1 Review of optimization concepts

We are going to review some basic concepts in optimization, and some of the mathematics required to understand optimization.

Most principled machine learning begins with a formalization of the problem, and then the choice of algorithm to solve the formal problem. In essence, this is a focus on separating out the question from the answer when understanding a problem context. This can be thought of as a ‘control theory’ or ‘optimization’ perspective on the world: every soft specification can be formalized into a mathematical problem, which in turn can be solved.

Additionally, we’re going to build on this a little bit with a focus on the theory of optimization because we will frequently use human models. One of the most commonly used human models is as optimizers: humans are making decisions to maximize their utility. When we design learning algorithms around these human models, or to induce a certain interaction between these human models, it will often become a staged optimization: I optimize for my goal knowing that my actions will affect the human’s decision, and we assume the human’s decision is to optimize some function which depends on my actions, &c. It helps to understand optimality conditions, so oftentimes we can replace optimizations with a set of equality or inequality constraints.

Throughout this course, we will explore some other models for humans, but this is currently an area of active research, and humans-as-optimizers is one of the most developed frameworks, in part because it lends itself nicely to mathematical analysis.

Additionally, as a note for this course’s convention, we will focus on minimization. It will be good to remain consistent in this, since this will affect a lot of minus signs and such down the road. It is a more common convention to focus on minimization, at least in engineering. Whenever we have a maximization, we can write the equivalent minimization with the addition of one minus sign.

1.1 Examples

First, let’s go over some examples where optimization is used.

1.1.1 Loss-minimization machine learning

Given some data \( \{(X_i, Y_i)\}_{i \in [N]} \):

\[
\min_\theta \ell(X, Y; \theta)
\]

For example, in linear regression, we have the model \( f(x; \theta) = \theta^T x \), and the loss is simply \( \sum_{i=1}^N \| f(X_i; \theta) - Y_i \|^2_2 \).

1.1.2 Autonomous vehicles

We’ll cover a stylized car model, commonly used in the literature. The car is visualized in Figure 1.

\[
\begin{align*}
\dot{x} &= u_s \cos(\theta) \\
\dot{y} &= u_s \sin(\theta) \\
\dot{\theta} &= \frac{u_s}{L} \tan(u_\phi)
\end{align*}
\]

Here, \( x \in \mathbb{R} \) and \( y \in \mathbb{R} \) represent the location of the car on a 2D plane, and \( \theta \in [0, 2\pi) \) represents the heading of the car. \( u_s \) is the acceleration, \( u_\phi \) is the steering applied. \( L \) is the length of the car.

Depending on the make, model, and year of your stylized car, there will be variability in the allowable inputs. Here are common models studied theoretically.
Figure 1: A visualization of the car model commonly used in the literature. This image is taken from ‘Planning Algorithms’ by Steven M. LaValle.

- The tricycle allows \( u_s \in [-1, 1] \) and \( u_\phi \in [-\pi/2, \pi/2] \).
- The simple car allows \( u_s \in [-1, 1] \) and \( u_\phi \in (-\phi_{\text{max}}, \phi_{\text{max}}) \). This car has a minimum turning radius of \( \rho = L/\tan(\phi_{\text{max}}) \).
- The Reeds-Shepp car is like a simple car, only \( u \in \{-1, 0, 1\} \).
- The Dubins car only allows \( u_s \in \{0, 1\} \).

If we approximate this continuous system with forward Euler, using a step size of \( h \), we have the following discrete-time system:

\[
\begin{align*}
x[k+1] &= x[k] + hu_s[k] \cos(\theta[k]) \\
y[k+1] &= y[k] + hu_s[k] \sin(\theta[k]) \\
\theta[k+1] &= \theta[k] + h \frac{u_s[k]}{L} \tan(u_\phi[k])
\end{align*}
\]

For brevity, we let \( x[k] = (x[k], y[k], \theta[k]) \) and \( u[k] = (u_s[k], u_\phi[k]) \), as per typical control-theory convention.

We suppose we want to minimize some cost \( J \), e.g. reach some destination as quickly and/or fuel efficiently as possible, and also want to avoid some obstacles \( S^c \), e.g. don’t hit pedestrians, stay on the road. This leads to an optimization problem.

\[
\min_{u,x} J(x) \\
\text{subject to } x[k+1] = f(x[k], u[k]) \\
x \in S
\]

For those who are more interested in this application, you can visit Planning Algorithms by Steven M. LaValle for more detailed discussion.

1.1.3 Targeted advertising

Suppose we have \( N \) potential customers for our product. We’ll let \( [N] = \{1, 2, \ldots, N\} \) denote the set of users.

We can target user \( i \in [N] \) with an amount of advertising \( u_i \in [0, 1] \), and suppose that the marketing team knows that for advertisement level \( u_i \) on user \( i \) causes \( p_i(u_i) \in [0, 1] \) probability of conversion to a sale of a single unit. (Users only buy 0 or 1 units.)

A single unit sale creates profit \( \pi > 0 \) for the company, and each total unit of advertising costs \( b > 0 \).

The problem of maximizing the expected profits can be phrased as follows:

\[
\max_u \pi \sum_{i \in [N]} p_i(u_i) + b \sum_{i \in [N]} u_i \\
\text{subject to } u_i \in [0, 1] \text{ for } i \in [N]
\]
1.1.4 Consumer decisions

Suppose there are $N$ goods, and a user chooses to consume $x = (x_1, \ldots, x_N)$ of each good. If the user does so, they receive $u(x)$ utility. To maximize their utility subject to a budget constraint, we get:

$$\max_x u(x)$$

subject to $p^Tx \leq I$

Here $p$ is the price vector and $I$ is the total income of the consumer.

1.2 Real numbers

Throughout this course, and throughout most applications in practice, we will be optimizing across $\mathbb{R}^n$. Let’s quickly review what $\mathbb{R}^n$ is.

$\mathbb{R}$ is the set of real numbers. This can be thought of as just all the points on the number line.

$\mathbb{R}^n$ is a vector of $n$ real numbers. This can be thought of as the mathematical version of an array of length $n$.

So, $x \in \mathbb{R}^n$ can be written $(x_1, \ldots, x_n)$. (The most common convention is to use 1-indexing in this setting.)

So, $\mathbb{R}^1$ is just the real number line again. $\mathbb{R}^2$ is a vector with 2 numbers. The key thing that defines the dimensionality is how many orthogonal directions you can go.

So, $\mathbb{R}^0$ has only 1 point (null, which is analogous to the empty array) and 0 directions to go in.

$\mathbb{R}^1$ has an infinite number of points, and 2 directions to go in (plus and minus).

$\mathbb{R}^2$ has an infinite number of points, and an infinite number of directions to go in, but only 2 orthogonal directions.

