# CEE5735/6735 Midterm: Grey Box Option Pricing in Financial Markets

Gabriele Manganelli

**Abstract**

The quantitative structure of modern finance has evolved through the study of stochastic processes and stochastic calculus. Pioneering work by Black and Scholes (1973) [1], later extended by Merton (1974) [2], introduced a framework to determine the fair price of a financial contract known as a call option. A call option gives the holder the right, but not the obligation, to buy a financial asset at a predetermined price on a set future date. This agreement, where the right can be exercised only at maturity, is known as a *European call option.*While the Black-Scholes model remains foundational in financial theory, its assumptions—such as constant volatility and frictionless markets—limit its ability to capture real-world dynamics. In this project, we explore how to address these limitations in the case of perfect knowledge of future stock prices by using Machine Learning (ML) models, specifically Gaussian Process Regression (GPR), to model the residuals between Black-Scholes pricing and actual market prices. The methodology involves splitting the option price time series at a threshold time to maturity $T_{\text{split}} = 0.1$. Residuals are calibrated on the past segment ($T > T_{\text{split}}$), where consistent discrepancies reveal systematic underestimation of market prices by the Black-Scholes model. Residuals predicted by the ML model are then added to Black-Scholes prices for the future segment ($T \leq T_{\text{split}}$), significantly improving option price accuracy for the Apple options analyzed. This approach demonstrates the potential of combining mechanistic pricing models with data-driven corrections.

# 1 Introduction

**System of Interest.** We are interested in the financial market, specifically the pricing of call options. An option is a financial derivative that gives the buyer the right, but not the obligation, to buy (call) an asset at a predetermined price at expiration. The classic Black-Scholes (BS) model provides a mechanistic description of this process based on continuous stochastic calculus, involving assumptions about market behavior such as constant volatility and no arbitrage opportunities.

**Aspect to be Modeled.** While BS provides a foundational mechanistic approach to pricing European call options, its simplifying assumptions—such as constant volatility, no transaction costs, and log-normal stock price distributions—fall short in capturing real-world complexities. This gap results in systematic pricing errors, especially as the time to maturity $T \to 0$, where market dynamics often deviate significantly from BS assumptions. For simplicity, we treat all options as European and disregard potential changes introduced by American-style exercise rights, as their impact on our systematic correction analysis is negligible.

**Modeling Approach: Grey Box.** Building on Kim's use of Gaussian Process Regression (GPR) for American options [3], we apply a grey-box approach that combines the mechanistic Black-Scholes model (clear-box) with GPR (black-box). For Apple options, this method captures and corrects a systematic underpricing in Black-Scholes predictions, as evidenced by consistently positive residuals. Assuming perfect knowledge of the underlying stock price, the corrections bring the model closer to observed market prices. A real-world application, such as in trading, involves forecasting future stock prices, with option prices calculated based on these predictions. However, the underlying stock price model is crucial. Geometric Brownian Motion, for example, has limitations. It fails to account for jumps in stock prices, mean reversion, and stochastic volatility (e.g., the Heston model), which may lead to unsatisfying results.

# 2 Mathematical Formulation

## 2.1 Clear Box Component (Black-Scholes)

The Black-Scholes model provides an explicit formula for pricing European call options. An option grants the holder the right—but not the obligation—to buy (call) an asset, such as a stock, at a predetermined *strike price* ($K$) on a specified future date. The *time to maturity* ($T$) refers to the time remaining, typically expressed in years or days, between today and the option's expiration date. The formula is derived using stochastic calculus under assumptions such as constant *volatility* ($\sigma$)—defined as the annualized standard deviation of the underlying asset price returns—and no transaction costs.

The call option price, $C_{\text{BS}}$, is expressed as:

$$C_{\text{BS}} = SN(d_1) - Ke^{-rT}N(d_2), \tag{1}$$

where $S$ is the current price of the underlying asset, $r$ is the risk-free interest rate, and $N(\cdot)$ is the cumulative distribution function of the standard normal distribution.

The terms $d_1$ and $d_2$ are given by:

$$d_1 = \frac{\ln\left(\frac{S}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}}, \tag{2}$$

$$d_2 = d_1 - \sigma\sqrt{T}. \tag{3}$$

The model's reliance on simplifying assumptions, such as constant $\sigma$, limits its accuracy in real-world markets, especially near expiration ($T \to 0$), where pricing dynamics often deviate significantly.

## 2.2 Black Box Component (Gaussian Process Regression)

To model the discrepancies between Black-Scholes predictions and actual market data, we employ **Gaussian Process Regression (GPR)**. The residuals between market prices and Black-Scholes predictions are defined as:

$$\text{Residual} = C_{\text{market}} - C_{\text{BS}}, \tag{4}$$

where $C_{\text{market}}$ represents the market option price, and $C_{\text{BS}}$ is the Black-Scholes prediction. These residuals are modeled as:

$$\text{Residual} = f(\mathbf{x}) + \varepsilon, \tag{5}$$

where:

- $\mathbf{x}$ includes input features such as *moneyness* $S/K$ (stock-to-strike ratio) and *time to maturity* $T$,
- $f(\mathbf{x})$ is the latent function governed by a Gaussian Process (GP), and
- $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$ is Gaussian noise.

