# Summary of OCS Slides

Philipp Gabler

## 1 Introduction

**General Form.** A general minimization problem has the form

$$\min_x f(x) \quad \text{s.t. } x \in X,$$

for a *constraint set* $X \subseteq \mathbb{R}^n$ (often given by some *constraint functions* and an *objective function* $f : X \to \mathbb{R}$. We want to find an optimal value or *minimizer* $x^* \in X$ such that

$$f(x^*) \leq f(x), \quad \forall x \in X.$$

**Types of Optimization Problems.**

1. (a) Discrete: $X$ is a discrete set, also called *interger programming*.
   (b) Continuous: $X$ is continuous (ie. uncountable)
2. (a) Linear: Objective functions and constraints are all linear:
   $$\min_x c^\top x, \quad \text{s.t. } Ax \leq b,\ x \geq 0.$$
   Constraints describe a polyhedron. Efficiently solvable.
   (b) Quadratic: Objective function is quadratic, constraints linear:
   $$\min_x \frac{1}{2} x^\top Q x + c^\top x, \quad \text{s.t. } Ax \leq b,\ Ex = d.$$
   If $Q$ is positive semidefinite, the objective is convex and the problem is polynomially solvable.
   (c) Nonlinear: no further constraints.
3. (a) Unconstrained: Optimal solution searched in full $\mathbb{R}^n$. Easier to characterize, and usually to solve.
   (b) Constrained: Optimal solution in an admissible region, usually more difficult to setup/characterize.

**Convexity.** A set $X$ is convex, if for all $x, y \in X$ and $\alpha \in [0, 1]$:

$$\alpha x + (1 - \alpha) y \in X.$$

This means that $X$ contains all convex combinations of points from it.

**Convex Functions.** If $X$ is a convex set, then $f : \mathbb{R} \to \mathbb{R}$ is called convex if for all $x, y \in X$ and $\alpha \in [0, 1]$:

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

This means that no points lie below any tangent.

**Level Sets.** For an objective function $f : X \to \mathbb{R}$, and $c \in \mathbb{R}$, the sets

$$S_c(f) = \{x \in X : f(x) = c\}$$

are called *level sets* of $f$. They can be convex even if $f$ is not!

**Definiteness.** A matrix $Q$ is called *positive semidefinite* if $x^\top Qx \geq 0$ for all $x$. $Q$ is called *positive definite* if $x^\top Qx > 0$ for all $x \neq 0$. Sometimes this is written as $Q \succeq 0$ and $Q \succ 0$.

**Local and Global Minima.** A point $x^*$ is called an *unconstrained global minimum* of $f$ if for all $x$

$$f(x^*) \leq f(x).$$

$x^*$ is called an *unconstrained local minimum* of $f$ if it is minimal in some neighbourhood; i.e., there is an $\epsilon > 0$ such that

$$f(x^*) \leq f(x) \quad \forall x \text{ with } \|x^* - x\| \leq \epsilon.$$

For *constrained* minima, we just require additionally that $x \in X \subset \mathbb{R}^n$.

**First Order Neccessary Condition for Optimality.** In a small neighbourhood of $x^*$, we can by Taylor expansion write $f$ as

$$f(x) = f(x^* + \Delta x) = f(x^*) + \nabla f(x^*)^\top \Delta x + o(\|\Delta x\|).$$

Since $x^*$ is a local minimum, $f(x^* + \Delta x) - f(x^*) \geq 0$, and we have

$$f(x^*) + \nabla f(x^*)^\top \Delta x - f(x^*) = \nabla f(x^*)^\top \Delta x \geq 0.$$

Since wlog. we can choose $\Delta x$ to have the opposite sign, it holds also that

$$\nabla f(x^*)^\top \Delta x \leq 0,$$

so $\nabla f(x^*)^\top \Delta x = 0$, which, since $\Delta x$ is arbitrary, implies that $\nabla f(x^*) = 0$.

