Summary of OCS Slides

Philipp Gabler

November 17, 2017

Contents

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

Secondly, sections marked with a star are additional material that is maybe useful, but was not covered completely (or at all) in this course. It is mainly taken from the more theoretical course on convex optimization.

1 Mathematical Preliminaries

1.1 Basic Math. Always remember these:

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

Futhermore, it is useful to know that

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \Rightarrow A^{-1} = \frac{1}{\det A} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix},$$

for

$$\det A = ab - cd \neq 0,$$

as well as the Taylor expansion around x_0:

$$f(x) = f(x_0) + \nabla f(x_0)^{\mathrm{T}}(x - x_0) + \frac{1}{2}(x - x_0)^{\mathrm{T}} \nabla^2 f(x_0)(x - x_0) + o(\|x - x_0\|^2).$$

1.2 Definiteness. A symmetric matrix Q is called *positive semidefinite* if $x^{\mathrm{T}} Q x \geq 0$ for all x, and *positive definite* if $x^{\mathrm{T}} Q x > 0$ 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.

The notation can also be used for general inequalities between quadratic forms: $a \preceq Q \preceq b \Leftrightarrow \forall x : a \leq x^{\mathrm{T}} Q x \leq b$, etc.

1.3* Norms. As usual, we have the ℓ_p-norms for $x \in \mathbb{R}^N$:

$$\|x\|_p = \left(\sum_{1 \leq i \leq N} |x_i|^p \right).$$

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

Special cases of this are the *Euclidean norm* ($p = 2$) and the *Manhatton norm* ($p = 1$):

$$\|x\|_2 = \sqrt{x^{\mathrm{T}} x} \quad \text{and} \quad \|x\|_1 = \sum_{1 \le i \le N} |x_i|.$$

The Manhattan norm is used mostly in regularization terms to achieve sparse values. In the limit, there are the ℓ_∞- or *maximum norm* and the ℓ_0-quasinorm:

$$\|x\|_\infty = \max_i \{|x_i|\} \quad \text{and} \quad \|x\|_0 = \operatorname{card}\{i \mid x_i \neq 0\}.$$

The ℓ_0-norm counts the number of nonzero entries in a vector, and is the strongest possible sparsity constraint. It usually is replaced by the ℓ_1-norm in practical applications.

For matrices in $\mathbb{R}^{m \times n}$, we can consider so-called *operator norms*. An operator $A : X \to Y$ between normed spaces is called bounded by a number C if for all $x \in X$

$$\|Ax\|_Y \le C \|x\|_X.$$

The operator norm is the smallest bound of it:

$$\|A\|_{X,Y} = \sup_x \frac{\|Ax\|_Y}{\|x\|_X} = \sup_{\|x\|_X \le 1} \|Ax\|_Y.$$

If A can be described by a matrix, we can use p-norms again and write $\|A\|_{p,q}$.

Based on the operator norm, we can form the *dual norm* of a norm $\|\cdot\|$:

$$\|y\|_* = \sup_{\|x\| \le 1} y^{\mathrm{T}} x,$$

which is the operator norm of y when treated as a dual vector. If $\|\cdot\|$ is an ℓ_p-norm, then the dual is the ℓ_q-norm with $\frac{1}{q} + \frac{1}{p} = 1$.

A further family of matrix norms are the *Schatten p-norms*. For a matrix $A \in \mathbb{R}^{M \times N}$ with rank R, consider the singular value decomposition

$$A = U \Sigma V^{\mathrm{T}},$$

where U, V are orthogonal and $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_R, 0, \ldots, 0) \in \mathbb{R}^{m \times n}$. It holds that $\sigma_i(A) = \sqrt{\lambda_i(A A^{\mathrm{T}})}$. Then

$$\|A\|_{S_p} = \left(\sum_{1 \le i \le R} \sigma_i(A)^p \right).$$

This norm is based solely on the eigenvalues; it measures the stretching of the matrix (described by Σ), while ignoring the rotation part. Special cases here are the *nuclear norm* ($p = 1$)

$$\|A\|_* = \sum_{1 \le i \le R} \sigma_i(A),$$

the *Frobenius norm* ($p = 2$)

$$\|A\|_F = \sqrt{\sum_{1 \le i \le R} \sigma_i^2(A)} = \sqrt{\mathrm{tr}(AA^{\mathrm{T}})} = \sqrt{\sum_{ij} a_{ij}^2}$$

as well as the limit cases of the *spectral norm* and the rank:

$$\|A\|_\infty = \max\{\sigma_1, \ldots, \sigma_R\} \quad \text{and} \quad \|A\|_0 = \mathrm{rank}(A).$$

1.4 Convex Sets. 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.

If C_1 and C_2 are convex sets, then also $C_1 \cap C_2$ and $C_1 + C_2$ (Minkowski sum) are. Examples of convex sets are convex hulls, planes, halfspaces, polyhedra, norm balls, and cones.

1.5 Level Sets. For a 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!

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

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

This means that no points lie below any tangent. Equivalent to this, for sufficiently differentiable functions, are the first-order condition

$$f(x) \ge f(y) + \nabla f(y)^{\mathrm{T}} (x - y), \quad \forall x, y \in X$$

and the second-order condition

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

We speak about *strict convexity* when the inequalities are replaced by strict ones.

