APM462 Notes

Jenci Wei

Summer 2023

Contents

1 Foundations 3

2 Finite Dimensional Unconstrained Optimization Problems 5

3 Convex Optimization Problems 6

4 Lagrange Multipliers 9

5 Kuhn-Tucker 11

6 Duality 13

7 Newton's Method 14

8 Method of Steepest Descent 15

9 Conjugate Directions Method 16

10 Augmented Lagrangian Method 17

11 Calculus of Variations 19

1 Foundations

A function $f : \mathbb{R} \to \mathbb{R}$ is in C^k if it has $k$derivatives and its kth derivative $f^{(k)}$ is continuous.

Theorem 1.1 (Mean Value Theorem). *Let $f : \mathbb{R} \to \mathbb{R}$ be a C^1 function. Then for $a, b \in \mathbb{R}$, the slope of the secant line from $(a, f(a))$ to $(b, f(b))$, given by $\frac{f(b)-f(a)}{b-a}$, is equal to the derivative of f at c, where $c = a + \theta(b-a)$ for some $\theta \in (0,1)$:*

$$\frac{f(b) - f(a)}{b - a} = f'(c) \iff f(x+h) = f(x) + hf'(x + \theta h)$$

where we let $b = x + h$ and $a = x$

Theorem 1.2 (1st order Taylor). *Let $g : \mathbb{R} \to \mathbb{R}$ be a C^1 function. Then for $x, h \in \mathbb{R}$,*

$$g(x+h) = g(x) + hg'(x) + o(h)$$

where "little o" notation indocates that the function $o(h)$ goes to zero faster than h, i.e.

$$\lim_{h \to 0} \frac{o(h)}{h} = 0$$

Theorem 1.3 (2nd order MVT). *Let $f : \mathbb{R} \to \mathbb{R}$ be a C^2 function. Then for $a, h \in \mathbb{R}$:*

$$f(a+h) = f(a) + hf'(a) + \frac{1}{2}h^2 f''(a + \theta h)$$

for some $\theta \in (0,1)$

Theorem 1.4 (2nd order Taylor). *Let $g : \mathbb{R} \to \mathbb{R}$ be a C^2 function. Then for $x, h \in \mathbb{R}$:*

$$g(x+h) = g(x) + hg'(x) + \frac{1}{2}h^2 g''(x) + o(h^2)$$

The **gradient** of a function $f : \mathbb{R}^n \to \mathbb{R}$ at $x = (x_1, \ldots, x_n)$, if it exists, is the (unique) vector $\nabla f(x) \in \mathbb{R}^n$ with the property that

$$\lim_{|v| \to 0} \frac{f(x+v) - f(x) - \nabla f(x) \cdot v}{|v|} = 0$$

It can be written as

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

- The gradient of f at x is the vector that points in the direction of greatest rate of change of f at x

The rate of change of f at x in the direction of unit vector v, also called the **directional derivative** of f in the direction v at x, is

$$D_v f(x) = f_v(x) \triangleq \left. \frac{d}{ds} \right|_{s=0} f(x + sv) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x) v_i = \nabla f(x) \cdot v = |\nabla f(x)| \cos\theta$$

where $\theta \in [0, \pi]$ is the angle between $\nabla f(x)$ and v, as derived from the law of cosines.

The **level sets** (or **level curves**) of a function $f : \mathbb{R}^n \to \mathbb{R}$ are defined as the sets

$$L_c = \{x \in \mathbb{R}^n \mid f(x) = c\}$$

for $c \in \mathbb{R}$

- Take a differentiable curve $r(t) = (x_1(t), \ldots, x_n(t)) \subset L_c$ for $t \in [0, \infty)$

- The derivative $\frac{dr}{dt} = \left(\frac{d}{dt}x_1(t), \ldots, \frac{d}{dt}x_n(t)\right)$ is a **tangent vector** on the surface $S = L_c$
- A tangent vector satisfies $\nabla f(x) \cdot \frac{dr}{dt} = 0$

Theorem 1.5 (1st order MVT in $\mathbb{R}^n$)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^1 function. Then for $x, v \in \mathbb{R}^n$:*

$$f(x+v) = f(x) + \nabla f(x + \theta v) \cdot v$$

for some $\theta \in (0, 1)$

Theorem 1.6 (1st order Taylor in $\mathbb{R}^n$)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^1 function. Then for $x, v \in \mathbb{R}^n$:*

$$f(x+v) = f(x) + \nabla f(x) \cdot v + o(|v|)$$

The **Hessian** of f at $x \in \mathbb{R}^n$, $\nabla^2 f(x)$, is the $n \times n$ matrix

$$\nabla^2 f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x)\right)\Bigg|_{1 \le i,j \le n}$$

which is a symmetric matrix, since $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$ by Clairaut's theorem.

Theorem 1.7 (2nd order MVT in $\mathbb{R}^n$)**.** *Let $f : \mathbb{R}^n \to R$ be a C^2 function. Then for $x, v \in \mathbb{R}^n$:*

$$f(x+v) = f(x) + \nabla f(x) \cdot v + \frac{1}{2} v^\top \nabla^2 f(x + \theta v) v$$

for some $\theta \in (0, 1)$

Theorem 1.8 (2nd order Taylor in $\mathbb{R}^n$)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^2 function. Then for $x, v \in \mathbb{R}^n$:*

$$f(x+v) = f(x) + \nabla f(x) \cdot v + \frac{1}{2} v^\top \nabla^2 f(x) v + o(|v|^2)$$

Theorem 1.9 (Implicit Function Theorem)**.** *Suppose that $F : \mathbb{R}^{n+k} \to \mathbb{R}^k$ is C^1 such that*

1. *At some fixed $(x_0, y_0) \in \mathbb{R}^{n+k}$, we have $F(x_0, y_0) = 0$*
2. *The Jacobian at (x_0, y_0) is invertible: $\det[D_y F(x_0, y_0)] \neq 0$*

Then there exists a small ball $B_r(x_0) \subset \mathbb{R}^n$ such that for

$$x \in B_r(x_0) \implies \exists! f(x) \in B_r(y_0)$$

and it satisfies the system

$$F(x, f(x)) = 0$$

2 Finite Dimensional Unconstrained Optimization Problems

We say f has a **local minimum** at $x_0 \in \Omega$ if it has a neighbourhood $B_\epsilon^\Omega(x_0)$ such that

$$f(x_0) \leq f(x) \quad \forall x \in B_\epsilon^\Omega(x_0)$$

where $\epsilon > 0$ and $B_\epsilon^\Omega(x_0) = \{x \in \Omega \mid |x - x_0| < \epsilon\}$ is the open ball in Ω of radius ϵ, centered at x_0

