How Overfit Is Your Search?

A Controlled Study of the Probability of Backtest Overfitting

Eugen Soloviov

July 2026

Abstract

A parameter search reports the best of many backtested configurations; the question that decides whether the report means anything is not whether the winner looks good in sample but whether the *act of selecting it* generalizes out of sample. The Probability of Backtest Overfitting (PBO), estimated by Combinatorially Symmetric Cross-Validation (CSCV), scores exactly that selection procedure. We run it on three controlled, fully seeded regimes with known ground truth. Over $T = 1000$ observations, $N = 200$ configurations, and $S = 16$ blocks — all $C(16, 8) = 12{,}870$ symmetric train/test splits per matrix, 60 Monte-Carlo matrices per regime — CSCV cleanly separates skill from luck. Under a pure null of 200 independent zero-edge strategies the in-sample winner is a coin flip out of sample: PBO $= 0.476 \pm 0.137$, its annualized in-sample Sharpe of 1.98 collapsing to 0.06 out of sample. Plant a genuine per-observation edge of 0.15 (annualized 2.38) in 20 of the 200 strategies and PBO falls to 0.001: the selected in-sample Sharpe of 3.73 retains 2.34 out of sample, essentially the true edge. A moving-average crossover grid of 170 configurations swept over a pure random walk — a search with no edge to find — returns PBO $= 0.463 \pm 0.223$ with an in-sample Sharpe of 0.97 decaying to 0.04: statistically indistinguishable from the null, which is the correct diagnosis. An edge-strength sweep traces PBO monotonically from 0.518 at zero annualized edge down through 0.205, 0.028, 0.001 to 0.000 as the planted annualized Sharpe rises from 0 to 3.17, with the selected out-of-sample Sharpe rising in lockstep. The teaching point is the null: PBO's reference value is not 0 but 0.5, because picking the in-sample best out of pure noise lands you at the out-of-sample median by symmetry. PBO near 0.5 therefore means *no out-of-sample skill in the selection*, i.e. overfitting; PBO near 0 means the search is trustworthy. We place PBO alongside the Deflated Sharpe Ratio, which deflates the winner rather than scoring the selection, and state the method's limitations — synthetic data, a single performance metric, correlated grids that cap PBO near 0.5 rather than above it, and the block count S as a modeling choice — plainly.

1 Introduction

A systematic trading study almost always ends with a selection. The researcher sweeps a grid of parameters, ranks the configurations by a backtest statistic — typically the Sharpe ratio — and reports the best one. The reported number is then, silently, the maximum of many noisy estimates, and the maximum of many zero-mean draws is not zero-mean. Bailey et al. (2014) call a literature built on unreported selection pseudo-mathematical; the mechanism is not subtle, but its consequences are routinely misread as evidence of skill.

There are two structurally different ways to correct for this, and they answer different questions. The first deflates the winner: given that N trials were run, how impressive is the single best Sharpe ratio, really? This is the route of the Deflated Sharpe Ratio (Bailey and López de Prado, 2014), which raises the significance threshold for the chosen strategy to the expected maximum a search of N zero-skill trials would have produced, building on the Probabilistic Sharpe Ratio (Bailey and

López de Prado, 2012) and paralleled by the multiple-testing haircuts of Harvey and Liu (2015) and the whole-model resampling of White's Reality Check (White, 2000). The second route does not deflate any single number; it asks whether the *selection procedure itself* generalizes: if we had split the data differently, would the in-sample winner still have been good out of sample? This is the question the Probability of Backtest Overfitting answers.

PBO, introduced by Bailey et al. (2017) and estimated non-parametrically by Combinatorially Symmetric Cross-Validation (CSCV), scores the search rather than the winner. Over many symmetric train/test partitions of the same performance matrix, CSCV repeatedly selects the in-sample-best configuration and records where that configuration lands in the out-of-sample ranking of the field. If the selection is overfit, the in-sample winner is out-of-sample mediocre; PBO is the probability that it lands in the bottom half. The subtle and instructive feature of this construction is its null value. Under no genuine skill, the in-sample winner is, by exchangeability, equally likely to land anywhere out of sample, so it sits at the out-of-sample median half the time: the null PBO is 0.5, not 0. A PBO near one half is therefore not a benign "uncertain" verdict — it is the signature of a search that has learned nothing that survives resampling.

