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.

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

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 $\mathbf{x}$.

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

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:

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

The Hessian is given by

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

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.

Solution: Let

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

for $\lambda_1, \ldots, \lambda_k \geq 0$ and $f_1, \ldots, f_k : \mathcal{X} \to \mathbb{R}$ convex functions. Let $\alpha \in [0, 1]$ and $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. Then we have to show according to the definition of convex functions that

$$f(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y}).$$

According to the definition of f, we have

$$f(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}) = \lambda_1 f_1(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}) + \ldots + \lambda_k f_k(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y})$$

Since f_i is convex for $1 \leq i \leq k$, for the functions f_i holds that

$$f_i(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}) \leq \alpha f_i(\mathbf{x}) + (1 - \alpha) f_i(\mathbf{y}).$$

Now we multiply the inequality above with the nonnegative value λ_i. Note that here it becomes obvious why the coefficients have to be nonnegative. Multiplying an inequality with a negative value changes the direction of the inequality, that is a $\leq$ becomes a $\geq$ and vice versa. However, here we have only nonnegative values λ_i and thus, multiplying with the coefficient keeps the inequality intact:

$$\begin{aligned} \lambda_i f_i(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}) &\leq \lambda_i(\alpha f_i(\mathbf{x}) + (1 - \alpha) f_i(\mathbf{y})) \\ &= \alpha\lambda_i f_i(\mathbf{x}) + (1 - \alpha)\lambda_i f_i(\mathbf{y}). \end{aligned} \tag{1}$$

We can now put these inequalities together to derive the convexity of the function f:

$$\begin{aligned}
f(\alpha\mathbf{x}+(1-\alpha)\mathbf{y}) &= \lambda_1 f_1(\alpha\mathbf{x}+(1-\alpha)\mathbf{y}) + \ldots + \lambda_k f_k(\alpha\mathbf{x}+(1-\alpha)\mathbf{y}) \\
&= \sum_{i=1}^{k} \lambda_i f_i(\alpha\mathbf{x}+(1-\alpha)\mathbf{y}) \\
&\leq \sum_{i=1}^{k} \alpha\lambda_i f_i(\mathbf{x}) + (1-\alpha)\lambda_i f_i(\mathbf{y}) && \text{apply Eq. (1)} \\
&= \alpha \sum_{i=1}^{k} \lambda_i f_i(\mathbf{x}) + (1-\alpha) \sum_{i=1}^{k} \lambda_i f_i(\mathbf{y}) \\
&= \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}). && \text{apply definition of } f
\end{aligned}$$

This concludes what we wanted to show.

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.

Solution: Let $g : \mathbb{R}^d \to \mathbb{R}^k$, $g(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ be an affine map, and let $f : \mathbb{R}^k \to \mathbb{R}$ be a convex function. Then we have to show according to the definition of convex functions that

$$\begin{aligned}
& f(g(\alpha\mathbf{x}+(1-\alpha)\mathbf{y})) \leq \alpha f(g(\mathbf{x})) + (1-\alpha) f(g(\mathbf{y})) \\
\Leftrightarrow\ & f(A(\alpha\mathbf{x}+(1-\alpha)\mathbf{y})+\mathbf{b}) \leq \alpha f(A\mathbf{x}+\mathbf{b}) + (1-\alpha) f(A\mathbf{y}+\mathbf{b}),
\end{aligned}$$

where the last inequality derives from the definition of g. The function g is a linear function. Linear functions satisfy for any scalar α and vectors $\mathbf{x}$ and $\mathbf{y}$ the criterion

$$g(\alpha\mathbf{x}+\mathbf{y}) = \alpha g(\mathbf{x}) + g(\mathbf{y}).$$

The linearity of g follows from the linearity of the matrix multiplication:

$$\begin{aligned}
A(\alpha\mathbf{x}+(1-\alpha)\mathbf{y})+\mathbf{b} &= \alpha A\mathbf{x} + (1-\alpha)A\mathbf{y} + \mathbf{b} && \text{(linearity)} \\
&= \alpha A\mathbf{x} + (1-\alpha)A\mathbf{y} + (\alpha+1-\alpha)\mathbf{b} \\
&= \alpha A\mathbf{x} + (1-\alpha)A\mathbf{y} + \alpha\mathbf{b} + (1-\alpha)\mathbf{b} \\
&= \alpha(A\mathbf{x}+\mathbf{b}) + (1-\alpha)(A\mathbf{y}+\mathbf{b})
\end{aligned}$$

