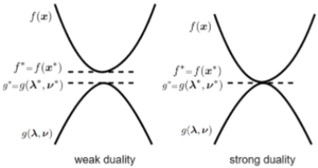
Optimization in Machine Learning

Nonlinear programs and Lagrangian



$f(\boldsymbol{x})$
$f^*=f(\boldsymbol{x}^*)$
$g^*=g(\boldsymbol{\lambda}^*,\boldsymbol{\nu}^*)$
$g(\boldsymbol{\lambda},\boldsymbol{\nu})$
weak duality
$f(\boldsymbol{x})$
$f^*=f(\boldsymbol{x}^*)$
$g^*=g(\boldsymbol{\lambda}^*,\boldsymbol{\nu}^*)$
$g(\boldsymbol{\lambda},\boldsymbol{\nu})$
strong duality

Learning goals

- Lagrangian for general constrained optimization
- Geometric intuition for Lagrangian duality
- Properties and examples

NONLINEAR CONSTRAINED OPTIMIZATION

In the previous lecture, we introduced the general primal LP of the form

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^d} \quad & \boldsymbol{c}^T\boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \\ & \boldsymbol{G}\boldsymbol{x} = \boldsymbol{h} \end{aligned}$$

and discussed how its dual formulation can be related to a function $\mathcal{L}$ of the form

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{c}^T\boldsymbol{x} + \alpha^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) + \beta^T(\boldsymbol{G}\boldsymbol{x} - \boldsymbol{h})$$

NONLINEAR CONSTRAINED OPTIMIZATION

Given the general form of a constraint optimization problem

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^d} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g_i(\mathbf{x}) \leq 0 \, i = 1, \ldots, k, \\ & h_j(\mathbf{x}) = 0 \, j = 1, \ldots, l, . \end{aligned}$$

with not necessarily linear functions $f : \mathbb{R}^d \to \mathbb{R}$, $g_i : \mathbb{R}^d \to \mathbb{R}$, $h_j : \mathbb{R}^d \to \mathbb{R}$ and assuming $f, g, h \in \mathcal{C}^2$, it is tempting to mirror this construction and define

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) := f(\mathbf{x}) + \sum_{i=1}^{k} \alpha_i g_i(\mathbf{x}) + \sum_{j=1}^{l} \beta_j h_j(\mathbf{x})$$

NONLINEAR CONSTRAINED OPTIMIZATION

with multipliers $\alpha_i \geq 0$ and unconstrained β_i. As we will see later, this construction, called the Lagrange function (or Lagrangian) associated with the primal nonlinear optimization problem, will be key to defining Lagrangian duality. The multipliers α_i and β_i are called Lagrange multipliers.

CONSTRAINED PROBLEMS: THE DIRECT WAY

Similar to LPs, however, certain constraint problems can be solved in a direct way using methods from calculus and linear algebra.

Example 1:

$$\begin{aligned} \min_{x \in \mathbb{R}} \quad & 2 - x^2 \\ \text{s.t.} \quad & x - 1 = 0 \end{aligned}$$

In this trivial example, we would simply resolve the constraint

$$\begin{aligned} x - 1 &= 0 \\ x &= 1 \end{aligned}$$

and insert it into the objective:

$$x^* = 1, \qquad f(x^*) = 1$$

CONSTRAINED PROBLEMS: THE DIRECT WAY

Example 2:

$$\begin{array}{ll} \min\limits_{\mathbf{x}\in\mathbb{R}^2} & -2 + x_1^2 + 2x_2^2 \\ \text{s.t.} & x_1^2 + x_2^2 - 1 = 0 \end{array}$$

We solve the problem by resolving the constraint

$$x_1^2 = 1 - x_2^2$$

and inserting it into the objective function

$$\begin{aligned} f(x_1, x_2) &= -2 + x_1^2 + 2x_2^2 \\ &= -2 + (1 - x_2^2) + 2x_2^2 = -1 + x_2^2. \end{aligned}$$

We turned the optimization problem into a one-dimensional, unconstrained optimization problem that has a minimum at $x_2 = 0$. Plugging $x_2 = 0$ into constraint, we get $x_1 = \pm 1$.

However, this direct way is not possible in most cases.

