Informal Notes on Optimization Theory

Junnan He

August 20, 2017

Contents

Chapter 1

Unconstraint Optimization

1.1 Optimization in $\mathbb{R}$

Do more vector calculus, implicit function theorem, theorem of maximum and simple example of existence of equilibrium.

1.1.1 Existence of Extrema

Given a continuous function $f(x) : \mathbb{R} \to \mathbb{R}$. In general, such a function need not have a global maximum nor minimum (e.g. let $f(x) = x^2 \sin(x)$) (pic). One reason is that its domain has no "limit" and the value of the function can get arbitrarily large. Therefore finding the maximum of these functions is an ill-posed question and we should restrict our discussion to problems guarenteed to have a solution. One way to guarenteed the existence of maxima (or minima) is through Weierstrass Theorem.

Theorem 1.1.1. *Let $\mathbb{D} \subsetneq \mathbb{R}$ be a closed and bounded interval, and let $f : \mathbb{D} \to \mathbb{R}$ be a continuous function on $\mathbb{D}$. Then f attains a maximum and a minimum on $\mathbb{D}$.*

This theorem puts a restriction so that the domain is "limited" in exactly the right way for a continuous function to have a maximum and a minimum. Intuitively, the reason is that when a continuous function whose domain is a closed

and bounded interval, its image in the co-domain (or range) is also a closed and bounded interval. Apparently a closed and bounded interval always has a max and a min.

Notice that the extrema can appear on the boundary of the domain (the two ends of a closed bounded interval). However if the global extrema lie in the interior of the domain, we say they are interior extrema. (pic)

1.1.2 Characteristics of Extrema

When we study the maxima of a function f, we can always apply the same arguments to study the function $-f$ and study the minima of the function f. Therefore we need to discuss only one of them and the results would apply to the other.

Consider an objective function f with an interior maximum. Suppose f is also continuously differentiable, then from elementary calculus, we know that at the maximum x^*, the derivative of the function vanishes. I.e. (pic)

$$f'(x^*) = \frac{\mathrm{d}f}{\mathrm{d}x}(x^*) = 0.$$

The reason is that in the interior of the domain, if at a point x where $f'(x) > 0$, we can move x to $x + \delta$ for some infinitesimal δ so that $f(x + \delta) > f(x)$.

More over, if the function is smooth enough, we can even distinguish local maximum from minimum through the second derivative. (pic)

When $f'(x^*) = 0$, if

$$f''(x^*) = \frac{\mathrm{d}^2 f}{dx^2}(x^*) < 0,$$

this implies f' is decreasing near x^*, which means $f'(x^* - \delta) > 0$ and $f'(x^* + \delta) < 0$. In otherwords, the function f is increasing to the left of x^* and decreasing to the right. Hence x^* is a local maximum.

One can also see this through a Taylor expansion of f near x^*

$$f(x) \approx f(x^*) + f'(x^*)(x - x^*) + \frac{1}{2} f''(x^*)(x - x^*)^2 + \dots$$

The second term vanishes since $f'(x^*) = 0$. Hence the function is approximately a quadratic function near x^*. Elementary algebra tells us that if the coefficient $f''(x^*)/2$ of the quadratic term is negative, then x^* is the maximum of the quadratic function.

1.1.3 Exercises

Exercise 1.1.1. *Is it possible for a real valued function $f : \mathbb{D} \to \mathbb{R}$, whose domain $\mathbb{D}$ is closed and bounded, to have no global maximum, global minimum, local maximum nor local minimum?*

Exercise 1.1.2. *Prove the Weierstrass Theorem.*

1.2 Optimization in $\mathbb{R}^n$

In multi-dimensional optimization problems, the existence of a maximum can also be guarenteed by a more general version of the Weierstrass Theorem. If we consider a continuous function $f : \mathbb{D} \to \mathbb{R}$ where $\mathbb{D} \subsetneq \mathbb{R}^n$ is a bounded close set, both maxima and minima are guarenteed to exist.

1.2.1 Multivariate Taylor Approximation

The maxima in higher dimansions also have a similar characterization. Suppose the function f near a critical point x^* can be approximated by its Taylor expansion

$$f(x) \approx f(x^*) + \mathrm{D}f(x^*)^{\mathrm{T}}(x - x^*) + \frac{1}{2}(x - x^*)^{\mathrm{T}} \mathrm{D}^2 f(x^*)(x - x^*) + \dots$$

Notice that in higher dimension, the derivative of a function $\mathrm{D}f$ is a vector and

$\mathrm{D}f^{\mathrm{T}}$ denotes its tranpose. For example, if $f : \mathbb{R}^2 \to \mathbb{R}$,

$$\mathrm{D}f(x^*) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x^*) \\ \frac{\partial f}{\partial x_2}(x^*) \end{bmatrix} = \begin{bmatrix} f_1(x^*) \\ f_2(x^*) \end{bmatrix}, \text{ and } Df^{\mathrm{T}}(x^*) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x^*) & \frac{\partial f}{\partial x_2}(x^*) \end{bmatrix} = \begin{bmatrix} f_1(x^*) & f_2(x^*) \end{bmatrix}.$$

According to the Taylor expansion, the first order linear approximation of f is

$$f(x) \approx f(x^*) + \begin{bmatrix} \frac{\partial f}{\partial x_1}(x^*) & \frac{\partial f}{\partial x_2}(x^*) \end{bmatrix} \begin{bmatrix} x_1 - x_1^* \\ x_2 - x_2^* \end{bmatrix}.$$

It can be seen that for functions on the $\mathbb{R}^2$ plane, the linear approximation is just the approximation by a tengent plane. If the function reaches a maximum at x^*, the tangent plane must lie flat and parallel to the domain. Otherwise if the tangent plane is tilted, one can always increase the value of the function by moving x^* towards the upward directions on the tangent plane. (pic) The tangent plane is flat at x^* if and only if the derivative of f vanishes. I.e. the **first order condition (FOC)** that

$$\mathrm{D}f = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x^*) \\ \frac{\partial f}{\partial x_2}(x^*) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which means in no direction is the function f increasing.

The second derivative $\mathrm{D}^2 f$ of a multivariate real valued function is also called the Hessian matrix. It is obtained by differentiating each component of $\mathrm{D}f$ as a multivariate function. I.e. for functions on the plane, the second derivative at x^* is

$$\mathrm{D}^2 f(x^*) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x^*) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x^*) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x^*) & \frac{\partial^2 f}{\partial x_2^2}(x^*) \end{bmatrix}.$$

When the function is smooth enough, the order of partial differentiation does not matter and hence the Hessian matrix is a symmetric matrix.

When the first derivative vanishes, the Taylor expansion near x^* is approximately

$$f(x) \approx f(x^*) + \frac{1}{2}(x - x^*)^{\mathrm{T}} \mathrm{D}^2 f(x^*)(x - x^*) + \dots$$

If the second derivative satisfies the condition that for any non-zero $v \in \mathbb{R}^n$, v^{T}, $\mathrm{D}^2 f(x^*)v < 0$, then locally $f(x) < f(x^*)$ for any $x \neq x^*$. In other words, x^* is a local maximum. This condition is called **negative definiteness**.

Definition 1.2.1. *An $n \times n$ symmetric matrix H is negative definite if for any non-zero vectors $v \in \mathbb{R}^n$, $v^{\mathrm{T}} H v < 0$.*

When the function f is smooth near x^* and it safisfies both the FOC ($\mathrm{D}f(x^*) = 0$) and the **second order condition (SOC)** that $\mathrm{D}^2 f(x^*)$ is negative definite, then x^* is a local maximum.

1.2.2 Gradient Vector

The derivative $\mathrm{D}f(x^*)$ is also called the gradient vector. It is the vector that points at the direction in which the function increase f increases the fastest. In other words, consider the linear approximation

$$f(x + v) \approx f(x) + Df(x)^{\mathrm{T}} v,$$

it can be shown that if we hold the length of v fixed, $Df(x)^{\mathrm{T}} v$ is maximized when $Df(x) = \delta v$ for some $\delta > 0$. This is the consequence of the Cauchy-Schwarz inequality below.

Proposition 1.2.1. *Fix a non-zero vector $v \in \mathbb{R}^n$, for any $w \neq 0$,*

$$\frac{v^{\mathrm{T}} w}{||w||} \leq \frac{v^{\mathrm{T}} v}{||v||}.$$

Therefore it is intuitive to see that at a local maximum, there is no way to

increase the function and hence $\mathrm{D}f = 0$.

However, the same equation $\mathrm{D}f = 0$ holds when we are at a minimum where there is no way to decrease the function. In order to tell which one is the case, we need to use the second order conditions. We can restrict our attention to the approximation of the function in an arbitrary direction v, and assuming that the first order conditions are satisfied,

$$f(x^* + \delta v) \approx f(x^*) + \frac{1}{2}\delta^2 v^{\mathrm{T}} \mathrm{D}^2 f(x^*) v.$$

Notice that this is a familiar quadratic function in δ whose second order coefficient is $\frac{1}{2}v^{\mathrm{T}}\mathrm{D}^2 f(x^*)v$. When the second order coefficient $\frac{1}{2}v^{\mathrm{T}}\mathrm{D}^2 f(x^*)v$ is negative for every direction v, there is no direction in which the function can be increased, then. In otherwords, the matrix $\mathrm{D}^2 f(x^*)$ being negative definite implies that x^* is a local maximum.

1.2.3 The Envelope Theorem

Although we can usually fully characterize the maxima in terms of FOCs and SOCs, we may not be able to solve them analytically. In application, we usually want to maximize an objection function $f(x;\alpha)$ under a vector of exogeneously given parameters α. The FOCs of the problem could be a system of highly nonlinear equations that are difficult to solve, and we can only numerically approxi--mate the solution $x^*(\alpha)$ given the parameter α. However, we sometimes want to know what happens to the maximized value $M(\alpha) := f(x^*(\alpha);\alpha)$ as the exogeneous parameter changes locally. It turns out we can usually do so without analytically solving for $x^*(\alpha)$. This technique is called the *envelope theorem*.

Intuitively, when α changes, it not only directly affect the maximized value through the second argument in f, it also would affect indirectly the maximized value through tilting the solution x^*. However the envelope theorem says the

indirect effect is negligible, because at the maximum x^*, a small deviation in x can neither increase nor decrease the value of the objective function.

Theorem 1.2.1. *Suppose the objective function $f(x;\alpha)$ is smooth in both arguments, and for each $\alpha \in \mathbb{R}^n$ the maximization problem has a solution $x^*(\alpha)$ that lies in the interior of the domain of $x \in \mathbb{D} \subseteq R^m$. Let $M(\alpha) := f(x^*(\alpha), \alpha)$. Then*

$$\mathrm{D}M(\alpha) = \left(f_{\alpha_1}(x^*(\alpha);\alpha), \ldots, f_{\alpha_n}(x^*(\alpha);\alpha)\right)^{\mathrm{T}}.$$

It is easy to offer an immediate sketch proof. Observe that

$$M_{\alpha_i}(\alpha) = f_{\alpha_i}(x^*(\alpha);\alpha) + \sum_{j=1}^{m} f_{x_j}(x^*(\alpha);\alpha)\frac{\partial x_j^*}{\partial \alpha_i}(\alpha).$$

However $f_{x_j}(x^*(\alpha);\alpha) = 0$ for all j due to the FOCs of the maximization problem. It follows that

$$M_{\alpha_i}(\alpha) = f_{\alpha_i}(x^*(\alpha);\alpha)$$

for every $i = 1, \ldots, n$.

1.2.4 Exercises

Exercise 1.2.1. *State and prove the Weierstrass Theorem in $\mathbb{R}^n$.*

Exercise 1.2.2. *Are there any $n \times n$ real matrix H, that satisfies $v^{\mathrm{T}}Hv < 0$ for all $v \neq 0$ but that H is not symmetric?*

Exercise 1.2.3. *Prove the Cauchy-Schwarz inequality and then the fact that $\mathrm{D}f(x)$ is the direction of fastest increase of f at x.*

Exercise 1.2.4. *Suppose a firm produces $L^{1/4}K^{1/2}$ units of good with L units of labour and K units of capital. The wage per unit labour is w, rent per unit capital is r and the unit price for the firm's output is p all given exogeneously. Show that*

the maximum is given by

$$L^* = \frac{p^4}{64r^2w^2} \text{ and } K^* = \frac{p^4}{32r^3w}.$$

What is the change of profit for an infinitesimal change in output price?

Exercise 1.2.5. *Construct a function f such that at a critical point x^*, all the FOCs are satisfied but the Hessian matrix has some negative eigenvalues and some positive eigenvalues. What does the function look like near x^*?*

1.3 Matrices*

The definition for negative definiteness for the Hessian matrix H is convenient to state, but to check if the conditions are satisfied, we need to be able to argue that for any vector v, $v^{\mathrm{T}}Hv < 0$. This may not be as easy. In this section we provide some basic concepts that is useful for characterizing matrix properties such as negative definiteness.

1.3.1 Eigenvalues and Eigenvectors

It turns out that in order to analyse the properties of a matrix, it is usually enough to study the matrix's action on a few specific vectors. These vectors are called *eigenvectors*.

Definition 1.3.1. *Let M be any $n \times n$ matrix and let v be a non-zero vector that satisfies the equation*

$$Mv = \lambda v$$

for some scalar λ. The non-zero vector v is an eigenvector *of the matrix M and the scalar λ is an* eigenvalue *of the matrix M.*

Given a matrix M we can first find its eigenvector and eigenvalue pairs $\{v_i, \lambda_i\}_{i=1}^n$. These vectors are particularly useful, because for any vector x, one can usually

write x as a linear combination of the eigenvectors. I.e. $x = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n$ for scalars $\alpha_1 \ldots \alpha_n$. Then the multiplication of x by M is simply

$$Mx = \alpha_1 M v_1 + \alpha_2 M v_2 + \cdots + \alpha_n M v_n = \alpha_1 \lambda_1 v_1 + \alpha_2 \lambda_2 v_2 + \cdots + \alpha_n \lambda_n v_n.$$

This is particularly convenient for studying many properties of a matrix. Therefore, we shall now restrict our attention to studying eigenvectors.

