ITESO Universidad Jesuita de Guadalajara
Convex Optimization

Exam

Prof.: Juan Diego Sánchez T.

April 6, 2022

1.		2.	
3.		4.	
5.		6.	
7.		8.	
		Total	

*Name:*_____

1. (20 pt) **Warm-up**: Let the following optimization exercises:

   (a) (5 pt) Consider the linear optimization problem (linear program)

   $$\begin{array}{ll} \min_{x\in\mathbb{R}^n} & c^T x \\ \text{s. t.} & Ax \leq b. \end{array} \tag{1}$$

   Prove that the Lagrangian of (1) is $L(x,u) = \left(c + A^T u\right)^T x - b^T u$ for $u \geq 0$. For this case, $b \in \mathbb{R}^m$; then, the dimensions of c and A follow from $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$. Show that the Lagrange dual function (The Wolfe's dual) of (1) is

   $$g(u) = \min_{x\in\mathbb{R}^n} L(x,u) = -b^T u$$

   and, then, write the complete dual optimization problem. This dual problem is also a linear program, but with m variables.

   (b) (5 pt) Consider the quadratic optimization problem

   $$\begin{array}{ll} \min_{x\in\mathbb{R}^n} & \frac{1}{2}x^T Q x + c^T x \\ \text{s. t.} & Ax = b, x \geq 0 \end{array} \tag{2}$$

   with $Q \succ 0$ and symmetric. Note that the Lagrangian of (2) is $L(x,u,v) = \frac{1}{2}x^T Q x + c^T x - u^T x + v^T(Ax - b)$ for $u \geq 0$ and any v. For this case, $b \in \mathbb{R}^m$; then, the dimensions of Q, c and A follow from $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$. Show that the Lagrange dual function (The Wolfe's dual) of (2) is

   $$g(u,v) = \min_{x\in\mathbb{R}^n} L(x,u,v) = -\frac{1}{2}\left(c - u + A^T v\right)^T Q^{-1}\left(c - u + A^T v\right) - b^T v$$

   and, then, write the complete dual optimization problem.

   (c) (5 pt) Consider the quadratic optimization problem

   $$\begin{array}{ll} \min_{x\in\mathbb{R}^n} & \frac{1}{2}x^T Q x + c^T x \\ \text{s. t.} & Ax = 0 \end{array} \tag{3}$$

   with $Q \succ 0$ and symmetric. Show that, by the KKT conditions, for this convex problem with no inequality constraints, x is a solution if satisfies

   $$\begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ v \end{bmatrix} = \begin{bmatrix} -c \\ 0 \end{bmatrix} \tag{4}$$

   for some v. That is, the system (4) provides the optimality conditions for (3).

(d) (5 pt) Consider the quadratic optimization problem

$$\begin{array}{ll} \min_{x\in\mathbb{R}^n} & \frac{1}{2}x^TQx + c^Tx \\ \text{s. t.} & Ax = b \end{array} \tag{5}$$

with $Q \succ 0$. Find the the optimality conditions for (5).

2. (10 pt) Consider the following optimization problem:

$$\begin{aligned} &\min_{w,b,e} \mathcal{P}(w,e) = \frac{1}{2}w^Tw + \gamma\frac{1}{2}\sum_{k=1}^{N} e_k^2 \\ &\text{s. t. } y_k = w^T\varphi\left(x_k\right) + b + e_k, \quad k = 1,\ldots,N. \end{aligned} \tag{6}$$

where $\{x_k, y_k\}_{k=1}^N$ represents a training set with input data $x_k \in \mathbb{R}^n$, the output data given $y_k \in \mathbb{R}$, $e_k \in \mathbb{R}^n$ are slack variables, and the feature maps have the form $\varphi(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$. Then, the model's parameters are $w \in \mathbb{R}^m$ and $b \in \mathbb{R}$. Finally, $\gamma > 0$.

