Due: 21 Dec 2020

The code for this problem can be found at this Github link.

Suppose we want to approximate the function $g(x) = 1 - \cos(x)$ on the interval $[0, \pi/2]$ with the function $\text{ReLU}(ax - b)$ where a and b are to be determined. We take 6 training points $x_j = \pi j/10$, $j = 0, 1, 2, 3, 4, 5$ and set up the following loss function:

$$f(a, b) = \frac{1}{12} \sum_{j=0}^{5} [\text{ReLU}(ax_j - b) - g(x_j)]^2.$$

1. **Stationary points.** The set of stationary points of f (i.e., the set of points where $\nabla f = 0$) consists of the global minimizer and a flat region. Describe this set analytically using equalities and inequalities and show it in a figure. Provide an analytic formula for the global minimizer of f. What is the global minimum of f?

Solution. To begin, we compute the partial derivatives of f.

$$\begin{aligned}
\frac{\partial f}{\partial a} &= \frac{1}{6} \sum_{j=0}^{5} x_j [\text{ReLU}(ax_j - b) - g(x_j)] \text{ReLU}'(ax_j - b) \\
\frac{\partial f}{\partial b} &= -\frac{1}{6} \sum_{j=0}^{5} [\text{ReLU}(ax_j - b) - g(x_j)] \text{ReLU}'(ax_j - b)
\end{aligned}$$

where the derivative of the ReLU is defined as

$$\text{ReLU}'(x) = \begin{cases} 1 & x > 0 \\ 0 & x < 0. \end{cases}$$

In order to guarantee that the partials are equal to 0, we look for values of a and b that make the terms in the interior equal to 0. First, we study the ReLU' factor and notice that

$$\text{ReLU}'(ax_j - b) = 0 \iff ax_j - b < 0$$

so $b > ax_j$ for all j in order to zero out every term in both sums. Since $x_0 = 0$, we know $b > 0$. When $a > 0$, we need $b > \max_j ax_j = \frac{\pi}{2} a$ and when $a < 0$, we need $b > \min_j ax_j = 0$ since the inequality flips when $a < 0$. Naturally, when $a = 0$ we simply need $b > 0$ (as $\text{ReLU}(-b) = 0$ for $b > 0$), which squares up with the above characterization. This flat region can be seen in Figure 1 below. To find the minimizer, **UNFINISHED**

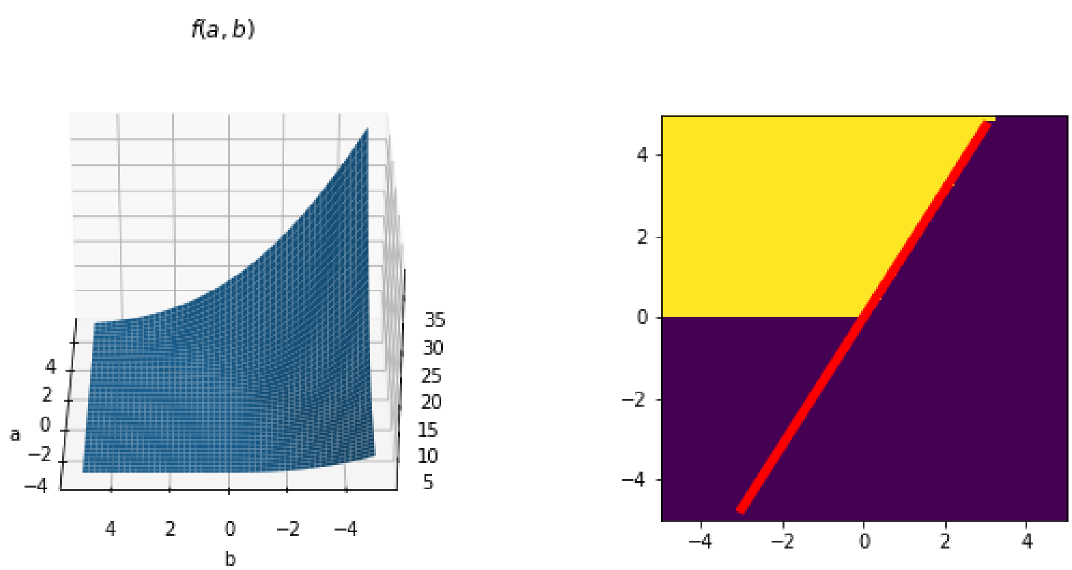

Figure 1: Left: surface of $f(a,b)$ near the origin. Right: flat region $\nabla f = 0$. The yellow region is the flat region, the purple region is the rest of the function, and the red line is the boundary $b = \frac{\pi}{2}a$.

2. **Gradient descent.** Take $a = 1$ and $b = 0$ as the initial guess for gradient descent with constant stepsize. What is the minimal stepsize α^* such that the iterates end up in the flat region? Suppose we take $\alpha = 0.99\alpha^*$ and run gradient descent. Will the iterates approach the global minimizer? Either way, explain why. Propose a stepsize trying to make it as large as possible such that the iterates will necessarily converge to the global minimizer and give a rationale for your choice.

Solution. The gradient descent iteration is $x_{k+1} = x_k - \alpha\nabla f(x_k)$. Here, $x = [a,b]^\intercal \in \mathbb{R}^2$.**UNFINISHED**

3. **Stochastic gradient descent.** As above, take $a = 1$ and $b = 0$ as the initial guess. Use a simple stochastic gradient descent with a single training point chosen randomly for approximating the gradient of f at each step. Find a strategy for stepsize reduction such that the stochastic gradient descent will converge to the global minimizer.

Solution. See `prob1.ipynb` for my stochastic gradient descent code. I tested the following schedules:

- Reciprocal: $\alpha_k = L/k$
- Power: $\alpha_k = L2^{-k-1}$
- Exponential $\alpha_k = Le^{-rk}$ where r controls the speed of the decay
- Drop: $\alpha_k = Lr^{\lfloor k/M \rfloor}$ where M is the number of steps we take per drop

In all of these, L is a constant initial learning rate. These all satisfy $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$. **UNFINISHED**