You can mix and choose any subproblems as long as you solve a total of at least 10 points. You are encouraged to solve problems beyond 10 points, but no extra credit will be given.

1 Basics of conjugate, and quasiconvexity (5 points)

1.1 Log-sum-exp function [1 point]

Derive the conjugate of the log-sum-exp function $f(x) = \log(\sum_{i=1}^n e^{x_i})$. This can be broken down to a few steps.

1. To calculate the conjugate, we need to maximize $y^\intercal x - f(x)$. Given log-sum-exp is convex, we can maximize by first order condition. Show that a vanishing gradient with respect to x implies the following.

$$y_i = \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}, \qquad\qquad i = 1, \ldots, n,$$

   Also show that this condition has solution if and only if $y \succ 0$ and $\mathbf{1}^\intercal y = 1$.

2. Substitute back into $y^\intercal x - f(x)$ to obtain that the conjugate of log-sum-exp is negative entropy.

$$f^*(y) = \sum_{i=1}^n y_i \log y_i, \qquad\qquad y \succeq 0, \mathbf{1}^\intercal y = 1.$$

   Here we slightly extended the domain to allow $y_i = 0$, and take the convention that $0 \log 0 = 0$.

3. Also show that the domain of f^* coincides with the domain above, i.e. $\sup_x y^\intercal x - f(x)$ is finite if and only if $\mathbf{1}^\intercal y = 1$ and $y \succeq 0$. (Hint: consider $x = -te_i$ and $x = t\mathbf{1}$.)

1.2 Properties of conjugates [2 points]

Solve at least two of the following problems. These are taken from Chapter 3 and Exercise 3.39 of [1].

1. (Affine transformation.) Check that, for $A \in \mathbf{R}^{n\times n}$ non-singular and $b \in \mathbf{R}^n$, the conjugate of $g(x) = f(Ax+b)$ is $g^*(y) = f^*(A^{-\intercal}y) - b^\intercal A^{-\intercal}y$, with $\mathbf{dom}\, g^* = A^\intercal \mathbf{dom}\, f^*$.

2. (Sum of independent functions.) Show that, for f_1 and f_2 convex, $f(u, v) = f_1(u) + f_2(v)$ has conjugate $f^*(w, z) = f_1^*(w) + f_2^*(z)$.

3. (Conjugate of convex plus affine function). Let $g(x) = f(x) + c^\intercal x + d$, with f convex. Express g^* in terms of f^* and c, d.

4. (Conjugate of perspective). Epxress the conjugate of the perspective of a convex function f in terms of f^*.

5. (Conjugate and minimization.) $f(x, z)$, convex in x, z), define $g(x) = \inf_z f(x, z)$. Express g^* in terms of f^*. As an application, express the conjugate of $g(x) = \inf_z\{h(z) \mid Az + b = x\}$, where h is convex, in terms of h^*, A and b.

1.3 Practice on conjugates (1 point)

Derive the conjugates of the following functions. Solve at least two. This is exercise 3.36 in [1].

1. Max function. $f(x) = \max_{i=1,\ldots,n} x_i$ on $\mathbf{R}^n$.

2. Sum of largest elements. $f(x) = \sum_{i=1}^r x_{[i]}$ on $\mathbf{R}^n$.

3. Piecewise-linear function on $\mathbf{R}$. $f(x) = \max_{i=1,\ldots,m}(a_i x + b_i)$ on $\mathbf{R}$. You can assume that the a_i are sorted in increasing order, i.e., $a_i \leq \cdots \leq a_m$, and that none of the functions $a_i x + b_i$ is redundant, i.e., for each k there is at least one x with $f(x) = a_k x + b_k$.
4. Power function. $f(x) = x^p$ on $\mathbf{R}_{++}$, where $p > 1$. Repeat for $p < 0$.
5. Negative geometric mean. $f(x) = -(\prod x_i)^{1/n}$ on $\mathbf{R}^n_{++}$.
6. Negative generalized logarithm for second-order cone. $f(x,t) = -\log(t^2 - x^\intercal x)$ on $\{(x,t) \in \mathbf{R}^n \times \mathbf{R} \mid \|x\|_2 < t\}$.

1.4 Quasiconvexity (1 point)

1. Recall that a function f is quasiconvex if and only if $\mathbf{dom}\, f$ is convex and $f(\theta x + (1-\theta)y) \leq \max\{f(x), f(y)\}$ for any $x, y \in \mathbf{dom}\, f$ and $\theta \in [0,1]$.

   Use this to show that the cardinality of a non-negative vector is quasiconcave, and the rank of a positive semidefinite matrix is quasiconcave.

