Summary of OCS Slides

Philipp Gabler

March 27, 2017

Contents

First of all, the material by the course of Dimitri P. Bertsekas cover pretty much the same material. What a coincidence.

1 Introduction

1.1 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

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

1.2 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.

1.3 Definiteness. A symmetric matrix Q is called *positive semidefinite* if $x^\top Q x \geq 0$ for all x, and *positive definite* if $x^\top Q x > 0$ for all $x \neq 0$. Sometimes this is written as $Q \succeq 0$ and $Q \succ 0$.[1] In the case of $Q \in \mathbb{R}^{2\times 2}$, we can use the following criteria:

1. $Q \succeq 0 \Leftrightarrow \det Q \geq 0, Q_{11} \geq 0, Q_{22} \geq 0$
2. $Q \succ 0 \Leftrightarrow \det Q > 0, Q_{11} > 0 \Leftrightarrow$ all eigenvalues of Q are positive.

[1] `https://en.wikipedia.org/wiki/Positive-definite_matrix`

1.4 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.

1.5 Convex Functions. If X is a convex set, then $f : X \to \mathbb{R}$ 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. Equivalent to this are the first-order condition

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y), \quad \forall x, y \in X$$

and the second-order condition

$$\nabla^2 f(x) \succeq 0 \quad \forall x \in X.$$

1.6 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!

1.7 Basic Derivatives. Always remember these:

$$\nabla_x \frac{1}{2} x^\top Q x + c^\top x = Qx + c,$$
$$\nabla_x \frac{1}{2} \|Ax - b\| = A^\top (Ax - b).$$

1.8 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 $x \in X \subset \mathbb{R}^n$.

1.9 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 Δx 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 Δx is arbitrary, implies that $\nabla f(x^*) = 0$.

1.10 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 Δx 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.

1.11 Sufficient Condition for Optimality. If for a point x^* we have

$$\nabla f(x^*)^\top = 0 \text{ and } \nabla^2 f(x^*) \succ 0,$$

(no "semi-"!), then x^* is a strict unconstrained local minimum of f.

1.12 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 $f(y^*) < f(x^*)$. But 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

2.1 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 $d^{(k)}$.

2.2 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 useful. Corresponds to local approximation by a quadratic surface (see Summary 4.2).
3. Levenberg-Marquart method: $D^{(k)} = (\nabla^2 f(x^{(k)}) + \lambda I)^{-1}$. Tries to fix problems with Newton's method by regularization (see Summary 5.3).
4. Diagonal scaling: $D^{(k)} = \text{diag}(d_1^{(k)}, \ldots, d_n^{(k)})$. 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 (see Summary 5.2.

2.3 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 $\lim_{k\to\infty} f(x^{(k)}) > f(x^*)$, e.g., when the step sequence in the limit oscillates between two values on the opposite sides of a "bowl" (cf. Summary 3.1).

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.

2.4 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 (see Figure 1).

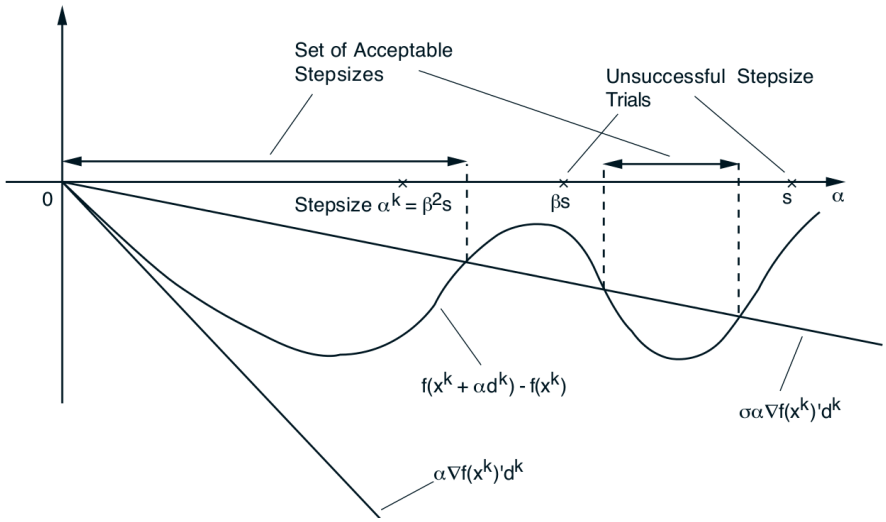

Figure 1: Graphical interpretation of Armijo rule.

