t m(s)\,ds. \end{aligned} \tag{16}$$

Since $\theta(0) = \theta_0$, we obtain:

$$\theta(t) = \theta_0 - \alpha \int_0^t m(s)\,ds. \tag{17}$$

This expresses $\theta(t)$ as a functional of $m(t)$, which we now determine.

Consider the linear SDE for $m(t)$ with a time-dependent forcing term:

$$dm(t) = [-(1-\beta)m(t) + u\nabla f(\theta(t))]\,dt + u\sigma\,dW(t). \tag{18}$$

Rewrite it in standard form:

$$dm(t) + (1-\beta)m(t)\,dt = u\nabla f(\theta(t))\,dt + u\sigma\,dW(t). \tag{19}$$

To solve this, apply the integrating factor $e^{\int_0^t (1-\beta)\,ds} = e^{(1-\beta)t}$. Multiply through by $e^{(1-\beta)t}$:

$$e^{(1-\beta)t}\,dm(t) + (1-\beta)e^{(1-\beta)t}m(t)\,dt = ue^{(1-\beta)t}\nabla f(\theta(t))\,dt + u\sigma e^{(1-\beta)t}\,dW(t). \tag{20}$$

Recognize the left-hand side as the differential of a product:

$$\begin{aligned} d[e^{(1-\beta)t}m(t)] &= e^{(1-\beta)t}\,dm(t) + (1-\beta)e^{(1-\beta)t}m(t)\,dt \\ &= ue^{(1-\beta)t}\nabla f(\theta(t))\,dt + u\sigma e^{(1-\beta)t}\,dW(t). \end{aligned} \tag{21}$$

Integrate both sides from 0 to t, with $m(0) = 0$, this simplifies to:

$$\begin{aligned} e^{(1-\beta)t}m(t) - e^{(1-\beta)\cdot 0}m(0) &= u\int_0^t e^{(1-\beta)s}\nabla f(\theta(s))\,ds + u\sigma\int_0^t e^{(1-\beta)s}\,dW(s), \\ e^{(1-\beta)t}m(t) &= u\int_0^t e^{(1-\beta)s}\nabla f(\theta(s))\,ds + u\sigma\int_0^t e^{(1-\beta)s}\,dW(s), \\ m(t) &= u\int_0^t e^{-(1-\beta)(t-s)}\nabla f(\theta(s))\,ds + u\sigma\int_0^t e^{-(1-\beta)(t-s)}\,dW(s), \end{aligned} \tag{22}$$

where the exponent is adjusted using $e^{(1-\beta)s}/e^{(1-\beta)t} = e^{-(1-\beta)(t-s)}$.

The expression for $m(t)$ depends on $\theta(s)$ via $\nabla f(\theta(s))$, where:

$$\theta(s) = \theta_0 - \alpha\int_0^s m(\tau)\,d\tau. \tag{23}$$

Thus, the complete solution is:

$$\begin{cases} m(t) = u\displaystyle\int_0^t e^{-(1-\beta)(t-s)}\nabla f(\theta(s))\,ds + u\sigma\int_0^t e^{-(1-\beta)(t-s)}\,dW(s), \\ \theta(t) = \theta_0 - \alpha\displaystyle\int_0^t m(s)\,ds. \end{cases} \tag{24}$$

This integral form encapsulates the coupled dynamics, with $\nabla f(\theta(t))$ linking the equations and the stochastic term $\int e^{-(1-\beta)(t-s)}\,dW(s)$ as an Itô integral. ∎

Theorem A.7. *Consider the unified momentum estimator $m(t)$ defined by the stochastic differential equation (SDE) from Lemma A.4, with solution given in Lemma A.6. Let the bias be defined relative to the expected true gradient:* $\mathrm{Bias}(m(t)) = \mathbb{E}[m(t)] - \mathbb{E}[\nabla f(\theta(t))]$. *Assuming that the gradient $\nabla f(\theta(t))$ is bounded and Lipschitz continuous, the asymptotic bounds (as $t \to \infty$) for the bias and variance of $m(t)$ as an estimator are given by:*

$$\|\mathrm{Bias}(m(t))\|^2 \leq \left(\frac{u^2 \alpha L G}{(1-\beta)^3} + \frac{u^2 \alpha L \sigma}{\sqrt{2}(1-\beta)^{2.5}} + \left(\frac{u}{1-\beta} - 1 \right) G \right)^2, \tag{25}$$

where L is the Lipschitz constant, G bounds the gradient norm $\|\nabla f(\theta(t))\|$, and the second term inside the parenthesis explicitly captures the parameter-shift bias induced by the stochastic noise σ.

$$\mathrm{Var}(m(t)) \leq \frac{u^2 \sigma^2}{1-\beta} + \frac{2u^2 V^2}{(1-\beta)^2}, \tag{26}$$

where σ^2 is the total variance of the stochastic gradient noise, and V^2 conservatively bounds the variance of the true gradient sequence, i.e., $\mathrm{Var}(\nabla f(\theta(t))) \leq V^2$.

Proof. We compute the bias and variance of $m(t)$ relative to $\mathbb{E}[\nabla f(\theta(t))]$.

1. Bias Calculation

Consider the unified momentum update rule:

$$m_t = \beta m_{t-1} + u g_t, \quad \theta_t = \theta_{t-1} - \alpha m_t, \tag{27}$$

where $\beta \in [0, 1)$ represents the decay or momentum factor, $u \in [1-\beta, 1]$ is a scaling parameter controlling the gradient contribution, $\alpha > 0$ is the learning rate, and $g_t = \nabla f(\theta_t) + \epsilon_t$ with $\epsilon_t \sim \mathcal{N}(0, \sigma^2 I)$.

In continuous time, the expectation of $m(t)$ is:

$$\mathbb{E}[m(t)] = u \int_0^t e^{-(1-\beta)(t-s)} \mathbb{E}[\nabla f(\theta(s))]\, ds, \tag{28}$$

since the stochastic term has zero mean:

$$\mathbb{E}\left[u\sigma \int_0^t e^{-(1-\beta)(t-s)}\, dW(s) \right] = 0. \tag{29}$$

The squared bias is defined as:

$$\begin{aligned} (\mathrm{Bias}(m(t)))^2 &= (\mathbb{E}[m(t)] - \mathbb{E}[\nabla f(\theta(t))])^2 \\ &= \left(u \int_0^t e^{-(1-\beta)(t-s)} \mathbb{E}[\nabla f(\theta(s))]\, ds - \mathbb{E}[\nabla f(\theta(t))] \right)^2. \end{aligned} \tag{30}$$

We assume ∇f is Lipschitz continuous with constant $L > 0$:

$$\|\nabla f(\theta) - \nabla f(\phi)\| \leq L \|\theta - \phi\|, \quad \forall \theta, \phi. \tag{31}$$

Given $u \geq 1 - \beta$, so $\frac{u}{1-\beta} \geq 1$. From the continuous-time dynamics $\frac{d\theta}{dt} = -\alpha m(t)$, integrating from s to t ($s < t$) yields:

$$\theta(s) - \theta(t) = \alpha \int_s^t m(u)\, du. \tag{32}$$

To bound $\mathbb{E}[\|\theta(s) - \theta(t)\|]$, we must first bound the magnitude of the momentum $\mathbb{E}[\|m(u)\|]$ considering both the gradient drift and the noise diffusion:

$$m(u) = u \int_0^u e^{-(1-\beta)(u-v)} \nabla f(\theta(v))\, dv + u\sigma \int_0^u e^{-(1-\beta)(u-v)}\, dW(v). \tag{33}$$

Taking the expectation of the norm and applying Jensen's inequality to the stochastic term:

$$\begin{aligned}\mathbb{E}[\|m(u)\|] &\leq u\int_0^u e^{-(1-\beta)(u-v)}\mathbb{E}[\|\nabla f(\theta(v))\|]\,dv + \mathbb{E}\left[\left\|u\sigma\int_0^u e^{-(1-\beta)(u-v)}\,dW(v)\right\|\right] \\ &\leq \frac{uG}{1-\beta} + \sqrt{u^2\sigma^2\frac{1-e^{-2(1-\beta)u}}{2(1-\beta)}} \\ &\leq \frac{uG}{1-\beta} + \frac{u\sigma}{\sqrt{2(1-\beta)}} := M.\end{aligned} \tag{34}$$

Thus, taking the expected norm for the parameter difference:

$$\mathbb{E}\left[\|\theta(s)-\theta(t)\|\right] \leq \alpha\int_s^t \mathbb{E}[\|m(u)\|]\,du \leq \alpha M(t-s). \tag{35}$$

Rewrite the bias by splitting the integral:

$$\begin{aligned}\text{Bias}(m(t)) = u\int_0^t e^{-(1-\beta)(t-s)}\left(\mathbb{E}[\nabla f(\theta(s))] - \mathbb{E}[\nabla f(\theta(t))]\right)\,ds \\ + \mathbb{E}[\nabla f(\theta(t))]\left(u\int_0^t e^{-(1-\beta)(t-s)}\,ds - 1\right).\end{aligned} \tag{36}$$

Compute the second integral:

$$u\int_0^t e^{-(1-\beta)(t-s)}\,ds = u\frac{1-e^{-(1-\beta)t}}{1-\beta}. \tag{37}$$

Apply the triangle inequality and the bounded gradient assumption:

$$\|\text{Bias}(m(t))\| \leq \underbrace{\left\|u\int_0^t e^{-(1-\beta)(t-s)}\left(\mathbb{E}[\nabla f(\theta(s))] - \mathbb{E}[\nabla f(\theta(t))]\right)\,ds\right\|}_{:=I_1} + \left|u\frac{1-e^{-(1-\beta)t}}{1-\beta} - 1\right|G. \tag{38}$$

Bound I_1 using Lipschitz continuity and our bound M:

$$\begin{aligned}\|I_1\| &\leq u\int_0^t e^{-(1-\beta)(t-s)}\mathbb{E}\big[\|\nabla f(\theta(s)) - \nabla f(\theta(t))\|\big]\,ds \\ &\leq uL\alpha M\int_0^t e^{-(1-\beta)(t-s)}(t-s)\,ds.\end{aligned} \tag{39}$$

Evaluate the integral:

$$\begin{aligned}\int_0^t e^{-(1-\beta)(t-s)}(t-s)\,ds &= \int_0^t e^{-(1-\beta)\tau}\tau\,d\tau \\ &= \frac{1}{(1-\beta)^2} - \left(\frac{t}{1-\beta} + \frac{1}{(1-\beta)^2}\right)e^{-(1-\beta)t} \leq \frac{1}{(1-\beta)^2},\end{aligned} \tag{40}$$

thus $\|I_1\| \leq \frac{u\alpha LM}{(1-\beta)^2}$. Then:

$$\|\text{Bias}(m(t))\| \leq \frac{u\alpha LM}{(1-\beta)^2} + \left|u\frac{1-e^{-(1-\beta)t}}{1-\beta} - 1\right|G. \tag{41}$$

As $t\to\infty$, $e^{-(1-\beta)t}\to 0$, giving $\left|u\frac{1-e^{-(1-\beta)t}}{1-\beta} - 1\right| \to \frac{u}{1-\beta} - 1$. Substitute $M = \frac{uG}{1-\beta} + \frac{u\sigma}{\sqrt{2(1-\beta)}}$ and square the bound:

$$\begin{aligned}\|\text{Bias}(m(t))\|^2 &\leq \left(\frac{u\alpha LM}{(1-\beta)^2} + \left(u\frac{1-e^{-(1-\beta)t}}{1-\beta} - 1\right)G\right)^2 \\ &\leq \left(\frac{u\alpha L}{(1-\beta)^2}\left(\frac{uG}{1-\beta} + \frac{u\sigma}{\sqrt{2(1-\beta)}}\right) + \left(\frac{u}{1-\beta} - 1\right)G\right)^2 \\ &= \left(\frac{u^2\alpha LG}{(1-\beta)^3} + \frac{u^2\alpha L\sigma}{\sqrt{2}(1-\beta)^{2.5}} + \left(\frac{u}{1-\beta} - 1\right)G\right)^2.\end{aligned} \tag{42}$$

2. Variance Calculation
The fluctuation $m(t) - \mathbb{E}[m(t)]$ is:

$$m(t) - \mathbb{E}[m(t)] = \underbrace{u \int_0^t e^{-(1-\beta)(t-s)} \left[\nabla f(\theta(s)) - \mathbb{E}[\nabla f(\theta(s))]\right] ds}_{\mathcal{T}_{\text{Grad Diff}}} + \underbrace{u\sigma \int_0^t e^{-(1-\beta)(t-s)} \, dW(s)}_{\mathcal{T}_{\text{Noise Diff}}}. \tag{43}$$

Using the inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, the variance becomes:

$$\begin{aligned} \text{Var}(m(t)) &= \mathbb{E}\Big[\|m(t) - \mathbb{E}[m(t)]\|^2\Big] \\ &\leq 2\,\mathbb{E}\big[\|\mathcal{T}_{\text{Grad Diff}}\|^2\big] + 2\,\mathbb{E}\big[\|\mathcal{T}_{\text{Noise Diff}}\|^2\big]. \end{aligned} \tag{44}$$

The noise variance term is derived using the Itô isometry:

$$\begin{aligned} 2\,\mathbb{E}\big[\|\mathcal{T}_{\text{Noise Diff}}\|^2\big] &= 2u^2\sigma^2 \mathbb{E}\left[\left(\int_0^t e^{-(1-\beta)(t-s)} \, dW(s)\right)^2\right] \\ &= 2u^2\sigma^2 \int_0^t e^{-2(1-\beta)(t-s)} \, ds \quad \text{(Itô isometry)} \\ &= 2u^2\sigma^2 \frac{1 - e^{-2(1-\beta)t}}{2(1-\beta)} \leq \frac{u^2\sigma^2}{1-\beta}. \end{aligned} \tag{45}$$

For the gradient variance term, we apply the Cauchy-Schwarz inequality to properly bound the squared norm of the integral:

$$\begin{aligned} 2\,\mathbb{E}\big[\|\mathcal{T}_{\text{Grad Diff}}\|^2\big] &= 2u^2\,\mathbb{E}\left[\left\|\int_0^t e^{-(1-\beta)(t-s)/2} \cdot e^{-(1-\beta)(t-s)/2} \left[\nabla f(\theta(s)) - \mathbb{E}[\nabla f(\theta(s))]\right] ds\right\|^2\right] \\ &\leq 2u^2\,\mathbb{E}\left[\left(\int_0^t e^{-(1-\beta)(t-s)} \, ds\right)\left(\int_0^t e^{-(1-\beta)(t-s)} \left\|\nabla f(\theta(s)) - \mathbb{E}[\nabla f(\theta(s))]\right\|^2 ds\right)\right] \\ &\leq \frac{2u^2}{1-\beta} \int_0^t e^{-(1-\beta)(t-s)} \text{Var}(\nabla f(\theta(s))) \, ds. \end{aligned} \tag{46}$$

By Assumption A.2, $\text{Var}(\nabla f(\theta(s))) \leq V^2$, yielding:

$$\begin{aligned} 2\,\mathbb{E}\big[\|\mathcal{T}_{\text{Grad Diff}}\|^2\big] &\leq \frac{2u^2V^2}{1-\beta} \int_0^t e^{-(1-\beta)(t-s)} \, ds \\ &\leq \frac{2u^2V^2}{1-\beta}\left(\frac{1}{1-\beta}\right) = \frac{2u^2V^2}{(1-\beta)^2}. \end{aligned} \tag{47}$$

Combining both bounds, the total variance is bounded by:

$$\text{Var}(m(t)) \leq \frac{u^2\sigma^2}{1-\beta} + \frac{2u^2V^2}{(1-\beta)^2}. \tag{48}$$

■

B. Method Derivation (Section 3 in main paper)

B.1. Optimal Linear Filter Derivation for Gradient Estimation (Main paper Section 3.1)

In the stochastic gradient descent (SGD) process, given the sequence of gradients $\{g_i\}_{i=1}^{t}$, our objective is to estimate $\hat{g}_t$, which incorporates information from both historical gradients and the current gradient. The Optimal Linear Filter provides a mechanism to minimize the mean squared error in this estimation. We start by constructing $\hat{g}_t$ as a simple average and then refine it using the properties of the Optimal Linear Filter.

$$\begin{aligned}\hat{g}_t &= \frac{1}{t}\sum_{i=1}^{t} g_i = \frac{1}{t}\left(\sum_{i=1}^{t-1} g_i + g_t\right) = \frac{1}{t}\sum_{i=1}^{t-1} g_i + \frac{1}{t} g_t \\ &= \frac{1}{t}\left[(t-1)\cdot\frac{1}{t-1}\sum_{i=1}^{t-1} g_i\right] + \frac{1}{t} g_t \\ &= \frac{t-1}{t}\,\bar{g}_{1:t-1} + \frac{1}{t} g_t,\end{aligned} \tag{49}$$

where $\bar{g}_{1:t-1} = \frac{1}{t-1}\sum_{i=1}^{t-1} g_i$ denoting the averaging of the gradient under different θ_t to differentiate $\bar{g}_t$ which in fixed parameter θ.

To better capture historical information, we replace the arithmetic mean $\bar{g}_{1:t-1}$ with the momentum term $\widehat{m}_t$. Here we substitute the iteration m_{t-1} with m_t because of the absence of m_0 and the ease of implementation. Thus, we rewrite $\hat{g}_t$ as follows:

$$\begin{aligned}\hat{g}_t &\approx \frac{t-1}{t}\widehat{m}_t + \frac{1}{t} g_t \\ &= \left(1 - \frac{1}{t}\right)\widehat{m}_t + \frac{1}{t} g_t \\ &= \widehat{m}_t - K_t\widehat{m}_t + K_t g_t \\ &= \widehat{m}_t + K_t(g_t - \widehat{m}_t),\end{aligned} \tag{50}$$

where $K_t = \frac{1}{t}$ serves as an initial estimation gain that balances the influence of $\widehat{m}_t$ and g_t.

