Machine Learning B (2025)
Home Assignment 2

Yasin Baysal, cmv882

Contents

1 Convex geometry (25 points)

Response to Question 1. We consider an arbitrary positive integer $d \geq 1$ and an arbitrary set of n points given by $S = (x_1, \ldots, x_n) \subset \mathbb{R}^d$ with $n > d$. We then let $\mathrm{Co}(S)$ denote the convex hull of the set S and have to prove the following property:

(P1) Any $x \in \mathrm{Co}(S)$ can be written as a convex combination of at most $d+1$ points in S.

Proof. Suppose that $S = (x_1, \ldots, x_n) \subset \mathbb{R}^d$ with $n > d$ and fix any point

$$x \in \mathrm{Co}(S) = \left\{ x = \sum_{i=1}^{n} \theta_i x_i \;\middle|\; \theta_i \geq 0 \ \forall i,\ \sum_{i=1}^{n} \theta_i = 1 \right\}.$$

By the definition of the convex hull, there exists weights $\theta_i \geq 0$ with $\sum_{i=1}^{n} \theta_i = 1$, so that

$$x = \sum_{i=1}^{n} \theta_i x_i$$

represents the convex combination of the set of points given by S. Now, to reduce the convex combinations to conic combinations so that we can invoke Lemma 1 from the assignment paper on cones, we first define the "lifted" points in one higher dimension by

$$x_i \mapsto y_i = (x_i, 1) \in \mathbb{R}^{d+1}, \quad x \mapsto y = (x, 1) \in \mathbb{R}^{d+1},$$

where we used that if $S \subset \mathbb{R}^d$, then $S \subset \mathbb{R}^d$ has a subset $\{\theta_i \in \mathbb{R}^{d+1} \colon \theta_{d+1} = 1\}$ (i.e., we identifty $\mathbb{R}^d$ as an affine hyperplane). Therefore, we can set

$$S' = S \times \{1\} = \{(x_i, 1) \colon i = 1, \ldots, n\} \subset \mathbb{R}^{d+1}.$$

Then, we see that y can be written as

$$y = (x, 1) = \left(\sum_{i=1}^{n} \theta_i x_i, \sum_{i=1}^{n} \theta_i \right) = \sum_{i=1}^{n} \theta_i (x_i, 1) = \sum_{i=1}^{n} \theta_i y_i$$

with all $\theta_i \geq 0$. Hence, y lies in the cone generated by the embedding of S. Hereafter, by the definition of the conic hull, we know that

$$\mathrm{Cone}(S') = \left\{ \sum_{i=1}^{n} \theta_i (x_i, 1) \;\middle|\; (x_i, 1) \in S', \theta_i \geq 0, \forall i \right\},$$

and since we have shown that $y = (x, 1) = \sum_{i=1}^{n} \theta_i y_i$ with all $\theta_i \geq 0$, it follows that $y \in \mathrm{Cone}(S') = \mathrm{Cone}(\{y_1, \ldots, y_n\})$. Thus, we have exhibited y as a conic combination of the y_i, which is exactly the setup needed to invoke Lemma 1 on the cone generated by $\{y_i\}$.

First, we let $n \geq d+1$. Accordingly, by the given Lemma 1 applied with dimension $d+1$ to the cone $\mathrm{Cone}(S')$ stated above, there is subset $\mathcal{I} \subseteq \{1, \ldots, n\}$ with cardinality $|\mathcal{I}| \leq d+1$ and nonnegative coefficients $\{\theta_i\}_{i \in \mathcal{I}}$, such that for all $i \in \mathcal{I}$, we get that

$$y = \sum_{i \in \mathcal{I}} \theta_i y_i, \quad \text{with} \quad y = (x, 1), \quad y_i = (x_i, 1),$$

where at most $d+1$ of θ_i are nonzero. Looking at the first d coordinates gives us that

$$x = \sum_{i \in \mathcal{I}} \theta_i x_i,$$

since the first d entries of the left side $y = (x, 1)$ are exactly the vector $x \in \mathbb{R}^d$, while the first d entries of each $\theta_i y_i = \theta_i(x_i, 1)$ on the right side are $\theta_i x_i$. Hereafter, by looking at the last (i.e., the $(d+1)$-th) coordinate, it gives us that