What the primary literature does not provide is a *controlled* calibration: experiments in which the ground truth is known by construction, so that the estimator can be watched separating skill from luck rather than argued about. That is the gap this paper fills. We make no claim of a new estimator; the contribution is calibration evidence on synthetic regimes with known truth, and an honest account of what PBO does and does not tell a practitioner.

Concretely, we run CSCV on three regimes and one sweep (Section 3), all seeded and reproducible from one command:

1. **Null calibration.** $N = 200$ independent zero-edge strategies. The in-sample winner is a coin flip out of sample: PBO $= 0.476 \pm 0.137$, its annualized in-sample Sharpe of 1.98 collapsing to 0.06. This is the anchor: PBO ≈ 0.5 is what pure luck looks like.

2. **A planted edge.** A genuine per-observation edge of 0.15 (annualized 2.38) in 20 of the 200 strategies drives PBO to 0.001; the selected out-of-sample Sharpe of 2.34 is essentially the planted truth.

3. **An overfit grid on noise.** A moving-average crossover grid of 170 configurations swept over a pure random walk — no edge exists to be found — returns PBO $= 0.463 \pm 0.223$, its in-sample Sharpe of 0.97 decaying to 0.04. The estimator correctly refuses to distinguish this seductive search from the null.

4. **An edge sweep.** As the planted annualized Sharpe climbs from 0 to 3.17, PBO descends monotonically from 0.518 to 0.000 and the selected out-of-sample Sharpe tracks the true edge — a calibration curve from "pure overfit" to "fully trustworthy."

Section 2 states the CSCV/PBO algorithm exactly as we implement it, Section 3 the experimental design, Section 4 the numbers, and Sections 5–6 the interpretation and the limitations. A companion script verifies every numeric claim in this manuscript against the run's JSON output.

2 CSCV and the probability of backtest overfitting

We follow Bailey et al. (2017) throughout; the equation numbers we cite are those printed in that source. The construction is deliberately model-free: it consumes only a matrix of realized performance and returns a probability.

The performance matrix. Collect the N configurations' return series into a real matrix $\mathbf{M}$ of order $T \times N$: column n is the per-observation return vector of configuration n over the T synchronous time steps. Two conditions are required. First, $\mathbf{M}$ must be a true matrix — the same T rows for every column, observations synchronous across all N trials. Second, the chosen performance metric must be estimable on sub-samples of a column, since CSCV re-estimates it on halves of the record. Our metric is the per-split Sharpe ratio $\hat{\mu}/\hat{\sigma}$, annualized for reporting by $\sqrt{252}$ (we treat 252 observations as one year of daily data); the framework is metric-agnostic and admits Sortino, Jensen's alpha, or the Probabilistic Sharpe Ratio unchanged.

Symmetric splits. Partition the T rows of $\mathbf{M}$ into an even number S of disjoint, contiguous, equal-size blocks $\mathbf{M}_s$, each of order $(T/S) \times N$. Form every combination of $S/2$ blocks as a training set and take its complement as the test set. There are

$$C_S = \binom{S}{S/2} \tag{1}$$

such combinations (their Eq. 2.3); with $S = 16$ this is $C(16, 8) = 12{,}870$ symmetric train/test splits, each using exactly half the record to train and the disjoint other half to test. Symmetry — every split's complement is also a split — is what makes the estimator balanced and is the "S" in CSCV.

Select in sample, rank out of sample. For each combination $c \in C_S$, join the chosen blocks in original row order into the training matrix $\mathbf{J}$ and its complement into the test matrix $\bar{\mathbf{J}}$, each of order $(T/2) \times N$. Compute the performance vector on $\mathbf{J}$ and let n^* be the in-sample-optimal configuration (their step of choosing the column whose IS rank is highest). Then evaluate the *same* configuration n^* out of sample on $\bar{\mathbf{J}}$ and record its relative rank among the N configurations,

$$\bar{\omega}_c = \frac{\bar{r}^c_{n^*}}{N+1} \in (0, 1), \tag{2}$$

where $\bar{r}^c_{n^*} \in \{1, \ldots, N\}$ is the out-of-sample rank of the in-sample winner (rank N = best). If the selection generalizes, $\bar{\omega}_c$ tends toward 1; if it does not, the winner falls toward the middle or bottom of the field.

