Example Problem to show the Massive Variance produced by Adam at the initial Iterations :

We analyze the minimization of $f(x) = \frac{1}{2}x^2$ **with simulated noisy gradients.**

Initial Conditions:

- **Initial Position:** $x_0 = 1.0$
- **Initial Moments:** $\mathbf{m}_0 = 0, \mathbf{v}_0 = 0$

Hyperparameters (Corrected):

- **Learning Rate (**LR**):** 0.001
- **Momentum coefficient (**β_1**):** 0.9
- **Second moment coefficient (**β_2**):** 0.99
- **Epsilon (**ϵ**):** 10^{-8}

Step (t)	Gradient ($\mathbf{g}_t$) (Input)
1	$\mathbf{g}_1 = 2.0$ (Outlier/High Noise)
2	$\mathbf{g}_2 = 1.0$ (Stable)
3	$\mathbf{g}_3 = 1.0$ (Stable)
4	$\mathbf{g}_4 = 1.0$ (Stable)

2. Step-by-Step Calculation Comparison (4 Iterations)

Iteration 1 :

Calculation	Formula	Adam	Madam
Gradient $\mathbf{g}_1$	2	2	2
Momentum $\mathbf{m}_1$	$0.9(0) + 0.1(\mathbf{g}_1)$	0.2	0.2
Squared Term	$\mathbf{g}_1^2$ (Adam) / $\mathbf{m}_1^2$ (MSAM)	4.0	0.04
Raw $\mathbf{v}_1$	$0.01(\text{Sq. Term})$	(0.01) * 4 = 4	(0.01) * (0.04) = 0.0004

Correction C_1	$1/(1-0.99^1)$	**100**	**100**
$\hat{\mathbf{v}}_1$ **(Bias-Corr.)**	$C_1 \cdot \mathbf{v}_1$	**4**	**0.04**

Iteration 2 :

Calculation	**Formula**	**Adam**	**Madam**
Gradient $\mathbf{g}_2$	**1**	**1**	**1**
Momentum $\mathbf{m}_2$	$0.9(0.2) + 0.1(1.0)$	**0.28**	**0.28**
Squared Term	$\mathbf{g}_2^2 = 1.0$ **/** $\mathbf{m}_2^2 \approx 0.0784$	**1.0**	0.0784
Raw $\mathbf{v}_2$	$0.99(\mathbf{v}_1) + 0.01(\text{Sq. Term})$	**0.0496**	**0.0082**
Correction C_2	$1/(1-0.99^2)$	**50.25(approx)**	**50.25**
$\hat{\mathbf{v}}_2$ **(Bias-Corr.)**	$C_2 \cdot \mathbf{v}_2$	**2.49(approx)**	**0.41**

Iteration 3 :

Calculation	**Adam** $\mathbf{v}_3$ **/** $\hat{\mathbf{v}}_3$	**MADAM**
$\mathbf{m}_3$	**0.352**	**0.352**
Squared Term	$\mathbf{g}_3^2 = 1.0$	$\mathbf{m}_3^2 = 0.352^2 \approx \mathbf{0.124}$
Raw $\mathbf{v}_3$	**0.0591**	**0.0094**
Correction C_3	$1/(1-0.99^3) \approx \mathbf{33.67}$	**33.67**
$\hat{\mathbf{v}}_3$ **(Bias-Corr.)**	$33.67 \cdot 0.0591 \approx \mathbf{1.99}$	$33.67 \cdot 0.0094 \approx \mathbf{0.316}$

