CIS 520: Machine Learning **Oct 02, 2017**

Constrained Optimization and Lagrangian Duality

Lecturer: Shivani Agarwal

Disclaimer: These notes are designed to be a supplement to the lecture. They may or may not cover all the material discussed in the lecture (and vice versa).

Outline

- Preliminaries: convex sets and convex functions
- Constrained optimization
- Projected gradient descent
- Lagrangian duality
- Karush-Kuhn-Tucker (KKT) optimality conditions

1 Preliminaries: Convex Sets and Convex Functions

Let us start by recalling the notions of convex sets and convex functions.

1.1 Convex Sets

Definition 1 (Convex sets)**.** A set $\mathcal{C} \subseteq \mathbb{R}^d$ is *convex* if the line segment between any two points in $\mathcal{C}$ lies in $\mathcal{C}$, i.e. if for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}$ and all $\theta \in [0, 1]$, we have

$$\theta \mathbf{x}_1 + (1 - \theta)\mathbf{x}_2 \;\in\; \mathcal{C} \,.$$

See Figure 1 for an illustration of both convex and non-convex sets.

1.1.1 Examples of Convex Sets

Example 1 (Hyperplanes)**.** Let $\mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}$, and let $\mathcal{C}$ be the set of all points on the hyperplane with coefficients given by $\mathbf{a}, b$:

$$\mathcal{C} = \left\{ \mathbf{x} \in \mathbb{R}^d \;\middle|\; \mathbf{a}^\top \mathbf{x} + b = 0 \right\}.$$

Then $\mathcal{C}$ is a convex set.

Example 2 (Halfspaces)**.** Let $\mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}$, and let $\mathcal{C}$ be the halfspace given by

$$\mathcal{C} = \left\{ \mathbf{x} \in \mathbb{R}^d \;\middle|\; \mathbf{a}^\top \mathbf{x} + b \leq 0 \right\}.$$

Then $\mathcal{C}$ is a convex set.

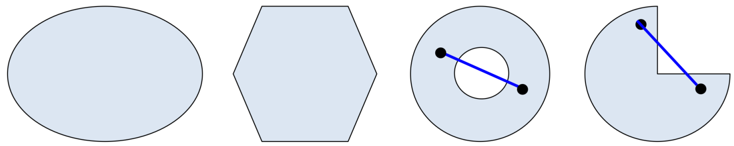

Figure 1: Examples of (left, second-left) convex and (right, second-right) non-convex sets in $\mathbb{R}^2$.

Example 3 (Norm balls)**.** Let $\|\cdot\|$ be any norm on $\mathbb{R}^d$ (such as the Euclidean norm $\|\cdot\|_2$), and let $\mathbf{x}_0 \in \mathbb{R}^d$, $r > 0$. Let $\mathcal{C}$ be the corresponding norm ball of radius r centered at $\mathbf{x}_0$:

$$\mathcal{C} = \left\{ \mathbf{x} \in \mathbb{R}^d \;\middle|\; \|\mathbf{x} - \mathbf{x}_0\| \leq r \right\}.$$

Then $\mathcal{C}$ is a convex set.

Example 4 (Polyhedra)**.** Let $\mathcal{C}$ be a polyhedron, i.e. the intersection of a finite number of halfspaces and hyperplanes:

$$\mathcal{C} = \left\{ \mathbf{x} \in \mathbb{R}^d \;\middle|\; \mathbf{a}_i^\top \mathbf{x} \leq b_i, i = 1, \ldots, m;\; \mathbf{c}_j^\top \mathbf{x} = d_j, j = 1, \ldots, n \right\}.$$

Then $\mathcal{C}$ is a convex set.

Example 5 (Intersection of convex sets)**.** Let $\mathcal{C}_1, \ldots, \mathcal{C}_n \subseteq \mathbb{R}^d$ be convex sets. Then their intersection

$$\mathcal{C} = \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_n$$

is also a convex set. This property extends to infinite intersections.