Similarly, $\mathbb{R}^n$ has an infinite number of points, a infinite number of directions, and $n$ orthogonal directions.

1.3 First-order approximations

In $\mathbb{R}$: $f(y) \approx f(x) + f'(x)(y - x)$.

More generally in $\mathbb{R}^n$: $f(y) \approx f(x) + \langle \nabla f(x), y - x \rangle$.

Note that here, $x$ and $y$ are points, $\nabla f(x)$ is a vector representing an operator, and $y - x$ is a direction with magnitude.

Also note that $\nabla f(x)$ is pretty magical. It can take in any direction and tell you the derivative along that direction. This lets you handle the fact that for $n \geq 2$, $\mathbb{R}^n$ has an infinite number of directions.

When $y - x$ is small, this is actually a pretty good approximation.

We can visualize this for functions $\mathbb{R}^2 \to \mathbb{R}$.

1.4 Unconstrained optimization

Let’s consider the unconstrained optimization problem. Suppose we’re given some function $f : \mathbb{R}^n \to \mathbb{R}$. Then, in optimization, we want to find $\inf_{x \in \mathbb{R}^n} f(x)$. Additionally, we sometimes want to find a $x^{opt} \in \mathbb{R}^n$ such that $f(x^{opt}) = \inf_{x \in \mathbb{R}^n} f(x)$. Note that $\inf_{x \in \mathbb{R}^n} f(x)$ always exists, while, in general, $x^{opt}$ might not always exist and might not be unique if it exists.

We’ll begin with some definitions of the general concepts in unconstrained optimization. Much of this content is drawn from [Bertsekas, 2016, Chapter 1], with some of the notation also borrowed from [Rockafellar, 1997].

**Definition 1.1** (Topological properties). For a point $x \in \mathbb{R}^n$ and radius $r > 0$, we define the open ball $B_r(x) = \{y : \|x - y\|_2 < r\}$.

For a set $A \subseteq \mathbb{R}^n$, we say $A$ is open if for every $x \in A$, there exists $\epsilon > 0$ such that $B_\epsilon(x) \subseteq A$.

For a set $A \subseteq \mathbb{R}^n$, we say $A$ is closed if $\mathbb{R}^n \setminus A$ is open.

**Definition 1.2** (Minima). Let $f : \mathbb{R}^n \to \mathbb{R}$.

A vector $x$ is an unconstrained local minimum of $f$ if there exists some $\epsilon > 0$ such that $f(x) \leq f(y)$ for all $y \in B_\epsilon(x)$. This is strict if the inequality holds strictly for all $y \neq x$.

A vector $x$ is an unconstrained global minimum of $f$ if $f(x) \leq f(y)$ for all $y \in \mathbb{R}^n$. This is strict if the inequality holds strictly for all $y \neq x$.

**Definition 1.3** (Differentiation and stationary points). Let $f : \mathbb{R}^n \to \mathbb{R}$.

The directional derivative of $f$ is defined as:

$$f'(x; y) = \lim_{\lambda \to 0} \frac{f(x + \lambda y) - f(x)}{\lambda}$$
We say $f'(x; y)$ is the directional derivative of $f$ at $x$ in direction $y$.

Recall $e_i$ denotes the unit vector in the $i$th coordinate. If $f'(x; e_i) = f'(x; -e_i)$, we define the partial derivative at $x$ with respect to $x_i$ as:

$$
\frac{\partial f}{\partial x_i}(x) = f'(x; e_i)
$$

If $f'(x; e_i) \neq f'(x; -e_i)$, this partial derivative is not defined.

If there exists a vector $x^*$ such that $\langle x^*, y \rangle = f'(x; y)$ for all $y$, then we say $f$ is differentiable at $x$. If such an $x^*$ exists, it must necessarily equal the partial derivatives with respect to each coordinate. Thus, when $f$ is differentiable at $x$, we define the gradient $\nabla f$ at $x$ as:

$$
\nabla f(x) = x^* = \left( \frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right)
$$

We say $f$ is differentiable if $f$ is differentiable at $x$ for all $x \in \mathbb{R}^n$. If $\nabla f$ is continuous, we say $f$ is continuously differentiable.

We can also define partial second derivatives. If we introduce the shorthand $g = \frac{\partial f}{\partial x_j}$, then:

$$
\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \frac{\partial}{\partial x_i} \left( \frac{\partial f}{\partial x_j}(x) \right) = \frac{\partial g}{\partial x_i}(x)
$$

Note that this definition requires the existence of all the partial derivatives.

If all the partial second derivatives exist, then we say $f$ is twice differentiable at $x$, and we define the Hessian $\nabla^2 f$ element-wise:

$$
(\nabla^2 f(x))_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x)
$$

We say $f$ is twice differentiable if $f$ is twice differentiable at $x$ for all $x \in \mathbb{R}^n$. If $\nabla^2 f$ is continuous, we say $f$ is twice continuously differentiable.

We say $x$ is a stationary point of $f$ if $\nabla f(x) = 0$. Note that this implicitly requires that $\nabla f(x)$ exists.

With these definitions, we note some basic properties.

**Proposition 1.4.** If $f$ is twice-differentiable at $x$, then $\nabla^2 f(x)$ is symmetric, i.e. $\nabla^2 f(x) = (\nabla^2 f(x))^\top$.

We also write out necessary and sufficient conditions for optimality.

**Proposition 1.5** (Necessary optimality conditions). Let $x^{\text{opt}}$ be an unconstrained local minima of $f : \mathbb{R}^n \to \mathbb{R}$. Suppose $f$ is continuously differentiable in an open set $A$ containing $x^{\text{opt}}$. Then $\nabla f(x^{\text{opt}}) = 0$.

Furthermore, if $f$ is twice continuously differentiable within $A$, then $\nabla^2 f(x^{\text{opt}}) \succeq 0$.

**Proof.** Since $x^{\text{opt}}$ is an unconstrained local minimum, we have that $f'(x^{\text{opt}}; y) \geq 0$ for all $y$. Fix any $y \neq 0$, note that $\langle \nabla f(x^{\text{opt}}), y \rangle = f'(x^{\text{opt}}; y) \geq 0$ by Proposition 1.4, and $-\langle \nabla f(x^{\text{opt}}), y \rangle = \langle \nabla f(x^{\text{opt}}), -y \rangle \geq 0$, which implies that $\langle \nabla f(x^{\text{opt}}), y \rangle = 0$. This held for any $y$, which implies that $\nabla f(x^{\text{opt}}) = 0$.

