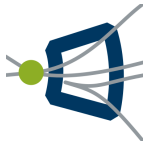


Data Analysis Block Course: Statistics and Machine learning



Adaptive Moment Estimation (Adam) - Solutions

Prof. Dr. Mikhail Mikhasenko

Assistants: Hendric Jonas, Christoph Tönnis



29 September 2026

Indico schedule

RUB Moodle

Use $f(x) = (x - 3)^2$, $x_0 = m_0 = v_0 = 0$, $\alpha = 0.1$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. All iterations below carry the unrounded values forward. Displayed decimals are rounded; the positive ϵ is retained in the updates.

Solution 1 - Iteration 1

1. **Gradient.** $g_1 = 2(0 - 3) = -6$.
2. **First moment.** $m_1 = 0.9(0) + 0.1(-6) = -0.6$.
3. **Second moment.** $v_1 = 0.999(0) + 0.001(-6)^2 = 0.036$.
4. **Bias correction.** The initial shrinkage is removed:

$$\hat{m}_1 = \frac{-0.6}{0.1} = -6 \quad \hat{v}_1 = \frac{0.036}{0.001} = 36$$

5. **Parameter update.** The negative gradient makes x increase:

$$x_1 = 0 - 0.1 \frac{-6}{6 + 10^{-8}} \approx 0.099999999833$$

To six decimals, $x_1 \approx 0.100000$. It is slightly smaller than 0.1 because $\epsilon > 0$. Ordinary gradient descent at the same learning rate would instead give $x_1 = 0 - 0.1(-6) = 0.6$; Adam rescales the gradient using its moment estimates.

Solution 2 - Iteration 2

The gradient is evaluated at the updated parameter x_1 . The moment estimates retain their previous values:

$$\begin{aligned} g_2 &= 2(x_1 - 3) \approx -5.800000000333 \\ m_2 &= 0.9(-0.6) + 0.1g_2 \approx -1.120000000033 \\ v_2 &= 0.999(0.036) + 0.001g_2^2 \approx 0.069604000004 \end{aligned}$$

The correction denominators are $1 - 0.9^2 = 0.19$ and $1 - 0.999^2 = 0.001999$:

$$\begin{aligned} \hat{m}_2 &= \frac{m_2}{0.19} \approx -5.894736842281 \\ \hat{v}_2 &= \frac{v_2}{0.001999} \approx 34.819409706787 \\ x_2 &= x_1 - 0.1 \frac{\hat{m}_2}{\sqrt{\hat{v}_2} + 10^{-8}} \approx 0.199897292585 \end{aligned}$$

Both corrected moments reflect the gradient history; $\hat{m}_2$ need not equal the current gradient g_2 .

Solution 3 - Iteration 3

Starting with $x_2 \approx 0.199897292585$:

$$\begin{aligned}g_3 &= 2(x_2 - 3) \approx -5.600205414830 \\m_3 &= 0.9m_2 + 0.1g_3 \approx -1.568020541513 \\v_3 &= 0.999v_2 + 0.001g_3^2 \approx 0.100896696692\end{aligned}$$

Now $1 - 0.9^3 = 0.271$ and $1 - 0.999^3 = 0.002997001$, so

$$\begin{aligned}\hat{m}_3 &= \frac{m_3}{0.271} \approx -5.786053658719 \\\hat{v}_3 &= \frac{v_3}{0.002997001} \approx 33.665886895650 \\x_3 &= x_2 - 0.1 \frac{\hat{m}_3}{\sqrt{\hat{v}_3} + 10^{-8}} \approx 0.299618476549\end{aligned}$$

The completed results, rounded to six decimal places, are:

Quantity	Iteration 1	Iteration 2	Iteration 3
Previous x_{t-1}	0.000000	0.100000	0.199897
g_t	-6.000000	-5.800000	-5.600205
m_t	-0.600000	-1.120000	-1.568021
v_t	0.036000	0.069604	0.100897
$\hat{m}_t$	-6.000000	-5.894737	-5.786054
$\hat{v}_t$	36.000000	34.819410	33.665887
Updated x_t	0.100000	0.199897	0.299618

As a check, the objective decreases from $f(x_0) = 9$ to approximately 8.410000, 7.840575 and 7.292060. The three iterates move toward the minimum at 3. These three steps illustrate the update; they do not establish a general convergence guarantee for Adam.

Solutions - Conceptual questions

1. **Answer b: exponentially weighted moving average of the gradients.** Recent gradients receive more weight than older ones. The signed first moment smooths the direction of the update and retains information from preceding iterations.
2. **Answer b: exponentially weighted moving average of the squared gradients.** This is a second **raw** moment, not a variance: it does not subtract the squared mean. It records gradient magnitude without cancellation between positive and negative gradients.
3. Squaring makes both signs contribute positively and gives large gradient magnitudes a larger contribution to v_t . For a fixed numerator, a larger $\sqrt{\hat{v}_t} + \epsilon$ reduces the magnitude of the update. The actual update depends on the numerator as well, so a large current gradient does not automatically imply a smaller step. In the first iteration, ignoring ϵ , the normalized direction is $g_1/|g_1|$ for $g_1 \neq 0$.

4. Starting at zero makes the early uncorrected moving averages too small in magnitude for a constant gradient. Expanding the recurrences gives

$$m_t = (1 - \beta_1) \sum_{k=1}^t \beta_1^{t-k} g_k$$

$$v_t = (1 - \beta_2) \sum_{k=1}^t \beta_2^{t-k} g_k^2$$

Their weights sum to $1 - \beta_1^t$ and $1 - \beta_2^t$. Dividing by these sums removes the effect of the zero initialization. For a constant gradient g , the corrected values are exactly g and g^2 . For a changing gradient, they remain weighted averages of the history, rather than exact values of the current gradient and its square.

Key concepts

Concept	Meaning
Gradient evaluation	Use x_{t-1} before updating the parameter
First moment	Smoothed signed gradient
Second raw moment	Smoothed squared gradient
Bias correction	Normalize the finite history's exponential weights
Adaptive rescaling	Divide the first moment by the square root of the second
Stabilizer	Add ϵ outside the square root