When Interpolation Is Loss-Blind: Quantile Regression at the Interpolation Boundary

Anonymous authors
Paper under double-blind review

Abstract

Quantile regression turns a flexible predictor into prediction intervals by fitting several conditional quantiles, and the intervals are informative only when these fits differ across levels. Models that can fit every training response exactly are usually regularized by a ridge penalty, or weight decay, which makes the penalized coefficients unique. In this article, we show that uniqueness does not separate the levels on an explicit range of penalties. For fixed features whose training feature matrix has full row rank, we derive for each quantile level a closed-form penalty boundary, computed with one linear solve, at or below which the fit of that level equals the minimum-norm interpolant, which does not depend on the level. Theoretical results show that two distinct levels share a fit exactly when the common penalty is at or below both boundaries, with or without an unpenalized intercept, and that the same certificate applies to the linear quantile heads of a network at any local minimum of training at which the learned feature matrix has full row rank. With Gaussian features and noise and a feature count growing faster than the sample size, the boundary shrinks at a fixed response scale, and extinct fits have coverage converging to one half. For bounded responses and features fixed before sampling, validation above the boundary over a fixed penalty grid controls the excess pinball risk up to the approximation error of the features. Through simulations, we confirm the predicted coverage transition at the boundary. On four regression data sets and a simulated model, the weight decay 10^{-4} used to train a network lies below the boundaries computed from its learned features in 499 of 500 splits, and the five trained quantile heads have mean test coverage between 0.482 and 0.538 on every data set, at nominal levels from 0.1 to 0.9. Refitting the heads with a validated weight decay far above the boundary moves the pooled mean coverage at the levels 0.1 and 0.9 to 0.166 and 0.825.

1 Introduction

Quantile regression turns a predictor into a description of predictive uncertainty: the fitted conditional 10th and 90th percentiles, for example, form an 80% prediction interval (Koenker & Bassett, 1978; Koenker & Hallock, 2001; Koenker, 2005). Neural quantile models, kernel quantile regression, quantile regression forests, gradient boosting, and linear quantile heads on random or learned features are standard choices for this task (Taylor, 2000; Cannon, 2011; Takeuchi et al., 2006; Li et al., 2007; Meinshausen, 2006; Friedman, 2001; Tagasovska & Lopez-Paz, 2019; Gasthaus et al., 2019; Lim et al., 2021). Such models often have more parameters than observations and can fit every training response exactly. Their fitted percentiles are useful only when they differ across levels. If the fitted 10th and 90th percentiles coincide, the interval has zero width even when the true percentiles are far apart.

An interpolating fit makes the check loss zero at every level, and it minimizes the unregularized objective for all quantile levels at once (Sivakumar & Banerjee, 2017). For overparameterized Gaussian linear models with symmetric sub-Gaussian noise, Bai et al. (2021) show that zero-initialized subgradient descent with decreasing step sizes has a minimum-norm limit with coverage near one half at levels above one half. A ridge penalty, also called weight decay, makes the penalized coefficients unique and is standard in kernel and neural quantile regression (Takeuchi et al., 2006; Li

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.

et al., 2007; Tu et al., 2024). Whether a positive penalty also separates the levels has not been characterized. The difficulty is that the check loss is not differentiable at zero: its subgradient interval can absorb the gradient of the penalty, and an interpolating fit can remain optimal over an interval of positive penalties. Exact regularization of linear programs guarantees that such an interval exists (Friedlander & Tseng, 2008, Corollary 2.3). Solution-path algorithms for kernel quantile regression compute it numerically, because they stop when every observation is interpolated (Li et al., 2007), and their knots move piecewise linearly with the quantile level (Rosset, 2009). These results do not give the interval in closed form, do not decide whether distinct levels can coincide outside it, and do not relate it to the penalties used in practice or to features learned by a network.

In this article, we study ridge-penalized quantile regression in models with a fixed feature map, including random features and a linear output layer on learned features. Let $y \in \mathbb{R}^n$ be the training responses and let $X \in \mathbb{R}^{n \times p}$ be the feature matrix, of rank n, which makes $X\beta = y$ solvable for every y. For a level $\tau \in (0, 1)$ and a penalty $\lambda > 0$, the fitted coefficient vector $\widehat{\beta}(\lambda, \tau)$ minimizes $\sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top \beta) + \lambda \|\beta\|_2^2 / 2$, where $\rho_\tau(u) = u\{\tau - \mathbf{1}(u < 0)\}$ is the check loss and $x_i^\top$ is the ith row of X. The key observation is that the dual variables of the check loss range over the box $[\tau - 1, \tau]^n$, and the interpolating fit remains optimal exactly while the vector $\lambda(XX^\top)^{-1}y$ stays in this box. This gives, for each level, a closed-form penalty boundary computed from one linear solve.

This article makes three contributions. First, we characterize exactly when the penalized fit loses its dependence on the quantile level. At or below the boundary of level τ, the fit equals the minimum-norm interpolant, which is the exact fit with the smallest coefficient norm. We call this state extinction, and the fit is then loss-blind: it does not depend on the level of the check loss. Two distinct levels have identical fits at a common penalty if and only if both are extinct. For a single common predictor, the largest marginal calibration error over the requested levels is at least half their range. These identities hold with an unpenalized intercept and use only the realized features. They thus also apply to the linear quantile heads of a network at any local minimum of training at which the learned feature matrix has full row rank, whatever the architecture that produced the features.

Second, we determine the scale of the boundary. For Gaussian features and noise with $n/p \to 0$, $s_{n,p}\sqrt{2 \log n}\, \bar{\lambda}(\tau)$ converges in probability to $\min\{\tau, 1 - \tau\}$ uniformly in τ, where $\bar{\lambda}(\tau)$ is the boundary of level τ and $s_{n,p}$ is the marginal standard deviation of the response. The first level to become extinct as the penalty decreases thus approaches the median. In the usual averaged-loss parameterization, the boundary is a weight decay of order $\{ns_{n,p}\sqrt{\log n}\}^{-1}$. With an unpenalized intercept, every fixed level is extinct with probability tending to one at penalties of smaller order, and the common fit of the extinct levels has coverage converging to one half. Bai et al. (2021) reached the analogous conclusion for the minimum-norm interpolant of unregularized quantile regression. For non-Gaussian features, the concentration of the signal across features can change the scale.

Third, we examine the consequences for penalty selection and for trained models. Above the boundary, validation over a fixed penalty grid controls the excess pinball risk up to the approximation error of the features when responses and features are bounded and the features are fixed before sampling; the guaranteed penalty lies above the target levels' boundaries with high probability. Through simulations, we confirm the coverage transition predicted by the theory. On four public regression data sets and a simulated model, the features learned by a two-hidden-layer network trained with weight decay 10^{-4} have median boundaries $\bar{\lambda}_c(0.5)/n$ between 9.0×10^{-3} and 3.2×10^{-2}, on the scale of the weight decay, whereas random Fourier features on the same data have median boundaries between 2.2×10^{-7} and 1.8×10^{-5}. The five trained quantile heads then nearly coincide, with mean test coverage between 0.482 and 0.538 on every data set, at nominal levels from 0.1 to 0.9. Exact refits of the heads on the learned features keep coverage near one half up to about 30 times the boundary and approach the nominal levels only at larger weight decays, and raising the training weight decay to 10^{-2} separates the refitted heads without calibrating them.

Our results complement several other literatures. The interpolation literature studies prediction and implicit bias for least-squares and classification losses (Belkin et al., 2019; Hastie et al., 2022; Muthukumar et al., 2021; Lai & Muthukumar, 2025), and studies of regularized quantile regression address estimation accuracy, coefficient paths, and computation (Rosset & Zhu, 2007; Takeuchi et al., 2009; Tu et al., 2024; Tang et al., 2026; Wang & Feng, 2024). Other approaches ensure the proper ordering of quantile curves (He, 1997; Chernozhukov et al., 2010; Cannon, 2018; Narayan et al., 2024), and calibration methods correct coverage bias (Romano et al., 2019; Gibbs et al., 2025). These approaches target the ordering or the coverage of fitted quantiles and not the coincidence of

penalized fits at distinct levels. Thresholds have also been studied for linear programs with Gaussian objective coefficients (Friedlander et al., 2025).

The remainder of the article is organized as follows. Section 2 gives the exact extinction and level-separation results, including intercepts and learned features. Section 3 analyzes the boundary and the resulting coverage under Gaussian random features. Section 4 gives the accuracy guarantee above the boundary. Section 5 reports the numerical studies, and Section 6 discusses scope and use. Appendix A proves the main results and gives finite-sample Gaussian bounds and the sub-Gaussian extension. The Supplementary Materials give the Gaussian limits at general dimension ratios, the non-Gaussian results, further deterministic extensions, and the numerical details.

2 EXTINCTION: THE INTERVAL, THE ORDER, AND EXACT SEPARATION

Throughout, $\|\cdot\|_2$, $\|\cdot\|_\infty$, I_d, $\mathbf{1}_n$, and $\mathbf{1}(\cdot)$ denote the Euclidean norm, maximum absolute coordinate norm, $d \times d$ identity matrix, vector of n ones, and an indicator; $\mathbb{E}$ and $\mathbb{P}$ denote expectation and probability; and $u \wedge v = \min\{u, v\}$ and $u \vee v = \max\{u, v\}$. Let $X \in \mathbb{R}^{n\times p}$ be the rank-n fixed design, $p \geq n$, and put $K = XX^\top$ and $Q = K^{-1}$. For $\tau \in (0, 1)$ and $\lambda > 0$, define

$$\widehat{\beta}(\lambda, \tau) = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top \beta) + \frac{\lambda}{2} \|\beta\|_2^2 ,$$

and let

$$a = Qy, \qquad \widehat{\beta}_0 = X^\top a, \qquad \mathcal{B}_\tau = [\tau - 1, \tau]^n.$$

Here $\widehat{\beta}_0$ is the minimum-norm vector among all β satisfying $X\beta = y$, and we call a the interpolation vector. For the averaged objective $n^{-1} \sum_i \rho_\tau(y_i - x_i^\top \beta) + \gamma \|\beta\|_2^2 /2$ with weight decay $\gamma > 0$, the minimizer is $\widehat{\beta}(n\gamma, \tau)$.

A two-observation example shows the phenomenon. Suppose $X = I_2$ and $y = (3, -1)^\top$, which gives $a = \widehat{\beta}_0 = y$. The objective separates over the two coordinates, and $\widehat{\beta}_0$ is optimal exactly when zero is a subgradient of the objective at $\widehat{\beta}_0$, that is, when $0 \in [-\tau, 1-\tau]+3\lambda$ and $0 \in [-\tau, 1-\tau]-\lambda$. Hence $\widehat{\beta}(\lambda, \tau) = \widehat{\beta}_0$ if and only if $\lambda \leq \tau/3 \wedge (1 - \tau)$. The levels $1/2$ and $3/4$ thus give identical coefficients for every $\lambda \in (0, 1/6]$, although the penalty is positive.

For $\lambda > 0$ and $\tau \in (0, 1)$, let $g_{\lambda,\tau}$ minimize $(g-\lambda a)^\top K(g-\lambda a)$ over $g \in \mathcal{B}_\tau$. By convex duality and subgradient optimality (Rockafellar, 1970; Boyd & Vandenberghe, 2004), this projection determines the coefficients and the training residuals, as the following proposition shows.

Proposition 1. *For every $\lambda > 0$ and $\tau \in (0, 1)$, the minimizers $\widehat{\beta}(\lambda, \tau)$ and $g_{\lambda,\tau}$ are unique, and*

$$\widehat{\beta}(\lambda, \tau) = \lambda^{-1} X^\top g_{\lambda,\tau}, \qquad y - X\widehat{\beta}(\lambda, \tau) = K(a - \lambda^{-1} g_{\lambda,\tau}).$$

The exact fit remains optimal exactly when the unconstrained dual point λa already lies in the subgradient box $\mathcal{B}_\tau$. Define, for $i \leq n$,

$$b_i(\tau) = \begin{cases} \tau/a_i, & a_i > 0, \\ (1 - \tau)/|a_i|, & a_i < 0, \\ +\infty, & a_i = 0, \end{cases} \qquad \bar{\lambda}(\tau) = \min_{i\leq n} b_i(\tau).$$

Theorem 1. *For every $\tau \in (0, 1)$ and every $\lambda > 0$,*

$$\widehat{\beta}(\lambda, \tau) = \widehat{\beta}_0 \qquad \Longleftrightarrow \qquad 0 < \lambda \leq \bar{\lambda}(\tau).$$

If $a \neq 0$, then $\bar{\lambda}(\tau) > 0$ for every $\tau \in (0, 1)$. If $a = 0$, then $\bar{\lambda}(\tau) = +\infty$ for every $\tau \in (0, 1)$.

Theorem 1 shows that the fitted coefficients do not merely approach the minimum-norm interpolant as the penalty decreases: they equal it over an interval of strictly positive penalties. We call the level τ extinct at the penalty λ when $\widehat{\beta}(\lambda, \tau) = \widehat{\beta}_0$, and we call $\bar{\lambda}(\tau)$ the extinction boundary. Above the boundary, the projection in Proposition 1 lies on a face of the box, and at least one training residual is nonzero.

For $u \in \mathbb{R}$, write $u^+ = \max(u, 0)$, put $A_+ = \max_{i\leq n} a_i^+$ and $A_- = \max_{i\leq n}(-a_i)^+$, and use the convention $c/0 = +\infty$ for $c > 0$. The following theorem gives the shape of the boundary and the order of extinction.

Theorem 2. *For every $\tau \in (0,1)$,*

$$\bar{\lambda}(\tau) = \frac{\tau}{A_+} \wedge \frac{1-\tau}{A_-}.$$

If $A_+ > 0$ and $A_- > 0$, then $\bar{\lambda}$ is concave and piecewise linear, with unique maximizer $\tau^\dagger = A_+/(A_+ + A_-)$ and maximum $\bar{\lambda}(\tau^\dagger) = (A_+ + A_-)^{-1}$. If $A_+ = A_- = 0$, then every level is extinct at every $\lambda > 0$. If $A_+ + A_- > 0$ and $0 < \lambda \le (A_+ + A_-)^{-1}$, then

$$\{\tau \in (0,1) : \lambda \le \bar{\lambda}(\tau)\} = [\lambda A_+,\, 1 - \lambda A_-] \cap (0,1),$$

with the convention that $[\lambda A_+, 1)$ or $(0, 1-\lambda A_-]$ replaces this interval when $A_- = 0$ or $A_+ = 0$, respectively.

When both signs occur in a, the graph of $\bar{\lambda}$ is a triangle. Its apex $\tau^\dagger$ is the first level to become extinct as the penalty decreases, and the extinct set grows outward from $\tau^\dagger$ toward the extreme quantiles. In the two-observation example, $A_+ = 3$, $A_- = 1$, and $\tau^\dagger = 3/4$. Because Theorem 2 expresses every boundary through A_+ and A_-, all boundaries cost one linear solve for a.

Corollary 1. *Let $\mathcal{T} \subset (0,1)$ be nonempty and suppose $0 < \lambda \le \inf_{\tau\in\mathcal{T}} \bar{\lambda}(\tau)$. Then $x^\top \widehat{\beta}(\lambda,\tau) = x^\top \widehat{\beta}_0$ for every $x \in \mathbb{R}^p$ and every $\tau \in \mathcal{T}$.*

Below their joint boundary, the fitted quantile curves at all levels in $\mathcal{T}$ coincide at every covariate value. Corollary 4 in Appendix A quantifies the cost for a finite set of levels. The largest calibration error of the common predictor over the levels is at least half the difference between the largest and the smallest level. Under moment conditions, its squared quantile error, averaged over the levels, is at least the mean squared deviation of the target quantiles from their average. Under a lower bound $\underline{f}$ on the conditional density between two target quantiles, the summed excess pinball risk of these two levels is at least $\underline{f}/4$ times the mean squared gap between the quantiles. The following theorem shows that coincidence cannot occur above the boundary.

Theorem 3. *Let $\lambda > 0$ and let $\tau, \tau' \in (0,1)$ with $\tau \ne \tau'$. Then*

$$\widehat{\beta}(\lambda,\tau) = \widehat{\beta}(\lambda,\tau') \qquad \Longleftrightarrow \qquad 0 < \lambda \le \bar{\lambda}(\tau) \wedge \bar{\lambda}(\tau'),$$

and in this case $\widehat{\beta}(\lambda,\tau) = \widehat{\beta}(\lambda,\tau') = \widehat{\beta}_0$.

Theorem 3 shows that exact coincidence of two levels certifies extinction: the two fitted coefficient vectors differ if and only if the common penalty exceeds $\bar{\lambda}(\tau) \wedge \bar{\lambda}(\tau')$. When a single observation attains the minimum that defines $\bar{\lambda}(\tau)$, only its residual becomes nonzero just above the boundary, and Proposition 2 in Appendix A gives the exact departure.

Intercepts and learned features. Most implementations leave the intercept unpenalized. With an intercept $\alpha \in \mathbb{R}$, let $\mathcal{M}_c(\lambda,\tau)$ be the set of minimizers of $\sum_{i=1}^n \rho_\tau(y_i - \alpha - x_i^\top\beta) + \lambda\,\|\beta\|_2^2/2$ over (α,β), and let $\{\widehat{\alpha}_c(\lambda,\tau), \widehat{\beta}_c(\lambda,\tau)\}$ be any measurable selection from it. Put $\alpha_Q = \mathbf{1}_n^\top Q y / \mathbf{1}_n^\top Q \mathbf{1}_n$ and $a_c = Q(y - \alpha_Q \mathbf{1}_n)$, and let $\bar{\lambda}_c(\tau)$ be defined as $\bar{\lambda}(\tau)$ with a replaced by a_c. Corollaries 27 and 26 and Proposition 46 in the Supplementary Materials show that Theorems 1–3 continue to hold. The set $\mathcal{M}_c(\lambda,\tau)$ equals $\{(\alpha_Q, X^\top a_c)\}$, the minimum-norm interpolating pair, exactly when $0 < \lambda \le \bar{\lambda}_c(\tau)$, and we then call the level extinct. Two distinct levels share a fitted pair exactly when both are extinct.

