

An option-implied benchmark for prediction-market binaries

Fedor Tolmatskiy¹, Aleksei Demin¹

¹Faculty of Computational Mathematics and Cybernetics, Shenzhen MSU-BIT University, Shenzhen, China

1220230018@smbu.edu.cn, demin@smbu.edu.cn

September 2026

Abstract

We measure the disagreement between two types of platforms that price the same state-contingent payoff: option exchanges, where a barrier probability is implicit in the chain, and prediction markets, where it is quoted directly. Every fifteen minutes a collector extracted risk-neutral probabilities from full option chains (Deribit for BTC and ETH, index chains through Yahoo Finance for S&P 500 and Nasdaq 100), compared them with the mid quotes of Kalshi binary contracts on the same barrier and the same expiry, captured the whole order book of every contract that produced a signal, and followed every contract to settlement. The sample contains 4.03 million contract-snapshots, 221 thousand signals and 1 665 settled events. Four findings.

First, on crypto neither forecast is the better one: the Brier scores of the two differ by 0.0003 on both underlyings, a clustered bootstrap interval on that difference covers zero, and the two venues agree on average, the signed mean discrepancy being a fifth of a point against a mean absolute gap of two to three points. On crypto, therefore, $\delta = P_{\text{chain}} - P_{\text{Kalshi}}$ cannot be read as an error of the prediction market. On the indices the picture is different: the discrepancy is large, mostly because the index chain is delayed by fifteen minutes, and the ranking splits, the option chain scoring better on the S&P 500 and worse on the Nasdaq 100, a split that 88 and 80 settled events cannot yet confirm.

Second, δ is hardest to trade where it looks largest. The expected payoff grows with the size of δ while the realised payoff does not follow, which is the signature of the winner's curse.

Third, the most favourable configuration in the sample is the ETH threshold series: at one contract of one dollar notional per signal, 2 502 trades cost \$1 517 in premium and fees and returned +\$3.64, that is +0.24% on the money staked and +1.04 percentage points per event bet.

Fourth, that result is not established, and a larger position does not rescue it. At one contract per signal the clustered bootstrap interval is $[-12.4; +14.1]$ points, the estimate stands 0.15 standard errors from zero where 1.645 would be required, and a single event out of 349 carries more than the entire positive total. Sized at the position the stored book supports, on average \$27 392 per event, the same series turned \$9.56 million of turnover into +\$15.7 thousand, or +0.16% realised against +7.2% expected, and every other series of the four underlyings lost money at that size. A telescoping decomposition of δ into the smile, the forward-spot basis and the residual level puts the largest loss on the basis

group on both crypto underlyings, that is, on artefacts of the data rather than on information. This is a measurement, not a trading signal.

JEL: G13, G14, C53. **Keywords:** risk-neutral density, implied volatility smile, binary options, prediction markets, calibration, market microstructure.

1 Introduction

A Kalshi binary contract pays one dollar if the price of an underlying lies inside a stated range, or above a stated level, at a stated moment, and nothing otherwise. Its quote is therefore a probability stated by a market. Exactly the same probability is implicit in the option chain on the same underlying with the same expiry, as the strike derivative of the call price. Two venues price one payoff, and the difference between their answers is either an error of one of them or a premium.

Deciding which requires a third source of truth, and this study has one: every contract is followed to settlement, so both prices can be scored against the outcome. That is the difference between asking whether two numbers disagree and asking which of them is right.

The question also cannot be answered from an archive. Prediction-market order books are not published historically by the venue, and a reconstruction from trades would lose exactly the quantity that decides executability, namely the depth available at the moment a discrepancy appears. The measurement therefore has to be made forward in time, and the design below is a consequence of that constraint: a fixed snapshot every fifteen minutes, a full chain slice, a full book on every signal, and a settlement record.

Contributions. The main one is a benchmark built from the whole chain rather than from a single option. The forward and the discount factor are extracted from put-call parity, the smile is interpolated in volatility coordinates, and the barrier probability carries the smile-slope term explicitly. This yields a telescopic decomposition of the discrepancy into the contributions of the smile, the forward-spot basis and the residual level, which turns the simplifications that a coarser benchmark makes silently into measurable quantities (Section 3.5). Both prices are then scored against settled outcomes (Sections 5 and 6), and executability is measured on the stored order books at the venue's fee schedule (Sections 7 and 8).

Related work. The mapping from option prices to a risk-neutral distribution is due to Breeden and Litzenberger (1978): the second derivative of the call price in strike is the discounted density, and the first derivative is the exceedance probability we need. The practical step, interpolating the implied volatility smile rather than the price curve and returning to prices before differentiating, goes back to Shimko (1993) and is set out as a procedure by Malz (2014), with a review of the estimation choices and their pitfalls in Figlewski (2018). The sensitivity of a digital payoff to the slope of the smile is standard (Gatheral, 2006). On the prediction-market side, Wolfers and Zitzewitz (2004) state the conditions under which a contract price may be read as a probability, Berg et al. (2008) compare the accuracy of such prices with polls over long samples, and Snowberg et al. (2013) survey the forecasting applications. The standard test in that literature is retrospective, applied after outcomes are known; an option-implied benchmark is available continuously before settlement, which is what makes a forward measurement possible at all.

Structure. Section 2 describes the data and the design of the collection. Section 3 builds the benchmark and bounds its error, and Section 4 defines the signal, the position size and the inference procedure. The results then follow one line of questions: how large the discrepancy is (Section 5), which of the two forecasts is closer to the outcome (Section 6), whether the discrepancy can be traded (Section 7), and how much data a profit would take to establish (Section 8). Section 9 concludes. Appendix A gives the decomposition in full, and Appendix B collects the checks and supplementary results that do not change any conclusion.

2 Data

2.1 Contracts and the payoff map

Two families of Kalshi contract are used for each underlying. A threshold contract pays one dollar if $S_T > K$, which is a cash-or-nothing call with unit payoff; a range contract pays one dollar if $K_1 < S_T \leq K_2$, which is the difference of two such calls. Both are quoted in cents and settle against a published reference price at a stated moment. Events are hourly and daily. All contracts of one event share one settlement price, a fact that governs the statistical treatment in Section 4.4.

The barrier and the settlement moment are taken from the Kalshi contract, and the horizon T of (6) is measured to that moment. The option slice is the first expiry at or after it: 08:00 UTC on Deribit, up to a day after an hourly crypto event, and 16:00 New York on the index chains, up to six hours after an index event. The volatility and the forward are read from that slice without adjustment to the Kalshi moment; the forward’s part of this difference is what the basis term of (7) isolates (Section 7.2). On crypto Kalshi settles on the CF Benchmarks real-time index averaged over the final minute, and the Deribit chain on Deribit’s own index; the two are composites of the same spot venues and track each other closely. On the indices both sides reference the same published level.

Table 1: What each side of the comparison settles against. RTI is the CF Benchmarks real-time index, whose settlement value is the mean of sixty prints in the final minute; NY is 09:30-16:00 New York. On the indices both sides reference the same published level, on crypto they do not: Kalshi settles on the CF Benchmarks index and the chain is quoted against Deribit’s own.