$$1 = \sum_{i \in \mathcal{I}} \theta_i,$$

since the $(d+1)$-st coordinate of $y = (x, 1)$ is 1, while each $y_i = (x_i, 1)$ has $(d+1)$-st coordinate 1. But $\theta_i \geq 0$ and $\sum_{i \in \mathcal{I}} \theta_i$, so this is exactly a convex combination of the $\{x_i\}_{i \in \mathcal{I}}$. Since $|\mathcal{I}| \leq d+1$, we have finally shown that every $x \in \mathrm{Co}(S)$ can be written as a convex combination of at most $d+1$ points of S. This completes the proof of (P1). □

Response to Question 2. No, **(P1)** does not hold if we replace at most $d+1$ points with just at most d points, which we will justify in the following proof by contradiction:

Proof by contradiction. First, recall (see page 2 in Matoušek (2002)) that in $\mathbb{R}^d$, any affinely independent family can have at most $d+1$ members. In particular, there do exist families of exactly $d+1$ affinely independent points—for example the vertices of a standard d-simplex.

Let $\{x_1, x_2, \ldots, x_{d+1}\} \subset \mathbb{R}^d$ be any affine basis of $\mathbb{R}^d$, i.e. the $d+1$ points are affinely independent. Then, we define

$$x := \frac{1}{d+1} \sum_{i=1}^{d+1} x_i.$$

Here, we clearly see that $x \in \mathrm{Co}(S)$ (i.e., that x exhibits a convex combination of the points in S), since the coefficients are nonnegative and sum to 1, as we see that

$$x = \sum_{i=1}^{d+1} \underbrace{\frac{1}{d+1}}_{\geq 0} x_i, \quad \sum_{i=1}^{d+1} \frac{1}{d+1} = (d+1)\frac{1}{d+1} = 1.$$

Now suppose for the sake of contradiction that x could also be written as a convex combination of only d of these points. That is, assume there exist nonnegative $\alpha_1, \ldots, \alpha_d$ with

$\sum_{i=1}^{d} \alpha_i = 1$, such that

$$x = \sum_{i=1}^{d} \alpha_i x_i.$$

Since, now both $x = \sum_{i=1}^{d} \alpha_i x_i$ and $\frac{1}{d+1} \sum_{i=1}^{d+1} x_i = \frac{1}{d+1} \left(\sum_{i=1}^{d} x_i + x_{d+1} \right)$, then we get that

$$\sum_{i=1}^{d} \alpha_i x_i = \frac{1}{d+1} \sum_{i=1}^{d} x_i + \frac{1}{d+1} x_{d+1} \Leftrightarrow \sum_{i=1}^{d} \left(\alpha_i - \frac{1}{d+1} \right) x_i - \frac{1}{d+1} x_{d+1} = 0.$$

Then, we set $\lambda_{d+1} = -\frac{1}{d+1}$ and $\lambda_i = \alpha_i - \frac{1}{d+1}$, such that we now get

$$\sum_{i=1}^{d} \left(\alpha_i - \frac{1}{d+1} \right) x_i - \frac{1}{d+1} x_{d+1} = 0 \Leftrightarrow \sum_{i=1}^{d} \lambda_i x_i + \lambda_{d+1} x_{d+1} = 0,$$

while we also see that

$$\sum_{i=1}^{d+1} \lambda_i = \sum_{i=1}^{d} \left(\alpha_i - \frac{1}{d+1} \right) + \lambda_{d+1} = \sum_{i=1}^{d} \alpha_i - \frac{d}{d+1} - \frac{1}{d+1} = \sum_{i=1}^{d} \alpha_i - 1 = 1 - 1 = 0,$$

but not all λ_i's are zero, since $\lambda_{d+1} = -\frac{1}{d+1} \neq 0$. Hence, the relation

$$\sum_{i=1}^{d+1} \lambda_i x_i = 0, \quad \sum_{i=1}^{d+1} \lambda_i = 0,$$

is exactly an affine dependence among $x_1, \ldots, x_{d+1}$. We began by assuming $x_1, \ldots, x_{d+1}$ were affinely independent. The above gives a nontrivial affine dependence among the $d+1$ points of S (a contradiction), so it is not possible to express x using only d points of S.

