Lecture V: Constrained Optimization, Integration

1. Constrained Optimization

The general problem of constrained optimization can be written as

$$(1) \qquad \max_{x \in \mathbb{R}^N} f(x) \quad g(x) \leq b \quad h(x) = c$$

$f : \mathbb{R}^N \to \mathbb{R}$ is a real-valued function, $g : \mathbb{R}^N \to \mathbb{R}^K$, $h : \mathbb{R}^N \to \mathbb{R}^M$ are vector-vaued functions, and $b \in \mathbb{R}^K$ and $c \in \mathbb{R}^M$ are constants. Recall

$$g(x) \leq b \equiv g_k(x) \leq b_k \quad \text{and} \quad h(x) = c \equiv h_m(x) = c_m$$

The function f is called the **objective**, the functions that comprise g are called the **inequality constraints**, and the functions that comprise h are called **equality constraints**.

Example 1. What is the maximum of $f(x) = x^2$ on $[0, 1]$? We can write this as

$$\max_{x \in \mathbb{R}} f(x) \quad g(x) \le b$$

where $g : \mathbb{R} \to \mathbb{R}^2$ is given by $g(x) = (-x, x)$ and $b = (0, 1)$. A more common way to write this, however, is

$$\max_{x \in \mathbb{R}} x^2 \quad \text{s.t.} \quad 0 \le x \le 1$$

In this case, the maximum occurs at $x = 1$ (note $f(x)$ is strictly increasing). $\square$

1.1. One Equality Constraint. Suppose that we only have a single equality constraint:

$$\max_{x \in \mathbb{R}^N} f(x) \quad h(x) = c$$

with $c \in \mathbb{R}$ some scalar. With unconstrained optimization, we look for candidate extrama by solving for x using the FOC. In this case, however, solving for the FOC doesn't immediately help because the candidate points might not be in the constraint. What to do?

- Suppose that we could solve for x_i for some i explicitly as a function of c and the other variables. WLOG let this be x_1; that is, given $h(x) = c$ we find φ s.t.

$$x_1 = \varphi(x_{-1}) \iff h(x) = c$$

  Then we can pug in x_1 and we have

$$\max_{x \in \mathbb{R}^N} f(\varphi(x_{-1}), x_{-1})$$

  The FOC for this transformed problem is

  (2) $\qquad 0 = Df(\varphi(x_{-1}^*), x_{-1}^*) = D_{x_1} f(\varphi(x_{-1}^*), x_{-1}^*) D_{x_{-1}} \varphi(x_{-1}^*) + D_{x_{-1}} f(\varphi(x_{-1}^*), x_{-1}^*)$

  and any x_{-1}^* satisfying Equation (2) for the unconstrained problem gives a candidate point of the constrained problem: $(\varphi(x_{-1}^*), x_{-1}^*)$ with $\varphi(x_{-1}^*) = x_1^*$.

- Given the above, if there was a general way of finding φ we'd be good. This is not possible, but we don't really need to solve for φ *explicitly*. Looking at the FOC, all the expressions are computable except for $D_{x_{-1}} \varphi(x_{-1}^*)$. Looking back at the constraint, if we plug in φ, we can write any x^* satisfies

$$h(\varphi(x_{-1}^*), x_{-1}^*) - c = 0$$

  and now you should be reminded of the IFT!

- In other words, even when we cannot find an explicit expression for φ, we know implicitly such a function exists. Following the theorem, we can compute

$$D_{x_{-1}} \varphi(x_{-1}^*) = -(D_{x_1} h(x^*))^{-1} (D_{x_{-1}} h(x^*))$$

  Plugging this into Equation (2), we find

$$0 = -D_{x_1} f(x^*)(D_{x_1} h(x^*))^{-1}(D_{x_{-1}} h(x^*)) + D_{x_{-1}} f(x^*)$$
$$D_{x_{-1}} f(x^*) = \left[D_{x_1} f(x^*)(D_{x_1} h(x^*))^{-1} \right] D_{x_{-1}} h(x^*)$$

Let $\lambda_1 \equiv D_{x_1} f(x^*)(D_{x_1} h(x^*))^{-1}$ and we have any optimum must satisfy

$$D_{x_{-1}} f(x^*) = \lambda_1 D_{x_{-1}} h(x^*)$$

This should be a very familiar expression!

- Indeed, what we just found is almost exactly the method of Lagrangians you know and love. What's missing? Well, remember 1 was chosen arbitrarily, so we can write down the above expression $2, \dots, N$ and find the corresponding expression with $\lambda_2, \dots, \lambda_N$. However, the gradients at x^* will not change, which implies that $\lambda_i = \lambda_j$. Hence we can simply write

$$D_x f(x^*) = \lambda D_x h(x^*)$$

for any local optimum, which is exactly the method of Lagrange multipliers.

Example 2. Consider $f(x, y) = \sqrt{x \cdot y}$, $h(x, y) = x^2 + y^2$, and $c = 1$. Solve

$$\max_{x,y}(xy)^{1/2} \quad \text{s.t.} \quad x^2 + y^2 = 1$$

We know that $(Df) = \lambda(Dh)$ for some λ, or

$$\begin{bmatrix} \frac{1}{2} x^{-1/2} y^{1/2} \\ \frac{1}{2} x^{1/2} y^{-1/2} \end{bmatrix} = \lambda \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$

$$\begin{bmatrix} x^{-1/2} y^{1/2} \\ x^{1/2} y^{-1/2} \end{bmatrix} = 4\lambda \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\frac{x^{-1/2} y^{1/2}}{x^{1/2} y^{-1/2}} = \frac{x}{y}$$

$$x^2 = y^2$$

The last step is to use the constraint, which must hold at the optimum as well:

$$x^2 + x^2 = 1 \implies x = y = \sqrt{1/2} \quad \text{or} \quad x = y = -\sqrt{1/2}$$

Note that we can't mix and match roots here lest the objective be $\sqrt{}$ a negative number. $\qquad\square$

1.2. Multiple Equality Constraints.

Theorem 1 (Lagrange Multipliers). *Let $f : \mathbb{R}^N \to \mathbb{R}$ and $h : \mathbb{R}^N \to \mathbb{R}^M$ be continuously differentiable and x^* be a solution to*

(3)
$$\max_{x \in \mathbb{R}^N} f(x) \quad \text{s.t.} \quad h(x) = c$$

*for a given $c \in \mathbb{R}^M$. If $\text{rank}(Dh(x^*)) = M \leq N$ (this rank condition is called a **non-degenerate constraint qualification**; why is a NDCQ required?[1]) then $\exists \lambda^* \in \mathbb{R}^M$ s.t.*

$$Df(x^*) = Dh(x^*)\lambda^*$$

$$\frac{\partial f(x^*)}{\partial x_k} = \sum_{i=1}^M \frac{\partial h_i(x^*)}{\partial x_k} \lambda_i^*$$

---

[1]There cannot be more non-redundant constraints than there are variables, lest the problem have no solution!

