SHOULD YOU TRAIN FOR DECISIONS? A DECISION-ADEQUACY TEST AND STEIN SHRINKAGE FOR PREDICT-THEN-OPTIMIZE

Anonymous authors
Paper under double-blind review

ABSTRACT

A cost model for an optimization problem can be fitted by least squares and then optimized (two-stage) or trained on the decisions it induces (decision-focused). Which is better depends on an unknown misspecification, so the two are mixed by cross-validation or by hand, with no guarantee. We ask instead whether the model is decision-adequate: would decision-focused training change its decisions? We test this with a Hausman-type statistic, the gap between the two fits in the geometry of the decision loss rather than the Wald geometry, which overweights directions that change no decision. Smoothing the SPO+ loss gives both fits a closed-form covariance, and the same gap is the unknown coefficient of the risk along the path between them, so one statistic says whether to move and how far: a closed-form James-Stein weight that, asymptotically in the risk of that geometry, is never much worse than the better fit and, under a regime condition, beats the decision-focused fit. In preregistered experiments the test has size 0.051 under correct specification and rejects in most splits on two electricity markets where decision-focused training cuts regret by up to a factor of six. In practice this gives a p-value for decision adequacy and, for methods that interpolate between the fits, a closed-form step that requires no refits and no prior knowledge of which fit is better. Every failed criterion is reported: the shrinkage gain is small, the rule trails cross-validation where the smoothing scale is coarse, and the test over-rejects under heavy tails at small samples and reads skewed noise as misspecification.

1 INTRODUCTION

In a predict-then-optimize pipeline a model predicts the costs of an optimization problem from a context, and the decision solves that problem at the predicted costs. Our running example is an electricity trader who each day chooses, within a capacity, which of the day's half-hour slots to use; a slot is worth its price, and prices are predicted from features of the day such as load and wind generation. The decision is a knapsack solved at the predicted prices, and the loss is the regret: the value given up by acting on the prediction rather than on the true prices. Two electricity markets supply our real data.

The price model can be fitted in two ways. The two-stage way fits it by least squares and optimizes afterwards; the decision-focused way trains it on the decisions it induces, through a convex surrogate of the regret such as SPO+ (Elmachtoub & Grigas, 2022). Theory says which is better when the misspecification is known: least squares under correct specification (Hu et al., 2022; Elmachtoub et al., 2023), the decision loss otherwise (Elmachtoub et al., 2025), and under local misspecification, which shrinks with the sample size so that neither fit wins outright, it depends on the degree and direction (Lan et al., 2025).

In practice the degree is not observed. Methods that move between the two fits, by a blending weight, a trust region, a blended gradient or a selector (Loke et al., 2026; Yang et al., 2025; Jeon et al., 2025; Iglesias et al., 2026), set the step by cross-validation or by hand, which costs refits and held-out data and gives no guarantee on the decisions. None of them answers the prior question: would decision-focused training change the decisions at all? If not, the difference between the fits is noise.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

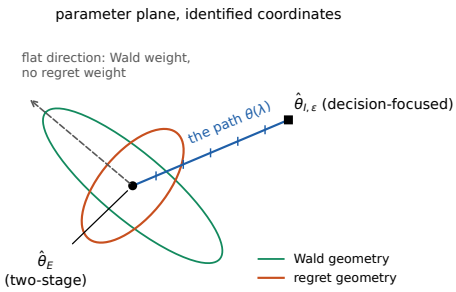

Figure 1: The two fits, the path between them and the two geometries; the dashed direction changes no decision, so it has Wald weight but no regret weight.

We make that question a hypothesis test and let its statistic set the step. A model is *decision-adequate* when decision-focused training would not change its decisions in the population. Under this null the two fits estimate the same parameter, so a quadratic form in their difference is a test statistic: a Hausman-type test (Hausman, 1978) that compares least squares with the decision-focused fit. The same quadratic form is the unknown coefficient of the risk along the segment between the fits, so it also sets how far to move, through a James-Stein weight that, asymptotically and in the risk of the decision geometry, is never much worse than the better fit and, under a regime condition (in the Hausman case, at least four effective directions), beats the decision-focused fit.

One choice makes or breaks this construction: the geometry in which the difference is measured (Figure 1). The Wald geometry standardises each direction by its own noise. Take a knapsack item that is almost never worth using: almost no decision depends on its coefficient, yet the Wald statistic counts that direction like any other, and because the decision loss is nearly flat there, its noise is estimated badly and the statistic becomes unstable. The geometry of the decision loss, its Hessian in the parameters, gives that direction weight near zero and the largest weight to the directions in which decisions turn. Our first design used the Wald geometry; Section 3.2 records what went wrong.

Contributions. (1) A smoothed SPO+ loss, convex and Fisher consistent under symmetric noise at every scale, that gives both fits a closed-form sandwich covariance (Section 3.1). (2) A decision-adequacy test in the regret geometry, consistent against every decision-relevant alternative, with a free and a calibrated reference (Sections 3.2 and 3.3). (3) A closed-form James-Stein step with an oracle inequality against the better fit, dominance over the decision-focused fit at effective dimension at least four in the Hausman case, a safe variant from two and a last-layer version (Sections 3.4 to 3.5). (4) A preregistered evaluation, every failed criterion reported: a toy, two PyEPO benchmarks, three kinds of null, two electricity markets and a neural predictor (Section 4). (5) For practice, a p-value for decision adequacy and, for methods that interpolate between the fits, a closed-form step that requires no refits and no prior knowledge of which fit is better (Section 6).

2 Setting

We observe i.i.d. pairs (z_i, c_i), $i = 1, \dots, n$, of a context $z \in \mathbb{R}^p$ and a cost vector $c \in \mathbb{R}^d$, and use a fixed feature map $\phi(z) \in \mathbb{R}^{q_0}$ with an intercept. The decision solves a linear program over a polytope $X \subset \mathbb{R}^d$: $x^*(c) \in \arg\min_{x \in X} c^\top x$, with value $z^*(c) = \min_{x \in X} c^\top x$. The predictor is linear, $\hat{c} = \Theta\phi(z)$, with $\theta = \mathrm{vec}(\Theta) \in \mathbb{R}^q$. In the running example X is the knapsack polytope, c the negative slot prices and $\phi(z)$ features of the day such as load and wind generation.

The two fits. The two-stage estimator is least squares, $\hat{\theta}_E = \arg\min_\theta \sum_i |c_i - \Theta\phi_i|^2$. The decision-focused estimator minimises the SPO+ loss of Elmachtoub & Grigas (2022), $\ell_I(\theta; z, c) = -z^*(2\hat{c} - c) + 2\hat{c}^\top x^*(c) - z^*(c)$, a convex surrogate of the regret $c^\top x^*(\hat{c}) - z^*(c)$. Some directions of θ never change a decision (when every decision uses the same number of items, as in top-k selection, adding one constant to every predicted cost leaves every $\arg\min$ in place). We compare the fits only on the complementary *identified coordinates*, with projector Π (formally, Π removes the

lineality space of the normal fan of X; Appendix G.1), and anchor every decision-focused estimator at $\hat{\theta}_E$ on the other directions, which changes neither its decisions nor its loss.

Decision adequacy. Let θ_E^* and θ_I^* be the population targets of the two fits and $\Delta = \Pi(\theta_I^* - \theta_E^*)$. The model is *decision-adequate* when $\Delta = 0$: it may be misspecified, but not in a way that decision-focused training would exploit. This is our null, $\mathrm{H}_0 : \Delta = 0$. It implies equal population decisions but is not implied by them: a rescaling of the identified part changes no decision and moves Δ (Remark S8-H). And it concerns the targets, not regret: under skewed noise with a correct mean the test rejects although least squares is decision-optimal (Section 4.1).

3 THE METHOD

3.1 SMOOTHING THE SPO+ LOSS

A Hausman-type test needs the joint covariance of the two fits, and the step needs the Hessian of the decision loss. Neither has a closed form for the raw SPO+ loss, whose empirical Hessian vanishes almost everywhere, nor for the regret, which is piecewise constant in θ. One device removes both obstacles: entropic smoothing of the value function. For $\epsilon > 0$ let

$$z_\epsilon(w) = -\epsilon \log \sum_{v \in V(X)} \exp(-w^\top v/\epsilon), \tag{1}$$

the sum over the vertices of X, and $x_\epsilon = \nabla z_\epsilon$, the Gibbs average of the vertices (Berthet et al., 2020); as $\epsilon \to 0$ they tend to z^* and x^*. The sum is computed exactly in every problem of this paper, by recursions for top-k selection and for paths and by enumeration or a dynamic programme for the knapsacks (Appendix A). The smoothed SPO+ loss smooths both terms at the same scale,

$$\ell_\epsilon(\theta; z, c) = -z_\epsilon(2\hat{c} - c) + 2\hat{c}^\top x_\epsilon(c) - z_\epsilon(c), \tag{2}$$

and the smoothed decision-focused fit $\hat{\theta}_{I,\epsilon}$ is its anchored empirical minimiser. The scale is $\epsilon = \epsilon_0 s_r$, with s_r the pooled residual standard deviation of least squares, and $\epsilon_0 = 0.1$ was fixed before any benchmark ran (its effect: Section 4.3).

Lemma 1 (Lemma S1, proved in Appendix G). *z_ϵ is concave, and ℓ_ϵ is convex and C^∞ in θ with the positive semidefinite Hessian $4\phi\phi^\top \otimes (-\nabla^2 z_\epsilon(2\hat{c} - c))$, whose expectation is positive definite on the identified coordinates under Assumption A. If the conditional law of $c - \Theta_0\phi$ given z is centrally symmetric, the smoothed population risk is minimised at θ_0 for every ϵ, so H_0 holds for the raw and the smoothed pair alike; under a conditional density and a unique SPO+ target, the smoothed target converges to it as $\epsilon \to 0$.*

In words: smoothing keeps what matters in SPO+ (convexity, Fisher consistency under symmetric noise) and adds a Hessian; because both terms are smoothed at one scale, Fisher consistency holds at every ϵ, so the scale is a nuisance, not a bias. With the Hessian the two fits are jointly asymptotically normal, with a sandwich covariance built from the empirical scores (Theorem 1-S, Proposition 1-S); the difference $\hat{\Delta} = \Pi(\hat{\theta}_{I,\epsilon} - \hat{\theta}_E)$ has covariance $V_\Delta = V_{EE} + V_{II} - V_{EI} - V_{IE}$.

3.2 THE TWO GEOMETRIES

Two matrices can weight $\hat{\Delta}$. The *Wald geometry* V_Δ^+ standardises it by its own noise; it is the geometry of inference on θ. The *regret geometry* W is the Hessian at $\hat{\theta}_E$ of the smoothed empirical decision loss $\mathcal{L}_\epsilon(\theta) = \frac{1}{n}\sum_i c_i^\top x_\epsilon(\Theta\phi_i)$, projected on the identified coordinates and reduced to its positive semidefinite part (the dropped mass is reported per cell). To second order, moving θ by Δ costs $\frac{1}{2}\Delta^\top W \Delta$ in regret: W prices a displacement in decisions.

The two disagree exactly where decisions care least: a direction in which the decision loss is flat has W-weight near zero and, since the sandwich is near singular there, V_Δ^+-weight near infinity. Our first design worked in V_Δ^+. On our toy its shrinkage constant was 1.7 to 8.5 times the admissible bound of the regret geometry, and on the benchmarks the calibrated critical values of its statistic varied by orders of magnitude across replicates of one cell, tracking the condition number of the Hessian (Figure 5, Appendix A); the comparison is across two designs, an observation rather than a theorem.

3.3 THE TEST

The regret-weighted decision-adequacy statistic is $T_W = n\,\hat{\Delta}^\top W \hat{\Delta}$. It needs no pseudo-inverse and no rank rule; the sandwich enters only its reference distribution.

Proposition 1 (Proposition S7, proved in Appendix G)**.** *Under the assumptions of Theorem 1-S with a consistent plug-in W: (a) under H_0, $T_W \to_d \sum_j \omega_j Z_j^2$, with Z_j independent standard normal and ω_j the eigenvalues of WV_Δ. (b) Under a fixed alternative, $T_W/n \to_p \Delta^\top W \Delta$, positive whenever $W^{1/2}\Delta \neq 0$: the test is consistent against every decision-relevant alternative and blind to the others by design. (c) A Satterthwaite reference (Satterthwaite, 1946) matching the first two moments of this limit is exact when the nonzero ω_j are equal. (d) The null-calibrated reference refits both estimators on 199 pseudo-samples $c_i^* = \hat{\Theta}_E \phi_i + s_i e_i$ (e_i the least-squares residuals, s_i random signs, ϵ and W fixed) and rejects above the 190th order statistic, a wild bootstrap whose level and consistency are proved under conditionally symmetric errors (Proposition S7-d). (e) Under the local alternatives of Assumption L-S the noncentrality of T_W is $\delta^\top W \delta$, the coefficient of Theorem 1: test and rule read the same quantity.*

In words: under decision adequacy T_W is a weighted sum of squared normals; the calibrated reference, primary here, is exact in the limit but costs 199 refits, and the moment-matched one is free but approximate. The null is symmetric, which is why the sign flip reproduces it exactly (Lemma 1) and why the test reads skewed noise as misspecification.

3.4 THE STEP

The path and its risk. Let $\theta(\lambda) = \hat{\theta}_E + \lambda \hat{\Delta}$ for $\lambda \in [0, 1]$: the two-stage fit at $\lambda = 0$, the decision-focused fit at $\lambda = 1$. We study it in the local limit experiment of Lan et al. (2025) (Assumption L-S), where the identified difference of the targets is $\delta/\sqrt{n}$, so that neither fit wins outright.

Theorem 1 (Theorem S2, the risk along the path)**.** *Under Assumption L-S the asymptotic W-weighted risk of $\theta(\lambda)$ is*

$$\rho_W(\lambda) = (1-\lambda)^2(\delta^\top W \delta + \mathrm{tr}(WV_{EE})) + \lambda^2 \mathrm{tr}(WV_{II}) + 2\lambda(1-\lambda)\mathrm{tr}(WV_{EI}), \tag{3}$$

a parabola in λ, minimised at $\lambda^* = (\delta^\top W \delta + \mathrm{tr}(WV_{EE}) - \mathrm{tr}(WV_{EI}))/(\delta^\top W \delta + \mathrm{tr}(WV_\Delta))$, which tends to one as $|\delta| \to \infty$ and is positive under the null whenever least squares is inefficient.

In words: the risk along the path is a parabola whose minimiser moves toward the decision-focused fit as the misspecification grows; its one unknown, $\delta^\top W \delta$, is the noncentrality of the test statistic (Proposition 1(e)). Since $\mathbb{E}[T_W] \to \delta^\top W \delta + \mathrm{tr}(WV_\Delta)$, the quantity $T_W - \mathrm{tr}(W\hat{V}_\Delta)$ estimates it without bias in the limit, and plugging it in gives the step.

The rule. With $\hat{q} = \max\{T_W - \mathrm{tr}(W\hat{V}_\Delta), 0\}$,

$$\hat{\lambda} = \mathrm{clip}\Big(\frac{\hat{q} + \mathrm{tr}(W\hat{V}_{EE}) - \mathrm{tr}(W\hat{V}_{EI})}{\hat{q} + \mathrm{tr}(W\hat{V}_\Delta)}, 0, 1\Big), \qquad \hat{\theta}_{\mathrm{SURE}} = \hat{\theta}_E + \hat{\lambda}\hat{\Delta}. \tag{4}$$

Writing $\tau_W = \mathrm{tr}(W\hat{V}_\Delta)$, $C = V_{II} - V_{IE}$ and $\kappa_W = \mathrm{tr}(WC)$, the rule reads $\hat{\theta}_{\mathrm{SURE}} = \hat{\theta}_{I,\epsilon} - w\hat{\Delta}$ with $w = \min\{1, \kappa_W / \max(T_W, \tau_W)\}$: a positive-part James-Stein weight that shrinks the decision-focused fit toward the two-stage fit, its constant κ_W set in the regret geometry. The name refers to Stein's unbiased risk estimate of the fixed-λ risk, which $\hat{\lambda}$ minimises (Remark S3-N).

Theorem 2 (Theorem S3: oracle inequality and dominance)**.** *Under Assumption L-S with $B = V_\Delta^{1/2} W V_\Delta^{1/2} \succ 0$ and $\kappa_W > 0$, let ν_W be the largest eigenvalue of $B^{-1/2}(AB)_s B^{-1/2}$ with $A = V_\Delta^{1/2} W C V_\Delta^{-1/2}$, $(AB)_s$ the symmetric part of AB, and $\bar{\tau} = \max(\tau_W, \kappa_W)$. (c) Oracle inequality. For every local alternative μ, the asymptotic W-risk of $\hat{\theta}_{\mathrm{SURE}}$ exceeds $\min_{\lambda \in [0,1]} \rho_W(\lambda)$, hence that of the better endpoint, by at most $4\kappa_W \nu_W\, \mathbb{E}[\mathbf{1}\{T_W > \bar{\tau}\}/T_W]$, which is $O(|\mu|^{-2})$ and at most $4\kappa_W\nu_W/\bar{\tau}$. (b) Dominance. If $\kappa_W \geq 4\nu_W$, then $\hat{\theta}_{\mathrm{SURE}}$ has strictly smaller asymptotic W-risk than $\hat{\theta}_{I,\epsilon}$ for every μ. In the Hausman case $V_{IE} = V_{EE}$ this reads $\mathrm{tr}(WV_\Delta)/\lambda_{\max}(WV_\Delta) \geq 4$: an effective dimension of at least four, and the constant four is exact.*

In words: the rule is never much worse than the better of the two fits, with a remainder that vanishes as the misspecification grows, and it needs no condition for this. When the regime condition holds (in the Hausman case, at least four effective directions), it is strictly better than the decision-focused fit, for every degree of misspecification.

A safer constant and a smaller one. The Stein argument of Theorem 2(b) covers every constant a in $w = \min\{1, a/\max(T_W, \tau_W)\}$ up to $\tau_{\max} = 2(\kappa_W - 2\nu_W)$, which is positive exactly when the effective dimension exceeds two in the Hausman case (Theorem S6). *SURE-safe* uses $a = \min(\kappa_W, \tau_{\max})$: it dominates the decision-focused fit whenever $\tau_{\max} > 0$, equals SURE from effective dimension four, returns the decision-focused fit when $\tau_{\max} \le 0$, and shrinks less under adequacy. *JS-W*, added after the first benchmark round, uses $a = \tau_{\max}/2$ in Hansen's form $\min\{1, a/T_W\}$ and dominates whenever $\kappa_W > 2\nu_W$ (Corollary S6-JS). We recommend SURE, the preregistered primary; JS-W where the effective dimension may lie between two and four, at a price under adequacy at small n; and SURE-safe below two.

The procedure in one place. Algorithm 1 collects the method: two fits, one Hessian, one statistic, and three estimators that differ only in a constant.

Algorithm 1 Decision-adequacy shrinkage from two fits (compact form; full form: Algorithm 2)

Input: the fits $\hat{\theta}_E$, $\hat{\theta}_{I,\epsilon}$; the sandwich blocks $\hat{V}_{EE}$, $\hat{V}_{II}$, $\hat{V}_{EI}$; the regret weight W; the number of refits B.
Output: the calibrated p-value, $\hat{\lambda}$, the constants $(\kappa_W, a_{\text{safe}}, a_{\text{JS}})$ and the three estimators.

1: $\hat{\Delta} \leftarrow \Pi(\hat{\theta}_{I,\epsilon} - \hat{\theta}_E)$; $\hat{V}_\Delta \leftarrow \hat{V}_{EE} + \hat{V}_{II} - \hat{V}_{EI} - \hat{V}_{EI}^\top$; $C \leftarrow \hat{V}_{II} - \hat{V}_{EI}^\top$; $T_W \leftarrow n\hat{\Delta}^\top W \hat{\Delta}$; $\tau_W \leftarrow \operatorname{tr}(W\hat{V}_\Delta)$; $\kappa_W \leftarrow \operatorname{tr}(WC)$.
2: $\nu_W \leftarrow \lambda_{\max}(B_W^{-1/2}(AB_W)_s B_W^{-1/2})$ with $B_W = \hat{V}_\Delta^{1/2} W \hat{V}_\Delta^{1/2}$, $A = \hat{V}_\Delta^{1/2} W C \hat{V}_\Delta^{-1/2}$; $\tau_{\max} \leftarrow 2(\kappa_W - 2\nu_W)$; $a_{\text{safe}} \leftarrow \min(\kappa_W, \tau_{\max})$; $a_{\text{JS}} \leftarrow \kappa_W - 2\nu_W$.
3: **for** $b = 1, \ldots, B$: draw Rademacher signs s_i, set $c_i^* = \hat{\Theta}_E \phi_i + s_i e_i$, refit both estimators with ϵ and W held fixed, and compute T_W^{*b}.
4: $p \leftarrow B^{-1}\#\{b : T_W^{*b} \ge T_W\}$ (the test rejects at level 0.05 when T_W exceeds the 190th of $B = 199$ ordered T_W^{*b}).
5: $\hat{q} \leftarrow \max(T_W - \tau_W, 0)$; $\hat{\lambda} \leftarrow \operatorname{clip}((\hat{q} + \operatorname{tr}(W\hat{V}_{EE}) - \operatorname{tr}(W\hat{V}_{EI}))/(\hat{q} + \tau_W), 0, 1)$.
6: the three estimators $\hat{\theta}_a = \hat{\theta}_{I,\epsilon} - \min\{1, a/\max(T_W, \tau_W)\}\hat{\Delta}$ with $a = \kappa_W$ (SURE), a_{safe} (SURE-safe) and a_{JS} (JS-W, with T_W in place of the max), each equal to $\hat{\theta}_{I,\epsilon}$ unless $a > 0$ and $\kappa_W > 0$.
7: **return** p, $\hat{\lambda}$, $(\kappa_W, a_{\text{safe}}, a_{\text{JS}})$ and the three estimators.

Lines 1, 2, 5 and 6 need no refit; line 3, the calibrated reference, is the whole cost of the method, and the moment-matched reference of Proposition 1(c) replaces lines 3 and 4 by a closed form.

3.5 The last layer of a neural network

A learned representation is a feature map, so test and rule apply to the last layer of a network when the representation does not depend on the sample the heads are fitted on. We split the training sample into halves; on one, a multilayer perceptron with hidden widths $(64, h)$ is trained by mean squared error, and its last hidden layer $h(z)$, dead units dropped, gives $\phi(z) = (h(z), 1)$; on the other, with ϕ fixed, the pipeline above runs unchanged: the least-squares head is *FT-none*, the smoothed SPO+ head *FT-full*, and $\hat{\theta}_E + \hat{\lambda}\hat{\Delta}$ is *SURE-FT*. We swap the halves and average; conditional on the representation every statement above applies (Remark 1).

4 Experiments

Every experiment was preregistered: grids, seeds, estimators, criteria and thresholds were committed and timestamped before any result existed, and every outcome, PASS or FAIL, is reported (Tables 31, 32 and 9 list them with their conditions); claim identifiers appear as pointers in parentheses. Regret is the normalized regret of Tang & Khalil (2024) on held-out data, the summed regret divided by the summed optimal values; a *cell* is one benchmark, noise model, degree of misspecification and sample size; cell values are means over seeds; and *the better endpoint* of a cell is the smaller of the two fits' cell means.

Table 1: Rejection rates at the nominal 0.05 of the four references; within a row the references are computed on identical draws. The calibrated T_W test is the test of this paper; F is the first preregistered design (Wald statistic, F reference). A Part is a preregistered experiment; n/a: not computed in that Part. Every cell of every Part: Table 8.

Null (degree one unless stated)	tests	cal. T_W	cal. Wald	Satt.	F	pointer
Correct specification, regret-weighted test (Part 2E)	1200	0.051	0.048	0.014	0.575	E1 PASS
Symmetric nulls N1 and N2, toy and knapsack (Part 2H)	1200	0.0375	0.050	0.012	0.257	H1 PASS
Misspecification in the lineality space (N4), toy	300	0.043	0.037	0.013	0.057	reported
Heavy-tailed omitted identified component (N2-s), toy	300	0.097	0.060	0.023	0.060	H1-s FAIL
N2-s at $n = 5000, 10000$, fresh seeds	200	0.035	0.025	0.020	0.020	H1-s-n PASS
Real features, empirical residual law (energy prices, $C = 12, 48$)	200	0.030	n/a	0.105	n/a	R5 PASS

Problems. A top-k toy as a mechanism check; two PyEPO benchmarks (Tang & Khalil, 2024), shortest path on a 5×5 grid (96 identified coordinates) and a ten-item knapsack (60), with polynomial misspecification of degree $\{1, 2, 4, 6\}$ (degree one is correct specification), multiplicative or Gaussian noise and ten seeds per cell; and two real markets, the PredOpt energy-price knapsack on Irish prices (Mandi et al., 2024; Ifrim et al., 2012), 48 half-hour items at capacities 12, 24, 48 and 96, and a knapsack over the 24 hours of Nord Pool day-ahead prices (Lago et al., 2021) at capacities 6, 12, 24 and 48, ten splits each. Baselines: PyEPO's SPO+, PFYL (Berthet et al., 2020), DBB (Vlastelica Pogančić et al., 2020), DDR (Loke et al., 2026) with its cross-validated weight, and CV-select, the 5-fold cross-validated choice between the two-stage fit and SPO+.

4.1 Does the test have the right size and power?

Yes, except under heavy tails at small samples, and at the cost of the refits; the free moment-matched reference is conservative on the benchmarks and oversized on real features (Table 1). Under correct specification the calibrated T_W test has pooled size 0.051 over 1200 null replicates of the twelve benchmark cells, no cell above 0.08, and rejects every degree-six replicate at $n = 1000$ (E1, E2). On the same draws the F-referenced Wald test of our first design rejects in 0.575 of null replicates and the uncorrected chi-square reference in 0.696; the calibrated Wald test has the right size (0.048), but on the knapsack its critical values have a coefficient of variation across replicates of 0.61 to 9.02, against 0.34 to 0.49 for T_W (Figure 5).

A second experiment misspecified the model in three ways that leave the targets equal (heteroskedastic symmetric noise, an omitted symmetric component, a shift in directions that change no decision): the calibrated test has pooled size 0.0375, no cell above 0.08, while the F-referenced Wald test rejects in 0.453 on the knapsack (H1). It fails under one heavy-tailed null, an omitted cubed-Gaussian component in the identified coordinates, with size 0.097 at $n \leq 2000$ and 0.035 at $n = 5000$ and 10000 on fresh seeds (H1-s FAIL, H1-s-n). On real features, with the empirical residual law as the null, the calibrated test's size is 0.030 (R5).

Asymmetry and cost. The null is symmetric: with a correct mean and skewed noise least squares is decision-optimal, yet the targets differ, and the test rejects in 0.98 to 1.00 of toy replicates and in 0.14 to 0.96 on the knapsack at skewness 6.18, at no cost there (the decision-focused fit is within 0.9845 to 1.0012 of the two-stage fit); on skewed real data a rejection says that the targets differ, not that the mean is wrong. The rule is free once the fits and the sandwich exist; the calibrated reference costs 199 refits, 26 to 737 seconds per replicate on the benchmarks against 1.3 to 3.6 for CV-select (Table 4).

4.2 Does the rule track the better fit?

Yes where its condition holds, and the misses are where the theory puts them. Figure 2 places every benchmark cell by the share of replicates in which the dominance condition of Theorem 2(b) holds and by the regret of the rule relative to the better endpoint. Under adequacy (degree one) SURE is at the two-stage fit in all 14 cells, within a ratio of 1.0014 (D1). Where the condition holds in at least 0.7 of replicates, 32 of the 56 cells, SURE is within 1.0043 of the better endpoint (D2). SURE-safe is within 1.0168 in all 48 cells with $n \geq 500$ (D3). Every miss of SURE in these cells is a knapsack cell at degree four or six with $n \leq 500$ and effective dimension below four, up to 1.0654 at $n = 200$.

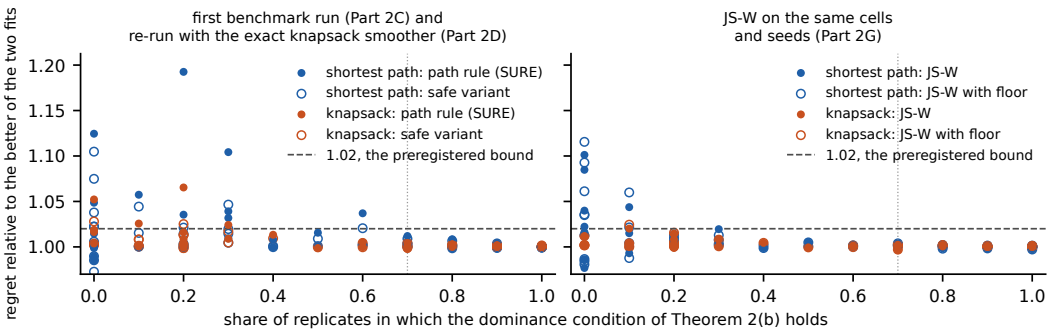

Figure 2: The regime map, one point per benchmark cell: left SURE and SURE-safe (Parts 2C and 2D), right JS-W on the same cells and seeds (Part 2G); dashed: the preregistered bound 1.02. The misses sit at the left, where the condition rarely holds.

Table 2: Irish energy-price knapsack, shared linear head, normalized test regret, cell means over ten splits: decision-focused training pays at every capacity and the shrinkage rows sit on the decision-focused row (last-layer rows: Table 10).

	energy-price knapsack, capacity			
Test regret	12	24	48	96
Shared linear head ($q_M = 9$)				
two-stage (ETO)	0.2894	0.2623	0.2101	0.1311
smoothed decision-focused ($\hat{\theta}_{I,\epsilon}$)	0.2321	0.1889	0.1362	0.0858
$\hat{\theta}_{\mathrm{SURE}}$	0.2326	0.1902	0.1364	0.0860
PyEPO SPO+ (chosen by CV-select in every split)	0.2325	0.1899	0.1357	0.0835
calibrated T_W rejection share	0.9	0.9	0.9	1.0
effective dimension	1.18	1.48	1.52	1.55

JS-W, run later on the same cells and seeds, is within 1.02 of the better endpoint in 23 of the 24 cells where the condition holds in fewer than 0.7 of replicates (maximum 1.0202; G1). It pays under adequacy at $n = 200$, where it is 1.0202 and 1.0146 of the two-stage fit against 1.0005 and 1.0014 for SURE, as Corollary S6-JS predicts.

4.3 Is it better than cross-validation?

Not reliably, and the reason is the smoothing scale, not the rule. Against CV-select, SURE is within 1.02 in 35 of the 40 benchmark cells with $n \geq 1000$ and below it in 31. The five misses are all at degree six under multiplicative noise, where SURE sits on its smoothed endpoint ($\hat{\lambda}$ between 0.96 and 0.99) and that endpoint trails PyEPO's SPO+ by 2 to 11 percent, the largest miss being 1.1096 on the knapsack (D5 FAIL). Under Gaussian noise at degree six the smoothed endpoint beats SPO+ by 4 to 5 percent on the shortest path, and SURE beats cross-validation by that margin. In the five missing cells PFYL is the best column; DDR with its cross-validated weight is 0.996 to 1.044 of the two-stage fit across the 56 cells.

The preregistered scale $\epsilon_0 = 0.1$ is too coarse for these cells: a sweep on fresh seeds closes the gap at $\epsilon_0 = 0.02$, where SURE is 0.97 to 0.99 of cross-validated selection in four of the five cells at no cost under adequacy (0.9996 to 1.0003 of the two-stage fit at every scale), and fails its claim in the fifth (F1 FAIL). Of three preregistered rules for the scale, the rule on the eigen-mass of W failed, the in-sample SPO+ loss rule passed on the benchmarks and failed on the real prices, and the fixed 0.02 with JS-W passed on fresh seeds (G3, G4 FAIL; G9 to G11, R1c FAIL; G7, G8). We recommend no rule: every main table keeps 0.1, the record supports 0.02 to 0.05, and the question is open (Table 9).

4.4 Real electricity markets

On both markets decision-focused training pays, the test says so, and the shrinkage adds almost nothing on top. With one linear head shared by the 48 Irish items (Table 2), the smoothed decision-

focused head has 0.65 to 0.80 of the two-stage regret and the calibrated test rejects in 0.9 to 1.0 of splits. SURE follows the path to its end ($\hat{\lambda}$ at least 0.99, within 0.7 percent of the smoothed endpoint) and is within 1.02 of cross-validated selection except at capacity 96, where the smoothed endpoint trails SPO+ (1.0302, R1 FAIL). The effective dimension of this head is 1.18 to 1.55, below two, so the dominance theory is silent: SURE-safe and JS-W return the decision-focused fit, and only the oracle inequality covers SURE (R2, R4).

Item-specific heads (56 to 120 parameters) raise the effective dimension to 4.4 to 7.0 and give the shrinkage room. The smoothed head now has 0.90 to 0.96 of the two-stage regret, the test rejects in every split at capacities 24 to 96 and in 0.40 at 12, and JS-W has 1.0004 to 1.0155 of the smoothed endpoint's regret, below it in 6 of 40 splits: no gain (R4d FAIL). Two mechanisms explain this. Least squares is not efficient under the skewed price noise (residual skewness -3.29 in the cost units of the oracle, the cost being the negative of the price), so the Hausman identity $V_{IE} = V_{EE}$ fails and the regime condition of Theorem 2 fails at capacities 24 to 96, where κ_W/ν_W is 1.05 to 3.84 although the effective dimension exceeds four (at 12 it holds in 0.7 of splits). And the regret keeps falling to the decision-focused endpoint, still falling from step 0.75 to 1 in 31 of 40 splits (the case $\alpha > 1$ of Corollary S5(iii)), so a rule confined to the path can at best match the endpoint; SURE, whose constant is larger, retreats further and pays for it at capacity 12 (1.0464 of the better endpoint, R1d FAIL). PFYL is 0.89 to 0.92 of the two-stage fit and the best column at two capacities; carrying test and path to its Fenchel-Young surrogate is future work.

On the Nord Pool prices, preregistered as a second market, decision-focused training with item-specific heads cuts the regret by a factor of two to six (0.18 to 0.56 of the two-stage regret), and the test rejects in every split of every cell. Only at capacity 48 does the effective dimension exceed four in most splits; there JS-W's own condition $\kappa_W > 2\nu_W$ holds in every split and JS-W has 0.9978 of the smoothed endpoint's regret, below it in 6 of 10 splits: a gain of 0.2 percent, the order Corollary S6-JS predicts (R4s). SURE misses the better endpoint at capacity 12 (1.0620, R1s FAIL). At capacities 6 and 12, where the smoothed endpoint trails SPO+ by 8 and 34 percent, both rules miss cross-validated selection (1.0769 and 1.4269, R3s FAIL). At 48 both are 0.84 of it (full tables: Appendix D).

4.5 What the last layer of a network can and cannot do

The test says whether the last layer should move; only the experiments say whether moving it is enough. On the cross-fitted representation of Section 3.5 the test rejects in 1 of 60 tests under adequacy, where SURE-FT returns the two-stage head (within 1.0002 of FT-none), and the rejection share and the step rise with the degree (NL5, NL3). But on three data sets the best last-layer head trails the network trained end to end on the decision loss, by 15.9 to 38.4 percent at degree six on the shortest path, 1.8 to 7.0 percent on the Irish prices and a factor above four on the Warcraft maps (Table 10), while the test rejects in 0.80 to 1.00 of replicates (NL2, R3b, W2 FAIL). The representation, not the head, carries the misspecification. The data-chosen step is within 1.02 of the best fixed step, chosen in hindsight, in 7 of 8 cells (the miss is 1.0247; NL4 FAIL).

5 Related work

The theory this answers. SPO+ and its Fisher consistency are due to Elmachtoub & Grigas (2022), fast rates to Hu et al. (2022), the dominance of least squares under correct specification to Elmachtoub et al. (2023), a dissection of misspecification to Elmachtoub et al. (2025), and local misspecification, the setting of Theorems 1 and 2, to Lan et al. (2025); all condition on the degree and direction that we estimate, test and shrink by. Liu (2026) asks our question in tutorial form; Cameron et al. (2022) show empirically when learning before optimizing fails.

The nearest neighbours. Decision-driven regularization (Loke et al., 2026), decision-focused fine-tuning (Yang et al., 2025), gradient blending (Jeon et al., 2025), pipeline selection (Iglesias et al., 2026) and cost regularization (Spitzer & Sanson, 2026) move between the two fits with a step set by cross-validation or by hand; the trust-region radius of Yang et al. (2025) is the hand-set analogue of our step, and their two-fold scheme is our cross-fitting. Our path estimator is this family's closed-form member, with a test of whether any step is needed. Qi et al. (2026) integrate estimation with a conditional model, Bennouna et al. (2025) make the surrogate robust to misspecification,

Candogan & Foussoul (2026) design the predictor class, and Zheng et al. (2026) test a deployed decision after distributional shift; the surveys of Mandi et al. (2024) and Sadana et al. (2025) list the choice between pipelines as open.

Shrinkage and smoothing. Our rule is a positive-part James-Stein weight with its constant set in the regret geometry, in the line of efficient shrinkage toward a robust estimator in the Hausman setting (Hansen, 2016; 2017), data pooling (Gupta & Kallus, 2022), averaging under local misspecification (Cheng et al., 2019) and focused criteria (Claeskens & Hjort, 2003; Hjort & Claeskens, 2003; DiTraglia, 2016), whose focus here is the decision; Li & Ignatiadis (2025) use the same template for prediction-powered inference. The smoothing is the Gibbs form of the perturbed optimizer of Berthet et al. (2020) and the entropic Fenchel-Young loss of Blondel et al. (2020); Wilder et al. (2019) smooth with a quadratic term, Mandi & Guns (2020) with a log barrier, and the projector of Lee et al. (2026) is the hard-active-set form of our W.

6 Limitations and what changes in practice

What the guarantees are. Large-sample statements in the W-weighted risk of a local limit experiment, relative to the endpoints of the path. Their transfer to regret needs Assumption R, which fails numerically on our polynomial family (Corollary S5(iii)), and on the real prices the regret still falls from step 0.75 to 1 in 31 of 40 splits; so every claim of Section 4 is checked on regret, and the estimator is not claimed to beat decision-focused training or cross-validation in every cell.

What the test is. Its reference is exact under conditionally symmetric noise, the condition for the Fisher consistency of SPO+, so it reads asymmetry as misspecification; both real data sets are skewed and carry no size claim. It over-rejects under a heavy-tailed omitted component at $n \leq 2000$. The moment-matched reference is free but approximate, conservative on the benchmarks and oversized on real features; the calibrated one costs 199 refits. The theory is for a linear predictor on a fixed feature map; SURE-FT is conditional on a cross-fitted representation and excludes end-to-end training.

