UNIVERSITAT DE VALENCIA

FACULTAT DE CIÈNCIES ECONÒMIQUES I EMPRESARIALS

DEPARTAMENT D'ECONOMIA FINANCERA I MATEMÀTICA

TESIS DOCTORAL

LA TEORÍA DEL CONTROL ÓPTIMO EN TIEMPO DISCRETO. MODELOS FINANCIEROS APLICADOS A LAS PENSIONES DE JUBILACIÓN

Presentada por:
D. Robert Meneu Gaya

Dirigida por:
Dr. D. Francisco Muñoz Murgui

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ÍNDICE

INTRODUCCIÓN

Este trabajo de investigación pretende ser una contribución dentro del campo de la optimización dinámica y, más concretamente, dentro de la teoría del control óptimo cuando el tiempo se toma como variable discreta. Se persigue también realizar alguna aportación en la modelización y resolución de problemas de tipo económico y financiero mediante la aplicación de éstos métodos matemáticos.

La importancia del entorno dinámico en el que se mueve la economía es evidente, de ahí la necesidad de incorporar al análisis económico instrumentos matemáticos que permitan relacionar variables y modelizar comportamientos desde un punto de vista dinámico. La ventaja del análisis dinámico sobre el estático deriva de la presencia de una variable más en el problema, el tiempo, lo cual permite encontrar soluciones mejores ya que se tiene en cuenta cómo pueden afectar las decisiones de hoy al mañana y en el análisis del presente se consideran las situaciones previstas en el futuro. De esta manera, la decisión tomada se compone de decisiones a lo largo del tiempo.

En el análisis económico dinámico, al igual que en el análisis estático, se ha venido utilizando instrumental matemático de una forma bastante habitual. Las ecuaciones dinámicas son el elemento matemático básico porque con ellas se recoge el comportamiento de las variables en el tiempo. La conjunción del análisis dinámico con la optimización matemática permite hablar de optimización dinámica cuya aplicabilidad a la economía responde al objetivo de repartir recursos escasos entre usos alternativos en un entorno de interrelaciones dinámicas entre las variables.

Los problemas susceptibles de resolverse a través de la optimización dinámica son diversos y, entre ellos, destaca el problema de control óptimo. Este problema se caracteriza por incluir variables de decisión (de control) que afectan, a través de la ecuaciones dinámicas, a otras variables (de estado) y cuyos valores se eligen de forma que optimicen algún funcional objetivo. Tales problemas admiten distintos métodos de resolución siendo la llamada teoría del control óptimo el más específico de ellos, aunque también destaca uno más general: la programación dinámica.

El análisis dinámico en tiempo discreto ha ido históricamente por detrás del análisis en tiempo continuo debido a que requiere un uso más complejo de formulaciones matemáticas de tipo analítico. En el ámbito de la teoría del control óptimo se puede citar a Boltyanskii (1978) como el intento más sólido de presentar una teoría en tiempo discreto paralela a la del tiempo continuo desarrollada 20 años antes. No obstante, creemos que es necesario un esfuerzo de esquematización y homogeneización para clarificar las semejanzas y diferencias de esta teoría con otras ya conocidas, sobre todo para aquél que, como nosotros, hace uso de la teoría del control óptimo discreto con fines aplicados.

El objetivo de este trabajo es doble. A nivel teórico, se pretende resaltar los aspectos principales de la teoría del control óptimo en tiempo discreto de manera que resulte un instrumento matemático asequible para el estudio de problemas financieros y económicos. A nivel aplicado, se persigue el estudio de problemas de tipo financiero relacionados con las pensiones de jubilación a través de su modelización y posterior resolución utilizando la teoría del control óptimo discreto. De esta manera, se completa el objetivo planteado a nivel teórico en la medida que se demuestra la utilidad del control óptimo discreto en el ámbito de las aplicaciones económicas.

Los contenidos de este trabajo están organizados en seis capítulos. Los cuatro primeros incluyen principalmente material de tipo teórico y los dos últimos desarrollan modelos matemáticos concretos para cubrir el objetivo planteado a nivel aplicado.

En la parte teórica no se ha pretendido realizar un manual de la teoría del control óptimo en tiempo discreto, sino que se ha recogido aquellos aspectos más relevantes para presentar las formulaciones propias de este método matemático de una forma clara, al mismo tiempo que rigurosa, con el objetivo de que sirvan de base para su uso posterior a nivel aplicado. Lo más destacado de esta parte es la clarificación de las interrelaciones que se dan entre los dos enfoques que existen para abordar el problema, todo ello sintetizado a través de esquemas comprensibles que

esperamos sean de utilidad. En cambio, se ha evitado profundizar en los aspectos teóricos no fundamentales para no recargar la exposición y, sobre todo, pensando en cubrir el objetivo planteado a nivel aplicado.

El primer capítulo inserta la teoría del control óptimo en tiempo discreto en el ámbito más general de la optimización dinámica, destacando la interrelación con el cálculo de variaciones y la programación dinámica. En este primer capítulo se analizan, además, las distintas alternativas para enunciar el problema, tanto a nivel de función objetivo como de restricciones y de sistema dinámico.

El capítulo segundo constituye la parte principal a nivel teórico. Partiendo del enunciado del problema recogido en el capítulo anterior se desarrollan las dos vías de enfocar la obtención de las condiciones de óptimo del problema: a través de la programación no lineal y a través del principio del máximo. La interrelación entre ambas vías para un enunciado simplificado del problema queda totalmente demostrada en uno de los teoremas del capítulo, constituyendo una de las principales aportaciones en el ámbito teórico.

La metodología seguida al enunciar y demostrar los teoremas aplicables al problema de control óptimo discreto ha consistido en tomar como punto de referencia los teoremas equivalentes en programación no lineal y realizar posteriormente las identificaciones correspondientes entre ambos tipos de problemas. De esta manera, y sin perder rigurosidad, no ha sido necesario realizar largas demostraciones matemáticas en la medida que dichas demostraciones descansan sobre otras ya conocidas en programación no lineal y de fácil consulta en cualquier manual.

En el capítulo tres se exponen aspectos complementarios relacionados con el núcleo de la teoría que se ha desarrollado en el capítulo dos. Se estudian conceptos como la existencia de solución, las condiciones de suficiencia, la estabilidad y sensibilidad de la solución óptima y la ampliación de la teoría al campo estocástico. El nivel de profundización en estos aspectos secundarios no es excesivo pero

creemos que resulta interesante incluirlos para tener un conocimiento más amplio de las posibilidades de este instrumento matemático. La lista de extensiones de la teoría del control óptimo discreto no se cierra con las que se incluyen en este capítulo y, en todo caso, siempre habrá que esperar nuevas aportaciones teóricas en el futuro.

El cuarto capítulo estudia de forma algo más específica las peculiaridades de dos tipos de problemas: los lineales y los lineales-cuadráticos. Su estudio por separado se justifica de cara a las aplicaciones que se desarrollan posteriormente, puesto que en ellas aparecen modelos que responden a la estructura lineal y lineal-cuadrática. La segunda parte del capítulo conecta la teoría con la práctica al centrarse en los métodos de resolución del problema y, en especial, en los métodos numéricos dada la limitación que suele aparecer al intentar la resolución analítica. La investigación en el campo de los métodos numéricos de resolución es, posiblemente, la que ofrece un mayor dinamismo en la actualidad. En este capítulo dedicaremos una parte a comentar los métodos numéricos más utilizados en programación no lineal y se hará una breve referencia a los todavía incipientes métodos numéricos específicos para el problema de control óptimo discreto.

La parte aplicada consta de dos capítulos en los que se desarrollan aspectos financieros relacionados con las pensiones de jubilación, siempre desde un punto de vista dinámico como corresponde al instrumento matemático que se va a utilizar. Los objetivos de los modelos que se plantean son diversos. En unos, no interesa tanto los resultados que se obtienen sino la validez del método matemático utilizado y la resolución analítica de las condiciones de óptimo. En otros modelos, además, los resultados concretos alcanzados son novedosos y revisten cierta importancia.

Nuestro interés por las pensiones de jubilación es un reflejo de la preocupación más general que se observa en el conjunto de la sociedad, donde la discusión acerca del sistema privado y público de pensiones está de permanente actualidad. Esta motivación no sólo está presente en el campo científico sino que se extiende, prácticamente, a todos los ámbitos sociales. Estas observaciones son igualmente aplicables a la mayoría de los países de nuestro entorno.

en uno de sus aspectos. En concreto, se pretende determinar a través de qué instrumento canalizar el ahorro con vistas a la jubilación dadas las características fiscales de las distintas alternativas existentes. Los resultados que se obtienen en esta aplicación no son nuevos dado que las relaciones que aparecen al solucionar el modelo se pueden conseguir también a través de ecuaciones de equilibrio financiero sin utilizar modelos de optimización. La novedad que aporta la aplicación está, por tanto, en el instrumento matemático utilizado, es decir, en el uso de la teoría del control óptimo en tiempo discreto. En la medida que los resultados son los ya conocidos, esta aplicación sirve para darles mayor consistencia a la vez que se observa la coherencia del modelo y del instrumental matemático utilizado.

En el capítulo sexto se desarrolla una aplicación referida a la parte pública de las pensiones de jubilación. En este caso, no sólo destacamos la aplicabilidad del instrumental matemático sino también los resultados obtenidos, en la medida que aportan elementos nuevos en el campo de la viabilidad futura del sistema público de pensiones. El contenido de esta aplicación intenta ser una aportación al esfuerzo intelectual que se está produciendo y que va dirigido a la construcción de modelos que sean adecuados para el análisis de la evolución de las principales variables del sistema público de pensiones y de las posibles vías de solución de los problemas a los que está sometido.

En este último capítulo se abordan dos tipos de modelos con el objetivo común de hacer frente a las necesidades futuras del sistema de pensiones público. Con ese fin, las variables de control de cada modelo se deben desviar lo menos posible de unos valores deseados (bajo un criterio de aproximación cuadrático) y se permite, como elemento fundamental del modelo, la formación de un fondo de capital cuyos rendimientos pasan a constituir una parte de los ingresos del sistema.

El primer modelo parte de supuestos simples acerca de la evolución de las variables exógenas, lo cual permite solucionar el modelo en términos analíticos. Otra característica de este modelo es que el fondo de capital es una variable de estado sometida a una condición terminal, es decir, la formación óptima del fondo de

capital se debe llevar a cabo de manera que al final del horizonte de planificación la cuantía del fondo sea la deseada. Así pues, este modelo podría servir de base si la decisión política fuera la de formar un fondo de capital durante un periodo transitorio para que, de forma permanente o hasta que se tome la decisión contraria, los rendimientos del capital pasen a formar parte de los ingresos del sistema aliviando así la carga a soportar por las cotizaciones.

El segundo tipo de modelos considera al fondo de capital como una variable de estado con valor final libre. El objetivo que se persigue en esta aplicación es determinar la evolución de las variables de control fundamentales del sistema (tipo de cotización y tasa de reemplazo de pensiones) para que sea viable el sistema de reparto ante el cambio demográfico que se espera, todo ello permitiendo la formación de un fondo de capital que tendrá la función de suavizar el impacto demográfico sobre las variables de control.

La aplicación numérica de este modelo se concreta al caso español tomando como punto de partida la proyección demográfica realizada por el Instituto de Demografía y estableciendo hipótesis y escenarios sobre el valor de los parámetros. En consecuencia, la resolución del modelo se realiza en términos numéricos, no analíticos, dado el grado de concreción exigido a las variables exógenas. Este tipo de modelos es el apropiado si la decisión política es la de mantener el sistema de reparto pero incorporando reformas necesarias, entre las que destaca la de permitir la existencia de un fondo de capital.

Un tema que nos ha preocupado particularmente y al que hemos dedicado esfuerzo en este trabajo ha sido el de los efectos de cambios en las condiciones de partida de los modelos. La posibilidad de adaptar el modelo ante cambios en los supuestos es una muestra de su flexibilidad, lo cual es una característica deseable que hemos intentado explotar. En cuanto a las aplicaciones numéricas, también nos ha interesado observar cómo cambia la solución si se toman otros valores para los parámetros u otras trayectorias temporales para las variables exógenas. Este tipo de análisis de post-optimización permite completar la información que proporciona la

solución numérica concreta.

Los resultados obtenidos tras la aplicación numérica no se pueden instaurar por decreto pero sirven de referencia para tomar las medidas de política económica adecuadas. En este sentido, se elaboran unas consideraciones acerca de cómo conseguir que las variables en la realidad se ajusten a los valores óptimos obtenidos, para lo cual hay que acudir a las reformas técnicas del sistema de reparto que habitualmente se pueden leer en la mayoría de los estudios sobre este tema y que, básicamente, vienen recogidas en el llamado Pacto de Toledo. Los cálculos que se realizan son sólo una primera aproximación ya que los datos publicados oficialmente son insuficientes para este tipo de análisis.

En todo caso, esta aplicación se inserta plenamente dentro de la polémica acerca de la viabilidad futura del sistema de pensiones público y esperamos que los resultados que aquí se obtienen constituyan una aportación a tener en cuenta a la hora de introducir las reformas ineludibles que se avecinan.

Finalmente, la tesis se cierra con una exposición de las conclusiones más relevantes derivadas de la investigación realizada y una relación de la bibliografía utilizada. Tenemos que señalar que la revisión de la literatura no ha pretendido ser exhaustiva, sino que se ha centrado en los artículos y trabajos cuya influencia y relevancia es ampliamente reconocida.

CAPÍTULO I:

OPTIMIZACIÓN DINÁMICA EN TIEMPO DISCRETO

I.1.- INTRODUCCIÓN

La optimización matemática es el instrumento más adecuado para dar respuesta al problema central en economía de distribuir recursos escasos entre usos alternativos de una manera óptima. Cuando este problema se plantea a lo largo de un horizonte temporal, de forma que los valores de las variables en un periodo o en un instante están relacionados con los valores de esas mismas variables en otros periodos o instantes, estamos ante la optimización dinámica.

La optimización dinámica no se caracteriza sólo por el hecho de optimizar a lo largo del tiempo, ya que esto se podría resolver a través de un conjunto de problemas estáticos sucesivos (optimización estática multietápica), ni tampoco por las técnicas de resolución ya que, aunque la optimización dinámica posea técnicas específicas de resolución, bajo ciertas condiciones, como veremos a lo largo de este trabajo, se pueden utilizar técnicas del análisis estático para resolver problemas dinámicos. La característica relevante para clasificar a un problema como dinámico es la interrelación entre las variables en distintos periodos o instantes de tiempo. En términos matemáticos el instrumento que refleja esta interrelación es el operador derivada (en tiempo continuo) o el operador diferencia (en tiempo discreto).

Como suele ocurrir en otros campos de la ciencia, la optimización dinámica ha ido desarrollándose con el tiempo, ampliando el tipo de problemas a los que se aplica y mejorando las técnicas generales y concretas de resolución. En este desarrollo se aprovecha la ventaja de que la optimización estática, desarrollada cronológicamente antes, ofrece campos de investigación susceptibles de ser abordados tarde o temprano por la dinámica. Desde los primeros problemas de optimización dinámica (como el problema isoperimétrico, el de la distancia más corta entre dos puntos o el de la braquistocrona) se observa una tendencia a ampliar la casuística de los problemas abordados a medida que surgen nuevas investigaciones. Por lo general, estas investigaciones tienen como referencia las desarrolladas previamente en optimización estática. Así, van emergiendo nuevas partes de la optimización dinámica como la optimización en tiempo discreto (cuando

la variable tiempo es discreta), la optimización dinámica estocástica (con variables estocásticas) y la teoría de juegos diferenciales (cuando hay más de un sujeto decisor).

Por lo que se refiere a las técnicas de resolución, la primera teoría que intentó abordar de una forma sistemática la resolución del problema de optimización dinámica fue el Cálculo de Variaciones en el s. XVII, teoría que se aplicó a problemas económicos desde los años 20 de este siglo. Las limitaciones de este método impulsan el desarrollo de métodos más generales y surge así la Teoría del Control Óptimo y la Programación Dinámica.

La Teoría del Control Óptimo tiene como principal resultado el Principio del Máximo atribuido a L.S. Pontryagin y otros en 1956, aunque Hestenes en 1950 lo había utilizado pero de una forma menos general y sin llamarlo con ese nombre. La Programación Dinámica se desarrolla paralelamente en el tiempo y en sus orígenes destacan Bellman y Isaacs; los principales resultados de este método como el Principio de Óptimo o la ecuación de Bellman son también de 1956.

Sin embargo, se ha publicado recientemente un interesante artículo (Pesch y Bulirsch, 1994) que revela las aportaciones pioneras de Carathéodory. Este artículo considera que los principales resultados en optimización dinámica (Principio del Máximo y ecuación de Bellman) se pueden derivar de los trabajos de Carathéodory de 1935 e incluso de 1926 por lo que este autor debería ser considerado como un punto de partida en el estudio de la Teoría del Control Óptimo y de la Programación Dinámica.

Estas primeras aportaciones a la optimización dinámica se desarrollaron en tiempo continuo debido a las ventajas analíticas de los instrumentos matemáticos teóricos que requiere. La alternativa de tratar la variable independiente tiempo como una variable continua o una variable discreta es importante tanto a nivel teórico como a nivel práctico y existen varios motivos por los cuales es deseable desarrollar los métodos matemáticos de la optimización dinámica en tiempo discreto, destacando

entre ellos la propia naturaleza de los problemas a resolver. Este sería el caso de algunos tipos de aplicaciones económicas que incorporan variables significativas únicamente si se evalúan en un periodo de tiempo, "variables-flujo"; simplificar el tratamiento de estas variables como si fueran "variables-fondo", valoradas en instantes temporales, en aras a una modelización del problema más operativa puede resultar a veces excesivo, es decir, la formulación continua de relaciones discretas puede no llevar siempre a aproximaciones aceptables.

De esta manera, en Borrell (1985) se destaca que conviene desarrollar modelos discretos o continuos según la naturaleza del problema a resolver. Así, se argumenta la necesidad de la versión discreta con el objetivo de resolver problemas de gran dimensión a través del ordenador, cuando las soluciones no se pueden explicitar sino sólo resolver por métodos numéricos; y también, para modelizar problemas en los que el valor de una variable en un periodo no depende sólo del valor en el periodo anterior sino también del valor en periodos precedentes (Sethi y Thompson, 1981).

Una vez enunciado un problema en tiempo discreto, aplicar mediante traslación la teoría y los métodos de resolución desarrollados en tiempo continuo no es válido de una forma general, por ello son necesarias teorías y métodos específicos en tiempo discreto, tanto en lo que se refiere al Cálculo de Variaciones, como a la Teoría del Control Óptimo y la Programación Dinámica.

Así pues, este trabajo tiene como uno de sus objetivos avanzar en la construcción de una teoría de la optimización dinámica en tiempo discreto. En concreto, se centra en el planteamiento y resolución de problemas de optimización deterministas y con un solo sujeto decisor. Este tipo de problemas se enfocan, sobre todo, a través de la Teoría del Control Óptimo y se estudian también las interrelaciones entre las dos formas de deducir las condiciones de óptimo en el problema: el Principio del Máximo, en su versión discreta; y la Programación no Lineal.

Antes de profundizar en ello conviene situar la Teoría del Control Óptimo dentro de los métodos de optimización dinámica en tiempo discreto, y hacer una breve referencia a todos ellos. Estos métodos surgen por analogía con los correspondientes métodos en tiempo continuo y son el Cálculo de Variaciones (C.V.), la Teoría del Control Óptimo (C.O.) y la Programación Dinámica (P.D.).

I.2.- CÁLCULO DE VARIACIONES EN TIEMPO DISCRETO

I.2.1.- El problema más elemental

Un problema simple de optimización dinámica en tiempo discreto se puede enunciar de la siguiente manera:

$$Max \quad F = \sum_{t=0}^{T-1} F_t(x_{t+1}, x_t) \qquad [I]$$

donde el objetivo es encontrar los valores de la sucesión de números reales:

$$\{ x_t \in \mathbb{R};\ t=0,1,2,...,T \}$$

que maximiza aquel sumatorio. La dinamicidad del problema viene del hecho que la función intermedia F_t depende de dos elementos de la sucesión, es decir de la variable evaluada en dos periodos de tiempo. El Problema [I] es el equivalente, en tiempo discreto, al problema de C.V. clásico en tiempo continuo.

Sin embargo, este instrumento no ha sido apenas desarrollado en la literatura existente hasta el momento ni lo va a ser en el futuro, a pesar de que su equivalente en tiempo continuo se remonta al s.XVII (algunos ejemplos de problemas que más tarde se resolvieron a través del C.V. son mucho más antiguos, como el isoperimétrico y el de la distancia más corta entre dos puntos). Así pues, a pesar de ser una extensión natural del caso continuo no ha sido estudiado. Creemos que las siguientes razones ayudan a explicar este hecho:

a) Los problemas a los que se aplicó el C.V., en una primera etapa, no tenían necesidad de un análisis en tiempo discreto debido a su propia naturaleza.

b) Cuando en los años 20 surgieron las aplicaciones económicas del C.V. quedaba justificado el salto al caso discreto. Sin embargo, al realizar este paso se observa que el problema resultante queda dentro del campo de la programación clásica.

c) El desarrollo de la teoría del C.O. aparcó definitivamente cualquier extensión del C.V. porque éste resultaba ser ya un instrumento superado.

Las pocas referencias existentes al C.V. en tiempo discreto, entre las que destacan las de Tapiero (1977) y Tu (1991), tratan este instrumento muy de pasada, como una simple extensión del caso continuo, y sólo puede servirnos a nivel ilustrativo.

El Problema **[I]** se puede resolver, tal y como hacen Tapiero y Tu, utilizando paralelismos con las técnicas de resolución en tiempo continuo, en concreto la ecuación de Euler. Sin embargo, se pueden utilizar también técnicas más antiguas de programación clásica, aunque de esta manera se pierde la visión dinámica del problema ya que éstas son técnicas estáticas. Para ello hay que razonar de la siguiente manera: la variable dinámica x_t se puede entender como una sucesión de $T+1$ variables estáticas, cada una de las cuales representa el valor de la variable dinámica en cada periodo de tiempo. De esta manera, el Problema **[I]** se relee como un problema de programación clásica con $T+1$ variables estáticas. A partir de aquí, y con la única condición de que las funciones intermedias sean diferenciables, se pueden aplicar las conocidas condiciones necesarias de máximo relativo, que proporcionan el conjunto de puntos críticos. Estas condiciones se expresan igualando a cero el gradiente de la función objetivo, obteniéndose para el Problema **[I]** el sistema de ecuaciones siguiente:

$$\frac{\partial F_0}{\partial x_0}=0$$

$$\frac{\partial F_0}{\partial x_1}+\frac{\partial F_1}{\partial x_1}=0$$

$$\cdots$$

$$\frac{\partial F_{T-2}}{\partial x_{T-1}}+\frac{\partial F_{T-1}}{\partial x_{T-1}}=0$$

$$\frac{\partial F_{T-1}}{\partial x_T}=0$$

Englobando las T-1 igualdades intermedias del sistema anterior, se obtiene:

$$\frac{\partial F_0}{\partial x_0}=0 \qquad (1a)$$

$$\frac{\partial F_{t-1}}{\partial x_t}+\frac{\partial F_t}{\partial x_t}=0 \qquad t=1,2,...,T-1 \quad (1b)$$

$$\frac{\partial F_{T-1}}{\partial x_T}=0 \qquad (1c)$$

Las ecuaciones (1) representan las condiciones necesarias de máximo local. La ecuación (1b) se puede entender como la equivalente en tiempo discreto a la condición de Euler en tiempo continuo. Se observa que al desarrollarla se obtiene una ecuación dinámica de la forma:

$$h(x_{t+1},x_t,x_{t-1})=0 \qquad t=1,2,...,T-1$$

es decir, una ecuación en diferencias de segundo orden. Para resolverla se necesita de dos condiciones de transversalidad que vienen dadas por las ecuaciones (1a) y (1c).

La condición suficiente de máximo local también se conoce en programación clásica y hace referencia a la concavidad de la función objetivo en un punto crítico x', condición que queda asegurada si la siguiente matriz Hessiana de la función objetivo representa una forma cuadrática definida negativa:

$$HF(x') = \begin{pmatrix} \frac{\partial^2 F(x')}{\partial x_1^2} & \cdots & \frac{\partial^2 F(x')}{\partial x_1 \partial x_T} \\ \cdots & & \cdots \\ \frac{\partial^2 F(x')}{\partial x_T \partial x_1} & \cdots & \frac{\partial^2 F(x')}{\partial x_T^2} \end{pmatrix}$$

El hecho de que se utilicen técnicas estáticas para resolver un problema dinámico no significa que éste carezca de entidad propia ya que modelizar y presentar los resultados con los instrumentos propios de las variables dinámicas discretas (ecuaciones en diferencias) supone una ventaja de concreción respecto a tratar la variable de estado en cada periodo de tiempo como una variable estática distinta. Relacionado con esto, en Tapiero (1977) se reflexiona sobre si se necesita un procedimiento para la optimización de sistemas dinámicos llegándose a la conclusión de que sí ya que, aunque la solución se pueda encontrar con técnicas de optimización clásica, éste es un método ineficiente porque la estructura dinámica de los problemas no se utiliza como una ventaja en la optimización clásica y por tanto se realiza un esfuerzo de cálculo innecesario. Además, destaca implicaciones conceptuales de la visión dinámica a la hora de comprender mejor la plasmación de interrelaciones económicas en un modelo dinámico.

Por tanto, se puede aprovechar las ventajas, en lo que a técnicas de resolución se refiere, de la equivalencia de nuestro problema con otro ya conocido como el de programación clásica, pero sin perder de vista que se trata de un problema dinámico, es decir, la solución se debe interpretar como una variable dinámica y no como un conjunto de variables estáticas.

El anterior enunciado simple del Problema **[I]** del Cálculo de Variaciones admite algunas extensiones que no suponen cambios cualitativos importantes en cuanto al método de resolución apropiado. Dado que, en todos los casos, las técnicas de programación clásica son satisfactorias para afrontar estas modificaciones, sólo se van a comentar brevemente.

I.2.2.- Condiciones de contorno

Existen distintos tipos de problemas según se conozcan o no los valores iniciales y finales de las variables de estado. Así, se dice que el enunciado **[I]** es un problema de punto inicial y final libres porque no se parte de valores conocidos inicial, x_0; ni final, x_T. De forma similar puede aparecer el problema de punto inicial fijo o el problema de punto final fijo. Es fácil observar que si se fija el punto inicial desaparece la condición (1a) y si se fija el punto final desaparece la condición (1c) al existir en cada caso una variable menos en el problema. El enunciado más habitual es el de punto inicial conocido y punto final libre:

$$x_0 = \bar{x}_0 \qquad (1d)$$
$$x_T \in \mathbb{R}$$

Este problema tiene como condiciones necesarias (1b) y (1c), junto con (1d).

I.2.3.- Función residual

A veces, sobre todo en aplicaciones económicas, suele aparecer en la función objetivo algún tipo de función que depende del estado final del sistema y que tiene la interpretación de una función residual $S(x(T))$. Si se introduce en el Problema **[I]** en forma de un término adicional de la función objetivo el único cambio que se produce es sobre la condición (1c) que ahora se transforma en:

$$\frac{\partial F_{T-1}}{\partial x_T}+\frac{dS}{dx_T}=0 \qquad (1c')$$

I.2.4.- Variable n-dimensional

Si aparece más de una variable, es decir $x\in\mathbb{R}^n$, el conjunto de condiciones (1) debe exigirse para cada una de las variables. Basta por tanto con reinterpretar x_t como un vector $(x_t^1, \ldots, x_t^n)$ y al mismo tiempo reescribir las condiciones necesarias como:

$$\frac{\partial F_0}{\partial x_0^i}=0 \qquad i=1,\ldots,n$$

$$\frac{\partial F_{t-1}}{\partial x_t^i}+\frac{\partial F_t}{\partial x_t^i}=0 \qquad t=1,\ldots,T-1 \;;\; i=1,\ldots,n$$

$$\frac{\partial F_{T-1}}{\partial x_T^i}=0 \qquad i=1,\ldots,n$$

Lo que se tiene es, por tanto, un sistema de n ecuaciones en diferencias de orden 2 con un conjunto de $2n$ condiciones de contorno.

I.2.5.- Desfases de orden mayor que 1

Si las funciones intermedias dependen de la variable de estado desfasada en más de un periodo, las condiciones (1) tienen más sumandos. Suponemos que aparecen $s+1$ desfases en las funciones intermedias, es decir son de la forma:

$$F_t\,(x_{t+1},\, x_t,\, x_{t-1},\, \ldots,\, x_{t-s})$$

y suponemos también que se conocen los $s+1$ primeros valores de la sucesión: x_0, x_1, ..., x_s conocidos. Entonces se tiene el Problema **[II]**:

$$Max \quad \sum_{t=s}^{T-1} F(x_{t+1}, x_t, x_{t-1}, ..., x_{t-s}) = \sum_{t=s}^{T-1} F_t \qquad \textbf{[II]}$$

$$x_0, x_1, ..., x_s \quad \textit{conocidos}$$

Las condiciones necesarias asociadas son:

$$\frac{\partial F_{t-1}}{\partial x_t} + \frac{\partial F_t}{\partial x_t} + ... + \frac{\partial F_{t+s}}{\partial x_t} = 0 \qquad t=s+1,...,T-1$$

$$\frac{\partial F_{T-1}}{\partial x_T} = 0$$

Hay que tener en cuenta que en la primera igualdad, para que la notación sea consistente, se anulan los sumandos en los que el subíndice de la función F sea mayor que T-1. Se observa que la solución forma una ecuación en diferencias de orden $s+2$.

I.2.6.- Restricciones sobre las variables

Es también habitual en las aplicaciones económicas que los valores de las variables no sean del todo libres sino que deban pertenecer a un determinado subconjunto. En la medida en que este subconjunto se pueda delimitar por restricciones de igualdad, el problema se puede resolver a través del método de Lagrange, esto es, sin salirse de la programación clásica. El problema, en el caso de m ($m<T+1$) restricciones de igualdad funcionalmente independientes, admite el siguiente enunciado:

$$Max \quad \sum_{t=0}^{T-1} F(x_{t+1}, x_t) = \sum_{t=0}^{T-1} F_t$$

$$s.a: \quad g_1(x_T, x_{T-1}, \ldots, x_0) = 0 \qquad \text{[III]}$$

$$\vdots$$

$$g_m(x_T, x_{T-1}, \ldots, x_0) = 0$$

El problema **[III]** se resuelve con programación clásica. Para ello se define la siguiente función Lagrangiana:

$$L(x_T, x_{T-1}, \ldots, x_0, \mu_1, \ldots, \mu_m) = \sum_{t=0}^{T-1} F_t + \sum_{j=1}^{m} \mu_j g_j$$

El vector $(\mu_1, \ldots, \mu_m)$ es el de multiplicadores de Lagrange. Las condiciones necesarias de máximo local se obtienen al igualar a cero las derivadas parciales primeras de la función Lagrangiana. Si la estructura que adoptan las restricciones es adecuada se llega a una ecuación en diferencias de segundo orden como en el caso más simple; sin embargo, en el caso general, esto no queda asegurado y se tiene que resolver un sistema de ecuaciones simultáneas.

Por otra parte, las derivadas de segundo orden de la función Lagrangiana forman la matriz Hessiana que representa a una forma cuadrática restringida que, en el caso que sea definida negativa, determina con carácter suficiente el máximo local.

I.2.7.- Un problema más general

Las consideraciones introducidas en los epígrafes anteriores permiten establecer como enunciado más general de un problema de C.V. en tiempo discreto el siguiente:

$$\begin{aligned} Max \quad & \sum_{t=0}^{T-1} F(x_{t+1}, x_t) + S(x_T) \\ s.a: \quad & x_t \in X_t \qquad t = 1, \ldots, T \\ & x_0 = \bar{x}_0 \end{aligned} \qquad \text{[IV]}$$

donde x_t ($x_t \in \mathbb{R}^n$, $\forall t$) son las variables dependientes y X_t ($X_t \subset \mathbb{R}^n$, $\forall t$) son los subconjuntos definidos por restricciones de igualdad. El objetivo es encontrar aquella sucesión en $\mathbb{R}^n$ de entre las sucesiones admisibles (sucesiones que cumplan las restricciones y cuyo valor inicial sea el fijado) que maximice la función objetivo. La solución se obtiene construyendo la función lagrangiana, a partir de la cual, mediante las correspondientes derivadas parciales, se obtiene un sistema de ecuaciones que, según la forma de las restricciones, puede ser también un conjunto de ecuaciones en diferencias de segundo orden con sus condiciones de transversalidad.

I.3.- CONTROL ÓPTIMO EN TIEMPO DISCRETO

I.3.1.- Consideraciones previas

El problema de C.O. es un problema de optimización dinámica en el que aparecen variables de estado sobre las que el sujeto decisor no puede influir directamente sino sólo a través de otras variables, llamadas de control, que son las variables de decisión. Ambos tipos de variables son funciones del tiempo, pero mientras las variables de control toman valores dictados por la decisión del sujeto decisor, las variables de estado siguen unas sendas temporales dictadas por un sistema dinámico (conjunto de ecuaciones dinámicas) en cuya expresión intervienen las variables de control. Como en todo problema de optimización dinámica, el sujeto debe encontrar, con todas las limitaciones que se le planteen (restricciones), los mejores valores que puede asignar a las variables de decisión para optimizar algún funcional objetivo.

Este problema pretende ser equivalente en tiempo discreto al problema de control óptimo desarrollado originariamente en tiempo continuo y como tal debe admitir restricciones sobre las variables de control. Por otra parte, como veremos, se puede identificar como un problema de extremos de funciones en un subconjunto del espacio euclídeo n-dimensional. Más concretamente, si se dan condiciones de diferenciabilidad de las funciones y si el subconjunto en el que se encuentran los extremos está definido a través de igualdades y/o desigualdades, el problema se reduce a uno de programación matemática (en general de programación no lineal - P.N.L.- dado que la programación clásica no permite restricciones de desigualdad).

Efectivamente, el problema de C.O. discreto se puede enfocar como uno de extremos de funciones y, por tanto, de optimización estática de manera análoga a lo realizado con el Cálculo de Variaciones. Para ello, se interpreta cada variable no como una variable dinámica sino como un conjunto de variables estáticas: una variable dinámica en tiempo discreto es una sucesión finita $\{x_0, x_1, \ldots, x_T\}$ y queda determinada si se determinan todos los valores de esa sucesión.

Así pues, el problema se puede abordar tomando como base la teoría del control óptimo en tiempo continuo o a partir de la programación no lineal. Ambas aproximaciones han sido estudiadas y puestas en relación a lo largo de los últimos años por diferentes autores. Aunque el desarrollo formal se realiza con más detalle en el siguiente capítulo conviene plasmar de forma breve los rasgos generales de cada una de ellas y su interrelación:

a) La **aproximación a través de la teoría del C.O.** trata a las variables como dinámicas y estudia los supuestos bajo los cuales se puede enunciar para el caso discreto un Principio del Máximo (P.M.) similar al del caso continuo. Dicho P.M. contiene, entre otras, una condición necesaria de máximo representada por la maximización de una función, llamada función hamiltoniana, respecto a las variables de control, condición que caracteriza a esta aproximación.

b) La **aproximación a través de la P.N.L.** trata a las variables como estáticas y desarrolla las condiciones de máximo aplicando las condiciones de Kuhn y Tucker (K-T). Se trata de estudiar bajo qué supuestos dichas condiciones son necesarias y/o suficientes para la maximización. Dichas condiciones no recogen ninguna que maximice la función hamiltoniana.

c) La **interrelación entre ambas aproximaciones** trata de establecer una correspondencia del tipo si y sólo si entre las condiciones de K-T y la maximización de la función hamiltoniana. Para lograr esta correspondencia se establecen supuestos cada vez menos restrictivos acerca de las funciones que intervienen en el enunciado del problema y acerca de los conjuntos sobre los que están definidos dichas funciones.

Dado que el estudio de la teoría del Control Óptimo en tiempo discreto es nuestro principal objetivo, conviene especificar la nomenclatura y el enunciado del problema que se va a seguir en los próximos capítulos. En cuanto a la nomenclatura básica se tiene:

- La **variable independiente tiempo**, t, es discreta, es decir, se habla de periodos de tiempo y no de instantes de tiempo. Se parte siempre del instante 0. Desde ese instante hasta el instante 1 transcurre el periodo 0, de la misma forma desde el instante t-1 hasta el instante t transcurre el periodo t-1. Los valores de las variables se asignan al principio del periodo y son constantes en el periodo. Salvo que se indique otra cosa, se supondrá que el proceso de optimización termina en el instante T, es decir, se consideran T periodos.

- La **variable de estado** en un periodo t es una variable n-dimensional representada por $x_t=(x_t^1,\ldots,x_t^i,\ldots,x_t^n)$. El total de vectores de variables de estado en un problema con T periodos es $T+1$, $\{x_0,x_1,\ldots,x_t,\ldots,x_T\}$, y por tanto el vector que incluye todas las variables de estado es $X=(x_0,x_1,\ldots,x_t,\ldots,x_T)$, es decir, un vector con $(T+1)n$ componentes.

- La **variable de control** en un periodo t es una variable m-dimensional representada por $u_t=(u_t^1,\ldots,u_t^j,\ldots,u_t^m)$. El total de vectores de variables de control en un problema con T periodos es T, $\{u_0,u_1,\ldots,u_t,\ldots,u_{T-1}\}$, de donde el vector que contiene todas las variables de control es $U=(u_0,\ldots,u_t,\ldots,u_{T-1})$, vector con Tm componentes.

En la Figura I.1 se recoge la situación de las variables en el tiempo.

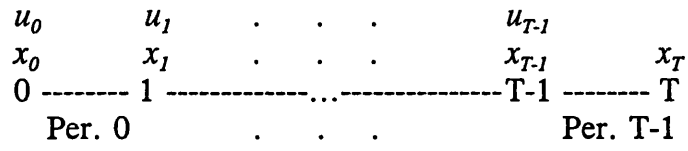

Figura I.1.: Variables de estado y de control en C.O. discreto.

Una vez introducida la notación básica se pasa a continuación a especificar las partes que intervienen en el enunciado de un problema de C.O. discreto y que son:

- El sistema dinámico.
- Las restricciones sobre los dos tipos de variables.
- Las condiciones de transversalidad.
- La función objetivo.

I.3.2.- El sistema dinámico

El sistema dinámico establece las relaciones de cambio dinámico de las variables de estado según los valores asignados a las variables de control, es decir, establece la dinámica de las variables de estado. Estas relaciones se modelizan mediante ecuaciones diferenciales en tiempo continuo y mediante ecuaciones en diferencias en tiempo discreto. La dinámica de las variables de estado, esto es, cómo las variables de estado cambian de valor en un periodo en función de los valores de todas las variables en el periodo anterior, viene descrita por funciones vectoriales de la forma $f_t(x_t, u_t)$ definidas de $\boldsymbol{R}^n x \boldsymbol{R}^m$ sobre $\boldsymbol{R}^n$ y que son continuamente diferenciables respecto a todas las variables.

Los sistemas dinámicos en tiempo discreto que se van a considerar son del tipo:

$$x_{t+1} - x_t = f_t(x_t, u_t) \qquad t = 0, 1, \ldots, T-1 \qquad (2)$$

donde cada función vectorial se compone de n funciones reales, es decir, $f_t = (f_t^1, \ldots, f_t^i, \ldots, f_t^n)$.

Existen autores[1] que expresan el sistema dinámico de una forma más general que (2), esto es:

[1]Se puede citar entre otros a Arkin y Evstigneev, Boltyanskii, Chow, Nahorsky, Ravn y Valqui, Tu, etc.

$$x_{t+1} = f_t(x_t, u_t) \qquad t=0,1,...,T-1 \qquad (3)$$

y otros autores[2] prefieren enunciar el sistema dinámico como:

$$f_t(x_{t+1}, x_t, u_t) = 0 \qquad t=0,1,...,T-1 \qquad (4)$$

Las formas (3) y (4) tienen la ventaja de incluir relaciones dinámicas más generales que las posibles con (2) pero tienen el inconveniente de apartarse del paralelismo que (2) tiene con el caso continuo donde el sistema dinámico es:

$$\frac{dx(t)}{dt} = f(x(t), u(t), t) \qquad (5)$$

Cuando más adelante se utilice el sistema dinámico se seguirá la expresión (2) dado que es la más utilizada en la literatura consultada.

I.3.3.- Las restricciones

Las restricciones en el problema de control, como en cualquier problema de optimización, determinan el subconjunto en el que se debe realizar la búsqueda de los óptimos. Relacionado con las restricciones y con el sistema dinámico se pueden definir tres conceptos.

> *Definición I.1.:* **Control admisible** es aquel valor de las variables de control que satisface todas las restricciones y que da lugar a estados que son compatibles con las restricciones y con el sistema dinámico. Utilizando una nomenclatura propia de problemas de optimización

[2]Entre los que se puede citar a Dechert, Lin y Yang, etc.

diremos que un control admisible equivale a un control factible.

Definición I.2.: **Estado alcanzable** en un periodo t es aquel valor de las variables de estado que cumple las restricciones y que se puede alcanzar desde el periodo t-1 a través de un control admisible y mediante la aplicación del sistema dinámico.

Definición I.3.: **Soluciones factibles** son todos aquellos valores de las variables de estado y de control que, cumpliendo las restricciones, verifican conjuntamente el sistema dinámico.

Las restricciones de un problema de C.O. pueden adoptar múltiples formas.

- Una primera distinción de las restricciones se puede establecer según afecten a variables de estado, de control o a ambas al mismo tiempo.

- En segundo lugar, pueden ser estáticas en el sentido que afecten a variables evaluadas en un mismo periodo de tiempo o dinámicas si relacionan variables en más de un periodo de tiempo.

- Por último, las restricciones se pueden plasmar a través de igualdades, desigualdades o a través de subconjuntos más generales.

Existe la posibilidad de encontrar una notación general que incluya todos los tipos de restricciones posibles a través de una función vectorial que dependa de todas las variables que aparecen en el problema. Para ello las variables dinámicas discretas se descomponen en sucesiones de variables estáticas de la forma habitual.

Suponiendo que en el problema haya s restricciones, se necesita una función vectorial h para expresar esas restricciones de la siguiente forma:

$$h(x_0,...,x_T,u_0,...,u_{T-1})\in H\subset\mathbb{R}^s \qquad (6)$$

o equivalentemente:

$$(x_0,...,x_T,u_0,...,u_{T-1})\in D\subset\mathbb{R}^r \qquad (6')$$

donde $h:\mathbb{R}^r \longrightarrow \mathbb{R}^s$ y D es el subconjunto de puntos que cumplen (6). A su vez r es el número total de variables de estado y control, es decir, $r=(T+1)n+Tm$.

La restricción se puede escribir de forma más compacta como:

$$h(X,U)\in H \qquad (7)$$

o:

$$(X,U)\in D \qquad (7')$$

La expresión (7) es una expresión general de cualquier restricción. Sin embargo, muchos autores creen conveniente especificar de forma separada ciertas expresiones que recojan sólo un tipo de restricciones que, eso sí, sean frecuentes en la mayoría de los enunciados de un problema de control. De esta forma se sacrifica la generalidad de la expresión (7) a cambio de diferenciar claramente distintos tipos de restricciones, lo cual tiene la ventaja de que permite establecer condiciones de óptimo mucho más operativas.

En Canon, Cullum y Polack (1970) se distingue, por ejemplo, otros dos tipos de restricciones; las restricciones de control que afectan a las variables de control en cada periodo de tiempo y las restricciones de estado-espacio que afectan a las

variables de estado en cada periodo de tiempo, es decir:

$$u_t \in U_t \; ; \quad U_t \subset \mathbb{R}^m \quad t=0,1,...,T-1 \qquad (8)$$

$$x_t \in X_t \; ; \quad X_t \subset \mathbb{R}^n \quad t=0,1,...,T \qquad (9)$$

Así mismo incluyen la expresión general (7) para recoger otro tipo de restricciones llamadas restricciones de trayectoria. De esta manera, el conjunto factible queda reducido a subconjuntos para ambos tipos de variables (regiones de control y regiones de estado). Esta división permite que ciertas restricciones muy frecuentes se puedan englobar fácilmente en un conjunto u otro. Para ver esto pueden servir los siguientes ejemplos.

Un ejemplo de restricciones del tipo (7) son las restricciones que expresan limitaciones de utilización de recursos a lo largo del periodo de planificación. Así sea u_t el vector control de cantidades empleadas de m recursos en un periodo t, y sea d el vector que tiene como componentes las cantidades disponibles para cada uno de ellos a lo largo de los T periodos, entonces, en ausencia de otras restricciones del tipo (7), se puede expresar el conjunto D de la siguiente manera:

$$D = \{(u_0,...,u_{T-1}) \in \mathbb{R}^{Tm} \; / \; \sum_{t=0}^{T-1} u_t^j \leq d^j \; ; \quad j=1,...,m\}$$

Un ejemplo de restricciones del tipo (8) es el caso de controles acotados, entonces el subconjunto U_t adopta una expresión similar a esta:

$$U_t = U_t^1 \times ... \times U_t^m \; , \quad con \quad U_t^j = \{u_t^j \in \mathbb{R} \; / \; u_{\min}^j \leq u_t^j \leq u_{\max}^j\}$$

Como ejemplo de restricciones del tipo (9) puede servir la exigencia de trabajar en el ortante positivo para muchas de las variables de estado que se utilizan

en aplicaciones económicas, en ese caso el conjunto X_t es el siguiente:

$$X_t = \{x_t \in \mathbb{R}^n \ / \ x_t \geq 0\}$$

En Boltyanskii (1978) se utiliza únicamente las restricciones (8) y (9), y en sus desarrollos posteriores distingue dos casos para la región de control: cuando no depende de la variable de estado (U_t) y cuando sí que depende ($U_t(x_t)$).

Otros autores se han decantado por limitar el tipo de restricciones que aparecen en el problema con el objetivo de ganar en operatividad. Por ejemplo en Tapiero (1977) se distingue sólo un tipo de restricciones en el enunciado general del problema: las restricciones de control. Además no permite que éstas dependan del tiempo y por tanto las engloba en una expresión del tipo $u_t \in U$.

En Sethi y Thompson (1981) se utiliza sólo la restricción de control del tipo (8) mientras que en Hwang y Fang (1966) solamente se utilizan restricciones de la forma $h(u_0,\ldots,u_{T-1}) \in H$ que es una expresión menos general que la (7).

Un paso más en la delimitación del tipo de restricciones se da cuando se procede a concretar los subconjuntos mediante igualdades o desigualdades. Este paso supone una merma de generalidad en el problema pero permite la obtención de resultados más potentes. De hecho es imprescindible si se quiere seguir la línea de la programación matemática para la resolución del problema. Si esta concreción no se produce se llega a resultados cuya validez depende de condiciones más difíciles de comprobar. Afortunadamente, los modelos económicos permiten que un conjunto de oportunidades quede totalmente definido a través de un sistema de igualdades y desigualdades.

El tratamiento más riguroso de toda la casuística que se puede plantear en este sentido se recoge en el libro de Boltyanskii (1978). Este autor incluye restricciones de igualdad y desigualdad a la hora de concretar tanto las regiones de

control como las de estado y también considera el caso general donde las regiones vienen dadas a través de subconjuntos. Otros autores reducen el tipo de restricciones en sus enunciados básicos para llegar a resultados más manejables.

Una nota común a todos ellos es que prescinden de las restricciones de trayectoria. Así, además de Boltyanskii, autores como Kleindorfer y otros (en Tapiero, 1977), Nahorsky, Ravn y Valqui (1984), Lin y Yang (1989), etc., sólo consideran las restricciones del tipo (8) y (9) con expresiones como $g(x_t)=0$, $g(u_t)=0$ o la muy utilizada $g(x_t,u_t)=0$. Como se ve ninguna de estas expresiones incluye a variables en distintos periodos de tiempo (restricciones de trayectoria), lo cual supone una simplificación. Sin embargo, tal y como se indica en Nahorski y otros (1984), existen ciertos tipos de restricciones de trayectoria, en concreto las que adoptan la forma general:

$$\sum_{t=0}^{T-1} g_t(x_t,u_t)\in H\subset\mathbb{R}^s$$

que pueden transformarse en restricciones de estado-espacio por medio de la creación de variables de estado adicionales que cumplan:

$$\tilde{x_0}=0 \quad y \quad \tilde{x_{t+1}}-\tilde{x_t}=g_t(x_t,u_t) \qquad t=0,\ldots,T-1$$

y la consecuente redefinición del problema.

I.3.4.- Las condiciones de contorno

Un primer nivel de condiciones de contorno son las restricciones que se pueden imponer o no al valor de las variables de estado en los periodos inicial y final. En general, se puede incluir dentro del conjunto de restricciones de estado-espacio ya que las restricciones $x_0\in X_0$ y $x_T\in X_T$ son también condiciones de

transversalidad. Sin embargo es bastante habitual, y muchos autores así lo reflejan, que los conjuntos X_0 y X_T sean todo el espacio vectorial (condiciones de contorno libres) o tengan un solo punto (condiciones de contorno fijadas). Según esto el problema de control óptimo puede ser:

- de puntos inicial y final conocidos.
- de puntos inicial y final libres.
- de punto inicial libre y final fijado.
- de punto inicial fijado y final libre.

La elección de una posibilidad u otra no es importante dado que los problemas se resuelven de forma similar, sin embargo preferimos la última posibilidad por ser la más habitual en la literatura, por tanto en el enunciado del problema de control el conjunto X_0 está formado sólo por aquel punto cuyas componentes son los valores de las variables de estado en el momento inicial, es decir:

$$X_0 = \{\bar{x}_0\} = (\bar{x}_0^1,\ldots,\bar{x}_0^n) \qquad (10)$$

dejando la condición de contorno final, si existe, englobada dentro de las restricciones más generales del tipo (9).

Por paralelismo con la teoría del C.O. en tiempo continuo se puede considerar un segundo nivel de condiciones de contorno que hace referencia a si se fija o no la duración del horizonte de planificación. Esto equivale a determinar si el número de periodos (T) que aparece en el problema de control es fijo o libre, dado que siempre se parte del periodo cero y los periodos son de amplitud uno. Esta distinción es importante debido a que la mayoría de estudios y resultados obtenidos se han hecho para el caso de tiempo fijo y extender los métodos al caso de tiempo libre no es trivial. No obstante, se hará una breve extensión al caso en que sea libre, es decir, cuando sea una variable más del problema.

I.3.5.- La función objetivo

La función objetivo (o funcional objetivo[3]) es la que se encarga de evaluar las soluciones óptimas (función de costes, beneficios, etc.). Efectivamente, el sistema dinámico puede evolucionar de distintas maneras posibles según la elección hecha sobre las variables de control, incluso si el valor final de las variables de estado está fijado. La función objetivo sirve de referencia para elegir de entre todas ellas la mejor, en el sentido de que lleve a un máximo o un mínimo de dicha función. Se trata de una función real en la que intervienen, en general, todas las variables del problema. Ello sugiere una función del tipo $F(X, U)$ pero de nuevo ésta es una expresión demasiado general para la función objetivo.

La forma más habitual de enunciar esta función es como suma de funciones intermedias (F_t), dado que lo más corriente es que en cada periodo se evalúe el objetivo (beneficios, costes, etc.). De esta forma el objetivo total es la suma de los objetivos en cada periodo, es decir:

$$F(X,U) = \sum_{t=0}^{T-1} F_t(x_t,u_t)+F_T(x_T) \qquad (11)$$

La función F_T se puede interpretar en algunas aplicaciones económicas como un valor residual ya que depende del estado final del sistema.

Este planteamiento, llamado de Bolza, es uno de los más utilizados en la literatura consultada y tiene la ventaja de separar el carácter del objetivo en el último periodo del carácter en periodos anteriores lo cual suele ser habitual en aplicaciones económicas.

---

[3]En nomenclatura propia del análisis dinámico es más propio utilizar "funcional objetivo" porque los argumentos de las funciones que intervienen son, a su vez, funciones (dependen de la variable independiente tiempo). Sin embargo, dado que el problema es reducible a uno de extremos, se puede mantener la expresión "función objetivo".

Existen otras formas de enunciar la función objetivo que también son habituales y que, aunque parecen distintas a la (11) en realidad no lo son dado que se puede pasar de un enunciado a otro haciendo algún cambio de variable y combinando la expresión (11) con el sistema dinámico (2). Supongamos la función objetivo en la llamada forma de Mayer:

$$F(X,U) = G_T(x_T) \qquad (12)$$

utilizada entre otros por Sethi y Thompson (1981). Se puede demostrar el siguiente teorema:

Teorema I.1.: Las formas (11) y (12) de expresar la función objetivo son equivalentes si se redefine de forma adecuada el vector de variables de estado y el sistema dinámico.

Demostración: (12) implica (11) sin más que tomar las funciones intermedias F_t iguales a 0 para $t=0,\ldots,T\text{-}1$ y hacer $G_T=F_T$.

(11) implica (12) si se crea una variable discreta x^0 con las siguientes características:

$$x_{t+1}^0 - x_t^0 = F_t(x_t,u_t) \qquad t=0,1,\ldots,T-1$$

$$x_0^0 = 0$$

entonces (11) se transforma en:

$$F(X,U) = \sum_{t=0}^{T-1} (x_{t+1}^0 - x_t^0) + F_T(x_T) =$$

$$= x_T^0 + F_T(x_T) = G_T(\chi_T)$$

$$con \quad \chi_T = (x_T^0, x_T^1, ..., x_T^n)$$

con el sistema dinámico y las condiciones de transversalidad correspondientes siguientes:

$$\chi_{t+1} - \chi_t = \varphi_t(x_t, u_t) \qquad t = 0, 1, ..., T-1$$

$$\chi_0 = \bar{\chi}_0$$

$$con \quad \varphi_t = (F_t, f_t) \quad y \quad \bar{\chi}_0 = (0, \bar{x}_0)$$

dando lugar a una función objetivo propia de la expresión (12).

Una tercera forma de expresar el funcional objetivo es la de Lagrange:

$$F(X,U) = \sum_{t=0}^{T-1} F_t(x_t, u_t) \qquad (13)$$

para la cual también se verifica el siguiente teorema:

Teorema I.2.: La expresión (13) es equivalente a cualquiera de las dos anteriores (expresiones (11) y (12)).

Demostración: (13) implica (11): basta con hacer $F_T(x_T)=0$.

(12) implica (13): basta con realizar las siguientes transformaciones:

$$F_t(x_t,u_t) = 0 \qquad t=0,1,...,T-2$$

$$F_{T-1}(x_{T-1},u_{T-1}) = G_T(x_T)$$

La demostración del teorema se completa aplicando el teorema I.1..

Pese a que estas tres formas son las más habituales para expresar la función objetivo, no se puede ignorar que existen otras formas de expresión que no se pueden reducir a (11). Un ejemplo es cuando queremos minimizar la máxima desviación de una función K_t respecto a un valor deseado K que se produce en un periodo. En este caso la función objetivo es:

$$F(X,U) = Max\ \{|K_t(x_t,u_t)-K|,\ t=0,...,T-1\}$$

I.3.6.- Enunciado estándar del problema

Una vez hecho el repaso de las distintas formas de expresar cada componente del problema de control, es conveniente tomar un enunciado estándar del problema sin perjuicio de hacer referencia ocasional a otros enunciados que, por algún motivo, sean interesantes. Recopilando lo visto en anteriores epígrafes aparece el siguiente enunciado:

$$
\begin{aligned}
&Max \quad F(X,U) = \sum_{t=0}^{T-1} F_t(x_t,u_t)+F_T(x_T) \\
&s.a: \\
&\quad x_{t+1}-x_t = f_t(x_t,u_t) \qquad t=0,\ldots,T-1 \\
&\quad u_t \in U_t \subset \mathbb{R}^m \qquad t=0,\ldots,T-1 \\
&\quad x_t \in X_t \subset \mathbb{R}^n \qquad t=0,\ldots,T \\
&\quad h(X,U) \in H \subset \mathbb{R}^s \\
&\quad X_0 = \{\bar{x}_0\} \\
&\quad T \text{ conocido}
\end{aligned}
\qquad [V]
$$

Es decir, el problema estándar consiste en una función objetivo que se puede descomponer en suma de funciones intermedias y una función residual, un sistema dinámico como un conjunto de ecuaciones en diferencias, unas restricciones de control, de estado-espacio y de trayectoria, un punto inicial fijo y final libre y una duración del proceso conocida.

Para el Problema **[V]** se establece la siguiente definición de solución óptima:

> *Definición I.4.:* La **solución óptima** de un problema de control está formada por unas sucesiones de las variables de control $\{u_t^*;\ t=0,\ldots,T\text{-}1\}$ y las correspondientes sucesiones de las variables de estado $\{x_t^*;\ t=0,\ldots,T\}$ de tal forma que, siendo compatibles con el sistema dinámico (2), con las restricciones (7), (8) y (9) y con la condición inicial (10), se consiga la desigualdad:
>
> $$F(X^*,U^*) \geq F(X,U)$$
>
> para todo vector (X,U) que cumpla (2), (7), (8), (9) y (10), es decir,

para toda solución factible.

I.3.7.- Enunciado en forma canónica

El enunciado cambia si se adopta otra forma de expresar alguna de las cuatro partes que componen el problema de control: la función objetivo, el sistema dinámico, las restricciones o las condiciones de transversalidad.

En la mayoría de los casos el cambio del enunciado no supone cambios cualitativos importantes de cara a la posterior resolución del problema. Sin embargo, es conveniente destacar la llamada forma canónica del problema de control óptimo discreto, que aparece cuando la función objetivo anterior se pasa a la forma de Mayer (12) y se realizan los cambios ya vistos en el sistema dinámico y en las condiciones de transversalidad. El enunciado en forma canónica es:

$$
\begin{aligned}
&Max \quad F(X,U) = G_T(\chi_T) \\
&s.a: \\
&\chi_{t+1}-\chi_t = \varphi_t(\chi_t,u_t) \qquad t=0,...,T-1 \\
&u_t \in U_t \subset \mathbb{R}^m \qquad t=0,...,T-1 \\
&\chi_t \in X_t \subset \mathbb{R}^{n+1} \qquad t=1,...,T \\
&h(X,U) \in H \subset \mathbb{R}^s \\
&\chi_0 = \bar{\chi}_0 \\
&T \text{ conocido}
\end{aligned}
\qquad [VI]
$$

La ventaja de la forma canónica, como se indica en Canon, Cullum y Polack (1970) es que permite un tratamiento geométrico del problema y esto sirve para utilizar técnicas de perturbaciones alrededor de las sucesiones óptimas con el fin de determinar condiciones necesarias de óptimos. Estas técnicas, sin embargo, no son más eficientes que las que se derivan de la programación no lineal, según estos

mismos autores.

I.3.8.- Enunciado del problema de tiempo final libre

El problema de tiempo final libre introduce un cambio cualitativo importante como es el hecho de que no se especifica la duración del periodo de planificación, siendo ésta una variable más del problema. El enunciado del problema de tiempo final libre es:

$$
\begin{aligned}
&Max \quad F(X,U) = \sum_{t=0}^{T-1} F_t(x_t,u_t)+F_T(x_T) \\
&s.a: \\
&\quad x_{t+1}-x_t = f_t(x_t,u_t) \qquad t=0,...,T-1 \\
&\quad u_t \in U_t \subset \mathbb{R}^m \qquad t=0,...,T-1 \\
&\quad x_t \in X_t \subset \mathbb{R}^n \qquad t=1,...,T \\
&\quad h(X,U) \in H \subset \mathbb{R}^s \\
&\quad x_0 = \bar{x}_0 \\
&\quad T \text{ libre}
\end{aligned}
\qquad \text{[VII]}
$$

La forma más fácil de abordar el problema es introducir T como una variable adicional y resolver el problema de tiempo final libre a través de la resolución de múltiples subproblemas de tiempo final fijo de la forma estándar, aunque cada uno de ellos con un valor distinto de T. En cada subproblema se encontrará una solución que depende de T. La solución al problema de tiempo final libre será la solución del subproblema para el cual se alcanza un mayor valor de la función objetivo, es decir, aquel valor T^* con su correspondiente solución que cumpla la desigualdad:

$$F(X^*(T^*),U^*(T^*)) \geq F(X^*(T),U^*(T))$$

para todo $T \in N$.

Este método de resolución que es evidente en teoría resulta inabordable en la práctica dado que T puede tomar infinitos valores. Por eso se debe acompañar de un estudio del valor de la función objetivo como función de la variable T.

