Descomposición y Coordinación Paralela para Solucionar un Modelo de Planeación de la Operación de Sistemas de Generación y Transmisión Eléctrica con Altas Penetraciones de Fuentes Intermitentes, Almacenamiento Energético y Tecnologías de Redes Inteligentes

María del Pilar Buitrago Villada

Universidad Nacional de Colombia
Facultad de Ingeniería y Arquitectura,
Departamento de Ingeniería Eléctrica Electrónica y Computación
Manizales, Colombia
Año 2021

Parallel Decomposition and Coordination to Solve an Operation Planning Model for the Electricity Generation and Transmission Systems with High Penetrations of Intermittent Sources, Energy Storage, and Smart Grid Technologies

María del Pilar Buitrago Villada

Universidad Nacional de Colombia
Facultad de Ingeniería y Arquitectura,
Departamento de Ingeniería Eléctrica Electrónica y Computación
Manizales, Colombia
Año 2021

Descomposición y Coordinación Paralela para Solucionar un Modelo de Planeación de la Operación de Sistemas de Generación y Transmisión Eléctrica con Altas Penetraciones de Fuentes Intermitentes, Almacenamiento Energético y Tecnologías de Redes Inteligentes

María del Pilar Buitrago Villada

Tesis presentada como requisito parcial para optar al título de:
Doctor en Ingeniería Automática

Director:
Ph.D. Carlos Edmundo Murillo Sánchez

Línea de Investigación:
Análisis de sistemas de potencia eléctrica
Grupo de Investigación:
Potencia Energía y Mercados - GIPEM

Universidad Nacional de Colombia
Facultad de Ingeniería y Arquitectura,
Departamento de Ingeniería Eléctrica Electrónica y Computación
Manizales, Colombia
Año 2021

A Dios.
A mis padres, mis hermanos y a Marianita.

Agradecimientos

El presente trabajo de investigación doctoral recibió financiamiento por parte del Ministerio de Ciencias (Colciencias) bajo el programa de Becas de Doctorados Nacionales convocatoria 727 de 2015.

Deseo expresar mis más profundos agradecimientos a mi director, el profesor Carlos Edmundo Murillo Sánchez, por su paciencia y dedicación en la dirección de esta tesis. Al Departamento de Ingeniería Eléctrica Electrónica y Computación y a la Facultad de Ingeniería y Arquitectura por su apoyo económico para la realización de la pasantía en el Centro de Energía de la Universidad de Chile. Al Dr. Marcelo Matus Acuña por su asesoría durante mi estancia doctoral allí.

Gracias a mis colegas y amigos del Grupo de Investigación en Potencia Energía y Mercados-GIPEM de la Universidad Nacional de Colombia y a los profesores Francisco Abel Roldán Hoyos, Belizza Janet Ruíz Mendoza y Jorge Fernando Gutiérrez Gómez (Q.E.P.D.), por su apoyo constante y por las charlas que ayudaron a enriquecer este trabajo.

Resumen

Este trabajo presenta una metodología de solución de un modelo estocástico usado en la planeación de la operación de sistemas eléctricos de potencia con alta penetración de fuentes de energía renovable, respuesta de la demanda y sistemas de almacenamiento energético, que además incluye de manera explícita el modelo AC de la red de transmisión. El impacto que tiene el modelo AC sobre la apropiada asignación y valoración de los recursos del sistema de potencia, en el contexto de mercados multi-dimensionales, se evidencia a través de un estudio comparativo simulando un caso de prueba de tamaño real. Resolver de forma directa un problema de las dimensiones que puede alcanzar la formulación propuesta requiere mucho tiempo, grandes esfuerzos de cálculo y recursos informáticos. Por tal motivo, se exploraron dos estrategias para explotar la estructura matemática del problema y abordar su solución usando técnicas de descomposición: La descomposición por Relajación Lagrangiana con lagrangiano Aumentado (RLA) y la Descomposición Generalizada de Benders (DGB). Entre estas, se implementó efectivamente DGB en su versión multicorte con una modificación en la formulación de los subproblemas mediante variables penalizadas. El algoritmo fue acelerado con una técnica de estabilización inspirada en los métodos de haz con región de confianza y el cómputo en paralelo de los subproblemas. Otras medidas de aceleración adicionales fueron diseñadas a partir de observaciones en la evolución de algunos parámetros durante los experimentos. El desempeño de la técnica DGB se validó a través de pruebas experimentales en dos casos de diferente tamaño: el sistema IEEE de 30 barras y el sistema de potencia colombiano de 96 barras. Los resultados sugieren que el esquema de solución propuesto es apropiado para tratar de forma eficiente un problema de optimización de tamaño real como el sistema de potencia colombiano. Una asignación de cantidades de potencia y reservas bastante aproximada fue reflejada en una desviación cercana al 0,005 % en el costo óptimo comparado con la solución de referencia; además del buen desempeño computacional dado por la reducción del 88 % del tiempo de cálculo con respecto a la solución de referencia (sin descomposición), generando un avance en el estado de arte de este campo de estudio.

Palabras clave: Descomposición y coordinación, despacho económico seguro, fuentes de energía renovable, generación y transmisión de potencia, optimización numérica de gran escala, planeación del sistema de potencia, procesamiento en paralelo.

Abstract

This work presents a solution methodology for a stochastic model used in the operational planning of electric power systems with high penetration of renewable sources, demand response, and energy storage systems, which also explicitly includes the AC model of the network. The AC model impacts the correct allocation and assessment of power system resources in the context of multi-dimensional markets, demonstrated through a comparative study simulating a real-size test case. Solving in a direct way a high dimensional problem that could be reached through the proposed formulation requires a lot of time, great calculation effort, and computer resources. For this reason, two strategies were explored to exploit the mathematical structure of the problem and approach its solution by decomposition techniques: Augmented Lagrangian Relaxation decomposition (ALR) and Generalized Benders Decomposition (GBD). Among these, multi-cut GBD was effectively implemented with a modification in the subproblems formulation through penalized variables. The algorithm was accelerated with a trust-region stabilization technique and the parallel computing of subproblems. Other additional acceleration measures were designed from observations of the evolution of some parameters during the experiments. The performance of the GBD technique was validated through experimental tests in two different-sized test cases: the IEEE 30-bus system and the Colombian 96-bus power system. The results suggest the effectiveness of the proposed solution scheme to efficiently solve a real-size optimization problem like the Colombian power system. A quite approximate power and reserve quantities allocation was reflected in an optimal cost deviation close to 0.005 %, compared with the reference solution; in addition to the good computational performance given by the 88 % reduction in the calculation time relative to the reference solution (without decomposition), generating an advance in the state of the art of this field of study.

Keywords: decomposition and coordination, large-scale numerical optimization, parallel processing, power generation and transmission, power system planning, renewable energy sources, security economic dispatch.

Contenido

Lista de Figuras

Lista de Tablas

Lista de símbolos para MPSSOPF

La nomenclatura usada en todo el documento para referirse a los componentes del modelo de flujo óptimo de potencia multi-periodo con restricciones de seguridad, en adelante MPSSOPF (*Multi-Period Stochastic Security constrained Optimal Power Flow*), se indica a continuación. Otros símbolos específicos se declararán donde sean utilizados. Para abreviar, se hará referencia al estado/flujo de potencia post-contingente k del escenario j en el tiempo t, simplemente como el estado/flujo de potencia post-contingente tjk.

Índices

Símbolo	Término
i	Índice de la potencia inyectada (generadores, almacenamiento y cargas despachables).
j	Índice correspondiente a los escenarios.
k	Índice de los casos post-contingencia ($k = 0$ para el caso base).
t	Índice correspondiente al periodo de tiempo.

Conjuntos

Símbolo	Término
I^{tjk}	Índice de todas las unidades disponibles para despachar en el estado post-contingente tjk.
J^t	Conjunto de índices de todos los escenarios considerados en el tiempo t.
K^{tj}	Conjunto de índices para todas las contingencias consideradas en el escenario j en el tiempo t.
T	Conjunto de índices del periodo de tiempo en el horizonte de planeación.

Constantes

Símbolo	Término
P_{min}^{tijk}	Límite inferior en la potencia activa para la unidad i en el estado post-contingente tjk.
P_{max}^{tijk}	Límite superior en la potencia activa para la unidad i en el estado post-contingente tjk.
Q_{min}^{tijk}	Límite inferior en la potencia reactiva para la unidad i en el estado post-contingente tjk.
Q_{max}^{tijk}	Límite superior en la potencia reactiva para la unidad i en el estado post-contingente tjk.
s_0^i	Energía almacenada inicial (esperada) en la unidad de almacenamiento i.
α	Para casos de contingencia, la fracción de tiempo que es gastada en el caso base antes que la contingencia ocurra ($\alpha = 0$ significa que el periodo entero es gastado en la contingencia).
γ^t	Probabilidad de pasar al periodo t sin desviarse de la ruta central a una contingencia en periodos $1 \ \ldots \ t-1$.
δ_{max+}^i	Límite superior en la reserva de potencia activa de rampa de seguimiento de carga para la unidad i.
δ_{max-}^i	Límite inferior en la reserva de potencia activa de rampa de seguimiento de carga para la unidad i.
Δ	Longitud del periodo de tiempo de planeación, típicamente 1 hora.
δ_+^i	Límite físico superior de rampa para la unidad i para la transición desde el caso base ($k = 0$) a los casos de contingencia.
δ_-^i	Límite físico inferior de rampa para la unidad i para la transición desde el caso base ($k = 0$) a los casos de contingencia.
η_{in}^i	Eficiencia de carga para la unidad de almacenamiento i.
η_{out}^i	Eficiencia de descarga para la unidad de almacenamiento i.
η_{loss}^i	Fracción de energía almacenada perdida por hora en la unidad de almacenamiento i.
τ_i^+	Tiempo mínimo de arranque para la unidad i en un número de periodos.
τ_i^-	Tiempo mínimo de parada para la unidad i en un número de periodos.
$\Phi^{tj_2j_1}$	Probabilidad de transición al escenario j_2 en el periodo t dado que se procede del escenario j_1 en el periodo $t-1$.

Símbolo	Término
ψ^{tjk}	Probabilidad de la contingencia k en el escenario j en el tiempo t (ψ^{tj0} es la probabilidad del caso base).
ψ_{α}^{tjk}	Probabilidad de la contingencia k en el escenario j en el tiempo t, ajustado por α.

Variables

Símbolo	Término
p^{tjk}	Vector de potencia activa para el flujo de potencia post-contingente tjk.
p^{tijk}	Potencia activa inyectada por la unidad i en el estado post-contingente tjk.
p_{+}^{tijk}	Desviación ascendente de la cantidad de potencia activa contratada para la unidad i en el estado post-contingente tjk.
p_{-}^{tijk}	Desviación descendente de la cantidad de potencia activa contratada para la unidad i en el estado post-contingente tjk.
p_{c}^{ti}	Cantidad contratada de potencia activa para la unidad i en el tiempo t.
p_{sc}^{tijk}	Carga de potencia de la unidad i en el estado post-contingente tjk.
p_{sd}^{tijk}	Descarga de potencia de la unidad i en el estado post-contingente tjk.
q^{tjk}	Vector de potencia reactiva para el flujo de potencia post-contingente tjk.
q^{tijk}	Potencia reactiva inyectada por la unidad i en el estado post-contingente tjk.
r_{+}^{ti}	Cantidad ascendente de reserva de contingencia provista por la unidad i en el tiempo t.
r_{-}^{ti}	Cantidad descendente de reserva de contingencia provista por la unidad i en el tiempo t.
δ_{+}^{ti}	Reserva ascendente de rampa de seguimiento de carga necesaria de la unidad i en el tiempo t, para la transición al tiempo $t+1$.
δ_{-}^{ti}	Reserva descendente de rampa de seguimiento de carga necesaria de la unidad i en el tiempo t, para la transición al tiempo $t+1$.
s_{+}^{ti}	Límite superior en la energía almacenada en la unidad de almacenamiento i al final del periodo t.
s_{-}^{ti}	Límite inferior en la energía almacenada en la unidad de almacenamiento i al final del periodo t.

Símbolo	**Término**
V^{tjk}	Vector de magnitudes de tensión para el flujo de potencia post-contingente tjk.
θ^{tjk}	Vector de ángulos de tensión para el flujo de potencia post-contingente tjk.
u^{ti}	Estado de comisionamiento (binario) para la unidad i en el periodo t, 1 si la unidad está conectada, 0 de otra forma.
v^{ti}	Estado de arranque (binario) para la unidad i en el periodo t, 1 para eventos de arranque en el periodo t, 0 de otra forma.
w^{ti}	Estado de parada (binario) para la unidad i en el periodo t, 1 para eventos de parada en el periodo t, 0 de otra forma.

1. Introducción

La planeación de la operación de los sistemas de potencia eléctrica en mercados desregulados tiene como objetivo principal garantizar, de manera anticipada, la disponibilidad y suficiencia de todos los recursos de generación necesarios para suplir la demanda al menor costo posible, cumpliendo con las restricciones de seguridad y confiabilidad impuestas por el sistema de transmisión.

Las restricciones de seguridad abarcan las restricciones físicas y operacionales de los elementos del sistema de potencia. Las restricciones que tienen que ver con los generadores consideran los tiempos mínimos de arranque y parada, los límites superior e inferior de generación, el rango de toma de carga, los rangos de rampa, la contribución a las reservas rodantes, etc. Por su parte, las restricciones operativas del sistema de transmisión comprenden los límites de tensión, los límites de capacidad de transmisión, entre otros. Otras consideraciones de seguridad están relacionadas con la capacidad del sistema para soportar la ocurrencia de posibles contingencias y requieren de la provisión de reservas operativas zonales, rodantes y no rodantes, por parte de algunas unidades de generación.

El mercado de día en adelanto, o del día siguiente, tiene a la programación de la operación del sistema como parte inicial del proceso de establecimiento de las cantidades de generación para cada hora del siguiente día operativo. En esta etapa, el operador del sistema recibe las ofertas de los comercializadores, las ofertas de los generadores y la programación de transacciones bilaterales para cerrar el mercado. Este proceso se realiza mediante estudios de comisión de unidades y de despacho económico con restricciones de seguridad [1].

El problema de comisión de unidades establece las decisiones de programación (*on/off*) de un conjunto de unidades de generación para un horizonte temporal definido. Por su parte, el problema de despacho económico define la cantidad de generación necesaria para suplir la demanda, y otras restricciones operativas del sistema, que debe ser despachada en cada periodo de tiempo para cada generador *online*, generalmente determinado por una solución del problema de comisión de unidades. En el proceso se minimiza el costo total de producción o se maximiza el beneficio social de los participantes del mercado. Los precios de la energía y las reservas en cada intervalo de tiempo y en cada localización también se pueden derivar del problema de despacho económico.

Tradicionalmente, la ocurrencia de eventos discretos, como la desconexión de recursos de generación o de los componentes de transmisión, y los errores en los pronósticos de carga constituyeron las principales fuentes de incertidumbre en el proceso de planeación del sistema de potencia. La incertidumbre ocacionada por las contingencias se modela mediante una distribución de probabilidad basada en datos históricos y los posibles redespachos que tengan lugar son soportados a través de una capacidad de generación adicional, en términos de reserva operativa, mientras que la incertidumbre en el pronóstico de carga ha sido disminuida gracias al uso de técnicas modernas de predicción. Por lo tanto, el modelo de planeación es equivalente a uno determinístico dado el grado de precisión con el que se pueden conocer la capacidad de generación disponible y la carga.

En los últimos años la mayor fuente de incertidumbre proviene de la inclusión incremental de Fuentes de Energía Renovable (FER) en la matriz energética. Dada la variabilidad e intermitencia en su producción de energía no es fácil programar su operación, en contraste con las fuentes de generación convencionales [2]. En general, la variabilidad de las FER puede aumentar los requisitos de reserva para acomodar los posibles erorres entre el valor predicho y la producción real de potencia y ocasionar rampas más frecuentes de lo habitual en algunos generadores. Adicionalmente, las FER generan impactos en los precios como lo señala [3], tales como: precios de energía más bajos en promedio debido al costo marginal casi nulo de las FER, aumento de la volatilidad en los precios, precios de energía negativos debido a los esquemas de incentivos para las FER o precios más altos para los servicios de reserva.

Los operadores del sistema de potencia, los participantes del mercado, los reguladores y los investigadores en el ámbito académico trabajan en conjunto para entender los cambios necesarios en la operación del sistema de potencia ante los desafíos impuestos por las características inciertas en la producción de potencia por parte de las FER. La discusión gira en torno a cómo el incremento de FER puede impactar los mercados de electricidad y si el diseño actual del mercado es suficiente para soportar la operación del sistema de potencia. Con respecto a esto último, en [4] se mencionan algunos ejemplos de cambios del diseño del mercado implementados, o considerados para su implementación, tales como la introducción de nuevos productos en el mercado del día en adelanto como el de rampa flexible adoptado por el operador independiente del sistema en California (CAISO), o el producto de capacidad de rampa implementado por MISO (*Midcontinent ISO*); o remunerar la capacidad térmica flexible a través de mercados intradiarios para que los participantes del mercado ajusten sus posiciones a medida que evolucionan los pronósticos de energía eólica, solar y de carga, como en el noroeste de Europa (Francia, Bélgica, Países Bajos y Alemania). A parte de la inclusión de nuevos productos de mercado también se han redefinido servicios auxiliares, como en el caso de ERCOT (*Electric Reliability Council of Texas*) [4].

Entonces, los mecanismos del mercado de electricidad, incluídos los mercados de servicios

auxiliares, pueden ser replanteados para compensar de forma eficiente a los productores de energía de tal forma que se incetive la flexibilidad operativa de la red. Sin esta flexibilidad se menoscaba la confiabilidad de una red eléctrica que obtenga un mayor porcentaje de su energía de las FER y no es posible un mayor despliegue de las mismas [5]. En aras de contribuir con la flexibilidad exigida al sistema, se ha considerado incorporar recursos como la respuesta de la demanda, la gestión del almacenamiento de energía, la generación distribuída, entre otros, que permitan mantener el balance generación-demanda.

La respuesta de la demanda aprovecha el hecho de que una cantidad importante de la demanda de electricidad sea elástica [6], lo que significa que parte de esta se puede diferir en el tiempo. Cargas como los vehículos eléctricos, los sistemas de calefacción, de ventilación y de aire acondicionado son demandas elásticas (o flexibles), que pueden representar un porcentaje considerable de la demanda total en algunos países. Una cantidad suficiente de demanda flexible se asemeja a un tipo de capacidad de reserva rodante, funcionando como almacenamiento virtual que retrasa el uso de la reserva del lado de la generación [7]. De hecho, algunos operadores del sistema están estructurando las reglas de mercados de servicios auxiliares, tal que las reservas provistas tradicionalmente por los generadores puedan participar junto con la respuesta de la demanda [8]. Esto provee un medio para disminuir los efectos económicos, técnicos y ambientales adversos de mitigar la variabilidad de las FER usando generadores convencionales de respuesta rápida al proveer un tipo de reserva flexible. Adicional a esto, la demanda flexible constituye una oportunidad para aliviar la congestión en transmisión en sistemas cuya generación base es poco flexible. Sin embargo, si no se modela adecuadamente la respuesta de la demanda, cuando esta se despliegue en condiciones de escasez de la capacidad de emergencia, la respuesta de la demanda puede aparecer como una simple reducción de carga o una especie de suministro gratuito, lo que podría conducir a la supresión de precios, contrario a las señales necesarias durante estas condiciones operativas [4].

En cuanto a la implementación efectiva de los dispositivos de almacenamiento de energía, esta trae beneficios adicionales como la provisión de reservas operativas y de servicios auxiliares basados en la capacidad (habilidad para proveer energía a demanda) [5], el brindar soporte al funcionamiento de los sistemas de transmisión evitando la necesidad de mejoras específicas y/o la adición de nuevos recursos de generación [9], o el arbitraje, que consiste en comprar electricidad cuando es barata y venderla de nuevo a la red cuando se vuelve cara. Una variedad de tecnologías de almacenamiento incluye: almacenamiento de energía por bombeo, almacenamiento de energía por aire comprimido, sistemas de baterías, capacitores, almacenamiento térmico, entre otros.

Por su parte, la suposición convencional de que la energía siempre fluye de la red de transmisión a la red de distribución está siendo cambiada gracias a una cantidad importante de recursos de generación distribuída que se ha conectado a la red de distribución. Los recursos

de generación distribuída pueden incluir generadores de respaldo como motores diesel, celdas de combustible, paneles solares fotovoltaicos instalados en techo, como también la respuesta de la demanda. Una penetración significativa de estos recursos puede hacer que se incremente el nivel de incertidumbre ya que una parte de esos recursos no es observable para el operador de la red de transmisión. Incluso el concepto de contingencia está cambiando de ser binario (encendido/apagado o conectado/desconectado) a continuo [10]. Por ejemplo, una variación de varios GW en la carga o en la generación del sistema en un período relativamente corto podría pasar a verse como un funcionamiento normal del sistema, cuando tradicionalmente ha sido considerado como un evento de emergencia o anormal. Entonces, las herramientas para análisis de contingencias requerirán modificaciones para capturar los impactos en la confiabilidad ocasionados por una penetración significativa de generación distribuída [4].

En el contexto descrito, el enfoque clásico de planificación operativa basado en un esquema determinístico convencional puede ser insuficiente. Lo anterior, ha alentado cambios en el paradigma de la programación de la operación de los sistemas de potencia, que no sólo debe mantener el balance generación-demanda, sino que además puede tener en cuenta varios problemas particulares:

1. Un esquema consistente para despachar y redespachar la generación de potencia con generación variable, que tenga un enfoque apropiado para manejar la incertidumbre de la inyección de potencia de las FER, por ejemplo, a través de múltiples escenarios de realización de energía renovable.

2. El establecimiento de un esquema operativo sujeto a consideraciones de seguridad, mediante una serie de contingencias ponderadas por probabilidad, y la planeación para estados post-contingentes, representando apropiadamente las restricciones de la red de transmisión.

3. El modelado de mecanismos de mercados multi-producto, co-optimizando simultáneamente energía y servicios auxiliares, que garanticen el costo mínimo y la operación segura de la red. Entre estos servicios auxiliares se podrían considerar, a parte de las reservas tradicionales, los aportes a reserva y rampa provenientes de dispositivos de almacenamiento despachado centralmente y de mecanismos de la respuesta de la demanda, facilitada por las nuevas tecnologías de redes inteligentes.

4. La selección de técnicas de optimización adecuadas para manejar la gran dimensionalidad del problema, dada entre otros, por el número de variables y restricciones del modelo, como consecuencia de los múltiples escenarios de realización de generación renovable y de los estados post-contingentes, que deben incorporarse para modelar la incertidumbre.

Generalmente, cada uno de estos problemas se resuelve individualmente y de forma secuencial. En el proceso se revisa que el despacho inicial calculado por una herramienta de Flujo

Óptimo de Potencia (OPF, por sus siglas en inglés) se adapte a restricciones adicionales, por ejemplo, restricciones de seguridad en el caso de las contingencias. Esta práctica no garantiza que los procesos de verificación sean introducidos de una forma que se preserve la optimización del problema general, ni que se pueda aprovechar la información suministrada por el OPF, como los precios marginales locales para valorar correctamente los productos de energía y reserva.

1.1. Antecedentes

En la literatura se encuentran propuestas para llevar a cabo la programación de la operación de sistemas de potencia, algunas en el contexto del mercado de día en adelanto, que abordan uno o varios de los problemas indicados anteriormente. La Tabla **1-1** presenta un resumen de varias propuestas disponibles en la literatura, indicando los aspectos que hacen parte de la formulación del problema.

La mayoría de estas propuestas son estocásticas multi-periodo y co-optimizan la producción de energía y la asignación de las reservas, con excepción de [11–14, 16], configurando un mercado multi-producto. De otra parte, un número importante de estos estudios tiene en cuenta restricciones que permiten programar una cantidad suficiente de reservas para la operación segura del sistema, a excepción de [13, 14]. Adicionalmente, otras formulaciones consideraron los costos de la generación de potencia reactiva [24], el costo de la carga no atendida [21, 25], el costo del almacenamiento [16, 23, 27] o el costo de la respuesta de la demanda [14].

Algunos estudios se enfocaron en tratar la respuesta de la demanda acoplada con la operación de FER's como en [11, 12, 14, 27], mientras que en [7] se tomó como medida para proveer reservas, ofreciendo grados de libertad adicionales en términos de deslastre de carga. En ese caso, la demanda respondía a las contingencias y no a los precios del mercado.

En cuanto al modelado de los dispositivos de almacenamiento de energía, encontraron que no solo facilita la integración de FER en la red, como en [16, 26], sino que además ayuda a reducir la congestión y los costos de rampa, aunque puede incrementar los gases de efecto invernadero producidos por las plantas convencionales [23]. Al igual que en el caso de la respuesta de la demanda, se apunta a la necesidad de incluir el almacenamiento de energía como otra categoría de reserva [19].

En general, se usaron como casos de prueba sistemas de potencia que varían desde 5 barras hasta 118 barras, con excepción de [19] que utilizó un sistema de 2.556 barras, y de [13] donde simularon sistemas de 30, 118, 1.354 y 13.659 barras con periodos de tiempo entre 240 y 8.760 horas. Algunas pruebas realizadas sobre sistemas de potencia pequeños resultaron en problemas de optimización de gran tamaño debido al número de escenarios y/o contin-

Tabla 1-1.: Caracterización de los estudios revisados para la planeación de la operación de sistemas eléctricos de potencia

Referencia	**Multi-Producto**	**Reservas**	**AE**	**RD**	**Seguro**	**Modelo de Red**
Martinez [11]	X	✓	X	✓	X	-
Papavasiliou [12]	X	✓	X	✓	X	-
Kourounis [13]	X	X	✓	X	X	AC
Bukhsh [14]	✓	X	X	✓	X	DC
López-Salgado [15]	✓	✓	X	X	X	DC
Gomes [16]	✓	✓	✓	X	X	-
Banshwar [17]	✓	✓	X	X	X	AC
Bouffard [18]	✓	✓	X	X	✓	-
Liu [19]	✓	✓	X	X	✓	DC
Sharifzadeh [20]	✓	✓	X	X	✓	DC
Lamadrid [21]	✓	✓	X	X	✓	AC
Amjady [22]	✓	✓	X	X	✓	AC
Virasjoki [23]	✓	✓	✓	X	X	DC
Murillo-Sánchez [24]	✓	✓	X	X	✓	AC
Zhang [25]	✓	✓	X	X	✓	DC
Parastegari [26]	✓	✓	✓	X	X	-
Karangelos [7]	✓	✓	X	✓	✓	DC
Murillo-Sánchez [27, 28]	✓	✓	✓	✓	✓	AC-DC

AE = Almacenamiento de Energía
RD = Respuesta de la Demanda

gencias consideradas. Por ejemplo, Zhang *et al.* [25] analizaron casos de prueba con más de 3 millones de restricciones y más de 1 millón de variables, usando sistemas de potencia de 6 y 20 barras, con más de 30 escenarios.

Ante la expectativa de tener que resolver un problema de tales dimensiones, algunos estudios recurrieron a simplificaciones para lograr que la solución se calculara con una cantidad de tiempo razonable. Varias alternativas abogaron por: técnicas de reducción de escenarios [12, 25, 26] o de generación de escenarios [20], horizontes de programación diarios con

pocos periodos de tiempo representando varias horas [23, 27], no modelar la red de transmisión [11, 12, 16, 18, 26, 27] o usar la forma linealizada de la red de corriente alterna, conocido como modelo DC, [7, 14, 15, 19, 20, 23, 25, 27] en lugar del modelo no lineal AC.

Entre las propuestas que atacaron el problema con metodologías de solución de dos etapas [12, 14, 16, 19, 20, 22], tres de ellas [14, 19, 22] pasaron de un modelo estocástico a uno determinístico equivalente mediante algunas transformaciones en las restricciones y el uso de intervalos de confianza, facilitando su solución. Algunos estudios señalan la necesidad de emplear métodos de descomposición como en [11, 15, 18]. De hecho en [12, 27] se resuelve el problema descomponiéndolo mediante Relajación Lagrangiana, en ambos casos para sistemas de pequeña dimensión; y en [13] se explotó la estructura multi-periodo del problema para lograr su solución por descomposición a través de una aproximación basada en el complemento de Schur. Por último, un método heurístico de solución fue empleado en [25].

De la revisión de la literatura se observa que se han hecho importantes esfuerzos para programar la operación del sistema de cara a los nuevos desafíos impuestos por la penetración de cantidades importantes de FER. Los estudios revisados han atacado uno o varios problemas en conjunto, enfatizando que tanto el almacenamiento como la demanda flexible deberían considerarse como otra categoría de reservas. También se revelan algunos efectos secundarios, tanto positivos como negativos, resultantes de la implementación de las mencionadas tecnologías en la operación del sistema. Por último, algunas propuestas apuntan a la necesidad de hacer simplificaciones en el modelo o utilizar técnicas de optimización eficientes para manejar la gran dimensionalidad que puede alcanzar un problema de optimización, inclusive a partir de sistemas de potencia de tamaño pequeño o moderado.

1.2. Motivación e interés por el tema

La programación de la operación segura y confiable de los sistemas de transmisión y generación que incorporen en su matriz energética FER, depende críticamente de la disponibilidad de herramientas que posibiliten la toma de decisiones para asignar y valorar apropiadamente los recursos del sistema eléctrico de potencia ante incertidumbre. Por tanto, se requiere el conocimiento de todas las variables involucradas en el problema.

En la literatura se han encontrado propuestas que usan de forma preferencial el modelo lineal de la red de transmisión, conocido como el modelo DC, con el fin de reducir el tamaño y la complejidad del problema. Adicionalmente, esta simplificación permite el uso de técnicas de programación lineal o cuadrática, que conducen a una rápida solución, y carece de los problemas de convergencia de su contraparte AC [29]. No obstante, la formulación basada en el modelo DC tiene como desventajas el hecho de no proveer información de la potencia

reactiva ni de la magnitud de las tensiones de barra, que son necesarias para definir en forma correcta ciertas restricciones del sistema, tales como los flujos en las líneas de transmisión, las que muestran la debilidad real del sistema al usar límites aproximados. Además, favorece la distorsión de precios especialmente cuando el sistema está estresado y es precisamente bajo estas circunstancias cuando los precios correctos permiten identificar la localización de debilidades existentes en la red.

En contraste, aunque la formulación del problema basado en el modelo AC de la red de transmisión es más completa, el problema se vuelve más desafiante cuando se incorporan las restricciones no lineales de tal modelo. Existen consideraciones ingenieriles genuinas para incluir el modelo de flujo de red AC, puesto que algunas restricciones pueden ser modeladas con mayor precisión. Por ejemplo, los límites en el flujo de líneas de transmisión son mejor expresados en términos del rango de MVA del transformador, o en términos de una corriente máxima. Sin embrago, tanto los MVA reales como la corriente dependen de sus componentes ortogonales de potencia activa y reactiva. Otro ejemplo es dado por los límites de tensión; predecirlos linealmente en términos de inyecciones de potencia activa no es suficientemente preciso. Finalmente, hay componentes en la red que los flujos DC no son capaces de modelar, como los taps de los transformadores.

Por otra parte, evaluar el desempeño de los marcos de optimización estocástico más completos bajo el modelo AC, usando sistemas de potencia de tamaño real, ha resultado prohibitivo puesto que la complejidad computacional del problema aumenta exponencialmente al aumentar la escala del sistema. Incluso, problemas de programación para sistemas de potencia de tamaño moderado se vuelven inmanejables. Para abordar estos desafíos computacionales en aplicaciones de mayor escala, se ha dirigido la atención de la comunidad científica al desarrollo de algoritmos de solución basados en la descomposición del problema original, por ejemplo, en subproblemas sencillos de menor tamaño que se resuelvan por separado, ya sea en paralelo o secuencialmente.

De la revisión de la literatura se establece que sigue siendo un reto resolver el problema de la programación estocástica de la operación sujeta a diferentes factores, tales como la respuesta de la demanda, la administración dispositivos de almacenamiento y el procuramiento de reservas localizadamente, teniendo en cuenta las restricciones de la red de transmisión bajo el modelo AC y su correspondiente aplicación a sistemas de potencia reales.

Si bien la propuesta formulada en [27], reconocida como la más completa en cuanto a multidimensionalidad y número de variables [30], acomete la mayoría de los problemas mencionados anteriormente y su formulación contempla el modelo AC de la red de transmisión, sólo se incluyen algunas discusiones del algoritmo para su solución AC. En cambio, ese artículo se limita a presentar los resultados de convergencia de la formulación bajo el modelo DC

para una versión modificada del sistema IEEE de 118 barras con 24 periodos de tiempo y 4 escenarios de viento, alcanzando una dimensión de más de 462 mil variables y más de 819 mil restricciones. Recientemente, en [28] se amplió el tamaño del problema a 500 escenarios de viento. Estos dos ejemplos ayudan a ilustrar las potencialidades de la formulación y el gran desafío computacional de manejar simultáneamente múltiples periodos de tiempo, contingencias y escenarios de realización de energía renovable. Con este nivel de detalle el problema resultante es de gran magnitud, incluso para la implementación bajo el modelo DC. Esto justifica que no haya sido puesto en práctica un solucionador para el problema completo con modelo AC, que es más complejo y de dimensiones difíciles de manejar.

Resolver el problema de optimización como está formulado en [27] constituiría una opción que desde la academia trata de abordar el problema de forma íntegral, al considerar un buen número de cuestiones que podrían afectar la operación de un sistema de potencia a futuro, algunas de ellas relacionadas con la interacción de los nuevos recursos como los sistemas de almacenamiento y respuesta de la demanda. También puede verse como una herramienta de apoyo útil para otras investigaciones, por ejemplo, en la evaluación de diseños de mercados que requieran un análisis detallado de la acción que recíprocamente se ejerce entre estas nuevas tecnologías con las ya existentes.

1.3. Objetivos del trabajo

1.3.1. Objetivo general

Implementar un algoritmo efectivo de descomposición y coordinación para la solución de un modelo estocástico para la programación de la operación de sistemas eléctricos de potencia con alta penetración de FER, almacenamiento de energía y respuesta de la demanda, que incorpore el modelo AC de la red, para alcanzar la solución de problemas de tamaño real.

1.3.2. Objetivos específicos

1. Desarrollar un programa de cómputo para resolver el problema de programación de la operación en [27] que incorpore las restricciones no lineales de los flujos óptimos de potencia AC, para generar soluciones patrón con el fin de validar los resultados del solucionador a programar.

2. Comparar varias estrategias de descomposición y coordinación, para definir la estrategia que provea mayor efectividad y velocidad de convergencia del problema de optimización.

3. Aplicar por lo menos uno de los algoritmos estudiados a la solución de problemas de tamaño real, para demostrar su valor agregado sobre las herramientas disponibles actualmente.

1.4. Estructura de la tesis

En el **Capítulo 1** se introduce el tema de investigación mediante la exposición de los antecedentes, la motivación e interés por el tema y los objetivos perseguidos, además de presentar la estructura del documento.

La formulación del planeador estocástico multi-periodo para la programación de la operación de sistemas de potencia es introducida en el **Capítulo 2**. Conceptualmente, el planeador ofrece la capacidad de modelar las restricciones de la red de transmisión de ambos modelos de red (AC y DC), aunque en la práctica solo se ha implementado el modelo DC de la red. También se exponen los detalles de modelado de cadenas hidrológicas, que no hacen parte de la formulación original de [27].

La discusión sobre la importancia de la elección de uno u otro modelo de red, en un entorno de mercados competitivos, se aborda a través de un análisis comparativo de los resultados del planeador estocástico usando tanto el modelo AC como el DC, se presenta en el **Capítulo 3**.

El **Capítulo 4** explora las principales técnicas de descomposición para problemas de programación estocástica no lineal encontradas en la literatura. Una revisión de las herramientas de cómputo paralelo complementarias para potencializar la rápida solución del problema de optimización concluye este capítulo.

Las propuestas para la implementación del planeador estocástico en su versión no lineal se han definido en el **Capítulo 5**. Una primera aproximación para la solución del problema mediante Relajación Lagrangiana con lagrangiano Aumentado (RLA) en su forma tradicional es discutido. Luego, se expone una propuesta basada en técnicas de RLA con una regla de actualización de los multiplicadores usando información de segundo orden. Por último, se presenta la propuesta basada en la Descomposición Generalizada de Benders (DGB) con subproblemas reformulados para ser siempre factibles y algunas estrategias de aceleración. Entre estas, se hace énfasis en las herramientas de cómputo paralelo al final del Capítulo.

El **Capítulo 6** esta dedicado a la presentación del procedimiento de solución, los diferentes experimentos realizados y su respectivo análisis. La primera parte trata sobre las suposiciones de modelado y la metodología de validación del modelo a través de la adaptación de la formulación incorporando las restricciones no lineales del modelo AC sobre un modelo de OPF extensible. Las simulaciones de la estrategia implementada se efectuaron sobre un

caso de prueba de pequeña dimensión, como el caso IEEE de 30 barras y sobre un caso de tamaño real como el sistema interconectado colombiano de 96 barras. Ambos casos de prueba permitieron estudiar las particularidades del modelo de descomposición y evaluar las estrategias de aceleración. Por último, se presentan las **Conclusiones y Trabajo Futuro** derivados de esta investigación.

Los **Anexos** presentan la formulación matemática detallada del mecanismo de almacenamiento, además de información de los elementos del sistema eléctrico de potencia colombiano y del caso IEEE de 30 barras que sirvieron como casos de prueba.

2. Planeador Estocástico Multi-Periodo para la Operación de una Red Eléctrica

La versión estocástica del modelo de OPF multi-periodo con restricciones de seguridad (en adelante MPSSOPF) usado en este trabajo para la programación de la operación del sistema de potencia eléctrica, corresponde al propuesto por Murillo-Sánchez *et al.* [27]. El problema de asignación y valoración de recursos como energía y reserva, en el contexto del mercado de día en adelanto, es resuelto en un único marco de co-optimización a través de múltiples escenarios para manejar la incertidumbre.

Esta formulación combina varios problemas de programación matemática que generalmente se resuelven por separado, de forma secuencial, durante la planeación de la operación del sistema de potencia, tales como:

1. La comisión de unidades en su versión estocástica que determine la programación óptima de la generación, abordando condiciones de incertidumbre en algunos parámetros del problema como la generación de fuentes intermitentes.

2. El problema del OPF con una representación completa de las restricciones no lineales de la red de transmisión, mediante el modelo AC. La versión simplificada o modelo DC también puede ser usada.

3. El manejo de dos tipos de incertidumbre presentes en la operación del sistema de potencia, ocasionados por:

    - La intermitencia y variabilidad en la producción de energía de las FER. Este tipo de incertidumbre puede ser descrita como un conjunto de distribuciones de probabilidad aproximadas por un grupo de estados del sistema con probabilidades asociadas, cada uno con una realización específica de los parámetros inciertos.
    - La ocurrencia de eventos de baja probabilidad o contingencias (desconexión de una línea de transmisión, la salida de operación de una unidad de generación, etc.). Simultáneamente se modela cada contingencia creíble como la modificación del OPF del caso base, incluyendo restricciones adicionales que limiten las desviaciones de generación de potencia desde el despacho del caso base. Así, se garantiza

la planeación para estados post-contingentes y el establecimiento de un esquema operativo sujeto a consideraciones de seguridad del tipo *N-1*. En consecuencia, cualquier acción siguiendo una contingencia debe ser determinada con la solución de un nuevo problema, ya que desde el punto terminal en el árbol de transición representando el estado contingente, no es posible determinar que la transición a otro estado contingente desde ese punto sea factible.

4. El procuramiento apropiado, tanto en cantidad como en precio, de servicios auxiliaries como las reservas necesarias para mantener la seguridad y confiabilidad del sistema eléctrico. MPSSOPF calcula dos tipos de reserva, determinadas por locación, de forma endógena y no pre-especificada por zona como se hace en estudios convencionales. Ambas reservas son co-optimizadas junto con las demás variables de optimización como parte del problema. El primer tipo, las reservas de contingencias, garantiza la factibilidad de los posibles redespachos en las contingencias. El segundo, las reservas de rampa de seguimiento de carga, garantiza la cantidad suficiente de rampa para el cambio de toma de carga.

5. La operación de dispositivos de almacenamiento despachado centralmente, cuyo prototipo es flexible para adaptarse a diferentes tipos de sistemas de almacenamiento. El mecanismo para el cálculo de almacenamiento tiene en cuenta el valor residual esperado de la energía almacenada en estados terminales, valor incluido en la función objetivo del problema.

6. La modelación de demanda capaz de responder a señales del mercado, como los precios, facilitada por las tecnologías de redes inteligentes (*smart grid*). Esta flexibilidad en la demanda puede modelarse como generadores negativos, con cuota de energía consumida preestablecida para el horizonte de planeación de la misma forma que se hace con los dispositivos de almacenamiento.

7. Asignación de cantidades óptimas de potencia contratada, superando la suposición de que las cantidades contratadas corresponden simplemente a los despachos de un caso base (sin contingencias). El uso de variables de cantidades contratadas junto con desigualdades que involucran los redespachos, las variables de reserva y los despachos base y post-contingentes, ofrece una mayor flexibilidad al operador del sistema de potencia en la toma de decisiones.

Esta formulación ha evolucionado a partir de diferentes estudios, como [21], [31], [32], [33], enfatizando alguna de las características del modelo. En todos los casos, el modelo de la red de transmisión usado ha sido el DC.

La siguiente sección ofrece una descripción conceptual de la estructura del problema de optimización, antes de abordar la formulación matemática del problema general.

2.1. Estructura del problema de programación estocástico para MPSSOPF

La Figura **2-1** ilustra un horizonte temporal de planeación de tres horas, dos escenarios de generación de energía renovable y tres contingencias, y servirá para describir la estructura del problema de optimización bajo estudio.

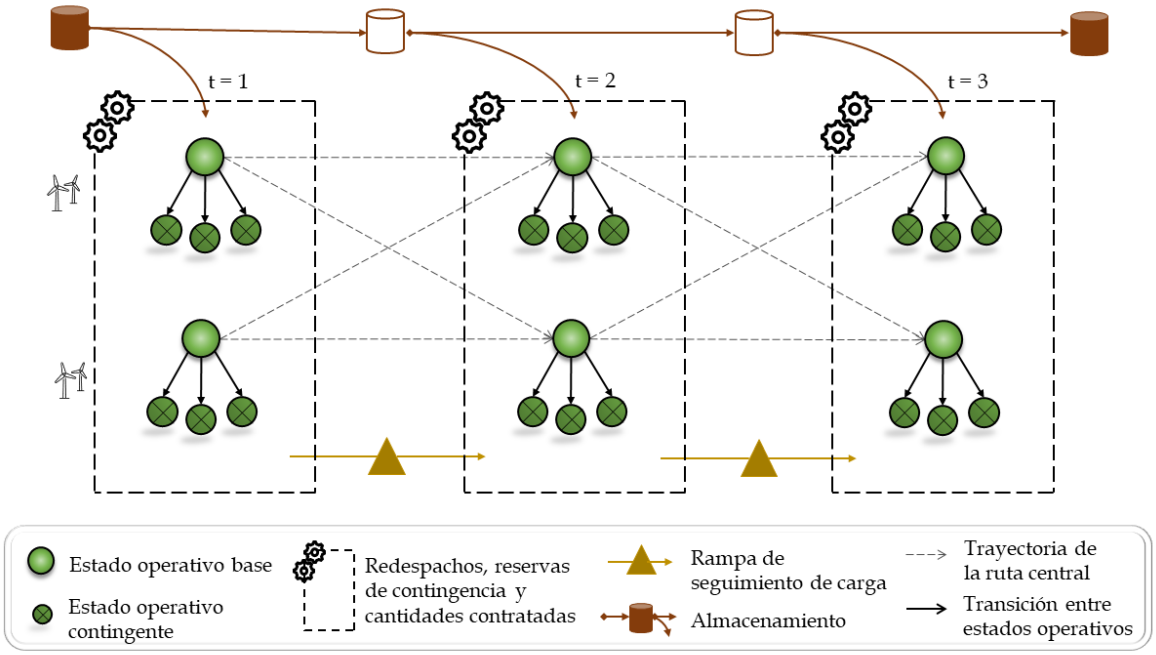

Figura 2-1.: Estructura general del problema de optimización estocástico multi-periodo

Cada estado operativo base es simbolizado por un círculo y los estados operativos post-contingentes por un círculo cruzado. Un estado operativo esta definido por las variables y las restricciones propias de un OPF. El estado operativo contingente es una copia modificada del estado operativo base, reflejando los cambios producidos por la contingencia y con los costos de generación apropiadamente escalados por la probabilidad de ocurrencia de la contingencia. El problema considera un estado operativo para cada periodo de tiempo t, escenario j de generación renovable y contingencia o caso base k.

El estado operativo base ($k = 0$) está enlazado con los estados operativos contingentes de un escenario y un periodo de tiempo determinado, mediante límites físicos de rampa, restricciones de reserva y de redespachos que hacen factible la transición desde el estado operativo base a los estados contingentes. Las reservas de contingencia, ascendente (r_+) y descendente (r_-), definen la máxima desviación desde la generación contratada requerida para cubrir

cualquiera de los estados operativos posibles en una etapa del horizonte temporal.

De una etapa temporal t a la etapa $t+1$ siguiente solo hay transiciones entre estados operativos base, representadas gráficamente por líneas discontinuas. Cada transición tiene una probabilidad asociada y contempla restricciones de rampa de seguimiento de carga. En cada periodo t del horizonte temporal de planeación se definen dos cantidades de reserva de rampa de seguimiento de carga (icono triángular), ascendente (δ_+) y descendente (δ_-), para cada generador, necesarias para garantizar suficiente capacidad de rampa para seguir los cambios en las condiciones del sistema. También se definen las cotas superior (s_+) e inferior (s_-) de las variables de almacenamiento, para cada periodo de tiempo, que dependen de la energía almacenada/descargada en el periodo precedente y son ajustadas por las pérdidas en las eficiencias de carga/descarga.

El modelo de optimización estocástico multietapa clásico tiene una estructura en forma de árbol. La explosión combinatórica que generalmente ocurre con árboles ortodoxos de escenarios, en programación estocástica multi-periodo, se puede evitar utilizando el concepto de recombinación de escenarios, como de hecho se asume en la presente formulación. Bajo este enfoque de programación estocástico, se encuentra un plan operativo óptimo siguiendo una ruta o envolvente central de alta probabilidad, tomando en cuenta un gran número de trayectorias sin sacrificar factibilidad estricta. Esta ruta central está definida por los estados operativos base en cada periodo y sus correspondientes transiciones factibles entre periodos. La propagación período a período está dirigida por una cadena de Markov, a través de una matriz de transición que relaciona el vector de probabilidades del estado operativo base en el tiempo t al vector de probabilidad de escenario en el tiempo $t+1$ [34].

Un aspecto central del modelo es la idea del contrato óptimo de energía del día en adelanto, cantidad que no necesariamente corresponde al despacho en alguno de los casos base, sino a la *postura contractual óptima de día en adelanto* frente a todas las cosas que podrían pasar al día siguiente. Es a partir de esta cantidad contratada que se definen los redespachos necesarios en cada flujo de potencia contemplado en ese período.

El resultado de una corrida del *software* comprende, para cada oferente y período en el horizonte, *i)* Una potencia óptima contratada de día previo; *ii)* Rangos de reserva rodante contratada, tanto hacia arriba como hacia abajo, a partir de la potencia contratada de día previo; *iii)* Rangos de rampa de seguimiento de carga tanto positiva como negativa. Los tres productos se fijan en cantidades y precios para cada oferente, implementando de hecho un mercado multidimensional con cinco productos: energía, cantidades de reservas rodantes incrementales y decrementales, y cantidades de rampa de seguimiento de carga positivas y negativas. Adicionalmente el *software* produce toda la información relacionada con cada OPF contemplado en el problema: despachos específicos, precios nodales, tensiones, ángulos

y multiplicadores de Lagrange asociados con cada restricción.

2.2. Formulación matemática de MPSSOPF

El problema de optimización general es del tipo no lineal mixto-entero, con variables continuas θ, V, p, q, p_c, p_+, p_-, r_+, r_-, δ_+, δ_-, p_{sc}, p_{sd}, s_+, s_- y variables binarias u, v y w, agrupadas en la variable de optimización x. En adelante, la simbología usada en la formulación del problema corresponde a la **Lista de símbolos para MPSSOPF**.

2.2.1. Función objetivo

La función objetivo (2-1) minimiza el costo esperado del sistema, constituido por cinco componentes: el costo de generación considerando un conjunto de escenarios estocásticos de generación renovable por FER y un conjunto de escenarios de contingencia para un horizonte temporal, típicamente de 24 horas, en un esquema de planificación diaria; el costo del redespacho sobre las posibles desviaciones de potencia con respecto a las cantidades contratadas; el costo de dos tipos de reservas operativas; el costo de las rampas de toma de carga; el valor de la energía residual del sistema de almacenamiento y el costo de arranque y parada.

$$\min_{x} \; f_p\left(p, p_+, p_-\right) + f_r\left(r_+, r_-\right) + f_\delta\left(p\right) + f_{lf}\left(\delta_+, \delta_-\right) + f_s\left(p_{sc}, p_{sd}\right) + f_{uc}\left(v, w\right) \tag{2-1}$$

Cada componente, expresado en términos de las variables de optimización, se describe a continuación.

1. **Costo del despacho y redespacho de potencia activa**.
   El costo de despacho corresponde al costo esperado de producción o de oferta, calculado para cada generador en cada estado operativo. Las unidades térmicas tienen costos de producción asociados con el costo del combustible (sólido, líquido o gaseoso) y el costo de mantenimiento, entre otros; mientras que las unidades hidroeléctricas y FER se asumen con un costo de generación prácticamente nulo. La función de costos puede ser lineal, lineal a tramos o cuadrática.

   Por su parte, el costo de redespacho es un costo incremental impuesto a las posibles desviaciones del despacho, hacia arriba o hacia abajo, desde las cantidades contratadas en el mercado del día en adelanto. Este además funciona como un indicador de la renuencia de los generadores a variar su producción de potencia desde las cantidades

contratadas en el mercado.

$$f_p\left(p, p_+, p_-\right) = \sum_{t\in T}\sum_{j\in J^t}\sum_{k\in K^{tj}} \psi_\alpha^{tjk} \sum_{i\in I^{tjk}} \left[C_P^{ti}\left(p^{tijk}\right) + C_{P_+}^{ti}\left(p_+^{tijk}\right) + C_{P_-}^{ti}\left(p_-^{tijk}\right)\right] \quad (2\text{-}2)$$

La probabilidad de contingencia ψ_α^{tjk} está ajustada por la fracción del periodo temporal t que es gastado en el caso base antes de que la contingencia ocurra. Esta puede ser determinada apropiadamente para un caso base ($k = 0$) o un caso contingente de acuedo con lo siguiente:

$$\psi_\alpha^{tjk} = \begin{cases} \psi^{tj0} + \alpha \sum_{\kappa\in K^{tj}\neq 0} \psi^{tj\kappa}, & k = 0 \\ (1-\alpha)\psi^{tjk}, & \forall k \in K^{tj} \neq 0 \end{cases} \quad (2\text{-}3)$$

2. **Costo de las reservas de contingencia**.
El costo de reservas para eventos de baja probabilidad de ocurrencia se determina sobre las cantidades de reserva de contingencia asignadas a cada unidad de generación por periodo de tiempo, ponderado por la probabilidad γ^t de llegar a la t-ésima etapa.

$$f_r\left(r_+, r_-\right) = \sum_{t\in T}\gamma^t \sum_{i\in I^t}\left[C_{R_+}^{ti}\left(r_+^{ti}\right) + C_{R_-}^{ti}\left(r_-^{ti}\right)\right] \quad (2\text{-}4)$$

3. **Costo de rampa de seguimiento de carga (uso y desgaste)**.
Es el costo asociado con el desgaste o fallos acelerados de los componenetes de las unidades de generación cuando están sometidos a tensiones y fatiga, producto de la variación de la salida de potencia de las unidades.

$$f_\delta\left(p\right) = \sum_{t\in T}\gamma^t \sum_{\substack{j_1\in J^{t-1} \\ j_2\in J^t}} \Phi^{tj_2j_1} \sum_{i\in I^{tj_20}} C_\delta^i\left(p^{tij_20} - p^{(t-1)ij_10}\right)^2 \quad (2\text{-}5)$$

Este costo está afectado por la probabilidad de llegar a la t-ésima etapa (γ^t) y por la probabilidad de la transición al escenario j_2 en el periodo t desde el escenario j_1 en el periodo $t-1$ ($\Phi^{tj_2j_1}$).

4. **Costo de reserva de rampa de seguimiento de carga**.
Este costo, ponderado por probabilidad γ^t de llegar a la t-ésima etapa en cada transición individual, representa las ofertas del mercado para un potencial producto de reserva de rampa, sobre las capacidades máximas y mínimas de rampa procuradas en la programación óptima, para satisfacer la variabilidad de la carga neta de una hora a otra.

$$f_{lf}\left(\delta_+, \delta_-\right) = \sum_{t\in T}\gamma^t \sum_{i\in I^t}\left[C_{\delta_+}^{ti}\left(\delta_+^{ti}\right) + C_{\delta_-}^{ti}\left(\delta_-^{ti}\right)\right] \quad (2\text{-}6)$$

5. **Valor de la energía residual almacenada esperada en estados terminales**.
   Es el valor asociado a la cantidad esperada de energía almacenada residual en estados terminales, independientemente del estado en el que ocurra.

$$f_s\left(p_{sc}, p_{sd}\right) = -\left(C_{sc}^T p_{sc} + C_{sd}^T p_{sd}\right) \tag{2-7}$$

6. **Costo de arranque y parada**.
   Los costos de arranque (en frío o en caliente) y de parada se asocian a unidades de generación térmica. El costo de arranque está relacionado con el desgaste que sufre la máquina durante el proceso de arranque y al costo del combustible usado durante tal proceso.

$$f_{uc}\left(v, w\right) = \sum_{t \in T} \gamma^t \sum_{i \in I^t} \left(C_v^{ti} v^{ti} + C_w^{ti} w^{ti}\right) \tag{2-8}$$

2.2.2. Restricciones

La minimización (2-1) está sujeta a restricciones que pueden ser de dos tipos, según las variables vinculadas. Existen restricciones en función de las variables propias de cada uno de los estados operativos, por ejemplo, aquellas que describen el comportamiento de la red de transmisión (2-9)-(2-11), y los límites de variables tales como (2-12)-(2-34). Las restricciones que vinculan diferentes estados operativos, denominadas restricciones acopladoras, aseguran que la transición desde un punto operativo al siguiente sea factible. Típicamente, modelan límites como la capacidad de rampa de las unidades de generación (2-18)-(2-22), el cambio de energía almacenada (2-23)-(2-32), o las restricciones de arranque y parada de las unidades de generación (2-35).

1. **Restricciones estándar del OPF**.
   Las restricciones (2-9) - (2-11) son no lineales y nominalmente no-convexas en las variables V y θ. Sin embargo, en sistemas reales las no convexidades pueden ser poco marcadas, por ejemplo, gracias a las pérdidas de transmisión.

   - Ecuaciones no lineales de balance de potencia activa y reactiva.

$$g_P^{tjk}\left(\theta^{tjk}, V^{tjk}, p^{tjk}\right) = 0 \tag{2-9}$$
$$g_Q^{tjk}\left(\theta^{tjk}, V^{tjk}, q^{tjk}\right) = 0 \tag{2-10}$$

   - Límites de transmisión de potencia. Esta restricción consiste de dos conjuntos, uno para cada extremo de la línea de transmisión. Los límites de transmisión generalmente se expresan en términos de la potencia aparente, pero también se

pueden estar expresados en términos de la potencia activa o del flujo de corriente.

$$h^{tjk}\left(\theta^{tjk}, V^{tjk}\right) \leq 0 \tag{2-11}$$

- Límites en las magnitudes y ángulos de tensión.

$$\theta^{tjk}_{min} \leq \theta^{tjk} \leq \theta^{tjk}_{max} \tag{2-12}$$
$$V^{tjk}_{min} \leq V^{tjk} \leq V^{tjk}_{max} \tag{2-13}$$
$$\tag{2-14}$$

2. **Restricciones de contingencia**.
La Figura **2-2** ilustra la estructura de reservas, definiendo las variables de desviación de potencia y las de reserva para el generador i en el periodo t.