We say f has a **strict local minimum** at $x_0 \in \Omega$ if there is an $\epsilon > 0$ such that

$$f(x_0) < f(x) \quad \forall x \in B_\epsilon^\Omega(x_0) \setminus \{x_0\}$$

We say that f has a **global minimum** at $x_0 \in \Omega$ if

$$f(x_0) \leq f(x) \quad \forall x \in \Omega$$

A direction v is **feasible** at $x_0 \in \Omega$ if

$$x_0 = sv \in \Omega \quad \forall 0 \leq s \leq \overline{s}$$

for some $\overline{s} > 0$

Theorem 2.1 (1st order necessary conditions for a local minimum). *Let f be a C^1 function on $\Omega \subseteq \mathbb{R}^n$. If $x_0 \in \Omega$ is a local minimum of f, then*

$$\nabla f(x_0) \cdot v \geq 0$$

for every feasible direction v

- If we move away a little bit from a minimum point, the function can only get larger

Lemma 2.2 (Zero gradient). *if $x_0 \in \Omega^o$ is a local minimum of f, then*

$$\nabla f(x_0) = 0$$

Theorem 2.3 (2nd order necessary conditions for a local minimum). *Let f be a C^2 function on $\Omega \subseteq \mathbb{R}^n$. If $x_0 \in \Omega$ is a local minimum of f on Ω, then*

1. $\nabla f(x_0) \cdot v \geq 0$ *for all feasible directions v*
2. *If $\nabla f(x_0) \cdot v = 0$ for some feasible direction v, then $v^\top \nabla^2 f(x_0) v \geq 0$*

A symmetric square matrix A is **positive definite** if the following equjivalent conditions hold:

- $v^\top A v > 0$ for all $v \in \mathbb{R}^n$
- All eigenvalues of A are positive
- Determinants of all leading minors are positive (Sylvestor's criterion)

A symmetric square matrix is **positive semidefinite** if $v^\top A v \geq 0$ for all $v \in \mathbb{R}^n$ or equivalently if all its eigenvalues are nonnegative

- The analogue to Sylvester's criterion is no longer true (counterexample: zero matrix)

Theorem 2.4 (2nd order sufficient conditions for a local minimum at an interior point). *Let $f \in C^2$ be a function on $\Omega \subseteq \mathbb{R}^n$ and let $x_0 \in \Omega^o$. If the following conditions hold:*

1. $\nabla f(x) = 0$
2. $\nabla^2 f(x_0) > 0$*, i.e.* $\nabla^2 f(x_0)$ *is positive definite*

then x_0 is a strict local minimum point

3 Convex Optimization Problems

A set $\Omega \subseteq \mathbb{R}^n$ is **convex** if for every $x_1, x_2 \in \Omega$ and every $s \in [0, 1]$,

$$sx_1 + (1 - s)x_2 \in \Omega$$

- For any two points in Ω, the line segment coneecting them is contained in Ω
- The sum of convex sets $A + B$ is convex
- The intersection of convex sets $A \cap B$ is convex
- The scaled convex set $c\Omega$ for $c > 0$ is convex

Let $\Omega \subseteq \mathbb{R}^n$ be a convex set. A function $f : \Omega \to \mathbb{R}$ is **convex** if for every $x_1, x_2 \in \Omega$ and $s \in [0, 1]$,

$$f(sx_1 + (1 - s)x_2) \leq sf(x_1) + (1 - s)f(x_2)$$

- The secant line connecting $(x_1, f(x_1))$ and $(x_2, f(x_2))$ is above the graph of f
- The function is **strictly convex** if we have strict inequality
- The function is **concave** if we switch the inequalities

Lemma 3.1. *$f : \mathbb{R} \to \mathbb{R}$ is convex iff f' is increasing (i.e. $f'' \geq 0$)*

Convex Function Properties

- If f_is are convex, then $\sum a_i f_i(x)$ is convex for $a_i \geq 0$
- The maximum of convex functions $max(f_1(x), \ldots, f_n(x))$ is convex
- If f is convex and $c \in \mathbb{R}$, then the c-sublevel set of f, $\Gamma_c = \{x \in \Omega \mid f(x) \leq c\}$, is convex
  - Γ_c is the set of all preimage points
- If set Ω is convex and $f(x, y)$ is convex over Ω^2, then $\sup y \in \Omega f(x, y)$ is convex over Ω

Lemma 3.2 (C^1 characterization of convexity)**.** *Let $f : \Omega \to \mathbb{R}$ be a C^1 function, then the following are equivalent:*

1. *f is convex on Ω*
2. *$f(y) \geq f(x) + \nabla f(x) \cdot (y - x)$ for all $x, y \in \Omega$*

- Second point is saying that for each $x \in \Omega$, $f(y)$ lies above the tangent line/plane to f at x

Lemma 3.3 (C^2 characterization of convexity)**.** *Let $f : \Omega \to \mathbb{R}$ be a C^2 function, then the following are equivalent:*

1. *f is convex on Ω*
2. *$\nabla^2 f(x) \geq 0$ for all $x \in \Omega$ (i.e. $\nabla^2 f(x)$ is positive definite)*

Lemma 3.4 (Set of minimizers is convex)**.** *Let $\Omega \subseteq \mathbb{R}^n$ be a convex set. Suppose $f : \Omega \to \mathbb{R}$ is a convex function which has a minimum, i.e. the set of minimal points $\Gamma = \left\{x \in \Omega \mid f(x) = \min\limits_{\Omega} f\right\}$ is nonempty. Then*

1. *Γ is a convex set*
2. *Any local minimum is a global minimum*

A function $f : \Omega \to \mathbb{R}$ on convex $\Omega \subseteq \mathbb{R}^n$ is **locally convex** if for any $z \in \Omega$, there is some radius $r_z > 0$ such that f is convex on the restricted set $\Omega \cap B_{r_z}(z)$

Lemma 3.5. *Let $f \in C^1$ on $\Omega \subseteq \mathbb{R}^n$. Suppose $\exists x_0 \in \Omega^o$ such that*

1. *$\nabla f(x_0) = 0$*
2. *f is locally convex at x_0*

Then x_0 is a local minimum.