2.5 Termination of Gradient Methods. Gradient methods are not automatically convergent, so we need some stopping criterion. The standard approach is to terminate iteration based on the norm of the gradient:

$$\|\nabla f(x^{(k)})\| \leq \epsilon,$$

for some reasonable $\epsilon > 0$. Since the absolute sizes of the gradients are not neccessarily meaningful, a better criterion is

$$\frac{\|\nabla f(x^{(k)})\|}{\|\nabla f(x^{(0)})\|} \leq \epsilon.$$

Assuming we have diagonal scaling, we can also use $\|D^{(k)}\nabla f(x^{(0)})\| \leq \epsilon$.

If $\nabla^2 f(x)$ is positive definite, we have a strongly convex problem, and the norm of the gradient actually bounds the distance to a local minimum $\|x - x^*\|$.

3 Convergence Analysis

3.1 Gradient Related Condition. A sequence of descent directions $\{d^{(k)}\}$ is called *gradient related* to a step sequence $\{x^{(k)}\}$, if for any subsequence $\{x^{(k)}\}_{k\in\mathcal{K}}$ converging to a nonstationary point, the corresponding subsequence $\{d^{(k)}\}_{k\in\mathcal{K}}$ is bounded and satisfies

$$\limsup_{k\in\mathcal{K}\to\infty} \nabla f(x^{(k)})^\top d^{(k)} < 0.$$

The interpretation of this is that in the limit, $d^{(k)}$ is still a descent direction (see Summary 2.1).

This is, for example, satisfied for $d^{(k)} = -D^{(k)}\nabla f(x^{(k)})$, when the eigenvalues of $D^{(k)}$ are bounded between zero and a positive constant. It fails if the directions get more and more orthogonal to the gradient; a counterexample would be a sequence which oscillates in the limit, due to a badly chosen step size (there, all finite directions are descent directions, but they get "worse" in the limit).

3.2 Lipschitz continuity. A continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is called *Lipschitz continuous*, if there is value $L \geq 0$, called *Lipschitz constant*, such that for all $x, y \in \mathbb{R}^n$ and $t \in [0, 1]$:

$$\left\|\nabla f(x + ty) - \nabla f(x)\right\| \leq Lt\|y\|,$$

or more generally

$$\left\|f(x) - f(y)\right\| \leq L|x - y|$$

for some norm.

Intuitively, a Lipschitz continuous function is limited in how fast it can change: there exists a Lipschitz constant such that, for every pair of points on

the graph of this function, the absolute value of the slope of the line connecting them is not greater than this constant.[2]

A function $\mathbb{R} \to \mathbb{R}$ is Lipschitz continuous with $L = \sup_x |g'(x)|$ if and only if it has a bounded first derivative.

3.3 Descent Lemma. Let $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz continuous with Lipschitz constant L. Then for all x, y, we have the following quadratic upper bound on the objective function at x:

$$f(x+y) \leq f(x) + \nabla f(x)^\top y + \frac{L}{2}\|y\|^2 .$$

Derivation: let $g(t) = f(x+ty)$ for $t \in [0,1]$.