What the rule is. Below effective dimension four the step of SURE is variable and its dominance has no bite; JS-W and SURE-safe cover dimensions above two and return the decision-focused fit below, and JS-W pays under adequacy at small n. The constants pass through $V_\Delta^{-1/2}$ and are sensitive when it is ill-conditioned (SURE uses traces only). On the scored heads of the real markets SURE was 0.9908 to 1.0620 of the better fit (Nord Pool, capacities 48 and 12) and JS-W gained at most 0.2 percent.

What the smoothing is. The preregistered scale 0.1 is too coarse at degree six under multiplicative noise and at the smallest real capacities, and the rule for the scale is open. The smoothing needs an enumerable vertex family or a dynamic programme; a barrier smoothing for polytopes given by inequalities satisfies the same lemma and is exploratory (Appendix B).

What changes in practice. Fit least squares; fit the smoothed SPO+ loss from that point; compute the sandwich, W and T_W (Algorithm 1; no refits or held-out data except 199 refits for a calibrated p-value; Appendix A). Report the p-value and the effective dimension and use SURE (SURE-safe below effective dimension two, JS-W where it may lie between two and four), which sits at the point of the path with the smallest variance when T_W is small (the two-stage fit in the Hausman case) and moves toward the decision-focused fit as T_W grows. A non-rejection with little power is not proof of adequacy (Section 4.4). For heads on a learned representation, read a rejection as "the last layer should move", not "moving it is enough".

AI USE STATEMENT

Large language models were used in this work for the following purposes, matching the disclosure on the submission form. *To aid and polish writing, and to draft sections of the paper*: Claude drafted the abstract, the contribution list, the paper skeleton and the section drafts; the authors revised and approved the final text. *For retrieval and discovery*: Claude was used for locating related work and for checking the reference list against the sources; the authors read every reference. *For research ideation and execution*: the idea and the design rulings were developed by the authors; Claude Code wrote the code and the preregistration records (Amendments 9 and 11, the two dated notes and the freeze of each Part, each recording a ruling of the authors with its timing), and ran the experiments under those rulings; a review of the draft by Claude prompted the revision rounds. *For proving mathematical claims*: Claude Code wrote the theorem statements and the proofs; the authors verified them. The authors made every design decision and ruling, set the preregistered criteria, read every run report, verified the proofs and are responsible for the final text. Safeguards: every theorem carries a status (proved, assumed or cited); the key results of Theorems S2, S3 and S6 and of Proposition S7 have numerical verifications independent of the proofs, recorded in the frozen results; every number in the paper is generated from frozen summaries by the registry, with a build-time audit; every criterion was frozen and timestamped before its run, and every failure is reported. The synthetic data come from PyEPO's generators and the preregistered toy generator, which are code, not model-generated content; the real data are public benchmark files. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The synthetic data come from PyEPO's generators and the preregistered toy generator. The energy-price knapsack is the public knapsack instance of the PredOpt benchmark (Mandi et al., 2024) on the price series of Ifrim et al. (2012) (file hash recorded and checked at every load); the Warcraft maps are the public dataset of Vlastelica Pogančić et al. (2020). The work involves no human subjects and no personal data. Nothing is released beyond the code and the result files.

REPRODUCIBILITY STATEMENT

Every experiment was preregistered: criteria, grids, seeds and thresholds were committed and timestamped with a third-party anchor (OpenTimestamps) before any related result existed; evaluation code was frozen before it read any result file; rule corrections were made by numbered append-only amendments, and the round-by-round history (the amendments, the dated notes to the Parts, the time box of the Warcraft experiment and the re-run of the frozen seeds) is Appendix F. A first design preceded every experiment above: a full-space analytic Wald statistic for the nonsmooth SPO+ estimator with the shrinkage constant of the Wald geometry, preregistered, run and closed against the kill line frozen in the project charter (size FAIL, 0.56, 0.22 and 0.10 at $n = 200$, 500 and 1000; power PASS; the two regret criteria PASS; the weight criterion FAIL; an assumption gate that failed); no verdict of that design enters this paper, and its frozen Wald constant, 35 to 47 on the toy, is the numerator of the factor 1.7 to 8.5 of Section 3.2 (Table 31, first block; Appendix F). Every number in this paper traces to a named frozen summary through a machine-checked registry, and an automated audit fails the build on any numeral without a source. The code, the frozen result files, the preregistration with every verdict, and the timestamp proofs are in the anonymous supplementary repository, with absolute paths rewritten to relative ones.

REFERENCES

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, Third edition, 2003.

Omar Bennouna, Jiawei Zhang, Saurabh Amin, and Asuman E. Ozdaglar. Contextual optimization under model misspecification: A tractable and generalizable approach. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 3749–3775, 2025.

Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. Learning with differentiable perturbed optimizers. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Mathieu Blondel, André F. T. Martins, and Vlad Niculae. Learning with Fenchel-Young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

G. E. P. Box. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics*, 25(2):290–302, 1954.

Chris Cameron, Jason Hartford, Taylor Lundy, and Kevin Leyton-Brown. The perils of learning before optimizing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36(4), pp. 3708–3715, 2022. arXiv:2106.10349.

Ozan Candogan and Ayoub Foussoul. Deep flow networks: Optimization-friendly surrogates for integer predict-then-optimize. In *International Conference on Machine Learning*, 2026.

Xu Cheng, Zhipeng Liao, and Ruoyao Shi. On uniform asymptotic risk of averaging GMM estimators. *Quantitative Economics*, 10(3):931–979, 2019.

Gerda Claeskens and Nils Lid Hjort. The focused information criterion. *Journal of the American Statistical Association*, 98(464):900–916, 2003.

Francis J. DiTraglia. Using invalid instruments on purpose: Focused moment selection and averaging for GMM. *Journal of Econometrics*, 195(2):187–208, 2016.

Adam N. Elmachtoub and Paul Grigas. Smart "Predict, then Optimize". *Management Science*, 68 (1):9–42, 2022.

Adam N. Elmachtoub, Henry Lam, Haofeng Zhang, and Yang Zhao. Estimate-then-optimize versus integrated-estimation-optimization versus sample average approximation: A stochastic dominance perspective. *arXiv preprint arXiv:2304.06833*, 2023.

Adam N. Elmachtoub, Henry Lam, Haixiang Lan, and Haofeng Zhang. Dissecting the impact of model misspecification in data-driven optimization. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1594–1602, 2025. arXiv:2503.00626.

Vishal Gupta and Nathan Kallus. Data pooling in stochastic optimization. *Management Science*, 68 (3):1595–1615, 2022.

Bruce E. Hansen. Efficient shrinkage in parametric models. *Journal of Econometrics*, 190(1):115–132, 2016.

Bruce E. Hansen. Stein-like 2SLS estimator. *Econometric Reviews*, 36(6–9):840–852, 2017.

Jerry A. Hausman. Specification tests in econometrics. *Econometrica*, 46(6):1251–1271, 1978.

Nils Lid Hjort and Gerda Claeskens. Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899, 2003.

Nils Lid Hjort and David Pollard. Asymptotics for minimisers of convex processes. Preprint, arXiv:1107.3806, 1993.

Yichun Hu, Nathan Kallus, and Xiaojie Mao. Fast rates for contextual linear optimization. *Management Science*, 68(6):4236–4245, 2022.

Georgiana Ifrim, Barry O'Sullivan, and Helmut Simonis. Properties of energy-price forecasts for scheduling. In *Principles and Practice of Constraint Programming*, Lecture Notes in Computer Science, pp. 957–972. Springer, 2012.

Caio de Próspero Iglesias, Kimberly Villalobos Carballo, and Dimitris Bertsimas. Prescribe-then-select: Adaptive policy selection for contextual stochastic optimization. *Transactions on Machine Learning Research*, 2026. arXiv:2509.08194.

Haeun Jeon, Hyunglip Bae, Chanyeong Kim, Yongjae Lee, and Woo Chang Kim. Prediction loss guided decision-focused learning. *arXiv preprint arXiv:2509.08359*, 2025.

George G. Judge and Mary Ellen Bock. *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. North-Holland, 1978.

Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, and Rafał Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021.

Haixiang Lan, Luofeng Liao, Adam N. Elmachtoub, Christian Kroer, Henry Lam, and Haofeng Zhang. The bias-variance tradeoff in data-driven optimization: A local misspecification perspective. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2510.18215.

Junhyeong Lee, Sangjin Jin, and Yongjae Lee. Decision-focused learning via tangent-space projection of prediction error. In *International Conference on Machine Learning*, 2026. arXiv:2605.01361.

Sida Li and Nikolaos Ignatiadis. Prediction-powered adaptive shrinkage estimation. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Mo Liu. Decision-focused learning: When and why traditional prediction models fail. *arXiv preprint arXiv:2606.21773*, 2026.

Regina Y. Liu. Bootstrap procedures under some non-i.i.d. models. *The Annals of Statistics*, 16(4): 1696–1708, 1988.

Gar Goei Loke, Qinshen Tang, Yangge Xiao, and Xun Zhang. Decision-driven regularization: A blended model for learning and optimization. *arXiv preprint arXiv:2608.15124*, 2026.

Enno Mammen. Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, 21(1):255–285, 1993.

Jayanta Mandi and Tias Guns. Interior point solving for LP-based prediction+optimisation. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7272–7282, 2020.

Jayanta Mandi, Emir Demirović, Peter J. Stuckey, and Tias Guns. Smart predict-and-optimize for hard combinatorial optimization problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 1603–1610, 2020.

Jayanta Mandi, James Kotary, Senne Berden, Maxime Mulamba, Víctor Bucarey, Tias Guns, and Ferdinando Fioretto. Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research*, 80:1623–1701, 2024.

Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pp. 2111–2245. Elsevier, 1994.

Meng Qi, Paul Grigas, and Zuo-Jun Max Shen. Integrated conditional estimation-optimization. *Operations Research*, 74(3):1604–1625, 2026. arXiv:2110.12351.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Utsav Sadana, Abhilash Chenreddy, Erick Delage, Alexandre Forel, Emma Frejinger, and Thibaut Vidal. A survey of contextual optimization methods for decision-making under uncertainty. *European Journal of Operational Research*, 320(2):271–289, 2025.

F. E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110–114, 1946.

Victor Spitzer and Francois Sanson. Managing solution stability in decision-focused learning with cost regularization. *arXiv preprint arXiv:2601.21883*, 2026.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Bo Tang and Elias B. Khalil. PyEPO: A PyTorch-based end-to-end predict-then-optimize library for linear and integer programming. *Mathematical Programming Computation*, 16(3):297–335, 2024.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Marin Vlastelica Pogančić, Anselm Paulus, Vít Musil, Georg Martius, and Michal Rolínek. Differentiation of blackbox combinatorial solvers. In *International Conference on Learning Representations*, 2020.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Xianchao Xie, S. C. Kou, and Lawrence D. Brown. SURE estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association*, 107(500):1465–1479, 2012.

Jiaqi Yang, Enming Liang, Zicheng Su, Zhichao Zou, Peng Zhen, Jiecheng Guo, Wanjing Ma, and Kun An. DFF: Decision-focused fine-tuning for smarter predict-then-optimize with limited data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39(25), pp. 26868–26876, 2025. arXiv:2501.01874.

Minxing Zheng, Holly Wiberg, and Shixiang Zhu. Deciding when to decide: Testing operational suboptimality under distributional shift. *arXiv preprint arXiv:2608.29465*, 2026.

A Remarks and recorded notes

Remark 1 (The last-layer extension within the theory). Conditional on the representation h learned on the other half of the sample, $\phi(z) = (h_{\mathrm{active}}(z), 1)$ is a fixed feature map for the half on which the heads are fitted, and Theorem 1-S, Proposition 1-S, Theorems S2, S3 and S6 and Proposition S7 apply without change to the pair of heads with q_0 the number of active units plus one and $q_M = \dim M \cdot q_0$ ($q_M = 144$ on the 5×5 grid when all eight units are active). Cross-fitting makes the representation independent of the head sample; the average over the two directions is the reported column, and the two-direction spread is reported. The fixed-step rows $\theta(\lambda)$ are points on the same path, so the oracle inequality of Theorem S3 is a statement about SURE-FT against the best of them in the limit. Nothing is claimed about end-to-end training.

The Wald test on the benchmarks, recorded amendment. After the knapsack regret cells of the second round had been written and before any size cell had run, the per-replicate records showed the calibrated Wald test conservative at degree one with critical values spread over several orders of magnitude within one cell, the spread following the condition number of the smoothed Hessian; the uncorrected and F-referenced Wald tests over-rejected in the same replicates. The amendment moving the test to the regret geometry was recorded at that moment, with the observation and its timing, and the second round continued untouched; the third round (the T_W test) was frozen after it. The size cells of the second round recorded the mechanism at 100 replicates: mean calibrated critical values of 146 to 250342 across the twelve cells (Table 14).

The scissors (C4), recorded note. The recorded reading of C4 is the sign pattern of the cell-mean gap between $\hat{\theta}_{I,\epsilon}$ and ETO across degrees. Under the literal nonincreasing clause applied to the gap itself, the gap from degree one to degree two rises in three shortest-path cells of the first round before turning negative (multiplicative noise at $n = 100$; Gaussian noise at $n = 100$ and $n = 200$); the verdict stands under the sign reading and the note is part of the record.

The knapsack arm of the first round. With a Monte Carlo perturbed smoothing (32 common random numbers per observation) the sandwich Hessian was indefinite, its positive part near singular, and the plug-in effective dimension below two in every cell; the shrinkage rule was erratic there. This is a negative result of that smoothing, kept in the record; the second round replaced it by the exact entropic smoothing over the enumerated feasible set.

An observation on the decision-loss minimiser. An exploratory reference point, the L-BFGS minimiser of the smoothed empirical decision loss started at $\hat{\theta}_{I,\epsilon}$ (scale matched, since decisions are invariant to positive scaling), was the best column at degree six and $n \geq 1000$ on shortest path in the first round. No claim is attached.

Computation of the smoothing. For top-k selection the vertex sum of z_ϵ is an elementary symmetric polynomial computed by an $O(dk)$ recursion in log space; for the shortest path on a grid it is the soft Bellman recursion over the arcs; for the ten-item knapsack it is the log-sum-exp over the enumerated feasible subsets (at most 1024 for ten items), and for the electricity knapsacks a dynamic programme in log space over the capacity grid. Every smoothed fit is by L-BFGS with a Newton polish on the automatic-differentiation Hessian.

Numerical verification of Proposition 1. Parts (a) and (c) were verified on the toy problem at $n = 200000$ before the test was frozen: the empirical law of T_W against the eigenvalue mixture and the Satterthwaite quantile; the records are in the supplementary repository.

The Wald test on the benchmarks, per cell. The F-type reference is T/c_n against $\chi^2_{\hat{r}}$ with $c_n = (n-1)/(n-\hat{r}-2)$. In the twelve degree-one size cells of the second round (100 replicates each) it rejected 0.35 to 0.81; the calibrated Wald test held its size (52 rejections of 1200), and its mean critical value ran from 146 to 250342 across cells against chi-square quantiles of 78 and 120, the largest values in the knapsack cells whose median condition number reached 12754; where the Hessian is well conditioned the calibrated Wald test is correct.

Related work, continued: shrinkage and testing. The rule is a positive-part James-Stein weight whose constant is set in the regret geometry; its analysis follows the Stein unbiased risk estimate as used by Xie et al. (2012) and the efficient shrinkage of Hansen (2016) between an efficient and a robust estimator, which is the Hausman (1978) setting our test generalises. The moment-matched reference of Proposition 1 is the Satterthwaite approximation (Satterthwaite, 1946; Box, 1954); the null-calibrated reference is a wild bootstrap with Rademacher weights (Liu, 1988; Mammen, 1993). The risk hump of the pretest estimator between the two endpoints is classical (Judge & Bock, 1978) and is why we shrink rather than switch. *Related shrinkage methods in more detail.* Gupta & Kallus (2022) choose a shrinkage amount in stochastic optimization by an unbiased estimate of the out-of-sample objective, across decoupled subproblems without covariates; we shrink along the path between two estimators of one contextual model. In econometrics Hansen (2016; 2017) shrink an efficient estimator towards a robust one, Cheng et al. (2019) average under local misspecification, and the focused criteria of Claeskens & Hjort (2003), Hjort & Claeskens (2003) and DiTraglia (2016) are the template of our W-weighted risk, whose focus is the decision; Li & Ignatiadis (2025) use the same template for prediction-powered inference. *Datasets.* The field's recurring real datasets are the Irish energy prices (Mandi et al., 2020; 2024), the Warcraft maps (Vlastelica Pogančić et al., 2020; Berthet et al., 2020) and the CORA matching (Wilder et al., 2019); Elmachtoub & Grigas (2022) and Hu et al. (2022) report synthetic experiments only; we use the first two and, through PyEPO (Tang & Khalil, 2024), the generators of Elmachtoub & Grigas (2022). *The regret kernel as a projector.* Lee et al. (2026) show that under a fixed active set the regret gradient is the prediction error e projected onto the tangent space of the active constraints in the curvature metric, $P_H e$, and train with it; integrating, the regret there is $e^\top P_H e/2$, so their P_H is the hard-active-set form of our per-context regret kernel, of which W is the Gibbs-smoothed population version, used here as the metric of a test and a shrinkage rule rather than as a training signal; the normal directions they filter include, on our polytopes, the lineality space that Π removes; their quadratic scale and injection strength are tuning inputs, as our ϵ_0 is preregistered.

Two closed designs. Wald-geometry James-Stein (WG-JS), the positive-part James-Stein rule on the Wald statistic with the constant set in the Wald geometry, was frozen, tested and closed before the present design; a projected variant of it, the one named in the Reproducibility statement, was smoke-tested only and abandoned the same day, its preregistration never frozen; the records of both are part of the supplementary repository.

What changes in practice, continued. The long form of the last paragraph of Section 6. *Practitioners who build predict-then-optimize pipelines* today either default to the two-stage model or train

a decision-focused surrogate because the literature reports gains, without knowing whether the gain exists for their instance. The test settles this from two fits and one Hessian (the calibrated reference adds refits), and where it rejects, the closed-form step replaces the choice between the endpoints. *Authors of decision-focused learning papers* report regret tables across misspecification levels that mix cells where the population decisions cannot change with cells where they can; only the second kind can show a gain. Reporting the decision-adequacy test and the regret-geometry effective dimension next to the table separates them: on our regime map (Figure 2) the calibrated T_W test of Part 2E rejects in 0.02 to 0.08 of replicates at degree one and in every replicate of its four power cells at degree six (the calibrated Wald test of Part 2D rejects in 0.90 of one of them), and in the 32 of 56 second-round cells where the effective-dimension condition holds in at least 0.7 of replicates the estimator is within 1.0043 of the better endpoint. *Authors of methods that interpolate between the two objectives* carry a knob that sets how far to move from the two-stage model: a blending weight, a trust-region radius, a mixing schedule or a selector fitted on held-out data (Section 5). None of these is wrong; the knob now has a closed form with an oracle inequality (Theorem 2) and needs no refits. *Practitioners who fine-tune a pretrained predictor for decisions* can apply the test to the last layer: it measures whether fine-tuning that layer would change the decisions, keeps the two-stage head where it would not (SURE-FT equals FT-none at degree one, NL3) and sets the step where it would. It does not say whether moving the last layer is enough: it rejects both where the heads are close to the end-to-end networks and where they are far behind (Section 4.5).

Anyone shrinking toward a decision-relevant target should set the constant in the geometry of the decision loss: on our synthetic problem the constant of Wald-geometry James-Stein, our first design, was 35 to 47 against the admissible bound of the regret geometry, 4.5 to 27.1, that is 1.7 to 8.5 times too large cell by cell, an observation across two designs, not a theorem (Figure 4, which plots κ_W itself).

What the procedure costs. The rule is free once the two fits and the sandwich exist; the calibrated reference costs 199 refits (26 to 737 seconds per replicate on the benchmarks, against 1.3 to 3.6 for CV-select; Table 4); the refit-free Satterthwaite reference is conservative on the benchmarks (0.012 to 0.023), not on the real features (0.17 at $C = 48$).

Who does not change practice. Users whose sample size is not well above the identified dimension: the sandwich needs it, and on shortest path ($q_M = 96$) the estimator tracked the better endpoint only from $n = 1000$ on (within 1.0156 in all 16 cells, against up to 1.1925 below). And users who only want to know which endpoint is better, are willing to pay for refits, and need no guarantee or diagnostic: cross-validated selection stayed within 3.3 percent of the better of the two-stage and SPO+ fits in every first-round shortest-path cell and within 2.3 percent in every second-round cell (up to 10.8 percent in the first-round knapsack cells at the smallest n).

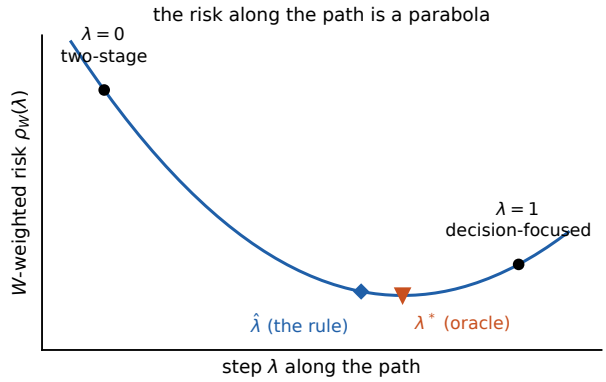

Figure 3: The risk along the path of Section 3.4: the asymptotic regret-weighted risk is a parabola whose coefficient the test statistic estimates; the rule moves to the estimated minimiser.

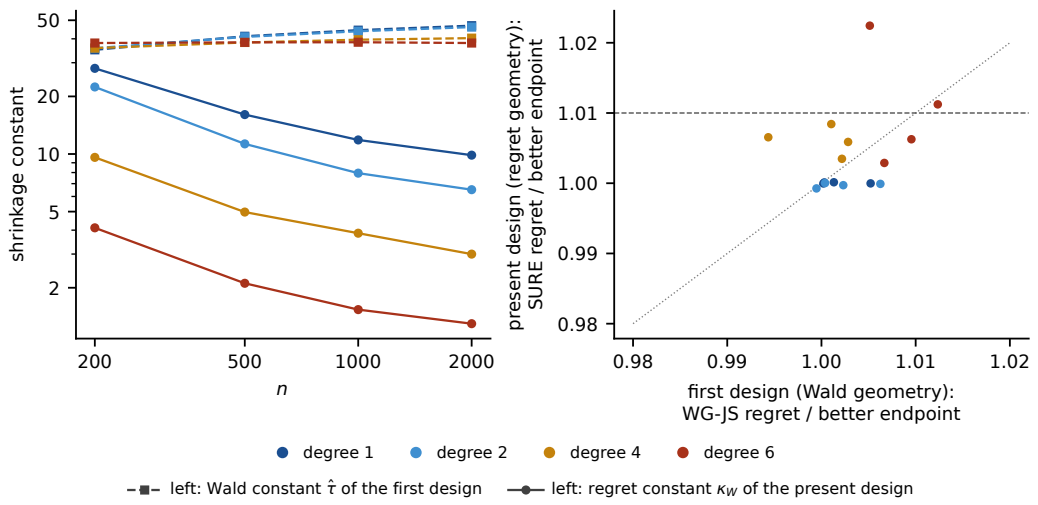

Figure 4: The two geometries on the toy problem. Left: the constant frozen by Wald-geometry James-Stein (WG-JS, dashed) against the regret-geometry constant κ_W of the present design (solid), by degree and n. Right: the regret ratio to the better endpoint of WG-JS against that of $\hat{\theta}_{\mathrm{SURE}}$, by cell; dotted, equal ratios; dashed, the preregistered tolerance for $\hat{\theta}_{\mathrm{SURE}}$.

The procedure in one place

Algorithm 1 in Section 3.4 is the compact form of the procedure; this is the full form.

Algorithm 2 Decision-adequacy shrinkage from two fits, the full form of Algorithm 1

Input: the two fits $\hat{\theta}_E$, $\hat{\theta}_{I,\epsilon}$ on n observations; the sandwich blocks $\hat{V}_{EE}$, $\hat{V}_{II}$, $\hat{V}_{EI}$; the regret weight W; the number of refits B.

Output: the p-value of the calibrated T_W test; $\hat{\lambda}$; κ_W, a_{safe}, a_{JS}; $\hat{\theta}_{\mathrm{SURE}}$, $\hat{\theta}_{\mathrm{safe}}$ and the JS-W estimate $\hat{\theta}_{\mathrm{JS}}$.

1: $\hat{\Delta} \leftarrow \Pi(\hat{\theta}_{I,\epsilon} - \hat{\theta}_E)$; $\hat{V}_\Delta \leftarrow \hat{V}_{EE} + \hat{V}_{II} - \hat{V}_{EI} - \hat{V}_{EI}^\top$; $C \leftarrow \hat{V}_{II} - \hat{V}_{EI}^\top$.
2: $T_W \leftarrow n\hat{\Delta}^\top W \hat{\Delta}$; $\tau_W \leftarrow \mathrm{tr}(W\hat{V}_\Delta)$; $\kappa_W \leftarrow \mathrm{tr}(WC)$.
3: $\nu_W \leftarrow \lambda_{\max}\big(B_W^{-1/2}(AB_W)_s B_W^{-1/2}\big)$ with $B_W = \hat{V}_\Delta^{1/2} W \hat{V}_\Delta^{1/2}$, $A = \hat{V}_\Delta^{1/2} W C \hat{V}_\Delta^{-1/2}$.
4: $\tau_{\max} \leftarrow 2(\kappa_W - 2\nu_W)$; $a_{\mathrm{safe}} \leftarrow \min(\kappa_W, \tau_{\max})$; $a_{\mathrm{JS}} \leftarrow \kappa_W - 2\nu_W$.
5: **for** $b = 1, \ldots, B$: draw Rademacher signs s_i, set $c_i^* = \hat{\Theta}_E \phi_i + s_i e_i$, refit both estimators with ϵ and W held fixed, and compute T_W^{*b}.
6: $p \leftarrow B^{-1}\#\{b : T_W^{*b} \geq T_W\}$ (the test rejects at level 0.05 when T_W exceeds the 190th of $B = 199$ ordered T_W^{*b}).
7: $\hat{q} \leftarrow \max(T_W - \tau_W, 0)$; $\hat{\lambda} \leftarrow \mathrm{clip}\big((\hat{q} + \mathrm{tr}(W\hat{V}_{EE}) - \mathrm{tr}(W\hat{V}_{EI}))/(\hat{q} + \tau_W), 0, 1\big)$.
8: $\hat{\theta}_{\mathrm{SURE}} \leftarrow \hat{\theta}_E + \hat{\lambda}\hat{\Delta}$ (equivalently $\hat{\theta}_{I,\epsilon} - \min\{1, \kappa_W / \max(T_W, \tau_W)\}\hat{\Delta}$).
9: $\hat{\theta}_{\mathrm{safe}} \leftarrow \hat{\theta}_{I,\epsilon} - \min\{1, a_{\mathrm{safe}} / \max(T_W, \tau_W)\}\hat{\Delta}$ if $\tau_{\max} > 0$ and $\kappa_W > 0$, else $\hat{\theta}_{I,\epsilon}$.
10: $\hat{\theta}_{\mathrm{JS}} \leftarrow \hat{\theta}_{I,\epsilon} - \min\{1, a_{\mathrm{JS}}/T_W\}\hat{\Delta}$ if $a_{\mathrm{JS}} > 0$ and $\kappa_W > 0$, else $\hat{\theta}_{I,\epsilon}$.
11: **return** p, $\hat{\lambda}$, $(\kappa_W, a_{\mathrm{safe}}, a_{\mathrm{JS}})$, $(\hat{\theta}_{\mathrm{SURE}}, \hat{\theta}_{\mathrm{safe}}, \hat{\theta}_{\mathrm{JS}})$.

Lines 1 to 4 and 7 to 10 use no refit; line 5 is the whole cost of the calibrated reference. The Satterthwaite reference of Proposition 1(c) replaces lines 5 and 6 by a closed form in $W\hat{V}_\Delta$.

Cost

In Table 4 both noise families and all degrees are pooled. In the Part 2D rows (the second round) the calibrated column is the calibrated Wald test; in the Part 2E rows (the size cells of the third round) it is the calibrated T_W test and the calibrated Wald test on shared refits, the column W excludes the rule, and SPO+ and CV-select were not run. The Satterthwaite reference has no timer of its own in the record: it is a closed form in the sandwich and W, with no refit. The rule is free once the two

fits and the sandwich exist: in the third round the rule together with the evaluation of the test-set regret took at most 0.053 seconds at the median of any benchmark and n. The cost of the method is the calibrated reference, which is B refits (Table 4); the Satterthwaite reference needs none and was conservative at these n (Table 8). The real-price arm took 21.8 CPU hours in all (Table 3), the calibrated test of one linear-head record 76 to 368 seconds. The Warcraft arm took 58084 to 58741 seconds of wall clock per seed, a figure that includes the hours in which the machine was asleep (Appendix D).

Table 3: The cost of the real-data arm on the energy prices as recorded (Part 4R), per cell. A linear-head record holds four smoothing scales, the calibrated test at the primary scale and the comparators; a last-layer record holds the two cross-fitting directions and the end-to-end networks.

Arm	Capacity	records	median seconds per record	CPU hours
linear head	12	10	91	0.26
linear head	24	10	136	0.38
linear head	48	10	233	0.64
linear head	96	10	376	1.05
last layer	12	10	136	0.38
last layer	24	10	179	0.50
last layer	48	10	313	0.84
last layer	96	10	481	1.31
real-feature null	12	100	97	9.73
real-feature null	48	100	242	6.75

Table 4: Seconds per replicate: the median over the replicates of the frozen cell files at each benchmark and n, as timed by the frozen drivers under pools of up to 18 processes (fit: the smoothed SPO+ fit; calibrated: the calibrated test, 199 refits; SPO+: one SPO+ fit; CV-select: 5-fold selection between the two-stage fit and SPO+).

Record	Benchmark	n	fit	sandwich	W, rule	calibrated	SPO+	CV-select
Part 2D	knapsack	200	0.14	0.011	0.016	26.2	0.56	1.31
Part 2D	knapsack	500	0.21	0.018	0.028	41.3	0.52	1.27
Part 2D	knapsack	1000	0.33	0.030	0.047	61.8	0.51	1.27
Part 2D	knapsack	2000	0.57	0.061	0.095	119	0.53	1.34
Part 2D	knapsack	5000	1.40	0.14	0.20	260	0.57	1.42
Part 2D	shortest path	5000	1.13	1.00	1.80	431	1.43	3.64
Part 2D	shortest path	10000	2.24	1.77	3.33	737	1.36	3.42
Part 2E	toy	500	0.021	0.019	0.030	8.78	not run	not run
Part 2E	toy	1000	0.018	0.018	0.029	10.4	not run	not run
Part 2E	toy	2000	0.046	0.048	0.081	17.9	not run	not run
Part 2E	knapsack	500	0.28	0.015	0.023	61.1	not run	not run
Part 2E	knapsack	1000	0.39	0.029	0.042	88.2	not run	not run
Part 2E	knapsack	2000	0.79	0.063	0.092	175	not run	not run
Part 2E	shortest path	500	0.17	0.086	0.17	63.5	not run	not run
Part 2E	shortest path	1000	0.19	0.16	0.31	74.7	not run	not run
Part 2E	shortest path	2000	0.29	0.32	0.63	126	not run	not run

GLOSSARY OF SYMBOLS AND ACRONYMS

Table 5: Symbols.

Symbol	Meaning
z, $\phi(z)$, c	context, feature map (with an intercept), cost vector
$\mathcal{X}$, $x^*(c)$, $z^*(c)$	feasible set, optimal decision and optimal value of $\min_{x\in\mathcal{X}} c^\top x$
Θ, θ	the linear predictor $\hat{c} = \Theta\phi(z)$ and its vectorisation
z_ϵ, x_ϵ, ℓ_ϵ	entropic smoothing of z^*, its gradient, the smoothed SPO+ loss
ϵ, ϵ_0, s_r	smoothing scale, $\epsilon = \epsilon_0 s_r$ with s_r the pooled residual standard deviation of the two-stage fit
$\hat{\theta}_E$, $\hat{\theta}_{I,\epsilon}$	the two-stage (least squares) fit and the minimiser of the smoothed SPO+ loss
Π, q_M	projection on the identified coordinates and their number
$\hat{\Delta}$, Δ	$\Pi(\hat{\theta}_{I,\epsilon} - \hat{\theta}_E)$ and its population counterpart
$\hat{V}_{EE}$, $\hat{V}_{II}$, $\hat{V}_{EI}$, $\hat{V}_\Delta$	sandwich blocks of the two fits and the covariance of $\hat{\Delta}$
H_ϵ	Hessian of the smoothed SPO+ risk
W	regret weight: Hessian of the smoothed decision loss at $\hat{\theta}_E$, projected, positive semidefinite part
T_W, τ_W	$n\hat{\Delta}^\top W\hat{\Delta}$ and $\mathrm{tr}(W\hat{V}_\Delta)$
$\theta(\lambda)$, $\hat{\lambda}$, $\rho_W(\lambda)$	the path $\hat{\theta}_E + \lambda\hat{\Delta}$, the estimated step, the asymptotic W-risk along the path
C, κ_W, ν_W	$V_{II} - V_{IE}$, $\mathrm{tr}(WC)$, the eigenvalue bound of Theorem 2
$\tau_{\max}$, a_{safe}, a_{JS}	$2(\kappa_W - 2\nu_W)$, $\min(\kappa_W, \tau_{\max})$, $\kappa_W - 2\nu_W$
effective dimension	$\mathrm{tr}(B)/\lambda_{\max}(B)$ with $B = V_\Delta^{1/2} W V_\Delta^{1/2}$
H_0	decision adequacy, $\Delta = 0$
B (count)	number of refits of the calibrated reference, 199

Table 6: Acronyms and row names.

Name	Meaning
ETO	estimate then optimize: the two-stage least-squares fit
IEO, IEO$_\epsilon$	integrated estimation and optimization: the decision-focused fit; with the subscript, the smoothed SPO+ minimiser
SPO+	the convex surrogate of the decision loss of Elmachtoub & Grigas (2022)
SURE	Stein's unbiased risk estimate; as a row name, the path estimator $\hat{\theta}_{\mathrm{SURE}}$ with the constant κ_W
SURE-safe	the rule with the constant a_{safe}
JS-W, JS-W(floor)	the James-Stein rule in the regret geometry with the constant a_{JS}, in Hansen's form; the variant with the floor τ_W
WG-JS (DAS in the records)	Wald-geometry James-Stein, the closed first design and negative control
CV-select	5-fold cross-validated selection between the two-stage fit and SPO+
PFYL, DBB, DDR	perturbed Fenchel-Young loss; differentiation of blackbox solvers; the blended model of Loke et al. (2026)
MSE, MLP	mean squared error; multilayer perceptron, trained with early stopping on a validation split (settings in the preregistration)
MLP-MSE, MLP-SPO+, MLP-PFYL, MLP-DBB	end-to-end networks trained with the named loss
FT-none, FT-full, SURE-FT, JS-W-FT	last-layer heads: the two-stage head, the smoothed SPO+ head, and the two rules applied to the pair
FT-all	fine-tuning of every layer for a fixed budget, without early stopping
L-BFGS	limited-memory Broyden-Fletcher-Goldfarb-Shanno, the optimiser of the smoothed fits
PyEPO, PredOpt	the benchmark library of Tang & Khalil (2024); the benchmark suite of Mandi et al. (2024)
N1 to N4, N2-s	the null generators of Part 2H (Section 4.1)
P, C, D, E, F, G, H, R, W, NL with a number	preregistered claims of Parts 1C, 2C, 2D, 2E, 2F, 2G and 2G-b, 2H, 4R, 3W and 3N
PASS, FAIL, VOID	a claim met, not met, or with its precondition not met, as recorded

B EXPLORATORY: BARRIER SMOOTHING FOR A POLYTOPE GIVEN BY INEQUALITIES

Exploratory and labelled as such: no claim, no threshold and no verdict is attached to this section; it was frozen with its code before its run, after the authors lifted its time box and set its grid, before any of its code or results existed (Section 6). For $X = \{x : Gx \le h,\ Ex = f\}$ whose vertices are the decisions, the barrier smoothing is $z_\beta(w) = \min_x [w^\top x - \beta \sum_i \log(h_i - g_i^\top x)]$ subject to $Ex = f$ (Mandi & Guns, 2020), solved by a damped Newton method from the analytic centre; $x_\beta = \nabla z_\beta$ by the envelope theorem and $\nabla^2 z_\beta = \mathrm{d}x_\beta/\mathrm{d}w$ by the implicit function theorem on the optimality system (Remark S1-G). The loss is the smoothed SPO+ loss of Section 3.1 with z_β and x_β in both terms and $\beta = \beta_0 s_r$, the convention of ϵ; nothing downstream changes: the same fit, sandwich, regret weight, rules and test. Two instances: the top-k polytope of the toy problem ($0 \le x \le 1$, $\mathbf{1}^\top x = k$) and the flow polytope of the 5×5 shortest path, neither of which needs the barrier (both have exact entropic smoothings), which is what allows the comparison on the same data (Table 7). Each of the 88863000 Newton solves of the run converged to the frozen tolerance, in at most 14 iterations.

What the rows show. At degree one SURE and JS-W with W_β are within 0.9989 to 1.0007 of ETO at the three scales, as the entropic rows are. At degree six the barrier smoothing is the coarser of the two at equal nominal scale: its minimiser has 1.022 to 1.032 (toy) and 1.300 to 1.348 (shortest path) times the regret of the entropic one at $\beta_0 = 0.1$, and 1.003 and 1.133 to 1.134 at $\beta_0 = 0.02$; the two nominal scales are not matched in bias. The rules stay within 1.0135 of the better endpoint in every barrier row. At scale 0.1 the positive semidefinite projection drops 0.000 to 0.100 of the mass of W_β against 0.022 to 0.320 for the entropic weight, and the median condition number of the smoothed Hessian is 2.2 to 15.2 against 2.6 to 1028. The calibrated T_W test with the barrier ($\beta_0 = 0.1$, 199 refits) rejected in 0 of 20 tests in the two degree-one toy cells at $n = 1000$ and $n = 2000$, a count too small to say more than that it is compatible with the nominal level.

C EXPLORATORY: PREDICTION-SPACE SURE FOR THE END-TO-END PAIR

Beyond the theory and labelled as such: for the two end-to-end networks (MSE and smoothed SPO+) let $c_E(z)$ and $c_I(z)$ be their predictions and $c(\lambda) = c_E + \lambda(c_I - c_E)$. On a calibration split held out from both networks, the per-context kernel $K_i = \nabla_u^2 [c_i^\top x_\epsilon(u)]$ at $u = c_E(z_i)$ (positive semidefinite part) plays the role of W, a paired bootstrap of both networks over ten resamples of the remaining training data gives the covariances of the predictions, and one pooled $\hat{\lambda}$ follows from the frozen rule with the traces and the quadratic form summed over the calibration contexts. Test regret along $\lambda \in \{0, 0.1, \ldots, 1\}$ and at $\hat{\lambda}$, at $n = 5000$, degrees $\{1, 2, 4, 6\}$, five seeds: the pooled $\hat{\lambda}$ is 0.28, 0.28, 0.72 and 0.78 at degrees one, two, four and six, and the blended prediction is at or below the better network in every cell. No claim.