Theorem 3.6 (Optimality)**.** *Given a C^1 function $f : \Omega \to \mathbb{R}$ where $\Omega \subseteq \mathbb{R}^n$ is convex.*

1. *If x_0 is a local minimum, then $\nabla f(x_0) \cdot (x - x_0) \geq 0$ for all $x \in \Omega$*
2. *If f is also convex, then x_0 is a global minimum*

Lemma 3.7. *Let $f : \Omega \to \mathbb{R}$ be a convex function on a convex, compact set $\Omega \subset \mathbb{R}^n$. Then the maximum of f is (also) attained on the boundary of Ω, i.e.*

$$\max_{\Omega} f = \max_{\partial\Omega} f$$

Theorem 3.8 (Young Inequality)**.** *Let $a, b \in \mathbb{R}$, then*

$$|ab| \leq \frac{1}{p}|a|^p + \frac{1}{q}|b|^q$$

for any $0 < p, q$ such that $1/p + 1/q = 1$

Convex Optimization Subject to Bounds

- Suppose we want to minimize a C^1 function $f : \Omega \to \mathbb{R}$ where the domain is

  $$\Omega = \{x \in \mathbb{R}^n : a_i \leq x_i \leq b_i, i = 1, \dots, n\}$$

  i.e. a hyperrectangle
- If a point $x^* \in \partial\Omega$ is a local minimum, then

  $$\begin{cases} \frac{\partial f(x^*)}{\partial x_i} \geq 0, & \text{if } x_i^* = a_i \\ \\ \frac{\partial f(x^*)}{\partial x_i} \leq 0, & \text{if } x_i^* = b_i \\ \\ \frac{\partial f(x^*)}{\partial x_i} = 0, & \text{if } a_i < x_i^* < b_i \end{cases}$$

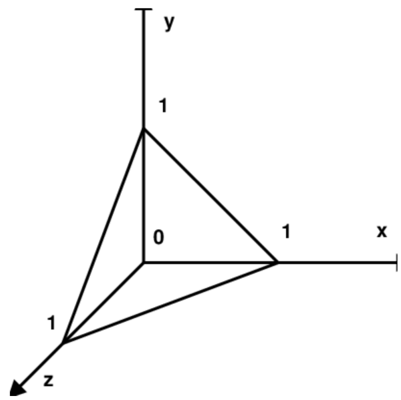
- If a point $x^* \in \Omega^o$ is a local minimum, then $\nabla f(x^*) = 0$
- If f is furthermore convex, then the above are necessary too (i.e. replace *then* with *iff*)

Convex Optimization Over a Simplex

- Suppose we want to minimize a C^1 function $f : \Omega \to \mathbb{R}$ where the domain is

  $$\Omega = \left\{ x \in \mathbb{R}^n \mid x_i \geq 0, \sum_i x_i = r \right\}$$

  for some $r > 0$



- Notice that $\Omega = \partial\Omega$, so it must be the case that $x^* \in \partial\Omega$
- If a point $x^* \in \Omega$ is a local minimum, then

$$x_i > 0 \implies \frac{\partial f(x^*)}{\partial x_i} \leq \frac{\partial f(x^*)}{\partial x_j} \quad \forall j \neq i$$

- If f is furthermore convex, then the above is necessary too (i.e. replace *then* with *iff*)

Theorem 3.9 (Extreme Value Theorem)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous, and $K \subseteq \mathbb{R}^n$ compact. Then the problem $\min\limits_{x \in K} f(x)$ has a solution, i.e.*

$$\exists x_0 \in K \text{ such that } f(x_0) = \min_{x \in K} f(x)$$

Theorem 3.10 (Local Extreme Value Theorem)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous. Suppose $\exists a \in \mathbb{R}^n$ such that $f(x) \geq f(a)$ whenever $|x - a| > R > 0$, then*

$$\min_{x \in \mathbb{R}^n} f(x) = \min_{x \in B_R(a)} f(x)$$

- If we have a ball, where all values outside the ball are greater than one value within the ball
- Then the minimum of the function is attained within the ball

4 Lagrange Multipliers

A **surface** is the zero set of a collection of functions $h_i : \mathbb{R}^n \to \mathbb{R}$, namely

$$M = \{x \in \mathbb{R}^n \mid h_1(x) = \cdots = h_k(x) = 0\}$$

- The functions are required to be C^1, C^2, etc. as necessary

A C^k**-differentiable curve** on a surface $M \subset \mathbb{R}^n$ is a C^k function $x : (a,b) \to \mathbb{R}^n$ such that $x(s) \in M$

- Let $x(s)$ be a differentiable curve on M that passes through the point $x_0 \in M$
- WLOG let $x(0) = x_0$

The **tangent vectors** v to M at x_0 are vectors which are velocity vectors of curves $x(s)$ on M through x_0:

$$v = \frac{d}{ds}\bigg|_{s=0} x(s)$$

- The vector v is "generated" by the curve $x(s)$

The set of all tangent vectors to M at x_0 is the **tangent space** to M at x_0, and is denoted by $T_{x_0}M$:

$$T_{x_0}M = \left\{ v \in \mathbb{R}^n \mid v = \frac{d}{ds}\bigg|_{s=0} x(s) \text{ for some differentiable curve } x(s) \in M \text{ such that } x(0) = x_0 \right\}$$

Given a surface $M = \{x \in \mathbb{R}^n \mid h_1(x) = \cdots = h_k(x) = 0\}$, a point $x_0 \in M$ is a **regular point** of M (or of the constraints h_i) if $\nabla h_1(x_0), \ldots, \nabla h_k(x_0)$ are linearly independent

- When there is only 1 constraint h, x_0 is regular if $\nabla h(x_0) \neq 0$

Lemma 4.1 (Tangent space at regular point)**.** *Let x_0 be a point on the surface*

$$M = \{x \in \mathbb{R}^n \mid h_1(x) = \cdots = h_k(x) = 0\}$$

If the gradient vectors $\{\nabla h_1(x_0), \ldots, \nabla h_k(x_0)\}$ are linearly independent, then

$$T_{x_0}M = T_{x_0} \triangleq \{v \in \mathbb{R}^n \mid v \cdot \nabla h_1(x_0) = \cdots = v \cdot \nabla h_k(x_0) = 0\}$$