$$\begin{aligned}
f(x+y) - f(x) &= g(1) - g(0) \\
&= \int_0^1 \frac{\mathrm{d}g}{\mathrm{d}t}(t)\,\mathrm{d}t \\
&= \int_0^1 \nabla f(x+ty)^\top y \,\mathrm{d}t \\
&= \int_0^1 \nabla f(x)^\top y \,\mathrm{d}t + \int_0^1 \left(\nabla f(x+ty)^\top - \nabla f(x)^\top\right) \cdot y \,\mathrm{d}t \\
&\overset{(1)}{\leq} \nabla f(x)^\top y + \int_0^1 \left\|\nabla f(x+ty) - \nabla f(x)\right\| \cdot \|y\| \,\mathrm{d}t \\
&\overset{(2)}{\leq} \nabla f(x)^\top y + \|y\| \int_0^1 Lt\|y\| \,\mathrm{d}t \\
&= \nabla f(x)^\top y + \frac{L}{2}\|y\|^2 ,
\end{aligned}$$

where (1) is an application of the Cauchy-Schwarz theorem ($\langle x, y\rangle \leq \|x\| \cdot \|y\|$), and (2) from Lipschitz continuity.

3.4 Interpretation of Descent Lemma. We can express the lemma in terms of a step sequence $\{x^{(k)}\}$ by substituting $\{x \mapsto x^{(k)}, y \mapsto (x - x^{(k)})\}$:

$$f(x) \leq f(x^{(k)}) + \nabla f(x^{(k)})^\top (x - x^{(k)}) + \frac{L}{2}\left\|x - x^{(k)}\right\|^2 .$$

This is actually a local upper bound of f at x by a quadratic function, which we can optimize analytically:

$$\begin{aligned}
\nabla f(x^{(k)}) + L(x - x^{(k)}) &\overset{!}{=} 0 \\
\Rightarrow x = x^{(k)} - \frac{1}{L}\nabla f(x^{(k)})&
\end{aligned}$$

Convergent step size methods relate to this fact.

[2] `https://en.wikipedia.org/wiki/Lipschitz_continuity`

3.5 Convergence with Constant Step Size. Assume f is Lipschitz continuous, and $\{x^{(k)}\}$ is a sequence generated by a gradient method with gradient related $d^{(k)} \neq 0$. By that we have

$$\begin{aligned} f(x^{(k)} + \alpha^{(k)} d^{(k)}) - f(x^{(k)}) &\leq \overbrace{\nabla f(x^{(k)})^\top d^{(k)}}^{<0} \alpha^{(k)} + \frac{1}{2}(\alpha^{(k)})^2 L \left\| d^{(k)} \right\|^2 \\ &= \underbrace{\alpha^{(k)} \left(\frac{1}{2} \alpha^{(k)} L \left\| d^{(k)} \right\|^2 - \left| \nabla f(x^{(k)}) d^{(k)} \right| \right)}_{A}. \end{aligned}$$

We first calculate the optimal step size $\bar{\alpha}^{(k)} = \min_{\alpha^{(k)}} A$:

$$\begin{aligned} \frac{\partial A}{\partial \alpha^{(k)}} &= (\alpha^{(k)})^2 L \left\| d^{(k)} \right\|^2 - \left| \nabla f(x^{(k)})^\top d^{(k)} \right| \stackrel{!}{=} 0 \\ \Rightarrow \bar{\alpha}^{(k)} &= \left\| d^{(k)} \right\| - \frac{\left| \nabla f(x^{(k)})^\top d^{(k)} \right|}{(\alpha^{(k)})^2 L \left\| d^{(k)} \right\|^2} \end{aligned}$$

