Lagrange duality

In the previous lecture, we derived the KKT conditions for minimizing a convex function under convex inequality constraints and/or affine equality constraints. This involved introducing additional variables $\boldsymbol{\nu}$ and $\boldsymbol{\lambda}$. Here we will provide an alternative perspective on these problems and provide a bit more intuition as to how to interpret these additional variables.

The Lagrangian

We again consider an optimization program of the form

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}}\ & f(\boldsymbol{x}) \\ \text{subject to}\ \ & g_m(\boldsymbol{x}) \leq 0, \quad m = 1, \ldots, M \\ & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \end{aligned} \tag{1}$$

We will focus on the case where the objective function f and the inequality constraints g_m are convex, and the equality constraints are affine (note that for equality constraints, convexity is equivalent to being affine). However, in general much of what we have to say applies to arbitrary (nonconvex) problems as well so we will be clear when we are or are not assuming convexity. We will take the domain of all of the g_m to be all of $\mathbb{R}^N$ below; this just simplifies the exposition, we can easily replace this with the intersections of the dom g_m. We will also assume that the feasible set

$$\mathcal{C} = \{\boldsymbol{x} \ : \ g_m(\boldsymbol{x}) \leq 0 \ m = 1, \ldots, M, \ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$$

is non-empty and a subset $\mathbb{R}^N$.

The **Lagrangian** takes the constraints in the program above and integrates them into the objective function. Specifically, the Lagrangian associated with (1) is

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f(\boldsymbol{x}) + \sum_{m=1}^{M} \lambda_m g_m(\boldsymbol{x}) + \boldsymbol{\nu}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}).$$

For reasons that will become clearer below, the $\boldsymbol{x}$ above are referred to as **primal variables**, and the $\boldsymbol{\lambda}, \boldsymbol{\nu}$ as either **dual variables** or **Lagrange multipliers**.

The Lagrangian allows us to transform the *constrained* optimization problem in (1) into an *unconstrained* one. Specifically, suppose for the moment that we are interested in a problem of the form in (1) but without equality constraints. Consider the problem given by

$$\underset{\boldsymbol{x} \in \mathbb{R}^N}{\text{minimize}} \;\; f(\boldsymbol{x}) + \sum_{m=1}^{M} \lambda_m g_m(\boldsymbol{x}). \tag{2}$$

To get some intuition, suppose that we set the $\lambda_1, \ldots, \lambda_M$ to be very large (positive) numbers. In this case, violating any of the constraints (allowing $g_m(\boldsymbol{x}) > 0$) will result in a very large penalty being added to the objective function, so that by setting the corresponding λ_m to be large we will eventually guarantee that the resulting solution will satisfy the desired constraints.

The problem here is that large values of λ_m not only avoid the setting where $g_m(\boldsymbol{x}) > 0$, but actually encourages $g_m(\boldsymbol{x}) \ll 0$ (since we can potentially benefit by not just satisfying the constraints but by exceeding them by a large margin).

This raises a natural question: can we set $\boldsymbol{\lambda}$ so that the solution to the unconstrained problem (2) is the same as the constrained problem (1)? Here we will provide an answer in the case where the objective function f and the constraints $g_1, \ldots, g_M$ are both convex and differentiable.

Suppose that $\boldsymbol{x}^\star$ is a solution to the constrained problem (1) (again without the equality constraints). If we want $\boldsymbol{x}^\star$ to be a solution to (2), then a necessary and sufficient condition is for the $\boldsymbol{\lambda}$ to obey

$$\nabla L(\boldsymbol{x}^\star, \boldsymbol{\lambda}) = \nabla f(\boldsymbol{x}^\star) + \sum_{m=1}^{M} \lambda_m \nabla g_m(\boldsymbol{x}^\star) = \mathbf{0}. \tag{3}$$

At this point you might want to compare (3) with condition (K4) from the second two examples from the previous lecture. (Hint: they are the same!)

If we knew $\boldsymbol{x}^\star$ already, finding a $\boldsymbol{\lambda}$ that would make the unconstrained and constrained problems equivalent (meaning that they both have the same solution $\boldsymbol{x}^\star$) would just amount to finding a $\boldsymbol{\lambda}$ such that (3) holds. Unfortunately, this might not seem to be particularly useful since $\boldsymbol{x}^\star$ is what we are trying to find to begin with.

To see how we might compute a $\boldsymbol{\lambda}$ that makes the unconstrained and constrained problems equivalent, we will need to begin our first exploration of one of the deepest and most important ideas of optimization: **duality**.