1.7 Strong Convexity. A function $f : X \to \mathbb{R}$ is *μ-strongly convex*[2] (for $\mu > 0$) if for all $x, y \in X$ and $\alpha \in [0, 1]$:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{\mu}{2}\alpha(1-\alpha)\left\|x - y\right\|^2,$$

or equivalently for differentiable functions:

$$f(x) \geq f(y) + \nabla f(y)^{\mathrm{T}}(y - x) + \frac{\mu}{2}\left\|y - x\right\|^2 \quad \forall x, y \in X,$$

and twice differentiable functions:

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

The latter is equivalent to requiring that the minimum eigenvalue of $\nabla^2 f(x)$ is at least μ for all x. This definition approaches the definition for strict convexity as $\mu \to 0$, and is identical to the definition of a convex function when $\mu = 0$. On $X = \mathbb{R}$, f is μ-strongly convex if $f''(x) \geq \mu > 0$ for all x.

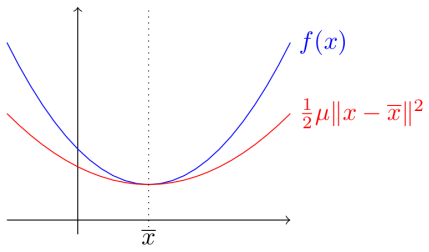

Figure 1: Illustration of the lower bound provided by strong convexity.

Strong convexity means that the function can be bounded from below by a quadratic function: if $\nabla f(x^*) = 0$, then

$$f(x) \geq f(x^*) + \frac{\mu}{2}\|x - x^*\|^2;$$

see Figure 1.

1.8 Lipschitz continuity. A 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 $x, y \in \mathbb{R}^n$

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

for some norm. This is equivalent to

$$\left\|f(x + ty) - f(x)\right\| \leq Lt\left\|y\right\| \quad \forall t \in [0, 1].$$

[2] `https://en.wikipedia.org/wiki/Convex_function#Strongly_convex_functions`

Lipschitz continuity implies continuity.

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.[3]

A differentiable 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.

1.9* Lipschitz-continuous gradients and strong convexity. Let us denote the following classes of *covex* functions:

- $\mathcal{F}_L^{n,k}$: convex n times continuously differentiable functions with L-Lipschitz continuous k-th derivative,
- $\mathcal{S}_\mu^n$: n times continuously differentiable μ-strongly convex functions, and
- $\mathcal{S}_{L,\mu}^{n,k}$: n times continuously differentiable μ-strongly convex functions with L-Lipschitz continuous k-th derivative.

(Where $\mathcal{S}_{L,0}^{n,k} = \mathcal{F}_L^{n,k}$, obviously.)

First, there is a genenal second-order characterization for $f \in \mathcal{S}_{\mu,L}^{2,1}$:

$$\mu \preceq \nabla^2 f(x) \preceq L.$$

Then, we have a range of properties of the classes of convex functions with Lipschitz-continuous gradient and of μ-strongly convex differentiable functions, which are the ones we usually deal with in first-order gradient methods. There are a couple of remarkable symmetries between properties of them, listed below; in each case except the last, the whole statement holds for $f \in \mathcal{S}_{\mu,L}^{1,1}$, but both sides hold individually in $\mathcal{S}_\mu^1$ and $\mathcal{F}_L^{1,1}$, respectively.

1. Differences in gradients are bounded by differences in arguments:

$$\mu\|x-y\|^2 \le \langle \nabla f(x) - \nabla f(y), x-y\rangle \le L\|x-y\|^2$$

2. The descent lemma (Summary 4.2) and its converse, which bound the first order approximation around a point:

$$f(y) \ge f(x) + \langle \nabla f(x), y-x\rangle + \frac{\mu}{2}\|y-x\|^2$$
$$f(y) \le f(x) + \langle \nabla f(x), y-x\rangle + \frac{L}{2}\|y-x\|^2$$

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

3. Their "inverses", bounding the first order approximation by the gradients:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2$$
$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\mu} \|\nabla f(x) - \nabla f(y)\|^2$$

4. A bound on the difference between gradients (the left inequality is known as *co-coercivity*):

$$\frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \frac{1}{\mu} \|\nabla f(x) - \nabla f(y)\|^2$$

5. Finally, we have a combination of bounds with L and μ:

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L} \|x - y\|^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|^2$$

2 Introduction to Optimization

2.1 General Form of Optimization Problems. 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

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

2.2 Types of Optimization Problems.

1. (a) Discrete: X is a discrete set, also called *integer programming*.
   (b) Continuous: X is continuous (ie. uncountable)
2. (a) Linear: Objective functions and constraints are all linear:

$$\min_x c^{\mathrm{T}} x, \quad \text{s.t. } Ax \leq b, \ x \geq 0.$$

   Constraints describe a polyhedron. Polynomially solvable.

   (b) Quadratic: Objective function is quadratic, constraints linear:

$$\min_x \frac{1}{2} x^{\mathrm{T}} Q x + c^{\mathrm{T}} x, \quad \text{s.t. } Ax \leq b, \ Ex = d.$$

   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.

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

2.4 First Order Neccessary Condition for Optimality. If f is continuously differentiable, then in a small neighbourhood of x^*, we can by Taylor expansion write f as

$$f(x) = f(x^* + \Delta x) = f(x^*) + \nabla f(x^*)^{\mathrm{T}} \Delta x + o(\|\Delta x\|).$$

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

$$f(x^*) + \nabla f(x^*)^{\mathrm{T}} \Delta x - f(x^*) = \nabla f(x^*)^{\mathrm{T}} \Delta x \geq 0.$$

Since we can equally choose Δx to have the opposite sign, it holds also that

$$\nabla f(x^*)^{\mathrm{T}} \Delta x \leq 0,$$

so $\nabla f(x^*)^{\mathrm{T}} \Delta x = 0$, which, since Δx is arbitrary, implies that $\nabla f(x^*) = 0$, which is the neccessary condition. A point which has this property is called a *stationary point*. In "normal" cases, it is either a local optimum or a saddle point.

