E0 270 Machine Learning **Jan 17, 2015**

Optimization Tutorial 2

Newton's Method, Karush-Kuhn-Tucker (KKT) Conditions

Lecture by Harikrishna Narasimhan

Note: This tutorial shall assume background in elementary calculus and linear algebra.

In the first part of the tutorial, we introduced the problem of unconstrained optimization, provided necessary and sufficient conditions for optimality of a solution to this problem, and described the gradient descent method for finding a (locally) optimal solution to a given unconstrained optimization problem. We now describe another method for unconstrained optimization, namely Newton's method, that has better convergence guarantees than the gradient descent method in many settings. We will then start with the more challenging problem of constrained optimization, where we shall look at the Karush-Kuhn-Tucker (KKT) conditions for optimality of a solution to a constrained optimization problem.

1 Preliminaries

We begin with some preliminaries on convex sets and convex functions.

Convex set. A set $A \subseteq \mathbb{R}^d$ is said to be convex if for any $\mathbf{x}_1, \mathbf{x}_2 \in A$ and $\alpha \in (0, 1)$, the point $\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2$ is also in A.

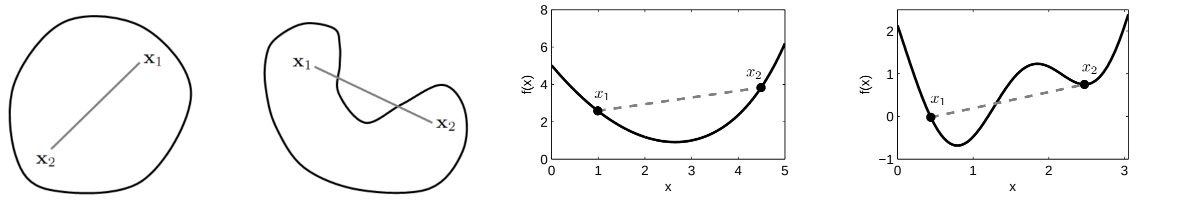
Thus a set in $\mathbb{R}^d$ is convex if every point on the line segment joining two given points in the set is also in the set. See Figure 1 for an example of a convex set.

Convex function. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be convex if for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ and $\alpha \in (0, 1)$, $f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha)f(\mathbf{x}_2)$.

From the above definition, it can be seen that the plot of a convex function between any two points will always lie (weakly) below the line joining the two points (see Figure 1 for an example). Moreover, one can show that a function is convex if and only if its Hessian is positive semi-definite at all points. A key property that follows from this is that *all local minimizers of a convex function are also its global minimizers*. Consequently, unlike a general optimization problem, for problems involving an unconstrained minimization of a convex function or a constrained minimization of a convex function over a convex constraint set (referred to as convex optimization problems), the global optimal solution to the problem can be found efficiently. We will see examples of such problems both in the current lecture and the next one.

2 Newton's Method

Following up on our discussion in our last lecture, we now present the Newton's method for unconstrained optimization. Let us first briefly review the gradient descent method for finding the minimizer of a function $f : \mathbb{R}^d \to \mathbb{R}$ in C^1. In this method, one starts with an initial point $\mathbf{x}_1 \in \mathbb{R}^d$ and at each iteration t obtains a new point by moving in the direction of the negative gradient of f at the current point: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)$, where $\eta_t > 0$ is the step-size at iteration t. As seen earlier, for appropriate choices of η_t's, this procedure always produces a decrease in the value of f after each iteration; this is evident from the first-order Taylor expansion of f, which tells us that for a small η_t, $f(\mathbf{x}_{t+1}) \approx f(\mathbf{x}_t) - \eta_t \nabla f(\mathbf{x}_t)^\top \nabla f(\mathbf{x}_t) < f(\mathbf{x}_t)$. Notice that, in fact, one can use a more general update rule $\mathbf{x}_{t+1} = \mathbf{x}_t + \eta_t \mathbf{d}_t$, with the direction $\mathbf{d}_t \in \mathbb{R}^d$ chosen such that $\nabla f(\mathbf{x}_t)^\top \mathbf{d}_t < 0$, and using the same argument, still guarantee a decrease in function value after



