— *Solution notes* —

Seventh week practicals in Machine learning 1 – 2025 – Paper 1

1 Deriving the stationary and KKT conditions. (October)

The purpose of this exercise is to show the reasoning behind the Lagrangian and KKT conditions from form the method for solving inequality-constrained optimization problems introduced in the lectures.

Consider the constrained optimization problem:

$$\max_{\mathbf{x}\in\mathbb{R}^2} f(\mathbf{x}) \quad \text{subject to} \quad g(\mathbf{x}) \geq 0.$$

We aim to find candidate points for a global maximum. The restriction $g(\mathbf{x}) \geq 0$ gives us the **primal feasibility** KKT condition. Restricting the domain to this region gives us two types of candidate points: those on the boundary $g(\mathbf{x}) = 0$ (Case I) and those in the region $g(\mathbf{x}) > 0$ (Case II).

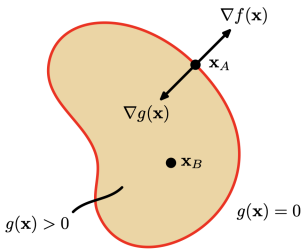
(a) (Case I) If there is a critical point $\mathbf{x}_0$ on the boundary, it means that an unconstrained maximum would occur either on the boundary or outside the feasible region, and the boundary restricts us from reaching that point.

(i) In which direction does $\nabla_{\mathbf{x}} f(\mathbf{x}_0)$ point in this case?

(ii) What about $\nabla_{\mathbf{x}} g(\mathbf{x}_0)$?

(iii) How are the directions of both gradients related? Can you express $\nabla_{\mathbf{x}} f(\mathbf{x}_0)$ in terms of $\nabla_{\mathbf{x}} g(\mathbf{x}_0)$ using a scalar parameter μ?

Answer:



(i) $\nabla_{\mathbf{x}} f(\mathbf{x}_0)$ points in the direction of the greatest increase of f from $\mathbf{x}_0$, which is perpendicular to the level curve $g(\mathbf{x}) = 0$, pointing outside the region $g(\mathbf{x}_0) > 0$.

The reason why it is perpendicular to the level curve is as follows: if the direction of maximum increase was not perpendicular to the level curve, then the gradient would have a nonzero component along the curve. This would mean we could move along the curve to increase the value of f, which contradicts the fact that $\mathbf{x}_0$ is a maximum.

(ii) $\nabla_{\mathbf{x}} g(\mathbf{x}_0)$ is perpendicular to the level curve $g(\mathbf{x}) = 0$, pointing towards the region where $g(\mathbf{x}) > 0$.

The fact that it is perpendicular to the level curve follows directly from the definition of a level curve. Using a similar reasoning as before, if the gradient were not perpendicular, one could increase the value of g by moving along the curve, which is impossible since g is constant (0) on the curve.

(*iii*) The two gradients are anti-parallel, that is:

$$\nabla_{\mathbf{x}} f(\mathbf{x}_0) + \mu \nabla_{\mathbf{x}} g(\mathbf{x}_0) = \mathbf{0}, \quad \mu \geq 0.$$

(*b*) (Case II) If there is a critical point $\mathbf{x}_0$ inside the feasible region, then the inequality constraint does not restrict the location of the maximum of f.

(*i*) What is the value of $\nabla_{\mathbf{x}} f(\mathbf{x}_0)$ in this case?

(*ii*) Can you express this condition using a similar equation as in (a.iii)?

Answer:

(*i*) $\nabla_{\mathbf{x}} f(\mathbf{x}_0) = \mathbf{0}$.

(*ii*) This can be written as:

$$\nabla_{\mathbf{x}} f(\mathbf{x}_0) + \mu \nabla_{\mathbf{x}} g(\mathbf{x}_0) = \mathbf{0}, \quad \mu = 0.$$

(*c*) Combine your results from (a.iii) and (b.ii) into a single condition that all candidate points must satisfy. Congratulations! You have derived the **stationarity condition** for the Lagrangian and the **dual feasibility** KKT condition.

