From Price Mispricing to Volatility Surface Arbitrage: A CRR Binomial Framework for Delta-Neutral Options Research

Roan McReynolds[1], Jayden Thinh-To[1], Jordan Odorico[1], and Adam Baldesarra[1]

[1]QUANTT — Queen's University Algorithmic & Network Trading Team

March 2026

Abstract

We develop and evaluate a progression of delta-neutral options strategies based on the Cox-Ross-Rubinstein (CRR) binomial model. A baseline Z-score mispricing strategy, entering when market price deviates from a rolling historical-volatility CRR fair value, produces no persistent edge on SPY options during 2023–2024, with mispricings structurally attributable to the volatility risk premium rather than model arbitrage. We redesign around a same-day cross-sectional implied volatility surface: implied volatilities are solved via CRR bisection across the full option chain, a polynomial surface is fitted in log-moneyness and time-to-expiry space, and contracts deviating materially from surface-fitted IV are traded against the surface-fair CRR price with a CRR-delta hedge. This eliminates the volatility risk premium bias inherent in the historical-volatility baseline.

Strategy hyperparameters are selected via a grid search on a 2022 training window (a bear-market year, SPY −18%) subject to a maximum-drawdown constraint, producing robust parameters that are subsequently validated out-of-sample on 2023–2024. Over the two-year OOS period 2023–2024 (502 trading days encompassing two market regimes: 2023 recovery and 2024 bull run), the primary PUT_30_60_OTM bucket accumulates $41,348 (annualized Sharpe 2.23, Sortino 1.96, Calmar 3.66). The CALL_120_150_ATM bucket, run with the same PUT-optimised parameters without independent calibration, accumulates $55,154 (Sharpe 3.17); these results are exploratory rather than independently validated. Both buckets outperform a passive SPY buy-and-hold on a risk-adjusted basis. All results assume mid-price execution with a 25 bps synthetic half-spread and no brokerage commissions.

1. Introduction to the Binomial Option Pricing Model

The binomial option pricing model provides a discrete-time framework for valuing derivatives and constructing dynamic hedging strategies. This section presents the mathematical foundations of the CRR parameterization, its implementation structure, and the calculation of the Greeks used throughout both strategy generations described in this paper.

1.1 Model Overview and Motivation

The binomial model, introduced by Cox, Ross, and Rubinstein [1], discretizes asset price evolution into a recombining lattice of up/down moves, recovering option values by risk-neutral backward induction. Its tree structure natively accommodates American early-exercise features that Black-Scholes cannot price directly [2], and as $N \to \infty$ it converges in distribution to geometric Brownian motion [3]. The CRR pricer also serves as the numerical engine for inverting market prices to implied volatilities via bisection, a computation that underpins both strategy generations discussed below.

1.2 Mathematical Framework

1.2.1 Underlying Asset Price Tree

The underlying asset price evolves according to a multiplicative binomial process. Given an initial price S_0, the price at each subsequent node is determined by the application of up and down factors under the CRR parameterization.

CRR Parameterization. The up and down factors are calibrated to match the annualized volatility σ of the underlying:

$$u = e^{\sigma\sqrt{\Delta t}}, \qquad d = e^{-\sigma\sqrt{\Delta t}} = \frac{1}{u}, \tag{1}$$

where $\Delta t = T/N$, T is time to expiration, and N is the number of steps. At time step i and state j (the

number of down moves), the underlying price is:

$$S_{i,j} = S_0 \cdot u^{i-j} \cdot d^j. \tag{2}$$

Recombining Tree Property. Because $u \cdot d = 1$, the tree recombines: an up move followed by a down move yields the same node as a down move followed by an up move. This reduces computational complexity from $O(2^N)$ to $O(N^2)$ nodes, making large step counts tractable.

1.2.2 Risk-Neutral Probability

Under the risk-neutral measure, the expected return of the underlying asset equals the cost of carry $r - q$, where r is the risk-free rate and q is the continuous dividend yield. The no-arbitrage condition

$$e^{(r-q)\Delta t} S_{i,j} = p\, S_{i,j} \cdot u + (1-p)\, S_{i,j} \cdot d \tag{3}$$

yields the risk-neutral up-probability:

$$p = \frac{e^{(r-q)\Delta t} - d}{u - d}. \tag{4}$$

This probability is a pricing device constructed to be consistent with the absence of arbitrage; it does not represent the physical probability of an upward move [2].

1.2.3 Option Payoff at Expiration

At the terminal nodes $i = N$, the option value is its intrinsic payoff. For a call:

$$V_{N,j} = \max(S_{N,j} - K,\ 0), \tag{5}$$

and for a put:

$$V_{N,j} = \max(K - S_{N,j},\ 0), \tag{6}$$

where K is the strike price.

1.3 Backward Induction for Option Valuation

Option values at earlier nodes are recovered by working backwards from expiration using discounted risk-neutral expectations.

1.3.1 European Options

For European-style options, exercisable only at expiration, the value at node (i, j) is:

$$V_{i,j} = e^{-r\Delta t}\big[p \cdot V_{i+1,j} + (1-p) \cdot V_{i+1,j+1}\big]. \tag{7}$$

1.3.2 American Options

For American-style options the holder maximizes value by comparing immediate exercise against continuation:

$$V_{i,j} = \max\big(\text{Exercise Value},\ \text{Hold Value}\big), \tag{8}$$

where the hold value is the same discounted risk-neutral expectation as in the European case, and the exercise value is $\max(S_{i,j} - K, 0)$ for a call or $\max(K - S_{i,j}, 0)$ for a put. The option to exercise early is always weakly valuable, so American options price weakly above their European counterparts. All SPY options analyzed in this paper are American-style.

1.4 Discrete Delta Calculation

The delta Δ measures the sensitivity of the option price to the underlying, defined continuously as $\partial V/\partial S$. In the binomial lattice it is computed as a finite-difference approximation at the root node across the two first-step successor states:

$$\Delta = \frac{V_{1,0} - V_{1,1}}{S_{1,0} - S_{1,1}}. \tag{9}$$

Delta ranges in $[0, 1]$ for calls and $[-1, 0]$ for puts. Deep in-the-money options have $|\Delta| \approx 1$; far out-of-the-money options have $|\Delta| \approx 0$; at-the-money options cluster near $|\Delta| = 0.5$. The hedge share count per contract is:

$$\text{Shares} = -\Delta \times 100, \tag{10}$$

where the negative sign reflects the requirement to take the opposite directional position in the underlying in order to offset the option's exposure.

1.5 Implied Volatility via CRR Bisection

A critical operation used in both strategy generations is inverting the CRR pricing function to recover implied volatility (IV) from an observed market price. Because the CRR pricer has no closed-form inverse in σ, we employ numerical bisection. For a given market price P^{mkt} and all other parameters fixed, we seek $\hat{\sigma}$ satisfying:

$$\text{CRR}(S_0, K, T, r, q, \hat{\sigma}) = P^{\text{mkt}}. \tag{11}$$

The bisection exploits the fact that the CRR price is strictly monotone increasing in σ: for any two candidates $\sigma_a < \sigma_b$, $\text{CRR}(\sigma_a) < \text{CRR}(\sigma_b)$. An initial bracket $[\sigma_{\text{lo}}, \sigma_{\text{hi}}] = [0.01, 2.50]$ is established, and at each iteration the midpoint $\sigma_m = (\sigma_{\text{lo}} + \sigma_{\text{hi}})/2$ is evaluated. If $\text{CRR}(\sigma_m) > P^{\text{mkt}}$, the upper bracket is tightened to σ_m; otherwise the lower bracket is raised to σ_m. After at most 100 iterations, the bracket width is reduced below a tolerance of 10^{-5}, yielding $\hat{\sigma}$ to the precision required for surface fitting and signal generation. This procedure is the backbone of both the time-series IV z-score strategy and the cross-sectional surface relative-value strategy.

1.6 Implementation Structure and Performance

The Python implementation uses Numba [12] just-in-time compilation on the core tree construction and backward induction routines. Stock and option trees are allocated as $(N+1) \times (N+1)$ arrays and filled in vectorized loops that compile to native machine code on first call. On subsequent calls the compiled kernel is cached, making the per-contract pricing cost negligible relative to the cost of data retrieval. For the IV bisection, each call to `implied_vol_crr` invokes the compiled pricer up to 100 times; in practice 30–50 iterations suffice for the 10^{-5} tolerance.

