1 Convexity I

1.1

Proof
Suppose f_1 and f_2 are convex on G. By convexity of f_1 and f_2, we have

$$f_1(\theta x + (1-\theta)z) \le \theta f_1(x) + (1-\theta) f_1(z)$$
$$f_2(\theta x + (1-\theta)z) \le \theta f_2(x) + (1-\theta) f_2(z)$$

where $x, z \in G$, $\theta \in [0,1]$. Consider the sum of f_1 and f_2 on G

$$\begin{aligned} f(\theta x + (1-\theta)z) &= f_1(\theta x + (1-\theta)z) + f_2(\theta x + (1-\theta)z) \\ &\le \theta[f_1(x) + f_2(x)] + (1-\theta)[f_1(z) + f_2(z)] \\ &= \theta f(x) + (1-\theta) f(z) \end{aligned}$$

1.2

1.2.1

Proof
Since g_i are convex on G, we have

$$g_i(\theta x + (1-\theta)z) \le \theta g_i(x) + (1-\theta) g_i(z) \ \forall i$$

where $x, z \in G$, $\theta \in [0,1]$.
Since h is monotone increasing in all components, we have $h(\cdots, x_{i,1}, \cdots) \le h(\cdots, x_{i,2}, \cdots)$ if $x_{i,1} <= x_{i,2}$ for any i. We therefore have

$$\begin{aligned} f(\theta x + (1-\theta)z) &= h(g_1(\theta x + (1-\theta)z), \cdots, g_n(\theta x + (1-\theta)z)) \\ &\le h(\theta g_1(x) + (1-\theta) g_1(z), \cdots, \theta g_n(x) + (1-\theta) g_n(z)) \\ &\le \theta h(g_1(x), \cdots, g_n(x)) + (1-\theta) h(g_1(z), \cdots, g_n(z)) \\ &= \theta f(x) + (1-\theta) f(z) \end{aligned}$$

where line 2 to line 3 is by the convexity of h.

1.2.2

Proof
Since g_i are affine functions, they satisfy linearity property

$$g_i(ax + by) = q g_i(x) + b g_i(y)$$

We therefore have

$$\begin{aligned} f(\theta x + (1-\theta)z) &= h(g_1(\theta x + (1-\theta)z), \cdots, g_n(\theta x + (1-\theta)z)) \\ &= h(\theta g_1(x) + (1-\theta) g_1(z), \cdots, \theta g_n(x) + (1-\theta) g_n(z)) \\ &\le \theta h(g_1(x), \cdots, g_n(x)) + (1-\theta) h(g_1(z), \cdots, g_n(z)) \\ &= \theta f(x) + (1-\theta) f(z) \end{aligned}$$

where line 2 to line 3 is again by the convexity of h.

1.2.3

Proof
Since g_i are convex on G, we have

$$g_i(\theta x + (1-\theta)z) \geq \theta g_i(x) + (1-\theta)g_i(z) \ \forall i$$

where $x, z \in G$, $\theta \in [0, 1]$.
Since h is monotone decreasing in all components, we have $h(\cdots, x_{i,1}, \cdots) \leq h(\cdots, x_{i,2}, \cdots)$ if $x_{i,1} >= x_{i,2}$ for any i. We therefore have

$$\begin{aligned}
f(\theta x + (1-\theta)z) &= h(g_1(\theta x + (1-\theta)z), \cdots, g_n(\theta x + (1-\theta)z)) \\
&\leq h(\theta g_1(x) + (1-\theta)g_1(z), \cdots, \theta g_n(x) + (1-\theta)g_n(z)) \\
&\leq \theta h(g_1(x), \cdots, g_n(x)) + (1-\theta)h(g_1(z), \cdots, g_n(z)) \\
&= \theta f(x) + (1-\theta)f(z)
\end{aligned}$$

where line 2 to line 3 is by the convexity of h.

1.3

Proof
The maximum is achieved either inside the polyhedron or on its boundaries. We will discuss the two case. Let x be a point where maximum of f is achieved.

- Suppose x is inside P
  Any line that passes through x must intersect the polyhedron at two points, denoted x_1 and x_2. Then we know that $x = \theta x_1 + (1-\theta)x_2$ for some θ. By convexity of f we have

  $$f(x) = f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta)f(x_2) \leq \max\{f(x_1), f(x_2)\}$$

  However, since x is strictly inside P (assumed that if x is a maximum, then it is not on boundary), $f(x_1)$ and $f(x_2)$ must be strictly smaller than $f(x)$ as they are both on the boundary of P. Thus a contradiction with the above equation. Therefore, maximum must be achieved at boundaries of P.