D PREREGISTERED CLAIMS AND VERDICTS

The rounds. Every experiment was preregistered before any result existed and every verdict is in Tables 31 and 32. The criteria that were not met: on the toy, the bootstrap arm's size and the 1 percent adequacy criterion (P1b, P3); on the benchmarks, the first-round smoothing (C1, C1b, C2, C3), $\hat{\theta}_{\mathrm{SURE}}$ against cross-validation in five cells (D5), the scale sweep (F1) and two of the three scale rules (G3, G4; R1c); the last layer against end-to-end networks on three data sets (NL2, R3b, W2) and against the hand-set step (NL4, W3); the heavy-tailed null at $n \le 2000$ (H1-s); on the real prices, $\hat{\theta}_{\mathrm{SURE}}$ against cross-validated selection with the shared head at one capacity (R1), the oracle inequality with the item-specific head at one capacity (R1d) and the shrinkage gain (R4d); on the second market, the oracle inequality at one capacity (R1s) and both rules against cross-validated selection at two capacities (R3s; Tables 23 and 24); three regime claims were void (NL1, R4, W4).

Tables and sentences of the main text before the readability rewrite of 2026-09-25. The three tables below stood in the main text with these captions; the sentences that follow carry numbers absent from the rewritten main text and are kept here verbatim, each with the section it stood in.

Table 7: Part 5L, exploratory, no claim: the entropic rows at $\epsilon_0 = 0.1$ and the barrier rows at $\beta_0 \in \{0.1, 0.05, 0.02\}$ on the same data, cell means (toy: 10 seeds; shortest path, multiplicative noise: 5 seeds; test set 10000). Ratios of test regret to ETO, and of the smoothed minimiser (IEO) to a fresh PyEPO SPO+ fit; $\hat{\lambda}$ of SURE; the mass of W dropped by the positive semidefinite projection; the effective dimension; the rejection share of the Satterthwaite reference. The calibrated T_W test with the barrier at $\beta_0 = 0.1$ rejected in 0 of 20 tests in the degree-one toy cells at $n = 1000$ and $n = 2000$.

Benchmark	degree	n	smoothing	IEO/ETO	SURE/ETO	safe/ETO	JS-W/ETO	IEO/SPO+	$\hat{\lambda}$	dropped W	eff. dim.	Satt. rej.
toy	1	500	entropic 0.1	1.0057	0.9999	1.0000	1.0005	0.9965	0.075	0.138	7.8	0.00
			barrier 0.1	1.0022	1.0000	1.0000	0.9998	0.9930	0.134	0.005	15.2	0.00
			barrier 0.05	1.0033	1.0001	1.0001	1.0002	0.9941	0.095	0.064	10.4	0.00
			barrier 0.02	1.0049	0.9999	1.0000	1.0005	0.9957	0.071	0.286	6.9	0.00
toy	1	1000	entropic 0.1	1.0017	1.0000	1.0000	0.9997	0.9946	0.012	0.087	11.2	0.00
			barrier 0.1	1.0002	0.9999	0.9999	0.9998	0.9932	0.044	0.001	19.0	0.00
			barrier 0.05	1.0004	0.9999	0.9999	0.9998	0.9934	0.037	0.024	14.6	0.00
			barrier 0.02	1.0015	0.9998	0.9998	1.0002	0.9944	0.013	0.252	8.5	0.00
toy	1	2000	entropic 0.1	1.0010	1.0000	1.0000	1.0002	0.9916	0.090	0.022	15.6	0.00
			barrier 0.1	1.0006	1.0000	1.0000	0.9999	0.9912	0.093	0.000	21.6	0.00
			barrier 0.05	1.0001	0.9999	0.9999	0.9997	0.9907	0.089	0.004	19.5	0.00
			barrier 0.02	1.0008	1.0001	1.0001	0.9996	0.9914	0.073	0.128	12.3	0.00
toy	6	500	entropic 0.1	0.9293	0.9406	0.9406	0.9357	0.9935	0.705	0.126	6.5	0.50
			barrier 0.1	0.9494	0.9503	0.9498	0.9497	1.0150	0.986	0.012	4.9	1.00
			barrier 0.05	0.9375	0.9422	0.9420	0.9405	1.0023	0.893	0.055	5.9	1.00
			barrier 0.02	0.9319	0.9445	0.9443	0.9377	0.9964	0.698	0.247	6.3	0.60
toy	6	1000	entropic 0.1	0.9227	0.9291	0.9290	0.9265	0.9973	0.877	0.099	5.9	1.00
			barrier 0.1	0.9509	0.9509	0.9509	0.9509	1.0278	0.999	0.006	4.4	1.00
			barrier 0.05	0.9354	0.9365	0.9365	0.9360	1.0111	0.970	0.033	5.6	1.00
			barrier 0.02	0.9253	0.9319	0.9319	0.9294	1.0001	0.862	0.180	6.4	1.00
toy	6	2000	entropic 0.1	0.9123	0.9158	0.9158	0.9148	0.9914	0.941	0.081	6.4	1.00
			barrier 0.1	0.9417	0.9418	0.9417	0.9417	1.0233	0.999	0.007	4.7	1.00
			barrier 0.05	0.9253	0.9261	0.9261	0.9259	1.0054	0.983	0.028	5.9	1.00
			barrier 0.02	0.9152	0.9178	0.9177	0.9170	0.9945	0.943	0.150	6.4	1.00
shortest path	1	1000	entropic 0.1	1.0081	1.0003	1.0003	1.0004	0.9940	0.066	0.320	7.3	0.00
			barrier 0.1	1.0009	0.9989	0.9989	0.9992	0.9869	0.240	0.100	23.2	0.00
			barrier 0.05	1.0024	0.9992	0.9992	0.9992	0.9885	0.157	0.038	28.3	0.00
			barrier 0.02	1.0054	0.9993	0.9993	0.9994	0.9914	0.094	0.113	16.4	0.00
shortest path	1	5000	entropic 0.1	1.0022	1.0000	1.0000	1.0005	0.9832	0.042	0.142	9.9	0.00
			barrier 0.1	1.0003	1.0000	1.0000	1.0000	0.9814	0.218	0.091	25.8	0.00
			barrier 0.05	1.0014	1.0002	1.0002	1.0005	0.9825	0.133	0.014	36.6	0.00
			barrier 0.02	1.0014	1.0001	1.0001	1.0007	0.9824	0.082	0.015	26.3	0.00
shortest path	6	1000	entropic 0.1	0.6179	0.6283	0.6253	0.6216	1.0216	0.887	0.291	3.6	1.00
			barrier 0.1	0.8329	0.8329	0.8329	0.8329	1.3772	1.000	0.070	3.3	1.00
			barrier 0.05	0.7710	0.7708	0.7711	0.7710	1.2747	0.988	0.050	4.8	1.00
			barrier 0.02	0.7007	0.7063	0.7035	0.7022	1.1585	0.959	0.113	5.0	1.00
shortest path	6	5000	entropic 0.1	0.6968	0.6996	0.6990	0.6987	1.0493	0.979	0.194	5.0	1.00
			barrier 0.1	0.9060	0.9060	0.9060	0.9060	1.3643	1.000	0.051	4.4	1.00
			barrier 0.05	0.8554	0.8553	0.8554	0.8554	1.2881	0.997	0.025	6.2	1.00
			barrier 0.02	0.7896	0.7895	0.7894	0.7891	1.1891	0.989	0.043	6.2	1.00

Table 8: Rejection rates at the nominal 0.05 of the four references; within a row the references are computed on identical draws. The calibrated T_W test is the test of this paper; F is the first preregistered design (Wald statistic, F reference). The Part 2D row (the Wald design) and the Part 2E row are different null draws of the same twelve cells; Part 2E re-ran the calibrated Wald test on its own draws (0.048). A Part is a preregistered experiment; n/a: not computed in that Part.

Null (degree one unless stated)	tests	cal. T_W	cal. Wald	Satt.	F	claim
Correct specification, regret-weighted test (Part 2E)	1200	0.051	0.048	0.014	0.575	E1 PASS
Correct specification, Wald design (Part 2D)	1200	n/a	0.043	n/a	0.596	D4 PASS
Heteroskedastic noise and omitted component (N1, N2), toy	600	0.037	0.057	0.012	0.060	
N1, N2, knapsack (both noise models)	600	0.038	0.043	0.012	0.453	
N1 and N2 pooled (the claim)	1200	0.0375				H1 PASS
Misspecification in the lineality space (N4), toy	300	0.043	0.037	0.013	0.057	reported
Heavy-tailed omitted identified component (N2-s), toy	300	0.097	0.060	0.023	0.060	H1-s FAIL
N2-s at $n = 5000, 10000$, fresh seeds	200	0.035	0.025	0.020	0.020	H1-s-n PASS
Correct specification, in-sample loss rule's scale (Part 2G-c)	1200	0.051	n/a	n/a	n/a	E1-r PASS
Real features, empirical residual law (energy prices, $C = 12, 48$)	200	0.030	n/a	0.105	n/a	R5 PASS

Table 9: The five cells in which $\hat{\theta}_{\mathrm{SURE}}$ at the preregistered scale misses cross-validated selection (D5), through four later experiments and the real prices; values below one beat cross-validation. The scale sweep (Part 2F) and the fixed scale (Part 2G-b) ran on fresh seeds, the two scale rules (Parts 2G, 2G-c) on the seeds of the first run.

Ratio to cross-validated selection (degree six, multiplicative noise)	knapsack, $n =$ 1000	2000	5000	shortest path, $n =$ 5000	10000	claim
$\hat{\theta}_{\mathrm{SURE}}$ at the preregistered scale 0.1 (Part 2D)	1.1082	1.0908	1.1096	1.0243	1.0230	D5 FAIL
smoothing gap $\hat{\theta}_{I,\epsilon}$ / SPO+ at 0.1	1.0807	1.0601	1.0751	1.0551	1.0456	
Scale sweep: $\hat{\theta}_{\mathrm{SURE}}$ at 0.02	1.0368	0.9864	0.9852	0.9725	0.9821	F1 FAIL
smoothing gap at 0.02	0.9893	0.9834	0.9839	0.9685	0.9812	
Scale chosen by the dropped-mass rule: $\hat{\theta}_{\mathrm{SURE}}$	1.0841	1.0721	1.0586	1.0175	1.0201	G3 FAIL
JS-W at that scale	1.0695	1.0699	1.0580	1.0171	1.0181	G4 FAIL
Fixed scale 0.02: JS-W	0.9977	0.9795	0.9910	0.9611	0.9618	G7 PASS
JS-W at 0.02 over ETO in the degree-one counterparts	1.0020	1.0016	1.0005	1.0008	0.9995	G8 PASS
Scale chosen by the in-sample SPO+ loss rule: JS-W (G9 to G11)	0.9990	0.9887	0.9856	0.9721	0.9701	PASS
Real energy prices, capacity $C =$	12	24	48	96		
$\hat{\theta}_{\mathrm{SURE}}$ at 0.1	1.0005	1.0016	1.0052	1.0302		R1 FAIL
JS-W at the dropped-mass rule's scale	0.9941	0.9930	1.0022	1.0185		R1b PASS
JS-W at the in-sample SPO+ loss rule's scale	0.9983	0.9947	1.0039	1.0272		R1c FAIL

Table 10: Irish energy-price knapsack (48 items, ten splits) and Warcraft (monotone, five seeds): decision-focused training pays at every capacity, the shrinkage rows sit on the decision-focused row, and the last-layer rows trail the end-to-end network (outcomes in Table 32).

Test regret	energy-price knapsack, capacity 12	24	48	96	Warcraft ($\times 10^{-4}$)
Shared linear head ($q_M = 9$)					
two-stage (ETO)	0.2894	0.2623	0.2101	0.1311	
smoothed decision-focused ($\hat{\theta}_{I,\epsilon}$)	0.2321	0.1889	0.1362	0.0858	
$\hat{\theta}_{\mathrm{SURE}}$	0.2326	0.1902	0.1364	0.0860	
PyEPO SPO+ (chosen by CV-select in every split)	0.2325	0.1899	0.1357	0.0835	
calibrated T_W rejection share	0.9	0.9	0.9	1.0	
effective dimension	1.18	1.48	1.52	1.55	
Last layer, cross-fitted representation					
FT-none	0.2893	0.2424	0.2026	0.1334	2.854
FT-full	0.2258	0.1822	0.1351	0.0842	2.756
SURE-FT	0.2302	0.1831	0.1354	0.0845	2.636
best fixed step (hindsight)	0.2468	0.1974	0.1492	0.0967	2.446
end-to-end two-stage network	0.2716	0.2331	0.1950	0.1324	1.255
end-to-end decision-focused network	0.2110	0.1734	0.1326	0.0805	0.579
calibrated T_W rejection share	0.9	1.0	1.0	1.0	1.0
effective dimension	1.36	1.44	1.63	1.84	4.03

From Section 3.5. Fine-tuning a last layer for decisions has a step size (a trust-region radius, a blending weight, a number of epochs), set by hand or by cross-validation in the methods we compare with; we report next to SURE-FT the fixed steps $\lambda \in \{0.25, 0.5, 0.75\}$, the pretest, end-to-end networks trained by MSE and by the smoothed SPO+ loss, and cross-validated selection between them, with a capacity sweep and a decision-focused representation as diagnostics (Appendix E).

From Section 4.2. JS-W, preregistered after these results, ran on the same cells and seeds (Part 2G, right panel): in the 24 cells where it holds in fewer than 0.7 of replicates it is within 1.02 of the better endpoint in 23 (G1 PASS; $\hat{\theta}_{\mathrm{SURE}}$: 20, max 1.0654), on the toy within 1.0059 in all 16 cells (G5 PASS) and on the last layer of Section 4.5 within 1.0036 (G6 PASS; SURE-FT up to 1.0472).

From Section 4.2. The price is paid under adequacy at small n: JS-W is within 1.0012 of ETO in the 12 degree-one cells with $n \geq 500$ (G2 PASS) and at 1.0202 and 1.0146 at $n = 200$, where $\hat{\theta}_{\mathrm{SURE}}$ is at 1.0005 and 1.0014 (the null residual, Corollary S6-JS).

From Section 4.1. The Satterthwaite reference, which needs no refits, is conservative on the synthetic benchmarks (0.012 to 0.023) and not on the real features.

From Section 4.3. (iii) The fixed 0.02 with JS-W on fresh seeds (Part 2G-b): below cross-validation in all five cells, within 1.0020 of ETO in their degree-one counterparts (G7, G8 PASS).

From Section 4.3. (iv) An in-sample rule (Part 2G-c), the largest scale whose hard SPO+ training loss is within one percent of the smallest among the three scales and the SPO+ fit: it returns 0.1 in every degree-one replicate, so the test keeps its size (E1-r PASS); JS-W at its scale is within 1.0004 of cross-validation in all 40 cells and below it in 37 (G9 to G11 PASS); on the real prices it stays at 0.1 and the miss at $C = 96$ repeats (R1c FAIL).

From Section 4.4. The effective dimension is 5.1, 4.4, 4.8 and 7.0; $\hat{\theta}_{I,\epsilon}$ has 0.90 to 0.96 of ETO's regret against 0.65 to 0.80 with the shared head (the richer map absorbs part of the misspecification), and the calibrated test rejects in every split at 24 to 96, in 0.40 at 12, where W drops 0.38 to 0.51 of its mass.

From Section 4.4. JS-W has 1.0004 to 1.0155 of the smoothed endpoint's regret and is below it in 6 of 40 splits, with 10 exact ties (R4d FAIL).

From Section 4.4. Least squares is not efficient under the skewed price noise, so the Hausman identity $V_{IE} = V_{EE}$ fails (the Hausman gap $\|V_{EI} - V_{EE}\|/\|V_{EE}\|$ is 0.38 to 1.25) and the regime condition never holds: the ratio κ_W/ν_W of Theorem 2 is 1.05 to 3.84 at 24 to 96, below four, although its Hausman-case form, the effective dimension, exceeds four.

From Section 4.4. And the regret keeps falling all the way to the decision-focused endpoint: at the steps 0, 0.25, 0.5, 0.75, 1 of the path it is smallest at the endpoint, still falling from 0.75 to 1, in 31 of the 40 scored splits and in all 40 shared-head splits, the case $\alpha > 1$ of Corollary S5(iii), so a rule confined to the path can at best match the endpoint, and $\hat{\theta}_{\mathrm{SURE}}$, whose constant is larger, retreats further than JS-W and pays for it at $C = 12$, where the oracle inequality fails (R1d, 1.0464); the safety claim and the claim against cross-validation pass (R2d, R3d; Table 32; second-order numbers in Appendix D).

From Section 4.4. PFYL is 0.89 to 0.92 of ETO, the best column at $C = 24$ and 96, within 1.4 percent of it at 12 and 48; carrying test and path from the smoothed SPO+ surrogate to the Fenchel-Young one is future work.

From Section 4.4. On a second market, the Nord Pool day-ahead prices with a knapsack over the 24 hours (Part 4S), the calibrated test rejects in every split of every cell and decision-focused training cuts the regret by a factor of two to six with item-specific heads (0.18 to 0.56 of ETO's regret on the $p = 34$ head).

From Section 4.4. In numbers: the effective dimension exceeds four in at least seven of ten splits at one capacity only, $C = 48$ with the $p = 114$ head (4.35; at most 3.8 at the other capacities on any head); there JS-W has 0.9978 of the smoothed endpoint's regret, below it in 6 of 10 splits (R4s PASS).

From Section 4.4. $\hat{\theta}_{\mathrm{SURE}}$ is within 1.0053 of the better endpoint except at $C = 12$ (1.0620, R1s FAIL), $\hat{\theta}_{\mathrm{safe}}$ within 1.0029 (R2s PASS), and both rules miss cross-validated selection at $C = 6$ and 12 (1.0769 and 1.4269; 1.0767 and 1.3449; R3s FAIL), where the smoothed endpoint trails SPO+ by 8 and 34 percent, the scale effect of Section 4.3 at its largest; at $C = 48$ both are 0.84 of cross-validated selection.

Sentences removed from the rewritten main text for the page limit (2026-09-25). Each is kept here verbatim with the section it stood in.

From Section 1. Our first design used the Wald geometry, and its record, a shrinkage constant 1.7 to 8.5 times the admissible bound of the regret geometry on our toy problem and calibrated critical values that varied by orders of magnitude across replicates, is what led to the design in this paper.

From Section 4. Problems. A top-k toy ($d = 10$, $k = 3$) as a mechanism check; two PyEPO benchmarks (Tang & Khalil, 2024), shortest path on a 5×5 grid (96 identified coordinates) and knapsack with ten items (60), with polynomial misspecification of degree $\{1, 2, 4, 6\}$ (degree one is correctly specified) and two noise models, multiplicative and additive Gaussian, ten seeds per cell; and two real markets, the PredOpt energy-price knapsack on Irish prices (Mandi et al., 2024; Ifrim et al., 2012) with 48 half-hour items at capacities 12, 24, 48 and 96, and a knapsack over the 24 hours of the Nord Pool day-ahead prices (Lago et al., 2021) at capacities 6, 12, 24 and 48, ten splits each. Baselines are PyEPO's SPO+, PFYL (Berthet et al., 2020) and DBB (Vlastelica Pogančić et al.,

2020), the blended model DDR (Loke et al., 2026) with its cross-validated weight, and *CV-select*, 5-fold cross-validated selection between the two-stage fit and SPO+ by validation regret.

From Section 4.1. The moment-matched reference is conservative here (0.014).

From Section 4.1. On the real features of the energy prices, with the empirical residual law as the null, the calibrated test has size 0.030 (R5); the moment-matched reference rejects 0.04 and 0.17 there, so it is not conservative on real features and we do not recommend it for them.

From Section 4.2. On the toy problem (Table 11 in the appendix) the same pattern holds over 16 cells, and the calibrated test has size 0.02 to 0.08 at every sample size.

From Section 4.4. On the last layer, SURE-FT is 6.7 to 12.6 percent below the best hand-set step (R3).

Sentences whose numerals left the main text in its final revision (2026-09-25). Each is kept here verbatim with the place it stood in.

From the abstract of the rewritten main text. In preregistered experiments the test has size 0.051 under symmetric nulls, the rule is within 0.43 percent of the better fit where its condition holds in most replicates, and on two electricity markets, where decision-focused training cuts regret by 20 percent to a factor of six, the test rejects in 90 to 100 percent of splits with a shared head.

From the shortened abstract. In preregistered experiments the test has size 0.051 under symmetric nulls and rejects in most splits on two electricity markets where decision-focused training cuts regret by 20 percent to a factor of six.

From Section 6 (shortened version). On the real markets the shrinkage gained at most 0.2 percent and cost 4.6 percent at one capacity.

Every preregistered claim with its recorded verdict; nothing is omitted.[1]

Toy (Part 1C). P1 PASS, P1b FAIL, P2 PASS, P3 FAIL, P4 PASS, P5 PASS (Table 26). Per-cell ratio of $\hat{\theta}_{\mathrm{SURE}}$ to the better endpoint, S3(b) frequency and effective dimension: Table 11.

Table 11: Toy problem, per cell: ratio of $\hat{\theta}_{\mathrm{SURE}}$ to the better endpoint, S3(b) frequency, effective dimension, mean $\hat{\lambda}$.

degree	n	SURE / min(ETO, IEO$_\epsilon$)	S3(b) frequency	effective dimension	$\hat{\lambda}$ mean
1	200	1.0000	0.55	5.4	0.049
1	500	1.0001	1.00	8.7	0.047
1	1000	1.0001	1.00	11.7	0.068
1	2000	1.0000	1.00	15.5	0.061
2	200	0.9999	0.52	4.9	0.039
2	500	0.9997	0.96	7.5	0.067
2	1000	1.0000	1.00	9.9	0.143
2	2000	0.9993	1.00	12.7	0.235
4	200	1.0065	0.45	4.8	0.218
4	500	1.0084	0.96	7.6	0.507
4	1000	1.0059	1.00	9.3	0.692
4	2000	1.0035	1.00	10.9	0.835
6	200	1.0224	0.33	4.2	0.482
6	500	1.0112	0.70	5.5	0.761
6	1000	1.0063	0.78	5.3	0.886
6	2000	1.0029	0.87	6.0	0.943

First benchmark round (Part 2C). C1 FAIL, C1b FAIL, C2 FAIL, C3 FAIL, C4 PASS (Table 27); the shortest-path cells at $n \leq 2000$ enter the regime map from this record (Table 12).

[1] In the repository and its frozen records the full-space negative control is named DAS; in this paper it is Wald-geometry James-Stein (WG-JS).

The first round, as recorded. The first frozen round (shortest path at $n \leq 2000$; knapsack with a Monte Carlo perturbed smoothing) is reported as recorded in Table 27: the scissors exist in both noise models at every n (C4), while the uniform claims failed: $\hat{\theta}_{\mathrm{SURE}}$ was within 1.02 of the better endpoint in 48 of 80 cells, all of them where the effective dimension exceeded four, and the knapsack arm with the perturbed sandwich was unusable (effective dimension 1.3 to 1.9 in every cell, $\hat{\theta}_{\mathrm{SURE}}$ up to 43 percent above $\hat{\theta}_{I,\epsilon}$), a negative result of that smoothing which we keep in the record. The F reference did not control size (0.2 to 1.0 at degree one on shortest path).

Table 12: Shortest path, first round (record): S3(b) frequency, effective dimension, and the ratios of $\hat{\theta}_{\mathrm{SURE}}$ and $\hat{\theta}_{\mathrm{safe}}$ to the better endpoint.

noise	degree	n	S3(b) frequency	effective dimension	SURE / min	SURE-safe / min
mult	1	100	0.0	2.8	0.9987	1.0225
mult	1	200	0.2	3.6	0.9999	1.0034
mult	1	500	0.4	4.8	0.9996	0.9997
mult	1	1000	1.0	7.4	1.0003	1.0003
mult	1	2000	1.0	8.7	1.0006	1.0006
mult	2	100	0.0	2.6	1.0035	1.0379
mult	2	200	0.2	3.5	0.9992	1.0016
mult	2	500	0.2	3.7	0.9992	1.0000
mult	2	1000	0.8	6.4	0.9984	0.9985
mult	2	2000	0.9	7.6	0.9989	0.9989
mult	4	100	0.0	2.0	0.9849	1.0058
mult	4	200	0.1	3.0	1.0574	1.0002
mult	4	500	0.3	4.1	1.0318	1.0145
mult	4	1000	0.7	4.8	1.0120	1.0090
mult	4	2000	0.7	5.5	1.0037	1.0032
mult	6	100	0.0	2.3	1.1245	0.9853
mult	6	200	0.2	2.8	1.1925	1.0144
mult	6	500	0.2	3.5	1.0354	1.0212
mult	6	1000	0.4	3.8	1.0080	1.0010
mult	6	2000	0.3	4.4	1.0095	1.0046
gauss	1	100	0.1	2.7	0.9999	1.0443
gauss	1	200	0.1	2.7	1.0000	1.0154
gauss	1	500	0.7	5.2	1.0002	0.9999
gauss	1	1000	0.8	5.9	1.0000	1.0003
gauss	1	2000	1.0	8.2	0.9999	0.9999
gauss	2	100	0.0	2.0	1.0004	1.1048
gauss	2	200	0.0	2.9	1.0004	1.0222
gauss	2	500	0.1	3.2	0.9996	1.0008
gauss	2	1000	1.0	5.5	0.9996	0.9996
gauss	2	2000	1.0	6.7	0.9991	0.9991
gauss	4	100	0.0	2.2	0.9898	1.0749
gauss	4	200	0.0	3.2	1.0146	0.9897
gauss	4	500	0.6	4.6	1.0370	1.0205
gauss	4	1000	0.8	5.8	1.0084	1.0067
gauss	4	2000	0.9	6.9	1.0038	1.0038
gauss	6	100	0.0	2.1	1.0486	0.9727
gauss	6	200	0.3	3.2	1.1042	1.0465
gauss	6	500	0.3	3.9	1.0391	1.0196
gauss	6	1000	0.5	4.5	1.0156	1.0091
gauss	6	2000	0.7	5.0	1.0044	1.0040

Second benchmark round (Part 2D). D1 PASS, D2 PASS (32 of 56 cells qualify), D3 PASS, D4 PASS, D5 FAIL (Table 28); the regime map by cell in Table 13; the size and power cells with the F reference in Table 14.

The smoothing scale (Part 2F). A preregistered sweep of $\epsilon_0 \in \{0.02, 0.05, 0.1, 0.2\}$ on the five cells where D5 failed and their degree-one counterparts, ten seeds, the four scales sharing the data of each seed (Table 15). Its one claim, F1 ($\hat{\theta}_{\mathrm{SURE}} \leq 1.02\times$ cross-validated selection at the scale

Table 13: Second round, per cell: S3(b) frequency, effective dimension, and the ratios of $\hat{\theta}_{\mathrm{SURE}}$ and $\hat{\theta}_{\mathrm{safe}}$ to the better endpoint.

benchmark	noise	degree	n	S3(b) frequency	effective dimension	SURE / min	SURE-safe / min
shortest path	mult	1	5000	1.0	10.1	1.0001	1.0001
shortest path	mult	1	10000	1.0	11.2	1.0000	1.0000
shortest path	mult	2	5000	0.9	9.2	1.0006	1.0003
shortest path	mult	2	10000	1.0	9.5	1.0004	1.0004
shortest path	mult	4	5000	0.7	6.4	1.0012	1.0007
shortest path	mult	4	10000	0.9	7.1	0.9991	0.9992
shortest path	mult	6	5000	0.5	5.0	1.0009	1.0009
shortest path	mult	6	10000	0.6	5.1	1.0024	1.0018
shortest path	gauss	1	5000	1.0	9.0	0.9999	0.9999
shortest path	gauss	1	10000	1.0	11.0	1.0003	1.0003
shortest path	gauss	2	5000	1.0	8.8	1.0005	1.0005
shortest path	gauss	2	10000	1.0	11.4	0.9994	0.9994
shortest path	gauss	4	5000	0.8	7.6	1.0021	1.0018
shortest path	gauss	4	10000	0.9	8.8	1.0005	1.0005
shortest path	gauss	6	5000	0.6	5.5	1.0030	1.0024
shortest path	gauss	6	10000	0.8	6.8	1.0011	1.0010
knapsack	mult	1	200	0.1	3.3	1.0005	1.0082
knapsack	mult	1	500	0.8	5.6	0.9999	1.0002
knapsack	mult	1	1000	0.9	7.1	0.9999	1.0000
knapsack	mult	1	2000	0.9	8.9	1.0001	1.0000
knapsack	mult	1	5000	0.9	8.3	1.0000	1.0000
knapsack	mult	2	200	0.2	3.8	0.9999	0.9987
knapsack	mult	2	500	0.6	4.7	0.9998	0.9996
knapsack	mult	2	1000	0.5	5.2	0.9988	0.9995
knapsack	mult	2	2000	0.9	7.4	1.0014	1.0014
knapsack	mult	2	5000	1.0	9.0	1.0005	1.0005
knapsack	mult	4	200	0.2	3.4	1.0654	1.0251
knapsack	mult	4	500	0.3	4.3	1.0089	1.0047
knapsack	mult	4	1000	0.6	4.4	1.0052	1.0044
knapsack	mult	4	2000	0.8	5.6	1.0027	1.0021
knapsack	mult	4	5000	0.9	7.1	1.0005	1.0005
knapsack	mult	6	200	0.1	3.3	1.0257	1.0081
knapsack	mult	6	500	0.3	3.7	1.0242	1.0168
knapsack	mult	6	1000	0.0	3.0	1.0164	1.0020
knapsack	mult	6	2000	0.0	3.8	1.0038	1.0027
knapsack	mult	6	5000	0.1	3.8	1.0013	1.0004
knapsack	gauss	1	200	0.2	3.2	1.0014	1.0130
knapsack	gauss	1	500	0.7	5.2	1.0002	1.0010
knapsack	gauss	1	1000	0.7	5.7	1.0000	1.0004
knapsack	gauss	1	2000	1.0	10.6	0.9998	0.9999
knapsack	gauss	1	5000	1.0	12.2	1.0001	1.0001
knapsack	gauss	2	200	0.2	3.5	0.9992	0.9987
knapsack	gauss	2	500	0.7	5.0	0.9977	0.9990
knapsack	gauss	2	1000	0.9	6.6	1.0020	1.0023
knapsack	gauss	2	2000	1.0	8.0	0.9999	0.9999
knapsack	gauss	2	5000	1.0	10.2	1.0016	1.0016
knapsack	gauss	4	200	0.2	4.2	1.0160	1.0132
knapsack	gauss	4	500	0.4	4.4	1.0134	1.0095
knapsack	gauss	4	1000	0.7	5.2	1.0043	1.0043
knapsack	gauss	4	2000	0.8	6.4	1.0029	1.0028
knapsack	gauss	4	5000	0.9	7.2	1.0009	1.0009
knapsack	gauss	6	200	0.0	3.6	1.0522	1.0280
knapsack	gauss	6	500	0.0	2.9	1.0188	1.0037
knapsack	gauss	6	1000	0.2	3.6	1.0082	1.0028
knapsack	gauss	6	2000	0.2	3.5	1.0035	1.0006
knapsack	gauss	6	5000	0.1	3.6	1.0018	1.0013

Table 14: Second round, size cells at degree one: rejection rates of the calibrated Wald test, the F reference and the uncorrected chi-square reference, and the ratio of the calibrated critical value to the F quantile.

benchmark	noise	n	calibrated test	F reference	chi-square	crit. ratio (cal. / F)
shortest path	mult	500	0.03	0.69	0.96	1.53
shortest path	mult	1000	0.03	0.49	0.65	1.35
shortest path	mult	2000	0.05	0.35	0.46	1.23
shortest path	gauss	500	0.06	0.79	0.98	1.66
shortest path	gauss	1000	0.02	0.70	0.85	1.45
shortest path	gauss	2000	0.05	0.49	0.57	1.29
knapsack	mult	500	0.04	0.58	0.70	1293.74
knapsack	mult	1000	0.06	0.56	0.63	2.75
knapsack	mult	2000	0.03	0.36	0.38	1.93
knapsack	gauss	500	0.03	0.81	0.92	2778.74
knapsack	gauss	1000	0.03	0.71	0.76	4.81
knapsack	gauss	2000	0.08	0.62	0.65	3.97

with the smallest smoothing gap, in each of the five degree-six cells; the gap is a test-set quantity, so the rule, frozen before the run, selects the scale on the test set, and the tables of this paper keep the preregistered 0.1), FAIL as recorded: the smallest gap is at $\epsilon_0 = 0.02$ in every cell, where four cells are within (all four below cross-validated selection outright) and the knapsack cell at $n = 1000$ is at 1.0368, through the step ($\hat{\theta}_{\mathrm{SURE}}$ at 1.0480 of its better endpoint, $\hat{\lambda}$ with a standard deviation of 0.199 across seeds against at most 0.023 in the other cells) and not through the endpoint, which is below SPO+ there. Recorded without a criterion: the degree-six gap $\hat{\theta}_{I,\epsilon}$ / SPO+ is 0.9685 to 0.9893 at $\epsilon_0 = 0.02$, 1.0013 to 1.0178 at 0.05, 1.0456 to 1.0807 at the preregistered 0.1 and 1.1395 to 1.2016 at 0.2; $\hat{\theta}_{\mathrm{SURE}}$ / ETO at degree one is 0.9996 to 1.0003 at every scale; the calibrated T_W test rejects 6 of 200 degree-one tests; every fit and every null refit converged. No scale was selected after seeing F1.

Third round, the regret-weighted test (Part 2E). E1 PASS, E2 PASS (Table 29); E3 reported; the size cells with every reference and the spread of the critical values in Table 16.

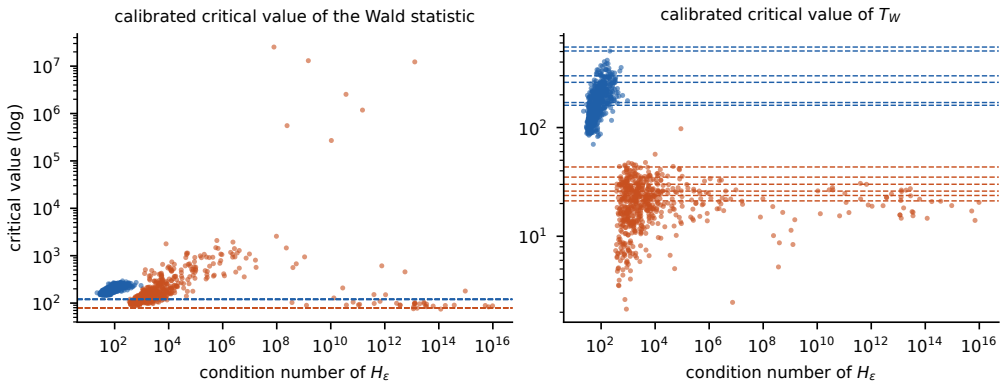

Figure 5: The null tails on identical draws (on knapsack the Wald critical value has a coefficient of variation across replicates of 0.61 to 9.02 and its logarithm correlates up to 0.37 with the log condition number of H_ϵ, a share of 0.09 to 0.45 of replicates above 1e4, while the T_W critical value has a coefficient of variation of 0.34 to 0.49 and a correlation of at most 0.12): per replicate of the twelve degree-one size cells of the third round, the calibrated critical value of the Wald statistic (left) and of T_W (right) against the condition number of $\hat{H}_\epsilon$; dashed lines are the chi-square quantile (left) and the mean Satterthwaite quantile (right); blue shortest path, orange knapsack.