To find the eigenvectors and eigenvalues, we shall apply the definition and see that

$$Mv - \lambda v = 0 \Rightarrow (M - \lambda I)x = 0,$$

where I is the $n \times n$ identity matrix. Since v must be a non-zero vector, this implies the the matrix $M - \lambda I$ is singular. Basic linear algebra says that the determinent of a singular matrix is zero. Hences we solve for the equation

$$\det(M - \lambda I) = 0$$

to find the eigenvalues. The above equation is called the *characteristic equation* of the matrix M. The characteristic equation for an $n \times n$ matrix M is, in general, a polynomial in λ of degreen n. And the roots of the polynomial are the eigenvalues of M. In general, eigenvalues of a matrix can be a complex number.

After we solve for the characteristic equation and obtain the value of an eigenvalue λ, we plug it back into the equation

$$(M - \lambda I)v = 0$$

This now becomes a system of n equations with n unknowns $(v_1, v_2, \ldots, v_n) = v$. One can apply the row reduction (or other methods) to solve for the vector v associated with λ. Notice that the solution is never unique, because if v is an eigenvector, so is αv for any $\alpha \neq 0$ (why?). However, these two eigenvectors are

equivalent, because they point towards the same direction (up to a change of sign). However, if a set of eigenvectors each is associated with a different eigenvalue, then this set of eigenvectors must be linearly independent (why?).

1.3.2 Negative Definite Matrices

A negative definite definite matrix M is, by definition, a real symmetric matrix. It has the property that its transpose equals itself $M^{\mathrm{T}} = M$. This makes left and right mutiplication of the matrix be the same. I.e.

$$(x^{\mathrm{T}}M)^{\mathrm{T}} = (x^{\mathrm{T}}M^{\mathrm{T}})^{\mathrm{T}} = Mx.$$

Hence when we consider an eigenvector v associated with the eigenvalue λ,

$$v^{\mathrm{T}}Mv = (Mv)^{\mathrm{T}}v = \lambda v^{\mathrm{T}}v.$$

Notice that $v^{\mathrm{T}}v$ is the inner product of v with itself so it is non-negative. Since v is an eigenvector and must be non-zero, $v^{\mathrm{T}}v > 0$. This means $v^{\mathrm{T}}Mv < 0$ if and only if $\lambda < 0$.

Therefore given a symmetric matrix M, for an arbitrary non-zero vector x, if we can decompose x into the eigenvectors of M that $x = \alpha_1 v_1 + \ldots \alpha_n v_n$, and if for every two eigenvectors $v_i \neq v_j$ we have $v_i^{\mathrm{T}} v_j = 0$, then

$$x^{\mathrm{T}}Mx = \sum_{i,j=1}^{n} \lambda_i \alpha_i v_i^{\mathrm{T}} \alpha_j v_j = \lambda_1 \alpha_1^2 v_1^{\mathrm{T}} v_1 + \ldots \lambda_n \alpha_n^2 v_n^{\mathrm{T}} v_n. \tag{1.1}$$

The sum is negative if and only if all the eigenvalues are negative. Fortunately, these assumptions are all true.

Theorem 1.3.1. *Let M be a real, $n \times n$ symmetric matrix, then all its eigenvalues are real numbers. Moreover, one can choose a set of n linearly independent eigenvectors $\{v_i\}_{i=1}^n$ such that for each two $v_i \neq v_j$, $v_i^{\mathrm{T}} v_j = 0$.*

Because there are n linearly independent eigenvectors, the set of eigenvectors spans $\mathbb{R}^n$ and hence every $x \in \mathbb{R}^n$ can be written as a linear combination of the v_i's. The later condition that $v_i \neq v_j \Rightarrow v_i^{\mathrm{T}} v_j = 0$ is called *orthogonality*. It means each two of the vectors are perpendicular to one another.

Now with the above theorem, equation (1.1) implies the following characterization of negative definite matrices.

Corollary 1.3.1. *A real symmetric matrix M is negative if and only if all its eigenvalues are negative.*

The above Corollary provides an efficient method of checking the negative definiteness of an $n \times n$ matrix. However, for small matrices, there are easier methods. First recall that from matrix algebra,

Theorem 1.3.2. *If M is an $n \times n$ matrix, then the sum of the n eigenvalues of M is the trace of M and the product of the n eigenvalues is the determinant of A.*

A corollary characterizing 2×2 negative definite matrices follows immediately.

Corollary 1.3.2. *Let M be a 2×2 real symmetric matrix. It is negative definite if and only if* $\det(M) > 0$ *and* $\mathrm{tr}(M) < 0$.

1.3.3 Exercises

Exercise 1.3.1. *Show that if v is an eigenvector of M, so is αv for any $\alpha \neq 0$.*

Exercise 1.3.2. *Show that if for a set of eigenvectors of a matrix M, each is associated with a different eigenvalue, then this set of eigenvectors must be linearly independent.*

Exercise 1.3.3. *Find the eigenvalues and the corresponding eigenvectors of the matrix*

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 4 & -17 & 8 \end{bmatrix}.$$

Exercise 1.3.4. *Show that if M is a real $n \times n$ symmetric matrix, then all its eigenvalues are real numbers.*

Hint: show first that if λ (v) is a complex eigenvalue (eigenvector), then so is its conjugate $\bar{\lambda}$ ($\bar{v}$); then show that if M is symmetric then $\lambda v^{\mathrm{T}} \bar{v} = \bar{\lambda} v^{\mathrm{T}} \bar{v}$.

Exercise 1.3.5. *Show that for a $n \times n$ symmetric matrix M, if any two of its n eigenvectors are associated with two distinct eigenvalues, then each pair of the eigenvectors are orthogonal to each other.*

Exercise 1.3.6. *For 2×2 matrices, show that the trace of a matrix equals the sum of the eigenvalues, and that the determinent of a matrix equals the product of the eigenvalues.*

Hint: use the characteristic polynomial.

Chapter 2

Constraint Optimization

Frequently in application, we want to maximize some objective function but there are certain resource constraint binds us away from unconstraint optimal. For example, a firm might want to choose an efficient resource allocation so as to minimize operational cost, but needs to meet a certain production quantity level. In this case, the objective function to be maximized is the negative of the cost function, and the constraint is that output level must be greater than or equal to the targeted level.

Mathematically, the problem is phrased as

$$\begin{aligned} &\max_{x\in\mathbb{D}} f(x) : \mathbb{R}^n \to \mathbb{R} \\ \text{subject to: } &g_1(x) \geq 0; \\ &g_2(x) \geq 0; \\ &\dots \\ &g_k(x) \geq 0. \end{aligned}$$

Apparently, if all constraints are not binding, meaning that at the maximum, all constraints hold with strict inequality, then the solution is the same as that in the unconstraint problem. On the other hand, if some are binding but others are

not, this problem is then the same as finding the maximum when those binding constraints hold with *equality* and there is no other constraints. In the following, we will first introduce the method of Lagarangian in the case where there is only one binding constraint.

2.1 Single Constraint Optimization

2.1.1 Geometry of the Level Set

Consider the problem

$$\max_{x \in \mathbb{D}} f(x) : \mathbb{R}^n \to \mathbb{R}$$
$$\text{subject to: } g(x) \geq 0$$

where we consider only the case that at the maximum, the constraint is binding $g(x^*) = 0$.

Suppose the functions f, g are smooth and well-behaved. The level set $\{x \in \mathbb{R}^n | g(x) = 0\}$ is a manifold. For 2 dimensional problems when $x \in \mathbb{R}^2$, the level set is a smooth curve on the plane (pic). One can immediately see that the the set $S := \{x | g(x) \geq 0\}$ lies on one side of the level set, and the level set is the boundary of S. For this reason, we use the intuitive notation ∂S to denote the level set. By definition, all points x on this boundary has the value $0 = g(x)$. However, their value under the objective function $f(x)$ varies.

Recall that the derivative of a function is a vector that points toward the direction of fastest increase. The vector $\mathrm{D}g(x)$ always points towards the direction in which g increases the fastest and $-\mathrm{D}g(x)$ the direction g decreases the fastest. Since a movement along ∂S has exactly zero increase, the gradient $\mathrm{D}g(x)$ for each x on the ∂S must be perpendicular to the ∂S. Intuitively, when one moves towards the steepest direction, the moving path and the level curve forms a right

angle (pic for equal-altitude curves/contour lines). The vectors that point towards movements along ∂S are called tangent vectors of the level set. If the set S lies on one side of ∂S, then $\mathrm{D}g(x)$ always points towards this side. A movement towards any direction v from ∂S such that $v^{\mathrm{T}}\mathrm{D}g(x) \geq 0$ corresponds to a movement towards the interior of S from its boundary. Apparently any such movement is permitted by the constraint $g(x) \geq 0$ (pic).

2.1.2 The Maxima and the Lagarange Multiplier

One can also consider the derivatives of the function f on ∂S and $\mathrm{D}f(x)$ always points towards the direction in which f increases the fastest at x. For any $x \in \partial S$, let v be a tangent vector of ∂S at x. An infinitesimal movement from x towards v is allowed by the constraint. If it happens that $v^{\mathrm{T}}\mathrm{D}f(x) > 0$, then this infinitesimal movement not only satisfies the constraint, but also increases the function $f(x)$ marginally (why?). This implies that such an x cannot be a local maximum on the constraint. In otherwords, if for some $x \in \partial S$, $\mathrm{D}f(x)$ is not perpendicular to ∂S, then such an x cannot be a local maximum (pic).

Even if for some $x \in \partial S$ such that $v^{\mathrm{T}}\mathrm{D}f(x) = 0$ for every tangent vector v of ∂S at x, we still need to be careful. This is because $v^{\mathrm{T}}\mathrm{D}g(x) = 0$ for every tangent vector v as discussed previously. If $\mathrm{D}g(x)^{\mathrm{T}}\mathrm{D}f(x) > 0$, then one can move towards $\mathrm{D}g(x)$ satisfying the constraint and increases f (pic).

It turns out if we want to find the maximum, we only need to rule out these two situations. In a 2-dimensional problem, one can see that if $x \in \partial S$, if $\mathrm{D}f(x)$ is perpendicular to ∂S, it either points in the same direction as $\mathrm{D}g(x)$ or the opposite. If it is the opposite direction, then x is a local maximum of f (pic). In terms of an equation, we write

$$\mathrm{D}f(x) + \lambda \mathrm{D}g(x) = 0$$

for some $\lambda > 0$. The λ is called a *Lagarange Multiplier*. Moreover, since the constraint is binding, we have also $g(x) = 0$. When $x \in \mathbb{R}^n$, this system of equations

$$\begin{cases} 0 = \mathrm{D}f(x) + \lambda \mathrm{D}g(x) \\ 0 = g(x) \\ 0 < \lambda \end{cases}$$

has $n+1$ equations (n FOCs and 1 constraint) and $n+1$ unknowns. Solving the systems would let us find the maxima. Solving the above system of equations to find the maxima is called the *method of Lagarange Multipliers*.

2.1.3 Exercises

Exercise 2.1.1. *Suppose a function f is smooth in a neighbourhood of x and $\mathrm{D}f(x) \neq 0$. Show that for any v such that $v^{\mathrm{T}}\mathrm{D}f(x) > 0$, then f is increasing in the direction v from x. In otherwords, given x and v, $f(x + \delta v)$ as a function of δ is increasing at $\delta = 0$.*

Exercise 2.1.2. *Show that two vectors $v, w \in \mathbb{R}^n$ are perpendicular if and only if $v^{\mathrm{T}}w = 0$.*

Exercise 2.1.3. *Maximize $2x + 3y$ subject to the constraint $x^2 + y^2 = 1$ using the method of Lagarange Multiplier(s).*

Exercise 2.1.4. *Find the closest point(s) to the origin subject to the constraint $x^2 + xy + y^2 \geq 1$.*

2.2 Karush-Kuhn-Tucker Theorem and Envelope Theorem

2.2.1 A Karush-Kuhn-Tucker Theorem Under Convexity

Generally when when the objective function f has a restricted domain $\mathbb{D}$, a maximum x^* can appear on either the interior of $\mathbb{D}$ or the boundary $\partial\mathbb{D}$. If it is in the interior, the first order conditions must be satisfied at x^*. If it is on the boundary, due to the restricted domain, one can only move in a direction v towards the interior of $\mathbb{D}$. However since x^* is a local maximum, any of such movements cannot increase the value of f. In otherwords, we must have $v^{\mathrm{T}}\mathrm{D}f(x) \leq 0$. In fact, one can show the above intuitive description is a necessary condition.

Proposition 2.2.1. *Suppose $\mathbb{D} \subsetneq \mathbb{R}^n$ has smooth boundaries. Suppose x^* is a local maximizer of a smooth function f on $\mathbb{D}$, then for any v pointing into $\mathbb{D}$ (i.e. $\delta v + x^* \in \mathbb{D}$ for all $\delta > 0$ sufficiently small), $v^{\mathrm{T}}\mathrm{D}f(x^*) \leq 0$.*

It turns out that the above condition is "almost sufficient" as well, and in many application contexts, the sufficiency is guarenteed. For example, in economics it is usually assumed that the objective function f is concave and the domain $\mathbb{D}$ is convex. A concave function has only one maximum when the domain is concave (why?). As mentioned before, if it appears in the interior, it must be that $\mathrm{D}f(x^*) = 0$. If it appears on the boundary, then a movement v towards the interior must not increase f. Under the convexity assumptions this is both necessary and sufficient.

