Homework 4

Convex Optimization 10-725/36-725

Due Friday November 2 at 11:59pm
Submit your work as a single PDF on Gradescope, including the source code.
Make sure to prepare your solution to each problem on a separate page.
On Gradescope, please select source code along with the corresponding problem.
Please choose either Q1 or Q2 (Score = max(Q1, Q2) + Q3 + Q4).

Total: 61 points

1 Nuclear norm, duality, and matrix completion (17 pts) [Po-wei & Akash]

In this problem, we will take a look at convex optimization problems involving *matrix* variables. We touched on these a little bit in class, and while they might seem somewhat abstract, these sorts of problems are very much connected to real-world applications like Netflix's movie recommendation engine (you are welcome to ask us for further details here).

Let $X \in \mathbb{R}^{m \times n}$ be a matrix. The *trace norm* (also known as the *nuclear norm*) of a matrix X, which we write as $\|X\|_{\mathrm{tr}}$, can be defined as the sum of the singular values of X.

(a, 5pts) Show that computing the trace norm of a matrix, i.e., computing $\|X\|_{\mathrm{tr}}$, can be expressed as the following (convex) optimization problem:

$$\begin{aligned} \underset{Y \in \mathbb{R}^{m \times n}}{\text{maximize}} \quad & \mathrm{tr}(X^T Y) \\ \text{subject to} \quad & \begin{bmatrix} I_m & Y \\ Y^T & I_n \end{bmatrix} \succeq 0, \end{aligned} \tag{1}$$

where I_p is the $p \times p$ identity matrix. (By the way, problem (5) is a semidefinite program; more on this in part (d) below.)

Hint: recall that the trace norm is the dual of the operator norm; also, think about using the "Schur complement" identity here. A good reference for this is Section A.5.5 in the "Convex Optimization" book, by Stephen Boyd and Lieven Vandenberghe.

> Proof:
> We know that the dual norm of the operator norm is trace norm
> $\|X\|_{tr} = sup\{tr(X^T Y) | \|Y\|_{op} \le 1\}$ (`https://en.wikipedia.org/wiki/Dual_norm`)
>
> We can look at the problem again, by "Schur complement", we know that
> $I_n - Y^T I_m^{-1} Y \succeq 0$, i.e, $Y^T Y \preceq I_n$, i.e., the largest singular value of Y is less than 1
> And we also know that $\|Y\|_{op}$ equals the largest singular value of Y(why?),
> therefore, the constraint is the same as $\|Y\|_{op} \le 1$
> so we know computing $\|X\|_{tr}$ can be expressed as the optimization problem

(b, 5pts) Show that the dual problem associated with (5) can be expressed as

$$
\begin{aligned}
\underset{\substack{W_1 \in \mathbb{S}^m, \\ W_2 \in \mathbb{S}^n}}{\text{minimize}} \quad & \operatorname{tr}(W_1) + \operatorname{tr}(W_2) \\
\text{subject to} \quad & \begin{bmatrix} W_1 & (1/2)X \\ (1/2)X^T & W_2 \end{bmatrix} \succeq 0,
\end{aligned}
\tag{2}
$$

where, just to remind you, $\mathbb{S}^p$ is the space of $p \times p$ real, symmetric matrices.

Hint: even though this is a matrix-variate problem, you can just go through the same steps as if it were a vector-variate problem (form the Lagrangian, maximize it over the primal variables, etc.) Recall that a succint way to write the inner product between matrices A, B is $\operatorname{tr}(A^T B)$ (this is just the usual inner product, once we unravel them into long vectors).