These certificates use only the realized feature matrix, and hence they also apply to features learned from the responses, provided that the quantile heads are optimal given those features. Let $T \subset (0,1)$ be finite, and consider a model with trunk parameters ϑ, features $\phi_\vartheta(z) \in \mathbb{R}^p$ for a covariate z, and one linear head $(\alpha_\tau, w_\tau) \in \mathbb{R} \times \mathbb{R}^p$ for each $\tau \in T$, trained with a penalty $\lambda > 0$ by minimizing

$$\sum_{\tau\in T} \left[\sum_{i=1}^n \rho_\tau\{y_i - \alpha_\tau - \phi_\vartheta(z_i)^\top w_\tau\} + \frac{\lambda}{2}\,\|w_\tau\|_2^2 \right] + \Omega(\vartheta), \tag{1}$$

where Ω is any real-valued function of the trunk, for instance a weight penalty. When the loss of each level is averaged over the n observations, weight decay γ on the head weights corresponds to $\lambda = n\gamma$.

Corollary 2. *Let $\{\vartheta^*, (\alpha^*_\tau, w^*_\tau)_{\tau\in T}\}$ be a local minimizer of* (1)*. Suppose that the matrix X_{ϑ^*} with rows $\phi_{\vartheta^*}(z_i)^\top$ has rank n, and compute α_Q, a_c, and $\bar\lambda_c$ from $X = X_{\vartheta^*}$ and y. Then $(\alpha^*_\tau, w^*_\tau) = (\alpha_Q, X^\top_{\vartheta^*} a_c)$ if and only if $\lambda \le \bar\lambda_c(\tau)$. For distinct $\tau, \tau' \in T$,*

$$(\alpha^*_\tau, w^*_\tau) = (\alpha^*_{\tau'}, w^*_{\tau'}) \qquad \Longleftrightarrow \qquad \lambda \le \bar\lambda_c(\tau) \wedge \bar\lambda_c(\tau').$$

Beyond full row rank of the learned feature matrix, Corollary 2 places no condition on the trunk, its penalty, or how the features were learned. Once training reaches such a local minimum, the heads of all levels whose boundaries are at least λ are identical. The same conclusion holds for heads refitted exactly on a fixed trunk, and the boundary can be checked after training with one linear solve on the learned features.

3 The extinction boundary under random features

The fixed-design results locate the boundary after the data are observed. Under Gaussian random features, its shape and scale are determined asymptotically before the data are observed, which identifies the penalties that make the fit loss-blind. Assume that X has independent $N(0, 1/p)$ entries, that $\varepsilon \sim N(0, I_n)$ is independent of X, and that

$$y = X\theta + \sigma\varepsilon, \qquad \sigma > 0, \qquad \theta \in \mathbb{R}^p. \tag{2}$$

The coefficient vector θ is nonrandom and may depend on n and p without a bound on its norm, and $s_{n,p} = (\sigma^2 + \|\theta\|_2^2/p)^{1/2}$ is the marginal response standard deviation. The row normalization gives $\mathbb{E}\,\|x_i\|_2^2 = 1$, and X has rank n almost surely when $p \ge n$. We set $\tau^\dagger = 1/2$ on the event $A_+A_- = 0$, whose probability tends to zero.

Theorem 4. *Assume model* (2)*. If $n \to \infty$ and $n/p \to 0$, then*

$$\sup_{\tau\in(0,1)} \left| s_{n,p}\sqrt{2\log n}\,\bar\lambda(\tau) - \{\tau \wedge (1-\tau)\} \right| \longrightarrow 0$$

in probability. In addition, in probability,

$$\tau^\dagger \longrightarrow \frac{1}{2}, \qquad s_{n,p}\sqrt{2\log n}\,\bar\lambda(\tau^\dagger) \longrightarrow \frac{1}{2}.$$

The triangular limit holds because A_+ and A_- are both asymptotic to $s_{n,p}\sqrt{2\log n}$: when $p \gg n$, the interpolation vector is close to the response vector coordinatewise, and the extremes of n Gaussian responses have this size. Rotational invariance of the design gives this coordinatewise control for every signal norm. Let $E_\lambda = \{\tau \in (0,1) : \lambda \le \bar\lambda(\tau)\}$ be the set of extinct levels at the penalty λ, and fix $0 < \tau_L \le \tau_U < 1$.

Corollary 3. *Assume model* (2) *with $n \to \infty$ and $n/p \to 0$, and let $\lambda_n > 0$ be deterministic.*

- *(i) If $\lambda_n s_{n,p}\sqrt{\log n} \to 0$, then $\mathbb{P}\{[\tau_L, \tau_U] \subseteq E_{\lambda_n}\} \to 1$.*
- *(ii) If $\lambda_n s_{n,p}\sqrt{\log n} \to \infty$, then $\mathbb{P}\{E_{\lambda_n} = \emptyset\} \to 1$.*
- *(iii) If $\lambda_n = c/\{s_{n,p}\sqrt{2\log n}\}$ for a constant $c > 0$, then, for $c < 1/2$, the set E_{λ_n} is nonempty with probability tending to one and its endpoints converge in probability to c and $1-c$; for $c > 1/2$, $\mathbb{P}\{E_{\lambda_n} = \emptyset\} \to 1$.*

Corollary 3 identifies $\{s_{n,p}\sqrt{\log n}\}^{-1}$ as the critical penalty for the summed loss. In the averaged-loss parameterization, the critical weight decay is of order $\{n s_{n,p}\sqrt{\log n}\}^{-1}$. For standardized responses and unit-norm features, the median becomes extinct for weight decays below about 1.6×10^{-3} at $n = 100$ and 1.3×10^{-4} at $n = 1000$ by the limit in Theorem 4; the levels 0.1 and 0.9 become extinct below one fifth of these values.

Without an intercept, model (2) gives every fitted coefficient vector coverage exactly one half at a new observation, because the test residual is centered Gaussian. Practical fits estimate an unpenalized intercept, which we now include. Let $\mu_0 \in \mathbb{R}$ and consider

$$y = \mu_0 \mathbf{1}_n + X\theta + \sigma\varepsilon, \tag{3}$$

with X, θ, σ, and ε as in (2), fitted with an unpenalized intercept. Let (X_*, Y_*) be an independent test pair with $X_* \sim N(0, I_p/p)$ and $Y_* = \mu_0 + X_*^\top \theta + \sigma\varepsilon_*$, where $\varepsilon_* \sim N(0,1)$ is independent of X_*. For a pair (α, β) fitted from the training sample, its marginal coverage is $C(\alpha, \beta) = \mathbb{P}_*(Y_* \leq \alpha + X_*^\top \beta)$, where $\mathbb{P}_*$ is the probability over the test pair given the training sample. The following theorem shows that the common fit of the extinct levels has coverage close to one half, and it locates the penalties at which the levels are extinct.

Theorem 5. *Assume model* (3) *with* $n \to \infty$ *and* $n/p \to 0$*, let* $0 < \tau_L \leq \tau_U < 1$ *be fixed, and let* $\lambda_n > 0$ *be deterministic. Then, in probability,*

$$\sup_{\tau \in (0,1)} \left| s_{n,p}\sqrt{2\log n}\,\bar{\lambda}_c(\tau) - \{\tau \wedge (1-\tau)\} \right| \longrightarrow 0, \qquad C(\alpha_Q, X^\top a_c) \longrightarrow \frac{1}{2}.$$

(i) If $\lambda_n s_{n,p}\sqrt{\log n} \to 0$*, then with probability tending to one every* $\tau \in [\tau_L, \tau_U]$ *is extinct at* λ_n*, and on this event* $C\{\widehat{\alpha}_c(\lambda_n, \tau), \widehat{\beta}_c(\lambda_n, \tau)\} = C(\alpha_Q, X^\top a_c)$ *for every such* τ*.*

(ii) If $\lambda_n s_{n,p}\sqrt{\log n} \to \infty$*, then with probability tending to one no level in* $(0,1)$ *is extinct at* λ_n*.*

Theorem 5 shows that the extinct fit is a median-type predictor. At every penalty below its boundary, each level has the common fit $(\alpha_Q, X^\top a_c)$, whose coverage converges to one half. At penalties below the boundaries of both 0.1 and 0.9, the interval between these nominal quantiles has zero width. Bai et al. (2021) found coverage near one half for the minimum-norm interpolant of unregularized quantile regression, with the intercept inside the norm. The theorem gives the analogous conclusion for ridge fits with an unpenalized intercept at every positive penalty below the boundary, and the boundary of each level marks the penalty at which its coverage can begin to move toward the nominal value. As the penalty grows further, the fit approaches intercept-only quantile regression, whose marginal coverage is close to τ; Section 5 shows this transition.

For non-Gaussian features, the boundary scale can also depend on how the signal is spread across features. Independent sub-Gaussian features give the same limit when the signal variance is negligible (Remark 1) or when no single coefficient dominates the response extremes (Theorem 10). For independent sign-valued features with k_n equal nonzero coefficients and fixed signal variance, the normalizing scale moves from the noise standard deviation to the response standard deviation as $k_n/\log n$ increases (Theorem 11). Appendix A gives finite-sample bounds (Section A.16), and the Supplementary Materials give the scale at general ratios n/p (Section S.1), second-order fluctuations (Section S.1.3), and a penalized constant feature (Corollary 6).

4 Accuracy above the boundary

Below the boundary the quantile curves collapse. Under the following conditions, validation over penalties above the boundary controls the excess pinball risk and the calibration error. Let $\mathcal{Z}$ be a covariate space and $\psi : \mathcal{Z} \to \mathbb{R}^p$ a feature map fixed before sampling with $\|\psi(Z)\|_2 \leq \kappa$ for the random covariate $Z \in \mathcal{Z}$. Consider $Y = \mu(Z) + \sigma U$ with $U \sim \mathrm{Unif}[-1,1]$ independent of Z, $\mu(z) = \mathbb{E}(Y \mid Z = z)$, $\sigma > 0$, and $|Y| \leq M$. Let $(Z_i, Y_i)_{i=1}^n$ be independent copies of (Z, Y), put $x_i = \psi(Z_i)$ and $y_i = Y_i$, and assume that X has rank n almost surely. For L target levels $\tau_\ell \in (0,1)$, assume that the conditional quantiles $q_\ell(z) = \mu(z) + \sigma(2\tau_\ell - 1)$ equal $\psi(z)^\top \beta_\ell^*$ with $\|\beta_\ell^*\|_2 \leq B$. Write $\mathbb{E}_*$ and $\mathbb{P}_*$ for expectation and probability over an independent test observation, and for $b \in \mathbb{R}^p$ put $R_\ell(b) = \mathbb{E}_*\rho_{\tau_\ell}\{Y - \psi(Z)^\top b\}$ and $C_\ell(b) = |\mathbb{P}_*\{Y \leq \psi(Z)^\top b\} - \tau_\ell|$. For $\delta \in (0,1)$, define

$$t_n = \sqrt{\frac{2\log(4L/\delta)}{n}}, \qquad e_n = 4\kappa B(4/\sqrt{n} + t_n), \qquad \lambda_n = \frac{n e_n}{2B^2}, \qquad r_{\min} = \sigma \min_{\ell \leq L} \tau_\ell(1-\tau_\ell).$$

For each level, let $\Lambda_\ell \subset [\lambda_n, \infty)$ be a finite grid constructed from the training data that contains λ_n, with at most N candidates over all levels, and let $\widehat{\lambda}_\ell$ minimize the pinball loss on m further independent validation copies of (Z, Y).

Theorem 6. *Assume the preceding conditions and* $e_n < r_{\min}(1 - t_n)$*. Then, with probability at least* $1 - \delta$*, simultaneously for* $\ell \leq L$*,*

$$\lambda_n > \bar{\lambda}(\tau_\ell), \qquad R_\ell\{\widehat{\beta}(\widehat{\lambda}_\ell, \tau_\ell)\} - R_\ell(\beta_\ell^*) \leq D_{n,m}, \qquad C_\ell\{\widehat{\beta}(\widehat{\lambda}_\ell, \tau_\ell)\} \leq \sqrt{D_{n,m}/\sigma},$$

where $D_{n,m} = 5e_n/4 + 2(M + 3\kappa B)\{\log(4N/\delta)/(2m)\}^{1/2}$.

Theorem 6 bounds the excess risk by the risk at the penalty λ_n and the cost of choosing among N candidates with m validation observations. For fixed constants the bound is $O(n^{-1/2} + \sqrt{\log N/m})$, and the squared calibration error is at most $D_{n,m}/\sigma$. The theorem uses standard complexity and validation bounds (Bartlett & Mendelson, 2002), and its sample-size condition is conservative. Its role here is to show that, with probability at least $1 - \delta$, the penalty λ_n at which the bound holds lies strictly above the boundaries of the target levels, because a norm-bounded interpolant cannot have small risk under noise. A search that starts at the boundary thus loses no candidate needed for the bound. Theorem 12 removes the coefficient bound from the search. Over a fixed dyadic grid screened at the observed boundary, with predictions clipped to a known response bound, it gives the same conclusion up to the approximation error of the features for bounded, possibly heteroskedastic responses. A bounded conditional density then also bounds the calibration error. Neither result asserts a faster rate than existing analyses of kernel quantile regression (Steinwart & Christmann, 2011; Wang & Feng, 2024).

5 Numerical studies

We ask three questions. Does the coverage of ridge-penalized quantile regression change at the boundary as Theorem 5 predicts? Where does the boundary lie, relative to the weight decay used in training, for random and learned features on real data? Does the boundary predict when the quantile heads of trained networks coincide, and which weight decays give distinct and calibrated heads? Every study uses the levels $0.1, 0.25, 0.5, 0.75, 0.9$, an unpenalized intercept, exact extinct fits from Theorem 1 and its intercept version, and a dual solver that certifies every convex fit by its primal–dual gap; Section S.29 gives the designs and the numerical certificates.

Coverage below and above the boundary. We simulate model (3) with $\mu_0 = 1$, $\sigma = 1$, $\|\theta\|_2^2/p = 1$, and $p = n^2$ for $n \in \{100, 300, 1000\}$, using 120 independent repetitions for each n. For each repetition, level, and normalized penalty $c = \lambda s_{n,p}\sqrt{2\log n}$ on a grid from 0.02 to 100, we compute the exact marginal coverage of the fitted pair. Figure 1 agrees with the coverage limit, part (i), and the boundary limit of Theorem 5. For $c \leq 0.1$, each level is extinct in at least 104 of the 120 repetitions for every n, and the mean coverage of all five levels lies between 0.500 and 0.503. The coverage of the common extinct fit concentrates at one half as n grows: its standard deviation over repetitions is 0.042, 0.024, and 0.013 for $n = 100$, 300, and 1000. Each level keeps coverage near one half while it is extinct and moves toward its nominal value only after the penalty exceeds its boundary. At $c = 100$ and $n = 1000$, the mean coverages are 0.104, 0.254, 0.501, 0.749, and 0.899. The mean normalized boundary approaches $\tau \wedge (1 - \tau)$ slowly, from 0.125 to 0.114 at level 0.1 and from 0.563 to 0.541 at the median as n increases from 100 to 1000. Tables S.18 and S.19 give the complete results.

Learned and random features on real data. We use four public regression data sets and one simulated heteroskedastic model: diabetes (Efron et al., 2004) and California housing from scikit-learn (Pedregosa et al., 2011), and puma8NH and house_16H from OpenML. In each of 100 random splits per data set, we standardize the covariates and responses with the $n = 128$ training observations, keep 128 validation observations and up to 1000 test observations, and fit two feature maps. The first is a Gaussian random Fourier feature map with 2048 features and the median-distance bandwidth. The second is the last hidden layer of a two-hidden-layer ReLU network of width 256 with five linear quantile heads, trained end to end by 2000 full-batch Adam steps on the sum over levels of the averaged pinball losses plus $(\gamma/2)$ times the squared norm of the trunk parameters and head weights, with $\gamma = 10^{-4}$ and unpenalized head intercepts. This coupled weight decay matches (1) with $\lambda = n\gamma$. We compare three sets of heads on the same trained features: the trained heads, exact refits of each head at the training weight decay, and exact refits at the weight decay selected for each level on the validation sample from the grid $10^{-6}, 10^{-5.5}, \ldots, 10^{1}$. Boundaries are reported as $\bar{\lambda}_c(\tau)/n$, on the scale of γ. Pooled values average over the 500 splits of the five data sets, and mean pinball losses also average over the five levels, on the standardized response scale. Table 1 and Figure 2 give three findings. First, with training weight decay 10^{-4}, the median boundary of the learned features at the level 0.5 is 90 to 324 times the training weight decay, whereas that of random features ranges from 2.2×10^{-7} to 1.8×10^{-5}. The exact refits of all five heads at the training weight decay

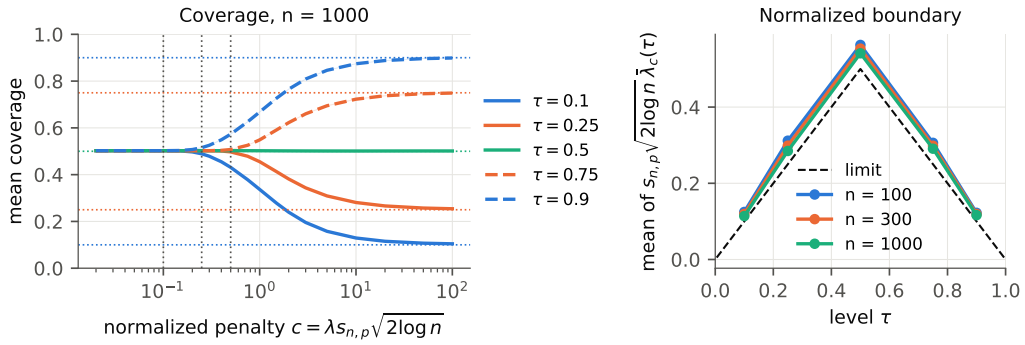

Figure 1: Mean coverage against the normalized penalty for $n = 1000$, with dotted lines at the limiting boundaries $\tau \wedge (1-\tau)$ (left), and mean normalized boundary against the level (right), under Gaussian random features with an intercept.