Series	Expiry	Kalshi settles on	Option slice for P_{chain}
KXBTC, KXBTC	hourly, on the hour UTC	BTC-USD RTI	Deribit, 08:00 UTC
KXETH, KXETHD	hourly, on the hour UTC	ETH-USD RTI	Deribit, 08:00 UTC
KXINX	daily, 16:00 NY	S&P 500 level	Yahoo ^SPX, 16:00 NY
KXINXU	hourly, 10:00-16:00 NY	S&P 500 level	Yahoo ^SPX, 16:00 NY
KXNASDAQ100	daily, 16:00 NY	Nasdaq 100 level	Yahoo ^NDX, 16:00 NY
KXNASDAQ100U	hourly, 10:00-16:00 NY	Nasdaq 100 level	Yahoo ^NDX, 16:00 NY

Table 2: Symbols used throughout

Symbol	Meaning
S_T	settlement price of the underlying
F, D	forward and discount factor to the expiry, from put-call parity
S	spot price of the underlying at the snapshot
$\sigma(K)$	implied volatility at strike K ; $d\sigma/dK$ its slope
$P_{\text{chain}}(K)$	option-implied probability of the contract event
P_{Kalshi}	mid of the Kalshi spread, on the Yes probability scale
δ	discrepancy, $\delta = P_{\text{chain}} - P_{\text{Kalshi}}$
$N(\cdot), \varphi(\cdot)$	standard normal distribution function and density
$\mathcal{V}$	vega of the call in forward form
y	realised outcome of the contract, 1 for yes and 0 for no

2.2 Sources and period

Option chains come from Deribit (Deribit, 2026) for BTC and ETH, which quotes continuously, and from Yahoo Finance (Yahoo Finance, 2026) for S&P 500 and Nasdaq 100, which quotes in US market hours. Kalshi quotes, order books and settlements come from the venue API (Kalshi, 2026a). The window opens on 12 September 2026 at 11:13 UTC and closes on 29 September at 02:35 UTC for BTC, ETH and Nasdaq 100, and on 28 September at 23:10 UTC for S&P 500.

The index chains close while the prediction market keeps quoting. Outside the session the chain is frozen at its last value while Kalshi moves, so an unrestricted comparison would measure the staleness of the feed and nothing else. All index statistics in this paper are therefore restricted to 09:30-16:00 New York on weekdays; the restriction is part of the method rather than a robustness check.

2.3 Notation

Table 2 collects the symbols. All quantities live on the Yes probability scale, so a payoff expressed in percentage points is a payoff per dollar of notional. A positive δ means the model prices the event above the market.

2.4 Unfitness flags

A contract is recorded in every snapshot but is not treated as a signal when any of seven conditions holds: the quote is one-sided or missing; the mid is below 5 or above 95 cents,

Table 3: Coverage over the window

	BTC	ETH	S&P 500	Nasdaq 100
Snapshots taken	1 359	1 359	865	789
Snapshots failed for the pair	51	49	506	621
Events (Kalshi expiries)	754	755	96	89
Events settled	748	749	88	80
Distinct contracts	137 576	216 940	5 571	31 160
Contract-snapshots	801 368	983 960	364 195	1 882 960
Contract-snapshots with a trading signal	47 511	19 748	60 310	93 402
Contract-snapshots with a known outcome	97 695	50 456	17 125	23 870
First-signal trades with a known outcome	5 067	4 137	2 270	4 917

where a one-cent tick and the little room left for an edge dominate; the barrier lies outside the quoted strike range, so the probability would require extrapolation; the staleness of the option feed exceeds the measured discrepancy, by the criterion of Section 3.6; the forward-spot basis is anomalous; less than half an hour remains to settlement; or the slice fails the consistency checks of Appendix B.1. A row may carry several flags at once. On BTC the barrier-outside-chain flag fires on 60 % of rows, the one-sided-quote flag on 83 %, and the near-expiry flag on 31 %.

2.5 Coverage

Table 3 states what the window contains and where it fails.

A snapshot is nominally taken every fifteen minutes; the median interval was 16.1 minutes on crypto and 16.2 to 16.9 on the indices, and the irregularity of that cadence is treated in Appendix B.1. Index pairs fail in more than half of the snapshots, almost always because Yahoo returns no two-sided quotes outside US market hours; the counts above are over the whole window, while every index statistic below is restricted to the session. The number of independent outcomes is the number of settled events, about 750 per crypto underlying and 88 and 80 on the indices, and the index figures below should be read as order-of-magnitude statements for that reason. The row of contract-snapshots with a known outcome is in-session for the indices and therefore much smaller than their total.

3 Benchmark: from the option chain to a probability

3.1 Risk-neutral valuation of a binary claim

Absence of arbitrage is equivalent to the existence of an equivalent martingale measure $\mathbb{Q}$ under which the price of any claim is the discounted expectation of its payoff, with no assumption about the dynamics of the underlying. Under that measure the call price $c_t(K)$ and the probability this paper needs are linked by one relation,

$$\frac{\partial c_t}{\partial K} = -D_t \mathbb{Q}(S_T > K), \quad (1)$$

so the first derivative of the call price in strike, divided by minus the discount factor, is the probability of the threshold event. This is the corollary we use of Breeden and Litzenberger (1978), who state the result one derivative further, for the density; the step to (1) and its regularity conditions are standard and are not reproduced. Economically (1) is the

limit of a vertical call spread: long the strike K , short the strike $K + h$, divided by h , reproduces the payoff of the binary contract as $h \rightarrow 0$. A range contract is the difference of two such probabilities,

$$\mathbb{Q}(K_1 < S_T \leq K_2) = \mathbb{Q}(S_T > K_1) - \mathbb{Q}(S_T > K_2). \quad (2)$$

3.2 Forward and discount from put-call parity

Neither the risk-free rate nor a dividend or funding yield is needed separately if the calculation is done in forward parametrisation (Black, 1976). Parity states

$$C(K) - P(K) = D(F - K), \quad (3)$$

so a linear regression of the observed price difference on the strike, over strikes with two-sided quotes, returns DF as the intercept and $-D$ as the slope. Both parameters are therefore read from the same chain slice that supplies the volatilities, with no external rate input, and the dispersion of the residuals is a direct diagnostic of the quality of the slice. The simpler alternative, taking the spot and a postulated rate, is not discarded silently: it is retained as the reference point P_{spot} of the decomposition below, so the cost of that simplification is measured rather than assumed. When Deribit offers too few two-sided pairs the forward falls back to the exchange index price and D is set to one; over horizons of hours the resulting error is below quote precision.

3.3 Inverting volatility

In forward form the call price is

$$c^{\text{mod}}(K, \sigma) = D [F N(d_1) - K N(d_2)], \quad (4)$$

$$d_{1,2} = \frac{\ln(F/K)}{\sigma\sqrt{T}} \pm \frac{\sigma\sqrt{T}}{2},$$

and its vega

$$\mathcal{V} = \frac{\partial c^{\text{mod}}}{\partial \sigma} = D F \varphi(d_1) \sqrt{T} > 0 \quad (5)$$

is strictly positive, so the map $\sigma \mapsto c^{\text{mod}}$ is strictly monotone and the implied volatility of a quoted price is unique. It is found by bisection. The stopping rule is the width of the bracket rather than agreement in price, because vega vanishes away from the forward and a price tolerance would accept a wide interval of volatilities in the wings.

The role of Black and Scholes (Black and Scholes, 1973) here is a change of coordinates, not a model of the world. Nothing below assumes lognormality; the formula is used only because it is a bijection between a price and a number in which interpolation is well posed.

3.4 Smile and the step to probability

The exact statement (1) requires $c(K)$ on a continuum, whereas a chain supplies a dozen to a few dozen quotes, and barriers fall between strikes. A finite difference on the listed grid does not help: on an hourly contract the strike step can exceed one standard deviation of the settlement distribution, so the difference quotient estimates the probability

Table 4: Three benchmarks on one slice; $\bar{\sigma} = \sigma(K_F)$ is the volatility at the quoted strike nearest the forward

Probability	Centre	Volatility
P_{spot}	spot S	flat, $\bar{\sigma}$
P_{flat}	forward F	flat, $\bar{\sigma}$
P_{chain}	forward F	smile, $\sigma(K)$

away from the barrier, while a smaller step would amplify quote noise, and the step is not ours to choose.

Interpolation is therefore performed in volatility coordinates (Shimko, 1993; Malz, 2014; Figlewski, 2018), where the curve is shallow and smooth, rather than in prices, where an interpolant easily violates monotonicity and convexity. The quotes are converted to implied volatilities, a smooth curve $\sigma(K)$ is fitted through them, and the price is defined as $c(K) := c^{\text{mod}}(K, \sigma(K))$, which passes through every quote exactly; nothing is extrapolated beyond the quoted strikes. The strike now enters both directly and through $\sigma(K)$, so its derivative is total, and dividing it by $-D$ gives (Appendix A.1)

$$P_{\text{chain}}(K) = N(d_2) - F \varphi(d_1) \sqrt{T} \frac{d\sigma}{dK}. \quad (6)$$

The first term is what a flat smile would give; the second is the slope correction (Gatheral, 2006). Its sign is known: with a left-skewed smile, $d\sigma/dK < 0$, the correction is positive, so a flat-volatility benchmark understates the probability.