Furthermore

$$T_{x_0} = \operatorname{span}\{\nabla h_1(x_0), \ldots, \nabla h_k(x_0)\}^{\perp}$$

- Otherwise $T_{x_0}M \subseteq T_{x_0}$

Lemma 4.2 (Gradient normal to tangent)**.** *Let $f, h_1, \ldots, h_k$ be C^1 functions defined on an open set $\Omega \subset \mathbb{R}^n$. Suppose x_0 is a local minimum of f on $M = \{x \in \Omega \mid h_1(x) = \cdots = h_k(x) = 0\}$. Then $\nabla f(x_0) \perp T_{x_0}M$, i.e. $\nabla f(x_0) \cdot v = 0$ for all $v \in T_{x_0}M$*

Theorem 4.3 (First order necessary conditions for min: equality constraints)**.** *Let $f, h_1, \ldots, h_k$ be C^1 functions on the open set $\Omega \subset \mathbb{R}^n$ to $\mathbb{R}$. Suppose $x_0 \in M = \{h_1(x) = \cdots = h_k(x) = 0\}$ is a regular point which is a local minimum of f on M. Then there exists $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$, called Lagrange multipliers, such that*

$$\nabla f(x_0) + \lambda_1 \nabla h_1(x_0) + \cdots + \lambda_k \nabla h_k(x_0) = 0$$

Theorem 4.4 (Second order necessary conditions for min: equality constraints)**.** *Let $f, h_1, \ldots, h_k$ be C^2 functions on an open set $\Omega \subset \mathbb{R}^n$. Let $x_0 \in M = \{x \in \Omega \mid h_1(x) = \cdots = h_k(x) = 0\}$ be a regular point of the constraints h_i. Suppose x_0 is a local min of f on M. Then there exists Lagrange multipliers $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$, such that*

1. *The first order condition holds, i.e.*

$$\nabla f(x_0) + \lambda_1 \nabla h_1(x_0) + \cdots + \lambda_k \nabla h_k(x_0) = 0$$

2. *The following second order condition hold:*

$$\nabla^2 f(x_0) + \lambda_1 \nabla^2 h_1(x_0) + \cdots + \lambda_k \nabla^2 h_k(x_0)$$

   is positive semidefinite on the tangent space $T_{x_0}M$

- Does *not* require positive semidefiniteness on entire space $\mathbb{R}^n$

Theorem 4.5 (Second order sufficient conditions for min: equality constraints)**.** *Given $f, h_i \in C^2$ on an open subset $\Omega \subset \mathbb{R}^n$, suppose there exist $x_0 \in M = \{x \mid h_i(x) = 0\}$ and $\lambda \in \mathbb{R}^n$ such that*

1. $\nabla f(x_0) + \sum_i \lambda_i \nabla h_i(x_0) = 0$ *(Lagrange muyltipliers)*
2. $\nabla^2 f(x_0) + \sum_i \lambda_i \nabla^2 h_i(x_0) > 0$ *on* $T_{x_0}M$ *(positive definite)*

Then x_0 is a strict local minimum of f on M

- x_0 is *not* required to be regular
- Knowing $T_{x_0}M$ may be hard if x_0 is not regular, but we can easily obtain its superset T_{x_0} using our formula

5 Kuhn-Tucker

Let $\Omega \subseteq \mathbb{R}^n$ be an open set, and $f, h_1, \dots, h_k, g_1, \dots, g_l : \Omega \to \mathbb{R}$

- Suppose we want to minimize $f(x)$ on the feasible set defined by

$$S = \{x \in \Omega \mid h_1(x) = \cdots = h_k(x) = 0, g_1(x) \leq 0, \dots, g_l(x) \leq 0\}$$

- h_i give equality constraints and g_j give inequality constraints
- Suppose x_0 satisfies all the constraints. An inequality constraint $g_j(x) \leq 0$ is **active** at x_0 if $g_j(x_0) = 0$. Otherwise, it is **inactive**
- All the equality constraints are active at x_0
- Order the inequality constraints so the first $l' \leq l$ are active and the remaining inactive at x_0. Then x_0 is a **regular point** of the constraints if the gradients of the active constraints

$$\nabla h_1(x_0), \dots, \nabla h_k(x_0), \nabla g_1(x_0), \dots, \nabla g_{l'}(x_0)$$

  are linearly independent

Theorem 5.1 (First order necessary conditions for min: inequality constraints)**.** *Let Ω be an open subset of $\mathbb{R}^n$, and let $f, h_i, g_j \in C^1(\Omega)$. Suppose $x_0 \in \Omega$ is a local minimum of f and a regular point of the constraints. Then there exists $\lambda_i \in \mathbb{R}, \mu_j \geq 0$ such that*

1. $\nabla f(x_0) + \sum_{i=1}^{k} \lambda_i \nabla h_i(x_0) + \sum_{j=1}^{l} \mu_j \nabla g_j(x_0) = 0$ *(Lagrange multiplers)*
2. $\mu_j g_j(x_0) = 0$ *for all* $1 \leq j \leq l$ *(complementary slackness), or equivalently*

$$\sum_{j=1}^{l} \mu_j g_j(x_0) = 0$$

- Because $\mu_j \geq 0$, (2) implies that the μ_j for inactive g_j must be zero

Theorem 5.2 (Second order necessary conditions for min: ineqliaty constraints)**.** *Let Ω be an open subset of $\mathbb{R}^n$, and let $f, h_i, g_j \in C^2(\Omega)$. Let x_0 be a local minimum of f subject to the constraints $h_i(x) = 0, g_j(x) \leq 0$, and a regular point on those constraints. Then there exists $\lambda_i \in \mathbb{R}$ and $\mu_j \geq 0$ such that*

1. *The first order constraints are satisfied:*

$$\nabla f(x_0) + \sum_{i=1}^{k} \lambda_i \nabla h_i(x_0) + \sum_{j=1}^{l} \mu_j \nabla g_j(x_0) = 0$$

2. $\mu_j g_j(x_0) = 0$ *for all* $1 \leq j \leq l$
3. $L(x_0) \triangleq \nabla^2 f(x_0) + \sum_{i=1}^{k} \lambda_i \nabla^2 h_i(x_0) + \sum_{j=1}^{l} \mu_j \nabla^2 g_j(x_0)$ *is positive semidefinite on the tangent space* $T_{x_0}\tilde{M}$ *of the active constraints*

