Hints for OCS Questions

Julian Wolf Philipp Gabler

December 10, 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.*

- 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 $f(x^*) \leq f(x)$ for all $x \in X$, and locally, if this holds in a neighbourhood 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, and 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:
$$\min_x c^{\mathrm{T}} x, \quad \text{s.t. } Ax \leq b,\ x \geq 0,$$
for some constraint matrix A and vector b.
- Quadratic: objective function is quadratic, constraints linear:
$$\min_x \frac{1}{2} x^{\mathrm{T}} Q x + c^{\mathrm{T}} x, \quad \text{s.t. } Ax \leq b,\ Ex = d,$$
for some constraint matrices A and E, and vectors b and d.

- Convex: objective function f and constraint set X are convex:

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

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

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

  (see Figure 1).
- 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

$$\begin{aligned} 0 &\leq f(x^* + \Delta x) - f(x^*) \\ &\leq f(x^*) + \nabla f(x^*)^{\mathrm{T}} \Delta x - f(x^*) \\ &= \nabla f(x^*)^{\mathrm{T}} \Delta x. \end{aligned}$$

  Since we can equally choose Δx to have the opposite sign, we can do the same proof using $f(x^* - \Delta x)$, which results in

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

  Combining both inequalities implies $\nabla f(x^*)^{\mathrm{T}} \Delta x = 0$. Since Δx is arbitrary, that implies that $\nabla f(x^*) = 0$, which is the neccessary condition.

- Second order neccessary condition: $\nabla^2 f(x^*)$ must be positive semidefinite.
- For f convex: assume $\nabla f(x^*) = 0$, but x^* were not optimal, so there is a y^* with $f(y^*) < f(x^*)$. Then, by the first order convexity condition of f (see Summary: "Convex Functions"),

$$\begin{aligned} f(y^*) &\geq f(x^*) + \underbrace{\nabla f(x^*)^{\mathrm{T}}}_{=0}(y^* - x^*) \\ &= f(x^*); \end{aligned}$$

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

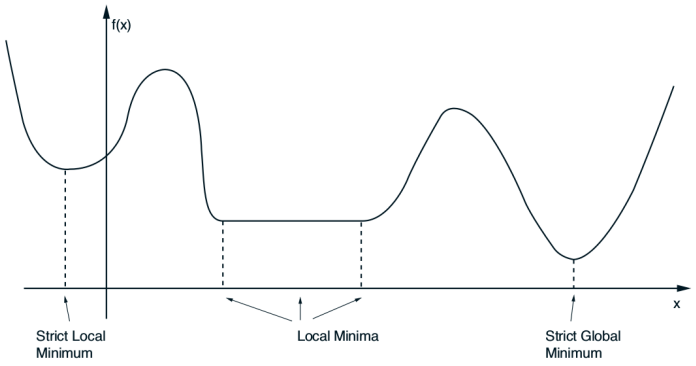

Figure 1: Local/global minima.

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

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

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. It is positive semidefinite if α, β, and $\alpha\beta$ are nonnegative. See Figure 2: the second case is not bounded from below; the fourth case is a saddle surface.

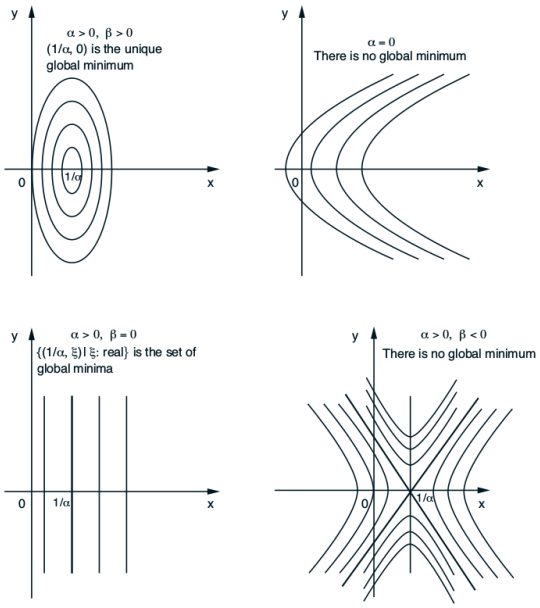

Figure 2: Different scenarios for Q.