3.5 Decomposition of the discrepancy

Formula (6) can be re-evaluated on the same stored slice with some of its ingredients switched off. Each probability rests on two choices, the centre of the distribution and the volatility, and Table 4 lists three versions, each differing from its neighbour in exactly one of them.

Then identically

$$\begin{aligned} \delta &= P_{\text{chain}} - P_{\text{Kalshi}} \\ &= \underbrace{(P_{\text{chain}} - P_{\text{flat}})}_{\text{skew}} + \underbrace{(P_{\text{flat}} - P_{\text{spot}})}_{\text{basis}} \\ &\quad + \underbrace{(P_{\text{spot}} - P_{\text{Kalshi}})}_{\text{level}}. \end{aligned} \quad (7)$$

Each term is the change in probability caused by one ingredient. The skew term is what the smile adds at a fixed centre: both the difference between the volatility at the barrier and at the forward, and the slope correction of (6). The basis term is the shift of the centre from the spot to the forward at a fixed volatility. The level term is the disagreement that remains even for the simplest benchmark, which uses neither the smile nor parity. The source of a signal is the term of largest absolute value. The attribution depends on the order in which the refinements are removed; the order chosen, shape, then basis, then level, makes P_{spot} the simplest possible benchmark. For a range contract each term is the difference of its values at the two barriers.

One term has a simple size. To first order

$$\text{basis} \approx \varphi(d_2(K, \bar{\sigma}, F)) \frac{\ln(F/S)}{\bar{\sigma} \sqrt{T}}, \quad (8)$$

Table 5: Error budget of the benchmark, percentage points unless stated, over the same rows as Table 6. Lag rows apply (9) to the measured gap between the option slice and the Kalshi quote; feed rows add the fifteen-minute delay of the index chain. The smile row is taken from Table 6; the fee row is the rounding the venue adds to the model fee on one contract.

Source	BTC	ETH	S&P 500	Nasdaq 100
Chain-to-quote lag, median (seconds)	1.0	0.9	1.9	4.9
$ \Delta P_{\text{chain}} $ at that lag, median	0.52	0.63	0.24	0.26
$ \Delta P_{\text{chain}} $ at that lag, 90th percentile	1.39	1.40	0.66	0.75
$ \Delta P_{\text{chain}} $ at a 15-minute feed delay, median	—	—	3.17	3.39
$ \Delta P_{\text{chain}} $ at a 15-minute feed delay, 90th percentile	—	—	11.07	9.30
Smile-slope term in the mean discrepancy	−0.01	−0.05	+0.41	+0.78
Fee rounding on one contract	0.47	0.48	0.62	0.55
Mean $ \delta $ over those rows, for comparison	2.08	2.67	12.21	11.80

and since the log basis $\ln(F/S)$ grows roughly in proportion to T , the term grows as $\sqrt{T}$: it is below quote precision on hourly events and material on multi-day ones. A large basis term on a short contract is therefore a sign of a stale or inconsistent slice rather than of information, which is how Section 7.2 reads it. Full expressions for all three terms are given in Appendix A.

3.6 Feed staleness

If the option slice is Δt old, the log forward has moved by a quantity with standard deviation $\sigma \sqrt{\Delta t}$, and at the money the probability responds to it with sensitivity $\varphi(0)/(\sigma \sqrt{T})$. The expected displacement is then

$$\mathbb{E} |\Delta P_{\text{chain}}| \approx \frac{1}{\pi} \sqrt{\frac{\Delta t}{T}}, \quad (9)$$

independent of volatility, which cancels. Short contracts are therefore where our own probability is noisiest, and staleness exceeding the discrepancy is one of the unfitness flags.

Table 5 sizes this and the other sources of error on the sample. The chain and the Kalshi quote are taken seconds apart, which (9) turns into a fifth to two thirds of a point at the median. The index chain adds a feed delay of about fifteen minutes, which gives three points at the median and nine to eleven at the ninetieth percentile, the order of the index discrepancy itself. The error of the smile slope and the consistency checks on each slice are described in Appendix B.1.

4 Signal, book and inference

4.1 Reconstructing the ask and the discrepancy

Kalshi publishes bids on both sides only, so the ask on one side is reconstructed by reflecting the opposite-side bid,

$$\text{ask} = 1 - \text{bid}_{\text{opposite}}, \quad (10)$$

and the mid P_{Kalshi} is the middle of the resulting spread on the Yes probability scale. The discrepancy is $\delta = P_{\text{chain}} - P_{\text{Kalshi}}$.

4.2 Signal condition and fee

A contract becomes a signal when the discrepancy exceeds half the spread and no unfitness flag is set,

$$|\delta| > \frac{1}{2} (\text{ask} - \text{bid}), \quad (11)$$

which is the condition for the model price to lie outside the quoted spread rather than inside it. The taker fee on an order of C contracts filled at a price of P dollars is $f(C, P) = \lceil 100 \mu 0.07 C P(1 - P) \rceil / 100$ dollars, rounded up to the cent, with a fee multiplier $\mu = 1$ on every series used here (Kalshi, 2026b). A single contract therefore pays at least one cent, which is one percentage point of its dollar notional, and the per-contract tables below charge that rounded fee on every trade. A fill that sweeps several book levels is charged level by level, which bounds the cost from above because the venue rounds once per order; that is how Section 4.3 evaluates it. Crypto series carry no maker fee, while the index series charge one at a quarter of the taker rate, which matters for the resting-order question left open in Section 9. Note that (11) compares the discrepancy with the spread but not with the fee, which is why the smallest bucket in Section 7.1 has a negative promise by construction.

4.3 Capacity and position size

For every signal the whole book is stored at full depth, with offers reconstructed by (10). Let $\bar{p}(x)$ be the all-in price per contract of buying x contracts: the book is swept from the best offer upwards, and the fee of Section 4.2 is charged on each level at that level's own price. Since the levels are sorted, $\bar{p}(x)$ is non-decreasing in x . The expected value of the position under the model is

$$V(x) = x (P_{\text{chain}} - \bar{p}(x)). \quad (12)$$

Two sizes are reported. The *capacity* $x^* = \max\{x : \bar{p}(x) \leq P_{\text{chain}}\}$ is the largest position whose expected value is still non-negative; at x^* the expected profit is zero by construction, so capacity is an upper bound on what could be deployed, not a recommended size. The *optimal size* $\hat{x} = \arg \max_x V(x)$ stops at the last level whose own price, fee included, is below P_{chain} , hence $\hat{x} \leq x^*$. Capacity near zero means the signal could not be traded at all; large capacity with small $\hat{x}$ means it could be traded only far below the visible depth. Both quantities are recomputed from the stored book, so they can be revised under another target probability or fee schedule.

The traded side is the one with the larger $V(\hat{x})$. On 28 % of BTC and 33 % of ETH signals neither side has non-negative expected value at any size, so the sized results of Section 7.1 cover about seventy per cent of crypto signals. On 1.0 and 1.1 % the side opposite to the signal is chosen, which a single consistent quote rules out: in those rows the book, fetched seconds after the quote, sits six to seven points away from it on both sides. They carry 0.7 and 1.1 % of the sized turnover.

4.4 Unit of observation and bootstrap

All contracts that settle against one price have almost perfectly dependent outcomes: twenty contracts decided by one settlement value are one bet, not twenty. The unit of observation is therefore the settlement moment rather than the Kalshi event, because the two contract families of one underlying expire on the same hour and are decided by the same number: KXBTC and KXBTCd at 12:00 UTC read the same index print, as do KXINX and KXINXu at the close. Clustering on the event would count them as two draws. The result of cluster i is the sum of the payoffs of its signalled contracts, each bought once at its first signal, in percentage points of one dollar per contract:

$$X_i = \sum_{c \in i} (\mathbf{1}[\text{settlement favours the side bought}] - \text{price}_c - \text{fee}_c). \quad (13)$$

For a series of m clusters, $\hat{\mu}$ and $\hat{\sigma}$ are the sample mean and standard deviation of $X_1, \dots, X_m$. Within one series a cluster is one expiry, so the two coincide there; they differ only where series are pooled, as in Tables 6 and 7 and in the buckets of Section 7.1. Moving from the event to the settlement moment widens those intervals by 7 to 22 % on the Brier difference and by up to a third on the payoff buckets, and changes no sign and no statement about whether an interval excludes zero.

