Nonparametric Machine Learning Methods for Equity Option Pricing

Bohan Jiang
Department of Computer Science
Western University
London, ON, Canada
E-mail: bjiang53@uwo.ca

***Abstract*—Despite the explosive growth in the market for financial options in the recent decades, theoretical parametric option pricing models, such as the Black-Scholes options pricing model ("BSOPM"), fail to accurately capture and price the dynamics influencing real-world equity options prices. As a result, we propose a nonparametric approach in pricing real-world options, in which we utilize Artificial Neural Networks ("ANN") and vast quantities of historical options data to accurately determine the relation of various variables with respect to option prices. Specifically, we contribute to the existing literature by developing and comparing nonparametric pricing models trained against various option types and exercise styles. We determine that the main difference in neural network architecture in pricing various option types revolve around the chosen activation functions. Furthermore, our evaluation results show that our trained ANNs perform significantly better than both the BSOPM and the Binomial Options Pricing Model ("BOPM") in pricing real-world American and European-style put options. Our findings lead to a future proposal in which ANNs are utilized to calibrate a theoretical options price to a real-world options price.**

Keywords—Artificial Neural Network (ANN), Option Pricing, Black-Scholes Options Pricing Model (BSOPM), Binomial Options Pricing Model (BOPM)

I. Introduction

Options are financial instruments that represents the right, but not the obligation to buy or sell an underlying security on (European-exercise style) or before (American-exercise style) an expiration date. There exists two types of options: call options and put options. A call option grants the option holder the right to buy the underlying asset at a fixed price K, called the strike price, if the underlying asset price S is above the strike price when the option is exercised. Conversely, a put option grants the option holder the right to sell the underlying asset at the strike price K, if and only if the underlying asset price S is below the strike price when the option is exercised. The payoff of both types of options when exercised can be characterized with the two following equations:

$$C = Max(0, S - K) \quad (1)$$

$$P = Max(0, K - S) \quad (2)$$

Where C represents the payoff of a call option, and P represents the payoff of a put option. Hence, the minimum value of an option at expiry is 0, and the profit to the option holder can be calculated as the difference between the payoff of the option at expiry and the premium paid to purchase the option. The profit and loss diagram of a typical call and put option to the option holder is shown in Figures 1 and 2.

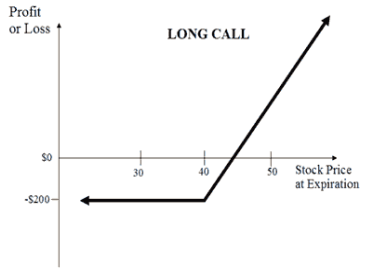

Fig. 1. Profit and loss diagram of the holder of a call option at expiry.

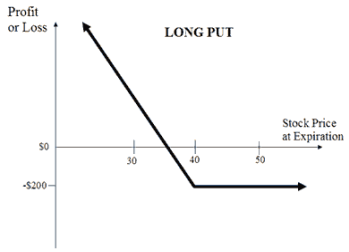

Fig. 2. Profit and loss diagram of the holder of a put option at expiry.

Financial options have experienced explosive growth in their popularity and usage. Indeed, in 2018, the Options Clearing Corporation reported that over 5.24 billion option contracts were traded in the US.[1] This growth in the market for options can be traced to the seminal work in the field of option pricing done by Black & Scholes (1973) [1] and Merton, (1973) [2], where they constructed a closed-form

[1] Data Retrieved from: https://www.theocc.com/Newsroom/Press-Releases/2020/OCC-Clears-Over-4-9-Billion-Contracts-in-2019

parametric expression for European option pricing, known as the Black-Scholes option pricing model ("BSOPM"). Using the version of the BSOPM formula which incorporates continuous dividends, then the premium of a European call option can be determined by the following equation:

$$C(S,K,\sigma,T,r,q) = S \cdot \Phi(d_1) - Ke^{-rT} \cdot \Phi(d_2) \quad (3)$$

$$d_1 = \frac{\ln\left(\frac{S}{K}\right)+\left(r-q+\frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}} \quad (4)$$

$$d_2 = \frac{\ln\left(\frac{S}{K}\right)+\left(r-q-\frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}} = d_1 - \sigma\sqrt{T} \quad (5)$$

In the above formula, $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution. Hence, the premium of a European call option can be determined by 6 input variables: S is the price of the underlying asset, K is the strike price of the option, σ is the annualized standard deviation of the stock's returns (also known as implied volatility), T is the time-to-maturity of the option, r is the annualized and continuously compounded risk-free interest rate, and q is the continuous dividend yield of the underlying asset. Similarly, the put option is priced by the Black-Scholes model [1] as follows:

$$P(S,K,\sigma,T,r,q) = Ke^{-rT} \cdot \Phi(-d_2) - S \cdot \Phi(-d_1) \quad (6)$$

Therefore, BSOPM describes a stationary nonlinear relationship between a theoretical option price and six input variables. However, there exists significant discrepancies between real-world option market prices and theoretical option prices. This pricing discrepancy is due to the assumptions inherent in the BSOPM. Specifically, the BSOPM assumes that stock prices are continuous and follow geometric Brownian motion, that both r and σ are constant over the life of the option, and that the market has no arbitrage opportunity (i.e. there is no way to make a riskless profit) and is frictionless (i.e. no transaction cost and no taxes) [2]. Indeed, an implied volatility surface (i.e. a 3D graph of implied volatility against strike and maturity) generated with options priced by the BSOPM is flat, whereas the typical shape of the implied volatility curve for real-world options on equity underlying is skewed (Figure 3). Generally, implied volatility is lowest for at-the-money ("ATM") options (i.e. options where the strike price K is around the stock price S), and higher for both in-the-money ("ITM") options (i.e. where the strike price K is lower (higher) than the stock price S for calls (puts)), and out-of-the-money ("OTM") options (i.e. where the strike price K is higher (lower) than the stock price S for calls (puts).

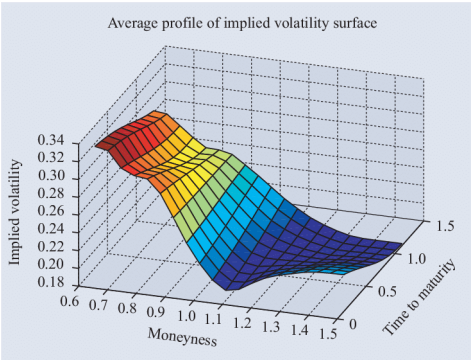

Fig. 3. Example of an implied volatility surface. Moneyness of an option is stock price divided by strike price (S/K).

Similarly, theoretical American-style option pricing most prominently utilizes the Binomial option pricing model ("BOPM") formalized by Cox, Ross, and Rubinstein in 1979 [3]. The BOPM is a numerical method which traces the evolution of the option's key underlying variables in discrete-time by the means of a binomial lattice (tree), for a given number of time steps between the valuation and expiration dates. Then, valuation of the option is performed iteratively, starting at each of the final nodes (i.e. potential values of the option at expiry), and then working backwards through the tree towards the first node (i.e. valuation date). A diagram illustrating the binomial lattice is shown below.

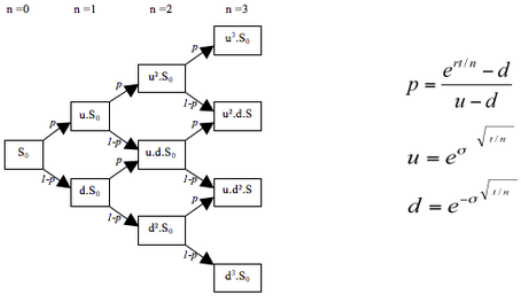

Fig. 4. Example of a binomial lattice used in the BOPM.

Despite being more computationally intensive than a closed-form equation such as the BSOPM, the advantage of such a valuation method relies in its ability to value an option at any given point in time, thus enabling the model to accurately value American options that may be exercised prior to expiry. Nevertheless, the BOPM is subject to the same stationarity assumptions as in the BSOPM, and the same discrepancies still exist between this model and real-world option prices.

With recent advancements in machine learning techniques, nonparametric option pricing methods have been proposed as an alternative pricing approach to traditional parametric pricing models such as the BSOPM or the BOPM. This method involves a data-driven approach in which historical data is utilized to determine the relations of various variables with respect to option prices. Various nonparametric

models have been developed using Deep Learning [4] and discovered to have greater pricing accuracy than theoretical pricing models on real-world option prices.[2] Nevertheless, the options explored in the literature remain largely focused on accurately pricing European call options generated by the BSOPM [6]. As a result, the scope of research in the literature is constrained in terms of the option type (i.e. calls and puts), option exercise style (i.e. European-style and American-style), as well as the availability of either computational resources or vast quantities of real-world option price data.

Hence, the goal of this project is to develop various nonparametric option pricing models across option types, option exercise styles, and across various types of options price data (i.e. theoretical prices generated from the BSOPM/BOPM as well as real-world options price data). This project will also seek to compare the accuracy of the developed nonparametric pricing models to the theoretical pricing models (i.e. BSOPM, BOPM) on real-world option prices. Finally, this research seeks to explore differences in model architectures and techniques used when pricing these different options, in order to improve the understanding of the various factors that produce an accurate option pricing model.

The novelty of this research involves constructing nonparametric models using extensive real-world equity underlying data across both option types and exercise-styles. In fact, the optionality for early exercise has greatly increased the difficulty in accurately pricing American options, even using theoretical option pricing models such as BOPM with stationarity assumptions. The construction of a pricing model that accurately prices real-world American options would further option pricing theory in understanding the specific pricing differences between European and American options.

Furthermore, this project will also focus on evaluating the accuracy of developed models against theoretical pricing models on real-world data model and comparing differences in model architecture across different option types.

This report will begin by evaluating and drawing upon the work done on nonparametric option pricing already produced in the related literature. This synthesis will serve as the basis on which we will build upon to further the research done in options pricing with nonparametric models.

II. Related Work

Due to the recent advances in machine learning techniques and widespread access to computational resources, nonparametric methods for estimating option prices have been proposed as a more computationally efficient alternative to traditional arbitrage-based pricing formulas such as the BSOPM [5]. Indeed, these methods are particularly useful when the underlying asset's price dynamics are unknown, or when the pricing equation associated with the no-arbitrage condition cannot be solved analytically.

A. A Nonparametric Approach to Pricing and Hedging Derivative Securities Via Learning Networks (1994)[5]

Hutchinson, Lo and Poggio generated a two-year training set of daily call option prices using the BSOPM, and estimated models utilizing four popular nonparametric methods: ordinary least squares (as a comparison benchmark), radial basis function networks ("RBFs"), multilayer perceptron networks ("MLPs"), and projection pursuit. The best performing models in fitting the Black-Scholes model, as determined by out-of-sample R^2 values, was the MLP model with a single hidden layer containing four units with a mean R^2 of 99.48, and the RBF model with 4 multiquadric centers and an output sigmoid with a mean R^2 of 99.95.

B. A new hybrid parametric and machine learning model with homogeneity hint for European-style index option pricing (2016) [8]

Das and Padhy introduce a homogeneity hint prior to the training of their nonparametric machine learning models by categorizing their options data based on the moneyness (i.e. in-the-money, at-the-money, out-of-the-money) and the time-to-maturity (i.e. short-term, long-term) of the option contract, in order to reduce their models' forecasting error on India's Bank NIFTY Index's European-style call options. The option prices were first computed using a parametric model (i.e. BSOPM/Monte-Carlo methods) on historical data, and the output price and historical data was then used to train the categorized nonparametric models to forecast the next day's option price. The nonparametric methods used were support vector regression ("SVR") and an Extreme Learning Machine ("ELM") with a single hidden layer, in order to overcome the slow learning speed of feed-forward neural networks. The results achieved imply that the SVR model was most accurate in predicting the next-day option price, as measured by an average root-mean-square error ("RMSE") of 0.0028. Moreover, this model has the most accurate predictions for short time-to-maturity options and at-the-money options, with poorest performance for long time-to-maturity and in-the-money options.

