Reinforcement Learning para la Optimización de Portafolios

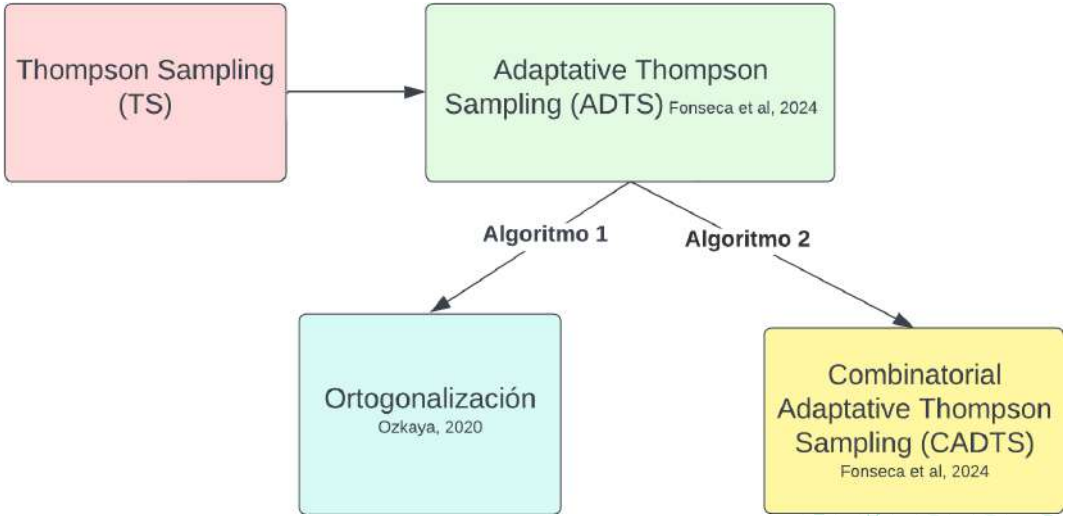
Camilo Díaz Ardila - Ana María Patrón Piñerez

January 30, 2025

Optimización de portafolios

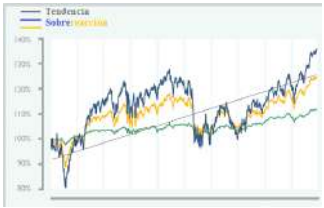
- **Objetivo:** Determinar la mejor combinación de activos (acciones, bonos, fondos, etc.) dentro de un portafolio, para maximizar el ratio retorno riesgo.
- **Desafíos:** la dinámica/no estacionariedad de los mercados financieros y la correlación entre activos.

Reinforcement Learning para la Optimización de Portafolios



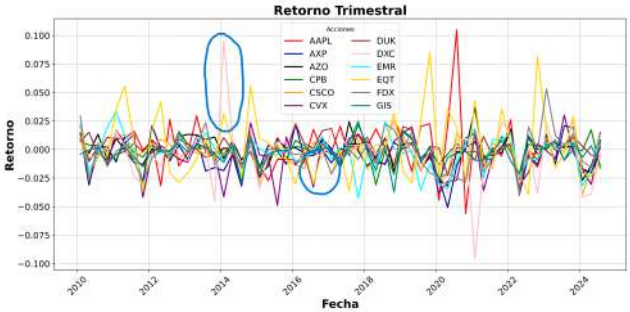
Motivación

- Métodos tradicionales: teoría de portafolios de Markowitz y el Modelo de Precios de Activos de Capital (CAPM)
- Multi-Armed Bandit (Thompson Sampling):
  - Incorpora trade off entre exploración y explotación para toma de decisiones secuenciales bajo incertidumbre = equilibrar riesgo y retorno (Huo y Fu, 2017).
  - TS supone distribuciones estacionarias → Se incumple en el mercado financiero.
  - Extensiones Adaptative Thompson Sampling (ADTS) y Combinatorial Adaptative Thompson Sampling (CADTS) (Fonseca, Silva y Castro, 2024)



Contexto

- Métodos tradicionales
- Multi-Armed Bandit (Thompson Sampling):
  - Extensiones: Adaptative Thompson Sampling (ADTS) y Combinatorial Thompson Sampling (CADTS) (Fonseca, Silva y Castro, 2024)
- Además, portafolios pueden estar correlacionados entre sí.