Table 1: Median boundaries $\bar{\lambda}_c(0.5)/n$, percentage of splits in which the exact refits of all five heads at the training weight decay are extinct, and mean test coverage of the trained heads and of the refits with validated weight decay.

	Boundary $\lambda_c(0.5)/n$		Extinct	Trained heads		Refit, validated γ	
Data	learned	random	at 10^{-4}	$\tau = 0.1$	$\tau = 0.9$	$\tau = 0.1$	$\tau = 0.9$
Simulated	3.2×10^{-2}	1.8×10^{-5}	100%	0.483	0.512	0.194	0.767
California	9.0×10^{-3}	2.2×10^{-7}	99%	0.505	0.538	0.163	0.822
puma8NH	2.4×10^{-2}	2.9×10^{-6}	100%	0.482	0.502	0.157	0.842
house_16H	1.5×10^{-2}	2.0×10^{-6}	100%	0.501	0.524	0.152	0.847
Diabetes	1.7×10^{-2}	8.9×10^{-7}	100%	0.496	0.518	0.163	0.848

are extinct in 499 of the 500 splits. Second, the trained heads nearly coincide, as Corollary 2 predicts for an exact local minimum. Adam does not reach an exact minimizer: the median over splits of the largest absolute training residual is between 0.070 and 0.118 on the standardized scale, whereas the extinct refits interpolate. Yet the mean spread of the five trained predictions at a test point is at most 0.076, their mean test coverage lies between 0.482 and 0.538 for every data set and level, and their pooled mean pinball loss, 0.287, exceeds that of the training quantiles of the response, 0.279. Third, the refits with validated weight decay move the pooled coverage at the levels 0.1 and 0.9 from 0.493 and 0.519 to 0.166 and 0.825 and lower the mean pinball loss to 0.202. Linear quantile regression has pooled coverages 0.141 and 0.863 and mean pinball loss 0.214, and random features with a validated weight decay have mean pinball loss 0.191.

A weight decay above the boundary is necessary but not sufficient for calibrated heads. Exact refits on the learned features keep mean coverage at the levels 0.1 and 0.9 between 0.491 and 0.520 up to 30 times the boundary of each level and reach 0.236 and 0.758 only between 100 and 300 times the boundary (Table S.21), whereas in the simulation of Figure 1, the coverage at the level 0.1 is 0.196 at 26 times the mean boundary. Raising the training weight decay to 10^{-2} separates the refitted heads but does not calibrate them. In this second arm, run on the simulated model, California housing, and puma8NH, the median boundary of the learned features falls to between 1.6×10^{-6} and 1.8×10^{-4}, below the training weight decay, and no refit at the training weight decay is extinct, yet the trained heads and these refits keep mean coverage between 0.481 and 0.546 at every level. For the refits, the search that fits the extinct candidates once selects the same weight decay as the full search in every split, feature map, and level. It avoids 55% of the solver fits on the learned features but only 8.4% of the solver time, because the solver time is concentrated at the penalties above the boundary.

Sections S.20.2–S.28 of the Supplementary Materials report additional studies of validation with the extinct candidates represented by one fit, reinvestment of saved fits under arithmetic and elapsed-time budgets, screening under nonuniform noise, Gaussian boundary scaling, the predictive cost of collapse, boundary diagnostics, and comparisons across aspect ratios and repeated real-data splits.

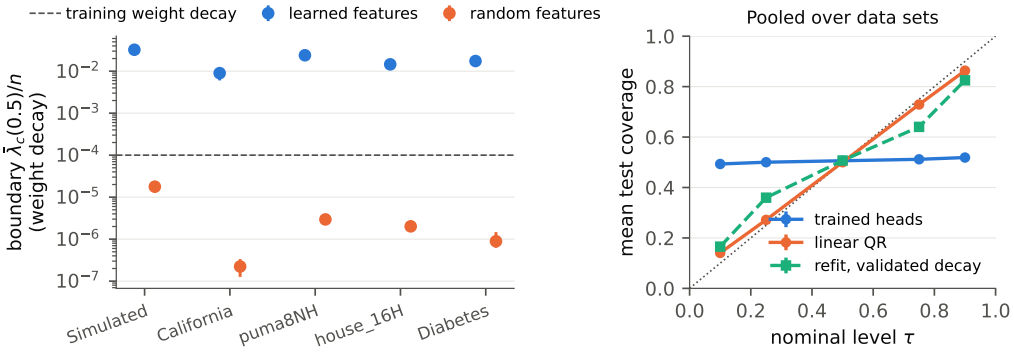

Figure 2: Medians and quartiles of the boundary $\bar{\lambda}_c(0.5)/n$ over splits, with the training weight decay as a dashed line (left), and pooled mean test coverage, with bars of two Monte Carlo standard errors (right). QR denotes quantile regression.

6 Discussion

A ridge penalty makes the penalized coefficients unique, but uniqueness does not separate quantile levels. For features whose training matrix has full row rank, Theorems 1–3 give a certificate, computed with one linear solve, of when the levels share the minimum-norm interpolant, and Corollary 2 applies it to the heads of trained networks. Theorem 5 shows what this costs: under Gaussian random features with $n/p \to 0$, every extinct level has coverage converging to one half. Under these features, the boundary is a weight decay of order $\{ns_{n,p}\sqrt{\log n}\}^{-1}$. On four regression data sets and a simulated model, the weight decay 10^{-4} used to train a network lies below the boundaries of its learned features in 499 of 500 splits, and the trained quantile heads nearly coincide.

These findings suggest three uses. Before a penalty search, the boundary of each level separates the penalties that give the common extinct fit, which is computed once, from the penalties that give fits specific to that level. A penalty shared by all requested levels leaves every level nonextinct exactly when it exceeds the largest boundary (Proposition 3). For exact head refits after training, the boundary computed from the learned features determines whether the heads are distinct. Distinct heads need not be calibrated: in our experiments, exact refits on the learned features kept coverage near one half up to about 30 times the boundary and approached the nominal levels only at weight decays more than 100 times the boundary. When a model must interpolate, separating the levels requires penalties above the boundary. Conformal calibration (Romano et al., 2019) gives marginal interval coverage under exchangeability, even from collapsed base fits, but does not restore their level-dependent shapes.

The results have limits. The limits in Section 3 assume independent Gaussian features and Gaussian noise independent of the design with $n/p \to 0$. The Supplementary Materials treat general ratios n/p, and the non-Gaussian results show that the scale can depend on how the signal is spread across features. Corollary 2 holds at local minima and for heads refitted on a fixed trunk, and end-to-end training approaches this regime only approximately. Our network experiments use one architecture, 128 training observations, and full-batch Adam, and other architectures, sample sizes, and optimizers may place the boundary differently. Early stopping and other implicit regularization prevent interpolation by different means and are outside our analysis. The accuracy guarantees above the boundary are conservative, and how to choose penalties above the boundary that target calibration remains open.

AI use statement

Generative AI tools were used to assist numerical-code development and review, related-literature searches, and prose editing. The authors take responsibility for the final text.

Ethics statement

This paper studies a mathematical property of penalized estimators. It uses simulated data and public regression benchmarks distributed with scikit-learn (diabetes and California housing) and OpenML (puma8NH and house_16H). It collects no new human-participant data and makes no clinical decision claim.

Reproducibility statement

Appendix A gives complete proofs of the main results. The Supplementary Materials give the supporting deterministic results, numerical settings, and additional artifact checks. The numerical sections of the Supplementary Materials specify the scripts, seeds, and optimality certificates, and the manifest `results/summary/numerical_manifest.json` in the code archive identifies every recorded artifact. The supplementary code archive contains the scripts, recorded outputs, generated table and figure sources, and validators used for the reported results.

References

Yu Bai, Song Mei, Huan Wang, and Caiming Xiong. Understanding the under-coverage bias in uncertainty estimation. In *Advances in Neural Information Processing Systems 34*, pp. 18307–18319, 2021.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002. URL https://jmlr.org/papers/v3/bartlett02a.html.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

Alex J. Cannon. Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Computers & Geosciences*, 37(9):1277–1284, 2011. doi: 10.1016/j.cageo.2010.07.005.

Alex J. Cannon. Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic Environmental Research and Risk Assessment*, 32(11):3207–3225, 2018. doi: 10.1007/s00477-018-1573-6.

Victor Chernozhukov, Iván Fernández-Val, and Alfred Galichon. Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125, 2010. doi: 10.3982/ECTA7880.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004. doi: 10.1214/009053604000000067.

Michael P. Friedlander and Paul Tseng. Exact regularization of convex programs. *SIAM Journal on Optimization*, 18(4):1326–1350, 2008. doi: 10.1137/060675320.

Michael P. Friedlander, Sharvaj Kubal, Yaniv Plan, and Matthew S. Scott. Average-case thresholds for exact regularization of linear programs, 2025. URL https://arxiv.org/abs/2510.13083. arXiv:2510.13083.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

Jan Gasthaus, Konstantinos Benidis, Yuyang Wang, Syama Sundar Rangapuram, David Salinas, Valentin Flunkert, and Tim Januschowski. Probabilistic forecasting with spline quantile function RNNs. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1901–1910. PMLR, 2019.

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Correcting the coverage bias of quantile regression, 2025. URL https://arxiv.org/abs/2511.00820. arXiv:2511.00820, version 2.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022. doi: 10.1214/21-AOS2133.

Xuming He. Quantile curves without crossing. *The American Statistician*, 51(2):186–192, 1997. doi: 10.1080/00031305.1997.10473959.

Roger Koenker. *Quantile Regression*. Cambridge University Press, Cambridge, 2005.

Roger Koenker and Gilbert Bassett, Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. doi: 10.2307/1913643.

Roger Koenker and Kevin F. Hallock. Quantile regression. *Journal of Economic Perspectives*, 15 (4):143–156, 2001. doi: 10.1257/jep.15.4.143.

Kuo-Wei Lai and Vidya Muthukumar. General loss functions lead to (Approximate) interpolation in high dimensions. *Journal of Machine Learning Research*, 26(244):1–72, 2025.

Youjuan Li, Yufeng Liu, and Ji Zhu. Quantile regression in reproducing kernel Hilbert spaces. *Journal of the American Statistical Association*, 102(477):255–268, 2007. doi: 10.1198/016214506000000979.

Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69, 2021.

Taman Narayan, Serena Lutong Wang, Kevin Robert Canini, and Maya R. Gupta. Expected pinball loss for quantile regression and inverse CDF estimation. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=Eg8Rnb0Hdd.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

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

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems 32*, pp. 3543–3553, 2019.

Saharon Rosset. Bi-level path following for cross validated solution of kernel quantile regression. *Journal of Machine Learning Research*, 10(86):2473–2505, 2009.

Saharon Rosset and Ji Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, 35 (3):1012–1030, 2007. doi: 10.1214/009053606000001370.

Vidyashankar Sivakumar and Arindam Banerjee. High-dimensional structured quantile regression. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3220–3229. PMLR, 2017.

Ingo Steinwart and Andreas Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 2011. doi: 10.3150/10-BEJ267.

Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning. In *Advances in Neural Information Processing Systems 32*, pp. 6417–6428, 2019.

Ichiro Takeuchi, Quoc V. Le, Timothy D. Sears, and Alexander J. Smola. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7:1231–1264, 2006.

Ichiro Takeuchi, Kaname Nomura, and Takafumi Kanamori. Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression. *Neural Computation*, 21(2):533–559, 2009. doi: 10.1162/neco.2008.10-07-628.

Qian Tang, Yuwen Gu, and Boxiang Wang. fastkqr: A fast algorithm for kernel quantile regression. *Journal of Computational and Graphical Statistics*, 35(1):395–405, 2026. doi: 10.1080/10618600.2025.2541004.

James W. Taylor. A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, 19(4):299–311, 2000.

Shanshan Tu, Yunzhang Zhu, Yoonkyung Lee, Qiuyu Gu, and Haozhen Yu. Cross validation for penalized quantile regression with a case-weight adjusted solution path, 2024. URL `https://arxiv.org/abs/1902.07770v2`. arXiv:1902.07770v2.

Caixing Wang and Xingdong Feng. Optimal kernel quantile learning with random features. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50419–50452. PMLR, 2024.

A PROOFS

The proofs establish the fixed-design identities, the Gaussian boundary limits, and the population risk guarantee before the auxiliary extensions.

A.1 PROOF OF PROPOSITION 1

For any $r \in \mathbb{R}$,

$$\rho_\tau(r) = \sup_{g \in [\tau-1, \tau]} gr.$$

Thus the primal objective is

$$\sup_{g \in \mathcal{B}_\tau} \left\{ g^\top (y - X\beta) + \frac{\lambda}{2} \|\beta\|_2^2 \right\}.$$

The objective is closed, proper, and strongly convex for $\lambda > 0$. Fenchel duality gives

$$\max_{g \in \mathcal{B}_\tau} \left\{ g^\top y - \frac{1}{2\lambda} g^\top K g \right\}.$$

Because $K \succ 0$, completing the square gives the unique minimizer of $(g - \lambda a)^\top K (g - \lambda a)$ over $\mathcal{B}_\tau$. The first-order condition in β gives

$$\lambda \widehat{\beta}(\lambda, \tau) = X^\top g_{\lambda,\tau}.$$

Substitution gives the residual identity

$$y - X\widehat{\beta}(\lambda, \tau) = y - \lambda^{-1} X X^\top g_{\lambda,\tau} = K(a - \lambda^{-1} g_{\lambda,\tau}).$$

This completes the proof.

A.2 PROOF OF THEOREM 1

By Proposition 1 and injectivity of $X^\top$, $\widehat{\beta}(\lambda, \tau) = \widehat{\beta}_0$ is equivalent to $g_{\lambda,\tau} = \lambda a$. The projection identity makes this equivalent to $\lambda a \in \mathcal{B}_\tau$. Equivalently,

$$\tau - 1 \le \lambda a_i \le \tau \qquad \text{for every } i \le n.$$

For $a_i > 0$, this condition is $\lambda \le \tau / a_i$. For $a_i < 0$, it is $\lambda \le (1 - \tau)/|a_i|$. For $a_i = 0$, it imposes no restriction. Taking the minimum gives the claimed boundary.

For $a \neq 0$, the boundary is positive and finite. For $a = 0$, $y = 0$, all caps are infinite, and $\widehat{\beta}_0 = 0$ minimizes the objective for every $\lambda > 0$. This completes the proof.

A.3 PROOF OF THEOREM 2

Empty inner minima below are interpreted as $+\infty$. The definition of $\bar{\lambda}$ gives

$$\bar{\lambda}(\tau) = \min \left\{ \min_{i: a_i > 0} \frac{\tau}{a_i}, \min_{i: a_i < 0} \frac{1 - \tau}{|a_i|} \right\} = \frac{\tau}{A_+} \wedge \frac{1 - \tau}{A_-},$$

with the stated convention for a zero denominator. If $A_+ > 0$ and $A_- > 0$, the minimum is concave and piecewise linear, with one increasing and one decreasing term. They meet when

$$\frac{\tau}{A_+} = \frac{1 - \tau}{A_-},$$

which gives

$$\tau^\dagger = \frac{A_+}{A_+ + A_-}$$

and value $(A_+ + A_-)^{-1}$.

If $A_+ = A_- = 0$, then $a = 0$ and $\bar{\lambda}(\tau) = +\infty$ at every level. Suppose $A_+ + A_- > 0$. For fixed $0 < \lambda \le (A_+ + A_-)^{-1}$, the inequality $\lambda \le \bar{\lambda}(\tau)$ is equivalent to

$$\tau \ge \lambda A_+, \qquad \tau \le 1 - \lambda A_-.$$

This gives the claimed interval. For $A_- = 0 < A_+$, the second constraint is absent, giving $[\lambda A_+, 1) \cap (0, 1)$. The case $A_+ = 0 < A_-$ is symmetric. This completes the proof.

A.4 Proof of Corollary 1

Because $\lambda \le \bar{\lambda}(\tau)$ for every $\tau \in \mathcal{T}$, Theorem 1 gives

$$\widehat{\beta}(\lambda, \tau) = \widehat{\beta}_0 \qquad (\tau \in \mathcal{T}).$$

Multiplying by any covariate vector x gives the fitted-value identity. This completes the proof.