C. The Case of Deep Learning for Option Pricing (2017) [4]

Culkin and Das trained a fully-connected feed-forward deep learning neural network to reproduce the BSOPM by simulating 300,000 call option prices. The out-of-sample RMSE achieved was 0.0112, with an R^2 of 0.9982. The architecture of the deep learning model consisted of 4 hidden layers of 100 neurons each, with the activation functions at each hidden layer determined to be: LeakyReLU, ELU, ReLU, ELU, and an exponential output function for the neural net to be non-negative with certainty. A dropout rate of 25% was also used at each hidden layer to ameliorate overfitting issues. The training batch size was 64 with 10 total training epochs to fit a total of 31,101 coefficients for the deep learning model.

[2] See the work done by [4], [5], [6], [7] and [8].

D. *Deep Learning Calibration of Option Pricing Models (2019) [9]*

Itkin contends that ANN option pricing models for call option price C must be at least C^1 or better C^2, to allow for differentiability required in the derivation of option Greeks for industry application. This implies choosing non-linear yet differentiable activation functions for ANN construction and training, with the final activation function of the ANN involving a non-negative output. Itkin highlights the ELU activation function:

$$R(z) = \begin{cases} z, & z > 0 \\ \alpha(e^z - 1), & z \leq 0 \end{cases} \quad (7)$$

with hyperparameter $\alpha = 1$ as a potential choice, since the function has a zero-centered distribution with a smooth first derivative, as shown in Figure 5 below.

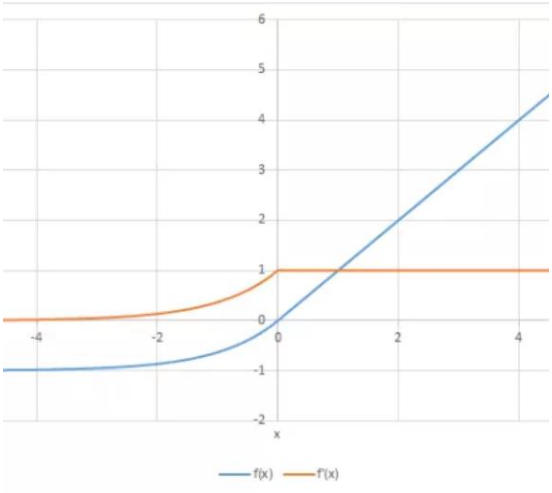

Fig. 5. ELU Activation Function (f(x)) and its first derivative (f'(x)).

However, the ELU has a second derivative that jumps at $z = 0$, and so a modified version of ELU was proposed ("MELU"):

$$R(z) = \begin{cases} \frac{0.5z^2 + az}{z+b}, & z > 0 \\ \alpha(e^z - 1), & z \leq 0 \end{cases} \quad (7)$$

with $a = 1 - 2\alpha$ and $b = -2 + \frac{1}{\alpha}$. Following this analysis, an ANN model was trained on 300,000 BSOPM-generated call option prices, where $0.001 \leq C \leq 10$. The feed-forward ANN had 128 nodes and 4 total layers, and was trained for 15 epochs with a batch size of 64 using the Adam optimizer. The activation functions for each layer were: LeakyReLU with $\alpha = 1$, MELU, MELU, and Softplus minus 0.5. Both MELU layers had $\alpha = 0.49$ as hyperparameters, and all activation functions used are C^2 to allow for first and second order option Greeks. The out-of-sample mean-squared-error ("MSE") of the model was 9.7E-04.

E. *Analysis and Research Gap*

The related literature has identified nonparametric option pricing models as good estimators of parametric option pricing models, and potentially more effective pricing models in situations where the no-arbitrage condition cannot be solved analytically or when the underlying asset's price dynamics are unknown [5]. The most accurate machine learning method in estimating option prices was the construction of an ANN and its variants (RBF, MLP in [5], ELM in [8], and ANN in both [4] and [9]). A similar methodology was applied across the literature, particularly in the usage of the parametric BSOPM model to generate and fit call option prices, and in the performance measurement method (R^2, MSE, RMSE). Major differences involved the architecture of the ANN and its associated hyperparameters, as well as the model training approach (e.g. homogeneity hint in [8]).

Consequently, the related literature is solely focused on European call options, and does not address or compare any differences in nonparametric model training across option type (call vs. put) and option exercise style (European vs. American). Moreover, the samples used are relatively small (ranging from 26K – 300K data points), and only [8] utilizes real-world historical options in the training process, which may differ significantly from the option prices estimated by parametric models. This project seeks to address this gap by constructing nonparametric models for both European and American-style options, generated by both the BSOPM and BOPM models respectively, as well as collected from real-world historical prices, with a large sample of over 20M data points per option style and type.

III. Research Objectives

This project has the following research objectives:

- *O1*: Generate European option prices using the BSOPM and develop various nonparametric models in order to infer the BSOPM, as a means to provide a performance benchmark in selecting the most accurate model architectures for pricing European options.
- *O2*: Generate American option prices using the BOPM and develop various nonparametric models in order to infer the BOPM, as a means to provide a performance benchmark in selecting the most accurate model architectures for pricing American options.
- *O3*: Utilize the subset of accurate model architectures to train and predict on real-world European equity option prices for varying option types, moneyness, and times-to-expiration, in order to accurately infer real-world European-style option prices.
- *O4*: Utilize the subset of accurate model architectures to train and predict on real-world American equity option prices for varying option types, moneyness, and times-to-expiration, in order to accurately infer real-world American-style option prices.

- *O5*: Determine if introducing a homogeneity hint on options data significantly improves nonparametric model out-of-sample accuracy by conducting performance comparisons between models trained with and without homogeneity hint both on generated and real-world options data.
- *O6*: Conduct architecture and performance comparisons between the best performing nonparametric models for call and put options, as well as for American and European style options. In particular, conduct analysis on pricing accuracy of existing parametric models (i.e. BSOPM, BOPM) versus their respective fitted nonparametric models on real-world option prices.
- *Significance*: The above objectives form a rigorous sequential process that seeks to determine and understand how best to construct and interpret an accurate nonparametric option-pricing model, by isolating factors such as model architecture, types of input data, option type, moneyness, and times-to-expiry. The achievement of each objective above will help clarify and improve the understanding of the various factors that produces an accurate option-pricing model.

IV. Research Methodology

A. Data Generation

In order to isolate for nonparametric model architecture in a deterministic manner, the BSOPM [2] and BOPM [3] are first implemented programmatically in order to generate over 20M prices for each style (European-style and American-style, respectively for BSOPM and BOPM) and type of option over a wide range of parameters (Figure 6).

Parameter	Range	Increment
Spot Price (S)	1	N/A
Strike Price (K)	\$0.01 - \$10	**\$0.01:** \$0.01 - \$0.10 and \$0.95 - \$1.05 **\$0.05:** \$0.10 - \$2.00 **\$1.00:** \$2.00 to \$10.00
Maturity (T)	0 years - 2.055 years	30 days
Dividend Yield (q)	0% - 14%	1%
Risk-free rate (r)	0% - 9%	1%
Volatility (σ)	0% - 800%	10%

Fig. 6. The range of parameters used to simulate 20,217,600 call and put option prices using the BSOPM and BOPM models. The "Increment" column represents the increment value within the bounds of each respective range.

Note that in the range of parameters described in Figure 6, we fix spot price at 1 and vary the strike price instead at various increments below and above the spot price. This effect produces a consistent range of moneyness values (S/K), and we explain its usefulness in the following section.

In order to incorporate dividends, the extended version of the BSOPM is implemented to incorporate a constant dividend yield. A similar implementation is done for the BOPM, and thus only American puts option prices are generated from the BOPM to avoid redundancy. Indeed, note that because the BOPM is implemented with a dividend yield input instead of with discrete dividends, then the American call price calculated by the BOPM would be exactly equivalent to the call price determined by the BSOPM. Consequently, only American put option prices generated by the BOPM would differ from those generated by the BSOPM.

B. Data Transformations and Preprocessing

In order to prepare the generated dataset for training, calls and puts are divided into two separate datasets in order to fulfill the objective of training both option types separately, and in order to introduce a homogeneity hint as per [8]. Furthermore, we utilize the fact that the BSOPM function is linear homogenous in (S, K):

$$\frac{V(S,K)}{K} = V(\frac{S}{K}, 1) \tag{8}$$

to normalize both the stock price S and the option price V by dividing both values by the strike price K. This is a standard transformation (performed in [3], [4], and [8]) which reduces the number of inputs into our nonparametric models to the five input variables $(\frac{S}{K}, T, q, r, \sigma)$, along with the output now involving the normalized option price $\frac{V}{K}$. A snapshot of the transformed dataset can be seen in Figure 7 below.

	maturity	risk-free	div_yield	implied_vol	moneyness	norm_option_price
0	1.643836	0.08	0.0	0.0	0.869565	0.000000
1	1.643836	0.08	0.0	0.1	0.869565	0.041120
2	1.643836	0.08	0.0	0.2	0.869565	0.085522
3	1.643836	0.08	0.0	0.3	0.869565	0.129591
4	1.643836	0.08	0.0	0.4	0.869565	0.173128

Fig. 7. Snapshot of the transformed BSOPM-generated Call Options dataset. Moneyness represents the normalized Stock price S/K, and norm_option_price represent the normalized option price V/K.

Finally, all generated datasets are prepared for training by being randomly divided into training, validation, and testing splits in proportions of 72%, 18%, and 10%, respectively.

C. Data Collection & Preprocessing

For real-world option prices, we collect daily historical options data from the OptionsMetrics database provided by the Wharton Research Data Services [10]. Over 42M daily European-style options prices were collected from equity indices (e.g. RUT, SPX) from 2008 – 2011, along with their respective underlying asset prices and dividend yields. Similarly, over 222M daily American-style option prices were collected on individual equities traded on various US exchanges in 2020, along with their respective underlying asset prices and discrete dividend payments.

In order to incorporate these discrete dividends into the option price, we change the underlying asset price by subtracting the present value of all dividend payments within the lifetime of the option from the current date's underlying asset price. This is shown in the below equation:

$$F_{0,T} = S_0 - \sum_{t=1}^{T} D_t e^{-rt} \tag{9}$$

The above equation shows that the forward stock price $F_{0,T}$ at time 0 is equal to the current stock price less the discounted value of all future dividends within the lifetime of the option. Consequently, the dividend yield for these options are therefore zero after this adjustment.

In order to determine the appropriate risk-free rate for each option, the interest rate of various maturities on the risk-free zero-coupon yield curve for each date was collected. A snapshot of the collected zero-coupon yield curve data is shown in Figure 8.

	date	days	rate
0	2012-09-04	7	0.184616
1	2012-09-04	14	0.206066
2	2012-09-04	15	0.207648
3	2012-09-04	43	0.326578
4	2012-09-04	78	0.348339

Fig. 8. Snapshot of the risk-free zero-coupon rates for 2012-09-04.

Subsequently, the corresponding risk-free rate was determined by first matching the date of the option to the zero-coupon yield-curve date. Then, the appropriate risk-free rate is determined by matching the time-to-maturity of the option to the risk-free yield curve via linear interpolation.

D. Exploratory Data Analysis

This section describes the various generated and collected options data used in the model training process, and provides descriptive statistics and distribution plots for each dataset.

