Lagrange multipliers

Gatsby induciton week
Roman Pogodin

September 26, 2018

Preliminaries

This section is a very short version of the great book [1] (mainly its first part). The goal is to give a formal view on constraint optimisation. It's not necessary for understanding the material, but should clarify a lot of small intermediate steps that will appear in the TN/ML/kernels courses.

We'd like to solve problems of the form

$$\begin{aligned} \min \quad & f(x), \\ \text{s.t.} \quad & f_i(x) \leq 0, \ i = 1, \dots, m \\ & h_i(x) = 0, \ i = 1, \dots, p. \end{aligned} \tag{1}$$

Denote the intersection of their domains as

$$\mathcal{D} = \text{dom} f \cap (\cap \text{dom} f_i) \cap (\cap \text{dom} h_i).$$

One can interpret the constraints f_i and h_i as penalties, obtaining the *Lagrangian*:

$$L(x, \lambda, \nu) = f(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x), \tag{2}$$

where λ and ν are so-called *Lagrange multipliers* (or dual variables). Why is it useful? Denote the *Lagrange dual function* as

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu). \tag{3}$$

We have that for every $\lambda \geq 0$, every ν and every *feasible* x (such that the constraints in Equation 1 are satisfied) the following is **always** true:

$$g(\lambda, \nu) \leq L(x, \lambda, \nu) \leq f(x). \tag{4}$$

The left inequality is trivial; the right one is true as $h_i(x) = 0$ and $\lambda_i f_i(x) \leq 0$ for any feasible point.

We haven't yet said that any of the used functions is "good", although we used the fact that $\mathcal{D} \neq \varnothing$. In practice, it means that if one can find $g(\lambda, \nu)$, maximising it would give a lower bound on $f(x)$. Moreover, $g(\lambda, \nu)$ is always a concave function (we'll define it soon; the proof uses only the definitions of concavity and the dual function), which means that there is only one global maximum.

The question is how one can find all three variables x^*, λ^*, ν^* such that $g(\lambda^*, \nu^*) = f(x^*)$. First, we need to define "good" functions.

Convex functions. A function $f(x),\ x \in \mathbb{R}^n$ is *convex* if

$$\alpha f(x) + (1-\alpha) f(y) \geq f(\alpha x + (1-\alpha y)),\ \alpha \in [0,1],\ x, y \in \mathbb{R}^n.$$

An example is drawn in Figure 1. A function $f(x)$ is called *concave* if $-f(x)$ is convex.

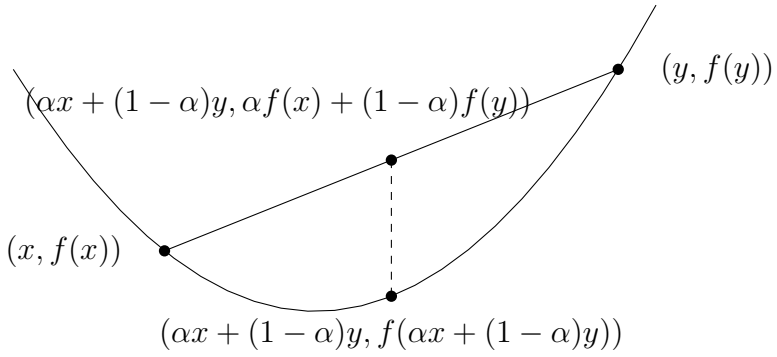

Figure 1: Graph of a convex function

Convex problems. Problems of the form Equation 1 for which $f(x), f_i(x)$ are convex and $h_i(x)$ are affine (forming a constraint $Ax = b$) are called convex. In other words, we take an *objective function* $f(x)$ that is convex on a convex set and minimise it on this set (recall that a set is called convex if the edge between any two points of the set belongs to it as well).

A problem is called concave, if $f(x)$ is concave (and the rest is convex).

Slater's condition. Denote $p^* = \min f(x),\ d^* = \max g(\lambda, \nu)$ (letters stand for primal and dual, optimisation is done w.r.t. the mentioned constraints). Consider problems of the form

$$\begin{aligned} \min \quad & f(x), \\ \text{s.t.} \quad & f_i(x) \leq 0,\ i = 1, \dots, m \\ & Ax = b. \end{aligned} \tag{5}$$

with convex f, f_i. The *Slater's condition* states that if $\exists x \in \operatorname{relint} \mathcal{D}$ for which all $f_i < 0$, then $d^* = p^*$. Relative interior of a set is defined as

$$\operatorname{relint} \mathcal{D} = \{x \in \mathcal{D} | B(x, r) \cap \operatorname{aff} \mathcal{D} \subseteq \mathcal{D} \text{ for some } r > 0\},$$

where $B(x, r)$ is an x-centred ball of radius r and $\operatorname{aff} \mathcal{D}$ is the affine hull of $\mathcal{D}$.