Therefore our assumption was false: no convex combination of only d of the points x_i can equal x. In particular, one cannot always write an arbitrary point of $\mathrm{Co}(S)$ as a convex combination of merely d points, since there is a point of $\mathrm{Co}(S)$ (i.e., the point x we specifically have chosen) that cannot be written using the at most d points. Hence, the property **(P1)** fails if "at most $d+1$" is replaced by "at most d". □

2 Convex functions (25 points)

Response to Question 3. Let $f \colon \mathbb{R}^d \to \mathbb{R}$ be a convex function.

(1) *Proof.* Recall that a set $S \subseteq \mathbb{R}^{d+1}$ is convex if for any two points in S, every convex combination (i.e., the line segment between any two points) of them also lies in S. Let

$$(x_1, t_1), (x_2, t_2) \in \mathrm{Epi}(f), \text{ where } \mathrm{Epi}(f) = \left((x, t) \in \mathbb{R}^d \times \mathbb{R} \mid f(x) \leq t \right).$$

Also, let $\theta \in [0,1]$. By the definition of the epigraph stated above, we first have that

$$f(x_1) \leq t_1 \text{ and } f(x_2) \leq t_2.$$

Next, we consider the convex combination given by

$$(x_\theta, t_\theta) = (\theta x_1 + (1-\theta)x_2, \theta t_1 + (1-\theta)t_2)\,.$$

Since f is a convex function by assumption, Jensen's inequality gives us that

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta)f(x_2) \leq \theta t_1 + (1-\theta)t_2 = t_\theta.$$

Hence, we get that $f(x_\theta) \leq t_\theta$, which then means that the convex combination $(x_\theta, t_\theta) \in \text{Epi}(f)$. Because this holds for every choice of $(x_1,t_1),(x_2,t_2) \in \text{Epi}(f)$ and also every $\theta \in [0,1]$, we conclude that $\text{Epi}(f)$ is a convex set as desired. □

(2) *Proof.* Suppose that the epigraph of the function $f\colon \mathbb{R}^d \to \mathbb{R}$ given by

$$\text{Epi}(f) = \big((x,t) \in \mathbb{R}^d \times \mathbb{R} \mid f(x) \leq t\big)$$

is a convex set. We must show that for all $x_1, x_2 \in \mathbb{R}^d$ and all $\theta \in [0,1]$, it holds that

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta)f(x_2).$$

First, let $(x_1,y_1),(x_2,y_2) \in \text{Epi}(f)$, so that $f(x_1) \leq y_1$ and $f(x_2) \leq y_2$. For any $\theta \in [0,1]$, convexity of the epigraph $\text{Epi}(f)$ gives us that

$$(1-\theta)(x_1,y_1) + \theta(x_2,y_2) = ((1-\theta)x_1 + \theta x_2, (1-\theta)y_1 + \theta y_2) \in \text{Epi}(f),$$

and by using $f(x_1) \leq y_1$ and $f(x_2) \leq y_2$ from before, this implies that

$$f((1-\theta)x_1 + \theta x_2) \leq (1-\theta)y_1 + \theta y_2.$$

This holds for any $(x_i,y_i) \in \text{Epi}(f)$. In particular, let us choose $y_1 = f(x_1)$ and $y_2 = f(x_2)$, which certainly lie in the epigraph. Now, this simply yields

$$f((1-\theta)x_1 + \theta x_2) \leq (1-\theta)f(x_1) + \theta f(x_2),$$

which is exactly the definition of convexity of the function f. Since x_1, x_2 and $\theta \in [0,1]$ were chosen arbitrarily, f is therefore a convex function as desired. □

Response to Question 4. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a convex function, $x \in \mathbb{R}^d$ and $t \in \mathbb{R}_{++}$, where $\mathbb{R}_{++}$ denotes the set of strictly positive real-values.