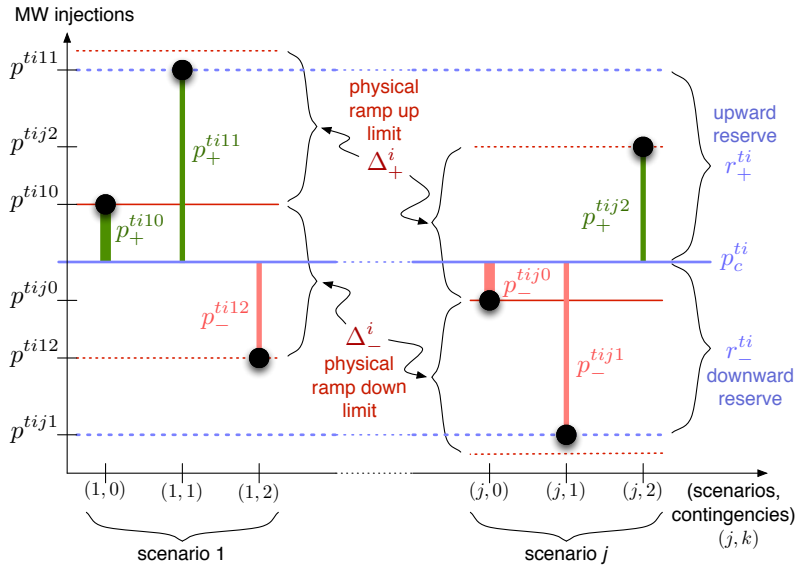

Figura 2-2.: Estructura de las reservas para el generador i en el periodo t.
Tomado de [27].

- Variables de despacho y redespacho contratadas.
Las reservas de contingencias, hacia arriba (r_+) y hacia abajo (r_-), asignadas a cada generador se definen respectivamente por la máxima y mínima desviación con respecto a las cantidades contratadas, que pueden resultar de los redespachos en una contingencia.

$$0 \leq p^{tijk}_{+} \leq r^{ti}_{+} \leq R^{ti}_{Pmax+} \tag{2-15}$$
$$0 \leq p^{tijk}_{-} \leq r^{ti}_{-} \leq R^{ti}_{Pmax-} \tag{2-16}$$
$$p^{tijk} - p^{ti}_{c} = p^{tijk}_{+} - p^{tijk}_{-} \tag{2-17}$$

- Límites de rampa en transiciones desde los casos base a los casos de contingencia. Limitan la rapidez con la cual puede cambiar la salida de un generador de un caso base a un caso post-contingente.

$$-\Delta_{-}^{i} \leq p^{tijk} - p^{tij0} \leq \Delta_{+}^{i}, k \neq 0 \tag{2-18}$$

3. **Restricciones inter-temporales**.

- Límites de rampa de seguimiento de carga y reservas.
  Estas restricciones garantizan suficiente capacidad de rampa disponible para seguir los cambios en las condiciones del sistema entre los despachos de los casos base ($k = 0$), en periodos de tiempo consecutivos.

  La reserva de rampa de seguimiento de carga positiva (δ_+) corresponde a la diferencia entre el menor despacho entre los estados base en el periodo t y el mayor despacho entre los estados base en el periodo $t+1$, representando la capacidad máxima de aumento de rampa necesaria en un sistema optimizado en ausencia de una contingencia. La reserva de rampa de seguimiento de carga negativa (δ_-) se define de forma semejante:

$$0 \leq \delta_{+}^{ti} \leq \delta_{max+}^{ti} \tag{2-19}$$
$$0 \leq \delta_{-}^{ti} \leq \delta_{max-}^{ti} \tag{2-20}$$
$$p^{tij_2 0} - p^{(t-1)ij_1 0} \leq \delta_{+}^{(t-1)i},\ j_1 \in J^{t-1},\ j_2 \in J^{t} \tag{2-21}$$
$$p^{(t-1)ij_1 0} - p^{tij_2 0} \leq \delta_{-}^{(t-1)i},\ j_1 \in J^{t-1},\ j_2 \in J^{t} \tag{2-22}$$

- Restricciones de almacenamiento.
  La inyección de potencia asociada a cada unidad de almacenamiento esta dividida entre dos variables, una de carga (p_{sc}) para los valores negativos de la inyección y una de descarga (p_{sd}) para los positivos, y cada uno puede estar afectado por una eficiencia de carga y de descarga, respectivamente.

$$p^{tijk} = p_{sc}^{tijk} + p_{sd}^{tijk} \tag{2-23}$$
$$p_{sc}^{tijk} \leq 0 \tag{2-24}$$
$$p_{sd}^{tijk} \geq 0 \tag{2-25}$$

  El incremento neto de la energía almacenada debido a la carga/descarga de la unidad i en el estado tjk es definido como:

$$s_{\Delta}^{tijk} \equiv -\Delta \left(\eta_{in}^{i} p_{sc}^{tijk} + \frac{1}{\eta_{out}^{i}} p_{sd}^{tijk} \right) \tag{2-26}$$

Los límites inferior y superior en las cantidades de energía almacenada, para cada unidad y periodo de tiempo a través de todos los escenarios, aseguran que todos los eventos posibles de carga y descarga asociados con los estados en la ruta central sean factibles con respecto a los límites en la capacidad de energía de la unidad de almacenamiento. El límite inferior en la energía almacenada al final del periodo t depende del límite inferior en el tiempo $t-1$ y la máxima reducción de esa energía en los escenarios base en el tiempo t. De una forma análoga se puede establecer el valor del límite superior en la energía almacenada al final del periodo t. La Figura **2-3** es un ejemplo en el que los límites superior e inferior de una unidad de almacenamiento dada se extienden a sus límites de capacidad. En esta, los rombos representan la energía almacenada esperada y las líneas punteadas corresponden a los despachos del almacenamiento, los cuales pueden variar o no por escenario.

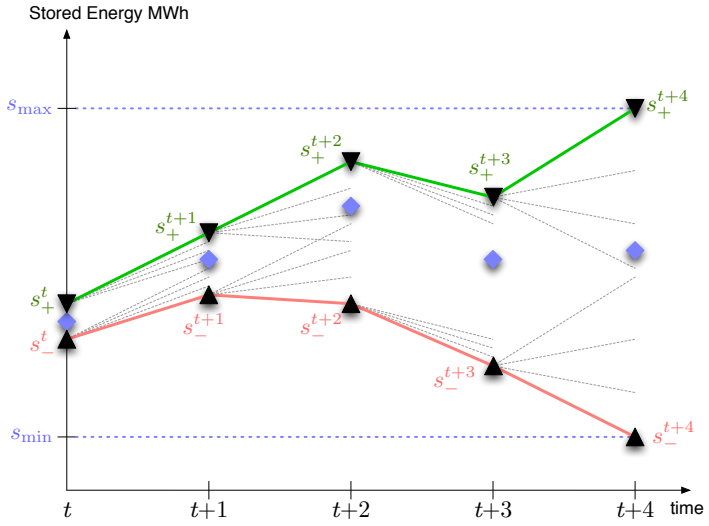

Figura 2-3.: Límites superior e inferior en las variables de estado de almacenamiento de una unidad. Tomado de [27].

$$s_-^{ti} \geq S_{min}^{ti} \tag{2-27}$$

$$s_+^{ti} \leq S_{max}^{ti} \tag{2-28}$$

$$s_-^{ti} \leq s_-^{(t-1)i} + s_\Delta^{tij0} - \Delta \frac{\eta_{loss}^i}{2} \left(s_-^{ti} + s_-^{(t-1)i} \right) \tag{2-29}$$

$$s_+^{ti} \leq s_+^{(t-1)i} + s_\Delta^{tij0} - \Delta \frac{\eta_{loss}^i}{2} \left(s_+^{ti} + s_+^{(t-1)i} \right) \tag{2-30}$$

$$S_{min}^{ti} \leq s_-^{(t-1)i} + \alpha s_\Delta^{tij0} + (1-\alpha)\, s_\Delta^{tijk}, k \neq 0 \tag{2-31}$$

$$S_{max}^{ti} \leq s_+^{(t-1)i} + \alpha s_\Delta^{tij0} + (1-\alpha)\, s_\Delta^{tijk}, k \neq 0 \tag{2-32}$$

Esta formulación para el almacenamiento permite tomar la decisión óptima entre usar

el almacenamiento para hacer arbitrage o mitigar la incertidumbre proveniente de las FER, a costa de no saber el valor almacenado en cada momento del tiempo sino sus límites inferior y superior.

Dado que el mecanismo de seguimiento del almacenamiento esperado es matemáticamente complejo se ha dejado para ilustración del lector al final de la tesis en el Anexo A.

4. **Comisión de unidades**.

   - Límites de inyección y comisionamiento.
     Los límites de inyección de potencia activa y reactiva están dados por unas cantidades máxima y mínima, condicionado por el estado operativo u de la unidad de generación. Si el generador no esta comisionado, es decir $u = 0$, el generador no inyecta potencia al sistema.

$$u^{ti} P_{min}^{tijk} \leq p^{tijk} \leq u^{ti} P_{max}^{tijk} \tag{2-33}$$
$$u^{ti} Q_{min}^{tijk} \leq q^{tijk} \leq u^{ti} Q_{max}^{tijk} \tag{2-34}$$

   - Eventos de arranque y parada.
     Para cada unidad de generación, las variables binarias v^{ti} y w^{ti} permiten determinar en que momento la unidad es encendida o apagada, tomando un valor de uno según corresponda.

$$u^{ti} - u^{(t-1)i} = v^{ti} - w^{ti} \tag{2-35}$$

   - Tiempos mínimos de arranque y parada.
     Esta restricción asegura que no se de un cambio en las decisiones binarias de arranque y parada hasta que no se alcancen los tiempos mínimos de encendido y apagado de las unidades de generación.

$$\sum_{y=t-\tau_i^+}^{t} v^{yi} \leq u^{ti}, \qquad \sum_{y=t-\tau_i^-}^{t} w^{yi} \leq 1 - u^{ti} \tag{2-36}$$

   - Restricciones de integralidad.
     Estas restricciones permiten que el modelo capture la naturaleza discreta de las variables de decisión binarias.

$$u^{ti} \in \{0,1\}, \quad v^{ti} \in \{0,1\}, \quad w^{ti} \in \{0,1\} \tag{2-37}$$

2.2.3. Sistemas de almacenamiento de energía

La formulación del almacenamiento es general y se adapta a cualquier tipo de tecnología incluyendo: el almacenamiento por bombeo de agua, almacenamiento térmico, almacenamiento por unidades de aire comprimido, baterías sin una trayectoria definida de carga-descarga, celdas de combustible, entre otros. Las plantas hidráulicas con embalse y las cargas flexibles también se pueden modelar como sistemas de almacenamiento de energía.

Las inyecciones de potencia de las unidades de almacenamiento al sistema, tanto en casos base como contingentes, se dividen en una parte positiva (potencia de descarga, p_{sd}) y en una parte negativa (potencia de carga, p_{sc}), con una eficiencia diferente para cada parte. Adicionalmente, cada unidad de almacenamiento tiene dos variables para los límites superior (s^t_+) e inferior (s^t_-) en la cantidad de energía almacenada por periodo. Los valores de estas variables tienen una dependencia temporal, es decir, el límite superior o inferior del almacenamiento en el periodo t depende del despacho de las unidades en los periodos precedentes, ajustados por las eficiencias de carga y descarga, así como de la energía inicial almacenada en $t = 0$. Lo anterior garantiza que se respeten los límites de las unidades de almacenamiento de energía en cualquier realización de los escenarios considerados. Esta característica es importante para la administración central del almacenamiento, como se realiza en este trabajo.

El uso óptimo del almacenamiento para hacer arbitraje depende en parte del costo final asignado a la energía almacenada. Si su costo es nulo, la energía almacenada siempre se utiliza, a menos que exista un costo alto para descargar energía que proporcione un umbral alto para la descarga. Por ejemplo, cuando los precios nodales en el estado terminal son muy bajos no sería óptimo descargar energía y lo mejor sería esperar hasta un período posterior cuando el precio sea superior. Lo mismo se puede argumentar para el proceso de carga, si existe un costo bajo que proporcione un umbral bajo para la carga. Entonces es óptimo cargar la unidad de almacenamiento si el precio está por debajo del costo de carga. Dado el caso de que el precio este entre los dos umbrales, lo óptimo es no hacer nada y guardar la energía almacenada para usarla más tarde.

En cuanto al uso del almacenamiento en estados contingentes, este debe respetar los límites de almacenamiento y está sujeto a un factor α que especifíca la fracción del tiempo t que es gastada en el caso base, antes que ocurra la contingencia. Las unidades de almacenamiento pueden tener energía residual almacenada en estados contingentes, dado que estos son puntos finales en el árbol de transición.

Por último, se permite elegir si la operación del almacenamiento será o no cíclica. Si la operación es cíclica, la primera hora del horizonte de planeación sucede a la última, de tal manera que la energía inicial corresponde a la energía residual en la hora final.

2.2.4. Cargas con flexibilidad temporal

Las cargas con flexibilidad temporal o demandas despachables, que son sensibles al precio, se pueden representar de forma estándar como generadores con inyecciones negativas de potencia activa y con costos asociados negativos. Estas cargas deberan cumplir con una cuota de consumo a lo largo del horizonte temporal. El límite mínimo de potencia de este generador ficticio puede ser igual a una fracción del valor negativo de la cuota de consumo y el límite máximo igual a cero, permitiendo que la carga sea completamente reducida en función de los precios.

Otro aspecto a considerar en el modelado de las cargas flexibles, cuando se utiliza el modelo AC de las restricciones de la red, tiene que ver con su despacho de potencia reactiva. Si bien las cargas flexibles se modelan como generadores, el despacho reactivo no tomaría cualquier valor dentro de sus límites máximo y mínimo definido como lo haría cualquier generador. En su lugar, el modelo asume que las cargas flexibles mantienen un factor de potencia constante, lo que sería un comportamiento mas acorde con lo esperado para una carga normal.

2.3. Modelado de cadenas hidrológicas

La operación de las plantas hidroeléctricas está fuertemente influenciada por la configuración del sistema hídrico donde se encuentran instaladas. Algunas plantas, denominadas controlables o despachables, pueden almacenar las contribuciones de diversos afluentes en embalses y optimizar su uso de acuerdo al costo de oportunidad. Otras, en cambio, deben turbinar todo el agua que reciben en el momento, como las plantas filo de agua o de pasada, y se denominan no despachables. Las plantas hidroeléctricas pueden estar ubicadas de manera consecutiva (o en cascada), desde la perspectiva de las contribuciones hídricas de entrada y salida. El agua liberada en la planta aguas arriba contibuye al flujo de entrada de la planta aguas abajo y por ende, controla su producción de energía.

En este trabajo se modela la dependencia de producción de energía existente entre plantas ubicadas en cadenas hidráulicas, donde las plantas con embalse ubicadas río arriba controlan la producción de energía de las plantas sin embalse río abajo. En ese caso, se utiliza un conjunto adicional de restricciones lineales como (2-38) para los generadores en esas cadenas, para cada estado operativo.

$$p^{tijk}_{downstream} - c^{i}_{f}\ p^{tijk}_{upstream} \leq 0 \tag{2-38}$$

donde, $p^{tijk}_{upstream}$ es la potencia activa de la planta i, ubicada río arriba, en el flujo de potencia tjk; $p^{tijk}_{downstream}$ es la potencia activa de la planta i, ubicada río abajo, en el flujo de potencia tjk; y c^{i}_{f} es el factor de control de la planta i ubicada río arriba.

Es preciso anotar que no se incluyen de forma explícita restricciones sobre cuestiones hidráulicas, tales como relación caudal-potencia o balance hídrico, volumen máximo y mínimo de los embalses, etc. Se asume que las plantas hidráulicas con embalse tienen asignada una cuota energética para el horizonte de planeación.

2.4. Comparación de la complejidad de diferentes formulaciones para la programación de la generación

Esta sección ilustra las diferencias existentes en el modelo matemático resultante de tres formulaciones para el problema de la programación de la generación, basados en el modelo de Coordinación Hidro-Térmica (CHT) de [35], el OPF seguro ante contingencias (SCOPF) de [13] extendido para abarcar múltiples periodos de tiempo (MP-SCOPF), y la formulación de MPSSOPF, con modelos AC y DC, expuestos en las secciones anteriores.

Tanto CHT, MP-SCOPF y MPSSOPF-NL consideran las restricciones del modelo AC de la red de transmisión y sólo MPSSOPF-L usa el modelo DC. Aunque MPSSOPF-L originalmente puede formularse con variables binarias para comisionamiento de unidades, se omitieron estas para poder comparar sus dimensiones con las de su contraparte no lineal. En consecuencia, CHT es la única formulación conteniendo variables binarias. Este no contempla contingencias y el acoplamiento temporal se da a través de las ecuaciones de balance hídrico de embalses.

En primer lugar, se compara el tamaño de las formulaciones mencionadas mediante un caso específico como el constituido por el sistema de potencia de tres barras, tres generadores, tres líneas de transmisión y tres cargas, presentado en la Figura **2-4**. El generador de la barra 1 es hidroeléctrico y tiene un embalse, los otros dos generadores son térmicos. Para todos los casos, el periodo de planeación es de veinticuatro horas y en los estudios que modelan contingencias, se considera una sola contingencia.

La Figura **2-5** compara el número de variables y de restricciones en cada modelo. Entre estos, MPSSOPF-NL cuenta con el mayor número de variables y de restricciones. Igualmente se observa que el modelo lineal de la misma formulación resulta en un problema de mayores dimensiones que las dos restantes (CHT y MP-SCOPF). Lo anterior puede conducir a que MPSSOPF-NL requiera una cantidad considerable de cálculo para alcanzar la solución del problema, en comparación con las otras formulaciones, dada la relación directa entre el tamaño del problema y el tiempo de solución, que aumenta consecuentemente.

Las estructuras matemáticas del jacobiano de las condiciones de optimalidad de cada problema se presentan en la Figura **2-6**. Una sola representación gráfica para MPSSOPF-L y

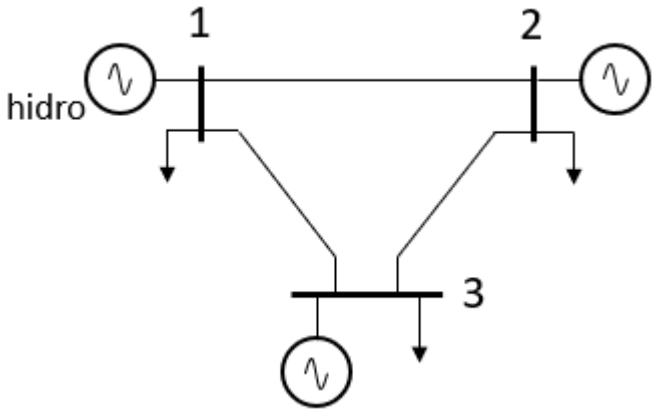

Figura 2-4.: Sistema de potencia de 3 barras

MPSSOPF-NL es necesaria para esquematizar dicha estructura matricial y corresponde a MPSSOPF en la mencionada figura. Por su propia naturaleza, estos problemas exhiben una matriz jacobiana rala, es decir, la mayoría de sus elementos son cero. En los tres casos se observan estructuras genéricas que se replican y que corresponden al mismo tipo de restricciones sobre variables de diferentes momentos en el tiempo. De igual forma, se observan estructuras originadas por las restricciones intertemporales que se extienden para incluir variables de más de un periodo temporal.

La matriz para CHT tiene estructura diagonal por bloques, en el que cada bloque simboliza un OPF para el tiempo t, con un borde horizontal que representa las restricciones intertemporales del balance hídrico en el embalse de la barra 1.

La estructura para MP-SCOPF es similar a la de CHT, con la particularidad de que cada bloque diagonal en la matriz principal además tiene una estructura diagonal, dada por un OPF para cada contingencia y caso base en el mismo periodo de tiempo t, con bloque vertical representando a las variables de potencia activa y a las magnitudes de tensión en las barras de tensión controlada, cantidades que deben ser iguales en cada periodo t. El borde horizontal de la matriz principal pertenece a las restricciones intertemporales de rampa.

Por su parte, la matriz jacobiana para MPSSOPF tiene estructura diagonal por bloques con borde horizontal y borde vertical, estructura que es conocida como forma de flecha. Cada bloque diagonal de la matriz principal contiene todos los OPF del mismo periodo de tiempo t, dispuestos igualmente en bloques diagonales, uno por cada OPF de un caso base o contingente, para cada escenario de planeación. Junto a estos se distingue un borde vertical, asociado a las variables de reserva de contingencia, y unos pequeños bloques adosados a los bloques diagonales que representan las restricciones en las variables de redespacho de potencia y de carga y descarga de las unidades de almacenamiento. Con respecto a los bloques

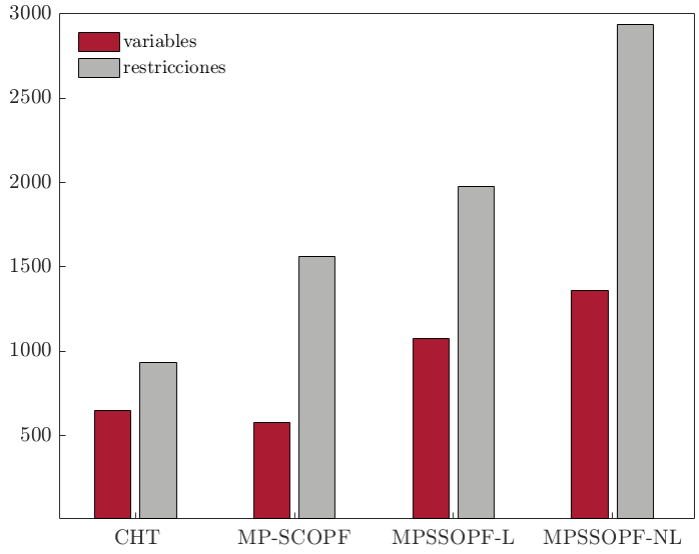

Figura 2-5.: Comparación del tamaño de las diferentes formulaciones para el problema de programación de la generación en el sistema de potencia de tres barras

en los bordes horizontal y vertical de la matriz principal, estos están relacionados con restricciones intertemporales y sus respectivas variables, tales como: las reservas de rampa de seguimiento de carga, las restricciones del mecanismo de almacenamiento y las restricciones del comisionamiento de unidades, si estas últimas están incluidas en la formulación.

De lo anteriormente expuesto se evidencia que MPSSOPF resulta mucho más complejo que CHT y MP-SCOPF y que el procedimiento de solución, para su versión no lineal (MPSSOPF-NL), requerirá una cantidad importante de recursos computacionales y tiempo de ejecución gracias al número de variables y restricciones que puede contener un problema de optimización con esta formulación. A saber, a la fecha no se han resuelto problemas de dimensiones importantes con un modelo tan complejo como MPSSOPF-NL, lo que representaría un avance en el estado del arte.

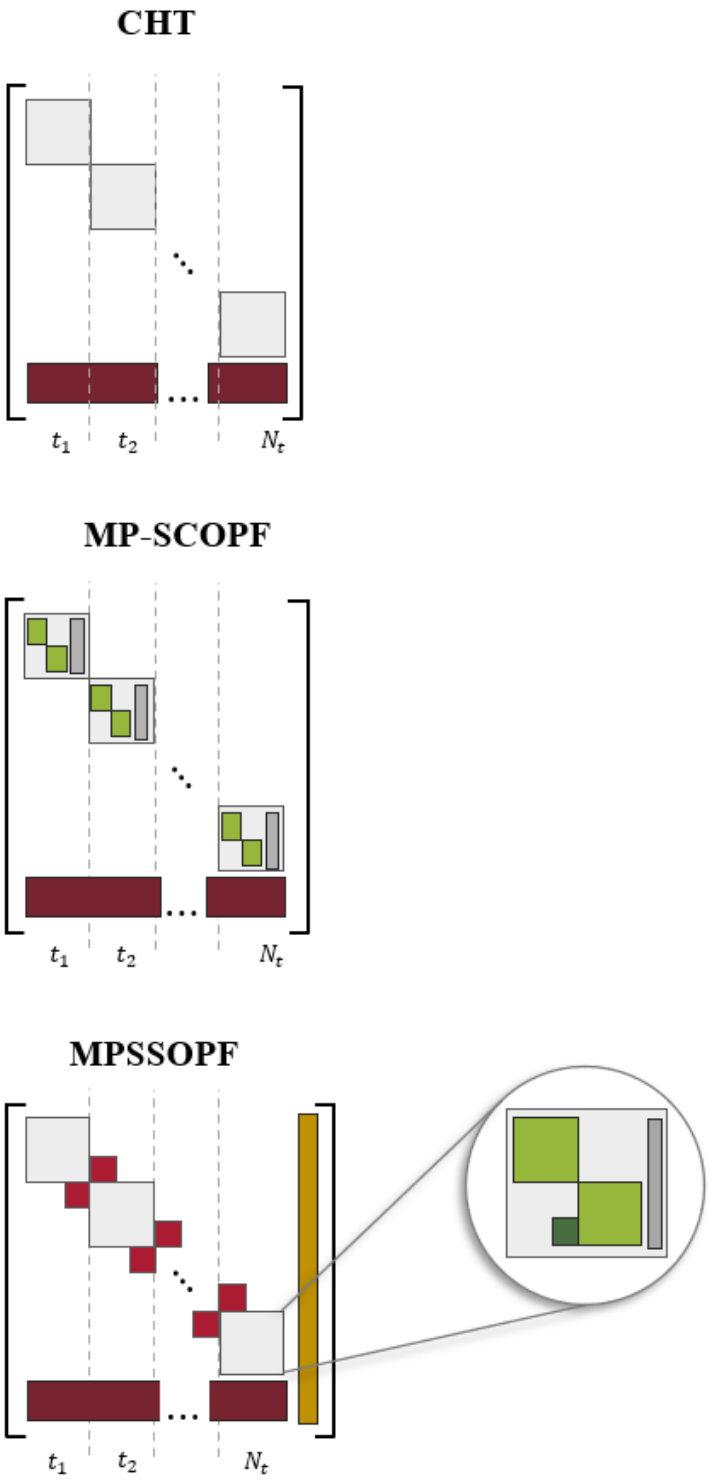

Figura 2-6.: Comparación de estructuras matriciales de tres diferentes formulaciones para el problema de programación de la generación de potencia

3. Importancia de Usar un Modelo AC o DC de la Red de Transmisión en el Cierre de un Mercado Multidimensional

Este capítulo es reproducción parcial del artículo [36], publicado en la revista *IEEE Latin America Transactions*. Se omite todo lo relacionado con la presentación del marco de optimización estocástico multi-periodo, introducido de forma extensa en el Capítulo 2.

Abstract - *As the penetration of renewable energy sources increases, their variability and uncertainty have pushed the development of secure stochastic approaches to solving the secure day-ahead and intra-day multi-period optimal power flow. Some of these approaches, in addition to settling the energy market, also procure other products that generators can offer such as spinning reserves and ramping capability, the latter of which becomes even more important under high penetration of renewable sources. The resulting large scale nonlinear minimization problem makes using the more simple DC flow model of the network appealing, at least in the day-ahead stage. This work focuses on the impact of using an AC vs. a DC model of the network in the results of these multi-dimensional markets, using the Colombian system as a test case. The study implies that although the overall energy allocations may not change much from one model to another, other products may exhibit very different allocations under different network models.*

Index Terms - *Electricity markets, operation planning under uncertainty, optimal scheduling, power system modeling, wind energy integration.*

3.1. Introducción

La perspectiva de tener sistemas de generación y transmisión con una penetración cada vez más alta de fuentes variables de energía ha llevado a considerar diferentes mecanismos para asegurar algunos servicios auxiliares que se vuelven especialmente importantes bajo tal con-

dición. En particular, dicha variabilidad y la necesidad de la seguridad operativa requieren de cantidades apropiadas de reserva rodante (para enfrentar contingencias) y de capacidad de rampa de seguimiento de carga (para enfrentar la variabilidad de las fuentes no despachables renovables). Un enfoque para determinar la asignación de estos recursos entre los participantes en el mercado es la co-optimización de las asignaciones de energía, reserva rodante y reserva de rampa en un mismo problema de optimización. El problema resultante es más complejo que una coordinación hidrotérmica tradicional por el hecho de tener más variables de decisión, y si además se incluye el modelo de la red (para respetar límites de transmisión), las restricciones de seguridad (para asegurar que la ocurrencia de contingencias no deje al sistema en un estado operativo no factible) y la realización de diferentes escenarios de generación de fuentes variables, dicho problema de optimización es de un tamaño formidable. Ante un problema como ese, se vuelve deseable el uso de modelos simplificados de red tal como el modelo DC en vez de un modelo completo de corriente alterna AC. Sin embargo, como el problema de optimización resultante llevará a cabo el cierre de un mercado multidimensional con diferentes tipos de productos, se genera la siguiente inquietud: ¿Es la asignación de cantidades de los diferentes productos entre los participantes sensible al tipo de modelo de red utilizado en la optimización? Se trata de una pregunta acerca de la justicia en la asignación de cantidades en el cierre de un mercado multidimensional y por lo tanto una pregunta fundamental para la creación de posibles mercados de servicios auxiliares.

Entre las propuestas de co-optimización existentes hay modelos que además de considerar la incertidumbre proveniente de las contingencias, también enfrentan la incertidumbre en la demanda [20, 22, 24] y en la generación de fuentes de energía renovable (FER) [19–21, 27]. En algunas formulaciones se considera la provisión de reservas rodantes, que pueden ser localizadas [21, 24, 27, 33, 37–42] o no [7, 19, 20, 22, 43] y en ciertos casos suministrada por los consumidores [7, 40, 41, 44, 45]. Algunos de estos modelos de optimización implementan diferentes variaciones de un flujo óptimo multi-periodo estocástico seguro (MPSSOPF), bien sea bajo el modelo de red AC [21, 22, 27, 33, 37, 39, 46] o el DC [7, 20, 27, 38, 43, 44].

La adición del modelo de red en un modelo estocástico típicamente se basa en la replicación de la red, donde cada instancia de la misma representa un estado operativo posible; por lo tanto, el número de variables en el modelo tiene un componente que es proporcional al número de réplicas de la red. La complejidad inherente al modelo de la red es un componente fundamental de la complejidad del modelo total. Por ello es importante saber si es factible el utilizar la representación DC del flujo en la red, resultando en un modelo que se puede resolver más fácilmente; o si el cierre más preciso del mercado exige el uso de un modelo AC de la red, requiriendo la solución de un complejo programa no lineal.

Para contestar esta pregunta en el contexto del mercado colombiano de energía eléctrica, se ha implementado un modelo [27] reconocido como uno de los más completos en cuanto a

multi-dimensionalidad y número de variables [30], tanto en versión DC como en versión AC y se ha procedido a comparar sus resultados en una representación de 96 barras del sistema colombiano derivada de datos reales. El estudio usa datos históricos de demanda horaria, disponibilidad hídrica, perfiles de viento y precios de oferta en el mercado colombiano.

Este tipo de estudio explícitamente enfocado en mercados multi-dimensionales de energía, reserva y rampa, con flujo óptimo estocástico seguro y alta penetración de renovables, no ha sido llevado a cabo aún. Existen desarrollos que enfocan el MPSSOPF de gran escala [13], pero no necesariamente en el marco de mercados multi-dimensionales y menos en el contexto de evaluar las consecuencias de usar el modelo DC o AC. Es en dicho aspecto que el presente trabajo aporta resultados nuevos y de gran incidencia para la estructuración de posibles mercados multidimensionales en el futuro.

3.2. Caso de estudio

El caso de estudio corresponde a un modelo del sistema interconectado colombiano basado en la configuración de 2017 [47]. Este sistema de potencia está constituido por 96 barras, 49 plantas generadoras, 206 líneas de transmisión y 19 unidades de almacenamiento de energía[1].

Tabla 3-1.: Características del sistema de potencia colombiano de 96 barras

Topología	Red de 96 barras - 220 y 500 kV
Generación	16.310 MW potencia activa pico instalada
	10.295 MW plantas hidroeléctricas
	3.850 MW plantas de gas natural
	1.365 MW plantas de carbón
	800 MW Granja eólica
Demanda	9.116 MW potencia activa pico
	3.527 MVAR potencia reactiva pico
Almacenamiento	19 unidades hidroeléctricas con embalse
	9.385 MW de capacidad
Transmisión	206 elementos

Algunos generadores del modelo resultan de colapsar las unidades de generación individuales que conforman una central, representando toda la capacidad de generación de esa barra. Por otro lado, partiendo de los datos de operación y condiciones de hidrología para todos los días del año 2014, se pudo realizar una agrupación de estos datos por medio de un algoritmo

[1]Información más detallada en el Anexo B

de clusterización, encontrando que 16 días típicos representan el comportamiento global del sistema para todo el año 2014. A partir de los datos obtenidos, para el caso de estudio se seleccionó uno de estos días típicos más representativos en términos de carga y disponibilidad hídrica.

El horizonte temporal diario considerado incluyó 24 períodos de tiempo de una hora. El perfil horario de demanda, ilustrado en la Fig. **3-1**, se construyó a partir de la información de carga horaria por barra de 2014, reportada por XM, el operador del mercado de energía eléctrica en Colombia.

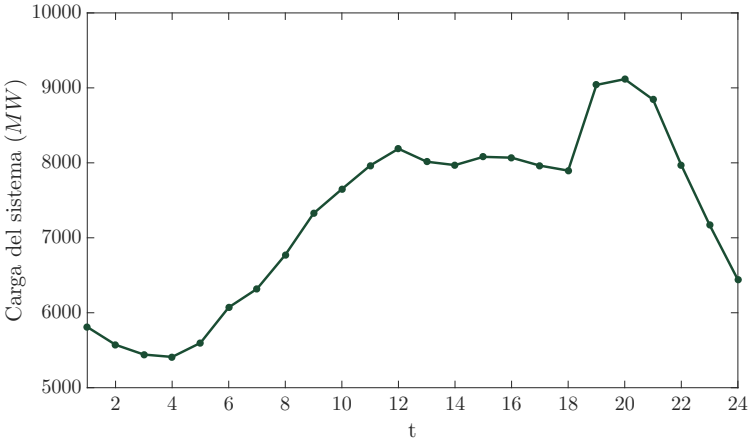

Figura 3-1.: Perfil de demanda diario del sistema de 96 barras.

Para introducir incertidumbre en generación renovable se ubica una granja eólica en el nodo Copey, con capacidad instalada de 800MW. La incertidumbre en la generación de potencia eólica es modelada a través de 14 escenarios de viento, cada uno representando una realización de energía renovable. Estos escenarios fueron construidos a partir de la velocidad de viento medido a 10m de altura por la estación meteorológica del Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) en Puerto Bolívar. Se tomaron perfiles históricos típicos para los tipos de día representados por el perfil de carga y la disponibilidad hídrica. Se precisó la transformación de la información tomada de una base de datos histórica de los años 2006-2014 con el fin de estimar la velocidad de viento a 50m, que es una altura más usual para una turbina eólica. Posteriormente, esta velocidad de viento transformada se aplicó a la curva de producción de una turbina Nordex N60/1300. Por último, la producción de potencia eólica obtenida se normalizó con respecto a la capacidad de la turbina, configurando un factor que se utilizó para modular la producción de la granja de viento. Cada escenario diario contiene 24 puntos, uno por período de tiempo, representando el nivel de producción normalizada de potencia eólica, como se ilustra en la Fig. **3-2**.

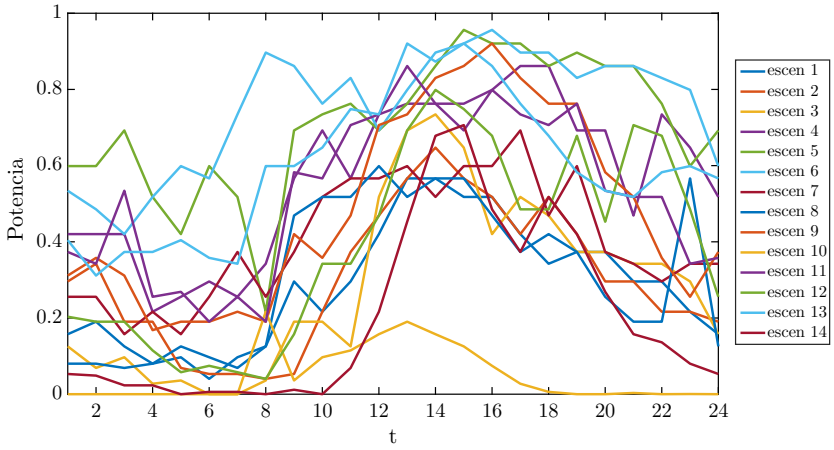

Figura 3-2.: Perfiles temporales de generación eólica con potencia normalizada.

Para esta simulación se incluyen nueve contingencias simples, todas con la misma probabilidad de ocurrencia, de las cuales siete corresponden a desconexiones de un elemento del sistema de transmisión (línea o transformador) y dos a la disminución de la capacidad de generación en las barras TEBSA y Guavio. La mayoría de las contingencias críticas fueron tomadas de [47]. Esto da lugar al modelado de $10 \times 14 \times 24 = 3360$ flujos de carga individuales en el problema.

Adicionalmente, para el análisis detallado de los resultados el sistema de potencia se ha subdividido en cinco áreas operativas denominadas: Costa Atlántica, Antioquia, Sur, Centro y Oriente, tal como se muestra en la Fig. **3-3**. Esta subdivisión corresponde en gran medida con la que se encuentra en [47] con criterios adicionales de agrupación basados en la cercanía geográfica y eléctrica de las barras en la red.

3.2.1. Suposiciones de modelado seguidas en este estudio

A continuación se enumeran suposiciones adicionales de modelado que se siguieron en el estudio:

1. Por simplicidad, se asume que la comisión de unidades fue establecida de antemano. En consecuencia, las variables binarias, las restricciones y los costos relacionados con la comisión de unidades no son tenidos en cuenta.

2. Los precios de las ofertas de generación se obtienen de datos históricos del 2014; para cada día, se tiene el precio promedio de oferta por tipo de combustible.

3. Para los precios de oferta de las reservas rodantes y las rampas de seguimiento de carga, se utilizaron precios positivos pero despreciables en comparación con las ofertas

Figura 3-3.: Áreas del sistema eléctrico colombiano.

de energía, con el ánimo de que las cantidades contratadas de estos otros productos en la subasta fuesen "ajustadas" (contratar la mínima cantidad realmente necesaria) pero al mismo tiempo con precios resultantes equivalentes al costo de oportunidad de la energía.

4. Aunque el *software* permite modelar el almacenamiento para una operación cíclica, de tal manera que la última hora del horizonte temporal anteceda a la primera hora del periodo de planeación, se descarta tomar esta opción.

3.3. Procedimiento de solución

La implementación bajo el modelo DC usada fue modificada para poder modelar cadenas hidráulicas. Salvo por esa modificación y el código de preparación de datos de entrada, el *software* es esencialmente como MOST [48]. Este fue el punto de partida para la implementación en AC, introduciendo modificaciones para ser resuelto mediante el paquete OPF generalizado de MATPOWER 7.0.1. El gran número de variables y costos de usuario adicionales, y las restricciones necesarias para implementar el modelo, se especificaron utilizando el mecanismo existente para ello en el flujo óptimo generalizado de `MATPOWER`. Estos códigos han sido utilizados en la solución de ambas versiones del modelo. Para resolver el modelo con red DC se empleó el paquete comercial Gurobi 9.0.0 [49] y para el modelo AC se usó

Tabla 3-2.: Resumen de características de los problemas resueltos

Item	**Modelo AC**	**Modelo DC**
Número de variables	1.438.152	950.952
Número de restricciones	2.978.992	1.885.632
Costo de operación	$ 22.826.371	$ 22.621.304

IPOPT 3.12.1 [50].

3.3.1. Resultados numéricos

En estas sección, si bien las cantidades asignadas de cada producto son individuales a cada planta, en varias de las comparaciones se agrupan las plantas generadoras en zonas operativas para mayor facilidad, sin afectar las conclusiones generales. La Tabla **3-2** contiene un resumen de la información del problema resuelto, discriminado por cada modelo, correspondiente al tamaño del problema, representado por el número de variables y restricciones, y el costo de operación esperado.

La complejidad del problema estocástico no lineal se evidencia por el número de variables y restricciones adicionales, aproximadamente 50 % mayor que su contraparte lineal. En cuanto al valor de la función objetivo, la diferencia entre ambos modelos fue inferior al 1 %.

La Figura **3-4** presenta el despacho total, por tipo de combustible: agua (hidro), gas natural para plantas de generación de ciclo simple (gn) y de ciclo combinado (gncc), carbón y viento. Se evidencian diferencias importantes en los despachos horarios entre el modelo DC y el modelo AC. Al examinar la energía total diaria por tipo de combustible, tanto a nivel sistémico como de manera diferenciada por zona operativa en la Fig. **3-5**, es claro que a nivel sistémico no hay mayor diferencia en la asignación energética por tipo de combustible. Por otro lado, a nivel zonal, se encuentra que las diferencias en las zonas Sur, Centro y Oriente son en realidad mínimas al punto de la insignificancia, mientras que entre las zonas Costa Atlántica y Antioquia, sucede que un bloque importante de generación de ciclo combinado presente en Antioquia en el modelo DC es transferido a la zona Costa Atlántica en AC, particularmente a la planta TEBSA, manteniéndola con un factor de utilización alto. Este fenómeno sucede en la vida real en la operación típica del sistema colombiano [51]. Salvo por esta diferencia, dictada por la red, las asignaciones de energía a los oferentes se guían esencialmente por los precios de oferta.

En la Fig. **3-6** se presentan las asignaciones de rampa de seguimiento de carga por tipo de combustible, discriminadas por zona operativa. Cabe resaltar que la franja verde, corres-

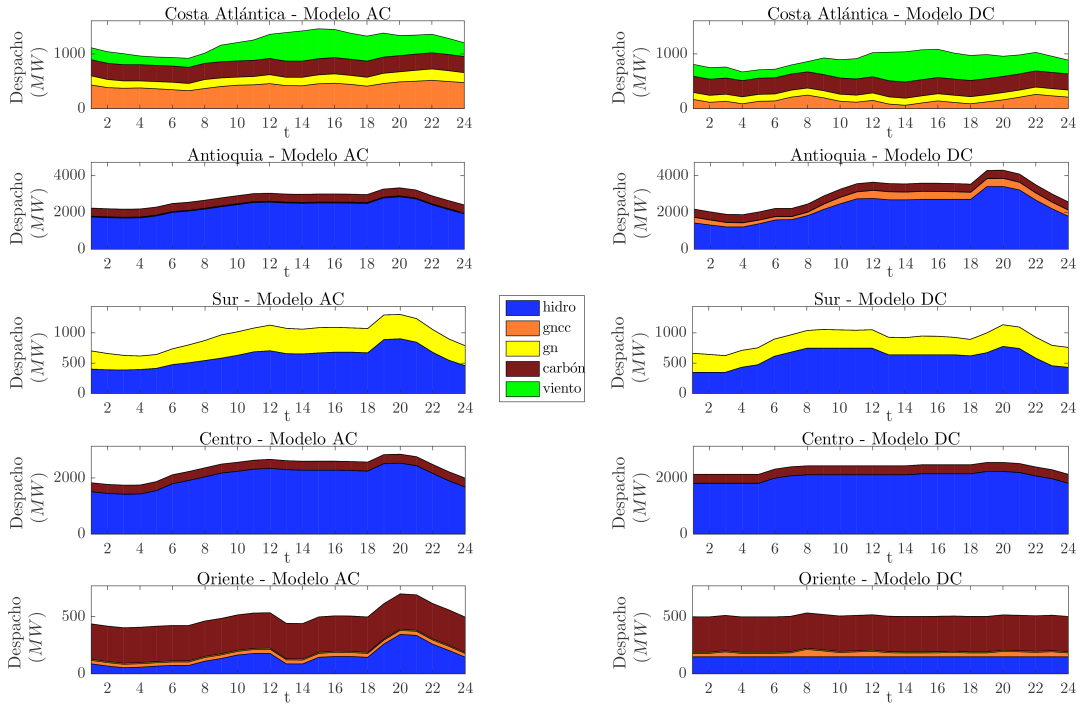

Figura 3-4.: Despacho por tipo de combustible para cada área operativa a lo largo del horizonte de planeación.

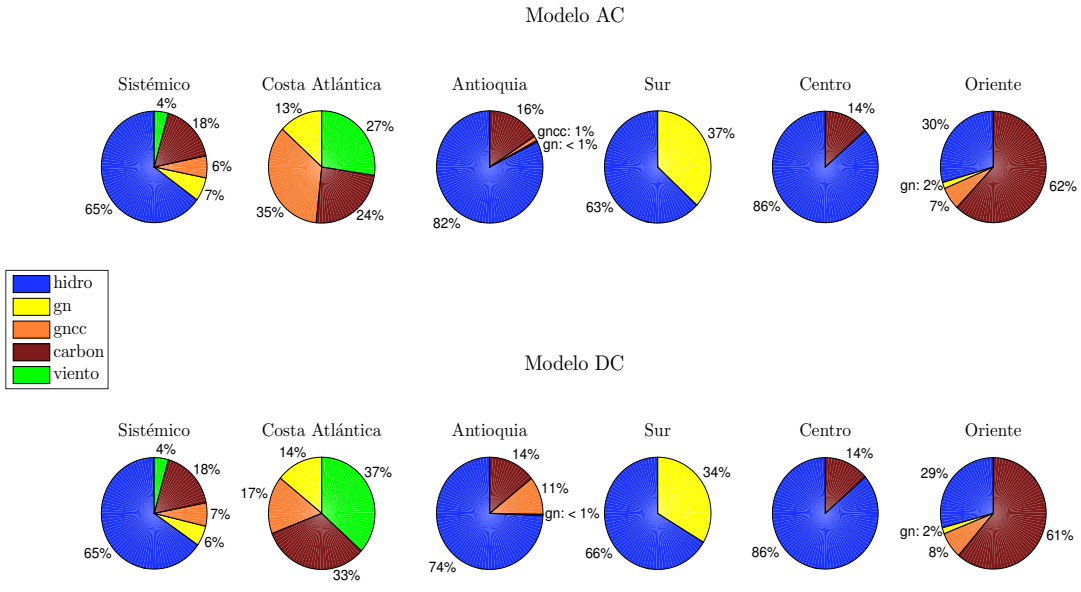

Figura 3-5.: Energía total de área y sus proporciones por combustible.

pondiente a la "reserva de rampa" asignada al recurso eólico, debe ser interpretada como la flexibilidad que se requiere del resto del sistema para poder aceptar dicha generación.

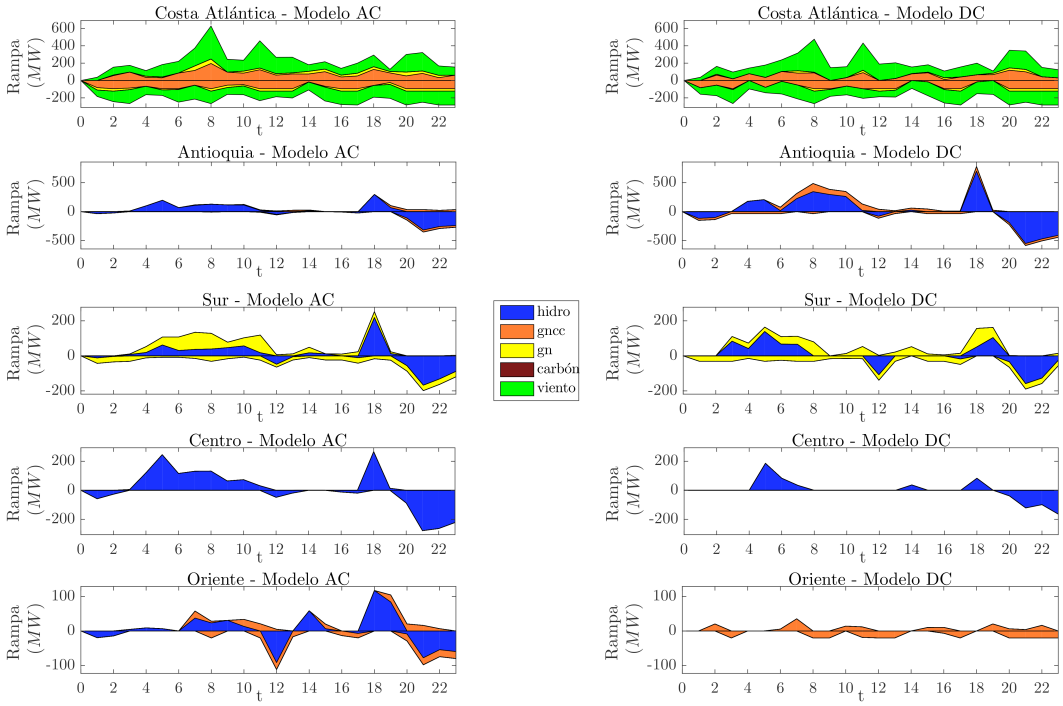

Figura 3-6.: Reserva de rampa de seguimiento de carga por tipo de combustible para cada área operativa.

La mayor parte de los requerimientos de rampa son asignados a las plantas hidroeléctricas, observándose que para el modelo DC las reservas de rampa recaen principalmente en la zona de Antioquia mientras que en el modelo AC están más uniformemente distribuídas en todas las zonas operativas con generación hidráulica. De esta forma, el modelo AC requiere de una distribución de este recurso a lo largo de todo el sistema, en contraste con el modelo DC, en el que la asignación se focaliza mayoritariamente en una sola zona del sistema.

La Figura **3-7** presenta las asignaciones de reserva rodante de contingencia, inicialmente de manera sistémica y después discriminadas por zona operativa. En el total sistémico se evidencia que el modelo AC resulta en una asignación no trivial de reserva de contingencia a las turbinas de gas de ciclo sencillo, a expensas de las reservas asignadas previamente a las plantas de ciclo combinado en el modelo DC. Al ver los resultados discriminados por zonas, se puede apreciar que en DC las contingencias parecen requerir siempre de redespachos negativos en las hidroeléctricas de la zona Centro y redespachos positivos en las hidroeléctricas de Antioquia y del Sur, además de redespachos positivos en plantas de gas en las zonas Sur, Antioquia y Costa Atlántica. Por otro lado, al ver la situación en el modelo AC, la zona Centro no siempre es redespachada de manera negativa, la zona Oriente asume cantidades importantes de reserva rodante y en la Costa Atlántica también lo hacen las turbinas de ciclo sencillo. Las diferencias son, sin embargo, más abundantes que sólo lo expuesto, llevando a

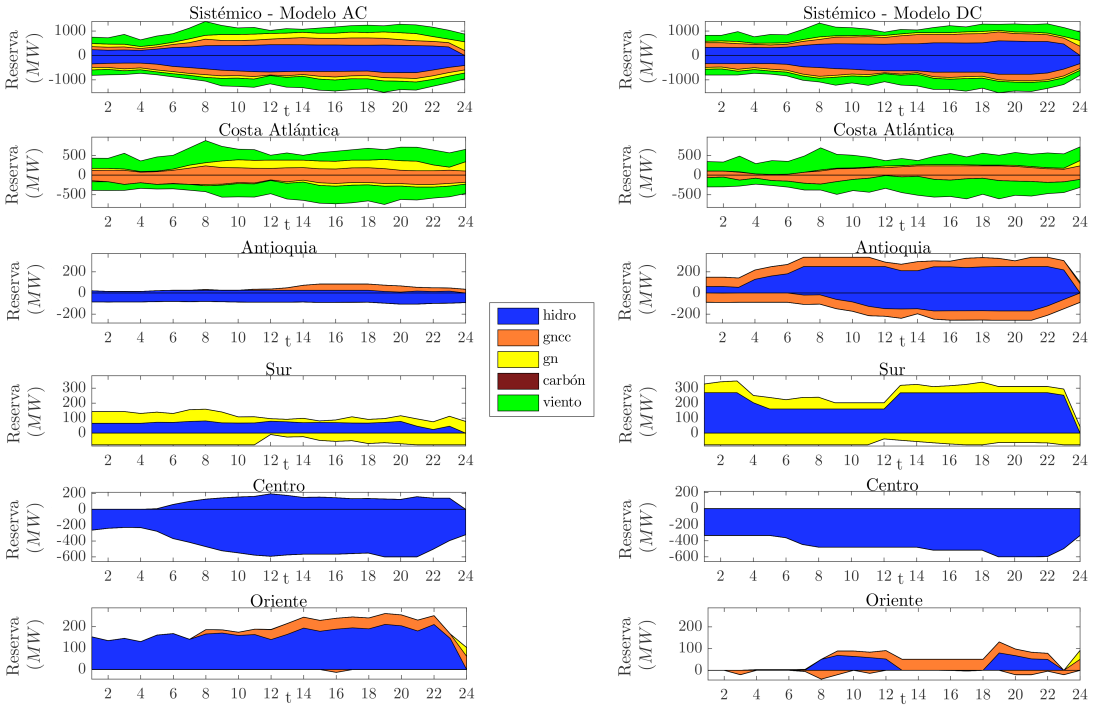

Figura 3-7.: Reserva rodante de contingencia por tipo de combustible para cada área operativa.

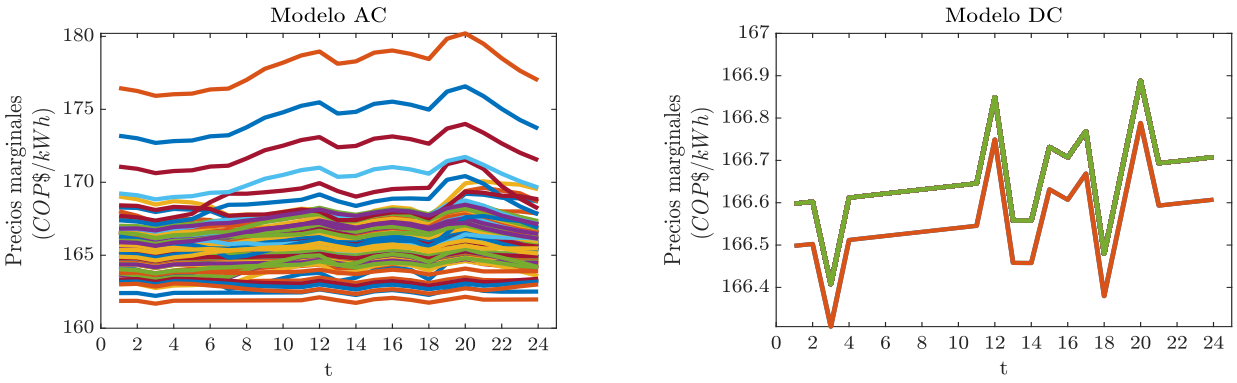

Figura 3-8.: Precios nodales en las barras del sistema a lo largo del horizonte.

la conclusión de que las consideraciones de red AC conducen a una asignación muy diferente de las cantidades requeridas del recurso de reserva rodante de contingencia entre el modelo DC y el AC.

La Figura **3-8** muestra los precios nodales esperados a lo largo del horizonte para las barras del sistema usando los dos modelos de red. La variedad de precios existentes en el modelo AC es notoria, comparada con la casi nula variabilidad en el caso DC. Evidentemente, las pérdidas y los acoplamientos activo/reactivo en el caso AC generan unas diferencias importantes entre

estos precios. La diferencia porcentual entre el precio más alto y el más bajo puede llegar a ser tanto como +11 %. a pesar de no existir ramas congestionadas en el modelo.

3.4. Conclusiones

La conclusión principal del estudio es que en un mercado de día previo multidimensional en el que los servicios distintos a la energía se ofrecen esencialmente al costo de oportunidad, el cierre de mercado en las cantidades del producto de energía es casi igual usando el modelo DC o AC, pero la asignación de otros productos tales como reserva rodante o reserva de rampa de seguimiento de carga puede variar sustancialmente en cuanto a cuáles son los oferentes escogidos. El modelo AC reprodujo algunos aspectos que son conocidos en la operación real del sistema colombiano, como por ejemplo el alto factor de utilización de la planta TEBSA en 2014. De manera general, el modelo de optimización con modelo AC suele distribuír entre más participantes y mayores extensiones geográficas las asignaciones de los productos secundarios, enfatizando la correcta distribución geográfica de la asignación. La diversidad de precios nodales de energía en el modelo AC comparado con el DC implica una mejor resolución en la comparación entre competidores, pues evidentemente se toman en cuenta la contribución a pérdidas y los acoplamientos entre potencia activa y reactiva. Cabe abundar que estas conclusiones se mantienen al considerar otros días típicos operativos del sistema colombiano, tales como días típicos en época de sequía y por lo tanto con menor disponiblidad hídrica.