I.3.9.- Enunciado del problema de tiempo infinito

El denominado problema de tiempo infinito aparece cuando el horizonte de planificación no es finito. Este problema es habitual en ciertas aplicaciones económicas, por ejemplo el problema de la explotación de recursos naturales, donde la obtención de la política óptima debe tener en cuenta las generaciones futuras. El enunciado del problema de horizonte infinito es:

$$
\begin{aligned}
&Max \quad F(X,U) = \sum_{t\in\mathbb{N}} F_t(x_t,u_t) + F_T(x_T) \\
&s.a: \\
&\quad x_{t+1} - x_t = f_t(x_t,u_t) \qquad t\in\mathbb{N} \\
&\quad u_t \in U_t \subset \mathbb{R}^m \qquad t\in\mathbb{N} \\
&\quad x_t \in X_t \subset \mathbb{R}^n \qquad t\in\mathbb{N} \\
&\quad h(X,U) \in H \subset \mathbb{R}^s \\
&\quad x_0 = \bar{x}_0
\end{aligned}
\qquad \text{[VIII]}
$$

con la particularidad de que las variables del problema forman parte de un espacio de dimensión infinita, es decir, $(X,U)=\{(x_t,u_t);t\in N\}$.

Abordar la resolución de este tipo de problemas requiere tratar con espacios de dimensión infinita dado que hay infinitas variables, lo que deja a este problema al margen de los anteriores. Sin embargo, se puede aproximar la solución del problema a través de la resolución de subproblemas de la forma estándar para

valores T_k concretos que formen parte de una sucesión $\{T_k;\ T_k \in \mathbb{N}\}$ que tienda a infinito. Para cada T_k aparece una solución óptima $(X^*(T_k), U^*(T_k))$ y un valor de la función objetivo $F(X^*(T_k), U^*(T_k))$. Por tanto, a medida que T_k va tomando valores se podrá construir una sucesión de soluciones óptimas y una sucesión de valores óptimos de la función objetivo.

La clave para poder encontrar la solución del problema de tiempo infinito es que los valores óptimos de la función objetivo que van apareciendo para cada valor de T_k formen una sucesión convergente. En ese caso el límite de la sucesión de soluciones forma la solución del problema de tiempo infinito, es decir:

$$(X^*, U^*) = \lim_{T_k \to \infty} (X^*(T_k), U^*(T_k))$$

$$F(X^*, U^*) = \lim_{T_k \to \infty} F(X^*(T_k), U^*(T_k))$$

Las condiciones para que esto ocurra se recogen en los teoremas de existencia de solución para el problema de control óptimo discreto con tiempo infinito, entre los que se puede citar los que aparecen en Keerthi y Gilbert (1985). Alternativamente, se pueden seguir métodos de resolución basados en los multiplicadores de Lagrange (Dechert, 1982) o establecer condiciones necesarias y suficientes de óptimo sin reducir el problema a una sucesión de problemas de tiempo finito (Michel, 1990).

I.4.- PROGRAMACIÓN DINÁMICA EN TIEMPO DISCRETO

I.4.1.- Fundamentos de la Programación Dinámica

La Programación Dinámica (P.D.) es una metodología más que una técnica para resolver problemas. Su característica definitoria es la de abordar un problema de varias variables a través de su descomposición en subproblemas de una sola variable. Esta descomposición permite llegar a establecer una "ecuación de optimalidad" cuya solución es la del problema original. La forma de resolver esta ecuación no es una característica propia de la P.D., sino que se debe resolver por cualquier técnica que resulte apropiada para ese problema en concreto. Así pues, la metodología de la P.D. consiste en:

1) Descomponer el problema original y plantear la "ecuación de optimalidad".

2) Resolver dicha ecuación.

La formulación original de la P.D. se debe a Isaac y Bellman (desde el año 1951) quienes la desarrollaron como una técnica para aplicar en problemas de optimización, de ahí términos como el de "ecuación de optimalidad", "principio de optimalidad", "función valor óptimo", etc. El libro básico de partida sobre P.D. es el de Bellman (1957) donde se recoge las primeras formulaciones de este método.

La visión inicial de la P.D. como una técnica de optimización se ha ido superando poco a poco y hoy en día se considera más como una metodología que se puede aplicar también a problemas que no son de optimización[4].

Según esta metodología un problema se puede abordar a través de la

[4]Entre los manuales recientes que insisten más en el aspecto metodológico que en el instrumental de la Programación Dinámica se puede citar a Sniedovich (1992).

programación dinámica si se puede formular mediante una ecuación de óptimo que se obtiene aplicando el siguiente principio de optimización condicional: "Un problema con dos variables de decisión se puede descomponer en dos problemas, cada uno de ellos con una variable de decisión". En el caso de problemas de optimización este principio se puede establecer de la siguiente forma:

$$Opt_{(x,y)\in Z\subset(X\times Y)} \; F(x,y) = Opt_{x\in X'} \, (Opt_{y\in Y'(x)} \; F(x,y))$$

donde X e Y son los conjuntos de oportunidades de las variables x e y; Z es el conjunto de puntos factibles, y se han definido los conjuntos:

$$X' = \{ \, x\in X \; / \; (x,y)\in Z, \; y\in Y \, \}$$
$$Y'(x) = \{ \, y\in Y \; / \; (x,y)\in Z \, \}$$

Es decir, el problema de P.D. se caracteriza por su descomposición en varios problemas con menos variables de decisión. La forma general de abordarlo tiene ahora dos etapas:

1ª. Para cada $x\in X'$, optimizar $F(x,y)$ para la variable y; y obtener $y=f(x)$.

2ª. Con ese valor de y sustituir en la función objetivo y calcular el valor óptimo de x (x^*) para $x\in X'$, valor que determina a su vez $y^*=f(x^*)$.

Esto es sólo un método general de resolver el problema, la forma concreta de obtener los valores óptimos no es relevante para caracterizar a un problema como de P.D., es más, un método de optimización puede ser apropiado para unos casos y no para otros, todo depende de la naturaleza del problema a resolver.

I.4.2.- Enunciado y tratamiento del problema

En el contexto de este trabajo interesa estudiar, aunque sea brevemente, cómo se aborda el problema de optimización dinámica en tiempo discreto desde el punto de vista de la P.D..

El problema de optimización dinámica que tenemos planteado se trata bajo el método de la P.D. como un problema de decisión multietápica, esto es, un problema en el que la decisión tomada en un periodo depende de la tomada en periodos anteriores y afecta, a su vez, a la de periodos posteriores. Este tipo de problemas admiten la descomposición en varios subproblemas, cada uno de ellos en un periodo de tiempo y admite, por tanto, la aplicación de la metodología propia de la P.D.. Partamos, por tanto, del siguiente problema:

$$Max \quad F(X,U) = \sum_{t=0}^{T-1} F_t(x_t,u_t)+F_T(x_T)$$

$$s.a:$$

$$x_{t+1}-x_t = f_t(x_t,u_t) \qquad t=0,...,T-1 \qquad [IX]$$

$$u_t \in U_t \subset \mathbb{R}^m \qquad t=0,...,T-1$$

$$x_t \in X_t \subset \mathbb{R}^n \qquad t=0,...,T$$

$$x_0 = \bar{x}_0$$

Para llegar a establecer la ecuación de óptimo hay que observar que el sistema dinámico determina las variables de estado una vez elegidos los valores de los controles para cada valor inicial x_0. Como consecuencia, la función objetivo, una vez fijado el valor inicial, sólo depende de las variables de control; esto permite formular las siguientes igualdades:

$$Max_{u_t \in U'_t} \quad F_T(x_T) + \sum_{t=0}^{T-1} F_t(x_t, u_t) =$$

$$= Max_{u_0 \in U'_0, \ldots, u_{T-1} \in U'_{T-1}} \quad F_T(x_T) + F_0(x_0, u_0) + \ldots + F_{T-1}(x_{T-1}, u_{t-1}) =$$

$$= Max_{u_0 \in U'_0} \quad [F_0(x_0, u_0) + \quad Max_{u_1 \in U'_1, \ldots, u_{T-1} \in U'_{T-1}} \quad F_T(x_T) + \sum_{t=1}^{T-1} F_t(x_t, u_t)]$$

siendo:

$$U_t^{/} = \{ u_t \in U_t \; / \; x_{t+1} = x_t + f_t(x_t, u_t) \in X_{t+1} \}$$

De esta forma queda establecida la "ecuación de optimalidad" siguiente:

$$Max_{u_0, \ldots, u_{T-1}} \quad F_T(x_T) + \sum_{t=0}^{T-1} F_t(x_t, u_t) =$$

$$= Max_{u_0} \quad [F_0(x_0, u_0) + \; Max_{u_1, \ldots, u_{T-1}} \quad F_T(x_T) + \sum_{t=1}^{T-1} F_t(x_t, u_t)]$$

ecuación que se puede ir extendiendo sin más que repetir el proceso sobre el problema de maximización que queda entre corchetes y seguir así hasta llegar al último periodo.

Se puede apreciar que la ecuación obtenida responde al conocido "Principio de optimalidad" de Bellman con el que este autor resume el método de la P.D.. Efectivamente, si nos situamos en un periodo de tiempo determinado, la "ecuación de optimalidad" hallada exige que las decisiones sean óptimas con independencia de la elección hecha en periodos anteriores.

La forma de llegar desde la ecuación a la solución es un problema que suele ser difícil de resolver y depende del tipo de problema concreto. Este, sin embargo, no es el objetivo del presente trabajo.

I.4.3.- Comparación entre la P.D. y las técnicas variacionales

Para acabar este capítulo es conveniente plasmar las diferencias y semejanzas que existen entre las tres técnicas citadas: la Programación Dinámica, por una parte; y las técnicas variacionales (Cálculo de Variaciones y Control Óptimo), por otra. En Tapiero (1977) se destaca que las técnicas variacionales son métodos equivalentes mientras no aparezcan restricciones de desigualdad, sin embargo existen importantes diferencias entre estas técnicas y la P.D.. Tapiero destaca el hecho de que las técnicas variacionales llevan a óptimos locales (aunque bajo ciertas condiciones pueden ser óptimos globales) mientras que la P.D. utiliza un principio que consigue óptimos globales; sin embargo, la aplicación de las técnicas para obtener el valor de las variables depende de cada problema mientras que en los métodos variacionales se sigue el mismo procedimiento. Se argumenta, por tanto, que cada problema de P.D. y su solución son una teoría en sí mismos mientras que esto no ocurre en los métodos variacionales. Finalmente, destaca que la P.D. es un método más general mientras que los otros no se pueden aplicar en ciertos casos de problemas en tiempo discreto, cuestión que se verá con más detalle en el próximo capítulo cuando se estudie detenidamente la teoría del Control Óptimo en tiempo discreto.

Existen otras diferencias importantes entre ambos métodos. Una ventaja de la P.D. es su alto rango de aplicabilidad, dado que no tienen porque aparecer supuestos sobre la función objetivo ni sobre el subconjunto de soluciones factibles. Por contra, resulta imposible estandarizar los métodos numéricos de resolución basados en el "principio de optimalidad" porque se resuelve según sea el problema. Por otra parte, y de cara a la eficiencia en la resolución, el hecho de que la dimensión del problema se reduzca es sólo una ventaja aparente ya que la contrapartida es que las funciones que van apareciendo (cada vez de menos variables) son progresivamente más difíciles de manejar y contienen más parámetros.

En cambio, los métodos basados en la teoría del Control Óptimo sí que se pueden estandarizar, sobre todo si el problema se enfoca a través de la P.N.L., donde existe *software* muy avanzado. Si se utilizan algoritmos específicos de C.O. discreto, que son todavía incipientes, se aumenta la eficiencia, aunque en última instancia todo depende del problema a resolver. En una posición conciliadora se sitúa Chow (1992), quien desarrolla una técnica numérica que se basa en el método de multiplicadores de Lagrange y que toma algunos elementos característicos tanto de la P.D. como del Control Óptimo.

CAPÍTULO II:

LA TEORÍA DEL CONTROL ÓPTIMO EN TIEMPO DISCRETO

II.1.- INTRODUCCIÓN

En el epígrafe I.3.6. se ha establecido un enunciado estándar del problema de C.O. en tiempo discreto seguido de otros enunciados que por unos motivos u otros hemos creído conveniente plasmar. El objetivo de este capítulo es obtener los métodos matemáticos analíticos que se utilizan para resolver los problemas allí planteados. Dedicaremos un capítulo posterior a complementar este análisis a través del estudio de los métodos numéricos y su concreción en algoritmos de resolución.

Desde los primeros trabajos sobre control óptimo discreto se pueden observar dos grandes aproximaciones al problema: una, a través de la Programación No Lineal (P.N.L.); y otra, a través del Principio del Máximo discreto (P.M.)[5]. La interrelación entre ambas aproximaciones fue brevemente comentada por Holtzman en un artículo (Holtzman, 1966) y más claramente establecida a medida que se producían nuevas investigaciones (se puede destacar el libro de Canon, Cullum y Polack, 1970; el de Boltyanski, 1978; y artículos como los de Nahorski y Ravn, 1988a; Lin y Yang, 1989; y Ravn, 1990, 1991). El desarrollo posterior tanto de la Programación Matemática como de la Teoría del Control Óptimo permite establecer interrelaciones más enriquecedoras.

En el epígrafe II.2. se parte de establecer la equivalencia entre el enunciado del problema de C.O. discreto y el de extremos de funciones en un subconjunto del espacio euclídeo, lo cual justifica la aplicación de las técnicas de la Programación Matemática (y en especial de la P.N.L.) al problema objeto de estudio en el caso en que se den condiciones de diferenciabilidad y el subconjunto quede definido a través de igualdades y/o desigualdades. Estas técnicas se recogen brevemente en el siguiente epígrafe. A continuación, en el epígrafe II.4., se estudia el P.M. discreto por ser la técnica más específica de la Optimización Dinámica en tiempo discreto. Este capítulo acaba destacando algunas interrelaciones entre ambas aproximaciones.

---

[5]Entre los trabajos pioneros que utilizan un P.M. discreto se puede citar el de Rozonoer (1959). Entre los primeros trabajos que siguen la aproximación a través de la P.N.L., en Holtzman (1966) se cita uno de J.B. Rosen publicado en 1957.

La equivalencia entre los dos problemas es una característica que se explotará más adelante para realizar algunas extensiones del problema de C.O. discreto (capítulo III) y para aplicarle métodos numéricos de la P.N.L. (capítulo IV). En definitiva, esta equivalencia permite aplicar las técnicas analíticas y numéricas ya muy desarrolladas de la P.N.L. al problema de C.O. discreto, siendo esta la vía más eficiente, a nuestro juicio, de enfocar el problema.

II.2. EQUIVALENCIA ENTRE EL ENUNCIADO DE UN PROBLEMA DE EXTREMOS DE UNA FUNCIÓN Y EL DE C.O. DISCRETO

En este epígrafe se demuestra la equivalencia entre un problema de C.O. discreto enunciado en forma estándar y un problema de extremos de una función. El Problema [I] es el de C.O. discreto de partida.

$$
\begin{aligned}
&Max \quad F(X,U)=\sum_{t=0}^{T-1} F_t(x_t,u_t)+F_T(x_T)) \\
&s.a: \\
&x_{t+1}-x_t=f_t(x_t,u_t) \qquad t=0,...,T-1 \\
&u_t \in U_t \subset \mathbb{R}^m \qquad t=0,...,T-1 \\
&x_t \in X_t \subset \mathbb{R}^n \qquad t=0,...,T \\
&h(X,U) \in H \subset \mathbb{R}^s \\
&X_o=\{\bar{x}_0\} \\
&T \text{ conocido}
\end{aligned}
\qquad [I]
$$

El primer paso para demostrar la equivalencia es definir un único vector de variables que englobe a todas las variables de estado y de control, es decir:

$$z=(X,U)=(x_0,...,x_T,u_0,...,u_{T-1}) \in \mathbb{R}^r \qquad (1)$$

La notación es la misma que en el capítulo anterior, esto es:

$$r=(T+1)\cdot n+T\cdot m$$

$$x_t=(x_t^1,...,x_t^n)\in\mathbb{R}^n \qquad t=0,...,T$$

$$u_t=(u_t^1,...,u_t^m)\in\mathbb{R}^m \qquad t=0,...,T-1$$

En segundo lugar, el conjunto de oportunidades, sin tener en cuenta el sistema dinámico, se puede construir como:

$$S=(X_0\times...\times X_T\times U_0\times...\times U_{T-1})\cap D \qquad (2)$$

donde nuevamente D es el subconjunto de puntos de $\mathbb{R}^r$ que cumplen las restricciones de trayectoria. A su vez el conjunto S se puede suponer que se descompone como intersección de dos conjuntos:

1°) Un conjunto definido por restricciones de desigualdad del tipo mayor o igual[6], $G(z)\geq 0$; y,

2°) Un subconjunto $Z\in\mathbb{R}^r$ que no se puede definir a través de igualdades o desigualdades.

Por tanto, el conjunto S admite la siguiente expresión:

$$S = \{ z\in\mathbb{R}^r \mid G(z)\geq 0,\ z\in Z \} \qquad (3)$$

donde G: $\mathbb{R}^r \longrightarrow \mathbb{R}^p$ es una función vectorial cuyas funciones componentes son $G_l(z)$, $l=1,2,...,p$; si se supone que hay p restricciones.

[6]Esto no supone ninguna pérdida de generalidad ya que las restricciones de igualdad o del tipo menor o igual se pueden pasar a restricciones del tipo mayor o igual con transformaciones simples.

En tercer lugar, el sistema dinámico, $x_{t+1}-x_t=f_t(x_t,u_t)$, se puede releer como un conjunto de Tn restricciones de igualdad que, utilizando una notación más general, se puede escribir como:

$$Q(z)=0 \qquad (4)$$

donde se tiene que Q: $\mathbb{R}^r \longrightarrow \mathbb{R}^{Tn}$, siendo:

$$Q(z)=(Q_0(z),...,Q_t(z),...,Q_{T-1}(z))$$

$$Q_t(z)=(Q_t^1(z),...,Q_t^i(z),...,Q_t^n(z)) \qquad t=0,...,T-1$$

$$Q_t^i(z)=x_{t+1}^i-x_t^i-f_t^i(x_t,u_t) \qquad i=1,...,n$$

Por último, la función objetivo se puede representar simplemente como:

$$F(z)$$

utilizando la definición del vector z dada en (1).

Haciendo todas estas transformaciones sobre el Problema **[I]** se pasa al Problema **[I']**.

$$\begin{aligned} &Max \quad F(z) \\ &s.a: \\ &\quad G(z)\geq 0 \\ &\quad Q(z)=0 \\ &\quad z\in Z \end{aligned} \qquad [I']$$

Este problema responde al enunciado de un problema de extremos de una función en un subconjunto del espacio euclídeo r-dimensional. Si además las funciones F, G y Q son diferenciables (en general no lineales) y Z es un subconjunto abierto no vacío de $\mathbb{R}^r$ se puede interpretar este problema como uno de Programación no Lineal.[7]

De esta manera queda demostrado que un problema de C.O. discreto se reduce a uno de extremos. Para completar la equivalencia a que hace referencia el título de este epígrafe hay que demostrar ahora que un problema de extremos de funciones se puede enunciar como uno de C.O. discreto. Para ello se toma el Problema **[II]** como el problema de extremos de partida.

$$\begin{aligned} &Max \quad F(y) \\ &s.a: \\ &\quad G(y)\geq 0 \\ &\quad y\in Y\subset\mathbb{R}^q \end{aligned} \qquad [II]$$

donde $G:\mathbb{R}^q \longrightarrow \mathbb{R}^p$.

[7]Para no recargar excesivamente la notación, prescindiremos de especificar el dominio de definición de las funciones como un subconjunto del espacio inicial sobre el que se construyen. No obstante, al plantear los conjuntos de oportunidades sobre las variables argumento de tales funciones siempre habrá que asumir que dichos conjuntos son compatibles con los dominios de las funciones.

Dado que este es un problema estático, basta considerarlo un problema dinámico con un solo periodo ($T=1$). Además se considera que y es el vector de variables de control y se toma una única variable de estado (ficticia) con valor constante cero. Para llegar al enunciado de un problema de C.O. óptimo discreto hay que plantear las siguientes equivalencias:

$$X=(x_0,x_1)\in\mathbb{R}^2$$

$$U=u_0=y\in\mathbb{R}^q$$

$$F(X,U)=F(y)$$

$$f_0=0$$

$$U_0=\{u_0\in\mathbb{R}^q/G(u_0)\geq 0,u_0\in Y\}$$

$$X_0=\{0\}\ ,\ D=\mathbb{R}^{q+2}$$

Se consigue por tanto el Problema [II'] con la estructura del Problema [I] de C.O. discreto en forma estándar:

$$\begin{aligned} &Max \quad F(x_0,x_1,u_0)=F(u_0)\\ &s.a:\\ &\quad x_1-x_0=0 \qquad\qquad [II']\\ &\quad u_0\in U_0\subset\mathbb{R}^q\\ &\quad x_0=0 \end{aligned}$$

Tras demostrar esta equivalencia surge la siguiente pregunta: ¿Es necesario construir una teoría propia para resolver el problema de C.O. discreto o este esfuerzo resulta innecesario dado que se puede considerar un caso especial de la teoría de extremos de funciones?. Distintos autores se han hecho esta misma pregunta y, en general, se decantan por la opinión de construir una teoría propia.

A pesar de esta equivalencia, hay que hacer constar, como se resalta en Borrell (1985) que el problema de control tiene ciertas peculiaridades que se derivan del hecho de que sus funciones tienen una estructura dinámica y esto se debe aprovechar a la hora de especificar las condiciones de óptimo por lo que no se debe englobar el problema de C.O. discreto en la teoría de extremos de funciones, es decir, se debe sacar ventaja de la dinamicidad del problema tanto para expresar las condiciones de óptimo como para plantearse cuestiones específicamente dinámicas que no tendrían sentido en un problema de extremos más general. Boltyanskii (1978) también pone de relieve estas consideraciones y añade que muchas aplicaciones requieren que el problema se formule como un problema donde las variables de estado se controlen por variables de decisión; además puede servir para obtener una variante simplificada del problema continuo.

En nuestra opinión, abandonar el tratamiento del problema como uno específico y subsumirlo en uno de extremos puede resultar empobrecedor, tanto en lo que se refiere a la interpretación del problema, como a la limitación que ello supondría para buscar algoritmos más eficientes de resolución del problema. Efectivamente, si se desarrolla la teoría del C.O. discreto, creemos que la forma separable de la función objetivo y la forma tan estructurada de las restricciones (donde cada variable de estado y de control entra a formar parte de un número muy reducido de restricciones) favorecerá la búsqueda de métodos numéricos de resolución más eficientes que los existentes, por ejemplo, en P.N.L..

No obstante, lo que sí resulta conveniente es aprovechar la equivalencia entre ambos problemas para validar las condiciones necesarias y suficientes de máximo en el problema de control a través de las respectivas condiciones en P.N.L.. Además, mientras se investiga en los métodos numéricos específicos de este problema, siempre se puede utilizar los paquetes informáticos existentes en el mercado propios de la P.N.L.; cosa que, hoy por hoy, parece lo más eficiente. Por estas consideraciones, es conveniente dedicar un epígrafe a la teoría más amplia de extremos de funciones y más concretamente a la P.N.L..

II.3. PROGRAMACIÓN NO LINEAL

II.3.1. Aspectos generales de la P.N.L.

Los Problemas [III] y [III'] corresponden a uno de extremos de una función en un subespacio del espacio euclídeo n-dimensional:

$$\begin{aligned} &Max \quad F(z) \\ &s.a: \\ &\quad g(z)\geq 0 \\ &\quad z\in Z \end{aligned} \qquad [III]$$

$$\begin{aligned} &Max \quad F(z) \\ &s.a: \\ &\quad z\in C \end{aligned} \qquad [III']$$

donde:

$$C = \{ z\in\mathbb{R}^n \,/\, g(z)\geq 0,\ z\in Z \}$$

siendo F y g funciones del tipo:

$$F: \mathbb{R}^n \longrightarrow \mathbb{R} \ \text{ y } \ g: \mathbb{R}^n \longrightarrow \mathbb{R}^m$$

En Takayama (1985) se resume en cinco las cuestiones que se pueden plantear a la hora de abordar la resolución del problema:

1ª) ¿Es el conjunto de oportunidades C no vacío, es decir, existe algún punto factible?.

2ª) ¿Existe solución óptima, es decir, un punto factible z^* tal que $F(z^*) \geq F(z)$ $\forall z \in C$?

3ª) ¿Cuáles son las características de ese punto z^*?.

4ª) ¿Porqué es la solución óptima?.

5ª) ¿Existen algoritmos para encontrar todas las soluciones?.

Las dos primeras cuestiones hacen referencia a la existencia de solución óptima, para lo cual se enuncian teoremas que aseguran la existencia de máximo (mínimo) de una función real definida en un subconjunto de $\boldsymbol{R}^n$. El teorema clásico de existencia de máximo y mínimo es el teorema de Weierstrass:

> *Teorema II.1.:* Si la función $F(z)$ es continua en un subconjunto C compacto, la función admite máximo y mínimo en C.

Las hipótesis del teorema II.1. no se pueden relajar más si se quiere asegurar la existencia de máximo y mínimo, sin embargo la hipótesis de continuidad se puede relajar por la de semicontinuidad superior en un problema de maximizar o por la de semicontinuidad inferior en un problema de minimizar. Por ello el teorema de existencia anterior puede quedar de la siguiente manera:

> "Si la función $F(z)$ es semicontinua superiormente en el conjunto compacto C existe máximo del problema en C."

La cuestión 3ª y colateralmente la 4ª son las más importantes a nivel teórico y las que han ocupado una mayor parte en la literatura sobre optimización matemática, sobre todo desde que autores como Fritz John (en 1948) y Kuhn y

Tucker (en 1951) aportaron diversos tipos de características que deben cumplir los puntos óptimos. A raíz de sus trabajos fueron apareciendo muchos otros en la línea de refinar las condiciones de óptimo de forma que en los años 70 se llegaron a establecer condiciones de óptimo más relajadas para el problema de extremos. Un resumen de este proceso se desarrollará más adelante y se puede seguir en algunos manuales clásicos de programación no lineal[8].

Por último, la 5ª cuestión, es decir, la búsqueda de algoritmos para encontrar todas las soluciones, ha sido muy tratada a raíz de la obtención del algoritmo del símplex en programación lineal. Esta es la línea de investigación más seguida actualmente en programación no lineal, debido en parte al desarrollo de ordenadores cada vez más potentes que permiten, a través de paquetes informáticos basados en alguno de estos algoritmos, resolver problemas de programación no lineal. En el capítulo cuarto se comentan algunos de estos algoritmos aplicados al problema de control óptimo discreto.

Veamos a continuación algunas consideraciones teóricas que hay que tener presentes a la hora de resolver el Problema **[III]**.

El problema de encontrar el máximo absoluto de un problema de P.N.L. se desarrolló, inicialmente por parte de Karush en 1939[9], como una ampliación de la programación clásica (donde todas las restricciones son de igualdad). Este autor reduce las restricciones de desigualdad a igualdades mediante la resta del cuadrado de un número real:

[8]Entre los muchos manuales sobre P.N.L. merece la pena destacar, entre los pioneros, a Mangasarian (1969) debido a su rigurosidad y claridad al mismo tiempo; y entre los más recientes a Bazaraa, Sherali y Shetty (1993), que incorpora también métodos numéricos de resolución.

[9]El estudio de este autor se recoge en su tesis doctoral y sólo se publicó en la Universidad de Chicago de ahí que permaneciera bastante ignorado hasta finales de los años 60 (un pequeño resumen de la evolución de los enfoques dados a la P.N.L. se puede seguir en Takayama, 1985)

$$g_j(z)-\delta_j^2=0 \qquad j=1,...,m$$

y a continuación aplica el teorema clásico de Lagrange tomando δ_j como variables adicionales del problema. La obtención de las condiciones de óptimo se basa, igual que en programación clásica, en el cálculo diferencial: derivabilidad y diferenciabilidad de funciones y teorema de la función implícita.

Otra aproximación al Problema **[III]** se puede realizar a través de desigualdades de punto de silla utilizando la teoría de conjuntos convexos, funciones cóncavas y convexas y teoremas de separación. En esta aproximación no se exige diferenciabilidad de las funciones que intervienen en el problema y por tanto es una aproximación más general. Según este enfoque la condición de punto de silla es suficiente para obtener un máximo global pero sólo es necesaria bajo hipótesis adicionales de convexidad y de ciertas condiciones de regularidad sobre las restricciones llamadas también cualificación de restricciones. Una limitación de esta aproximación es la poca operatividad de las desigualdades de punto de silla de cara a la obtención de los puntos concretos que las satisfacen, aunque en problemas de C.O. discreto la condición de punto de silla puede dar lugar a un principio del máximo como se hace en Valqui (1987). Para que las condiciones de punto de silla ganen en eficacia hay que introducir la hipótesis de diferenciabilidad.

Efectivamente, si se parte de funciones diferenciables y de un conjunto Z abierto y no vacío se llega a la aproximación más conocida y más desarrollada en los años 50 y 60: la Programación no Lineal. Esta aproximación da lugar a un conjunto de igualdades y desigualdades llamadas condiciones de cuasi-punto de silla (Takayama, 1985) o condiciones de punto estacionario (Mangasarian, 1969) que se deducen de las desigualdades de punto de silla y de la característica de diferenciabilidad de las funciones, recíprocamente las condiciones de punto estacionario implican las desigualdades de punto de silla bajo hipótesis de

convexidad o concavidad[10].

En P.N.L. Fritz John en 1948 y con más éxito Kuhn y Tucker en 1951, formularon inicialmente las condiciones de óptimo bajo hipótesis de diferenciabilidad. La divergencia entre ambos autores se encuentra en la definición de la función lagrangiana, lo que a su vez lleva a la exigencia de cualificación de restricciones para Kuhn y Tucker sin que esto aparezca en Fritz John[11]. A partir de estas aportaciones fueron apareciendo en la década de los 60 distintos trabajos en la línea de relajar los supuestos bajo los cuales las condiciones de punto estacionario de Kuhn y Tucker son necesarias y/o suficientes. Este relajamiento se produjo introduciendo conceptos como el de pseudoconcavidad y cuasiconcavidad de funciones, conceptos definidos tras la publicación del citado artículo de Kuhn y Tucker. Por su importancia de cara a la resolución de nuestro problema de control, en el siguiente epígrafe mencionaremos las aportaciones de algunos autores al caso particular en que aparecen restricciones mixtas, esto es, de igualdad y desigualdad.

Establezcamos a continuación las técnicas de la programación no lineal tomando como base las aportaciones de Kuhn y Tucker.

Sea de nuevo el Problema **[III]** o **[III']** donde las funciones F y g se suponen diferenciables respecto a z en Z y se parte de que el conjunto Z es abierto y no vacío. El objetivo es obtener el máximo global de dicho problema.

---

[10]Dado que el objetivo de este trabajo no es realizar un análisis riguroso de la P.N.L. no creemos conveniente demostrar aquí estas implicaciones, que se puede ver en alguno de los manuales citados.

[11]Las condiciones de punto estacionario de Fritz John no necesitan ninguna cualificación de restricciones para ser condiciones necesarias de máximo global porque se parte de una función lagrangiana del tipo:

$$L(z,\lambda_0,\lambda)=\lambda_0 F(z)+\lambda g(z)$$

En realidad la cualificación de restricciones de Kuhn y Tucker garantiza que $\lambda_0 \neq 0$ lo cual permite pasar de las condiciones de Fritz John a las de Kuhn y Tucker.

Definición II.1. (M.G.): **Máximo global** del Problema **[III]** es aquel punto $z^* \in C$ que cumple:

$$F(z^*) \geq F(z) \quad \forall\, z \in C$$

Para llegar a encontrar este punto a través de ciertas características, Kuhn y Tucker expresaron las condiciones de punto estacionario o simplemente "condiciones de Kuhn y Tucker" que se enuncian a continuación:

Definición II.2. (K-T): Un punto $z' \in Z$ cumple las **condiciones de Kuhn y Tucker** si existe un vector $\lambda' \in \mathbb{R}^m$ tal que:

$$\nabla F(z') + \lambda' \nabla g(z') = 0 \qquad (K\text{-}T,\ a)$$

$$g(z') \geq 0 \qquad (K\text{-}T,\ b)$$

$$\lambda' g(z') = 0 \qquad (K\text{-}T,\ c)$$

$$\lambda' \geq 0 \qquad (K\text{-}T,\ d)$$

La condición *(K-T, a)*, llamada de estacionariedad, se puede escribir también en términos de la función lagrangiana de programación clásica de la siguiente manera:

$$\nabla_z L(z', \lambda') = 0$$

$$[\ L(z,\lambda) = F(z) + \lambda g(z)\ ]$$

Tal condición, junto con las otras tres, *(K-T, b)* de factibilidad, *(K-T, c)* de holgura complementaria, y *(K-T, d)* de no negatividad, no son por sí solas necesarias ni suficientes para llegar a *M.G.*, por eso se necesitan condiciones adicionales. Las siguientes definiciones y teoremas aclaran bajo qué supuestos un punto que satisface

K-T cumple también ***M.G.*** (***K-T*** son suficientes) y, recíprocamente, bajo qué otros supuestos un punto que cumple ***M.G.*** cumple también ***K-T*** (***K-T*** son necesarias).

Definición II.3.: Una **restricción** $g_j(z) \geq 0$ se dice que es **activa** en un punto z' si se cumple con igualdad: $g_j(z')=0$.

Teorema II.2.: Sean F y las componentes de g funciones diferenciables respecto a z en Z. Sea g_A el subvector de g formado por las restricciones activas en un punto z'. Entonces si F y las componentes de g_A son funciones cóncavas respecto de z en Z, ***K-T*** implica ***M.G.***.

Definición II.4. (C.R.): Un punto $z' \in C$ satisface la **cualificación de restricciones** si se cumple una de las dos siguientes condiciones:

a) z' pertenece al conjunto interior de C, es decir, no cumple ninguna restricción con igualdad (ninguna es activa).

b) Si se saturan algunas restricciones en z', es decir,

$$g_j(z')=0 \quad \forall j \in A \subset \{1,2,...,m\}$$

entonces para cualquier punto $z'' \in Z$ que cumpla

$$\nabla g_j(z')(z''-z') \geq 0 \quad \forall j \in A$$

debe existir una función

$$h(t):[0,1] \mapsto Z$$

diferenciable en 0 con las siguientes condiciones:

i) $h(0)=z'$
ii) $h(t)\in C$
iii) $\nabla h(0)=\alpha(z''-z')$ *para algún* $\alpha>0$

Teorema II.3.: Si un punto satisface ***(M.G.)*** y en ese punto se cumple la cualificación de restricciones entonces también se cumple ***(K-T)***.

La dificultad que conlleva la verificación de esta condición ***(C.R.)*** llevó a varios autores a enumerar condiciones que pudieran sustituirla en el sentido que también aseguraran que las condiciones ***(K-T)*** fueran necesarias de óptimo global. Algunas de estas nuevas cualificaciones de restricciones son tan poco operativas como la de Kuhn y Tucker[12], sin embargo otras son más fáciles de comprobar y por tanto más importantes desde un punto de vista operativo[13]. De todas ellas hay que destacar la llamada condición de regularidad, que se recoge en la definición II.5..

Definición II.5.: Un punto cumple la **condición de regularidad** o de rango completo si en ese punto el rango de la matriz jacobiana de las funciones g_A es igual al número de restricciones saturadas, en otras palabras, si los gradientes de las restricciones saturadas en ese punto forman un sistema de vectores linealmente independientes.

Las hipótesis de los teoremas II.2. y II.3. se resumen en la Figura II.1.

[12]Por ejemplo, las de Slater y Karlin (Mangasarian, 1969).

[13]Entre éstas se puede citar la de Arrow-Hurwicz-Uzawa, la de Abadie y la de Evans; todas ellas recogidas en Mangasarian (1969).

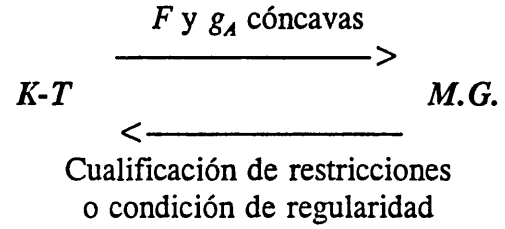

*Figura II.1: Hipótesis para que **K-T** sean condiciones necesarias y/o suficientes para M.G..*

II.3.2.- Incorporación de restricciones de igualdad

Como paso necesario de cara a establecer las condiciones necesarias y suficientes de óptimo en P.N.L. que sirvan para el problema de C.O. discreto conviene ver los efectos que se producen al considerar en el enunciado del problema restricciones específicas de igualdad, dado que en el problema de control el sistema dinámico se expresa a través de igualdades.

El enunciado correspondiente da lugar al Problema **[IV]** o **[IV']**.

$$\begin{aligned} &Max \quad F(z) \\ &s.a: \\ &\quad g(z) \geq 0 \\ &\quad q(z) = 0 \\ &\quad z \in Z \end{aligned} \qquad \textbf{[IV]}$$

$$\begin{aligned} &Max \quad F(z) \\ &s.a: \\ &\quad z \in C' \end{aligned} \qquad \textbf{[IV']}$$

donde:

$$C'=\{z\in\mathbb{R}^n / g(z)\geq 0,\ q(z)=0,\ z\in Z\}$$

siendo la nueva función q del tipo $q:\mathbb{R}^n \longrightarrow \mathbb{R}^p$.

Para relacionar el Problema [IV] con el Problema [III] basta con observar que una igualdad equivale a dos desigualdades:

$$q_k(z)=0 \leftrightarrow \begin{matrix} q_k(z)\geq 0 \\ -q_k(z)\geq 0 \end{matrix}$$

Debido a estas nuevas restricciones, se debe adaptar al nuevo enunciado las condiciones de Kuhn y Tucker y las hipótesis exigidas en los teoremas II.2. y II.3.. No obstante, todos los cambios se pueden realizar sin desdoblar las restricciones de igualdad para no aumentar todavía más la dimensión del problema. De esta manera los cambios que se introducen son los siguientes:

- En la función lagrangiana se deberían introducir $2p$ multiplicadores no negativos asociados a las nuevas $2p$ restricciones de desigualdad, pero esto es equivalente a introducir p multiplicadores no restringidos en signo asociados a las p nuevas restricciones de igualdad. Las condiciones de Kuhn y Tucker recogen las p igualdades adicionales.

- En el teorema de suficiencia de Kuhn y Tucker se exige a las restricciones de desigualdad que sean cóncavas, lo cual significa que las funciones $q_k(z)$ y $-q_k(z)$ deben ser cóncavas, de donde se llega a que $q_k(z)$ deben ser lineales.

- En la condición de regularidad y dado que las restricciones de igualdad se

saturan siempre, la matriz jacobiana afecta a las funciones componentes de g_A y de q.

Hechas estas observaciones se puede enunciar a continuación las condiciones de Kuhn y Tucker y los supuestos bajo los cuales dichas condiciones son necesarias y/o suficientes de máximo global en el problema con restricciones de igualdad.

Las condiciones de Kuhn y Tucker se obtienen definiendo previamente la lagrangiana:

$$L(z,\lambda,\mu)=F(z)+\lambda g(z)+\mu q(z)$$
$$con \quad \lambda\in\mathbb{R}^m \quad y \quad \mu\in\mathbb{R}^p$$

Ahora, aplicando la definición II.2., un punto $z'\in Z$ satisface las condiciones de Kuhn y Tucker o es punto estacionario de la lagrangiana si existen multiplicadores (λ,μ) de forma que se cumple simultáneamente:

i) $\nabla F(z')+\lambda\nabla g(z')+\mu\nabla q(z')=0$

ii) $g(z')\geq 0$

$q(z')=0$

iii) $\lambda g(z')=0$

iv) $\lambda\geq 0$

A partir de estas condiciones y de las observaciones anteriores se puede enunciar el siguiente teorema de suficiencia para el Problema **[IV]**:

> *Teorema II.4.:* Si F y las componentes de g_A son funciones cóncavas y las componentes de q son funciones lineales, todo ello respecto a z en Z, entonces si un punto cumple las anteriores condiciones de Kuhn

y Tucker es máximo global del problema.

Para que las condiciones i) a iv) de Kuhn y Tucker sean necesarias, se debe especificar la condición de regularidad que en el caso de restricciones de igualdad viene dada por la definición II.6.. A su vez, esta definición permite enunciar el teorema II.5..

Definición II.6.: Un punto cumple la **condición de regularidad en el Problema [IV]** si el rango de la matriz jacobiana de las funciones componentes de g_A y de q en ese punto es igual al número de restricciones de desigualdad saturadas en ese punto más el número de restricciones de igualdad.

Teorema II.5.: Un punto que es máximo global del Problema **[IV]** y que cumple la definición II.6. cumple también las condiciones i) a iv), de Kuhn y Tucker.

La condición de regularidad dada en la definición II.6. se puede sustituir por alguna otra cualificación de restricciones adaptada al caso de restricciones de igualdad. También se puede sustituir por otras cualificaciones específicas al caso de restricciones de igualdad y desigualdad, entre las que destaca la de Mangasarian y Fromovitz enunciada en 1967 (recogida también en Mangasarian, 1969) y que se reproduce en la definición II.7..

Definición II.7.: Un punto z' satisface la **cualificación de restricciones en el Problema [IV]** si cumple:

a) rango $(J(q(z')) = p$

b) existe un punto $y \in \mathbb{R}^n$ tal que:

$$J(q(z'))y = 0$$
$$J(g_A(z'))y < 0$$

donde J indica la matriz jacobiana.

La condición dada por la definición II.7. es menos fuerte que la condición de regularidad enunciada en la definición II.6 aunque también menos operativa. Sin embargo, es la condición que utilizan más recientemente autores como Ravn (Ravn, 1990, 1991).

Como último paso para establecer de forma definitiva las condiciones de óptimo hay que utilizar conceptos como el de cuasiconcavidad (enunciado por primera vez por Nikaido en 1954) y el de pseudoconcavidad (enunciado por Mangasarian en 1965) de una función. Estos conceptos son menos fuertes que el de concavidad y por tanto permiten enunciar los anteriores teoremas de forma que resulten aplicables a un conjunto de funciones más extenso. La relajación de los supuestos bajo los cuales se cumplen los teoremas ha ido produciéndose sucesivamente a lo largo de los años 60 y 70 utilizando estos conceptos más generales que el de concavidad. En Bazaraa y otros (1993) por ejemplo, se formulan los teoremas que recogen tales supuestos.

Aquí destacamos, mediante el teorema II.6., las hipótesis y resultados equivalentes al enunciado del teorema II.4..

> *Teorema II.6.:* Sean F, las componentes de g y las componentes de q funciones diferenciables respecto a z en el conjunto abierto y no vacío Z. Si existe un punto z' que cumple las condiciones i) a iv), de Kuhn y Tucker, y en ese punto F es pseudocóncava, las componentes de g_A son cuasicóncavas, las de q_V son cuasicóncavas y las de q_W son cuasiconvexas; entonces ese punto es máximo global del problema.

En este enunciado q_V y q_W son los subvectores de la función vectorial q asociados a un multiplicador μ mayor que cero y menor que cero respectivamente.

Algunas de las cualificaciones de restricciones existentes también se pueden enunciar de una forma más relajada utilizando estos conceptos, sin embargo no ocurre esto con la condición de regularidad (definición II.6.) ni con la condición de Mangasarian y Fromovitz (definición II.7.) que son las más utilizadas por su operatividad.

Para acabar este epígrafe se esquematiza en la Figura II.2. las implicaciones entre las condiciones de Kuhn y Tucker y la condición de máximo global teniendo en cuenta las restricciones de igualdad y los conceptos introducidos de cuasiconcavidad y pseudoconcavidad:

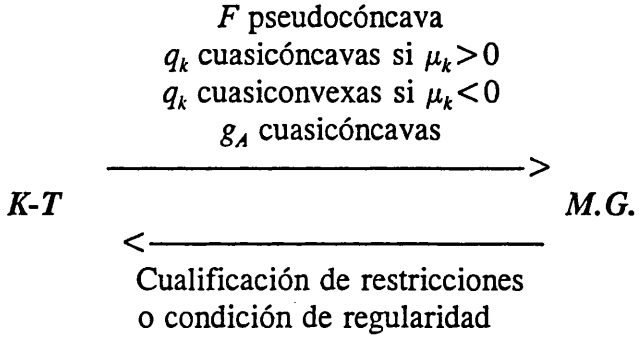

Figura II.2.: Hipótesis menos fuertes para que K-T sean condiciones necesarias y/o suficientes de M.G. en el problema con restricciones de igualdad.

II.3.3. Aplicación de las condiciones de óptimo de P.N.L. a un problema de C.O. discreto

Dado que se ha demostrado la equivalencia entre el problema de extremos de funciones y por tanto de Programación no Lineal y el problema de Control Óptimo discreto (epígrafe II.2.) se puede aplicar las condiciones de óptimo de la P.N.L. (epígrafes II.3.1. y II.3.2.) al problema de C.O. discreto en forma estándar enunciado en el capítulo I. Sea el Problema **[I]** de C.O. discreto en forma estándar:

$$Max \quad F(X,U)=\sum_{t=0}^{T-1} F_t(x_t,u_t)+F_T(x_T)$$

$$s.a:$$

$$x_{t+1}-x_t=f_t(x_t,u_t) \qquad t=0,...,T-1$$

$$u_t \in U_t \subset \mathbb{R}^m \qquad t=0,...,T-1 \qquad \textbf{[I]}$$

$$x_t \in X_t \subset \mathbb{R}^n \qquad t=0,...,T$$

$$h(X,U) \in H \subset \mathbb{R}^s$$

$$X_0=\{x_0\}$$

$$T \ \ conocido$$

El conjunto de oportunidades, sin tener en cuenta el sistema dinámico es:

$$S=(X_0 \times X_1 \times ... \times X_T \times U_0 \times ... U_{T-1}) \cap D$$

donde nuevamente D es el subconjunto de puntos de $\mathbb{R}^r$ que cumplen las restricciones de trayectoria. Se supone que el conjunto S se puede expresar a través de restricciones de desigualdad, de la condición inicial y de un conjunto Z abierto y no vacío que no se puede reducir a desigualdades, es decir:

$$S = \{(X,U)\in\mathbb{R}^r / \ g(X,U)\geq 0;\ x_0=\bar{x}_0;\ (X,U)\in Z\}$$

donde g es una función vectorial de p funciones componentes con r variables reales cada una ($r=(T+1)n+Tm$).

Por tanto, el Problema **[I]** de C.O. discreto también puede adoptar el enunciado del Problema **[V]** con $T(n+m)$ variables (dado que x_0 es conocido) y $Tn+p$ restricciones:

$$Max \quad F(X,U)=\sum_{t=0}^{T-1} F_t(x_t,u_t)+F_T(x_T)$$

$$s.a:$$

$$x_{t+1}-x_t=f_t(x_t,u_t) \quad t=0,\ldots,T-1 \qquad \textbf{[V]}$$

$$g(X,U)\geq 0$$

$$(X,U)\in Z$$

$$x_0 \ y \ T \ \ conocidos$$

El Problema **[V]** se puede tratar mediante P.N.L. donde:

- el vector de variables, $z\in\mathbb{R}^{T(n+m)}$, es:

$$z=(x_1^1,\ldots,x_1^n,\ldots,x_T^1,\ldots,x_T^n,u_0^1,\ldots,u_0^m,\ldots,u_{T-1}^1,\ldots,u_{T-1}^m)$$

- las restricciones de igualdad son:

$$q_t^i(z)=f_t^i(x_t,u_t)-x_{t+1}^i+x_t^i=0 \qquad t=0,\ldots,T-1 \ , \ i=1,\ldots,n$$

- las restricciones de desigualdad son:

$$g_l(z)\geq 0 \qquad l=1,...,p$$

- y la función lagrangiana adopta la expresión:

$$L(z,\lambda,\mu)=F(z)+\sum_{t=0}^{T-1}\mu_{t+1}(f_t-x_{t+1}+x_t)+\sum_{l=1}^{p}\lambda_l g_l$$

Ahora, en forma desarrollada, las condiciones de estacionariedad de Kuhn y Tucker para el Problema [V] son las siguientes:

$$\frac{\partial F_t}{\partial x_t^i}-\mu_t^i+\mu_{t+1}^i+\sum_{k=1}^{n}\mu_{t+1}^k\frac{\partial f_t^k}{\partial x_t^i}+\sum_{l=1}^{p}\lambda_l\frac{\partial g_l}{\partial x_t^i}=0 \qquad (5,\ a)$$

$$t=1,...,T-1\ ,\ i=1,...,n$$

$$\frac{\partial F_T}{\partial x_T^i}-\mu_T^i+\sum_{l=1}^{p}\lambda_l\frac{\partial g_l}{\partial x_T^i}=0 \qquad i=1,...,n \qquad (5,\ b)$$

$$\frac{\partial F_t}{\partial u_t^j}+\sum_{k=1}^{n}\mu_{t+1}^k\frac{\partial f_t^k}{\partial u_t^j}+\sum_{l=1}^{p}\lambda_l\frac{\partial g_l}{\partial u_t^j}=0 \qquad (5,\ c)$$

$$t=0,...,T-1\ ,\ j=1,...,m$$

Igualmente aparecen las condiciones de factibilidad siguientes:

$$g_l \geq 0 \qquad l=1,...,p \qquad (6,\ a)$$

$$f_t^i - x_{t+1}^i + x_t^i = 0 \qquad t=0,...,T-1\ ,\ i=1,...,n \qquad (6,\ b)$$

También las condiciones de no negatividad:

$$\lambda_l \geq 0 \qquad l=1,...,p \qquad (7)$$

Y, por último, las condiciones de holgura complementaria:

$$\lambda_l g_l = 0 \qquad l=1,...,p \qquad (8)$$

Estas condiciones se suelen aplicar a través de una función llamada hamiltoniana que es la siguiente:

$$H_t(x_t, u_t, \mu_{t+1}) = F_t(x_t, u_t) + \sum_{i=1}^{n} \mu_{t+1}^i f_t^i(x_t, u_t) \qquad t=0,...,T-1$$

Esta función es habitual en optimización dinámica tanto en tiempo discreto como continuo y su importancia es crucial para relacionar las condiciones de óptimo derivadas de la P.N.L. y aquéllas que se obtendrán desde el P.M.. Utilizando la función hamiltoniana, y como paso intermedio, se tiene:

$$\frac{\partial H_t}{\partial x_t^i}=\frac{\partial F_t}{\partial x_t^i}+\sum_{k=1}^{n}\mu_{t+1}^k\frac{\partial f_t^k}{\partial x_t^i} \qquad t=0,...,T-1\ ,\ i=1,...,n$$

$$\frac{\partial H_t}{\partial u_t^j}=\frac{\partial F_t}{\partial u_t^j}+\sum_{k=1}^{n}\mu_{t+1}^k\frac{\partial f_t^k}{\partial u_t^j} \qquad t=0,...,T-1\ ,\ j=1,...,m$$

$$\frac{\partial H_t}{\partial \mu_{t+1}^i}=f_t^i \qquad t=0,...,T-1\ ,\ i=1,...,n$$

Entonces, algunas de las condiciones de Kuhn y Tucker del Problema [V] se pueden expresar mediante la función hamiltoniana, en concreto las condiciones (5, a), (5, c) y (6, b) pasan a ser, respectivamente:

$$\frac{\partial H_t}{\partial x_t^i}-\mu_t^i+\mu_{t+1}^i+\sum_{l=1}^{p}\lambda_l\frac{\partial g_l}{\partial x_t^i}=0 \qquad t=1,...,T-1\ ,\ i=1,...,n \qquad (5,\ a')$$

$$\frac{\partial H_t}{\partial u_t^j}+\sum_{l=1}^{p}\lambda_l\frac{\partial g_l}{\partial u_t^j}=0 \qquad t=0,...,T-1\ ,\ j=1,...,m \qquad (5,\ c')$$

$$\frac{\partial H_t}{\partial \mu_{t+1}^i}-x_{t+1}^i+x_t^i=0 \qquad t=0,...,T-1\ ,\ i=1,...,n \qquad (6,\ b')$$

Por último, reordenando los términos, agrupando las expresiones y pasando a una notación vectorial[14] se consigue las condiciones de Kuhn y Tucker del Problema [V] de C.O. discreto:

[14]Obsérvese el abuso de notación al expresar la derivada parcial de una función vectorial, puesto que debería expresarse a través de una matriz Jacobiana. Sin embargo, cuando no haya lugar a confusión utilizaremos la derivada parcial para presentar los resultados de una manera más homogénea.

$$\frac{\partial H_t}{\partial x_t}+\lambda\cdot\frac{\partial g}{\partial x_t}=-(\mu_{t+1}-\mu_t) \quad t=1,...,T-1 \quad (KT\text{-}COD\ a)$$

$$\frac{\partial F_T}{\partial x_T}+\lambda\frac{\partial g}{\partial x_T}=\mu_T \quad (KT\text{-}COD\ b)$$

$$\frac{\partial H_t}{\partial u_t}+\lambda\frac{\partial g}{\partial u_t}=0 \quad t=0,...,T-1 \quad (KT\text{-}COD\ c)$$

$$g\geq 0\ ,\ \lambda g=0\ ,\ \lambda\geq 0 \quad (KT\text{-}COD\ d)$$

$$\frac{\partial H_t}{\partial \mu_{t+1}}=x_{t+1}-x_t \quad t=0,...,T-1 \quad (KT\text{-}COD\ e)$$

Para llegar a especificar si estas condiciones son necesarias y/o suficientes de máximo global hay que partir de una serie de supuestos que tienen su justificación en el epígrafe II.3.2. y que, a continuación, se resumen:

1°) Las funciones F_t, F_T, f_t^i y g_j son diferenciables en el conjunto de oportunidades respecto a todas sus variables, lo que implica, a su vez, la diferenciabilidad de la función hamiltoniana.

2°) La función objetivo es pseudocóncava.

3°) Las funciones f_t^i son cuasicóncavas si $\mu_t^i>0$ o cuasiconvexas si $\mu_t^i<0$.

4°) Las funciones g_j que corresponden a restricciones saturadas son cuasicóncavas.

5°) Se cumple alguna cualificación de restricciones, por ejemplo la condición de regularidad o la de Mangasarian y Fromovitz.

Ahora se tienen los siguientes teoremas:

Teorema II.7.: Sea z^* un punto máximo global del Problema **[V]** de C.O. discreto. Si se cumple el supuesto 1°) y en z^* se verifica el supuesto 5°), entonces z^* cumple las condiciones de Kuhn y Tucker expresadas en ***(KT-COD)***.

Teorema II.8.: Sea $z' \in Z$ un punto que verifica las condiciones ***(KT-COD)***. Si en z' se satisfacen los supuestos 2°), 3°) y 4°), entonces z' es un máximo global del Problema **[V]** de C.O. discreto.

Los teoremas II.7. y II.8. son una simple traslación respectivamente de los teoremas II.5. y II.6. de P.N.L. al Problema **[V]** de C.O. discreto, traslación que es posible dada la equivalencia demostrada entre ambos problemas.

A continuación se considera un caso especial de restricciones de desigualdad de forma que dependan de las variables en un mismo periodo de tiempo. Para ello las restricciones de trayectoria se deben poder englobar dentro de las restricciones de control y de estado-espacio lo cual, aunque pueda restar generalidad, es bastante seguido por muchos autores.

Tras eliminar las restricciones de trayectoria, el conjunto de oportunidades se puede definir a través de unas restricciones de desigualdad de la forma $g_t(x_t, u_t) \geq 0$, de la condición inicial y de un conjunto Z que no se pueda reducir a restricciones de desigualdad. La función g_t es una función vectorial, es decir, se considera que en cada periodo hay varias restricciones de desigualdad, cada una de las cuales tiene la característica de tener como argumentos las variables en un mismo periodo de tiempo.

El Problema **[VI]** es ahora similar al anterior pero, al eliminar las restricciones de trayectoria, es posible establecer unas condiciones de estacionariedad más simples.

$$Max \quad \sum_{t=0}^{T-1} F_t(x_t,u_t)+F_T(x_T)$$

$$s.a:$$

$$x_{t+1}-x_t=f_t(x_t,u_t) \qquad t=0,...,T-1$$

$$g_t(x_t,u_t)\geq 0 \qquad t=0,...,T-1$$

$$x_0 \ y \ T \ conocidos$$

[VI]

Utilizando la función hamiltoniana y en notación vectorial, las condiciones de K-T para el Problema [VI] son:

$$\frac{\partial H_t}{\partial x_t}+\lambda_t\frac{\partial g_t}{\partial x_t}=-(\mu_{t+1}-\mu_t) \qquad t=1,...,T-1 \qquad (9,\ a)$$

$$\frac{\partial F_T}{\partial x_T}=\mu_T \qquad (9,\ b)$$

$$\frac{\partial H_t}{\partial u_t}+\lambda_t\frac{\partial g_t}{\partial u_t}=0 \qquad t=0,...,T-1 \qquad (9,\ c)$$

$$g_t\geq 0\ ,\ \lambda_t g_t=0\ ,\ \lambda_t\geq 0 \qquad t=0,...,T-1 \qquad (9,\ d)$$

$$\frac{\partial H_t}{\partial \mu_{t+1}}=x_{t+1}-x_t \qquad t=0,...,T-1 \qquad (9,\ e)$$

El enunciado de las condiciones (9,a) a (9,e) es posible gracias a la forma que han adoptado las restricciones de desigualdad y tiene la ventaja de que se pueden separar en cada periodo de tiempo. Esto permite, como veremos en el siguiente epígrafe, una mejor comparación entre la aproximación a través de la P.N.L. y la aproximación con el P.M..

Antes de acabar este epígrafe es interesante hacer algunas **observaciones.**

La primera hace referencia a las **condiciones de transversalidad**. Como se comentó en el capítulo I el problema admite distintos enunciados en función de si se fijan o no los puntos iniciales y/o finales. En este sentido, y referido al Problema [V], la segunda condición de Kuhn y Tucker (K-T, b) se puede entender como la condición de transversalidad asociada a un valor final no especificado de la variable de estado, en cambio si se fija este valor de x_T la segunda condición desaparece. De la misma forma, si la condición inicial no se especifica en el enunciado del problema, es decir, si x_0 es libre, en su lugar aparece una condición de transversalidad que la sustituye. Esta condición, de tipo estacionario, se obtiene al derivar la lagrangiana respecto a x_0 e igualar a cero. En el Problema [VI] la condición correspondiente a x_0 libre sería:

$$\frac{\partial H_0}{\partial x_0}+\lambda_0\frac{\partial g_0}{\partial x_0}=-\mu_1$$

Otra observación hace referencia al **carácter del conjunto de oportunidades** S. Se ha partido de la hipótesis que este conjunto se puede delimitar a través de un conjunto de igualdades y desigualdades y a través de un conjunto Z abierto y no vacío para que tenga sentido hablar de diferenciabilidad en todos los puntos. Detrás de este supuesto está la convicción de que en la mayoría de los problemas y sobre todo en las aplicaciones económicas Z es todo el espacio $\mathbb{R}^r$, es decir, el conjunto de oportunidades se puede delimitar totalmente con restricciones de igualdad o desigualdad. Sin embargo, se puede plantear otras formas más generales del conjunto de oportunidades abandonando así la P.N.L.. Si el conjunto S es convexo y la función objetivo es diferenciable, la condición necesaria que se obtiene es del tipo:

$$\nabla F(z')(z-z')\leq 0 \qquad \forall z\in S$$

que es también suficiente si F es cóncava en z'. Si el conjunto S es todavía más general se puede utilizar como condición necesaria la no existencia de direcciones factibles de mejora (ver Bazaraa, Sherali y Shetty, 1993).

II.4.- RESOLUCIÓN A TRAVÉS DEL PRINCIPIO DEL MÁXIMO CLÁSICO

II.4.1. Enunciado del Principio del Máximo discreto clásico

Siguiendo a Boltyanskii (1978), la gran popularidad del P.M. de Pontryagin en el caso continuo hizo dirigir los esfuerzos de los investigadores hacia la obtención de condiciones de óptimo en el problema discreto en forma también de un Principio del Máximo. Se enunció, en consecuencia, un Principio del Máximo discreto análogo al continuo, aunque su validez no quedó asegurada automáticamente ya que en el paso de una variable continua a una discreta se producen errores de aproximación y de concepto. Pese a ello, este primer enunciado del P.M. discreto es uno de los principales puntos de partida en la obtención de las condiciones de óptimo del problema de C.O. discreto. Veamos como surge el P.M. discreto a partir del continuo.

Para obtener el P.M. continuo se parte del Problema **[VII]** de control óptimo en tiempo continuo:

$$\begin{aligned} &Max \quad F_T(x(T),T)+\int_0^T F(x(t),u(t))\,dt \\ &s.a: \\ &\frac{dx(t)}{dt}=f(x(t),u(t)) \\ &g(x(t),u(t))\geq 0 \\ &x(0) \;\; y \;\; T \;\; conocidos \end{aligned} \qquad \textbf{[VII]}$$

donde ahora $t\in\mathbb{R}$, $x(t)\in\mathbb{R}^n$, $u(t)\in\mathbb{R}^m$ y g: $\mathbb{R}^{n+m}$ —> $\mathbb{R}^p$

Para establecer las condiciones necesarias de máximo hay que partir de la función hamiltoniana:

$$H(x(t),\ u(t),\ \mu(t)) = F(x(t),\ u(t),\ t) + \mu(t)f(x(t),\ u(t),\ t)$$

o también de la función hamiltoniana ampliada:

$$H'(x(t),\ u(t),\ \mu(t)) = H + \lambda(t)g(x(t),\ u(t),\ t)$$

donde $\mu(t)$ es el vector de n variables de coestado y $\lambda(t)$ es el vector de p multiplicadores dinámicos.