$\lambda^* = (\lambda_1^*, \ldots, \lambda_M^*)$ *are called **Lagrange multipliers**. (NB: Here Dh is N × M.)*

Note the above is equivalent to stating that if x^* solves Equation (3) then $\exists \lambda^*$ s.t. (x^*, λ^*) satisfy the FOC for

(4) $$\mathcal{L}(\lambda, x) = f(x) - \lambda^T (h(x) - c)$$

In the above equation $\mathcal{L}$ is called the **Lagrangian**, so the result from Theorem 1 that $Df(x^*) = Dh(x^*)\lambda^*$ is referred to as the FOC for the Lagrangian.

Example 3. Let $f(x, y, z) = xyz$, $h_1(x, y, z) = x^2 + y^2$, $h_2(x, y, z) = x + z$, and $c_1 = c_2 = 1$. Solve

$$\max_{x,y,z} f(x, y, z) \quad \text{s.t.} \quad h(x, y, z) = c$$

$$\max_{x,y,z} xyz \quad \text{s.t.} \quad \begin{bmatrix} x^2 + y^2 \\ x + z \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Now we can apply the theorem:

$$\begin{bmatrix} yz \\ xz \\ xy \end{bmatrix} = \begin{bmatrix} 2x & 1 \\ 2y & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}$$

$$xy = \lambda_2$$

$$\frac{xz}{2y} = \lambda_1$$

$$yz = \frac{x^2 z}{y} + xy$$

$$y^2 z = x^2 z + xy^2$$

To finish we need to leverage the constraints: $z = x - 1$ and $y^2 = 1 - x^2$, so

$$(1 - x^2)(1 - x) = x^2(1 - x) + x(1 - x^2)$$

$$0 = (1 - x - 3x^2)(1 - x)$$

The solutions are given by $x = \left(1, \dfrac{1 \pm \sqrt{13}}{6} \right)$. Note that $x = 1$ cannot be a solution. In this case, $y = z = 0$ (this yields an objective of 0, which we can readily check this is not a max, but it is also an issue for how we solved the problem, where $y \neq 0$l is required). For the other two points:

$$x = \frac{1 + \sqrt{13}}{6} \quad z = \frac{5 - \sqrt{13}}{6} \quad y = -\sqrt{1 - \frac{14 + 2\sqrt{13}}{36}} = -\frac{\sqrt{22 - 2\sqrt{13}}}{6}$$

$$x = \frac{1 - \sqrt{13}}{6} \quad z = \frac{5 + \sqrt{13}}{6} \quad y = -\sqrt{1 - \frac{14 - 2\sqrt{13}}{36}} = -\frac{\sqrt{22 + 2\sqrt{13}}}{6}$$

(technically we get 4, but the other two give a negative value for the objective, which is even smaller than 0). With some tedious algebra (or plugging numbers into a computer) you find the function is maximized at the second point. (NB: There is a SOC you can also check; it's coming in Section 2.) $\qquad \square$

1.3. Inequality Constraints. The trick is to realize we can think of inequality constraints as either

1. **Binding**, that is, the optimum occurs along the boundary of the constraint. In this case we can treat them as equality constraints.

2. **Non-binding**, in which case we can de facto ignore them.

In other words, we can solve a problem with inequality constraints in almost the exact way we would a problem with equality constraints. Consider $\max_x f(x)$ s.t. $g(x) \leq b$; one way to approach this:

- Solve for $x^* : Df(x^*) = Dg(x^*)\lambda$ like in a constrained problem assuming $g(x^*) = b$.

- Solve for $x^* : Df(x^*) = 0$ like in an unconstrained problem and check $g(x^*) < b$.

Once you have all the candidate points, check which is largest to determine the max.

Example 4. Consider maximizing $f(x, y) = \sqrt{xy}$ subject to $x + y \leq 1$, $x, y \geq 0$.

$$Df(x, y) = \begin{bmatrix} \frac{1}{2}x^{-1/2}y^{1/2} \\ \frac{1}{2}x^{1/2}y^{-1/2} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}\lambda = Dg(x, y)\lambda$$

$$x = y$$

assuming $x + y = 1$ at the optimum (note that if x or y are 0 then $f(x, y) = 0$, which we can readily check is not a max). Here the candidate point is $(x, y) = (1/2, 1/2)$; now for the unconstrained problem, we note that f is strictly increasing if $x, y > 0$, which means that any such point cannot be an unconstrained max (we can increase either x or y slightly and make the objective bigger). To recap:

- Solving assuming binding constraints we find $x = 0$ or $y = 0$ cannot be max, and the Lagrange multiplier method gives $(x, y) = (1/2, 1/2)$ as a possible maximum.

- Solving assuming non-binding constraints yields no solution.

Hence the only candidate point is $(x, y) = (1/2, 1/2)$. $\qquad\qquad\square$

2. Karush-Kuhn-Tucker (KKT) Conditions

Theorem 2 (Karush-Kuhn-Tucker). *Let $f : \mathbb{R}^N \to \mathbb{R}$ be a real-valued function, $g : \mathbb{R}^N \to \mathbb{R}^K$, $h : \mathbb{R}^N \to \mathbb{R}^M$ vector-valued functions, and $b \in \mathbb{R}^K$ and $c \in \mathbb{R}^M$ constants. Suppose that x^* is such that the constraint qualification is satisfied (non-degenerate) and that it solves*

$$\max_{x \in \mathbb{R}^N} f(x) \quad s.t. \quad g(x) \leq b \quad h(x) = c$$

Then there exist $\mu \in \mathbb{R}^K_+$ (non-negative Lagrange multipliers for the inequality constants), $\lambda \in \mathbb{R}^M$ s.t.