Logits and PBO. Map each relative rank to a logit (their Eq. 2.4 region),

$$\lambda_c = \ln \frac{\bar{\omega}_c}{1 - \bar{\omega}_c}, \tag{3}$$

so that $\lambda_c > 0$ iff the in-sample winner is above the out-of-sample median and $\lambda_c \leq 0$ iff it is at or below it. The Probability of Backtest Overfitting is the fraction of splits in which selection fails to beat the median,

$$\mathrm{PBO} = \frac{1}{C_S} \sum_{c \in C_S} \mathbf{1}\{\lambda_c \leq 0\} = \mathrm{Prob}[\bar{\omega}_c \leq \tfrac{1}{2}], \tag{4}$$

the discrete form of their integral $\int_{-\infty}^{0} f(\lambda)\, d\lambda$ over the logit distribution f. This is a Monte-Carlo estimate of the abstract quantity of their Definition 2.2 — the probability that the in-sample-optimal strategy ranks below the out-of-sample median — computed non-parametrically from data rather than from a distributional model.

Why the null is one half. Suppose no configuration has genuine skill: the N columns of $\mathbf{M}$ are exchangeable draws from the same noise process. Then on any split the in-sample winner n^* is, by symmetry, equally likely to occupy any out-of-sample rank, so $\bar{r}^c_{n^*}$ is (in expectation across splits) uniform on $\{1, \ldots, N\}$, $\bar{\omega}_c$ is uniform on $(0, 1)$, and λ_c is symmetric about 0. The fraction of splits with $\lambda_c \leq 0$ therefore converges to $\frac{1}{2}$ (Bailey et al., 2017). Selecting the in-sample winner out of pure noise is literally a coin flip relative to the field. This is the fixed point every reading of PBO must start from: PBO ≈ 0.5 is the *no-skill* verdict, PBO ≈ 0 is trustworthy, and only a systematic PBO *above* $\frac{1}{2}$ (in-sample success predicting out-of-sample failure) signals something worse than noise — an anti-persistent or cost-driven inversion that our zero-edge searches, sitting at $\frac{1}{2}$, do not manufacture.

Degradation and loss, two companion diagnostics. CSCV yields, for each split, the pair $(R^c_{n^*}, \bar{R}^c_{n^*})$ — the in-sample winner's in-sample and out-of-sample performance. Regressing the second on the first across all splits,

$$\bar{R}^c_{n^*} = \alpha + \beta\, R^c_{n^*} + \varepsilon^c, \tag{5}$$

gives the *performance-degradation slope* β; the paper reports $\beta < 0$ in most cases it studies, "more in-sample performance $\Rightarrow$ less out-of-sample performance." In our controlled setting this negative slope is primarily regression to the mean induced by in-sample selection — it appears even in the memoryless iid null ($\beta = -0.20$) — with the serial-memory effects Bailey et al. (2014) discuss for real financial series an additional contributor. Separately, the *probability of loss* is the share of splits in which the selected strategy's out-of-sample performance is negative, $\text{Prob}[\bar{R}^c_{n^*} < 0]$. These answer different questions from PBO and from each other — rank-consistency (low PBO) does not imply positive out-of-sample return, and a negative slope appears even for genuine edges — so we report all three and read them together, not as substitutes.

Relation to deflating the winner. PBO and the Deflated Sharpe Ratio (Bailey and López de Prado, 2014) attack the same selection bias from opposite directions. DSR is parametric and single-number: it corrects the significance threshold of the one chosen Sharpe ratio for the number of trials N, using an extreme-value estimate of the expected maximum under the null. CSCV is non-parametric and procedure-level: it resamples the train/test split and scores how often selection survives, so the combination count C_S plays for CSCV the role the trial count N plays for DSR — both convert "how hard did we search?" into a correction. The DSR paper frames the two as explicitly complementary. López de Prado (2018) operationalizes the same combinatorial idea inside a purging/embargo scheme as Combinatorial Purged Cross-Validation; Arian et al. (2024) benchmark CSCV and CPCV against walk-forward and k-fold in a synthetic controlled environment and find the combinatorial methods dominate on false-discovery control. We return to the DSR comparison in Section 5.