1.2 Convex Functions

Definition 2 (Convex functions.)**.** Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a convex set. A function $f : \mathcal{C} \to \mathbb{R}$ is *convex* if for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}$ and all $\theta \in [0, 1]$, we have

$$f\big(\theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2\big) \;\leq\; \theta f(\mathbf{x}_1) + (1-\theta) f(\mathbf{x}_2),$$

and is *strictly convex* if whenever $\mathbf{x}_1 \neq \mathbf{x}_2$ and $\theta \in (0, 1)$, we have strict inequality:

$$f\big(\theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2\big) \;<\; \theta f(\mathbf{x}_1) + (1-\theta) f(\mathbf{x}_2).$$

We say f is *concave* if $-f$ is convex, and is *strictly concave* if $-f$ is strictly convex.

In other words, f is convex if the chord joining any two points $(\mathbf{x}_1, f(\mathbf{x}_1))$ and $(\mathbf{x}_2, f(\mathbf{x}_2))$ lies above the graph of the function f. See Figure 2 (left) for an illustration.

1.2.1 Characterizations of Convex Functions

Proposition 1 (First-order condition for convexity)**.** Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a convex set, and let $f : \mathcal{C} \to \mathbb{R}^d$ be differentiable (i.e. its gradient ∇f exists at each point in $\mathcal{C}$). Then f is convex if and only if for all $\mathbf{x}, \mathbf{x}_0 \in \mathcal{C}$, we have

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0),$$

and is strictly convex if and only if for all $\mathbf{x} \neq \mathbf{x}_0 \in \mathcal{C}$, we have strict inequality:

$$f(\mathbf{x}) > f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0).$$

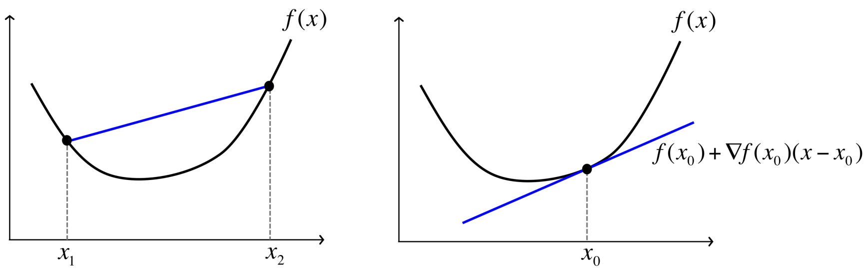

Figure 2: Example of a convex one-dimensional function f. Left: The chord joining any two points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ must lie above the graph of f. Right: The first-order Taylor approximation of f at any point x_0 must lie below the graph of f.

In other words, f is convex if and only if the first-order Taylor approximation of f at any point $\mathbf{x}_0$, given by the linear function $f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0)$, lies below the graph. See Figure 2 (right).

Proposition 2 (Second-order condition for convexity)**.** Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a convex set, and let $f : \mathcal{C} \to \mathbb{R}^d$ be twice differentiable (i.e. its Hessian or second derivative $\nabla^2 f$ exists at each point in $\mathcal{C}$). Then f is convex if and only if its Hessian is positive semi-definite, i.e. for all $\mathbf{x} \in \mathcal{C}$, we have

$$\nabla^2 f(\mathbf{x}) \succeq 0\,.$$

If $\nabla^2 f(\mathbf{x}) \succ 0$ for all $\mathbf{x} \in \mathcal{C}$, then f is strictly convex (the converse however is not true).

1.2.2 Examples of Convex Functions

Example 6 (Affine functions)**.** Let $\mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}$, and let $f : \mathbb{R}^d \to \mathbb{R}$ be the affine function defined by

$$f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + b\,.$$

Then f is both a convex function and a concave function.

Example 7 (Quadratic functions)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a symmetric matrix, and let $\mathbf{b} \in \mathbb{R}^d, c \in \mathbb{R}$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be the quadratic function defined by

$$f(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c\,.$$

Then f is convex if and only if $\mathbf{A}$ is positive semi-definite, i.e. $\mathbf{A} \succeq 0$.

Example 8 (Common one-dimensional functions)**.**

1. *Exponentials:* For any $a \in \mathbb{R}$, the function $f(x) = e^{ax}$ is convex on $\mathbb{R}$.
2. *Powers:* The function $f(x) = x^a$ is convex on $\mathbb{R}_{++}$ for $a \geq 1$ or $a \leq 0$, and is concave for $0 \leq x \leq 1$.
3. *Powers of absolute value:* The function $f(x) = |x|^p$ is convex on $\mathbb{R}$ for $p \geq 1$.
4. *Logarithm:* The function $f(x) = \log x$ is concave on $\mathbb{R}_{++}$.
5. *Entropy:* The function $f(x) = -x \log x$ is concave on $\mathbb{R}_{++}$ (or on $\mathbb{R}_+$, with $f(0)$ defined as 0).