The Lagrange dual function

We can think of the unconstrained optimization problem (2) as actually representing a family of different optimization problems (depending on $\boldsymbol{\lambda}$). For any fixed $\boldsymbol{\lambda}$, imagine solving (2) and computing the minimal value of the objective function – we can think of this as actually defining a function that maps $\boldsymbol{\lambda} \in \mathbb{R}^M$ to $\mathbb{R}$. Specifically, returning to the case where we have both inequality and equality constraints, the **Lagrange dual function** $d(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is the minimum[1] of the Lagrangian over all $\boldsymbol{x} \in \mathbb{R}^N$:

$$\begin{aligned} d(\boldsymbol{\lambda}, \boldsymbol{\nu}) &= \inf_{\boldsymbol{x} \in \mathbb{R}^N} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \\ &= \inf_{\boldsymbol{x} \in \mathbb{R}^N} \left(f(\boldsymbol{x}) + \sum_{m=1}^{M} \lambda_m g_m(\boldsymbol{x}) + \boldsymbol{\nu}^{\mathrm{T}} (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \right). \end{aligned}$$

Note that since the dual is the pointwise infimum of a family of affine functions in $\boldsymbol{\lambda}, \boldsymbol{\nu}$, the Lagrange dual function is **always concave**, regardless of whether or not f, g_m, and equality constraints are convex. While we will not stress this much here, this is a remarkable fact and can be very useful when dealing with nonconvex problems.

A key fact about the dual function is that it can provide a lower bound on the optimal value of the original program. In the discussion below, we assume throughout that $\boldsymbol{\nu}$ and $\boldsymbol{\lambda} \geq 0$ are arbitrary. Our main claim is that if $p^\star = f(\boldsymbol{x}^\star)$ is the optimal value for (1),[2] then we have

$$d(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^\star.$$

[1] We are writing inf instead of min here since we in general cannot be sure that the minimum exists. It very well may be that $d(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is $-\infty$.

[2] We use $p^\star$ instead of $f^\star$ to indicate the optimal value of the *primal* problem, which we will soon be opposing to the optimal value of the *dual* problem.

This is very easy to show. Specifically, for any feasible point $\boldsymbol{x}'$, we must have $g_m(\boldsymbol{x}') \leq 0$ for all m and also $\boldsymbol{A}\boldsymbol{x}' = \boldsymbol{b}$, and hence

$$\sum_{m=1}^{M} \lambda_m g_m(\boldsymbol{x}') + \boldsymbol{\nu}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x}' - \boldsymbol{b}) \leq 0.$$

From this we have that

$$L(\boldsymbol{x}', \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f(\boldsymbol{x}'),$$

meaning that

$$d(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\boldsymbol{x} \in \mathbb{R}^N} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq L(\boldsymbol{x}', \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f(\boldsymbol{x}').$$

Since this holds for all feasible $\boldsymbol{x}'$, including the minimizer of (1), we have $d(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^\star$.

The (Lagrange) dual problem

Given that $d(\boldsymbol{\lambda}, \boldsymbol{\nu})$ provides a lower bound on $p^\star$, if you wanted to get an idea of what $p^\star$ looks like (for example, to see if you are close to convergence), it is natural to see how large you can make this lower bound. This gives rise to what we call the **(Lagrange) dual problem** of (1):

$$\underset{\boldsymbol{\lambda}, \boldsymbol{\nu}}{\text{maximize}} \;\; d(\boldsymbol{\lambda}, \boldsymbol{\nu}) \quad \text{subject to} \quad \boldsymbol{\lambda} \geq \boldsymbol{0}. \tag{4}$$

The dual optimal value $d^\star$ is

$$d^\star = \sup_{\boldsymbol{\lambda} \geq \boldsymbol{0}, \boldsymbol{\nu}} \; d(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \sup_{\boldsymbol{\lambda} \geq \boldsymbol{0}, \boldsymbol{\nu}} \; \inf_{\boldsymbol{x} \in \mathbb{R}^N} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}).$$

Since $d(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^\star$, we know that

$$d^\star \leq p^\star.$$

The quantity $p^\star - d^\star$ is called the **duality gap**. If $p^\star = d^\star$, then we say that (1) and (4) exhibit **strong duality**.

We will soon discuss when strong duality holds, but first, why is it important? Suppose that $\boldsymbol{x}^\star$ is a solution to the original constrained problem (1) – which we will call the **primal problem** to distinguish it from the dual problem – and suppose that $(\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star)$ is a solution to the dual problem (4). It turns out that if we have strong duality, then $(\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star)$ is exactly what we need to make $\boldsymbol{x}^\star$ the solution to the unconstrained problem (2).