(a) A convex set (b) A non-convex set (c) A convex function (d) A non-convex function

Figure 1: Example of convex and non-convex sets, and convex and non-convex functions.

each iteration. We will call any direction that satisfies the mentioned condition as a *descent direction* at $\mathbf{x}_t$. Clearly, $\mathbf{d}_t = -\nabla f(\mathbf{x}_t)$ used in the gradient descent method satisfies this condition. We now describe an optimization procedure that uses a different descent direction at each iteration, namely Newton's method, which we shall see has better convergence properties than gradient descent in many settings.

Newton's method is based on the idea that a function $f : \mathbb{R}^d \to \mathbb{R}$ in C^2 can be approximated locally at any point by a quadratic function (given by the second-order Taylor expansion of f), and operates by minimizing a sequence quadratic approximations to f obtained at successive iterates. Let us assume that the given function $f \in C^2$ has a local minimizer $\mathbf{x}^*$ for which the Hessian of f is positive definite (p.d.), and observe that by continuity of the second partial derivatives of f, the Hessian of f at any point in a small neighborhood of $\mathbf{x}^*$ will also be p.d. Then given an initial point $\mathbf{x}_1$ in such a neighborhood of the minimizer $\mathbf{x}^*$, Newton's method computes a new point at each iteration t by minimizing the quadratic approximation of f at the current point:

$$\mathbf{x}_{t+1} \in \operatorname*{argmax}_{\mathbf{x} \in \mathbb{R}^d} \left\{ f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x} - \mathbf{x}_t) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{x} - \mathbf{x}_t) \right\}.$$

This gives us $\mathbf{x}_{t+1} = \mathbf{x}_t - \big[\nabla^2 f(\mathbf{x}_t)\big]^{-1} \nabla f(\mathbf{x}_t)$, where it can be shown using the positive definiteness of the Hessian $\nabla^2 f(\mathbf{x}_1)$ at the initial point that the Hessian $\nabla^2 f(\mathbf{x}_t)$ at each subsequent point is also p.d. and hence invertible. The method is outline below.

Newton's Method

Input: $f : \mathbb{R}^d \to \mathbb{R}$
Initialize: $\mathbf{x}_1 \in \mathbb{R}^d$
Parameter: T
for $t = 1$ **to** T
  $\mathbf{x}_{t+1} = \mathbf{x}_t - \big[\nabla^2 f(\mathbf{x}_t)\big]^{-1} \nabla f(\mathbf{x}_t)$
end for
Output: $\mathbf{x}_{T+1}$

The above update can be seen as moving in a direction $\mathbf{d}_t = -\big[\nabla^2 f(\mathbf{x}_t)\big]^{-1} \nabla f(\mathbf{x}_t)$ with unit step-size. Further, by positive definiteness of $\nabla^2 f(\mathbf{x}_t)$, we have that $\nabla f(\mathbf{x}_t)^\top \mathbf{d}_t = -\nabla f(\mathbf{x}_t)^\top \big[\nabla^2 f(\mathbf{x}_t)\big]^{-1} \nabla f(\mathbf{x}_t) < 0$, and hence $\mathbf{d}_t$ is indeed a descent direction. In fact, if the Hessian of f satisfies certain assumptions and the initial point $\mathbf{x}_1$ is sufficiently close to the minimizer $\mathbf{x}^*$, then it can be shown that these updates will produce a sequence of points with successively lower values of f and which converge to $\mathbf{x}^*$ [1].

