Logistic regression and convex analysis

Pierre Gaillard, Alessandro Rudi

March 12, 2019

In this class, we will see logistic regression, a widely used classification algorithm. Contrary to linear regression, there is no closed-form solution and one needs to solve it thanks to iterative convex optimization algorithms. We will then see the basics of convex analysis.

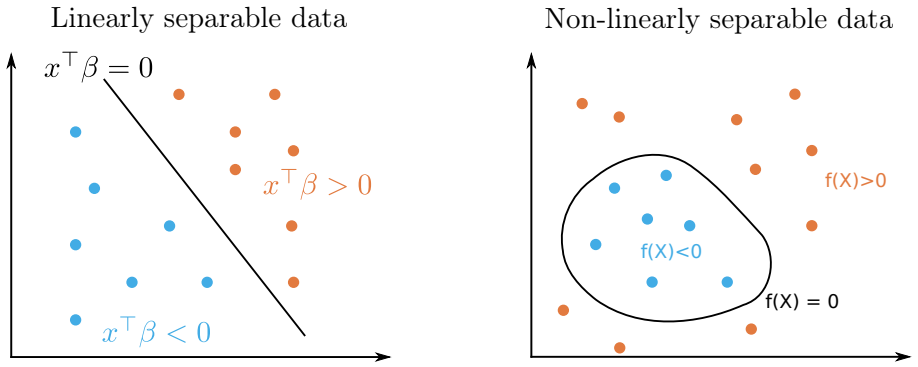
1 Logistic regression

We will consider the binary classification problem in which one wants to predict outputs in $\{0,1\}$ from inputs in $\mathbb{R}^d$. We consider a training set $D_n := \big\{(X_i, Y_i)\big\}_{1\leqslant i\leqslant n}$. The data points (X_i, Y_i) are i.i.d. random variables and follow a distribution $\mathcal{P}$ in $\mathcal{X}\times\mathcal{Y}$. Here, $\mathcal{X}=\mathbb{R}^d$ and $\mathcal{Y}=\{0,1\}$.

Goal We would like to use a similar algorithm to linear regression. However, since the outputs Y_i are binary and belong to $\{0,1\}$ we cannot predict them by linear transformation of the inputs X_i (which belong to $\mathbb{R}^d$). We will thus classify the data thanks to classification rules $f:\mathbb{R}^d\mapsto\mathbb{R}$ such that:

$$f(X_i)\begin{cases}\geqslant 0 & \Rightarrow \quad Y_i=+1\\ <0 & \Rightarrow \quad Y_i=0\end{cases},$$

to separate the data into two groups. In particular, we will consider linear functions f of the form $f_\beta : x\mapsto x^\top\beta$. This assumes that the data are well-explained by a linear separation (see figure below).



Of course, if the data does not seem to be linearly separable, we can use similar tricks that we used for linear regression (polynomial regression, kernel regression, splines,...). We search a feature map $x\mapsto\phi(x)$ into a higher dimensional space in which the data are linearly separable.

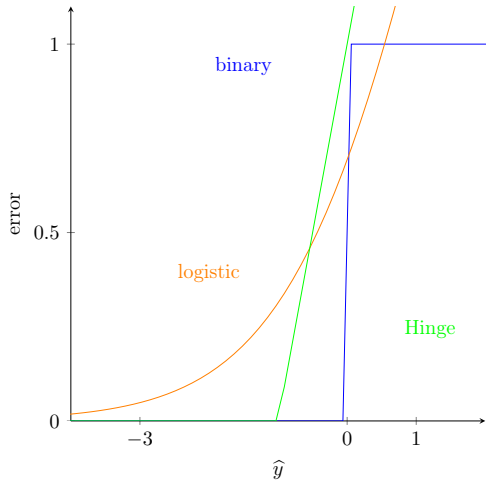


Figure 1: Binary, logistic and Hinge loss incured for a prediction $\widehat{y} := x^\top \beta$ when the true observation is $y = 0$.