**Covariance Kernels.** The covariance structure of the GP is defined by a composite kernel, $k(\mathbf{x}, \mathbf{x}')$, which determines the correlation between residuals at two input points $\mathbf{x}$ and $\mathbf{x}'$. Two combinations of kernels were considered:

1. **Composite Kernel 1: RBF + Matérn ($\nu = 0.5$) + White Noise**
   The composite kernel is expressed as:

$$k_{\text{Composite-1}}(\mathbf{x}, \mathbf{x}') = C \cdot [k_{\text{RBF}}(\mathbf{x}, \mathbf{x}') + k_{\text{Matern-0.5}}(\mathbf{x}, \mathbf{x}')] + k_{\text{White}}(\mathbf{x}, \mathbf{x}'), \tag{6}$$

where:

$$k_{\text{RBF}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\ell^2}\right), \tag{7}$$

$$k_{\text{Matern-0.5}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|}{\ell}\right), \tag{8}$$

$$k_{\text{White}}(\mathbf{x}, \mathbf{x}') = \sigma_n^2 \delta(\mathbf{x}, \mathbf{x}'), \tag{9}$$

and $\delta(\mathbf{x}, \mathbf{x}')$ is the Dirac delta function. [1]

[1] The **Matérn kernel** is a generalization of the RBF kernel, characterized by an additional smoothness parameter $\nu$. The general form of the Matérn kernel is:

$$k(\mathbf{x}, \mathbf{x}') = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{\sqrt{2\nu}\|\mathbf{x} - \mathbf{x}'\|}{\ell}\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}\|\mathbf{x} - \mathbf{x}'\|}{\ell}\right),$$

where $\Gamma(\cdot)$ is the gamma function, $K_\nu(\cdot)$ is the modified Bessel function of the second kind, $\ell$ is the length scale, and $\|\mathbf{x} - \mathbf{x}'\|$ is the Euclidean distance. For $\nu = 0.5$, the kernel simplifies to the exponential form shown above.

2. **Composite Kernel 2: Dot Product + RBF + White Noise**
   The composite kernel is expressed as:

$$k_{\text{Composite-2}}(\mathbf{x}, \mathbf{x}') = C \cdot [k_{\text{DotProduct}}(\mathbf{x}, \mathbf{x}') + k_{\text{RBF}}(\mathbf{x}, \mathbf{x}')] + k_{\text{White}}(\mathbf{x}, \mathbf{x}'), \tag{10}$$

   where:

$$k_{\text{DotProduct}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \left(\mathbf{x} \cdot \mathbf{x}' + \sigma_0^2\right). \tag{11}$$

   The parameters in the composite kernels are defined as follows:

   - $\sigma_f^2$: Signal variance, representing the scale of variations in the function $f(\mathbf{x})$,
   - $\ell$: Length scale, controlling the smoothness of the kernel,
   - $\sigma_n^2$: Noise variance, modeling the observation noise,
   - $C$: Overall scaling factor of the composite kernel,
   - $\sigma_0^2$: Offset parameter in the Dot Product kernel.
   - $\|\mathbf{x} - \mathbf{x}'\|$: Euclidean distance between input vectors $\mathbf{x}$ and $\mathbf{x}'$.

**Kernel Selection.** Each composite kernel was evaluated based on the negative Mean Squared Error (MSE) on the validation set. The kernel achieving the lowest validation MSE was selected for each option, with the learned hyperparameters detailed in the results section, Table **??**.

# 3 Dataset and Preprocessing

The dataset was sourced from the Wharton Research Data Services (WRDS) platform, covering options data for Apple (AAPL) stock from August 31, 2022, to August 31, 2023. We selected **three options** with the highest trading volume during this period to ensure better liquidity, reduced bid-ask spreads, and minimized price distortions caused by individual irrational trades.

The selected options were:

1. **AAPL 230616C185000**: Call option with a strike price of \$185, expiring on June 16, 2023.
2. **AAPL 230721C195000**: Call option with a strike price of \$195, expiring on July 21, 2023.
3. **AAPL 230721C200000**: Call option with a strike price of \$200, expiring on July 21, 2023.

Each option identifier follows the format **[Ticker] YYMMDD [Option Type] [Strike Price × 1000]**. For example, in **AAPL 230616C185000**, the string **230616** denotes the expiration date (June 16, 2023), **C** indicates a call option, and **185000** corresponds to a strike price of \$185, scaled by a factor of 1000. This scaling was removed during preprocessing to ensure consistency with the numerical representation of the stock price.

## 3.1 Preprocessing Steps

To prepare the data for modeling, the following steps were applied:

- **Mid-Market Price**: Option prices were calculated as the average of the bid and ask prices:

$$C_{\text{market}} = \frac{\text{Bid} + \text{Ask}}{2}.$$

- **Derived Features**: Key features were calculated to align with the Black-Scholes model:
  - *Time to Maturity (T)*: Expressed in years, calculated as the difference between the current date and the option's expiration date.
  - *Moneyness*: Defined as the ratio of the stock price ($S$) (underlying of the option) to the strike price ($K$):

$$\text{Moneyness} = \frac{S}{K}.$$

- **Black-Scholes Predictions**: Option prices were estimated using the Black-Scholes model 1. Residuals were then calculated as:

$$\text{Residual} = C_{\text{BS}} - C_{\text{market}}.$$

To ensure data quality:

- Options with $T > 0$ and valid implied volatilities ($\sigma > 0$) were retained.
- Data with missing or invalid values in key fields, such as stock prices, bid-ask spreads, and implied volatilities, were removed.

