AI511-ML: Constrained Optimization*

Aniruddh Kishore Budhgavi
aniruddh.kishore@iiitb.ac.in

13 October 2022

1 Definitions

1.1 Convex functions

Let us consider a function $f(x), f : D_1 \rightarrow D_2$ where x may be a vector. f is a convex function if $\forall x_1, x_2 \in D_1$, $\forall \alpha \in [0, 1]$, we have:

$$\alpha f(x_1) + (1 - \alpha) f(x_2) \geq f(\alpha x_1 + (1 - \alpha) x_2)$$

Or, the line joining $f(x_1)$ and $f(x_2)$ lies above the function curve between x_1 and x_2. A visual representation of the same is given below:

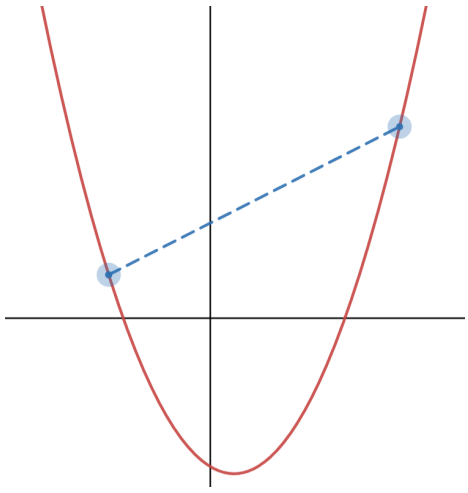

Figure 1: Visualization of convexity

A key property of convex functions is that **there is only one local minimum – which is the global minimum.**

1.2 The general convex optimization problem

Minimize the convex function $f_0(x)$[1]
Subject to the convex inequality constraints $f_i(x) \leq 0, i \in \{1, 2, ..., m\}$
And the affine equality constraints $h_j(x) = 0, j \in \{1, 2, ..., n\}$

An affine function is a function of the form $h(x) = a^T x + b$.

We will explore the main concepts of this below.

1.3 The unconstrained case

Minimize the convex function $f_0(x)$

This is pretty straightforward. We can use gradient descent (or its many variations, like momentum, RMSprop, Adam), or, if we can analytically solve the below closed-form equation, that works too. Just find x^* such that

$$\nabla_x f(x)\big|_{x=x^*} = 0$$

2 Equality constraints

2.1 A single equality constraint

Given in the below plot is a line equality constraint $h(x, y) = 0$ and the **level curves** of the function $f(x, y)$ which we wish to minimize. Each level curve represents the set of points in the domain where f has the same value i.e. each level curve represents the solution to $f(x, y) = k$ for a different value of k.

[1]Henceforth, the notation used here shall be consistent with *Boyd and Vandenberghe: Convex Optimization*.

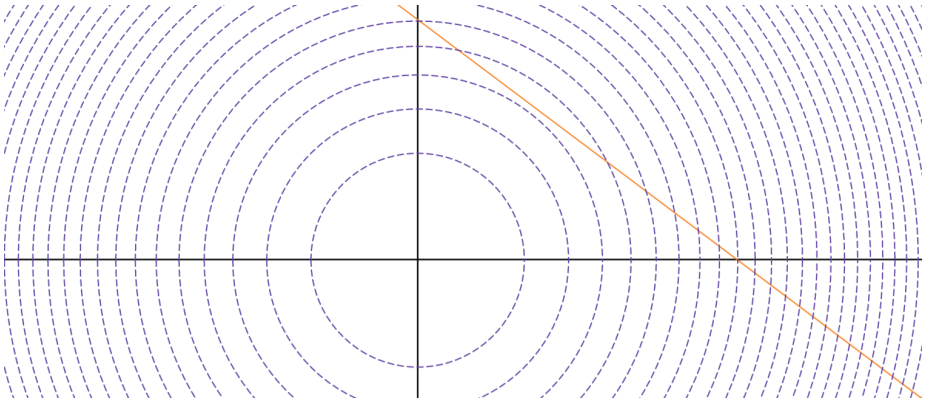

Figure 2: Level curves and a single equality constraint

The function to minimize is $f(x, y) = x^2 + y^2$. The *unconstrained* optimum is at $(0, 0)$, but with a constraint...