Tomado de Refinitiv, 2024

Contribución

En adición a las correcciones que logra ADTS y CADTS, la extensión que proponemos trata de:

1. Eliminar las correlaciones entre activos **vía** → proceso de ortogonalización previo.
2. Incorporar la información disponible sobre el pasado para tomar decisiones en el presente **vía** → modifición de la actualización de ADTS.

El Problema Multi-Armed Bandit

El Problema Multi-Armed Bandit

- Contamos con K armas $\{1, \ldots, K\}$, cada uno asociado a una **probabilidad desconocida de éxito** θ_k.
- Nuestro objetivo es maximizar la recompensa acumulada esperada:

$$\max \mathbb{E}\left[\sum_{t=1}^{T} r_{k_t}\right],$$

donde:

  - $r_{k_t} \in \{0, 1\}$ es la recompensa del brazo k_t en el tiempo t,
  - T es el horizonte temporal.

- Necesitamos decidir, en cada paso t, qué brazo tirar balanceando:
  - **Exploración**: Probar armas para aprender sus valores θ_k.
  - **Explotación**: Elegir el brazo que maximiza la recompensa esperada.

Componentes del Modelo Bayesiano: Introducción (Russo et al, 2020)

- Queremos modelar la **distribución posterior** $p(\theta_k \mid r)$, que refleja nuestra creencia actual sobre θ_k después de observar r.
- Según la regla de Bayes:

$$p(\theta_k \mid r) = \frac{p(r \mid \theta_k) \cdot p(\theta_k)}{p(r)}.$$

- Elementos clave:
  - $p(\theta_k)$: la **distribución prior**, que representa nuestras creencias iniciales.
  - $p(r \mid \theta_k)$: la **distribución estructural** o verosimilitud, que describe la probabilidad de los datos observados.
  - $p(r)$: la **evidencia**, que normaliza la distribución posterior.

Componentes del Modelo Bayesiano: Definiciones

1. **Distribución prior** $\rightarrow p(\theta_k)$:
   - Usamos una **distribución Beta(α, β)**, donde:

$$p(\theta_k) \propto \theta_k^{\alpha-1}(1-\theta_k)^{\beta-1}.$$

2. **Distribución estructural** $\rightarrow p(r \mid \theta_k)$:
   - Supongamos r condicional a θ_k sigue una distribución Bernoulli con parámetro θ_k:

$$p(r \mid \theta_k) = \theta_k^r(1-\theta_k)^{1-r}.$$

3. **Distribución Posteriori** $\rightarrow p(\theta_k \mid r)$:

$$\begin{aligned} p(\theta_k \mid r) &\propto p(\theta_k) \cdot p(r \mid \theta_k) \\ &\propto \theta_k^{\alpha-1}(1-\theta_k)^{\beta-1} \cdot \theta_k^r(1-\theta_k)^{1-r} \\ &\propto \theta_k^{\alpha-1+r}(1-\theta_k)^{\beta-1+(1-r)} \sim \text{Beta}(\alpha+r, \beta+(1-r)). \end{aligned}$$

El algoritmo (TS)

Algorithm 1 Thompson Sampling (TS)

Input: $K \geq 2$ armas

1: **for** $t = 1, \ldots, T$ **do**
2: **for** $k = 1, \ldots, K$ **do**
3: $\theta_k(t) = \mathcal{B}(\alpha_k, \beta_k) \leftarrow$ samplear.
4: **end for**
5: $I(t) = \arg\max_k \theta_k(t)$; $r_{I(t)} \leftarrow$ observar
6: **if** $k = I(t)$ **then**
7: $\alpha_k \leftarrow \alpha_k + r_k,\ \beta_k \leftarrow \beta_k + (1 - r_k)$
8: **end if**
9: **end for**