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.

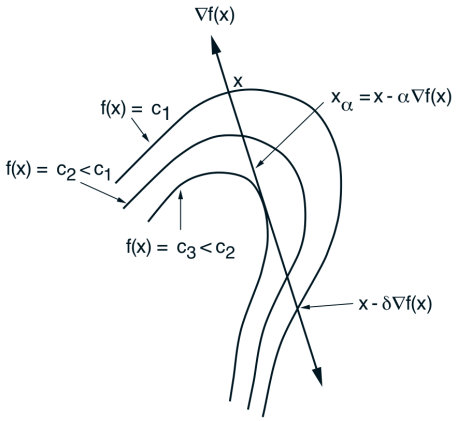

Figure 3: Illustration of a descent direction.

- 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
  $$\nabla f(x^{(k)})^{\mathrm{T}} d^{(k)} = -\nabla f(x^{(k)})^{\mathrm{T}} D^{(k)} \nabla f(x^{(k)}) < 0,$$
  which follows directly from the assumption of $D^{(k)}$ being positive definite (since $x^{\mathrm{T}} D^{(k)} x > 0$ for all x).

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), but easy to evaluate.

- $D^{(k)} = (\nabla^2 f(x^{(k)}))^{-1}$: 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. Hessian might also become singular.
- $d^{(k)} = \left(\frac{\partial^2 f(x^{(k)})}{\partial (x_i)^2}\right)^{-1}$: 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 an initial step size s, a reduction factor $0 < \beta < 1$, and a scalar $0 < \sigma < 1$, and choose $m^{(k)}$ as the first nonnegative integer for which

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

Then, set $\alpha^{(k)} = \beta^{m^{(k)}} s$. Thus, the step sizes are chosen such that the energy decrease is sufficiently large – see Figure 4.

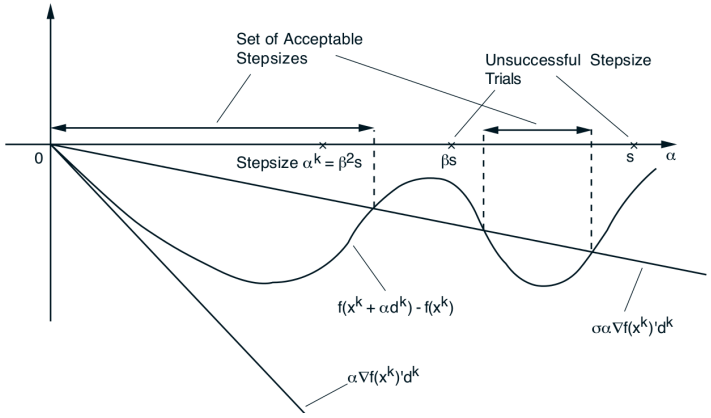

Figure 4: Graphical representation of the idea of the Armijo rule.

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 around 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

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

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:

$$\begin{aligned}
\big\|\nabla f(x + t(y-x)) - \nabla f(x)\big\| &\le L\big\|x + t(y-x) - x\big\| \\
&= L\big\|t(y-x)\big\| \\
&= Lt\|y - x\|.
\end{aligned}$$