Theorem 5.3 (Second order sufficient conditions for min: inequality constraints)**.** *Let $\Omega \subseteq \mathbb{R}^n$ be an open set, and let $f, h_i, g_j \in C^2(\Omega)$ be functions. Suppose there is a point $x_0 \in \Omega$ (not necessarily regular) satisfying the constraints $h_i(x) = 0$ for $i = 1, \dots, k$ and $g_j(x) \leq 0$ for $j = 1, \dots, l$, and $\lambda_i \in \mathbb{R}$, $\mu_j \geq 0$ such that*

1. $\nabla f(x_0) + \sum_{i=1}^{k} \lambda_i \nabla h_i(x_0) + \sum_{j=1}^{l} \mu_j \nabla g_j(x_0) = 0$ *(Lagrange multiplers)*

2. *$\mu_j g_j(x_0) = 0$ for all j (complementary slackness)*
3. *The Lagrangian matrix*

$$L(x_0) = \nabla^2 f(x_0) + \sum_{i=1}^{k} \lambda_i \nabla^2 h_i(x_0) + \sum_{j=1}^{l} \mu_j \nabla^2 g_j(x_0)$$

is positive definite on the vector space $\tilde{\tilde{T}}_{x_0}$ to the strongly active constraints *at x_0 (i.e. $\mu_j > 0$)*

Then x_0 is a strict local minimum of f subject to the constraints.

- A constraint g_j is **strongly active** at x_0 if $\mu_j > 0$
- Arranging the constraints so that the first l'' are the strongly active ones, we have

$$\tilde{\tilde{T}}_{x_0} = \{v \in \mathbb{R}^n \mid \nabla h_i(x_0) \cdot v = 0 \text{ for } i = 1, \ldots, k, \nabla g_j(x_0) \cdot v = 0 \text{ for } j = 1, \ldots, l''\}$$

KKT Conditions

1. Necessary: finding candidates
   - If x_* is a local minimum and a regular point, then there exist λ, μ such that
     (a) $\nabla_x L(x_*, \lambda, \mu) = 0$
     (b) $\mu_j \geq 0$ for all j and $\mu_j = 0$ for the inactive constraints, i.e. $g_j(x_*) < 0$
2. Sufficient: verifying candidates
   - Suppose that there exist λ, μ such that
     (a) x_* is a regular point, i.e. $\nabla h_i(x_*), \nabla g_i(x_*)$ are linearly independent
     (b) $\nabla_x L(x_*, \lambda, \mu) = 0$
     (c) $\mu_j > 0$ for the active constraints, i.e. $g_j(x_*) = 0$
     (d) $\mu_j = 0$ for the inactive constraints, i.e. $g_j(x_*) < 0$
     (e) $\nabla_x^2 L(x_*, \lambda, \mu)$ is positive definite on the tangent space to the active constraints
   - Then x_* is a local minimum

6 Duality

For the minimization problem

$$\min f(x) \text{ with } \{x \in D \subseteq \mathbb{R}^n \mid h_1(x) = \cdots = h_k(x) = 0\}$$

We consider the **Lagrange dual function** $g(\lambda) : \mathbb{R}^k \to \mathbb{R}$

$$g(\lambda) = \inf_{x \in D} f(x) + \sum_{i=1}^{k} \lambda_i h_i(x)$$

- This function is *concave*

Theorem 6.1 (Weak duality)**.** *Suppose that $p_* = f(x_*)$ (i.e. the minimum value), then*

$$g(\lambda) \leq p_*$$

for all $\lambda \in \mathbb{R}^k$

- When $d_* \triangleq \max_{\lambda} g(\lambda) = p_*$, then we would have **strong duality**
- The difference $p_* - d_* \geq 0$ is the **duality gap**

Theorem 6.2 (KKT implies strong duality)**.** *If we have the KKT conditions (last page), then*

$$\min_{x \in constraints} f(x) = \max_{\lambda, \mu \in \mathbb{R}^{k_1 + k_2}} \inf_{x \in \mathbb{R}^n} L(x, \lambda, \mu)$$

7 Newton's Method

Newton's Method in 1D

- Given a function $f : \mathbb{R} \to \mathbb{R}$ which is unimodal (i.e. one local min) on some interval $I \subseteq \mathbb{R}$, we want to find its min
- Given a starting point $x_0 \in I$, consider the quadratic approximation

$$q(x) \triangleq f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2$$

- We search for the minimum of q, which gives

$$x_1 = x_0 - \frac{f'(x_0)}{f''(x_0)}$$

Algorithm (1D)

1. Pick some starting point $x_0 \in I$
2. Let

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)} \triangleq x_n - \frac{g(x_n)}{g'(x_n)}$$

3. Repeat until convergence

Order of Convergence of a Sequence

- Let $\{x_n\}$ converge to x_*, with $x_i \neq x_*$
- The order of convergence of $\{x_n\}$ is the largest p such that

$$\lim_{n \to \infty} \frac{|x_{n+1} - x_*|}{|x_n - x_*|^p} < \infty$$

Newton's Method Convergence

- Let $g \in C^2(I)$, or equivalently $f \in C^3(I)$
- Suppose $x_* \in I$ satisfies $g(x_*) = 0$ and $g'(x_*) \neq 0$
- If x_0 is sufficiently close to x_*, then the sequence $\{x_n\}$ generated by Newton's method converges to x_*
- The order of convergence is at least 2
- May converge to any critical point (i.e. including a saddle point or local max)

Newton's Method in $\mathbb{R}^n$

- Let $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ be a C^3 function on an open set Ω
- Let x_* be a local minimum of f, with $\nabla^2 f(x_*) > 0$ (i.e. second order sufficient condition for x_* to be a local min)
- Then, given x_0 sufficiently close to x_*, the Newton's method algorithm defined by

$$x_{n+1} = x_n - \left[\nabla^2 f(x_n)\right]^{-1} \nabla f(x_n)$$

  converges to x_* with order at least 2

8 Method of Steepest Descent

Introduction

- Update our guess from some starting point x_0 with

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

- Define α_k to be the number $\alpha_k > 0$ such that

$$f(x_k - \alpha_k \nabla f(x_k)) = \min_{s>0} f(x_k - s\nabla f(x_k))$$

Convergence

- Let $f \in C^1(\Omega)$ and $x_0 \in \Omega$
- Let $\{x_k\}$ be generated by the method of steepest descent, i.e.