We want to know how far $\hat{\mu}$ could move had a different set of events been drawn, using nothing but the sample itself. This is what the bootstrap does (Efron, 1979, 1981): draw m clusters with replacement, recompute the mean, repeat $B = 1000$ times, and take the 2.5 and 97.5 percentiles of the replicates as the 95 % interval. Clusters are drawn whole, which is the cluster form of the method (Davison and Hinkley, 1997; Field and Welsh, 2007); drawing rows instead would copy one settlement many times and make the interval artificially narrow.

Example: ETH thresholds. This series has $m = 349$ clusters, one per expiry since a single series is involved, and 2502 signalled contracts, about seven per cluster. Its X_i have mean $\hat{\mu} = +1.04$ and standard deviation $\hat{\sigma} = 127.7$ percentage points, reported with the other series in Table 11. In money, the average event bet at one dollar per contract earned 1.04 cents, and 349×1.04 cents is the +\$3.64 total quoted in the abstract. The dispersion exceeds 100 points because X_i sums several contracts decided by one settlement price: when they all settle the same way, an event wins or loses several dollars at once. The standard error of the mean is $\hat{\sigma}/\sqrt{m} = 127.7/\sqrt{349} \approx 6.8$, so the estimate lies 0.15 standard errors from zero. The bootstrap interval is $[-12.4; +14.1]$; the normal approximation $\hat{\mu} \pm 1.96 \hat{\sigma}/\sqrt{m}$ gives $[-12.4; +14.4]$, so no conclusion depends on the choice of method. The interval contains zero: on 349 events the sample cannot distinguish this series from one with no edge, and the positive total is fragile, since the single best event exceeds it. The same $\hat{\mu}$ and $\hat{\sigma}$ enter the profit horizon of Section 8.

Every other series and every interval in the tables below is computed the same way; where a table reports payoff per contract, the statistic is the ratio of summed payoffs

Table 6: Mean discrepancy $\delta = P_{\text{chain}} - P_{\text{Kalshi}}$ over all rows with both prices, percentage points, with 95 % clustered bootstrap intervals, and the same mean with a flat volatility in place of the smile.

Underlying	n	Mean	Median	95 % interval	Flat σ	Smile term
BTC	117 199	+0.13	+0.17	[−0.01; +0.21]	+0.14	−0.01
ETH	59 284	−0.24	−0.06	[−0.35; −0.15]	−0.19	−0.05
S&P 500	207 171	+4.89	+5.99	[+3.38; +6.44]	+4.47	+0.41
Nasdaq 100	407 001	+2.48	+3.18	[+0.99; +4.18]	+1.70	+0.78

to summed contract counts over the resampled events. On crypto neighbouring events are close to independent, the lag-one autocorrelation of X_i running from 0.00 to −0.07 across the four series, so a block bootstrap (Künsch, 1989) would not widen those intervals. On the index threshold series it is +0.56 and +0.62: consecutive hourly expiries of one session share a move, and the index intervals below are therefore optimistic by an amount this design cannot measure with 77 and 67 events.

5 How large the discrepancy is

Table 6 averages δ over every row with both prices, with the intervals of Section 4.4.

On crypto the two venues agree on average: the mean wedge is a fifth of a point or less, of opposite sign on the two underlyings, while the typical gap of two to three points cancels in the mean. On the indices the wedge is large and the model stands above the market on 65 and 59 % of rows. Much of that size is the fifteen-minute delay of the index chain, which by Table 5 displaces the probability by about three points at the median and ten at the ninetieth percentile, against a mean $|\delta|$ of twelve. The delay does not explain the sign, since its error has mean zero; that would require the index to drift during the delay, which we do not measure.

The smile barely matters on crypto, where dropping it moves the mean by a twentieth of a point or less. On the indices it accounts for a tenth to a third of the wedge, with the expected sign. Treating the quoted ladder of an event as a distribution (Appendix B.6) gives the same picture: the centres agree to a few hundredths of a percent, and the prediction market prices a distribution 4 % wider on crypto and 10 % wider on the S&P 500.

6 Calibration: which forecast is closer to the outcome

Both forecasts are scored against realised settlement on every row with a known outcome, indices in trading hours only. With $y = 1$ if the contract settled yes and $y = 0$ otherwise, the Brier score (Brier, 1950) is the mean squared error of the forecast, $\mathbb{E}[(P - y)^2]$, averaged over rows; log-loss, $-\mathbb{E}[y \ln P + (1 - y) \ln(1 - P)]$, penalises confident misses more heavily. Lower is better for both (Table 7).

On crypto the market is ahead by 0.0003 in Brier score on both underlyings, and by log-loss as well, but the interval on that difference covers zero: the two forecasts are close and the sample does not rank them. That is enough for the claim this paper needs, since whatever δ is, it is not evidence that

Table 7: Brier score and log-loss of the two forecasts. Lower is better; the better of the pair is in bold. The difference is model minus market in Brier score, with a 95 % bootstrap interval on the paired difference, clustered on the settlement moment as Section 4.4 defines it.

Underlying	n	Brier			95 % interval	log-loss	
		model	P_{chain}	market mid		model	market
BTC	97 695	0.09520	0.09492	+0.0003	[−0.0010; +0.0021]	0.3145	0.3128
ETH	50 456	0.10714	0.10687	+0.0003	[−0.0019; +0.0028]	0.3621	0.3508
S&P 500	17 125	0.12434	0.13353	−0.0092	[−0.0185; −0.0019]	0.4126	0.4155
Nasdaq 100	23 870	0.14468	0.12804	+0.0166	[+0.0048; +0.0295]	0.5576	0.4030

the prediction market is the party in error, but it is less than a claim that the market forecasts better. On the indices the ranking splits and both differences exclude zero, the model winning on the S&P 500 by 0.0092 and losing on the Nasdaq 100 by 0.0166; with 88 and 80 settled events, and with the event dependence of Section 4.4, both remain order-of-magnitude statements. A lower Brier score is also not a profitable rule: in the half-to-two-hour band where the S&P 500 chain scores better than the market, acting on the difference lost money (Appendix B.2). Brier score rewards calibration and resolution together, so these comparisons rank the forecasts, not their calibration alone.

How close the forecasts are can be stated exactly. With $P_{\text{chain}} = P_{\text{Kalshi}} + \delta$,

$$\text{Brier}(P_{\text{chain}}) - \text{Brier}(P_{\text{Kalshi}}) = \mathbb{E}[\delta^2] + 2 \mathbb{E}[\delta (P_{\text{Kalshi}} - y)] . \quad (14)$$

The first term is the penalty for departing from the market, the second the reward for departing in the direction in which the market errs, so δ improves on the market only if the second outweighs the first. Both are computed on the rows of Table 7, and neither is small. On BTC $\mathbb{E}[\delta^2] = 0.00155$, an RMS of δ of 3.9 percentage points, against a Brier gap of 0.00029; the cross term is −0.00126, so 81 % of the penalty is paid back. On ETH the two are 0.00303 and −0.00275, a repayment of 91 %. The model therefore departs from the market largely in the direction in which the market errs, but not by enough to cover the noise of departing. The indices split the same way: on the S&P 500 the cross term −0.03442 exceeds the penalty 0.02523, which is the whole of the model’s advantage there, while on the Nasdaq 100 it repays only a third of 0.02517 and the model loses by 0.0166.

A rule that ignores the model and buys yes wherever the mid lies between 0.2 and 0.6 tests the venue directly; it loses on both crypto underlyings, −2.8 points per contract on BTC and −8.2 on ETH with intervals below zero, the market being calibrated inside that band to within a point or two and the loss equal to the cost of entry (Appendix B.5).