1. *$Df(x^*) = Dg(x^*)\mu + Dh(x^*)\lambda$. This is still the first order condition of the Lagrangian; in this case,*

$$\mathcal{L}(\mu, \lambda, x) = f(x) - \mu^T(g(x) - b) - \lambda^T(h(x) - c)$$

2. *$(g_j(x^*) - b_j)\mu_j = 0$ for each $j = 1, \ldots, K$; this is called the **complementary slackness** condition.*

Complementary slackness states that either the constraint binds with equality or that we can ignore it for the purposes of the FOC (i.e. either $g_j(x^*) = b$ or the corresponding $\mu_j = 0$). Note as we did when solving equality constraints, we typically have to leverage the fact that $h(x^*) = c$ and $g(x^*) \leq b$ to arrive at a solution.

Remark 1. The result of the above theorem is often known as the Kuhn-Tucker conditions, named after two mathematicians who showed the theorem in in the 1950s. However, Karush had already proven this result more than a decade earlier in his (apparently unpublished) masters' thesis. □

2.1. Second Order Conditions. Consider the problem in Equation (1) and suppose the functions are twice continuously differentiable, so the Lagrangian

$$\mathcal{L}(\mu, \lambda, x) = f(x) - \mu^T(g(x) - b) - \lambda^T(h(x) - c)$$

is twice continuously differentiable as well. Suppose (μ^*, λ^*, x^*) is s.t. x^* is a solution to the problem and the KKT conditions are satisfied. Further suppose the constraints $g_1, \ldots, g_L$ are binding and the other $K - L$ constrains are non-binding. Then the Hessian of $\mathcal{L}$ with respect to x is negative definite on the set

$$V = \left\{ v : D\widetilde{g}(x^*)^T v = 0 \quad \text{and} \quad Dh(x^*)^T v = 0 \right\}$$

with $\widetilde{g} = (g_1, \ldots, g_L)$. That is,

$$v^T D_x^2 \mathcal{L}(\mu^*, \lambda^*, x^*) v < 0$$

for any $v \in V$. To check this condition, consider the **_bordered Hessian_**:

$$H = \begin{bmatrix} 0 & 0 & D\widetilde{g}(x^*)^T \\ 0 & 0 & Dh(x^*)^T \\ D\widetilde{g}(x^*) & Dh(x^*) & D_x^2 \mathcal{L}(\mu^*, \lambda^*, x^*) \end{bmatrix}$$

You will note the bordered Hessian is basically the Hessian of the Lagrangian, but it is so-called because it looks like we are "bordering" $D_x^2 \mathcal{L}$ with $D\widetilde{g}, Dh$. In any case, it is sufficient to check that the last $N - L - M$ leading principal minors of H:

- For a **max**, they alternate signs, with the sign of the $(2(L + M) + 1)$th principal minor equal to $(-1)^{L+M+1}$ (note the last $N - L - M$ leading principal minors are from $2(L + M) + 1$ to $2(L + M) + (N - L - M) = N + L + M$).

- For a **min**, all their signs equal $(-1)^{L+M}$.

You will note we are discarding the first $2(L + M)$ leading principal minors. Why?[2]

---

[2]The bordered Hessian cannot be positive or negative definite due to the matrix of zeros in the top left corner (second derivatives with respect to the multipliers are a 0). That discards checking the first $L + M$ leading principal minors. What about the next? Intuitively, we can think that binding conditions impose restrictions of parameters. Remember we motivated our sketch of the proof of the Lagrange multiplier method by invoking the IFT. Why did we do this? Because if for each constraint we express a variable in terms of the others, we can solve the modified unconstrained problem. In other words, each constraint removes a variable from the problem, and there goes the other $L + M$ degrees of freedom.

2.2. Application to Economics. Consider a consumer with wealth $w > 0$ maximizing Cobb-Douglas utility $U(x, y) = x^\alpha y^{1-\alpha}$ for goods $x, y \in \mathbb{R}$ with prices p_x, p_y such that $x, y \geq 0$, $p_x, p_y > 0$.

$$\max_{x,y} x^\alpha y^{1-\alpha} \quad \text{s.t.} \quad x, y \geq 0 \quad \text{and} \quad p_x x + p_y y \leq w$$

Note that $f(x, y) = x^\alpha y^{1-\alpha}$, $g(x, y) = (-x, -y, p_x x + p_y y)$, and $b = (0, 0, w)$. We know from Theorem 2 that any maximum satisfies

$$\begin{bmatrix} \alpha x^{\alpha-1} y^{1-\alpha} \\ (1-\alpha) x^\alpha y^{-\alpha} \end{bmatrix} = \begin{bmatrix} -1 & 0 & p_x \\ 0 & -1 & p_y \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}$$

and $\mu_1 x = \mu_2 y = (p_x x + p_y y - w) \mu_3 = 0$. Let us consider cases: If $x, y \neq 0$ then $\mu_1 = \mu_2 = 0$ and from the FOC we find

$$\frac{\alpha x^{\alpha-1} y^{1-\alpha}}{(1-\alpha) x^\alpha y^{-\alpha}} = \frac{p_x}{p_y}$$

$$\alpha p_y y = (1-\alpha) p_x x$$

If $p_x x + p_y y < w$ then since f is strictly increasing in either argument, we can find some ε s.t. $p_x(x + \varepsilon) + p_y(y + \varepsilon) < w$ but $f(x + \varepsilon, y + \varepsilon) > f(x, y)$. Therefore the constraint is binding, and

$$w = \frac{\alpha}{1-\alpha} p_y y + p_y y$$

$$y = \frac{(1-\alpha) w}{p_y}$$

$$x = \frac{\alpha w}{p_y}$$

Note if x or y are 0 then the objective is 0, which is lower than the objective when both $x, y > 0$ (as is the case above). Hence the only candidate point is given by

$$(x, y) = \left(\frac{(1-\alpha) w}{p_y}, \frac{\alpha w}{p_y} \right)$$

In this case, the Lagrangian was

$$\mathcal{L}(\mu, x) = x^\alpha y^{1-\alpha} - x \mu_1 - y \mu_2 - (p_x x + p_y y - w) \mu_3$$

