UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

OPTIMIZACIÓN DE OPERACIONES DE COMPRAVENTA DE USD MEDIANTE EL USO DE METODOLOGÍAS DE APRENDIZAJE POR REFUERZO

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN TECNOLOGÍAS DE LA INFORMACIÓN

TOMÁS GERARDO ROJAS SALVO

PROFESOR GUÍA:
ALEXANDRE BERGEL

MIEMBROS DE LA COMISIÓN:
FELIPE BRAVO MÁRQUEZ
JÉRÉMY BARBAY
LUIS SILVESTRE QUIROGA

SANTIAGO DE CHILE
2022

Resumen

El presente proyecto de tesis tuvo por objetivo el estudio de metodologías de aprendizaje por refuerzo (RL) y la evaluación de su aplicabilidad a la toma de decisiones financieras, considerando los beneficios, dificultades y consecuencias de estas mediante una resolución algorítmica y automatizada.

El aprendizaje por refuerzo corresponde a un área del aprendizaje de máquina donde se calibra un comportamiento de forma autónoma en base a recompensas y penalizaciones experimentadas en estados de una realidad.

La problemática que se buscó analizar, y sobre la cual se aplicaron las metodologías estudiadas, consistió en la fragmentación de la forma más rentable[1] del volumen de compraventa de dólares estadounidenses en el mercado local por parte de una institución financiera, esto con el objetivo de mitigar el impacto de sus operaciones en el precio y así evitar pérdidas ocasionadas en posteriores transacciones efectuadas a un valor menos beneficioso.

Se modeló el precio del dólar[2](USDCLP) y se programó un agente algorítmico para ejecutar decisiones de compra y venta de la divisa, considerando los riesgos y retornos de las operaciones con el objetivo de maximizar su utilidad. Estas decisiones se calibraron en base a la experiencia y aprendizaje autónomo del agente, las que de forma acumulada permitieron determinar una política óptima de compraventa de dólares.

La política óptima de decisiones se generó mediante modelos basados en los métodos de Q-Learning y SARSA, algoritmos de RL a los que se les instauró el impacto de una decisión presente sobre un posible estado venidero.

Complementariamente se aplicó una metodología que permitiese generar precios futuros del USDCLP en base a fundamentos económicos y empíricos, como también una métrica que cuantificase el riesgo de mercado y su exposición frente a las decisiones del algoritmo.

Se ejecutó el modelo en un escenario hipotético, simplificado, durante un periodo de tiempo definido y con finita cantidad de dólares a ser transados, simulando una jornada de actividades financieras cotidianas con el objetivo de analizar su desempeño. El algoritmo proveyó una política óptima constituida por la cantidad de dólares a comprar o vender para cada configuración de tiempo, precio de la divisa y nivel de inventario de los estados remanentes.

De los resultados, estos demuestran la viabilidad del uso de metodologías de RL en la problemática bajo estudio. La algorítmica evidencia su capacidad de aprendizaje frente a estrategias convencionales al identificar decisiones favorables de forma consistente, alcanzando a su vez mayores beneficios. Si bien el modelo desarrollado corresponde a una prueba de concepto sobre su potencial aplicación a la industria financiera local, permite de forma favorable secundar las percepciones de una alternativa válida para disminuir los tiempos de análisis, como también explorar y descubrir decisiones desapercibidas por los analistas.

[1]Relación entre los beneficios obtenidos por un vehículo de inversión frente al costo de subscribirlo.

[2]El presente documento hará referencia única y exclusivamente al precio de la divisa estadounidense medida en pesos chilenos, a menos que se especifique una paridad de cambio diferente.

Tabla de Contenido

Capítulo 1

Introducción

1.1. Descripción de mercado del USDCLP

La divisa estadounidense corresponde a uno de los activos con mayor liquidez transados en el mercado local, con cifras que rondan los US$ 1.000 millones en transacciones diarias y donde Datatec representa a la fecha la plataforma con mayor volumen de operaciones en nuestro mercado, abarcando el 95 % de las negociaciones. Los principales bancos, fondos de inversión, compañías de seguros, empresas del sector real e instituciones de gobierno compran y venden USDCLP a través de dicho sistema.

Las transacciones de la divisa reciben el indicativo de dólar *spot*, haciendo referencia al dólar intercambiado en la mínima unidad operativa de tiempo por un pago en moneda local al contado, a diferencia de los *derivados*[3] de la moneda cuya contraparte recibe una promesa de pago futuro, una prima u otro medio de liquidación.

En la figura 1.1 se puede observar la elevada demanda por transacciones de USDCLP de los últimos 10 años, como también el nivel mostrado por la paridad de cambio.

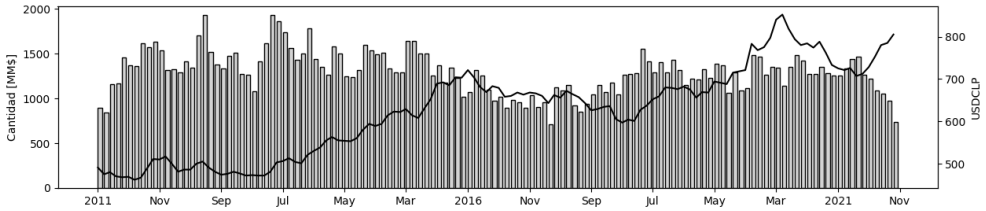

Figura 1.1: Media mensual de transacciones diarias, fuente: Datatec

Las altas variaciones de cotización de la divisa se explican por su sensibilidad a factores de mercado como el precio de la libra de cobre, la cotización de la canasta de dólares[4] a nivel

[3]Un derivado financiero corresponde a un vehículo de inversión que basa su valor en otro activo.

[4]La apreciación o depreciación del USD a nivel mundial se mide utilizando como referencia la paridad del dólar frente a un conjunto de monedas. El índice más referido corresponde al DXY.

mundial, el diferencial de tasas de interés con EE.UU. y la percepción de riesgo del mercado local y global[5].

Para el presente proyecto y la problemática a resolver se considera el dólar *spot*, aunque dicha solución podría extrapolarse a los *derivados* de la divisa.

1.2. Descripción del problema

En la actualidad nacional, los mecanismos de inversión son extensos y variados. Una de las alternativas más simples y accesibles corresponde al mercado de divisas, siendo el dólar estadounidense el producto más destacado al interior de éste. Se utiliza como intermediario en inversiones en el extranjero, así como también con un carácter plenamente especulativo.

Como intermediario, las personas compran y venden dólares para adquirir bienes y servicios nominados en dicha moneda con el fin que estos generen un retorno futuro. Desde una perspectiva especulativa, un individuo puede comprar dólares a la espera de un aumento en su valor y así, en una posterior venta, recibir una ganancia o experimentar una pérdida.

Tal como las personas corrientes compran y venden USD, lo hacen también las instituciones financieras, aunque a una velocidad, complejidad y volúmenes inmensamente superiores.

Un volumen de compra muy elevado impulsará inmediatamente su precio al alza, esto producto de una menor cantidad de dólares disponible para el resto de los interesados. De lo anterior, en sentido opuesto, será el caso de un volumen elevado de venta.

Un individuo común no tiene el poder suficiente de influir en este mercado, mientras que instituciones como bancos[6] y fondos de pensiones, al mantener inversiones por centenares de millones de dólares, presionan[7] inevitablemente el precio de esta divisa en sus operaciones de compra y de venta.

El precio de la moneda también es susceptible a los efectos de la economía local e internacional. Eventos políticos y económicos, publicación de indicadores financieros, cambios en las industrias de *commodities*, requerimientos normativos y decisiones de política monetaria son ejemplos de alteraciones en las expectativas de sus involucrados, las que se traducen en variaciones en la oferta y demanda por dólares, como en su valor.

La presente tesis busca abstraer los desafíos referidos a un algoritmo que dotase de un balance entre la contrapartida de beneficios y riesgos en la compraventa de la divisa mentada.

[5] El indicador más representativo corresponde al VIX, el que se desprende de una ponderación sobre la volatilidad implícita en contratos de opciones.

[6] Las instituciones bancarias ofrecen al mercado productos financieros denominados *forwards* de tipo de cambio. Estos consisten en una promesa de compra o venta futura de la moneda a un precio fijado en el presente. Este derivado obliga a la banca, por normativa de descalce cambiario, a compensar dichas operaciones con la adquisición o venta de la divisa previo al vencimiento del producto. Lo anterior, en tiempo de estrés financiero, presiona a los bancos a una vertiginosa demanda o liquidación de la moneda según corresponda.

[7] Estas imperfecciones de mercado, en sistemas con elevado número de transacciones, suelen corregirse conforme avanza una determinada cantidad de tiempo.

El problema por enfrentar consiste en estructurar la compra o venta de una cantidad finita y significativa de dólares en un horizonte de tiempo acotado, buscando reducir el impacto del volumen de la transacción en el precio y considerando los riesgos financieros de mercado. Lo señalado, con la finalidad de maximizar la rentabilidad de la operación.

En una compraventa financiera existe el riesgo que un volumen elevado presione el precio de un activo de forma perniciosa, transformándose en una dificultad para los participantes que continúan operando en la misma dirección. Debido a lo anterior, las instituciones son cautas en los volúmenes a comprar o vender para así evitar actuar de forma contraproducente. Para volúmenes lo suficientemente elevados, estás fragmentan los montos en distintas operaciones.

Esta problemática se aborda en la actualidad mediante heurísticas, donde los riegos inherentes a cada posible combinación de decisiones, la exposición a fluctuaciones de mercado y el impacto de las acciones en escenarios futuros eleva la complejidad en su resolución.

La solución de este problema consiste en una política óptima de decisiones, es decir, en la cantidad precisa de dólares a comprar o vender para cada configuración de estados, siendo estos la combinación de tres elementos: el precio del USDCLP, la cantidad restante de USD a ser comprada o vendida y el tiempo remanente para culminar las operaciones.

La problemática se origina en instituciones financieras de alto patrimonio. Éstas deben cumplir con objetivos de compra o venta de dólares en el transcurso de un periodo de operación, esto con el fin de satisfacer la demanda de clientes por la moneda, métricas y necesidades internas, adquirir activos nominados en dicha divisa, cumplir con normativas y descalces cambiarios, entre diversas actividades que involucran al tipo de cambio.

Lo anterior establece un marco temporal en la adquisición de dólares, donde la mesa de dinero de la institución respectiva, o un símil equivalente, debe adquirir o liquidar una cantidad determinada de la divisa al precio más rentable[8].

El analista financiero se enfrenta entonces a la problemática de decidir la cantidad y el momento idóneo en la compra o venta de dólares, sujeto a la restricción de disponer de un tiempo finito para ejecutar el número de operaciones necesarias. Este inicia su jornada con un determinado monto a comprar o vender, el que deberá al término del período haber comprado o vendido al valor más beneficioso en su totalidad.

Determinar el momento óptimo de compra y de venta supone una elevada dificultad debido a la diversidad de factores que influyen en el precio de un activo financiero. Estos se pueden dividir en factores conocidos como fundamentales, es decir, en aquellos alineados a los fundamentos de la economía y en factores operativos, vinculados a necesidades de los participantes que alteran la liquidez y disponibilidad de la divisa[9].

El analista financiero puede decidir diversas estrategias para cumplir con el objetivo. Este puede fraccionar los instantes en los que opera y con esto reducir los volúmenes de compra y de venta. Una mayor cantidad de operaciones permitirá reducir el volumen transado y, por

[8]La rentabilidad se mide por periodos y se ajusta a los ciclos económicos y financieros. De todas formas, en su expresión más simple, aspira a la venta al máximo valor de la jornada y a la compra en su mínimo valor.
[9]Uno de los mayores exponentes de este factor consiste en la problemática de impacto de mercado, donde un agente financiero influye, por el elevado volumen de venta y compra de un activo, en el precio de este.

consiguiente, un menor impacto en el precio. Lo anterior no se materializa sin un costo, siendo este el riesgo que la divisa se revalorice por variaciones de mercado de forma desfavorable en el intertanto.

También, puede escoger la compra o venta de dólares en un reducido número de operaciones, mitigando el riesgo de mercado, aunque incrementando el riesgo de impacto del volumen en el precio, donde cada una de las consecuentes compras o ventas por parte de la institución podrían realizarse a un precio menos fructuoso.

Cada una de estas estrategias involucran riesgos y potenciales retornos en el transcurso de la jornada de operación.

El enrevesado consiste en la limitante para un individuo de poder sistematizar dichas estrategias, seleccionando el conjunto de decisiones óptimas en base a una resolución estructurada. La cantidad de elementos involucrados en el proceso de decisión dificulta el poder trazar extensamente las implicancias de una decisión, lo que conlleva a ejecutar acciones subóptimas por parte de las instituciones.

Para enfrentar el problema señalado, se hace uso de un modelo estocástico el cual permite simular la trayectoria del USDCLP. Sus datos son utilizados de entrada a la algorítmica de aprendizaje por refuerzo, permitiendo calibrar un aprendizaje que proporcione las decisiones idóneas de compraventa de dólares según una abstracción de riesgo de mercado y considerando los efectos de volumen en los precios.

El presente trabajo tiene por objetivo el desarrollo de un tomador de decisiones capaz de ejecutar un conjunto óptimo de acciones según su percepción de los datos disponibles. La exhaustiva precisión en la predicción del USDCLP y en la cuantificación de riesgos de mercado se encuentra fuera del alcance de esta tesis, abocándose a modelos estándar de la economía financiera y sirviendo así, como recursos para evaluar la viabilidad de las técnicas de aprendizaje por refuerzo en la problemática atingente.

1.3. Justificación del trabajo de tesis

El presente proyecto de tesis recibe justificación en la exploración de alternativas que permitan fortalecer los análisis económicos referente a los riesgos y beneficios inmersos en operaciones financieras. Si bien el trabajo aborda la particularidad de un problema vinculado al mercado de divisas, siendo este representado por la paridad USDCLP, su estudio y análisis puede ser extrapolado a diversos desafíos que involucren una compensación entre retornos y riesgos sobre decisiones sucesivas.

La inmensa cantidad de datos en un cúmulo de relaciones endógenas y exógenas hacen de los análisis financieros y económicos un reto cada vez más difícil para los analistas, donde la exploración de nuevas metodologías aquieta una necesidad dominante en la industria.

Adicionalmente, la constante convivencia con la aleatoriedad de los elementos financieros requiere de métodos que permitan incorporar condicionales y eventos probabilísticos en su resolución, características propias de los algoritmos de aprendizaje por refuerzo. A lo anterior

se suma la fortaleza de estos métodos en la exploración secuencial de soluciones, donde el descubrimiento de reglas y estrategias óptimas de negocio concluyen en resultados significativamente atractivos e incluso, fuera de la previsibilidad humana.

Se espera que técnicas como las propuestas en la presente tesis cobren cada vez más auge en la industria, llevando así los análisis a un nivel más robusto e incluso, autónomo.

1.4. Resultados esperados

Desde la perspectiva técnica se persigue un resultado correspondiente a la cantidad de dólares a ser vendidos o comprados para una configuración de precio del USDCLP, un respectivo saldo de dólares y un remanente de tiempo para la completitud del objetivo. Es imperante que los resultados sean proporcionados de forma sencilla y efectiva, orientándose a una representación matricial o en forma de grilla que permita resumir y relacionar las decisiones a sus estados respectivos.

Referente al desempeño del algoritmo propuesto, este se contrasta con una situación base comparativa fundada en una heurística común y recurrente del mercado financiero. Se espera que el rendimiento de la mecánica de aprendizaje por refuerzo exceda a su contraparte basal, demostrando la habilidad de aprender de las señales recibidas y plasmar el conocimiento adquirido en decisiones coherentes.

Debido a que la orientación del presente documento se encuentra en evaluar la factibilidad del aprendizaje por refuerzo como alternativa a las metodologías actuales de trabajo, con focalización en los métodos de RL más no en el perfeccionamiento de los procedimientos económicos y financieros, es previsible un desempeño abierto a espacios de mejora en la interpretación del sistema sobre el que se desenvuelve. Una evaluación positiva de los resultados se concentra en identificar la capacidad de aprender del agente ficticio, demostrando mejoras progresivas y sustanciales en su desempeño frente a la iteración de las simulaciones. Frente a un cumplimiento de lo anterior, se subentiende una base sólida sobre la cual una mejora en la calidad de métricas económicas y financieras traerán consigo un superior entendimiento de las señales del mundo hacia el agente algorítmico y por consiguiente, decisiones más coherentes con su rendimiento.

Alusivo a los beneficios esperados en el corto plazo sobre el presente estudio y solución se encuentran:

- Concluir si las técnicas de aprendizaje por refuerzo son una alternativa promisoria para el problema abordado.
- Disponer de un mínimo producto viable (MVP) que permita interactuar y comprender los distintos elementos involucrados, como también servir de soporte para desarrollos futuros.

En el mediano y largo plazo se espera que soluciones como las presentadas en este documento permitan:

- Habilitar aplicaciones sencillas en su operación y que concedan rápidas recomendaciones de decisión.
- Complementar y fortalecer los estudios financieros de una mesa de dinero.
- Proporcionar métodos que automaticen procesos de análisis.
- Preceptuar estructuras lógicas con el objetivo de inferir conocimiento.

1.5. Objetivo principal y específicos

El objetivo principal de esta tesis consiste en aplicar y evaluar las metodologías de aprendizaje por refuerzo Q-learning y SARSA para determinar una política óptima de compra y venta de dólares en un horizonte temporal finito. Una política de compraventa corresponde para este ejercicio, como se ha señalado, a la cantidad de dólares a ser comprados o vendidos según el precio de la divisa, la cantidad de USD restante por adquirir o ser liquidados y el tiempo restante para completar dicha meta.

Para el desarrollo de este proyecto se persiguen los siguientes objetivos específicos:

- La aplicación de un modelo de generación de precios de USDCLP que incorpore los efectos de mercado, la aleatoriedad de estos y las decisiones de un algoritmo. Los grandes volúmenes de compra o venta presionan los niveles de precio de la divisa, efectos que deben ser capturados en la valorización de esta misma.
- El desarrollo de una función de recompensa que permita representar de forma objetiva los beneficios y riesgos en las decisiones efectuadas por el algoritmo. El postergar la compra o venta de dólares trae consigo el riesgo de apreciaciones o depreciaciones de la moneda producto de eventos económicos ajenos a las decisiones de los operadores de los mercados financieros.
- El modelamiento de los distintos estados cohesionados con la realidad. Una de las mayores dificultades de este objetivo se atribuye a la continuidad del precio del USDCLP y su expresión en estados discretos.
- La implementación de métricas de evaluación sobre las metodologías de aprendizaje por refuerzo que posibiliten discriminar la efectividad de estos. Definir una situación basal comparativa y contrastar los resultados con respecto a dichos valores.

1.6. Sustento de la metodología de aprendizaje por refuerzo

El objetivo modelado presenta sus bases en decisiones estratégicas sucesivas realizadas por un analista financiero a través de heurísticas, las que son complementadas mediante el uso de análisis estadísticos, numéricos, económicos y financieros.

Lo anterior proporciona atractivo en la metodología de aprendizaje por refuerzo ante su lógica en la resolución de problemas mediante mecánicas secuenciales. Su origen se remonta a la conclusión exacta de modelos de programación dinámica, los que, debido a su complejidad y la necesidad de técnicas numéricas en su desarrollo, acercaron su lógica a procedimientos *bootstraping* o recursivos de solución.

Esto último presenta el beneficio de incorporar en los modelos causalidad de eventos y relaciones no lineales entre sus componentes, significativas ventajas al momento de describir comportamientos económicos. A esto se suma la flexibilidad que un modelo RL presenta, en conjunto con la rapidez en su resolución y la capacidad por describir un inconmensurable número de decisiones y seleccionar la más fructuosa.

1.7. Elementos complementarios

La mecánica de aprendizaje por refuerzo requiere el poder interpretar el entorno, donde para lograrlo hace uso, como será detallado más adelante, de estados y de una función de recompensa.

Para la problemática, esta interpretación requiere de la inferencia del precio del USDCLP, como a su vez de métricas de riesgo y de impacto en las órdenes de compra que permitan al algoritmo interpretar la realidad que se busca modelar.

Para la estimación del precio del USDCLP se utiliza un modelo de no arbitraje fundado en las dinámicas de un movimiento browniano geométrico (GMB). Su elección se sustenta en su robustez teórica, su acotado número de parámetros y su capacidad para incorporar aleatoriedad [10] en la estimación de precios, siendo este un elemento significativo en las dinámicas estudiadas. El comportamiento estocástico del dólar frente a cualquier moneda o unidad de valorización corresponde a una de las limitantes que hacen de su estimación una actividad excesivamente compleja. Por estocástico se infiere que su comportamiento futuro considera al azar como elemento intrínseco del sistema, donde modelos basados en el GMB permiten describir trayectorias considerando dicha aleatoriedad.

En la estimación del riesgo financiero sobre una decisión de compra o de venta se aplica la métrica del Valor en Riesgo, destacada por su simplicidad y robustez financiera. Su uso permite asignar un valor a la pérdida esperada sobre una decisión ejecutada.

[10] El modelo GBM seleccionado incorpora una única fuente de incertidumbre, siendo esta la responsable de abarcar la completa aleatoriedad de mercado.

