Chapter 3, Part 1: Constrained Optimization

Advanced Topics in Statistical Machine Learning

Eugenio Clerico

Hilary 2026

eugenio.clerico@stats.ox.ac.uk

Constrained Optimization

- Much of machine learning requires us to perform **optimization**, e.g. minimizing the empirical risk
- This is often subject to **constraints** on the variables
- In this lecture, we will go through some essential basic results in constrained optimization
- In particular, we will be covering the concept of **duality** and showing how constrained optimization problems all have a **convex dual problem** form that can often be useful exploited
- This will form the basis for support vector machines (SVMs)

The Primal Problem

Consider a general constrained optimization problem with objective function $f : \mathbb{R}^n \to \mathbb{R}$, and m inequality and r equality constraints:

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0 && i = 1, \ldots, m \\ & h_j(x) = 0 && j = 1, \ldots r. \end{aligned}$$

- This is known as the **primal problem** and we denote its (primal) optimum value as $p^* = f\left(x^*\right)$[1]
- Any $x : g_i(x) \leq 0 \; \forall i, h_j(x) = 0 \; \forall j$ is known as a **primal feasible** point

[1] Throughout the lecture we implicitly assume the minimiser exists and is unique...

A Naive Approach

In principle, we could convert this to an unconstrained problem by instead minimizing

$$\tilde{f}(x) := f(x) + \sum_{i=1}^{m} I_{-}\left(g_i(x)\right) + \sum_{j=1}^{r} I_0\left(h_j(x)\right),$$

where

$$I_{-}(u) = \begin{cases} 0, & u \leq 0 \\ \infty, & u > 0 \end{cases}$$

$$I_0(u) = \begin{cases} 0, & u = 0 \\ \infty, & u \neq 0 \end{cases}$$

However, this is clearly impractical from the perspective of performing the optimization

The Lagrangian

The Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^r \to \mathbb{R}$ is defined as

$$L(x, \alpha, \beta) := f(x) + \sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{j=1}^{r} \beta_j h_j(x).$$

where the vectors $\alpha \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^r$ are our Lagrange multipliers, sometimes known as **dual variables**

Now it turns out that if $\alpha \succeq 0$,[2] then the Lagrangian is a lower bound on $\tilde{f}(x)$, that is

$$L(x, \alpha, \beta) \leq \tilde{f}(x) \quad \forall x \in \mathbb{R}^n, \beta \in \mathbb{R}^r, \alpha \in \mathbb{R}^m : \alpha \succeq 0$$

[2] By this we mean that each $\alpha_i \geq 0$

Lower Bound Interpretation of the Lagrangian

Note that we have the following

$$\sup_{\alpha_i \in \mathbb{R}^+} \alpha_i g_i(x) = \begin{cases} 0, & g_i(x) \leq 0 \\ \infty, & g_i(x) > 0 \end{cases} = I_-(g_i(x))$$

$$\sup_{\beta_j \in \mathbb{R}} \beta_j h_j(x) = \begin{cases} 0, & h_j(x) = 0 \\ \infty, & h_j(x) \neq 0 \end{cases} = I_0(h_j(x))$$

And thus

$$\begin{aligned} \tilde{f}(x) &= f(x) + \sum_{i=1}^{m} \sup_{\alpha_i \in \mathbb{R}^+} \alpha_i g_i(x) + \sum_{j=1}^{r} \sup_{\beta_j \in \mathbb{R}} \beta_j h_j(x) \\ &= \sup_{\alpha_i \in \mathbb{R}^+,\ \beta_j \in \mathbb{R},\ \forall i,j} f(x) + \sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{j=1}^{r} \beta_j h_j(x) \\ &= \sup_{\alpha \succeq 0, \beta} L(x, \alpha, \beta) \end{aligned}$$

The Dual Problem

We now have that the primal problem can be solved using the unconstrained minimax problem

$$p^* = \inf_{x \in \mathcal{D}} \tilde{f}(x) = \inf_{x \in \mathcal{D}} \sup_{\alpha \succeq 0, \beta} L(x, \alpha, \beta)$$

The so-called **dual form** of the problem **switches the order** of these optimizations:

$$d^* = \sup_{\alpha \succeq 0, \beta} \inf_{x \in \mathcal{D}} L(x, \alpha, \beta)$$

The **max-min inequality** now guarantees that $d^* \leq p^*$. This result is known as **weak duality**