Note here only the budget constraint is binding, so the bordered Hessian is

$$\begin{bmatrix} 0 & p_x & p_y \\ p_x & \alpha(\alpha-1) x^{\alpha-2} y^{1-\alpha} & \alpha(1-\alpha) x^{\alpha-1} y^{-\alpha} \\ p_y & \alpha(1-\alpha) x^{\alpha-1} y^{-\alpha} & -\alpha(1-\alpha) x^\alpha y^{-\alpha-1} \end{bmatrix}$$

Hence we have two variables and one binding constraint, so we check the last $2 - 1 = 1$ leading principal

minors, which is the third leading principal minor, or just the matrix itself. The determinant is

$$
0 \cdot \det \begin{bmatrix} \alpha(\alpha-1)x^{\alpha-2}y^{1-\alpha} & \alpha(1-\alpha)x^{\alpha-1}y^{-\alpha} \\ \alpha(1-\alpha)x^{\alpha-1}y^{-\alpha} & -\alpha(1-\alpha)x^{\alpha}y^{-\alpha-1} \end{bmatrix}
$$

$$
- p_x \cdot \det \begin{bmatrix} p_x & \alpha(1-\alpha)x^{\alpha-1}y^{-\alpha} \\ p_y & -\alpha(1-\alpha)x^{\alpha}y^{-\alpha-1} \end{bmatrix}
$$

$$
+ p_y \cdot \det \begin{bmatrix} p_x & \alpha(\alpha-1)x^{\alpha-2}y^{1-\alpha} \\ p_y & \alpha(1-\alpha)x^{\alpha-1}y^{-\alpha} \end{bmatrix}
$$

$$
= -p_x \cdot \left[-p_x \alpha(1-\alpha)x^{\alpha}y^{-\alpha-1} - p_y\alpha(1-\alpha)x^{\alpha-1}y^{-\alpha} \right]
$$

$$
+ p_y \cdot \left[p_x\alpha(1-\alpha)x^{\alpha-1}y^{-\alpha} + p_y\alpha(1-\alpha)x^{\alpha-2}y^{1-\alpha} \right]
$$

$$
> 0
$$

Finally, $L + M + 1 = 1 + 1 = 2$, so $(-1)^2 > 0$ means the 3rd leading principal minor must be positive for a max, and that is exactly what we found.

3. Envelope Theorem

Consider the function

$$
f(x;\theta) = x^\theta - x
$$

for $\theta \in (0,1)$. From the FOC, $\theta x^{\theta-1} = 1$, and we can parametrize the maximum for any value of θ, $x^*(\theta) = \theta^{1/(1-\theta)}$.

$$
f(x^*(\theta);\theta) = (x^*(\theta))^\theta - x^*(\theta) = \theta^{\theta/(1-\theta)} - \theta^{1/(1-\theta)}
$$

If we are interested in how f changes with respect to θ, we can take this monster derivative. However, differentiating θ^θ is beyond the ability of any mortal,[3] so instead I will do some wrangling. Note

$$
\frac{d}{d\theta}f(x^*(\theta);\theta) = (\partial_x f(x^*(\theta);\theta)) \frac{dx^*(\theta)}{d\theta} + \partial_\theta f(x^*(\theta);\theta)
$$

From the FOC, however, we know $\partial_x f(x^*(\theta);\theta) = 0$ for each θ, meaning (writing just x^* for ease of notation),

$$
\frac{d}{d\theta}f(x^*;\theta) = \partial_\theta f(x^*;\theta) = \partial_\theta \left(\exp(\theta \log(x)) + x \right)_{x=x^*} = (x^*)^\theta \log(x^*)
$$

This trick of leveraging the FOC to obtain the derivative of the optimized objective with respect to a parameter is called the **envelope theorem**. Graphically, we can see that
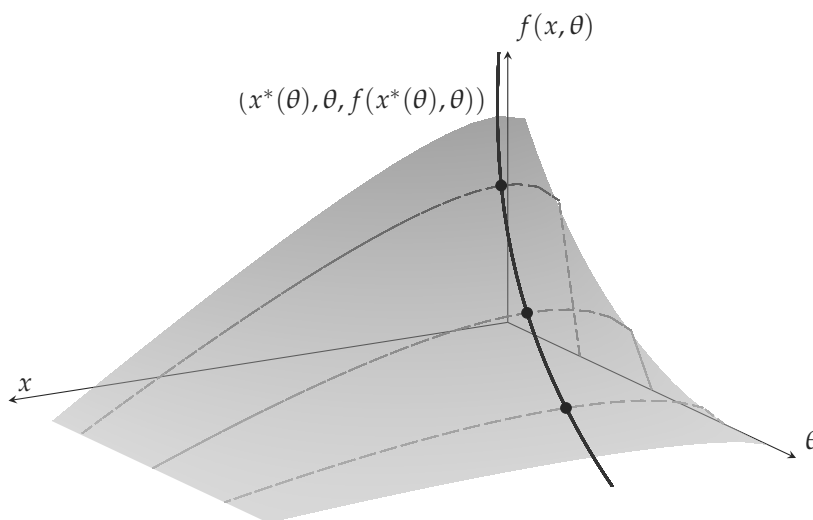
---

[3]This is a joke, of course: Everyone knows you can type "derive θ^θ" into wolframalpha.com ⤤ to get the answer.