Finalmente para interpretar el impacto de una orden de compra se hace uso de la regla empírica de la *raíz cuadrada*, correspondiendo a una relación empleada con frecuencia en la industria producto de su sencillez y efectividad, a pesar de carecer de fundamentos teóricos robustos.

1.8. Plan de trabajo

A continuación, se detalla el plan de trabajo ejecutado para alcanzar los objetivos indicados.

- Se realizó una exploración de la mecánica de trabajo de distintas mesas de dinero del mercado nacional.
- A partir de lo anterior se establecieron los *elementos esenciales de la solución* descritas en numeral 2.4 de este documento.
- Se investigaron propuestas similares en la industria y su complementariedad con los métodos de aprendizaje por refuerzo a ser analizados.
- Se desarrolló un MVP que contuviese una base de modelación del USDCLP, la capacidad de inferir los conceptos de riesgo-beneficio y la noción de impacto de una orden de compra.
- Se evaluó la capacidad de aprendizaje del algoritmo, como también la interpretabilidad y calidad de los resultados entregados.

1.9. Trabajos relacionados

La problemática de ejecuciones de tamaño óptimo de operaciones financieras ha sido investigada por diversos autores y a través de distintas metodologías.

Bertsimas y Lo (1998) dieron un inicio formal a estos estudios a través del uso de programación dinámica para encontrar una solución de forma cerrada en la minimización de costos de ordenes de compraventa sobre un periodo definido. Investigaciones posteriores extendieron dichos estudios, incorporando abstracciones de la realidad económica financiera con mayor nivel de complejidad, destacando las publicaciones de Huberman y Stanzl (2001, 2005).

Estos trabajos establecieron sólidas bases de modelamiento sobre escenarios financieros de incertidumbre, riesgo y competitividad, aunque sin estar ausentes de las dificultades que representa el modelar sistemas de elevada complejidad a través dichas metodologías, como la demanda de recursos computacionales.

La aplicación de técnicas de aprendizaje por refuerzo ha sido una evolución atractiva sobre estas investigaciones. Nevmyvaka, Feng, y Kearns (2006) profundizaron estos análisis al publicar uno de los primeros trabajos investigativos sobre esta temática, incorporando la metodología de RL sobre una escala de milisegundos en las transacciones de la bolsa de valores NASDAQ.

En el periodo contemporáneo a la elaboración de la presente tesis, la incorporación de métodos de aprendizaje profundo ha sido la siguiente orientación en la revisión de nuevos análisis sobre la optimización de operaciones financieras. Estos últimos permiten abordar escenarios con mayor número de dimensiones e interactividad, paradigmas de los sistemas financieros. Ning, Lin, Jaimungal (2018) perfeccionaron modelos Deep Q-Network, los que incorporan técnicas de Q-Learning y aprendizaje profundo en el modelamiento del problema, obteniendo resultados satisfactorios y validando estas técnicas como una alternativa fehaciente.

Capítulo 2

Marco teórico

2.1. Recapitulación

Modelar los mercados financieros se ha transformado en una actividad creciente y dinámica. La velocidad con la que muchos productos reaccionan a los distintos eventos económicos y sociales, en conjunto con el vertiginoso aumento del acceso a datos y la complejidad de una gran diversidad de activos han llevado a la industria a una búsqueda persistente de metodologías que permitan fortalecer los análisis.

Durante el siglo XX, el desarrollo de modelos matemáticos y estadísticos que permitiesen estructurar eventos probabilísticos para medir riegos, relaciones entre los productos financieros y precios de mercado observó uno de sus mayores apogeos. Ilustrados exponentes como Fama, Itō, Feynman, Sharpe, Markowitz, Black, Scholes, Merton, Jorion entre muchos otros ampliaron el desarrollo de las matemáticas, física y estadística a la aplicación financiera.

Con el pasar del tiempo, nuevas corrientes complementarias de análisis se fueron masificando, destacando el análisis fundamental y el análisis técnico.

- El análisis técnico se orienta a predecir tendencias y describir patrones considerando la serie histórica de precios. Su estructura depende de la hipótesis de mercados eficientes, la que supone que cualquier fluctuación de mercado que involucre a un activo se incorporará inmediatamente en su precio.
- El análisis fundamental por su parte persigue el identificar el verdadero valor de un activo financiero incorporando los eventos económicos, coyunturales, contables y financieros que puedan afectar la relación de su oferta o demanda.

Con el desarrollo de la computación, algoritmos de *trading* fueron incrementando sus adeptos. La implementación de reglas precisas en la ejecución de órdenes de compraventa facilitó la automatización de operaciones en función de señales especificas en indicadores de mercado.

A medida que la rapidez de cálculo de los computadores fue alcanzando niveles cada vez más significativos en la industria, también lo fue en los algoritmos, dando paso al *trading* de alta

frecuencia y donde la escala de tiempo de las negociaciones se orientó a los microsegundos.

En la actualidad, otros de los hitos ha sido la incorporación de técnicas de aprendizaje de máquina (ML) a las problemáticas financieras. Las técnicas de ML decantan en los mercados financieros por análisis y metodologías no supervisadas. Lo anterior debido a su capacidad de relacionar distintos componentes de la industria y describir patrones de comportamiento de sus participantes y productos. Por su parte las metodologías supervisadas se utilizan en menor medida debido a la necesidad de contar con un conocimiento *a priori* estable y consistente sobre el cual calibrar un aprendizaje, mientras que en la práctica los eventos financieros no son del todo consistentes y en muchos casos poco repetitivos a cabalidad.

Por último, la implementación de técnicas de aprendizaje por refuerzo ha ganado popularidad en la resolución de problemáticas secuenciales financieras, esto debido a la naturalidad en la implementación de relaciones entre periodos de tiempo e incorporando elementos metodológicos semejantes, como factores de descuento y costos de oportunidad.

El aprendizaje por refuerzo permite una gran flexibilidad, donde los distintos avances numéricos en materia de modelamiento financiero pueden ser incorporados a un tomador de decisiones automatizado con el objetivo de optimizar y resolver complejas reglas de negocio.

2.2. Situación base y aspectos de mercado

Las divisas son compradas o vendidas por instituciones financieras en un mercado regulado a través de plataformas de negociación como lo son las bolsas de comercio. Lo esencial en su operativa se encuentra en el mecanismo de órdenes de compraventa. Este registra los precios ofertados por el comprador de la moneda y los precios demandados por el tenedor de esta. En la jerga se identifica como BID al precio más alto que el comprador está dispuesto a pagar y ASK al precio más bajo al que el vendedor está dispuesto a vender.

En un libro de ordenes tradicional se clasifican las distintas ofertas y demandas según su precio, concretando la operación cuando el BID y el ASK acuerdan el mismo valor. Mientras esto no se produzca, la diferencia entre estos valores se identificará como BID-ASK *spread* y denota una métrica de la liquidez del mercado.

La única posibilidad para un comprador de conseguir la divisa de forma inmediata radica en ofrecer un precio igual al valor ASK. De lo contrario su operación se mantendrá retrasada y posicionada según su valor ofertado, a la espera que el resto de las ordenes por delante culminen.

De igual forma un oferente de la moneda podrá venderla de forma inmediata si ofrece un precio igual al valor BID, de lo contrario su operación también se mantendrá en espera. Ambas partes pueden corregir los precios ofertados y demandados con el fin de mejorar su posición en el libro de ordenes o concretar la operación de forma inmediata.

Lo señalado fuerza que las operaciones se concreten de forma secuencial, lo que facilita la adopción de un modelo de aprendizaje por refuerzo en su formulación, como será detallado en los siguientes capítulos. En la práctica, los conceptos de BID y el ASK pueden ser simplificados a la acción de compra o de venta realizada por el agente financiero, mientras estas se desarrollen de forma ordenada y sucesiva.

Con la claridad del contexto en que las ofertas de compraventa se realizan en un orden sucesivo, el problema a abordar tiene su origen en el impacto en el volumen de dichas órdenes.

El primer elemento por resaltar corresponde a la cantidad de unidades a ser vendidas. Para volúmenes elevados, la posibilidad de concretar el monto de una compra o venta en una sola operación es reducida. Es de elevada dificultad encontrar a una contraparte que requiera comprar o vender volúmenes significativos en el momento requerido, por lo que dicha operación es materializada en distintas órdenes del libro de compraventa.

Señalado esto y para detallar la problemática, los extremos situacionales se enfrentan a los siguientes elementos favorables y desfavorables.

- Segmentar un elevado volumen de compra o venta en reducidas operaciones significará reducir la exposición a factores de mercado que pudiesen mover el precio de forma adversa. Este beneficio ocurre a expensas de presionar el precio de la divisa al introducir una elevada cantidad monetaria al sistema en una acotada unidad de tiempo, donde las restantes operaciones a tranzar deberán hacerse con un efecto reductor en el rendimiento proporcionado.

- Por el contrario, ventas segmentadas en numerosas operaciones de reducido volumen tendrán el beneficio de un nulo impacto en el precio como consecuencia de la operación. Pero lo anterior incrementará la exposición al riesgo de un movimiento desfavorable producto de las fluctuaciones de mercado en el transcurso que las distintas operaciones secuenciales se materializan.

Los analistas desarrollan estrategias para segmentar la cantidad de operaciones y volúmenes con el objetivo de mitigar la exposición al riesgo de mercado y al impacto del precio de la divisa producto de la acción del mismo participante.

Cada participante actúa de forma racional y táctica con el objetivo de reducir sus riesgos y aumentar sus beneficios. Para ello, utilizan relaciones numéricas y estadísticas para aproximar el efecto de un determinado volumen en el precio. El análisis de series de tiempo, de modelos matemáticos, el análisis técnico y fundamental asisten a la estimación del precio de la divisa.

El agente financiero no solo debe efectuar lo anterior para la primera operación, sino que debe repetir el procedimiento para cada una de estas que decida segmentar. Es de uso habitual la resolución del efecto del impacto en cada decisión mediante heurística, limitando la trayectoria de decisiones a un conjunto reducido y en muchos casos ineficiente.

Al considerar la amplia cantidad de combinaciones de estados se vislumbra la significativa dificultad en la selección del conjunto idóneo de decisiones. Debido a lo anterior, se propone la aplicación de técnicas de aprendizaje por refuerzo que permitan la selección de aquellas decisiones adecuadas para una combinación de factores determinada, considerando los riesgos y beneficios de estas y su impacto en decisiones en el espacio muestral de acciones futuras.

2.3. Solución desarrollada

La presente tesis busca determinar una política óptima de decisiones. Una política de decisión estriba en la acción a seleccionar para cada estado del problema. Como se indicó y será detallado, los estados particulares de esta problemática se componen por una determinada cantidad restante de USDCLP a ser comprados o vendidos, su precio unitario y el tiempo remanente para realizar dichas operaciones.

Una política corresponde entonces a la cantidad de USDCLP a seleccionar para vender o comprar para cada nivel de precio de dólar, considerando a su vez la cantidad de dólares faltantes para completar el objetivo y los minutos, horas o días que se disponen.

La unidad de tiempo permite definir de forma natural los distintos estados señalados y por consiguiente la estructura de solución del problema. Producto de la existencia de centavos en la cotización del precio de la divisa y de la profundidad del mercado en la que esta habita, como también para facilitar el análisis, el tiempo se discretiza en unidades pequeñas de observación. De lo anterior, la política proporciona la cantidad a comprar o vender de la moneda para intervalos de tiempo finito restantes para completar la meta, los que se ajustan a conveniencia del análisis.

Finalmente, la solución corresponde, dentro de todas las políticas factibles, a aquella que cumpla con los criterios de optimalidad. Estos últimos consideran los riesgos y beneficios de cada decisión que conforman a la política óptima, tanto de forma local como global, siguiendo la resolución de las metodologías Q-Learning y SARSA.

En la práctica, lo anterior permite disponer de una función que relacione cada configuración de estado posible con la acción óptima que se deba ejecutar, para así obtener el mayor beneficio acumulado.

La solución desarrollada analiza un espacio factible y discreto de posibles precios futuros del dólar en pesos. Sobre estos explora diversas secuencias y combinaciones de decisiones de compra o venta de la divisa, considerando el impacto del volumen en el precio como también el riesgo futuro de cada acción ejecutada.

Procede posteriormente a ponderar los beneficios de cada secuencia de decisión, referida como política, y determinar así aquella trayectoria más beneficiosa. El conjunto de acciones perteneciente a esta última corresponderá a la representación de la solución.

A modo de simplificación de la algorítmica, considerar al agente ficticio en un estado particular P con valor del USDCLP de $ 700. Suponer que dicho estado se vincula de forma sucesiva a un estado A con valor de la divisa en $ 702 y a un estado B con valor en $ 699. Los estados A y B forman parte de distintas trayectorias convergentes en P, ramificándose producto de los efectos de mercado y en complemento con la acción ejecutada en dicho nodo. Triadas equivalentes y enlazadas se repetirán de forma exhaustiva una vez iniciada la mecánica y hasta finalizar las simulaciones, calibrándose diferentes trayectorias y discriminando aquella más beneficiosa, identificada como solución de la problemática.

Lo anterior será registrado según su distribución de probabilidad de ocurrencia para los niveles de inventario de dólar y tiempo residual correspondiente, permitiendo así almacenar un significativo número de configuraciones y seleccionar las acciones más rentables durante el desarrollo de la simulación.

En 4.1 se procederá a detallar las dinámicas presentes en la solución con una simplificada trayectoria de ejemplificación.

2.4. Integración de los elementos esenciales de la solución

Como ya fue indicado, la solución propuesta se basa en la combinación de cuatro elementos: Un algoritmo de aprendizaje por refuerzo a través de las metodologías de Q-Learning y SARSA, un modelo de estimación de precios de USDCLP derivado del movimiento browniano geométrico, una relación empírica del impacto del volumen transado en el precio de la moneda y una abstracción de riesgo financiero medido a través del Valor en Riesgo (VaR).

La integración de estos elementos se adecua de la siguiente forma. El modelo de estimación de precios calibra sus parámetros a la realidad coyuntural regente con el objetivo de volver

fidedignas sus salidas. Estas últimas sirven de entrada al algoritmo de RL, describiendo distintos estados sobre el cual recoger aprendizaje.

Los modelos financieros basados en el movimiento browniano geométrico incorporan en su estructura distribuciones de probabilidad que describen la posibilidad que un nivel de precio se materialice. Esto último es recibido como entrada por el modelo de RL, adecuando así su estrategia.

El agente ficticio, representado por el algoritmo de aprendizaje por refuerzo, responde a incentivos para ejecutar sus decisiones. Estos son definidos por una función de recompensa que se ajusta en relación con la medida entregada por el VaR. Este último cuantificará la exposición al riesgo de mercado y financiero identificado las posibles pérdidas monetarias que puede incurrir por cada unidad de USDCLP que no compre o venda en el estado actual, esto ante posibles movimientos contraproducentes en el precio de la divisa por concepto de volatilidad de mercado.

Complementando a lo anterior, el modelo proporciona *feedback* a los estados de acuerdo con la acción que ejerza sobre éstos. Según la cantidad de dólares vendidos o comprados este podrá incidir en el precio de la divisa. Lo anterior se gestiona por una relación empírica de mercado, permitiendo ajustar el precio teórico proporcionado en función de las decisiones de la mecánica de RL.

La sinergia entre los distintos componentes permite al algoritmo experimentar una elevada cantidad de complejos escenarios sobre los que recolectar un aprendizaje. El agente ficticio se enfrenta a múltiples trayectorias de sucesos donde percibirá el impacto de sus decisiones al culminar cada una de estas. Esto le permite discriminar y seleccionar las mejores decisiones para cada configuración del sistema, las que identificará como la política óptima.

2.5. Totalidad de un tipo de operación a ser modelado

Se considera en el problema a modelar que las operaciones sean únicamente de compra o venta. Esto quiere decir que el modelo permita evaluar para un determinado monto de USD objetivo la totalidad de las operaciones en una única dirección, ya sea de compra o de venta.

Para una institución que busca comprar una determinada cantidad de dólares[11], una acción de venta será contraproducente para el objetivo de completar la cantidad requerida, con excepción que este decidiese especular sobre la trayectoria del USDCLP y asumir o cubrir parte del riesgo frente a una ganancia potencial. Esta última propiedad no es incorporada en la algorítmica del aprendizaje por refuerzo, esto es para enfocar la toma de decisión a lo fundamental del impacto de volumen en las decisiones.

2.6. Supuesto de simetría en las operaciones de venta y compra

En la presente tesis, modelar operaciones de compra o de venta no difieren en su metodología. En la realidad, este supuesto no se mantiene a cabalidad. La volatilidad de los activos aumenta conforme a escenarios de pérdida, evento que se exacerba y se observa en mayor cuantía durante operaciones de venta [12].

La aplicación de esta simetría se hace coherente con el objetivo que se busca alcanzar, pudiendo prescindir de los efectos que vuelven característica una operación de venta frente a una de compra de dólares y mantener un modelo robusto.

[11] Equivalente comportamiento ocurre en la direccionalidad opuesta.

[12] Esto último se ha visto mitigado con la incorporación de derivados con fines especulativos y de cobertura, acercando la capacidad para tomar posición en ambas direccionalidades.

Capítulo 3

Concepción de la solución

3.1. Requisitos y restricciones a la solución

El entorno de trabajo de una mesa de dinero es dinámico y donde el análisis se desarrolla a un ritmo vertiginoso. La incertidumbre es un factor preponderante con el cual los analistas financieros deben convivir de forma constante. Debido a lo anterior, y complementado con los sondeos realizados en el plan de trabajo, se agrupan los siguientes elementos como requisito de la solución a analizar.

- Estrategia: El modelo debe proveer una estrategia de compraventa según los datos proporcionados de forma coherente y donde se perciba un aprendizaje continuo.
- Fiabilidad: Los métodos deben ser consistentes, robustos y resilientes a las condiciones de mercado y ser capaces de proporcionar recomendaciones que mejoren la capacidad analítica vigente.
- Rapidez: La algorítmica debe ser rápida en su calibración y ejecución. El mercado financiero se mantiene constantemente en actualización, por lo que la metodología debe ser capaz de incorporar nuevos datos y generar resultados fiables con una demora reducida.
- Flexibilidad: Símil al requerimiento previo, la solución propuesta debe poder integrar nuevos elementos que permitan enriquecer el análisis. Complementario a la velocidad de actualización de la industria financiera, la amplia variedad y nueva incorporación de indicadores relevantes deben permitir al modelo adaptarse de forma sencilla y fluida.
- Táctica: La solución debe ser capaz de considerar una amplia cantidad de escenarios, reglas y el impacto de las acciones ejecutadas desde un estado a los siguientes.

3.2. Perfiles de usuario

Los actores a darle uso a la solución propuesta corresponden a un perfil perito en aspectos económicos, financieros y de análisis de coyuntura. La solución debe dirigirse a dicha pericia y donde los conceptos técnicos computacionales deben enfocarse a lo relevante y sustancial.

La solución será operada por el perfil anterior durante las etapas de calibración y asignación de parámetros, como de estimación de la política óptima a través de la solución propuesta.

Estos procesos se realizarán en distintos momentos de una jornada laboral promedio según la urgencia de los requerimientos, donde los objetivos pueden abarcar unas pocas horas hasta un cúmulo de días para su completitud.

3.3. Introducción al aprendizaje por refuerzo

La solución propuesta hace uso de la metodología de aprendizaje por refuerzo en su desarrollo y aplicación. Los problemas modelados como RL se expresan en un marco que considera un agente ficticio encargado de la toma de decisiones y de un ambiente que produce información y describe el estado de un entorno.

Este agente ficticio interactúa con el ambiente ejecutando acciones en los estados que lo conforman, recibiendo así una señal por parte de estos en forma de recompensa y calibrando su comportamiento en respuesta de esta última, pudiendo entregar a su vez un *feedback* a dicho estado e influenciar el ecosistema.

El algoritmo asigna un valor a los distintos estados en relación con el retorno o recompensa que proporcionen, esto según la acción ejercida en cada uno de estos. El valor de cada estado se relaciona a los estados que lo suceden, fortaleciendo así una causalidad en las decisiones del algoritmo y en sus beneficios futuros.

Lo anterior es repetido tantas veces como sea requerido con el fin que la retroalimentación ajuste el comportamiento del agente ficticio al conjunto de decisiones más idóneas para la resolución efectiva de su calibración. Esto último se conoce como aprendizaje.

3.3.1. Definición de estados

El agente ficticio aprende con la finalidad de resolver un objetivo de la forma más eficiente posible. Dicho objetivo se compone de estados, correspondientes a distintas situaciones o eventos que permiten alcanzar su cumplimiento. Los estados refieren de cierta forma a caminos que este agente debe recorrer, los que se revelan a medida que tome decisiones en cada uno de ellos.

Los estados de este problema se conforman por tres elementos:

- El precio del USD en pesos chilenos.

- El tiempo transcurrido desde el comienzo de la evaluación.
- La cantidad de dólares restantes por comprar o vender.

Para iniciar su aprendizaje, el agente comienza en un estado inicial. El precio del USD en el mencionado estado corresponde a una variable exógena del modelo de RL, siendo esta una condición de inicio del sistema y observable en el mercado financiero. Su posterior valor en los estados siguientes se modela mediante un proceso estocástico calibrado con datos históricos, incorporando los efectos de volumen en las transacciones de estados precedentes.