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

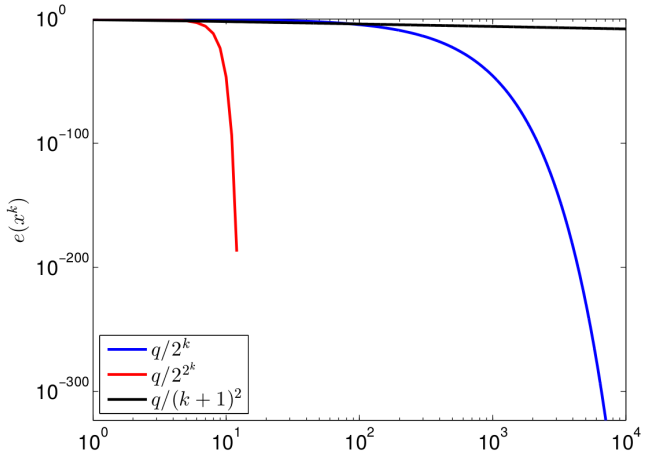

Figure 5: Graphical representation of linear, superinear and sublinear convergence

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 a fast second order method based on the sequence

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

It is also scale invariant: if we used a transformation of variables $Sy = x$, such that $(f \circ S)(y) = f(x)$, where S is a nonsingular affine transformation, the generated steps will remain the same.

Derivation including scale invariance proof: first, observe that by chain rule,

$$\nabla(f \circ S)(y) = S^{\mathrm{T}}(\nabla f \circ S)(y),$$
$$\nabla^2(f \circ S)(y) = S^{\mathrm{T}}(\nabla^2 f \circ S))(y)S.$$

Now, the second-order approximation of $f \circ S$ around $y^{(k)}$ becomes

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

We can then set $y^{(k+1)}$ as the minimizer of $T^{(k)}$ to get the update rule, like above:

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

Here we can substitute back $x^{(i)}$ for $Sy^{(i)}$ to get the formula from above:

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

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

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

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

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

In the positive semindefinite case, Nesterov's method is optimal; in the positive definite (i.e., strongly convex) case, also the conjugate gradient and heavy ball methods are optimal (they are equivalent in this case).

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

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

What is the relation to solving linear system of equations?

The step scheme for the CG method 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 applying the Gram-Schmidt procedure on the gradient vectors at the successive steps.

For a quadratic problem, we set

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

and calculate the descent directions as

$$d^{(0)} = -g^{(0)},$$

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

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

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

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: inspired by physics. A ball uses the momentum it gained beforehand to overcome small increases or flat areas on its way; this is achieved by including the difference to the second-last term:

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

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

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

- Both algorithms are optimal for strongly convex functions, with

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

  for Nesterov's method and

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

  for the heavy-ball method (for optimal choices of parameters). Note that while the latter rate appears to be better, they are asymptotically of the same order.

- However, only Nesterov's method is optimal for non-strong problems, since the step size parameters cannot be chosen correctly in the other case (since $\mu = 0$). It overcomes this difficulty by a dynamic choice of $\beta^{(k)}$, leading to a still optimal rate of

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

7 Constrained 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 neccessary condition for optimality for a convex constraint set X is

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

ie., the gradient in all feasible directions must be increasing.

For illustration, see Figure 6: 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.

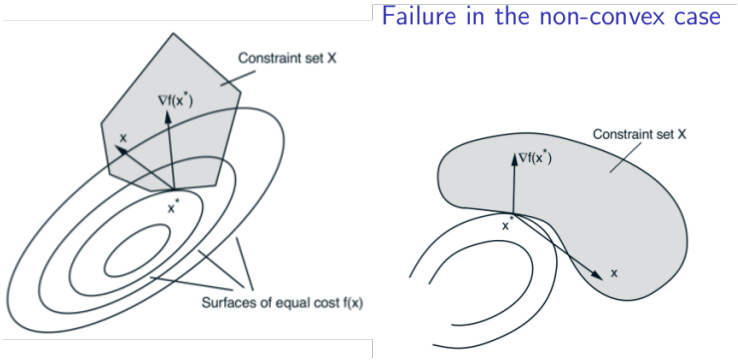

Figure 6: Differences between convex and non-convex sets

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?*

Let $z \in \mathbb{R}^n$. The projection of z onto X is defined by

$$\mathrm{proj}_X(z) = \min_x \|x - z\| = \min_x \frac{1}{2}\|x - z\|^2 \quad \text{s.t. } x \in X$$

It is the point in X which is closest to z.