To achieve an optimal balance, we define $\hat{g}_t$ as a weighted combination of $\widehat{m}_t$ and g_t, aiming to minimize the variance of $\hat{g}_t$. Assuming independence between $\widehat{m}_t$ and g_t, we express the variance as:

$$\begin{aligned}\mathrm{Var}(\hat{g}_t) &= \mathrm{Var}((1-K_t)\widehat{m}_t + K_t g_t) \\ &= (1-K_t)^2\mathrm{Var}(\widehat{m}_t) + K_t^2\mathrm{Var}(g_t).\end{aligned} \tag{51}$$

To find the optimal K_t, we take the derivative of $\mathrm{Var}(\hat{g}_t)$ with respect to K_t and set it to zero:

$$\begin{aligned}\frac{\mathrm{dVar}(\hat{g}_t)}{\mathrm{d}K_t} &= -2(1-K_t)\mathrm{Var}(\widehat{m}_t) + 2K_t\mathrm{Var}(g_t) = 0, \\ (1-K_t)\mathrm{Var}(\widehat{m}_t) &= K_t\mathrm{Var}(g_t),\end{aligned} \tag{52}$$

solving for K_t gives:

$$K_t = \frac{\mathrm{Var}(\widehat{m}_t)}{\mathrm{Var}(\widehat{m}_t) + \mathrm{Var}(g_t)}. \tag{53}$$

The final expression for K_t indicates that the optimal interpolation coefficient is the ratio of the variance of the momentum term to the total variance. This embodies the Optimal Linear Filter's principle: optimally combining historical estimates with new observations to minimize estimation error due to stochastic noise in the gradient signal.

B.2. Variance Correction (Correction factor in main paper Section 3.1)

The momentum term m_t in stochastic gradient descent is defined as:

$$m_t = (1-\beta_1)\sum_{i=1}^{t}\beta_1^{t-i} g_i, \tag{54}$$

which means that m_t is a weighted sum of past gradients, where the weights decrease exponentially over time according to the factor β_1.

To accurately estimate the variance of m_t using the variance of g_t, we derive a correction factor under the assumption that the stochastic gradients g_t are independent with bounded variance σ_g^2.

Each weighted gradient term $\beta_1^{t-i} g_i$ has a variance of $\beta_1^{2(t-i)} \sigma_g^2$, because the variance scaling factor becomes $\beta_1^{2(t-i)}$ in the variance computation due to the quadratic nature of the variance operator.

Given that m_t is a sum of these weighted terms and assuming independence among g_i, the variance of m_t is the sum of the variances of all weighted gradients:

$$\sigma_{m_t}^2 = (1-\beta_1)^2 \sigma_g^2 \sum_{i=1}^{t} \beta_1^{2(t-i)}. \tag{55}$$

The factor $(1-\beta_1)^2$ appears from the multiplication factor $(1-\beta_1)$ in the definition of m_t, which also applies to the variance calculation.

The summation $\sum_{i=1}^{t} \beta_1^{2(t-i)}$ forms a geometric series:

$$\sum_{i=1}^{t} \beta_1^{2(t-i)} = \frac{1-\beta_1^{2t}}{1-\beta_1^2}. \tag{56}$$

As $t \to \infty$ and given that $\beta_1 < 1$, we find that $\beta_1^{2t} \to 0$, so the series converges to:

$$\sum_{i=1}^{t} \beta_1^{2(t-i)} \approx \frac{1}{1-\beta_1^2}. \tag{57}$$

Substituting back, we obtain the long-term variance of m_t as:

$$\sigma_{m_t}^2 = \frac{(1-\beta_1)^2}{1-\beta_1^2} \sigma_g^2 = \frac{1-\beta_1}{1+\beta_1} \sigma_g^2. \tag{58}$$

Thus, the correction factor we derived is:

$$\left(\frac{1-\beta_1}{1+\beta_1}\right) \cdot (1-\beta_1^{2t}). \tag{59}$$

This correction factor $\left(\frac{1-\beta_1}{1+\beta_1}\right) \cdot (1-\beta_1^{2t})$ allows us to adjust the variance of the EMA gradient to accurately estimate the variance of the momentum gradient m_t using the original variance σ_g^2. This adjustment reflects the effect of exponentially decaying weights in m_t, yielding a more stable gradient estimate with reduced noise over time.

B.3. Fusion of Gaussian Distributions (Main paper Section 3.2)

In this section, we address the fusion of two Gaussian distributions to produce a more reliable gradient estimate in the stochastic gradient descent (SGD) process. This fusion combines information from both the historical momentum term $\widehat{m}_t$ and the current gradient g_t, resulting in an estimate with reduced uncertainty. Here, "fusion" refers to finding an optimal combined distribution that minimizes mean-square error by utilizing both sources of information.

Consider the following two Gaussian distributions:

- The momentum term $\widehat{m}_t$ follows a normal distribution with mean μ_m and variance σ_m^2, denoted as $\widehat{m}_t \sim \mathcal{N}(\mu_m, \sigma_m^2)$.
- The current gradient g_t follows a normal distribution with mean μ_g and variance σ_g^2, denoted as $g_t \sim \mathcal{N}(\mu_g, \sigma_g^2)$.

Linear estimator perspective. Before presenting the probability density product approach, we first derive the fused estimate $\hat{g}_t$ from a weighted linear combination perspective. We define $\hat{g}_t$ as:

$$\hat{g}_t = (1-K_t)\widehat{m}_t + K_t g_t, \tag{60}$$

where K_t is a variance-based weighting coefficient:

$$K_t = \frac{\sigma_m^2}{\sigma_m^2 + \sigma_g^2}, \quad \text{thus} \quad 1-K_t = \frac{\sigma_g^2}{\sigma_m^2 + \sigma_g^2}. \tag{61}$$

Assuming independence between $\widehat{m}_t$ and g_t, the expectation of $\hat{g}_t$ is:

$$\mathbb{E}[\hat{g}_t] = (1 - K_t)\mu_m + K_t\mu_g = \frac{\sigma_g^2\mu_m + \sigma_m^2\mu_g}{\sigma_m^2 + \sigma_g^2}. \tag{62}$$

The variance of $\hat{g}_t$ becomes:

$$\begin{aligned} \mathrm{Var}(\hat{g}_t) &= (1 - K_t)^2\sigma_m^2 + K_t^2\sigma_g^2 \\ &= \left(\frac{\sigma_g^2}{\sigma_m^2 + \sigma_g^2}\right)^2 \sigma_m^2 + \left(\frac{\sigma_m^2}{\sigma_m^2 + \sigma_g^2}\right)^2 \sigma_g^2 \\ &= \frac{\sigma_g^4\sigma_m^2 + \sigma_m^4\sigma_g^2}{(\sigma_m^2 + \sigma_g^2)^2} = \frac{\sigma_m^2\sigma_g^2}{\sigma_m^2 + \sigma_g^2}. \end{aligned} \tag{63}$$

Probability density product derivation. We now show that the same fused Gaussian distribution arises from multiplying the two individual Gaussian probability densities:

$$N(\widehat{m}_t; \mu_m, \sigma_m) \cdot N(g_t; \mu_g, \sigma_g) = \frac{1}{2\pi\sigma_m\sigma_g} \exp\left(-\frac{(\hat{g}_t - \mu_m)^2}{2\sigma_m^2} - \frac{(\hat{g}_t - \mu_g)^2}{2\sigma_g^2}\right). \tag{64}$$

To derive the fused form, we simplify the exponent by completing the square:

$$\begin{aligned} \text{Exponent} &= -\frac{(\hat{g}_t - \mu_m)^2}{2\sigma_m^2} - \frac{(\hat{g}_t - \mu_g)^2}{2\sigma_g^2} \\ &= -\frac{\sigma_g^2(\hat{g}_t - \mu_m)^2 + \sigma_m^2(\hat{g}_t - \mu_g)^2}{2\sigma_m^2\sigma_g^2} \\ &= -\frac{\left(\hat{g}_t - \frac{\sigma_g^2\mu_m + \sigma_m^2\mu_g}{\sigma_m^2 + \sigma_g^2}\right)^2}{\frac{2\sigma_m^2\sigma_g^2}{\sigma_m^2 + \sigma_g^2}} + \frac{(\mu_m - \mu_g)^2}{2(\sigma_m^2 + \sigma_g^2)}. \end{aligned} \tag{65}$$

Ignoring constant terms, we identify the resulting fused distribution:

$$\mu_{\hat{g}_t} = \frac{\sigma_g^2\mu_m + \sigma_m^2\mu_g}{\sigma_m^2 + \sigma_g^2}, \quad \sigma_{\hat{g}_t}^2 = \frac{\sigma_m^2\sigma_g^2}{\sigma_m^2 + \sigma_g^2}. \tag{66}$$

Equivalence and insight. This demonstrates that the PDF product view yields the same fused mean and variance as the minimum mean-square error (MMSE) linear estimator. The fused mean $\mu_{\hat{g}_t}$ is closer to the distribution with smaller variance, indicating greater trust in more certain estimates [10]. The fused variance $\sigma_{\hat{g}_t}^2$ is always less than either original variance, demonstrating the variance reduction benefit of fusion.

This equivalence between statistical estimation and probabilistic fusion confirms the theoretical soundness of the SGDF method. It validates the fusion-based formulation from both a Bayesian and a signal processing perspective.

B.4. Modulating Observation Variance through Power Scaling

The preceding analysis yields the optimal linear fusion gain under the MMSE criterion

$$K_t \;=\; \frac{\sigma_m^2}{\sigma_m^2 + \sigma_g^2}, \tag{67}$$

which balances the historical momentum estimate and the instantaneous stochastic gradient according to their uncertainties.

In practice, the variance estimates (especially σ_g^2) can be noisy or biased due to mini-batch stochasticity and nonstationarity. To improve robustness while remaining consistent with our convergence analysis, we adopt a power-scaled gain

$$\tilde{K}_t \;=\; K_t^{\gamma}, \qquad \gamma = \tfrac{1}{2}, \tag{68}$$

since $K_t \in [0,1]$ element-wise, the scaled gain still satisfies $\|\tilde{K}_t\|_\infty \leq 1$, which is the sole requirement for our convergence guarantees; therefore, the theoretical results remain unchanged.

With $K_t = \frac{\sigma_m^2}{\sigma_m^2+\sigma_g^2}$, the choice $\gamma = \frac{1}{2}$ gives

$$\tilde{K}_t \;=\; \sqrt{K_t} \;=\; \sqrt{\frac{\sigma_m^2}{\sigma_m^2+\sigma_g^2}}. \tag{69}$$

We show that $\tilde{K}_t$ can be written in the same variance-fusion form as K_t by introducing an *effective* observation variance $\tilde{\sigma}_g^2$ such that

$$\tilde{K}_t \;=\; \frac{\sigma_m^2}{\sigma_m^2+\tilde{\sigma}_g^2}. \tag{70}$$

Equating the two expressions and solving for $\tilde{\sigma}_g^2$:

$$\begin{aligned}
\frac{\sigma_m^2}{\sigma_m^2+\tilde{\sigma}_g^2} &= \sqrt{\frac{\sigma_m^2}{\sigma_m^2+\sigma_g^2}}, \\
\frac{\sigma_m^2+\tilde{\sigma}_g^2}{\sigma_m^2} &= \sqrt{\frac{\sigma_m^2+\sigma_g^2}{\sigma_m^2}} = \sqrt{1+\frac{\sigma_g^2}{\sigma_m^2}}, \\
1+\frac{\tilde{\sigma}_g^2}{\sigma_m^2} &= \sqrt{1+\frac{\sigma_g^2}{\sigma_m^2}}, \\
\frac{\tilde{\sigma}_g^2}{\sigma_m^2} &= \sqrt{1+\frac{\sigma_g^2}{\sigma_m^2}} - 1.
\end{aligned} \tag{71}$$

Therefore, according to Eq. (71), we have:

$$\tilde{\sigma}_g^2 \;=\; \sigma_m^2\left(\sqrt{1+\tfrac{\sigma_g^2}{\sigma_m^2}} - 1\right).$$

Therefore, the scaled gain $\tilde{K}_t = \sqrt{K_t}$ is equivalently a standard variance-based fusion gain with a reparameterized (effective) observation variance:

$$\tilde{K}_t \;=\; \frac{\sigma_m^2}{\sigma_m^2+\tilde{\sigma}_g^2}, \qquad \tilde{\sigma}_g^2 \;=\; \sigma_m^2\left(\sqrt{1+\tfrac{\sigma_g^2}{\sigma_m^2}} - 1\right). \tag{72}$$

This view shows that power-scaling preserves the fusion structure while introducing a controlled regularization against overconfident (noisy) instantaneous gradient observations.

C. Convergence analysis in convex online learning case (Theorem 3.2 in main paper).

Assumption C.1. Variables are bounded: $\exists D, D_\infty$ such that $\forall t, \|\theta_t - \theta^*\|_2 \leq D, \|\theta_t - \theta^*\|_\infty \leq D_\infty$. Gradients are bounded: $\exists G, G_\infty$ such that $\forall t, \|g_t\|_2 \leq G, \|g_t\|_\infty \leq G_\infty$. The interpolation parameter satisfies $K_{t,i} \in [0, 1]$. Furthermore, we assume the interpolation parameter sequence has sublinear total variation, i.e., $\sum_{t=1}^{T-1} |K_{t,i} - K_{t+1,i}| \leq \mathcal{O}(\sqrt{T})$.

Definition C.2. Let $f_t(\theta_t)$ be the loss at time t and $f_t(\theta^*)$ be the loss of the best possible strategy at the same time. The cumulative regret $R(T)$ at time T is defined as:

$$R(T) = \sum_{t=1}^{T} \left(f_t(\theta_t) - f_t(\theta^*)\right) \tag{73}$$

Definition C.3. A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if for all $x, y \in \mathbb{R}^d$ and for all $\lambda \in [0, 1]$,

$$\lambda f(x) + (1 - \lambda) f(y) \geq f(\lambda x + (1 - \lambda) y) \tag{74}$$

Also, notice that a convex function can be lower bounded by a hyperplane at its tangent.

Lemma C.4. *If a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex, then for all $x, y \in \mathbb{R}^d$,*

$$f(x) - f(y) \leq \nabla f(x)^T (x - y) \tag{75}$$

The above lemma can be used to upper bound the regret, and our proof for the main theorem is constructed by substituting the hyperplane with SGDF update rules.

We define $g_t \triangleq \nabla f_t(\theta_t)$ and $g_{t,i}$ as the i^{th} element. Let $\hat{g}_t$ be the effective update direction.

Lemma C.5. *Let gradients be bounded by $|g_{t,i}| \leq G_\infty$. For any $T \geq 1$, the sum of squared bounded elements discounted by $\sqrt{t}$ is strictly bounded by:*

$$\sum_{t=1}^{T} \frac{g_{t,i}^2}{\sqrt{t}} \leq 2G_\infty^2 \sqrt{T} \tag{76}$$

Proof. Since $g_{t,i}^2 \leq G_\infty^2$, we can bound the summation using the integral test. Since $1/\sqrt{t}$ is a monotonically decreasing function for $t \geq 1$:

$$\begin{aligned} \sum_{t=1}^{T} \frac{g_{t,i}^2}{\sqrt{t}} &\leq G_\infty^2 \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \\ &\leq G_\infty^2 \left(1 + \int_1^T \frac{1}{\sqrt{t}} dt\right) \\ &= G_\infty^2 \left(1 + 2\sqrt{T} - 2\right) \leq 2G_\infty^2 \sqrt{T} \end{aligned} \tag{77}$$

■

Lemma C.6. *Let the exponential moving average be $m_{t,i} = \beta_1 m_{t-1,i} + (1-\beta_1) g_{t,i}$, with bias-correction $\widehat{m}_{t,i} = m_{t,i}/(1-\beta_1^t)$. Under the update rule $\hat{g}_{t,i} = \widehat{m}_{t,i} + K_{t,i}(g_{t,i} - \widehat{m}_{t,i})$, the effective update direction is bounded by $|\hat{g}_{t,i}| \leq G_\infty$.*

Proof. By mathematical induction, since $m_{t,i}$ is a convex combination of past bounded gradients, we have $|m_{t,i}| \leq (1 - \beta_1^t) G_\infty$. Therefore, the bias-corrected momentum satisfies $|\widehat{m}_{t,i}| \leq G_\infty$. The update direction is a convex combination $\hat{g}_{t,i} = K_{t,i} g_{t,i} + (1 - K_{t,i}) \widehat{m}_{t,i}$. Since both components are bounded by G_∞ and $K_{t,i} \in [0, 1]$, we rigorously have $|\hat{g}_{t,i}| \leq G_\infty$.