2 Geometric interpretation of duality and the strong duality under Slater's condition (5 points)

Geometric interpretation of Lagrange duality helps with our intuitive understanding, and it provides a way to prove the strong duality under Slater's condition using separating hyperplanes. The following are adapted from Chapter 5.3 and Exercise 5.22 of [1].

2.1 Weak duality via set of values (1 point)

Consider the generic optimization problem

$$\begin{array}{ll} \min\limits_{x} & f_0(x) \\ \text{s.t.} & f_i(x) \leq 0 \quad i = 1, \ldots, m, \\ & h_j(x) = 0 \quad j = 1, \ldots, p \end{array} \tag{1}$$

with variable $x \in \mathbf{R}^n$, and f_0 is defined on domain $\mathcal{D}$.

A simple geometric interpretation of the dual function in terms of the set

$$\mathcal{G} = \{(f_1(x), \ldots, f_m(x), h_1(x), \ldots, h_p(x), f_0(x)) \in \mathbf{R}^m \times \mathbf{R}^p \times \mathbf{R} \mid x \in \mathcal{D}\},$$

which is the set of values taken on by the constraint and objective functions. The optimal value p^* can be expressed in terms of $\mathcal{G}$ as

$$p^* = \inf\{t \mid (u, v, t) \in \mathcal{G},\ u \preceq 0,\ \nu = 0\}.$$

To evaluate the dual function at (λ, ν), we minimize the affine function

$$(\lambda, \nu, 1)^\intercal (u, v, t) = \sum_{i=1}^{m} \lambda_i u_i + \sum_{i=1}^{p} \nu_i v_i + t$$

over $(u, v, t) \in \mathcal{G}$, i.e. we have

$$g(\lambda, \nu) = \inf\{(\lambda, \nu, 1)^\intercal (u, v, t) \mid (u, v, t) \in \mathcal{G}\}.$$

In particular, we see that if the infimum is finite, then the inequality

$$(\lambda, \nu, 1)^\intercal (u, v, t) \geq g(\lambda, \nu)$$

defines a supporting hyperplane to $\mathcal{G}$. This is sometimes referred to as a nonvertical supporting hyperplane, because the last component of the normal vector is nonzero.

Show that for feasible variables and values, i.e. $\lambda \succeq 0$, $u \preceq 0$, and $v = 0$, the weak duality $p^* \geq g(\lambda, \nu)$ holds.

2.2 Epigraph form of weak and strong duality (1 point)

To have a closer relation to strong duality, instead of the value set $\mathcal{G}$, we could also consider the epigraph of the optimization problem (1), defined as $\mathcal{A} \subset \mathbf{R}^m \times \mathbf{R}^p \times R$,

$$\mathcal{A} = \mathcal{G} + (\mathbf{R}^m_+ \times \{0\} \times \mathbf{R}_+).$$

More explicitly,

$$\mathcal{A} = \{(u, v, t) \mid \exists x \in \mathcal{D},\ f_i(x) \leq u_i,\ i = 1, \ldots, m,\ h_i(x) = v_i,\ i = 1, \ldots, p,\ f_0(x) \leq t\}.$$

Intuitively, $\mathcal{A}$ is like an epigraph of $\mathcal{G}$, such that it includes all the points in $\mathcal{G}$, as well as points that are "worse", i.e., those with larger objective or inequality constraint function values.

We can express the optimal value in terms of $\mathcal{A}$ as

$$p^* = \inf\{t \mid (0, 0, t) \in \mathcal{A}\}.$$

To evaluate the dual function at a point (λ, ν) with $\lambda \succeq 0$ (this is needed because inequalities in $\mathcal{A}$ requires the sign of λ to be correct), we can minimize the affine function $(\lambda, \nu, 1)^\intercal (u, v, t)$ over $\mathcal{A}$: if $\lambda \succeq 0$, then

$$g(\lambda, \nu) = \inf\{(\lambda, \nu, 1)^\intercal (u, v, t) \mid (u, v, t) \in \mathcal{A}\}.$$

If the infimum is finite, then

$$(\lambda, \nu, 1)^\intercal (u, v, t) \geq g(\lambda, \nu)$$

defines a nonvertical supporting hyperplane to $\mathcal{A}$.

Show that $p^* \geq g(\lambda, \nu)$, the weak duality lower bound. (Hint: $(0, 0, p^*)$ is in the boundary of $\mathcal{A}$.)

What is the condition for strong duality? Show that strong duality holds if and only if there exists a nonvertical supporting hyperplane to $\mathcal{A}$ at its boundary point $(0, 0, p^*)$.