Neural (Part 3N). NL1 VOID, NL2 FAIL, NL3 PASS, NL4 FAIL, NL5 PASS (Table 30); every cell in Table 25. Under the capped optimiser budget the gradient rule was met in 0.45, 0.45 and

Table 15: The smoothing scale (Part 2F, frozen), per cell and scale: the smoothing gap, the ratios of $\hat{\theta}_{\mathrm{SURE}}$ to cross-validated selection, to ETO and to the better endpoint, the calibrated T_W rejections at that cell, $\hat{\lambda}$ and the effective dimension; * marks the scale with the smallest gap.

bench	degree	n	ϵ_0	$\hat{\theta}_{I,\epsilon}$ / SPO+	$\hat{\theta}_{\mathrm{SURE}}$ / CV	$\hat{\theta}_{\mathrm{SURE}}$ / ETO	$\hat{\theta}_{\mathrm{SURE}}$ / better	T_W rej. (of 10)	$\hat{\lambda}$	eff. dim
ks	6	1000	0.02*	0.9893	1.0368	0.5062	1.0480	10	0.872	2.8
ks	6	1000	0.05	1.0178	1.0280	0.5019	1.0100	10	0.959	3.3
ks	6	1000	0.1	1.0807	1.0927	0.5335	1.0111	10	0.969	3.6
ks	6	1000	0.2	1.2016	1.2081	0.5898	1.0054	10	0.987	3.7
ks	6	2000	0.02*	0.9834	0.9864	0.5157	1.0030	10	0.967	3.3
ks	6	2000	0.05	1.0047	1.0081	0.5271	1.0034	10	0.982	3.3
ks	6	2000	0.1	1.0601	1.0621	0.5553	1.0019	10	0.986	3.7
ks	6	2000	0.2	1.1564	1.1607	0.6068	1.0037	10	0.991	3.4
ks	6	5000	0.02*	0.9839	0.9852	0.5530	1.0014	10	0.991	3.4
ks	6	5000	0.05	1.0087	1.0086	0.5661	0.9999	10	0.993	4.1
ks	6	5000	0.1	1.0751	1.0770	0.6045	1.0018	10	0.993	4.4
ks	6	5000	0.2	1.1788	1.1804	0.6625	1.0013	10	0.996	4.1
sp	6	5000	0.02*	0.9685	0.9725	0.5977	1.0041	10	0.967	4.1
sp	6	5000	0.05	1.0048	1.0059	0.6182	1.0011	10	0.975	4.3
sp	6	5000	0.1	1.0551	1.0564	0.6492	1.0013	10	0.984	4.5
sp	6	5000	0.2	1.1430	1.1442	0.7032	1.0010	10	0.986	4.2
sp	6	10000	0.02*	0.9812	0.9821	0.6534	1.0010	10	0.978	4.6
sp	6	10000	0.05	1.0013	1.0018	0.6664	1.0005	10	0.987	5.0
sp	6	10000	0.1	1.0456	1.0463	0.6961	1.0007	10	0.989	4.6
sp	6	10000	0.2	1.1395	1.1402	0.7585	1.0006	10	0.989	5.0
ks	1	1000	0.02	0.9980	0.9998	1.0002	1.0002	0	0.015	3.8
ks	1	1000	0.05	0.9972	0.9998	1.0001	1.0001	0	0.012	4.4
ks	1	1000	0.1	0.9963	0.9996	1.0000	1.0000	0	0.001	6.7
ks	1	1000	0.2*	0.9955	0.9996	1.0000	1.0000	0	0.006	10.8
ks	1	2000	0.02	0.9902	0.9971	0.9996	0.9996	1	0.043	4.2
ks	1	2000	0.05	0.9895	0.9975	1.0000	1.0000	0	0.022	5.7
ks	1	2000	0.1	0.9888	0.9976	1.0001	1.0001	0	0.023	6.4
ks	1	2000	0.2*	0.9878	0.9975	1.0000	1.0000	0	0.013	12.8
ks	1	5000	0.02*	0.9891	0.9985	1.0001	1.0007	0	0.032	4.9
ks	1	5000	0.05	0.9896	0.9983	1.0000	1.0000	0	0.018	7.0
ks	1	5000	0.1	0.9899	0.9984	1.0000	1.0000	1	0.039	10.5
ks	1	5000	0.2	0.9904	0.9983	1.0000	1.0000	1	0.042	16.2
sp	1	5000	0.02	0.9803	1.0003	1.0003	1.0003	0	0.060	6.0
sp	1	5000	0.05	0.9797	1.0000	1.0000	1.0000	0	0.130	8.4
sp	1	5000	0.1	0.9799	0.9999	0.9999	0.9999	0	0.148	11.3
sp	1	5000	0.2*	0.9797	1.0000	1.0000	1.0000	0	0.126	18.0
sp	1	10000	0.02	0.9785	0.9999	0.9999	0.9999	0	0.058	7.3
sp	1	10000	0.05	0.9787	0.9998	0.9998	0.9998	1	0.075	8.9
sp	1	10000	0.1	0.9786	0.9997	0.9997	0.9997	1	0.090	11.9
sp	1	10000	0.2*	0.9784	0.9998	0.9998	0.9998	1	0.156	19.2

Table 16: Third round, size cells at degree one: rejection rates of the calibrated T_W test, the Satterthwaite reference, the calibrated Wald test and the F reference on identical null draws, and the coefficient of variation of the two calibrated critical values across replicates.

benchmark	noise	n	T_W cal.	Satterthwaite	Wald cal.	F	crit. cv T_W	crit. cv Wald
shortest path	mult	500	0.03	0.00	0.06	0.64	0.22	0.05
shortest path	mult	1000	0.02	0.00	0.03	0.49	0.21	0.05
shortest path	mult	2000	0.05	0.03	0.04	0.30	0.21	0.03
shortest path	gauss	500	0.06	0.00	0.04	0.84	0.25	0.06
shortest path	gauss	1000	0.05	0.00	0.06	0.73	0.24	0.04
shortest path	gauss	2000	0.07	0.02	0.05	0.49	0.19	0.04
knapsack	mult	500	0.05	0.01	0.05	0.53	0.49	0.61
knapsack	mult	1000	0.04	0.01	0.05	0.44	0.49	0.78
knapsack	mult	2000	0.07	0.05	0.03	0.31	0.39	0.90
knapsack	gauss	500	0.08	0.03	0.04	0.80	0.42	9.02
knapsack	gauss	1000	0.05	0.00	0.06	0.72	0.39	7.03
knapsack	gauss	2000	0.04	0.02	0.07	0.61	0.34	7.38

Table 17: The third round's size cells, read against Proposition S7-d without any re-run: the empirical 0.95 quantile of the 100 original T_W values of a cell, $q_{0.95}(T_W)$, against the mean and the median of its 100 calibrated critical values (crit.). A quantile from 100 draws rests on its fifth and sixth largest values and is noisy; the ratio runs from 0.89 to 1.40 and is above one in most benchmark cells, in line with rejection counts of 2 to 8 per 100.

Benchmark	noise	n	$q_{0.95}(T_W)$	crit., mean	crit., median	$q_{0.95}$ / median	rej. of 100
toy	Gaussian	200	24.01	23.39	23.08	1.04	2
toy	Gaussian	500	14.93	16.96	16.70	0.89	3
toy	Gaussian	1000	14.54	14.51	14.60	1.00	8
toy	Gaussian	2000	12.48	12.99	12.88	0.97	6
shortest path	Gaussian	500	268.45	258.02	249.21	1.08	6
shortest path	Gaussian	1000	194.19	178.39	172.25	1.13	5
shortest path	Gaussian	2000	142.77	127.79	127.47	1.12	7
shortest path	multiplicative	500	248.50	219.35	214.87	1.16	3
shortest path	multiplicative	1000	167.89	168.71	163.48	1.03	2
shortest path	multiplicative	2000	126.35	121.36	117.01	1.08	5
knapsack	Gaussian	500	28.74	21.96	21.05	1.37	8
knapsack	Gaussian	1000	22.97	19.93	19.88	1.16	5
knapsack	Gaussian	2000	19.55	17.45	17.71	1.10	4
knapsack	multiplicative	500	32.70	26.51	27.21	1.20	5
knapsack	multiplicative	1000	30.41	21.61	21.66	1.40	4
knapsack	multiplicative	2000	26.38	20.20	21.70	1.22	7

0.85 of the degree-six fits at $n = 2000, 5000, 10000$, and at width 16 in 0.00 of fits and 0.04 of null refits. At $n = 2000$ and 5000 the L-BFGS cap of 200 iterations binds in every direction, the Newton polish runs 5.65 and 4.65 steps on average, the ratio of the final gradient norm to the gradient rule has median 17.5 and 14.6 and maximum 5.0×10^5 and 2.2×10^5, and the median condition number of $\hat{H}_\epsilon$ is 2.3×10^8 and 9.8×10^{12}, against 5.3×10^3 at degree one and $n = 5000$. The regret of FT-full in the directions that met the rule and in those that did not is 0.0809 against 0.0807 at $n = 5000$ (0.0721 against 0.0722 at $n = 10000$), so the residual gradient lies in directions with little decision relevance; at $n = 2000$ it is 0.0900 against 0.1160, with FT-none at 0.1219 against 0.1413 in the same split, so those are the harder samples and not an effect of the cap. In the second and third benchmark rounds every smoothed fit met the rule in every cell, and the null refits in at least 0.988 of a cell.

Neural addendum arms (Parts 3N-b and 3N-c; reported rows, no claim). Table 18: the end-to-end networks trained by DBB and by PFYL on the same architecture, the smoothed SPO+ network at widths 4 and 16 next to the width-8 network of Part 3N, and the last-layer rows with the representation and the head both trained on the full sample without cross-fitting (labelled "full"; outside the theory, since the representation is not independent of the head sample, so the sandwich and the rule are plug-ins without a guarantee). Against PFYL, which is the best end-to-end network at degree six, SURE-FT would be 1.265 and 1.193 at $n = 5000$ and 10000; the frozen NL2 verdict is against the better of the MSE and smoothed SPO+ networks and stands as recorded. On the full sample the plug-in step is within 1.9 percent of its better endpoint in 15 of 16 cells and at 1.1205 at degree six, $n = 2000$, where the smoothed fit met the gradient rule in 0.40 of seeds; FT-full on the full sample is 0.957 to 1.015 of the cross-fitted one at degrees one to four.

Warcraft, monotone variant (Part 3W). W2 FAIL, W3 FAIL, W4 VOID (Table 32). Decisions are restricted to the paths that move right, down or diagonally down-right, because on the original 8-connected graph the SPO+ argument $2\hat{c} - c$ has negative entries and neither SPO+ nor its smoothing is defined; the monotone optimum equals the shipped one on 0.807 of the test maps, regret is against the monotone oracle, and the numbers are not comparable with published ones. The experiment was frozen with its claims, five seeds and a time box (freeze commit `8b9a4a9` in the supplementary repository); its pool was suspended by system sleep, outlasted the box, and was first recorded as not run. After the pool had finished and before its result file was read, Amendment 11 lifted the time box; the pool's log, which prints the regret of every row rounded to four decimals, was visible before that ruling, and no rule, threshold or claim was changed. W3 is scored, as its frozen code computes

Table 18: Neural addendum arms, per cell (test regret): MLP-DBB and MLP-PFYL; the smoothed SPO+ network at widths 4, 8 and 16; FT-full and SURE-FT cross-fitted (cf, Part 3N) and on the full sample (full, no cross-fitting).

noise	degree	n	MLP-DBB	MLP-PFYL	SPO+$_\epsilon$ h=4	h=8	h=16	FT-full (cf)	FT-full (full)	SURE-FT (cf)	SURE-FT (full)
mult	1	2000	0.2285	0.1664	0.1685	0.1619	0.1603	0.1600	0.1586	0.1590	0.1581
mult	1	5000	0.2297	0.1612	0.1633	0.1595	0.1593	0.1589	0.1602	0.1583	0.1598
mult	1	10000	0.2306	0.1591	0.1649	0.1575	0.1577	0.1574	0.1567	0.1572	0.1565
mult	2	2000	0.2206	0.1132	0.1243	0.1104	0.1099	0.1114	0.1113	0.1108	0.1113
mult	2	5000	0.2144	0.1049	0.1135	0.1040	0.1039	0.1072	0.1026	0.1071	0.1028
mult	2	10000	0.2350	0.1061	0.1183	0.1051	0.1050	0.1047	0.1063	0.1047	0.1063
mult	4	2000	0.2567	0.0738	0.1006	0.0736	0.0719	0.0749	0.0724	0.0772	0.0737
mult	4	5000	0.2144	0.0739	0.0898	0.0715	0.0708	0.0733	0.0725	0.0740	0.0729
mult	4	10000	0.2195	0.0701	0.0846	0.0688	0.0690	0.0709	0.0706	0.0713	0.0709
mult	6	2000	0.2640	0.0754	0.1313	0.0754	0.0717	0.1043	0.0846	0.1120	0.0948
mult	6	5000	0.2089	0.0669	0.0908	0.0676	0.0657	0.0808	0.0754	0.0846	0.0759
mult	6	10000	0.2389	0.0609	0.0826	0.0622	0.0606	0.0721	0.0697	0.0727	0.0698
gauss	1	5000	0.1927	0.1252	0.1290	0.1239	0.1239	0.1231	0.1227	0.1228	0.1222
gauss	2	5000	0.1991	0.0921	0.0989	0.0911	0.0915	0.0908	0.0902	0.0905	0.0902
gauss	4	5000	0.2880	0.0855	0.1118	0.0845	0.0844	0.0863	0.0853	0.0866	0.0855
gauss	6	5000	0.2896	0.1181	0.1442	0.1187	0.1162	0.1265	0.1231	0.1295	0.1235

Table 19: The capacity sweep of Part 3N at $n = 5000$ (reported, no claim): the ratios of SURE-FT and SURE-safe to the better head, the step with its standard deviation across seeds, the effective dimension, the frequency of the S3(b) condition and the rejection share of the last-layer test, by degree and width.

degree	width	SURE-FT / better head	SURE-safe / better head	$\hat{\lambda}$ (sd)	eff. dim	S3(b)	T_W rej.
1	4	1.0008	1.0008	0.11 (0.17)	6.50	0.85	0.10
1	8	0.9999	0.9999	0.04 (0.05)	10.29	1.00	0.00
1	16	0.9999	1.0006	0.03 (0.03)	11.33	0.95	0.05
6	4	1.0037	1.0000	0.88 (0.11)	2.96	0.20	1.00
6	8	1.0472	1.0057	0.69 (0.31)	3.93	0.45	1.00
6	16	1.1372	1.0036	0.24 (0.31)	2.15	0.10	0.50

it, against the mean of the per-seed best fixed step; NL4 and R3 are scored against the fixed step with the smallest cell-mean regret, and W3 fails under either reading. The smoothed end-to-end network has 0.1606 of the regret of the PyEPO SPO+ network, an observation with no claim; the regret weight of the head loses 0.12 to 0.74 of its eigen-mass in the positive semidefinite projection, and with a single fitted scale no scale rule was or could be applied. SPO+ is undefined on the cyclic grid graph because $2\hat{c} - c$ has negative entries, so decisions are the monotone paths and regret is against the monotone oracle (Table 10, last column). All regrets there are below 4×10^{-4} and the last-layer rows lie within the noise of five seeds; the ratio to the end-to-end network (4.5485, W2 FAIL) does not (its time box was lifted after the fact: Reproducibility statement).

The energy instance against the benchmark (Part 4R). The instance uses the benchmark's data file, items, weights and split rule (Mandi et al., 2024); the capacities (12, 24, 48 and 96 here; 60, 120 and 180 there), the decision-focused loss (the smoothed SPO+ here, SPO+ and others there) and the reported normalization differ from the benchmark's (Section 4.4). Table 20 gives the absolute test regrets of the two-stage head, PyEPO's SPO+, PFYL and DBB on the shared linear head in both normalizations, from the frozen summary of Part 4R; no published number is quoted. On the real features the Satterthwaite reference rejects in 0.04 of the 100 null tests at $C = 12$ and in 0.17 at $C = 48$ (0.105 pooled; Table 8).

Table 20: Energy-price knapsack, shared linear head: absolute test regret, cell means over ten splits (Part 4R). Normalized: the sum of the regrets over the sum of the optimal values, the regret of every claim; relative: the mean over days of the regret divided by the optimal value, the benchmark's metric.

	normalized regret, capacity				relative regret, capacity			
Method	12	24	48	96	12	24	48	96
two-stage (ETO)	0.2894	0.2623	0.2101	0.1311	0.2510	0.2487	0.2058	0.1294
PyEPO SPO+	0.2325	0.1899	0.1357	0.0835	0.2100	0.1843	0.1336	0.0816
PFYL	0.4612	0.2749	0.2102	0.0964	0.5440	0.3110	0.2251	0.0987
DBB	0.4042	0.3355	0.2592	0.1967	0.4080	0.3431	0.2615	0.1986

The item-specific head (Part 4R-d). R1d FAIL, R2d PASS, R3d PASS, R4d FAIL (Table 32; the scored arm per capacity in Table 10). Part 4R in everything but the feature map: arms A (item indicators, $p = 56$), B (plus the slopes interacted with a Fourier basis of order two in the half-hour index, $p = 88$) and C (order four, $p = 120$), a rank guard that dropped no column, and three notes recorded before any run: the fit budget of this Part (2600 L-BFGS iterations against 200 in Part 4R, because p is six to thirteen times larger; the Newton polish and the gradient rule unchanged), qualification on converged fits only (an arm qualifies at a capacity iff its main-scale fit meets the gradient rule and its effective dimension exceeds four in at least seven of ten splits; the first qualifying arm of A, B, C is scored), and Arm C conditional, run only at the capacities where neither A nor B qualified. Table 21 gives the scored arm per capacity and Table 22 the nine arm-capacity cells. Arm C was triggered at $C = 12$ only, where A qualified in no split and B in four; at 24, 48 and 96 Arm A qualified and C did not run. The main-scale fit met the gradient rule in every split of every cell, so the exclusion by convergence that a synthetic stand-in had led us to expect at $C = 12$ did not occur; the null refits met the rule in 0.99 of draws on B and 0.97 on C at $C = 12$ and in every draw elsewhere. On Arm B, not scored, JS-W has 1.0052, 1.0033, 1.0028 and 0.9994 of the regret of $\hat{\theta}_{I,\epsilon}$ and is below it in 3, 3, 5 and 6 of ten splits as the effective dimension rises from 4.0 to 10.2. On the scored arms the constant a_{JS} of JS-W is positive in 37 of 40 splits (the premise of Corollary S6-JS); of the 10 exact ties with $\hat{\theta}_{I,\epsilon}$, three are splits where it is non-positive and the rule returns the endpoint and seven are splits where no decision moves; the weight is 0.004 to 0.025 at 24 to 96 and 0.06 to 0.15 at 12, so at 24 to 96 the rule moves at most 2.5 percent of the way toward the two-stage endpoint and the hard regret changes by -0.3 to 2.0 percent; the per-split ratio at 12 ranges from 0.987 to 1.059 (the JS-W fields of the cell files). Moved here from Section 4.4: the regret is smallest at the endpoint, still falling from 0.75 to 1, in 7, 8, 10 and 6 of ten scored splits at $C = 12$, 24, 48 and 96; at 24 to 96 the constant of JS-W is 0.00 to 0.24 of its Hausman-case value $\mathrm{tr}B - 2\lambda_{\max}B$ (median 0.12); T_W is 41 to 264 and 6.9 to 17.6 times a_{JS} (at 24 to 96 and at 12); the risk reduction of Corollary S6-JS, a_{JS}^2/T_W, is at most 0.004 of τ_W (0.011 to 0.060 at 12); and the fall of the regret to the endpoint is the case $\alpha > 1$ of Corollary S5(iii) in the direction of the first design's norm-ratio record. Reported without criterion: $\hat{\theta}_{I,\epsilon}$ / ETO is 0.90 to 0.99 over the nine cells, against 0.65 to 0.80 with the shared head, whose two-stage regrets are larger; the calibrated T_W test rejects in every split on Arm A and on Arm B at 24, 48 and 96, in 0.70 on B and 0.40 on C at $C = 12$, read under the frozen rule as the map absorbing part of the misspecification, next to $\hat{\theta}_{I,\epsilon}$ / ETO, not as a failure of the test; $\hat{\lambda}$ is 0.80 to 0.97 for $\hat{\theta}_{\mathrm{SURE}}$ and 0.90 to 0.995 for JS-W; S3(b) holds in 0.70 of splits on C at 12, 0.20 on B at 48, 0.10 on B at 12 and in no split elsewhere, although the effective dimension exceeds four: the two are different conditions in the general case, and the qualification reads the effective dimension, as frozen; W drops 0.38 to 0.51 of its mass at $C = 12$, 0.22 to 0.26 at 24 and 0.02 to 0.05 at 48 and 96; PFYL / ETO is 0.89 to 0.93 and DBB / ETO 1.20 to 1.57; JS-W at the dropped-mass rule's scale is 0.9756 to 1.0296 of cross-validated selection, the rule choosing 0.1 in 0.1 to 0.9 of the splits of a cell. The pool ran 13 hours on 18 workers with no failed replicate; the machine idle-slept for about six hours of it, which lengthened the recorded seconds per replicate and entered no result.

Table 21: Energy-price knapsack with an item-specific head (Part 4R-d), the scored arm per capacity: test regret, cell means over ten splits (PyEPO normalized regret). Verdicts as recorded: R1d FAIL at $C = 12$, R2d PASS, R3d PASS, R4d FAIL.

	energy-price knapsack, capacity			
Test regret	12	24	48	96
Item-specific head, scored arm (C, $p = 120$, at $C = 12$; A, $p = 56$, at 24, 48, 96)				
two-stage (ETO)	0.1751	0.1449	0.1070	0.0647
smoothed decision-focused ($\hat{\theta}_{I,\epsilon}$)	0.1583	0.1340	0.0965	0.0621
$\hat{\theta}_{\mathrm{SURE}}$	0.1656	0.1345	0.0978	0.0620
$\hat{\theta}_{\mathrm{safe}}$	0.1640	0.1345	0.0969	0.0621
JS-W	0.1607	0.1345	0.0967	0.0621
CV-select	0.1645	0.1358	0.0965	0.0617
PFYL	0.1605	0.1296	0.0975	0.0597
calibrated T_W rejection share	0.4	1.0	1.0	1.0
effective dimension	5.10	4.43	4.84	7.02
S3(b) share	0.70	0.00	0.00	0.00
κ_W/ν_W (median)	4.38	2.58	2.96	2.60
Hausman gap (median)	0.88	0.56	0.52	0.44

Table 22: Part 4R-d, the item-specific head: qualification and the dominance ratio in the nine arm-capacity cells ($p = 56$, 88, 120 for arms A, B, C; ten splits each; above 4 counts the splits whose main-scale fit meets the gradient rule and whose effective dimension exceeds four; every main-scale fit met the rule). The scored arm at a capacity is the first of A, B, C that qualifies; Arm C ran at $C = 12$ only, where neither A nor B qualified.

arm	C	eff. dim. (min, max)	above 4	qual.	scored	$\hat{\theta}_{I,\epsilon}$ / ETO	JS-W / $\hat{\theta}_{I,\epsilon}$	below (ties)	S3(b)	cal. rej.
A	12	2.41 (2.00, 3.30)	0 of 10	no	no	0.9859	1.0031	0 (8)	0.00	1.00
A	24	4.43 (3.31, 5.21)	8 of 10	yes	yes	0.9250	1.0034	2 (4)	0.00	1.00
A	48	4.84 (1.24, 6.93)	9 of 10	yes	yes	0.9023	1.0019	1 (1)	0.00	1.00
A	96	7.02 (2.81, 8.61)	9 of 10	yes	yes	0.9590	1.0004	0 (5)	0.00	1.00
B	12	3.95 (2.61, 5.24)	4 of 10	no	no	0.9003	1.0052	3 (0)	0.10	0.70
B	24	4.94 (2.27, 6.94)	7 of 10	yes	no	0.9167	1.0033	3 (2)	0.00	1.00
B	48	6.79 (3.67, 10.18)	8 of 10	yes	no	0.9464	1.0028	5 (4)	0.20	1.00
B	96	10.19 (4.03, 13.08)	10 of 10	yes	no	0.9048	0.9994	6 (3)	0.00	1.00
C	12	5.10 (3.36, 6.90)	8 of 10	yes	yes	0.9038	1.0155	3 (0)	0.70	0.40

The second market (Part 4S). R1s FAIL, R2s PASS, R3s FAIL, R4s PASS at the one qualifying capacity (Table 32; the scored arm per capacity in Table 23, the sixteen arm-capacity cells in Table 24). The Nord Pool day-ahead prices of the open-access benchmark of Lago et al. (2021), six years of hourly prices with the system load and the wind forecast as exogenous series: each day is one observation, the 24 hours are the items, one integer weight per hour drawn once from the benchmark's set, item value the weight times the price, capacities 6, 12, 24 and 48; the shared head S ($p = 11$: the two exogenous series, the price one and seven days earlier, weekday indicators) and the arms A, B, C of Part 4R-d adapted to 24 items ($p = 34$, 74, 114); ten random splits of the days, with one chronological split reported without criterion; preregistered before the file was opened, run before submission after five notes recorded before the run (the summary stage; arms B and C conditional, run only where neither S nor A qualifies; days with 23 or 25 hours dropped and counted; per-replicate checkpoints; the wall-clock estimate and the cutoff). The file has 2184 calendar days, none with 23 or 25 hours; 2177 days are kept. Neither S nor A qualified at any capacity (no split above four effective dimensions), so B and C ran at all four; only C at $C = 48$ qualifies (4.35, above four in 7 of 10 splits); at $C = 6$ the main-scale fit of B and C missed the gradient rule in every split (excluded by convergence, their effective dimension reported). On the scored arms the ratio κ_W/ν_W has medians 2.02, 1.42, 1.69 and 3.01 and the Hausman gap 1.11, 1.15, 2.66 and 1.61 at $C = 6$, 12, 24, 48. At $C = 48$ on arm C the constant a_{JS} is positive in every split, the weight is 0.002 to 0.006 and the risk reduction that Corollary S6-JS predicts, a_{JS}^2/T_W, is 0.0003 to 0.0019 of τ_W, against the recorded gain of 0.2 percent. The calibrated test rejects in every split of every cell. PFYL's recorded value on the shared head, 11.03 of ETO at $C = 12$ (3.40 and 3.72 at 24 and 48), is the comparator's value under its fixed configuration, not re-run; on the item-specific arms it is 0.24 to 0.75. The residual skewness of the standardised costs is -0.70 to -0.34 with the shared head and -5.48 to -0.95 on the item-specific arms. The pool ran 7 hours on 18 workers, 176 replicates, no failed replicate; the cutoff did not fire.

Table 23: Nord Pool knapsack (Part 4S), the scored arm per capacity: test regret, cell means over ten random splits (PyEPO normalized regret). Verdicts as recorded: R1s FAIL at $C = 12$, R2s PASS, R3s FAIL at $C = 6$ and 12, R4s PASS at $C = 48$, the one qualifying capacity.

	Nord Pool knapsack, capacity			
Test regret	6	12	24	48
Scored arm (A, $p = 34$, at $C = 6$, 12, 24, where no arm qualifies; C, $p = 114$, at 48)				
two-stage (ETO)	0.0337	0.0351	0.0280	0.0151
smoothed decision-focused ($\hat{\theta}_{I,\epsilon}$)	0.0061	0.0195	0.0106	0.0050
$\hat{\theta}_{\mathrm{SURE}}$	0.0061	0.0207	0.0107	0.0050
$\hat{\theta}_{\mathrm{safe}}$	0.0061	0.0195	0.0106	0.0050
JS-W	0.0061	0.0196	0.0106	0.0050
CV-select	0.0057	0.0145	0.0110	0.0059
PyEPO SPO+	0.0057	0.0145	0.0110	0.0059
PFYL	0.0106	0.0180	0.0169	0.0086
calibrated T_W rejection share	1.0	1.0	1.0	1.0
effective dimension	2.18	2.06	1.99	4.35
S3(b) share	0.00	0.00	0.00	0.10
κ_W/ν_W (median)	2.02	1.42	1.69	3.01
Hausman gap (median)	1.11	1.15	2.66	1.61

Table 24: Part 4S, the sixteen arm-capacity cells (p = 11, 34, 74, 114 for arms S, A, B, C; ten random splits each; above 4 counts the splits whose main-scale fit meets the gradient rule and whose effective dimension exceeds four, with the number of fits that missed the rule). The scored arm at a capacity is the first of S, A, B, C that qualifies (C at $C = 48$), Arm A where none does. PFYL / ETO is the comparator's recorded value under its fixed configuration, not re-run.

arm	C	eff. dim. (min, max)	above 4 (missed)	qual.	$\hat{\theta}_{I,\epsilon}$ / ETO	JS-W / $\hat{\theta}_{I,\epsilon}$	below (ties)	S3(b)	cal. rej.	PFYL / ETO
S	6	1.52 (1.30, 1.75)	0 of 10 (0)	no	0.6786	1.0000	0 (10)	0.00	1.00	0.90
S	12	1.42 (1.29, 1.57)	0 of 10 (0)	no	0.9629	1.0000	0 (10)	0.00	1.00	11.03
S	24	1.18 (1.08, 1.49)	0 of 10 (0)	no	0.9767	1.0000	0 (10)	0.00	1.00	3.40
S	48	1.48 (1.36, 1.64)	0 of 10 (0)	no	0.9304	1.0000	0 (10)	0.00	1.00	3.72
A	6	2.18 (1.65, 2.58)	0 of 10 (0)	no	0.1816	1.0000	0 (10)	0.00	1.00	0.31
A	12	2.06 (1.38, 3.53)	0 of 10 (0)	no	0.5567	1.0010	0 (8)	0.00	1.00	0.51
A	24	1.99 (1.58, 2.64)	0 of 10 (0)	no	0.3804	0.9997	1 (9)	0.00	1.00	0.60
A	48	2.22 (1.60, 2.63)	0 of 10 (0)	no	0.5455	1.0000	0 (10)	0.00	1.00	0.75
B	6	3.22 (1.67, 3.89)	0 of 10 (10)	no	0.1639	1.0004	4 (2)	0.00	1.00	0.24
B	12	3.76 (2.30, 5.16)	4 of 10 (0)	no	0.3754	1.0333	4 (1)	0.00	1.00	0.35
B	24	3.49 (1.21, 5.25)	3 of 10 (0)	no	0.3838	1.0051	4 (1)	0.60	1.00	0.43
B	48	4.01 (3.18, 4.94)	5 of 10 (0)	no	0.2866	0.9965	6 (0)	0.10	1.00	0.58
C	6	1.15 (1.00, 1.55)	0 of 10 (10)	no	0.2274	1.0000	0 (10)	0.00	1.00	0.28
C	12	3.80 (2.47, 5.79)	5 of 10 (0)	no	0.3197	1.0308	4 (0)	0.30	1.00	0.33
C	24	3.80 (2.15, 4.64)	5 of 10 (0)	no	0.3474	1.0162	0 (1)	0.10	1.00	0.42
C	48	4.35 (3.23, 5.59)	7 of 10 (0)	yes	0.3330	0.9978	6 (1)	0.10	1.00	0.57

The second-round rule (Part 2G) and the scale at 0.02 on fresh seeds (Part 2G-b). G1, G2, G5, G6 PASS; G3, G4 FAIL; G7, G8 PASS (Table 32; Figure 2, right; Table 9). Part 2G reuses the frozen seeds of the cells it extends and reproduces the frozen rows of ETO, $\hat{\theta}_{I,\epsilon}$, $\hat{\theta}_{\mathrm{SURE}}$ and $\hat{\theta}_{\mathrm{safe}}$ before adding its own: 2710 of 2720 replicates reproduce every frozen row bitwise, and ten, the worst-conditioned of their cells, to 0.0004 in regret; no claim reads a row that failed to reproduce. JS-W is Hansen's form without the floor; the floored member of the family of Theorem S6 is reported next to it and differs in the fourth decimal except under adequacy at small n, where it cannot return ETO. The dropped-mass rule was capped at the preregistered 0.1 by a dated note before the pool started, after the timing stage on unregistered seeds had shown it choosing 0.2. At 0.02 on the fresh seeds of Part 2G-b, W drops 0.36 to 0.49 of its mass in the five cells of Table 9.

The hard-loss rule, the third attempt at the scale (Part 2G-c). G9, G10, G11 and E1-r PASS; R1c FAIL (Table 32; Table 9). Disclosed in its freeze as motivated by the record of the scale sweep, where the hard SPO+ training loss of the smoothed minimiser minus that of the SPO+ fit had the sign pattern of the regret gap. The rule fits the smoothed minimiser at $\epsilon_0 \in \{0.02, 0.05, 0.1\}$ on the training sample, computes the hard SPO+ training loss of each and of the SPO+ fit, and takes the largest scale, among the fits that meet the gradient rule, whose loss is within one percent of the smallest of the four; it reads no test set, no regret and no sandwich. One case was left open by that definition (fits meet the gradient rule and none is within the tolerance); it was completed before the freeze by the smallest hard loss, and that branch was taken in four cells only, all knapsack at degree six under multiplicative noise (0.3 of replicates at $n = 200$, 0.1 at three larger n). On the frozen seeds the rule returns 0.1 in every replicate at degrees one and two, moves in 0.14 of the benchmark replicates at degree four and in 0.61 at degree six (0.12 on the toy); the frozen rows at 0.1 reproduce bitwise in regret in every replicate. On the real prices the in-sample curve is flat (the 0.1 minimiser is within 0.0004 of the smallest loss at every capacity), the rule stays at 0.1 in all 40 splits, and JS-W at its scale repeats the miss at $C = 96$ (1.0272; $\hat{\theta}_{\mathrm{SURE}}$ 1.0302), where the effective dimension is below two and the rule returns the smoothed endpoint: the in-sample proxy does not see that endpoint effect. Reported without criterion: $\hat{\theta}_{\mathrm{SURE}}$ at the rule's scale passes and fails the same claims; a 5-fold cross-validated choice of the scale agrees with the rule in 0.34 to 0.53 of the benchmark replicates by degree and brings the ratio at $C = 96$ to 1.0149. In the 12 size cells the rule chose 0.1 in all 1200 replicates, so E1-r is the recorded count of E1, 61 of 1200: the claim says that the rule leaves the test alone under adequacy and nothing about a scale chosen under misspecification.

The Wald test and its recorded failure mode (from the main text of the first version). The first frozen test was the Wald statistic $T = n\hat{\Delta}^\top \hat{V}_\Delta^+ \hat{\Delta}$ on the eigenvectors retained by a fixed rank rule, with an F-type reference (asymptotically exact, Proposition S4) and the same null-calibrated refer-

ence applied to T. On the benchmarks the F reference over-rejected at degree one; the calibrated Wald test held its size (0.02 to 0.08 per cell at 100 replicates), but its critical value varied across the replicates of one cell by orders of magnitude where the smoothed Hessian is near singular (Tables 14 and 16), and only the Satterthwaite reference was conservative (0.014). The mechanism, read from the per-replicate records before any size cell had run and recorded as an amendment (Appendix A): where the smoothed Hessian has a near-flat direction (Remark S8), the pseudo-inverse of the sandwich amplifies it, the null copies of T explode along it, and their upper order statistic is dominated by the explosion. Directions in which the decision loss is flat carry no decision relevance, and V_{Δ}^{+} weights them most; T_W weights them by W, which is to say not at all. The Wald test stays in the record and is reported next to T_W on identical null draws in every cell (Table 8).

The Wald test under the non-trivial null (Part 2H). On the draws of Table 8 the F-referenced Wald test of the first design, whose size at degree one under correct specification already failed on the benchmarks (0.575), rejects in 0.453 of the knapsack replicates under heteroskedastic noise or an omitted symmetric component, where the calibrated T_W test holds 0.038 and the calibrated Wald test 0.043; on the toy all four references are within 0.012 to 0.060. The collapse is that of the reference, not of the calibration. Moved from Section 4.1: on the toy the preregistered N2 component fell almost entirely in the lineality space, so an addendum arm with it in the identified coordinates (N2-s) was preregistered after that pool had finished and before any of its results was read; a further arm, preregistered before its run, tested whether the over-rejection is a small-sample effect (H1-s-n).

Table 25: Neural, per cell: last-layer heads, the fixed step with the smallest cell-mean regret, the end-to-end networks, the S3(b) frequency and the rejection share of the last-layer T_W test.

noise	degree	n	FT-none	FT-full	SURE-FT	best fixed (cell mean)	MLP-MSE	MLP-SPO+$_\epsilon$	FT-all	S3(b)	T_W rej.
mult	1	2000	0.1590	0.1600	0.1590	0.1589	0.1602	0.1619	0.1599	0.90	0.05
mult	1	5000	0.1583	0.1589	0.1583	0.1583	0.1603	0.1595	0.1592	1.00	0.00
mult	1	10000	0.1573	0.1574	0.1572	0.1572	0.1568	0.1575	0.1568	1.00	0.00
mult	2	2000	0.1108	0.1114	0.1108	0.1106	0.1117	0.1104	0.1108	0.40	0.10
mult	2	5000	0.1072	0.1072	0.1071	0.1070	0.1031	0.1040	0.1027	0.65	0.20
mult	2	10000	0.1048	0.1047	0.1047	0.1046	0.1069	0.1051	0.1052	0.90	0.45
mult	4	2000	0.0797	0.0749	0.0772	0.0747	0.0773	0.0736	0.0726	0.10	0.60
mult	4	5000	0.0782	0.0733	0.0740	0.0737	0.0778	0.0715	0.0718	0.60	0.85
mult	4	10000	0.0754	0.0709	0.0713	0.0714	0.0756	0.0688	0.0689	0.65	1.00
mult	6	2000	0.1326	0.1043	0.1120	0.1059	0.1087	0.0754	0.0833	0.05	0.80
mult	6	5000	0.1045	0.0808	0.0846	0.0826	0.0963	0.0676	0.0722	0.45	1.00
mult	6	10000	0.0995	0.0721	0.0727	0.0744	0.0944	0.0622	0.0638	0.20	1.00
gauss	1	5000	0.1227	0.1231	0.1228	0.1228	0.1224	0.1239	0.1226	0.95	0.10
gauss	2	5000	0.0905	0.0908	0.0905	0.0904	0.0906	0.0911	0.0902	1.00	0.10
gauss	4	5000	0.0896	0.0863	0.0866	0.0864	0.0892	0.0845	0.0846	0.80	1.00
gauss	6	5000	0.1488	0.1265	0.1295	0.1285	0.1462	0.1187	0.1212	0.40	0.95

The per-round claim tables with their full conditions.

Table 26: Toy problem (Part 1C, frozen): criteria and verdicts as recorded.

Criterion	Condition	Recorded	Verdict
P1 size (F reference, degree 1)	in [0.03, 0.08] at $n = 500, 1000, 2000$	0.060 / 0.080 / 0.080 / 0.060 at $n = 200/500/1000/2000$	PASS
P1b size (bootstrap arm)	in [0.03, 0.08] at $n = 200, 500$	0.000 / 0.000 / 0.010 / 0.060	FAIL
P2 power (degree 6, $n = 500$)	at least 0.90	1.000	PASS
P3 SURE vs better endpoint	ratio at most 1.01 in all 16 cells	max 1.0224 (degree 6, $n = 200$); 14 of 16 within	FAIL
P4 SURE vs ETO at degree 1	ratio at most 1.01	max 1.0001	PASS
P5 best-of-both count	at most 1.01 in at least 14 of 16 cells	14 of 16	PASS

Table 27: First benchmark round (Part 2C, frozen): claims and verdicts as recorded.

Claim	Condition	Recorded	Verdict
C1 SURE vs better endpoint	at most 1.02 in every cell (80)	48 of 80 within; max 1.4333	FAIL
C1b SURE-safe vs better endpoint	at most 1.02 in every cell	63 of 80 within; max 1.1048	FAIL
C2 SURE vs ETO at degree 1	at most 1.01 in every degree-1 cell (20)	17 of 20 within; max 1.0628	FAIL
C3 SURE beats CV-select (shortest path)	at least 15 of 20 cells per noise model	mult 11 of 20; gauss 15 of 20	FAIL
C4 the scissors (shortest path)	sign of $\mathrm{IEO}_\epsilon - \mathrm{ETO}$ flips with the degree at every n	both noise models, every n (sign reading; Appendix note)	PASS

Table 28: Second benchmark round (Part 2D, frozen): claims and verdicts as recorded.

Claim	Condition	Recorded	Verdict
D1 adequacy	SURE at most 1.01 $\times$ ETO in every degree-1 cell (14)	max 1.0014	PASS
D2 regime	SURE at most 1.02 $\times$ min(ETO, IEO_ϵ) where S3(b) holds in at least 0.7 of replicates	32 of 56 cells qualify; max 1.0043	PASS
D3 safety	SURE-safe at most 1.05 $\times$ min(ETO, IEO_ϵ) in every cell with $n \geq 500$ (48)	max 1.0168	PASS
D4 calibrated size and power	pooled size in [0.03, 0.08], no cell above 0.15; power at least 0.9	pooled 0.043 (F reference 0.596); max cell 0.08; min power 0.90	PASS
D5 against cross-validation	SURE at most 1.02 $\times$ CV-select in every cell with $n \geq 1000$ (40) and below it in at least half	max 1.1096; wins 31 of 40	FAIL

Table 29: Third round (Part 2E, frozen): the regret-weighted test, claims and verdicts as recorded.