2.5 Second Order Neccessary Condition for Optimality. If f is twice continuously differentiable, then by second order Taylor expansion, we get

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

From this follows that $\Delta x^{\mathrm{T}} \nabla^2 f(x^*) \Delta x \geq 0$. Since Δx is arbitrary, this means that $\nabla^2 f(x^*)$ must be positive semidefinite. The interpretation of this is that the function must be convex in some neighbourhood for for a point to be "at least as good as its neighbours"; if it is strictly convex, we get the following:

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

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

(no "semi-"!), then x^* is a strict unconstrained local minimum of f. (The reason for this is that in this case, f is locally strictly convex, and thus a stationary point is neccessaryly an optimum.)

2.7 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 of x^* being minimal in a local environment. Therefore x^* must also be a global minimum.

Furthermore, the neccessary condition for minima, $\nabla f(x^*) = 0$, for convex functions becomes a sufficient condition: assume $\nabla f(x^*) = 0$, but x^* were not optimal, so there is a y^* with $f(y^*) < f(x^*)$. Then, by convexity of f,

$$f(y^*) \geq f(x^*) + \left\langle \nabla f(x^*), y^* - x^* \right\rangle = f(x^*);$$

this contradicts the assumption, therefore x^* must be optimal.

2.8* Existence of Minimizers. A crucial condition for the general existence of minimizers is the following: a function $f : X \subset \mathbb{R}^N \to \mathbb{R} \cup \{\infty\}$ is *lower semicontinuous (lsc)*, if for all sequences $x_k \in X$ converging to $x \in X$ we have that

$$f(x) \leq \liminf_{k \to \infty} f(x_k).$$

If f is such, we can prove by the Heine-Borel theorem that if X is closed and bounded, f has a global minimizer.

If X is not bounded, we cannot use a compactness argument, but need an additional condition: f is called *coercive* if for every sequence $x_k \in X$ with $\|x_k\| \to \infty$, $f(x_k) \to \infty$. Now, if f is lower semicontinuous and coercive, and X is closed and non-empty, then f also has a global minimizer.

3 Gradient Methods

3.1 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

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

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

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

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

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 5.2).
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 6.3).
4. Diagonal scaling: $D^{(k)} = \mathrm{diag}(d_1^{(k)}, \ldots, d_n^{(k)})$. 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 $D^{(k)} = (\nabla g(x^{(k)} \nabla g(x^{(k)})^{\mathrm{T}})^{-1}$. This is related to the pseudo-inverse (see Summary 6.2.

3.3 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 4.1).

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.

3.4 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

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

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 2).

3.5 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 $\|x - x^*\|$.

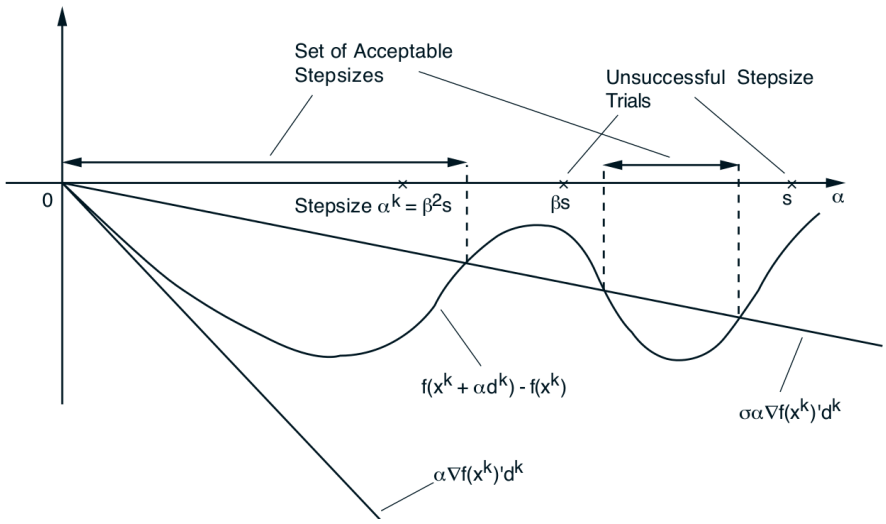

Figure 2: Graphical interpretation of Armijo rule.

4 Convergence Analysis

4.1 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

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

The interpretation of this is that in the limit, $d^{(k)}$ is still a descent direction (see Summary 3.1).

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).

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

$$f(y) \le f(x) + \left\langle \nabla f(x), y - x \right\rangle + \frac{L}{2}\|y - x\|^2,$$

Proof: let $g(t) = f(x + t(y - x))$, so that $g(0) = f(x)$ and $g(1) = f(y)$. Then

$$\begin{aligned}
f(y) &= f(x) + f(y) - f(x) \\
&= f(x) + g(1) - g(0) \\
&= f(x) + \int_0^1 g'(t)\,\mathrm{d}t \\
&= f(x) + \int_0^1 \nabla f(x + t(y - x))^{\mathrm{T}}(y - x)\,\mathrm{d}t \\
&= f(x) + \int_0^1 \big\langle \nabla f(x), y - x \big\rangle\,\mathrm{d}t + \int_0^1 \big\langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \big\rangle\,\mathrm{d}t \\
&\overset{(1)}{\leq} f(x) + \big\langle \nabla f(x), y - x \big\rangle + \int_0^1 \big\|\nabla f(x + t(y - x)) - \nabla f(x)\big\| \cdot \big\|y - x\big\|\,\mathrm{d}t \\
&\overset{(2)}{\leq} f(x) + \big\langle \nabla f(x), y - x \big\rangle + \big\|y - x\big\| \int_0^1 Lt\big\|y - x\big\|\,\mathrm{d}t \\
&= f(x) + \big\langle \nabla f(x), y - x \big\rangle + \frac{L}{2}\big\|y - x\big\|^2,
\end{aligned}$$

where (1) is an application of the Cauchy-Schwarz theorem ($\langle x, y\rangle \leq \|x\| \cdot \|y\|$), and (2) follows from Lipschitz continuity of the gradient (Summary 1.8).