1) BSOPM-genereated European-style options

Descriptive statistics for the 20M+ data points generated by the BSOPM for both calls and puts with the range of parameters given in Figure 6 are shown in the below Figure 9.

	strike_price	time_to_expiration	risk_free_rate	div_yield	implied_vol	BSOPM_price	moneyness	norm_option_price
count	2.021760e+07	2.021760e+07	2.021760e+07	2.021760e+07	2.021760e+07	2.021760e+07	2.021760e+07	2.021760e+07
mean	1.584375e+00	1.027397e+00	4.500000e-02	7.000000e-02	4.000000e+00	7.122040e-01	5.592229e+00	4.959052e+00
std	2.101809e+00	6.164384e-01	2.872281e-02	4.320494e-02	2.338090e+00	2.951614e-01	1.460555e+01	1.359011e+01
min	1.000000e-02	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e-01	0.000000e+00
25%	4.375000e-01	4.931507e-01	2.000000e-02	3.000000e-02	2.000000e+00	5.947420e-01	6.202652e-01	3.266245e-01
50%	1.005000e+00	1.027397e+00	4.500000e-02	7.000000e-02	4.000000e+00	8.402176e-01	9.950495e-01	7.468952e-01
75%	1.612500e+00	1.561644e+00	7.000000e-02	1.100000e-01	6.000000e+00	9.210949e-01	2.291667e+00	1.907816e+00
max	1.000000e+01	2.054795e+00	9.000000e-02	1.400000e-01	8.000000e+00	1.000000e+00	1.000000e+02	1.000000e+02

	strike	maturity	risk-free	div_yield	implied_vol	BSOPM_price	moneyness	norm_option_price
count	2.021760e+07	2.021760e+07	2.021760e+07	2.021760e+07	2.021760e+07	2.021760e+07	2.021760e+07	2.021760e+07
mean	1.584375e+00	1.027397e+00	4.500000e-02	7.000000e-02	4.000000e+00	1.293777e+00	5.592229e+00	6.990554e-01
std	2.101809e+00	6.164384e-01	2.872281e-02	4.320494e-02	2.338090e+00	1.952633e+00	1.460555e+01	3.138570e-01
min	1.000000e-02	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e-01	0.000000e+00
25%	4.375000e-01	4.931507e-01	2.000000e-02	3.000000e-02	2.000000e+00	1.611731e-01	6.202652e-01	5.161978e-01
50%	1.005000e+00	1.027397e+00	4.500000e-02	7.000000e-02	4.000000e+00	7.455374e-01	9.950495e-01	8.540716e-01
75%	1.612500e+00	1.561644e+00	7.000000e-02	1.100000e-01	6.000000e+00	1.299631e+00	2.291667e+00	9.323470e-01
max	1.000000e+01	2.054795e+00	9.000000e-02	1.400000e-01	8.000000e+00	1.000000e+01	1.000000e+02	1.000000e+00

Fig. 9. Descriptive Statistics for BSOPM-generated calls (top) and puts (bottom).

Note that in the above figure 9, BSOPM_price is the unnormalized price of the option as generated by the BSOPM, and is removed prior to the usage of the dataset. As expected, the statistics for all inputs to the model are the same, and the only difference lies within the actual option prices. Next, we observe the following distribution of features for the BSOPM-generated calls in Figure 10 and the BSOPM-generated puts in Figure 11.

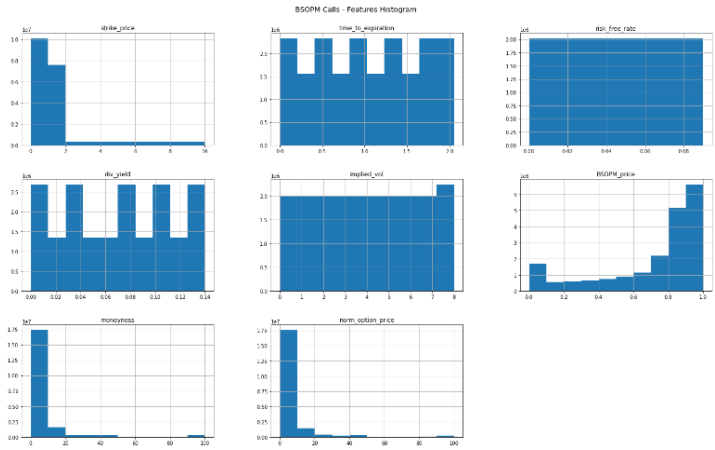

Fig. 10. BSOPM-generated calls Feature Distribution Histograms

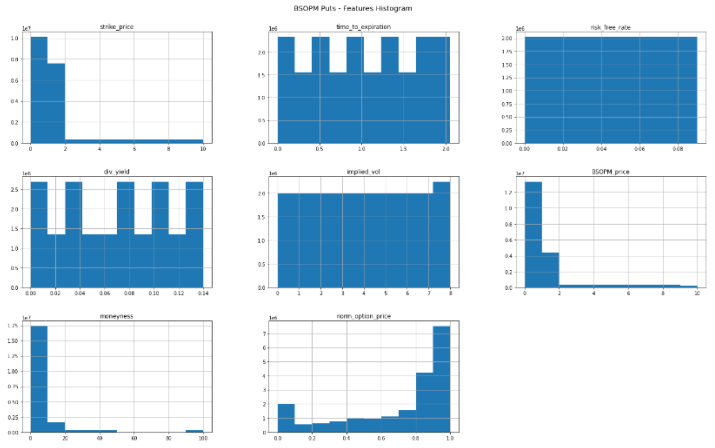

Fig. 11. BSOPM-generated puts Feature Distribution Histograms

As expected, the method used to generate these prices has caused the distribution of the risk-free-rate, dividend yield, implied volatility, and maturity of the option to be roughly uniform. In contrast, the strike price of the option has a high number of values between 0 and 1; due to the spot price being set to 1, the large number of strikes generated between 0 and 1 are used to simulate ATM options, as well as ITM calls and OTM puts. This in turn generates a moneyness distribution that resembles the highly positively skewed distribution of the strike price. We also note the expected differences in the distribution of the BSOPM_price and norm_option_price for call and put options. For the call options, the BSOPM_price grows close to 1 for deep ITM calls (since the minimum strike price is 0.01). Contrastingly, the BSOPM_price for the put options grows close to 10 for deep ITM puts (since the maximum strike price is close to 10). Hence, this causes a normalized option price range of 0 to 100 (e.g. deep ITM calls with strike 0.01) for calls, and a normalized option price range of 0 to 1 (since a put option cannot have a greater intrinsic value than its strike price) for puts.

Finally, we observe the dataset by plotting the moneyness (S/K) of the option versus its normalized option price (V/K), as shown in Figures 12 and 13.

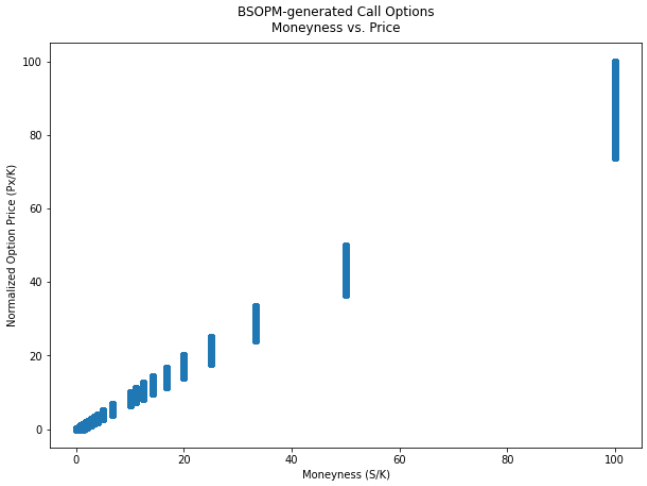

Fig. 12. Moneyness vs. normalized price for BSOPM-generated calls

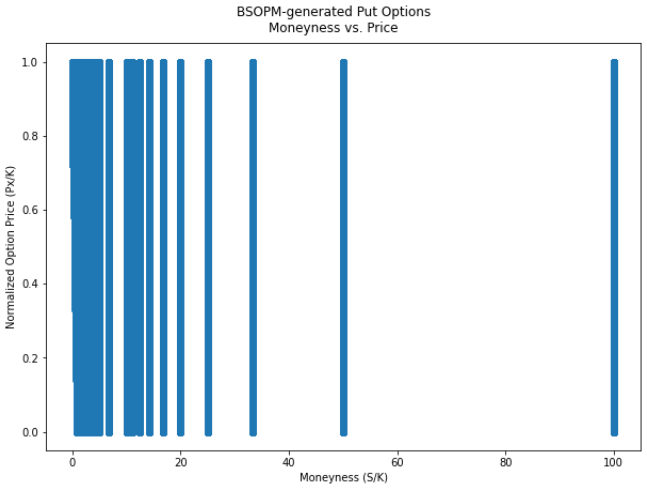

Fig. 13. Moneyness vs. normalized price for BSOPM-generated puts

It can be seen that the moneyness vs. price plot for call options ressembles the payoff profile of a call option. For ITM calls at increasing moneyness, a vertical range of option prices are generated due to changes in maturity, dividend yield, and implied volatility. For the put option plot, we see that for a given moneyness, the normalized option prices span the entire y-axis, ranging from 0 to 1. Although these put options may be OTM, the generated BSOPM option price can still be large relative to the strike price for varying amounts of implied volatility, dividend yield, and maturity.

2) BOPM-generated American-style puts

Descriptive statistics for the 20M+ data points generated by the BOPM model for theoretical American-style put prices with the range of parameters given in Figure 6 are shown in the below Figure 14.

	strike	maturity	risk-free	div_yield	implied_vol	bs_price	moneyness	norm_option_price
count	2.021760e+07	2.021760e+07	2.021760e+07	2.021760e+07	2.021760e+07	2.021760e+07	2.021760e+07	2.021760e+07
mean	1.584375e+00	1.027397e+00	4.500000e-02	7.000000e-02	4.000123e+00	1.407730e+00	5.592229e+00	7.630326e-01
std	2.101809e+00	6.164384e-01	2.872281e-02	4.320494e-02	2.337879e+00	2.126424e+00	1.460555e+01	3.506896e-01
min	1.000000e-02	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e-02	0.000000e+00	1.000000e-01	0.000000e+00
25%	4.375000e-01	4.931507e-01	2.000000e-02	3.000000e-02	2.000000e+00	1.733633e-01	6.202652e-01	5.495645e-01
50%	1.005000e+00	1.027397e+00	4.500000e-02	7.000000e-02	4.000000e+00	8.012147e-01	9.950495e-01	9.219458e-01
75%	1.612500e+00	1.561644e+00	7.000000e-02	1.100000e-01	6.000000e+00	1.417357e+00	2.291667e+00	1.008258e+00
max	1.000000e+01	2.054795e+00	9.000000e-02	1.400000e-01	8.000000e+00	1.333319e+01	1.000000e+02	1.333319e+00

Fig. 14. Descriptive Statistics for BOPM-generated American-style puts.

We note that the bs_price represents the unnormalized price of the option, and is removed prior to the usage of the dataset. Next, we observe the following distribution of features for the BOPM-generated dataset in Figure 15.

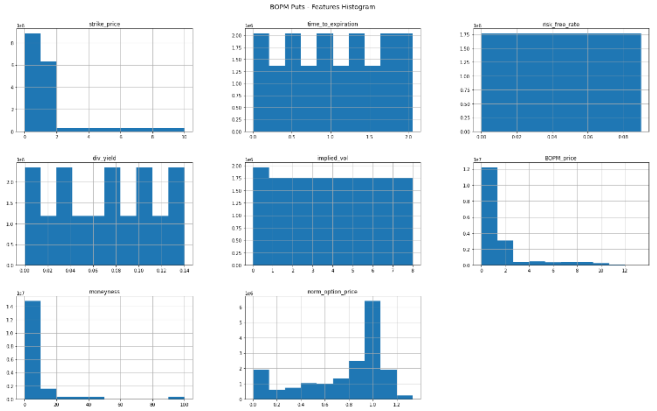

Fig. 15. BOPM-generated puts Feature Distribution Histograms

The feature distribution for these BOPM-generated American-style put options greatly resembles the distribution observed for the BSOPM-generated puts since the same range of parameters from Figure 6 were used for data generation. However, we should note the normalized option price distribution with a price range from 0 to 1.33. Although a put option cannot have intrinsic value greater than its strike price, this discrepancy can be explained by the fact that a non-zero dividend yield on the American-style option increases the premium of the option relative to its European-style counterpart, since dividend payouts usually cause a drop in the stock price by the dividend amount on the ex-dividend date. This early exercise advantage for American-style options is thus captured by the BOPM model. Finally, the Moneyness vs. Price plot for the BOPM-generated prices are plotted in Figure 16.