A.5 Population consequences of level collapse

Because a single common predictor can only yield one marginal coverage probability, it is mathematically impossible for it to be exactly calibrated across multiple distinct target levels. This limitation requires no moment assumptions. Under moment conditions, the following corollary quantifies the resulting errors in squared quantile distance and, for two levels, gives an excess pinball risk bound under additional conditional quantile and density lower-bound conditions. Let (X_*, Y_*) be a new test observation independent of the training sample, and let $\mathbb{E}_*$ and $\mathbb{P}_*$ denote expectation and probability over this new observation conditional on the training sample. For levels $0 < \tau_1 < \cdots < \tau_L < 1$, let q_k be a conditional τ_k-quantile of Y_* given X_*, and put $\bar{q} = L^{-1} \sum_{k=1}^{L} q_k$. Define

$$\mathcal{R}_k(b) = \mathbb{E}_* \left[\{X_*^\top b - q_k(X_*)\}^2 \right], \qquad \mathcal{C}_k(b) = \left| \mathbb{P}_* \{Y_* \le X_*^\top b\} - \tau_k \right|,$$

$$\mathcal{P}_k(b) = \mathbb{E}_* \left[\rho_{\tau_k}(Y_* - X_*^\top b) - \rho_{\tau_k}\{Y_* - q_k(X_*)\} \right].$$

Write F_x for the conditional distribution function of Y_* given $X_* = x$, and write f_x for its density when it exists.

Corollary 4. *Suppose $0 < \lambda \le \min_{k \le L} \bar{\lambda}(\tau_k)$. If $\mathbb{E}_* \|X_*\|_2^2 < \infty$ and each $q_k(X_*)$ is square-integrable, then*

$$\frac{1}{L} \sum_{k=1}^{L} \mathcal{R}_k\{\widehat{\beta}(\lambda, \tau_k)\} = \mathbb{E}_* \left[\{X_*^\top \widehat{\beta}_0 - \bar{q}(X_*)\}^2 \right] + \frac{1}{L^2} \sum_{1 \le k < \ell \le L} \mathbb{E}_* \left[\{q_k(X_*) - q_\ell(X_*)\}^2 \right]$$

and

$$\max_{k \le L} \mathcal{R}_k\{\widehat{\beta}(\lambda, \tau_k)\} \ge \frac{1}{L^2} \sum_{1 \le k < \ell \le L} \mathbb{E}_* \left[\{q_k(X_*) - q_\ell(X_*)\}^2 \right],$$

whereas, without moment conditions,

$$\max_{k \le L} \mathcal{C}_k\{\widehat{\beta}(\lambda, \tau_k)\} \ge \frac{\tau_L - \tau_1}{2}.$$

The mean calibration error also satisfies

$$\frac{1}{L} \sum_{k=1}^{L} \mathcal{C}_k\{\widehat{\beta}(\lambda, \tau_k)\} \ge \frac{1}{L} \sum_{k=1}^{\lfloor L/2 \rfloor} (\tau_{L+1-k} - \tau_k). \tag{4}$$

If $L = 2$, $\mathbb{E}_* \|X_*\|_2 + \sum_{k=1}^{2} \mathbb{E}_* |q_k(X_*)| < \infty$, and, for almost every x,

$$F_x\{q_k(x)\} = \tau_k \quad (k = 1, 2), \qquad f_x(t) \ge \underline{f} > 0 \quad \{q_1(x) \le t \le q_2(x)\},$$

then

$$\sum_{k=1}^{2} \mathcal{P}_k\{\widehat{\beta}(\lambda, \tau_k)\} \ge \frac{\underline{f}}{4} \mathbb{E}_* \left[\{q_2(X_*) - q_1(X_*)\}^2 \right].$$

Corollary 4 translates the mathematical event of coefficient coincidence into explicit statistical consequences. Because the quantile-error lower bounds are strictly positive under the stated moment conditions when the true target quantiles differ with positive probability, extinction prevents the simultaneous recovery of multiple quantile levels. The bound in (4) is sharp over all common coverage probabilities, with equality when the common coverage is a median of the target levels, and it applies directly to the mean absolute empirical calibration error when the levels are evaluated on any shared set of test observations.

Proof. Corollary 1 gives $\widehat{\beta}(\lambda, \tau_k) = \widehat{\beta}_0$ for every k. Put $h = X_*^\top \widehat{\beta}_0$. The pointwise identities

$$\frac{1}{L}\sum_{k=1}^{L}\{h - q_k(X_*)\}^2 = \{h - \bar{q}(X_*)\}^2 + \frac{1}{L}\sum_{k=1}^{L}\{q_k(X_*) - \bar{q}(X_*)\}^2,$$

$$\frac{1}{L}\sum_{k=1}^{L}\{q_k(X_*) - \bar{q}(X_*)\}^2 = \frac{1}{L^2}\sum_{1\le k<\ell\le L}\{q_k(X_*) - q_\ell(X_*)\}^2$$

give the risk identity after taking $\mathbb{E}_*$. The maximum is at least the average, giving the first inequality.

Put $c_* = \mathbb{P}_*(Y_* \le h)$. The common coefficient gives $\mathcal{C}_k\{\widehat{\beta}(\lambda, \tau_k)\} = |c_* - \tau_k|$. Hence

$$\max_{k\le L}|c_* - \tau_k| \ge \frac{\tau_L - \tau_1}{2},$$

which proves the calibration bound.

For $k \le \lfloor L/2 \rfloor$, the triangle inequality gives

$$|c_* - \tau_k| + |c_* - \tau_{L+1-k}| \ge \tau_{L+1-k} - \tau_k.$$

Summing proves (4). Equality holds when c_* is a median of the requested levels, proving sharpness over $c_* \in [0, 1]$. Replacing c_* by the common empirical coverage proves the test-sample assertion.

For the last conclusion, let $q_\tau = q_k$ when $\tau = \tau_k$, and define $\mathcal{L}_{\tau,x}(t) = \mathbb{E}_*[\rho_\tau(Y_* - t) - \rho_\tau\{Y_* - q_\tau(x)\} \mid X_* = x]$. The integrand is bounded by $|t - q_\tau(x)|$ in absolute value and is integrable under the first-moment conditions. The pinball-loss identity gives

$$\mathcal{L}_{\tau,x}(t) - \mathcal{L}_{\tau,x}\{q_\tau(x)\} = \int_{q_\tau(x)}^{t}\{F_x(s) - \tau\}\, ds.$$

Let $\widetilde{g}$ be the projection of $g = x^\top \widehat{\beta}_0$ onto $[q_1(x), q_2(x)]$. Projection cannot increase either excess risk, hence

$$\begin{aligned}
&\sum_{k=1}^{2}[\mathcal{L}_{\tau_k,x}(g) - \mathcal{L}_{\tau_k,x}\{q_k(x)\}] \\
&\quad\ge \int_{q_1(x)}^{\widetilde{g}}\{F_x(s) - \tau_1\}\, ds + \int_{\widetilde{g}}^{q_2(x)}\{\tau_2 - F_x(s)\}\, ds \\
&\quad\ge \frac{f}{2}\left[\{\widetilde{g} - q_1(x)\}^2 + \{q_2(x) - \widetilde{g}\}^2\right] \ge \frac{f}{4}\{q_2(x) - q_1(x)\}^2.
\end{aligned}$$

Taking expectation over X_* proves the last bound. This completes the proof. □

A.6 Proof of Theorem 3

By Proposition 1 and injectivity of $X^\top$, coefficient equality is equivalent to

$$g_{\lambda,\tau} = g_{\lambda,\tau'}.$$

Suppose this equality holds, write $\bar{g}$ for the common vector, and put $v = K(\lambda a - \bar{g})$. The projection variational inequality for $\mathcal{B}_\tau$ gives

$$v^\top(h - \bar{g}) \le 0 \qquad (h \in \mathcal{B}_\tau).$$

Feasible coordinate increases give $v_i \le 0$ when $\bar{g}_i < \tau$; feasible decreases give $v_i \ge 0$ when $\bar{g}_i > \tau - 1$. Thus $v_i > 0$ requires $\bar{g}_i = \tau$, and $v_i < 0$ requires $\bar{g}_i = \tau - 1$. The same implications hold with τ' in place of τ. Because $\tau \ne \tau'$, both signs of v_i are excluded. Hence $v = 0$, and nonsingularity of K gives $\bar{g} = \lambda a$ and

$$\lambda a \in \mathcal{B}_\tau \cap \mathcal{B}_{\tau'}.$$

Theorem 1 converts the two inclusions into $0 < \lambda \le \bar{\lambda}(\tau) \wedge \bar{\lambda}(\tau')$ and gives $\widehat{\beta}(\lambda, \tau) = \widehat{\beta}_0 = \widehat{\beta}(\lambda, \tau')$.

The reverse implication follows from the same theorem. This completes the proof.

A.7 Proof of Corollary 2

Fix $\vartheta = \vartheta^*$ and $\tau \in T$. For $(\alpha, w) \in \mathbb{R} \times \mathbb{R}^p$, put

$$h_\tau(\alpha, w) = \sum_{i=1}^{n} \rho_\tau\{y_i - \alpha - \phi_{\vartheta^*}(z_i)^\top w\} + \frac{\lambda}{2}\|w\|_2^2.$$

Changing only the head of level τ changes the objective (1) by the change in h_τ. Because $\Omega(\vartheta^*)$ is finite and the given point is a local minimizer of (1), the pair $(\alpha_\tau^*, w_\tau^*)$ is a local minimizer of h_τ. The function h_τ is convex, and every local minimizer of a convex function is a global minimizer. Hence $(\alpha_\tau^*, w_\tau^*) \in \mathcal{M}_c(\lambda, \tau)$, where $\mathcal{M}_c(\lambda, \tau)$ is the set of minimizers defined before Corollary 27 with $X = X_{\vartheta^*}$.

Apply Section S.30 with the matrix Z of unpenalized columns equal to $\mathbf{1}_n$, for which $P_Z^Q y = \alpha_Q \mathbf{1}_n$, $a_Z = a_c$, and $\bar{\lambda}_Z = \bar{\lambda}_c$. Theorem 19 then gives the equivalence of $\lambda a_c \in [\tau - 1, \tau]^n$ and $\lambda \le \bar{\lambda}_c(\tau)$. If $\lambda \le \bar{\lambda}_c(\tau)$, Corollary 27 gives $\mathcal{M}_c(\lambda, \tau) = \{(\alpha_Q, X_{\vartheta^*}^\top a_c)\}$, and hence $(\alpha_\tau^*, w_\tau^*) = (\alpha_Q, X_{\vartheta^*}^\top a_c)$. For the reverse implication, if $(\alpha_\tau^*, w_\tau^*) = (\alpha_Q, X_{\vartheta^*}^\top a_c)$, the same corollary gives $\lambda a_c \in [\tau - 1, \tau]^n$, which is equivalent to $\lambda \le \bar{\lambda}_c(\tau)$.

Let $\tau, \tau' \in T$ be distinct. If $\lambda \le \bar{\lambda}_c(\tau) \wedge \bar{\lambda}_c(\tau')$, the first part gives equal heads. If the heads are equal, their common value lies in $\mathcal{M}_c(\lambda, \tau) \cap \mathcal{M}_c(\lambda, \tau')$, and Proposition 46 gives $\lambda \le \bar{\lambda}_c(\tau) \wedge \bar{\lambda}_c(\tau')$. This completes the proof.

The second equivalence concerns the whole head. Above the boundary, two levels can have the same weight vector and different intercepts. For the feature matrix I_2, $y = (1, 0)^\top$, and $\lambda = 1$, we have $\alpha_Q = 1/2$, $a_c = (1/2, -1/2)^\top$, and $\bar{\lambda}_c(1/4) = \bar{\lambda}_c(3/4) = 1/2 < \lambda$. The optimality conditions in the proof of Proposition 46 show that the unique minimizers at the levels $1/4$ and $3/4$ are $\{1/4, (1/4, -1/4)^\top\}$ and $\{3/4, (1/4, -1/4)^\top\}$, with subgradient vectors $(1/4, -1/4)^\top$ at both levels.

A.8 Statement and proof of Proposition 2

Fix $\tau \in (0, 1)$ with $\lambda_0 = \bar{\lambda}(\tau) < \infty$, let i_* be an index attaining this minimum, and put $q_{i_* i_*} = e_{i_*}^\top Q e_{i_*}$. The following proposition describes this departure.

Proposition 2. *Assume that i_* is the unique minimizing index. Then there exists $\eta > 0$ such that, for $\lambda_0 < \lambda < \lambda_0 + \eta$,*

$$\widehat{\beta}(\lambda, \tau) = \widehat{\beta}_0 - \frac{(\lambda - \lambda_0) a_{i_*}}{\lambda q_{i_* i_*}} X^\top Q e_{i_*}.$$

Also,

$$y - X\widehat{\beta}(\lambda, \tau) = \frac{(\lambda - \lambda_0) a_{i_*}}{\lambda q_{i_* i_*}} e_{i_*}.$$

Proof. Put $j = i_*$ and $\bar{g}_j = \lambda_0 a_j$. Then $q_{jj} > 0$, and $\lambda_0 a_k \in (\tau - 1, \tau)$ for $k \ne j$. For $\lambda > \lambda_0$ near λ_0, define

$$t(\lambda) = \frac{a_j - \bar{g}_j/\lambda}{q_{jj}}, \qquad \beta(\lambda) = \widehat{\beta}_0 - t(\lambda) X^\top Q e_j.$$

Then

$$y - X\beta(\lambda) = t(\lambda) e_j.$$

For $a_j > 0$, $\bar{g}_j = \tau$ and $t(\lambda) > 0$; for $a_j < 0$, $\bar{g}_j = \tau - 1$ and $t(\lambda) < 0$. In both cases, $\bar{g}_j$ is the jth check-loss subgradient.

Set

$$g(\lambda) = \lambda Q\{y - t(\lambda) e_j\}.$$

Then $g_j(\lambda) = \bar{g}_j$, while continuity gives $g_k(\lambda) \in [\tau - 1, \tau]$ for $k \ne j$ in a sufficiently small neighborhood. Thus $g(\lambda)$ is an admissible subgradient vector, and

$$\lambda\beta(\lambda) = \lambda X^\top Q\{y - t(\lambda) e_j\} = X^\top g(\lambda).$$

Optimality gives $\beta(\lambda) = \widehat{\beta}(\lambda, \tau)$. Because

$$t(\lambda) = \frac{(\lambda - \lambda_0)a_j}{\lambda q_{jj}},$$

the two identities in the proposition follow. This completes the proof. □

Under (2), $a \mid X$ has a Gaussian density with covariance $\sigma^2 Q^2 \succ 0$. A tie between finite coordinate caps at a fixed level lies on one of finitely many hyperplanes in a, each of probability zero. Unique binding thus holds almost surely on any fixed finite level grid, but need not hold at $\tau^\dagger$, where the two branches meet.

A.9 A COMMON BOUNDARY FOR A SET OF LEVELS

For any nonempty set of target levels $T \subset (0, 1)$, define $\Lambda_n(T) = \sup_{\tau \in T} \bar{\lambda}(\tau)$. The following proposition shows that a common penalty avoids extinction at every level in T exactly when it exceeds $\Lambda_n(T)$.

Proposition 3. *Assume $T \subset (0, 1)$ is nonempty and compact. If $\lambda > \Lambda_n(T)$, then*

$$\widehat{\beta}(\lambda, \tau) \neq \widehat{\beta}_0 \qquad \textit{for every } \tau \in T.$$

If $0 < \lambda \leq \Lambda_n(T)$, then at least one level in T is extinct.

Proof. For $a = 0$, every level is extinct and $\Lambda_n(T) = +\infty$. For $a \neq 0$, continuity on compact T gives a maximizer τ_* of $\bar{\lambda}$. If $\lambda > \Lambda_n(T)$, every level is nonextinct by Theorem 1. If $\lambda \leq \Lambda_n(T)$, the same theorem gives extinction at τ_*. This completes the proof. □

A.10 PROOF OF THEOREM 4

Put $\rho_n = \sqrt{n/p}$ and $L_n = \sqrt{2 \log n}$, and let Z denote a standard Gaussian variable with density ϕ. All limits are taken along $n \to \infty$ with $\rho_n \to 0$. For a matrix A, $s_{\min}(A)$ and $s_{\max}(A)$ denote its smallest and largest singular values, and $\|A\|_{\mathrm{op}} = s_{\max}(A)$ is its Euclidean operator norm. For matrices, we also write $\|A\|_2 = \|A\|_{\mathrm{op}}$. For symmetric A, $\lambda_{\min}(A)$ denotes its smallest eigenvalue.

Step 1: Gaussian maxima. The coordinates y_i are independent $N(0, s_{n,p}^2)$ variables. Put $M_+ = \max_{i \leq n} y_i$ and $M_- = \max_{i \leq n}(-y_i)$. We claim that

$$\frac{M_\pm}{s_{n,p} L_n} \longrightarrow 1 \qquad \text{in probability.}$$

Fix $\delta \in (0, 1)$. The Gaussian tail bound $\mathbb{P}(Z > t) \leq e^{-t^2/2}$ for $t \geq 0$ and a union bound give

$$\mathbb{P}\{M_+ > (1 + \delta) s_{n,p} L_n\} \leq n\, e^{-(1+\delta)^2 \log n} = n^{-2\delta - \delta^2} \longrightarrow 0.$$

For the lower bound, put $t_n = (1 - \delta) L_n$ and take n large enough that $t_n \geq \sqrt{2}$. The lower tail bound $\mathbb{P}(Z > t) \geq \phi(t)(t^{-1} - t^{-3}) \geq \phi(t)/(2t)$ for $t \geq \sqrt{2}$ (Vershynin, 2018, Proposition 2.1.2) and independence give

$$\mathbb{P}\{M_+ < (1 - \delta) s_{n,p} L_n\} = \{1 - \mathbb{P}(Z > t_n)\}^n \leq \exp\left\{-\frac{n\, \phi(t_n)}{2 t_n}\right\} = \exp\left\{-\frac{n^{\delta(2-\delta)}}{2\sqrt{2\pi}\, t_n}\right\} \longrightarrow 0,$$

because $\delta(2 - \delta) > 0$ and $t_n = O(\sqrt{\log n})$. Symmetry gives the same bounds for M_-.