El impacto inmediato para el diseño de mercados multidimensionales, es que si se desea hacer una asignación a nivel geográfico general para los productos secundarios en el mercado de día anterior, es deseable cerrar dichos mercados usando un modelo AC en aras de la justicia de la asignación. Si se insiste en utilizar un modelo DC, es necesario emplear mecanismos adicionales que garanticen la imparcialidad en la asignación de los productos secundarios.

Para estudios futuros, se puede investigar el efecto de la existencia de congestión en el sistema de transmisión y el del comportamiento estratégico en las ofertas de los productos secundarios. Sin embargo, el efecto de la congestión, siguiendo la premisa de que los resultados deben contrastarse con algo que exista en la realidad, debe estudiarse sobre un sistema existente en el cual la congestión tenga efectos conocidos, lo cual no favorece el uso del sistema colombiano para ello.

4. Técnicas de Descomposición para Problemas de Programación no Lineal

Un problema de optimización estocástico tiene parámetros inciertos modelados como variables aleatorias que harán parte de la formulación del problema, bien sea en la función objetivo o en las restricciones. Resolver uno de estos problemas considerando explicitamente todas las realizaciones del parámetro incierto puede resultar poco práctico, debido a la alta carga computacional que esto implica. En su lugar, en el enfoque de escenarios se usa un conjunto finito de realizaciones discretas, o escenarios, con el objetivo de conseguir una solución que funcione bien para cualquier realización del parámetro incierto. Cada escenario contiene una descripción completa de los valores inciertos en una realización y tiene asociada una ponderación positiva, acorde con su probabilidad de ocurrencia.

Algunos problemas estocásticos de la vida real son demasiado complejos para resolverlos de forma directa. Esta complejidad surge del crecimiento exponencial del tamaño del problema con el número de escenarios incorporados al modelo. Un problema de grandes dimensiones está descrito por un gran número de variables y restricciones. Para su solución se requiere de mucho cálculo, almacenamiento de datos y excesivo tiempo computacional [52]. En la práctica, los problemas de programación de grandes dimensiones no se resuelven en su formulación completa y en su lugar, se recurre a técnicas de descomposición para facilitar su solución.

Las técnicas de descomposición son viables cuando el problema tiene una estructura divisible. La solución por descomposición se fundamenta en la división del problema original en dos o más subproblemas relacionados jerárquicamente, que se deben resolver de forma coordinada. En comparación con el problema original, los subproblemas son más fáciles de resolver, de comprender y de simplificar. Al tratar el problema inicial en un espacio de variables y restricciones reducido se busca conseguir eficiencia computacional. Además, permite adaptar algoritmos de solución a cada tipo de subproblema y ejecutar los cálculos en paralelo.

Los métodos de descomposición se pueden agrupar en dos categorías básicas: la descomposición primal y la descomposición dual. Los métodos de descomposición primal tratan problemas con variables de *complicación*, es decir, variables que al ser fijadas temporalmente facilitan la división del problema en subproblemas de menor tamaño y más sencillos de resolver. En cambio los métodos de descomposición dual tratan problemas con restricciones de

complicación que al ser relajadas, enviándolas a la función objetivo o al lagrangiano penalizadas por parámetros, facilitan la descomposición del problema. Los dos enfoques pueden considerarse como complementarios. Este Capítulo ofrece una revisión en detalle de los mismos.

En lo que se refiere al problema de optimización bajo estudio, se ha asumido que la programación de las unidades de generación en el comisionamiento de unidades se determinó de antemano. Por lo tanto, al ser el problema de tipo no lineal con variables continuas, los métodos de descomposición expuestos en breve no harán referencia a variables binarias.

Algunos componentes de los métodos de descomposición se formulan con base en las propiedades de la teoría dual en programación no lineal, las cuales serán expuestas a continuación.

4.1. Nociones de teoría dual en programación no lineal

Un problema de optimización general, en el que las no linealidades puedan estar tanto en la función objetivo como en las restricciones, es introducido por (4-1).

$$\begin{aligned} &\min_{x} \ f(x) \\ &\text{s.t. } g_i(x) = 0 \qquad i = 1, ..., n \\ &\qquad h_j(x) \leq 0 \qquad j = 1, ..., m \\ &\qquad x \in X \end{aligned} \tag{4-1}$$

donde, X es un conjunto convexo compacto, $f(x)$, $g_i(x)$ y $h_j(x)$ son funciones convexas.

En programación no lineal, la relación entre los problemas primal y dual se fundamenta en las condiciones de optimalidad. Los conceptos enunciados a continuación están basados en [53].

La función lagrangiana de (4-1) se introduce aquí para su posterior uso en la definición de las condiciones de optimalidad.

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{n} \lambda_i g_i(x) + \sum_{j=1}^{m} \mu_j h_j(x) \tag{4-2}$$

Condiciones necesarias de primer orden (Karush-Kuhn-Tucker)

Sea (x^*) un óptimo local de (4-1). Existen escalares μ_j^* y λ_i^*, llamados multiplicadores de Karush-Kuhn-Tucker (KKT), tales que:

$\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0$
$\nabla_{\lambda,\mu} \mathcal{L}(x^*, \lambda^*, \mu^*) = 0$
$\mu_j^* \geq 0$
$\mu_j^* = 0 \qquad \forall j \notin A(x)$

Se denota por $A(x) = \{j | h_j(x) \geq 0\}$ al conjunto de restricciones activas. En (x^*) las restricciones activas se pueden tratar como igualdades. Las restricciones son continuamente diferenciables en (x^*) y sus jacobianos son linealmente independientes.

Explícitamente, para este problema las condiciones necesarias de primer orden KKT se escriben como:

$$\nabla f(x^*) + \sum_{i=1}^{n} \lambda_i^* \nabla g_i(x^*) + \sum_{j \in A(x^*)} \mu_j^* \nabla h_j(x^*) = 0 \tag{4-3}$$

$$g_i(x^*) = 0 \qquad i = 1, ..., n \tag{4-4}$$

$$\mu_j^* h_j(x^*) = 0 \qquad \forall j \in \{1, ..., m\} \tag{4-5}$$

$$\mu_j^* \geq 0 \qquad \forall j \in \{1, ..., m\} \tag{4-6}$$

Un problema de programación no lineal como (4-1) tiene otro problema no lineal estrechamente asociado y se conoce como problema dual lagrangiano. Bajo ciertas suposiciones de convexidad, los problemas primal y dual tiene valores óptimos iguales de la función objetivo, lo que sugiere que es posible resolver el primal a través de su dual.

El problema dual lagrangiano se establece como:

$$\begin{aligned} &\text{máx}\, \theta(\lambda, \mu) \\ &\quad s.t.\ \mu \geq 0 \end{aligned} \tag{4-7}$$

donde, $\theta(\lambda, \mu) = \text{ínf}_{x \in X}\, \mathcal{L}(x, \lambda, \mu)$.

Es necesario introducir dos teoremas importantes en teoría de dualidad para problemas no lineales.

Teorema 1 (Teorema de dualidad débil) *Sea (x) solución factible de (4-1), tal que $x \in X$, $g(x) = 0$ y $h(x) \leq 0$; además (λ, μ) es solución factible de (4-7) tal que $\mu \geq 0$. Entonces, $f(x) \geq \theta(\lambda, \mu)$.*

Prueba.

Por la definición de θ y dado que $x \in X$, se tiene que

$$\theta(\lambda, \mu) = \inf_{x \in X} (f(x) + \lambda^T g(x) + \mu^T h(x)) \leq f(x) + \lambda^T g(x) + \mu^T h(x) \leq f(x)$$

ya que $\mu \geq 0$, $h(x) \leq 0$ y $g(x) = 0$. Esto completa la prueba.

El teorema de dualidad débil señala que, en un punto factible, la función objetivo del problema primal será mayor que la función objetivo del problema dual. En otras palabras, el valor objetivo de cualquier solución factible al problema dual produce una cota inferior en el valor objetivo de cualquier solución factible del problema primal. Otras observaciones importantes se desprenden de este teorema.

Corolario 1 $\inf_{x \in X} \mathcal{L}(x, \lambda, \mu) \geq \sup\ \theta(\lambda, \mu)$

Corolario 2 *Si* $f(x^*) = \theta(\lambda^*, \mu^*)$*, entonces* (x^*) *y* (λ^*, μ^*) *resuelven los problemas primal y dual respectivamente.*

Corolario 3 *Si* $\inf_{x \in X} \mathcal{L}(x, \lambda, \mu) = -\infty$*, entonces* $\theta(\lambda, \mu) = -\infty$

Corolario 4 *Si* $\sup \theta(\lambda, \mu) = \infty$*, entonces el problema primal no tiene solución factible.*

Si se mantiene estrictamente la desigualdad en el teorema de dualidad débil, existe una brecha de dualidad. La ausencia de la brecha de dualidad se puede garantizar bajo ciertas suposiciones de convexidad y otras condiciones dadas por el teorema de dualidad fuerte.

Teorema 2 (Teorema de dualidad fuerte) *Existe un punto* $(\hat{x})$*, tal que* $h(\hat{x}) \leq 0$ *y* $g(\hat{x}) = 0$*. Entonces,* $\inf_{x \in X} \mathcal{L}(x) = \sup \theta(\lambda, \mu)$

Además, si el ínfimo tiene valor finito, entonces el supremo es alcanzado en (λ^*, μ^*)*, con* $\mu^* \geq 0$*. Si el ínfimo es alcanzado en* (x^*)*, entonces* $\mu^* h(x^*) = 0$

La prueba de este teorema se encuentra detallada en [53].

4.2. Descomposición por Relajación Lagrangiana

La idea básica en métodos Lagrangianos es relajar las restricciones de acople, ubicándolas en la función objetivo, para obtener problemas sencillos que se resuelvan con más eficiencia. Estos métodos resultan atractivos al exhibir rangos de convergencia lineal o superlineal [54]. Los métodos tipo Relajación Lagrangiana (RL) resuelven el dual del problema original. Si el problema inicial es convexo, la solución del primal se obtiene a partir de la solución del dual. En el caso no convexo, la solución del dual genera una cota inferior (para problemas de minimización) a la solución del problema primal y entonces, se hace necesario adoptar algún

procedimiento para encontrar una solución primal factible a partir de la solución del dual. La aproximación dual basada en RL presenta unas desventajas asociadas, en algunos casos, con la no unicidad de la solución de los subproblemas relajados [55]. Para problemas muy grandes con muchas restricciones acopladoras, estas dificultades hacen que la aproximación dual no sea práctica.

Una alternativa consiste en introducir una función de penalización en la función objetivo, para regularizar el método. La versión regularizada es conocida como Relajación Lagrangiana con lagrangiano Aumentado (RLA). La función de penalización más común es la cuadrática y originalmente fue adoptada para el método de multiplicadores por Hestenes [56] e independientemente por Powell [57], en una forma diferente pero equivalente. Aunque este enfoque tiene ventajas como la simplicidad y estabilidad del método de multiplicadores, la posibilidad de partir de un multiplicador arbitrario o el hecho de no tener que resolver un problema maestro; el problema sigue siendo difícil de resolver ya que el lagrangiano aumentado no es separable. Esto ha motivado el desarrollo de diversas técnicas de descomposición que van desde usar aproximaciones lineales de la función lagrangiana, aproximaciones cuadráticas diagonales, o una serie de problemas auxiliares involucrando el lagrangiano aumentado.

Para describir el método general de descomposición por RLA, se considera el problema de programación no lineal convexo y estructurado por bloques (4-8), en el que $x = (x_1, x_2, ..., x_L)$ es una partición de variables de decisión y la función objetivo, las restricciones y el conjunto X también se puede partir como x:

$$\begin{aligned} \min_{x} \ & \sum_{i=1}^{L} f_i(x_i) \\ \text{s.t.} \ & \sum_{i=1}^{L} A_i x_i = b \\ & x_i \in X_i, \qquad i = 1, 2, ..., L \end{aligned} \tag{4-8}$$

donde, cada función f_i es convexa, $A = (A_1, ..., A_L) \in \mathbb{R}^{m \times l}$ es una partición apropiada de la matriz A con A_i como matriz $m \times l_i$, y $b \in \mathbb{R}^m$ es un vector.

El lagrangiano aumentado asociado con (4-8) tiene la forma:

$$\begin{aligned} \mathcal{L}(x, \lambda) &= F(x) + \lambda^T (b - Ax) + \frac{c}{2} \|b - Ax\|^2 \\ &= \lambda^T b + F(x) - \lambda^T Ax + \frac{c}{2} \|b - Ax\|^2 \end{aligned} \tag{4-9}$$

con

$$F(x) = \sum_{i=1}^{L} f_i(x_i)$$

$$Ax = \sum_{i=1}^{L} A_i x_i$$

donde, el vector de multiplicadores $\lambda \in \mathbb{R}^m$ es asociado a las restricciones $Ax = b$, y c es el coeficiente de penalización. Existe un $c > 0$ lo suficientemente grande tal que la curvatura del Hesiano del lagrangiano proyectada en el espacio factible es positiva.

El problema dual correspondiente es:

$$\max_{\lambda \in R^m} q(\lambda) \tag{4-10}$$

donde q es el funcional dual:

$$q(\lambda) = \inf_{x \in X} \mathcal{L}(x, \lambda) \tag{4-11}$$

con $X = X_1 \times X_2 \times ... \times X_L$.

Las relaciones entre el problema primal (4-8) y el dual (4-10) se fundamentan en la teoría dual [53]. Para cada solución óptima $\hat{x}$ del primal y cada solución óptima $\hat{\lambda}$ del dual, se cumple que $F(\hat{x}) = q(\hat{\lambda})$. Además, para cada solución $\hat{\lambda}$ de (4-10) un punto $\hat{x} \in X$ es solución de (4-8), si y solo sí:

$$\mathcal{L}(\hat{x}, \hat{\lambda}) = \min_{x \in X} \mathcal{L}(x, \hat{\lambda}) \tag{4-12}$$

Una ventaja importante de (4-12) sobre el dual lagrangiano habitual es su suficiencia para obtener la solución primal cuando se conoce la solución dual. El problema dual puede resolverse de forma iterativa como se indica a continuación. En primer lugar se calcula una solución x^w con un λ^w fijo:

$$x^w = \arg\min_{x \in X} \mathcal{L}(x, \lambda^w) \tag{4-13}$$

Luego, la secuencia de multiplicadores λ^w es generada de acuerdo con la expresión siguiente, que es equivalente a una regla de máximo ascenso dual:

$$\lambda^{w+1} = \lambda^w + c(b - Ax^w), \qquad w = 1, 2, ..., L \tag{4-14}$$

La secuencia $\{\lambda^w\}$ generada por el método de multiplicadores converge a la solución $\hat{\lambda}$ de (4-10). Si la secuencia de λ^w generada se vuelve acotada, entonces el método garantiza un

límite para el valor óptimo del problema de optimización [54].

No obstante, una desventaja importante del método de multiplicadores es que el lagrangiano aumentado no es separable con respecto a las variables primales, debido al término cuadrático de penalización en (4-11). Aunque estos no poseen las propiedades de descomposición de los lagrangianos ordinarios, algunas transformaciones del problema original a otra forma equivalente que se preste mejor al enfoque de descomposición pueden ser empleadas para facilitar la división del problema inicial en L problemas independientes en x_i.

1. Aplicar de forma iterativa un método tipo Jacobiano no lineal [55] para minimizar (4-9), introduciendo funciones $\mathcal{L}_i(x, \widetilde{x}, \lambda)$ con parámetros adicionales $\widetilde{x}$ que se actualizan iterativamente.

   $$\mathcal{L}_i(x_i, \widetilde{x}, \lambda) = f_i(x_i) + \lambda^T b - \lambda^T A_i x_i + \frac{c}{2}\|b - A_i x_i - \sum_{j \neq i} A_j \widetilde{x}_j\|^2 \tag{4-15}$$

   A partir de funciones como (4-15) se construye un problema equivalente de (4-11) que es divisible y facilita el cálculo de x^w de forma descompuesta. Cada lagrangiano aumentado (4-15) se minimiza con respecto a las decisiones asociadas con el subconjunto de variables x_i, mientras las demás decisiones, $\widetilde{x}_j, j \neq i$, permanecen fijas.

   El sub-algoritmo implementado al principio del método de multiplicadores además de calcular x^w actualiza el punto de referencia $\widetilde{x}^w$ con algún tamaño de paso $\tau \in (0, 1)$. Este algoritmo converge para τ suficientemente pequeño, aunque puede ser lento.

   $$\widetilde{x}^{w+1} = (1 - \tau)\widetilde{x}^w + \tau x^w \tag{4-16}$$

   Un ejemplo de aplicación de esta alternativa es el método de aproximación cuadrático diagonal de Ruszczyński, expuesto en su formulación general para problemas de optimización convexa en [55].

2. El Principio del Problema Auxiliar del Lagrangiano Aumentado propuesto por Cohen y Zhu (1984) [58], encuentra la solución de un problema de optimización restringido mediante una secuencia de problemas auxiliares que involucran al lagrangiano aumentado. En (4-9) se linealiza el término cuadrático y se adiciona un término de regularización de una función auxiliar $K(x)$, dado por:

   $$\frac{1}{\epsilon}[K(x) - K(x^w) - \langle \nabla K(x^w), x \rangle] \tag{4-17}$$

   Tal que la función lagrangiana aumentada modificada queda formulada como:

   $$\mathcal{L}(x, \lambda) = \lambda^T b + F(x) + Ax^T(\lambda^w + c(b - Ax^w)) + \frac{1}{\epsilon}[K(x) - K(x^w) - x^T \nabla K(x^w)] \tag{4-18}$$

Si la función $K(x)$ es separable, por ejemplo, $K(x) = \sum_{i=1}^{L} K_i(x_i)$, la función lagrangiana modificada también lo será.

Una cuestión de importancia práctica tiene que ver con la selección del valor inicial del multiplicador y la secuencia de parámetros de penalización. Bertsekas [59] hace una serie de sugerencias que ayudan en ambos casos. Para el primero, se sugiere explotar algún tipo de conocimiento previo para seleccionar λ^0 tan cercano como sea posible al valor óptimo λ^*. Con respecto al parámetro de penalización c, su valor inicial c^0 no debe ser demasiado grande para no causar un mal condicionamiento del problema inicial sin restricciones. Una buena práctica consiste en seleccionar un c^0 de valor moderado, eventualmente producto de experimentación previa. Tampoco se debe incrementar demasiado lento o demasiado rápido durante las primeras iteraciones. Un incremento demasiado lento no ayuda a mejorar una tasa de convergencia pobre, mientras que un incremento demasiado rápido puede ocasionar mal condicionamiento del problema. Se recomienda incrementar c^w a través de la fórmula $c^{w+1} = \beta c^w$ con $\beta \geq 1$ escalar. Otra posibilidad es usar un c^w diferente para cada restricción e incrementar por un cierto factor el parámetro de penalización de la restricción con la mayor violación. Otro factor clave esta relacionado con la actualización de la variable dual en cada iteración, como se trata en la siguiente sección.

4.2.1. Métodos de actualización de multiplicadores

Varios métodos han sido empleados para la actualización de los multiplicadores en la solución del problema dual. Estos buscan acelerar el proceso de convergencia que usualmente es lento. Los métodos más comunes son aproximaciones basadas en subgradiente, planos cortantes y métodos de haz. Estos se exponen a continuación con información basada en [60].

1. **Subgradiente**

   Esta aproximación consiste en actualizar la variable dual de forma proporcional al subgradiente de la función dual.

   $$\lambda^{w+1} = \lambda^w + \kappa^w \frac{\nabla_\lambda q(\lambda)|_{x^w}}{\|\nabla_\lambda q(\lambda)|_{x^w}\|} \tag{4-19}$$

   donde, w es el índice de iteración, κ es el tamaño de paso en cada iteración y $\nabla_\lambda q(\lambda)|_{x^w}$ representa al subgradiente de la función dual evaluado en x^w, por ejemplo, $\nabla_\lambda q(\lambda)|_{x^w} = b - Ax^w$ en la función dual (4-11). Un tamaño de paso muy pequeño produce una lenta convergencia. En cambio, un tamaño de paso muy grande puede hacer que el método oscile y por lo tanto no converja. En algunas ocasiones, no se emplea la normalización

del subgradiente en (4-19). El valor inicial y subsecuente del tamaño de paso determina cuán rápido converge el algoritmo. El tamaño de paso elegido puede depender, por ejemplo, de los datos del problema o de una estimación del óptimo de la función dual en cada iteración.

El método de actualización basado en subgradiente es el más utilizado dada su simplicidad y baja sobrecarga computacional. No obstante, es computacionalmente ineficiente y oscilante haciedo difícil establecer un criterio de parada. Generalmente el criterio de parada se basa en la ejecución de un cierto número de iteraciones establecido de antemano o en que la diferencia entre el valor de la función objetivo entre dos iteraciones sucesivas esté dentro de una tolerancia definida.

2. **Planos cortantes**

Este método se fundamenta en la reconstrucción de la función dual mediante hiperplanos. En cada iteración w se obtiene un hiperplano tangente a la función dual en un punto dado por $(\lambda^w, q(\lambda^w))$, con $q(\lambda^w)$ como el valor de la función dual evaluada en λ^w. La aproximación lineal del problema dual, incluyendo una serie de hiperplanos adicionados en cada iteración, que se debe resolver es como:

$$\lambda^{w+1} = \arg \max_{z,\lambda} z \tag{4-20}$$

$$\text{s.t. } z \leq q(\lambda^i) + (\nabla_\lambda q(\lambda)|_{x^i})^T(\lambda - \lambda^i) \qquad i = 1, ..., w \tag{4-21}$$

donde, i es el índice de iteración, $\nabla_\lambda q(\lambda)|_{x^i}$ es el subgradiente de la función dual con respecto a λ evaluada en x^i y z representa la aproximación a la función dual. De esta forma, la complejidad del problema dual a resolver en cada iteración aumenta debido a que las restricciones del problema aumentan con cada hiperplano que es agregado.

A diferencia de las técnicas basadas en subgradiente se puede establecer un criterio de parada a partir del valor óptimo de la función dual y su aproximación en cada iteración, como:

$$\frac{z - q(\lambda^w)}{q(\lambda^w)} \leq \epsilon \tag{4-22}$$

El método de planos cortantes converge al óptimo pero lentamente y además es oscilante. En las primeras iteraciones la amplitud de las oscilaciones puede ser muy grande y son proporcionales al tamaño del dominio de definición de las variables duales, que debe ser grande para garantizar que el método encuentre una solución.

3. **Método de Haz**

Los métodos de haz buscan estabilizar los métodos de planos cortantes incluyendo un término cuadrático que penalice el alejamiento de la mejor solución obtenida, llamada centro de estabilidad. Este corresponde al vector de multiplicadores que ha permitido obtener la mejor solución hasta la iteración w, de acuerdo con un crecimiento nominal calculado en cada iteración. El centro de estabilidad se actualiza siempre que la diferencia entre el valor de la función dual en la iteración w y el valor de la función dual para el centro de estabilidad de la iteración $w-1$ es mayor o igual a un porcentaje del crecimiento nominal predicho. Cuando esto sucede se dice que se ha dado un paso serio o ascendente, ya que se ha obtenido un mayor valor de la función dual. De lo contrario, se dice que se ha dado un paso nulo.

La forma nominal del método de haz queda definida por la expresión:

$$\lambda^{w+1} = \arg\max_{z,\lambda} z - c\|\lambda - \hat{\lambda}\|^2 \tag{4-23}$$

$$\text{s.t. } z \leq q(\lambda^i) + (\nabla_\lambda q(\lambda)|_{x^i})^T(\lambda - \lambda^i) \qquad i = 1, ..., w \tag{4-24}$$

donde, $\hat{\lambda}$ es el centro de estabilidad y $c \geq 0$ es el parámetro de penalización.

El crecimiento nominal se calcula como:

$$\delta^w = z^{*w} - c\|\lambda^* - \hat{\lambda}\|^2 - q(\lambda^w) \tag{4-25}$$

El criterio para establecer si la mejora ha sido significativa es el siguiente:

$$q(\lambda^w) - q(\hat{\lambda}) \geq m\delta^{w-1} \tag{4-26}$$

La selección del parámetro de penalización c es un factor critico para el buen desempeño del método. Si su valor es muy grande, la aproximación del problema dual conseguida no es buena porque el problema resultante puede ser muy convexo. Por el contrario, si su valor es muy pequeño el método es inestable por parecerse al método de planos cortantes del que procede. Una alternativa puede ser emplear un parámetro de penalización que dependa del contador de iteraciones para que esta sea incrementada a lo largo de las iteraciones. Por ello, los métodos de haz pueden resultar complicados al requerir el ajuste cuidadoso de sus parámetros para que sea eficiente.

Las aproximaciones descritas requieren un número significativo de iteraciones para alcanzar la convergencia. El diseño de algoritmos apropiados para mejorar el desempeño de los métodos de actualización de variables duales, y por ende, las propiedades de convergencia de los

métodos basados en RL, han sido reportados en algunos trabajos. Aquí se mencionan solo las aplicaciones específicas en alguno de los estudios para programación de la generación de potencia.

El estudio de coordinación hidrotérmica de corto plazo de Jimenez y Conejo [61], presenta una mejora en la iteración dual basada en el método de planos cortantes, al introducir un control dinámico de la región de confianza de la variable dual. Otro estudio en la misma área, conducido por Bento *et al.* [62], propone un ajuste dinámico en el tamaño de paso para la actualización de los multiplicadores, de acuerdo con el valor de la función dual, evitando la necesidad de elegir un conjunto predetermiado de parámetros. Un algoritmo de RL para la programación de la generación en un mercado de día en adelanto presentado en [63], usa multiplicadores iniciales mejorados y un mecanismo de ajuste adaptable de los multiplicadores en los que el tamaño de paso es inversamente proporcional a la norma Euclidiana de los balances de potencia y de reserva rodante. El método RL aplicado a la comisión de unidades por Feng y Liao en [64], consigue mejorar el desempeño del método de subgradiente incluyendo un factor de amortiguación y un tamaño de paso diferente para cada restricción, basado en el concepto de factor de escala, derivado de información histórica de las restricciones de acople específicas.

El mecanismo de actualización del tamaño de paso es un componente clave para estos métodos. En algunos casos se trata de cubrir el enfoque clásico en el que se depende en gran medida de la experiencia del usuario para determinar esos valores y en su lugar, optan por tomar información generada por el problema mismo para ajustar el tamaño de paso.

4.2.2. Algoritmo para RLA

El proceso de descomposición mediante RLA requiere de dos procedimientos. El primero es la solución del problema primal relajado, o el problema dual. El segundo corresponde a la actualización de los multiplicadores con alguno de los métodos de la Sección 4.2.1.

Algoritmo 4.1 Descomposición por RLA

1: Inicializar $w = 0$, el vector de multiplicadores λ^w y el parámetro de penalización c^w.
2: Obtener x_i^w al resolver cada subproblema $\min_{x_i} \mathcal{L}_i(x_i, \lambda^w)$ con λ^w fijo.
3: Actualizar λ^w con alguna regla de ascenso (Sección 4.2.1),
4: **si** se cumple el criterio de convergencia **entonces**
5: Parar
6: **si no**
7: Hacer $w = w + 1$
8: Ir a 2
9: **fin si**

Como criterio de parada se puede optar por chequear en cada iteración la disminunción de la norma del subgradiente o asumir un criterio de terminación en el que se ejecuta un número específico de iteraciones

4.3. Descomposición de Benders

J.F. Benders propuso en 1962 un método de descomposición para resolver los problemas de programación de tipo mixto-entero (MIP) [65], difíciles de resolver debido a la naturaleza entera de algunas variables que hacen que se pierda la convexidad de la región factible. El método de Descomposición de Benders (DB) es aplicable a una amplia gama de problemas en los que se identifiquen variables de *complicación*, esto es, variables que al ser fijadas temporalmente como parámetros hacen que el problema de optimización sea más sencillo de resolver. Por ejemplo, las variables enteras, las variables que ocasionen no linealidades en restricciones o las variables de acople que impidan que un problema se pueda dividir en dos o más subproblemas, se consideran como variables de *complicación*.

A través de esa distinción de variables se identifican estructuras en el problema que permiten su separación en dos problemas independientes, llamados problema maestro y subproblema. Específicamente se hace referencia a la estructura de la matriz de restricciones, que adquiere una configuración diagonal por bloques, como se muestra en la Figura **4-1** *b)*, luego de reordenar las variables de acuerdo con su clasificación. La solución mediante DB implica un proceso iterativo en el que el problema maestro y el subproblema se resuelven de forma alternada e intercambian información a través de los *cortes de Benders* para alcanzar el punto óptimo del problema original.

El enfoque de DB fue generalizado por Geoffrion [66] a una clase más amplia de programas en los que el subproblema parametrizado ya no necesita ser un programa lineal. En este caso, se hace uso de la teoría de la dualidad convexa no lineal para derivar los cortes equivalentes a los de Benders clásico. El trabajo de Geoffrion también demuestra la convergencia del procedimiento de Descomposición Generalizada de Benders (DGB), o algunas de sus variantes, bajo varios conjuntos de condiciones. Esta generalización es particularmente atractiva para problemas no lineales. Al tratar problemas no lineales, los fundamentos teóricos de la formulación de DB clásico, basados en la teoría dual en programación lineal, no son del todo válidos. La región factible puede no estar definida por un poliedro y no siempre se puede hablar de puntos extremos o direcciones extremas. Si además la región factible resulta ser no convexa, la solución óptima encontrada puede ser local mas no necesariamente global. Ya que muchos problemas en ingeniería son convexos solamente en la región donde la solución de interés está localizada, los algoritmos de solución recurren a la suposición matemática de convexidad fuerte, que no es restrictiva desde un punto de vista práctico. En otros casos, la región factible se vuelve convexa por el solo hecho de fijar un subconjunto de variables,

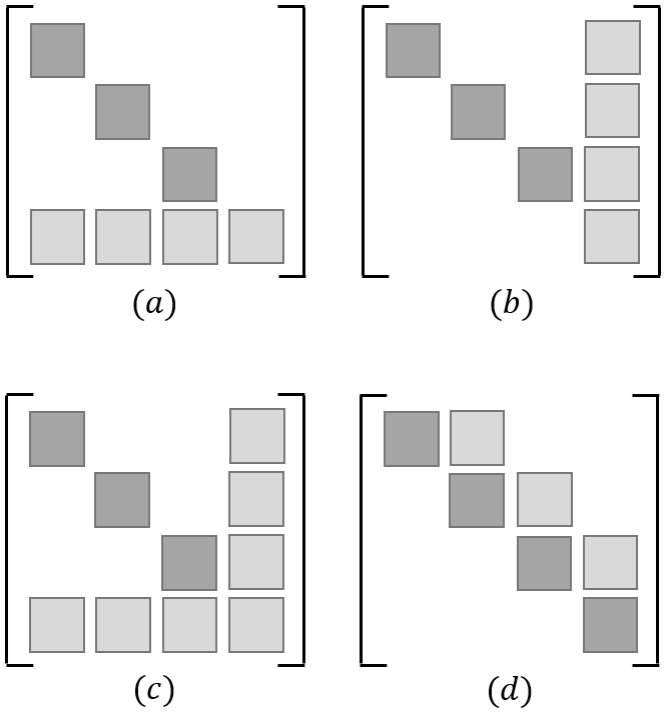

Figura 4-1.: Estructuras matriciales por bloques: (a) diagonal por bloques con borde horizontal, (b) diagonal por bloques con borde vertical, (c) en forma de flecha, (d) banda diagonal

como en el caso de las variables de *complicación* [66].

La DGB ha sido empleada en varios estudios de planeación en sistemas eléctricos de potencia, incorporando ciertas simplificaciones como: el uso de modelos de flujo de potencia AC modificado [35] o linealizado [67], o el modelo DC de la red de transmisión [68]; la relajación de las variables binarias para acotarlas en un intervalo real [69]; o suposiciones como la operación del sistema a un factor de potencia constante haciendo de la potencia reactiva una variable dependiente de la potencia activa [70]; o combinándolo con otros métodos para su aceleración como *Progressive Hedging* [71], un algoritmo iterativo de generación de columna y restricción [72], o un algoritmo de aproximación exterior [73].

4.3.1. Deducción del subproblema y del problema maestro para DGB

Siguiendo la premisa del método de DGB, se identifica un grupo de variables de *complicación* que permita distinguir algún tipo de estructura en el problema original que facilite su división en dos problemas que serán resueltos de forma iterativa: un problema propondrá el valor de las variables de *complicación* y el otro lo retroalimentará a través de información dual,

indicando como cambiar el valor de las variables de *complicación* en la siguiente iteración con el fin de alcanzar una solución óptima. El primer problema es conocido como el maestro y el segundo como el subproblema.

Considere el problema no lineal (4-27) dado por:

$$\begin{aligned} \min_{x,y}\ & f(x,y) \\ \text{s.t. } & g_i(x,y) = 0 \qquad i = 1,...,n \\ & h_j(x,y) \leq 0 \qquad j = 1,...,m \\ & x \in X \\ & y \in Y \end{aligned} \tag{4-27}$$

donde, y corresponde al conjunto de variables de *complicación*. Si se fija el valor de las variables y a $\hat{y}$, el subproblema queda definido como:

$$\begin{aligned} \min_{x}\ & f(x,\hat{y}) \\ \text{s.t. } & g_i(x,\hat{y}) = 0 \qquad i = 1,...,n \\ & h_j(x,\hat{y}) \leq 0 \qquad j = 1,...,m \\ & x \in X \end{aligned} \tag{4-28}$$

La construcción del problema maestro se realiza a través de tres procedimientos, según Geoffrion [66].

1. **Proyección.**
   La idea clave que permite que (4-27) sea visto como un problema en el espacio generado por las variables y es el concepto de proyección. La proyección del problema original en y se escribe como:

   $$\begin{aligned} \min_{y}\ & v(y) \\ \text{s.t. } & y \in Y \end{aligned} \tag{4-29}$$

   donde,

   $$\begin{aligned} v(y) = \inf_{x}\ & f(x,y) \\ \text{s.t. } & g_i(x,y) = 0 \qquad i = 1,...,n \\ & h_j(x,y) \leq 0 \qquad j = 1,...,m \\ & x \in X \end{aligned} \tag{4-30}$$

   El problema auxiliar (4-30) es equivalente al subproblema (4-28).

También se debe definir el conjunto de los valores de y para los que (4-29) es factible.

$$V = \{y : g_i(x, y) = 0, h_j(x, y) \leq 0, \text{ para algún } x \in X\} \tag{4-31}$$

Entonces, $Y \cap V$ es la proyección de la región factible de (4-27) en el espacio de y.

Una serie de observaciones se derivan de esta proyección y se recogen en el siguiente teorema para demostrar la equivalencia entre (4-27) y (4-29).

Teorema 3 *El problema (4-27) es no factible o no acotado si lo mismo ocurre con (4-29). Si (x^*, y^*) es una solución óptima de (4-27), y^* tiene que ser una solución óptima para (4-29). Si y^* es una solución óptima en (4-29) y x^* es el ínfimo en (4-30) cuando se ha fijado y, entonces (x^*, y^*) es una solución óptima del problema original (4-27).*

La dificultad con (4-29) es que tanto la función v como el conjunto V sólo se conocen implícitamente a través de sus definiciones. Benders optó por un método de planos cortantes para construir aproximaciones a v y V, conocidos como *cortes de Benders* como se verá más adelante.

2. **Representación dual de V.**
La representación de V a partir de su dual natural se basa en el siguiente teorema.

Teorema 4 *Suponiendo que X es un conjunto convexo no vacío y que $g_i(x, y)$ y $h_j(x, y)$ son convexas en X para valores fijos de $y \in Y$. Asumiendo además que existe un conjunto cerrado tal que para cada $y \in Y$ se tiene,*

$$Z_y \equiv \{z \in \mathbb{R}^m : g_i(x, y) = 0, h_j(x, y) \leq z, x \in X\}.$$

Entonces, debe existir un punto $\bar{y} \in Y$ que esté en el conjunto V tal que exista un punto z para que el conjunto anterior sea acotado y no vacío.

Esto ocurre cuando el sistema tiene solución:

$$0 \geq \mathcal{L}_*(x, y, \bar{\lambda}, \bar{\mu}), \qquad \forall(\bar{\lambda}, \bar{\mu}) \in \Theta$$

$$\text{donde, } \Theta = \left\{ \bar{\mu} \in \mathbb{R}^m, \bar{\lambda} \in \mathbb{R}^n : \bar{\mu} \geq 0, \sum_{j=1}^{m} \bar{\mu}_j = 1 \right\} \tag{4-32}$$

y donde $\mathcal{L}_*(x, y, \bar{\lambda}, \bar{\mu}) = \bar{\lambda}^T g(x, y) + \bar{\mu}^T h(x, y)$

$\mathcal{L}_*(x, y, \bar{\lambda}, \bar{\mu})$ es la representación dual de V.

Geoffrion [66] ofrece una definición alternativa de multiplicador óptimo generalizado, usando la norma L_1 de las condiciones de holgura complementaria, que además admite multiplicadores con valor cero:

$$\sum_{j=1}^{m} \bar{\mu_j} = 0$$

Los conceptos anteriores sustentan la construcción de los *cortes de factibilidad*. Estos se introducen en el problema maestro cuando la solución del subproblema resulta no factible para un valor fijo de y. Los *cortes de factibilidad* tienen el propósito de orientar al algoritmo en la búsqueda de soluciones factibles para el subproblema.

Los *cortes de factibilidad* introducidos en el problema maestro tendrán la forma:

$$\mathcal{L}_*(x^*, y, \lambda^*, \mu^*) \leq 0 \tag{4-33}$$

Esta función es equivalente a la representación dual de V con la condición adicional de ser menor que cero, evitando que el dual tenga un valor infinito si el subproblema resulta no factible.

3. **Representación dual de v.**
   La representación dual de la función v se basa en el siguiente teorema.

   Teorema 5 *Suponiendo que X es un conjunto convexo no vacío y que $f(x, y)$, $g_i(x, y)$ y $h_j(x, y)$ son convexas en X para cada $y \in Y$. Suponiendo además que para cada $\bar{y} \in Y \cap V$ fijo, se cumple al menos una de las siguientes condiciones;*
   i) $v(\bar{y})$ es finito y (4-28) posee un vector multiplicador óptimo.
   ii) $v(\bar{y})$ es finito, $g_i(x, \bar{y})$, $h_j(x, \bar{y})$ y $f(x, \bar{y})$ son continuos en X, X es cerrado. Entonces, el valor óptimo de (4-30) es igual al de su dual en $y \in Y \cap V$, es decir,

$$v(y) = \sup \theta(\lambda, \mu) \tag{4-34}$$

   El dual del problema (4-30) se obtiene de la definición del dual lagrangiano en (4-7). La representación dual del problema sobre el espacio generado por y se usa para derivar los llamados *cortes de optimalidad*.

Bajo los supuestos anteriores, los tres procedimientos enunciados (4-29), (4-33) y (4-34) producen el problema maestro equivalente:

$$\begin{aligned} &\min_{y \in Y} \ [\sup_{\mu \geq 0} \ [\inf_{x \in X} \mathcal{L}(x, y, \lambda, \mu)]] \\ &s.t. \inf_{x} \mathcal{L}_*(x, y, \bar{\lambda}, \bar{\mu}) \leq 0 \qquad \forall (\bar{\lambda}, \bar{\mu}) \in \Theta \end{aligned} \tag{4-35}$$

O, usando la definición de supremo como la mínima cota superior y añadiendo una variable escalar auxiliar β:

$$\min_{y \in Y} \ \beta \tag{4-36}$$

$$\text{s.t.} \ \inf_{x} \mathcal{L}(x, y, \bar{\lambda}, \bar{\mu}) \leq \beta \qquad \forall \bar{\lambda}, \bar{\mu} \geq 0 \tag{4-37}$$

$$\inf_{x} \mathcal{L}_*(x, y, \bar{\lambda}, \bar{\mu}) \leq 0 \qquad \forall (\bar{\lambda}, \bar{\mu}) \in \Theta \tag{4-38}$$

Aquí β representa el valor óptimo del subproblema dentro del problema maestro. El conjunto de restricciones (4-37) representa los cortes de optimalidad y el conjunto (4-38) los cortes de factibilidad. Por dualidad débil se evidencia que el corte de optimalidad acota por debajo al problema original (4-27), implicando que las soluciones del problema maestro son cotas inferiores del problema original.

El problema maestro reconstruye la región factible del problema original empleando el lagrangiano de las restricciones del subproblema. Si el lagrangiano es lineal en y, el problema maestro construirá la región factible del subproblema mediante hiperplanos. Enumerar todos los cortes (4-38) y (4-37) no conduce a un procedimiento de solución práctico. Para superar esta limitación, Benders propuso relajar los cortes de factibilidad y optimalidad y usar un enfoque iterativo para la solución del problema. La estrategia de descomposición consiste en:

1. Resolver el problema maestro (4-36) con un subconjunto de cortes (4-37) o (4-38) para obtener un valor de prueba de las variables de *complicación* $\hat{y}$ y una cota inferior válida ($\underline{F}$) en el costo óptimo del problema original, dado que el problema maestro es una relajación del mismo.

2. Los valores de las variables de *complicación* calculados entran como parámetros al subproblema (4-30), que se resuelve para generar los cortes de factibilidad (4-38) si el subproblema es no acotado, o los cortes de optimalidad (4-37) si el subproblema resulta factible y acotado. En ese último caso, el subproblema provee una cota superior válida ($\overline{F}$) para el costo óptimo del problema original.

3. Estos cortes se insertan en el problema maestro y el proceso se repite resolviendo de forma alternada el problema maestro y el subproblema hasta encontrar una solución óptima.

La cota superior tiende a ser monotonamente decreciente, mientras que la cota inferior tiende a ser monotonamente creciente al incorporar cortes de optimalidad al problema maestro. Cuando el problema de optimización es no convexo, ambas cotas se acercan pero no llegan a tener el mismo valor dando lugar a la llamada brecha de dualidad. De lo contrario, las dos cotas coinciden en el mismo valor al encontrar la solución óptima del problema original. El

algoritmo finaliza cuando la diferencia entre ambas cotas sean menor o igual a una tolerancia preestablecida ϵ:

$$f(\bar{x}, \bar{y})) - \beta \leq \epsilon \tag{4-39}$$

donde, $\bar{y}$ y $\bar{x}$ son soluciones del problema maestro y del subproblema, respectivamente, y ϵ es la tolerancia de convergencia.

En el peor de los casos, se deberán construir un gran número de cortes (infinitos potencialmente) para lograr la convergencia del algoritmo. En la práctica, es más común emplear una cantidad menor de cortes con precisiones aceptables en los resultados.

Un ejemplo práctico para la aplicación del procedimiento iterativo de DGB es presentado a continuación.

Ejemplo ilustrativo

Considere el problema de optimzación no lineal (4-40):

$$\begin{aligned} \min_{x,y} \ & 6xy + 2x \\ \text{s.t. } & 3y - x^2 \leq 6 \\ & 2y^2 - xy \leq 1 \\ & 1 \leq x \leq 15 \\ & y \geq 0 \end{aligned} \tag{4-40}$$

Para este problema y es la variable de *complicación*. Con $y = \hat{y}$ se define el subproblema como (4-41).

$$\begin{aligned} v(y) = \min_{x} \ & 6xy + 2x \\ \text{s.t. } & 3y - x^2 \leq 6 \\ & 2y^2 - xy \leq 1 \\ & 1 \leq x \leq 15 \\ & y = \hat{y} : \lambda \end{aligned} \tag{4-41}$$

donde, λ es valor óptimo de la variable dual asociada con la restricción de igualdad que fija el valor de la variable de *complicación* y.

Por su parte, el problema maestro (4-42) queda formado como:

$$\begin{aligned} \min_{y} \ & \beta \\ \text{s.t. } & \beta \geq v(y) + \lambda(y - \hat{y}) \\ & y \geq 0 \end{aligned} \tag{4-42}$$

El proceso iterativo de solución por DGB es el siguiente:

Inicialización:

Hacer $w = 1$ contador de iteraciones, $\epsilon = 0{,}01$ tolerancia de convergencia, $y^{(1)} = 1$ y $\underline{F}^{(1)} = -\infty$.

Iteración 1:

- Solución del subproblema:

$$\begin{aligned} v(y) = \min_{x} \; & 6xy + 2x \\ \text{s.t. } & 3y - x^2 \leq 6 \\ & 2y^2 - xy \leq 1 \\ & 1 \leq x \leq 15 \\ & y = 1 : \lambda \end{aligned}$$

La solución es $x^{(1)} = 1$, con un valor de la función objetivo $v(y)^{(1)} = 8$ y $\lambda^{(1)} = 30$. La cota superior actualizada es $\overline{F}^{(1)} = v(y)^{(1)} = 8$

- Chequeo de convergencia:

$\overline{F}^{(1)} - \underline{F}^{(1)} = 8 - -\infty = \infty$

Incrementar el contador de iteraciones $w = w + 1 = 2$

Iteración 2:

- Solución del problema maestro:

$$\begin{aligned} \min_{y} \; & \beta \\ \text{s.t. } & \beta \geq 8 + 30(y - 1) \\ & y \geq 0 \end{aligned}$$

La solución es $y^{(2)} = 0$, $\beta^{(2)} = -22$. La cota inferior actualizada es $\underline{F}^{(2)} = \beta^{(2)} = -22$

- Solución del subproblema:

$$\begin{aligned} v(y) = \min_{x} \; & 6xy + 2x \\ \text{s.t. } & 3y - x^2 \leq 6 \\ & 2y^2 - xy \leq 1 \\ & 1 \leq x \leq 15 \\ & y = 0 : \lambda \end{aligned}$$

La solución es $x^{(2)} = 0$, con un valor de la función objetivo $v(y)^{(2)} = 2$ y $\lambda^{(2)} = 6$. La cota superior actualizada es $\overline{F}^{(2)} = v(y)^{(2)} = 2$

- Chequeo de convergencia:

  $\overline{F}^{(2)} - \underline{F}^{(2)} = 2 - -22 = 24$

  Incrementar el contador de iteraciones $w = w + 1 = 3$

Iteración 3:

- Solución del problema maestro:

$$\begin{aligned} \min_{y} \ & \beta \\ \text{s.t. } & \beta \geq 8 + 30(y-1) \\ & \beta \geq 2 + 6y \\ & y \geq 0 \end{aligned}$$

  La solución es $y^{(3)} = 0$, $\beta^{(3)} = 2$. La cota inferior actualizada es $\overline{F}^{(3)} = \beta^{(3)} = 2$

- Solución del subproblema:

$$\begin{aligned} v(y) = \min_{x} \ & 6xy + 2x \\ \text{s.t. } & 3y - x^2 \leq 6 \\ & 2y^2 - xy \leq 1 \\ & 1 \leq x \leq 15 \\ & y = 0 : \lambda \end{aligned}$$

  La solución es $x^{(3)} = 0$, con un valor de la función objetivo $v(y)^{(3)} = 2$ y $\lambda^{(3)} = 6$. La cota superior actualizada es $\overline{F}^{(3)} = v(y)^{(3)} = 2$

- Chequeo de convergencia:

  $\overline{F}^{(3)} - \underline{F}^{(3)} = 2 - 2 = 0$

  El algoritmo converge con solución óptima $x^* = 1$, $y^* = 0$ y valor óptimo de la función objetivo igual a 2.

La convergencia del algoritmo para este ejemplo se ilustra mediante la Figura **4-2**, en la que se observa que la cota superior y la cota inferior convergen al mismo valor.

4.3.2. Algoritmo para DGB

El procedimiento de solución del método DGB se implementa a través del Algoritmo 4.2. La solución inicial de las variables de *complicación* del *Paso 1* se puede obtener de la solución de un problema como el maestro, sin cortes y dejando fijo el valor de la función objetivo del subproblema. En un sentido práctico, para la construcción de los cortes de optimalidad en el *Paso 2* se usan las variables duales entregadas por cualquier *software* comercial y no es necesario resolver el dual del subproblema.

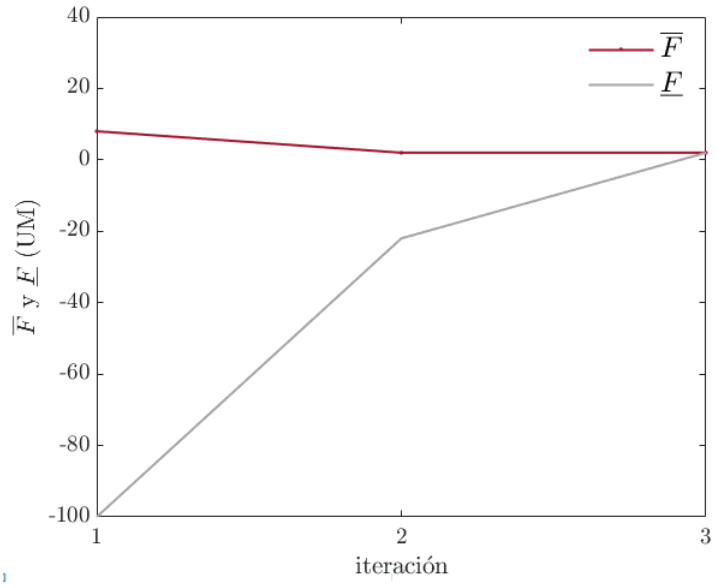

Figura 4-2.: Evolución de las cotas en el algoritmo DGB para el ejemplo ilustrativo

4.3.3. Estrategias para la aceleración de DGB

El método de DGB presenta deficiencias que le impiden proporcionar una solución satisfactoria en un tiempo razonable. Entre estas se pueden mencionar la alta carga computacional del problema maestro, el excesivo tiempo computacional de las iteraciones, la adición de cortes deficientes, algunas iteraciones iniciales ineficaces, el comportamiento oscilante de las soluciones y su lenta convergencia final (*tailing-off effect*) [74, 75].

Diferentes estrategias de aceleración han sido propuestas buscando reducir el costo computacional de cada iteración, o el número de iteraciones, o ambos. La disminución en el tiempo de solución de cada iteración se puede lograr, por ejemplo, calculando soluciones subóptimas del problema maestro y/o del subproblema, o manejando la dimensión y complejidad del problema maestro a través de la compresión o administración de cortes, la generación de cortes más eficientes, o empleando múltiples cortes por iteración en lugar de uno sólo. Otras estrategias consiguen reducir el número de iteraciones, y por ende el tiempo total de cómputo, estabilizando el método en la vecindad de la solución óptima de manera que no se generen oscilaciones. A continuación se describen algunas de las estrategias mencionadas.

1. **Soluciones subóptimas del problema maestro o del subproblema**

   La propuesta se basa en encontrar soluciones subóptimas del problema maestro (o del subproblema) sólo durante las primeras iteraciones, dado que al principio la relajación es débil y no se necesita una solución muy precisa [74].

Algoritmo 4.2 DGB

Hacer $y^1 \in Y$, $w = 1$, $\epsilon \geq 0$, $\overline{F} = \infty$ y $\underline{F} = -\infty$.
Resolver el subproblema (4-30) con $y = y^w$ ▷ *Paso 1.*
si solución del subproblema es óptima **entonces**
    $\overline{F} = \text{mín}(\overline{F}, f(x^w, y^w))$
    si $|\overline{F} - \underline{F}| \leq \epsilon$ **entonces** ▷ Chequeo de convergencia.
        Terminar.
    si no
        $w = w + 1$
        Generar el corte de optimalidad (4-37).
        Ir al *Paso 2.*
    fin si
si no
    Generar un corte de factibilidad (4-38).
    Ir al *Paso 2.*
fin si
Resolver el problema maestro (4-43) con cortes (4-37) o (4-38) ▷ *Paso 2.*
Obtener y^* y β^*
$\underline{F} = \text{máx}(\underline{F}, \beta^*)$
si Problema maestro es no factible **entonces**
    Terminar.
si no
    Ir al *Paso 1.*
fin si

2. **Selección y compresión de cortes**

   Otra alternativa explora la adición parcial de cortes, a través de un mecanismo de selección y compresión. Cuando el algoritmo llegue a una iteración en la que el número de cortes sea igual a un número máximo preestablecido, se seleccionan los cortes que estén inactivos y se descartan, dejando sólo los cortes activos ya que son indispensables. Si el número de cortes aún es grande, se comprime la información de varios cortes para formar uno como una combinación convexa de ellos. Por ejemplo, se pueden agrupar varios subproblemas para agregar un corte por grupo como en [76]. Aunque se puede acelerar la convergencia del algoritmo se corre el riesgo de perder información como lo menciona [77].

3. **Administración dinámica de cortes**

   En cada iteración se acumula un número importante de cortes definiendo el modelo.

Si se tiene un comportamiento inestable este fenómeno se amplifica. En este último caso, las oscilaciones producidas cerca del punto óptimo hacen que muchos cortes generados sean similares y el problema maestro además de crecer en tamaño, se vuelve mal condicionado. Esto hace que el problema maestro sea muy difícil de resolver. Por ende, una buena cantidad de los cortes generados no contribuyen a la convergencia del algoritmo pero si lo ralentizan.

Reducir el número de cortes en el problema maestro es viable puesto que en cualquier solución óptima del mismo, el número de restricciones activas nunca excede el número de variables de decisión, así que los cortes inactivos pueden ser eliminados. Sin embargo, cabe la posibilidad de remover cortes necesarios o agregar cortes inútiles dado que no hay una forma confiable de identificarlos, por lo que estas técnicas son consideradas heurísticas. Los cortes deben eliminarse con poca frecuencia porque la eliminación de restricciones puede impactar de forma negativa a los solucionadores de optimización disponibles. Por ejemplo, el tener que regenerar cortes ya eliminados hace que el algoritmo entre en un funcionamiento cíclico [78].

4. **Versión multicorte**

Una formulación alternativa a la versión estándar monocorte, consiste en agregar un corte por cada subproblema en cada iteración. Esta versión denominada formulación multicorte, busca reforzar más rápidamente el problema maestro. Sin embargo, debe considerarse que el tamaño del problema maestro crece con mayor rapidez al agregar más cortes por iteración, por lo que la compensación entre el número de iteraciones y el tiempo de cálculo depende del problema. Adicionalmente, debido a que los dos métodos no siguen necesariamente el mismo camino, puede darse el caso que el método monocorte funcione mejor que el método multicorte. Birge y Louveaux [79] propusieron como regla general, preferir la formulación multicorte cuando el número de subproblemas no es mucho mayor que el número de restricciones en el problema maestro.

La variante multicorte del problema maestro para la DGB de la Sección 4.3.1, esta representado por (4-43), con $s \in S$ como el conjunto de índices de subproblemas.

$$
\begin{aligned}
&\min_{y} f(y) + \sum_{s \in S} \rho_s \beta_s \\
&s.t. \quad h(y) \leq 0 \\
&\qquad \inf_{x_s} \mathcal{L}(x_s, y, \lambda_s, \mu_s) \leq \beta_s \qquad \forall s \in S \\
&\qquad \inf_{x_s} \mathcal{L}_*(x_s, y, \lambda_s, \mu_s) \leq 0 \qquad \forall s \in S \\
&\qquad y \in Y
\end{aligned}
\tag{4-43}
$$

El Algoritmo 4.2 es válido para la opción multicorte cambiando el problema maestro del *Paso 3* por (4-43) y la generación de los cortes de optimalidad del *Paso 2*.

5. **Regularización y Estabilización**

Un método tipo planos cortantes como Benders tiene buenas propiedades de convergencia para funciones convexas en forma de "V". Sin embargo, para funciones generales, el algoritmo puede presentar inestabilidades y mal comportamiento numérico cuando este se aproxima a la solución óptima [80]. El proceso de solución puede llegar a exhibir un comportamiento oscilatorio cerca del punto óptimo, ocasionando una lenta convergencia del algoritmo. Por lo tanto, resulta conveniente adoptar un mecanismo que estabilice la descomposición, acelerando su convergencia.

Los métodos de haz (*bundle method*) han sido propuestos para estabilizar algoritmos del tipo planos cortantes. La idea general detrás de estos métodos consiste en encontrar puntos de solución (llamados candidatos) que se mantengan cercanos al mejor punto encontrado (llamado centro de estabilidad), siendo este el que ha producido un decremento sustancial de la función objetivo. Un punto candidato se puede convertir en el centro de estabilidad si satisface la siguiente prueba de descenso:

$$f(y^{w+1}) \leq f(\hat{y}^w) - m\delta_{w+1} \tag{4-44}$$

donde, δ_{w+1} representa el decremento nominal y $m \in (0,1)$ es un parámetro dado.