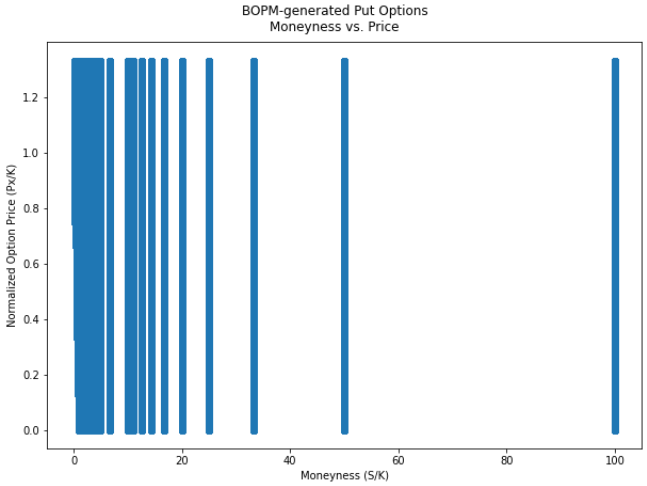

Fig. 16. Moneyness vs. normalized price for BOPM-generated puts

As expected, this moneyness vs. normalized price plot for BOPM-generated puts is nearly identical to the same plot in Figure 13 describing the BSOPM-generated puts.

3) Real-world European-style options

The descriptive statistics for the ~19.5M collected real-world European-style calls and ~19.3M real-world European -style puts can be seen in Figure 17s and 18.

	impl_volatility	ttm	rf_rate	div_yield	norm_option_px	moneyness
count	1.951907e+07	1.951907e+07	1.951907e+07	1.951907e+07	1.951907e+07	1.951907e+07
mean	3.190501e-01	2.757920e-01	1.143619e-02	2.272784e-02	1.595032e-01	1.109145e+00
std	2.811891e-01	3.730556e-01	9.285427e-03	1.913311e-02	3.837918e-01	4.155575e-01
min	1.001400e-02	2.739726e-03	0.000000e+00	0.000000e+00	1.562500e-06	6.600000e-02
25%	1.654630e-01	6.301370e-02	2.616870e-03	1.538829e-02	6.308725e-03	9.504457e-01
50%	2.399720e-01	1.424658e-01	8.411990e-03	1.878660e-02	6.177570e-02	1.040644e+00
75%	3.626900e-01	3.342466e-01	1.995690e-02	2.353788e-02	1.882500e-01	1.178644e+00
max	2.999984e+00	4.287671e+00	4.826287e-02	3.918256e-01	3.640900e+01	3.756070e+01

Fig. 17. Descriptive Statistics for real-world European-style calls.

	impl_volatility	ttm	rf_rate	div_yield	norm_option_px	moneyness
count	1.936927e+07	1.936927e+07	1.936927e+07	1.936927e+07	1.936927e+07	1.936927e+07
mean	3.314351e-01	2.686945e-01	1.127394e-02	2.077591e-02	3.894383e-02	1.183750e+00
std	2.739190e-01	3.733584e-01	9.211919e-03	1.546242e-02	6.751257e-02	6.021614e-01
min	1.002500e-02	2.739726e-03	0.000000e+00	0.000000e+00	2.155172e-06	1.531000e-01
25%	1.766780e-01	5.753425e-02	2.582730e-03	1.534495e-02	1.097222e-03	9.836855e-01
50%	2.567940e-01	1.342466e-01	8.129868e-03	1.859810e-02	1.052239e-02	1.077706e+00
75%	3.864190e-01	3.232877e-01	1.972026e-02	2.283843e-02	4.695013e-02	1.236244e+00
max	2.999987e+00	4.287671e+00	4.820561e-02	1.901557e-01	8.457500e-01	3.756070e+01

Fig. 18. Descriptive Statistics for real-world European-style puts.

It can be seen that the European-style put and call options statistics have parameters ranges that are greatly similar to their BSOPM-generated European-style counterparts. Subsequently, we observe the features distribution of these European-style options data in Figure 19 and 20.

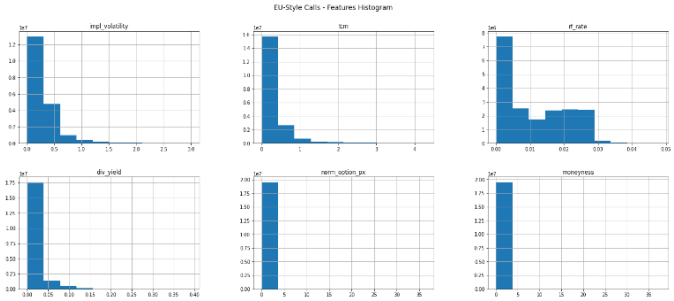

Fig. 19. European-style calls Feature Distribution Histograms

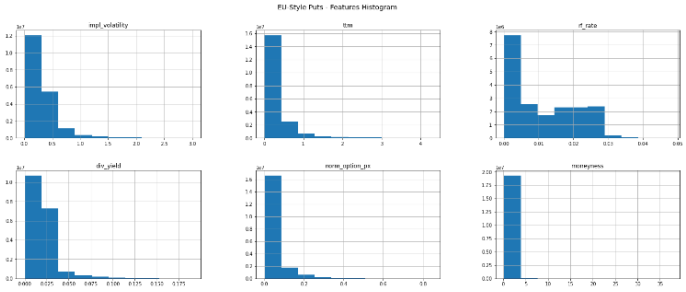

Fig. 20. European-style puts Feature Distribution Histograms

Unlike the prices generated by the parametrized pricing models, the features for the collected data are much more positively skewed. This is expected as real-market listed options on equity underlying is often standardized, and extreme values for certain parameters (i.e. risk-free rate, dividend-yield) is not typically experienced frequently within the time frame of the collected data (2008 – 2020). Subsequently, we observe the moneyness vs. normalized price plot for European-style options in Figures 21 and 22.

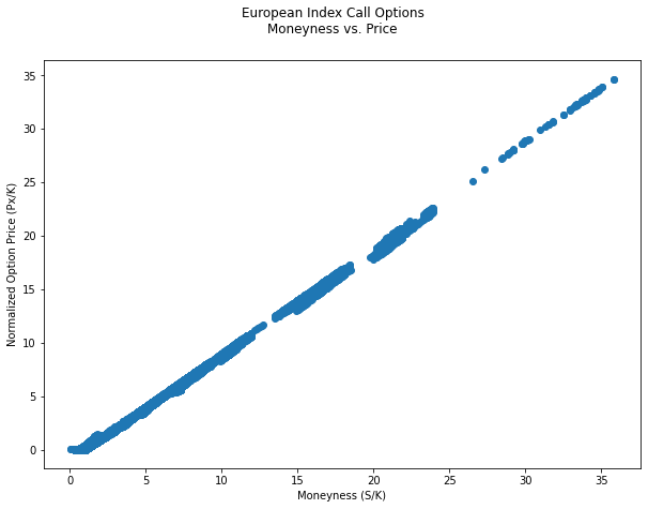

Fig. 21. Moneyness vs. normalized price for EU-style calls

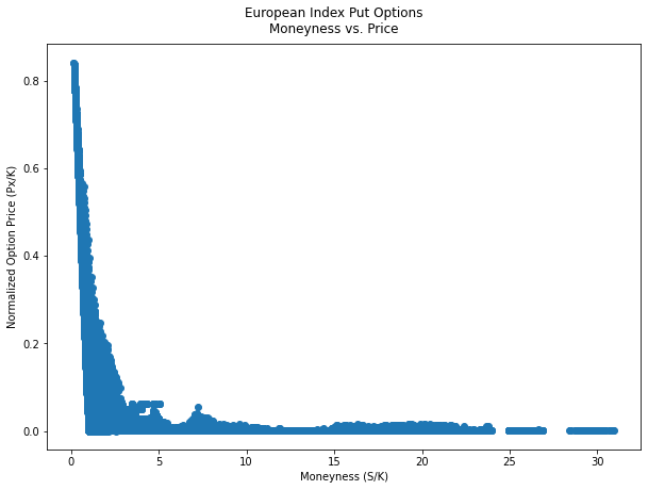

Fig. 22. Moneyness vs. normalized price for EU-style puts

For collected real-world European options data, the moneyness vs. normalized option price plot for European-style calls and puts closely reassemble the payoff diagrams for calls and puts, as shown in Figures 1 and 2. As the call option goes ITM (along the x-axis), the value of the call increases nearly linearly with the moneyness. For puts, the value of the put option increases nearly linearly as the put goes ITM (when the moneyness decreases to the range between 0 and 1).

4) Real-world American-style options

Although over 222M rows of real-world American style options data was collected, only a subset of 23M data points were used for both American-style calls and puts, in order to reduce training time, computation requirements, and match the length of other utilized datasets. The descriptive statistics for American call options are shown in Figure 23, and the same statistics for American put options are shown in Figure 24.

	impl_volatility	secid	ttm	rf_rate	div_yield	norm_option_px	moneyness
count	2.300059e+07	2.300059e+07	2.300059e+07	2.300059e+07	23000590.0	2.300059e+07	2.300059e+07
mean	7.009399e-01	1.432679e+05	3.747926e-01	4.655874e-03	0.0	1.386789e-01	1.206080e+00
std	4.897254e-01	4.247381e+04	4.765190e-01	5.059414e-03	0.0	1.614814e-01	1.038520e+00
min	1.000100e-02	5.139000e+03	2.739726e-03	8.947500e-04	0.0	1.123596e-05	7.447412e-05
25%	3.785660e-01	1.087640e+05	6.849315e-02	1.737424e-03	0.0	1.833333e-02	8.463226e-01
50%	5.546520e-01	1.247710e+05	1.726027e-01	2.221926e-03	0.0	7.500000e-02	1.032000e+00
75%	8.564640e-01	2.033350e+05	4.986301e-01	5.060380e-03	0.0	2.091463e-01	1.314589e+00
max	2.999998e+00	2.150740e+05	2.983562e+00	1.798196e-02	0.0	9.913043e-01	1.113100e+02

Fig. 23. Descriptive Statistics for real-world American-style calls.

	impl_volatility	secid	ttm	rf_rate	div_yield	norm_option_px	moneyness
count	2.300000e+07	2.300000e+07	2.300000e+07	2.300000e+07	23000000.0	2.300000e+07	2.300000e+07
mean	7.220020e-01	1.441252e+05	3.677346e-01	4.633494e-03	0.0	3.102542e-01	1.147577e+00
std	4.844037e-01	4.071769e+04	4.627733e-01	5.019915e-03	0.0	9.862428e-01	1.030550e+00
min	1.000900e-02	1.009580e+05	2.739726e-03	8.947500e-04	0.0	1.428571e-06	4.116667e-03
25%	3.978340e-01	1.079260e+05	7.671233e-02	1.748239e-03	0.0	1.538462e-02	7.981081e-01
50%	5.853540e-01	1.282520e+05	1.753425e-01	2.238765e-03	0.0	8.814103e-02	9.974483e-01
75%	8.774490e-01	1.888080e+05	4.876712e-01	5.038823e-03	0.0	3.201887e-01	1.262400e+00
max	2.999999e+00	2.137110e+05	2.983562e+00	1.798196e-02	0.0	3.281150e+02	3.286000e+02

Fig. 24. Descriptive Statistics for real-world American-style puts.

Note that the real-world American-style options are similar to their simulated counterparts, and that the dividend yield for all American-style options are set to 0, after incorporating discrete dividends in the method described in equation (9). Additionally, we also observe the features distribution of these American-style options data in Figure 25 and 26.

