**CS292A Convex Optimization: Gradient Methods and Online Learning** **Spring 2019**

# Lecture 9: May 7

*Lecturer: Yu-Xiang Wang* *Scribes: Mert Kosan, Zexi Huang*

**Note**: *LaTeX template courtesy of UC Berkeley EECS dept.*

**Disclaimer**: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

## 9.1 KKT conditions

For general optimization problem:

$$\min_x f(x), \text{subject to } h_i(x) \leq 0, i = 1, ..., m, l_j(x) = 0, j = 1, ..., r \tag{9.1}$$

The Karush-Kuhn-Tucker conditions or KKT conditions are:

- Stationarity: $0 \in \partial \left( f(x) + \sum_{i=1}^m u_i h_i(x) + \sum_{j=1}^r v_j l_j(x) \right)$
- Complementary slackness: $u_i \cdot h_i(x) = 0, \forall i$
- Primal feasibility: $h_i(x) \leq 0, l_j(x) = 0, \forall i, j$
- Dual feasibility: $u_i \geq 0, \forall i$

**Theorem 9.1** *Sufficiency and necessity of KKT conditions for characterizing optimal solutions:*

- *Necessity: If $x^*, u^*$ and $v^*$ are optimal solutions for primal and dual with zero duality gap (e.g., convex problem and there exists $x$ strictly satisfying non-affine inequality constraints), then $x^*, u^*$ and $v^*$ satisfy the KKT conditions.*
- *Sufficiency: If $x^*, u^*$ and $v^*$ satisfy the KKT conditions, then $x^*, u^*$ and $v^*$ are optimal solutions for primal and dual.*

**Proof:**

- Necessity (Optimality ⇒ KKT)

  Let $x^*$ and $u^*, v^*$ be primal and dual solutions with zero duality gap (strong duality holds). Then

$$\begin{aligned} f(x^*) &= g(u^*, v^*) \\ &= \min_x f(x) + \sum_{i=1}^m u_i^* h_i(x) + \sum_{j=1}^r v_j^* l_j(x) \\ &\leq f(x^*) + \sum_{i=1}^m u_i^* h_i(x^*) + \sum_{j=1}^r v_j^* l_j(x^*) \\ &\leq f(x^*) \end{aligned} \tag{9.2}$$

which means all the inequalities are equality here. We learn:

- From the second inequality being equality, we know the point $x^*$ minimizes $L(x, u^*, v^*)$ over $x \in \mathbb{R}$. Hence the subdifferential of $L$ must contain 0 at $x = x^*$, i.e., stationarity holds.
- From the first inequality being equality, we must have $\sum_{i=1}^m u_i^* h_i(x^*) = 0$, i.e., complementary slackness holds.
- Primal and dual feasibility hold as the optimal solutions are feasible.

Therefore, we have the optimal solutions satisfy the KKT conditions

- Sufficiency (KKT ⇒ Optimality)

  Consider $x^*, u^*, v^*$ that satisfy the KKT conditions, then

$$\begin{aligned} g(u^*, v^*) &= f(x^*) + \sum_{i=1}^m u_i^* h_i(x^*) + \sum_{j=1}^r v_j^* l_j(x^*) \\ &= f(x^*) \end{aligned} \tag{9.3}$$

  where the first equality holds from stationarity, and the second holds from complementary slackness and primal feasibility.

  Therefore, we have the points that satisfy the KKT conditions are optimal solution for the problem.

■

Note that this KKT conditions are for characterizing global optima. There are other versions of KKT conditions that deal with local optima.