Proof for Weak Duality

$$
\begin{aligned}
& \forall x, \alpha, \beta, && \inf_{x'} L(x', \alpha, \beta) \leq L(x, \alpha, \beta) \\
\implies \quad & \forall x, \alpha, \beta && \inf_{x'} L(x', \alpha, \beta) \leq \sup_{\alpha' \succeq 0, \beta'} L(x, \alpha', \beta') \\
\implies \quad & \forall x && \sup_{\alpha \succeq 0, \beta} \inf_{x'} L(x', \alpha, \beta) \leq \sup_{\alpha \succeq 0, \beta} L(x, \alpha, \beta) \\
\implies \quad & && \sup_{\alpha \succeq 0, \beta} \inf_{x} L(x, \alpha, \beta) \leq \inf_{x} \sup_{\alpha \succeq 0, \beta} L(x, \alpha, \beta)
\end{aligned}
$$

The Lagrange Dual Function

We can more formally define the dual problem by first defining the **Lagrange dual function** (or just "dual function") as

$$\phi(\alpha, \beta) = \inf_{x \in \mathcal{D}} L(x, \alpha, \beta)$$

A **dual feasible** pair (α, β) is a pair where $\alpha \succeq 0$ and the Lagrangian is bounded from below, i.e. $\phi(\alpha, \mu) > -\infty$

The **dual problem** is now

$$\begin{aligned} &\text{maximize} && \phi(\alpha, \beta) \\ &\text{subject to} && \alpha \succeq 0 \end{aligned}$$

We thus find the **largest lower bound** to original (primal) problem

$$d^* = \sup_{\alpha \succeq 0, \beta} \phi(\alpha, \beta)$$

noting that $\phi(\alpha, \beta) \leq p^* \ \forall \alpha, \beta$

Why Use the Dual?

- In general, $\tilde{f}(x)$ is very difficult to work with as it equals ∞ for any input that does not satisfy the constraints
- If we can calculate, $\phi(\alpha, \beta)$ we can exploit the fact that it is concave: it is a pointwise infimum of affine functions of (α, β)

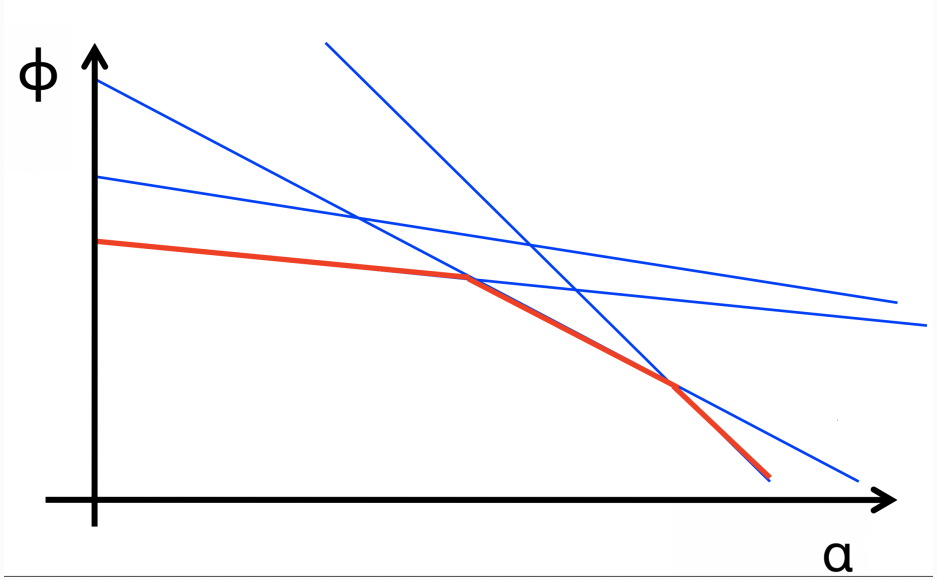
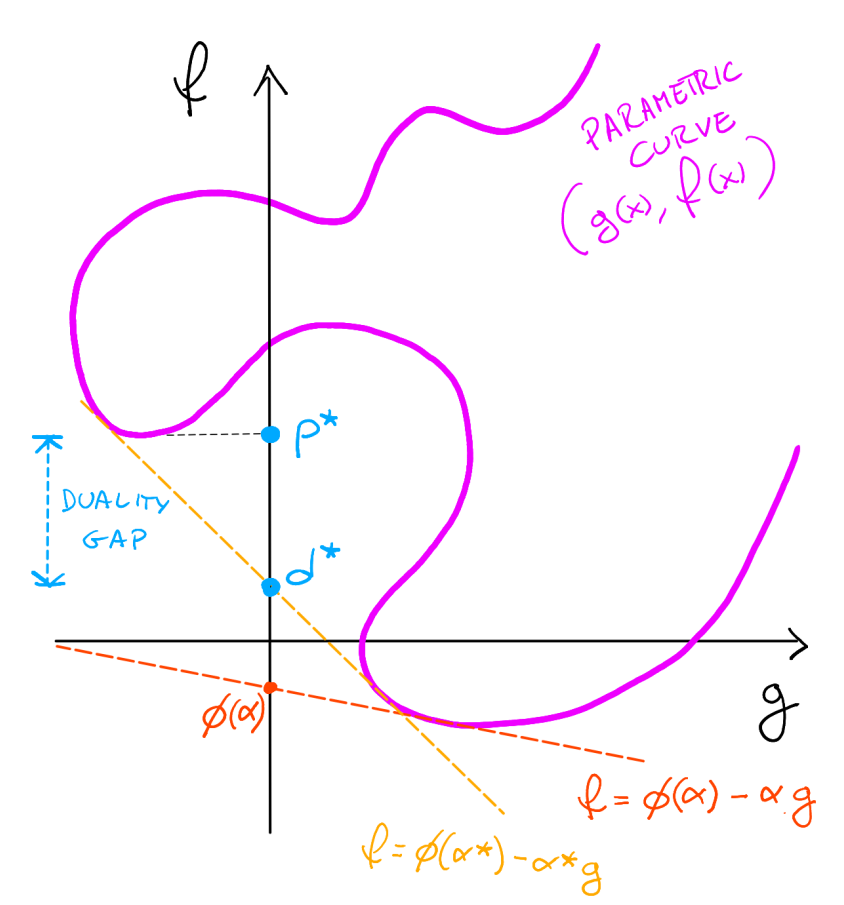