Note that the problem (6) can be written as:

$$\min_{w,b,e} \mathcal{P}(w,e) = \frac{1}{2}w^Tw + \gamma\frac{1}{2}\sum_{k=1}^{N}\left[y_k - \left(w^T\varphi\left(x_k\right) + b\right)\right]^2.$$

Thus, the problem (6) is related to the so-called least squares support vector machines LS-SVM.

(a) Show that the Lagrangian of the problem (6) is given by:

$$\mathcal{L}(w,b,e;\alpha) = \frac{1}{2}w^Tw + \gamma\frac{1}{2}\sum_{k=1}^{N} e_k^2 - \sum_{k=1}^{N}\alpha_k\left\{w^T\varphi\left(x_k\right) + b + e_k - y_k\right\}$$

(b) Since the problem has not inequality constraints, the KKT optimality follows directly from the first-order conditions provided by the gradient of the Lagrangian $\mathcal{L}(w,b,e;\alpha)$. Then, show that:

- $\nabla_w\mathcal{L} = 0$ implies $w = \sum_{k=1}^N \alpha_k\varphi\left(x_k\right)$.
- $\frac{\partial\mathcal{L}}{\partial b} = 0$ implies $\sum_{k=1}^N \alpha_k = 0$.
- $\frac{\partial\mathcal{L}}{\partial e_k} = 0$ implies $\alpha_k = \gamma e_k$ for $k = 1,\ldots,N$.
- $\frac{\partial\mathcal{L}}{\partial \alpha_k} = 0$ implies $w^T\varphi\left(x_k\right) + b + e_k - y_k = 0$ for $k = 1,\ldots,N$.

(c) Define adequate vector variables, such that the optimization problem reduces to a set of linear equations which must be solved for α and b.

3. (10 pt) Consider the following optimization problem:

$$\begin{aligned} &\min_{w,b,e} \mathcal{P}(w,e) = \frac{1}{2}w^Tw + \gamma\frac{1}{2}\sum_{k=1}^{N} e_k^2 \\ &\text{s. t.} \quad y_k\left[w^T\varphi\left(x_k\right) + b\right] = 1 - e_k, \quad k = 1,\ldots,N. \end{aligned} \tag{7}$$

where $y_k \in \{-1, 1\}$ is the response (target) variable.

Then conduct an analysis of the problem (7) by applying the steps (a) to (d) of the problem (6) with, possibly, their respective modifications. In the step (d), explain how the new problem (7) is related to the classification problem. Finally, compare KKT matrix system obtained for this case with that of the problem (6).

4. (20 pt) Let $\{x_i, y_i\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}$ be the training data where $x_1, \ldots, x_N$ are deterministic points (fixed design) and $y_i = f(x_i) + e_i$ with $f : \mathbb{R}^d \to \mathbb{R}$ an unknown real-valued smooth function and $e_1, \ldots, e_N$ uncorrelated random errors with $E[e_i] = 0, E[e_i^2] = \sigma_e^2 < \infty$. The model for regression is given as $f(x) = w^T \varphi(x) + b$ where $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{n_h}$ denotes a potentially infinite ($n_h = \infty$) dimensional feature map.

   Then, consider the cost functions:

   - Tikhonov:

$$\min_{w,b,e_i} \mathcal{J}_T(w, e) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^N e_i^2 \text{ s.t. } w^T \varphi(x_i) + b + e_i = y_i, \quad \forall i = 1, \ldots, N$$

   - Morozov's discrepancy principle, where the minimal 2-norm of w realizing a fixed noise level σ^2 is to be found:

$$\min_{w,b,e_i} \mathcal{J}_M(w) = \frac{1}{2} w^T w \text{ s.t. } \begin{cases} w^T \varphi(x_i) + b + e_i = y_i, \forall i = 1, \ldots N \\ N\sigma^2 = \sum_{i=1}^N e_i^2 \end{cases}$$

   - Ivanov regularization amounts at solving for the best fit with a 2-norm on w smaller than π^2. The following modification is considered in this case

$$\min_{w,b,e_i} \mathcal{J}_I(e) = \frac{1}{2} e^T e \text{ s.t. } \begin{cases} w^T \varphi(x_i) + b + e_i = y_i, \forall i = 1, \ldots N \\ \pi^2 = w^T w \end{cases}$$

     The use of the equality (instead of the inequality) can be motivated in a kernel machine context as these problems are often ill-conditioned and result in solutions on the boundary of the trust region $w^T w \leq \pi^2$.