Then, for general $\alpha^{(k)}$ with $\epsilon \leq \alpha^{(k)} \leq (2-\epsilon)\bar{\alpha}^{(k)}$, we have

$$\begin{aligned} f(x^{(k)} + \alpha^{(k)} d^{(k)}) - f(x^{(k)}) &\leq \alpha^{(k)} \left(\frac{1}{2} L \left\| d^{(k)} \right\|^2 - \left| \nabla f(x^{(k)})^\top d^{(k)} \right| \right) \\ &\leq \alpha^{(k)} \left(\frac{1}{2}(2-\epsilon) \frac{\left| \nabla f(x^{(k)})^\top d^{(k)} \right|}{(\alpha^{(k)})^2 L \left\| d^{(k)} \right\|^2} L \left\| d^{(k)} \right\|^2 - \left| \nabla f(x^{(k)})^\top d^{(k)} \right| \right) \\ &= \alpha^{(k)} \left(\left| \nabla f(x^{(k)})^\top d^{(k)} \right| - \frac{\epsilon}{2} \left| \nabla f(x^{(k)})^\top d^{(k)} \right| - \left| \nabla f(x^{(k)})^\top d^{(k)} \right| \right) \\ &= \alpha^{(k)} \underbrace{\left(-\frac{\epsilon}{2} \left| \nabla f(x^{(k)})^\top d^{(k)} \right| \right)}_{\leq 0}; \end{aligned}$$

and by the reverse,

$$\begin{aligned} f(x^{(k)} + \alpha^{(k)} d^{(k)}) - f(x^{(k)}) &\geq \alpha^{(k)} \left(\frac{\epsilon}{2} \left| \nabla f(x^{(k)})^\top d^{(k)} \right| \right) \\ &\geq \frac{\epsilon^2}{2} \left| \nabla f(x^{(k)})^\top d^{(k)} \right|, \end{aligned}$$

where the last step results from the assumption about $\alpha^{(k)}$. Thus, $f(x^{(k)}) \geq f(x^{(k+1)}) \geq \cdots$; in each step, the function is decreased by at least $\frac{\epsilon^2}{2} \left| \nabla f(x^{(k)})^\top d^{(k)} \right|$.

Convergence to a stationary point follows by contradiction: assume a subsequence $\{x^{(k)}\}_{k \in \mathcal{K}}$ converged to a point $\bar{x}$ which is non-stationary (ie., for which $\nabla f(\bar{x}) \neq 0$). From above, we know that

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) - f(x^{(k)}) \to 0$$

(assuming f is bounded below), and by that

$$\left|\nabla f(x^{(k)})^\top d^{(k)}\right| \to 0$$

This would however contradict to $d^{(k)}$ being gradient related, since it implies

$$\limsup_{k\in\mathcal{K}\to\infty} \nabla f(x^{(k)})^\top d^{(k)} = 0$$

Therefore, every accumulation point $\bar{x}$ of $\{x^{(k)}\}$ must be stationary ($\nabla f(\bar{x}) = 0$).

3.6 Convergence with Armijo Rule.

TODO

3.7 Convergence Rates. Important for practical problems, to compare different algorithms. Is usually measured in terms of a step-dependent error function $e : \mathbb{R}^n \to \mathbb{R}$ with $e(x^*) = 0$. Common choices are

$$e(x) = \left\|x - x^*\right\|, \text{ or}$$
$$e(x) = f(x) - f(x^*)$$

We are interested in the asymptotic behaviour of e, in terms of how much better the method improves for every step. We have:

1. Sublinear convergence, if

$$\limsup_{k\to\infty} \frac{e(x^{(k+1)})}{e(x^{(k)})} = 1.$$

2. Linear convergence, if

$$\limsup_{k\to\infty} \frac{e(x^{(k+1)})}{e(x^{(k)})} \leq \beta \in (0,1).$$

3. Superlinear convergence, if

$$\limsup_{k\to\infty} \frac{e(x^{(k+1)})}{e(x^{(k)})} = 0.$$

   (which does not imply that the method does not converge!)

Alternatively, we can compare e to a standard sequence of powers: if there exist $q > 0$, $\beta \in (0,1)$, and $p \geq 1$ such that for all k

$$e(x^{(k)}) \leq q\beta^{p^k},$$

we have linear convergence if $p = 1$, and superlinear convergence of order p if $p > 1$. The latter is equivalent to

$$\limsup_{k\to\infty} \frac{e(x^{(k+1)})}{e(x^{(k)})^p} < \infty.$$

derivations for some rates? analysis of quadratic model?