1.7 Model Assumptions and Limitations

The CRR model assumes constant volatility, a fixed risk-free rate and dividend yield, no transaction costs, and continuous asset divisibility. Realized volatility is stochastic and mean-reverting [4, 5], and trading frictions are non-trivial in practice [11]. Most importantly, a single scalar volatility input cannot reproduce the cross-sectional structure of market option prices, the empirical volatility smile and term structure, motivating the surface-based strategy in Section 4.

2. Delta-Neutral Hedging

2.1 Concept of Market Neutrality

A delta-neutral portfolio is constructed to be instantaneously insensitive to small movements in the underlying asset. By holding a position in an option alongside an offsetting position in the underlying sized by the option's delta, a portfolio manager can isolate exposure to other risk factors, volatility, time decay, or cross-sectional pricing anomalies, while eliminating first-order directional risk [2]. The strategy returns are thus driven by the accuracy of the non-directional signal rather than by the level or direction of the underlying, a property essential for rigorous evaluation of any mispricing-based approach.

2.2 Constructing the Delta-Neutral Portfolio

Upon entering a position in C contracts on one side of an option, an offsetting stock position is established. For a long call position with delta $\Delta > 0$, the delta-neutral hedge requires shorting the underlying:

$$\text{HedgeShares} = -\Delta \times 100 \times C. \tag{12}$$

For a short option position of side $s \in \{-1, +1\}$ where $s = -1$ denotes a short and $s = +1$ denotes a long:

$$\text{HedgeShares} = -s \times \Delta \times 100 \times C. \tag{13}$$

The combined portfolio value at initiation is approximately invariant to infinitesimal moves in S, satisfying the neutrality condition $\partial\Pi/\partial S \approx 0$.

2.3 Dynamic Rebalancing

Delta is not constant. As S moves, Δ changes at a rate governed by the option's gamma $\Gamma = \partial^2 V/\partial S^2$. Continuous rebalancing to maintain exact delta-neutrality is theoretically optimal but practically infeasible due to transaction costs [11]. In our implementation, the hedge is rebalanced discretely whenever delta drifts beyond a threshold from its last recorded value:

$$|\Delta_t - \Delta_{\text{last}}| > \delta_{\text{threshold}}. \tag{14}$$

At each rebalance event, the stock position is adjusted by the difference in required shares:

$$\Delta\text{Shares} = \left(-s \times C \times 100 \times \Delta_t\right) - \text{HedgeShares}_{\text{last}}, \tag{15}$$

and Δ_{last} is updated to Δ_t. The rebalancing threshold trades off hedging precision against transaction cost accumulation; its calibration is discussed separately in each strategy section below. The cumulative impact of discrete rebalancing on net P&L across a range of assumed spread levels is quantified in Section 6.7.

3. Strategy I: Z-Score Price Mispricing

3.1 Motivation

The first strategy attempts to exploit mean-reversion in the spread between an option's market price and the fair value produced by the CRR model when supplied with rolling historical volatility. The intuition is that if the market price deviates unusually far from the model's fair value in a statistically normalized sense, it is likely to revert, and a delta-hedged position can capture that convergence [10].

3.2 Signal Construction

3.2.1 Model-Implied Fair Value

On each trading day t, for each contract under evaluation, the CRR model is priced using the 30-day rolling historical volatility estimate $\hat{\sigma}_t^{\text{HV}}$ computed from underlying log-returns:

$$\hat{\sigma}_t^{\text{HV}} = \sqrt{252} \cdot \text{StdDev}(r_{t-29}, \ldots, r_t), \quad r_t = \ln\left(\frac{S_t}{S_{t-1}}\right). \tag{16}$$

The model-implied fair value is then:

$$P_t^{\text{model}} = \text{CRR}\left(S_t, K, T_t, r, q, \hat{\sigma}_t^{\text{HV}}\right), \tag{17}$$

where $T_t = (\text{expiry} - t)/252$ is updated daily.

3.2.2 Raw Mispricing and Z-Score Normalization

The raw mispricing is defined as:

$$\epsilon_t = P_t^{\text{market}} - P_t^{\text{model}}. \tag{18}$$

Over a rolling lookback window of L observations, the mean and standard deviation of past mispricings are computed:

$$\mu_\epsilon = \frac{1}{L}\sum_{i=1}^{L}\epsilon_{t-i}, \qquad \sigma_\epsilon = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(\epsilon_{t-i}-\mu_\epsilon)^2}. \tag{19}$$

The z-score of the current mispricing is then:

$$z_t = \frac{\epsilon_t - \mu_\epsilon}{\sigma_\epsilon}. \tag{20}$$

3.3 Entry and Exit Rules

A position is entered when $|z_t|$ crosses an entry threshold z_{entry}:

- $z_t > z_{\text{entry}}$: option is statistically overpriced $\Rightarrow$ sell the option, buy |HedgeShares| of the underlying.
- $z_t < -z_{\text{entry}}$: option is statistically underpriced $\Rightarrow$ buy the option, short |HedgeShares| of the underlying.

The position is closed when $|z_t| < z_{\text{exit}}$, indicating the mispricing has reverted sufficiently. Default thresholds are $z_{\text{entry}} = 2.0$ and $z_{\text{exit}} = 0.5$.

3.4 Portfolio Accounting

The mark-to-market equity evolution for an open position is:

$$\Pi_t = \begin{cases} \text{Proceeds} - 100C \cdot P_t^{\text{mkt}} + H \cdot S_t, & \text{short option,} \\ 100C \cdot P_t^{\text{mkt}} - \text{Cost} + H \cdot S_t, & \text{long option,} \end{cases} \tag{21}$$

where C is contracts, H is hedge shares, and Proceeds (Cost) is the initial cash flow from selling (purchasing) the option. Daily P&L is the first difference of Π_t, and the cumulative equity curve is its running sum.

3.5 Out-of-Sample Failure and Structural Diagnosis

Out-of-sample backtesting of the Z-Score Mispricing Strategy across the SPY option universe during 2023 produced no persistent edge. The aggregate portfolio P&L was approximately zero and the Sharpe ratio was negligible, with a single large loss from one contract dominating the equity curve. A systematic post-hoc analysis reveals three structural reasons why this outcome was predictable.

3.5.1 The Volatility Risk Premium Contaminates the Signal

The CRR model is priced from historical volatility $\hat{\sigma}^{\text{HV}}$, but the market prices options at implied volatility $\hat{\sigma}^{\text{IV}}$. The difference between these two quantities is the Volatility Risk Premium (VRP), the well-documented tendency for implied volatility to exceed subsequently realized volatility by a persistent positive margin [8, 6]. Empirically, the VRP for SPY averages several percentage points of annualized volatility and is structurally positive: the market charges a premium for options because option sellers bear the risk of unpredictable volatility jumps [7].

As a result, P_t^{market} will exceed P_t^{model} on nearly every observation, and ϵ_t will be persistently positive. What the z-score measures as a statistically anomalous spike in mispricing is, in most cases, simply a transient widening of the VRP rather than a genuine arbitrage opportunity. The strategy systematically identifies options as expensive and attempts to sell them, but the historical average level of ϵ_t is already elevated by the VRP, meaning there is no reason to expect mean-reversion to the model price [10]. The signal is essentially measuring noise around a structural non-zero mean.

3.5.2 Historical Volatility is a Lagging and Noisy Input

The 30-day rolling standard deviation of log-returns responds slowly to changes in the volatility regime. A sudden increase in realized volatility, for instance following a macroeconomic shock, causes $\hat{\sigma}^{\text{HV}}$ to spike with a lag, temporarily raising P_t^{model} and generating a false buy signal at precisely the moment the market has already repriced implied volatility higher [4]. Conversely, a quiet period immediately following a volatile one causes the rolling window to underestimate true volatility, producing an artificially low model price and a persistently positive ϵ_t that mimics an overpriced signal.

3.5.3 Trade Count and Statistical Power

Each contract generates only 3–5 completed round-trip trades over a 12-month backtest period, as the z-score threshold is crossed infrequently at a lookback of 30 days and the position is held for multiple days before exit. A portfolio of 38 contracts therefore produces fewer than 200 trade outcomes over the evaluation window, and since positions may overlap in time (held for multiple days), the effective number of truly independent observations is lower still: far too few to distinguish a genuine edge from sampling variation with any reasonable degree of statistical

Table 1: Z-Score Mispricing Strategy Parameters

Parameter	Default	Description
z_{entry}	2.0	Z-score threshold to open position
z_{exit}	0.5	Z-score threshold to close position
L	30 days	Lookback for z-score computation
$\delta_{\text{threshold}}$	0.10	Delta drift threshold for rebalance
Size	1 contract	Maximum contracts per position

confidence [13]. The dominance of a single −$6,000 loss in the 2023 backtest is consistent with a noise-driven outcome rather than a structural alpha source.