Answer:

$$\nabla_{\mathbf{x}} f(\mathbf{x}_0) + \mu \nabla_{\mathbf{x}} g(\mathbf{x}_0) = \mathbf{0}, \quad \mu \geq 0.$$

(*d*) Use your findings so far to derive the **complementary slackness** KKT condition.

Answer: For a candidate point $\mathbf{x}_0$, either it lies on the boundary $g(\mathbf{x}_0) = 0$ (Case I), or it lies inside the feasible region, where $\mu = 0$ (Case II). Therefore, at least one of $g(\mathbf{x}_0)$ and μ must be zero, leading to the **complementary slackness** condition:

$$\mu \, g(\mathbf{x}_0) = 0.$$

(*e*) How would we define the Lagrangian for the following optimization problems? *Hint: Think about whether* $\nabla_{\mathbf{x}} f$ *and* $\nabla_{\mathbf{x}} g$ *would be paralell or anti-paralell in each case.*

(*i*) $\min_{\mathbf{x}\in\mathbb{R}^2} f(\mathbf{x})$ subject to $g(\mathbf{x}) \geq 0$.

(*ii*) $\max_{\mathbf{x}\in\mathbb{R}^2} f(\mathbf{x})$ subject to $g(\mathbf{x}) \leq 0$.

(*iii*) $\min_{\mathbf{x}\in\mathbb{R}^2} f(\mathbf{x})$ subject to $g(\mathbf{x}) \leq 0$.

Answer:

(*i*) $\mathcal{L}(\mathbf{x}, \mu) = f(\mathbf{x}) - \mu g(\mathbf{x})$ (we are now interested in the direction of maximum **decrease** of f, which is $-\nabla_{\mathbf{x}} f$, which is anti-paralell to $\nabla_{\mathbf{x}} g$).

(*ii*) $\mathcal{L}(\mathbf{x}, \mu) = f(\mathbf{x}) - \mu g(\mathbf{x})$ (now $\nabla_{\mathbf{x}} g$ points in the opposite direction as in the original problem).

(*iii*) $\mathcal{L}(\mathbf{x}, \mu) = f(\mathbf{x}) + \mu g(\mathbf{x})$ ($\nabla_{\mathbf{x}} g$ points in the opposite direction as in the original problem, but we consider $-\nabla_{\mathbf{x}} f$, so the gradients are still anti-paralell).

Seventh week practicals in Machine Learning 1 – 2025 – Paper 1

2 Inequality-constrained optimization (October)

Consider the following 1D optimization problem:

$$\min_{x\in\mathbb{R}} \frac{1}{2}(x-2)^2 \quad \text{subject to} \quad x \geq 4.$$

Of course, this is a very simple problem that can be solved just by evaluating the function in the critical point of the parabola ($x = 2$) and in the extreme of the domain ($x = 4$). However, we will take the long way around to further demystify the method of Lagrange multipliers.

(a) Formulate the Lagrangian of the problem using a Lagrange multiplier $\mu \geq 0$ for the inequality constraint.

Answer: The Lagrangian is:

$$\mathcal{L}(x,\mu) = \frac{1}{2}(x-2)^2 - \mu(x-4), \quad \mu \geq 0.$$

Note that, as this is a minimization problem, the term with the Lagrangian multiplier is negative, as $-\nabla_{\mathbf{x}} f$ and $\nabla_{\mathbf{x}} g$ will be parallel (pointing in the same direction) at the potential minimums located in the boundary.

(b) Derive the stationarity condition and express the primal variable x in terms of the dual variable μ.

Answer: The derivative of the Lagrangian w.r.t. x is

$$\frac{\partial \mathcal{L}}{\partial x} = x - 2 - \mu.$$

Setting it to zero, we obtain the stationary condition

$$x = 2 + \mu.$$

(c) Write down the complementary slackness condition.