As a result, we get with respect to g that:

$$\begin{aligned}
g(\alpha\mathbf{x}+(1-\alpha)\mathbf{y}) &= A(\alpha\mathbf{x}+(1-\alpha)\mathbf{y})+\mathbf{b} \\
&= \alpha(A\mathbf{x}+\mathbf{b}) + (1-\alpha)(A\mathbf{y}+\mathbf{b}) \\
&= \alpha g(\mathbf{x}) + (1-\alpha) g(\mathbf{y}).
\end{aligned}$$

If we apply now the function f to the equality above and use the convexity of f, then we can conclude what we wanted to show:

$$\begin{aligned}
f(g(\alpha\mathbf{x}+(1-\alpha)\mathbf{y})) &= f\left(\alpha g(\mathbf{x}) + (1-\alpha) g(\mathbf{y})\right) \\
&\leq \alpha f\left(g(\mathbf{x})\right) + (1-\alpha) f\left(g(\mathbf{y})\right).
\end{aligned}$$

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?

Solution: In order to conduct gradient descent, we need the derivative:

$$f(x) = (x-2)^2+1$$
$$f'(x) = 2(x-2)$$

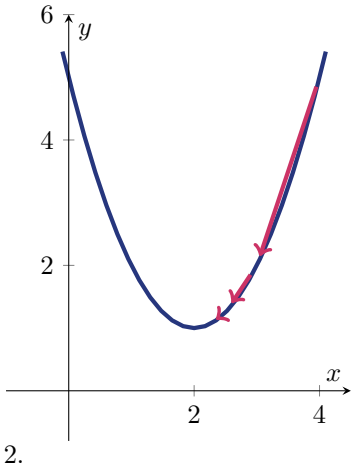
The gradient descent update rules are defined as

$$x_{t+1} = x_t - \eta f'(x_t).$$

We conduct now two gradient descent steps for the stated scenarios:

1. $x_0 = 4$**, step size** $\eta = \frac{1}{4}$

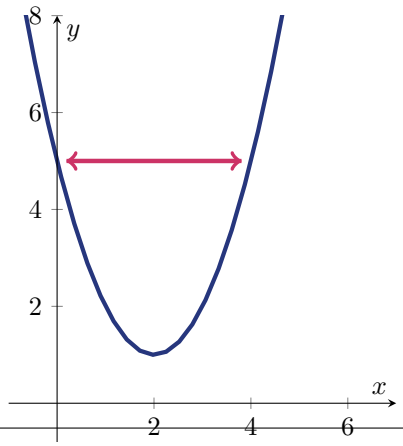
$$x_1 = x_0 - \eta f'(x_0) = 4 - \frac{1}{4}4 = 3$$
$$x_2 = x_1 - \eta f'(x_1) = 3 - \frac{1}{4}2 = 2.5$$
$$x_3 = x_2 - \eta f'(x_2) = 2.5 - \frac{1}{4} = 2.25$$



The iterates are slowly converging to the minimum $x^* = 2$.

2. $x_0 = 4$**, step size** $\eta = 1$

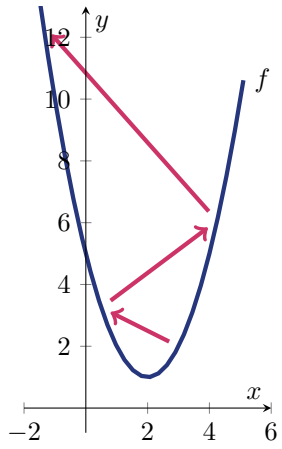
$$x_1 = x_0 - \eta f'(x_0) = 4 - 4 = 0$$
$$x_2 = x_1 - \eta f'(x_1) = 0 - (-4) = 4$$
$$x_3 = x_2 - \eta f'(x_2) = 4 - 4 = 0$$



The iterates are oscilliating between the values 0 and 4. Hence, the iterates will never converge when using a step-size of $\eta = 1$. The step-size is too large.

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

$$x_1 = x_0 - \eta f'(x_0) = 3 - \frac{5}{4}2 = \frac{1}{2}$$
$$x_2 = x_1 - \eta f'(x_1) = \frac{1}{2} - \frac{5}{4}(-3) = 4.25$$
$$x_3 = x_2 - \eta f'(x_2) = 4.25 - \frac{5}{4}4.5 = -1.375$$



The iterates are oscilliating and the function values are diverging (going to infinity). Every gradient step is actually increasing the objective function since the step size is far too large.

