Assignment 4

Corrections: [1] Formula (3) of Exercise 4 has been corrected to have $J_n^x)^{-1}$ instead of just J. [2] formula (3) corrected again to have $g(x_n) - a$

1. Consider applying the simple Newton method to minimizing

$$f(x) = \sqrt{1 + x^2} \; .$$

"Simple" means the step size is always equal to one. Show that f is strictly convex but not "uniformly convex" ($|f''(x)| \to 0$ as $x \to \infty$) Show that $x_n \to x_* = 0$ if x_0 is small enough but $|x_n| \to \infty$ if x_0 is too large.

2. Consider the problem of fitting a time series to a sum of simple oscillations with frequencies ω_j and amplitudes A_j. The loss function is

$$L(A_1, \omega_1, \cdots, A_d, \omega_d) = \sum_{j=1}^{m} \left(Y_j - \sum_{k=1}^{d} A_k \sin(\omega_k t_j) \right)^2 \; . \qquad (1)$$

Suppose that L has a local minimizer with distinct frequencies ($\omega_j \neq \omega_k$ if $j \neq k$). Show that L has more than 100 local minima if $d > 4$ frequencies are used and more than 1000 local minima if more than $d > 6$ frequencies are used.

3. The *Gauss Newton* method is a way to solve optimization problems involving sums of squares. Such problems come up in data fitting and modeling. Suppose a model has parameters $x = (x_1, \cdots, x_d)$ and makes predictions

$$y_j = g_j(x) \; , \quad j = 1, \cdots, m \; .$$

Suppose you measure Y_j and want to identify (estimate, learn) the values of the parameters x. Least squares parameter estimation is

$$x_* = \arg\min_x \sum_{j=1}^{m} \left(Y_j - g_j(x) \right)^2 \; .$$

The loss function is the sum of squares error

$$f(x) = \sum_{j=1}^{m} \left(Y_j - g_j(x) \right)^2 = \left(Y - g(x) \right)^T \left(Y - g(x) \right) \; . \qquad (2)$$

The Gauss Newton method is like Newton's method in that it uses a local model of the loss function. For Gauss Newton, it replaces the nonlinear functions g_k with linear approximations

$$g_j(\overline{x} + x') \approx g_j(\overline{x}) + \sum_{j=1}^{d} \frac{\partial g_j}{\partial x_k}(\overline{x}) \left(x'_k - \overline{x}_k \right) \; .$$

In matrix/vector form, let J be the jacobian of g and write

$$g(\overline{x} + x') \approx g(\overline{x}) + J(\overline{x})\,(\,x' - \overline{x}) = \widetilde{g}_{\overline{x}}(x')\ .$$

Assume that $m > d$ (more data than parameters) and that J has rank d. We define $\widetilde{g}_{\overline{x}}(x')$ using $\widetilde{g}_{\overline{x}}$ instead of g in (2). The Gauss Newton iteration is

$$x_{n+1} = \arg\min_x \widetilde{f}_{x_n}(x)\ .$$

Answer the following questions about the Gauss Newton iteration.

(a) Define the search direction as $p_n = x_{n+1} - x_n$. Is p_n a descent direction for f at x_n?

(b) Is the Gauss Newton method affine invariant?

(c) Is the Gauss Newton method locally quadratically convergent?

(d) Does the Gauss Newton method have faster local convergence if the model fits the data better?

4. There is a form of Newton's method for solving systems of nonlinear equations. Suppose you have d equations $g_j(x) = a_j$ involving d unknowns x_k. Suppose that $g(x_*) = a$ and the jacobian $J(x_*) = Dg(x_*)$ is non-singular. Newton's method is the iteration

$$x_{n+1} = x_n - J(x_n)^{-1}\left[\,g(x_n) - a\right]\ . \qquad (3)$$

(a) Consider the case $d = 1$. Show that the Newton iteration is equivalent to the geometric method from Calculus, where you try to find x_* with $g(x_*) = a$ using the intersection of the tangent line at x_n with the line $y = a$.

(b) Show that if you apply this Newton's method to $g(x) = \nabla]f(x) = 0$, you get the Newton's method for optimization.

(c) Show that Newton's method (3) is locally quadratically convergent as long as $J(x_*)$ is nonsingular.