Example 9 (Norms)**.** Let $\|\cdot\|$ be any norm on $\mathbb{R}^d$, and let $f : \mathbb{R}^d \to \mathbb{R}$ be the corresponding norm function:

$$f(\mathbf{x}) = \|\mathbf{x}\|\,.$$

Then f is convex.

Example 10 (Log-sum-exp)**.** The function $f : \mathbb{R}^d \to \mathbb{R}$ defined as

$$f(\mathbf{x}) = \log\left(e^{x_1} + \cdots + e^{x_d}\right)$$

is convex on $\mathbb{R}^d$.

Example 11 (Non-negative weighted sums of convex functions)**.** Let $f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R}$ be convex functions, and let $w_1, \ldots, w_n \geq 0$. Then the function $f : \mathbb{R}^d \to \mathbb{R}$ defined as

$$f(\mathbf{x}) = w_1 f_1(\mathbf{x}) + \cdots + w_n f_n(\mathbf{x})$$

is also convex. This property extends to infinite sums and integrals.

Example 12 (Composition with affine functions)**.** Let $f : \mathbb{R}^d \to \mathbb{R}$ and let $\mathbf{A} \in \mathbb{R}^{d \times p}, \mathbf{b} \in \mathbb{R}^d$. Define $g : \mathbb{R}^p \to \mathbb{R}$ as

$$g(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})\,.$$

If f is convex, then g is also convex; if f is concave, then g is also concave.

Example 13 (Pointwise maximum/supremum of convex functions)**.** Let $f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R}$ be convex functions. Then the function $f : \mathbb{R}^d \to \mathbb{R}$ defined as

$$f(\mathbf{x}) = \max\left(f_1(\mathbf{x}), \ldots, f_n(\mathbf{x})\right)$$

is also convex. Similarly, if $\{f_\alpha\}_\alpha$ is an infinite family of convex functions $f_\alpha : \mathbb{R}^d \to \mathbb{R}$, then the function $g : \mathbb{R}^d \to \mathbb{R}$ defined as

$$g(\mathbf{x}) = \sup_\alpha f_\alpha(\mathbf{x})$$

is also convex.

1.2.3 Sublevel Sets of Convex Functions

Example 14 (Sublevel sets)**.** Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a convex set and let $f : \mathcal{C} \to \mathbb{R}$ be a convex function. Then for every $\alpha \in \mathbb{R}$, the *α-sublevel set* of f, given by

$$H_\alpha = \left\{\mathbf{x} \in \mathcal{C} \;\middle|\; f(\mathbf{x}) \leq \alpha\right\},$$

is a convex set. Note that norm balls are a special case: they are simply the sublevel sets of the corresponding (convex) norm function!

2 Constrained Optimization

In many settings, we need to minimize a function $f : \mathbb{R}^d \to \mathbb{R}$ over some *constraint set* $\mathcal{C} \subseteq \mathbb{R}^d$:

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathcal{C}\,. \end{array} \tag{1}$$

This is a *constrained* optimization problem. A point $\mathbf{x}^* \in \mathcal{C}$ is said to be a *global minimizer* of this problem if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{C}$. A point $\mathbf{x}_0 \in \mathcal{C}$ is said to be a *local minimizer* if there exists $\epsilon > 0$ such that $f(\mathbf{x}_0) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{C}$ with $\|\mathbf{x} - \mathbf{x}_0\|_2 < \epsilon$.

If the constraint set $\mathcal{C}$ is a convex set and the objective function f is a convex function on $\mathcal{C}$, then the above problem is said to be a *convex* optimization problem.[1] In this case, any local minimizer is also a global minimizer.

[1] Note that maximizing a concave function is equivalent to minimizing a convex function, and therefore any problem involving maximization of a concave function over a convex constraint set is also a convex optimization problem.