For the second half of this proof, suppose $f$ is twice continuously differentiable. Fixing a vector $y$, we have the Taylor series expansion for any $\alpha \in \mathbb{R}$:

$$
f(x^{\text{opt}} + \alpha y) = f(x^{\text{opt}}) + \alpha \langle \nabla f(x^{\text{opt}}), y \rangle + \frac{\alpha^2}{2} y^\top \nabla^2 f(x^{\text{opt}}) y + o(\alpha^2)
$$

By the first half of this proof, we know $\nabla f(x^{\text{opt}}) = 0$, so, using optimality:

$$
0 \leq \frac{f(x^{\text{opt}} + \alpha y) - f(x^{\text{opt}})}{\alpha^2} = \frac{1}{2} y^\top \nabla^2 f(x^{\text{opt}}) y + \frac{o(\alpha^2)}{\alpha^2}
$$

Taking the limit $\alpha \to 0$ yields $y^\top \nabla^2 f(x^{\text{opt}}) y \geq 0$, and this held for arbitrary $y$, so $\nabla^2 f(x^{\text{opt}}) \succeq 0$.

**Proposition 1.6** (Sufficient optimality conditions). Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable in some open set $A$. Suppose a vector $x \in S$ satisfies $\nabla f(x) = 0$ and $\nabla^2 f(x) \succeq 0$. Then, $x$ is a strict unconstrained local minimum of $f$.

Even stronger, there exists scalars $\gamma > 0$ and $\epsilon > 0$ such that:

$$
f(y) \geq f(x) + \frac{\gamma}{2} \| y - x \|^2 \text{ for all } y \in B_\epsilon(x)
$$

4
Proof. Let $\lambda$ be the smallest eigenvalue of $\nabla^2 f(x)$. We know that $\lambda > 0$ since $\nabla^2 f(x) > 0$. Furthermore, from the eigenvalue decomposition of $\nabla^2 f(x)$, we can see that $y^T \nabla^2 f(x)y \geq \lambda \|y\|_2^2$ for any $y \in \mathbb{R}^n$. The second order Taylor series is, for any $y$:

$$f(x + y) - f(x) = \langle \nabla f(x), y \rangle + \frac{1}{2} y^T \nabla^2 f(x) y + o(\|y\|_2^2)$$

$$\geq \frac{\lambda}{2} \|y\|_2^2 + o(\|y\|_2^2)$$

$$= \left( \frac{\lambda}{2} + o(\|y\|_2^2) \right) \|y\|_2^2$$

Since $\lim_{\|y\|_2 \to 0} o(\|y\|_2^2)/\|y\|_2^2 = 0$, we can pick $\epsilon > 0$ and $\gamma > 0$ such that, for all $\|y\|_2 < \epsilon$, we have:

$$\left( \frac{\lambda}{2} + o(\|y\|_2^2) \right) \geq \frac{\gamma}{2}$$

This is our desired result. 

1.5 Gradient descent

Now, we quickly discuss some algorithms for solving unconstrained optimization. This material is also drawn from [Bertsekas, 2016, Chapter 1].

First, we discuss gradient descent. At a high level, we want our algorithm to pick a sequence $(x_0, x_1, \ldots)$ such that $f(x_{k+1}) < f(x_k)$. The hope is that this will continue to a local minima. The question is: can we find a direction of descent using local information? Gradient descent is built on the assumption that we can.

Generally, gradient descent algorithms take step sizes $\alpha_k > 0$ and assign iterates:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

Why should this work nicely? For brevity, we’ll write $x^+ = x - \alpha \nabla f(x)$. Then, the first order Taylor series yields:

$$f(x^+) = f(x) + \langle \nabla f(x), x^+ - x \rangle + o(\|x^+ - x\|_2)$$

$$= f(x) - \alpha \|\nabla f(x)\|_2^2 + o(\alpha \|\nabla f(x)\|_2)$$

For small $\alpha$, this provides a descent direction.

More generally, we can think of gradient-based descent methods. Note that above, if we pick any direction $d$ such that $\langle \nabla f(x), d \rangle < 0$, and consider $x^+ = x + \alpha d$, we have:

$$f(x^+) = f(x) + \alpha \langle \nabla f(x), d \rangle + o(\alpha)$$

Commonly, we choose $d = -D \nabla f(x)$ for some positive definite matrix $D$. These are known as quasi-Newton methods. Note that, in this case, we always have $\langle \nabla f(x), d \rangle = -\nabla f(x)^T D \nabla f(x) < 0$. Also, note that gradient descent is a special case where $D = I$. Another common choice is a diagonal matrix $D$.

Finally, if we assume $\nabla^2 f(x_k) > 0$ for $k = 0, 1, \ldots$, we can take $D_k = (\nabla^2 f(x_k))^{-1}$, we have Newton’s method. This algorithm is interesting in its own regard, because if we take the quadratic approximation of $f$ around $x$, we get:

$$f(x^+) \approx f(x) + \langle \nabla f(x), x^+ - x \rangle + \frac{1}{2} (x^+ - x)^T \nabla^2 f(x)(x^+ - x)$$

To minimize the right-hand side, we would take the derivative with respect to $x^+$ and set it to zero:

$$\nabla f(x) + \nabla^2 f(x)(x^+ - x) = 0$$

We would then want:

$$x^+ = x - (\nabla^2 f(x))^{-1} \nabla f(x)$$

This provides a general motivation for using:

$$x^+ = x - \alpha (\nabla^2 f(x))^{-1} \nabla f(x)$$
1.6 Equality-constrained optimization and Lagrange multipliers

Let’s begin mulling over duality in earnest. Suppose we have the equality constrained optimization:

\[
\min \ f(x) \\
\text{subject to } h_i(x) = 0 \text{ for } i = 1, \ldots, m
\]

In this section, \( f : \mathbb{R}^n \to \mathbb{R} \) and \( h_i : \mathbb{R}^n \to \mathbb{R} \) for \( i = 1, \ldots, m \), and we assume that \( f \) and \((h_i)_{i \in [m]}\) are all continuously differentiable. For brevity, we will define \( h : \mathbb{R}^n \to \mathbb{R}^m \) such that \( h(x) = (h_1(x), \ldots, h_m(x)) \). This material is drawn from [Bertsekas, 2016, Chapter 3].

**Definition 1.7** (Feasible vectors). For the optimization problem in (1), we define the feasible set as \( \{ x : h(x) = 0 \} \).

**A point \( x \) such that \( h(x) = 0 \) is called a feasible point.**

**Proposition 1.8** (Lagrange multiplier theorem – necessary conditions). Suppose \( x^{opt} \) is a local minimum of \( f \) subject to \( h(x) = 0 \) (where \( f \) and \( h \) are both continuously differentiable), and the constraint gradients \( \nabla h_1(x^{opt}), \ldots, \nabla h_m(x^{opt}) \) are linearly independent.