For unconstrained problems, the KKT conditions reduce to subgradient optimality condition, i.e.,

$$0 \in \partial f(x^*) \tag{9.4}$$

And for general convex problems, it's equivalent to the subgradient optimality condition of the sum of objective function and indicator functions of the constraints, i.e.,

$$0 \in \partial f(x^*) + \sum_{i=1}^m \mathcal{N}_{\{h_i \leq 0\}}(x^*) + \sum_{j=1}^r \mathcal{N}_{\{l_j = 0\}}(x^*) \tag{9.5}$$

## 9.2 Examples

***Quadratic with equality constraints***

Consider $Q \succeq 0$,

$$\min_x \frac{1}{2} x^T Q x + c^T x \ \text{ subject to } Ax = 0 \tag{9.6}$$

KKT Conditions:

- $L(x,u) = \frac{1}{2} x^T Q x + c^T x + u^T A x$
- Stationary: $\nabla L(x,y) = 0 \Rightarrow Qx + c + A^T u = 0$
- Complementary Slackness: $\emptyset$
- Feasibility: $Ax = 0$

If we combine stationary and feasibility, we will have:

$$\begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} = \begin{pmatrix} -c \\ 0 \end{pmatrix}$$

***Water Filling***

$$\min_x - \sum_{i=1}^n \log(\alpha_i + x_i) \ \text{ subject to } x \geq 0, \ \ 1^T x = 1 \tag{9.7}$$

KKT Conditions:

- $L(x,u,v) = -\sum_{i=1}^n \log(\alpha_i + x_i) + u^T(-x) + v(1^T x - 1)$
- Stationary: $\nabla L(x,u,v) = 0 \Rightarrow \frac{-1}{\alpha_i + x_i} - u_i + v = 0, \ i = 1, ..., n$
- Complementary Slackness: $u_i \cdot x_i = 0, \ i = 1, ..., n$
- Feasibility: $x \geq 0, \ \ 1^T x = 1, \ \ u \geq 0$

***Support vector machines***

$$\min_{\beta,\beta_0,\xi} \quad \frac{1}{2}||\beta||_2^2 + C\sum_{i=1}^{n}\xi_i \text{ subject to } \xi_i \geq 0, \quad y_i(x_i^T\beta + \beta + 0) \geq 1 - \xi_i, \quad i = 1, ...n \tag{9.8}$$

KKT Conditions:

- Stationary: $0 = \sum_{i=1}^n w_i y_i, \quad \beta = \sum_{i=1}^n w_i y_i x_i, \quad w = C1 - v$
- Complementary slackness: $v_i\xi_i = 0, \quad w_i(1 - \xi_i - y_i(x_i^T\beta + \beta_0)) = 0, \quad i = 1, ..., n$

So we know $\beta = \sum_{i=1}^n w_i y_i x_i$ at optimality and $w_i$ is nonzero only if $y_i(x_i^T\beta + \beta_0) = 1 - \xi_i$. Then such points i are called the **support points**:

- For support point $i$, if $\xi_i = 0$, then $x_i$ lies on edge of margin, and $w_i \in (0, C]$
- For support point $i$, if $\xi \neq 0$, then $x_i$ lies on wrong side of margin, and $w_i = C$

***Lasso Support Recovery***

$$\min_{\beta,\alpha} \frac{1}{2}||X\beta - y||^2 + \lambda||\alpha||_1 \text{ subject to } \alpha = \beta \tag{9.9}$$

KKT Conditions:

- $L(\beta, \alpha, u) = \min_{\beta,\alpha} \frac{1}{2}||X\beta - y||^2 + \lambda||\alpha||_1 + u^T(\alpha - \beta)$
- Stationary: $\nabla L(\beta, \alpha, u) = 0 \Rightarrow X^T(X\beta - y) - u = 0$

$$-u \in \lambda\partial||\alpha||_1 = \left\{ \begin{array}{ll} [-\lambda, \lambda], & \text{if } \alpha = 0 \\ \lambda(\alpha), & \text{otherwise} \end{array} \right\}$$

- Feasibility: $\alpha = \beta$

And we have the followings conditions,

$$X^T(X_s\beta_s^* - y) = u$$

$$u_i \in \left\{ \begin{array}{ll} [-\lambda, \lambda], & \text{if } \beta_i^* = 0 \\ \lambda sign(\beta_i^*), & \text{otherwise} \end{array} \right\}$$

where $\beta_i^*$ is the optimal $\beta_i$.