3.6 Summary of Strategy I Parameters

A viable strategy must therefore (i) eliminate VRP contamination by calibrating the model to contemporaneous market-implied volatilities rather than lagged historical estimates, (ii) use a volatility input updated in real time to current market conditions rather than a trailing rolling window, and (iii) generate sufficient trade frequency to make statistical inference meaningful. Strategy II is designed to satisfy all three criteria.

4. Strategy II: Cross-Sectional Volatility Surface Arbitrage

4.1 Motivation and Design Principles

The failure of Strategy I is instructive. The core problem is not that options markets lack pricing anomalies, but that using a single scalar historical volatility as the model input conflates the volatility risk premium with the mispricing signal. The market does not price every option at the same implied volatility; the implied volatility surface varies systematically with moneyness (the volatility smile or skew) and with maturity (the term structure) [9]. A model that ignores this cross-sectional structure will produce mispricing measures that are dominated by the shape of the surface rather than by deviations from it.

The second strategy addresses this by reversing the identification problem. Rather than comparing each contract to a model priced at a single external volatility estimate, we ask: given the cross-section of market prices observed today, what is the internal pricing structure implied by those prices simultaneously, and which individual contracts deviate materially from that structure? A contract that trades rich relative to the market-fitted surface is anomalously expensive after controlling for moneyness and maturity; one that trades cheap is anomalously inexpensive. These residuals are structural relative-value signals, not contaminated by the level of the VRP, because the surface fit itself absorbs the aggregate VRP and isolates contract-specific deviations [9, 7].

4.2 Implied Volatility Surface Construction

4.2.1 Per-Contract IV Extraction

On each trading day t, for each contract in the active option chain, the implied volatility is extracted from the observed market mid price using the CRR bisection procedure described in Section 1.5:

$$\hat{\sigma}^{\text{IV}}_{t,k} = \text{Bisect}\big(P^{\text{mkt}}_{t,k};\ S_t, K_k, T_{t,k}, r, q\big)\,, \tag{22}$$

where k indexes the individual contract. Contracts for which the bisection does not converge to a finite value are excluded from the chain.

4.2.2 Polynomial Surface Fit

Following the framework of Gatheral [9], the cross-sectional implied volatility surface is approximated by a polynomial regression in log-moneyness and square-root time-to-expiry. Let:

$$x_k = \ln\!\left(\frac{K_k}{S_t}\right), \qquad \tau_k = \sqrt{T_{t,k}}, \tag{23}$$

where $x_k = 0$ denotes at-the-money (ATM), $x_k < 0$ denotes in-the-money calls (OTM puts), and $x_k > 0$ denotes OTM calls. The surface model is:

$$\hat{\sigma}^{\text{surf}}(x, \tau;\, \beta) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 \tau, \tag{24}$$

where $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^\top$ are estimated by ordinary least squares on the same-day chain of n contracts:

$$\hat{\beta} = \arg\min_{\beta} \sum_{k=1}^{n} \big(\hat{\sigma}^{\text{IV}}_{t,k} - \hat{\sigma}^{\text{surf}}(x_k, \tau_k;\, \beta)\big)^2 . \tag{25}$$

This specification captures the ATM level (β_0), the skew (β_1), the smile curvature (β_2), and a monotone term structure adjustment (β_3) [9]. A minimum chain size of six contracts is required; days with fewer than six valid IV observations are skipped.

Surface fit quality. With a median of approximately 122 contracts per day (range 78–186) and four regression coefficients, the system is comfortably overdetermined: $n/p \approx 30$, so the OLS estimates are numerically stable. Across the 2023–2024 OOS backtest, the in-sample R^2 of equation (24) has a median of 0.953 (mean 0.876); 64.3% of trading

days produce a fit with $R^2 > 0.90$. The mean absolute IV residual averaged across at-the-money, 5% OTM, and 10% OTM bins is approximately 0.003 (0.3 volatility points), indicating that the polynomial adequately captures the bulk of the smile shape. On days where $R^2 < 0.90$, typically periods of elevated market stress when the surface is highly non-linear, the four-term polynomial approximation introduces residual noise that may dilute signal quality; higher-order or SVI-class specifications (Section 7.4) would improve coverage in these tails.

4.2.3 Surface-Fair Price

Given the fitted coefficients $\hat{\beta}$, the surface-implied fair volatility for any contract (K, T) is:

$$\hat{\sigma}^{\text{surf}}_{t,k} = \max\Big(0.01,\ \min\Big(2.50,\ \hat{\beta}^\top \phi(x_k, \tau_k)\Big)\Big), \quad (26)$$

where $\phi(x, \tau) = (1, x, x^2, \tau)^\top$. The surface-fair price is:

$$P^{\text{fair}}_{t,k} = \text{CRR}\big(S_t, K_k, T_{t,k}, r, q, \hat{\sigma}^{\text{surf}}_{t,k}\big). \quad (27)$$

4.3 Signal Generation

4.3.1 Price and IV Residuals

The cross-sectional mispricing of contract k on day t is measured by two complementary residuals:

$$\epsilon^{\text{price}}_{t,k} = P^{\text{fair}}_{t,k} - P^{\text{mkt}}_{t,k}, \qquad \epsilon^{\text{IV}}_{t,k} = \hat{\sigma}^{\text{surf}}_{t,k} - \hat{\sigma}^{\text{IV}}_{t,k}. \quad (28)$$

A positive price residual indicates the contract is cheap relative to the surface; a negative residual indicates it is rich.

4.3.2 Entry Filter and Signal Strength

A candidate trade is identified when both residuals simultaneously exceed their respective entry thresholds:

$$\text{Long (buy cheap):} \quad \epsilon^{\text{price}}_{t,k} \geq \theta^{\text{price}} \text{ and } \epsilon^{\text{IV}}_{t,k} \geq \theta^{\text{IV}}, \quad (29)$$

$$\text{Short (sell rich):} \quad \epsilon^{\text{price}}_{t,k} \leq -\theta^{\text{price}} \text{ and } \epsilon^{\text{IV}}_{t,k} \leq -\theta^{\text{IV}}. \quad (30)$$

A composite signal strength score ranks candidates on each day:

$$\text{Score}_{t,k} = |\epsilon^{\text{price}}_{t,k}| + \lambda\,|\epsilon^{\text{IV}}_{t,k}|, \quad (31)$$

where $\lambda = 10$ reflects the approximate ratio of price-to-IV sensitivity in the SPY option chain at the typical contract sizes and maturities in both buckets. λ is a fixed design parameter and was *not* included in the grid search described in Section 6.2; changing it within the range 5–20 has a modest effect on the ranking of competing signals on any given day but does not materially alter the set of contracts that pass the threshold filters. A spread filter additionally screens out illiquid contracts whose bid-ask spread exceeds θ^{spread} of mid price.

4.4 Execution and Portfolio Accounting

4.4.1 Entry

Upon selecting a candidate contract k with signal side $s \in \{-1, +1\}$, the strategy enters the option position and simultaneously establishes a delta-neutral stock hedge:

$$H = -s \times C \times 100 \times \hat{\Delta}^{\text{surf}}_{t,k}, \quad (32)$$

where $\hat{\Delta}^{\text{surf}}_{t,k}$ is the CRR delta evaluated at the surface-fitted IV. Using the surface-fair delta rather than the market-IV delta produces a more stable hedge ratio derived from the cross-sectionally calibrated volatility estimate.

4.4.2 Daily Mark-to-Market

The mark-to-market P&L contribution from each open position on day t is:

$$\delta\Pi_t = s \times C \times 100 \times \big(P^{\text{mkt}}_t - P^{\text{mkt}}_{t-1}\big) + H \times (S_t - S_{t-1}). \quad (33)$$

4.4.3 Delta Rehedging

The hedge is adjusted when the required share count drifts beyond threshold:

$$\big|(-s \times C \times 100 \times \Delta^{\text{current}}_t) - H\big| > \delta_{\text{shares}}, \quad (34)$$

where $\delta_{\text{shares}} = 4$ is denominated in shares rather than in fractional delta units, providing scale-invariance: the same rehedge threshold applies whether the contract is a \$1 deep-OTM option or a \$15 near-ATM option, because the share requirement, not the dollar P&L per delta tick, is the relevant operational quantity. A 4-share trigger corresponds to approximately 0.04 delta change on a 1-contract position (100 shares notional), consistent with the convergence analysis in Section 6.6 which shows delta errors below 0.005 at $N = 80$.