Figure 1: Graphical Illustration of the Envelope Theorem
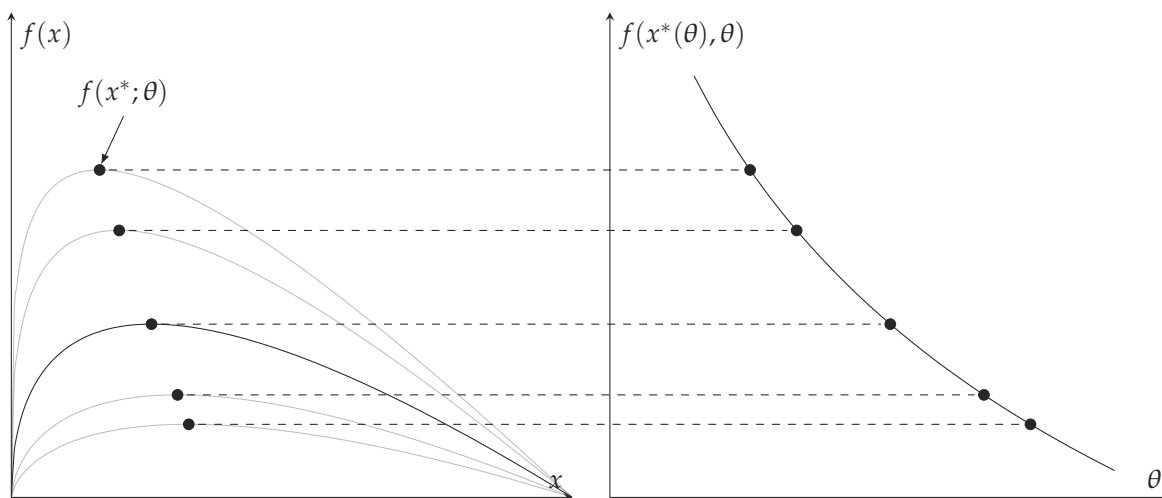
In x–θ space, we can see that $f(x^*(\theta), \theta)$ **envelops** $f(x, \theta)$ at each $(x^*(\theta), \theta)$ (hence the name).

Figure 2: Graphical Illustration of the Envelope Theorem Ctd.

Theorem 3 (Envelope Theorem). *Let $f : \mathbb{R}^N \times \Theta \to \mathbb{R}$ and $h : \mathbb{R}^N \times \Theta \to \mathbb{R}^N$ be continuously differentiable and $(x^*(\theta), \lambda^*(\theta))$ satisfy the FOC of the Lagrangian for the optimization problem at a given $\theta \in \Theta$:*

$$\max_x f(x; \theta) \quad s.t. \quad h(x; \theta) = c$$

*(NB: θ may be a vector of parameters, not just a single parameter.) The **value function** is the maximized value of the objective at each θ, given by*

$$V(\theta) \equiv f(x^*(\theta); \theta)$$

If $V(\theta)$ is continuously differentiable at θ then

$$D_\theta V(\theta) = \Big[D_\theta f(x; \theta) - D_\theta h(x; \theta) \lambda \Big]_{(x, \lambda) = (x^*(\theta), \lambda^*(\theta))}$$

$$\frac{\partial V(\theta)}{\partial \theta_k} = \frac{\partial f(x^*(\theta), \theta)}{\partial \theta_k} - \sum_{m=1}^{M} \lambda_m^*(\theta) \frac{\partial h_m(x^*(\theta), \theta)}{\partial \theta_k}$$

where $\lambda_m^*(\theta)$ is the mth entry of $\lambda^*(\theta)$ and similarly for $h_m(\cdot)$.

Note that in our example, there is no constraint so the second term in the theorem is just 0. (Alternatively, if θ does not appear in the constraint, the second term drops out of the expression as well.)

Remark 2. As we showed the envelope theorem for unconstrained optimization by leveraging the FOC, we can do the same for constrained optimization. Note, dispensing with the stars,

$$D_\theta V(\theta) = (D_\theta x(\theta)) D_x f(x(\theta); \theta) + D_\theta f(x(\theta); \theta)$$

Now recall the FOC:

$$D_x f(x(\theta); \theta) = D_x h(x(\theta); \theta) \lambda(\theta)$$

Plugging this in,

$$D_\theta V(\theta) = (D_\theta x(\theta)) D_x h(x(\theta); \theta) \lambda(\theta) + D_\theta f(x(\theta); \theta)$$

To finish, differentiate the constraint $h(x(\theta); \theta) = c$ with respect to θ to get

$$-D_\theta h(x(\theta); \theta) = (D_\theta x(\theta)) D_x h(x(\theta); \theta)$$

by the chain rule. The right-hand appears in our expression for $D_\theta V(\theta)$; plug in $-D_\theta h(x(\theta); \theta)$ and you get the envelope theorem.   □

3.1. Application to Economics. Consider a consumer minimizing expenditures subject to a minimum utility level u, with utility function $U(x, y) = x^\alpha y^{1-\alpha}$ for goods $x, y \in \mathbb{R}$ with prices p_x, p_y such that $x, y \geq 0$, $p_x, p_y > 0$.

$$\min_{x,y} p_x x + p_y y \quad \text{s.t.} \quad x^\alpha y^{1-\alpha} \geq u$$

If the constraint is not binding at (x, y), that is, $x^\alpha y^{1-\alpha} > u$, then $\exists \varepsilon > 0$ s.t. $(x - \varepsilon)^\alpha (y - \varepsilon)^{1-\alpha} > u$. Since $p_x(x - \varepsilon) + p_y(y - \varepsilon) < p_x x + p_y y$, this point cannot be a minimum. Thus the constraint is binding, and we can apply the envelope theorem. The value function of this problem is called the ***expenditure function***, given by

$$e(p_x, p_y, u) = p_x x(p_x, p_y, u) + p_y y(p_x, p_y, u)$$

with $x(\cdot), y(\cdot)$ the so-called Hicksian demand functions. We can see that to find gradient of the expenditure function with respect to price we can apply the envelope theorem! Noting p_x, p_y are not in the constraint,
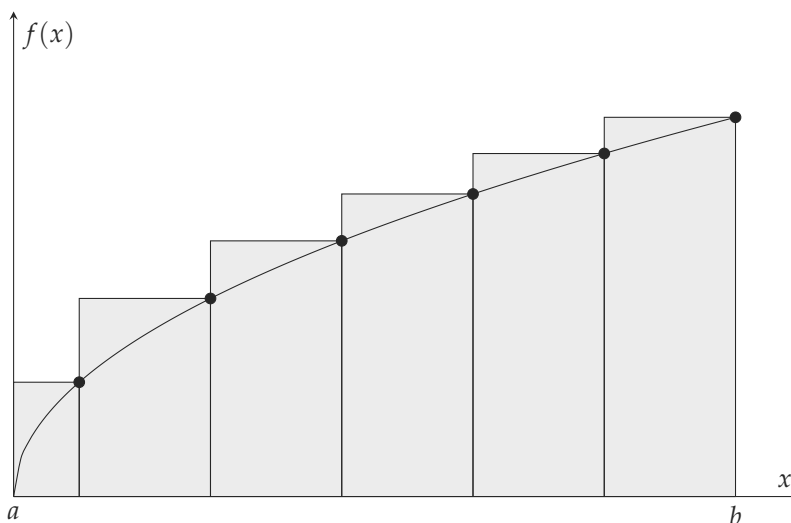
$$D_{(p_x, p_y)} e(p_x, p_y, u) = \begin{bmatrix} x(p_x, p_y, u) \\ y(p_x, p_y, u) \end{bmatrix}$$