Si f disminuye por al menos una fracción m del decremento predicho por el modelo, el centro de estabilidad se moverá a y^{w+1}. La iteración en la que se actualiza el centro de estabilidad se considera como paso serio. Por el contrario, si el punto candidato no ofrece un decrecimiento nominal significativo de f, el centro de estabilidad se conserva y esa iteración se considera un paso nulo. Si bien el algoritmo puede generar pasos nulos con un valor en la función objetivo menor que en un paso serio, su disminución no se considera suficientemente bueno en términos del decremento nominal. Esto es importante cuando los cortes que definen la región factible son inexactos, es decir, han sido construidos a partir de la solución subóptima del subproblema.

El decremento nominal o esperado es definido por:

$$\delta_{w+1} = f(\hat{y}^w) - v_w(y^{w+1}) \tag{4-45}$$

Tres variantes del método de haz han sido las más empleadas: proximal, región de confianza y nivel. Cada variante se caracteriza por un parámetro de estabilización, actualizado en cada iteración por un procedimiento que puede ser heurístico y que hace

variable el desempeño del método. En teoría, las tres variantes son equivalentes como lo demuestran Bonnans *et al.* mediante el Teorema 10.7 en [80].

a) **Método de haz proximal**
Un término cuadrático es agregado a la función objetivo del problema maestro para mantener la solución cercana al centro de estabilidad actual $\hat{y}^w$.

$$\min_{y \in Y} f(y) + \frac{1}{2}\pi^w \|y - \hat{y}^w\|_w^2 \tag{4-46}$$

donde, $\pi^w \geq 0$ es el parámetro de estabilización. Una regla general para actualizar este parámetro, en un paso serio, es conocida como *forma inversa*. A diferencia de otras, esta regla no es completamente heurística, está soportada por análisis convexo y teoría cuasi-Newton (su deducción detallada se puede consultar en [80]).

$$\frac{1}{\pi_{w+1}} = \frac{1}{\pi_w} + \frac{\langle y^{w+1} - y^w, \lambda^{w+1} - \lambda^w \rangle}{\|\lambda^{w+1} - \lambda^w\|^2} \tag{4-47}$$

donde λ^{w+1} y λ^w son las variables duales usadas para la construcción de los cortes, en dos iteraciones o pasos serios sucesivos.

Aunque el costo computacional por iteración puede incrementar, debido a que el problema maestro a resolver es cuadrático en lugar de lineal, el procedimiento en general es compensado por un número de iteraciones que se reduce significativamente.

b) **Método de haz con región de confianza**
En lugar de minimizar el modelo sobre un conjunto fijo, posiblemente grande, se define un conjunto factible que varíe a lo largo de las iteraciones, donde el modelo sea considerado confiable, es decir, una región de confianza.

Teniendo un parámetro $r^w \geq 0$, se define una región centrada en $y = \hat{y}^w$ con radio r^w. El problema estabilizado es:

$$\min_{y \in Y} f(y) \qquad s.t \quad \|y - \hat{y}^w\| \leq r^w \tag{4-48}$$

El parámetro $r^w \to 0$ cuando $w \to \infty$, y escencialmente varía en el intervalo $[r^w, r_{HI}]$, donde r_{HI} es una cota superior para el radio r^w.

Comúnmente, la norma usada para definir la región de confianza es la Euclidiana. Sin embargo, Linderoth y Wright [81] propusieron una formulación para un método de haz con región de confianza usando la norma infinito. En este caso, la región de confianza tiene forma de caja y queda definida por la restricción:

$$-r^w \ \mathbb{1} \leq y - y^w \leq r^w \ \mathbb{1} \tag{4-49}$$

donde, $\mathbb{1} = (1, 1, ..., 1)^T$ y $r^w \geq 0$.

Una regla de actualización simple para aumentar r^w encontrada en [81] es:

$$r^{w+1} = \text{mín}(r_{HI}, 2r^w) \tag{4-50}$$

c) **Método de haz de nivel**
En este método, el siguente iterato y^{w+1} se calcula como el punto más cercano al mejor iterato actual $\hat{y}^w$, dentro de un cierto conjunto y no como el mínimo del modelo usual del método de Benders. Para ello, la función objetivo se cambia por un término cuadrático definido por una norma:

$$\min_{y \in Y} \frac{1}{2}\|y - \hat{y}^w\|^2 \qquad s.t. \quad v_w(y) \leq L_w \tag{4-51}$$

Este método tienen la ventaja de asegurar una cota inferior para DB a través del parámetro de nivel L_w, especialmente cuando el problema maestro es no factible como lo señala [82]. A su vez, la diferencia entre las cotas inferior y superior define un criterio de parada simple.

En cuanto al parámetro de nivel L_w, este puede ser actualizado usando reglas simples, tal como:

$$L_w = \lambda v_w^{up} + (1 - \lambda) v_w^{low}, \qquad 0 \leq \lambda \leq 1 \tag{4-52}$$

6. **Uso de herramientas de cómputo paralelo para potenciar las técnicas de descomposición**

Las herramientas de cómputo paralelo son útiles si la división del problema original da origen a muchos subproblemas que se puedan resolver de forma independiente, es decir, que la comunicación entre ellos sea poca o nula. Estos se resuelven simultáneamente en diferentes procesadores o máquinas de cómputo. Usualmente, un procesador designado como el principal coordina a los otros procesadores esclavos encargados de resolver cada subproblema. El procesador principal distribuye la solución de las variables dadas por el problema maestro entre los procesadores esclavos para que ejecuten

una tarea (la parte del problema que les corresponda). Estos a su vez, devuelven al procesador principal la información dual obtenida al resolver los subproblemas que luego será utilizada para la formación de los cortes. Algunas estrategias de solución pueden incluir la asignación dinámica de tareas, es decir, el siguiente procesador disponible toma el siguiente subproblema hasta que todos los subproblemas son resueltos; o considerar la integración de algunos subproblemas, si estos se pueden calcular secuencialemente, y resolverlos en el mismo procesador cuando el número de subproblemas es considerablemente mayor que los procesadores disponibles.

5. Propuestas para la Solución por Descomposición del Planeador Estocástico MPSSOPF-NL

Modelar la planeación de la operación de un sistema eléctrico de potencia mediante una formulación como la del Capítulo 2, puede llegar a configurar un problema de programación de grandes dimensiones (millones de variables y restricciones), incluso para sistemas de tamaño modesto considerando pocos escenarios y contingencias. Si adicionalmente se modelan las restricciones de los flujos de potencia bajo el modelo AC, aumenta el grado de complejidad del problema resultante que será difícil de resolver.

No obstante, se observan en el problema algunas características que hacen posible la implementación de técnicas de descomposición para su solución. El problema tiene restricciones de acople, como (2-18) -(2-35), que al ser relajadas y agregadas a la función objetivo facilitarían el uso del método de RLA con alguna manipulación del mismo para lograr la separación del problema original en subproblemas independientes. Por otra parte, las variables representando las inyecciones de potencia activa, presentes tanto en restricciones que describen cada estado operativo como en restricciones acopladoras, pueden ser manejadas como un subconjunto de las variables iniciales bajo la filosofía de la DGB para dividir el problema original en un problema maestro lineal y un subproblema conformado por OPFs independientes.

Este Capítulo presenta las estrategias propuestas para la solución del problema estocástico multiperiodo descrito en el Capítulo 2, basadas en los métodos de RLA y DGB, de acuerdo con lo discutido en el Capítulo 4. Cabe anotar que el problema de optimización a resolver es del tipo continuo, en ausencia de todo lo relacionado con la comisión de unidades que es de naturaleza entera-mixta. La aplicación de cada método requiere algunas adaptaciones adicionales, que también son descritas en cada caso. La implementación de los algoritmos de descomposición en plataformas de cómputo paralelo se ha planteado como un complemento a las propuestas referidas en las primeras secciones, razón por la cual al final del Capítulo se encuentra una sección dedicada a exponer las herramientas para cómputo paralelo, tanto en *hardware* como en *software*.

5.1. Descomposición por RLA para MPSSOPF-NL

En [27] se bosquejó la solución del problema por descomposición mediante Relalación Lagrangiana con Lagrangiano Aumentado regularizado al estilo de [58,83] y con los algoritmos derivados a partir de esas referencias fundamentales, en particular, los denominados métodos de subgradiente y de haz. La estratégia básica para dividir la función lagrangiana involucra la duplicación de las variables de inyección de potencia activa y un mecanismo de coordinación de precios a través de variables duales que deben ser actualizadas en cada iteración.

Todas las inyecciones de potencia activa se duplican dando origen a dos conjuntos: p^{tjk} y s^{tjk}. El conjunto s^{tjk} es usado en cada OPF AC representando un estado operativo (caso base o contingencia), de tal manera que todas las restricciones no lineales del OPF (2-9)-(2-11) están en función de las variables θ^{tijk}, V^{tijk}, s^{tijk} y q^{tijk}. El otro conjunto, p^{tjk}, es usado en un programa cuadrático de coordinación central conteniendo todas las restricciones de acople (2-15)-(2-32), restricciones sustitutas del balance de potencia activa para limitar el espacio de búsqueda de p y los costos originales, en función de las variables p, p_c, p_+, p_-, r_+, r_-, δ_+, δ_-, s_+, s_-. La duplicación de variables implica la adición de restricciones que garanticen la igualdad de las dos versiones de las variables:

$$s^{tijk} - p^{tijk} = 0 \tag{5-1}$$

El lagrangiano aumentado asociado con el problema de optimización (2-1)-(2-32), incluyendo la restricción (5-1), está definido por:

$$\begin{aligned}\mathcal{L}(x_c, s, \lambda) =& f(x_c) + \sum_{t\in T}\sum_{j\in J^t}\sum_{k\in K^{tj}}\sum_{i\in I^{tjk}} \lambda^{tijk}(s^{tijk} - p^{tijk}) \\ &+ \sum_{t\in T}\sum_{j\in J^t}\sum_{k\in K^{tj}}\sum_{i\in I^{tjk}} \frac{c}{2}(s^{tijk} - p^{tijk})^2\end{aligned} \tag{5-2}$$

donde, x_c es un subconjunto de x conteniendo solo variables continuas, $f(x_c)$ son los términos de la función objetivo (2-1) que involucran solamente variables continuas x_c, λ^{tijk} es el vector de multiplicadores de *Kuhn-Tucker* de la restricción (5-1) y c es el coeficiente del término de aumentación cuadrático.

A la función (5-2) se añade un término de regularización que evite que la búsqueda se aleje del valor óptimo encontrado en la iteración previa.

$$\begin{aligned}\mathcal{L}(x_c, s, \lambda) =& f(x_c) + \sum_{t\in T}\sum_{j\in J^t}\sum_{k\in K^{tj}}\sum_{i\in I^{tjk}} \lambda^{tijk}(s^{tijk} - p^{tijk}) \\ &+ \sum_{t\in T}\sum_{j\in J^t}\sum_{k\ inK^{tj}}\sum_{i\in I^{tjk}} \frac{c}{2}(s^{tijk} - p^{tijk})^2 \\ &+ \sum_{t\in T}\sum_{j\in J^t}\sum_{k\in K^{tj}}\sum_{i\in I^{tjk}} \left[\frac{b}{2}(p^{tijk} - \hat{p}^{tijk})^2 + \frac{b}{2}(s^{tijk} - \hat{s}^{tijk})^2\right]\end{aligned} \tag{5-3}$$

donde, b es el coeficiente del término de regularización cuadrático que tiene un efecto amortiguador en la trayectoria de los multiplicadores, y $\hat{p}^{tijk}$ y $\hat{s}^{tijk}$ son los valores de las variables en la iteración previa.

La función dual queda definida como:

$$\begin{aligned} &q(\lambda) = \min_{x_c,s,\lambda} \mathcal{L}(x_c, s, \lambda) \\ &\text{s.t. Ecs. (2-9)-(2-32)} \end{aligned} \tag{5-4}$$

El término de aumentación cuadrático impide que la función dual pueda ser separada, es por ello que se reemplaza por su linealización en torno a los valores de la iteración previa ($c(\hat{s}^{tijk} - \hat{p}^{tijk})(s^{tijk} - p^{tijk})$). Luego, se reordenan los términos de la función objetivo en (5-3) agrupándolos de acuerdo con las copias de las inyecciones de potencia activa.

$$\begin{aligned} \mathcal{L}(x_c, s, \lambda) =& f(x_c) \\ &+ \sum_{t\in T}\sum_{j\in J^t}\sum_{k\in K^{tj}}\sum_{i\in I^{tjk}} \left[\frac{b}{2}(s^{tijk})^2 + \left[\lambda^{tijk} - b\hat{s}^{tijk} + c(\hat{s}^{tijk} - \hat{p}^{tijk})\right] s^{tijk}\right] \\ &+ \sum_{t\in T}\sum_{j\in J^t}\sum_{k\in K^{tj}}\sum_{i\in I^{tjk}} \left[\frac{b}{2}(p^{tijk})^2 - \left[\lambda^{tijk} - b\hat{p}^{tijk} + c(\hat{s}^{tijk} - \hat{p}^{tijk})\right] p^{tijk}\right] \\ &+ \sum_{t\in T}\sum_{j\in J^t}\sum_{k\in K^{tj}}\sum_{i\in I^{tjk}} \frac{b}{2}((\hat{s}^{tijk})^2 + (\hat{p}^{tijk})^2) \end{aligned} \tag{5-5}$$

Ahora, la función dual se puede separar en dos problemas independientes:

$$\begin{aligned} q(\lambda) = &\min_{s,\lambda} q_1(s, \lambda) \qquad + \qquad \min_{x_c,\lambda} q_2(x_c, \lambda) \\ &\text{s.t. Ecs. (2-9)-(2-13)} \quad \text{s.t. Ecs. (2-15)-(2-32)} \end{aligned} \tag{5-6}$$

con

$$q_1(s, \lambda) = \sum_{t\in T}\sum_{j\in J^t}\sum_{k\in K^{tj}}\sum_{i\in I^{tjk}} \left[\frac{b}{2}(s^{tijk})^2 + \left[\lambda^{tijk} - b\hat{s}^{tijk} + c(\hat{s}^{tijk} - \hat{p}^{tijk})\right] s^{tijk}\right] \tag{5-7}$$

$$\begin{aligned} q_2(x_c, \lambda) =& f(x_c) \\ &+ \sum_{t\in T}\sum_{j\in J^t}\sum_{k\in K^{tj}}\sum_{i\in I^{tjk}} \left[\frac{b}{2}(p^{tijk})^2 - \left[\lambda^{tijk} - b\hat{p}^{tijk} + c(\hat{s}^{tijk} - \hat{p}^{tijk})\right] p^{tijk}\right] \end{aligned} \tag{5-8}$$

En (5-7) y en (5-8) se han eliminado los términos constantes de (5-5) porque no afectan la minimización.

De esta forma es posible descomponer el problema en un conjunto de OPF AC individuales y en un programa cuadrático central. Los OPF AC, aunque complejos por su no linealidad,

normalmente se pueden solucionar con métodos eficientes tales como los de punto interior, gracias a los costos convexos y a que las pérdidas en la red tienden a hacer un poco más convexas las restricciones de balance de potencia. Por otro lado, el programa cuadrático central, aunque grande, puede ser resuelto para problemas de tamaño real con solucionadores para problemas de programación cuadrática (QP) o programación lineal (LP) disponibles comercialmente.

El proceso complementario de coordinación sucede solamente al nivel de las dos copias de los conjuntos de inyecciones de potencia activa, p^{tjk} y s^{tjk}, forzando su igualdad a través de los costos de coordinación λ^{tjk}, variables duales que deben ser actualizadas en cada iteración dual. Entonces, los OPF sólo exhibirán los costos de coordinación cuadráticos (5-7), mientras que el problema central tendrá los costos originales más los costos de coordinación de acuerdo con (5-8).

Con respecto a la actualización de las variables duales λ^{tjk} se usa un paso de máximo ascenso basado en el subgradiente del lagrangiano con tamaño de paso adaptado. Este esquema de primer orden es fácil de implementar y especialmente apropiado para problemas no diferenciables. El gradiente de la función dual con respecto al multiplicador recupera la restricción de igualdad (5-1) y la regla de actualización del multiplicador en la iteración $w+1$ queda definida por:

$$\lambda_{w+1}^{tijk} = \lambda_{w+1}^{tijk} + \kappa_w (s_w^{tijk} - p_w^{tijk}) \tag{5-9}$$

con κ_w como el tamaño de paso en la iteración w.

La Figura **5-1** esquematiza el proceso de solución mediante RLA, mientras que el Algoritmo 5.1 lista el procedimiento de solución para MPSSOPF-NL descrito en esta sección.

Desempeño del algoritmo RLA con paso de ascenso de primer orden

En la práctica, esta primera aproximación a la solución del problema no lineal mediante descomposición por RLA no resultó suficientemente eficiente para asegurar el éxito en la solución de cualquier problema que sea físicamente factible. La Figura **5-2** ejemplifica la lenta convergencia e inestabilidad del método tomando como caso de estudio el sistema interconectado colombiano con tres escenarios de viento y diez contingencias simples, para un horizonte temporal de veinticuatro horas. Tanto el valor inicial de los multiplicadores de coordinación λ^{tijk} como del vector de inyecciones de potencia activa en el problema central p^{tijk} se obtuvieron de una solución del problema con un modelo DC de la red de transmisión.

El análisis se hace sobre la progresión de las normas L_∞ y L_2 de la restricción de coordinación $s^{tjk} - p^{tjk} = 0$, que sirven como criterio de parada. La simulación se detuvo después de

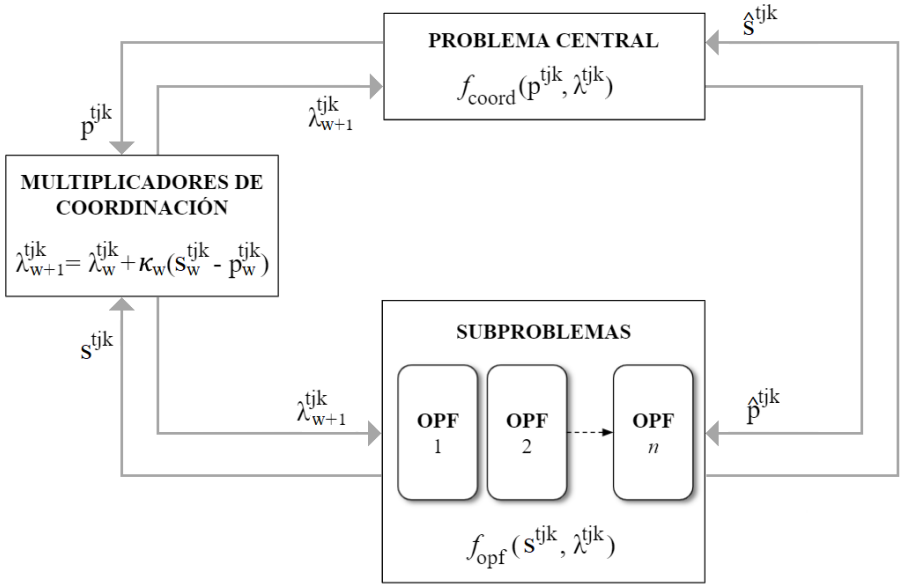

Figura 5-1.: Esquema de Descomposición con RLA para MPSSOPF-NL

Algoritmo 5.1 Descomposición por RLA para MPSSOPF-NL

1: Inicializar el contador de iteraciones $w = 0$, la tolerancia ϵ, el tamaño de paso κ_w y los parámetros de penalización b y c.
2: Resolver MPSSOPF-L para obtener el valor incial de λ_0^{tijk} y p_0^{tijk}
3: Hacer una copia de los OPF que estarán definidos en función de s^{tijk}. Cambiar los coeficientes de la función de costos de los OPF: el coeficiente cuadrático por $b/2$ y el coeficiente lineal por $\lambda_0^{tijk} - bp_0^{tijk}$.
4: Resolver los OPF para obtener el valor incial de s_0^{tijk}
5: Calcular el error inicial $s_0^{tijk} - p_0^{tijk}$
6: **repetir**
7: Hacer $\hat{p}^{tijk} = p_w^{tijk}$ y $\hat{s}^{tijk} = s_w^{tijk}$
8: Linealizar los términos de aumentación cuadrático en torno a los valores de la iteración previa $c(\hat{s}^{tijk} - \hat{p}^{tijk})(s_w^{tijk} - p_w^{tijk})$.
9: Construir los términos de regularización adicionales $\frac{1}{2}(p_w^{tijk} - \hat{p}^{tijk})^2$ y $\frac{1}{2}(s_w^{tijk} - \hat{s}^{tijk})^2$
10: Cambiar los coeficientes de la función de costos de los OPF según (5-7). Resolver los OPF para obtener s_w^{tijk}.
11: Cambiar los costos de coordinación de problema central según (5-8). Resolver el problema central para obtener p_w^{tijk}
12: Calcular el error $s_w^{tijk} - p_w^{tijk}$
13: Actualizar λ^{w+1} con un método tipo subgradiente (Sección 4.2.1) con κ_w
14: Hacer $w = w + 1$
15: **hasta que** $|s^{tijk} - p^{tijk}| \leq \epsilon$

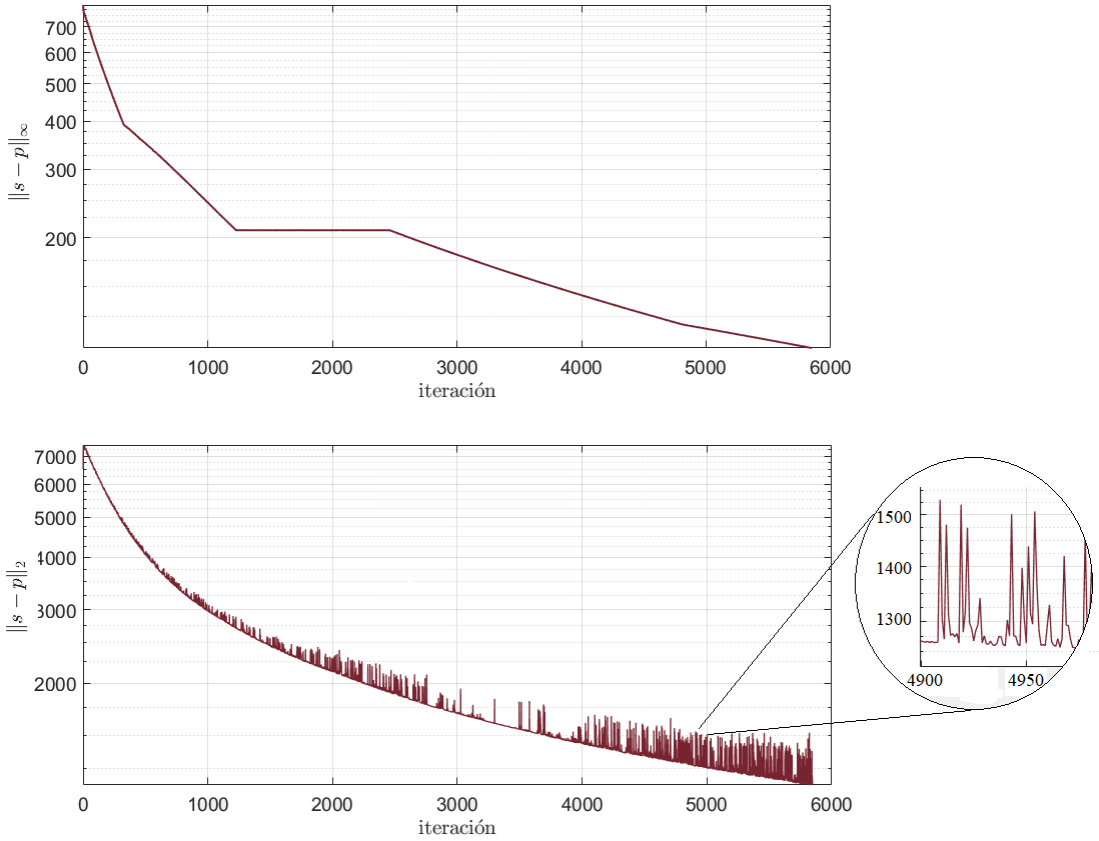

Figura 5-2.: Ejemplo de lenta convergencia del método de descomposición por RLA para MPSSOPF-NL aplicado al caso colombiano de 96 barras

seis mil iteraciones, ejecutadas aproximadamente en cinco días de procesamiento continuo, sin alcanzar la convergencia. Se observa un decrecimiento lento en la norma L_∞, indicando que la máxima diferencia entre las copias de potencia activa disminuyó a lo largo de las iteraciones de forma continua, mientras que en un intervalo de mil quinientas iteraciones en las que esa norma tuvo un valor relativamente constante. Por su parte, la norma L_2 sufrió múltiples oscilaciones cuya magnitud creció de forma apreciable en las últimas iteraciones, evidenciando la inestabilidad de las soluciones halladas por el método de RLA para el caso de estudio. Se identificó que el error $(s-p)$ fue distribuído entre varias unidades de generación y los despachos de los generadores en iteraciones consecutivas también tuvieron comportamiento oscilatorio, ya que en una iteración la diferencia entre s y p tuvo signo positivo y en la siguiente iteración tuvo signo negativo. Por su parte, el costo dual calculado en cada iteración fue creciente en las primeras iteraciones y luego se mantuvo alrededor del mismo rango de valores, dado por un comportamiento oscilatorio con oscilaciones que en su mayoría fueron de pequeña magnitud (0,002 %). Puesto que se dieron discrepancias importantes en los despachos entre las dos copias de las inyecciones de potencia activa, especialmente en las últimas iteraciones registradas, sin afectar de manera significativa el costo dual se presume

que el problema de optimización puede resultar muy plano.

Varios experimentos adicionales fueron realizados cambiando el valor de los coeficientes del término de aumentación cuadrático c y del coeficiente de regularización b, en todos los casos conservando la proporción $b < 2c$. Estos experimentos permitieron constatar que el algoritmo es poco sensible a cambios en c, pero no así para cambios en b. El coeficiente b se aumentó progresivamente partiendo de $b = 0,048$, que fue el utilizado en el experimento previo, y se observó que las normas disminuían más rápido pero igualmente con un comportamiento oscilatiorio, con oscilaciones que iban aumentando su magnitud conforme el valor de b se hacia más grande. Ante valores muy grandes del coeficiente de regularización b (entre 1.700 y 2.500), la norma L_∞ desciende con mayor rapidez y con pocas oscilaciones pero estas son de mayor magnitud que las observadas con un valor de b más pequeño. La norma L_2 y el costo dual igualmente presentan oscilaciones de gran amplitud.

El comportamiento oscilatorio del algoritmo de RLA con actualización de variables duales basado en técnicas de subgradiente fue consistente en todos los experimentos realizados. Esta fue una de las razones por las que, previo al presente trabajo, solamente se hayan reportado implementaciones de la formulación de [27] con las restricciones de la red de transmisión del modelo DC.

5.2. RLA con actualización de variables duales basada en técnicas de sensibilidad

Se puede pensar en la Relajación Lagrangiana como una especie de método de multiplicadores para encontrar los mínimos del problema de optimización que usa un método de tipo subgradiente, que es de primer orden, para la actualización de las variables duales. Pero como en algunos casos un método de actualización de primer orden puede resultar poco eficiente, surge la pregunta si puede ser posible derivar un método de orden superior para hallar $\Delta\lambda$ y acelerar el método de Relajación Lagrangiana. Un ejemplo sencillo, dado por el problema de despacho económico con pérdidas, servirá para ilustrar este supuesto.

El problema de despacho económico con pérdidas esta definido por (5-10),

$$\begin{aligned} \min_{p_g} \; f = & \sum_{i=1}^{n_g} f_i(p_{g_i}) \\ & \text{s.t. } P_D + P_{loss} - \sum_{i=1}^{n_g} p_{g_i} = 0 \end{aligned} \tag{5-10}$$

con, p_{g_i} como la generación de potencia de cada generador, P_D como la demanda total del

sistema y P_{loss} como las pérdidas de potencia.

De forma general, las pérdidas del sistema se pueden aproximar por la fórmula de pérdidas de *Kron*, válida alrededor de un punto operativo,

$$P_{loss}(p_g) = \frac{1}{2} p_g^T B p_g + c^T p_g + d \tag{5-11}$$

donde, los coeficientes $B \in \mathbb{R}^{n_g \times n_g}$, $c \in \mathbb{R}^{n_g \times 1}$ y d dependen del punto de operación del sistema. Estos pueden calcularse analíticamente o estimarse utilizando regresión lineal a partir de muchos puntos de muestra obtenidos de perturbar el despacho al rededor de un punto de operación y registar las pérdidas para cada experimento.

La función lagrangiana del problema (5-10) está dada por,

$$\mathcal{L}(p_g, \lambda) = \sum_{i=1}^{n_g} f_i(p_{g_i}) + \lambda(P_D + P_{loss} - \sum_{i=1}^{n_g} p_{g_i}) \tag{5-12}$$

Las condiciones de optimalidad de primer orden para este lagrangiano son

$$\nabla_{p_g} \mathcal{L}(p_g, \lambda) = \nabla_{p_g} f + \lambda(\nabla_{p_g} P_{loss} - \mathbb{1}) = 0 \tag{5-13}$$

$$\nabla_{\lambda} \mathcal{L}(p_g, \lambda) = (P_D + P_{loss} - \textstyle\sum_{i=1}^{n_g} p_{g_i}) = 0 \tag{5-14}$$

donde, $\mathbb{1} = [1, ..., 1]^T$, $\nabla_{p_g} f = H p_g + h$ si el costo es cuadrático, y $\nabla_{p_g} P_{loss} = B p_g + c$.

Estas ecuaciones se resuelven a través de un método numérico sencillo denominado búsqueda lambda. Dado un valor λ^w en la iteración w, el vector p_g se puede calcular fácilmente de (5-13), que al especificar sus derivadas y reordenar términos queda expresada como :

$$\nabla_{p_g} \mathcal{L}(p_g, \lambda) = (H + \lambda^w B) p_g + h + \lambda^w (c - \mathbb{1}) = 0 \tag{5-15}$$

Para corroborar que ese valor de p_g^w encontrado satisface la restricción de balance de potencia, que necesita hacerse cero, se calcula el error ϵ^w en esta restricción:

$$\epsilon^w = P_D + \frac{1}{2} (p_g^w)^T B p_g^w + c^T p_g^w + d - \sum_{i=1}^{n_g} p_{g_i}^w = 0 \tag{5-16}$$

El valor del error calculado ϵ^w tiene dos interpretaciones. Sí $\epsilon^w > 0$, significa que no se está generando suficiente potencia para suplir la demanda y las pérdidas en el sistema. Entonces, la estrategia para incrementar la generación será incrementar el valor de λ en la siguiente iteración. Por el contrario, sí $\epsilon^w < 0$, significa que hay un exceso de generación y el valor de λ deberá disminuir en la próxima iteración.

El proceso iterativo transcurre entre la solución de (5-15) con un λ dado para calcular el vector de generación p_g, y la actualización de λ de acuerdo con el error ϵ mediante:

$$\lambda^{w+1} = \lambda^w + \kappa\epsilon^w \tag{5-17}$$

donde, κ es un un factor de paso. Esta regla de actualización corresponde al clásico método del subgradiente usando el lagrangiano con respecto a λ, que en este caso es la misma función de error (5-16).

Luego, para saber cuál es la sensibilidad de λ a los cambios en generación Δp_g, se debe encontrar la sensibilidad de la ecuación de error (5-16) a λ tal que permita determinar un $\Delta\lambda$ que lleve el error de generación a cero en la próxima iteración.

Sea $F = \nabla f + \lambda(\nabla P_{loss} - \mathbb{1}) = 0$, donde ∇ es el gradiente con respecto a p_g. Entonces, la variación con respecto a Δp_g, $\Delta\lambda$ es:

$$\Delta F = \nabla^2 f \Delta p_g + \lambda\nabla^2 P_{loss}\Delta p_g + (\nabla P_{loss} - \mathbb{1})\Delta\lambda = 0 \tag{5-18}$$

Despejando Δp_g de esta ecuación y reemplazando los términos correspondientes,

$$\begin{aligned}\Delta p_g &= (\nabla^2 f + \lambda\nabla^2 P_{loss})^{-1}[(\mathbb{1} - \Delta P_{loss})\Delta\lambda] \\ &= [(H + \lambda B)^{-1}(\mathbb{1} - Bp_g - c)]\Delta\lambda\end{aligned} \tag{5-19}$$

La ecuación (5-19) proporciona un vector con las sensibilidades de los p_{g_i} a λ, indicando cuanto cambia cada generador ante un cambio en λ. Adicionalmente, el cambio en generación total esta dado por:

$$\sum \Delta p_{g_i} = \mathbb{1}^T \Delta p_g = \mathbb{1}^T[(H + \lambda B)^{-1}(\mathbb{1} - Bp_g - c)]\Delta\lambda \tag{5-20}$$

Si los cambios en generación Δp_g corresponden exactamente con el desbalance de potencia ϵ, λ se puede actualizar en la iteración w como:

$$\lambda^{w+1} = \lambda^w + \Delta\lambda_w \tag{5-21}$$

donde,

$$\Delta\lambda_w = \frac{\epsilon^w}{\mathbb{1}^T[(H + \lambda B)^{-1}(\mathbb{1} - Bp_g - c)]} \tag{5-22}$$

Lo anterior equivale a usar un método de Newton en la actualización de λ.

La Tabla **5-1** compara los resultados de convergencia de la solución por relajación lagrangiana del despacho económico con pérdidas expuesto, cuando se actualizan los multiplicadores con el método clásico del subgradiente (5-17), que es de primer orden, y con técnicas de sensibilidad (5-21), que es un método de segundo orden.

Tabla 5-1.: Tabla comparativa de la convergencia de dos métodos de actualización de los multiplicadores para RL

Iteración	Error con regla de actualización de $\Delta\lambda$:	
	(5-17)	(5-21)
0	-4.82822	-4.82822
1	-1.65505	0.05314
2	-0.56472	-0.00073
3	-0.19238	0.00001
4	-0.06550	
5	-0.02230	
6	-0.00759	
7	-0.00258	
8	-0.00088	
9	-0.00030	

Estos resultados muestran que la solución por RL con actualización de multiplicadores por técnicas de sensibilidad converge mucho más rápido en comparación con el clásico método del subgradiente, con un error en la restricción de balance de potencia expresado como (5-16) de menor magnitud y que disminuye con mayor velocidad.

En consecuencia, y de acuerdo con lo expuesto anteriormente, se propone cambiar la iteración dual del método de RLA basado en subgradiente para incorporar información de segundo orden en la actualización de los multiplicadores con el fin de mejorar las características de convergencia en el esquema para RLA de la Sección 5.1. De hecho, en muchos casos los algoritmos para programación no lineal más eficientes utilizan información de segundo orden, lo que evidencia la conexión fuerte y aprovechable que estos tienen con los cálculos de análisis de sensibilidad, especialmente en aquellos algoritmos basados en el método de Newton [84]. Aunque la mayoría de las técnicas de análisis de sensibilidad encontradas en la literatura están dedicadas al estudio de la sensibilidad de la solución frente a las variaciones de un parámetro, se pueden usar sus principios básicos para aplicarlos a un análisis de sensibilidad con respecto a un subconjunto de variables del problema.

Dado que las condiciones de optimalidad de *Karush-Kuhn-Tucker* (KKT) de un problema de programación no lineal suponen, en su versión de segundo orden, que las funciones involucradas sean continuamente diferenciables dos veces, es razonable extraer de estas la información de sensibilidad con respecto al subconjunto de variables seleccionado. En primer

lugar, se considera el problema de programación no lineal definido en (5-23) para estudiar la sensibilidad del despacho a precios nodales, partiendo de la solución de un OPF.

$$\begin{aligned} f(p,z,s,x) = \min_{p,z,s,x} \; & \frac{1}{2}\begin{bmatrix} p \\ z \end{bmatrix}^T H \begin{bmatrix} p \\ z \end{bmatrix} + h^T \begin{bmatrix} p \\ z \end{bmatrix} + \frac{1}{2}[s-p]^T C[s-p] \\ & + \frac{1}{2}[p-\hat{p}]^T B_p[p-\hat{p}] + \frac{1}{2}[s-\hat{s}]^T B_s[s-\hat{s}] \\ & + \frac{1}{2}\left[B - A\begin{bmatrix} p \\ z \end{bmatrix}\right]^T C_1 \left[B - A\begin{bmatrix} p \\ z \end{bmatrix}\right] + \frac{1}{2} g(s,x)^T C_2 g(s,x) \end{aligned} \tag{5-23}$$

$$\text{s.t. } s - p = 0 \tag{5-24}$$

$$g(s,x) = 0 \tag{5-25}$$

$$A\begin{bmatrix} p \\ z \end{bmatrix} - B \leq 0 \tag{5-26}$$

donde p,$s \in \mathbb{R}^{n_g}$ son vectores de las copias de todas las inyecciones de potencia activa del problema cuadrático central y de cada OPF, respectivamente; $z \in \mathbb{R}^{n_z}$ es el vector que representa las variables adicionales del problema de optimización central, como las desviaciones ascendentes y descendentes en la inyección de potencia activa, y $x \in \mathbb{R}^{n_x}$ con $n_x = 2n_b + n_g$, es el vector con las demás variables de los OPF (θ,V,Q_g).

B_p, $B_s \in \mathbb{R}^{n_g \times n_g}$ son las matrices de coeficientes positivos del término cuadrático de regularización, $C \in \mathbb{R}^{n_g \times n_g}$ es la matriz de coeficientes positivos del término cuadrático de aumentación y $C_1 \in \mathbb{R}^{n_a \times n_a}$ y $C_2 \in \mathbb{R}^{(2n_b+2n_g)\times(2n_b+2n_g)}$ son las matrices de coeficientes de las otras funciones cuadráticas de penalización.

La restricción (5-24) es la restricción de coordinación que obliga a la igualdad de ambas copias de potencia activa, mientras que (5-25) agrupa las restricciones no lineales de balance de flujo de potencia nodal, los límites en flujos de líneas, los límites de tensión y límites de generación; y (5-26) es el conjunto de restricciones lineales que contiene restricciones inter-temporales e inter-flujo contenidas en el problema cuadrático central.

La función lagrangiana asociada con (5-23)-(5-26) está definida por:

$$\mathcal{L}_{(p,z,s,x,\lambda_c,\lambda_g,\mu)} = f(p,z,s,x) + \lambda_c^T[s-p] + \lambda_g^T g(s,x) + \mu^T\left[A\begin{bmatrix} p \\ z \end{bmatrix} - B\right] \tag{5-27}$$

donde, $\lambda_c \in \mathbb{R}^{n_g}$, $\lambda_g \in \mathbb{R}^{2n_b+2n_g}$ y $\mu \in \mathbb{R}^{n_a}$ son los vectores de los multiplicadores de *Kuhn-Tucker* asociados con las restricciones (5-24) a (5-26).

Si p^*,z^*,s^*,x^* representan un mínimo local de (5-23), las condiciones de KKT se mantienen para ese mínimo, es decir, existen multiplicadores μ^*,λ_c^*, λ_g^* tal que:

$$\nabla_{(p,z)}\mathcal{L}_{\left(p^*,s^*,x^*,z^*,\mu^*,\lambda_c^*,\lambda_g^*\right)} = 0 \tag{5-28}$$

$$\nabla_{(s,x)}\mathcal{L}_{\left(p^*,s^*,x^*,z^*,\mu^*,\lambda_c^*,\lambda_g^*\right)} = 0 \tag{5-29}$$

Estas condiciones necesarias de primer orden garantizan la optimalidad local de la solución. Los gradientes del Lagrangiano (5-27) con respecto a los conjuntos de variables definidos son:

$$\nabla_{(p,z)}\mathcal{L}_{\left(p^*,z^*,s^*,x^*,\mu^*,\lambda_c^*,\lambda_g^*\right)} = H\begin{bmatrix}p\\z\end{bmatrix} + h + \begin{bmatrix}-C\left[s-p\right]\\0_{n_z}\end{bmatrix} + \begin{bmatrix}B_p\left[p-\hat{p}\right]\\0_{n_z}\end{bmatrix} - A^T C_1\left[B - A\begin{bmatrix}p\\z\end{bmatrix}\right] + A^T\mu - \begin{bmatrix}\lambda_c\\0_{n_z}\end{bmatrix} = 0 \tag{5-30}$$

$$\nabla_{(s,x)}\mathcal{L}_{\left(p^*,z^*,s^*,x^*,\mu^*,\lambda_c^*,\lambda_g^*\right)} = \begin{bmatrix}C\left[s-p\right]\\0_{n_x}\end{bmatrix} + \begin{bmatrix}B_s\left[s-\hat{s}\right]\\0_{n_x}\end{bmatrix} + \begin{bmatrix}\frac{\partial g}{\partial s}\\\frac{\partial g}{\partial x}\end{bmatrix}^T C_2 g\left(s,x\right) + \begin{bmatrix}\lambda_c\\0_{n_x}\end{bmatrix} + \frac{\partial g^T}{\partial_{s,x}}\lambda_g = 0 \tag{5-31}$$

Diferenciando las condiciones KKT de primer orden en (5-30) y (5-31), con respecto a cada conjunto de variables y reagrupando los términos, se obtiene:

$$\left[\begin{bmatrix}C+B_p & 0\\0 & 0\end{bmatrix} + H + A^T C_1 A\right]\begin{bmatrix}\Delta p\\\Delta z\end{bmatrix} + A^T\Delta\mu - \begin{bmatrix}I_{n_g}\\0_{n_x}\end{bmatrix}\Delta\lambda_c - \begin{bmatrix}C & 0\\0 & 0\end{bmatrix}\begin{bmatrix}\Delta s\\\Delta x\end{bmatrix} = 0 \tag{5-32}$$

$$\left[\begin{bmatrix}C+B_s & 0\\0 & 0\end{bmatrix} + \frac{\partial g^T}{\partial_{s,x}}C_2\frac{\partial g}{\partial_{s,x}} + \sum_i\left(C_2^i g_i + \lambda_{gi}\right)\nabla_{s,x}^2 g_i\right]\begin{bmatrix}\Delta s\\\Delta x\end{bmatrix} + \begin{bmatrix}I_{n_g}\\0\end{bmatrix}\Delta\lambda_c + \frac{\partial gT}{\partial_{s,x}}\Delta\lambda_g - \begin{bmatrix}C & 0\\0 & 0\end{bmatrix}\begin{bmatrix}\Delta p\\\Delta z\end{bmatrix} = 0 \tag{5-33}$$

Las ecuaciones (5-32) y (5-33) se organizan en forma matricial como se enuncia en (5-34).

$$\mathcal{M}\begin{bmatrix}\Delta p\\\Delta z\\\Delta s\\\Delta x\end{bmatrix} + \left[\begin{array}{c|c|c}-I_{n_g} & 0 & A^T\\0 & & \\\hline I_{n_g} & \frac{\partial g^T}{\partial_{s,x}} & 0\\0 & & \end{array}\right]\begin{bmatrix}\Delta\lambda_c\\\Delta\lambda_g\\\Delta\mu\end{bmatrix} = 0 \tag{5-34}$$

donde $\mathcal{M}$ es la matriz de segundas derivades de la función Lagrangiana (5-27), y está estructurada como:

$$\mathcal{M} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \tag{5-35}$$

donde,

$$M_{11} = \begin{bmatrix} C + B_p & 0 \\ 0 & 0 \end{bmatrix} + H + A^T C_1 A$$

$$M_{22} = \begin{bmatrix} C + B_s & 0 \\ 0 & 0 \end{bmatrix} + \frac{\partial g^T}{\partial_{s,x}} C_2 \frac{\partial g}{\partial_{s,x}} + \sum_i \left(C_2^i g_i + \lambda_{gi} \right) \nabla_{s,x}^2 g_i$$

$$M_{12} = \begin{bmatrix} -C & 0 \\ 0 & 0 \end{bmatrix}$$

$$M_{21} = M_{12}^T$$

Puesto que las relaciones de interés se dan entre Δp, Δs y $\Delta \lambda_c$, el sistema en 5-34 se puede reducir a la expresión dada por (5-36).

$$[\Delta s - \Delta p] = - \begin{bmatrix} -I_{n_g} & 0 & I_{n_g} & 0 \end{bmatrix} \mathcal{M}^{-1} \begin{bmatrix} -I_{n_g} \\ 0 \\ I_{n_g} \\ 0 \end{bmatrix} \Delta \lambda_c \tag{5-36}$$

De forma equivalente,

$$[\Delta s - \Delta p] = \widetilde{\mathcal{M}} \Delta \lambda_c \tag{5-37}$$

La matriz $\widetilde{\mathcal{M}}$ en la expresión (5-37) caracteriza la sensibilidad de la diferencia entre las dos copias de potencia activa (coordinación) a los precios nodales (multiplicadores λ_c).

Es posible reducir el trabajo involucrado en calcular $\mathcal{M}^{-1}$ al eliminar la información de las restricciones no activas en el lagrangiano si se cumplen las siguientes condiciones: *i)* sólo las primeras r restricciones de (5-26) están activas, y *ii)* se mantiene la suficiencia de segundo orden, la independencia lineal y las condiciones de holgura complementaria estricta. La holgura complementaria y la continuidad implican que $\mu = 0$ para las restricciones no activas. Además, pequeños cambios en las variables no harán que las restricciones cambien de no

activa a activa.

El valor de $\Delta\lambda_c$ en (5-37) será utilizado para actualizar los multiplicadores en el método RLA. El algoritmo propuesto tiene similitud con el algoritmo de búsqueda lambda y se esquematiza en el Algoritmo (5.2).

Algoritmo 5.2 Algoritmo búsqueda lambda con información de segundo orden

1: Resolver el problema central sin pérdidas y los subproblemas. Obtener p^0, z^0 s^0, x^0 y λ^0.
2: Iniciar el contador de iteraciones $w = 0$
3: **repetir**
4: Calcular la matriz de sensibilidad $\mathcal{M}$ (5-35), considerando sólo las restricciones activas en (5-24).
5: Calcular $[\Delta s^w - \Delta p^w]$.
6: Resolver (5-37) para obtener $\Delta\lambda_c^w$.
7: Calcular $\lambda^{w+1} = \lambda^w + \Delta\lambda_c^w$.
8: Actualizar los costos del problema central y de cada OPF con los valores en λ^{w+1}.
9: Resolver el problema central y los subproblemas. Obtener p^{w+1},z^{w+1},s^{w+1},x^{w+1}.
10: Hacer $w = w + 1$.
11: **hasta que** $||(s^w - p^w)||_2 \leq \varepsilon$

5.3. Descomposición Generalizada de Benders para MPSSOPF-NL

El método de DGB es aplicable a un problema en el que se identifica un conjunto de variables, que al ser fijadas como parámetros, permita la división del problema de optimización en dos problemas que se puedan resolver de manera independiente. Las variables de inyección de potencia activa (p) en el problema (2-1)-(2-32) están presentes en la función objetivo a través de los costos de despacho y de los costos de rampa de seguimiento de carga, mediante funciones aditivas ponderadas por una probabilidad; pero también hacen parte de restricciones de acople como las de límites de rampa en transiciones desde los casos base a los casos contingentes (2-18) o las restricciones intertemporales de reserva de rampa de seguimiento de carga (2-21)-(2-22). Si tales variables se fijan temporalmente, el problema inicial se puede dividir en un problema maestro lineal conteniendo las restricciones (2-15)-(2-32) y varios subproblemas no lineales restringidos por (2-9)-(2-11) correspondientes a un OPF por cada estado operativo (caso base o contingencia). De esta manera el problema maestro se puede resolver con cualquier solucionador lineal o cuadrático disponible y los subproblemas no lineales mediante el paquete de OPF generalizado de `MATPOWER`.

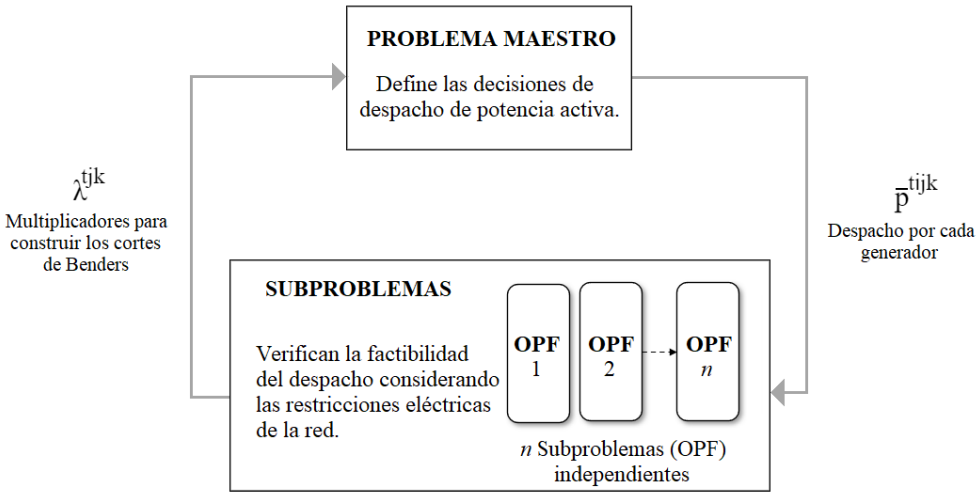

Figura 5-3.: Esquema de DGB para MPSSOPF-NL.

En este esquema de descomposición, el problema maestro realizará las propuestas para los despachos de los generadores, mientras que los subproblemas verifican la factibilidad de tal despacho considerando las restricciones eléctricas de la red. Aunque el problema maestro y los subproblemas se resuelven individualmente, entre ellos existe una comunicación dada por los cortes de Benders como se indicó en el marco teórico de la DGB, desarrollado en el Capítulo 4. La Figura **5-3** esboza de forma simplificada la descomposición por DGB del problema de optimización bajo estudio.

Una vez se ha identificado a p como la variable de *complicación*, se dividen las variables en dos conjuntos x_m y x_{sp}.

$$x_m = (p, p_+, p_-, p_c, r_+, r_-, \delta_+, \delta_-, s_+, s_-) \tag{5-38}$$

$$x_{sp} = (\theta, V, q) \tag{5-39}$$

El conjunto x_m (5-38) comprende las variables del problema maestro, incluyendo la variable de *complicación*. Las demás variables del modelo se dejan en el conjunto x_{sp} (5-39) y corresponden a aquellas del OPF bajo el modelo AC de la red de transmisión. Estas a su vez se pueden dividir en subconjuntos de variables, uno para cada subproblema conformado por un OPF independiente. Luego del ordenamiento de las variables se definen el subproblema y el problema maestro.

El subproblema esta constituido por un conjunto de OPFs independientes que serán modificados como se indica a continuación. Debido a que p entra como un parámetro en el problema de OPF, se disminuyen los grados de libertad necesarios para determinar el sistema. En compensación, se introducen variables penalizadas en las ecuaciones de balance de potencia activa y reactiva (2-9)-(2-10), como en [35] para la solución del problema de

coordinación hidrotérmica mediante DGB, o en [85] para manejar las no factibilidades de las restricciones de tensión y de seguridad de transmisión.

La extensión penalizada de cada OPF AC es formulado como:

$$\widetilde{\beta} = \min_{z_p, z_{px}, z_q, z_{qx}} (R_p z_p + R_{px} z_{px} + R_q z_q + R_{qx} z_{qx}) \tag{5-40}$$

$$\text{s.t. } g_{p_n}(\theta, V, z_p, z_{px}) = 0, \qquad n = 1, ..., n_b \tag{5-41}$$

$$g_{q_n}(\theta, V, q, z_q, z_{qx}) = 0, \qquad n = 1, ..., n_b \tag{5-42}$$

$$h_l(\theta, V) \leq 0, \qquad l = 1, ..., n_l \tag{5-43}$$

$$\theta_{min} \leq \theta \leq \theta_{max} \tag{5-44}$$

$$V_{min} \leq V \leq V_{max} \tag{5-45}$$

$$Q_{min} \leq q \leq Q_{max} \tag{5-46}$$

$$p = \bar{p} \quad : \lambda_{p_n} \tag{5-47}$$

$$z_p, z_{px}, z_q, z_{qx} \geq 0 \tag{5-48}$$

donde, $\bar{p}$ es el despacho propuesto por el problema maestro, z_p, z_{px}, z_q y z_{qx} son las variables representando déficit/exceso de potencia activa/reactiva, respectivamente; R_p, R_{px}, R_q y R_{qx} son los vectores de costo de penalización debido al déficit/exceso de potencia activa/reactiva; y n_b y n_l corresponden al número de barras y de líneas en el sistema, respectivamente.

Las ecuaciones de balance de potencia activa y reactiva modificadas quedan expresadas en términos de las variables penalizadas y del subconjunto de variables x_{sp}, así:

$$\begin{aligned} g_{p_n}(\theta, V, z_p, z_{px}) =& V_n \sum_{b=1}^{nb} V_b Y_{n,b} \cos(\theta_n - \theta_b - \gamma_{n,b}) + p_{d_n} - \sum_{i \in n} \bar{p} - z_{p_n} \\ &+ z_{px_n} = 0 \end{aligned} \tag{5-49}$$

$$\begin{aligned} g_{q_n}(\theta, V, q, z_q, z_{qx}) =& V_n \sum_{b=1}^{nb} V_b Y_{n,b} \sin(\theta_n - \theta_b - \gamma_{n,b}) + q_{d_n} - \sum_{i \in n} q - z_{q_n} \\ &+ z_{qx_n} = 0 \end{aligned} \tag{5-50}$$

Este OPF AC modificado resulta en soluciones siempre factibles, dado que bajo cualquier condición de despacho se permite cerrar el balance de potencia. En consecuencia, no se requiere la construcción de cortes de factibilidad, lo que resulta ventajoso en programación no lineal, ya que no siempre es posible conocer cuando un problema de optimización es no factible o que el minimizador no lineal no pueda alcanzar la solución correcta.

Cada subproblema queda constituido por un OPF del tipo (5-40)-(5-48), para cada estado operativo tjk. La función objetivo (5-40) minimiza el costo de ser incapaz de proveer potencia activa o reactiva al sistema, siempre que las variables z_p, z_{px}, z_q y z_{qx} sean no nulas. Esto sucede cuando la generación propuesta y la red de transmisión no puedan satisfacer el balance nodal. A diferencia de la formulación clásica del OPF, las restricciones en los límites de potencia activa deben ser respetados en el problema maestro para que las restricciones intertemporales puedan ser resueltas sin hacer simplificaciones.

Los subproblemas retroalimentan al problema maestro mediante la construcción de los cortes de optimalidad. Dichos cortes usan información dual (λ_{p_n}) asociada a las restricciones de balance de potencia activa en los subproblemas, dado que sólo en estas restricciones se involucra la variable p. Los multiplicadores indican el rango de cambio del costo de cada subproblema causado por un cambio unitario en p. Los cortes proveen una aproximación lineal del costo de cada subproblema cerca del punto de solución encontrado por el problema maestro. Con esta información mantenida a lo largo de las iteraciones, el problema maestro debe mejorar el despacho propuesto.

Un requisito para la aplicación de DGB es que el subproblema sea convexo, de lo contrario, el corte de optimalidad linealizado puede eliminar soluciones factibles. Como lo señalan Sifuentes y Vargas [35], aunque las no convexidades pueden surgir debido a la presencia de funciones seno y coseno en las ecuaciones de balance de potencia, se observan algunas condiciones que pueden prevenir la ocurrencia de no convexidades en la solución de los subproblemas, tales como: las pequeñas diferencias angulares a lo largo del sistema que en la práctica se presentan, las variables penalizadas $z_{p,q}$ y la magnitud de su penalización que reducen las zonas no convexas, el establecimiento de los límites angulares entre -90 y 90 grados y las pérdidas de transmisión.