7 Can the discrepancy be traded

7.1 Promise against realisation

For every signal we record what the trade promised at entry, $|\delta|$ less half the spread and the fee, since entry is at the offer, and what it returned after settlement. Two modes are reported: every snapshot treated as an independent hypothesis, and one trade per contract taken at its first signal. The second is closer to execution, since it excludes re-entering

Table 8: Expected and realised payoff, percentage points of one dollar per contract

Underlying	Mode	n	Expected	Realised
BTC	every snapshot a hypothesis	42 025	+0.89	−0.21
BTC	one trade per contract, first signal	5 067	+2.77	−0.75
ETH	every snapshot a hypothesis	18 117	+1.26	+0.13
ETH	one trade per contract, first signal	4 137	+3.22	−0.51
S&P 500	one trade per contract, first signal	2 270	+1.98	−5.88
Nasdaq 100	one trade per contract, first signal	4 917	+1.60	−12.58

the same contract at the following snapshot.

In every row of Table 8 the realisation falls short of the promise, and at one trade per contract every underlying now loses: the crypto losses, half a point and three quarters of a point per contract, are of the order of the rounded fee itself. In money, entering at the position size with the highest expectation on the real book: BTC 3 791 trades and \$23.6 m turnover for \$1.62 m expected against −\$211 k realised; ETH 3 035 trades and \$10.5 m turnover for \$743 k expected against −\$67 k realised; S&P 500 −\$565 k on \$2.4 m; Nasdaq 100 −\$1.11 m on \$6.8 m turnover. Both sizings charge the exchange fee of Section 4.2; at one contract that fee is a full cent, which is why a per-contract result of half a point is not distinguishable from the cost of trading.

Figure 1 splits the same comparison by the size of the discrepancy at the signal. The gap between promise and realisation widens with $|\delta|$, which is the signature of the winner’s curse (Capen et al., 1971): selecting on the largest $|\delta|$ also selects the largest estimation error, and (9) shows how large that error can be on short contracts. The smallest bucket has a negative promise, because the signal condition (11) compares the discrepancy with the spread and not with the fee, so discrepancies below roughly two percentage points are consumed by the fee before the model is tested at all. No bucket on any underlying realises a positive payoff with an interval above zero. Five realise a negative one with an interval below zero, all of them on the indices: the S&P 500 at 5-10 pp, −10.00 points with $[-14.74; -4.47]$, and above 10 pp, −6.80 with $[-13.08; -1.74]$; the Nasdaq 100 at 2-5 pp, −8.70 with $[-20.05; -0.74]$, at 5-10 pp, −10.59 with $[-13.15; -7.59]$, and above 10 pp, −18.37 with $[-26.90; -10.42]$. On crypto no bucket separates from zero in either direction.

7.2 Which component drives a signal

The decomposition (7) is evaluated on every signal, and each signal is attributed to its dominant term.

In Table 9 the basis group is the worst on both crypto underlyings, losing 1.30 and 1.08 percentage points per contract against −1.12 and −0.52 for the level group, and on the Nasdaq 100 it is the worst by a wide margin, −18.64 points on a hit rate of 0.37. Only two groups in the table are right less than half the time, and both are basis groups: 0.37 on the Nasdaq 100 and 0.48 on BTC. A basis discrepancy is a symptom of a stale or inconsistent slice rather than of information, and the mismatch of moments that produces it is set out in Table 1, so this is a property of our data and not of the venue. Excluding it is the one filter the evidence supports without reservation, and it is also the term that a bench-

Table 9: Signals by dominant component, one trade per contract

Underlying	Component	Signals	Mean $ \delta $, pp	Hit rate	Payoff, pp
BTC	level	3 819	5.57	0.57	−1.12
BTC	basis	558	3.06	0.48	−1.30
BTC	skew	690	4.04	0.56	+1.73
ETH	level	3 287	6.98	0.58	−0.52
ETH	basis	382	3.21	0.51	−1.08
ETH	skew	468	5.31	0.53	+0.04
S&P 500	level	1 987	18.11	0.76	−6.64
S&P 500	basis	109	12.92	0.72	+4.28
S&P 500	skew	174	11.67	0.59	−3.65
Nasdaq 100	level	3 715	17.47	0.65	−12.23
Nasdaq 100	basis	318	16.13	0.37	−18.64
Nasdaq 100	skew	884	15.77	0.51	−11.90

Table 10: Book capacity and dollar expectation by size of the discrepancy, threshold series: median capacity on the signal side and mean expectation at the optimal size. The choice of side is described in Section 4.3.

Bucket	BTC (KXBTCD)		ETH (KXETHD)	
	capacity, \$	expectation, \$	capacity, \$	expectation, \$
0-2 pp	0	13	0	10
2-5 pp	7 167	83	1 775	46
5-10 pp	13 606	433	3 231	326
> 10 pp	11 656	1 305	2 085	652

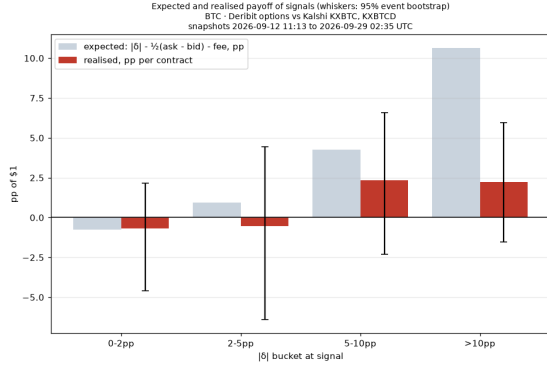
mark built on the spot rather than on the forward cannot see at all. The S&P 500 row is the exception, +4.28 points on 109 trades, which is one group of eleven settled events and carries no weight against the other three.

The skew group is the only one with a positive payoff on crypto, +1.73 percentage points on 690 BTC trades and +0.04 on 468 ETH trades, and it is negative on both indices. The hypothesis that the shape of the smile carries information the prediction market lacks is therefore neither confirmed nor refuted, which is the honest reading of samples of a few hundred trades with dispersion of the order of the effect.

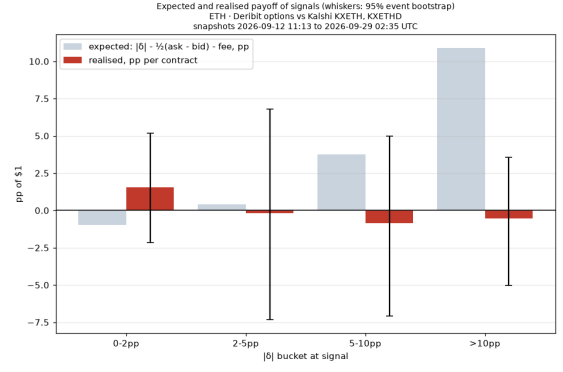
7.3 Speed and liquidity

Two execution constraints could explain the gap between promise and realisation. The first is speed. The median signal is gone by the next snapshot, the median time to the first signal-free observation being about eighteen to nineteen minutes on all four underlyings (Appendix B.3). A signal therefore cannot be traded on a fifteen-minute clock; whether it survives the seconds an order takes is below the resolution of this design, and the backtest assumes a fill at the stored book.

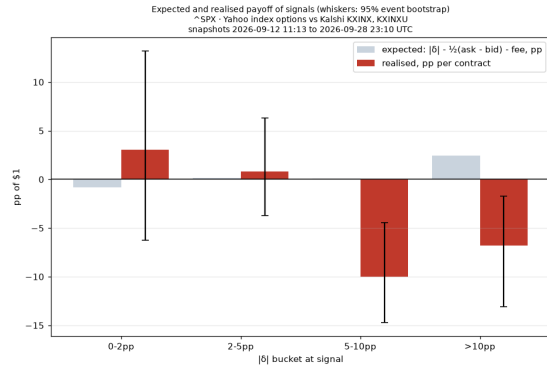
The second is liquidity: large discrepancies might appear only where the book is empty. On crypto the stored books do not support this (Table 10). Capacity is near zero only in the 0-2 pp bucket, where the fee already exceeds the discrepancy, and runs to thousands of dollars in every other bucket, while the expectation per signal rises by two orders of magnitude. What thins is the best bid on the signal side: in the largest bucket it is below \$20 on 20 % of BTC and 40 % of ETH signals, so the largest discrepancies rest on a less reliable mid. On crypto the gap between promise and realisation is therefore not an execution constraint.