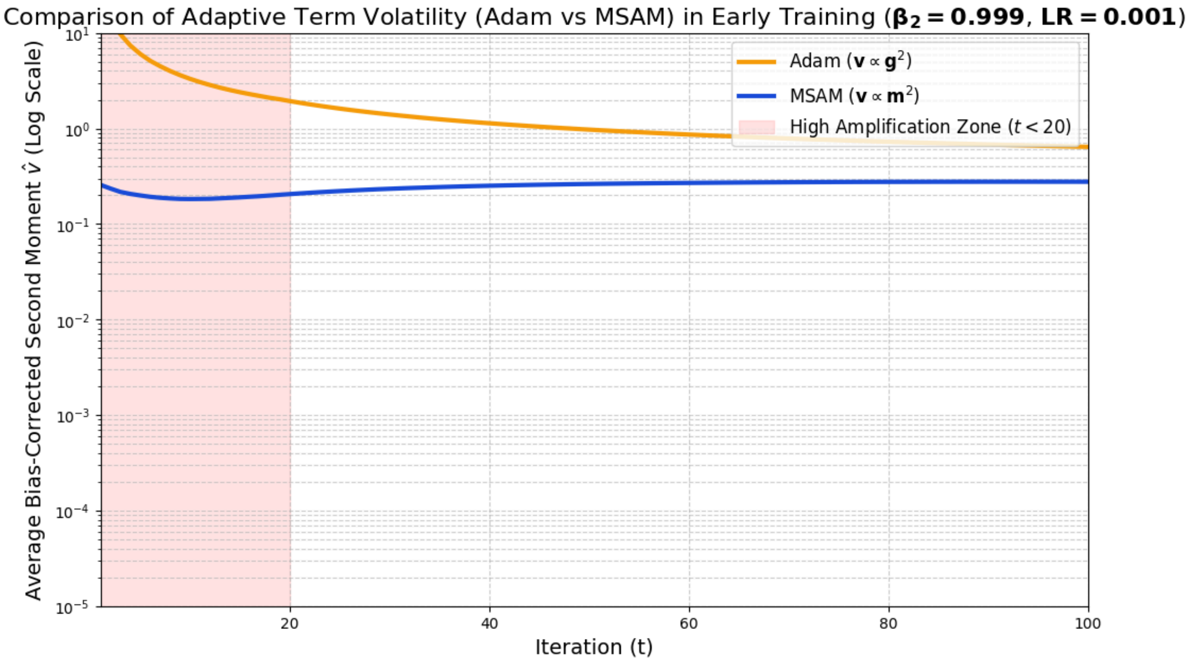
Iteration 4 :

Calculation	**Adam** $\mathbf{v}_3$ **/** $\hat{\mathbf{v}}_3$	**MADAM**
$\mathbf{m}_4$	**0.4168**	**0.4168**
Squared Term	$\mathbf{g}_4^2 = 1.0$	$\mathbf{m}_4^2 = 0.416^2 \approx \mathbf{0.174}$
Raw $\mathbf{v}_4$	**0.0685**	**0.011**
Correction C_4	$1/(1-0.99^5) \approx \mathbf{25.38}$	**25.38**

$\hat{v}_4$ **(Bias-Corr.)**	$25.38 \cdot 0.0685 \approx \mathbf{1.74}$	$25.38 \cdot 0.011 \approx \mathbf{0.28}$

Summary :

Step (t)	**Correction Factor (C_t)**	**Adam $\hat{v}_t$ (Unstable, g^2)**	**MSAM $\hat{v}_t$ (Smooth, m^2)**
1	**100**	**4**	**0.04**
2	**50.25**	**2.49**	**0.41**
3	**33.67**	**1.99**	**0.316**
4	**25.38**	**1.74**	**0.28**



On Rosenbrock :

After the initial turbulent start (where the bias correction C_t is huge), C_t approaches 1, and the $\hat{v}$ terms for both optimizers settle into a similar trend, tracking the overall curvature of the Rosenbrock function.

To properly illustrate the volatility you are asking about, we must examine the absolute difference in the critical early steps ($t < 20$) on a linear scale.

Analysis of $t = 1$ on Rosenbrock $X_0 = (-1.2, 1.0)$

Let's use the actual gradient from the Rosenbrock function at your starting point and the standard $\beta_2 = 0.999$:

- Initial Gradient $\mathbf{g}_1 \approx [-209.6, -88.0]$
- Bias Correction Factor $C_1 = 1/(1 - 0.999^1) = 1000$

Optimizer	Raw Input to V	Raw v1 (Scaled)	Bias-Corrected v^1
Adam	$\mathbf{g}_1^2 \approx 43900$	$43900 \times 0.001 =$	$43.9 \times 1000 = \mathbf{43,900}$
MSAM	$\mathbf{m}_1^2 \approx 439$	$439 \times 0.001 = 0$	$0.439 \times 1000 = \mathbf{439}$

The initial $\hat{\mathbf{v}}$ calculated by Adam is 100 times larger than MSAM's. This massive $\hat{\mathbf{v}}$ tells Adam to take a vanishingly small step, effectively halting progress, whereas MSAM takes a much more appropriate step based on the smoother momentum.