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

$$f(x) \leq f(x^{(k)}) + \nabla f(x^{(k)})^{\mathrm{T}}(x - x^{(k)}) + \frac{L}{2}\big\|x - x^{(k)}\big\|^2.$$

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.

The upper bound of provided by the descent lemma is a global upper bound on the second-order error of the Taylor expansion. In this way, it is exactly dual to μ-strong convexity (Summary 1.7) in the differential formulation, which provides the analog lower bound.

4.4 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

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

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

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

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

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

and by the reverse,

$$f(x^{(k)}) - f(x^{(k)} - \alpha^{(k)} d^{(k)}) \geq \alpha^{(k)} \left(\frac{\epsilon}{2} \left| \nabla f(x^{(k)})^{\mathrm{T}} d^{(k)} \right| \right)$$
$$\geq \frac{\epsilon^2}{2} \left| \nabla f(x^{(k)})^{\mathrm{T}} d^{(k)} \right|,$$

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 an amount of at least $\frac{\epsilon^2}{2} \left| \nabla f(x^{(k)})^{\mathrm{T}} d^{(k)} \right|$.

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

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

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

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

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

4.5 Convergence with Armijo Rule.

TODO

4.6 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?

5 Newton's Method and Variants

5.1 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.

Close to a local minimum, $\alpha^{(k)} = 1$ will work well. However, when we are far from a local minimum, we run into problems:

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.

5.2 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

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

Compare this to the descent lemma (Summary 4.3) – 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:

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

5.3 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

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

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

5.4 Scale Invariance. Newton's method has the property that it is invariant under affine coordinate changes; ie., if we exchange $x = Sy$ for a nonsingular S, the generated steps will remain the same. Proof: First, observe that

$$\begin{aligned}\nabla_y f(Sy) &= S^{\mathrm{T}} \nabla f(Sy),\\ \nabla_y^2 f(Sy) &= S^{\mathrm{T}} \nabla^2 f(Sy) S.\end{aligned}$$

Now, the second-order approximation around $Sy^{(k)}$ becomes

$$\begin{aligned}f^{(k)}(Sy) = f(Sy^{(k)}) &+ (S^{\mathrm{T}} \nabla f(Sy))^{\mathrm{T}} (y - y^{(k)})\\ &+ \frac{1}{2}(y - y^{(k)})^{\mathrm{T}} \nabla^2 f(Sy^{(k)})(y - y^{(k)}),\end{aligned}$$

which we can minimize at $y^{(k+1)}$ to get the update rule, like above:

$$\begin{aligned}&\nabla f^{(k)}(y^{(k+1)}) = S^{\mathrm{T}} \nabla f(Sy^{(k)}) + S^{\mathrm{T}} \nabla^2 f(Sy^{(k)}) S (y^{(k+1)} - y^{(k)}) \stackrel{!}{=} 0\\ \Rightarrow \quad &y^{(k+1)} = y^{(k)} - \left(S^{\mathrm{T}} \nabla^2 f(Sy^{(k)}) S\right)^{-1} S^{\mathrm{T}} \nabla f(Sy^{(k)})\\ &\qquad = y^{(k)} - S^{-1} \nabla^2 f(Sy^{(k)})^{-1} (S^{\mathrm{T}})^{-1} S^{\mathrm{T}} \nabla f(Sy^{(k)})\\ \Rightarrow \quad &Sy^{(k+1)} = Sy^{(k)} - \nabla^2 f(Sy^{(k)})^{-1} \nabla f(Sy^{(k)})\\ \Leftrightarrow \quad &x^{(k+1)} = x^{(k)} - \nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)}).\end{aligned}$$

5.5 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

$$0 = g(x^*) = g(x^{(k)}) + \nabla g(x^{(k)})^{\mathrm{T}} (x^* - x^{(k)}) + o\left(\left\|x^* - x^{(k)}\right\|\right);$$

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

$$\begin{aligned}x^{(k)} - x^* - (\nabla g(x^{(k)})^{\mathrm{T}})^{-1} g(x^{(k)}) = x^{(k+1)} - x^* &= o\left(\left\|x^* - x^{(k)}\right\|\right)\\ \Rightarrow \quad \lim_{k\to\infty} \frac{\left\|x^{(k+1)} - x^*\right\|}{\left\|x^{(k)} - x^*\right\|} &= 0,\end{aligned}$$

therefore we have superlinear convergence.

5.6 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[4] or a combination of a dampening factor with a trust region[5].

6 Least Squares Optimization & Model Fitting

6.1 General Form. We want to minimize

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

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

Often, a least squares problem is a model fitting problem, where

$$g_i(\theta) = h(x_i, \theta) - \hat{y}_i$$

is the loss for a single sample for a model function h with parameters θ, samples x_i, and targets $\hat{y}_i$.

