# Maths for Intelligent Systems
# Exercise 6

Marc Toussaint
Machine Learning & Robotics lab, U Stuttgart
Universitätsstraße 38, 70569 Stuttgart, Germany

January 18, 2019

## 1 Convergence proof

a) Given a function $f:\ \mathbb{R}^n \to \mathbb{R}$ with $f_{\min} = \min_x f(x)$. Assume that its Hessian—that is, the eigenvalues of $\nabla^2 f$—are lower bounded by $m > 0$ and upper bounded by $M > m$, with $m, M \in \mathbb{R}$. Prove that for any $x \in \mathbb{R}^n$ it holds

$$f(x) - \frac{1}{2m}|\nabla f(x)|^2 \le f_{\min} \le f(x) - \frac{1}{2M}|\nabla f(x)|^2 \ .$$

Tip: Start with bounding the 2nd-order Taylor expansion. Then consider the minima of these bounds. Note, it also follows:

$$|\nabla f(x)|^2 \ge 2m(f(x) - f_{\min}) \ .$$

b) Consider backtracking line search with Wolfe parameter $\varrho_{\text{ls}} \le \frac{1}{2}$, and step decrease factor $\varrho_\alpha^-$. First prove that line search terminates the latest when $\frac{\varrho_\alpha^-}{M} \le \alpha \le \frac{1}{M}$, and then it found a new point $y$ for which

$$f(y) \le f(x) - \frac{\varrho_{\text{ls}}\varrho_\alpha^-}{M}|\nabla f(x)|^2 \ .$$

From this, using the result from a), prove the convergence equation

$$f(y) - f_{\min} \le \Big[1 - \frac{2m\varrho_{\text{ls}}\varrho_\alpha^-}{M}\Big]\ (f(x) - f_{\min}) \ .$$

a)

$$f(x) + \nabla f(x)^\top (y-x) + \frac{m}{2}(y-x)^2 \le f(y) \tag{1}$$

$$\le f(x) + \nabla f(x)^\top (y-x) + \frac{M}{2}(y-x)^2 \tag{2}$$

$$f(x) - \frac{1}{2m}|\nabla f(x)|^2 \le f_{\min} \le f(x) - \frac{1}{2M}|\nabla f(x)|^2 \tag{3}$$

$$|\nabla f(x)|^2 \ge 2m(f(x) - f_{\min}) \tag{4}$$

b) Substitute $y = x - \alpha\nabla f(x)$ in the 2nd-order taylor, and use $\alpha \le \frac{1}{M}$:

$$f(y) \le f(x) - \alpha|\nabla f(x)|^2 + \frac{M\alpha^2}{2}|\nabla f(x)|^2 \tag{5}$$

$$\le f(x) - \frac{\alpha}{2}|\nabla f(x)|^2 \tag{6}$$

$$\le f(x) - \varrho_{\text{ls}}\alpha|\nabla f(x)|^2 \tag{7}$$

If follows:

$$f(y) \le f(x) - \frac{\varrho_{\text{ls}}\varrho_\alpha^-}{M}|\nabla f(x)|^2 \tag{8}$$

$$f(y) - f_{\min} \le f(x) - f_{\min} - \frac{\varrho_{\text{ls}}\varrho_\alpha^-}{M}|\nabla f(x)|^2 \tag{9}$$

$$\le f(x) - f_{\min} - \frac{2m\varrho_{\text{ls}}\varrho_\alpha^-}{M}(f(x) - f_{\min}) \tag{10}$$

$$\le \Big[1 - \frac{2m\varrho_{\text{ls}}\varrho_\alpha^-}{M}\Big]\ (f(x) - f_{\min}) \tag{11}$$

# 2 Backtracking Line Search

Consider the functions

$$f_{\text{sq}}(x) = x^\top C x \ , \tag{12}$$

$$f_{\text{hole}}(x) = 1 - \exp(-x^\top C x) \ . \tag{13}$$

with diagonal matrix $C$ and entries $C(i,i) = c^{\frac{i-1}{n-1}}$, where $n$ is the dimensionality of $x$. We choose a conditioning[1] $c = 10$. To plot the function for $n = 2$, you can use gnuplot calling

```
set isosamples 50,50
set contour
f(x,y) = x*x+10*y*y
#f(x,y) = 1 - exp(-x*x-10*y*y)
splot [-1:1][-1:1] f(x,y)
```

a) Implement gradient descent with backtracking, as described on page 42 (Algorithm 2 Plain gradient descent). Test the algorithm on $f_{\text{sq}}(x)$ and $f_{\text{hole}}(x)$ with start point $x_0 = (1,1)$. To judge the performance, create the following plots:

- The function value over the number of function evaluations.
- For $n = 2$, the function surface including algorithm's search trajectory. If using gnuplot, store every evaluated point $x$ and function value $f(x)$ in a line (with $n+1$ entries) in a file 'path.dat', and plot using

```
unset contour
splot [-3:3][-3:3] f(x,y), 'path.dat' with lines
```

b) Play around with parameters. How does the performance change for higher dimensions, e.g., $n = 100$? How does the performance change with $\rho_{\text{ls}}$ (the Wolfe stop criterion)? How does the alternative in step 3 work?

c) Newton step: Modify the algorithm simply by multiplying $C^{-1}$ to the step. How does that work?

(The Newton direction diverges (is undefined) in the concave part of $f_{\text{hole}}(x)$. We're cheating here when always multiplying with $C^{-1}$ to get a good direction.)

[1] The word "conditioning" generally denotes the ratio of the largest and smallest Eigenvalue of the Hessian.