Then there exists a unique vector \( \lambda^{opt} = (\lambda_1^{opt}, \ldots, \lambda_m^{opt}) \in \mathbb{R}^m \), called the Lagrange multiplier vector, such that:

\[
\nabla f(x^{opt}) + \sum_{i=1}^m \lambda_i^{opt} \nabla h_i(x^{opt}) = 0
\]

Let \( V(x^{opt}) = \{ y : (\nabla h_i(x^{opt}), y) = 0 \text{ for } i = 1, \ldots, m \} \) denote the subspace of first order feasible variations. If \( f \) and \( h \) are twice continuously differentiable as well, then:

\[
y^\top \left( \nabla^2 f(x^{opt}) + \sum_{i=1}^m \lambda_i^{opt} \nabla^2 h_i(x^{opt}) \right) y \geq 0 \text{ for all } y \in V(x^{opt})
\]

**Definition 1.9** (Regularity). For the optimization in (1) where \( h \) is continuously differentiable, a feasible vector \( x \) is regular if the constraint gradients \( \nabla h_1(x), \ldots, \nabla h_m(x) \) are linearly independent.

Before we prove the proposition, we prove one useful lemma.

**Lemma 1.10.** Consider the optimization in (1) and suppose \( f \) and \( h \) are continuous. Let \( x^{opt} \) be a local minimum, and take \( \epsilon > 0 \) such that \( f(x^{opt}) \leq f(y) \) for all \( y \in S = \{ x : \| x - x^{opt} \|_2 \leq \epsilon \} \). (Such an \( \epsilon \) exists by the definition of a local minimum.)

Let \( F^k(x) = f(x) + \frac{k_1}{2} \| h(x) \|_2^2 + \frac{\alpha}{2} \| x - x^{opt} \|_2^2 \) and, for each \( k = 1, \ldots, m \), let \( x^k \) be an optimal solution to:

\[
\min \ F^k(x) \\
\text{subject to } x \in S
\]

(A classical result in analysis, Weierstrass’ extreme value theorem, ensures that a continuous function is bounded and attains its extrema on a compact set. This implies that each \( x^k \) is well-defined.)

Then the sequence \( (x^k)_k \) converges to \( x^{opt} \).

**Proof.** We have that:

\[
F^k(x^k) = f(x^k) + \frac{k_1}{2} \| h(x^k) \|_2^2 + \frac{\alpha}{2} \| x^k - x^{opt} \|_2^2 \leq F^k(x^{opt}) = f(x^{opt})
\]

This, along with the fact that \( f \) is bounded, implies that \( \lim_{k \to \infty} \| h(x^k) \|_2 = 0 \). (We can see this by contradiction. Suppose \( \lim \sup_{k} \| h(x^k) \|_2 = \epsilon > 0 \), and let \( \gamma = \inf_{x \in S} f(x) \). Then, \( \| h(x^k) \|_2 \geq \epsilon / 2 \) infinitely often, i.e. there exists a subsequence \( (x^{n_k})_k \) such that \( F^{n_k}(x^{n_k}) \geq \gamma + \frac{\epsilon}{2} \frac{\alpha}{2} \). This is unbounded as \( n_k \to \infty \), violating Equation (2).)

Next, we show that every convergent subsequence of \( x^k \) converges to \( x^{opt} \). (We know a convergent subsequence exists by compactness.) Take any convergent subsequence \( (x^{n_k})_k \), and let \( \bar{x} \) be its limit. By continuity of \( h \) and closedness of \( S \), \( h(\bar{x}) = 0 \) and \( \bar{x} \in S \). Note that Equation (2) implies \( f(x^k) + \frac{\alpha}{2} \| x^k - x^{opt} \|_2^2 \leq f(x^{opt}) \), and, in the limit, using the continuity of the left-hand side, this means \( f(\bar{x}) + \frac{\alpha}{2} \| \bar{x} - x^{opt} \|_2^2 \leq f(x^{opt}) \).

Since \( \bar{x} \) is a feasible point and within \( S \), we have \( f(x^{opt}) \leq f(\bar{x}) \). This can only be the case when \( \| \bar{x} - x^{opt} \|_2 = 0 \), i.e. \( \bar{x} = x^{opt} \).

If every convergent subsequence of \( x^k \) converges to \( x^{opt} \), then \( x^k \) converges to \( x^{opt} \). (Suppose \( x^k \) did not converge to \( x^{opt} \). Then, \( \lim \sup_{k} \| x^k - x^{opt} \|_2 = \epsilon > 0 \), and we can find a subsequence that is always at least \( \epsilon / 2 \) away from \( x^{opt} \). This is a sequence in a compact set, so it must have a convergent subsequence. This is a subsequence of the original sequence \( (x^k)_k \), so it must converge to \( x^{opt} \), a contradiction.) 

\[\square\]
The nice thing about this lemma is that for $k$ sufficiently large, $x^k$ is in the interior of $S$, i.e. $x^k \in B_r(x^{\text{opt}})$. Thus, $x^k$ is an unconstrained local minimum of $F^k$ when $k$ is sufficiently large. Thus, we can work with our unconstrained necessary optimality conditions.

Proof of Proposition 1.8. For sufficiently large $k$, we can use the first order necessary condition of optimality in Proposition 1.5 to get:

$$0 = \nabla F^k(x^k) = \nabla f(x^k) + k \sum_{i=1}^m h_i(x^k) \nabla h_i(x^k) + \alpha (x^k - x^{\text{opt}}) \quad (3)$$

By a slight abuse of notation, let’s define $\nabla h(x^k)$ as the matrix whose $i$th column is $\nabla h_i(x^k)$. Since $x \mapsto \nabla h(x)$ is continuous and $(\nabla h_i(x^{\text{opt}}))_i$ are linearly independent, so are $(\nabla h_i(x^k))_i$ for $k$ sufficiently large. (The eigenvalues of a matrix are a continuous function of its entries.) Thus, $\nabla h(x^k) \in \mathbb{R}^{n \times m}$ has rank $m$ for $k$ sufficiently large. It follows that $\nabla h(x^k)\nabla h(x^k)$ is invertible. Also, note that we can write Equation (3):

$$0 = \nabla f(x^k) + k \nabla h(x^k)g(x^k) + \alpha (x^k - x^{\text{opt}}) \quad (4)$$

Left-multiplying Equation (4) by $(\nabla h(x^k)\nabla h(x^k))^{-1}\nabla h(x^k)^\top$ yields:

$$0 = (\nabla h(x^k)^\top\nabla h(x^k))^{-1}\nabla h(x^k)^\top (\nabla f(x^k) + \alpha (x^k - x^{\text{opt}})) + k \nabla h(x^k)$$

Since these derivatives are assumed to be continuous, we can take the limit as $k \to \infty$, and note that $x^k \to x^{\text{opt}}$. We can see that $k \nabla h(x^k)$ approaches a limit $\lambda^{\text{opt}}$:

$$\lambda^{\text{opt}} = \lim_{k \to \infty} k h(x^k) = -((\nabla h(x^{\text{opt}})^\top\nabla h(x^{\text{opt}}))^{-1}\nabla h(x^{\text{opt}})^\top\nabla f(x^{\text{opt}})$$

Thus, returning to Equation (3), and taking the limit as $k \to \infty$:

$$\nabla f(x^{\text{opt}}) + \nabla h(x^{\text{opt}})\lambda^{\text{opt}} = 0$$

This proves the first order Lagrange multiplier condition.