2.3 Examples of value sets and epigraphs (1 point)

For at least one of the following optimization problems, (1) draw a sketch of the sets

$$\mathcal{G} = \{(u, t) \mid \exists x \in \mathcal{D},\ f_0(x) = t,\ f_1(x) = u\}, \quad \mathcal{A} = \{(u, t) \mid \exists x \in \mathcal{D},\ f_0(x) \leq t,\ f_1(x) \leq u\},$$

(2) give the dual problem, and (3) solve the primal and dual problems. Is the problem convex? Is Slater's condition satisfied? Does strong duality hold? The domain of the problem is $\mathbf{R}$ unless otherwise stated.

1. Minimize x such that $x^2 \leq 1$.
2. Minimize x such that $x^2 \leq 0$.
3. Minimize x such that $|x| \leq 0$.
4. Minimize x such that $f_1(x) \leq 0$ where

$$f_1(x) = \begin{cases} -x + 1, & x \geq 1; \\ x, & -1 \leq x \leq 1; \\ -x - 2, & x \leq -1. \end{cases}$$

5. Minimize x^3 such that $-x + 1 \leq 0$.
6. Minimize x^3 such that $-x + 1 \leq 0$ with domain $\mathcal{D} = \mathbf{R}_+$.

2.4 Proof of strong duality under Slater's condition (2 points)

The geometric interpretations, in particular the epigraph description, allows us to prove strong duality for convex optimization problems under Slater's condition using separating hyperplane theorem.

Consider the canonical convex problem on variable $x \in \mathbf{R}^n$ and f_0 has domain $\mathcal{D}$:

$$\begin{array}{ll} \min\limits_{x} & f_0(x) \\ \text{s.t.} & f_i(x) \leq 0 \quad i = 1, \ldots, m, \\ & Ax = b, \end{array} \tag{2}$$

with $f_0, f_1, \ldots, f_m$ convex.

Recall that Slater's condition is that there exists an x in the relative interior of $\mathcal{D}$ such that $f_i(x) < 0$ for $i = 1, \ldots, m$, and $Ax = b$. Such a point is called *strictly feasible.*

For simplify the proof, we make two additional assumptions: first that $\mathcal{D}$ has nonempty interior (hence the relative interior of $\mathcal{D}$ is just the interior of $\mathcal{D}$), and second, that $\operatorname{rank} A = p$, i.e. A has full row rank.

We can assume that p^* is finite without loss of generality. Indeed, since there is a feasible point by Slater's condition, so we can only have $p^* = -\infty$ or p^* is finite. If $p^* = -\infty$, then $d^* = -\infty$ by weak duality.

1. Argue that $\mathcal{A}$ is convex for (2). Show that the following set is disjoint from $\mathcal{A}$.

$$\mathcal{B} = \{(0, 0, s) \in \mathbf{R}^m \times \mathbf{R}^p \times R \mid s < p^*\}.$$

2. Use separating hyperplane theorem to show that there exists $(\tilde{\lambda}, \tilde{\nu}, \mu) \neq 0$ and $\alpha \in \mathbf{R}$ such that

$$\begin{aligned} (u, v, t) \in \mathcal{A} &\implies \tilde{\lambda}^\intercal u + \tilde{\nu}^\intercal v + \mu t \geq \alpha, \\ (u, v, t) \in \mathcal{B} &\implies \tilde{\lambda}^\intercal u + \tilde{\nu}^\intercal v + \mu t \leq \alpha. \end{aligned}$$

   Argue that $\tilde{\lambda} \succeq 0$ and $\mu \geq 0$. (Hint: consider $\tilde{\lambda}^\intercal u + \mu t$ in $\mathcal{A}$.)

3. Simplify the result on $\mathcal{B}$ to $\mu p^* \leq \alpha$. Then plug this into the result on $\mathcal{A}$ to obtain, for any $x \in \mathcal{D}$,

$$\sum_{i=1}^{m} \tilde{\lambda}_i f_i(x) + \tilde{\nu}^\intercal (Ax - b) + \mu f_0(x) \geq \mu p^*. \tag{3}$$

4. Assume $\mu > 0$, then we can divide by μ. Use Lagrangian to show that the above inequality can be written as

$$L(x, \frac{\tilde{\lambda}}{\mu}, \frac{\tilde{\nu}}{\mu}) \geq p^*,$$

   for all $x \in \mathcal{D}$. Then argue $g(\lambda^*, \nu^*) \geq p^*$ by minimizing over x, where we define $\lambda^* = \frac{\tilde{\lambda}}{\mu}$ and $\nu^* = \frac{\tilde{\nu}}{\mu}$.

   Together with weak duality, we have $p^* = g(\lambda^*, \nu^*)$. This shows that for $\mu > 0$, strong duality holds, and the dual optimum is attained.