El siguiente sistema hamiltoniano dinámico expresa las relaciones entre las variables dinámicas:

$$\frac{dx(t)}{dt} = \frac{\partial H'(t)}{\partial \mu(t)} \quad \leftrightarrow \quad \dot{x} = H'_{\mu}$$

$$\frac{d\mu(t)}{dt} = -\frac{\partial H'(t)}{\partial x(t)} \quad \leftrightarrow \quad \dot{\mu} = -H'_{x}$$

La resolución de este sistema permite obtener las variables $x(t)$ y $\mu(t)$ una vez conocida la variable $u(t)$ y aplicar las condiciones de transversalidad. Para encontrar los controles $u(t)$ y, en su caso, las condiciones de transversalidad se debe tener en cuenta el denominado Principio del Máximo de Pontryagin (Pontryagin y otros, 1962) que, para el Problema **[VII]**, se enuncia a través del teorema II.9..

> *Teorema II.9.:* Sea $u^*(t)$ una función continua que resuelve el Problema **[VII]** y sea $x^*(t)$ la trayectoria óptima asociada a tales variables de control. Entonces, existe una función continua no nula $\mu(t)$ tal que:

a) Si $U(t)$ es el conjunto de controles admisibles o factibles, se cumple para todo $t \in [0,T]$ y para todo $u(t) \in U(t)$:

$$H(x^*(t),\ u^*(t),\ \mu(t),\ t) \geq H(x^*(t),\ u(t),\ \mu(t),\ t)$$

b) Se cumple el sistema hamiltoniano dinámico en $(x^*(t),\ u^*(t))$.

c) Se cumple la condición de transversalidad asociada a un valor final libre de la variable de estado:

$$\frac{\partial F_T(x(T),T)}{\partial x(T)} = \mu(T)$$

Este enunciado del Principio del Máximo se puede trasladar al caso de tiempo discreto si el problema se enuncia sin restricciones de trayectoria[15]. Sea, por tanto, el Problema **[VI]** de C.O. discreto ya enunciado en II.3.3.:

$$Max \quad \sum_{t=0}^{T-1} F_t(x_t,u_t) + F_T(x_T)$$

$$s.a:$$

$$x_{t+1} - x_t = f_t(x_t,u_t) \qquad t=0,...,T-1$$

$$g_t(x_t,u_t) \geq 0 \qquad t=0,...,T-1$$

$$x_0 \ \ y \ \ T \ \ conocidos$$

[VI]

La función hamiltoniana asociada en el periodo t es:

[15]Esto no siempre supone una pérdida de generalidad porque como hemos visto en el epígrafe I.3.3. ciertas restricciones de trayectoria se pueden eliminar creando variables de estado adicionales.

$$H_t(x_t,u_t,\mu_{t+1})=F_t(x_t,u_t)+\mu_{t+1}f_t(x_t,u_t)$$

Y la hamiltoniana ampliada en el periodo t:

$$H'_t(x_t,u_t,\mu_{t+1},\lambda_t)=H_t+\lambda_t g_t(x_t,u_t)$$

El sistema hamiltoniano está formado ahora por las siguientes ecuaciones en diferencias:

$$x_{t+1}-x_t=\frac{\partial H'_t}{\partial \mu_{t+1}} \qquad (10,\ a)$$

$$\mu_{t+1}-\mu_t=-\frac{\partial H'_t}{\partial x_t} \qquad (10,\ b)$$

Estas son equivalentes, en tiempo discreto, a las ecuaciones diferenciales que forman el sistema hamiltoniano en tiempo continuo. Ahora el enunciado del principio del máximo a validar es el siguiente:

Teorema II.10. (Principio del Máximo discreto): Sea $U^*=\{u_t^*;t=0,\ldots,T\text{-}1\}$ una sucesión solución del Problema **[VI]** y $X^*=\{x_t^*;t=0,\ldots,T\text{-}1\}$ la solución asociada para las variables de estado. Entonces, existe una sucesión no nula $\mu(t)$ tal que:

a) Si U_t es el conjunto de controles admisibles o factibles, se cumple para todo $t=0,1,\ldots,T$ y para todo $u_t\in U_t$:

$$H(x_t^*, u_t^*, \mu_{t+1}) \geq H(x_t^*, u_t, \mu_{t+1}) \qquad (P.M., a)$$

b) Se cumple el sistema hamiltoniano dinámico en (x^*_t, u^*_t).

c) Se cumple la siguiente condición de transversalidad asociada a un valor final libre de la variable de estado:

$$\frac{\partial F_T}{\partial x_T} = \mu_T \qquad (P.M., b)$$

Este P.M. discreto ha sido enunciado en forma de condición necesaria por analogía al P.M. continuo. Sin embargo, los errores que se cometen al pasar del caso continuo al discreto impiden considerar este enunciado como una condición necesaria, por lo que hay que considerarlo sólo como un punto de partida. Lo que interesa a partir de aquí es estudiar bajo que condiciones el anterior P.M. discreto es válido y en consecuencia sirve, al igual que el continuo, como condición necesaria de óptimo global. Las formas de demostrar o dar validez a este P.M. en tiempo discreto es lo que nos ocupa en el siguiente epígrafe.

II.4.2. Validaciones del Principio del Máximo discreto

El Principio del Máximo discreto como condición necesaria para obtener un control óptimo ha tenido dos vías de demostración. La primera, a través de **argumentos geométricos** similares a los utilizados en el Principio del Máximo continuo, se basa en estudiar las características que se pueden deducir de trayectorias de control próximas a las trayectorias óptimas, esto es lo que se llama "técnicas de perturbaciones".

La segunda aproximación, a través de la **Programación no Lineal**, trata de estudiar bajo qué condiciones se produce la equivalencia entre el P.M. discreto

(junto con el sistema hamiltoniano) y las condiciones de estacionariedad de Kuhn y Tucker; si se da esta equivalencia se puede aplicar las técnicas de la P.N.L. para llegar a la condición de máximo global.

En otras palabras, para relacionar el principio del máximo con la condición de máximo global se puede seguir dos caminos: uno, directo, mediante **argumentos geométricos**; y otro indirecto a través de las **condiciones de Kuhn y Tucker**. En ambas aproximaciones se exigen condiciones adicionales sobre la estructura del problema para que el principio del máximo sea condición necesaria de máximo global. El requerimiento matemático de la primera aproximación es más complejo pero es necesario para ampliar la aplicabilidad del P.M. discreto ya que existen problemas donde los supuestos de diferenciabilidad, convexidad, cualificación de restricciones, etc. que se exigen en la segunda aproximación no se dan, aunque éste no es el caso de las aplicaciones económicas que aquí se desarrollan.

La **aproximación geométrica** se basa en técnicas similares a las del Principio del Máximo en tiempo continuo. A pesar de que algunos autores trasladaron inicialmente de forma automática este principio al caso discreto otros autores pusieron de manifiesto que esto conduce a errores en ciertos tipos de sistemas discretos[16]. El primero que intenta acotar la validez del principio del máximo discreto es Rozonoer (1959) que demuestra su validez en sistemas discretos lineales. Las investigaciones posteriores se orientaron a resaltar la importancia de la convexidad del conjunto que forman los posibles valores de los estados y de la función objetivo, llamado conjunto de estados extendidos alcanzables[17]. En este sentido los artículos de Halkin y Holtzman son los más conocidos. La condición importante para dar validez al P.M. discreto necesita de las definiciones II.8. y II.9.

---

[16]Entre los autores que utilizaron de forma incorrecta el P.M. en tiempo discreto se puede citar a Fang y Wang (*The Discrete Maximum Principle. John Wiley, Nueva York, 1964*), aunque previamente (en 1962) Katz había publicado dos artículos en la misma línea. Un resumen esclarecedor de estos inicios del P.M. discreto se puede seguir en Nahorski, Ravn y Valqui (1984).

[17]El tratamiento del valor de la función objetivo como un estado más proviene del enunciado del problema en la forma de Mayer.

Definición II.8.: Un conjunto $A \subset \mathbb{R}^n$ es **direccionalmente convexo** en la dirección del vector $d \in \mathbb{R}^n$ si $x+\lambda d \in A \quad \forall x \in \mathrm{Co}(A)$ para algún $\lambda \geq 0$, donde $\mathrm{Co}(A)$ es la envoltura lineal convexa del conjunto A.

Esta definición implica que cualquier combinación lineal convexa de puntos del conjunto A, se puede desplazar en la dirección d y llegar a un punto de dicho conjunto. El siguiente ejemplo gráfico muestra, en $\mathbb{R}^2$, un conjunto no convexo pero direccionalmente convexo en la dirección del vector (0,1).

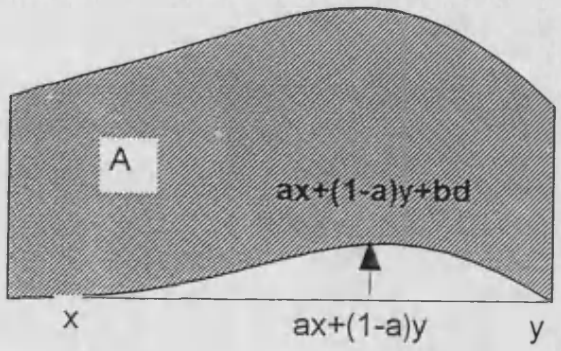

Gráfico II.1.: Convexidad direccional.

Definición II.9.: Se llama **conjunto de estados extendidos alcanzables** desde un estado previo x_t, a un subconjunto de $\mathbb{R}^{n+1}$ donde cada elemento tiene una primera componente que es el valor que puede alcanzar la suma de las funciones intermedias hasta el periodo siguiente y el resto de componentes son los valores que pueden alcanzar las variables de estado. Estos valores van apareciendo al aplicar controles admisibles sobre la función

intermedia y el sistema dinámico. Es decir, llamando $V_{t+1}(x_t)$ a este conjunto de estados extendidos alcanzables desde x_t se tiene:

$$V_{t+1}(x_t) = \{ (x^0_{t+1}, x_{t+1}) \in \mathbb{R}^{n+1} \ / \ u_t \in U_t(x_t) \}$$

donde:

$$x^0_{t+1} = \sum_{k=0}^{t} F_k(x_k, u_k)$$

$$x_{t+1} = x_t + f_t(x_t, u_t)$$

$$U_t(x_t) = \{ u_t \ / \ g_t(x_t, u_t) \geq 0 \}$$

Con estas definiciones se tiene la condición de convexidad direccional, establecida por primera vez en Holtzman (1966).

Condición de Convexidad Direccional (C.D.): El conjunto de estados extendidos alcanzables desde un estado previo en cada periodo de tiempo, $V_{t+1}(x_t)$, debe ser direccionalmente convexo en la dirección del primer vector de la base canónica.

Ahora podemos enunciar el siguiente teorema:

Teorema II.11. (1ª validación del P.M. discreto): Si se cumple la condición ***C.D.*** entonces es válido el enunciado del teorema II.10.

Se puede ver la demostración pionera en Holtzman (1966) tomando como base de partida la de Halkin en 1965 aunque la demostración más rigurosa de este teorema se encuentra en Boltyanskii (1978) donde se realiza para distintos casos según sea la expresión de la región de control. Este teorema es aplicable no sólo en

el Problema **[VI]** sino también en aquellos casos en que la región de control no se pueda reducir a igualdades y desigualdades. Para entender la importancia de ***C.D.*** resulta también interesante las consideraciones que se recogen en Nahorski, Ravn y Valqui (1984) y en Nahorski y Ravn (1988b) dado que ilustran gráficamente las implicaciones de la convexidad direccional del conjunto de estados extendidos alcanzables para asegurar la existencia de un hiperplano soporte a dicho conjunto y que pase por el estado extendido óptimo, hecho éste que permite la demostración del P.M. discreto.

Esta aproximación geométrica ha sido objeto de estudio por otros autores a partir de los años 80 con el objetivo de relajar la hipótesis de convexidad direccional. Uno de los estudios más importantes se debe a autores ya citados (Nahorski, Ravn y Valqui, 1984; Nahorski y Ravn, 1988b), los cuales presentaron un principio del máximo generalizado que no parte de ***C.D.*** sino que se basa en la construcción de una hamiltoniana generalizada que no es lineal en la variable de coestado.

En los artículos publicados por Halkin y Holtzman se reconoce que ***C.D.*** es sólo una condición suficiente para dar validez al P.M. discreto, condición que, además, facilita su demostración. Por tanto aunque el P.M. pueda no cumplirse si esta condición no se da, es posible que se cumpla bajo otras condiciones como efectivamente demuestra Kleindorfer (en Tapiero, 1977) o en general a través de la segunda vía de validar el P.M.: la Programación no Lineal.

La segunda vía para validar el principio del máximo se realiza a través de las condiciones de Kuhn y Tucker y tiene la ventaja de que no necesita instrumentos matemáticos distintos a los utilizados en P.N.L.. Los inconvenientes lógicos aparecen porque el problema de control debe cumplir las condiciones adecuadas ya vistas en el epígrafe II.2 para poderse tratar como un problema de programación matemática, condiciones que se cumplen habitualmente en las aplicaciones económicas. Las implicaciones que se deben dar siguen el esquema de la Figura II.3..

$$P.M. \underset{2^a}{<\!\!-\!\!-} K\text{-}T \underset{1^a}{<\!\!-\!\!-} M.G.$$

Figura II.3.: El principio del máximo a través de K-T.

La implicación 1ª ha sido estudiada en el epígrafe II.3.2. y se cumple bajo la hipótesis de cualificación de restricciones o alguna condición de regularidad.

Ahora, para analizar la implicación 2ª hay que realizar un estudio de la interrelación entre el bloque de condiciones (10)-***P.M.*** (sistema hamiltoniano dinámico y maximización de la hamiltoniana) y el bloque de condiciones (9) (condiciones de Kuhn y Tucker). Efectivamente, la implicación 2ª se puede desglosar tal y como se recoge en la figura II.4..

(10, a), (10, b), (***P.M.***, a), (***P.M.***, b) <—— (9, a) hasta (9, e)

Figura II.4.: Desglose de la implicación 2ª.

Dado que la función hamiltoniana es igual, independientemente de si el problema se toma como de P.N.L. o de C.O. discreto, se observa fácilmente que el sistema de ecuaciones en diferencias (10) junto con las condiciones de transversalidad (***P.M.***, b) equivalen a (9, a), (9, b) y (9, e).

Sin embargo, para que la equivalencia entre ambas aproximaciones fuera total se debería dar la equivalencia entre la condición (***P.M.***, a) por una parte y las condiciones (9, c) y (9, d) por otra. Para ello, los controles que maximicen la hamiltoniana deben cumplir también las condiciones de K-T expresadas en (9, c) y

(9, d) y viceversa. En otras palabras, el siguiente problema de maximizar la hamiltoniana entre los controles admisibles:

$$\begin{array}{ll} Max & H_t(x_t^*, u_t, \mu_{t+1}) \\ s.a: & g_t(x_t^*, u_t) \geq 0 \\ para & t=0,...,T-1 \end{array} \qquad \text{[VIII]}$$

debe tener como condiciones necesarias y suficientes para resolverlo las siguientes:

$$\frac{\partial H_t}{\partial u_t} + \lambda_t \frac{\partial g_t}{\partial u_t} = 0 \qquad t=0,...,T-1 \qquad (9\ c)$$

$$g_t \geq 0\ ,\ \lambda_t g_t = 0\ ,\ \lambda_t \geq 0 \qquad t=0,...,T-1 \qquad (9\ d)$$

aunque, para que la implicación 2ª se cumpla basta con que las dos condiciones (9, c) y (9, d) sean suficientes de (*P.M.*, b).

Para ello hay que observar que este problema de maximizar la hamiltoniana respecto a los controles admisibles es un problema no lineal (dado que el problema de control de partida también lo es) y por tanto se puede abordar a través de las condiciones de Kuhn y Tucker, condiciones que se obtienen construyendo previamente su función lagrangiana:

$$L(u_t) = H_t + \lambda_t g_t$$

sobre la que se aplica las condiciones de óptimo:

$$\nabla L(u_t)=0 \qquad (11,\ a)$$

$$\lambda_t \geq 0,\ g_t \geq 0,\ \lambda_t g_t=0 \qquad (11,\ b)$$

Las condiciones (11, a) y (11, b) coinciden con (9, c) y (9, d). Es decir, las condiciones de Kuhn y Tucker del problema de maximizar la hamiltoniana respecto a las variables de control constituyen una parte de las condiciones de K-T del problema de C.O. discreto.

Para completar la demostración de la implicación 2ª basta con observar bajo qué supuestos estas condiciones de Kuhn y Tucker son suficientes de máximo global de la hamiltoniana. Recordando el teorema de suficiencia en P.N.L. se puede enunciar el teorema II.12.:

> *Teorema II.12.:* Si la función hamiltoniana es pseudocóncava respecto de u_t y las funciones componentes de g_A son cuasicóncavas también respecto de u_t entonces los controles que cumplen (9, c) y (9, d) son máximos globales de la función hamiltoniana.
>
> *Demostración:* Basta con aplicar el teorema II.6. al problema de maximizar la hamiltoniana y observar que la función objetivo es la hamiltoniana y que no hay restricciones de igualdad.

De esta manera queda completada la demostración de las implicaciones recogidas en la figura II.4., que se puede resumir en forma de teorema (teorema II.13.).

> *Teorema II.13. (2ª validación del P.M. discreto):* Si se cumple la hipótesis de cualificación de restricciones o alguna condición de regularidad en el problema **[VI]** y si la función hamiltoniana es pseudocóncava respecto de u_t y las funciones componentes de g_A son

cuasicóncavas también respecto de u_t entonces es válido el enunciado del teorema II.10.

Kleindorfer y otros (en Tapiero, 1977) exponen otra vía para demostrar la segunda implicación de la Figura II.3. a través de las condiciones de Kuhn y Tucker del problema de maximización de la hamiltoniana. Esta demostración es menos importante por la propia naturaleza de los supuestos, que la hacen prácticamente trivial.

Teorema II.14.: Las condiciones de K-T son suficientes de máximo global si se dan los tres supuestos siguientes:

a) Existencia de máximo global de la hamiltoniana entre los controles admisibles (compacidad del conjunto de controles admisibles y semicontinuidad superior de la hamiltoniana[18]).

b) Cualificación de restricciones en el problema de maximizar la hamiltoniana.

c) La solución del sistema formado por las condiciones de Kuhn y Tucker de este problema es única.

Demostración: Efectivamente bajo los dos primeros supuestos las condiciones de K-T son necesarias de máximo para la hamiltoniana y dado que estas condiciones tienen solución única, ésta es también la solución del problema de maximización de la hamiltoniana dado que en este problema existe máximo global.

El P.M. discreto clásico exige, como acabamos de ver, el cumplimiento de

[18]En realidad estos autores exigen la continuidad de la hamiltoniana aunque este supuesto se puede relajar por el de semicontinuidad superior.

hipótesis adicionales acerca de los conjuntos y funciones que aparecen en el problema, hipótesis que dependen de la aproximación que se utilice. Esto se puede resumir en la figura II.5..

Aproximación geométrica

Máximo global ——— Convexidad direccional ———> Principio del máximo

Aproximación a través de K-T

Máximo global ———> Condiciones de K-T ———> Principio del máximo

Condiciones de K-T: Cualificación de restricciones; H_t pseudocóncava; g_A cuasicóncavas

Figura II.5.: Las dos vías para validar el Principio del Máximo en tiempo discreto.

II.4.3.- Comparación entre las condiciones de Kuhn y Tucker y el Principio del Máximo como condiciones necesarias de máximo global

La aproximación a través de las condiciones de Kuhn y Tucker se puede seguir si se dan las condiciones para reducir un problema de extremos a un problema de programación matemática dado que el problema de control óptimo se puede tratar como un problema de extremos; para ello las funciones deben ser diferenciables y el conjunto de puntos factibles se debe expresar a través de igualdades, desigualdades y a través de un conjunto abierto. Esta aproximación permite englobar la teoría del C.O. en tiempo discreto dentro de otra mucho más conocida y

desarrollada como es la programación matemática con la consiguiente ventaja que ello supone, tanto a la hora de expresar las características que deben cumplir los máximos globales como a la hora de establecer algoritmos y métodos numéricos para obtener esos puntos. El principal inconveniente de esta aproximación es la gran dimensión del problema que se deriva del hecho de considerar a cada variable en cada periodo de tiempo como una variable estática, con lo que la dimensión del problema crece muy rápidamente a medida que aumenta el número de periodos de tiempo que forman parte del problema, además esta aproximación no aprovecha la estructura especial del problema de control con lo que la ventaja que supone disponer de algoritmos de resolución de problemas no lineales queda un poco mermada al tratarse de algoritmos generales y no específicos al problema de control lo que se traduce en tiempos de cálculo mayores.

El P.M. es más específico de la optimización dinámica y se basa en resolver un problema de optimización de dimensión menor en cada periodo de tiempo (el problema de maximizar la hamiltoniana) gracias a que sí aprovecha la estructura del problema de control para desglosarlo en varios problemas de menor dimensión. Sin embargo, los algoritmos de resolución no han sido todavía suficientemente estudiados y es de esperar que vayan surgiendo algoritmos más eficientes para este problema que los disponibles en P.N.L..

En cuanto a la validez de ambos métodos, se ha estudiado que las condiciones de K-T necesitan de la cualificación de restricciones como hipótesis adicional para ser condiciones necesarias de máximo global. Por otra parte, el P.M. necesita también hipótesis adicionales. La más habitual es la de convexidad direccional de los estados extendidos alcanzables, aunque en la Figura II.5. se recoge otra vía de dar validez al Principio del Máximo: a través de las condiciones de K-T.

Ambas aproximaciones, pese a ser distintas en su origen, se pueden relacionar como se ha visto en el epígrafe II.4.2 en el sentido que bajo ciertas hipótesis las condiciones de Kuhn y Tucker implican la maximización de la hamiltoniana.

CAPÍTULO III:

EXTENSIONES

III.1.- INTRODUCCIÓN

El capítulo anterior se ha centrado en el estudio del Principio del Máximo como condición necesaria de óptimo global en el problema general de C.O. discreto. En este capítulo se hará referencia a aspectos colaterales que merece la pena destacar para tener una visión más amplia de las posibilidades que encierra este tipo de técnicas. Hay que decir que las extensiones que se desarrollan en este capítulo son puramente matemáticas y que el campo de extensiones no se cierra con las que aquí se enumeran limitándonos a las que creemos más importantes o que pueden servir para las aplicaciones incluidas en los últimos capítulos.

Destacamos, en primer lugar, consideraciones acerca de la existencia de la solución y acerca de condiciones de suficiencia para un punto[19]. Analizar los puntos que cumplen el P.M. sólo significa que si existe una solución óptima ésta formará parte de esos puntos; sin embargo, en ningún caso se puede asegurar que la solución encontrada a través del P.M. sea la óptima dado que es posible que no exista óptimo. En otros casos, puede haber varios puntos que cumplan las condiciones necesarias y se necesitan condiciones de segundo orden para concluir cuál o cuáles de ellos son efectivamente óptimos. No obstante, si se sabe que existe solución óptima y si el P.M. tiene una única solución se podrá asegurar que ésa es la solución óptima buscada. Los teoremas de existencia y de suficiencia se abordan en el siguiente epígrafe.

Tanto la existencia como la suficiencia son aspectos puramente matemáticos ligados a la resolución de un problema genérico. En las aplicaciones económicas no se suele plantear estas dos cuestiones debido a que la modelización del comportamiento económico se lleva a cabo con una lógica que lleva implícito el cumplimiento de ambos requisitos, quedando como cuestión fundamental la búsqueda del punto que cumple las condiciones necesarias.

[19]En el contexto dinámico en el que nos encontramos un punto equivale a una trayectoria.

Una extensión de sumo interés en este tipo de problemas es el denominado análisis de sensibilidad. Este análisis complementa la información que se persigue con la obtención de la solución ya que permite aproximar la nueva solución o, al menos, observar en qué dirección se desplazará la solución original al producirse pequeños cambios en el valor de algún parámetro del problema.

El capítulo finaliza con un epígrafe donde se recoge, a nivel introductorio, dos extensiones que marcan líneas de investigación recientes y de desarrollo futuro. La primera trata la extensión de la teoría del C.O. discreto al campo estocástico. La segunda hace referencia a la relajación de la condición de convexidad direccional para dar validez al P.M. discreto.

III.2.- TEOREMAS DE EXISTENCIA Y SUFICIENCIA

III.2.1.- Estudio de la existencia de solución óptima

Los teoremas de existencia especifican bajo qué condiciones se puede llegar a aplicar el teorema de Weierstrass que, como se sabe, garantiza la existencia de óptimo global en un problema de extremos. Este teorema exige dos condiciones: continuidad de la función objetivo y conjunto de oportunidades compacto y no vacío. Estas dos condiciones vienen referidas al problema tomado como uno de extremos (de gran dimensión) pero, por operatividad, interesa especificar otras que las sustituyan y que se apliquen en cada periodo. Para ello tomemos el Problema **[I]** de C.O. discreto:

$$Max \quad \sum_{t=0}^{T-1} F_t(x_t,u_t)+F_T(x_T)$$

$$s.a:$$

$$x_{t+1}-x_t=f_t(x_t,u_t) \qquad t=0,...,T-1$$

$$g_t(x_t,u_t)\geq 0 \qquad t=0,...,T-1$$

$$x_0 \;\; y \;\; T \;\; conocidos$$

[I]

Sobre este problema se enuncia el teorema de existencia que se enuncia a continuación.

Teorema III.1.: Si las funciones F_t, funciones intermedias; F_T, la función residual; y f_t son continuas y si $U_t(x_t)$ es compacto y no vacío para $t=0,\ldots,T-1$, donde $U_t(x_t) = \{ u_t / g_t(x_t, u_t) \geq 0 \}$; entonces se cumple el teorema de Weierstrass, es decir, existe solución óptima.

Demostración: Si la función intermedia F_t es continua y si la función residual F_T es continua entonces la función objetivo también lo es

dado que la suma de funciones continuas es otra función continua. Así queda demostrada la 1ª condición del teorema de Weierstrass, la continuidad de la función objetivo.

La segunda condición del teorema, conjunto de oportunidades compacto y no vacío, es equivalente a la existencia de trayectorias factibles. Para demostrarla, partiendo de un valor para x_0 conocido, se tiene que, por hipótesis, u_0 tomará valores en un conjunto compacto y no vacío; de donde se puede deducir que x_1 también tomará valores en un conjunto compacto y no vacío teniendo en cuenta el sistema dinámico y la continuidad de f_t. De esta manera, el conjunto de oportunidades en el periodo inicial es compacto y no vacío. Con el valor x_1 se repite el proceso hasta llegar al último periodo. Como el vector de variables se puede particionar por periodos, el conjunto de oportunidades del problema inicial es la intersección de los conjuntos de oportunidades de todos los periodos y por tanto es compacto. Además es no vacío porque por hipótesis todos los conjuntos U_t son no vacíos. Así queda demostrada la segunda condición del teorema de Weierstrass.

Este teorema se ha enunciado y demostrado en referencia al Problema [I] de control óptimo. Para otros enunciados más completos se pueden seguir teoremas y demostraciones análogas (ver por ejemplo Boltyanskii, 1978).

También se han estudiado teoremas de existencia en otros problemas más especiales como el de control óptimo discreto en tiempo infinito (Keerthi y Gilbert, 1985) y el caso no diferenciable (Dolezal, 1988).

El problema de la existencia de solución en aplicaciones económicas suele obviarse dado que la construcción del modelo atiende a unas consideraciones económicas que garantizan que existe solución óptima. Sin embargo, tal y como señalan Seierstad y Sydsaeter (1987), la lógica en la construcción del modelo no

sustituye a la demostración matemática de la existencia de solución, e incluso dicha demostración ayuda a dar consistencia al modelo.

III.2.2.- El Principio del Máximo como condición suficiente

En la búsqueda de algún teorema de suficiencia se toma como punto de partida el Principio del Máximo. Aunque dicho principio se enuncia originalmente como condición necesaria, se trata de estudiar los requisitos adicionales que aseguren que la condición de maximización de la hamiltoniana y el sistema hamiltoniano dinámico lleven en efecto a un óptimo global, pasándose así a un Principio del Máximo suficiente.

Una primera posibilidad es a través de las condiciones de K-T y por tanto bajo el supuesto de que el problema de C.O. discreto sea reducible a uno de Programación Matemática. En este caso, y si seguimos el Problema [I], se tiene el siguiente teorema:

> *Teorema III.2.:* Sea (X', U') un punto que maximiza la hamiltoniana y cumple el sistema hamiltoniano dinámico. Si en ese punto se cumple la cualificación de restricciones para g_t y si se cumplen las tres condiciones siguientes:
>
> - Función objetivo pseudocóncava[20].
>
> - Funciones f_t^i cuasicóncavas (si el multiplicador asociado es positivo) o cuasiconvexas (si es negativo).
>
> - Restricciones de desigualdad saturadas (componentes de g_A)

[20]Este supuesto no se puede establecer en términos de las funciones en cada periodo, es decir, no se puede sustituir por la condición de que F_t y F_T sean pseudocóncavas ya que la suma de funciones pseudocóncavas no es, en general, otra función pseudocóncava. Sí se puede exigir en cambio que F_t y F_T sean cóncavas ya que la suma es cóncava y por tanto pseudocóncava.

cuasicóncavas.

Entonces ese punto es el máximo global del Problema [I].

Demostración: La demostración se realiza interpretando el problema como uno de P.N.L. El esquema de relaciones que se debe demostrar es similar al de la figura II.3. del capítulo anterior pero en sentido opuesto, es decir:

$$P.M. \xrightarrow[1^a]{} K\text{-}T \xrightarrow[2^a]{} M.G.$$

Figura III.1.: El P.M. discreto como condición suficiente.

La implicación 1ª tiene dos partes. Por un lado el sistema hamiltoniano dinámico y la condición de transversalidad implican algunas de las condiciones de K-T dado que son totalmente equivalentes como ya se vio en el capítulo II (en concreto (10, a), (10, b) y (P.M., b) implican (9, a), (9, b) y (9, e)).

En segundo lugar queda estudiar cuándo la maximización de la hamiltoniana para los controles admisibles implica las condiciones (9,c) y (9,d). Esto es lo mismo que establecer cuándo estas condiciones son necesarias del problema no lineal que en realidad es el de maximización de la hamiltoniana. Ahora bien, en el epígrafe II.3.2. ha quedado establecido que las condiciones (9,c) y (9,d) son en realidad las condiciones de K-T del problema de maximizar la hamiltoniana y por tanto esas condiciones serán necesarias si se cumple la cualificación de restricciones para este problema (es decir una cualificación de restricciones que sólo afecta a las restricciones

g_t). Esta cualificación de restricciones es la que se cumple por hipótesis y podría sustituirse por la condición de regularidad que, aunque es más fuerte, es más operativa.

En cuanto a la implicación 2ª, hay que estudiar bajo qué supuestos las condiciones de K-T (todas las condiciones (9) del capítulo II) son también condiciones suficientes de M.G. del problema de C.O. discreto. Siguiendo en este caso el teorema de suficiencia en P.N.L. (teorema II.6) se establecen como supuestos adicionales las tres condiciones de partida que aparecen en el enunciado del teorema (son hipótesis de partida) y, por tanto, queda demostrada la implicación 2ª.

La figura III.2. resume los supuestos que se deben dar:

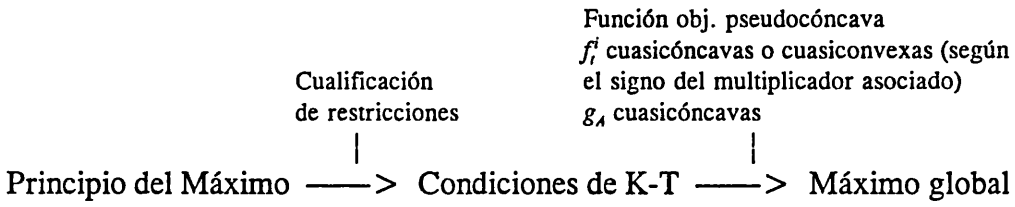

Figura III.2.: Condiciones adicionales de suficiencia.

Aunque no se den las circunstancias que permitan subsumir el problema de C.O. discreto dentro de la Programación Matemática todavía se puede enunciar otros teoremas de suficiencia aprovechando la equivalencia con un problema de extremos de funciones.

En este sentido Boltyanskii (1978) enuncia un teorema de suficiencia para el caso en que la región de control U_t sea un conjunto convexo que no dependa de la

variable de estado. En ese caso, además de la maximización de la hamiltoniana y el cumplimiento del sistema hamiltoniano dinámico se debe exigir que la función intermedia sea cóncava, que las funciones f_t sean lineales y que no aparezcan restricciones sobre los estados.[21]

Valqui (1987) establece un teorema de suficiencia todavía más global. Este autor presenta un conjunto de principios del máximo distintos al clásico con la ventaja de llegar a la condición de máximo global a través de desigualdades de punto de silla y por tanto sin necesidad de exigir supuestos adicionales a las funciones que intervienen en el problema. La desventaja de estos principios del máximo es que en lugar de maximizar la hamiltoniana en cada periodo respecto a los controles admisibles (T subproblemas con m variables) se maximiza una función en cada periodo respecto a los controles y estados factibles ($T+1$ subproblemas con $n+m$ variables).

Valqui elabora una familia de principios del máximo con el objetivo de conseguir las desigualdades de punto de silla de la optimización matemática para el problema de C.O. discreto. A partir de ahí, es casi inmediato pasar a la condición de M.G. (ver en cualquier manual de P.N.L., por ejemplo en Mangasarian, 1969). Lo que se hará a continuación es adaptar estos principios al Problema **[I]**. Para ese problema se toma la siguiente función Lagrangiana:

$$L(X,U,\mu,\lambda)=\sum_{t=0}^{T-1} F_t(x_t,u_t)+F_T(x_T)+$$

$$+\sum_{t=0}^{T-1} \mu_{t+1}(f_t(x_t,u_t)-x_{t+1}+x_t)+\sum_{t=0}^{T-1} \lambda_t g_t(x_t,u_t)$$

[21]Este teorema se enuncia en Boltyanskii para el caso de minimizar. En él se podrían relajar más algunos supuestos si se consideran conceptos de pseudoconcavidad y cuasiconcavidad que este autor no utiliza.

y a partir de ella se enuncia el siguiente teorema que proviene de la teoría de extremos de funciones:

> *Teorema III.3.:* Si existe un par admisible (X',U') y vectores μ' y λ' $(\lambda' \geq 0)$ de dimensiones apropiadas tales que, para todo par (X,U), para todo μ y para todo $\lambda \geq 0$, se verifica:
>
> $$L(X',U',\mu,\lambda) \geq L(X',U',\mu',\lambda') \geq L(X,U,\mu',\lambda') \qquad (1)$$
>
> entonces (X',U') es máximo global del Problema **[I]**.

La demostración de este teorema aparece en cualquier manual clásico de extremos de funciones o de P.N.L. (Mangasarian, 1969; Takayama, 1985).

La anterior función $L(X,U,\mu,\lambda)$ se puede separar por periodos gracias a la forma que adoptan las restricciones de desigualdad. Si quitamos los argumentos de las funciones para simplificar la notación se tiene:

$$L=\sum_{t=0}^{T-1} L_t+F_T \qquad (2)$$

donde:

$$L_t=F_t+\mu_{t+1}(f_t-x_{t+1}+x_t)+\lambda_t g_t \qquad t=0,...,T-1$$

Introduciendo ahora la función hamiltoniana conocida:

$$H_t=F_t+\mu_{t+1}f_t \qquad t=0,...,T-1$$

se observa la siguiente relación:

$$L_t=H_t-\mu_{t+1}(x_{t+1}-x_t)+\lambda_t g_t \qquad (3)$$

Esto permite expresar L en función de la hamiltoniana sustituyendo la expresión (3) en (2). Si se sustituye y se hacen operaciones se llega a:

$$L=H_0+\mu_1 x_0+\lambda_0 g_0+\sum_{t=1}^{T-1}[H_t+(\mu_{t+1}-\mu_t)x_t+\lambda_t g_t]+F_T-\mu_T x_T$$

función que es separable por periodos. Con todos estos antecedentes se puede enunciar el siguiente teorema:

Teorema III.4.: Si existe un par admisible (X',U') del problema **[I]** y vectores μ' y $\lambda' \geq 0$ tales que:

1.- u_0' maximiza $[\,H_0+\mu_1'x_0+\lambda_0'g_0\,]$ para los u_0 factibles.

2.- (x_t', u_t') maximiza $[\,H_t+(\mu_{t+1}'-\mu_t')x_t+\lambda_t'g_t\,]$ para los (x_t, u_t) factibles y ello para $t=1,...,T$-1.

3.- x_T' maximiza $[\,F_T-\mu_T'x_T\,]$ para los x_T factibles.

4.- $\lambda_t'g_t(x_t',u_t')=0$ para $t=0,...,T$-1.

Entonces (X',U') es un máximo global del problema **[I]**.

Demostración: Las condiciones 1,2 y 3 del teorema garantizan la segunda desigualdad que se recoge en la expresión (1) aplicando la descomposición que se ha hecho de la función L por periodos. Por otra parte, la primera desigualdad de esa expresión se demuestra de la siguiente manera. Por definición de la función lagrangiana se parte de:

$$L_t(x_t',u_t',\mu_{t+1},\lambda_t)=F_t+\mu_{t+1}(f_t-x_{t+1}'+x_t')+\lambda_t g_t(x_t',u_t') \qquad (4)$$

Dado que el punto (X',U') es factible, cumple la restricción de igualdad por lo que en el segundo sumando se puede reemplazar μ por μ' sin que cambie el resultado. Como el tercer sumando es mayor o igual que cero, si se sustituye por una expresión igual a cero se conseguirá una desigualdad del tipo mayor o igual, esa expresión viene dada por la condición cuarta del teorema. Con todo ello se llega a:

$$(4)\geq F_t+\mu_{t+1}'(f_t-x_{t+1}'+x_t')+\lambda_t' g_t(x_t',u_t')=L_t(x_t',u_t',\mu_{t+1}',\lambda_t')$$

con lo cual se tiene:

$$L(X',U',\mu,\lambda)=\sum_{t=0}^{T-1} L_t(x_t',u_t',\mu_{t+1},\lambda_t)+F_T\geq$$

$$\geq\sum_{t=0}^{T-1} L_t(x_t',u_t',\mu_{t+1}',\lambda_t')+F_T=L(X',U',\mu',\lambda')$$

lo que demuestra la primera desigualdad de la expresión (2). Por todo ello queda demostrado que (X', U', μ', λ') es un punto de silla de la

función L y aplicando el teorema III.3. es también un máximo global.

El teorema III.4. es el equivalente al que aparece en Valqui (1987) pero para el problema **[I]**. Se trata según este autor de un principio del máximo suficiente que no exige condiciones adicionales a las funciones, ni siquiera que sean diferenciables. Sin embargo, una crítica a nuestro juicio es que exige resolver $T+1$ problemas, T-1 de los cuales tienen $n+m$ variables en lugar del principio clásico del máximo que sólo resuelve T problemas con m variables. Este intercambio entre operatividad y generalidad de los teoremas es algo que ya ha aparecido en otros epígrafes de este trabajo: si se quiere obtener condiciones más potentes, más manejables y más operativas tiene que ser a costa de disminuir su rango de aplicabilidad.

A partir de este principio del máximo, Valqui impone supuestos cada vez más restrictivos que le permiten obtener otros principios del máximo más simples. Al final, bajo supuestos adecuados de separabilidad, diferenciabilidad y pseudoconcavidad llega al principio del máximo clásico como condición suficiente de máximo global, pero esto es ya similar a lo desarrollado por la vía de las condiciones de K-T.

III.3.- ANÁLISIS DE SENSIBILIDAD Y DE ESTABILIDAD EN EL PROBLEMA DE C.O. DISCRETO

III.3.1.- Introducción

Es muy habitual que cuando se modelizan los comportamientos que se quieren estudiar no se conozcan los valores exactos de ciertos parámetros, o que estos parámetros estén sujetos a perturbaciones. Es de crucial importancia conocer cómo estas perturbaciones afectan a las variables de control del problema para de este modo poder tomar las decisiones adecuadas. Este estudio se conoce como análisis de sensibilidad y estabilidad de las soluciones respecto a los parámetros del problema. Este análisis afecta tanto a la solución (variables de estado y de control) como al valor óptimo de la función objetivo.

Este análisis no hay que confundirlo con el problema de optimización de parámetros aunque en cierto sentido puede resultar complementario. Este otro problema consiste en obtener el valor óptimo que debe tener un parámetro "a" que afecta a las funciones y que es desconocido. La solución, como apunta Tapiero (1977), pasa por transformar dicho parámetro en una variable dinámica de estado con valor constante, es decir, con una ecuación en diferencias del tipo a_{t+1}-a_t=0 y con valores iniciales y finales iguales pero libres. Por tanto, este análisis calcula el valor óptimo de un parámetro desconocido y no se ocupa de ver los efectos sobre la solución de perturbaciones en el valor de ese parámetro, es decir no es un análisis de sensibilidad.

El análisis de sensibilidad y estabilidad forma parte de cualquier metodología de resolución de un problema matemático hasta el punto que la resolución no quedaría completa sin algunos comentarios acerca de la sensibilidad y estabilidad de la solución.

Aunque la principal utilidad del análisis de sensibilidad es la de aproximar soluciones cercanas a la inicial con diferentes datos, las técnicas de perturbaciones

también se utilizan para obtener condiciones de óptimo, resultados de dualidad, algoritmos de resolución, etc..

El análisis de sensibilidad se ocupa de analizar perturbaciones locales o infinitesimales en algún dato del problema mientras que el análisis de estabilidad analiza perturbaciones finitas[22]. En el primer caso, el objetivo será encontrar las derivadas de la solución o de la función objetivo respecto al vector de perturbaciones; y, en el segundo caso, se trata de obtener cuánto puede cambiar la solución óptima o el valor de la función objetivo ante pequeñas variaciones en dicho vector. Evidentemente, ambos tipos de análisis están relacionados; por ejemplo, si se puede obtener las derivadas se podrá analizar perturbaciones finitas aproximando mediante la fórmula de Taylor.

La posibilidad de estandarizar el análisis de sensibilidad y estabilidad depende del tipo de problema matemático que se estudia. En programación lineal se ha conseguido esta estandarización gracias a la dominancia del método simplex y a las características apropiadas de las funciones; sin embargo, en programación no lineal esto no ha sido posible debido a la gran casuística que se puede plantear y es un campo en el que continuamente van apareciendo nuevos resultados. En C.O. discreto, los resultados de sensibilidad y estabilidad se basan en los obtenidos en P.N.L. (Malanowski, 1991), dada la equivalencia entre ambos tipos de problemas ya comentada.

III.3.2.- Análisis de sensibilidad en P.N.L.

Los resultados básicos de sensibilidad y estabilidad en P.N.L. se pueden ver en Fiacco (1983) donde, además, se presentan técnicas numéricas para calcular las derivadas direccionales respecto a los parámetros de perturbaciones. Aquí sólo vamos a recoger los principales resultados utilizando el punto de vista que resulta

[22]El análisis de estabilidad no se refiere, por tanto, a la estabilidad del sistema dinámico sino a la estabilidad de la solución, es decir, si la nueva solución ante un cambio finito en algún dato se puede acotar. La estabilidad del sistema dinámico no se plantea al estar en tiempo finito.

más fácil de entender a nuestro juicio y que consiste en combinar las condiciones de Kuhn y Tucker y el teorema de la función implícita. Este análisis se puede considerar como el clásico para estudiar la dependencia de la solución respecto de los parámetros y necesita de supuestos que restringen la aplicabilidad del análisis. Una aproximación más general y más completa, aunque también más compleja matemáticamente, se puede seguir en Levitin (1994).

Para desarrollar el análisis clásico se parte del Problema [II] de P.N.L. con un vector de parámetros o perturbaciones:

$$\begin{aligned} &Max \quad F(x,\theta) \\ &s.a: \quad h(x,\theta)=0 \\ &\qquad g(x,\theta)\geq 0 \end{aligned} \qquad [II]$$

donde $x\in\mathbb{R}^n$ y $\theta\in\mathbb{R}^t$, la función F es real con nxk variables y las funciones h y g son vectoriales (m igualdades y p desigualdades) y con nxk variables.

El análisis de sensibilidad trata de ver las condiciones para asegurar la existencia de un máximo local único que sea, además, una función de θ diferenciable en el valor inicial del parámetro que, sin pérdida de generalidad, puede establecerse en $\theta=0$. En segundo lugar, interesa la forma de llegar a calcular las derivadas direccionales que son las que informan del efecto de las perturbaciones locales sobre la solución. Estas derivadas difícilmente se podrán calcular directamente porque es raro disponer de la solución como función del vector de perturbaciones por lo que se hace necesario disponer de formulaciones para estas derivadas que se puedan someter a técnicas de resolución numérica. Cuando el problema inicial se resuelve a través de algún algoritmo numérico, es posible estimar resultados de sensibilidad a partir de la información que se obtiene a medida que el algoritmo se va acercando a la solución óptima.

La obtención de la solución inicial parte de las condiciones de Kuhn y Tucker

de primer orden, cuyas igualdades son las siguientes:

$$\nabla_x L(x,\mu,\lambda,\theta)=0$$

$$h(x,\theta)=0$$

$$\lambda g(x,\theta)=0$$

donde:

$$L=F(x,\theta)+\mu h(x,\theta)+\lambda g(x,\theta) \qquad (5)$$

A su vez, estas igualdades definen un sistema de $n+p+m$ ecuaciones implícitas con $n+p+m+k$ variables, que si se cumplen las hipótesis del teorema de la función implícita definen $n+p+m$ funciones de k variables diferenciables en el punto que define la solución de partida (la solución del problema para $\theta=0$). Estas funciones son $x(\theta)$, $\lambda(\theta)$ y $\mu(\theta)$ y permiten realizar el análisis de sensibilidad al ser diferenciables en $\theta=0$.

Así pues, las hipótesis que se plantean en los distintos teoremas de sensibilidad deben garantizar, por una parte, que las condiciones de Kuhn y Tucker den lugar a un único máximo local y, por otra, el cumplimiento del teorema de la función implícita para sistemas de ecuaciones. Aunque existen teoremas de sensibilidad más completos o que exigen condiciones menos fuertes vamos a destacar el siguiente por su claridad:

Teorema III.5.: Si en la solución inicial del Problema **[II]** (para $\theta=0$) se cumplen las siguientes hipótesis:

1.- Las funciones F, h y g admiten derivadas continuas de segundo orden respecto de x dos veces y respecto de x y de θ; y h y g admiten derivadas primeras continuas respecto de θ.

2.- Condiciones suficientes de segundo orden para máximo local.

3.- Condición de regularidad.

4.- Holgura complementaria estricta, es decir, el multiplicador asociado a una restricción de desigualdad saturada es estrictamente positivo. Además, el conjunto de multiplicadores debe ser único.

Entonces se puede asegurar la existencia, alrededor de $\theta=0$, de funciones diferenciables $x(\theta)$, $\lambda(\theta)$ y $\mu(\theta)$ de forma que $x(\theta)$ es el máximo local único del problema con multiplicadores asociados $\lambda(\theta)$ y $\mu(\theta)$. Además, para $\theta=0$, el valor de estas funciones coincide con la solución del problema sin perturbaciones y por tanto constituye el máximo local único.

Demostración: Las hipótesis 2 y 3 del teorema permiten asegurar que la solución del sistema de Kuhn y Tucker para $\theta=0$ es el único máximo local. Por otra parte, la hipótesis 1 garantiza que la matriz Jacobiana del sistema de ecuaciones implícitas esté bien definida y las hipótesis 2, 3 y 4 permiten, por último, asegurar la existencia de su inversa y así poder aplicar el teorema de la función implícita al sistema de ecuaciones de Kuhn y Tucker quedando demostrado el teorema.

Tras garantizar la existencia de una función diferenciable que permite obtener localmente la solución en función del vector de perturbaciones, hay que pasar al cálculo de las derivadas parciales y, como consecuencia, de las direccionales. El vector de derivadas parciales se puede calcular, al igual que en cualquier función implícita, haciendo referencia a la regla de la cadena una vez ha quedado asegurada la diferenciabilidad de las funciones. En este caso se plantea el encadenamiento:

$$\theta \longrightarrow (x,\ \mu,\ \lambda,\ \theta) \longrightarrow (\nabla L_x,\ h,\ \lambda g) = (0,\ldots,0)$$

Y se aplica la regla de la cadena para calcular las derivadas parciales deseadas en θ:

$$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \nabla^2_{xx}L & \nabla_x h & \nabla_x g & \nabla^2_{x\theta}L \\ \nabla_x h & 0 & 0 & \nabla_\theta h \\ \lambda\nabla_x g & 0 & g & \lambda\nabla_\theta g \end{pmatrix} \begin{pmatrix} \nabla_\theta x \\ \nabla_\theta \mu \\ \nabla_\theta \lambda \\ 1 \end{pmatrix}$$

donde cada elemento es, a su vez, una matriz de dimensión apropiada. Haciendo operaciones por bloques se llega a la expresión siguiente para las derivadas parciales:

$$\nabla_\theta y = -M^{-1}N$$

donde:

$$M = \begin{pmatrix} \nabla^2_{xx}L & \nabla_x h & \nabla_x g \\ \nabla_x h & 0 & 0 \\ \lambda\nabla_x g & 0 & g \end{pmatrix}$$

$$N = \begin{pmatrix} \nabla^2_{x\theta}L \\ \nabla_\theta h \\ \lambda\nabla_\theta g \end{pmatrix}$$

$$\nabla_\theta y = \begin{pmatrix} \nabla_\theta x \\ \nabla_\theta \mu \\ \nabla_\theta \lambda \end{pmatrix}$$

La matriz M es la Jacobiana del sistema de ecuaciones de Kuhn y Tucker respecto de x, λ y μ y admite inversa dadas las hipótesis del teorema III.5..

Las derivadas direccionales se calculan a partir de las derivadas parciales ya que las funciones son diferenciables al cumplirse el teorema de la función implícita.

A partir de las derivadas parciales se puede calcular una estimación de primer orden de los valores de la solución al producirse un cambio en el vector de parámetros aplicando la fórmula de Taylor, dada la diferenciabilidad de las funciones:

$$y(\theta) = y(0)+M(0)^{-1}N(0)\theta+R(\theta) \quad ; \quad \lim_{\theta\to 0}R(\theta)=0$$

De esta manera, se consigue aproximar la nueva solución ante perturbaciones en los valores de los parámetros. Evidentemente se pueden conseguir aproximaciones más exactas exigiendo un grado mayor de diferenciabilidad. En este sentido, para que las funciones $x(\theta)$, $\lambda(\theta)$ y $\mu(\theta)$ sean diferenciables de grado q, las condiciones del teorema III.5. se deben cumplir en grado $q+1$.

Existen otros teoremas de sensibilidad con condiciones menos fuertes, entre ellos cabe destacar la referencia que de Jittorntrum hace Malanowski para aplicarla al problema de C.O. discreto. Este teorema no exige la hipótesis 4 de holgura complementaria estricta pero a cambio sustituye la hipótesis 2 de suficiencia por la llamada condición de suficiencia fuerte de segundo orden que, en el caso de máximo local es la siguiente: la matriz hessiana de la función lagrangiana asociada al problema debe representar una determinada forma cuadrática restringida definida negativa, en concreto:

$$z^T \nabla^2_{xx} L\, z<0 \qquad \forall\, z\neq 0;\ z\in Z$$

$$Z=\{\, z\in\mathbb{R}^n \,/\, \nabla_x h z=0,\ \nabla_x g_k z=0,\ \forall k \,/\, g_k=0 \,\}$$

También es de utilidad el análisis de sensibilidad del valor de la función objetivo. Este análisis ha sido tratado muy frecuentemente en lo que hace referencia a perturbaciones en el término independiente de las restricciones, debido a su importancia en aplicaciones económicas. En efecto, la interpretación de los multiplicadores de Kuhn y Tucker como precios sombra se deriva del análisis de sensibilidad del valor de la función objetivo y es conocida desde hace tiempo (Dorfman, 1969). Sin embargo, este resultado es sólo un caso particular del análisis de sensibilidad más general que se obtiene a partir del teorema III.6.

Teorema III.6.: Dadas las hipótesis del teorema III.5. se cumple en un entorno de $\theta=0$ lo siguiente:

1.- El valor óptimo de la función objetivo coincide con el valor óptimo de la función lagrangiana: $F^*(\theta)=L^*(\theta)$.

2.- Las derivadas del valor óptimo de la función objetivo se calculan a través de las de la función lagrangiana:

$$\nabla_\theta F(\theta)^* = \nabla_\theta F+\mu\nabla_\theta h+\lambda\nabla_\theta g \qquad (6)$$

Demostración: La conclusión 1 se deduce de la expresión (5) de la función Lagrangiana y del hecho de que el punto óptimo cumple las condiciones de Kuhn y Tucker dadas las hipótesis del teorema.

La conclusión 2 se obtiene diferenciando respecto de θ ambos miembros de la igualdad que se recoge en la conclusión 1 y aplicando

las hipótesis.

Con la expresión (6) para las derivadas del valor óptimo de la función objetivo se puede aproximar por Taylor el nuevo valor de la función objetivo ante perturbaciones en los parámetros partiendo del valor inicial $\theta=0$:

$$F^*(\theta) = F^*(0)+\nabla_\theta F^*(0)\theta+R'(\theta) \quad ; \quad \lim_{\theta\to 0} R'(\theta)=0$$

Si la perturbación se encuentra sólo en el término independiente de las restricciones se tiene el caso particular más conocido ya que entonces las funciones son:

$$F(x,\theta) = F(x)$$
$$h(x,\theta) = h(x)-\theta_1$$
$$g(x,\theta) = g(x)-\theta_2$$

y la expresión de las derivadas parciales respecto al vector de parámetros queda, en este caso, de la siguiente manera:

$$\nabla_{\theta_1} F^* = -\mu$$
$$\nabla_{\theta_2} F^* = -\lambda$$

Evidentemente se pueden obtener otros resultados para otros casos particulares y también es posible llegar a formulaciones para derivadas de orden superior o derivadas cruzadas.

III.3.3.- Análisis de sensibilidad en C.O. discreto.

Para realizar el análisis de sensibilidad en el problema de C.O. discreto a nivel teórico y de forma general, hay que aplicar los resultados generales en P.N.L. al enunciado del problema de C.O. discreto. Podemos considerar, por ejemplo, el siguiente problema con el vector de perturbaciones:

$$Max \quad F(X,U,\theta)=\sum_{t=0}^{T-1} F_t(x_t,u_t,\theta)+F_T(x_T,\theta)$$

s.a:

$$x_{t+1}-x_t=f_t(x_t,u_t,\theta) \qquad t=0,...,T-1 \qquad \textbf{[III]}$$

$$g_t(x_t,u_t,\theta)\geq 0 \qquad t=0,...,T-1$$

$$x_0=r(\theta)$$

$$T \ \textit{conocido}$$

En consecuencia, el vector de variables, que en P.N.L. se ha llamado x, está formado ahora por el vector de estados y de controles en cada periodo de tiempo, es decir, (X, U). Y la función lagrangiana asociada es:

$$L(X,U,\mu,\lambda,\theta)=\sum_{t=0}^{T-1} F_t(x_t,u_t,\theta)+F_T(x_T,\theta)+$$

$$+\sum_{t=0}^{T-1} \mu_{t+1}(f_t(x_t,u_t,\theta)-x_{t+1}+x_t)+\sum_{t=0}^{T-1} \lambda_t g_t(x_t,u_t,\theta)$$

A su vez, el teorema III.5. se transforma en el siguiente:

Teorema III.7.: Si en la solución inicial del Problema **[III]** (para $\theta=0$) se cumplen las siguientes hipótesis:

1.- Las funciones F_t, F_T, f_t, y g_t admiten derivadas continuas de segundo orden respecto a todos sus argumentos (X, U y θ)[23].

2.- Se cumplen las condiciones suficientes de segundo orden para máximo local (por ejemplo, las hipótesis del teorema III.2.).

3.- Se cumple la condición de regularidad: los gradientes de las funciones componentes de g_A junto con los gradientes de las funciones f_t forman un sistema linealmente independiente.

4.- Se cumple la holgura complementaria estricta, es decir, el multiplicador asociado a una restricción de desigualdad saturada es estrictamente positivo.

Entonces se puede asegurar la existencia, alrededor de $\theta=0$, de funciones diferenciables $X(\theta)$, $U(\theta)$, $\lambda(\theta)$ y $\mu(\theta)$ de forma que $(X(\theta), U(\theta))$ es el máximo local único del problema con multiplicadores asociados $\lambda(\theta)$ y $\mu(\theta)$. Además, para $\theta=0$ el valor de estas funciones coincide con la solución del problema sin perturbaciones y por tanto constituye el máximo local único.

Análogamente se podría enunciar un teorema similar al III.6. pero concretado al problema de C.O. discreto.

Los teoremas adaptados al problema de control admiten distintas formulaciones, por ejemplo utilizando la condición de suficiencia fuerte de segundo orden como se hace en Malanowski (1991).

[23]Esta condición se ha relajado respecto a la 1 del teorema III.5. para no recargar excesivamente la notación.

Si se desea calcular el valor de las derivadas respecto al vector de perturbaciones o la estimación del vector de variables o de la función objetivo ante pequeños cambios en los datos se utilizarán las formulaciones obtenidas en P.N.L. pero con la función lagrangiana del problema de control óptimo.

Todos estos resultados difícilmente se pueden obtener analíticamente porque la solución no es normal que quede explícita. Lo lógico es obtenerlos a través de técnicas numéricas, tanto en P.N.L. como en C.O. discreto.

Lo más práctico es extraer información acerca de la sensibilidad a medida que el algoritmo que se utiliza para obtener la solución inicial vaya aproximándose al óptimo. En el proceso de obtención de dicho óptimo va apareciendo información relevante para el análisis de sensibilidad y estabilidad. Otra posibilidad es adaptar algún método numérico en P.N.L. para estimar los valores de las derivadas desarrolladas antes a nivel teórico, en este sentido existe una recopilación de métodos en Fiacco (1983, cap. 6 y 7). Las nuevas aportaciones que se recogen en Levitin (1994) también deben impulsar el desarrollo de nuevos métodos numéricos de optimización.

En C.O. discreto ocurre lo mismo con el agravante que sólo se podrán utilizar ciertos métodos numéricos de P.N.L., dada la gran dimensión de los problemas, o algún método específico de este tipo de problemas. A partir de estos métodos se puede llegar, como resultado adicional, a información sobre la sensibilidad. Esto es lo que ocurre, por ejemplo, en Vinter (1988) en el caso de que las perturbaciones sólo afecten a los términos independientes de las restricciones.

La dificultad de estos temas se refleja en que los paquetes informáticos en P.N.L. y C.O. discreto no recogen la posibilidad de análisis de sensibilidad generales. Sí que se puede analizar perturbaciones en los términos independientes de las restricciones y de tipo finito, de forma que se obtiene la nueva solución para el nuevo valor del parámetro (análisis de post-optimización). Otras veces se exigen condiciones especiales al problema para poder realizar el análisis, como linealidad,

convexidad, separabilidad, etc.. Es de esperar que, a medida que surjan nuevas investigaciones se vayan perfeccionando los paquetes informáticos en la línea de introducir el análisis de sensibilidad aunque resulta improbable que se llegue al nivel de estandarización que existe en programación lineal.

III.4.- OTRAS EXTENSIONES

III.4.1.- Aproximación al control óptimo estocástico en tiempo discreto

El problema se ha planteado hasta ahora de forma determinista, se conoce el valor de todos los parámetros del problema y las relaciones dinámicas entre las variables se pueden determinar de forma exacta. El análisis cambia sustancialmente si se introduce la incertidumbre, pasando así al problema de Control Óptimo Estocástico. Esta consideración aumenta el realismo del modelo pero la consecuencia que se deriva es que las técnicas de resolución se complican de forma importante, obteniéndose soluciones analíticas sólo en problemas muy sencillos, y acudiendo generalmente a técnicas numéricas (sobre todo simulaciones[24]) como única vía de observar las posibles trayectorias solución.

Hay veces, sin embargo, que el problema se puede reducir a uno determinista. Esto ocurre cuando se conoce la distribución de probabilidad de algún parámetro incierto del problema, en cuyo caso sus variables descriptivas (normalmente la media y la varianza) se transforman en nuevas variables de estado y su evolución en las nuevas ecuaciones dinámicas. Este problema se llama también pseudo-estocástico (Neck, 1984).

El caso más común de fuente de incertidumbre y que da lugar al problema de control estocástico propiamente dicho es aquél que afecta al sistema dinámico, es decir, el caso en que la evolución de las variables de estado depende de alguna componente aleatoria. En este caso, se utilizan elementos básicos de la modelización dinámica estocástica en tiempo discreto como los procesos estocásticos y las ecuaciones en diferencias estocásticas (ver por ejemplo en Chow, 1975; en Malliaris y Broch, 1982; y en Neck, 1984). La aplicación de estas técnicas a problemas de C.O. discreto ha sido tratada sobre todo en Tapiero (1988).