### 3.1.1 Residual Segmentation

Residuals were segmented into two distinct groups based on time to maturity ($T$):

- **Training Residuals ($T > 0.1$)**: Used for training and calibration of the Gaussian Process Regression (GPR) model.
- **Testing Residuals ($T \leq 0.1$)**: Reserved exclusively for testing the model's performance on near-expiration predictions.

This segmentation ensured a clear temporal separation between the data used for model training and the data reserved for evaluation. Importantly, no future information leaks into the training process. Figure 1 provides an overview of the residual behavior for one of the selected options, highlighting the discrepancies between Black-Scholes predictions and market prices.

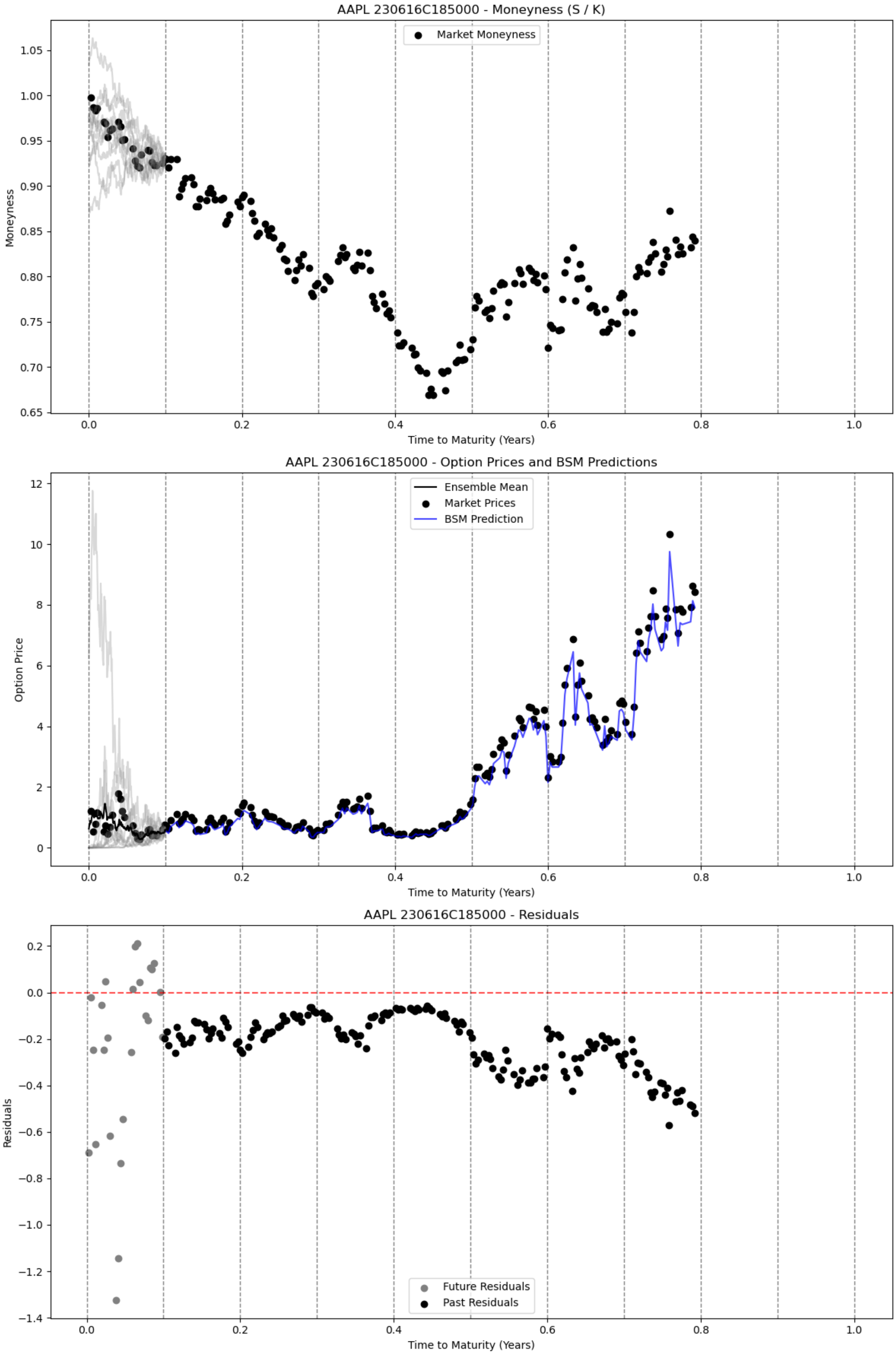


Figure 1: Visual representation of moneyness, option prices, and residuals for the AAPL 230616C185000 option. In the region $T \in [0, 0.1]$, stock prices were simulated using GBM with the earliest available values for the stock price $S_0$, implied volatility $\sigma$, and risk-free rate $r$. Option prices (grey shaded lines) were calculated using the Black-Scholes model on these simulated paths. The black line is the ensemble mean of these option prices, while grey residuals show the deviation between this mean and market prices. Black points represent the real market data, and for $T > 0.1$, the blue line shows Black-Scholes predictions using real stock prices and implied volatilities. These simulations highlight the unreliability of direct forecasting in $T \in [0, 0.1]$ due to erratic behavior of option prices near its expiration date, and GBM limitations. Only systematic corrections for $T > 0.1$ are analyzed further at first.

