MATH-329 Nonlinear optimization
Homework 4: Duality, penalty methods

Instructor: Nicolas Boumal
TAs: Quentin Rebjock, Radu Dragomir

Due: January 3, 2022, by 23h59

Typeset your answers (preferably using LaTeX) and submit a PDF on Moodle before the deadline. Late homework is not accepted. Where applicable, provide runnable code in separate files *and* in your report (see Moodle for how to include Matlab code in LaTeX). Justify all your answers. Writing quality and clarity as well as mathematical rigor affect grading substantially. Conciseness permitting, demonstrate you understand what is going on.

Heads up: in the theory questions below, as always, be very clear in your proofs. Point out where you use any assumptions and be explicit in the potential theorems or other facts you invoke.

Let A, B be two given symmetric matrices of size n. We want to solve the following nonlinearly constrained optimization problem:

$$\min_{x,y\in\mathbb{R}^n} f(x,y) \quad \text{subject to} \quad h(x,y) = 0, \tag{P}$$

where $f\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is defined by

$$f(x,y) = \frac{1}{2}x^\top Ax + \frac{1}{2}y^\top By$$

and $h\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^3$ is defined by

$$h(x,y) = \begin{bmatrix} 1 - x^\top x \\ 1 - y^\top y \\ x^\top y \end{bmatrix}.$$

1. Is the feasible set of (P) convex? (Justify.) Show that LICQ holds at all feasible points.

2. We call (P) the *target problem*. If we remove the constraint $x^\top y = 0$, we get a new problem called the *relaxed problem*. What is the optimal value of the relaxed problem? (You may reference results/exercises/examples in the lecture notes to support a short justification.) How do the optimal values of the target and relaxed problems compare? (Explain briefly.)

3. Work out an expression for the Lagrangian function $L(x, y, \mu) = f(x, y) + \mu^\top h(x, y)$ in the following format, where "…" are quantities that may not depend on x, y:

$$L(x, y, \mu) = \frac{1}{2} \begin{bmatrix} x^\top & y^\top \end{bmatrix} \begin{bmatrix} \cdots & \cdots \\ \cdots & \cdots \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \cdots . \tag{1}$$

You should arrange for the matrix to be symmetric.

4. Obtain an expression for the dual function $L_D(\mu) = \inf_{x,y \in \mathbb{R}^n} L(x, y, \mu)$. Make sure it explicitly shows a necessary and sufficient condition for $L_D(\mu)$ to be finite: it should be a condition you could check on a computer given A, B, μ. Use this to write down the dual problem of (P).

5. What is the optimal value of the dual problem? Comment in light of Q2 and weak duality. Hint: if a symmetric block-matrix is positive semidefinite, what can you say about its diagonal blocks?

6. The strong duality theorem in the lecture notes has some assumptions. Does (P) satisfy them? Argue explicitly for each one if it does, does not, or if we would need more information to tell. (If the latter, what additional piece of information do you need, and why can't we expect to have it before solving the problem?)

7. The augmented Lagrangian function for (P) with penalty weight $\beta > 0$ is

$$L_\beta(x, y, \mu) = f(x, y) + \mu^\top h(x, y) + \frac{\beta}{2} \|h(x, y)\|^2. \tag{2}$$

Give expressions for its gradients with respect to x and y, denoted by $\nabla_x L_\beta(x, y, \mu)$ and $\nabla_y L_\beta(x, y, \mu)$. No need to show your steps. Make these expressions nice enough to implement.

Let's make things concrete. From now on, we let $n = 5$ and the matrices A, B are given by:

```
A = [ 2     4     4     5     9
      4     4     8     7     5
      4     8     2     8     5
      5     7     8     6     5
      9     5     5     5     2 ];
B = [ 4     8     3     6     8
      8     4     4     5     2
      3     4     2     8     5
      6     5     8     4     7
      8     2     5     7     8 ];
```

8. Write code for a function which takes as input x, y, μ, β and returns $L_\beta(x, y, \mu)$, $\nabla_x L_\beta(x, y, \mu)$ and $\nabla_y L_\beta(x, y, \mu)$. You have lots of freedom in how to do this. For example, you may find it more practical to let $z = [x^\top, y^\top]^\top \in \mathbb{R}^{2n}$, so that your code takes as input z, μ, β and returns $L_\beta(z, \mu)$ and $\nabla_z L_\beta(z, \mu)$. You are strongly encouraged to check your gradient numerically (as we have done in previous weeks). You do not have to report on your tests, but if the gradient is wrong, you will have a hard time completing the rest of the homework.

9. Write code which takes as input β, μ and an initial guess for x, y, and which makes an honest attempt to minimize $L_\beta(x, y, \mu)$ with respect to x, y (with μ fixed). You may reuse code you have written for exercise sessions or homeworks (e.g., gradient descent, trust-region method...). You may even use Matlab's `fminunc` (requires Optimization Toolbox) or an equivalent in Python. Run your code with $\beta = 1.42$, $\mu = [1, 2, -3]^\top$ and initial guess $x = [1, 0, -1, 2, 1]^\top$, $y = [1, 2, 0, 1, 2]^\top$. Report the outputs x, y and the value $L_\beta(x, y, \mu)$.

10. The quadratic penalty function with penalty weight β is $F_\beta(x, y) = L_\beta(x, y, 0)$, so by the previous question you already have code to compute F_β and its gradient, and even to (try to) minimize it. Implement the quadratic penalty method: initialize x, y randomly, and run 9 iterations with β taking the values $1, 2, 4, 8, \ldots, 256$. Don't forget to warm-start. Report the following at each iteration of the quadratic penalty method: $x, y, f(x, y), h(x, y), \beta h(x, y)$. Comment briefly on the behavior, especially for $\|h(x, y)\|$ and $\beta h(x, y)$, in light of what we learned in class / lecture notes.

11. Do the same as above, but this time with the augmented Lagrangian method: initialize x, y randomly and $\mu = 0$, and run 9 iterations with β taking the values $1, 2, 4, 8, \ldots, 256$. Don't forget to warm-start. Report the following at each iteration of ALM: $x, y, f(x, y), h(x, y), \mu$. Comment briefly on the behavior, also comparing with the previous question.

12. Using ALM, you should have figured out Lagrange multipliers μ at what we hope is a global minimizer of (P). With your computed value of μ, is the Lagrangian $(x, y) \mapsto L(x, y, \mu)$ convex in (x, y)? Comment briefly in light of the strong duality theorem and comparing $f(x, y)$ to the optimal value of the dual problem: be nuanced.