WEEKLY PARTICIPATION 5: 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

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

- Using the above result, find the stepsize α that guarantees the fastest rate of convergence for gradient descent. What is that α, how does it relate to the smoothness constant (β) of the objective, and comment on this relationship.

- Argue that the corresponding rate of convergence using the optimal α satisfies $\|\boldsymbol{\omega_t} - \boldsymbol{\omega_\star}\|_2^2 = \left(1 - \frac{1}{10^6}\right)^{2t}$.

  This is atrocious: we would need to use more than 10^6 steps of gradient descent to get one digit of accuracy in the minimizer!

- Compute the condition number of the objective. Is the convergence rate of steepest descent on this problem significantly better than that of gradient descent?

- 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 algorithms that we will look at.