Answer:

$$\mu(x-4) = 0.$$

(d) Construct the Dual Lagrangian $\tilde{\mathcal{L}}(\mu) = \min_x \mathcal{L}(x,\mu)$ and formulate the dual problem.

Answer: We substitute the stationary condition $x = 2 + \mu$ into the Lagrangian:

$$\tilde{\mathcal{L}}(\mu) = \frac{1}{2}(\mu)^2 - \mu(\mu-2) = -\frac{1}{2}\mu^2 + 2\mu.$$

The dual problem is

$$\max_{\mu \geq 0} \tilde{\mathcal{L}}(\mu).$$

(*e*) Solve the dual problem to find μ^* and then compute the corresponding primal optimum x^*. Is the value located in the boundary or inside the constrained region?

Answer:

$$\frac{d\tilde{\mathcal{L}}}{d\mu} = -\mu + 2; \quad \frac{d\tilde{\mathcal{L}}}{d\mu} = 0 \Leftrightarrow \mu = 2,$$

Since the Dual Lagrangian is concave, the dual maximum occurs at $\mu^* = 2$.

Substituting μ^* in the stationary condition, we obtain

$$x^* = 2 + \mu^* = 4.$$

The optimal value is in the boundary $x - 4 = 0$.

(*f*) Verify that the obtained solution meets the KKT conditions.

Answer:

- Primal feasibility: $x^* - 4 = 0 \geq 0$.
- Dual feasibility: $\mu^* = 2 \geq 0$.
- Complementary slackness: $\mu^*(x^* - 4) = 2(4 - 4) = 0$.

Sixth week practicals in Machine learning 1 – 2025 – Paper 1

3 Kernel outlier Detection (October)

Consider the picture in Figure 2. The dots represent data items. Our task is to derive an algorithm that will detect the outliers (in this example there are 2 of them). To that end, we draw a circle rooted at location $\boldsymbol{a}$and with radius R. All data-cases that fall outside the circle are detected as outliers

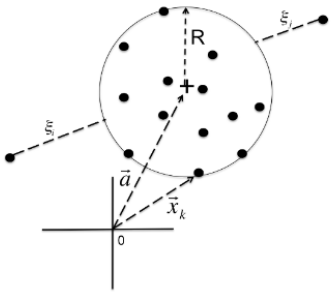


Figure 2: Kernel Outlier Detection

We will now write down the primal program that will find such a circle:

$$\min_{\boldsymbol{a},R,\boldsymbol{\xi}} R^2 + C\sum_{i=1}^{N}\xi_i$$

$$s.t.\ \forall i : \|\boldsymbol{x_i} - \boldsymbol{a}\|^2 \leq R^2 + \xi_i,\ \xi_i \geq 0$$

In words: we want to minimize the radius of the circle subject to the constraint that most data cases should lay inside it. Outliers are allowed to stay outside, but they pay a price proportional to their distance from the circle boundary and C. Answer the following questions:

(a) Introduce Lagrange multipliers for the constraints and write down the primal Lagrangian. Use the following notation: $\{\alpha_i\}$are the Lagrange multipliers for the first constraint and $\{\mu_i\}$for the second constraint.

Answer:

$$\mathcal{L}(\boldsymbol{a}, R, \xi, \alpha, \mu) = R^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left(R^2 + \xi_i - \|\boldsymbol{x}_i - \boldsymbol{a}\|^2\right) - \sum_{i=1}^{N}\mu_i\xi_i.$$

(b) Write down all stationary and KKT conditions. (Hint: take the derivative w.r.t. R^2instead of R).