Proposition 2.2.2. *Suppose a smooth function $f : \mathbb{D} \to \mathbb{R}$ is concave and $\mathbb{D} \subseteq \mathbb{R}^n$ is convex. Then $x^* \in \mathbb{D}$ is the maximum of f if and only if $v^{\mathrm{T}}\mathrm{D}f(x^*) \leq 0$ for all v pointing into $\mathbb{D}$.*

Observe that with a set of constraints $g_1, \ldots, g_k \geq 0$, if each g_i is a weakly concave function, then the set of x satisfying all these constraints is a convex set

(why?). When the constraints $g_i, \ldots, g_k$ are binding, a movement in the direction v is permitted if and only if it is pointing at a direction that increases every binding constraint. In other words, $v^{\mathrm{T}}\mathrm{D}g_i \geq 0$ for every g_i. However, at a local maximum x^*, it must be the case that moving towards any of such v's causes a decrease in f. In otherwords, we must have $v^{\mathrm{T}}\mathrm{D}f(x^*) \leq 0$. These two observation implies that $\mathrm{D}f(x^*)$ must be a linear combination of the gradients $\mathrm{D}g_i(x^*)$ (why?). So we write

$$\mathrm{D}f(x^*) = \sum_{i=1}^{k} \alpha_i \mathrm{D}g_i(x^*)$$

for some $\alpha_i \in \mathbb{R}$, $i = 1, \ldots, k$. Moreover, to ensure that if v is a permitted movement then $v^{\mathrm{T}}Df(x^*) \leq 0$, we impose the additional restriction that $\alpha_i = -\lambda_i$ where $\lambda_i \geq 0$ for $i = 1, \ldots, k$ (why?). Therefore

$$\mathrm{D}f(x^*) = -\sum_{i=1}^{k} \lambda_i \mathrm{D}g_i(x^*) \text{ for some } \lambda_i \geq 0, i = 1, \ldots, k.$$

In fact, similar to the previous propositions, this intuitive argument is also "amost necessary and sufficient". Formalising this intuition leads us to the following version of the Karush-Kuhn-Tucker Theorem which is the most commomly used version in economics.

Theorem 2.2.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a smooth concave function and $g_1, \ldots, g_k : \mathbb{R}^n \to \mathbb{R}$ be smooth and weakly concave. Suppose also that*

$$\{x \in \mathbb{R}^n | g_1(x) > 0, \ldots, g_k(x) > 0\} \neq \emptyset.$$

Then x^* is the unique maximum of f over

$$\mathbb{D} := \{x \in \mathbb{R}^n | g_1(x) \geq 0, \ldots, g_k(x) \geq 0\}$$

if and only if there exists $\lambda_1, \ldots, \lambda_k$ *such that*

$$\begin{cases} 0 = \mathrm{D}f(x^*) + \sum_{i=1}^{k} \lambda_i \mathrm{D}g_i(x^*); \\ 0 = \sum_{i=1}^{k} \lambda_i g_i(x^*); \\ 0 \leq \lambda_i \text{ for } i = 1, \ldots, k. \end{cases}$$

Since each λ_i is non-negative, the condition $0 = \sum_{i=1}^{k} \lambda_i g_i(x^*)$ simply means if g_i is not binding, then $\lambda_i = 0$. In most application, a constraint is not binding if and only if its corresponding $\lambda = 0$. The non-binding constraints are effactively erased by setting its corresponding Lagarange multiplier to zero. Therefore if we move away from x^* in the direction v, the movement is permitted if and only if $v^{\mathrm{T}} \lambda_i \mathrm{D}g_i(x^*) \geq 0$ for all i. Since $\mathrm{D}f(x^*) = -\sum_{i=1}^{k} \lambda_i \mathrm{D}g_i(x^*)$, this movement causes a loss in f. I.e.

$$v^{\mathrm{T}} \mathrm{D}f(x^*) = -\sum_{i=1}^{k} v^{\mathrm{T}} \lambda_i \mathrm{D}g_i(x^*) \leq 0.$$

2.2.2 A Karush-Kuhn-Tucker Theorem without Convexity

However, these convexity assumptions do not always hold. The Karush-Kuhn-Tucker Theorem can be stated more generally without these assumptions, but it reduces to a theorem that only states the necessity conditions for a local maximum.

Theorem 2.2.2. *Suppose* $f, g_1, \ldots, g_k : \mathbb{R}^n \to \mathbb{R}$ *are smooth functions. Let* $\mathbb{D} := \{x \in \mathbb{R}^n | g_1(x) \geq 0, \ldots g_k(x) \geq 0\}$ *for some* $k < n$*. Suppose it holds that*

1. $x^* \in \mathbb{D}$ *is a local maximum of* f *in* $\mathbb{D}$*;*
2. $g_i(x^*) = 0$ *for* $i = 1, \ldots, k$*;*
3. *the set of vectors* $\{\mathrm{D}g_1(x^*), \ldots, \mathrm{D}g_1(x^*)\}$ *is linearly independent.*

Then there exist multipliers $\lambda_1, \ldots, \lambda_k \geq 0$ *that*

$$\mathrm{D}f(x^*) = -\sum_{i=1}^{k} \lambda_i \mathrm{D}g_i(x^*).$$

The above theorem states that if we know which constraints are binding at x^*, and if we knows that at the maximum, the derivatives of the binding constraints are linearly independent, then we can find x^* by solving the system of equations

$$\begin{cases} 0 = \mathrm{D}f(x^*) + \sum_{i=1}^{k} \lambda_i \mathrm{D}g_i(x^*); \\ 0 = g_i(x^*) \text{ for } i = 1, \ldots, k; \\ 0 \leq \lambda_i \text{ for } i = 1, \ldots, k. \end{cases}$$

In application, if we cannot predict which set of constraints are binding, we might need to do "trial and error". We pick a set of constraints that we believe to be binding. Assume that the linearly independency condition holds. Write down the Lagarangian with the binding constraints

$$L(x, \lambda_1, \ldots, \lambda_k) = f(x) + \sum_{i=1}^{k} \lambda_i g_i(x).$$

Differentiate the Lagarangian with respect to x as well as the λ's and set them equal to 0, we obtain the system of FOCs

$$\begin{cases} 0 = \mathrm{D}f(x) + \sum_{i=1}^{k} \lambda_i \mathrm{D}g_i(x) & \\ 0 = g_i(x) & i = 1, \ldots, k. \end{cases}$$

If the set of constraints are rightly chosen, and the gradients of the constraints are linearly independent at the maximum, then the solutions to this system of equations include the maximum.

The linearly independency of the gradients of the constraints might be easily checked if the number of constraints is small. Sometimes it is possible to show that $\{\mathrm{D}g_1(x), \ldots, \mathrm{D}g_k(x)\}$ is linearly independent for all $x \in \mathbb{D}$, then this implies at the global maximum, the assumptions in the previous theorem is satisfied. Hence we can write down the Lagarangian and solve the system of equations. The global

maxima, if exist, will appear in the set of solutions.

In the exercises, there is one example where the linearly independency does not hold. The problem is that at the global maximum, $\mathrm{D}g = 0$ but $\mathrm{D}f \neq 0$. Hence there is no λ that satisfies the equation

$$\mathrm{D}f(x^*) = -\lambda \mathrm{D}g(x^*).$$

Therefore, the global maximum will not turn up in the solutions to the systems of FOCs of the Lagarangian.

2.2.3 Envelope Theorem Under Constraint

Consider the maximization

$$\max_{x \in \mathbb{D}} f(x; \alpha)$$
$$\text{subject to: } g(x; \alpha) \geq 0.$$

Suppose the constraint is binding at the maximum. We want to see how does the maximized value change in reaction to a change in α. This is, in fact very similar to the Envelop theorem in the unconstraint case.

Theorem 2.2.3. *Suppose the objective function $f(x; \alpha)$ and the constraint function $g(x; \alpha)$ are smooth in both arguments. Suppose that for each $\alpha \in \mathbb{R}^n$ the maximization problem has a solution $(x^*(\alpha), \lambda^*(\alpha))$. Let $M(\alpha) := f(x^*(\alpha), \alpha)$. Then*

$$\mathrm{D}M(\alpha) = f_\alpha(x^*; \alpha) + g_\alpha(x^*, \alpha)\lambda^*.$$

Notice that this theorem holds for n binding constraints. We just need to interpret g as a vector valued function and $g \geq 0$ as each of its entries is greater

than 0. The sketch proof is the following. By definition,

$$\mathrm{D}M(\alpha) = \mathrm{D}x^*(\alpha)f_x(x^*;\alpha) + f_\alpha(x^*;\alpha);$$
$$\mathrm{D}g(x^*(\alpha);\alpha) = \mathrm{D}x^*(\alpha)g_x(x^*;\alpha) + g_\alpha(x^*;\alpha) = 0.$$

Since at the maximum, KKT theorem says $f_x(x^*;\alpha) = -g_x(x^*;\alpha)\lambda^*$. Substitute in and obtain

$$\mathrm{D}M(\alpha) = g_\alpha(x^*;\alpha)\lambda^* + f_\alpha(x^*;\alpha).$$

The interpretation is the following. Consider the case with only one constraint g and α is a scaler. Similarly, $f_\alpha(x^*;\alpha)$ is the direct effect of a change of α to f, and $g_\alpha(x^*;\alpha)$ is the direct effect of a change of α to g. If g changes by g_α, the boundary of the constraint set $g \geq 0$ move "inside" or "outside" (depending if $g_\alpha < 0$ or > 0) according to the direction g_x (see pic). The amount of "movement" is $\frac{g_\alpha}{||g_x||^2}$ (why?). Therefore, when we keep the constraint binding, we have to move x^* by a vector $-\frac{g_\alpha(x^*;\alpha)}{||g_x(x^*;\alpha)||^2}g_x(x^*;\alpha)$. Hence this affects the value of f by

$$-\frac{g_\alpha(x^*;\alpha)}{||g_x(x^*;\alpha)||^2}g_x(x^*;\alpha)^{\mathrm{T}}f_x(x^*;\alpha) = \lambda\frac{g_\alpha(x^*;\alpha)}{||g_x(x^*;\alpha)||^2}g_x(x^*;\alpha)^{\mathrm{T}}g_x(x^*;\alpha) = g_\alpha(x^*;\alpha)\lambda^*.$$

In other words, a change in α directly change the constraint and x^* is forced to move. This movement results in a change of f value of the amount $g_\alpha(x^*;\alpha)\lambda^*$.

2.2.4 Exercises

Exercise 2.2.1. *Prove the two propositions in this section.*

Exercise 2.2.2. *Suppose f is a concave function over a convex domain. Show that if a maximum exist, it is the unique one.*

Exercise 2.2.3. *Suppose that $g_1, g_2 : \mathbb{R}^n \to \mathbb{R}$ are both weakly concave. Show that the set $\{x \in \mathbb{R}^n | g_1(x) \geq 0, g_2(x) \geq 0\}$ is convex.*

Exercise 2.2.4. *Let $w_0, \ldots, w_k$ be vectors in $\mathbb{R}^n$. Suppose that for every $v \in \mathbb{R}^n$ such that $v^{\mathrm{T}} w_i \geq 0$ for $i = 1, \ldots, k$, then $v^{\mathrm{T}} w_0 \leq 0$. Show that w_0 lies in the subspace spanned by $\{w_1, \ldots, w_k\}$.*

Hint: suppose the conclusion doesn't hold, then $w_0 = w^* + u$ where w^* is a linear combination of $w_1, \ldots, w_k$ and $u \neq 0$ satisfies $u^{\mathrm{T}} w_i = 0$ for all $i = 1, \ldots, k$. Derive a contradiction. (w^* is called the orthogonal projection of w_0 to the space spanned by $\{w_1, \ldots, w_k\}$.)

Exercise 2.2.5. *Show that if $\mathrm{D}f(x) = -\sum_{i=1}^{k} \lambda_i \mathrm{D}g_i(x)$ for $\lambda_i \geq 0$, then any movement from x permitted by the binding constraints $g_1, \ldots, g_k \geq 0$ cannot increase the value of f.*

Exercise 2.2.6. *Suppose $f := 2x^3 - 3x^2$ and $g(x) = (3 - x)^3$. Find the x^* that maximizes f subject to $g(x) \geq 0$.*

2.3 An Application

2.3.1 A Finite Horizon Optimization Example

Suppose a farmer owns a piece of land and grows crops. There is no market for trade and the grains are perishables, so he consumes some of his own harvest and sows the rest for the next period. So at each time t, he separates the total amount of harvest into consumption c_t and capital k_{t+1} sowed for the future. If the technology for production is such that if he has k_t amount of grains sowed from the previous period, the total amount of output is $f(k_t)$ for this period. Naturally, $f(k_t) \geq c_t + k_{t+1}$.

The farmer values only consumption. If he consumes an amount c_t at time t, he receives an amount of satisfaction of $u(c_t)$ where u is a real valued utility function. However, he discount future consumption at a rate $\beta \in (0, 1)$. In otherwords, from

his perspective at time zero, a consumption of c_t at time t would results in the amount of utility $\beta^t u(c_t)$.

Now the farmer would like to consider planning future production for T periods and nothing afterwards. Then he faces the dynamic planning problem

$$\begin{aligned} &\max_{k_{t+1},c_t} \sum_{t=0}^{T} \beta^t u(c_t) \\ \text{subject to: } & f(k_t) - c_t - k_{t+1} \geq 0, \text{ for } t = 0, \ldots, T; \\ & k_{t+1} \geq 0, \text{ for } t = 0, \ldots, T; \\ & c_t \geq 0, \text{ for } t = 0, \ldots, T; \\ & k_0 \text{ is given.} \end{aligned}$$

It would be both convenient and sensible to impose some regularity assumptions on the utility function u and the production function f. For instance, they should be both strictly increasing functions subject to decreasing return. For instance, u flats out due to satiation from consumption, and f flats out due to the the limited ability to plow or the limited fertility of the land. In otherwords, $u', f' > 0$ and $u'', f'' < 0$. Because of this assumption, we know that the first set of constraints $f(k_t) - c_t - k_{t+1} \geq 0$ is binding for every t (why?). Moreover, in economics, usually utility functions satisfies the condition that $\lim_{c \to 0^+} u'(c) = \infty$. When this is the case, we see that the other two type of constraints never bind except for $k_{T+1} \geq 0$. One should check that all the conditions for one of the Karush-Kuhn-Tucker Theorem is satisfied. To apply the theorem, we write down the following Lagarangian

$$L := \sum_{t=0}^{T} \beta^t \left(u(c_t) + \lambda_t \left(f(k_t) - c_t - k_{t+1} \right) \right)$$

where k_0 is given and $k_{T+1} = 0$.