A CLASSIC EXAMPLE: "MILKMAID PROBLEM"

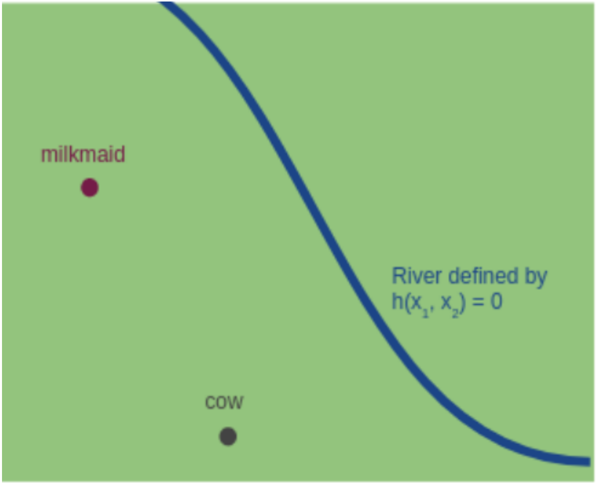
Question 1: Is a general recipe for solving general constraint nonlinear optimization problems?
Question 2: Can we understand this recipe geometrically?
Question 3: How does this relate to the Lagrange function approach?

For this purpose, we consider the classical “milkmaid problem”; the following example is taken from *Steuard Jensen, An Introduction to Lagrange Multipliers* (but the example works of course equally well with a “milk man”).

- Assume a milk maid is sent to the field to get the day’s milk
- The milkmaid wants to finish her job as quickly as possible
- However, she has to rinse out her bucket first in the nearby river.

A CLASSIC EXAMPLE: "MILKMAID PROBLEM"

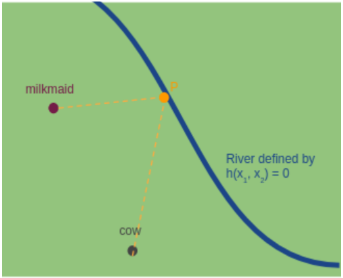
What is the best point *P* to rinse her bucket?



A CLASSIC EXAMPLE: "MILKMAID PROBLEM"

Let us put this into maths: The milkmaid wants to find the point P at the riverbank for which the total distance $f(P)$ is a minimum, with

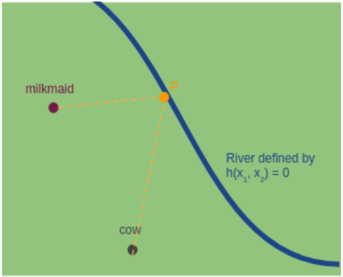
- $f(P)$ being defined as $f(P) := d(M, P) + d(P, C)$ (distance from M to P + distance from P point to C), and
- the river is described by $h(x_1, x_2) = 0$.



A CLASSIC EXAMPLE: "MILKMAID PROBLEM"

We want to solve the following optimization problem

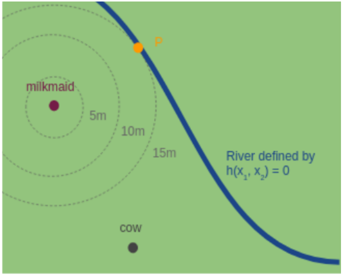
$$\begin{aligned} \min_{x_1, x_2} \quad & f(x_1, x_2) \\ \text{s. t.} \quad & h(x_1, x_2) = 0. \end{aligned}$$



A CLASSIC EXAMPLE: "MILKMAID PROBLEM"

Let us visualize how far the milkmaid could get for any fixed total distance $f(P)$.

Let us first assume we only care about the distance from M to P. We might picture this as a set of concentric circles. How far can the milkmaid get for a distance of 5, 10, 15, ... meters?

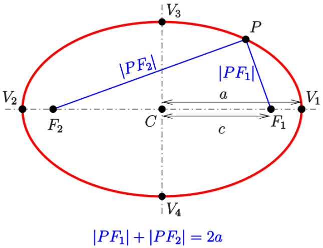


As soon as one of those circles was big enough to touch the river, we'd recognize the point where it touched as the closest riverbank point to M.

A CLASSIC EXAMPLE: "MILKMAID PROBLEM"

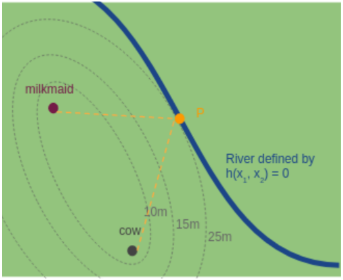
We use now the geometric definition of an ellipse:
Given two fixed points, F_1, F_2 called the foci and a distance $2a$ which is greater than the distance between the foci, the ellipse is the set of points P such that the sum of the distances $|PF_1|, |PF_2|$ is equal to $2a$:

$$E = \{P \in \mathbb{R}^2 | \, |PF_1| + |PF_2| = 2a\}$$



A CLASSIC EXAMPLE: "MILKMAID PROBLEM"

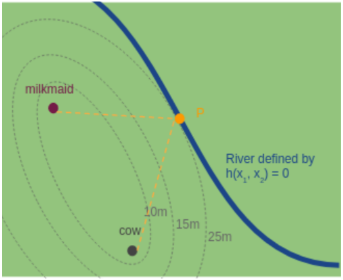
In the milkmaid problem, this means that the milkmaid could get to the cow by way of any point on a given ellipse with foci M and C in the same amount of time: the ellipses are curves of constant $f(P)$.



