Exam Continuous Optimization

24 January 2022, 13.30–16.30

This closed-book exam consists of 5 questions. Please start each question on a new page, write legibly, and hand in your work with the solutions in the correct order. Good luck!

1. (5 points) Let
$$f(x_1, x_2) = x_1^2 - x_2^2.$$
Compute the Newton direction $\Delta x_{\rm nt}$ of f at $(1, 2)$ and show this is not a descent direction.

Solution: We have
$$\nabla f(1,2) = \begin{pmatrix} 2 \\ -4 \end{pmatrix}$$
and
$$\nabla^2 f(1,2) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}.$$
So
$$\Delta x_{\rm nt} = -\nabla^2 f(1,2)^{-1} \nabla f(1,2) = -\begin{pmatrix} 0.5 & 0 \\ 0 & -0.5 \end{pmatrix} \begin{pmatrix} 2 \\ -4 \end{pmatrix} = \begin{pmatrix} -1 \\ -2 \end{pmatrix}$$
and
$$\nabla f(1,2)^{\mathsf{T}} \Delta x_{\rm nt} = \begin{pmatrix} 2 \\ -4 \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} -1 \\ -2 \end{pmatrix} = 6 \geq 0,$$
so $\Delta x_{\rm nt}$ is not a descent direction.

2. (10 points) Consider the equality constrained least squares problem
$$\begin{array}{ll} \text{minimize} & \|Ax - b\|_2^2 \\ \text{subject to} & Cx = d, \end{array}$$
Explain how (and why) we can use the KKT optimality conditions to solve this problem by solving a single linear system.

Solution: The objective function can be written as
$$x^{\mathsf{T}} A^{\mathsf{T}} A x - 2b^{\mathsf{T}} A x + b^{\mathsf{T}} b,$$
and hence is convex by the second-order condition for convexity (the Hessian $2A^{\mathsf{T}}A$ is positive semidefinite). It follows that Slater's condition holds if and only if the problem is feasible.

Since the functions defining the problem are differentiable, if Slater's condition holds, then a vector x is optimal if and only if there exists a vector ν such that the KKT conditions hold:

- $Cx = d$
- $2A^\mathsf{T}Ax - 2A^\mathsf{T}b + C^\mathsf{T}\nu = 0$

We can write this as the linear system

$$\begin{pmatrix} 2A^\mathsf{T}A & C^\mathsf{T} \\ C & 0 \end{pmatrix} \begin{pmatrix} x \\ \nu \end{pmatrix} = \begin{pmatrix} 2A^\mathsf{T}b \\ d \end{pmatrix}.$$

If this system has a solution, then Slater's condition holds and the KKT conditions hold, hence x is optimal. If this system does not have a solution, then the problem is infeasible.

3. (10 points) Derive the Lagrangian, Lagrange dual function, and the Lagrange dual problem of the following optimization problem in x and y:

$$\begin{aligned} \text{minimize} \quad & -\sum_{i=1}^{m} \log(y_i) \\ \text{subject to} \quad & y = b - Ax \end{aligned}$$

Solution: The Lagrangian is

$$L(x, y, \nu) = -\sum_{i=1}^{m} \log(y_i) + \nu^\mathsf{T}(y + Ax - b).$$

The dual function is

$$g(\nu) = \inf_{x,y} \left(-\sum_{i=1}^{m} \log(y_i) + \nu^\mathsf{T}(y + Ax - b) \right)$$

The terms in x are unbounded below if $A^\mathsf{T}\nu \neq 0$ and the terms in y are unbounded below unless $\nu \succ 0$. If $\nu \succ 0$, then the optimal y is given by $y_i = 1/\nu_i$. So the dual function is

$$g(\nu) = \begin{cases} \sum_{i=1}^{m} \log(\nu_i) + m - b^\mathsf{T}\nu & A^\mathsf{T}\nu = 0,\ \nu \succ 0 \\ -\infty & \text{otherwise.} \end{cases}$$