Adaptative Thompson Sampling (ADTS)

- Thompson Sampling supone que las distribuciones de las armas son **estacionarias en el tiempo**
- En el ámbito financiero, es natural considerar que las **distribuciones de los retornos** de las acciones cambian constantemente a lo largo del tiempo.
- **ADTS** busca relajar este supuesto, mediante incorporar la modelación de la distribución de retorno, como una **mezcla** entre una distribución de corto y largo plazo.
- La distribución de largo plazo es **estática** y corresponde a la misma de Thomson Sampling original, mientras que la de corto por definición es no **estacionaria**.

Adaptative Thompson Sampling (ADTS) - Ejemplo: Apple

- En ciertos contextos, puede ser una aproximación adecuada pensar en que la distribución del retorno de una acción puede estar determinada por una mezcla entre un componente de corto y largo plazo.

- Apple es una empresa muy consolidada en la mente de los consumidores, lo que puede ser un factor que afecta positivamente, en el largo plazo, los retornos de Apple.
- No obstante, en el corto plazo pueden ocurrir choques que afecten positiva o negativamente la imagen de la compañía. Por ejemplo, el fracaso en ventas de las gafas de realidad virtual.

Algoritmo: Thompson Sampling con Descuento Adaptativo (ADTS)

Algorithm 2 Thompson Sampling con Descuento Adaptativo (ADTS)

Input: $K \geq 2$ armas, $\gamma \in (0,1]$ factor de descuento, $w \in \{1, .., T\}$ ventana deslizante

for $t = 1, \ldots, T$ **do**
  for $k = 1, \ldots, K$ **do**
    Obtener las últimas w recompensas para el brazo k
    Definir α_k^w y β_k^w como el número de éxitos y fracasos pasados, respectivamente.
    $\theta_k(t) = \mathcal{B}(\alpha_k, \beta_k)$, $\tilde{\theta}_k(t) = \tilde{\mathcal{B}}(\alpha_k^w, \beta_k^w) \leftarrow$ doble sampleo.
  end for
  $I(t) = \arg\max_k \frac{\theta_k(t) + \tilde{\theta}_k(t)}{2}$, $r_{kt} \leftarrow$ observar
  for $k = 1, \ldots, K$ **do**
    $\alpha_k \leftarrow \gamma\alpha_k + r_{kt}$, $\beta_k \leftarrow \gamma\beta_k + (1 - r_{kt})$
  end for
end for
Output: Arrays de tamaño T con la política tomada, los retornos observados, dummies (éxito o fracaso).

Ejercicio Teórico

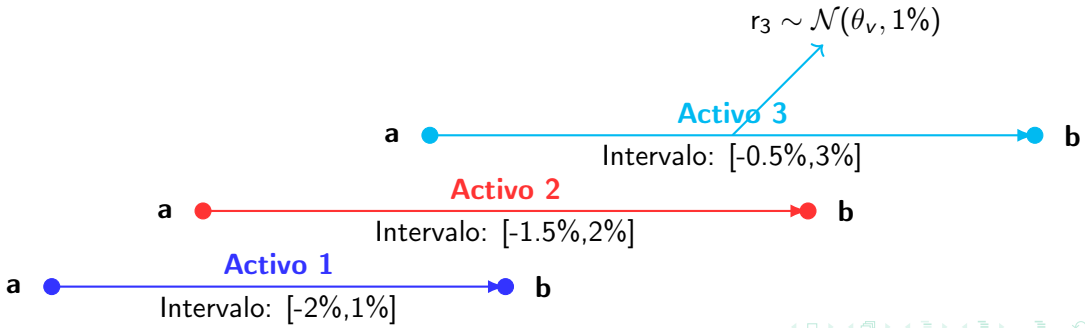
Simulación de Activos Financieros

- Simulamos un conjunto de **12 activos financieros** para un período de **7 años** (*1825 días hábiles*).
- Para cada activo, modelamos los retornos con dos componentes:
  - **Componente de largo plazo:**
    - Cada activo está asociado a una distribución **uniforme** sobre un intervalo fijo.
    - Este intervalo es *único y constante* durante los **1825 días**.
  - **Componente de corto plazo:**
    - Cada **121 días** las distribuciones de los retornos cambian.
    - Para cada ventana:
    - Sampleamos un parámetro de una distribución **uniforme**.
    - Usamos este parámetro para generar una distribución **normal**:
    - Media: el parámetro sampleado.
    - Varianza: **1%**.