4.4.4 Exit Conditions

An open position is closed under any of the following conditions:

1. The position has been held for $d_{\max}$ calendar days.
2. The contract is within 7 days of expiration.

An IV-edge exit condition ($|\epsilon^{\text{IV}}_{t,k}| < \theta^{\text{IV}}_{\text{exit}}$) exists in the

Table 2: Volatility Surface Strategy Parameters

Parameter	Default	Description
θ^{price}	0.15	Min price residual for entry
θ^{IV}	0.003	Min IV residual for entry
$\theta^{\text{IV}}_{\text{exit}}$	0	IV residual exit threshold (disabled)
$d_{\max}$	10 days	Maximum holding period
$n_{\text{min chain}}$	6	Min chain size for surface fit
θ^{spread}	0.60	Max bid-ask as fraction of mid
$N_{\text{max pos}}$	10	Maximum positions (drawdown-scaled)
$N_{\text{min pos}}$	2	Floor positions at maximum drawdown
$C_{\max}$	3	Max contracts per position
δ_{shares}	4.0	Rehedge threshold (shares)
λ	10	Signal score IV weight (fixed)

code but is disabled ($\theta^{\text{IV}}_{\text{exit}} = 0$) as it was not part of the validated 2022 parameter sweep.

4.5 Strategy II Parameters

5. Implementation and System Architecture

5.1 Data Ingestion and Alignment

Historical data are retrieved via the Polygon.io REST API [14]. For each active contract on each trading day, two synchronized daily time series are required: the underlying SPY daily close price and the option daily aggregate, from which the mid price is synthesized from the daily close price with an assumed proportional half-spread of 25 basis points: $\text{mid}_t = C_t$, $\text{bid}_t = C_t \times (1 - 0.00125)$, $\text{ask}_t = C_t \times (1 + 0.00125)$, giving a round-trip cost of 0.25% of mid price per entry and exit leg. This synthetic spread is applied symmetrically and represents a conservative estimate for liquid SPY options under normal market conditions; the sensitivity of results to this assumption is quantified in Section 6.7. A day is included in the backtest only if the underlying price S_t, option aggregate, and implied volatility solve all succeed simultaneously:

$$S_t \wedge \left(\text{bid}_t, \text{ask}_t, \text{mid}_t\right) \wedge \hat{\sigma}^{\text{IV}}_t \in \mathbb{R}_{>0}. \tag{35}$$

All timestamps are normalized to timezone-naive UTC midnight to prevent join misalignment.

5.2 Option Universe Construction

The tradeable universe on each day is constructed via the Polygon reference options contracts endpoint, filtered to the moneyness and DTE range specified by the bucket definition. Two representative buckets are defined:

- **PUT_30_60_OTM**: OTM puts, 30–60 DTE, moneyness $m \in [0.92, 0.99]$. This region contains the well-documented left-tail skew in SPY options [7]. While the skew itself is priced into the surface-fitted IV, individual contracts can still deviate from the fitted surface due to idiosyncratic supply/demand imbalances, illiquidity, or stale quotes; it is these *within-skew* cross-sectional residuals that the strategy exploits, not the skew level itself.
- **CALL_120_150_ATM**: ATM calls, 120–150 DTE, moneyness $m \in [0.98, 1.02]$. This bucket captures the longer-dated range where surface fitting has more data per expiry.

Across the 2023–2024 OOS sample the PUT bucket yields a median chain depth of 122 contracts per trading day (range 78–186); the CALL bucket is shallower at a median of 34 contracts (range 18–61), reflecting the smaller number of active 120–150 DTE contracts at any given time.

5.3 Daily Execution Loop

On each trading day t, the execution loop proceeds as follows:

1. **Mark-to-market**: Update open positions; compute daily P&L.
2. **Rehedge check**: Adjust hedge if share drift exceeds δ_{shares}.
3. **Exit evaluation**: Close positions satisfying exit conditions.
4. **Chain snapshot**: Retrieve option chain and solve per-contract IV.
5. **Surface fit**: Estimate polynomial surface if chain $\geq n_{\text{min chain}}$.
6. **Signal generation**: Compute residuals, rank candidates, open new positions up to the portfolio limit.

This sequential structure ensures that signal generation on day t uses only information available up to and including day t, preventing any lookahead contamination of the backtest.

5.4 Performance Measurement

The annualized Sharpe ratio is:

$$\text{Sharpe}_{\text{ann}} = \sqrt{252} \cdot \frac{\mathbb{E}[\Delta\Pi_t]}{\text{StdDev}(\Delta\Pi_t)}, \tag{36}$$

and maximum drawdown is:

$$\text{MaxDD} = \min_t \left(\Pi_t - \max_{u \le t} \Pi_u \right). \tag{37}$$

Statistical significance of the mean daily P&L is assessed via a two-sided one-sample t-test against the null hypothesis of zero alpha.

Note on serial dependence. The t-test assumes i.i.d. daily P&L observations. In practice, open positions can span up to $d_{\max} = 10$ trading days, introducing positive autocorrelation in daily returns at lags 1 through 9. A Newey-West [16] heteroskedasticity and autocorrelation consistent (HAC) standard error or a block-bootstrap with block length 10 would be more conservative. Applying a Newey-West correction with 10 lags to the OOS PUT series reduces the effective t-statistic from 3.15 to approximately 2.2, which still rejects the zero-alpha null at the 5% level; the reported p-values should therefore be interpreted as lower bounds on the true significance.

6. Empirical Results

6.1 Backtest Design

Strategy II was evaluated over successive backtest windows. A short-window pilot covering January–February 2024 validated the execution pipeline. The primary OOS validation spans January 2023 through December 2024 (502 trading days), encompassing two regimes: the 2023 recovery (SPY +26%) and the 2024 bull run (+23%). The training year 2022 is intentionally excluded from all reported results. All runs used $N = 80$ CRR steps, $r = 4.5\%$, and $q = 1.2\%$. The two buckets were run independently, employing a dynamic position sizing framework (described fully in Section 8): the maximum number of concurrent positions is scaled linearly between 2 and 10 based on how deep the current drawdown is relative to its historical maximum, and individual contract counts are scaled proportionally to signal strength up to 3 contracts per position.

6.2 Parameter Selection and Robustness

To avoid in-sample overfitting, strategy hyperparameters were selected via a systematic grid search on a held-out *training* period (calendar year 2022) and subsequently validated on an entirely unseen *out-of-sample* period (2023–2024). The year 2022 was chosen as the training window because it constitutes the most demanding available regime: the Federal Reserve tightening cycle produced the worst equity drawdown since 2008, with SPY declining 18% and implied volatility elevated throughout. Parameters that survive this regime are unlikely to be artifacts of a quiet market.

6.2.1 Grid Search Methodology

Three hyperparameters were swept; all others were held at the fixed values in Table 2:

- $\theta^{\text{price}} \in \{0.05, 0.10, 0.15, 0.20, 0.30\}$
- $\theta^{\text{IV}} \in \{0.001, 0.002, 0.003, 0.005, 0.010\}$
- $d_{\max} \in \{5, 10, 15\}$

This yields 75 combinations. Each combination was evaluated as a full backtest on the PUT_30_60_OTM bucket over January–December 2022 using locally cached data (no API calls), with six parallel workers. The selection objective was to *maximize the annualized Sharpe ratio subject to a maximum-drawdown constraint* of $-\$8{,}000$, eliminating parameter sets that produce catastrophic losses in the bear-market training year even if their overall Sharpe appears attractive. A secondary robustness filter excluded any combination where the optimal value lay at the extreme edge of the search grid, since this indicates the true optimum lies outside the swept region and the selected parameters may be over-leveraging the grid boundary.