Convergence rates. In many settings, Newton's method is known to have better convergence guarantees than the gradient descent method. We now discuss these guarantees for the case of convex functions [2]. In particular, if the Hessian of f at any point has eigen values that are bounded above and below by strictly positive values, then the number of iterations T taken by the gradient descent method to reach a solution $\mathbf{x}_T$ that has function value within an $\epsilon > 0$ of the minimum function value, i.e., for which $f(\mathbf{x}_T) - \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \leq \epsilon$, is at most $O(\log(1/\epsilon))$. On the other hand, if f satisfies an additional smoothness assumption and the initial point $\mathbf{x}_1$ is close enough to a minimizer of f, then Newton's method exhibits an exponentially faster convergence rate, requiring at most $O(\log\log(1/\epsilon))$ iterations to reach a solution whose function value is within an ϵ of the optimum value. Indeed it is not clear how one can obtain an initial point for Newton's method that is sufficiently close to a minimizer of f. In practice, starting from an arbitrary initial point, one can first run several iterations of the gradient descent method to obtain a point close to a minimizer of f and then use Newton's method to rapidly converge to the minimizer.

3 Constrained Optimization and KKT Optimality Conditions

We have so far discussed in much detail about unconstrained optimization problems, providing conditions for optimality of a solution to these problems, and describing iterative methods for solving these problems. However, in many real-life applications, one needs to optimize a function over a specified constraint set:

$$\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x}),$$

where $\mathcal{C} \subseteq \mathbb{R}^d$. The rest of the tutorial shall focus on such problems. We first extend the definition for a local minimizer, seen in the previous lecture, to constrained optimization problems.

Definition 1. *Let $\mathcal{C} \subseteq \mathbb{R}^d$. A point $\mathbf{x}^* \in \mathcal{C}$ is said to be a local minimizer of a function $f : \mathbb{R}^d \to \mathbb{R}$ over the constraint set $\mathcal{C}$ if there exists $\epsilon > 0$ such that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{C}$ within an ϵ-distance of $\mathbf{x}^*$, i.e., for all $\mathbf{x} \in \mathcal{C}$ such that $\|\mathbf{x} - \mathbf{x}^*\|_2 \leq \epsilon$.*

We shall be particularly interested in problems where the constraint set can be represented using a set of equality and inequality constraints:

$$\min_{\mathbf{x}\,\in\,\mathbb{R}^d} f(\mathbf{x}) \tag{OP1}$$

$$\begin{aligned} \text{s.t.} \quad h_i(\mathbf{x}) = 0, &\quad i \,=\, 1,\ldots,E, \\ g_j(\mathbf{x}) \leq 0, &\quad j \,=\, 1,\ldots,I, \end{aligned}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a function in C^1 and so are each of $h_i : \mathbb{R}^d \to \mathbb{R}$ and $g_j : \mathbb{R}^d \to \mathbb{R}$; also, f is assumed to take bounded values within the specified constraint set, i.e., $\exists B \in \mathbb{R}$ s.t. $f(\mathbf{x}) \geq B$, for all $\mathbf{x} \in \mathbb{R}^d$ that satisfy these constraints. In the rest of this lecture, we describe the Karush-Kuhn-Tucker (KKT) conditions for optimality of a solution to a problem of the above form. In the next lecture, we shall introduce a method for solving the above problem and explain the notion of duality in constrained optimization.

Recall that in an unconstrained problem, a necessary condition for a point to be a local minimizer of a function of interest is simply that the gradient of the function is zero. However, in a constrained problem setting, it is possible that the gradient of the given function is non-zero at every point that satisfies the specified constraints. Below, we provide a set of conditions, known as KKT conditions, that are necessarily satisfied by any local minimizer of a constrained optimization problem (assuming a mild regularity condition on the problem). Our explanation shall be inspired by the eloquent description provided for this topic in [3], and will be based on high-level intuitions rather than rigorous proofs.