4 Newton's Method and Variants

4.1 Basic Idea. A second order method, which is one of the fastest gradient methods. We generate the sequence $\{x^{(k)}\}$ based on

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)} \left(\nabla^2 f(x^{(k)})\right)^{-1} \nabla f(x^{(k)}),$$

where we assume that the direction $(\nabla^2 f(x^{(k)}))^{-1}\nabla f(x^{(k)}$ is defined and a descent direction.

This has three problems when we are far from a local minimum:

1. The Hessian can be singular – we what are reasonable approximations in this case?
2. The Hessian is convex – it attracts local maxima as well as minima. Thus, we must choose the step size carfully.
3. The direction might not be a descent direction.

Variants of the method deal with this, to ensure convergence globally while maintaining the fast convergence rate.

4.2 Derivation of Newton's Method from Taylor Expansion. By Taylor approximation: Given a point $x^{(k)}$, we can approximate a function $f \in \mathcal{C}^2$ locally as

$$f^k(x) = f(x^{(k)}) + \nabla f(x^{(k)})^\top (x - x^{(k)}) + \frac{1}{2}(x - x^{(k)})^\top \nabla^2 f(x^{(k)})(x - x^{(k)}).$$

Compare this to the Descent Lemma (Summary 3.4) – there, $\nabla^2 f$ is approximated by L, with some transformation of the metric.

We can analytically minimize this approximation, which gives the stated update rule:

$$\begin{aligned}
\nabla f^k(x) = \nabla f(x^{(k)}) + \nabla^2 f(x^{(k)})(x - x^{(k)}) &\stackrel{!}{=} 0 \\
\nabla^2 f(x^{(k)})(x - x^{(k)}) &= -\nabla f(x^{(k)}) \\
(x - x^{(k)}) &= -(\nabla^2 f(x^{(k)}))^{-1}\nabla f(x^{(k)}) \\
x = x^{(k+1)} &= x^{(k)} - (\nabla^2 f(x^{(k)}))^{-1}\nabla f(x^{(k)}).
\end{aligned}$$

4.3 Relation of Newton's Method to Equation Solving. The general for of Newton's method is not used for optimization, but for solving equations of the form $g(x) = 0$, where $g : \mathbb{R}^n \to \mathbb{R}^n \in \mathcal{C}^1$. For this problem, the method has the form

$$x^{(k+1)} = x^{(k)} - (\nabla g(x^{(k)})^\top)^{-1} g(x^{(k)}).$$

A method converging to a stationary point results from setting $g(x) = \nabla f(x)$, implying a symmetric matrix $\nabla g(x)^\top = \nabla^2 f(x)$.

4.4 Local convergence. For a local optimum x^* of f, we must have $\nabla f(x^*) = 0$, or, by the above relation to equation solving, $g(x^*) = 0$. Now, suppose $x^{(k)} \to x$ and $\nabla f(x^*)$ is nonsingular. Expanding g around x^*, we get

$$0 = g(x^*) = g(x^{(k)}) + \nabla g(x^{(k)})^\top (x^* - x^{(k)}) + o\big(\big\|x^* - x^{(k)}\big\|\big)\,;$$

multiplying from the left with $(\nabla g(x^{(k)})^\top)^{-1}$, this is

$$\begin{gathered} x^{(k)} - x^* - (\nabla g(x^{(k)})^\top)^{-1} g(x^{(k)}) = x^{(k+1)} - x^* = o\big(\big\|x^* - x^{(k)}\big\|\big) \\ \Rightarrow \quad \lim_{k\to\infty} \frac{\big\|x^{(k+1)} - x^*\big\|}{\big\|x^{(k)} - x^*\big\|} = 0, \end{gathered}$$

therefore we have superlinear convergence.