---

[24]Los aspectos generales de la técnica de la simulación en modelos dinámicos discretos se pueden consultar en Delaney y Vaccari (1989).

Las fuentes de incertidumbre se completan con aquélla que afecta a la observación de los valores de las variables de estado. Esto significa que los estados no siempre se pueden observar directamente sino que se observan ciertas variables que están relacionadas de forma estocástica con las de estado. La estimación del valor de las variables de estado recibe el nombre de *filtering*. Este problema ha sido tratado ampliamente sólo en el caso de sistema dinámico lineal (problemas lineales o lineales-cuadráticos) donde la estimación óptima de la variable de estado se conoce como filtro de Kalman. Con esa estimación se obtiene el control óptimo tratando el problema como uno determinista si no hay otras fuentes de incertidumbre o se pasa al problema de control óptimo estocástico propiamente dicho. En el caso de problemas distintos al lineal-cuadrático se pueden utilizar técnicas de aproximación basadas en el desarrollo de Taylor u otras más complejas.

El problema del control óptimo estocástico en tiempo discreto surge como confluencia de dos elementos: el control óptimo discreto determinista y los sistemas dinámicos discretos estocásticos. El primer elemento es el estudiado en el capítulo II y el segundo se reduce a tratar las ecuaciones en diferencias estocásticas.

Una **ecuación en diferencias estocástica** se obtiene a partir de una ecuación determinista al añadir una componente que, en cada periodo, es una variable aleatoria (ϵ_t) de media cero y varianza σ^2 y que es independiente de unos periodos a otros. Un ejemplo de ecuación en diferencias estocástica de primer orden es:

$$x_{t+1}-x_t = f(x_t)+\epsilon_t \qquad (7)$$

donde $f(x_t)$ es la media del incremento en la variable dada x_t. Evidentemente, la componente aleatoria no tiene porque aparecer de forma aditiva. La sucesión de variables aleatorias $\{\epsilon_t;\ t=0,1,\ldots,T-1\}$ es un proceso estocástico. A su vez, la sucesión $\{x_t;\ t=1,2,\ldots,T\}$ resulta ser un proceso estocástico de Markov (el valor de la variable en un periodo depende del valor en el periodo anterior, y por tanto, la información del pasado queda recogida en el valor de la variable en el presente).

La resolución de la ecuación (7) no da lugar a una relación determinista entre la variable de estado y el tiempo, y el objetivo es obtener su distribución de probabilidad para cada periodo. En la mayoría de los casos, sin embargo, esto no es posible y hay que conformarse con obtener medias, varianzas y covarianzas. No obstante, si ϵ_t se distribuye según una normal, x_0 es conocido de forma cierta y la ecuación en diferencias es lineal, entonces las variables de estado se distribuirán según una normal y el conocimiento de sus momentos es suficiente para conocer su distribución de probabilidad completamente. El cálculo de las medias, varianzas y covarianzas se puede realizar aplicando el método recursivo típico de las ecuaciones en diferencias y aplicando el operador esperanza. La solución analítica raramente es posible en tiempo discreto siendo más fácil obtenerla en tiempo continuo. La resolución por métodos numéricos, en cambio, es más exacta en tiempo discreto.

Para observar el paralelismo entre ecuaciones estocásticas diferenciales y en diferencias, la componente aleatoria que se añade en la ecuación en diferencias debe corresponder a los incrementos asociados a un proceso de Wiener:

$$x_{t+1}-x_t = f(x_t)+\Delta v_t \qquad \text{con } \Delta v_t \sim N(0,\sigma_t^2)$$

y si se pasa a un proceso de Wiener estándar:

$$\Delta x_t = f(x_t)+\sigma_t \Delta w_t \qquad \text{con } \Delta w_t \sim N(0,1) \qquad (8)$$

La expresión (8) es la que se obtiene si se discretiza una ecuación diferencial estocástica de Itô y se toma $\Delta t=1$[25]. La distribución para Δw_t inducirá una distribución para Δx_t. En el caso de que la varianza sea constante será una normal de media $f(x_t)$ y varianza σ^2. Aplicando el método recursivo se puede intentar solucionar la ecuación en diferencias, en el sentido de obtener una distribución de probabilidad para la variable x_t o al menos su media y varianza. Así, suponiendo que

[25]La ecuación diferencial estocástica de Itô es la más habitual en la modelización dinámica bajo incertidumbre aunque, recientemente, se han utilizado otras más completas que resultan apropiadas en ciertos cálculos financieros (ver en Blenman y otros, 1995).

se parte de $x_0=0$ se tiene:

$$E(x_1)=f(x_0)\ ,\ \ Var(x_1)=\sigma^2$$

$$E(x_2)=f(x_0)+f(x_1)\ ,\ \ Var(x_2)=2\sigma^2$$

$$\ldots$$

$$E(x_t)=f(x_0)+\ldots+f(x_{t-1})\ ,\ \ Var(x_t)=t\sigma^2$$

$$Cov(x_t,x_s)=\min\{t,s\}\ \sigma^2$$

La distribución de probabilidad depende de la función f. Así, si dicha función es lineal las variables x_t están distribuidas según una normal.

Existen otras formas de obtener ecuaciones en diferencias estocásticas que no se pueden reducir a la forma equivalente de Itô. En cualquier caso, la resolución parte de aplicar el método recursivo y observar si de las expresiones que se obtienen se puede extraer la distribución de probabilidad y/o la media y varianza del término general de la sucesión de variables aleatorias. Este proceso sólo es posible ante ecuaciones relativamente simples, siendo necesario utilizar técnicas más complejas en casos más generales (ver, por ejemplo, en Aoki, 1989).

La combinación del sistema dinámico estocástico y el resto de elementos de un problema de control óptimo determinista da lugar al control óptimo estocástico. El sistema dinámico describe la evolución de las variables de estado en función de las variables de control y de una componente estocástica. A su vez, la función objetivo, al depender de una variable de estado aleatoria, se optimizará en referencia a su valor medio. De esta manera el enunciado del problema puede adoptar la siguiente forma que, aunque no es la más general, mantiene el paralelismo con el problema determinista:

$$Max \quad E[\sum_{t=0}^{T-1} F_t(x_t,u_t)+F_T(x_T)]$$

$$s.a:$$

$$x_{t+1}-x_t=f_t(x_t,u_t)+\varepsilon_t \qquad t=0,...,T-1$$

$$g_t(x_t,u_t)\geq 0 \qquad t=0,...,T-1 \qquad \text{[IV]}$$

$$x_0 \quad conocido$$

$$T \quad conocido$$

$$\varepsilon_t \sim N(0,\sigma^2)$$

El Problema **[IV]** es sólo un posible enunciado dado que la componente aleatoria también puede afectar a las funciones intermedias F_t, a las funciones f_t, a las restricciones o incluso a la condición inicial en el caso de valor inicial libre. Por otra parte, puede haber incertidumbre en la observación de las variables de estado, la componente aleatoria puede no distribuirse como una normal, etc..

Siguiendo a Arkin y Evstigneev (1987) la definición correcta del problema desde un punto de vista matemático necesita distintos supuestos. Por una parte están los que se pueden considerar equivalentes a los del caso determinista (convexidad, regularidad, diferenciabilidad) y por otra, aquéllos específicos del caso estocástico que hacen referencia a la mesurabilidad de las funciones, definición de los espacios de probabilidad, etc..

La resolución del problema es compleja y, en términos analíticos, prácticamente imposible salvo en problemas limitados como el lineal-cuadrático o alguno más general en el caso continuo (ver en Tapiero, 1994). El Principio del Máximo estocástico es una más de las técnicas posibles. Parte de la definición de un Hamiltoniano que se construye igual que el del caso determinista pero que ahora dependerá de la variable aleatoria: $H_{t+1}(x_t, u_t, \lambda_{t+1}, \mu_t, \epsilon_t)$. El control óptimo es aquél que maximiza la función $E[H_{t+1}]$ tomada como función de u_t de entre los controles

admisibles. Otra técnica, también paralela al caso determinista, es la Programación Dinámica estocástica, que se basa en el conocido "Principio de optimalidad" de Bellman. Esta técnica necesita menos requisitos que el Principio del Máximo estocástico aunque no existe un algoritmo general para su aplicación. Un tratamiento a fondo de estas dos técnicas y su interrelación se puede seguir en Arkin y Evstigneev (1987).

La dificultad propia del problema de C.O. discreto aumenta todavía más en el caso estocástico, de ahí que estas técnicas no resulten operativas y se deban utilizar otras como la simulación o la reducción a un problema determinista.

III.4.2.- Un Principio del Máximo sin la condición de convexidad direccional

El hecho de que el Principio del Máximo en tiempo continuo no necesite de hipótesis sobre el conjunto de estados extendidos alcanzables ha estimulado los intentos para relajar la hipótesis de convexidad de dicho conjunto que se exige para dar validez al P.M. en tiempo discreto[26]. Así, se pasó pronto a la hipótesis de convexidad direccional. Otros intentos de relajación son más recientes y entre ellos se puede destacar a Nahorski y otros (1984) y a Vinter (1988). Hay que decir que estas aportaciones son puramente matemáticas ya que dicha relajación se hace a costa de aumentar la complejidad matemática de las formulaciones, sin apenas consecuencias para el aspecto aplicado.

La aportación de Nahorski, Ravn y Valqui es más interesante de comentar porque lleva consigo la definición de una función hamiltoniana distinta a la habitual y da lugar, para estos autores, a un principio del máximo generalizado.

El P.M. clásico parte de la función hamiltoniana típica que, como se sabe,

[26]No hay que olvidar que se puede dar validez al Principio del Máximo cuando el problema se reduce a uno de Programación Matemática y se dan condiciones apropiadas de cualificación de restricciones y convexidad.

es lineal respecto a la variable de coestado o, lo que es lo mismo, respecto al vector de multiplicadores de Kuhn y Tucker asociado al sistema dinámico. La función hamiltoniana que proponen estos autores es una función no lineal, donde en lugar de la variable de coestado aparece una función más general de la variable de estado ($\pi_t(x_t)$). Utilizando la nomenclatura de estos autores aparece la siguiente hamiltoniana:

$$NH_t(x_t,u_t,\pi_{t+1})=F_t(x_t,u_t)+\pi_{t+1}(f_t(x_t,u_t)) \qquad (9)$$

y con ella se enuncia el principio del máximo generalizado, que es el siguiente:

> *Teorema III.8.:* Si (X^*, U^*) es una solución óptima del problema **[I]** de C.O. discreto, entonces existe una sucesión de funciones $\pi_t^*(x_t)$ tal que u_t^* maximiza el hamiltoniano no lineal expresado en (9) para $t=0,\ldots,T-1$.

Se puede observar que el hamiltoniano clásico coincide con el generalizado en el caso en que se cumpla:

$$\pi_{t+1}(f_t) = \mu_{t+1}f_t = \mu_{t+1}(x_{t+1}-x_t)$$

por tanto el sistema hamiltoniano de ecuaciones en diferencias que se obtenía se puede generalizar teniendo en cuenta la siguiente equivalencia:

$$\mu_{t+1}=\frac{\partial \pi_{t+1}}{\partial x_{t+1}}$$

El problema, tal y como resaltan estos autores, es construir esta función π_t. La idea que aportan para que se cumpla el sistema hamiltoniano dinámico es tomar una familia de funciones del tipo:

$$\pi_t=\mu_t(x_t-x_{t-1})+k_t(x_t-x_{t-1})$$

donde el multiplicador es el mismo que en el P.M. clásico y la función k_t tiene una derivada parcial respecto a x_t igual a cero.

Mediante algunos ejemplos sencillos se encuentran las funciones π_t adecuadas, partiendo de que se conoce la solución de antemano. Cuando la solución no es conocida se necesita un algoritmo iterativo para encontrar estas funciones entre las del tipo cuadrático. Sin embargo, en Nahorski y Ravn (1988b) no se concreta este tipo de algoritmo, al tiempo que reconocen que esta herramienta no es flexible en la práctica, aunque su aportación teórica puede servir para crear algoritmos o mejorar los existentes de cara a la optimización de problemas multietápicos.

CAPÍTULO IV:

MÉTODOS DE RESOLUCIÓN DE PROBLEMAS DE CONTROL ÓPTIMO DISCRETO

IV.1.- INTRODUCCIÓN

Este capítulo sirve de conexión entre los capítulos de contenido más teórico desarrollados hasta ahora y los de carácter aplicado que se abordan posteriormente. Se trata de resolver el problema de C.O. discreto una vez se han estudiado las condiciones que debe cumplir el óptimo.

La resolución analítica es compleja debido a la dimensión que alcanza el problema en cuanto se pretende modelizar relaciones económicas o de otro tipo con cierto grado de aproximación a la realidad. Además, el tratamiento analítico en tiempo discreto es más complejo que en tiempo continuo.

La solución óptima en problemas dinámicos consiste en obtener funciones del tiempo que, en el caso discreto, son funciones discretas o sucesiones y, en este sentido, lo ideal siempre es encontrar el término general de la sucesión dado que, de esta manera, se consigue una relación válida entre las variables (de control y de estado) y el tiempo, para todo el horizonte de planificación. Sin embargo, esto sólo es posible en problemas de reducidas dimensiones y con funciones muy específicas (por ejemplo, las lineales). Este método de resolución analítico es, por tanto, el más deseable porque es el que más información aporta acerca de la solución, pero también el menos operativo porque es el que más exige al enunciado del problema. Además, la resolución analítica de ecuaciones en diferencias (el sistema dinámico hamiltoniano) sólo tiene un procedimiento generalizado para el caso de ecuaciones lineales mientras que las no lineales requiere técnicas más complejas que sólo son válidas bajo determinados supuestos[27]. Por otra parte, el uso del método analítico es totalmente inapropiado a medida que crece el número de variables y de periodos del problema.

Una vez asumidas las limitaciones del método analítico se hace necesario utilizar otros métodos de resolución del problema que se engloban en los llamados

[27]Un manual clásico de resolución de ecuaciones en diferencias es Goldberg (1964).

métodos numéricos. La resolución numérica apoyada en soporte informático es la mejor opción cuando no es posible la resolución analítica. Además, en tiempo discreto, el tratamiento numérico consigue soluciones más exactas que en tiempo continuo donde el proceso de discretización lleva implícito ya cierto error. Estos métodos se diferencian del anterior en que no buscan el término general de la sucesión solución sino que tratan de determinar cada uno de los elementos de la sucesión. De esta manera, se consigue una relación punto a punto entre la variable independiente y las variables de estado y control y no una relación válida para todo el horizonte de planificación. Sin embargo, la representación gráfica de las sucesiones puede servir para deducir ciertas consideraciones cualitativas de la solución, y además puede orientar una aproximación analítica por algún tipo de curva (*fitting*).

El método numérico puede ser en algunos problemas un método exacto. Efectivamente, si el control óptimo, como solución al problema de maximización de la hamiltoniana, se obtiene de forma analítica y si se parte de un valor inicial de las variables de estado, entonces la resolución numérica de las ecuaciones en diferencias llevará a una solución exacta sin más que aplicar la forma de recurrencia que lleva consigo toda ecuación en diferencias. El método numérico también es un método exacto en problemas lineales o cuadrático-lineales.

En problemas generales con gran número de periodos, de variables y de restricciones, hay que acudir a los métodos numéricos aproximados para resolver el problema. Los primeros métodos que se desarrollaron son los propios de la Programación Matemática (los de P.N.L.) y de la Programación Dinámica. Recientemente ha empezado a aparecer *software* específico para el problema de C.O. discreto.

Para la resolución de problemas que nosotros perseguimos creemos que lo más adecuado es el uso de *software* estándar de Programación no Lineal. La razón principal para ello es el hecho de que está más desarrollado, es más asequible y abarca una amplia variedad de problemas.

Los métodos basados en técnicas de Programación Dinámica no suelen estar estandarizados porque utilizan técnicas distintas según el tipo de problemas que resuelve; además, tienen unos requerimientos de memoria de almacenamiento de datos más exigentes.

Los métodos que utilizan la teoría del C.O. discreto o el P.M. discreto son todavía incipientes y tienen un uso muy limitado al ámbito de los investigadores que los están desarrollando. Como luego se comentará, es de esperar que estos métodos sean más eficientes que los generales de P.N.L. y que poco a poco se vaya generalizando su uso.

La resolución numérica se caracteriza por el uso de algoritmos para llegar a los valores óptimos de las variables. Los algoritmos constan de una serie de etapas en las que se realiza un conjunto de operaciones tras las cuales se llega a una aproximación a la solución. Repitiendo las etapas se consiguen sucesivas aproximaciones. Todo algoritmo contiene también un criterio de parada tras el que, o se da por buena la última aproximación conseguida, o se deduce que no se ha llegado a ninguna solución.

Las operaciones que se realizan en cada etapa definen el método numérico utilizado y de él dependen aspectos importantes como la convergencia, generalidad y eficiencia del algoritmo. La convergencia puede ser de distintos tipos pero, básicamente, implica que las aproximaciones al final de cada etapa sean cada vez mejores y el criterio de parada indica que se ha llegado a una solución aceptable. La generalidad del algoritmo supone que se pueda aplicar a un conjunto lo más amplio posible de problemas. Por último, la eficiencia hace referencia al tiempo de cálculo necesario para llegar a la solución.

La resolución numérica del problema de C.O. discreto se puede enfocar de dos maneras. Dado que es un problema equivalente al de programación matemática, la primera opción es utilizar algoritmos de P.N.L. con la ventaja de que están más estudiados y desarrollados. Sin embargo, en estos algoritmos tanto las variables de

control como las de estado son variables de decisión y las ecuaciones dinámicas se interpretan como restricciones adicionales, es decir, son métodos que no aprovechan la estructura específica del problema de C.O. discreto lo cual hace pensar que se puedan lograr otros algoritmos más eficientes.

La segunda opción es utilizar algoritmos específicos del problema de C.O. discreto que, aunque todavía poco desarrollados, deben ser más eficientes que los anteriores. Ambas opciones se relacionan ya que a menudo utilizan métodos numéricos comunes.

A continuación se exponen de forma breve algunos de los métodos numéricos en P.N.L. clasificados según el tipo de problemas a los que se aplican. La descripción de estos métodos no es exhaustiva y para una visión más profunda de éstos y otros métodos numéricos se puede consultar manuales como Fletcher (1991) y Bazaraa y otros (1993).

El epígrafe IV.3. está dedicado a dos casos especiales de problemas de C.O. discreto: los problemas lineales y los lineales-cuadráticos. En estos problemas las funciones utilizadas cumplen una serie de características deseables, lo cual ha propiciado la existencia de algoritmos generales de P.N.L. muy eficientes. Su estudio se justifica también por las aplicaciones que se desarrollan en los siguientes capítulos, las cuales se modelizan a través de este tipo de problemas. Por último, se comentan los métodos que se están utilizando para resolver el problema de C.O. discreto, se presentan algunas características del *software* disponible y se justifica la elección del mismo.

IV.2.- MÉTODOS NUMÉRICOS EN PROGRAMACIÓN NO LINEAL

IV.2.1.- Problemas sin restricciones.

La importancia de estos problemas radica en que suelen formar parte de algún paso dentro de algoritmos para problemas con restricciones. En el problema con una sola variable (*line search*) se utilizan métodos que realizan, en cada etapa, una acotación cada vez más estrecha del intervalo en el que se encuentra el óptimo (intervalo de incertidumbre). Para ello hay métodos que evalúan la función objetivo en un punto interior del intervalo y lo comparan con el valor en los extremos; otros métodos evalúan la derivada de la función en un punto interior y según su signo acotan el intervalo por un extremo o por el otro. Estos métodos requieren, para ser consistentes, alguna hipótesis acerca de la cuasiconcavidad o pseudoconcavidad (en el caso de máximo) de la función objetivo.

En los problemas sin restricciones pero con varias variables (*multidimensional search*) existen distintos métodos con la característica común que realizan dos operaciones en cada etapa k:

- Seleccionar una dirección $d_k \in \mathbb{R}^n$ apropiada en la que se debe desplazar el punto para mejorar la función (*search direction*).

- Seleccionar la magnitud $\lambda_k \in \mathbb{R}$ de ese desplazamiento (*step size*).

De esta manera en la etapa k se pasa de un punto x_k a otro x_{k+1} que cumple $x_{k+1}=x_k+\lambda_k d_k$.

La búsqueda del tamaño del desplazamiento en cada etapa (*step size*) se reduce a un problema no lineal sin restricciones en una dimensión. Efectivamente, se trata de resolver el siguiente problema de *line search*:

$$Max \quad F(x_k+\lambda_k d_k) \qquad [I]$$

Para resolver el problema de encontrar la dirección de búsqueda más apropiada (*search direction*) existen diversos métodos. Uno de los más básicos es el **método del gradiente**, que se basa en utilizar como dirección de búsqueda la del gradiente normalizada por ser la dirección de máximo crecimiento (si se minimiza se utiliza la dirección opuesta). Evidentemente ello está limitado al caso de función diferenciable. Tomando como base el método del gradiente se han desarrollado otros que han ido corrigiendo la dirección del gradiente para hacerla más eficiente y evitar algunos inconvenientes. Entre estos cabe citar el **método de Newton** (que utiliza el Hessiano) o **métodos de cuasi-Newton** (que utilizan aproximaciones al Hessiano)[28]. Por último cabe citar los **métodos del gradiente conjugado** que corrigen la dirección del gradiente utilizando una apropiada combinación lineal convexa del gradiente y de la dirección usada en la iteración anterior[29].

IV.2.2.- Problemas con restricciones

Un primer conjunto de métodos numéricos que resuelve el problema con restricciones se basa en pasar a un problema asociado sin restricciones a través de la incorporación de una función auxiliar a la función objetivo y de forma que la solución de este problema sea la misma que la del problema original. Esta técnica es la que se utiliza en los **métodos de penalización y en los métodos de barrera.**

En los **métodos de penalización** se añade una función de penalización cuyo objetivo es doble: no penalizar a los puntos factibles y penalizar a los puntos

[28]Los métodos de cuasi-Newton más utilizados actualizan la aproximación al hessiano en cada iteración, teniendo en cuenta siempre que el resultado debe ser una matriz simétrica y definida negativa. La fórmula de actualización más utilizada sigue siendo la llamada BFGS (Broyden, Fletcher, Goldfarb and Shanno) y data de 1970.

[29]De ellos destaca el de Fletcher y Reeves de 1964. Los métodos de gradiente conjugado son, en general, menos eficientes que los de cuasi-Newton pero, al requerir menos cálculos en cada iteración, son los más aconsejables en problemas de gran dimensión.

infactibles. Existen muchas funciones de este tipo y cada una de ellas da lugar a distintos algoritmos de penalización. Por ejemplo, dado el problema:

$$\begin{aligned} &Max \quad F(x) \\ &s.a: \\ &\quad h(x)=0 \\ &\quad g(x)\leq 0 \end{aligned} \qquad [II]$$

una de las funciones de penalización más habituales es la del valor absoluto:

$$\alpha(x) = \max \{0, g(x)\} + |h(x)|$$

o si se busca que sea diferenciable:

$$\alpha(x) = \max \{0, g(x)\}^2 + h^2(x)$$

Estas funciones de penalización dan lugar al siguiente problema sin restricciones:

$$Max \quad F(x) - \mu\alpha(x)$$

donde μ es un número arbitrariamente grande. Se observa que la función de penalización toma el valor cero en puntos factibles y valor positivo en puntos infactibles, incurriendo en este caso en una penalización total igual a $\mu\alpha(x)$ que disminuye el valor de la función objetivo.

Los métodos de penalización empiezan con un punto inicial infactible y un escalar $\mu > 0$ arbitrario. El siguiente paso es resolver el problema asociado sin

restricciones utilizando alguno de los métodos apropiados ya mencionados. Si tras resolverlo se obtiene un punto con una penalización total suficientemente pequeña se acepta como solución óptima y si no, se aumenta el valor del escalar. De esta manera se va acercando al punto óptimo desde puntos infactibles.

En los **métodos de barrera**, en cambio, el mecanismo es acercarse al punto óptimo desde puntos factibles. En estos métodos las restricciones de igualdad se deben incorporar a un conjunto X o, si es posible, utilizarlas para sustituir unas variables en función de las otras y trabajar con el problema con menos variables; esto es necesario porque los puntos que van apareciendo en cada paso son interiores al conjunto de oportunidades.

La función de barrera que se incorpora tiene el objetivo de establecer una barrera a los puntos que alcancen la frontera del conjunto de oportunidades desde el interior. Para ello dicha función debe tomar valores positivos para puntos interiores y el valor infinito para puntos frontera. Por ejemplo, dado el problema:

$$\begin{aligned} &Max \quad F(x) \\ &s.a: \\ &\quad g(x) \leq 0 \\ &\quad x \in X \end{aligned} \qquad \text{[III]}$$

una de las funciones de barrera que se utiliza es:

$$B(x) = -\frac{1}{g(x)}$$

que da lugar al siguiente problema sin restricciones:

$$\underset{x \in X}{Max} \quad F(x) - \mu B(x)$$

donde ahora μ es un número positivo arbitrariamente pequeño. Se observa que la función de barrera toma un valor positivo en puntos interiores y tiende a infinito a medida que el punto se acerca a la frontera del conjunto de oportunidades.

Los métodos de barrera empiezan con un punto interior del conjunto de oportunidades[30] y un escalar $\mu > 0$ arbitrario. Luego se soluciona el problema asociado sin restricciones mediante alguno de los métodos ya mencionados. Si tras resolverlo se obtiene un punto con un valor $\mu B(x)$ suficientemente pequeño se acepta como solución óptima, si no se disminuye el valor del escalar (manteniéndolo positivo) y se repiten las operaciones.

Un segundo método más eficiente y directo es el de **Programación Cuadrática Secuencial (SQP)**, mencionado por Wilson en 1963. La adaptación de este método a problemas de C.O. discreto es una de las líneas que parece más fructífera en la construcción de algoritmos específicos de C.O. discreto (por ejemplo, en Pantoja y Mayne, 1991; y en Fisher y Jennings, 1992).

La base del método es, en cada etapa, aproximar las restricciones por funciones lineales (primer término del desarrollo de Taylor) alrededor del punto resultante de la etapa anterior y la función objetivo por una función cuadrática (los dos primeros términos del desarrollo de Taylor pero con el hessiano de la función lagrangiana) también alrededor del punto resultante de la etapa anterior. Con ello se reduce el problema no lineal original a subproblemas cuadrático-lineales cuyas técnicas de resolución son generalmente exactas. Tras cada etapa se consigue una nueva solución que sirve de base para construir otro subproblema cuadrático-lineal, etc..

[30]La obtención de un punto interior de partida no suele ser una cuestión trivial, de ahí que existan algoritmos para conseguir un punto interior empezando desde cualquier otro punto.

Tomando como base el método SQP existen distintas maneras de modificarlo para aumentar su eficiencia o asegurar la convergencia de la solución. Entre ellas se puede citar el uso de aproximaciones al hessiano de la lagrangiana (por ejemplo con la fórmula BFGS) o utilizar funciones de penalización para realizar las aproximaciones cuadráticas.

En tercer lugar, se pueden citar los **métodos de direcciones factibles**. Estos métodos, iniciados por Zoutendijk en 1960, actúan sobre un punto factible desplazándolo a otro donde la función objetivo es mejor, para ello la dirección de desplazamiento debe tener la doble cualidad de ser factible y de mejora. Tras determinar la dirección apropiada, la segunda parte del algoritmo debe calcular el tamaño de ese desplazamiento, lo cual es un problema del tipo *step size* ya comentado.

Si se parte de nuevo del problema **[II]**, las direcciones de mejora en un punto factible, supuesta la diferenciabilidad de las funciones, son las que dan lugar a una derivada direccional positiva en ese punto y por tanto son direcciones $d \epsilon \mathbb{R}^n$ que cumplen:

$$\nabla F(x)d>0$$

Además, para que sean direcciones factibles, deben admitir un desplazamiento (aunque sea muy pequeño) que lleve a otro punto factible, es decir, se debe satisfacer la siguiente condición:

$$\exists \delta>0 \ / \ \ x+\lambda d \epsilon S \quad \forall \lambda \epsilon]0,\delta[$$

donde S es el conjunto de puntos factibles.

El problema de las direcciones factibles sólo se plantea en puntos de la

frontera del conjunto S dado que en puntos interiores todas las direcciones son factibles. Por tanto, para obtener estas direcciones sólo se tiene en cuenta las restricciones de desigualdad saturadas en el punto de partida (g_A) y las restricciones de igualdad. Suponiendo diferenciabilidad de estas funciones se tendrá que cumplir, en notación vectorial, la desigualdad e igualdad siguientes:

$$g_A(x+\lambda d)=g_A(x)+\lambda Jg_a(x)d+R_1(x,\lambda d)\leq 0$$
$$h(x+\lambda d)=h(x)+\lambda Jh(x)d+R_2(x,\lambda d)=0$$

donde R_1 y R_2 son los restos de Taylor (tienden a cero a medida que λ tiende a cero). Ahora, dado que x es factible y λ es un valor positivo suficientemente pequeño, la desigualdad se cumplirá si[31]:

$$Jg_A(x)d<0$$

mientras que la igualdad no se conseguirá para ninguna dirección no nula. Sin embargo, la condición:

$$Jh(x)d=0$$

que por sí sola llevaría a un punto infactible (a no ser que la función fuera lineal), se combina con un movimiento corrector hasta llevar a un punto factible.

En resumen, para el problema **[II]**, una dirección factible de mejora debe cumplir simultáneamente:

[31]Si la función es lineal la desigualdad que sigue no tiene porqué ser estricta porque el resto de Taylor se igualaría a cero.

$$\nabla F(x)d>0$$

$$Jg_A(x)d<0$$

$$Jh(x)d=0$$

y para determinar cuál de las direcciones que cumplen estas condiciones se debe escoger, se puede seguir el método de Zoutendijk que elige como dirección apropiada aquélla que es solución del siguiente programa lineal:

$$Max \quad z$$

$$s.a:$$

$$\nabla F(x)d-z\geq 0$$

$$Jg_A(x)d+Z\leq 0$$

$$Jh(x)d=0$$

$$-1\leq d_j\leq 1$$

donde Z es un vector de dimensión apropiada con componentes iguales a z y donde se han añadido cotas a las componentes de la dirección para que la solución sea acotada. Además, si es necesario, se debe realizar el movimiento corrector correspondiente antes comentado.

Tras obtener la dirección factible de mejora hay que resolver el siguiente problema de *line search*:

$$\begin{array}{l} Max \quad F(x+\lambda d) \\ 0\leq\lambda\leq\lambda_{max} \end{array}$$

donde para valores $\lambda>\lambda_{max}$ el punto es infactible.

Existen otros métodos para encontrar la dirección factible de mejora. Uno de ellos es el del **gradiente reducido generalizado (GRG)**, propuesto por Abadie y Carpentier en 1969. Este tipo de métodos se han adaptado posteriormente para resolver problemas de control óptimo. Mejoras posteriores de este método son las que se utilizan con más asiduidad en los algoritmos con soporte informático (por ejemplo, el método GRG2 de Lasdon y Waren de 1978).

El **método GRG2** distingue dos tipos de restricciones: las cotas y el resto de restricciones. Estas últimas siempre se pueden pasar a igualdades (si no lo son) añadiendo variables de holgura positivas. En un punto factible se puede hablar de tres tipos de variables (incluyendo las de holgura):

- variables básicas: aparecen cuando hay restricciones activas en ese punto y son las que se despejan en función de las otras variables y se sustituyen en la función objetivo.

- variables superbásicas: son no básicas y aquéllas donde no actúan las cotas. Pueden variar en cualquier dirección hasta que alcancen su cota superior o inferior.

- resto de variables no básicas: en ellas actúan las cotas. Se evalúa, a través del gradiente, si es conveniente que abandone su cota o no. Si no es conveniente su valor queda fijado y si es conveniente su valor puede variar hasta la otra cota.

Este método resuelve a continuación un problema reducido, es decir, un problema con menos variables: las superbásicas y las no básicas que pueden abandonar la cota. Este problema, sin restricciones, se resuelve a través de alguno de los métodos de *multidimensional search*, en concreto se realiza la siguiente elección:

- en la primera iteración utiliza como dirección de búsqueda la del gradiente.

- en las siguientes iteraciones utiliza métodos de cuasi-Newton si el problema

no es de grandes dimensiones o métodos del gradiente conjugado si es de grandes dimensiones. De esta manera aprovecha al máximo la eficiencia de cada uno de estos métodos.

Tras resolver el problema reducido se encuentra el valor de las variables básicas utilizando la ecuación correspondiente. Con el nuevo punto obtenido, que debe cumplir todas las ecuaciones, se repite el proceso. El criterio de parada, en caso de convergencia, aparece cuando la dirección de búsqueda está suficientemente próxima al vector nulo.

Junto con éste método de búsqueda de la solución óptima, hace falta un algoritmo de búsqueda de un punto factible inicial. Este problema se resuelve convirtiéndolo en uno de programación matemática donde, partiendo de un punto cualquiera no factible, se construye la función objetivo como suma de infactibilidades (utilizando las restricciones que no se satisfacen pasándolas todas a un miembro de forma que el resultado deba ser negativo) y como restricciones las que sí se cumplen. De esta manera minimizando la suma de infactibilidades se llega a un punto factible (si existe) con el que puede empezar el algoritmo principal.

IV.3.- PROBLEMAS LINEALES Y LINEALES-CUADRÁTICOS

IV.3.1.- Problemas lineales

Los problemas lineales se refieren a aquellos casos en los que la resolución del Principio del Máximo es un problema lineal. En otras palabras, la función hamiltoniana es lineal respecto a la variable de control y la región de control viene también definida a través de restricciones lineales en la variable de control. El problema de maximización de la hamiltoniana respecto a los controles factibles se convierte en un programa lineal. El algoritmo más utilizado en estos problemas es el del simplex que se debe a Dantzig y data de 1951.

El hecho de tener un programa lineal significa que el óptimo, si existe, se encuentra analizando sólo los puntos extremos del conjunto de oportunidades y que será o bien un punto extremo (en caso de solución única o de vértice) o una combinación lineal convexa de dos o más puntos extremos (en caso de solución múltiple). La solución dependerá, entre otras cosas, de los coeficientes de los controles en la función hamiltoniana y dado que, en general, dependen del tiempo, la solución del problema puede ir pasando de un punto extremo a otro de la región de control a medida que transcurren los periodos.

Un caso particular interesante es aquél en que la región de control se reduce a intervalos de acotación para cada variable de control en cada periodo. Si la función es lineal respecto a una variable acotada, el máximo (y el mínimo) de la función se encuentra en un extremo del intervalo de acotación según el signo del coeficiente de la variable de control en la función hamiltoniana. A medida que dicho coeficiente cambia de signo, la solución cambia de un extremo a otro del intervalo. Este tipo de solución se llama *bang-bang*.

Aunque el enunciado del problema lineal no es único, para observar el tipo de solución *bang-bang*, se parte del siguiente (Problema **[IV]**):

$$Max \quad \sum_{t=0}^{T-1} a_t x_t + b_t u_t$$

$$s.a:$$

$$x_{t+1} - x_t = A_t x_t + B_t u_t \qquad t=0,...,T-1 \qquad [IV]$$

$$m_t \leq u_t \leq M_t \qquad t=0,...,T-1$$

$$x_0 \ dado$$

donde a_t es un vector de dimensión n, b_t es de dimensión m, A_t son matrices nxn y B_t son nxm. A su vez, los valores mínimos y máximos de las variables de control en cada periodo vienen dados, respectivamente, por los vectores m_t y M_t. El enunciado con restricciones de estado, no necesariamente lineales, se puede reducir a éste modificando adecuadamente las cotas de los controles para que dichas cotas, junto con el sistema dinámico, sea compatible con las restricciones de estado. En ese caso las cotas son función de las variables de estado.

Este enunciado lleva a una hamiltoniana lineal respecto a la variable de control que, a su vez, está acotada. El problema de maximizar la hamiltoniana se reduce al siguiente:

$$Max \quad H_t = (a_t + \mu_{t+1} A_t) x_t + (b_t + \mu_{t+1} B_t) u_t$$

$$s.a.: \quad m_t^j \leq u_t^j \leq M_t^j \qquad j=1,...,m$$

El problema es separable dado que cada restricción sólo depende de una variable de control, es decir, son independientes entre sí. La solución depende del signo de los elementos de las sucesiones $b_t^j + \mu_{t+1} B_t^j = C_t^j$ obteniendo la típica expresión de los controles *bang-bang*:

$$u_t^j = \left\{ \begin{array}{l} m_t^j \ \ si \ \ C_t^j<0 \\ M_t^j \ \ si \ \ C_t^j>0 \end{array} \right\}$$

En consecuencia, la sucesión solución para una variable de control está formada por elementos que toman un valor que va pasando de un extremo a otro del intervalo correspondiente según el valor del elemento asociado de la sucesión C_t^j llamada también sucesión de cambio o *switching*.

Nótese que si en algún periodo y para algún control la función C_t^j es igual a cero, la solución para dicho control no queda determinada siendo indiferente la elección de un valor u otro dentro de su intervalo. Se trata, por tanto, de solución múltiple y es raro que se dé este caso debido a que el cambio de signo de C_t^j es poco probable que ocurra en valores de t naturales[32].

Si la linealidad afecta también a la variable de estado, el problema de C.O. discreto en su totalidad se reduce a uno de programación lineal y se puede resolver, de nuevo a través de métodos específicos. En consecuencia, se puede aprovechar el mayor desarrollo en soporte informático de los métodos numéricos en programación lineal, tanto del método símplex como de otros más recientes.

Además del método símplex que fue el primero que se desarrolló, se pueden citar algunas modificaciones posteriores que lo dotan de mayor eficiencia. El método del símplex revisado, por ejemplo, necesita ir actualizando en cada iteración una matriz de menor dimensión, ahorrando así tiempo de cálculo y necesidad de memoria. Para problemas con muchas restricciones de desigualdad es preferible seguir un método de conjunto activo que trata el problema, en una primera etapa,

[32] En control óptimo en tiempo continuo, el cambio de signo da lugar a saltos en la variable de control (función continua a tramos). La solución recibe el nombre de solución singular si la función *switching* se anula a lo largo de un intervalo de tiempo. En esos casos la resolución a través del Principio del Máximo continuo falla y se hace necesario utilizar principios de óptimo de segundo orden (uno de los más recientes se puede ver en Gift (1993)).

con sólo las restricciones activas desplazando el punto de un extremo a otro de la arista y observando, en una segunda etapa, si alguna restricción inactiva no se cumple, en cuyo caso se realiza el movimiento corrector correspondiente. Si el problema es de gran dimensión y la matriz de coeficientes tiene gran cantidad de ceros (como ocurre con el problema de C.O. discreto) existen métodos especiales para ganar en eficiencia entre los que destaca el de descomposición del símplex.

Otra ventaja de la linealidad es que asegura que se cumplen todos los supuestos de los teoremas que validan el principio del máximo, tanto a través de las condiciones de Kuhn y Tucker como a través de los argumentos geométricos, lo cual puede aprovecharse para desarrollar métodos numéricos específicos en C.O. discreto lineal que utilicen el principio del máximo.

Por último, hay que hacer notar que el tipo de solución *bang-bang* no es exclusivo de los problemas lineales. Como se ha descrito antes para el caso lineal, si la región de control viene dada a través de intervalos de acotación para cada variable de control en cada periodo, el problema de maximizar la hamiltoniana respecto a los controles es un problema separable. Sin embargo, el problema puede seguir siendo separable si la hamiltoniana lo es y para eso no es necesario que sea lineal.

En realidad, para que aparezca un tipo de solución *bang-bang* es suficiente con que la hamiltoniana sea creciente o decreciente para cada control a lo largo del intervalo de acotación. Efectivamente, el máximo de una función creciente (decreciente) que depende de una variable acotada se encuentra en el extremo superior (inferior) del intervalo. El hecho de que sea creciente o decreciente dependerá, en general, del periodo de tiempo considerado. En consecuencia, la sucesión óptima para cada control puede ir pasando de un extremo a otro del intervalo a medida que transcurren los periodos dando lugar a la solución de tipo *bang-bang*.

IV.3.2.- Problemas lineales-cuadráticos

Los problemas lineales-cuadráticos son aquéllos donde la función objetivo es cuadrática y el sistema dinámico y las restricciones son lineales, respecto a todas las variables del problema. La ventaja de estos problemas es que al transformarse en un problema de P.N.L. se tiene un problema de programación cuadrática y, en consecuencia, se dispone de algoritmos eficientes de resolución. Los métodos numéricos que se han citado en el epígrafe anterior funcionan bien o incluso se transforman en métodos exactos ante problemas cuadrático-lineales, debido a la simplicidad de las funciones y a lo bien determinado que está el óptimo en una función cuadrática.

Los **problemas de programación cuadrática en P.N.L.** necesitan pocos supuestos adicionales para asegurar la existencia de óptimo, así como para llegar a condiciones necesarias y suficientes. Aunque existen distintas formas de enunciar un problema cuadrático, utilizaremos el siguiente enunciado:

$$\begin{aligned} &Min \quad z^T Rz + dz \\ &s.a: \\ &\quad x_{t+1} - x_t = A_t x_t + B_t u_t \qquad t=0,\ldots,T-1 \qquad [V] \\ &\quad Gz \geq 0 \\ &\quad x_0 \ dado \end{aligned}$$

donde $z=(X, U)$ y R, d, A_t, B_t y G son vectores y matrices de dimensión adecuada para la conformidad de las operaciones enunciadas.

Un primer resultado evidente referido a la existencia de solución es que si la matriz R es semidefinida positiva, $d=0$ y existe una solución factible, entonces

existe una solución óptima[33]. El resultado es claro ya que al ser semidefinida positiva está acotada inferiormente y por tanto existe mínimo siempre que el conjunto de oportunidades sea no vacío. En segundo lugar, se tiene que las condiciones de Kuhn y Tucker son necesarias al cumplirse la cualificación de restricciones (son lineales) y son suficientes si R es semidefinida positiva.

Por tanto, lo único que hay que comprobar en estos problemas es el signo de la forma cuadrática representada por la matriz simétrica R, o lo que es lo mismo, la concavidad (para máximo) o la convexidad (para mínimo) de la función objetivo.

Un ejemplo de este tipo de problemas que conviene destacar, ya que se va a utilizarlo luego en las aplicaciones del capítulo VI, es el caso de minimizar las desviaciones cuadráticas de ciertas variables respecto de unos valores deseados en cada periodo. Por ejemplo, si estas variables sólo son las de control se tiene una función objetivo del tipo:

$$MIN \quad \sum_{t=0}^{T-1} \sum_{j=1}^{m} (u_t^j - \bar{u}_t^j)^2$$

donde u_t^j es el valor deseado de cada variable. En este tipo de problemas se redefinen las variables de control como la diferencia entre el control óptimo y su valor deseado. Así, utilizando la notación del Problema [V], se tiene que $d=0$ y la matriz R es semidefinida positiva, dado que es igual a:

$$R = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

donde cada bloque es de dimensión apropiada, según sea el número de variables de

[33]En Canon, Cullum y Polack (1970) se puede ver otros enunciados de problemas de programación cuadrática, así como teoremas de existencia más generales.

estado y de control. La aplicación de los algoritmos generales de P.N.L. antes reseñados o de algoritmos específicos para problemas cuadráticos resulta, en este caso, muy eficiente dada la sencillez de la matriz R.

Una vez analizado el problema, se aplica algún algoritmo de programación cuadrática[34], aunque el gran número de variables y restricciones es un inconveniente serio. Para reducir el número de variables y restricciones se puede resolver primero el sistema dinámico y obtener las variables de estado en función de las de control y del estado inicial, pero esto exige realizar cálculos adicionales. Otra posibilidad es aprovechar la estructura del problema, ya que al ser de C.O. discreto, las matrices deben tener gran cantidad de ceros correspondientes, en cada fila, a los coeficientes de las variables en los periodos de tiempo no considerados en esa restricción. Por otra parte, el uso del P.M. discreto es válido al tratarse de un sistema dinámico lineal, lo cual se puede aprovechar para desarrollar métodos numéricos específicos.

Uno de los algoritmos iniciales de programación cuadrática es una adaptación del método símplex de programación lineal y se debe a Wolfe y Dantzig. Para aplicarse, las restricciones deben adoptar la misma estructura que en el método símplex (restricciones de desigualdad y condiciones de no negatividad). Este método necesita actualizar en cada etapa una matriz inversa, necesidad que ha sido superada en algoritmos posteriores.

Los algoritmos más generales de programación cuadrática parten del problema con restricciones de igualdad. La linealidad de las restricciones permite entonces obtener unas variables en función de otras siguiendo algún método de eliminación. Se pasa así a un problema reducido sin restricciones donde el sistema que se obtiene tras aplicar las condiciones de primer orden es lineal. Si la matriz R es semidefinida positiva la solución del sistema proporciona un mínimo global y si

[34]Los algoritmos de programación cuadrática, además de utilizarse para resolver este tipo de problemas, sirven de base, como ya se ha visto, para desarrollar algoritmos en problemas más generales (no cuadráticos) de P.N.L.; por ejemplo, en el método SQP.

es definida positiva es único.

Cuando el problema tiene restricciones de desigualdad se utiliza una estrategia de conjunto activo que resuelve, en cada etapa, un problema con restricciones de igualdad (las que lo son, más las de desigualdad activas) para pasar a un punto mejor que también las cumpla. Si el nuevo punto no cumple alguna de las inactivas se resuelve un problema de *step size* para pasar a un punto factible donde el conjunto de restricciones activas habrá cambiado.

IV.4.- MÉTODOS NUMÉRICOS ESPECÍFICOS EN CONTROL ÓPTIMO DISCRETO

Estos métodos utilizan técnicas de la teoría del control óptimo discreto y, en concreto, el Principio del Máximo. Son métodos que están en pleno desarrollo actualmente y, lógicamente, deben ser más eficientes que los anteriores al aprovechar la estructura del enunciado del problema.

Los métodos que tratan de resolver numéricamente las condiciones del Principio del Máximo discreto junto con el sistema hamiltoniano dinámico presentan el inconveniente de que las variables son interdependientes lo cual complica bastante el uso de métodos de *shooting* para el valor inicial de la variable de coestado[35].

Los métodos más tradicionales utilizan algoritmos de programación dinámica diferencial que requieren gran capacidad de almacenamiento y tienen ciertas dificultades para tratar restricciones de carácter general. Como toda técnica de programación dinámica, se basa en cálculos iterativos de las variables y de la función objetivo.

La alternativa que se impone es la combinación de elementos del Principio del Máximo con métodos numéricos de P.N.L. antes comentados. Esta opción es la seguida por gran parte de los investigadores a la vista de los trabajos publicados. Así, Dunn y Bertsekas (1989) aplican el método de Newton por etapas al problema de C.O. discreto sin restricciones.

Pantoja y Mayne (1991) amplían la estrategia de los anteriores autores para incluir restricciones, aunque sólo de control. Transforman el problema de control en uno de programación matemática y, aprovechando su estructura, utilizan un algoritmo SQP para resolver en cada etapa, no el problema general de gran

[35]En este contexto, los métodos de *shooting* consisten en probar un valor inicial para la variable de coestado y resolver el sistema hamiltoniano dinámico hacia delante en el tiempo. Si la solución obtenida no cumple las condiciones terminales se revisa de forma apropiada el valor inicial elegido.

dimensión, sino varios subproblemas de menor dimensión. La aproximación cuadrática de la función objetivo se realiza utilizando el hamiltoniano y el cálculo de los multiplicadores mediante el sistema hamiltoniano dinámico que aparece en el principio del máximo.

Fisher y Jennings (1992) también convierten el problema original en uno de programación matemática en el que sólo se consideran como variables de decisión las de control, calculando las variables de estado a partir del sistema dinámico. Además, el resto de restricciones se transforman en una forma canónica estándar. El algoritmo permite hacer también optimización de parámetros. Una vez transformado el problema, se resuelve a través del algoritmo SQP utilizando la función hamiltoniana del Principio del Máximo para el cálculo de los gradientes de las funciones respecto al vector de parámetros.

Por último, resumimos las principales características del *software* susceptible de utilizarse en la resolución de problemas de C.O. discreto. En cuanto al *software* específico de C.O. discreto se puede citar el programa **DMISER3**. Es la versión en tiempo discreto de MISER3 y permite realizar también optimización de parámetros. La primera parte consta de un algoritmo que transforma el problema en uno de P.N.L., destacando en ese proceso la forma en la que los distintos tipos de restricciones se pasan a una forma canónica. Una vez pasado el problema a uno de P.N.L. se parametrizan los controles, realizando una partición del horizonte temporal y asignando a cada variable de control un valor constante (parámetro) para cada intervalo resultado de la partición. El problema que queda es aproximado al inicial y también es de P.N.L. donde las variables son los parámetros. Sobre este enunciado se aplica *software* estándar de P.N.L. y en concreto la subrutina NLPQL de Schittkowski de 1985 que es un método SQP.

En cuanto a *software* general de P.N.L. destacamos **GAMS** y **LINGO**. Pese a ser programas no específicos de C.O. discreto tienen un lenguaje de modelización y unos resolutores apropiados para tratar dicho problema. En efecto, el lenguaje de modelización permite introducir problemas estructurados de gran dimensión mediante

el uso de subíndices sobre las variables, creando de esta manera vectores y matrices. Así, tomando como ejemplo el programa **LINGO**, tras crear el conjunto de variables con subíndices la función *@SUM* permite introducir la función objetivo tecleando simplemente la función intermedia y la función *@FOR* hace lo mismo con el sistema dinámico o las restricciones. En cuanto a los métodos numéricos de resolución, **GAMS** permite escoger el resolutor de una librería mientras que **LINGO** evalúa las características del problema y escoge el método más adecuado (normalmente del tipo GRG2 en modelos no lineales).

En las aplicaciones numéricas que se desarrollan en los próximos capítulos, cuando no es posible la resolución analítica, se utiliza el programa informático **LINGO** en su versión para problemas de grandes dimensiones (versión 2.1 de 1995). La accesibilidad de este programa, el desarrollo que han alcanzado los métodos numéricos que utiliza y la disponibilidad de una versión reciente lo han hecho preferible a los otros *software* citados. Además, dispone de un editor de pantalla completa propio (en **GAMS** hay que utilizar el del sistema operativo), es posible incorporar datos de variables exógenas en el modelo desde hojas de cálculo (función *@IMPORT*) o trasladar la solución a una hoja de cálculo (función *@EXPORT*), realizar análisis de post-optimización (análisis *What if*) o crear ficheros por lotes que automaticen varias operaciones.

CAPÍTULO V:

APLICACIÓN A LA ELECCIÓN ÓPTIMA DE INSTRUMENTOS DE AHORRO DE COMPLEMENTO A LA JUBILACIÓN

V.1.- INTRODUCCIÓN

V.1.1.- Marco en el que se desarrolla la aplicación

En este capítulo se desarrolla una aplicación financiera que tiene como principal resultado la obtención de unas relaciones que permiten valorar desde el punto de vista financiero-fiscal y para un inversor individual distintos instrumentos financieros de ahorro a largo plazo (en el modelo este plazo lo marca la edad de jubilación) y especialmente los planes de pensiones. Dichos instrumentos se comparan con los de ahorro convencional (sin ventajas fiscales), dando lugar a estrategias dinámicas óptimas de ahorro y a la formulación de la rentabilidad financiero-fiscal para cada uno de ellos.

El modelo que se plantea atiende especialmente a los aspectos financieros y fiscales de los instrumentos de ahorro, sin profundizar en otros aspectos tratados en la literatura como los jurídicos, sindicales, y sobre todo los macroeconómicos. Entre estos últimos destaca el estudio, a nivel teórico y empírico, de los efectos económicos de los planes de pensiones tales como la influencia sobre el binomio consumo-ahorro y sobre las decisiones de jubilación anticipada y merecen que se les dedique unos breves comentarios en esta introducción.

La cuestión de si los planes de pensiones hacen aumentar el ahorro de una economía o no es una de las más discutidas. Desde el punto de vista de la teoría económica (tomando como punto de partida el **modelo de ciclo vital** del ahorro[36]) se concluye que el efecto sobre el ahorro privado es indeterminado ya que aparecen dos efectos de signo contrario asociados a la mayor rentabilidad neta de los planes de

[36]Ver en Ando y Modigliani (1963).

pensiones[37]: el **efecto renta**, derivado de unos mayores ingresos netos a lo largo de la vida, que hace aumentar tanto el consumo actual como el futuro (menor ahorro) y el **efecto sustitución**, como consecuencia de un mayor "precio" relativo del consumo actual, que lo disminuye a cambio de consumo futuro (mayor ahorro). Aún en el caso de que el efecto total sobre el ahorro privado sea positivo queda por ver el comportamiento del ahorro público. Si los menores ingresos fiscales derivados de los planes de pensiones se compensan con menor gasto público, el ahorro público puede mantenerse, dando lugar a un efecto global positivo sobre el ahorro. Pero este comportamiento austero del sector público es dudoso. Si, por otra parte, la compensación a las ventajas fiscales anteriores viene a través del aumento de impuestos, se produciría una caída en la renta disponible de los individuos a lo largo del ciclo de vida con lo que resultaría más difícil que la combinación del efecto renta y sustitución fuera favorable al ahorro.

Otros modelos teóricos basados en el modelo de ciclo vital consideran dos tipos de ahorro: uno deducible (aportaciones a planes de pensiones) y otro que no lo es. Se puede citar a Daly (1981) y Ko (1988) que amplían un modelo previo de Atkinson (1971) formulado para estudiar los efectos de la imposición sobre el capital. El modelo básico se puede formular como uno de control óptimo bien en tiempo discreto o continuo. Se toma como variables de control el consumo y la aportación al plan de pensiones en cada periodo y como variables de estado el ahorro acumulado en el fondo de pensiones y en el resto de activos. La función objetivo a maximizar es la suma descontada de la utilidad del consumo en cada periodo y de la utilidad descontada de la riqueza en el último periodo. El sistema dinámico indica cómo se produce la acumulación del fondo de pensiones y del resto del ahorro teniendo en cuenta las características fiscales del plan de pensiones. Bajo los supuestos habituales, la estrategia

[37]La mayor rentabilidad neta de los planes de pensiones deriva de sus ventajas fiscales. La posibilidad de aplazar el pago del impuesto supone, en la mayoría de los casos, que una parte de los ingresos tributen a un tipo marginal menor con lo que los impuestos que se pagan a lo largo del ciclo vital son menores.

óptima de un individuo es realizar aportaciones y recuperaciones al plan de pensiones de forma que su tipo impositivo marginal sea creciente a lo largo de su ciclo de vida[38]. Al mismo tiempo, se produce un mayor ahorro a lo largo del ciclo vital con planes de pensiones que sin ellos debido a que las ventajas fiscales permiten destinar fondos adicionales tanto al ahorro como al consumo. Estos modelos anticipan un mayor ahorro privado pero no explican la evolución del ahorro agregado de la economía. De nuevo el menor ahorro público que se deriva de la menor recaudación fiscal que se produce con los planes de pensiones puede llevar a un menor ahorro agregado.

Los estudios empíricos más conocidos hacen referencia al caso americano donde los planes de pensiones (IRAs), existentes desde 1975 para los trabajadores que no estaban cubiertos por otro tipo de planes de pensiones privados, se generalizaron a partir del año 1981 al resto de trabajadores. Estos estudios, basados en técnicas econométricas, tratan de determinar si las aportaciones a los IRAs suponen creación de ahorro neto o no. En el primer caso, las aportaciones provendrían de un menor consumo y, en el segundo caso, tendrían su origen en un trasvase de otras formas de ahorro o en un simple ahorro impositivo. Los distintos estudios econométricos realizados con datos de EE.UU. no parecen despejar dudas sobre este aspecto e incluso llegan a conclusiones contrarias.

Hubbard (1984) estima un modelo de ahorro de ciclo vital y llega a conclusiones de tipo individual, no agregado. Los resultados muestran que las aportaciones a los IRAs y a otros instrumentos de ahorro con ventajas fiscales aumentan el ahorro individual y esto depende directamente de los tipos impositivos marginales. Venti y Wise iniciaron estudios de este tipo en 1986 y sus resultados más conocidos se recogen

[38]La estrategia óptima se vería modificada si se incluyen restricciones habituales como la limitación a la aportación al plan de pensiones, la penalización o imposibilidad de realizar recuperaciones antes de los 65 años, etc.. Un tipo marginal creciente es necesario para compensar el diferente trato fiscal de los rendimientos de cada periodo del plan de pensiones y del activo sin ventajas fiscales; todo ello para que a un individuo le sea indiferente aportar en un periodo al plan de pensiones o hacerlo en el periodo siguiente, invirtiendo durante ese periodo en el activo no deducible (Daly, 1981).

en Venti y Wise (1990) donde destacan que las dos terceras partes de las aportaciones a los IRAs provienen de un menor consumo y por tanto sí que son efectivos para aumentar la tasa de ahorro de una economía. Otros estudios reflejan un papel más neutral de los planes de pensiones e incluso los hay que llegan a la conclusión de que su efecto sobre el ahorro es negativo. Entre estos se puede citar a Burman, Cordes y Ozanne (1990). En Gravelle (1991) se realiza una revisión histórica de esta controversia. En el caso español no existen estudios empíricos sobre la materia debido a los pocos años transcurridos desde la implantación de los planes de pensiones y a que no han acumulado un volumen de capital significativo.

Otro de los efectos económicos estudiados hace referencia a las decisiones de jubilación anticipadas y a las repercusiones sobre la oferta de trabajo. Estos estudios son abundantes en lo que se refiere a las decisiones inducidas por las pensiones de jubilación de un sistema de Seguridad Social público, pero se puede extender a planes de pensiones de prestación definida[39] ya que este tipo de planes incluyen unas prestaciones futuras en función de la edad, último salario, años de servicio, etc., similares a las variables que se tienen en cuenta para determinar la pensión de jubilación en un sistema de Seguridad Social. Entre estos estudios se puede citar el de Stock y Wise (1990) quienes, tras realizar una revisión de las formas de analizar la decisión de jubilarse, presentan un modelo nuevo que se basa en una función que valora los beneficios de retirarse en un periodo dado. Sus conclusiones, al igual que en otros estudios, apoyan la hipótesis de que los planes de pensiones de prestación definida, al igual que las pensiones de jubilación públicas distorsionan la oferta de trabajo en función de las condiciones requeridas para disfrutar de las prestaciones. Un completo panorama de este tipo de efectos, tanto derivados de las pensiones públicas como de los planes privados se puede ver en Fabel (1994).

[39]En España los planes de pensiones individuales no pueden ser de prestación definida. Sí en cambio los de empresa.

V.1.2.- Objeto de la aplicación

El modelo que se desarrolla en este capítulo estudia los planes de pensiones individuales (PPIs) en su aspecto financiero-fiscal y no se ocupa de los efectos económicos antes comentados. La novedad del modelo que aquí se desarrolla frente a otros trabajos que estudian este mismo aspecto de los planes de pensiones reside en la utilización de la teoría del Control Óptimo en tiempo discreto como método de optimización a partir del cual se obtienen los resultados. Así, por ejemplo, en Vidal (1994a) se utiliza el método analítico del factor ganancia para comparar los planes de pensiones con el ahorro convencional, llegando básicamente a los mismos resultados que con nuestro método de optimización. Lo novedoso de esta aplicación no está, por tanto, en los resultados que se obtienen sino en el instrumento matemático utilizado para ello. Dado que los resultados son los ya conocidos, lo que se consigue es, en todo caso, darles mayor consistencia pero lo más destacable es, como se ha dicho, observar la aplicabilidad de la teoría del Control Óptimo en tiempo discreto.

El problema que se plantea es el de determinar si es conveniente colocar en un plan de pensiones una unidad monetaria de ahorro adicional ganada en un periodo dado o, por el contrario, es mejor colocarla en un instrumento de ahorro convencional (sin ventajas fiscales), todo ello con fines de complementar los ingresos del periodo de jubilación y atendiendo a un criterio financiero-fiscal. La solución, como veremos, depende básicamente del perfil temporal de tipos impositivos marginales que se esperen para el individuo del problema, que a su vez está asociado con su perfil temporal de ingresos.

Obsérvese que en el modelo no se abordan cuestiones como el volumen de inversión óptimo (la alternativa ahorro-consumo se suele tratar a través de modelos de ciclo vital) o el plazo óptimo de la inversión (que viene dado por la edad de jubilación). Tampoco se pretende evaluar el ahorro de alguno de estos dos tipos:

- El ahorro que no persigue fines de complemento a la jubilación.

- El ahorro que no atiende al criterio financiero-fiscal.

Por tanto, el modelo parte de que se dispone de una unidad monetaria adicional y se desea ahorrar de cara a la jubilación de forma que se consiga la mejor rentabilidad financiero-fiscal.

