Hints for OCS Questions

Julian Wolf Philipp Gabler

November 16, 2017

1 Basics

Question 1. *What is the definition of a general mathematical optimization problem? Give an example and explain the notion of an objective function, a constraint set, an optimal solution and the definition of the level sets of a function.*

Draw level lines and arrows

- General form:
$$\min_x f(x) \quad \text{s.t. } x \in X,$$
- Objective function f, constraint set X.
- A solution x^* is globally optimal if $f(x^*) \leq f(x), \forall x \in X$ (locally, if this holds in an environment in X around x^*).
- Level sets: for $c \in \mathbb{R}$, the sets
$$S_c(f) = \{x \in X : f(x) = c\}$$
of values with equal outcome.

Question 2. *Explain nonlinear programming, linear programming, quadratic programming, convex programming and give examples. What is the definition of a convex set and a convex function? Give examples for convex sets and convex functions.*

- Linear: objective functions and constraints are all linear:
$$\min_x c^{\mathrm{T}} x, \quad \text{s.t. } Ax \leq b,\ x \geq 0.$$
- 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.

- Nonlinear: no further requirements – objective function and constraints may be arbitrary. Usually used if not known whether the problem is convex. Not much theory available in general form.
- Convex set: 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 function: $f : X \to \mathbb{R}$ is 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.

Question 3. *What is the difference between local and global minima. Give examples. Give the first order necessary condition of optimality and prove it. What is the second order necessary condition of optimality? Show that for differentiable convex functions, the first order necessary condition of optimality becomes sufficient.*

- A point x^* is called an global minimum of f if for all $x \in X$

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

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

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

- First order neccessary condition: 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.
- Second order neccessary condition: $\nabla^2 f(x^*)$ must be positive semidefinite.

- 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.

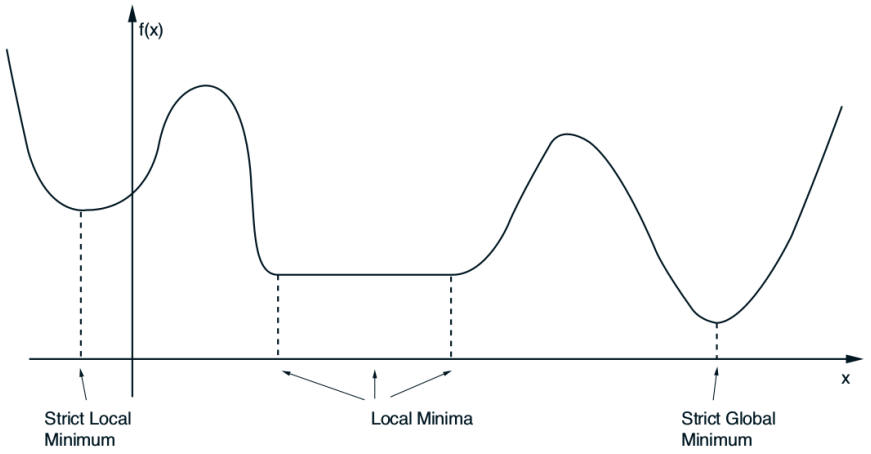

Figure 1: Local/global minima.

Question 4. *Discuss the optimality conditions for a quadratic optimization problem of the form*

$$\min_x \frac{1}{2} x^T Q x - b^T x.$$

When is this problem convex and what does convexity imply? Give a simple example in 2D showing different realizations of Q.

Example for $f : \mathbb{R} \to \mathbb{R}$: in this case, $f(x) = \frac{1}{2} q x^2 + bx$. We have:

- First order necessessary optimality condition: $\nabla f(x^*) = qx - b = 0$.
- Second order necessessary optimality condition: $\nabla^2 f(x^*) = q \geq 0$.

Since this is a parabola, we have three cases:

1. $q > 0$: Unique minimum exists as minimum of upwards parabola.
2. $q = 0$: Horizontal line, every point is locally minimal.
3. $q < 0$: Downwards parabola, not bounded from below, so no minimum.

Example for $f : \mathbb{R} \to \mathbb{R}^2$:

- First order necessessary optimality condition: $\nabla f(x^*) = Qx - b = 0$

- Second order necessessary optimality condition: $\nabla^2 f(x^*) = Q$ is positive semidefinite.