Claim	Condition	Recorded	Verdict
E1 size	calibrated T_W rejection at degree 1 pooled over 12 benchmark cells in [0.03, 0.08], no cell above 0.15	pooled 0.051; max cell 0.08	PASS
E2 power	calibrated T_W rejection at least 0.9 in each of 4 cells at degree 6, $n = 1000$	min 1.00	PASS
E3 (reported)	pooled size of the other references on the same replicates	Satterthwaite 0.014; calibrated Wald 0.048; F 0.575	reported

Table 30: Neural (Part 3N, frozen): claims and verdicts as recorded.

Claim	Condition	Recorded	Verdict
NL1 regime	SURE-FT at most 1.02 × min(FT-none, FT-full) where S3(b) holds in at least 0.7	4 of 12 cells qualify; max 1.0002	VOID
NL2 competitiveness	SURE-FT at most 1.05 × min(MLP-MSE, MLP-SPO+$_\epsilon$) for $n \geq 5000$ (8)	max 1.2519	FAIL
NL3 adequacy	SURE-FT at most 1.01 × FT-none at degree 1 (3)	max 1.0002	PASS
NL4 hand-set step	SURE-FT at most 1.02 × the fixed step with the smallest cell-mean regret, $n \geq 5000$ (8)	max 1.0247	FAIL
NL5 size	calibrated T_W rejection at degree 1 in [0.0, 0.15] (60 tests)	1 of 60 (0.017)	PASS

E THE FIRST THREE ROUNDS AND THE NEURAL ROUND IN DETAIL

This appendix keeps the account of the toy, the three benchmark rounds and the neural round as it stood before the revision rounds of Section 4; Table 31 lists every claim of those rounds and Table 32 every claim of the revision rounds (Parts 2G, 2G-b, 2G-c, 2H with its addendum arm, 4R, 4R-d and 3W).

Every experiment below was preregistered: grids, seeds, estimators, criteria and thresholds were committed and timestamped with a third-party anchor before any result existed, and every verdict is reported as recorded (Table 31; the per-round tables with the full conditions in Appendix D). Regret is the normalized regret of Tang & Khalil (2024) on a held-out test set of 10000 contexts; cell values are means over seeds; the better endpoint of a cell is $\min$ of the cell means of ETO and $\hat{\theta}_{I,\epsilon}$.

E.1 MECHANISM CHECK ON A TOY PROBLEM

Top-k selection ($d = 10$, $k = 3$, $p = 5$), the PyEPO polynomial family of degree $\{1, 2, 4, 6\}$, Gaussian noise, $n \in \{200, 500, 1000, 2000\}$, 100 replicates per cell (Table 26 in the appendix). In the first frozen round, whose test was the Wald statistic with the F reference, size and power were as preregistered, $\hat{\theta}_{\mathrm{SURE}}$ recovered ETO under adequacy to four decimals and was within one percent of the better endpoint in 14 of 16 cells, the two misses (within 2.3 percent) being the cells of lowest effective dimension, as Theorem 2 predicts; the bootstrap-referenced arm was conservative at small n and its criterion failed. The calibrated T_W test of the third round has size 0.02, 0.03, 0.08 and 0.06 at $n = 200, 500, 1000, 2000$ on the same problem.

E.2 BENCHMARKS: THE REGIME MAP AND THE TWO TESTS

Two PyEPO benchmarks: shortest path on the 5×5 grid ($d = 40$, $q_M = 96$; soft Bellman smoothing) and knapsack with ten items ($d = 10$, $q_M = 60$; entropic smoothing over the enumerated feasible subsets), degrees $\{1, 2, 4, 6\}$, two noise models (multiplicative uniform and additive symmetric Gaussian), ten seeds per cell. Baselines: PyEPO's SPO+, DBB (Vlastelica Pogančić et al., 2020), PFYL (Berthet et al., 2020), DDR (Loke et al., 2026) with its cross-validated weight, cross-validated selection between ETO and SPO+, the pretest estimators, Wald-geometry James-Stein (WG-JS, the negative control of Appendix A) where the sandwich has full rank, and the oracle best-of-two.

The first and second rounds: exact smoothing and the calibrated Wald test. The first frozen round (C4 PASS, the uniform claims and the perturbed knapsack smoothing failed) is in Appendix D. The second replaced the knapsack smoother by the exact entropic one, removed the rounding step of PyEPO's knapsack generator so that degree one is correctly specified (the rounded generator is kept as a robustness row), extended shortest path to $n \in \{5000, 10000\}$, and made the null-calibrated Wald reference primary. Of its claims (Table 31; full conditions in Table 28), D1 to D4 PASS and D5 FAIL. Under adequacy $\hat{\theta}_{\mathrm{SURE}}$ is at ETO in all 14 degree-one cells (max ratio 1.0014); where

Table 31: Every preregistered claim of the toy, the three benchmark rounds and the neural round, with its verdict as recorded; the first block is the pre-pivot design, closed 2026-09-15 (Reproducibility statement).

Round	Claim	Recorded	Verdict
Pre-pivot design, closed 2026-09-15 (full-space Wald statistic, constant of the Wald geometry; Phase B)			
Phase B	B1 size of the analytic statistic in [0.03, 0.08]	0.56 / 0.22 / 0.10 at n = 200 / 500 / 1000	FAIL
	B2 power at degree 6, $n = 500$, at least 0.9	1.00	PASS
	B3 DAS / IEO at most 1.02 in every cell	max 1.0123	PASS
	B4 DAS / ETO at most 1.01 at degree 1	max 1.0052	PASS
	B5 mean weight at degree 6, $n = 500$, at most 0.15	0.282	FAIL
	A1.3 Hausman-form gate, ρ_H at most 0.15 (Amendment 2, left unratified)	0.2102 at $n = 2000$	FAIL
Toy	P1 size (F reference)	0.080 / 0.080 / 0.060 at n = 500 / 1000 / 2000	PASS
	P1b size (bootstrap arm)	0.000 / 0.000 at n = 200 / 500	FAIL
	P2 power	1.000	PASS
	P3 SURE / better endpoint at most 1.01	max 1.0224; 14 of 16 cells within	FAIL
	P4 SURE / ETO at degree 1	max 1.0001	PASS
	P5 at least 14 of 16 cells within 1.01	14 of 16	PASS
First	C1 SURE / better endpoint at most 1.02	48 of 80 cells; max 1.4333	FAIL
	C1b SURE-safe / better endpoint at most 1.02	63 of 80 cells; max 1.1048	FAIL
	C2 SURE / ETO at degree 1 at most 1.01	17 of 20 cells; max 1.0628	FAIL
	C3 SURE beats CV-select in 15 of 20 cells	mult 11, gauss 15 of 20	FAIL
	C4 the scissors	sign flips with the degree at every n	PASS
Second	D1 SURE / ETO at degree 1 at most 1.01	max 1.0014	PASS
	D2 SURE / better endpoint at most 1.02 where S3(b) holds	32 of 56 cells qualify; max 1.0043	PASS
	D3 SURE-safe / better endpoint at most 1.05	max 1.0168	PASS
	D4 calibrated Wald size and power	size 0.043; min power 0.90	PASS
	D5 SURE / CV-select at most 1.02	max 1.1096; wins 31 of 40	FAIL
Third	E1 calibrated T_W size in [0.03, 0.08]	0.051; max cell 0.08	PASS
	E2 calibrated T_W power at least 0.9	min 1.00	PASS
Neural	NL1 SURE-FT / better head where S3(b) holds	4 of 12 cells qualify	VOID
	NL2 SURE-FT / better network at most 1.05	max 1.2519	FAIL
	NL3 SURE-FT / FT-none at degree 1 at most 1.01	max 1.0002	PASS
	NL4 SURE-FT / fixed step with the smallest cell mean at most 1.02	max 1.0247	FAIL
	NL5 last-layer T_W size in [0.0, 0.15]	0.017	PASS

Table 32: Every preregistered claim of the revision rounds, with its verdict as recorded. Part 2G is a second round on the frozen grids and seeds, disclosed as motivated by the frozen record; Part 2G-b is the second attempt at the scale question, on fresh seeds; the N2-s arm was added by a dated note after the Part 2H pool had finished and before any of its results was read; the Part 3W time box was lifted after its pool had finished and before its file was read; Part 4R-d was frozen before its run and is read with three notes recorded before any run (Appendix D).

Round	Claim	Recorded	Verdict
Part 2G	G1 JS-W within 1.02 of the better endpoint in at least 20 of the 24 low-S3(b) cells	23 of 24; max 1.0202	PASS
	G2 JS-W within 1.01 of ETO at degree one, $n \geq 500$ (12 cells)	max 1.0012; $n = 200$: 1.0202, 1.0146	PASS
	G3 $\hat{\theta}_{\mathrm{SURE}}$ at the dropped-mass rule within 1.02 of cross-validated selection in the five D5 misses	max 1.0841	FAIL
	G4 JS-W at the rule within 1.02 of cross-validated selection in 40 cells, below in at least 20	max 1.0699; below in 29	FAIL
	G5 JS-W within 1.01 of the better endpoint in the 16 toy cells	max 1.0059	PASS
	G6 JS-W-FT within 1.02 of the better head in 12 neural cells	max 1.0036	PASS
Part 2G-b	G7 JS-W at 0.02 within 1.02 of cross-validated selection in the five cells on fresh seeds, below in at least three	max 0.9977; below in 5	PASS
	G8 JS-W at 0.02 within 1.01 of ETO in the five degree-one cells	max 1.0020	PASS
Part 2G-c	G9 JS-W at the hard-loss rule within 1.02 of cross-validated selection in 40 cells, below in at least 20	max 1.0004; within in 40; below in 37	PASS
	G10 JS-W at the hard-loss rule within 1.01 of ETO in the 12 degree-one cells	max 1.0012	PASS
	G11 the same two conditions on fresh seeds, five degree-six and five degree-one cells	max 1.0031 and 1.0016	PASS
	R1c JS-W at the hard-loss rule within 1.02 of cross-validated selection on the real prices	max 1.0272	FAIL
	E1-r calibrated T_W size at the scale chosen by the hard-loss rule in [0.03, 0.08], no cell above 0.15	61 of 1200; cell max 0.08	PASS
Part 2H	H1 calibrated T_W size under N1 and N2 in [0.03, 0.08], no cell above 0.15	45 of 1200; cell max 0.08	PASS
	H2 $\hat{\theta}_{\mathrm{SURE}}$ and JS-W within 1.01 of ETO under N1 and N2, $n \geq 1000$	max 1.0001, 1.0010	PASS
N2-s	H1-s the same band on the three N2-s cells	29 of 300; cell max 0.15	FAIL
	H2-s $\hat{\theta}_{\mathrm{SURE}}$ and JS-W within 1.01 of ETO, $n \geq 1000$	max 0.9999, 1.0000	PASS
	H1-s-n the N2-s generator at $n = 5000$ and 10000, pooled size in [0.03, 0.08]	7 of 200	PASS
Part 4R	R1 $\hat{\theta}_{\mathrm{SURE}}$ within 1.02 of cross-validated selection in each capacity	max 1.0302	FAIL
	R1b JS-W at the rule within 1.02 of cross-validated selection	max 1.0185	PASS
	R2 $\hat{\theta}_{\mathrm{safe}}$ within 1.05 of the better endpoint	max 1.0000	PASS
	R3 SURE-FT within 1.02 of the fixed step with the smallest cell-mean regret	max 0.9327	PASS
	R3b SURE-FT within 1.05 of the better end-to-end network	max 1.0913	FAIL
	R4 SURE-FT within 1.02 of the better head where S3(b) holds in at least 0.7 of directions	0 cells qualify	VOID
	R5 calibrated T_W size on the real features in [0.02, 0.10]	6 of 200	PASS
Part 4R-d	R1d $\hat{\theta}_{\mathrm{SURE}}$ within 1.02 of the better endpoint at each capacity, item-specific head, scored arm	1.0464, 1.0038, 1.0139, 0.9983 at $C = 12, 24, 48, 96$	FAIL
	R2d $\hat{\theta}_{\mathrm{safe}}$ within 1.05 of the better endpoint at each capacity	1.0359, 1.0036, 1.0043, 0.9998	PASS
	R3d $\hat{\theta}_{\mathrm{SURE}}$ and JS-W within 1.02 of cross-validated selection at each capacity	$\hat{\theta}_{\mathrm{SURE}}$ 1.0066, 0.9908, 1.0144, 1.0044; JS-W 0.9770, 0.9903, 1.0023, 1.0065	PASS
	R4d JS-W within 1.01 of $\hat{\theta}_{I,\epsilon}$ at each qualifying capacity (all four qualify) and below it in at least half of the pooled splits	1.0155, 1.0034, 1.0019, 1.0004; below in 6 of 40	FAIL
Part 4S	R1s $\hat{\theta}_{\mathrm{SURE}}$ within 1.02 of the better endpoint at each capacity, Nord Pool, scored arm (A, $p = 34$, at $C = 6$, 12, 24; C, $p = 114$, at 48)	1.0002, 1.0620, 1.0053, 0.9908 at $C = 6, 12, 24, 48$	FAIL
	R2s $\hat{\theta}_{\mathrm{safe}}$ within 1.05 of the better endpoint at each capacity	1.0029, 1.0004, 0.9997, 0.9950	PASS
	R3s $\hat{\theta}_{\mathrm{SURE}}$ and JS-W within 1.02 of cross-validated selection at each capacity	$\hat{\theta}_{\mathrm{SURE}}$ 1.0769, 1.4269, 0.9757, 0.8375; JS-W 1.0767, 1.3449, 0.9703, 0.8435	FAIL
	R4s JS-W within 1.01 of $\hat{\theta}_{I,\epsilon}$ at each qualifying capacity (one qualifies, $C = 48$) and below it in at least half of the pooled splits	0.9978; below in 6 of 10	PASS
Part 3W	W2 SURE-FT within 1.05 of the better end-to-end network	ratio 4.5485	FAIL
	W3 SURE-FT within 1.02 of the mean of the per-seed best fixed step	ratio 1.0776	FAIL
	W4 SURE-FT within 1.02 of the better head if S3(b) holds in at least 0.7 of directions	S3(b) 0.50; ratio 0.9564	VOID

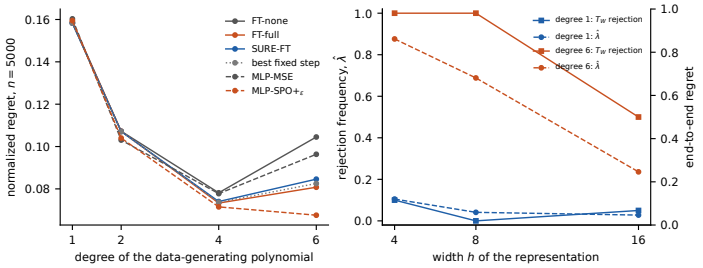

Figure 6: Neural. Left: regret by degree at $n = 5000$ for the last-layer heads and the end-to-end networks. Right: the capacity sweep, the rejection share of the last-layer test and $\hat{\lambda}$ by width at degrees one and six.

the S3(b) condition holds in at least 0.7 of replicates, 32 of the 56 cells, $\hat{\theta}_{\mathrm{SURE}}$ is within 1.0043 of the better endpoint; $\hat{\theta}_{\mathrm{safe}}$ is within 1.0168 of it in all 48 cells with $n \geq 500$; the calibrated Wald test has pooled size 0.043 at degree one with power at least 0.90 at degree six. D5 fails in five cells (35 of 40 within 1.02), all at degree six under multiplicative noise, where $\hat{\theta}_{\mathrm{SURE}}$ sits at its smoothed endpoint ($\hat{\lambda}$ 0.96 to 0.99) and the endpoint itself trails PyEPO's SPO+ by 2 to 11 percent (Table 14 in the appendix). A preregistered sweep of the smoothing scale on those five cells and their degree-one counterparts (Table 15 in the appendix) shows the gap falling with the scale, 0.97 to 0.99 at $\epsilon_0 = 0.02$ against 1.05 to 1.08 at the preregistered 0.1, at no adequacy cost ($\hat{\theta}_{\mathrm{SURE}}$ over ETO 0.9996 to 1.0003 at degree one at every scale); its one claim failed in one of five cells (1.0368 of cross-validated selection, through the step, not the endpoint). At the scale the preregistered rule of that claim selects, the one with the smallest gap on the test set (0.02 in all five cells), $\hat{\theta}_{\mathrm{SURE}}$ is below cross-validated selection in four of the five cells (0.9725 to 0.9864); in the fifth (knapsack, multiplicative noise, degree six, $n = 1000$) the endpoint is below SPO+ and the miss is the step, $\hat{\theta}_{\mathrm{SURE}}$ at 1.0480 of its better endpoint with a standard deviation of 0.199 in the plug-in step across seeds. The mechanism: at the preregistered scale the endpoint trails SPO+ and the step sits on it; at the small scale the endpoint is repaired and what remains is the variance of the plug-in step at effective dimension below four, the case Theorem S6 covers and where $\hat{\theta}_{\mathrm{safe}}$ is the estimator to use. The rule was frozen before the run, no scale was selected after the fact, and every table of this paper keeps the preregistered $\epsilon_0 = 0.1$. Figure 2 (left) draws every cell of both rounds.

The third round: the regret-weighted test. The regret-grid diagnostics of the second round exposed the mechanism of Section 3.3 before any size cell had run; the amendment that moved the test to the regret geometry was recorded at that moment, and the third round was frozen after it, with claims E1 (calibrated T_W size pooled over the same twelve degree-one cells), E2 (power at degree six, $n = 1000$, in four cells) and E3 (the other references on identical null draws, reported), conditions in Table 29: E1 and E2 PASS. The calibrated T_W test rejects 61 of the 1200 degree-one replicates (pooled size 0.051, no cell above 0.08) and every degree-six replicate at $n = 1000$; on the same null draws the calibrated Wald test has pooled size 0.048, the Satterthwaite reference 0.014 (conservative at these n), the F reference 0.575 and the uncorrected chi-square reference 0.696. The two calibrated tests differ in their critical values, not their size: on knapsack the Wald critical value varies across replicates by orders of magnitude and tracks the condition number of $\hat{H}_\epsilon$, while the T_W critical value is tight and does not (Figure 5 and Table 16 in the appendix); on shortest path, where the Hessian is well conditioned, both are tight.

Reading the regime map. The frequency of the S3(b) condition rises with n on both benchmarks and the ratios fall with it; the cells that miss the 1.02 line are the small-n knapsack cells at degrees four and six and the first-round shortest-path cells at $n \leq 500$, where the effective dimension is below four. At degree two and $n = 1000$ the calibrated T_W test rejects in 0.25 to 0.70 of replicates, the regime in which $\hat{\lambda}$ sits inside the segment. On the same draws the calibrated Wald test rejects in 0.40 to 0.80 (0.95 to 1.00 at degree six, against 1.00): T_W gives no weight to directions in which the decision loss is flat, so an alternative that lives there is invisible to it by design.

E.3 NEURAL: SURE FINE-TUNING OF THE LAST LAYER

Shortest path on the 5×5 grid, the network and the cross-fitted pipeline of Section 3.5 with $h = 8$, degrees $\{1, 2, 4, 6\}$, $n \in \{2000, 5000, 10000\}$, multiplicative noise, ten seeds per cell (Gaussian noise at $n = 5000$ as a sensitivity arm). The frozen claims NL1 to NL5 are those of Table 31, with the full conditions in Table 30: NL3 and NL5 PASS, NL2 and NL4 FAIL, NL1 VOID.

What holds. On the learned representation the calibrated T_W test keeps its size (1 rejection in 60 tests at degree one) and SURE-FT returns the two-stage head under adequacy (ratio to FT-none at most 1.0002). The rejection share and the step rise with the degree: 0.00 to 0.05 and $\hat{\lambda}$ of 0.02 to 0.06 at degree one, 0.10 to 0.45 and 0.07 to 0.29 at degree two, 0.60 to 1.00 and 0.30 to 0.79 at degree four, 0.80 to 1.00 and 0.51 to 0.89 at degree six. The data-chosen step is within 1.02 of the fixed step with the smallest cell-mean regret, chosen with hindsight, in 7 of 8 cells, while that step moves from 0.25 to 0.75 across the cells. *What fails.* NL1 is void: the S3(b) condition holds in at least 0.7 of replicate-directions in only 4 of the 12 cells, so the dominance statement has no bite at this width; the oracle inequality and SURE-safe apply, and SURE-safe is within 1.0095 of the better head in every cell. NL4 fails in one cell (1.0247). NL2 fails at degree six: SURE-FT is within 1.0393 of the better end-to-end network at degrees one and two and 3.5 to 4.8 percent above it at degree four, but at degree six it is 1.2519 and 1.1682 of it at $n = 5000$ and 10000, and FT-full, the best head, is 38.4, 19.5 and 15.9 percent above the end-to-end decision-focused network at $n = 2000, 5000, 10000$ (Figure 6, left). With an MSE representation the last layer cannot recover what the representation did not learn (the degree-six head fits meet the gradient rule in 0.45 to 0.85 of directions under the capped budget, the residual gradient lying in directions with little decision relevance, Appendix D). On a learned representation the effective dimension is below four in every degree-six cell under multiplicative noise, so SURE-safe, within 1.0095 of the better head in every cell, is the estimator to deploy there, with SURE-FT reserved for the cells where the condition holds.

The learning diagnostics. Training the representation on the smoothed decision loss brings FT-full to within 4.3 percent of the end-to-end decision-focused network at degree six, yet the last-layer test still rejects in 1.00 of directions: a rejection says that the last layer should move, not that moving it is enough. Width helps every head at degree six (FT-full 0.1829, 0.0808 and 0.0740 at widths 4, 8 and 16) and changes nothing under adequacy; at width 16 the head problem is nearly degenerate at this sample size (the gradient rule met in 0.00 of fits, rejection share 0.50; SURE-FT at 1.1372 of the better head and SURE-safe at 1.0036, Table 19), a finding about identification, not adequacy (Figure 6, right). *Comparators and the cost of cross-fitting* (two preregistered addendum arms with no claim; Table 18 in the appendix). PFYL is within 0.979 to 1.039 of the better of the two end-to-end networks in every cell and is the best end-to-end network at degree six (0.979 to 0.999 of the smoothed SPO+ network); DBB is 1.43 to 3.84 times the better network. The smoothed SPO+ network at width 16 is 0.951 to 1.004 of the width-8 network, so the end-to-end curve flattens where the last-layer curve does. Training the representation and the head on the full sample, without cross-fitting and outside the theory, brings FT-full to 0.811 to 0.973 of its cross-fitted value at degree six and leaves it 11.7 to 12.2 percent above the end-to-end decision-focused network under multiplicative noise (3.7 under Gaussian): cross-fitting explains part of the degree-six gap and the representation the rest.

F HISTORY OF THE PREREGISTRATION

The two paragraphs below are the round-by-round history of the preregistration, moved from the reproducibility statement without change: the first design and its closure, then every amendment and dated note that bears on the reading of the results.

A first design preceded every experiment above: a full-space analytic Wald statistic for the non-smooth SPO+ estimator with the shrinkage constant of the Wald geometry, preregistered and run (Phase B: size FAIL, 0.56, 0.22 and 0.10 at $n = 200$, 500 and 1000; power PASS; the two regret criteria PASS; the weight criterion FAIL), and an assumption gate that failed (A1.3). It was closed on 2026-09-15 against the kill line frozen in the project charter, and the present design started from that record after a projected variant, adopted and abandoned the same day after a smoke test only (`PIVOT_NOTE_2` in the supplementary material); its diagnosis, a shrinkage constant calibrated in the wrong geometry, is the origin of Section 3.4. No verdict of that design enters this paper; its

frozen Wald constant, 35 to 47 on the toy, is the numerator of the factor 1.7 to 8.5 of Section 3.2 (Table 31, first block).

Amendment 9 moved the test from the Wald geometry to the regret geometry and amended the submission rule after 40 regret cells of the second round had been read and before any size cell existed; the calibrated Wald claim it hedged against (D4) passed, so the amendment did not bind. *Amendment 11* recorded the authors' rulings on an internal review after every first-round result was frozen: it announced the revision rounds (Parts 2G, 2H, 4R), each frozen before its run, disclosed Part 2G as a second round on grids whose results were known, and lifted the time box of the Warcraft experiment after its pool had finished and before its result file was read; the pool's log, which prints regrets rounded to four decimals, was visible before that ruling, and no rule, threshold or claim was changed. *The note to Part 2G* (dated, before the Part 2G pool started, no result existing): JS-W was set to Hansen's form without the floor, the floored variant kept as a reported row, and the scale rule was capped at 0.1, after the theory freeze and the timing stage on unregistered seeds. *The note to Part 2H* (dated, after the Part 2H pool finished and before any of its result files was read): the claims H1 and H2 were kept as frozen, the toy N2 component was recorded as lying in the lineality space, and the arm N2-s was added with its own claims. *Part 2G-b* (a second-round Part with its own freeze, made after the results of Part 2G were read and before its own run): the fixed scale 0.02 with JS-W was preregistered as the second attempt at the scale question after two failed claims, on fresh seeds. *Part 5L* (the exploratory barrier rows of Appendix B, no claim) was announced by Amendment 11 with a time box of six hours of work; the authors lifted the box and set its grid before any of its code or results existed, and it was frozen with its code before its run. *Part 2G-c* (frozen after every record above was read and before its own run): the in-sample hard-loss rule was preregistered as the third attempt at the scale question, disclosed as motivated by the hard-loss gaps recorded in the scale sweep; one case the rule's definition left open was completed by the recorder before the freeze and its frequency is reported (four cells); its claim on the real prices failed and is reported. *A second note to Part 2H* (frozen before its run) added the large-n cells of the N2-s arm with the claim H1-s-n and its reading fixed in advance; H1-s stands as FAIL. *Part 4R-d* (the item-specific head on the real prices) was preregistered with its claims R1d to R4d, the thresholds of Parts 4R and 2G-c and the reading of every outcome, frozen with its code before its run and run on the authors' go; three notes recorded before any run fixed the fit budget of this Part (2600 L-BFGS iterations, a mechanical change recorded with its reason), qualification on converged fits only, and Arm C as conditional on the effective dimension of arms A and B, which triggered it at $C = 12$ only; a fourth note, recorded while the pool ran and before the summary stage had read any result, added a cutoff option to that stage, which was not used. The non-convergence at $C = 12$ that a synthetic stand-in had led us to expect did not occur on the real prices: the main-scale fit met the gradient rule in every split of every arm. The machine idle-slept during part of the pool; the replicates are deterministic given their seeds, so this entered no result. *Part 4S* (a second market for the linear pipeline, the Nord Pool day-ahead prices) was preregistered on 2026-09-21 as a second market for the review discussion and run before submission on the authors' instruction, after five notes recorded before the run (the summary stage; arms B and C conditional on S and A; days with 23 or 25 hours dropped and counted, none found; per-replicate checkpoints; the wall-clock estimate with a cutoff that did not fire): 176 replicates, no failure; its results were read before the paper changed, and the readings fixed in 4S.5 applied. The proof of Proposition S7-d was added after every size experiment was frozen; it changes the status of the calibrated reference from assumed to proved and no experiment. In a re-run from the frozen seeds ten of 2720 replicates reproduced the frozen record to 0.0004 in regret rather than bitwise (the worst-conditioned of their cells). Every number in this paper traces to a named frozen summary through a machine-checked registry, and an automated audit fails the build on any numeral without a source. The code, the frozen result files, the preregistration with every verdict, and the timestamp proofs are in the anonymous supplementary repository, with absolute paths rewritten to relative ones.

G STATEMENTS, ASSUMPTIONS AND PROOFS

This appendix states every assumption and proves every statement the main text uses. The statements keep the names under which the main text cites them (Lemma S1, Theorem 1-S, Proposition 1-S, Theorems S2 and S3, Proposition S4, Corollary S5, Theorem S6, Proposition S7); a main-text statement is a part of the statement of the same name here. Table 33 gives the status of each claim: *proved* here, *assumed* (stated as an assumption and not proved), or *cited*; the last column names

the test file of the supplementary repository that checks the claim numerically, independently of the proof.

Table 33: Status of every claim of this appendix; the test files are in the folder `tests` of the supplementary repository.

Claim	Status	Depends on	Numerical verification
Assumption A (moments)	assumed		
Assumptions B and C (density; unique SPO+ target)	assumed; used only in Lemma S1(d)		
Assumption C-ϵ (nondegenerate V_Δ)	assumed		
Lemma 5 (Stein's lemma)	cited (Stein, 1981)		
Lemma S1(a) to (c) (smoothing; convexity; Fisher consistency)	proved	A for (b)	`test_smooth.py`
Lemma S1(d) (limit $\epsilon \to 0$)	proved (one convexity result cited)	A, B, C	`test_smooth.py`
Theorem 1-S (joint normality; the Wald statistic)	proved (M-estimation theorems cited)	A, C-ϵ	`test_smooth.py`
Proposition 1-S (sandwich consistency)	proved (uniform law cited)	Theorem 1-S	`test_smooth.py`
Assumption L-S (local sequence)	assumed; a sufficient condition is sketched		
Theorem S2 (risk along the path)	proved	L-S	`test_sure.py`
Theorem S3 (SURE: exact risk, dominance, oracle inequality)	proved	L-S, Lemma 5	`test_sure.py` (b), (c), Hausman case
Proposition S4 (Wald test, F reference)	(a) to (c) proved; (d) cited, motivation only	Theorem 1-S, Proposition 1-S	`test_smooth.py`
Assumptions D and R (regret expansion; regret consistency)	assumed		
Corollary S5 (regret transfer)	proved under the assumed D (and R for (i)); limitation stated	L-S, D, R	
Theorem S6 (largest safe constant; SURE-safe)	proved	L-S, Lemma 5	`test_sure_safe.py` (c)
Proposition S7 (the T_W test)	(a), (b), (e) proved; (c) cited, an approximation; (d) proved for the sign-flip reference as implemented (Proposition S7-d: entropic smoothing, conditionally symmetric errors under H_0)	Theorem 1-S, Proposition 1-S, Corollary S5	`test_prop_s7.py` (a), (c), (d)

G.1 Notation and standing assumptions

Data and decision problem. The data (z_i, c_i), $i = 1, \dots, n$, are i.i.d. copies of (z, c) with context $z \in \mathbb{R}^p$ and cost vector $c \in \mathbb{R}^d$. A fixed feature map $\phi = \phi(z) \in \mathbb{R}^{q_0}$ contains an intercept. The feasible set $X \subset \mathbb{R}^d$ is a polytope with vertex set $V(X)$, $R_X := \max_{x \in X} |x|$. The oracle is $x^*(c) \in \arg\min_{x \in X} c^\top x$; ties are broken by a fixed rule, so x^* is a single-valued map $\mathbb{R}^d \to V(X)$ (ties occur only on a Lebesgue-null set, which is irrelevant under Assumption B below).

Normal fan. For $v \in V(X)$ let $N_v := \{w \in \mathbb{R}^d : w^\top (v - v') \le 0 \text{ for all } v' \in V(X)\}$ be the (closed) normal cone of v, so that $x^*(w) = v$ whenever $w \in \operatorname{int} N_v$. The cones $\{N_v\}$ tile $\mathbb{R}^d$. For $v \ne v'$ put $n_{vv'} := (v - v')/|v - v'|$ and $H_{vv'} := \{w : n_{vv'}^\top w = 0\}$; then $\partial N_v \subset \bigcup_{v' \ne v} H_{vv'}$.

Lineality space and identified coordinates. Let $L := \{w \in \mathbb{R}^d : w^\top x \text{ is constant on } X\}$ (the lineality space of the normal fan; $L = \{0\}$ iff X is full-dimensional) and $M := L^\perp$, the direction

space of the affine hull of X. Every $v - v'$ lies in M. For $\ell \in L$ one has $x^*(w + \ell) = x^*(w)$ and, for every $x \in X$, $(w + \ell)^\top x = w^\top x + \text{const}$. Let P_M be the orthogonal projector onto M and $\Pi := I_{q_0} \otimes P_M$.

Linear predictor and vectorisation. $\hat{c} = \Theta\phi$ with $\Theta \in \mathbb{R}^{d \times q_0}$; $\theta := \mathrm{vec}(\Theta) \in \mathbb{R}^q$, $q = dq_0$, column-major. Then

$$\hat{c} = (\phi^\top \otimes I_d)\,\theta, \qquad \frac{\partial \hat{c}}{\partial \theta} = \phi^\top \otimes I_d, \qquad (\phi \otimes I_d)\,g = \mathrm{vec}(g\phi^\top) \quad (g \in \mathbb{R}^d), \tag{5}$$

so that everything below is linear in θ. The unidentified parameter directions are $\mathrm{range}(I - \Pi) = \mathbb{R}^{q_0} \otimes L$: adding $\mathrm{vec}(\ell a^\top)$, $\ell \in L$, $a \in \mathbb{R}^{q_0}$, to θ shifts $\hat{c}$ by $(a^\top\phi)\ell \in L$ and changes neither the decision $x^*(\hat{c})$ nor the SPO+ loss. We write $q_M := q_0 \dim M$ for the number of identified coordinates.

Losses, scores, Hessians.

$$s_E(\theta) := \nabla_\theta \ell_E = -2(\phi \otimes I_d)(c - \Theta\phi), \qquad H_E := \nabla^2_\theta R_E = 2\,\mathbb{E}[\phi\phi^\top] \otimes I_d. \tag{6}$$

SPO+ loss (IEO): with $\psi(u) := \max_{x \in X}(-u)^\top x = -u^\top x^*(u)$ (convex, piecewise linear, R_X-Lipschitz, $\nabla\psi(u) = -x^*(u)$ off the hyperplanes $H_{vv'}$),

$$\ell_I(\theta; z, c) := \max_{x \in X}(c - 2\hat{c})^\top x + 2\hat{c}^\top x^*(c) - c^\top x^*(c) = \psi(2\hat{c} - c) + 2\hat{c}^\top x^*(c) - c^\top x^*(c), \tag{7}$$

convex and piecewise linear in θ, with the subgradient

$$s_I(\theta) := 2(\phi \otimes I_d)\big(x^*(c) - x^*(2\Theta\phi - c)\big), \tag{8}$$

which is the gradient whenever $2\Theta\phi - c \notin \bigcup H_{vv'}$. Note $s_I(\theta) \in \mathrm{range}(\Pi)$ always, since $x^*(c) - x^*(\cdot) \in M$.

$R_E(\theta) := \mathbb{E}\ell_E(\theta)$, $R_I(\theta) := \mathbb{E}\ell_I(\theta)$, with minimisers θ^*_E and θ^*_I. R_I is constant along $\mathrm{range}(I - \Pi)$, so θ^*_I is at best unique modulo $\mathbb{R}^{q_0} \otimes L$; we adopt throughout the *anchoring convention*

$$\theta^*_I := \Pi\theta^*_I + (I - \Pi)\theta^*_E, \qquad \hat{\theta}_I \in \hat{\theta}_E + \mathrm{range}(\Pi) \text{ minimises the empirical SPO+ risk,} \tag{9}$$

so that $\hat{\Delta} := \hat{\theta}_I - \hat{\theta}_E \in \mathrm{range}(\Pi)$ identically. (A subgradient method started at $\hat{\theta}_E$ preserves equation 9 exactly because $s_I \in \mathrm{range}(\Pi)$; any other minimiser is projected onto $\hat{\theta}_E + \mathrm{range}(\Pi)$, which does not change its empirical SPO+ risk.) The same convention is used for the smoothed pair below. For the nonsmooth pair, *decision adequacy* is the hypothesis $\mathrm{H}_{0,I} : \Pi\theta^*_I = \Pi\theta^*_E$; the hypothesis H_0 of the main text is its counterpart for the smoothed pair (Definition S).

Assumption A (moments)**.** $\mathbb{E}|\phi|^4 < \infty$, $\mathbb{E}|c|^4 < \infty$, and $\mathbb{E}[\phi\phi^\top]$ is positive definite.

Assumption B (density; used only in Lemma S1(d))**.** For a.e. z, $c \mid z$ has a Lebesgue density on $\mathbb{R}^d$. Under this assumption the vector $2\hat{c} - c$ lies on $\bigcup H_{vv'}$ with probability zero for every θ, so ℓ_I is a.s. differentiable at every θ with gradient equation 8.

Assumption C (uniqueness of the SPO+ target; used only in Lemma S1(d))**.** $\Pi\theta^*_I$ is unique.

Assumption C-ϵ (nondegeneracy for the smoothed pair)**.** The limit covariance V^ϵ_Δ of Theorem 1-S is positive definite on $\mathrm{range}(\Pi)$.

Lemma 5 (Stein's lemma; Stein, 1981, Lemma 2)**.** *If $U \sim \mathcal{N}(\mu, I_r)$ and $g : \mathbb{R}^r \to \mathbb{R}^r$ is Lipschitz, then $\mathbb{E}[g(U)^\top(U - \mu)] = \mathbb{E}[\nabla \cdot g(U)]$.*

Throughout, $z^*(w) := \min_{x \in X} w^\top x = -\psi(w)$ is the value function (concave, piecewise linear, R_X-Lipschitz, $\nabla z^*(w) = x^*(w)$ off the hyperplanes $H_{vv'}$), $N_X := |V(X)|$, and $\xi_i := (z_i, c_i)$.

G.2 S.1 Smoothings and the smoothed loss

Definition S (Smoothed value functions)**.** For $\epsilon > 0$:

(E) *Entropic (exact) smoothing.* $z_\epsilon(w) := -\epsilon \log \sum_{v \in V(X)} \exp(-w^\top v/\epsilon)$, with the Gibbs weights $p_v(w) := \exp(-w^\top v/\epsilon)/\sum_{v'} \exp(-w^\top v'/\epsilon)$. For top-$k$ the sum is the elementary symmetric polynomial $e_k(\exp(-w/\epsilon))$ and is computed by the $O(dk)$ recursion in log space; for the shortest path on a DAG it is the soft dynamic programme (log-sum-exp Bellman recursion over paths).

(G) *Gaussian perturbation (fallback).* $z_\epsilon(w) := \mathbb{E}[z^*(w + \epsilon Z)]$, $Z \sim \mathcal{N}(0, I_d)$ (Berthet et al., 2020); in the implementation the expectation is replaced by an average over M common random numbers.

In both cases $x_\epsilon(w) := \nabla z_\epsilon(w)$. The *smoothed SPO+ loss* is

$$\ell_\epsilon(\theta; z, c) := -z_\epsilon(2\hat{c} - c) + 2\hat{c}^\top x_\epsilon(c) - z_\epsilon(c), \qquad \hat{c} = \Theta\phi(z), \tag{10}$$

i.e. the SPO+ loss equation 7 of the regularised decision problem, with population risk $R_\epsilon(\theta) := \mathbb{E}\,\ell_\epsilon(\theta; z, c)$, minimiser θ^*_ϵ under the anchoring convention equation 9, and the hypothesis of decision adequacy of the main text is $\mathrm{H}_0 : \Pi\theta^*_\epsilon = \Pi\theta^*_E$. The main text uses the entropic smoothing (E) throughout.