4.5 Global Convergence by Diagonal Modifications. To solve the problems mentioned above, when we are far away from an optimum, we can add a diagonal matrix $\Delta^{(k)}$ to the Hessian, such that $\nabla^2 f(x^{(k)}) + \Delta^{(k)} \succ 0$. In that way, we the Newton equation

$$\left(\nabla^2 f(x^{(k)}) + \Delta^{(k)}\right) d^{(k)} = -\nabla f(x^{(k)})$$

can be solved, and $d^{(k)}$ is a descent direction. Some possibilities for $\Delta^{(k)}$ are simply a large enough multiple of I, or more advanced methods like modified Cholesky factorization[3] or a combination of a dampening factor with a trust region[4].

5 Least Squares Optimization and Model Fitting

5.1 General Form. We want to minimize a function

$$f(x) = \frac{1}{2}\big\|g(x)\big\|^2 = \frac{1}{2}\sum_i \big\|g_i(x)\big\|^2\,,$$

for a continuously differentiable function g with components g_i, which can be linear or nonlinear. This is equivalent to solving the problem $g(x) = 0$.

Usually, a least squares problem is a model fitting problem, where $g_i(x) = z_i - h(y_i, x)$, for a model function h with parameters x, samples y_i, and targets and z_i.

[3] `https://www.gnu.org/software/gsl/manual/html_node/Modified-Cholesky-Decomposition.html`

[4] `https://en.wikipedia.org/wiki/Trust_region`

5.2 Gauss-Newton Method. To get a step method, we replace $g(x)$ by a local approximation around $x^{(k)}$:

$$\tilde{g}(x, x^{(k)}) = g(x^{(k)}) + \nabla g(x^{(k)})^\top (x - x^{(k)}).$$

Then we have a quadratic problem:

$$\begin{aligned} x^{(k+1)} &\in \operatorname*{argmin}_{x \in \mathbb{R}^n} \frac{1}{2} \left\| \tilde{g}(x, x^{(k)}) \right\|^2 \\ &= \operatorname*{argmin}_{x \in \mathbb{R}^n} \frac{1}{2} \left\| g(x^{(k)}) + \nabla g(x^{(k)})^\top (x - x^{(k)}) \right\|^2 \end{aligned}$$

$$\begin{aligned} \Rightarrow \qquad 0 &\stackrel{!}{=} \frac{1}{2} \Big(\left\| g(x^{(k)}) \right\|^2 + 2(x - x^{(k)})^\top \nabla g(x^{(k)}) g(x^{(k)}) \\ &\qquad + (x - x^{(k)})^\top \nabla g(x^{(k)}) \nabla g(x^{(k)})^\top g(x^{(k)}) \Big) \\ \Rightarrow \quad x^{(k+1)} &= x^{(k)} - \left(\nabla g(x^{(k)}) \nabla g(x^{(k)})^\top \right)^{-1} \nabla g(x^{(k)}) g(x^{(k)}), \end{aligned}$$

assuming that $(\nabla g(x^{(k)}) \nabla g(x^{(k)})^\top$ is invertible.

5.3 Levenberg-Marquardt Method. To ensure that $\nabla g(x^{(k)}) \nabla g(x^{(k)})^\top$ is invertible, and the descent direction is actually gradient related, one can use a modified iteration scheme

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)} \left(\nabla g(x^{(k)}) \nabla g(x^{(k)})^\top + \delta^{(k)} I \right)^{-1} \nabla g(x^{(k)}) g(x^{(k)}),$$

where $\alpha^{(k)}$ can be chosen by the Armijo rule, and $\delta^{(k)} > 0$ large enough.

5.4 Connection of Gauss-Newton to Newton. The target function of a least squares problem, $f(x) = \frac{1}{2} \left\| g(x) \right\|^2$, with components g_i, has derivatives

$$\begin{aligned} \nabla f(x) &= \nabla g(x)^\top g(x), \\ \nabla^2 f(x) &= \nabla g(x) \nabla g(x)^\top + \sum_i \nabla^2 g_i(x) g_i(x). \end{aligned}$$

Neglecting the second-order part in the Hessian, this reduces to a form equivalent to Newton's method, but saving the computation of the Hessian (intuitively, $\nabla^2 g \approx (\nabla g)^2$).