Loss function To minimize the empirical risk, it remains to choose a loss function to assess the performance of a prediction. A natural loss is the *binary loss*: 1 if there is a mistake ($f(X_i) \neq Y_i$) and 0 otherwise. The empirical risk is then:

$$\widehat{\mathcal{R}}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i \neq \mathbb{1}_{X_i^\top \beta \geqslant 0}} .$$

This loss function is however not convex neither in β. The minimization problem $\min_\beta \widehat{\mathcal{R}}_n(\beta)$ is extremely hard to solve. The idea of logistic regression consists in replacing the binary loss with another similar loss function which is convex in β. This is the case of the *Hinge loss* and of the logistic loss $\ell : \mathbb{R} \times \{0,1\} \to \mathbb{R}_+$. The latter assigns to a linear prediction $\widehat{y} = x^\top \beta$ and an observation $y \in \{0,1\}$ the loss

$$\ell(\widehat{y}, y) := y \log\left(1 + e^{-\widehat{y}}\right) + (1-y)\log\left(1 + e^{\widehat{y}}\right). \tag{1}$$

The binary loss, Hinge loss and logistic loss are plotted in Figure 1.

Definition 1.1 (Logistic regression estimator)**.** *The logistic regression estimator is the solution of the following minimization problem:*

$$\widehat{\beta}_{(logit)} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell(X_i^\top \beta, Y_i) ,$$

where ℓ is the logistic loss defined in Equation (1)*.*

The advantage of the logistic loss with respect to the Hinge loss is that it has a probabilistic interpretation by modeling $\mathbb{P}(Y = 1|X)$, where (X, Y) is a couple of random variables following the law of (X_i, Y_i). We will see more on this in the lecture on Maximum Likelihood.

Computation of $\widehat{\beta}_{(\text{logit})}$ Similarly to OLS, we may try to analytically solve the minimization problem by canceling the gradient of the empirical risk. Since $\frac{\partial \ell(\widehat{y},y)}{\partial \widehat{y}} = \sigma(\widehat{y}) - y$, where $\sigma : z \mapsto$

$\frac{1}{1+e^{-z}}$ is the logistic function, we have:

$$\nabla \widehat{\mathcal{R}}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} X_i\big(\sigma(X_i^\top \beta) - Y_i\big) = \frac{1}{n} X\big(Y - \sigma(X\beta)\big)$$

where $X := (X_1, \dots, X_n)^\top$, $Y := (Y_1, \dots, Y_n)$, and $\sigma(X\beta)_i := \sigma(X_i^\top \beta)$ for $1 \leqslant i \leqslant n$. Bad news: the equation $\nabla \widehat{\mathcal{R}}_n(\beta) = 0$ has no closed-form solution. It needs to be solved through iterative algorithm (gradient descent, Newton's method,...). Fortunately, this is possible because the logistic loss is convex in its first argument. Indeed,

$$\frac{\partial^2 \ell(\widehat{y}, y)}{\partial \widehat{y}} = \sigma(\widehat{y})\sigma(-\widehat{y}) > 0\,.$$

The loss is strictly convex, the solution is thus unique. In this class and the next one, we will see tools and methods to solve convex optimization problems.

Regularization Similarly to linear regression, logistic regression may over-fit the data (especially when $p > n$). One needs then to add a regularization such as $\lambda\|\beta\|_2^2$ to the logistic loss.

2 Convex analysis

We will now see notions of convex analysis to solve convex optimization problems such as the one of logistic regression. For more details on this topic, we refer to the monograph Boyd and Vandenberghe, 2004. This class and the next one, we will see two aspects:

- convex analysis: properties of convex functions and convex optimization problems
- convex optimization: algorithms (gradient descent, Newton's method, stochastic gradient descent, ...)

Convexity is a crucial notion in many fields of mathematics and computer sciences. In machine learning, convexity allows to get well-defined problems with efficient solutions. A typical example is the problem of *empirical risk minimization*:

$$\widehat{f}_n \in \operatorname*{arg\,min}_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell\big(f(X_i), Y_i\big) + \lambda\Omega(f)\,, \qquad (*)$$

where $D_n = \big\{(X_i, Y_i)\big\}_{1 \leqslant i \leqslant n}$ is the data set, $\mathcal{F}$ is a convex set of predictors $f : \mathcal{X} \mapsto \mathbb{R}$, $\widehat{y} \mapsto \ell(\widehat{y}, y)$ are convex loss functions for all $y \in \mathcal{Y}$ and Ω is a convex penalty ($\|\cdot\|_2, \|\cdot\|_1, \dots$).

Convexity will be useful to analyze

- the statistical properties of the solution $\widehat{f}_n$ and its generalization error (i.e., its risk):

$$\mathcal{R}(\widehat{f}_n) := \mathbb{E}\big[\ell(f(X), Y)\big|D_n\big]$$

- get efficient algorithms to solve the minimization problem $(*)$ and find $\widehat{f}_n$.