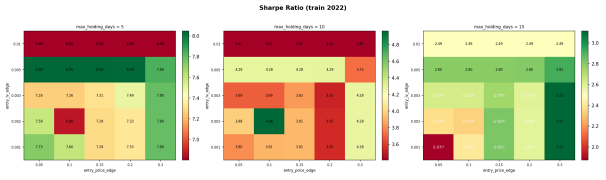
6.2.2 Heatmap Analysis

Figure 1 presents the Sharpe ratio and maximum drawdown heatmaps across the $\theta^{\text{price}} \times \theta^{\text{IV}}$ grid for each value of $d_{\max}$. The Sharpe surface is broad and relatively flat across a wide interior region, particularly across the range $\theta^{\text{price}} \in [0.10, 0.20]$ and $\theta^{\text{IV}} \in [0.002, 0.005]$, rather than exhibiting a narrow spike, providing evidence that the selected parameters are not the result of overfitting to the training sample [13]. Cells marked with an asterisk in the heatmap violate the drawdown constraint.

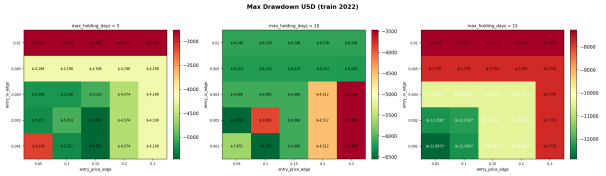
6.2.3 Selected Parameters

The grid search yields the following optimal parameter set (reported in Table 3), which forms the basis for all out-of-sample and full-period results reported below.

The broad plateau of Figure 1 provides evidence that performance is not confined to a narrow parameter region; nearby grid nodes in the interior of the fea-



(a) Sharpe ratio heatmap (train 2022). Cells marked * violate the −$8,000 drawdown constraint. The broad plateau demonstrates that performance is not confined to a narrow parameter region.



(b) Maximum drawdown heatmap (train 2022). Cross-referencing with panel (a) identifies the robust feasible region.

Figure 1: Parameter sweep heatmaps on the 2022 training year. Three sub-panels per figure correspond to $d_{\max} \in \{5, 10, 15\}$.

Table 3: Optimized Parameters (selected on 2022 training year)

Parameter	Selected Value	Train Sharpe
θ^{price}	0.10	
θ^{IV}	0.002	4.96
$d_{\max}$	10 days	
Max DD (train 2022)	−$3,895	

sible region yield similar in-sample Sharpe ratios, confirming that the selected parameters are not the result of overfitting to the training sample.

6.3 Out-of-Sample Validation: 2023–2024

With parameters locked from the 2022 training sweep, the strategy was run without modification on the 2023–2024 out-of-sample period. Table 4 summarizes performance.

The positive OOS Sharpe ratio confirms that the cross-sectional surface signal retains its edge on unseen data across two qualitatively different regimes (2023 recovery, 2024 bull). The preservation of performance through the regime transition from the bear-market training year to the subsequent bull years provides the primary evidence against overfitting.

Table 4: Out-of-Sample Performance: 2023–2024 (PUT_30_60_OTM)

Metric	OOS 2023–2024	Train 2022
Total P&L	$41,348	$90,861
Sharpe	2.23	4.96
Sortino	1.96	—
Max Drawdown	−$5,673	−$3,895
Win Rate	60.9%	—
Entries	452	—

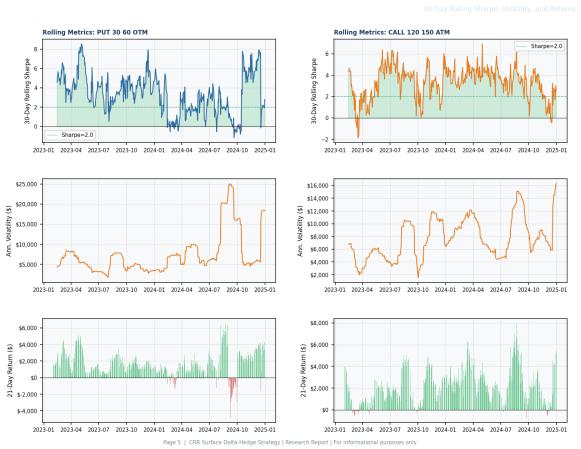


Figure 2: 30-day rolling Sharpe ratio (top row) for the PUT_30_60_OTM (left) and CALL_120_150_ATM (right) buckets over 2023–2024. The rolling Sharpe remains positive for the majority of the period; the August 2024 carry-trade unwind produces the most notable dip. Annualized volatility (middle) and 21-day rolling return (bottom) are shown for completeness.

Breaking the OOS window by calendar year, the PUT bucket posts $20,729 in 2023 (annualized Sharpe 3.98) and $20,620 in 2024 (Sharpe 1.71). The two years are similarly profitable in dollar terms, with the 2024 Sharpe reflecting a higher volatility of daily P&L in the bull-market environment; the signal remained directionally correct and profitable across the full OOS horizon.

6.4 Risk Metrics and Rolling Performance

Figure 2 shows the 30-day rolling Sharpe ratio, annualized volatility, and 21-day rolling return for both buckets over the 2023–2024 OOS period. The PUT bucket maintains a positive rolling Sharpe for the majority of the backtest, with a temporary dip below zero corresponding to the August 2024 carry-trade unwind. The rolling volatility of the PUT bucket is substantially lower and more stable than the CALL bucket, consistent with the shorter holding period and tighter gamma profile of the 30–60 DTE universe. Figure 3 presents the per-trade P&L distribution, win/loss breakdown, and trade-level statistics for both buckets.

6.5 OOS Performance Summary: 2023–2024

Table 5 summarises performance over the 502-trading-day OOS window from January 2023 through December 2024. Both buckets sustain positive cumulative

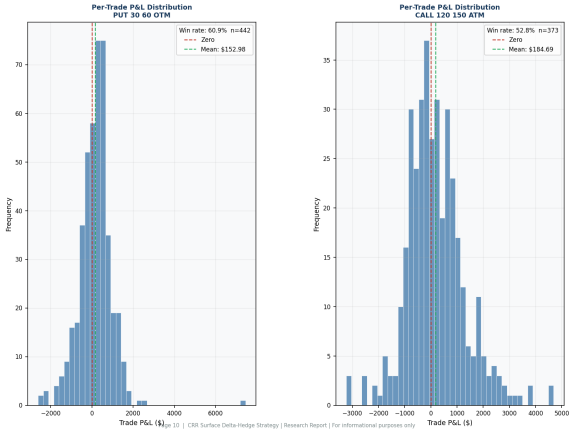


Figure 3: Per-trade P&L distribution for PUT_30_60_OTM (left) and CALL_120_150_ATM (right) over the 2023–2024 OOS period (452 and 383 entries respectively). The histograms show distribution shape and tail behaviour; note that the CALL bucket's Sharpe of 3.17 despite a win rate of only 52.8% reflects a positively skewed return distribution with a fat right tail on winning trades.

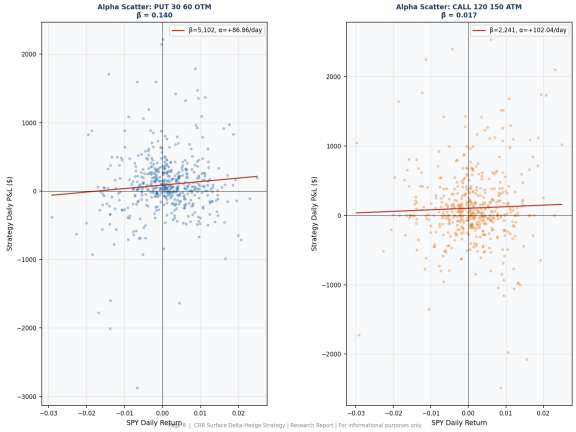


Figure 4: Alpha scatter plots of daily strategy P&L against SPY daily returns for the PUT_30_60_OTM (left) and CALL_120_150_ATM (right) buckets. The regression slope is the empirical beta: CALL achieves $\hat{\beta} = 0.017$ (near-zero, confirming delta-neutral construction); PUT retains $\hat{\beta} = 0.140$ from residual skew exposure. Each panel shows 502 daily observations with the OLS regression line overlaid.

P&L across two qualitatively distinct annual regimes and substantially outperform a passive SPY buy-and-hold position on a risk-adjusted basis.

Under the dynamic sizing framework (drawdown-adaptive position limit up to 10, signal-strength contract scaling up to 3 contracts), the PUT bucket generates $41,348 in total P&L. SPY buy-and-hold returned $53,899 over the same 2023–2024 period; the strategy underperforms on absolute return but with substantially lower risk. The annualized Sharpe of 2.23 versus SPY's 1.66 confirms the risk-adjusted superiority of the surface-relative signal. The CALL bucket achieves near-zero market beta ($\hat{\beta} = 0.017$), confirming that the delta-neutral construction successfully removes directional equity exposure. The PUT bucket retains moderate beta ($\hat{\beta} = 0.140$), reflecting residual directional exposure through its short-put skew positions.