Lemma S1 (Properties of the smoothing and of ℓ_ϵ)**.** *Let $\epsilon > 0$ be fixed and z_ϵ be either smoothing of Definition S.*

(a) Value function. *z_ϵ is concave and C^∞ on $\mathbb{R}^d$; $x_\epsilon(w) \in X$ (in case (E), $x_\epsilon(w) = \sum_v p_v(w)v$, the soft selection; in case (G), $x_\epsilon(w) = \mathbb{E}[x^*(w + \epsilon Z)]$); for $\ell \in L$, $z_\epsilon(w + \ell) = z_\epsilon(w) + \ell^\top v_0$ ($v_0 \in V(X)$ arbitrary, the constant of the lineality space) and $x_\epsilon(w + \ell) = x_\epsilon(w)$; $z^*(w) - \epsilon \log N_X \le z_\epsilon(w) \le z^*(w)$ in case (E), so that $|z_\epsilon - z^*| \le \epsilon \log |V(X)|$, and $|z_\epsilon(w) - z^*(w)| \le \epsilon R_X \mathbb{E}|Z| \le \epsilon R_X \sqrt{d}$ in case (G). The Hessian is*

$$-\nabla^2 z_\epsilon(w) = \frac{1}{\epsilon}\operatorname{Cov}_{p(w)}(v) \quad \text{(E)}, \qquad -\nabla^2 z_\epsilon(w) = -\frac{1}{\epsilon}\,\mathbb{E}\big[x^*(w + \epsilon Z)\, Z^\top\big] \quad \text{(G)},$$

symmetric positive semidefinite, with null space containing L, $\|\nabla^2 z_\epsilon\| \le R_X^2/\epsilon$ in case (E) and $\le R_X\sqrt{d}/\epsilon$ in case (G), and positive definite on M for every w (in case (E) because $p(w)$ has full support on $V(X)$, whose affine span has direction space M; in case (G) because ψ is not affine along any line in direction $h \in M \setminus \{0\}$ and the Gaussian law charges every open set).

(b) Convexity and smoothness of the loss. *$\ell_\epsilon(\cdot; z, c)$ is convex and C^∞ in θ, $6R_X|\phi|$-Lipschitz, with*

$$\begin{aligned} s_\epsilon(\theta) &:= \nabla_\theta \ell_\epsilon = 2(\phi \otimes I_d)\big(x_\epsilon(c) - x_\epsilon(2\hat{c} - c)\big) \in \operatorname{range}(\Pi), \\ \nabla^2_\theta \ell_\epsilon &= 4\,(\phi\phi^\top) \otimes \big(-\nabla^2 z_\epsilon(2\hat{c} - c)\big) \succeq 0, \end{aligned} \tag{11}$$

*$|s_\epsilon| \le 4R_X|\phi|$ and $\|\nabla^2_\theta \ell_\epsilon\| \le 4|\phi|^2 R_X^2/\epsilon$ (case (E)); moreover $\ell_\epsilon \ge 0$, with $\ell_\epsilon = 2\epsilon \operatorname{Ent}(p(c)) \ge 0$ when $\hat{c} - c \in L$ in case (E) (*Ent *the Shannon entropy of the Gibbs weights). Under Assumption A, R_ϵ is convex, C^2, with $\nabla R_\epsilon = \mathbb{E}s_\epsilon$ and Hessian $H_\epsilon(\theta) := 4\,\mathbb{E}[(\phi\phi^\top) \otimes (-\nabla^2 z_\epsilon(2\Theta\phi - c))]$, positive definite on $\operatorname{range}(\Pi)$ for every θ; hence $\Pi\theta^*_\epsilon$ is unique and the anchored minimiser is well defined. We write $H_\epsilon := H_\epsilon(\theta^*_\epsilon)$.*

(c) Fisher consistency for every ϵ. *If $\mathbb{E}[c \mid z] = \Theta_0\phi(z)$ and the conditional law of $e := c - \Theta_0\phi$ given z is centrally symmetric ($e \mid z \overset{d}{=} -e \mid z$), then $\nabla R_\epsilon(\theta_0) = 0$ and $\theta^*_\epsilon = \theta_0 = \theta^*_E$ (modulo $\mathbb{R}^{q_0} \otimes L$), so H_0 holds for every $\epsilon > 0$ and for both smoothings. The consistent smoothing of both terms of equation 10 is what makes this work: with $x^*(c)$ in place of $x_\epsilon(c)$ in the linear term the population gradient at θ_0 would be $2\mathbb{E}[(\phi \otimes I_d)(x^*(c) - x_\epsilon(\Theta_0\phi - e))] \neq 0$ in general.*

(d) The $\epsilon \to 0$ limit. *Under Assumptions A and B, $R_\epsilon \to R_I$ uniformly on compact sets as $\epsilon \to 0$, and $\Pi\theta^*_\epsilon \to \Pi\theta^*_I$ whenever $\Pi\theta^*_I$ is unique (Assumption C). In particular H_0 for all small ϵ and $\mathrm{H}_{0,I}$ are compatible; they are not equivalent at a fixed ϵ (only Fisher consistency (c) gives H_0 exactly). The same holds for the empirical risks $R_{n,\epsilon} \to R_{n,I}$ at fixed n.*

Proof. (a) Case (E): $w \mapsto \log \sum_v \exp(-w^\top v/\epsilon)$ is a log-sum-exp of affine maps, hence convex and C^∞; its gradient is $-\sum_v p_v v/\epsilon$ and its Hessian $\operatorname{Cov}_p(v)/\epsilon^2$ (the standard exponential-family identities), which gives the formulas after multiplying by $-\epsilon$. $x_\epsilon(w) = \sum_v p_v(w)v$ is a convex combination of vertices, so it lies in X. Directly, $\max_v e^{-w^\top v/\epsilon} \le \sum_v e^{-w^\top v/\epsilon} \le N_X \max_v e^{-w^\top v/\epsilon}$ gives $z^*(w) - \epsilon \log N_X \le z_\epsilon(w) \le z^*(w)$. For $\ell \in L$, $\ell^\top v$ is the same for all $v \in X$, so

$p(w+\ell)=p(w)$ and $z_\epsilon(w+\ell)=z_\epsilon(w)+\ell^\top v_0$. $\|\mathrm{Cov}_p(v)\|\le \mathbb{E}_p|v|^2\le R_X^2$. The null space: $\mathrm{Cov}_p(v)\ell=\mathbb{E}_p[(v-\bar v)(v-\bar v)^\top\ell]=0$ because $(v-\bar v)^\top\ell=0$. Positive definiteness on M: for $h\in M\setminus\{0\}$, $h^\top\mathrm{Cov}_p(v)h=\mathrm{Var}_p(h^\top v)>0$ because p charges every vertex and $h^\top v$ is not constant on $V(X)$ (otherwise $h\in L$). Case (G): $z_\epsilon=z^*\ast\varphi_\epsilon$ is a Gaussian convolution of a Lipschitz concave function, hence concave and C^∞; $\nabla z_\epsilon(w)=\mathbb{E}[\nabla z^*(w+\epsilon Z)]=\mathbb{E}[x^*(w+\epsilon Z)]$ (differentiation under the expectation, z^* Lipschitz, the non-differentiability set is Lebesgue-null), a mixture of vertices, hence in X; $\nabla^2 z_\epsilon(w)=\nabla_w\mathbb{E}[x^*(w+\epsilon Z)]=\epsilon^{-1}\mathbb{E}[x^*(w+\epsilon Z)Z^\top]$ by Stein's identity applied coordinatewise ($\mathbb{E}[f(w+\epsilon Z)Z]=\epsilon\,\mathbb{E}[\nabla f(w+\epsilon Z)]$ for Lipschitz f), with norm at most $R_X\mathbb{E}|Z|/\epsilon$; symmetry follows from $\nabla^2 z_\epsilon$ being a Hessian. Invariance along L is inherited from z^*. $|z_\epsilon(w)-z^*(w)|\le\mathbb{E}|z^*(w+\epsilon Z)-z^*(w)|\le\epsilon R_X\mathbb{E}|Z|$. Positive definiteness on M: $-h^\top\nabla^2 z_\epsilon(w)h=\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbb{E}\,\psi(w+th+\epsilon Z)|_{t=0}$; the function $t\mapsto\psi(u+th)$ is convex piecewise linear with at least one kink for every u when $h\in M\setminus\{0\}$ (its slope is $-h^\top x^*(u+th)$, which takes at least two values because $h^\top x$ is not constant on X), and the Gaussian smoothing of a convex function with a kink of positive size at a point charged by the (positive) Gaussian density has positive second derivative.

(b) $\theta\mapsto 2\hat c-c$ is affine, $-z_\epsilon$ is convex, and $2\hat c^\top x_\epsilon(c)$ is linear: ℓ_ϵ is convex and C^∞. The chain rule with equation 5 gives equation 11 (a factor 2 from $\partial(2\hat c-c)/\partial\hat c$ in the gradient and 4 in the Hessian); $x_\epsilon(c)-x_\epsilon(2\hat c-c)\in X-X\subset M$ gives the range statement, $|x_\epsilon(c)-x_\epsilon(\cdot)|\le 2R_X$ the score bound, and the Lipschitz constant is that of equation 7. Nonnegativity: by concavity of z_ϵ, $z_\epsilon(2\hat c-c)\le z_\epsilon(c)+2x_\epsilon(c)^\top(\hat c-c)$, so $\ell_\epsilon\ge 2\big(c^\top x_\epsilon(c)-z_\epsilon(c)\big)$; in case (E) the Gibbs variational identity $z_\epsilon(c)=\sum_v p_v c^\top v-\epsilon\,\mathrm{Ent}(p)$ makes the right side $2\epsilon\,\mathrm{Ent}(p(c))\ge 0$, with equality in the concavity step iff $2\hat c-c-c\in L$; in case (G), $c^\top x_\epsilon(c)-z_\epsilon(c)=\mathbb{E}[(c-(c+\epsilon Z))^\top x^*(c+\epsilon Z)]+c^\top(x_\epsilon(c)-\mathbb{E}x^*(c+\epsilon Z))\ge\mathbb{E}[z^*(c)-z^*(c+\epsilon Z)+\epsilon Z^\top x^*(c+\epsilon Z)]$, and $z^*(c)\le c^\top x^*(c+\epsilon Z)$ gives the same lower bound ≥ 0. Under Assumption A the derivatives are dominated ($|s_\epsilon|\le 4R_X|\phi|$, $\|\nabla^2\ell_\epsilon\|\le 4|\phi|^2R_X^2/\epsilon$, both square integrable resp. integrable), so differentiation under $\mathbb{E}$ is justified and $R_\epsilon\in C^2$ with the stated Hessian; $a^\top H_\epsilon(\theta)a=4\mathbb{E}[(A\phi)^\top(-\nabla^2 z_\epsilon)(A\phi)]=0$ for $a=\mathrm{vec}(A)$ forces $P_M A\phi=0$ a.s. by (a), hence $P_M A\,\mathbb{E}[\phi\phi^\top]A^\top P_M=0$ and $a\in\mathbb{R}^{q_0}\otimes L$. A convex function that is strictly convex on the identified coordinates has a unique identified minimiser if it has one; existence: $R_\epsilon\ge 0$ and, on $\mathrm{range}(\Pi)$, R_ϵ is coercive because $\ell_\epsilon\ge 2(c^\top x_\epsilon(c)-z_\epsilon(c))+$ [z_ϵ-Bregman term] and the Bregman term of the strictly concave z_ϵ grows linearly along every ray in M (the slope of $-z_\epsilon$ along $h\in M$ tends to $-\min_{x\in X}h^\top x>-\max$, i.e. $-z_\epsilon(u+th)$ is eventually affine with slope $-\min_x h^\top x$ which differs from the initial slope), so $R_\epsilon(\theta)\to\infty$ along every identified ray.

(c) Conditionally on z, with $m:=\Theta_0\phi$: at θ_0, $2\hat c-c=m-e$ and $c=m+e$, so $s_\epsilon(\theta_0)=2(\phi\otimes I_d)(x_\epsilon(m+e)-x_\epsilon(m-e))$, whose conditional expectation vanishes because e and $-e$ have the same conditional law. Hence $\nabla R_\epsilon(\theta_0)=0$; by convexity θ_0 minimises R_ϵ, and by the uniqueness in (b) $\Pi\theta^*_\epsilon=\Pi\theta_0$. $\theta^*_E=\theta_0$ because the conditional mean is correctly specified. The last claim is immediate from the same computation.

(d) For a.e. c (no ties, Assumption B) $x_\epsilon(c)\to x^*(c)$ as $\epsilon\to 0$: in case (E) $p_v(c)\to\mathbf{1}\{v=x^*(c)\}$; in case (G) $x^*(c+\epsilon Z)\to x^*(c)$ a.s. and is bounded. With the bounds of (a), $\ell_\epsilon(\theta;\xi)\to\ell_I(\theta;\xi)$ for a.e. ξ and every θ, dominated by $2R_X(|2\hat c-c|+|c|)+\epsilon\log N_X$ (resp. $+\epsilon R_X\sqrt{d}$), integrable under Assumption A; so $R_\epsilon\to R_I$ pointwise. Pointwise convergence of finite convex functions is uniform on compacts (Rockafellar, 1970, Theorem 10.8), and the argmin of a convex function with a unique minimiser is continuous under uniform-on-compacts convergence (the level sets of R_I below $R_I(\theta^*_I)+1$ are bounded on $\mathrm{range}(\Pi)$ by uniqueness and convexity, and the minimisers of R_ϵ eventually lie in them). The empirical statement is the same argument with the empirical measure. □

G.3 S.2 Joint asymptotics and the sandwich for the smooth pair

Theorem 1-S (Joint asymptotic normality, smooth loss; the Wald statistic)**.** *Let Assumption A hold, let $\epsilon>0$ be fixed, let $\hat\theta_{I,\epsilon}\in\hat\theta_E+\mathrm{range}(\Pi)$ minimise the empirical smoothed SPO+ risk $R_{n,\epsilon}(\theta):=$*

$n^{-1}\sum_i \ell_\epsilon(\theta;\xi_i)$, *and let Assumption C-*ϵ *hold. Then*

$$\sqrt{n}\begin{pmatrix}\hat\theta_E-\theta_E^*\\ \Pi(\hat\theta_{I,\epsilon}-\theta_\epsilon^*)\end{pmatrix}\xrightarrow{d}\mathcal{N}(0,V^\epsilon),$$
$$V^\epsilon=D_\epsilon^+\Sigma^\epsilon D_\epsilon^+,\quad D_\epsilon=\mathrm{diag}(H_E,H_\epsilon),\quad \Sigma^\epsilon=\mathrm{Cov}\big(s_E(\theta_E^*),s_\epsilon(\theta_\epsilon^*)\big), \tag{12}$$

with H_ϵ^+ *the Moore-Penrose inverse on* $\mathrm{range}(\Pi)$. *No smoothness assumption on the conditional law of c is needed: the smoothing supplies it. Moreover, let* $\hat\Delta:=\hat\theta_{I,\epsilon}-\hat\theta_E$, $V_\Delta^\epsilon:=V_{II}^\epsilon+V_{EE}^\epsilon-V_{IE}^\epsilon-V_{EI}^\epsilon$ *and* $r:=\mathrm{rank}(V_\Delta^\epsilon)$ *(*$=q_M$ *under Assumption C-*ϵ*).*

(a) Under H_0, $\sqrt{n}\,\hat\Delta\xrightarrow{d}Z_\Delta\sim\mathcal{N}(0,V_\Delta^\epsilon)$.

(b) Let $\hat V_\Delta\xrightarrow{p}V_\Delta^\epsilon$ *and let* $\hat r$ *be the rank rule* $\hat r:=\#\{j:\lambda_j(\hat V_\Delta)>\epsilon_{\mathrm{rank}}\,\lambda_{\max}(\hat V_\Delta)\}$ *with the fixed* $\epsilon_{\mathrm{rank}}=10^{-6}$, *and* $\hat V_\Delta^+$ *the pseudo-inverse on the retained eigenvectors. If* $\hat r\xrightarrow{p}r$ *(automatic when* $\hat V_\Delta$ *has exact zeros on* $\mathbb{R}^{q_0}\otimes L$, *as the sandwich of Proposition 1-S does), then under* H_0

$$T:=n\,\hat\Delta^\top\hat V_\Delta^+\hat\Delta\xrightarrow{d}\chi^2_r.$$

(c) Under a fixed alternative $\Delta^*:=\Pi(\theta_\epsilon^*-\theta_E^*)\neq 0$, $T/n\xrightarrow{p}\Delta^{*\top}(V_\Delta^\epsilon)^+\Delta^*>0$, *so the test that rejects when* $T>\chi^2_{r,1-\alpha}$ *is consistent.*

Rank deficiency of V_Δ^ϵ *in the full parameter space arises exactly from the directions* $\mathbb{R}^{q_0}\otimes L$ *that decisions do not identify (the common-shift directions of top-k, the node-potential directions of shortest path); those directions carry* $\hat\Delta=0$ *and are excluded by the pseudo-inverse.*

Proof. Identified coordinates. Let $B_M\in\mathbb{R}^{q\times q_M}$ be an orthonormal basis of $\mathrm{range}(\Pi)$. Under equation 9 the smoothed IEO estimator is $\hat\theta_{I,\epsilon}=\hat\theta_E+B_M\hat\beta$ with $\hat\beta\in\mathbb{R}^{q_M}$ minimising $\beta\mapsto\sum_i\ell_\epsilon(\hat\theta_E+B_M\beta;\xi_i)=\sum_i\ell_\epsilon(B_M\beta+(I-\Pi)\hat\theta_E;\xi_i)$, where the last equality uses the invariance of ℓ_ϵ along $\mathbb{R}^{q_0}\otimes L$ (Lemma S1(a)); its population counterpart has the unique minimiser $\beta^*:=B_M^\top\theta_\epsilon^*$ (Lemma S1(b)). The stacked parameter is $\vartheta:=(\theta_E,\beta)\in\mathbb{R}^{q+q_M}$; both losses are convex in ϑ, the stacked empirical objective is $M_n(\vartheta):=n^{-1}\sum_i[\ell_E(\theta_E;\xi_i)+\ell_\epsilon(B_M\beta;\xi_i)]$, and $(\hat\theta_E,\hat\beta)$ minimises it; $\mathsf{s}_\epsilon:=(s_E,B_M^\top s_\epsilon)$ is the stacked score.

Two routes, both standard. *M-estimation* (van der Vaart, 1998, Theorem 5.23, applied to the stacked objective): the stacked loss is Lipschitz in ϑ with a square-integrable constant ($2|\phi|(|c|+|\Theta\phi|)$ and $6R_X|\phi|$ locally), differentiable at ϑ^* for every ξ, and the stacked population risk is C^2 at ϑ^* with the nonsingular Hessian $\mathsf{D}_\epsilon=\mathrm{diag}(H_E,B_M^\top H_\epsilon B_M)$ (Lemma S1(b)); consistency of $\hat\vartheta$ follows from convexity and the uniqueness of the identified minimiser (Lemma S1(b); Newey & McFadden, 1994, Theorem 2.7). The theorem gives $\sqrt{n}(\hat\vartheta-\vartheta^*)=-\mathsf{D}_\epsilon^{-1}n^{-1/2}\sum_i\mathsf{s}_\epsilon(\vartheta^*;\xi_i)+o_p(1)$ and the CLT concludes. *Convex-process route.* The basic corollary of Hjort & Pollard (1993) states: if $A_n(s)$, $s\in\mathbb{R}^m$, are random convex functions with $A_n(s)=s^\top U_n+\frac12 s^\top\mathsf{D}s+r_n(s)$, where D is symmetric positive definite, $U_n=O_p(1)$, and $r_n(s)\to0$ in probability for each fixed s, then the minimiser α_n of A_n satisfies $\alpha_n=-\mathsf{D}^{-1}U_n+o_p(1)$. Take $A_n(s):=\sum_i[\ell(\vartheta^*+s/\sqrt{n};\xi_i)-\ell(\vartheta^*;\xi_i)]$ with $\ell=\ell_E+\ell_\epsilon\circ B_M$, which $\sqrt{n}(\hat\vartheta-\vartheta^*)$ minimises, and decompose it into $s^\top U_n$ with $U_n:=n^{-1/2}\sum_i\mathsf{s}_\epsilon(\vartheta^*;\xi_i)$, the population term $n[R(\vartheta^*+s/\sqrt{n})-R(\vartheta^*)]=\frac12 s^\top\mathsf{D}_\epsilon s+o(1)$ (R_E is exactly quadratic and $R_\epsilon\in C^2$ by Lemma S1(b)), and the centred sum of the increments $D_i(t):=\ell(\vartheta^*+t;\xi_i)-\ell(\vartheta^*;\xi_i)-t^\top\mathsf{s}_\epsilon(\vartheta^*;\xi_i)$ at $t=s/\sqrt{n}$. The increment of the squared loss is $t^\top(\phi\phi^\top\otimes I_d)t$, with second moment $O(|t|^4)$ under Assumption A, and the increment of the smoothed loss is $D_\epsilon(t):=\ell_\epsilon(\theta^*+t)-\ell_\epsilon(\theta^*)-t^\top s_\epsilon(\theta^*)\in[0,2R_X^2|\phi|^2|t|^2/\epsilon]$ (Taylor with the Hessian bound of Lemma S1(b)), so that $\mathrm{Var}[D_\epsilon(t)]\le 4R_X^4\mathbb{E}|\phi|^4|t|^4/\epsilon^2=o(|t|^2)$. Hence the centred sum has variance at most $n\,\mathbb{E}[D(s/\sqrt{n})^2]=O(n^{-1})\to0$ and Chebyshev's inequality gives $r_n(s)\to0$ in probability for each fixed s. The scores $\mathsf{s}_\epsilon(\vartheta^*;\xi_i)$ are i.i.d., mean zero, with finite covariance $\Sigma:=\mathrm{diag}(I,B_M^\top)\,\Sigma^\epsilon\,\mathrm{diag}(I,B_M)$ (Assumption A gives $\mathbb{E}|s_E|^2<\infty$; $|s_\epsilon|\le 4R_X|\phi|$); the multivariate CLT gives $U_n\xrightarrow{d}\mathcal{N}(0,\Sigma)$ and the corollary gives $\sqrt{n}(\hat\vartheta-\vartheta^*)\xrightarrow{d}\mathcal{N}(0,\mathsf{D}_\epsilon^{-1}\Sigma\mathsf{D}_\epsilon^{-1})$. *Mapping back.* $\Pi(\hat\theta_{I,\epsilon}-\theta_\epsilon^*)=B_M(\hat\beta-\beta^*)$ and $B_M(B_M^\top H_\epsilon B_M)^{-1}B_M^\top=H_\epsilon^+$

because $\mathrm{range}(H_\epsilon) = \mathrm{range}(\Pi)$ by Lemma S1(b); since $s_\epsilon \in \mathrm{range}(\Pi)$, $B_M B_M^\top s_\epsilon = s_\epsilon$, which yields equation 12.

(a) $\sqrt{n}\hat{\Delta} = \sqrt{n}\,\Pi(\hat{\theta}_{I,\epsilon} - \theta^*_\epsilon) - \sqrt{n}\,\Pi(\hat{\theta}_E - \theta^*_E)$ under H_0; apply equation 12 and the continuous mapping theorem, giving $Z_\Delta = Z_I - Z_E$ with covariance V_Δ. (b) On the event $\hat{r} = r$, whose probability tends to one, the pseudo-inverse is a continuous function of $\hat{V}_\Delta$ on the manifold of rank-r PSD matrices, so $\hat{V}_\Delta^+ \xrightarrow{p} V_\Delta^+$; then $T \xrightarrow{d} Z_\Delta^\top V_\Delta^+ Z_\Delta$, which is χ^2_r because $Z_\Delta \in \mathrm{range}(V_\Delta)$ a.s. and $V_\Delta^{+1/2} Z_\Delta \sim \mathcal{N}(0, I_r)$ on that range. (c) $\hat{\Delta} \xrightarrow{p} \Delta^*$, $\hat{V}_\Delta^+ \xrightarrow{p} V_\Delta^+$, and V_Δ^+ is positive definite on $\mathrm{range}(\Pi) \ni \Delta^*$ by Assumption C-ϵ. □

Proposition 1-S (Analytic sandwich for the smooth pair)**.** *Under the assumptions of Theorem 1-S let*

$$\hat{H}_\epsilon := \frac{1}{n}\sum_{i=1}^n \nabla^2_\theta \ell_\epsilon(\hat{\theta}_{I,\epsilon}; \xi_i) = \frac{4}{n}\sum_{i=1}^n (\phi_i \phi_i^\top) \otimes \big(- \nabla^2 z_\epsilon(2\hat{\Theta}_{I,\epsilon}\phi_i - c_i)\big),$$

$$\hat{\Sigma}^\epsilon := \widehat{\mathrm{Cov}}\big((s_E(\hat{\theta}_E; \xi_i), s_\epsilon(\hat{\theta}_{I,\epsilon}; \xi_i))_{i \le n}\big),$$

with $\nabla^2 z_\epsilon$ from automatic differentiation of the closed form (E) (or the perturbed-oracle identity in case (G)), and $\hat{V}^\epsilon := \hat{D}_\epsilon^+ \hat{\Sigma}^\epsilon \hat{D}_\epsilon^+$, $\hat{D}_\epsilon = \mathrm{diag}(\hat{H}_E, \Pi\hat{H}_\epsilon\Pi)$. Then $\hat{H}_\epsilon \xrightarrow{p} H_\epsilon$, $\hat{\Sigma}^\epsilon \xrightarrow{p} \Sigma^\epsilon$, $\hat{V}^\epsilon \xrightarrow{p} V^\epsilon$ and $\hat{V}^\epsilon_\Delta \xrightarrow{p} V^\epsilon_\Delta$; in case (E) moreover $\hat{H}_\epsilon - H_\epsilon = O_p(n^{-1/2})$. Consequently Theorem 1-S(b),(c) apply with $\hat{V}^\epsilon_\Delta$, and in the local experiment of Assumption L-S below the plug-in statistics T_ϵ and T_W (Theorem S3) have the limits stated there. No finite differences and no bandwidth enter; the rate is the parametric one, against $O_p(n^{-2/5})$ for a finite-difference Hessian of the nonsmooth loss.

Proof. $\theta \mapsto \nabla^2_\theta \ell_\epsilon(\theta; \xi)$ is continuous and dominated on every bounded set by $4|\phi|^2 R_X^2/\epsilon$ (integrable under Assumption A), and $\hat{\theta}_{I,\epsilon}$ is consistent; the uniform law of large numbers over a compact neighbourhood of θ^*_ϵ (Newey & McFadden, 1994, Lemma 2.4) gives $\hat{H}_\epsilon \to H_\epsilon$. In case (E) the third derivative of ℓ_ϵ in θ is bounded by $8|\phi|^3\|\nabla^3 z_\epsilon\| \le c\,|\phi|^3 R_X^3/\epsilon^2$ (the third cumulant of the Gibbs law), so $\theta \mapsto \nabla^2 \ell_\epsilon$ is Lipschitz with an integrable constant, and $\hat{H}_\epsilon - H_\epsilon = [n^{-1}\sum_i \nabla^2 \ell_\epsilon(\theta^*_\epsilon; \xi_i) - H_\epsilon] + O_p(|\hat{\theta}_{I,\epsilon} - \theta^*_\epsilon|) = O_p(n^{-1/2})$ by the CLT for the first bracket (finite second moment under Assumption A) and Theorem 1-S. The scores are continuous in θ and dominated by square-integrable envelopes ($|s_\epsilon| \le 4R_X|\phi|$), so $\hat{\Sigma}^\epsilon \to \Sigma^\epsilon$ by the same uniform law. The map $(D, \Sigma) \mapsto D^+\Sigma D^+$ is continuous at fixed rank, and $\Pi\hat{H}_\epsilon\Pi$ has exact zeros on $\mathbb{R}^{q_0} \otimes L$ so that the rank rule of Theorem 1-S(b) is met. □

Reduction to identified coordinates. From here on we work in the identified coordinates and drop Π from the notation: θ stands for $\Pi\theta$, V_{EE} for $\Pi V_{EE}\Pi$, V_{IE} for $V_{IE}\Pi$, and inverses of matrices supported on $\mathrm{range}(\Pi)$ are Moore-Penrose inverses there. This is without loss for every statement below: $\hat{\Delta}$ and every estimator on the path live in $\hat{\theta}_E + \mathrm{range}(\Pi)$, and the unidentified component $(I - \Pi)\hat{\theta}_E$ is common to all of them. For weight matrices that are block diagonal with respect to $\mathrm{range}(\Pi) \oplus \mathrm{range}(I - \Pi)$ (this covers $W = I$, the regret weight W_L of Assumption D, which annihilates $\mathbb{R}^{q_0} \otimes L$, and $W = V_\Delta^+$), the common component contributes the same amount to every weighted risk and cancels from every risk difference. When X is full-dimensional nothing changes.

The Hausman case. Suppose that under H_0 the ETO estimator is asymptotically efficient for θ^* within a regular class of estimators that contains the smoothed IEO (for instance Gaussian noise with a correctly specified conditional mean, where ETO is the maximum-likelihood estimator). Then $\mathrm{Cov}(Z_E, Z_\Delta) = 0$ (Hausman, 1978, Lemma 2.1: otherwise $\hat{\theta}_E + a\hat{\Delta}$ would have smaller asymptotic variance than $\hat{\theta}_E$ for a suitable small a), i.e. $V_{IE} = V_{EE}$, so that $V_\Delta = V_{II} - V_{EE}$ and $C := \mathrm{Cov}(Z_I, Z_\Delta) = V_{II} - V_{IE} = V_\Delta$. This is the classical Hausman form. It is not assumed anywhere below; it is the special case $C = V_\Delta$.

G.4 S.3 The path and its exact asymptotic risk

Assumption L-S (Local misspecification, smooth pair)**.** $(P_n)_{n \ge 1}$ is a sequence of data-generating processes for (z, c) such that

(L1) Assumptions A and C-ϵ hold under every P_n with constants ($\mathbb{E}|\phi|^4$, $\mathbb{E}|c|^4$, $\lambda_{\min}\mathbb{E}[\phi\phi^\top]$, eigenvalue bounds on H_E, H_ϵ, V_Δ) uniform in n;

(L2) the population targets satisfy $\theta^*_{E,n} =: \theta_n \to \theta^*$ and $\theta^*_{\epsilon,n} = \theta_n + \delta/\sqrt{n}$ for a fixed $\delta \in \mathrm{range}(\Pi)$;

(L3) the conclusion of Theorem 1-S holds under P_n with the same limit covariance V^ϵ (evaluated at θ^* under $P = \lim P_n$): $\sqrt{n}(\hat\theta_E - \theta_n, \hat\theta_{I,\epsilon} - \theta^*_{\epsilon,n}) \xrightarrow{d} (Z_E, Z_I) \sim \mathcal{N}(0, V^\epsilon)$;

(L4) $W \succeq 0$ is block diagonal in the identified coordinates and the matrix $B = V_\Delta^{1/2} W V_\Delta^{1/2}$ of equation 15 is positive definite on $\mathbb{R}^r$ (i.e. $W \succ 0$ on $\mathrm{range}(\Pi)$; for a singular positive semidefinite W all statements hold on $\mathrm{range}(B)$ after replacing r by $\mathrm{rank}\, B$ when A maps $\mathrm{range}(B)$ into itself, e.g. in the Hausman case).

From here on the superscript ϵ is dropped: V, V_Δ, $C = V_{II} - V_{IE}$, $G = \delta + Z_\Delta$ and the quantities of equation 14 and equation 15 refer to the smooth pair.

This is the contiguity-style local framework of Lan et al. (2025), here imposed on the two targets rather than derived. A sufficient condition is the location family $c = \Theta_0\phi(z) + n^{-1/2}g(z) + \varepsilon$ with ε independent of z, symmetric, with $\mathbb{E}|\varepsilon|^4 < \infty$, and $\mathbb{E}|g(z)|^4 < \infty$: then $\theta_n = \theta_0 + n^{-1/2}\gamma_E$ with $\gamma_E = (\mathbb{E}[\phi\phi^\top]^{-1} \otimes I_d)\mathbb{E}[(\phi \otimes I_d)g(z)]$ exactly, and $\theta^*_{\epsilon,n} = \theta_0 + n^{-1/2}\gamma_I + o(n^{-1/2})$ with $\gamma_I = -H_\epsilon^+\, \partial_t \nabla R_\epsilon(\theta_0; t)|_{t=0}$ by the implicit function theorem applied to $\nabla R_\epsilon(\theta; t) = 0$ for the mean shift $t\, g(z)$ (available in its textbook form because $R_\epsilon \in C^2$ jointly in θ and the mean shift), so that $\delta = \Pi(\gamma_I - \gamma_E)$; and (L3) holds because the proof of Theorem 1-S goes through for triangular arrays with the Lindeberg CLT, all $o(1)$ terms being uniform under (L1). This sufficient condition is stated with a sketch only and is tagged *assumed*.

Limit experiment. Under Assumption L-S, with $Z_\Delta := Z_I - Z_E \sim \mathcal{N}(0, V_\Delta)$ and $G := \delta + Z_\Delta$,

$$\sqrt{n}(\hat\theta_E - \theta_n) \to Z_E, \qquad \sqrt{n}(\hat\theta_{I,\epsilon} - \theta_n) \to \delta + Z_I, \qquad \sqrt{n}\,\hat\Delta \to G.$$

Measuring every estimator against the smoothed IEO target $\theta^*_{\epsilon,n}$, the centred limit errors of the two endpoints and of an estimator $\hat\theta_{I,\epsilon} - w\,\hat\Delta$ with a weight $w = w(G) \in [0,1]$ are

$$E_{\mathrm{IEO}} = Z_I, \qquad E_{\mathrm{ETO}} = Z_E - \delta = Z_I - G, \qquad E_w = Z_I - w\,G. \tag{13}$$

For a symmetric positive semidefinite W the *asymptotic weighted risk* of an estimator with centred limit error E is $\rho_W(E) := \mathbb{E}[E^\top W E]$ (the risk of the limit law; it coincides with the trimmed-loss definition of Hansen (2016), $\lim_{\zeta\to\infty} \liminf_n \mathbb{E}[\min\{\zeta, n(\hat\theta - \theta^*_{\epsilon,n})^\top W(\hat\theta - \theta^*_{\epsilon,n})\}]$).

Standardised coordinates. Let $r = \mathrm{rank}\, V_\Delta$, $V_\Delta^{1/2}$ and $V_\Delta^{-1/2}$ the symmetric square roots on $\mathrm{range}(V_\Delta)$, and

$$\begin{aligned} U := V_\Delta^{-1/2} G \sim \mathcal{N}(\mu, I_r), \quad \mu := V_\Delta^{-1/2}\delta, \\ C := \mathrm{Cov}(Z_I, Z_\Delta) = V_{II} - V_{IE}, \qquad \mathbb{E}[Z_I \mid G] = C V_\Delta^{-1}(G - \delta), \end{aligned} \tag{14}$$

the last by joint Gaussianity of (Z_I, G). Define the $r \times r$ matrices

$$A := V_\Delta^{1/2}\, W\, C\, V_\Delta^{-1/2}, \qquad A_s := \tfrac{1}{2}(A + A^\top), \qquad B := V_\Delta^{1/2}\, W\, V_\Delta^{1/2} \succeq 0. \tag{15}$$

Note $\mathrm{tr}(A) = \mathrm{tr}(A_s) = \mathrm{tr}(WC)$ and, for every $u \in \mathbb{R}^r$, $u^\top A u = u^\top A_s u$.

A risk identity. For any weight $w = w(U)$, from equation 13, $\rho_W(E_w) = \mathbb{E}[Z_I^\top W Z_I] - 2\mathbb{E}[w\, G^\top W Z_I] + \mathbb{E}[w^2 G^\top W G]$. Conditioning on G (equivalently on U) and using equation 14, $\mathbb{E}[w\, G^\top W Z_I] = \mathbb{E}[w\, G^\top W C V_\Delta^{-1}(G - \delta)] = \mathbb{E}[w\, U^\top A(U - \mu)]$ and $\mathbb{E}[w^2 G^\top W G] = \mathbb{E}[w^2 U^\top B U]$, with A, B as in equation 15 (substitute $G = V_\Delta^{1/2}U$, $\delta = V_\Delta^{1/2}\mu$). Hence

$$\rho_W(E_w) - \rho_W(E_{\mathrm{IEO}}) = \mathbb{E}[w^2\, U^\top B U] - 2\,\mathbb{E}[w\, U^\top A(U - \mu)], \qquad \rho_W(E_{\mathrm{IEO}}) = \mathrm{tr}(W V_{II}), \tag{16}$$

and the case $w \equiv 1$ gives, by direct Gaussian moments,

$$\rho_W(E_{\mathrm{ETO}}) - \rho_W(E_{\mathrm{IEO}}) = \mathrm{tr}(B) - 2\,\mathrm{tr}(A) + \mu^\top B \mu. \tag{17}$$

Path and traces. $\hat{\theta}(\lambda) := \hat{\theta}_E + \lambda\hat{\Delta}$, $\lambda \in [0, \lambda_{\max}]$. Define the scalars

$$\begin{gathered} a_E := \operatorname{tr}(WV_{EE}), \quad a_{EI} := \operatorname{tr}(WV_{EI}) = \operatorname{tr}(WV_{IE}), \quad a_{II} := \operatorname{tr}(WV_{II}), \\ \tau_W := \operatorname{tr}(WV_\Delta) = \operatorname{tr} B, \quad \kappa_W := \operatorname{tr}(WC) = a_{II} - a_{EI} = \operatorname{tr} A, \end{gathered} \tag{18}$$

and the loss-adapted statistic $T_W := G^\top WG = U^\top BU$, with $\mathbb{E}[T_W] = \delta^\top W\delta + \tau_W = \mu^\top B\mu + \tau_W$. The rule of Section 3.4 is $\hat{q} := (T_W - \tau_W)_+$ and $\hat{\lambda} := \operatorname{clip}\big((\hat{q}+a_E-a_{EI})/(\hat{q}+\tau_W),\, 0,\, \lambda_{\max}\big)$ (with the plug-in traces and the empirical $T_W = n\hat{\Delta}^\top W\hat{\Delta}$ in the finite-sample version; by Proposition 1-S they converge jointly to the limit-experiment quantities).