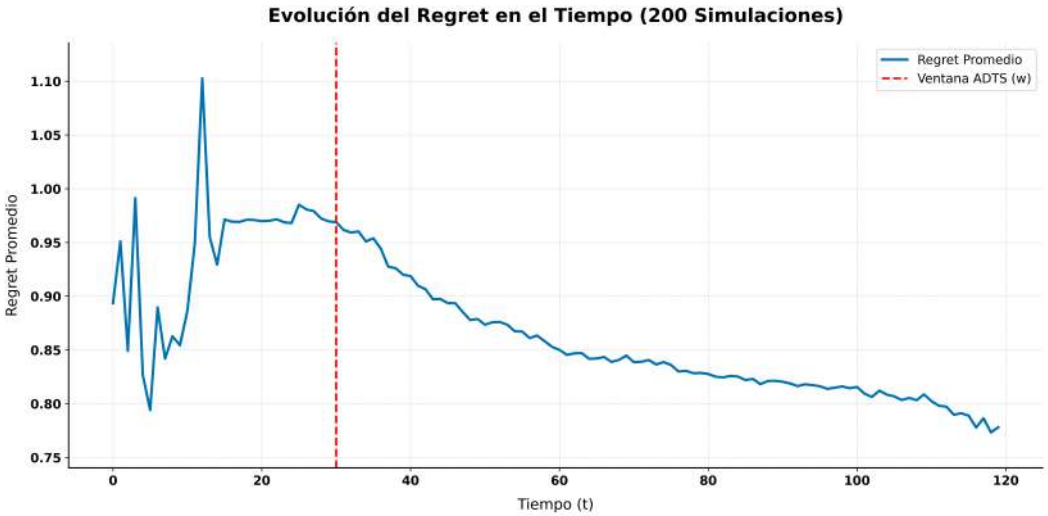
Ejercicio Teórico

Retornos como una "mezcla" de largo y corto plazo

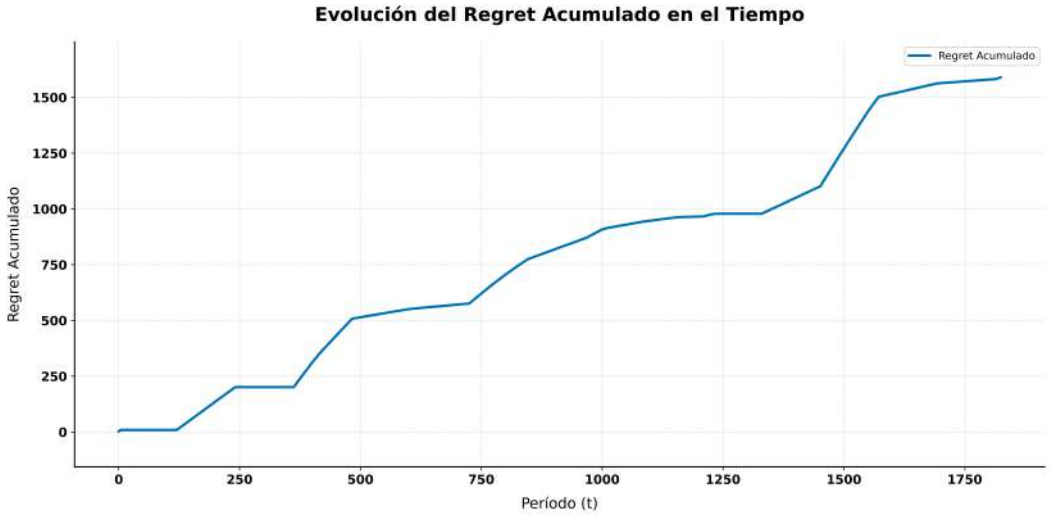
Cada 121 días hábiles, la distribuciones cambian: se define la distribución del retorno de cada acción como una normal donde la media es muestreada de una distribución uniforme que permanece estática.