### 3.2 Gaussian Process Regression: Training, Validation, and Results

To correct residuals between market option prices and Black-Scholes predictions, we employed Gaussian Process Regression (GPR). The following methodology was adopted to ensure robust model selection while preserving the temporal structure of the data.

**Training and Validation Procedure.** Residuals were defined as:

$$\text{Residual} = C_{\text{market}} - C_{\text{BS}}, \tag{12}$$

where $C_{\text{market}}$ represents market option prices, and $C_{\text{BS}}$ are Black-Scholes predictions. The residuals were extracted for $T > 0.1$ and split into distinct training and validation sets to preserve the temporal order:

- **Training Set ($T > 0.3$)**: Residuals far from expiry were used for model training.
- **Validation Set ($0.1 < T \leq 0.3$)**: Residuals closer to expiry were held out to evaluate model performance and ensure generalization.

The split reflects the natural temporal ordering of option pricing data, where information closer to expiry ($T \to 0$) is more volatile and distinct from far-expiry data. The model was never exposed to the validation set during training.

The input features for the GPR model were:

- *Moneyness* ($S/K$): Ratio of stock price to strike price.
- *Time to Maturity* ($T$): Time remaining until expiry.

**Gaussian Process Framework.** The Gaussian Process models the residuals as:

$$\text{Residual} = f(\mathbf{x}) + \varepsilon, \tag{13}$$

where $\mathbf{x} = [S/K, T]$, $f(\mathbf{x})$ is a latent function with a zero-mean Gaussian prior, and $\varepsilon$ is Gaussian noise with variance $\sigma_n^2$.

The covariance function $k(\mathbf{x}, \mathbf{x}')$ is a composite kernel selected from two candidates:

1. **Composite Kernel 1: RBF + Matérn ($\nu = 0.5$) + White Noise**,
2. **Composite Kernel 2: Dot Product + RBF + White Noise**.

The parameters of each kernel, such as signal variance ($\sigma_f^2$), length scale ($\ell$), noise variance ($\sigma_n^2$), and scaling factor ($C$), were optimized to maximize the negative Mean Squared Error (MSE) on the validation set (therefore equivalent to "minimizing the loss function").

**Model Selection Metric.** The negative Mean Squared Error (MSE) served as the loss function during hyperparameter optimization:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{14}$$

where $y_i$ are the actual residuals and $\hat{y}_i$ are the GPR predictions. The MSE was evaluated using a temporal split with a fixed training and validation partition. The model achieving the lowest validation MSE was selected for each option.

The coefficient of determination ($R^2$) was computed as a diagnostic measure to assess the model fit, but it was not used in the selection process.

**Kernel Parameters.** The table below summarizes the optimal hyperparameters for each option. Parameters are displayed as rows for clarity and compactness.

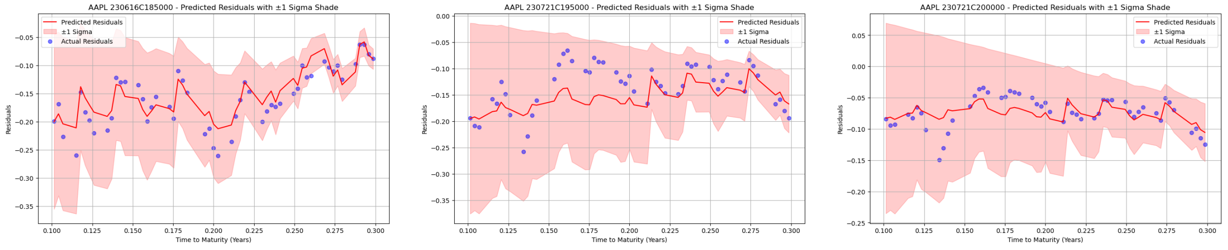
Table 1: Gaussian Process Hyperparameters for Selected Options

| | AAPL Options | | |
|---|---|---|---|
| **Parameter** | **230616C185000** | **230721C195000** | **230721C200000** |
| Best Kernel | RBF + Matérn ($\nu = 0.5$) | Dot Product + RBF | Dot Product + RBF |
| Alpha | $1.07 \times 10^{-6}$ | $1.09 \times 10^{-2}$ | $1.09 \times 10^{-2}$ |
| Signal Variance | 0.356 | 0.107 | 0.106 |
| RBF Length Scale | 3.14 | 1.29 | 1.78 |
| Matérn Length Scale | 54.04 | - | - |
| Noise Variance | $6.50 \times 10^{-5}$ | $1.0 \times 10^{-6}$ | $1.0 \times 10^{-6}$ |
| DotProduct ($\sigma_0$) | - | 1.53 | 1.20 |

**Residual Behavior Across Options.** The GPR model fits were evaluated on the validation set and exhibited varying degrees of success across the three options:

- **AAPL 230616C185000**: The model captured residual trends effectively with minimal error (Validation MSE = $4.36 \times 10^{-4}$, $R^2 = 0.82$).
- **AAPL 230721C195000**: The model showed a weaker fit (Validation MSE = $1.31 \times 10^{-3}$, $R^2 = 0.28$).
- **AAPL 230721C200000**: A moderate fit was achieved (Validation MSE = $3.51 \times 10^{-4}$, $R^2 = 0.44$).