Figure 1: $\phi(\alpha)$ is concave as a the infimum of a family of affine functions

Duality gap

The difference $p^* - d^*$ is called the **optimal duality gap**.



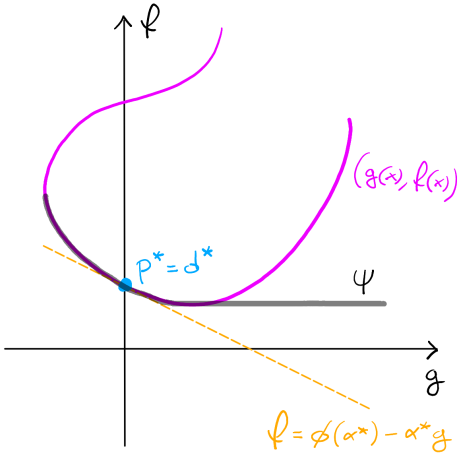
Case of a single ineq constraint
$\alpha \geq 0$ is real
no eq constraints (no β, no h).

Recall:
$p^\star = \inf_x \tilde{f}(x)$
$d^\star = \sup_\alpha \phi(\alpha)$

$$\begin{aligned}\phi(\alpha) &= \inf L(x, \alpha) \\ &= \inf_x \big(f(\alpha) + \alpha g(x)\big) \\ &= \sup \big\{c \,:\, f(x) + \alpha g(x) \geq c\,,\ \forall x\big\} \\ &= \sup \big\{c \,:\, f(x) \geq c - \alpha g(x)\,,\ \forall x\big\}\end{aligned}$$

Strong Duality

The *bump* around the vertical axis was the cause of the duality gap in the previous slide. If there were no bump we might be able to have $d^\star = p^\star$ (**strong duality**)! A condition preventing the bump is that the *left-lower boundary* ψ of the curve (g, f) is **convex**. This is the case if f and g are convex. Then, the supporting hyperplane theorem is the tool to use to show strong duality. This is the reasoning behind the proof of next slide's sufficient condition.