A CLASSIC EXAMPLE: "MILKMAID PROBLEM"

Therefore, to find the desired point P on the riverbank, we must simply find the **smallest ellipse** with M and C as foci, that intersects the curve of the river.

The optimum point P is the one where the respective ellipse is just **tangential** to the riverbank!



It is obvious from the picture that the “perfect” ellipse and the river are truly tangential to each other at the ideal point P!

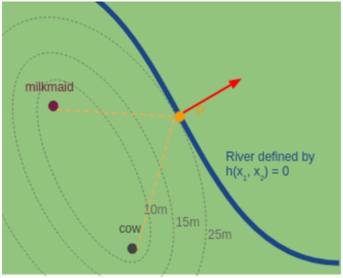
A CLASSIC EXAMPLE: "MILKMAID PROBLEM"

In multivariate calculus, the gradient ∇h of a function h is normal to a surface on which h is constant (∇h is normal to the “contour lines”).

In our case, we have two functions whose normal vectors are parallel:

$$\nabla f(P) = \beta \nabla h(P).$$

The multiplier β is necessary because the magnitudes of the two gradients may be different. β is called **Lagrange multiplier**.



LAGRANGE FUNCTION

In order to solve a problem with a single equality constraint, we require the following to be fulfilled:

$$\begin{aligned} \nabla f(\mathbf{x}^*) &= \beta \nabla h(\mathbf{x}^*), \quad \beta \in \mathbb{R} \\ h(\mathbf{x}^*) &= 0 \end{aligned}$$

where the first line requires that gradients are parallel, and the second requires that the constraint is met.

This can also be written as:

$$\begin{aligned} \nabla f(\mathbf{x}^*) + \beta \nabla h(\mathbf{x}^*) &= 0, \quad \beta \in \mathbb{R} \\ h(\mathbf{x}^*) &= 0 \end{aligned}$$

LAGRANGE FUNCTION

If we define $\mathcal{L}(\mathbf{x}, \beta) := f(\mathbf{x}) + \beta h(\mathbf{x})$, then the point fulfilling the equations above is nothing but a stationary point of $\mathcal{L}$:

$$\begin{pmatrix} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^*, \beta) \\ \nabla_{\beta} \mathcal{L}(\mathbf{x}^*, \beta) \end{pmatrix} = \begin{pmatrix} \nabla f(\mathbf{x}^*) + \beta \nabla h(\mathbf{x}^*) \\ h(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

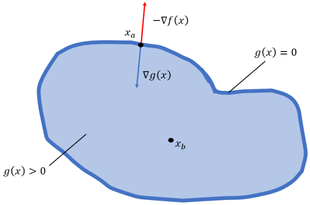
The function $\mathcal{L}$ is called **Lagrange function** or **Lagrangian**.

Note: In some books the Lagrangian is defined as $\mathcal{L}(\mathbf{x}, \beta) := f(\mathbf{x}) - \beta h(\mathbf{x})$. Both conventions obtain the same stationary points, but the sign of β is different.

LAGRANGE FUNCTION

The method can be extended to inequality constraints of the form $g(\mathbf{x}) \geq 0$. There are two possible cases for a solution:

- If the optimal solution x_b is inside the constraint region, the constraint is inactive, meaning that β can be set to zero.
- If the optimal solution x_a lies on the boundary $g(\mathbf{x}) = 0$, the negative gradient ∇f points in the opposite direction of the gradient of $g(\mathbf{x})$



LAGRANGE FUNCTION AND PRIMAL PROBLEM

The Lagrangian can be extended to general constrained optimization problems with k inequality constraints ($g(\mathbf{x}) \geq 0$) and l equality constraints ($g(\mathbf{x}) = 0$):

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) := f(\mathbf{x}) + \sum_{i=1}^{k} \alpha_i g_i(\mathbf{x}) + \sum_{j=1}^{l} \beta_j h_j(\mathbf{x})$$

with **Lagrange multipliers** $\alpha_i \geq 0$, β_i.

We can write the problem **equivalently** as

$$\min_{\mathbf{x}} \max_{\boldsymbol{\alpha} \geq 0, \boldsymbol{\beta}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}).$$

We call this problem **primal problem.**

Why is the above formulation equal to our initial (constrained) formulation of the optimization problem?

LAGRANGE FUNCTION AND PRIMAL PROBLEM

For simplicity, assume we only have a single inequality constraint.