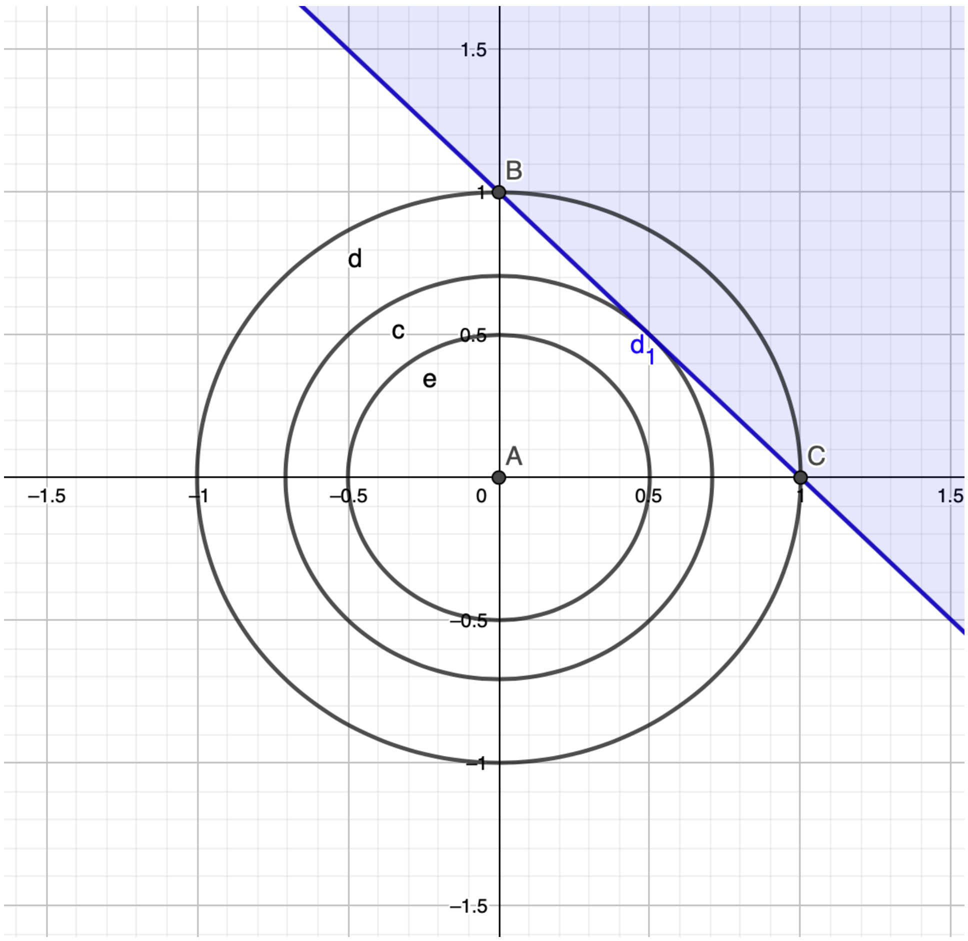
- Suppose x is on the boundary of P
  Assume the two vertices of the boundary are x_1 and x_2, we know that there is a θ such that $x = \theta x_1 + (1-\theta)x_2$. By convexity of f we have

  $$f(x) = f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta)f(x_2) \leq \max\{f(x_1), f(x_2)\}$$

  Therefore, maximum is achieved at one of x_1 and x_2 or both.