4. Integration

4.1. Motivation and Definition. The motivation is the problem of finding the area under a curve:

Figure 3: Area Under $f(x)$
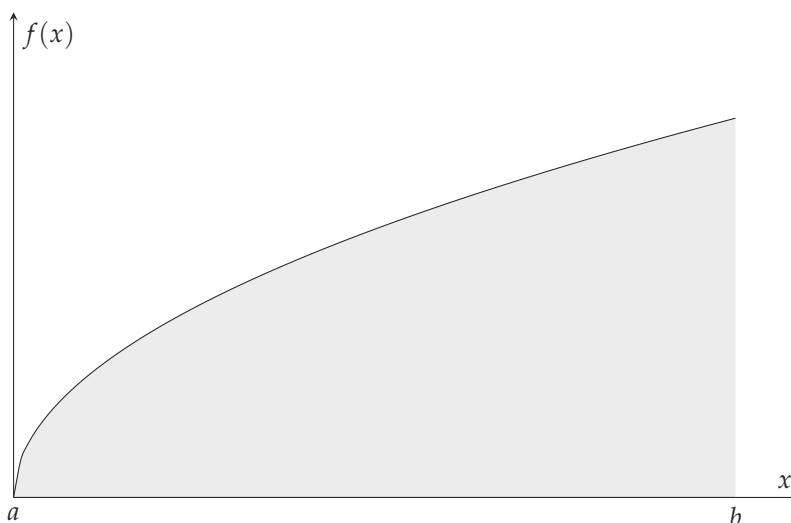
The strategy is to approximate the area under the curve with continuously tighter refinements:

Figure 4: Approximate Area Under $f(x)$

The above approximation is given by

$$\text{Area below } f \text{ from } a \text{ to } b \approx \sum_{i=1}^{m} (x_i - x_{i-1}) f(x_i)$$

We can intuit that as our approximations become tighter, in the limit, we will recover the area under the curve exactly. Let us do this formally.

Definition 1. Let $[a, b]$ be an interval on $\mathbb{R}$. A **partition** P_m is a finite set of points $\{x_0, \ldots, x_m\}$ s.t.

$$a = x_0 \leq x_1 \leq \ldots \leq x_{m-1} \leq x_m$$

We denote $x_i - x_{i-1} = \Delta x_i$.

Definition 2. Let P_m be a partition of $[a, b]$; P_n is a refinement of P_m if $P_m \subseteq P_n$.

Definition 3. Let $f : [a, b] \to \mathbb{R}$ be bounded on $[a, b]$ and for any partition P_m define

$$U_i \equiv \sup_{x \in (x_{i-1}, x_i)} f(x)$$

$$L_i \equiv \inf_{x \in (x_{i-1}, x_i)} f(x)$$

$$U(P_m, f) \equiv \sum_{i=1}^{m} U_i \Delta x_i$$

$$L(P_m, f) \equiv \sum_{i=1}^{m} L_i \Delta x_i$$

U and L are called the **upper Riemann sum** and **lower Riemann sum** of the partition P_m, respectively.

Definition 4. Let $f : [a, b] \to \mathbb{R}$ and consider $\mathcal{P}$ the set of all possible partitions of the interval $[a, b]$.

$$\inf_{P_m \in \mathcal{P}} U(P_m, f)$$

is called the **upper Riemann integral** and

$$\sup_{P_m \in \mathcal{P}} L(P_m, f)$$

is called the **lower Riemann integral**. If they are equal then f is **Riemann integrable** on $[a, b]$ and the **Riemann integral** of f is denoted

$$\int_a^b f(x)dx = \inf_{P_m \in \mathcal{P}} U(P_m, f) = \sup_{P_m \in \mathcal{P}} L(P_m, f)$$

Theorem 4. $f : [a, b] \to \mathbb{R}$ is Riemann integrable iff $\forall \varepsilon > 0 \ \exists P_m$ s.t. $U(P_m, f) - L(P_m, f) < \varepsilon$.

Theorem 5. Let $f : [a, b] \to \mathbb{R}$ and P_n be a refinement of P_m. Then

$$U(P_n, f) \leq U(P_m, f) \quad and \quad L(P_n, f) \geq L(P_m, f)$$

4.2. Properties and Theorems. Some useful properties of Riemann integrals (given $f, g : [a, b] \to \mathbb{R}$)

- Order of integration:
$$\int_a^b f(x)dx = -\int_b^a f(x)dx$$

- Zero-width interval:
$$\int_a^a f(x)dx = 0$$

- Constant multiple:
$$\int_a^b cf(x)dx = c\int_a^b f(x)dx$$

- Sum and difference:
$$\int_a^b (f \pm g)(x)dx = \int_a^b f(x)dx \pm \int_a^b g(x)dx$$