To see why, note that if we have strong duality then

$$\begin{aligned} f(\boldsymbol{x}^\star) &= d(\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star) \\ &= \inf_{\boldsymbol{x} \in \mathbb{R}^N} \left(f(\boldsymbol{x}) + \sum_{m=1}^{M} \lambda_m^\star g_m(\boldsymbol{x}) + \boldsymbol{\nu}^{\star \mathrm{T}} (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \right) \\ &\leq f(\boldsymbol{x}^\star) + \sum_{m=1}^{M} \lambda_m^\star g_m(\boldsymbol{x}^\star) + \boldsymbol{\nu}^{\star \mathrm{T}} (\boldsymbol{A}\boldsymbol{x}^\star - \boldsymbol{b}) \\ &\leq f(\boldsymbol{x}^\star). \end{aligned} \tag{5}$$

where the last inequality follows from the facts that we must have $\lambda_m^\star \geq 0$ and $g_m(\boldsymbol{x}^\star) \leq 0$ and that $\boldsymbol{A}\boldsymbol{x}^\star = \boldsymbol{b}$. Looking at this entire chain of inequalities, where the first and last term are both $f(\boldsymbol{x}^\star)$, means that

$$f(\boldsymbol{x}^\star) = \min_{\boldsymbol{x} \in \mathbb{R}^N} L(\boldsymbol{x}, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star) = L(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star).$$

In words, a solution to the primal problem $\boldsymbol{x}^\star$ is also a minimizer of $L(\boldsymbol{x}, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star)$.

Strong duality and the KKT conditions

We can also show that if the primal optimal value $p^\star = f(\boldsymbol{x}^\star)$ and the dual optimal value $d^\star = d(\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star)$ are equal

$$p^\star = d^\star,$$

then the KKT conditions hold for $\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star$.

Recall the string of inequalities (5) above, and note that since we started out and ended up with the same thing, we can replace the last two inequalities with equality. Since $\boldsymbol{x}^\star$ is feasible $\boldsymbol{A}\boldsymbol{x}^\star = \boldsymbol{b}$ and $g_m(\boldsymbol{x}^\star) \leq 0$, and so

$$\lambda_m^\star g_m(\boldsymbol{x}^\star) = 0, \quad m = 1, \ldots, M.$$

Also, since we know $\boldsymbol{x}^\star$ is a minimizer of $L(\boldsymbol{x}, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star)$ (second equality in (5)), which is an unconstrained convex function (with $\boldsymbol{\lambda}, \boldsymbol{\nu}$ fixed), the gradient with respect to $\boldsymbol{x}$ must be zero:

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star) = \nabla f(\boldsymbol{x}^\star) + \sum_{m=1}^{M} \lambda_m^\star \nabla g_m(\boldsymbol{x}^\star) + \boldsymbol{A}^{\mathrm{T}} \boldsymbol{\nu}^\star = \boldsymbol{0}.$$

Thus strong duality immediately leads to the KKT conditions holding at the solution.

We can also go the other way. If you can find $\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star$ that obey the KKT conditions, not only do you know that you have a primal optimal point on your hands, but also we have strong duality (and $\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star$ are dual optimal). For if KKT holds,

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star) = \boldsymbol{0},$$

meaning that $\boldsymbol{x}^\star$ is a minimizer of $L(\boldsymbol{x}, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star)$, i.e.

$$L(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star) \leq L(\boldsymbol{x}, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star),$$

thus

$$\begin{aligned} d(\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star) &= L(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star) \\ &= f(\boldsymbol{x}^\star) + \sum_{m=1}^{M} \lambda_m^\star g_m(\boldsymbol{x}^\star) + \boldsymbol{\nu}^{\star \mathrm{T}}(\boldsymbol{A}\boldsymbol{x}^\star - \boldsymbol{b}) \\ &= f(\boldsymbol{x}^\star), \quad \text{(by KKT)}, \end{aligned}$$

and we have strong duality.

The upshot of this is that the conditions for strong duality are essentially the same as those under which the KKT conditions are necessary. As we discussed in the last set of notes, perhaps the most commonly encountered such condition is *Slater's condition*. Informally, Slater's condition simply says that the feasible set has a non-empty relative interior. We re-state this condition here as

Slater's condition: There exists at least one $\bar{\boldsymbol{x}}$ such that for each inequality constraint g_m, either g_m is affine or

$$g_m(\bar{\boldsymbol{x}}) < 0.$$

That is, there is an $\bar{\boldsymbol{x}}$ that is *strictly* feasible for all non-affine constraints.