5. (Optional.) For the case $\mu = 0$, we show contradiction. Apply $\mu = 0$, our condition Eqn (3) becomes

$$\sum_{i=1}^{m} \tilde{\lambda}_i f_i(x) + \tilde{\nu}^\intercal (Ax - b) \geq 0.$$

By Slater's condition, we have a point $\tilde{x}$ that is strictly feasible.

Use $\tilde{x}$ and $(\tilde{\lambda}, \tilde{\nu}, \nu) \neq 0$ to show that

$$\tilde{\lambda} = 0, \mu = 0, \tilde{\nu} \neq 0.$$

This implies that for all $x \in \mathcal{D}$, we have $\tilde{\nu}^\intercal (Ax - b) \geq 0$. Since A is full row rank, we know $A^\intercal \nu \neq 0$.

Use that $\tilde{x}$ satisfy $\tilde{\nu}^\intercal (A\tilde{x} - b) = 0$, and that $\tilde{x}$ is in the interior of $\mathcal{D}$, to show that there is $x \in \mathcal{D}$ such that $\tilde{\nu}^\intercal (Ax - b) < 0$.

But this contradicts our previous result that $\tilde{\nu}^\intercal (Ax - b) \geq 0$ for all $x \in \mathcal{D}$.

3 Principal component analysis (5 points)

This is adapted from Lecture 3 slides of Tibshirani [2].

Principal component analysis (PCA) is a commonly used tool in data analysis to select a few vectors that can best represent a high dimensional dataset. PCA is a low rank approximation problem for a given dataset $X \in \mathbf{R}^{n \times p}$ where n is the number of data points and p is the number of features:

$$\begin{array}{ll} \min\limits_{R} & \|X - R\|_F^2 \\ \text{s.t.} & \operatorname{rank}(R) = k \end{array} \tag{4}$$

Here k is the rank, or number of representative vectors, of the desired low-rank approximation R of X, and $\|A\|_F^2 = \sum_{ij} A_{ij}^2$ is the Frobenius norm of a matrix.

The PCA problem in its current form 4 is not convex. However, we know from practice, that we can compute the PCA problem directly via singular value decomposition (SVD). Given $X = UDV^\intercal$, the solution is $R = U_k D_k V_k^\intercal$, where U_k and V_k are the first k columns of U and V respectively, and D_k is the first k diagonal elements of D. Recall that D_k are non-negative have a decreasing order, so R is the reconstruction of X's k components with the largest singular values, i.e. k principle components.

We will recast the PCA problem into a convex problem in the following steps.

1. Instead of using R as the variable, we can consider a projection matrix Z with rank k that maps X to R. So the objective is now $\|X - XZ\|_F^2$, and the equality constraint is on $\operatorname{rank}(Z) = k$ and Z is an orthogonal projection (therefore symmetric). Next, show that you can recast the problem into the following form, where $S = X^\intercal X$.

$$\begin{array}{ll} \min\limits_{Z \in \mathbf{S}^p} & -\operatorname{tr}(SZ) \\ \text{s.t.} & \operatorname{rank}(Z) = k, \quad Z \text{ is a projection.} \end{array} \tag{5}$$

Hint: the Frobenius norm can be written in terms of trace, $\|A\|_F^2 = \operatorname{tr}(AA^\intercal)$, and use the cyclic property of trace ($\operatorname{tr}(ABC) = \operatorname{tr}(CAB) = \operatorname{tr}(BCA)$) and that Z is symmetric and a projection matrix so $Z^2 = Z$.

2. Now, use the fact that a matrix is a projection if and only if its eigenvalues are in $\{0, 1\}$, and a projection matrix is an orthogonal projection if and only if it is symmetric, to show that the constraint set is

$$C = \{Z \in \mathbf{S}^p \mid \lambda_i(Z) \in \{0, 1\},\ i = 1, \ldots, p,\ \mathrm{tr}(Z) = k\}.$$

   A side note: Recall that singular vectors of SVD gives an orthonormal basis for column and row spaces of a matrix. In particular, the first k columns of V is a basis for the k-dimensional subspace of the row space of X with largest singular values. So, we should take $Z = V_k V_k^\mathsf{T}$, if we use the SVD method.