For the second order condition, we again used the unconstrained optimality conditions:

$$\nabla^2 F^k(x^k) = \nabla^2 f(x^k) + k \nabla h(x^k)\nabla h(x^k)^\top + k \sum_{i=1}^m h_i(x^k) \nabla^2 h_i(x^k) + \alpha I$$

Since for any $\alpha > 0$ and sufficiently large $k$, $x^k$ is the unconstrained optimum, we know that $\nabla^2 F^k(x^k) \succeq 0$.

If we take any $y \in V(x^{\text{opt}})$, i.e. $y$ such that $(\nabla h_i(x^{\text{opt}}), y) = 0$ for $i = 1, \ldots, m$, we let $y^k$ be the projection of $y$ onto the nullspace of $\nabla h(x^k)^\top$, i.e.:

$$y^k = y - \nabla h(x^k)(\nabla h(x^k)^\top\nabla h(x^k))^{-1}\nabla h(x^k)^\top y$$

This is a nice trick that allows us to treat $y$ as if it were orthogonal to $\nabla h(x^k)$, by replacing it with $y^k$ and taking a limit when we need to recover $y$. (Note that, by the relevant continuity, $y^k \to y$.)

Since $\nabla^2 F^k(x^k)$ is positive semidefinite and $(\nabla h_i(x^k), y^k) = 0$ for $i = 1, \ldots, m$:

$$0 \leq (y^k)^\top \nabla^2 F^k(x^k)y^k = (y^k)^\top \left( \nabla^2 f(x^k) + k \sum_{i=1}^m h_i(x^k) \nabla^2 h_i(x^k) \right) y^k + \alpha \|y^k\|_2^2$$

We take the limit as $k \to \infty$, and see that, from the first part of this proof, $k h_i(x^k) \to \lambda_i^{\text{opt}}$. Thus:

$$0 \leq y^\top \left( \nabla^2 f(x^{\text{opt}}) + \sum_{i=1}^m \lambda_i^{\text{opt}} \nabla^2 h_i(x^{\text{opt}}) \right) y + \alpha \|y\|_2^2 \text{ for all } y \in V(x^{\text{opt}})$$

This held for any $\alpha > 0$, so:

$$0 \leq y^\top \left( \nabla^2 f(x^{\text{opt}}) + \sum_{i=1}^m \lambda_i^{\text{opt}} \nabla^2 h_i(x^{\text{opt}}) \right) y \text{ for all } y \in V(x^{\text{opt}})$$

We are done.
Motivated by what happened above, we can define the Lagrangian.

**Definition 1.11 (Lagrangian).** The Lagrangian \( L : \mathbb{R}^{n+m} \rightarrow \mathbb{R} \) of optimization (1) is defined as:

\[
L(x, \lambda) = f(x) + \lambda^\top h(x)
\]

For a fixed \( \lambda \), let \( g_\lambda(x) = L(x, \lambda) \), and similarly \( g_\lambda(\lambda) = L(x, \lambda) \) for a fixed \( \lambda \). Then we define \( \nabla_x L(x, \lambda) = \nabla g_\lambda(x) \) and \( \nabla_\lambda L(x, \lambda) = \nabla g_\lambda(\lambda) \). Additionally, \( \nabla^2_{xx} L(x, \lambda) = \nabla^2 g_\lambda(x) \).

Thus, we can nicely summarize our necessary optimality conditions using the Lagrangian. We have that the first order necessary condition in Proposition 1.8 can be written:

\[
\nabla_x L(x^{\text{opt}}, \lambda^{\text{opt}}) = 0
\]

We have the equality constraint \( h(x^{\text{opt}}) = 0 \) summarized as:

\[
\nabla_\lambda L(x^{\text{opt}}, \lambda^{\text{opt}}) = 0
\]

The second order necessary condition is:

\[
y^\top \nabla^2_{xx} L(x^{\text{opt}}, \lambda^{\text{opt}}) y \geq 0 \text{ for all } y \in V(x^{\text{opt}})
\]

Thus, every local minimum \( x^{\text{opt}} \), which is regular, satisfies these 3 equations along with its associated Lagrangian multiplier vector.

Let’s look at the converse condition.

**Proposition 1.12 (Lagrange multiplier theorem – sufficient conditions).** Suppose \( f \) and \( h \) are twice continuously differentiable, and let \( x^{\text{opt}} \in \mathbb{R}^n \) and \( \lambda^{\text{opt}} \in \mathbb{R}^m \) satisfy:

\[
\begin{align*}
\nabla_x L(x^{\text{opt}}, \lambda^{\text{opt}}) &= 0 \\
\nabla_\lambda L(x^{\text{opt}}, \lambda^{\text{opt}}) &= 0 \\
y^\top \nabla^2_{xx} L(x^{\text{opt}}, \lambda^{\text{opt}}) y &> 0 \text{ for all } y \neq 0 \text{ such that } \langle \nabla h(x^{\text{opt}}), y \rangle = 0
\end{align*}
\]

Then \( x^{\text{opt}} \) is a strict local minimum of \( f \) subject to \( h(x) = 0 \). Even stronger, there exists \( \gamma > 0 \) and \( \epsilon > 0 \) such that:

\[
f(x) \geq f(x^{\text{opt}}) + \frac{\gamma}{2} \| x - x^{\text{opt}} \|^2 \text{ for all } x \text{ such that } h(x) = 0 \text{ and } \| x - x^{\text{opt}} \|_2 < \epsilon
\]

Note that regularity of \( x^{\text{opt}} \) is not required here. Before we prove Proposition 1.12, we’ll quickly note one lemma that essentially gives us our desired result. Essentially, Lemma 1.13 is exactly what we need to convert the assumed conditions in Proposition 1.12 into an equivalent setting where \( x^{\text{opt}} \) is an unconstrained optimum, allowing us to invoke the sufficiency conditions of Proposition 1.6.