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

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

which yields

$$\begin{aligned}
\tilde{f}(x, x^{(k)}) &= \frac{1}{2}\left\|\tilde{g}(x, x^{(k)})\right\|^2 \\
&= \frac{1}{2}\tilde{g}(x, x^{(k)})^{\mathrm{T}}\tilde{g}(x, x^{(k)}) \\
&= \frac{1}{2}\left(g(x^{(k)})^{\mathrm{T}} + (x - x^{(k)})^{\mathrm{T}}\nabla g(x^{(k)})\right)\left(g(x^{(k)}) + \nabla g(x^{(k)})^{\mathrm{T}}(x - x^{(k)})\right) \\
&= \frac{1}{2}\Big(g(x^{(k)})^{\mathrm{T}}g(x^{(k)}) \\
&\quad + g(x^{(k)})^{\mathrm{T}}\nabla g(x^{(k)})^{\mathrm{T}}(x - x^{(k)}) + (x - x^{(k)})^{\mathrm{T}}\nabla g(x^{(k)})g(x^{(k)}) \\
&\quad + (x - x^{(k)})^{\mathrm{T}}\nabla g(x^{(k)})\nabla g(x^{(k)})^{\mathrm{T}}(x - x^{(k)})\Big) \\
&= \frac{1}{2}\left\|g(x^{(k)})\right\|^2 + 2(x - x^{(k)})^{\mathrm{T}}\nabla g(x^{(k)})g(x^{(k)}) \\
&\quad + (x - x^{(k)})^{\mathrm{T}}\nabla g(x^{(k)})\nabla g(x^{(k)})^{\mathrm{T}}(x - x^{(k)}).
\end{aligned}$$

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

[5] https://en.wikipedia.org/wiki/Trust_region

(Note that ∇g, as the Jacobian, is a matrix!) Then we can turn this into a quadratic problem, yielding the next step as the minimum of the approximation

$$x^{(k+1)} \in \operatorname*{argmin}_{x \in \mathbb{R}^n} \tilde{f}(x, x^{(k)}),$$

which we can solve as

$$\begin{aligned} 0 &\stackrel{!}{=} \nabla_{x^{(k+1)}} f(x^{(k+1)}, x^{(k)}) \\ &= \nabla g(x^{(k)})^{\mathrm{T}} g(x^{(k)}) + \nabla g(x^{(k)}) \nabla g(x^{(k)})^{\mathrm{T}} (x^{(k+1)} - x^{(k)}) \\ \Rightarrow \qquad x^{(k+1)} &= x^{(k)} - \left(\nabla g(x^{(k)}) \nabla g(x^{(k)})^{\mathrm{T}} \right)^{-1} \nabla g(x^{(k)})^{\mathrm{T}} g(x^{(k)}), \end{aligned}$$

assuming that $\nabla g(x^{(k)}) \nabla g(x^{(k)})^{\mathrm{T}}$ is invertible. If g is linear, this method reduces to the Moore-Penrose pseudoinverse and solves the problem in one step.

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

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

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

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

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

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, $\nabla^2 g \approx (\nabla g)^2$).

6.5 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.

6.6 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

$$\xi_i \in \operatorname*{argmin}_x \sum_{j=1}^{i} \lambda^{i-j} \left\| z_j - C_j x \right\|^2.$$

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

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

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

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

6.7 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*:

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

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

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

7 Accelerated Gradient Methods

7.1* General Lower Bounds for Convex Functions. A general first-order gradient method generates steps of the form

$$x^{(k+1)} \in x^{(0)} + \operatorname{span}\{\nabla f(x^{(0)}), \ldots, \nabla f(x^{(k)})\}.$$

We can give the following general lower bounds for convex functions:

- Smooth convex functions: for any k with $1 \le k \le \frac{1}{2}(n-1)$, $L > 0$, and $x^{(0)} \in \mathbb{R}^n$, there is a function $f \in \mathcal{F}_L^{\infty,1}$ such that for any first order method we have

  1. $\left\|x^{(k)} - x^*\right\| \ge \frac{1}{8}\left\|x^{(0)} - x^*\right\|$
  2. $f(x^{(k)}) - f(x^*) \ge \frac{3L\left\|x^{(0)} - x^*\right\|}{32(n+1)^2}$,

- Strongly convex functions: for any $x^{(0)} \in \mathbb{R}^n$, $\mu > 0$, and $L > \mu$, there is a function $f \in \mathcal{S}_{\mu,L}^{\infty,1}$ such that for any first order method we have

  1. $\left\|x^{(k)} - x^*\right\| \ge \left(\frac{\sqrt{L/\mu} - 1}{\sqrt{L/\mu} + 1}\right)^k \left\|x^{(0)} - x^*\right\|$
  2. $f(x^{(k)}) - f(x^*) \ge \frac{\mu}{2}\left(\frac{\sqrt{L/\mu} - 1}{\sqrt{L/\mu} + 1}\right)^{2k} \left\|x^{(0)} - x^*\right\|^2$

  Here, the value $L/\mu \ge 1$ is called the *condition number* of the function and often abbreviated as Q_f.

7.2 Q-Conjugacy. Given a positive definite matrix Q, a set of nonzero vectors $d^{(0)}, \dots, d^{(k-1)}$ are called Q-*conjugate* if for all $i \neq j$

$$(d^{(i)})^{\mathrm{T}} Q d^{(j)} = 0.$$

We can interpret this as orthogonality in the metric given by Q, which induces an inner product $\langle \cdot \rangle_Q$. Since the $d^{(i)}$ are linearly independent in the Q-metric, they are also linearly independent in the normal metric.

