

1. Consider the classification problem **one** versus **seven** with the MNIST dataset from the previous homework. Use the same settings as for the stochastic gradient descent, i.e., the quadratic dividing hypersurface

$$x^\top Wx + v^\top x + b,$$

the quadratic test function

$$q(x_j; \mathbf{w}) := y_j \left(x^\top Wx + v^\top x + b \right),$$

and the loss function

$$f(\mathbf{w}) = \frac{1}{n} \sum_{j=1}^n \log \left(1 + e^{-q(x_j; \mathbf{w})} \right) + \frac{\lambda}{2} \|\mathbf{w}\|^2.$$

Here \mathbf{w} denotes the $d^2 + d + 1$ -dimensional vector of coefficients of $\{W, v, b\}$.

Implement

- (a) Deterministic and Stochastic Nesterov (experiment with various batch sizes). Its deterministic version is given by Eqs. (61)-(62) in `Optimization.pdf`;
- (b) Deterministic and Stochastic Adam (experiment with various batch sizes). Its deterministic version is proposed in a paper by D. P. Kingma and J. L. Ba “Adam: A Method for Stochastic Optimization” where ADAM is introduced: <https://arxiv.org/pdf/1412.6980.pdf>.

You can use a constant step size. Run the stochastic optimizers for the same number of epochs (if you have n data points and your batch size is m then $\text{round}(n/m)$ timesteps is one epoch). Compare the performance of these optimizers with each other and with the stochastic gradient descent from the previous homework. Write a report containing graphs of the objective function and the norm of its gradient versus the iteration number. Which stochastic optimizer do you find the most efficient? Which batch size do you recommend? Which step size do you recommend?

Solution: TODO

2. Consider the KKT system

$$\begin{bmatrix} G & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} -\mathbf{p} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{g} \\ 0 \end{bmatrix}$$

where G is $d \times d$ symmetric positive definite and A is $m \times d$ and has linearly independent rows. Show that the matrix

$$K := \begin{bmatrix} G & A^\top \\ A & 0 \end{bmatrix}$$

is of *saddle-point type*, i.e., it has d positive eigenvalues and m negative ones. *Hint: Find matrices X and S (S is called the **Schur complement**) such that*

$$\begin{bmatrix} G & A^\top \\ A & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ X & I \end{bmatrix} \begin{bmatrix} G & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} I & X^\top \\ 0 & I \end{bmatrix}.$$

Then use Sylvester’s Law of Inertia (look it up!) to finish the proof.

Solution: TODO

3. Consider an equality-constrained quadratic program QP

$$\begin{aligned} \min \quad & \frac{1}{2} \mathbf{x}^\top G \mathbf{x} + \mathbf{c}^\top \mathbf{x} \\ \text{subject to} \quad & A \mathbf{x} = \mathbf{b}. \end{aligned}$$

The matrix G is symmetric. Assume that A is full rank (i.e., its rows are linearly independent) and $Z^\top G Z$ is positive definite where Z is a basis for the null-space of A , i.e., $AZ = 0$.

- (a) Write the KKT system for this case in the matrix form.

Solution: TODO

- (b) Show that the matrix of this system K is invertible. *Hint: assume that there is a vector $\mathbf{z} := (\mathbf{x}, \mathbf{y})^\top$ such that $K\mathbf{z} = 0$. Consider the quadratic form $\mathbf{z}^\top K \mathbf{z}$, use logical reasoning and algebra, and arrive at the conclusion that then $\mathbf{z} = 0$.*

Solution: TODO

- (c) Conclude that there exists a unique vector $(\mathbf{x}^*, \boldsymbol{\lambda}^*)^\top$ that solves the KKT system. Note that since we have only equality constraints, the positivity of $\boldsymbol{\lambda}$ is irrelevant.

Solution: TODO

4. Consider the following quadratic program with inequality constraints:

$$\begin{aligned} \min \quad & f(x, y) = (x - 1)^2 + (y - 2.5)^2 \\ \text{subject to} \quad & x - 2y + 2, -x - 2y + 6, -x + 2y + 2, x, y \geq 0 \end{aligned}$$

- (a) Plot the level sets of the objective function and the feasible set.

Solution: TODO

- (b) What is the exact solution? Find it analytically with the help of your figure.

Solution: TODO

- (c) Suppose the initial point is $(2, 0)$. Initially, constraints 3 and 5 are active, hence start with $\mathcal{W} = \{3, 5\}$. Work out all iterations of the active-set method analytically. The arising linear systems should be very easy to solve. For each iteration, you need to write out the set \mathcal{W} , the KKT system, its solution, i.e., (p_x, p_y) , the vector of Lagrange multipliers, and the current iterate (x_k, y_k) . Plot all iterates on your figure. There should be a total of 5 iterations.

| |
|-----------------------|
| Solution: TODO |
|-----------------------|