Dynamic Momentum Recalibration in Online Gradient Learning

Conference Presentation

Zhipeng Yao [1, 2] ⋆ Rui Yu [2] † Guisong Chang [3] Ying Li [1] Yu Zhang [1] Dazhou Li [1] †

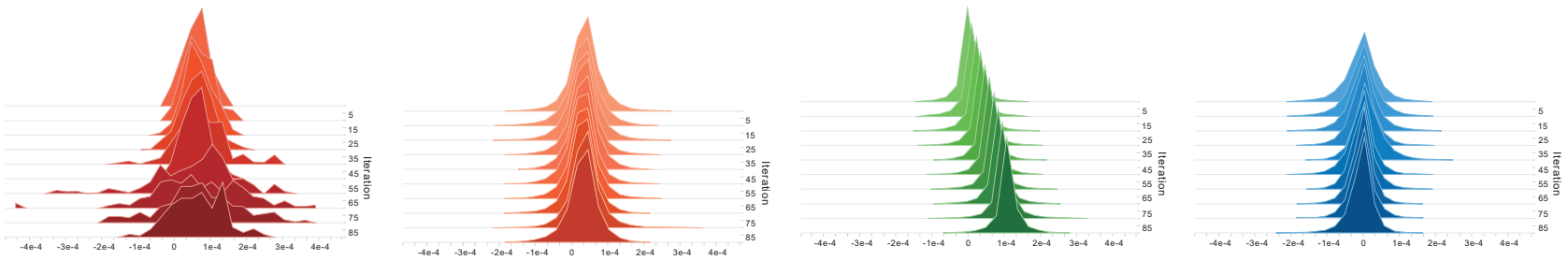
[1]Shenyang University of Chemical Technology

[2]University of Louisville [3]Northeastern University

Motivation

Standard momentum relies on static coefficients, trapping optimizers in a rigid trade-off between gradient bias (which causes plateaus) and variance (which drives instability).

- **Gradient Bias:** Excessive smoothing causes the optimizer to lag, trapping it in suboptimal plateaus and slowing convergence.
- **Gradient Variance:** Unregulated stochastic noise drives instability in the optimization path, causing oscillations that hinder final settling.



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

Figure 1: 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.

The Gradient Estimation Dilemma

Stochastic optimization faces a fundamental challenge: the gradient **bias-variance trade-off**. We unify prominent momentum strategies into a single framework:

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

where u scales the current gradient. This encapsulates **EMA** ($u = 1 - \beta$) and **CM** ($u = 1$). For any gradient estimator $\hat{g}_t = \mathcal{A}(g_1, ..., g_t)$, the estimation of the mean square error decomposes as **bias** and **variance**:

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

By reformulating momentum in continuous time (SDEs), we derive asymptotic bounds for bias and variance. **The key finding:** Static choices of u and β lock the estimator into a rigid trade-off.

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

$$\text{Var}(m(t)) \leq \frac{u^2 \sigma^2}{1-\beta} + \frac{2u^2 V^2}{(1-\beta)^2}. \tag{4}$$

The Gradient Estimation Dilemma

Table 1: Bias and variance bounds for different momentum estimators.

Method	**Bias Bound**	**Variance Bound**	**Limit** ($\beta \to 1$)
SGD ($\beta = 0$)	0	$\sigma^2 + 2V^2$	N/A
EMA ($u = 1 - \beta$)	$\mathcal{O}\left(\frac{1}{1-\beta}\right)$	$(1-\beta)\sigma^2 + 2V^2$	Bias $\to \infty$, Var $\to 2V^2$
CM ($u = 1$)	$\mathcal{O}\left(\frac{1}{(1-\beta)^3}\right)$	$\frac{\sigma^2}{1-\beta} + \frac{2V^2}{(1-\beta)^2}$	Bias $\to \infty$, Var $\to \infty$

The Dilemma: Structurally reducing variance inevitably amplifies bias, while minimizing bias exposes the estimator to higher variance. *Can we design an adaptive gain that resolves this?*

Optimal Linear Filter

SGDF reframes gradient estimation as a **Signal Processing** task, utilizing the **Minimum Mean Square Error (MMSE)** to dynamically fuse historical momentum and new observations.
We model the estimate as a linear interpolation between the bias-corrected momentum $\hat{m}_t$ and the noisy gradient g_t:

$$\hat{g}_t = \hat{m}_t + K_t(g_t - \hat{m}_t). \tag{5}$$

Minimizing the $\hat{g}_t$ variance $\mathrm{Var}(\hat{g}_t)$ yields the **Optimal Gain** K_t:

$$K_t = \frac{\mathrm{Var}(\hat{m}_t)}{\mathrm{Var}(\hat{m}_t) + \mathrm{Var}(g_t)}. \tag{6}$$

This is equivalent to fusing two Gaussian distributions $\mathcal{N}(\mu_m, \sigma_m^2)$ and $\mathcal{N}(\mu_g, \sigma_g^2)$, resulting in a posterior with **reduced uncertainty**:

$$\mathbb{E}[\hat{g}_t] = \frac{\sigma_g^2 \mu_m + \sigma_m^2 \mu_g}{\sigma_m^2 + \sigma_g^2}, \quad \mathrm{Var}(\hat{g}_t) = \frac{\sigma_m^2 \sigma_g^2}{\sigma_m^2 + \sigma_g^2}. \tag{7}$$

Intuitively, it shifts trust toward historical momentum under high noise and favors new observations when reliable, stabilizing training to guarantee.