Proof. The domain of the function $g\colon \mathbb{R}^d \times \mathbb{R}_{++} \to \mathbb{R}$ defined by $g(x,t) = tf(\frac{x}{t})$ is

$$\text{Dom}(g) = \{(x,t) \in \mathbb{R}^d \times \mathbb{R}_{++} \mid \tfrac{x}{t} \in \text{Dom}(f), t > 0\}.$$

Take any two points in the domain of g, say $(x_1, t_1), (x_2, t_2) \in \mathrm{Dom}(g)$ and weight $\theta \in [0, 1]$. In particular, $t_1, t_2 > 0$ and $x_i/t_i \in \mathrm{Dom}(f)$. Now, form their convex combination by

$$\begin{aligned}(x, t) &= \theta(x_1, t_1) + (1-\theta)(x_2, t_2) \\ &= (\theta x_1 + (1-\theta)x_2, \theta t_1 + (1-\theta)t_2).\end{aligned}$$

Here, we see that $t = \theta t_1 + (1-\theta)t_2 > 0$ and

$$\frac{x}{t} = \frac{\theta x_1 + (1-\theta)x_2}{\theta t_1 + (1-\theta)t_2} \in \mathrm{Dom}(f),$$

so $(x, t) \in \mathrm{Dom}(g)$. Now, we can write out $g(x, t)$ as

$$\begin{aligned}g(x,t) &= tf\left(\frac{x}{t}\right) \\ &= (\theta t_1 + (1-\theta)t_2) f\left(\frac{\theta x_1 + (1-\theta)x_2}{\theta t_1 + (1-\theta)t_2}\right) \\ &= (\theta t_1 + (1-\theta)t_2) f\left(\frac{\theta t_1}{\theta t_1 + (1-\theta)t_2}\frac{x_1}{t_1} + \frac{(1-\theta)t_2}{\theta t_1 + (1-\theta)t_2}\frac{x_2}{t_2}\right) \\ &= (\theta t_1 + (1-\theta)t_2) f\left(\lambda\frac{x_1}{t_1} + (1-\lambda)\frac{x_2}{t_2}\right),\end{aligned}$$

where we set $\lambda = \frac{\theta t_1}{\theta t_1 + (1-\theta)t_2}$ and $1 - \lambda = \frac{(1-\theta)t_2}{\theta t_1 + (1-\theta)t_2}$ for $\lambda \in [0, 1]$. By convexity of f, we get

$$f\left(\lambda\frac{x_1}{t_1} + (1-\lambda)\frac{x_2}{t_2}\right) \leq \lambda f\left(\frac{x_1}{t_1}\right) + (1-\lambda) f\left(\frac{x_2}{t_2}\right).$$

Substituting this back into the expression for the function $g(x, t)$, it then gives us that

$$\begin{aligned}g(x,t) &= (\theta t_1 + (1-\theta)t_2) f\left(\lambda\frac{x_1}{t_1} + (1-\lambda)\frac{x_2}{t_2}\right) \\ &\leq (\theta t_1 + (1-\theta)t_2)\left[\lambda f\left(\frac{x_1}{t_1}\right) + (1-\lambda) f\left(\frac{x_2}{t_2}\right)\right] \\ &= (\theta t_1 + (1-\theta)t_2)\lambda f\left(\frac{x_1}{t_1}\right) + (\theta t_1 + (1-\theta)t_2)(1-\lambda) f\left(\frac{x_2}{t_2}\right) \\ &= (\theta t_1 + (1-\theta)t_2)\frac{\theta t_1}{\theta t_1 + (1-\theta)t_2} f\left(\frac{x_1}{t_1}\right) + (\theta t_1 + (1-\theta)t_2)\frac{(1-\theta)t_2}{\theta t_1 + (1-\theta)t_2} f\left(\frac{x_2}{t_2}\right) \\ &= \theta t_1 f\left(\frac{x_1}{t_1}\right) + (1-\theta)t_2 f\left(\frac{x_2}{t_2}\right) \\ &= \theta g(x_1, t_1) + (1-\theta) g(x_2, t_2).\end{aligned}$$

which is exactly the convexity inequality for g. Since our choice of $(x_1, t_1), (x_2, t_2) \in \mathrm{Dom}(g)$ and $\theta \in [0, 1]$ were arbitrary, we conclude that g is also convex as desired. $\square$