Assume an **x does break** the inequality constraint, i.e., $g(\mathbf{x}) > 0$. Then

$$\max_{\alpha \geq 0} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}) = \max_{\alpha \geq 0} f(\mathbf{x}) + \alpha g(\mathbf{x}) = \infty,$$

because we can pick $\alpha = \infty$ to drive the objective value to infinity.

If otherwise $g(\mathbf{x}) \leq 0$,

$$\max_{\alpha \geq 0} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}) = \max_{\alpha \geq 0} f(\mathbf{x}) + \alpha g(\mathbf{x}) = f(\mathbf{x}),$$

because $\alpha g(\mathbf{x}) \leq 0$ for any $\alpha \geq 0$, and $g(\mathbf{x}) = 0$ if $\alpha = 0$.

LAGRANGE FUNCTION AND PRIMAL PROBLEM

We end up with

$$\min_{x} \max_{\alpha \geq 0} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}) = \begin{cases} \infty & \text{if } g(\mathbf{x}) > 0 \\ \min_x f(\mathbf{x}) & \text{if } g(\mathbf{x}) \leq 0 \end{cases}$$

which corresponds to our original formulation.

A similar argument holds for the equality constraint $h(\mathbf{x})$.

EXAMPLE: LAGRANGE FUNCTION FOR QP'S

We consider quadratic programming

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\top \boldsymbol{Q}\mathbf{x} \\ \text{such that} \quad & h(\mathbf{x}) := \boldsymbol{C}\mathbf{x} - \boldsymbol{d} = 0, \end{aligned}$$

with $\boldsymbol{Q} \in \mathbb{R}^{d\times d}$ symmetric, $\boldsymbol{C} \in \mathbb{R}^{l\times d}$, $\boldsymbol{d} \in \mathbb{R}^{n}$.

The Lagrange function is

$$\mathcal{L}(\mathbf{x}, \beta) = \frac{1}{2}\mathbf{x}^\top \boldsymbol{Q}\mathbf{x} + \beta^\top (\boldsymbol{C}\mathbf{x} - \boldsymbol{d}).$$

EXAMPLE: LAGRANGE FUNCTION FOR QP'S

For the calculation of the stationary points we calculate the gradient of the Lagrange function:

$$\nabla \mathcal{L}(\mathbf{x}, \beta) = \begin{pmatrix} \frac{\partial L}{\partial \mathbf{x}} \\ \frac{\partial L}{\partial \boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{Qx} + \boldsymbol{C}^\top \beta \\ \boldsymbol{Cx} - \boldsymbol{d} \end{pmatrix} = \mathbf{0}$$

This is a linear system with $n + p$ equations, which we write as

$$\begin{pmatrix} \boldsymbol{Q} & \boldsymbol{C}^\top \\ \boldsymbol{C} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \beta \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \boldsymbol{d} \end{pmatrix}.$$

It follows that finding stationary points of the Lagrange function corresponds to the solution of a system of linear equations and can be solved efficiently and stable with the help of suitable matrix decompositions.

EXAMPLE: LAGRANGE FUNCTION FOR LASSO

Lasso regression:

$$\begin{aligned} \min_{\theta} \quad & \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 \\ \text{s.t.} \quad & \|\boldsymbol{\theta}\|_1 - t \leq 0. \end{aligned}$$

We formulate the primal problem with the help of the Lagrangian:

$$\min_{\theta} \max_{\alpha \geq 0} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}) = \min_{\theta} \max_{\alpha \geq 0} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \alpha \left(\|\boldsymbol{\theta}\|_1 - t \right) \right\}.$$

The explicit derivation of the dual of the Lasso can be found in ▸ Osbourne et al.,2000.

LAGRANGE DUALITY

The **dual problem** of the above problem results when we reverse minimization and maximization:

$$\max_{\boldsymbol{\alpha} \geq 0, \boldsymbol{\beta}} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

We define the **dual function** $g(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

We hold $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ constant and optimize **x** (unconstrained optimization problem!). In a second step, we maximize $\boldsymbol{\alpha}, \boldsymbol{\beta}$.

Note: This definition of duality is a generalization of the duality for Linear programming. For LPs, the two definitions are the same.

LAGRANGE DUALITY

Important characteristics of the dual problem:

- The dual problem is **always convex**. Many methods are therefore based on the solution of the dual problem.
- **Weak duality always** applies, i.e.

$$g(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \leq f(\mathbf{x}^*)$$

- If the primal problem is convex, we have **strong duality** in general[1], i.e.

$$g(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = f(\mathbf{x}^*)$$

[1] **slater's condition** must be fulfilled. Read more here `http://www.cs.cmu.edu/~ggordon/10725-F12/slides/15-duality.pdf`.