The set of FOCs are

$$\begin{cases} u'(c_t) - \lambda_t = 0 & \text{for differentiating } c_t,\ t = 0, \ldots, T; \\ -\lambda_t + \beta\lambda_{t+1} f'(k_{t+1}) = 0 & \text{for differentiating } k_{t+1},\ t = 0, \ldots, T-1; \\ f(k_t) - c_t - k_{t+1} = 0 & \text{for differentiating } k\lambda_t,\ t = 0, \ldots, T. \end{cases}$$

The three sequences of equations can be simplified to

$$u'(f(k_t) - k_{t+1}) = \beta u'(f(k_{t+1}) - k_{t+2}) f'(k_{t+1}),\ \text{where } t = 0, \ldots, T-1.$$

This is the so called *Euler equation* (although Euler was mainly working on the continuous time version of dynamic optimization problems) and it describes the intertemperal trade-offs between consumption and investment decisions.

For simplicity let's pick the following parametric form of the relevant functions

$$\begin{cases} u(c) := \ln(c); \\ f(k) := k^\alpha \qquad \alpha \in (0, 1). \end{cases}$$

The log-utility function belongs to the class called *constant relative risk aversion* (CRRA) utility functions and the production function exhibits decreasing return to scale in capital. These two type of parametric forms are frequently used in economic modelling.

Under the parametric assumption, the Euler equation becomes

$$\begin{aligned} \frac{1}{k_t^\alpha - k_{t+1}} &= \beta \frac{1}{k_{t+1}^\alpha - k_{t+2}} \alpha k_{t+1}^{\alpha-1} \\ \Rightarrow \frac{k_{t+2}}{k_{t+1}^\alpha} &= 1 + \alpha\beta - \alpha\beta \frac{k_t^\alpha}{k_{t+1}} \qquad \text{where } t = 0, \ldots, T-1. \end{aligned}$$

Generally to solve a finite period optimization problem, we solve the maximization problem at the last period first. Then given the solution to the last period, we

solve the second-to-last period maximization and so on... This is called the method of *backwards induction.*

The last period decision problem has solution $k_{T+1} = 0$ as discussed before. We plug it into the last Euler equation $\frac{k_{T+1}}{k_T^\alpha} = 1 + \alpha\beta - \alpha\beta\frac{k_{T-1}^\alpha}{k_T}$ to get

$$0 = 1 + \alpha\beta - \alpha\beta\frac{k_{T-1}^\alpha}{k_T} \Rightarrow \frac{k_T}{k_{T-1}^\alpha} = \frac{\alpha\beta}{1+\alpha\beta} = \alpha\beta\frac{1-\alpha\beta}{1-(\alpha\beta)^2}.$$

We plug this into the second to last Euler equation $\frac{k_T}{k_{T-1}^\alpha} = 1 + \alpha\beta - \alpha\beta\frac{k_{T-2}^\alpha}{k_{T-1}}$ to get

$$\frac{\alpha\beta}{1+\alpha\beta} = 1 + \alpha\beta - \alpha\beta\frac{k_{T-2}^\alpha}{k_{T-1}} \Rightarrow \frac{k_{T-1}}{k_{T-2}^\alpha} = \frac{\alpha\beta + (\alpha\beta)^2}{1+\alpha\beta+(\alpha\beta)^2} = \alpha\beta\frac{1-(\alpha\beta)^2}{1-(\alpha\beta)^3}.$$

Similarly,

$$\frac{k_{T-2}}{k_{T-3}^\alpha} = \frac{\alpha\beta + (\alpha\beta)^2 + (\alpha\beta)^3}{1+\alpha\beta+(\alpha\beta)^2+(\alpha\beta)^3} = \alpha\beta\frac{1-(\alpha\beta)^3}{1-(\alpha\beta)^4}, \text{ etc...}$$

One plugs in the pattern into the Euler equation and check (check by yourself!) that for $t = 1, \ldots, T+1$,

$$\frac{k_t}{k_{t-1}^\alpha} = \alpha\beta\frac{1-(\alpha\beta)^{T-t+1}}{1-(\alpha\beta)^{T-t+2}}.$$

Since k_0 is given at the beginning, we sequentially solve for all k_t by

$$\begin{aligned}
k_1 &= \frac{k_1}{k_0^\alpha}k_0^\alpha = \alpha\beta\frac{1-(\alpha\beta)^T}{1-(\alpha\beta)^{T+1}}k_0^\alpha; \\
k_2 &= \frac{k_2}{k_1^\alpha}k_1^\alpha = \alpha\beta\frac{1-(\alpha\beta)^{T-1}}{1-(\alpha\beta)^T}k_1^\alpha = \alpha\beta\frac{1-(\alpha\beta)^{T-1}}{1-(\alpha\beta)^T}\left(\alpha\beta\frac{1-(\alpha\beta)^T}{1-(\alpha\beta)^{T+1}}k_0^\alpha\right)^\alpha; \\
&\text{etc ...}
\end{aligned}$$

This is the solution to our dynamic planning problem. At each period after harvest, k_{t+1} is given by the above sequence, and the remaining outputs are the consumption c_t.

2.3.2 Exercises

Exercise 2.3.1. *Check that the conditions for Karush-Kuhn-Tucker Theorem under convexity are satisfied.*

Exercise 2.3.2. *Re-do the example with general CRRA utility function* $u(c) := c^\rho/\rho$ *where* $\rho < 1$*. What is the Euler equation? Try to solve the optimization for* $T = 2$ *and* $T = 3$*.*

Exercise 2.3.3. *Solve the example with* $f(k) = k^\alpha$ *and* $u(c) = \rho c$ *for* $\alpha \in (0, 1)$*,* $\rho > 0$*.*

Exercise 2.3.4. *Suppose* $f(k) = \alpha k$ *with* $\alpha > 0$ *in the example and the utility function is* $u(c) = c^\rho/\rho$ *with* $\rho < 1$*. Solve the optimization problem.*

Chapter 3

Infinite Horizon Optimization

3.1 Discrete Time Dynamic Programming

3.1.1 Value Function and Functional Equation

There are many optimization problems that are posed in finite time horizons, but many of them are just as sensible, if not better to be posed in infinite time. Consider the farmer's example in the previous chapter. What if the farmer wants to plan for really long term, and takes $T \to \infty$?

To this end, we consider the following problem. Given k_0,

$$\max_{\{k_{t+1},c_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t u(c_t)$$

$$\text{subject to: } k_{t+1} + c_t \leq f(k_t), \text{ for } t = 0, 1, \ldots$$

From the previous chapter, the solution to the finite horizon problem states

$$k_{t+1} = \alpha\beta \frac{1 - (\alpha\beta)^{T-t+1}}{1 - (\alpha\beta)^{T-t+2}} k_t^{\alpha}$$

for $t = 0, 1, \ldots T$ and $c_t = f(k_t) - k_{t+1}$. Naively if we simply take $T \to \infty$, the

solution becomes

$$k_{t+1} = \alpha\beta k_t^{\alpha}.$$

The form of this relation suggests that when T becomes large, the optimal choice of k_{t+1} depends only on k_t. Why is it so?

Recall that in solving the finite time optimization problem, we solved the last period first. When the last period is optimized, we moved to the second to last period etc... The similar intuition can be applied here. Suppose at time t, for any value k that we can pick for $t+1$, the future optimization problem has been solved and the maximized value is given by $\beta^{t+1}v(k)$. I.e. let $k_{t+1} = k$,

$$v(k) := \max_{k_{\tau+t+1}\in[0,f(k_{\tau+t})]} \sum_{\tau=0}^{\infty} \beta^{\tau} u(f(k_{\tau+t+1}) - k_{\tau+t+2}).$$

The function v is called the *value function* of the problem starting from $t+1$. However, notice that we can relable the indices and write

$$v(k) = \max_{k_{\tau+1}\in[0,f(k_{\tau})]} \sum_{\tau=0}^{\infty} \beta^{\tau} u(f(k_{\tau}) - k_{\tau+1})$$

by labeling the time $t+1, t+2, t+3\ldots$ by $0, 1, 2\ldots$ etc. It is readily seen that if the initial amount of capital are the same, the value function for time $t+1$ is the same as the value function for time 0, and hence also the same for any other time. This is simply because in infinite time horizon, at any time t, if it is given that $k_t = k$, the optimization problem from time t and on to the infinite future is the same (pic). Therefore, the solution of the infinite horizon optimization should take the form that the next period capital k_{t+1} should depend only on the initial amout of capital k_t in this period. Most importantly, this relation between k_{t+1} and k_t should be the same across time. In other words, for any $t = 0, 1, \ldots$, $k_{t+1} = h(k_t)$ for some function h. This function h that always return the optimal amount of capital for the next period is called the *policy function*.

Also, since the value function is the same across time, we can consider the following alternative optimization problem.

$$\max_{c_0,\ k} u(c_0) + \beta v(k)$$
$$\text{subject to: } f(k_0) \geq c_0 + k$$
$$c_0, k \geq 0$$

where k_0 is given. If we can apply the backward induction method, let $v(k)$ be the optimized sum of infinite future stream of utilities given k, then solving the above maximization problem gives us the optimal level of capital k given k_0. In otherwords, if we know the function v, we can find the policy function by solving an easy 2 period maximization problem. Since future optimization is assumed to be done and has total payoff $v(k)$, the maximized value of the above problem should be the maximized value of the infinite sequence problem. In other words, the maximized value v as a function of initial capital should satisfies the following equation.

$$v(k_0) = \max_{0 \leq k \leq f(k_0)} u(f(k_0) - k) + \beta v(k).$$

As intuitive as it is, we shall prove that under very general assumptions, the maximized value of the sequential dynamic optimization problem *is a* solution to the functional equation. Consider the sequential optimization problem (SP) in general

$$\sup \sum_{t=0}^{\infty} \beta^t F(x_t, x_{t+1}) \tag{SP}$$
$$\text{subject to: } x_{t+1} \in \Gamma(x_t), \text{ for } t = 0, 1, \ldots;$$
$$x_0 \text{ is given.}$$

F is the period by period payoff function, such as $u(f(k_t) - k_{t+1})$ in the previous

example. Γ is a set-valued function. $\Gamma(x_t)$ returns the set of feasible k_{t+1}. In the previous example, it corresponds to $\Gamma(k_t) = [0, f(k_t)] \ni k_{t+1}$. Notice that the domain of x can be any abstract spaces. Under the same notation, the corresponding *functional equation* (FE) is

$$v(x) = \sup_{y \in \Gamma(x)} F(x, y) + \beta v(y). \tag{FE}$$

Observe that in this equation, u and f are known functions, k is a dummy variable. The only unknow in the equation is the function v, hence the name "functional equation". If v is allowed to take values in the extended realine (i.e. including $\pm\infty$), then immediately $v = \pm\infty$ are two solutions. However these solutions are not sensible, and we will focus on finding real-valued solutions (that does not take values $\pm\infty$).

We say an infinite sequence $\mathbf{x}(x_0) := (x_0, x_1, \dots)$ is a *feasible sequence* from x_0 if $x_{t+1} \in \Gamma(x_t)$ for $t = 0, 1, \dots$. Given an SP, the partial sum of the first T returns of $\mathbf{x}$ is denoted by $S_T(\mathbf{x}) := \sum_{t=0}^{T} \beta^t F(x_t, x_{t+1})$.

We shall see first that if the SP is well-defined, then its maximized value is *a* solution to the FE.

Definition 3.1.1. *A sequential optimization problem is well-defined if*

1. $\Gamma(x)$ *is not empty for all* x*;*
2. *for all initial value* x_0 *and feasible* $\mathbf{x}(x_0)$, $\lim_{T\to\infty} S_T(\mathbf{x}(x_0))$ *exists (with the possibility of* $+/-\infty$*).*

If a sequential optimization problem is well-defined, we can denote $\lim_{T\to\infty} S_T(\mathbf{x}(x_0))$ by $S(\mathbf{x}(x_0))$. For each initial condition x_0, denotes the optimized value of the problem SP by $v^*(x_0) := \sup_{\mathbf{x}(x_0)} S(\mathbf{x}(x_0))$. It is apparent that if for each x_0, $v^*(x_0)$ is bounded above and below, then v^* is a real-valued function. We shall show that if the SP is well-defined and has a real-valued v^*, then v^* satisfies the FE.

Theorem 3.1.1. *If a sequential optimization problem is well-defined and its optimized value v^* is a real-valued function, then its optimized value v^* is a solution to the corresponding functional equation.*

Proof. Because v^* is real valued, $v^*(x_0)$ is finite for any x_0, this also means for any $x \in \Gamma(x_0)$, $v^*(x)$ is also finite. The definition of v^* implies $v^*(x_0) \geq F(x_0, x) + \beta S(\mathbf{x}(x))$ where $(x_0, \mathbf{x}(x))$ is any feasible sequence that starts with $(x_0, x, \dots)$.

Now we fix an $x \in \Gamma(x_0)$. Finiteness of $v^*(x)$ means for any $\epsilon > 0$, there is a feasible sequence starting from x, i.e. $\mathbf{x}(x)$, such that $v^*(x) \leq S(\mathbf{x}(x)) + \epsilon$. By definition of $v^*(x_0)$, we have

$$\begin{aligned} v^*(x_0) &\geq F(x_0, x) + \beta S(\mathbf{x}(x)) \\ &\geq F(x_0, x) + \beta v^*(x) - \beta\epsilon. \end{aligned}$$

This argument holds for every ϵ, therefore $v^*(x_0) \geq F(x_0, x) + \beta v^*(x)$. Now see that $x \in \Gamma(x_0)$ is arbitrary, it follows that

$$v^*(x_0) \geq \max_{x \in \Gamma(x_0)} F(x_0, x) + \beta v^*(x).$$

To prove the reversed inequality, finiteness of $v^*(x_0)$ means that for any $\epsilon > 0$ there is a feasible sequence $(x_0, \mathbf{x}(x))$ where $x \in \Gamma(x_0)$ such that $v^*(x_0) \leq S((x_0, \mathbf{x}(x))) + \epsilon$. Therefore

$$\begin{aligned} v^*(x_0) &\leq S((x_0, \mathbf{x}(x))) + \epsilon \\ &= F(x_0, x) + \beta S(\mathbf{x}(x)) + \epsilon \\ &\leq F(x_0, x) + \beta v^*(x) + \epsilon \\ &\leq \max_{x' \in \Gamma(x_0)} (F(x_0, x') + \beta v^*(x')) + \epsilon. \end{aligned}$$

Because ϵ is arbitrary we have shown that $v^*(x_0) = \max_{x' \in \Gamma(x_0)} (F(x_0, x') + \beta v^*(x'))$.