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

- $\{x_k\}$ is bounded in Ω, i.e. there exists a compact set $K \subset \Omega$ such that $x_k \in K$ for all $k \in \mathbb{N}$
- Then every convergent subsequence of $\{x_k\}$ converges to a citical point $x_* \in K$ of f, i.e. $\nabla f(x_*) = 0$

Quadratic Case for Gradient Descent

- Take the method of steepest descent for a quadratic function f, i.e.

$$f(x) = \frac{1}{2}x^\top Qx - b^\top x$$

  where $b, x \in \mathbb{R}^n$
- et $\lambda = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = \Lambda$ be the ordered eigenvalues of Q
  - Notice that $\lambda \geq 0$ since Q is PD
- Since f is strictly convex and Q is PD, there exists a unique global maximum point $x_* = Q^{-1}b$
- Let

$$q(x) = \frac{1}{2}(x - x_*)^\top Q(x - x_*) \geq 0$$

- The steepest descent algorithm becomes

$$\begin{aligned} x_{k+1} &= x_k - \alpha_k g(x_k) \\ &= x_k - \frac{\|g(x_k)\|^2}{g(x_k)^\top Q g(x_k)} g(x_k) \end{aligned}$$

- By the Kantorovich inequality

$$q(x_{k+1}) \leq \left(\frac{\Lambda - \lambda}{\Lambda + \lambda}\right)^2 q(x_k)$$

- Order of convergence is at least 1
- Rate of convergence depends on the ratio Λ/λ
  - When $\Lambda \approx \lambda$, gradient descent is fast

9 Conjugate Directions Method

Introduction

- We study quadratic problems of the form

$$f(x) = \frac{1}{2}x^\top Qx - b^\top x$$

- For other functions, we can get a similar result by iteratively approximating f quadratically (i.e. by the second order Taylor expandsion)

Q-Orthogonality

- Two vectors d, d' are called **Q-orthogonal** or **Q-conjugate** if

$$d^\top Qd' = 0$$

- A finite set of vectors $d_0, d_1, \ldots, d_k$ are called a **Q-orthogonal set** if

$$\forall i \neq j, d_i^\top Qd_j = 0$$

- If $Q = I_{n\times n}$, then Q-orthogonality is the same as the usual orthogonality
- Let d and d' be two eigenvectors of Q with distinct eigenvalues. Then d and d' are Q-orthogonal
- For a $n \times n$ symmetric matrix Q, there is a set of n vectors which is Q-orthogonal

Conjugate Directions Method

- If Q is PD and symmetric, and $d_0, d_1, \ldots, d_k$ is a set of nonzero Q-orthogonal vectors, then $d_0, d_1, \ldots, d_k$ is linearly independent
- Then the n nonzero Q-orthogonal vectors, $d_0, \ldots, d_{n-1}$ form a basis of $\mathbb{R}^n$
- We can then write our minimum as

$$x_* = \alpha_0 d_0 + \alpha_1 d_1 + \cdots + \alpha_{n-1}d_{n-1}$$

- Solving for the α_i, we can take the dot product with d_i to get

$$d_i^\top Qx_* = d_i^\top Q(\alpha_0 d_0 + \alpha_1 d_1 + \cdots + \alpha_{n-1}d_{n-1}) = \alpha_i d_i^\top Qd_i$$

- Therefore

$$\alpha_i = \frac{d_i^\top Qx_*}{d_i^\top Qd_i}$$

Algorithm

- Consider the sequence $\{x_k\}$ generated by

$$x_{k+1} = x_k + \alpha_k d_k$$

where

$$\alpha_k = -\frac{g_k^\top d_k}{d_k^\top Qd_k}$$

with

$$g_k = Qx_k - b$$

- Then $x_n = x_*$, i.e. we converge to the minimum in n steps
- Bounds

$$q(x_{k+1}) \leq \max_{\lambda\in\text{eigenvalues of } Q}[1 + \lambda P_k(\lambda)]^2 q(x_0)$$

where P_k is any polynomial of degree k

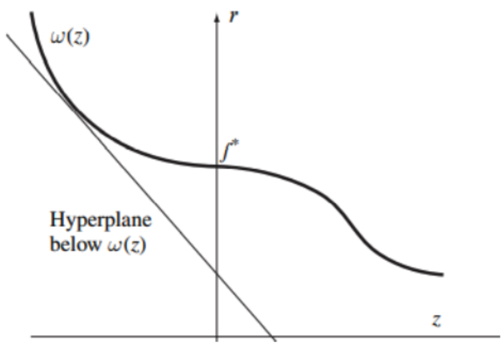
10 Augmented Lagrangian Method

Geometric Viewpoint of Duality

- For minimization problem
$$f_* = \min_x f(x) \quad \text{with} \quad g_i(x) \leq 0$$
the dual problem is
$$\phi_* \triangleq \max_\mu \phi(\mu) = \max_\mu \inf_x \left\{ f(x) + \mu^\top g(x_*) \right\}$$
- Weak duality $\phi_* \leq f_*$ is always true
- Strong duality $\phi_* = f_*$ follows by the KKT conditions

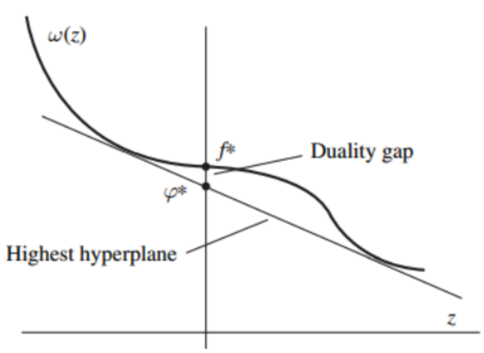
Primal Function
$$\omega(z) \triangleq \inf_{x \in \Omega} \{ f(x) : g_i(x) \leq z_i \} \quad \text{where} \quad z \in \mathbb{R}^p$$

- When $z = 0$, then we have the *original minimization problem*
- For single constraint $p = 1$, the intercept with the vertical axis is f_*



Hyperplane

- For each multiplier μ, consider the hyperplane $r + \mu^\top z = c$
- Suppose that x_* minimizes
$$\phi(\mu) = \inf_x \left\{ f(x) + \mu^\top g \right\} = f(x_*) + \mu^\top g(x_*)$$
- At maximizer μ_*, the hyperplane $r + \mu_*^\top z = \phi(\mu_*) = \phi_*$ for $r = f(x_*), z = g(x_*)$ hits the vertical axis at ϕ_*