Nearly all of the optimization problems that we will encounter in this course will satisfy this condition. There are, however, convex problems that do not. As a simple example, let $\boldsymbol{p}_1 = [1, 0]^{\mathrm{T}}$ and $\boldsymbol{p}_2 = [-1, 0]^{\mathrm{T}}$ and consider the constraints

$$\begin{aligned} g_1(\boldsymbol{x}) &= \|\boldsymbol{x} - \boldsymbol{p}_1\|_2^2 - 1 \leq 0 \\ g_2(\boldsymbol{x}) &= \|\boldsymbol{x} - \boldsymbol{p}_2\|_2^2 - 1 \leq 0. \end{aligned}$$

Note that the only $\boldsymbol{x}$ satisfying both constraints is $\boldsymbol{x} = 0$ and there are no *strictly* feasible points.

Certificates of (sub)optimality

One potential application of the above facts is to serve as a way of measuring how far away we are from finding an optimal solution to our optimization problem. To see this recall that any dual feasible[3] $(\boldsymbol{\lambda}, \boldsymbol{\nu})$ gives us a lower bound on $p^\star$, since $d(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^\star$. Thus, if we have a primal feasible $\boldsymbol{x}$, then we know that

$$f(\boldsymbol{x}) - p^\star \;\leq\; f(\boldsymbol{x}) - d(\boldsymbol{\lambda}, \boldsymbol{\nu}).$$

We will refer to $f(\boldsymbol{x}) - d(\boldsymbol{\lambda}, \boldsymbol{\nu})$ as the duality gap for the primal/dual (feasible) variables $\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}$. We know that

$$p^\star \in [d(\boldsymbol{\lambda}, \boldsymbol{\nu}), f(\boldsymbol{x})], \quad \text{and likewise} \quad d^\star \in [d(\boldsymbol{\lambda}, \boldsymbol{\nu}), f(\boldsymbol{x})].$$

If we are ever able to reduce this gap to zero, then we know that $\boldsymbol{x}$ is primal optimal, and $\boldsymbol{\lambda}, \boldsymbol{\nu}$ are dual optimal.

There are certain kinds of "primal-dual" algorithms that produce a series of (feasible) points $\boldsymbol{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\nu}_k$ at every iteration. We can then use

$$f(\boldsymbol{x}_k) - d(\boldsymbol{\lambda}_k, \boldsymbol{\nu}_k) \leq \epsilon,$$

as a stopping criteria, and know that our answer would yield an objective value no further than ϵ from optimal.

[3] We simply need $\boldsymbol{\lambda} \geq \mathbf{0}$ for $(\boldsymbol{\lambda}, \boldsymbol{\nu})$ to be dual feasible.

Lagrange duality examples

1. **Inequality LP**. Calculate the dual of

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}}\ \langle \boldsymbol{x}, \boldsymbol{c}\rangle \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}.$$

Answer: The Lagrangian is

$$\begin{aligned} L(\boldsymbol{x}, \boldsymbol{\lambda}) &= \langle \boldsymbol{x}, \boldsymbol{c}\rangle + \sum_{m=1}^{M} \lambda_m \left(\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x} - b_m\right) \\ &= \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{b} + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}. \end{aligned}$$

This is a linear functional in $\boldsymbol{x}$ — it is unbounded below unless

$$\boldsymbol{c} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} = \boldsymbol{0}.$$

Thus

$$\begin{aligned} d(\boldsymbol{\lambda}) &= \inf_{\boldsymbol{x}} \left(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{b} + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}\right) \\ &= \begin{cases} -\langle \boldsymbol{\lambda}, \boldsymbol{b}\rangle, & \boldsymbol{c} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} = \boldsymbol{0} \\ -\infty, & \text{otherwise.} \end{cases} \end{aligned}$$

So the Lagrange dual program is

$$\begin{aligned} \underset{\boldsymbol{\lambda}\in\mathbb{R}^M}{\text{maximize}}\ -\langle \boldsymbol{\lambda}, \boldsymbol{b}\rangle \quad \text{subject to} \quad & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} = -\boldsymbol{c} \\ & \boldsymbol{\lambda} \geq \boldsymbol{0}. \end{aligned}$$

2. **Least-squares**. Calculate the dual of

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \ \|\boldsymbol{x}\|_2^2 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},$$

where $\boldsymbol{A}$ is an $M \times N$ matrix. Check that the duality gap is zero.