2.3 Computing the Gradients

1. What is the Jacobian of the squared Euclidean norm $f(\mathbf{x}) = \|\mathbf{x}\|^2$?

Solution: Given a vector $\mathbf{x} \in \mathbb{R}^d$, then the squared Euclidean norm is defined as:

$$\|\mathbf{x}\|^2 = \sum_{i=1}^{d} x_i^2.$$

We compute the partial derivative with respect to x_k, treating the terms x_i as constants for $i \neq k$:

$$\frac{\partial}{\partial x_k}\|\mathbf{x}\|^2 = \frac{\partial}{\partial x_k}\sum_{i=1}^{d} x_i^2 = \frac{\partial}{\partial x_k}x_k^2 = 2x_k.$$

Hence, the Jacobian is given by

$$\frac{\partial}{\partial \mathbf{x}}\|\mathbf{x}\|^2 = \begin{pmatrix}2x_1 & \dots & 2x_d\end{pmatrix} = 2\mathbf{x}^\top.$$

Correspondingly, we can denote the gradient now as the transposed of the Jacobian:

$$\nabla\|\mathbf{x}\|^2 = 2\mathbf{x}.$$

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 $x \in \mathbb{R}$?

Solution: We write the function f as a vector of one-dimensional functions:

$$f(x) = \mathbf{b} - \mathbf{a}x = \begin{pmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{pmatrix} = \begin{pmatrix} b_1 - a_1 x \\ \vdots \\ b_n - a_n x \end{pmatrix}.$$

The derivative of the one-dimensional functions f_i is given by

$$\frac{\partial}{\partial x} f_i(x) = \frac{\partial}{\partial x}(b_i - a_i x) = -a_i.$$

Hence, the Jacobian of f is equal to the vector

$$\frac{\partial}{\partial x} f(x) = -\mathbf{a}.$$

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)?

Solution: There are multiple ways to derive the Jacobian of this function. I believe the shortest, but not necessarily most obvious way is to use here the result from the exercise above and to employ the matrix product definition given by the outer-product in the column-times-row scheme:

$$A\mathbf{x} = A_{\cdot 1} x_1 + \ldots + A_{\cdot d} x_d.$$

Now we can apply the linearity of the partial derivative of f:

$$\begin{aligned} \frac{\partial}{\partial x_k} f(\mathbf{x}) &= \frac{\partial}{\partial x_k}(\mathbf{b} - A\mathbf{x}) = \frac{\partial}{\partial x_k}\mathbf{b} - \frac{\partial}{\partial x_k} A\mathbf{x} = \mathbf{0} - \frac{\partial}{\partial x_k}(A_{\cdot 1} x_1 + \ldots + A_{\cdot d} x_d) \\ &= -\frac{\partial}{\partial x_k} A_{\cdot 1} x_1 - \ldots - \frac{\partial}{\partial x_k} A_{\cdot k} x_k - \ldots - \frac{\partial}{\partial x_k} A_{\cdot d} x_d \\ &= -\frac{\partial}{\partial x_k} A_{\cdot k} x_k \\ &= -A_{\cdot k}, \end{aligned}$$

where we applied for the partial derivatives the rule which we derived in the previous exercise for the Jacobian of a function from a scalar to a vector.

Now the question is how we have to arrange the partial derivatives to form the Jacobian. We can either look up in the slides how that goes, or we remember from the lecture that the dimensionality of the Jacobian is swapping the dimensionality from the input- and output space. Our function f maps from the d-dimensional space to the n-dimensional space. Hence, the dimensionality of the Jacobian is $(n \times d)$, the same like our matrix A. Thus, the n-dimensional partial derivatives have to be concatenated horizontally:

$$\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = \begin{pmatrix} -A_{\cdot 1} & \ldots & -A_{\cdot d} \end{pmatrix} = -A.$$

4. What is the gradient of the function $f : \mathbb{R}^d \to \mathbb{R}$, $f(\mathbf{x}) = \|\mathbf{b} - A\mathbf{x}\|^2$?