> Proof:
> Let $Z = \begin{bmatrix} W_1 & W_3 \\ W_4 & W_2 \end{bmatrix}$, $A = \begin{bmatrix} I_m & Y \\ Y^T & I_n \end{bmatrix}$, the lagranian can be expressed as:
> $L(Y, Z) = \operatorname{tr}(X^T) + <Z, A>$, where $Z \in \mathbb{S}^{m+n}$ is psd
> $L(Y, Z) = \operatorname{tr}(X^T) + <Z, A> \geq f(Y)$, therefore, $\max L(Y, Z) \geq f^*$
>
> $L(Y, Z) = \operatorname{tr}(X^T) + \operatorname{tr}(ZA) = \operatorname{tr}(X^T) + \operatorname{tr}(W_1) + \operatorname{tr}(W_2) + \operatorname{tr}(W_3 Y^T) + \operatorname{tr}(W_4 Y) = <$
> $X, Y> + \operatorname{tr}(W_1) + \operatorname{tr}(W_2) + <W_3, Y> + <W_4^T, Y>$
>
> take derivative against Y of $L(Y, Z)$, we get
>
> $X + W_3 + W_4^T = 0$, since Z is symmetric, $W_3 = W_4^T$, i.e., $W_3 = -\frac{1}{2}X$
> $\min_Z \max_Y L(Y, Z) = \operatorname{tr}(W_1) + \operatorname{tr}(W_2)$
> So, the dual problem is:
>
> $$
> \begin{aligned}
> \underset{\substack{W_1 \in \mathbb{S}^m, \\ W_2 \in \mathbb{S}^n}}{\text{minimize}} \quad & \operatorname{tr}(W_1) + \operatorname{tr}(W_2) \\
> \text{subject to} \quad & \begin{bmatrix} W_1 & (1/2)X \\ (1/2)X^T & W_2 \end{bmatrix} \succeq 0,
> \end{aligned}
> \tag{3}
> $$
>
> Note:
> We can show that $\begin{bmatrix} W_1 & (1/2)X \\ (1/2)X^T & W_2 \end{bmatrix} \succeq 0 \iff \begin{bmatrix} W_1 & -(1/2)X \\ -(1/2)X^T & W_2 \end{bmatrix} \succeq 0$
> consider $v = \begin{bmatrix} v_1 & v_2 \end{bmatrix}$, first statement is saying that, $\forall v_1, v_2 \neq 0$,
> $v^T \begin{bmatrix} W_1 & (1/2)X \\ (1/2)X^T & W_2 \end{bmatrix} v \geq 0$, therefore,
> $v_1^T W_1 v_1 + \frac{1}{2} v_2^T X^T v_1 + \frac{1}{2} v_1^T X v_2 + v_2^T W_2 v_2 = v_1^T W_1 v_1 + v_2^T W_2 v_2 + v_2^T X^T v_1 \geq 0$
>
> Now, let $v_1^* = -v_1$, we would have, $\forall v_1^*, v_2 \neq 0$
> $v_1^{*T} W_1 v_1^* + v_2^T W_2 v_2 - v_2^T X^T v_1^* = v_1^T W_1 v_1 + v_2^T W_2 v_2 + v_2^T X^T v_1 \geq 0$, this is saying that,
>
> $$
> \begin{bmatrix} W_1 & -(1/2)X \\ -(1/2)X^T & W_2 \end{bmatrix} \succeq 0
> $$
>
> In the same logic, we can prove the other direction.

(c, 2pts) Show that the optimal values for problems (5) and (7) are equal to each other, and that both optimal values are attained.

> Proof:
> Need to show that at the solution, the constraint is strictly satisfied, i.e, positive definite

(d, 5pts) In the *matrix completion problem*, we want to find a matrix $X \in \mathbb{R}^{m \times n}$ of low rank that is close, in a squared error sense, to some observed matrix $Z \in \mathbb{R}^{m \times n}$. We do not assume that all of the entries of Z are observed, so we will look at the squared error over Z's observed entries only, which we store in a set Ω of (observed) row and column indices. Putting all this together leads us to the following (convex) optimization problem:

$$\underset{X \in \mathbb{R}^{m \times n}}{\text{minimize}} \quad \sum_{(i,j) \in \Omega} (X_{ij} - Z_{ij})^2 + \lambda \|X\|_{\text{tr}}, \tag{4}$$

with tuning parameter $\lambda > 0$.

Show that problem (4) can be expressed as a semidefinite program (SDP) of the form

$$\begin{aligned} \underset{x \in \mathbb{R}^p}{\text{minimize}} \quad & c^T x \\ \text{subject to} \quad & x_1 A_1 + \cdots + x_p A_p \preceq B, \end{aligned}$$

for some fixed $c, B, A_i, \ i = 1, \ldots, p$.

Hint: you may start by considering the following constrained problem:

$$\begin{aligned} \underset{X \in \mathbb{R}^{m \times n}}{\text{minimize}} \quad & \|X\|_{\text{tr}} \\ \text{subject to} \quad & \sum_{(i,j) \in \Omega} (X_{ij} - Z_{ij})^2 \leq s, \end{aligned}$$

which is equivalent to our original penalized problem, for some $s > 0$; now you will probably need to use each of the above parts (in different ways) here to recast the above in SDP form.