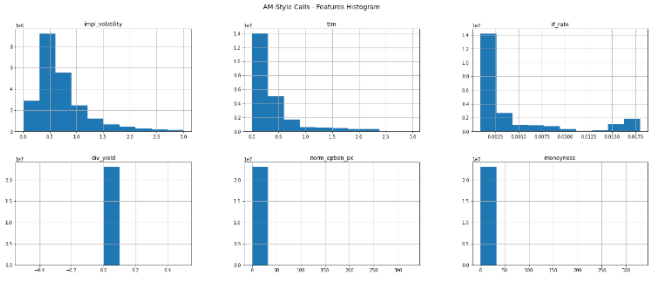

Fig. 25. American-style calls Feature Distribution Histograms

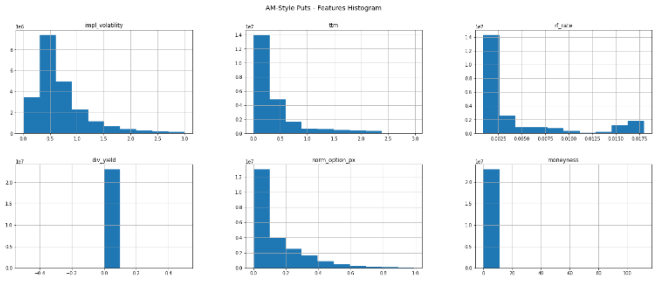

Fig. 26. American-style puts Feature Distribution Histograms

Much like the real-world European-style options data, the real-world American-style options data is also very positively skewed, with a majority of the features occurring within a small range of frequently seen values. Finally, we plot the moneyness vs. normalized price for American-style options in Figures 27 and 28.

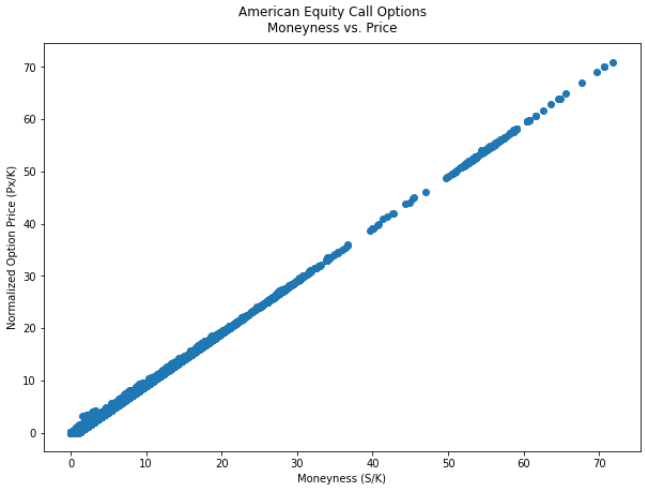

Fig. 27. Moneyness vs. normalized price for AM-style calls

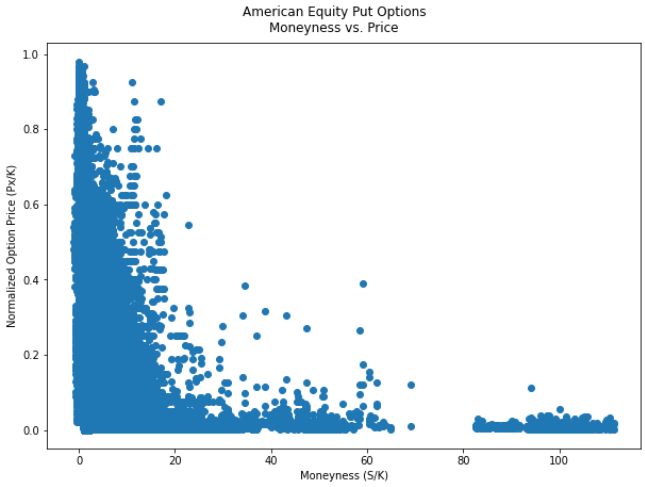

Fig. 28. Moneyness vs. normalized price for AM-style puts

The moneyness vs. normalized price plot for American-style options looks extremely similar to that of the real-world European-style options. However, we note the greater variance in put prices as moneyness decreases. This can be explained by the early-exercise premium for American puts as previously described, as well as a greater premium paid in the real-world for OTM puts relative to ATM put options (i.e. volatility surface convexity).

E. Traditional Supervised Regression Methods

Firstly, as in [5], traditional machine learning techniques for supervised regression are first estimated on the BSOPM-generated option prices in order to provide a comparison benchmark for the accuracy of other nonparametric models. In particular, Lasso, Ridge, ElasticNet regressions, as well as Stochastic Gradient Descent ("SGD") are used on BSOPM-generated call options. Prior to model fitting, in order to improve model performance and introduce potentially important interactions between model features, the dataset was transformed by adding second-order polynomial features, resulting in an augmented dataset with 20 features. Moreover, all datasets were standardized prior to model fitting, as done in [8]. All regression methods were applied to both the normal standardized dataset with 5 features and the dataset with polynomial features, and each model is fitted with MSE as the objective loss function to minimize.

During the model training process, 10-fold cross-validation grid search in firstly performed in order to determine the most optimal model-specific hyperparameters. The hyperparameters of the best performing model (based on cross-validation error on the validation partition) is then used to fit the entire training and validation set, and the complete model is then used to estimate the testing set, at which point the following out-of-sample performance metrics are calculated: MSE, RMSE, R^2, Explained Variance, Max Error, Mean Absolute Error, and Median Absolute Error.

Finally, the best performing model within these supervised regression methods was selected and used to fit the BSOPM-generated put options, following the same standardization and grid search process for hyperparameter tuning. This test was performed in order to determine if the same supervised regression model type is generalizable to put options.

F. Artifical Neural Network Methods

In the next step, a feed-forward ANN is constructed for training on the BSOPM-generated European-style calls and puts, as well as the BOPM-generated American-style puts. The baseline model architecture was inspired by the promising architecture determined in [4] with an input layer of size 5, an output layer of size 1, and 4 hidden layers. An adaptive learning rate for each optimizer was used: the learning rate was initially instantiated at 0.001, and subsequently reduced by a factor of 10 after 3 epochs of unimproved training loss. Furthermore, the hyperparameters of the model were tuned using 3-fold cross-validation grid search for each dataset within the hyperparameter ranges specified in Figure 29. Moreover, additional models were trained with custom activation functions such as the MELU activation function introduced in [9], in order to preserve activation function differentiability and a non-negative output in the final layer. The most accurate ANN architectures and model hyperparameters were determined for each dataset using various out-of-sample error metrics measured on the test set, including MSE, RMSE, R^2, Explained Variance, Max Error, Mean Absolute Error, and Median Absolute Error.

Hyperparameter	Range
Batch Size	[128, 512, 256, 2048, 3000]
Optimizer	['sgd', 'adam','rmsprop']
Dropout Rate	[0, 0.1, 0.2]
Activation	['sigmoid', 'elu', 'tanh', 'custom1', 'custom2']
Normalize before training	[True, False]
Number of Neurons	[200, 400, 600]
Training Epochs	[20,50,100,200,400]

Fig. 29. ANN Grid Search for Hyperameter Tuning Ranges

The grid-search procedure consisted of a broad search across all hyperparameters in the given combinations above, for a total of 6750 fitted models per grid-search. Note that since option values cannot be negative, the activation function in the final layer was ensured to be non-negative (i.e. 'ReLU' was used by default). Aside from the traditional activation functions, the 'custom1' and 'custom2' activation functions represent the sequence of activation functions inspired by the literature in [4] and [9], respectively. The 'custom1' activation function involves using the following activation functions in sequence: LeakyReLU, ELU, ReLU, ELU, and exp() (exponential function). Similarly, the 'custom2' activation functions involve using the following activation functions in sequence: LeakyReLU, MELU (as described in equation (7), MELU, MELU, Softplus – 0.5. Softplus - 0.5 is described as follows:

$$f(x) = \ln(1 + e^{x}) - 0.5 \tag{8}$$

As an activation function in the final layer, this activation function is non-negative, and can also ensure that fractional values of x close to 0 are outputted closer to 0.

Subsequently, the most accurate ANN architectures and hyperparameters determined from fitting the BSOPM and BOPM-generated option prices were then implemented as baseline models to train on their respective real-world European and American-style call and put options. These models were also tuned using 3-fold cross-validation grid search, and alternative activation functions as suggested in [9] were also implemented and tests. Out-of-sample test set error metrics were collected for each of these models in order to determine the best model architecture and hyperparameters to use for each dataset.

G. Homogeneity Hint Introduction

In order to meet objective O5, a homogeneity hint as described in [8] was introduced in order to determine if such a training technique had the potential to improve nonparametric model accuracies. This homogeneity hint was introduced to the BSOPM-generated European-style options data as well as the real-world American-style options data, as a means to determine the potential accuracy improvements of such a training method on both option styles and on both theoretical and real-world option prices.

The homogeneity hint was introduced by first classifying the aforementioned datasets into 6 categories (by moneyness and maturity). The classification thresholds can be seen in Table 1 below, and are consistent with the thresholds seen in [8].

TABLE I. HOMOGENEITY HINT CLASSIFICATION THRESHOLDS

Classification Thresholds by Moneyness and Maturity	
Moneyness	***Threshold***
ATM	0.95 <= Moneyness <=1.05
ITM Calls/OTM Puts	Moneyness > 1.05
OTM Calls/ITM Puts	Moneyness < 0.95
Maturity	***Threshold***
Short-dated	Maturity >= 60 days (0.1644 years)
Long-dated	Maturity < 60 days (0.1644 years)

The distribution of option prices after classifying all datasets by moneyness and maturity according to the above thresholds can be seen in Figure 9 below.

	Call Option Prices by Moneyness and Maturity						
	Calls						
	Short Maturity (<= 60 Days)			**Long Maturity (> 60 Days)**			
Dataset	**ITM**	**ATM**	**OTM**	**ITM**	**ATM**	**OTM**	**Total**
BSOPM	656,100	243,000	656,100	7,873,200	2,916,000	7,873,200	**20,217,600**
American Options	4,519,288	1,969,412	4,772,496	5,387,805	1,152,700	5,198,299	**23,000,000**
European Options	4,633,080	3,451,804	2,524,791	4,676,518	1,900,894	2,331,987	**19,519,074**

	Put Option Prices by Moneyness and Maturity						
	Puts						
	Short Maturity (<= 60 Days)			**Long Maturity (> 60 Days)**			
Dataset	**ITM**	**ATM**	**OTM**	**ITM**	**ATM**	**OTM**	**Total**
BOPM	656,100	243,000	656,100	7,873,200	2,916,000	7,873,200	**20,217,600**
BSOPM	656,100	243,000	656,100	7,873,200	2,916,000	7,873,200	**20,217,600**
American Options	5,131,586	2,783,890	6,992,255	6,293,260	1,613,202	7,185,807	**30,000,000**
European Options	1,476,797	3,278,751	6,127,145	1,765,409	1,873,059	4,848,113	**19,369,274**

Fig. 30. Distribution of Calls and Puts for all datasets after Homogeneity Hint Classification

After the BSOPM-generated calls and puts and the real-world American-style options calls and puts are classified, 6 ANNs are trained for each dataset, one for each classification type (i.e. combination of moneyness and maturity), for a total of 24 ANNs. The architecture of the ANNs used to fit each dataset is the previously determined optimal model architectures for each of the BSOPM-generated calls and puts as well as the real-world American-style calls and puts.

Finally, the out-of-sample error metrics across each classification type is examined for each dataset, in order to identify potential significant discrepancies in ANN forecast accuracy depending on the option classification type. Moreover, performance is also compared to the respective generalized models' performances for each option style and

type, in order to determine if the homogeneity hint introduced actually helped improve nonparametric model accuracy.

H. Results Evaluation

In order to evaluate the effectiveness of the trained nonparametric pricing models in pricing real-world options, an accuracy comparison was then conducted by comparing the out-of-sample error metrics of the most accurate ANNs for each type and style of option to the error metrics of the BSOPM and BOPM parametric models by respectively pricing the same set of real-world European and American option prices. Consequently, this test result allows us to evaluate the performance of nonparametric option pricing models vs. their traditional parametric counterparts in pricing real-world option prices.