We will focus on constraint sets $\mathcal{C}$ described as the intersection of a finite number of inequality and equality constraints defined by some functions $g_i : \mathbb{R}^d \to \mathbb{R}, i = 1, \ldots, m$ and $h_j : \mathbb{R}^d \to \mathbb{R}, j = 1, \ldots, n$ as follows:

$$\mathcal{C} = \left\{ \mathbf{x} \in \mathbb{R}^d \;\middle|\; g_i(\mathbf{x}) \leq 0, i = 1, \ldots, m;\; h_j(\mathbf{x}) = 0, j = 1, \ldots, n \right\}.$$

The resulting constrained optimization problem is then written as

$$\begin{array}{lll} \text{minimize} & f(\mathbf{x}) & \\ \text{subject to} & g_i(\mathbf{x}) \leq 0, & i = 1, \ldots, m \\ & h_j(\mathbf{x}) = 0, & j = 1, \ldots n\,. \end{array} \tag{2}$$

If the objective function f and inequality constraint functions g_i are all convex and the equality constraint functions h_j are all affine, then the above problem is a convex optimization problem.

Example 15. Consider the one-dimensional minimization problem

$$\begin{array}{ll} \text{minimize} & (x-3)^2 \\ \text{subject to} & x - 2 \leq 0\,. \end{array}$$

This is a convex optimization problem with a unique global minimum at $x^* = 2$. (To see this, simply draw a graph of the function $f(x) = (x-3)^2$ and observe the behavior of this function over the constraint set $\mathcal{C} = (-\infty, 2]$.) In fact, this is a simple example of a convex *quadratic program* (convex quadratic objective, affine constraints). We will consider larger convex quadratic programs below.

For convex constrained optimization problems, many specialized solution methods have been developed; often, these are tailored to different classes of optimization problems (such as linear programs, quadratic programs, semi-definite programs, etc). Below we quickly look at one type of general-purpose solution method, namely, that of projected gradient descent. We will then discuss the important concept of Lagrangian duality and the associated Karush-Kuhn-Tucker (KKT) optimality conditions for constrained optimization problems. In some cases, the dual optimization problem, which is always convex, may be easier to solve than the original (primal) optimization problem (and can be used to indirectly solve the primal). In some special cases, the KKT conditions can be solved analytically in order to solve the primal problem. Often, the specialized solvers one might use can also be viewed as trying to solve the KKT conditions.

3 Projected Gradient Descent

As noted above, *projected gradient descent* (PGD) is one type of general-purpose solution method for (convex) constrained optimization problems.

The basic idea is quite simple. Consider a convex optimization problem of the form (1), with convex f and $\mathcal{C}$. A standard gradient descent approach, which can be used to find a local minimum for an unconstrained optimization problem, is problematic here since the iterates $\mathbf{x}^t$ may fall outside the constraint set $\mathcal{C}$. What PGD does is simply correct for this situation: on each iteration t, it first applies a standard gradient descent step to obtain an intermediate point $\widetilde{\mathbf{x}}^{t+1}$ that might fall outside $\mathcal{C}$, and then *projects* this point back to the constraint set $\mathcal{C}$ by finding a point $\mathbf{x}^{t+1}$ in $\mathcal{C}$ that is closest (in terms of Euclidean distance) to $\widetilde{\mathbf{x}}^{t+1}$:

Algorithm **Projected Gradient Descent (PGD)**

Inputs: Objective function $f : \mathbb{R}^d \to \mathbb{R}$; constraint set $\mathcal{C} \subseteq \mathbb{R}^d$
Parameter: Number of iterations T
Initialize: $\mathbf{x}^1 \in \mathbb{R}^d$
For $t = 1, \ldots, T$:
– Select step-size $\eta_t > 0$
– $\widetilde{\mathbf{x}}^{t+1} \leftarrow \mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t)$
– $\mathbf{x}^{t+1} \leftarrow \operatorname{argmin}_{\mathbf{x}\in\mathcal{C}} \|\mathbf{x} - \widetilde{\mathbf{x}}^{t+1}\|_2$
Output: $\mathbf{x}^{T+1}$

For convex $\mathcal{C}$, the point $\mathbf{x}^{t+1}$ after projection is always closer to the minimizer in $\mathcal{C}$ than is the result of the gradient descent step $\widetilde{\mathbf{x}}^{t+1}$, and for suitable choices of the step-sizes η_t (e.g. $\eta_t = 1/t$), PGD converges to a (global) minimizer of the (convex) constrained optimization problem.