In more concrete form: if $Q = \mathrm{diag}(\alpha, \beta)$, $b^{\mathrm{T}} = [1, 0]$, then

$$f(x) = \frac{1}{2}(\alpha x^2 + \beta y^2) + x.$$

Q is positive definite if all eigenvalues, which in this case are α and β, are positive. See Figure 2. The second case is not bounded from below; the fourth case is a saddle surface.

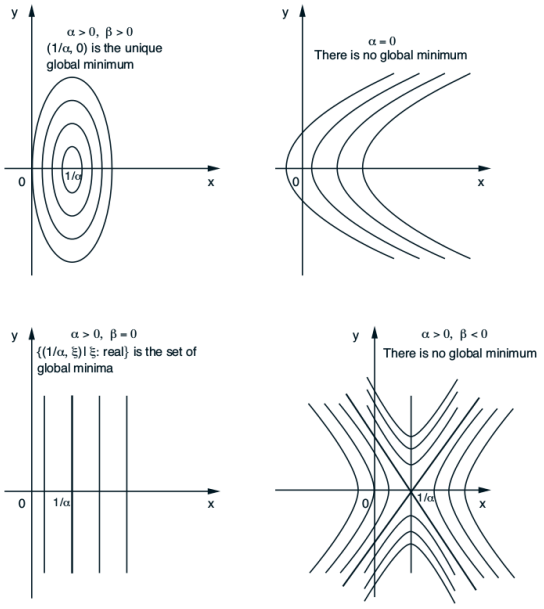

Figure 2: Different scenarios for Q.

2 Gradient Methods, Optimality

Question 5. *What is a descent direction? Draw a simple example explaining the properties of a descent direction. Give the general form of a gradient method and show that $d^{(k)} = -D^{(k)}\nabla f(x_k)$ with $D^{(k)}$ symmetric and positive definite is a descent direction.*

- $d^{(k)}$ is a descent direction if

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

  Interpretation: $d^{(k)}$ goes more downhill than uphill; see Figure 3.

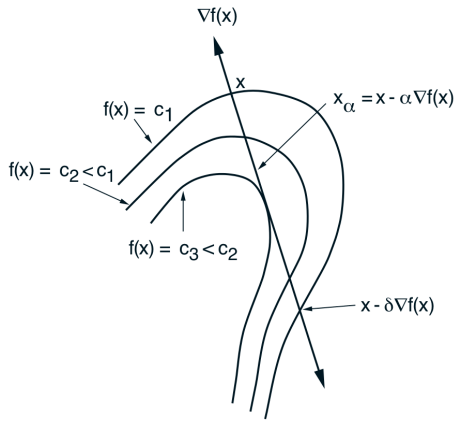

Figure 3: Illustration of a descent direction.

- General form of gradient method:
  1. Choose an initial vector $x^{(0)} \in \mathbb{R}^n$
  2. Choose a descent direction $d^{(k)}$ that satisfies $\nabla f(x^{(k)})^{\mathrm{T}} d^{(k)} < 0$
  3. Choose a positive step size $\alpha^{(k)}$
  4. Compute the new vector as
  $$x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}$$
  5. Set $k = k + 1$ and goto 2, until some termination criterion is fulfilled
- We have
  $$\nabla f(x^{(k)})^{\mathrm{T}} d^{(k)} = -\nabla f(x^{(k)})^{\mathrm{T}} D^{(k)} \nabla f(x^{(k)}) < 0,$$
  by direct application of the definition of positive definiteness.

Question 6. *Give three different standard choices for descent directions based on choosing the scaling matrix $D^{(k)}$ and discuss their numerical performance. What is the Armijo step size selection rule? Draw an example explaining the set of acceptable step sizes.*

Possible descent scalings:

- $D^{(k)} = I$: steepest descent, slow convergence if the level lines are elongated (leads to zig-zagging).
- $D^{(k)} = \nabla^2 f(x^{(k)})$: Newton's method, very fast convergence near minima. Unstable in wrt. initial values (may diverge). Requires recalculation of inverse of Hessian in every step – very expensive in large dimensions.

- $d^{(k)} = \left(\frac{\partial^2 f(x^{(k)})}{\partial (x_i)^2}\right)^2$: diagonal scaling, an approximation of Newton's method, but usually not worth it (scales only in diagonal directions).
- $d^{(k)} = \left(\nabla g(x^{(k)}) \nabla g(x^{(k)})^{\mathrm{T}}\right)^{-1}$: Gauss-Newton method, for least squares problems with design function g. Good performance; again calculation of inverse, but not of the Hessian.