V. Results

Since 10% of each dataset is preprocessed and reserved as a testing set, the out-of-sample testing error across various error metrics (i.e. MSE, RMSE, R^2, Explained Variance, Max Error, Mean Absolute Error, and Median Absolute Error) is used as the primary way of identifying model accuracy. Moreover, we also plot the predicted option values against the actual option values, as well as the distribution of residuals (i.e. residuals = predicted – actual option prices) to observe any heteroskedasticity or excessive variance in the predictions of the model. Additionally, we also plot the residuals against various features such as moneyness, implied volatility, and maturity, in order to isolate and observe any differences in prediction error by feature.

A. Generated Option Prices

1) Traditional Supervised Regression Methods

Figure 31 shows the training MSE and the out-of-sample test error metrics of the various supervised regression models after grid search is performed and the full model with the tuned hyperparameters is trained using both training and validation partitions on the BSOPM-generated call options dataset.

Error Metrics for Supervised Regression Methods on BSOPM-generated Calls					
Model Type	**Ridge Regression - Quadratic Features, Standardized**	**Lasso Regression - Quadratic Features, Standardized**	**ElasticNet Regression - Quadratic Features, Standardized**	**Stochastic Gradient Descent - Quadratic Features, Standardized**	**Stochastic Gradient Descent - Standardized**
Error Metrics					
Training MSE	1.395E-01	1.395E-01	1.414E-01	1.395E-01	8.191E-01
Test MSE	1.442E-01	1.442E-01	1.475E-01	1.439E-01	8.401E-01
RMSE	3.798E-01	3.798E-01	3.841E-01	3.793E-01	9.166E-01
R2	9.992E-01	9.992E-01	9.992E-01	9.992E-01	9.955E-01
Explained Variance	9.992E-01	9.992E-01	9.992E-01	9.992E-01	9.955E-01
Max Error	7.663E+00	7.662E+00	7.604E+00	7.637E+00	1.738E+01
Mean Absolute Error	1.829E-01	1.828E-01	1.863E-01	1.827E-01	3.996E-01
Median Absolute Error	1.024E-01	1.024E-01	1.043E-01	1.023E-01	2.271E-01

Fig. 31. Heatmap Summary of Test Errors of Supervised Regression Methods on BSOPM-generated calls. Green represents the lowest value, and red represents the highest value.

From the above results, it can be seen that introducing second-order polynomial features to the dataset helped to improve model performance, as the SGD model without quadratic features had the poorest performance on all error metrics by a significant margin. Furthermore, it can be seen that the SGD model with quadratic features scored the highest on all test set measures except for max error (albeit by a slim margin), and thus remains the most accurate model fitted using traditional supervised regression methods. We now plot the model's test set predictions against the true values, as well as the distribution of residuals of the model's predictions in Figure 32.

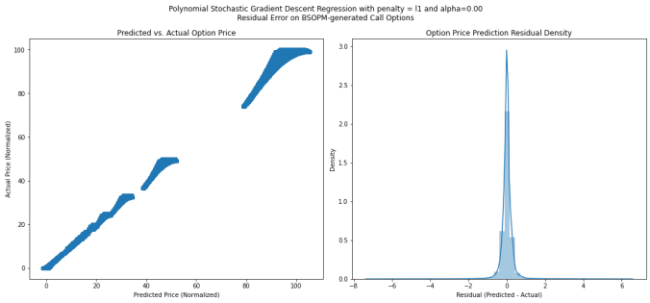

Fig. 32. Residual density of polynomial SGD Regression on BSOPM-generated calls.

Since a model with perfect predictive capabilities would showcase all predictions lying on the $y = x$ line for the predicted vs. actual value plot, as well as a concentrated residual density exactly at 0, we can see that this SGD model has substantive prediction error, particularly as the true normalized option price gets large. Moreover, despite a symmetric residual density graph, we see long left and right tails in the distribution, thus highlighting infrequent large prediction errors.

Subsequent to finding the best model architecture for BSOPM-generated calls, the SGD model with quadratic features was fitted to the dataset containing BSOPM-generated puts, in order to determine if this supervised regression method is generalizable between option types. Prior to fitting the entire dataset, grid search was performed against the put options dataset to tune the model's hyperparameters to best suit the put option dataset. The model was then fitted on the entire training and validation partitions. The error metrics on the BSOPM-generated puts are summarized in Table II, and the same residual distribution plot is shown in Figure 33.

TABLE II. Error Metrics of SGD w/ Quadratic Features on BSOPM-generated Put Options

Error Metric	***Value***
Training MSE	2.507E-02
Test MSE	2.509E-02
RMSE	1.584E-01
R^2	7.456E-01
Explained Variance	7.456E-01
Max Error	1.004E+00
Mean Absolute Error	1.091E-01
Median Absolute Error	7.119E-02

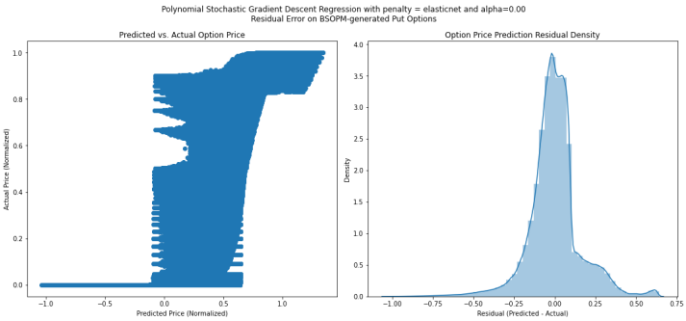

Fig. 33. Residual density of polynomial SGD Regression on BSOPM-generated puts.

With a test set R^2 of 0.745 and mean absolute error of 0.183, it can be seen from the error metrics and the plotted residuals that the SGD model with polynomial features does a poor job of fitting and predicting put option values for the BSOPM-generated put options dataset. Hence, it was found that although the SGD method had a fair performance accuracy in predicting BSOPM-generated call options, the same supervised regression method performance could not be replicated across option types. Finally, the SGD model fitted on BSOPM-generated calls' out-of-sample R^2 of 0.9992 and MSE of 0.1439 now becomes the performance benchmark for subsequent ANN-based models.

2) Artifical Neural Network Methods

After running grid-search to tune ANN hyperparameters for all three datasets containing generated option prices, the test error metrics of the best performing models for each dataset is shown in Figure 34, and the model architecture details for each dataset is shown in Figure 35.

Error Metrics for Best-performing ANNs on Simulated Option Prices			
Dataset	**BSOPM - Calls**	**BSOPM - Puts**	**BOPM - Puts**
Error Metric			
Last Epoch Train Loss (MSE)	3.261E-06	1.103E-06	2.000E-05
Last Epoch CV Loss (MSE)	3.106E-06	1.655E-07	9.563E-07
Overall Test Loss (MSE)	3.121E-06	1.657E-07	9.567E-07
Test RMSE	1.767E-03	4.070E-04	9.781E-04
R2	1.000E+00	1.000E+00	1.000E+00
Explained Variance	1.000E+00	1.000E+00	1.000E+00
Max Error	3.134E-02	7.429E-03	2.372E-02
Mean Absolute Error	9.943E-04	2.845E-04	7.340E-04
Median Absolute Error	5.276E-04	2.026E-04	5.881E-04

Fig. 34. Error Metrics for Best-performing ANNs on Simulated Option Prices.

Best-performing ANN Architecture, Simulated Option Prices			
Hyperparameter	**BSOPM - Calls**	**BSOPM - Puts**	**BOPM - Puts**
Batch Size	2048	2048	2048
Optimizer	adam	adam	adam
Dropout Rate	0	0	0
Activation	custom2	custom1	elu
Normalize before training	FALSE	FALSE	FALSE
Number of Neurons	200	200	200
Training Epochs	100	100	100

Fig. 35. Hyperparameters for best-performing ANNs on Simulated Option Prices.

It can be seen from the error metrics table that the fitted ANNs all perform very accurately on out-of-sample test set evaluation: all three ANNs perform much better than the supervised regression SGD benchmark in fitting their respective datasets. Furthermore, the best-performing model architectures of all three ANNs are similar, with the only difference being the set of activation functions used. Specifically, a diagram of the ANN fitting the BOPM-generated puts is shown in Figure 36, and a diagram of the ANN fitting the BSOPM-generated puts is shown in Figure 37.

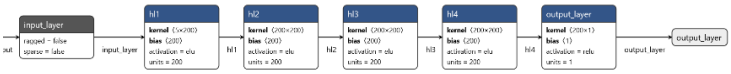

Fig. 36. Diagram of the ANN fitting BOPM-generated puts. The last activation function is "relu" by default to guarantee a non-negeative output.

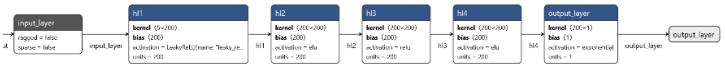

Fig. 37. Diagram of the ANN fitting BSOPM-generated puts. The activations used are the "custom1" series of functions previously described.

The training loss history of the ANN fitting BSOPM - generated calls, BSOPM-generated puts, and BOPM-generated puts can be seen in Figures 38, 39, and 40, respectively.

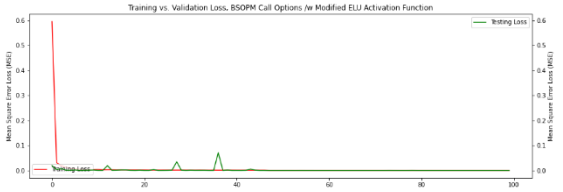

Fig. 38. Training and Validation Loss History, BSOPM-generated calls.

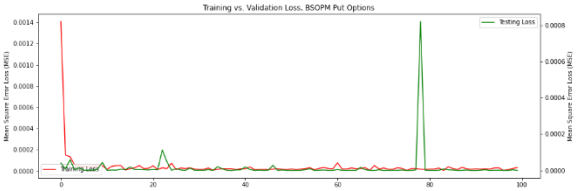

Fig. 39. Training and Validation Loss History, BSOPM-generated puts.

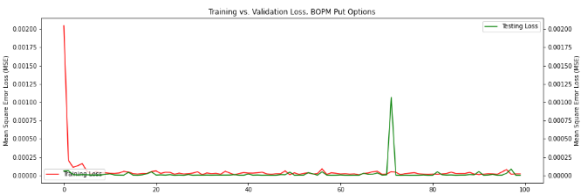

Fig. 40. Training and Validation Loss History, BOPM-generated puts.

It can be seen that all ANNs converge after 100 epochs, with cross-validation MSEs ending below training MSEs, despite temporary "spikes" in validation loss during training. In order to verify the predictive ability of the trained ANNs, we plot the residual distributions of these ANNs in the following figures.

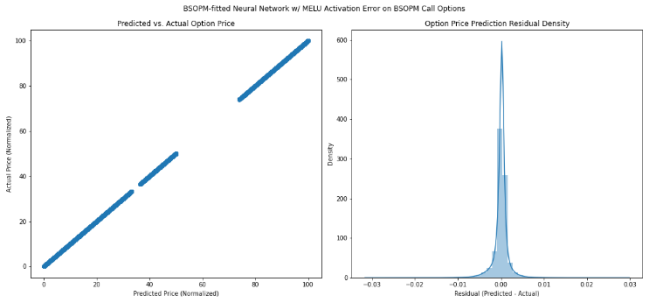

Fig. 41. Residual density of ANN trained on BSOPM-generated calls.

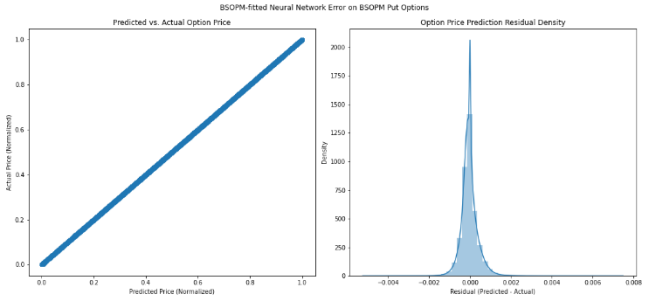

Fig. 42. Residual density of ANN trained on BSOPM-generated puts.