This condition seems to be trivial, but it doesn't always hold (e.g. [1], p. 280, example 5.21).

Karush-Kuhn-Tucker (KKT) conditions. If x^* and λ^*, ν^* are the solutions of the primal and dual problems and also $d^* = p^*$, then the KKT conditions for differentiable f, f_i, h_i are

$$\begin{aligned} \frac{\partial}{\partial x} L(x^*, \lambda^*, \nu^*) &= 0, \\ f_i(x^0) &\leq 0, \ i = 1, \dots, m, \\ h_i(x) &= 0, \ i = 1, \dots, p, \\ \lambda_i &\geq 0, \ i = 1, \dots, m, \\ \lambda_i f_i(x^*) &= 0, \ i = 1, \dots, m. \end{aligned} \tag{6}$$

The first condition means that x^* minimises the Lagrangian, the next three imply that all the constraints (both primal and dual) are satisfied. The third one, called *complementary slackness*, means that λ_i has to be zero when the corresponding constraint f_i is strictly negative (otherwise it would penalise the Lagrangian).

These conditions are necessary, meaning that they hold for any solution of a problem with zero duality gap. However, for convex problems this conditions are also sufficient. Hence, it's enough to find a point that satisfies all of them.

How do I use it? In practice, a lot of problems are convex and usually do satisfy the Slater's conditions. Hence, it's enough to find when $L'_x = 0$ (another notation for partial derivatives) and then match all the conditions. Let's illustrate it with a simple example.

Quadratic problem example. For $b > a$, consider

$$\begin{aligned} \min_x \quad & x^2, \\ \text{s.t.} \quad & (x - a)(x - b) \leq 0. \end{aligned} \tag{7}$$

It's a convex problem that satisfies the Slater's condition. The solution is trivial, but let's get it formally. The Lagrangian is

$$L(x, \lambda) = x^2 + \lambda(x - a)(x - b).$$

Setting $L'_x = 0$, we get

$$x = \frac{\lambda(a + b)}{2(1 + \lambda)}.$$

Complementary slackness allows us to find possible λ:

$$\lambda(x - a)(x - b) = 0 \quad \Rightarrow \quad \frac{\lambda}{4(1 + \lambda)^2}(\lambda(b - a) - 2a)(\lambda(a - b) - 2b) = 0.$$

It gives three possible λ along with corresponding x:

$$x = 0,\ \lambda = 0;$$
$$x = a,\ \lambda = \frac{2a}{b-a};$$
$$x = b,\ \lambda = \frac{2b}{a-b}.$$

When $a < 0 < b$, the last two solutions result in negative λ, hence $x = 0$. The same solution holds when either a or b is zero.

If $b < 0$, the first solution results in $(x-a)(x-b) = ab > 0$ and the second one gives $\lambda < 0$. Hence, $x = b$.

If $a > 0$, again the first solution violates the x constraint and the last one violates the λ constraint.

Overall, we first set L'_x to zero and then used complementary slackness in order to select possible values of the dual variable. The unique solution is determined by satisfying the primal and dual constraints. This is more or less the general algorithm.

Infinite-dimensional problems. Somewhat less rigorously, we can use Lagrange multipliers for the problems of the form

$$\begin{aligned} \min_p \quad & \int F(x, p(x))dx, \\ \text{s.t.} \quad & \int p(x) f_i(x) dx = c_i,\ i = 1, \dots, m. \end{aligned} \tag{8}$$

In such calculus of variations problems, the function $p(\cdot)$ belongs to some function class (e.g. continuously differentiable on $[a, b]$) and F is also sufficiently differentiable.

Similarly to the finite-dimensional case, we'll study

$$L(p, \lambda) = \int F(x, p(x))dx + \sum_i \lambda_i \left(\int p(x) f_i(x) dx - c_i \right).$$

Instead of the usual derivative, consider the variation of the functional L w.r.t. δp:

$$\frac{\delta L}{\delta p(x)} = \frac{\delta F(x, p(x))}{\delta p(x)} + \sum_i \lambda_i f_i(x).$$

Setting it to zero and then choosing λ_i that satisfy the constraints would give the solution.

References

[1] Boyd, Stephen; Vandenberghe, Lieven. *Convex Optimization.* Cambridge University Press, 2004.