Step 2: coordinatewise control of the interpolation vector. By rotational invariance of the Gaussian design, we may take $\theta = \|\theta\|_2 e_1$, where e_1 is the first coordinate vector of $\mathbb{R}^p$. Write

$$X = p^{-1/2}[z, W_-], \qquad B = p^{-1} W_- W_-^\top, \qquad K = B + p^{-1} z z^\top.$$

Here $W_- \in \mathbb{R}^{n \times (p-1)}$, and $z \sim N(0, I_n)$ is independent of B. Apply Davidson & Szarek (2001, Theorem 2.13) to W_- and $[z, W_-]$ with deviation $\sqrt{n}$. For a constant C, with probability tending to one,

$$\left\|K^{-1} - I_n\right\|_{\mathrm{op}} \leq C\rho_n, \qquad \left\|B^{-1} - I_n\right\|_{\mathrm{op}} \leq C\rho_n, \qquad \left\|B^{-1}\right\|_{\mathrm{op}} \leq 2.$$

Here B is eventually invertible because $n/p \to 0$.

Put $u_n = \sqrt{2\log(2n^2)}$. Conditional on B and X, respectively, the coordinates of $(B^{-1} - I_n)z$ and $(K^{-1} - I_n)\varepsilon$ are centered Gaussian. The preceding spectral bounds give standard deviations at most $C\rho_n$. Gaussian tails and a union bound give the first three bounds with probability tending to one:

$$\left\|(B^{-1} - I_n)z\right\|_\infty \le C\rho_n u_n, \qquad \left\|(K^{-1} - I_n)\varepsilon\right\|_\infty \le C\rho_n u_n,$$
$$\|z\|_\infty \le u_n, \qquad \|z\|_2^2 \le 2n.$$

For the last inequality, the Gaussian moment generating function gives

$$\mathbb{P}\{\|z\|_2^2 > 2n\} \le \exp\{-n/2\}\mathbb{E}\exp\{\|z\|_2^2/4\} = \exp\{-(1-\log 2)n/2\}.$$

Put $h_n = p^{-1}z^\top B^{-1}z$. The Sherman–Morrison identity gives

$$K^{-1}z = \frac{B^{-1}z}{1+h_n}, \qquad K^{-1}z - z = \frac{(B^{-1} - I_n)z - h_n z}{1+h_n}.$$

On the preceding event, $0 \le h_n \le 4n/p = 4\rho_n^2$. Hence

$$\left\|K^{-1}z - z\right\|_\infty \le C\rho_n u_n + 4\rho_n^2 u_n.$$

Because

$$a - y = \frac{\|\theta\|_2}{\sqrt{p}}(K^{-1}z - z) + \sigma(K^{-1} - I_n)\varepsilon,$$

putting $b_n = \|\theta\|_2/\sqrt{p}$ gives

$$\|a - y\|_\infty \le C(b_n + \sigma)\rho_n u_n + 4b_n\rho_n^2 u_n =: \kappa_n, \qquad \frac{\kappa_n}{s_{n,p}L_n} \longrightarrow 0.$$

Step 3: boundary convergence. Steps 1 and 2 give $M_+ \wedge M_- > \kappa_n$ with probability tending to one. On this event, $A_+ = \max_{i\le n} a_i > 0$, $A_- = \max_{i\le n}(-a_i) > 0$, and

$$|A_+ - M_+| \vee |A_- - M_-| \le \|a - y\|_\infty \le \kappa_n.$$

Because $b_n/s_{n,p} \le 1$ and $(b_n + \sigma)/s_{n,p} \le \sqrt{2}$, Steps 1 and 2 give

$$\frac{A_+}{s_{n,p}L_n} \longrightarrow 1, \qquad \frac{A_-}{s_{n,p}L_n} \longrightarrow 1 \qquad \text{in probability,}$$

and $A_+A_- > 0$ with probability tending to one. Put $r_+ = s_{n,p}L_n/A_+$ and $r_- = s_{n,p}L_n/A_-$ on this event. Theorem 2 gives

$$s_{n,p}L_n\bar{\lambda}(\tau) = (r_+\tau) \wedge \{r_-(1-\tau)\}.$$

For every $\tau \in (0,1)$,

$$|(r_+\tau) \wedge \{r_-(1-\tau)\} - \{\tau \wedge (1-\tau)\}| \le |r_+ - 1| \vee |r_- - 1|,$$

because $|u \wedge v - u' \wedge v'| \le |u - u'| \vee |v - v'|$ and $\tau \vee (1-\tau) \le 1$. The right side converges to zero uniformly in τ. The identities $\tau^\dagger = A_+/(A_+ + A_-)$ and $\bar{\lambda}(\tau^\dagger) = (A_+ + A_-)^{-1}$ from Theorem 2 give the remaining limits. This completes the proof.

A.11 Proof of Corollary 3

Put $L_n = \sqrt{2\log n}$. The proof of Theorem 4 gives $A_\pm/(s_{n,p}L_n) \to 1$ in probability and $\mathbb{P}(A_+A_- > 0) \to 1$. On $A_+A_- > 0$, Theorem 2 makes $\lambda \le \bar{\lambda}(\tau)$ equivalent to

$$\lambda A_+ \le \tau, \qquad \lambda A_- \le 1 - \tau.$$

Thus $E_\lambda = \{\tau \in (0,1) : \lambda A_+ \le \tau \le 1 - \lambda A_-\}$, which is empty exactly when $\lambda(A_+ + A_-) > 1$.

(i) If $\lambda_n s_{n,p}\sqrt{\log n} \to 0$, the ratio limits give $\lambda_n A_\pm \to 0$. Hence $\lambda_n A_+ \le \tau_L$ and $\lambda_n A_- \le 1 - \tau_U$ with probability tending to one.

(ii) If $\lambda_n s_{n,p}\sqrt{\log n} \to \infty$, then $\lambda_n(A_+ + A_-) \to \infty$ in probability, giving $E_{\lambda_n} = \emptyset$ with probability tending to one.

(iii) For $\lambda_n = c/(s_{n,p}L_n)$,

$$\lambda_n A_+ = c\,\frac{A_+}{s_{n,p}L_n} \longrightarrow c, \qquad 1 - \lambda_n A_- \longrightarrow 1 - c, \qquad \lambda_n(A_+ + A_-) \longrightarrow 2c$$

in probability. For $c < 1/2$, E_{λ_n} is nonempty with probability tending to one, and its endpoints converge to c and $1 - c$. For $c > 1/2$, $\lambda_n(A_+ + A_-) > 1$ with probability tending to one, giving $E_{\lambda_n} = \emptyset$. This completes the proof.

A.12 Proof of Theorem 5

Put $\rho_n = \sqrt{n/p}$, $L_n = \sqrt{2\log n}$, $u_n = \sqrt{2\log(2n^2)}$, $s = s_{n,p}$, and $v = X\theta + \sigma\varepsilon$, which gives $y = \mu_0 \mathbf{1}_n + v$. The pair (X, v) follows model (2). Hence Steps 1 and 2 of the proof of Theorem 4 apply to v: for a constant C, with probability tending to one,

$$\|Q - I_n\|_{\mathrm{op}} \le C\rho_n, \qquad \|Qv - v\|_\infty \le \kappa_n, \tag{5}$$

where κ_n is defined in that proof and satisfies $\kappa_n/(sL_n) \to 0$, and $\max_{i\le n}(\pm v_i)/(sL_n) \to 1$ in probability. Because $v \sim N(0, s^2 I_n)$, the moment generating function bound in Step 2 of that proof, applied to v/s, gives $\|v\|_2^2 \le 2ns^2$ with probability at least $1 - \exp\{-(1-\log 2)n/2\}$. Write $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ for the smallest and largest eigenvalues of a symmetric matrix. On the first event in (5), for all large n, $\lambda_{\min}(K) \ge 1/2$, $\lambda_{\max}(K) \le 2$, and $\|Q\|_{\mathrm{op}} \le 2$.

Step 1: coordinatewise control of $Q\mathbf{1}_n$. Fix i. Let X_{-i} be X without row i and put $K_{-i} = X_{-i}X_{-i}^\top$. For a vector of ones $\mathbf{1}_{n-1}$, the block inverse of K gives

$$(Q\mathbf{1}_n)_i = Q_{ii}\left(1 - x_i^\top w_i\right), \qquad w_i = X_{-i}^\top K_{-i}^{-1}\mathbf{1}_{n-1}, \qquad \|w_i\|_2^2 = \mathbf{1}_{n-1}^\top K_{-i}^{-1}\mathbf{1}_{n-1}.$$

The vector w_i depends only on X_{-i}, and $x_i \sim N(0, I_p/p)$ is independent of X_{-i}. On the event $G_i = \{\lambda_{\min}(K_{-i}) \ge 1/2\}$, which is determined by X_{-i}, the conditional distribution of $x_i^\top w_i$ given X_{-i} is centered Gaussian with variance at most $2n/p = 2\rho_n^2$. Hence

$$\mathbb{P}\left(G_i \cap \{|x_i^\top w_i| > \sqrt{2}\rho_n u_n\}\right) \le 2e^{-u_n^2/2} = n^{-2},$$

and a union bound over $i \le n$ shows that, with probability at least $1 - n^{-1}$, every i with G_i satisfies $|x_i^\top w_i| \le \sqrt{2}\rho_n u_n$. Eigenvalue interlacing gives $\lambda_{\min}(K_{-i}) \ge \lambda_{\min}(K) \ge 1/2$ for every i on the first event in (5). Hence, with probability tending to one,

$$\|Q\mathbf{1}_n\|_\infty \le \max_{i\le n} Q_{ii}\left(1 + |x_i^\top w_i|\right) \le 2\left(1 + \sqrt{2}\rho_n u_n\right). \tag{6}$$

Step 2: the intercept. Because $y = \mu_0\mathbf{1}_n + v$,

$$\alpha_Q - \mu_0 = \frac{\mathbf{1}_n^\top Q v}{\mathbf{1}_n^\top Q \mathbf{1}_n}.$$

On the first event in (5), $\mathbf{1}_n^\top Q\mathbf{1}_n \ge n/\lambda_{\max}(K) \ge n/2$, and

$$|\mathbf{1}_n^\top Q v| \le |\mathbf{1}_n^\top v| + \|(Q - I_n)\mathbf{1}_n\|_2 \, \|v\|_2 \le |\mathbf{1}_n^\top v| + C\rho_n\sqrt{n}\, \|v\|_2 .$$

The coordinates of v are independent $N(0, s^2)$ variables. Hence $\mathbf{1}_n^\top v \sim N(0, ns^2)$, and $|\mathbf{1}_n^\top v| \le s\sqrt{n}\, u_n$ with probability at least $1 - n^{-2}$. With the bound on $\|v\|_2^2$, this gives, with probability tending to one,

$$|\alpha_Q - \mu_0| \le 2s\left(\frac{u_n}{\sqrt{n}} + \sqrt{2}C\rho_n\right) =: s\eta_n, \qquad \eta_n \longrightarrow 0. \tag{7}$$

Step 3: boundary convergence. The definition of a_c and $y = \mu_0\mathbf{1}_n + v$ give

$$a_c = Q\{v - (\alpha_Q - \mu_0)\mathbf{1}_n\} = Qv - (\alpha_Q - \mu_0)Q\mathbf{1}_n.$$

By (5), (6), and (7), with probability tending to one,

$$\|a_c - v\|_\infty \le \kappa_n + 2s\eta_n\left(1 + \sqrt{2}\rho_n u_n\right).$$

Because $\eta_n \to 0$, $\rho_n \to 0$, and $u_n/L_n \to \sqrt{2}$, the right side divided by sL_n tends to zero. The maps $u \mapsto \max_i u_i^+$ and $u \mapsto \max_i(-u_i)^+$ are 1-Lipschitz in $\|\cdot\|_\infty$. Hence $A_{c,+} = \max_i a_{c,i}^+$ and $A_{c,-} = \max_i(-a_{c,i})^+$ differ from $\max_i v_i^+$ and $\max_i(-v_i)^+$ by at most $\|a_c - v\|_\infty$, and $A_{c,\pm}/(sL_n) \to 1$ in probability. With the identification $a_Z = a_c$ for $Z = \mathbf{1}_n$, Corollary 26 gives $\bar\lambda_c(\tau) = \tau/A_{c,+} \wedge (1-\tau)/A_{c,-}$, and the argument in Step 3 of the proof of Theorem 4 proves the uniform limit of the boundary.

Step 4: coverage of the extinct pair. Put $\widehat{\beta}_{c,0} = X^\top a_c$. Given the training sample, $Y_* - \alpha_Q - X_*^\top \widehat{\beta}_{c,0} = \mu_0 - \alpha_Q + X_*^\top(\theta - \widehat{\beta}_{c,0}) + \sigma\varepsilon_*$ is Gaussian with mean $\mu_0 - \alpha_Q$ and variance

$$\omega^2 = \sigma^2 + \frac{\left\|\theta - \widehat{\beta}_{c,0}\right\|_2^2}{p}.$$

Hence $C(\alpha_Q, \widehat{\beta}_{c,0}) = \Phi\{(\alpha_Q - \mu_0)/\omega\}$, where Φ is the standard normal distribution function, and

$$\left|\Phi\left(\frac{\alpha_Q - \mu_0}{\omega}\right) - \frac{1}{2}\right| \le \frac{|\alpha_Q - \mu_0|}{\sqrt{2\pi}\,\omega}.$$

Let $P_X = X^\top Q X$ be the orthogonal projection onto the row space of X. Because $\widehat{\beta}_{c,0}$ lies in this row space, $\left\|\theta - \widehat{\beta}_{c,0}\right\|_2 \ge \|(I_p - P_X)\theta\|_2$. For $\theta \neq 0$, the distribution of X is invariant under right multiplication by a fixed orthogonal matrix, and hence $\mathbb{E}\,\|P_X\theta\|_2^2 = \|\theta\|_2^2\,\mathbb{E}\,\mathrm{tr}(P_X)/p = \|\theta\|_2^2\, n/p$. Markov's inequality gives $\|P_X\theta\|_2^2 / \|\theta\|_2^2 \to 0$ in probability. Writing $o_{\mathbb{P}}(1)$ for a random quantity that converges to zero in probability, we obtain

$$\omega^2 \ge \sigma^2 + \frac{\|\theta\|_2^2}{p}\{1 - o_{\mathbb{P}}(1)\} \ge s^2\{1 - o_{\mathbb{P}}(1)\}.$$

The same bound holds trivially for $\theta = 0$. Combining it with (7) gives $|C(\alpha_Q, \widehat{\beta}_{c,0}) - 1/2| \le \eta_n\{1 + o_{\mathbb{P}}(1)\}/\sqrt{2\pi} \to 0$ in probability, which proves the second limit.

Step 5: the subcritical regime. Assume $\lambda_n s\sqrt{\log n} \to 0$. Then $\lambda_n A_{c,\pm} = \lambda_n s L_n \cdot A_{c,\pm}/(sL_n) \to 0$ in probability. Hence, with probability tending to one, $A_{c,+}A_{c,-} > 0$, $\lambda_n A_{c,+} \le \tau_L$, and $\lambda_n A_{c,-} \le 1 - \tau_U$. On this event, $\lambda_n \le \bar{\lambda}_c(\tau)$ for every $\tau \in [\tau_L, \tau_U]$, and Corollary 27 gives $\mathcal{M}_c(\lambda_n, \tau) = \{(\alpha_Q, \widehat{\beta}_{c,0})\}$ at every such level. Every selection $\{\widehat{\alpha}_c(\lambda_n, \tau), \widehat{\beta}_c(\lambda_n, \tau)\}$ then equals the extinct pair, which proves part (i).

Step 6: the supercritical regime. On the event $A_{c,+}A_{c,-} > 0$, whose probability tends to one by Step 3, the formula $\bar{\lambda}_c(\tau) = \tau/A_{c,+} \wedge (1-\tau)/A_{c,-}$ gives $\sup_\tau \bar{\lambda}_c(\tau) = (A_{c,+} + A_{c,-})^{-1}$. If $\lambda_n s\sqrt{\log n} \to \infty$, Step 3 gives $\lambda_n(A_{c,+} + A_{c,-}) \to \infty$ in probability, and hence $\lambda_n > \bar{\lambda}_c(\tau)$ for every τ with probability tending to one. This proves part (ii) and completes the proof.

A.13 Proof of Theorem 6

The proof of Theorem 6 establishes both nonextinction and a risk bound from the stated population model. The coefficient bound is a conclusion of empirical minimization.

Proof of Theorem 6. Let P_n denote the training empirical mean and P the population mean. Write $\ell_\ell(b) = \rho_{\tau_\ell}\{Y - \psi(Z)^\top b\}$ and $h_{\ell,b} = \ell_\ell(b) - \ell_\ell(\beta_\ell^*)$. For $\|b\|_2 \le 3B$, the coefficient difference has norm at most $4B$. Symmetrization and the contraction inequality for the 1-Lipschitz check loss give

$$\mathbb{E}\sup_{\|b\|_2 \le 3B} |(P_n - P)h_{\ell,b}| \le 4\mathbb{E}\sup_{\|v\|_2 \le 4B}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \psi(Z_i)^\top v\right| \le \frac{16\kappa B}{\sqrt{n}},$$

where the ϵ_i are independent uniform signs. The last step follows from $\mathbb{E}\,\|\sum_i \epsilon_i\psi(Z_i)\|_2 \le \kappa\sqrt{n}$. These are the symmetrization and Lipschitz-composition bounds in Bartlett & Mendelson (2002). Because $|h_{\ell,b}| \le 4\kappa B$, replacing one observation changes the supremum by at most $8\kappa B/n$. The bounded-differences inequality and a union bound imply

$$\sup_{\ell \le L,\, \|b\|_2 \le 3B} |(P_n - P)h_{\ell,b}| \le e_n \tag{8}$$

with probability at least $1 - \delta/4$.