Steepest Ascent

- Dual problem is
$$\phi_* \triangleq \max_\lambda \phi(\lambda) = \max_\lambda \inf_x \left\{f(x) + \lambda^\top g(x)\right\}$$
- Applying steepest ascent:
$$\lambda_{k+1} = \lambda_k + a_k \nabla_\lambda \phi(\lambda_k)$$
where a_k is the most ideal step size
$$a_k = \arg\max_{s \geq 0} \{\lambda_k + s\nabla_\lambda \phi(\lambda_k)\}$$

Augmented Lagrangian

- Start with a primal problem
$$\min f(x) \quad \text{with} \quad h(x) = 0$$
- The **Augmented Lagrangian** is
$$l_c(x, \lambda) \triangleq f(x) + \lambda^\top h(x) + \frac{c}{2}|h(x)|^2$$
for $c > 0$
- As $c \to \infty$, the cost function becomes more *convex-like*
- Assume that the second-order sufficiency conditions for a local minimum are satisfied at x_*, λ_*. Then there is a c_* such that for all $c \geq c_*$, the augmented Lagrangian $l_c(x, \lambda_*)$ has a local minimum point at x_*
- Let A and B be $n \times n$ symmetric matrices. Suppose that A is PD and B is PSD on the subspace $Bx = 0$. Then there is a c_* such that for all $c \geq c_*$, the matrix $A + cB$ is PD

ALM (Version 1) Algorithm

1. Starting with λ_k, we find x_{k+1} by minimizing
$$x_{k+1} = \arg\min_x f(x) + \lambda_k^\top h(x) + \frac{\alpha_k}{2}|h(x)|^2$$
where α_k is some sequence of our choice, e.g. $\alpha_k = c^k$
2. Update the multiplier
$$\lambda_{k+1} = \lambda_k + \alpha_k h(x_{k+1})$$

Primal Function and Augmented Lagrangian

$$\begin{aligned}\min_x l_c(x, \lambda_k) &= \min_x \left(f(x) + \lambda_k^\top h(x) + \frac{c}{2}|h(x)|^2\right) \\ &= \min_{x,y} \left\{f(x) + \lambda_k^\top y + \frac{c}{2}|y|^2 : y = h(x)\right\} \\ &= \min_y \left\{\omega(y) + \lambda_k^\top y + \frac{c}{2}|y|^2\right\}\end{aligned}$$

- If x_k is the minimizer, then $y_k \triangleq h(x_k)$ is the minimizer for
$$F(y, \lambda) \triangleq \omega(y) + \lambda^\top y + \frac{c}{2}|y|^2$$

ALM (Version 2) Algorithm

1. Find y_k from $\nabla_y F(y_k, \lambda_k) = 0$
2. Evaluate slope $\lambda_{k+1} = -\nabla_y \omega(y_k)$
3. Find y_{k+1} as in (1)

11 Calculus of Variations

A **functional** is a function $F : \mathcal{A} \to \mathbb{R}$ which maps a function $u(\cdot) \in \mathcal{A}$ to a real number $F[u(\cdot)]$

A **test function** $v(\cdot)$ is a C^1 function on $[a, b]$ such that $v(a) = v(b) = 0$

- Similar to the notion of "feasible direction"

Lemma 11.1 (Fundamental Lemma of the Calculus of Variations)**.** *Suppose g is continuous on $[a, b]$. If*

$$\int_a^b g(x)v(x)dx = 0 \qquad \forall v \in C^1, v(a) = v(b) = 0$$

then $g(x) = 0$ for all $x \in [a, b]$

General Class of Problems in the Calculus of Variations

- Let $\mathcal{A} = \{u : [a, b] \to \mathbb{R} \mid u \in C^1, u(a) = A, u(b) = B\}$
- Consider the minimization problem

$$\begin{aligned} &\text{Minimize } F[u(\cdot)] = \int_a^b \mathcal{L}(x, u(x), u'(x))dx \\ &\text{subject to } u(\cdot) \in \mathcal{A} \end{aligned}$$

- $\mathcal{L}(x, z, p) : [a, b] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is the **Lagrangian function**
- Denote

$$\mathcal{L}_z(x, z, p) = \frac{\partial \mathcal{L}}{\partial z}(x, z, p)$$
$$\mathcal{L}_p(x, z, p) = \frac{\partial \mathcal{L}}{\partial p}(x, z, p)$$

Variational Derivative

- Given a function $u(\cdot) \in \mathcal{A}$, suppose there exists a function $g(\cdot) : [a, b] \to \mathbb{R}$ such that

$$\left.\frac{d}{ds}\right|_{s=0} F[u(\cdot) + sv(\cdot)] = \int_a^b g(x)v(x)dx$$

  for all test functions v, then the function $g(\cdot)$ is the **variational derivative** of F at $u(\cdot)$, and is denoted by

$$\frac{\delta F}{\delta u}(u)(\cdot) \quad \text{or} \quad \frac{\delta F}{\delta u}(u) \quad \text{or} \quad \frac{\delta F}{\delta u}$$

First Order Condition

- Let $\mathcal{A} = \{u : [a, b] \to \mathbb{R} \mid u \in C^1, u(a) = A, u(b) = B\}$ and suppose $u_*(\cdot) \in \mathcal{A}$ minimizes F over $\mathcal{A}$, and $u_*(\cdot) + v(\cdot) \in \mathcal{A}$ for all test functions v.
- If $\frac{\delta f}{\delta u}(u_*)(\cdot)$ exists and is continuous, then it must equal to the zero function on $[a, b]$:

$$\frac{\delta f}{\delta u}(u_*)(\cdot) \equiv 0$$

Euler-Lagrange Equation

- Let $\mathcal{A} = \{u : [a,b] \to \mathbb{R} \mid u \in C^1, u(a) = A, u(b) = B\}$ and $F[u(\cdot)] = \int_a^b \mathcal{L}(x, u(x), u'(x))dx$ for some Lagrangian function $\mathcal{L}(x, z, p) \in C^2$. Suppose $u \in C^1$ on $[a, b]$. Then

$$\frac{\delta F}{\delta u}(u)(\cdot)$$

exists, is continuous, and is given by the equation

$$\frac{\delta F}{\delta u}(u)(x) - \frac{d}{dx}\mathcal{L}_p(x, u(x), u'(x)) + \mathcal{L}_z(x, u(x), u'(x))$$