Given a set of linearly independent vectors $\xi^{(0)}, \dots, \xi^{(k)}$, we can construct a set of mutually Q-conjugate vectors $d^{(0)}, \dots, d^{(k)}$ spanning the same space by the *Gram-Schmidt procedure*. Therefore, we first define the projection operator onto a vector d (in the Q-metric):

$$\mathrm{proj}_d(\xi) = \frac{\langle d, \xi \rangle_Q}{\langle d, d \rangle_Q} d = \frac{d^{\mathrm{T}} Q \xi}{d^{\mathrm{T}} Q d} d.$$

Then we can iteratively orthogonalize the $\xi^{(i)}$ by removing all projections on the already orthogonalized ones:

$$\begin{aligned} d^{(0)} &= \xi^{(0)} \\ d^{(k)} &= \xi^{(k)} - \sum_{j=0}^{k-1} \mathrm{proj}_{d^{(j)}}(\xi^{(k)}) \\ &= \xi^{(k)} - \sum_{j=0}^{k-1} \frac{(d^{(j)})^{\mathrm{T}} Q \xi^{(k)}}{(d^{(j)})^{\mathrm{T}} Q d^{(j)}} d^{(j)}. \end{aligned}$$

7.3 Conjugate Direction Methods. For *quadratic problems* of the form

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2} x^{\mathrm{T}} Q x - b^{\mathrm{T}} x,$$

these methods accelerate the slow convergence of gradient methods, while still avoiding the overhead of second-order methods, like Newton's. Originally, they were developed for symmetric and positive definite Q, but they also work for solving linear systems of the form $Qx = b$, where Q is not neccessarily symmetric, but just invertible.

For a given set of Q-conjugate descent directions $d^{(0)}, \dots, d^{(k)}$, the conjugate direction method minimizing a quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ is given by the gradient method

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

where $\alpha^{(k)}$ is obtained by line minimization and $x^{(0)}$ is arbitrary.

This method generates steps which minimize the function over a sequence of expanding linear submanifolds $M^{(k)}$ of the domain of f:

$$x^{(k+1)} = \operatorname*{argmin}_{x \in M^{(k)}} f(x),$$

where

$$M^{(k)} = x^{(0)} + \operatorname{span}(d^{(0)}, \ldots, d^{(k)}).$$

Since by construction, the $d^{(i)}$ span the same space as the $x^{(i)}$, $x^{(n)}$ minimizes the function over its whole domain; hence, the method terminates after at most n steps (in theory, assuming exact arithmetic).

7.4 Conjugate Gradient Method. The conjugate gradient method (CG) is the most important conjugate direction method for quadratic problems. The step scheme is

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

where $\alpha^{(k)}$ is chosen by line minimization, and the descent directions are obtained by simply applying the Gram-Schmidt procedure on the gradient vectors at the successive steps: we set $\xi^{(i)} = -g^{(i)}$ for

$$g^{(k)} = \nabla f(x^{(k)}) = Qx^{(k)} - b,$$

and calculate the descent directions $d^{(i)}$ as stated above:

$$\begin{aligned} d^{(0)} &= -g^{(0)}, \\ d^{(k)} &= -g^{(k)} + \sum_{j=0}^{k-1} \frac{(g^{(k)})^{\mathrm{T}} Q d^{(j)}}{(d^{(j)})^{\mathrm{T}} Q d^{(j)}} d^{(j)}. \end{aligned}$$

Since all but one of the coefficients are zero, the sum can be simplified to

$$d^{(k)} = -g^{(k)} + \beta^{(k)} d^{(k-1)},$$

where $\beta^{(k)}$ is given by

$$\beta^{(k)} = \frac{(g^{(k)})^{\mathrm{T}} g^{(k)}}{(g^{(k-1)})^{\mathrm{T}} g^{(k-1)}}.$$

This method terminates with an optimal solution after at most n steps (as argued above), and is the optimal first order method for quadratic problems, for which we have the easy characterization $L = \lambda_{\max}(Q)$ and $\mu = \lambda_{\min}(Q)$: if Q is positive definite, and $0 \preceq \mu \preceq Q \preceq L$, then we have a strongly convex problem with

$$\left\|x^{(k)} - x^*\right\| \leq 2\sqrt{L/\mu} \left(\frac{\sqrt{L/\mu} - 1}{\sqrt{L/\mu} + 1}\right)^k \left\|x^{(0)} - x^*\right\|.$$

If Q is just positive semidefinite with $0 \preceq Q \preceq L$, we have a smooth problem with

$$f(x^{(k)}) - f(x^*) \leq \frac{L\left\|x^{(0)} - x^*\right\|^2}{2(2k+1)^2}.$$

7.5 Application of CG to Nonlinear Problems. For general minimization problems, we can adapt the idea of the above method as follows: $\alpha^{(k)}$ is still chosen by line minimization, but we generate the descent directions via

$$d^{(k)} = -\nabla f(x^{(k)}) + \beta^{(k)} d^{(k-1)},$$

where $\beta^{(k)}$ is computed as

$$\beta^{(k)} = \frac{\nabla f(x^{(k)})^{\mathrm{T}}(\nabla f(x^{(k)}) - \nabla f(x^{(k-1)})}{\nabla f(x^{(k-1)})^{\mathrm{T}} \nabla f(x^{(k-1)})}.$$

Here, we can't anymore say anything about convergence. It might be neccessary to sometimes restart the procedure with $\beta^{(k)} = 0$.