Regret Promedio por período en la venta de tiempo (200 simulaciones)



Regret Acumulado (200 Simulaciones)



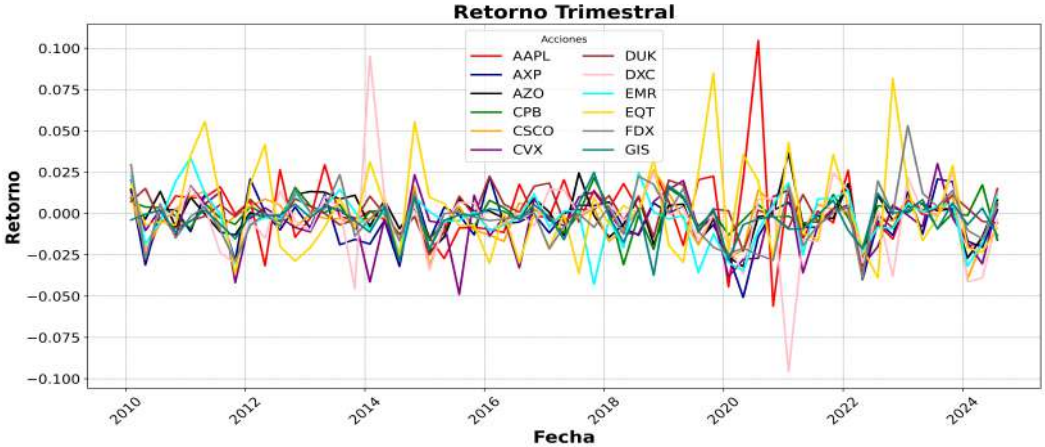
Descripción de los datos

- Empresas con las mayores capitalizaciones de mercado en Estados Unidos (índice S&P 500)
- Retornos de 12 acciones entre enero de 2010 y octubre de 2024 (Fuente: Refinitiv).

Descriptivas de los Retornos

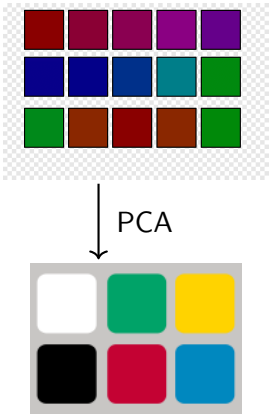
Acción	Obs	Media	Min	25%	50%	75%	Max	Std
AAP	373K	0.02%	-35.04%	-0.88%	0.04%	1.00%	16.56%	2.13%
AAPL	373K	0.11%	-12.86%	-0.75%	0.09%	1.04%	11.98%	1.76%
AXP	373K	0.07%	-14.82%	-0.71%	0.07%	0.92%	21.88%	1.83%
AZO	373K	0.09%	-15.94%	-0.63%	0.08%	0.84%	11.78%	1.47%
CPB	373K	0.02%	-12.37%	-0.64%	0.05%	0.71%	10.11%	1.37%
CSCO	373K	0.04%	-16.21%	-0.67%	0.04%	0.80%	15.95%	1.64%
CVX	373K	0.03%	-22.12%	-0.76%	0.06%	0.82%	22.74%	1.70%
DUK	373K	0.03%	-11.50%	-0.54%	0.06%	0.65%	12.30%	1.19%
DXC	373K	0.04%	-30.47%	-0.98%	0.06%	1.05%	42.08%	2.81%
EMR	373K	0.04%	-18.96%	-0.75%	0.06%	0.86%	16.33%	1.68%
EQT	373K	0.05%	-18.66%	-1.33%	-0.02%	1.43%	37.32%	2.75%
FDX	373K	0.05%	-21.40%	-0.84%	0.06%	0.95%	15.53%	1.88%