■

Lemma C.7 (Bounded Total Variation)**.** *Assume the gradient sequence is bounded such that $\|\nabla f_t\|_\infty \leq G_\infty$ for all t, and the hyperparameters $\beta_1, \beta_2 \in [0, 1)$ are constants. If the power scaling factor is scheduled as $\gamma_t = \gamma_0/\sqrt{t}$, the effective interpolation gain $\tilde{K}_t = K_t^{\gamma_t}$ satisfies:*

$$\sum_{t=1}^{T-1} |K_{t+1}^{\gamma_{t+1}} - K_t^{\gamma_t}| \leq \mathcal{O}(\sqrt{T}) \tag{78}$$

Proof. Decomposition of the total variation using the triangle inequality.

$$\sum_{t=1}^{T-1} |K_{t+1}^{\gamma_{t+1}} - K_t^{\gamma_t}| \leq \sum_{t=1}^{T-1} \left(|K_{t+1}^{\gamma_{t+1}} - K_{t+1}^{\gamma_t}| + |K_{t+1}^{\gamma_t} - K_t^{\gamma_t}| \right)$$
$$= \sum_{t=1}^{T-1} \underbrace{|K_{t+1}^{\gamma_{t+1}} - K_{t+1}^{\gamma_t}|}_{\text{Part (A)}} + \sum_{t=1}^{T-1} \underbrace{|K_{t+1}^{\gamma_t} - K_t^{\gamma_t}|}_{\text{Part (B)}} \tag{79}$$

Bound Part (A) representing the variation in the exponent γ_t. By the Mean Value Theorem for $f(\gamma) = x^\gamma$ where $x \in [\delta, 1]$ and $\delta = \frac{\varepsilon}{G_\infty^2 + \varepsilon}$ (with $\varepsilon > 0$ being a small constant), there exists $\xi_t \in [\gamma_{t+1}, \gamma_t]$ such that:

$$|K_{t+1}^{\gamma_{t+1}} - K_{t+1}^{\gamma_t}| = |K_{t+1}^{\xi_t} \ln(K_{t+1})| \cdot |\gamma_{t+1} - \gamma_t|$$
$$\leq C_\varepsilon \cdot \gamma_0 \left(\frac{1}{\sqrt{t}} - \frac{1}{\sqrt{t+1}} \right) \tag{80}$$

Summing over t yields a telescoping series:

$$\sum_{t=1}^{T-1} |K_{t+1}^{\gamma_{t+1}} - K_{t+1}^{\gamma_t}| \leq C_\varepsilon \gamma_0 \sum_{t=1}^{T-1} \left(\frac{1}{\sqrt{t}} - \frac{1}{\sqrt{t+1}} \right)$$
$$= C_\varepsilon \gamma_0 \left(1 - \frac{1}{\sqrt{T}} \right) = \mathcal{O}(1) \tag{81}$$

Bound Part (B) representing the variation in the base K_t. Assuming $\gamma_0 \leq 1$, and using the Lipschitz continuity of x^{γ_t} on $[\delta, 1]$ with $\gamma_t \in (0, 1]$, we observe:

$$|K_{t+1}^{\gamma_t} - K_t^{\gamma_t}| \leq \left(\sup_{x \in [\delta, 1]} \gamma_t x^{\gamma_t - 1} \right) \cdot |K_{t+1} - K_t|$$
$$\leq \gamma_t \delta^{\gamma_t - 1} \cdot |K_{t+1} - K_t| \tag{82}$$

Even if $|K_{t+1} - K_t| = \mathcal{O}(1)$ due to constant β_2 and lack of smoothness, the decay of γ_t ensures:

$$\sum_{t=1}^{T-1} |K_{t+1}^{\gamma_t} - K_t^{\gamma_t}| \leq \text{Const} \cdot \sum_{t=1}^{T-1} \gamma_t$$
$$= \text{Const} \cdot \gamma_0 \sum_{t=1}^{T-1} \frac{1}{\sqrt{t}}$$
$$\leq 2 \cdot \text{Const} \cdot \gamma_0 \sqrt{T} = \mathcal{O}(\sqrt{T}) \tag{83}$$

Combine the bounds to achieve the final sublinear variation.

$$\sum_{t=1}^{T-1} |K_{t+1}^{\gamma_{t+1}} - K_t^{\gamma_t}| \leq \mathcal{O}(1) + \mathcal{O}(\sqrt{T})$$
$$= \mathcal{O}(\sqrt{T}) \tag{84}$$

■

Theorem C.8. *Assume that Assumption C.1 holds, and $\beta_1 \in [0, 1)$. Let the learning rate be $\alpha_t = \alpha/\sqrt{t}$. For all $T \geq 1$, SGDF achieves the following cumulative regret bound:*

$$R(T) \leq \sum_{i=1}^{d} \left(\frac{D_\infty^2}{2\alpha} + \alpha G_\infty^2 \frac{1 + \beta_1}{1 - \beta_1} \right) \sqrt{T} + \sum_{i=1}^{d} \frac{\beta_1 G_\infty D_\infty}{1 - \beta_1} \left(2 + \sum_{t=1}^{T-1} |K_{t,i} - K_{t+1,i}| \right) \tag{85}$$

Under the assumption that the total variation of the interpolation parameter is sublinear, i.e., $\sum_{t=1}^{T-1} |K_{t,i} - K_{t+1,i}| \leq \mathcal{O}(\sqrt{T})$, *we have $R(T) \leq \mathcal{O}(\sqrt{T})$. Consequently, the average regret converges to zero:* $\lim_{T \to \infty} \frac{R(T)}{T} = 0$.

Proof. Using Lemma C.4, we lower bound the convex functions to establish the regret connection:

$$f_t(\theta_t) - f_t(\theta^*) \leq \langle g_t, \theta_t - \theta^* \rangle = \sum_{i=1}^{d} g_{t,i}(\theta_{t,i} - \theta_i^*) \tag{86}$$

From Algorithm 1, the update direction incorporates bias correction and interpolation:

$$\hat{g}_{t,i} = K_{t,i} g_{t,i} + (1 - K_{t,i})\widehat{m}_{t,i} \implies g_{t,i} = \hat{g}_{t,i} + (1 - K_{t,i})(g_{t,i} - \widehat{m}_{t,i}) \tag{87}$$

Recall the momentum definition $m_{t,i} = \beta_1 m_{t-1,i} + (1-\beta_1) g_{t,i}$, giving $g_{t,i} - m_{t,i} = \frac{\beta_1}{1-\beta_1}(m_{t,i} - m_{t-1,i})$. Also, $\widehat{m}_{t,i} = m_{t,i}/(1-\beta_1^t)$. Expanding the difference accurately yields:

$$\begin{aligned} g_{t,i} - \widehat{m}_{t,i} &= (g_{t,i} - m_{t,i}) + \left(m_{t,i} - \frac{m_{t,i}}{1-\beta_1^t} \right) \\ &= \frac{\beta_1}{1-\beta_1}(m_{t,i} - m_{t-1,i}) - \frac{\beta_1^t}{1-\beta_1^t} m_{t,i} \end{aligned} \tag{88}$$

Substituting this back, we decompose the inner product into three parts:

$$\sum_{t=1}^{T} g_{t,i}(\theta_{t,i} - \theta_i^*) = \underbrace{\sum_{t=1}^{T} \hat{g}_{t,i}(\theta_{t,i} - \theta_i^*)}_{\text{(A)}} + \underbrace{\frac{\beta_1}{1-\beta_1} \sum_{t=1}^{T} (1 - K_{t,i})(m_{t,i} - m_{t-1,i})(\theta_{t,i} - \theta_i^*)}_{\text{(B)}} - \underbrace{\sum_{t=1}^{T} (1 - K_{t,i}) \frac{\beta_1^t}{1-\beta_1^t} m_{t,i}(\theta_{t,i} - \theta_i^*)}_{\text{(C)}} \tag{89}$$

For part (A), using the parameter update rule $\theta_{t+1,i} = \theta_{t,i} - \alpha_t \hat{g}_{t,i}$, we have:

$$\hat{g}_{t,i}(\theta_{t,i} - \theta_i^*) = \frac{(\theta_{t,i} - \theta_i^*)^2 - (\theta_{t+1,i} - \theta_i^*)^2}{2\alpha_t} + \frac{\alpha_t}{2} \hat{g}_{t,i}^2 \tag{90}$$

Summing over T and noting $\alpha_t = \alpha/\sqrt{t}$, we get a telescoping sum:

$$\begin{aligned} \sum_{t=1}^{T} \hat{g}_{t,i}(\theta_{t,i} - \theta_i^*) &\leq \frac{D_\infty^2}{2\alpha_1} + \sum_{t=2}^{T} \frac{D_\infty^2}{2} \left(\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} \right) + \frac{\alpha}{2} \sum_{t=1}^{T} \frac{\hat{g}_{t,i}^2}{\sqrt{t}} \\ &\leq \frac{D_\infty^2 \sqrt{T}}{2\alpha} + \alpha G_\infty^2 \sqrt{T} \quad \text{(Using Lemma C.5 and C.6)} \end{aligned} \tag{91}$$

For part (B), let $C_{t,i} = 1 - K_{t,i}$. We apply Summation by Parts (Abel transformation):

$$\begin{aligned} &\sum_{t=1}^{T} C_{t,i}(m_{t,i} - m_{t-1,i})(\theta_{t,i} - \theta_i^*) \\ &= C_{T,i} m_{T,i}(\theta_{T,i} - \theta_i^*) - C_{1,i} m_{0,i}(\theta_{1,i} - \theta_i^*) - \sum_{t=1}^{T-1} m_{t,i} \left[C_{t+1,i}(\theta_{t+1,i} - \theta_i^*) - C_{t,i}(\theta_{t,i} - \theta_i^*) \right] \end{aligned} \tag{92}$$

Since $m_{0,i} = 0$, the boundary term is bounded by $G_\infty D_\infty$. We rigorously expand the difference inside the summation:

$$\begin{aligned} &C_{t+1,i}(\theta_{t+1,i} - \theta_i^*) - C_{t,i}(\theta_{t,i} - \theta_i^*) \\ &= C_{t+1,i}(\theta_{t+1,i} - \theta_{t,i}) + (C_{t+1,i} - C_{t,i})(\theta_{t,i} - \theta_i^*) \\ &= -(1 - K_{t+1,i})\alpha_t \hat{g}_{t,i} + (K_{t,i} - K_{t+1,i})(\theta_{t,i} - \theta_i^*) \end{aligned} \tag{93}$$

Taking the absolute value and substituting back, we obtain:

$$\begin{aligned} |\text{(B)}| &\leq \frac{\beta_1}{1-\beta_1} \left(G_\infty D_\infty + \sum_{t=1}^{T-1} |m_{t,i}| \cdot |1 - K_{t+1,i}| \alpha_t |\hat{g}_{t,i}| + \sum_{t=1}^{T-1} |m_{t,i}| \cdot |K_{t,i} - K_{t+1,i}| \cdot |\theta_{t,i} - \theta_i^*| \right) \\ &\leq \frac{\beta_1}{1-\beta_1} \left(G_\infty D_\infty + \alpha G_\infty^2 \sum_{t=1}^{T-1} \frac{1}{\sqrt{t}} + G_\infty D_\infty \sum_{t=1}^{T-1} |K_{t,i} - K_{t+1,i}| \right) \\ &\leq \frac{\beta_1}{1-\beta_1} \left(G_\infty D_\infty + 2\alpha G_\infty^2 \sqrt{T} + G_\infty D_\infty \sum_{t=1}^{T-1} |K_{t,i} - K_{t+1,i}| \right) \end{aligned} \tag{94}$$

For part (C), the bias correction residual can be tightly bounded. Recall from Lemma C.6 that $|m_{t,i}| \leq (1-\beta_1^t)G_\infty$. Substituting this into the expression allows us to exactly cancel the denominator:

$$\begin{aligned} |\text{(C)}| &\leq \sum_{t=1}^{T} |1 - K_{t,i}| \frac{\beta_1^t}{1-\beta_1^t} |m_{t,i}||\theta_{t,i} - \theta_i^*| \\ &\leq \sum_{t=1}^{T} 1 \cdot \frac{\beta_1^t}{1-\beta_1^t}(1-\beta_1^t) G_\infty D_\infty \\ &= G_\infty D_\infty \sum_{t=1}^{T} \beta_1^t \leq \frac{\beta_1}{1-\beta_1} G_\infty D_\infty \end{aligned} \tag{95}$$

Summing parts (A), (B), and (C) over all d dimensions and grouping the terms by $\sqrt{T}$ and the interpolation total variation, we obtain the highly condensed cumulative regret:

$$R(T) \leq \sum_{i=1}^{d} \left[\left(\frac{D_\infty^2}{2\alpha} + \alpha G_\infty^2 \frac{1+\beta_1}{1-\beta_1} \right) \sqrt{T} + \frac{\beta_1 G_\infty D_\infty}{1-\beta_1} \left(2 + \sum_{t=1}^{T-1} |K_{t,i} - K_{t+1,i}| \right) \right] \tag{96}$$

Given Lemma C.7 that $\sum |K_{t,i} - K_{t+1,i}| \leq \mathcal{O}(\sqrt{T})$, the cumulative regret satisfies $R(T) \leq \mathcal{O}(\sqrt{T})$.
To prove convergence, we evaluate the average regret as $T \to \infty$:

$$\lim_{T\to\infty} \frac{R(T)}{T} \leq \lim_{T\to\infty} \frac{\mathcal{O}(\sqrt{T})}{T} = \lim_{T\to\infty} \mathcal{O}\left(\frac{1}{\sqrt{T}} \right) = 0 \tag{97}$$

■

D. Convergence analysis for non-convex stochastic optimization (Theorem 3.3 in main paper).

We have relaxed the assumption on the objective function, allowing it to be non-convex, and adjusted the criterion for convergence from the statistic $R(T)$ to $\mathbb{E}(T)$. Let's briefly review the assumptions and the criterion for convergence after relaxing the assumption:

Assumption D.1.

- A1 Bounded variables (same as convex). $\|\theta - \theta^*\|_2 \leq D,\ \forall \theta, \theta^*$ or for any dimension i of the variable, $\|\theta_i - \theta_i^*\|_2 \leq D_i,\ \forall \theta_i, \theta_i^*$
- A2 The noisy gradient is unbiased. For $\forall t$, the random variable ζ_t is defined as $\zeta_t = g_t - \nabla f(\theta_t)$, ζ_t satisfy $\mathbb{E}[\zeta_t] = 0$, $\mathbb{E}\left[\|\zeta_t\|_2^2\right] \leq \sigma^2$, and when $t_1 \neq t_2$, ζ_{t_1} and ζ_{t_2} are statistically independent, i.e., $\zeta_{t_1} \perp \zeta_{t_2}$.
- A3 Bounded gradient and noisy gradient. At step t, the algorithm can access a bounded noisy gradient, and the true gradient is also bounded. $i.e.\ ||\nabla f(\theta_t)|| \leq G,\ ||g_t|| \leq G,\ \forall t > 1$.
- A4 The property of function. The objective function $f(\theta)$ is a global loss function, defined as $f(\theta) = \lim_{T \longrightarrow \infty} \frac{1}{T} \sum_{t=1}^{T} f_t(\theta)$. Although $f(\theta)$ is no longer a convex function, it must still be a L-smooth function, i.e., it satisfies (1) f is differentiable, ∇f exists everywhere in the domain; (2) there exists $L > 0$ such that for any θ_1 and θ_2 in the domain, (first definition)

$$f(\theta_2) \leq f(\theta_1) + \langle \nabla f(\theta_1), \theta_2 - \theta_1 \rangle + \frac{L}{2} \|\theta_2 - \theta_1\|_2^2 \tag{98}$$

or (second definition)

$$\|\nabla f(\theta_1) - \nabla f(\theta_2)\|_2 \leq L \|\theta_1 - \theta_2\|_2 \tag{99}$$

This condition is also known as *L - Lipschitz*.

Definition D.2. The criterion for convergence is the statistic $\mathbb{E}(T)$:

$$\mathbb{E}(T) = \min_{t=1,2,\ldots,T} \mathbb{E}_t \left[\|\nabla f(\theta_t)\|_2^2\right] \tag{100}$$

When $T \to \infty$, if the amortized value of $\mathbb{E}(T)$, $\mathbb{E}(T)/T \to 0$, we consider such an algorithm to be convergent, and generally, the slower $\mathbb{E}(T)$ grows with T, the faster the algorithm converges.

Definition D.3. Define ξ_t as

$$\xi_t = \begin{cases} \theta_t & t = 1 \\ \theta_t + \frac{\beta_1}{1-\beta_1}(\theta_t - \theta_{t-1}) & t \geq 2 \end{cases} \tag{101}$$

Lemma D.4. *Let f be an L-smooth function. Then, for any points ξ_t and θ_t, the following inequality holds:*

$$f(\xi_{t+1}) - f(\xi_t) \leq \frac{L}{2} \|\xi_t - \theta_t\|_2^2 + L \|\xi_{t+1} - \xi_t\|_2^2 + \langle \nabla f(\theta_t), \xi_{t+1} - \xi_t \rangle \tag{102}$$

Proof. Since f is an L-smooth function,

$$\|\nabla f(\xi_t) - \nabla f(\theta_t)\|_2^2 \leq L^2 \|\xi_t - \theta_t\|_2^2 \tag{103}$$

Thus,

$$
\begin{aligned}
& f\left(\xi_{t+1}\right)-f\left(\xi_t\right) \\
& \leq\left\langle\nabla f\left(\xi_t\right), \xi_{t+1}-\xi_t\right\rangle+\frac{L}{2}\left\|\xi_{t+1}-\xi_t\right\|_2^2 \\
& =\left\langle\frac{1}{\sqrt{L}}\left(\nabla f\left(\xi_t\right)-\nabla f\left(\theta_t\right)\right), \sqrt{L}\left(\xi_{t+1}-\xi_t\right)\right\rangle+\left\langle\nabla f\left(\theta_t\right), \xi_{t+1}-\xi_t\right\rangle+\frac{L}{2}\left\|\xi_{t+1}-\xi_t\right\|_2^2 \\
& \leq \frac{1}{2}\left(\frac{1}{L}\left\|\nabla f\left(\xi_t\right)-\nabla f\left(\theta_t\right)\right\|_2^2+L\left\|\xi_{t+1}-\xi_t\right\|_2^2\right)+\left\langle\nabla f\left(\theta_t\right), \xi_{t+1}-\xi_t\right\rangle+\frac{L}{2}\left\|\xi_{t+1}-\xi_t\right\|_2^2 \\
& \leq \frac{1}{2 L}\left\|\nabla f\left(\xi_t\right)-\nabla f\left(\theta_t\right)\right\|_2^2+L\left\|\xi_{t+1}-\xi_t\right\|_2^2+\left\langle\nabla f\left(\theta_t\right), \xi_{t+1}-\xi_t\right\rangle \\
& \leq \frac{1}{2 L} L^2\left\|\xi_t-\theta_t\right\|_2^2+L\left\|\xi_{t+1}-\xi_t\right\|_2^2+\left\langle\nabla f\left(\theta_t\right), \xi_{t+1}-\xi_t\right\rangle \\
& =\frac{L}{2} \underbrace{\left\|\xi_t-\theta_t\right\|_2^2}_{(1)}+L \underbrace{\left\|\xi_{t+1}-\xi_t\right\|_2^2}_{(2)}+\underbrace{\left\langle\nabla f\left(\theta_t\right), \xi_{t+1}-\xi_t\right\rangle}_{(3)}
\end{aligned}
\tag{104}
$$

■

Theorem D.5. *Consider a non-convex optimization problem. Suppose Assumption D.1 are satisfied, and let $\alpha_t = \alpha/\sqrt{t}$. For all $T \geq 1$, SGDF achieves the following guarantee:*

$$
\mathbb{E}(T) \leq \frac{C_7 \alpha^2(\log T+1)+C_8}{\alpha \sqrt{T}} \tag{105}
$$

where $\mathbb{E}(T)=\min_{t=1,2,\ldots,T} \mathbb{E}_{t-1}\left[\left\|\nabla f\left(\theta_t\right)\right\|_2^2\right]$ denotes the minimum of the squared-paradigm expectation of the gradient, α is the learning rate at the 1-th step, C_7 are constants independent of d and T, C_8 is a constant independent of T, and the expectation is taken w.r.t all randomness corresponding to g_t.