Solution: Here, we can apply now the chainrule to the inner function $g(\mathbf{x}) = \mathbf{b} - A\mathbf{x}$ and the outer function $h(\mathbf{y}) = \|\mathbf{y}\|^2$. From the exercises before, we know the gradients of both functions:

$$\nabla_{\mathbf{x}} g(\mathbf{x}) = \left(\frac{\partial}{\partial \mathbf{x}}(\mathbf{b} - A\mathbf{x})\right)^\top = (-A)^\top = -A^\top$$
$$\nabla_{\mathbf{y}} h(\mathbf{y}) = \left(\frac{\partial}{\partial \mathbf{y}}\|\mathbf{y}\|^2\right)^\top = 2\mathbf{y}$$

In the chain rule, the inner and outer gradients are multiplied. You can either look up the definition or deduce how the gradients have to be multiplied from the dimensionalities. The gradient of a function to the real values has the same dimensionality like the input space. Hence, we have a look how we can multiply the inner and outer gradients such that we get a d-dimensional vector. This is only the case if we multiply the gradient of the inner function with the gradient of the outer function. Therewith, we get:

$$\begin{aligned}\nabla_{\mathbf{x}} h(g(\mathbf{x})) &= \nabla_{\mathbf{x}} g(\mathbf{x}) \cdot \nabla_{g(\mathbf{x})} h(g(\mathbf{x})) \\ &= -A^\top(2(\mathbf{b} - A\mathbf{x})) \\ &= -2A^\top(\mathbf{b} - A\mathbf{x}).\end{aligned}$$

5. What is the gradient of the function $f : \mathbb{R}^{d\times r} \to \mathbb{R}$, $f(X) = \|D - YX^\top\|^2$, where $D \in \mathbb{R}^{n\times d}, Y \in \mathbb{R}^{n\times r}$?

Solution: Let's have first a look at the dimensionality of the resulting gradient. Since the function f is mapping to the real values, the dimensionality of the gradient is the same as the one of the input space: $(n \times r)$. Since we do not know any gradients subject to matrices yet, we divide the problem and compute the gradient row-wise. Every row of X is mapped to the corresponding row of the gradient:

$$\nabla_X f(X) = \begin{pmatrix} - & \nabla_{X_{1\cdot}} f(X) & - \\ & \vdots & \\ - & \nabla_{X_{d\cdot}} f(X) & - \end{pmatrix}. \tag{2}$$

The gradient with regard to row $X_{k\cdot}$ of the function f is equal to

$$\begin{aligned}\nabla_{X_{k\cdot} f(X)} &= \nabla_{X_{k\cdot}} \|D - YX^\top\|^2 \\ &= \nabla_{X_{k\cdot}} \sum_{i=1}^{d} \|D_{\cdot i} - YX_{i\cdot}^\top\|^2 \qquad (3) \\ &= \nabla_{X_{k\cdot}} \|D_{\cdot k} - YX_{k\cdot}^\top\|^2, \qquad (4)\end{aligned}$$

where Eq. (4) derives from the linearity of the gradient. Eq. (3) follows from the fact that the squared Frobenius norm (matrix L_2-norm) is the sum of the squared Euclidean norms over all column- or row-vectors of a matrix. That is for any matrix $A \in \mathbb{R}^{n\times d}$ we have

$$\|A\|^2 = \sum_{i=1}^{d}\sum_{j=1}^{n} A_{ji}^2 = \sum_{i=1}^{d} \|A_{\cdot i}\|^2 = \sum_{j=1}^{n} \|A_{j\cdot}\|^2.$$

We can denote the gradient of the term in Eq. (4), as we have derived it in the previous exercise. We only have to keep in mind that we derived the gradient in the previous exercise subject to a

column-vector and here we have the gradient with regard to the row vector $X_{k\cdot}$. Hence, we have to transpose the result from the previous exercise to get the gradient for our row-vector:

$$\nabla_{X_{k\cdot}} \|D_{\cdot k} - Y X_{k\cdot}^\top\|^2 = (-2Y^\top (D_{\cdot k} - Y X_{k\cdot}^\top))^\top = -2(D_{\cdot k} - Y X_{k\cdot}^\top)^\top Y.$$

We insert this result now in Eq. (2) and obtain the final result:

$$\begin{aligned}
\nabla_X f(X) &= \begin{pmatrix} - & \nabla_{X_{1\cdot}} f(X) & - \\ & \vdots & \\ - & \nabla_{X_{d\cdot}} f(X) & - \end{pmatrix} \\
&= \begin{pmatrix} -2(D_{\cdot 1} - Y X_{1\cdot}^\top)^\top Y \\ \vdots \\ -2(D_{\cdot d} - Y X_{d\cdot}^\top)^\top Y \end{pmatrix} \\
&= -2 \begin{pmatrix} (D_{\cdot 1} - Y X_{1\cdot}^\top)^\top \\ \vdots \\ (D_{\cdot d} - Y X_{d\cdot}^\top)^\top \end{pmatrix} Y \\
&= -2(D - Y X^\top)^\top Y.
\end{aligned}$$

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.