Chapter 3, Part 1: Constrained Optimization

Advanced Topics in Statistical Machine Learning

Tom Rainforth

Hilary 2024

rainforth@stats.ox.ac.uk

Constrained Optimization

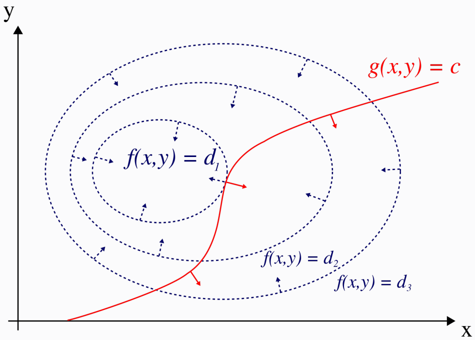
- Much of machine learning requires us to perform **optimization**, e.g. minimizing the empirical risk
- This is often subject to **constraints** on the variables
- In this lecture, we will go through some essential basic results in constrained optimization
- In particular, we will be covering the concept 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)

Lagrange Multipliers

Consider the following optimization problem:

$$\text{minimize } f(x) \quad \text{subject to } h(x) = 0$$

At the optimum x^*, $\nabla_x f(x)|_{x=x^*} = -\nu \nabla_x h(x)|_{x=x^*}$ for some scalar ν, known as a **Lagrange Multiplier**



[Source: Wikipedia]

This is equivalent to finding the saddle points[1] of the **Lagrangian**

$$L(x, \nu) := f(x) + \nu h(x)$$

by noting that

$$\nabla_{x,\nu} L(x, \nu) = 0 \iff \begin{cases} \nabla_x f(x) = -\nu \, \nabla_x h(x) \\ h(x) = 0 \end{cases}$$

Unfortunately, this no longer necessarily applies in the more general case where we also have inequality constraints

[1] Note these must be saddle points, not minima or maxima, as $L(x, \nu)$ is constant over ν for all $x : h(x) = 0$.

The Primal Problem

Consider a general constrained optimization problem with objective function $f_0 : \mathbb{R}^n \to \mathbb{R}$, and m inequality and r equality constraints:

$$
\begin{aligned}
\text{minimize} \quad & f_0(x) & \\
\text{subject to} \quad & f_i(x) \leq 0 & \qquad i = 1, \ldots, m \\
& h_j(x) = 0 & \qquad j = 1, \ldots r.
\end{aligned}
$$

- This is known as the **primal problem** and we denote its (primal) optimum value as $p^* = f_0\left(x^*\right)$
- Any $x : f_i(x) \leq 0 \; \forall i, h_j(x) = 0 \; \forall j$ is known as a **primal feasible** point

A Naive Approach

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

$$\tilde{f}(x) := f_0(x) + \sum_{i=1}^{m} I_{-}\left(f_i(x)\right) + \sum_{j=1}^{r} I_0\left(h_j(x)\right),$$

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 still defined in this setting, namely

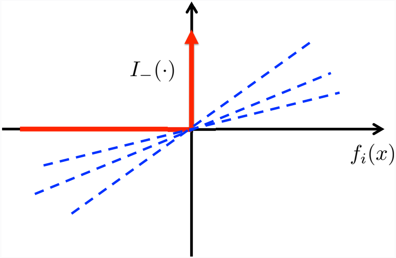
$$L(x, \lambda, \nu) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{r} \nu_j h_j(x).$$

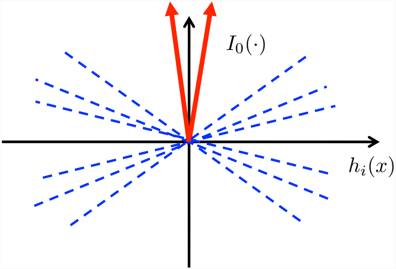
where the vectors $\lambda \in \mathbb{R}^m$ and $\nu \in \mathbb{R}^r$ are our Lagrange multipliers, sometimes known as **dual variables**

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

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

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





Different blue lines represent different values of λ_i and ν_i for left and right plots respectively. We see that regardless of these values, we have a lower bound on $I_-(u)$ and $I_0(u)$ respectively.