El problema maestro queda conformado por una función objetivo (5-51) correspondiente a la minimización de todos los costos considerados en el problema inicial (2-2)-(2-7) más el costo total esperado del subproblema dado por la sumatoria de las variables β^n. Cabe anotar que la DGB presentada corresponde a la versión multicorte (Sección 4.3.3, numeral 4), en la que hay n_{flow} variables β, uno por cada OPF representando estados operativos en el problema de optimización. El problema está sujeto a las restricciones lineales (2-15)-(2-32), a los límites de las variables x_m (5-55), incluyendo los límites en las inyecciones de potencia activa, y a un conjunto de restricciones (5-54) que son los cortes de optimalidad construidos a partir de la información dual aportada por los subproblemas. En cada iteración del método se deberán incorporar al problema maestro varios cortes para la versión multicorte de DGB, los cuales

se van acumulando en cada iteración. Con esto, el problema maestro es:

$$\min_{x_m} f(x_m) + \sum_{n=1}^{n_{flow}} \alpha^n \beta^n \tag{5-51}$$

$$\text{s.t. Ecs. (2-15) - (2-32)} \tag{5-52}$$

$$\text{cortes añadidos en iteraciones previas;} \tag{5-53}$$

$$\beta^n \geq f(x_{sp}^n) + \sum_{n=1}^{n_{flow}} (\lambda_p^n)^T (p^n - \bar{p}^n) \qquad n = 1, ..., n_{flow}, \text{ cortes nuevos} \tag{5-54}$$

$$x_{m,min} \leq x_m \leq x_{m,max} \tag{5-55}$$

La solución x_m de este problema en la próxima iteración será $\bar{x}_m$ y en este conjunto está $\bar{p}$, que propone el despacho de potencia activa dentro de los subproblemas inmediatamente subsecuentes.

El problema maestro y los subproblemas se resuelven iterativamente hasta que se cumpla un criterio de convergencia. Usualmente, la distancia entre una cota inferior y una superior del costo del problema original igualando a una tolerancia preestablecida se elije como criterio de parada. La cota inferior $\underline{F}$ esta dada por el costo del problema maestro, mientras que la cota superior $\overline{F}$ es provista por el costo de los subproblemas más el costo del problema maestro, sin tener en cuenta el valor de las variables β. O lo que es lo mismo, la cota superior se puede representar por la sumatoria de las variables β y la inferior por la sumatoria de los costos óptimos de los subproblemas $\widetilde{\beta}$. En cualquier caso, el criterio de parada está dado por (5-56).

$$\frac{|\overline{F} - \underline{F}|}{\underline{F}} \leqq \epsilon \tag{5-56}$$

En caso de optar por la clásica versión monocorte de la DGB, se designa una sola variable β para el costo total esperado de todos los escenarios, ponderando el costo de cada OPF por su probabilidad de ocurrencia, así:

$$\beta = \sum_{n=1}^{n_{flow}} \alpha^n \beta^n \tag{5-57}$$

Toda la información dual proporcionada por los subproblemas se recogerá en una sola función lineal, o corte de optimalidad, por cada iteración del algoritmo. Tanto los multiplicadores como los valores óptimos de cada OPF se deben ponderar por la probabilidad de ocurrencia del escenario correspondiente antes de ser agregados en una sola ecuación.

$$\beta \geq \sum_{n=1}^{n_{flow}} \alpha^n \widetilde{\beta}^n + \sum_{n=1}^{n_{flow}} \alpha^n (\lambda_p^n)^T (p^n - \bar{p}^n) \qquad n = 1, ..., n_{flow} \tag{5-58}$$

La formulación del problema maestro en su versión monocorte, esta dada por:

$$
\begin{aligned}
&\min_{x_m} \; f(x_m) + \beta \\
&\text{s.t. Ecs. (2-15) - (2-32)} \\
&\qquad \text{cortes añadidos en iteraciones previas;} \\
&\qquad \beta \geq \sum_{n=1}^{n_{flow}} \alpha^n \widetilde{\beta}^n + \sum_{n=1}^{n_{flow}} \alpha^n (\lambda_p^n)^T \left(p^n - \bar{p}^n\right) \qquad n = 1, ..., n_{flow}, \text{ cortes nuevos} \\
&\qquad x_{m,min} \leq x_m \leq x_{m,max}
\end{aligned}
\tag{5-59}
$$

El Algoritmo 5.3 corresponde a la versión multicorte para la DGB del problema de optimización bajo estudio. Por último, para acelerar la convergencia del algoritmo, se implementarán estrategias como las descritas en la Sección 4.3.3 del Capítulo 4.

Algoritmo 5.3 DGB para MPSSOPF-NL

1: Hacer $w = 1$, $tol = \epsilon$, $\overline{F} = 0$ y $\underline{F} = -\infty$.
2: Obtener el valor de prueba $\bar{p} = p^w$ resolviendo (2-1)-(2-32) sin las restricciones de la red (2-9)-(2-11) y con $\sum p - \sum P_D = 0$
3: Resolver cada OPF con $p = \bar{p}$ para obtener λ_p^{tjk}.
4: Actualizar $\overline{F}^w = \sum_{n=1}^{n_{flow}} \widetilde{\beta}^n$
5: Resolver el problema maestro (5-51)-(5-55) después de adicionar cortes (5-54).
6: Hacer $w = w + 1$.
7: Obtener el valor de prueba $\bar{p} = p^w$
8: Actualizar $\underline{F}^w = \sum_{n=1}^{n_{flow}} \beta^n$.
9: **si** se cumple (5-56) **entonces**
10: Parar.
11: **si no**
12: Ir a 3.
13: **fin si**

Es preciso hacer una anotación en cuanto a los precios nodales marginales, obtenidos de los precios sombra de las restricciones de balance de potencia. Estos precios nodales deben averiguarse mediante un procedimiento como el usado a nivel de los mercados de electricidad dado que en los OPF no se usan los costos de despacho del problema original sino los de penalización en las variables de exceso o déficit de potencia activa. En este, los despachos de potencia se obtienen de alguna forma y luego el cálculo de los precios nodales se hace aparte como un problema de optimización incremental. Por ejemplo, se crea un programa lineal o cuadrático en torno al despacho óptimo y se resuelve con un método de Newton o un programa lineal cuya solución es practicamente igual al despacho óptimo tomado como punto incial, pero con los precios nodales calculados sobre las bases correctas. Cabe anotar

que no se recomienda utilizar un método de punto interior para encontrar esos precios nodales cuando se parte de un punto de solución, ya que el reinicio de los métodos de punto interior no es trivial.

5.3.1. Medidas propuestas para la aceleración del algoritmo de DGB

Algunas medidas de aceleración expuestas en la Sección 4.3.3 serán adoptadas para el algoritmo DGB propuesto, buscando reducir tanto el tiempo de cómputo del problema maestro como el de los subproblemas y por ende, del proceso total.

En general, el número de restricciones del problema maestro será mucho mayor que el número de subproblemas, representados por un OPF independiente para cada estado operativo (caso base o contingencia). Por lo tanto, la versión multicorte del Algoritmo 4.2 fue elegida tras comprobar las recomendaciones de [79]. Aprovechando que cada OPF es completamente independiente, se explora la posibilidad de reducir el tiempo de cómputo total del subproblema al resolver los OPF en paralelo a través de un equipo multinúcleo, en lugar de hacerlo en forma secuencial usando un solo núcleo de procesamiento.

Las oscilaciones e inestabilidades durante las últimas iteraciones son comunes en el método de DGB. Esto motivó la incorporación de un método de estabilización, en este caso, el de haz con región de confianza empleando la norma infinito al estilo de [81]. Esta región en forma de caja guarda similitud con las restricciones en los límites de generación de potencia activa, ofreciendo la posibilidad de modelar la región de confianza al hacer modificaciones menores sobre tales límites, en lugar de incluir restricciones adicionales.

En cada iteración w del método DGB estabilizado se modificará el radio de la región de confianza r^w, el cual puede aumentar en iteraciones mayores o disminuir en iteraciones menores. Una iteración mayor es aquella en la que se logra un decremento en el valor de la función objetivo de al menos una fracción $m \in (0, \frac{1}{2})$ del decremento predicho, calculado como la diferencia entre la aproximación de la función objetivo del subproblema $\tilde{f}_{sp}$, obtenida de la solución del problema maestro, y el valor de la función objetivo de la iteración previa f_{sp}^{w-1}. En caso contrario, se dice que la iteración es menor. En una iteración mayor, el radio de la región de confianza aumenta por un factor c y puede variar en el intervalo dado por el valor actual del radio y una cota superior designada por r_{HI}, que es un valor dependiente del problema a resolver. En estas iteraciones además se actualiza el centro de estabilidad $\tilde{p}$. En las iteraciones menores, se disminuye el radio de la región de confianza de acuerdo con alguna regla preestablecida, tal como (5-60).

Los valores iniciales de variables necesarias para el método estabilizado se obtienen de la solución del problema maestro y del subproblema en la primera iteración, tales como el cen-

Algoritmo 5.4 Algoritmo DGB estabilizado por región de confianza

1: Hacer $w = 1$, $tol = \epsilon$, $\overline{F} = 0$ y $\underline{F} = -\infty$ ▷ Para DGB
2: Elegir $m \in (0, \frac{1}{2})$, r_{HI}, $\tilde{p}^0$, r^0, $c \geq 0$. ▷ Para estabilización
3: Obtener el valor de prueba $\bar{p} = p^w$ resolviendo (2-1)-(2-32) sin las restricciones de la red (2-9)-(2-11) y con $\sum p - \sum P_D = 0$
4: **repetir**
5: Resolver cada OPF con $p = \bar{p}$ para obtener λ_p^{tjk}. ▷ Subproblemas
6: Actualizar $\overline{F}^w = \sum_{n=1}^{n_{flow}} \widetilde{\beta}^n$
7: Hacer $f_{sp}^w = \overline{F}^w$
8: **si** $w \geq 2$ **entonces** ▷ Estabilización
9: Hacer $\tilde{f}_{sp}^w = \underline{F}^w$
10: **si** $f_{sp}^w - f_{sp}^{w-1} \leq m(\tilde{f}_{sp}^w - f_{sp}^{w-1})$ **entonces** ▷ Iteración mayor
11: $r^{w+1} = min(r_{HI}, c \cdot r^w)$
12: $\tilde{p}^w = p^w$ ▷ Nuevo centro de estabilidad
13: **si no** ▷ Iteración menor

$$r_h = min(1, r^w)\frac{f_{sp}^w - f_{sp}^{w-1}}{f_{sp}^{w-1} - \tilde{f}_{sp}^w} \tag{5-60}$$

14: **si** $r_h \in (1, 3])$ **entonces**
15: $r^{w+1} = r^w \frac{1}{min(r_h, 4)}$
16: **fin si**
17: **fin si**
18: Actualizar las restricciones en los límites de p,
19: $\widetilde{p}^w - r^w \leq p \leq \widetilde{p}^w + r^w$
20: **fin si**
21: Resolver (5-51)-(5-55) después de adicionar cortes (5-54). ▷ Problema maestro
22: Obtener el valor de prueba $\bar{p} = p^w$
23: Actualizar $\underline{F}^w = \sum_{n=1}^{n_{flow}} \beta^n$.
24: Hacer $w = w + 1$
25: **hasta que** se cumple (5-56)

tro de estabilidad $\tilde{p}^0$ y el valor de la función objetivo del subproblema f_{sp}^0, respectivamente. En cuanto al radio inicial de la región de confianza r^0 este puede ser suministrado como un dato de entrada o calculado como un porcentaje del rango de potencia de los generadores, y en cualquier caso es específico para cada sistema de potencia. El Algoritmo 5.4 resume el proceso para DGB estabilizado incluyendo la actualización de la región de confianza.

Por último, otra medida de aceleración consiste en relajar las tolerancias numéricas de factibilidad y optimalidad de las condiciones de optimalidad de primer orden del solucionador

usado para resolver el problema maestro, y la tolerancia de factibilidad primal (violación de restricciones) del solucionador usado para resolver los subproblemas no lineales. Al relajar las tolerancias se busca reducir el número de iteraciones que toma cada solucionador para converger a una solución óptima, requiriendo menos tiempo de cálculo. Cabe anotar que aún relajadas, las soluciones obtenidas son ingenierilmente significativas.

El diagrama de flujo **5-4** esquematiza la implantación de las medidas de aceleración descritas.

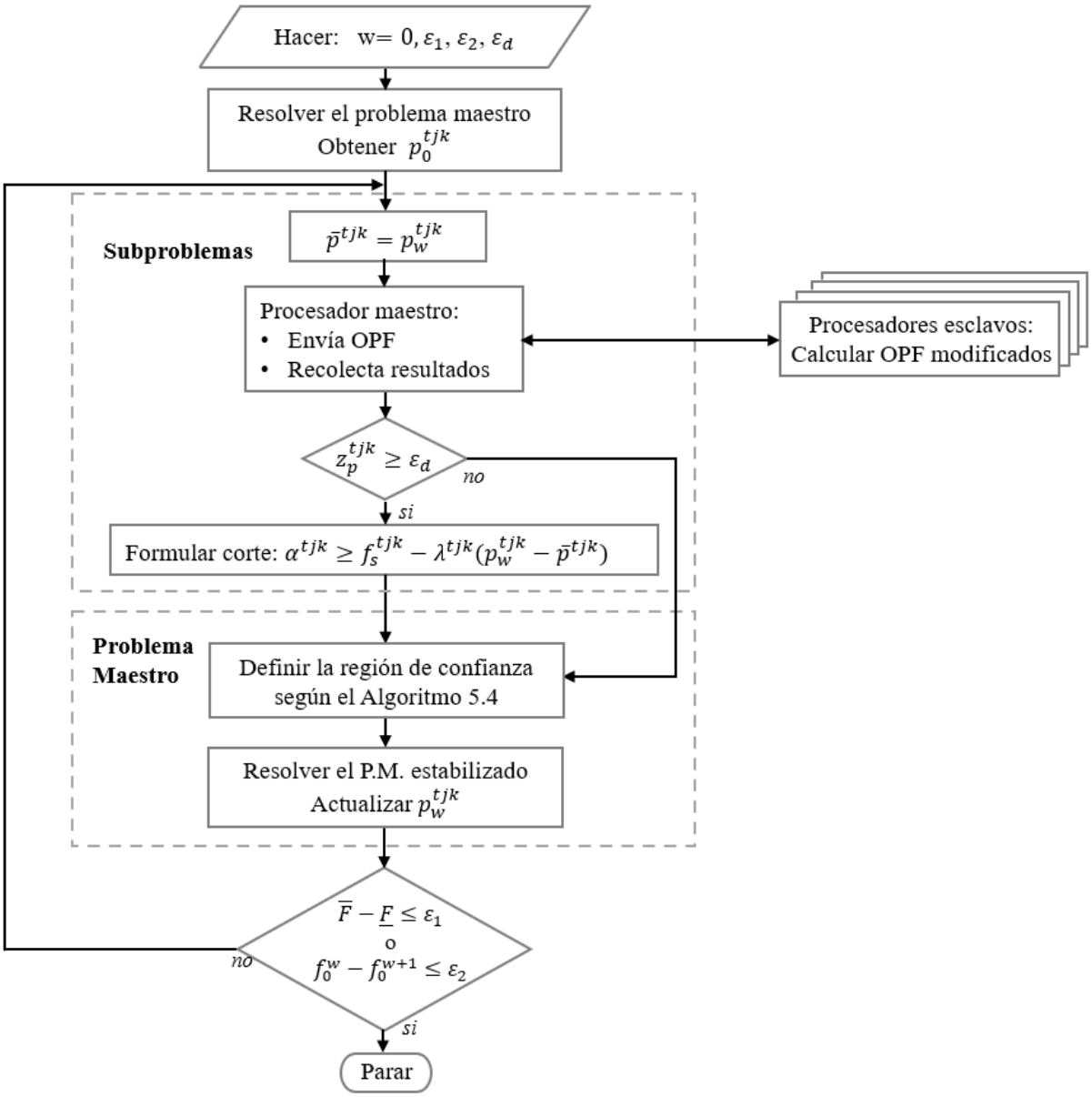

Figura 5-4.: Diagrama de flujo para DGB con técnicas de aceleración e implementación paralela.

5.4. Hardware y software para cómputo paralelo de los subproblemas

El problema de planeación de la operación de sistemas de potencia bajo estudio es un problema de gran escala y el procesamiento de su solución es intensivo, demanda mucho tiempo y recursos de cómputo. No solo se requiere una metodología de solución eficiente, como la descomposición, sino también arquitecturas de cómputo eficientes. Es por ello que las propuestas de descomposición de las secciones anteriores se complementarán y potenciarán con una implementación computacional que explote la división de los subproblemas al nivel de un OPF independiente, para disminuir el tiempo de cálculo.

Existe una tendencia al uso simultáneo de múltiples recursos de cómputo para resolver un problema computacional y acelerar su solución, como se ha reportado en artículos de revisión sobre actividades de investigación en sistemas de potencia [3, 86–88]. Para implementar algoritmos en paralelo, además del *hardware* propio de cómputo paralelo, es indispensable contar con un *software* específico para la administración del sistema y programación del trabajo, a través de procesos que se encarguen de la comunicación entre procesadores o nodos, la asignación de tareas, los protocolos de acceso a memoria, etc.

5.4.1. Hardware

Los sistemas de cómputo paralelo, típicamente están conformados por un solo equipo con múltiples procesadores y/o procesadores multinúcleo, o por un conjunto arbitrario de estaciones de trabajo o computadores (llamados nodos) interconectados por una red de alta velocidad, conocido como clúster. Ambas plataformas se determinan por las características de sus componentes básicos, tales como: uno o varios dispositivos de procesamiento que pueden ser multinúcleo, número de nodos, arquitectura de memoria compartida o distribuída, tipo de red o interconexión, capacidad de almacenamiento, etc.

Principales características de los equipos multinúcleo para cómputo paralelo

Los componentes de un computador multinúcleo más relevantes para aplicaciones de cómputo paralelo son la memoria y el procesador. La arquitectura de la memoria principal del computador es compartida entre todos los elementos de procesamiento al existir un único canal (bus) de comunicación a la memoria. Los accesos a la memoria local son, en general, más rápidos que los accesos a la memoria global. En arquitectura de memoria compartida cada elemento de procesamiento accede a la memoria con una misma latencia, con un mismo canal, y ancho de banda. La latencia es el tiempo, medido en número de ciclos de reloj, transcurrido entre el envío de una petición de lectura de memoria en una posición determinada hasta el momento en que los datos se transmiten a la salida del módulo de memoria.

El ancho de banda de memoria se refiere a la velociad a la cual se puede leer o almacenar datos en una memoria de semiconductor expresada en bytes/s o alguno de sus múltiplos.

Por su parte, el procesador multinúcleo es un procesador con múltiples unidades de procesamiento contenidas en el mismo circuito integrado. Los núcleos individuales pueden ejecutar múltiples instrucciones en paralelo, mejorando el desempeño en la ejecución de un proceso. En el mejor de los casos, el rendimiento obtenido por el uso de procesadores multinúcleo pueden generar factores de aceleración cercanos al número de núcleos, si la fracción del problema a resolver cabe en la memoria caché del procesador, evitando el uso de la memoria principal del sistema que es mucho más lenta.

La memoria caché, de tamaño menor que la memoria principal, es un tipo de memoria intermedia muy rápida que está dentro del procesador, a la que éste tiene un acceso directo. Los procesadores tienen varios niveles de caché comúnmente designados como L1, L2 y L3, que son diferentes en tamaño y rapidez. La memoria caché L1 es la más pequeña y la más rápida porque se encuentra a nivel de cada unidad de procesamiento; la memoria caché L2 tiene mayor capacidad que L1 y su función es la de comunicar los diferentes núcleos, lo que la hace una memoria compartida; y por último, la memoria caché L3 es la más grande pero la más lenta de todas. Cuando se ejecuta un programa, la memoria caché copia los datos contenidos en la RAM cercanos a la localidad de memoria en la que esté trabajando la CPU. Si los datos necesarios no exceden la capacidad de la memoria caché, una copia completa de estos residirá en la memoria caché y la CPU ejecutará sus procesos a la velocidad interna del procesador. De lo contrario, parte de la información requerida por el procesador no se encontrará en la memoria caché y tendrá que ser buscada en la memoria principal (RAM) para poder acceder a esa información, con la correspondiente pérdida de rendimiento por los tiempos de latencia.

Es por ello que, las aplicaciones paralelas ejecutadas en computadores con memoria de arquitectura compartida no siempre alcanzan el valor máximo teórico de aceleración, ya que esta se encuentra condicionada por la capacidad que tenga la memoria caché para contener todos los datos necesarios durante la ejecución de la aplicación paralela, evitando los accesos simultáneos de los núcleos de procesamiento a la memoria principal para obtener información.

Componentes principales de un clúster

Un clúster es un conjunto de equipos de cómputo independientes, conectados mediante una red de comunicaciones, de tal forma que el conjunto opera como un solo computador. La red de comunicaciones puede considerar tecnologías tan simples como Ethernet o tecnologías de alta velocidad como Fast Ethernet, Myrinet, InfiniBand, entre otros. Cada equipo conectado al clúster es conocido como nodo. Los nodos tienen su propia memoria local y funcionan

de forma independiente, por lo que no existe un espacio de direcciones global. Entonces, un clúster puede ser visto como un equipo con memoria distribuída. Además, los cambios realizados sobre la memoria local no afectan la memoria de otros nodos. Cuando un nodo requiere acceder a información de otro nodo, la forma de comunicación entre nodos debe estar definida de antemano por el programador al igual que la sincronización entre tareas.

Un clúster puede ser homogéneo, si esta conformado por equipos de iguales características, en cuanto a *hardware* y sistema operativo, o por el contrario ser heterogéneo. En el primer caso, el balance de carga, es decir, la distribución de las tareas o procesos entre nodos, puede ser má fácil de lograr y en ese caso el clúster sería más eficiente. La eficiencia de un clúster heterogéneo depende del tipo de problema y de su granularidad, y de si esta última es homogénea o heterogénea. Puede ser posible que para un cierto tipo de problema sea más eficiente un clúster heterogéneo, por ejemplo, cuando se tiene un problema en el que uno de los gránulos es de mayor tamaño y en ese caso sería más eficiente tener un nodo con una mayor capacidad para resolverlo, mientras que los otros gránulos se pueden procesar con nodos de menores especificaciones. Si la granularidad del problema es homogénea, un clúster heterogéneo tiende a ser ineficiente cuando la diferencia entre nodos es muy grande, debido a que la asignación de tareas se vuelve más difícil de optimizar.

Para que el clúster efectivamente funcione como un único equipo se requiere una interfaz única de acceso al sistema, operando entre el sistema operativo y las aplicaciones, conocido como *midleware*. Este recibe los trabajos a ejecutarse en el clúster y los distribuye de tal forma que el proceso total se ejecute rápido y no se den cuellos de botella en el sistema.

5.4.2. Software

La plataforma de programación `MATLAB`, sobre la cual están codificados el modelo estudiado y sus algoritmos de solución propuestos, dispone de un ambiente para cómputo paralelo en equipos con procesadores multinúcleo o en clúster de computadores. Porciones independientes de un programa, o una tarea que se ejecuta de forma repetitiva con diferentes datos de entrada se puede correr en los diferentes núcleos de un computador a través del *Parallel Computing Toolbox -PCT* [89] y escalarlo para correr en muchos computadores mediante el *MATLAB Distributed Computing Server - MDCS* [90].

Ejecutar un trabajo paralelo, codificado en `MATLAB`, en un computador multinúcleo requiere la instalación de PCT en el equipo, el uso de funciones de cómputo paralelo de `MATLAB` (*parfor, batch, spmd*) y algunas modificaciones menores en el código del programa. Cada núcleo del procesador se considerará como un trabajador y este ejecutará las tareas localmente. Realizar la misma labor en un clúster de computadores no es tan simple como en el caso anterior y requiere consideraciones adicionales. Además de PCT se necesita la instalación de MDCS,

configurando un ambiente de cómputo paralelo de MATLAB compuesto de tres partes [90]:

1. Una instalación cliente de MATLAB a través de la cual se da la interacción de los usuarios con el clúster. En ella es donde se crean las tareas y los trabajos que posteriormente serán enviados al clúster para su procesamiento. Para que este cliente se comunique con los nodos de trabajo de MDCS, debe tener instalado PCT y una licencia adecuada.

2. Los trabajadores que son los nodos de procesamiento donde se reciben y completan las tareas. Generalmente no hay interacción directa del usuario con los nodos trabajadores.

3. El programador de trabajos encargado de administrar los trabajos, tareas y resultados. Los clientes de MATLAB se comunican con el programador de trabajos para enviar trabajos y tareas. Luego, este transmite las tareas a los trabajadores disponibles y una vez completada su tarea, devuelven los resultados al programador de trabajos, quien transmite los resultados a la sesión del cliente donde el usuario los recibe. Tanto el programador de tareas como el administrador de licencias se ejecutan en el nodo designado como principal.

El *software* MDCS incluye el servicio o "*daemon*" *mdce*, que debe ejecutarse en todos los equipos registrados en un programador de trabajos, para habilitar la ejecución del programador de trabajos o de un trabajador. Este proceso base además asegura que el programador de trabajos y los procesos que controla siempre se estén ejecutando. El servicio *mdce* no se usa con programadores de trabajos de terceros.

Para cualquier sistema de cómputo que se use, el paradigma de programación es del tipo maestro-esclavo, que emplea dos tipos de trabajadores (procesadores/nodos), un trabajador maestro y múltiples trabajadores esclavos. En esta estrategia de balanceo de carga computacional entre los trabajadores, el trabajador maestro dirige la ejecución de las tareas y efectúa la recolección de los datos procesados devueltos por los trabajadores esclavos. En general, este esquema proporciona un buen equilibrio de carga, no requirere comunicación mutua entre esclavos y, por lo tanto, tiene una implementación muy simple. Sin embargo, surge un inconveniente, que se incrementa con el número de procesadores, y está relacionado con la sobrecarga del trabajador maestro cuando varios procesadores envían simultáneamente resultados o solicitudes de tareas adicionales.

Finalmente, dado que cada subproblema ejecuta la misma sección de programa, correspondiente a un OPF, pero con datos de entrada diferentes, la función más apropiada de MATLAB para transformar de serial a paralelo la porción del código que calcula los OPF es *spmd (single program multiple data)*. Basta con cambiar la función *for* por la función *spmd*, establecer los procesos de envío y recepción de información entre los trabajadores esclavos y el trabajador maestro, y extraer los resultados locales en el trabajador maestro para su uso en otras secciones del código.

6. Experimentación

Las técnicas de descomposición presentadas en el Capítulo 5 se aplicaron a la solución del planeador de la operación formulado según el Capítulo 2, para dos sistemas de potencia de diferente tamaño: el caso IEEE de 30 barras y el sistema de potencia interconectado colombiano de 96 barras.

En términos generales, este Capítulo tiene como propósito evaluar el funcionamiento de las técnicas de descomposición propuestas. En primer lugar, se definen y enumeran las suposiciones de modelado asumidas para ayudar en la interpretación de los resultados. Las soluciones generadas por DGB se validan por comparación con la solución sin descomposición del mismo problema de optimización, que será la solución de referencia. Luego de ser validados, los resultados de las diferentes simulaciones permitirán analizar el desempeño específico del método DGB. Igualmente, se estudia la eficiencia de las medidas de aceleración sugeridas, implementadas en una plataforma de cómputo paralelo, de acuerdo con una métrica de desempeño establecida. La parte final del Capítulo está dedicada a la presentación y análisis de las pruebas preliminares realizadas a la propuesta de descomposición por RLA con actualización de variables duales basadas en técnicas de sensibilidad.

6.1. Suposiciones de modelado

Las diferentes suposiciones de modelado asumidas en la ejecución de las pruebas se enumeran a continuación:

- El valor inicial de las variables de *complicación* p^{tijk} para el método DGB se obtiene de una solución del problema sin considerar la red de transmisión.
- El problema de optimización no es cíclico, es decir, no se tiene en cuenta ninguna transición de la hora veinticuatro a la hora uno.
- Se asume que la comisión de unidades se estableció previamente y, por lo tanto, no se tendrán en cuenta las variables, restricciones y costos propios de ese estudio.
- Las variables duales de la restricción $p = \bar{p}$ tiene un valor nulo, porque la función de costos de los OPF no considera los costos de despacho. Entonces, la ecuación de balance de potencia activa es la única restricción que revela la sensibilidad a la variación de la

potencia activa en los OPF. En consecuencia, las variables duales para la construcción de los cortes se asumen iguales a los precios nodales de las barras en las que se conecta cada generador, pero con signo negativo.

- Para el caso IEEE de 30 barras los coeficientes de costos para las variables de déficit de potencia activa y reactiva fueron de 1×10^4 USD/MVA, mientras que los coeficientes de costos para las variables de exceso de potencia activa y reactiva fueron 1×10^2 USD/MVA y 1×10^3 USD/MVA respectivamente. En el sistema de potencia colombiano, los coeficientes de los costos para las variables de penalización de déficit de potencia activa y déficit y exceso de potencia reactiva se fijaron en 1×10^4 KCOP/MVA, mientras que los costos de exceso de potencia activa se fijaron en 1×10^3 KCOP/MVA.

- En el método de estabilización con región de confianza se estimó un radio inicial de la región de confianza equivalente al 25 % del rango entre los límites máximo y mínimo de potencia de cada generador, y un radio máximo igual al triple del radio inicial. Estos valores son producto de ensayos previos y por lo tanto fueron determinados experimentalmente.

- El criterio de convergencia del método DGB tiene dos componentes. El primero, dado por la expresión (5-56), mide la distancia relativa entre la cota superior y la inferior con una tolerancia de 1×10^{-3} pu. En esta expresión, la cota superior corresponde a la sumatoria de los costos de los subproblemas y la cota inferior a la sumatoria de las variables β representando los subproblemas dentro del problema maestro. El segundo componente, mide el cambio absoluto en el costo del problema maestro entre iteraciones y tiene una tolerancia de 1×10^{-6}USD (o KCOP en el caso colombiano).

- La tolerancia de factibilidad primal para los problemas no lineales, así como la tolerancia de optimalidad y facibilidad de las condiciones de primer orden para el problema lineal, se establecieron en 1×10^{-4} pu. Al relajar las tolerancias se busca reducir el tiempo de cálculo, obteniendo soluciones que sean significativas desde un punto de vista ingenieril.

6.2. Especificaciones del software y hardware empleados

El *software* en el que se codificó cada algoritmo de solución fue `MATLAB` versión 2018a y el paquete de herramientas de simulación `MATPOWER` [91] versión 7.1. Los solucionadores `GUROBI` versión 9.1.1 [49] y `IPOPT` versión 3.12.5 [50] se emplearon para la solución del problema lineal y de los subproblemas no lineales, respectivamente. Por último, la herramienta para cómputo paralelo de `MATLAB Parallel Computing Toolbox` versión 6.12 hizo posible la implementación del código paralelo para la solución de los OPFs.

Se utilizó una plataforma de cómputo paralelo para la solución masiva de los subproblemas independientes constituida por una estación de trabajo multinúcleo con arquitectura de memoria compartida, marca `DELL` modelo Precision 7750 con procesador Intel Xeon W-10885M de 8 núcleos, memoria caché de 16MB y 2,4GHz; memoria RAM de 128GB; disco duro de 2TB y sistema operativo Windows 10 Pro. La capacidad de memoria de cada nivel de la memoria caché es de 512KB para L1, 2MB para L2 y 16MB para L3.

6.3. Validación de la solución obtenida por descomposición

La metodología de validación de los resultados de la descomposición se basa en la comparación de estos con los resultados del problema sin descomposición, que será la solución de referencia. La información de interés está compuesta por el costo total, los despachos horarios de potencia activa y reactiva así como las cantidades de reserva de contingencia y de rampa de seguimiento de carga, con resolución horaria. En algunos casos, también resulta pertinente comparar las restricciones que están activas en la solución y los multiplicadores de *Kuhn-Tucker* de las mismas, cuando se presentan resultados con congestión de líneas de transmisión.

La solución de referencia, o solución patrón, se generará con la implementación de MPSSOPF-NL mediante el paquete OPF generalizado de `MATPOWER`, haciendo primero unas modificaciones sobre el código original para MPSSOPF-L [48] el cual esta codificado en `MATLAB` y hace uso de los paquetes propios de `MATPOWER` [91]. Para tal fin, se aprovecharon las capacidades de la arquitectura extensible del paquete OPF generalizado y se prescindió de la dimensión de la comisión de unidades ya que el paquete OPF sólo maneja variables continuas. La idea básica consiste en hacer copias del caso base original, modificándolas para tener en cuenta los cambios que dan lugar a cada una de las contingencias consideradas y agrupar todos estos sistemas en una gran red con múltiples islas.

En primer lugar, se debe modificar la estructura completa del problema de optimización estocástico lineal para hacer coincidir las variables del problema original con la estructura del OPF de `MATPOWER`, así:

$x = (V^{tjk}, \theta^{tjk}, p^{tijk}, q^{tijk}, z^{tjk})$

donde, z^{tjk} representa las variables de usuario adicionales, en este caso:

$z^{tjk} = (p_c^t, p_+^{tjk}, p_-^{tjk}, r_+^t, r_-^t, \delta_+^t, \delta_-^t, p_{sc}^{tjk}, p_{sd}^{tjk}, s_+^t, s_-^t)$

La formulación convencional minimiza los costos de generación, que pueden ser o no lineales. La formulación extendida adiciona una función generalizada de costos cuadráticos definida

por el usuario:

$$\min_{x,z} f(x) + f_{user}(x, z) \tag{6-1}$$

Para este estudio en particular, el primer término representa los costos de generación, escalados apropiadamente por las probabilidades, de todos los estados operativos (caso base y contingencias) en el problema de planeación. El segundo término modela los costos de redespacho, de rampa de seguimiento de carga, de reservas de contingencia y de rampa de seguimiento de carga, y el valor del almacenamiento en estados terminales.

$$\min_{x,z} \sum_{t\in T} \sum_{j\in J^t} \sum_{k\in K^{tj}} \psi_\alpha^{tjk} \sum_{i=1}^{n_g} f_p(p^{tijk}) + \frac{1}{2}\omega^T H\omega + C\omega \tag{6-2}$$

El vector ω es derivado desde las variables de optimización en un procedimiento de varias etapas que el solver de `MATPOWER` implementa internamente; con ello se busca flexibilidad para tratar una amplia gama de funciones de costos. Mayor información al respecto puede encontrarse en [92], sección 6.3.3.

Las restricciones acopladoras como (2-18)-(2-22), (2-23)-(2-32), y las demás restricciones lineales, se pueden trasladar al modelo general para restricciones lineales agrupadas genéricamente por la matriz A y los vectores l y u, como se indica:

$$l \leq A \begin{bmatrix} x \\ z \end{bmatrix} \leq u \tag{6-3}$$

Es preciso anotar que en el caso de contar con cargas flexibles, el paquete OPF de `MATPOWER` se encarga de generar una restricción de igualdad que haga constante el factor de potencia para cada generador negativo en el sistema. Las restricciones (6-3), junto con las restricciones estándar del OPF y las restricciones para el factor de potencia constante de las cargas flexibles, si las hay, completan la formulación del problema para ensamblar una única gran red, que puede ser resuelta mediante el solucionador OPF generalizado de `MATPOWER`.

6.4. Métrica de desempeño de la aplicación en paralelo

Existen varios indicadores que miden el desempeño de un programa en paralelo. Entre ellos, se calcularán dos indicadores en función del tiempo de ejecución tales como la aceleración, que es el más simple y usado, y la eficiencia. El desempeño de la implementación paralela se evalúa sobre los tiempos de ejecución de cada simulación variando el número de núcleos de procesamiento. El tiempo paralelo incluye el tiempo neto de ejecución de cada OPF y el tiempo indirecto gastado por los procesos paralelos como sincronización, comunicación de

datos, inicio y terminación de tareas, etc., entre el núcleo principal y los núcleos esclavos; y la competencia por acceso a memoria en procesadores con arquitectura compartida.

También cabe anotar que los costos computacionales del algoritmo para la solución descompuesta y del algoritmo para la solución sin descomposición, medido en términos de la cantidad de recursos de cómputo (memoria y número de operaciones en punto flotante) requeridos para ejecutar el algoritmo, no han sido evaluados puesto que estos no son comparables. En el caso no descompuesto, las matrices son muy grandes y el costo computacional es cúbico en el tamaño de estas matrices; mientras que, en el caso descompuesto, hay que repetir varias veces la solución de problemas más sencillos. Entonces, el costo computacional de un algoritmo no es necesariamente mayor o menor que el costo computacional del otro.

6.4.1. Aceleración

La aceleración, también conocida como ganancia en velocidad, se define como la relación entre el tiempo de ejecución en serie t_s para resolver un problema y el tiempo de ejecución del algoritmo en paralelo t_p para resolver el mismo problema en n procesadores [93]:

$$A_{cc} = \frac{t_s}{t_p} \tag{6-4}$$

En teoría, el máximo valor de aceleración que puede alcanzar un algoritmo en paralelo es n, si todos los procesadores tienen igual potencia de cálculo. Un valor de aceleración igual a n indica que el tiempo de ejecución en paralelo con n procesadores es n veces menor que el tiempo de ejecución serial. En la práctica, el valor de la aceleración será menor a n considerando que la competencia por el acceso a memoria, en procesadores con arquitectura compartida, y los tiempos de sincronización y comunicación de los procesadores causan un cuello de botella en el proceso de solución en paralelo.

6.4.2. Eficiencia

La eficiencia se define como la relación entre la aceleración y el número de procesadores.

$$E_p = \frac{A_{cc}}{n} = \frac{t_s}{nt_p} \tag{6-5}$$

Básicamente, la eficiencia mide la fracción de tiempo durante la cual un procesador es empleado de manera útil. El máximo valor de eficiencia que en teoría se puede conseguir es 1, representando un 100 % de uso útil de los procesadores.

6.5. Caso de prueba 1: Sistema de potencia IEEE 30 barras modificado

El sistema de potencia IEEE de 30 barras modificado es presentado en la Fig **6-1**. El sistema consta de treinta barras, seis generadores convencionales (térmicos e hidroeléctrico) y cuarenta y un líneas de transmisión. Las modificaciones consistieron en adicionar un generador eólico en la barra 6 y tres cargas con flexibilidad horaria en las barras 7, 15 y 21 con flexibilidad de 30 %, 30 % y 10 %, respectivamente, de acuerdo con [94]. Tanto el generador hidroeléctrico como las cargas flexibles son considerados como almacenadores y modelados como tal.

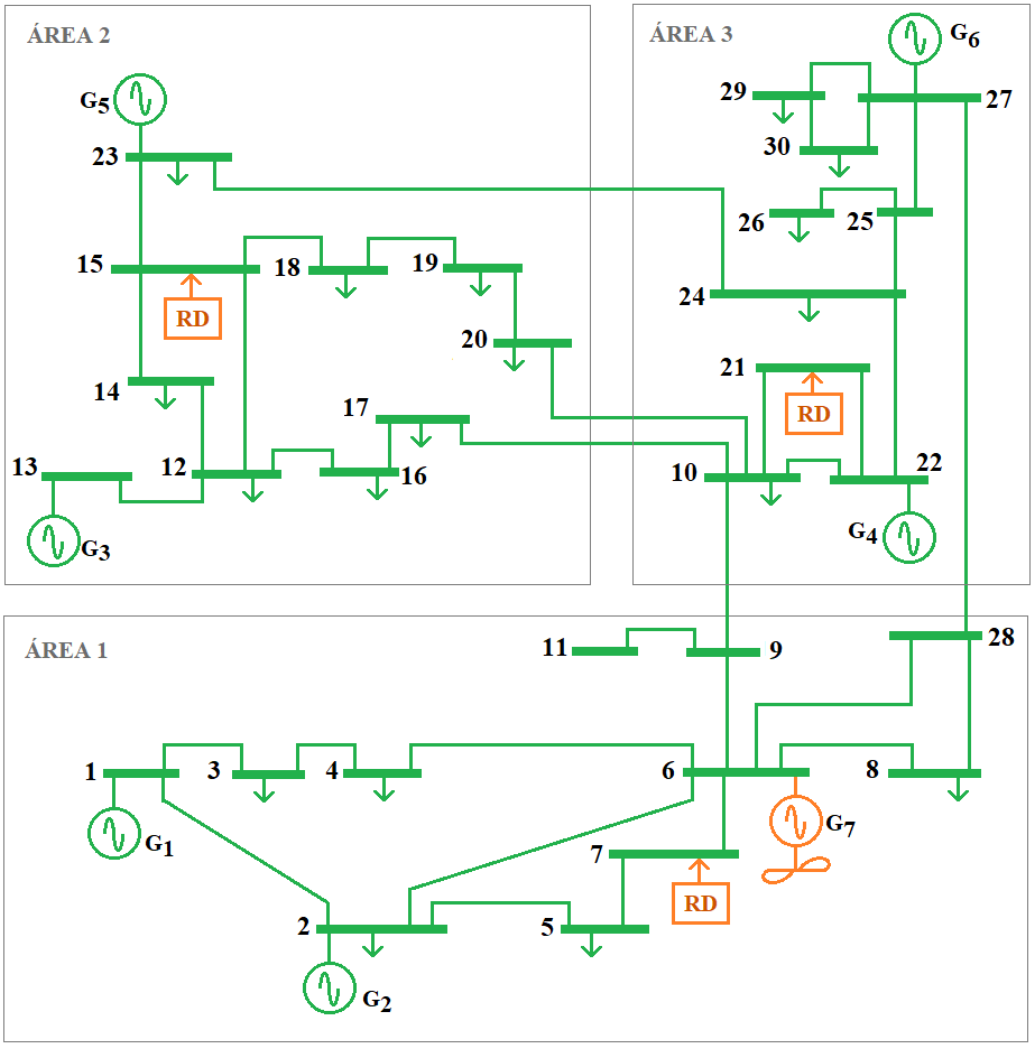

Figura 6-1.: Diagrama unifilar del sistema de potencia IEEE de 30 barras modificado

Los parámetros de los componentes del sistema, algunos tomados de [95, 96], así como los

perfiles de demanda de potencia y de viento tomado de [94], se encuentran en el Anexo C. La Figura **6-2** corresponde al perfil de demanda horario para un periodo de planeación de veinticuatro horas.

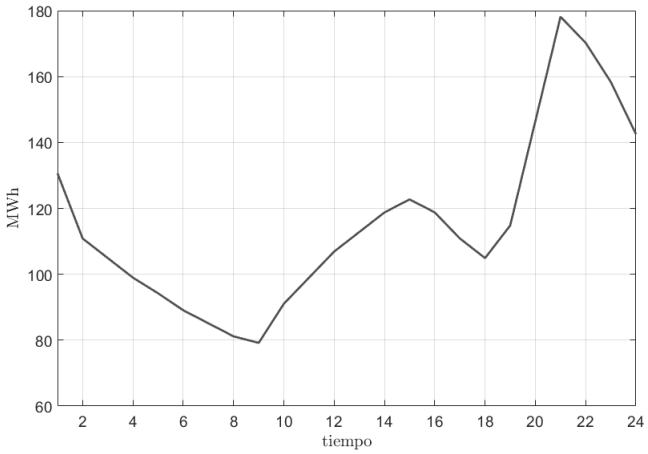

Figura 6-2.: Perfil horario de demanda para el caso de prueba 1. Sistema de potencia IEEE de 30 barras

Las contingencias consideradas se presentan en la Tabla **6-1**, todas tienen la misma probabilidad de ocurrencia (1×10^{-5}).

Tabla 6-1.: Lista de contingencias del tipo N-1 para el caso de prueba 1. Sistema de potencia IEEE de 30 barras

Cont ID	Descripción
1	Aumento del 5 % de la demanda en todas las barras.
2	Desconexión de la línea 9-10 (interconexión áreas 1-3).
3	Desconexión de la línea 10-17 (interconexión áreas 2-3).
4	Desconexión de la línea 23-24 (interconexión áreas 2-3).
5	Desconexión de la línea 25-27.
6	Desconexión de la línea 6-28.
7	Desconexión del generador en la barra 1.

6.5.1. Validación de la solución por DGB del sistema IEEE de 30 barras

La solución del problema de planeación del sistema IEEE de 30 barras, para un horizonte temporal de veinticuatro horas, seis escenarios de viento y siete contingencias, obtenida a

través de DGB se compara con la solución de referencia calculada para el mismo problema. La solución por descomposición no incluyó ninguna de las medidas de aceleración descritas en la Sección 4.3.3, y en consecuencia, la solución de los OPF fue serial.

El costo óptimo del problema de optimización para la solución de referencia fue de USD 9.551,0742 y para la solución descompuesta de USD 9.551,4662, esto es, 0,004 % mayor que la solución de referencia. En cuanto al tiempo de procesamiento, la solución de referencia requirió de 113,0461s y la solución por descomposición de 7.051,7079s en 54 iteraciones. El tiempo de solución del problema maestro a lo largo de las iteraciones estuvo en el rango de 2,3587s a 13,6275s, mientras que en promedio, los OPF de una iteración toman 116,6700s.

La evolución en el valor de las cotas cotas inferior y superior en la solución del problema mediante el método de DGB, aplicado a este caso de estudio, se muestra en la Figura **6-3** y constituye el primer criterio de convergencia. En esta figura se observa que en un principio la cota superior disminuye rápidamente y después se aprecian varias oscilaciones en esta como producto de los cambios abruptos en el nivel de generación de algunas unidades de una iteración a otra. Las cotas convergen a un valor cercano a cero, que correspondió a las penalizaciones por el exceso de potencia activa residual que fue igual a $6,67 \times 10^{-5} MW$.

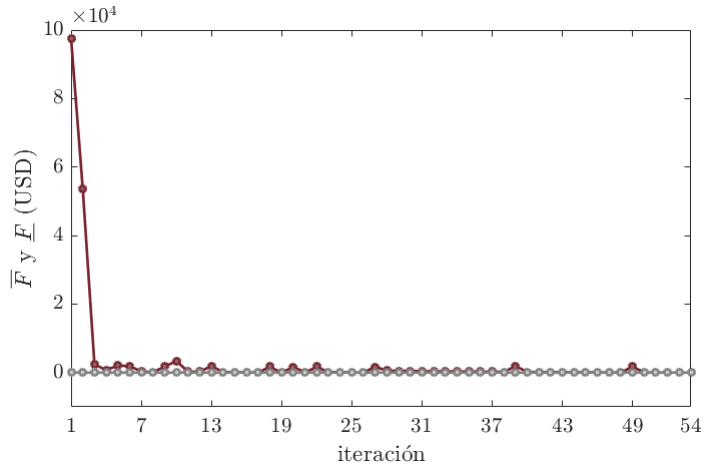

Figura 6-3.: Convergencia de DGB para el sistema de potencia IEEE de 30 barras

El costo óptimo de la función objetivo, que corresponde al costo del problema original calculado con las variables de solución en cada iteración, se grafica en la Figura **6-4**, y sirve para evaluar el segundo criterio de convergencia basado en el cambio absoluto de este costo óptimo entre dos iteraciones sucesivas. El costo óptimo del problema es creciente aproximándose al valor final con algunas oscilaciones después de la décima iteración, que no son apreciables gráficamente porque fueron de magnitud menor, variando entre USD 1×10^{-1} y USD 1×10^{-3}.

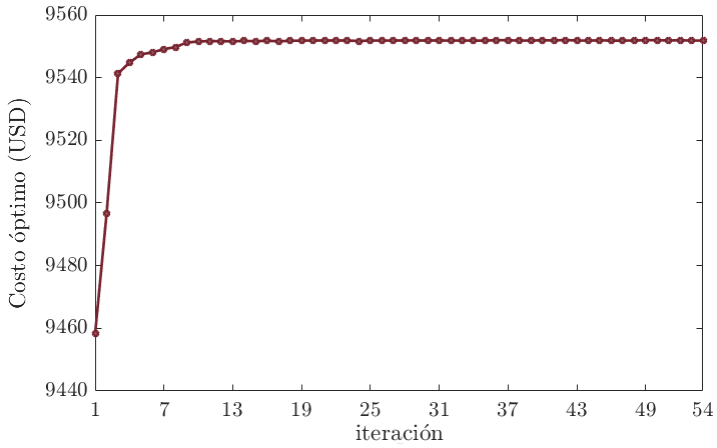

Figura 6-4.: Evolución del costo óptimo en la solución por DGB para el sistema de potencia IEEE de 30 barras

Las Figuras **6-5** a **6-7** muestran la comparación, por generador, de los despachos, las reservas de rampa y las reservas de contingencia entre la solución descompuesta y la solución AC de referencia. Dado que las demandas flexibles de las barras 7, 15 y 21 se modelan como generadores, estas han sido etiquetadas en las gráficas como Generador 8, Generador 9 y Generador 10, respectivamente.

El despacho por unidad de generación en la Figura **6-5** es muy similar en ambas soluciones. De hecho, las mayores diferencias en los despachos esperados se dan en los últimos generadores que representan a las demandas flexibles, especialmente en la demanda de la barra 21 (Generador 10), que tiene asignado el menor porcentaje de flexibilidad. En la solución sin descomposición, la demanda de la barra 21 busca cumplir con su cuota de consumo de energía principalmente durante las horas 3 a 12, que son horas valle; mientras que esa misma carga en la solución descompuesta también intenta cumplir su cuota de consumo mayoritariamente en esa franja horaria, pero adicionalmente traslada una porción pequeña del consumo de energía para las horas 16 a 20, segundo valle en la curva de demanda, donde hay un poco más de generación por parte de unidades de menor costo de producción, en comparación con las horas valle, como la hidroeléctrica (Generador 2) y eólica (Generador 7). Un comportamiento similar se advierte en la demanda flexible de la barra 15 (Generador 9). En general, todas las demandas flexibles cumplen con la cuota de energía variable que les ha sido asignada. Aparte de los generadores mencionados, entre los generadores convencionales el que presenta las mayores discrepancias en el despacho esperado es el generador 2 entre las horas 20 a 24, horas de máxima demanda de potencia, donde la mayor diferencia se dio en la hora 24 en la que el despacho de la solución descompuesta fue 1,2 % mayor, con respecto al de la solución sin descomposición, cantidad que resulta insignificante y no es gráficamente apreciable.

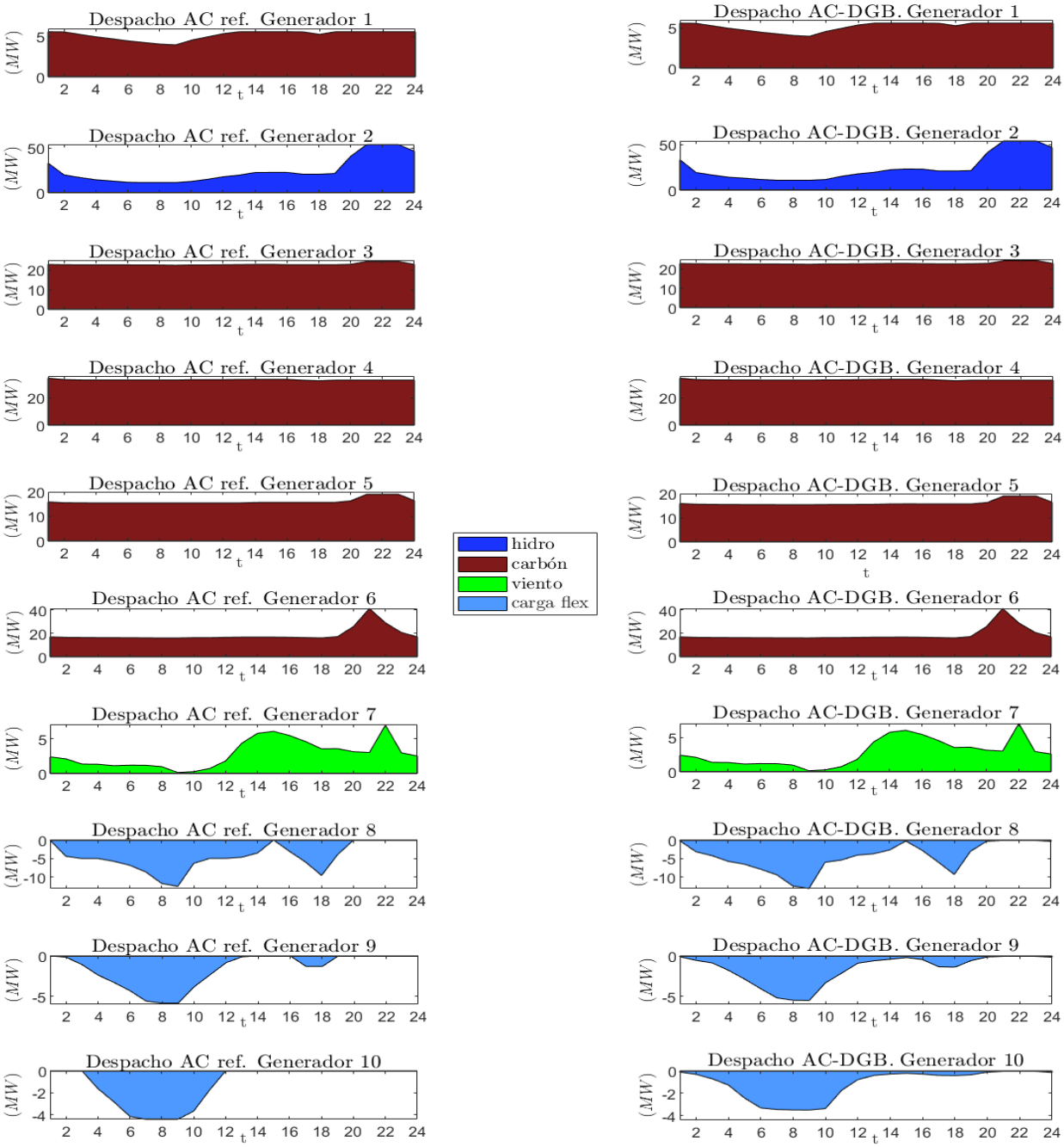

Figura 6-5.: Despachos horarios por unidad, sistema de potencia IEEE de 30 barras. Comparación entre la solución AC de referencia y la solución AC-DGB

En cuanto a la reserva de rampa de seguimiento de carga de la Figura **6-6**, la mayoría de las cantidades asignadas a los generadores convencionales en las dos soluciones son, en términos generales, muy similares. Se observa en la gráfica que la mayoría de la rampa ascendente necesaria para el cambio de toma de carga hacia la hora de máxima demanda se provee

a través de generadores térmicos. Nuevamente, las diferencias más notorias se dan en las demandas flexibles, sobre todo en la demanda de la barra 21 (Generador 10) donde la máxima diferencia entre las dos soluciones es del orden de 1,3MW (2,29MW para la solución descompuesta y 3,61MW para la solución de referencia). Las cargas ejercen su función de arbitraje de energía la mayor parte del tiempo, pero hay momentos en los cuales proveen pequeñas cantidades de rampa, por ejemplo, la demanda en la barra 7 (Generador 8) en las horas 12 a 14 y la demanda de la barra 21 (Generador 10) en las horas 9 a 10. De otra parte, dado que el generador eólico no es despachable, la reserva de rampa que se le asigna debe ser interpretada como la flexibilidad requerida del resto de generadores para aceptar dicha generación eólica. La rampa negativa del generador eólico en las horas 21 a 23, está relacionada con la disminución en su producción de potencia, que no es absorbida por los otros generadores, que estarían obligados a proveer rampa ascendente, porque la carga en este caso disminuye al mismo tiempo que la generación eólica.

En la Figura **6-7** se observan diferencias en las asignaciones de reserva rodante de contingencia en varios generadores entre ambas soluciones. El generador 2 siempre requiere redespachos positivos, aunque las cantidades por hora son muy diferentes y no mayores a 3MW. Según la gráfica, la solución descompuesta tiende a asignar pequeñas cantidades en las cargas de las barras 15 (Generador 9) y 21 (Generador 10) a lo largo del periodo de planeación como reserva rodante ante contingencias, mientras que en la solución de referencia es la demanda de la barra 7 (Generador 8) la que más participación tiene en la reserva rodante de las tres demandas flexibles.

Por otra parte, en la solución de referencia se detectaron líneas con los límites de transmisión activos en el punto óptimo, para varios flujos de potencia en las horas 21, 22 y 23, en el caso base y en cuatro contingencias, las cuatro primeras listadas en la Tabla **6-1**. En la solución obtenida por el método DGB se identificaron correctamente la mayoría de las líneas con la restricción de transmisión activa, como se indica en la Tabla **6-2**. Los flujos de potencia se identifican con los índices t, j, k con t indicando el tiempo, j el escenario de generación eólica y k la contingencia (o caso base si $k = 0$). Las líneas se identifican con los índices de las barras entre las que está conectada.