3 Lagrange duality (25 points)

Response to Question 5. We have to derive the Lagrange dual problem of the primal linear program given by

$$\begin{aligned} \min_{w \in \mathbb{R}^d} \quad & c^\top w \\ \text{s.t.} \quad & Aw = b \\ & w \succeq 0, \end{aligned}$$

where $A \in \mathbb{R}^{p \times d}$, $b \in \mathbb{R}^p$ and $c \in \mathbb{R}^d$. We define the Lagrangian of the above problem to be

$$\begin{aligned} \mathcal{L}(w, \lambda, \nu) &= c^\top w + \lambda^\top (-w) + \nu^\top (Aw - b) \\ &= \left(c + A^\top \nu - \lambda\right)^\top w - \nu^\top b, \end{aligned}$$

where $\lambda \in \mathbb{R}^d$ and $\nu \in \mathbb{R}^p$ are Lagrange multipliers for the inequality and equality constraints, respectively. The Lagrange dual function $\phi \colon \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ is defined to be

$$\begin{aligned} \phi(\lambda, \nu) &= \inf_{w \in \mathbb{R}^d} \mathcal{L}(w, \lambda, \nu) \\ &= \inf_{w \in \mathbb{R}^d} \left(\left(c + A^\top \nu - \lambda\right)^\top w - \nu^\top b \right) \\ &= -\nu^\top b + \inf_{w \in \mathbb{R}^d} \left(\left(c + A^\top \nu - \lambda\right)^\top w \right). \end{aligned}$$

This is easily determined, as a linear function is bounded below only when it is identically zero. If $c + A^\top \nu - \lambda \neq 0$, then $\inf_{w \in \mathbb{R}^d} \left(c + A^\top \nu - \lambda\right)^\top w = -\infty$. To keep $\phi(\lambda, \nu) > -\infty$, we must have $c + A^\top \nu - \lambda = 0 \Leftrightarrow c + A^\top \nu = \lambda$, so the dependence on w drops out and

$$\phi(\lambda, \nu) = \begin{cases} -b^\top \nu & \text{if } c + A^\top \nu = \lambda \\ -\infty & \text{otherwise.} \end{cases}$$

Thus, the Lagrange dual problem is defined to be

$$\begin{aligned} \max_{\lambda, \nu} \quad & \phi(\lambda, \nu) \\ \text{s.t.} \quad & \lambda \succeq 0. \end{aligned}$$

Since we know that $\phi(\lambda, \nu) = -\nu^\top b$ and $\lambda = c + A^\top \nu$ for ϕ to be finite (since otherwise the term $\inf_{w \in \mathbb{R}^d}((c + A^\top \nu - \lambda)^\top w)$ would be $-\infty$), it gives us the equivalent formulation

$$\begin{aligned} \max_{\lambda, \nu} \quad & -\nu^\top b \\ \text{s.t.} \quad & c + A^\top \nu \succeq 0, \end{aligned}$$

which is the Lagrange dual problem of the original primal linear program.

Response to Question 6. Suppose that the primal optimization problem is convex and has differentiable objective and constraint functions. We have to show that if there exists a solution to the KKT conditions, then *strong duality* holds for this optimization problem.

Proof. We consider the convex optimization problem given by

$$\begin{array}{lrcll} \min\limits_{w\in\mathbb{R}^d} & f(w) & & & \\ \text{s.t.} & f_i(w) & \leq & 0, & i=1,\dots,p \\ & g_i(w) & = & 0, & i=1,\dots,q, \end{array}$$

where f and each are convex and continuously differentiable, while each g_i is affine. Furthermore, we define the Lagrangian $\mathcal{L}\colon \mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ of the above problem to be

$$\mathcal{L}(w,\lambda,\nu) = f(w) + \sum_{i=1}^{p} \lambda_i f_i(w) + \sum_{i=1}^{q} \nu_i g_i(w),$$

where $\lambda \in \mathbb{R}^p, \nu \in \mathbb{R}^q$ are dual variables. The Lagrange dual function $\phi\colon \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ is

$$\phi(\lambda,\nu) = \inf_{w\in\mathbb{R}^d} \mathcal{L}(w,\lambda,\nu) = \inf_{w\in\mathbb{R}^d} \left(f(w) + \sum_{i=1}^{p} \lambda_i f_i(w) + \sum_{i=1}^{q} \nu_i g_i(w) \right).$$