5. *Equality constraints* are equations that a point $x \in \mathbb{R}^d$ must satisfy exactly in order to have $x \in \mathcal{F}$ (the feasible set). This exercise describes gradient descent for equality constrained optimization. We suppose the equality constraints involve equations

$$g_j(x) = a_j\ ,\quad j = 1, \cdots, m\ .$$

We suppose the g_j are differentiable and have gradient vectors that are linearly independent. This is expressed in matrix terms using the function g that takes $\mathbb{R}^d$ to $\mathbb{R}^m$.

$$g(x) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_m(x) \end{pmatrix}\ .$$

The jacobian of this function is an $m! \times d$ matrix $B(x) = Dg(x)$. We suppose $\mathcal{F}$ is defined by the constraints $g_j(x) = a_j$ only:

$$\mathcal{F} = \{\, x \mid g(x) = a \,\} \ .$$

We always suppose $\text{rank}(B(x)) = m$ for all $x \in \mathcal{F}$. A vector $p \in \mathbb{R}^d$ is *tangent* to $\mathcal{F}$ at a point $x \in \mathcal{F}$ if

$$\frac{d}{ds} g(x + sp)\Big|_{s=0} = 0 \ .$$

You may assume the following theorem, which is related to the *implicit function theorem*: If p is tangent to $\mathcal{F}$ at x and if s is small enough, then there is $y(x) \in \mathcal{F}$ with $y(s) = x + sp + O(s^2)$. This implies that

$$\frac{d}{ds} y(s)\Big|_{s=0} = p \ . \tag{4}$$

The vector space of all p tangent to $\mathcal{F}$ at x is the *tangent space* to $\mathcal{F}$ at x and is written T_x. The equality constrained optimization problem is

$$\min_{x \in \mathcal{F}} f(x) \ .$$

(a) What is the dimension of T_x, assuming that $B(x)$ has full rank?

(b) The *orthogonal projection* of $\nabla f(x)$ onto T_x is the tangent vector p that solves

$$\min_{p \in T_x} \|p - \nabla f(x)\|_2^2 \ .$$

Find a way to find this p using the QR factorization of B.

(c) Show that it is possible to express the projection p from part (b) as

$$p = \nabla f(x) - \sum_{j=1}^{m} \lambda_j \nabla g_j(x) \ .$$

Find the *normal equations* that the vector λ satisfies.

(d) The following are related. Take the first ones as hints for the last. Show that the projection p at x_* is zero if

$$x_* = \arg \min_{x \in \mathcal{F}} f(x) \ .$$

Show that x is not a local minimizer of f in $\mathcal{F}$ if $p \neq 0$. Show that $-p$ is a descent direction for f within $\mathcal{F}$ at x if $p \neq 0$. Show that if $p \neq 0$ and s is small enough, then $f(y(-s)) < f(x)$.

(e) Show that if x_* is a constrained minimizer of f, then there are numbers λ_j (*Lagrange multipliers*) so that

$$\nabla f(x_*) = \sum_{j=1}^{m} \lambda_j \nabla g(x_*) \ .$$

(f) *Nonlinear projection* to $\mathcal{F}$ means finding $y \in \mathcal{F}$ that is close to $x + sp$. If the functions g defining $\mathcal{F}$ are nonlinear then $\mathcal{F}$ is likely to be curved so $x + sp$ is not in $\mathcal{F}$. The "projection" is not uniquely determined if $\mathcal{F}$ is curved. One kind of projection looks for variables w_j so that

$$y = x + sp + \sum_{j=1}^{m} w_j \nabla g_j(x) \in \mathcal{F} \ .$$

The goal is to find w so that $g(y) = a$. This is a nonlinear system of equations. There are m equations and m unknowns (the w_j), which may be written as $h(w) = 0$, where h represents m functions of m variables. Show that Newton's method for finding w needs only first derivatives. Write $w_{n+1} = H(w_n)$ as the Newton iteration mapping. Show that if $\|w\|$ and s are small then $\|D_w H(w)\|$ is small so that linearized analysis suggests that the Newton iteration succeeds in finding w with $h(w) = 0$, using initial guess $w_0 = x + sp$.

(g) Combine these parts to suggest a gradient descent method for finding x_* by searching on $\mathcal{F}$. The algorithm will have an *outer iteration*, which goes from $x_n \in \mathcal{F}$ to $x_{n+1} \in \mathcal{F}$ and has $f(x_{n+1}) < f(x_n)$ unless the projection of $\nabla f(x_n)$ onto T_{x_n} is zero. Each outer step involves an *inner iteration* that goes from $x_n - s_n p_n$ to $x_{n+1} \in \mathcal{F}$ using nonlinear projection.