**Second Order Neccessary Condition for Optimality.** By second order Taylor expansion, we get

$$\begin{aligned} 0 &\leq f(x^* + \Delta x) - f(x^*) \\ &= f(x^*) + \underbrace{\nabla f(x^*)^\top \Delta x}_{=0} + \frac{1}{2}\Delta x^\top \nabla^2 f(x^*)\Delta x + o(\|\Delta x\|^2) - f(x^*) \\ &= \frac{1}{2}\Delta x^\top \nabla^2 f(x^*)\Delta x + o(\|\Delta x\|^2). \end{aligned}$$

From this follows that $\Delta x^\top \nabla^2 f(x^*)\Delta x \geq 0$. Since $\Delta x$ is arbitrary, this means that $\nabla^2 f(x^*)$ must be positive semidefinite.

A point which has this property is called a *stationary point*. In "normal" cases, it is either a local optimum or a saddle point.

**Sufficient Condition for Optimality.** If for a point $x^*$ we have $\nabla f(x^*)^\top = 0$ and $\nabla^2 f(x^*)$ positive *definite* (no "semi-"!), then $x^*$ is a strict unconstrained local minimum of f.

**Minima of Convex Functions.** For a convex function $f$, local minima are also global minima: suppose $x*$ were a local, but not global minimum. Then there must be some $y^* \neq x^*$ with $f(y^*) < f(x^*)$. By convexity, we have for all $\alpha \in [0, 1)$:

$$f(\alpha x^* + (1-\alpha)y^*) < \alpha f(x^*) + (1-\alpha)f(y^*) < f(x^*),$$

which contradicts the assumption, so $x^*$ must also be a global minimum.

Furthermore, the neccessary condition for minima, $\nabla f(x^*) = 0$, for convex functions becomes a sufficient condition.

why this?

# 2 Gradient Methods

**Basic Idea.** To find a minimum of $f$, we construct a sequence $x^{(k)}$ such that for all $k$, $f(x^{(k+1)}) < f(x^{(k)})$. To do that, we choose an initial $x^{(0)}$ and then set

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}.$$

Here $\alpha^{(k)}$ is some step size, and $d^{(k)}$ is a *descent direction* which must satisfy

$$\frac{\partial f}{\partial d^{(k)}}(x^{(k)}) = \nabla f(x^{(k)})^\top d^{(k)} < 0,$$

where $\frac{\partial f}{\partial d^{(k)}}$ is the directional derivative in direction $d^{(k)}$.

**Matrix-Scaled Gradients.** Given the above form, one can choose $d^{(k)} = -D^{(k)}\nabla f(x^{(k)})$ for a positive definite $D^{(k)}$:

$$\nabla f(x^{(k)})^\top d^{(k)} = -\nabla f(x^{(k)})^\top D^{(k)} \nabla f(x^{(k)}) < 0,$$

by the definition of positive definiteness.

1. Steepest descent: $D^{(k)} = I$. Simple, but slow convergence.
2. Newton's method: $D^{(k)} = (\nabla^2 f(x^{(k)}))^{-1}$. Fast convergence, but $\nabla^2 f(x^{(k)})$ needs to be positive definite to be invertible. Corresponds to local approximation by a quadratic surface (see below).
3. Levenberg-Marquart method: $D^{(k)} = (\nabla^2 f(x^{(k)}) + \lambda I)^{-1}$. Tries to fix problems with Newton's method by regularization.
4. Diagonal scaling: $D^{(k)} = \operatorname{diag}(d_1^{(k)}, \ldots, d_n^{(k)})$. E.g. approximating Newton's method with
$$d_i^{(k)} = \left(\frac{\partial^2 f}{\partial x_i^2}(x^{(k)})\right)^{-1}.$$
5. Gauss-Newton method: For a nonlinear least-squares problem $f(x) = \frac{1}{2}\|g(x)\|^2$, we can choose $D^{(k)} = (\nabla g(x^{(k)} \nabla g(x^{(k)})^\top)^{-1}$. This is related to the pseudo-inverse.

references to later sections

**Step Size Selection.** To ensure convergence and performance, $\alpha^{(k)}$ needs to be chosen with care. Theoretically, it is not enough set it such that $f(x^{(k+1)}) < f(x^{(k)})$; there are counterexamples for which this holds, but $\lim_{,\to\infty} f(x^{(k)}) > f(x^*)$, e.g., when the step sequence in the limit oscillates between two values on the opposite sides of a "bowl".