Since $f(x) = \frac{1}{2}\|x - z\|^2$, we have $\nabla f(x) = x - z$. By the (neccessary and sufficient) optimality condition we get that

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

$$\Leftrightarrow \quad (z-x)^{\mathrm{T}}(x-x^*) \leq 0.$$

If X is a linear subspace of $\mathbb{R}^n$, this reduces to

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

since for all $x \in X$, both $x^* + x$ and $x^* - x$ are in X.

A linear subspace $Y \subseteq X$ is a subset of X which is also a vector space, i.e., it is an abelian group with respect to addition and behaves ring-like with scalar multiplication. In particular, it contains the zero vector.

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 $\bar{x}$.*

Given a vector x, a feasible direction at x is a vector d such that $x + \alpha d$ is feasible for all sufficiently small $\alpha > 0$; i.e., $x + \alpha d \in X$ (see Figure 7).

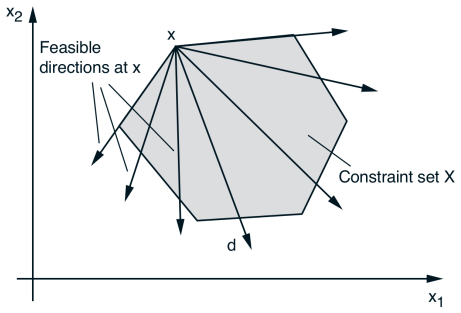

Figure 7: Illustration of feasible directions at x for a convex constraint set.

A *feasible direction method* starts with a feasible vector, and generates a sequence of feasible points $\{x^{(k)}\}$ by

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

where $d^{(k)}$ are feasible directions at $x^{(k)}$ which are also descent directions, and the $\alpha^{(k)}$ are step sizes satisfying $x^{(k)} + \alpha^{(k)} d^{(k)} \in X$.

Since X is assumed to be convex, the feasible directions at $x^{(k)}$ can be written in the form

$$d^{(k)} = \gamma(\bar{x}^{(k)} - x^{(k)}), \quad \gamma > 0,$$

where $\bar{x}^{(k)}$ is some feasible vector different from $x^{(k)}$. Assuming that $x^{(k)} + \alpha^{(k)} d^{(k)} \in X$, the steps can be expressed as

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

with $\alpha^{(k)} \in (0, 1]$, $\bar{x}^{(k)} \in X$, and $\nabla f(x^{(k)})^{\mathrm{T}}(\bar{x}^{(k)} - x^{(k)}) < 0$.

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.*

The conditional gradient method generates the point $\bar{x}^{(k)}$ by finding the descent direction furthest away from $x^{(k)}$ in the constraint set:

$$\bar{x}^{(k)} \in \operatorname*{argmin}_{x \in X} \nabla f(x^{(k)})^{\mathrm{T}}(x - x^{(k)}),$$
$$x^{(k+1)} = x^{(k)} + \alpha^{(k)}(\bar{x}^{(k)} - x^{(k)})$$

for a step size $\alpha^{(k)} \in (0, 1]$. Finding $\bar{x}^{(k)}$ is then a linear subproblem over a convex set, which should be easier to solve (especially in the case that X is a simplex).

The gradient projection method generates a feasible direction method with

$$\bar{x}^{(k)} = \operatorname{proj}_X(x^{(k)} - s^{(k)} \nabla f(x^{(k)})),$$
$$x^{(k+1)} = x^{(k)} + \alpha^{(k)}(\bar{x}^{(k)} - x^{(k)}),$$

for a step size $\alpha^{(k)} \in (0, 1]$ and positive scalars $s^{(k)}$. Thus, we take a step in the negative gradient direction (as in steepest descent), which is then projected onto X, to obtain the feasible vector $\bar{x}^{(k)}$. This method needs to solve a quadratic subproblem, which may be more complex, but typically results in better convergence rate.