Note that in order to implement PGD in practice, one must be able to efficiently compute projections onto the constraint set $\mathcal{C}$; this can be done easily for certain types of constraint sets, but can be harder for others.

4 Lagrangian Duality

We now turn to discuss the important concept of Lagrangian duality associated with constrained optimization problems. In particular, any constrained optimization problem has an associated *(Lagrange) dual optimization problem*, which is always convex. Sometimes, the dual problem is easier to solve, and can be used to obtain useful information about the original (primal) problem; in some cases, it can even be used to obtain a solution to the primal problem.

Consider a constrained optimization problem of the form shown in Eq. (2). We will refer to this as the *primal* problem and will denote its constraint set by $\mathcal{C}_P$:

$$\mathcal{C}_P = \left\{ \mathbf{x} \in \mathbb{R}^d \;\middle|\; g_i(\mathbf{x}) \le 0, i = 1, \ldots, m;\; h_j(\mathbf{x}) = 0, j = 1, \ldots, n \right\}.$$

Any point $\mathbf{x} \in \mathcal{C}_P$ is said to be *primal feasible*. Denote by p^* the optimal value of the primal problem:

$$p^* = \inf_{\mathbf{x}\in\mathcal{C}_P} f(\mathbf{x}).$$

Note that p^* may or may not be achieved, i.e. there may or may not exist a point $\mathbf{x}^* \in \mathcal{C}_P$ with value $f(\mathbf{x}^*) = p^*$.

In Lagrangian duality, one introduces *dual variables* λ_i and ν_j associated with each of the inequality constraints $g_i(\mathbf{x}) \le 0$ and equality constraints $h_j(\mathbf{x}) = 0$, respectively, and augments the objective function f by adding the constraint functions multiplied by the corresponding dual variables to obtain the *Lagrangian* function. Formally, the *Lagrangian* function $\mathcal{L} : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ associated with the optimization problem (2) is defined as

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^{n} \nu_j h_j(\mathbf{x}).$$

The dual variables λ_i and ν_j are also referred to as the *Lagrange multipliers* associated with the inequality and equality constraints, respectively.

Next, one defines the *(Lagrange) dual function* $\phi : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ over the dual variables as

$$\phi(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}\in\mathbb{R}^d} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}),$$

and the associated *dual optimization problem* over the dual variables as

$$\begin{array}{ll} \text{maximize} & \phi(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ \text{subject to} & \lambda_i \geq 0 \quad i = 1, \ldots, m\,. \end{array} \tag{3}$$

This dual problem is always a convex optimization problem, even if the primal problem is not convex (note that the dual function is a pointwise infimum of affine functions of $(\boldsymbol{\lambda}, \boldsymbol{\nu})$, and therefore concave). Let $\mathcal{C}_D$ denote the constraint set of this dual problem:

$$\mathcal{C}_D = \left\{(\boldsymbol{\lambda}, \boldsymbol{\nu}) \in \mathbb{R}^m \times \mathbb{R}^n \;\middle|\; \lambda_i \geq 0, i = 1, \ldots, m\right\}.$$

Any point $(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is said to be *dual feasible*. Denote by d^* the optimal value of the dual problem:

$$d^* = \sup_{(\boldsymbol{\lambda}, \boldsymbol{\nu}) \in \mathcal{C}_D} \phi(\boldsymbol{\lambda}, \boldsymbol{\nu})\,.$$

Again, d^* may or may not be achieved. Now, it can be verified that

$$d^* \leq p^*\,.$$

Indeed, to see this, consider any primal feasible $\mathbf{x}$ and any dual feasible $(\boldsymbol{\lambda}, \boldsymbol{\nu})$. Then we have

$$\begin{aligned} \phi(\boldsymbol{\lambda}, \boldsymbol{\nu}) &= \inf_{\mathbf{x}' \in \mathbb{R}^d} \mathcal{L}(\mathbf{x}', \boldsymbol{\lambda}, \boldsymbol{\nu}) \\ &\leq \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \\ &= f(\mathbf{x}) + \sum_{i=1}^m \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^n \nu_j h_j(\mathbf{x}) \\ &\leq f(\mathbf{x})\,, \quad \text{since } h_j(\mathbf{x}) = 0,\ g_i(\mathbf{x}) \leq 0, \text{ and } \lambda_i \geq 0.\,. \end{aligned}$$