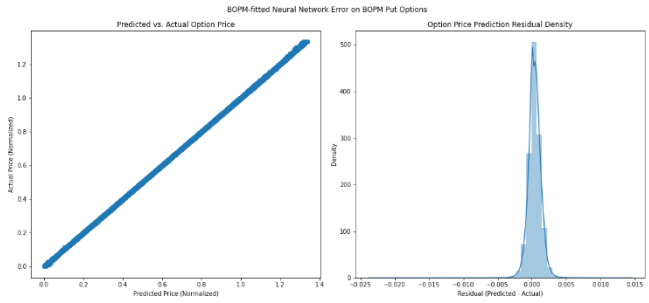

Fig. 43. Residual density of ANN trained on BOPM-generated puts.

In these residual density plots, we see that the predicted vs. actual plot looks near-perfect for all three models. Additionally, despite the residuals in all three figures being concentrated at zero, there exists some variance in the residuals. The residuals of the ANN pricing BSOPM-generated puts form a slight positive skew, whereas the residuals of the ANN pricing BOPM-generated puts form a slight negative skew. Nevertheless, it was shown that these trained ANNs were able to price theoretically generated option prices from parametric pricing models for both European and American options with a great degree of accuracy.

B. Real-world Option Prices

The error metrics of the best-performing ANNs fitting real-world American and European style calls and puts are shown in Figure 44. The respective model architectures for each of these ANNs are shown in Figure 45.

Error Metrics for Best-performing ANNs on Real-world Option Prices				
Dataset	**European-style Calls**	**European-style Puts**	**American-style Calls**	**American-style Puts**
Error Metric				
Last Epoch Train Loss (MSE)	2.606E-07	1.102E-07	7.395E-05	2.137E-05
Last Epoch CV Loss (MSE)	2.517E-07	8.345E-08	7.578E-05	2.128E-05
Overall Test Loss (MSE)	2.448E-07	6.502E-08	7.177E-05	2.165E-05
Test RMSE	4.947E-04	2.550E-04	8.472E-03	4.653E-03
R2	1.000E+00	1.000E+00	9.999E-01	9.992E-01
Explained Variance	1.000E+00	1.000E+00	9.999E-01	9.992E-01
Max Error	9.623E-02	1.941E-01	1.834E+00	3.665E-01
Mean Absolute Error	2.055E-04	9.245E-05	2.271E-03	1.149E-03
Median Absolute Error	1.314E-04	5.199E-05	7.186E-04	1.856E-04

Fig. 44. Error Metrics for best-performing ANNs on Real-world Option Prices.

Best-performing ANN Architecture, Real-world Option Prices				
Hyperparameter	**European-style Calls**	**European-style Puts**	**American-style Calls**	**American-style Puts**
Batch Size	2048	2048	2048	2048
Optimizer	adam	adam	adam	adam
Dropout Rate	0	0	0	0
Activation	custom2	custom2	elu	custom2
Normalize before training	FALSE	FALSE	FALSE	FALSE
Number of Neurons	200	200	200	200
Training Epochs	100	100	100	100

Fig. 45. Hyperparameters for best-performing ANNs on Real-world Option Prices.

Once again, all out-of-sample error metrics of the best-performing ANNs for each dataset all outperform the supervised regression benchmark. In fact, several of these fitted ANNs performed better on real-world data than on the generated options data: the ANNs fitting the real-world European-style calls and puts both have lower test MSEs than the ANNs fitting the BSOPM-generated European-style calls and puts. Furthermore, apart from changes in the activation functions across each model, the architecture of the best-performing ANNs are found to be the same across both real-world and generated data. The "custom2" series of activation functions also seems to perform particularly well when applied to real-world options data.

Figures 46, 47, 48, and 49 show the training and validation history for each of these fitted ANN models.

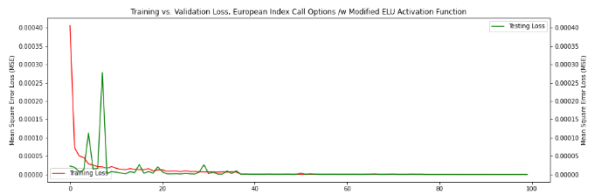

Fig. 46. Training and Validation Loss History, European-style calls.

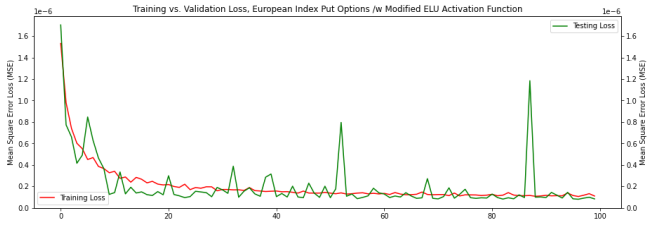

Fig. 47. Training and Validation Loss History, European-style puts.

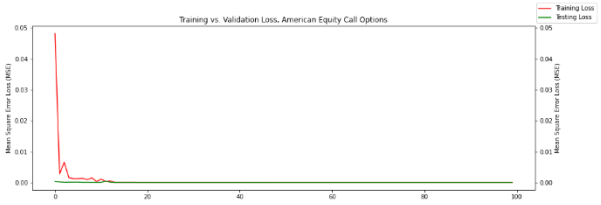

Fig. 48. Training and Validation Loss History, American-style calls.

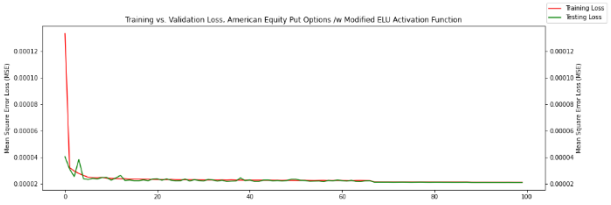

Fig. 49. Training and Validation Loss History, American-style puts.

In each of the above training loss plots for each ANN, we see validation and training loss steadily decrease and converge by the 100th epoch. Since the validation and out-of-sample test loss for these ANNs are below the training loss, there seems to be no issue of these ANNs overfitting the dataset. Subsequently, we show the residual distribution for each of these models in Figures 50, 51, 52, and 53 in order to observe model prediction accuracy in greater detail.

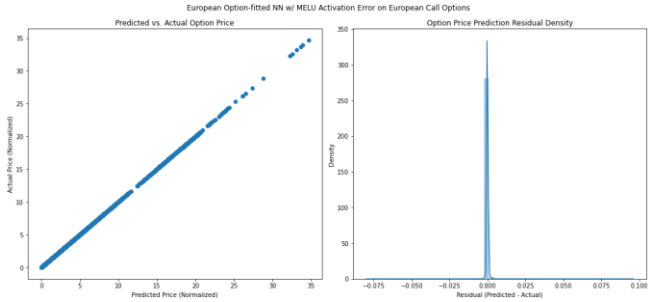

Fig. 50. Residual density of ANN trained on European-style calls.

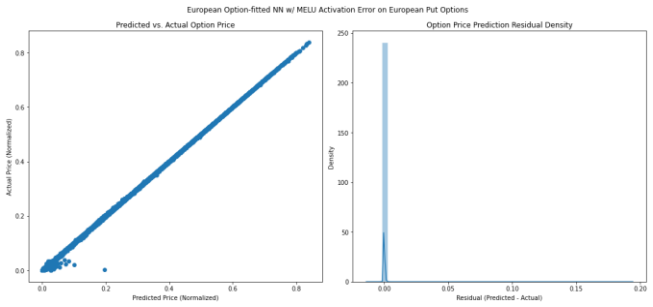

Fig. 51. Residual density of ANN trained on European-style puts.

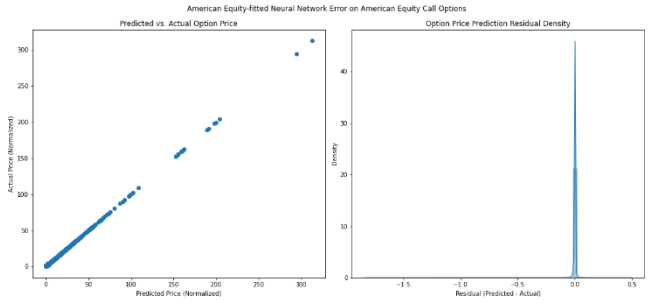

Fig. 52. Residual density of ANN trained on American-style calls.

Fig. 53. Residual density of ANN trained on American-style puts.

Figure 50 shows a near-perfect prediction plot for the ANN fitting European-style calls: there is little variance in the prediction vs. actual value scatter plot, and the distribution of the residuals are concentrated and centered around 0. The plot for European-style puts show that the model makes errors when the actual option price is low, and the residual density is positively skewed, thus indicating that the model tends to predict values that are too large. The residual density plot of the American-style calls also looks near-perfect, but the residual density is now negatively skewed, indicating that the model tends to predict values that are too small. Finally, the plot for the American-style puts look the worst, as there is considerable variance and heteroskedasticity exhibited in the predicted vs. actual price plot. Nevertheless, the residual density plot is centered, so there is no tendency for the model to predict values that are too low or too high.

Finally, we plot the residual error by several model features (i.e. maturity, implied volatility, and moneyness) in

an attempt to discover if these ANNs perform differently for different ranges of feature parameters.

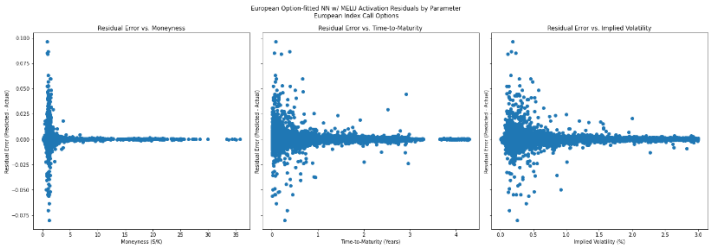

Fig. 54. Residuals by feature, ANN trained on European-style calls.

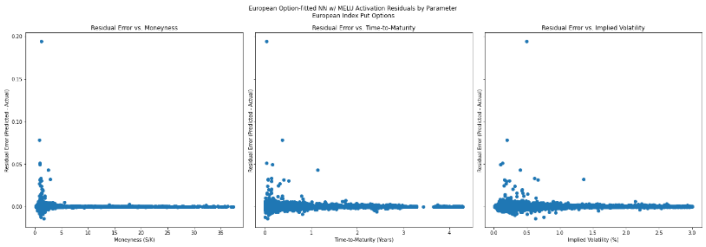

Fig. 55. Residuals by feature, ANN trained on European-style puts.

Fig. 56. Residuals by feature, ANN trained on American-style calls.

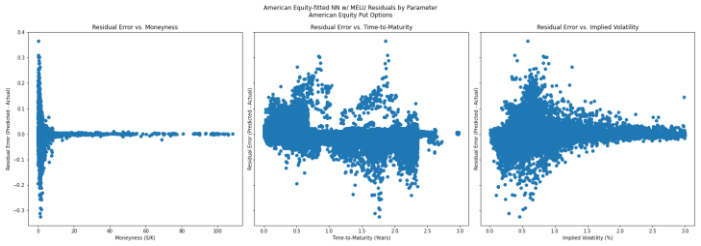

Fig. 57. Residuals by feature, ANN trained on American-style puts.

In the case of European-style calls, it can be observed that although the residuals are centered around 0, there is significant variance in the residuals when implied volatility, maturity, and moneyness are low. For European-style puts, the skew of the residuals shows that residuals tend to be positive. For American-style calls, the opposite is true, with residuals having a tendency of being negative. Moreover, there is significant variance in the residuals when moneyness is low and when implied volatility increases higher. Finally, for American-style puts, there is significant variance in the residuals when moneyness is low, when implied volatility is between 0.4 – 1.2, and at all times to maturity.

In sum, it was found that real-world call option prices for both European-style and American-style options can be accurately priced by ANNs with a great degree of accuracy. Contrastingly, put options are harder to price accurately, with American-style put options exhibiting the most pricing difficulty for ANNs.