5.5 Incremental Gradient Methods. If a least squares problem is a model fitting problem, where $g_i(x) = z_i - h(y_i, x)$, for samples y_i, and targets and z_i, we can see each g_i as a data block, and $g = (g_1, \ldots, g_m)$ as the data set. (In machine learning terminology, this corresponds to stochastic gradient descent with minibatches for "data blocks".)

When this data set is large, the Gauss-Newton iterations are costly (because the matrix to be inverted becomes large). Sometimes, for example in real-time applications, the data samples also are not provided in advance, but only incrementally. Therefore, we can use the following scheme to calculate updates based on single samples:

$$\begin{aligned} \xi_0^{(k)} &= x^{(k)} \\ \xi_i^{(k)} &= \xi_{i-1}^{(k)} - \alpha^{(k)} \nabla g_i(\xi_{i-1}^{(k)}) g_i(\xi_{i-1}^{(k)}) \\ x^{(k+1)} &= \xi_m^{(k)}. \end{aligned}$$

This amounts to

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)} \sum_i \nabla g_i(\xi_{i-1}^{(k)}) g_i(\xi_{i-1}^{(k)}),$$

where we base the update on the intermediate $\xi_i^{(k)}$ instead of using the full gradient

$$\nabla f(x^{(k)}) = \sum_i \nabla g_i(x^{(k)}) g_i(x^{(k)}).$$

This converges very fast in the first iterations, but needs further restictions on the step size to ensure global convergence.

5.6 Kalman Filter. If the model (ie. the functions g_i) is linear, one Gauss-Newton iteration is enough to find the least squares estimate (this amounts to the analytical solution using the Moore-Penrose pseudoinverse). However, we can develop an incremental method for this, avoiding the matrix inversion. This method is called *Kalman filter*.

Suppose $g_i(x) = z_i - C_i x_i$, with $z_i \in \mathbb{R}^r$ and $C_i \in \mathbb{R}^{r \times n}$ (the model parameters). We are interested in a method for finding

$$\xi_i \in \operatorname*{argmin}_x \sum_{j=1}^{i} \lambda^{i-j} \left\| z_j - C_j x \right\|.$$

The optimal solution is then $x^* = \xi_m$. For $\lambda = 1$, this is the least squares fit; for $\lambda < 1$, we "decay" the importance of "older" samples. The sequence of ξ_i can be generated iteratively by

$$\begin{aligned} \xi_i &= \xi_{i-1} + H_i^{-1} C_i^\top (z_i - C_i \xi_{i-1}), \quad \text{with} \\ H_i &= \lambda H_{i-1} + C_i^\top C_i, \\ H_0 &= 0, \quad \xi_0 \text{ arbitrary.} \end{aligned}$$

(H_i and C_i are relatively small, compared to $\nabla f(x^{(k)})$.)

Example: for a linear model $x(t) = l + ft$, with parameters f, l, we use $C = [1, t]$.

5.7 Extended Kalman Filter. If the g_i are nonlinear, we can use a linearization at the last available iteration ξ_{i-1} to get the *extended Kalman filter*:

$$\tilde{g}_i(x, \xi_{i-1}) = g_i(\xi_{i-1}) + \nabla g_i(\xi_{i-1})^\top (x - \xi_{i-1}),$$
$$\xi_i \in \operatorname*{argmin}_x \sum_{j=1}^{i} \lambda^{i-j} \left\| \tilde{g}_j(x, \xi_{i_1}) \right\|.$$

The iteration scheme stays the same, but the model parts must be adapted to

$$\tilde{z}_i = \tilde{g}_i(z_i, \xi_{i-1}),$$
$$C_i = -\nabla g_i(\xi_{i-1})^\top.$$

6 Accelerated Gradient Methods

6.1 Lower Bound for Quadratic Optimization.