Armijo rule: it is not sufficient to simply ensure that $f(x^{(k+1)}) < f(x^{(k)})$. 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

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

Thus, the step sizes are chosen such that the energy decrease is sufficiently large – see Figure 4.

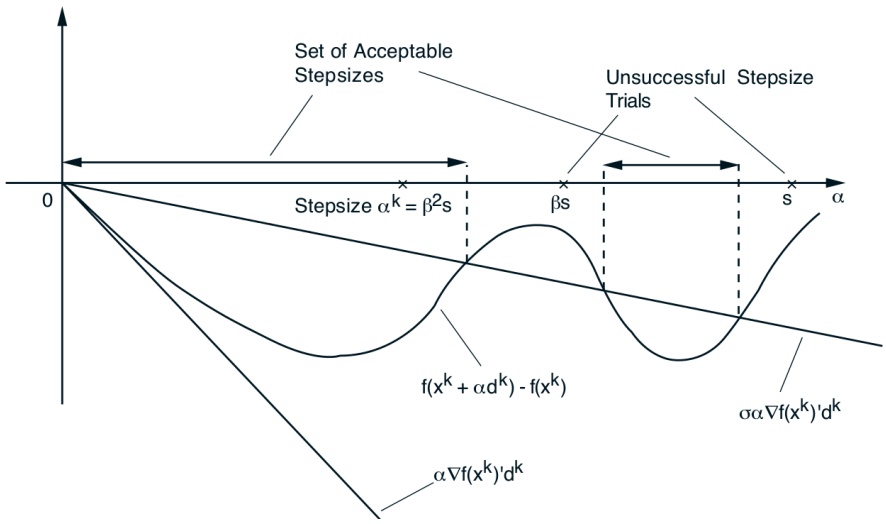

Figure 4: Graphical representation of the idea of the Armijo rule.

3 Convergence

Question 7. *What is a Lipschitz continuous gradient, and what is the descent lemma? Give the proof of the descent lemma.*

A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ has Lipschitz continuous gradient if there is a Lipschitz constant $L \geq 0$, such that for all $x, y \in \mathbb{R}^n$

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

or equivalently

$$\big\|\nabla f(x+ty) - \nabla f(x)\big\| \le Lt\big\|y\big\| \quad \forall t \in [0,1]$$

for some norm.

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:

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

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

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

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

Question 8. *What is a rate of convergence? Explain linear, superlinear, and sublinear convergence and give examples.*

We measure convergence in terms of asymtotic rates of a step-dependent error function $e : \mathbb{R}^n \to \mathbb{R}$ with $e(x^*) = 0$. The usual choices are

$$\begin{aligned}
e(x) &= \big\|x - x^*\big\|, \text{ or} \\
e(x) &= f(x) - f(x^*)
\end{aligned}$$

Classes of convergence rates (see Figure 5):

1. Sublinear: $\limsup_{k\to\infty} \frac{e(x^{(k+1)})}{e(x^{(k)})} = 1$. Example: $e(x^{(k)}) \sim 1/(k+1)^2$ (black curve; this does not imply that the method does not converge!).
2. Linear: $\limsup_{k\to\infty} \frac{e(x^{(k+1)})}{e(x^{(k)})} \le \beta \in (0,1)$. Example: $e(x^{(k)}) \sim 1/2^k$ (blue curve)

3. Superlinear: $\limsup_{k\to\infty} \frac{e(x^{(k+1)})}{e(x^{(k)})} = 0$. Example: $e(x^{(k)}) \sim 1/2^{2^k}$ (red curve).

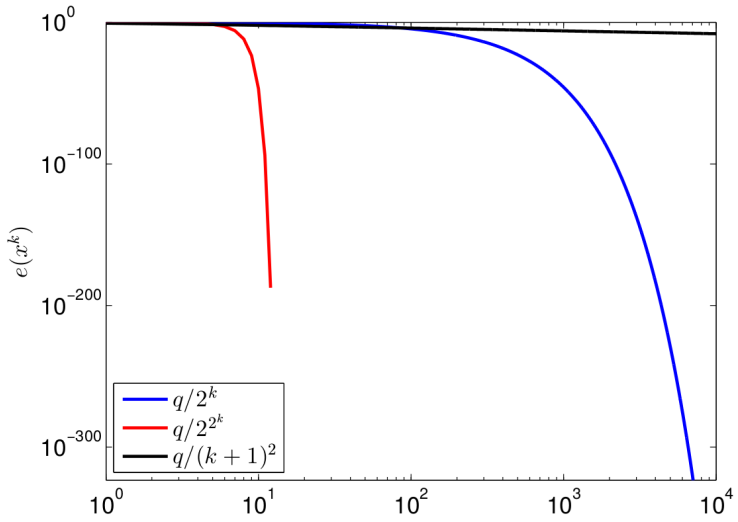

Figure 5: Graphical representation of linear, superinear and sublinear convergence