So the dual problem is

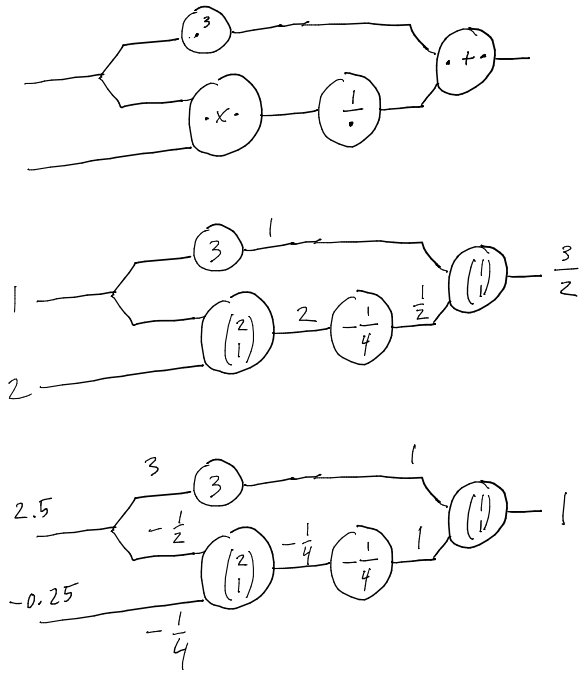
$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{m} \log(\nu_i) + m - b^\mathsf{T}\nu \\ \text{subject to} \quad & A^\mathsf{T}\nu = 0, \\ & \nu \succ 0. \end{aligned}$$

4. (a) (7 points) Consider the function

$$F(x, y) = x^3 + \frac{1}{xy},$$

where we view addition, multiplication, taking the reciprocal, and taking the third power as elementary functions. Show how $\nabla F(1, 2)$ is computed using reverse-mode automatic differentiation by drawing the appropriate diagrams.

Solution:



So,

$$\nabla F(1,2) = \begin{pmatrix} 2.5 \\ -0.25 \end{pmatrix}$$

(b) (3 points) Give the cost function for training a neural network using supervised learning, and explain why reverse-mode automatic differentiation is used here.

Solution: The cost function is

$$C(W^{(1)}, \ldots, W^{(L)}, b^{(1)}, \ldots, b^{(L)}) = \frac{1}{2N} \sum_{i=1}^{N} \|f(x_i) - y_i)\|_2^2,$$

where $(x_1, y_1), \ldots, (x_N, y_N)$ is the training set, and where $f(x_i)$ represents the evaluation at x_i of the neural network with weights $W^{(1)}, \ldots, W^{(L)}$ and biases $b^{(1)}, \ldots, b^{(L)}$. Note

that in practice we sum over a random batch of training samples as opposed to all training samples.
The cost function is a highly nested function of the weights and biases, and it maps a high dimensional space into a one dimensional space. For (stochastic) gradient descent we need to compute many gradients of this function. For this reverse-mode (since the input dimensional is much larger than the output dimension) automatic differentiation is much faster than symbolic differentiation.

5. Consider the barrier method for an optimization problem of the form

$$\begin{aligned} &\text{minimize} && f_0(x) \\ &\text{subject to} && f_i(x) \leq 0, \quad i = 1, \ldots, m, \end{aligned}$$

where the functions $f_0, \ldots, f_m$ are convex and twice continuously differentiable. Assume the problem has an optimal solution x^* with objective p^*. Assume furthermore the problem is strictly feasible.

(a) (7 points) Show that if $x^*(t)$ is optimal for the centering problem with parameter t, then

$$f_0(x^*(t)) - p^* \leq \frac{m}{t}.$$

Solution: The centering problem is

$$\text{minimize} \quad tf_0(x) - \sum_{i=1}^{m} \log(-f_i(x)).$$

The stationarity condition for the centering problem is

$$t\nabla f_0(x) + \sum_{i=1}^{m} \frac{1}{-f_i(x)} \nabla f_i(x) = 0.$$

Define

$$\lambda_i = \frac{1}{-tf_i(x^*(t))}.$$

Then $\lambda \geq 0$, so

$$\begin{aligned} p^* &= f_0(x^*) \\ &\geq g(\lambda) \\ &= \inf_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \right) \\ &= \frac{1}{t} \inf_x \left(tf_0(x) + \sum_{i=1}^{m} \frac{1}{-f_i(x^*(t))} f_i(x) \right) \\ &= \frac{1}{t} \left(tf_0(x^*(t)) - m \right) \\ &= f_0(x^*(t) - \frac{m}{t}. \end{aligned}$$

Alternatively, one can give a proof completely on the primal side using the first-order conditions for convexity for the functions $f_0, \ldots, f_m$.

(b) (3 points) Suppose $m = 1000$ and we apply the barrier method with parameters $\mu = 2$, $\epsilon = 1$, and with 1 as the initial value for t. After approximately how many outer iterations does the barrier method terminate?

Solution: At each outer iteration we multiple t by μ, so after approximately 10 iterations we get $m/t < \epsilon$.

End of test