KKT optimality conditions for equality constraints. Let us start with a simple optimization problem where the goal is to minimize a function f subject to a single equality constraint $h(\mathbf{x}) = 0$. An example of one such problem is shown in Figure 2. It is easy to see that the gradient of the constraint function h at any point on the constraint surface is perpendicular to the surface (as with points $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ in the figure). The optimality of such a point would depend on how the gradient of h at the point is oriented with respect to the negative gradient of the function f being optimized. For instance, consider a candidate point $\mathbf{x}'$ on the constraint surface where $\nabla h(\mathbf{x}')$ is neither parallel nor anti-parallel to $-\nabla f(\mathbf{x}')$. This would mean that the vector $-\nabla f(\mathbf{x}')$ has a non-zero component along a direction perpendicular to $\nabla h(\mathbf{x}')$ and hence along the surface of h. This also means that by taking an appropriate small step along the constraint surface one can decrease the value of f at $\mathbf{x}'$, clearly implying that $\mathbf{x}'$ is not a locally optimal solution for the given optimization problem. We illustrate this fact more clearly in Figure 2, where the negative gradient of f at a point $\mathbf{x}_1$ on the constraint surface has a component perpendicular to $\nabla h(\mathbf{x}_1)$, as a result of which one can traverse appropriately along the constraint surface (in this case to the right of $\mathbf{x}_1$) and reduce the value of f; on the other hand, the point $\mathbf{x}_2$, for which $\nabla h(\mathbf{x}_2)$ and $-\nabla f(\mathbf{x}_2)$ lie in the same direction, turns out to be an optimal solution to the problem. It can therefore be seen that a necessary condition for a point $\mathbf{x}^*$ (that satisfies the specified equality constraint) to be a local minimizer of the given optimization problem is that $\nabla h(\mathbf{x}^*)$ is either parallel or anti-parallel to the negative gradient $-\nabla f(\mathbf{x}^*)$, i.e., there exists $\mu \in \mathbb{R}$ such that $-\nabla f(\mathbf{x}^*) = \mu \nabla h(\mathbf{x}^*)$; the scalar μ is known as a *Lagrange multiplier.*

Example 1. *Consider a two-dimensional optimization problem where we seek to minimize a function $f(\mathbf{x}) = -x_1 - x_2$ subject to $h(\mathbf{x}) = x_1^2 + x_2^2 - 1 = 0$. This is a convex optimization problem with a unique optimal*

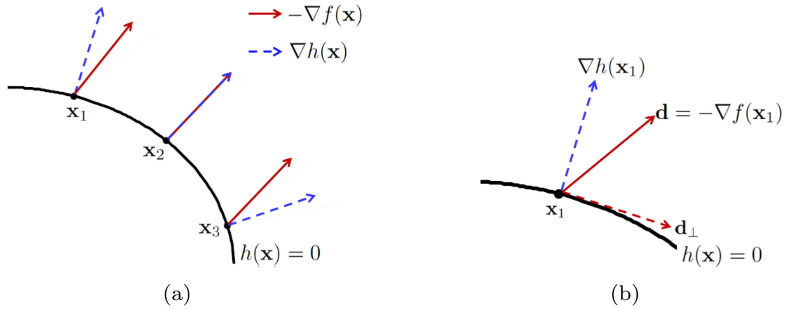


Figure 2: Illustration of KKT conditions for a two-dimensional optimization problem involving minimization of a function f subject to a single equality constraint $h(\mathbf{x}) = 0$. All three points $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ shown in (a) satisfy the given constraint. Notice that the gradients of h at $\mathbf{x}_1$ and $\mathbf{x}_3$ are neither parallel nor anti-parallel to the corresponding negative gradients of f at these points; consequently, by moving appropriately along the constraint surface one obtain points with lower values of f compared to $\mathbf{x}_1$ and $\mathbf{x}_3$. This is clearly seen in (b), where the vectors $\nabla h(\mathbf{x}_1)$ and $\mathbf{d} = -\nabla f(\mathbf{x}_1)$ are at acute angles with each other; thus $\mathbf{d}$ has a non-zero component in a direction perpendicular to $\nabla h(\mathbf{x}_1)$ and along the constraint surface, due to which moving along the constraint surface to the right of $\mathbf{x}_1$ yields a lower value of f. At point $\mathbf{x}_2$, on the other hand, both gradients lie in the same direction, and therefore no point on the constraint surface in a neighborhood around $\mathbf{x}_2$ has a lower function value; in fact, this point is the unique optimal solution to the given problem.