Answer: The first three conditions are obtained by setting the derivative of the primal Lagrangian to zero with respect to R^2, $a$and ξ_i.

$$\frac{\partial \mathcal{L}}{\partial R^2} = 1 - \sum_{i=1}^{N} \alpha_i$$

$$\frac{\partial \mathcal{L}}{\partial R^2} = 0 \implies \sum_{i=1}^{N} \alpha_i = 1.$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}} = -2 \sum_{i=1}^{N} \alpha_i (\boldsymbol{x}_i - \boldsymbol{a})^T$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{a}} = 0 \implies \sum_{i=1}^{N} \alpha_i \boldsymbol{x}_i = \boldsymbol{a} \sum_{i=1}^{N} \alpha_i \implies \boldsymbol{a} = \sum_{i=1}^{N} \alpha_i \boldsymbol{x}_i$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \mu_i$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \implies C - \alpha_i - \mu_i = 0, \forall i$$

Note that the three conditions above are **not** KKT conditions. The KKT conditions for this model are given below:

- $R^2 + \xi_i - \|\boldsymbol{x}_i - \boldsymbol{a}\|^2 \geq 0 \forall i.$
- $\xi_i \geq 0 \forall i.$
- $\alpha_i \geq 0 \forall i.$
- $\mu_i \geq 0 \forall i.$
- $\alpha_i \left(R^2 + \xi_i - \|\boldsymbol{x}_i - \boldsymbol{a}\|^2\right) = 0 \forall i.$
- $\mu_i \xi_i = 0 \forall i.$

(*c*) Use these conditions to derive which data-cases $\boldsymbol{x_i}$will have $\alpha_i > 0$and which ones will have $\mu_i > 0$.

Answer: The complementary slackness conditions are:

Inside the ball: $\xi_i = 0,\ R^2 + \xi_i - \|\boldsymbol{x}_i - \boldsymbol{a}\|^2 > 0 \implies \alpha_i = 0 \implies \mu_i = C > 0$

Outside the ball: $\xi_i > 0,\ R^2 + \xi_i - \|\boldsymbol{x}_i - \boldsymbol{a}\|^2 = 0 \implies \mu_i = 0 \implies \alpha_i = C > 0$

On the ball: $\xi_i = 0,\ R^2 + \xi_i - \|\boldsymbol{x}_i - \boldsymbol{a}\|^2 = 0 \implies \mu_i \geq 0,\ \alpha_i \geq 0,\ \mu_i + \alpha_i = C$

(*d*) Derive the dual Lagrangian and specify the dual optimization problem. Kernelize the problem, i.e. write the dual program only in terms of kernel entries and Lagrange multipliers.

Answer: We use conditions a, b, and c to eliminate R^2, a, and $\xi_i$from the primal Lagrangian to obtain the dual representation.

$$\mathcal{L}(\mathbf{a}, R, \xi, \alpha, \mu) = R^2 + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i \left(R^2 + \xi_i - \|\boldsymbol{x}_i - \boldsymbol{a}\|^2\right) - \sum_{i=1}^{N} \mu_i \xi_i$$

$$= R^2 + C\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\alpha_i \boldsymbol{x}_i^T\boldsymbol{x}_i - 2\sum_{i=1}^{N}\alpha_i \boldsymbol{x}_i^T\boldsymbol{a} + \sum_{i=1}^{N}\alpha_i \boldsymbol{a}^T\boldsymbol{a}$$

$$- \sum_{i=1}^{N}\alpha_i R^2 - \sum_{i=1}^{N}\alpha_i\xi_i - \sum_{i=1}^{N}\mu_i\xi_i$$

$$= \left(R^2 - \sum_{i=1}^{N}\alpha_i R^2\right) + \sum_{i=1}^{N}(C - \alpha_i - \mu_i)\xi_i$$

$$-2\left(\sum_{i=1}^{N}\alpha_i \boldsymbol{x}_i^T\right)\boldsymbol{a} + \left(\sum_{i=1}^{N}\alpha_i\right)\boldsymbol{a}^T\boldsymbol{a} + \sum_{i=1}^{N}\alpha_i \boldsymbol{x}_i^T\boldsymbol{x}_i$$

$$= \sum_{i=1}^{N}\alpha_i \boldsymbol{x}_i^T\boldsymbol{x}_i - \boldsymbol{a}^T\boldsymbol{a}$$

$$= \sum_{i=1}\alpha_i\|\boldsymbol{x}_i\|^2 - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j \boldsymbol{x}_i^T\boldsymbol{x}_j$$