El componente de tiempo transcurrido define el orden y secuencia de cada estado, cuya magnitud considera el tiempo consumido desde el estado inicial. La diferencia de esta magnitud entre estados se define de forma constante y cuyo valor es proporcionado por el analista financiero como una condición de inicio, identificada como *paso de tiempo* o dt. Esta diferencia guarda relación con los siguientes elementos del modelo y permite generar de forma natural lo siguiente:

- Segmentación de estados, donde cada uno de estos se encuentra a la misma distancia temporal. El tiempo que transcurre entre estados es homogéneo, definido por el *paso de tiempo*, con la finalidad de estandarizar el aprendizaje.
- Valorización del USDCLP, permitiendo una discretización de su precio y una mayor facilidad en su incorporación al modelo de aprendizaje por refuerzo. El tiempo transcurrido en la innovación del precio se distancia según esta diferencia temporal.
- Segmentación temporal de las políticas, las que responden a los estados utilizados durante el proceso de aprendizaje. Dado que una política corresponde a las acciones a ejecutar en los distintos estados utilizados en la calibración del modelo, la diferencia temporal entre estados condiciona la precisión del aprendizaje. Una reducción en el *paso de tiempo* implica una ocurrencia entre estados más frecuente y consigo, un aumento en la cantidad de acciones aprendidas.

3.3.2. Estado inicial, iteración de estados y estado terminal

Cada iteración o episodio, como se explicará más adelante, tiene en su composición distintos estados conectados de forma secuencial. El agente ficticio comienza en un estado inicial, el que por defecto tiene un tiempo transcurrido de 0 y una cantidad de dólares restantes igual a la totalidad del objetivo. El componente correspondiente al precio del USDCLP es a su vez un parámetro de inicio recogido del entorno.

El siguiente estado define su parámetro de tiempo transcurrido como el valor del estado precedente aumentado en el *paso de tiempo*. La componente de dólares remanentes se determina a su vez del mismo parámetro del estado anterior, disminuido en la acción ejecutada en aquel estado. El precio del USD del estado se determina por un modelo estocástico de divisas, el que incorpora el efecto de volumen por la compra o venta de dólares en el estado previo.

Lo anterior se repite hasta que el tiempo transcurrido sea igual al tiempo total disponible o hasta que la cantidad de dólares remanente llegue a 0 en algún estado. Esto ultimo define

al estado final, siendo este el estado que marcará el fin de una trayectoria o episodio de aprendizaje. En el caso que la cantidad de dólares remanentes sea distinta de 0 en un estado final, la acción ejecutada en dicho estado será la cantidad de dólares remanentes. Esto refiere a que el agente liquida el total de la operación en el último estado al precio vigente con la finalidad de completar el objetivo.

3.3.3. Acciones

En cada estado el agente ficticio ejecuta una acción definida de un conjunto de acciones disponibles. Para el problema a modelar, este tendrá la capacidad de comprar o vender una cantidad finita y entera de dólares, siendo capaz esta de ser 0 y significando en dicho caso una postergación de la operación para el periodo siguiente.

El agente no podrá comprar o vender una cantidad superior de dólares a su objetivo, mientras que en el estado final estará obligado a comprar o vender, según sea el caso, la cantidad restante de dólares.

Las acciones son unitarias y seleccionadas dentro de una región factible. Cada acción se ejecuta una única vez en cada estado en relación con la trayectoria experimentada.

3.3.4. Política de decisión

En el ejercicio de aprendizaje, el algoritmo se enfrenta a distintas situaciones sobre las cuales tomar una decisión y avanzar al estado siguiente. La cadena de decisiones que este ejecuta se conoce como política de decisión, la que representa de forma organizada las distintas acciones que este seleccionará para su ejecución.

3.3.5. Recompensa

La acción realizada por el agente en cada estado tiene un valor representativo que posibilita medir el beneficio o pérdida frente a las distintas alternativas. Dicho valor se conoce como recompensa, la cual permite cuantificar la calidad de una decisión y discriminar las más beneficiosas.

Al ejecutar el agente ficticio una acción, este compara el resultado de la recompensa recibida para determinar qué tan buena elección fue. Distintas decisiones entregan idealmente distintas recompensas sobre los estados experimentados. Una vez recorrido todos los estados o alcanzado un estado terminal, la mecánica consolida todas las recompensas en un cúmulo, permitiendo así concluir que tan efectivas fueron sus vivencias para completar su aprendizaje.

3.3.6. Exploración e intensificación

Los estados se experimentan de forma secuencial, siendo en muchos casos el tiempo en el que ocurren la forma tradicional de segmentación. La ilusión de tomar decisiones que produzcan altas recompensas en estados iniciales podrá traer como consecuencia estados futuros que solo permitan obtener reducidos beneficios. El agente debe buscar maximizar su beneficio global, es decir su retorno, a pesar de que ciertos estados o decisiones puedan parecer atractivas de forma local.

Para lograr esto último, el agente ficticio se enfrenta en cada etapa del aprendizaje a la disyuntiva de la exploración o intensificación. La exploración corresponde a la actividad de privilegiar el descubrimiento de estados con elevada recompensa, a expensas de obtener provecho de los estados conocidos. Por su lado, la intensificación atañe a la reiteración de aquellos estados conocidos y con altas recompensas, dejando de lado lo desconocido.

La capacidad de aprender resulta de un balance entre la exploración e intensificación. El agente debe recorrer y experimentar reiterativamente aquellos estados con mayor recompensa debido al aporte que generan al retorno de la actividad, como también a la promesa que estos trasladen al agente a nuevos estados que mantengan un nivel de beneficio atractivo. Pero también el agente debe experimentar estados no relacionados que permitan descubrir nuevas oportunidades, pudiendo desembocar estas en trayectorias incluso más beneficiosas que las ya consabidas.

3.3.7. Rescatando lo aprendido

Finalmente, el aprendizaje se refleja en la política de decisión. El algoritmo al experimentar tantas combinaciones de estados como le sea posible, dentro del límite de la factibilidad, irá seleccionado las acciones que le permitan obtener el mayor retorno, buscando mejorar en cada vivencia dicho parámetro al identificar las distintas recompensas y ajustar las decisiones tomadas con anterioridad.

3.3.8. Aprendizaje cíclico

Una vez que el agente experimente todos los estados posibles, ya sea por alcanzar un estado final o una condición de término, este podrá volver a comenzar desde el estado inicial su proceso de aprendizaje, alternando las decisiones en la búsqueda de mejorar su entendimiento del entorno. Cada una de estas iteraciones se identifican como episodios y permiten conocer el valor real de cada estado.

3.4. Interacción entre los distintos componentes

Los elementos indicados se presentan en el siguiente diagrama, destacando los principales componentes y sus relaciones.

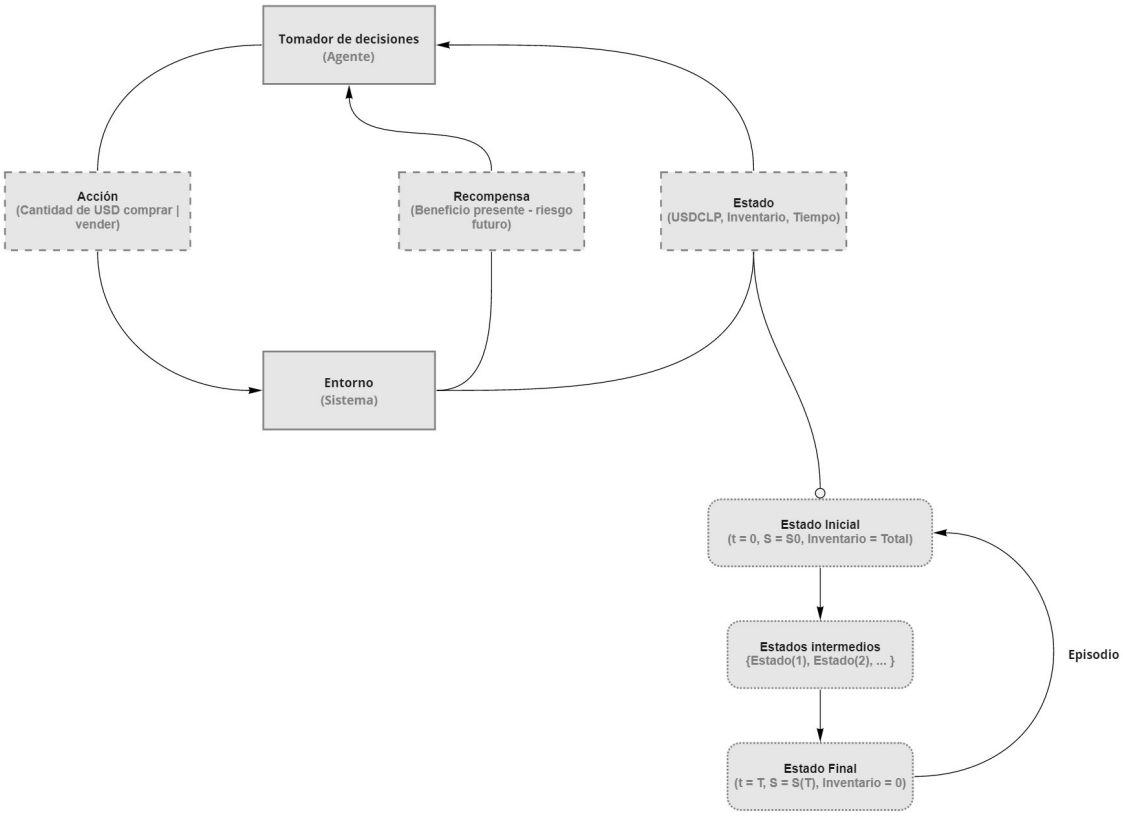

Figura 3.1: Diagrama de integración del modelo de aprendizaje por refuerzo

El universo es identificado como entorno o sistema. Este interactúa con el tomador de decisiones a través de los distintos estados. El conjunto de acciones ejecutadas sobre estos se identifica, según se ha señalado, como política de decisión, correspondiendo estas al conjunto de acciones del agente ficticio en cada estado experimentado. Una vez ejecutada una acción el estado proveerá una recompensa, permitiéndole así jerarquizar sus decisiones y establecer las más beneficiosas. Esto último se identifica como política óptima.

Los estados se componen por un estado inicial, siendo este el comienzo del proceso de aprendizaje. El agente ficticio observará un precio inicial del USDCLP del entorno, como también asumirá un objetivo de compra o venta de una determinada cantidad de dólares identificada como inventario y un tiempo máximo para cumplir dicho objetivo. Tanto el nivel inicial de inventario como el tiempo máximo de operación serán proporcionados, entre otros, por el analista financiero como parámetros del modelo.

El agente ficticio ejecuta una determinada acción en el estado inicial, dentro de los límites establecidos por el analista financiero y plasmados como restricciones en el modelo, y recibe una recompensa. Este último valor puede ser significativo, reducido o incluso negativo, per-

mitiéndole al agente interpretar que tan beneficiosa fue la decisión ejercida. La recompensa corresponde a un valor comparable resultante de interpretar el beneficio de ejecutar la acción en el estado actual y sus riegos implicados sobre los estados futuros.

Una vez recibida la recompensa por parte del estado hacia el agente, este avanza al estado siguiente. Este nuevo estado es condicionado por la acción del agente en el estado previo, la evolución natural de este último y los movimientos económicos y financieros ocurridos en el sistema. En este estado el precio del dólar responde a los movimientos de la economía modelados por variables representativas y por la presión del volumen de las operaciones ejecutadas en el estado que lo precede. El inventario de dólares se establecerá al reducir el nivel del estado previo en la acción de compra o venta realizada, mientras que el tiempo reflejará un avance en dt unidades.

Una vez avanzado a este estado el agente ejecutará nuevamente una acción, obteniendo una recompensa y repitiendo el ciclo de aprendizaje hasta llegar a un estado terminal. El conjunto de estados contenidos entre el estado inicial y terminal se identifica como episodio, el cual tendrá asociado un conjunto de acciones o decisiones señaladas como política.

Los distintos episodios se relacionan de forma sinérgica, siendo esta una de las bases del aprendizaje secuencial y de modelos de aprendizaje por refuerzo como Q-Learning o SARSA. Durante el primer episodio el agente no presenta mayor conocimiento del entorno o sistema y lo descubre a través de la experiencia de recorrer los distintos estados y las decisiones que ejerza sobre estos. Debido a la nula veteranía del agente en esta etapa de su aprendizaje las acciones responden a comportamientos mayormente aleatorios en la búsqueda de un mayor entendimiento del marco que lo rodea. Comprar o vender determinadas cantidades de dólares a un precio establecido no representará mayor trascendencia durante dicho episodio, esto debido a la ausencia de una comparativa representativa para proporcionar un valor al estado y la acción correspondiente.

Una vez finalizado el primer episodio, el agente dispondrá de un reducido entendimiento de la realidad que lo enmarca y de las decisiones idóneas para el cumplimiento de su objetivo. Pero la mecánica de aprendizaje reside en la iteración. Una vez culminado cada uno de estos episodios, la experiencia es almacenada por parte del agente, mejorando así la comprensión del efecto de sus decisiones sobre cada estado percibido.

Es importante destacar que la experiencia se almacena en forma de valores comparables considerando, como se explicará más adelante, la relevancia de las acciones o probabilidad de transición en los estados a considerar. Como será indicado, el valor de la recompensa permite una comparativa entre estados de forma aislada, pero no considerará la interacción entre estos. De forma sencilla, esta experiencia debe almacenar el efecto de una acción sobre la política de decisión, es decir, recoger el retorno de una acción sobre el estado actual ajustado por el beneficio recibido sobre los estados consecutivos. Esto es lo que permite dotar al agente con una herramienta que considere la recompensa de una acción en el presente, complementada por la causalidad de esta sobre los estados futuros.

Al enfrentar cada nuevo episodio dispondrá de una experiencia más desarrollada y de una base comparativa superior en el valor de estados y de las consecuencias de las acciones. El agente podrá establecer y priorizar los distintos beneficios y riesgos asociados a un determinado nivel

de USDCLP, la probabilidad de transición a un nuevo precio, los efectos de un determinado volumen sobre la divisa y las posibles ganancias y pérdidas al anticipar o postergar una decisión de compra o de venta.

3.5. Valorización de estados y su relación a las acciones

Si bien la función de recompensa permite jerarquizar las acciones en relación de un indicador de beneficio, su valor es local y no presenta relación con estados futuros. La métrica utilizada para comparar lo beneficioso de un estado o de las acciones asociadas a este corresponden a las relaciones que se exponen a continuación.

Como fue señalado, una acción sobre un estado particular puede retornar una atractiva recompensa, pero posteriormente los estados sucesivos vinculados a dicha decisión podrán entregar reducidos beneficios o incluso pérdidas. Lo anterior hace necesario el poder valorizar un estado y las acciones con sus eventos futuros.

Referente a lo anterior, y según lo señalado por Bellman (1957), el valor de un estado podrá definirse como el valor esperado intrínseco de las recompensas de los estados venideros, dado un estado s particular y una política de decisión π, según la expresión siguiente:

$$V_t^{\pi}(s) = \mathrm{E}_t^{\pi}\left[\sum_{i=0}^{T-t-1} \gamma R(S_{t+i}, a_{t+i}, S_{t+i+1}) \mid S_t = s\right] \tag{3.1}$$

La expectativa indicada vincula el retorno de un estado futuro a la probabilidad de trasladarse a dicho estado desde un estado presente, medida en π.

Por su parte, el factor γ representa un factor de descuento[13] que permite la comparativa entre los retornos de diferentes estados que lo suceden. Adicionalmente, la política de decisión π identifica, como se ha señalado, al conjunto de decisiones futuras realizadas por el agente a partir del estado s.

Lo señalado permite establecer un entorno de trabajo formal para la problemática estudiada. (3.1) establece una metodología para valorizar los distintos estados del problema, correspondiendo al retorno esperado de las decisiones de compraventa de dólares de los estados sucesivos, con respecto a la política o conjunto de acciones ejercidas.

La formalización anterior establece una base de análisis sobre la que se desarrollarán las metodologías bajo estudio. Una de las principales limitantes de (3.1) radica en el uso de probabilidades de transición para la comunicación directa entre estados, valores difíciles de precisas en la práctica.

De forma similar a (3.1), se define el valor de un estado-acción como la simultaneidad de

[13]El factor de descuento es un concepto de consideración en economía y finanzas por ser una herramienta que facilita la comparativa del valor de los activos, bienes y productos financieros a través del costo de oportunidad en distintos eventos y unidades de tiempo.

ejecutar una acción a en un estado s, siguiendo una política π como acciones sucesivas:

$$Q_t^{\pi}(s) = \mathrm{E}_t^{\pi}\left[\sum_{i=0}^{T-t-1} \gamma R(S_{t+i}, a_{t+i}, S_{t+i+1}) \mid S_t = s, A_t = a\right] \tag{3.2}$$

La anterior relación puede ser separada, obteniéndose:

$$Q_t^{\pi}(s) = \mathrm{E}_t^{\pi}\left[R(S_t, a_t, S_{t+1})\right] + \mathrm{E}_t^{\pi}\left[\sum_{i=1}^{T-t-1} \gamma R(S_{t+i}, a_{t+i}, S_{t+i+1}) \mid S_t = s, A_t = a\right] \tag{3.3}$$

El primer término hace referencia a la recompensa recibida en el estado actual, mientras que el segundo término cuantifica el valor del estado particular en tiempo presente con respecto a eventos futuros. Esto proporciona facilidad para simplificar la notación de la forma:

$$Q_t^{\pi}(s) = \mathrm{E}_t^{\pi}\left[R_t(s, a, s')\right] + \gamma \mathrm{E}_t^{\pi}\left[V_{t+1}^{\pi}(s')\right] \tag{3.4}$$

Esta relación permite trabajar la problemática desde la perspectiva de la acción ejecutada en cada estado, proporcionando facilidad para una ejecución computacional bajo un régimen de simulaciones. Al utilizar esta metodología se proporciona la cantidad de dolares a comprar o vender para cada configuración de estados, desprendiéndose la necesidad de probabilidades de transición entre estos, correspondiendo a una significativa ventaja de (3.2) en el modelamiento de sistemas financieros.

Ya sea que la relación utilizada para medir el valor de un estado sea V o Q, su utilidad radica en la posibilidad de consolidar el valor de una trayectoria futura desde el estado actual.

De lo anterior se desprende la existencia de 2 mecanismos que permiten darle un valor a un estado particular, el de valorización a través del *valor del estado* (V) o el del *valor estado - acción* (Q). Si bien la recompensa proporciona el valor en bruto del beneficio de cada estado, esta no es utilizada de forma directa para discriminar estados o acciones, sino que dicha comparativa se realiza mediante Q o V por la capacidad de incorporar el itinerario de recompensas de cada política de decisión.

Q y V no son utilizadas derechamente en los modelos de estudio de la presente tesis, sino como será detallado más adelante, permiten dar origen, en particular Q, a las metodologías de Q-Learning y SARSA.

3.6. Política de decisión y su relación con las metodologías de aprendizaje estructurado

Como se ha detallado previamente, la política de decisión corresponde al conjunto de acciones o decisiones materializadas por el agente en cada estado particular, actividad relevante para una valorización en Q. Adicionalmente, su aplicación podrá vincularse a probabilidades de ocurrencia de eventos ligados a estados y acciones de un sistema, siendo esto de suma relevancia para una valorización en V.

La política de decisión concierne a una función que asigna una acción a un estado específico. Esta corresponde a la función $\pi(s)$ que retorna la acción a en s. La política de decisión se origina de los Procesos de Decisión de Markov (MDP), que como se indicará en una sección posterior, corresponden a la metodología precursora a RL.

Si un agente persigue una política π en el tiempo t, entonces $\pi(a \mid s)$ corresponde a la probabilidad que $A_t = a$ si $s_t = s$. Lo anterior señala que cada acción del agente en un determinado estado estará gobernada por una distribución de probabilidad en π, siendo esta alimentada por la viabilidad del agente de posicionarse en el determinado estado s.

Disponer de aquella política que maximiza los beneficios para el sistema permite resolver el problema en Q o en V, donde su determinación se realiza mediante la acción repetitiva de distintas políticas y su posterior jerarquización, siendo los distintos episodios las instancias para su definición.

Entonces, para la problemática de estudio, la política de decisión corresponde a la totalidad de dólares a ser comprados o vendidos para cada estado experimentado. Esta política, como se detallará más adelante, será dependiente a la trayectoria ejecutada por el algoritmo.

Relevante es el concepto de aprendizaje secuencial que experimenta el agente. Tras la ocurrencia de cada episodio el agente busca mejorar el entendimiento de su entorno, correspondiendo a un proceso donde las acciones cobrarán mayor sentido según la recompensa proporcionada. La selección de estados y sus acciones representativas entregarán mayor congruencia según el retorno asociado y la distribución de probabilidad correspondiente.

Los estados no experimentados tienen por definición una valorización de 0[14], mientras que aquellos experimentados por primera vez asignan el valor de la recompensa de dicha configuración, esto ante la ausencia de una trayectoria futura generada.