Los instrumentos financieros seleccionados para este tipo de ahorro son, en principio, dos: los planes de pensiones individuales[40] (PPI) y las formas convencionales de ahorro en el sentido que no tienen ventajas fiscales (AC). Más tarde se extenderá el análisis a los fondos de inversión (FI), muy desarrollados en nuestro país. Las componentes diferenciadoras de los instrumentos financieros son la rentabilidad y la fiscalidad. En cuanto al riesgo se supondrá que no hay diferencias relevantes y en cuanto a la liquidez la única restricción será que ambos instrumentos se puedan adquirir en cualquier periodo y se puedan recuperar al final del horizonte de planificación, es decir, en el periodo de jubilación.

Los planes de pensiones se someten al marco legal establecido principalmente por la Ley 46/1984 reguladora de las Instituciones de Inversión Colectiva, la Ley 8/1987 de Regulación de los Planes y Fondos de Pensiones y el Real Decreto 1307/1988 por el que se aprueba el Reglamento de Planes y Fondos de Pensiones. De entre las características de los planes de pensiones sólo vamos a extraer las relevantes fiscalmente de cara a la modelización matemática, que son las tres siguientes:

1. Las inversiones o aportaciones al plan desgravan de la base

[40]Se hace referencia a los Planes de Pensiones del sistema individual debido a la naturaleza del problema a resolver.

imponible[41].

2. Los intereses del fondo de pensiones se acumulan y no están sujetos a gravamen.

3. La recuperación del ahorro a partir de los 65 años está sujeta a impuestos[42].

El fondo de pensiones se compara con un activo sin ventajas fiscales, es decir, un activo donde las aportaciones y retiradas no tienen efectos fiscales y donde los intereses se perciben al final de cada periodo en que se divide el horizonte de planificación y están sujetos a gravamen.

[41]Aunque esto sólo es cierto teniendo en cuenta ciertos límites, en el modelo se supondrá que las características del individuo son tales que la desgravación por la unidad monetaria adicional que se aporta es total.

[42]Existe la posibilidad de recibir el fondo acumulado en forma de capital y/o renta periódica con tratamiento fiscal específico en cada caso. La elección de una u otra alternativa no es relevante en el planteamiento teórico del modelo.

V.2.- EL MODELO GENERAL

V.2.1.- Consideraciones generales y supuestos

El modelo que sirve de base para la obtención de resultados tiene puntos de conexión con el de Sethi y Thompson de 1970 aplicado a la gestión de saldos de tesorería[43]. El modelo se plantea adoptando los siguientes supuestos generales:

S.1. Tiempo discreto, modelo determinista y periodos de tiempo de un año. Esto significa que todas las variables se evalúan al final o al principio de cada año.

S.2. El horizonte temporal de la inversión es igual al número de años que faltan hasta la edad de 65, dado que se trata de un tipo de ahorro con fines de complemento de pensión.

S.3. Se plantea la inversión de una unidad monetaria ganada en cada periodo (y por tanto sujeta a impuestos). La colocación de esta unidad monetaria en un instrumento u otro no afecta al tipo impositivo de ese periodo (se considera exógeno puesto que depende de los ingresos habituales sujetos a impuestos[44] y de la función de impuestos, todo ello exógeno al modelo).

S.4. Los impuestos son progresivos. El tipo impositivo al que se grava

[43]Se puede consultar este modelo en Sethi y Thompson (1981) o en Biayna (1985). Los puntos de conexión se refieren sobre todo a la forma de la función objetivo (maximizar fondos al final del horizonte de planificación), región de control en forma de intervalo y existencia de dos tipos de instrumentos financieros (tesorería y títulos-valor en aquél caso y ahorro deducible y no deducible en el nuestro).

[44]Los ingresos habituales pueden estar disminuidos por las aportaciones al plan de pensiones que se realicen al margen de la unidad monetaria cuya inversión se está planteando en el modelo.

la unidad monetaria es, por tanto, el tipo marginal, que será distinto en cada periodo en función de la evolución temporal de los ingresos sujetos a impuestos.

S.5. La situación del inversor es tal que no supera los límites para deducir las aportaciones al PPI de la Base Imponible: si la unidad monetaria adicional se invierte en el PPI, queda totalmente exenta de impuesto.

S.6. No se permiten desinversiones de ninguno de los instrumentos financieros antes del último periodo ni saldos negativos, es decir, el fondo acumulado en cada activo siempre debe ser mayor o igual que cero.

S.7. No existen costes de transacción en las inversiones ni en las desinversiones finales.

S.8. El marco tributario es estable: no se producen cambios relevantes en el tratamiento fiscal de las aportaciones y las recuperaciones de fondos.

En el modelo se introducen las variables y los parámetros con la siguiente notación:

T: Horizonte de planificación. Debido a la forma de colocar las variables en el eje temporal, T es igual a 65 menos la edad del inversor más 1.

x_t: Importe acumulado en el PPI al principio del periodo t.

z_t: Importe acumulado en el ahorro convencional, AC, al principio del periodo t.

A_t': Aportación al PPI al final del periodo t (1-A_t' es la aportación al AC antes de impuestos).

i_t': Rentabilidad del PPI durante el periodo t.

i_t: Rentabilidad del AC durante el periodo t.

r_t: Tipo marginal del individuo en el periodo t (ver supuestos S.3. y S.4.). Se evalúa al final del periodo.

I_p: Impuesto que se paga al recuperar el ahorro acumulado en los instrumentos financieros utilizados.

Se considera que el instante 1 es la edad inicial del sujeto decisor. Así, si un sujeto quiere empezar a ahorrar al cumplir los 45 años, se tendrá un problema con 21 periodos contando el periodo 0 pero con 20 decisiones de ahorro dado que al cumplir los 65 años no hay decisión de ahorro. Se considera también que al final de cada periodo se realizan las aportaciones a los distintos instrumentos de ahorro tras lo cual se pagan los impuestos correspondientes a ese periodo y se calculan las variables de estado.

V.2.2.- Construcción del modelo

El modelo general tiene las componentes siguientes:

a) Las variables de estado y de control

Las variables de estado, x_t y z_t, son las cantidades acumuladas en los dos instrumentos de ahorro considerados: el PPI y el AC respectivamente. Se trata de un tipo de ahorro deducible en el momento de la aportación en el caso del PPI y de un ahorro no deducible en el caso del AC. La variable de control, A_t', es la aportación al PPI y es una variable acotada porque se plantea la mejor alternativa de ahorro para una unidad monetaria marginal. No aparece el consumo como variable de control ya que se ha considerado exógeno y, como consecuencia, el modelo se aparta de los convencionales de ciclo vital. Esto se debe a que el problema a resolver es distinto: no se trata de una decisión óptima ahorro-consumo sino una elección óptima de tipo de ahorro.

b) La función objetivo

El objetivo del inversor es llegar a los 65 años con el mayor volumen de ahorro posible en los instrumentos financieros. Además el importe acumulado se minora por el impuesto que se paga al recuperarlo. Este impuesto se considera al margen del impuesto normal del último periodo y su notación es I_p. La función objetivo que resulta es la siguiente:

$$Max \quad x_T+z_T-I_p$$

La recuperación del ahorro convencional no está sujeta a impuestos mientras que

la del plan de pensiones sí. Independientemente de si se recupera en forma de capital o de renta periódica el impuesto a pagar se puede aproximar al resultado de multiplicar la cuantía del PPI por un tipo impositivo (r_p)[45]. La función objetivo se transforma en:

$$Max \quad x_T(1-r_p)+z_T$$

c) El sistema dinámico y las condiciones de contorno

El sistema dinámico indica la evolución en el tiempo de las variables de estado. El importe que se paga en impuestos en cada periodo tendrá una parte exógena que no interviene en el modelo y una parte que sí que hay que tener en cuenta. Esta parte es la que corresponde a la unidad monetaria marginal ganada en cada periodo. Dadas las características fiscales de cada instrumento, el impuesto será el tipo marginal por la unidad monetaria ganada menos la aportación al PPI más los intereses que ha generado el ahorro convencional. Dados los supuestos del modelo se deduce el siguiente sistema dinámico de ecuaciones en diferencias con las correspondientes condiciones de contorno:

$$x_{t+1}-x_t = i_t'x_t+A_t' \qquad t=0,1,...,T-1$$

$$x_0 \text{ conocido, } x_T \text{ libre}$$

$$z_{t+1}-z_t = i_tz_t+1-A_t'-r_t(1-A_t'+i_tz_t) \qquad t=0,1,...,T-1$$

$$z_0 \text{ conocido, } z_T \text{ libre}$$

[45]La forma de calcular r_p sí que dependerá de cómo se produce la recuperación del PPI. Si se hace en forma de capital equivaldrá a una combinación lineal convexa del tipo marginal y del tipo medio según los años de permanencia en el fondo (renta irregular). Si se hace en forma de renta periódica será aproximadamente una media de los tipos marginales de los años en que se percibirá.

d) Las restricciones

Atendiendo a los supuestos del modelo, existen restricciones de no negatividad para todas las variables del modelo. También hay que tener en cuenta que al cumplir los 65 años no se realizan aportaciones y que las aportaciones que se plantean son de una unidad de ahorro adicional, es decir:

$$x_t \geq 0,\ z_t \geq 0 \qquad t=0,1,...,T$$

$$0 \leq A_t' \leq 1 \qquad t=0,1,...,T-2$$

$$A_{T-1}'=0$$

Obsérvese que el hecho de que la aportación esté entre cero y uno permite, en principio, que la inversión se reparta entre ambos instrumentos de ahorro; sin embargo, el enunciado final del problema resulta ser lineal con control acotado lo que llevará a una solución óptima en los extremos del intervalo.

La conjunción de los anteriores elementos da lugar al Problema **[I]** de C.O. discreto que se expone a continuación:

$$Max \quad x_T(1-r_p)+z_T$$

s.a:

$$x_{t+1}-x_t = i_t'x_t+A_t' \qquad t=0,1,...,T-1$$

$$z_{t+1}-z_t = (i_tz_t+1-A_t')\,(1-r_t) \qquad t=0,1,...,T-1 \qquad \text{[I]}$$

$$x_0,\ z_0 \quad \textit{conocidos}$$

$$x_t\geq 0,\ z_t\geq 0 \qquad t=0,1,...,T$$

$$0\leq A_t'\leq 1 \qquad t=0,1,...,T-2$$

$$A_{T-1}'=0$$

V.2.3.- Resolución a través del Principio del Máximo discreto

La aplicación del P.M. discreto al Problema **[I]** requiere previamente la definición de la función hamiltoniana. Este problema tiene una función objetivo que consta sólo de una función residual y su sistema dinámico es lineal. Por otra parte, se observa que las condiciones de no negatividad se cumplen en cualquier periodo sin más que inspeccionar el sistema dinámico. Como consecuencia de todo ello, la función hamiltoniana, H_t, es la siguiente:

$$H_t = \mu_{t+1}\,(i_t'x_t+A_t') + \lambda_{t+1}\,(i_tz_t+1-A_t')\,(1-r_t)$$

Se trata de una hamiltoniana lineal en la variable de control que, a su vez, es una variable restringida a tomar valores en un intervalo de acotación. Las condiciones de óptimo, descritas en el Capítulo II, se expresan a continuación aplicadas al Problema **[I]**:

a) El sistema hamiltoniano dinámico, compuesto por:

a.1) el sistema dinámico de ecuaciones en diferencias:

$$x_{t+1}-x_t = i_t'x_t+A_t' \qquad t=0,1,...,T-1$$

$$z_{t+1}-z_t = (i_tz_t+1-A_t')\,(1-r_t) \qquad t=0,1,...,T-1$$

a.2) las siguientes ecuaciones en diferencias:

$$\mu_{t+1}-\mu_t=-\mu_{t+1}i_t'$$

$$\lambda_{t+1}-\lambda_t=-\lambda_{t+1}(i_t-i_tr_t)$$

b) Las condiciones de contorno:

$$x_0=\bar{x}_0$$

$$z_0=\bar{z}_0$$

$$\mu_T=1-r_p$$

$$\lambda_T=1$$

c) El Principio del Máximo: Los controles óptimos deben maximizar la función hamiltoniana respecto a los controles admisibles.

La solución de a) y b) se puede obtener de forma recursiva empezando por el último periodo y observando el término general de la sucesión que aparece. Las soluciones para las variables de coestado son:

$$\mu_t = (1-r_p) \prod_{k=t}^{T-1} (1+i_k')$$

$$\lambda_t = \prod_{k=t}^{T-1} (1+i_k(1-r_k))$$

La interpretación económica de estas variables de coestado ha sido estudiada desde el inicio de la teoría del control óptimo, siendo el trabajo pionero el de Dorfman (1969), ya citado en III.3.2.. Su valor indica cómo afecta a la función objetivo las variaciones marginales en las variables de estado en el periodo t. En nuestro caso, se pueden interpretar de la siguiente manera:

μ_t: Aumento que se produce al recuperar la inversión en T si al final del periodo t-1 (o al principio del periodo t) hay en el PPI una u.m. adicional.

λ_t: Aumento que se produce al recuperar la inversión en T si al final del periodo t-1 (o al principio del periodo t) hay en el activo sin ventajas fiscales una u.m. adicional.

Si los tipos de interés se suponen constantes en el tiempo, las soluciones adoptan la siguiente forma:

$$\mu_t = (1-r_p)(1+i')^{T-t} \qquad (1)$$

$$\lambda_t = \prod_{k=t}^{T-1} (1+i(1-r_k)) \qquad (2)$$

Para obtener la solución a la condición c) se debe calcular previamente la derivada parcial de la función hamiltoniana respecto a la variable de control:

$$\frac{\partial H_t}{\partial A_t'} = \mu_{t+1} - \lambda_{t+1}(1-r_t) = p_t'$$

donde p_t' es la denominada función *switching* o función de cambio.

Dado que el problema es de control lineal y acotado, la solución a la condición c) está en los extremos del intervalo. De esta manera, aparece el tipo de solución *bang-bang* característica de los problemas lineales (ver epígrafe IV.3.1.).

Si la derivada parcial es positiva, la función hamiltoniana es creciente y la solución está en el extremo superior del intervalo, lo que en este contexto significa que es conveniente aportar la unidad monetaria en el PPI. Si es negativa el efecto es el contrario, es decir, lo conveniente es no aportar al PPI (valor de la variable igual a cero).

En consecuencia, la solución se puede describir de la siguiente manera:

si $\mu_{t+1} > \lambda_{t+1}(1-r_t)$ *aportar al PPI al final del periodo t.*

si $\mu_{t+1} < \lambda_{t+1}(1-r_t)$ *no aportar al PPI al final del periodo t.*

En términos económicos, dicho resultado dice que es conveniente aportar al PPI si el valor adicional en T que produce una unidad monetaria bruta aportada en t supera al valor adicional que se produciría si se invierte una unidad monetaria neta en t en el activo sin ventajas fiscales. Por tanto es una desigualdad que compara dos expresiones que incorporan todos los efectos fiscales de ambos instrumentos.

Por otra parte, el cociente entre el primer y el segundo miembro de la

desigualdad es lo que se llama el factor de ganancia del periodo t (Collins, 1980):

$$g_t^{PPI} = \frac{\mu_{t+1}}{\lambda_{t+1}(1-r_t)}$$

Y utilizando esta terminología se deduce que si el factor de ganancia es mayor que uno es preferible el PPI y si es menor que uno la conclusión es la contraria.

Si se sustituye las expresiones para las variables de coestado obtenidas anteriormente en la función *switching* se tiene:

$$p_t' = (1-r_p) \prod_{k=t+1}^{T-1} (1+i_k') - (1-r_t) \prod_{k=t+1}^{T-1} (1+i_k(1-r_k))$$

De donde se deduce que los factores que afectan a la decisión de inversión (al signo de p_t') son los siguientes:

- Rentabilidades financieras de los dos instrumentos.

- Años que faltan hasta la edad de 65.

- Tipos impositivos marginales de todos los periodos que faltan hasta la edad de 65 años (lo cual depende del perfil temporal de ingresos y de los parámetros de fiscalidad).

De todo ello se concluye que la valoración de los PPIs diferirá de un sujeto a otro en función de características personales como la edad y su estructura temporal de ingresos (o de tipos impositivos).

V.2.4.- Resultados adicionales

Las anteriores formulaciones permiten obtener como resultado adicional la rentabilidad financiero-fiscal de la aportación al final del periodo t al PPI (R_t^{PPI})[46]. Igualando la valoración de la aportación al PPI con la rentabilidad neta de impuestos se tiene:

$$(1-r_p)\prod_{k=t+1}^{T-1}(1+i_k') = (1-r_t)\,(1+R_t^{PPI})^{T-t-1}$$

Despejando se llega a la expresión de la rentabilidad financiero-fiscal de la aportación al PPI en el periodo t.

$$R_t^{PPI} = \left[\left(\frac{1-r_p}{1-r_t}\right)\prod_{k=t+1}^{T-1}(1+i_k')\right]^{\frac{1}{T-t-1}} - 1$$

En el caso de rentabilidades financieras constantes en el tiempo el resultado es:

$$R_t^{PPI} = \left(\frac{1-r_p}{1-r_t}\right)^{\frac{1}{T-t-1}}(1+i') - 1 \qquad (3)$$

[46]Aquí entendemos la rentabilidad financiero-fiscal como la rentabilidad neta de impuestos que se obtiene en la operación financiera de ahorro. No se adopta, por tanto, la otra acepción que la define como la rentabilidad que debería ofrecer el activo sin ventajas fiscales para que le fuera indiferente al inversor elegir una u otra forma de ahorro.

De esta expresión se extrae como característica interesante de los planes de pensiones el hecho de que se puede alcanzar en un determinado periodo una rentabilidad financiero-fiscal superior a la rentabilidad financiera, siempre que el tipo marginal del periodo pasivo (el que grava la recuperación del fondo) sea inferior al tipo marginal del periodo de la aportación.

Otro resultado adicional se consigue al descomponer la comparación entre el PPI y el ahorro convencional en dos partes. Efectivamente el cociente que define al factor de ganancia se puede descomponer en el producto de dos cocientes:

$$g_t^{PPI} = \frac{1-r_p}{1-r_t} \frac{\prod_{k=t+1}^{T-1}(1+i_k')}{\prod_{k=t+1}^{T-1}(1+i_k(1-r_k))}$$

Ambos cocientes miden respectivamente los efectos impositivo y de acumulación:

- el efecto impositivo puede ser mayor o menor que uno y puede variar en cualquier sentido a lo largo de los periodos;

- el efecto acumulación es mayor que uno, ya que no es de esperar una rentabilidad financiera en el ahorro convencional netamente superior a la del PPI, y es creciente en el tiempo.

Para que el PPI fuera preferible al ahorro convencional en un periodo t (factor ganancia mayor que uno) sería suficiente con que el efecto impositivo fuera mayor que uno (tipo marginal pasivo menor al del periodo t) pero esto no sería necesario, siempre y cuando el efecto acumulación compensara al impositivo.

V.3.- GENERALIZACIONES DEL MODELO

V.3.1.- Generalización del modelo con otros instrumentos de ahorro

El modelo básico descrito en el epígrafe V.2. para los planes de pensiones puede aplicarse, con las modificaciones adecuadas, a otros instrumentos de ahorro que han entrado en competencia con los planes de pensiones en el segmento del ahorro-pensión. La iliquidez de los PPIs junto con el hecho de que la recuperación del fondo esté sujeta a gravamen son desventajas respecto a otras formas de ahorro (por ejemplo los fondos de inversión y algunas formas de seguro) que son más líquidas y cuyas recuperaciones pueden quedar libres de impuestos si se cumplen ciertos requisitos. Este efecto puede incluso compensar la ventaja de las desgravaciones de las aportaciones, propias del plan de pensiones.

Para estudiar en qué casos ocurrirá esto, se procede a continuación a aplicar el modelo anterior al caso de los fondos de inversión (FI). La elección de los FI y no de otras formas de ahorro se debe a la popularidad que este instrumento de ahorro ha alcanzado en los últimos años. Además, los supuestos que se adoptan en el modelo pueden cumplirse también si se eligen otras formas de ahorro.

Pese a que los FI existían con anterioridad, el marco legal que realmente los tipifica se remonta a la Ley 46/1984 de Instituciones de Inversión Colectiva. Con posterioridad aparece nueva legislación que los regula de forma sistemática y los va adaptando a las nuevas condiciones económicas[47]. Las principales características fiscales que se pueden extraer de cara a la modelización matemática son las tres siguientes:

---

[47]Real Decreto 1346/1985 que recoge el Reglamento de las instituciones de inversión colectiva, modificaciones de dicho Reglamento (Reales Decretos 1393/1990 y 686/1993) y Ley 24/1988 reguladora del Mercado de Valores. Aparte de esta legislación son importantes las disposiciones aparecidas en los sucesivos Presupuestos Generales del Estado.

1. Las aportaciones no desgravan en el momento de producirse.

2. Los rendimientos del fondo se acumulan y no están sujetos a gravamen hasta que se recuperan[48].

3. Cuando se recupera el fondo existen ventajas fiscales. De ellas, la única relevante que interesa en el modelo, por el momento, es la exención total que se produce si han transcurrido más de 15 años desde la aportación.

Con estas características se puede adaptar el modelo presentado en el epígrafe V.2.2. al caso de los FI[49], manteniendo los supuestos de tipo general, e incorporando otros:

S.9. Los fondos de inversión que se consideran en el análisis invierten en activos financieros similares a los del PPI con el objetivo de limitar la incertidumbre sobre la rentabilidad financiera de estos instrumentos de ahorro de forma que sea comparable a la de un PPI.

S.10. El sujeto está en disposición de lograr las máximas ventajas fiscales, es decir, consigue la exención total cuando recupera las aportaciones.

Como consecuencia de las características y los supuestos anteriores, las diferencias que surgen al ver el efecto impositivo son dos:

[48]Esto es cierto para los fondos de inversión de acumulación que son los mayoritarios y a los que se aplica el modelo.

[49]Aspectos como las diferencias de liquidez y seguridad entre los planes de pensiones y los fondos de inversión no pueden recogerse en el modelo ya que sólo se incorporan las características de rentabilidad y fiscalidad.

1.- El impuesto de cada periodo no queda minorado por las aportaciones al FI.

2.- El impuesto final no existe[50].

Estas mismas características se pueden dar, bajo ciertas circunstancias, en otras formas de ahorro como el seguro de capital diferido. Por una parte, las aportaciones o primas no están exentas. Por otra, se puede conseguir la exención impositiva total al recuperar el ahorro acumulado. Efectivamente, en el seguro de capital diferido (cuyo tratamiento fiscal se recoge en el Real Decreto 1841/1991), se consigue si han transcurrido más de 22 años desde la primera prima (bajo el sistema simplificado); aunque, en contrapartida a este mayor periodo, el ahorro se puede recuperar totalmente en el primer periodo pasivo (en los PPIs esta posibilidad supondría un mayor tipo impositivo medio y en el caso de los FI una muy probable pérdida de ventajas fiscales). Adicionalmente, el seguro de capital diferido disfrutará de una rentabilidad superior al añadir los beneficios de los que no sobrevivan. La diferencia estriba en el mayor riesgo de este instrumento de ahorro[51].

Al introducir los nuevos efectos fiscales se llega al Problema **[II]** que, aunque referido a los FI, se puede extender a los seguros de capital diferido bajo las circunstancias adecuadas antes comentadas.

---

[50]Esto es consecuencia del supuesto *S.10.*. El sujeto del modelo está en una situación ideal ya que dispone del capital acumulado en el FI de una forma libre de impuestos. Bajo la legislación actual eso significa que siempre vende participaciones con más de 15 años de antigüedad.

[51]Un estudio completo sobre la variedad y características de estos y otros instrumentos de ahorro-previsión puede verse en Gálvez y otros (1992).

$$Max \quad y_T+z_T$$

s.a:

$$y_{t+1}-y_t = i_t''y_t+A_t'' \qquad t=0,1,...,T-1$$

$$z_{t+1}-z_t = (i_t z_t+1)\,(1-r_t)-A_t'' \qquad t=0,1,...,T-1 \qquad \text{[II]}$$

$$y_0,\ z_0 \quad \textit{conocidos}$$

$$y_t \geq 0,\ z_t \geq 0 \qquad t=0,1,...,T$$

$$0 \leq A_t'' \leq 1 \qquad t=0,1,...,T-2$$

$$A_{T-1}''=0$$

La notación es similar a la seguida en el Problema **[I]**. Las nuevas variables son:

y_t: Importe acumulado en el fondo de inversión al principio del periodo t.

i_t'': Rentabilidad del FI durante el periodo t.

A_t'': Aportación al FI al final del periodo t.

La aplicación del P.M. discreto sigue los mismos pasos que en el epígrafe anterior.

La función hamiltoniana adopta la siguiente forma:

$$H_t = \omega_{t+1} (i_t'' y_t + A_t'') + \lambda_{t+1} [(i_t z_t + 1) (1 - r_t) - A_t'']$$

Las condiciones de óptimo son:

a) El sistema hamiltoniano dinámico. Está compuesto por:

a.1) el sistema dinámico de ecuaciones en diferencias que aparece en el Problema **[II]**.

$$y_{t+1} - y_t = i_t'' y_t + A_t'' \qquad t=0,1,...,T-1$$

$$z_{t+1} - z_t = (i_t z_t + 1) (1 - r_t) - A_t'' \qquad t=0,1,...,T-1$$

a.2) las siguientes ecuaciones en diferencias:

$$\omega_{t+1} - \omega_t = -\omega_{t+1} i_t''$$

$$\lambda_{t+1} - \lambda_t = -\lambda_{t+1}(i_t - i_t r_t)$$

b) Las condiciones de contorno:

$$y_0 = \bar{y}_0$$

$$z_0 = \bar{z}_0$$

$$\omega_T = 1$$

$$\lambda_T = 1$$

c) El Principio del Máximo: Los controles óptimos deben maximizar la función hamiltoniana respecto a los controles admisibles.

Las soluciones que aparecen ahora para las variables de coestado son:

$$\omega_t = \prod_{k=t}^{T-1} (1+i_k'') \qquad (4)$$

$$\lambda_t = \prod_{k=t}^{T-1} (1+i_k(1-r_k))$$

Ahora la variable de coestado, ω_t, se interpreta como el aumento que se produce al recuperar la inversión en T si al final del periodo t-1 (o al principio del periodo t) hay en el FI una unidad monetaria adicional.

Para que se cumpla c) se calcula la derivada parcial respecto a la variable de control:

$$\frac{\partial H_t}{\partial A_t''} = \omega_{t+1} - \lambda_{t+1} = p_t''$$

donde p_t'' es la nueva función *switching*.

Razonando igual que se hizo en el modelo básico se tiene ahora que la estrategia de ahorro en el fondo es la siguiente:

si $\omega_{t+1} > \lambda_{t+1}$ *aportar al FI al final del periodo t.*

si $\omega_{t+1} < \lambda_{t+1}$ *no aportar al FI al final del periodo t.*

La interpretación económica de la solución es también lógica: es conveniente aportar al FI si el valor adicional en T que produce una u.m. invertida en t supera al valor adicional que se produciría si se invierte esa u.m. en t en el activo sin ventajas fiscales.

Al construir el factor de ganancia, g_t^{FI}, se observa que no existe efecto impositivo por lo que siempre es mayor que uno (exceptuando el caso poco probable que la rentabilidad del ahorro convencional excediera en mucho la del fondo de inversión) y creciente en el tiempo:

$$g_t^{FI} = \frac{\omega_{t+1}}{\lambda_{t+1}} = \frac{\prod_{k=t+1}^{T-1}(1+i_k'')}{\prod_{k=t+1}^{T-1}(1+i_k(1-r_k))}$$

La rentabilidad financiero-fiscal de la aportación en t al Fondo de Inversión (R_t^{FI}) entendida como la rentabilidad neta de impuestos que se obtiene en la operación financiera de ahorro da lugar a la igualdad:

$$\prod_{k=t+1}^{T-1}(1+i_k'') = (1+R_t^{FI})^{T-t-1}$$

Despejando se tiene:

$$R_t^{FI} = [\prod_{k=t+1}^{T-1} (1+i_k'')]^{\frac{1}{T-t-1}} - 1$$

Y en el caso de tipos de interés constantes en el tiempo se llega a:

$$R^{FI} = i'' \qquad (5)$$

Esta relación de igualdad es válida para todos los periodos. Esta característica no se da en planes de pensiones, donde la relación depende del periodo t y no tiene porque ser de igualdad (ambas rentabilidades pueden diferir en cualquier sentido).

Los resultados obtenidos en el modelo con PPI, problema **[I]**, y en el modelo con FI, problema **[III]**, comparan cada alternativa de ahorro, por separado, con un instrumento de ahorro sin ventajas fiscales. Sin embargo, dado que este instrumento de ahorro convencional es el mismo en las dos aplicaciones, puede servir de puente para comparar el PPI y el FI entre sí. Efectivamente, la variable de coestado λ_t es la misma en ambos modelos, lo cual permite enlazar las estrategias óptimas de ahorro de los dos modelos para determinar cuál de los dos instrumentos es más conveniente, dando lugar a la siguiente estrategia:

$$si \quad \mu_{t+1} > \omega_{t+1}(1-r_t) \quad \textit{aportar al PPI al final del periodo } t$$

$$si \quad \mu_{t+1} < \omega_{t+1}(1-r_t) \quad \textit{aportar al FI al final del periodo } t$$

Desarrollando las variables de coestado y en el caso de rentabilidades financieras constantes en el tiempo la estrategia es:

$$si \quad (1-r_p)\,(1+i')^{T-t-1} > (1-r_t)\,(1+i'')^{T-t-1} \qquad aportar\ al\ PPI$$

$$si \quad (1-r_p)\,(1+i')^{T-t-1} < (1-r_t)\,(1+i'')^{T-t-1} \qquad aportar\ al\ FI$$

Se observa que los dos factores claves para determinar la mejor alternativa de ahorro son:

1. Las diferencias de rentabilidades financieras. Dependen del mercado financiero, de las comisiones de las entidades gestoras y depositarias, de la calidad de la gestión de los activos, de la fiscalidad de las instituciones de inversión colectiva, etc.. En este sentido, el entorno financiero en el que actúan las gestoras es similar en un PPI y en un FI (atendiendo al supuesto S.9.) por lo que, a largo plazo, es de esperar que las diferencias de rentabilidad financiera sean escasas[52].

2. De las diferencias entre los tipos impositivos marginales del periodo pasivo y del periodo de la aportación. Esto depende del perfil de ingresos esperados por el individuo en su etapa activa y pasiva y de los cambios en la fiscalidad.

En el caso de rentabilidades financieras iguales, el individuo tendrá que comparar si el tipo impositivo marginal del periodo actual es superior o inferior al que espera tener tras los 65 años. Si es superior ahorrará a través de un PPI y si es inferior a través de un FI. Es decir, el perfil temporal de tipos impositivos futuros se convierte en la principal variable que determina la estrategia de ahorro óptima.

[52]Tal vez se espera una pequeña prima favorable a los PPI debido a que la gestora está exenta del impuesto de sociedades mientras que la del FI paga un 1%. Un seguro de capital diferido disfrutará, en este aspecto, de una ventaja siempre y cuando se consiga la supervivencia; pero con el inconveniente de que la empresa aseguradora paga un tipo del 35% en el impuesto de sociedades.

Las mismas conclusiones se obtienen utilizando la terminología del factor de ganancia y desglosando los efectos impositivo y de acumulación. La expresión que se obtiene ahora es:

$$g_t = \frac{\mu_{t+1}}{\omega_{t+1}(1-r_t)} = \frac{1-r_T}{1-r_t} \frac{\prod_{k=t+1}^{T-1}(1+i_k')}{\prod_{k=t+1}^{T-1}(1+i_k'')}$$

Se observa, por tanto, que un PPI no siempre resulta ser la mejor alternativa para complementar la pensión de jubilación pese a estar especialmente creado para esta forma de ahorro. Aparte de esta conclusión, no hay que olvidar que existe otro inconveniente en los PPIs: la rigidez o falta de liquidez. Esto supone que las otras formas de ahorro se convierten en complementarias de un PPI incluso aunque éste sea mejor desde el punto de vista financiero-fiscal analizado en el modelo[53]. Sin embargo, no hay que descartar otro efecto no recogido en el modelo y que juega a favor de los planes de pensiones: el fenómeno de ilusión monetaria. Este fenómeno es la mayor percepción del ahorro fiscal presente que del pago fiscal futuro y es de suponer que es mayor cuanto mayor sea el tipo impositivo presente del sujeto[54].

El modelo se ha descrito suponiendo que se dispone de una unidad monetaria adicional para ahorrar y, por tanto, las conclusiones son válidas desde un punto de vista marginalista ya que esa unidad de ahorro no afecta al tipo impositivo marginal.

---

[53]Algunos autores ven en la iliquidez de los PPI una ventaja porque responde mejor al objetivo de complementar los ingresos tras la jubilación. Por otra parte, la liquidez de las otras formas de ahorro puede llevar asociada la pérdida de ventajas fiscales.

[54]Existen numerosos estudios empíricos en el caso americano que destacan una relación directa muy evidente entre tipos impositivos marginales y aportaciones a los IRAs, por ejemplo Long (1988, 1990) y O'Neil y Thompson (1987, 1988).

Extender las conclusiones para cantidades de ahorro no marginales no es, en principio, correcto porque sí que afectarían al tipo marginal. Sin embargo, no se produciría grandes distorsiones debido a que entran en juego dos efectos que tienden a compensarse:

1°) La aportación a un PPI reduce el tipo impositivo del periodo, lo que ofrece una ventaja adicional para este instrumento frente a los demás.

2°) Su recuperación provoca un aumento del tipo impositivo tras los 65 años, lo cual perjudica a los PPIs.

Ambos efectos son contrarios aunque el 2° será posiblemente superior al 1°, *ceteris paribus*. El argumento es que la rentabilidad del PPI se espera que supere la deflactación de la tarifa (la inflación), es decir, son de esperar rentabilidades reales positivas. Entonces el capital que se recupera es superior en términos reales al que se aportó y por tanto el impacto sobre el tipo impositivo final (efecto 2°) será mayor que el que tuvo sobre el inicial (efecto 1°). En el paso de lo marginal a lo absoluto se desprende, por tanto, una ligera pérdida competitiva para los PPIs.

V.3.2.- Revisión del modelo con planes de pensiones

Las conclusiones a las que se ha llegado indican que la fiscalidad no es suficientemente alentadora para canalizar a través de un PPI la mayor parte del ahorro con fines de complemento a la pensión pública de jubilación, y más si se tiene en cuenta la iliquidez de este instrumento[55]. Existen propuestas de distinto signo para potenciar los PPIs, desde reformas a fondo del sistema de pensiones públicas hasta reformas de

[55]Una muestra clara de la preocupación de las entidades promotoras de fondos de pensiones por la falta de liquidez es la reciente aparición de fondos en los que se posibilita la liquidez mediante fórmulas de préstamo especialmente concebidas.

tipo técnico.

Las propuestas más de fondo pasan por una reforma del sistema de pensiones de jubilación de la Seguridad Social. A veces se argumenta que el sistema público debería quedar reducido a pensiones mínimas o, incluso, a pensiones no contributivas; dejando el resto para ser cubierto a nivel privado, a través de planes de pensiones individuales y de empresa. El modelo británico se adapta a esta propuesta al mantener un nivel de cobertura básico (pensiones públicas mínimas) y un nivel complementario que puede ser público o privado. Esta propuesta impulsaría la creación de más fondos de pensiones pero no corregiría las desventajas anteriormente citadas de iliquidez y posiblemente de fiscalidad.

Otras propuestas de reforma son más técnicas e intentan aumentar la ventaja comparativa de los PPIs frente a los otros instrumentos de ahorro. Existen varias maneras de plantear estas reformas técnicas y aquí vamos a estudiar los efectos de dos de ellas (una visión más completa así como la materialización práctica se puede seguir en Vidal, 1994a).

El primer modelo de reforma es el americano. Actúa en la línea de disminuir la rigidez permitiendo el reembolso de una parte del fondo acumulado a cambio de una penalización en forma de porcentaje del capital retirado (este porcentaje puede ser fijo o variable). Esta modificación se puede introducir en nuestro modelo cambiando el supuesto de que el horizonte de planificación son los 65 años y reformulando la función objetivo. En este caso la penalización disminuiría la rentabilidad financiero-fiscal aunque dotaría a este tipo de ahorro de una mayor liquidez.

El segundo modelo es el británico. Actúa eximiendo del impuesto a una parte de los reembolsos finales. Esta reducción impositiva se puede plantear fija o variable (en función de la edad, tipo impositivo, cuantía, etc.). Con ello se produce un aumento de

la rentabilidad financiero-fiscal que hace aumentar la ventaja de los PPIs sobre los otros instrumentos, intentando así compensar la falta de liquidez. La modelización del problema incluyendo esta modificación es posible reformulando la función objetivo del modelo básico.

Dado el Problema [I], los cambios que se introducen en la reforma que dota de mayor liquidez a los PPIs a cambio de una penalización (modelo americano) sólo afectan a la función objetivo, que pasa a ser:

$$Max \quad z_{T'}+x_{T'}(1-\beta)(1-r_{T'})$$

donde T' es ahora la edad del inversor cuando se produce la retirada de fondos menos la edad del inversor en el momento de la aportación más uno. Por otra parte, β indica la penalización por la retirada anticipada (en tanto por uno y sobre la cantidad retirada). Suponiendo rentabilidades constantes para no recargar la notación, la función de coestado que se utiliza para obtener la estrategia óptima de ahorro es ahora:

$$\mu_t' = (1-\beta)(1-r_{T'}) \ (1+i')^{T'-t} \qquad (6)$$

y la rentabilidad financiero-fiscal de una unidad monetaria aportada al final del periodo t y retirada en T' debe ser, *ceteris paribus*, menor que antes de la reforma debido a la penalización:

$$R_t^{PPI'} = \left(\frac{(1-\beta)(1-r_{T'})}{1-r_t}\right)^{\frac{1}{T'-t-1}} (1+i') - 1 < R_t^{PPI} \qquad (7)$$

donde la desigualdad es válida si han transcurrido el mismo número de años en los dos modelos y los tipos marginales son iguales en el periodo T' y en el periodo pasivo.

Las expresiones (6) y (7) son las que sustituyen a la (1) y (3) respectivamente. Se pueden comparar con las obtenidas para el FI, expresiones (4) y (5), siempre y cuando el nuevo periodo T' permita una enajenación de participaciones de más de 15 años de antigüedad. La comparación lleva a una pérdida comparativa de los PPIs frente a los FI debido a la penalización. Como es lógico la ventaja de la liquidez da lugar a una menor rentabilidad financiero-fiscal. Sin embargo, también es cierto que la liquidez de los otros instrumentos de ahorro puede llevar consigo la pérdida de ventajas fiscales con lo que las expresiones (4) y (5), referidas al FI, se modifican. Este efecto se puede medir con nuestro modelo básico.

Por ejemplo, en el caso del FI, si todas las participaciones no se reembolsan tras 15 años sino tras un número menor de años ($m \leq 15$)[56] no se produce la exención impositiva total, y el valor del fondo acumulado en T' queda disminuido por un término que indica el pago del impuesto. La función objetivo del Problema **[II]** queda:

$$z_{T'} + y_{T'} (1 - c_m)$$

donde c_m es un término variable que indica el pago impositivo que corresponde a una unidad monetaria del FI reembolsada en T'. Este término depende de m, de la rentabilidad financiera y de los tipos medio y marginal del sujeto en T', debido al tratamiento fiscal existente para los reembolsos de un FI. Ahora la variable de coestado que sirve para realizar las comparaciones es:

[56] La variable m hace referencia a un número de años a efectos fiscales, es decir, es el número entero por exceso que corresponde a los años efectivos de antigüedad de las participaciones.

$$\omega_t^{/} = (1-c_m)(1+i^{//})^{T'-t} \qquad (8)$$

Y la rentabilidad financiero-fiscal de una unidad monetaria invertida al final del periodo t y retirada en T' es:

$$R_t^{FI'} = (1-c_m)^{\frac{1}{T'-t-1}} (1+i^{//}) - 1 \qquad (9)$$

Las expresiones (8) y (9) sustituyen a (4) y (5) respectivamente. En ellas m sería igual a T'-t-1, aunque se aprovecharían más las ventajas fiscales del FI con sólo atrasar un día el momento de la recuperación del fondo, con lo cual $m=T'-t$.

El ahorro convencional no pierde por un reembolso anticipado por lo que la variable de coestado asociada, λ_t (recogida en la expresión (2)), es la misma salvo el nuevo valor del periodo T'. Naturalmente la desventaja del ahorro convencional se reducirá respecto a los PPIs y FI aunque, probablemente, no llegará a anularse debido al peso del efecto acumulación. Por su parte, la comparación entre PPI y FI tendrá como nuevos elementos de análisis el importe de la penalización por la retirada anticipada del PPI y el número de años de permanencia de las aportaciones al FI.

Bajo el segundo tipo de reformas, las que aumentan las ventajas fiscales del reembolso durante el periodo pasivo (modelo británico), de nuevo es la función objetivo la que recoge los cambios. En concreto, cambia el término que se encarga de minorar el fondo acumulado en T. Si se supone que una parte constante α de la cuantía recuperada está exenta de impuesto, la función objetivo a maximizar del Problema **[I]** se sustituye por la siguiente:

$$x_T\,(1-(1-\alpha)r_p) + z_T$$

lo que equivale a decir que los importes recuperados pagan un tipo impositivo menor. El análisis que sigue es el mismo que en el epígrafe V.2. La diferencia es que en las distintas formulaciones el tipo marginal final aparece multiplicado por el factor (1-α). En consecuencia la estrategia óptima de ahorro es ahora:

$$si \quad \mu''_{t+1} > \lambda_{t+1}(1-r_t) \qquad aportar\ al\ PPI$$

$$si \quad \mu''_{t+1} < \lambda_{t+1}(1-r_t) \qquad no\ aportar\ al\ PPI$$

donde la nueva variable de coestado, con rentabilidades constantes, es:

$$\mu''_t = [1-r_p(1-\alpha)]\ (1+i')^{T-t} \qquad (10)$$

Y la nueva rentabilidad financiero-fiscal es mayor a la de antes de la reforma:

$$R_t^{PPI''} = \left(\frac{1-r_p(1-\alpha)}{1-r_t}\right)^{\frac{1}{T-t-1}} (1+i') - 1 > R_t^{PPI} \qquad (11)$$

Las expresiones (10) y (11) son las que sustituyen ahora a (1) y (3) respectivamente. Esta reforma supone una ventaja adicional sobre el activo sin ventajas fiscales y sobre el FI, ya que actúa sobre el efecto impositivo (es como una reducción del impuesto del periodo pasivo).

V.3.3.- Introducción de una componente estocástica

En un modelo a largo plazo como el descrito en el epígrafe V.2. existen numerosas fuentes de incertidumbre que pueden afectar a la solución del modelo. El objetivo ahora es considerar alguna de ellas para observar la influencia que puede tener sobre la solución. No se trata de realizar un tratamiento exhaustivo de cómo afecta la incertidumbre al modelo sino sólo ilustrar con algún ejemplo el manejo de componentes estocásticas en sistemas dinámicos discretos.

En primer lugar, se considera que el tipo impositivo marginal de cada periodo no se conoce con certeza pero se supone que se conoce su distribución de probabilidad. Se tiene de esta manera un problema pseudo-estocástico (ver epígrafe III.4.1.) porque las ecuaciones en diferencias no son estocásticas pero sí lo es un parámetro. El problema se resuelve planteando nuevas ecuaciones en diferencias, que serán deterministas, pero donde las incógnitas son la media y la varianza de las variables correspondientes. Teniendo en cuenta sólo la comparación entre el PPI y el FI, recuérdese que la estrategia de ahorro óptima es:

$$si \quad \mu_{t+1} > \omega_{t+1}(1-r_t) \quad aportar\ al\ PPI$$

$$si \quad \mu_{t+1} < \omega_{t+1}(1-r_t) \quad aportar\ al\ FI$$

Ahora, suponiendo una distribución normal para el tipo marginal con una varianza creciente linealmente en función del periodo de tiempo[57], se tiene:

[57]Esta es una posibilidad de las muchas que se podrían contemplar y supone que el tipo marginal se distribuye como una normal cuya media se obtiene a partir de la estructura temporal de ingresos y de la función impositiva. La varianza es creciente porque cuanto más nos alejamos del periodo actual más desviaciones se pueden producir respecto al valor medio.

$$r_t \rightarrow N(\bar{r}_t, \sigma^2 t)$$

Si se supone que el tipo marginal del periodo pasivo es el del último periodo activo y los tipos de interés son constantes, la ecuación en diferencias que hay que resolver para la variable de coestado asociada al PPI (μ_t) es:

$$\mu_{t+1} - \mu_t = -\mu_{t+1} i'$$
$$\mu_T = 1 - r_T$$
$$r_t \rightarrow N(\bar{r}_t, \sigma^2 t)$$

La componente estocástica se encuentra en la condición terminal de la variable de coestado. Reordenando y tomando esperanzas se tiene:

$$(1+i') E[\mu_{t+1}] = E[\mu_t]$$
$$E[\mu_T] = 1 - \bar{r}_T$$

Esto es una ecuación en diferencias determinista cuya solución por el método recursivo proporciona la esperanza de la variable de coestado:

$$E[\mu_t] = (1-\bar{r}_T)(1+i')^{T-t}$$

De esta manera, se observa que la expresión obtenida para la esperanza de la variable de coestado coincide con la expresión de dicha variable bajo un punto de vista determinista. Respecto a la varianza se puede razonar de la misma manera llegando a

la ecuación en diferencias siguiente:

$$(1+i')^2 \; var\; [\mu_{t+1}] = var\; [\mu_t]$$

$$var\; [\mu_T] = \sigma^2 T$$

Su solución es:

$$var\; [\mu_t] = \sigma^2 T\; (1+i')^{2(T-t)}$$

Por otra parte, la variable de coestado asociada a un FI (w_t) no depende del tipo impositivo marginal pero dicha variable se multiplica por el término ($1-r_t$) para compararse con el PPI y deducir la estrategia óptima. Por tanto, hay que estudiar la influencia de la componente estocástica sobre la variable de coestado ajustada por ese término. El resultado se obtiene simplemente aplicando la definición de media y varianza y es:

$$E\; [\omega_{t+1}(1-r_t)] = \omega_{t+1}(1-\bar{r}_t) = (1+i'')^{T-t-1}\; (1-\bar{r}_t)$$

$$var\; [\omega_{t+1}(1-r_t)] = \omega^2_{t+1}\sigma^2 t = (1+i'')^{2(T-t-1)}\; \sigma^2 t$$

De nuevo, se obtiene el mismo resultado que en el caso determinista si se sustituye la variable por su media. Nótese que si la rentabilidad financiera es igual en el PPI y el FI, la varianza del valor neto de impuestos es superior en el PPI como consecuencia de la hipótesis de que la varianza del tipo marginal es creciente en el tiempo. Se observa también que la estrategia óptima no diferirá respecto a la del caso determinista si se escogen los valores medios como sustitutos de las variables estocásticas ya que se obtiene:

$$si \quad (1-\bar{r}_T)\,(1+i')^{T-t-1} > (1-\bar{r}_t)\,(1+i'')^{T-t-1} \qquad \textit{aportar al PPI}$$

$$si \quad (1-\bar{r}_T)\,(1+i')^{T-t-1} < (1-\bar{r}_t)\,(1+i'')^{T-t-1} \qquad \textit{aportar al FI}$$

Sin embargo, la varianza de ambos miembros de la desigualdad es distinta, lo cual se puede utilizar si se desea ajustar la estrategia óptima. Efectivamente, en ausencia de ajustes si ambos miembros son iguales sería indiferente un tipo de instrumento u otro porque ambos llevarían al mismo valor neto de impuestos como media; sin embargo, el valor obtenido a través de un PPI tiene una varianza mayor y, en este sentido, está sometido a un mayor riesgo. Esto podría justificar realizar un ajuste a través de la suma de un término al segundo miembro de la desigualdad que debería ser positivo y decreciente en el tiempo para recoger las anteriores consideraciones.

El segundo y último parámetro que vamos a considerar estocástico es el tipo de interés. Bajo el supuesto que las rentabilidades son iguales en todos los instrumentos de ahorro, si el tratamiento fiscal de los rendimientos de cada periodo es también el mismo (están exentos en el PPI y en el FI) no habrá ningún efecto que distorsione la estrategia óptima de ahorro seguida en el caso determinista. Por ello, lo único relevante es estudiar cómo se produce la acumulación de capital cuando el tipo de interés es estocástico.

La complejidad de las formulaciones aconseja, si se quiere llegar a resultados analíticos, partir de una ecuación en diferencias simple (con término independiente igual a cero). La siguiente ecuación representa como se produce la acumulación de un fondo de capital partiendo de un valor inicial cierto y sin aportaciones ni detracciones durante los periodos:

$$x_t = x_{t-1}(1+i_t) \qquad t=1,2,...,T$$

$$x_0 = \bar{x}_0$$

El tipo de interés de cada periodo es estocástico a partir de un tipo inicial cierto y el supuesto que se adopte acerca de su distribución de probabilidad será determinante para deducir la correspondiente evolución de la media y la varianza de la variable x_t (capital acumulado en t). Tres posibilidades, entre las múltiples que se podrían plantear, son:

1.- $i_t = i+u_t \qquad u_t \rightarrow N\ [0,\sigma^2]$

2.- $i_t = i+u_t \qquad u_t \rightarrow N\ [0,\sigma^2 t]$

3.- $i_t = i_{t-1}+u_t \ , \ i_0=i \qquad u_t \rightarrow N\ [0,\sigma^2]$

En los tres casos el tipo de interés sigue una normal con los siguientes parámetros descriptivos:

	E (i_t)	var (i_t)	cov (i_t,i_{t-k})
1	i	σ^2	0
2	i	$\sigma^2 t$	0
3	i	$\sigma^2 t$	$\sigma^2(t-k)$

Tabla V.1.: Parámetros descriptivos del tipo de interés.

El fondo acumulado, en cambio, no sigue una distribución normal. Sin embargo, es posible obtener su media, varianza y covarianza aplicando el método recursivo en los dos primeros casos:

	E (x_t)	var (x_t)	cov (x_t, x_{t-k})
1	$x_0(1+i)^t$	$\sigma^2 t x_0^2(1+i)^{2t-2}$	$\sigma^2(t-k)x_0^2(1+i)^{2t-k-2}$
2	$x_0(1+i)^t$	$\sigma^2 t(t+1)x_0^2(1+i)^{2t-2}/2$	$\sigma^2(t-k)(t-k+1)x_0^2(1+i)^{2t-k-2}/2$

Tabla V.2.: Parámetros descriptivos del fondo de capital.

El tercer caso y otros que se pudieran plantear con correlación entre los tipos de interés de los periodos resultan más complejos de resolver analíticamente por lo que haría falta acudir a técnicas de simulación.

V.4.- APLICACIÓN NUMÉRICA

A continuación se desarrolla un ejemplo numérico de esta aplicación, en concreto se utiliza el modelo general, problema **[I]**, el modelo con FI, problema **[III]**, y la segunda propuesta de reforma de los PPIs. La característica principal del sujeto es que su estructura temporal de ingresos en el periodo activo tiene la forma de una campana de Gauss (se emplea por ejemplo en Ko, 1988)[58]. Se supone también que los ingresos del periodo pasivo son tales que los tipos impositivos a los que se verá sometido coinciden con los del último periodo activo.

La función de ingresos que se adopta para el periodo activo es la siguiente:

$$Y_t = A\, e^{\frac{-(t-M)^2}{p}}$$

donde los parámetros indican:

A: renta máxima que se percibe en términos nominales (el valor máximo de la campana de Gauss).

M: periodo en que se percibe la renta máxima.

p: es un coeficiente que indica el apuntamiento de la campana de Gauss. Cuanto menor sea *p* mayor es el apuntamiento, es decir, más aumenta (disminuye) el ingreso antes (después) de la edad de máximo ingreso. Este valor se puede obtener si se conoce la renta inicial, *A* y *M*.

[58]Este tipo de función de ingresos se asigna, en Vidal (1994a), al colectivo de profesionales libres o trabajadores por cuenta propia.

La función de impuestos para determinar el tipo marginal es:

$$I_t = a_t Y_t^b$$

El parámetro b indica la progresividad del impuesto ($b > 1$) y se supone constante en el tiempo. El coeficiente a_t indica la magnitud del impuesto y va evolucionando en el tiempo a medida que cambia la presión fiscal. Se supondrá que la única forma de disminuir o aumentar la magnitud del impuesto a lo largo del tiempo es deflactando más o menos que la inflación la tabla del IRPF. Así, partiendo de un valor conocido para a_0, el coeficiente a_t evolucionará según la trayectoria siguiente:

$$a_t = a_0(1+d)^{t(1-b)}$$

donde d es el coeficiente que indica la deflactación de la tarifa en tanto por uno y se supone constante en el tiempo[59]. Para estar ante un marco tributario estable, en el sentido que capitales económicamente equivalentes en el tiempo sean gravados al mismo tipo impositivo, es necesario, entre otras cosas, que el valor de d sea igual al de la inflación.

Los valores numéricos que se toman son[60]:

---

[59]Esta trayectoria se ha deducido a partir de las igualdades que aparecen al suponer que un aumento en tanto por uno de la renta de cuantía d debe suponer el mismo aumento del impuesto, es decir:

$$I_t = a_t Y_t^b \qquad I_{t+1} = a_{t+1} Y_{t+1}^b$$

$$Y_{t+1} = Y_t(1+d) \qquad I_{t+1} = I_t(1+d)$$

[60]Los valores de a_0 y b se consiguen ajustando la tarifa del I.R.P.F. español del año 94 a través de la función exponencial correspondiente. El valor del coeficiente R^2 es superior a 0,99.

* Edad del individuo: 35 años. Supone un valor $T=31$.

* $x_0=y_0=z_0=0$

* $a_0=0{,}00572$, $b=1{,}4485$, $d=0{,}04$

* $i=i'=i''=0{,}07$ constantes en el tiempo.

* $A=18000$ (miles de pesetas).

* $M=30$ (edad de máximo ingreso en términos nominales: 64 años).

* $Y_0=3000$ (miles de pesetas).

* $\alpha=0{,}1$

La estructura temporal de ingresos y de tipos impositivos marginales se recoge en el gráfico V.1.

El valor de p que resulta es 502,3. Los tipos impositivos marginales inicial, máximo y final son 31,1%, 43,2% y 38,9%, respectivamente.

En el gráfico V.2. se recogen las funciones relevantes de cara a elegir la estrategia de ahorro óptima (variables de coestado). Para que estas funciones sean comparables, las asociadas al FI y al AC se han multiplicado por el factor $(1-r_t)$, dando lugar a las variables de coestado ajustadas.

Dicho gráfico permite identificar cuál es la mejor alternativa en cada periodo en función de la curva que sea superior, lo que da la solución del problema. Obsérvese que las funciones son discretas, aunque se han representado como si fueran contínuas para una mejor visualización de las trayectorias.

En el ejemplo planteado se prefiere el FI hasta el periodo 9 (43 años) y a partir de ahí el PPI. En caso de exención de una parte del impuesto final el periodo que marca el cambio de instrumento de ahorro es el 4 (38 años).

Así pues, el sujeto del ejemplo no utilizará los Planes de Pensiones hasta los 44 años si su único criterio de decisión es el de obtener la máxima rentabilidad financiero-fiscal.

En el ejemplo numérico también se pueden obtener las rentabilidades financiero-fiscales de las aportaciones. Se muestra, como ejemplo, las que se obtienen en el primer periodo y en el periodo de máximo tipo marginal (54 años):

	PPI	FI	PPI'
35 años	6,58%	7%	6,79%
54 años	7,71%	7%	8,32%

Tabla V.3.: Rentabilidades financiero-fiscales en el ejemplo numérico.

Nótese como este procedimiento operativo permite obtener una amplia información sobre los resultados deseados y la evolución de los mismos a partir de las diferentes hipótesis que se realicen para modelizar las situaciones reales.

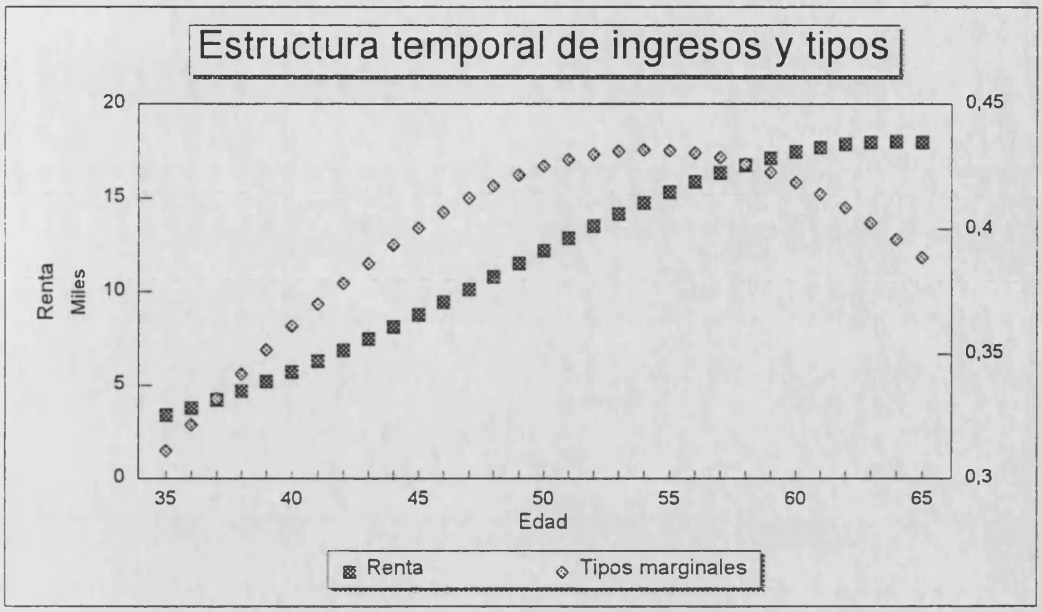

Gráfico V.1.: Evolución temporal de ingresos y tipos impositivos.

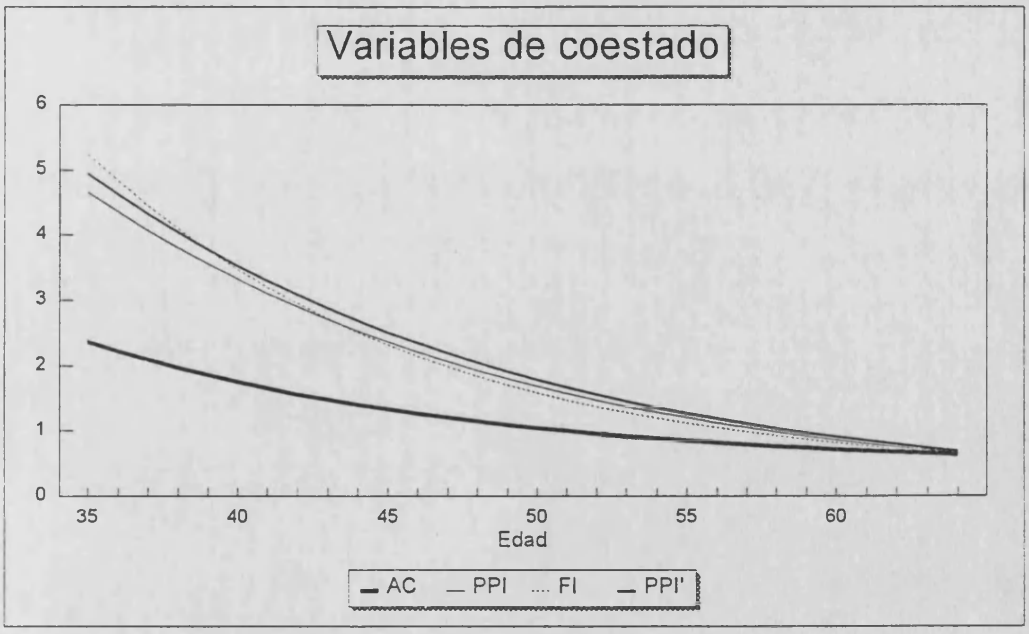

Gráfico V.2.: Variables de co-estado ajustadas.

CAPÍTULO VI:

APLICACIONES A LA FINANCIACIÓN DE LAS PENSIONES DE JUBILACIÓN PÚBLICAS CON UN FONDO DE CAPITAL

VI.1.- CARACTERÍSTICAS GENERALES DE LA FINANCIACIÓN DE LAS PENSIONES PÚBLICAS

VI.1.1.- Problemática en la financiación de los sistemas de pensiones públicos

La problemática de la financiación de las pensiones de jubilación públicas se sitúa dentro del contexto más general en el que se enmarca el sistema de protección pública, hecho éste que es común a la mayoría de los países desarrollados. Hace unos años el problema fundamental era determinar el ámbito de cobertura del sistema. La generalización del Estado del Bienestar se tradujo en todos los países desarrollados en un efecto protector que abarcaba cada vez a una parte creciente de la población y a una lista también creciente de prestaciones sociales. Este proceso de generalización se enfrenta con varios problemas, uno de los cuales es el de financiación del Estado del Bienestar y más concretamente de la Seguridad Social (S.S.) como organismo que atiende gran parte de las prestaciones sociales.