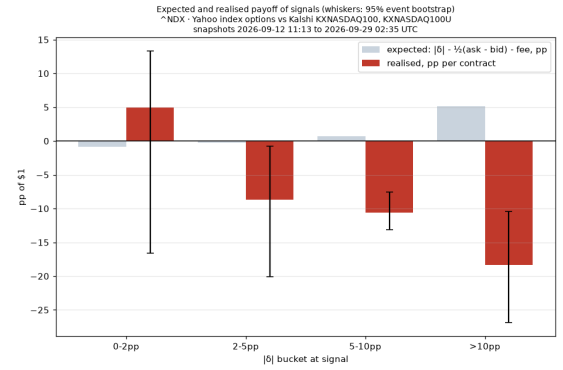
(a) BTC



(b) ETH



(c) S&P 500



(d) Nasdaq 100

Figure 1: Expected and realised payoff by the size of the discrepancy at the signal. The promise grows with the bucket on the crypto pairs and, with one inversion, on the indices; the realisation follows it on none.

Table 11: Profit horizon at current estimates; $\hat{\mu}$ and $\hat{\sigma}$ per event bet

Underlying	Series	Events	$\hat{\mu}$, pp	$\hat{\sigma}$, pp	$P(\text{profit})$ now	N_{95}	Stake, \$
BTC	ranges	338	-10.36	68.0	0.00	unreachable	1 732
BTC	thresholds	331	-0.89	194.4	0.47	unreachable	69 607
ETH	ranges	356	-6.94	69.9	0.03	unreachable	2 756
ETH	thresholds	349	+1.04	127.7	0.56	40 570	27 392
S&P 500	thresholds	77	-170.92	492.8	0.00	unreachable	19 465
Nasdaq 100	thresholds	67	-915.90	1 401.7	0.00	unreachable	92 391

On the indices the book is the constraint. Above 5 pp most signals rest on less than \$50 of capacity and about half on a best bid under \$20 (Appendix B.4), so large index discrepancies are prices at which almost nothing could have been bought, and their losses cannot be separated into model error and execution artefact with these data.

8 Profit horizon

Because the unit of a bet is an event, the question of whether a strategy is profitable is a question about the number of events. Take the sum of first-signal payoffs of all signalled contracts of one event as a single draw, with mean $\hat{\mu}$ and dispersion $\hat{\sigma}$. The total over N such bets is positive with 95 % probability once

$$N > N_{95} = \left(\frac{1.645 \hat{\sigma}}{\hat{\mu}} \right)^2. \quad (15)$$

Table 11 applies (15) to each series. Five of the six have a negative $\hat{\mu}$ and no horizon at all. The only positive case,

ETH thresholds, would need 40 570 events, more than five years at the current rate. Even that figure is optimistic: $\hat{\sigma}$ enters it squared while estimated on the same sample, and the series was chosen as the best of six, so it shows the scale of observation required rather than a promise.

The $\hat{\mu}$ and $\hat{\sigma}$ columns count one dollar per contract, while the stake column is the sized position $\hat{x}$ of (12), so the two must not be multiplied. On ETH thresholds the product would read \$99 thousand, against a sized outcome of +\$16 thousand, which with -\$82 thousand on the range series reconciles with the -\$67 thousand of Section 7.1. The sized position is largest where the promised edge is largest (Table 10), which is the winner's curse in dollars. The two crypto series also do not share a calendar, since the threshold series lost a day of collection to a data incident.

9 Conclusion

A benchmark built from the full option chain, with the forward from parity and the smile slope carried explicitly, can be measured against a prediction market forward in time. On crypto neither forecast beats the other and the two venues agree on average, so a discrepancy is not evidence that the prediction market is wrong. The discrepancy is hardest to trade where it looks largest: the promised payoff grows with its size much faster than the realised one, which is the winner's curse. Only the ETH threshold series earns a positive return, which is not established and which a larger position

does not rescue. The signals that lose most on crypto are those driven by the forward-spot basis, which points to stale data rather than information. The benchmark is therefore a diagnostic rather than a trading signal. Only taker execution is tested; whether an edge survives a resting limit order remains open.

Data and code. The collector runs every fifteen minutes on a small server and writes a single SQLite database: option slices, Kalshi quotes with model probabilities and flags, order books and capacity, settlements, and a heartbeat of snapshots. Analysis runs separately on a copy of that database and regenerates every table and figure above. The code is open under the MIT licence at <https://github.com/mr-tolmatskiy/kalshi-option-benchmark>; market data belongs to the venues and is not redistributed.

References

- Berg, J. E., Nelson, F. D., and Rietz, T. A. (2008). Prediction market accuracy in the long run. *International Journal of Forecasting*, 24(2), 285-300. doi:10.1016/j.ijforecast.2008.03.007.
- Black, F. (1976). The pricing of commodity contracts. *Journal of Financial Economics*, 3(1-2), 167-179. doi:10.1016/0304-405X(76)90024-6.
- Black, F., and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3), 637-654. doi:10.1086/260062.
- Breeden, D. T., and Litzenberger, R. H. (1978). Prices of state-contingent claims implicit in option prices. *Journal of Business*, 51(4), 621-651. doi:10.1086/296025.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1-3.
- Capen, E. C., Clapp, R. V., and Campbell, W. M. (1971). Competitive bidding in high-risk situations. *Journal of Petroleum Technology*, 23(6), 641-653. doi:10.2118/2993-PA.
- Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press. doi:10.1017/CBO9780511802843.
- Deribit (2026). API v2 documentation. <https://docs.deribit.com> (accessed 29 September 2026).
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1), 1-26. doi:10.1214/aos/1176344552.
- Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9(2), 139-158. doi:10.2307/3314608.
- Field, C. A. and Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society Series B*, 69(3), 369-390. doi:10.1111/j.1467-9868.2007.00593.x.
- Figlewski, S. (2018). Risk-neutral densities: a review. *Annual Review of Financial Economics*, 10, 329-359. doi:10.1146/annurev-financial-110217-022944.
- Gatheral, J. (2006). *The Volatility Surface: A Practitioner's Guide*. Hoboken: Wiley.
- Kalshi (2026a). Trade API v2 documentation. <https://docs.kalshi.com> (accessed 29 September 2026).
- Kalshi (2026b). Fee schedule. <https://kalshi.com/docs/kalshi-fee-schedule.pdf> (version of 7 July 2026, accessed 29 September 2026).
- Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3), 1217-1241. doi:10.1214/aos/1176347265.
- Malz, A. M. (2014). A simple and reliable way to compute option-based risk-neutral distributions. *Federal Reserve Bank of New York Staff Reports*, no. 677. doi:10.2139/ssrn.2449692.
- Shimko, D. (1993). Bounds of probability. *Risk*, 6(4), 33-37.
- Snowberg, E., Wolfers, J., and Zitzewitz, E. (2013). Prediction markets for economic forecasting. In G. Elliott and A. Timmermann (eds.), *Handbook of Economic Forecasting*, vol. 2, 657-687. Amsterdam: Elsevier. doi:10.1016/B978-0-444-53683-9.00011-6.
- Wolfers, J., and Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, 18(2), 107-126. doi:10.1257/0895330041371321.
- Yahoo Finance option chains for S&P 500 and Nasdaq 100, accessed through the yfinance library. <https://finance.yahoo.com>, <https://github.com/ranaroussi/yfinance> (accessed 29 September 2026).