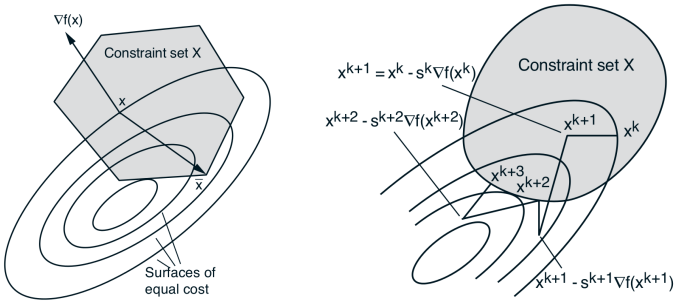

Figure 8: Left: illustration of finding the feasible direction in at point x in the conditional gradient method. $\bar{x}$ is the furthest point in X that lies in a descent direction. Right: illustration of the gradient projection method for $\alpha^{(k)} = 1$. Inside X, projection is trivial and we only do a normal gradient step.

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?*

The affine scaling method is based on the scaled gradient projection method, by which we generate the steps using

$$\bar{x}^{(k)} = \operatorname*{argmin}_{x \in X} \nabla f(x^{(k)})^{\mathrm{T}}(x - x^{(k)}) + \frac{1}{s^{(k)}}(x - x^{(k)})^{\mathrm{T}} H^{(k)}(x - x^{(k)}).$$

Since the problem is linear, we can calculate the exact solution of this quadratic intermediate problem, and get an update of

$$x^{(k+1)} = -x^{(k)} - \alpha^{(k)}(H^{(k)})^{-1}(I - A^{\mathrm{T}}(A(H^{(k)})^{-1}A^{\mathrm{T}})^{-1}A(H^{(k)})^{-1})c.$$

Now $\alpha^{(k)}$ has to be chosen such that $x^{(k+1)} > 0$. This can be done by using e.g. $\alpha^{(k)} = 0.99\bar{\alpha}^{(k)}$, where $\bar{\alpha}^{(k)}$ is the largest step size for which $x^{(k+1)} \geq 0$ (i.e., it is still feasible). Furthermore, $x^{(0)}$ needs to be (strictly) feasible. $H^{(k)}$ is for different reasons chosen as

$$H^{(k)} = \left(\mathrm{diag}(x_1^{(k)}, \ldots, x_n^{(k)})\right)^{-2}.$$

Since $x^{(k)} > 0$, if $s^{(k)}$ is small enough, $\bar{x}^{(k)} > 0$, and we can drop the inequality constraint $x \geq 0$.

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.*

Consider the general equality constrained problem

$$\min_x f(x), \quad \text{s.t. } h_i(x) = 0,\ i = 1, \ldots, m.$$

If x^* is a local minimum, and $\nabla h_1(x^*), \ldots, \nabla h_i(x^*)$ are linearly independent, then there exist unique scalars $\lambda_1^*, \ldots, \lambda_m^*$, called Lagrange multipliers, such that

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) = 0.$$

This means that at the optimal value, the gradient of the cost function belongs to the subspace spanned by the gradients of the constraint functions. Equivalently, $\nabla f(x^*)$ is orthogonal to the space of feasible variations at x^*, which is the the set of directions d such that $h(x^* + d) = 0$ (i.e., the constraint is still fulfilled). See Figure 9 for an illustration of this condition.