4 Newton's method

Question 9. *Show that the plain form of Newton's method can be derived from a second order Taylor approximation of the objective function. Show that Newton's method is invariant with respect to an affine change of the coordinate system.*

Newton's method is base on the sequence

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

By Taylor approximation, given a point $x^{(k)}$, we can approximate a function $f \in \mathcal{C}^2$ locally as

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

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

rule:

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

Scale invariance: if we exchange $x = Sy$ for a nonsingular S, the generated steps will remain the same. 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:

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

5 Least squares problems

Question 10. *What are linear (and nonlinear) least squares problems? Give an example. What is the Gauss-Newton method, and what is its relation to Newton's method?*

We want to minimize

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

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 $\hat{y}_i$. If we take, e.g., a linear predictor, we would have the form $g_i(\theta) = \theta^{\mathrm{T}} x_i - \hat{y}_i$, which we can put together in matrix form as $g(\theta) = Ax - y$ with the so-called model matrix A.

To derive the Gauss-Newton method, we replace g by a local approximation around $x^{(k)}$:

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

This leads to a step method of the form

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

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

To compare with Newton's method, look at the 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 full Hessian (intuitively, $\nabla^2 g \approx (\nabla g)^2$).

Question 11. *What is a Kalman filter? How does it relate to an optimization problem? What is an extended Kalman filter?*

If a least-squares model (ie. the functions g_i) is linear, one Gauss-Newton iteration would be enough to find the optimal solution (this amounts to the analytical solution using the Moore-Penrose pseudoinverse). However, we can develop an incremental method for this, avoiding the matrix inversion, and instead using one sample at a time.

Concretely, for $g_i(x) = z_i - C_i x_i$, we have a step method

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

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

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

$$\begin{aligned}
\tilde{z}_i &= g_i(\xi_{i-1}) + \nabla g_i(\xi_{i-1})^{\mathrm{T}} (z_i - \xi_{i-1}), \\
C_i &= -\nabla g_i(\xi_{i-1})^{\mathrm{T}}.
\end{aligned}$$

6 Accelerated gradient methods

Question 12. *What is the lower bound of first order methods on quadratic problems? What is an optimal algorithm for quadratic problems?*

A quadratic minimization problem

$$\min_x f(x) = \frac{1}{2}x^{\mathrm{T}}Qx - b^{\mathrm{T}}x,$$

is automatically convex for positive semidefinite Q. The lower bounds for first-order gradient methods then depend only on the eigenvalues of Q, since they fix $L = \lambda_{\max}$ and $\mu = \lambda_{\min}$: if $Q \succeq 0$ (just positive semidefinite), then

$$f(x^{(k)}) - f(x^*) \geq \frac{3\lambda_{\max}\left\|x^{(0)} - x^*\right\|}{32(n+1)^2}.$$

If $Q \succ 0$ (positive definite, therefore a strongly convex problem), then

$$\left\|x^{(k)} - x^*\right\| \geq \left(\frac{\sqrt{\lambda_{\max}/\lambda_{\min}} - 1}{\sqrt{\lambda_{\max}/\lambda_{\min}} + 1}\right)^k \left\|x^{(0)} - x^*\right\|.$$

An optimal first order method for these cases is the conjugate gradient method.

Question 13. *Write down the conjugate gradient (CG) method and specialize the algorithm for solving a least-squares problem of the form*

$$\min_x f(x) = \frac{1}{2}x^{T}Qx - b^{T}x.$$

What is the relation to solving linear system of equations?

We set

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

and calculate the descent directions as

$$\begin{aligned} d^{(0)} &= -g^{(0)}, \\ d^{(k)} &= d^{(k)} = -g^{(k)} + \beta^{(k)}d^{(k-1)}, \end{aligned}$$

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

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

When solving a linear system $Ax = y$, we are essentially also looking for an x which minimizes the expression $\|Ax - y\|$. This is a special form of a quadratic problem, which we can solve using CG.