Suppose we have $w^* \in \mathbb{R}^d$, $\lambda^* \in \mathbb{R}^p$ and $\nu^* \in \mathbb{R}^q$ satisfying the KKT conditions given by

$$\begin{aligned} f_i(w^*) &\leq 0, && i=1,\dots,p \\ g_i(w^*) &= 0, && i=1,\dots,q \\ \lambda^* &\succeq \mathbf{0} \\ \lambda_i^* f_i(w^*) &= 0, && i=1,\dots,p \\ \nabla_w f(w^*) + \sum_{i=1}^{p} \lambda_i \nabla_w f_i(w^*) + \sum_{i=1}^{q} \nu_i \nabla_w g_i(w^*) &= \mathbf{0}. \end{aligned}$$

Since f and f_i are convex and g_i affine, then $w \mapsto \mathcal{L}(w,\lambda^*,\nu^*)$ is convex for fixed (λ^*,ν^*). We know that a convex, differentiable function attains its global minimum at any point where its gradient vanishes. By stationarity, w^* is a global minimizer of $\mathcal{L}(\cdot,\lambda^*,\nu^*)$, so

$$\phi(\lambda^*,\nu^*) = \inf_{w\in\mathbb{R}^d} \mathcal{L}(w,\lambda^*,\nu^*) = \mathcal{L}(w^*,\lambda^*,\nu^*) = f(w^*) + \sum_{i=1}^{p} \lambda_i^* f_i(w^*) + \sum_{i=1}^{q} \nu_i^* g_i(w^*).$$

But $f_i(w^*) \leq 0$ for $i=1,\dots,p$, $g_i(w^*) = 0$ for $i=1,\dots,q$ and $\lambda_i^* f_i(w^*) = 0$ for $i=1,\dots,p$ (complementary slackness), so $\mathcal{L}(w^*,\lambda^*,\nu^*) = f(w^*)$ and $\phi(\lambda^*,\nu^*) = f(w^*)$. By definition,

$$d^* = \max_{\lambda\succeq 0,\nu} \phi(\lambda,\nu) \leq \min_{w\in\mathbb{R}^d} f(w) = p^*,$$

where d^* and p^* are the optimal primal and dual values, respectively. We have derived a dual-feasible (λ^*, ν^*), such that $\phi(\lambda^*, \nu^*) = f(w^*)$. Thus, we get that

$$d^* \geq \phi(\lambda^*, \nu^*) = f(w^*) \geq p^*,$$

since $d^* = \max_{\lambda \succeq 0, \nu} \phi(\lambda, \nu) \Rightarrow d^* \geq \phi(\lambda, \nu)$ and $f(w^*) \geq \min_{w \in \mathbb{R}^d} f(w) = p^*$ for a feasible solution w^*. Combining this with the trivial weak duality $d^* \leq p^*$, it gives us that

$$p^* = d^*.$$

Hence, the duality gap is zero. Equivalently, strong duality holds for this optimization problem as desired, and (w^*, λ^*, ν^*) is optimal for both the primal and dual problem. $\square$

4 SVM (25 points)

We consider a dataset $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ such that $x_i \in \mathbb{R}^m$ and $y_i \in \{-1, +1\}$ for all i. Here, we suppose that the points are linearly separable. Then, we consider the following alternative to the SVM problem formulation, which was shown in the lecture:

Let $w = \sum_{i=1}^n \lambda_i y_i x_i$, where $\lambda_i \in \mathbb{R}$ for all i. Instead of maximizing the margin of seperation, we aim to "sparsify" the representation of w by minimizing the number of non-zero λ_i's, approximately achieved by minimizing $\sum_{i=1}^n \lambda_i^2$. The resulting optimization problem is

$$\begin{aligned} \min_{\lambda_1, \ldots, \lambda_n, b \in \mathbb{R}} \quad & \frac{1}{2} \sum_{i=1}^n \lambda_i^2 \\ \text{s.t.} \quad & 1 - y_i \left(\sum_{j=1}^n \lambda_i y_j x_j^\top x_i + b \right) \leq 0 \qquad i = 1, \ldots, n \\ & \lambda_i \geq 0, \qquad i = 1, \ldots, n. \end{aligned} \tag{Sparse SVM}$$

