08 - Optimization

Ethan Graham

20th June 2023

Basics

Definition of 1-D Derivative

$f'(x_0) = \frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{f(x_0+\Delta x)-f(x_0)}{\Delta x} \rightarrow$ this is the slope of the tangeant to f at x_0

Convex vs. Non-Convex

- **Convex:** Only one minimum. Second derivative is ≥ 0
- **Non-Convex:** There are several local minima and one global minimum

Minimizing a Convex Function

To minimize a convex function it suffices to find x^* such that $f'(x^*) = 0$. If we know f' then we can often solve this as an equation.

Example: $f(x) = x^2 + 2x + 2$ which yields $\frac{df(x)}{dx} = 2x + 2$ and thus we can easily compute $f'(x^*) = 0 \Leftrightarrow x^* = -1$ our **global minimum** for this function.

When the minimum cannot be found with a closed-form solution (like above) we use the derivative: $x_k = x_{k-1} - \eta \frac{df(x_{k-1})}{dx}$ Where η is the step-size of each iteration. We often refer to this as *learning rate*

How many iterations are required for finding the minimum is determined by initialization x_0 and choice of η

This method provides no guarantees when applied to a non-convex function

This method applies to multivariate functions as well *recall cs328.*

Minimum of a multivariate convex function: $\nabla f(\mathbf{x}^*) = \mathbf{0}$

Gradient Descent

Is the same as one of the previous formulae, but adapted for N-dimendional space. $\mathbf{x}_k \leftarrow \mathbf{x}_{k-1} - \eta \nabla f$ We set a *stopping criterion* of some sort. - Thresholding a change in function value - Thresholding a change in parameters i.e. $\|\mathbf{x}_k - \mathbf{x}_{k-1}\| < \delta$

Theoretical Justification

- $f(\mathbf{x} + \mathbf{dx}) \approx f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{dx}$
- $f(\mathbf{x} - \eta \nabla f(\mathbf{x})) \approx f(\mathbf{x}) - \eta \|\nabla f(\mathbf{x})\|^2 < f(\mathbf{x})$

It is very important to choose a good η value. - η too large: we can overstep the local minimum - η too small: may converge very slowly

Conjugate Gradient

Take the search direction to be a weighted average of the gradient vector and the previous search directions $\rightarrow$ **faster convergence**