Question 14. *Explain the difference between the heavy-ball algorithm and Nesterov's algorithm. What are the rates of convergence of those algorithms on strongly convex problems?*

- Heavy-ball: idea is like in physics: a ball uses its momentum it gained beforehand to overcome small increases or flat areas of its way:

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

- Nesterov: instead of calculating the gradient at the current point, use a gradient step based at the point extrapolated from the momentum:

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

- The heavy-ball-method is only optimal for strongly convex functions, since otherwise the step size parameters cannot be chosen correctly. Nesterov overcomes this difficulty by a dynamic choice of $\beta^{(k)}$.
- Both algorithms are optimal for strongly convex functions; Nesterov is also optimal for general smooth convex functions.

7 Contrained optimization

Question 15. *Give an example showing the necessary optimality condition for minimizing a differentiable function over a convex set. Why does it fail in case the feasible set is non-convex?*

- the gradient $\nabla f(x^*)$ makes an angle ≤ 90 in all feasable points
- this condition is in general not reachable

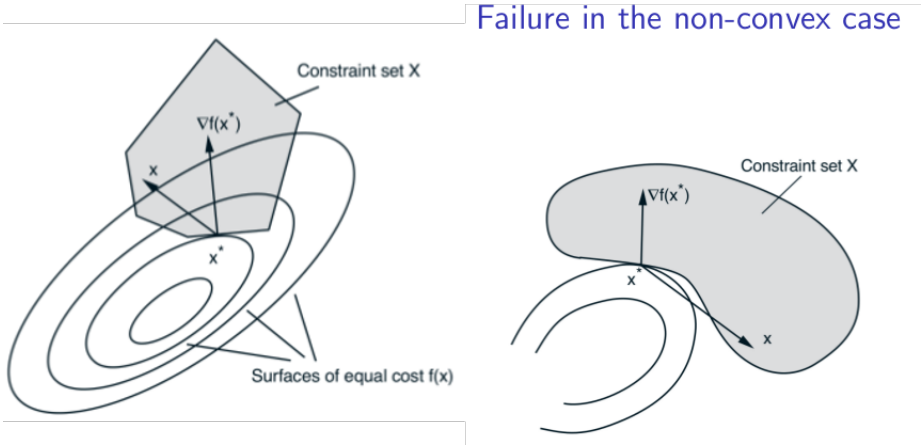

Figure 6: Graphical representation of convex and non-convex set

Question 16. *What is a projection on a convex set? Give the optimality condition and specialize the condition to the case where the convex set is a subspace. What are the properties of a subspace?*

- z is a fixed vector, find vector x^* in a closed convex set X
- $min_x f(x) = ||x - z||^2$

Question 17. *What is a feasible direction? Give an example. What is the general form of a feasible direction method? Also give an alternative form of the feasible direction based on a feasible vector $\bar{x}$.*

middle/end of pages slide 10 - start in interior and just take small steps $- >$ we can ignore constraint under these conditions

- Given a feasible vector x, a feasible direction at x is a vector d such that the vector $x + \alpha d$ is feasible for all sufficiently small $\alpha > 0$.
- a feasable method generates starts at x^0 and generates multiple such points x^{k+1}

Question 18. *Explain the conditional gradient method and the projected gradient method. What is different? For both methods draw a simple example showing how the feasible directions are computed.*

- Conditional gradient solves subproblem with linear cost, gradient projection method solves quadratic cost fct
- The conditional gradient method generates the point $\overline{x}^k$ by finding a feasible point which is furthest way from x^k along the negative gradient direction $-\nabla f(x^k)$.

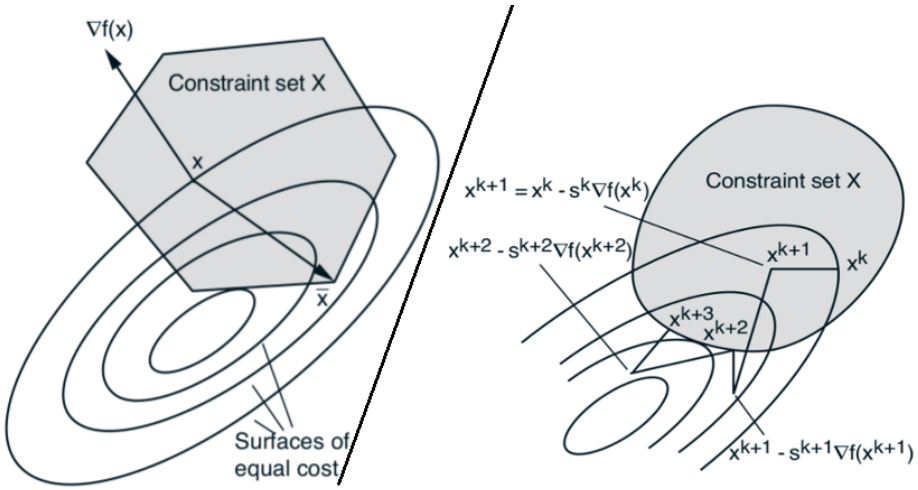

Figure 7: Graphical representation of conditional (left) and projective (right) method