Answer: The Lagrangian is

$$L(\boldsymbol{x}, \boldsymbol{\nu}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} - \boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{b} + \boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}.$$

This is quadratic in $\boldsymbol{x}$ and will attain its minimum for

$$\boldsymbol{x} = -\frac{1}{2}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}.$$

Thus

$$\begin{aligned} d(\boldsymbol{\nu}) &= \frac{1}{4}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\nu} - \frac{1}{2}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} \\ &= -\frac{1}{4}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\nu}, \end{aligned}$$

and the Lagrange dual problem is

$$\underset{\boldsymbol{\nu}\in\mathbb{R}^M}{\text{maximize}} -\frac{1}{4}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\nu}.$$

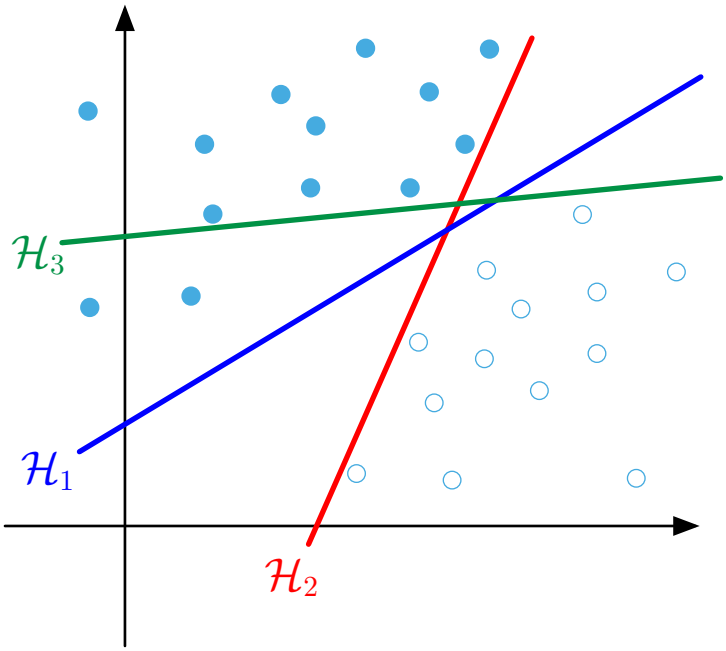
Note that this will be maximized when $-\frac{1}{2}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}^\star = \boldsymbol{b}$, which, when substituted into the dual problem yields

$$-\frac{1}{4}\boldsymbol{\nu}^{\star\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}^\star + \frac{1}{2}\boldsymbol{\nu}^{\star\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}^\star = \frac{1}{4}\boldsymbol{\nu}^{\star\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}^\star = \left\| -\frac{1}{2}\boldsymbol{A}^T\boldsymbol{\nu}^\star \right\|_2^2.$$

Since $\boldsymbol{x}^\star = -\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}^\star/2$ is primal feasible, we have found a $(\boldsymbol{x}^\star, \boldsymbol{\nu}^\star)$ such that $f(\boldsymbol{x}^\star) = d(\boldsymbol{\nu}^\star)$, so we see first hand that we have strong duality.

3. Support vector machines

Consider the following fundamental binary classification problem. We are given points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M \in \mathbb{R}^N$ with class labels $y_1, \ldots, y_M$, where $y_m \in \{-1, +1\}$. We would like to find a hyperplane (i.e., affine functional) which *separates* the classes:



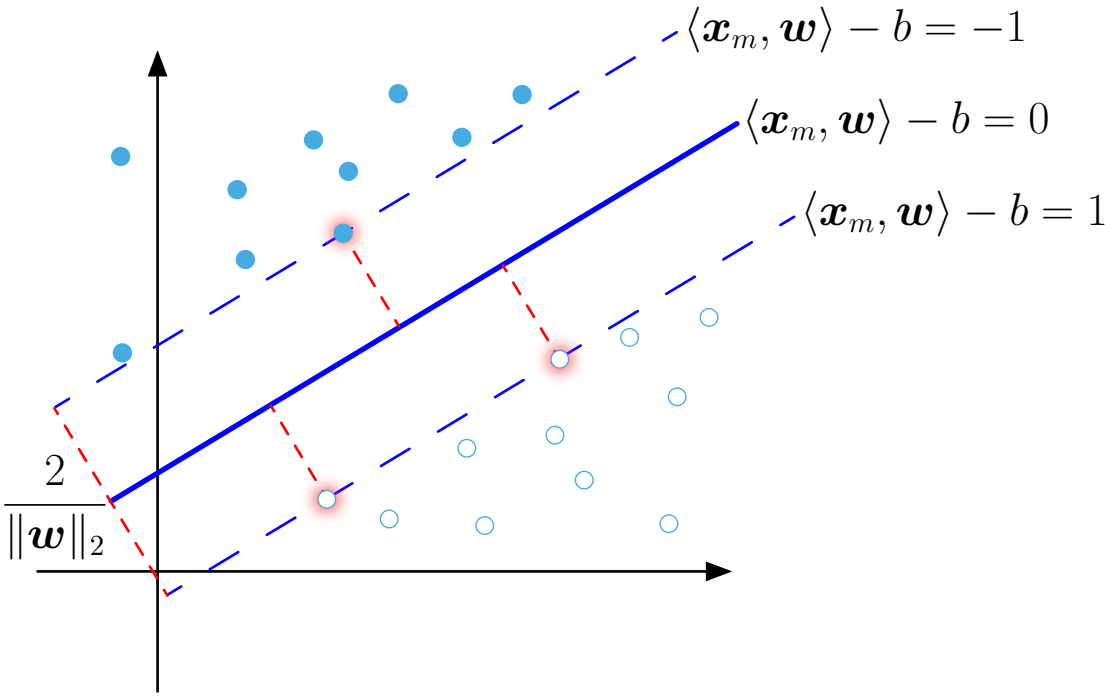
$\mathcal{H}_1$ and $\mathcal{H}_2$ above both separate the classes in $\mathbb{R}^2$, but $\mathcal{H}_3$ does not. While separating the classes is obviously desirable, we still need a good method to choose from among the many hyperplanes that do separate the classes – and some will perform better than others. Support vector machines (SVMs) take the one with *maximum margin*, i.e., we choose the hyperplane that maximizes the distance to the closest point in either class.