Ambos algoritmos identificaron correctamente las mismas líneas de transmisión contra sus límites, si el flujo de potencia correspondía a un caso base. En cuanto a los flujos de potencia en las contingencias, en ocho flujos particulares se dieron resultados discordantes. En cinco de tales flujos, la solución descompuesta detectó líneas operando contra sus límites que no se reportaron en la solución de referencia, dado que los flujos de potencia en esas líneas fue menor. En los tres flujos restantes sucedió lo opuesto, y fue en la solución de referencia que se detectaron líneas que no se reportaron en la solución descompuesta aunque con multiplicadores muy pequeños de 4×10^{-6} USD/MVA, razón por la cual es posible que no aparecieran

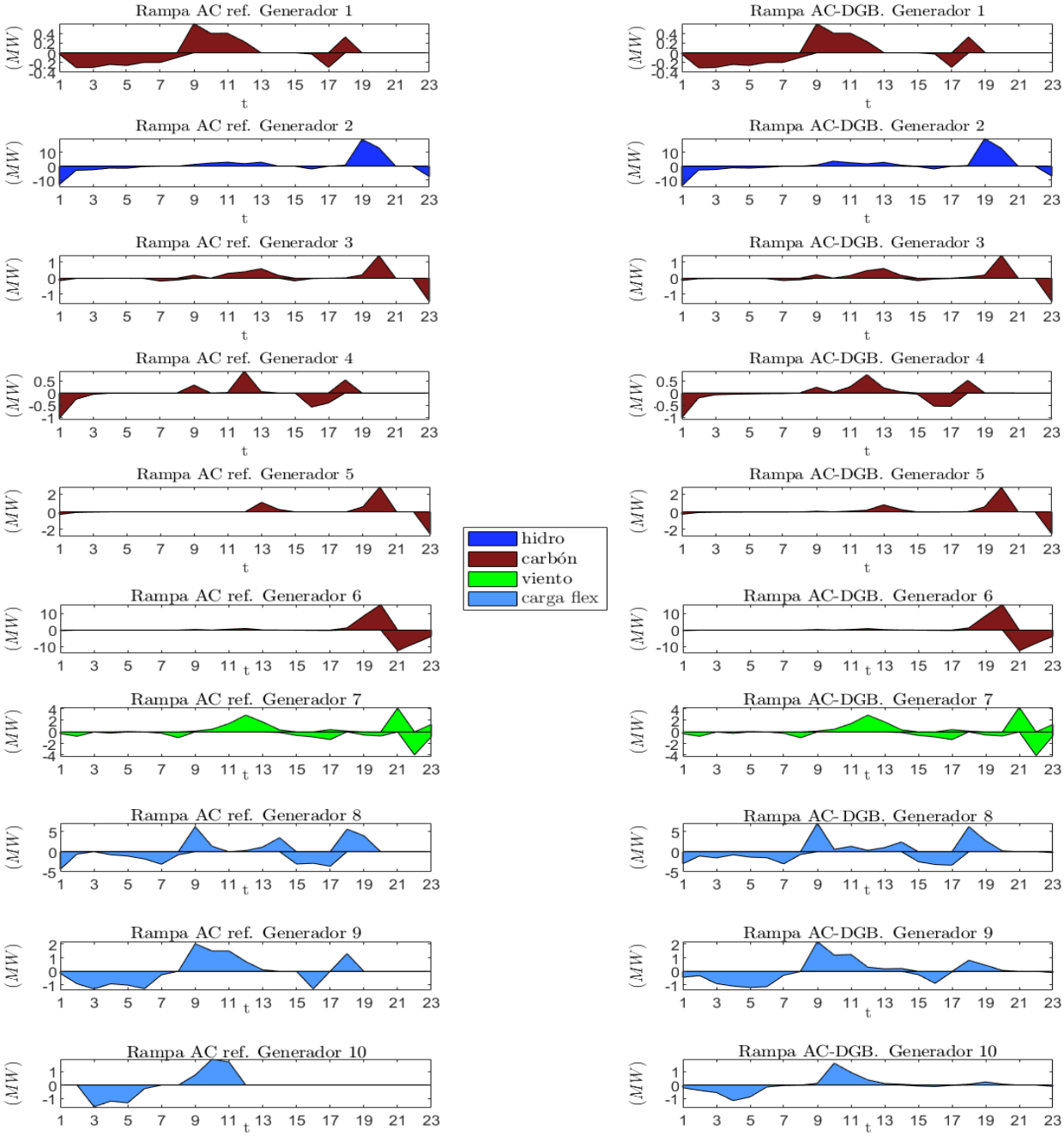

Figura 6-6.: Reservas de rampa por unidad, sistema de potencia IEEE de 30 barras. Comparación entre la solución AC de referencia y la solución AC-DGB

en la solución por DGB. En otros flujos además de identificarse la misma línea operando contra sus límites (línea 6-8), en la solución por descomposición se reportaron otras líneas adicionales pero con multiplicadores muy pequeños comparados con los de la mencionada restricción para la línea 6-8 (por lo menos dos órdenes de magnitud menos). La línea de

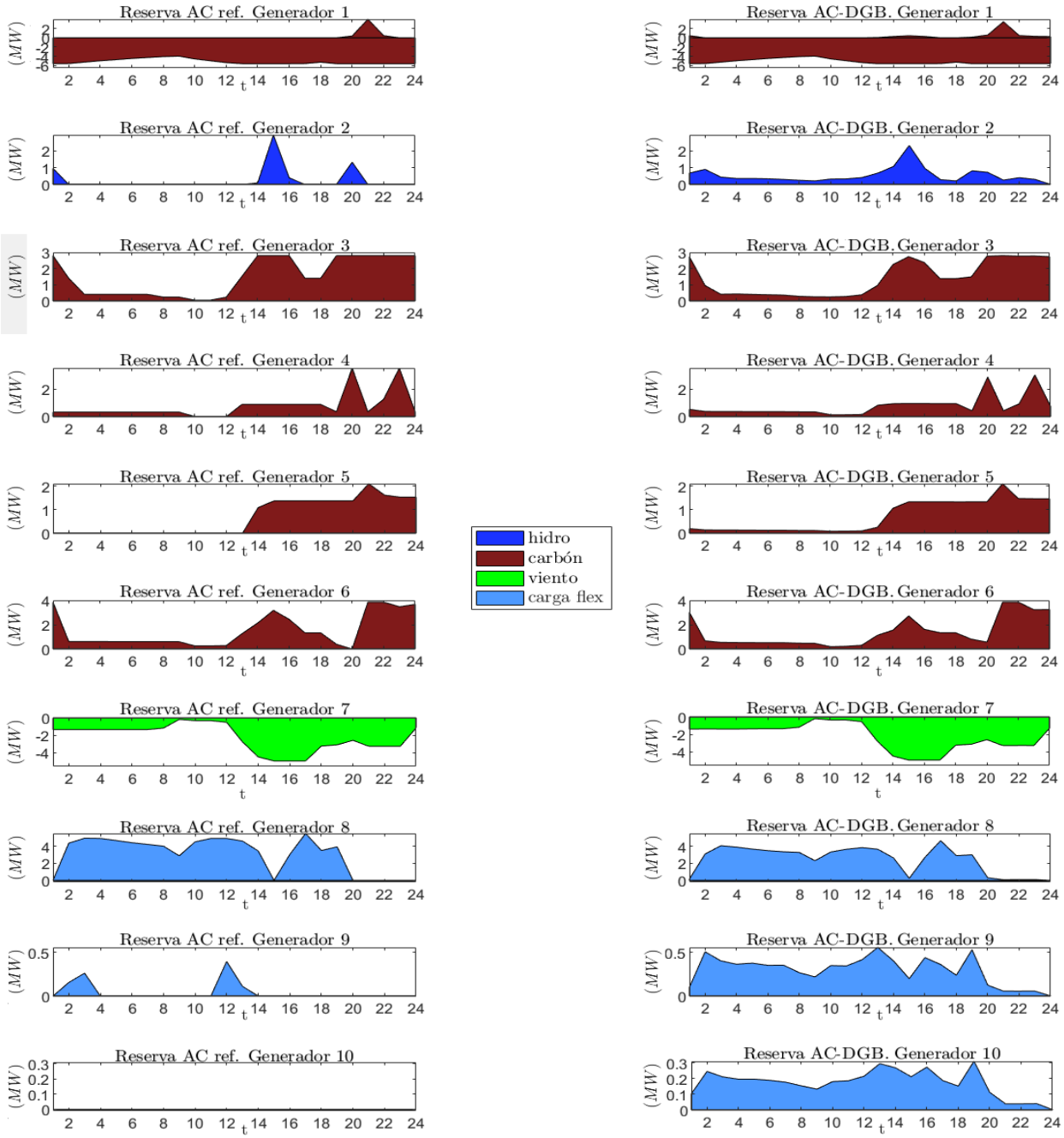

Figura 6-7.: Reserva de contingencia por unidad, sistema de potencia IEEE de 30 barras. Comparación entre la solución AC de referencia y la solución AC-DGB

transmisión que conecta las barras 6 y 8 fue la que presentó los mayores valores en los multiplicadores, en un rango de 0,0660 USD/MVA a 1,0405 USD/MVA para flujos en casos base y en el rango de 2,1520 USD/MVA a 15,8019 USD/MVA para casos contingentes. Los demás multiplicadores de las restricciones de límites activas tuvieron valores inferiores a la unidad.

Tabla 6-2.: Identificación de líneas de transmisión operando contra sus límites de potencia en el sistema IEEE de 30 barras

ID Flujo	**ID Línea de transmisión**	
(t,j,k)	**AC de referencia**	**DGB**
21,1,0	6-8, 21-22	6-8, 21-22
21,2,0	6-8, 21-22	6-8, 21-22
21,3,0	6-8, 21-22	6-8, 21-22
21,4,0	6-8, 21-22	6-8, 21-22
21,5,0	6-8, 21-22	6-8, 21-22
21,6,0	6-8, 21-22	6-8, 21-22
21,1,1	6-8, 21-22, 15-23	6-8, 21-22, 15-23
21,2,1	6-8, 21-22, 15-23	6-8, 21-22, 15-23
21,3,1	6-8, 21-22, 15-23, 25-27	6-8, 21-22, 15-23, 25-27
21,4,1	6-8, 21-22, 15-23	6-8, 21-22, 15-23
21,5,1	6-8, 21-22, 15-23	6-8, 21-22, 15-23
21,6,1	6-8, 21-22, 15-23	6-8, 21-22, 15-23
21,4,2	—	6-8, 23-24, 25-27
21,5,2	—	6-8, 23-24, 25-27
21,4,3	6-8	6-8, 15-23, 23-24, 25-27
21,4,4	6-8	6-8, 21-22, 15-23, 25-27
21,5,4	6-8	6-8, 21-22, 15-23, 25-27
21,6,4	6-8	6-8, 23-24, 25-27
22,1,0	6-8	6-8
22,2,0	6-8	6-8
22,3,0	6-8	6-8
22,4,0	6-8	6-8
22,5,0	6-8	6-8
22,6,0	6-8	6-8
22,3,1	6-8	—
22,4,1	6-8	—
22,5,1	6-8	—
22,6,1	6-8, 21-22	6-8, 21-22
22,4,4	—	15-23
23,4,4	—	15-23
23,5,4	—	15-23

6.5.2. Solución por descomposición con medidas de aceleración del caso IEEE de 30 barras

Este apartado presenta los resultados de convergencia del algoritmo DGB complementado con las técnicas de aceleración descritas en la Sección 4.3.3, esto es, la solución en paralelo de los subproblemas individuales y la estabilización de DGB mediante un método de haz con región de confianza. La Figura **6-8** muestra la evolución de las cotas en el proceso iterativo de solución. Estas convergen a un valor cercano a cero, que correspondió a las penalizaciones por el exceso de potencia activa residual que fue igual a $5,86 \times 10^{-5} MW$. El número de iteraciones disminuye considerablemente, comparado con la solución por DGB sin técnicas de aceleración, pasando de 54 iteraciones a 15. El tiempo de cómputo también se redujo drásticamente, siendo el tiempo de solución por DGB estabilizado y con cómputo paralelo de 476,6495s (cerca de 15 veces menos). El tiempo promedio de cómputo de los OPF fue de 27,9874s, mientras que el tiempo de cómputo del problema maestro estuvo en el rango de 2,4993s a 4,1289s.

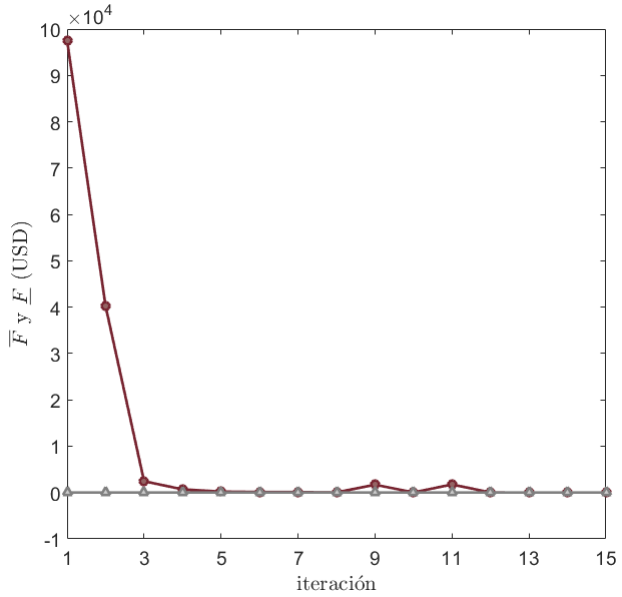

Figura 6-8.: Convergencia de DGB estabilizado y con cómputo paralelo de los subproblemas. Sistema de potencia IEEE de 30 barras

De otra parte, el costo óptimo del problema en esta prueba fue USD 9.551,4661, prácticamente igual al de la solución descompuesta sin la adopción de medidas de aceleración. Cabe mencionar que los valores por generador de cantidades como el despacho, las reservas de rampa y las reservas de contingencia guardan mucha similitud con las presentadas en la sección anterior, razón por la cual se omite su comparación aquí. Sobre la identificación de líneas operando contra sus límites, estos resultados fueron semejantes a lo expuesto en la

Tabla **C-1**, pero con reportes adicionales en la línea 15-23 para los flujos (22,2,4) y (23,2,4) en la solución por DGB con multiplicadores de valor insignificante.

6.5.3. Sumario de los resultados del caso IEEE de 30 barras

Las pruebas del algoritmo para DGB propuesto para la solución del problema de planeación de la operación del caso de prueba IEEE de 30 barras, partieron de una implementación multicorte sin medidas de aceleración. Los resultados conseguidos se validaron por comparación con los de una solución de referencia obtenida de resolver el problema de optimización sin descomposición.

Los costos óptimos de las dos soluciones se compararon encontrando que su diferencia fue mínima, al punto de la insignificancia, por lo tanto se puede decir que ambos problemas llegaron al mismo punto óptimo. El algoritmo DGB tomó casi dos horas para alcanzar la convergencia, tiempo que excedió en gran medida al tiempo de cómputo de la solución de referencia. Este se vio impactado de forma negativa por las oscilaciones en la cota superior que ralentizaron el proceso de convergencia. En las primeras iteraciones, el problema maestro no cuenta con mucha información sobre la factibilidad del despacho propuesto, considerando las restricciones de la red de transmisión, y se presentaron cambios abruptos en los despachos propuestos en algunos generadores en iteraciones sucesivas. Esto ocasionó que las variables penalizadas en los OPF respondieran para cerrar el balance de potencia, también con variaciones en algunos generadores, que se reflejaron en el costo de los subproblemas y por lo tanto en la cota superior. Al final, la solución óptima se alcanzó con valores muy pequeños en las variables de exceso de potencia activa ($6{,}67\times10^{-5}$), y con valores de las demás variables penalizadas iguales a cero, lo que es consistente con la formulación dada a los OPF.

El despacho esperado fue muy similar en las dos soluciones, salvo por las cargas flexibles en algunas horas. Sin embargo, todas las cargas flexibles cumplieron con su cuota de consumo en ambas soluciones de forma razonable al hacerlo en horas donde la demanda de potencia es baja. Por otra parte, la reserva de rampa de seguimiento de carga igualmente guardó similitudes en las soluciones de los generadores convencionales y algunas discrepancias en las asignaciones a las cargas flexibles. Las diferencias más destacadas surgieron de las asignaciones de reserva de contingencia en cada solución, principalmente para el generador 2 y las demandas flexibles. En general, se observó que las unidades con las diferencias más destacadas fueron aquellas que también se consideran como almacenadores, esto es, las unidades hidroeléctricas y las demandas flexibles. Estas tienen acoplamientos adicionales a los de los otros generadores, dados por las restricciones inter temporales del mecanismo de almacenamiento lo que pudo contribuir a que su asignación fuera diferente.

Los resultados de uno y otro algoritmo fueron consistentes en identificar aspectos propios de

la operación de este sistema de potencia, como la existencia de líneas operando contra sus límites de transmisión de potencia. Específicamente, todos los flujos correspondientes a los casos base en los que se presentó esa situación fueron reportados tanto por la solución de referencia como por la solución obtenida por DGB. Por su parte, en los casos contingentes se identificaron diferencias en los flujos de potencia a través de las líneas entre las dos soluciones, de tal manera que en algunos flujos la potencia transmitida disminuyó, con respecto a la solución de referencia, y en otros aumentó. Esto reveló la existencia de diferencias en los despachos para los casos contingentes, lo que a su vez afectó las cantidades de reserva de contingencia asignadas.

En conclusión, se puede considerar que la solución por descomposición mediante DGB es eficiente, dado que el costo óptimo y las cantidades asignadas de potencia y reserva a cada generador, son bastante aproximadas a las de la solución de referencia. Además, se evidenció consistencia en la identificación de aspectos que se pueden dar en la operación del sistema.

La segunda parte del experimento consistió en implementar medidas de aceleración en el algoritmo para DGB con el fin de reducir el tiempo de cálculo, el cual se redujo de forma significativa aunque no fue mejor que el tiempo de la solución de referencia. Posiblemente la competencia por el acceso simultáneo a la memoria principal por parte de los núcleos de procesamiento, operando en paralelo para resolver los OPF, ralentizó el proceso de cálculo impidiendo que se consiguiera mayor aceleración. Otra anotación sobre estos resultados es el efecto que tiene la dimensionalidad del problema maestro en los tiempos de solución. El tamaño del problema maestro creció con cada iteración, ya que se adicionaron tantos cortes como flujos de potencia se configuraron en el problema de optimización, y estos se acumularon a lo largo del proceso iterativo. Es por ello que, al disminuir el número de iteraciones, al final se procesó un problema maestro de dimensiones menores que el resuelto al final del procedimiento de solución por DGB sin medidas de aceleración, como lo demostró la reducción del rango de tiempo de solución del problema maestro en ambos casos.

6.6. Caso de prueba 2: Sistema de potencia colombiano de 96 barras

Algunos detalles del sistema de potencia colombiano se ofrecieron en el Capítulo 3 e información complementaria del mismo se recopila en el Anexo B. La Figura **6-9** corresponde al diagrama unifilar de una representación del sistema interconectado colombiano basado en la configuración de 2017 (Anexo XIII de [47]), con algunas modificaciones que consistieron en colapsar varias barras en una sola barra equivalente, colapsar varios generadores en un único generador equivalente y omitir las líneas de transmisión de las interconexiones internacionales con Ecuador y Venezuela. Los nombres de identificación de las barras del sistema se consignan en la Tabla **B-1**.

Este sistema de potencia de noventa y seis barras cuenta con cuarenta y nueve plantas de generación, doscientas seis líneas de transmisión y diecinueve unidades de almacenamiento de energía, que corresponden a generadores hidroeléctricos con embalse. Para facilitar el análisis detallado de los resultados, el sistema interconectado se dividió en cinco regiones o áreas operativas denominadas: Costa Atlántica, Antioquia, Sur, Centro y Oriente, acorde con [47] y con criterios adicionales de agrupación basados en la cercanía geográfica y eléctrica de las barras en la red.

6.6.1. Especificación de los datos de entrada

La información necesaria para planear la operación del sistema en un horizonte temporal de veinticuatro horas está dada por el perfil de demanda y los perfiles de generación eólica, ambos con resolución horaria, y una lista de posibles contingencias. La curva de demanda del sistema colombiano para un día típico se ha introducido previamente en el Capítulo 3, Figura **3-1**. La Tabla **B-2** en el Anexo B lista los valores horarios, tanto para potencia activa como reactiva. La información de carga horaria por barra es del año 2014 y fue reportada por XM. La información de tipo estocástico está representada por los escenarios de generación eólica y las contingencias. Esta información es presentada en detalle en el Anexo B.

Entradas estocásticas: escenarios de generación eólica y contingencias

La generación variable del caso colombiano es suministrada por una unidad de generación eólica ficticia, emplazada en la barra Copey, con una capacidad instalada de 800MW. Se consideraron quince escenarios de viento construidos a partir de medidas de velocidad de viento obtenidos de una base de datos histórica del IDEAM, correspondiente a los años 2006 a 2014. El procedimiento de transformación de estos datos fue explicado en la Sección 3.2

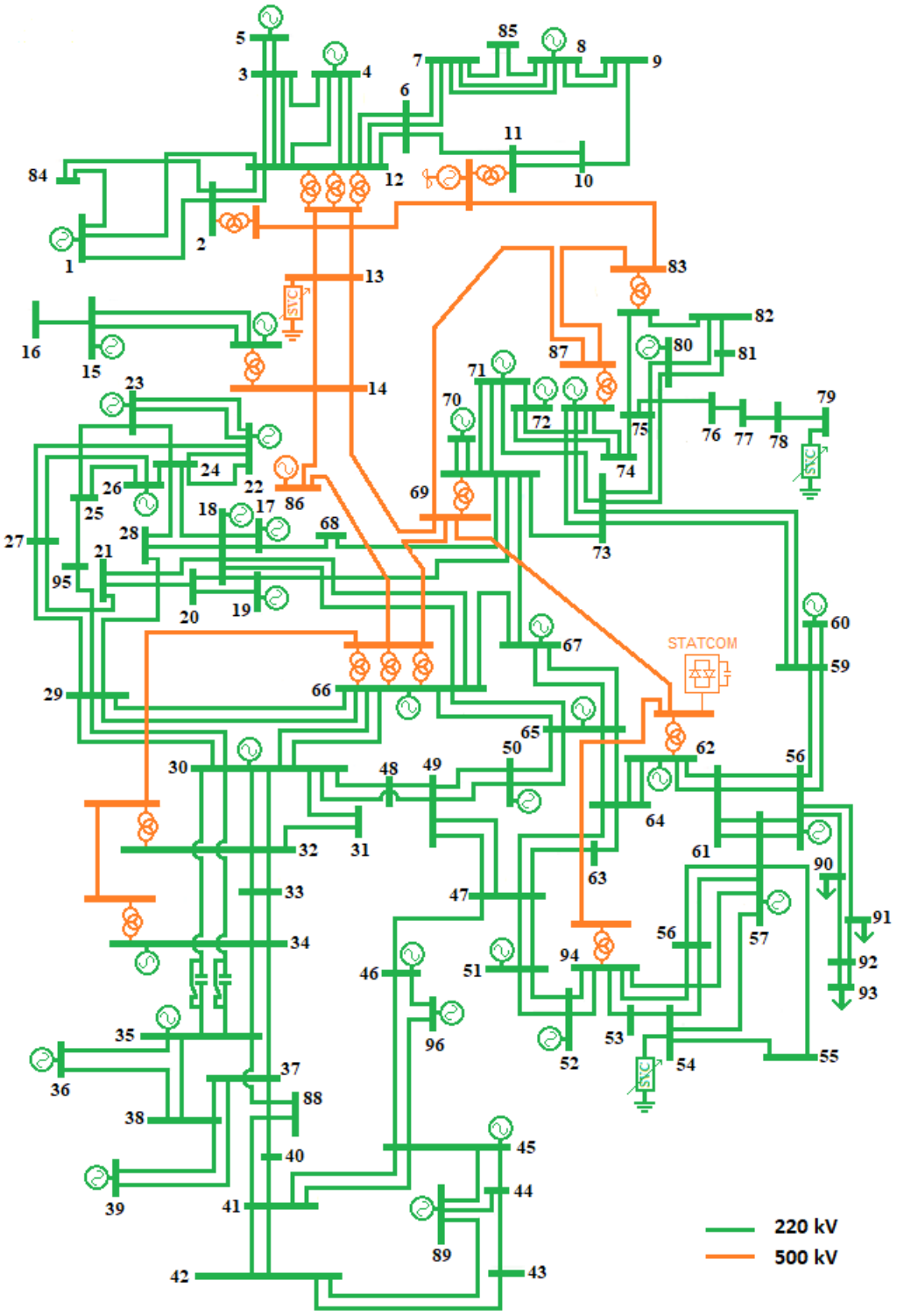

Figura 6-9.: Diagrama unifilar del sistema de potencia colombiano de 96 barras

Caso de estudio del Capítulo 3. La Tabla **B-5** del Anexo B contiene los datos de generación normalizados de la unidad eólica para los quince escenarios y veinticuatro periodos de tiempo.

Las diez contingencias sencillas (N-1) contempladas, todas con la misma probabilidad de ocurrencia de 1×10^{-5}, son relacionadas en la Tabla **6-3**. De estas, tres representan la deconexión de unidades de generación, disminuyendo la producción de potencia total de barra; seis, la desconexión de una línea de transmisión y una la salida de un transformador. La mayoría de las contingencias fueron tomadas de [47].

Tabla 6-3.: Lista de contingencias del tipo N-1 para el caso de prueba 2. Sistema de potencia colombiano

ID	Descripción
1	Desconexión de la línea Cerromatoso - Urrá, 220kV
2	Desconexión de la línea Primavera - Bacatá, 500kV
3	Desconexión de la línea Chivor - Guavio, 220kV
4	Desconexión de la línea Guaca - Mesa, 220kV
5	Desconexión de la línea Ocaña - La Loma, 500kV
6	Disminución del 75 % de la generación de TEBSA
7	Desconexión de la línea Sabanalarga - Chinú, 220kV
8	Salida del transformador en Altamira - Betania, 220kV
9	Disminución del 75 % de la generación de Guavio
10	Disminución del 75 % de la generación de Termo Flores

Dimensiones del problema de optimización con modelo AC

El problema de optimización con modelo AC, que proveerá la solución de referencia para la posterior validación de la solución por descomposición, consta de veinticuatro horas, quince escenarios de viento y once estados operativos por escenario (un caso base más diez contingencias), y da lugar al modelado de $24 \times 15 \times 11 = 3{,}960$ flujos de potencia individuales. El problema esta conformado por 1.871.854 variables y 3.948.090 restricciones, de las cuales 1.560.570 son lineales y 2.387.520 son no lineales.

6.6.2. Validación de la solución por DGB del sistema de potencia colombiano

En primer lugar, se obtuvo una solución del problema de optimización sin descomposición que será la referencia para validar los resultados de la solución por descomposición. Por otro lado, el algoritmo de descomposición está basado en la versión multicorte de DGB y no hace uso de ninguna de las estrategias de aceleración propuestas, en consecuencia, en esta sección

los subproblemas se resuelven en serie.

El tamaño del problema maestro y de los subproblemas generados por el método DGB para el caso colombiano, se resume en la Tabla **6-4**. El problema maestro lineal retiene el 51 % de las variables y el 60 % de las restricciones del problema original. El número de restricciones se incrementará en cada iteración de manera proporcional al número de OPFs. De otra parte, el subproblema quedó conformado por 3.960 OPFs no lineales y totalmente independientes.

Tabla 6-4.: Tamaño del problema maestro y los subproblemas del método DGB para el caso de prueba 2. Sistema de potencia colombiano

Problema maestro		**Subproblema**	
		Número de OPF	3.960
Variables	921.454	Variables	**2.475.000**
		OPF	954.360
		Penalización	1.520.640
Restricciones	1.560.570	Restricciones	**2.585.880**
	+cortes	lineales	194.040
		no lineales	2.387.520

En cuanto a los resultados, el costo óptimo de la función objetivo de la solución de referencia fue KCOP 21.434.540 y de la solución con descomposición fue de KCOP 21.435.516. La diferencia porcentual entre ambos resultados fue del 0,0045 %, lo que se puede considerar aceptable. Por su parte, el tiempo total de cálculo de la solución de referencia fue de 172.503s (47h 55min 4s), mientras que la solución del problema descompuesto tomó 64.761s (17h 59min 21s). Para este último, el tiempo de solución del problema maestro estuvo en el rango de 293,91s a 3.342,11s, y el tiempo promedio de solución de los subproblemas en una iteración fue de 373,05s. El tiempo de solución del problema maestro se incrementa conforme crece el tamaño del problema al acumular múltiples cortes (restricciones) por iteración. De acuerdo con estas observaciones preliminares, la descomposición *per se* disminuyó cerca de tres veces el tiempo de solución.

Luego, se compararon otras cantidades de interés por área operativa: potencia activa despachada y reservas de contingencia y de rampa de seguimiento de carga. En la Figura **6-10** se observa que las diferencias en la potencia activa despachada por área son muy pequeñas. Sin embargo, a nivel de cada unidad de generación individual las diferencias más destacables se dieron en la primera hora del día. Comparado con la solución de referencia, unidades como Chivor y Sogamoso generaron más potencia, 35MW y 21MW respectivamente, que en valores porcentuales equivalen respectivamente a un 7,7 % y un 100 % más que en la solución de referencia. Ese despacho de Sogamoso en la primera hora es ligeramente apreciable en

la gráfica del área Oriente. En cambio, generadores como Porce 3 y Guavio tuvieron menor producción de potencia en la solución descompuesta con 46MW y 12MW respectivamente, representando un 14 % y un 1,7 % menos que la potencia producida a esa hora en la solución de referencia. En otras horas se observaron efectos complementarios entre unidades de generación dentro de la misma área operativa, como en el caso de Termo Barranquilla con Cartagena y Candelaria, y, de forma parcial, en el caso de Chivor y Guavio.

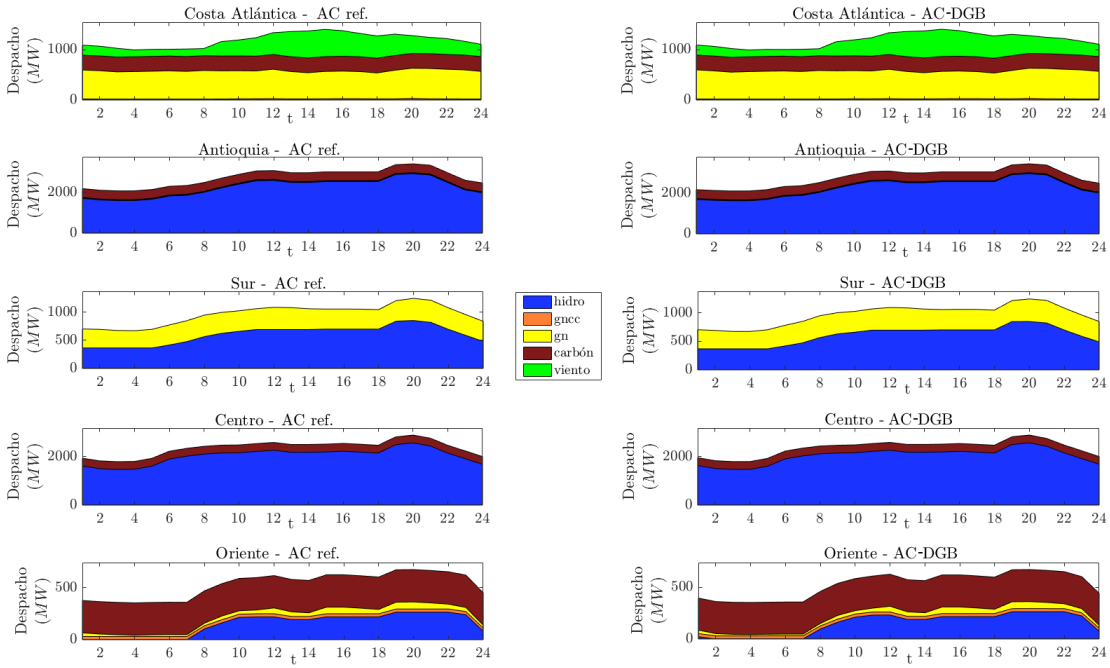

Figura 6-10.: Despachos horarios por área, sistema de potencia colombiano. Comparación entre la solución AC de referencia y la solución AC-DGB

Por su parte, las reservas de rampa de seguimiento de carga presentan algunas diferencias en todas las áreas, sobre todo en la asignación de las unidades hidroeléctricas en Antioquia en t igual a 1 y 22 con asignaciones menores en la solución descompuesta; y en Oriente para el generador Sogamoso en t igual a 1 y 12 con reservas de rampa mayores en la solución descompuesta y en t igual a 9 con asignación de reserva de rampa menor. En esta área también se observaron diferencias en la asignación al generador Barranca en varias horas del periodo de planeación (Figura **6-11**).

En lo que se refiere a las reservas de contingencias, presentadas en la Figura **6-12**, existen diferencias apreciables en las cantidades de reserva para unidades de generación de gas natural en algunas horas, especialmente en las áreas Sur y Oriente; y en las cantidades asignadas a las unidades hidroeléctricas, de reserva hacia arriba en Antioquia y Oriente, y de reserva hacia abajo en el área Centro. Entre estas últimas, las máximas diferencias en la reserva

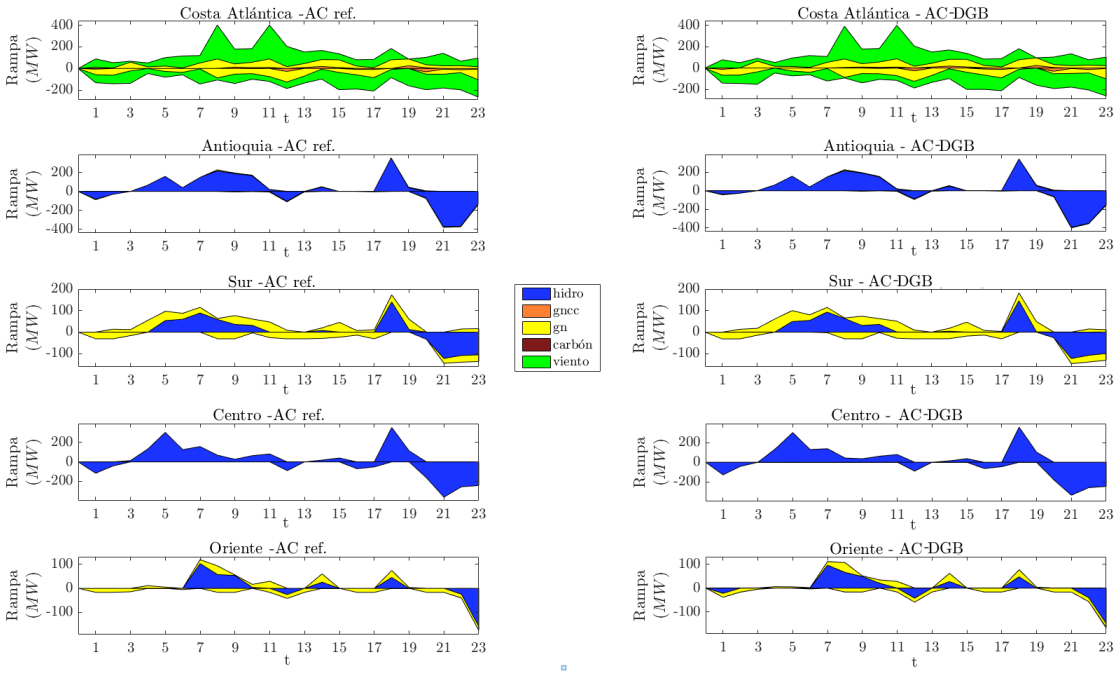

Figura 6-11.: Reservas de rampa por área, sistema de potencia colombiano. Comparación entre la solución AC de referencia y la solución AC-DGB

de contingencia hacia arriba se presentaron en unidades de generación como San Carlos, ubicada en el área de Antioquia, con 31,69MW menos en la hora 1 que equivale al 6 % de la asignación dada a ese generador por la solución de referencia; y el generador Chivor, ubicado en el área Centro, con 30,56MW menos en la hora 12 equivalente al 25 % de la cantidad programada por la solución de referencia. En cuanto a las reservas de contingencia hacia abajo, las mayores diferencias se presentaron en el generador Guavio, ubicado en el área Centro, con 23,88MW más en la hora 6 que equivale al 6 % de la asignación de reserva hacia abajo en la solución de referencia. Un resultado particular se observó en Oriente con la asignación en la solución por descomposición de una cantidad de reserva de contingencia hacia arriba para la unidad hidroeléctrica en Sogamoso, con un máximo de 7,6MW en la hora 20, mientras que en la solución de referencia no se asignó reserva de contingencia a esa unidad de generación. En cuanto a las unidades térmicas, las mayores diferencias en las cantidades de reserva hacia abajo se dieron en Termo Barranquilla en la hora 2 con 15,15MW que corresponde a un 22 % menos del valor proporcionado por la solución de referencia; y Termo Emcali en la hora 2 con 13,69MW, para un 50 % adicional a la asignación dada por la solución de referencia.

Por otra parte, no se detectaron restricciones activas en los límites de transmisión en ninguno de los flujos de potencia, tanto en casos base como contingentes, en ninguna de las dos soluciones.

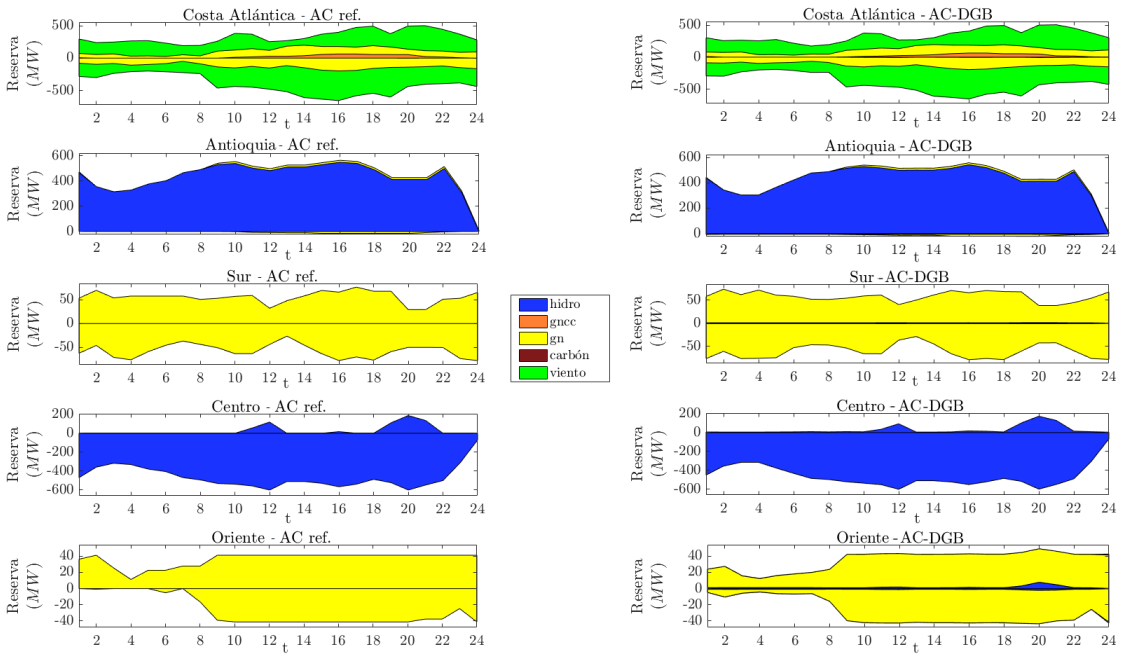

Figura 6-12.: Reserva de contingencia por área, sistema de potencia colombiano. Comparación entre la solución AC de referencia y la solución AC-DGB

La solución por descomposición del problema de planeación es bastante aproximada a la solución de referencia, puesto que la diferencia entre los costos óptimos de ambas soluciones fue despreciable y puede considerarse como suficiente para aplicaciones en las que el tiempo de cálculo sea un factor determinante. Además, resultados como los obtenidos para las reservas de rampa y las reservas de contingencias mostraron que el problema de optimización resultó ser muy plano, es decir, es posible hacer cambios significativos en los despachos al rededor de la solución óptima sin que esto tenga un impacto grande en el costo óptimo del problema. En la formulación de MPSSOPF se usan muchas variables auxiliares para crear los conjuntos de desigualdades que definen las reservas y puede darse el caso que en una solución muchas de esas variables, que no tienen un costo asociado, no estén activas (sean libres) y se pueden mover sin afectar el costo óptimo del problema pero si la definición de las rampas o las reservas de contingencia. Por otra parte, existen variables que teniendo un costo se pueden mover sin afectar mucho el costo total, como los despachos en las contingencias, dado que esos costos multiplicados por la probabilidad de las contingencias no resultan significativos.

6.6.3. Convergencia de la solución por DGB

La Figura **6-13** muestra la evolución de las cotas inferior y superior en el proceso iterativo de solución para DGB sin técnicas de aceleración. En la gráfica se omite la primera iteración ya que el valor de la cota superior en ese punto dista tanto del valor en la segunda iteración que no se logra visualizar su comportamiento en las iteraciones sucesivas. El algoritmo convergió

a la solución óptima en veinticinco iteraciones. Sobre la evolución de las cotas, se puede observar que la cota inferior permanece cercana a cero, mientras que la cota superior es decreciente y se aproxima progresivamente y sin oscilaciones a la cota inferior.

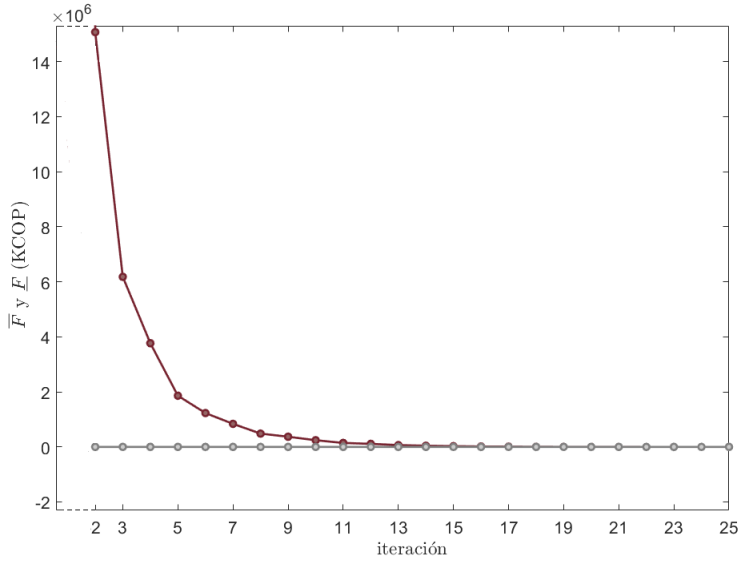

Figura 6-13.: Convergencia del método de DGB sin medidas de aceleración para el sistema de potencia colombiano

Por su parte, el costo óptimo de la función objetivo en cada iteración, se grafica en la Figura **6-14**, y sirve para evaluar el segundo criterio de convergencia basado en el cambio absoluto de este costo óptimo entre dos iteraciones sucesivas. El costo óptimo del problema es creciente aproximándose al valor final sin oscilaciones.

6.6.4. Efecto del procesamiento paralelo de los subproblemas en el tiempo de solución

Esta simulación tiene como propósito estudiar el impacto que tiene el procesamiento en paralelo de los subproblemas sobre el tiempo total de cálculo del problema de optimización con DGB. El énfasis del análisis se hará sobre el tiempo de solución y no en los resultados de despacho de potencia y reservas, dado que son iguales a los obtenidos en la sección anterior.

La Figura **6-15** presenta los tiempos de ejecución del algoritmo DGB al aumentar el número de núcleos en el computador multinúcleo. El mejor tiempo de procesamiento, conseguido con ocho núcleos, fue de 11.521,93s. En general, el tiempo de cómputo es decreciente al aumentar el número de núcleos resolviendo subproblemas en paralelo, aunque el decrecimiento no es proporcional al número de núcleos empleados. La mayor tasa de descenso fue de 4.000s

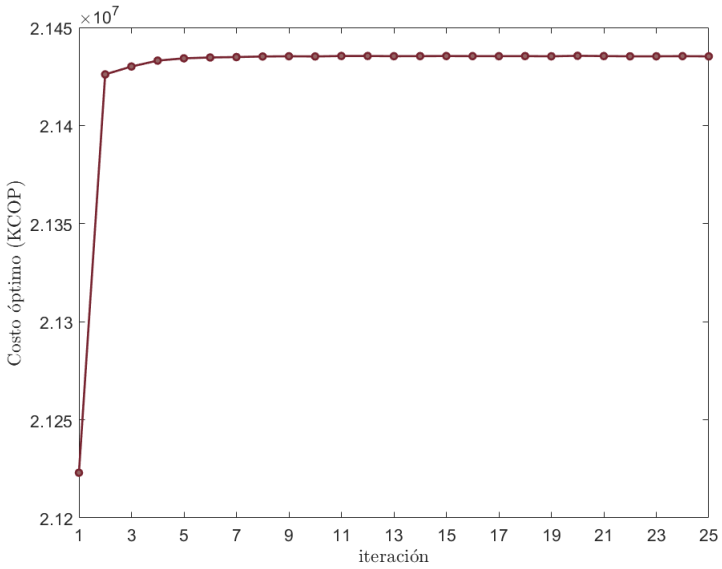

Figura 6-14.: Evolución del costo óptimo en la solución por DGB sin medidas de aceleración para el sistema de potencia colombiano

aproximadamente y se logró cuando se usaron tres núcleos de procesamiento. A partir de ese punto, el rango de reducción del tiempo de cómputo no mejoró como se esperaría al usar más núcleos de procesamiento, incluso en una de las pruebas el tiempo aumentó al usar cinco núcleos. Algunos factores pueden influir para que la reducción en el tiempo de solución sea deficiente, principalmente la competencia por el acceso a memoria en procesadores con arquitectura de memoria compartida y, en menor medida, a la sobrecarga producida por la sincronización y comunicación entre el núcleo principal y los esclavos.

La Figura **6-16** muestra la aceleración y eficiencia por número de procesadores, calculados con las ecuaciones (6-4) y (6-5). El cambio en el valor de la aceleración con el número de procesadores adicionados para el cálculo en paralelo de los subproblemas es poco, esta cambia de 2,770 a 3,207 variando de 3 a 8 núcleos, y en ese caso se puede decir que la aceleración permanece casi constante. Esto es un indicio de que el principal obstáculo para conseguir mayores aceleraciones al incrementar el número de núcleos es la competencia por el acceso a memoria física común en procesadores de memoria compartida. El drástico aumento en el tráfico de memoria al agregar más procesadores, impone una limitante en la escalabilidad del algoritmo, implementado en un computador con procesador multinúcleo, para incrementar la aceleración con la adición de más recursos. De forma análoga, la eficiencia calculada resultó afín a lo ya mencionado y disminuye al agregar núcleos de procesamiento.

La Figura **6-17** corresponde a la tabla de procesos del `monitor de recursos`, herramienta de `Windows` que muestra información detallada acerca del uso de memoria. Este enumera todos los procesos actualmente en ejecución y divide la cantidad de memoria usada por cada

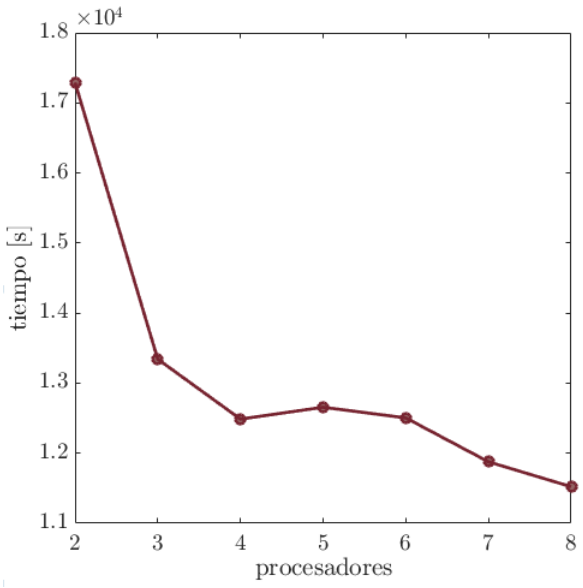

Figura 6-15.: Tiempo total de cómputo *vs* cantidad de núcleos de procesamiento. Algoritmo DGB con cálculo paralelo de los subproblemas. Sistema de potencia colombiano

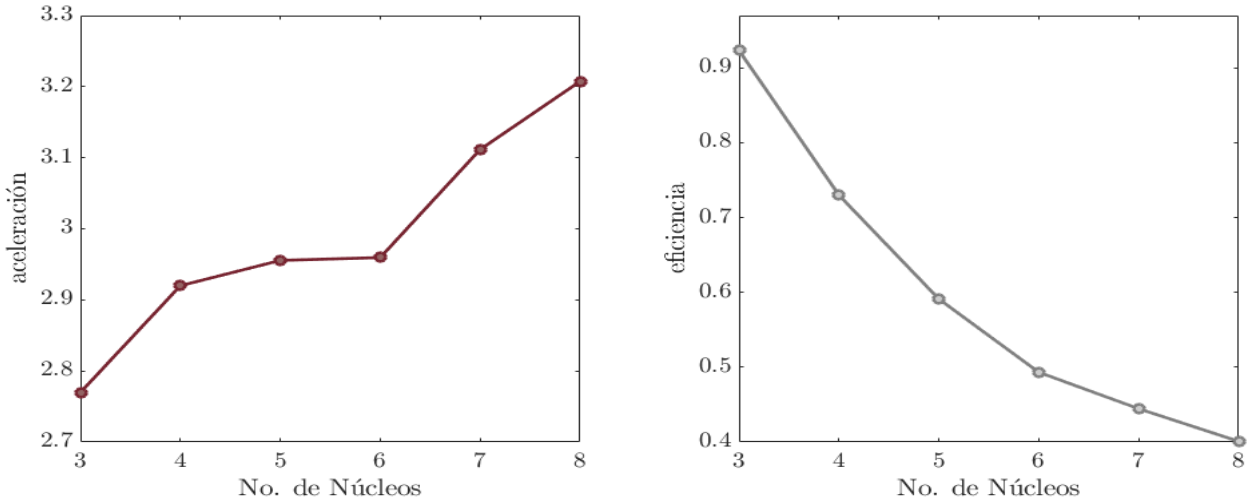

Figura 6-16.: Aceleración y eficiencia calculados para el algoritmo de DGB con cómputo paralelo de los subproblemas, variando la cantidad de núcleos de procesamiento. Sistema de potencia colombiano

proceso en varias categorias. Las columnas de interés en la tabla son: la columna *Image*, identifica el nombre del archivo ejecutable del proceso; la columna *Commit*, muestra la cantidad de memoria que el sistema operativo ha reservado para el proceso; la columna *Working Set*, muestra la cantidad de memoria física usada actualmente por el proceso, la que adicionalmente se divide en dos categorías de memoria, una es la cantidad de memoria física que

es compartida con otros procesos (columna *Shareable*) y la otra es la cantidad de memoria que no es compartida con otros procesos (columna *Private*). Esta última brinda una medida bastante precisa de la cantidad de memoria que se necesita para ejecutar una aplicación.

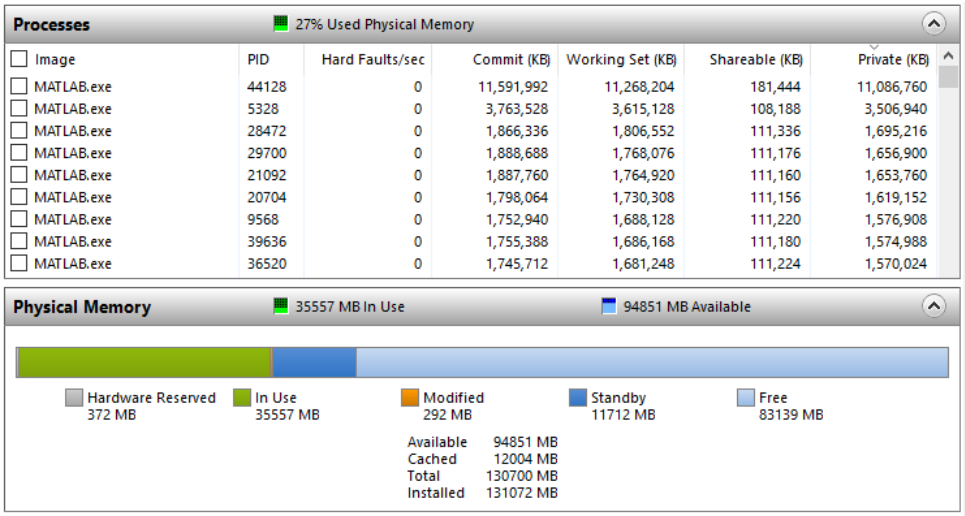

Figura 6-17.: Información detallada del uso de la memoria, dividida por categorias, en el cómputo paralelo de los OPFs para DGB. Sistema de potencia colombiano

En la gráfica se observan nueve procesos con el nombre `MATLAB.exe`, el primero corresponde a la sesión cliente de `MATLAB` y los ocho restantes a cada uno de los trabajadores en el ambiente de cómputo paralelo, uno por cada núcleo del procesador. Una inspección al uso de la memoria del computador multinúcleo durante el proceso de cálculo de los OPF en paralelo, reveló que efectivamente la memoria total requerida, aproximadamente 25,94GB, excede por mucho a la capacidad de memoria caché del procesador (16MB), que es más rápida. Esto estaría forzando a que los núcleos del procesador deban acceder de forma simultánea a la memoria principal, que es más lenta y por consiguiente, todo el proceso se ralentiza.

6.6.5. Efecto de la estabilización con región de confianza

Ahora, se pone en práctica el Algoritmo 5.4 para el método de DGB con estabilización para conseguir una mejora adicional en el rendimiento. La selección de un punto de estabilización y la definición de una región de confianza en cada iteración, pueden modificar la calidad de la solución del problema. En consecuencia, además de analizar los tiempos de solución, se validará el costo óptimo, los despachos de potencia activa, las reservas de contingencia y de rampa de seguimiento de carga, y potencialmente algunas restricciones en transmisión, si es del caso.

La evolución de las cotas superior e inferior, usando el número máximo de núcleos de procesamiento se muestra en la Figura **6-18**. La progresión de las cotas a lo largo de las iteraciones es igual al de la Figura **6-13**, y además se observa una reducción en el número de iteraciones, pasando de veinticinco para DGB a trece para DGB estabilizado.

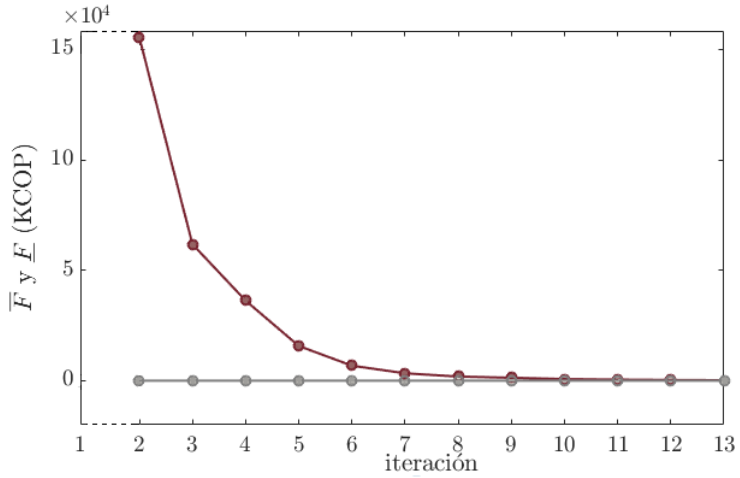

Figura 6-18.: Convergencia del método DGB estabilizado con región de confianza para una solución del sistema de potencia colombiano

La Figura **6-19** esquematiza los tiempos de solución variando el número de núcleos de procesamiento. El mejor tiempo conseguido con el máximo número de núcleos fue de 8.905,87s, menor en un 22,7 % con respecto al tiempo registrado en el experimento previo con el mismo número de núcleos, que fue de 11.521,9300. Al igual que en el experimento anterior, al aumentar el número de núcleos de procesamiento la aceleración aumenta, aunque no de forma importante si se tiene en cuenta la cantidad de núcleos, y la eficiencia disminuye como se evidencia en la Figura **6-20**. En este caso, la aceleración calculada aunque fue mayor a la del experimento anterior no alcanza los valores esperados.