7.6 Heavy Ball Method. In machine learning terminology, this is the same as using a *momentum term*. We use a multi-step iteration scheme

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)} \nabla f(x^{(k)}) + \beta^{(k)}(x^{(k)} - x^{(k-1)}).$$

The idea behind this comes from the physical system of a "heavy ball" with friction: we descretize the differential equation (for a given f):

$$\ddot{x}(t) + \gamma \dot{x}(t) + \nabla f(x(t)) = 0$$

using finite differences:

$$\begin{aligned} \ddot{x}(t) &\approx \frac{x^{(k+1)} - 2x^{(k)} + x^{(k-1)}}{h^2}, \\ \dot{x}(t) &\approx \frac{x^{(k+1)} - x^{(k)}}{h}, \\ \nabla f(x(t)) &\approx \nabla f(x^{(k)}). \end{aligned}$$

We can arrange these terms to get the above form, where α and β are to be chosen. If f is a quadratic problem, we can solve for them analytically: let

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)}(\underbrace{Qx^{(k)} - b}_{r^{(k)}}) + \beta^{(k)}(\underbrace{x^{(k)} - x^{(k-1)}}_{p^{(k)}}).$$

Now, we insert this back into the function:

$$\begin{aligned} (\alpha^{(k)}, \beta^{(k)}) = \operatorname*{argmin}_{\alpha,\beta} \frac{1}{2}&(x^{(k)} - \alpha r^{(k)} + \beta p^{(k)})^{\mathrm{T}} Q (x^{(k)} - \alpha r^{(k)} + \beta p^{(k)}) \\ &- b^{\mathrm{T}}(x^{(k)} - \alpha r^{(k)} + \beta p^{(k)}). \end{aligned}$$

The right hand side has a gradient of

$$\begin{pmatrix} -(r^{(k)})^{\mathrm{T}} Q (x^{(k)} - \alpha r^{(k)} + \beta p^{(k)}) + b^{\mathrm{T}} r^{(k)} \\ (p^{(k)})^{\mathrm{T}} Q (x^{(k)} - \alpha r^{(k)} + \beta p^{(k)}) + b^{\mathrm{T}} p^{(k)} \end{pmatrix},$$

which when set to zero leads to a linear system in α, β with solutions

$$\alpha^{(k)} = \frac{\left\|r^{(k)}\right\|^2 \langle Qp^{(k)}, p^{(k)}\rangle - \langle r^{(k)}, p^{(k)}\rangle \langle Qr^{(k)}, p^{(k)}\rangle}{\langle Qr^{(k)}, r^{(k)}\rangle \langle Qp^{(k)}, p^{(k)}\rangle - \langle Qr^{(k)}, p^{(k)}\rangle^2},$$
$$\beta^{(k)} = \frac{\left\|r^{(k)}\right\|^2 \langle Qr^{(k)}, p^{(k)}\rangle - \langle r^{(k)}, p^{(k)}\rangle \langle Qr^{(k)}, r^{(k)}\rangle}{\langle Qr^{(k)}, r^{(k)}\rangle \langle Qp^{(k)}, p^{(k)}\rangle - \langle Qr^{(k)}, p^{(k)}\rangle^2}.$$

Incidentially, the steps generated by this choice of parameters are exactly those generated by the conjugated gradient method: remember that there, the current step was constructed by Gram-Schmidt out of all previous iterations, of which only the last remained.

The following statement can be given globally for strongly convex functions: if $f \in \mathcal{S}^{2,1}_{\mu,L}$ for $\mu > 0$, and α, β are chosen as

$$\alpha = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2},$$
$$\beta = \left(\frac{\sqrt{L/\mu} - 1}{\sqrt{L/\mu} + 1}\right)^2,$$

then for every $\epsilon > 0$ there is a $c > 0$ such that for all k

$$\left\| \begin{pmatrix} x^{(k+1)} - x^* \\ x^{(k)} - x^* \end{pmatrix} \right\| \leq c(\sqrt{\beta} + \epsilon)^k \left\| \begin{pmatrix} x^{(k)} - x^* \\ x^{(k-1)} - x^* \end{pmatrix} \right\|.$$

By this, the heavy ball method is optimal in the strongly convex case, but does not work well if $\mu = 0$ (then, we can't choose the parameters in the shown way).

7.7 Nesterov's Method. As seen above, the heavy ball method works well for strongly convex differentiable functions, but in general cases, it is hard to set the parameters to get good convergence rates. Nesterov's algorithm is a very similar method, which however requires the objective function just to be smooth and convex, and yields an optimal convergence rate of order $\mathcal{O}(1/k^2)$.

For the algorithm, let $f \in \mathcal{F}^{1,1}_L$. Set $x^{(-1)} = x^{(0)} = y^{(0)}$ arbitrary, and $\alpha = 1/L$. Then:

$$y^{(k)} = x^{(k)} + \beta^{(k)}(x^{(k)} - x^{(k-1)}),$$
$$x^{(k+1)} = y^{(k)} - \alpha \nabla f(y^{(k)}).$$

In essence, this is the same idea as in the heavy ball/momentum method, just that the gradient is evaluated at the extrapolated point $y^{(k)}$ (which is "where the ball would have rolled to"), instead of $x^{(k)}$ itself.

Furthermore, the parameter $\beta^{(k)}$ ("overrelaxation") is chosen in a different, dynamic way, which however does not depend on the steps and converges to 1: for optimal convergence speed in general, one can use

$$t^{(k+1)} = \frac{1+\sqrt{1+4(t^{(k)})^2}}{2},$$
$$\beta^{(k)} = \frac{t^{(k)}-1}{t^{(k+1)}}$$

with $t^{(-1)} = t^{(0)} = 0$. Alternatively, the easier

$$\beta^{(k)} = \frac{k-1}{k+2}$$

ensures convergence, too. Using either choice we achieve a convergence rate

$$f(x^{(k)}) - f(x^*) \leq \frac{2L\left\|x^{(0)} - x^*\right\|^2}{(k+1)^2},$$

so the method is optimal.

