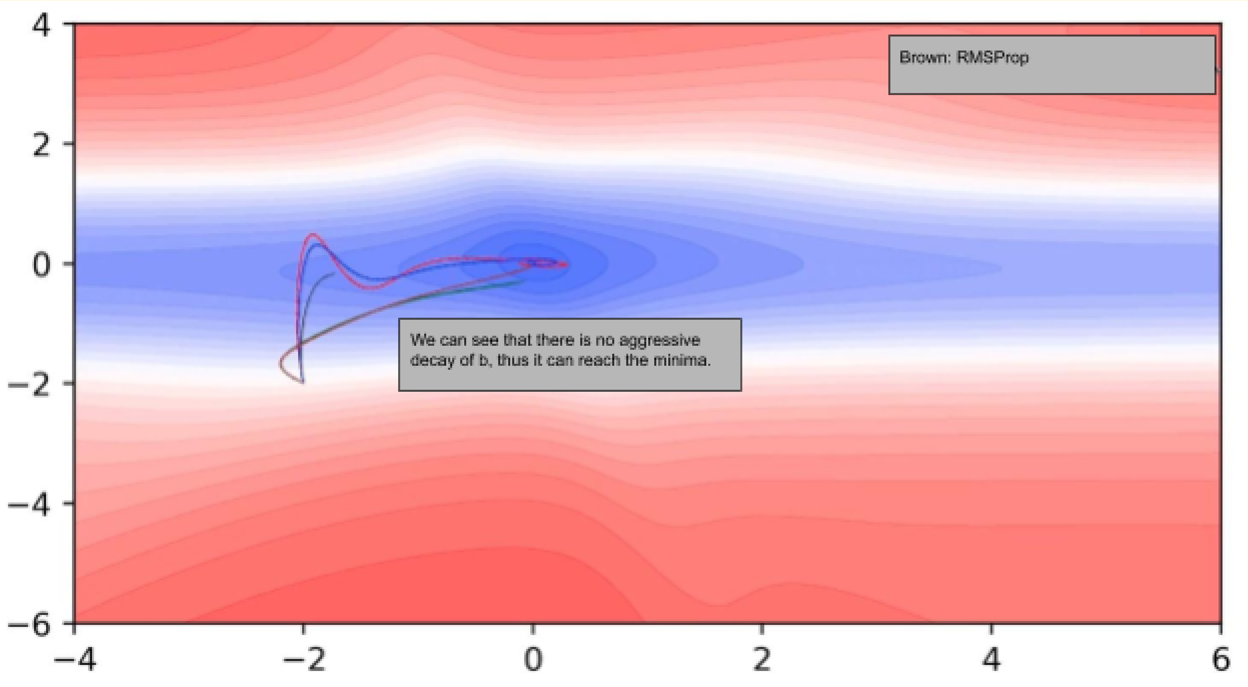
PadhAI: Variants of Gradient Descent

One Fourth Labs

Running and Visualizing RMSProp

Can we overcome aggressively decaying denominators?

1. Intuition: Why not decay the denominator and prevent its rapid growth?
2. We can consider the RMSProp algorithm
   a. $\upsilon_t = \beta * \upsilon_{t-1} + (1-\beta)(\nabla \omega_t)^2$
      i. Here we are taking an exponentially decaying sum
      ii. Let $\beta = 0.9$ and consider the 4th iteration υ_4
      iii. $\upsilon_0 = 0$
      iv. $\upsilon_1 = 0.1(\nabla \omega_1)^2$
      v. $\upsilon_2 = (0.9)(0.1)(\nabla \omega_1)^2 + 0.1(\nabla \omega_2)^2$
      vi. $\upsilon_3 = (0.9)^2(0.1)(\nabla \omega_1)^2 + (0.9)(0.1)(\nabla \omega_2)^2 + 0.1(\nabla \omega_3)^2$
      vii. $\upsilon_4 = (0.9)^3(0.1)(\nabla \omega_1)^2 + (0.9)^2(0.1)(\nabla \omega_2)^2 + (0.9)(0.1)(\nabla \omega_3)^2 + 0.1(\nabla \omega_4)^2$
      viii. We can see from this that our value υ_4 is much smaller than in the case of Adagrad, due the history of the gradients being multiplied by the decay ratio.
      ix. The relative difference between dense and sparse features is still maintained.
   b. $\omega_{t+1} = \omega_t - \frac{\eta}{\sqrt{(\upsilon_t)} + \varepsilon} \nabla \omega_t$
      i. This is the same as in Adagrad
3. Let's visualise RMSProp in 2D



We can see that there is no aggressive decay of b, thus it can reach the minima.

4. Adagrad got stuck when it was close to convergence (it was no longer able to move in the vertical (b) direction because of the decayed learning rate)
5. RMSProp overcomes this problem by being less aggressive on the decay