Aunque los resultados parecieran sugerir que la aplicación de un método de estabilización contribuye en la aceleración de la solución del problema descompuesto, el incremento conseguido en la aceleración del método de DGB estabilizado con respecto al experimento anterior, resulta ser es un producto indirecto del efecto que la dimensionalidad del problema maestro tiene en el tiempo general de cómputo. Como ya se ha mencionado anteriormente, el tamaño del problema maestro crece con cada iteración debido al número de cortes acumulados. Entonces, cuando el tamaño del problema maestro es más pequeño, como en las primera iteraciones, se necesita menos memoria y el uso de la memoria caché es más eficiente, y es a través de ese mecanismo que se obtiene mejor aceleración. De no ser porque el tamaño del problema cambia con la estabilización, al requerirse menos iteraciones, no habría mayor

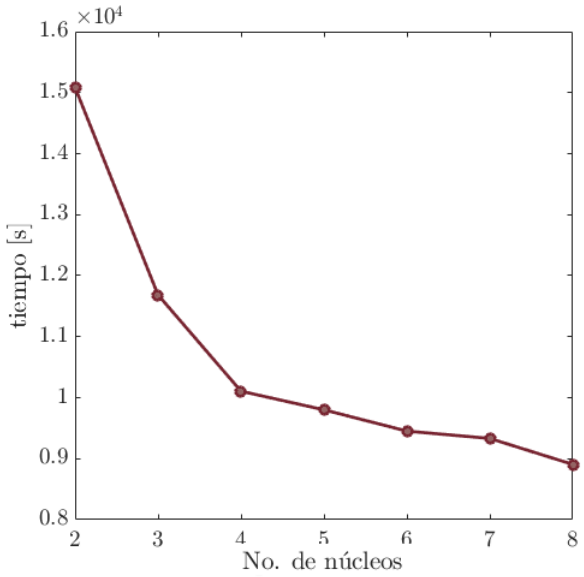

Figura 6-19.: Tiempo total de cómputo *vs* cantidad de núcleos de procesamiento. Algoritmo DGB estabilizado con cálculo paralelo de los subproblemas. Sistema de potencia colombiano

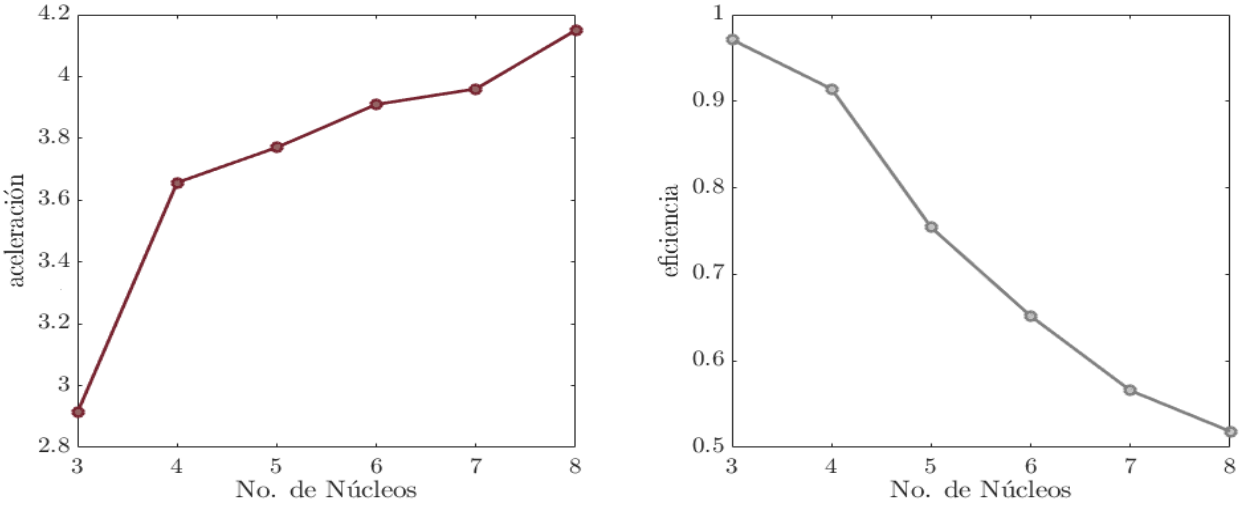

Figura 6-20.: Aceleración y eficiencia calculados para el algoritmo DGB estabilizado con cómputo paralelo de los subproblemas, variando la cantidad de núcleos de procesamiento. Sistema de potencia colombiano

aceleración.

Por otra parte, la inspección de las gráfica para los despachos horarios y las reservas de rampa de seguimiento de carga y de contingencias mostró que estas cantidades son muy similares a lo hallado en el apartado 6.6.2. La inspección numérica de los resultados permitió constatar

que las diferencias entre la solución por DGB con estabilización y la solución de referencia son, en términos generales, muy parecidos a los obtenidos con DGB sin estabilización. Por ejemplo, en el despacho esperado las máximas diferencias también se dieron en la hora 1. Los generadores Chivor y Sogamoso aumentaron su generación en 34,35MW (7,46 %) y en 24,44MW (100 %), respectivamente; de forma similar, generadores como Porce 3 y Guavio generaron menos potencia con 49,17MW (14,76 %) y 10,11MW (1,34 %), respectivamente. Esos valores fueron muy cercanos a los encontrados en la solución por DGB sin estabilización. Con respecto a las reservas de contingencia y de rampa de seguimiento de carga, se encontraron hallazgos análogos a lo ya mencionado al comparar los resultados con y sin estabilización. Al igual que en la solución por DGB sin medidas de aceleración, no se detectaron restricciones activas en los límites de transmisión en ninguno de los flujos de potencia, tanto en casos base como contingentes.

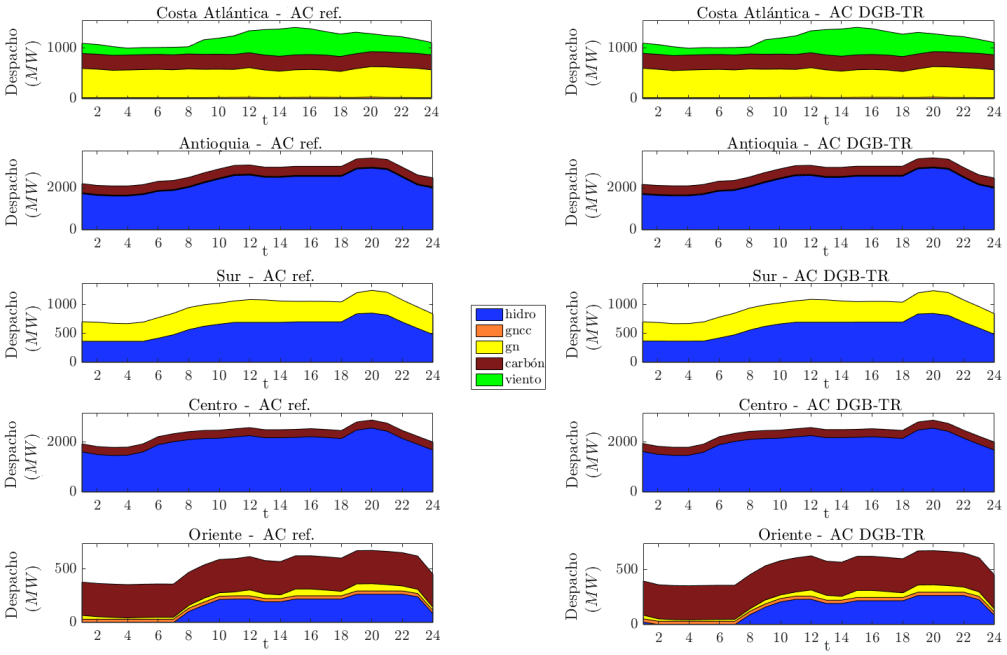

Figura 6-21.: Despacho horario por área, sistema de potencia colombiano. Comparación entre la solución AC de referencia y la solución AC-DGB estabilizado

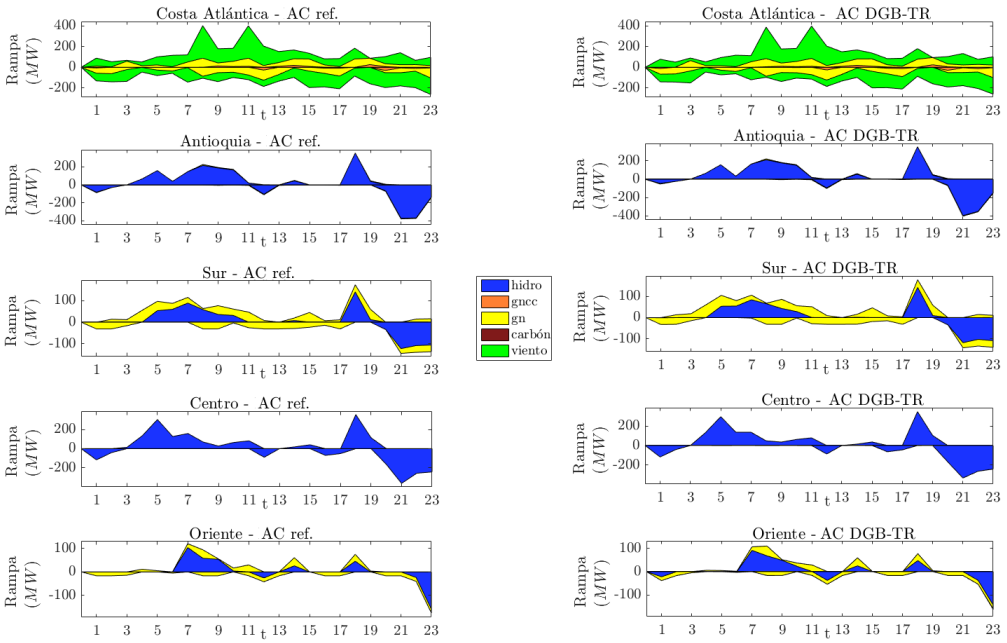

Figura 6-22.: Reservas de rampa por área, sistema de potencia colombiano. Comparación entre la solución AC de referencia y la solución AC-DGB estabilizado

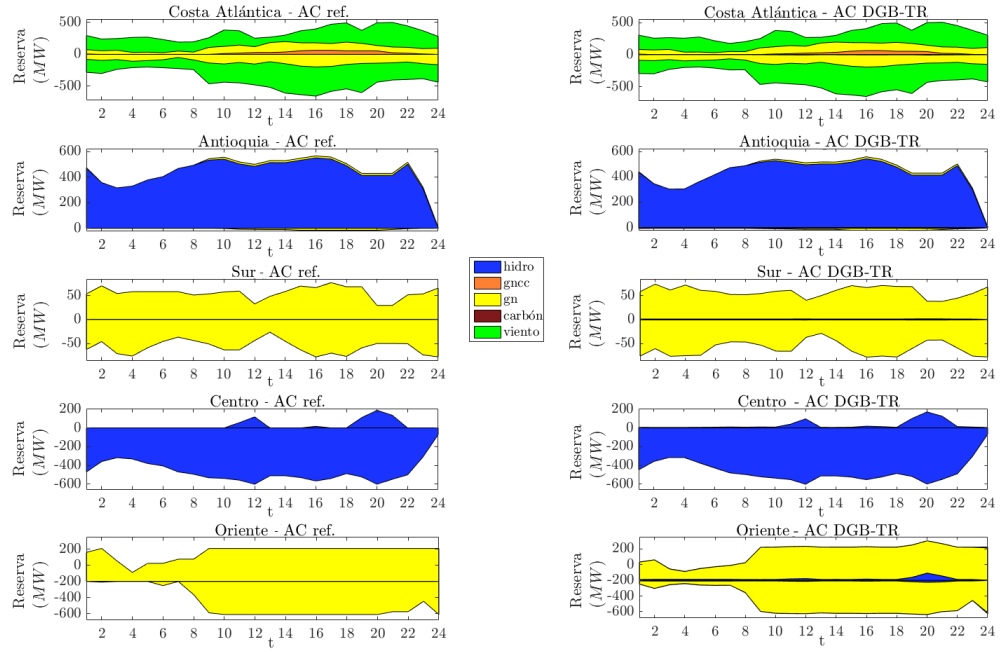

Figura 6-23.: Reserva de contingencia por área, sistema de potencia colombiano. Comparación entre la solución AC de referencia y la solución AC-DGB estabilizado

6.6.6. Medidas de aceleración adicionales

La inspección realizada sobre las variables de la solución por descomposición permitió diseñar dos mecanismos de aceleración adicionales. El primero busca reducir el tamaño del problema maestro al controlar los cortes adicionados por iteración, a partir de observaciones realizadas en las variables penalizadas por OPF. El segundo pretende disminuir el tiempo de solución del subproblema al evitar resolver OPF que no generan cortes de Benders, de acuerdo con los multiplicadores de las restricciones de balance de potencia, evaluados, iteración tras iteración, en una ventana de observación establecida. Ambos mecanismos se explican a continuación.

Evaluación de las variables penalizadas por OPF

Para el problema particular de programación de la operación del sistema de potencia colombiano, el valor de las variables penalizadas de exceso de potencia activa y las variables de déficit y exceso de potencia reactiva fue cero en todas las iteraciones, a diferencia de la variable de déficit de potencia activa que disminuye en cada iteración. En la Figura **6-24** cada punto representa la sumatoria del déficit de potencia activa en todas las barras de cada subproblema en la iteración 15 de la solución por descomposición sin medidas de aceleración, tomada como ejemplo. La suma de los déficit de potencia activa en todos los OPF's fue de 2,73MW con un costo total de KCOP 27.324, cerca de 0,12 % del costo óptimo de la función objetivo en esa iteración. Adicionalmente, se observó que de los 3.960 flujos óptimos, 3.599 incurrieron en un déficit total de potencia activa por flujo menor o igual a 0,001MW con un impacto insignificante en el costo óptimo del problema. Los 361 flujos óptimos restantes presentaron un déficit de potencia activa superior o igual a este umbral.

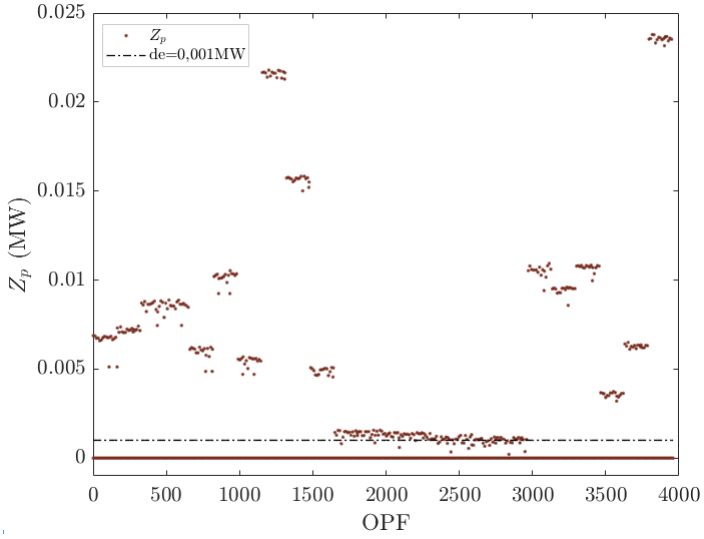

Figura 6-24.: Déficit total de potencia activa por OPF para una iteración del algoritmo de DGB, sistema de potencia colombiano

Dado que el efecto de esas cantidades residuales sobre el costo óptimo de la función objetivo es muy pequeño, se puede justificar la introducción de un mecanismo que prevenga la formación de un corte de optimalidad cuando el déficit total de potencia activa iguale o supere un umbral preestablecido. De esta forma, se reduce el número de restricciones en el problema maestro sin afectar significativamente la delimitación del espacio de búsqueda de la solución. Por ejemplo, si en una iteración del algoritmo de DGB se encuentra algo similar a lo expuesto por la Figura **6-24**, y con un umbral establecido en el déficit de potencia activa acumulado por OPF igual a 0,001MW, en la siguiente iteración solo se agreagrían 361 cortes al problema maestro en lugar de 3.960 cortes. En este caso particular, el umbral admisible para que el déficit total de potencia activa de un OPF amerite la adición de un corte de optimalidad al problema maestro se definió en 0,001MW, de acuerdo con lo observado experimentalmente.

Discusión sobre la generación de variables duales cero en las ecuaciones de balance de potencia

Se observó que los multiplicadores de las ecuaciones de balance de potencia generados por algunos OPF son cero a partir de cierta iteración, en su mayoría para estados operativos contingentes ($k \neq 0$). Un caso particular se da con los estados operativos correspondientes a la disminución del 75 % de la generación de Guavio, en todos los periodos de tiempo y en las hora con menor producción de generación eólica de ciertos escenarios, que requirieron la adición de cortes casi hasta el final del proceso iterativo. La Figura **6-25** representa a través de puntos los OPF en los que se generó un corte por iteración, para cuatro horas diferentes del horizonte de planeación.

La reiterada producción de multiplicadores iguales a cero en las mencionadas restricciones en algunos OPFs, indica que para esos subproblemas ya se había encontrado una solución óptima en alguna iteración previa. En consecuencia, se abre la posibilidad de conseguir una disminución adicional del tiempo global de cómputo si se dejan de resolver los flujos de potencia para los cuales se contó un número determinado de iteraciones con multiplicadores iguales a cero en las restricciones de balance de potencia.

Entonces, la simulación anterior se repitió agregando: i) una ventana de observación en las iteraciones para determinar cuáles subproblemas dejan de aportar cortes de optimalidad durante un número determinado de iteraciones, y en tal caso dejar de resolver esos subproblemas en iteraciones posteriores; y ii) un umbral en el déficit de potencia activa igual a 0,001MW por OPF, para decidir cuando generar un corte de Benders para un flujo en particular. El tiempo total de solución empleando los ocho núcleos de procesamiento y una ventana de observación de cinco iteraciones, valor definido de forma experimental, fue de 7.813,28s que, en comparación con los 8.905,87s de la simulación inicial, representa una reducción adicional en el tiempo de cómputo de aproximadamente 12,2 %.

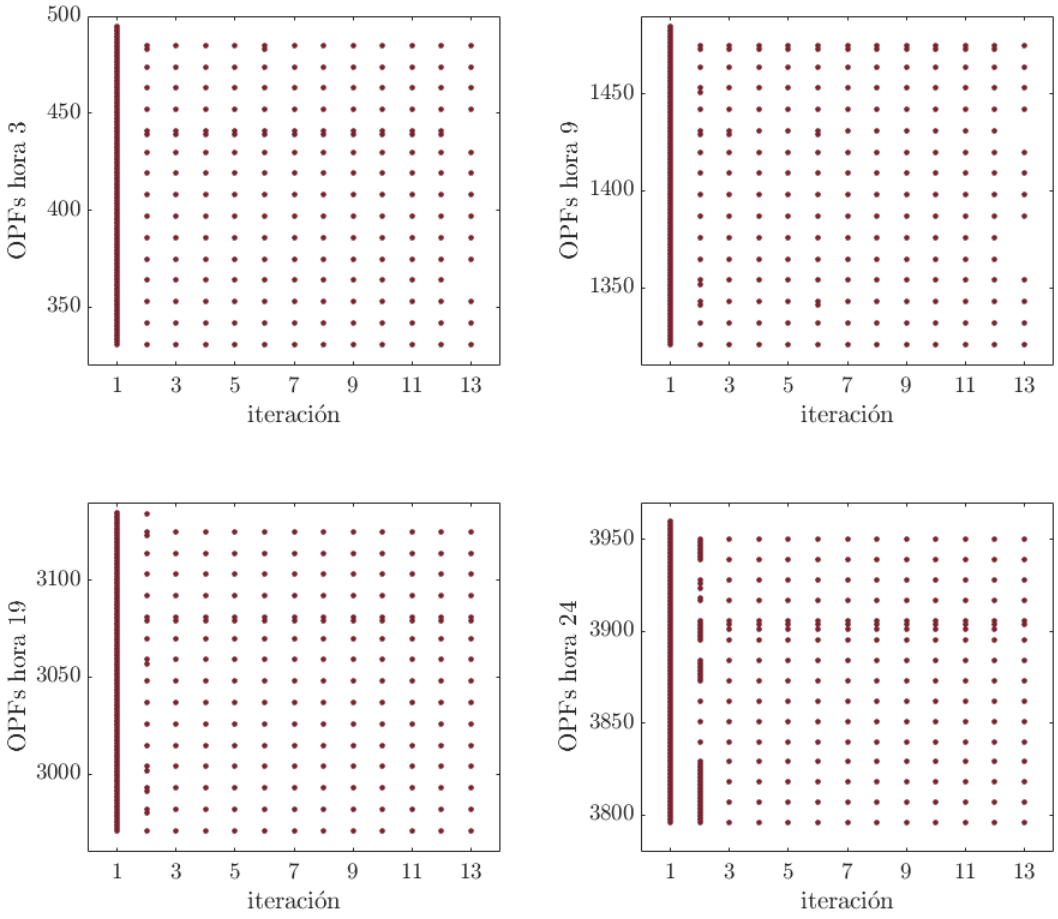

Figura 6-25.: Ejemplo de identificación de OPFs generando cortes por cada iteración. Sistema de potencia colombiano

6.6.7. Sumario de los resultados del caso colombiano

Los experimentos de esta sección tuvieron como propósito validar el funcionamiento del algoritmo DGB para resolver el problema bajo estudio usando como caso de prueba un sistema de potencia de tamaño real, como el sistema de potencia colombiano. Después, diferentes estrategias de aceleración se aplicaron progresivamente para potenciar el desempeño del algoritmo DGB, tales como el cómputo paralelo y la estabilización por región de confianza.

El problema de planeación formulado como se indica en el Capítulo 2, puede considerarse como de grandes dimensiones por el número de variables y restricciones que lo conforman (aproximadamente dos millones de variables y cuatro millones de restricciones). Bajo la perspectiva de la descomposición este problema se dividió en dos problemas de menor tamaño que el problema original. En el caso del problema maestro, este quedó constituido por un millón de variables e, inicialmente, con un millón y medio de restricciones, dimensión que fue creciendo con los cortes acumulados por iteración. Este problema aunque pequeño com-

parado con el problema original sigue siendo retador. Por su parte, el subproblema estuvo conformado por 3.960 OPFs independientes, en el que cada OPF es de menores dimensiones.

Un primer experimento, sin incluir ningún mecanismo de aceleración, demostró que solo con utilizar una técnica de descomposición se logró reducir cerca de una tercera parte el tiempo necesario para obtener una solución óptima. Al igual que en el caso de pequeña dimensión, los costos óptimos fueron muy aproximados, con una diferencia inferior al 0,005 % y las discrepancias en las cantidades asignadas de algunos productos tuvieron comportamientos variados por área y por tecnología de generación. Estas discrepancias mostraron que el problema de optimización es demasiado plano, de tal forma que los cambios significativos en algunos despachos, por ejemplo en una contingencia, en torno al punto de solución óptimo no tuvieron un efecto significativo en el costo óptimo del problema.

Un segundo experimento introdujo el procesamiento paralelo de los subproblemas. Se determinó el rango de aceleración y de eficiencia para la plataforma de cómputo paralelo ante una variación del número de núcleos de procesamiento. El rango de aceleración aumentó muy poco al incrementar la cantidad de núcleos utilizados, mientras que la eficiencia tuvo un rango de decrecimiento importante. Ambos parámetros pusieron de manifiesto la existencia de una limitante en la escalabilidad del algoritmo, ocasionada por la competencia en el acceso a la memoria común por parte de los núcleos del procesador en un equipo con arquitectura de memoria compartida. La aceleración máxima alcanzada con ocho núcleos fue de 3,207 y la eficiencia varió en el rango de 92,3 % (con tres núcleos) y 40 % (con ocho núcleos). También se demostró que el bajo rendimiento de la aplicación en paralelo se originó por el uso poco eficiente de la memoria caché, según la inspección realizada en el uso de la memoria durante el proceso de solución, haciendo que los procesadores tuvieran que acceder simultáneamente a la memoria principal que es mucho más lenta.

El tercer experimento adicionó una técnica de estabilización por región de confianza al algoritmo DGB. El primer efecto de la estabilización se dio por la reducción en el número de iteraciones, con respecto a las iteraciones en el segundo experimento. La aceleración en el procesamiento del algoritmo con cómputo paralelo y estabilización fue de 4,15, ligeramente superior al del experimento anterior. El rango en la eficiencia también se incrementó con respecto al experimento previo, entre un 97,12 % (con tres núcleos) y un 51,87 % (con ocho núcleos). No obstante, aquí también se observó la relación existente entre el número de núcleos con los valores obtenidos de la aceleración y de la eficiencia, cuya tasa de cambio fue muy inferior a lo esperado, lo que se puede atribuir a la competencia por el acceso a memoria en equipos con arquitectura de memoria compartida como el usado. Si bien, la adopción de un método de estabilización mejoró el desempeño del algoritmo de descomposición al reducir el número de iteraciones necesarias para alcanzar la convergencia, no puede pensarse que este contribuye directamente con la aceleración del algoritmo ya que en realidad este fenómeno

se dio a través de un mecanismo indirecto, dado por la reducción del tamaño del problema maestro como producto de la estabilización.

A partir de varias observaciones realizadas sobre algunas variables generadas por los OPF, se adoptaron dos mecanismos de aceleración complementarios al procesamiento en paralelo de los OPFs y la estabilización del método de DGB. El primer mecanismo estuvo encaminado a reducir el tamaño del problema maestro al controlar los cortes adicionados por iteración y surgió de la verificación del valor de las variables penalizadas en el primer experimento. Dado que en los OPFs las únicas variables penalizadas con valor diferente a cero fueron las de déficit de potencia activa, y que en la mayoría de los OPFs estas tuvieron un valor total inferior a 0,001MW, además con un costo acumulado por iteración sin un efecto significativo sobre el costo del problema original, se determinó que se podía establecer un umbral en el déficit total de potencia activa por OPF para la selección de cortes a partir de tales variables. El segundo mecanismo surgió de la revisión de los multiplicadores en las restricciones de balance de potencia por OPF, necesarios para construir los cortes de Benders, en los resultados del tercer experimento. Se encontró que algunos OPF dejaron de generar las variables duales para esas restricciones a partir de cierta iteración y hasta el final del proceso. Partiendo de este hallazgo se estableció el uso de una ventana de observación en las iteraciones para determinar a partir de qué punto se debía dejar de calcular esos OPF que ya no aportaban cortes al problema maestro. Con ello se logró una reducción adicional de 12,2 % en el tiempo de cómputo, con respecto al mejor tiempo en el tercer experimento.

6.7. Evaluación del esquema de solución por RLA con regla de actualización de segundo orden

Durante las pruebas ejecutadas para comprobar el desempeño del esquema de descomposición por RLA con la regla de actualización de variables duales mediante un método de segundo orden, se encontraron algunas dificultades prácticas para su implementación. Un análisis numérico de la matriz de segundas derivadas $\mathcal{M}$ en (5-35) en cada caso de prueba, proporcionó indicios acerca de algunas características de la formulación del problema que propiciaron un mal funcionamiento.

El análisis numérico requiere el conocimiento de los valores y vectores propios de cada matriz, por ejemplo, para determinar si la matriz esta mal condicionada, indicar si la matriz es singular, la dimensión del espacio nulo, entre otros. En cuanto al número de condición de una matriz A dada, este se calcula como la relación entre el valor propio más grande (σ_1) y

el valor propio más pequeño (σ_n), relativo a la norma L_2, como:

$$\kappa(A) = \frac{\sigma_1}{\sigma_n} \tag{6-6}$$

Cuando el número de condición es grande, la matriz de coeficientes A está mal condicionada [97]. El número de condición además indica cuán sensible es la respuesta a perturbaciones en los datos de entrada y a los errores de redondeo hechos durante el proceso de solución. Ante un problema con mal condicionamiento se plantean dos alternativas según [97]. La primera consiste en verificar que la matriz sea de rango completo, es decir, no deben existir filas o columnas linealmente dependientes. La segunda alternativa consiste en perturbar un poco la matriz para volverla bien condicionada, aunque ese ruido puede tener efectos importantes sobre la solución. Ambas se usarán en caso de ser necesario.

En primera instancia, se realizó una prueba sobre el sistema de potencia IEEE de 30 barras para un horizonte temporal de veinticuatro horas, dos escenarios de viento y una contingencia. La matriz de segundas derivadas $\mathcal{M}$ tuvo un número de condición muy alto ($7,029 \times 10^{10}$) y valores singulares positivos. Entonces, la matriz de segundas derivadas de este modelo es singular. Igualmente, se trató de evitar la singularidad incluyendo pequeñas perturbaciones en las submatrices pero no se logró una corrección efectiva. Por otra parte, se observó que los $\Delta\lambda$ de los OPF correspondientes a los casos base eran de menor magnitud comparados con los de los OPF en los casos contingentes y, que entre los primeros, varios cambiaban de signo iteración tras iteración mientras que los segundos tuvieron signo consistentes. Esto evidencia inestabilidades en el proceso de actualización de las variables duales originadas por el mal condicionamiento de la matriz $\mathcal{M}$. Por su parte, el número de condición calculado para la matriz de sensibilidad $\widetilde{\mathcal{M}}$ fue de 1,1972 y los valores propios de esta matriz fueron pequeños en el rango de $5,7 \times 10^{-9}$ y -0,0021.

Solo con el fin observar la incidencia de la magnitud de la probabilidad de la contingencia en el cálculo de $\Delta\lambda$ y en la convergencia del algoritmo, se condujo otro experimento aumentando el valor de la probabilidad a 1×10^{-1}. Los valores de $\Delta\lambda$ de los casos base y contingentes tuvieron valores muy similares y, al igual que en el experimento anterior, se presentaron cambios de signo en $\Delta\lambda$ de los OPF de casos base de una iteración a la siguiente pero en pocos casos. Para esta prueba las normas L_2 y L_∞ del error en $s - p$ disminuyeron rápidamente en las primeras diez iteraciones y luego de forma mucho más lenta hasta alcanzar la convergencia al cabo de 728 iteraciones. La norma L_2 presentó oscilaciones cuando su valor se hizo pequeño (< 1).

Luego, un experimento fue conducido sobre el caso de prueba construido a partir del sistema de potencia colombiano considerando dos escenarios de viento y una contingencia. Al igual que en el experimento con el caso IEEE de 30 barras, para este caso de estudio la matriz $\mathcal{M}$ resultó mal condicionada al tener un número de condición muy grande, de hecho con valor

infinito. La matriz $\mathcal{M}$ resultó ser de rango deficiente puesto que tiene filas linealmente dependientes dadas por la multiplicidad de generadores conectados por nodo en algunas barras del sistema.

Para los dos casos de estudio, otros experimentos conducentes a la correcta sintonización de los coeficientes de aumentación c y regularización b dejaron en evidencia lo sensible que es esta regla de actualización de segundo orden a la elección de estos valores, sobre todo al coeficiente de regularización b. Ninguno de estos experimentos logró tener un efecto sobre la matriz de segundas derivadas de tal forma que esa matriz se volviera no singular.

Se puede concluir que el mal condicionamiento hace parte integral de la formulación del problema bajo estudio, en parte por la disparidad en los costos dada por el valor de las probabilidades de las contingencias que hacen que existan en las matrices coeficientes con valores muy pequeños que puedan estar ocasionando el mal condicionamiento; y por otra parte, en sistemas con múltiples generadores por nodo la matriz de segundas derivadas no es de rango completo. En este tipo de problemas las técnicas de algebra lineal convencionales no son directamente aplicables, dado que el tratamiento numérico de la matriz de segundas derivadas $\mathcal{M}$ es difícil. En el tratamiento de este tipo de problemas resultan apropiados los métodos de regularización (o estabilización) numérica basados en descomposición QR o descomposición en valores singulares (SVD, por sus síglas en inglés) truncados para descartar los valores singulares pequeños, diferentes de cero o no, que provocan una deficiencia numérica en el rango.

Finalmente, se encontró que para un sistema de gran tamaño, el cálculo directo de $\Delta\lambda$ es costoso computacionalmente hablando debido a las dimensiones del problema, haciendo que además se necesite implementar algún método de descomposición pero a nivel del cálculo matricial.

7. Conclusiones y Trabajo Futuro

7.1. Conclusiones

En esta tesis se presentó una metodología de solución para un modelo de programación estocástica para la planeación de la operación de sistemas eléctricos de potencia con altas penetraciones de FER, respuesta de la demanda y administración central de sistemas de almacenamiento energético, que además incluye de manera explícita el modelo AC de la red de transmisión.

El problema de la programación de la operación de un sistema de potencia con una formulación como la expuesta en el Capítulo 2 es mucho más complejo que otras propuestas encontradas en la literatura para la programación del despacho de la generación como la Coordinación Hidro-Térmica (CHT) y el OPF seguro ante contingencias (MP-SCOPF). Lo anterior se demostró en la Sección 2.4 mediante la comparación de las dimensiones (número de variables y de restricciones) y la estructura matricial del jacobiano de las condiciones de optimalidad, de cada una de estas dos formulaciones y MPSSOPF, en sus versiones lineal y no lineal. La planeación del despacho para un sistema de potencia pequeño, como el de tres barras, con MPSSOPF quedó definido con un 88 % más de restricciones y un 136 % más de variables que la formulación que le sigue en tamaño que fue MP-SCOPF. Por su parte, la estructura matricial del jacobiano para MPSSOPF resultó ser más compleja al tener bloques diagonales, que a su vez están compuestos por bloques diagonales con bloques de acoplamiento adicionales, y unos bloques verticales y horizontales de acoplamiento correspondientes a las restricciones intertemporales, estructura que no se evidenció en las demás formulaciones. A la fecha, un problema con las características de MPSSOPF, definidas por su gran dimensionalidad y por su estructura matemática tan compleja, que además considere las restricciones de la red de transmisión del modelo AC, no había sido resuelto.

Utilizar un modelo AC de la red de transmisión en lugar del modelo DC, tiene un impacto sobre la apropiada asignación y valoración de algunos recursos del sistema eléctrico de potencia como las reservas, cuando estos se ofrecen al costo de oportunidad, como se demostró en el Capítulo 3. La evaluación de las consecuencias de usar un modelo AC o un modelo DC resultó relevante para determinar que con el modelo AC se hizo una correcta asignación de cantidades de reserva, distribuída geográficamente, además, se reprodujeron aspectos conocidos de la operación de un sistema de potencia real, como el sistema de potencia colombiano,

lo que no sucedió con el modelo DC de la red de transmisión.

Aunque incorporar el modelo AC en el problema de planeación puede resultar ventajoso por las razones expuestas, resolver de forma directa un problema de la envergadura que puede alcanzar MPSSOPF requiere mucho tiempo, grandes esfuerzos de cálculo y recursos informáticos. Tiempos de cálculo como los obtenidos para el caso colombiano, considerando una cantidad moderada de escenarios de generación eólica y contingencias, que estuvieron al rededor de 48 horas por simulación, no resultan prácticos para una aplicación de esta formulación en el contexto de un mercado de día en adelanto, y pueden hacer que se descarte su uso. Por tal motivo, se exploraron diferentes estrategias que explotaron la estructura matemática del problema para abordar su solución usando técnicas de descomposición.

Una primera dirección de búsqueda se centró en la aplicación de un esquema basado en relajación lagrangiana con lagrangiano aumentado (RLA). La resolución del problema dual basado en métodos de tipo subgradiente con tamaño de paso adaptado para la actualización de los multiplicadores, aunque atractivo por su simplicidad, resultó en una lenta convergencia y comportamiento oscilatorio de las soluciones, tanto del problema central como de los subproblemas. Estas dificultades hacen que tal esquema de primer orden no sea el adecuado para el tipo de problemas que se pretenden resolver.

Dado lo anterior, surgió la idea de acelerar el proceso de convergencia del método RLA cambiando la regla de actualización de las variables duales por una regla basada en información de segundo orden, dada por el Hessiano de la función lagrangiana. Las pruebas preliminares dejaron al descubierto algunas dificultades para su exitosa implementación. La matriz de segundas derivadas parciales resultó mal condicionada, básicamente, por dos motivos principales. Las cantidades que están afectadas por las probabilidades, que son muy pequeñas (1×10^{-5}), hacen que algunas submatrices tengan valores demasiado pequeños y la matriz se vuelve singular. Por otra parte, en sistemas de potencia en los que se conecta más de un generador por nodo, como en el caso del sistema de potencia colombiano, la matriz tiene filas linealmente dependientes y por tanto la matriz es singular. A partir de estas observaciones se determinó la necesidad de considerar mecanismos adicionales de regularización numérica como los métodos de descomposición QR o SVD para tratar la matriz de segundas derivadas con rango deficiente, lo que se sugiere como trabajo futuro.

Luego, se exploró otra variante que permitió tomar ventaja de la existencia de variables de *complicación* en la formulación matemática del problema de optimización, haciendo de este un buen candidato para su solución mediante Descomposición Generalizada de Benders (DGB). Específicamente, las variables representando las inyecciones de potencia activa al ser tratadas como parámetros permitieron la descomposición del problema original en un problema maestro lineal y en varios subproblemas no lineales, cada uno correspondiente a

un OPF independiente. Separar la parte no lineal de la parte lineal fue posible gracias a una reformulación de los OPF incorporando variables de penalización en la función objetivo y en las restricciones, tal que se restituyeran los grados de libertad perdidos al fijar el conjunto de variables de potencia activas dentro de cada subproblema. Este artificio además de ayudar a cerrar el balance de potencia activa y reactiva en los OPF, evitó tener que resolver problemas factibles auxiliares para generar cortes de factibilidad.

La estrategia de solución por DGB se potenció al implementar diferentes paradigmas de aceleración señalados en la literatura, para reducir tanto el número de iteraciones como el tiempo de solución de una parte del problema, y por ende, el tiempo de convergencia global, tales como: i) una versión multicorte del algoritmo para DGB; ii) una técnica de estabilización inspirada en los métodos de haz con región de confianza, con esta última en forma de caja; y iii) el cómputo en paralelo de los subproblemas en una plataforma paralela como un computador de múltiples núcleos y arquitectura de memoria compartida. Dos medidas de aceleración adicionales se diseñaron a partir de observaciones realizadas en la evolución de algunos parámetros, iteración trás iteración. La primera medida permitió un filtrado de los cortes antes de agregarlos al problema maestro, con base en un umbral mínimo en el déficit total de potencia activa por OPF como criterio de selección. La segunda medida disminuyó el tiempo de cómputo de los subproblemas al dejar de calcular los OPF que generaron multiplicadores en las restricciones de balance de potencia iguales a cero, alcanzando una reducción de 12,2 %, con respecto al experimento previo, y de 88 % con respecto a la solución sin medidas de aceleración.

El desempeño del algoritmo DGB con técnicas de aceleración fue ilustrado mediante experimentos computacionales en dos sistemas de potencia de diferentes tamaños: el caso IEEE de 30 barras y una representación del sistema de potencia colombiano de 96 barras. La codificación del algoritmo se realizó con *software* académico especializado de código abierto como `MATPOWER` y su solución mediante *software* comercial de uso libre académico como `GUROBI` y `IPOPT`. La plataforma paralela elegida fue un servidor multinúcleo con arquitectura de memoria compartida operando bajo un paradigma de programación tipo maestro-esclavo, con los esclavos usados para resolver los subproblemas, y asignación dinámica de tareas a los núcleos esclavos.

La eficacia y validez de esta técnica de descomposición para problemas de grandes dimensiones, como el caso de prueba basado en el sistema de potencia colombiano, quedaron demostradas en la práctica con la reducción del tiempo de cálculo y la validación de los resultados comparándolos con los de una solución directa sin descomposición, tomada como solución referencia. El tiempo de procesamiento para el sistema de potencia colombiano se redujo cerca de 20 veces, con respecto al tiempo de la solución de referencia. En contraste, el tiempo de cómputo para el sistema IEEE de 30 barras, que se puede considerar como

pequeño, fue mayor para la solución con DGB comparado con la solución de referencia, lo que indica que para ese tipo de problemas la descomposición no mejora su velocidad de procesamiento si el cálculo en paralelo de los subproblemas se da en entornos de memoria compartida. En los dos casos de estudio considerados, se encontró que la estrategia propuesta produjo un costo óptimo con una desviación mínima con respecto a la solución de referencia. Estos resultados se consideraron aceptables teniendo en cuenta que el problema en principio es no lineal, de gran dimensión, sumamente plano y que el problema se resolvió en tiempos razonables.

La implementación en paralelo de la solución de los OPFs en un equipo con arquitectura de memoria compartida como la usada, aunque aportó un cierto nivel de aceleración en la ejecución de los experimentos, no logró llegar a valores cercanos de aceleración que en teoría se podían alcanzar, dado que el problema por ser de grandes dimensiones requirió para su cálculo el uso de memoria que excedía la capacidad de la memoria caché del equipo que es más rápida y entonces, la memoria compartida se convirtió en un obstáculo para el rendimiento del método de DGB. Por lo anterior, un clúster de nodos individuales podría ser una estrategia que mejore el rendimiento de la aplicación ya que este es de memoria distribuída.

A lo largo de los experimentos, tanto con RLA como con DGB, se identificaron problemas relacionados con la gran disparidad entre los costos presentes en el problema original debido a las probabilidades de contingencia que son pequeñas. Esto creó problemas de diferente índole. Por ejemplo, los costos que son pequeños por estar multiplicados por la probabilidad de contingencia, al ser introducidos en la función de condiciones de optimalidad de primer orden pueden dar números muy pequeños, no porque se haya cumplido realmente de forma satisfactoria las condiciones de optimalidad de primer orden, sino porque los gradientes de la función de costos son demasiado pequeños por culpa de las probabilidades. Esto afectó el método de RLA de segundo orden propuesto al dejar mal condicionada la matriz de sensibilidad. También representa un inconveniente para un posible y futuro método de punto interior paralelizado y descompuesto porque, al igual que en el método de RLA de segundo orden, van a surgir matrices mal condicionadas numéricamente por esas cantidades tan pequeñas. Otra fuente adicional de problemas en el desempeño o implementación éxitosa de los algoritmos fue la naturaleza del problema que resultó ser demasiado plano. Esto tuvo muchas consecuencias en particular para RLA porque indujo oscilaciones en el método del subgradiente y causó mal condicionamiento numérico en el método de segundo orden.

En general, los resultados sugirieron que el esquema de solución por DGB propuesto es eficiente para tratar el problema de optimización estudiado, con una asignación de cantidades de potencia y reservas bastante aproximada además del buen desempeño computacional en cuanto al tiempo de cálculo, generando un avance en el estado de arte de este campo de estudio.

7.2. Contribuciones de la tesis

A continuación se listan las contribuciones realizadas con el desarrollo de esta tesis al área de planeación de la operación de sistemas eléctricos de potencia:

1. Durante el desarrollo de la tesis se reconoció cuán numéricamente difícil es el problema MPSSOPF, que a parte de ser de gran tamaño, tiene estructura matemática compleja y resultó ser demasiado plano, presentando retos numéricos importantes. Estas dificultades frustraron la implementación del método de descomposición con RLA, con actualización de variables duales basado en métodos de subgradiente y en técnicas de sensibilidad de segundo orden, y además puede ser un problema a la hora de llevar a cabo la descomposición del paso de Newton en el método de punto interior. Sin embargo, el método de Descomposición Generalizada de Benders si logró resolver este problema con un nivel de precisión aceptable.

2. El análisis comparativo del uso de MPSSOPF con las restricciones de red de los modelos AC y DC en un estudio de mercados multi-dimensionales de energía, reserva y rampa en el contexto del día en adelanto y usando un sistema de potencia de tamaño real. Este estudio demostró el impacto positivo de una modelación completa de la red de transmisión con modelo AC en la apropiada asignación y valoración de algunos recursos del sistema eléctrico de potencia como las reservas, cuando estos se ofrecen al costo de oportunidad, y la reproducción de aspectos conocidos en la operación de un sistema de potencia real que no fueron revelados cuando se uso el modelo DC de la red de transmisión.

3. Una estrategia de descomposición basada en DGB que facilita la solución de problemas de planeación de la operación con FER, almacenamiento energético y demanda flexible, mediante la división del problema estocástico no lineal inicial en un problema lineal y varios subproblemas no lineales, estos últimos al nivel de un OPF independiente que se pueden resolver en paralelo. Esta estrategia permitiría que los problemas de optimización sean escalables y manejables, y en teoría se pueda manejar una cantidad importante de escenarios de generación renovable y contingencias en sistemas de cómputo de altas prestaciones con múltiples núcleos de procesamiento o nodos independientes, pero con arquitectura de memoria distribuída.

4. Una aplicación construida sobre lenguajes computacionales de libre uso y distribución como `MATPOWER`, que se puede modificar y adaptar a problemas específicos para ser implementado en un proceso secuencial de solución o en plataformas de cómputo paralelo.

5. Una formulación incipiente para la actualización de los multiplicadores en la iteración dual del método de descomposición por RLA, basado en análisis de sensibilidad o información de segundo orden de las condiciones de optimalidad del lagrangiano aumentado del problema de optimización. Esfuerzos posteriores pueden retomarse a partir del análisis desarrollado aquí.

7.3. Trabajo Futuro

Partiendo de los resultados obtenidos y del análisis de otras alternativas experimentales exploradas, se identificaron direcciones de investigaciones futuras que permitan extender el presente trabajo:

- Implementar la solución de los subproblemas del método de DGB en un clúster con varios nodos de procesamiento independientes, dado que la memoria en estas plataformas de cálculo es distribuída y el acceso a la memoria física no es común. De esta manera se mejoraría el rango de aceleración y la eficiencia en la solución del método de DGB al evitar los cuellos de botella por el acceso simultáneo a la memoria local que se presentan en equipos con arquitectura de memoria compartida. Adicionalmente, también se podría escalar el algoritmo usando una cantidad importante de nodos.

- Usar herramientas efectivas de álgebra lineal numérica para descomponer el cálculo del paso de Newton en el método de punto interior, empleando métodos conocidos, como el complemento de Schur, y acelerar su procesamiento mediante cómputo paralelo. Estas herramientas aprovechan la estructura matricial del jacobiano de las condiciones de optimalidad para problemas de optimización de gran escala, que no solo es simétrica y rala (la mayoría de entradas son cero), sino que también es estructurada por bloques. En el caso de MPSSOPF la estructura matricial tiene bloques diagonales representando las restricciones de cada estado operativo y bloques fuera de la diagonal conteniendo las restricciones de acople. Un ejemplo de este tipo de aplicaciones esta dado por la implementación paralela para el aprovechamiento de este tipo de estructuras matriciales en optimización de gran escala, usando el método de punto interior primal dual, estudiado en [98] pero para estructuras matriciales más sencillas que MPSSOPF.

- Integrar la dimensión de la comisión de unidades al esquema de descomposición propuesto con DGB. Un buen número de estudios han abordado la solución de comisión de unidades mediante DB dado que las variables binarias se pueden tratar como variables de *complicación*. En este caso, los límites de potencia reactiva (2-34) deben ser trasladados al problema maestro en virtud de la variable binaria de estado de comisionamiento u. Adicionalmente, se deben investigar aspectos de convexidad en los cortes debido a que el dominio de los despachos ya no es convexo, sino que será $0 \cup [P_{min}, P_{max}]$, y ese

rango es no convexo. Lo anterior implica tener más precaución con la forma en la que se hacen los cortes.

- Explorar como implementar la formulación para la actualización de variables duales con un método de segundo orden, como el propuesto en la Sección 5.2, en el problema de la planeación de la operación para lograr una solución por RLA en un tiempo de solución razonable. Es posible aplicar algún método de regularización numérica, por ejemplo, métodos de descomposición canónica como QR o SVD para tratar el sistema de ecuaciones cuando la matriz de segundas derivadas sea de rango deficiente. Incluir herramientas de algebra lineal como las mencionadas en el punto anterior serán de gran utilidad para el cálculo eficiente de grandes matrices de sensibilidad, lo que también debería considerarse.

A. Formulación Matemática del Mecanismo de Almacenamiento

Los detalles del término $f_s(\bullet)$ de la función(2-1), específicamente los términos de (2-7) relacionados con el valor residual esperado de energía almacenada en estados terminales, se presentan a continuación y son tomados de [99].

Primero, para cada recurso de almacenamiento i, se requiere una vía eficiente de calcular la cantidad esperada de energía almacenada al principio y al final de cada periodo t, para cada escenario j. Esto se denotará por los $n_{Jt} \times 1$ vectores S_I^{ti} y S_F^{ti} respectivamente, donde n_{Jt} es el número de escenarios en el periodo t.

La energía almacenada s_F^{tij0} en la unidad i al final del periodo t en el estado base del escenario j, puede ser calculada de forma determinista desde la energía almacenada al principio s_I^{tij0} y las inyecciones de potencia en tal estado, donde se asume que las pérdidas son proporcionales al promedio de energía almacenada durante el periodo. Usando la definición en (2-26), esta relación puede ser expresada como sigue:

$$s_F^{tij0} = s_I^{tij0} + s_\Delta^{tij0} - \Delta\eta_{loss}\frac{s_I^{tij0} + s_F^{tij0}}{2} \tag{A-1}$$

$$= \beta_1^i s_I^{tij0} + \beta_2^i s_\Delta^{tij0} \tag{A-2}$$

donde,

$$\beta_1^i \equiv \frac{1 - \Delta\frac{\eta_{loss}^i}{2}}{1 + \Delta\frac{\eta_{loss}^i}{2}} \qquad \beta_2^i \equiv \frac{1}{1 + \Delta\frac{\eta_{loss}^i}{2}} \tag{A-3}$$

Suponiendo que ocurre una contingencia en una fracción α a través de un periodo determinado, la energía almacenada esperada al momento en que ocurre la contingencia s_α^{tijk} es expresada como:

$$s_\alpha^{tijk} = s_I^{tijk} + \alpha\left(s_F^{tij0} - s_I^{tijk}\right) \tag{A-4}$$

Entonces, las pérdidas son más complicadas de calcular y son iguales a:

$$s_{loss}^{tijk} = \Delta\eta_{loss}^{i}\left[\alpha\frac{s_I^{tijk}+s_\alpha^{tijk}}{2}+(1-\alpha)\,\frac{s_\alpha^{tijk}+s_F^{tijk}}{2}\right] \tag{A-5}$$

$$= \Delta\eta_{loss}^{i}\left[\alpha\frac{s_I^{tij0}+s_F^{tij0}}{2}+(1-\alpha)\,\frac{s_I^{tijk}+s_F^{tijk}}{2}\right] \tag{A-6}$$

donde, (A-6) sigue directamente de (A-4) y (A-5), teniendo presente que $s_I^{tijk} = s_I^{tij0}$.

En este caso, la energía almacenada en la unidad i al final del periodo t en el estado jk puede ser calculada de forma determinista desde la energía inicial almacenada y las inyecciones en estados $j0$ y jk como sigue:

$$s_F^{tijk} = s_I^{tijk}+\alpha s_\Delta^{tij0}+(1-\alpha)\,s_\Delta^{tijk}-s_{loss}^{tijk} \tag{A-7}$$

$$= \alpha\left[s_I^{tij0}+s_\Delta^{tij0}-\Delta\eta_{loss}^{i}\frac{s_I^{tij0}+s_F^{tij0}}{2}\right]+(1-\alpha)\left[s_I^{tijk}+s_\Delta^{tijk}-\Delta\eta_{loss}^{i}\frac{s_I^{tijk}+s_F^{tijk}}{2}\right] \tag{A-8}$$

$$= \alpha\left[\beta_1^i s_I^{tij0}+\beta_2^i s_\Delta^{tij0}\right]+(1-\alpha)\left[s_I^{tijk}+s_\Delta^{tijk}-\Delta\eta_{loss}^{i}\frac{s_I^{tijk}+s_F^{tijk}}{2}\right] \tag{A-9}$$

$$= \beta_5^i s_I^{tijk}+\beta_4^i s_\Delta^{tij0}+\beta_3^i s_\Delta^{tijk} \tag{A-10}$$

donde,

$$\beta_3^i \equiv \left(\frac{1}{1-\alpha}+\Delta\frac{\eta_{loss}^i}{2}\right)^{-1} = \frac{1-\alpha}{1+(1-\alpha)\,\Delta\frac{\eta_{loss}^i}{2}} \tag{A-11}$$

$$\beta_4^i \equiv \frac{\alpha}{1-\alpha}\beta_2^i\beta_3^i = \frac{\alpha}{\left(1+\Delta\frac{\eta_{loss}^i}{2}\right)\left(1+(1-\alpha)\,\Delta\frac{\eta_{loss}^i}{2}\right)} \tag{A-12}$$

$$\beta_5^i \equiv \frac{\beta_1^i}{\beta_2^i}\left(\beta_3^i+\beta_4^i\right) = \left(1-\Delta\frac{\eta_{loss}^i}{2}\right)\frac{\alpha+(1-\alpha)\left(1+\Delta\frac{\eta_{loss}^i}{2}\right)}{\left(1+\Delta\frac{\eta_{loss}^i}{2}\right)\left(1+(1-\alpha)\,\Delta\frac{\eta_{loss}^i}{2}\right)} \tag{A-13}$$

Sean G_k^{ti} y H_k^{ti} las matrices que contienen las eficiencias de carga y de descarga, respectivamente, relacionadas con el cambio correspondiente en la energía almacenada desde el principio hasta el fin del periodo t, en el estado jk, de la unidad de almacenamiento i. Específicamente, los elementos g_{jl}^{ti} y h_{jl}^{ti} en la fila j y la columna l de los conjuntos G_k^{ti} y H_k^{ti} son como se indica.

$$g_{jl}^{ti} = \begin{cases} -\Delta\eta_{in}^{i} & \text{donde la columna } l \text{ corresponde a } p_{sc}^{tijk} \\ 0 & \text{de otra forma} \end{cases} \tag{A-14}$$

$$h_{jl}^{ti} = \begin{cases} -\Delta\frac{1}{\eta_{out}^{i}} & \text{donde la columna } l \text{ corresponde a } p_{sd}^{tijk} \\ 0 & \text{de otra forma} \end{cases} \tag{A-15}$$

La razón para mantener G_k^{ti} y H_k^{ti} separados es hacer posible el uso de diferentes precios para representar tanto la ganancia por el incremento de la cantidad de almacenamiento residual como la perdida por la reducción de la cantidad de almacenamiento residual. La necesidad de usar diferentes precios para valorar la carga y la descarga es soportada por dos eventos posibles. En primer lugar, la energía almacenada puede no ser usada en un estado terminal dado si existe un tiempo mejor para usarla (esperando un alto precio en el horizonte). En segundo lugar, almacenar energía adicional en un estado terminal dado puede no ser lo deseable si hay un mejor tiempo para hacerlo (esperando un bajo precio en el horizonte).