To restate this, we want to find a $\boldsymbol{w} \in \mathbb{R}^N$ and $b \in \mathbb{R}$ such that

$$\begin{aligned} \langle \boldsymbol{x}_m, \boldsymbol{w} \rangle - b &\geq 1, \quad \text{when } y_m = 1, \\ \langle \boldsymbol{x}_m, \boldsymbol{w} \rangle - b &\leq -1, \quad \text{when } y_m = -1. \end{aligned}$$

Of course, it is possible that no separating hyperplane exists; in this case, there will be no feasible points in the program above. It is straightforward, though, to modify this discussion to allow "mislabeled" points.

In the formulation above, the distance between the two (parallel) hyperplanes is $2/\|\boldsymbol{w}\|_2$:



Thus maximizing this distance is the same as minimizing $\|\boldsymbol{w}\|_2$.

This leads to the program

$$\begin{aligned} &\underset{\boldsymbol{w}\in\mathbb{R}^N,\ b\in\mathbb{R}}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ &\text{subject to} \ \ y_m(b - \langle \boldsymbol{x}_m, \boldsymbol{w} \rangle) + 1 \ \leq 0, \quad m = 1, \ldots, M. \end{aligned}$$

This is a linearly constrained quadratic program, and is clearly

convex. The Lagrangian is

$$\begin{aligned} L(\boldsymbol{w}, b, \boldsymbol{\lambda}) &= \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{m=1}^{M} \lambda_m \left[y_m(b - \langle \boldsymbol{x}_m, \boldsymbol{w} \rangle) + 1 \right] \\ &= \frac{1}{2}\|\boldsymbol{w}\|_2^2 + b\, \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{y} - \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{w} + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{1}, \end{aligned}$$

where $\boldsymbol{X}$ is the $N \times M$ matrix

$$\boldsymbol{X} = \begin{bmatrix} y_1 \boldsymbol{x}_1 & y_2 \boldsymbol{x}_2 & \cdots & y_M \boldsymbol{x}_M \end{bmatrix}.$$

The dual function is

$$d(\boldsymbol{\lambda}) = \inf_{\boldsymbol{w}, b} \left(\frac{1}{2}\|\boldsymbol{w}\|_2^2 + b\, \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{y} - \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{w} + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{1} \right).$$

Since b is unconstrained above, we see that the presence of $b\, \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{y}$ means that the dual will be $-\infty$ unless $\boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{y} = 0$. Minimizing over $\boldsymbol{w}$, we need the gradient equal to zero,

$$\nabla_{\boldsymbol{w}} L(\boldsymbol{w}, b, \boldsymbol{\lambda}) = \mathbf{0}, \quad \Rightarrow \quad \boldsymbol{w} - \boldsymbol{X}\boldsymbol{\lambda} = \mathbf{0}.$$

This means that we must have $\boldsymbol{w} = \boldsymbol{X}\boldsymbol{\lambda}$, which itself is a very handy fact as it gives us a direct passage from the dual solution to the primal solution. With these substitutions, the dual function is

$$d(\boldsymbol{\lambda}) = \begin{cases} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\lambda}\|_2^2 - \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{\lambda} + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{1}, & \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{y} = 0, \\ -\infty, & \text{otherwise.} \end{cases}$$

Thus, the dual SVM program is then

$$\begin{aligned} \underset{\boldsymbol{\lambda}}{\text{maximize}} \;\; & -\frac{1}{2}\|\boldsymbol{X}\boldsymbol{\lambda}\|_2^2 + \sum_{m=1}^{M} \lambda_m \\ \text{subject to} \;\; & \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\lambda} = 0, \quad \boldsymbol{\lambda} \geq \mathbf{0}. \end{aligned}$$