A Decomposition of the discrepancy in full

A.1 Derivation of the working formula

With $d_{1,2}$ as in (4), both depend on the strike in the same way, $\partial d_1 / \partial K = \partial d_2 / \partial K = -1 / (K\sigma\sqrt{T})$, and they satisfy the identity $F\varphi(d_1) = K\varphi(d_2)$. Differentiating (4) in the strike at fixed volatility,

$$\begin{aligned} \left. \frac{\partial c^{\text{mod}}}{\partial K} \right|_{\sigma} &= D \left[F\varphi(d_1) \frac{\partial d_1}{\partial K} - N(d_2) \right. \\ &\quad \left. - K\varphi(d_2) \frac{\partial d_2}{\partial K} \right] = -D N(d_2), \end{aligned} \quad (\text{A.1})$$

because the first and last terms cancel by the identity. Differentiating in volatility gives the vega (5), $\mathcal{V} = DF\varphi(d_1)\sqrt{T}$. When the volatility is itself a function of the strike, $c(K) = c^{\text{mod}}(K, \sigma(K))$, and the chain rule gives

$$\frac{dc}{dK} = -D N(d_2) + \mathcal{V} \frac{d\sigma}{dK}. \quad (\text{A.2})$$

Dividing by $-D$ and using (1) yields (6).

A.2 The three components

Write $d_{1,2}(K, \sigma, X) = \ln(X/K)/(\sigma\sqrt{T}) \pm \sigma\sqrt{T}/2$ for the quantities of (4) at a general centre X . The three benchmarks of Table 4 are

$$P_{\text{chain}} = N(d_2(K, \sigma(K), F)) - F\varphi(d_1(K, \sigma(K), F))\sqrt{T} \frac{d\sigma}{dK}, \quad (\text{A.3})$$

$$P_{\text{flat}} = N(d_2(K, \bar{\sigma}, F)), \quad (\text{A.4})$$

$$P_{\text{spot}} = N(d_2(K, \bar{\sigma}, S)), \quad (\text{A.5})$$

where S is the spot at the same timestamp. The skew term splits into two pieces,

$$\begin{aligned} \text{skew} = & \underbrace{N(d_2(K, \sigma(K), F)) - N(d_2(K, \bar{\sigma}, F))}_{\text{local level of the smile}} \\ & - \underbrace{F\varphi(d_1(K, \sigma(K), F))\sqrt{T} \frac{d\sigma}{dK}}_{\text{slope}}. \end{aligned} \quad (\text{A.6})$$

The first piece vanishes when the barrier coincides with the strike nearest the forward, and the term then reduces to the slope correction; away from the forward the two pieces need not share a sign.

The basis term is exactly

$$\text{basis} = N(d_2(K, \bar{\sigma}, F)) - N(d_2(K, \bar{\sigma}, S)), \quad (\text{A.7})$$

and since $d_2(K, \bar{\sigma}, F) - d_2(K, \bar{\sigma}, S) = \ln(F/S)/(\bar{\sigma}\sqrt{T})$, a first-order expansion gives (8). With $\ln(F/S) \approx bT$ for a carry rate b , the term is of order $\varphi b\sqrt{T}/\bar{\sigma}$. The same $\bar{\sigma}$ enters P_{flat} and P_{spot} , so the basis is a pure shift of the centre at fixed volatility.

The discount factor appears in none of the three terms, since it cancels in the probability, but it enters $\bar{\sigma}$ and $\sigma(K)$ through the inversion of (4); an error in D therefore reaches the decomposition inside the volatility rather than beside it.

A.3 Simplified benchmarks

Equation (7) also locates what a simpler construction discards. A benchmark that inverts a single call at the barrier strike takes the correct local volatility $\sigma(K)$ and so recovers the first piece of the skew term, but it drops the slope and does not extract the forward: it misses the slope component of the skew and the whole of the basis. Section 7.2 shows that the second omission is not harmless.

B Checks and supplementary results

Nothing in this appendix changes a conclusion of the main text.

B.1 Cadence and slice checks

The collector processes the four pairs in one loop, so the interval between snapshots varies: its median was 16.1 minutes on crypto and 16.2 and 16.9 on the S&P 500 and Nasdaq 100 pairs, and 76 to 92 % of intervals were twenty minutes or shorter. Recomputing the first-signal statistics on

Table 12: Index order books by bucket, both series pooled: median capacity on the signal side, share of signals below the \$50 threshold, and share whose best bid is below \$20.

Bucket	S&P 500			Nasdaq 100		
	capacity, \$	<\$50	bid <\$20	capacity, \$	<\$50	bid <\$20
0-2 pp	0.00	0.82	0.28	0.00	0.82	0.18
2-5 pp	51.34	0.49	0.20	21.90	0.52	0.15
5-10 pp	1.60	0.84	0.61	14.31	0.62	0.56
> 10 pp	47.56	0.52	0.50	47.98	0.51	0.58

regular snapshots only, those within twenty minutes of the previous one, changes nothing: the BTC payoff is -0.66 points on 4 718 trades against -0.75 on all 5 067, and the S&P 500 payoff -6.23 against -5.88 . Signal lifetimes move by less than a minute.

Every slice is checked for internal consistency: the exceedance probability must decrease in the barrier, interval probabilities must be non-negative, total variance $\sigma^2 T$ must increase with maturity, and call and put volatilities at one strike must agree. A violation marks the slice unfit (Section 2.4). The error of the slope $d\sigma/dK$, estimated on a finite strike grid, enters (6) with the factor $F\varphi(d_1)\sqrt{T}$ and is largest at the money.

B.2 Payoff by time to settlement

On crypto the realised payoff is near zero in the half-to-two-hour band, which holds most trades, and negative beyond six hours (Figure 2). The two-to-six-hour band is positive, $+4.49$ points on BTC, $[+0.04; +8.44]$, and $+9.63$ on ETH, $[+0.89; +16.62]$, but it rests on 17 and 16 settlement moments and nothing else in the paper points at it, so we report it without building on it. On the indices the payoff is negative at every horizon, including the half-to-two-hour band where the S&P 500 chain scores better than the market: a better forecast is not a profitable rule. The crypto bucket below half an hour ($+25.56$ points on ETH) comes from rows that leaked past the near-expiry flag, all from a single settlement moment; it has no interval and is excluded from every summary.

B.3 Signal persistence

Most contracts signal in one or two consecutive snapshots (Figure 3). An AR(1) fit on contracts with at least eight snapshots gives $\varphi = 0.59$ on BTC and 0.72 on the S&P 500, a half-life of 1.30 and 2.10 snapshots, or 21 and 34 minutes at the measured median interval.

B.4 Order books on the indices

Above the 2-5 pp bucket the index books are empty in the sense of the \$50 threshold: most signals rest on less than \$50 of capacity and about half on a best bid under \$20 (Table 12). Figure 4 shows the contrast with crypto, where capacity stays well above the threshold across the whole range of $|\delta|$.

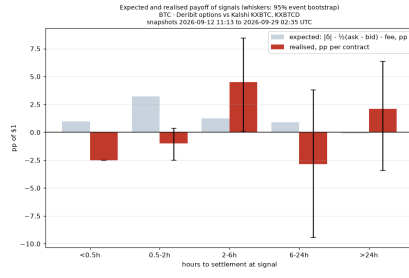
B.5 Reliability diagrams

Counted by snapshot (Figure 5), both crypto curves lie above the diagonal between 0.2 and 0.6, by up to 14 points on BTC. This is an artefact of the unit of observation: counted once per contract, the middle BTC buckets agree with their forecasts to within a few points. The index curves have not been recounted and should not be read as venue miscalibration.

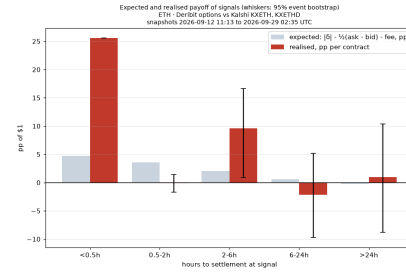
The displacement still suggests a rule that needs no model, buying yes wherever the mid lies between 0.2 and 0.6, and the rule tests the venue directly. It loses -2.8 points per contract on BTC, $[-5.4; -0.2]$, and -8.2 on ETH, $[-11.0; -5.4]$. Inside the band the market is calibrated to within a point or two, so the loss is the cost of entry: the average fill sits 1.7 and 4.6 points above the mid, and the rounded fee adds about two more. Together with Section 7.1, this shows the same result from both sides: a rule that selects on the discrepancy loses, and so does one that ignores it, because entry costs more than any disagreement that survives to settlement.