Problems

Task 1. (ridge regression) Consider the usual linear regression problem. We have a data matrix $X \in \mathbb{R}^{n \times m}$ (n objects and m features) and a vector of answers $y \in \mathbb{R}^n$. We want to find a vector w that solves the following:

$$\begin{aligned} \min_{w} \quad & \frac{1}{2}\|y - Xw\|_2^2, \\ \text{s.t.} \quad & \|w\|_2^2 \leq c. \end{aligned} \tag{9}$$

Is it a convex problem? Write down the Lagrangian and express w in terms of the dual variable λ. Using the KKT conditions, obtain an equation for evaluating λ. Why is the solution unique? **Hint:** you will need the SVD decomposition $X = USV^\top$.

Task 2. (Bregman divergence) Assume that we have a vector $y \in \mathbb{R}^n_{++}$ (meaning that all entries are positive). Solve the following problem:

$$\begin{aligned} \min_{x} \quad & \sum_{i=1}^{n} \left(x_i \log \frac{x_i}{y_i} - (x_i - y_i) \right), \\ \text{s.t.} \quad & \sum_{i=1}^{n} x_i = 1. \end{aligned} \tag{10}$$

Also assume that the objective is convex on the given set. Note that the non-negativity constraint on x is implicit (it won't affect optimisation; you can add it and see what happens) and actually enforced by the domain of $\log(\cdot)$.

Task 3. (entropy maximisation) In the TN course, you'll encounter the entropy maximisation problem:

$$\begin{aligned} \max_{p} \quad & -\int p(x) \log p(x) dx, \\ \text{s.t.} \quad & \int p(x) dx = 1, \\ & \int p(x) f(x) dx = a. \end{aligned} \tag{11}$$

Find the maximising distribution $p(x)$ using calculus of variations. Note that the non-negativity constraint is again implicit. What distribution do you obtain for $f(x) = x$ (so the expectation is fixed)? For $f(x) = x^2$?

Solutions

Task 1. For a data matrix $X \in \mathbb{R}^{n \times m}$ (n objects and m features) and a vector of answers $y \in \mathbb{R}^n$, we're solving the following problem:

$$\begin{aligned} \min_w \quad & \frac{1}{2}\|y - Xw\|_2^2, \\ \text{s.t.} \quad & \|w\|_2^2 \leq c. \end{aligned} \tag{12}$$

You can easily verify that this problem is convex. There is only one constraint, so the Lagrangian is

$$L(w, \lambda) = \|y - Xw\|_2^2 + \lambda(\|w\|_2^2 - c).$$

It's derivative w.r.t. w is

$$\frac{\partial}{\partial w} L(w, \lambda) = -2X^\top(y - Xw) + 2\lambda w.$$

We need to set it to zero, hence

$$X^\top y = X^\top X w + \lambda w = (X^\top X + \lambda)w.$$

The matrix on the right hand side is invertible for every positive λ, thus we get the well-known *ridge regression* solution:

$$w = (X^\top X + \lambda)^{-1} X^\top y.$$

That's only the first part of the KKT conditions. Let's check the rest of them:

$$\|w\|_2^2 = y^\top X (X^\top X + \lambda)^{-2} X^\top y \leq c.$$

Using the SVD decomposition $X = USV^\top$, we get

$$y^\top USV^\top V(S^2 + \lambda)^{-2} V^\top VSU^\top y \leq c.$$

As $V^\top V = I$, we have

$$y^\top US(S^2 + \lambda)^{-2} SU^\top y \leq c.$$

Denoting $b = U^\top y$, we get

$$\sum_{i=1}^m b_i^2 \frac{s_i^2}{(s_i^2 + \lambda)^2} \leq c.$$

If $\sum_{i=1}^m (b_i/s_i)^2 \leq c$, any non-zero λ would not satisfy complementary slackness (as both multipliers would be non-zero), hence $\lambda = 0$. The other way round, $\lambda = 0$ is only possible when $\sum_{i=1}^m (b_i/s_i)^2 \leq c$ as the constraint has to be satisfied.

When $\sum_{i=1}^m (b_i/s_i)^2 > c$ and thus $\lambda > 0$, complementary slackness implies that

$$\sum_{i=1}^m b_i^2 \frac{s_i^2}{(s_i^2 + \lambda)^2} = c,$$

so we can find the correct λ numerically.

It's clear that λ is a monotonically decreasing function of c. Therefore choosing c is equivalent to picking the corresponding λ and solving

$$\min_w \|y - Xw\|_2^2 + \lambda \|w\|_2^2$$

without constraints, which is sometimes called the *Tikhonov regularisation problem*. The smaller the λ, the bigger the allowed norm of w.