   Show that:

   (a) The conditions for optimality are

Condition	Tikhonov	Morozov	Ivanov
$\frac{\partial \mathcal{L}}{\partial w} = 0$	$w = \sum_{i=1}^N \alpha_i \varphi(x_i)$	$w = \sum_{i=1}^N \alpha_i \varphi(x_i)$	$2\xi w = \sum_{i=1}^N \alpha_i \varphi(x_i)$
$\frac{\partial \mathcal{L}}{\partial b} = 0$	$\sum_{i=1}^N \alpha_i = 0$	$\sum_{i=1}^N \alpha_i = 0$	$\sum_{i=1}^N \alpha_i = 0$
$\frac{\partial \mathcal{L}}{\partial e_i} = 0$	$\gamma e_i = \alpha_i$	$2\xi e_i = \alpha_i$	$e_i = \alpha_i$
$\frac{\partial \mathcal{L}}{\partial \alpha_i} = 0$	$w^T \varphi(x_i) + b + e_i = y_i,,$	$w^T \varphi(x_i) + b + e_i = y_i,,$	$w^T \varphi(x_i) + b + e_i = y_i$
$\frac{\partial \mathcal{L}}{\partial \xi} = 0$	–	$\sum_{i=1}^N e_i^2 = N\sigma^2$	$w^T w = \pi^2$

   for all $i = 1, \ldots, N$. The kernel-trick is applied as follows: $\varphi(x_k)^T \varphi(x_l) = K(x_k, x_l)$ for an appropriate kernel $K : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ in order to avoid explicit computations in the high dimensional feature space. Let $\Omega \in \mathbb{R}^{N \times N}$ be such that $\Omega_{kl} = K(x_k, x_l)$ for all $k, l = 1, \ldots, N$.

   (b) The Tikhonov conditions result in the following set of linear equations as classical:

$$\left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega + \frac{1}{\gamma} I_N \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right]$$

   (c) Re-organizing the sets of constraints of the Morozov and Ivanov scheme results in the following sets of linear equations where an extra nonlinear constraint relates the Lagrange multiplier ξ with the hyper-parameter σ^2 or π^2

   Morozov:

$$\left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega + \frac{1}{2\xi} I_N \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right], \text{ s.t. } N\sigma^2 = \alpha^T \alpha$$

(d) and,

Ivanov :

$$\left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \frac{1}{2\xi}\Omega + I_N \end{array}\right]\left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right], \text{ s.t. } \pi^2 = \alpha^T \Omega \alpha$$

.

5. (10 pt) Consider the following optimization problem:

$$\begin{aligned} \min_{w,b,\xi} \mathcal{P}(w,\xi) &= \frac{1}{2} w^T w + c \sum_{k=1}^{N} \xi_k \\ \text{s. t.} \quad & y_k \left[w^T \varphi(x_k) + b \right] \geq 1 - \xi_k, \quad k = 1, \ldots, N. \\ & \xi_k \geq 0, \quad k = 1, \ldots, N. \end{aligned} \tag{8}$$

where $y_k \in \{-1, 1\}$ is the response (target) variable, $\xi_k \in \mathbb{R}^n$ are slack variables, and $c > 0$. The remaining variables are defined as in (6).

(a) Explain (with math) how the optimization problem (8) is related to the binary classification problem. That is, deduce the optimization problem related to the linearly separable binary classification problem. After that, reformulate the optimization problem considering the possibility of separating the two sets with nonlinear functions. In addition, introduce slack variables (soft margin) to consider the possibility that some samples are misclassified.