□

This theorem holds even if v^* is not bounded and takes the value $+/-\infty$. However the discussion for these cases are not particularly relevant for the iteration method that we are going to introduce in later sections.

We have seen that the optimized value function v^* for a SP is a solution to the corresponding FE. Now instead of solving an optimization problem, we shall first pause and consider a related problem that asks to solve the corresponding functional equation. If the solution to the functional equation is indeed the maximized value of the optimization problem, then substitute this solution into the right hand side of FE gives rise to a simple static optimization problem whose solution is the policy function to the dynamic optimization problem.

3.1.2 The Example in Infinite Horizon

Recall the functional equation for this example

$$v(k_0) = \max_{0 \le k \le f(k_0)} u(f(k_0) - k) + \beta v(k). \tag{3.1}$$

Observe that if we do know v, then we can solve the optimization problem on the right hand side of (3.1). The optimal choice of $k^* \in (0, f(k_0))$ will depends on the value of k_0. On the other hand, if we do know the relation between the optimal choice of k^* for every k_0, we can back out the value function. Although in general, there is no method for directly solving v from the functional equation, in the example where we assumed $u(c) = \ln c$ and $f(k) = k^\alpha$, we have a candidate relation between k^* and k_0. Namely, the $T \to \infty$ limit of the policy function from the finite period example, $k^* = h(k_0) = \alpha\beta k_0^\alpha$. We shall guess that this policy function is correct and try to recover a solution to the FE.

Assuming v is differentiable, first we apply the Envelop theorem and see that $v'(k) = f'(k)u'(f(k) - h(k))$. Having guessed the policy function to be $h(k) =$

$\alpha\beta k^\alpha$, we substitute in and obtain

$$v'(k) = \frac{\alpha k^{\alpha-1}}{k^\alpha - \alpha\beta k^\alpha} = \frac{\alpha}{1-\alpha\beta}\frac{1}{k}.$$

Integrate both side we derive a candidate $v(k) = \frac{\alpha}{1-\alpha\beta}\ln k + C$ for some constant C. We now just need to verify if this function is a solution. To this end, consider the optimization problem on the right hand side of (3.1) which now becomes

$$\max_{0\le k\le f(k_0)} u(f(k_0)-k) + \beta\frac{\alpha}{1-\alpha\beta}\ln k + C.$$

The FOC of this problem is

$$-\frac{1}{k_0^\alpha - k^*} + \frac{\alpha\beta}{1-\alpha\beta}\frac{1}{k^*} = 0 \Rightarrow k^* = \alpha\beta k_0^\alpha,$$

confirms our guess for the policy function. Hence the maximized value of the right hand side of (3.1) is

$$\frac{\alpha}{1-\alpha\beta}\ln k_0 + \frac{1}{1-\beta}\left(\ln(1-\alpha\beta) + \frac{\alpha\beta}{1-\alpha\beta}\ln(\alpha\beta)\right) = v(k_0).$$

This verifies that our guess is *one* solution to the functional equation (3.1) with $C = \frac{1}{1-\beta}\left(\ln(1-\alpha\beta) + \frac{\alpha\beta}{1-\alpha\beta}\ln(\alpha\beta)\right)$. However, we know neither if there are other solutions nor if our solution is the value function for the sequence problem. In the following section, we shall introduce a practical algorithm that iteratively approximates a solution arbitrarily well. After that, we shall see that the algorithm satisfies the so called "Blackwell's conditions", and approximates the unique solution of the FE. When the solution is unique, Theorem 3.1.1 implies that it has to be the value function of the SP and hence our infinite horizon example is indeed solved.

3.1.3 Iteration Algorithm

By taking the limit of the finite horizon policy function, sometimes we can obtain the policy function of the infinite period problem (see exercises). But in practice one usually cannot solve the FE analytically for general F, Γ. Fortunately, there is an algorithm that at least heuristically approaches the solution very accurately.

We start with a random guess of the policy function $k_{t+1} = h_0(k_t)$. It could be anything as long as it is feasible. I.e. $0 \leq h_0(k) \leq f(k)$. Suppose for every time t, one always apply this policy function even though we know that it is unlikely to be optimal. This gives rise to a value function for each amount of initial capital k_0.

$$v_0(k_0) := \sum_{t=0}^{\infty} \beta^t u(f(k_t) - k_{t+1}).$$

If the infinite series converges, one can immediately verify that

$$v_0(k) = u(f(k) - g_0(k)) + \beta v_0(g_0(k)).$$

Consider that if we commit to blindly applying h_0 only for period $t \geq 1$. But at period 0 we try to choose the optimal k_1 given that we will follow h_0 afterwards. The problem becomes

$$\max_{k_1} u(f(k_0) - k_1) + \beta v_0(k_1)$$

for any given k_0. Since v_0 is known, this maximization problem can be solve for any initial k_0. Denote the optimal policy for choosing k_1 by $h_1(k_0)$. Using this policy gives us a better value function

$$v_1(k_0) := u(f(k_0) - h_1(k_0)) + \beta v_0(h_1(k_0))$$

in the sense that $v_1 \geq v_0$ (why?).

Now consider if we commit to blindly applying h_0 only for period $t \geq 2$, but we try to make the optimal decision at $t = 0, 1$ given that the decision will be made

through h_0 afterwards. In period 1 the optimization problem will be the same as described above. So if one chooses any k_1 at time 0, the value from period 1 and on will be given by $v_1(k_1)$. At $t = 0$, the problem becomes

$$\max_{k_1} u(f(k_0) - k_1) + \beta v_1(k_1)$$

for any exogeneous initial capital k_0. Since v_1 is known, this maximization problem can be solve for any initial k_0. Denote the optimal policy for choosing k_1 in this problem by $h_2(k_0)$. Using this policy gives us another value function

$$v_2(k_0) := u(f(k_0) - h_2(k_0)) + \beta v_1(h_2(k_0))$$

Similarly, it can be seen that $v_2 \geq v_1$.

The above process can be iterated for an arbitrary number of times. If we delay applying h_0 to period t and on, the optimal sequence would be to apply h_1 at period $t-1$, h_2 at period $t-2$ etc... And the total value of applying this sequence of policy functions until t and then adopt h_0 forever is denoted by $v_t(k_0)$. Ideally, if v^* is the solution to the functional equation, it should be true $v^* \geq v_t$ for any t, because v^* is the discounted utility when every decision is done optimally. Since $v_0 \leq v_1 \leq v_2 \ldots$ This sequence of functions is monotonically increasing and v_t and can be shown to approximate v^* well as $t \to \infty$ in many practical problems. It turns out the algorithm not only converges to v^* with a prespecified h_0, it converges even with an arbitrary initial guess of v_0 (so it is unnecessary to specify h_0 to begin with). Moreover, the convergence is "exponentially fast". We will show in the next section that the iterative approximation method indeed approximates the *unique* solution of the functional equation, and thus, the value function to the SP. This method of iteratively approximating the value function of the infinite horizon problem is called *dynamic programming*.

3.1.4 Exercises

Exercise 3.1.1. *Prove the following "almost-converse" to Theorem 3.1.1.*

Suppose a SP is well-defined and has a real-valued v^* and there is a real-valued v that solves the corresponding FE. If it is true that given any feasible sequence $(x_0, x_1, \dots) = \mathbf{x}$ with any initial condition x_0, $\lim_{T\to\infty} \beta^T v(x_T) = 0$, then $v = v^*$.

Exercise 3.1.2. *Use the finite horizon limit to guess the value function to the infinite horizon example with $f(k) = \alpha k$ and $u(c) = c^\rho/\rho$ for $\alpha > 0$, $\rho < 1$. Under what parameter restrictions is it a sensible solution to the corresponding FE?*

Exercise 3.1.3. *Solve the infinite horizon problem in the example with $f(k) = k^\alpha$ and $u(c) = \rho c$ for $\alpha \in (0, 1)$, $\rho > 0$. Which constraints are binding?*

3.2 Complete Spaces and Contraction Mappings*

3.2.1 Complete Metric Spaces

We have seen vectors in $\mathbb{R}^n$ as points, but for optimization problems it is usually convenient to view functions as points (or vectors) as well. In a space of points, the geometry is given by the distances between points. A *metric*, as defined below, is a measure of distance between points (functions) in a space.

Definition 3.2.1. *A metric on a space $\mathbb{X}$ is a real-valued function $d(.,.) : \mathbb{X}^2 \to \mathbb{R}$ satisfying the following properties:*

- *$d(x, y) \in [0, \infty)$ for all $x, y \in \mathbb{X}$ and $d(x, y) = 0$ iff $x = y$ (non-negativity);*
- *For any $x, y \in \mathbb{X}$, $d(x, y) = d(y, x)$ (symmetry);*
- *$d(x, z) \leq d(x, y) + d(y, z)$ for any $x, y, z \in \mathbb{X}$ (triangle inequality).*

A *metric space* is a set of points $\mathbb{X}$ together with a metric $d(.,.)$. When the space is $\mathbb{R}^n$, where $x := (x_1, \dots, x_n)$ there are many well-defined metrics such as

- the Euclidean metric: $d_{Euclid}(x, x') := \sqrt{\sum_{i=1}^{n}(x_i - x_i')^2}$;
- the chessboard metric (the maximum norm) $d_{\max}(x, x') := \max_i |x_i - x_i'|$.

These metrics usually have analogues in functional spaces. Let $\mathbb{R}_+$ be the positive reals, and let $\mathbb{F}(\mathbb{R}_+)$ be the space of functions on the positive reals. We can define similar distances for these spaces

- the Euclidean metric: $d_{Euclid}(f, g) := \sqrt{\int_{\mathbb{R}_+} |f(x) - g(x)|^2 \mathrm{d}x}$;
- the sup-norm $d_{\sup}(f, g) := \sup_{\mathbb{R}_+} |f(x) - g(x)|$.

For many families of continuous functions (for example, the family of bounded continuous functions), the above definition do form well-defined metrics. When we have a well-defined metric space, we can start to talk about concepts such as convergence and limits.

Definition 3.2.2. *Let $(\mathbb{X}, d)$ be a metric space. A sequence $\{x_n\}_{n\in\mathbb{N}}$ in $\mathbb{X}$ converges to $x \in \mathbb{X}$, denoted by $x_n \to x$ as $n \to \infty$, if for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $d(x_n, x) < \epsilon$ whenever $n > N$.*

Intuitively, convergence means the tail of the sequence gets closer and closer to a limit point in the space. This definition has to be carefully distinguished from the definition of a *Cauchy sequence*.

Definition 3.2.3. *Let $(\mathbb{X}, d)$ be a metric space. A sequence $\{x_n\}_{n\in\mathbb{N}}$ in $\mathbb{X}$ is called Cauchy if for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $d(x_m, x_n) < \epsilon$ whenever both $m, n > N$*

These two definitions are not equivalent. The key distinction is that a Cauchy sequence need not converge to a limit inside the space (examples ?). In otherwords, a Cauchy sequence need not be convergent, but a convergent sequence must be Cauchy (why?). As mathematicians have discovered, that a metric space can be difficult to work with if its Cauchy sequences may not converge. To distinguish the "good" spaces from these "bad" ones, they coined the term *completeness*.

Definition 3.2.4. *A metric space is complete if every Cauchy sequence converges.*

3.2.2 The Contraction Mapping Theorem

Let $(\mathbb{X}, d)$ be a complete metric space, a contraction mapping is a transformation $T : \mathbb{X} \to \mathbb{X}$, that shrinks the distance between any two points with a given factor. The contraction mapping theorem states that such a transformation has a unique fixed point as the limit of all other points under iteration of T. This theorem is of critical importance in the course of dynamic programming, because in many models, the functional equation can be interpreted as contraction mapping on the space of (potential value) functions. As a consequence, the value function can be well-approximated by iteration of T, even if its analytic expression is not available.

Definition 3.2.5. *Let $(\mathbb{X}, d)$ be a complete metric space. A mapping $T : \mathbb{X} \to \mathbb{X}$ is a contraction with modulus $\beta \in (0, 1)$ if for every $x, y \in \mathbb{X}$, $d(Tx, Ty) \leq \beta d(x, y)$.*

A fixed point of a function T is just a solution to the equation $T(x^*) = x^*$. As stated in the theorem below, a contraction has a unique fixed point. Moreover, the proof of the theorem is constructive, and the unique fixed point x^* is given by $x^* = \lim_{n\to\infty} T^n(x)$ for any x in the space.