- Sum over intervals:

$$\int_a^b f(x)dx + \int_b^c f(x)dx = \int_a^c f(x)dx$$

- Domination: If $f(x) \le g(x)$ for each $x \in [a, b]$ then

$$\int_a^b f(x)dx \le \int_a^b g(x)dx$$

- Upper and lower bounds: If $L \le f(x) \le M$ for each $x \in [a, b]$ then

$$L(a - b) \le \int_a^b f(x)dx \le U(a - b)$$

Theorem 6. *Let $f : [a, b] \to \mathbb{R}$ be Riemann integrable on $[a, b]$ and $g : f([a, b]) \to \mathbb{R}$ be a continuous function. Then $g \circ f$ is Riemann integrable on $[a, b]$.*
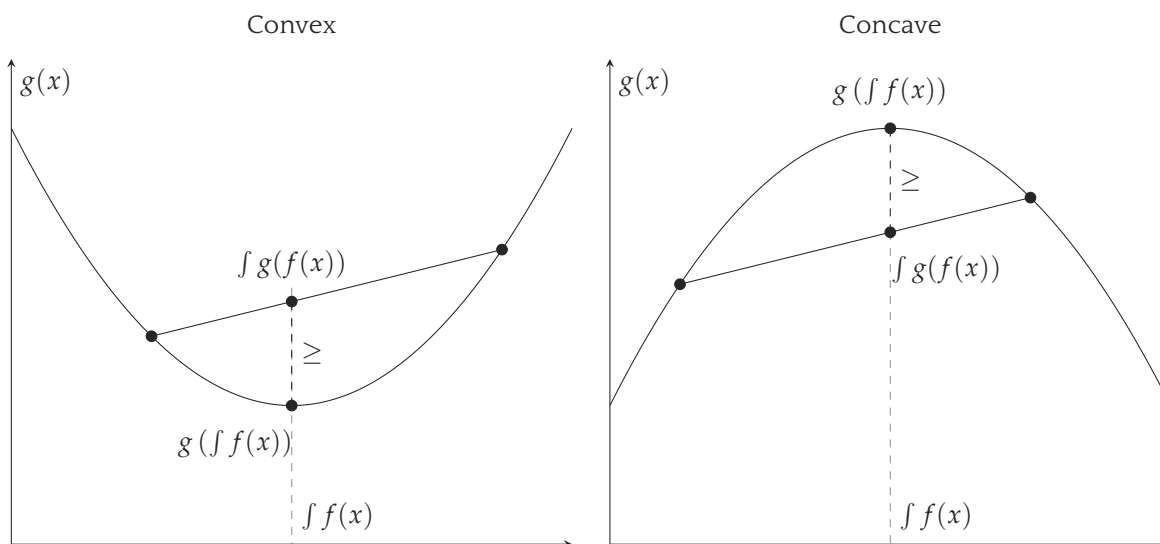
Corollary 1. *Let $f, g : [a, b] \to \mathbb{R}$ be Riemann integrable on $[a, b]$. Then*

- *$f \cdot g$ is Riemann integrable on $[a, b]$.*

- *$|f|$ is Riemann integrable on $[a, b]$.*

Theorem 7 (Jensen's Inequality). *Let $f : [a, b] \to \mathbb{R}$ be Riemann integrable on $[a, b]$ and take $g : \mathbb{R} \to \mathbb{R}$.*

- *If g is convex then $\int_a^b g(f(x))dx \ge g\left(\int_a^b f(x)dx\right)$.*

- *If g is concave then $\int_a^b g(f(x))dx \le g\left(\int_a^b f(x)dx\right)$.*

Figure 5: Visual Intuition for Jensen's Inequality

Corollary 2. *Let $f : [a, b] \to \mathbb{R}$ be Riemann integrable on $[a, b]$. Then $\left|\int_a^b f(x)dx\right| \le \int_a^b |f(x)|dx$.*

4.3. Fundamental Theorem of Calculus.

Definition 5. Let $f : [a, b] \to \mathbb{R}$ be Riemann integrable on $[a, b]$ and $F : [a, b] \to \mathbb{R}$ be differentiable on (a, b) and continuous on $[a, b]$. F is called the **_anti-derivative_** of f if

$$F'(x) = f(x)$$

for each $x \in (a, b)$.

Remark 3. Anti-derivatives are not unique. I'm sure you have lost points in some exam in your life beause you forgot the so-callled constant of integration! In general if F is an anti-derivative of f then $F + G$ is also an anti-derivative of f provided G is not a function of the same variable as f (i.e. G is not a function of x or G is some constant). $\qquad\qquad\square$

Remark 4. One common alternative notation for an integral is

$$\int_a^b f(x)dx = \int_a^b dF(x)$$

with F an anti-derivative of f. The intuition is given by $dF(x) = f(x)dx$, since $F'(x) = f(x)$. $\qquad\square$

Theorem 8 (Fundamental Theorem of Calculus). *Let $f : [a, b] \to \mathbb{R}$ Riemann integrable continuous on $[a, b]$.*

1. *$F(x) \equiv \int_a^x f(t)dt$ is continuous on $[a, b]$, differentiable on (a, b), and an anti-derivative of f. That is,*

$$F'(x) = \frac{d}{dx} \int_a^x f(t)dt = f(x)$$

2. *For any F that is an anti-derivative of f we have*

$$\int_a^b f(x)dx = F(b) - F(a)$$

Remark 5. Why is continuity of f required? We want the function

$$F(x) \equiv \int_a^x f(t)dt$$

to be differentiable, and if f is not continuous at some point c then F will not be differentiable at that point. However, note that even if f is not continuous, F might be continuous: Discontinuities in f introduce a kink in F, and as long as the set of discontinuities is "small" then the function will continue to be integrable and F will remain continuous. (I believe the technical condition is that f has to be continuous "almost everywhere.") As an example, take $f : [0, 2] \to \mathbb{R}$

$$f(x) = 1(x \le 1) + 2(x > 1)$$
$$\int_0^2 f(x)dx = \int_0^2 1(x \le 1) + 2(x > 1)dx$$
$$= \int_0^2 1(x \le 1)dx + \int_0^2 2(x > 1)dx$$
$$= \int_0^1 1dx + \int_1^2 2dx$$
$$= x \cdot 1(x \le 1) + (2x - 1) \cdot 2(x > 1)$$

which has a kink at 1 but remains continuous. A counterexample is the function $f : [0, 1] \to \mathbb{R}$

$$f(x) = 1(x \in \mathbb{Q})$$

That is, 1 at every rational number and 0 elsewhere. This is discontinuous everywhere in $[0, 1]$ and it's actually not Riemann integrable (the $\sup$ and $\inf$ of any partition's interval will be 1 and 0, respectively). $\square$