solution at $\mathbf{x}^* = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]$. *Clearly, this point satisfies the given equality constraint. The gradients of* f *and* g *at this point are then given by*

$$\nabla f(\mathbf{x}^*) = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \quad \text{and} \quad \nabla h(\mathbf{x}^*) = \begin{bmatrix} 2x_1^* \\ 2x_2^* \end{bmatrix}.$$

One can now verify that $\mathbf{x}^*$ *does indeed satisfy the necessary condition for optimality* $-\nabla f(\mathbf{x}^*) = \mu \nabla h(\mathbf{x}^*)$ *when the Lagrange multiplier* $\mu = \frac{1}{\sqrt{2}}$.

The above optimality condition can be generalized to a problem with more than one equality constraint. Suppose we are interested in a minimization problem involving a function f and a set of E equality constraints $h_i(\mathbf{x}) = 0$. Then a necessary condition for any point $\mathbf{x}^*$ (that satisfies the specified constraints and a mild regularity condition w.r.t. the constraints) to be an optimal solution of this problem is that the negative gradient of f at $\mathbf{x}^*$ is a linear combination of the gradients of the constraint functions h_i's at $\mathbf{x}^*$, i.e., there exists Lagrange multipliers $\mu_1, \ldots, \mu_E \in \mathbb{R}$ such that $-\nabla f(\mathbf{x}^*) = \sum_{i=1}^{E} \mu_i \nabla h_i(\mathbf{x}^*)$. Notice that this condition along with the E equality constraints gives us a set of $d + E$ equations with $d + E$ unknowns. In settings where the provided condition is also sufficient for optimality of $\mathbf{x}^*$ (examples of which, we shall see later in this section), one can find a local minimizer for the given problem by simply solving the given system of equations. The resulting optimization technique is called the *method of Lagrange multipliers*.

KKT Optimality Conditions for Inequality Constraints. We next move to optimization problems with inequality constraints. Let us start with a simple example where we wish to minimize a function f subject to a single inequality constraint $g(\mathbf{x}) \leq 0$. Figure 3 contains one such problem. There are now two possible cases for a candidate solution $\mathbf{x}'$ to this problem. The point $\mathbf{x}'$ could be in the interior of the constraint region (i.e., $g(\mathbf{x}') < 0$), in which case, the necessary condition for this point to be a local minimizer is same as in the unconstrained setting and is simply that $\nabla f(\mathbf{x}') = \mathbf{0}$. Alternatively, the point could be on the boundary of constraint region (i.e., $g(\mathbf{x}') = 0$), in which case by the same arguments that we had previously for equality constraints, $\mathbf{x}'$ cannot be a local minimizer if the vectors $\nabla g(\mathbf{x}')$ and $-\nabla f(\mathbf{x}')$ are neither parallel nor anti-parallel to each other. Moreover, with inequality constraints, $\mathbf{x}'$ cannot be a local minimizer even when $\nabla g(\mathbf{x}')$ is anti-parallel to $-\nabla f(\mathbf{x}')$ as in this case, one can take a short step into the constraint region in the direction pointed to by the negative gradient of f and decrease the function value (see Figure 3(b) for a clear illustration). Hence, for $\mathbf{x}'$ to be a local minimum, we require that both $\nabla g(\mathbf{x}')$ and $-\nabla f(\mathbf{x}')$ be in the same direction, i.e., that $-\nabla f(\mathbf{x}') = \lambda g(\mathbf{x}')$ for some Lagrange multiplier $\lambda > 0$. Note that in both case, $\lambda g(\mathbf{x}^*) = 0$. Combining the two cases, we have the following necessary conditions for local optimality of point $\mathbf{x}^*$ to the given problem: $\exists\, \lambda \geq 0$ such that $\lambda g(\mathbf{x}^*) = 0$ and $-\nabla f(\mathbf{x}^*) = \lambda g(\mathbf{x}^*)$.