Theorem 3.2.1. *If T is a contraction on the complete metric space $(\mathbb{X}, d)$, then T has a unique fixed point in $\mathbb{X}$.*

Proof. Let x_0 be any point in $\mathbb{X}$. We define the sequence $\{x_n\}_{n\in\mathbb{N}}$ iteratively by $x_n := T^n(x_0) := T(x_{n-1})$. For any m, n that $m \geq n > N$, the distance between x_n

and x_m is bounded above,

$$\begin{aligned} d(x_n, x_m) &= d(T^n x_0, T^m x_0) \\ &\leq \beta^n d(x_0, T^{m-n} x_0) \\ &\leq \beta^n \left[d(x_0, Tx_0) + d(Tx_0, T^2 x_0) + \cdots + d(T^{m-n-1} x_0, T^{m-n} x_0) \right] \\ &\leq \beta^n \sum_{i=0}^{m-n-1} \beta^i d(x_0, x_1) \\ &\leq \frac{\beta^n}{1-\beta} d(x_0, x_1) < \frac{\beta^N}{1-\beta} d(x_0, x_1). \end{aligned}$$

Therefore the sequence $\{x_n\}_{n \in \mathbb{N}}$ is Cauchy, and hence has a limit in the complete metric space. Because T is continuous (why?), it follows from continuity that

$$x^* := \lim_{n \to \infty} x_{n+1} = \lim_{n \to \infty} T x_n = T \lim_{n \to \infty} x_n = T(x^*)$$

is a fixed point of T.

To see the fixed point is unique, let both $x, y \in \mathbb{X}$ be fixed points. Because $0 \leq d(x, y) = d(Tx, Ty) \leq \beta d(x, y)$, it must be the case that $x = y$. □

For our purpose, the above theorem is mainly used before applying dynamic programming to solve for the value function. Recall our algorithm from the previous section that

$$v_{t+1}(x) = \sup_{y \in \Gamma(x)} F(x, y) + \beta v_t(y).$$

Iteratively applying this algorithm gives us a sequence of functions v_t for $t = 0, 1, \ldots$. To apply the contraction mapping theorem in this motivating problem, we can interpret each function v_t as a point in some function space $\mathbb{F}$, and interpret the supremum operator as the transformation $T : \mathbb{F} \to \mathbb{F}$. I.e.,

$$v_{t+1}(x) = (T v_t)(x) := \sup_{y \in \Gamma(x)} F(x, y) + \beta v_t(y).$$

When T is a contraction on $\mathbb{F}$, then we know that there is a unique solution to the FE, and it can be approximated by the algorithm of dynamic programming. If T is not a contracting mapping, then the algorithm may not converge and even if it does, the limit may not be the value function.

In general, showing T is a contraction on $\mathbb{F}$ can be tricky. In our motivating problem above, whether T is a contraction depends on F, Γ and β. In general, it is often convenient to check instead the *Blackwell's sufficient conditions* that imply a contraction.

Before formally stating the conditions, we need to first define an appropriate metric space $(\mathbb{F}, d)$. Let $\mathbb{D}$ be the domain of our state variables (i.e. the x's and y's over which the supremum is taken). For any two real-valued functions f and g on $\mathbb{D}$, we define the distance to be

$$d(f, g) := \sup_{x \in \mathbb{D}} |f(x) - g(x)| .$$

Suppose $\phi : \mathbb{D} \to \mathbb{R}$ be a reference function. Let $\mathbb{F} := \mathbb{F}(\phi)$ be a space of all functions on $\mathbb{D}$ such that $d(f, \phi) < \infty$ for every $f \in \mathbb{F}$. It is left as an exercise that this is a well-defined metric and the corresponding metric space is complete.

Theorem 3.2.2. *For any complete metric space of functions $(\mathbb{F}(\phi), d)$ with $d(.,.)$ given as above. Let T be a well-defined self-map on $\mathbb{F}$. Then T is a contraction with modulus β and has a unique fixed point in $\mathbb{F}$ if there exists $\beta \in (0, 1)$ such that for any $f, g \in \mathbb{F}$ and any $c \in \mathbb{R}$,*

- *$f \leq g$ implies $Tf \leq Tg$ (monotonicity);*
- *$T(f + c) \leq Tf + \beta c$ (discounting).*

If in addition to the above conditions, $T(\phi + c) \in \mathbb{F}$ for every $c \in \mathbb{R}$, then T is guarenteed to be well-defined.

Proof. By definition, $d(f,g) \in [0,\infty)$. Observe that $f \leq g + |f-g| \leq g + d(f,g)$. By monotonicity and discounting,

$$Tf \leq T\left(g + d(f,g)\right) \leq Tg + \beta d(f,g) \Rightarrow Tf - Tg \leq \beta d(f,g).$$

Similarly, $Tg - Tf \leq \beta d(f,g)$. Hence

$$d(Tf,Tg) = \sup_{\mathbb{D}} |Tf - Tg| \leq \beta d(f,g).$$

Hence we have shown T is a contraction with modulus β. Since we have implicitly assumed that $T : \mathbb{F} \to \mathbb{F}$ is well-defined, we conclude that T is a contraction on $\mathbb{F}$. The unique fixed point follows from the contraction mapping theorem.

To show T is well defined, let $f \in \mathbb{F}$. By definition of $\mathbb{F}$, there exists some large enough constant c such that

$$\phi - c \leq f \leq \phi + c.$$

Monotonicity implies $T(\phi - c) \leq Tf \leq T(\phi + c)$. Since both the $T(\phi - c)$ and $T(\phi + c)$ are in $\mathbb{F}$ by assumption, it follows that $Tf \in \mathbb{F}$. □

3.2.3 The Example Revisited

In this section, we will apply the contraction mapping theorem to our infinite horizon example with $u(c) = \ln c$ and $f(k) = k^\alpha$ for some $\alpha \in (0,1)$. By cheching the Blackwell's conditions, the dynamic programming method is shown to be a contraction and hence the FE has a unique solution. Therefore the value function obtained in the previous section is indeed this unique solution to the FE, and hence it is also the solution to the SP by Theorem 3.1.1.

We will first provide some bounds to the value function. An upper bound can

be easily obtained by

$$\begin{aligned}\max_{k_{t+1}\in[0,k_t^\alpha]}\sum_{t=0}^{\infty}\beta^t\ln(k_t^\alpha-k_{t+1}) &\le \sum_{t=0}^{\infty}\beta^t\ln(\bar{k}_t^\alpha) \quad \text{(where } \bar{k}_{t+1}:=\bar{k}_t^\alpha \text{ and } \bar{k}_0:=k_0\text{)}\\ &= \sum_{t=0}^{\infty}\beta^t\alpha^t\ln k_0^\alpha\\ &= \frac{\alpha}{1-\alpha\beta}\ln k_0.\end{aligned}$$

To obtain a lower bound,we choose the policy function $h(k) := \frac{1}{2}k^\alpha$. Then the sequence of capital becomes $k_0, \frac{1}{2}k_0^\alpha, \frac{1}{2}(\frac{1}{2}k_0^\alpha)^\alpha, \ldots$. Let $\underline{k}_t := \left(\frac{1}{2}\right)^{-\alpha^{-1}+\sum_{i=0}^t\alpha^{i-1}} k_0^{\alpha^t}$

$$\begin{aligned}\max_{k_{t+1}\in[0,k_t^\alpha]}\sum_{t=0}^{\infty}\beta^t\ln(k_t^\alpha-k_{t+1}) &\ge \sum_{t=0}^{\infty}\beta^t\ln\left(\frac{1}{2}\underline{k}_t^\alpha\right)\\ &= \sum_{t=0}^{\infty}\beta^t\ln\left(\frac{1}{2}\left(\frac{1}{2}\right)^{-1+\sum_{i=0}^t\alpha^i}k_0^{\alpha^{t+1}}\right)\\ &= \sum_{t=0}^{\infty}\beta^t\ln\left(\frac{1}{2}\right)^{\sum_{i=0}^t\alpha^i}+\sum_{t=0}^{\infty}\beta^t\ln\left(k_0^{\alpha^{t+1}}\right)\\ &= C+\frac{\alpha}{1-\alpha\beta}\ln k_0,\end{aligned}$$

where C is some constant not depending on k_0 (it is easy to check that C is negative).Therefore, the value function $v^*(k)$ to the SP is bounded by

$$-|C|+\frac{\alpha}{1-\alpha\beta}\ln k \le v^*(k) \le |C|+\frac{\alpha}{1-\alpha\beta}\ln k.$$

Let $\phi(k) := \frac{\alpha}{1-\alpha\beta}\ln k$ be our reference function. As before, we define the distance between two functions to be $d(f,g) := \sup|f-g|$. Let $\mathbb{F}$ be the space of all real-valued functions on $(0,\infty)$ that has finite distance to ϕ, i.e. $d(f,\phi)<\infty$ for all $f\in\mathbb{F}$. We have seen that this is a complete metric space.

To see that the solution to the FE is unique, we need to first check the Blackwell

conditions. As before, for any $f \in \mathbb{F}$, let

$$Tf := \sup_{y \in (0, x^\alpha]} \ln(x^\alpha - y) + \beta f(y).$$

It is left as an exercise to check it is well-defined, i.e. $d(T\phi, \phi) := |T\phi - \phi| < \infty$. The check the Blackwell conditions, let $f \leq g$. Monotonicity follows from

$$\begin{aligned} Tf &:= \sup_{y \in (0, x^\alpha]} \ln(x^\alpha - y) + \beta f(y) \\ &\leq \sup_{y \in (0, x^\alpha]} \ln(x^\alpha - y) + \beta g(y) =: Tg. \end{aligned}$$

Discounting follows from

$$\begin{aligned} T(f + c) &:= \sup_{y \in (0, x^\alpha]} \ln(x^\alpha - y) + \beta(f(y) + c) \\ &\leq \left(\sup_{y \in (0, x^\alpha]} \ln(x^\alpha - y) + \beta f(y) \right) + \beta c = Tf + \beta c. \end{aligned}$$

Therefore, we have checked the Blackwell's sufficient conditions and shown that T is indeed a contraction in the space $(\mathbb{F}(\phi), d)$ and has a unique fixed point. We have already seen in the previous section that

$$v^*(k) := \frac{\alpha}{1 - \alpha\beta} \ln k + \frac{1}{1 - \beta} \left(\ln(1 - \alpha\beta) + \frac{\alpha\beta}{1 - \alpha\beta} \ln(\alpha\beta) \right)$$

is a fixed point of T (and hence now the unique one). We conclude by Theorem 3.1.1 that it is exactly the solution to the SP.

3.2.4 Exercises

Exercise 3.2.1. *Show that any convergent sequence must be Cauchy. And give an example where a Cauchy sequence is not convergent.*

Exercise 3.2.2. *Show that a contraction function is continuous.*

Exercise 3.2.3. *Suppose $\mathbb{X} \subseteq \mathbb{R}^n$. A weak contraction on $\mathbb{X}$ is a function $f : \mathbb{X} \to \mathbb{X}$ such that $d(f(x), f(y)) < d(x, y)$ for any $x, y \in \mathbb{X}$. Show that if $\mathbb{X}$ is closed and bounded then there exist a unique fixed point of f and it is given by $\lim_{n\to\infty} f^n(x)$ for any $x \in \mathbb{X}$.*

Exercise 3.2.4. *Let $\mathbb{F}$ be the space of bounded continuous functions on $\mathbb{R}$. Show that $(\mathbb{F}, d_{\sup})$ is a complete metric space.*

Exercise 3.2.5. *Fix any reference function ϕ and let $d(.,.)$ be the sup-metric. Let $\mathbb{F}(\phi)$ be the space of all continuous functions on $\mathbb{D}$ such that $d(\phi, f) := \sup_{\mathbb{D}} |\phi - f| < \infty$ for every $f \in \mathbb{F}$. Show that $(\mathbb{F}, d)$ is a complete metric space .*

Exercise 3.2.6. *Solve the maximization problem $\sup_{y\in(0,x^\alpha]} \ln(x^\alpha - y) + \beta\phi(y)$ where $\phi(y) := \frac{\alpha}{1-\alpha\beta} \ln y$. Show that the supremum differs from ϕ by at most a constant.*

Exercise 3.2.7. *Suppose $F(.,.)$ is a bounded function (i.e. $|F| < M < \infty$ for some fixed M). Show that The iteration algorithm converges to the value function $v(x_0) := \sup \sum_{t=0}^{\infty} \beta^t F(x_t, x_{t+1})$ for all x_0 (without using the contraction mapping theorem). What type of functions should be your first guess?*

Exercise 3.2.8. *The method in the above exercise can also be applied in the example ($u(c) = \ln c$, and $f(k) = k^\alpha$). Without using the contraction mapping theorem, show that the iteration algorithm converges to the value function. What type of functions should be your first guess?*

3.3 Miscellaneous*

3.3.1 The Euler Equation Approach

An alternative way to approach the SP directly is by what is called the Euler Equation approach. This approach eventually transforms the problem into a

system of difference equations. Then one can apply techniques from the difference equations to analyse the property of the optimal policy function.

Consider the sequential optimization problem again

$$\begin{aligned} &\sup_{x_1,x_2,\ldots} \sum_{t=0}^{\infty} \beta^t F(x_t, x_{t+1}) \\ \text{subject to: } &x_{t+1} \in [0, f(x_t)] \qquad \text{for } t = 0, 1, \ldots \\ &x_0 \text{ is given.} \end{aligned}$$

Suppose the problem has almost been solved and that the optimal choice for $x_1^*, \ldots, x_t^*$ and $x_{t+2}^*, x_{t+3}^* \ldots$ are all known. We only need to look for the last unknown x_{t+1}^*. Due to the additive separability of the objective function, the problem boils down to solving a two period problem

$$\begin{aligned} &\sup_y F(x_t^*, y) + \beta F(y, x_{t+2}^*) \\ \text{subject to: } &y \in [0, f(x_t^*)] \\ &x_{t+2}^* \in [0, f(y)]. \end{aligned}$$

Suppose that F is strictly concave, increasing in the first argument and decreasing in the second. If we have enough concavity that guarentees an interior solution, we can write the FOC of the two period problem directly and set to 0. I.e.

$$F_2(x_t^*, x_{t+1}^*) + \beta F_1(x_{t+1}^*, x_{t+2}^*) = 0 \text{ for } t = 0, 1, \ldots$$

because the same arguments hold for all time t. When F has given analytical form, the above set of equations is a system of 2nd order difference equation whose initial condition x_0 is given. In general, to solve a 2nd order difference equation, one needs two boundary conditions, the initial and the ending conditions. For our purpose,

a sufficient boundary condition at $t \to \infty$ is

$$\lim_{t\to\infty} \beta^t F_1(x_t^*, x_{t+1}^*)^{\mathrm{T}} x_t^* = 0.$$

This limiting condition is called the *transversality condition* (TVC). Notice that the transpose indicates the first argument of F could be a vector. In fact, all the theorems and proofs in this section works even if x is a multivariate vector.