Observe that when the function is minimized while maintaining the constraint, the level curve is tangent to the constraint line. In other words, at the optimum (x^*, y^*) , the normal to the level curve is parallel to the normal to the constraint.

Note: The normal to the level curve (or lower/higher dim equivalent) of a function is given by $\nabla_x f(x)$. This does not contradict the fact that the direction of steepest ascent of the surface of $z = f(x)$ is given by $\nabla_x f(x)$

Therefore,

$$\nabla_{x,y} f(x, y)\big|_{x^*,y^*} = \lambda \nabla_{x,y} h(x, y)\big|_{x^*,y^*}$$

This, combined with the constraint conditions, gives us the desired point (x^*, y^*).

2.2 Multiple equality constraints

Now, observe this figure below. We wish to minimize $f(x, y, z) = x^2 + y^2 + z^2$ subject to two equality constraints given in the form of planes.

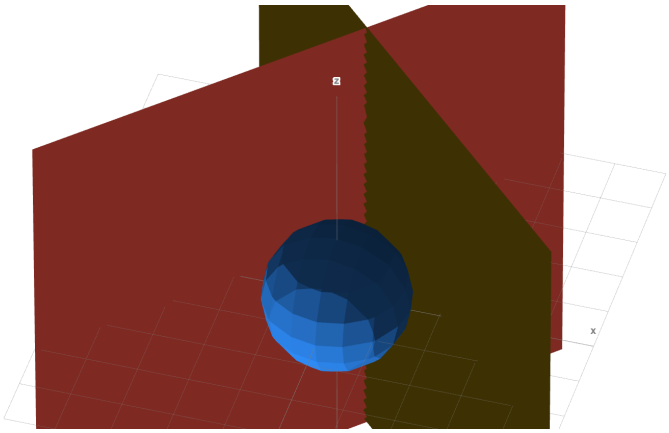

Figure 3: Multiple equality constraints

Our aim is to minimize the radius of that blue sphere, while ensuring that there exists a point on the sphere that is there on **both** the planes. Pay attention now – **one point** that is simultaneously on **both planes** – not one point *each* on each plane.
Let's choose a less confusing perspective for this:

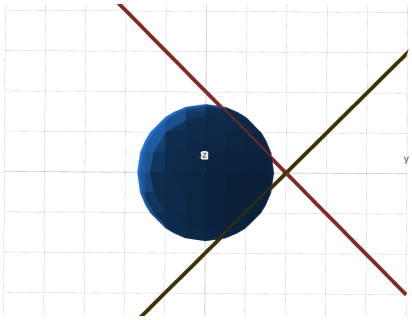

Figure 4: Multiple equality constraints – top view

From this view, we see that at the optimal (x^*, y^*, z^*), f needs to be tangent to the **intersection** of h_1 and h_2 (the line defined by the intersection of the two planes).

In other words, the normal to f at the optimal point, is a linear combinations of the normals to h_1 and h_2. Mathematically,

$$\nabla_{x,y,z} f(x,y,z)\big|_{x^*,y^*,z^*} = \lambda_1 \nabla_{x,y,z} h_1(x,y,z)\big|_{x^*,y^*,z^*} + \lambda_2 \nabla_{x,y,z} h_2(x,y,z)\big|_{x^*,y^*,z^*}$$

while also satisfying the constraints.

2.3 The Lagrangian

We can combine the expressions for the objective function f_0 and the affine equality constraints h_i to obtain a single expression. The expression is:

$$\mathcal{L}(x, \lambda) = f_0(x) + \sum_{i=1}^{n} \lambda_i h_i(x)$$

$\mathcal{L}$ is called the **Lagrangian**, (yes, that one, from the Physics class in which we were all sleeping or mute-ending[2]) and the λ_is are called the **Lagrange Multipliers**. For now, these multipliers are just a curiosity, but they're significant when we come to inequality constraints (and in SVM).

