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

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

the quadratic test function

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

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 $\mathsf{round}(n/m)$ 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

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

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: Starting with the ansatz suggested by the hint, note that

$$\begin{bmatrix} I & 0 \\ X & I \end{bmatrix} \begin{bmatrix} G & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} I & X^\top \\ 0 & I \end{bmatrix} = \begin{bmatrix} G & 0 \\ XG & S \end{bmatrix} \begin{bmatrix} I & X^\top \\ 0 & I \end{bmatrix} = \begin{bmatrix} G & GX^\top \\ XG & XGX^\top + S \end{bmatrix}$$

hence to achieve equality with K in the bottom left block, let $X = AG^{-1}$. It remains to achieve equality in the bottom right block; note that

$$XGX^\top + S = AG^{-1}G(AG^{-1})^\top + S = AG^{-1}A^\top + S$$

and hence we take $S = -AG^{-1}A^\top$ to achieve equality with K. Hence, we have written

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

Since K is symmetric (because G is symmetric), by Sylvester's law of inertia, we have that K has the same inertia as the symmetric matrix

$$\begin{bmatrix} G & 0 \\ 0 & S \end{bmatrix}.$$

We now claim that the eigenvalues of the above matrix are exactly the eigenvalues of G and S combined. This follows from the multiplicativity of the determinant, since,

$$\begin{aligned} \det\left(\begin{bmatrix} G - \lambda I_d & 0 \\ 0 & S - \lambda I_m \end{bmatrix}\right) &= \det\left(\begin{bmatrix} G - \lambda I_d & 0 \\ 0 & I_m \end{bmatrix} \begin{bmatrix} I_d & 0 \\ 0 & S - \lambda I_m \end{bmatrix}\right) \\ &= \det\left(\begin{bmatrix} G - \lambda I_d & 0 \\ 0 & I_m \end{bmatrix}\right) \det\left(\begin{bmatrix} I_d & 0 \\ 0 & S - \lambda I_m \end{bmatrix}\right) = \det\left(G - \lambda I_d\right) \det\left(S - \lambda I_m\right). \end{aligned}$$

Hence, the characteristic polynomial of the above matrix is the product of the characteristic polynomials of G and S, so the eigenvalues of the above matrix are exactly the eigenvalues of G and S combined.

It remains to show that all m eigenvalues of S are negative. Toward this end, consider

$$x^\top Sx = -x^\top AG^{-1}A^\top x = -(A^\top x)^\top G^{-1}(A^\top x).$$

Then since G^{-1} is symmetric positive definite,

$$x^\top Sx = -(A^\top x)^\top G^{-1}(A^\top x) \leq 0$$

with equality if and only if $A^\top x = 0$, hence S is negative definite, meaning all eigenvalues of S are negative. Hence, referring back to Sylvester's law of inertia, K has d positive eigenvalues and m negative ones.

3. Consider an equality-constrained quadratic program QP

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

The matrix G is symmetric. Assume that A is full rank (i.e., its rows are linearly independent) and $Z^\top GZ$ 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: From the above, we define the Lagrangian

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \frac{1}{2}\mathbf{x}^\top G\mathbf{x} + \mathbf{c}^\top \mathbf{x} - \boldsymbol{\lambda}^\top (A\mathbf{x} - \mathbf{b})$$

in which case

$$\nabla_{\mathbf{x}} \mathcal{L} = G\mathbf{x} + \mathbf{c} - A^\top \boldsymbol{\lambda}.$$

Taking $\nabla_{\mathbf{x}} \mathcal{L} = 0$, we obtain the system

$$\begin{aligned} G\mathbf{x} + \mathbf{c} - A^\top \boldsymbol{\lambda} &= 0 \\ A\mathbf{x} - \mathbf{b} &= 0 \end{aligned}$$

which we can rewrite as

$$\begin{aligned} G(-\mathbf{x}) + A^\top \boldsymbol{\lambda} &= \mathbf{c} \\ A(-\mathbf{x}) &= -\mathbf{b} \end{aligned}$$

or equivalently,

$$\begin{bmatrix} G & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} -\mathbf{x} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{c} \\ -\mathbf{b} \end{bmatrix}.$$