**Observations.** For time to maturity close to $T = 0.3$ , the GPR predictions across all options exhibit similar oscillatory patterns, particularly in the rightmost section of the plots. As $T$ decreases (moving left), the residuals display greater variability, and the predictions start to diverge (see Fig.2). While the GPR captures the general structure of the residuals, it struggles to follow some of the finer peaks and troughs, particularly in the second and third options, indicating a slight underfitting effect.



(a) AAPL 230616C185000

(b) AAPL 230721C195000

(c) AAPL 230721C200000

Figure 2: **GPR Training Fits for Residuals.** The red line shows GPR predictions, while shaded areas reflect uncertainty ($\pm 1\sigma$).

The GPR approach demonstrated varying performance, and could benefit from further kernel refinement (use different signal variances per kernel type) or additional input features may improve model accuracy.

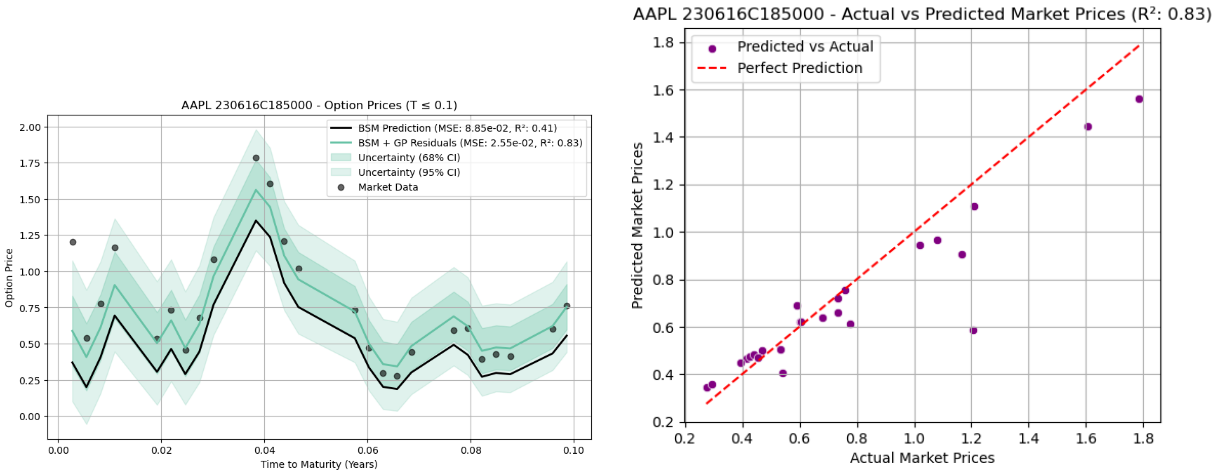
## 3.3 Corrected Option Prices

The GPR-predicted residuals were added to the Black-Scholes prices to obtain corrected option prices. Figures 3 and 4 illustrate the results for the three analyzed AAPL options.

The GPR corrections significantly improved alignment with market data, reducing residuals and tightening confidence intervals. Notably, the systematic underestimation present in the raw Black-Scholes predictions was effectively mitigated.

- For **AAPL 230616C185000** (see Fig. 3), the corrected prices closely track market values, achieving an $R^2$ of 0.83 compared to 0.41 for the baseline Black-Scholes model.
- For **AAPL 230721C195000** and **AAPL 230721C200000** (see Fig. 4), the improvements are similar, with $R^2$ increasing to 0.89 and 0.92, respectively.

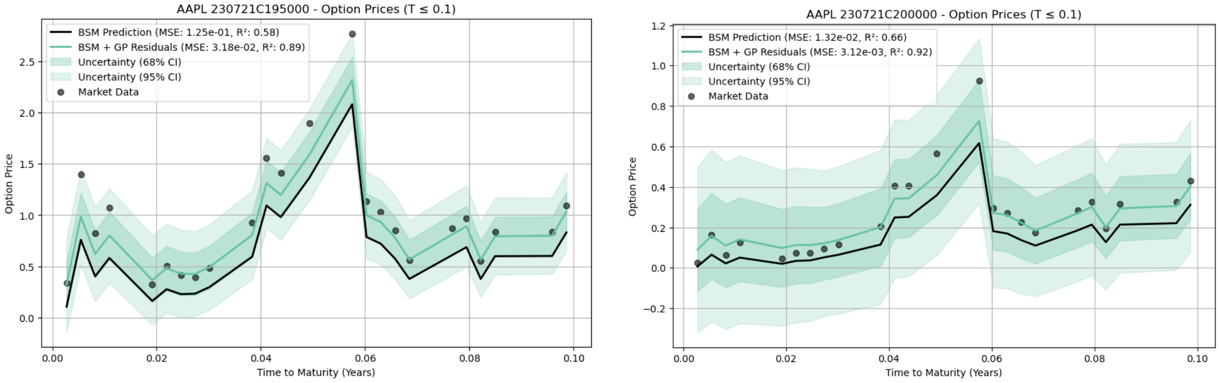
These results demonstrate that the GPR model effectively captures residual patterns across all three options, uniformly enhancing the Black-Scholes predictions.



(a) Corrected prices for AAPL 230616C185000.

(b) Correction residuals for AAPL 230616C185000.