Usando esas matrices, (A-2) puede ser expresada para el vector S_F^{ti} como una función determinista de S_I^{ti} y las inyecciones como:

$$S_F^{ti} = \beta_1^i S_I^{ti} + \beta_2^i \left(G_0^{ti} + H_0^{ti}\right) x \tag{A-16}$$

De otra parte, la energía almacenada esperada en cada escenario al principio del periodo t depende de los valores correspondientes al final del periodo $t-1$ y las probabilidades de transición. Sea σ^t igual al vector de probabilidades de cada uno de los escenarios base al final del periodo $t-1$, condicionado por la llegada al final de tal periodo sin la ocurrencia de una contingencia.

$$\sigma^t = \frac{1}{\gamma^t}\psi^{(t-1)J0} = \frac{1}{\gamma^t}\begin{bmatrix} \psi^{(t-1)10} \\ \psi^{(t-1)20} \\ \vdots \\ \psi^{(t-1)n_{Jt-1}0} \end{bmatrix} \tag{A-17}$$

Si además $[a]$ denota la matriz diagonal con el vector a en la diagonal principal, entonces la relación entre S_I^{ti} y $S_F^{(t-1)i}$ puede ser expresada como:

$$\left[\Phi^t\sigma^t\right] S_I^{ti} = \Phi^t \left[\sigma^t\right] S_F^{(t-1)i} \tag{A-18}$$

En otras palabras,

$$S_I^{ti} = D^{ti} S_F^{(t-1)i} \tag{A-19}$$

donde,

$$D^{ti} = \begin{cases} 1_{n_{jt}\times 1} & t = 1 \\ \left[\Phi^t \sigma^t\right]^{-1} \Phi^t \left[\sigma^t\right] & t \neq 1 \end{cases} \tag{A-20}$$

Agrupar los vectores S_I^{ti} y S_F^{ti} para todas las unidades (i desde 1 hasta n_s) permite que la relación anterior sea expresada en términos de la matriz D^t, formada al ubicar el termino D^{ti} a lo largo de la diagonal, y los vectores G_k^{ti} y H_k^{ti} en forma vertical.

$$D^t = \begin{bmatrix} D^{t1} & 0 & \cdots & 0 \\ 0 & D^{t2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D^{tn_s} \end{bmatrix}, \qquad G_k^t = \begin{bmatrix} G_k^{t1} \\ G_k^{t2} \\ \vdots \\ G_k^{tn_s} \end{bmatrix}, \qquad H_k^t = \begin{bmatrix} H_k^{t1} \\ H_k^{t2} \\ \vdots \\ H_k^{tn_s} \end{bmatrix} \tag{A-21}$$

Así mismo, los valores escalares β_n^{ti} son convertidos en matrices diagonales $B_n^{ti} \equiv \beta_n^i I_{n_{Jt}\times n_{Jt}}$ y agrupados para formar:

$$B_n^t = \begin{bmatrix} B_n^{t1} & 0 & \cdots & 0 \\ 0 & B_n^{t2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_n^{tn_s} \end{bmatrix} \tag{A-22}$$

La expresión completa para todas las unidades de almacenamiento, en todos los escenarios, en el periodo t, puede ser expresado como:

$$S_I^t = D^t S_F^{(t-1)} \tag{A-23}$$

$$S_F^t = B_1^t S_I^t + B_{\text{"}}^t \left(G_0^t + H_0^t\right) x \tag{A-24}$$

Las relaciones en (A-23) y (A-24) implican que la energía almacenada esperada en cualquier punto del horizonte de planeación puede ser expresada como una función lineal de la energía inicial almacenada esperada s_0 y la inyección de potencia activa en x, específicamente la inyección de las unidades de almacenamiento.

$$S_I^t = L_I^t s_0 + \left(M_g^t + M_h^t\right) x \tag{A-25}$$

$$S_F^t = L_F^t s_0 + \left(N_g^t + N_h^t\right) x \tag{A-26}$$

Las siguientes expresiones recursivas son usadas para calcular L_I^t, L_F^t, M_g^t, M_h^t, N_g^t y N_h^t.

$$L_I^t = D^t L_F^{(t-1)} = D^t B_1^{(t-1)} L_I^{(t-1)} \tag{A-27}$$

$$L_F^t = B_1^t L_I^t = B_1^t D^t \; L_F^{(t-1)} \tag{A-28}$$

$$M_g^t = D^t \; N_g^{(t-1)} \tag{A-29}$$

$$M_h^t = D^t \; N_h^{(t-1)} \tag{A-30}$$

$$N_g^t = B_1^t M_g^t + B_2^t G_0^t \tag{A-31}$$

$$N_h^t = B_1^t M_h^t + B_2^t H_0^t \tag{A-32}$$

donde, $L_I^1 = D^1$ y $M_g^1 = M_h^1 = 0$.

Si las filas de cada uno de estos vectores y matrices son ordenadas y fraccionadas por escenarios (no por unidades de almacenamiento), se pueden denotar los *j-ésimos* componentes resultantes, cuya *i-ésima* fila corresponde a la unidad de almacenamiento i, con una barra, por ejemplo $\bar{S}_F^{tj}$, $\bar{S}_I^{tj}$, $\bar{G}_k^{tj}$, $\bar{H}_k^{tj}$, $\bar{L}_I^{tj}$, $\bar{L}_F^{tj}$, $\bar{M}_g^{tj}$, $\bar{M}_h^{tj}$, $\bar{N}_g^{tj}$ y $\bar{N}_h^{tj}$. Puede notarse para las matrices B, que el componente $\bar{B}_n^{tj}$ correspondiente es justo la matriz diagonal $[\beta_n]$, con el elemento β_n^i individual en la diagonal. Usando esta notación, el vector $\bar{S}_F^{n_t j}$ representando la energía almacenada residual esperada para todas las unidades en un escenario base j, al final del último periodo n_t del horizonte de planeación, puede ser escrito como una función de estas matrices.

$$\begin{aligned}
\bar{S}_F^{n_t j} &= [\beta_1]\, \bar{S}_I^{n_t j} + [\beta_2] \left(\bar{G}_0^{n_t j} + \bar{H}_0^{n_t j}\right) x \\
&= [\beta_1] \left(\bar{L}_I^{n_t j} s_0 + \left(\bar{M}_g^{n_t j} + \bar{M}_h^{n_t j}\right) x\right) + [\beta_2] \left(\bar{G}_0^{n_t j} + \bar{H}_0^{n_t j}\right) x \\
&= [\beta_1]\, \bar{L}_I^{n_t j} s_0 + \left([\beta_1]\, \bar{M}_g^{n_t j} + [\beta_2]\, \bar{G}_0^{n_t j} + [\beta_1]\, \bar{M}_h^{n_t j} + [\beta_2]\, \bar{H}_0^{n_t j}\right) x
\end{aligned} \tag{A-33}$$

Igualmente, la energía almacenada residual esperada al final del periodo t para cualquier escenario j y contingencia k es expresada como:

$$\begin{aligned}
\bar{S}_F^{tjk} &= [\beta_5]\, \bar{S}_I^{tj} + [\beta_4]\, \bar{S}_\Delta^{tj0} + [\beta_3]\, \bar{S}_\Delta^{tjk} \\
&= [\beta_5]\, \bar{S}_I^{tj} + \left([\beta_4] \left(\bar{G}_0^{tj} + \bar{H}_0^{tj}\right) + [\beta_3] \left(\bar{G}_k^{tj} + \bar{H}_k^{tj}\right)\right) x \\
&= [\beta_5] \left(\bar{L}_I^{tj} s_0 + \left(\bar{M}_g^{tj} + \bar{M}_h^{tj}\right) x\right) + \left([\beta_4] \left(\bar{G}_0^{tj} + \bar{H}_0^{tj}\right) + [\beta_3] \left(\bar{G}_k^{tj} + \bar{H}_k^{tj}\right)\right) x \\
&= [\beta_5]\, \bar{L}_I^{tj} s_0 + \left([\beta_5]\, \bar{M}_g^{tj} + [\beta_4]\, \bar{G}_0^{tj} + [\beta_3]\, \bar{G}_k^{tj} + [\beta_5]\, \bar{M}_h^{tj} + [\beta_4]\, \bar{H}_0^{tj} + [\beta_3]\, \bar{H}_k^{tj}\right) x
\end{aligned} \tag{A-34}$$

La cantidad esperada total de energía almacenada a través de los estados no contingentes al final del horizonte de planeación está dada por:

$$s_F^{n_t} = \frac{1}{\gamma^{(n_t+1)}} \sum_{j \in J^{n_t}} \psi^{n_t j 0}\, \bar{S}_F^{n_t j} \tag{A-35}$$

donde $\gamma^{(n_t+1)} = \sum_j \psi^{n_t j 0}$. Esta expresión puede ser usada en la construcción de términos de la función objetivo.

Regresando al valor de la energía esperada restante en esados terminales, dado por (2-7), si se usa un precio único para cada unidad de almacenamiento i para valorar todas las contribuciones a tal energía restante esperada, sin importar el estado en el que este ocurra, entonces el valor dado por (2-7) sería ese precio multiplicado por la suma ponderada por probabilidad de la energía en cada estado, modificada por la eficiencia de salida. En concreto, el precio se relaciona con el valor de cada MW de energía recuperable, en oposición a la energía almacenada.

$$f_s(x) = C_s^T \left([\eta_{out}^{n_t}] \sum_{j \in J^{n_t}} \psi^{n_t j 0} \bar{S}_F^{n_t j} + \sum_{t \in T} [\eta_{out}^{t}] \sum_{j \in J^t} \sum_{k \in K^{tj} \neq 0} \psi^{tjk} \bar{S}_F^{tjk} \right) \tag{A-36}$$

No obstante, resulta útil clasificar los estados del sistema en tres categorías: estados contingentes terminales, estados bases terminales al final del horizonte y estados no terminales (estados base precediendo el último perido). Esto abre la posibilidad de valorar de manera diferente las contribuciones realizadas a la energía almacenada terminal esperada en cada una de estas categorías de estados. También puede ser útil diferenciar entre el valor obtenido al aumentar la energía almacenada terminal esperada y el valor perdido al disminuirla.

Esto conduce al diseño actual basado en los cinco modelos de precios que surgen de las consideraciones anteriores,

$$F_s(x) = C_s^T (A_1 s_0 + A_2 x + A_3 x) + C_{sc0}^T A_4 x + C_{sd0}^T A_5 x + C_{sck}^T A_6 x + C_{sdk}^T A_7 x \tag{A-37}$$

donde,

$$A_1 = [\eta_{out}^{n_t}] [\beta_1^{n_t}] \sum_{j \in J^{n_t}} \psi^{n_t j 0} \bar{L}_I^{n_t j} + \sum_{t \in T} [\eta_{out}^{t}] [\beta_5^{t}] \sum_{j \in J^t} \left(\sum_{k \in K^{tj} \neq 0} \psi^{tjk} \right) \bar{L}_I^{tj} \tag{A-38}$$

$$A_2 = [\eta_{out}^{n_t}] [\beta_1^{n_t}] \sum_{j \in J^{n_t}} \psi^{n_t j 0} \bar{M}_g^{n_t j} + \sum_{t \in T} [\eta_{out}^{t}] \sum_{j \in J^t} \left(\sum_{k \in K^{tj} \neq 0} \psi^{tjk} \right) \left([\beta_5^{t}] \bar{M}_g^{tj} + [\beta_4^{t}] \bar{G}_0^{tj} \right) \tag{A-39}$$

$$A_3 = [\eta_{out}^{n_t}] [\beta_1^{n_t}] \sum_{j \in J^{n_t}} \psi^{n_t j 0} \bar{M}_h^{n_t j} + \sum_{t \in T} [\eta_{out}^{t}] \sum_{j \in J^t} \left(\sum_{k \in K^{tj} \neq 0} \psi^{tjk} \right) \left([\beta_5^{t}] \bar{M}_h^{tj} + [\beta_4^{t}] \bar{H}_0^{tj} \right) \tag{A-40}$$

$$A_4 = [\eta_{out}^{n_t}]\,[\beta_2^{n_t}] \sum_{j \in J^{n_t}} \psi^{n_t j 0} \bar{G}_0^{n_t j} \tag{A-41}$$

$$A_5 = [\eta_{out}^{n_t}]\,[\beta_2^{n_t}] \sum_{j \in J^{n_t}} \psi^{n_t j 0} \bar{H}_0^{n_t j} \tag{A-42}$$

$$A_6 = \sum_{t \in T} \left[\eta_{out}^{t}\right]\,\left[\beta_3^{t}\right] \sum_{j \in J^t} \sum_{k \in K^{tj} \neq 0} \psi^{tjk} \bar{G}_k^{tj} \tag{A-43}$$

$$A_7 = \sum_{t \in T} \left[\eta_{out}^{t}\right]\,\left[\beta_3^{t}\right] \sum_{j \in J^t} \sum_{k \in K^{tj} \neq 0} \psi^{tjk} \bar{H}_k^{tj} \tag{A-44}$$

Si se usa $\bar{A}_n$ para representar la versión de A_n con todas las columnas removidas excepto para aquellas correspondientes a las inyecciones de carga y descarga relevantes (p_{sc} para $n = 2, 4, 6$ y p_{sd} para $n = 3, 5, 7$), el costo de la energía inicial y final almacenada se puede expresar como

$$f_s\left(s_0, p_{sc}, p_{sd}\right) = -\left(C_{s0}^T s_0 + C_{sc}^T p_{sc} + C_{sd}^T p_{sd}\right) \tag{A-45}$$

donde,

$$C_{s0} = A - 1^T C_s \tag{A-46}$$

$$C_{sc} = \bar{A}_2^T C_s + \bar{A}_4^T C_{sc0} + \bar{A}_6^T C_{sck} \tag{A-47}$$

$$c_{sd} = \bar{A}_3^T C_s + \bar{A}_5^T C_{sd0} + \bar{A}_7^T C_{sdk} \tag{A-48}$$

B. Descripción del Sistema Eléctrico Interconectado Colombiano de 96 Barras

La información necesaria para modelar la operación del sistema eléctrico colombiano es presentada en esta sección. Solo se incluyen los datos que se pueden hacer públicos y que se encuentran en las páginas de XM y la UPME.

Los nombres identificando las barras del sistema, correspondientes con la numeración en el diagrama unifilar, se consignan en la Tabla **B-1**.

B.1. Demanda de energía eléctrica

La demanda de energía eléctrica corresponde a la carga horaria por barra utilizada por XM para la operación del mercado para un día típico del año 2014. El perfil horario de demanda de potencia activa y reactiva es presentado en la Tabla **B-2**.

B.2. Capacidad de generación instalada

La capacidad instalada es igual a la del año 2014 con la adición de un generador eólico ubicado en Copey. El sistema modelado cuenta con una capacidad instalada de 16.310MW, de los cuales un 63.12 % corresponde a generación hidráulica, 23.60 % a térmica a gas natural, 8.37 % a térmica a carbón y 4.91 % a eólica. La generación máxima por unidad de generación se presenta en la Tabla **B-3**.

La mayor parte de la generación hidroeléctrica se ubica en las áreas de Antioquia (47.5 %) y Centro (30.6 %), mientras que la generación térmica es mayoritariamente usada en el área Costa Atlántica (49.5 %), Antioquia (18.7 %) y Oriente (17.2 %).

Tabla B-1.: Identificación de las barras del sistema de potencia colombiano

No.	Nombre	No.	Nombre	No.	Nombre
1	Cartagena	33	Cartago	65	La Dorada
2	Bolivar	34	San Marcos	66	San Carlos
3	Barranquilla	35	Yumbo	67	La Sierra
4	TEBSA	36	Alto Anchicaya	68	Malena
5	Flores	37	Juanchito	69	Primavera
6	Fundación	38	Pance	70	Termocentro
7	Santa Marta	39	Salvajina	71	Merieléctrica
8	Guajira	40	Paez	72	Barranca
9	Cuestecita	41	San Bernardino	73	Guatiguara
10	Valledupar	42	Jamondino	74	Bucaramanga
11	Copey	43	Mocoa	75	Palos
12	Sabanalarga	44	Altamira	76	Toledo
13	Chinú	45	Betania	77	Samore
14	Cerromatoso	46	Mirolindo	78	Banadia
15	Urrá	47	La Mesa	79	Caño limón
16	Urabá	48	La Enea	80	Tasajero
17	Jaguas	49	San Felipe	81	Cúcuta
18	Guatapé	50	Miel	82	San Mateo2
19	Playas	51	La Guaca	83	Ocaña
20	Oriente	52	Paraiso	84	Bosque
21	Envigado	53	San Mateo	85	Bonda
22	Porce2	54	Tunal	86	Porce3
23	Salto EPM	55	La Reforma	87	Sogamoso
24	Barbosa	56	Circo	88	Alferez
25	Bello	57	Guavio	89	Quimbo
26	La Tasajera	58	Chivor	90	Jaguey
27	Occidente	59	Sochagota	91	Corocora
28	Miraflores	60	Paipa	92	Quifa
29	Ancon Sur	61	Torca	93	Rubiales
30	Esmeralda	62	Bacatá	94	Nva. Esperanza
31	La Hermosa	63	Balsillas	95	Guayabal
32	La Virginia	64	Noroeste	96	Tuluní

B.2.1. Caracterización de la generación hidroeléctrica

El modelado del sistema hidráulico del caso colombiano contempló embalses, plantas en cadenas híidricas con o sin embalse y filtraciones río abajo que pueden ser parcialmente

Tabla B-2.: Perfil horario de demanda para el sistema de potencia colombiano

Hora	**PD** [MW]	**QD** [MVar]	**Hora**	**PD** [MW]	**QD** [MVar]	**Hora**	**PD** [MW]	**QD** [MVar]
1	6277,3	2493,5	9	7844,4	3298,0	17	8482,5	3697,8
2	6048,8	2445,0	10	8150,0	3463,8	18	8371,3	3498,7
3	5917,7	2419,9	11	8458,7	3605,3	19	9319,9	3473,1
4	5887,5	2404,6	12	8665,1	3712,4	20	9451,7	3383,0
5	6120,0	2395,0	13	8458,2	3625,5	21	9184,1	3220,0
6	6649,6	2457,3	14	8440,4	3637,7	22	8357,4	2905,0
7	6882,5	2663,0	15	8579,4	3738,7	23	7560,6	2729,4
8	7328,2	2992,2	16	8582,4	3747,0	24	6848,0	2617,5

Fuente: XM

recuperadas. En cadenas hidráulicas las plantas sin embalse que están localizadas río abajo de otras con embalse, se asumen controladas en su producción por las descargas que ocurran río arriba. En ese caso, también se asumió que existe sólo una planta de control río arriba para las plantas no despachables. Se modelaron dos cadenas hidrológicas: la cadena PAGUA (Paraiso-Guaca) de 600MW, la cadena SALACO (El Charquito-Tequendama-San Antonio-Limonar-La Junca-La Tinta) de 374MW.

Los datos de disponibilidad hídrica usados corresponden a los de un día típico representativo del año 2014 determinado mediante un algoritmo de clusterización. El cálculo de las cuotas de energía de cada planta hidroeléctrica, para el día típico seleccionado, tiene en cuenta los aportes hidrológicos por afluente alimentando cada planta y el factor de producción correspondiente. Toda la información fue obtenida de bases de datos de la Unidad de Planeación Minero Energética - UPME.

B.2.2. Oferta de generación basada en el mercado

La información de costos de generación disponibles proviene de bases de datos de XM y permiten modelar una función de costos del tipo lineal. En entornos de mercados de energía las ofertas para la generación generalmente se estructuran en bloques con un precio determinado y no como un costo polinomial. Los coeficientes de costo disponibles son un punto de partida para construir los bloques de oferta, como se indica más adelante, y que `MATPOWER` simulará como funciones de costos lineales convexas por partes.

En primer lugar, los coeficientes de costo se modificaron ligeramente para diferenciar el costo de producción de las unidades por tipo de tecnología de generación, como se muestra en la

Tabla B-3.: Máxima potencia activa por generador en el sistema de potencia colombiano

Generador	**Pmax** [MW]	**Generador**	**Pmax** [MW]
1. A.Anchicayá	375	26. Paipa	99
2. Amoyá	75.60	27. Paraiso	446
3. B.Anchicayá	74	28. Playas	414
4. Barranca	137	29. Porce3	660
5. Betania	540	30. Porce2	405
6. Termo Bquilla	380	31. Proeléctrica	90
7. TEBSA	697	32. San Carlos	1240
8. Calima	132	33. San Francisco	25,27
9. Candelaria	312	34. Salvajina	285
10. Cartagena	187	35. Sogamoso	791
11. Gecelca3	164	36. Tasajera	306
12. Gecelca3.2	273	37. Tasajero1	155,32
13. Prado	51	38. Tasajero2	155,32
14. Chivor	1000	39. Termo centro	282
15. Termo Dorada	51	40. Termo EMCali	229,07
16. Termo Flores	621,99	41. Termo Sierra	489
17. Guaca	528	42. Termo Valle	205
18. GuadalupeIII	162,67	43. Urrá	340
19. GuadalupeIV	201	44. Termo Zipa	222,23
20. Guajira	296	45. STAT Bacatá	0
21. Guatapé	560	46. SVC Chinú	0
22. Guavio	1200	47. SVC Caño limón	0
23. Jaguas	170	48. SVC Tunal	0
24. Merilectrica	169	49. Copey	800
25. Miel	396		

Fuente: XM

Tabla **B-4**. Entre las unidades térmicas, las unidades de generación a gas natural en ciclo combinado son las de más alto costo, seguidas de las unidades a gas natural en ciclo sencillo y las unidades a carbón. Los generadores hidroeléctricos y eólicos son los de menor costo de producción entre todas las tecnologías.

El proceso para generar tres bloques de oferta por generador comienza por calcular un pequeño coeficiente de costo cuadrático, tal que la diferencia entre los costos marginales evaluados en los límites máximo y mínimo de la potencia activa sea menor o igual al 5 %. La función de costos cuadrática se usa para evaluar el costo marginal de cada generador en los

Tabla B-4.: Coeficientes de la función de costos lineal de generación para los generadores del sistema de potencia colombiano

Generador	**Costo** [KCOP/MW]	**Generador**	**Costo** [KCOP/MW]
1. A.Anchicayá	105,6522	26. Paipa	130,5000
2. Amoyá	116,0870	27. Paraiso	100,4348
3. B.Anchicayá	117,3913	28. Playas	101,7491
4. Barranca	164,2857	29. Porce3	95,2174
5. Betania	97,8261	30. Porce2	103,0435
6. Termo Bquilla	150,0000	31. Proeléctrica	167,1429
7. TEBSA	180,0000	32. San Carlos	90,0000
8. Calima	114,7826	33. San Francisco	120,0000
9. Candelaria	152,8571	34. Salvajina	109,5652
10. Cartagena	161,4286	35. Sogamoso	93,9130
11. Gecelca3	124,0000	36. Tasajera	108,2609
12. Gecelca3.2	126,1667	37. Tasajero1	132,6667
13. Prado	118,7000	38. Tasajero2	134,8333
14. Chivor	92,6087	39. Termo centro	195,0000
15. Termo Dorada	170,0000	40. Termo EMCali	155,7143
16. Flores	185,0000	41. Termo Sierra	190,0000
17. Guaca	99,1304	42. Termo Valle	158,5714
18. GuadalupeIII	113,4783	43. Urrá	106,9565
19. GuadalupeIV	110,8697	44. Termo Zipa	137,0000
20. Guajira	128,3333	45. STAT Bacatá	0,0000
21. Guatapé	96,5217	46. SVC Chinú	0,0000
22. Guavio	91,3043	47. SVC Caño limón	0,0000
23. Jaguas	112,1739	48. SVC Tunal	0,0000
24. Merilectrica	200,0000	49. Copey	40,0000
25. Miel	104,3478		

valores al extremo derecho del intervalo de potencia que define cada bloque. Esos intervalos se determinan de acuerdo con el límite inferior de potencia activa P_{min} de cada generador, así:

1. Si $P_{min} \neq 0$, el primer intervalo corresponde a $(0, P_{min}]$. Los dos restantes se dividen en partes iguales entre P_{min} y P_{max}.

2. Si $P_{min} = 0$, el intervalo $[P_{min}, P_{max}]$ se divide en tres partes iguales.

B.2.3. Definición de los costos de las reservas

Por su parte, los costos de reserva de contingencia se ajustaron al 10 % del costo lineal de cada unidad de generación, mientras que el costo de reserva de rampa de seguimiento de carga se ajustó al 5 % del costo lineal.

B.3. Perfiles de viento

La incertidumbre en la generación de potencia eólica se modela mediante quince escenarios de viento, construidos como se indicó en la Sección 3.2. La Tabla **B-5** contiene los datos de producción de potencia eólica normalizada para cada escenario, para un periodo de tiempo de veinticuatro horas.

Tabla B-5.: Perfiles horarios de viento para el sistema de potencia colombiano

Esc.	**Hora**							
	1	2	3	4	5	6	7	8
1	0,157	0,190	0,126	0,080	0,097	0,040	0,097	0,126
2	0,296	0,342	0,190	0,190	0,069	0,053	0,053	0,040
3	0,126	0,069	0,097	0,027	0,036	0,000	0,000	0,036
4	0,420	0,420	0,420	0,216	0,256	0,296	0,256	0,342
5	0,599	0,599	0,692	0,518	0,420	0,599	0,518	0,216
6	0,534	0,485	0,420	0,518	0,599	0,566	0,734	0,897
7	0,256	0,256	0,157	0,216	0,157	0,256	0,373	0,256
8	0,080	0,080	0,069	0,080	0,126	0,097	0,069	0,126
9	0,311	0,358	0,311	0,168	0,190	0,190	0,216	0,190
10	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,216
11	0,373	0,342	0,534	0,256	0,269	0,190	0,256	0,190
12	0,203	0,190	0,190	0,115	0,058	0,075	0,058	0,040
13	0,404	0,311	0,373	0,373	0,404	0,358	0,342	0,599
14	0,053	0,049	0,023	0,023	0,000	0,006	0,006	0,000
15	0,000	0,097	0,190	0,080	0,126	0,157	0,296	0,256

Continuación...

Esc.	**Hora**							
	9	10	11	12	13	14	15	16
1	0,296	0,216	0,296	0,420	0,566	0,566	0,518	0,518
2	0,053	0,216	0,373	0,469	0,566	0,647	0,566	0,518
3	0,190	0,190	0,126	0,518	0,692	0,734	0,647	0,420
4	0,566	0,692	0,566	0,734	0,763	0,763	0,763	0,799
5	0,692	0,734	0,763	0,692	0,763	0,862	0,956	0,921
6	0,862	0,763	0,830	0,692	0,799	0,897	0,921	0,956
7	0,373	0,518	0,566	0,566	0,599	0,518	0,599	0,599
8	0,469	0,518	0,518	0,599	0,518	0,566	0,566	0,469
9	0,420	0,358	0,469	0,706	0,734	0,830	0,862	0,921
10	0,036	0,097	0,115	0,157	0,190	0,157	0,126	0,075
11	0,583	0,566	0,706	0,734	0,862	0,763	0,692	0,799
12	0,157	0,342	0,342	0,469	0,692	0,799	0,749	0,678
13	0,599	0,647	0,749	0,734	0,921	0,873	0,921	0,862
14	0,012	0,000	0,069	0,216	0,453	0,678	0,706	0,485
15	0,420	0,311	0,469	0,420	0,469	0,566	0,678	0,647

Esc.	**Hora**							
	17	18	19	20	21	22	23	24
1	0,420	0,342	0,373	0,373	0,296	0,296	0,216	0,157
2	0,420	0,518	0,420	0,296	0,296	0,216	0,216	0,190
3	0,518	0,469	0,373	0,373	0,342	0,342	0,296	0,157
4	0,862	0,862	0,692	0,692	0,469	0,734	0,647	0,518
5	0,921	0,862	0,897	0,862	0,862	0,763	0,599	0,692
6	0,897	0,897	0,830	0,862	0,862	0,830	0,799	0,599
7	0,692	0,469	0,599	0,373	0,342	0,296	0,342	0,342
8	0,373	0,420	0,373	0,256	0,190	0,190	0,566	0,126
9	0,830	0,763	0,763	0,583	0,518	0,358	0,256	0,373
10	0,027	0,006	0,000	0,000	0,003	0,000	0,000	0,000
11	0,734	0,706	0,763	0,534	0,518	0,518	0,342	0,358
12	0,485	0,485	0,678	0,453	0,706	0,678	0,485	0,256
13	0,763	0,678	0,583	0,534	0,518	0,583	0,599	0,566
14	0,373	0,518	0,420	0,269	0,157	0,136	0,080	0,053
15	0,518	0,469	0,420	0,342	0,256	0,216	0,243	0,342

C. Datos del Sistema de Potencia IEEE de 30 Barras

Los diferentes datos necesarios para modelar el caso IEEE de 30 barras se presentan en esta Sección, indicando las fuentes de las que fueron tomados y algunas modificaciones realizadas sobre los datos originales.

Los parámetros de las líneas de transmisión son tomados de [95] redondeados cerca de 0,01 y son presentados en la Tabla **C-1**. Todos los datos están en una base de 100MVA.

Tabla C-1.: Parámetros de línea para el sistema de potencia IEEE de 30 barras

No.	F_{BUS}	T_{BUS}	**R**	**X**	**B**
1	1	2	0,0192	0,0575	0,0264
2	1	3	0,0452	0,1852	0,0204
3	2	4	0,0570	0,1737	0,0184
4	3	4	0,0132	0,0379	0,0042
5	2	5	0,0472	0,1983	0,0209
6	2	6	0,0581	0,1763	0,0187
7	4	6	0,0119	0,0414	0,0045
8	5	7	0,0460	0,1160	0,0102
9	6	7	0,0267	0,0820	0,0085
10	6	8	0,0120	0,0420	0,0045
11	6	9	0,0000	0,2080	0,0000
12	6	10	0,0000	0,5560	0,0000
13	9	11	0,0000	0,2080	0,0000
14	9	10	0,0000	0,1100	0,0000
15	4	12	0,0000	0,2560	0,0000
16	12	13	0,0000	0,1400	0,0000

Continuación...

No.	F_{BUS}	T_{BUS}	**R**	**X**	**B**
17	12	14	0,1231	0,2559	0,0000
18	12	15	0,0662	0,1304	0,0000
19	12	16	0,0945	0,1987	0,0000
20	14	15	0,2210	0,1997	0,0000
21	16	17	0,0824	0,1932	0,0000
22	15	18	0,1070	0,2185	0,0000
23	18	19	0,0639	0,1292	0,0000
24	19	20	0,0340	0,0680	0,0000
25	10	20	0,0936	0,2090	0,0000
26	10	17	0,0324	0,0845	0,0000
27	10	21	0,0348	0,0749	0,0000
28	10	22	0,0727	0,1499	0,0000
29	21	22	0,0116	0,0236	0,0000
30	15	23	0,1000	0,2020	0,0000
31	22	24	0,1150	0,1790	0,0000
32	23	24	0,1320	0,2700	0,0000
33	24	25	0,1885	0,3292	0,0000
34	25	26	0,2544	0,3800	0,0000
35	25	27	0,1093	0,2087	0,0000
36	28	27	0,0000	0,3690	0,0000
37	27	29	0,2198	0,4153	0,0000
38	27	30	0,3202	0,6027	0,0000
39	29	30	0,2399	0,4533	0,0000
40	8	28	0,0636	0,2000	0,0214
41	6	28	0,0169	0,0599	0,0065

Los límites de potencia activa, la localización de los generadores y los coeficientes de la función de costos de generación son tomados de [96]. Los límites de potencia reactiva se calculan a partir de las capacidades de P_{max} de los generadores en [95]. Las capacidades de rampa por tipo de combustible son modificadas de acuerdo con los lineamientos para el sistema de

Tabla C-2.: Límites de potencia y rampas para los generadores del sistema de potencia IEEE de 30 barras

Gen.	**Tipo**	P_{max}	P_{min}	Q_{max}	Q_{min}	$\mathbf{R}_{AGC}$	$\mathbf{R}_{10}$	$\mathbf{R}_{30}$	$\mathbf{R}_{Q}$
1	carbón	80	0	150,0	-20	0,56	5,60	11,20	170,0
2	hidro	80	0	60,0	-20	8,00	40,00	80,00	80,0
13	carbón	40	0	44,7	-15	0,28	2,80	5,60	59,7
22	carbón	50	0	62,5	-15	0,35	3,50	7,00	77,5
23	carbón	30	0	40,0	-10	0,21	2,10	4,20	50,0
27	carbón	55	0	48,7	-15	0,38	3,85	7,70	63,7
6	eólico	50	0	35,0	-25	5,00	25,00	50,00	60,0
7	dem_flex	0	-22,8	0	-10,9	2,80	11,40	22,80	10,9
15	dem_flex	0	-8,2	0	-2,5	0,82	4,10	8,20	2,5
21	dem_flex	0	-17,5	0	-11,2	1,75	8,75	15,50	11,2

P_{max} y P_{min} en MW
Q_{max} y Q_{min} en MVAR
R_{AGC}: rango de rampa para seguimiento de cargaAGC, en MW/min
R_{10}: rango de rampa para reservas de 10 minutos, en MW
R_{30}: rango de rampa para reservas de 30 minutos, en MW
R_{Q}: rango de rampa para potencia reactiva (escala de tiempo de 2 seg.), en $MVAr/min$

prueba usado en la comisión de unidades de organizaciones regionales de transmisión (RTO, por sus síglas en inglés) de FERC (*Federal Energy Regulatory Commission*) [100]. Las Tablas **C-2** y **C-3** contienen toda la información descrita.

Otra información fue tomada de Falsafi *et al.* [94] y corresponde a la demanda de potencia activa para un horizonte temporal de 24 horas (Tabla **C-4**); un perfil de viento base (Esc 1) del que se derivaron otros cinco escenarios con diferente distribución horaria (Tabla **C-5**); y la ubicación de las demandas flexibles en las barras 7, 15 y 21 con porcentajes de flexibilidad previsto para cada carga.

Tabla C-3.: Coeficientes de la función de costos de generación para los generadores del sistema de potencia IEEE de 30 barras

Generador Bus	a_i [\$/MW^2]	b_i [\$/MW]	c_i [\$]
1	0,0200	2,00	0
2	0,0000	1,00	0
13	0,0250	3,00	0
22	0,0625	1,00	0
23	0,0250	3,00	0
27	0,0083	3,25	0
6	0,0000	0,00	0
7	0,0000	0,00	0
15	0,0000	0,00	0
21	0,0000	0,00	0

Función de costos: $a_i P_{g_i}^2 + b_i P_{g_i} + c_i$

Tabla C-4.: Perfil horario de demanda para el sistema de potencia IEEE de 30 barras

Hora	**PD** [MW]	**QD** [MVar]	**Hora**	**PD** [MW]	**QD** [MVar]	**Hora**	**PD** [MW]	**QD** [MVar]
1	130,64	74,99	9	79,17	45,61	17	110,84	63,40
2	104,91	63,40	10	91,05	51,80	18	104,91	60,30
3	110,84	60,30	11	98,97	56,44	19	114,80	65,72
4	104,91	56,44	12	106,89	61,08	20	146,47	84,27
5	94,22	54,12	13	112,82	64,94	21	178,15	102,06
6	89,07	51,03	14	118,76	68,04	22	170,23	97,42
7	85,11	48,71	15	122,72	70,35	23	158,35	90,46
8	81,15	46,39	16	118,76	68,04	24	142,52	81,95

Tabla C-5.: Perfiles de viento para el sistema de potencia IEEE de 30 barras

Esc.	**Hora**							
	1	2	3	4	5	6	7	8
1	25,262	19,458	11,260	10,076	8,460	7,485	6,167	5,381
2	6,568	5,059	1,802	2,217	1,354	1,946	2,282	1,399
3	3.789	2,335	4,729	4,534	3,130	2,395	1,727	1,184
4	13,389	9,534	4,729	5,240	5,076	4,266	4,564	4,843
5	15,157	11,675	7,770	5,240	3,553	4,491	3,207	1,184
6	1,263	0,973	0,225	0,202	0,000	0,075	0,062	0,000

Esc.	**Hora**							
	9	10	11	12	13	14	15	16
1	4,665	4,332	6,588	8,665	10,657	12,534	14,619	16,287
2	1,726	2,252	3,689	4,852	6,394	6,518	8,771	9,772
3	0,187	0,433	0,791	1,386	2,025	2,005	1,900	1,303
4	4,012	3,292	5,468	5,979	8,526	11,281	13,450	15,635
5	3,219	3,162	5,007	5,979	8,100	10,779	13,888	14,984
6	0,047	0,000	0,461	1,906	4,796	8,398	10,380	7,818

Esc.	**Hora**							
	17	18	19	20	21	22	23	24
1	14,619	11,886	9,738	13,205	27,873	50,000	43,049	34,653
2	10,087	5,586	5,843	4,886	9,477	15,000	14,637	11,782
3	0,439	0,832	0,974	1,585	0,279	7,500	0,430	6,238
4	13,157	10,697	8,082	11,357	23,971	41,500	34,439	20,792
5	13,450	10,222	8,764	11,357	23,971	38,000	25,829	23,910
6	5,409	6,181	4,090	3,565	4,460	6,500	3,444	1,733

Bibliografía

[1] J. Lin and F. H. Magnago, "Desing, structure and operation of an electricity market," in *Electricity markets: Theories and applications*, I. Press, Ed. Wiley, 2017, ch. 7, pp. 173–209.

[2] P. Pinson, "Wind energy: Forecasting challenges for its operational management," *Statist. Sci.*, vol. 28, no. 4, pp. 564–585, Nov. 2013.

[3] Y. Wang, Z. Zhou, C. Liu, and A. Botterud, "Systematic evaluation of stochastic methods in power system scheduling and dispatch with renewable energy." [Online]. Available: https://www.osti.gov/biblio/1307654

[4] E. Ela, C. Wang, S. Moorty, K. Ragsdale, J. O'Sullivan, M. Rothleder, and B. Hobbs, "Electricity markets and renewables: A survey of potential design changes and their consequences," *IEEE Power and Energy Magazine*, vol. 15, no. 6, pp. 70–82, 2017.

[5] P. Denholm, E. Ela, B. Kirby, and M. Milligan, "The role of energy storage with electricity renewable generation," NREL, Tech. Rep, Tech. Rep., Jan. 2010. [Online]. Available: http://www.nrel.gov/docs/fy10osti/47187.pdf

[6] M. Kefayati and R. Baldick, "Harnessing demand flexibility to match renewable production using localized policies," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2012, pp. 1105–1109.

[7] E. Karangelos and F. Bouffard, "Towards full integration of demand-side resources in joint forward energy/reserve electricity markets," *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 280–289, 2012.

[8] M. Motalleb, M. Thornton, E. Reihani, and R. Ghorbani, "A nascent market for contingency reserve services using demand response," *Applied Energy*, vol. 179, pp. 985 – 995, 2016.

[9] Y. Degeilh and G. Gross, "Stochastic simulation of utility-scale storage resources in power systems with integrated renewable resources," *IEEE Transactions on Power Systems*, vol. 30, no. 3, pp. 1424–1434, 2015.

[10] E. Litvinov, F. Zhao, and T. Zheng, "Electricity markets in the United States: Power industry restructuring processes for the present and future," *IEEE Power and Energy Magazine*, vol. 17, no. 1, pp. 32–42, 2019.

[11] G. Martínez, J. Liu, B. Li, J. L. Mathieu, and C. L. Anderson, "Enabling renewable resource integration: The balance between robustness and flexibility," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2015, pp. 195–202.

[12] A. Papavasiliou and S. S. Oren, "Large-scale integration of deferrable demand and renewable energy sources," *IEEE Transactions on Power Systems*, vol. 29, no. 1, pp. 489–499, 2014.

[13] D. Kourounis, A. Fuchs, and O. Schenk, "Toward the next generation of multiperiod optimal power flow solvers," *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 4005–4014, 2018.

[14] W. A. Bukhsh, C. Zhang, and P. Pinson, "An integrated multiperiod OPF model with demand response and renewable generation uncertainty," *IEEE Transactions on Smart Grid*, vol. 7, no. 3, pp. 1495–1503, 2016.

[15] C. J. López-Salgado, A. Helseth, O. Añó, and D. M. Ojeda-Esteybar, "Stochastic daily hydrothermal scheduling based on decomposition and parallelization," *International Journal of Electrical Power Energy Systems*, vol. 118, p. 105700, 2020.

[16] I. Gomes, H. Pousinho, R. Melício, and V. Mendes, "Stochastic coordination of joint wind and photovoltaic systems with energy storage in day-ahead market," *Energy*, vol. 124, pp. 310 – 320, 2017.

[17] A. Banshwar, N. K. Sharma, Y. R. Sood, and R. Shrivastava, "Market based procurement of energy and ancillary services from renewable energy sources in deregulated environment," *Renewable Energy*, vol. 101, pp. 1390 – 1400, 2017.

[18] F. Bouffard and F. D. Galiana, "Stochastic security for operations planning with significant wind power generation," *IEEE Transactions on Power Systems*, vol. 23, no. 2, pp. 306–316, 2008.

[19] F. Liu, Z. Bie, S. Liu, and T. Ding, "Day-ahead optimal dispatch for wind integrated power system considering zonal reserve requirements," *Applied Energy*, vol. 188, pp. 399–408, feb 2017.

[20] H. Sharifzadeh, N. Amjady, and H. Zareipour, "Multi-period stochastic security-constrained OPF considering the uncertainty sources of wind power, load demand and equipment unavailability," *Electric Power Systems Research*, vol. 146, pp. 33–42, may 2017.

[21] A. J. Lamadrid, T. Mount, R. Zimmerman, C. E. Murillo-Sanchez, and L. Anderson, "Alternate mechanisms for integrating renewable sources of energy into electricity markets," in *2012 IEEE Power and Energy Society General Meeting*, 2012, pp. 1–8.

[22] N. Amjady, J. Aghaei, and H. A. Shayanfar, "Stochastic multiobjective market clearing of joint energy and reserves auctions ensuring power system security," *IEEE Transactions on Power Systems*, vol. 24, no. 4, pp. 1841–1854, 2009.

[23] V. Virasjoki, P. Rocha, A. S. Siddiqui, and A. Salo, "Market impacts of energy storage in a transmission-constrained power system," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 4108–4117, 2016.

[24] C. E. Murillo-Sánchez, R. D. Zimmerman, C. L. Anderson, and R. J. Thomas, "A stochastic, contingency-based security-constrained optimal power flow for the procurement of energy and distributed reserve," *Decision Support Systems*, vol. 56, pp. 1 – 10, 2013.

[25] J. Zhang, J. D. Fuller, and S. Elhedhli, "A stochastic programming model for a day-ahead electricity market with real-time reserve shortage pricing," *IEEE Transactions on Power Systems*, vol. 25, no. 2, pp. 703–713, 2010.

[26] M. Parastegari, R.-A. Hooshmand, A. Khodabakhshian, and A.-H. Zare, "Joint operation of wind farm, photovoltaic, pump-storage and energy storage devices in energy and reserve markets," *International Journal of Electrical Power & Energy Systems*, vol. 64, pp. 275 – 284, 2015.

[27] C. E. Murillo-Sánchez, R. D. Zimmerman, C. L. Anderson, and R. J. Thomas, "Secure planning and operations of systems with stochastic sources, energy storage, and active demand," *IEEE Transactions on Smart Grid*, vol. 4, no. 4, pp. 2220–2229, 2013.

[28] A. J. Lamadrid, D. Muñoz-Alvarez, C. E. Murillo-Sánchez, R. D. Zimmerman, H. Shin, and R. J. Thomas, "Using the Matpower Optimal Scheduling Tool to test power system operation methodologies under uncertainty," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 3, pp. 1280–1289, 2019.

[29] H. Wang, C. E. Murillo-Sanchez, R. D. Zimmerman, and R. J. Thomas, "On computational issues of market-based optimal power flow," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 1185–1193, 2007.

[30] F. Capitanescu, "Critical review of recent advances and further developments needed in ac optimal power flow," *Electric Power Systems Research*, vol. 136, pp. 57 – 68, 2016.

[31] T. D. Mount, A. J.Lamadrid, W. Y. Jeon, R. D. Zimmerman, and C. E. Murillo-Sánchez, "How will customers pay for the smart-grid," in *30th Annual Eastern Conference on Regulated Industries, Skytop*, Jan. 2011.

[32] A. J. Lamadrid, T. Mount, R. Zimmerman, and C. E. Murillo-Sanchez, "Harnessing the renewable generation potential," in *30th USAEE/IAEE North American Conference*, 2011.

[33] A. J. Lamadrid, D. L. Shawhan, C. E. Murillo-Sánchez, R. D. Zimmerman, Y. Zhu, D. J. Tylavsky, A. G. Kindle, and Z. Dar, "Stochastically optimized, carbon-reducing dispatch of storage, generation, and loads," *IEEE Transactions on Power Systems*, vol. 30, no. 2, pp. 1064–1075, 2015.

[34] C. Küchler and S. Vigerske, *Decomposition of Multistage Stochastic Programs with Recombining Scenraio Trees.* Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II, Institut für Mathematik, 2007.

[35] W. S. Sifuentes and A. Vargas, "Hydrothermal scheduling using benders decomposition: Accelerating techniques," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 1351–1359, 2007.

[36] M. d. P. Buitrago-Villada, S. García-Marín, J. E. Zuluaga-Orozco, and C. E. Murillo-Sánchez, "On the importance of using an ac or dc network model in the multi-period secure stochastic optimal power flow for settling a multidimensional day-ahead market," *IEEE Latin America Transactions*, vol. 19, no. 12, pp. 2003–2010, May 2021. [Online]. Available: https://latamt.ieeer9.org/index.php/transactions/article/view/4794

[37] A. Fuchs, J. Garrison, and T. Demiray, "A security-constrained multi-period OPF for the locational allocation of automatic reserves," in *2017 IEEE Manchester PowerTech*, 2017, pp. 1–6.

[38] A. Street, A. Brigatto, and D. M. Valladão, "Co-optimization of energy and ancillary services for hydrothermal operation planning under a general security criterion," *IEEE Transactions on Power Systems*, vol. 32, no. 6, pp. 4914–4923, 2017.

[39] J. Chen, T. D. Mount, J. S. Thorp, and R. J. Thomas, "Location-based scheduling and pricing for energy and reserves: a responsive reserve market proposal," *Decision Support Systems*, vol. 40, no. 3, pp. 563 – 577, 2005, challenges of restructuring the power industry.

[40] F. D. Galiana, F. Bouffard, J. M. Arroyo, and J. F. Restrepo, "Scheduling and pricing of coupled energy and primary, secondary, and tertiary reserves," *Proceedings of the IEEE*, vol. 93, no. 11, pp. 1970–1983, 2005.

[41] J. M. Arroyo and F. D. Galiana, "Energy and reserve pricing in security and network-constrained electricity markets," *IEEE Transactions on Power Systems*, vol. 20, no. 2, pp. 634–643, 2005.

[42] J. Wang, M. Shahidehpour, and Z. Li, "Contingency-constrained reserve requirements in joint energy and ancillary services auction," *IEEE Transactions on Power Systems*, vol. 24, no. 3, pp. 1457–1468, 2009.

[43] W. Wei, F. Liu, S. Mei, and Y. Hou, "Robust energy and reserve dispatch under variable renewable generation," *IEEE Transactions on Smart Grid*, vol. 6, no. 1, pp. 369–380, 2015.

[44] F. Bouffard, F. D. Galiana, and A. J. Conejo, "Market-clearing with stochastic security - part I: formulation," *IEEE Transactions on Power Systems*, vol. 20, no. 4, pp. 1818–1826, 2005.

[45] K. Van den Bergh and E. Delarue, "Energy and reserve markets: interdependency in electricity systems with a high share of renewables," *Electric Power Systems Research*, vol. 189, p. 106537, 2020.

[46] Y. Z. Li, K. C. Li, P. Wang, Y. Liu, X. N. Lin, H. B. Gooi, G. F. Li, D. L. Cai, and Y. Luo, "Risk constrained economic dispatch with integration of wind power by multi-objective optimization approach," *Energy*, vol. 126, pp. 810–820, 2017.

[47] Unidad de Planeación Minero Energética- UPME, "Plan de expansión de referencia generación-transmisión 2017-2031," pp. 1–381, 2018. [Online]. Available: https://www1.upme.gov.co

[48] R. D. Zimmerman and C. E. Murillo-Sánchez, " MATPOWER Optimal Scheduling Tool (MOST) User's Manual." 2020. [Online]. Available: https://matpower.org/docs/MOST-manual.pdf

[49] Gurobi Optimization LLC, "Gurobi Optimizer Reference Manual version 9.0.0." 2020. [Online]. Available: http://www.gurobi.com

[50] A. Wächter and L. T. Biegler, "On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming," *Mathematical Programming*, vol. 106, no. 1, pp. 25–57, 2006.

[51] Generadora y comercializadora de energía del Caribe - GECELCA S.A. E.S.P. (2014) Informe de gestión 2014. Accessed 2020-08-06. [Online]. Available: https://www.gecelca.com.co/_Descargas/_publico/_Transparencia/INFORME%20DE%20GESTION.pdf

[52] G. Cohen, *Optimisation des Grands Systèmes.* Cours du DEA MMME, Universitè de Paris I, 2004, pp. 1–116.

[53] M. A. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming. Theory and algorithms*, 3rd ed. Wiley-Interscience, 2006, vol. 1, pp. 257–298.

[54] D. P. Bertsekas, "Multipliers Methods: A Survey," *Automatica*, vol. 12, no. 7, pp. 135–145, 1976.

[55] A. Ruszczyński, "On convergence of an augmented lagrangian decomposition method for sparse convex optimization," *Mathematics of operations research*, vol. 20, pp. 634–656, 1995.

[56] M. R. Hestenes, "Multipliers and gradient methods," *Journal of Optimization Theory and Applications*, vol. 4, pp. 303–320, 1969.

[57] M. J. D. Powell, "A method for nonlinear constraints in minimization problems," *Optimization (R.Fletchr, ed.), Academic Press, New York*, vol. 4, pp. 283–298, 1969.

[58] G. Cohen and D. Zhu, *Decomposition Coordination Methods in Large Scale Optimization Problems: The Nondifferentiable Case and the Use of Augmented Lagrangians.* Advances in Large Scale Systems. JAI Press Inc., 1984, vol. 1, pp. 203–266.

[59] D. P. Bertsekas, *Nonlinear programming*, 2nd ed., 1999, ch. 4, pp. 201–205.

[60] A. J. Conejo, E. Castillo, R. Mínguez, and R. García-Bertrand, *Decomposition Techniques in Mathematical Programming. Engineering and Science Applications.* Springer, 2006, vol. 1, pp. 195–205.

[61] N. Redondo and A. Conejo, "Short-term hydro-thermal coordination by lagrangian relaxation: solution of the dual problem," *IEEE Transactions on Power Systems*, vol. 14, no. 1, pp. 89–95, 1999.

[62] P. Bento, S. Mariano, M. Calado, and L. Ferreira, "A novel lagrangian multiplier update algorithm for short-term hydro-thermal coordination," *Energies*, vol. 13, no. 24, pp. 728–742, 2020.

[63] W. Ongsakul and N. Petcharaks, "Fast lagrangian relaxation for constrained generation scheduling in a centralized electricity market," *International Journal of Electrical Power and Energy Systems*, vol. 30, no. 1, p. 46–59, 2008.

[64] X. Feng and Y. Liao, "A new lagrangian multiplier update approach for lagrangian relaxation based unit commitment," *IEEE Transactions on Power Systems*, vol. 34, no. 8, pp. 857–866, 2006.

[65] J. Benders, "Partitioning procedures for solving mixed-variables programming problems," *Numer. Math*, vol. 4, p. 238–252, 1962.

[66] A. Geoffrion, "Generalized Benders decomposition," *J Optim Theory Appl*, vol. 10, no. 4, pp. 237—260, 1972.

[67] H. Kim, S. Lee, S. Han, W. Kim, K. Ok, and S. Cho, "Integrated generation and transmission expansion planning using generalized Bender's decomposition method," in *2015 IEEE International Conference on Computational Intelligence Communication Technology*, 2015, pp. 493–497.

[68] Z. Li, W. Wu, B. Zhang, and B. Wang, "Decentralized multi-area dynamic economic dispatch using modified generalized Benders decomposition," *IEEE Transactions on Power Systems*, vol. 31, no. 1, pp. 526–538, 2016.

[69] N. Alguacil and A. J. Conejo, "Multiperiod optimal power flow using Benders decomposition," *IEEE Transactions on Power Systems*, vol. 15, no. 1, pp. 196–201, 2000.

[70] K. Chung, B. H. Kim, J. Lee, T. Oh, and J. Lee, "Transmission-security constrained optimal dispatch scheduling using generalized Benders decomposition," in *2009 Transmission Distribution Conference Exposition: Asia and Pacific*, 2009, pp. 1–4.

[71] M. Majidi-Qadikolai and R. Baldick, "A generalized decomposition framework for large-scale transmission expansion planning," *IEEE Transactions on Power Systems*, vol. 33, no. 2, pp. 1635–1649, 2018.

[72] A. Moreira, A. Street, and J. M. Arroyo, "An adjustable robust optimization approach for contingency-constrained transmission expansion planning," *IEEE Transactions on Power Systems*, vol. 30, no. 4, pp. 2013–2022, 2015.

[73] M. R. Ansari, N. Amjady, and B. Vatani, "Stochastic security-constrained hydrothermal unit commitment considering uncertainty of load forecast, inflows to reservoirs and unavailability of units by a new hybrid decomposition strategy," *IET Generation, Transmission Distribution*, vol. 8, no. 12, pp. 1900–1915, 2014.

[74] R. Rahmaniani, T. G.Crainic, M. Gendreau, and W. Rei, "The Benders decomposition algorithm: A literature review," Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation - CIRRELT, Tech. Rep., 2016.

[75] D. W. Watkins and D. C. McKinney, "Decomposition methods for water resources optimization models with fixed costs," *Advances in Water Resources*, vol. 21, no. 4, pp. 283 – 295, 1998.

[76] S. Trukhanov, L. Ntaimo, and A. Schaefer, "Adaptive multicut aggregation for two-stage stochastic linear programs with recourse," *European Journal of Operational Research*, vol. 206, no. 2, pp. 395 – 406, 2010.

[77] R. Rahmaniani, T. G. Crainic, M. Gendreau, and W. Rei, "The Benders decomposition algorithm: A literature review," *European Journal of Operational Research*, vol. 259, no. 3, pp. 801–817, 2017.

[78] R. Pacqueau, F. Soumis, and L. Hoang, "A fast and accurate algorithm for stochastic integer programming, applied to stochastic shift scheduling," École Polytechnique de Montréal, Tech. Rep. [Online]. Available: https://www.gerad.ca/en/papers/G-2012-29

[79] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*, 2nd ed. Springer-Verlag New York, 2011, vol. 1, pp. 198–202.

[80] J. F. Bonnans, J. C. Gilbert, C. Lemarechal, and C. A. Sagastizabal, *Numerical Optimization. Theoretical and Practical Aspects*, 2nd ed. Springer, 2006, pp. 137–154.

[81] J. Linderoth and S. Wright, "Decomposition algorithms for stochastic programming on a computational grid," *Computational Optimization and Applications*, no. 24, pp. 207 – 250, 2003.

[82] S. Zaourar and J. Malik, "Quadratic stabilization of Benders decomposition," *HAL - archives-ouvertes*, no. hal-01181273, pp. 1 – 27, 2014.

[83] G. Cohen, *Auxiliary problem principle and decomposition of optimization problems.* J Optim Theory Appl, 1980, vol. 32, pp. 277–305.

[84] A. V. Fiacco, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, 1st ed., August 1983, vol. 165.

[85] H. Ma and S. M. Shahidehpour, "Unit commitment with transmission security and voltage constraints," *IEEE Transactions on Power Systems*, vol. 14, no. 2, pp. 757–764, 1999.

[86] R. C. Green, L. Wang, and M. Alam, "Applications and trends of high performance computing for electric power systems: Focusing on smart grid," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 922–931, 2013.

[87] ——, "High performance computing for electric power systems: Applications and trends," in *2011 IEEE Power and Energy Society General Meeting*, 2011, pp. 1–8.

[88] S. K. Khaitan, "A survey of high-performance computing approaches in power systems," in *2016 IEEE Power and Energy Society General Meeting (PESGM)*, 2016, pp. 1–5.

[89] MATLAB, *Parallel Computing Toolbox, User's guide*, MatWorks, Inc., 2018.

[90] ——, *MATLAB Distributed Computing Server, System administrator's guide*, MatWorks, Inc., 2018.

[91] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "Matpower: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 12–19, 2011.

[92] R. D. Zimmerman and C. E. Murillo-Sánchez, "Matpower User's Manual version 7.0." 2019. [Online]. Available: https://matpower.org/docs/MATPOWER-manual-7.1.pdf

[93] Lawrence Livermore National Laboratory. (2021) Introduction to parallel computing tutorial. [Online]. Available: https://hpc.llnl.gov/training/tutorials/introduction-parallel-computing-tutorial

[94] H. Falsafi, A. Zakariazadeh, and S. Jadid, "The role of demand response in single and multi-objective wind-thermal generation scheduling: A stochastic programming," *Energy*, vol. 64, pp. 853–867, 2014.

[95] O. Alsac and B. Stott, "Optimal load flow with steady-state security," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-93, no. 3, pp. 745–751, 1974.

[96] R. Ferrero, S. Shahidehpour, and V. Ramesh, "Transaction analysis in deregulated power systems using game theory," *IEEE Transactions on Power Systems*, vol. 12, no. 3, pp. 1340–1347, 1997.

[97] P. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems*, 1998, ch. 1. Setting the Stage, pp. 1–17.

[98] J. Gondzio and A. Grothey, "Exploiting structure in parallel implementation of interior point methods for optimization," *Computational Management Science*, vol. 6, no. 2, pp. 135–160, 2009.

[99] R. D. Zimmerman and C. E. Murillo-Sánchez, "Multi-period SuperOPF (SuperOPF 2.0) User's Manual." 2013.

[100] FERC, "FERC RTO Unit Commitment Test System," Federal Energy Regulatory Commission, Tech. Rep.