3. Note that the feasible set C is not convex. But we can do a convex relaxation. Define $\mathcal{F}_k = \mathrm{conv}\{C\}$. Show that

$$\mathcal{F}_k = \{Z \in \mathbf{S}^p \mid 0 \preceq Z \preceq \mathbf{I},\ \mathrm{tr}(Z) = k\}.$$

   $\mathcal{F}_k$ is called a *Fantope* of order k. Conclude that we have recasted the PCA problem into the following convex problem:

$$\min_{Z \in \mathcal{F}_k(Z)} \quad -\mathrm{tr}(SZ) \tag{6}$$

4. Recall that an affine function over a polytope always achieves its optima at the vertices of the polytope. (In fact it is more general, the maximum of a convex function over a polytope always achieves its optima at the vertices of the polytope.) Use this fact to show that our recast of the PCA problem in (6) is equivalent to the original problem in (4), i.e. their solutions are the same. This is a famous result (see Fan (1949), On a theorem of Weyl conerning eigenvalues of linear transformations).

5. Implement in code to check for randomly generated matrix X that the solution to the PCA problem via (6) and the solution using SVD yields the same answers. Food for thought: for cases in practice, when n is large and k is small, which method tend be faster?

4 Sparse problems, basis pursuit and lasso (5 points)

It is commonly a need in practice that we want a sparse solution to a problem, i.e. a solution with few number of nonzeros, or even specify the number of nonzeros.

4.1 Basis pursuit (1 point)

One of the simplest forms of such problems is a sparse constrained minimization problem, where given data $X \in \mathbf{R}^{n \times p}$, $p > n$, and label $y \in \mathbf{R}^n$, we want the sparsest solution fo $X\beta = y$:

$$\begin{aligned} \min_{\beta} \quad & \|\beta\|_0 \\ \text{s.t.} \quad & X\beta = y \end{aligned} \tag{7}$$

Here the $\|\beta\|_0$ denote the number of nonzeros of vector $\beta \in \mathbf{R}^n$, i.e. the cardinality of β. It is often called the "zero norm", but it should be noted that it is not actually a norm. So this problem is not convex.

We can relax the zero norm into an actual norm, the ℓ_1 norm. This gives the ℓ_1 approximation problem, often called *basis pursuit*:

$$\begin{aligned} \min_{\beta} \quad & \|\beta\|_1 \\ \text{s.t.} \quad & X\beta = y \end{aligned} \tag{8}$$

1. Show that the convex hull of the "unit norm ball" of the zero norm, i.e. $\{z : \|z\|_0 \leq 1\}$, is the unit norm ball of the ℓ_1 norm, therefore justifying the relaxation from zero norm to ℓ_1 norm.

2. Show that we can reformulate the basis pursuit problem as a linear program:

$$\begin{array}{ll} \min\limits_{\beta,\, z} & \mathbf{1}^\intercal z \\ \text{s.t.} & \beta \leq z, \\ & -\beta \leq z, \\ & X\beta = y \end{array} \tag{9}$$

4.2 Regressor selection problem and Lasso (3 points)

Another common form of the sparse problem is the sparse regressor problem: given vector $y \in \mathbf{R}^n$ and matrix $X \in \mathbf{R}^{n\times p}$, we would like to find a linear combination of k or less columns of X that best fit y. This can be expressed as

$$\begin{array}{ll} \min\limits_{\beta} & \|y - X\beta\|_2^2 \\ \text{s.t.} & \|\beta\|_0 \leq k \end{array} \tag{10}$$

In general, this is a hard combinatorial problem. But with the power of computers, we can solve this for small problems.

1. One straightforward approach to solve the sparse regressor problem (10) is by checking every possible sparsity pattern in β with k nonzeros. For a fixed sparsity pattern, we can find the optimal β by solving a least-squares problem, i.e., minimizing $\left\|\tilde{X}\tilde{\beta} - y\right\|_2$, where $\tilde{X}$ is the submatrix of X that keeps the columns corresponding to the sparsity pattern, and $\tilde{\beta}$ is the subvector with the nonzero components of β. Solving this for each of the $\frac{n!}{k!(n-k)!}$ possible sparsity patterns, then select the smallest objective value, yields the solution.

   Implement in code an algorithm that solve the sparse regressor problem exactly, following the procedure above. Generate a random example of small size, e.g. $X \in \mathbf{R}^{10\times 20}$, $\beta \in \mathbf{R}^{20}$, $y \in \mathbf{R}^{10}$. Then solve the sparse regressor problem exactly for k going from 10 to 1. Plot, for each k, the minimal cardinality of β on the y axis, and $\|y - X\beta\|_2$ on the x axis.

2. To make this into a problem that we can solve at scale, we can again take the relaxation of the zero norm to the ℓ_1 norm. This yields the lasso problem:

$$\begin{array}{ll} \min\limits_{\beta} & \|y - X\beta\|_2^2 \\ \text{s.t.} & \|\beta\|_1 \leq s \end{array} \tag{11}$$

   To solve this, we can further put this into penalized form below.

$$\min_{\beta} \quad \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{12}$$

   Then we can scan the λ parameter and trace out the optimal tradeoff curve between the objective $\|X\beta - y\|_2$ and sparsity $\|\beta\|_1$.