Theorem 3.3.1. *Suppose a well-defined SP is given by a concave and smooth F return function and $\Gamma(x) = [0, f(x)]$ for some non-negative function f. Suppose also that for an initial value x_0 the SP has finite maximized values $v^*(x_0)$. If a feasible sequence $\mathbf{x}^* = (x_0^*, x_1^*, \dots)$ satisfies $x_0^* = x_0$, the Euler equation*

$$F_2(x_t^*, x_{t+1}^*) + \beta F_1(x_{t+1}^*, x_{t+2}^*) = 0 \text{ for } t = 0, 1, \dots;$$

and the transversality condition

$$\lim_{t\to\infty} \beta^t F_1(x_t^*, x_{t+1}^*)^{\mathrm{T}} x_t^* = 0,$$

then $v^*(x_0) = \lim_{T\to\infty} \sum_{t=0}^{T} \beta^t F(x_t^*, x_{t+1}^*)$.

Proof. Observe that it is sufficient to show that for any feasible $\mathrm{x} = (x_0, x_1, \dots)$,

$\lim_{T\to\infty}\sum_{t=0}^{T}\beta^t F(x_t^*,x_{t+1}^*)\geq\lim_{T\to\infty}\sum_{t=0}^{T}\beta^t F(x_t,x_{t+1})$. This is so because

$$\begin{aligned}
&\lim_{T\to\infty}\sum_{t=0}^{T}\beta^t\left(F(x_t^*,x_{t+1}^*)-F(x_t,x_{t+1})\right)\\
\geq&\lim_{T\to\infty}\sum_{t=0}^{T}\beta^t\left(F(x_t^*,x_{t+1}^*)-\left[F(x_t^*,x_{t+1}^*)+(x_t-x_t^*)F_1(x_t^*,x_{t+1}^*)+(x_{t+1}-x_{t+1}^*)F_2(x_t^*,x_{t+1}^*)\right]\right)\\
=&\lim_{T\to\infty}\sum_{t=0}^{T}\beta^t\left[(x_t^*-x_t)F_1(x_t^*,x_{t+1}^*)+(x_{t+1}^*-x_{t+1})F_2(x_t^*,x_{t+1}^*)\right]\\
=&\lim_{T\to\infty}\left[\left(\sum_{t=0}^{T-1}\beta^t(x_{t+1}^*-x_{t+1})\left[F_2(x_t^*,x_{t+1}^*)+\beta F_1(x_t^*,x_{t+1}^*)\right]\right)+\beta^T(x_{T+1}^*-x_{T+1})F_2(x_T^*,x_{T+1}^*)\right]\\
=&\lim_{T\to\infty}-\beta^{T+1}(x_{T+1}^*-x_{T+1})F_1(x_{T+1}^*,x_{T+2}^*)\\
\geq&\lim_{T\to\infty}-\beta^{T+1}x_{T+1}^*F_1(x_{T+1}^*,x_{T+2}^*)=0.
\end{aligned}$$

From the first line to the second, we applied the linear approximation to $F(x_t,x_{t+1})$ and the inequality is due to concavity of F. From third to fourth, we used the fact that $x_0^*=x_0$ and regrouped the sum. We applied the Euler equation to obtain the fifth line, and because $F_1>0$ and $x_{T+1}\geq 0$ due to feasibility, we obtains the sixth in which the TVC is imposed. □

One notices that in the prove, the role of TVC is to ensure that

$$\lim_{T\to\infty}\beta^T(x_{T+1}^*-x_{T+1})F_2(x_T^*,x_{T+1}^*)\geq 0.$$

Since $F_2\leq 0$, we can interpret the limit as that TVC ensures $\beta^T x_{T+1}\geq\beta^T x_{T+1}^*$ in the limit. Recall from the finite period model, leaving unused resource to $T+1$ period is wasteful. The above inequality just says that any feasible sequence will "waste" more resource than the optimal one asymptotically.

3.3.2 One-Shot Deviation Principle

The sequential optimization problem we have been considering has the property that the period by period return function stays unchanged. However, in the context of our previous example, the farmer planning for future consumption, he can probably foresee that the utility he gets from a certain amount of consumption changes as he grows old. In many applications, it would be sensible to allow a time-varing return function.

More generally, we should also allow the past sequence affect future return functions. Take purchasing auto-insurance as an example. The insurance premium depends on whether one has purchased auto-insurance before, how long has one had auto-insurance for, and whether one has had accidents in the past. If a decision maker wants to maximizes some sort of objective function, she will need to take into consideration how would past histories affect future period by period return function.

The above problem can be modelled by a single-person decision tree with the return function depending on the history alone (why?) (pic). Consider an infinite decision tree. A feasible sequence $\mathbf{x} = (x_0, x_1, \dots)$ is a path along the tree whose root is x_0. A history up till period t is denoted by $\mathbf{x}_t = (x_0, x_1, \dots, x_t)$. For each history $\mathbf{x}_t$, the decision tree branches out to a set of nodes $\Gamma(\mathbf{x}_t) \ni x_{t+1}$. The return function depends on the history. In other words, the decision maker at $\mathbf{x}_t$ and chooses $x_{t+1} \in \Gamma(\mathbf{x}_t)$ receives utility $F(\mathbf{x}_t, x_{t+1})$. Like before, the objective is to maximize discounted utility

$$\sup_{\mathbf{x}=(x_0,x_1,\dots)} \sum_{t=1}^{\infty} \beta^t F(\mathbf{x}_t, x_{t+1})$$

for some $\beta \in (0, 1)$. The regularity condition we impose is that the period by period return is bounded. I.e. $|F| \leq M$ for some large number M.

We have generalized the sequential optimization problem to a large extent but

maybe the most intuitive way to think of such a problem is still a sort of backwards induction. Suppose all the future decision problem has been solved at the history $\mathbf{x}_T$, and one needs to make one decision to choose $x_{T+1} \in \Gamma(\mathbf{x}_T)$. Mathematically,

$$\max_{x_{T+1}} \sum_{t=1}^{T} \beta^t F(\mathbf{x}_t, x_{t+1}) + v(\mathbf{x}_T, x_{T+1})$$
$$\text{subject to: } x_{T+1} \in \Gamma(\mathbf{x}_T).$$

The value function $v(\mathbf{x}_T, x_{T+1})$ represent the maximized infinite discounted sum of utility with initial history $(\mathbf{x}_T, x_{T+1})$. This is a simple two-period maximization problem and can often be solved easily. Denotes the optimal choice to this problem by $h(\mathbf{x}_t)$. Because the above set-up is valid for every feasible sequence of finite length, any policy function of this problem is simply a function that assigns to every history $\mathbf{x}_t$ a choice in $\Gamma(\mathbf{x}_t)$. Namely, $h(\mathbf{x}_t) \in \Gamma(\mathbf{x}_t)$. A policy function is also called a *strategy* in the game theory literature.

Analytically finding a policy function can be difficult. However, given a candidate policy function, it is not too difficult to verify if it is an optimal one. A sequence infinite sequence given by some initial history $\mathbf{x}_t$ and following a feasible policy function afterwards is denoted by $\mathbf{x}(\mathbf{x}_t; h) := (\mathbf{x}_t, h(\mathbf{x}_t), h(h(\mathbf{x}_t)), \dots)$The definition of optimality is the similar as before.

Definition 3.3.1. *A feasible policy function h^* is optimal if for any other policy function h and every finite period history $\mathbf{x}_t$,*

$$S(\mathbf{x}(\mathbf{x}_t; h^*)) := \sum_{s=0}^{\infty} \beta^{t+s} F(\mathbf{x}_{t+s}, h^*(\mathbf{x}_{t+s})) \geq \sum_{s=0}^{\infty} \beta^{t+s} F(\mathbf{x}_{s+t}, h(\mathbf{x}_{s+t})) =: S(\mathbf{x}(\mathbf{x}_t; h)).$$

It worthes to emphasis that when the optimization problem is history dependent, optimality of h^* means the policy function is optimal even at histories that would not be reached by h^*, and that given such a history is reached, h^* renders higher total payoff than any other policy functions. (For people familiar with sequential

games, this is similar to the concept of subgame perfection.) Therefore if a feasible policy function h is not optimal, it means there are some histories $\mathbf{x}_t$ (reached by h or not) following which one can deviate from h in the subsequencial optimization problem and obtains higher total payoffs given $\mathbf{x}_t$. Intuitively as one may guess, sometimes a simutaneous deviation from h at multiple histories may be necessary to improve upon h. (pic) Yet this intuition fails in this type of setings. We will see that if a policy function h is not optimal, then there is a history at which we need only deviate once and follows h afterwards to make an improvement. This is the so called *one-shot deviation principle*. As a result, the following definition is is properly named even though it is defining for the nonexistence of "one-shot improvement".

Definition 3.3.2. *A policy function h is called* unimprovable *if there is no feasible finite history $\mathbf{x}_t$ and $x_{t+1} \in \Gamma(\mathbf{x}_t)$ such that*

$$S(\mathbf{x}((\mathbf{x}_t, x_{t+1}); h)) > S(\mathbf{x}(\mathbf{x}_t; h)).$$

We are now ready to prove the one-shot deviation principle. Essentially the reason is because "the tail" of a problem matters so little, if one can improve upon a policy function, the improvement can be made in finite time. And because of the additive form of the objective function, this implies some improvement must have been made at a particular history, i.e. the time of the one-shot deviation.

Theorem 3.3.2. *If F is bounded, then a feasible policy function to a well-defined SP is optimal if and only if it is unimprovable.*

Proof. It is straight forward to see optimality implies unimprovability. The following proves the converse.

Suppose a feasible h is not optimal at some history $\mathbf{x}_t$, then this means there

exists some feasible sequence $\hat{\mathbf{x}} = (\mathbf{x}_t, x_{t+1}, x_{t+2} \dots)$ starting with $\mathbf{x}_t$ such that

$$S(\mathbf{x}(\mathbf{x}_t; h)) + 2\epsilon < S(\hat{\mathbf{x}})$$

for some $\epsilon > 0$. Because F is bounded and $\beta \in (0, 1)$, this means there is some $T > t$ such that $\sum_{s=T}^{\infty} \beta^s |F| < \epsilon/2$. This means

$$S(\mathbf{x}(\mathbf{x}_t; h)) + 2\epsilon < S(\hat{\mathbf{x}}) \leq S(\mathbf{x}(\hat{\mathbf{x}}_T; h)) + \epsilon.$$

In other words $S(\mathbf{x}(\mathbf{x}_t; h)) + \epsilon < S(\mathbf{x}(\hat{\mathbf{x}}_T; h))$. If $S(\mathbf{x}(\hat{\mathbf{x}}_{T-1}; h)) < S(\mathbf{x}(\hat{\mathbf{x}}_T; h))$, then there is a profitable one-shot deviation from h at the history $\hat{\mathbf{x}}_{T-1}$. Otherwise, we have $S(\mathbf{x}(\mathbf{x}_t; h)) < S(\mathbf{x}(\hat{\mathbf{x}}_{T-1}; h))$, and we can compare $S(\mathbf{x}(\hat{\mathbf{x}}_{T-2}; h))$ and $S(\mathbf{x}(\hat{\mathbf{x}}_{T-1}; h))$. It is clear that this argument can be applied repeatedly and it has to stop before we reach $t < T$. When it stops, a profitable one-shot deviation is found. □

3.3.3 Kelly's Strategy: An Alternative Objective Function

The previous analysis heavily relie on the discounted additively separable objective function. However, sometimes it may be reasonable to consider an alternative objective function.

Consider the case where a professional gambler enters into a casino. She is equipted with a specific device that helps her to count the cards on the table and calculates the remaining deck. With the device her expected value of betting is positive. She has an initial amount of money s_0 in her pocket and is not allowed to borrow from the bank or anyone else. Her has time to stay in the casino for as long as she like and loosely speaking, the goal is to earn as much as possible. Is there some sort of "best strategy" in this situatioin? And if there is, in what sense is such a strategy the best strategy (i.e. what is the objective function)?

To simplified the analysis, let's suppose one can bet any positive amout of

money in a given round. And there are only two outcomes with a fixed probability of winning in each round. To put it differently, suppose there is a sequence of i.i.d gambles and each has 2 outcomes. When it wins, it pays in total $d+1$ dollars for each dollar bet, and when it loses it pays nothing. The probability of winning is p and the probability of losing is $1-p$ that $p(d+1) > 1$. And the initial endownment is s_0 dollars in her pocket.

It might be a little surprising to realize that there is a betting strategy that can win more money than any other (essentially different) strategies with probability one as the number of games goes to infinity. This strategy is that we are going to introduce is called *Kelly's strategy.*

Observe that if there is an optimal strategy the strategy should be betting a fixed certain fraction of the bankroll, since the strategy should be the same regardless of the unit in which one calculates her bankroll (why?). Suppose that one spends f fraction of the bankroll each time she gambles. After n games, one wins w number of games and loses l times, the bankroll after n games is

$$s_n = s_0(1-f)^l(1+fd)^w$$

For each n, taking log and divide by n is a monotonic transformation. Through this we obtain the average growth rate of the bankroll $\frac{1}{n}\ln(s_n/s)$. Observe that although the bankroll is random, the average growth rate converges to a constant for large n.

$$\frac{1}{n}\ln(s_n/s) = \frac{l}{n}\ln(1-f) + \frac{w}{n}\ln(1+fd) \to (1-p)\ln(1-f) + p\ln(1+fd).$$

It is optimized when

$$pd(1-f) = (1-p)(1+fd) \Rightarrow f = \frac{p(d+1)-1}{d}.$$

This is the Kelly's strategy when there are two outcomes. The objective function for this analysis, as we have seen, is the asymptotic growth rate of the bankroll. Under such an objective function, the stochastic element of this problem disappear in the limit. The strong law of large number says any other essentially different strategy would result in a less amount of winnings with overwhelming probability. In the language of risk theory, we could use the first order stochastic dominance in comparing any two strategies and Kelly's strategies is the optimal in this sense.