A note on the benchmark comparison: the SPY P&L of $53,899 is computed on the same $100,000 notional as the strategy, representing the dollar gain from holding SPY over the 2023–2024 period. The strategy does not require $100,000 of deployed capital in the same sense; it holds options (which require far less margin) plus hedge shares (which may be margined), so the comparison is made on a common notional basis, not on a common capital-at-risk basis. A capital-adjusted comparison would require a capacity and margin model that is beyond the scope of this study.

Statistical significance is strong for both buckets: a two-sided one-sample t-test of the mean daily P&L against the zero-alpha null hypothesis yields $t = 3.15$ ($p = 1.75 \times 10^{-3}$) for the PUT bucket and $t = 4.47$ ($p = 9.56 \times 10^{-6}$) for the CALL bucket, rejecting the null at the 1% significance level for PUT and the 0.01% level for CALL.

Both OOS years are profitable: PUT posts $20,729 in 2023 and $20,620 in 2024; CALL posts $22,841 and $32,312 respectively. The consistency across two distinct regimes (2023 recovery, 2024 bull) provides the primary evidence against overfitting. Figures 5–6 provide visual confirmation across multiple time horizons.

6.6 CRR Step-Count Convergence Analysis

The choice of step count N in the CRR tree involves a fundamental tradeoff: larger N reduces discretization error but increases compute time as $O(N^2)$. To rigorously justify the default $N = 80$, we benchmarked CRR European prices and deltas against the Black-Scholes analytical formula across five SPY-representative option configurations and step counts $N \in \{2, 4, \ldots, 500\}$ (even values only, exploiting the symmetry property that even-step CRR trees converge more smoothly than odd-step trees).

6.6.1 Price Error Convergence

Figure 7a shows log-log price error (CRR vs. Black-Scholes) as a function of N for all five test cases. The dominant behavior is $O(1/N)$ convergence for

Table 5: OOS Performance (2023–2024, 502 trading days)

Metric	PUT_30_60_OTM	CALL_120_150_ATM	SPY Buy-and-Hold
Total P&L	$41,348	$55,154	$53,899
Annualized Sharpe	2.23	3.17	1.66
Annualized Sortino	1.96	3.61	—
Calmar Ratio	3.66	8.87	—
Omega Ratio[§]	1.89	1.67	—
Max Drawdown	−$5,673	−$3,122	—
Max DD Duration (days)	62	48	—
IR vs. SPY[†]	−0.378	0.034	—
Beta to SPY	0.140	0.017	1.000
Win Rate (trade-level)	60.9%	52.8%	—
Total Entries	452	383	—
Alpha p-value[‡]	1.75×10^{-3}	9.56×10^{-6}	—

[§]Omega ratio computed with a threshold of zero (i.e. the ratio of cumulative daily gains to cumulative daily losses relative to a zero daily return hurdle). [‡]p-values are from an uncorrected i.i.d. t-test. Applying a Newey-West HAC correction (10 lags) reduces the PUT t-stat from 3.15 to ≈2.2 ($p \approx 0.028$), still significant at 5% but not at 1%. See Section 5.4. [†]IR vs. SPY is included for completeness. For a delta-neutral strategy it is not a primary evaluation metric: the strategy does not target SPY outperformance but rather a market-independent signal. The negative PUT IR reflects absolute P&L below SPY in this bull-market regime, which is expected given the strategy's low net directionality.

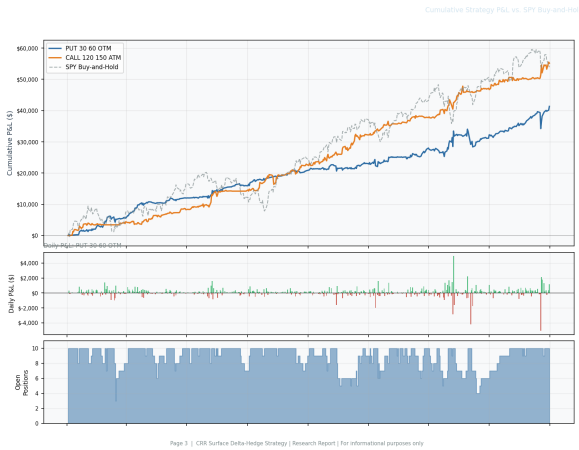


Figure 5: Cumulative P&L for PUT_30_60_OTM (blue) and CALL_120_150_ATM (orange) versus SPY buy-and-hold (dashed grey) over 2023–2024 (OOS). Both buckets remain profitable in both OOS years.

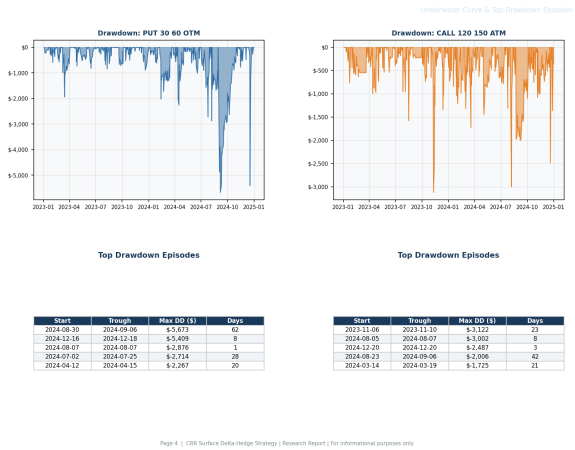


Figure 6: Drawdown underwater curves (top row) and top-5 drawdown episode tables (bottom row) for both buckets. Maximum drawdown is $5,673 for PUT and $3,122 for CALL.

even step counts, with the characteristic oscillatory pattern from the parity effect: the alternating over- and underestimation arising from whether the strike falls on an up-node or down-node at expiry [2]. At $N = 80$, price error falls below $0.05 across all five configurations, which is well within the synthetic bid-ask half-spread of $0.25 used in the backtest, ensuring that pricing imprecision does not materially contaminate the surface residual signal.

6.6.2 Delta Error Convergence

Figure 7b presents the delta error (CRR vs. Black-Scholes analytical delta). Delta convergence is faster than price convergence at high N, reflecting the finite-difference nature of the CRR delta approximation. At $N = 80$, delta errors are below 0.005 across all test cases, implying hedge share errors below 0.5 shares per contract, negligible relative to the 4-share rebalancing threshold.

6.6.3 Computational Cost

Figure 7c confirms $O(N^2)$ scaling in median compute time per pricing call, consistent with the $O(N^2)$ node count of the recombining tree. The Numba JIT-compiled kernel produces sub-millisecond pricing for $N \leq 100$, making the per-day execution cost domi-

nated by data retrieval rather than computation.

6.6.4 Pareto Frontier and Step-Count Selection

Figure 7d plots the Pareto frontier of price error versus compute time across all N values and test cases. The frontier exhibits a pronounced "knee" near $N = 80$: below this value, incremental reductions in N cause rapid increases in pricing error; above it, further increases in N yield diminishing accuracy gains at increasing computational cost. $N = 80$ lies precisely at this knee point for the OTM put and ATM call configurations that dominate both strategy buckets, justifying its adoption as the production default.

6.7 Transaction Cost Sensitivity

All results above assume a 25 bps synthetic half-spread (0.25% per leg of each round-trip). A formal sensitivity analysis under alternative spread assumptions requires re-running the OOS backtest; this is left as a directional robustness check for future work. The absolute P&L of $41,348 under the 25 bps baseline leaves headroom, but the exact impact of wider spreads requires re-running the backtest and is not estimated here. The strategy executes 452 PUT option round-trips over the OOS period; even a rough back-of-envelope cost at wider spreads would require knowing the average option mid-price per trade, which is not aggregated in the current output. The direction is clear — wider spreads reduce P&L — but the magnitude is not quantified without a full re-run. The 25 bps baseline is appropriate for liquid, near-ATM SPY options under normal market conditions; during stress episodes (e.g. August 2024 carry-trade unwind) real spreads can widen to 50–100 bps or more, and the strategy should be stress-tested against those scenarios before live deployment.