Put $\gamma = \lambda/n$ and $J_{\ell,\gamma}(b) = P_n\ell_\ell(b) + \gamma\,\|b\|_2^2/2$. The conditional quantile minimizes population pinball risk, hence $Ph_{\ell,b} \ge 0$. For $\|b\|_2 = 3B$ and $\gamma \ge \lambda_n/n = e_n/(2B^2)$, (8) gives

$$J_{\ell,\gamma}(b) - J_{\ell,\gamma}(\beta_\ell^*) \ge -e_n + \frac{\gamma}{2}(9B^2 - B^2) \ge e_n > 0.$$

The objective is strictly convex and coercive. If its minimizer were outside the radius-$3B$ ball, the segment from β_ℓ^* to that minimizer would intersect the sphere at a point with objective at most $J_{\ell,\gamma}(\beta_\ell^*)$, a contradiction. Thus every $\widehat{\beta}(\lambda, \tau_\ell)$ with $\lambda \geq \lambda_n$ has norm below $3B$. Empirical optimality at λ_n and (8) now give

$$R_\ell\{\widehat{\beta}(\lambda_n, \tau_\ell)\} - R_\ell(\beta_\ell^*) \leq e_n + \frac{\lambda_n B^2}{2n} = \frac{5}{4} e_n. \tag{9}$$

Let $r_\ell = \sigma\tau_\ell(1-\tau_\ell)$. Under the uniform-noise model, $P\ell_\ell(\beta_\ell^*) = r_\ell$ and $0 \leq \ell_\ell(\beta_\ell^*) \leq 2r_\ell$. Hoeffding's inequality and a union bound give

$$P_n \ell_\ell(\beta_\ell^*) \geq r_\ell(1 - t_n) > e_n \qquad (\ell \leq L)$$

with probability at least $1 - \delta/4$. On the intersection with (8), an interpolating b with norm below $3B$ would satisfy

$$0 \leq P h_{\ell,b} \leq P_n h_{\ell,b} + e_n = -P_n \ell_\ell(\beta_\ell^*) + e_n < 0.$$

Thus $\widehat{\beta}(\lambda_n, \tau_\ell)$ cannot interpolate. Theorem 1 proves $\lambda_n > \bar{\lambda}(\tau_\ell)$ for all levels.

Condition on a training sample satisfying these conclusions. Write $\widehat{R}_\ell(\lambda)$ for the mean validation pinball loss of $\widehat{\beta}(\lambda, \tau_\ell)$. All candidate losses lie in $[0, H]$, where $H = M + 3\kappa B$. Hoeffding's inequality over the at most N candidates gives

$$\max_{\ell, \lambda \in \Lambda_\ell} |\widehat{R}_\ell(\lambda) - R_\ell\{\widehat{\beta}(\lambda, \tau_\ell)\}| \leq H\sqrt{\frac{\log(4N/\delta)}{2m}}$$

with conditional probability at least $1 - \delta/2$. Compare the validation minimizer with the included λ_n and use (9). This proves the risk bound with total probability at least $1 - \delta$.

For the calibration bound, fix z, put $q = q_\ell(z)$, and let F_z be the uniform conditional distribution function. Direct integration gives

$$\mathbb{E}\{\rho_{\tau_\ell}(Y - t) - \rho_{\tau_\ell}(Y - q) \mid Z = z\} = \int_q^t \{F_z(s) - \tau_\ell\}\, ds \geq \sigma\{F_z(t) - \tau_\ell\}^2.$$

Equality holds within $[\mu(z) - \sigma, \mu(z) + \sigma]$; beyond this interval the integral increases while the squared coverage difference stays constant. Apply this inequality at $t = \psi(z)^\top \widehat{\beta}$, integrate over Z, and use Cauchy–Schwarz. The squared marginal calibration error is at most the excess risk divided by σ. This completes the proof. □

For $\eta > 0$, a concrete class has covariate $Z = (\zeta_0, \ldots, \zeta_{p-3})$ with independent $\zeta_j \sim \mathrm{Unif}[-\pi, \pi]$, and features $(1, \cos\zeta_0, \sin\zeta_0)$ followed by $\eta\sqrt{2/(p-3)}\cos\zeta_j$, $j = 1, \ldots, p-3$. For fixed $b_1, b_2 \in \mathbb{R}$ and $\mu(Z) = b_1\cos\zeta_0 + b_2\sin\zeta_0$, one may take $\kappa = \sqrt{2 + 2\eta^2}$, $B = \sqrt{b_1^2 + b_2^2 + \sigma^2}$, and $M = \sqrt{b_1^2 + b_2^2} + \sigma$, which remain bounded as p increases, and the independent continuous feature coordinates give full row rank for $p \geq n + 3$.

For the concrete trigonometric class, $\|\psi(Z)\|_2^2 \leq 2 + 2\eta^2$. The coefficient vector of q_ℓ is $(\sigma(2\tau_\ell - 1), b_1, b_2, 0, \ldots, 0)^\top$, whose norm is at most B. For $p - 3 \geq n$, the first n independent nuisance columns form a square matrix with independent continuously distributed entries. Its determinant is a nonzero polynomial in those entries, hence it equals zero with probability zero. This proves the stated rank property. The feature map is specified before observations are drawn and does not use response data. The theorem does not impose the isotropic Gaussian scaling used for the extinction-boundary limit.

A.14 Sub-Gaussian features

For a random variable V, write $\|V\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}\exp(V^2/t^2) \leq 2\}$. On rank-deficient designs, we set $\bar{\lambda}(\tau) = 0$ and $\tau^\dagger = 1/2$; under the conditions below, the probability of rank deficiency tends to zero by the Gram-matrix bound in the proof.

Remark 1 (Sub-Gaussian features). Assume $n \to \infty$ and $n/p \to 0$. The entries of $\sqrt{p}\,X$ are independent, have mean zero and variance one, and satisfy $\sup_{n,p,i,j} \left\|\sqrt{p}\,X_{ij}\right\|_{\psi_2} \le L < \infty$. Assume also that $\|\theta\|_2 / \sqrt{p} \to 0$. If ε remains Gaussian and independent of X, then Theorem 4 and Corollary 3 hold with the same limits.

Remark 1 states that sub-Gaussian features give the Gaussian boundary limit when the total signal variance is asymptotically negligible, and the following argument proves it.

Put $W = \sqrt{p}\,X$, $\rho_n = \sqrt{n/p}$, and $L_n = \sqrt{2\log n}$. The signal condition in Remark 1 gives $s_{n,p}/\sigma \to 1$. The independent isotropic rows of $W^\top$ have uniformly bounded sub-Gaussian norms. By Vershynin (2018, Theorem 4.6.1), with probability at least $1-2\exp(-cn)$, for constants $C, c > 0$ depending only on L,

$$1 - C\rho_n \le s_{\min}(X) \le s_{\max}(X) \le 1 + C\rho_n.$$

Because $\rho_n \to 0$,

$$\left\|K^{-1} - I_n\right\|_{\mathrm{op}} \le C\rho_n, \qquad s_{\max}(X) \le 2$$

eventually on this event, with $\left\|K^{-1}\right\|_{\mathrm{op}} \le 2$ and full row rank.

Let X_{-i} omit row i and let P_{-i} project onto its row space, including on rank-deficient designs. On the full-rank event, block inversion gives

$$(K^{-1}X\theta)_i = \frac{x_i^\top (I_p - P_{-i})\theta}{x_i^\top (I_p - P_{-i})x_i}, \qquad x_i^\top (I_p - P_{-i})x_i = \{(K^{-1})_{ii}\}^{-1} \ge \frac{1}{2}.$$

Conditionally on X_{-i}, the numerator is a sum of independent centered sub-Gaussian variables. Because $\|(I_p - P_{-i})\theta\|_2 \le \|\theta\|_2$, Vershynin (2018, Theorem 2.6.3) and a union bound give

$$\mathbb{P}\left\{\max_{i\le n} |x_i^\top (I_p - P_{-i})\theta| > CL\,\|\theta\|_2 \sqrt{\frac{\log(2n^2)}{p}}\right\} \le \frac{1}{n}.$$

For $\theta = 0$ the inequality is immediate. Hence $\left\|K^{-1}X\theta\right\|_\infty / L_n \to 0$ in probability under $\|\theta\|_2 / \sqrt{p} \to 0$.

On the singular-value event, the coordinates of $(K^{-1} - I_n)\varepsilon$ are conditionally centered Gaussian with standard deviations at most $C\rho_n$. A union bound gives

$$\|a - \sigma\varepsilon\|_\infty \le 2CL\,\|\theta\|_2 \sqrt{\frac{\log(2n^2)}{p}} + C\sigma\rho_n\sqrt{2\log(2n^2)}$$

with probability tending to one. Dividing by L_n makes both terms tend to zero. The Gaussian maximum calculation in Step 1 of the proof of Theorem 4 now gives

$$\frac{A_+}{s_{n,p}L_n} \longrightarrow 1, \qquad \frac{A_-}{s_{n,p}L_n} \longrightarrow 1$$

in probability. Hence $A_+A_- > 0$ with probability tending to one, and Theorem 2 gives the boundary and peak limits. The proof of Corollary 3 applies because $s_{n,p} \to \sigma > 0$. This completes the proof.

A.15 A BOUNDARY TO THE SUB-GAUSSIAN EXTENSION

Signal variance tending to zero is sufficient for Remark 1. It cannot be replaced by a bounded signal variance without further distributional restrictions. The next result gives a bounded-coordinate example. Define the boundary to be zero on rank-deficient designs, as in Remark 1. In the following proposition, e_1 is the first coordinate vector of $\mathbb{R}^p$. For a random sequence V_n and a deterministic sequence $c_n > 0$, $V_n = O_{\mathbb{P}}(c_n)$ means that V_n/c_n is bounded in probability, and $V_n = o_{\mathbb{P}}(c_n)$ means that this ratio converges to zero in probability.

Proposition 4. *Let the entries of $\sqrt{p}\,X$ be independent symmetric signs, let $\varepsilon \sim N(0, I_n)$ be independent of X, and let $y = X\theta + \sigma\varepsilon$, where $\theta = r\sqrt{p}\,e_1$ and $r, \sigma > 0$ are fixed. If $n \to \infty$ and $n/p \to 0$, then, in probability,*

$$\sup_{\tau\in(0,1)} \left|\sigma\sqrt{2\log n}\,\bar{\lambda}(\tau) - \{\tau \wedge (1-\tau)\}\right| \longrightarrow 0.$$

Here $s_{n,p}^2 = \sigma^2 + r^2$. Thus scaling by the response standard deviation, as in the Gaussian theorem, does not give the same triangular limit. The limit depends on the signal distribution, despite convergence of the Gram matrix to the identity in operator norm.

Proof. Write $L_n = \sqrt{2\log n}$ and $\rho_n = \sqrt{n/p}$. The singular-value argument in Appendix A.14 gives $\left\|K^{-1}\right\|_{\mathrm{op}} \le 2$ and $\left\|K^{-1} - I_n\right\|_{\mathrm{op}} \le C\rho_n$ with probability tending to one. Use the row-deleted projections P_{-i} defined there. On $s_{\min}(X_{-i})^2 \ge 1/2$,

$$\|P_{-i}\theta\|_2^2 = (X_{-i}\theta)^\top (X_{-i}X_{-i}^\top)^{-1} X_{-i}\theta \le 2r^2(n-1),$$

because every coordinate of $X\theta$ equals r or $-r$. Conditioning on X_{-i}, sub-Gaussian concentration and a union bound give

$$\max_{i\le n} |x_i^\top P_{-i}\theta| = O_{\mathbb{P}}(r\rho_n L_n).$$

On the row-deleted event, each conditional tail at $Cr\sqrt{n\log(2n^2)/p}$ is at most n^{-2} for sufficiently large C. Interlacing gives all row-deleted events on the full singular-value event. Block inversion and $|x_i^\top\theta| = r$ imply

$$\left\|K^{-1}X\theta\right\|_\infty \le 2r + 2\max_{i\le n} |x_i^\top P_{-i}\theta| = o_{\mathbb{P}}(L_n).$$

The same conditional Gaussian bound as in Appendix A.14 gives $\left\|(K^{-1} - I_n)\varepsilon\right\|_\infty = O_{\mathbb{P}}(\rho_n L_n)$. Thus $\|a - \sigma\varepsilon\|_\infty / L_n \to 0$ in probability. The signed maxima of $a/(\sigma L_n)$ converge to one, and Theorem 2 gives the uniform limit. This completes the proof. □

A.16 Finite-sample Gaussian boundary bounds

The asymptotic triangular limit established in Theorem 4 relies on design concentration, response maxima, and coordinatewise control of the interpolation vector. To provide a fully explicit, finite-sample counterpart to this limit, the following result assumes the true signal is bounded by $\|\theta\|_2 \le R_0$ for some constant $R_0 \ge 0$. This boundedness assumption allows us to derive a closed-form bound by controlling the operator norm of $K^{-1} - I_n$.

Fix $\alpha \in (0,1)$, and let Φ denote the standard Gaussian distribution function. Define

$$t_\alpha = \sqrt{2\log(6/\alpha)}, \qquad \eta_\alpha = \frac{\sqrt{n} + t_\alpha}{\sqrt{p}}, \qquad D_\alpha = (1-\eta_\alpha)^{-2} - 1,$$

$$v_\alpha = \sqrt{2\log(6n/\alpha)}, \qquad q_\alpha^- = \Phi^{-1}\left\{(\alpha/12)^{1/n}\right\}, \qquad q_\alpha^+ = \Phi^{-1}\left\{(1-\alpha/12)^{1/n}\right\},$$

and

$$\kappa_\alpha = D_\alpha\{(1+\eta_\alpha)R_0 + \sigma v_\alpha\}, \qquad \underline{A}_\alpha = \sigma q_\alpha^- - \kappa_\alpha, \qquad \overline{A}_\alpha = \sqrt{\sigma^2 + R_0^2/p}\, q_\alpha^+ + \kappa_\alpha.$$

The following theorem gives confidence bounds for the realized positive and negative extremes and hence for the full extinction boundary.

Theorem 7. *Assume model* (2) *and* $\|\theta\|_2 \le R_0$. *If* $\eta_\alpha < 1$ *and* $\underline{A}_\alpha > 0$, *then, with probability at least* $1-\alpha$,

$$\underline{A}_\alpha \le A_+ \wedge A_- \le A_+ \vee A_- \le \overline{A}_\alpha$$

and, simultaneously for every $\tau \in (0,1)$,

$$\frac{\tau \wedge (1-\tau)}{\overline{A}_\alpha} \le \bar\lambda(\tau) \le \frac{\tau \wedge (1-\tau)}{\underline{A}_\alpha}.$$

Theorem 7 gives ranges of penalties that lie inside or strictly outside the extinct interval with probability at least $1-\alpha$. Penalties at or below the lower bound are extinct, while penalties strictly above the upper bound are nonextinct; the bounds need not classify penalties between them. This is a boundary-location result; it characterizes when the fitted coefficients collapse, but it does not, by itself, guarantee accurate calibration or population quantile recovery.

On the event in Theorem 7, let $E_\lambda = \{\tau \in (0,1) : \lambda \le \bar\lambda(\tau)\}$. An interval whose left endpoint exceeds its right endpoint is interpreted as empty. The following corollary bounds the location and height of the peak and the set of extinct levels.

Corollary 5. *Under the conditions of Theorem 7, with probability at least* $1-\alpha$*,*

$$\frac{\underline{A}_\alpha}{\underline{A}_\alpha+\overline{A}_\alpha} \le \tau^\dagger \le \frac{\overline{A}_\alpha}{\underline{A}_\alpha+\overline{A}_\alpha}, \qquad \left|\tau^\dagger - \frac{1}{2}\right| \le \frac{\overline{A}_\alpha-\underline{A}_\alpha}{2(\overline{A}_\alpha+\underline{A}_\alpha)},$$

$$\frac{1}{2\overline{A}_\alpha} \le \bar{\lambda}(\tau^\dagger) \le \frac{1}{2\underline{A}_\alpha},$$

and, for every $\lambda > 0$*,*

$$[\lambda\overline{A}_\alpha, 1-\lambda\overline{A}_\alpha] \cap (0,1) \subseteq E_\lambda \subseteq [\lambda\underline{A}_\alpha, 1-\lambda\underline{A}_\alpha] \cap (0,1).$$

In particular, for fixed $0 < \tau_L \le \tau_U < 1$*,*

$$\lambda \le \frac{\tau_L \wedge (1-\tau_U)}{\overline{A}_\alpha} \quad \Longrightarrow \quad [\tau_L, \tau_U] \subseteq E_\lambda,$$

whereas $\lambda > 1/(2\underline{A}_\alpha)$ *implies* $E_\lambda = \emptyset$*.*

Corollary 5 separates levels certified as extinct from levels certified as nonextinct without replacing the exact realized boundary. For the averaged empirical loss, every boundary and penalty in the corollary is divided by n, while $\tau^\dagger$ is unchanged.