3 Experimental design

All experiments are generated and analyzed by a single Python harness (`scripts/run_all.py`, with the estimator in `scripts/pbo.py`), under Python 3.14.6 with NumPy 2.4.3. Every random stream is seeded, so every number below is bit-reproducible by one command. The CSCV estimator is fully vectorized over the 12,870 splits: block sums and sums of squares are formed once per matrix and combined by a $C_S \times S$ indicator, so the whole split ensemble is a pair of matrix products. Ground

truth is known by construction in each regime — we know which configurations have skill because we planted them (or planted none).

Unless noted, each regime uses $T = 1000$ observations (about four years of daily data), $N = 200$ configurations, $S = 16$ blocks, and 60 Monte-Carlo matrices, and we report the mean and standard deviation of PBO across those 60 matrices together with the mean degradation slope, mean probability of out-of-sample loss, and the mean in-sample and out-of-sample Sharpe of the selected strategy (annualized by $\sqrt{252}$).

Regime 1: null (zero edge). Each matrix is $T \times N$ of independent standard-normal returns. Every true Sharpe ratio is exactly zero; the in-sample winner is pure luck. This regime calibrates the estimator: we expect PBO ≈ 0.5 and the selected strategy's out-of-sample Sharpe to collapse toward zero.

Regime 2: planted edge. We draw the same standard-normal matrix and add a constant per-observation drift $s_{\text{true}} = 0.15$ to 20 of the 200 columns, giving those strategies a genuine annualized Sharpe of 2.38 (derived, $0.15\sqrt{252}$). A real, robust edge exists and is shared across a block of configurations, so resampling the split should not dislodge it; we expect PBO ≈ 0.

Regime 3: overfit grid on noise. This is the trap the method is built to catch. We generate one pure random walk (increments iid standard normal) and sweep a moving-average crossover grid over it: for fast window $f \in \{2, 4, \ldots, 20\}$ and slow window $s \in \{24, 30, \ldots, 120\}$ with $s > f$, the strategy holds position $\text{sign}(\text{MA}_f - \text{MA}_s) \in \{-1, 0, +1\}$ and earns the strictly causal next-bar return (the position formed at t earns the return at $t+1$). The grid yields $K = 170$ configurations. Because the underlying series is a driftless random walk, no configuration has any true edge; the in-sample winner is a lucky window pair. A naive analyst, seeing a smooth equity curve for the best crossover, would call it a strategy. We expect PBO near the null 0.5 — and, crucially, *not* above it, because the grid's configurations are strongly correlated (neighboring windows hold nearly the same positions), which is a modeling point we take up in Section 6.

Sweep: PBO versus edge strength. To trace the calibration curve, we repeat the planted-edge construction with per-observation edges $s_{\text{true}} \in \{0, 0.03, 0.06, 0.10, 0.15, 0.20\}$ (annualized 0 to 3.17), 20 planted among 200, recording mean PBO and the selected out-of-sample Sharpe at each level.

Overfit-grid detail. For one seeded random-walk grid we also report the full single-matrix diagnostics — PBO over all 12,870 splits, the best in-sample Sharpe in the grid, the median out-of-sample Sharpe of the selected strategy, the degradation slope, the probability of loss, and the median logit — to show the mechanics on a concrete case rather than an average.

4 Results

4.1 The three regimes: PBO separates skill from luck

Table 1 is the core result. The three regimes are engineered to be visually indistinguishable to a naive analyst — each produces a best-in-class equity curve with a respectable in-sample Sharpe — yet PBO assigns them sharply different verdicts, and the verdicts are correct by construction.