2 Convexity II

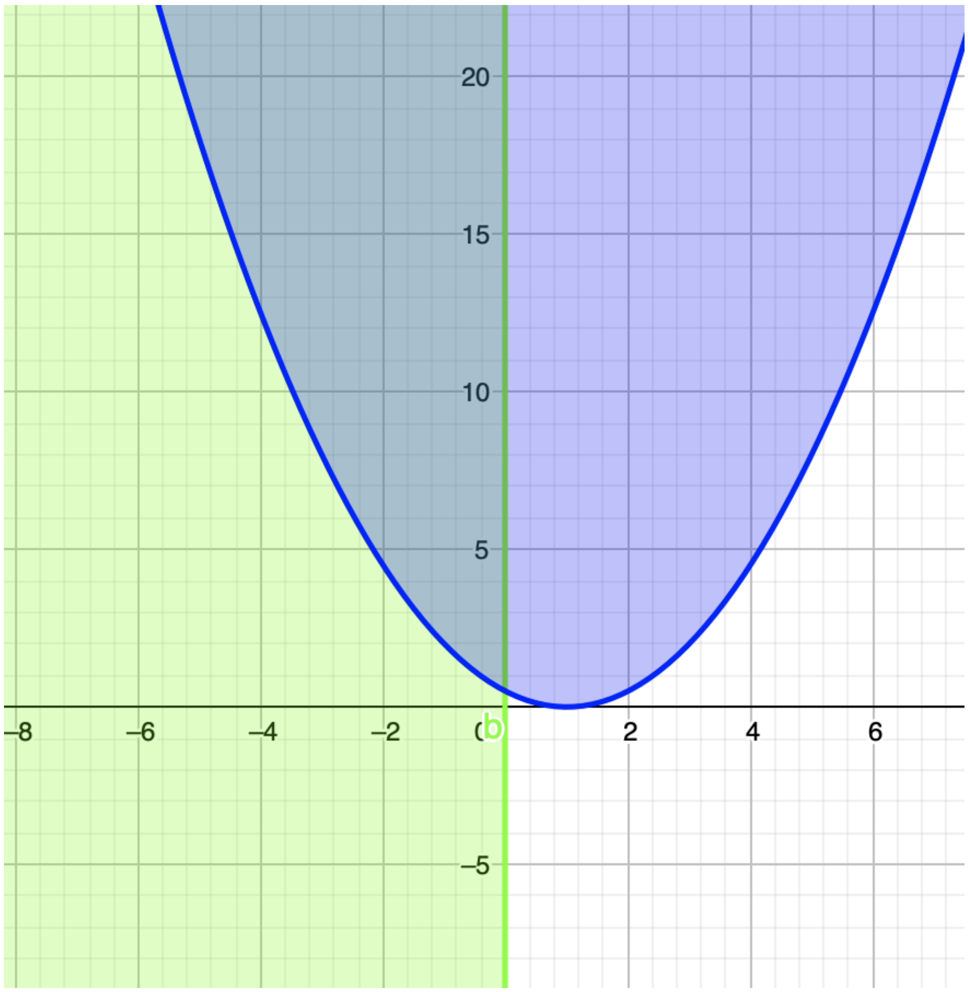
2.1



Feasible region is shaded bluish purple and the intersection $d_1(0.5, 0.5)$ is the optimal (x_1*, x_2*)

2.2

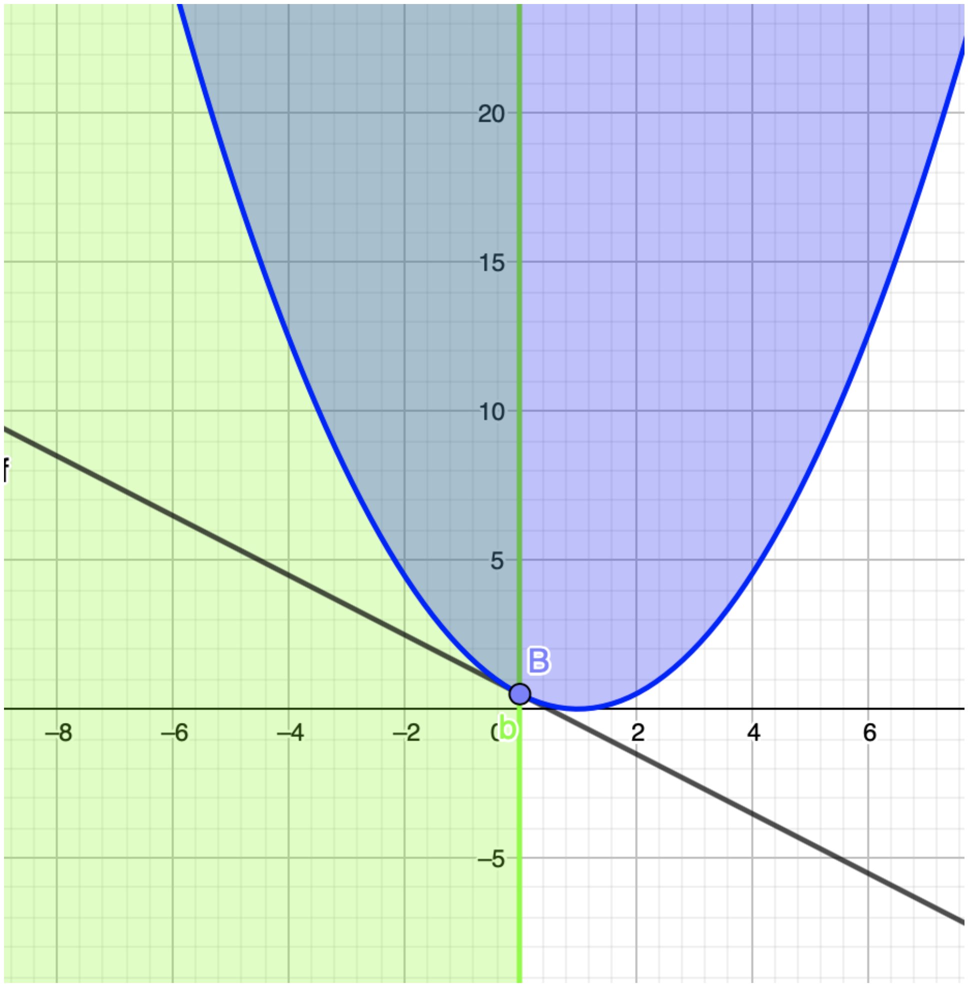
For region R, consider $g(x) = 1 - x_1 - x_2 = c$ where c is a constant. Then $x_1 = 1-c-x_2$. Therefore, we have $f(x) = x_1^2+x_2^2 = 2x_2^2+2(c-1)x_2+(c-1)^2$. This function achieves minimum at $x_2 = \frac{1-c}{2}$, which yields $f(x)_min = \frac{(c-1)^2}{2}$. Notice that $f(x)$ is unbounded above. Therefore, for a given c, all points above $\frac{(c-1)^2}{2}$ is achievable.
For region F, we need $g(x) \leq 0$, which means $y = g(w) \leq 0$. The purple region