Figure 3: **Corrected Option Prices and Residuals for AAPL 230616C185000.** (a) GPR corrections added to Black-Scholes prices. (b) Remaining residuals after correction.



(a) AAPL 230721C195000.

(b) AAPL 230721C200000.

Figure 4: **Corrected Option Prices for AAPL 230721C195000 and AAPL 230721C200000.** GPR corrections improve alignment with market data, with uncertainty estimates shown.

# 4 Conclusion

This work analyzed Apple options by correcting residuals between Black-Scholes (BS) predictions and market prices. The study tackled the static pricing problem using perfect knowledge of underlying stock prices, with daily updated volatility and interest rates. While this is not forecasting, as the residuals were added on top of a BS baseline option price that assumes perfect knowledge of stock prices, the same Gaussian Process Regression (GPR) calibration approach could be applied in a forecasting scenario. The method showed moderate success but struggled to capture finer patterns in the data. Future directions involve exploring more complex kernels and additional features for the GPR. An application to forecasting would involve refining the geometric Brownian motion (GBM) model, which underpins BS pricing, by incorporating jump-diffusion dynamics and the evolution of volatility (e.g. Heston model).

# A Appendix: Simulated Paths and Option Price Behavior

In this appendix, we provide a graphical analysis of simulated stock price paths and their corresponding Black-Scholes option prices for the three analyzed Apple options. These figures illustrate how option prices behave under the Geometric Brownian Motion (GBM) model in the short time-to-maturity region ($T \in [0, 0.1]$yrs) and serve as a visual reference for the challenges in option price forecasting.

## GBM Simulated Paths: Methodology

To model forecasting, Stock price paths for the region ($T \in [0, 0.1]$yrs) were simulated using the GBM model:

$$dS_t = rS_t dt + \sigma S_t dW_t,$$

where $S_t$ is the stock price, $r$ is the risk-free rate, $\sigma$ is the implied volatility, and $dW_t$ is a Wiener process. For each option, parameters $S_0$, $\sigma$, and $r$ were set to the values observed at $T = (0.1 - \frac{1}{252})$yrs. The Black-Scholes formula was then applied to compute option prices along these simulated paths, generating the shaded grey paths in "Option Prices and BSM Predictions" plots (Fig. 5, 6, 7). For comparison:

Residuals were computed as:

$$\text{Residual} = C_{\text{market}} - C_{\text{BS}},$$

where $C_{\text{BS}}$ represents the Black-Scholes price and $C_{\text{market}}$ is the observed market price. Residuals highlight discrepancies between market prices and the simulated mean option prices.

The black ensemble mean provides an illustration of what option price forecasts might look like under GBM. However, these paths were not used for forecasting in this study, as real stock prices were applied for baseline Black-Scholes predictions. Since predicting the behavior of option prices in the short-term region ($T \in [0, 0.1]$) can be challenging, future investigations could focus on more stable time regions, such as $T \in [0.4, 0.5]$, to begin testing forecasting strategies.