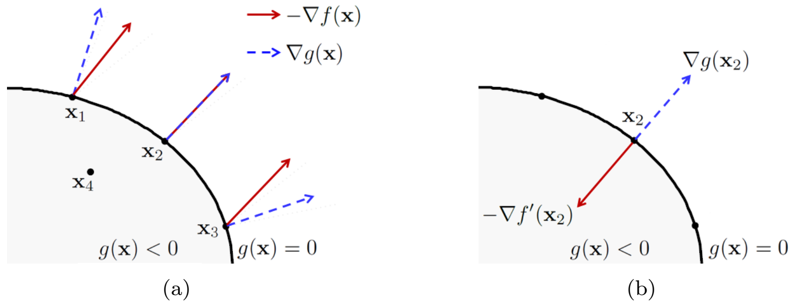


Figure 3: Illustration of KKT conditions for two-dimensional optimization problems where one wishes to minimize of a function f in (a) and f' in (b) subject to a single inequality constraint $g(\mathbf{x}) \leq 0$. The points $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ satisfy this constraint with strict equality and lie on the boundary of the constraint region, while $\mathbf{x}_4$ is in the interior. The point $\mathbf{x}_2$ is the unique optimal solution to the problem described in (a), with the gradient of the constraint function g at $\mathbf{x}_2$ being parallel to the negative gradient of f. However, in the case of (b), the gradient of g is in a direction opposite to the negative gradient of the function f' being optimized; thus one can find a point in the interior of the constraint set that has lower value of f' than $\mathbf{x}_2$.

Example 2. *Consider a two-dimensional (convex) optimization problem where we wish to minimize a function $f(\mathbf{x}) = x_1 + x_2$ subject to $g(\mathbf{x}) = x_1^2 + x_2^2 - 1 \leq 0$. The unique optimal solution for this problem is the point $\mathbf{x}^* = \left[-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right]$. Note that $\mathbf{x}^*$ satisfies the given constraint with strict equality, i.e., $g(\mathbf{x}^*) = 0$. The gradients of f and h at this point are then given by*

$$\nabla f(\mathbf{x}^*) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \nabla g(\mathbf{x}^*) = \begin{bmatrix} 2x_1^* \\ 2x_2^* \end{bmatrix}.$$

It can now be verified that for the positive Lagrange multiplier $\lambda = \frac{1}{\sqrt{2}}$, $\mathbf{x}^*$ does indeed satisfy the necessary conditions for optimality given above: $\lambda g(\mathbf{x}^*) = 0$ and $-\nabla f(\mathbf{x}^*) = \lambda \nabla g(\mathbf{x}^*)$.

For an optimization problem with more than one inequality constraint, namely that of minimizing a function f subject to constraints $g_j(\mathbf{x}) \leq 0$, $j = 1, \ldots, I$, a point $\mathbf{x}^*$ (that satisfies the specified constraints and a mild regularity condition) is a local minimizer if there exists Lagrange multipliers $\lambda_1, \ldots, \lambda_I \in \mathbb{R}$ with each $\lambda_i \geq 0$ such that $\lambda_j g_j(\mathbf{x}^*) = 0$ for all $j = 1, \ldots, I$ and $-\nabla f(\mathbf{x}^*) = \sum_{j=1}^{I} \lambda_j \nabla g_j(\mathbf{x}^*)$. Indeed these conditions can be generalized to an optimization problem with both equality and inequality constraints:

Proposition 1 (KKT conditions). Let $\mathbf{x}^*$ be a local minimizer of the optimization problem given in OP1. Also, assume that $\mathbf{x}^*$ satisfies a mild regularity condition w.r.t. the given constraints. Then there exists a set of Lagrange multipliers $\mu_1, \ldots, \mu_E \in \mathbb{R}$ and $\lambda_1, \ldots, \lambda_I \in \mathbb{R}$ such that:

$$\begin{aligned} h_i(\mathbf{x}^*) &= 0, \qquad i = 1, \ldots, E \\ g_j(\mathbf{x}^*) &\leq 0, \qquad j = 1, \ldots, I \\ \lambda_j &\geq 0, \qquad j = 1, \ldots, I \\ \lambda_j g_j(\mathbf{x}^*) &= 0, \qquad j = 1, \ldots, I \\ \nabla f(\mathbf{x}^*) + \sum_{i=1}^{E} \mu_i \nabla h_i(\mathbf{x}^*) + \sum_{j=1}^{I} \lambda_j \nabla g_j(\mathbf{x}^*) &= \mathbf{0}. \end{aligned}$$

These are called the Karush-Kuhn-Tucker (KKT) conditions, with the fourth condition known as the complementary slackness condition.

We have so far seen that KKT conditions are necessary for local optimality of a solution to a constrained optimization problem (provided a regularity condition is satisfied). It turns out that, under specific assumptions on the function f being optimized and the constraint functions h_i's and g_j's, these conditions are also sufficient for optimality. This is the case, for example, when f is a convex function and each h_i and g_j is affine. A more general characterization of problems where KKT conditions are sufficient for optimality can be found in standard text books (e.g., see Slater's condition in [2]).

4 Next Lecture

In the next class, we shall describe the projected gradient descent method for finding the minimizer of a constrained (convex) optimization problem. We will then discuss the concept of a dual problem for any given constrained optimization problem, which often provides a better understanding of the problem and in many cases is useful in developing efficient solvers for the original problem.

References

[1] D. Luenberger and Y. Ye. *Linear and Nonlinear Programming.* Springer, 3nd edition, 2008.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[3] C.M. Bishop. *Pattern Recognition and Machine Learning,* Appendix E. Springer, 2006.

Practice Exercise Questions

1. **Convexity.** Which among the following are convex optimization problems:

$$\min_{x_1,x_2} \; x_1 - x_2^2 \quad \text{s.t.} \quad x_1 + x_2 = 1;$$
$$\min_{x_1,x_2} \; x_1 + x_2 \quad \text{s.t.} \quad x_1 - x_2^2 \geq 1;$$
$$\min_{x_1,x_2} \; x_1 + x_2 \quad \text{s.t.} \quad x_1 - x_2^2 = 1.$$

2. **KKT conditions and method of Lagrange multipliers.** In a supervised learning algorithm that you are designing for your course project, you are required to compute the projection of a point $\mathbf{u} \in \mathbb{R}^d$ on to a hyperplane $\{\mathbf{x} \in \mathbb{R}^d \,|\, \mathbf{a}^\top \mathbf{x} = b\}$, where $\mathbf{a} \in \mathbb{R}^d$ and $b \in \mathbb{R}$. This projection operation can be framed as the following constrained optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \; \|\mathbf{x} - \mathbf{u}\|_2^2 \quad \text{s.t.} \quad \mathbf{a}^\top \mathbf{x} = b.$$

Derive the KKT conditions for the minimizer $\mathbf{x}^*$ of the above problem. Are these conditions sufficient to guarantee optimality of $\mathbf{x}^*$? Solve the resulting system of equations to obtain a closed-form expression for $\mathbf{x}^*$. If the Lagrange multiplier associated with the constraint $\mathbf{a}^\top \mathbf{x} = b$ at the minimizer $\mathbf{x}^*$ is 0, what can you conclude about $\mathbf{u}$?