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

Solution: From (a), recall that

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

and so for arbitrary $\mathbf{z} \coloneqq (\mathbf{x}, \mathbf{y})^\top$ it follows that

$$\begin{aligned} \mathbf{z}^\top K \mathbf{z} = \begin{bmatrix} \mathbf{x}^\top & \mathbf{y}^\top \end{bmatrix} \begin{bmatrix} G & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} &= \mathbf{x}^\top (G\mathbf{x} + A^\top \mathbf{y}) + \mathbf{y}^\top A\mathbf{x} \\ &= \mathbf{x}^\top G\mathbf{x} + \mathbf{x}^\top A^\top \mathbf{y} + \mathbf{y}^\top A\mathbf{x}. \end{aligned}$$

Further, since $\mathbf{x}^\top A^\top \mathbf{y}$ is a scalar,

$$\mathbf{x}^\top A^\top \mathbf{y} = (\mathbf{x}^\top A^\top \mathbf{y})^\top = \mathbf{y}^\top A\mathbf{x}$$

and so

$$\mathbf{z}^\top K \mathbf{z} = \mathbf{x}^\top G\mathbf{x} + 2\mathbf{y}^\top A\mathbf{x}. \tag{1}$$

Now assume that $\mathbf{z}$ is such that $K\mathbf{z} = 0$. Then

$$0 = K\mathbf{z} = \begin{bmatrix} G & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} G\mathbf{x} + A^\top \mathbf{y} \\ A\mathbf{x} \end{bmatrix},$$

so we have both that $G\mathbf{x} + A^\top \mathbf{y} = 0$ and $A\mathbf{x} = 0$. Further, $\mathbf{z}^\top K\mathbf{z} = \mathbf{z}^\top(0) = 0$. Substituting these facts into (1), we find that

$$0 = \mathbf{z}^\top K\mathbf{z} = \mathbf{x}^\top G\mathbf{x} + 2\mathbf{y}^\top A\mathbf{x} = \mathbf{x}^\top G\mathbf{x} + 2\mathbf{y}^\top(0) = \mathbf{x}^\top G\mathbf{x}$$

and so $\mathbf{x}^\top G\mathbf{x} = 0$.

Now let Z be a basis for the null-space of A. In particular, since $A\mathbf{x} = 0$, we can write $\mathbf{x} = Z\hat{\mathbf{x}}$, a linear combination of the columns of Z. Then

$$0 = \mathbf{x}^\top G\mathbf{x} = (Z\hat{\mathbf{x}})^\top G(Z\hat{\mathbf{x}}) = \hat{\mathbf{x}}^\top (Z^\top GZ)\hat{\mathbf{x}}.$$

Hence, since $\hat{\mathbf{x}}^\top (Z^\top GZ)\hat{\mathbf{x}} = 0$ and $Z^\top GZ$ is assumed to be positive definite, it must be that $\hat{\mathbf{x}} = 0$, and likewise, $\mathbf{x} = Z\hat{\mathbf{x}} = 0$.

It remains to show that $\mathbf{y} = 0$ also. This follows from

$$0 = G\mathbf{x} + A^\top \mathbf{y} = A^\top \mathbf{y}$$

and the fact that A has linearly independent rows, so $A^\top$ has linearly independent columns, meaning $\mathbf{y} = 0$ since $A^\top \mathbf{y} = 0$.

Hence, $K\mathbf{z} = 0$ has only the trivial solution $\mathbf{z} = 0$, meaning K is invertible.

(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: Since K is invertible, we can simply compute