El desequilibrio financiero tiene su origen principal en el aumento del gasto derivado del aumento en la protección y se ve agravado por las tendencias demográficas que disminuyen año a año el ratio activos-pasivos. Factores coyunturales de recesión económica contribuyen todavía más a aumentar el desequilibrio debido a las elevadas tasas de desempleo, afectando negativamente tanto en la parte de ingresos como de gastos. Algunos cálculos realizados bajo hipótesis simples en nuestro país prevén que en el año 2000 el déficit del actual sistema supondría el 1,7% del PIB (avance del estudio de la Fundación BBV).

El problema de la financiación se superpone con la necesidad de armonizar los sistemas de S.S. en los países de la Unión Europea. La convergencia en esta materia es necesaria como respuesta a otras medidas como la libre circulación de trabajadores y también para impedir que la existencia de diferentes sistemas se traduzca en ventajas comparativas para los países con menor protección. Pero esta convergencia se debe conseguir sin ampliar las divergencias en otros campos, como

por ejemplo, aumentando los déficits públicos. De todas formas, las recomendaciones del Consejo de las Comunidades Europeas en esta materia (Recomendaciones 92/441/CEE y 92/442/CEE) tienen en cuenta la autonomía y el respeto a las peculiaridades del sistema instaurado en cada país.

El ámbito de la protección pública abarca una lista amplia de prestaciones: cobertura sanitaria, prestaciones por desempleo, pensiones, servicios sociales, etc.. Aunque parte de la problemática es común a todas ellas, también hay características específicas que hace difícil un análisis común en términos matemáticos, como se pretende. Por eso la aplicación que aquí se desarrolla se centra en una parte concreta del sistema como son las pensiones de jubilación. Es en este campo donde el cambio demográfico afecta de una forma más clara, y éste será uno de los elementos que formará parte de la modelización.

En el estudio de las pensiones de jubilación públicas, el aspecto de la financiación no es el único ni, posiblemente, el más importante a tener en cuenta. Junto con el análisis de los métodos de financiación, cabe destacar el de los efectos económicos de estas pensiones y las decisiones óptimas de jubilación voluntaria anticipada. Las técnicas utilizadas en estos estudios también han sido distintas: modelos teóricos de equilibrio de una economía, modelos econométricos, simulaciones, métodos de optimización, etc..

La teoría del Control Óptimo discreto también puede jugar un papel como técnica matemática útil en el estudio de alguna parte del problema. Hemos elegido la parte de la financiación por ser un tema más afín con nuestra línea de investigación. Se pretende establecer un sistema de financiación que sea estable a largo plazo teniendo en cuenta el efecto demográfico. Para ello se incorpora un fondo de capital como fuente de financiación adicional, que permite distribuir entre más generaciones el efecto del cambio demográfico. En Durán (1986) se realiza también un estudio de equilibrio financiero teniendo en cuenta la evolución demográfica (con parámetros estables) de un grupo de individuos representativos pero sin utilizar métodos de optimización.

Para enmarcar el problema se parte de los esquemas más habituales que se pueden plantear para cubrir las pensiones de jubilación. Una distinción aceptada es aquélla que diferencia tres pilares: los regímenes públicos obligatorios, cuya vocación es garantizar un mínimo básico; los regímenes profesionales, que van ligados al desarrollo de un trabajo; y el ahorro individual y voluntario.

A partir de esta clasificación, la tipología concreta de regímenes de jubilación difiere bastante entre países, sobre todo en lo que se refiere al régimen profesional. Estas diferencias no siempre son fáciles de detectar ya que, como se señala en Roberts (1993), la terminología que se usa para unificar conceptos puede llevar a confusiones al encerrar un mismo término situaciones distintas.

Lo más habitual es que el régimen público obligatorio siga un esquema de reparto y el ahorro individual y voluntario uno de capitalización. El régimen profesional soporta más diferencias entre los países ya que puede gestionarse de forma pública (caso español), privada (regímenes complementarios franceses) o a elección del trabajador (caso británico). A su vez, el régimen profesional, sea cual sea su forma, se puede basar en el reparto o en la capitalización.

El sistema español de pensiones se compone de un sistema público obligatorio en régimen de reparto y de un sistema privado voluntario en régimen de capitalización. El sistema público incluye tanto la parte básica (pensiones no contributivas y complementos a mínimos de las pensiones contributivas) como la parte profesional (pensiones contributivas no afectadas por complementos a mínimos). Nuestro campo de estudio se centra en el sistema de financiación de las pensiones contributivas.

El **sistema de reparto** consiste en que los trabajadores y empresarios de un periodo aportan una parte de sus rentas, las cuales, junto a las aportaciones del Estado con cargo a los presupuestos, financian las pensiones de los jubilados de ese periodo. Bajo este sistema no hay creación de un fondo de capital puesto que las aportaciones igualan a las pensiones en cada periodo. Para que el método de reparto

sea estable en el tiempo debe haber, a su vez, una cierta estabilidad en la relación entre cotizantes y pensionistas y una cierta igualdad entre las revalorizaciones de las pensiones y los aumentos en las bases de cotización de empresarios y trabajadores (salarios). Estas condiciones de estabilidad se pueden romper por factores socio-económicos (coyunturales) como los cambios en la tasa de actividad o en la tasa de paro o por factores demográficos (de tipo estructural) como el envejecimiento de la población. Los factores coyunturales se solucionan ajustando las aportaciones del Estado en función del nivel de actividad económica o recurriendo a préstamos, pero los factores estructurales necesitan reformas más de fondo. Este fenómeno demográfico es común a muchos países desarrollados y la necesidad de resolverlo, junto con el proceso de armonización en el seno de la Unión Europea, es una de las principales causas que impulsan proyectos de reforma de la S.S..

El **sistema de capitalización** consiste, al igual que ocurre en los fondos de pensiones privados, en la equivalencia financiero-actuarial entre las aportaciones del periodo activo y las pensiones del periodo pasivo para una misma persona. Bajo este sistema se acumula un fondo de capital cuyos rendimientos determinan las pensiones. A nivel financiero, la teoría argumenta un enunciado muy simple: "en situaciones estacionarias el esquema de capitalización es mejor que el de reparto si la tasa de rendimiento del capital es mayor que la tasa de crecimiento de la masa salarial (base total de cotización) y es peor en caso contrario". Evidentemente, es posible un sistema mixto en el que exista un esquema de reparto (por ejemplo los cotizantes de un periodo pagan la pensión mínima de los jubilados de ese periodo) y uno de capitalización (los cotizantes pagan otra cantidad para financiar la parte de la pensión que exceda del mínimo cuando se jubilen).

En la situación actual y teniendo en cuenta la tendencia demográfica, la teoría parece aconsejar el paso a un sistema de capitalización, que es neutral respecto al cambio demográfico. Efectivamente, la rentabilidad del capital en términos reales en España, incluso descontando gastos de gestión, ha sido netamente superior al aumento de los salarios reales en el periodo 1987-1994 y si esta tendencia se mantiene en el futuro, junto con una población cotizante prácticamente estancada,

el resultado es una ventaja comparativa para la capitalización frente al reparto. Este fenómeno es común, a grandes rasgos, al resto de países de nuestro entorno. Sin embargo, las reformas planteadas en la mayoría de estos países no abandonan el sistema de reparto sino que simplemente lo adaptan. Se observa, por tanto, un alejamiento entre lo que aconseja la teoría y la práctica. En el fondo, lo que ocurre es que el enunciado teórico es excesivamente simple y el sistema de capitalización no está exento de problemas:

- Se puede dar una disminución de la rentabilidad del capital por debajo del nivel óptimo en la economía a medida que crece el volumen acumulado (exceso de capitalización). De esta manera, se anularía la ventaja de la capitalización sobre el reparto. Esto es posiblemente lo que ocurriría bajo un sistema de capitalización total. Este efecto se notaría más si la generación de activos fuera numerosa por lo que el sistema de capitalización no es tan neutral respecto al cambio demográfico como se dice.

- La rentabilidad del capital está sujeta a incertidumbre. En la medida que esta incertidumbre sea superior a la del sistema de reparto (derivada de una evolución de la población ocupada distinta a la esperada) se estará aumentando la incertidumbre de las pensiones futuras. Este peligro depende del desarrollo del mercado financiero y del grado de estabilidad interno y externo dada la gran sensibilidad de este tipo de mercados.

- No se pueden introducir elementos redistributivos que sí que están presentes en un sistema de reparto. Si acaso existe una redistribución de los que viven menos años hacia los que viven más años. Aunque el sistema actual de reparto (al menos en España) ofrece dudas sobre si resulta progresivo o regresivo, es innegable que el reparto permite incluir elementos redistributivos si se desea.

- Existen problemas añadidos en la transición de un sistema de reparto a uno de capitalización. El paso de un sistema a otro sólo se puede llevar a cabo

con la intervención del Estado, dados los elevados costes durante la etapa transitoria. Si además el país tiene problemas adicionales de endeudamiento o tiene otras prioridades de gasto (infraestructuras, reconversiones industriales, etc.) el cambio de sistema es prohibitivo ya que en la Unión Europea el recurso a un aumento del déficit es contrario a los criterios de convergencia[61].

- Por último, existen problemas de tipo político que explican la renuncia general a adoptar el sistema de capitalización (teoría de la elección pública[62]). En primer lugar, si el fondo de capital se gestiona a nivel público se corre el riesgo de utilizarlo para otros fines, como la financiación de déficits o la expansión del gasto para ganar elecciones, con el riesgo de no satisfacer las pensiones futuras. En segundo lugar, y más importante, aparecen los riesgos políticos de la transición. Efectivamente, los pensionistas actuales, los activos que están próximos a la jubilación y los activos de rentas bajas difícilmente darán su apoyo político a una reforma de este tipo.

Como se desprende de este análisis preliminar, la solución a toda la problemática es compleja y debe combinar distintas alternativas. Las que más eco han tenido son las que van encaminadas a adaptar el sistema de reparto. Sin embargo, también se oyen voces que plantean el cambio a un sistema de capitalización o mixto.

En este contexto se enmarca la propuesta de formación de un fondo de capital (sin capitalización), que puede aliviar los efectos del cambio demográfico sin que se renuncie a los elementos redistributivos, aunque su cuantía debe ser limitada (sistema mixto) para que no aparezca, de forma significativa, el resto de los efectos negativos antes comentados.

---

[61]El avance del informe sobre pensiones de la Fundación BBV cifra en 2,5 veces el P.I.B. el coste del paso brusco a un sistema de capitalización. Si dicho sistema se implanta sólo para los que entran en el mercado laboral, el coste sería del 5% del P.I.B. hasta su maduración.

[62]La influencia de la elección pública sobre las pensiones puede consultarse en Verbon (1993).

VI.1.2.- Propuestas de reforma del esquema de financiación

Las propuestas de reforma son muy variadas. La mayoría no se plantean un cambio de esquema sino una simple reestructuración del de reparto. En el otro extremo hay propuestas de abandono del sistema por uno de capitalización total (privado o público) o por uno no obligatorio. En el medio, hay propuestas que prefieren una situación mixta, que se caracteriza por la existencia de una parte básica, que se financia con el sistema de reparto, y una parte complementaria, que lo hace a través de un sistema de capitalización.

El paradigma del sistema de capitalización es Chile. A principios de los años 80 se instauró la reforma de la S.S. que supuso la privatización del sistema de pensiones de jubilación. Los activos nuevos están obligados a abrir una cuenta de ahorro-pensión en una institución privada (el sistema era voluntario para los trabajadores activos antes de la reforma). Estas instituciones invierten el capital y el individuo lo recupera en forma de pensión periódica tras la jubilación. Así pues, el sistema privado de capitalización sustituye completamente al público de reparto.

El sistema mixto es el existente en Reino Unido. Allí existe una pensión básica universal que se financia con las aportaciones a la S.S.. Existe otra parte obligatoria que puede ser pública si todas las cotizaciones se pagan a la S.S. (da lugar a los llamados SERPs o State Earnings Related Pension) o privada si una parte se paga a una entidad privada (como en Chile o a través de un plan de pensiones de empresa). Por último, existe una tercera parte voluntaria a través de un plan de pensiones privado. En este sistema mixto, la capitalización sustituye y complementa al reparto.

Los ejemplos de sistema de reparto son los más abundantes, entre ellos el sistema español. Las pensiones de jubilación son públicas y siguen el esquema de reparto. Los planes de pensiones privados son voluntarios y por tanto no sustituyen al sistema público sino que sólo lo complementan.

Las opiniones a favor y en contra del sistema de reparto se suceden continuamente, tanto dentro como fuera de nuestro país. Entre ellas cabe destacar la del Premio Nobel de Economía de 1992, Gary S. Becker, quien pronostica una quiebra en el sistema de reparto argumentando razones demográficas. Es partidario de un sistema privado de capitalización como el de Chile excepto para las pensiones no contributivas que deben correr a cargo del Estado. El periodo transitorio sería necesario para evitar un coste excesivo y sería voluntario para los actuales trabajadores activos, que optarían entre seguir en el sistema antiguo u obtener una especie de bono por las cotizaciones realizadas hasta ahora e invertirlas en un fondo de pensiones privado. Esta opinión coincide básicamente con la aportada por el F.M.I. en la reunión anual de Madrid de 1994. Entre los partidarios de seguir con el reparto se destaca el coste excesivo que supondría el paso a un sistema de capitalización (entre 2 y 2,5 veces el PIB), siendo preferible reformar el sistema actual con medidas técnicas que lo hagan viable en el futuro. Esta última opción es la que actualmente sostienen en nuestro país los partidos políticos (Pacto de Toledo) y los sindicatos. En la misma línea parecen ir algunos trabajos recientes: Barea (1995), el avance del estudio de la fundación BBV (cuya versión final está prevista para Enero de 1996) y el estudio de la Dirección General de Planificación y Ordenación Económica de la Seguridad Social.

La realidad es que, aunque las propuestas de reforma del sistema de reparto para transformarlo en un sistema de capitalización o mixto están bien fundamentadas teóricamente, son pocos los países que parecen estar dispuestos a llevarla a cabo. Los problemas antes comentados del sistema de capitalización y el coste político de la transición explican esta divergencia entre teoría y práctica[63]. Sin embargo, también es verdad que se asiste a una potenciación de los planes de pensiones privados ante el peligro de que el mantenimiento del sistema de reparto sólo se pueda llevar a cabo a través de menores pensiones en el futuro. Como propuesta intermedia está la de crear un fondo de capital pero sin el esquema típico del sistema

[63]Sobre los límites políticos de las reformas del sistema público de pensiones se puede leer un interesante artículo de opinión de J.L. Oller (Expansión, 19-V-95).

de capitalización.

Haciendo un breve repaso de las reformas recientes[64] en el marco de un sistema de reparto en el caso español, cabe citar la **Ley 26/1985 de racionalización de la estructura y de la acción protectora de la S.S.** como primer intento de envergadura de garantizar la protección social en el marco de un equilibrio financiero estable. Las medidas adoptadas tienen como objetivo frenar el crecimiento en el gasto, corregir algunos defectos en la protección y racionalizar la estructura del sistema. Las medidas concretas que se adoptaron fueron la revalorización automática de las pensiones, ampliar el periodo de carencia para tener derecho a una pensión de 10 a 15 años, establecer un nuevo método de cálculo de la base reguladora de las pensiones que tuviera en cuenta 8 años en vez de 2, redistribuir los recursos destinados a la protección familiar y englobar a ciertos regímenes especiales en el régimen general.

La ampliación del periodo de carencia y de los años de cálculo de la base reguladora refuerzan el carácter contributivo de las pensiones, lo cual ayuda a la reducción del "fraude" (disminuye el fenómeno de compra de pensiones) y a contener el gasto en pensiones ya que, en la práctica, la pensión media de los nuevos jubilados sobre la base media de cotización disminuyó con la reforma como muestra la tabla VI.1.

[64]Una visión más histórica puede consultarse en Gómez Sala (1994).

	Pensión media de altas / base media mensual
1985	67,86%
1989	55,00%
1993	58,45%

Tabla VI.1.: Pensión media mensual de las altas por base media mensual de cotización (Régimen general). Elaboración propia. Fuente: Proyecto de Presupuestos de la S.S. para 1994.[65]

Otro dato que demuestra el efecto de la reforma sobre las nuevas pensiones es la disminución que se produjo en la pensión media mensual de jubilación de las altas en el Régimen General en el año 86 respecto al 85 (-1,11%).

La siguiente reforma que merece destacarse es la llamada **reforma financiera de la S.S.** (se inicia en 1989), que consiste en asignar un destino a los recursos del sistema. Las cotizaciones y aportaciones del Estado que hasta entonces tenían un destino global pasan a tener cada vez más un carácter finalista. Este es el inicio de una mejora en la estructura financiera de la S.S.. A grandes rasgos, el objetivo es que los gastos contributivos se financien con cotizaciones y los gastos no contributivos con aportaciones del Estado. La parte con cargo al Estado ha ido creciendo continuamente en términos relativos incluso antes de 1989 y lo ha seguido haciendo después, avanzando así hacia una mayor racionalización de la estructura financiera. La aparición de pensiones no contributivas en 1991 (Ley 26/1990) y la universalización de la Sanidad exigió profundizar en el carácter finalista de los recursos para que el aumento del gasto no repercutiera en las cotizaciones. Hay que

[65]El dato de 1993 corresponde a una estimación para la base media y a la pensión media hasta Junio.

decir que en la actualidad la parte contributiva sigue teniendo superávit pero se está avanzando en el equilibrio presupuestario por separado de la parte contributiva y la no contributiva. Esto se puede observar en la tabla VI.2., donde se recogen las cifras de los tres últimos presupuestos y el porcentaje de alivio que se podría introducir en las cotizaciones a cambio del correspondiente aumento impositivo. La parte contributiva recoge cotizaciones sociales de trabajadores y empresarios en la parte de ingresos y prestaciones contributivas en la parte de gastos. La parte no contributiva incluye como ingresos las transferencias del Estado y como gastos las prestaciones no contributivas, los complementos a mínimos[66], la asistencia sanitaria y los servicios sociales. El resto de ingresos y gastos se han repartido proporcionalmente dentro de cada programa o capítulo.

	Superávit parte contributiva (mill. ptas.)	Superávit/ total cuotas
1993	916.074	14,50%
1994	823.826	12,85%
1995	624.933	9,43%

Tabla VI.2.: Superávit de la parte contributiva=Déficit de la parte no contributiva. Elaboración propia. Datos del Proyecto de presupuestos de la S.S. para 1994 y 1995.

En el año 95 se ha avanzado en la racionalización financiera por medio de una sustitución de cuotas por impuestos indirectos, lo que explica gran parte del descenso en el superávit de la parte contributiva previsto para este año. En la medida en que el superávit de la parte contributiva supone un alivio para el sistema

[66]El porcentaje que representan los complementos a mínimos dentro del total de pensiones se ha supuesto en un 10,3% en 1994 y en un 10% en 1995, tras observar su evolución desde 1983.

impositivo, se puede seguir avanzando en esta sustitución, sin embargo también hay que considerar la alternativa que se plantea en nuestra aplicación, es decir, la utilización de una parte del superávit para la formación de un fondo de capital que evite aumentos en la cotización (o disminuciones en las pensiones) cuando la demografía sea desfavorable.

Las medidas de reforma para financiar el sistema de reparto que se han dado en otros países de la Unión Europea se recogen en la monografía sobre la protección social en Europa en 1993 (Comisión de las comunidades europeas, 1994) y se pueden resumir en las siguientes:

- Por la parte de los ingresos, se ha recurrido en ocasiones al aumento de las cotizaciones, a la eliminación de los topes de las bases de cotización, a la creación de cotizaciones o impuestos específicos para financiar algún aspecto de la protección social y a la sustitución de cotizaciones sociales por alguna figura del sistema general de impuestos para favorecer el empleo.

- Por la parte del gasto, se han recortado prestaciones de forma directa (reducción de la cuantía) o indirecta (endurecimiento de los requisitos para tener derecho a ellas). También se suspendió ocasionalmente en algunos países la revalorización automática de las prestaciones.

- Una reforma más de fondo consiste en reducir el ámbito de aplicación del sistema de reparto. Esto se dio en el Reino Unido al permitir que la pensión de jubilación, en la parte que supera la pensión básica, se pueda contratar en el sector privado y por tanto bajo el sistema de capitalización. Otros países, sin llegar a esta privatización, han potenciado las planes de pensiones privados.

Además de estas medidas y de aquéllas que no son objeto de controversia (reducción del desempleo, aumento de la tasa de actividad, lucha contra el fraude, mejora de la gestión, etc.) existen otras propuestas que son objeto de estudio y que

aparecen en las conclusiones del pacto de Toledo sobre pensiones (Vicente, 1995):

- Retrasar la edad efectiva de jubilación, promoviendo fórmulas flexibles de jubilación, favoreciendo trabajos más adecuados para las personas mayores de 65 años, compatibilizando la jubilación y el trabajo a tiempo parcial, etc.

- Completar la reforma de la estructura financiera aumentando las aportaciones del Estado y, posiblemente, modificar la relación existente entre cotizaciones de trabajadores y empresarios. El acercamiento a la estructura de ingresos media de la U.E. supondría para España aumentar la parte a cargo del Estado y de los trabajadores y disminuir la parte a cargo de empresarios.

- Aumentar la vinculación entre la cuantía de la pensión y las cotizaciones pagadas durante la etapa activa (tener en cuenta más de los 8 años actuales para calcular la base reguladora, aumentar el periodo de carencia y el número de años para acceder al 100% de la base).

- Avanzar en la simplificación de regímenes hasta pasar a solamente dos: trabajadores por cuenta ajena y por cuenta propia.

- Crear reservas de nivelación que permita acumular fondos en años de expansión económica y liberarlos en años de recesión. Esta misma política se puede utilizar para el caso de expansión y recesión demográfica[67].

Para analizar correctamente las distintas medidas que se proponen se han de tener en cuenta los efectos, tanto financieros como económicos, del cambio demográfico así como tener una idea de los efectos generales de la protección social

[67]La posibilidad de la creación de reservas de nivelación se recoge en Jiménez (1994) y es una de las conclusiones del Pacto de Toledo, siempre referida al caso de la coyuntura económica. La ampliación para englobar el efecto del cambio demográfico es uno de los objetivos de la aplicación que se desarrolla en los siguientes epígrafes.

y su financiación sobre las variables económicas.

Los efectos del cambio demográfico en su aspecto financiero han sido estudiados en el caso español por Barea y Fernández (1994) llegando a cifrar en un 50% el aumento en términos reales de la carga a financiar por habitante en el capítulo de pensiones (un 30% si se tiene en cuenta el conjunto de la protección social). Otros trabajos relacionan el cambio demográfico con la Seguridad Social no sólo en su aspecto financiero. Así Schmähl (1990) describe la complejidad de estas relaciones y la dificultad en su análisis. Autores como Peters (1991), Boadway y otros (1991), López García (1994) y Artus (1994) realizan análisis teóricos para observar los cambios en variables como el ahorro, salarios, tipos de interés, etc. para maximizar la función de bienestar social ante el tránsito demográfico. Por último, el paso de los modelos teóricos a las simulaciones se da en Auerbach y Kotlikoff (1985), Blanchet y Kessler (1991), López García (1992) y Montalvo y Quesada (1994); llegando a conclusiones no siempre coincidentes: por ejemplo, Montalvo y Quesada prevén una disminución del bienestar (consumo per cápita) mientras que Auerbach y Kotlikoff anuncian un aumento del bienestar (medido en gasto de un adulto en consumo y ocio).

Para analizar las propuestas de reforma, también hay que tener en cuenta los efectos económicos de la protección social o de su financiación. Existen abundantes estudios al respecto, aunque sus conclusiones no siempre son compatibles entre sí. A pesar de ello, todos estos estudios son una guía útil para orientar las medidas de reforma de la S.S. y para tomar medidas complementarias que potencien los efectos positivos de dichas reformas.

Las técnicas utilizadas en estos estudios son variadas: estudios teóricos de equilibrio general, técnicas econométricas, simulaciones, estudios comparativos entre países, etc.. Por otra parte, también hay diversidad en cuanto al tipo de efectos económicos estudiados: efectos de las cotizaciones e impuestos sobre la oferta y demanda de trabajo, sobre el déficit público, sobre la competitividad de un país, efectos de las prestaciones sobre la oferta de trabajo y sobre las variables

económicas de una economía (salarios, tipos de interés, bienestar, etc.).

Gran parte de los modelos que se manejan en los estudios de tipo teórico parten del modelo de generaciones superpuestas (*overlapping generations model*), desarrollado inicialmente por Samuelson (1958) y Diamond (1965), aunque estos autores no introdujeron la Seguridad Social en el análisis. Este modelo se utilizó posteriormente para ver los efectos macroeconómicos de la Seguridad Social en Feldstein (1974), Samuelson (1975), Burbidge (1983) y Craig y Batina (1991) entre otros. En España lo han utilizado autores como López García (1986, 1991) para estudiar los efectos de los métodos de financiación y Hercé (1986) para ver cómo afecta la Seguridad Social a la oferta de trabajo y capital.

También existen estudios que enmarcan los efectos de la S.S. en un modelo de ciclo vital de consumo-ahorro. Este modelo prevé que aumentos en las prestaciones provocan mayor consumo al aumentar la renta disponible a lo largo del ciclo vital. Sin embargo, los individuos pueden pensar que, tarde o temprano, esto irá seguido de un aumento de impuestos con lo cual no habrá efectos sobre la alternativa ahorro-consumo (equivalencia ricardiana). Estos resultados teóricos han sido contrastados a nivel empírico con conclusiones desiguales (ver, por ejemplo, Diamond y Hausman, 1984 y Wilcox, 1989).

En cuanto a trabajos aplicados al caso español cabe señalar, en primer lugar, los referidos a los efectos de una reducción en las cotizaciones empresariales sobre la demanda de trabajo, por medio de modelos econométricos (Escobedo, 1991) o por medio de modelos de equilibrio general (Polo y Sancho, 1990). Y, en segundo lugar, el estudio de la influencia de las pensiones sobre la oferta de trabajo (Martín y Moreno, 1990; López García, 1986, 1990; y Hercé, 1986).

Algunos resultados interesantes de estos trabajos son:

- La reducción de las cuotas empresariales a la S.S. a cambio de un aumento en la financiación del Estado (vía IVA o IRPF) o en las cotizaciones de

trabajadores persigue el objetivo de reducir la tasa de paro aumentando la cantidad demandada de trabajo, lo cual beneficiaría al sistema al aumentar el número de cotizantes. Sin embargo, el aumento de las cotizaciones de trabajadores se tendería a compensar con mayores salarios nominales para evitar una merma en su poder adquisitivo, lo que, a su vez, podría anular el efecto de la reducción de las cuotas de empresarios y no generarse nuevo empleo. Por otra parte, el aumento de las aportaciones del Estado trae consigo el problema del aumento de impuestos. Se argumenta que el IVA es el más neutral sobre el trabajo pero al ser un impuesto indirecto se estarían pagando los gastos sociales de forma proporcional y no progresiva, además de sus efectos inflacionistas. La alternativa de un aumento del IRPF encontraría resistencias generalizadas. El trabajo econométrico de Escobedo (1991) recoge un efecto total escaso sobre la demanda de trabajo debido a la traslación de la reducción de las cuotas empresariales a los salarios reales. Creemos, sin embargo, que la situación del mercado laboral en los años de análisis del anterior estudio (1975-1983) es bastante distinta del actual donde han disminuido ciertas rigideces y la reivindicación salarial ha pasado a un segundo término. El modelo de equilibrio general analizado en Polo y Sancho (1990) da resultados más optimistas por cuanto, a no ser que la traslación a los salarios reales sea total, se produce una disminución en la tasa de desempleo y un aumento en la inversión. El aumento en la aportación del Estado a la S.S. puede provocar un aumento del déficit público aunque éste se ve atenuado por el aumento en la imposición directa que se produce al aumentar el nivel de actividad. En Barea y otros (1995) se recogen las principales conclusiones de varios estudios referidos a este tema, desde el pionero de Zabalza de 1987 que no deducía efectos positivos relevantes hasta otros más recientes que pronostican un efecto positivo de las reducciones en las cuotas empresariales y su sustitución por imposición indirecta.

- Aumentar la relación entre las pensiones recibidas y las cotizaciones pagadas pretende disminuir las perturbaciones sobre la oferta de trabajo inherentes al sistema de S.S.. Efectivamente, la S.S. puede afectar a la

decisión de trabajo y de retiro anticipado ya que redistribuye la renta de los trabajadores a lo largo de su ciclo vital, gravándole en los periodos activos (cotizaciones) y subvencionándole en los pasivos (pensiones). En un sistema de capitalización puro no se producen estas alteraciones porque los beneficios futuros compensan exactamente las cotizaciones pagadas. Este efecto se ha traducido, sobre todo antes de la reforma de 1985, en el fenómeno de compra de pensiones: los trabajadores estaban dispuestos a cotizar más en los dos últimos periodos porque eran los que servían de base al cálculo de la pensión. Estas perturbaciones están ampliamente fundamentadas en la literatura. En el caso español se puede citar a nivel teórico a Hercé (1986) que analiza el efecto de sustitución intertemporal en la oferta de trabajo; y a nivel empírico a Martín y Moreno (1990), donde se destaca un pequeño efecto positivo del sistema de S.S. sobre la oferta de trabajo en el periodo 1964-1984. Por tanto, una medida que vinculara más las pensiones a las cotizaciones tendería a reducir las distorsiones intertemporales en la oferta de trabajo aunque el efecto global presumiblemente sería de una pequeña reducción en la oferta de trabajo.

- Otros efectos económicos son más difíciles de estudiar a nivel teórico y se acude a la comparación internacional; esto ocurre, por ejemplo, en el estudio de los efectos de la protección social sobre la competitividad y el empleo de un país. El enunciado teórico que se argumenta es que la protección social y su financiación a través de cotizaciones es un elemento que resta competitividad a un país. Este enunciado parece cumplirse a simple vista en la comparación EEUU-Japón-Europa (Izquierdo y Cuevas, 1994) aunque haría falta un estudio más profundo que tuviera en cuenta otros factores. De hecho, en el ámbito europeo, no se llega a esa conclusión. Efectivamente, en la monografía de la Comisión de las Comunidades Europeas dedicada a la protección social en Europa en 1993, se recoge una correlación positiva evidente entre gastos en protección social por habitante y coste de la mano de obra por trabajador; sin embargo, dicha correlación no se traslada al nivel de empleo o de competitividad (medida a través de las exportaciones). Así

pues, no se confirma la hipótesis de una influencia negativa de la protección social sobre el empleo y la competitividad de un país. La conclusión a la que se llega en dicha monografía es que el crecimiento en los gastos en protección social debe ir asociado a un crecimiento económico y, en ese caso, no se producirán distorsiones. Sin embargo, si se producen cambios que alteren el equilibrio entre protección social y desarrollo económico (como por ejemplo, los cambios demográficos sin aumentos paralelos en la productividad) habrá que replantear el equilibrio a través de menores prestaciones o de menores salarios y beneficios para recuperar la senda de crecimiento.

VI.1.3.- Financiación a través de fondos de capital

Existen otras propuestas más radicales de reforma del esquema de financiación del sistema de pensiones. Nos estamos refiriendo a las que propugnan un mayor protagonismo de los rendimientos del capital en la financiación del sistema de pensiones. Estas propuestas tienen en común la constitución de un fondo de capital cuyos rendimientos pasan a formar parte de los ingresos del sistema. Sin embargo, difieren en otros aspectos ya que este fondo de capital puede ser privado o público. En el primer caso, se trata de privatizar una parte del sistema de pensiones (en el Reino Unido existe esta posibilidad para la parte complementaria de las pensiones de jubilación) y supondría el paso a un sistema de capitalización. En el segundo caso, no hay privatización y el sistema puede ser de capitalización o no.

En el ámbito privado se enmarca la Ley de Planes y Fondos de Pensiones de 1988, cuyo objetivo es fomentar el ahorro y complementar las pensiones públicas[68]. Este tipo de instrumento, sin embargo, tiene un discutible efecto potenciador del ahorro (a no ser que existan considerables ventajas fiscales), ya que se produce más

[68]Aunque en el Capítulo V se pudo obtener que existen otros instrumentos financieros que pueden ser más adecuados como fórmula de ahorro-pensión para un individuo.

una sustitución por otro tipo de ahorro que una creación de ahorro neto. Este parece ser el caso español ya que el volumen acumulado es modesto (aproximadamente 1,5 billones de ptas. a finales de 1994), lo que hace prever medidas de potenciación, sobre todo de tipo fiscal y de aumento en su liquidez.

La sustitución total del sistema público por el privado se ha argumentado en ocasiones, aunque es previsible que diera lugar a una tasa de ahorro y una capitalización por encima de las óptimas. Además, para garantizar las pensiones de los actuales pensionistas y de los activos a punto de jubilarse se tendría que acudir a su provisión pública (aumento de impuestos o del endeudamiento). Entre los partidarios de este cambio de sistema se puede citar a José Piñera (ex-ministro de Trabajo de Chile que implantó la reforma en ese país) y en cuyo libro se incluye un prólogo de Pedro Schwartz en el que se dan algunas recomendaciones para pasar a un sistema privado de capitalización en España (Piñera, 1995). El mismo autor del libro está actualmente preparando un informe completo sobre cómo llevar a cabo la reforma del sistema de pensiones en España.

Las propuestas más cercanas a la aplicación que se va a desarrollar persiguen un sistema mixto. Puede ser mixto en el sentido de que tiene una forma de reparto (por ejemplo la financiación de una pensión mínima) y una de capitalización (lo que excede de la pensión mínima) o puede ser mixto en el sentido más general de que una parte de los ingresos provienen de un fondo de capital que no se ha constituido necesariamente bajo un esquema de capitalización.

Teniendo en cuenta aspectos puramente financieros, la capacidad de generar recursos de un fondo de capital depende de los tipos de interés y la capacidad de un esquema de reparto depende de la expansión de la masa salarial de cotización. Aproximadamente, lo anterior equivale a decir que la capacidad financiera del esquema de reparto es mejor si la población cotizante crece a una tasa superior al tipo de interés real y peor en caso contrario. La existencia de una tendencia demográfica regresiva, factores coyunturales de recesión económica y tipos de interés reales positivos, han llevado a afirmar que, en la situación actual, es más

rentable el sistema con fondos de capital que el de reparto.

El anterior razonamiento es válido bajo dos consideraciones importantes: sólo se refiere al aspecto financiero y sólo se comparan sistemas ya maduros. No se tienen en cuenta todavía los efectos de tipo macroeconómico ni las consecuencias que tendría el paso del sistema de reparto actual al sistema con fondo de capital (establecimiento de periodos transitorios, transferencias intergeneracionales, etc.). Estas consideraciones se recogen de una forma descriptiva en Disney (1994).

La vía habitual para tratar los efectos económicos de los métodos de financiación es, de nuevo, el modelo de generaciones superpuestas. Bajo los supuestos restrictivos de este tipo de modelos se demuestra que la tasa de interés tiende a ser menor y la de salarios mayor bajo un esquema con fondos de capital que bajo un esquema de reparto. López García (1986) además de reflejar estas consideraciones estudia a nivel teórico lo que puede ocurrir en el paso de un sistema a otro. Una de sus principales conclusiones es la aparición de paradojas del siguiente tipo: para llegar a una situación estacionaria superior a otra (en el sentido de Pareto) es posible que algunas generaciones intermedias tengan que salir perdiendo durante el periodo transitorio. De esta manera, además del problema de obtener el mejor esquema de financiación de pensiones (la mejor situación estacionaria), aparece el problema de llegar a ese esquema de la mejor manera posible (la mejor transición).

A las anteriores consideraciones hay que sumarle las de tipo político. La formación de un fondo de capital exige costes transitorios y la manera de financiarlos hay que evaluarla en términos políticos. El recurso al endeudamiento significa trasladar la carga a las generaciones futuras además de poner en peligro el objetivo de déficit público; mientras que el aumento de impuestos, de cotizaciones o la disminución de pensiones lleva consigo costes políticos que sólo se pueden minimizar si se parte de un amplio consenso en esta materia.

La creación de un fondo de capital se puede plantear si la rentabilidad de la capitalización es superior a la del reparto (tipo de interés real superior al crecimiento

de la población cotizante) y si la tasa de ahorro de partida de la economía es inferior a la óptima. Dados ambos supuestos, su cuantía debería ser moderada para evitar una caída en su rentabilidad real y el periodo transitorio suficientemente dilatado para reducir la dureza del ajuste para las generaciones intermedias. La distribución de la carga es una cuestión a resolver, lo cual no es fácil. Hay que determinar:

- Sobre qué colectivo recae: pueden ser los pensionistas (rebajando sus pensiones), los empresarios y trabajadores (subiendo sus cotizaciones respectivas) o todos los contribuyentes (subida de impuestos).

- Sobre cuantas generaciones se reparte: establecimiento de un periodo transitorio adecuado.

Situados en este punto llega el momento de discutir el papel que puede jugar la optimización matemática y más concretamente la teoría del C.O. discreto en la formación del fondo de capital, es decir, en el paso de un sistema de reparto a uno mixto.

La optimización matemática se utiliza en el modelo teórico de generaciones superpuestas pero de forma estática. Los ejercicios de estática comparativa que se realizan a partir de este tipo de modelos orientan los efectos de ciertos cambios sobre las variables del modelo y sirven para fundamentar teóricamente las ventajas de unos esquemas sobre otros. La programación lineal se ha utilizado en un marco simple de combinación de esquemas de financiación (Enríquez de Salamanca, 1986). Cuando se plantea el paso de un esquema a otro, sin embargo, aparece un problema dinámico, puesto que la transición debe hacerse también de una forma óptima. En este punto la optimización dinámica puede jugar un papel destacado.

En este sentido la única referencia que conocemos en el caso español es la de Enríquez de Salamanca (1989) donde utiliza la programación dinámica y la teoría del C.O. tanto en tiempo continuo como discreto. Son modelos de aproximación cuadrática de ciertas variables a determinados valores deseados, es decir, estudian

las trayectorias que deben seguir las variables de control (porcentaje de cotización, tasa de reemplazo de pensiones, tipo de revalorización, etc.) para que se aproximen lo más posible a ciertos valores deseados, que se determinan según ciertas hipótesis acerca de la evolución demográfica, incremento de la pensión media, incremento de la cotización media, incremento de las reservas (fondo de capital), etc.. Este modelo, por tanto, parte de una evolución del fondo de capital deseado y esto supone una limitación porque su volumen queda mediatizado por estos valores deseados. Por ello, los modelos que desarrollamos aquí suponen un avance al someter el volumen del fondo de capital sólo a una condición terminal (primera aplicación) o dejarla como variable libre (segunda aplicación).

La **primera aplicación** (epígrafe VI.2.) tiene como variable de control el tipo de cotización y, por tanto, los pensionistas no entran a formar parte del ajuste. Partimos de algunos supuestos, entre los que destaca el tener que llegar al final del periodo transitorio con una determinada cuantía para el fondo de capital. La descripción del modelo exige la formación del fondo y la cuestión es cómo llevarla a cabo periodo a periodo, desviándose lo menos posible de un tipo de cotización deseado que será, por hipótesis, el de reparto. En esta aplicación utilizamos supuestos simplificadores sobre la evolución demográfica para llegar a soluciones analíticas.

En la **segunda aplicación** (epígrafe VI.3.) se plantea un modelo más completo. Las variables de control son el porcentaje de cotización y la pensión media expresada como una parte del salario (lo que se llama porcentaje de atribución o tasa de reemplazo de pensiones ya que éstas reemplazan al salario). Por tanto, se reparte la carga entre cotizantes y pensionistas. A su vez, el volumen del fondo de capital que resulta al final del periodo no está determinado y surge como resultado del proceso de optimización. El modelo parte de unos porcentajes de cotización y de atribución de pensiones deseados y hay que minimizar las desviaciones cuadráticas respecto a esos valores deseados. En esta segunda aplicación se utilizan supuestos más específicos para obtener la evolución de las variables exógenas con lo que la solución no es analítica sino numérica.

La cuantía del fondo se puede establecer en términos absolutos (en nuestra segunda aplicación) o en términos relativos, bien sea por trabajador, por pensionista, como una parte de la masa salarial (en nuestra primera aplicación) o de la masa de pensiones.

En los próximos epígrafes se describen las aplicaciones que tienen como objetivo determinar, mediante la teoría del C.O. en tiempo discreto, cómo se debe llevar a cabo la acumulación óptima de un fondo de capital a lo largo de un periodo. Ambos modelos son más adecuados para un sistema que no sea de capitalización pues no se vinculan las cotizaciones con las pensiones. Son a largo plazo y, por lo tanto, el número de cotizantes y de pensionistas no depende del ciclo económico sino de la estructura demográfica que, a su vez, se ha determinado de forma completamente exógena al modelo. Esto último significa que los modelos que se desarrollan asumen la hipótesis de aumento exógeno de la población, típica de los modelos neoclásicos de crecimiento. Existen, no obstante, otras líneas más recientes en el ámbito de la teoría del crecimiento económico que hacen depender la componente demográfica (especialmente la natalidad) de variables económicas o, incluso, que endogeneizan el crecimiento de la población (Becker y otros, 1990)[69].

El objetivo, en última instancia, es mostrar la utilidad de la teoría del C.O. en tiempo discreto en modelos dinámicos de optimización. Somos conscientes que las hipótesis enunciadas para la aplicación numérica del modelo, aunque las consideramos razonables, son discutibles. Esto, sin embargo, sólo afecta a los valores concretos de la solución pero no anula, a nuestro juicio, la validez del modelo y la utilización de la teoría del C.O. discreto como instrumento adecuado para el análisis de problemas dinámicos.

[69]Este tipo de modelos suelen anticipar un efecto negativo de la S.S. sobre la natalidad, aunque existen muchos efectos colaterales. El descenso en el número y/o en la calidad de los hijos se produce por varias causas: las cotizaciones aumentan el coste efectivo de los hijos, éstos dejan de ser una inversión para la vejez de los padres y, por último, bajo un sistema de reparto los padres no se benefician tanto de una inversión en hijos porque éstos ayudarán a mantener a otros ancianos (efecto de riesgo moral). Una amplia revisión de esta literatura se realiza en Ehrlich y Lui (1994).

VI.2.- MODELO DE FORMACIÓN ÓPTIMA DE UN FONDO DE CAPITAL

VI.2.1.- Supuestos y planteamiento

El objetivo que se plantea es la creación de un fondo de capital a lo largo de un periodo transitorio de forma que, al final del mismo, los rendimientos de ese fondo supongan un alivio para los cotizantes posteriores. Para ello, las generaciones del periodo transitorio deberán realizar un esfuerzo financiero extra. Dadas las tendencias de la población, este fondo permitirá nivelar, en parte, las cotizaciones de ambas etapas. El modelo resuelve el problema de cómo llevar a cabo la formación óptima de ese fondo una vez elegidos una serie de parámetros.

En este modelo inicial nos centraremos en el aspecto matemático de la resolución de un problema de C.O. discreto, por lo que se parte de supuestos simples y de un conjunto de parámetros conocidos. De esta manera, se consiguen soluciones analíticas que permiten deducir los factores que influyen en la formación del fondo así como realizar análisis de sensibilidad de los parámetros. La solución se debe aceptar con las cautelas propias de un modelo tan simplificado ya que no se consideran otros aspectos que también influirían en la solución. En el modelo que se desarrolla en el epígrafe VI.3. se destacará más el aspecto aplicado, por lo que se realizarán hipótesis más detalladas y se dedicarán amplios comentarios a la forma de llevar a la práctica la solución.

Los supuestos de este primer modelo son:

S1: Se supone que se ha tomado una decisión política de considerar los rendimientos de capital como una parte del esquema de financiación de las pensiones de jubilación. En concreto se desea que al final de un periodo transitorio el fondo de capital suponga una parte constante de la masa salarial de cotización. Esta parte se fija según un nivel deseado, pero de forma que el volumen del fondo no sea demasiado grande para que no tenga efectos distorsionadores sobre el mercado de capitales.

S2: La acumulación de capital no sigue un esquema de capitalización individual sino que se realiza con cotizaciones superiores al gasto en pensiones durante el periodo transitorio hasta llegar al nivel de madurez deseado.

S3: El número de pensionistas y de cotizantes, la pensión media y el salario base de cotización medio crecen exponencialmente a una tasa constante en todo el periodo. Este supuesto no es necesario en el modelo pero se establece para llegar a soluciones analíticas.

S4: La tasa de rendimiento del capital es constante y tiene ya deducidos los gastos de gestión. Este supuesto tampoco es necesario pero permite simplificar las formulaciones.

S5: La cotización durante el periodo transitorio está sujeta a limitaciones para que no aparezcan efectos económicos colaterales (los derivados de una mayor carga impositiva o de un mayor endeudamiento). Estas limitaciones indican las posibilidades financieras de la economía.

S6: En la formación del fondo de capital se persigue el objetivo de que el tipo de cotización se desvíe lo menos posible respecto a un valor deseado a lo largo de todas las generaciones intermedias. Este valor deseado coincide con el tipo de cotización del esquema de reparto, aunque podría elegirse otro. De esta manera, la formación del fondo de capital se realiza desviándose lo menos posible del esquema de reparto.

S7: No aparecen algunos peligros que se extraen de la teoría de la elección pública. Por ejemplo, no se contempla la posibilidad de que se cambie la decisión de acumular el fondo en algún periodo intermedio en función de cambios políticos. Es decir, existe consenso político en este aspecto y este consenso se mantiene a lo largo del periodo transitorio.

La ecuación en diferencias que recoge la evolución del fondo de capital o

nivel de reservas es:

$$K_{t+1}-K_t = rK_t+c_{t+1}W_{t+1}C_{t+1}-p_{t+1}P_{t+1} \qquad (1)$$

donde:

r es la rentabilidad real del fondo de capital descontados los gastos de gestión.

K_t es la cuantía del fondo de capital en el periodo t.

P_t es el número de pensionistas de jubilación en el periodo t.

C_t es el número de cotizantes en el periodo t.

p_t es la pensión media en el periodo t.

W_t es el salario medio de cotización (base media) en el periodo t.

c_t es el tipo medio de cotización en el periodo t. Incluye la parte de empresarios, trabajadores y Estado.

La ecuación (1) se transforma expresando el fondo de capital como múltiplo de la masa salarial de cotización. Para ello se hace uso del supuesto *S3*, que implica una evolución de cada variable según las siguientes ecuaciones:

$$W_{t+1} = W_t(1+i_W)$$

$$C_{t+1} = C_t(1+i_C)$$

$$p_{t+1}=p_t(1+i_p)$$

$$P_{t+1} = P_t(1+i_P)$$

siendo i_W, i_C, i_p e i_P las tasas de crecimiento de cada variable.

La ecuación (1) se divide por la masa salarial de cotización del periodo $t+1$, es decir, por el término:

$$W_{t+1}C_{t+1} = W_tC_t(1+i_W)(1+i_C) = W_tC_t(1+n)$$

donde n indica la tasa de crecimiento de la masa salarial de cotización. El resultado de la división es:

$$k_{t+1} - \frac{1}{1+n}k_t = \frac{r}{1+n}k_t + c_{t+1} - c^r_{t+1} \qquad (2)$$

donde:

$$k_t = \frac{K_t}{W_tC_t}$$

es la parte que representa el fondo de capital sobre la masa salarial y:

$$c^r_t = \frac{p_tP_t}{W_tC_t}$$

es el tipo de cotización bajo un esquema de reparto.

Realizando transformaciones en la ecuación (2) se llega a:

$$k_{t+1} - k_t = (\frac{r-n}{1+n})k_t + c_{t+1} - c^r_{t+1} \qquad (3)$$

que es la ecuación en diferencias fundamental del modelo. Indica la evolución del fondo de capital como parte de la masa salarial de cotización.

En un sistema de reparto el tipo de cotización es el de reparto y por tanto no se crea un fondo de capital por lo que k_t es igual a cero en todo periodo.

La expresión (3) es válida para cualquier periodo, tanto en los periodos de la etapa transitoria (cuando se forma el fondo), como en los periodos de la etapa madura (cuando el fondo se mantiene constante). Esta segunda etapa no se tiene en cuenta en el modelo porque el horizonte de planificación es el último periodo de la etapa transitoria, pero es interesante ver lo que ocurre en la etapa madura y así observar bajo qué condiciones la existencia de un fondo de capital de estas características produce un alivio en las cotizaciones futuras. Atendiendo al supuesto *S1*, en nuestro sistema maduro el fondo de capital es una parte constante de la masa salarial, es decir $k_t=\beta$, y por tanto, sustituyendo en (3) y llamando T al periodo final de la etapa transitoria se tiene:

$$c_{t+1}=c^{r}_{t+1}-\beta\frac{r-n}{1+n} \qquad t\geq T$$

La conclusión es que el tipo de cotización será inferior al de reparto si y sólo si la rentabilidad del capital supera al crecimiento de la masa salarial de cotización en la fase de madurez[70]. Esta condición es previsiblemente la que se dará en España dentro de unos 25 años, cuando la población cotizante tienda a decrecer por factores demográficos. Por otra parte, la rentabilidad real del capital ha estado en los últimos años por encima del aumento de los salarios y de las bases de cotización, y no es de esperar que esto cambie, a no ser que el volumen del fondo acumulado fuera tan elevado que presionara a la baja los tipos de interés de una forma significativa.

Por tanto, parece ser que es conveniente la creación de este fondo pero eso

[70]Este resultado es similar a otros que comparan el rendimiento del capital con la tasa de crecimiento de la población (López García, 1991). Las diferencias se deben a las distintas transformaciones que se han realizado para plantear la ecuación final.

sólo se puede hacer a costa de pagar un tipo de cotización superior al de reparto durante el periodo transitorio. Efectivamente, si se desea formar este fondo desde una situación de partida en la que no existe ($k_0=0$) se deberá pasar por periodos en los que $k_{t+1}>k_t$, es decir, periodos donde el tipo de cotización sea superior al de reparto según la ecuación (3).

El esfuerzo de cada periodo se puede repartir de muchas formas a lo largo de la etapa transitoria. La forma que adopte la función objetivo será relevante de cara a elegir una u otra distribución del esfuerzo. Un objetivo razonable es el de repartir ese esfuerzo entre todas las generaciones intermedias de forma que nos alejemos del esquema de reparto lo menos posible (supuesto *S6*). Una decisión de esta naturaleza lleva a una función objetivo del tipo:

$$Min \quad \sum_{t=0}^{T-1} (c_{t+1}-c_{t+1}^{r})^2 \qquad (4)$$

Si se desea dar más peso a unos periodos que a otros se puede multiplicar la expresión elevada al cuadrado por una función decreciente en el tiempo (mayor esfuerzo cuanto más cerca del final del periodo transitorio) o una creciente (mayor esfuerzo en las primeras generaciones).

El esfuerzo de cada periodo está limitado para que los efectos económicos no actúen de forma relevante (supuesto *S5*). Este límite se materializa a través de un tipo de cotización máximo que puede ser constante en todos los periodos o variable, por ejemplo si depende del tipo de cotización de reparto. Una posibilidad es que el tipo de cotización máximo de cada periodo (c_t^m) se exprese como el de reparto más una parte constante (c^e), lo que indica que, en cada periodo, los trabajadores, empresarios y el Estado tienen una capacidad financiera para aportar c^e puntos de cotización por encima del de reparto sin que ello cause distorsiones relevantes sobre las variables y parámetros exógenos del modelo. La restricción queda definida de la siguiente manera:

$$c_t \leq c_t^r + c^e = c_t^m \qquad (5)$$

Esta limitación, que indica las posibilidades financieras de la economía, se relaciona estrechamente con la duración del periodo transitorio (T) y con el grado de sistema mixto a alcanzar (β) para determinar el conjunto de oportunidades del problema de optimización, o en otras palabras, para determinar si es posible o no alcanzar el objetivo. Así, si las posibilidades financieras de la economía son pequeñas, se deberá alargar el periodo transitorio o ser menos optimistas en cuanto a la parte a financiar con el capital que se desea alcanzar. De lo contrario, es posible que, aunque las posibilidades financieras se exploten al máximo, no se alcance el objetivo en T: el problema de optimización será infactible.

La interrelación entre estas tres variables permite hacer ejercicios como el cálculo del tiempo mínimo de duración del periodo transitorio dados β y c^e; o el cálculo del máximo peso de los rendimientos de capital dados c^e y T, etc..

Las expresiones (3), (4) y (5), junto con las condiciones iniciales y finales para el fondo de capital forman el enunciado del modelo propuesto:

$$Min \quad \sum_{t=0}^{T-1} (c_{t+1} - c_{t+1}^r)^2$$

s.a:

$$k_{t+1} - k_t = \frac{r-n}{1+n} k_t + c_{t+1} - c_{t+1}^r \qquad t=0,...,T-1 \qquad [I]$$

$$k_0 = 0 \quad k_T = \beta > 0$$

$$c_{t+1} \leq c_{t+1}^m \qquad t=0,...,T-1$$

El problema tiene como variable de control al tipo de cotización y como variable de estado al fondo de capital o nivel de reservas. El tipo de cotización de

reparto es una variable exógena que se determina en base a escenarios sobre la evolución demográfica y económica futura. Se trata de un problema de C.O. discreto con valores inicial y final fijos para la variable de estado. A su vez, adopta la forma de un problema cuadrático-lineal con una función objetivo convexa y un conjunto de oportunidades que puede ser vacío o no en función del valor de algunos parámetros por lo que no está asegurada la existencia de mínimo global. Por último, existen varios parámetros sobre los que se puede realizar análisis de sensibilidad como el tipo de cotización máxima, la parte deseada para la cuantía del fondo, la rentabilidad del fondo y la duración del periodo transitorio.

VI.2.2.- Resolución del modelo a través del P.M. discreto

La solución parte de construir la función hamiltoniana del Problema **[I]**, que es:

$$H_t = -(c_{t+1}-c_{t+1}^r)^2+\mu_t(\frac{r-n}{1+n}k_t+c_{t+1}-c_{t+1}^r)+\lambda_t(c_{t+1}-c_{t+1}^m)$$

A continuación se aplica el Principio del Máximo discreto, cuyas condiciones de óptimo son las siguientes:

$$\frac{\partial H_t}{\partial c_{t+1}} = -2(c_{t+1}-c_{t+1}^r)+\mu_t+\lambda_t = 0 \qquad (6)$$

$$\frac{\partial H_t}{\partial k_t}=\mu_t(\frac{r-n}{1+n}) = -(\mu_t-\mu_{t-1}) \qquad (7)$$

$$\lambda_t(c_{t+1}-c_{t+1}^m) = 0 \quad ; \quad \lambda_t \geq 0$$

Condiciones que, junto con las restricciones, el sistema dinámico y las condiciones iniciales y finales, forman el conjunto de condiciones necesarias de

óptimo del problema. La resolución se obtiene considerando dos casos posibles:

Caso 1: Solución interior a la región de control

Si la solución se da en puntos interiores de la región de control, es decir, si las posibilidades financieras no se utilizan al máximo, se tiene que $\lambda_t=0$, y combinando (6) y el sistema dinámico (3) aparecen las siguientes expresiones:

$$c_{t+1} = \frac{1}{2}\mu_t + c^r_{t+1} \qquad (8)$$

$$k_{t+1} - k_t = \frac{r-n}{1+n}k_t + \frac{1}{2}\mu_t \qquad (9)$$

Por otra parte, la expresión (7) es una ecuación en diferencias lineal de primer orden con coeficientes y término independiente constante, cuya solución es:

$$\mu_t = \mu_0\left(\frac{1+n}{1+r}\right)^t \qquad (10)$$

El parámetro que falta por determinar es μ_0 que se obtiene sustituyendo esta expresión en el sistema dinámico (9) y resolviendo la ecuación en diferencias junto con las condiciones inicial y final. La sustitución de (10) en (9) y la reordenación de términos da lugar a:

$$k_{t+1} - \frac{1+r}{1+n}k_t = \frac{1}{2}\mu_0\left(\frac{1+r}{1+n}\right)^{-t} \qquad (11)$$

Esta ecuación en diferencias es lineal pero con término variable. Afortunadamente, es posible obtener la solución analítica de (11) por el método recursivo y utilizando nociones de series numéricas. La solución particular,

utilizando la condición inicial $k_0=0$, es:

$$k_t = \frac{\mu_0}{2} \frac{b^{2t}-1}{b^{t-1}(b^2-1)} \qquad (12)$$

donde $b=(1+r)/(1+n)$ debe ser distinto de 1.

La trayectoria temporal, dada por la expresión (12), tiene valor final conocido ($k_T=\beta$) de lo cual se deduce el valor de μ_0:

$$\mu_0 = 2\beta \frac{b^{T-1}(b^2-1)}{b^{2T}-1} \qquad (13)$$

Este valor permite completar la solución para todas las variables. Efectivamente, sustituyéndolo en (10) se tiene la trayectoria de la variable de coestado, que a su vez lleva a la solución para el tipo de cotización (al sustituir en (8)) y para el fondo de capital (al sustituir en (12)).

Al estudiar las trayectorias-solución se deduce que la relación entre r y n es crucial para determinar la forma de la solución. Dado que $r>n$ para que tenga sentido la acumulación de capital, el parámetro b es mayor que 1 y en consecuencia la variable de coestado es positiva pero decreciente. El tipo de cotización resultante es mayor que el de reparto pero acercándose a él por lo que el esfuerzo de cotización será mayor en los primeros periodos que en los últimos.

Si la solución obtenida para el tipo de cotización satisface la restricción que indica las posibilidades financieras de la economía en cada periodo, se habrá llegado a la solución óptima para la variable de control y por lo tanto serán óptimas las trayectorias para el resto de variables. De lo contrario la solución no se encuentra en puntos interiores de la región de control y habrá que replantear el proceso de

optimización (caso 2). La condición para que la solución esté en puntos interiores se puede especificar en función de los parámetros de partida. Efectivamente, dado que la diferencia entre el tipo de cotización óptimo y el de reparto es decreciente mientras que la diferencia entre el tipo de cotización máximo y el de reparto es constante (e igual a c^e), es suficiente con que el tipo de cotización óptimo sea menor que el máximo en el primer periodo, de donde se tiene la condición:

$$c_1=\frac{1}{2}\mu_0+c_1^r \leq c_1^m=c_1^r+c^e \qquad (14)$$

Simplificando la expresión (14) y sustituyendo el valor de μ_0 dado en (13) se llega a la siguiente relación entre los parámetros:

$$\beta\frac{b^{T-1}(b^2-1)}{b^{2T}-1} \leq c^e \qquad (15)$$

Esta relación liga los valores de los parámetros para que la solución se encuentre en puntos interiores, y por tanto, para que sea válida la solución obtenida bajo este supuesto.

Caso 2: Soluciones frontera

El caso de solución frontera se caracteriza por la saturación de la restricción (5) en algún periodo, si bien hay que señalar que no es posible mantener la hipótesis de que el tipo de cotización alcanza el tipo máximo admisible en periodos intermedios o finales porque la curva de tipos máximos es paralela y superior a la del tipo de reparto. Entonces, si el tipo de cotización óptimo no alcanza el máximo en los periodos iniciales ya nunca lo alcanzará porque es una curva que va aproximándose a la senda del tipo de reparto, dado que la variable de coestado sigue siendo decreciente, al igual que en el *Caso 1*.

Analíticamente, el caso de solución frontera se da si la solución es factible y los valores de los parámetros no cumplen la expresión (15). En este caso, el tipo de cotización óptimo se sitúa en el máximo hasta algún periodo intermedio, descendiendo a partir de entonces de forma que se cumpla la condición final. Un caso extremo sería aquel en el que el máximo se mantiene hasta el final del periodo transitorio, caso éste del que se deduciría la condición de problema infactible. En efecto, bajo este caso extremo el tipo de cotización es el máximo en todo periodo y, por tanto, la trayectoria (3) para el fondo de capital quedaría:

$$k_{t+1}-k_t = \frac{r-n}{1+n}k_t+c^e \qquad (16)$$

La solución particular de la ecuación (16) utilizando la condición inicial $k_0=0$, es:

$$k_t = \frac{1-b^t}{1-b}c^e \qquad (17)$$

La sucesión para k_t que se deduce de (17) es creciente ($b>1$), pero si en T no alcanza el valor β el problema será infactible: el periodo transitorio es demasiado corto o las posibilidades financieras de la economía son demasiado escasas para el nivel de sistema mixto que se pretende alcanzar. Por tanto, la condición para que se de una solución factible en puntos frontera es que se cumpla la siguiente desigualdad:

$$\frac{1-b^T}{1-b}c^e \geq \beta \qquad (18)$$

Si (18) se cumple con igualdad la solución es cotizar al tipo máximo sin necesidad de resolver ningún problema de optimización. Si (18) se cumple con

desigualdad estricta pero no se verifica (15) la solución es cotizar al máximo hasta algún periodo intermedio y luego cotizar por debajo del máximo. El objetivo es encontrar este periodo intermedio. Llamándolo t' se tiene que a partir de (16) y de (11) aparece respectivamente:

$$k_{t+1}-bk_t = c^e \qquad \forall\, t\le t' \qquad k_0=0 \qquad (19)$$

$$k_{t+1}-bk_t = \frac{1}{2}\mu_0 b^{-t} \qquad \forall\, t\ge t' \qquad k_T=\beta \qquad (20)$$

La solución a ambas ecuaciones en diferencias es, respectivamente:

$$k_t = \frac{1-b^t}{1-b}c^e \qquad \forall\, t\le t' \qquad (21)$$

$$k_t=\beta b^{t-T}-\frac{\mu_0}{2(b^2-1)}\frac{b^{2(T-t)}-1}{b^{2T-t-1}} \qquad \forall\, t>t' \qquad (22)$$

La igualdad entre (21) y (22) en t' junto con la ecuación:

$$c^e = \frac{1}{2}\mu_0 b^{-t'}$$

permite calcular los valores de μ_0 y t'.