   Implement a CVXPY program that solves the penalized form of the lasso problem. For the generated X and y, plot the tradeoff curve with $\|X\beta - y\|_2$ on the x axis and $\|\beta\|_1$ on the y axis. Describe any feature you observed. What is the smallest value of $\|\beta\|_1$ where you obtain cardinality $\|\beta\|_0 = k$, for $k = 1, 2, \ldots, 10$?

3. To compare the solution we get from the lasso problem and the exact solution to the sparse regressor problem, we can fix the sparsity pattern we get for a given λ value, and then solve the constrained least squares problem where $\left\|\tilde{X}\tilde{\beta} - y\right\|_2$ is minimized for the given sparsity pattern. Solve this for each point on the tradeoff curve obtained by scanning λ. Then plot on the same figure as the sparse regressor problem, where $\|X\beta - y\|_2$ is the x axis and cardinality of β is the y axis. Compare what the exact solution of the sparse regressor problem and the solution of the lasso procedure.

4.3 Lasso dual (1 point)

We calculate the dual to the lasso problem.

1. The dual to the lasso problem in penalized form (12) is just a scalar, so it is not very useful. Rewrite the lasso problem into the following form and calculate its dual.

$$\begin{array}{ll} \min\limits_{\beta,\, z} & \dfrac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 \\ \text{s.t.} & z - X\beta = 0 \end{array} \tag{13}$$

   Show that the dual problem (after getting rid of constants in the objective) is

$$\begin{array}{ll} \min\limits_{u} & \|y - u\|_2^2 \\ \text{s.t.} & \|X^\intercal u\|_\infty \leq \lambda \end{array} \tag{14}$$

2. Check that Slater's condition holds, so we have strong duality. But note that because we got rid of constants in the objective, the optimal value of the dual problem is not the same as the primal problem. This allows us to use KKT conditions.

   Use KKT conditions to show that, if u is the dual solution, then the primal solution β satisfies $X\beta = y - u$. So the lasso fit is just the dual residual.

4.4 Related facts (Optional)

Although minimizing the cardinality of a vector exactly is hard, certain related functions are easier to handle. Here we consider a few.

1. (Length.) The length of a vector $x \in \mathbf{R}^n$ is the largest index of a nonzero component, i.e. $f(x) = \max\{i \mid x_i \neq 0\}$. The length of a zero vector is zero. Show that this function is quasiconvex on $\mathbf{R}^n$. (Hint: observe that its sublevel sets are subspaces.)

2. (Density, not sparsity.) If the variable β satisfy that it is non-negative, e.g. a probability vector, then recall from Problem 1.4 that the cardinality of β is quasiconcave. (Similarly, the rank of a positive semidefinite matrix is also quasiconcave.) Show that we could use this to solve the reverse problem of (7), i.e. the dense constrained optimization problem that maximize $\|\beta\|_0$, using convex optimization. (Hint: recall that quasiconvex problems can be solved via iterative bisection, each time solving a feasibility convex problem.)

3. (Diversity.) The sum of the k largest components of a vector is $f(x) = \sum_{i=1}^k x_{[i]}$, with $x_{[1]} \geq x_{[2]} \geq \cdots \geq x_{[k]}$ are the components of x sorted in decreasing order. Show that its sublevel set is convex, i.e. contraints of the form $f(x) \leq \alpha$ is convex. (Hint: the sublevel set is equivalent to a finite set of linear inequalities, each of the form $x_{i_1} + \cdots + x_{i_k} \leq \alpha$.)

As a result, we can promote diversity via convex constraints. For example, if x is a probability vector, and we would like the largest k components to not take up more than 80%, then we can constrain $\sum_{i=1}^{k} x_{[i]} \leq 0.8$. This can be used to diversity a portfolio in investment, for example.

5 Support vector machines (5 points)

We used support vector machines to get familiar with CVXPY in the last problem set. Here, we continue to use it as a good example to see how we can rewrite convex problems, and use duality and KKT conditions to analyze a problem.