$$K^{-1} \begin{bmatrix} \mathbf{c} \\ -\mathbf{b} \end{bmatrix}$$

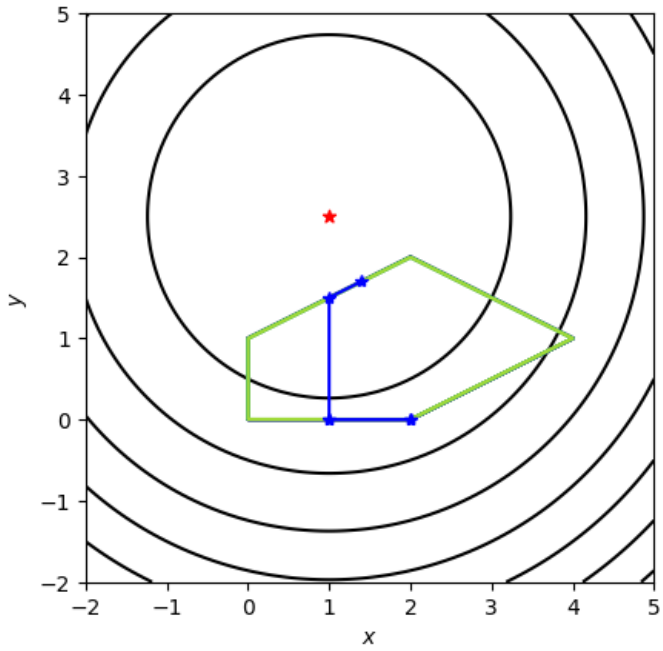
which is a concatenation of $-\mathbf{x}^*$ and $\boldsymbol{\lambda}^*$.

4. Consider the following quadratic program with inequality constraints:

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

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

Solution: We plot the level sets of the objective function and the boundary of the feasible set below. The corresponding code can be found here.



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

Solution: From the above figure, we see that solution lies on the line $x - 2y + 2 = 0$. Solving for x, we find $x = 2y - 2$, and so we seek to find

$$\min f(2y-2, y) = \min(2y-3)^2 + (y-2.5)^2$$

which occurs where

$$0 = \frac{d}{dy}\left[(2y-3)^2 + (y-2.5)^2\right] = 4(2y-3) + 2(y-2.5) = 10y - 17,$$

so $y^* = 17/10$ and hence $x^* = 14/10$.

(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: We first define $\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$ so that we may write

$$f(x,y) = (x-1)^2 + (y-2.5)^2 = x^2 + y^2 - 2x - 5y + 29/4$$
$$= \mathbf{x}^\top \mathbf{x} + \begin{bmatrix} -2 \\ -5 \end{bmatrix}^\top \mathbf{x} + 29/4.$$

For convenience, we will define a new function $\hat{f}$ with the same minimizer as f, obtained by ignoring the constant $29/4$ and scaling by $1/2$, that is,

$$\hat{f}(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{x} + \begin{bmatrix} -1 \\ -5/2 \end{bmatrix}^\top \mathbf{x}.$$

We will now derive the KKT system for each subproblem of the active-set method with arbitrary $\mathcal{W}$. Note that

$$\hat{f}(\mathbf{x}_k + \mathbf{p}) = (\mathbf{x}_k + \mathbf{p})^\top (\mathbf{x}_k + \mathbf{p}) + \begin{bmatrix} -1 \\ -5/2 \end{bmatrix}^\top (\mathbf{x}_k + \mathbf{p})$$
$$= \mathbf{p}^\top \mathbf{p} + \left(\mathbf{x}_k + \begin{bmatrix} -1 \\ -5/2 \end{bmatrix} \right)^\top \mathbf{p} + \hat{f}(\mathbf{x}_k) = \mathbf{p}^\top \mathbf{p} + \nabla \hat{f}_k^\top \mathbf{p} + \hat{f}(\mathbf{x}_k),$$

from which we obtain the subproblem

$$\begin{aligned} &\min g(\mathbf{p}) \coloneqq \mathbf{p}^\top \mathbf{p} + \nabla \hat{f}_k^\top \mathbf{p} \\ &\text{subject to } A_{\mathcal{W}} \mathbf{p} = 0. \end{aligned}$$

Towards this end, we define the Lagrangian

$$\mathcal{L}(\mathbf{p}, \boldsymbol{\lambda}) \coloneqq g(\mathbf{p}) - \boldsymbol{\lambda}^\top A_{\mathcal{W}} \mathbf{p} = \mathbf{p}^\top \mathbf{p} + \nabla \hat{f}_k^\top \mathbf{p} - \boldsymbol{\lambda}^\top A_{\mathcal{W}} \mathbf{p}$$

in which case

$$\nabla_{\mathbf{p}} \mathcal{L} = \mathbf{p} + \nabla \hat{f}_k - A_{\mathcal{W}}^\top \boldsymbol{\lambda}.$$