Table 1: The three controlled regimes, each summarized over 60 Monte-Carlo matrices with $T = 1000$, $S = 16$ ($C(16, 8) = 12{,}870$ splits per matrix). PBO is reported as mean $\pm$ standard deviation across the 60 matrices; the remaining columns are means. The selected strategy's Sharpe ratios are annualized by $\sqrt{252}$. Ground truth: the null and overfit-grid searches have no true edge (correct verdict: PBO near 0.5); the planted regime has a genuine edge in 20 of 200 configurations (correct verdict: PBO near 0). The overfit grid uses $N = 170$ correlated configurations rather than 200 independent ones.

	Null (zero edge)	Planted edge	Overfit grid (noise)
Ground truth	no skill	real edge	no skill
Configurations N	200	200	170
PBO	0.476 ± 0.137	0.001	0.463 ± 0.223
Degradation slope β	-0.20	-0.52	-0.95
Prob[OOS loss]	0.48	0.00	0.47
Selected IS Sharpe (ann.)	1.98	3.73	0.97
Selected OOS Sharpe (ann.)	0.06	2.34	0.04

The null regime is the anchor. Selecting the best of 200 zero-edge strategies produces an average annualized in-sample Sharpe of 1.98 — a number that would headline a pitch deck — but it evaporates to 0.06 out of sample, and PBO sits at 0.476, just 0.024 below the theoretical one half — about a fifth of one Monte-Carlo standard deviation (0.137), i.e. sampling noise around the exact value the exchangeability argument above predicts. The out-of-sample-loss probability is 0.48: the selected strategy is a coin flip to lose money out of sample, exactly as a no-skill winner should be. This is the estimator reproducing its own null: picking the in-sample champion out of noise buys nothing that survives the split.

The planted regime is the clean positive. A per-observation edge of 0.15 in 20 configurations gives PBO = 0.001: in essentially every one of the 12,870 splits, the in-sample winner is also out-of-sample above median. The mechanics are visible in the Sharpe pair. The selected in-sample Sharpe is 3.73 — inflated above the true annualized edge of 2.38 by the same selection bias that fools the naive analyst — but out of sample it settles to 2.34, within rounding of the planted truth. The out-of-sample-loss probability is 0.00. A real edge shared across a block of configurations is a broad plateau in the performance surface, and resampling the split cannot knock the winner off a plateau; that is precisely the geometry PBO rewards.

The overfit grid is the result that matters most, because it is the case a practitioner actually encounters. A moving-average crossover grid over a pure random walk yields an in-sample Sharpe of 0.97 and a smooth equity curve, yet PBO is 0.463 ± 0.223 — statistically indistinguishable from the null and nowhere near zero — and the out-of-sample Sharpe is 0.04 with a loss probability of 0.47. CSCV correctly reads the search as no better than picking the luckiest of 170 noise draws, which is exactly what it is. Note the wide standard deviation (0.223, versus 0.137 for the null): with 170 heavily correlated configurations there are effectively far fewer independent bets, so PBO is noisier matrix-to-matrix, a point we return to in Section 6.

One honest nuance is the degradation slope β, which is negative in all three regimes (-0.20, -0.52, -0.95) and is *not* monotone in overfitting — the genuine-edge regime has a steeper slope than the null. The slope's sign is a generic feature of selecting on a shared sample and does not by itself classify a search; PBO's *level* does. We therefore treat the slope as a descriptive companion, never as the headline verdict.

Table 2: PBO versus planted edge strength, 20 planted among $N = 200$, $T = 1000$, $S = 16$. Each row is a mean over Monte-Carlo matrices. The planted edge is reported annualized ($s_{\text{true}}\sqrt{252}$); the selected strategy's out-of-sample Sharpe is likewise annualized. PBO descends monotonically from the no-edge anchor of 0.518 to 0.000 as the edge grows, and the realized out-of-sample Sharpe rises to meet, then slightly exceed, the planted truth. Each point averages 30 Monte-Carlo matrices.

Planted Sharpe (annual)	PBO	Selected OOS Sharpe (ann.)
0.00	0.518	−0.05
0.48	0.435	0.19
0.95	0.205	0.81
1.59	0.028	1.65
2.38	0.001	2.48
3.17	0.000	3.29

4.2 The edge-strength sweep: a calibration curve

Table 2 traces PBO as the planted edge strengthens from nothing to an annualized Sharpe of 3.17. The curve is monotone and steep, and it is anchored exactly where the theory says: at zero edge PBO is 0.518, on top of one half; the selected strategy's out-of-sample Sharpe is −0.05, indeed a coin flip around zero.