Precio de las Acciones
Acciones
AAPL
AXP
AZO
CPB
CSCO
CVX
DUK
DXC
EMR
EQT
FDX
GIS
Log del Precio
Fecha
2010
2012
2014
2016
2018
2020
2022
2024


Retorno Trimestral
Acciones
AAPL
AXP
AZO
CPB
CSCO
CVX
DUK
DXC
EMR
EQT
FDX
GIS
Retorno
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100
2010
2012
2014
2016
2018
2020
2022
2024
Fecha

Ortogonalización

- Hay correlación entre los retornos de diferentes portafolios en un momento de tiempo específico.
- Se corrige mediante un proceso de ortogonalización por PCA.
- Partimos de K portafolios iniciales correlacionados entre sí y obtenemos $K' < K$ portafolios independientes (Ozkaya y Wang, 2020).



Algorithm 3 Ortogonalización

Input: $K \geq 2$ portafolios iniciales (armas), T número de períodos, $T_1 < T$ ventana de tiempo rodante (sliding window), Matriz de retornos de tamaño $T \times K$

for $t = 1, 2, \ldots, T$ **do**
  if $t \leq T$ **then**
    Calcular la matriz de covarianza Σ_t como $\Sigma_{t-T_1,t}$
    Aplicar PCA a Σ_t y Normalizar los *eigenvectores*
    Calcular $k_t = \{\text{median}(\lambda_t) = I_{\lambda_{t,i}}\} \cdot i$, donde $\lambda_{t,i}$ se refiere a la i-ésima componente de la matriz de valores propios ordenada de manera descendente.
    Obtener la t-ésima fila de la matriz de portafolios ortogonales.
  end if
end for
Tomar $K' = min\{k_t\}$.
Para cada fila obtenida en $t = T_1, .., T$, seleccionar las K' primeras columnas y agregar las filas verticalmente.
Output: Matriz de retornos de tamaño $(T - T_1) \times K'$

ADTS

Algorithm 4 Thompson Sampling con Descuento Adaptativo (ADTS)

Input: $K' \geq 2$ armas, $\gamma \in (0, 1]$ factor de descuento, $w \in \{1, .., T\}$ ventana deslizante

for $t = 1, \ldots, T$ **do**
  for $k = 1, \ldots, K'$ **do**
    Obtener las últimas w recompensas para el brazo k
    Definir α_k^w y β_k^w como el número de éxitos y fracasos pasados, respectivamente.
    $\theta_k(t) = \mathcal{B}(\alpha_k, \beta_k)$, $\tilde{\theta}_k(t) = \tilde{\mathcal{B}}(\alpha_k^w, \beta_k^w) \leftarrow$ doble sampleo.
  end for
  $I(t) = \arg\max_k \frac{\theta_k(t) + \tilde{\theta}_k(t)}{2}$, $r_{kt} \leftarrow$ observar
  for $k = 1, \ldots, K'$ **do**
    $\alpha_k \leftarrow \gamma\alpha_k + r_{kt}$, $\beta_k \leftarrow \gamma\beta_k + (1 - r_{kt})$
  end for
end for
Output: Arrays de tamaño T con la política tomada, los retornos observados, dummies (éxito o fracaso).

Combinatorial Adaptive Thompson Sampling (CADTS)

Algorithm 5 Combinatorial Adaptive Discounted Thompson Sampling

K Número de armas, $s \in [0, 1]$ Valor del paso de pesos posibles
$\gamma \in (0, 1]$ Factor de descuento, $w \in \{1, .., T\}$ Tamaño de la ventana deslizante

Generar los superarmas factibles del portafolio (S dados K y s):

1: **for** $k = 1,...,K$ **do**
2: $\quad w_k \leftarrow [0, s, 2s, \ldots, 1]$
3: **end for**
4: $S \leftarrow \left\{\text{Combinaciones}(w_k, i) \mid \sum_{k=1}^{K} w_k = 1\right\}$
5: **Ejecutar ADTS sobre el conjunto de superarmas S (dados γ y w)**
6: **Output:** Arrays de tamaño T con la política tomada, los retornos observados, dummies (éxito o fracaso).