We can find the optimal value (x^*) by finding the **critical points** (also sometimes referred to as **stationary points**) of the Lagrangian. That is, solve

$$\nabla_{x,\lambda} \mathcal{L}(x, \lambda)\big|_{x^*, \lambda^*} = 0$$

3 Introducing Inequality Constraints

We now come to the full constrained convex optimization problem: Minimize the convex function $f_0(x)$
Subject to the convex inequality constraints $f_i(x) \leq 0, i \in \{1, 2, ..., m\}$
And the affine equality constraints $h_j(x) = 0, j \in \{1, 2, ..., n\}$

A slight point on notation: We denote $f_0(x^*)$, the constrained optimal value of f_0, to be p^* (p meaning primal). We denote the set of x for which f_0 is defined and all the constraints are satisfied to be the **feasible region**, $\mathcal{D}$. $x \in \mathcal{D}$ implies that x satisfies the constraints and is in the domain of f_0.

3.1 The Lagrangian in the General Case

$$\mathcal{L}(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{n} \nu_j h_j(x)$$

3.2 The Lagrange Dual Function and Weak Duality

The Lagrange Dual Function is defined as

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} \mathcal{L}(x, \lambda, \nu)$$

[2]Historical reference: This document was prepared in one of the semesters shortly after the COVID-19 pandemic stabilized in India. It was not uncommon for students to just log in to the classroom videoconference meeting, put that tab/process on mute and then go do something else. Hence the term, mute-ending

Now, suppose we constrain ourselves to $\lambda_i \geq 0 \; \forall \; i > 0$[3] .
Under this constraint, g is unbounded below (fiddle with $\mathcal{L}$ to see why).

Then, since $f_i(x) \leq 0$, we have $\lambda_i . f_i(x) \leq 0 \; \forall \; i > 0$ which implies that

$$\mathcal{L}(x, \lambda, \nu) \leq f_0(x), \; \forall x \in \mathcal{D}, \; \lambda \geq 0$$

Further, since g is the infimum of $\mathcal{L}$ over x,

$$g(\lambda, \nu) \leq \mathcal{L}(x, \lambda, \nu) \forall x \in \mathcal{D}, \; \lambda \geq 0$$

This implies that

$$g(\lambda, \nu) \leq f_0(x) \forall x \in \mathcal{D}, \; \lambda \geq 0$$

This is not just a minor detail of some sort. If you continue down this train of thought, recalling that $f_0(x^*) = p^*$, the constrained optimal value,

$$g(\lambda, \nu) \leq p^*, \; \lambda \geq 0$$

This is not all! If we try to maximize g and let $d^* = g(\lambda^*, \nu^*)$, the optimal value of g over the constraint $\lambda \geq 0$, then

$$d^* \leq p^*$$

In other words, **the dual optimal is less than or equal to the primal optimal**. This result is known as **weak duality**.

Note: For convex problems, the Lagrange Dual Function is **concave**. *Boyd et. al.* provide a mathematical explanation for this, but I think exploring **this** visualization will help you understand better. The visualization has a convex function to be minimized and two inequality constraints with their multipliers given in sliders as a and b. The inequality constraints are the red and blue planes, and the feasible region is the region cut by these planes that is closest to the camera. Adjust the values of a and b and see the change in the graph for the Lagrangian for those values of λ wrt x, y. In particular, see what happens to the minimum of the Lagrangian within the feasible region.

3.3 Strong Duality and the KKT Conditions

When would d^* be equal to p^* ? This is an interesting question.

$$d^* = \mathcal{L}(x_{d^*}, \lambda^*, \nu^*) \leq \mathcal{L}(x^*, \lambda^*, \nu^*) \leq f_0(x^*) = p^*$$

Now, if $d^* = p^*$,

$$p^* = d^* \leq \mathcal{L}(x^*, \lambda^*, \nu^*) \leq p^* = d*$$

Which implies that

$$p^* = \mathcal{L}(x^*, \lambda^*, \nu^*)$$

[3]This is a bit wordy. I'll just write $\lambda \geq 0$ to mean the above henceforth.

