TMA4180 Optimization I
Project 1: Unconstrained optimization

Anton Evgrafov

February 12, 2018

1 Introduction

In this project we will take a look at a problem of fitting a given model to some observations or data. Our data is given as a set of points $z_i \in \mathbb{R}^n$, together with "labels" $w_i \in \{-1, +1\}$, $i = 1, \dots, m$. We will be interested in separating the points with labels $w_i = +1$ from those with labels $w_i = -1$ with the help of a second order polynomial.

1.1 Model 1

Any symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$ and a vector $c \in \mathbb{R}^n$ define an ellipsoid

$$S_{A,c} := \{\, z \in \mathbb{R}^n \mid (z-c)^{\mathrm{T}} A (z-c) \leq 1 \,\}. \tag{1}$$

If we drop the restriction on the positive definiteness of A (but keep the requirement on the symmetry, since the expression defining $S_{A,c}$ depends only on $(A + A^{\mathrm{T}})/2$), we can also model paraboloids, half-planes, hyperboloids etc. Our goal will be to find the pair (A, c) such that $z_i \in S_{A,c}$, if $w_i > 0$, and $z_i \notin S_{A,c}$, otherwise. Thus we have a problem with $n(n+1)/2 + n$ unknowns, as only half of the matrix coefficients is necessary.

Of course, depending on the data, this may not be possible. Therefore, we will solve our problem as the least squares problem by first defining the residuals

$$r_i(A, c) := \begin{cases} \max\{(z_i - c)^{\mathrm{T}} A (z_i - c) - 1, 0\}, & \text{if } w_i > 0, \\ \max\{1 - (z_i - c)^{\mathrm{T}} A (z_i - c), 0\}, & \text{otherwise}, \end{cases} \tag{2}$$

and then utimately trying to minimize the sum of squares of those:

$$\min_{x=(A,c)\in\mathbb{R}^{n(n+1)/2+n}} f(x) = \sum_{i=1}^{m} r_i(A, c)^2. \tag{3}$$

1.2 Model 2

We proceed along the lines of the previous subsection, except we use a different definition of the separating set. Namely, given the symmetric matrix $A \in \mathbb{R}^N$ and a vector $b \in \mathbb{R}^n$, we put

$$\hat{S}_{A,b} := \{\, z \in \mathbb{R}^n \mid z^{\mathrm{T}} A z + b^{\mathrm{T}} z \leq 1 \,\}. \tag{4}$$

Note that in this model $0 \in \hat{S}_{A,b}$ for all (A, b), and therefore the data z_i with $w_i > 0$ should be centered around zero, if we want a good fit. Model 1 can take care of this "automatically" by using the centering parameter c.

Similarly to the previous case we put

$$\hat{r}_i(A, b) := \begin{cases} \max\{z_i^{\mathrm{T}} A z_i + b^{\mathrm{T}} z_i - 1, 0\}, & \text{if } w_i > 0, \\ \max\{1 - z_i^{\mathrm{T}} A z_i - b^{\mathrm{T}} z_i, 0\}, & \text{otherwise,} \end{cases} \tag{5}$$

which in turn allows us to define the least squares problem as:

$$\min_{x=(A,b)\in\mathbb{R}^{n(n+1)/2+n}} \hat{f}(x) = \sum_{i=1}^{m} \hat{r}_i(A, b)^2. \tag{6}$$

2 Questions

Question 1

Show that functions defined in (3) and (6) are once continuously differentiable as functions of the optimization variables $x \in \mathbb{R}^{n(n+1)/2+n}$, but typically not twice differentiable.

Verify that the gradient of the function in (6) is globally Lipschitz continuous.

Question 2

For both problems present an instance, where the globally optimal solution to the optimization problem exists but is not unique. Similarly, describe a case where the objective function is not coercive.

Question 3

Show that the objective function defined in (6) is convex. Hint: show that if a function $g : \mathbb{R}^\ell \to \mathbb{R}$ is convex and $h : \mathbb{R}^k \to \mathbb{R}^\ell$ is affine (a sum of a linear function and a constant), then the composition $g(h(\cdot)) : \mathbb{R}^k \to \mathbb{R}$ is convex.

Question 4

Implement the code evaluating the function values and the gradients of the objective functions defined in (3) and (6). Verify your implementation against a finite-difference approximation.

Question 5

Implement a couple of different unconstrained optimization algorithms capable of solving the least squares problems (3) and (6). Compare their behaviour on the two models. Also compare the behaviour of two different algorithms on each model. Discuss the results in your report.

Note: whereas the objective functions in out least squares problems are not necessarily two times differentiable at all points, they are two times differentiable at "most" points. As a result, whereas Newton's method may not be directly applicable or converge without additional "tricks," quasi-Newton algorithms may provide a viable alternative with fast local convergence.

3 Test problems

Given the geometric nature of the problem, it is relatively easy to construct test examples for these problems, and visualize the computed solutions for inspection.

Construction of test problems

For example, one can start by selecting the parameters (A, c) or (A, b). One can then generate a set of points with random coordinates z_i, and afterwards compute w_i from by testing the inclusion $z_i \in S_{A,c}$ (or $z_i \in \hat{S}_{A,b}$). It would be quite natural to expect the numerical methods to converge to the parameters used for constructing the data; however remember the non-uniqueness of solutions, or even the non-convexity in Model 1.

Alternatively one can try to approximate other sets, such as rectangles or other polytopes, using model sets $S_{A,c}$ or $\hat{S}_{A,b}$ with the help of our least squares problem. Similarly to the previous case, one can generate some random points, but then compute the labels w_i based on the inclusion to the desired rectangle/polytope.

Finally, one can try to for example model classification errors in the preparation of test data (measurement errors, perhaps) by introducing a small probability of "misclassifying" the inclusion $z_i \in S_{A,c}$ (or $z_i \in \hat{S}_{A,b}$) when computing the label w_i. This should tell us something about how robust our optimization approach to this problem is.

Experiment with different problem types/sizes (in terms of the number of data points) to see how the algorithms you have implemented perform in practice.

Visualization of the data/solutions

In two or three dimensions the data can be visualized as a scatter plot, coloured by the "labels" w_i. The computed model sets can be vizualized as a 0-level contour of the function $z \mapsto (z-c)^{\mathrm{T}}A(z-c) - 1$ or $z \mapsto z^{\mathrm{T}}Az + b^{\mathrm{T}}z - 1$.