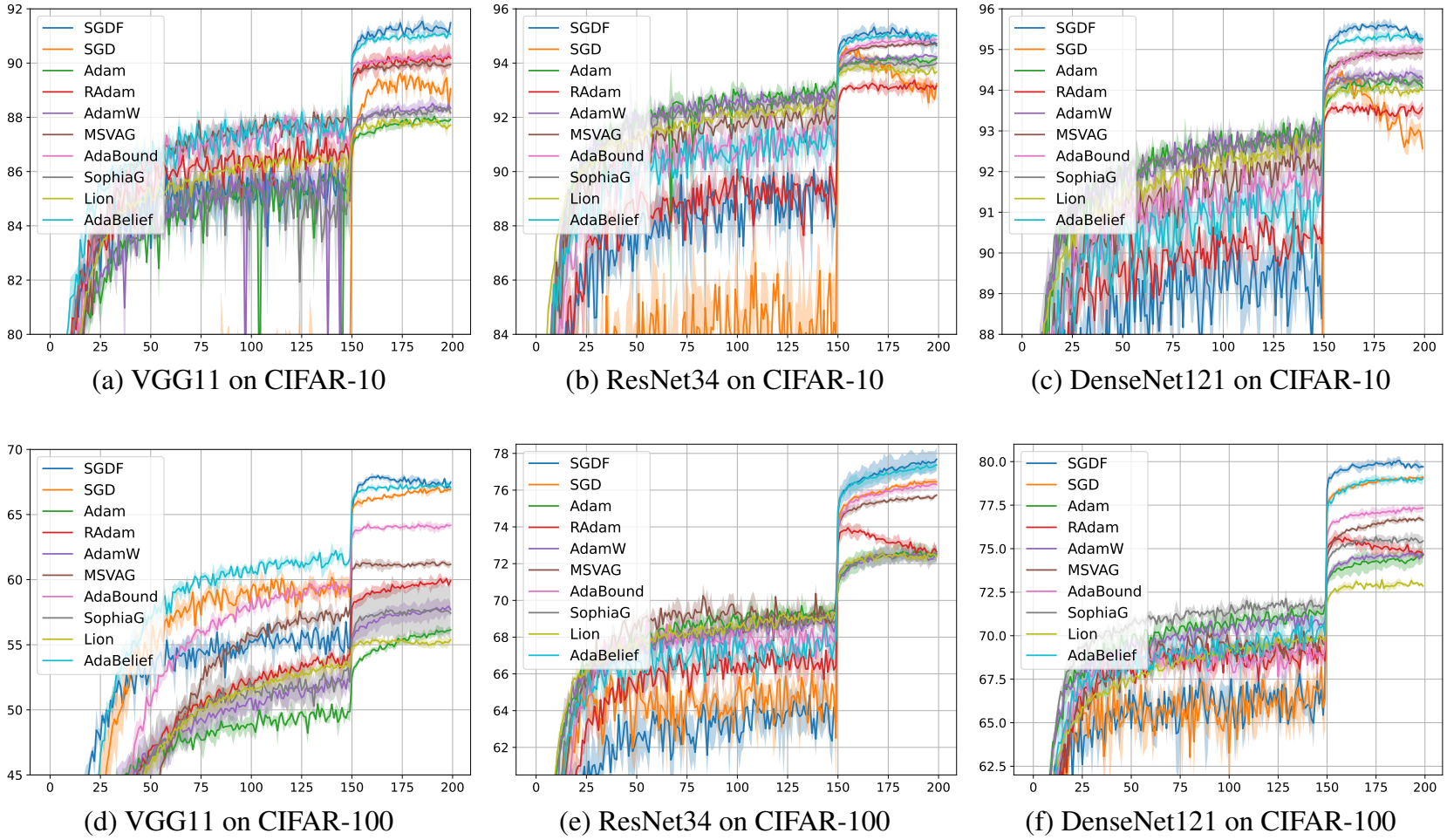
SGDF Algorithm

Algorithm 1 SGDF: Online Filter Estimate Gradient (element-wise).

Input: θ_0: initial parameter, $f(\theta)$: stochastic objective function
Parameter: $\{\alpha_t\}_{t=1}^{T}$: step size, $\{\beta_1, \beta_2\}$: attenuation coefficients, $\{\varepsilon\}$: numerical stability, $\{\gamma\}$: power scaling.
Output: θ_T: resulting parameters.