Theorem S2 (Exact asymptotic risk along the path; the oracle λ^*)**.** *Under Assumption L-S, for every fixed $\lambda \in \mathbb{R}$ the centred limit error of $\hat{\theta}(\lambda)$ relative to $\theta^*_{\epsilon,n}$ is $E_\lambda = Z_I - (1-\lambda)G = (1-\lambda)(Z_E - \delta) + \lambda Z_I$, and*

$$\begin{aligned} \rho_W(\lambda) := \rho_W(E_\lambda) &= (1-\lambda)^2\big(\delta^\top W\delta + a_E\big) + \lambda^2 a_{II} + 2\lambda(1-\lambda)a_{EI} \\ &= a_{II} - 2(1-\lambda)\kappa_W + (1-\lambda)^2\big(\delta^\top W\delta + \tau_W\big). \end{aligned} \tag{19}$$

(a) ρ_W is a convex quadratic in λ with leading coefficient $\delta^\top W\delta + \tau_W > 0$, minimised at

$$\begin{aligned} \lambda^* &= \frac{\delta^\top W\delta + a_E - a_{EI}}{\delta^\top W\delta + \tau_W} = 1 - \frac{\kappa_W}{\delta^\top W\delta + \tau_W}, \\ \rho_W(\lambda^*) &= \rho_W(\mathrm{IEO}) - \frac{\kappa_W^2}{\delta^\top W\delta + \tau_W} = \rho_W(\mathrm{ETO}) - (\lambda^*)^2\big(\delta^\top W\delta + \tau_W\big), \end{aligned} \tag{20}$$

with $\rho_W(\mathrm{IEO}) = \rho_W(1) = a_{II}$ and $\rho_W(\mathrm{ETO}) = \rho_W(0) = \delta^\top W\delta + a_E$ (consistent with equation 17).

(b) $\lambda^* \le 1$ iff $\kappa_W \ge 0$; $\lambda^* \ge 0$ iff $\delta^\top W\delta + a_E \ge a_{EI}$. In the Hausman case $V_{EI} = V_{EE}$ $\kappa_W = \tau_W$ and $\lambda^* = \delta^\top W\delta/(\delta^\top W\delta + \tau_W) \in [0,1)$, equal to 0 exactly at $\delta = 0$. Under H_0 ($\delta = 0$) in general, $\lambda^* = (a_E - a_{EI})/\tau_W$, which is positive whenever $\operatorname{tr}(WV_{EE}) > \operatorname{tr}(WV_{EI})$ (an inefficient ETO in the W-metric): some IEO weight then reduces variance.

(c) $\lambda^* \to 1$ as $|\mu| \to \infty$ at rate $\kappa_W/(\mu^\top B\mu)$, and $\rho_W(\mathrm{IEO}) - \rho_W(\lambda^*) = O(|\mu|^{-2})$.

Proof. $\hat{\theta}(\lambda) - \theta^*_{\epsilon,n} = (\hat{\theta}_E - \theta_n) - \delta/\sqrt{n} + \lambda\hat{\Delta}$ and $\sqrt{n}\hat{\Delta} \to G = \delta + Z_\Delta$, so the limit error is $Z_E - \delta + \lambda G = Z_I - (1-\lambda)G$. Then $\rho_W(\lambda) = \mathbb{E}[Z_I^\top WZ_I] - 2(1-\lambda)\mathbb{E}[G^\top WZ_I] + (1-\lambda)^2\mathbb{E}[G^\top WG]$ with $\mathbb{E}[Z_I^\top WZ_I] = a_{II}$, $\mathbb{E}[G^\top WZ_I] = \mathbb{E}[Z_\Delta^\top WZ_I] = \operatorname{tr}(W\operatorname{Cov}(Z_I, Z_\Delta)) = \kappa_W$ (δ is deterministic and Z_I centred) and $\mathbb{E}[G^\top WG] = \delta^\top W\delta + \tau_W$; expanding in the other basis gives the first form of equation 19. (a) to (c) are elementary from the second form: the vertex of $a_{II} - 2(1-\lambda)\kappa_W + (1-\lambda)^2 s$, $s := \delta^\top W\delta + \tau_W$, is at $1 - \lambda^* = \kappa_W/s$ with value $a_{II} - \kappa_W^2/s$, and $\rho_W(0) - \rho_W(\lambda^*) = s(\lambda^*)^2$ because $\rho_W(\lambda) - \rho_W(\lambda^*) = s(\lambda - \lambda^*)^2$. □

G.5 S.4 The SURE-tuned estimator

Theorem S3 (Risk of $\hat{\theta}_{\mathrm{SURE}}$; dominance; oracle inequality)**.** *Under Assumption L-S, with $\lambda_{\max} = 1$ and $\kappa_W > 0$, the rule equals*

$$1 - \hat{\lambda} = w(T_W), \qquad w(t) := \min\Big\{1,\, \frac{\kappa_W}{\max(t, \tau_W)}\Big\}, \qquad \hat{\theta}_{\mathrm{SURE}} = \hat{\theta}_E + \hat{\lambda}\hat{\Delta} = \hat{\theta}_{I,\epsilon} - w(T_W)\hat{\Delta}, \tag{21}$$

a James-Stein-type weight with the constant $\kappa_W = \operatorname{tr}(WC)$ and the loss-adapted statistic T_W floored at τ_W; in the Hausman case $\kappa_W = \tau_W$ and $w = \min(1, \tau_W/T_W)$ is the positive-part James-Stein weight with $\tau_W = \operatorname{tr}(WV_\Delta)$. Let $\tilde{A} := B^{-1/2}AB^{1/2}$ ($\operatorname{tr}\tilde{A} = \kappa_W$), $\tilde{A}_s := \frac{1}{2}(\tilde{A} + \tilde{A}^\top)$, $\nu_W := \lambda_{\max}(\tilde{A}_s) = \lambda_{\max}(B^{-1/2}(AB)_sB^{-1/2}) = \sup_{u \ne 0} u^\top(AB)_s u/(u^\top Bu)$ with $(AB)_s := \frac{1}{2}(AB + BA^\top)$, and $\bar{\tau} := \max(\tau_W, \kappa_W)$, $w_0 := \min(1, \kappa_W/\tau_W)$. Then:

(a) Exact risk (Stein). *With $w = w(T_W)$ and $w' = \mathrm{d}w/\mathrm{d}t$ at T_W ($= -\kappa_W/T_W^2$ on $\{T_W > \bar\tau\}$, 0 elsewhere),*

$$\begin{aligned}\rho_W(E_{\mathrm{SURE}}) - \rho_W(E_{\mathrm{IEO}}) &= \mathbb{E}\Big[w^2 T_W - 2w\kappa_W - 4w'\,U^\top (AB)_s U\Big] \\ &= \mathbb{E}\big[\mathbf{1}\{T_W \le \bar\tau\}\, w(wT_W - 2\kappa_W)\big] \\ &\quad - \kappa_W\, \mathbb{E}\Big[\mathbf{1}\{T_W > \bar\tau\}\Big(\frac{\kappa_W}{T_W} - \frac{4U^\top (AB)_s U}{T_W^2}\Big)\Big].\end{aligned} \tag{22}$$

(b) Dominance over IEO. *If*

$$\begin{gathered}\kappa_W \ge 4\nu_W, \qquad \textit{i.e.} \qquad \operatorname{tr}\tilde A \ge 4\lambda_{\max}(\tilde A_s), \\ \textit{which in the Hausman case reads} \quad \operatorname{tr}(WV_\Delta) \ge 4\lambda_{\max}(WV_\Delta),\end{gathered} \tag{23}$$

then for every $\mu \in \mathbb{R}^r$

$$\begin{aligned}\rho_W(E_{\mathrm{SURE}}) - \rho_W(E_{\mathrm{IEO}}) \le -\kappa_W(\kappa_W - 4\nu_W)\,\mathbb{E}\Big[\frac{\mathbf{1}\{T_W > \bar\tau\}}{T_W}\Big] \\ - \kappa_W w_0 \Pr(T_W \le \bar\tau) < 0.\end{aligned} \tag{24}$$

At equality, $\kappa_W = 4\nu_W$, the first term of the bound vanishes and the strict inequality comes from the second term, the positive-part region $\{T_W \le \bar\tau\}$, which has positive probability for every μ. The constant is exactly 4: the SURE choice sets the James-Stein constant at κ_W, and the range of constants a in the weight $\min(1, a/\max(t, \tau_W))$ for which the pointwise argument applies is $0 < a \le 2(\kappa_W - 2\nu_W)$ (Theorem S6(b),(c); Hausman form $0 < a \le 2(\operatorname{tr} B - 2\lambda_{\max}B)$); $a = \kappa_W$ lies in it iff equation 23. In the regret geometry ($W = W_L$, Hausman case) the condition is that the effective dimension $\operatorname{tr}(W_L V_\Delta)/\lambda_{\max}(W_L V_\Delta)$ is at least 4.

(c) Oracle inequality (no condition on κ_W beyond $\kappa_W > 0$). *For every μ, with the oracle taken over the range $\lambda \in [0,1]$ (its minimiser is λ^* whenever $\kappa_W \le \delta^\top W \delta + \tau_W$, in particular in the Hausman case, and 0 otherwise),*

$$\begin{aligned}\rho_W(E_{\mathrm{SURE}}) &\le \min_{\lambda\in[0,1]} \rho_W(\lambda) + 4\kappa_W\, \mathbb{E}\Big[\mathbf{1}\{T_W > \bar\tau\}\frac{U^\top (AB)_s U}{T_W^2}\Big] \\ &\le \min_{\lambda\in[0,1]} \rho_W(\lambda) + 4\kappa_W\nu_W\, \mathbb{E}\Big[\frac{\mathbf{1}\{T_W > \bar\tau\}}{T_W}\Big] \le \min_{\lambda\in[0,1]} \rho_W(\lambda) + \frac{4\kappa_W\nu_W}{\bar\tau},\end{aligned} \tag{25}$$

and the remainder is $O(|\mu|^{-2})$ as $|\mu| \to \infty$ (so $\rho_W(E_{\mathrm{SURE}}) - \min_{\lambda\in[0,1]} \rho_W(\lambda) \to 0$). In the Hausman case the last bound is $4\lambda_{\max}(B)$, against an ETO to IEO gap of $\operatorname{tr} B + \mu^\top B\mu$ (equation equation 17): the remainder is at most $4/\mathrm{effdim}$ of the H_0 gap.

(d) Convergence of $\hat\lambda$. *In the limit experiment $\hat\lambda \to_d \hat\lambda_\infty := 1 - w(T_W)$, a nondegenerate random variable; the plug-in $\hat q = (T_W - \tau_W)_+$ satisfies $\mathbb{E}[T_W - \tau_W] = \mu^\top B\mu = \delta^\top W\delta$ (unbiased before the positive part) and $\hat q/(\mu^\top B\mu) \to 1$ in probability as $|\mu| \to \infty$, so $\hat\lambda_\infty - \lambda^*(\mu) = O_p(|\mu|^{-1})$ along any sequence with $\mu^\top B\mu \to \infty$; under a fixed alternative $\Delta^* \ne 0$, $\hat\lambda \to 1$ and $\lambda^* \to 1$ (both estimators coincide with IEO to first order). $\hat\lambda$ does not converge to λ^* at a fixed local μ, which is the usual situation for Stein rules and the reason (c) is stated as an oracle inequality.*

(e) At H_0. *At $\mu = 0$, $\rho_W(E_{\mathrm{SURE}}) = \rho_W(E_{\mathrm{ETO}})|_{\mu=0} + \kappa_{0,W}$ with $\kappa_{0,W} := \mathbb{E}[2(1-w)U^\top A_s U - (1-w^2)U^\top BU]_{\mu=0}$, which in the Hausman case equals $\mathbb{E}[(1-w)^2 T_W]_{\mu=0} = \mathbb{E}[(T_W - \tau_W)_+^2/T_W]_{\mu=0} \ge 0$, with $T_W = \xi^\top B\xi$, $\xi \sim \mathcal{N}(0, I_r)$: the SURE estimator is not ETO at the null but ETO plus a bounded residual.*

(f) The $\lambda_{\max} = 2$ variant. *With the clip at 2 instead of 1, $1 - \hat\lambda = \kappa_W/\max(T_W, \tau_W)$ clipped to $[-1, 1]$; for $\kappa_W > 0$ this is equation 21 unchanged, so extrapolation beyond IEO occurs only when $\kappa_W = \operatorname{tr}(WV_{II}) - \operatorname{tr}(WV_{EI}) < 0$, where (b) offers no guarantee. The variant is exploratory.*

Proof. Form of the rule. With $\hat{q} = (T_W - \tau_W)_+$, $1 - \hat{\lambda}_{\mathrm{raw}} = 1 - \frac{\hat{q}+a_E-a_{EI}}{\hat{q}+\tau_W} = \frac{\tau_W-a_E+a_{EI}}{\hat{q}+\tau_W} = \frac{\kappa_W}{\max(T_W,\tau_W)}$ by equation 18 ($\tau_W - a_E + a_{EI} = a_{II} - a_{EI} = \kappa_W$ and $\hat{q} + \tau_W = \max(T_W, \tau_W)$); the clip of $\hat{\lambda}$ to $[0,1]$ is the clip of $1-\hat{\lambda}$ to $[0,1]$, and for $\kappa_W > 0$ the lower clip never binds, which gives equation 21. $E_{\mathrm{SURE}} = Z_I - w(T_W)G$ is equation 13 with the weight $w(T_W)$.

(a) The risk identity equation 16 holds for any weight $w(U)$: $\rho_W(Z_I - wG) - \rho_W(Z_I) = \mathbb{E}[w^2 U^\top BU] - 2\mathbb{E}[w\, U^\top A(U-\mu)]$. Apply Lemma 5 to $g(U) := w(U^\top BU)\, A^\top U$, which is Lipschitz: w is Lipschitz in t on $[0,\infty)$ with constant $\kappa_W/\bar{\tau}^2$, $|\nabla_U(U^\top BU)| = 2|BU|$, and on $\{T_W > \bar{\tau}\}$, $|\nabla w|\,|A^\top U| \le 2\kappa_W |BU||A||U|/T_W^2 \le 2\kappa_W |A| \lambda_{\max}(B)^{1/2}\lambda_{\min}(B)^{-1/2}/T_W \le c/\bar{\tau}$ (using $|BU| \le \lambda_{\max}(B)^{1/2} T_W^{1/2}$ and $|U| \le \lambda_{\min}(B)^{-1/2} T_W^{1/2}$), while w is constant on $\{T_W \le \bar{\tau}\}$; the divergence is $\nabla \cdot g = w \operatorname{tr} A + \sum_l (A^\top U)_l\, w'\, 2(BU)_l = w\kappa_W + 2w'\, U^\top ABU = w\kappa_W + 2w'\, U^\top (AB)_s U$. Hence $\mathbb{E}[w\, U^\top A(U-\mu)] = \kappa_W \mathbb{E}[w] + 2\mathbb{E}[w'\, U^\top (AB)_s U]$ and the first equality of equation 22 follows. The second substitutes w and w' on the two events: on $\{T_W > \bar{\tau}\}$, $w = \kappa_W/T_W$, $w' = -\kappa_W/T_W^2$, so $w^2 T_W - 2w\kappa_W - 4w' U^\top (AB)_s U = -\kappa_W^2/T_W + 4\kappa_W U^\top (AB)_s U/T_W^2$; on $\{T_W \le \bar{\tau}\}$, $w' = 0$ (the boundary has probability zero since T_W has a density).

(b) On $\{T_W > \bar{\tau}\}$: $U^\top (AB)_s U \le \nu_W\, U^\top BU = \nu_W T_W$ by the definition of ν_W (equivalently $\nu_W = \lambda_{\max}(B^{-1/2}(AB)_s B^{-1/2}) = \lambda_{\max}(\tilde{A}_s)$ since $B^{-1/2}(AB)_s B^{-1/2} = \frac{1}{2}(B^{-1/2}AB^{1/2} + B^{1/2}A^\top B^{-1/2}) = \tilde{A}_s$), so the bracket is $\le -(\kappa_W/T_W)(\kappa_W - 4\nu_W) \le 0$. On $\{T_W \le \tau_W\}$: $w = w_0$ and $w_0(w_0 T_W - 2\kappa_W) \le w_0(w_0\tau_W - 2\kappa_W)$, which equals $-\kappa_W^2/\tau_W = -\kappa_W w_0$ if $\kappa_W \le \tau_W$ and is $\le \tau_W - 2\kappa_W \le -\kappa_W = -\kappa_W w_0$ if $\kappa_W > \tau_W$. On $\{\tau_W < T_W \le \kappa_W\}$ (nonempty only if $\kappa_W > \tau_W$): $w = 1$ and $T_W - 2\kappa_W \le -\kappa_W = -\kappa_W w_0$. Taking expectations gives equation 24; strictness, also when $\kappa_W = 4\nu_W$, because $\kappa_W w_0 > 0$ and $\Pr(T_W \le \bar{\tau}) > 0$ for every μ (T_W has full support on $(0,\infty)$ when $B \succ 0$). For a general constant a in place of κ_W the bracket on $\{T_W > \bar{\tau}\}$ is $\le (a/T_W)(a - 2\kappa_W + 4\nu_W)$, nonpositive iff $a \le 2(\kappa_W - 2\nu_W)$; in the Hausman case $A = B$, $\tilde{A} = B$, $\nu_W = \lambda_{\max}(B)$ and $\kappa_W = \operatorname{tr} B$, and $\operatorname{tr} B = \operatorname{tr}(WV_\Delta)$, $\lambda_{\max}(B) = \lambda_{\max}(WV_\Delta)$ (B is similar to WV_Δ).

(c) Let $s := \mu^\top B\mu + \tau_W = \mathbb{E}[T_W]$ and $w^* := 1 - \lambda^* = \kappa_W/s$ (Theorem S2(a)). For the constant weight w^*, $\rho_W(\lambda^*) - \rho_W(\mathrm{IEO}) = \mathbb{E}[(w^*)^2 T_W - 2w^*\kappa_W] = (w^*)^2 s - 2w^*\kappa_W$ (no Stein term for a constant weight). Subtracting from equation 22,

$$\rho_W(E_{\mathrm{SURE}}) - \rho_W(\lambda^*) = \mathbb{E}\big[(w - w^*)\big((w + w^*)T_W - 2\kappa_W\big)\big] - 4\mathbb{E}[w'\, U^\top (AB)_s U].$$

Here $w^* := \min(1, \kappa_W/s)$ is the minimiser over $w = 1 - \lambda \in [0,1]$ of the convex quadratic $w \mapsto w^2 s - 2w\kappa_W$ (unconstrained minimiser κ_W/s), and the identity above holds for any constant w^* because $\mathbb{E}[T_W] = s$. The first expectation is nonpositive pointwise. On $\{T_W > \bar{\tau}\}$, $w = \kappa_W/T_W \le 1$: if $w^* = \kappa_W/s$ the product equals $\kappa_W^2(1/T_W - 1/s)(T_W/s - 1) = -\kappa_W^2(T_W - s)^2/(s^2 T_W) \le 0$; if $w^* = 1$ (i.e. $\kappa_W \ge s$) then $w - w^* \le 0$ and $(w + w^*)T_W - 2\kappa_W = T_W - \kappa_W > 0$. On $\{T_W \le \tau_W\}$, $w = w_0 \ge w^*$ (because $\tau_W \le s$ and $t \mapsto \min(1, \kappa_W/t)$ is nonincreasing) and $(w_0 + w^*)T_W - 2\kappa_W \le 2w_0\tau_W - 2\kappa_W = 2(\min(\tau_W, \kappa_W) - \kappa_W) \le 0$. On $\{\tau_W < T_W \le \kappa_W\}$, $w = 1 \ge w^*$ and $(1 + w^*)T_W - 2\kappa_W \le (w^* - 1)\kappa_W \le 0$. The Stein term is $-4\mathbb{E}[w' U^\top (AB)_s U] = 4\kappa_W \mathbb{E}[\mathbf{1}\{T_W > \bar{\tau}\} U^\top (AB)_s U/T_W^2] \le 4\kappa_W \nu_W \mathbb{E}[\mathbf{1}\{T_W > \bar{\tau}\}/T_W] \le 4\kappa_W\nu_W/\bar{\tau}$, which is equation 25. The rate: $\mathbb{E}[1/T_W\, \mathbf{1}\{T_W > \bar{\tau}\}] \le \mathbb{E}[1/\max(T_W, \bar{\tau})]$ and $T_W = (\mu + \xi)^\top B(\mu + \xi)$, $\xi \sim \mathcal{N}(0, I_r)$, satisfies $\Pr(T_W \le \mu^\top B\mu/2) \to 0$ exponentially in $|\mu|_B^2$ while $1/T_W \le 2/\mu^\top B\mu$ on the complement, so the expectation is $O(|\mu|^{-2})$. In the Hausman case $\kappa_W = \tau_W = \bar{\tau} = \operatorname{tr} B$ and $\nu_W = \lambda_{\max}(B)$, so $4\kappa_W\nu_W/\bar{\tau} = 4\lambda_{\max}(B)$.

(d) $\hat{\lambda} = 1 - w(T_W)$ is a continuous function of (T_W, the traces), so the plug-in version converges in distribution to $\hat{\lambda}_\infty$ by Proposition 1-S and the continuous mapping theorem. Unbiasedness is $\mathbb{E}[T_W] = \mu^\top B\mu + \tau_W$. With $T_W = \mu^\top B\mu + 2\mu^\top B\xi + \xi^\top B\xi$, $\hat{q}/(\mu^\top B\mu) = 1 + O_p(|\mu|^{-1})$, and $\hat{\lambda}_\infty = 1 - \kappa_W/(\hat{q} + \tau_W)$ versus $\lambda^* = 1 - \kappa_W/(\mu^\top B\mu + \tau_W)$ differ by $\kappa_W|\hat{q} - \mu^\top B\mu|/((\hat{q} + \tau_W)(\mu^\top B\mu + \tau_W))$, which is $O_p(|\mu|^{-1})$ (in fact $O_p(|\mu|^{-3})$, since the numerator is $O_p(|\mu|)$ and the denominator of order $|\mu|^4$). Under a fixed alternative $T_W/n \to \Delta^{*\top} W \Delta^* > 0$, so $w(T_W) \to 0$.

(e) $E_{\mathrm{SURE}} - E_{\mathrm{ETO}} = (1-w)G$, $\rho_W(E_{\mathrm{SURE}}) - \rho_W(E_{\mathrm{ETO}}) = 2\mathbb{E}[(1-w)U^\top A(U-\mu)] - \mathbb{E}[(1 - w^2)U^\top BU]$, continuous in μ, and at $\mu = 0$ the Hausman substitution $A = B$ (with $\kappa_W = \tau_W$, so $w = \min(1, \tau_W/T_W)$) gives $\mathbb{E}[(1-w)^2 U^\top BU] = \mathbb{E}[(1-w)^2 T_W]$; the integrand vanishes on $\{T_W \le \tau_W\}$ and equals $(T_W - \tau_W)^2/T_W$ on $\{T_W > \tau_W\}$.

(f) Immediate from the form of the rule: the lower clip of $1 - \hat{\lambda}$ at -1 replaces the clip at 0, and $\kappa_W / \max(T_W, \tau_W) \geq 0$ iff $\kappa_W \geq 0$. □

G.6 S.5 The Wald test with the F reference

Proposition S4 (Size and consistency of the Wald decision-adequacy test)**.** *Let $T_\epsilon := n\hat{\Delta}^\top (\hat{V}_\Delta^\epsilon)^+ \hat{\Delta}$ with the rank rule of Theorem 1-S(b), $c_n := (n-1)/(n-\hat{r}-2)$, and reject H_0 when $T_\epsilon / c_n > \chi^2_{\hat{r},1-\alpha}$. Under the assumptions of Theorem 1-S and Proposition 1-S:*

- *(a) under H_0, $T_\epsilon \to_d \chi^2_r$ and, since $c_n \to 1$, the rejection probability tends to α;*
- *(b) under Assumption L-S, $T_\epsilon \to_d \chi^2_r(|\mu|^2)$ (noncentral) and T_ϵ / c_n has the same limit;*
- *(c) under a fixed alternative $\Pi\theta^*_\epsilon \neq \Pi\theta^*_E$, $T_\epsilon / n \to_p \Delta^{*\top} (V^\epsilon_\Delta)^+ \Delta^* > 0$ and the test is consistent;*
- *(d)* (the factor, motivation only) *if $\sqrt{n}\hat{\Delta}$ were exactly Gaussian with covariance V_Δ and $\hat{V}_\Delta$ an independent Wishart estimate on $n-1$ degrees of freedom, T_ϵ would be Hotelling's T^2 with $\mathbb{E}[T_\epsilon] = (n-1)r/(n-r-2) = c_n r$; c_n matches the mean of T_ϵ to that of χ^2_r. The sandwich is neither independent of $\hat{\Delta}$ nor Wishart, so no finite-sample size statement is claimed; the adjustment is fixed in advance and its effect is reported (Appendix A).*

Proof. (a) and (c) are Theorem 1-S(b),(c) with the consistent $\hat{V}^\epsilon_\Delta$ of Proposition 1-S. (b) Under Assumption L-S, $\sqrt{n}\hat{\Delta} \to G = \delta + Z_\Delta$ and $\hat{V}^\epsilon_\Delta \to V_\Delta$, so $T_\epsilon \to G^\top V_\Delta^+ G = |U|^2 \sim \chi^2_r(|\mu|^2)$; $c_n \to 1$ by Slutsky. (d) is the classical Hotelling computation (Anderson, 2003, Section 5.2.2): $(n-r)T^2/((n-1)r) \sim F_{r,n-r}$ with mean $(n-r)/(n-r-2)$. □

G.7 S.6 Regret transfer

Decision loss. Let $\mathcal{L}(\theta) := \mathbb{E}[c^\top x^*(\Theta\phi(z))] = \mathbb{E}[m(z)^\top x^*(\Theta\phi(z))]$, $m(z) := \mathbb{E}[c \mid z]$, be the true decision loss, and θ^*_L its minimiser. Note that $\mathcal{L}$ depends on θ only through the events $\{\Theta\phi(z) \in N_v\}$: for fixed z it is piecewise constant in θ, and its smoothness comes from the law of $\phi(z)$ (the predicted cost crossing a normal-cone boundary as θ moves), *not* from the conditional law of c. This is why the regret expansion needs its own assumption.

Assumption D (second-order regret expansion along the local sequence)**.** Under Assumption L-S, $\mathcal{L}_n$ (the decision loss under P_n) has a minimiser $\theta^*_{L,n} = \theta_n + \delta_L/\sqrt{n} + o(n^{-1/2})$ for a fixed $\delta_L \in \mathrm{range}(\Pi)$, and there is a symmetric positive semidefinite W_L (the Hessian of $\mathcal{L}$ at θ^*, annihilating $\mathbb{R}^{q_0} \otimes L$) such that, uniformly over $|t| \leq K$ for every K,

$$n\big[\mathcal{L}_n(\theta_n + t/\sqrt{n}) - \mathcal{L}_n(\theta^*_{L,n})\big] = \tfrac{1}{2}(t-\delta_L)^\top W_L (t-\delta_L) + o(1).$$

A sufficient condition is that $\phi(z)$ has a density on its non-intercept coordinates, m is continuous, and $\mathcal{L}$ is C^2 near θ^* with $\mathcal{L}_n \to \mathcal{L}$ in C^2 (Lan et al., 2025, obtain such expansions in the unconstrained non-contextual case); it is tagged *assumed* here.

Assumption R (local regret-consistency of the smoothed SPO+)**.** $\delta_L = \delta$, i.e. the pseudo-true direction of the smoothed SPO+ coincides with the true-decision-loss direction along the local sequence. Under H_0 with a correctly specified conditional mean and symmetric noise, $\theta^*_E = \theta^*_\epsilon = \theta^*_L = \theta_0$ (the Bayes decision is $x^*(m(z)) = x^*(\Theta_0\phi)$), so $\delta = \delta_L = 0$ and R holds trivially; in general R is an assumption; its known limitation is Corollary S5(iii).

Regret of an estimator with limit error E. Under Assumptions L-S and D, for any estimator with centred limit error E (relative to $\theta^*_{\epsilon,n}$, as in equation 13), the excess regret satisfies

$$n\big[\mathcal{L}_n(\hat{\theta}) - \mathcal{L}_n(\theta^*_{L,n})\big] \xrightarrow{d} \tfrac{1}{2}\big(E-\eta\big)^\top W_L \big(E-\eta\big), \qquad \eta := \delta_L - \delta, \tag{26}$$

and we define the asymptotic excess regret $\mathcal{R}(\hat{\theta}) := \frac{1}{2}\mathbb{E}[(E-\eta)^\top W_L (E-\eta)] = \frac{1}{2}\rho_{W_L}(E) - \mathbb{E}[E^\top W_L \eta] + \frac{1}{2}\eta^\top W_L \eta$. Indeed, $\sqrt{n}(\hat{\theta} - \theta^*_{L,n}) = \sqrt{n}(\hat{\theta} - \theta^*_{\epsilon,n}) + (\delta - \delta_L) + o(1) \xrightarrow{d} E - \eta$ by (L2), (L3) and Assumption D, and the uniform expansion of Assumption D with the continuous

mapping theorem gives the limit of the excess regret (the estimators are $\sqrt{n}$-tight, so the expansion may be applied on $|t| \le K$ with K large). Expanding the quadratic form gives the three-term expression for $\mathcal{R}$.

Corollary S5 (Regret; limitation and the exploratory variant)**.** *Under Assumptions L-S and D (with δ_L, W_L of Assumption D and $\eta := \delta_L - \delta$ as in equation 26), for the estimator $\hat{\theta}_{\mathrm{SURE}}$ with weight matrix W in the rule and W_L in the regret,*

(i) under Assumption R ($\eta = 0$), $\mathcal{R}(\hat{\theta}) = \frac{1}{2}\rho_{W_L}(E)$ for every estimator, so Theorem S2, Theorem S3(b) and (c) with $W = W_L$ transfer without change to asymptotic excess regret: with $W = W_L$ in the rule, $\hat{\theta}_{\mathrm{SURE}}$ has strictly smaller regret than IEO for every δ when equation 23 holds with $W = W_L$, and is within $\frac{1}{2} \cdot 4\kappa_{W_L}\nu_{W_L}/\bar{\tau}$ of the oracle path point in regret for every δ;

(ii) in general, with $w = w(T_W)$ from equation 21,

$$\mathcal{R}(\hat{\theta}_{\mathrm{SURE}}) - \mathcal{R}(\hat{\theta}_{\mathrm{IEO}}) = \tfrac{1}{2}\big[\rho_{W_L}(E_{\mathrm{SURE}}) - \rho_{W_L}(E_{\mathrm{IEO}})\big] + \eta^\top W_L\, \mathbb{E}[w\, G],$$
$$\big|\eta^\top W_L \mathbb{E}[wG]\big| \le |\eta|_{W_L}\big(\mathbb{E}[w^2 G^\top W_L G]\big)^{1/2},$$

and when $W = W_L$ in the rule, $w^2 G^\top W_L G = w^2 T_W \le \kappa_W w_0$ pointwise, so the cross term is at most $|\eta|_{W_L}(\kappa_W w_0)^{1/2}$ ($= |\eta|_{W_L}\,\mathrm{tr}(W_L V_\Delta)^{1/2}$ in the Hausman case);

(iii) limitation*: on the polynomial family of our experiments the two directions are numerically found proportional with a norm ratio above one, $\delta_L = \alpha\delta$ with $\alpha > 1$, so $\eta = (\alpha - 1)\delta \neq 0$ under fixed misspecification and the cross term of (ii) is not zero; every regret statement under Assumption R is subject to it. The oracle path point in regret is then $\lambda_L^* = \lambda^*$ computed with δ_L in place of δ, which exceeds 1 when $\alpha > 1$ and $\delta^\top W_L \delta$ dominates the traces; the frozen path with $\lambda_{\max} = 1$ cannot reach it. The $\lambda_{\max} = 2$ variant of Theorem S3(f) is the recorded exploratory remedy, but under the frozen rule it extrapolates only when $\kappa_W < 0$ (Theorem S3(f)); an extrapolating rule would need an estimate of α from the smoothed decision loss $\mathcal{L}_{n,\epsilon}$ along $\hat{\Delta}$, which is not implemented and has no theorem.*

The plug-in W of Section 3.2 is the Hessian of the smoothed decision loss $\mathcal{L}_\epsilon(\theta) = \mathbb{E}[c^\top x_\epsilon(\Theta\phi)]$ at $\hat{\theta}_E$, whose consistency for $\nabla^2\mathcal{L}_\epsilon$ follows as in Proposition 1-S (third derivatives of z_ϵ bounded by cR_X^3/ϵ^2) and whose convergence to $W_L = \nabla^2\mathcal{L}$ as $\epsilon \to 0$ is part of Assumption D; the PSD projection is a continuous map at fixed rank and does not affect the limit when $W_L \succeq 0$.

Proof. (i) is immediate from equation 26. (ii) With $E_{\mathrm{SURE}} = Z_I - wG$ the difference of the middle terms of $\mathcal{R}$ is $-\mathbb{E}[(E_{\mathrm{SURE}} - E_{\mathrm{IEO}})^\top W_L \eta] = \mathbb{E}[wG^\top W_L \eta]$ by equation 13; the bound is Cauchy-Schwarz in the W_L inner product ($|\mathbb{E}[wG]|_{W_L} \le (\mathbb{E}[w^2 G^\top W_L G])^{1/2}$ by Jensen, with $|\eta|_{W_L}^2 := \eta^\top W_L \eta$). The pointwise bound on $w^2 T_W$: on $\{T_W > \bar{\tau}\}$, $w^2 T_W = \kappa_W^2/T_W \le \kappa_W^2/\bar{\tau} = \kappa_W w_0$; on $\{T_W \le \tau_W\}$, $w_0^2 T_W \le w_0^2 \tau_W \le \kappa_W w_0$ (both cases of w_0); on $\{\tau_W < T_W \le \kappa_W\}$, $T_W \le \kappa_W = \kappa_W w_0$. (iii) restates the numerical finding and Theorem S3(f). □

G.8 S.7 The largest safe constant and SURE-safe

Theorem S6 gives the largest shrinkage constant that keeps the dominance conclusion of Theorem S3(b), and the estimator that uses it, the second of the pair (SURE, SURE-safe).

Theorem S6 (The largest safe constant; SURE-safe)**.** *Under Assumption L-S with $\lambda_{\max} = 1$ and $\kappa_W > 0$, and with A, B, $\tilde{A}_s$, ν_W, τ_W, T_W, U as in Theorem S3, consider the family of weights*

$$w_a(t) := \min\Big\{1,\ \frac{a}{\max(t, \tau_W)}\Big\}, \qquad a > 0, \qquad \hat{\theta}_a := \hat{\theta}_{I,\epsilon} - w_a(T_W)\,\hat{\Delta}, \tag{27}$$

so that $a = \kappa_W$ is the SURE rule equation 21. Write $\bar{\tau}_a := \max(\tau_W, a)$, $w_{0,a} := \min(1, a/\tau_W)$, and define

$$\tau_{\max} := 2(\kappa_W - 2\nu_W), \qquad \textit{in the Hausman case} \qquad \tau_{\max} = 2\big(\mathrm{tr}\, B - 2\lambda_{\max}(B)\big), \tag{28}$$

which is positive iff $\kappa_W > 2\nu_W$, i.e. in the Hausman case iff the effective dimension $\operatorname{tr} B/\lambda_{\max}(B)$ exceeds 2. SURE-safe *is $\hat\theta_{\mathrm{safe}} := \hat\theta_{a_{\mathrm{safe}}}$ with $a_{\mathrm{safe}} := \min(\kappa_W, \tau_{\max})$ when $\tau_{\max} > 0$, and $\hat\theta_{\mathrm{safe}} := \hat\theta_{I,\epsilon}$ (weight 0) when $\tau_{\max} \le 0$; the plug-in version replaces $(\kappa_W, \nu_W, \tau_W)$ by the estimates of Proposition 1-S computed on the retained eigenvectors of $\hat V_\Delta^\epsilon$. Then:*

(a) Exact risk. *For every $a > 0$ and every μ,*

$$\rho_W(E_a) - \rho_W(E_{\mathrm{IEO}}) = \mathbb{E}\big[\mathbf{1}\{T_W \le \bar\tau_a\}\, w_a(w_a T_W - 2\kappa_W)\big] - \mathbb{E}\Big[\mathbf{1}\{T_W > \bar\tau_a\}\, \frac{a}{T_W}\Big(2\kappa_W - a - \frac{4\,U^\top (AB)_s U}{T_W}\Big)\Big], \tag{29}$$

with $E_a := Z_I - w_a(T_W)G$ the limit error of $\hat\theta_a$.

(b) $\tau_{\max}$ is the largest constant for which the pointwise Stein argument applies. *On $\{T_W > \bar\tau_a\}$ the bracket in equation 29 satisfies $2\kappa_W - a - 4U^\top(AB)_s U/T_W \ge \tau_{\max} - a$, with equality exactly when $B^{1/2}U$ lies in the top eigenspace of $\tilde A_s$. Hence the bracket is nonnegative for every U iff $a \le \tau_{\max}$; for $a > \tau_{\max}$ it is negative on a cone of U of positive probability under every μ, so the argument of Theorem S3(b) does not extend beyond $\tau_{\max}$.*

(c) Dominance. *If $\tau_{\max} > 0$ and $0 < a \le \min(\kappa_W, \tau_{\max})$, then for every $\mu \in \mathbb{R}^r$*

$$\rho_W(E_a) - \rho_W(E_{\mathrm{IEO}}) \le -\,a(\tau_{\max} - a)\,\mathbb{E}\Big[\frac{\mathbf{1}\{T_W > \bar\tau_a\}}{T_W}\Big] - \kappa_W\, w_{0,a}\, \Pr(T_W \le \bar\tau_a) < 0. \tag{30}$$

In particular SURE-safe dominates IEO for every local alternative whenever $\tau_{\max} > 0$, i.e. in the Hausman case whenever the effective dimension exceeds 2, with no further condition; when $\tau_{\max} \le 0$ it coincides with IEO.

(d) Relation to SURE; the price. *$a \mapsto w_a(t)$ is nondecreasing for every t, strictly increasing on $\{a < \max(t, \tau_W)\}$. If $\kappa_W \le \tau_{\max}$ (equivalently $\kappa_W \ge 4\nu_W$, the condition of Theorem S3(b)) then $a_{\mathrm{safe}} = \kappa_W$ and SURE-safe is SURE. If $\kappa_W > \tau_{\max} > 0$ then $w_{a_{\mathrm{safe}}}(T_W) \le w(T_W)$ pointwise, with strict inequality on $\{T_W > \tau_{\max}\}$: SURE-safe shrinks less towards ETO than SURE. At the null in the Hausman case ($\mu = 0$, $A = B$), $\rho_W(E_a) - \rho_W(E_{\mathrm{ETO}})|_{\mu=0} = \mathbb{E}[(1 - w_a(T_W))^2 T_W]_{\mu=0}$ is nonincreasing in a, so the residual of SURE-safe over ETO under decision adequacy is at least that of SURE: the guarantee of (c) is bought with part of the ETO gain whenever $\kappa_W > \tau_{\max}$.*

Proof. (a) The pre-Stein identity and Lemma 5 apply to $g(U) = w_a(U^\top BU)A^\top U$ exactly as in the proof of Theorem S3(a) (Lipschitz with a in place of κ_W), giving $\rho_W(E_a) - \rho_W(E_{\mathrm{IEO}}) = \mathbb{E}[w_a^2 T_W - 2w_a\kappa_W - 4w_a' U^\top (AB)_s U]$ with $w_a' = -a/T_W^2$ on $\{T_W > \bar\tau_a\}$ and 0 elsewhere. On $\{T_W > \bar\tau_a\}$ the integrand is $a^2/T_W - 2a\kappa_W/T_W + 4aU^\top(AB)_sU/T_W^2 = -(a/T_W)(2\kappa_W - a - 4U^\top(AB)_sU/T_W)$, which is equation 29.