The transition is informative. A genuine annualized edge of 0.48 — a real, if modest, strategy — barely moves PBO (0.435) and leaves the selected out-of-sample Sharpe at 0.19: an edge that small is swamped by selection over 200 configurations, and CSCV correctly declines to certify it. By an annualized edge of 0.95 PBO has dropped to 0.205 and the out-of-sample Sharpe to 0.81; by 1.59, PBO is 0.028 and the out-of-sample Sharpe 1.65; and from 2.38 onward PBO is at or below 0.001 with the out-of-sample Sharpe (2.48, then 3.29) tracking and slightly exceeding the planted level as selection begins to help rather than hurt. This is the shape a calibrated overfitting diagnostic should have: flat and high where edges are indistinguishable from luck, falling sharply through the region where a real edge starts to dominate the search noise, and pinned near zero once the edge is unmistakable.

4.3 Overfit-grid detail: the mechanics on one matrix

Averages can hide mechanics, so Table 3 opens up a single seeded random-walk grid. Here PBO is 0.573 — this particular draw is, if anything, slightly worse than a coin flip — computed over all 12,870 splits of the 170 configurations. The best in-sample Sharpe in the grid is an eye-catching 2.33 annualized, and its median out-of-sample Sharpe across splits is −0.22: the luckiest crossover in sample loses money out of sample at the median split. The degradation slope is −0.92, the probability of out-of-sample loss 0.63, and the median logit −0.25 (below zero, consistent with PBO > 0.5).

The gap between the best in-sample Sharpe (2.33) and its median out-of-sample outcome (−0.22) is the whole story of backtest overfitting in one matrix. The 2.33 is real in the sense that it happened; it is meaningless in the sense that selecting it does not transfer. PBO = 0.573 says so numerically, and it says so without any distributional assumption, any benchmark, or any knowledge that the underlying series was a random walk — it reads the failure of selection to generalize directly off the split ensemble.

Table 3: Full CSCV diagnostics for one seeded moving-average grid ($K = 170$ crossover configurations) over a single pure random walk, $T = 1000$, $S = 16$, all 12,870 splits. The best in-sample Sharpe is a compelling annualized 2.33; out of sample it is a median loss. Every diagnostic points the same way — the search has no edge — and the numbers are the raw ingredients behind the overfit-grid column of Table 1.

Diagnostic	Value
PBO (over 12,870 splits)	0.573
Best in-sample Sharpe (ann.)	2.33
Median OOS Sharpe of selected (ann.)	−0.22
Degradation slope β	−0.92
Prob[OOS loss]	0.63
Median logit	−0.25

5 Discussion

5.1 The null is the lesson

The single most important number in this paper is not a result but a reference: PBO's null value is 0.5. Almost every misreading of the diagnostic comes from importing the intuition of a p-value, where small is good and 0.5 is uninformative. For PBO, 0.5 is not uninformative — it is the precise signature of a search that overfits, because the in-sample winner drawn from noise is a coin flip out of sample by exchangeability (Section 2). Our three regimes make this operational. The null sits at 0.476, the overfit grid at 0.463, the genuine edge at 0.001. A practitioner who computed PBO on the overfit grid and saw 0.46 might, reading it as a p-value, feel reassured that it is "well below one"; the correct reading is that 0.46 is indistinguishable from the coin-flip null and the search should be discarded. The actionable rule is a distance from $\frac{1}{2}$, not a distance from 1: Bailey et al. (2017) suggest rejecting searches with PBO > 0.05, a demanding cutoff that our planted edge clears (0.001) and both no-edge searches fail decisively.

The edge sweep (Table 2) shows this is a smooth, calibrated transition rather than a threshold artifact: PBO does not jump from 0.5 to 0; it slides through 0.435, 0.205, 0.028 as the edge grows, and the selected out-of-sample Sharpe rises continuously alongside it. The region that should unsettle practitioners is the top of the curve, where genuine but small edges (annualized 0.48, PBO 0.435) are statistically indistinguishable from luck over a search of 200 configurations. PBO is honest about this: it declines to certify an edge the search cannot actually resolve, rather than manufacturing confidence.