Proof of Theorem 7. Write $X = W/\sqrt{p}$, where W has independent standard Gaussian entries. The Gaussian singular-value bound of Davidson & Szarek (2001, Theorem 2.13) gives

$$\mathbb{P}\left\{1-\eta_\alpha \le s_{\min}(X) \le s_{\max}(X) \le 1+\eta_\alpha\right\} \ge 1-\frac{\alpha}{3}.$$

On this event,

$$\left\|K^{-1}-I_n\right\|_{\mathrm{op}} \le D_\alpha, \qquad s_{\max}(X) \le 1+\eta_\alpha.$$

Because $a-y = (K^{-1}-I_n)X\theta + \sigma(K^{-1}-I_n)\varepsilon$,

$$\|a-y\|_\infty \le D_\alpha(1+\eta_\alpha)R_0 + \sigma \max_{i\le n} |e_i^\top (K^{-1}-I_n)\varepsilon|.$$

On this event, the coordinates of $(K^{-1}-I_n)\varepsilon$ are conditionally centered Gaussian with standard deviations at most D_α. A union bound gives

$$\mathbb{P}\left\{\max_{i\le n} |e_i^\top (K^{-1}-I_n)\varepsilon| > D_\alpha v_\alpha \;\middle|\; X\right\} \le 2n\exp(-v_\alpha^2/2) = \frac{\alpha}{3}.$$

Thus, outside an event of probability at most $2\alpha/3$,

$$\|a-y\|_\infty \le \kappa_\alpha.$$

Put $M_+ = \max_{i\le n} y_i$ and $M_- = \max_{i\le n}(-y_i)$. The coordinates of y are independent $N(0, s_{n,p}^2)$ variables. The definitions of q_α^- and q_α^+ give

$$\mathbb{P}\{M_+ < s_{n,p}q_\alpha^-\} = \mathbb{P}\{M_- < s_{n,p}q_\alpha^-\} = \frac{\alpha}{12}$$

and

$$\mathbb{P}\{M_+ > s_{n,p}q_\alpha^+\} = \mathbb{P}\{M_- > s_{n,p}q_\alpha^+\} = \frac{\alpha}{12}.$$

The four bounds hold jointly with probability at least $1-\alpha/3$. Intersecting with the preceding events and using $\sigma \le s_{n,p} \le \sqrt{\sigma^2 + R_0^2/p}$ and $\underline{A}_\alpha > 0$ gives

$$\underline{A}_\alpha \le A_+ \wedge A_- \le A_+ \vee A_- \le \overline{A}_\alpha.$$

Theorem 2 now yields, for every $\tau \in (0,1)$,

$$\bar{\lambda}(\tau) = \frac{\tau}{A_+} \wedge \frac{1-\tau}{A_-},$$

which gives the uniform bound. This completes the proof. □

Proof of Corollary 5. On the event in Theorem 7, Theorem 2 gives

$$\tau^\dagger = \frac{A_+}{A_+ + A_-}, \qquad \bar{\lambda}(\tau^\dagger) = \frac{1}{A_+ + A_-}.$$

The bounds on $A_\pm$ give both peak bounds. The inequalities

$$\lambda \overline{A}_\alpha \le \tau \le 1 - \lambda \overline{A}_\alpha$$

imply $\lambda A_+ \le \tau \le 1 - \lambda A_-$, giving the inner inclusion and the compact-level conclusion. For $\tau \in E_\lambda$, the lower bounds on $A_\pm$ give the outer inclusion. Its interval is empty when $\lambda > 1/(2\underline{A}_\alpha)$. This completes the proof. □

Supplementary Materials for "When Interpolation Is Loss-Blind: Quantile Regression at the Interpolation Boundary"

The main-result proofs precede these Supplementary Materials, which are organized in four groups. Sections S.1–S.3 extend Sections 3 and 4: Gaussian scales at general dimension ratios and their fluctuations, the effect of non-Gaussian signal concentration, and a prediction guarantee without a coefficient bound in the search. Sections S.4–S.19 give deterministic extensions of the extinction identities, including the prediction-space formulation, observation weights and quadratic penalties, convex residual losses, active-set formulas and certificates, stability, and joint or constrained fits; the results in the main manuscript use them only through the minimum-norm identity of Lemma 3. Sections S.20–S.29 give the validation results and the numerical studies, except the certified active-set audit, which follows the active-set formulas it checks. Section S.29 gives the designs, certificates, and complete results of the studies in Section 5. Sections S.30–S.34 treat unpenalized affine terms, which give the intercept results used in Sections 2 and 3, weighted affine terms, anisotropic features, arbitrary-rank designs, and boundary computation over a dual polytope.

S.1 Gaussian boundaries at general dimension ratios

Even when the number of features p is comparable to the sample size n, the Gaussian extinction boundary retains the same triangular limiting shape after normalization using the observed length of the interpolation vector. However, its overall scale deviates from the marginal response standard deviation. By decoupling the direction of the interpolation vector from its length, we establish a universal boundary limit for any dimension sequence $p \ge n$, including the square case $p = n$. When the excess dimension $p - n$ diverges, we also characterize this scale deterministically, even if the aspect ratio n/p approaches one.

Let X_{ij} be independent $N(0, 1/p)$ variables, let $\varepsilon \sim N(0, I_n)$ be independent of X, and let $y = X\theta + \sigma\varepsilon$, where $\theta \in \mathbb{R}^p$ is nonrandom and $\sigma > 0$ is fixed. No bound on $\|\theta\|_2$ is required in this section. Write $r_{n,p} = \|a\|_2 / \sqrt{n}$. For $c_n = n/p < 1$, define

$$v_{n,p}^2 = \frac{\|\theta\|_2^2 / p}{1 - c_n} + \frac{\sigma^2}{(1 - c_n)^3}.$$

Here $v_{n,p} > 0$ is the square root of this expression. The two terms separate the signal and noise contributions to the deterministic squared scale $v_{n,p}^2$. As $c_n \to 1$, the inverse Gram factor $(1-c_n)^{-3}$ of the noise variance grows faster than the factor $(1 - c_n)^{-1}$ of the signal variance.

Theorem 8. *For the Gaussian model specified above, if $n \to \infty$ and $p \ge n$, then, in probability,*

$$\sup_{\tau \in (0,1)} \left| r_{n,p} \sqrt{2 \log n}\, \bar{\lambda}(\tau) - \{\tau \wedge (1 - \tau)\} \right| \longrightarrow 0, \qquad \tau^\dagger \longrightarrow \frac{1}{2}.$$

If also $p - n \to \infty$, then, in probability,

$$\frac{r_{n,p}^2}{v_{n,p}^2} \longrightarrow 1, \qquad \sup_{\tau \in (0,1)} \left| v_{n,p} \sqrt{2 \log n}\, \bar{\lambda}(\tau) - \{\tau \wedge (1 - \tau)\} \right| \longrightarrow 0.$$

For either valid normalizer, the scaled supremum of the boundary over $(0, 1)$ converges to $1/2$.

The first conclusion of Theorem 8 describes the limiting shape of the boundary after multiplying it by $r_{n,p}\sqrt{2\log n}$, which uses the observed empirical length of a. This shape holds whether or not the length itself concentrates. The second conclusion provides a deterministic population scale that applies as long as the excess dimension $p-n$ diverges, even in the high-dimensional limit where $n/p \to 1$. When $n/p \to 0$, this generalized scale smoothly reduces to the $s_{n,p}$ scale derived in Theorem 4. However, for any strictly positive limiting ratio, the inverse Gram amplification factors persist. As we demonstrate later in Proposition 5, this length fails to concentrate around any deterministic scale when the signal is zero and $p-n$ stays fixed. The normalization using the observed length remains valid in this case.

Proof. For every deterministic orthogonal $n \times n$ matrix O, the pair $(OX, O\varepsilon)$ has the same distribution as (X, ε), and

$$\{(OX)(OX)^\top\}^{-1}(OX\theta + \sigma O\varepsilon) = O(XX^\top)^{-1}(X\theta + \sigma\varepsilon) = Oa.$$

Because $a \mid X$ is Gaussian with covariance $\sigma^2 Q^2 \succ 0$, $\mathbb{P}(a=0)=0$. Averaging its orthogonally invariant direction over the orthogonal group gives the uniform spherical distribution, independently of its length. Thus, for $Z \sim N(0, I_n)$ independent of $\|a\|_2$,

$$a \overset{d}{=} \|a\|_2 \, \frac{Z}{\|Z\|_2}.$$

The Gaussian tail bounds and independence give

$$\frac{\max_{i\le n} Z_i}{\sqrt{2\log n}} \longrightarrow 1, \qquad \frac{\max_{i\le n}(-Z_i)}{\sqrt{2\log n}} \longrightarrow 1, \qquad \frac{\|Z\|_2}{\sqrt{n}} \longrightarrow 1$$

in probability. For any $h \in (0,1)$ a union bound gives

$$\mathbb{P}\Big\{\max_i Z_i > (1+h)\sqrt{2\log n}\Big\} \le n\exp\{-(1+h)^2\log n\} \longrightarrow 0.$$

Writing Φ for the standard Gaussian distribution function, its lower tail bound gives

$$\mathbb{P}\Big\{\max_i Z_i \le (1-h)\sqrt{2\log n}\Big\} \le \exp\Big[-n\{1-\Phi((1-h)\sqrt{2\log n})\}\Big] \longrightarrow 0.$$

The same bounds apply to $-Z$, while $\mathrm{Var}(n^{-1}\sum_i Z_i^2) = 2/n$. Thus,

$$\frac{A_+}{r_{n,p}\sqrt{2\log n}} \longrightarrow 1, \qquad \frac{A_-}{r_{n,p}\sqrt{2\log n}} \longrightarrow 1.$$

Both signs occur with probability tending to one. Theorem 2 and

$$|\min(u_1,u_2) - \min(w_1,w_2)| \le \max\{|u_1-w_1|, |u_2-w_2|\}$$

give the uniform limit. The peak identities in Theorem 2 give the remaining conclusions for $r_{n,p}$.

We next assume $d = p-n \to \infty$. Put $G = \sqrt{p}\,X$, $W = GG^\top$, and $T = W^{-1}$, which gives $Q = pT$. The Wishart density used below is given in Muirhead (1982, Chapter 3); inverse spectral moments also arise in the fixed-ratio interpolation analysis of Hastie et al. (2022).

For $p \ge n$, let $g_1^\top$ be the first row of G and G_- the remaining $q = n-1$ rows. Put

$$B = G_-G_-^\top, \qquad z = B^{-1}G_-g_1, \qquad U_1 = g_1^\top(I_p - G_-^\top B^{-1}G_-)g_1.$$

Conditional orthogonal projections give $z \mid G_- \sim N(0, B^{-1})$ and independent $U_1 \sim \chi^2_{d+1}$. Because the latter distribution does not depend on G_-, U_1 is independent of (B, z). Block inversion gives

$$T = \begin{pmatrix} U_1^{-1} & -z^\top/U_1 \\ -z/U_1 & B^{-1} + zz^\top/U_1 \end{pmatrix}.$$

Integration against the Wishart density of B gives

$$f(z) \;\propto\; \int_{B\succ 0} |B|^{(p-q)/2} \exp\Big\{-\frac{1}{2}\,\mathrm{tr}\,\big[B(I_q + zz^\top)\big]\Big\}\, dB \;\propto\; (1+\|z\|_2^2)^{-(p+1)/2}.$$

This is the density of a standard Gaussian vector divided by the square root of an independent χ^2_{d+2} variable. Hence $\|z\|_2^2 \stackrel{d}{=} V_1/U_2$, with independent $V_1 \sim \chi^2_{n-1}$ and $U_2 \sim \chi^2_{d+2}$, both independent of U_1. Rotating ε to the first coordinate and using $(T^2)_{11} = (1 + \|z\|_2^2)/U_1^2$ gives

$$\frac{\sigma^2}{n}\varepsilon^\top Q^2 \varepsilon \stackrel{d}{=} \frac{\sigma^2 p^2}{n}\frac{V_0}{U_1^2}\left(1 + \frac{V_1}{U_2}\right), \qquad V_0 \sim \chi^2_n, \tag{S.1}$$

where V_0, V_1, U_1, U_2 are mutually independent. It follows that

$$\frac{\sigma^2 n^{-1}\varepsilon^\top Q^2 \varepsilon}{\sigma^2 (p/d)^3} \stackrel{d}{=} \frac{V_0}{n}\left(\frac{d}{U_1}\right)^2 \left\{\frac{d}{p} + \frac{n-1}{p}\frac{V_1}{n-1}\frac{d}{U_2}\right\} \longrightarrow 1$$

in probability, because all four chi-square ratios tend to one and $d/p + (n-1)/p \to 1$.

For the signal and cross terms, the same block inverse formula gives, for $d > 5$,

$$\mathbb{E}\,\mathrm{tr}\, T = \frac{n}{d-1}, \qquad \mathbb{E}\,\mathrm{tr}\, T^2 = \frac{n(p-1)}{d(d-1)(d-3)},$$

$$\mathrm{Var}(\mathrm{tr}\, T) = \frac{2n(p-1)}{d(d-1)^2(d-3)}, \qquad \mathbb{E}\,\mathrm{tr}\, T^3 = \frac{n(p-1)(p+n-1)}{(d+1)d(d-1)(d-3)(d-5)}.$$

The inverse chi-square moments and $\mathbb{E}\,\|z\|_2^2 = q/d$ give the first two formulas from $T_{11} = U_1^{-1}$ and $(T^2)_{11} = (1 + \|z\|_2^2)/U_1^2$. Orthogonal symmetry gives

$$\mathbb{E}T_{11}^2 = \frac{\mathbb{E}(\mathrm{tr}\, T)^2 + 2\mathbb{E}\,\mathrm{tr}\, T^2}{n(n+2)} = \frac{1}{(d-1)(d-3)},$$

which yields the variance formula. For the third moment,

$$(T^3)_{11} = \frac{(1 + \|z\|_2^2)^2}{U_1^3} + \frac{z^\top B^{-1} z}{U_1^2},$$

$$\mathbb{E}(1 + \|z\|_2^2)^2 = \frac{(p-1)(p-3)}{d(d-2)}, \qquad \mathbb{E}z^\top B^{-1} z = \mathbb{E}\,\mathrm{tr}\, B^{-2} = \frac{(n-1)(p-1)}{(d+1)d(d-2)}.$$

The last equality uses the second moment in dimension $n-1$. Taking expectations with $\mathbb{E}U_1^{-3} = \{(d-1)(d-3)(d-5)\}^{-1}$ gives $\mathbb{E}\,\mathrm{tr}\, T^3$.

Put $t_n^2 = \|\theta\|_2^2/p$, $b = QX\theta$, and $S = \|b\|_2^2/n$. In a singular value decomposition $X = UDV^\top$, the matrix V is uniform on the set of $p \times n$ matrices with orthonormal columns and is independent of D. Conditional spherical moments give

$$\mathbb{E}(\|b\|_2^2 \mid D) = t_n^2\, \mathrm{tr}\, Q, \qquad \mathrm{Var}(\|b\|_2^2 \mid D) \le \frac{2\,\|\theta\|_2^4}{p(p+2)}\,\mathrm{tr}\, Q^2.$$

If U_0 is uniform on the unit sphere in $\mathbb{R}^p$, these moments follow from

$$\mathbb{E}U_{0i}^2 = \frac{1}{p}, \qquad \mathbb{E}U_{0i}^4 = \frac{3}{p(p+2)}, \qquad \mathbb{E}U_{0i}^2 U_{0j}^2 = \frac{1}{p(p+2)} \quad (i \ne j).$$

A rotation by $\pi/4$ gives $\mathbb{E}U_{01}^4 = 3\mathbb{E}U_{01}^2 U_{02}^2$, and expansion of $\mathbb{E}(\sum_i U_{0i}^2)^2 = 1$ gives the constants. The trace formulas and conditional variance decomposition now imply

$$\mathbb{E}S = \frac{t_n^2 p}{d-1}, \qquad \mathrm{Var}(S) \le \frac{t_n^4}{n^2}\left\{2\mathbb{E}\,\mathrm{tr}\, Q^2 + \mathrm{Var}(\mathrm{tr}\, Q)\right\} \le \frac{C t_n^4 p^3}{nd^3}.$$

For $t_n > 0$, division by $(t_n^2 p/d)^2$ gives relative variance at most $Cp/(nd) = C(1/n + 1/d) \to 0$. For $t_n = 0$, $S = 0$. Hence

$$\frac{S - t_n^2 p/d}{v_{n,p}^2} \longrightarrow 0$$

in probability for every deterministic signal sequence.

The cross term is centered. Averaging over the noise and V gives

$$\mathbb{E}\left(\frac{2\sigma}{n}b^\top Q\varepsilon\right)^2 = \frac{4\sigma^2 t_n^2}{n^2}\mathbb{E}\,\mathrm{tr}\,Q^3 \le \frac{C\sigma^2 t_n^2 p^5}{nd^5}.$$

For $t_n > 0$, the inequality $v_{n,p}^4 \ge 4\sigma^2 t_n^2 p^4/d^4$ bounds the normalized second moment by $Cp/(nd) \to 0$; for $t_n = 0$, the cross term is zero. Because

$$r_{n,p}^2 = S + \frac{\sigma^2}{n}\varepsilon^\top Q^2\varepsilon + \frac{2\sigma}{n}b^\top Q\varepsilon, \qquad v_{n,p}^2 = \frac{t_n^2 p}{d} + \frac{\sigma^2 p^3}{d^3},$$

we obtain $r_{n,p}^2/v_{n,p}^2 \to 1$. Multiplication of the first uniform limit by $v_{n,p}/r_{n,p} \to 1$ proves the remaining assertions. This completes the proof. □

S.1.1 A FIXED EXCESS NUMBER OF FEATURES

A diverging excess dimension gives a deterministic scale without a spectral gap bounded away from zero. The next proposition shows that a fixed excess dimension does not give the same concentration, even when the signal is zero. Let $d \ge 0$ be a fixed integer, and let $U_1 \sim \chi_{d+1}^2$ and $U_2 \sim \chi_{d+2}^2$ be independent.