Inicialmente los estados comenzaran con un valor de reducida representatividad, la que aumentará conforme el agente experimente distintas políticas y distintos episodios, los que se ajustarán según la función de valorización Q o V.

Lo anterior viene a ser explicado por el descubrimiento del entorno que experimenta el agente. Para sistemas de considerable complejidad, la cantidad de estados será lo suficientemente elevada para que durante los primeros episodios las recompensas recibidas y el valor determinado de cada estado represente una cantidad ínfima del espacio muestral.

Conforme la cantidad de episodios aumente, también lo hará el número de estados descubiertos. Consecuencia de esto será un incremento en el número de estados experimentados y por consiguiente, una mayor representatividad en el valor descontado de la expectativa que da cuenta de la relevancia de la política asociada.

[14]Una técnica que fomenta la exploración del agente consiste en definir estados con un valor inferior a los retornados en los episodios iniciales, con el fin que el agente explore nuevos estados y prolifere su entendimiento del sistema de forma acelerada.

3.7. Relación entre la valorización de estados y de acciones - estados

Como ha sido mencionado, la función de valorización de estados $V_t^{\pi}(s)$ corresponde a la recompensa acumulada contabilizada desde el estado s, en un tiempo t y ejecutando un conjunto de acciones bajo una política π.

Por su parte, la función de valorización estado-acción $Q_t^{\pi}(s, a)$ considera, para un tiempo t y un estado s, la selección de una acción a y la posterior elección de una política π a ser ejecutada en los estados posteriores.

La principal diferencia entre las valorizaciones indicadas consiste en la aplicabilidad en su forma de cuantificar el valor de la realidad. $Q_t^{\pi}(s, a)$ captura el valor de una decisión sobre un estado particular. Su utilidad reside en la capacidad de parametrizar las acciones de forma explícita a la modelación, situación provechosa para las algorítmicas de Q-Learning y SARSA al aproximar Q mediante iteraciones de distintas decisiones en los estados que se revisan.

$V_t^{\pi}(s)$ por su parte expresa el valor de un estado condicionado a una política de decisión, pero donde las acciones se encuentran implícitas en el valor de estos. La conveniencia de V se aplica a eventos cuya valorización es factible de asociar a una probabilidad de transición entre los estados que los conforman.

El aprendizaje a través de V utiliza como herramientas la recompensa de posicionarse en un estado en particular y la factibilidad de alcanzar este. Cada configuración de estados tendrá asociado un retorno coherente con la eventualidad de ubicarse en dicho estado particular. Las acciones se encontrarán asignadas a cada estado como una configuración factible, donde la probabilidad de transición comunicará a los estados sucesivos. La política de decisión aprendida por el agente corresponderá a la configuración de estado – acción que proporcione mayor valor, siendo este condicionado por la factibilidad de los estados y retornos sucesivos.

Un aprendizaje en Q por su parte considera la acción del agente como comunicador entre estados, donde la probabilidad de transición aminora su relevancia al materializarse de forma subyacente. El valor de un estado se asignará al considerar una recompensa de acuerdo a la acción del agente sobre un estado en particular. Cada política de decisión considerará las consecuencias de una acción determinada y su incidencia sobre estados futuros.

De lo mencionado se desprende la relación numérica entre Q y V. Mientras el valor mensurado por medio del *valor estado - acción* considera la aplicación de una trayectoria de acciones en el retorno obtenido, la medida de *valor estado* considera el beneficio producido por la probabilidad de abarcar un conjunto de estados asociados a un grupo de acciones en particular. Lo anterior, según lo especificado por Sutton y Barto (1998), permite expresar los conceptos según la siguiente relación:

$$V_t^{\pi}(s) = \mathrm{E}_t^{\pi}\left[Q_t^{\pi}(s, a)\right] = \sum_a \pi(a \mid s)\ Q_t^{\pi}(s, a) \tag{3.5}$$

La métrica V corresponde a la esperanza [15] de todas las posibles acciones que permiten

[15] Valor esperado de una variable aleatoria.

alcanzar un estado en cuestión, considerando su probabilidad de materialización.

Para la problemática de la presente tesis, el disponer de probabilidades de transición entre estados presenta una elevada dificultad y carente de precisión en respuesta a la amplia cantidad de factores involucrados en su determinación. Debido a esto, el modelar la problemática en Q se hace la elección predilecta en escenarios financieros como el abordado, proporcionando un respaldo significativo a la aplicación de las presentes metodologías y un atractivo adicional a la mecánica de aprendizaje por refuerzo.

3.8. Optimización de la política de decisión

El valor óptimo para un estado corresponde al mayor valor obtenido producto de todas las políticas factibles. Dicho valor es representado por V^*, mientras que la política óptima, identificada como π^*, será aquella que permita su origen.

Complementariamente, las relaciones expuestas a continuación como su profundización pueden ser revisadas en Sutton y Barto (2018).

De esta forma, el valor óptimo V^* quedará expresado como:

$$V_t^*(s) := \max_{\pi} V_t^{\pi}(s), \forall s \in S \tag{3.6}$$

De igual forma, el valor óptimo de la función *acción-valor* se vincula a la política óptima π^* y su mayor valor generado para su selección. Lo anterior queda representado mediante:

$$Q_t^*(s,a) := \max_{\pi} Q_t^{\pi}(s,a), \forall s \in S \tag{3.7}$$

Vinculando las ecuaciones de valorización de estado y de *acción-valor*, podemos obtener su magnitud según la política óptima del sistema, es decir, el valor óptimo de estas. Para la ecuación de valoración de estado, esta es:

$$V_t^{\pi^*}(s) = \mathrm{E}_t^{\pi^*}\left[R_t(s,a,s')\right] + \gamma \mathrm{E}_t^{\pi^*}\left[V_{t+1}^{\pi^*}(s')\right] \tag{3.8}$$

Mientras que para la ecuación de *acción-valor* esta corresponde a:

$$Q_t^{\pi^*}(s,a) = \mathrm{E}_t^{\pi^*}\left[R_t(s,a,s')\right] + \gamma \mathrm{E}_t^{\pi^*}\left[V_{t+1}^{\pi^*}(s')\right] \tag{3.9}$$

Al ser dependiente esta última del valor del estado $V_t^{\pi^*}(s')$, su resolución queda vinculada a la solución de 2 procesos de optimización. Se aplica entonces la relación entre ambas funciones de optimalidad:

$$V_t^*(s) := \max_{\pi} Q_t^{\pi}(s,a) \tag{3.10}$$

Esto último permite expresar $Q_t^{\pi^*}(s,a)$ de la forma:

$$Q_t^{\pi^*}(s,a) = \mathrm{E}_t^{\pi^*}\left[R_t(s,a,s') + \gamma \max_{\pi} Q_t^{\pi}(s',a')\right] \tag{3.11}$$

La ecuación (3.11) corresponde a la ecuación de optimalidad de Bellman, siendo esta una relación no lineal entre sus componentes producto del operador de maximización. Su resolución

se logra usualmente a través de métodos numéricos utilizando metodologías de programación dinámica. Conforme el avance de técnicas de inteligencia artificial, los modelos de aprendizaje automático y su aplicación al desarrollo de metodologías de aprendizaje por refuerzo han permitido aproximar el resultado de la ecuación $Q_t^{\pi^*}(s, a)$, aplicación que se utiliza en el presente documento y que se detallará posteriormente.

3.9. Origen de la metodología de resolución

Las ecuaciones de optimalidad presentadas corresponden al objeto de estudio central de los Procesos de Decisión de Markov (MDP), donde los métodos de Programación Dinámica (DP) se focalizan en la solución exacta o numérica de estas ecuaciones.

Los MDP proporcionan un entorno de modelamiento de problemas de decisión donde la respuesta a los eventos permite incorporar un componente aleatorio fuera del control del tomador de decisiones. Por lo anterior, son ampliamente utilizados en modelar problemas con componentes secuenciales, desde la deuda en mora hasta juegos de azar.

Esta estructura considera estados sobre los cuales un tomador de decisiones ejecuta acciones que le permiten descubrir y desplazarse a nuevos estados, distinguiéndose estos por una recompensa obtenida.

La probabilidad de trasladarse entre estados depende de la acción seleccionada, la cual recursivamente se alimenta de los estados previos y las decisiones sobre estos, satisfaciendo así la propiedad de Markov. Esta última da cuenta que toda la información relevante de los eventos pasados se encuentra contenida en el estado actual:

$$P\left[S_{t+1} \mid S_t\right] = P\left[S_{t+1} \mid S_1, ..., S_t\right] \tag{3.12}$$

El objetivo de los Procesos de Decisión de Markov también se concentra en la búsqueda de una política óptima π, identificando la acción que el tomador de decisiones seleccionará en el estado correspondiente.

Debido a que la acción seleccionada en el estado s particular se encuentra determinada por π, la probabilidad de transición se transforma en:

$$P\left[S_{t+1} = s' \mid S_t = s, a_t = a\right] \rightarrow P\left[S_{t+1} = s' \mid S_t = s\right] \tag{3.13}$$

correspondiendo lo anterior a una matriz de transición de Markov.

Finalmente, la política de decisión se selecciona tomando como métrica una función de retorno acumulado, cuya expectativa quedará determinada por las probabilidades de transición del sistema, seleccionado π como aquella política que maximice la siguiente ecuación, análoga a la ecuación de optimalidad de Bellman:

$$V_t^{\pi}(s) = \mathrm{E}_t^{\pi}\left[\sum_{i=0}^{T-t-1} \gamma R_{a_t}(S_{t+i}, S_{t+i+1}) \mid S_t = s\right] \tag{3.14}$$

Lo anterior da cuenta de la homogeneidad en la abstracción de la realidad a modelar por las distintas metodologías. El origen se concentra en el beneficio resultante de estados con

características particulares y su relación entre los mismos a través de acciones que permiten interpretar la realidad que proporcionan, como también probabilidades que dan cuenta de su capacidad de ocurrencia.

Las distintas técnicas y metodologías utilizadas en resolver estas relaciones han sido el mérito de estudio de diferentes disciplinas, donde los métodos de Aprendizaje por Refuerzo han marcado hitos en los últimos años.

3.10. Métodos de aprendizaje por refuerzo

Las metodologías de aprendizaje por refuerzo persiguen resolver las ecuaciones de valorización de estados al igual que los modelos MDP y DP. La diferencia se traduce en la forma de procesar los datos y de diseñar las relaciones entre los componentes.

La valía de esta metodología consiste en la búsqueda de aproximar la ecuación de optimalidad de Bellman mediante técnicas de exploración de un entorno parcialmente desconocido, donde las acciones, beneficios y penalizaciones permiten descubrir una política óptima a través de una mejora continua en el entendimiento de $Q_t^{\pi^*}$ mediante experiencias iterativas de eventos simulados.

Mientras en una modelación convencional basada en DP se dispondrán de múltiples ecuaciones a ser resueltas por técnicas numéricas de forma exacta, un modelo tradicional de RL buscará la obtención de π^* mediante iteraciones y recursividad. Estos últimos asumen una percepción parcial del entorno, sustentando su accionar en muestras de datos proveniente de una distribución de eventos gobernada por el entorno que buscan comprender.

Existen distintas categorías de aprendizaje por refuerzo, destacando:

- Aplicación de modelamiento en el aprendizaje: La mecánica buscará aprender las probabilidades de transición y la recompensa de un estado, asistida por un modelo que facilite su entendimiento. Mientras una metodología de RL libre de modelamiento dispone de una función de recompensa para identificar y segmentar las combinaciones de estado - acción a través de una mejora sucesiva de $Q_t^{\pi^*}$, algoritmos de RL apoyados por modelos proporcionan conjuntos de reglas y ajustes adicionales para mejorar el entendimiento del entorno. Es importante destacar que disponer de modelos en la función de recompensa para abstraer conocimiento del sistema no induce a una categoría de modelamiento, esto debido a que dicho modelo abstraerá señales del entorno más no asistirá en la interpretación de los beneficios de estas.
- Aprendizaje coherente con la política de decisión (*on-policy*) y no limitado a estas (*off-policy*): Una metodología *on-policy* exige que la decisión iterativamente sucesiva corresponda a la trayectoria de política desarrollada, mientras que un método (*off-policy*) libera al tomador de decisiones a ejecutar una acción consecutiva no perteneciente a la política vigente. Lo anterior permite que el agente identifique durante el transcurso del episodio políticas que proporcionen un beneficio superior, seleccionando y corrigiendo la política encausada.

- Aprendizaje continuo (*on line*) y aprendizaje por lotes (*batch*): El aprendizaje por lotes segmenta el recorrido exploratorio del tomador de decisiones a fragmentos, los que culminados permitirán la actualización de $Q_t^{\pi^*}$. Por el contrario, un aprendizaje continuo actualiza dicha función en la unidad mínima para así proporcionar de forma constante perfeccionamiento en el aprendizaje del agente.

Adicionalmente se complementa la especificidad de las técnicas utilizadas en la aproximación de $Q_t^{\pi^*}$, dentro de las que se encuentran:

- Metodología de Monte Carlo: Aproximación del valor de $Q_t^{\pi^*}$ mediante un muestreo exploratorio de un conjunto N de trayectorias, sobre la cual se calcula la media muestral de $Q_t^{\pi^{(N)}}$.
- Metodología de búsqueda de políticas: Explorar utilizando métodos de gradiente y convenir una política adecuada. Su uso resalta en la búsqueda de políticas estocásticas $\pi_\theta(a|s)$ que define una distribución de probabilidad sobre un conjunto de decisiones $a \in A$, donde θ define parámetros de dicha distribución.
- Metodología de diferencia temporal: Aproximación del valor de $Q_t^{\pi^*}$ mediante un muestreo exploratorio donde la actualización se materializa en el siguiente *paso de tiempo* o dt. Esta metodología será la utilizada en la presente tesis.

3.11. Características método diferencia temporal

La metodología presenta beneficios y desventajas en el objetivo de aproximar el valor de $Q_t^{\pi^*}$, características que se describen a continuación.

- Rapidez en la resolución del problema modelado: El incremento en la frecuencia de actualización de la función $Q_t^{\pi^*}$ permite una convergencia con mayor celeridad a la solución óptima o a una estabilización de la función de recompensa.
- Oscilaciones durante el procedimiento: Debido a que la actualización considera en su aproximación una única observación correspondiente al estado vigente, esta puede presentar abruptas correcciones.
- Facilidad en una implementación continua: Debido a su actualización en lotes, su aplicabilidad a un aprendizaje *on line* es natural, pudiendo desplegar sus funciones en ambientes no terminales.
- Aplicabilidad sobre sistemas intensivos en el modelamiento bajo la propiedad de Markov: Esto se desprende de la consolidación del modelo en lotes que involucran una única observación la cual consolida los eventos que la originan.
- Presencia de sesgo en la metodología: El sesgo se origina producto de un valor inicial aleatorio, donde la metodología de aproximación converge parcialmente al verdadero valor de $Q_t^{\pi^*}$, brecha que disminuye conforme el aumento de episodios.

- Varianza no acumulativa: Producto de una actualización por lote, la varianza de un episodio no acumula la variabilidad de cada estado, originando trayectorias con reducida varianza frente a otras metodologías.

3.12. Justificación implementación metodología diferencia temporal

La elección de la metodología de diferencia temporal se sustenta en su facilidad de implementación y la rapidez en su resolución. Adicionalmente, su metodología de desarrollo, al ser intensiva en el uso de la propiedad de Markov, permite mayor afinidad con las problemáticas secuenciales financieras, cuya base teórica se sustenta en similar característica.

3.13. Definición metodología de diferencia temporal

El método de diferencia temporal (TD) pertenece a la metodología de aprendizaje por refuerzo libre de modelos cuya principal característica consiste en un muestreo en pequeñas unidades de tiempo sobre las cuales recolectar información y gestar un entrenamiento. Mientras distintos modelos de RL aguardan a la culminación del episodio para actualizar las ecuaciones de valorización y reflejar un aprendizaje, el método TD recoge la intuición que dicho proceso se puede materializar con un pequeño cúmulo de estados experimentados previo a la finalización del episodio.

La metodología considera la recompensa percibida en el estado actual y la valorización del estado sucesivo, algoritmo heredado de mecánicas iterativas de DP y que permiten aproximar una actualización de valorización del estado vigente mediante la observación siguiente.

Si bien disponer de una aproximación basada en una única observación trae como consecuencia la incorporación de un sesgo en la mecánica, este último puede ser disminuido de forma significativa mediante el incremento de episodios, presentando a su vez la ventaja de proporcionar una rápida convergencia a la solución.

Esta metodología aproxima $\mathrm{E}_t^\pi \left[R_t(s,a,s') + \gamma V_{t+1}^\pi(s')\right]$ mediante la diferencia de 2 observaciones, utilizando dicho valor en el cálculo de un error ε de la forma:

$$\varepsilon = R_t(s,a,s') + \gamma V_{t+1}^\pi(s') - V_t^\pi(s') \tag{3.15}$$

Lo anterior permite generar una regla de actualización de V_t según:

$$V_t(s) \leftarrow V(s) + \alpha \left[R_t(s,a,s') + \gamma V_{t+1}^\pi(s') - V_t^\pi(s)\right] \tag{3.16}$$

La incorporación del factor α permite ajustar el impacto del error ε en la actualización de la ecuación de valorización, refiriéndose a esta magnitud como la tasa de aprendizaje. Esta cuantifica la cantidad de error a ser aceptada y su ajuste en cada unidad de *paso de tiempo*, graduándose entre los valores de 0 y 1. Un valor elevado realiza ajustes más agresivos,

aceptando un mayor error, mientras que un valor más reducido corrige la aproximación de forma más conservadora.

Por su parte, s responde a la identificación del estado vigente mientras que s' representará al estado inmediatamente consecutivo. Complementariamente, el valor de a corresponde a la acción vinculada al estado s.

Como se observa en la última ecuación, el valor $V_t(s)$ se encuentra condicionado a la información disponible en el siguiente estado, beneficio significativo frente a métodos como Monte Carlo cuya actualización debe materializarse al final del episodio.

La metodología de diferencia temporal desarrolla su algorítmica sobre la ecuación de *valor estado*, permitiendo dar paso a diferentes ajustes y metodologías afines. Dentro de estas se encuentran los métodos de Q-Learning y SARSA, los que serán explicados a continuación.

3.14. Aplicación de metodología de diferencia temporal: Q-Learning - SARSA

Las metodologías por presentar incumben a la metódica central de la presente tesis en la evaluación de la aplicabilidad de técnicas de aprendizaje por refuerzo convencionales a la problemática financiera indicada. Estas corresponden a Q-Learning y SARSA

Su principal característica corresponde la de ser métodos de aprendizaje por refuerzo de modelo libre basados en la algorítmica de diferencia temporal, aprendiendo el valor de una acción en un estado en particular, es decir, el valor de Q_t^π. Para Procesos de Markov finitos estas metodologías convergen a una política óptima maximizando el valor esperado de la trayectoria descontada al estado vigente, identificando las acciones a ser consideradas en cada configuración de estados experimentados.

La diferencia por considerar entre las metodologías de Q-Learning y SARSA se concentra en la aplicación de un aprendizaje basado en técnicas *on-policy* y *off-policy*.

Como se señaló, un algoritmo *on-policy* asume que la política óptima será utilizada para generar los datos, por lo que el objetivo se centra en su aprendizaje. Por su parte un algoritmo *off-policy* asume que la política utilizada en un determinado conjunto de datos puede no ser óptima o incluso aleatoria. El propósito se concentra en aprender la política óptima cuando los datos son seleccionados bajo una política diferente.

Lo anterior, para una mecánica de diferencias finitas, permite establecer casos de uso en los pares de tuplas a ser consideradas. La regla de actualización de esta mecánica utiliza elementos de la forma $\{(s(a), s(a')'), (s(a')', s(a'')''), (s(a'')'', s(a''')'''), ...\}$ para la obtención del error de la innovación. Un aprendizaje *on-policy*, por construcción, debe establecer al último elemento de una tupla cualquiera como el primer elemento de la tupla siguiente. Por el contrario, un aprendizaje *off-policy* permite considerar, dentro del espacio de factibilidad, un elemento distinto al comienzo de la tupla subsecuente. Esta libertad le proporcionará al agente la flexibilidad de seleccionar una acción que maximice el valor de Q_t^π, característica

propia del método de Q-Learning.

Los métodos *on-policy* utilizan la regla de actualización aplicada a la ecuación óptima de acción-valor de la forma:

$$Q_t(s,a) \leftarrow Q(s,a) + \alpha \left[R_t(s,a,s') + \gamma Q^{\pi}_{t+1}(s',a') - Q^{\pi}_t(s,a) \right] \tag{3.17}$$

Esta ecuación define la metodología SARSA, Rummery y Niranjan (1994). Su característica radica en el valor de $Q^{\pi}_{t+1}(s',a')$, correspondiente al valor de Q sujeto a la acción a' en el estado s' siguiente, esto debido que los datos se seleccionaron siguiendo una política óptima.

