Optimization Algorithms
Weekly Exercises 8

Marc Toussaint
Learning & Intelligent Systems Lab, TU Berlin
Marchstr. 23, 10587 Berlin, Germany

Winter 2024/25

1 Convergence of Stochastic Gradient Descent

For a cost function $f(w) = \frac{1}{n}\sum_{i=1}^n f_i(w),\ w \in \mathbb{R}^d$, we are interested to show that, when iterating $w_{k+1} \leftarrow w_k - \alpha_k \nabla f_i(w_k)$ for random i, the gradient ∇f goes to zero. The typical assumption we make is Lipschitz continuity of the gradient, namely there exists a Lipschitz constant L such that

$$\|\nabla f(w) - \nabla f(\bar w)\| \le L\ \|w - \bar w\|\ ,$$

where $\|w\| = \sqrt{w^2}$ is the L_2-norm.
Based on this assumption, show that

a) For any $\delta \in \mathbb{R}^d$, the Hessian $\nabla^2 f(w)$ fulfills $\|\nabla^2 f(w)\delta\| \le L\|\delta\|$. (This can also be written as $\|\nabla^2 f(w)\|_2 \le L$, also means that the largest eigenvalue of $\nabla^2 f$ is $\le L$, and we have an upper bound on curvature.)

b) We have
$$f(w) \le f(\bar w) + \nabla f(\bar w)^\top (w - \bar w) + \frac{1}{2}\ L(w - \bar w)^2$$

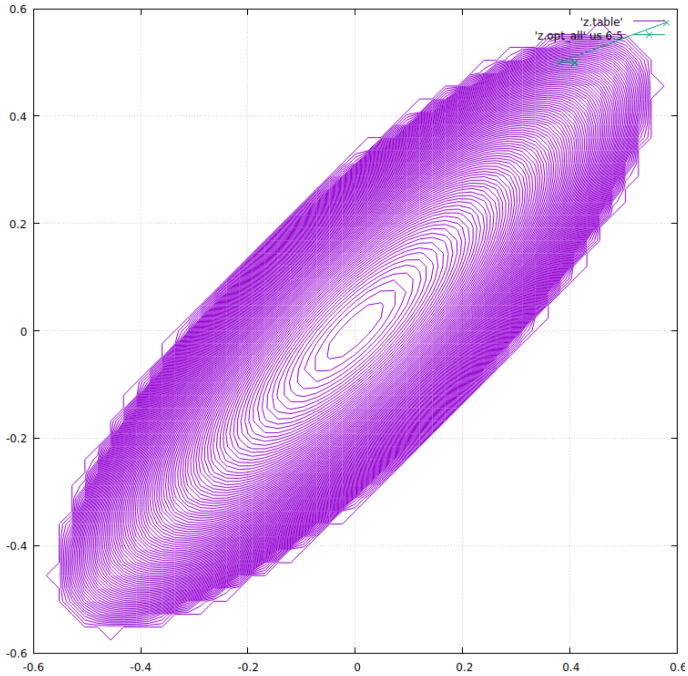
c) We have
$$\mathbb{E}\{f(w_{k+1})\} \le f(w_k) - \alpha_k \|\nabla f(w_k)\|^2 + \frac{1}{2}\ \alpha_k^2\ L\ \mathbb{E}\big\{\|\nabla f_i(w_k)\|^2\big\}$$

(We then often assume a given variance $\mathbb{E}\big\{\|\nabla f_i(w_k)\|^2\big\} = \sigma^2 + \|\nabla f(w_k)\|^2$ of the stochastic gradient and can continue convergence analysis as on the lecture slide.)

2 Bound Constraints

Consider the problem:
$$\min_{x\in\mathbb{R}^2} \frac{1}{2} x^\top A x \ \ \text{s.t.}\ \ x_2 \ge \frac{1}{2}\ , \quad \text{with } A = \begin{pmatrix} 200 & -160 \\ -160 & 200 \end{pmatrix}$$

Here a plot of isolines, and at the top right in green, a few steps of a Newton method that properly handles bound constraints:



a) Analytically compute the optimum for this problem. You may assume the constraint active. (For arbitrary positive definite A, the specific numbers are not important.)

b) Assume we are at location $x = (0, 1)$. In which direction does the gradient $-\nabla f$ point? (First compute it analytically, then plug in the 160,200 numbers of A). And in which direction does the Newton step $-\nabla^2 f^{-1} \nabla f$ point? (This should be obvious, without much computation.)

c) Assume we initialize our bound constrained Newton method (slide 13 of lecture 11) at $x = (0, 1)$, how many Newton iterations (where each iteration does line search in the determined direction δ), will it need until convergence. Illustrate roughly, where each step moves to.

d) Let us define $r(x_1) = f(x_1, x_2 = \frac{1}{2})$, which is the cost function on the hyperplane only. Given any point x_1 on the hyperplane, what is the Newton step within the hyperplane w.r.t. x_1? Is this the same as the (clipped) Newton step for $f(x_1, x_2)$ when deleting the off-diagonal terms from A (as our method does)?