In the first step, we expand the expression for the primal Lagrangian, in the second step we rearrange the terms, finally, we apply the conditions. Kernelize the problem:

$$\sum_{i=1}^{N}\alpha_i \boldsymbol{x}_i^T\boldsymbol{x}_i - \boldsymbol{a}^T\boldsymbol{a} = \sum_{i=1}\alpha_i\|\boldsymbol{x}_i\|^2 - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j \boldsymbol{x}_i^T\boldsymbol{x}_j$$

$$= \sum_{i=1}^{N}\alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}_i) - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}_j)\alpha_j$$

The dual program is:

$$argmax_\alpha \sum_{i=1}^{N}\alpha_i K_{ii} - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j K_{ij}$$

$$\text{with } \alpha_i \in [0, C]\ \forall i.$$

(*e*) The dual program will return optimal values for $\{\alpha_i\}$. Assume that at least one of these is such that $0 < \alpha_i < C$. In terms of the optimal values for α_i, compute the optimal values for the other dual variables $\{\mu_i\}$.

Then, solve the primal variables $\{\boldsymbol{a}, R, \boldsymbol{\xi}\}$(in that order) in terms of the dual variables $\{\mu_i, \alpha_i\}$. Note that you do not need to know the dual optimization program to solve this question.

Answer:

Note that when $0 < \alpha_i < C$, then $\boldsymbol{x}_i$must be on the ball. When $\alpha_i = 0$, then $\boldsymbol{x}_i$is on or inside the ball, and when $\alpha_i = C$then $\boldsymbol{x}_i$is on or outside the ball.

$$\mu_i^* = C - \alpha_i^*$$

$$\boldsymbol{a}^* = \sum_{i=1}^{N} \alpha_i^* \boldsymbol{x}_i$$

$$R^{*2} = \|\boldsymbol{x}_i - \boldsymbol{a}^*\|^2 \text{ (when } 0 < \alpha_i^* < C)$$

$$\xi_i^* = \begin{cases} \|\boldsymbol{x}_i - \boldsymbol{a}^*\|^2 - R^{*2} & \text{if } \alpha_i^* = C \\ 0 & \text{if } \alpha_i^* \in [0, C) \end{cases}$$

(f) Assume we have solved the dual program. We now want to apply it to new test cases. Describe a test in the dual space (i.e. in terms if kernels and Lagrange multipliers) that could serve to detect outliers. (Students who got stuck along the way may describe the test in primal space).

Answer: A new test case $\boldsymbol{x}_t$ is an outlier when:

$$\|\boldsymbol{x}_t - \boldsymbol{a}^*\|^2 > R^{*2}$$

$$\boldsymbol{x}_t^T \boldsymbol{x}_t - 2\boldsymbol{x}_t^T \boldsymbol{a}^* + \boldsymbol{a}^{*T} \boldsymbol{a}^* > R^{*2}$$

$$K(\boldsymbol{x}_t, \boldsymbol{x}_t) - 2\sum_{i=1}^{N} \alpha_i^* \boldsymbol{x}_i^T \boldsymbol{x}_t + \sum_{i,j} \alpha_i \alpha_j \boldsymbol{x}_i^T \boldsymbol{x}_j > R^{*2}$$

$$K(\boldsymbol{x}_t, \boldsymbol{x}_t) - 2\sum_{i=1}^{N} \alpha_i^* K(\boldsymbol{x}_i, \boldsymbol{x}_t) + \sum_{i,j} \alpha_i \alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) > R^{*2}$$

(g) What kind of solution do you expect if we use $C = 0$. And what solution if we use $C = \infty$?

Answer: When $C \to 0$ we expect $R \to 0$, and when $C \to \infty$ we expect R to be such that all data-cases are inside the ball.