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

$$\min_x \frac{1}{2}\|x - y\|^2, \quad \textit{s.t. } Ax = 0.$$

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

We assume $x, y \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$. The Lagrangian is

$$\begin{aligned} L(x, \lambda) &= \frac{1}{2}\|x - y\|^2 + \sum_i \lambda_i \sum_j A_{ij} x_j \\ &= \frac{1}{2}\|x - y\|^2 + \lambda^{\mathrm{T}} A x, \end{aligned}$$

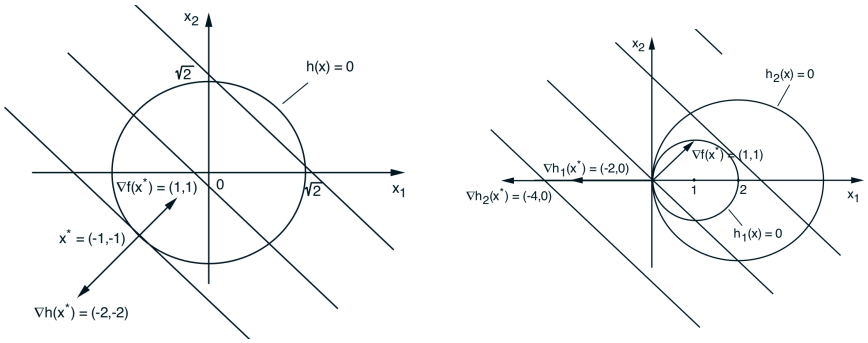

Figure 9: Left: illustration of Lagrange theorem for $f(x) = x_1 + x_2$ subject to $h(x) = x_1^2 + x_2^2 = 0$. Right: failure in the non-regular case for $h_1(x) = (x_1 - 1)^2 + x_2^2 - 1 = 0$, $h_2(x) = (x_1 - 2)^2 + x_2^2 - 4 = 0$, where the constraint gradients at x^* are collinear.

for $\lambda \in \mathbb{R}^m$. Its gradients are

$$\begin{aligned}\nabla_x L(x, \lambda) &= x - y + A^{\mathrm{T}}\lambda, \\ \nabla_\lambda L(x, \lambda) &= Ax.\end{aligned}$$

Since this is an equality constrained problem, the KKT conditions consist only of the stationarity condition (there is no μ):

$$\begin{aligned}x^* &= y - A^{\mathrm{T}}\lambda, \\ Ax^* &= 0.\end{aligned}$$

Replacing x^* in the second line gives us

$$\begin{aligned}A(y - A^{\mathrm{T}}\lambda) &= 0 \\ Ay &= AA^{\mathrm{T}}\lambda \\ \lambda &= (AA^{\mathrm{T}})^{-1}Ay,\end{aligned}$$

which we can put back to get the solution

$$x^* = y - A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}Ay.$$

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

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

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

We assume $x, y \in \mathbb{R}^n$, $a \in \mathbb{R}^n$, $b \in \mathbb{R}$. The Lagrangian is

$$L(x, \mu) = \frac{1}{2}\|x - y\|^2 + \mu(a^{\mathrm{T}}x - b)$$

for $\mu \in \mathbb{R}$. It has derivatives

$$\nabla_x L(x, \mu) = x - y + \mu a,$$
$$\nabla_\mu L(x, \mu) = a^{\mathrm{T}} x - b.$$

The stationarity condition $\nabla_x L(x^*, \mu) = 0$ gives us

$$x^* = y - \mu a.$$

Now, since this is an inequality constrained problem, we have to distinguish the case of the constraint being active or not. If it is inactive (i.e., $a^{\mathrm{T}} x^* - b < 0$), the KKT conditions require $\mu = 0$, which results in the solution $x^* = y$. This is the case where y is already in the half space, so the projection is trivial.

If the constraint is active, we know that $a^{\mathrm{T}} x^* = b$. In this case, the KKT conditions require $\mu \geq 0$. We insert the expression for x^* into the constraint:

$$a^{\mathrm{T}}(y - \mu a) = b$$
$$a^{\mathrm{T}} y - \mu a^{\mathrm{T}} a = b$$
$$\frac{a^{\mathrm{T}} y - b}{a^{\mathrm{T}} a} = \mu.$$

Putting μ back into the stationarity equation gives the projection

$$x^* = y - \frac{a^{\mathrm{T}} y - b}{a^{\mathrm{T}} a} a.$$

This is the case where an "outside" y is projected onto the closest point on the plane defined by a and b.