Since this inequality holds for all $\mathbf{x} \in \mathcal{C}_P$ and all $(\boldsymbol{\lambda}, \boldsymbol{\nu}) \in \mathcal{C}_D$, it follows that

$$d^* = \sup_{(\boldsymbol{\lambda}, \boldsymbol{\nu}) \in \mathcal{C}_D} \phi(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \inf_{\mathbf{x} \in \mathcal{C}_P} f(\mathbf{x}) = p^*\,.$$

Thus the optimal value of the dual problem always gives a lower bound on the optimal value of the primal problem; this property is called *weak duality*.

Under some settings, the primal and dual optimal values are in fact equal; this property, when it holds, is called *strong duality*:

$$d^* = p^*\,.$$

Strong duality does not always hold. However, if the primal problem is convex, i.e. if f and g_i are all convex functions and h_j are all affine functions, and if some additional conditions hold, then we do have strong duality. One such additional condition is *Slater's condition*, which requires that there exist a *strictly feasible* point $\mathbf{x}$ satisfying $g_i(\mathbf{x}) < 0, i = 1, \ldots, m$ and $h_j(\mathbf{x}) = 0, j = 1, \ldots, n$. In other words, if the primal problem is convex and Slater's condition holds, then we have strong duality (there are also other settings under which strong duality holds; we do not discuss these here). When strong duality holds, one can choose to solve the (convex) dual problem instead of the primal; on obtaining a dual solution $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$, unconstrained minimization of $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ over $\mathbf{x} \in \mathbb{R}^d$ then gives an optimal solution $\mathbf{x}^*$ to the primal problem.

Example 16. Consider the following convex quadratic program over $\mathbb{R}^d$ with equality constraints only:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\mathbf{x}^\top \mathbf{x} \\ \text{subject to} & \mathbf{C}\mathbf{x} = \mathbf{d}\,, \end{array}$$

where $\mathbf{C} \in \mathbb{R}^{n \times d}$, and $\mathbf{d} \in \mathbb{R}^n$. Here Slater's condition (trivially) holds, and therefore we have strong duality. Introducing dual variables $\boldsymbol{\nu} \in \mathbb{R}^n$, the Lagrangian function is given by

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

Minimizing over $\mathbf{x} \in \mathbb{R}^d$ gives the following dual function:

$$\phi(\boldsymbol{\nu}) = -\tfrac{1}{2}\boldsymbol{\nu}^\top \mathbf{C}\mathbf{C}^\top \boldsymbol{\nu} - \mathbf{c}^\top \boldsymbol{\nu} \,.$$

The dual optimization problem is then the following unconstrained convex quadratic program over $\boldsymbol{\nu} \in \mathbb{R}^n$:

$$\text{maximize} \quad -\tfrac{1}{2}\boldsymbol{\nu}^\top \mathbf{C}\mathbf{C}^\top \boldsymbol{\nu} - \mathbf{c}^\top \boldsymbol{\nu} \,.$$

This dual problem does not depend on the number of variables d in the primal problem. Thus if the number of variables d in the primal problem is significantly larger than the number of equality constraints n, then solving the dual problem can be more efficient than solving the primal directly. On obtaining the dual solution $\boldsymbol{\nu}^*$, the primal solution is given by $\mathbf{x}^* = -\mathbf{C}^\top \boldsymbol{\nu}^*$.

5 Karush-Kuhn-Tucker (KKT) Optimality Conditions

We now assume that the objective function f and constraint functions g_i, h_j are all differentiable.