**Lemma 1.13.** Let \( P \) and \( Q \) be two symmetric matrices. Suppose \( Q \) is positive semidefinite and \( P \) is positive definite on the nullspace of \( Q \), i.e. for any \( x \neq 0 \), \( x^\top Q x = 0 \) implies \( x^\top P x > 0 \). Then there exists a scalar \( \bar{c} \) such that \( P + cQ \succ 0 \) for all \( c > \bar{c} \).

**Proof.** Suppose not, i.e. for all \( \bar{c} \) there exists \( c > \bar{c} \) such that there is some \( x \neq 0 \) such that \( x^\top (P + cQ)x \leq 0 \). Thus, we can find a sequence of \( x^k \) such that \( \| x^k \|_2 = 1 \) and \( c_k \) strictly increasing with \( c_k \rightarrow \infty \) such that:

\[
(x^k)^\top P x^k + c_k (x^k)^\top Q x^k \leq 0
\]  

(5)

The sequence \( x^k \) is bounded, so there’s a convergence subsequence \( (x^{n_k})_k \), and, by continuity of the norm, its limit \( \bar{x} \) also satisfies \( \| \bar{x} \|_2 = 1 \). Thus, taking the limit superior of the inequality in (5):

\[
\bar{x}^\top P \bar{x} + \limsup_{k \to \infty} c_{n_k} (x^{n_k})^\top Q x^{n_k} \leq 0
\]

(6)

Since \( Q \succeq 0 \), we know that \( (x^{n_k})^\top Q x^{n_k} \geq 0 \), and therefore \( (x^{n_k})^\top Q x^{n_k} \to 0 \), since it is multiplied by the \( c_{n_k} \). Thus, \( \bar{x}^\top Q \bar{x} = 0 \), which, by assumption, implies \( \bar{x}^\top P \bar{x} > 0 \), which directly contradicts Equation 6.

Great, now let’s consider the augmented Lagrangian function.
**Definition 1.14 (Augmented Lagrangian).** We define the augmented Lagrangian function, with a scalar parameter \( c \) as:

\[
L_c(x, \lambda) = f(x) + \langle \lambda, h(x) \rangle + \frac{c}{2} ||h(x)||_2^2
\]

Note that the augmented Lagrangian is the Lagrangian of the following optimization:

\[
\min_x f(x) + \frac{c}{2} ||h(x)||_2^2
\]

subject to \( h(x) = 0 \) \hspace{1cm} (7)

Note this optimization is equivalent to (1), in the sense that the optimizers and optimal values are the same.

The gradient and Hessian with respect to \( x \) are:

\[
\nabla_x L_c(x, \lambda) = \nabla f(x) + \nabla h(x)(\lambda + ch(x))
\]

\[
\nabla^2_x L_c(x, \lambda) = \nabla^2 f(x) + \sum_{i=1}^m (\lambda_i + ch_i(x))\nabla^2 h_i(x) + c\nabla h(x)\nabla h(x)^\top
\]

**Proof of Proposition 1.12.** By assumption, \( \nabla_x L(x^{\text{opt}}, \lambda^{\text{opt}}) = \nabla f(x^{\text{opt}}) + \nabla h(x^{\text{opt}})\lambda^{\text{opt}} = 0 \) and \( \nabla_\lambda L(x^{\text{opt}}, \lambda^{\text{opt}}) = h(x^{\text{opt}}) = 0 \), so:

\[
\nabla_x L_c(x^{\text{opt}}, \lambda^{\text{opt}}) = \nabla f(x^{\text{opt}}) + \nabla h(x^{\text{opt}})(\lambda^{\text{opt}} + ch(x^{\text{opt}})) = \nabla_x L(x^{\text{opt}}, \lambda^{\text{opt}}) = 0
\]

Similarly,

\[
\nabla^2_{xx} L_c(x^{\text{opt}}, \lambda^{\text{opt}}) = \nabla^2 f(x^{\text{opt}}) + \sum_{i=1}^m \lambda_i^{\text{opt}}\nabla^2 h_i(x^{\text{opt}}) + c\nabla h(x^{\text{opt}})\nabla h(x^{\text{opt}})^\top
\]

\[
= \nabla^2 L(x^{\text{opt}}, \lambda^{\text{opt}}) + c\nabla h(x^{\text{opt}})\nabla h(x^{\text{opt}})^\top
\]

By assumption, \( y^\top \nabla^2_{xx} L(x^{\text{opt}}, \lambda^{\text{opt}}) y > 0 \) for all \( y \neq 0 \) such that \( y^\top \nabla h(x^{\text{opt}})\nabla h(x^{\text{opt}})^\top y = 0 \). Invoking Lemma 1.13 with \( P = \nabla^2_{xx} L(x^{\text{opt}}, \lambda^{\text{opt}}) \) and \( Q = \nabla h(x^{\text{opt}})\nabla h(x^{\text{opt}})^\top \), we have that there exists a \( \tilde{c} \) such that \( \nabla^2_{xx} L_c(x^{\text{opt}}, \lambda^{\text{opt}}) > 0 \) for all \( c > \tilde{c} \).

Thus, by Proposition 1.6, \( x^{\text{opt}} \) is an unconstrained local minimum of \( L_c(\cdot, \lambda^{\text{opt}}) \). In fact, there exists \( \gamma > 0 \) and \( \epsilon > 0 \) such that:

\[
L_c(x, \lambda^{\text{opt}}) \geq L_c(x^{\text{opt}}, \lambda^{\text{opt}}) + \frac{\gamma}{2} ||x - x^{\text{opt}}||_2^2 \text{ for all } x \text{ such that } ||x - x^{\text{opt}}||_2 < \epsilon
\]

Noting that when \( h(x) = 0, L_c(x, \lambda^{\text{opt}}) = f(x) \), this implies:

\[
f(x) \geq f(x^{\text{opt}}) + \frac{\gamma}{2} ||x - x^{\text{opt}}||_2^2 \text{ for all } x \text{ such that } h(x) = 0 \text{ and } ||x - x^{\text{opt}}||_2 < \epsilon
\]

Thus, \( x^{\text{opt}} \) is a strict local minimum of \( f \) subject to \( h(x) = 0 \). \( \square \)

This provides some of the theoretical underpinnings for some optimization algorithms: we can take the constrained minimization of (1) and implement it as an unconstrained optimization if \( \lambda^{\text{opt}} \) is known; in practice, \( \lambda^{\text{opt}} \) will not be known but can be approximated.