> Proof:
> we already know that $\|X\|_{tr} =$
>
> $$\begin{aligned} \underset{Y \in \mathbb{R}^{m \times n}}{\text{maximize}} \quad & \text{tr}(X^T Y) \\ \text{subject to} \quad & \begin{bmatrix} I_m & Y \\ Y^T & I_n \end{bmatrix} \succeq 0, \end{aligned} \tag{5}$$
>
> $=$
>
> $$\begin{aligned} \underset{\substack{W_1 \in \mathbb{S}^m, \\ W_2 \in \mathbb{S}^n}}{\text{minimize}} \quad & \text{tr}(W_1) + \text{tr}(W_2) \\ \text{subject to} \quad & \begin{bmatrix} W_1 & (1/2)X \\ (1/2)X^T & W_2 \end{bmatrix} \succeq 0, \end{aligned} \tag{6}$$
>
> Therefore, minimize $\|X\|_{tr}$ can be expressed as:
>
> $$\begin{aligned} \underset{\substack{W_1 \in \mathbb{S}^m, \\ W_2 \in \mathbb{S}^n, \\ X \in \mathbb{R}^{m \times n}}}{\text{minimize}} \quad & \text{tr}(W_1) + \text{tr}(W_2) \\ \text{subject to} \quad & \begin{bmatrix} W_1 & (1/2)X \\ (1/2)X^T & W_2 \end{bmatrix} \succeq 0 \end{aligned} \tag{7}$$
>
> $$\begin{bmatrix} sI & a \\ a^T & s \end{bmatrix} \succeq 0$$
>
> where $a_k = X_{ij} - Z_{ij}$, where $(i,j) \in \Omega$
> This is apparently a SDP

2 Sparse eigenvectors via the barrier method (17 pts) [Po-Wei]

In this problem we consider finding a sparse "eigenvector" of a matrix $A \succeq 0$.

(a, 1pt) Consider solving the optimization problem

$$\max_{x \in \mathbb{R}^n} \quad x^T A x$$
$$\text{subject to} \quad \|x\|_2 = 1$$
$$\|x\|_1 \leq C.$$

Is this problem convex? Why or why not?

(b, 1pt) Consider the matrix-variate problem

$$\max_{X \in \mathbb{S}^n} \quad \text{tr} A X$$
$$\text{subject to} \quad X \succeq 0$$
$$\text{tr} X = 1,$$
$$\text{rank}(X) = 1,$$
$$\sum_{i,j} |X_{ij}| \leq C^2.$$

Is this problem convex? Why or why not?

(c, 4pts) Prove that the problems in parts (a) and (b) are equivalent. Hint: if X is rank 1, then we can write it as $X = xx^T$ for some vector x.

(d, 2pts) Now, consider the matrix-variate problem

$$\max_{X \in \mathbb{S}^n} \quad \text{tr} A X$$
$$\text{subject to} \quad X \succeq 0$$
$$\text{tr} X = 1$$
$$\sum_{i,j} |X_{ij}| \leq C^2.$$

Is this problem convex? What is the relationship between this problem and that from part (b)?

(e, 4pts) Form the logarithmic barrier for the constraint $\sum_{i,j} |X_{ij}| \leq C^2$ and use this along with the logarithmic barrier for $X \succeq 0$ ($\phi(X) = \log \det X$) to modify the problem in part (d) to have a smooth objective and equality constraint.

(f, 5pts) Setup and describe the iterations of the barrier method for the problem in part (d), explicitly deriving the Newton update for $X^{(k)}$.

3   Second Order Methods for Logistic Regression (22 pts) [Akash]

In this problem, we will try to classify a viewer's age group from his movie ratings. We formulate the problem as a binary classification with output label $y \in \{0, 1\}^n$, corresponding to whether a person's age is under 40, and input features $X \in \mathbb{R}^{n \times (p+1)}$. Similarly to previous homework problems, the first column of X is taken to be 1_n to account for the intercept term. We will apply second order methods to solve the logistic regression problem.

We model each $y_i | x_i$ with the probabilistic model

$$\log \left(\frac{p_\beta(y_i = 1 | x_i)}{1 - p_\beta(y = 1 | x_i)} \right) = \log \left(\frac{\mu_i}{1 - \mu_i} \right) = (X\beta)_i,$$

for $i = 1, \ldots, n$, $\beta \in \mathbb{R}^{p+1}$. We use β_0 to denote the intercept and $\beta_{1:p}$ for the rest of the weights.