Some usual approaches to choosing the step size are:

1. Minimization rule: choose $\alpha^{(k)}$ such that the minimum in the descent direction is taken, i.e.
$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) = \min_{\alpha \geq 0} f(x^{(k)} + \alpha d^{(k)}).$$
2. Limited minimization rule: "heuristically" search the best $\alpha^{(k)}$ in some set, like in an interval $[0, s]$ or among some values $\{\beta^0 s, \beta^1 s, \ldots\}$ for some fixed $\beta \in (0, 1)$ and $s > 0$.
3. Constant step size: sometimes, an optimal (or good enough) value $\alpha^{(k)} = s$ can be computed from the objective function in advance.
4. Diminishing step size: choose a decreasing sequence with $\lim_{k\to\infty} \alpha^{(k)} = 0$ and $\sum_{k=0}^{\infty} \alpha^{(k)} = \infty$. The latter is sufficient to ensure convergence (we can never "run out of space" before the optimum is reached). This has good theoretical properties for some setups, but the convergence rate can be quite slow.

**Armijo Rule.** This is a special method for step size selection, which has nice theoretical properties (e.g. using it, the $x^{(k)}$ always converge to a stationary point). To apply it, we fix scalars $s$, $0 < \beta < 1$, and $0 < \sigma < 1$, and set $\alpha^{(k)} = \beta^{m_k} s$, where we choose $m_k$ as the first nonnegative integer for which

$$f(x^{(k)} + \beta^{m_k} s d^{(k)}) - f(x^{(k)}) \leq \sigma \beta^{m_k} s \nabla f(x^{(k)})^\top d^{(k)}.$$

The interpretation of this is that we try out the step sizes $\beta^{m_k} s$ in decreasing order, until we find one for which the decrease in the objective is sufficiently large.

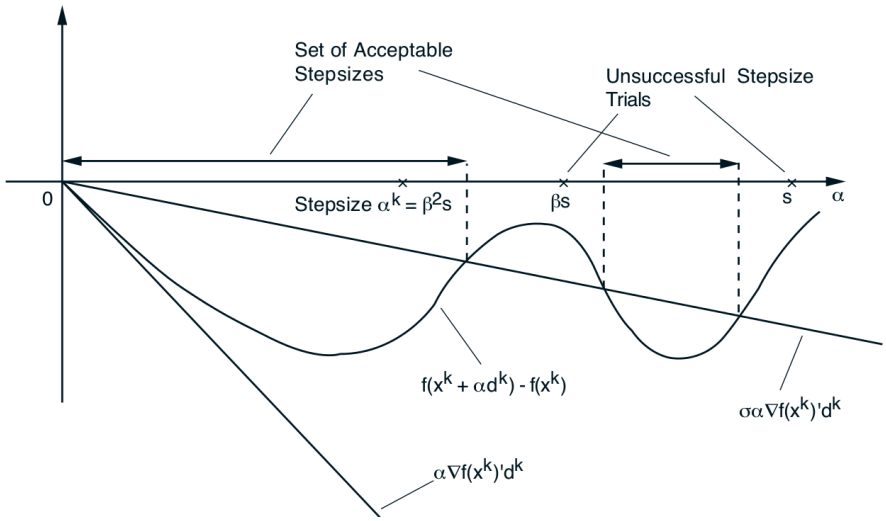


Figure 1: Graphical interpretation of Armijo rule.