4.4. Integration Rules. From the FTC we can derive a number of useful rules to help us evaluate integrals.

1. **Change of variables** or **substitution**. For $f, g : [a, b] \to \mathbb{R}$ with f Riemann integrable, g continuously differentiable, and $u = g(x)$,
$$\int_a^b f(g(x))g'(x)dx = \int_{g(a)}^{g(b)} f(u)du$$

   Example 5. Consider $f(x) = 2x\sqrt{1+x^2}$. What is $\int f(x)dx$? Set $u = 1 + x^2$ and, noting $du = 2xdx$,
$$\int 2x\sqrt{1+x^2}dx = \int \sqrt{u}du = \frac{2}{3}u^{3/2} + C = \frac{2}{3}\left(1+x^2\right)^{3/2} + C$$

   for any constant C. $\square$

2. **Leibniz' Rule**: Let f be continuous and a, b be continuously differentiable functions. Then
$$\frac{d}{dx}\int_{a(x)}^{b(x)} f(x, t)dt = f(x, b(x))\frac{db(x)}{dx} - f(x, a(x))\frac{da(x)}{dx} + \int_{a(x)}^{b(x)} \frac{\partial f(x, t)}{\partial x}dt$$

   Example 6. Compute $g'(x)$ where $g(x)$ is given by
$$g(x) = \int_{\sin x}^{x^2} (1 + t)dt$$

   Note that in this case, $f(x, t) = 1 + t$ does not depend on x, so we only have two terms:
$$g'(x) = (1 + x^2)2x - (1 + \sin x)\cos x$$

   $\square$

Remark 6. What happens if we replace $a(x), b(x)$ with $\pm\infty$? If we take limits of constant functions $a(x) = b(x) = c \to \infty$ we can gain additional insight from Leibniz' rule:
$$\lim_{c\to\infty} \frac{d}{dx}\int_{-c}^c f(x, t)dt = \lim_{c\to\infty} f(x, c)\frac{d(c)}{dx} - f(x, -c)\frac{d(-c)}{dx} + \int_{-c}^c \frac{\partial f(x, t)}{\partial x}dt$$

with c some constant. Hence $dc/dx = 0$, and we can conclude
$$\frac{d}{dx}\int_{-\infty}^{\infty} f(x, t)dt = \int_{-\infty}^{\infty} \frac{\partial f(x, t)}{\partial x}dt$$

This comes up in metrics in a few ways. For instance, suppose we have a random variable X with pdf f:
$$\frac{d}{d\theta}E[g(X, \theta)] = \frac{d}{d\theta}\int_{-\infty}^{\infty} g(x, \theta)f(x)dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta}g(x, \theta)f(x)dx = E\left[\frac{\partial}{\partial\theta}g(X, \theta)\right]$$

Another way it comes up is in maximum likelihood. Consider some random variable x_i and some parameter θ_0 and the log likelihood $\log(f(x_i|\theta_0))$. We maximize
$$\max_{\theta\in\Theta} E[\log(f(x_i|\theta))]$$

with Θ the parameter space and the expectation taken over the true likelihood $f(x_i|\theta_0)$. Note the FOC

$$\frac{d}{d\theta}E\left[\log(f(x_i|\theta))\right] = E\left[\frac{\partial}{\partial\theta}\log(f(x_i|\theta))\right] = E\left[\frac{\partial_\theta f(x_i|\theta)}{f(x_i|\theta)}\right] = 0$$

If $\theta = \theta_0$ the true value, we have

$$E\left[\frac{\partial_\theta f(x_i|\theta_0)}{f(x_i|\theta_0)}\right] = \int_{-\infty}^{\infty}\frac{\partial_\theta f(x_i|\theta_0)}{f(x_i|\theta_0)}f(x_i|\theta_0)dx_i$$

$$= \int_{-\infty}^{\infty}\partial_\theta f(x_i|\theta_0)dx_i$$

Now we apply Leibniz rule in reverse to find

$$\int_{-\infty}^{\infty}\partial_\theta f(x_i|\theta_0)dx_i = \frac{d}{d\theta}\int_{-\infty}^{\infty}f(x_i|\theta_0)dx_i = \frac{d}{d\theta}1 = 0$$

where $\int_{-\infty}^{\infty}f(x_i|\theta_0)dx_i = 1$ because f is a pdf. This shows that if $\theta = \theta_0$ then the FOC holds. $\qquad\square$

3. **Integration by Parts**: Let $u,v:[a,b]\to\mathbb{R}$ be continuously differentiable and s.t. $u(x)v'(x)$ is Riemann integrable. Then

$$\int_a^b u(x)v'(x)dx = \left[u(x)v(x)\right]_a^b - \int_a^b v(x)u'(x)dx$$

Remark 7. When I first encountered integration by parts, the formula seemed inscrutable. However, if you stare at it long enough, you will realize it's just the product rule:

$$\frac{d}{dx}u(x)v(x) = u'(x)v(x) + u(x)v'(x)$$

If you integrate both sides and bring the first term to the left then you get the formula. Now, the standard way of writing this formula (for mnemonic purposes) is

$$\int udv = uv - \int vdu$$

My Calculus teacher said a good mnemonic device for the right-hand side is to think "ultra-violet voodoo" in place of uv (ultra-violet) and $\int vdu$ (voodoo). Silly but it has helped me over the years (: $\quad\square$

Example 7. Evaluate the integral

$$\int_0^\infty \frac{x}{\theta}e^{-x/\theta}dx$$

Let $u = x$, $du = dx$, $dv = e^{-x/\theta}/\theta$, and $v = -e^{-x/\theta}$. Thus

$$\int_0^\infty \frac{x}{\theta}e^{-x/\theta}dx = \int_0^\infty udv = \left[uv\right]_0^\infty - \int_0^\infty vdu$$

$$= \left[-xe^{-x/\theta}\right]_0^\infty + \int_0^\infty e^{-x/\theta}dx$$

$$= \lim_{x\to\infty}\frac{-x}{e^{x/\theta}} + 0e^{0/\theta} + \left[-\theta e^{-x/\theta}\right]_0^\infty$$

$$= -\lim_{x\to\infty}\theta e^{-x/\theta} + \theta e^{0/\theta}$$

$$= \theta$$

$\qquad\square$