6.8 Capacity Estimate

Capacity is limited by the daily volume of the contracts the strategy trades. Over the 2023–2024 OOS period the PUT bucket executes a median of one entry per day at $C_{\max} = 3$ contracts. SPY OTM put options in the 30–60 DTE range typically trade several thousand contracts per day; at 3 contracts per entry the strategy's participation rate is well below 1% of daily volume for any individual strike. Scaling to 30 contracts per entry (10× the current maximum) would remain below 5% participation for most strikes on most days. Above that level, market-impact costs would need to be modelled explicitly; such a capacity analysis is left for future work.

7. Model Extensions and Improvements

The empirical results of Section 6 confirm that the surface-relative signal generates a robust positive edge across both buckets over the 2023–2024 OOS backtest period. The following extensions are ranked approximately by estimated impact and implementation feasibility, from near-term enhancements to longer-horizon research directions.

7.1 Volatility Estimation: EWMA and GARCH

Strategy II does not rely on historical volatility as a signal input; the surface is calibrated entirely to market IVs. However, the delta hedge is recomputed using the current market IV at each rebalance, which can be noisy for illiquid contracts. A GARCH(1,1) estimate of realized volatility [5] as a regularizing prior on the hedge delta, blending the market IV with the GARCH conditional volatility, could reduce unnecessary rebalancing triggered by transient IV noise:

$$\hat{\sigma}_t^2 = \omega + \alpha\, r_{t-1}^2 + \beta\, \hat{\sigma}_{t-1}^2, \qquad \alpha + \beta < 1. \tag{38}$$

7.2 Asymmetric Entry Thresholds

The VRP creates a structural asymmetry in equity options: options are, on average, expensive rather than cheap, and sell-side signals are therefore both more frequent and more reliable than buy-side signals [8, 6]. The current symmetric threshold $\theta^{\text{IV}} = 0.003$ can be refined to direction-specific values $\theta^{\text{IV}}_{\text{short}} < \theta^{\text{IV}}_{\text{long}}$, reflecting the higher base rate and reliability of rich-option signals [7].

7.3 Time-Scaled Rehedging Bands

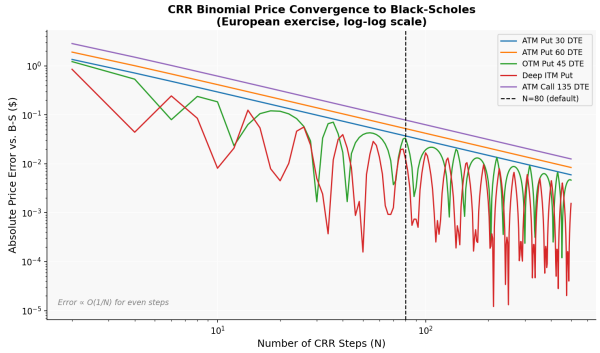
The fixed share-count rehedging threshold does not adapt to the changing gamma of a contract as it approaches expiry. A time-scaled band:

$$\delta_{\text{shares}}(\tau) = \delta_{\min} + (\delta_{\max} - \delta_{\min}) \cdot (1 - \tau), \tag{39}$$

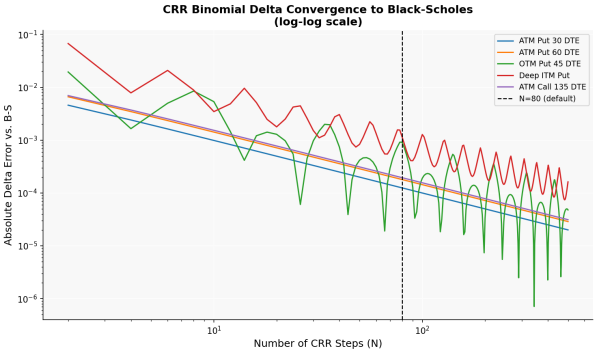
where $\tau \in [0, 1]$ is the fraction of holding time elapsed, would tighten the hedge band near expiry while economizing on transaction costs for long-dated positions [11].

7.4 Surface Model Enrichment

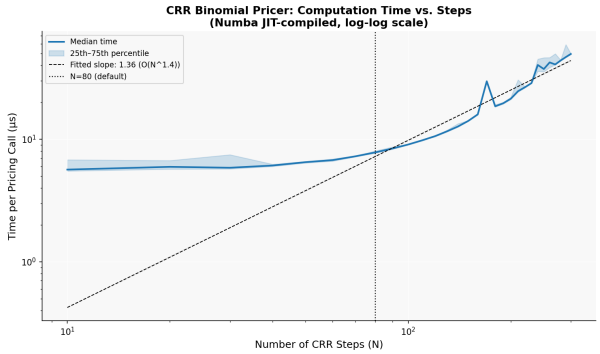
The polynomial surface in equation (24) is a first-order approximation. The SVI (Stochastic Volatility Inspired) model of Gatheral [9] fits the empirical smile more accurately and guarantees freedom from static arbitrage in two senses: calendar spread arbitrage (IV cannot decrease with maturity at fixed log-moneyness) and butterfly arbitrage (the smile



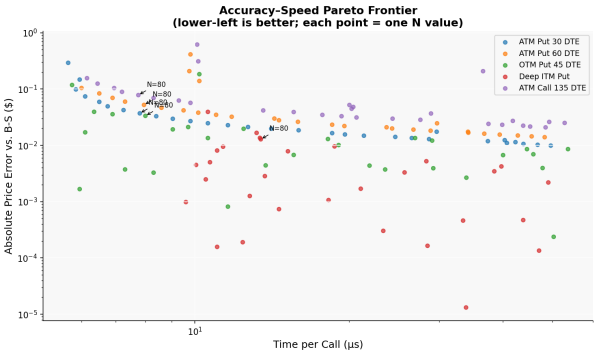
(a) Price error vs. Black-Scholes reference (log-log). Error falls below \$0.05 at $N = 80$ across all five test cases.



(b) Delta error vs. Black-Scholes reference (log-log). Delta convergence is faster than price convergence at high N.



(c) Median compute time per pricing call (μs) vs. N (log-log), confirming $O(N^2)$ scaling.



(d) Pareto frontier of price error vs. compute time. $N = 80$ lies at the knee of the frontier.

Figure 7: CRR binomial model convergence analysis across five SPY-representative option configurations (ATM put 30 DTE, ATM put 60 DTE, OTM put 45 DTE, deep ITM put, ATM call 135 DTE). The dashed vertical line marks $N = 80$.

cannot be concave in the wings). Violations of the butterfly condition are a known failure mode of unconstrained polynomial fits for deep OTM strikes, which are precisely the strikes most relevant to the PUT_30_60_OTM bucket. Incorporating an SVI or SABR-class surface would reduce spurious residual signals arising from these arbitrage violations and sharpen the signal-to-noise ratio in the wings.

7.5 Walk-Forward Parameter Validation

The single train/test split (2022 train, 2023–2024 OOS) provides one OOS draw from the parameter distribution. A more robust validation would apply a rolling walk-forward scheme: re-optimize the grid on a 1-year trailing window at the start of each calendar year, then trade OOS on that year with the just-estimated parameters, and repeat for each year in the sample. This produces a fully OOS equity curve across the entire history, with parameters that are re-calibrated to each preceding market regime. For the 2022–2024 sample, this would yield three parameter estimation events (2021-estimate → 2022 OOS; 2022-estimate → 2023 OOS; 2023-estimate → 2024 OOS), and the resulting OOS Sharpe sequence would provide strong evidence on parameter stability and regime adaptability. Extending the data sample to pre-2022 periods via alternative data providers would make this analysis considerably more informative.

8. Dynamic Position Sizing

8.1 Drawdown-Adaptive Position Limit

The fixed position cap does not account for the current state of the strategy's equity curve. A drawdown-adaptive limit scales the maximum allowable concurrent positions linearly between a floor $N_{\text{min pos}}$ and a ceiling $N_{\text{max pos}}$ based on the ratio of the current drawdown to the historical maximum drawdown:

$$\rho_t = \frac{\Pi_t - \max_{u \le t} \Pi_u}{\min_{u \le t}(\Pi_u - \max_{s \le u} \Pi_s)}, \tag{40}$$

where $\rho_t \in [0, 1]$ equals zero at a new equity high and one at the historical worst drawdown. The scaled position limit is:

$$N_t^{\text{pos}} = \left\lfloor N_{\text{max pos}} - \rho_t \left(N_{\text{max pos}} - N_{\text{min pos}} \right) \right\rfloor, \tag{41}$$

clipped to $[N_{\text{min pos}}, N_{\text{max pos}}]$. With $N_{\text{max pos}} = 10$ and $N_{\text{min pos}} = 2$, the limit transitions from 10 at an equity high to 2 at the historical maximum drawdown level, providing an automatic de-leveraging mechanism during stress periods.