If f and g are convex...
Let $\psi(u) = \inf\{f(x) : g(x) \leq u\}$
We can show that ψ is convex, hence no bump!
Let x_1, x_2, u_1, u_2 s.t. $g(x_1) \leq u_1$, $g(x_2) \leq u_2$
Fix $\theta \in [0, 1]$, then:

$$g(\theta x_1 + (1-\theta)x_2) \leq \theta u_1 + (1-\theta)u_2 \quad \text{by conv of } g$$
$$\psi(\theta u_1 + (1-\theta)u_2) \leq f(\theta x_1 + (1-\theta)x_2) \quad \text{by def of } \psi$$
$$\leq \theta f(x_1) + (1-\theta)f(x_2) \quad \text{by conv of } f$$

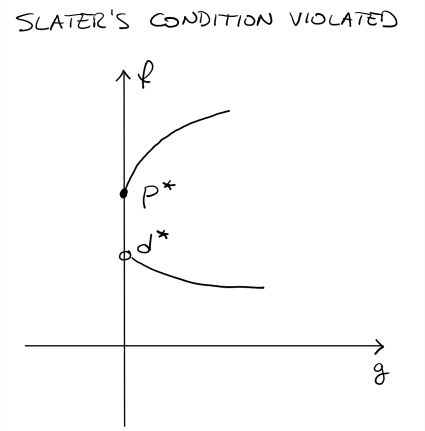
Taking inf for x_1 s.t. $g(x_1) \leq u_1$ and x_2 s.t. $g(x_2) \leq u_2$:

$$\psi(\theta u_1 + (1-\theta)u_2) \leq \theta\psi(u_1) + (1-\theta)\psi(u_2)$$

Strong Duality sufficient condition

- **strong duality**:
  $d^* = \sup_{\alpha \succeq 0, \beta} \inf_x L(x, \alpha, \beta) = \inf_x \sup_{\alpha \succeq 0, \beta} L(x, \alpha, \beta) = p^*$.
- Most common (but not only) **sufficient** condition is for **both**:

1. Primal problem is **convex**: f is convex, each $g_i(x)$ is a convex function and each $h_j(x)$ is affine (i.e. $h_j(x) = a_j^T x - b_j = 0$, such that we can represent the equality constraints as $Ax = b$)
2. **Slater's condition**: there exists a **strictly feasible** $x : f(x) < \infty; g_i(x) < 0 \ \forall i = 1, \ldots, m; h_j(x) = 0 \ \forall j = 1, \ldots, r$



Complementary Slackness

- Recall: the dual problem is generally **easier** to directly solve than the **primal**! If strong duality applies, we solve the dual and hence get a solution for the primal!
- When strong duality holds, we can use the dual problem to find both p^* and x^*, i.e. the solution of our original problem
- It also means that a condition called **complementary slackness** holds at the optimum: denoting $(\alpha^*, \beta^*) = \arg\max_{\alpha \succeq 0, \beta} \phi(\alpha, \beta)$, we have

$$\alpha_i^* g_i(x^*) = 0 \ \forall i$$

and thus

$$\begin{aligned} \alpha_i^* > 0 \quad &\implies \quad g_i(x^*) = 0, \\ g_i(x^*) < 0 \quad &\implies \quad \alpha_i^* = 0. \end{aligned}$$

Proof for Complementary Slackness

Denote by x^* the optimum solution of the original problem, and by (α^*, β^*) the solutions to the dual. Then strong duality implies

$$\begin{aligned} f(x^*) &= \phi(\alpha^*, \beta^*) \\ &= \inf_x \left(f(x) + \sum_{i=1}^{m} \alpha_i^* g_i(x) + \sum_{i=1}^{r} \beta_i^* h_i(x) \right) \\ &\leq f(x^*) + \sum_{i=1}^{m} \alpha_i^* g_i(x^*) + \sum_{i=1}^{r} \beta_i^* \underbrace{h_i(x^*)}_{=0} \\ &= f(x^*) + \sum_{i=1}^{m} \alpha_i^* g_i(x^*). \end{aligned}$$

Now as $\alpha_i^* \geq 0$ and $g_i(x^*) \leq 0$, none of the terms in the sum can be positive, so the inequality can only hold if each term is exactly zero, i.e. $\alpha_i^* g_i(x^*) = 0 \ \forall i$.

Optimality

If strong duality holds and the Lagrangian is differentiable, then $\nabla_x L(x, \alpha^*, \beta^*)|_{x=x^*} = 0$ as otherwise it would be possible to achieve a better dual solution by moving down the gradient