2.1 Convex sets

In this class, we will only consider finite dimensional Euclidean space (typically $\mathbb{R}^d$).

Definition 2.1 (Convex set)**.** *A set $K \subseteq \mathbb{R}^d$ is convex if and only if for all $x, y \in K$ $[x, y] \subset K$ (or equivalently for all $\alpha \in (0, 1)$, $\alpha x + (1-\alpha)y \in K$).*

Example 2.1. *Here are a few examples of convex sets*

- *Hyperplans:* $K = \{x \in \mathbb{R}^d : a^\top x = b, a \neq 0, b \in \mathbb{R}\}$
- *Half spaces:* $K = \{x \in \mathbb{R}^d : a^\top x \geqslant b, a \neq 0, b \in \mathbb{R}\}$
- *Affine subspaces:* $K = \left\{x \in \mathbb{R}^d : Ax = b, A \in M_d(\mathbb{R}), b \in \mathbb{R}\right\}$
- *Balls:* $\|x\| \leqslant R$, $R > 0$
- *Cones:* $K = \{(x, r) \in \mathbb{R}^{d+1}, \|x\| \leqslant r\}$
- *Convex polytopes: intersections of half spaces.*

Properties of Convex sets :

- stability by intersection (not necessarily countable)
- stability by affine transformation
- convex separation: if C, D are disjoints convex sets ($C \cap D = \emptyset$), then there exists a hyperplane which separates C and D:

$$\exists a \neq 0, b \in \mathbb{R} \text{ such that } C \subset \left\{a^\top x \geqslant b\right\} \text{ and } D \subset \left\{a^\top x \leqslant b\right\}.$$

  The inequalities are strict if C and D are compact. Exercise: show this property when C and D are compact (clue: define $(x, y) \in \arg\min_{x \in C, y \in D} \|x - y\|$).

Definition 2.2 (Convex Hull)**.** *Let $A \in \mathbb{R}^d$. The convex hull, denoted* $\mathrm{Conv}(A)$*, of A is the smallest convex set that contains A. In other words:*

$$\mathrm{Conv}(A) = \bigcap \left\{B \subseteq \mathbb{R}^d : A \subseteq B \text{ and } B \text{convex}\right\}$$

$$\left\{x \in \mathbb{R}^d : \exists p \geqslant 1, \alpha \in \mathbb{R}_+^p, \sum_{i=1}^p \alpha_i = 1 \text{ and } z_1, \ldots, z_p \in A \text{such that } x = \sum_{i=1}^p \alpha_i z_i\right\}$$

2.2 Convex functions

Definition 2.3 (Convex function)**.** *A function $f : D \subset \mathbb{R}^d \to \mathbb{R}$ with D convex is*

- *convex iff for all $x, y \in D$ and $0 \leqslant \alpha \leqslant 1$,*

$$f(\alpha x + (1-\alpha)y) \leqslant \alpha f(x) + (1-\alpha) f(y)$$

- *strictly convex iff for all $x, y \in D$ and $0 \leqslant \alpha \leqslant 1$,*

$$f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha) f(y)$$

- *μ-strongly convex if there exists $\mu > 0$ such that*

$$f(\alpha x + (1-\alpha)y) \leqslant \alpha f(x) + (1-\alpha) f(y) - \frac{\mu}{2}\|x\|^2$$

Proposition 2.1. *f is μ-strongly convex if and only if $x \mapsto f(x) - \frac{\mu}{2}\|x\|^2$ is convex.*

A few examples of useful convex functions:

- dimension $d = 1$: x, x^2, $-\log x$, e^x, $\log(1 + e^{-x})$, $|x|^p$ for $p \geqslant 1$, $-x^p$ for $p < 1$ and $x \geqslant 0$
- higher dimension $d \geqslant 1$: linear functions $x \mapsto a^\top x$, quadratic functions $x \mapsto x^\top Q x$ for Q semidefinite (symmetric) positive matrix (i.e., all eigenvalues are nonnegative, or for all x $x^\top Q x \geqslant 0$), norms, $\max\{x_1, \ldots, x_d\}$, $\log\left(\sum_{i=1}^d e_i^x\right)$

Characterization of convex functions

- if f is C^1: f convex $\Leftrightarrow \forall x, y \in D$ $f(x) \geqslant f(y) + f'(y)(x - y)$
- if f is twice differentiable: f convex $\Leftrightarrow \forall x \in D$ its Hessian is semi-definite positive ($f''(x) \geqslant 0$)