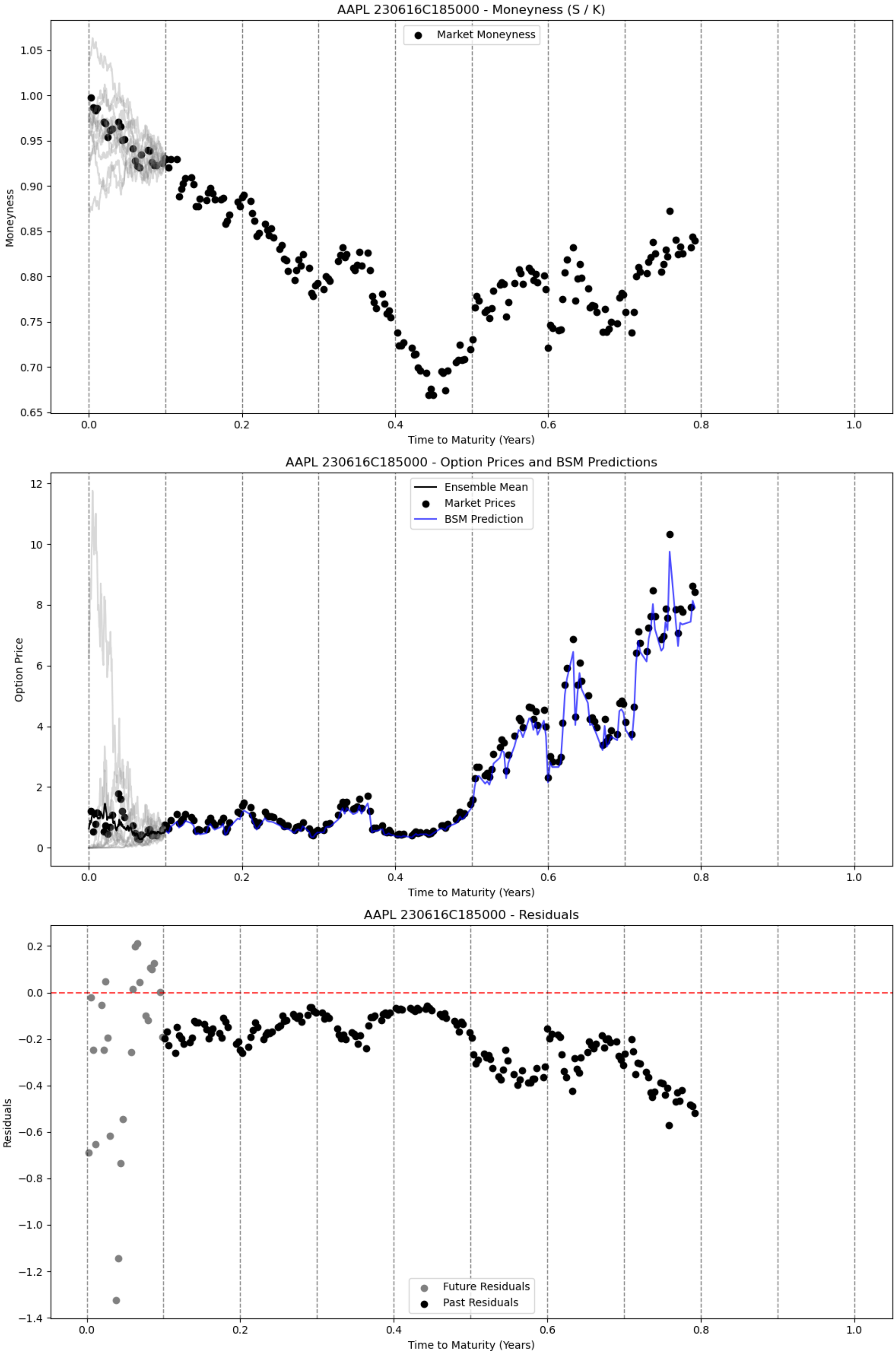


Figure 5: Visualization for option AAPL 230616C185000. The grey lines represent Black-Scholes option prices computed from GBM-simulated stock paths ($T \in [0, 0.1]$yrs). The black line shows the ensemble mean, and the blue line shows Black-Scholes predictions using real stock prices and implied volatilities for $T > 0.1$yrs. Residuals illustrate discrepancies between market and model option prices.

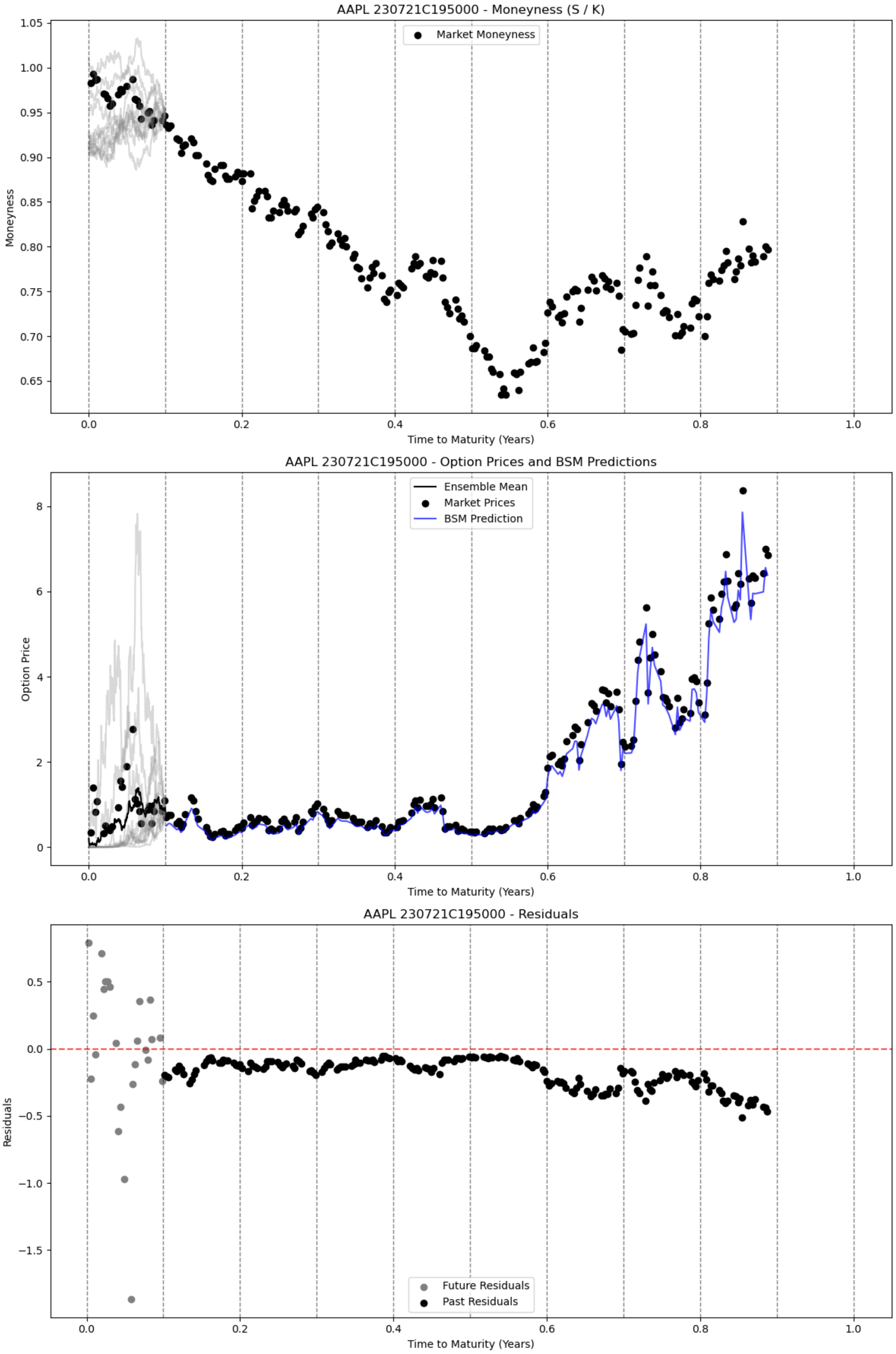


Figure 6: Visualization for option AAPL 230721C195000. GBM-simulated paths, Black-Scholes prices, and residuals highlight short-term deviations in option price behavior.