Using the shorthand $\nabla_x f(x^*) = \nabla_x f(x)|_{x=x^*}$ we thus have

$$\nabla_x f(x^*) + \sum_{i=1}^{m} \alpha_i^* \nabla_x g_i(x^*) + \sum_{i=1}^{r} \beta_i^* \nabla_x h_i(x^*) = 0$$

at the optimum if strong duality holds

The KKT Conditions

Combining everything together now gives the **KKT** conditions for a optimality of a tuple (x, α, β) if

$$\nabla_x f(x) + \sum_{i=1}^{m} \alpha_i \nabla_x g_i(x) + \sum_{i=1}^{r} \beta_i \nabla_x h_i(x) = 0$$

$$g_i(x) \leq 0,\ i = 1, \ldots, m,$$

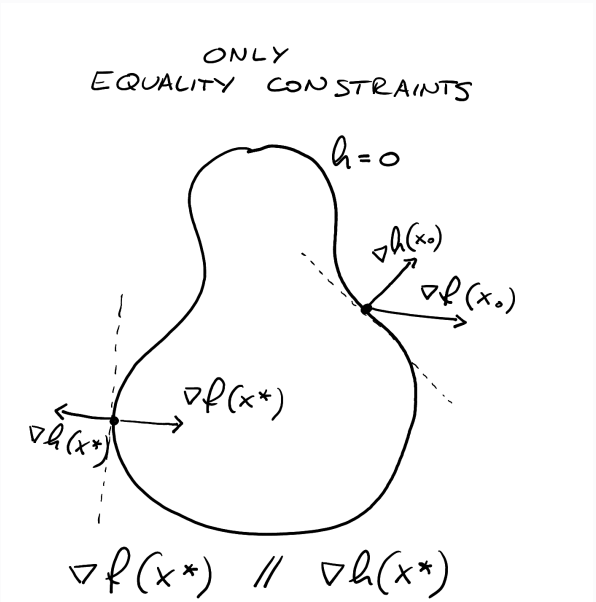
$$h_i(x) = 0,\ i = 1, \ldots, r,$$

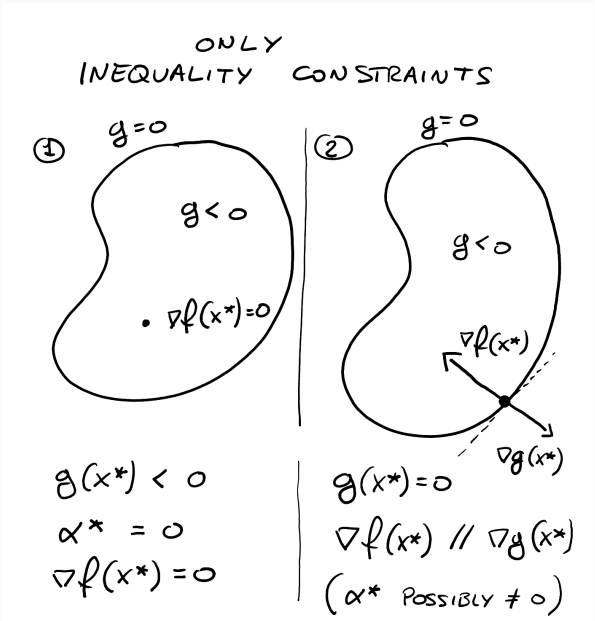
$$\alpha_i \geq 0,\ i = 1, \ldots, m,$$

$$\alpha_i g_i(x) = 0,\ i = 1, \ldots, m.$$

If a minimiser x^* exists, the KKT conditions are **sufficient and necessary** for global optimality if our problem is convex, satisfies Slater's condition, and has differentiable objective and constraint functions.

Some intuition





$$\nabla f(x^*) + \beta^* \nabla h(x^*) = 0 \qquad \nabla f(x^*) + \alpha^* \nabla g(x^*) = 0$$

The gradient is orthogonal to the level sets. If the gradient of f in x has a non-null component along the tangent to the curve $h = 0$, then x cannot be a stationary point. So, for equality constraint the gradients of f and h are parallel.

The same applies on the boundary of $g \leq 0$. But for a point to be stationary in the interior of $g \leq 0$, we need $\nabla f = 0$!

Recap

- Directly solving minimization problems with inequality (and equality) constraints is typically challenging
- All such problems have a **convex dual form** where we **maximize a lower bound** on the optimum with respect to the Lagrange multipliers (aka dual variables)
- If the dual form is itself tractable this can form a means of (approximately) solving the original optimization problem
- Many convex problems exhibit **strong duality**, such that the primal and dual problems have the same optima
- In such case solving the dual gives us an unconstrained optimisation problem $\inf_x L(x, \alpha^\star, \beta^\star)$
- We can use the KKT conditions to confirm global optimality in such cases

Further Reading

- Chapter 5 of Stephen P Boyd and Lieven Vandenberghe. **Convex optimization.** Cambridge university press, 2004, `https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf`