METHODS OF NONLINEAR OPTIMIZATION: HW#1

(1) The gradient descent method applied to an objective function that is not strongly convex can be very slow.

Let $f(x) = x^4$ when $|x| \leq 1$ and define $f(x)$ to be a quadratic polynomial when $|x| > 1$ so that f is C^2 on the whole line.

(i) Prove that $f \in C_L^{1,1}(\mathbb{R})$ with $L = 12$. Also prove that f is convex but not strongly convex (see Page 144, Problem 7.26 for a precise definition.)

Apply the gradient descent method to f with any initial guess $x_0 \in [-1, 1]$ and step size $1/L$. Call the iterates x_k.

(ii)* Prove that x_k goes to zero sub-linearly, i.e. it goes to zero but slower than ρ^k for any $\rho \in (0, 1)$.

For 40% extra credit, figure out the exact asymptotic of x_k, i.e. find a sequence θ_k in terms of elementary functions (e.g. $2/k$, $1/(k \log k)$) such that $x_k \sim \theta_k$, or $\lim_{k\to\infty} x_k/\theta_k = 1$.

For 60% partial credit, carry out a careful numerical experiment to (empirically) determine out the exact asymptotic of x_k. (Suggestion: do a log-log plot of the error versus k, using the `loglog()` function in Matlab.)

(2) (i) Prove that all three versions of the gradient descent method are invariant under an orthogonal change of coordinates, i.e. if $f : \mathbb{R}^n \to \mathbb{R}$ and $\bar{f} : \mathbb{R}^n \to \mathbb{R}$ is defined by $\bar{f}(\bar{x}) = f(U\bar{x} + v)$, for an orthogonal matrix U and a vector v, then the same gradient descent method applied to f and $\bar{f}$ with initial guesses x_0 and $\bar{x}_0 = U^T(x_0 - v)$ (respectively) results in iterates x_k and $\bar{x}_k$ that are related by $\bar{x}_k = U^T(x_k - v)$ for any $k \geq 0$.

(ii) This would in particular mean that the **rate of convergence** of the gradient descent method is insensitive to an orthogonal change of coordinates applied to the objective function. Why?

(iii) The invariance property would also imply that if we want to analyze the **rate of convergence** property of the gradient descent method applied to a quadratic polynomial $f(x) = \frac{1}{2}x^T Ax - b^T x + c$, one can without loss of generality assumes that A is diagonal and $b = 0$. (Of course, the constant c obviously plays no role, so we can also assume $c = 0$.)

(iv) Illustrate that, however, the gradient descent method can be very sensitive to affine change of coordinates. (For this purpose, it suffices to work with quadratic polynomials, and you can illustrate it with any one of the three versions of the gradient method.)

(3) Consider applying the gradient descent method to minimize $f(x) = \frac{1}{2}x^T Ax - b^T x + c$, $A \succ 0$, with a constant step size $\bar{t}$. Show that the rate of convergence, for almost all initial guesses x_0, is dependent only on $\bar{t}$, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$.

Date: April 6, 2020.

(i) For a given $A \succ 0$, what choice of $\bar{t}$ gives the fastest rate of convergence? From an algorithmic point of view, is it practical use such an optimal step size $\bar{t}$?

(ii) How is this rate of convergence compared to the rate of convergence to the one given by Lemma 4.11 in [Beck] pertaining to the gradient descent method with **exact line search**?

Thomas Yu, Department of Mathematics, Drexel University, 3141 Chestnut Street, 206 Korman Center, Philadelphia, PA 19104, U.S.A.

E-mail address: yut@drexel.edu