Parámetros Estrategias

Table 1: Parámetros Estrategias

Parámetro	**Ortogonalización**	**Combinatorial**	**Markowitz**
Acciones	12	12	12
K	4	78	-
T_1	200	-	200
γ	0.8	0.8	-
w	121	121	-
s	-	0.5	-
T	3530	3530	3530
Simulaciones	100	30	1

Comparación número de éxitos promedio y Métricas de Desempeño (2010-2024)

Table 2: Comparación número de éxitos promedio

Estrategia	# de Éxitos Promedio	# de Intentos	Éxitos Esperados Aleatorizando
Combinatorial	544.43	3530	45.26
Ortogonalización	1212.9	3530	882.75

Table 3: Métricas de desempeño (2010-2024)

Estrategia	Sharpe Ratio	Max Drawdown	Retorno Anualizado
Combinatorial	0.43	-75.31%	6.58%
Ortogonalización	0.64	-43.57%	7.31%
Markowitz	0.71	-36.02%	6.73%

Serie de Tiempo Valor Esperado Riqueza

Comentarios finales

- **Desempeño destacado en optimización de portafolios:** Las metodologías Ortogonalización con ADTS y Combinatorial con ADTS superan a modelos tradicionales como Markowitz bajo ciertas condiciones (hipótesis de componente de corto y largo plazo).
- **Definición del reward es clave:** Combinatorial presenta mayor desempeño al final, pero es muy riesgoso y su eficacia se ve limitada por ventanas pequeñas (por como definimos el retorno).

Comentarios finales

- **Potencial y perspectivas futuras:** El trabajo demuestra el potencial de los métodos y sugiere mejoras como incorporar nociones explícitas de riesgo o ajustar parámetros clave, como el factor de descuento y el número de períodos.
- Código disponible: Link repositorio en GitHub

Referencias

- * Ozkaya, G., & Wang, Y. (2023). Multi-Armed Bandit Approach to Portfolio Choice Problem. *Barcelona Graduate School of Economics*.
- * Fonseca, G., Silva, L., & Castro, P. (2024). Improving Portfolio Optimization Results with Bandit Networks.
- Fama, E. F., & French, K. R. (1993). Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, 33(1), 3–56.
- Chapelle, O., & Li, L. (2011). An Empirical Evaluation of Thompson Sampling. Recuperado de:
- Vermorel, J., & Mohri, M. (2005). Multi-armed Bandit Algorithms and Empirical Evaluation.
- Zhou, Y., & Zhu, J. (2015). Contextual Thompson Sampling for Combinatorial Cascading Bandits.
- Seldin, Y., Slivkins, A., & Tewari, A. (2014). Contextual Bandits with Similarity Information.
- Ding, Y., & Liang, F. (2021). Adaptive Thompson Sampling with Automatic Exploration and Exploitation Balance.
- Cao, L., Sun, R., Ma, T., & Liu, C. (2024). On Asymmetric Correlations and Their Applications in Financial Markets. *Journal of Financial Analysis*, 10(3), 45–65. Recuperado de: `https://doi.org/10.3390/jrfm16030187`
- Graczyk, M. B., & Queirós, S. M. D. (2024). Intraday Seasonalities and Nonstationarity of Trading Volume in Financial Markets: Individual and Cross-Sectional Features. *Journal of Financial Markets*, 12(4), 123–140. Recuperado de: `https://doi.org/10.1371/journal.pone.0165057`
- Curtó, J., de Zarzà, I., Roig, G., Cano, J. C., Manzoni, P., & Calafate, C. T. (2024). LLM-Informed Multi-Armed Bandit Strategies for Non-Stationary Environments. *Journal of Machine Learning Research*, 18(3), 45–67. Recuperado de: `https://doi.org/10.3390/electronics12132814`