Hence, setting this gradient equal to zero yields the KKT system

$$\begin{aligned} -\mathbf{p} + A_{\mathcal{W}}^\top \boldsymbol{\lambda} &= \nabla \hat{f}_k \\ -A_{\mathcal{W}} \mathbf{p} &= 0 \end{aligned}$$

which we rewrite in matrix form as

$$\begin{bmatrix} I & A_{\mathcal{W}}^\top \\ A_{\mathcal{W}} & 0 \end{bmatrix} \begin{bmatrix} -\mathbf{p} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \nabla \hat{f}_k \\ 0 \end{bmatrix}.$$

We now begin the iterations starting with initial point $\mathbf{x}_0 = \begin{bmatrix} 2 & 0 \end{bmatrix}^\top$, so that $\mathcal{W} = \{3, 5\}$ and the KKT system is

$$\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 2 & 1 \\ -1 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} -p_x \\ -p_y \\ \lambda_3 \\ \lambda_5 \end{bmatrix} = \begin{bmatrix} 1 \\ -5/2 \\ 0 \\ 0 \end{bmatrix}.$$

We find that this system has the solution $(p_x, p_y) = (0, 0)$ and $(\lambda_3, \lambda_5) = (-1, -1/2)$. Therefore, we remove 3 from $\mathcal{W}$ and keep $\mathbf{x}_1 = \mathbf{x}_0 = \begin{bmatrix} 2 & 0 \end{bmatrix}^\top$. We now begin the second iteration with $\mathcal{W} = \{5\}$, yielding the KKT system

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} -p_x \\ -p_y \\ \lambda_5 \end{bmatrix} = \begin{bmatrix} 1 \\ -5/2 \\ 0 \end{bmatrix}$$

We find that this system has the solution $(p_x, p_y) = (-1, 0)$ and $\lambda_5 = -5/2$. We then compute the step size

$$\alpha = \min \left\{ 1, \min_{i \notin \mathcal{W}, \mathbf{a}_i^\top \mathbf{p} < 0} \frac{\mathbf{b}_i - \mathbf{a}_i^\top \mathbf{x}_1}{\mathbf{a}_i^\top \mathbf{p}} \right\} = \min\{1, \min\{2, 4, 2\}\} = 1,$$

and hence $\mathbf{x}_2 = \mathbf{x}_1 + \alpha \mathbf{p} = \begin{bmatrix} 1 & 0 \end{bmatrix}^\top$. At the start of the third iteration, we still have $\mathcal{W} = \{5\}$, so our KKT system is

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} -p_x \\ -p_y \\ \lambda_5 \end{bmatrix} = \begin{bmatrix} 0 \\ -5/2 \\ 0 \end{bmatrix}$$

which has the solution $(p_x, p_y) = (0, 0)$ and $\lambda_5 = -5/2$, so we remove 5 from $\mathcal{W}$ and take $\mathbf{x}_3 = \mathbf{x}_2 = \begin{bmatrix} 1 & 0 \end{bmatrix}^\top$. Now at the start of the fourth iteration, we have $\mathcal{W} = \{\}$, yielding the KKT system

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -p_x \\ -p_y \end{bmatrix} = \begin{bmatrix} 0 \\ -5/2 \end{bmatrix}$$

with solution $(p_x, p_y) = (0, 5/2)$. Then computing the step size,

$$\alpha = \min\{1, \min\{1, 3/5\}\} = 3/5,$$

and hence $\mathbf{x}_4 = \mathbf{x}_3 + \alpha \mathbf{p} = \begin{bmatrix} 1 & 3/2 \end{bmatrix}^\top$. We also add 1 to $\mathcal{W}$. At long last we begin the fifth iteration with $\mathcal{W} = \{1\}$. This yields the KKT system

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -2 \\ 1 & -2 & 0 \end{bmatrix} \begin{bmatrix} -p_x \\ -p_y \\ \lambda_1 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix}$$

with solution $(p_x, p_y) = (2/5, 1/5)$ and $\lambda_1 = 2/5$. Since all Lagrange multipliers are positive, we are done. Computing the final step size,

$$\alpha = \min\{1, \min\{2\}\} = 1$$

we find that $\mathbf{x}_5 = \mathbf{x}_4 + \alpha \mathbf{p} = \begin{bmatrix} 7/5 & 17/10 \end{bmatrix}^\top$.