Considerando que los datos no fueron seleccionados siguiendo una política óptima, se deberá establecer una regla que permita la elección de una política. Una posibilidad considera forzar que la acción seleccionada en s' represente al máximo valor de Q. Esta característica define a la metodología de Q-Learning, Watkins (1989), correspondiendo a un método *off-policy* expresado como:

$$Q_t(s,a) \leftarrow Q(s,a) + \alpha \left[R_t(s,a,s') + \gamma \max_{a'} Q^{\pi}_{t+1}(s',a') - Q^{\pi}_t(s,a) \right] \tag{3.18}$$

Lo señalado con anterioridad expresa los orígenes, formación y estructura de las técnicas de aprendizaje por refuerzo a ser aplicadas en el presente documento. En las definiciones siguientes se detallarán los complementos a estas metodologías, las que permiten modelar el precio del USDCLP, los mecanismos de evaluación de riesgo y la abstracción del impacto de una decisión a un valor cuantificable.

3.15. Valorización de producto financieros mediante procesos estocásticos

La impredecibilidad en los activos económicos y financieros corresponde a una de las características más importantes de su modelamiento. Debido a la aleatoriedad propia de estos eventos, los modelos que buscan dar entendimiento a su comportamiento sustentan sus fundamentos en distribuciones de probabilidad. De lo anterior, se procederá a explicar la descomposición de los retornos financieros en componentes deterministas y estocásticos, permitiendo así su modelación.

Desde su origen, el objetivo de una inversión se traduce en producir un retorno positivo, cantidad relevante a ser medida en su forma porcentual más no en su nivel. Este retorno responde al crecimiento relativo en el valor de un activo. Para el presente documento, lo señalado corresponde al valor del dólar estadounidense medido con relación al peso chileno, conforme el avance de una determinada unidad de tiempo.

Considerando lo anterior, el retorno de un día i a un día $i+1$ fruto de mantener en posesión una divisa corresponde al cambio de valor en el tiempo de esta con respecto a su precio de referencia Y_i[16]:

$$R_i = \frac{Y_{i+1} - Y_i}{Y_i} \tag{3.19}$$

[16]Existen otros tipos de activos que incorporan en su retorno el devengo de dividendos o cupones.

Por su parte, se entienden los 2 primeros momentos estadísticos de los retornos financieros como la media y la desviación estándar de una muestra[17]. Lo anterior fuerza una suposición en que los retornos pueden ser representados por una distribución normal[18].

Asumiendo los retornos como una variable aleatoria ϕ descrita por una distribución Gaussiana, su valor en i puede ser representado por una modelación simple de la forma:

$$R_i = \mu + \sigma\phi \tag{3.20}$$

siendo μ la media de los retornos y σ su desviación estándar. Lo precedente señala que el retorno de un activo financiero durante un día i puede ser referido mediante la media de los retornos de una unidad de tiempo y la dispersión de estos, ajustada esta última por una variable aleatoria bajo una distribución normal.

Es relevante destacar que la relación indicada en (3.20) debe poder expresarse y ser robusta ante distintos horizontes de tiempo. Lo anterior restringe a que las magnitudes de media y varianza de retornos diarios puedan ser utilizadas para representar retornos en distintas unidades temporales, como semanas, meses o años. Para hacer factible lo indicado, se incorpora el factor de ajuste δ_t a los 2 momentos anteriores con la finalidad de poder representar esta graduación.

Al ser usualmente las magnitudes financieras expresadas de forma anualizada, esta proporcionalidad identificada como δ_t corresponde convencionalmente a fracciones de un año. Esto implica, a modo de ejemplificación, que el valor δ_t utilizado para reflejar un ajuste de retornos diarios a retornos anualizados corresponde a $\frac{1}{252}$, expresando la fracción de días laborales.

Esto proporciona un ajuste sencillo, pero a la vez robusto, permitiendo escalar los momentos y así expresar retornos coherentes con distintas unidades de tiempo a representar. Importante es observar cómo dicho ajuste coexiste con esta media y varianza, como también la modificación de proporcionalidad que establece.

Para el caso de la media de los retornos, esta es proporcional al intervalo de tiempo que se busca modelar. De lo anterior, la media de los retornos puede ser aproximada por $\mu\delta_t$, donde en ausencia de aleatoriedad, los retornos pueden componerse en la siguiente relación:

$$Y_{i+1} = Y_i(1 + \mu\delta t) \tag{3.21}$$

donde para M periodos sucesivos se obtendrá:

$$Y_M = Y_i(1 + \mu\delta t)^M \rightarrow Y_0 e^{Mlog(1+\mu\delta t)} \tag{3.22}$$

Aproximando $log(1+\mu\delta t)$ a $\mu\delta t$, el valor del activo Y en el periodo M queda expresado como:

$$Y_M = Y_0 e^{\mu M\delta t} \rightarrow Y_t = Y_0 e^{\mu t} \tag{3.23}$$

donde μ se conoce como tasa de crecimiento o tendencia.

[17] Esto debido a la imposibilidad de abarcar la totalidad de la población en un evento a ser medido.

[18] En la teoría dicha suposición permite simplificar las ecuaciones, a expensas de la evidencia empírica de eventos de cola más pronunciados que los que una distribución Normal es capaz de capturar.

Referente a la desviación estándar, se ha de suponer que esta escale con respecto a δt^{α}, mientras que la varianza escale con respecto a $\delta t^{2\alpha}$, siendo esta aditiva en su construcción.

El observar una cantidad finita de retornos entre un tiempo inicial 0 y t da cuenta de una cantidad $\frac{t}{\delta t}$ de estas magnitudes, cada una con varianza $\delta t^{2\alpha}$. Agregar cada una de estas varianzas en el número total de retornos proporciona una varianza total de $\frac{t}{\sigma t} \cdot \delta t^{2\alpha}$. Para obtener una varianza positiva, finita y estable en el tiempo, en el límite $\delta t \to 0$ el valor de α debe corresponder a $\frac{1}{2}$, motivo por el cual la desviación estándar escala con respecto a la raíz cuadrada del *paso de tiempo*.

Consolidando lo señalado referente a la composición del retorno, el precio de un activo financiero puede ser modelado entonces en tiempo discreto como:

$$R_i = \mu\delta t + \sigma\phi\sqrt{\delta t} \to Y_{i+1} = (1 + \mu\delta t)Y_i + Y_i\phi\sigma\sqrt{\delta t} \tag{3.24}$$

De lo anterior, identificado como un proceso de Weiner y detallado en Shreve (2008), se desprende que el crecimiento y la volatilidad ejercen efectos distintos en la trayectoria del activo. En intervalos de tiempo reducido, la volatilidad predomina en su incidencia, mientras que en horizontes de tiempo prolongados lo hace el factor de crecimiento.

El factor $\sigma\sqrt{\delta t}$, al ser escrito como dW, corresponde a una variable aleatoria distribuida de forma uniforme con media 0 y varianza dt, referida como proceso de Weiner. Esto permite obtener un modelo de ecuaciones diferenciales estocásticas (SDE) de la forma:

$$dY = \mu Y dt + \sigma Y dW \tag{3.25}$$

3.16. Generador de precios estocásticos de tipo de cambio

El modelo de aprendizaje por refuerzo se nutre de precios acordes con la realidad de los distintos estados con la finalidad que el agente aprenda de su comportamiento e infiera movimientos de la economía.

Para la obtención de estos precios, se propone utilizar un modeló estocástico de valorización de divisas que toma como base la probabilidad neutral al riesgo sobre el precio de las mismas, esto a través de un modelo de movimiento browniano geométrico (GBM) de la forma siguiente:

$$dY_t = \mu Y_t dt + \sigma Y_t dW_t \tag{3.26}$$

Este puede ser revisado en Karatzas y Shreve (1998), donde W_t corresponde a un proceso estándar de Wiener bajo la medida neutral al riesgo, cuya dinámica se sustenta en la creación de un portafolio de no arbitraje entre un activo riesgoso y un activo libre de riesgo.

Lo anterior tiene por finalidad reducir la incertidumbre y obtener una valorización del activo, donde Y representa al precio del producto financiero, el factor μ corresponde a la tendencia del activo, mientras que σ a la volatilidad en el precio de este. El GBM tiene como particularidad el no permitir que los precios de los activos tomen valores negativos.

Un proceso de Wiener W_t tiene diversas características, dentro de las que destaca que sus incrementos siguen una distribución normal con media 0 y varianza igual al *paso de tiempo*. Esto permite que el GMB disponga de un componente aleatorio, símil al comportamiento de los activos financieros en intervalos de tiempo reducidos.

La relación anterior puede ser resuelta en Y a través de la aplicación del lema de Itô, el cual permite la obtención de soluciones analíticas de ecuaciones diferenciales estocásticas, cuyo resultado para la valorización de tipo de cambio queda expresada por:

$$Y_t = Y_0 \; exp\{(r_b - r_a)t - \frac{\sigma^2 t}{2} + \sigma dW_t\} \tag{3.27}$$

con r_a y r_b correspondientes a las tasas libres de riesgo del mercado local y del mercado estadounidense respectivamente. Por su parte, Y_t representa al valor del tipo de cambio en el instante t, permitiendo asignar un valor del USDCLP en pesos chilenos a cada estado del modelo de RL a través de la relación indicada en (3.27).

La presencia de r_a y r_b surge de la utilización conceptual de instrumentos de renta fija en la cobertura del tipo de cambio a través de un portafolio resiliente al riesgo.

Tomando un estado inicial en la metodología de RL, el precio de la unidad de dólares estadounidenses medida en pesos chilenos estará proporcionada como una condición inicial del sistema. Posteriormente, cada estado asignará un precio de USDCLP medido a través del modelo de GMB (3.27).

Este modelo tiene la propiedad de calibrarse en función de la volatilidad histórica de precios del activo. Como parámetro inicial se asignará la volatilidad del USDCLP calculada según el intervalo de tiempo más representativo, sumado a las tasas de interés referenciales al periodo.

El modelo presentado en (3.27) permite entonces asignar un valor al USD en pesos chilenos en cada estado según dinámicas económicas capturadas por tasas de interés y volatilidad de mercado. A este precio se debe aplicar los efectos de la decisión del agente en materia de cantidad de unidades de dólares a vender o comprar, motivo a ser detallado a continuación.

3.17. Abstracción del impacto de volumen de una transacción

Una práctica común del mercado financiero consiste en estimar el impacto de una orden de mercado en el precio de la transacción de la forma siguiente:

$$\Delta P = \text{Costo de spread} + \eta\ \sigma\sqrt{\frac{J}{V}} \tag{3.28}$$

con σ como la volatilidad diaria del precio del activo, J el volumen de la transacción materializada, V el volumen diario de operaciones del producto financiero y η un factor de ajuste.

Esta metodología consiste en una aproximación, omitiendo efectos de nivel de tipo de cambio, capitalización de los mercados y fricciones referentes a la plataforma de transacción. Adicionalmente es importante destacar que dicha metodología es insuficiente frente a eventos de compraventa muy agresivos o *outliers*.

En la práctica el costo de *spread*[19], correspondiente a un efecto en la liquidez del activo, se puede obviar de la estimación, haciendo referencia a mercados con mayor profundidad. Referente al parámetro α, su valor se desprende de calibrar datos históricos.

Los modelos o relaciones que abstraen el impacto de una orden en el mercado pueden utilizar como *benchmark* a la métrica de VWAP para contrastar su efectividad. Esta cuantifica una ponderación del precio según el volumen de sus transacciones.

Los mercados ponen en evidencia que el impacto de una orden por volumen presenta una forma cóncava, presentando robustez frente a la madurez de la inversión, a lo que se suma una consistencia entre los distintos periodos de tiempo.

El impacto de una orden en el mercado puede ser descompuesta en dos fases. La primera reflejando un salto en la valorización, consolidando un efecto de concavidad en el movimiento del precio, mientras que posteriormente se materializa una segunda fase correspondiente a un decaimiento del efecto hasta un nivel medio.

Finalmente, el impacto de mercado da cuenta de la profundidad del mismo. Mercados con reducido número de transacciones diarias en el activo, o cuyo nivel de capitalización es reducido, enfrentan una mayor sensibilidad al impacto de elevados niveles de volumen en una compra o venta.

Para el modelo de aprendizaje por refuerzo se aplicó la relación anterior a la definición de estado como un *feedback* del agente. Se buscó condicionar el precio del USDCLP del estado siguiente a la acción del agente en el estado precedente. Lo anterior se materializa condicionando el factor J a la acción ejecutada por el algoritmo, derivando así una causalidad entre las decisiones y un impacto de estas sobre los estados del sistema.

De esta forma, el agente incorpora en su aprendizaje, es decir, en los valores Q del modelo

[19]Una motivación para su incorporación se vincula a robustecer efectos de fricción en el precio producto de reducidas transacciones sobre un periodo determinado.

de RL, las consecuencias de sus acciones sobre la realidad que experimenta.

3.18. Métricas de riesgo de mercado

Sobre la mecánica de RL el agente ficticio debe ser capaz de comprender el sentido de riesgo de sus decisiones. Mantener una significativa cantidad de dólares por comprar implica un riesgo superior producto de la posibilidad de disminución del precio de la divisa en estados subsecuentes. Es debido a lo anterior que dicha característica debe ser implementada en la función de recompensa del agente como una penalización medida a través de una métrica de riesgo coherente.

Una métrica de riesgo de mercado se caracteriza por ser una medida de incertidumbre del valor futuro de un activo financiero, es decir, una medida de dispersión de los retornos. Su propósito se concentra en sintetizar una potencial desviación de un respectivo valor esperado o *benchmark*.

Para medir la dispersión indicada se debe tener en consideración la dispersión individual, como de su dependencia frente a movimientos colectivos. La volatilidad y correlación son métricas de riesgo que suelen ser suficientes cuando los retornos se vinculan a distribuciones multivariadas normales, situación no visible en la coyuntura.

Una práctica inducida por los organismos reguladores consiste en la aplicación de métricas de riesgo financiero vinculadas a derivaciones del Valor en Riesgo[20] (VaR). A esta métrica se suma el VaR Condicional o Pérdida de Cola Esperada (ETL) como una metodología complementaria y ampliamente usada en la industria. En la presente tesis se aplicará la metodología VaR para la representación de riesgo de mantener una determinada cantidad de dólares de inventario en la actividad del agente.

El VaR corresponde a la pérdida monetaria esperada ante un nivel determinado de probabilidad. La distribución regente sobre los precios de la divisa proporcionará información referente a la cantidad de pesos chilenos que podrían reducirse en la rentabilidad del agente con cierto nivel de certeza. Esta métrica puede ser profundizada en Jorion (1996).

Dentro de las principales ventajas de esta magnitud se encuentra su capacidad para ser comparable entre distintos mercados y niveles de exposición. Esta medida puede ser obtenida a distintos niveles de granularidad, desde activos y transacciones individuales hasta complejos portafolios de inversión.

Es importante destacar que las métricas de riesgo de mercado se relacionan a los horizontes de inversión y al nivel de precisión requerido por los analistas financieros, como también del grupo institucional que disponga de los activos. Las instituciones bancarias presentan como giro de negocio la incorporación de riesgo financiero, cobrando así una prima por la gestión de este. Por su parte fondos de pensiones buscan mitigar su riesgo inflacionario debido al horizonte de sus carteras.

[20]La métrica VaR incurre en un riesgo inherente debido a su deficiencia en no ser necesariamente sub aditiva.

Las instituciones consideran distintas estrategias referentes a los riesgos financieros. Algunas tratan de mitigar gran parte de este[21], mientras que otras buscan regular y habilitar una mayor exposición ante la búsqueda de un mayor rendimiento. Es debido a esto que las áreas internas de gestión de riesgo de las distintas instituciones establecen limites en su exposición.

3.19. Definición de la métrica de Valor en Riesgo

El Valor en Riesgo se define por la pérdida que no se debiese[22] sobrepasar al mantener el activo o portafolio riesgoso a lo largo de un intervalo de tiempo definido. El VaR se encuentra compuesto por 2 parámetros, correspondientes al nivel de significancia[23] y el horizonte temporal sobre el cual se busca medir la desvalorización potencial.

En términos numéricos, el VaR se representa por su nivel de significancia α y el horizonte temporal h que considera en la evaluación. Un VaR 100α h-tiempo es el monto de pérdida en valor presente que se excederá con una probabilidad α en el intervalo de tiempo h.

Lo anterior queda expresado como:

$$P(B_{ht}Y_{h+t} - Y_t < r_{ht}) = \alpha \tag{3.29}$$

con B_{ht} como factor de descuento para comparar los niveles de precio. Para el cálculo del VaR, una práctica común es la de asumir retornos independientes e idénticamente distribuidos (i.i.d.) provenientes de una distribución normal con media μ y varianza σ^2. De esta forma, el VaR 100α corresponde al resultado de la siguiente relación:

$$\text{VaR}_{ht,\alpha} = (\Phi^{-1}(1-\alpha)\sigma_h - \mu_{ht})Y_t \tag{3.30}$$

donde Φ corresponde a la función de distribución normal estándar.

3.20. Utilización del VaR en la metodología de Aprendizaje por Refuerzo

El valor resultante del cálculo del VaR proporciona al algoritmo de aprendizaje por refuerzo una métrica de riesgo sobre la cantidad de USDCLP remanente a ser comprada o vendida.

Para el caso de un objetivo de compra [24] de una cantidad determinada de dólares, el no adquirir la totalidad de USDCLP en un estado particular supone como riesgo un aumento

[21] Debido a fricciones de mercado, indisponibilidad de instrumentos financieros y la velocidad de ajuste en los precios de los activos, se imposibilita una cobertura completa del riesgo de los mismos.

[22] La condicionalidad se refiere a los eventos *outlier* que el modelo no es capaz de capturar, como aquellos valores resultantes de la violación de los supuestos de este.

[23] Es de practica usual que el nivel de significancia sea establecido por un organismo regulador externo a la institución que busca medir su nivel de riesgo.

[24] De forma opuesta sobre un objetivo de venta de USDCLP.

del valor de la divisa en el estado siguiente por condiciones de mercado. Este riesgo será capturado por el VaR, indicando al agente el impacto sobre su decisión.

La metodología del VaR permite informar al agente de las consecuencias de sus acciones a través de la función de recompensa, ajustando el beneficio inmediato de la operación con una pérdida potencial.

Capítulo 4

Implementación de la solución

4.1. Ejemplificación de una trayectoria del problema

Se abordará una trayectoria preliminar ficticia, simplificada y sencilla para observar y comprender de mejor forma el comportamiento del agente ficticio, las decisiones factibles y el desarrollo de la problemática. La siguiente ejemplificación estará caracterizada en un escenario de **venta** de dólares. Como se señaló, la acción de compra mantiene los mismos objetivos, desafíos y riesgos, con la diferencia en la magnitud y dirección de los parámetros.

Consideremos una institución A que busca vender US$ 40 millones durante un tiempo máximo de 50 minutos. Por motivos de simplicidad de la ejemplificación, A podrá vender en bloques o paquetes de US$ 10 millones, con un precio ficticio inicial del USDCLP de $ 700.

4.1.1. Decisiones

Las decisiones plausibles por el agente consistirán en la cantidad de dólares a vender en cada estado o situación venidera. A podrá vender cualquier cantidad entera de paquetes de US$ 10 millones hasta cumplir su objetivo de US$ 40 millones. El tiempo avanzará en un *paso de tiempo* de 10 minutos.

4.1.2. Estados y estado inicial

Para definir un estado, se deberá abstraer este a sus unidades fundamentales identificando los elementos que hacen única cada configuración. En este problema, el estado estará compuesto por el tiempo disponible para cumplir el objetivo, el saldo de dólares restante de A a ser vendido y el valor del USD en cada unidad de tiempo.

El estado inicial, cómo se mencionó, considerará la siguiente configuración: El tiempo será de 50 minutos, el saldo de dólares de US$ 40 millones y el valor del USD en $ 700 para este ejercicio.

Así, el estado 0 en la presente ejemplificación será representado por [50, 40, 700].

¿Cómo se generarán los estados siguientes?

Como se mencionó, cada estado se encuentra compuesto por 3 elementos: el tiempo disponible, el saldo de dólares y el precio de estos últimos.

El tiempo es secuencial e independiente. Para la problemática se define en minutos, aunque puede ser especificado en unidades más o menos pequeñas. Continuando con el ejercicio, al posicionarse el agente con un tiempo de 50 minutos, el siguiente estado deberá, por definición, estar especificado con un tiempo reducido en 10 minutos, repitiendo la sustracción hasta el final del ejercicio, identificado como la culminación de un episodio en un estado con tiempo de 0 minutos. Este último valor podrá ser superior si el agente vende la totalidad de su saldo anticipadamente.

El segundo elemento de un estado, el saldo de dólares se encontrará completamente vinculado a la decisión ejecutada en el estado precedente, con la excepción por construcción del estado inicial 0.

Si en el estado 0 A decide vender US$ 10 millones, entonces el estado 1 por definición tendrá al componente de saldo de dólares con una cuantía de US$ 30 millones. Si en dicho estado 1 A decide vender US$ 20 millones, entonces el estado 2 tendrá a este componente de inventario con una cantidad de US$ 10 millones, reflejando así la cantidad de dólares restantes a ser vendidos.

Es importante destacar que no podrán existir estados con un saldo de dólares negativo. Como se mencionó, los estados con saldo de dólares de 0 darán por finalizado un episodio.

El tercer elemento de un estado corresponderá al valor del USD. Este valor es influenciado por 2 fuerzas, la decisión de A y el comportamiento de mercado.