Proof. According to Lemma D.4, we deal with the three terms (1), (2), and (3) separately.

Bounding Term (1): When $t=1$, $\left\|\xi_t-\theta_t\right\|_2^2=0$

When $t \geq 2$,

$$
\begin{aligned}
\left\|\xi_t-\theta_t\right\|_2^2 & =\left\|\frac{\beta_1}{1-\beta_1}\left(\theta_t-\theta_{t-1}\right)\right\|_2^2 \\
& =\frac{\beta_1^2}{\left(1-\beta_1\right)^2} \alpha_{t-1}^2\left\|\hat{g}_{t-1}\right\|_2^2 \\
& =\frac{\beta_1^2}{\left(1-\beta_1\right)^2} \alpha_{t-1}^2 \sum_{i=1}^d\left(\left(1-K_{t-1, i}\right)\left(\widehat{m}_{t-1, i}\right)^2+K_{t-1, i} g_{t-1, i}^2\right) \\
& \stackrel{(a)}{\leq} \frac{\beta_1^2}{\left(1-\beta_1\right)^2} \alpha_{t-1}^2 \sum_{i=1}^d G_i^2
\end{aligned}
\tag{106}
$$

Where (a) holds because for any t:

1. $\left|\widehat{m}_{t, i}\right| \leq \frac{1}{1-\beta_1^t} \sum_{s=1}^t\left(1-\beta_1\right) \beta_1^{t-s}\left|g_{s, i}\right| \leq \frac{1}{1-\beta_1^t} \sum_{s=1}^t\left(1-\beta_1\right) \beta_1^{t-s} G_i=G_i$.
2. $\left\|g_t\right\|_2 \leq G, \forall t$, or for any dimension of the variable i: $\left\|g_{t, i}\right\|_2 \leq G_i, \forall t$

Bounding Term (2): For the initial iteration $t = 1$, we have:

$$\begin{aligned}
\xi_2 - \xi_1 =& \theta_2 + \frac{\beta_1}{1-\beta_1}(\theta_2 - \theta_1) - \theta_1 \\
=& \frac{1}{1-\beta_1}(\theta_2 - \theta_1) \\
=& -\frac{\alpha_1}{1-\beta_1}(\hat{g}_1) \\
=& -\frac{\alpha_1}{1-\beta_1}\left(\frac{1-K_1}{1-\beta_1}m_1 + K_1 g_1\right) \\
=& -\frac{\alpha_1}{1-\beta_1}\frac{1-K_1}{1-\beta_1}\left(\beta_1 \cancelto{0}{m_0} + (1-\beta_1)g_1\right) - \frac{\alpha_1}{1-\beta_1}K_1 g_1 \\
=& -\frac{\alpha_1(1-K_1)}{1-\beta_1}g_1 - \frac{\alpha_1 K_1}{1-\beta_1}g_1 \\
=& -\frac{\alpha_1}{1-\beta_1}g_1
\end{aligned} \tag{107}$$

Consequently, the squared ℓ_2-norm can be bounded as follows:

$$\begin{aligned}
\|\xi_2 - \xi_1\|_2^2 =& \left\| -\frac{\alpha_1}{1-\beta_1}g_1 \right\|_2^2 \\
=& \frac{\alpha_1^2}{(1-\beta_1)^2}\|g_1\|_2^2 \\
=& \frac{\alpha_1^2}{(1-\beta_1)^2}\sum_{i=1}^{d} g_{1,i}^2 \\
\leq& \frac{\alpha_1^2}{(1-\beta_1)^2}\sum_{i=1}^{d} G_i^2
\end{aligned} \tag{108}$$

For subsequent iterations $t \geq 2$, the difference between consecutive auxiliary variables expands as:

$$\begin{aligned}
\xi_{t+1} - \xi_t =& \theta_{t+1} + \frac{\beta_1}{1-\beta_1}(\theta_{t+1} - \theta_t) - \theta_t - \frac{\beta_1}{1-\beta_1}(\theta_t - \theta_{t-1}) \\
=& \frac{1}{1-\beta_1}(\theta_{t+1} - \theta_t) - \frac{\beta_1}{1-\beta_1}(\theta_t - \theta_{t-1})
\end{aligned} \tag{109}$$

Recalling the parameter update rule, the difference $\theta_{t+1} - \theta_t$ is given by:

$$\begin{aligned}
\theta_{t+1} - \theta_t =& -\alpha_t \hat{g}_t \\
=& -\frac{\alpha_t(1-K_t)}{1-\beta_1^t}m_t - \alpha_t K_t g_t \\
=& -\frac{\alpha_t(1-K_t)}{1-\beta_1^t}(\beta_1 m_{t-1} + (1-\beta_1)g_t) - \alpha_t K_t g_t
\end{aligned} \tag{110}$$

Substituting this expression back into the expansion of $\xi_{t+1} - \xi_t$ and rearranging the terms, we obtain:

$$\begin{aligned}
&\xi_{t+1} - \xi_t \\
=&\frac{1}{1-\beta_1}\left(-\frac{\alpha_t(1-K_t)}{1-\beta_1^t}\left(\beta_1 m_{t-1} + (1-\beta_1)\, g_t\right) - \alpha_t K_t g_t\right) \\
&- \frac{\beta_1}{1-\beta_1}\left(-\frac{\alpha_{t-1}(1-K_{t-1})}{1-\beta_1^{t-1}} m_{t-1} - \alpha_{t-1} K_{t-1} g_{t-1}\right) \\
=&-\frac{\beta_1}{1-\beta_1} m_{t-1} \odot \left(\frac{\alpha_t(1-K_t)}{1-\beta_1^t} - \frac{\alpha_{t-1}(1-K_{t-1})}{1-\beta_1^{t-1}}\right) \\
&-\left(\frac{\alpha_t(1-K_t)}{1-\beta_1^t} + \frac{\alpha_t K_t}{1-\beta_1}\right) g_t + \frac{\beta_1 \alpha_{t-1} K_{t-1}}{1-\beta_1} g_{t-1}
\end{aligned} \tag{111}$$

Using the general inequality $\|A+B+C\|_2^2 \le 3\|A\|_2^2 + 3\|B\|_2^2 + 3\|C\|_2^2$, we have:

$$\begin{aligned}
\|\xi_{t+1} - \xi_t\|_2^2 \le\ & 3\left\|-\frac{\beta_1}{1-\beta_1} m_{t-1} \odot \left(\frac{\alpha_t(1-K_t)}{1-\beta_1^t} - \frac{\alpha_{t-1}(1-K_{t-1})}{1-\beta_1^{t-1}}\right)\right\|_2^2 \\
&+ 3\left\|-\left(\frac{\alpha_t(1-K_t)}{1-\beta_1^t} + \frac{\alpha_t K_t}{1-\beta_1}\right) g_t\right\|_2^2 + 3\left\|\frac{\beta_1 \alpha_{t-1} K_{t-1}}{1-\beta_1} g_{t-1}\right\|_2^2
\end{aligned} \tag{112}$$

To properly decouple the step size decay from the dynamic mask flipping, let $\eta_t = \frac{\alpha_t}{1-\beta_1^t}$. We can decompose the mask difference algebraically by adding and subtracting $\eta_{t-1} K_t$:

$$\begin{aligned}
\eta_t(1-K_t) - \eta_{t-1}(1-K_{t-1}) &= \eta_t - \eta_t K_t - \eta_{t-1} + \eta_{t-1} K_{t-1} \\
&= (\eta_t - \eta_{t-1}) - \eta_t K_t + \eta_{t-1} K_t - \eta_{t-1} K_t + \eta_{t-1} K_{t-1} \\
&= (\eta_t - \eta_{t-1})(1-K_t) + \eta_{t-1}(K_{t-1} - K_t)
\end{aligned} \tag{113}$$

Using the inequality $\|A+B\|_2^2 \le 2\|A\|_2^2 + 2\|B\|_2^2$, we bound the squared ℓ_2 norm of this difference. Since $K_{t,i} \in \{0,1\}$, we have $(1-K_{t,i})^2 \le 1$ and $(K_{t-1,i} - K_{t,i})^2 = |K_{t-1,i} - K_{t,i}|$:

$$\begin{aligned}
\left\|\frac{\alpha_t(1-K_t)}{1-\beta_1^t} - \frac{\alpha_{t-1}(1-K_{t-1})}{1-\beta_1^{t-1}}\right\|_2^2 &= \sum_{i=1}^{d} \left((\eta_t - \eta_{t-1})(1-K_{t,i}) + \eta_{t-1}(K_{t-1,i} - K_{t,i})\right)^2 \\
&\le 2\sum_{i=1}^{d} (\eta_{t-1} - \eta_t)^2 (1-K_{t,i})^2 + 2\sum_{i=1}^{d} \eta_{t-1}^2 (K_{t-1,i} - K_{t,i})^2 \\
&\le 2\sum_{i=1}^{d} (\eta_{t-1} - \eta_t)^2 + 2\eta_{t-1}^2 \sum_{i=1}^{d} |K_{t-1,i} - K_{t,i}|
\end{aligned} \tag{114}$$

Since η_t is monotonically decreasing, $\eta_{t-1} - \eta_t \ge 0$, and thus $(\eta_{t-1} - \eta_t)^2 \le \eta_1(\eta_{t-1} - \eta_t)$. Let $F_t = \sum_{i=1}^{d} |K_{t-1,i} - K_{t,i}|$ denote the total number of flipped mask bits at step t. To maintain a tight bound, we retain F_t:

$$\begin{aligned}
\left\|\frac{\alpha_t(1-K_t)}{1-\beta_1^t} - \frac{\alpha_{t-1}(1-K_{t-1})}{1-\beta_1^{t-1}}\right\|_2^2 &\le 2d\eta_1(\eta_{t-1} - \eta_t) + 2\eta_{t-1}^2 F_t \\
&\le 2d\eta_1\left(\frac{\alpha_{t-1}}{1-\beta_1^{t-1}} - \frac{\alpha_t}{1-\beta_1^t}\right) + 2F_t\left(\frac{\alpha_{t-1}}{1-\beta_1^{t-1}}\right)^2
\end{aligned} \tag{115}$$

Now, bounding the three terms of $\|\xi_{t+1}-\xi_t\|_2^2$ respectively. For the first term, since $|m_{t-1,i}| \le G_i$:

$$\begin{aligned}
&3\left\|\frac{\beta_1}{1-\beta_1}m_{t-1}\odot\left(\frac{\alpha_t(1-K_t)}{1-\beta_1^t}-\frac{\alpha_{t-1}(1-K_{t-1})}{1-\beta_1^{t-1}}\right)\right\|_2^2 \\
&\le 3\frac{\beta_1^2}{(1-\beta_1)^2}\left(\max_i G_i\right)^2\left(2d\eta_1\left(\frac{\alpha_{t-1}}{1-\beta_1^{t-1}}-\frac{\alpha_t}{1-\beta_1^t}\right)+2F_t\left(\frac{\alpha_{t-1}}{1-\beta_1^{t-1}}\right)^2\right) \\
&= 6d\frac{\beta_1^2\eta_1}{(1-\beta_1)^2}\left(\max_i G_i\right)^2\left(\frac{\alpha_{t-1}}{1-\beta_1^{t-1}}-\frac{\alpha_t}{1-\beta_1^t}\right)+6F_t\frac{\beta_1^2}{(1-\beta_1)^2}\left(\max_i G_i\right)^2\left(\frac{\alpha_{t-1}}{1-\beta_1^{t-1}}\right)^2
\end{aligned} \tag{116}$$

For the second term, notice that $1-\beta_1 \le 1-\beta_1^t$, which implies $\frac{1}{1-\beta_1} \ge \frac{1}{1-\beta_1^t}$. Since the dimensions masked by K_t and $1-K_t$ are completely disjoint, we safely bound the disjoint coefficients using the globally larger denominator $\frac{1}{1-\beta_1}$:

$$\begin{aligned}
3\left\|\left(\frac{\alpha_t(1-K_t)}{1-\beta_1^t}+\frac{\alpha_t K_t}{1-\beta_1}\right)g_t\right\|_2^2 &= 3\sum_{i=1}^d\left(\frac{\alpha_t(1-K_{t,i})}{1-\beta_1^t}+\frac{\alpha_t K_{t,i}}{1-\beta_1}\right)^2 g_{t,i}^2 \\
&\le 3\left(\frac{\alpha_t}{1-\beta_1}\right)^2\sum_{i=1}^d G_i^2
\end{aligned} \tag{117}$$

For the third term, we maintain the exact β_1 constant multiplier to prevent invalid bounding:

$$3\left\|\frac{\beta_1\alpha_{t-1}K_{t-1}}{1-\beta_1}g_{t-1}\right\|_2^2 \le 3\frac{\beta_1^2\alpha_{t-1}^2}{(1-\beta_1)^2}\sum_{i=1}^d G_i^2 \tag{118}$$

Therefore, bringing everything together, we obtain the corrected upper bound:

$$\begin{aligned}
\|\xi_{t+1}-\xi_t\|_2^2 \le\ & 6d\frac{\beta_1^2\eta_1}{(1-\beta_1)^2}\left(\max_i G_i\right)^2\left(\frac{\alpha_{t-1}}{1-\beta_1^{t-1}}-\frac{\alpha_t}{1-\beta_1^t}\right) \\
&+6F_t\frac{\beta_1^2}{(1-\beta_1)^2}\left(\max_i G_i\right)^2\left(\frac{\alpha_{t-1}}{1-\beta_1^{t-1}}\right)^2 \\
&+3\left(\frac{\alpha_t}{1-\beta_1}\right)^2\sum_{i=1}^d G_i^2+3\frac{\beta_1^2\alpha_{t-1}^2}{(1-\beta_1)^2}\sum_{i=1}^d G_i^2
\end{aligned} \tag{119}$$

Bounding Term (3): When $t=1$, referring to the case of $t=1$ in the previous subsection, we expand the inner product:

$$\begin{aligned}
\langle\nabla f(\theta_1), \xi_2-\xi_1\rangle &= \left\langle\nabla f(\theta_1), -\frac{\alpha_1}{1-\beta_1}g_1\right\rangle \\
&= \left\langle\nabla f(\theta_1), -\frac{\alpha_1}{1-\beta_1}\nabla f(\theta_1)\right\rangle+\left\langle\nabla f(\theta_1), -\frac{\alpha_1}{1-\beta_1}\zeta_1\right\rangle \\
&= -\frac{\alpha_1}{1-\beta_1}\|\nabla f(\theta_1)\|_2^2+\left\langle\nabla f(\theta_1), -\frac{\alpha_1}{1-\beta_1}\zeta_1\right\rangle
\end{aligned} \tag{120}$$

Note that we leave the inner product with the zero-mean noise ζ_1 exactly as it is, because it will vanish when taking the conditional expectation later.

When $t \ge 2$,

$$\begin{aligned}
&\langle\nabla f(\theta_t), \xi_{t+1}-\xi_t\rangle \\
=&\left\langle\nabla f(\theta_t), -\frac{\beta_1}{1-\beta_1}m_{t-1}\odot\left(\frac{\alpha_t(1-K_t)}{1-\beta_1^t}-\frac{\alpha_{t-1}(1-K_{t-1})}{1-\beta_1^{t-1}}\right)\right\rangle \\
&+\left\langle\nabla f(\theta_t), -\left(\frac{\alpha_t(1-K_t)}{1-\beta_1^t}+\frac{\alpha_t K_t}{1-\beta_1}\right)\nabla f(\theta_t)\right\rangle+\left\langle\nabla f(\theta_t), -\left(\frac{\alpha_t(1-K_t)}{1-\beta_1^t}+\frac{\alpha_t K_t}{1-\beta_1}\right)\zeta_t\right\rangle \\
&+\left\langle\nabla f(\theta_t), \frac{\beta_1\alpha_{t-1}K_{t-1}}{1-\beta_1}\nabla f(\theta_{t-1})\right\rangle+\left\langle\nabla f(\theta_t), \frac{\beta_1\alpha_{t-1}K_{t-1}}{1-\beta_1}\zeta_{t-1}\right\rangle
\end{aligned} \tag{121}$$

Let us deal with the terms after the equal sign separately.

