① vanilla gradient descent

→ we want to find gradient using the entire dataset

The ~~loss is~~ update rule is

$$w_{t+1} = w_t - \eta \frac{(X^T \cdot (X \cdot W - Y))}{N}$$

this algo takes time to converge but converges to minimum

sgd

here we take only one point

$$w_{t+1} = w_t - \eta \frac{2 \cdot x^T \cdot (xw - y)}{1}$$

(for a single datapoint)

for our dataset this doesnt ~~converge~~ converge for our dataset on multiple η and initial weight vectors

minibatch is picking some small batch size from our data set → out of 1000 points we pick for eg 32

this converges slower but doesnt have as much noise as sgd.

As we change batch size conversion becomes slower

Checked for 1, 8, 16, 32, 100, 1000

Dataset

$X = \text{random}(1500, 15)$

$y =$ some linear combination of columns

(2) We want to find the optimal η

Bisection line : $\eta \in [0, 1]$ are our high lo.

We want to find $\frac{\text{high}-\text{low}}{2}$

and keep doing these for intervals that perform best.

We use the loss $= e^x$

gradient e^x

$\boxed{\text{bisection says } \eta = 2.9 \times 10^{-8}}$

loss converges -

golden ~~selection~~ section rule :

using golden ratio to determine the η

$\varphi = \frac{1+\sqrt{5}}{2}$ $\quad \rho = \frac{2}{1+\sqrt{5}}$

0.618 $\quad 0.382$

$m_1 = a + 0.382(b-a)$

$m_2 = a + 0.618(b-a)$

this converses faster than bisection

$\eta = 0.276$

armijo:

~~actual a objective function &~~

Starting point $\beta = 0.9$

predicted decrement: $\alpha \cdot \nabla f(x)^T \cdot \nabla f x$

actual decrease $f(x) - f(x - \alpha \nabla f(x))$

Case I actual decrease $\geq \beta *$ pred decrease

we $\alpha = \eta$

else $\alpha = \alpha * 0.5$ & rerun

armijo $\eta = 0.5$

(3) $f(x) = x^2 \cos x + \sin x - x$

Constant step size → slow convergence

since we go slow on plains &

too fast on valleys

Decaying step size

$$\eta = \frac{\eta_0}{k}$$

faster convergence

$\boxed{\eta_0 = 0.03 \text{ to start with}}$

~~Bold driver~~ : → $\boxed{\eta_{new} = \frac{1}{1+kt}\eta_0}$

$\eta_0 = 0.01$

Bold driver:

Variable step size depending on norm of gradients.

if $e^{(x - \eta \cdot \text{gradient})} < e^{x}$;

$\eta \rightarrow 1.1 \cdot \eta$

else $\eta \rightarrow 0.5\eta$

Converges faster as we cover valleys slower

& else we get slower.