Initialize: $m_0 \leftarrow 0$, $s_0 \leftarrow 0$
while $t \leq T$ **do**
  $g_t \leftarrow \nabla f_t(\theta_{t-1})$
  $m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$
  $s_t \leftarrow \beta_2 s_{t-1} + (1-\beta_2)(g_t - m_t)^2$
  $\widehat{m}_t \leftarrow \dfrac{m_t}{1-\beta_1^t}, \widehat{s}_t \leftarrow \dfrac{(1-\beta_1)(1-\beta_1^{2t})\, s_t}{(1+\beta_1)(1-\beta_2^t)}$
  $K_t \leftarrow \dfrac{\widehat{s}_t}{\widehat{s}_t + (g_t - \widehat{m}_t)^2 + \varepsilon}$
  $\widehat{g}_t \leftarrow \widehat{m}_t + K_t^{\gamma}(g_t - \widehat{m}_t)$
  $\theta_t \leftarrow \theta_{t-1} - \alpha_t \widehat{g}_t$
end while
return θ_T

Classification on CIFAR



(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 2: Test accuracy ($[\mu \pm \sigma]$) on CIFAR.

Classification on ImageNet

Table 2: Comprehensive Benchmark Results for SGDF

Part A: Generalization Across CNN Architectures (ImageNet Top-1 Acc %)

Model	VGG11	VGG13	ResNet34	ResNet50	DenseNet121	DenseNet161
SGD	70.37	71.58	73.31	76.13	74.43	77.13
SGDF	**71.34**	**72.74**	**74.07**	**76.72**	**75.75**	**78.34**

Part B: Transfer Learning and Fine-tuning in Vision Transformer (ViT)

Model	Method	CIFAR10	CIFAR100	OxfordPets	OxfordFlowers	Food101	ImageNet	Avg
ViT-B/32	SGD	98.71	90.62	89.71	96.79	88.56	81.42	90.97
	SGDF	**98.74**	**91.44**	**92.68**	**97.17**	**89.35**	**81.52**	**91.81**
ViT-L/32	SGD	98.73	91.30	85.21	96.52	89.13	81.28	90.36
	SGDF	**98.83**	**92.20**	**91.96**	**96.79**	**90.04**	**81.38**	**91.87**

Part C: Comparison with State-of-the-Art Optimizers (ResNet18 on ImageNet)

Method	SGDF	SGD	PAdam	AdaBelief	AdaBound	Yogi	MSVAG	Adam	RAdam	AdamW
Top-1	**70.51**	70.23	70.07	70.08	68.13	68.23	65.99	63.79	67.62	67.93
Top-5	**89.69**	89.40	89.47	-	88.55	88.59	-	85.61	-	88.47

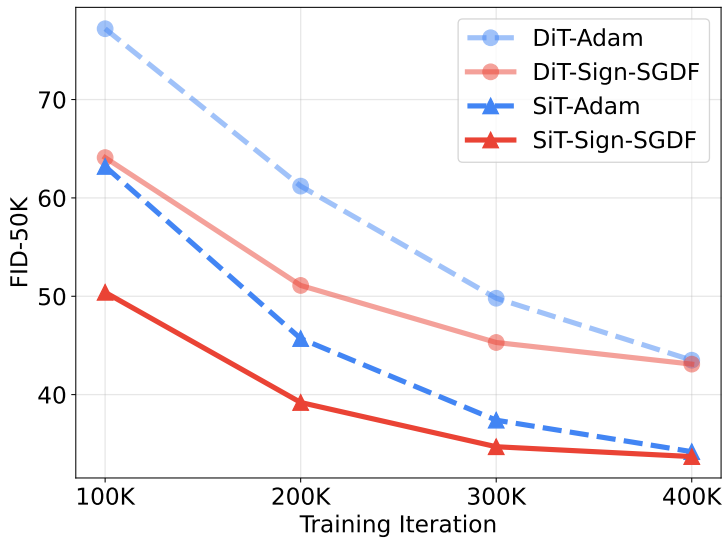
Extensibility

Table 3: Extensibility in Adaptive optimizer and Muon.

Part A: Classification Accuracy (%) in CIFAR-100			
Model	VGG11	ResNet34	DenseNet121
Filter-Adam	**62.64**	**73.98**	**74.89**
Vanilla-Adam	56.73	72.34	**74.89**

Part B: Generative Quality with Muon					
Method	FID ↓	sFID ↓	IS ↑	Pre. ↑	Rec. ↑
Adam	68.32	13.63	20.51	0.36	0.53
Muon + SGDF	**64.24**	**12.43**	**22.26**	**0.37**	**0.59**

Figure 3: Extensibility in Sign Descent.



Thanks!