De esta forma, el precio del USD incorpora las ecuaciones (3.27) y (3.28) y queda definido por la relación:

$$\text{USD}_t = \text{USD}_{t-1} + \text{efecto mercado}_t + \text{efecto decisión } A_{t-1} \tag{4.1}$$

Para nuestro problema, la condición inicial del valor del USD será de un precio ficticio de $ 700.

Como ya se indicó, un volumen significativo podrá alterar el valor del USD en periodos inmediatamente subsecuentes. Supongamos que A decide vender US$ 10 millones en el estado inicial 0 y que este corresponde a un volumen que no logra alterar el valor del USD. Esto significará en nuestro simplificado modelo que el valor del USD en el siguiente estado, es decir, en el estado 1, solo será consecuencia del valor del USD en el estado 0 y de un efecto del comportamiento del mercado.

$$\text{USD}_t = \text{USD}_{t-1} + \text{efecto mercado}_t \tag{4.2}$$

El avance a un estado siguiente, como fue definido, implicará un avance temporal de 10 minutos.

Continuando con lo anterior, el valor entonces del USD en el estado 1 corresponderá a su valor en el estado precedente, $700 en este caso, más un efecto de mercado. Esto último significará un valor de incremento o pérdida en el USD generado durante estos 10 minutos por efectos económicos y financieros.

Estos efectos de mercado son generados por eventos en la economía que se encuentran ajenos a las actividades de A. Eventos políticos, publicación de indicadores económicos, decisiones de los bancos centrales, comportamiento de los *commodities*. Son un sin número de actividades que impactan al valor de una divisa, más a una tan globalizada como lo es el dólar estadounidense, donde un intervalo de 10 minutos podrá contener relevantes variaciones siendo capturadas por σ en la ecuación (3.27).

Supongamos, para este ejercicio, que A decide vender US$ 10 millones en el estado inicial, que transcurrido el avance del estado 0 al estado 1 el valor de USD registra un aumento de $3 por efectos de la economía y que la acción de venta no genera un impacto de mercado. Esto quiere decir que el valor del USD en el estado 1 será de $ 703 debido a la ausencia de efecto en la acción ejecutada por A sobre el precio de la divisa.

Para complementar el ejemplo, supongamos que A en el estado 0 no hubiese vendido US$ 10 millones, sino que US$ 30 millones y este valor si hubiese impactado en el valor del dólar. Presumamos que dicho impacto fuese de $ 1. Ahora el valor del siguiente estado, es decir, aquel que cuantifica 10 minutos posteriores al estado 0, considerará un valor de USD de $ 704. La dinámica del precio incorpora ahora la acción del agente, como señalado en la relación (3.2) y ajustado por la ecuación (3.28).

¿Es este estado el mismo que el del ejercicio de compra de US$ 10 millones? La respuesta es negativa. Este estado si bien cuantifica a aquel con un decremento en el tiempo de 10 unidades desde el tiempo restante de 50 minutos, considera también a aquel donde el saldo de dólares de A se actualizó a US$ 10 millones remanentes por vender y donde el valor del USD se modificó a $ 704.

Desde el estado inicial 0, cuyo vector queda dado por [50, 40, 700], se puede llegar a un estado [10, 30, 703] como también a un estado [10,10, 704], como a muchos otros. La acción del agente afectará el descubrimiento de un nuevo estado a través del inventario restante por vender, su influencia en el precio de la divisa y la volatilidad de mercado.

Por continuidad del tiempo y la definición simplificada del problema, entre estados inmediatamente consecutivos la unidad de tiempo reflejará un incremento de 10 unidades, pero el saldo de dólares de A y el valor del USD dependerá de las decisiones de la institución. Como fue señalado, el valor del dólar dependerá a su vez de un componente externo a la decisión de A, identificado como el efecto de mercado y que responderá de forma estocástica.

4.1.3. Recompensa

La decisión de A sobre cada estado determina una recompensa. Esta le permite a A discriminar las distintas decisiones y seleccionar la más favorable. Para esta ejemplificación, una recompensa adecuada deberá premiar el beneficio de una determinada decisión y castigar el

riesgo asociado a esta.

La fuente de beneficio es directa, correspondiendo a la ganancia de cada venta de dólares en el estado sobre el que se posicionará el agente ficticio. Por otro lado, el riesgo se hace presente a través de la posible desvalorización del USDCLP en los estados venideros. La cantidad de dólares remanentes por vender materializan un riesgo a través de una posible disminución del valor de la divisa en un estado futuro, repercutiendo en una reducción de la ganancia del agente.

De lo anterior, la recompensa se encuentra entonces determinada por la acción del agente a través de:

- La cantidad de dólares a ser vendidos (influencia directa del estado presente y del estado previo)
- El precio del USDCLP (influencia indirecta del estado previo)

Como ha sido señalado, una cantidad elevada de dólares a ser vendidos presiona la desvalorización de este producto por el comportamiento fundamental de oferta y demanda. Una decisión de este tipo en un estado particular alterará la recompensa del agente en el estado siguiente a través de su afectación en el precio de la divisa.

La magnitud de la recompensa debe estar compuesta en unidades representativas entre el beneficio y el riesgo que la componen. La unidad monetaria de CLP por unidad de dólar es adecuada para este modelo, donde el beneficio estará dado por la cantidad de CLP obtenidos por la venta total de dólares en dicho estado, mientras que el riesgo será identificado por la dispersión histórica del precio de la divisa en un horizonte de tiempo relevante. Dicha dispersión identificará, de forma simple, un rango monetario sobre el cual el USDCLP podría desvalorizarse[25].

De esta forma, la recompensa obtenida de la decisión en cada estado quedará expresada de forma simplificada como:

$$\text{recompensa} = \text{Cantidad}_t \cdot \text{USD}_{\text{precio}_t} - \text{Cantidad}_t \cdot \sigma_{\text{USD}} \tag{4.3}$$

La magnitud $\text{Cantidad}_t \cdot \sigma_{USD}$ corresponderá a una simplificación con fines explicativos de dispersión de la divisa que, como se detallará más adelante, será reemplazada por la métrica de Valor en Riesgo.

4.1.4. Transición de estados

El estado inicial quedará determinado por 2 componentes establecidos por el analista financiero, el tiempo restante para culminar el episodio y la cantidad de dólares a ser vendidos, como también por un componente observado del entorno, como lo es el precio del USDCLP.

[25]Una medida de dispersión como la desviación estándar tiene su fortaleza en la simplicidad y en representar el comportamiento en las mismas unidades que el objeto que mide. Para el comportamiento de desvalorización del USDCLP, sin embargo, tiene la debilidad de considerar movimientos favorables de la divisa. Se utilizará esta medida priorizando los beneficios que entrega, tomando en consideración que sobre estima el riesgo de la moneda.

Este último componente, en su evolución, permitirá que los estados sean descubiertos de forma estocástica y que los episodios se conformen según la distribución de probabilidad que gobierna el precio del dólar. La mecánica de resolución de este problema mediante RL permitirá obtener resultados significativos, aproximando la probabilidad de transición del precio de la divisa a través de la trayectoria del agente, su calibración y sus experiencias.

Por otro lado, lo anterior requiere intensificar el número de episodios para la producción de aproximaciones cada vez más robustas, esto con la finalidad de minimizar la existencia de estados con reducida o nula participación, y sobre los cuales la posibilidad de extraer aprendizaje es reducida o inexistente.

Ciertos eventos económicos y financieros, abstraídos en las variables representativas del modelo de valorización de la divisa, presentan mayor probabilidad de ocurrencia. Lo anterior trae consigo niveles de USDCLP experimentados con mayor frecuencia, esto como resultado de una distribución de probabilidad implícita en los precios que induce a trasladarse a ciertos estados con mayor asiduidad.

4.2. Aplicando métricas de precio y recompensa acordes con la realidad

Continuando con lo señalado en la ejemplificación anterior, se procederá a robustecer la estructura de cálculo de las métricas de precio, como también de recompensa, a aquellas a ser utilizadas por el algoritmo de aprendizaje por refuerzo.

Como fue mencionado, el precio del USD en pesos chilenos responde a un proceso *browniano geométrico* teórico según lo indicado en (3.27), el que se ajusta de acuerdo con la regla empírica de impacto de volumen según lo señalado en (3.28). Con esto, el valor del USDCLP a utilizar en la algorítmica de RL corresponde a:

$$M = \begin{cases} 1, & \text{compra divisa} \\ -1, & \text{venta divisa} \end{cases} \tag{4.4}$$

$$Y_t = Y_{t-1}\ exp\{(r_b - r_a)t - \frac{\sigma^2 t}{2} + \sigma dW_t\} + M\ \eta\ \sigma\sqrt{\frac{J}{V}} \tag{4.5}$$

Si bien la ecuación precedente complejiza lo señalado en (4.2), mantiene similar lógica. Un componente estocástico que incorpora el precio del periodo antecesor, ajustado por el efecto del impacto de mercado de la última decisión del agente.

Para las operaciones de compra de dólares, el ajuste por el impacto de la decisión es positivo, aumentando el nivel del USDCLP. Para las operaciones de venta la corrección ejercida se presenta en sentido opuesto.

Adicionalmente, se robustece la función de recompensa indicada en (4.3). El objetivo de penalizar una decisión según los saldos de dólares restantes por comprar o vender se hace

coherente al utilizar métricas vinculadas a su dispersión histórica (σ^2_{USDCLP}), donde el Valor en Riesgo asume el rol de forma efectiva para un tomador de decisiones automatizado.

De esta forma, la función de recompensa queda dada por:

$$\text{recompensa} = -M(\text{Cantidad}_t \cdot \text{USD}_{\text{precio}_t} - \text{VaR}) \tag{4.6}$$

cuya magnitud resulta de materializar la operación en el estado vigente, ajustada por el riesgo sobre el saldo por transar en los estados futuros.

La función de maximización incorporada en el algoritmo de aprendizaje por refuerzo busca identificar las decisiones que proporcionen la mayor rentabilidad, es decir, un valor de recompensa superior, esto considerando los beneficios y penalizaciones expuestas.

4.3. Implementación de la solución

4.3.1. Identificación y calibración de los parámetros del modelo

Los parámetros del modelo responden a magnitudes obtenidas de forma directa o indirecta por parte del analista financiero. Dentro de los parámetros de libre disposición se encuentran:

- El tiempo, de la forma de tiempo terminal del modelo representado por T, y del incremento temporal a ser identificado por *dt* o *paso de tiempo*. El tiempo T identifica el tiempo máximo[26] que dispondrá un episodio. El *paso de tiempo* o *dt* identificará el incremento temporal entre estados, mientras que la razón T/dt dará cuenta del número máximo[27] de estados por episodio.
- Direccionalidad de las operaciones de compra o venta de dólares. El parámetro M discrimina ambas operaciones de forma excluyente, cuya magnitud se define en valores de 1 y -1 para una compra o venta respectivamente.
- Las *acciones* del agente ficticio se representan por un parámetro definido por el analista en la forma de un arreglo unidimensional. Las decisiones o acciones ejecutadas por el algoritmo corresponden a un conjunto de magnitudes discretas que identifican el número de paquetes a ser adquiridos o vendidos por el agente ficticio a ser entrenado.
- El parámetro *paquetes* establece la cantidad o corte de dólares disponible a ser comprados o vendidos. Al ser transado el dólar en volúmenes elevados por parte del mercado institucional, este admite por parte de las plataformas transaccionales la convención de ser operado en múltiplos estructurados.
- El *volumen* corresponde al volumen de transacciones a ser consideradas por el periodo a analizar. Dicho valor es utilizado por la regla de impacto de mercado, por lo que se mantiene en congruencia con el parámetro *dt*.

[26] Un episodio podrá finalizar con una variable t inferior a T en el caso que en dicho estado se compre o venta, según sea el caso, la totalidad del saldo de dólares.

[27] El número de estados podrá ser inferior a T/dt si el agente reduce la totalidad de su inventario con anticipación.

- *eta* (η) identifica una magnitud de corrección proporcionado por la regla de impacto de mercado para ajustar dicha magnitud a valores empíricos.
- La desviación estándar del precio del dólar es capturada por el parámetro *sigma*. Este valor se determina por la dispersión histórica del USDCLP coherente con el *paso de tiempo* a ser considerado.
- Las tasas domésticas (r_d) y extranjeras (r_f) dan cuenta del costo del dinero en un tiempo mínimo en dólares y en pesos chilenos respectivamente.
- La probabilidad P es utilizada para discriminar la exploración e intensificación del agente. Dicho valor define, bajo una distribución binomial, la probabilidad de escoger una acción aleatoria.
- *gamma* (γ) representa al factor de descuento de un evento sucesivo, siendo este el costo de oportunidad monetario en pesos chilenos para un intervalo de tiempo dt.
- El valor en riesgo, identificado a través del VaR, condensa el riesgo de una decisión. Su valor se calibra en relación con la dispersión histórica de los movimientos del USD de acuerdo con el horizonte de tiempo o *paso de tiempo* a considerar.
- La puntuación z sobre las observaciones de riesgo, la cual informa del número de desviaciones estándar por sobre o por debajo de la media de población de datos.

4.4. Desarrollo del algoritmo

A continuación, se plasmarán los postulados anteriores en procedimientos que permitan la resolución de la problemática. Se procederá a segmentar cada componente, detallar su estructura y características.

El lenguaje de programación utilizado corresponde a Python. Sobre este se definió la clase USDCLP encargada de controlar el precio del USD en pesos chilenos y la clase Agente, la que incorporó los siguientes métodos:

- *decisión*: encargado de proporcionar una acción válida según los parámetros establecidos.
- *innovación*: responsable de retornar nuevos estados y la recompensa asociada.
- *aprendizaje*: facultado para almacenar el valor de $Q_t^{\pi^*}$ y entregar la política óptima.

4.4.1. Activo Financiero - USDCLP

El precio inicial del dólar se establece como un parámetro suministrado por el analista financiero, identificado como S_0. Este se recoge de la coyuntura y responde a su valor vigente en pesos chilenos. De todas formas, la flexibilidad de la mecánica permite utilizar valores

hipotéticos para realizar análisis de escenario, suministrando el resto de los parámetros de forma consistente.

El precio del USD en el estado siguiente dependerá de su valor en el estado actual, ajustado por su dispersión histórica, la acción del agente, las tasas de interés de mercado representativas, el tiempo transcurrido entre estados y una variable aleatoria[28], de acuerdo a lo señalado en (4.5). Como ha sido explicado, lo anterior proporciona a la dinámica de precios el comportamiento de la economía y los efectos de las decisiones del algoritmo.

Se establece un vector de precios gobernado por S y circundante a S_0 con la finalidad de limitar la cantidad de precios del dólar a ser modelados. Los valores de precio fuera de limite se asignarán a los valores extremos.

El parámetro M condiciona la actividad de compra o de venta que se busca modelar, cuyo valor incide en ajustar el valor económico de la divisa a través de la regla de impacto de mercado. Esto último reduce el precio del USDCLP para operaciones de venta cuyo volumen sea significativo, mientras que para las acciones de compra proporciona un aumento de la paridad.

El parámetro *sigma* es proporcionado por la calibración, acorde con los periodos de análisis a ser considerados.

Como fue indicado, la dinámica ajusta a valores enteros y dentro de límite para el valor de la moneda. Si bien esto supone una disminución en la precisión del algoritmo y una alteración en la función de distribución de probabilidad, permite adaptar este precio a una grilla de valores de fácil manejo para los estados representados por los Q_{Values} respectivos. Una aplicación sencilla de lo señalado se observa en la línea 9 del algoritmo.

```
1  class USDCLP ():
2      def __init__(self, S0, sigma):
3          self.precio = S0
4          self.precios = np.zeros(S * 2 + 1).tolist() #El vector precio
           ↪  tendrá una distancia simetrica
5          for i in range(len(self.precios)):
6              self.precios[i] = self.precio - S + i #El vector precio
               ↪  abarcará S pesos superiores e inferiores a S0
7      def nuevo_precio(self, accion):#dinámica de precio indicada en (35)
8          self.precio = self.precio * np.exp((rf - rd)*dt - (sigma**2)*dt/2
           ↪  + (sigma)*np.sqrt(dt)*np.random.normal(0,1)) + M * sigma *
           ↪  1/2 * np.sqrt(accion*activos/volumen) * self.precio
9          self.precio = int(np.clip(np.ceil(self.precio), self.precios[0],
           ↪  self.precios[-1])) # El precio será entero y abarcará hasta
           ↪  los extremos
```

[28]El valor proviene de una distribución normal con media 0 y varianza 1.

Si a modo de ejemplificación consideramos un precio inicial S_0 de 700, un valor de σ de 0.007, un valor de r_f de 0.25 % y de r_d de 0.5 % para 25 estados de tiempo consecutivos, la trayectoria del precio del activo queda expresada en la Figura 4.1 al seguir la dinámica estocástica señalada. Para su cálculo se consideraron acciones aleatorias del agente dentro del dominio factible.

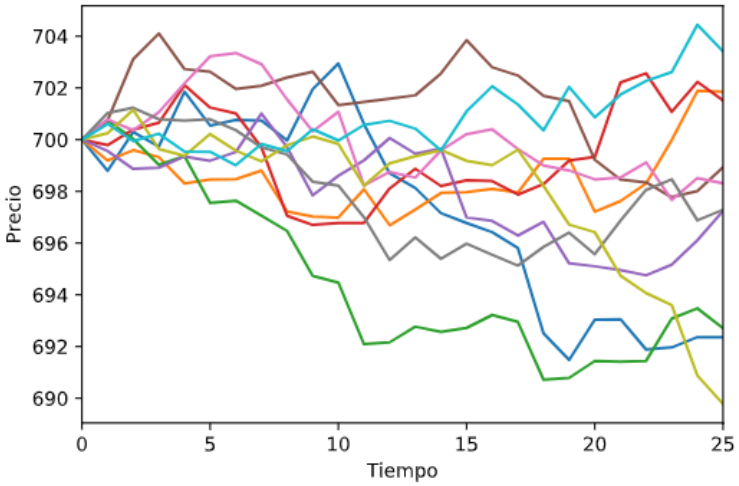

Figura 4.1: Precio modelo estocástico del USDCLP con $S_0 = 700$

La trayectoria de S_0 es gobernada por los parámetros r_d, r_f y $sigma$, representando la tasa de interés doméstica, la tasa de interés foránea y la dispersión histórica de la divisa. La diferencia de r_f y r_d se conoce como el diferencial de tasas de 2 economías y reproduce el costo de oportunidad de capitales[29]. El resto de los eventos económicos serán capturados por $sigma$ para proporcionar robustez al modelo.

Asumir 3 variables para describir los efectos la economía sobre una divisa es un supuesto elevado. Para un diferencial de tasas de magnitudes considerables existe un sesgo significativo por parte del modelo de precios, exacerbando apreciaciones o depreciaciones del tipo de cambio según corresponda. Esto último toma un rol protagónico frente a $sigma$, donde eventos adicionales como el precio de los *commodities* u otros tendrán una menor incidencia, distanciándose de lo observado de forma empírica.

A pesar de lo anterior, el modelo permite reducir dichas fricciones mediante un aumento en el número de simulaciones y proporcionar a un tomador de decisiones datos útiles para su aprendizaje.

Es importante destacar una dificultad adicional para un modelamiento basado en el aprendizaje por refuerzo. La amplía dispersión en los precios del activo y el sesgo en la frecuencia de ocurrencia de estados producto de la distribución del precio del dólar lleva consigo eventos que suceden con una baja probabilidad de ocurrencia, dificultando materializar un aprendizaje en el Q_{value} respectivo.

[29]De acuerdo a los eventos analizados por la paridad cubierta de tasas de interés de la economía nacional, un incremento de la tasa libre de riesgo local frente a la tasa libre de riesgo foránea impulsará un fortalecimiento de la moneda local.

Si bien lo anterior reduce la muestra sobre la cual recoger aprendizaje, el modelo de RL continúa proporcionando conocimiento a través del resto de los estados.

4.4.2. Agente - Parámetros

A continuación, se define al agente ficticio y sus correspondientes métodos encargados de efectuar el aprendizaje de las dinámicas del USD, siendo estos el centro de la metodología de RL. El agente comienza en el estado $t = 0$[30] y tiene como objetivo comprar o vender una cierta cantidad de paquetes discretos de USD. Para cumplir dicho objetivo, este podrá comprar o vender una cantidad finita de paquetes en cada estado del modelo. Por cada acción ejecutada sobre el correspondiente estado, el agente almacena la recompensa de dicha decisión, la cual siendo positiva o negativa impacta sobre la recompensa acumulada del episodio.

```
class Agente ():
    def __init__(self, tipo_aprendizaje = 'QLearning', paquetes = paquetes):
        self.tipo_aprendizaje = tipo_aprendizaje
        self.activo = USDCLP(S0, sigma)

        self.t = T
        self.paquetes = paquetes
        self.acciones = [0, 1, 2, 3] # Siempre se permitirán 0 elementos a
        ↪  ser transados, equivalente a postergar una decisión.

        self.recompensa_acumulada = 0
        self.q_value = np.zeros((T + 1, self.paquetes + 1, S * 2 + 2,
        ↪  len(self.acciones)))
```

El problema se modela considerando estados que se componen por:

- Un tiempo t donde se posiciona el agente.
- Un nivel de paquetes o inventario restante para completar el objetivo.
- Un precio S del USD asociado a cada estado particular.