Lower Bound Interpretation of the Lagrangian

More concretely we have the following

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

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

And thus

$$\begin{aligned} \tilde{f}(x) &= f_0(x) + \sum_{i=1}^{m} \sup_{\lambda_i \in \mathbb{R}^+} \lambda_i f_i(x) + \sum_{j=1}^{r} \sup_{\nu_j \in \mathbb{R}} \nu_j h_j(x) \\ &= \sup_{\lambda_i \in \mathbb{R}^+,\ \nu_j \in \mathbb{R},\ \forall i,j} f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{r} \nu_j h_j(x) \\ &= \sup_{\lambda \succeq 0, \nu} L(x, \lambda, \nu) \end{aligned}$$

The Dual Problem

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

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

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

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

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

Proof for Weak Duality

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

The Lagrange Dual Function

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

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

A **dual feasible** pair (λ, ν) is a pair where $\lambda \succeq 0$ and the Lagrangian is bounded from below, i.e. $g(\lambda, \mu) > -\infty$

The **dual problem** is now

$$\begin{aligned} \text{maximize} \qquad & g(\lambda, \nu) \\ \text{subject to} \qquad & \lambda \succeq 0 \end{aligned}$$

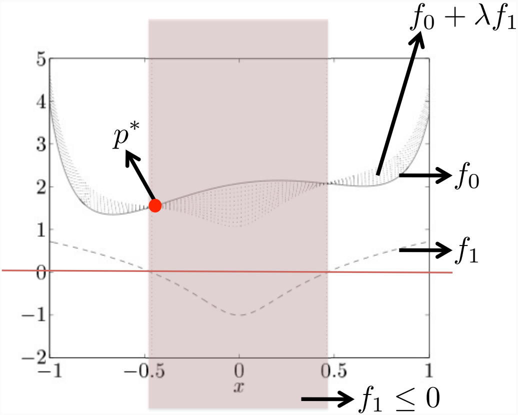
We thus find the **largest lower bound** to original (primal) problem

$$d^* = \sup_{\lambda \succeq 0, \nu} g(\lambda, \nu)$$

noting that $g(\lambda, \nu) \leq p^* \ \forall \lambda, \nu$

Lower Bound Interpretation of the Lagrangian

Simplest example: minimize $L(x, \lambda) = f_0(x) + \lambda f_1(x)$ w.r.t. x

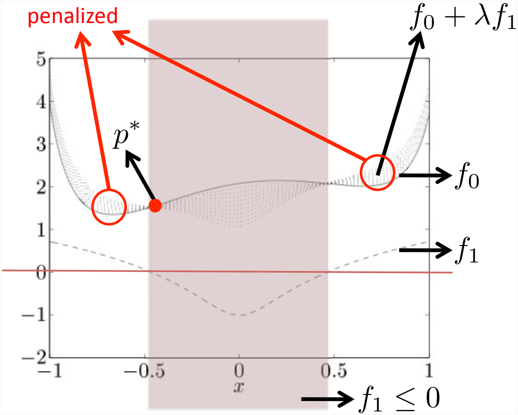


- Primal problem: $p^* = \inf_x \sup_{\lambda \succeq 0, \nu} L(x, \lambda, \nu)$
- Dual problem $d^* = \sup_{\lambda \succeq 0, \nu} g(\lambda, \nu)$ where $g(\lambda, \nu) = \inf_x L(x, \lambda, \nu)$
- p^* is minimum f_0 **in constrained set**

Figure from Boyd and Vandenberghe

Lower Bound Interpretation of the Lagrangian

Simplest example: minimize $L(x, \lambda) = f_0(x) + \lambda f_1(x)$ w.r.t. x

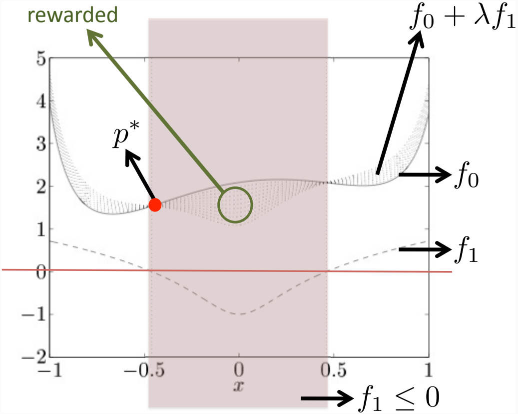


- Primal problem: $p^* = \inf_x \sup_{\lambda \succeq 0, \nu} L(x, \lambda, \nu)$
- Dual problem $d^* = \sup_{\lambda \succeq 0, \nu} g(\lambda, \nu)$ where $g(\lambda, \nu) = \inf_x L(x, \lambda, \nu)$
- p^* is minimum f_0 **in constrained set**