C. *Homogeneity Hint Introduction*

A homogeneity hint was introduced in the BSOPM-generated options data as well as the real-world American-style options data by categorizing the options in the dataset by moneyness and maturity, and an ANN was fitted for each of the option categories. The architecture of the ANN trained is the previously determined best performing model architecture for each respective dataset. The error metrics achieved from the homogeneity hint introduction on both BSOPM-generated data and real-world American-style options data can be seen in Figure 58 and 59.

	ANN Error Metrics for BSOPM-generated Options with Homogeneity Hint											
	Calls						Puts					
	Short Maturity (<= 60 Days)			Long Maturity (> 60 Days)			Short Maturity (<= 60 Days)			Long Maturity (> 60 Days)		
Error Metrics	ITM	ATM	OTM	ITM	ATM	OTM	ITM	ATM	OTM	ITM	ATM	OTM
Last Epoch Train Loss (MSE)	2.485E-03	1.074E-01	2.653E-02	1.471E-03	5.813E-01	1.867E-01	5.906E-06	1.693E-02	4.277E-02	3.789E-07	9.839E-07	3.608E-06
Last Epoch CV Loss (MSE)	5.758E-04	1.082E-01	2.656E-02	5.557E-05	5.809E-01	1.869E-01	2.977E-06	1.264E-03	4.291E-02	3.237E-07	3.163E-07	2.850E-06
Overall Test Loss (MSE)	5.690E-04	1.081E-01	2.604E-02	5.543E-05	5.805E-01	1.873E-01	2.975E-06	1.289E-03	4.181E-02	3.251E-07	3.164E-07	2.833E-06
Test RMSE	2.385E-02	3.288E-01	1.614E-01	7.445E-03	7.619E-01	4.328E-01	3.244E-05	3.244E-05	3.244E-05	3.244E-05	3.244E-05	3.244E-05
R2	1.000E+00	-6.788E-01	-3.129E-01	1.000E+00	-6.901E+00	-1.999E+00	9.999E-01	9.801E-01	-2.889E-01	1.000E+00	1.000E+00	1.000E+00
Explained Variance	1.000E+00	0.000E+00	0.000E+00	1.000E+00	0.000E+00	0.000E+00	1.000E+00	9.813E-01	-2.324E-06	1.000E+00	1.000E+00	1.000E+00
Max Error	2.837E-01	7.859E-01	6.690E-01	8.280E-02	1.042E+00	9.091E-01	7.607E-03	1.179E-01	7.421E-01	1.706E-02	7.072E-03	1.986E-02
Mean Absolute Error	1.127E-02	2.090E-01	7.878E-02	4.890E-03	7.121E-01	3.533E-01	1.424E-03	2.902E-02	9.681E-02	4.778E-04	4.307E-04	1.157E-03
Median Absolute Error	7.047E-03	4.167E-02	0.000E+00	3.020E-03	8.214E-01	3.748E-01	1.286E-03	2.549E-02	1.578E-07	4.458E-04	3.598E-04	7.838E-04

Fig. 58. ANN Error Metrics for BSOPM-generated options with Homogeneity Hint, highlighted from low (green) to high (red).

	ANN Error Metrics for American-style Options with Homogeneity Hint											
	Calls						Puts					
	Short Maturity (<= 60 Days)			Long Maturity (> 60 Days)			Short Maturity (<= 60 Days)			Long Maturity (> 60 Days)		
Error Metrics	ITM	ATM	OTM	ITM	ATM	OTM	ITM	ATM	OTM	ITM	ATM	OTM
Last Epoch Train Loss (MSE)	2.059E-05	3.582E-03	5.873E-04	4.059E+00	2.175E-02	5.978E-03	4.975E-06	1.750E-06	8.766E-04	7.334E-05	4.066E-05	2.824E-05
Last Epoch CV Loss (MSE)	2.015E-05	3.557E-03	5.840E-04	4.387E+00	2.169E-02	5.955E-03	4.757E-06	1.702E-06	8.760E-04	7.551E-05	4.094E-05	2.804E-05
Overall Test Loss (MSE)	1.993E-05	3.569E-03	5.878E-04	4.390E+00	2.168E-02	6.012E-03	5.486E-06	1.745E-06	8.791E-04	7.553E-05	3.898E-05	2.749E-05
Test RMSE	4.465E-03	5.974E-02	2.424E-02	2.095E+00	1.472E-01	7.754E-02	2.342E-03	1.321E-03	2.965E-02	8.691E-03	6.243E-03	5.243E-03
R2	9.999E-01	-2.043E+00	-4.470E-01	-2.008E-01	-2.340E+00	-6.369E-01	9.997E-01	9.985E-01	-4.834E-01	9.975E-01	9.942E-01	9.953E-01
Explained Variance	9.999E-01	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	9.997E-01	9.985E-01	7.772E-16	9.975E-01	9.942E-01	9.953E-01
Max Error	7.182E-01	4.000E-01	5.000E-01	3.127E+02	1.053E+00	1.550E+00	4.424E-01	8.442E-02	3.914E-01	5.121E-01	2.773E-01	4.459E-01
Mean Absolute Error	1.529E-03	4.895E-02	1.348E-02	8.568E-01	1.232E-01	4.836E-02	4.888E-04	2.672E-04	1.693E-02	2.960E-03	2.554E-03	1.435E-03
Median Absolute Error	7.427E-04	4.387E-02	5.833E-03	5.000E-01	1.042E-01	2.595E-02	2.062E-04	8.907E-05	8.130E-03	8.715E-04	1.146E-03	3.805E-04

Fig. 59. ANN Error Metrics for American-style options with Homogeneity Hint, highlighted from low (green) to high (red).

It can be observed from Figure 58 that the fitted ANNs perform best in pricing long-maturity puts, as well as short-maturity ITM puts, and long-maturity ITM calls. Conversely, the ANN does a poor job in pricing long-maturity ATM calls, as well as short-maturity ATM calls. When comparing the error metrics of the ANNs trained with homogeneity hint to the error metrics of the generalized ANNs for BSOPM-generated calls and puts (Figure 34), the introduction of the homogeneity hint does not seem to improve the prediction accuracy for each ANN model. In fact, the generalized ANN

models actually score lower in out-of-sample test MSEs than the same scores presented in Figure 34.

For real-world American-style options, the ANNs performs best in pricing long-maturity puts, short-maturity ITM and ATM puts, and short-maturity ITM calls. Conversely, the models do poorly in pricing ITM and ATM long-maturity calls. When compared to the error metrics of the generalized ANNs for American-style options (Figure 44), the out-of-sample test MSEs is again lower for the generalized models when compared to the same scores in Figure 44. As a result, it was empirically determined that the introduction of a homogeneity hint into the dataset prior to model training does not seem to improve the pricing accuracy of the trained ANNs.

D. Model Performance Comparison

Finally, we compare the pricing accuracy of our trained ANNs on real-world European-style and American-style options to the pricing accuracy of traditional standard parametric options pricing models, the BSOPM and the BOPM, in order to determine if nonparametric options pricing models can serve as more accurate pricing alternatives in capturing real-world option price dynamics. The error metrics for the European-style options comparison are shown in Figure 60, and the same metrics for American-style options is shown in Figure 61.

Accuracy Comparison on Real-World European-style Options				
	BSOPM Model		EU Equity-fitted NN	
Error	**Calls**	**Puts**	**Calls**	**Puts**
Mean Squared Error	5.16E-08	7.53E-08	2.45E-07	6.50E-08
Root Mean Squared Error	2.27E-04	2.74E-04	4.95E-04	2.55E-04
R2	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Explained Variance	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Max Error	5.47E-02	1.79E-01	9.62E-02	1.94E-01
Mean Absolute Error	4.46E-05	5.12E-05	2.06E-04	9.25E-05
Median Absolute Error	4.18E-08	2.70E-08	1.31E-04	5.20E-05

Fig. 60. Accuracy comparison on real-world European-style options, highlighted from low (green) to high (red).

Accuracy Comparison on Real-World American-style Options				
	BOPM Model		AM Equity-fitted NN	
Error	**Calls**	**Puts**	**Calls**	**Puts**
Mean Squared Error	3.87E-05	1.33E-02	7.18E-05	2.17E-05
Root Mean Squared Error	6.22E-03	1.15E-01	8.47E-03	4.65E-03
R2	9.99E-01	9.65E-01	1.00E+00	9.99E-01
Explained Variance	9.99E-01	9.65E-01	1.00E+00	9.99E-01
Max Error	1.81E+00	1.35E+02	1.83E+00	3.67E-01
Mean Absolute Error	1.28E-03	1.53E-03	2.27E-03	1.15E-03
Median Absolute Error	7.20E-06	5.13E-06	7.19E-04	1.86E-04

Fig. 61. Accuracy comparison on real-world American-style options, highlighted from low (green) to high (red).

For real-world European-style options, it can be observed that the BSOPM is more accurate that our trained ANN in the pricing of call prices. However, when pricing European-style puts, our trained ANN for European-style puts has a higher pricing accuracy than the BSOPM. For American-style options, it can be seen that our ANN for pricing American-style puts has a better accuracy than the BOPM, but our ANN for pricing American-style calls performs slightly worse than the BOPM.

VI. Conclusions

In this research project, we evaluated the viability of utilizing machine learning methods, in particular Artificial Neural Networks, to construct nonparametric models as a means to price financial options. The objective of this research was to augment the current literature by evaluating the viability of this pricing approach across both option types and option exercise styles while utilizing extensive amounts of both simulated and real-world options price data.

It was discovered that although traditional supervised regression techniques did not perform as well in approximating theoretical option prices, ANNs were found to be very accurate in its ability to price both simulated and real-world option prices. In particular, it was discovered that ANNs may produce more accurate prices for real-world put options than theoretical pricing models such as the BSOPM or the BOPM. From a machine learning standpoint, it was discovered that the main discrepancy in model architectures between ANNs trained to price different option types and exercise styles remained focused on the series of activation functions used in model construction. Finally, it was seen that the introduction of a homogeneity hint in the dataset prior to ANN training did not improve the pricing accuracy of the ANN models.

A. Future Work

In the future, we hope to evaluate a hybrid options pricing method, in which we utilize parametric options pricing models such as the BSOPM or the BOPM as a first step in determining an estimated option price, and then using that option price as an input feature when training the nonparametric pricing model. The goal of this project would be to use ANNs as means to calibrate between the theoretical options price and the observed real-world options price.

References

[1] F. Black and M. Scholes, "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, vol. 81, no. 3, pp. 637–654, 1973.

[2] R. C. Merton, "Theory of Rational Option Pricing," *The Bell Journal of Economics and Management Science*, vol. 4, no. 1, p. 141, 1973.

[3] J. C. Cox, S. A. Ross, and M. Rubinstein, "Option pricing: A simplified approach," *Journal of Financial Economics*, vol. 7, no. 3, pp. 229–263, 1979.

[4] R. Culkin and S. R. Das, "Machine learning in finance: the case of deep learning for option pricing," *Journal of Investment Management*, 2017.

[5] J. Hutchinson, A. Lo, and T. Poggio, "A Nonparametric Approach to Pricing and Hedging Derivative Securities Via Learning Networks," 1994.

[6] A. Deoda, "Option Pricing using Machine Learning techniques," *Indian Institute of Technology*, 2011.

[7] R. Chowdhury, M. Mahdy, T. N. Alam, G. D. A. Quaderi, and M. A. Rahman, "Predicting the stock price of frontier markets using machine

learning and modified Black–Scholes Option pricing model," *Physica A: Statistical Mechanics and its Applications*, vol. 555, p. 124444, 2020.

[8] S. P. Das and S. Padhy, "A new hybrid parametric and machine learning model with homogeneity hint for European-style index option pricing," *Neural Computing and Applications*, vol. 28, no. 12, pp. 4061–4077, 2016.

[9] A. Itkin, "Deep learning calibration of option pricing models: some pitfalls and solutions," 2019.

[10] W. R. D. Services, "Wharton Research Data Services," *WRDS*. [Online]. Available: https://wrds-www.wharton.upenn.edu/. [Accessed: 11-Apr-2021].