B.6 The quoted ladder as a distribution

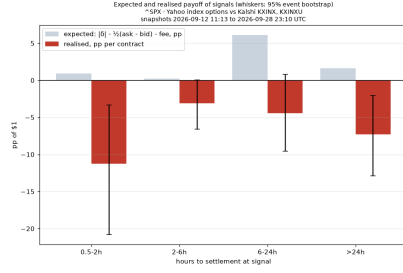
Treating the quoted ladder of an event as a distribution, the option-implied measure $\mathbb{Q}$ and the Kalshi measure $\mathbb{K}$ agree in location to within 0.03 % on BTC, ETH and the S&P 500, but $\mathbb{K}$ is wider by 3.9 % on BTC, 4.4 % on ETH and 10.3 % on the S&P 500, the crypto intervals excluding one. The ladder carries an overround of 16.4 and 16.8 points on crypto, 38.0 on the S&P 500 and 47.8 on the Nasdaq 100, where too few slices pass the checks of Section 2.4 to compute the moments.



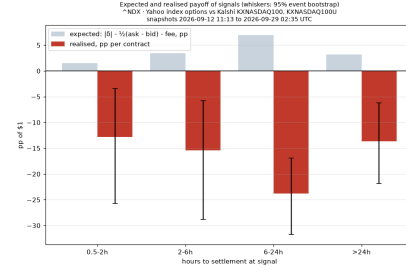
(a) BTC



(b) ETH

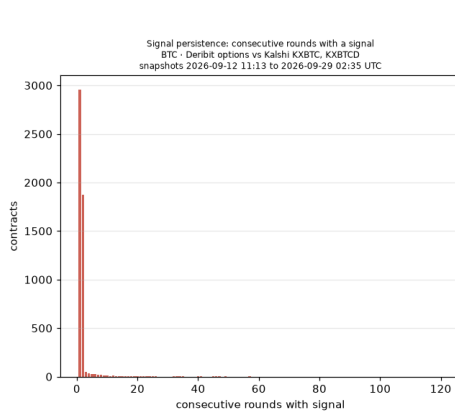


(c) S&P 500

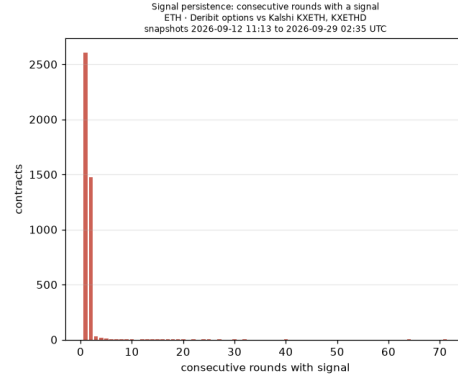


(d) Nasdaq 100

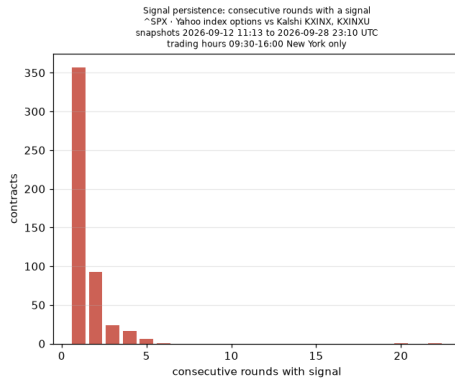
Figure 2: Expected and realised payoff by hours to settlement at the signal.



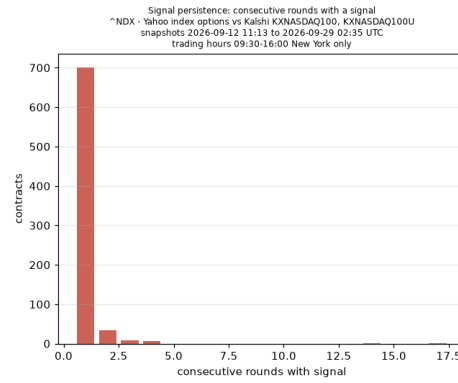
(a) BTC



(b) ETH

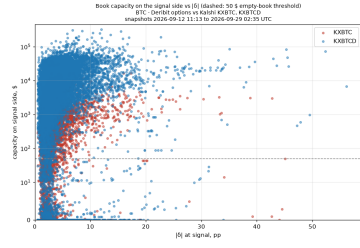


(c) S&P 500

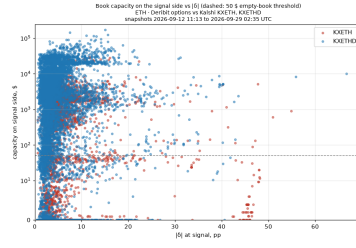


(d) Nasdaq 100

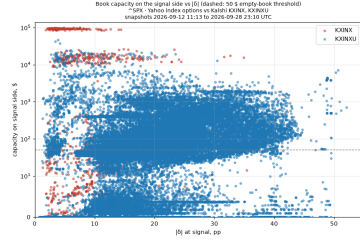
Figure 3: Number of consecutive snapshots in which a contract keeps producing a signal.



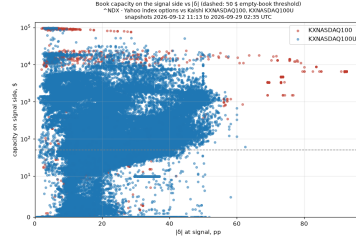
(a) BTC



(b) ETH

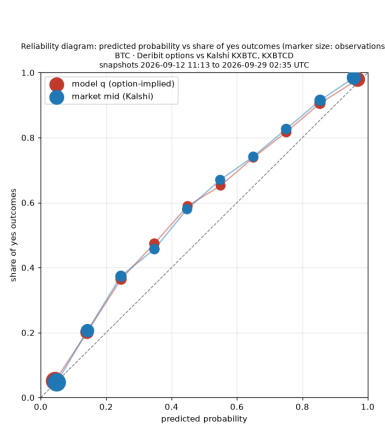


(c) S&P 500

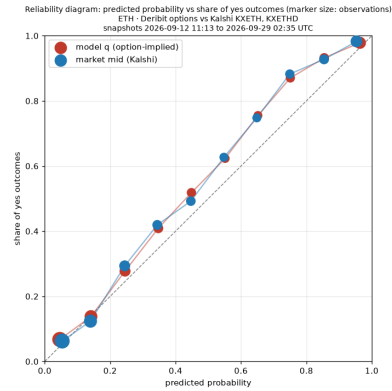


(d) Nasdaq 100

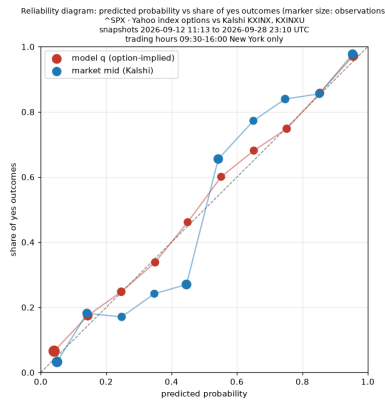
Figure 4: Book capacity on the signal side against $|\delta|$, logarithmic vertical axis; the dashed line marks the \$50 threshold.



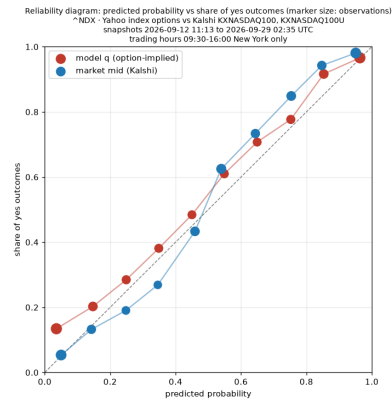
(a) BTC



(b) ETH



(c) S&P 500



(d) Nasdaq 100

Figure 5: Reliability diagrams over every contract-snapshot with a known outcome; marker area is proportional to the count.