is R, and the green region is F.

2.3

Using the stationary condition, we differentiate $L(x, \lambda)$ with respect to x_1 to get $2x_1^* - \lambda^* = 0$. Since we already know from part 1 that the optimal $(x_1^*, x_2^*) = (0.5, 0.5)$, we obtain that $\lambda^* = 1$. Therefore, easy to see that we need to plot point $(y, z) = (1 - 0.5 - 0.5, 0.5^2 + 0.5^2) = (0, 0.5)$, and the line $z + y = min(x_1^2 + x_2^2 + (1 - x_1 - x_2)) = 0.5^2 + 0.5^2 + (1 - 0.5 - 0.5) = 0.5$.



2.4

$$q(\lambda) = \min_x f(x) + \lambda g(x)$$

For a fixed x, we know that $f(x)+\lambda g(x)$ is affine with respect to λ, and therefore is concave. $q(\lambda)$ is equivalent to the minimum of a collection of concave functions (i.e. $f(x) + \lambda g(x)$ for different values of x), and therefore is also concave.

3 Support Vector Machine

3.1

The Lagrangian is the following:

$$\mathcal{L}(w, b, \epsilon, \alpha, \beta) \quad = \quad \frac{1}{2}w^T w + C\sum_{i=1}^{n}\epsilon_i + \sum_{i=1}^{n}\alpha_i[1 - \epsilon_i - y_i(w^T x_i + b)] - \sum_{i=1}^{n}\beta_i \cdot \epsilon_i$$

Apply KKT conditions, we have:
Primal feasibility: $y_i(w^T x_i + b) \geq (1 - \epsilon_i)$, $\epsilon_i \geq 0$, $\forall i$
Dual feasibility: $\alpha_i \geq 0$, $\beta \geq 0$, $\forall i$
Complementary slackness: $\alpha_i[1 - \epsilon_i - y_i(w^T x_i + b)] = 0$, $\beta_i \epsilon_i = 0$, $\forall i$
Stationary:

- Differentiate against w:

$$w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0$$

- Differentiate against b:

$$\sum_{i=1}^{n}\alpha_i y_i = 0$$

- Differentiate against ϵ_i:

$$\alpha_i + \beta_i = C, \forall i$$

Plug in the stationary conditions back into the Lagrangian, we will have the following:

$$\mathcal{L}(w^*, b^*, \epsilon^*, \alpha, \beta) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

The dual problem is maximizing the above with respect to all α and β. The dual feasibility constraint $\beta_i \geq 0$ can be modified to $C - \alpha_i \geq 0$ by one of the relations derived. Therefore, the dual problem will contain only α in the expression to be maximized as well as all the feasibility constraints.