Start by looking at the first term. We decouple the step size decay and dynamic mask flipping using the identity $\eta_t(1-K_t)-\eta_{t-1}(1-K_{t-1})=(\eta_t-\eta_{t-1})(1-K_t)+\eta_{t-1}(K_{t-1}-K_t)$ and Hölder's inequality ($|\langle A,B\rangle|\le\|A\|_\infty\|B\|_1$):

$$\begin{aligned}
&\left\langle \nabla f\left(\theta_t\right), -\frac{\beta_1}{1-\beta_1} m_{t-1} \odot \left(\frac{\alpha_t(1-K_t)}{1-\beta_1^t}-\frac{\alpha_{t-1}(1-K_{t-1})}{1-\beta_1^{t-1}}\right)\right\rangle \\
\le & \frac{\beta_1}{1-\beta_1}\left\|\nabla f\left(\theta_t\right)\right\|_\infty\left\|m_{t-1}\right\|_\infty\left\|\frac{\alpha_t(1-K_t)}{1-\beta_1^t}-\frac{\alpha_{t-1}(1-K_{t-1})}{1-\beta_1^{t-1}}\right\|_1 \\
\le & \frac{\beta_1}{1-\beta_1}\left(\max_i G_i\right)^2\left(\sum_{i=1}^d\left(\frac{\alpha_{t-1}}{1-\beta_1^{t-1}}-\frac{\alpha_t}{1-\beta_1^t}\right)\left(1-K_{t,i}\right)+\sum_{i=1}^d \frac{\alpha_{t-1}}{1-\beta_1^{t-1}}\left|K_{t-1,i}-K_{t,i}\right|\right) \\
\le & \frac{\beta_1}{1-\beta_1}\left(\max_i G_i\right)^2 \cdot d\left(\frac{\alpha_{t-1}}{1-\beta_1^{t-1}}-\frac{\alpha_t}{1-\beta_1^t}\right)+\frac{\beta_1}{1-\beta_1}\left(\max_i G_i\right)^2 \frac{\alpha_{t-1}}{1-\beta_1^{t-1}} F_t
\end{aligned} \tag{122}$$

where $F_t=\sum_{i=1}^d\left|K_{t-1,i}-K_{t,i}\right|$ represents the number of indices where the mask changes at step t.

For the second and third terms, notice that since $1-\beta_1^t\ge 1-\beta_1$, we have $\frac{\alpha_t(1-K_t)}{1-\beta_1^t}+\frac{\alpha_t K_t}{1-\beta_1}\ge\frac{\alpha_t}{1-\beta_1^t}(1-K_t+K_t)=\frac{\alpha_t}{1-\beta_1^t}$:

$$\begin{aligned}
&\left\langle \nabla f\left(\theta_t\right), -\left(\frac{\alpha_t(1-K_t)}{1-\beta_1^t}+\frac{\alpha_t K_t}{1-\beta_1}\right)\nabla f\left(\theta_t\right)\right\rangle+\left\langle \nabla f\left(\theta_t\right), -\left(\frac{\alpha_t(1-K_t)}{1-\beta_1^t}+\frac{\alpha_t K_t}{1-\beta_1}\right)\zeta_t\right\rangle \\
\le & -\frac{\alpha_t}{1-\beta_1^t}\left\|\nabla f\left(\theta_t\right)\right\|_2^2+\left\langle \nabla f\left(\theta_t\right), -\left(\frac{\alpha_t(1-K_t)}{1-\beta_1^t}+\frac{\alpha_t K_t}{1-\beta_1}\right)\zeta_t\right\rangle
\end{aligned} \tag{123}$$

For the fourth term (the cross-gradient term), applying Hölder's inequality directly would yield an un-decaying $\mathcal{O}(\alpha_{t-1})$ penalty. Instead, we use the basic inequality $2\langle a,b\rangle\le\|a\|_2^2+\|b\|_2^2$. Since $K_{t-1,i}\in\{0,1\}$, we have:

$$\begin{aligned}
\left\langle \nabla f\left(\theta_t\right), \frac{\beta_1\alpha_{t-1}K_{t-1}}{1-\beta_1}\nabla f\left(\theta_{t-1}\right)\right\rangle &= \frac{\beta_1\alpha_{t-1}}{1-\beta_1}\sum_{i=1}^d K_{t-1,i}\nabla f(\theta_t)_i\nabla f(\theta_{t-1})_i \\
&\le \frac{\beta_1\alpha_{t-1}}{2(1-\beta_1)}\sum_{i=1}^d K_{t-1,i}\left(\nabla f(\theta_t)_i^2+\nabla f(\theta_{t-1})_i^2\right) \\
&\le \frac{\beta_1\alpha_{t-1}}{2(1-\beta_1)}\left(\left\|\nabla f\left(\theta_t\right)\right\|_2^2+\left\|\nabla f\left(\theta_{t-1}\right)\right\|_2^2\right)
\end{aligned} \tag{124}$$

For the fifth term (the noise term involving ζ_{t-1}), taking the absolute value would similarly result in an $\mathcal{O}(\alpha_{t-1})$ error bound. To properly bound it, we decompose the inner product to leverage the expectation later:

$$\begin{aligned}
\left\langle \nabla f\left(\theta_t\right), \frac{\beta_1\alpha_{t-1}K_{t-1}}{1-\beta_1}\zeta_{t-1}\right\rangle &= \frac{\beta_1\alpha_{t-1}}{1-\beta_1}\left\langle \nabla f(\theta_{t-1}), K_{t-1}\odot\zeta_{t-1}\right\rangle \\
&\quad+\frac{\beta_1\alpha_{t-1}}{1-\beta_1}\left\langle \nabla f(\theta_t)-\nabla f(\theta_{t-1}), K_{t-1}\odot\zeta_{t-1}\right\rangle
\end{aligned} \tag{125}$$

Notice that when taking the expectation, the first part satisfies $\mathbb{E}_{t-2}\left[\langle\nabla f(\theta_{t-1}), K_{t-1}\odot\zeta_{t-1}\rangle\right]=0$ because both θ_{t-1} and K_{t-1} are independent of the noise realization ζ_{t-1}. For the second part, we apply the Cauchy-Schwarz inequality and utilize the L-smoothness of the objective function. Given that $\|\theta_t-\theta_{t-1}\|_2\le\alpha_{t-1}\|\hat{g}_{t-1}\|_2\le\alpha_{t-1}\sqrt{\sum_{i=1}^d G_i^2}$ and

$\|\zeta_{t-1}\|_2 \leq 2\sqrt{\sum_{i=1}^{d} G_i^2}$, we can bound this term strictly by $\mathcal{O}(\alpha_{t-1}^2)$:

$$\begin{aligned}
\frac{\beta_1\alpha_{t-1}}{1-\beta_1}\left\langle \nabla f(\theta_t) - \nabla f(\theta_{t-1}), K_{t-1} \odot \zeta_{t-1} \right\rangle &\leq \frac{\beta_1\alpha_{t-1}}{1-\beta_1}\left\| \nabla f(\theta_t) - \nabla f(\theta_{t-1}) \right\|_2 \left\| \zeta_{t-1} \right\|_2 \\
&\leq \frac{\beta_1\alpha_{t-1}}{1-\beta_1} L \left\| \theta_t - \theta_{t-1} \right\|_2 \left\| \zeta_{t-1} \right\|_2 \\
&\leq \frac{\beta_1\alpha_{t-1}}{1-\beta_1} L \left(\alpha_{t-1}\sqrt{\sum_{i=1}^{d} G_i^2} \right) \left(2\sqrt{\sum_{i=1}^{d} G_i^2} \right) \\
&= \frac{2L\beta_1}{1-\beta_1}\alpha_{t-1}^2 \sum_{i=1}^{d} G_i^2
\end{aligned} \tag{126}$$

Finally, combining everything together:

$$\begin{aligned}
& \left\langle \nabla f\left(\theta_t\right), \xi_{t+1} - \xi_t \right\rangle \\
\leq\ & \frac{\beta_1 d}{1-\beta_1}\left(\max_i G_i\right)^2 \left(\frac{\alpha_{t-1}}{1-\beta_1^{t-1}} - \frac{\alpha_t}{1-\beta_1^t} \right) + \frac{\beta_1}{1-\beta_1}\left(\max_i G_i\right)^2 \frac{\alpha_{t-1}}{1-\beta_1^{t-1}} F_t \\
& - \left(\frac{\alpha_t}{1-\beta_1^t} - \frac{\beta_1\alpha_{t-1}}{2(1-\beta_1)} \right) \left\| \nabla f\left(\theta_t\right) \right\|_2^2 + \frac{\beta_1\alpha_{t-1}}{2(1-\beta_1)} \left\| \nabla f(\theta_{t-1}) \right\|_2^2 \\
& + \left\langle \nabla f\left(\theta_t\right), -\left(\frac{\alpha_t(1-K_t)}{1-\beta_1^t} + \frac{\alpha_t K_t}{1-\beta_1} \right) \zeta_t \right\rangle \\
& + \frac{\beta_1\alpha_{t-1}}{1-\beta_1} \left\langle \nabla f(\theta_{t-1}), K_{t-1} \odot \zeta_{t-1} \right\rangle + \frac{2L\beta_1}{1-\beta_1}\alpha_{t-1}^2 \sum_{i=1}^{d} G_i^2
\end{aligned} \tag{127}$$

Summarizing the results Since we have already rigorously bounded the cross-inner product terms and the noise components in the previous subsection, we can proceed directly to summarizing the results.

First, taking the expectation $\mathbb{E}_t$ over the random distribution of $\zeta_1, \zeta_2, \ldots, \zeta_t$ on both sides of the inequality. Since the value of θ_t is independent of g_t, they are statistically independent of ζ_t, and $\mathbb{E}_t\left[\zeta_t\right] = 0$. Therefore, all inner products with ζ_t perfectly vanish:

$$\begin{aligned}
& \mathbb{E}_t\left[\left\langle \nabla f\left(\theta_t\right), -\left(\frac{\alpha_t(1-K_t)}{1-\beta_1^t} + \frac{\alpha_t K_t}{1-\beta_1} \right) \zeta_t \right\rangle \right] \\
=& \left\langle -\left(\frac{\alpha_t(1-K_t)}{1-\beta_1^t} + \frac{\alpha_t K_t}{1-\beta_1} \right) \nabla f\left(\theta_t\right), \cancelto{0}{\mathbb{E}_t\left[\zeta_t\right]} \right\rangle = 0
\end{aligned} \tag{128}$$

Combining the bounds from Term (1), Term (2), and the refined Term (3), for $t \geq 2$, we have:

$$\begin{aligned}
& \mathbb{E}_t\left[f\left(\xi_{t+1}\right) - f\left(\xi_t\right) \right] \\
\leq & \frac{L\beta_1^2}{2(1-\beta_1)^2}\alpha_{t-1}^2 \sum_{i=1}^{d} G_i^2 + 6dL\frac{\beta_1^2\eta_1}{(1-\beta_1)^2}(\max_i G_i)^2 \left(\frac{\alpha_{t-1}}{1-\beta_1^{t-1}} - \frac{\alpha_t}{1-\beta_1^t} \right) \\
& + \left(6F_t\frac{\beta_1^2}{(1-\beta_1)^2}(\max_i G_i)^2 + 3L\sum_{i=1}^{d} G_i^2 \right) \left(\frac{\alpha_{t-1}}{1-\beta_1^{t-1}} \right)^2 + 3L\left(\frac{\alpha_t}{1-\beta_1} \right)^2 \sum_{i=1}^{d} G_i^2 \\
& + \frac{\beta_1 d}{1-\beta_1}(\max_i G_i)^2 \left(\frac{\alpha_{t-1}}{1-\beta_1^{t-1}} - \frac{\alpha_t}{1-\beta_1^t} \right) + \frac{\beta_1}{(1-\beta_1)^2}(\max_i G_i)^2 \left(\frac{\alpha_{t-1}}{1-\beta_1^{t-1}} \right) F_t \\
& - \left(\frac{\alpha_t}{1-\beta_1^t} - \frac{\beta_1\alpha_{t-1}}{2(1-\beta_1)} \right) \mathbb{E}_t \left\| \nabla f(\theta_t) \right\|_2^2 + \frac{\beta_1\alpha_{t-1}}{2(1-\beta_1)} \mathbb{E}_{t-1} \left\| \nabla f(\theta_{t-1}) \right\|_2^2 \\
& + \frac{2L\beta_1}{1-\beta_1}\alpha_{t-1}^2 \sum_{i=1}^{d} G_i^2
\end{aligned} \tag{129}$$

To maintain the inequality and simplify the notation, we define the following iteration-independent constants. Note that $\frac{1}{1-\beta_1^t} \leq \frac{1}{1-\beta_1}$, and we absorb the new noise bound into C_1:

1. For α_{t-1}^2: $C_1 \triangleq \frac{L\beta_1^2}{2(1-\beta_1)^2}\sum_{i=1}^d G_i^2 + \frac{3L}{(1-\beta_1)^2}\sum_{i=1}^d G_i^2 + \frac{2L\beta_1}{1-\beta_1}\sum_{i=1}^d G_i^2$
2. For α_t^2: $C_2 \triangleq \frac{3L}{(1-\beta_1)^2}\sum_{i=1}^d G_i^2$
3. For the telescoping difference: $C_3 \triangleq \frac{6dL\beta_1^2\alpha_1}{(1-\beta_1)^3}(\max_i G_i)^2 + \frac{\beta_1 d}{1-\beta_1}(\max_i G_i)^2$
4. For the dynamic mask flipping terms (both linear and squared): $C_4 \triangleq \frac{\beta_1}{(1-\beta_1)^2}(\max_i G_i)^2 + \frac{6\beta_1^2}{(1-\beta_1)^4}(\max_i G_i)^2\alpha_1$

For $t = 1$:

$$\mathbb{E}_1\left[f(\xi_2) - f(\xi_1)\right] \leq \frac{L}{(1-\beta_1)^2}\alpha_1^2\sum_{i=1}^d G_i^2 - \frac{\alpha_1}{1-\beta_1}\mathbb{E}_1\left\|\nabla f(\theta_1)\right\|_2^2 \tag{130}$$

Summing up both sides of the inequality for $t = 1, 2, \ldots, T$:

Left side of the inequality (LHS)

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}_t\left[f\left(\xi_{t+1}\right) - f\left(\xi_t\right)\right] =& \mathbb{E}_T\left[f\left(\xi_{T+1}\right)\right] - \mathbb{E}_0\left[f\left(\xi_1\right)\right] \\ \geq& f\left(\theta^*\right) - f\left(\theta_1\right)\end{aligned} \tag{131}$$

Right side of the inequality (RHS)

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}_t[\text{RHS}] \leq& \sum_{t=2}^T C_1\alpha_{t-1}^2 + \sum_{t=2}^T C_2\alpha_t^2 + \sum_{t=2}^T C_3\left(\frac{\alpha_{t-1}}{1-\beta_1^{t-1}} - \frac{\alpha_t}{1-\beta_1^t}\right) + \sum_{t=2}^T C_4\frac{\alpha_{t-1}}{1-\beta_1^{t-1}}F_t \\ &+ \sum_{t=2}^T \frac{\beta_1\alpha_{t-1}}{2(1-\beta_1)}\mathbb{E}_{t-1}\left\|\nabla f(\theta_{t-1})\right\|_2^2 - \sum_{t=2}^T\left(\frac{\alpha_t}{1-\beta_1^t} - \frac{\beta_1\alpha_{t-1}}{2(1-\beta_1)}\right)\mathbb{E}_t\left\|\nabla f(\theta_t)\right\|_2^2 \\ &+ \frac{L}{(1-\beta_1)^2}\alpha_1^2\sum_{i=1}^d G_i^2 - \frac{\alpha_1}{1-\beta_1}\mathbb{E}_1\left\|\nabla f(\theta_1)\right\|_2^2\end{aligned} \tag{132}$$

We rigorously shift the index of the positive gradient expectation term to combine it with the negative counterpart. By shifting $k = t - 1$:

$$\begin{aligned}&\sum_{t=2}^T \frac{\beta_1\alpha_{t-1}}{2(1-\beta_1)}\mathbb{E}_{t-1}\left\|\nabla f(\theta_{t-1})\right\|_2^2 - \sum_{t=2}^T\left(\frac{\alpha_t}{1-\beta_1^t} - \frac{\beta_1\alpha_{t-1}}{2(1-\beta_1)}\right)\mathbb{E}_t\left\|\nabla f(\theta_t)\right\|_2^2 \\ =& \frac{\beta_1\alpha_1}{2(1-\beta_1)}\mathbb{E}_1\left\|\nabla f(\theta_1)\right\|_2^2 + \sum_{t=2}^{T-1}\frac{\beta_1\alpha_t}{2(1-\beta_1)}\mathbb{E}_t\left\|\nabla f(\theta_t)\right\|_2^2 \\ &- \sum_{t=2}^{T-1}\left(\frac{\alpha_t}{1-\beta_1^t} - \frac{\beta_1\alpha_{t-1}}{2(1-\beta_1)}\right)\mathbb{E}_t\left\|\nabla f(\theta_t)\right\|_2^2 - \left(\frac{\alpha_T}{1-\beta_1^T} - \frac{\beta_1\alpha_{T-1}}{2(1-\beta_1)}\right)\mathbb{E}_T\left\|\nabla f(\theta_T)\right\|_2^2 \\ \leq& \frac{\beta_1\alpha_1}{2(1-\beta_1)}\mathbb{E}_1\left\|\nabla f(\theta_1)\right\|_2^2 - \sum_{t=2}^T\left(\frac{\alpha_t}{1-\beta_1^t} - \frac{\beta_1(\alpha_{t-1}+\alpha_t)}{2(1-\beta_1)}\right)\mathbb{E}_t\left\|\nabla f(\theta_t)\right\|_2^2\end{aligned} \tag{133}$$

where the inequality holds because we drop the strictly negative term $-\frac{\beta_1\alpha_T}{2(1-\beta_1)}\mathbb{E}_T\|\nabla f(\theta_T)\|_2^2$ to upper-bound the expression.