Operations which preserve convexity

- supremum of a family $x \mapsto \sup_{i \in I} f_i(x)$
- linear combination with non-negative coefficients
- partial minimization: f convex on $C \times D \Rightarrow \quad y \mapsto \inf_{x \in C} f(x, y)$ is convex on D

Proposition 2.2. *If f is convex on D, then f is continuous on the interior of D. Furthermore, the epigraph of f $\big\{(x, t) \in D \times \mathbb{R}, f(x) \leqslant t\big\}$ is convex.*

Proposition 2.3 (Jensen's inequality)**.** *If f is convex. For all $x_1, \ldots, x_n \in D$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}_+$ such that $\sum_{i=1}^n \alpha_i = 1$ then*

$$f\left(\sum_{i=1}^n \alpha_i x_i\right) \leqslant \sum_{i=1}^n \alpha_i f(x_i)\,.$$

Jensen's inequality can be extended to infinite sums, integrals and expected values: if f is convex

- integral formulation: if $p(x) \geqslant 0$ on $S \subset D$ such that $\int_S p(x)dx = 1$ then

$$f\left(\int_S p(x) x dx\right) \leqslant \int_S p(x) f(x) dx\,.$$

- expected value formulation: if X is a random variable such that $X \in D$ almost surely and $\mathbb{E}[X]$ exists then if

$$f\big(\mathbb{E}\big[X\big]\big) \leqslant \mathbb{E}\big[f(X)\big]\,.$$

2.3 Unconstrained optimization problems

Let $f : \mathbb{R}^d \to \mathbb{R}$ convex and finite on $\mathbb{R}^d$. We consider the problem

$$\inf_{x \in \mathbb{R}^d} f(x)\,.$$

First, remark that we use the notation $\min_x f(x)$ only when the minimum is reached. If no point achieves the minimum, we use the notation $\inf_x f(x)$.

There are three possible cases

- $\inf_{x \in \mathbb{R}^d} f(x) = -\infty$: there is no minimum. For instance, $x \mapsto x$.
- $\inf_{x \in \mathbb{R}^d} f(x) > -\infty$ and the infimum is not reached. This is the case for instance for $x \mapsto \log(1 + e^{-x})$ or for $x \mapsto e^{-x}$.

– $\inf_{x\in\mathbb{R}^d} f(x) > -\infty$ and the minimum is reached and equals $\min_{x\in\mathbb{R}^d} f(x)$. This is the case for instance for coercive functions $\lim_{\|x\|\to\infty} f(x) = +\infty$.

Definition 2.4 (Local minimum)**.** *Let $f : D \to \mathbb{R}$ and $x \in D$. x is a local minimum if and only if there exists an open set $V \subset D$ such that $x \in V$ and $f(x) = \min_{x'\in V} f(x')$.*

Properties:

– f convex $\Rightarrow$ any local minimum is a global minimum.
– f strictly convex $\Rightarrow$ at most one minimum.
– f convex and C^1 then x is a minimum of f on $\mathbb{R}^d$ if and only if $f'(x) = 0$.

As we saw for linear regression, canceling the gradient provide an efficient solution to solve the minimization problem in closed form.

2.4 Constrained optimization problems

Let $f : D \mapsto \mathbb{R}^d$ convex and $C \subset D$ convex. We consider the constrained minimization problem

$$\inf_{x\in C} f(x)\,.$$

C is the constraint set. It is often defined as the intersection of sets of the form $\{h_i(x) = 0\}$ and $\{g_j(x) \leqslant 0\}$.

Example 2.2. *The minimization of a linear function over a compact $A \subset \mathbb{R}^d$. Let $A \subset \mathbb{R}^d$ compact (non-necessarily convex) and $a \in \mathbb{R}^d \neq 0$ then we can reformulate the non-convex minimization on A as a constrained convex optimization problem on* $\mathrm{Conv}(A)$

$$\min_{x\in A}\left\{a^\top x\right\} = \min_{x\in \mathrm{Conv}(A)}\left\{a^\top x\right\}$$

Lagrangian duality A useful notion to solve constrained optimization problems is Lagrangian duality. Assume that we are interested in the following constrained optimization problem:

$$\min_{x\in D} f(x) \quad \text{such that} \quad \begin{cases} h_i(x) = 0 & \text{for } i = 1,\dots,m \\ g_j(x) \leqslant 0 & \text{for } j = 1,\dots,r \end{cases}\,. \tag{P}$$

We denote by $D^* \subseteq D$ the set of points that satisfy the constraints. Remark that equality constraints $h_i(x) = 0$ can be rewritten as inequalities

$$h_i(x) \leqslant 0 \qquad \text{and} \qquad -h_i(x) \leqslant 0\,.$$

Contrary to unconstrained optimization problems, canceling the gradient does not necessarily provide a solution for constrained optimization problems. The basic idea of Lagrangian duality is to take the constraint D^* into account in the minimization problem by augmenting the objective function with a weighted sum of the constraint functions.