Given the solution $\boldsymbol{\lambda}^\star$ to the dual, we can take $\boldsymbol{w}^\star = \boldsymbol{X}\boldsymbol{\lambda}^\star$, and the classifier is

$$\begin{aligned} f(\boldsymbol{x}) &= \langle \boldsymbol{x}, \boldsymbol{w}^\star \rangle - b^\star \\ &= \langle \boldsymbol{x}, \boldsymbol{X}\boldsymbol{\lambda}^\star \rangle - b^\star \\ &= \sum_{m=1}^{M} \lambda_m^\star y_m \langle \boldsymbol{x}, \boldsymbol{x}_m \rangle - b^\star. \end{aligned}$$

Notice that the data $\boldsymbol{x}_m$ appear only through inner products with $\boldsymbol{x}$.

A key realization about the SVM is that the for the dual program, the objective function depends on the data $\boldsymbol{x}_m$ only through inner products, as

$$\|\boldsymbol{X}\boldsymbol{\lambda}\|_2^2 = \sum_{\ell=1}^{M}\sum_{m=1}^{M} y_\ell y_m \langle \boldsymbol{x}_\ell, \boldsymbol{x}_m \rangle.$$

This means that we can replace $\langle \boldsymbol{x}_\ell, \boldsymbol{x}_m \rangle$ with any "positive kernel function" $K(\boldsymbol{x}_\ell, \boldsymbol{x}_m) : \mathbb{R}^N \otimes \mathbb{R}^N \to \mathbb{R}$ – a positive kernel just means that the $M \times M$ matrix $K(\boldsymbol{x}_\ell, \boldsymbol{x}_m)$ is in S_+^M for all choices of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$.

For example, you might take

$$K(\boldsymbol{x}_\ell, \boldsymbol{x}_m) = (1 + \langle \boldsymbol{x}_\ell, \boldsymbol{x}_m \rangle)^2 = 1 + 2\langle \boldsymbol{x}_\ell, \boldsymbol{x}_m \rangle + \langle \boldsymbol{x}_\ell, \boldsymbol{x}_m \rangle^2.$$

This means we have replaced the inner product of two vectors with the inner product between two vectors which have been mapped into a higher-dimensional space:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \quad \rightarrow \quad \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_N \\ x_1^2 \\ x_2^2 \\ \vdots \\ x_N^2 \\ \sqrt{2}x_1x_2 \\ \vdots \\ \sqrt{2}x_{N-1}x_N \end{bmatrix}.$$

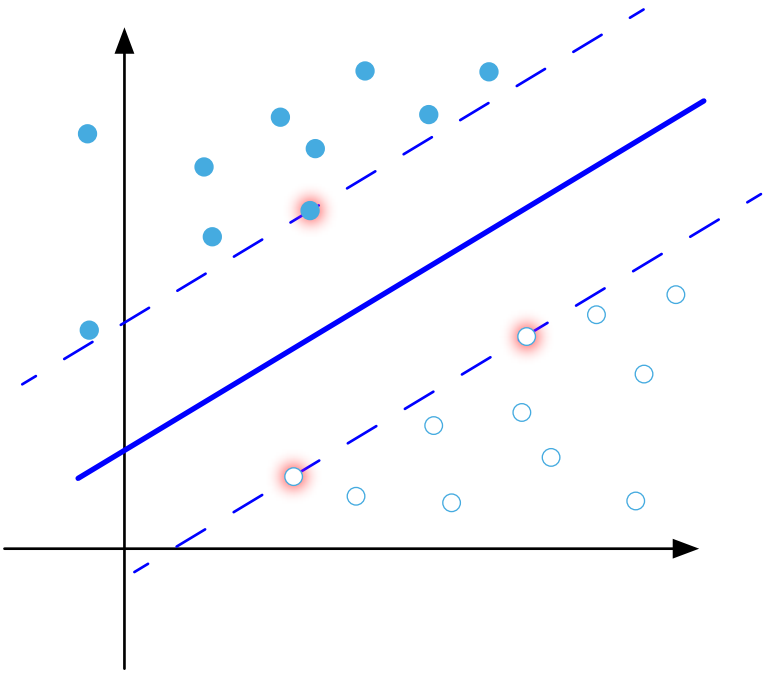
A set of linear constraints on the coordinates on the right, then, corresponds to a second order curve constraint (parabola, ellipse, hyperbola) on the coordinates on the left.

