WEEKLY PARTICIPATION 7: THE IMPORTANCE OF ACCOUNTING FOR CURVATURE

Consider the simple but illustrative optimization problem

$$\boldsymbol{\omega}_\star = \text{argmin}_{\boldsymbol{\omega}} \, \tfrac{1}{2}\|\mathbf{H}\boldsymbol{\omega}\|_2^2,$$

where $\mathbf{H} = \begin{pmatrix} \frac{1}{10} & 0 \\ 0 & 100 \end{pmatrix}$.

- Verify that the iterates of gradient descent using fixed step-size α and starting at $\boldsymbol{\omega}_0 = \mathbf{1}$ satisfy

$$\boldsymbol{\omega}_t = \left(\mathbf{I} - \alpha\mathbf{H}^2\right)^t \mathbf{1} = \begin{pmatrix} \left(1 - \frac{\alpha}{100}\right)^t \\ (1 - \alpha \times 10^4)^t \end{pmatrix}.$$

- According to our theory for gradient descent, what step size should we take to ensure linear convergence?
- In fact, the stepsize $\alpha = \frac{2}{10^{-2}+10^4}$ guarantees the fastest asymptotic rate of convergence for gradient descent *for this problem*. Verify that this choice of stepsize performs better than the standard choice from the theory, by giving a t for which $\|\boldsymbol{\omega}_t - \boldsymbol{\omega}_\star\|_2^2$ is smaller with this stepsize than the standard step size. (Yes, this t is large).

  How does the optimal step size for this problem compare to the step size from the theory for gradient descent?
- Using the optimal step size, how many steps of gradient descent would be required to get one digit of accuracy in the solution? What if we use the non-optimal stepsize from the theory? There's an important lesson here about the choice of stepsizes: the stepsize that is asymptotically optimal may perform poorly during the initial iterates. In practice, one prefers to change the stepsize over time.
- What is the first iterate of Newton's method?

Discussion. As you can imagine, since methods which do not account for curvature perform poorly on this convex, smooth, and unconstrained problem, they will probably perform poorly in general for more complicated problems. This motivates the introduction of optimization algorithms which try to account for curvature while still trying to be inexpensive, by only using gradient information.

Observe that, for this problem, Newton's method corresponds to a single step of "gradient descent" **where we scaled the gradient by different amounts in each coordinate**. This idea of independently choosing stepsizes for each coordinate is key in the most popular adaptive gradient descent algorithms that we will look at.