Definition 2.5 (Lagrangian)**.** *The Lagrangian associated to the optimization problem* (P) *is the function $\mathcal{L} : D \times \mathbb{R}^m \times \mathbb{R}^r_+$ defined by:*

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \lambda^\top h(x) + \mu^\top g(x)\,.$$

Definition 2.6 (Primal function)**.** *We define the primal function* $\bar{f} : D \to \mathbb{R} \cup \{+\infty\}$ *associated to* (P) *by, for all* $x \in D$

$$\bar{f}(x) = \sup_{\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^r_+} \mathcal{L}(x, \lambda, \mu) = \begin{cases} f(x) & \text{if } x \in D^* \\ +\infty & \text{otherwise} \end{cases} .$$

With these definitions, we remark that the optimization problem (P) can be re-written by using the primal function without the constraints

$$\begin{aligned} \inf_{x \in D^*} f(x) &= \inf_{x \in D} \bar{f}(x) \\ &= \inf_{x \in D} \sup_{\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^r_+} \mathcal{L}(x, \lambda, \mu) \,. \end{aligned} \qquad \text{(Primal problem)}$$

This optimization problem is thus called the *Primal problem.*

The *Dual problem* is obtained by exchanging inf and sup in the primal problem.

$$\sup_{\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^r_+} f^*(\lambda, \mu) = \sup_{\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^r_+} \inf_{x \in D} \mathcal{L}(x, \lambda, \mu) \,, \qquad \text{(Dual problem)}$$

where $f^* : (\lambda, \mu) \mapsto \inf_{x \in D} \mathcal{L}(x, \lambda, \mu)$ is the dual function. If f is convex this function is concave. Remark that the dual of the dual is the primal.

The denote by $D^* = \{x \in D : \bar{f}(x) < \infty\}$ the admissibility domain of the primal. Similarly we denote by $C^* = \{(\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}^r_+ : f^*(\lambda, \mu) > -\infty\}$ the admissibility domain of the dual. If there is no solution to the optimization problem (P) then $D^* = \emptyset$. If the problem is unbounded then $C^* = \emptyset$.

Link between the primal and the dual problems If they are not necessarily identical the primal and the dual problems have strong relationship. For any (λ, μ), $f^*(\lambda, \mu)$ provides a lower bound on the solution of (P). The dual problem finds the best lower bound.

Proposition 2.4 (Weak duality principle)**.** *We have the inequality*

$$d^* := \sup_{\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^r_+} \inf_{x \in D} \mathcal{L}(x, \lambda, \mu) \leqslant \inf_{x \in D} \sup_{\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^r_+} \mathcal{L}(x, \lambda, \mu) := p^* \,.$$

Therefore, the solution of the dual problem is always smaller than the solution of the primal. A good mnemonic to remember this inequality is "the largest dwarf is always smaller than the smallest giant".

Definition 2.7 (Dual gap and strong duality)**.** *The dual gap of the optimization problem is the difference between the primal and dual solutions:* $p^* - p^* \geqslant 0$*. We say that there is* strong duality *if* $p^* = d^*$*.*

If the duality gap is non-zero, the solutions of the primal and the dual problems are not really related. But when there is no gap, we say that there is strong duality. The two problems are equivalent (they share the same solutions). In this case, the existence of the solutions are related with the existence of saddle points of the Lagrangian. It is worth emphasizing that strong duality does not always holds.

When is there strong duality? Sometimes the dual problem is easier to solve than the primal problem. It is then useful to know if there is strong duality.

Definition 2.8 (Slater's condition)**.** *There exists a point $x_0 \in D$ strictly feasible*

$$\exists x_0 \in D \quad \textit{such that} \quad \begin{cases} \forall 1 \leqslant i \leqslant m & h_i(x) = 0 \\ \forall 1 \leqslant j \leqslant r & g_j(x) < 0 \end{cases} .$$

Theorem 2.5 (Strong duality)**.** *If the optimization problem* (P) *is convex: i.e.,*

- *f and D are convex,*
- *the equality constraint functions h_i are affine*
- *the inequality constraint functions g_j are convex*

and if Slater's condition holds than there is strong duality ($p^* = d^*$).