If we know the problem to be strongly convex, ie., $f \in \mathcal{S}^{1,1}_{\mu,L}$ with $\mu > 0$, we can use a constant value of

$$\beta^{(k)} = \frac{1-\sqrt{L/\mu}}{1+\sqrt{L/\mu}}$$

to get a linear rate

$$f(x^{(k)}) - f(x^*) \leq (1-\sqrt{L/\mu})^k \left(f(x^{(0)}) - f(x^*) + \frac{\mu}{2}\left\|x^* - x^{(0)}\right\|^2\right).$$

8 Constraint Optimization over Convex Sets

8.1 Basic Setup. Here we consider problems of the form

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

for a set $X \subset \mathbb{R}^n$ that is nonempty, closed, and convex. Usually, X is described by inequality or equality constraints (eg., $\left\|x-c\right\|_2 \leq 1$, or $x \geq 0$). When f is continuously differentiable, the algorithms considered are based on gradient methods for unconstrained minimization, but it will be harder to find fast methods.

The definitions of local and global minima stay essentially the same as above, with the difference that we require x to be *feasible*, ie., $x \in X$.

8.2 Conditions of Optimality. In the unconstrained case, we have observed that a local minimum is characterized by the fact that the cost function is increasing in every direction:

$$\nabla f(x^*)^{\mathrm{T}}(x - x^*) \geq 0, \quad \forall x \in \mathbb{R}^n.$$

This leads to the conclusion that $\nabla f(x^*) = 0$. In the constrained case, we need to consider the neccessary condition

$$\nabla f(x^*)^{\mathrm{T}}(x - x^*) \geq 0, \quad \forall x \in X;$$

ie., the gradient *in all feasible directions* must be less than 0. If f is convex over X, then this condition is also sufficient.

For illustration, see Figure 3: if a local minimum is at the boundary of X (which will be the most interesting case, otherwise the problem is not much different from an unconstrained one), we only need to consider directions (other points) inside the boundary, which for a convex set implies that the gradient makes an an angle of 90° or less with the feasible directions. In the non-convex case, this condition can fail.

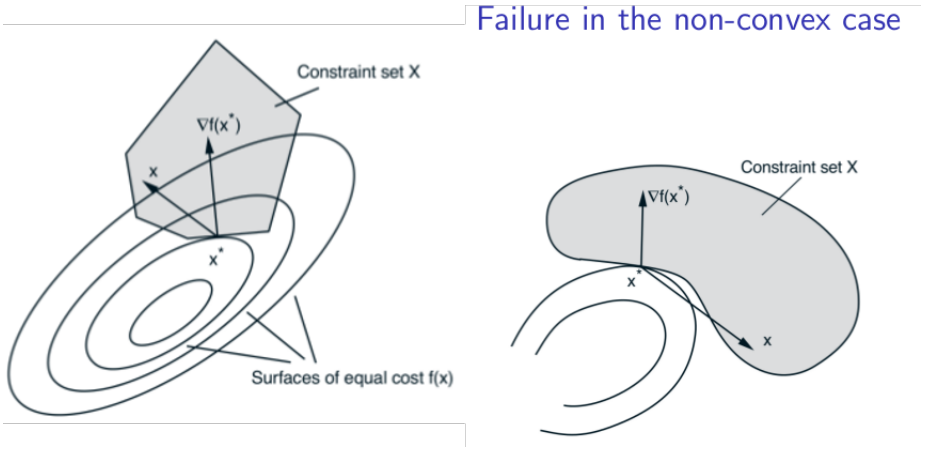

Figure 3: Illustraion of the neccessary condition for constraint local minima and when it fails.

8.3 Projection on a Convex Set. This operation is a subproblem of many constraint optimization algorithms. Let $z \in \mathbb{R}^n$. We call

$$\mathrm{proj}_X(z) = \min_{x \in X} \|x - z\|^2 \quad \text{s.t. } x \in X$$

the projection of z onto X; it is the point in X which is closest to z.

8.4 Projection Theorem. For a nonempty, closed $X \subseteq \mathbb{R}^n$, there is for every $z \in \mathbb{R}^n$ a unique $x^* = \mathrm{proj}_X(z)$; this follows from the strong convexity of the projection operator.

Furthermore, given z, $x^* = \mathrm{proj}_X(z)$ if and only if

$$(z - x^*)^{\mathrm{T}}(x - x^*) \leq 0, \quad \forall x \in X;$$

if X is a sub-vectorspace of $\mathbb{R}^n$, this reduces to

$$(z - x^*)^{\mathrm{T}} x \leq 0, \quad \forall x \in X.$$

The second condition follows from specializing the optimality condition to the projection: for $f(x) = \frac{1}{2}\|x - z\|^2$, we have $\nabla f(x) = x - z$. The optimility condition for this is

$$\begin{aligned} &\nabla f(x^*)^{\mathrm{T}}(x - x^*) \geq 0 \\ \Leftrightarrow\, &(x - z)^{\mathrm{T}}(x - x^*) \geq 0 \\ \Leftrightarrow\, &(z - x)^{\mathrm{T}}(x - x^*) \leq 0. \end{aligned}$$

The third condition follows from the fact that for all $x \in X$, both $x^* + x$ and $x^* - x$ are in X.

8.5 Quadratic Programming with Equality Constraints.

8.6 Feasible Direction Methods. Definition, alternative definition, step sizes, convergence

8.7 Conditional Gradient Method.

8.8 Gradient Projection Method. + scaled variants

8.9 Linear Programming with Equality Constraints. Intro, as a sequence of quadratic problems

8.10 Affine Scaling Method.

8.11 Linear Programming with Inequality Constraints.