(b) Show that the Lagrangian of the problem (8) is given by:

$$\mathcal{L}(w, b; \alpha) = \frac{1}{2} w^T w + c \sum_{k=1}^{N} \xi_k - \sum_{k=1}^{N} \alpha_k \left(y_k \left[w^T \varphi\left(x_k\right) + b \right] - 1 + \xi_k \right) - \sum_{k=1}^{N} \nu_k \xi_k.$$

(c) Calculate the dual cost function (The Wolfe Lagrangian) $\mathcal{D}(\alpha) = \min_{w,b} \mathcal{L}(w, b; \alpha)$.

(d) Show the dual problem in the Lagrange multipliers α_k (The Wolfe dual problem) as a quadratic programming formulation.

(e) For the problem (8), derive the KKT conditions:

- Stationary condition.
- Primal feasibility condition.
- Dual feasibility condition.
- Complementary slackness condition.

Then, compare the result obtained with the requirements of the KKT formulation with that given in the dual problem.

(f) Use the complementary slackness condition to prove the existence of some $\alpha_k = 0$. Then, confirm or refute the following conditions related to α_k:

$$\begin{aligned} \alpha_k = 0 &\Rightarrow y_k \left[w^T \varphi\left(x_k\right) + b \right] \geq 1 \\ \alpha_k = c &\Rightarrow y_k \left[w^T \varphi\left(x_k\right) + b \right] \leq 1 \\ 0 < \alpha_k < c &\Rightarrow y_k \left[w^T \varphi\left(x_k\right) + b \right] = 1 \end{aligned}$$

If those conditions hold true, rewrite the primal problem (8) and its Lagrangian function in terms of the so-called *Hinge loss*.

The property of having some $\alpha_k = 0$ is called *sparseness*.

(g) Present a rigorous method to find the value of b.

(h) Finally, provide the following refinements and explanations:

- Define $K(x_k, x_l) = \varphi(x_k)^T \varphi(x_l)$ for $k, l = 1, \ldots, N$ and write the dual problem in terms of $K(x_k, x_l)$.
- Show that the solution vector w has an expansion in terms of the training vectors x_k with $k = 1, \ldots, N$. Then, explain the effect of the sparseness on w.
- Those Lagrange multipliers such that $\alpha_k > 0$ are called *support values*, and those vectors x_k corresponding to the support values in the expansion for the solution vector w are called *support vectors*. Firstly, notice that, by definition, only the support values and vectors are relevant to provide a solution for the optimization problem (8).

6. (10 pt) Note that the problem (8) defines a support vector machine for solving binary classification problems. In this case, state appropriately (as a convex optimization problem) the analogous problem for the regression case. After that, present the solution to the proposed optimization problem using a solution which is analogous to the one used in the classification problem (8).

7. (10 pt) Consider the following optimization problem:

$$\begin{aligned} \min_{w,b,\xi} \mathcal{P}(w, \xi) &= \frac{1}{2} w^T w + c \left(\nu \varepsilon + \frac{1}{N} \sum_{k=1}^{N} (\xi_k + \xi_k^*) \right) \\ \text{s. t.} \quad & y_k - w^T \varphi(x_k) - b \leq \varepsilon + \xi_k, \quad k = 1, \ldots, N. \\ & w^T \varphi(x_k) + b - y_k \leq \varepsilon + \xi_k^*, \quad k = 1, \ldots, N. \\ & \xi_k, \xi_k^* \geq 0, \quad k = 1, \ldots, N. \end{aligned} \tag{9}$$

where $0 < \nu < 1$. The remaining variables are defined as usual.

For the problem (9), derive the KKT conditions:

- Stationary condition.
- Primal feasibility condition.
- Dual feasibility condition.
- Complementary slackness condition.

Use the previous conditions to present a complete and detailed development of the regression model related to the optimization problem (9).

8. (10 pt) Note that the problem (9) defines a support vector machine for solving regression problems. In this case, state appropriately (as a convex optimization problem) the analogous problem for the binary classification case. After that, present the solution to the proposed optimization problem using a solution which is analogous to the one used in the regression problem (9).