Idealy, with such a strategy one's wealth never reaches zero because there is no limitation on a minimum amount of bet. In reality where there is a minimum amount of bet, if s_0 is large enough, the probability of hitting zero diminishes very quickly.

3.3.4 Exercises

Exercise 3.3.1. *Prove that when the period by period return function does not depend on the entire history, but only the choice on the last period (sometimes called the state variable), then the one-shot deviation principle can be strengthened to a first-step deviation principle.*

Exercise 3.3.2. *In Kelly's strategy we only consider betting a fixed fraction of the bankroll. Why do we not need to consider other strategies where we place a different fraction when the bankroll is different?*

Exercise 3.3.3. *Show that Kelly's strategy first order stochastically dominates any other (essentially different) strategy.*

Exercise 3.3.4. *Consider a sequential optimization problem where an individual wants to maximize (expected) discounted sum of infinite stream of utility* $\sum_{t=0}^{\infty} \beta^t \ln(c_t)$ *where* ln *is the utility function and* c_t *is the amount of dollar she spends in consumption at* t*. The individual starts with initial amount of saving* s_0 *and she can save to or borrow from the bank with a constant period-by-period interest rate*

r. In each period, there is an investment opportunity available and she can invest any amount of money i_t into it. At each $t+1$ the project successes with probability p and she receives $d+1$ dollars for each dollar invested. When it fails with probability $1-p$ she receives nothing back. The probability of success is p is constant over time. In other words, her net wealth, denoted by $s_t :=$ saving $-$ borrowing, follows the law of motion

$$s_{t+1} = \begin{cases} (1+r)(s_t - c_t - i_t) + (d+1)i_t & \text{with probability } p; \\ (1+r)(s_t - c_t - i_t) & \text{with probability } 1-p. \end{cases}$$

Her only constraint is that her net wealth at t, $s_t :=$ saving$-$borrowing, has positive present value as $t \to \infty$. I.e.

$$\lim_{t\to\infty} \frac{s_t}{(1+r)^t} > 0.$$

If $1 + r < p(d + 1)$, what is an optimal strategy?

3.4 Optimal Control

3.4.1 Brachistochrone

We use the term *optimal control* to refer to optimization problems where the choice variable is a function in some continuous variable. For instance, the choice variable could be a function in space or time. Historically this type of problem first arise from the problem of Brachistochrone, from the Greek for "shortest time". It was a problem posed by Johann Bernoulli in 1696 to challenge the mathematics community in Europe of the time. The question supposes there are two points A, B with A at a higher altitute and B closer to the ground. What is the path between A, B such that a bead, starting from A with 0 initial velocity, slides along the path under only the influence of gravity to B in the shortest time? (pic)

Let $A = (A_1, A_2)$ and $B = (B_1, B_2)$. If we denotes the path by the function $y(x)$ where x is the horizonal distance from A and y is the (negative) vertical distance from A, then the problem is essentially asking to choosing a function $y(x)$ that minimizes the travelling time. By the conservation of energy, the kinetic energy of the ball at y equals the loss in potential energy. In otherwords,

$$\frac{1}{2}mv^2 = mgy \Rightarrow v = \sqrt{2gy},$$

where v is the speed at the height y. Along the curve $y(x)$, the distance given by an infinitesimal movement along the direction x is given by

$$\sqrt{1 + \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2}\mathrm{d}x.$$

The associated time spent is naturally

$$\sqrt{\frac{1 + \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2}{2gy}}\mathrm{d}x.$$

Therefore the total time spent is just an integral of x from A_1 to B_1 (pic),

$$\frac{1}{\sqrt{2g}}\int_{A_1}^{B_1}\sqrt{\frac{1+y'^2}{y}}\mathrm{d}x.$$

In otherwords, this integral as a whole the objective function, and we are trying to choose the variable y to minimizes it and the boundary values has to satisfies $y(A_1) = A_2$ and $y(B_1) = B_2$.

Johann Bernoulli's solution to this problem uses other physics such as Fermat's Principle and Snell's Law. However his method is limited in the sense that it cannot be applied to more general problems with a similar form. His brother, Jakob Bernoulli, later created a harder version of the Brachistochrone problem and used a new technique in solving it. That technique was later on refined by

Leonhard Euler and Joseph-Louis Lagrange and named *variational calculus*.

Generally, the problem can be written as

$$\max_y J[y] := \int_a^b F(y, y', x)\mathrm{d}x$$
$$\text{with boundary conditions: } y(a) = y_a;$$
$$y(b) = y_b.$$

The following is an analogue to the variational argument. Think of the problem where we try to maximize a finite dimensional function $f : \mathbb{R}^n \to \mathbb{R}$. When the function f is nice and smooth, at the maximum vector $x^* \in \mathbb{R}^n$, the derivative of f in any direction $v \in \mathbb{R}^n$ must be zero. Namely

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} f(x^* + \epsilon v)\bigg|_{\epsilon=0} = 0$$

for every $v \in \mathbb{R}^n$.

A similar stream of reasoning applies to $J[y]$, which is a function of y. As we have seen earlier, the function y is also a vector. If J is a nice and smooth function in y, at the maximum $\bar{y}$, we can take derivative of J in any direction η and the derivative must be zero. Mathematically, it is

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} J[\bar{y} + \epsilon\eta]\bigg|_{\epsilon=0} = 0. \tag{3.2}$$

However, an extra bit of care is needed here. We are trying to maximize J over the set of functions. With the motivating example above, the set of functions that we are considering should be taken to be smooth functions satisfying the boundary conditions. Therefore, we requires that $y+\epsilon\eta$ be smooth and satisfies the boundary conditions. It follows that η is smooth and $\eta(a) = \eta(b) = 0$. The equation (3.2)

must hold for all such η. Therefore,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\epsilon}J[\bar{y}+\epsilon\eta]\bigg|_{\epsilon=0} &= \int_a^b \frac{\mathrm{d}}{\mathrm{d}\epsilon}F(\bar{y}+\epsilon\eta,(\bar{y}+\epsilon\eta)',x)\bigg|_{\epsilon=0}\mathrm{d}x \\ &= \int_a^b F_1(\bar{y},\bar{y}',x)\eta + F_2(\bar{y},\bar{y}',x)\eta'\ \mathrm{d}x \\ &= \int_a^b \left(F_1(\bar{y},\bar{y}',x) - \frac{\mathrm{d}}{\mathrm{d}x}F_2(\bar{y},\bar{y}',x)\right)\eta(x)\ \mathrm{d}x,\end{aligned}$$

where the last equation is obtained using integration by parts and that $\eta(a) = \eta(b) = 0$. Because the above integral must be 0 for every η (that satisfies the boundary conditions), the integrand must be zero. Namely,

$$F_1(\bar{y},\bar{y}',x) - \frac{\mathrm{d}}{\mathrm{d}x}F_2(\bar{y},\bar{y}',x) = 0.$$

This equation is a necessary condition for $\bar{y}$ to be a maximum to $J[y]$, and it is called the *Euler-Lagrange equation*. Usually, applying the Euler-Lagrange equation to the maximization problem renders a system of differential equations with boundary conditions. Solving this system of equations would give us (a set of candidate) solutions to the maximization problem.

Applying the Euler-Lagrange equation to the Brachistochrone where

$$F(y,y',x) = \sqrt{\frac{1+y'^2}{y}},$$

and simplify, we derive

$$\frac{1}{2}(1+y'^2) + y''y = 0.$$

Solving this differential equation with boundary conditions $y(A_1) = A_2, y(B_1) = B_2$) and we see immediately that the solution is the curve of a *cycloid*.

3.4.2 Hamiltonians

A general optimal control problem (CP) is posed as

$$\begin{aligned} \max_y J[x] := & \int_a^b f(x,y,t)\mathrm{d}t \qquad \text{(CP)} \\ \text{subject to: } & x'(t) = g(x,y,t); \\ \text{with boundary conditions: } & x(a) = x_a; \\ & x(b) = x_b. \end{aligned}$$

The constraint $x'(t) = g(x,y,t)$ is called the *law of motion* for the state variable $x(t)$. The variable $y(t)$, as a function in t, is the *choice variable*. If y can be solved from the constraint $x'(t) = g(x,y,t)$ in x, x' and t, then one can substitute the expression for y into the objective function and the problem reduces to the one we saw in the previous section. However, if one cannot solve for the choice variable, the variational argument can still be applied to derive the Euler-Lagrange equation.

Consider the optimal path $\bar{x}$ (associated with $\bar{y}$ by the constraint), if we perturbe it in the direction η such that the boundary conditions are satisfied, the constraints becomes

$$\bar{x}'(t) + \epsilon\eta'(t) = g(\bar{x}(t) + \epsilon\eta(t), y(\eta;\epsilon,t), t) \tag{3.3}$$

where $y(\eta;\epsilon,t)$ is solve from the constraint equation given $\bar{x}, \eta$ and ϵ. This is essentially the resulting y determined by the constraint given a particular choice of x. Substitute both of them into the objective and we have

$$\int_a^b f(\bar{x}(t) + \epsilon\eta(t), y(\eta;\epsilon,t), t)\ \mathrm{d}t.$$

This is the value of the objective function when we differ from $\bar{x}$ by $\epsilon\eta$. The

derivative of this value in ϵ evaluated at $\epsilon = 0$ must be 0 for $\bar{x}$ to be a maximum. This gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\epsilon} J[\bar{x} + \epsilon\eta]\bigg|_{\epsilon=0} &= \int_a^b \frac{\mathrm{d}}{\mathrm{d}\epsilon} f(\bar{x}(t) + \epsilon\eta(t), y(\eta;\epsilon,t), t)\bigg|_{\epsilon=0} \mathrm{d}t \\ &= \int_a^b f_1(\bar{x},\bar{y},t)\eta + f_2(\bar{x},\bar{y},t)\frac{\partial y}{\partial \epsilon}(\eta;0,t)\ \mathrm{d}t = 0\end{aligned}$$

because $y(\eta;0,t) = \bar{y}$ by definition. If we differentiate the law of motion on both hand sides with respect to the constraint (3.3) and evaluates at $\epsilon = 0$,

$$\eta' = g_1(\bar{x},\bar{y},t)\eta + g_2(\bar{x},\bar{y},t)\frac{\partial y}{\partial \epsilon}(\eta;0,t).$$

Solve for $\frac{\partial y}{\partial \epsilon}$ and substitute into the integral to see that

$$\int_a^b f_1(\bar{x},\bar{y},t)\eta + \frac{f_2(\bar{x},\bar{y},t)}{g_2(\bar{x},\bar{y},t)}\left(\eta' - g_1(\bar{x},\bar{y},t)\eta\right)\ \mathrm{d}t = 0.$$

Defines $\bar{\lambda}(t) := -f_2(\bar{x},\bar{y},t)/g_2(\bar{x},\bar{y},t)$. Apply integration by parts and consequently

$$\int_a^b \left(f_1 + \bar{\lambda} g_1 - \bar{\lambda}'\right)\eta\ \mathrm{d}t = 0.$$

Since at $(\bar{x},\bar{y})$ the above equality has to hold for all η satisfying the boundary conditions, the Euler-Lagrange equation is now

$$\frac{\partial f}{\partial x} + \lambda\frac{\partial g}{\partial x} = \lambda'$$

where

$$\lambda(t) := -\frac{\partial f/\partial y}{\partial g/\partial y}$$

is called the *costate variable*. The above derivation can be summarized into the following theorem.

Theorem 3.4.1. *Consider the problem CP, the* Hamiltonian *to this problem is*

defined as

$$H(x, y, \lambda, t) = f(x, y, t) + \lambda g(x, y, t).$$

If the three functions $\bar{x}, \bar{y}$ *and* $\bar{\lambda}$ *together solves the maximization problem, then they must satisfy*

1. $\bar{y}$ *maximizes* $H(\bar{x}, ., \bar{\lambda}, t)$ *for every* t*;*

2. $\bar{\lambda}' = -\frac{\partial H}{\partial x}(\bar{x}, \bar{y}, \bar{\lambda}, t)$;

3. $\bar{x}' = \frac{\partial H}{\partial \lambda}(\bar{x}, \bar{y}, \bar{\lambda}, t)$;

4. $\bar{x}(a) = x_a$ *and* $\bar{x}(b) = x_b$.

The Euler-Lagrange equation is equivalent to the first and the second conditions (why?). The third one is simply the law of motion and the fourth the boundary conditions. In analogue to the Karush-Kuhn-Tucker Theorem, if certain concavity is assumed, then the above set of conditions becomes both necessary and sufficient. The following version is proved by Mangasarian in 1966.

Theorem 3.4.2. *If a Hamiltonian* H *is strictly concave in both* x *and* y *for every* t*, then the conditions in the previous theorem are both necessary and sufficient for the maximum* $(\bar{x}, \bar{y}, \bar{\lambda})$.

3.4.3 Exercises

Exercise 3.4.1. *Solve the differential equation* $\frac{1}{2}(1 + y'^2) + y''y = 0$ *with boundary conditions* $y(A_1) = A_2, y(B_1) = B_2$.

Exercise 3.4.2. *Find the minimum of the integral* $J[y] := \int_0^1 y^2 + y'^2 \mathrm{d}x$ *with boundary conditions* $y(0) = 0, y(1) = 1$.

Exercise 3.4.3. *Use variational calculus to show that the shortest distance between two points is the straight line.*

Exercise 3.4.4. *Show that the first and the second conditions in the Hamiltonian theorem is equivalent to the Euler-Lagrange equation.*

Exercise 3.4.5. *Show that the later version of the Euler-Lagrange equations implies the former if the law of motion is* $x' = g(x, y, t) = x^{\alpha} - y$ *for some* $\alpha > 0$.

Exercise 3.4.6. *Use the variational calculus method to show that the shortest distance between any two points on the Euclidean plane is the straight line.*