Figure from Boyd and Vandenberghe

Lower Bound Interpretation of the Lagrangian

Simplest example: minimize $L(x, \lambda) = f_0(x) + \lambda f_1(x)$ w.r.t. x

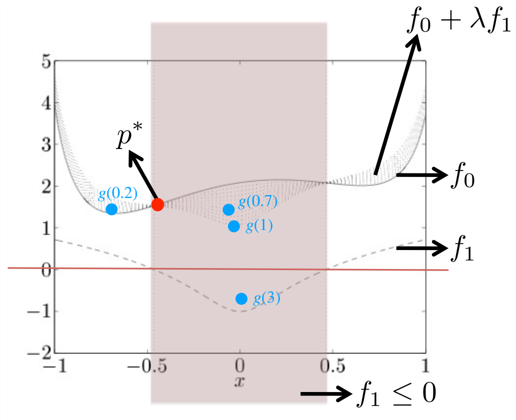


- Primal problem: $p^* = \inf_x \sup_{\lambda \succeq 0, \nu} L(x, \lambda, \nu)$
- Dual problem $d^* = \sup_{\lambda \succeq 0, \nu} g(\lambda, \nu)$ where $g(\lambda, \nu) = \inf_x L(x, \lambda, \nu)$
- p^* is minimum f_0 **in constrained set**

Figure from Boyd and Vandenberghe

Lower Bound Interpretation of the Lagrangian

Simplest example: minimize $L(x, \lambda) = f_0(x) + \lambda f_1(x)$ w.r.t. x



- Primal problem: $p^* = \inf_x \sup_{\lambda \succeq 0, \nu} L(x, \lambda, \nu)$
- Dual problem $d^* = \sup_{\lambda \succeq 0, \nu} g(\lambda, \nu)$ where $g(\lambda, \nu) = \inf_x L(x, \lambda, \nu)$
- p^* is minimum f_0 **in constrained set**

Figure from Boyd and Vandenberghe

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, $g(\lambda, \nu)$ we can exploit the fact that it is concave: it is a pointwise infimum of affine functions of (λ, ν)

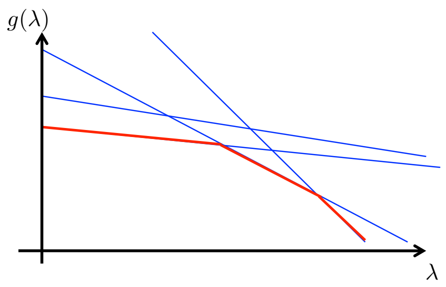


Figure 1: Example: Lagrangian with one inequality constraint, $L(x, \lambda) = f_0(x) + \lambda f_1(x)$, where x here can take one of four values

Strong Duality and Constraint Qualifications

- The difference $p^* - d^*$ is called the **optimal duality gap**.
- In some cases, the optimal duality gap is zero, i.e.

$$d^* = \sup_{\lambda \succeq 0, \nu} \inf_x L(x, \lambda, \nu) = \inf_x \sup_{\lambda \succeq 0, \nu} L(x, \lambda, \nu) = p^*.$$

  This is known as **strong duality**.
- The conditions under which this happens are known as **constraint qualifications**
- Most common (but not only) **sufficient** condition is for **both** the following to hold:
  1. Primal problem is **convex**: each $f_i(x)$ 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** input, i.e.

$$\exists x : f_0(x) < \infty; f_i(x) < 0 \ \forall i = 1, \ldots, m; h_j(x) = 0 \ \forall j = 1, \ldots, r$$

Complementary Slackness

- 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 **complimentary slackness** holds at the optimum: denoting $(\lambda^*, \nu^*) = \arg\max_{\lambda \succeq 0, \nu} g(\lambda, \nu)$, we have

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

  and thus

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

Proof for Complementary Slackness

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

$$\begin{aligned} f_0(x^*) &= g(\lambda^*, \nu^*) \\ &= \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^r \nu_i^* h_i(x) \right) \\ &\leq f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^r \nu_i^* \underbrace{h_i(x^*)}_{=0} \\ &= f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*). \end{aligned}$$

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

Optimality

If strong duality holds and the Lagrangian is differentiable, then $\nabla_x L(x, \lambda^*, \nu^*)|_{x=x^*} = 0$ 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

$$\nabla_x f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla_x f_i(x^*) + \sum_{i=1}^{r} \nu_i^* \nabla_x h_i(x^*) = 0$$

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, λ, ν) if

$$\begin{aligned} \nabla_x f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla_x f_i(x) + \sum_{i=1}^{r} \nu_i \nabla_x h_i(x) = 0 \\ f_i(x) \leq 0,\ i = 1, \ldots, m, \\ h_i(x) = 0,\ i = 1, \ldots, r, \\ \lambda_i \geq 0,\ i = 1, \ldots, m, \\ \lambda_i f_i(x) = 0,\ i = 1, \ldots, m. \end{aligned}$$

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.

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
- 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