(a) (2 pts) First, show that the negative (conditional) log likelihood (NLL) under the logistic probability model can be expressed as:

$$l(\beta) = - \sum_{i=1}^n \log(p_\beta(y_i | x_i)) = - \sum_{i=1}^n y_i (X\beta)_i + \sum_{i=1}^n \log(1 + \exp\{(X\beta)_i\}),$$

Hint: model $y_i | x_i$ as a bernoulli random variable with parameter $\mu_i = p_\beta(y_i = 1 | x_i)$.

(b) In this part of the problem, we will implement Newton's method for the nonpenalized logistic regression problem and see that in this context, Newton's method is known as Iteratively Reweighted Least Squares (IRLS).

(i, 1pt) Express the gradient and the Hessian of the NLL in terms of X, y, and μ.

> Proof:
> $\nabla l(\beta) = X^T(-y + \mu)$,
> $\nabla^2 l(\beta) = X^T diag(\mu) diag(1 - \mu) X$

(ii, 1pt) Write the Newton update for β using the above gradient and Hessian, using t to denote step-size.

> $\beta^{(k)} = \beta^{(k-1)} - t(X^T diag(\mu^{(k)}) diag(1 - \mu^{(k)}) X)^{-1} X^T (-y + \mu^{(k)})$

(iii, 1pts) Show that the Newton update has the form of a weighted least-squares estimation problem. This is why in this context, Newton's method is known as IRLS.

> $\beta^{(k)} = \beta^{(k-1)} + t(X^T S^{(k)} X)^{-1} X^T (y - \mu^{(k)}) = t(X^T S^{(k)} X)^{-1} [\frac{1}{t} X^T S^{(k)} X \beta^{(k-1)} + X^T (y - \mu_k)] = (X^T S^{(k)} X)^{-1} X^T (S^{(k)} X \beta^{(k-1)} + y - \mu^{(k)})$

(iv, 1pts) Given X and y, write out the steps for performing IRLS to estimate $\hat{\beta}$. Assume a fixed step size t in the steps.

(v, 3pts) Now, implement IRLS with backtracking using the movie data set on the website (in `hwk4_gs3.zip`, for this homework, not the second one). Initialize your weights with zeros. Use $\alpha = 0.01$ and shrink parameter $\beta = 0.9$ for backtracking. You can stop IRLS when the

change between consecutive objective values is less than 1e-6. Report both the train and test errors of classifying whether a person is under 40. Plot $f^{(k)} - f^\star$ versus k, where $f^{(k)}$ denotes the objective value at outer iterations k of IRLS, and the optimal objective value is $f^\star = 186.637$ on a semi-log scale (i.e. where the y-axis is in log scale).

(c) Now, we introduce regularization to the NLL to improve our test error. The problem becomes:

$$\min_{\beta \in \mathbb{R}^{p+1}} \quad l(\beta) + \lambda \|\beta_{1:p}\|_1 \tag{8}$$

(i, 2pts) Rewrite (8) as a problem that has a smooth criterion, aims to preserve the sparsity of the original problem, and can be solved by an interior point method. (Hint: consider introducing a new variable and corresponding inequality constraints).

$$\min_{\beta \in \mathbb{R}^{p+1}} \quad l(\beta) + \lambda \Sigma_{i=1}^{p} s_i$$
$$\text{subject to} \quad - s_i \leq \beta_i \leq s_i, \forall i = 1, ..., p$$

(ii, 3pts) Describe the iterations of the barrier method for the problem in (i), explicitly deriving the Newton updates.

$$\min_{\beta, s} t(l(\beta) + \lambda \Sigma_{i=1}^{p} s_i) - \Sigma_{i=1}^{p} log(\beta_i + s_i) - \Sigma_{i=1}^{p} log(s_i - \beta_i)$$

At $t = t^{(0)}$, we run Newton to find $x^{(0)} = x^*(t)$

(iii, 4pts) Implement the barrier method described above. For the barrier parameters t(the multiplier for the original criterion in (8)) and μ(the constant by which t increases at each outer iteration of the barrier method), use $\mu = 20$ and start with $t = 5$. A good number for m(the constant that, together with t, bounds the duality gap) is the number of constraints in the barrier problem. For backtracking during the Newton method steps, use $\alpha = 0.2$ and $\beta = 0.9$. You can use 1e-9 as the stopping threshold for both the Newton method and the barrier method. Remember to initialize the centering Newton method step with a strictly feasible point (hint: in MATLAB, one simple way to find such a point is with *linprog*).

Use the barrier method with $\lambda = 15$ to classify whether a viewer is under 40. Report the train and test classification errors. Report the number of zeros at the solution $\beta^\star$. Here we consider any number with absolute value under 1e-10 to be zero.

(iv, 4 pts) Now, implement proximal gradient descent with backtracking for logistic regression with lasso penalty. You can refer back to the group lasso problem in homework 2 to help you in your implementation.

Using $\lambda = 15$, run proximal gradient descent (with backtracking line search, with shrinkage parameter $\beta = 0.5$) on the training data.

To compare the convergence of both the barrier method and the proximal gradient descent, plot $f^{(k)} - f^\star$ versus k, where $f^{(k)}$ denotes the objective value at total number of iterations k of the algorithms, and the optimal objective value is $f^\star = 306.476$ on a semi-log scale (i.e. where the y-axis is in log scale). Plot these in the same figure.

To be clear, when counting total number of iterations, each backtracking iteration counts as 1 iteration (for prox gradient descent), and each inner Newton iteration counts as 1 iteration (for log-barrier method).

Attach all relevant code to the end of this problem.

4 Interior Point Methods for SVMs [Wenbo] (22 pts)

Overview   In this question we will continue playing with the same kernel support vector machine as we did in last homework. Recall that in homework 3 we have developed the dual optimization problem for support vector machine with radial basis function (RBF) kernel, and solved it using standard QP solver. This time, we will write our own solver, and solve the kernel-SVM using the barrier method and the primal-dual interior point method.

Following homework 3, we will be using the same dataset with inputs $X \in \mathbb{R}^{863 \times 2}$ and labels $y \in \{-1, 1\}^{863}$ from `data.txt` on the class website.

Kernel SVM   Recall from homework 3, the primal optimization of kernel SVM is

$$
\begin{aligned}
\underset{\beta, \xi_i}{\text{minimize}} \quad & \frac{1}{2} ||\beta||_2^2 + C \sum_{i=1}^{n} \xi_i \\
\text{subject to} \quad & y_i(\beta^T \phi(x_i) + \beta_0) \geq 1 - \xi_i \quad i = 1, \dots, n \\
& \xi_i \geq 0 \qquad\qquad\qquad\qquad\quad i = 1, \dots, n
\end{aligned}
$$

where $\xi_1, \dots, \xi_n$ are the slack variables. This is equivalent to solving the following dual optimization problem

$$
\begin{aligned}
\underset{w}{\text{maximize}} \quad & 1^T w - \frac{1}{2} w^T \tilde{K} w \\
\text{subject to} \quad & 0 \leq w \leq C1, \quad w^T y = 0
\end{aligned} \tag{9}
$$

where w is the dual variable, $K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle$, and $\tilde{K}_{ij} = y_i y_j K_{ij}$.

Again, recall that the support vectors correspond to the instances with the dual variable $w_i > 0$. The optimal primal vector β^* can be represented in terms of the dual optimal $w_i^*, i = 1, \dots, n$ as $\beta^* = \sum_{i=1}^{n} w_i^* y_i \phi(x_i)$. To solve for β_0^* from the dual, we can pick any instance j that lies on the margin, and compute β_0^* as

$$
\beta_0^* = y_j - (\beta^*)^T \phi(x_j) = y_j - \sum_{i=1}^{n} w_i^* y_i K_{ij}
$$

In practice, in order to reduce the variance, we can take the average as

$$
\beta_0^* = \frac{1}{|J|} \sum_{j \in J} \left(y_j - \sum_{i=1}^{n} w_i^* y_i K_{ij} \right) \tag{10}
$$

where J is the set of instances on the margin.

Lastly, the prediction for a given vector $x \in \mathbb{R}^d$ can be calculated as follows:

$$
\hat{y} = \text{sign} \left(\sum_{i=1}^{n} w_i^* y_i \langle \phi(x), \phi(x_i) \rangle + \beta_0^* \right) = \text{sign} \left(\sum_{i=1}^{n} w_i^* y_i K(x, x_i) + \beta_0^* \right) \tag{11}
$$

The RBF kernel is expresses as

$$
\langle \phi(x_i), \phi(x_j) \rangle = K_{ij} = \exp \left(-\gamma ||x_i - x_j||_2^2 \right)
$$

where we fix $\gamma = 500$.

Please submit your code as an appendix to this problem.

(a) **Barrier Method**

(i, 5pts) Derive the gradient and the Hessian of the dual criterion $f(w)$, and of the log barrier objective of the form $tf(w) + g(w)$, where $g(w)$ is the log barrier term which takes the role of the inequality constraint. (Note, the equality constraint remains unchanged in the log barrier problem!) Confirm that the Hessian only changes from that of the original version in the diagonal entries. (Hint: This may be helpful in understanding the update directions in the interior point implementation in part (b).)

(ii, Bonus question, 5pts) Implement the barrier method as described in class to solve the SVM dual. Note, you will be using an equality-constrained Newton method with backtracking, in the barrier method updates. Set $C = 0.1$. Using initial $t = 10000$ and $\mu = 4$, run your barrier method until $m/t < 10^{-8}$. For the inter Newton centering step, run until the Newton decrement $\frac{1}{2}\lambda(x)^2 < 10^{-16}$.

Several steps you should address:

1) Describe how to obtain an initial feasible point in our problem.

2) Describe how to obtain the equality-constrained Newton update direction.

3) Describe the stopping rule for in terms of the Newton decrement, using the equality-constrained Newton direction update. Note, your stopping criterion should be sufficiently small, for your inner Newton to have good precision.

4) Describe how to find an initial step size prior backtracking step. (Hint: the key is to choose a step size that would allow the upcoming update to be feasible; there are several ways to do this.)

5) For backtracking line search, you can use $\alpha = 0.01$ and $\beta = 0.5$, per notation in the lecture notes.

Plot the objective value versus iterations. Report the optimal objective value at the optimal training weights w^*: $f^* = \frac{1}{2}(w^*)^T \tilde{K} w^* - 1^T w^*$. Report the number of support vectors, considering only $w^* > 10^{-5}$. Use the optimal w^* to do prediction, report classification accuracy. A good check is to see if your answer agrees with (b) (or with standard QP solver).

(b) **Primal-Dual Interior Point Method**

(i, 2pts) For a given $t > 0$, list the perturbed KKT condition for this problem.

(ii, 2pts) Let $u_i \geq 0$ be the dual variable correspond to the inequality constraint $-w_i \leq 0, \forall i$ and $v_i \geq 0$ be the dual variable correspond to the inequality constraint $w_i \leq C, \forall i$. Let λ be the dual variable correspond to the equality constraint $y^T w = 0$. Write down the expression for the primal residual r_{prim}, the dual residual r_{dual} and the centrality residual r_{cent}.

(iii, 2pts) Define the residual vector $r(w, u, v, \lambda) = (r_{\text{dual}}, r_{\text{cent}}, r_{\text{prim}})$ and $z = (w, u, v, \lambda)$. The primal-dual interior point method is trying to find an update direction $\Delta z = (\Delta w, \Delta u, \Delta v, \Delta \lambda)$ such that $r(z + \Delta z) = 0$. However, in general $r(z + \Delta z) = 0$ corresponds to a system of nonlinear equations that does not admit closed-form solution, so instead we make a first-order approximation to $r(z + \Delta z)$ and solves the corresponding linear system: $r(z + \Delta z) \approx r(z) + Dr(z)\Delta z = 0$. Write down the linear system to be solved in order to get Δz. Δz is also known as the Newton direction.

(iv, 2pts) Primal-dual interior point method requires a strictly feasible point of (9) as a start point. Form a linear program that can be used to find such a strictly feasible initial point.

(v, 9pts) Implement the primal-dual interior point method to solve the minimization problem you formed in (i). We fix $C = 0.1$, $\mu = 1000$. You should stop your algorithm when both the following two conditions are met:

- $(||r_{\text{prim}}||_2^2 + ||r_{\text{dual}}||_2^2)^{1/2} \leq 10^{-6}$.

- The surrogate duality gap is less than or equal to 2×10^{-6}.

Here is a list of tips that might be helpful to you:

- Use the LP formed in the last question to find a strictly feasible initial point.
- After obtaining each Newton direction, it is important to compute a corresponding step size such that both the primal and the dual variables are feasible.
- For backtracking line search, you can use $\alpha = 0.01$ and $\beta = 0.5$ (Note this β corresponds to the parameter to shrink the step size, not the β in the primal SVM).
- It is important to tune to barrier parameter μ such that your algorithm stops within a fair amount of iterations. As a reference, the TA's implementation stops in 99 iterations.

Plot the objective value versus iterations. Report the function value at the optimal w^*: $f^* = \frac{1}{2}(w^*)^T \tilde{K} w^* - 1^T w^*$. Report the number of support vectors. Here we consider any w^* that is less than 10^{-5} to be 0. Use the optimal w^* to do prediction and report classification accuracy.

Please submit your code as an appendix to this problem.