Task 2. Assume that we have a vector $y \in \mathbb{R}^n_{++}$ (meaning that all entries are positive). We'll solve the following problem:

$$\begin{aligned} \min_x \quad & \sum_{i=1}^n \left(x_i \log \frac{x_i}{y_i} - (x_i - y_i) \right), \\ \text{s.t.} \quad & \sum_{i=1}^n x_i = 1. \end{aligned} \tag{13}$$

Formally, we're trying to minimise the Bregman divergence associated to the negative entropy on a simplex (the sum constraint). Informally, we're projecting y onto the set of probability distributions.

It's a convex problem as $x \log x$ is convex on the simplex. The Lagrangian is

$$L(x, \lambda) = \sum_{i=1}^n \left(x_i \log \frac{x_i}{y_i} - (x_i - y_i) + \lambda x_i \right) - \lambda n.$$

Setting it's derivative w.r.t. x_i to zero, we get

$$\frac{\partial}{\partial x_i} L(x, \lambda) = \log \frac{x_i}{y_i} + 1 - 1 + \lambda = 0 \quad \Rightarrow \quad x_i = y_i e^{-\lambda}.$$

Having and additional $x_i \geq 0$ constraint would result in $x_i = y_i \exp(-\lambda + \nu_i)$, but as $y_i > 0$ and thus $x_i > 0$, complementary slackness enforces $\nu_i = 0$.

Now we need to satisfy the KKT conditions. Check the simplex constraint:

$$1 = \sum_i x_i = e^{-\lambda} \sum_i y_i \quad \Rightarrow \quad e^{-\lambda} = \frac{1}{\|y\|_1} \text{ and } x = \frac{y}{\|y\|_1}.$$

We got that this problem is equivalent to simple re-normalisation. Note that we don't check complementary slackness as there is no inequality constraints.

The discussed problem arises in the optimisation algorithm called *Mirror Descent*. Check the Sebastien Bubeck's blog for a reference (links to part 1 and part 2).

Task 3. For the entropy maximisation problem

$$\begin{aligned} \max_p \quad & -\int p(x) \log p(x) dx, \\ \text{s.t.} \quad & \int p(x) dx = 1, \\ & \int p(x) f(x) dx = a, \end{aligned} \tag{14}$$

which is equivalent to

$$\begin{aligned} \min_p \quad & \int p(x) \log p(x) dx, \\ \text{s.t.} \quad & \int p(x) dx = 1, \\ & \int p(x) f(x) dx = a, \end{aligned} \tag{15}$$

the Lagrangian is

$$L(p, \lambda, \nu) = \int p(x) \log p(x) dx + \lambda \left(\int p(x) dx - 1 \right) + \nu \left(\int p(x) f(x) dx - a \right).$$

The non-negativity constraint is implicit, but it won't affect the solution. It's variation is:

$$\frac{\delta L}{\delta p(x)} = \log p(x) + 1 + \lambda + \nu f(x) = 0.$$

Hence,

$$p(x) = \exp\left(-\left(1 + \lambda + \nu f(x)\right)\right) = \exp(-1 - \lambda) \exp(-\nu f(x)).$$

From the first constraint we get that

$$\int \exp(-\nu f(x)) dx = \exp(1 + \lambda).$$

The second constraint results in

$$\frac{1}{\int \exp(-\nu f(x)) dx} \int \exp(-\nu f(x)) f(x) dx = a,$$

which does not have a general form.

If $f(x) = x$ and limits of integration are $[0, \infty)$, we have the exponential distribution with rate ν:

$$p(x) = \nu \exp(-\nu x).$$

Hence,

$$\int p(x) f(x) dx = \nu^{-1} = a \quad \Rightarrow \quad p(x) = \frac{1}{a} \exp\left(-\frac{x}{a}\right).$$

If $f(x) = x^2$ and the limits are $(-\infty, \infty)$, then we get the zero-mean Gaussian distribution with $\sigma^2 = \nu/2$:

$$p(x) = \frac{\sqrt{2\nu}}{\sqrt{2\pi}} \exp(-\nu x^2) = \frac{\sqrt{\nu}}{\sqrt{\pi}} \exp(-\nu x^2).$$

Hence,

$$\int p(x) x^2 dx = \sigma^2 = \frac{\nu}{2} = a \quad \Rightarrow \quad p(x) = \sqrt{\frac{2a}{\pi}} \exp(-2a x^2).$$