Response to Question 7. We focus on the (Sparse SVM) optimization problem, but ignore the constraints $\lambda_i \geq 0$; that is, we now consider the modified problem given by

$$\begin{aligned} \min_{\lambda_1, \ldots, \lambda_n, b \in \mathbb{R}} \quad & \frac{1}{2} \sum_{i=1}^n \lambda_i^2 \\ \text{s.t.} \quad & 1 - y_i \left(\textstyle\sum_{j=1}^n \lambda_j y_j x_j^\top x_i + b \right) \leq 0 \qquad i = 1, \ldots, n. \end{aligned}$$

From the lecture slides, we know that the (Linear SVM) optimization problem is given by

$$\begin{aligned} \min_{w \in \mathbb{R}^m, b \in \mathbb{R}} \quad & \frac{1}{2} ||w||^2 \\ \text{s.t.} \quad & y_i \left(w^\top x_i + b \right) \geq 1 \qquad i = 1, \ldots, n. \end{aligned}$$

Now, we will show that (Sparse SVM) coincides with an instance of the original (Linear SVM) optimization problem: First, let an n-dimensional vector be given by

$$\tilde{w} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix} \in \mathbb{R}^n$$

Since the (Sparse SVM) objective is function $\frac{1}{2}\sum_{i=1}^{n} \lambda_i^2 = \frac{1}{2}||\tilde{w}||^2$, we can already see that

$$\min_{\lambda \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{2} \sum_{i=1}^{n} \lambda_i^2 \Leftrightarrow \min_{\tilde{w} \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{2}||\tilde{w}||^2.$$

For each data point $x_i \in \mathbb{R}^m$, we define a new feature vector $\phi(x_i) \in \mathbb{R}^n$ by

$$\phi(x) = \begin{pmatrix} y_1 x_1^\top x_i \\ y_2 x_2^\top x_i \\ \vdots \\ y_n x_n^\top x_i \end{pmatrix}$$

for $i = 1, \ldots, n$. Since the (Sparse SVM) constraint for each i is given by

$$1 - y_i \left(\sum_{j=1}^{n} \lambda_j y_j x_j^\top x_i + b \right) \leq 0,$$

and we notice that

$$\sum_{j=1}^{n} \lambda_j y_j x_j^\top x_i = \tilde{w}^\top \begin{pmatrix} y_1 x_1^\top x_i \\ \vdots \\ y_n x_n^\top x_i \end{pmatrix} = \tilde{w}^\top \phi(x_i),$$

the constraint then becomes

$$1 - y_i \left(\tilde{w}^\top \phi(x_i) + b \right) \leq 0 \Leftrightarrow y_i \left(\tilde{w}^\top \phi(x_i) + b \right) \geq 1,$$

which is exactly the form of a (Linear SVM) margin constraint in the feature space $\phi(\cdot)$. Putting objective and constraints together, the (Sparse SVM) without $\lambda_i \geq 0$ becomes

$$\begin{aligned} \min_{\tilde{w} \in \mathbb{R}^m, b \in \mathbb{R}} \quad & \frac{1}{2}||\tilde{w}||^2 \\ \text{s.t.} \quad & y_i \left(\tilde{w}^\top \phi(x_i) + b \right) \quad \geq \quad 1 \qquad i = 1, \ldots, n. \end{aligned}$$

But that is precisely the primal form of a (Linear SVM) trained on the new dataset $\tilde{S} = \{(\phi(x_i), y_i)\}_{i=1}^n$. Thus, by construction there is a one-to-one map

$$(\lambda_1, \ldots, \lambda_n, b) \longleftrightarrow (\tilde{w}, b)$$

with $x_i \mapsto \phi(x_i)$, and the objectives and feasible regions coincide. Hence, solving the (Sparse SVM) is identical to solving the ordinary (Linear SVM) in the feature space ϕ.