Given labels $y \in \{-1, 1\}^n$, and a feature matrix $X \in \mathbf{R}^{n \times p}$, with rows $x_1, \ldots, x_n$, the support vector machine (SVM) problem is the following.

$$\begin{array}{ll} \min\limits_{\beta, \beta_0, \xi} & \frac{1}{2}\|\beta\|_2^2 + C\sum\limits_{i=1}^{n} \xi_i \\ \text{s.t.} & \xi_i \geq 0, \qquad i = 1, \ldots, n, \\ & y_i(x_i^\intercal \beta + \beta_0) \geq 1 - \xi_i, \quad i = 1, \ldots, n \end{array} \tag{15}$$

1. (*Hinge form rewrite.*) Rewriting a problem into different forms could yield insights. Rewrite the constraints in 15 as $\xi_i \geq \max\{0, 1 - y_i(x_i^\intercal \beta + \beta_0)\}$. Argue that we have equality at solution.

   Then plug this into the objective function. We have the *hinge form* of SVM.

   $$\min_{\beta, \beta_0} \quad \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^{n} [1 - y_i(x_i^\intercal \beta + \beta_0)]_+ \tag{16}$$

   Here $[a]_+ = \max 0, a$ is the hinge function.

   The hinge form of the SVM problem allows us to interpret it as margin maximization. Consider solving the SVM problem for classification, where the data points are partitioned into two classes, one class has $y_i = 1$, and the other has $y_i = -1$. Then the objective in (16) can be interpreted as finding a hyperplane defined by (β, β_0) (normal vector and intercept) that maximizes the margin between the two classes, with ℓ_2 regularization on β. Argue for yourself why SVM is said to maximize the margin between the two classes. (Hint: if $\sum_{i=1}^{n} [1 - y_i(x_i^\intercal \beta + \beta_0)]_+ = 0$ for all $+$ data points, i.e. with $y_i = +1$, then all $+$ data points satisfy $x_i^\intercal \beta + \beta_0 >= 1$, and if there is a $+$ point achieving equality, then $\beta^\intercal x + \beta_0 = 1$ defines a supporting hyperplane for the $+$ data points.)

2. (*Dual.*) Introduce the dual variables $v, w \geq 0$, write down the Lagrangian, and minimize over β, β_0, ξ to obtain the dual function:

   $$g(v, w) = \begin{cases} -\frac{1}{2} w^\intercal \tilde{X} \tilde{X}^\intercal w + \mathbf{1}^\intercal w, & w = C\mathbf{1} - v,\ w^\intercal y = 0; \\ -\infty, & \text{else.} \end{cases}$$

   Here $\tilde{X} = \mathrm{diag}(y)X$.

   Then eliminate the slack variable v to obtain the dual SVM problem:

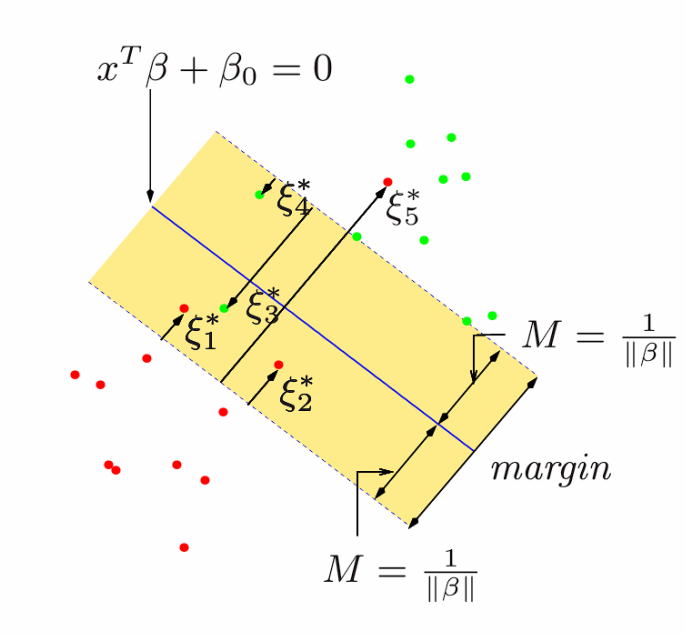

Figure 1 Figure illustrating the KKT conditions and the margin for the SVM problem. Taken from slides of [2] on KKT conditions.

$$\begin{aligned} \min_{w} \quad & -\frac{1}{2} w^\intercal \tilde{X}\tilde{X}^\intercal w + \mathbf{1}^\intercal w \\ \text{s.t.} \quad & 0 \le w \le C\mathbf{1}, \\ & w^\intercal y = 0 \end{aligned} \tag{17}$$

Argue that strong duality holds for the SVM problem if it is feasible.

3. (*KKT.*) Let us investigate the KKT conditions for the SVM problem. Recall that if strong duality holds for an optimization problem, then the primal and dual variables are optimal if and only if they satisfy the KKT conditions.