Suppose strong duality holds, i.e. $d^* = p^*$, and that the optimal values of both primal and dual problems are achieved, i.e. there exist $\mathbf{x}^* \in \mathcal{C}_P$ and $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \in \mathcal{C}_D$ with

$$f(\mathbf{x}^*) = p^* = d^* = \phi(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \,.$$

Then we have the following:

$$\begin{aligned}
f(\mathbf{x}^*) &= \phi(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \\
&= \inf_{\mathbf{x} \in \mathbb{R}^d} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \\
&\leq \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \\
&= f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* g_i(\mathbf{x}^*) + \sum_{j=1}^n \nu_j^* h_j(\mathbf{x}^*) \\
&\leq f(\mathbf{x}^*) \,, \quad \text{since } h_j(\mathbf{x}^*) = 0,\ g_i(\mathbf{x}^*) \leq 0, \text{ and } \lambda_i^* \geq 0.
\end{aligned}$$

Since the left-hand and right-hand sides of the above equation are the same, this implies that both inequalities above must hold with equality. In particular, this allows us to draw the following two conclusions: First, we have that the infimum

$$\inf_{\mathbf{x} \in \mathbb{R}^d} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$$

is achieved at $\mathbf{x}^*$, which implies that the gradient of $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ vanishes at $\mathbf{x}^*$:

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^n \nu_j^* \nabla h_j(\mathbf{x}^*) = \mathbf{0} \,.$$

Second, we have that

$$\lambda_i^* g_i(\mathbf{x}^*) = 0 \,, \quad i = 1, \ldots, m \,;$$

in other words, for each $i = 1, \ldots, m$, we have that either $\lambda_i^* = 0$ or $g_i(\mathbf{x}^*) = 0$ (or both). This latter property is known as *complementary slackness*.

Thus, to summarize, if strong duality holds and the optimal values of the primal and dual problems are achieved at $\mathbf{x}^*$ and at $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$, respectively, then the following 5 conditions, collectively called the *Karush-*

Kuhn-Tucker (KKT) conditions, must hold:

$$\begin{aligned} g_i(\mathbf{x}^*) &\leq 0\,, \quad i = 1, \ldots, m \quad &&\text{(primal feasibility)} \\ h_j(\mathbf{x}^*) &= 0\,, \quad j = 1, \ldots, n \quad &&\text{(primal feasibility)} \\ \lambda_i^* &\geq 0\,, \quad i = 1, \ldots, m \quad &&\text{(dual feasibility)} \\ \lambda_i^* g_i(\mathbf{x}^*) &= 0\,, \quad i = 1, \ldots, m \quad &&\text{(complementary slackness)} \\ \nabla f(\mathbf{x}^*) + \textstyle\sum_{i=1}^m \lambda_i^* \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^n \nu_j^* \nabla h_j(\mathbf{x}^*) &= \mathbf{0} &&\text{(gradient condition)} \end{aligned}$$

In other words, under strong duality, the above conditions are *necessary* for optimality of $\mathbf{x}^*$ and $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$.

If the primal problem is convex, then the KKT conditions are also *sufficient* for points $\mathbf{x}^*$ and $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ to be optimal; see [1] for details. Thus in particular, if the primal problem is convex and Slater's condition holds, then the KKT conditions are both necessary and sufficient for optimality.

In some special cases, the KKT conditions can be solved analytically in order to obtain a solution to the primal problem; in many other cases, some of the solvers one might use in practice can be viewed as trying to solve these conditions.

Example 17. Consider the following convex quadratic program over $\mathbb{R}^d$ with equality constraints only:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c \\ \text{subject to} & \mathbf{C}\mathbf{x} = \mathbf{d}\,, \end{array}$$

where $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a symmetric positive semi-definite matrix, $\mathbf{b} \in \mathbb{R}^d$, $c \in \mathbb{R}$, $\mathbf{C} \in \mathbb{R}^{n \times d}$, and $\mathbf{d} \in \mathbb{R}^n$. The KKT conditions are

$$\mathbf{C}\mathbf{x}^* = \mathbf{d}\,, \qquad \mathbf{A}\mathbf{x}^* + \mathbf{b} + \mathbf{C}^\top \boldsymbol{\nu}^* = \mathbf{0}\,,$$

and can be written as

$$\begin{bmatrix} \mathbf{A} & \mathbf{C}^\top \\ \mathbf{C} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^* \\ \boldsymbol{\nu}^* \end{bmatrix} = \begin{bmatrix} -\mathbf{b} \\ \mathbf{d} \end{bmatrix}.$$

Solving this system of $d + n$ linear equations in $d + n$ variables gives the optimal primal and dual variables $\mathbf{x}^*, \boldsymbol{\nu}^*$.

Acknowledgments

Part of the material in these notes is based on the textbook of Boyd and Vandenberghe [1].

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.