El tiempo es discreto y finito, culminando en el fin del horizonte temporal que el agente dispone para comprar o vender dólares, o en el estado donde el saldo de paquetes se reduzca completamente. El nivel de inventario considera los paquetes por comprar o vender del agente. El precio del USD, como se señaló, podrá ser influenciado por las decisiones ejecutadas.

Por su parte, las metodologías de RL consideran la recursividad de la relación de las decisiones, estados y recompensa en un valor comparable para determinar la importancia de cada

[30]Este estado estará a su vez representado por un valor de la divisa S_0 y un inventario restante definido por el total de dólares a comprar o vender.

evento en función de la acción ejecutada. Lo anterior se almacena en el Q_{value} del algoritmo, permitiendo así jerarquizar los distintos estados del modelo y de esta forma calibrar un aprendizaje del agente.

Para este desarrollo el agente almacena en un tensor de 4 dimensiones el respectivo Q_{value}. De esta forma, para los 3 elementos que conforman un estado (tiempo, inventario y precio) y la decisión ejecutada sobre estos, se asigna una dimensión correspondiente a cada ítem señalado y vinculada al Q_{value} respectivo.

Al comienzo del episodio, el tensor Q_{value} estará compuesto por el valor de 0[31] en todas sus dimensiones. A medida que el agente avance en su recorrido actualizará el Q_{value} respectivo de acuerdo a lo representado en (3.17) y (3.18), incorporando los beneficios o pérdidas de las trayectorias vinculantes.

[31] Algunas metodologías de RL asignan un valor inicial diferente para forzar la exploración del agente, aunque en la presente tesis se obviará dicha implementación por la marginalidad de su beneficio.

4.4.3. Agente - Decisión

El agente ficticio debe incorporar la capacidad de tomar decisiones por lo que para lograrlo se le adapta un método que le permite dicha facultad. El método *decisión* será el primer paso para que el agente determine una decisión racional.

Como fue señalado, la metodología de los algoritmos de RL consiste en utilizar el Q_{value} como medida comparativa del valor de una decisión en un determinado estado. Por ese motivo, el agente escoge aquella política que proporcione el mayor Q_{value} según las trayectorias experimentadas por el agente.

Adicionalmente, el agente deberá tener la facultad de explorar su entorno y evitar intensificar sobre óptimos locales. Esta competencia se le incorpora utilizando una variable aleatoria binomial. El agente selecciona una acción basada en la racionalidad del Q_{value} o escogerá una decisión aleatoria que permita diversificar la trayectoria del episodio y experimentar estados que pudieran no ser seleccionados inicialmente. Lo anterior corresponderá a las dinámicas de intensificación y exploración de RL y serán controlados a través de la mencionada variable a través de una probabilidad de ocurrencia.

```
def decisión(self, P = probabilidad):
    if np.random.binomial(1, P) == 1:
        self.accion = np.random.choice(self.acciones)
    else:
        self.valores_ = self.q_value[self.t, self.paquetes,
        ↪  self.activo.precios.index(self.activo.precio), :]
        #Si dos o mas acciones entregan el máximo qvalue, escoger dicha
        ↪  accion de forma aleatoria
        self.accion = np.random.choice([accion_ for accion_, valor_ in
        ↪  enumerate(self.valores_) if valor_ ==
        ↪  np.max(self.valores_)])

    if self.accion > self.paquetes:
        self.accion = self.paquetes
```

Aprovechando la estructura de tensor en la que se almacenará el Q_{value}, se selecciona aquella posición en la dimensión de las decisiones que maximiza dicho valor. A su vez, la incorporación de una distribución de probabilidad como discriminador entre explorar o intensificar permite regular la capacidad de descubrir nuevas trayectorias potencialmente beneficiosas para el agente.

La flexibilidad del modelo permite a su vez instaurar políticas de exploración inicial en las acciones de compra o venta de USD durante las primeras etapas de un episodio, entregando así una mayor dispersión en la experimentación, para luego disminuir dicha probabilidad y concentrar los esfuerzos de aprendizaje en aquellos recorridos más promisorios.

Adicionalmente se incorpora una restricción de integridad para no comprar o vender más paquetes de los que restan en inventario.

4.4.4. Agente - Innovación

La innovación corresponde a la transición a un nuevo estado por parte del agente ficticio. Este recorre los episodios en intervalos constantes de tiempo dt[32], donde en cada estado descubrirá un nuevo precio del USD, actualizará su registro de tiempo actual t como también su nivel inventario de la divisa.

En el afán del algoritmo por desarrollar su aprendizaje, debe disponer de una métrica comparativa sobre sus decisiones. Como fue expuesto con anterioridad, dicho concepto se materializa en la función de recompensa del agente según las relaciones exhibidas en (4.6).

Al completarse la transición de estados, la mecánica de RL proporciona dicha recompensa para así dar comienzo a la siguiente iteración del proceso.

El algoritmo define un Valor en Riesgo adaptado al inventario del estado a ser considerado. Lo anterior expondrá el riesgo asociado a la posible pérdida de valor de estos últimos en instancias futuras. La recompensa incorpora este riesgo en forma de unidades monetarias nacionales y permite calcular el beneficio o pérdida efectivo y potencial concretado por las decisiones del agente. Este valor corresponde a la diferencia entre el beneficio de la transacción y el riesgo sobre el inventario restante.

Como ha sido desarrollado con antelación, la recompensa de cada estado adquiere relevancia al ser considerada en conjunto con el resto de los estados que conforman una política de decisión, es decir, a través del Q_{value}.

```
def innovación(self):
    self.activo.nuevo_precio(self.accion)
    self.t = self.t - 1
    self.paquetes = self.paquetes - self.accion

    self.recompensa = activos * self.accion * self.activo.precio -
    ↪  activos * self.paquetes * z * np.sqrt(sigma)
```

La variable z establece y estandariza una referencia que, según lo ya señalado, permite incorporar de forma coherente los efectos de incertidumbre estadística en la métrica de VaR.

[32]También identificados como *paso de tiempo*.

4.4.5. Agente - Aprendizaje

El aprendizaje del agente ficticio se realiza mediante la metodología Q-Learning y SARSA, las que consideran el Q_{value} de la innovación y del estado actual para calibrar recursivamente la acción más favorable del agente. Cada estado almacena una recompensa representativa del episodio de forma secuencial, permitiendo una métrica del beneficio total alcanzado una vez culminado el estado final.

El agente ejecuta la metodología hasta reducir completamente su inventario de dólares o alcanzar el tiempo t de 0 [33], siendo este último la única variable independiente que permite de forma sencilla modelar la trayectoria del agente. Los otros elementos que dan conformidad a un estado, correspondiendo al nivel de inventario y al precio del USD, se descubrirán a medida que el agente avance en cada unidad dt de tiempo discreto.

Para una mayor simplicidad en la estructura del código, el tensor que describe el estado se descompone en sus elementos fundamentales, para así estructurar de forma sencilla la ecuación que da origen a las dinámicas de RL.

Cada valor Q_{value} del tensor será propenso a sufrir actualizaciones según la acción que ejecute el agente, permitiendo determinar una acción racional y descubrir la realidad a través de los estados que se le presentan.

```
    def aprendizaje(self, probabilidad):

        if(self.tipo_aprendizaje == 'SARSA'):
            self.decisión(probabilidad)
            self.accion_presente = self.accion

        while (self.t >= 0) and (self.paquetes >= 0):

            self.precio_actual =
            ↪  self.activo.precios.index(self.activo.precio)
            self.estado_actual = [self.t, self.paquetes, self.precio_actual]

            self.decisión(probabilidad)
            self.innovación()
            self.precio_futuro =
            ↪  self.activo.precios.index(self.activo.precio)
            self.recompensa_acumulada += self.recompensa

            if(self.tipo_aprendizaje == 'SARSA'):
                self.accion_futura = self.accion
                error = self.q_value[self.t, self.paquetes,
                ↪  self.precio_futuro, self.accion_futura] #SARSA
            elif(self.tipo_aprendizaje == 'QLearning'):
```

[33] En cuyo caso de existir un saldo de inventario se procede a la compra o venta total, según sea el caso, al precio de dólar vigente en dicho estado.

```
    error = np.max(self.q_value[self.t, self.paquetes,
    ↪  self.precio_futuro, :]) #Q-Learning

self.q_value[self.estado_actual[0], self.estado_actual[1],
↪  self.estado_actual[2], self.accion] += alfa *
↪  (self.recompensa
+ gamma * error
- self.q_value[self.estado_actual[0], self.estado_actual[1],
↪  self.estado_actual[2], self.accion])

if(self.tipo_aprendizaje == 'SARSA'):
    self.accion_presente = self.accion_futura
```

Capítulo 5

Prueba conceptual del modelo y metodología

Para la evaluación de la solución se establecieron variables que simularon un comportamiento económico tradicional, ajustándose a una configuración hipotética con el fin de evaluar el desempeño del método y su viabilidad en el proceso de aprendizaje.

La aptitud de aprender se midió a través de una recompensa acumulada. Esta registró el cúmulo total entre las distintas recompensas obtenidas en cada estado durante el transcurso de los distintos episodios, permitiendo así trazar el nivel de aprendizaje conforme el aumento de experiencias por parte del agente ficticio. Entendiendo la aleatoriedad en el rendimiento de un episodio particular, cada uno de estos se ejecutó una cantidad determinada de veces, procediendo a la obtención de la media de su recompensa y con esto, un valor más representativo en la calidad del aprendizaje.

Una evaluación satisfactoria del modelo y de su aplicabilidad a la problemática que atañe a la presente tesis refiere al cumplimiento de los *Requisitos y restricciones a la solución* expuestos en el capítulo 3, destacando el ítem de *Estrategia*. Con relación a esto, el evento más significativo de este proyecto consiste entonces en la determinación de la capacidad de aprendizaje de la metodología y de su comportamiento frente a una resolución heurística en similar escenario.

El rendimiento del agente, vinculado al nivel de recompensa alcanzado, no representa un interés superlativo para la presente evaluación. Esto es debido a su dependencia en la calidad de las funciones de recompensa y de modelamiento de la divisa, elementos con una elevada ponderación tanto financiera como económica y cuya efectividad puede ser mejorada por una precisión en la abstracción de la realidad.

Es debido a lo anterior que el presente documento tiene por objetivo evaluar la aplicación de técnicas de aprendizaje por refuerzo sobre una problemática financiera particular, permitiendo así discernir su aplicabilidad en la industria. La evaluación se concentrará en determinar la calidad y capacidad de aprendizaje del agente, su cumplimiento con los requerimientos mínimos expuestos y la coherencia de la solución, entendiendo que su desempeño puede ser mejorado por una mayor exactitud en la representatividad de la economía.

5.1. Parámetros del Ejercicio

En la subsecuente ejemplificación de la metodología se establecieron los siguientes parámetros:

```
N = -1 # Operaciones de venta de USD

probabilidad = 0.1 # probabilidad de exploración
gamma = 0.99 # factor de descuento
alfa = 0.9 # tasa de aprendizaje
sigma = 0.09 # volatilidad del USDCLP

activos = 10000 # cuota de dólares por paquete
volumen = 800000 # cantidad de transacciones en el periodo a evaluar
paquetes = 12 # homogéneos

z = 2.33 # nivel de confianza

S0 = 700 # precio inicial del activo
S = 3 # discretización del precio del activo en un espacio muestral de S *
↪  2 + 1 posibilidades
T = 4 # término de las operaciones, las que abarcan 4 días de transacciones
↪  en un régimen de innovación diario
dt = 1/252 # incremento de la innovación temporal, representando una
↪  convertibilidad diaria de tasas y factores anualizados

rf = 0.25/100 # tasa de referencia foránea
rd = 1.0/100 # tasa de referencia local
```

La probabilidad de 0.1 se aplicó a la metodología de decisión del agente. Con un 10 % de probabilidad el agente ejecutó decisiones aleatorias sobre un estado particular.

El factor de descuento *gamma* se estableció en 0.99 para este ejercicio, permitiendo calibrar el beneficio futuro con respecto al beneficio presente. Para las instituciones financieras el costo del dinero entre un reducido número de días es mínimo, por lo que un beneficio futuro no representa una mayor relevancia en escalas de tiempo pequeñas[34].

Los activos corresponden a 10000 y los paquetes a 12 en este ejercicio. Lo anterior solo con finalidad de flexibilizar la contabilización de los dólares a ser comprados o vendidos. El agente tendrá la capacidad de comprar o vender únicamente paquetes discretos.

El precio S_0 será de \$700. Adicionalmente, la cantidad de valores para el precio que permite el algoritmo se definió en 7, distribuidos simétricamente.

[34]En economía el factor de descuento representa, en conjunto con la tasa de interés, el costo del dinero en el tiempo. Para un modelo donde el tiempo evoluciona por periodos superiores a un día (*overnight*), el factor de descuento *gamma* a considerar decrecerá en relación con los días que se consideren entre estados. Esto debido a la tasa *overnight* y los costos de oportunidad de las instituciones financieras.

El tiempo estuvo definido en 4 unidades, las que por conveniencia representaron 4 días de actividad del agente ficticio, donde cada estado quedó identificado por incrementos diarios en la unidad temporal.

5.2. Ejecución del algoritmo

El algoritmo ejecutó 100 episodios registrando el aprendizaje en el tensor de Q_{Values}. Una vez finalizado un episodio, se procedió a reiniciar al agente. Este mantuvo el aprendizaje de los episodios que lo precedieron, perdurando los Q_{Values} respectivos. Esto último permitió al agente conservar la experiencia de una trayectoria y mejorar su desempeño seleccionando decisiones que maximizasen el Q_{Value} respectivo.

Por cada episodio se ejecutaron 200 repeticiones con el objetivo de disminuir el impacto de la aleatoriedad del modelo, calculando la media de la recompensa acumulada de la totalidad de las 100 trayectorias experimentadas.

Una vez completada una ejecución, se reinició la totalidad del algoritmo. El agente olvidó todo su aprendizaje y recorrió nuevamente los episodios con el objetivo de un nuevo muestreo de las trayectorias.

Lo anterior queda representado como:

```
episodios = 100
ejecuciones = 200

x = Agente('QLearning')
y = Agente('SARSA')

recompensa_media_x = np.zeros(episodios)
recompensa_media_y = np.zeros(episodios)

for i in range(ejecuciones):
    x.reinicio_entorno()
    y.reinicio_entorno()

    for j in range(0, episodios):
        x.aprendizaje(probabilidad)
        y.aprendizaje(probabilidad)

        recompensa_media_x[j] += x.recompensa_acumulada
        recompensa_media_y[j] += y.recompensa_acumulada

        x.reinicio_episodio()
        y.reinicio_episodio()
```

```
recompensa_media_x /= ejecuciones
recompensa_media_y /= ejecuciones
```

Por su parte, los métodos de reinicio del episodio y de su ejecución quedan definidos como parte del objeto Agente de la forma:

```
    def reinicio_episodio(self, t = 0, paquetes = paquetes):
        self.activo = USDCLP(S0, sigma)

        self.t = T
        self.paquetes = paquetes
        self.recompensa_acumulada = 0

    def reinicio_entorno(self, t = 0, paquetes = paquetes):
        self.q_value = np.zeros((T + 1, self.paquetes + 1, S * 2 + 2,
        ↪  len(self.acciones)))
```

5.3. Resultados del Ejercicio

Se presenta la recompensa media entre ejecuciones de episodios, observándose en la figura 5.1.

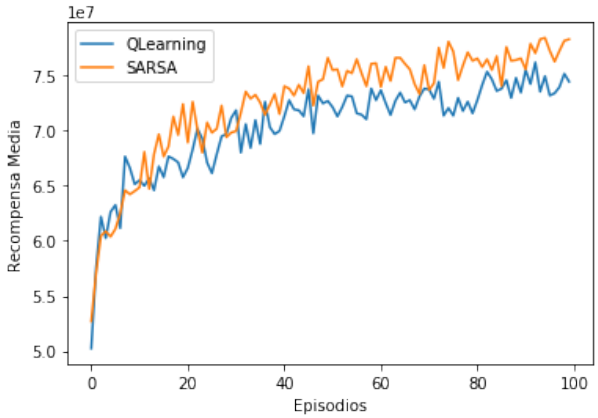

Figura 5.1: Recompensa media de la totalidad de episodios

De lo anterior se desprende un incremento en el aprendizaje del agente conforme el aumento en la cantidad de episodios ejecutados. Al disponer de una mayor experiencia a lo largo de las trayectorias, el agente es capaz de incrementar su entendimiento, seleccionando decisiones que aporten una mayor recompensa, la que se comunica entre los distintos estados y de forma implícita entre los distintos episodios a través del Q_{value}.

El aprendizaje observado por parte de la algorítmica SARSA se muestra superior frente a la metodología QLearning. Debido a la política *e-greedy* de este último, el agente ficticio es capaz de asumir y materializar acciones que incurren en un mayor riesgo y por consiguiente en una penalización frente a la mecánica SARSA.

Finalizado el episodio inicial ambas metodologías exhibieron una recompensa media reducida para posteriormente incrementar rápidamente dicha magnitud al avanzar en el número de episodios. Lo anterior se condice con la reducida experiencia y conocimiento del entorno por parte del agente en su comienzo, el que conforme expande sus vivencias a través de un mayor número de trayectorias potencia el entendimiento del sistema. Posteriormente y a medida que el número de episodios aumenta significativamente, la recompensa media se incrementa marginalmente, esto debido a que frente a la totalidad de conocimiento almacenado en los Q_{value} respectivos un nuevo entendimiento ve reducida su relevancia.

El resultado de la calibración para la metodología Q-Learning se representa en forma de un tensor *Tiempo x Inventario x Precio* según:

```
array([[[0., 0., 0., 0., 0., 0., 0.],
        [1., 1., 0., 0., 0., 0., 1.],
        [2., 2., 2., 0., 0., 0., 2.],
        [3., 0., 2., 0., 2., 1., 1.],
        [2., 0., 1., 3., 0., 0., 2.],
        [1., 1., 1., 0., 2., 3., 0.],
        [2., 3., 1., 0., 0., 3., 0.],
        [2., 2., 0., 0., 3., 0., 0.],
        [1., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.]],

        --- 1 matriz omitida ---

       [[0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [3., 3., 2., 0., 3., 2., 1.],
        [3., 3., 2., 3., 1., 2., 0.],
        [1., 0., 1., 0., 0., 0., 1.],
        [0., 1., 0., 0., 2., 1., 3.]],

       [[0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 3., 0., 0., 0.]]])
```

Al descomponer el tensor mostrado, cada matriz representa el tiempo en el cual se posiciona el agente, donde las filas de estas representan el nivel de inventario de USD a ser vendidos y las columnas al precio de la unidad de dólar vigente en dicho estado. El valor de cada elemento del tensor representa la acción óptima a ejercer para cada configuración de estados.

El algoritmo modelado proporciona una política que aproxima a la política óptima de operaciones, correspondiendo en este ejercicio a las instrucciones de venta de USD. Lo anterior proporciona la acción óptima que debe ejecutar un analista financiero en una determinada unidad de tiempo, al presentarse el correspondiente nivel de precio de la divisa y el inventario respectivo.

Como ya ha sido señalado, la política óptima anterior se obtiene del algoritmo de RL que considera la decisión más conveniente para una configuración secuencial de estados futuros. La metodología permite entregar un entorno de trabajo para un agente ficticio que le facilite tomar decisiones de compra y venta considerando los riesgos y beneficios asociados.

Por su parte la metodología SARSA entrega los siguientes resultados:

```
array([[[0., 0., 0., 0., 0., 0., 0.],
        [1., 1., 0., 0., 0., 0., 0.],
        [1., 0., 0., 0., 0., 0., 0.],
        [2., 0., 0., 0., 0., 2., 1.],
        [2., 1., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [2., 2., 0., 0., 1., 0., 0.],
        [0., 0., 1., 0., 0., 1., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [1., 0., 0., 0., 0., 3., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.]],

        --- 1 matriz omitida ---

       [[0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0.],
        [3., 2., 2., 2., 1., 0., 3.],
        [3., 1., 0., 2., 0., 0., 0.],
        [2., 0., 0., 0., 0., 0., 0.],
        [0., 2., 0., 0., 0., 3., 3.]],
```

```
[[0., 0., 0., 0., 0., 0., 0.],
 [0., 0., 0., 0., 0., 0., 0.],
 [0., 0., 0., 0., 0., 0., 0.],
 [0., 0., 0., 0., 0., 0., 0.],
 [0., 0., 0., 0., 0., 0., 0.],
 [0., 0., 0., 0., 0., 0., 0.],
 [0., 0., 0., 0., 0., 0., 0.],
 [0., 0., 0., 0., 0., 0., 0.],
 [0., 0., 0., 0., 0., 0., 0.],
 [0., 0., 0., 0., 0., 0., 0.],
 [0., 0., 0., 0., 0., 0., 0.],
 [0., 0., 0., 0., 0., 0., 0.],
 [0., 0., 0., 3., 0., 0., 0.]]])
```

Los resultados muestran decisiones que aspiran a una mayor recompensa por parte de la metodología Q-Learning con respecto a SARSA. Lo anterior se condice con la propiedad *on-policy* de la primera metodología, donde en cada estado prioriza decisiones orientadas a maximizar el Q_{Value} respectivo.