Expanding, we get

$$f_0(x^*) = f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{j=1}^{n} \nu_j^* h_j(x^*)$$

The terms involving ν evaluate to 0 anyway (equality constraints). If we eliminate $f_0(x^*)$, we get (flipping sides)

$$\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0$$

Now, since $\lambda \geq 0$ and $f_i(x) \leq 0 \; \forall \; i > 0$,

$$\lambda_i^* f_i(x^*) \leq 0 \; \forall i > 0$$

We conclude therefore that

$$\lambda_i^* f_i(x^*) = 0 \; \forall i > 0$$

The above property isn't insignificant. It's known as **complementary slackness** and will feature prominently in SVMs.

We've proven that this set of conditions is a necessary criterion. We can go in the reverse direction and also prove that for this case of convex optimization, it's a sufficient criterion. The complete set of all the criteria needed is summarized in the below section.

3.4 The Karush–Kuhn–Tucker (KKT) Conditions

Given the convex function $f_0(x)$
And the convex inequality constraints $f_i(x) \leq 0, i \in \{1, 2, ..., m\}$
Affine equality constraints $h_j(x) = 0, j \in \{1, 2, ..., n\}$

And given the Lagrangian $\mathcal{L}(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{n} \nu_j h_j(x)$

And given the Lagrange Dual Function $g(\lambda, \nu) = \inf_{x \in \mathcal{D}} \mathcal{L}(x, \lambda, \nu)$,

If $f_0(x^*) = p^*$

And if $g(\lambda^*, \nu^*) = d^*$

1. $f_i(x^*) \leq 0 \; \forall \; i \in \{1, 2, ..., m\}$ (Primal Feasibility, 1)
2. $h_j(x^*) = 0 \; \forall \; i \in \{1, 2, ..., n\}$ (Primal Feasibility, 2)
3. $\lambda^* \geq 0$ (Dual Feasibility)
4. $\lambda_i^* . f_i(x^*) = 0 \; \forall \; i \in \{1, 2, ..., n\}$ (Complementary Slackness)
5. $\nabla_x \mathcal{L}(x, \lambda, \nu)\big|_{x^*, \lambda^*, \nu^*} = 0$ (Stationarity)

Then, $p^* = d^*$ and they are the optimal values for the primal and dual problem.

3.5 Using the KKT conditions in practice

This is all well and good, but when it comes to solving problems, there's a long way to go.

One hint: Look at the **Complementary Slackness** condition.

- That condition means that for each inequality constraint, either the multiplier OR the inequality constraint have to evaluate to zero.
- If f_i is nonzero, then it basically means that the solution is inside the region of the inequality, thus meaning that the boundary need not be enforced. Further, the corresponding λ_i needs to equal zero. This term is eliminated from the Lagrangian.
- If f_i is zero, then the solution is at the boundary of this inequality, therefore λ_i need not be zero. However, since we know that the solution is at the boundary of this, we can replace the inequality constraint with an equality constraint.
- Performing this process for every i, we are left with only equality constraints, and this problem returns to the previous case and we solve accordingly.
- But how do we know which inequalities are "slack" and which ones are "taut" ? Ans: We don't. We have to take every case (i.e every subset) and give it a shot. There is no other easy *analytical* way. But note that this is a process that is at least $O(2^n)$ in terms of the number of inequalities.

4 Problems for TA to demonstrate

- Minimize $f_0(x, y, z) = x^2 + y^2 + z^2$ subject to the constraints $x + y = 3$ and $x - y = 3$.
- Minimize $f_0(x, y) = x^2 + y^2$ subject to the constraints $x + y - 1 \leq 0$ and $x - y + 2 \leq 0$.