Proofs, Exercises and Literature - Optimization

1 Proofs

1.1 Example: the Minimum of the Rosenbrock Function

In this example we apply FONC and SONC to find the minimizers of the Rosenbrock function

$$f(\mathbf{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

In order to apply FONC, we need to compute the gradient. We do so by computing the partial derivatives. The partial derivatives are computed by the same rules as you know it from computing the derivative of a one-dimensional function.

$$\begin{aligned}\frac{\partial}{\partial x_1} f(\mathbf{x}) &= 400x_1(x_1^2 - x_2) + 2(x_1 - 1)\\ \frac{\partial}{\partial x_2} f(\mathbf{x}) &= 200(x_2 - x_1^2)\end{aligned}$$

FONC says that every minimizer has to be a stationary point. Stationary points are the vectors at which the gradient of f is zero. We compute the set of stationary points by setting the gradient to zero and solving for $\mathbf{x}$.

$$\begin{aligned}\frac{\partial}{\partial x_2} f(\mathbf{x}) &= 200(x_2 - x_1^2) = 0 && \Leftrightarrow x_2 = x_1^2\\ \frac{\partial}{\partial x_1} f\begin{pmatrix} x_1 \\ x_1^2 \end{pmatrix} &= 2(x_1 - 1) = 0 && \Leftrightarrow x_1 = 1\end{aligned}$$

According to FONC we have a stationary point at $\mathbf{x} = (1, 1)$. Now we check with SONC if the stationary point is indeed a minimizer (it could also be a maximizer or a saddle point). SONC says that every stationary point whose Hessian is positive semi-definite is a minimizer. Hence, we require the Hessian, the second derivative of the Rosenbrock function. To that end, we compute the partial derivatives of the partial derivatives:

$$\begin{aligned}\frac{\partial^2}{\partial^2 x_1} f(\mathbf{x}) &= \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_1} f(\mathbf{x})\right) = 1200x_1^2 - 400x_2 + 2\\ \frac{\partial^2}{\partial^2 x_2} f(\mathbf{x}) &= \frac{\partial}{\partial x_2}\left(\frac{\partial}{\partial x_2} f(\mathbf{x})\right) = 200\\ \frac{\partial^2}{\partial x_1 \partial x_2} f(\mathbf{x}) &= \frac{\partial^2}{\partial x_2 \partial x_1} f(\mathbf{x}) = -400x_1\end{aligned}$$

The Hessian is given by

$$\begin{aligned}\nabla^2 f(\mathbf{x}) &= \begin{pmatrix} \frac{\partial^2}{\partial^2 x_1} f(\mathbf{x}) & \frac{\partial^2}{\partial x_1 x_2} f(\mathbf{x}) \\ \frac{\partial^2}{\partial x_2 x_1} f(\mathbf{x}) & \frac{\partial^2}{\partial^2 x_2} f(\mathbf{x}) \end{pmatrix}\\ &= 200\begin{pmatrix} 16x_1^2 - 2x_2 + 0.01 & -2x_1 \\ -2x_1 & 1 \end{pmatrix}\end{aligned}$$

We insert our stationary point $\mathbf{x}_0 = (1, 1)$ into the Hessian and get

$$\nabla^2 f(\mathbf{x}_0) = 200 \begin{pmatrix} 4.01 & -2 \\ -2 & 1 \end{pmatrix}$$

Now we check if the Hessian at the stationary point is positive definite. Let $\mathbf{x} \in \mathbb{R}^2$, then

$$\begin{aligned} \mathbf{x}^\top \nabla^2 f(\mathbf{x}_0)\mathbf{x} &= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 4.01 & -2 \\ -2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 4.01x_1 - 2x_2 \\ -2x_1 + x_2 \end{pmatrix} \\ &= 4.01x_1^2 - 2x_1x_2 - 2x_1x_2 + x_2^2 \\ &= 4.01x_1^2 - 4x_1x_2 + x_2^2 \\ &= (2x_1 - x_2)^2 + 0.01x_1^2 \geq 0 \end{aligned}$$

The last inequality follows because the sum of quadratic terms can not be negative. We conclude that the Hessian at our stationary point is positive semi-definite. As a result, FONC and SONC yield that $\mathbf{x} = (1, 1)$ is the only possible local minimizer of f.

2 Exercises

2.1 Convex Functions

1. Show that *nonnegative weighted* sums of *convex functions* are convex. That is, show for all $\lambda_1, \ldots, \lambda_k \geq 0$ and convex functions $f_1, \ldots, f_k : \mathcal{X} \to \mathbb{R}$, that the function

$$f(\mathbf{x}) = \lambda_1 f_1(\mathbf{x}) + \ldots + \lambda_k f_k(\mathbf{x})$$

   is convex.

2. If $g : \mathbb{R}^d \to \mathbb{R}^k$, $g(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ is an *affine map*, and $f : \mathbb{R}^k \to \mathbb{R}$ is a *convex function*, then the *composition*

$$f(g(\mathbf{x})) = f(A\mathbf{x} + \mathbf{b})$$

   is a convex function.

2.2 Numerical Optimization

1. Compute three gradient descent steps for the following optimization problem:

$$\min (x - 2)^2 + 1 \text{ s.t. } x \in \mathbb{R}$$

   Try the following combinations of initalizations and step sizes:

   1. $x_0 = 4$, step size $\eta = \frac{1}{4}$
   2. $x_0 = 4$, step size $\eta = 1$
   3. $x_0 = 3$, step size $\eta = \frac{5}{4}$

   Mark the iterates x_1, x_2 and x_3 in a plot of the objective function. What do you observe regarding the convergence of gradient descent methods? Does gradient descent always "descent" from an iterate?

2.3 Computing the Gradients

1. What is the Jacobian of the squared Euclidean norm $f(\mathbf{x}) = \|\mathbf{x}\|^2$?
2. What is the Jacobian of the function $f : \mathbb{R} \to \mathbb{R}^n$, $f(x) = \mathbf{b} - \mathbf{a}x$ for vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ and $x \in \mathbb{R}$?
3. What is the Jacobian of the function $f : \mathbb{R}^d \to \mathbb{R}^n$, $f(\mathbf{x}) = \mathbf{b} - A\mathbf{x}$, (A is an $(n \times d)$ matrix)?
4. What is the gradient of the function $f : \mathbb{R}^d \to \mathbb{R}$, $f(\mathbf{x}) = \|\mathbf{b} - A\mathbf{x}\|^2$?
5. What is the gradient of the function $f : \mathbb{R}^{d\times r} \to \mathbb{R}$, $f(X) = \|D - YX^\top\|^2$, where $D \in \mathbb{R}^{n\times d}, Y \in \mathbb{R}^{n\times r}$?

3 Recommended Literature

As always, the best exercise is to go through the lecture and see if you can follow the steps with pen and paper and to make the exercises. If you want a more general and extensive overview, the following material is recommended.

Linear Algebra and Optimization for Machine Learning by Charu C. Aggarwal

Sections 4.1-4.3 build up nicely the aspects of optimization from the one-dimensional case (univariate optimizattion) to higher dimensions (multivariate optimization). Section 4.6 gives an overview over computing gradients subject to vectors and matrices.