En tiempo continuo esto proporcionaría el instante en el que se produce el cambio de trayectoria dada la continuidad de las variables de estado y de control, pero en tiempo discreto el resultado será, probablemente, un valor no natural para el periodo t', por lo que el resultado será aproximado.

La solución también se puede obtener mediante métodos numéricos, a través

de paquetes informáticos que resuelven este tipo de problemas. Estos paquetes aproximan de forma bastante exacta la solución, y más en problemas de tipo cuadrático-lineal como es el planteado. Además, una función objetivo cuadrática (y convexa) y unas restricciones lineales caracterizan a todo mínimo local como global, por lo que no es necesario un estudio más detallado de las condiciones de suficiencia.

VI.2.3.- Comentarios acerca de la solución

En caso de existencia de solución se puede extraer una serie de características que valen tanto para el *Caso 1* como para el *Caso 2* y dependen de los valores de los parámetros r, n y β.

El esfuerzo de cotización en el periodo transitorio (medido como diferencia respecto a la cotización de reparto) es mayor cuanto mayor sea la diferencia entre r y n (cuanto más exceda de 1 el valor de b) y cuanto mayor sea β. Además, el esfuerzo será decreciente (dado $r>n$). En contrapartida, el alivio en la cotización en el periodo maduro depende también de la diferencia entre ambos parámetros y de β, como ha quedado establecido en el epígrafe VI.2.1.. Se trata, por tanto, de hacer un esfuerzo de cotización durante el periodo transitorio para que los intereses del fondo de capital supongan un alivio en las cotizaciones de los periodos futuros.

El parámetro β afecta a la solución de forma lineal por lo que si se quiere doblar el fondo de capital al final de la etapa transitoria habrá que doblar el esfuerzo financiero en todos los periodos intermedios y el alivio en los periodos futuros será también el doble. Por contra, otros parámetros como T y la diferencia entre r y n afectan de forma distinta. Así, cuanto mayor sea la etapa transitoria el esfuerzo anual de cotización es menor porque se reparte entre más periodos; y cuanto mayor sea la diferencia entre r y n más recae el esfuerzo sobre los primeros periodos que sobre los últimos dentro de la etapa transitoria.

Un momento adecuado para elegir una política de este tipo sería aquel en que

se esperan cambios demográficos que hagan disminuir la relación cotizantes-pensionistas. Bajo un sistema de reparto, ello llevaría a un aumento progresivo del tipo de cotización, aumento que se podría suavizar con la formación de un fondo de capital. En estas condiciones resulta también importante elegir un periodo transitorio adecuado para que el esfuerzo de cotización recaiga sobre una población cotizante en expansión y se llegue a la fase madura cuando el número de cotizantes esté en regresión. La duración de este periodo, además, debe ser compatible con las posibilidades financieras de la economía y con el valor del parámetro β a alcanzar.

La formación del fondo de capital se realiza con aportaciones extraordinarias y no pasando a un esquema de capitalización ya que, en ese caso, el modelo no sería válido. Efectivamente, con un esquema de capitalización puro hay una estrecha relación entre las cotizaciones realizadas y las pensiones recibidas, lo que significa que los parámetros que afectan a las variables que determinan el gasto en pensiones no son independientes de los que afectan a las que determinan el volumen de cotizaciones, como se considera en el modelo descrito. La igualdad financiero-actuarial del sistema de capitalización es una ecuación de equilibrio de tipo longitudinal en el sentido que es una igualdad a lo largo del tiempo para un individuo, mientras que la ecuación que se ha considerado en el modelo recoge un equilibrio transversal, esto es, un equilibrio en un periodo de tiempo para el total de la población. Esta distinción impide que se pueda recoger el esquema de capitalización en un modelo como el propuesto.

VI.2.4.- Aplicación numérica

A continuación se desarrolla una aplicación para ilustrar las características de la solución. En la medida de lo posible se tomarán datos del caso español.

Se toma el periodo transitorio 1996-2021 (T=25) al final del cual se debe haber formado un fondo de capital. Los valores de los parámetros que se especifican a continuación están justificados de forma más detallada en la próxima aplicación y se recogen en la tabla VI.3..

Tasas de incremento interanual	Periodo transitorio 1996-2021	Periodo de madurez 2021-2046
Pobl. cotizante (i_C)	0,35%	-0,7%
Pobl. pensionista (i_P)	0,81%	1,36%
Tipo de interés real	3%	
i_W y i_p	1,5%	

Tabla VI.3.: Valores de los parámetros en el ejemplo numérico.

Bajo estas condiciones el valor del parámetro b es 1,011 (en la etapa de madurez es igual a 1,022).

Se elige un valor para el parámetro β de forma que el volumen del fondo de capital no cause distorsiones en el mercado financiero y de forma que sea compatible con la duración del periodo transitorio y con las posibilidades financieras de la economía (expresadas a través de un tipo de cotización máximo por encima del de reparto). Para el ejemplo tomado, la relación entre β y c^e se puede representar gráficamente (gráfico VI.1.) y así identificar el tipo de solución según las desigualdades (15) y (18).

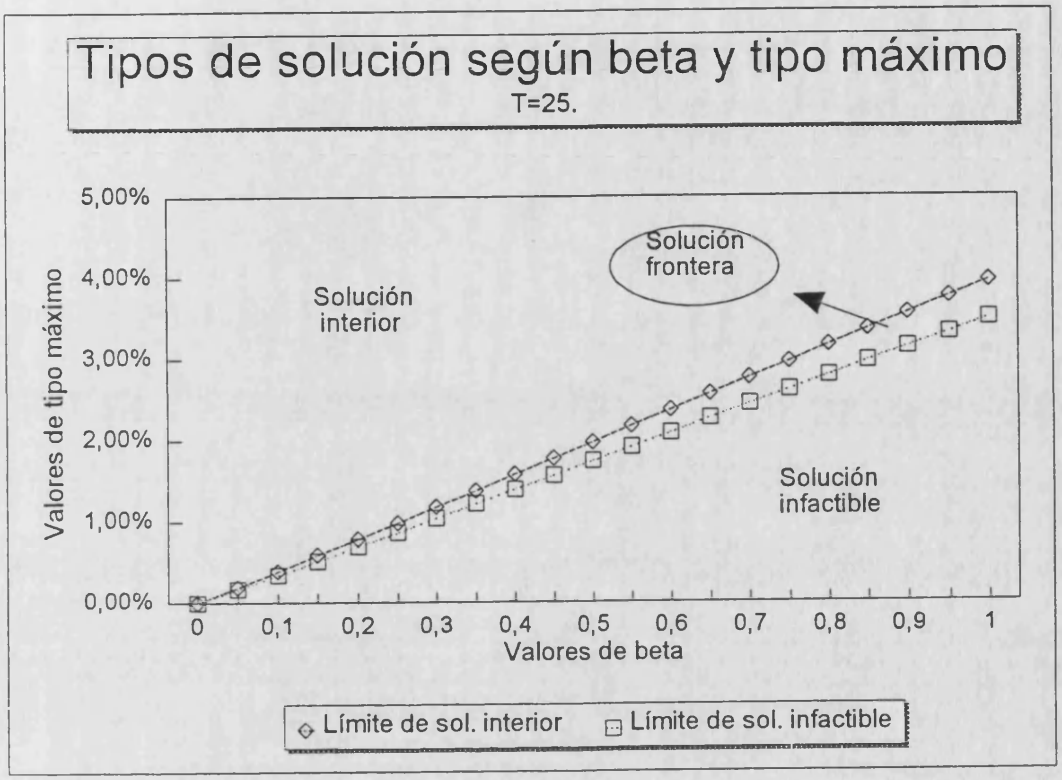

Gráfico VI.1.: Tipos de solución en el ejemplo numérico.

Supongamos que las posibilidades financieras de la economía y el mercado financiero permiten una solución interior para un valor $\beta=0,5$[71]. Este valor determina a su vez el valor inicial de la variable de coestado, en este caso $\mu_0=0,0395$.

El esfuerzo de cotización es variable y viene dado por la función:

[71]En el caso español, esto supone llegar al año 2021 con un fondo de capital de aproximadamente 19 billones de ptas. del año 1996. Se debe suponer que el mercado financiero y la rentabilidad del capital no se verán distorsionados por la existencia de este fondo.

$$c_t - c_t^r = \frac{1}{2} 0{,}0395 \cdot (1{,}011)^{-t+1} \qquad t \leq 25$$

Por tanto, es un esfuerzo decreciente cuyo valor inicial es de 1,97 puntos porcentuales. A su vez el alivio en la etapa de madurez es constante e igual a 1,09 puntos:

$$c_t^r - c_t = \beta \frac{r-n}{1+n} = \beta(b-1) = 0{,}0109 \qquad t > 25$$

La solución se admite siempre y cuando los 1,97 puntos de cotización extraordinaria máxima se puedan soportar por la economía[72], bien a través de aportaciones del Estado compatibles con otros objetivos económicos (déficit público) o bien con mayores aportaciones de trabajadores y/o empresarios que no distorsionen el mercado laboral ni depriman el consumo o la inversión.

En este sentido, hay que decir que una profundización en la racionalización de la estructura financiera de la Seguridad Social sería suficiente para conseguir estos recursos adicionales que se requieren en la etapa transitoria para formar el fondo de capital. Efectivamente, tal y como se resalta en la próxima aplicación de una forma más detallada, si las cotizaciones que actualmente se desvían a pagar complementos a mínimos y gasto sanitario se mantuvieran en el sistema contributivo sería equivalente a un aumento de 2,38 puntos del tipo de cotización, lo cual supera ampliamente los 1,97 puntos máximos requeridos en el ejemplo. Sin embargo, realizar este cambio de forma repentina puede ser excesivo y en ese caso habría que disminuir β (objetivo de sistema mixto) y/o aumentar T (duración del periodo transitorio).

[72]En el caso español, esto supone en 1996 algo menos de medio billón de ptas..

VI.3.- MODELO DE AJUSTE ÓPTIMO DEL PORCENTAJE DE COTIZACIÓN Y DE LA TASA DE REEMPLAZO DE PENSIONES CON FORMACIÓN DE UN FONDO DE CAPITAL

VI.3.1.- Supuestos y planteamiento

El modelo que se desarrolla pretende, al igual que el anterior, conseguir un equilibrio estable a largo plazo en la financiación de las pensiones de jubilación mediante la creación de un fondo de capital. Sin embargo, ahora el esfuerzo se desglosa en la parte de cotizantes y en la parte de pensionistas, que deben ajustarse lo más posible a unos valores deseados; por otra parte, el fondo de capital no está sometido a una condición terminal sino a restricciones sobre su cuantía para evitar efectos políticos y económicos adversos ya comentados. Adicionalmente, las variables exógenas, como la evolución del número de cotizantes y de pensionistas, no se toman en su versión simplificada (forma exponencial) sino que se concretan más.

Los supuestos del modelo son:

S1: Se parte de un porcentaje deseado de cotización sobre la masa salarial que incluye la parte de empresarios, trabajadores y Estado. El reparto entre las tres partes no queda reflejado en el modelo y se debe establecer a posteriori. Se supone que este porcentaje deseado es constante en el tiempo.

S2: También se parte de una tasa de reemplazo de pensiones deseada. Esta tasa es la parte que supone la pensión media respecto del salario medio de cotización y se supone constante en el tiempo.

S3: Se conoce la evolución de las otras variables del modelo como el número de cotizantes, el salario medio y el número de beneficiarios de la pensión de jubilación. Dado que el modelo es a largo plazo, el número de cotizantes y jubilados dependerá de la demografía y no de la coyuntura económica. Se trata, por tanto, de

variables exógenas.

S4: El horizonte de planificación es conocido y suficientemente dilatado como para incluir todo el impacto demográfico.

S5: La tasa de rentabilidad del capital sigue las consideraciones vistas en la anterior aplicación, entre ellas la de su constancia a lo largo del periodo de planificación.

S6: Para evitar las distorsiones del fondo de capital se imponen restricciones a su cuantía en términos de no poder exceder de un múltiplo de las pensiones de cada periodo.

S7: El objetivo es minimizar las desviaciones del volumen de cotizaciones y del volumen de pensiones respecto de sus valores deseados.

S8: No aparecen los peligros que se derivan de la teoría de la elección pública comentados en la aplicación del epígrafe VI.2..

El modelo logra el equilibrio financiero durante todo el horizonte de planificación por medio de desviaciones respecto a los valores deseados. Estas desviaciones dan lugar normalmente a desequilibrios en cada periodo cuyo resultado es la acumulación de un fondo de capital a lo largo de los periodos intermedios.

En base a las hipótesis del modelo se tiene el siguiente enunciado del problema:

$$Min \quad \sum_{t=1}^{T} [(c_t W_t C_t - c^* W_t C_t)^2 + (d_t W_t P_t - d^* W_t P_t)^2]$$

s.a:

$$K_{t+1} - K_t = rK_t + c_{t+1} W_{t+1} C_{t+1} - d_{t+1} W_{t+1} P_{t+1} \qquad t=0,...,T-1 \qquad \text{[II]}$$

$$K_0 = 0 \ , \ K_t \geq 0 \qquad t=1,...,T$$

$$K_t \leq a d_t W_t P_t \qquad t=1,...,T$$

Se ha seguido la misma nomenclatura que en el epígrafe VI.2.1. para cada variable y, además, c^* es el tipo de cotización deseado, d_t es la tasa de reemplazo en el periodo t, d^* es la tasa de reemplazo deseada y a es la parte sobre la masa de pensiones que, como máximo, puede alcanzar el fondo de capital según el supuesto *S6*.

El problema tiene dos variables de control, el tipo de cotización, c_t, y la tasa de reemplazo de pensiones, d_t; y una variable de estado, el fondo de capital o nivel de reservas, K_t. Se trata de un problema de C.O. discreto con valor inicial fijo y final libre para la variable de estado. A su vez, adopta la forma de un problema cuadrático-lineal con una función objetivo convexa y un conjunto de oportunidades no vacío (dado que el esquema de reparto cumple las restricciones), por lo que existirá mínimo global.

El problema **[II]** se puede entender como el modelo básico de estudio, pero sobre él es posible introducir ligeras modificaciones que resultan interesantes y que pasamos a enumerar:

M1) El modelo así enunciado lleva implícito un reparto igual de la carga entre cotizantes y pensionistas. Esto se puede cambiar fácilmente introduciendo parámetros de penalización estrictamente positivos (M y N) de forma que la función objetivo pasa a ser:

$$Min \quad \sum_{t=1}^{T} [M(c_t W_t C_t - c^* W_t C_t)^2 + N(d_t W_t P_t - d^* W_t P_t)^2]$$

Con este cambio se tiene que si $M=N$ la carga se reparte por igual entre ambos colectivos, mientras que si $M>N$ ($M<N$) la población pasiva deberá realizar un ajuste mayor (menor) que la activa. Los parámetros M y N deben ser elegidos por el sujeto decisor en base a los efectos económicos y políticos de cada alternativa. Los casos extremos se resuelven de la siguiente manera: si la decisión es que no recaiga esfuerzo alguno sobre la población pensionista entonces $d_t=d^*$ $\forall t$ y si se desea que no lo soporten los cotizantes hay que hacer $c_t=c^*$; en ambos casos el problema tendría una variable de control menos.

M2) Al igual que se hizo en la aplicación del epígrafe VI.2., se puede forzar una distribución del ajuste más intensa en los primeros periodos (introducción de una función creciente con el tiempo) o en los últimos (con una función decreciente) utilizando algún parámetro de preferencia temporal. En ausencia de estas funciones el modelo recoge implícitamente una mayor penalización para las desviaciones futuras debido a que los salarios suelen ser ligeramente crecientes en términos reales, y, en consecuencia, el ajuste será mayor en los primeros periodos.

M3) El modelo permite, mediante la introducción de nuevas restricciones, forzar una determinada trayectoria temporal para el tipo de cotización y la tasa de reemplazo de forma que sea más fácil, desde un punto de vista de política económica, llevarlas a la práctica. Se trata de acercar la solución óptima del modelo matemático al resultado de aplicar los instrumentos de política económica.

Por ejemplo, se puede exigir que las tasas y tipos sean constantes en el tiempo, lo cual significaría un ajuste inmediato en ambas variables a los

valores que, en ese caso, se interpretarían como los de equilibrio a largo plazo. Para ello, se seguiría la metodología comentada en el Capítulo III respecto a la optimización de parámetros, es decir, cada variable se consideraría como una de estado con valor constante. La función objetivo no cambiaría pero aparecería una nueva ecuación en diferencias dentro del sistema dinámico. Si la tasa de reemplazo se desea con valor constante se tiene:

$$d_{t+1}-d_t=0 \qquad d_T \text{ libre}$$

El mismo razonamiento habría que hacer si se exige que el tipo de cotización óptimo sea constante. En el caso en que ambas variables de control fueran parámetros a optimizar se tendría un problema dinámico pero sin variables de control, cuya solución podría ser fácilmente infactible a no ser que se relajaran mucho las restricciones sobre el tope máximo de la cuantía del fondo.

También se puede forzar una trayectoria temporal para el tipo de cotización y la tasa de reemplazo que sea constante a partir de cierto periodo intermedio. Se estaría exigiendo, en ese caso, que el ajuste en ambas variables finalizara en un periodo t' a partir del cual la trayectoria sería estable. Las restricciones a añadir serían:

$$c_{t+1}=c_t \qquad t\geq t'$$

$$d_{t+1}=d_t \qquad t\geq t'$$

VI.3.2.- Resolución a través del P.M. discreto

La función hamiltoniana del Problema [II] es:

$$H_t = -(c_{t+1}-c^*)^2 W_{t+1}^2 C_{t+1}^2 - (d_{t+1}-d^*)^2 W_{t+1} P_{t+1}^2 + \\ +\mu_t (rK_t + c_{t+1} W_{t+1} C_{t+1} - d_{t+1} W_{t+1} P_{t+1}) + \\ +\omega_{t+1} K_{t+1} + \upsilon_{t+1} (ad_{t+1} W_{t+1} P_{t+1} - K_{t+1})$$

Las condiciones de óptimo correspondientes, teniendo en cuenta el P.M. discreto desarrollado en el Capítulo II son:

$$\frac{\partial H_t}{\partial c_{t+1}} = -2(c_{t+1}-c^*) W_{t+1}^2 C_{t+1}^2 + \mu_t W_{t+1} C_{t+1} = 0 \qquad t=0,...,T-1$$

$$\frac{\partial H_t}{\partial d_{t+1}} = -2(d_{t+1}-d^*) W_{t+1}^2 P_{t+1}^2 - \mu_t W_{t+1} P_{t+1} + \upsilon_{t+1} a W_{t+1} P_{t+1} = 0 \qquad t=0,...,T-1$$

$$\frac{\partial H_t}{\partial K_t} = r\mu_t + \omega_t - \upsilon_t = -(\mu_t - \mu_{t-1}) \;\; ; \;\; \mu_T = 0 \qquad t=1,...,T$$

$$\omega_t K_t = 0 \;\; ; \;\; \omega_t \geq 0 \;\; ; \;\; K_t \geq 0 \qquad t=1,...,T$$

$$\upsilon_t (ad_t W_t P_t - K_t) = 0 \;\; ; \;\; \upsilon_t \geq 0 \;\; ; \;\; K_t \leq ad_t W_t P_t \qquad t=1,...,T$$

La resolución de estas ecuaciones junto con el sistema dinámico proporciona el porcentaje de cotización y la tasa de reemplazo de cada periodo que, en su conjunto, más se acercan a los valores deseados respetando todas las restricciones.

Si se manipulan las condiciones de óptimo se consiguen algunas características de la solución. Por ejemplo, de las dos primeras condiciones se llega

a:

$$c_{t+1} = c^* + \frac{\mu_t}{2W_{t+1}C_{t+1}}$$

$$d_{t+1} = d^* - \frac{\mu_t - a\upsilon_{t+1}}{2W_{t+1}P_{t+1}}$$

La interpretación que se puede hacer es que, en aquellos periodos en los que no se alcanza el máximo volumen del fondo posible ($\upsilon_t=0$), las desviaciones de las dos variables respecto a sus valores deseados tienen signos opuestos: si las cotizaciones son superiores a los valores deseados, entonces las pensiones son menores y viceversa. El signo de la variable de coestado μ_t es el dato relevante. Esta variable está sometida a una ecuación dinámica que se desprende del resto de condiciones de óptimo. Una resolución analítica exigiría concretar el tipo de funciones y la evolución temporal de las variables exógenas del modelo. Sin embargo, la evolución previsible de estas variables da lugar a una sucesión de puntos cuyo término general no se puede explicitar. En consecuencia resulta más adecuado utilizar la Programación no Lineal como técnica de resolución del modelo con ayuda de programas informáticos que, en general, aproximan bien la solución en problemas cuadrático-lineales como el descrito.

La resolución del modelo proporciona el porcentaje de cotización y la tasa de reemplazo óptimos en el sentido de acercarse lo más posible a sus niveles deseados. Estas dos variables de control son muy generales y queda por resolver el problema de cómo instrumentalizar los valores obtenidos. Es decir, el modelo resuelve una cuestión puramente financiera pero a la hora de poner en práctica la solución obtenida se debe tener en cuenta también cuestiones de tipo económico y político.

Efectivamente, por lo que hace referencia a las cotizaciones, el porcentaje

hay que repartirlo entre trabajadores, empresarios y Estado. En este sentido, se ha escrito mucho sobre la conveniencia de disminuir las cotizaciones empresariales a cambio de aumentar la de trabajadores y Estado (subida de impuestos). Esta nueva distribución del tipo de cotización se argumenta por sus efectos beneficiosos sobre la demanda de trabajo y también, en nuestro país, por avanzar hacia una distribución de las cuotas más acorde con la existente en la Unión Europea. Las distintas propuestas lanzadas sirven para orientar la decisión política sobre cómo debe ser la distribución del porcentaje de cotización obtenido por el modelo.

En cuanto a la tasa de reemplazo de pensiones, la forma de instrumentalizarla también debe ser determinada después de realizar la optimización. Dado que lo que se obtiene es un porcentaje medio, existe margen para incorporar, si se desea, elementos redistributivos. Por ejemplo, se puede aumentar la pensión de forma lineal o aumentar menos las pensiones altas (redistribución de la renta) o se puede disminuir las pensiones de los nuevos pensionistas a cambio de mantener la revalorización de las pensiones (redistribución intergeneracional). También cabría la posibilidad de afectar a cada régimen de la Seguridad Social de forma distinta pero logrando que la tasa media siga siendo la óptima.

Este modelo, por tanto, es una alternativa al esquema de financiación de las pensiones de jubilación en el sentido que propugna la formación de un fondo de capital, pero al mismo tiempo, necesita de propuestas complementarias para que el resultado obtenido se pueda instrumentalizar de forma que se incorporen los efectos deseados.

La formación de un fondo de capital que se deriva de este modelo permite distribuir parte de la carga intergeneracionalmente acumulando capital cuando la demografía es favorable y desacumulando cuando es desfavorable, anticipándose así a los problemas futuros causados por el envejecimiento de la población. Efectivamente, en España, las generaciones que llegarán a la edad de jubilación en los años 2000-2015 no van a ser muy numerosas y, al mismo tiempo, la tasa de actividad registrará una tendencia al alza por la llegada a las edades más activas de

las generaciones numerosas del *baby boom*. En consecuencia, se puede aprovechar ese periodo para formar un fondo de capital, que permitirá financiar parte de las pensiones de estas generaciones cuando lleguen a la jubilación en los años 2025-2040 desacumulando capital y no cargando así el peso sobre los activos de esos años. Para ver cómo surge esta estrategia, a continuación se aplica el modelo al caso español utilizando datos reales de la S.S., proyecciones demográficas y estableciendo algunos supuestos adicionales.

VI.3.3.- Aplicación del modelo al caso español en el periodo 1996-2056

VI.3.3.1.- Metodología y determinación de parámetros y variables exógenas

La aplicación con datos numéricos pretende comprobar la coherencia del modelo teórico y observar los rasgos principales de la solución. Más que los valores concretos de las variables, interesan cuestiones más generales como su tendencia dinámica o la diferencia entre los resultados cuando se parte de escenarios alternativos. Por este motivo, y dado que se enuncian hipótesis razonables para determinar las variables exógenas y los parámetros, las conclusiones de tipo general no variarán sustancialmente si dichas hipótesis resultan algo alejadas de las que en la realidad se manifiesten. Esto significa que el rango de validez de los resultados numéricos concretos está limitado al cumplimiento de las hipótesis mientras que las conclusiones generales que de ellos se derivan tienen una validez más amplia.

En la elección de hipótesis hay que tener en cuenta que el modelo es a muy largo plazo con lo que las fuentes de incertidumbre son numerosas y no compensa un excesivo esfuerzo por refinarlas ya que las desviaciones serán inevitables por muchos factores que se tengan en cuenta. La alternativa que parece más apropiada es la de plantear distintos escenarios a partir de unas hipótesis intermedias bien justificadas (esto es lo que se hace, por ejemplo, en las proyecciones demográficas a largo plazo).

La duración del periodo de estudio se ha tomado en 60 años (aproximadamente dos generaciones). Un periodo mayor hubiera sido más apropiado para englobar todo el cambio demográfico, sin embargo, resultaría excesivo para realizar proyecciones mínimamente fiables.

Para los datos de las variables exógenas (número de cotizantes y número de pensionistas) y a falta de proyecciones oficiales conocidas, se adopta la hipótesis general de que evolucionan principalmente en función de la componente demográfica. Por tanto, hay que concretar cómo se realiza la proyección demográfica y, posteriormente, cómo se pasa a la proyección para los cotizantes y los pensionistas.

Proyección demográfica

Para la proyección demográfica se dispone de una publicación del Instituto de Demografía (1994). Dicha proyección contiene los datos desagregados hasta el 2006, y a partir de ahí sólo se incluyen datos cada cinco años y por grupos quinquenales hasta el año 2026. Para completar la proyección hemos adoptado algunas de las hipótesis del Instituto de Demografía (I.D.) y hemos utilizado la misma metodología. Las proyecciones a un plazo tan largo están sometidas a mucha incertidumbre por lo que los resultados deben admitirse sólo a grandes rasgos. El año 2026 lo tomaremos como frontera en otros supuestos que estableceremos más adelante, pero es útil escoger un periodo más grande de tiempo porque el punto más bajo del cambio demográfico se espera alrededor de 20 años después. Los aspectos generales de nuestra proyección son:

- Hasta el año 2026 inclusive hemos utilizado la proyección del I.D. en su hipótesis media[73].

[73]Desde el año 2006 los datos son quinquenales. El paso a anuales se ha realizado por interpolación lineal.

- Para el periodo 2026-2056 hemos utilizado la hipótesis media del I.D. del quinquenio 2021-2026 para la mortalidad y la hipótesis alta del I.D. del año 2025 para la natalidad[74].

Los valores de las principales variables de la proyección se resumen en la tabla VI.4.

	1991-2026 Hipótesis media I.D.	2026-2056 Hipótesis propia
Población de partida	Censo provisional de 1991 ajustado	Proyección del I.D. para el año 2026
Mortalidad: esperanza de vida (año).	H. 74,41(91)-77,82(25) M. 80,76(91)-83,91(25)	Hombres 77,82 Mujeres 83,91
Natalidad: nºhijos (año) media y varianza (año)	1,326(1991)-1,8(2025) 28,92 y 29,25 (1991)	2,1 29,02 y 25,29
Saldo migratorio	+20.000 (hasta 2005)	0

Tabla VI.4.: Hipótesis de las proyecciones demográficas.

A partir de la proyección de población se obtienen las proyecciones para el número de cotizantes y el de pensionistas de jubilación. Dado que el número de cotizantes depende básicamente de la población entre 15 y 64 años y el número de pensionistas de la población de más de 65 años, presentamos los resultados por grandes grupos de edad (gráfico VI.2.). El cambio demográfico muestra una caída máxima del 56,4% en el ratio población de 15-64 años por población de más de 65

[74]Mantener la hipótesis media para la natalidad hubiera llevado a una población de 34,8 millones de personas al final del periodo, lo cual parece una cifra excesivamente baja.

(de 4,44 en 1996 a 1,93 en 2046).

Número de cotizantes

El número de cotizantes a la S.S. (número medio mensual de trabajadores en alta con efectos en cotización), que es aproximadamente el número de ocupados en la economía, es una variable que consideramos exógena en el modelo. No obstante, se debe tener en cuenta la interrelación entre esta variable, el producto nacional y la productividad del trabajo. Nuestra hipótesis, al ser un modelo a largo plazo, es que la fuerza de trabajo depende de la población entre 15 y 64 años, lo cual afectará al crecimiento del producto según evolucione la productividad, todo ello para mantener la interrelación entre las tres variables. Otra opción hubiera sido, por ejemplo, lanzar hipótesis sobre el crecimiento del producto y de la productividad y deducir la evolución del número de ocupados.

El número de ocupados o de cotizantes depende de la componente demográfica, de variables socio-culturales (tasa de actividad) y de variables económicas (tasa de paro). La tasa de actividad total apenas se ve afectada por el ciclo económico y, en los últimos años, se ha comportado de forma prácticamente constante en el tiempo. Debido a razones de cambios socio-culturales, se observa una tendencia a una disminución de la tasa de actividad de los hombres y un aumento en las mujeres y, por otra parte, una disminución de la tasa en edades jóvenes (menos de 25 años) y más mayores (más de 55) por un aumento en las edades intermedias.

Dado que es imposible anticipar el efecto de los futuros cambios socio-culturales, la elección de una u otra hipótesis acerca de la evolución de la tasa de actividad es igualmente criticable[75]. Una tasa global constante parece adecuarse a lo ocurrido en los últimos años pero no tiene en cuenta el cambio demográfico ni

[75]Existe algún estudio clásico (Bell, 1976) de tipo socioeconómico que intenta prever tendencias futuras en la actividad aunque referido a EEUU y ya algo alejado en el tiempo.

algunos hechos que previsiblemente impulsarán la tasa de actividad al alza. Otra posibilidad es la de mantener constantes las tasas de actividad por edades, y en ese caso el cambio en la estructura de edad de la población afecta a la tasa de actividad global. Así, a medida que los grupos de edad más numerosos lleguen a edades de más actividad la tasa global aumentará, pero posteriormente se producirá el efecto contrario al llegar a la jubilación. Mantener esta hipótesis supondría pasar de una tasa de actividad del 49% actual al 51% en diez años y bajar al 44% treinta años después[76].

Sin embargo, la tasa de actividad en el futuro parece que tenga una tendencia a aumentar si se desarrolla el trabajo a tiempo parcial, se entra en una fase de moderación salarial y se produce una cierta convergencia a las tasas europeas (cuya tasa media es un 15% superior a la española en 1991 según la monografía sobre la protección social en Europa en 1993).

En base a estas consideraciones recogemos como hipótesis más razonable la de una tasa de actividad creciente linealmente, partiendo del nivel de 1994 y tomando como punto de llegada en el año 2026 la de partida más un 15%[77]. Algunas cifras que se obtienen como resultado de aplicar esta metodología se muestran en la tabla VI.5.

[76]Los porcentajes son de activos sobre la población de 16 años o más.

[77]La tasa de actividad de partida, en nuestra aplicación, se obtiene dividiendo los activos (media de 1994 según la EPA) entre la población entre 15 y 64 años a 1 de Enero de 1994 según la proyección del I.D.. Esta tasa difiere de la tasa de actividad según la EPA porque el denominador del cociente es distinto, pero es la forma adecuada de calcularla dado que así transformará la población según el I.D. (a 1 de Enero) en activos según EPA (media anual).

Tasa de partida año 1994	Tasa de llegada años 2026-2056	Incremento anual
58,14%	66,86%	0,273 ptos.

Tabla VI.5.: Hipótesis sobre tasas de actividad.

Los activos así calculados se multiplican por uno menos la tasa de paro para obtener los ocupados, es decir, el número de cotizantes. En este contexto, hay que decir que las tasas de paro sí que están sometidas al ciclo económico, por lo que se hace necesario escoger adecuadamente las cifras correspondientes a la fase intermedia del ciclo y mantenerla constante para todo el periodo. Por contra, apenas se distingue ninguna tendencia diferenciadora por sexos y edades, es decir, el aumento o disminución del paro afecta a ambos sexos y a todos los grupos de edad.

En cuanto a elección de la fase intermedia del ciclo, se puede plantear un conjunto de escenarios dada la dificultad de prever las tendencias futuras del mercado laboral. En la proyección utilizamos como escenario medio la tasa de paro media del último ciclo económico, considerado éste como el tiempo que transcurre desde un mínimo a un máximo en la tasa de paro o desde un máximo a un mínimo. En el caso español el último mínimo se dio en el 4º trimestre de 1990 y el máximo en el 2º de 1994. La tasa de paro media de esos trimestres fue del 19,02%[78]. Los escenarios alternativos varían ese porcentaje en dos puntos al alza y a la baja. La tabla VI.6. resume los tres escenarios.

[78]Este porcentaje tiene en cuenta el número de cotizantes en el numerador y, por tanto, difiere de la tasa de paro según la EPA. Si se consideran los ocupados según la EPA en lugar de los cotizantes, el ciclo económico se alarga tres trimestres más (uno al principio y dos al final) y el porcentaje medio sube al 19,85%. Parece ser que la encuesta de población activa es el mejor indicador de la evolución de la ocupación pero no de la cifra global de ocupados que es lo que aquí interesa.

Escenario bajo	Escenario medio	Escenario alto
21%	19%	17%

Tabla VI.6.: Hipótesis sobre tasas de paro.

Con las tasas de actividad, los escenarios sobre tasas de paro y la proyección demográfica, se calcula la proyección de población ocupada realizando las operaciones aritméticas correspondientes. El resultado para el escenario medio se muestra en el gráfico VI.3.. El número medio de cotizantes que se deduce para 1996 es de 12,7 millones en la hipótesis media.

Número de pensiones de jubilación

El cálculo de la población pensionista (número medio mensual de pensiones de jubilación) no sigue el mismo criterio en todo el periodo dadas las características del sistema de pensiones español (posibilidad de jubilarse con menos de 65 años para los trabajadores en alta antes de 1967). Así, partiendo de un nivel inicial, los aumentos en la población pensionista (recordemos que sólo en su parte contributiva) tienen tres factores explicativos:

- Efecto población: aumento en la población de más de 65 años.

- Efecto protección o cobertura: es el aumento que se produce en el número de pensiones de jubilación contributivas de mayores de 65 años descontando el efecto población.

- Efecto menores de 65 años: aumento en el número de pensiones de jubilación contributivas descontando los dos primeros efectos.

Los dos primeros efectos son los importantes en una proyección a largo plazo. El efecto población se calcula a través de las proyecciones demográficas. La estimación del efecto protección en el futuro es más problemático. Un valor igual a cero significaría la madurez del sistema contributivo en mayores de 65 años y un valor positivo (negativo) significaría que el porcentaje de población de 65 años con derecho a pensión de jubilación es creciente (decreciente) lo que implica un sistema todavía no maduro. El cálculo del efecto protección en los años recientes da lugar a cifras distorsionadas debido a la incorporación de colectivos importantes (ONCE en 1991, ITP en 1992 y MUNPAL en 1993) que sobrevaloran el efecto, y a medidas coyunturales de seguimiento de pensionistas ya fallecidos (1994), que lo infravaloran. Esto hay que tenerlo en cuenta al observar la tabla VI.7., donde se desglosan los distintos efectos.

	1-1-92 / 1-1-91	1-1-93 / 1-1-92	1-1-94 / 1-1-93	1-1-95 / 1-1-94
Incremento total	2,87%	2,83%	4,41%	2,67%
Efecto población	2,67%	2,38%	2,54%	2,35%
Efecto protección	0,13%	0,17%	1,70%	-0,15%
Efecto < 65	0,06%	0,27	0,12%	0,46%

Tabla VI.7.: Descomposición del aumento en el número de pensiones de jubilación contributivas. Elaboración propia. Datos de los Boletines Informativos de la S.S. y de la proyección de población según el I.D..

La hipótesis que se tome para el efecto protección debe tener en cuenta hechos como la incorporación de la mujer al trabajo, que originará en un futuro un mayor porcentaje de personas de más de 65 años con derecho a pensión de jubilación; la reforma de 1985, cuyo efecto es contrario al anterior; las elevadas

tasas de paro de los últimos años, que también puede provocar la llegada a la jubilación de personas que no han cotizado los años suficientes, etc.. En general, es de esperar un efecto protección positivo porque el aumento en la cobertura de la población pensionista se produce con retardo respecto al aumento en la cobertura de la población cotizante y esto es lo que ha ocurrido en las últimas décadas (el porcentaje de trabajadores en alta respecto a la población activa ha pasado del 74% en 1972 al 91% en 1992). El supuesto que adoptaremos para el efecto protección será de 0,1% hasta el año 2026 y de 0% (sistema maduro) después.

El efecto menores de 65 años, aunque puede ser positivo durante algunos años, comenzará a ser negativo (los jubilados de menos de 65 años aumentarán menos que los de más de 65 años) hasta dar lugar a un número de jubilados que más o menos se mantendrá constante. Como consecuencia, el efecto menores de 65 años se establece en términos absolutos y se supone que el número de pensiones en menores de 65 años será de 340.000 en 1996[79]. Luego, debido a la llegada a los 60 años de las generaciones poco numerosas nacidas en la guerra civil y la posguerra mantendremos este valor constante hasta el año 2004, a partir del cual hemos supuesto un periodo de 6 años en los que desaparecen las jubilaciones con coeficiente reductor[80]. El número de jubilaciones sin coeficiente reductor lo mantenemos constante para el resto del periodo de análisis debido a que los regímenes que permiten la jubilación normal antes de los 65 están estancados o en recesión.

En el gráfico VI.3. se presentan, junto con el número de cotizantes, los resultados para el número de jubilados que se obtienen a partir de la proyección demográfica y de los supuestos antes comentados. El ratio cotizantes-jubilados sufre, bajo estas consideraciones, una caída del 46,6% (de 3,81 en 1996 a 2,04 en

[79]Esta previsión corresponde a un aumento anual del 5% en 1995 y 1996 a partir de la cifra media de 1994 que es de 308.285.

[80]La eliminación de este tipo de jubilaciones (el 90% del total) la hemos considerado lineal, es decir, a razón de 51.000 por año, hasta llegar al 2010 con 34.000 jubilados.

2046)[81].

En la tabla VI.8. se presentan algunos resultados obtenidos de las proyecciones que permiten separar el efecto demográfico (1ª y 3ª filas) del efecto del resto de supuestos planteados, así como comparar la evolución por grandes periodos.

	1996-2006	2006-2026	2026-2056
Pobl. 15-64 años	0,06%	-0,21%	-0,57%
Cotizantes	0,51%	0,22%	-0,57%
Pobl. 65 o más	1,39%	1,11%	0,38%
Jubilados	1,07%	0,92%	0,38%

Tabla VI.8.: Incrementos interanuales resultantes de las proyecciones.

[81]Obsérvese que esta caída es menor a la del efecto demográfico (56,4%) debido a los supuestos realizados para determinar los cotizantes y jubilados que, en este sentido, se pueden considerar supuestos ligeramente optimistas para el sistema.

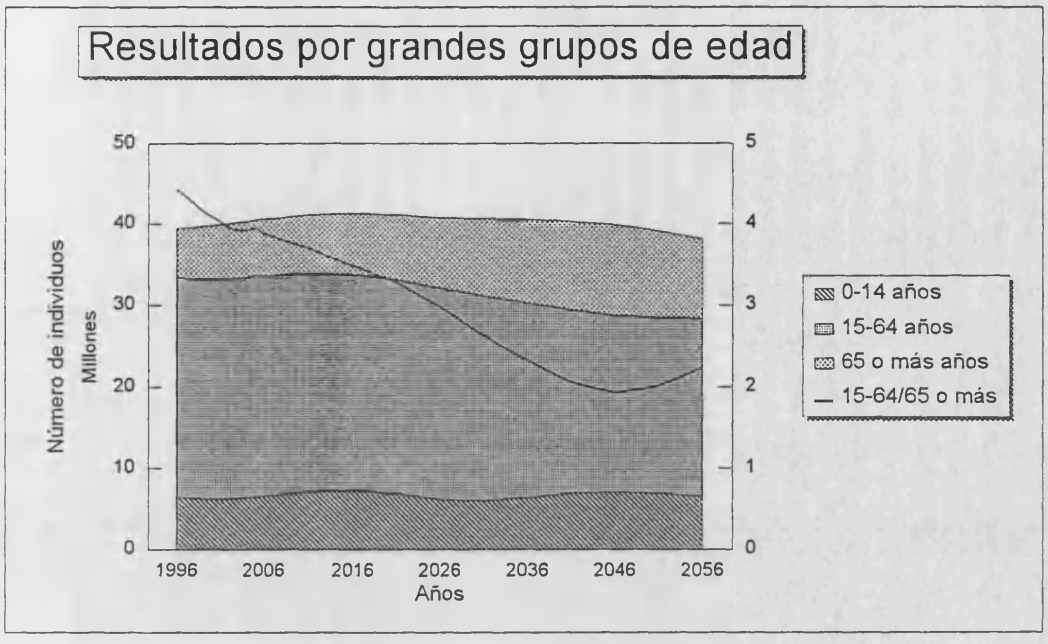

Gráfico VI.2.: Proyección de población

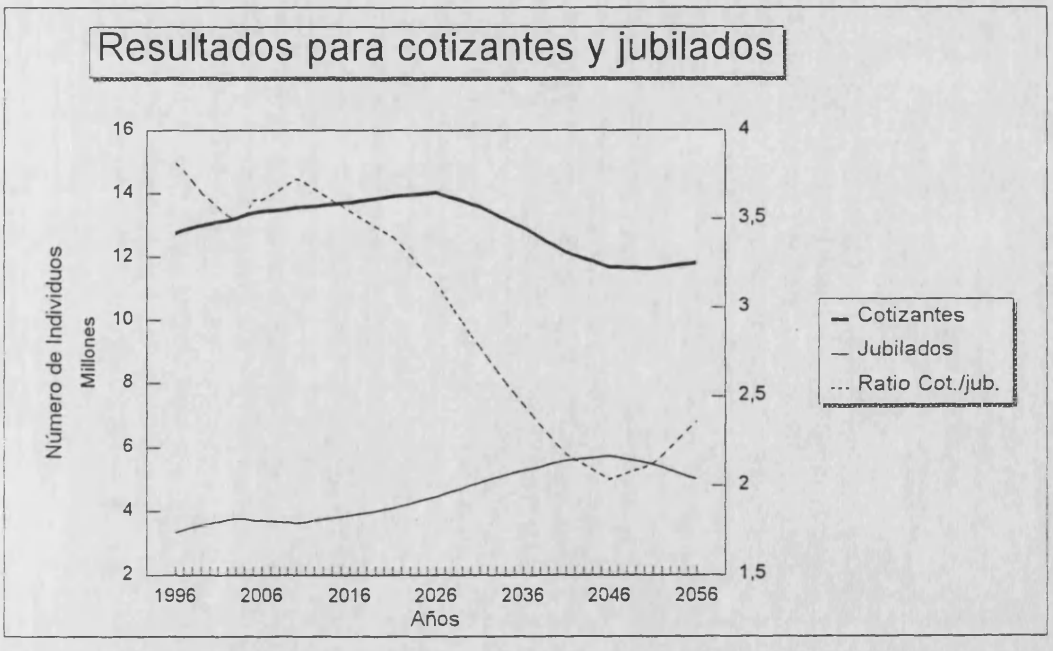

Gráfico VI.3.: Proyección de cotizantes y jubilados

Valores de los parámetros

En cuanto a la rentabilidad neta del capital se establecerá como hipótesis para todo el periodo un valor del 3% en términos reales, aunque se realizarán análisis de sensibilidad para valores del 2% y del 4%. La rentabilidad media obtenido en los últimos años por las instituciones de inversión colectiva en fondos de inversión o de pensiones ha superado esta cifra, por ello estamos considerando en nuestra hipótesis que se avanza hacia un periodo de menores tipos de interés reales debido a que el proceso de unión europea debe llevar a tipos de interés y tasas de inflación más moderados. El valor elegido es habitual en estudios a largo plazo en España, por ejemplo en Barea (1995).

La revalorización de los salarios o bases de cotización en términos reales ha sido algo superior al 2% anual en los últimos años, dato que es similar a la revalorización de la pensión media de jubilación en términos reales a pesar de que cada variable depende de factores distintos. Las bases de cotización aumentan en términos reales debido principalmente a un aumento de los salarios reales (convenios colectivos, antigüedad, deslizamientos, etc.) y a una aproximación entre salarios y bases de cotización. La pensión media aumenta por las diferencias entre las altas y las bajas (efecto sustitución). Se tomará como hipótesis un crecimiento de las bases de cotización del 1,5% real anual para todo el periodo[82], lo cual refleja un agotamiento en el acercamiento de los salarios a las bases de cotización y una cierta moderación salarial[83].

Los últimos parámetros que faltan por determinar son los valores deseados para el tipo de cotización y para la tasa de reemplazo de pensiones. Estos valores

[82]El incremento para 1996 es el mismo que para 1995 (4,51% nominal, como más tarde se razona).

[83]Si se supone que 2/3 de la productividad va a salarios reales se está considerando un aumento en la productividad del 2,25% para todo el periodo. Si este valor se suma al crecimiento del número de ocupados se obtiene la hipótesis implícita de crecimiento del PIB: 2,76% en 1996-2006, 2,47% en 2006-2026 y 1,68% en 2026-2056. El reciente informe de la Fundación BBV determina las variables de forma inversa: primero se supone un crecimiento del PIB del 3,5% hasta el 2000 y del 3% en adelante; después, se supone una productividad del 2,5%; y, por último, se deduce la evolución del número de ocupados.

son medios y por tanto se razona como si todos los cotizantes tuvieran el mismo tipo de cotización y todos los pensionistas la misma tasa de reemplazo. Para obtener estos valores medios deseados se pueden adoptar distintos criterios. La alternativa que vamos a seguir consta de los siguientes pasos:

1º) Se calculan los valores de partida (esperados para 1995) del tipo de cotización y de la tasa de reemplazo de pensiones. En nuestro caso, las variables que los determinan son las siguientes:

- Gasto en pensiones de jubilación: considera las transferencias a familias por pensiones de jubilación según el presupuesto de 1995 (de todos los regímenes y sólo las contributivas). No se tiene en cuenta, por tanto, ni las pensiones no contributivas ni los gastos de gestión, administración y control.

- Número medio de pensiones de jubilación contributivas: esta cifra no aparece en el presupuesto, pero sí el número estimado a principios y a final de año, por lo que se puede deducir el incremento esperado que es de 2,32%. Aplicando este porcentaje al número medio mensual de pensiones de 1994 se obtiene la cifra esperada para 1995.

- Número de cotizantes: se toma el número medio mensual real de 1994 más un 2,5%[84].

- Base media de cotización (salario): se calcula a través de una media de las bases de cotización esperadas de cada régimen ponderada por

[84]Este porcentaje representa el aumento medio observado durante el primer semestre de 1995 y es similar a la proyección realizada para el aumento de la población ocupada según el grupo de expertos de previsión económica del Ministerio de economía que lo cifra en un 2,4% (Síntesis de indicadores de coyuntura, Julio de 1995).

el número de cotizaciones[85].

Los valores de cada variable y los resultantes para el tipo de cotización y la tasa de reemplazo de pensiones para 1995 se resumen en la tabla VI.9..

	Datos 1995
Gasto en pensiones de jubil. (mill.) (a)	3.369.201
Número medio de cotizantes (b)	12.347.465
Número medio de pensionistas (c)	3.250.036
Base media de cotización (d)	1.763.383
Tipo de cotización medio (a/(b·d))	15,47%
Tasa de reemplazo media (a/(c·d))	58,79%

Tabla VI.9.: Cálculo de los parámetros de partida[86].

2°) Se resuelven a continuación dos problemas de C.O. discreto previos que recogen el esfuerzo máximo que debe recaer sobre cotizantes y pensionistas respectivamente para hacer frente al cambio demográfico si se permite la creación de un fondo de capital sin limitaciones y no se produce ninguna otra reforma. Llamando a los parámetros de partida anteriores c^p y d^p, los Problemas [III] y [IV]

[85]Este cálculo se realiza para 1994 sin tener en cuenta la cotización por jornadas reales y teóricas del régimen especial agrario (Anexo al informe económico-financiero del proyecto de presupuestos de la SS. 1994). El resultado se aumenta en un 4,51% para 1995 que corresponde a un aumento del volumen de cotizaciones normales de empresarios y trabajadores del 6,61% (tras homogeneizar la cifra del 94 y del 95) descontando el aumento previsto del número de cotizaciones del 2% (Presupuesto de la SS. Cifras y datos. 1995).

[86]Si el modelo incluyera también los gastos generales imputables a las pensiones de jubilación contributivas los parámetros de partida habría que recalcularlos. Lo mismo ocurriría si se deseara incluir las pensiones no contributivas.

siguientes dan como solución respectiva el tipo de cotización (c^m) y la tasa de reemplazo de pensiones (d^m) que se debería implantar durante todo el periodo para que el otro parámetro se mantuviera al nivel de partida y se hiciera frente al cambio demográfico esperado (se trata de porcentajes de máximo esfuerzo).

$$Min \quad \sum_{t=1}^{61} (cW_tC_t - c^PW_tC_t)^2$$

s.a: **[III]**

$$K_{t+1} - K_t = rK_t + cW_{t+1}C_{t+1} - d^PW_{t+1}P_{t+1} \qquad t=0,...,60$$

$$K_0=0 \ , \ K_t \geq 0 \qquad t=1,...,61$$

$$Min \quad \sum_{t=1}^{61} (dW_tP_t - d^PW_tP_t)^2$$

s.a: **[IV]**

$$K_{t+1} - K_t = rK_t + c^PW_{t+1}C_{t+1} - dW_{t+1}P_{t+1} \qquad t=0,...,60$$

$$K_0=0 \ , \ K_t \geq 0 \qquad t=1,...,61$$

Comparando las soluciones a ambos problemas con los valores iniciales de los parámetros se observa la magnitud del esfuerzo que debiera recaer sobre cada colectivo (cotizantes o pensionistas) si el otro se mantuviera al nivel de partida.

Las soluciones se han obtenido utilizando software estándar de P.N.L. (programa *LINGO* en su versión 2.1 de 1995 para problemas de grandes dimensiones) y para cada hipótesis sobre la fase intermedia del ciclo. Las tablas VI.10. y VI.11. recogen los valores óptimos de los parámetros. Unas características adicionales de las soluciones son que al resolver el Problema **[III]** se acumula un fondo de capital superior en cada año al que se consigue al solucionar el Problema

[IV], pero en ambos casos el importe máximo del fondo tiene lugar el mismo año y es igual a cero al finalizar el horizonte de planificación.

	c^m (c^p=15,47%)	d^m (d^p=58,79%)
Tasa de paro alta (21%)	20,03%	45,42%
Tasa de paro media (19%)	19,53%	46,57%
Tasa de paro baja (17%)	19,06%	47,72%

Tabla VI.10.: Tipos de cotización y tasas de reemplazo de pensiones de máximo esfuerzo en función de la tasa de paro para un rendimiento del capital del 3% en términos reales.

	c^m	d^m
r=2%	20,22%	44,99%
r=4%	18,92%	48,09%

Tabla VI.11.: Tipos de cotización y tasas de reemplazo de pensiones de máximo esfuerzo en función del rendimiento del capital para una tasa de paro del 19%.

Estas tablas dan, como resultado adicional, una aproximación a la sensibilidad de los tipos máximos ante variaciones en la tasa de paro y en la rentabilidad del capital. Se observa que variaciones en 2 puntos en la tasa de paro modifican el tipo de cotización máximo en 0,5 puntos aproximadamente (relación

directa) o la tasa de reemplazo máxima en 1,15 puntos (relación inversa). Por otro lado, variaciones en 1 punto en el rendimiento real del capital afectan en 0,65 puntos al tipo de cotización máximo (relación inversa) o en 1,55 puntos a la tasa de reemplazo máxima.

En términos relativos, la solución bajo las hipótesis medias indica que el tipo de cotización debe aumentar un 26,2% o, alternativamente, la tasa de reemplazo bajar un 20,8% para hacer frente al cambio demográfico si se crea un fondo de capital sin limitaciones. Estos cambios son muy inferiores a los que se exigirían sin fondo de capital (aumento máximo del tipo de cotización del 86,8% o disminución máxima de la tasa de reemplazo del 46,6%) aunque, por contra, el cambio debe ser inmediato.

3°) Con los valores obtenidos debe haber a continuación una decisión subjetiva de cómo repartir el ajuste y obtener así los porcentajes deseados. Se trata de elegir subjetivamente un valor (β) del intervalo [0,1] y obtener los porcentajes deseados como combinación lineal convexa de los valores de partida y los de máximo esfuerzo:

$$(c^*,d^*) = \beta(c^m,d^p)+(1-\beta)(c^p,d^m)$$

De esta manera cuanto más próximo esté β a 0 más esfuerzo recae sobre los pensionistas y cuanto más próximo esté a 1 más recae sobre los cotizantes. Si la decisión es que el reparto entre ambos grupos sea equilibrado el valor de β es 0,5 y los valores deseados correspondientes son $c^*=17,50\%$ y $d^*=52,68\%$.

VI.3.3.2.- Aplicación del modelo y resultados

Con las variables exógenas y los parámetros se pasa a resolver el modelo general. El modelo considera que hay limitaciones en la cuantía del fondo de capital y para adaptarse a ellas los tipos de cotización y tasas de reemplazo son variables

en el tiempo. La solución del modelo determina ambos porcentajes en cada periodo (variables de control) y adicionalmente la cuantía del fondo de capital (variable de estado) en términos reales. El modelo se resuelve para distintos supuestos acerca de la limitación del fondo de capital (parámetro a), empezando por el caso no restringido ($a=\infty$) cuya solución para el tipo de cotización y tasa de reemplazo es, obviamente, igual en cada periodo a los valores deseados. También incluimos la solución que se obtendría bajo un sistema de reparto ($a=0$) para comparar mejor el efecto que produce la consideración del fondo de capital.

El modelo a resolver es el descrito en el epígrafe VI.3.2. (Problema **[II]**):

$$Min \quad \sum_{t=1}^{61} [(c_t W_t C_t - c^* W_t C_t)^2 + (d_t W_t P_t - d^* W_t P_t)^2]$$

$$s.a:$$

$$K_{t+1} - K_t = rK_t + c_{t+1} W_{t+1} C_{t+1} - d_{t+1} W_{t+1} P_{t+1} \qquad t=0,...,60 \qquad \text{[II]}$$

$$K_0 = 0 \ , \ K_t \geq 0 \qquad t=1,...,61$$

$$K_t \leq a d_t W_t P_t \qquad t=1,...,61$$

Los valores de los parámetros son: $c^*=17{,}50\%$, $d^*=52{,}68\%$, $r=3\%$. Sobre a se toman distintas hipótesis y se escoge la hipótesis media para el número de cotizantes.

Un resumen de los resultados que se obtienen para cada variable del problema y para cuatro hipótesis sobre el parámetro a se recoge en los gráficos VI.4., VI.5. y VI.6..

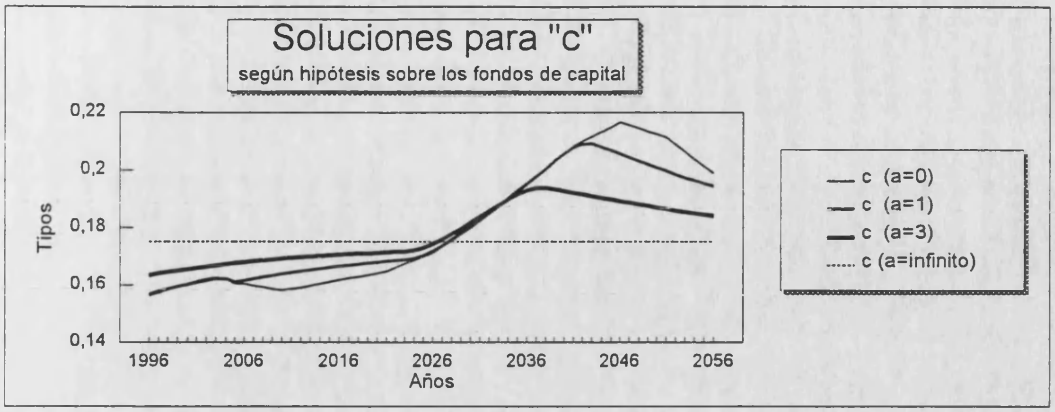

Gráfico VI.4.: Trayectorias óptimas de los tipos de cotización.

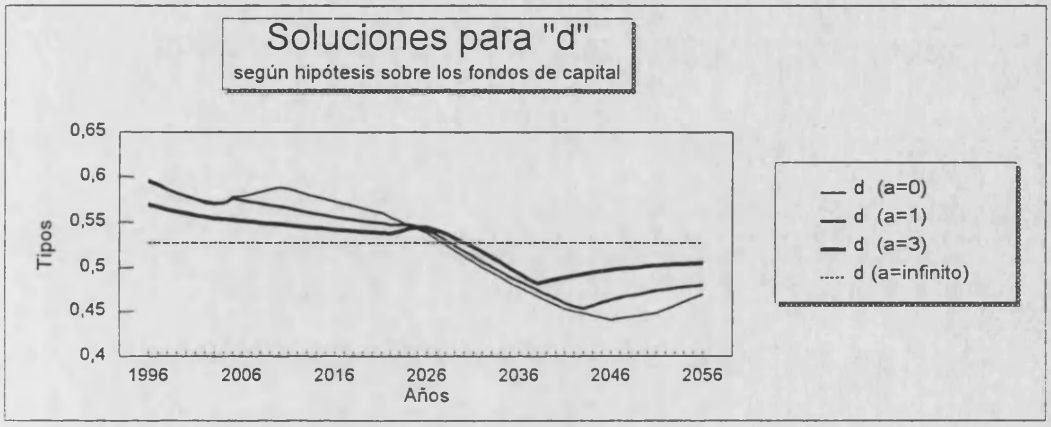

Gráfico VI.5.: Trayectorias óptimas de las tasas de reemplazo.

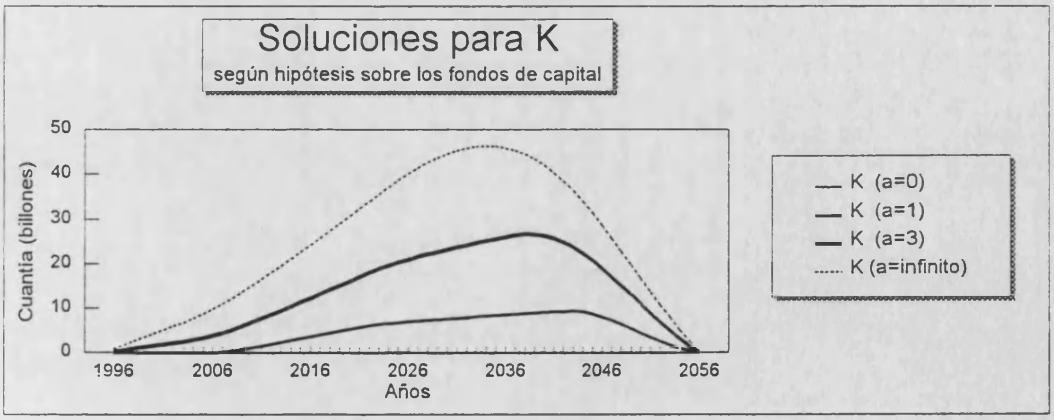

Gráfico VI.6.: Trayectorias óptimas del fondo de capital.

La tabla VI.12. recoge los principales resultados en términos absolutos: el ajuste inicial (diferencia entre los valores óptimos de 1996 y los iniciales de 1995), la magnitud total del ajuste (diferencia entre los valores óptimos del año de máximo esfuerzo y los de 1995) y la cuantía máxima del fondo de capital (en pesetas de 1996).

		$a=0$	$a=3$	$a=\infty$
Ajuste	c	0,18	0,90	2,03
inicial	d	0,94	-1,78	-6,11
Ajuste	c	6,22	3,93	2,03
total	d	-14,64	-10,55	-6,11
Cuantía máx. K		0	26,73 bill.	46,35 bill.