5.2 PBO scores the search; DSR deflates the winner

The two corrections are complementary, and the planted regime shows why one wants both. DSR (Bailey and López de Prado, 2014) would take the selected in-sample Sharpe of 3.73 and ask how impressive that single number is given that 200 trials were run, deflating it toward the expected maximum under the null. PBO never looks at the winner's magnitude; it asks whether the *procedure* of selecting an in-sample winner survives resampling, and answers 0.001. These are different questions with different failure modes. DSR can be fooled by a correlated grid if fed the raw trial count (the 170 crossovers are nowhere near 170 independent trials); PBO sidesteps trial counting entirely by resampling the split, but pays for it by being unable to distinguish a rank-consistent zero-return strategy from a rank-consistent profitable one — which is exactly why we report the

probability of loss alongside it. The clean division of labor: DSR prices the winner, PBO audits the selection, and the loss probability checks that rank-consistency actually pays. On our regimes they agree in direction, and their agreement across such different machinery is itself reassuring.

5.3 A real edge is a plateau

The reason the planted edge scores PBO = 0.001 while the overfit grid scores 0.463 is one of *redundancy*. A genuine edge shared across 20 of the configurations makes many of them good at once, so whichever one wins in sample, an equally-good one wins out of sample, and the rank is stable across splits. An overfit "edge" is a lone winner — one lucky configuration whose in-sample lead is noise — and noise is exactly what resampling the split knocks over: the in-sample leader is somewhere else out of sample. (A caveat on the geometry: in our design the 20 edge-carrying configurations are randomly scattered, not spatially adjacent, so the experiment varies edge *redundancy*, not parameter-space *adjacency*. The plateau-versus-spike picture below is the natural interpretation of that redundancy, not something the design tests directly.) This is the same plateau intuition that motivates preferring robust interior optima to knife-edge parameter settings, and PBO is, in effect, a quantitative redundancy detector: low PBO certifies that the winning *set*, not merely the winning point, generalizes. It is a useful mental model for why one should trust a search whose top configurations cluster in parameter space and distrust one whose winner is an isolated outlier.

6 Limitations

We state the boundaries of these results plainly.

- **Synthetic data, by design.** All three regimes are synthetic — iid Gaussian returns, a constant planted drift, and a driftless random walk — chosen so that ground truth is known exactly. That is the point of a calibration study and also its limit. Real returns have fat tails, volatility clustering, and structural breaks, and CSCV's sub-sample Sharpe estimates lean on approximate IID-normal behavior on each half; our experiments say nothing about market realism, only that the estimator behaves correctly when its assumptions hold.

- **We never observe** PBO > 0.5**.** PBO can in principle exceed one half — the perverse regime in which in-sample success actively predicts out-of-sample failure — but that requires an anti-persistence or transaction-cost structure that inverts the sign of the in-sample signal out of sample. We do not build such a structure, so our no-edge searches sit *at* $\frac{1}{2}$ rather than above it. Frictionless synthetic grids cannot exhibit the cost-driven inversion that produces the most alarming PBO values in practice; a realistic study with costs would be needed to exercise that region.

- **Correlated grids inflate PBO's sampling variance.** The overfit grid's 170 crossover configurations are strongly correlated, so the effective number of independent bets is far below 170. This inflates the variance of PBO across matrices (0.223 versus 0.137 for the independent null) — but does *not* push its level above one half (that requires the anti-persistence of the previous point; the uncorrelated iid null also sits at one half). It means the overfit grid's PBO is pinned near 0.5 rather than being driven higher. Correlation among trials is a genuine modeling feature of real parameter sweeps, not a flaw, but it does mean PBO's discriminating power lives entirely in the gap between ≈ 0.5 and ≈ 0, not in any excursion above 0.5.