### 1.7 Inequality-constrained optimization and Karush-Kuhn-Tucker optimality conditions

Generalizing from the previous setup, let’s consider optimizations with equality and inequality constraints:

\[
\min_x f(x)
\]

subject to \( h_i(x) = 0 \) for \( i = 1, \ldots, m \)

\[
g_j(x) \leq 0 \text{ for } j = 1, \ldots, r
\]

Again, we’ll suppose \( f, h_i, g_j \) are continuously differentiable functions from \( \mathbb{R}^n \) to \( \mathbb{R} \), and, again, we’ll allow \( h : \mathbb{R}^n \to \mathbb{R}^m \) and \( g : \mathbb{R}^n \to \mathbb{R}^r \) to be the vector of these function evaluations, i.e. \( h(x) = (h_1(x), \ldots, h_m(x)) \) and \( g(x) = (g_1(x), \ldots, g_r(x)) \). As before, we can define feasibility:
Definition 1.15 (Feasible vectors). For the optimization problem in (8), we define the feasible set as \( \{ x : h(x) = 0, g(x) \leq 0 \} \). A point \( x \) such that \( h(x) = 0 \) and \( g(x) \leq 0 \) is called a feasible point.

Definition 1.16 (Active inequality constraints). For a feasible point \( x \), the set of active inequality constraints is \( A(x) = \{ j : g_j(x) = 0 \} \subset [r] \). If \( j \notin A(x) \), we say constraint \( j \) is inactive at \( x \).

Intuitively, if \( x^{\text{opt}} \) is a local minimum of (8), then \( x^{\text{opt}} \) is a local minimum of:

\[
\min_x f(x) \\
\text{subject to } h_i(x) = 0 \text{ for } i = 1, \ldots, m \\
g_j(x) \leq 0 \text{ for } j \in A(x^{\text{opt}})
\]

(This requires continuity of \( g \).) In this sense, inactive constraints do not matter as far as necessary conditions are concerned, and do not factor into optimality conditions.

On the other hand, active constraints can be treated as equality constraints, in the following sense. If \( x^{\text{opt}} \) is a local minimum of (8), then \( x^{\text{opt}} \) is a local minimum of:

\[
\min_x f(x) \\
\text{subject to } h_i(x) = 0 \text{ for } i = 1, \ldots, m \\
g_j(x) = 0 \text{ for } j \in A(x^{\text{opt}})
\]

From Proposition 1.8, if \( x^{\text{opt}} \) is a local optimum of (9) and regular, then there exists Lagrangian multipliers \((\lambda_1^{\text{opt}}, \ldots, \lambda_m^{\text{opt}})\) and \((\mu_j^{\text{opt}})_{j \in A(x^{\text{opt}})}\) such that:

\[
\nabla f(x^{\text{opt}}) + \sum_{i=1}^m \lambda_i^{\text{opt}} \nabla h_i(x^{\text{opt}}) + \sum_{j \in A(x^{\text{opt}})} \mu_j^{\text{opt}} \nabla g_j(x^{\text{opt}}) = 0
\]

This is more commonly written with \( \mu_j^{\text{opt}} = 0 \) for \( j \notin A(x^{\text{opt}}) \), yielding:

\[
\nabla f(x^{\text{opt}}) + \sum_{i=1}^m \lambda_i^{\text{opt}} \nabla h_i(x^{\text{opt}}) + \sum_{j=1}^r \mu_j^{\text{opt}} \nabla g_j(x^{\text{opt}}) = 0
\]

\[\mu_j^{\text{opt}} = 0 \text{ for all } j \notin A(x^{\text{opt}})\]

We will also see that \( \mu_j^{\text{opt}} \geq 0 \) for all \( j \in [r] \).

Definition 1.17 (Regularity). For the optimization in (8) where \( g \) and \( h \) are continuously differentiable, a feasible vector \( x \) is regular if the equality constraint gradients \( \nabla h_1(x), \ldots, \nabla h_m(x) \) and the active inequality constraint gradients \( \nabla g_j(x)_{j \in A(x)} \) are linearly independent.

Definition 1.18 (Lagrangian). The Lagrangian \( L : \mathbb{R}^{n+m+r} \to \mathbb{R} \) of optimization (8) is defined as:

\[
L(x, \lambda, \mu) = f(x) + \lambda^\top h(x) + \mu^\top g(x)
\]

Proposition 1.19 (Karush-Kuhn-Tucker necessary conditions). Let \( x^{\text{opt}} \) be a local minimum of (8), where \( f, h, g \) are continuously differentiable. Furthermore, suppose \( x^{\text{opt}} \) is regular. Then, there exists unique Lagrange multiplier vectors \( \lambda^{\text{opt}} = (\lambda_1^{\text{opt}}, \ldots, \lambda_m^{\text{opt}}) \) and \( \mu^{\text{opt}} = (\mu_1^{\text{opt}}, \ldots, \mu_r^{\text{opt}}) \) such that:

\[
\nabla_x L(x^{\text{opt}}, \lambda^{\text{opt}}, \mu^{\text{opt}}) = 0
\]

\[\mu_j^{\text{opt}} \geq 0 \text{ for } j = 1, \ldots, r\]

\[\mu_j^{\text{opt}} = 0 \text{ for all } j \notin A(x^{\text{opt}})\]

If, in addition, \( f, h, g \) are twice continuously differentiable, then for any \( y \) such that \( \langle \nabla h_i(x^{\text{opt}}), y \rangle = 0 \) for \( i = 1, \ldots, m \) and \( \langle \nabla g_j(x^{\text{opt}}), y \rangle = 0 \) for all \( j \in A(x^{\text{opt}}) \), we have:

\[y^\top \nabla^2_{xx} L(x^{\text{opt}}, \lambda^{\text{opt}}, \mu^{\text{opt}}) y \geq 0\]
The proof is similar to the proof of Proposition 1.8, only we take the penalized cost $F^k(x) = f(x) + \frac{k}{2} \|h(x)\|^2 + \frac{k}{2} \sum_{j=1}^r \left(\max\{0, g_j(x)\}\right)^2 + \frac{\alpha}{2} \|x - x^{\text{opt}}\|^2$. (We only wish to penalize when $g(x) > 0$.) A consequence of this modification is that $\mu^{\text{opt}} \geq 0$.