Question 19. *What is the affine scaling method for solving an equality constrained LP? Show how the LP is solved based on solving a sequence of linearly constrained quadratic programs. Why can the inequality constraint be skipped?*

- iterative method $x^{k+1} = x^k + \alpha^k (H^k)^{-1}$(`big AHA formula`)
- affine scaling: choose $H^k = (X^k)^{-2}$, $X^k = diag(x_1, ..., x_n)$ leads to $y^{k+1} = y^k + \alpha^k (AX^k A')^{-1} b$, α^k ensures $x^{k+1} > 0$

Question 20. *What is the Lagrange multiplier theorem for equality constrained optimization problems? Draw a simple example and explain why the gradients of the constraint functions need to be linearly independent.*

- Interpretation 1: The gradient of the cost function $\nabla f(x^*)$ belongs to the subspace spanned by the gradients of the constraint functions $\nabla h_i(x^*)$
- Interpretation 2: The cost gradient $\nabla f(x^*)$ is orthogonal to the subspace of first order feasible directions
- for failure see figure 8. The Eigenvectors are lineary dependent, we loose one dimension and thus we can not optimize the problem (at least I think so)

Question 21. *Show how to solve the projection problem:*

$$\min_x \frac{1}{2} \|x - y\|^2$$

Write down the Lagrangian, give the KKT conditions and show how the problem is solved.

Question 22. *Show how to compute the projection onto a half space:*

$$\min_x \frac{1}{2}\|x-y\|^2 \quad s.t. \quad a^T x = b$$

Write down the Lagrangian, give the KKT conditions and show how the problem is solved.

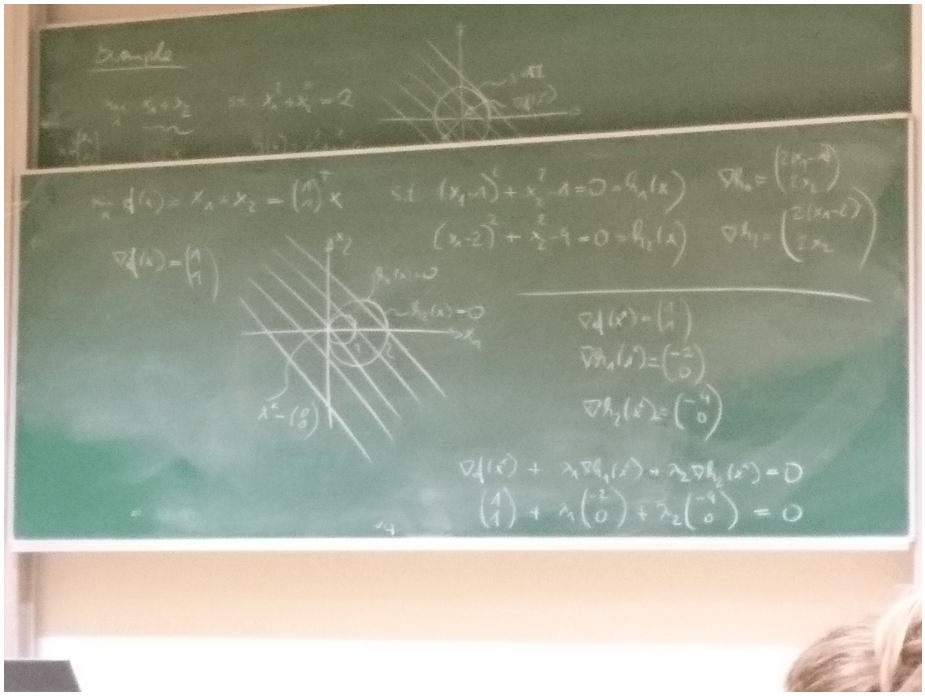

Figure 8: Example1, 24.01.2017

Figure 9: Example 2, 24.01.2017