Response to Question 8. We have to derive the Lagrange dual problem of the (Sparse SVM) optimization problem. First, we define the Lagrangian of the problem to be

$$\begin{aligned}\mathcal{L}(\lambda, b, \alpha, \mu) &= \frac{1}{2}\sum_{i=1}^{n}\lambda_i^2 + \sum_{i=1}^{n}\alpha_i\left[1 - y_i\left(\sum_{j=1}^{n}\lambda_i y_j x_j^\top x_i + b\right)\right] + \sum_{i=1}^{n}\mu_i(-\lambda_i)\\ &= \frac{1}{2}\lambda^\top\lambda + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i y_i y_j \lambda_j x_j^\top x_i - b\sum_{i=1}^{n}\alpha_i y_i - \mu^\top\lambda,\end{aligned}$$

where $\alpha_i \geq 0$ is Lagrange multiplier (dual variable) for $1 - y_i(\sum_{j=1}^{n}\lambda_i y_j x_j^\top x_i + b) \leq 0$ and $\mu_i \geq 0$ the one for $\lambda_i \geq 0 \Leftrightarrow -\lambda_i \leq 0$. The Lagrange dual function is then defined to be

$$\phi(\lambda, b, \alpha, \mu) = \inf_{\lambda\in\mathbb{R}^n, b\in\mathbb{R}} \mathcal{L}(\lambda, b, \alpha, \mu).$$

To obtain the infimum with respect to λ and b, the first-order partial derivative of the Lagrangian $\mathcal{L}$ with respect to these variables must be 0. We get that

$$\frac{\partial\mathcal{L}}{\partial b} = -\sum_{i=1}^{n}\alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{n}\alpha_i y_i = 0$$

$$\frac{\partial\mathcal{L}}{\partial\lambda_k} = \lambda_k - \sum_{i=1}^{n}\alpha_i y_i y_k x_k^\top x_i - \mu_k = 0 \Rightarrow \lambda_k = \sum_{i=1}^{n}\alpha_i y_i y_k x_k^\top x_i + \mu_k, \quad k = 1, \ldots, n.$$

Thus at the stationary point, we set

$$\lambda_k^* = \sum_{i=1}^{n}\alpha_i y_i y_k x_k^\top x_i + \mu_k, \quad \sum_{i=1}^{n}\alpha_i y_i = 0.$$

At $\lambda = \lambda^*$ and $b = b^*$ (with $\sum_{i=1}^{n}\alpha_i y_i = 0$), the Lagrangian then becomes

$$\begin{aligned}\mathcal{L}(\lambda^*, b^*, \alpha, \mu) &= \frac{1}{2}\sum_{j=1}^{n}(\lambda_j^*)^2 + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i y_i y_j \lambda_j^* x_j^\top x_i - \sum_{j=1}^{n}\mu_j\lambda_j^*\\ &= \sum_{i=1}^{n}\alpha_i + \frac{1}{2}\sum_{j=1}^{n}(\lambda_j^*)^2 - \sum_{j=1}^{n}\left(\sum_{i=1}^{n}\alpha_i y_i y_j x_j^\top x_i\right)\lambda_j^* - \sum_{j=1}^{n}\mu_j\lambda_j^*\\ &= \sum_{i=1}^{n}\alpha_i + \frac{1}{2}\sum_{j=1}^{n}(\lambda_j^*)^2 - \sum_{j=1}^{n}(\lambda_j^*)^2\\ &= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{j=1}^{n}(\lambda_j^*)^2.\end{aligned}$$

Hence, the Lagrange dual function now is given by

$$\phi(\alpha, \mu) = \inf_{\lambda\in\mathbb{R}^n, b\in\mathbb{R}} \mathcal{L}(\lambda, b, \alpha, \mu) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{j=1}^{n}\left(\sum_{i=1}^{n}\alpha_i y_i y_j x_j^\top x_i + \mu_j\right)^2.$$

Since we must also enforce the feasibility of the dual multipliers $\alpha_i \geq 0$, $\mu_j \geq 0$, and $\sum_{i=1}^n \alpha_i y_i = 0$, the Lagrange dual problem of the (Sparse SVM) optimization problem is

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^n, \mu \in \mathbb{R}^n} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{j=1}^n \left(\sum_{i=1}^n \alpha_i y_i y_j x_j^\top x_i + \mu_j \right)^2 & \\ \text{s.t.} \quad & \alpha_i \geq 0 & i = 1, \ldots, n \\ & \mu_j \geq 0 & j = 1, \ldots, n \\ & \sum_{i=1}^n \alpha_i y_i = 0. & \end{aligned}$$

References

Matoušek, J. (2002). *Lectures on Discrete Geometry*. Springer New York, NY, 1st edition.