5.4. Interpretación de los resultados

El objetivo de este modelo, como fue señalado, consistió en obtener el valor óptimo de compra o de venta de USD para los distintos niveles de inventario, precio del USDCLP y tiempo transcurrido.

La estructura utilizada para mostrar dichas combinaciones corresponden al tensor precedente, donde el cúmulo de matrices bidimensionales representan el tiempo transcurrido, y donde la dimensión fila y la dimensión columna de cada matriz representan el nivel de inventario y el precio del USDCLP respectivamente. El valor asociado a cada uno de estos componentes indica el monto de compra o venta de USD.

La anterior representación permite de forma sencilla expresar los distintos estados que conforman las trayectorias de decisiones y la selección de acciones asociadas por la metodología. A modo de ejemplo, si se considera el vector resultante de la selección [2, 10, :] para la metodología Q-Learning se obtiene:

```
[3., 3., 2., 3., 1., 2., 0.]
```

Esta porción del tensor proporciona como información qué restando 10 paquetes de USD a ser vendidos, una vez transcurrida 1 unidad de tiempo desde el inicio del ejercicio, la configuración de decisiones es la proporcionada con anterioridad para los niveles de precio modelados. A modo de explicación y considerando la configuración indicada, para un nivel de precio igual o inferior a $697 la cantidad de paquete de dólares a ser vendida debe ser de 3 unidades, mientras que para un precio de $ 701 se deberá vender 1 unidad de paquete de dólares, siendo estas instrucciones proporcionadas por la optimalidad del algoritmo.

El tiempo transcurrido, como variable independiente, nos permite conectar los distintos estados. Continuando con el tensor anterior, el estado inicial queda identificado por:

```
[0., 0., 0., 3., 0., 0., 0.]
```

Este vector indica qué al comenzar la jornada de operaciones, la decisión a considerar corresponde a la venta de 3 paquetes de unidades de USD. Con excepción del precio de $ 700 indicado en la dimensión columna correspondiente, el resto de los elementos registra un valor de 0. Esto se explica para la configuración de inventario correspondiente al total de dólares por vender y la totalidad de tiempo por transcurrir, donde el único precio factible de USDCLP corresponde a la condición inicial, manifestada en el centro del vector mencionado.

El tensor anterior permite a su vez el disponer de combinaciones secuenciales estructuradas a través del tiempo. Prosiguiendo con la ejemplificación, la configuración [4, 12, 3] → [3, 9, 2] expresa la transición entre el estado inicial, materializando la decisión óptima de vender 3 paquetes de dólares y el estado posicionado en la siguiente unidad temporal, el que registra un inventario de 9 unidades a un precio del USDCLP de $ 699, cuya decisión óptima conlleva la venta de 2 paquetes de dólares informados a través del valor asignado a dicho índice.

Cada configuración y combinación de dimensiones permite obtener la cantidad óptima a transar. La política óptima se devela al posicionarse en las distintas coordenadas del tensor en forma secuencial. Una vez situado en un índice [X, Y, Z], cuyo valor A representa a una decisión óptima del estado particular, el siguiente estado queda indicado por [X - 1, Y - A, Z]. La dimensión Z permite discriminar según el precio observado en el mercado la decisión óptima a seleccionar.

5.5. Heurística como metodología comparativa

Se utilizó a modo de comparativa una modificación de la metodología, donde el agente tomó decisiones aplicando una heurística común, basada esta en la minimización del riesgo y la aplicación de decisiones con menor aversión a este únicamente frente a señales de mercado específicas y beneficiosas.

Lo anterior queda representado por la estrategia de vender la mínima cantidad de dólares al posicionarse el precio del USDCLP por debajo del promedio histórico de la simulación [35].

Esto permite tener un resultado base comparativo para el desempeño del algoritmo y revisar si efectivamente un aprendizaje basado en RL supera a las decisiones heurísticas tradicionales y comunes.

[35] Se establece el promedio de la simulación y no del episodio para capturar la mayor cantidad de trayectorias posibles del precio del dólar.

Para lograrlo, se modificó la clase USD y el método *decisión* del agente de la forma:

```
class USDCLP ():
    #(...)
        self.precios_historicos = [self.precio]
    def nuevo_precio(self, accion):
        #(...)
        self.precios_historicos.append(self.precio)

def decisión(self):
        if(self.activo.precio * N <
        ↪  (sum(self.activo.precios_historicos)/len(self.activo.precios_historicos)))
        ↪  * N: # N discrimina la acción de compra o venta
            self.accion = max(self.acciones)
        else:
            self.accion = 1

        if self.accion > self.paquetes:
            self.accion = self.paquetes
```

Al ejecutar las trayectorias, la recompensa media entre ejecuciones de episodios quedó dada de la siguiente forma:

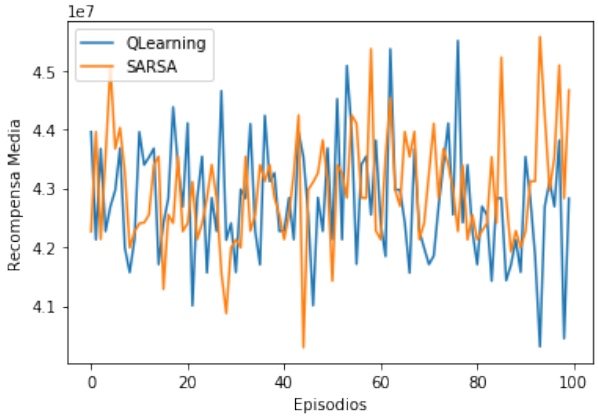

Figura 5.2: Recompensa media de la totalidad de episodios

La recompensa media de las 200 ejecuciones muestra que el agente mantiene una cota de rendimiento por debajo de la metodología basada en Q_{Values}, con una elevada dispersión en el rendimiento e imposibilitada de mantenerse rentable en el largo plazo.

Capítulo 6

Recapitulación y conclusiones

El desarrollo de la presente tesis proporcionó un entorno de estudio sobre técnicas tradicionales de aprendizaje por refuerzo a problemáticas financieras, concerniendo estas a decisiones de compraventa de una divisa considerando los riesgos de mercado y el impacto de sus volúmenes en las ordenes de operación.

Observando los resultados expuestos en el capítulo antecesor, la metodología demostró el cumplimiento del objetivo principal del presente trabajo, correspondiendo este a la capacidad de generar una política óptima de decisiones de acuerdo con la configuración de un sistema de precios para el USDCLP, su nivel de inventario y el tiempo dispuesto para el cumplimiento de la tarea.

Las metodologías Q-Learning y SARSA demostraron a través de la evolución de la media del cúmulo de recompensa la capacidad para aprender del agente sobre la simulación experimentada. Esta última consideró variables tradicionales para la metodología de RL, como también una representación cotidiana para los elementos económicos y financieros.

Este aprendizaje se mostró consistente conforme el aumento en el número de episodios, destacando el perfeccionamiento del entendimiento del entorno frente a un acotado número de parámetros y estructuras, mostrando así la competencia de la metodología para inferir las reglas de negocio más convenientes y concretarlas en una política óptima.

Considerando la situación base comparativa, se procedió a seleccionar una heurística que capturase elementos tradicionales observados en la industria, siendo el sesgo hacia comportamientos históricos un referente en las operaciones por parte de los analistas financieros. Para materializar lo anterior, se definió una política que transase el mínimo número de paquetes durante cada estado, con la excepción de un traspaso[36] en el precio de la divisa frente a su media histórica, en cuyo caso el agente operase el límite de sus posibilidades. Como se observó de los resultados, la política anterior introdujo un sesgo significativo que imposibilitó alcanzar de forma consistente un rendimiento elevado, haciendo de su desempeño un acto con significativa dispersión.

[36]Para el caso de una operación de venta, dicho traspaso refiere a un valor de la divisa por sobre su media histórica. En un evento de compra, se hace referencia al comportamiento opuesto.

Referente a los restantes objetivos específicos descritos en el capítulo 1, la metodología logró plasmar y ser compatible con estados coherentes, cohesionados y representativos de la realidad. Se estableció una función de recompensa capaz de equilibrar los riesgos futuros de una decisión en conjunto con los beneficios de un actuar en el presente, complementado estos con una función de modelamiento del USDCLP que permitiese describir trayectorias para el entrenamiento de la algorítmica.

Adicionando a lo anterior, la solución analizada concretó de forma satisfactoria los requerimientos propuestos en el capítulo 3. Como se señaló, esta proveyó de una política óptima estratégica al considerar un elevado y complejo número de escenarios en su determinación. A su vez, la metodología permitió alcanzar la solución con elevada celeridad, a lo que se sumó su flexibilidad a través de su construcción modular.

Con respecto a la fiabilidad de los resultados, la solución presentó limitantes en la abstracción del entorno financiero y económico con respecto a la realidad. La función de recompensa y de modelamiento en el precio del dólar proporcionaron directrices para discernir y discriminar el sistema percibido por el agente, aunque careciendo de un amplio abarque de condiciones financieras significativas. En la metodología se consideró el riesgo económico en una única variable, mientras que el riesgo potencial se vinculó a una métrica sencilla como el VaR y la implicancia de volumen en una regla empírica, descartando así elementos más robustos para simular y capturar señales y eventos de mercado.

Es por lo señalado que la investigación y consideración de los métodos de aprendizaje por refuerzo analizados satisficieron el objetivo encauzado, aunque habilitando un espacio de mejora destinado a depurar la percepción de la realidad financiera del agente.

Lo anterior tuvo el propósito de disminuir la complejidad financiera y económica al mínimo funcional con el objetivo de evaluar la capacidad de aprendizaje del agente, absteniéndose de distorsiones producto de la complejidad de modelos económicos y financieros.

6.1. Conclusiones complementarias del presente trabajo

Una vez finalizado el ejercicio y analizadas las observaciones se desprenden diversas conclusiones a ser consolidadas en las siguientes temáticas.

6.1.1. Materialización de aprendizaje

El algoritmo fue capaz de aprender en el acotado sentido que la estructura de RL refiere, demostrando empíricamente que el aprendizaje por refuerzo efectivamente entrega resultados prometedores en modelar dinámicas de compra y venta de activos financieros que siguen un comportamiento estocástico de precios.

6.1.2. Discriminación en riesgo-retorno

El agente ficticio tuvo competencia en reconocer y discriminar decisiones mediante una compensación entre penalizaciones y retornos. Mediante la aplicación de la función de recompensa el agente pudo materializar acciones al identificar el beneficio y riesgo de su decisión y el impacto futuro que esta generó.

6.1.3. Flexibilidad, modularidad y adaptabilidad de la solución

La algorítmica acepta la extensión modular de sus componentes facilitando así sustancialmente el perfeccionamiento del análisis. Debido a que el agente calibra sus decisiones de acuerdo con su percepción de los estados que experimenta, el proporcionar mayor veracidad en estos se traduce en un progreso del aprendizaje. Aplicar una modelación del precio del USDCLP más representativa requiere actualizar únicamente dicho objeto en el modelo, característica apetecida por los usuarios.

Complementariamente y dado que la interpretación del riesgo se materializa en la función de recompensa del agente, esta se puede extender para incorporar nuevos eventos que le permitan interpretar de mejor forma la dinámica de los estados. Adicionar representaciones de los movimientos económicos de distintas variables como el precio del cobre o la oferta de dólares interbancario podrían potenciar los resultados del aprendizaje del agente.

Adicionalmente, la estructura de tensor permite una gran flexibilidad en el almacenamiento de las acciones del agente y su representación. Para este ejercicio se orientó el resultado a un tensor de tres dimensiones que representó el estado del agente y cuyo valor entregó la acción óptima de este. Esto permitió el poder tener una representación de las distintas configuraciones de estados y de la decisión más favorable para un uso sencillo por parte del analista financiero.

6.1.4. Impacto de la aleatoriedad en el aprendizaje

Los resultados del ejercicio presentaron dispersión en el aprendizaje proporcionado por el componente aleatorio presente en la dinámica utilizada para modelar el precio del USD. Lo anterior conllevó a la necesidad de incrementar el número de episodios para mitigar su efecto.

6.1.5. Capacidad para analizar acotados dominios del problema

Para este ejercicio se utilizaron valores acotados para el precio S de una unidad de dólar. Incrementar la precisión a fracciones de dólares trae consigo una merma en la capacidad de revisitar estados, reduciendo la posibilidad de obtener aprendizaje.

6.1.6. Estados con limitada experiencia y extracción de aprendizaje

Configuraciones de la matriz de Q_{Values} finalizaron la calibración con valor de 0, siendo este su valor inicial. Lo anterior se desprendió del hecho que el algoritmo experimentó de forma nula o muy acotada dichos eventos. Esto supone una condición intrínseca del modelo debido a que existe un número de estados sobre los cuales el agente no genera aprendizaje.

Lo mencionado se relaciona con la distribución de precios utilizada para modelar el nivel del USDCLP y su varianza, donde algunas configuraciones exhiben una baja probabilidad de ocurrencia.

6.1.7. Comparativa frente a situación basal

Las metodologías de aprendizaje por refuerzo sobrepasan en rendimiento a la situación base aleatoria. La interconexión entre las experiencias de los distintos episodios permitió al agente explorar trayectorias de forma efectiva, mantener un aprendizaje y ejecutar decisiones que permitiesen incrementar su recompensa acumulada.

6.1.8. Capacidad para modelar eventos probabilísticos

La dinámica de RL incorpora en su interior probabilidades de transición entre estados, las que son encontradas de forma iterativa, aunque subyacente, mediante la calibración del Q_{Value}. Esto supone una metodología relevante y beneficiosa en la aproximación de dichas probabilidades debido a la dificultad que representa el poder identificarla mediante modelos analíticos.

6.1.9. Algoritmo de caja negra

La metodología de aprendizaje por refuerzo hereda los problemas de interpretabilidad presentes en las técnicas más comunes de aprendizaje de máquina. Debido a la compleja iteración de los componentes y su relación entre los distintos episodios e iteraciones se dificulta el trazar de forma precisa un resultado, como sensibilizar la causalidad de sus componentes.

6.1.10. Reducido número de variables explicativas

La economía es un proceso de interacciones entre colectividades significativamente complejo. Capturar y simular dicha información en acotadas variables como las tasas de interés y la volatilidad histórica del USDCLP induce a una modelación de la dinámica de precio con reducida representatividad de la realidad.

6.1.11. Rapidez de cálculo de resultados

Los métodos de Q-Learning y SARSA al aplicar las técnicas de Diferencias Temporales aseguran una rapidez en la resolución, la cual escalará con el número de episodios y ejecuciones. Lo anterior favorece a las dinámicas de una mesa de dinero.

6.1.12. Complejidad de modelar precios intradías

La dinámica de precios y su simulación se dificulta cuando los intervalos tienen duraciones cada vez más reducidas, acercándose al orden de pequeños fragmentos de un día laboral, esto debido al impacto de la volatilidad y costos de oportunidad que se alejan de las dinámicas tradicionales.

6.2. Trabajo futuro

La aplicación de algoritmos de aprendizaje de máquina en la industria financiera ha fortalecido las técnicas y tecnologías cuantitativas aplicada a esta ciencia durante los últimos años. Desde la detección de patrones de comportamiento en la rentabilidad de los activos hasta la optimización de portafolios de inversión, los métodos de ML han potenciado y permitido soluciones altamente sofisticadas y creativas.

Los métodos de aprendizaje por refuerzo no han sido la excepción, permitiendo resolver problemáticas secuenciales que involucran la toma de decisión con una amplia cuantía de posibilidades e interacción entre sus componentes.

Para el presente trabajo, la mecánica de aprendizaje por refuerzo alcanzó resultados satisfactorios, proporcionando una política de decisión a ser aplicada en distintos escenarios representados por los estados del entorno. Es importante destacar que dicha política de decisión admite una mejora continua producto del perfeccionamiento de la función de recompensa y la interpretación de los estados a través de la función de modelación del precio del USDCLP.

Medrar el desarrollo de la función de modelamiento del dólar en pesos permitirá progresar el entendimiento de los estados, permitiendo al agente recabar un aprendizaje más fidedigno del comportamiento de la divisa y, por consiguiente, decisiones amparadas en una distribución de probabilidad reducida en su sesgo. La incorporación de modelos con mayor robustez a escenarios *intraday*, como el extender la incorporación de redes neuronales y *deep learning* en la predicción del nivel del USDCLP podría incorporar avances altamente prometedores.

De igual forma ampliar la función de recompensa a la incorporación de una mayor gama de eventos económicos permitirá mejorar el entendimiento de las señales de costo-beneficio de las decisiones seleccionadas.

Lo anterior complementará el desarrollo del presente trabajo a políticas que presenten resultados plausibles y coherentes con las dinámicas financieras regentes.

Glosario

A continuación, se presenta un catálogo de palabras con la finalidad de homogeneizar la terminología específica utilizada en la presente tesis.

Activo Financiero: Título por el que cual un comprador obtiene derecho a recibir un ingreso futuro.

Aprendizaje por Refuerzo: Especialización del aprendizaje automático la cual busca determinar acciones a seleccionar por un agente ficticio en un entorno dado con el fin de maximizar alguna noción de beneficio.

Aprendizaje de Máquina: Especialización de las ciencias de la computación y una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan.

Apreciación (divisa): Aumento del precio de una moneda con respecto a otra o una referencia.

ASK: Precio más bajo al que el vendedor está dispuesto a transar un activo.

BID: Precio más alto que el comprador está dispuesto a transar un activo.

Commodities: Bienes genéricos que no representan mayor diferencia como resultado productivo.

Descalce Cambiario: Diferencia entre los ingresos y egresos monetarios producto de la tenencia de activos o deuda en una moneda distinta a la funcional.

Depreciación (divisa): Disminución del precio de una moneda con respecto a otra o una referencia.

Diferencial de Tasas de Interés: Sustracción entre tasas de interés que cuantifican distintas magnitudes financieras.

Divisa: Moneda extranjera referida a la unidad de cuenta del país de que se trata.

Estocástico: Procesos cuya evolución en el tiempo tiene un componente aleatorio.

Heurísticas: Técnicas de indagación y descubrimiento basadas en la pericia y experiencia.

Movimiento Browniano Geométrico: Modelo de amplio uso en finanzas el cual permite representar el precio de algunos activos que presentan componentes aleatorios.

Libro de órdenes: Registro de órdenes de compra y venta de valores u activos financieros en la totalidad del rango de precio.

Outlier: Observación numéricamente distante del resto de los datos.

Paridad de Cambio: Relación de intercambio entre dos monedas extranjeras.

Política (aprendizaje por refuerzo): Conjunto de estrategias que utiliza un algoritmo para decidir qué acciones llevar a cabo.

Política Monetaria: Actividad pública orientada a estabilizar la moneda de una localidad.

Prima: Importe a pagar por un determinado activo financiero.

Retorno: Recompensa producto del beneficio recibido o esperado.

SARSA: Algoritmo rígido de aprendizaje de políticas de decisión sobre procesos de decisión de Markov.

Spot: Referido a los mercados financieros considera a aquellas transacciones realizadas con un pago contra entrega.

Trading: Actividad especulativa sobre instrumentos financieros con el objetivo de obtener un beneficio.

Transacción: Trato, convenio o negocio materializado entre contrapartes.

Q-Learning: Algoritmo flexible de aprendizaje de políticas de decisión sobre procesos de decisión de Markov.

Bibliografía

[1] S. Sutton, R., & G. Barto, A. (2018). Reinforcement Learning: An Introduction (2.nd ed.). Bradford Books.

[2] F.Dixon, M., Halperin, I., & Bilokon, P. (2020). Machine Learning in Finance (1.th ed.). Springer.

[3] Graesser, L., & Keng, W. L. (2019). Foundations of Deep Reinforcement Learning (1.th ed.). Addison-Wesley Professional.

[4] López de Prado M. (2018). Advances in Financial Machine Learning (1.th ed.). Wiley.

[5] Jorion, P. (2006). Value at Risk: The New Benchmark for Managing Financial Risk (1.th ed.). McGraw-Hill Education.

[6] Hull, J. (2018). Risk Management and Financial Institutions (5.th ed.). Wiley. Wiley.

[7] Hull, J. (2021). Options, Futures, and Other Derivatives. (11.th ed.). Pearson.

[8] Röman, J. (2017). Analytical Finance: Volume I: The Mathematics of Equity Derivatives, Markets, Risk and Valuation (1.th ed.). Palgrave Macmillan.

[9] Kenyon, C., & Stamm, R. (2012). Discounting, LIBOR, CVA and Funding: Interest Rate and Credit Pricing (1.th ed.). Palgrave Macmillan.

[10] Gatheral, J., No-dynamic-arbitrage & market impact, Quantitative Finance 10(7) 749–759 (2010). [Gatheral and Schied] Jim Gatheral and Alexander Schied, Optimal Trade Execution under Geometric

[11] Gatheral J., & Schied A., Optimal Trade Execution under Geometric Brownian Motion in the Almgren and Chriss Framework, International Journal of Theoretical and Applied Finance 14(3) 353–368 (2011).