- **The block count S is a modeling choice.** We fix $S = 16$ (12,870 splits). Larger S yields more splits and finer resolution at the cost of shorter, noisier training halves per block; smaller S the reverse. We did not sweep S, and CSCV inherits the general cross-validation tension that the number and size of folds trades bias against variance. The purged and embargoed variant of López de Prado (2018) is the principled treatment for serially dependent, overlapping-label settings that our block-contiguous splits do not address.
- **A single performance metric.** We score every split by the Sharpe ratio. The framework is metric-agnostic, and a path-dependent metric (return over maximum drawdown) or a higher-moment-aware one (the Probabilistic Sharpe Ratio) could rank configurations differently and yield a different PBO on the same matrix. Our conclusions are Sharpe-specific; we did not test metric sensitivity.

7 Conclusion

The Probability of Backtest Overfitting asks the right question about a parameter search: not "is the winner good?" but "does selecting the winner generalize?" On three controlled regimes with known ground truth, Combinatorially Symmetric Cross-Validation answers cleanly. It pins a pure null at PBO = 0.476, the coin-flip value the theory demands, with the selected in-sample Sharpe of 1.98 collapsing to 0.06 out of sample. It drives a genuine planted edge to PBO = 0.001, with the out-of-sample Sharpe of 2.34 recovering essentially the true edge of 2.38. And it correctly refuses to be impressed by a moving-average grid swept over a random walk, scoring PBO = 0.463 — indistinguishable from the null — even as that search advertises an in-sample Sharpe of 0.97 and, on a single seed, a best configuration reaching an annualized 2.33 that turns into a median out-of-sample loss of -0.22. The edge sweep ties the picture together: PBO descends monotonically from 0.518 to 0.000 as the annualized edge climbs from 0 to 3.17, a calibration curve from pure overfit to fully trustworthy.

The lesson to carry away is the null. PBO is not a p-value; its reference point is 0.5, not 0, because the in-sample winner drawn from noise is a coin flip out of sample. A search with PBO near one half has learned nothing that survives resampling, however seductive its equity curve, and a search with PBO near zero has found an edge that lives on a plateau rather than a spike. Read together with the Deflated Sharpe Ratio — which deflates the winner where PBO audits the selection — and with the probability of loss — which checks that rank-consistency actually pays — PBO gives a practitioner a direct, assumption-light answer to the question that decides whether a backtest search means anything at all.

Reproducibility. All code, tests, and outputs accompany this paper: `scripts/run_all.py` regenerates `results/results.json` from fixed seeds (Python 3.14.6, NumPy 2.4.3) with the CSCV estimator in `scripts/pbo.py`; `scripts/check_paper_numbers.py` verifies every numeric claim in this manuscript against that file and fails on any mismatch; `tests/` contains deterministic invariant tests for the estimator (the CSCV combinatorics, the null-near-one-half property, monotonicity of PBO in edge strength, and the overfit-grid degradation).

References

Hamid R. Arian, Daniel Norouzi Mobarekeh, and Luis A. Seco. Backtest overfitting in the machine learning era: A comparison of out-of-sample testing methods in a synthetic controlled

environment. *Knowledge-Based Systems*, 305:112477, 2024. doi: 10.1016/j.knosys.2024.112477.

David H. Bailey and Marcos López de Prado. The sharpe ratio efficient frontier. *Journal of Risk*, 15(2):3–44, 2012. doi: 10.21314/JOR.2012.255.

David H. Bailey and Marcos López de Prado. The deflated sharpe ratio: Correcting for selection bias, backtest overfitting and non-normality. *Journal of Portfolio Management*, 40(5):94–107, 2014. doi: 10.3905/jpm.2014.40.5.094.

David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. 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, 2014. doi: 10.1090/noti1105.

David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. The probability of backtest overfitting. *Journal of Computational Finance*, 20(4):39–69, 2017. doi: 10.21314/JCF.2016.322.

Campbell R. Harvey and Yan Liu. Backtesting. *Journal of Portfolio Management*, 42(1):13–28, 2015. doi: 10.3905/jpm.2015.42.1.013.

Marcos López de Prado. *Advances in Financial Machine Learning*. John Wiley & Sons, Hoboken, NJ, 2018. ISBN 978-1-119-48208-6.

Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000. doi: 10.1111/1468-0262.00152.