   Show that the following are the KKT conditions for the primal and dual variables (β, β_0, v, w) of the SVM problem:

$$v, w \ge 0, \quad w^\intercal y = 0, \quad w = C\mathbf{1} - v, \quad \beta = \sum_{i=1}^{n} w_i y_i x_i$$
$$v_i \xi_i = 0, \quad w_i(1 - \xi_i - y_i(x_i^\intercal \beta + \beta_0)) = 0, \quad i = 1, \ldots, n$$

4. (*Support points.*) The KKT conditions give further insights into the solution. At optimality, we have $\beta = \sum_{i=1}^{n} w_i y_i x_i$, a linear combination of the data points. By complementary slackness, w_i is nonzero only when $y_i(x_i^\intercal \beta + \beta_0) = 1 - \xi_i$. Points i satisfying this are called *support points*.

   Show that a support point i satisfies the following property: if $\xi_i = 0$, then x_i is on the edge of the margin, and $w_i \in (0, C]$; if $\xi_i \neq 0$, then x_i is on the wrong side of the margin, and $w_i = C$. See Fig 1 for an illustration.

From this observation, we see that only the support points are included in the optimal solution β. Although the KKT conditions do not give us the solution directly, we see that it gives us insights about the solution. In particular, it gives us a way to speed up the SVM problem: eliminate the non-support points before performing the optimization.

5. (Optional.) Generate random data points in 2d for easy visualization (e.g. two standard normal distributions centered on $(1, 0)$ and $(-1, 0)$), and solve the SVM problem using CVXPY. Then eliminate non-support points to see that indeed you yield the same answer.

6 Practice on duality (5 points)

6.1 A simple example (2 points)

This is exercise 5.1 in [1].

Consider the optimization problem with variable $x \in \mathbf{R}$:

$$\begin{array}{ll} \min\limits_{x} & x^2 + 1 \\ \text{s.t.} & (x-2)(x-4) <= 0 \end{array} \tag{18}$$

1. (Analysis of primal problem.) Give the feasible set, the optimal value, and the optimal solution.
2. (Lagrangian and dual function.) Plot the objective $x^2 + 1$ versus x. On the same plot, show the feasible set, optimal point and value, and plot thee Lagrangian $L(x, \lambda)$ versus x for a few positive values of λ. Verify the lower bound property ($p^* \geq \inf_x L(x, \lambda$ for $\lambda \geq 0$). Derive and sketch the Lagrange dual function g.
3. (Lagrange dual problem.) State the dual problem, and verify that it is a concave maximization problem. Find the dual optimal value and dual optimal solution λ^*. Does strong duality hold?
4. (Sensitivity analysis.) Let $p^*(u)$ denote the optimal value of the problem

$$\begin{array}{ll} \min\limits_{x} & x^2 + 1 \\ \text{s.t.} & (x-2)(x-4) <= u, \end{array} \tag{19}$$

   as a function of the parameter u. Plot $p^*(u)$. Verify that $\frac{dp^*}{du}(0) = -\lambda^*$.

6.2 Optimality condition of QCQP (2 points)

This is exercise 5.26 in [1].

Consider the QCQP with variable $x \in \mathbf{R}^2$:

$$\begin{array}{ll} \min\limits_{x_1, x_2} & x_1^2 + x_2^2 \\ \text{s.t.} & (x_1 - 1)^2 + (x_2 - 1)^2 \leq 1, \\ & (x_1 - 1)^2 + (x_2 + 1)^2 \leq 1 \end{array} \tag{20}$$

1. Sketch the feasible set and level sets of the objective. Find the optimal point x^* and optimal value p^*.
2. Give the KKT conditions. Do there exist Lagrange multipliers λ_1^* and λ_2^* that prove that x^* is optimal?
3. Derive and solve the Lagrangeu dual problem. Does strong duality hold?

6.3 Optimality conditions for LP (1 point)

Prove (without using any linear programming code) that the optimal solution of the following LP is unique, and given by $x^* = (1, 1, 1, 1)$. This is exercise 5.28 of [1].

$$\begin{array}{ll} \min & 47x_1 + 93x_2 + 17x_3 - 93x_4 \\ \text{s.t.} & \begin{bmatrix} -1 & -6 & 1 & 3 \\ -1 & -2 & 7 & 1 \\ 0 & 3 & -10 & -1 \\ -6 & -11 & -2 & 12 \\ 1 & 6 & -1 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \preceq \begin{bmatrix} -3 \\ 5 \\ 8 \\ -7 \\ 4 \end{bmatrix} \end{array} \tag{21}$$

References

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004. [Online]. Available: http://www.stanford.edu/~boyd/cvxbook/

[2] R. Tibshirani, "Course on convex optimization," 2019. [Online]. Available: https://www.stat.cmu.edu/~ryantibs/convexopt/