Tabla VI.12.: Principales resultados de la aplicación del modelo en términos absolutos.

La tabla VI.13. sintetiza los resultados en términos relativos: ajuste inicial, variaciones interanuales desde 1996 hasta el año de máximo esfuerzo y cuantía máxima del fondo de capital en % del PIB si este crece al 2% anual.

	$a=0$	$a=3$	$a=\infty$
Ajuste c	1,18%	5,79%	13,12%
inicial d	1,60%	-3,03%	-10,39%
Ajuste c	0,65%	0,41%	0%
interanual d	-0,60%	-0,40%	0%
K max. (% P.I.B.)	0	16,18%	30,36%

Tabla VI.13.: Principales resultados de la aplicación del modelo en términos relativos.

Las soluciones muestran un intercambio entre magnitud y rapidez del ajuste. Así, la solución de reparto puro ($a=0$) realiza el ajuste al mismo ritmo que evoluciona la demografía y, por tanto, es un ajuste gradual pero profundo. En el caso opuesto, la existencia de reservas financieras no limitadas requiere un ajuste inmediato aunque de magnitud menor[87]. Si se limita el fondo de capital las soluciones son intermedias (el año de máximo esfuerzo se sitúa en el 2038, ocho años antes que bajo el sistema de reparto puro).

El modelo se puede resolver incorporando algunas de las modificaciones que se han planteado en el epígrafe VI.3.1. o cambiando alguno de los valores de los parámetros sujetos a una decisión política. A continuación se analizan otros tres modelos con el objetivo de comparar los distintos tipos de ajuste que surgen. Los modelos modificados tratan de incluir los siguientes efectos:

- Exigir un ajuste menor en los primeros periodos que en los últimos. Esto se consigue penalizando más las desviaciones de los últimos periodos. Para ello se introduce una función creciente con el tiempo dentro del sumatorio de la función

[87]La dificultad de esta política, aparte de exigir una reforma radical, está en los posibles efectos económicos adversos de un fondo de capital excesivamente dimensionado que puede llevar a una economía sobrecapitalizada.

objetivo, quedando:

$$\sum_{t=1}^{61} (1+b)^{t-1} \left[(c_t W_t C_t - c^* W_t C_t)^2 + (d_t W_t P_t - d^* W_t P_t)^2\right]$$

- Cambiar la distribución del ajuste entre cotizantes y pensionistas. En concreto, se ha resuelto el modelo exigiendo un mayor ajuste a los pensionistas que a los cotizantes. En términos matemáticos esto se logra disminuyendo el valor del parámetro β.

- Exigir trayectorias estables a partir de cierto periodo para las variables de control. Esto responde al objetivo de llegar a una fase de madurez con tipos de cotización y tasas de reemplazo constantes para dotar al sistema de estabilidad. Si el periodo a partir del cual se exige la estabilidad es t', el modelo a resolver es:

$$Min \quad \sum_{t=1}^{61} \left[(c_t W_t C_t - c^* W_t C_t)^2 + (d_t W_t P_t - d^* W_t P_t)^2\right]$$

$$s.a:$$

$$K_{t+1} - K_t = rK_t + c_{t+1} W_{t+1} C_{t+1} - d_{t+1} W_{t+1} P_{t+1} \qquad t=0,...,60$$

$$K_0 = 0 \ , \ K_t \geq 0 \qquad t=1,...,61 \qquad [V]$$

$$K_t \leq a d_t W_t P_t \qquad t=1,...,61$$

$$c_{t+1} - c_t = 0 \qquad t \geq t'$$

$$d_{t+1} - d_t = 0 \qquad t \geq t'$$

Los valores numéricos concretos que se han tomado son: $b=0{,}02$ en el primer modelo modificado, $\beta=0{,}25$ en el segundo y $t'=31$ (año 2026) en el tercero. La solución de los tres modelos y la comparación con el modelo original, todo ello para $a=3$, se recoge en los gráficos VI.7 a VI.9.

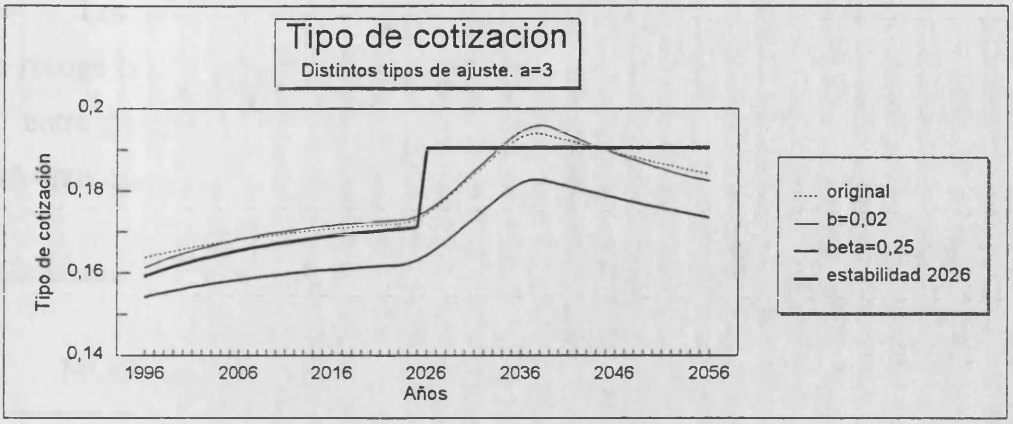

Gráfico VI.7.: Trayectorias óptimas de los tipos de cotización.

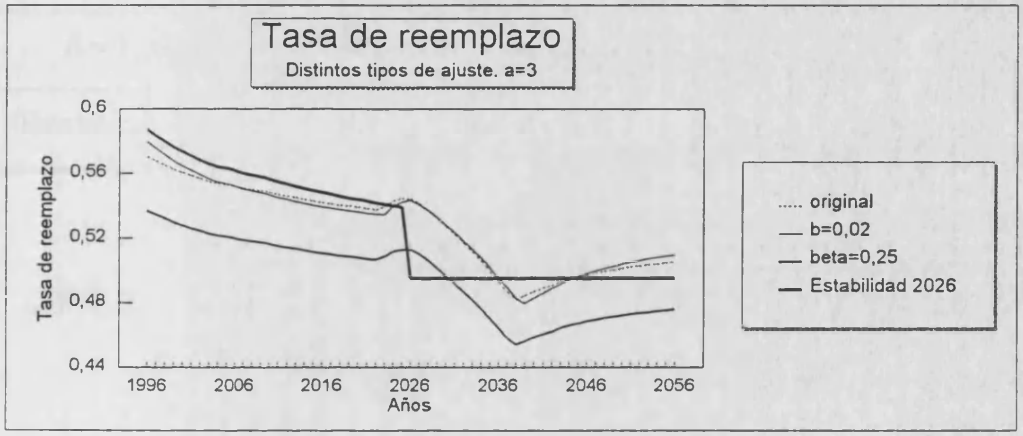

Gráfico VI.8.: Trayectorias óptimas de las tasas de reemplazo.

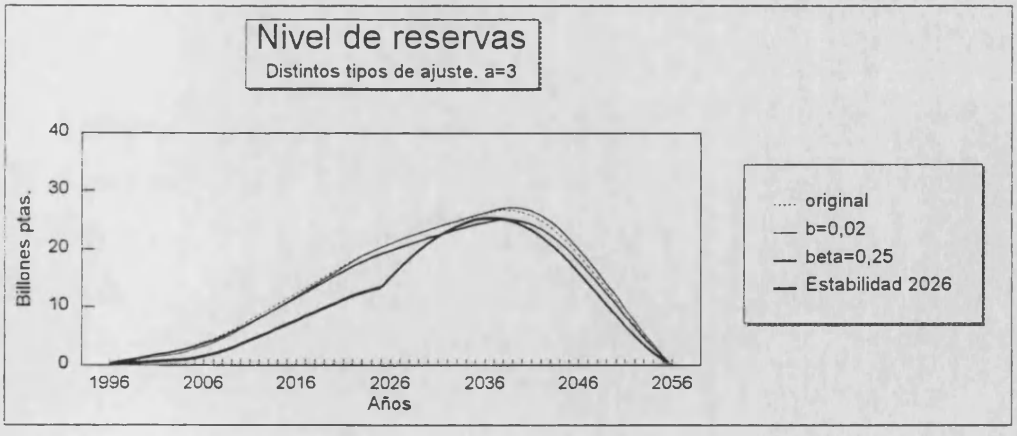

Gráfico VI.9.: Trayectorias óptimas del fondo de capital.

Los resultados en términos relativos se presentan en la tabla VI.14.. En ella se recoge la variación del año de máximo esfuerzo respecto de los valores iniciales y, entre paréntesis, la variación interanual desde 1996 hasta el año de máximo esfuerzo que es el 2038 excepto en el caso de trayectorias estables que es el 2026.

MODELOS	Incremento tipo de cotización	Disminución tasa de reemplazo	Cuantía máxima K como % del PIB
Original	25,39%(0,41%)	17,94%(0,40%)	16,56%
b=0,02	26,72%(0,47%)	18,43%(0,45%)	16,61%
β=0,25	18,12%(0,41%)	22,70%(0,40%)	15,60%
Estabilidad 2026	23,12%(0,60%)	15,74%(0,56%)	16,13%

Tabla VI.14.: Principales resultados de los modelos modificados con a=3.

El resultado de los modelos permite encontrar la solución óptima según el tipo de ajuste deseado:

- Bajo un sistema de reparto puro el ajuste es gradual, lento (hasta el año 2046) y de magnitud amplia: disminución máxima de 14,6 puntos en la pensión media como porcentaje del salario medio y aumento en 6,2 puntos en el tipo de cotización. Estas diferencias se recortarían al final del periodo considerado al mejorar las curvas demográficas.

- Con fondos de capital se puede alterar esta forma de realizar el ajuste. En primer lugar, cuantas menos limitaciones se impongan al volumen del fondo más rápido será el ajuste y menor su magnitud máxima (tablas VI.12. y VI.13. y gráficos VI.4. a VI.6.). En segundo lugar, si se exige que el ajuste finalice en un determinado año (a partir del cual las tasas y los tipos

son estables), el ajuste se realizará de forma gradual hasta dicho año donde se producirá un ajuste brusco, que será mayor cuanto más corto sea el periodo de ajuste (gráficos VI.7. a VI.9.). Adicionalmente, el ajuste puede repartirse de múltiples formas entre cotizantes y pensionistas (parámetro β) y también entre generaciones (parámetro b).

Estas consideraciones son importantes porque los instrumentos que se utilicen para reformar el sistema de pensiones son adecuados o no en función del tipo de ajuste deseado.

Más que los valores concretos de las variables hay que observar, como ya hemos mencionado antes, su tendencia dinámica y la comparación entre los resultados bajo distintos escenarios. Las conclusiones relevantes son, en consecuencia, la magnitud aproximada del ajuste que requiere el cambio demográfico y cómo este ajuste se puede suavizar a través de la creación de reservas financieras.

No se han tenido en cuenta ciertos efectos económicos de la acumulación de capital, es decir, variables como el aumento en los salarios, la tasa de rentabilidad del capital o la tasa de paro se han supuesto exógenas y no en función del nivel de capitalización, como parece aconsejar la teoría (por ejemplo en Burbidge, 1983; y López García, 1991). Creemos, sin embargo, que esto no afecta sustancialmente a la tendencia observada por las variables, debido a dos motivos:

- Primero, porque si se limita el fondo de capital se limita, a su vez, este tipo de efectos económicos.

- Segundo, porque la consecuencia de estos efectos a nivel financiero hubiera tendido a compensarse según los estudios teóricos: a mayor capitalización, menores tipos de interés pero también menores tasas de paro.

Los comentarios anteriores no suponen, a nuestro juicio, un obstáculo para dar validez al modelo y, por tanto, al uso de la teoría del C.O. en tiempo discreto

para estudiar el problema de la financiación de las pensiones de jubilación públicas. A partir de ahí, el modelo puede servir de base para hacer estudios similares a éste que partan de hipótesis distintas (por ejemplo, con otras proyecciones para el número de cotizantes y de pensionistas) o que reduzcan el problema al caso de las pensiones del Régimen General de la S.S. o que lo amplíen al caso de otras clases de pensiones, etc..

VI.3.4.- Instrumentalización de la solución

La decisión de establecer un nivel de reservas para hacer frente al cambio demográfico debe ir acompañada de otras decisiones que tengan como objetivo el adecuar los tipos de cotización y las tasas de reemplazo a los valores óptimos de cada periodo. Estas variables, pese a ser variables de control en el contexto del modelo matemático, no son instrumentos reales de decisión en el ámbito de la política económica, sobre todo en lo que se refiere a la tasa de reemplazo. Por tanto, la solución del modelo se debe interpretar como un objetivo intermedio a alcanzar, y lo que queda por determinar es el conjunto de instrumentos concretos que se deben utilizar para conseguir el objetivo. Los instrumentos que vamos a analizar son bastante citados en la literatura[88] y también aparecen, aunque no concretados, en las conclusiones del Pacto de Toledo (Vicente, 1995).

El tipo de cotización se divide en la parte de empresarios, trabajadores y Estado (vía impuestos). En ausencia de reformas, el tipo se mantendría constante al nivel de 1995 lo cual no es compatible con la solución del modelo. El principal instrumento que se maneja para aumentar el tipo de cotización es la **profundización en la reforma de la estructura financiera de la S.S.**. Esto significa distinguir las fuentes de financiación de la parte contributiva (sustitución de rentas) y de la no contributiva (redistribución de renta). La idea es que la parte de sustitución de rentas

[88]Existe abundante literatura divulgativa sobre propuestas de reforma de la S.S., aunque una limitación bastante general de estos trabajos es que no suelen cuantificar los efectos de las reformas que plantean. De entre ellos cabe citar la monografía dirigida por José Barea (Barea y otros, 1995), que destaca, no sólo por ser reciente, sino por el detalle con que concreta las propuestas.

se financie con cotizaciones y la parte redistributiva con impuestos, lo cual supondría un aumento en la aportación del Estado dado el superávit actual de la parte contributiva[89]. Los efectos cuantitativos de esta propuesta sobre el tipo de cotización se recogen a continuación, distinguiendo dos niveles de actuación.

En un primer nivel, dentro de las pensiones de jubilación, la parte de complementos a mínimos que todavía se financia con cotizaciones debe pasar a financiarse con impuestos. En el presupuesto de 1995 esto significaría una aportación extraordinaria del Estado de aproximadamente 171.000 millones de ptas.. Si esta aportación se realiza sin disminuir la aportación de trabajadores y empresarios el resultado sería equivalente a un aumento del tipo de cotización desde el nivel inicial (15,47%) hasta el 16,26%.

En un segundo nivel, si las cotizaciones que en 1995 se destinan a financiar la Sanidad se sustituyen por impuestos y se desvían a la parte contributiva de la Seguridad Social, se dispondrá de recursos adicionales para financiar los gastos contributivos (de los cuales el 52,6% son pensiones de jubilación). La cuantía resultante es de 347.000 millones, que permite pasar a un tipo de cotización equivalente del 17,07%, siempre y cuando el aumento en la aportación del Estado se realice sin disminuir las cotizaciones[90].

Si se tiene en cuenta los dos niveles de racionalización comentados en los dos párrafos anteriores, los recursos adicionales que debe aportar el Estado vía mayores impuestos permite situar el tipo de cotización en el 17,85%. Esta reforma es adecuada para aplicarla a un modelo cuya solución óptima requiera un ajuste gradual. Veamos cuatro ejemplos de ajustes:

---

[89]El superávit de la parte contributiva se sitúa en 1,1 billones según Barea y otros (1995) en el presupuesto de 1994, aunque nuestros cálculos dan una cifra inferior (0,82 billones en 1994 y 0,62 en 1995) debido a distintas metodologías de cálculo. El Ministerio de Trabajo y Seguridad Social lo cifra en 0,19 billones en 1995 debido básicamente a que no incluye los complementos a mínimos como gasto no contributivo (Presupuestos de la Seguridad Social. Cifras y datos. 1995).

[90]Recientemente, el F.M.I. ha recomendado la creación de un impuesto para financiar los gastos sanitarios.

- En el caso de un sistema sin reservas financieras el aumento de impuestos se puede repartir hasta el año 2029, quedando pendiente una subida de 3,84 puntos en el tipo de cotización (hasta el 21,69% en el año 2046), pudiendo descender más tarde. Se cubre el 38,2% del ajuste necesario.

- En el caso de fondo de capital limitado (a=3), esta reforma se puede implantar de forma progresiva (hasta el año 2028 aproximadamente). A partir de entonces el tipo de cotización debe aumentar todavía 1,55 puntos (hasta el 19,40% en el año 2038), aunque luego puede disminuir. Se cubre el 60,5% del ajuste.

- Si la decisión es que recaiga un mayor ajuste sobre los pensionistas que sobre los cotizantes (β=0,25), esta reforma es prácticamente suficiente para realizar el ajuste de la parte de cotizantes (cubre un 84,7%) y sólo se necesita un aumento de 0,43 puntos.

- Si se desean tipos estables a partir del año 2026 (con a=3), esta reforma es suficiente para cubrir todo el periodo transitorio y una parte del ajuste brusco de ese año, siendo necesario un aumento adicional en dicho año de 1,2 puntos (hasta el 19,05%). Se cubre el 66,4% del ajuste.

La tasa de reemplazo depende de la pensión media y de los salarios. En ausencia de reformas es muy posible que tienda a ser creciente ya que la pensión media crecerá más que los salarios. Esto es así porque se ha supuesto moderación salarial y por tanto el crecimiento de los salarios en el pasado (que afecta al crecimiento futuro de la pensión media vía pensión media de las altas) es superior al crecimiento en el futuro. Esto contrasta con la solución óptima, que exige disminuciones en la tasa de reemplazo, por lo que hay que analizar las propuestas de reforma.

Una de las reformas que se maneja es el **aumento en la edad de jubilación**[91,92]. Esta propuesta no debe entenderse como un aumento obligatorio en la edad de jubilación sino como un retraso en el comienzo del pago de las pensiones públicas. Así, un trabajador de 65 años puede optar por seguir trabajando los años adicionales o por retirarse, teniendo en cuenta que debe disponer de un ahorro privado suficiente hasta que empiece a percibir las pensiones públicas unos años después. Aparte de otras implicaciones, los efectos de esta medida sobre la financiación de las pensiones son positivos por cuanto disminuye el número de jubilados que cobran pensión, pero en el otro sentido también es verdad que se puede producir una disminución del número de cotizantes[93] o un aumento en la pensión media[94].

Un análisis simple sería aquel en el que sólo se tiene en cuenta el efecto sobre el número de jubilados. Los cálculos realizados dan como resultado una disminución del número de perceptores del 5,21% por cada año que se retrase el derecho a percibir la pensión de jubilación (antes de la desaparición de las jubilaciones anticipadas este porcentaje será aproximadamente de 4,7%)[95]. Con este análisis simple, las tasas de reemplazo óptimas quedan aumentadas en la misma medida que disminuye el número de pensionistas, lo cual permite que el ajuste que

[91]La propuesta concreta en Barea y otros (1995) es "aumentar la edad de jubilación a 70 años, excepto para aquellos trabajos que requiriesen una edad menor de jubilación" (pág. 51). La cuantificación que en ese trabajo se realiza de dicha propuesta se refiere al ratio años de pensionista por años de cotizante que pasaría de 0,37 a 0,2.

[92]Villagarcía (1995) ha estudiado las variables socioeconómicas que influyen en el paso a la jubilación a través de un modelo econométrico. Sus resultados son que existen ciertos colectivos, como los trabajadores por cuenta propia y los que tienen un alto nivel de educación-renta, que prolongarían su vida laboral si ello fuera posible.

[93]Si los mayores de 65 años que siguen trabajando no cotizan y, como parece lógico, ocupan un puesto de trabajo que iría a una persona más joven

[94]Si en el cálculo de la pensión inicial se considera la base de cotización de los años adicionales de trabajo y dicha base es superior a la de antes.

[95]En estos cálculos se ha utilizado la proyección del I.D. para los años 1996-2026 y el cociente entre el porcentaje de personas entre 65 y 69 sobre los mayores de 65 años y los jubilados entre 65 y 69 años sobre el total de jubilados mayores de 65 años a 1-1-95.

requiere la solución óptima no se traduzca en menores pensiones.

Este tipo de política provoca saltos bruscos en la tasa de reemplazo por lo que no se acomoda bien a las soluciones óptimas que requieren ajustes graduales y sí, en cambio, a las que incluyen ajustes repentinos. Veamos dos ejemplos:

- Si se opta por una reforma radical desde el primer periodo permitiendo un nivel de reservas sin limitaciones, la tasa óptima es igual a la deseada (52,68%) y este descenso respecto de la tasa inicial (que es del 10,4%) se conseguiría aumentando en dos años la edad de jubilación sin modificar la cuantía de las pensiones. El inconveniente es que este tipo de reforma debería conocerse con antelación por lo que resulta el instrumento ideal para los modelos cuyas soluciones óptimas contengan ajustes bruscos en periodos intermedios, es decir, para los modelos que exigen acabar con una tasa de reemplazo estable.

- Efectivamente, si se exige estabilidad a partir del año 2026 (con $a=3$) el descenso repentino en la tasa de reemplazo en ese año es de 4,41 puntos (8,2%), lo cual se consigue con holgura aumentando dos años la edad de jubilación: se debería anunciar, por tanto, un aumento en dos años en la edad de jubilación para los nacidos después de 1960. De esta manera, la reforma ya no se puede considerar radical ya que hay tiempo para que los actuales trabajadores ahorren lo suficiente si no quieren prolongar la edad de retiro esos dos años.

El análisis realizado ha sido el más simple posible, dado que el alargamiento en la edad de jubilación puede llevar a un aumento en la pensión inicial o a una disminución de cotizaciones. Por otra parte, la dinámica de la tasa de reemplazo, como se ha comentado, es creciente en ausencia de reformas, por lo que el resultado obtenido sólo es válido si se llevan a cabo otras reformas cuya consecuencia sea el mantenimiento de la tasa de reemplazo al nivel de partida. Estas reformas adicionales son las que pasamos a analizar.

Efectivamente, otros instrumentos que se manejan por el lado de las prestaciones se pueden englobar en uno más general: la disminución de la pensión media. Existen dos formas posibles: disminuir la pensión inicial de entrada al sistema o revalorizar por debajo de la inflación las pensiones existentes. Los instrumentos concretos que se citan son, entre otros, los siguientes:

- **Ampliar el periodo de referencia para el cálculo de la base reguladora.** Se trata de calcular la base reguladora, que sirve de base para calcular la pensión de las altas, en función de la base de cotización de un periodo de tiempo más amplio que el actual (ocho años). Esto, además de aumentar la relación entre lo cotizado y lo recibido, da lugar a una reducción de la base reguladora y por tanto de la pensión media de las altas. Este mismo instrumento se puso en práctica tras la reforma de 1985 (se paso de dos a ocho años) y el resultado fue un aumento en la pensión media de las altas en 1986 de sólo el 0,76% nominal.

- Asignar el porcentaje de la base reguladora que supone la pensión de las altas de manera que se reduzca en términos medios. Se enuncian dos posibilidades: **aumentar el número de años para llegar al 100% de la base reguladora** (actualmente en 35) y **pasar a un sistema proporcional** (actualmente con 15 años cotizados se obtiene el 60% de la base reguladora y los otros 20 años dan lugar al otro 40%).

- **Revalorizar las pensiones menos que la inflación.** Esta propuesta tendría efectos muy potentes porque afectaría a todas las pensiones y no sólo a las altas pero también enormes costes desde un punto de vista político. Sin embargo, existe cierto margen para llevarla a cabo si se introducen elementos redistributivos como la congelación de las pensiones máximas, menor aumento de las pensiones más elevadas, etc..

La forma de cuantificar este tipo de reformas parte de la ecuación (23) que relaciona la pensión media con la dinámica de altas y bajas y con la revalorización de las existentes.

$$p_{t+1}P_{t+1}=p_tP_t(1+i)+a_{t+1}A_{t+1}-b_{t+1}B_{t+1} \qquad (23)$$

donde las nuevas variables son la pensión media de las altas (a_t), el número de altas (A_t), la pensión media de las bajas (b_t), el número de bajas (B_t) y la revalorización en términos reales de las existentes (i).

Las dos primeras reformas comentadas inciden sobre la forma de cálculo de la pensión media de las altas (a_t) reduciéndola, el tercer tipo de reforma disminuye el parámetro i. En todos los casos se produce una disminución de la pensión media global. La tabla VI.15 muestra en tres columnas el impacto de diversas reformas. La primera recoge la variación que se produciría en la pensión media de las altas del año 96 al implantar cada reforma respecto a la situación actual. La segunda columna indica, en términos nominales, la variación de la pensión media de las altas de 1996 respecto a la de 1995 con cada reforma (una posible limitación de tipo político a las reformas es que esta pensión no disminuya en términos nominales como en la reforma de 1985). La última columna muestra el aumento de la pensión media global en términos reales del año 1996 respecto a 1995, con el objetivo de compararla con la solución óptima. También se han incluido los valores si no se introduce ninguna reforma (sistema actual). Las cifras obtenidas son sólo una primera aproximación al verdadero efecto de cada reforma ya que los cálculos se han basado en supuestos generales y en datos publicados oficialmente.

Sistema	Variación en a_t respecto al actual sistema	Variación en a_t de 1996 respecto a 1995 (nominal)	Variación en p_t de 1996 respecto a 1995 (real)
actual	0%	6,34%	2,26%
pr35[96]	-4,51%	1,54%	1,83%
pr40	-11,41%	-5,79%	1,18%
actual 40[97]	-7,02%	-1,13%	1,59%
ba15[98]	-6,82%	-0,92%	1,61%
ba30	-18,98%	-13,85%	0,45%
pr40-ba30	-28,22%	-23,67%	-0,43%
" i=-0,5%	-28,22%	-23,67%	-0,91%

Tabla VI.15.: Porcentajes de aumento de la pensión media de las altas y global según varias hipótesis. Elaboración propia.[99]

[96]El sistema pr35 (pr40) determina el % de la base reguladora que supone la pensión inicial de una forma proporcional con 35 años (40 años) cotizados para alcanzar el 100%. Los cálculos se han realizado con los datos de los años cotizados de las altas con hecho causante en 1992.

[97]Sistema actual modificado para alcanzar el 100% con 40 años cotizados. Los 15 primeros permiten obtener el 50% de la base reguladora y los otros 25 dan lugar al 50% restante de forma proporcional. Es una reforma intermedia a la de pasar directamente a un sistema proporcional.

[98]El sistema ba15 (ba30) calcula la base reguladora en función de los últimos 15 (30) años cotizados actualizando las bases de cotización excepto en los dos últimos años. Los cálculos se han realizado como si la media de los nuevos pensionistas hubieran tenido una base de cotización que se revalorizara un 2% anual en términos reales y utilizando un 7% en la actualización (estos porcentajes son aproximadamente las medias anuales de aumento de las bases de cotización en los últimos 15 años).

[99]Los porcentajes se obtienen utilizando la fórmula (23). La pensión media de 1995 se ha calculado aumentando la media de 1994 en un 6,84% anual (este porcentaje es el que se ha dado en el primer semestre de 1995). Para 1996 el aumento ha sido el mismo menos medio punto ante la previsión de una menor inflación (4% frente a 4,5% en 1995). La pensión media de altas y bajas se calcula a partir de la pensión media manteniendo las proporciones existentes entre las tres a 1 de Julio de 1993. La dinámica de altas y bajas se obtiene con la proyección demográfica del I.D..

Con este análisis se tiene una idea del efecto cuantitativo de cada reforma en términos de aumento de la pensión media del primer año tras la reforma. Evidentemente se pueden combinar entre sí aunque en la tabla sólo aparezca la combinación entre los sistemas "pr40" y "ba30". Las reformas se deben introducir paulativamente y, por tanto, son instrumentos adecuados para realizar ajustes graduales. Se observa que bajo el sistema actual la pensión media crece más que los salarios reales (2,26% frente a 1,5%) por lo que la tasa de reemplazo es creciente. El sistema "pr40" llevaría a una tasa de reemplazo decreciente (1,18% frente a 1,5%), tal y como exige las soluciones óptimas[100], pero a costa de una disminución nominal de las pensiones de las altas, lo cual puede ser peligroso en base a los efectos políticos. La alternativa de revalorizar menos que la inflación las pensiones elevadas tampoco parece muy popular. La única salida para aliviar los efectos de las reformas es introducirlas gradualmente junto con el aumento en la edad de jubilación.

Lanzar una propuesta concreta válida para todo el periodo que combine las distintas reformas exige realizar una proyección de la pensión media a largo plazo, ejercicio éste realmente atractivo pero laborioso y sujeto a muchas incertidumbres. Haría falta disponer de la pensión media por edades (están publicadas por grupos quinquenales), de las tasas de mortalidad esperadas, además de la proyección de jubilados realizada y de hipótesis sobre aumentos reales futuros de la pensión media de las altas (que dependerían no sólo del aumento esperado de salarios sino también del efecto de la desaparición de las jubilaciones con coeficiente reductor y de la tendencia a completar cada vez más los periodos de cotización[101]).

El modelo elegido determina los instrumentos de reforma más adecuados, los

[100]Si la tasa de reemplazo óptima disminuye en cuantía x (en tanto por uno), la reforma debe ser tal que la pensión media crezca a la tasa $(1+\Delta w)(1-x)-1$ (en tanto por uno). Por ejemplo, una disminución de la tasa de reemplazo del 0,4% y unos salarios creciendo al 1,5% implican que la pensión media sólo podrá subir un 1,094% en términos reales.

[101]En Achurra y Quílez (1990) se cuantifican dichos efectos en 1989, llegándose a la conclusión que la pensión media de las altas está infravalorada aproximadamente un 9% por las jubilaciones anticipadas y un 8% por no estar jubilándose con 35 años cotizados.

cuales, a su vez, generan unos costes políticos que no hay que menospreciar. En consecuencia, los efectos políticos determinan indirectamente el modelo a seguir. Una opción razonable parece ser la creación de un fondo de capital moderado con estabilidad a partir del año 2026 ya que requiere un ajuste gradual en los primeros 30 años (1ª etapa) seguido de un ajuste brusco (2ª etapa) para el que se ha tenido tiempo suficiente de prepararse. Un valor del parámetro a igual a 3 con $\beta=0,5$ reparte prácticamente al 50% el ajuste entre ambas etapas y entre cotizantes y pensionistas.

Con estos valores de los parámetros el modelo pronostica un aumento del 23,12% en el tipo de cotización y una caída del 15,74% en la tasa de reemplazo. El incremento en el tipo de cotización sería cubierto en dos terceras partes por un aumento de impuestos para lograr el equilibrio de la parte contributiva y la no contributiva (aproximadamente 2 puntos de IVA), y el tercio restante con un aumento del tipo a cargo de empresarios y trabajadores (1,1 puntos de cotización). La mitad de estos aumentos se debe producir de forma gradual hasta el 2026 y la otra mitad de forma brusca en dicho año. El descenso en la tasa de reemplazo se repartiría a partes iguales entre menores pensiones (instauración progresiva del sistema "pr40-ba30" con moderación de las pensiones más altas hasta el 2026) y menos pensionistas (aumento de dos años en la edad de jubilación en el 2026).

Así pues, el mantenimiento del sistema público de pensiones puede asegurarse, no sin esfuerzo, con una adecuada combinación de reformas graduales y otras más bruscas pero anunciadas con suficiente antelación.

Esto es sólo un ejemplo de ajuste con el uso de unos determinados instrumentos de reforma, pero indica que si el tema se afronta de forma decidida y rápida se puede conseguir la viabilidad, no sólo financiera sino también política, del sistema público de pensiones.

CONCLUSIONES

Las conclusiones de la tesis recogen el grado de cumplimiento de los objetivos propuestos en su inicio, así como las limitaciones a las que están sujetos los resultados obtenidos y las posibilidades que quedan abiertas para futuras investigaciones.

El objetivo principal planteado ha sido presentar los aspectos más relevantes de la teoría del control óptimo en tiempo discreto de una forma homogénea y rigurosa para su posterior uso como instrumento matemático aplicado al planteamiento y resolución de modelos financieros relacionados con las pensiones de jubilación.

El objetivo a nivel teórico se ha cubierto fundamentalmente a lo largo del capítulo segundo en la medida en que las formulaciones recogidas relacionan la teoría del control óptimo discreto con la programación no lineal, más conocida y utilizada en el entorno de modelos económicos. Presentar las formulaciones del control óptimo discreto como un caso particular de la programación no lineal resulta ventajoso de cara a su comprensión y a su resolución a través de métodos numéricos más estandarizados. A su vez, el hecho de incluir el principio del máximo discreto como vía alternativa de trabajo ha permitido observar las relaciones entre ambos procedimientos lo cual es enriquecedor desde un punto de vista teórico.

Nuestra aportación en este nivel teórico ha sido sobre todo de clarificación en la interrelación entre los dos enfoques. Para ello se ha enunciado un conjunto de teoremas que ha permitido representar en las Figuras II.2. y II.5. las distintas maneras de dar validez a las condiciones necesarias de óptimo del problema. Otro resultado interesante ha sido la obtención completa de dichas condiciones para un determinado enunciado del problema, tanto a través de la programación no lineal como a través del principio del máximo.

En el resto de capítulos teóricos se ha buscado, sobre todo, avanzar en el conocimiento de aspectos colaterales que permiten dotar a la teoría del control óptimo discreto de una visión más amplia como instrumento matemático.

En el capítulo tres se han desarrollado extensiones puramente matemáticas, aunque sin ánimo de ser exhaustivos ya que el campo de ampliaciones no se cierra con las allí enumeradas. En este capítulo se han presentado teoremas referidos a temas como la existencia de solución, condiciones suficientes de máximo global, análisis de sensibilidad, etc.. La metodología seguida nos ha permitido trasladar teoremas de la programación no lineal al problema de control óptimo discreto, obteniéndose así los principales resultados de una forma clara y rigurosa.

El capítulo cuarto se ha ocupado de estudiar los problemas que aparecen al pasar del planteamiento teórico de las condiciones de óptimo a su resolución. Las limitaciones del tratamiento analítico de este tipo de problemas aconseja pasar a métodos numéricos de resolución. En este capítulo se han comentado brevemente estos métodos en programación no lineal, así como las posibilidades que emergen hoy en día para lograr métodos específicos en el problema de control óptimo discreto. Por último, se han destacado los problemas lineales y lineales-cuadráticos, debido tanto a las características deseables que cumplen las funciones como al hecho de que en los modelos financieros tratados en la parte aplicada se utilizan enunciados de este tipo.

En la parte de aplicaciones se han planteado distintos objetivos. La resolución de modelos en términos analíticos ha pretendido observar el manejo de las condiciones de óptimo que surgen al aplicar el principio del máximo. El caso lineal ha sido el tratado en el capítulo cinco donde se ha estudiado un problema de elección óptima de instrumentos de ahorro con fines de complemento a la pensión de jubilación. Los resultados, ya conocidos a través de otros trabajos, se han ratificado mediante la teoría del control óptimo discreto.

En la primera parte del capítulo seis se ha desarrollado un modelo nuevo con el objetivo de determinar el tipo de cotización óptimo en el ámbito de un problema de formación de un fondo de capital con condición terminal. Este problema, de tipo lineal-cuadrático, queda resuelto también de forma analítica utilizando el principio del máximo.

Un objetivo más ambicioso, a nivel aplicado, ha sido el de estudiar un problema de permanente actualidad como es el de la viabilidad futura del sistema de pensiones público (sólo de jubilación) en términos de mantener el sistema de reparto pero incorporando un fondo de capital en un horizonte temporal que incluya el impacto demográfico.

La aplicación numérica realizada sirve también de ilustración en el uso de paquetes informáticos basados en métodos numéricos de programación no lineal y, en concreto, en el uso de LINGO. El buen comportamiento en cuanto a tiempo de cálculo, lenguaje de modelización, exportación e importación de datos a hoja de cálculo, etc. han sido algunas de las características a resaltar.

En cuanto a los resultados obtenidos en esta aplicación hay que destacar la cuantificación que se realiza del efecto del cambio demográfico en términos de ajuste necesario para el tipo de cotización y la tasa de reemplazo de pensiones. El ajuste que se extrae del simple cambio demográfico se compara con el obtenido tras aplicar diferentes modelos de desviaciones cuadráticas mínimas en el contexto de la teoría del control óptimo.

La flexibilidad del modelo original se pone de manifiesto al permitir la incorporación de diversas modificaciones que pueden resultar interesantes para una distribución del ajuste que tenga en cuenta distintos efectos de tipo político.

Las conclusiones más relevantes muestran como la incorporación del fondo de capital es un elemento nivelador intergeneracional que permite reducir la magnitud del ajuste para las generaciones más desfavorecidas por el cambio demográfico. La contrapartida es un ajuste mayor en los primeros periodos. La cuantía máxima del fondo de capital es el parámetro que refleja este intercambio, a mayor fondo de capital más igualación del ajuste entre generaciones.

Los resultados numéricos concretos del modelo indican que, bajo las hipótesis consideradas como más razonables, el tipo de cotización debería aumentar un 40%

y la pensión media como parte del salario medio disminuir un 25% hasta el año 2046 si no se incorporan fondos de capital en el sistema. Esta solución exige profundizar mucho en las reformas técnicas que habitualmente se proponen para adaptar el sistema de reparto a las nuevas circunstancias de manera que se consiga un aumento interanual aproximado del 0,65% en el tipo de cotización y una disminución del 0,6% en la tasa de reemplazo de pensiones, aunque con la ventaja de que el periodo transitorio es muy dilatado. Si el ajuste sólo recayera sobre cotizantes o pensionistas las cifras anteriores se multiplicarían aproximadamente por dos para el colectivo correspondiente.

Los modelos que hemos desarrollado incorporan la creación de un fondo de capital cuyos rendimientos forman parte de los ingresos del sistema. Incorporar este fondo de una manera no traumática (limitando su cuantía) permite disminuir el ajuste anterior de forma importante. Por ejemplo, si la cuantía del fondo no excede de tres veces la masa de pensiones, el ajuste anterior se reduce aproximadamente en un tercio: el tipo de cotización debe aumentar un 25% y la pensión media disminuir un 18%. Claro está que cuanto más se reduzca el ajuste, más corto es el periodo transitorio y más se deben esforzar las generaciones iniciales. En definitiva, lo que se produce es una redistribución intergeneracional, tendiendo a igualarse las cotizaciones y pensiones a lo largo del periodo de estudio. El caso extremo es el de un fondo de capital sin restricciones que permite igualar totalmente las variables, aunque sus efectos económicos y políticos pueden ser excesivos.

La creación de un fondo de capital es una medida complementaria al resto de instrumentos de reforma que son necesarios para adaptar el sistema de reparto. Su existencia permite que el efecto del resto de reformas pueda ser menos intenso a largo plazo aunque es necesario introducirlas más pronto.

En el último epígrafe del capítulo seis se ha realizado un primer intento de cuantificación del efecto de algunas de las reformas que se plantean para adaptar el sistema de reparto. Los cálculos se basan en hipótesis simples y en datos publicados oficialmente. La disponibilidad de datos más pormenorizados permitiría un mayor

grado de exactitud, por lo que los valores obtenidos sólo son válidos como una primera aproximación. En cualquier caso, se concluye que es posible lograr la viabilidad del sistema de reparto pero combinando varios tipos de reformas para que la transición no resulte excesivamente costosa para ningún colectivo.

En cuanto a las limitaciones del modelo se puede hablar de las que afectan al planteamiento teórico y aquéllas que inciden sobre la aplicación numérica.

El modelo teórico pretende conseguir el equilibrio financiero del sistema de pensiones a largo plazo y para ello existen variables como el tipo de interés, el nivel de salarios y el número de cotizantes y pensionistas, que se han tomado como parámetros o como variables exógenas. Esto puede suponer una limitación respecto a otros modelos porque son variables que pueden verse afectadas por las que sí que forman parte del modelo como el nivel del fondo de capital, el tipo de cotización o la tasa de reemplazo.

Esta limitación existe en menor medida en los modelos de generaciones superpuestas ya que el salario y el tipo de interés son variables endógenas. Sin embargo, esto se realiza a cambio de introducir hipótesis muy restrictivas para que el modelo sea manejable: los individuos viven dos periodos, uno de cotizante y otro de pensionista; crecimiento de la población constante; etc.. En consecuencia, estos modelos son más apropiados para observar relaciones teóricas entre las variables económicas pero no para realizar aplicaciones numéricas de tipo financiero como hemos pretendido en nuestro modelo.

Lo ideal sería integrar los dos tipos de modelos, por ejemplo tomando el tipo de interés y el salario como funciones del resto de variables del modelo, aunque la especificación de este tipo de funciones requeriría de un profundo análisis previo. En todo caso, esta sería una línea de estudio posterior que queda abierta.

Las limitaciones a las que está sometida la aplicación numérica son significativas debido fundamentalmente a la amplitud del horizonte temporal

escogido. Es evidente que tomar valores de parámetros o de variables exógenas a 60 años vista puede resultar excesivo, lo cual afecta al grado de validez de los resultados. Sin embargo, el uso de la técnica de la post-optimización para observar cómo cambia la solución para otros valores de los parámetros u otras trayectorias de las variables exógenas es siempre posible. En el trabajo sólo hemos realizado algún cálculo de este tipo en lo referente a la tasa de paro y el tipo de interés, pero cabría la opción de ampliar mucho más este análisis.

Por otra parte, existen ciertas variables que no aparecen directamente en el modelo cuyo comportamiento debe ser el esperado para dar validez a los resultados. Una tasa de natalidad y de actividad creciente y una evolución del PIB y de la productividad compatible con la proyección de ocupados son ejemplos de hipótesis cuya variación afectaría a los resultados y, por tanto, se deberían tomar medidas complementarias para que la evolución de estas variables fuera la pronosticada o mejor.

El modelo sólo proporciona la evolución en el tiempo que deben seguir las variables de control y de estado pero no dice nada acerca de cómo lograr ese objetivo. Queda abierto, por tanto, todo un abanico de posibilidades a la hora de materializar esa solución:

- Elección de los instrumentos para conseguir el tipo de cotización y tasa de reemplazo óptimos (en el trabajo se dan algunas orientaciones).

- Diferenciar o no las reformas por colectivos: según el régimen de la Seguridad Social, el nivel de renta, etc..

- La gestión del fondo de capital: privada o pública, tipo de activos en los que se invierte, etc..

- Seguimiento del modelo: sería necesario, como se recomienda en las conclusiones del Pacto de Toledo, establecer un periodo de tiempo regular para, en

su caso, adaptar las variables a los nuevos acontecimientos.

Los responsables políticos son los que, en última instancia, deben valorar, como ya han hecho otros países, la gravedad del problema financiero a largo plazo del sistema público de pensiones y extraer las consecuencias correspondientes en aras a su viabilidad futura. En este sentido, los modelos aquí presentados, al margen de los valores numéricos obtenidos, creemos que tienen la suficiente flexibilidad, generalidad y rigor para poder utilizarse como una buena herramienta de trabajo.

REFERENCIAS BIBLIOGRÁFICAS

Achurra, J.L. y Quílez, M.T. (1990): "Las pensiones en 1989". *Revista de Seguridad Social*, pags. 47-68.

Ando, y Modigliani (1963): "The Lyfe-cycle Hiphotesis of saving: Agregate implications and tests". *American Economic Review*, 53, pags. 55-84.

Aoki, M. (1989): *Optimization of Stochastic Systems. Topics in Discrete-Time Dynamics*. Academic Press. London.

Artus, P. (1994): "Financiamiento de las jubilaciones, ahorro y crecimiento". *Revista internacional de la Seguridad Social*, 47, pags. 3-17.

Atkinson, A.B. (1971): "Capital Taxes, the Redistribution of Wealth and Individual Savings". *The Review of Economic Studies*, 38, pags. 209-227.

Arkin, V.I. and Evstigneev, I.V. (1987): *Stochastic Models of Control and Economic Dynamics*. Academic Press. London.

Auerbach, A.J. and Kotlikoff, L.J. (1985): "Simulating Alternative Social Security Responses to the Demographic Transition". *National Tax Journal*, 38, pags. 153-168.

Barea, J. y Fernández, M. (1994): "Evolución demográfica y gasto en protección social en España". *Revista del Instituto de Estudios Económicos*, 1-2, pags. 453-471.

Barea, J. y otros (1995): *El sistema de pensiones en España: análisis y propuestas para su viabilidad*. Círculo de empresarios. Madrid.

Bazaraa, M.S.; Sherali, H.D. and Shetty, C.M. (1993): *Nonlinear Programming. Theory and Algorithms*. John Wiley and sons. London.

Becker, G., Murphy, K. y Tamura, R. (1990): "Human Capital, Fertility and Economic Growth". *Journal of Political Economy*, 98, pags. S12-S37.

Bell, D. (1976): *El advenimiento de la sociedad post-industrial: Un intento de prognosis social.* Alianza editorial. Madrid.

Bellman, R. (1957): *Dynamic Programming.* Princenton University Press. Princenton.

Biayna, A. (1985): "Modelo de control óptimo del saldo de tesorería". *Cuadernos de economía*, 13, pags. 251-265.

Blanchet, D. and Kessler, D. (1991): "Optimal Pension Funding with Demographic Inestability and Endogenous Returns on Investments". *Journal of Population Economics*, 4, pags. 137-154.

Blenman, L.P. y otros (1995): "An alternative approach to Stochastic Calculus for Economic and Financial Models". *Journal of Economic Dynamics and Control*, 19(3), pags. 553-568.

Boadway, R., Marchand, M. and Pestieau, P. (1991): "Pay-as-you-go Social Security in a Changing Enviroment". *Journal of Population Economics*, 4, pags. 257-280.

Boltyanskii, V.G. (1978): *Optimal Control of Discrete Systems.* Halsted Press. Jerusalem.

Borrell, M. (1985): *Teoría del control óptimo.* Hispano Europea S.A.. Barcelona.

Burbidge, J.B. (1983): "Social Security and Savings Plans in Overlapping-Generations models". *Journal of Public Economics*, 21, pags. 79-92.

Burman, L.; Cordes, J.; and Ozanne, L. (1990): "IRAs and National Savings". *National Tax Journal*, 43, pags. 259-283.

Canon, M.D., Cullum, C.D. and Polak, E. (1970): *Theory of Optimal Control and Mathematical Programming*. McGraw Hill. New York.

Chow, G.C. (1975): *Analysis and Control of Dynamic Economic Systems*. John iley. New York.

Chow, G.C. (1992): "Dynamic Optimization without Dynamic Programming". *Economic Modelling*, Enero 1992, pags. 3-9.

Collins, R. (1980): "Estimating the Benefits of IRA's: A Simulation Approach". *Journal of Consumers Affairs*, 14-1.

Comisión de las Comunidades Europeas (1994): *La protección social en Europa en 1993*. Oficina de publicaciones de las comunidades europeas. Luxemburgo.

Craig, B. and Batina, R.G. (1991): "The Effects of Social Security in a Life Cycle Family Labor Supply Simulation Model". *Journal of Public Economics*, 46, pags. 199-226.

Daly, M.J. (1981): "The role or registered retirement savings plans in a life-cycle model". *Canadian Journal of Economics*, 15-3, pags. 409-421.

Dechert, W.D. (1982): "Lagrange Multipliers in Infinite Horizon Discrete Time Optimal Control Models". *Journal of Mathematical Economics*, 9, pags. 285-302.

Delaney, W. and Vaccari, E. (1989): *Dynamic Models and Discrete Event Simulation*. Marcel Dekker. New York.

Diamond, P.A. (1965): "National Debt in a Neoclassical Growth Model". *American Economic Review*, 55, pags. 1126-1150.

Diamond, P.A. and Hausman, J.A. (1984): "Individual Retirement and Savings Behavior". *Journal of Public Economics*, 23, pags. 81-114.

Disney, R. (1994): "Optimal Financing of Social Security: a Brief Review of the Issues". *Workshop sobre la financiación de las pensiones. IVIE.*

Dolezal, J. (1988): "Necessary Conditions for Pareto Optimality in Nondifferentiable Discrete Control Problems". *Control and Cybernetics*, 17-2/3, pags. 213-223.

Dorfman, R. (1969): "An Economic Interpretation of Optimal Control Theory". *American Economic Review*, 59, pags. 817-831.

Dunn, J.C. y Bertsekas, D. (1989): "Efficient Implementations of Newton's Method for Unconstrained Optimal Control Problems". *Journal of Optimization Theory and Applications*, 63, pags. 23-38.

Durán Heras, A. (1986): "Características de la población y equilibrio financiero del sistema de pensiones". *Investigaciones económicas*, 10-1, pags. 97-126.

Ehrlich, I. y Lui, F.T. (1994): "El problema de la población y el crecimiento: una revisión de la literatura desde Malthus hasta los actuales modelos de población endógena y de crecimiento endógeno". *Cuadernos económicos*, 58, pags. 189-223.

Enríquez de Salamanca, R. (1986): "Combinación óptima de los métodos financieros de un sistema de pensiones". *Investigaciones económicas*, 10-1, pags. 127-140.

Enríquez de Salamanca, R. (1989): "Política de financiación óptima en la Seguridad Social". *Investigaciones económicas*, 13-3, pags. 485-516.

Escobedo, M.I. (1991): "Un análisis empírico de los efectos finales producidos sobre el empleo industrial por el sistema de financiación de la Seguridad Social española 1975-1983". *Investigaciones económicas*, 15-1, pags. 169-192.

Fabel, O. (1994): *The Economics of Pensions and Variable Retirement Schemes*. John Wiley and Sons. Chichester.

Feldstein, M. (1974): "Social Security, Induced Retirement and Aggregate Capital Accumulation". *Journal of Political Economy*, 82, pags. 905-926.

Fiacco, A.V. (1983): *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*. Academic Press. New York.

Fisher, M.E. and Jennings, L.S. (1992): "Discrete-Time Optimal Control Problems with General Constraints". *ACM Transactions on Mathematical Software*, 18-4, 401-413.

Fletcher, R. (1991): *Practical Methods of Optimization*. John Wiley ans Sons. Chichester.

Galvez, Fernández y Verges (1992): *Manual del seguro de vida*. Expansión. Madrid.

Gift, S.J.G. (1993): "Second-Order Optimality Principle for Singular Optimal Control Problems". *Journal of Optimization Theory and Applications*, 76-3, pags. 477-484.

Goldberg, S. (1964): *Introducción a las ecuaciones en diferencias finitas*. Marcombo. Barcelona.

Gómez Sala, J.S. (1994): "El largo camino hacia la racionalización de las pensiones públicas". *Cuadernos de Ciencias Económicas y Empresariales*, 26, pags. 47-69.

Gravelle, J.G. (1991): "Do individual Retirement Accounts Increase Savings?". *Journal of Economic Perspectives*, 5-2, pags. 133-148.

Hercé San Miguel, J.A. (1986): "Presupuesto de Seguridad Social y oferta de factores en una economía de generaciones sucesivas". *Investigaciones económicas*, 10-1, pags. 37-64.

Holtzman, J.M. (1966): "Convexity and the Maximum Principle for Discrete Systems". *IEEE Transactions on Automatic Control*, 11-1, pags. 30-35.

Hubbard, R.G. (1984): "Do IRAs and Keoghs Increase Saving?". *National Tax Journal*, 37, pags. 43-54.

Hwang, C.L. y Fan, L.T. (1966): "A Discrete Version of Pontryagin's Maximum Principle".

Instituto de demografía (1994): *Proyección de la población española*. CESIC. Madrid.

Izquierdo, G. y Cuevas, A. (1994): "La financiación de la protección social y la competitividad: implicaciones para España". *Revista del Instituto de Estudios Económicos*, 1-2, pags. 473-504.

Jiménez Fernández, A. (1994): "La Seguridad Social en 1994. Un presupuesto en tiempos de crisis". *Presupuesto y gasto público*, 12, pags. 83-106.

Keerthi, S.S. and Gilbert, E.G. (1985): "An Existence Theorem for Discrete-time Infinite-Horizon Optimal Control Problems". *IEEE Transactions on Automatic Control*, 30-9, pags. 907-909.

Ko, C. (1988): *Effects of IRA's on resource allocation*. Tesis Doctoral. University of Maryland.

Levitin, E.S. (1994): *Perturbation Theory in Mathematical Programming and its Applications*. John Wiley and Sons. Chichester.

Lin, J.Y. and Yang, Z.H. (1989): "A Discrete Optimal Control Problem for Descriptor Systems". *IEEE Transactions on Automatic Control*, 34-2, pags. 177-181.

LINDO Systems Inc. (1992): *LINGO, Optimization Modelling Language*. LINDO Systems Inc.. Chicago.

Long, J. (1988): "Taxation and IRA Participation: Re-examination and conformation". *National Tax Journal*, 41, pags. 585-590.

Long, J. (1990): "Marginal Tax Rates and IRA Contributions". *National Tax Journal*, 43, pags. 143-153.

López García, M.A. (1986): "Pensiones de la Seguridad Social y bienestar: un análisis de los periodos transitorios". *Investigaciones económicas*, 10-1, pags. 65-95.

López García, M.A. (1990): "Efectos de las pensiones de la Seguridad Social sobre la oferta de trabajo en España: un comentario". *Investigaciones económicas*, 14-2, pags. 305-310.

López García, M.A. (1991): "Sobre la reforma de la Seguridad Social: ¿Capitalización o fondos de capital?". *Investigaciones económicas*, 15-3, pags. 505-530.

López García, M.A. (1992): "El tránsito del reparto a los fondos de capital de la Seguridad Social: Un modelo de simulación". *Revista Española de Economía*, 9-2, pags. 197-225.

López García, M.A. (1994): "Demographic Change, Social Security and Capital Funds". *Workshop sobre la financiación de las pensiones. IVIE.*

Malanowski, K. (1991): "Stability and Sensitivity Analysis of Discrete Optimal Control Problems". *Problems of Control and Information Theory*, 20-3, pags. 187-200.

Malliaris, A.G. and Brock, W.A. (1982): *Stochastic Methods in Economics and Finance*. North-Holland. Amsterdam.

Mangasarian, O.L. (1969): *Nonlinear Programming*. McGraw Hill. New York.

Martín Marcos, A. y Moreno Martín, L. (1990): "Efectos de las pensiones de la Seguridad Social sobre la oferta de trabajo en España". *Investigaciones económicas*, 14-2, pags. 291-303.

Michel, P. (1990): "Some clarifications on the Transversality Condition". *Econometrica*, 58-3, 705-723.

Ministerio de Economía y Hacienda (1995): *Síntesis de indicadores económicos.* Dirección General de Previsión y Coyuntura. Julio.

Ministerio de Trabajo y Seguridad Social (1994): *Proyecto de presupuestos de la Seguridad Social para 1994. Informe económico-financiero y anexo.* Volumen V, tomos 1 y 2. Madrid.

Ministerio de Trabajo y Seguridad Social (1995): *Presupuestos de la Seguridad Social. Cifras y datos. Ejercicio 1995.* Dirección general de planificación y ordenación económica de la Seguridad Social. Madrid.

Montalvo, J.G. and Quesada, J. (1994): "Some notes on Growth and Population Aging". *Workshop sobre la financiación de las pensiones. IVIE.*

Nahorski, Z., Ravn, H.F. and Valqui, R.V. (1984): "The Discrete-time Maximum Principle: a Survey and Some New Results". *International Journal of Control*, 40-3, pags. 533-554.

Nahorski, Z. and Ravn, H.F. (1988a): "Introductory Remarks. Some Connections Between Mathematical Programming, Continuous and Discrete Time Optimal Control Theory". *Control and Cybernetics*, 17-2/3, pags. 107-113.

Nahorski, Z. and Ravn, H.F. (1988b): "The Upper Boundary Approach to Constrained Discrete Time Optimal Control". *Control and Cybernetics*, 17-2/3, pags. 145-172.

Neck, R. (1984): "Stochastic Control Theory and Operational Research". *European Journal of Operational Research*, 17, pags. 283-301.

O'Neil, C.J. and Thompson, G.R. (1987): "Participation in Individual Retirement Accounts: An Empirical Investigation". *National Tax Journal*, 40, pags. 617-624.

O'Neil, C.J. and Thompson, G.R. (1988): "Taxation and IRA Participation: A Response to Long". *National Tax Journal*, 41, pags. 591-594.

Pantoja, J.F.A. de O. and Mayne, D.Q. (1991): "Sequential Quadratic Programming Algorithm for Discrete Optimal Control Problems with Control Inequality constraints". *International Journal of Control*, 53-4, pags. 823-836.

Pesch, H.J. and Bulirsch, R. (1994): "The Maximum Principle, Bellman's Equation and Carathéodory's Work". *Journal of Optimization Theory and Applications*, 80-2, pags. 199-225.

Peters, W. (1991): "Public Pensions in transition: An Optimal Policy Path". *Journal of Population Economics*, 4, pags. 155-175.

Piñera, J. (1995): *Sin miedo al futuro. ¿Es posible la reforma de las pensiones en España?*. Noesis. Madrid.

Polo, C. y Sancho, F. (1990): "Efectos económicos de una reducción de las cuotas Empresariales a la Seguridad Social". *Investigaciones económicas*, 14-3, pags. 407-424.

Pontryagin, L.S., Boltyanskii, V.G., Gamkrelidze, R.V. and Mishchenko, E.F. (1962): *The Mathematical Theory of Optimal Processes*. Interscience Publishers. New York.

Ravn, H.F. (1990): "Comments on 'A Discrete Optimal Control Problem for Descriptor Systems'". *IEEE Transactions on Automatic Control*, 35-8, pags. 985-987.

Ravn, H.F. (1991): "Correction to: 'Comments on A Discrete Optimal Control Problem for Descriptor Systems'". *IEEE Transactions on Automatic Control*, 36-6, pag. 767.

Roberts, L. (1993): "Las jubilaciones complementarias: hacia una definición de los términos". *Revista internacional de Seguridad Social*, 46, pags. 57-74.

Rozonoer, L.T. (1959): "The Maximum Principle of L.S. Pontryagin in Optimal System Theory". *Automation and Remote Control*, 20, pags. 1519-1532.

Samuelson, P.A. (1958): "An Exact Consuption-loan Model of Interest with or without the Social Contrivance of Money". *Journal of Political Economy*, 66, pags. 467-482.

Samuelson, P.A. (1975): "Optimum Social Security in a Life-cycle Growth Model". *International Economic Review*, 16, pags. 539-544.

Schmähl, W. (1990): "Demographic Change and Social Security: Some Elements of a Complex Relationship". *Journal of Population Economics*, 3, pags. 159-177.

Seierstad, A. y Sydsaeter, K. (1987): *Optimal Control Theory with Economic Applications*. North Holland. Amsterdam.

Sethi, S.P. y Thompson. G.L. (1981): *Optimal Control Theory: Applications to Management Science*. Martinus Nijdorf Publishing. Boston.

Sniedovich, M. (1992): *Dynamic Programming*. Marcel Dekker. New York.

Stock, J.H. and Wise, D.A. (1990): "Pensions, the Option Value of Work, and Retirement". *Econometrica*, 58-5, pags. 1151-1180.

Takayama, A. (1985): *Mathematical Economics*. Cambridge University Press. New York.

Tapiero, C.S. (1977): *An Optimum and Stochastic Control Approach*. Vol II. Gordon and Breach Science Publishers. London.

Tapiero, C.S. (1988): *Applied Stochastic Models and Control in Management*. North Holland. Amsterdam.

Tapiero, C.S. (1994): "Applicable Stochastic Control: from Theory to Practice". *European Journal of Operations Research*, 73, pags. 209-225.

Tu, P.N.V. (1991): *Introductory Optimization Dynamics*. Segunda edición. Springer-Verlag. Berlín.

Valqui, R.V. (1987): "On the Sufficiency of the Linear Maximum Principle for Discrete-time Control Problems". *Journal of Optimization Theory and Applications*, 54-3, pags. 583-589.

Venti, S.F. and Wise, D.A. (1990): "Have IRAs increased U.S. saving?: Evidence from Consumer Expenditure Surveys". *The Quaterly Journal of Economics*, 105-3, pags. 661-698.

Verbon, H. (1993): "Public Pensions: The Role of Public Choice and Expectations". *Journal of Population Economics*, 6, pags. 123-135.

Vicente, A. (1995): "El pacto de Toledo". *Actuarios*, 12, pags. 32-35.

Vidal Meliá, C. (1994a): *Alternativas individuales de ahorro-pensión: planes de pensiones*. Tesis Doctoral. Universitat de València.

Vidal Meliá, C. (1994b): "La necesidad de reformar los planes de pensiones individuales". *Tribuna fiscal*, 50, pags. 48-57.

Villagarcía, T. (1995): "Análisis econométrico del tránsito a la jubilación para trabajadores de edad avanzada". *Investigaciones económicas*, 19(1), pags. 65-81.

Vinter, R.B. (1988): "Optimality and Sensitivity of Discrete Time Processes". *Control and Cybernetics*, 17-2/3, pags. 191-211.

Wilcox, D.W. (1989): "Social Security Benefits, Consuption Expenditure, and the Life Cycle Hypothesis". *Journal of Political Economy*, 97-2, pags. 288-304.