Now, integrating the $t = 1$ expectation term from the initialization, we observe that since $\beta_1 < 1$, the combined coefficient for $t = 1$ is strictly negative:

$$-\frac{\alpha_1}{1-\beta_1}\mathbb{E}_1\left\|\nabla f(\theta_1)\right\|_2^2 + \frac{\beta_1\alpha_1}{2(1-\beta_1)}\mathbb{E}_1\left\|\nabla f(\theta_1)\right\|_2^2 = -\frac{\alpha_1(2-\beta_1)}{2(1-\beta_1)}\mathbb{E}_1\left\|\nabla f(\theta_1)\right\|_2^2 \leq -\frac{\alpha_1}{2(1-\beta_1)}\mathbb{E}_1\left\|\nabla f(\theta_1)\right\|_2^2 \tag{134}$$

For $t \geq 2$, we assume the effective learning rate is chosen to be small enough such that this coefficient bounds the descent effectively: $\frac{\alpha_t}{1-\beta_1^t} - \frac{\beta_1(\alpha_{t-1}+\alpha_t)}{2(1-\beta_1)} \geq \frac{\alpha_t}{2(1-\beta_1^t)}$. Combining this with the bounds for the telescoping difference sum ($\leq \frac{\alpha_1}{1-\beta_1}$)

and the mask flipping bounded by S_{total}, we have:

$$f(\theta^*) - f(\theta_1) \leq (C_1 + C_2) \sum_{t=1}^{T} \alpha_t^2 + C_3 \frac{\alpha_1}{1-\beta_1} + C_4 \frac{\alpha_1}{1-\beta_1} S_{\text{total}} - \sum_{t=1}^{T} \frac{\alpha_t}{2(1-\beta_1^t)} \mathbb{E}_t \left[\|\nabla f(\theta_t)\|_2^2 \right] \tag{135}$$

Rearranging the terms to isolate the gradient norm:

$$\sum_{t=1}^{T} \frac{\alpha_t}{2(1-\beta_1^t)} \mathbb{E}_t \left[\|\nabla f(\theta_t)\|_2^2 \right] \leq (C_1 + C_2) \sum_{t=1}^{T} \alpha_t^2 + f(\theta_1) - f(\theta^*) + C_3 \frac{\alpha_1}{1-\beta_1} + C_4 \frac{\alpha_1}{1-\beta_1} S_{\text{total}} \tag{136}$$

Let $C_7 \triangleq 2(C_1 + C_2)$ and let C_8 absorb all the non-decaying initial constants, $C_8 \triangleq 2\left(f(\theta_1) - f(\theta^*) + C_3 \frac{\alpha_1}{1-\beta_1} + C_4 \frac{\alpha_1}{1-\beta_1} S_{\text{total}} \right)$. Note that since $1 - \beta_1^t \leq 1$, we universally have $\frac{1}{1-\beta_1^t} \geq 1$. Therefore, multiplying both sides by 2 yields:

$$\sum_{t=1}^{T} \alpha_t \mathbb{E}_t \left[\|\nabla f(\theta_t)\|_2^2 \right] \leq \sum_{t=1}^{T} \frac{\alpha_t}{1-\beta_1^t} \mathbb{E}_t \left[\|\nabla f(\theta_t)\|_2^2 \right] \leq C_7 \sum_{t=1}^{T} \alpha_t^2 + C_8 \tag{137}$$

Extracting the minimum over the trajectory $\mathbb{E}(T) = \min_{t=1,\ldots,T} \mathbb{E}_{t-1} \left[\|\nabla f(\theta_t)\|_2^2 \right]$:

$$\mathbb{E}(T) \sum_{t=1}^{T} \alpha_t \leq \sum_{t=1}^{T} \alpha_t \mathbb{E}_t \left[\|\nabla f(\theta_t)\|_2^2 \right] \leq C_7 \sum_{t=1}^{T} \alpha_t^2 + C_8 \Longrightarrow \mathbb{E}(T) \leq \frac{C_7 \sum_{t=1}^{T} \alpha_t^2 + C_8}{\sum_{t=1}^{T} \alpha_t} \tag{138}$$

Since $\alpha_t = \alpha/\sqrt{t}$, we apply the standard integral bound for the squared step size summation:

$$\sum_{t=1}^{T} \alpha_t^2 = \alpha^2 \sum_{t=1}^{T} \frac{1}{t} \leq \alpha^2 \left(1 + \int_1^T \frac{1}{x} dx \right) = \alpha^2 (\log T + 1) \tag{139}$$

For the linear step size summation, we can strictly lower bound it by replacing each term with the minimum term in the sequence:

$$\sum_{t=1}^{T} \alpha_t = \alpha \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \geq \alpha \sum_{t=1}^{T} \frac{1}{\sqrt{T}} = \alpha \frac{T}{\sqrt{T}} = \alpha \sqrt{T} \tag{140}$$

Substituting the upper bound of the numerator and the lower bound of the denominator yields the final finite-time convergence bound:

$$\mathbb{E}(T) \leq \frac{C_7 \alpha^2 (\log T + 1) + C_8}{\alpha \sqrt{T}} = \mathcal{O}\left(\frac{\log T}{\sqrt{T}} \right) \tag{141}$$

■

E. Detailed Experimental Supplement

We performed extensive comparisons with other optimizers, including SGD [25], Adam [14], RAdam [21] and AdamW [22], and so on. The experiments include: (a) image classification on CIFAR dataset [15] with VGG [31], ResNet [11] and DenseNet [13], and image recognition with VGG, ResNet, and DenseNet on ImageNet [6].

E.1. Image classification with CNNs on CIFAR

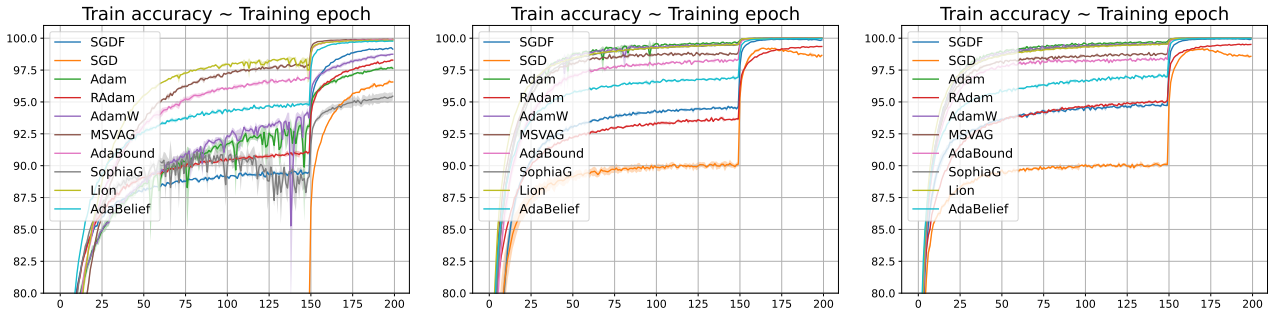
For all experiments, the model is trained for 200 epochs with a batch size of 128, and the learning rate is multiplied by 0.1 at epoch 150. We performed extensive hyperparameter search as described in the main paper. Here, we report both training and test accuracy in Fig. 1 and Fig. 2. Detailed experimental parameters we place in Tab. 1. We summarize the mean best test accuracies and their standard deviations for each algorithm in Tab. 2. The best results are highlighted in bold font. SGDF not only achieves the highest test accuracy but also a smaller gap between training and test accuracy compared with other optimizers. We ran each experiment three times with different seeds $\{0, 1, 2\}$ to ensure the robustness of the results.

Table 1. Hyperparameters used for CIFAR-10 and CIFAR-100 datasets.

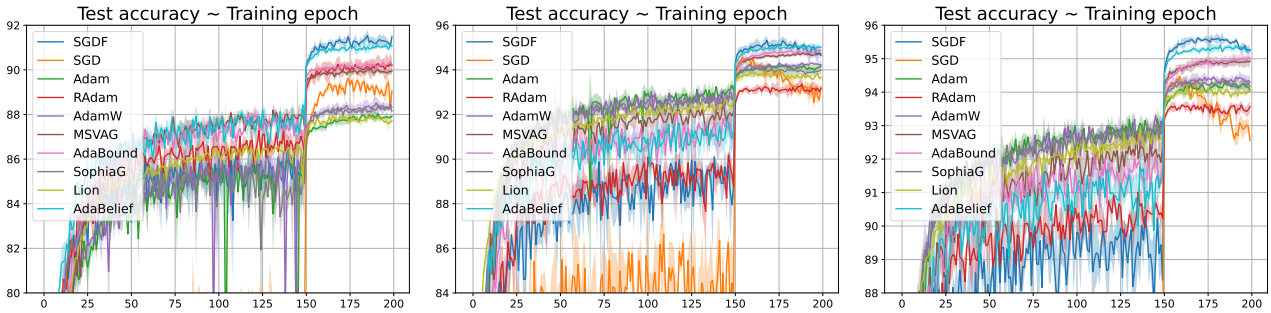
Optimizer	Learning Rate	β_1	β_2	Epochs	Schedule	Weight Decay	Batch Size	ε
SGDF	0.5	0.9	0.999	200	StepLR	0.0005	128	1e-8
SGD	0.1	0.9	-	200	StepLR	0.0005	128	-
Adam	0.001	0.9	0.999	200	StepLR	0.0005	128	1e-8
RAdam	0.001	0.9	0.999	200	StepLR	0.0005	128	1e-8
AdamW	0.001	0.9	0.999	200	StepLR	0.01	128	1e-8
MSVAG	0.1	0.9	0.999	200	StepLR	0.0005	128	1e-8
AdaBound	0.001	0.9	0.999	200	StepLR	0.0005	128	-
Sophia	0.0001	0.965	0.99	200	StepLR	0.1	128	-
Lion	0.00002	0.9	0.99	200	StepLR	0.1	128	-
AdaBound	0.001	0.9	0.999	200	StepLR	0.0005	128	1e-8

Table 2. Test Accuracies for CIFAR-10 and CIFAR-100 across different models and algorithms.

Algorithm	CIFAR-10			CIFAR-100		
	VGG11	ResNet34	DenseNet121	VGG11	ResNet34	DenseNet121
SGDF	**91.61**$_{\pm 0.21}$	**95.33**$_{\pm 0.19}$	**95.74**$_{\pm 0.06}$	**68.12**$_{\pm 0.15}$	**77.69**$_{\pm 0.64}$	**80.17**$_{\pm 0.19}$
SGD	89.83$_{\pm 0.05}$	94.62$_{\pm 0.07}$	94.52$_{\pm 0.03}$	63.48$_{\pm 0.39}$	76.88$_{\pm 0.12}$	78.77$_{\pm 0.27}$
Adam	88.12$_{\pm 0.10}$	94.30$_{\pm 0.06}$	94.37$_{\pm 0.17}$	56.27$_{\pm 0.32}$	72.81$_{\pm 0.45}$	74.67$_{\pm 0.45}$
AdamW	88.59$_{\pm 0.20}$	94.42$_{\pm 0.00}$	94.61$_{\pm 0.06}$	58.09$_{\pm 0.69}$	72.74$_{\pm 0.45}$	74.96$_{\pm 0.10}$
RAdam	90.47$_{\pm 0.34}$	93.41$_{\pm 0.21}$	93.75$_{\pm 0.04}$	60.20$_{\pm 0.37}$	74.08$_{\pm 0.35}$	75.82$_{\pm 0.28}$
MSVAG	90.08$_{\pm 0.13}$	94.79$_{\pm 0.08}$	95.01$_{\pm 0.12}$	61.55$_{\pm 0.23}$	75.75$_{\pm 0.06}$	76.84$_{\pm 0.13}$
Lion	88.04$_{\pm 0.06}$	93.97$_{\pm 0.10}$	94.26$_{\pm 0.02}$	55.59$_{\pm 0.15}$	72.79$_{\pm 0.14}$	73.41$_{\pm 0.10}$
SophiaG	88.53$_{\pm 0.04}$	94.15$_{\pm 0.26}$	94.53$_{\pm 0.13}$	58.01$_{\pm 1.85}$	72.83$_{\pm 0.18}$	75.81$_{\pm 0.23}$
AdaBound	90.41$_{\pm 0.12}$	94.93$_{\pm 0.12}$	95.06$_{\pm 0.13}$	64.51$_{\pm 0.15}$	76.37$_{\pm 0.29}$	77.43$_{\pm 0.18}$
AdaBelief	91.24$_{\pm 0.04}$	95.18$_{\pm 0.01}$	95.44$_{\pm 0.04}$	67.59$_{\pm 0.03}$	77.47$_{\pm 0.34}$	79.20$_{\pm 0.16}$



(a) VGG11 on CIFAR-10 (Training) (b) ResNet34 on CIFAR-10 (Training) (c) DenseNet121 on CIFAR-10 (Training)



(d) VGG11 on CIFAR-10 (Test) (e) ResNet34 on CIFAR-10 (Test) (f) DenseNet121 on CIFAR-10 (Test)

Figure 1. Training (top row) and test (bottom row) accuracy of CNNs on CIFAR-10 dataset. We report confidence interval ($[\mu \pm \sigma]$) of 3 independent runs.

E.2. Image Classification on ImageNet

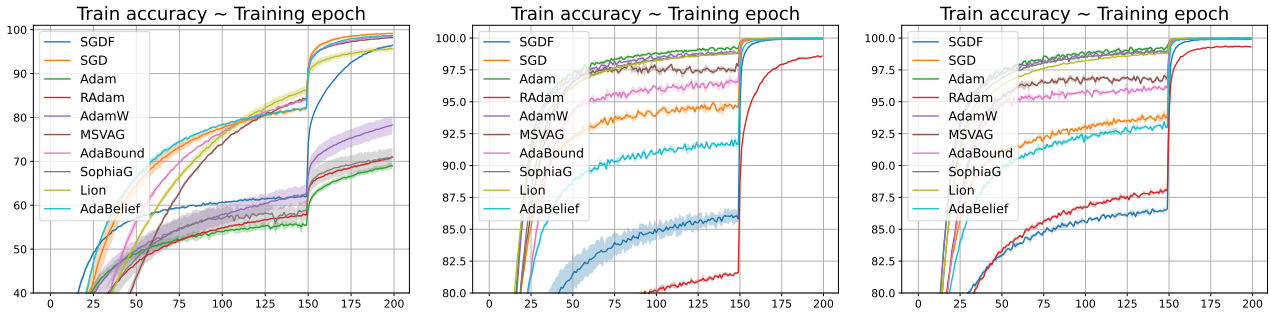
We experimented with a VGG / ResNet / DenseNet on ImageNet classification task. For SGDF and SGD, we set the initial learning rate of 0.5 same as CIFAR experiments. The weight decay is set as 10^{-4} for both cases to match the settings in [21]. Here β_1 serves to capture the gradient mean. The more closer β_1 is to 1, the longer the moving window is and the wider the historical mean is captured. Since ImageNet dataset has more iterations than CIFAR dataset, we directly set $\beta_1 = 0.5$ to prevent K_t from being overly influenced by the historical mean gradient. For sure, setting β_1 to 0.9, consistent with CIFAR experiments can also be superior to SGD, and adjusting β_1 to 0.5 or 0.9 according to the size of the dataset and batch size can bring better results. Detailed experimental parameters we place in Tab. 3. As shown in Fig. 3, SGDF outperformed SGD.

Table 3. Hyperparameters used for ImageNet.

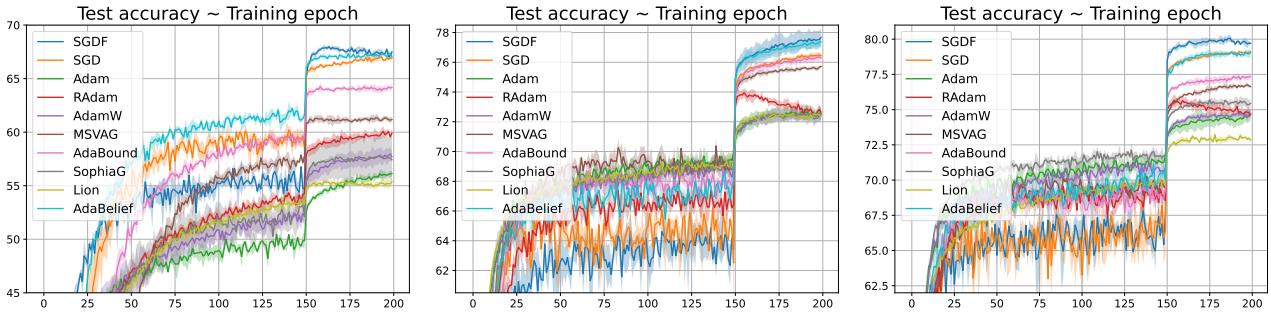
Optimizer	Learning Rate	β_1	β_2	Epochs	Schedule	Weight Decay	Batch Size	ε
SGDF	0.5	0.5	0.999	100/90	Cosine	0.0001	256	1e-8
SGD	0.1	0.9	-	100/90	Cosine	0.0001	256	-

E.3. Objective Detection on PASCAL VOC

We conducted object detection experiments on the PASCAL VOC dataset [9]. The model used in these experiments was pre-trained on the COCO dataset [20], obtained from the official website. We trained this model on the VOC2007 and VOC2012 trainval dataset (17K) and evaluated it on the VOC2007 test dataset (5K). The utilized model was Faster-RCNN [29] with FPN, and the backbone was ResNet50 [11]. We train 4 epochs and adjust the learning rate decay by a factor of 0.1 at the last epoch. We show the results on PASCAL VOC [9]. Object detection with a Faster-RCNN model[29]. Detailed experimental parameters we place in Fig. 4. The detection examples are shown in Fig. 4. These results also illustrate that our method is still efficient in object detection tasks.