## 9.3 Constrained and Lagrange forms

Constrained form, where $t \in \mathbb{R}$ is a tuning parameter,

$$\min_x f(x), \text{subject to } h(x) \leq t \tag{9.10}$$

For the optimal solution $x_C^*$, by stationarity of the KKT conditions we have

$$0 \in \partial f(x_C^*) + u \cdot \partial h(x_C^*) \tag{9.11}$$

Lagrange form, where $\lambda \geq 0$ is a tuning parameter,

$$\min_x f(x) + \lambda h(x) \tag{9.12}$$

For the optimal solution $x_L^*$, the subdifferential condition requires:

$$0 \in \partial f(x_L^*) + \lambda \cdot \partial h(x_L^*) \tag{9.13}$$

From these we see that $x_C^*$ and $x_L^*$ are characterized by the same equations. More rigorously, we have

- If the constrained problem is strictly feasible, then strong duality holds, and there exists some $\lambda \geq 0$ such that any solution $x^*$ of the constrained problem minimizes

  $$f(x) + \lambda \cdot (h(x) - t) \tag{9.14}$$

  so $x^*$ is also a solution for the Lagrange problem.

- If $x^*$ is a solution for the Lagrange problem, then the KKT conditions for the constrained problem are satisfied by taking $t = h(x^*)$, so $x^*$ is also a solution for the constrained problem.

## 9.4 Usage of Duality and KKT conditions

- Duality gap: zero duality gap implies optimality as

  $$f(x) - f(x^*) \leq f(x) - g(u, v) \tag{9.15}$$

  This can also be a stopping criterion in algorithms.

- Solving the primal via the dual: Under strong duality, we know any primal solution $x^*$ solves

  $$\min_x f(x) + \sum_{i=1}^m u_i^* h_i(x) + \sum_{j=1}^r v_j^* l_j(x) \tag{9.16}$$

  Often, solutions of this unconstrained problem can be expressed explicitly, thus giving an explicit characterization of primal solutions from dual solutions. Further, if this solution is unique, we know it must be the primal solution $x^*$.

  Example:

  $$\min_x \sum_{i=1}^n f_i(x_i), \text{subject to } a^T x = b \tag{9.17}$$

where each $f_i$ is smooth, strictly convex. We can find the dual function as

$$g(v) = bv - \sum_{i=1}^{n} f_i^*(a_i v) \tag{9.18}$$

where $f_i^*$ is the conjugate of $f_i$. The dual problem is

$$\max_v bv - \sum_{i=1}^{n} f_i^*(a_i v) \Leftrightarrow \min_v \sum_{i=1}^{n} f_i^*(a_i v) - bv \tag{9.19}$$

This is a convex minimization problem with scalar variable, and is thus much easier to solve than the primal.

Also, this additive form of objective function often comes from i.i.d data points, and is usually solved by stochastic gradient descent. Meanwhile, the dual can usually be solved with stochastic coordinate descent.

## 9.5 Dual norms, conjugate functions and dual cones

### 9.5.1 Dual norms

**Definition 9.2** *Dual norm $||x||_*$:*

$$||x||_* = \max_{||z|| \le 1} z^T x \tag{9.20}$$

Examples:

- $l_p$ norm dual: $(||x||_p)_* = ||x||_q$, where $1/p + 1/q = 1$
- Trace norm dual: $(||X||_{tr})_* = ||X||_{op} = \sigma_1(X)$

**Theorem 9.3** *Dual norm of dual norm is the original norm, i.e.,*

$$||x||_{**} = ||x|| \tag{9.21}$$

**Proof:** Consider the problem

$$\min_y ||y||, \text{subject to } y = x \tag{9.22}$$

whose optimal value is $||x||$. Lagrangian:

$$L(y, u) = ||y|| + u^T(x - y) = ||y|| - y^T u + x^T u \tag{9.23}$$

Maximizing $L$ while maintaining primal feasibility (which by definition of $|| \cdot ||_*$ requires $||u||_* \le 1$), we have the dual problem as

$$\max_u u^T x, \text{subject to } ||u||_* \le 1 \tag{9.24}$$

The optimal value of the dual problem is, again, by definition of $|| \cdot ||_*$, $||x||_{**}$. Thus, by strong duality we have

$$||x|| = ||x||_{**} \tag{9.25}$$

■

### 9.5.2 Conjugate functions

**Definition 9.4** *Conjugate function $f^*$ of $f$:*

$$f^*(y) = \max_x y^T x - f(x) \tag{9.26}$$

Properties:

- Always convex regardless of convexity of $f(x)$
- Fenchel's inequality: $f(x) + f^*(y) \geq x^T y$
- Conjugate of conjugate: $f^{**} \leq f$
- If $f$ is closed and convex, then
  - $f^{**} = f$
  - $x \in \partial f^*(y) \Leftrightarrow y \in \partial f(x) \Leftrightarrow f(x) + f^*(y) = x^T y$
- If $f(u, v) = f_1(u) + f_2(v)$, then $f^*(w, z) = f_1^*(w) + f_2^*(z)$

Examples:

***Simple quadratic***

$$f(x) = \frac{1}{2}x^T Q x$$
$$f^*(y) = \frac{1}{2}y^T Q^{-1} y$$

***Indicator function***

$$f(x) = I_C(x)$$
$$f^*(y) = I_C^*(y) = \max_{x \in C} y^T x$$

***Norm***

$$f(x) = ||x||$$
$$f^*(y) = I_{z:||z||_* \leq 1}(y)$$

***Lasso Dual***

Remember the lasso problem:

$$\min_\beta \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1$$

Its dual function is equal to $f^*$. First, we will change primal into another format:

$$\min_{\beta,z} \frac{1}{2}||y-z||_2^2 + \lambda||\beta||_1 \text{ subject to } z = X\beta$$

where dual function (so $f^*$) will be the following:

$$\begin{aligned} g(u) &= \min_{\beta,z} \frac{1}{2}||y-z||_2^2 + \lambda||\beta||_1 + u^T(z - X\beta) \\ &= \frac{1}{2}||y||_2^2 - \frac{1}{2}||y-u||_2^2 - I_{v:||v||_\infty \leq 1}(X^T u/\lambda) \end{aligned} \tag{9.27}$$

Conjugates also appear frequently in derivation of dual problems, via

$$-f^*(u) = \min_x f(x) - u^T x \tag{9.28}$$

in minimization of the Lagrangian.

For example, consider

$$\min_x f(x) + g(x) \tag{9.29}$$

The dual problem is

$$\max_u -f^*(-u) - g^*(-u) \tag{9.30}$$

We can find the dual of indicator functions and norms with the last equation.

Dual formulation can also help us by “shifting” a linear transformation between one part of the objective and another. Consider

$$\min_x f(x) + g(Ax) \tag{9.31}$$

By reparameterizing $z = Ax$, we have the dual problem as

$$\max_u -f^*(A^T u) - g^*(-u) \tag{9.32}$$

### 9.5.3 Dual cones and polar cones

**Definition 9.5** *For a cone $K \subseteq \mathbb{R}^n$, its dual cone $K^*$:*

$$K^* = \{y : y^T x \geq 0, \forall x \in K\} \tag{9.33}$$

Examples:

***Linear subspace***: The dual cone of a linear subspace $V$ is its orthogonal complement $V^\perp$.

***Norm cone***: The dual cone of the norm cone is the norm cone of its dual norm.

$$K = \{(x,t) \in R^{n+1} : ||x|| \leq t\}$$

$$K^* = \{(y,s) \in R^{n+1} : ||y||_* \leq s\}$$

***Positive semidefinite cone***: The convex cone $S_+^n$ is self-dual meaning $(S_+^n)^* = S_+^n$.

For the cone constrained problem

$$\min_x f(x), \text{subject to } Ax \in K \tag{9.34}$$

By denoting $K^*$ as the dual cone of $K$, we can write the dual problem as

$$\max_u -f^*(A^T u), \text{subject to } u \in K^* \tag{9.35}$$

**Definition 9.6** *For a cone $K \subseteq \mathbb{R}^n$, its polar cone $K^o$:*

$$K^o = \{y : y^T x \le 0, \forall x \in K\} \tag{9.36}$$

Following from the Moreau Decomposition, we have for any $x$,

$$x = \text{Proj}_K(x) + \text{Proj}_{K^o}(x) \tag{9.37}$$