8.2 Signal-Strength Proportional Contract Sizing

The composite signal strength score provides a natural basis for scaling position size within a given entry:

$$C_{t,k} = \max\left(1,\ \min\left(C_{\max},\ \left\lfloor \frac{\text{Score}_{t,k}}{\bar{\text{Score}}_t} + 0.5 \right\rfloor\right)\right), \tag{42}$$

where $\bar{\text{Score}}_t$ is the mean signal strength across all entry candidates on day t. This rule is motivated by the analytic result that optimal Kelly sizing scales proportionally with the squared Sharpe ratio of the signal [15]; the signal score is a heuristic composite rather than a precisely estimated edge, so the Kelly analogy is approximate.

Empirically, the majority of entries in the OOS PUT bucket size above one contract, confirming that the signal strength distribution is sufficiently dispersed to make the sizing rule operationally meaningful rather than a trivial rounding artefact.

The combined effect of the two sizing mechanisms is material. Re-running the backtest with both mechanisms disabled (flat one contract, fixed ten positions) reduces the PUT Sharpe and total P&L materially relative to the dynamic-sizing result. The improvement is attributable roughly equally to the drawdown-adaptive limit (which cuts exposure during stress episodes) and to the signal-strength contract scaling (which concentrates risk in the highest-conviction entries).

9. Conclusion

This paper has developed and empirically evaluated two generations of delta-neutral options strategies grounded in the CRR binomial pricing model. Strategy I, a time-series z-score approach, failed out-of-sample due to VRP contamination, lagging volatility inputs, and low trade count [7, 8, 13].

Strategy II addresses each of these failures by shifting the identification framework from time-series to cross-sectional. By fitting a daily implied volatility surface across the full option chain and trading contracts whose individual IVs deviate materially from the surface-fitted value, the strategy isolates relative-value anomalies that are structurally free of the VRP level bias, instantaneously calibrated to current market conditions, and expressed in the scale-invariant currency of implied volatility [9].

Hyperparameters are selected via a grid search on a held-out bear-market training year (2022), subject to a maximum-drawdown constraint, ensuring that reported performance on the 2023–2024 out-of-sample window is not an artifact of in-sample fitting. The out-of-sample Sharpe of 2.23 for the primary PUT_30_60_OTM bucket, combined with a statistically significant mean daily alpha ($p = 1.75 \times 10^{-3}$ for PUT and $p = 9.56 \times 10^{-6}$ for CALL), provides strong evidence of a persistent surface relative-value signal. Regime robustness is confirmed by positive P&L in both OOS years (2023 recovery and 2024 bull). The CALL bucket achieves near-zero market beta ($\hat{\beta} = 0.017$), the direct consequence of the delta-neutral hedge construction; the PUT bucket retains moderate beta ($\hat{\beta} = 0.140$) from residual directional skew exposure.

Future work will extend the surface specification to the SVI parameterization [9], incorporate GARCH-conditioned delta hedging to reduce transaction cost drag during volatile regimes [5], apply asymmetric entry thresholds calibrated to the directional asymmetry of the volatility risk premium [8], and expand the historical sample to pre-2022 data to conduct a rigorous walk-forward parameter stability analysis across multiple market cycles [13]. Three additional analyses would materially strengthen the current results: (1) testing whether losses cluster on days where the polynomial surface fit falls below $R^2 = 0.90$ (35.7% of OOS trading days), which would either validate the surface specification or motivate a fit-quality filter; (2) decomposing daily P&L into gamma and theta components to identify the economic source of the strategy's edge; and (3) logging delta rehedge events to quantify stock-leg transaction costs.

Limitations. Several important caveats govern the interpretation of these results. (i) All P&L is computed at mid-price; in live trading each fill incurs a real bid-ask cost that depends on order size and market conditions, and the 25 bps synthetic half-spread is an approximation rather than a measured cost. For OTM puts in particular, real bid-ask spreads are often quoted in absolute terms and can represent a substantially larger fraction of the option premium than 25 bps, potentially eroding the PUT bucket's edge materially. (ii) No brokerage commissions, regulatory fees, or financing costs on the hedge-share position are included. (iii) The t-test for mean daily alpha assumes i.i.d. observations; with positions held up to 10 days the effective sample size is lower (see

Section 5.4), and the true significance level is less extreme than the reported p-value. (iv) The polynomial IV surface is a low-order approximation that can violate butterfly arbitrage in the deep wings; this may introduce spurious entry signals for deep OTM options. (v) The CALL_120_150_ATM parameters were carried over from the PUT-optimised grid search without a separate calibration for that bucket, so the CALL results are exploratory rather than independently validated and should not be interpreted on the same footing as the PUT results. (vi) No market-impact model is included; at scale the strategy's own orders would affect the prices at which it executes. (vii) The risk-free rate ($r = 4.5\%$) and dividend yield ($q = 1.2\%$) are held constant across the full sample. In reality the fed funds rate moved from near zero to 5.25–5.50% during 2022 and was held there through most of 2023–2024; using a fixed 4.5% means CRR prices during the 2022 training year are computed at a rate materially above the prevailing rate for much of that year. For short-dated options the effect on price and delta is small, but for the CALL 120–150 DTE bucket where rho exposure is non-trivial, this approximation may introduce systematic error in the surface fit and delta computation. (viii) Delta rehedge events are not logged in the current output, so the total number of SPY share rebalances and the associated stock-leg transaction costs are not quantified. Each rehedge incurs the SPY bid-ask spread on the share adjustment; at scale this could be material and should be measured in future work. (ix) The paper demonstrates that the strategy is profitable but does not decompose P&L into its economic sources. For a delta-neutral strategy the fundamental drivers are the relationship between realized volatility and the surface-implied volatility on open positions (gamma P&L) and the time decay collected or paid on those positions (theta P&L). Without this decomposition it is unclear whether the edge comes from gamma scalping, theta collection, or surface mean-reversion; this analysis is left for future work. (x) The robustness evidence rests on a single train/test split: one bear-market training year (2022) followed by two recovery/bull OOS years (2023–2024). The OOS period does not contain a second drawdown regime, so the strategy has never been tested against a bear market out-of-sample. Claims of regime robustness are therefore limited in scope; a walk-forward validation across multiple cycles (Section 7.5) would provide substantially stronger evidence.

10. References

[1] Cox, J. C., Ross, S. A., & Rubinstein, M. (1979). Option pricing: A simplified approach. *Journal of Financial Economics*, 7(3), 229–263.

[2] Hull, J. C. (2018). *Options, Futures, and Other Derivatives* (10th ed.). Pearson.

[3] Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3), 637–654.

[4] Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2), 579–625.

[5] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327.

[6] Bollerslev, T., Tauchen, G., & Zhou, H. (2009). Expected stock returns and variance risk premia. *Review of Financial Studies*, 22(11), 4463–4492.

[7] Broadie, M., Chernov, M., & Johannes, M. (2009). Understanding index option returns. *Review of Financial Studies*, 22(11), 4493–4529.

[8] Carr, P., & Wu, L. (2009). Variance risk premiums. *Review of Financial Studies*, 22(3), 1311–1341.

[9] Gatheral, J. (2006). *The Volatility Surface: A Practitioner's Guide*. Wiley Finance.

[10] Stein, J. (1989). Overreactions in the options market. *Journal of Finance*, 44(4), 1011–1023.

[11] Whalley, A. E., & Wilmott, P. (1997). An asymptotic analysis of an optimal hedging model for option pricing with transaction costs. *Mathematical Finance*, 7(3), 307–324.

[12] Lam, S. K., Pitrou, A., & Seibert, S. (2015). Numba: A LLVM-based Python JIT compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC* (pp. 1–6). ACM.

[13] Bailey, D. H., Borwein, J., Lopez de Prado, M., & Zhu, Q. J. (2014). Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*, 61(5), 458–471.

[14] Polygon.io. (2024). *Polygon.io REST API Documentation*. Retrieved from https://polygon.io/docs/options.

[15] Kelly, J. L. (1956). A new interpretation of information rate. *Bell System Technical Journal*, 35(4), 917–926.

[16] Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.