(a) VGG11 on CIFAR-100 (Training) (b) ResNet34 on CIFAR-100 (Training) (c) DenseNet121 on CIFAR-100 (Training)



(d) VGG11 on CIFAR-100 (Test) (e) ResNet34 on CIFAR-100 (Test) (f) DenseNet121 on CIFAR-100 (Test)

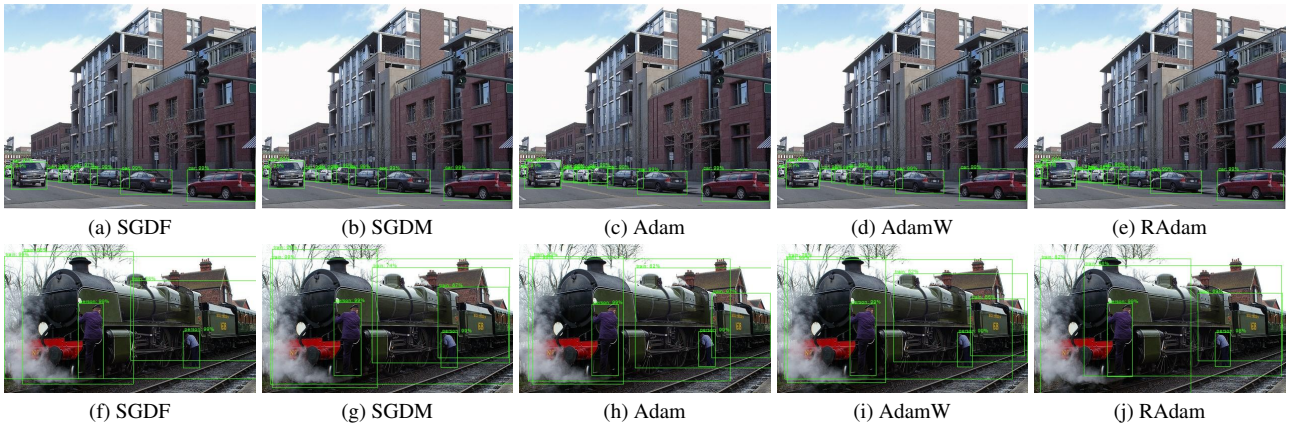
Figure 2. Training (top row) and test (bottom row) accuracy of CNNs on CIFAR-100 dataset. We report confidence interval ($[\mu \pm \sigma]$) of 3 independent runs.

(a) Training accuracy (top-1) (b) Test accuracy (top-1)

Figure 3. Training and test accuracy (top-1) of ResNet18 on ImageNet.

Table 4. Hyperparameters for object detection on PASCAL VOC using Faster-RCNN+FPN with different optimizers.

Optimizer	Learning Rate	β_1	β_2	Epochs	Schedule	Weight Decay	Batch Size	ε
SGDF	0.01	0.9	0.999	4	StepLR	0.0001	2	1e-8
SGD	0.01	0.9	-	4	StepLR	0.0001	2	-
Adam	0.0001	0.9	0.999	4	StepLR	0.0001	2	1e-8
AdamW	0.0001	0.9	0.999	4	StepLR	0.0001	2	1e-8
RAdam	0.0001	0.9	0.999	4	StepLR	0.0001	2	1e-8



(a) SGDF (b) SGDM (c) Adam (d) AdamW (e) RAdam

(f) SGDF (g) SGDM (h) Adam (i) AdamW (j) RAdam

Figure 4. Detection examples using Faster-RCNN + FPN trained on PASCAL VOC.

E.4. Image Generation

The stability of optimizers is crucial, especially when training Generative Adversarial Networks (GANs). If the generator and discriminator have mismatched complexities, it can lead to imbalance during GAN training, causing the GAN to fail to converge. This is known as model collapse. For instance, Vanilla SGD frequently causes model collapse, making adaptive optimizers like Adam and RMSProp the preferred choice. Therefore, GAN training provides a good benchmark for assessing optimizer stability. For reproducibility details, please refer to the parameter table in Tab. 6.

We evaluated the Wasserstein-GAN with gradient penalty (WGAN-GP) [30]. Using well-known optimizers [2, 37], the model was trained for 100 epochs. We then calculated the Frechet Inception Distance (FID) [12] which is a metric that measures the similarity between the real image and the generated image distribution and is used to assess the quality of the generated model (lower FID indicates superior performance). Five random runs were conducted, and the outcomes are presented in Tab. 5. Results for SGD and MSVAG were extracted from Zhuang *et al*. [39].

Table 5. FID score of WGAN-GP.

Method	SGDF	Adam	RMSProp	RAdam	Fromage	Yogi	AdaBound	SGD	MSVAG
FID	88.7 ± 4.9	**78.6** ± 4.8	109.2 ± 14.5	93.4 ± 8.3	101.5 ± 28.9	138.7 ± 21.2	119.8 ± 24.6	250.3 ± 30.1	239.7 ± 5.2

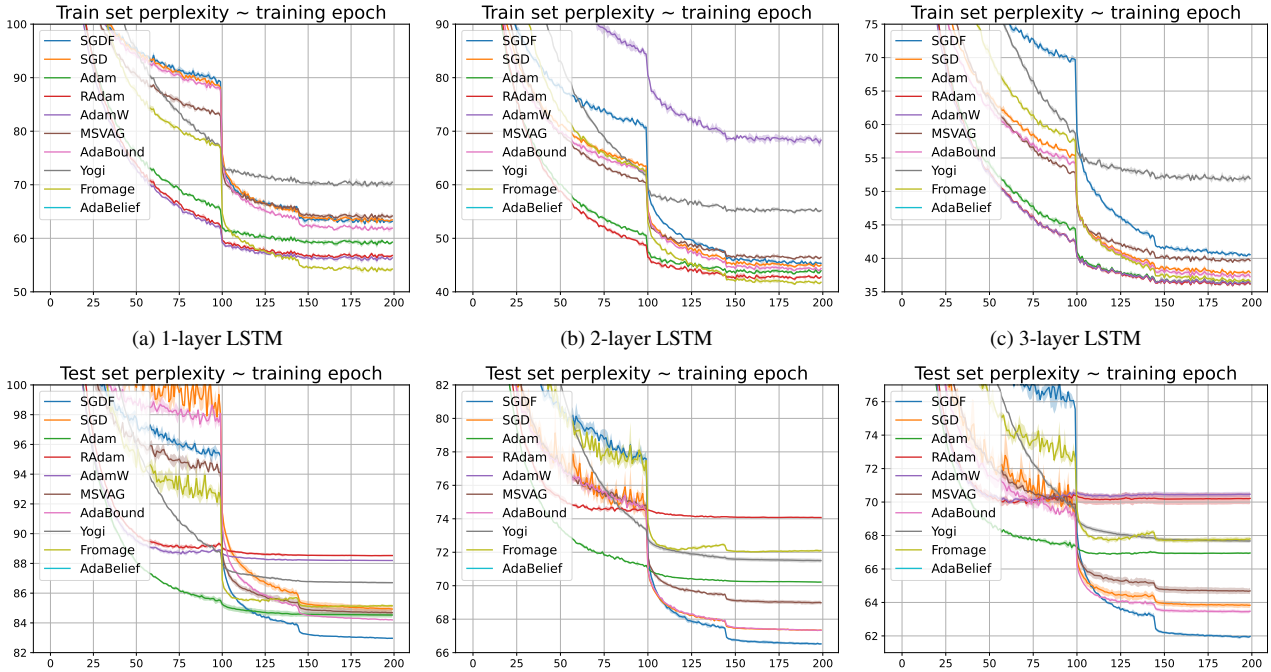
Experimental results demonstrate that SGDF significantly enhances WGAN-GP model training, achieving a FID score higher than vanilla SGD and outperforming most adaptive optimization methods. The integration of an Optimal Linear Filter in SGDF facilitates smooth gradient updates, mitigating training oscillations and effectively addressing the issue of pattern collapse.

Table 6. Hyperparameters for Image Generation Tasks.

Optimizer	Learning Rate	β_1	β_2	Epochs	Batch Size	ε
SGDF	0.01	0.5	0.999	100	64	1e-8
SGD	0.01	0.5	-	100	64	-
Adam	0.0002	0.5	0.999	100	64	1e-8
AdamW	0.0002	0.5	0.999	100	64	1e-8
Fromage	0.01	0.5	0.999	100	64	1e-8
RMSProp	0.0002	0.5	0.999	100	64	1e-8
AdaBound	0.0002	0.5	0.999	100	64	1e-8
Yogi	0.01	0.5	0.999	100	64	1e-8
RAdam	0.0002	0.5	0.999	100	64	1e-8

E.5. LSTM on Language Modeling

We experiment with LSTM [23] on the Penn-TreeBank dataset [24]. For SGDF, we apply gradient clipping with values of 0.1 for the LSTM 1-layer, 0.15 for 2-layer, and 0.25 for 3-layer. The learning rate is decayed by multiplying it by 0.1 at epochs [100, 145] using StepLR scheduling, as detailed in Tab. 7 alongside other hyperparameters. We report the mean and standard deviation across 3 independent runs with random seeds $\{0, 1, 2\}$ to ensure result robustness. As shown in Fig. 5, which presents both training and test perplexity curves for 1-layer, 2-layer, and 3-layer LSTM architectures, SGDF consistently outperforms other optimizers by achieving the lowest perplexity across all model configurations, highlighting its effectiveness for sequence modeling tasks. Notably, Adabelief performs poorly under standard parameters, failing to match the performance of other optimizers.



(a) 1-layer LSTM (b) 2-layer LSTM (c) 3-layer LSTM

(d) 1-layer LSTM (e) 2-layer LSTM (f) 3-layer LSTM

Figure 5. Training (top row) and test (bottom row) perplexity on Penn-TreeBank dataset, lower is better.

Table 7. Hyperparameters used for LSTM.

Optimizer	Learning Rate	β_1	β_2	Epochs	Schedule	Weight Decay	Batch Size	ε
SGDF	60	0.9	0.999	200	StepLR	1.2e-6	20	1e-8
SGD	30	0.9	-	200	StepLR	1.2e-6	20	-
Adam	0.001	0.9	0.999	200	StepLR	1.2e-6	20	1e-8
RAdam	0.001	0.9	0.999	200	StepLR	1.2e-6	20	1e-8
AdamW	0.001	0.9	0.999	200	StepLR	1.2e-6	20	1e-8
MSVAG	30	0.9	0.999	200	StepLR	1.2e-6	20	1e-8
AdaBound	0.001	0.9	0.999	200	StepLR	1.2e-6	20	-
Yogi	0.01	0.9	0.999	200	StepLR	1.2e-6	20	1e-3
Fromage	0.01	0.9	0.999	200	StepLR	1.2e-6	20	-
Adabelief	0.001	0.9	0.999	200	StepLR	1.2e-6	20	1e-8

E.6. Post-training in ViT.

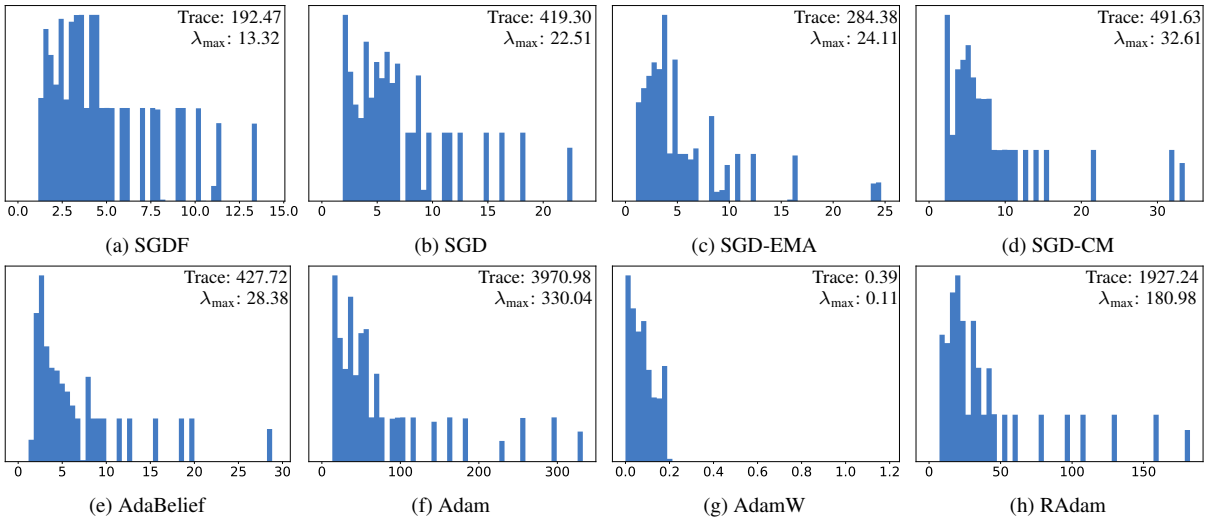
To evaluate SGDF's performance, we used Vision Transformers (ViT) [7] on six benchmark datasets: CIFAR-10, CIFAR-100, Oxford-IIIT-Pets [27], Oxford Flowers-102 [26], Food101 [3], and ImageNet-1K. Two ViT variants, ViT-B/32 and ViT-L/32, pretrained on ImageNet-21K, were selected. For fine-tuning, we replaced the original MLP classification head with a new fully connected layer, tailored to the dataset categories. All Transformer backbone weights were retained, preserving the rich representations learned from ImageNet-21K. We increased the image resolution (*e.g.*, from 224×224 to 384×384) to improve accuracy, while adjusting positional encoding through 2D interpolation to match the new resolution. For optimization, SGDF was compared to SGD with momentum as a baseline (We research learning set $\{$ 0.001, 0.003, 0.01, 0.03$\}$ same as [7]. For ours method, we're not tuning and just mirror the hyperparameter in the CIFAR experiments.), using cosine learning rate decay and no weight decay. A batch size of 512 and global gradient clipping (norm of 1) were used to prevent gradient explosion. All experiments were trained uniformly for 10 epochs and the random seed is set as the current year. We set the random seed to $\{0, 1, 2\}$. We summarized the hyperparameter in Tab. 8.

Table 8. Hyperparameters used for fine-tuning ViT.

Optimizer	Learning Rate	β_1	β_2	Epochs	Schedule	Weight Decay	Batch Size	ε	Resolution
SGDF	0.5	0.9	0.999	10	Cosine	0	512	1e-8	384
SGD	0.03	0.9	-	10	Cosine	0	512	-	384

E.7. Top Eigenvalues of Hessian and Hessian Trace

We computed the Hessian spectrum of ResNet-18 trained on the CIFAR-100 dataset for 200 epochs using more optimization methods: SGDF, SGD, SGD-EMA, SGD-CM, Adabelief, Adam, AdamW, and RAdam. We employed power iteration [35] to compute the top eigenvalues of Hessian and Hutchinson's method [36] to compute the Hessian trace. Histograms illustrating the distribution of the top 50 Hessian eigenvalues for each optimization method are presented in Fig. 6. SGDF brings lower eigenvalue and trace of the hessian matrix, which explains the fact that SGDF demonstrates better performance than SGD as the categorization category increases. Note that Fig. 6g shows that AdamW achieves very low hessian matrix eigenvalues and traces, but the final test set accuracy is about 4% lower than the other methods, and that AdamW's unique decouple weight decay changes the nature of the converged solution (We apply decoupled weight decay to other algorithms and similar results occur).



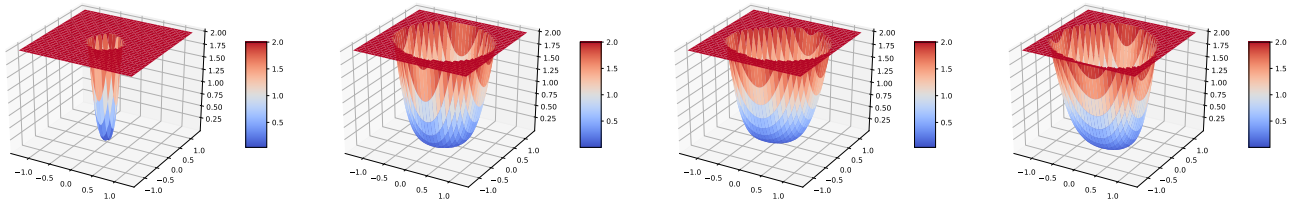
(a) SGDF (b) SGD (c) SGD-EMA (d) SGD-CM

(e) AdaBelief (f) Adam (g) AdamW (h) RAdam

Figure 6. Histogram of Top 50 Hessian Eigenvalues.

E.8. Visualization of Landscapes

We visualized the loss landscapes of models trained with SGD, SGDM, SGDF, and Adam using the ResNet-18 model on CIFAR-100, following the method in [18]. All models are trained with the same hyperparameters for 200 epochs. As shown in Fig. 7, SGDF finds flatter minima. Notably, the visualization reveals that Adam is more prone to converge to sharper minima.



(a) Adam (b) SGD (c) SGDM (d) SGDF

Figure 7. Visualization of loss landscape. Adam converges to sharp minima.

E.9. Computational Cost Analysis

Tab. 9 reports the per-parameter arithmetic cost of several optimizers. We count elementwise multiplications, additions/subtractions, divisions, and square roots as unit-cost operations. Red numbers indicate the additional overhead of *coupled* weight decay, while green numbers indicate the smaller overhead of *decoupled* weight decay. Compared with plain stochastic gradient methods, adaptive optimizers (e.g., Adam and SGDF) incur extra operations for moment estimation and normalization. Their optimized variants (AdamW and optimized SGDF) reduce compute by removing redundant elementwise divisions and using decoupled weight decay.

Optimizer	**Per-Parameter FLOPs**	**Operation Breakdown**
SGD	$\approx$ **2 ops/param** (+2 ops) (+1 op)	$\{1\times,\ 1+\}$
SGDM	$\approx$ **4 ops/param** (+2 ops) (+1 op)	$\{2\times,\ 2+\}$
Adam	**16 ops/param** $\rightarrow$ **14 ops/param** (+2 ops) (+1 op)	$\{7\times,\ 5+,\ 3\div,\ 1\sqrt{}\} \rightarrow \{7\times,\ 4+,\ 2\div,\ 1\sqrt{}\}$
SGDF	**22 ops/param** $\rightarrow$ **20 ops/param** (+2 ops) (+1 op)	$\{10\times,\ 6+,\ 2-,\ 2\div,\ 1\sqrt{}\} \rightarrow \{10\times,\ 6+,\ 1-,\ 1\div,\ 1\sqrt{}\}$

Table 9. Arithmetic cost and operation breakdown of optimizers. Red: Coupled Weight Decay, Green: Decoupled Weight Decay.