We can also introduce slack variables to convert this to the equality case:

$$\min_{x, z} f(x)$$

subject to $h_i(x) = 0$ for $i = 1, \ldots, m$

$$g_j(x) + z^2 = 0 \text{ for } j = 1, \ldots, r$$

Proposition 1.20 (Second order sufficiency conditions). If $f, h, g$ are twice continuously differentiable and $x^{\text{opt}} \in \mathbb{R}^n, \lambda^{\text{opt}} \in \mathbb{R}^m, \mu^{\text{opt}} \in \mathbb{R}^r$ satisfy:

$$\nabla_x L(x^{\text{opt}}, \lambda^{\text{opt}}, \mu^{\text{opt}}) = 0$$

$$h(x^{\text{opt}}) = 0$$

$$g(x^{\text{opt}}) \leq 0$$

$$\mu^{\text{opt}} \geq 0$$

$$\mu_j^{\text{opt}} = 0 \text{ for all } j \notin A(x^{\text{opt}})$$

Furthermore, assume that for any $y \neq 0$ such that $\langle \nabla h_i(x^{\text{opt}}), y \rangle = 0$ for $i = 1, \ldots, m$ and $\langle \nabla g_j(x^{\text{opt}}), y \rangle = 0$ for all $j \in A(x^{\text{opt}})$, we have:

$$y^T \nabla^2_{xx} L(x^{\text{opt}}, \lambda^{\text{opt}}, \mu^{\text{opt}}) y > 0$$

Still yet, assume $\mu_j^{\text{opt}} > 0$ for all $j \in A(x^{\text{opt}})$. Then, $x^{\text{opt}}$ is a strict local minimum of $f$ subject to $h(x) = 0, g(x) \leq 0$.

This can be proved using the sufficient conditions for the equality case on optimization (10).

We quickly discuss two more relevant things in general, constrained optimization before ending this section.

Proposition 1.21 ( Sufficiency condition for global optimum). Suppose $X \subset \mathbb{R}^n$ is given and we wish to solve:

$$\min_x f(x)$$

subject to $x \in X$

$$g_j(x) \leq 0 \text{ for } j = 1, \ldots, r$$

If $x^{\text{opt}}$ is a feasible vector, which, together with $\mu^{\text{opt}} = (\mu_1^{\text{opt}}, \ldots, \mu_r^{\text{opt}})$, satisfies:

$$\mu^{\text{opt}} \geq 0$$

$$\mu_j^{\text{opt}} = 0 \text{ for all } j \notin A(x^{\text{opt}})$$

$$x^{\text{opt}} \in \arg \min_{x \in X} L(x, \mu^{\text{opt}})$$

Then, $x^{\text{opt}}$ is a global minimum of the problem.

Proof. We have:

$$f(x^{\text{opt}}) = f(x^{\text{opt}}) + (\mu^{\text{opt}})^T g(x^{\text{opt}})$$

$$= \min_{x \in X} f(x) + (\mu^{\text{opt}})^T g(x)$$

$$\leq \min_{x \in X, g(x) \leq 0} f(x) + (\mu^{\text{opt}})^T g(x)$$

$$\leq \min_{x \in X, g(x) \leq 0} f(x)$$

By our assumptions, $(\mu^{\text{opt}})^T g(x^{\text{opt}}) = 0$, so the first equality follows. The second equality is by assumption; the next inequality is from contracting the feasible set, and the last inequality arises by noting that $\mu^{\text{opt}} \geq 0$ and $g(x) \leq 0$ implies $(\mu^{\text{opt}})^T g(x) \leq 0$.  \qed
Finally, we note a generalization of KKT conditions known as Fritz John optimality conditions. Note that in Proposition 1.19, we required $x^{\text{opt}}$ to be regular. Fritz John optimality conditions do not require such an assumption. Consider, again, the equality constrained optimization:

$$ \min_x f(x) $$

subject to $h_i(x) = 0$ for $i = 1, \ldots, m$

Proposition 1.8 states that if $x^{\text{opt}}$ is a local minimum and regular, there exists a unique $\lambda^{\text{opt}}$ such that:

$$ \nabla f(x^{\text{opt}}) + \sum_{i=1}^m \lambda_i^{\text{opt}} \nabla h_i(x^{\text{opt}}) = 0 $$

This is equivalent to stating that, if $x^{\text{opt}}$ is a local minimum of (1), then there exists scalars $\mu_0^{\text{opt}}, \lambda_1^{\text{opt}}, \ldots, \lambda_m^{\text{opt}}$ not all 0, such that $\mu_0^{\text{opt}} \geq 0$ and:

$$ \mu_0^{\text{opt}} \nabla f(x^{\text{opt}}) + \sum_{i=1}^m \lambda_i^{\text{opt}} \nabla h_i(x^{\text{opt}}) = 0 $$

If $x^{\text{opt}}$ is regular, then we can simply take $\mu_0^{\text{opt}} = 1$; if $x^{\text{opt}}$ is not regular, then $(\nabla h_i(x^{\text{opt}}))_{i \in [m]}$ are linearly dependent and we can take $\mu_0^{\text{opt}} = 0$. Necessary conditions that involve a nonnegative multiplier for the cost gradient are known as Fritz John necessary conditions. We provide one particular example now.

**Proposition 1.22** (Fritz John necessary conditions). Let $f, h, g$ be continuously differentiable functions and let $x^{\text{opt}}$ be a local minimum of:

$$ \min_x f(x) $$

subject to $h_i(x) = 0$ for $i = 1, \ldots, m$

$$ g_j(x) \leq 0 \text{ for } j = 1, \ldots, r $$

Then, there exists a scalar $\mu_0^{\text{opt}}$ and multipliers $\lambda_1^{\text{opt}}, \ldots, \lambda_m^{\text{opt}}$ and $\mu_1^{\text{opt}}, \ldots, \mu_r^{\text{opt}}$ satisfying:

$$ \mu_0^{\text{opt}} \nabla f(x^{\text{opt}}) + \sum_{i=1}^m \lambda_i^{\text{opt}} \nabla h_i(x^{\text{opt}}) + \sum_{j=1}^r \mu_j^{\text{opt}} \nabla g_j(x^{\text{opt}}) = 0 $$

$$ \mu^{\text{opt}} \geq 0 $$

Furthermore, $\mu_0^{\text{opt}}, \lambda_1^{\text{opt}}, \ldots, \lambda_m^{\text{opt}}, \mu_1^{\text{opt}}, \ldots, \mu_r^{\text{opt}}$ are not all 0, and in every neighborhood $U$ of $x^{\text{opt}}$ there is an $x \in U$ such that $\lambda_i^{\text{opt}} h_i(x) > 0$ for all $i$ with $\lambda_i^{\text{opt}} \neq 0$ and $\mu_j^{\text{opt}} g_j(x) > 0$ for all $j$ with $\mu_j^{\text{opt}} > 0$.

Note that continuity implies complementary slackness, i.e. $\mu_j^{\text{opt}} g_j(x^{\text{opt}}) = 0$ for all $j$. If $\mu_j^{\text{opt}} > 0$, then $g_j(x) > 0$ arbitrarily close to $x^{\text{opt}}$, so, taking a sequence to $x^{\text{opt}}$ and invoking continuity, we see that $g_j(x^{\text{opt}}) \geq 0$, which means $g_j(x^{\text{opt}}) = 0$.

**References**