(b) $U^\top(AB)_sU \le \nu_W U^\top BU = \nu_W T_W$ with equality iff $B^{1/2}U$ is a top eigenvector of $\tilde A_s = B^{-1/2}(AB)_sB^{-1/2}$ (Rayleigh quotient), so $2\kappa_W - a - 4U^\top(AB)_sU/T_W \ge 2\kappa_W - a - 4\nu_W = \tau_{\max} - a$ with equality on that eigenspace. The eigenspace has positive probability under $U \sim \mathcal{N}(\mu, I_r)$ for any neighbourhood cone, which gives the last claim.

(c) On $\{T_W > \bar\tau_a\}$, (b) gives the first term of equation 30. On $\{T_W \le \tau_W\}$, $w_a = w_{0,a}$ and $w_{0,a}(w_{0,a}T_W - 2\kappa_W) \le w_{0,a}(w_{0,a}\tau_W - 2\kappa_W) \le w_{0,a}(a - 2\kappa_W) \le -\kappa_W w_{0,a}$, using $w_{0,a}\tau_W \le a$ and $a \le \kappa_W$. On $\{\tau_W < T_W \le a\}$ (nonempty only if $a > \tau_W$), $w_a = 1$ and $T_W - 2\kappa_W \le a - 2\kappa_W \le -\kappa_W = -\kappa_W w_{0,a}$. Taking expectations gives equation 30; strictness because $\Pr(T_W \le \bar\tau_a) > 0$ for every μ (T_W has full support on $(0,\infty)$ when $B \succ 0$). For $a = a_{\mathrm{safe}} = \min(\kappa_W, \tau_{\max})$ the hypotheses hold whenever $\tau_{\max} > 0$.

(d) Monotonicity of w_a in a is immediate from equation 27; $\kappa_W \le \tau_{\max}$ iff $\kappa_W \le 2\kappa_W - 4\nu_W$ iff $\kappa_W \ge 4\nu_W$. For the null residual, $E_a - E_{\mathrm{ETO}} = (1 - w_a)G$ and the computation of Theorem S3(e) holds for any weight; in the Hausman case at $\mu = 0$ it gives $\mathbb{E}[(1 - w_a)^2 T_W]$, nonincreasing in a because $1 - w_a$ is. □

Plug-in and use. $\hat{a}_{\text{safe}} = \min(\hat{\kappa}_W, \hat{\tau}_{\max})$ with $\hat{\tau}_{\max} = 2(\hat{\kappa}_W - 2\hat{\nu}_W)$ from the estimated matrices (general-C form; the Hausman form $2(\operatorname{tr}\hat{B} - 2\lambda_{\max}\hat{B})$ is a reported diagnostic), $\hat{\kappa}_W = \operatorname{tr}(\hat{W}\hat{C})$; by Proposition 1-S and the continuous mapping theorem the plug-in weight converges to $w_{a_{\text{safe}}}(T_W)$ when $B \succ 0$. The two estimators are reported side by side (SURE primary, SURE-safe secondary) in every table; neither is chosen after the results. The numerical check in the Hausman limit experiment at effective dimensions 3, 5 and 10 (`tests/test_sure_safe.py`) is: at effective dimension 3 the worst excess of SURE over IEO on the δ grid is $+2.1$ percent and that of SURE-safe is -0.04 percent; at 5 and 10 the two estimators coincide.

G.9 S.8 The regret-weighted decision-adequacy test

The test of the main text moves decision adequacy from the Wald geometry of Proposition S4 to the regret geometry of Theorems S2 and S3. Let $W \succeq 0$ be the regret weight of Section 3.2 (block diagonal in the identified coordinates as in Assumption L-S) and $\hat{W}$ its plug-in at $\hat{\theta}_E$, consistent for W by the plug-in statement of Corollary S5. The statistic is

$$T_W := n\,\hat{\Delta}^\top \hat{W} \hat{\Delta}, \tag{31}$$

the quadratic form of the rule ($\hat{q} = (T_W - \operatorname{tr}(\hat{W}\hat{V}_\Delta))_+$). It uses neither the sandwich nor a pseudo-inverse nor the rank rule. Its null hypothesis is decision adequacy on the decision-relevant coordinates, $\mathrm{H}_{0,W} : W^{1/2}\Pi(\theta^*_\epsilon - \theta^*_E) = 0$, which is implied by H_0; the size statement below is proved under H_0 (Remark S8 records what is and is not claimed in between).

Proposition S7 (The regret-weighted test)**.** *Under the assumptions of Theorem 1-S and Proposition 1-S, with $\hat{W} \to_p W$, let $\omega_1 \geq \cdots \geq \omega_{q_M} \geq 0$ be the eigenvalues of WV_Δ (those of $B = V_\Delta^{1/2} W V_\Delta^{1/2}$; the nonzero ones coincide), and write $\tau_W = \operatorname{tr}(WV_\Delta) = \sum_j \omega_j$.*

(a) Null law. *Under* H_0, $T_W \to_d \sum_j \omega_j Z_j^2$ *with* $Z_1, Z_2, \ldots$ *independent standard normal; the limit has mean* τ_W *and variance* $2\operatorname{tr}((WV_\Delta)^2) = 2\sum_j \omega_j^2$.

(b) Consistency. *Under a fixed alternative* $\Delta^* := \Pi(\theta^*_\epsilon - \theta^*_E)$ *with* $W^{1/2}\Delta^* \neq 0$, $T_W/n \to_p \Delta^{*\top} W \Delta^* > 0$, *so the test with any critical value that is* $O_p(1)$ *(the references of (c) and (d)) rejects with probability tending to one: the test is consistent against every decision-relevant alternative.*

(c) Moment-matched reference (Satterthwaite, 1946; Box, 1954). *Define* $\hat{\nu}$ *as the ratio* $\operatorname{tr}(\hat{W}\hat{V}_\Delta)^2 / \operatorname{tr}((\hat{W}\hat{V}_\Delta)^2)$ *and reject when* $T_W > (\operatorname{tr}(\hat{W}\hat{V}_\Delta)/\hat{\nu})\,\chi^2_{\hat{\nu},1-\alpha}$. *The scaled chi-square matches the mean and the variance of the limit in (a);* $\hat{\nu} \to_p \nu := \tau_W^2 / \sum_j \omega_j^2$, *and* ν *satisfies* $1 \leq \tau_W/\omega_1 \leq \nu \leq \operatorname{rank}(WV_\Delta)$, *with* $\nu = \tau_W/\omega_1$ *(the effective dimension of Theorem S3, Hausman form) iff the nonzero* ω_j *are equal, and then the reference is the exact limit* $\omega_1\chi^2_\nu$. *In general the reference is an approximation of the limit law and the rejection probability under* H_0 *tends to a value near, not equal to,* α; *no finite-sample claim is made.*

(d) Null-calibrated reference. *Let* $e_i := c_i - \hat{\Theta}_E\phi_i$, s_i *independent Rademacher signs, and* $c_i^* := \hat{\Theta}_E\phi_i + s_i e_i$ *(the pseudo-sample); refit* $\hat{\theta}^*_E$ *and* $\hat{\theta}^*_{I,\epsilon}$ *on* (ϕ_i, c_i^*) *with* ϵ *fixed, and set* $T_W^* := n\hat{\Delta}^{*\top}\hat{W}\hat{\Delta}^*$ *with* $\hat{W}$ *held at its original-sample value.* Proved *in Proposition S7-d (a wild bootstrap with Rademacher weights; Liu, 1988; Mammen, 1993): under* H_0 *with conditionally symmetric errors and the moment conditions of Theorem 1-S, conditionally on the data,* $T_W^* \to_d \sum_j \omega_j Z_j^2$ *in probability, so that the* $(B+1)(1-\alpha)$*-th order statistic of* B *draws (*$B = 199$*, the 190th) is a critical value with asymptotic level* α *(up to the finite-*B *randomisation, which is exact under exchangeability). The sufficient conditions are: (i) in the pseudo-population (conditional on the data) decision adequacy holds exactly at* $\hat{\Theta}_E$, *because* $c^* - \hat{\Theta}_E\phi = se$ *is conditionally symmetric given* (ϕ, e) *and Lemma S1(c) applies with* $\Theta_0 = \hat{\Theta}_E$; *(ii) the conditional covariance of the stacked scores at* $\hat{\Theta}_E$ *equals the empirical score covariance, since both scores are odd in* s: $-2\,\phi \otimes se$ *and* $2\,\phi \otimes \big(x_\epsilon(\hat{c} + se) - x_\epsilon(\hat{c} - se)\big) = 2s\,\phi \otimes \big(x_\epsilon(\hat{c} + e) - x_\epsilon(\hat{c} - e)\big)$; *(iii) a conditional central limit theorem and uniform law of large numbers for the smoothed loss along the lines of*

the proof of Theorem 1-S, applied conditionally, which is Proposition S7-d; the bootstrap Hessian $\frac{4}{n}\sum_i \phi_i\phi_i^\top \otimes (-\nabla^2 z_\epsilon(\hat{c}_i - s_i e_i))$ *converges to* H_ϵ *under* H_0 *by the symmetry of* e*.*

(e) Local alternatives. *Under Assumption L-S,* $T_W \to_d (\mu+U)^\top B(\mu+U)$ *with* $U \sim \mathcal{N}(0, I_r)$ *and* $\mu = V_\Delta^{-1/2}\delta$ *in the standardised coordinates of Theorem S3, so the limit has mean* $\delta^\top W\delta + \tau_W$*: its noncentrality is the* $\delta^\top W\delta$ *of Theorem S2, the quantity the SURE rule estimates by* $\hat{q}$ *and the leading term of* λ^**. The test and the shrinkage rule read the same quadratic form.*

Proof. (a) By Theorem 1-S(a), $\sqrt{n}\hat{\Delta} \to_d Z_\Delta \sim \mathcal{N}(0, V_\Delta)$ under H_0; with $\hat{W} \to_p W$, Slutsky and the continuous mapping theorem give $T_W = (\sqrt{n}\hat{\Delta})^\top \hat{W}(\sqrt{n}\hat{\Delta}) \to_d Z_\Delta^\top W Z_\Delta$. Write $Z_\Delta = V_\Delta^{1/2}U$ with $U \sim \mathcal{N}(0, I_{q_M})$; then $Z_\Delta^\top W Z_\Delta = U^\top B U$, and diagonalising the symmetric PSD matrix B gives $\sum_j \omega_j Z_j^2$ with $Z = Q^\top U$ standard normal. $V_\Delta^{1/2} W V_\Delta^{1/2}$ and WV_Δ have the same nonzero eigenvalues (similar up to the null space), and the moments are those of a weighted sum of independent χ_1^2 variables.

(b) $\hat{\Delta} \to_p \Delta^*$ (consistency of both estimators, Theorem 1-S) and $\hat{W} \to_p W$ give $T_W/n \to_p \Delta^{*\top} W \Delta^* = |W^{1/2}\Delta^*|^2 > 0$; a critical value that is $O_p(1)$ is exceeded with probability tending to one.

(c) Matching $\mathbb{E}[a\chi_\nu^2] = a\nu = \tau_W$ and $\mathrm{Var}[a\chi_\nu^2] = 2a^2\nu = 2\sum_j \omega_j^2$ gives $a = \sum_j \omega_j^2/\tau_W$ and $\nu = \tau_W^2/\sum_j \omega_j^2$ (Satterthwaite, 1946; Box, 1954). The bounds: with m nonzero eigenvalues, $\sum \omega_j^2 \geq \tau_W^2/m$ (Cauchy-Schwarz) gives $\nu \leq m$ with equality iff the nonzero ω_j are equal; $\sum \omega_j^2 \leq \omega_1\tau_W$ gives $\nu \geq \tau_W/\omega_1 \geq 1$, again with equality iff the nonzero ω_j are equal, in which case $a = \omega_1$ and the limit in (a) is exactly $\omega_1\chi_\nu^2$. Consistency of $\hat{\nu}$ is the continuous mapping theorem applied to $\hat{W} \to_p W$ and $\hat{V}_\Delta \to_p V_\Delta$ (Proposition 1-S).

(d) Statement (i) is Lemma S1(c) applied to the pseudo-population, in which $c^* = \hat{\Theta}_E\phi + se$ is a correctly specified linear model with the conditionally symmetric error se (symmetric because s is Rademacher and independent of (ϕ, e)); hence the least-squares and smoothed SPO+ targets of the pseudo-population coincide at $\hat{\Theta}_E$, which is (ii)'s centring. The displayed identity in (ii) is $x_\epsilon(\hat{c}+se) - x_\epsilon(2\hat{c}-(\hat{c}+se)) = x_\epsilon(\hat{c}+se) - x_\epsilon(\hat{c}-se)$, odd in s, so that with $s^2 = 1$ the conditional covariance of the stacked scores is the empirical average of the outer products of the fixed vectors, the empirical score covariance at $\hat{\Theta}_E$. Given (iii), the conditional argument of Theorem 1-S yields $\sqrt{n}\hat{\Delta}^* \to_d \mathcal{N}(0, V_\Delta)$ in probability and (a) applies conditionally with $\hat{W}$ fixed; the order-statistic critical value is then a consistent estimate of the $1-\alpha$ quantile of the limit, up to the finite-B randomisation. (iii) is Proposition S7-d.

(e) Under Assumption L-S, $\sqrt{n}\hat{\Delta} \to_d G = \delta + Z_\Delta$ and $\hat{W} \to_p W$, so $T_W \to_d G^\top W G = (\mu + U)^\top B(\mu + U)$ in the coordinates of Theorem S3; the mean is $\delta^\top W\delta + \mathrm{tr}(WV_\Delta)$, the computation of Theorem S2 ($\mathbb{E}[G^\top W G]$ in its proof). □

Remark S8 (What is claimed and what is not; the recorded mechanism). (i) The size statement (a) is proved under H_0, which implies $\mathrm{H}_{0,W}$. For a parameter in $\mathrm{H}_{0,W} \setminus \mathrm{H}_0$ ($\Delta^* \neq 0$ with $W\Delta^* = 0$) the leading term of T_W vanishes but $n\Delta^{*\top}(\hat{W} - W)\Delta^*$ is $O_p(\sqrt{n})$ in general, so no statement is made there; such alternatives are decision-irrelevant to first order and the test is not designed to detect them. (ii) The relation to the rule: in the Hausman case the SURE weight is $w = \min(1, \tau_W/\max(T_W, \tau_W))$, so $\hat{\theta}_{\mathrm{SURE}} = \hat{\theta}_E$ exactly when $T_W \leq \tau_W$, the mean of the null law; the test asks whether T_W exceeds a null quantile of the same law. (iii) The mechanism behind the unstable critical values of the calibrated Wald test (Section 3.3): in the Wald geometry every retained direction is standardised by its own estimated variance, so a direction in which the decision loss is nearly flat (an item almost never in the optimal knapsack for any context; an arc rarely on the shortest path) enters with unit weight while its variance estimate carries the factor $\hat{H}_\epsilon^{-1}$ in that direction, whose relative error is unbounded as the curvature vanishes; the null copies of the Wald statistic then explode in the replicates where $\hat{H}_\epsilon$ is near singular. In the regret geometry the same

direction is weighted by its curvature under W, which is near zero there. This is an explanation of the recorded diagnostic, not a theorem.

Numerical verification of (a) and (c) (`tests/test_prop_s7.py`). On the toy problem at degree one and $n = 200000$ (seeds outside every preregistered range, 96 replicates, every smoothed fit meeting the gradient rule), the two-sample Kolmogorov-Smirnov test of the empirical law of T_W against Monte Carlo draws of the eigenvalue mixture $\sum_j \hat{\omega}_j Z_j^2$, $\hat{\omega}$ pooled over the replicates, has $p = 0.98$; the rejection frequency of T_W at the mixture's 0.95 quantile is 0.031; the Satterthwaite quantile is within a factor 1.0002 of the Monte Carlo quantile of the mixture; and $\hat{\nu} = 48.4$ against the effective dimension 33.2 of Theorem S3 and the rank 54. The check is of the limit law and of the approximation at one spectrum, not a finite-sample size statement.

G.10 S.9 Additions of the revision: JS-W, general smoothings, the name SURE, the two nulls

Corollary S6-JS (JS-W: the James-Stein constant in the regret geometry)**.** *Under the assumptions of Theorem S6 let* $a_{\mathrm{JS}} := \kappa_W - 2\nu_W = \tau_{\max}/2$ *(Hausman case:* $\operatorname{tr} B - 2\lambda_{\max}(B)$*; isotropic:* $p-2$*),* $w_{\mathrm{JS}}(t) := \min\{1, a_{\mathrm{JS}}/t\}$ *and* $\hat{\theta}_{JS\text{-}W} := \hat{\theta}_{I,\epsilon} - w_{\mathrm{JS}}(T_W)\hat{\Delta}_\epsilon$ *when* $a_{\mathrm{JS}} > 0$*,* $\hat{\theta}_{I,\epsilon}$ *otherwise. (a) JS-W returns ETO exactly when* $T_W \le a_{\mathrm{JS}}$ *and is the classical positive-part James-Stein rule in the isotropic Hausman case. (b)* $a_{\mathrm{JS}} \in (0, \min(\kappa_W, \tau_{\max})]$ *iff* $\kappa_W > 2\nu_W$*. (c) If* $\kappa_W > 2\nu_W$ *then for every* μ

$$\rho_W(E_{JS\text{-}W}) - \rho_W(E_{\mathrm{IEO}}) \le -a_{\mathrm{JS}}^2\, \mathbb{E}\Big[\frac{\mathbf{1}\{T_W > a_{\mathrm{JS}}\}}{T_W}\Big] - (\kappa_W + 2\nu_W)\Pr(T_W \le a_{\mathrm{JS}}) < 0,$$

and the same argument gives dominance for every constant in $(0, \tau_{\max}]$*; the floored member* $w = \min\{1, a_{\mathrm{JS}}/\max(T_W, \tau_W)\}$ *of the family of Theorem S6 (JS-W(floor) in the tables) dominates by Theorem S6(c). (d)* a_{JS} *is the midpoint of the admissible range: it maximises the coefficient* $a(\tau_{\max} - a)$ *of the Stein term of the bound, and minimises the risk integrand itself only in the isotropic case; no risk-optimality is claimed in general. (e)* $a_{\mathrm{JS}} \le a_{\mathrm{safe}} \le \kappa_W$*; in the Hausman case* $w_{\mathrm{JS}} \le w$ *pointwise, so the residual of JS-W over ETO under decision adequacy is at least that of SURE: the wider dominance region is bought with part of the ETO gain. (f) SURE's constant* κ_W *is admissible iff* $\kappa_W \ge 4\nu_W$*, JS-W's iff* $\kappa_W > 2\nu_W$*: four against two.*

Proof. (b) $a_{\mathrm{JS}} \le 2a_{\mathrm{JS}} = \tau_{\max}$, and $a_{\mathrm{JS}} \le \kappa_W$ because $\nu_W = \lambda_{\max}(\tilde{A}_s) \ge \operatorname{tr}\tilde{A}_s/r = \kappa_W/r > 0$. (c) w_{JS} is Lipschitz on $(0, \infty)$, so Stein's lemma applies as in Theorem S3(a) with $w' = -a_{\mathrm{JS}}/T_W^2$ on $\{T_W > a_{\mathrm{JS}}\}$ and 0 below. On $\{T_W \le a_{\mathrm{JS}}\}$ the integrand is $T_W - 2\kappa_W \le a_{\mathrm{JS}} - 2\kappa_W = -(\kappa_W + 2\nu_W)$; on $\{T_W > a_{\mathrm{JS}}\}$ it is $-(a_{\mathrm{JS}}/T_W)(2\kappa_W - a_{\mathrm{JS}} - 4U^\top(AB)_s U/T_W) \le -(a_{\mathrm{JS}}/T_W)(\tau_{\max} - a_{\mathrm{JS}})$ by the Rayleigh bound of Theorem S6(b), which does not involve the floor. (d) $a \mapsto a(\tau_{\max} - a)$ is maximal at $\tau_{\max}/2$; in the isotropic case the Rayleigh bound is an equality. (e), (f) from Theorem S6(d) and the definitions. □

Numerical check in the Hausman limit experiment at effective dimensions 2.5, 3, 5 and 10 (`tests/test_sure_safe.py`; the expectation was committed before the first run): the worst excess of JS-W over IEO on the δ grid is -0.0009, -0.0028, -0.0137 and -0.0457 of the IEO risk, equal to four decimals for the floored variant; the residual over ETO at the null, relative to the ETO risk, is 0.6844, 0.5016, 0.2813, 0.1263 for JS-W, 0.7474, 0.5682, 0.3196, 0.1389 with the floor and 0.1695, 0.1228, 0.1153, 0.0670 for SURE.

Remark S1-G (Lemma S1(b) and (c) for a general smoothing; the barrier smoothing). Let $\tilde{z}$ be concave and C^2 with $\nabla\tilde{z} \in \operatorname{conv} X$, affine along L, with $-\nabla^2\tilde{z}$ positive definite on M and bounded, and let both terms of the loss use $\tilde{z}$. Then parts (b) and (c) of Lemma S1 hold (with the Hessian bound $4|\phi|^2 \sup\|\nabla^2\tilde{z}\|$ and the lower bound $2(c^\top\tilde{x}(c) - \tilde{z}(c))$ in place of nonnegativity), and with them Theorem 1-S, the consistency part of Proposition 1-S, Theorems S2, S3, S6, Corollary S6-JS and Proposition S7: their proofs use the smoothing only through (b), (c) and the Hessian bound. The value bounds of (a) and the limit of (d) are not inherited. The log-barrier smoothing $\tilde{z}_\beta(w) = \min_{x\in X}[w^\top x - \beta\sum_i \log(b_i - a_i^\top x)]$ of a bounded polytope with nonempty relative interior satisfies the conditions (unique interior minimiser; envelope theorem; implicit function theorem; $N^\top\mathcal{H}N \succeq (\beta/s_{\max}^2)N^\top\sum_i a_i a_i^\top N$), with bias $0 \le w^\top\tilde{x}_\beta - z^* \le m\beta$. It appears in this paper only in the exploratory rows of Appendix B, with no claim.

Remark S3-N (The name SURE and the constant). $\hat{\lambda}$ minimises the unbiased estimate of the *fixed-λ* risk of Theorem S2; it is not the minimiser of an unbiased estimate of the risk of the data-dependent rule. The Stein correction for the data-dependent weight, the divergence term of Theorem S3(a), is what moves the constant of the bound from κ_W to $\kappa_W - 2\nu_W$.

Remark S8-H (The two nulls and the verbal definition). H_0 ($\Delta^* = 0$) implies $\mathrm{H}_{0,W}$ ($W^{1/2}\Delta^* = 0$) and equal population decisions (the two predictions differ by an element of L). Equal decisions do not imply H_0: a positive rescaling of the identified part of the prediction leaves every decision unchanged. Separately, a correctly specified mean with skewed noise makes ETO decision-optimal while $\nabla R_\epsilon(\theta_0) \neq 0$ in general, so $\Delta^* \neq 0$: there the test rejects although no training can improve the decisions. The test is a test of H_0 and reads asymmetry of the noise as misspecification (measured in Table 8 and Appendix E).

G.11 S.10 The sign-flip reference: condition (iii) of Proposition S7(d), proved

The reference is the one implemented: $e_i := c_i - \hat{\Theta}_E\phi_i$, $s_1, \ldots, s_n$ i.i.d. Rademacher and independent of the data $D_n := (\phi_i, c_i)_{i\le n}$, $c_i^* := \hat{c}_i + s_i e_i$ with $\hat{c}_i := \hat{\Theta}_E\phi_i$, both estimators refitted on (ϕ_i, c_i^*) with ϵ and $\hat{W}$ held at their original-sample values, the entropic smoothing. $\Pr^*$, $\mathbb{E}^*$ and Var^* denote the conditional law, mean and variance given D_n.

Proposition S7-d (The sign-flip reference: conditional limit law, level and consistency)**.** *Let Assumptions A and C-ϵ hold, let $\epsilon > 0$ be fixed and the smoothing be entropic, and let $\hat{W} \to_p W$ positive semidefinite.*

(a) Under H_0 in the form of Lemma S1(c) *($\mathbb{E}[c \mid z] = \Theta_0\phi$ and $c - \Theta_0\phi$ conditionally centrally symmetric given z): with V_Δ the limit covariance of Theorem 1-S,*

$$\sup_x \left| \Pr{}^*\big(\sqrt{n}\, B_M^\top \hat{\Delta}^* \le x\big) - \Pr\big(\mathcal{N}(0, V_\Delta) \le x\big) \right| \to_p 0,$$
$$\sup_t \left| \Pr{}^*\big(T_W^* \le t\big) - \Pr\Big(\sum\nolimits_j \omega_j Z_j^2 \le t\Big) \right| \to_p 0,$$

with ω_j the eigenvalues of WV_Δ as in Proposition S7(a). Hence, if $\sum_j \omega_j > 0$, the $(1-\alpha)$ quantile of the conditional law of T_W^* converges in probability to the $(1-\alpha)$ quantile of the limit law of T_W, and the test that rejects when T_W exceeds it has asymptotic level α; the $(B+1)(1-\alpha)$-th order statistic of B conditional draws estimates that quantile with the usual finite-B randomisation.

(b) Under a fixed alternative *with $W^{1/2}\Delta^* \neq 0$ (and no symmetry assumed): the conditional law of T_W^* is tight in probability, so the critical value is $O_p(1)$ and Proposition S7(b) gives consistency.*

(c) A data-dependent scale. *If $\epsilon = \hat{\epsilon}$ is a measurable function of D_n taking values in a fixed compact subset of $(0, \infty)$ (the implemented $\epsilon_0 \hat{s}_r$, with ϵ_0 fixed or chosen from a finite set by a rule that reads D_n only), then $\hat{\epsilon}$ is a constant under $\Pr^*$ and the conditional statements of (a) and (b) hold with V_Δ, ω evaluated at $\hat{\epsilon}$, because every bound used below is uniform over such a compact set. What this does* not *give is the unconditional law of T_W itself at a data-chosen scale: Proposition S7(a) is for a fixed ϵ, it extends to $\hat{\epsilon} \to_p \epsilon$ by the same uniformity, and for a choice among finitely many scales that does not settle on one of them the size of the test is measured (the claim E1-r of the third scale round, Appendix E), not proved.*

Proof. Work in the identified coordinates $\beta = B_M^\top\theta$ of Theorem 1-S, anchored at $\hat{\theta}_E$; write $t \in \mathbb{R}^{q_M}$ for a local parameter and $g_{E,i} := -2B_M^\top(\phi_i \otimes e_i)$, $g_{I,i} := 2B_M^\top(\phi_i \otimes I_d)\big(x_\epsilon(\hat{c}_i + e_i) - x_\epsilon(\hat{c}_i - e_i)\big)$, the original-sample scores of the two losses at $\hat{\theta}_E$ (for ℓ_ϵ: equation 11 with $c_i = \hat{c}_i + e_i$ and $2\hat{c}_i - c_i = \hat{c}_i - e_i$), and $g_i := (g_{E,i}, g_{I,i})$.

Step 1, the ETO refit. Least squares is linear in the response: $\hat{\Theta}_E^* - \hat{\Theta}_E = \big(\sum_i s_i e_i \phi_i^\top\big)\big(\sum_i \phi_i\phi_i^\top\big)^{-1}$ exactly, so $\sqrt{n}\, B_M^\top(\hat{\theta}_E^* - \hat{\theta}_E) = -\hat{H}_{E,M}^{-1}\, n^{-1/2} \sum_i s_i g_{E,i}$ with $\hat{H}_{E,M} := B_M^\top\big(\frac{2}{n}\sum_i \phi_i\phi_i^\top \otimes I_d\big)B_M$, and $\mathbb{E}^*$ of it is zero: the refit is exactly centred at $\hat{\Theta}_E$.

Step 2, the smoothed pseudo-risk. $R^*_{n,\epsilon}(\theta) := n^{-1}\sum_i \ell_\epsilon(\theta;\phi_i,c^*_i)$ is convex and C^∞ (Lemma S1(b)), with Hessian bounded by $\frac{4}{n}\sum_i|\phi_i|^2R_X^2/\epsilon$ and third derivative bounded by $\frac{c}{n}\sum_i|\phi_i|^3R_X^3/\epsilon^2$ (Lemma S1(b); proof of Proposition 1-S, case (E)), both free of s. At $\hat\theta_E$, $2\hat c_i - c^*_i = \hat c_i - s_ie_i$, so the score of observation i is $2(\phi_i\otimes I_d)\big(x_\epsilon(\hat c_i+s_ie_i)-x_\epsilon(\hat c_i-s_ie_i)\big)$, which is odd in s_i: in the identified coordinates it equals $s_ig_{I,i}$ (condition (i): Lemma S1(c) in the pseudo-population, whose error s_ie_i is symmetric given D_n). Hence $\mathbb{E}^*\nabla R^*_{n,\epsilon}(\hat\theta_E)=0$. The Hessian at $\hat\theta_E$ is $D^*_n := B_M^\top\big[\frac{4}{n}\sum_i\phi_i\phi_i^\top\otimes(-\nabla^2 z_\epsilon(\hat c_i - s_ie_i))\big]B_M$, with

$$\mathbb{E}^*D^*_n = \bar D_n := B_M^\top\Big[\frac{2}{n}\sum_i\phi_i\phi_i^\top\otimes\big(-\nabla^2 z_\epsilon(\hat c_i-e_i)-\nabla^2 z_\epsilon(\hat c_i+e_i)\big)\Big]B_M,$$

the symmetrised Hessian. Each entry of D^*_n is an average of conditionally independent terms bounded by $4|\phi_i|^2R_X^2/\epsilon$, so Var^* of each entry is at most $n^{-2}\sum_i 16|\phi_i|^4R_X^4/\epsilon^2 = O_p(1/n)$ under $\mathbb{E}|\phi|^4<\infty$, and $D^*_n-\bar D_n\to 0$ in $\Pr^*$-probability, in probability. $\bar D_n$ is positive definite on the identified coordinates whenever $\sum_i\phi_i\phi_i^\top$ is, by the argument of Lemma S1(b) with the empirical measure ($-\nabla^2 z_\epsilon(w)$ is positive definite on M for every w). Its limit: $w\mapsto\nabla^2 z_\epsilon(w)$ is Lipschitz with constant cR_X^3/ϵ^2, $\hat\Theta_E\to_p\Theta^*_E$, and the summands are dominated by $4|\phi|^2R_X^2/\epsilon$, so by the uniform law of large numbers used in Proposition 1-S $\bar D_n\to_p\bar D := B_M^\top 2\mathbb{E}\big[\phi\phi^\top\otimes(-\nabla^2 z_\epsilon(m-e)-\nabla^2 z_\epsilon(m+e))\big]B_M$ with $m:=\Theta^*_E\phi$, $e:=c-m$, positive definite. Under the null of (a), $\Theta^*_E=\Theta_0$ and $e\mid z$ is symmetric, so the two halves have the same expectation and $\bar D = B_M^\top H_\epsilon(\theta_0)B_M$, the Hessian of Theorem 1-S ($2\Theta_0\phi - c = m-e$).

Step 3, the convex-process expansion, conditionally. For fixed D_n put $A^*_n(t) := n\big[R^*_{n,\epsilon}(\hat\theta_E + B_Mt/\sqrt n) - R^*_{n,\epsilon}(\hat\theta_E)\big]$, a convex function of t whose minimiser is $\sqrt n\,B_M^\top(\hat\theta^*_{I,\epsilon}-\hat\theta_E)$ (the refit is anchored at the refitted ETO; the two anchors differ on the unidentified directions only, which $B_M^\top$ removes). Taylor's formula with the bounds of Step 2 gives

$$A^*_n(t) = t^\top U^*_{I,n} + \tfrac12 t^\top D^*_n t + r^*_n(t), \qquad U^*_{I,n} := n^{-1/2}\sum_i s_ig_{I,i},$$

$$|r^*_n(t)| \le \frac{c\,R_X^3}{6\,\epsilon^2}\,|t|^3\,n^{-1/2}\,\frac1n\sum_i|\phi_i|^3,$$

the last bound deterministic in s; it tends to zero in probability because $\frac1n\sum_i|\phi_i|^3 = O_p(1)$. The basic corollary of Hjort & Pollard (1993) (proof of Theorem 1-S) is a statement about a sequence of random convex functions; it is applied under $\Pr^*$ along data sequences on which $\bar D_n\to\bar D$, $\frac1n\sum_i|\phi_i|^3$ stays bounded and the covariance and Lindeberg quantities of Step 4 converge (with $D^*_n-\bar D$ absorbed in the remainder, which is $o_{\Pr^*}(1)$ for each fixed t by Step 2). Every subsequence of n has a further subsequence along which these hold almost surely, which turns the conclusion into convergence in probability of the conditional laws (the distance in (a) is a bounded function of D_n, and a sequence converges in probability iff every subsequence has an almost surely convergent further subsequence). The conclusion is $\sqrt n\,B_M^\top(\hat\theta^*_{I,\epsilon}-\hat\theta_E) = -\bar D^{-1}U^*_{I,n} + o_{\Pr^*}(1)$.

Step 4, the conditional central limit theorem. $U^*_n := n^{-1/2}\sum_i s_ig_i$ is a sum of conditionally independent, centred vectors with $\mathrm{Var}^*(U^*_n) = \frac1n\sum_i g_ig_i^\top$, the empirical covariance of the stacked scores at $\hat\theta_E$, exactly (condition (ii): $s_i^2=1$). By the uniform law of Proposition 1-S and $\hat\Theta_E\to_p\Theta^*_E$ it converges in probability to $\tilde\Sigma := \mathbb{E}[gg^\top]$ at θ^*_E, which under the null of (a) is the stacked score covariance Σ^ϵ of Theorem 1-S at θ_0. Lindeberg: for Rademacher weights and fixed vectors, $\sum_i\mathbb{E}^*\big[|s_ig_i|^2n^{-1}\mathbf{1}\{|s_ig_i|>\eta\sqrt n\}\big] = n^{-1}\sum_i|g_i|^2\mathbf{1}\{|g_i|>\eta\sqrt n\}$, which vanishes as soon as $\max_i|g_i|\le\eta\sqrt n$; and $\max_i|g_i|^2/n\to_p 0$ because $|g_{I,i}|\le 4R_X|\phi_i|$, $|g_{E,i}|\le 2|\phi_i||e_i|$, and $\mathbb{E}|\phi|^4<\infty$, $\mathbb{E}|c|^4<\infty$ give $\mathbb{E}[|\phi|^2+|\phi|^2|e|^2]<\infty$ (the maximum of n i.i.d. integrable variables is $o(n)$ almost surely; the estimation error in e_i is $|\hat\Theta_E-\Theta^*_E||\phi_i|$). The multivariate Lindeberg-Feller theorem, applied conditionally along the subsequences of Step 3, gives $U^*_n\to_d\mathcal N(0,\tilde\Sigma)$ under $\Pr^*$, in probability.

Step 5, conclusion. By Steps 1, 3 and 4, $\sqrt n\,B_M^\top\hat\Delta^* = -\bar D^{-1}U^*_{I,n} + \hat H^{-1}_{E,M}U^*_{E,n} + o_{\Pr^*}(1)$ is conditionally asymptotically normal with covariance $\tilde V_\Delta := (H^{-1}_{E,M},\ -\bar D^{-1})\,\tilde\Sigma\,(H^{-1}_{E,M},\ -\bar D^{-1})^\top$

(blocks in the order ETO, IEO); under the null of (a), $\bar{D} = H_{\epsilon,M}$ and $\tilde{\Sigma} = \Sigma^{\epsilon}$, so $\tilde{V}_\Delta = V_\Delta$, the sandwich of Theorem 1-S. Pólya's theorem gives the uniform statement. With $\hat{W}$ fixed given D_n and $\hat{W} \to_p W$, the continuous mapping theorem applied conditionally gives the law of T_W^*, as in the proof of Proposition S7(a); the limit law is continuous and strictly increasing on $(0, \infty)$ when $\sum_j \omega_j > 0$, so the quantiles converge, and the level follows from Proposition S7(a). (b) Under a fixed alternative nothing in Steps 1 to 4 uses the null except the identification of the limits: $\bar{D}$ is positive definite and $\tilde{\Sigma}$ finite, so T_W^* has a conditional limit law $\sum_j \tilde{\omega}_j Z_j^2$ with $\tilde{\omega}$ the eigenvalues of $W\tilde{V}_\Delta$, its quantile is $O_p(1)$, and Proposition S7(b) applies. (c) Given D_n, $\hat{\epsilon}$ is a constant; the constants R_X^2/ϵ and cR_X^3/ϵ^2 are bounded on a compact subset of $(0, \infty)$, the uniform laws hold uniformly over it (the summands are Lipschitz in ϵ there with integrable envelopes), so Steps 1 to 5 hold with $\bar{D}$, $\tilde{\Sigma}$ evaluated at $\hat{\epsilon}$. The last sentence of (c) records what is not proved. □

Conditions used. Assumption A ($\mathbb{E}|\phi|^4, \mathbb{E}|c|^4 < \infty$, $\mathbb{E}\phi\phi^\top \succ 0$); Assumption C-$\epsilon$; the entropic smoothing at a fixed (or, conditionally, data-dependent and bounded away from zero and infinity) scale, for the third derivative bound; for the level statement (a), the null in the form of Lemma S1(c): a correctly specified conditional mean with conditionally symmetric errors given ϕ. Under a null with $\Delta^* = 0$ but asymmetric errors or a misspecified mean, $\tilde{V}_\Delta$ need not equal V_Δ and no level statement is made; those cases (N1, N2, N2-s, N4) and the asymmetric case N3 are measured (Table 8).

Numerical checks of Proposition S7-d (`tests/test_prop_s7.py`). On one data set of the toy problem at degree one and $n = 200000$ (a seed outside every preregistered range), the conditional law of T_W^* over 2000 sign draws with $\hat{W}$ held fixed, every refit meeting the gradient rule, was compared with Monte Carlo draws of the eigenvalue mixture at $\hat{\omega}$: the two-sample Kolmogorov-Smirnov test has $p = 0.567$, and the ratio of the 0.95 quantiles is 0.9925. The frozen size cells of the third round, read without any re-run, are in Table 17. The first check is of the conditional limit law at one large n, not a finite-sample size statement; the size statements are the preregistered claims of Table 8.