Both **Adam** and **SGDF** reduce per-parameter cost through algebraic simplification and bias-correction restructuring. For **Adam**, the update follows $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$, $v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$, and $\theta_{t+1} = \theta_t - \eta \frac{m_t/(1-\beta_1^t)}{\sqrt{v_t/(1-\beta_2^t)}+\epsilon}$. In **AdamW**, redundant elementwise divisions are avoided by precomputing $\text{step\_size} = \eta \frac{\sqrt{1-\beta_2^t}}{1-\beta_1^t}$ and using decoupled weight decay: $\theta \leftarrow (1-\eta\lambda)\theta - \text{step\_size} \frac{m_t}{\sqrt{v_t}+\epsilon}$. This reduces roughly two operations per parameter (16 $\rightarrow$ 14) while preserving the update.

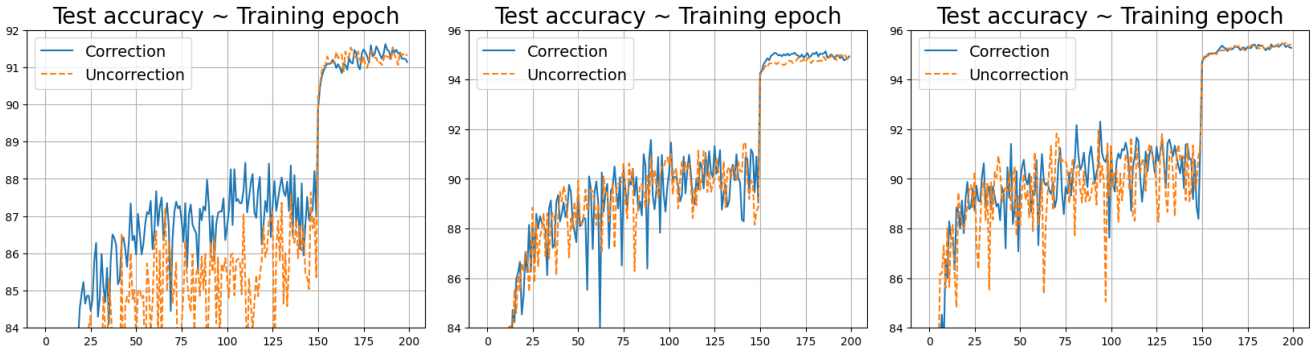
Inspired by this design, **SGDF** refines adaptive updates via the residual-variance estimate. Let $\widehat{m}_t = m_t/(1-\beta_1^t)$ and $\widehat{s}_t = s_t/c_{2,t}$, where $c_{2,t} = \frac{(1+\beta_1)(1-\beta_2^t)}{(1-\beta_1)(1-\beta_1^{2t})}$. The gain is $K_t = \frac{\widehat{s}_t}{\widehat{s}_t + (g_t - \widehat{m}_t)^2 + \epsilon}$, and the filtered gradient is $g'_t = \widehat{m}_t + K_t^{\gamma}(g_t - \widehat{m}_t)$, yielding $\theta_{t+1} = \theta_t - \eta\, g'_t$. A direct elementwise implementation explicitly forms $\widehat{s}_t = s_t/c_{2,t}$ (one elementwise division) and evaluates $(g_t - \widehat{m}_t)$ twice (once in K_t, once in g'_t), resulting in about 22 operations per parameter.

Our optimized implementation avoids explicitly forming $\widehat{s}_t$ by substituting $\widehat{s}_t = s_t/c_{2,t}$ into K_t and multiplying numerator and denominator by $c_{2,t}$, obtaining the equivalent form $K_t = \frac{s_t}{s_t + c_{2,t}((g_t - \widehat{m}_t)^2 + \epsilon)}$. This removes one elementwise division (replacing it with a scalar multiplication). In addition, we reuse the residual $r_t = g_t - \widehat{m}_t$ across the gain computation and the final filtered update, eliminating one redundant elementwise subtraction. With decoupled weight decay, the per-parameter cost is reduced from about 22 $\rightarrow$ 20 operations.

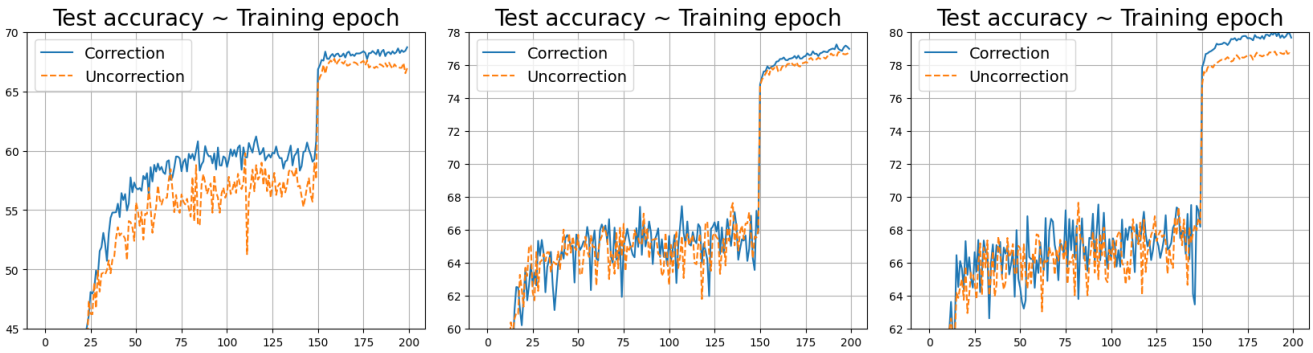
To complement the theoretical operation analysis, we further conducted an empirical runtime evaluation to quantify the real-world computational efficiency of different optimizers. Each optimizer was benchmarked on FP16 mixed-precision

Model	SGDM (h)	Adam (h)	SGDF (h)
VGG13	45.71	45.90	$47.78_{\downarrow 0.63}$
ResNet101	35.11	35.32	$39.77_{\downarrow 1.59}$
DenseNet161	57.68	58.03	$64.81_{\downarrow 1.99}$

Table 10. Total training time (h) for 100 epochs with batch size 256 on FP16 AMP in 224^2 Pixel ImageNet.



(a) VGG11 on CIFAR-10 (b) ResNet34 on CIFAR-10 (c) DenseNet121 on CIFAR-10



(d) VGG11 on CIFAR-100 (e) ResNet34 on CIFAR-100 (f) DenseNet121 on CIFAR-100

Figure 8. SGDF with or without the correction factor. The curve shows the accuracy of the test.

training for 100 epochs with a batch size of 256 across representative CNN backbones (VGG13, ResNet101, DenseNet161) on ImageNet. The measured wall-clock times are summarized in Tab. 10.

E.10. Ablation Study

We derived a correction factor $(1-\beta_1)(1-\beta_1^{2t})/(1+\beta_1)$ from the geometric progression to correct the variance of by the correction factor. So we test the SGDF with or without correction in VGG, ResNet, DenseNet on CIFAR. We report both test accuracy in Fig. 8. It can be seen that the SGDF with correction exceeds the uncorrected one. To better observe the effect of static momentum coefficients on the gradient estimation, while comparing our time-varying SGDF. We use VGG [31] because it is a very standard network with no modules that interfere with the gradient, allowing for a better representation of the optimizer's update mechanism. We trained it with different SGD-based methods: Vanilla SGD, SGD with EMA, SGD with Filter, and SGD with CM. Then, we plot curve in Fig. 9 and use kernel density estimates of gradient values distribution over the first 100 iterations in Fig. 10.

From Fig. 9, applying SGD with EMA and Filter, convergence is faster than vanilla SGD. EMA has less fluctuation in test curves. WF demonstrates higher test accuracy with the same training set accuracy and reduced generalization gap. On the other hand, CM is slow to converge and results fluctuate because of the larger bias and variance.

Fig. 10a shows high variance and uneven gradient values distribution in Vanilla SGD, resulting in training oscillations

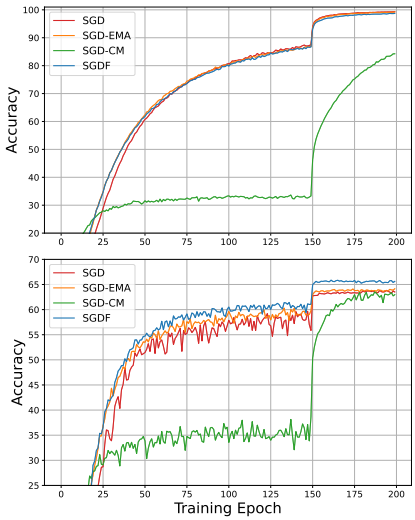
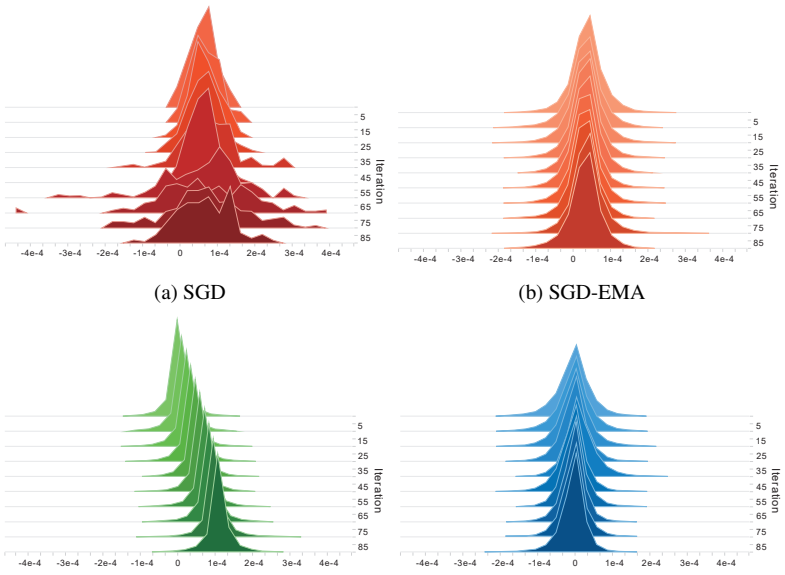


Figure 9. Train the VGG model on the CIFAR-100 dataset using the same initial learning rate of 0.1, and multiply it by a factor of 0.1 at the 150th epoch.



(a) SGD (b) SGD-EMA

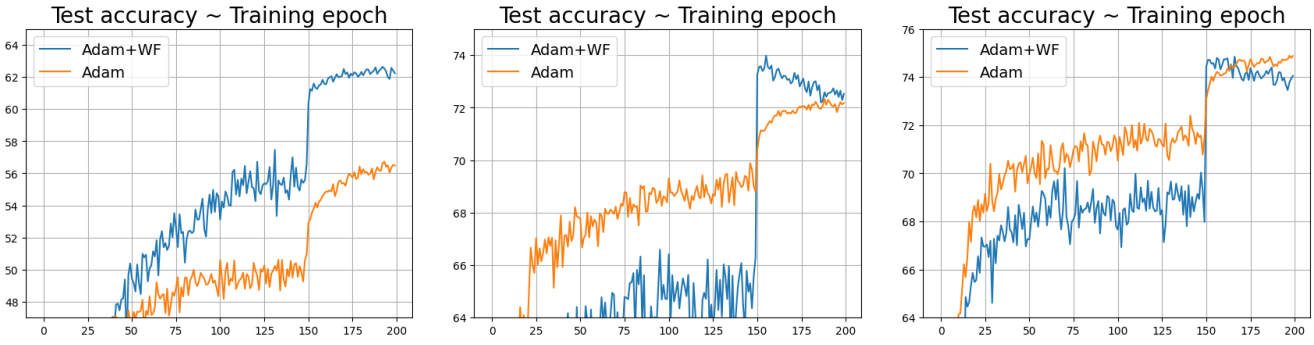
(c) SGD-CM (d) SGDF

Figure 10. The gradient histogram of the VGG on the CIFAR-100 dataset. The x-axis is the gradient value and the height is the frequency. SGD trains the VGG without BN, the variance of the gradient fluctuates dramatically and the update is unstable.

that hinder stable convergence. In contract, Fig. 10b and Fig. 10d shows concentrated gradient distribution and not distorted. Especially, Fig. 10c shows that SGD-CM smooths values fluctuations but introduces *gradient shift*, causing bias and variance over time. Previous research highlights that momentum struggle to adapt to variations in the curvature of the objective function, potentially causing deviation in updates [8, 38].

E.11. Extensibility of Filter-Estimated Gradients

The study involves evaluating the vanilla Adam optimization algorithm and its enhancement with an Optimal Linear Filter on the CIFAR-100 dataset. Fig. 11 contains detailed test accuracy curves for both methods across different models. The results indicate that the adaptive learning rate algorithms exhibit improved performance when supplemented with the proposed first-moment filter estimation. This suggests that integrating an Optimal Linear Filter with the Adam optimizer may improve performance.



(a) VGG11 on CIFAR-100 (b) ResNet34 on CIFAR-100 (c) DenseNet121 on CIFAR-100

Figure 11. Test accuracy of CNNs on CIFAR-100 dataset. We train vanilla Adam and Adam combined with Optimal Linear Filter.

E.12. Classical Momentum Discussion

In our framework, the EMA-Momentum is treated as a low-pass filter, in the nature of noise reduction. Cutkosky *et al*. [5] also proves the property that EMA-Momentum cancels out noise, further supporting our analysis. We further discuss classical momentum here.

Theoretically, we show that momentum converges faster than SGD in the setting of u-strong acceleration, but deep learning optimization does not always conform to this. Leclerc *et al*. [17] tested the classical momentum at different learning rates, taking the momentum factor $\{0, 0.5, 0.9\}$. It is empirically found that it is at small learning rates that the classical momentum speeds up the convergence of training losses. That is, SGD-CM can be either better or worse than SGD. In addition, Kunstner *et al*. [16] found that the classical momentum can only show an advantage over SGD when the batch size increases and approaches the full gradient, at which point the noise introduced by random sampling is almost non-existent. In our proof, we mentioned that SGD-CM introduces both bias and variance, but with a full gradient, SGD-CM does not introduce noise and only causes the gradient to produce bias.

We have not analyzed the nature of bias and variance for convergent solutions, but a certain amount of bias may lead to better results when the noise is reduced, and intuitively this may help the algorithm to discuss saddle points or local minima and converge to flatter regions, in a similar nature to the implicitly flat regularity introduced by noise [34]. Because the algorithm converges, the gradient at the position of convergence must be stable, and the classical momentum accumulation gradient, with its large values, must go to a smooth plateau in order to avoid oscillations. Also, it is implied that the gradient bias may not produce irretrievable results, since the bias decreases as the gradient converges, and the direction of the gradient may be more important. Sign SGD [1] takes sign for the gradient, which also converges, and only needs to be applied to the cosine learning rate decay.

Our overall opinion is that CM does not accelerate SGD, but brings better generalization. Early deep learning optimizations focused on reducing the noise introduced by SGD, resulting in several variance reduction algorithms, where reducing variance increases the speed of convergence [4]. The noise introduced by CM hinders convergence, but bias brings better generalization. Thus, the above empirical observation that the momentum method can only be accelerated at small learning rates is due to the reduced step size of SGD, which naturally slows down the convergence rate. Whereas the bias from CM offsets the effect of the reduced step size, and the step size reduces the variance of the gradient sequence. This also implies why deep learning uses warm-up to make the gradient more stable in the pre-training period[21].

References

[1] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR, 2018. 33

[2] Jeremy Bernstein, Arash Vahdat, Yisong Yue, and Ming-Yu Liu. On the distance between two neural networks and the stability of learning. *Advances in Neural Information Processing Systems*, 33:21370–21381, 2020. 27

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 29

[4] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018. 33

[5] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized sgd. In *International conference on machine learning*, pages 2260–2268. PMLR, 2020. 33

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 24

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 29

[8] Timothy Dozat. Incorporating nesterov momentum into adam. *International Conference on Learning Representations Workshop, 2016*, 2016. 32

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 25

[10] Ramsey Faragher. Understanding the basis of the kalman filter via a simple and intuitive derivation [lecture notes]. *IEEE Signal processing magazine*, 29(5):128–132, 2012. 9

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 24, 25

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 27

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 24

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 24

[15] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 24

[16] Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between sgd and adam on transformers, but sign descent might be. *arXiv preprint arXiv:2304.13960*, 2023. 33

[17] Guillaume Leclerc and Aleksander Madry. The two regimes of deep network training. *arXiv preprint arXiv:2002.10376*, 2020. 33

[18] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018. 30

[19] Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019. 2

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *European Conference on Computer Vision (ECCV)*, 2014. 25

[21] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*, 2020. 24, 25, 33

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 24

[23] Xiaolei Ma, Zhimin Tao, Yinhai Wang, Haiyang Yu, and Yunpeng Wang. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies*, 54:187–197, 2015. 28

[24] Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Using Large Corpora*, 273:31, 1994. 28

[25] Robbins Sutton Monro. a stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951. 24

[26] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 29

[27] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 29

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

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Neural Information Processing Systems (NIPS)*, 2015. 25

[30] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 27

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014. 24, 31

[32] Mandt Stephan, Matthew D Hoffman, David M Blei, et al. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017. 1, 2

[33] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013. 1

[34] Ning Yang, Chao Tang, and Yuhai Tu. Stochastic gradient descent introduces an effective landscape-dependent regularization favoring flat solutions. *Physical Review Letters*, 130(23):237101, 2023. 33

[35] Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. *Advances in Neural Information Processing Systems*, 31, 2018. 29

[36] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W. Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *International Conference on Big Data*, 2020. 29

[37] Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. *Advances in neural information processing systems*, 31, 2018. 27

[38] Jian Zhang and Ioannis Mitliagkas. Yellowfin and the art of momentum tuning. *arXiv preprint arXiv:1706.03471*, 2017. 32

[39] Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar C Tatikonda, Nicha Dvornek, Xenophon Papademetris, and James Duncan. Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. *Advances in neural information processing systems*, 33: 18795–18806, 2020. 27