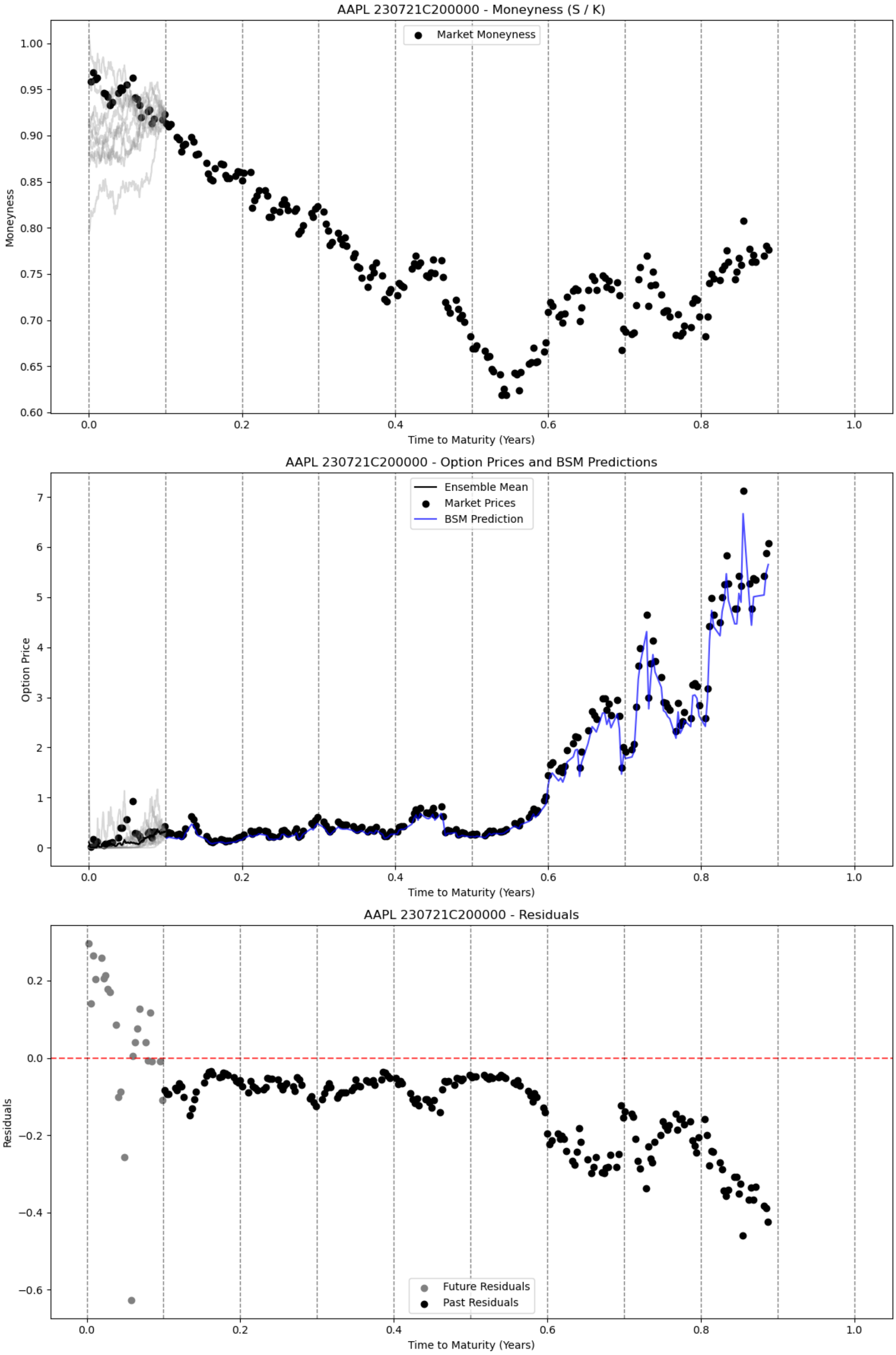


Figure 7: Visualization for option AAPL 230721C200000. Similar to the previous options, this figure illustrates GBM-based simulated prices and residual behavior.

These figures emphasize that while GBM provides a useful theoretical baseline for option pricing, its assumptions of constant volatility and smooth price evolution are inadequate for capturing short-term dynamics. Models incorporating jump diffusion processes or stochastic volatility (e.g. the Heston model) could offer more accurate representations of option price behavior near expiry.

# References

[1] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973.

[2] Robert C Merton. On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, 29(2):449–470, 1974.

[3] Chiwan Kim. Black-scholes grey box modeling for american options using gaussian process regression. *Finance Journal*, 45(4):123–135, 2020.