This equation is called the **Euler-Lagrange equation** of F

Functionals of Vector-Valued Functions

- Want to minimize F whose input is $u : [a,b] \to \mathbb{R}^N$
- Then $\mathcal{L}$ is a function $[a,b] \times \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$
- The partial derivatives $\mathcal{L}_{\mathbf{z}}$ and $\mathcal{L}_{\mathbf{p}}$ are vectors
- The resulting Euler-Lagrangian equation is a vector equation, or equivalently a system of N equations:

$$-\frac{d}{dx}\mathcal{L}_{\mathbf{p}}(x, u(x), u'(x)) + \mathcal{L}_{\mathbf{z}}(x, u(x), u'(x)) = 0$$

Isoperimetric Constraints

- Suppose we are given two functionals

$$F[u(\cdot)] = \int_a^b \mathcal{L}^F(x, u(x), u'(x))dx$$
$$G[u(\cdot)] = \int_a^b \mathcal{L}^G(x, u(x), u'(x))dx$$

where $\mathcal{L}^F, \mathcal{L}^G \in C^2$ are Lagrangian functions (one for F and one for G)

- Consider the problem (called an isometric problem):

$$\begin{aligned} \text{minimize } & F[u(\cdot)] \\ & u \in \mathcal{A} \\ \text{subject to } & G[u(\cdot)] = c \end{aligned}$$

where $\mathcal{A} = \{u : [a,b] \to \mathbb{R} \mid u \in C^1, u(a) = A, u(b) = B\}$

- Suppose $u_*(\cdot)$ is a "regular point of G", i.e. $\frac{\delta G}{\delta u}(u_*) \neq 0$, which is a minimizer
- The first order necessary conditions say there exists a $\lambda \in \mathbb{R}$ such that

$$\frac{\delta F}{\delta u}(u_*) + \lambda\frac{\delta G}{\delta u}(u_*) = 0$$

or equivalently that $u_*(\cdot)$ is a minimizer of $F + \lambda G$ for some $\lambda \in \mathbb{R}$

- This gives

$$-\frac{d}{dx}\left(\mathcal{L}^F + \lambda\mathcal{L}^G\right)_p + \left(\mathcal{L}^F + \lambda\mathcal{L}^G\right)_z = 0$$

Holonomic Constraints

- Suppose we are given a functional

$$F[x(\cdot), y(\cdot), z(\cdot)] = \int_a^b \mathcal{L}(t, x(t), y(t), z(t), x'(t), y'(t), z'(t))dt$$

where the Lagrangian $\mathcal{L}(t, z_1, z_2, z_3, p_1, p_2, p_3) : [a, b] \times \mathbb{R}^6 \to \mathbb{R}$ is C^2

- Consider the following optimization problem, called a holonomic problem:

$$\begin{aligned} \text{minimize } & F[x(\cdot), y(\cdot), z(\cdot)] \\ \text{subject to } & (x(\cdot), y(\cdot), z(\cdot)) \in \mathcal{A} \\ & H(x(\cdot), y(\cdot), z(\cdot)) \equiv 0 \end{aligned}$$

where $\mathcal{A} = \{(x, y, z) : [a, b] \to \mathbb{R}^3 \in C^1 \mid x(a) = A_1, y(a) = A_2, z(a) = A_3, x(b) = B_1, y(b) = B_2, z(b) = B_3\}$, and $H(x, y, z) : R^3 \to \mathbb{R}$ is a function (not a functional)

- The curve $(x(t), y(t), z(t))$ lies on the surface $H(x, y, z) = 0$ in $\mathbb{R}^3$
- Suppose $(x_*(\cdot), y_*(\cdot), z_*(\cdot))$ is a "regular point of H", i.e. $\nabla H(x_*(t), y_*(t), z_*(t)) \neq 0$ for all $t \in [a, b]$, which is a minimizer
- The first order necessary conditions say there exists a function $\lambda : [a, b] \to \mathbb{R}$ such that the following Euler-Lagrange equations hold at $(x_*(\cdot), y_*(\cdot), z_*(\cdot))$, for $t \in [a, b]$:

$$\begin{bmatrix} -\frac{d}{dt}\mathcal{L}_{p_1} + \mathcal{L}_{z_1} \\ \\ -\frac{d}{dt}\mathcal{L}_{p_2} + \mathcal{L}_{z_2} \\ \\ -\frac{d}{dt}\mathcal{L}_{p_3} + \mathcal{L}_{z_3} \end{bmatrix} + \lambda(t) \begin{bmatrix} H_x \\ \\ H_y \\ \\ H_z \end{bmatrix} = 0$$

- Notice that λ depends on t, because we apply Lagrange multipliers at each point t

Isoperimetric Constraints (Multiple Constriaints)

- Objective:

$$\begin{aligned} \text{minimize } & F[u(\cdot)] \\ & u \in \mathcal{A} \\ \text{subject to } & G_i[u(\cdot)] = c_i \quad i = 1, \dots, n \end{aligned}$$

where $\mathcal{A} = \{u : [a, b] \to \mathbb{R} \mid u \in C^1, u(a) = A, u(b) = B\}$

- Let

$$\mathcal{L} = \mathcal{L}^F + \sum_{i=1}^{n} \lambda_i \mathcal{L}_i^G$$

- Then the Euler-Lagrange equation is

$$-\frac{d}{dx}\mathcal{L}_p + \mathcal{L}_z = 0$$

Second Variation

- To find minimizer, we ask for

$$D^2F(u_*, v) \triangleq \frac{d^2}{d\epsilon^2}F(u_* + \epsilon v)\bigg|_{\epsilon=0} > 0$$

- In the case $F[u(\cdot)] = \int_a^b \mathcal{L}(x, u(x), u'(x))dx$, we have

$$D^2F(u_*, v) = \int_a^b v^2\mathcal{L}_{zz} + 2vv'\mathcal{L}_{zp} + (v')^2\mathcal{L}_{pp}dx$$

Second Variation Sufficiency

- Suppose y is a stationary/critical point of F, i.e. the variational derivative is zero
- Suppose also that there exists a constant $c > 0$ such that

$$D^2F(y, v) \geq c\int_0^1 (v')^2dx$$

- Then y is a local minimizer