Many kernels are possible. The advantage is that to train and use the classifier, you never have to explicitly move to the higher-dimensional space – you just need to be able to compute $K(\boldsymbol{x}_\ell, \boldsymbol{x}_m)$ for any pair of inputs in $\mathbb{R}^N$. A popular choice of kernel is

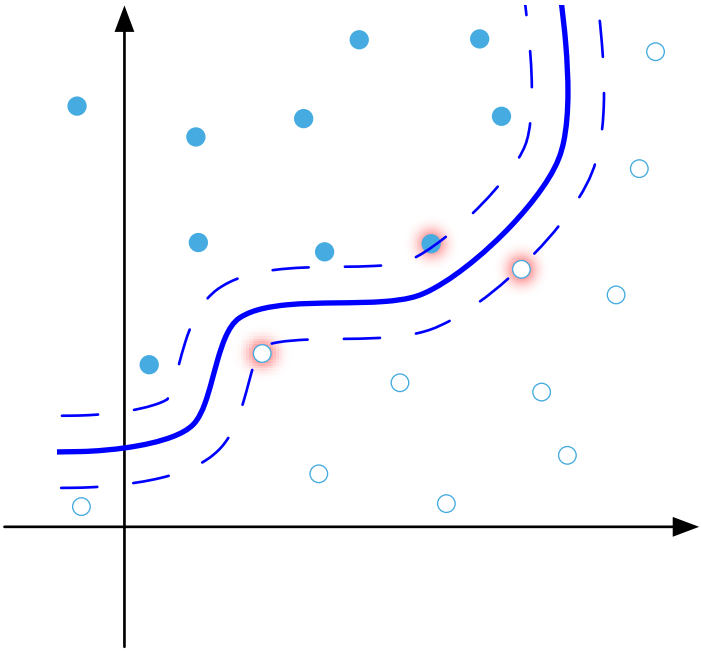
$$K(\boldsymbol{x}_\ell, \boldsymbol{x}_m) = \exp\left(-\gamma \|\boldsymbol{x}_\ell - \boldsymbol{x}_m\|_2^2\right).$$

This is a perfectly valid positive kernel, and it is straightforward to compute it for any pair of inputs. But it corresponds to mapping the $\boldsymbol{x}_m$ into an infinite dimensional space, then finding a separating hyperplane.

Here is an example of a linear classifier in a higher-dimensional space:



that results in a nonlinear classifier in a lower-dimensional space:



4. **Minimum norm.** Calculate the dual of

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \ \|\boldsymbol{x}\| \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},$$

where $\|\cdot\|$ is an arbitrary (valid) norm and $\boldsymbol{A}$ is an $M \times N$ matrix.

Answer: We start with the Lagrangian:

$$\begin{aligned} L(\boldsymbol{x}, \boldsymbol{\nu}) &= \|\boldsymbol{x}\| + \boldsymbol{\nu}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \\ &= \|\boldsymbol{x}\| - \boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{b} + (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu})^{\mathrm{T}}\boldsymbol{x} \end{aligned}$$

and so

$$\begin{aligned} d(\boldsymbol{\nu}) &= -\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{b} + \inf_{\boldsymbol{x}} \left(\|\boldsymbol{x}\| + (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu})^{\mathrm{T}}\boldsymbol{x} \right) \\ &= -\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{b} + \inf_{\boldsymbol{x}} \left(\|\boldsymbol{x}\| - \|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}\|_* \|\boldsymbol{x}\| \right) \\ &= -\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{b} + \inf_{\boldsymbol{x}} \left((1 - \|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}\|_*) \|\boldsymbol{x}\| \right). \end{aligned}$$

where the second equality above comes from the generalized Cauchy-Schwarz inequality for dual norms: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \geq -\|\boldsymbol{x}\| \, \|\boldsymbol{y}\|_*$ and there is always a $\boldsymbol{x}$ such that $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = -\|\boldsymbol{x}\| \, \|\boldsymbol{y}\|_*$. If $\|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}\|_* > 1$, then we can drive the expression on the right to $-\infty$ by scaling $\boldsymbol{x}$, so we know

$$d(\boldsymbol{\nu}) = -\infty \quad \text{if} \ \ \|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}\|_* > 1.$$

For $\|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}\|_* \leq 1$, the expression inside the inf above is ≥ 0 for all $\boldsymbol{x}$ and we can make it equal to 0 (minimize it) for $\boldsymbol{x} = 0$. Thus

$$d(\boldsymbol{\nu}) = \begin{cases} -\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{b}, & \|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}\|_* \leq 1 \\ -\infty, & \text{otherwise}, \end{cases}$$

and the dual program is

$$\underset{\boldsymbol{\nu}\in\mathbb{R}^M}{\text{maximize}} \ -\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{b} \quad \text{subject to} \quad \|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}\|_* \leq 1.$$