Exercise session notes - Week 9

Newton Method + Self-concordant

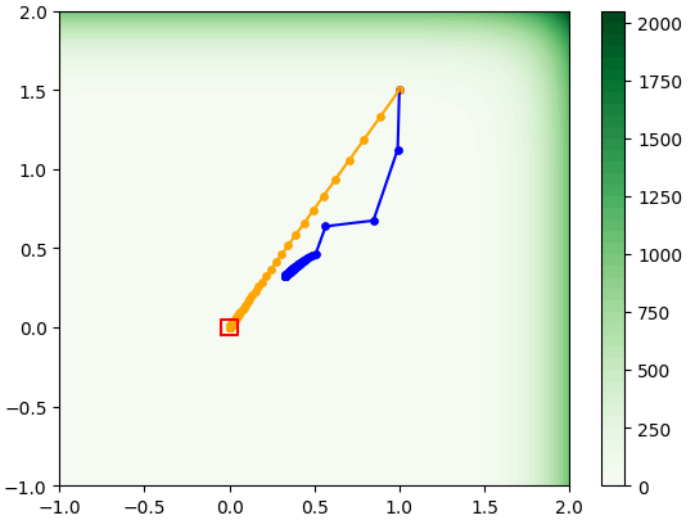
This week we studied the Newton Method and compared it with the Gradient Descent. Let f be a m-strongly convex function, i.e.

$$\nabla^2 f(x) \succeq mI$$

with Hessian L-Lipschitz continuous, i.e.

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\| \leq L \cdot \|x - y\|$$

Recall that in the Gradient Descent we use a step direction of $\Delta x = -\nabla f(x)$, so we move in the direction in which the function is decreasing the most. The Newton method instead uses a step direction of $\Delta x = -(\nabla^2 f(x))^{-1} \cdot \nabla f(x)$, so we move faster when the curvature of the function is small (the Hessian is "small") and we move slower when the curvature of f is large (the Hessian is "large"). So in particular we can distinguish two phases of the algorithm: one which is usually slow, just to reach a region that is very close to the optimal value, called Damped Phase; and the other one which is quadratically convergent and moves very fast towards the minimum, called Pure Newton Phase.
The main difference between Gradient and Newton method can be seen in the function $f(x) = x^{10} + y^{10}$:



Here the optimal value is at $x^* = (0, 0)$, we start the algorithms with initial point $x^{(0)} = (1, 1.5)$, the blue line corresponds to the Gradient Descent and the orange line corresponds to the Newton Method. Here we can clearly see that the Newton Method moves slower than the Gradient descent until some range (Damped Phase) and then faster to reach the optimal value (Pure Newton Phase). In both cases we stopped the algorithms after 100 steps.
We distinguished the different Phases also in the runtime of Newton Method: it reaches $f(x^{(k)}) - f^* \leq \varepsilon$ after

$$\underbrace{\frac{f(x^{(0)}) - f^*}{\gamma}}_{\text{Damped}} + \underbrace{\log\log\frac{2m^3}{L\varepsilon}}_{\text{Pure Newton}}$$

steps.
Instead of considering strongly convex functions with Lipschitz continuous Hessian, we additionally studied self-concordant functions.

Definition. A function $f : \mathbb{R} \to \mathbb{R}$ is self concordant if it is convex, 3 times differentiable and

$$|f'''(x)| \leq 2(f''(x))^{3/2} \qquad \forall x$$

A function $f : \mathbb{R}^n \to \mathbb{R}$ is self concordant if it is convex, 3 times differentiable and the function

$$f_{x,d} : t \mapsto f(x + td)$$

is self concordant for any $x, d \in \mathbb{R}^n$.

Remark. For self-concordant functions f we have

$$f(x) - f^* \leq \lambda(x)^2$$

where $\lambda(x) = \nabla f(x)^\top \cdot (\nabla^2 f(x))^{-1} \cdot \nabla f(x)$ is the Newton Decrement. Additionally the runtime of Newton Method applied to f is

$$\underbrace{\frac{f(x^{(0)}) - f^*}{\gamma}}_{\text{Damped}} + \underbrace{\log\log\frac{1}{\varepsilon}}_{\text{Pure Newton}}$$

Why did we consider such a strange class of functions? Assume we have a function f that is (1) m-strong convex and (2) with L-Lipschitz continuous Hessian. Let us define the third derivative of f in direction v as

$$f'''(x)[v] = \lim_{t \to 0} \frac{\nabla^2 f(x + tv) - \nabla^2 f(x)}{t}$$

Then by (2) we have

$$\|f'''(x)[v]\| \leq L \|v\|$$

and using scalar products with a vector w we get

$$\left|w^\top f'''(x)[v]w\right| \leq L \|v\| \|w\|^2$$

Now, the left hand side of the inequality is invariant under affine transformations of the variables, but the right hand side is not. Therefore we can choose the norm $\|\cdot\|_{\nabla^2 f(x)}$ and pick $v = w$ so that

$$\left|v^\top f'''(x)[v]v\right| \leq L \cdot \left|v^\top \nabla^2 f(x) v\right|^{3/2}$$

This leads to our definition of self concordant functions.