3.2

- If $y_i(w^T x_i + b) < 1$:
  We have $\epsilon_i > 0$ which is a positive penalty. Since $\beta_i \epsilon_i = 0$, we know that $\beta_i = 0$. Since $\alpha_i + \beta_i = C$, we know that $\alpha_i = C \neq 0$

- If $y_i(w^T x_i + b) = 1$:
  We have $\epsilon_i = 0$. Since $\beta_i \epsilon_i = 0$, we know that $\beta_i \geq 0$. Since $\alpha_i + \beta_i = C$, we know that $0 \leq \alpha_i \leq C$. In this case, a_i can be 0, but it also means that the optimal of α_i can be in this large range and it is not necessary that we pick a zero α in the quadratic programming process.

4 Kernels

4.1

We can define the Hilbert space using either kernel function as below:

$$H_k = \text{span}(k(\cdot, x) : x \in X)$$

where $k(x, y) = x^T y$.
By Representer theorem, we know that there exists an optimal solution that maximizes L, and is of the following form:

$$f^* = \sum_{i=1}^{n} a_i k(\cdot, x_i)$$

where a_i are some constants.

4.2

4.2.1

Since k_1 and k_2 are valid kernels, we know that feature space mapping ϕ_1 and ϕ_2 exists and

$$\begin{aligned} k_1(x, z) &=< \phi_1(x), \phi_1(z) > \\ k_2(x, z) &=< \phi_2(x), \phi_2(z) > \end{aligned}$$

then

$$\begin{aligned} k(x, z) &= \alpha k_1(x, z) + \beta k_2(x, z) \\ &=< \sqrt{\alpha}\phi_1(x), \sqrt{\alpha}\phi_1(z) > + < \sqrt{\beta}\phi_2(x), \sqrt{\beta}\phi_2(z) > \\ &= \left\langle \sqrt{\alpha}\phi_1(x)\sqrt{\beta}\phi_2(x), \sqrt{\alpha}\phi_1(z)\sqrt{\beta}\phi_2(z) \right\rangle \end{aligned}$$

which represents the concatenation of feature spaces.

4.2.2

By Mercer's theorem, we can write the kernel functions in the following way

$$\begin{aligned} k_1(x, z) &= \sum_{i=1}^{\infty} a_i \phi_i(x)\phi_i(z) \\ k_2(x, z) &= \sum_{i=1}^{\infty} b_i \psi_i(x)\psi_i(z) \end{aligned}$$

Therefore,

$$\begin{aligned} k(x,z) &= \left(\sum_{i=1}^{\infty} a_i \phi_i(x)\phi_i(z)\right)\left(\sum_{i=1}^{\infty} b_i \psi_i(x)\psi_i(z)\right) \\ &= \sum_{i,j} a_i b_j \phi_i(x)\psi_i(x)\ \psi_i(z)\phi_i(z) \end{aligned}$$

Let $f_k(\cdot) = \phi_i(\cdot)\psi_j(\cdot)$ and $m_k = a_i b_j$ such that each ordered pair (i,j) is mapped to a unique k. Then we have

$$k(x,z) = \sum_{k=1}^{\infty} m_k f_k(x) f_k(z)$$

which corresponds to the feature map $\Phi(x) = [\cdots, \sqrt{m_k} f_k(x), \cdots]$ using ordinary dot product as the inner product.

4.2.3

We can find the feature map to be $\Phi(x) = [f(x)]$

4.2.4

Suppose $f(k_1(x,z)) = \sum_{i=0}^{n} a_i k_1(x,z)^i$, then the new kernel is just a linear combination of monomials. Each monomial is a product of several valid kernels (in this case, product of k_1). Therefore, by results of the previous proofs, $k(x,z) = f(k_1(x,z))$ is a valid kernel.

4.2.5

Consider $e^{2ax^T z}$, we can apply Taylor expansion as the following:

$$e^{2ax^T z} = \sum_{i=0}^{n} (2a)^i \frac{(x^T z)^i}{i!}$$

which is a positive (coefficients are positive) linear combination of the valid kernel $k(x,z) = x^T z$. Therefore, $e^2 ax^T z$ is a valid kernel. Then

$$e^{-a|x-z|^2} = e^{-ax^T x} e^{-az^T z} e^{2a(x^T z)}$$

By results of earlier proofs, $e^{-ax^T x} e^{-az^T z} = f(x)g(z)$ is a valid kernel, and therefore, the product overall is a valid kernel.