In this case, we can solve the dual problem to find a solution of the primal problem.

Example 2.3. *Let us compute the dual of the following linear programing problem over the set $D = \mathbb{R}^d_+$*

$$\min_{x \geqslant 0:\ Ax = b} c^\top x \,,$$

where A is a $m \times d$ matrix and $b \in \mathbb{R}^m$. The constraints can be written as $Ax - b = 0$. We can thus define the Lagrangian $\mathcal{L} : (x, \lambda) \in \mathbb{R}^d_+ \times \mathbb{R}^{\text{'}} \to c^\top x + \lambda^\top (b - Ax)$ and re-write the primal problem with the Lagrangian

$$\begin{aligned} \min_{x \geqslant 0:\ Ax = b} c^\top x &= \min_{x \geqslant 0} \sup_{\lambda \in \mathbb{R}^m} \left\{ c^\top x + \lambda^\top (b - Ax) \right\} \\ &= \min_{x \geqslant 0} \sup_{\lambda \in \mathbb{R}^m} \left\{ b^\top \lambda + x^\top (c - A^\top \lambda) \right\} . \end{aligned}$$

By Slaters condition (the problem is convex since the objective function is convex and the equality constraints are affine), there is strong duality. We can thus swap the min *and the* sup*, we get*

$$\begin{aligned} \min_{x \geqslant 0,\ Ax = b} c^\top x &= \sup_{\lambda \in \mathbb{R}^m} \min_{x \geqslant 0} \left\{ b^\top \lambda + x^\top (c - A^\top \lambda) \right\} \\ &= \sup_{\lambda \in \mathbb{R}^m : A^\top \lambda \leqslant c} \left\{ b^\top \lambda \right\} . \end{aligned}$$

The latter is the dual formulation of the problem.

Optimality condition Now, we see conditions that play the same role as canceling the gradients for unconstrained optimization problems. These conditions will be useful to find equations to compute analytically the solution of (P).

Assume that the functions f, h_i and g_j are all differentiable. Let x^* and (λ^*, μ^*) be any primal and dual solutions and assume that there is strong duality (no duality gap). Then, we have

1. By definition x^* minimizes $\mathcal{L}(x, \lambda^*, \mu^*)$ over x. Therefore, its gradient must be canceled at x^*, i.e.,

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h(x^*) + \sum_{j=1}^{r} \mu_j \nabla g_j(x^*) = 0 \qquad \text{(KKT1)}$$

2. Since $x^* \in D^*$ and $(\lambda^*, \mu^*) \in C^*$ are feasible we have

$$\begin{aligned} h_i(x^*) &= 0 \quad \forall 1 \leqslant i \leqslant m \\ g_j(x^*) &\leqslant 0 \quad \forall 1 \leqslant j \leqslant r \\ \mu_j^* &\geqslant 0 \quad \forall 1 \leqslant j \leqslant r\,. \end{aligned} \tag{KKT2}$$

3. The *complementary condition* holds

$$\mu_j^* g_j(x^*) = 0 \quad \forall 1 \leqslant j \leqslant m\,. \tag{KKT3}$$

   Otherwise, we can improve μ^* by setting $\mu_j^* = 0$ since $g_j(x^*) \leqslant 0$ and (λ^*, μ^*) maximizes $\mathcal{L}(x^*, \lambda, \mu) = f(x^*) + \sum_i \lambda_i h_i(x^*) + \sum_j \mu_j g_j(x^*)$.

These conditions (KKT1-3) are called the Karush-Kuhn-Tucker (KKT) conditions. When the primal problem is convex (see Thm. 2.5) these conditions are also sufficient.

Theorem 2.6. *If there is strong duality then*

$$\left.\begin{array}{c} x^* \text{ is a solution of the primal problem} \\ (\lambda^*, \mu^*) \text{ is a solution of the dual problem} \end{array}\right\} \Leftrightarrow \textit{(KKT) conditions are satisfied.}$$

The KKT conditions play an important role in optimization. In some cases, it is possible to solve them analytically. Many optimization methods are conceived for solving the KKT conditions.

References

Boyd, S. and L. Vandenberghe (2004). *Convex optimization*. Cambridge university press.