## **Running and Visualizing Adam**

Does it make sense to use a cumulative history of gradients?

1. We have already looked at a algorithms that make use of a history term
   1. Momentum based GD: Makes use of the history of the gradients
      1. Here, history is used to calculate the current update
   2. RMSProp: Makes use of the history of the square of the gradients
      1. Here, history is used to adjust the learning-rate
   3. Can we combine these two ideas?
   4. Yes, in the form of Adam, which uses both of those history terms
2. Adam
   * 1. This is very similar to the history that Momentum based GD maintains
     2. It’s a running sum of all the updates done
     3. This is similar to the history that RMSProp maintains
     4. It is used to regulate the learning-rate
     5. Here, the first history is used to make the update, ensuring that the history of derivatives is used to calculate the current update
     6. The second derivative is used to regulate the learning rate based on density or sparsity of the feature
   1. In addition to the above points, Adam performs bias correction by using the following equations
      1. It ensures that the training is smoother and also prevents erratic updates in beginning of training.
3. 