Proposition 5. *Under the Gaussian model of Theorem 8, if $\theta = 0$ and $p = n + d$, then*

$$\frac{r_{n,n+d}^2}{\sigma^2 n^3} \xrightarrow{d} \frac{1}{U_1^2 U_2}.$$

There is no deterministic sequence $b_n > 0$ for which $r_{n,n+d}^2/b_n \to 1$ in probability.

The random-length normalization in Theorem 8 remains valid in this case. The proposition concerns concentration of the length, not the existence of a limiting boundary shape.

Proof. By (S.1),

$$\frac{r_{n,n+d}^2}{\sigma^2 n^3} \overset{d}{=} \left(\frac{p}{n}\right)^2 \frac{V_0}{n}\frac{n^{-1} + V_1/(nU_2)}{U_1^2}.$$

Because $V_0/n, V_1/n, p/n \to 1$ and U_1, U_2 have fixed distributions, the limit follows. If $r_{n,n+d}^2/b_n \to 1$, take a subsequence on which $b_n/(\sigma^2 n^3)$ converges in $[0,\infty]$. A zero or infinite limit contradicts the positive finite limit above, and a finite positive limit would make that limit degenerate. Thus no such sequence exists. This completes the proof. □

S.1.2 A PENALIZED CONSTANT FEATURE

A constant feature permits nonzero conditional quantiles even when the conditional mean has no intercept. The following corollary shows that this feature does not change the first-order boundary scale whenever the excess number of features tends to infinity. The dimension ratio may approach one.

For a positive integer m, let $\mathbf{1}_m$ denote the vector of ones in $\mathbb{R}^m$. Let $X = [p^{-1/2}\mathbf{1}_n, Z]$, where the entries of $Z \in \mathbb{R}^{n\times(p-1)}$ are independent $N(0, 1/p)$. Let $y = Z\theta_{-1} + \sigma\varepsilon$, where $\theta_{-1} \in \mathbb{R}^{p-1}$ is nonrandom, $\sigma > 0$ is fixed, and $\varepsilon \sim N(0, I_n)$ is independent of Z. The coefficient of the constant feature is penalized, and its true conditional-mean value is zero. Define $r_{n,p} = \left\|(XX^\top)^{-1}y\right\|_2/\sqrt{n}$ and $v_{n,p}$ as above with $\|\theta\|_2^2 = \|\theta_{-1}\|_2^2$.

Corollary 6. *Assume $n \to \infty$ and $p - n \to \infty$. Then the conclusions of Theorem 8 hold for the constant-feature model.*

Corollary 6 applies to the feature construction in the Gaussian predictive experiment. The deterministic extinction identity applies to every realized full-rank design; the corollary gives its asymptotic scale for this distribution.

Proof. Put $B = ZZ^\top$ and $a_B = B^{-1}y$. For $\widetilde{X} = \sqrt{p/(p-1)}Z$ and $\widetilde{\theta} = \sqrt{(p-1)/p}\theta_{-1}$,

$$y = \widetilde{X}\widetilde{\theta} + \sigma\varepsilon, \qquad (\widetilde{X}\widetilde{X}^\top)^{-1}y = \frac{p-1}{p}a_B, \qquad \frac{\left\|\widetilde{\theta}\right\|_2^2}{p-1} = \frac{\|\theta_{-1}\|_2^2}{p}.$$

Put $d = p-n$. The signal and noise contributions to the squared scale of a_B equal their counterparts in $v_{n,p}^2$ multiplied, respectively, by

$$\frac{pd}{(p-1)(d-1)} \longrightarrow 1, \qquad \frac{p-1}{p}\left(\frac{d}{d-1}\right)^3 \longrightarrow 1.$$

Theorem 8, applied with excess dimension $d-1 \to \infty$, gives

$$\frac{\|a_B\|_2^2}{nv_{n,p}^2} \longrightarrow 1, \qquad \frac{\max_i(a_B)_i}{v_{n,p}\sqrt{2\log n}} \longrightarrow 1, \qquad \frac{\max_i\{-(a_B)_i\}}{v_{n,p}\sqrt{2\log n}} \longrightarrow 1$$

in probability. Rotational invariance of a_B and its length bound give $\mathbf{1}_n^\top a_B = O_{\mathbb{P}}(\sqrt{n}\, v_{n,p})$. The Sherman–Morrison identity yields

$$a - a_B = -\frac{B^{-1}\mathbf{1}_n\,\mathbf{1}_n^\top a_B}{p + \mathbf{1}_n^\top B^{-1}\mathbf{1}_n},$$

and if $\limsup n/p < 1$, the Gaussian singular-value bound gives $\left\|B^{-1}\right\|_{\mathrm{op}} = O_{\mathbb{P}}(1)$, and the direct bound is

$$\|a - a_B\|_\infty \le \|a - a_B\|_2 \le \frac{\sqrt{n}\,\left\|B^{-1}\right\|_{\mathrm{op}}|\mathbf{1}_n^\top a_B|}{p} = O_{\mathbb{P}}(v_{n,p}).$$

For general $d \to \infty$, put $u = \mathbf{1}_n/\sqrt{n}$, $q = u^\top B^{-1}u$, and $w = B^{-1}u/q - u$. Then $u^\top w = 0$. With $h = (nq/p)/(1+nq/p) \in [0,1]$, the same identity gives

$$a - a_B = -h(u^\top a_B)(u+w), \qquad u^\top a_B = O_{\mathbb{P}}(v_{n,p}).$$

Rotate u to the first coordinate and partition the Gaussian matrix defining B. The Schur-complement calculation used in (S.1), now with $p-1$ columns, gives

$$\|w\|_2^2 \overset{d}{=} \frac{V}{U}, \qquad V \sim \chi_{n-1}^2, \quad U \sim \chi_{d+1}^2, \quad V \perp U, \qquad \|w\|_2 = O_{\mathbb{P}}\left(\sqrt{n/d}\right).$$

The pair (B, y) is invariant under every orthogonal transformation that fixes u. The same is true of the component of $a - a_B$ orthogonal to u. Conditional on its nonzero length, its direction is uniform on the unit sphere of $u^\perp$. The Gaussian representation of this direction gives a maximum coordinate of order $O_{\mathbb{P}}\{\sqrt{\log n/n}\}$. Because $0 \le h \le 1$, these bounds imply

$$\frac{\|a - a_B\|_2}{\sqrt{n}\, v_{n,p}} = O_{\mathbb{P}}\left(n^{-1/2} + d^{-1/2}\right), \qquad \frac{\|a - a_B\|_\infty}{v_{n,p}\sqrt{2\log n}} = O_{\mathbb{P}}\left((n\log n)^{-1/2} + d^{-1/2}\right).$$

Both bounds tend to zero. Thus the length and signed-maximum limits for a_B also hold for a. Theorem 2 gives the boundary and peak limits under both normalizations. This completes the proof. □

S.1.3 Fluctuations at the Gaussian boundary

The first-order triangle does not describe which levels remain extinct when the penalty is close to its maximum. We now determine the fluctuations of the first extinct level and distinguish its boundary from the median boundary. The normalization by the observed length remains valid even when the excess feature dimension is fixed.

Use the Gaussian model and $r_{n,p}$ of Theorem 8. For $n \ge 3$, let $u_n = \Phi^{-1}(1 - 1/n)$, where Φ is the standard Gaussian distribution function. Let G_+ and G_- be independent variables with distribution function $x \mapsto \exp\{-\exp(-x)\}$. On the event that a has only one sign, define $\tau^\dagger = 1/2$ for the following asymptotic statements. This event has probability 2^{1-n} and does not affect the limits.

Theorem 9. *If $n \to \infty$ and $p \geq n$, then jointly,*

$$\left(u_n \left\{ \frac{A_+}{r_{n,p}} - u_n \right\}, u_n \left\{ \frac{A_-}{r_{n,p}} - u_n \right\} \right) \xrightarrow{d} (G_+, G_-),$$

$$4u_n^2(\tau^\dagger - 1/2) \xrightarrow{d} G_+ - G_-,$$

$$u_n^2 \left\{ 2r_{n,p}u_n \sup_{\tau \in (0,1)} \bar{\lambda}(\tau) - 1 \right\} \xrightarrow{d} -\frac{G_+ + G_-}{2},$$

$$u_n^2 \{ 2r_{n,p}u_n \bar{\lambda}(1/2) - 1 \} \xrightarrow{d} -\max(G_+, G_-).$$

The first extinct level differs from the median by order $1/\log n$. Its scaled difference has the logistic distribution, with distribution function $x \mapsto (1 + e^{-x})^{-1}$. The largest boundary depends on the sum of the two extreme fluctuations, whereas the median boundary depends on their maximum. These are different second-order limits, although their first-order limits agree.

Proof. The spherical representation in the proof of Theorem 8 gives

$$\frac{a}{r_{n,p}} \overset{d}{=} \frac{\sqrt{n}}{\|Z\|_2} Z, \qquad Z \sim N(0, I_n).$$

For $t > 0$, integration by parts gives the Gaussian tail bounds

$$\frac{t}{1+t^2}\,\phi(t) \leq 1 - \Phi(t) \leq \frac{\phi(t)}{t}, \qquad \phi(t) = (2\pi)^{-1/2} e^{-t^2/2}.$$

Thus $u_n \to \infty$, $u_n^2/(2\log n) \to 1$, and, for fixed $x \in \mathbb{R}$,

$$n\{1 - \Phi(u_n + x/u_n)\} = \frac{1 - \Phi(u_n + x/u_n)}{1 - \Phi(u_n)} \longrightarrow e^{-x}.$$

For fixed $x, y \in \mathbb{R}$, the upper and lower tail events are disjoint for all large n. Independence of the coordinates gives

$$\begin{aligned} &\mathbb{P}\left\{ \max_i Z_i \leq u_n + x/u_n,\ \max_i (-Z_i) \leq u_n + y/u_n \right\} \\ &\quad = [1 - \{1 - \Phi(u_n + x/u_n)\} - \{1 - \Phi(u_n + y/u_n)\}]^n \longrightarrow \exp\{-e^{-x} - e^{-y}\}. \end{aligned}$$

Also $\|Z\|_2/\sqrt{n} = 1 + O_{\mathbb{P}}(n^{-1/2})$. Each Gaussian maximum is $O_{\mathbb{P}}(u_n)$, hence replacing Z by $\sqrt{n}Z/\|Z\|_2$ changes either centered and scaled maximum by $O_{\mathbb{P}}(u_n^2/\sqrt{n}) = o_{\mathbb{P}}(1)$. The probability that either Gaussian maximum is nonpositive is at most 2^{1-n}. This proves the first joint limit for A_+ and A_-.

Write $H_{n,\pm} = u_n(A_\pm/r_{n,p} - u_n)$. On the event that both signs occur, Theorem 2 gives

$$4u_n^2(\tau^\dagger - 1/2) = \frac{2(H_{n,+} - H_{n,-})}{2 + (H_{n,+} + H_{n,-})/u_n^2},$$

$$2r_{n,p}u_n \sup_\tau \bar{\lambda}(\tau) = \left\{ 1 + \frac{H_{n,+} + H_{n,-}}{2u_n^2} \right\}^{-1},$$

$$2r_{n,p}u_n \bar{\lambda}(1/2) = \left\{ 1 + \frac{\max(H_{n,+}, H_{n,-})}{u_n^2} \right\}^{-1}.$$

The remaining limits follow by expansion at zero and the first joint limit. For the logistic distribution, $E_\pm = e^{-G_\pm}$ are independent exponential variables of mean one. Hence

$$\mathbb{P}(G_+ - G_- \leq x) = \mathbb{P}(E_- \leq e^x E_+) = \int_0^\infty (1 - e^{-e^x t}) e^{-t}\, dt = \frac{1}{1 + e^{-x}}.$$

This completes the proof. □

For fixed $t \in \mathbb{R}$, put $\lambda_n(t) = \{r_{n,p}(2u_n + t/u_n)\}^{-1}$ for all sufficiently large n. This is a random normalization for describing the boundary, not a prescription for minimizing prediction risk. Recall that E_λ is the set of extinct levels.

Theorem 9 gives

$$\mathbb{P}\{E_{\lambda_n(t)} \neq \varnothing\} \longrightarrow \mathbb{P}(G_+ + G_- \leq t), \qquad \mathbb{P}\{1/2 \in E_{\lambda_n(t)}\} \longrightarrow \exp\{-2e^{-t/2}\}. \tag{S.2}$$

The first probability concerns collapse at some level, and the second concerns collapse at the median. Equation (S.2) resolves these probabilities in a shrinking neighborhood that the first-order triangle cannot distinguish.

To prove (S.2), on the event that both signs occur the interval formula in Theorem 2 yields

$$\begin{aligned} E_{\lambda_n(t)} \neq \varnothing \quad &\Longleftrightarrow \quad H_{n,+} + H_{n,-} \leq t, \\ 1/2 \in E_{\lambda_n(t)} \quad &\Longleftrightarrow \quad \max(H_{n,+}, H_{n,-}) \leq t/2. \end{aligned}$$

The sum and maximum of G_+, G_- have continuous distribution functions. Apply Theorem 9 and use independence for the second probability.

A population normalizer requires more control of the length than a first-order limit. The next proposition gives the additional rate. All constants in its probability order are uniform over the deterministic signal sequence.

Proposition 6. *In the Gaussian model, if $n \to \infty$ and $d = p - n \to \infty$, then*

$$\frac{r_{n,p}}{v_{n,p}} - 1 = O_{\mathbb{P}}(n^{-1/2} + d^{-1/2}).$$

If also $d/(\log n)^2 \to \infty$, then Theorem 9 and the limits in (S.2) *hold with $v_{n,p}$ in place of $r_{n,p}$.*

A diverging excess dimension suffices for the first-order population scale. The second-order conclusion above requires its fluctuations to be smaller than $1/\log n$. No deterministic second-order normalization is asserted here when the excess dimension is fixed or grows more slowly.

Proof. Let $b = QX\theta$, and put

$$S = \frac{\|b\|_2^2}{n}, \qquad T_0 = \frac{\sigma^2}{n}\varepsilon^\top Q^2 \varepsilon, \qquad W_0 = \frac{2\sigma}{n} b^\top Q\varepsilon, \qquad s_0 = \frac{\|\theta\|_2^2}{d}, \qquad v_0 = \frac{\sigma^2 p^3}{d^3}.$$

Thus $r_{n,p}^2 = S + T_0 + W_0$ and $v_{n,p}^2 = s_0 + v_0$. The trace moments established in the proof of Theorem 8 give, for $d > 5$,

$$\mathbb{E}S = s_0 \frac{d}{d-1}, \qquad \mathrm{Var}(S) \leq C s_0^2 (1/n + 1/d), \qquad \mathbb{E}W_0^2 \leq C s_0 v_0 (1/n + 1/d).$$

When $s_0 = 0$, both S and W_0 equal zero. For $s_0 > 0$, Chebyshev's inequality and $2\sqrt{s_0 v_0} \leq s_0 + v_0$ give

$$\frac{S - s_0 + W_0}{s_0 + v_0} = O_{\mathbb{P}}(n^{-1/2} + d^{-1/2}).$$

By (S.1), with the independent chi-square variables defined there,

$$\frac{T_0}{v_0} \overset{d}{=} \frac{V_0}{n} \left(\frac{d}{U_1}\right)^2 \left\{\frac{d}{p} + \frac{n-1}{p} \frac{V_1}{n-1} \frac{d}{U_2}\right\}.$$

The chi-square means and variances imply

$$\frac{V_0}{n} = 1 + O_{\mathbb{P}}(n^{-1/2}), \qquad \frac{V_1}{n-1} = 1 + O_{\mathbb{P}}(n^{-1/2}), \qquad \frac{d}{U_j} = 1 + O_{\mathbb{P}}(d^{-1/2}), \quad j = 1, 2.$$

The two weights inside braces sum to $1 - 1/p$ and are bounded by one. Their product expression is thus $1 + O_{\mathbb{P}}(n^{-1/2} + d^{-1/2})$. Combining the noise and signal bounds proves

$$\frac{r_{n,p}^2}{v_{n,p}^2} - 1 = O_{\mathbb{P}}(n^{-1/2} + d^{-1/2}).$$

The square-root identity gives the stated rate. If $d/(\log n)^2 \to \infty$, then $u_n^2(r_{n,p}/v_{n,p} - 1) = o_{\mathbb{P}}(1)$. In particular,

$$u_n \left\{ \frac{A_\pm}{v_{n,p}} - u_n \right\} = \frac{r_{n,p}}{v_{n,p}} H_{n,\pm} + u_n^2 \left(\frac{r_{n,p}}{v_{n,p}} - 1 \right) = H_{n,\pm} + o_{\mathbb{P}}(1).$$

Use these two expansions in the boundary identities from the preceding proofs. This establishes every stated replacement without an independence assumption between the length and its estimation error. This completes the proof. □

S.2 NON-GAUSSIAN FEATURES AND SIGNAL CONCENTRATION

Under non-Gaussian features, the extinction boundary scale need not be determined solely by the magnitude of the true signal variance; how that signal is distributed across the individual features can also matter. In this section, we first establish a condition under which the boundary scale remains identical to the Gaussian scale derived in Theorem 4. We then analyze a model with independent sign-valued features to quantify how the scale changes when this condition is violated.

Let $X_{ij} = \xi_{ij}/\sqrt{p}$, where the ξ_{ij} are independent, have mean zero and variance one, and satisfy $\sup_{n,p,i,j} \|\xi_{ij}\|_{\psi_2} \le L < \infty$. Let $y = X\theta + \sigma\varepsilon$, where $\theta \in \mathbb{R}^p$ is deterministic, $\sigma > 0$ is fixed, and $\varepsilon \sim N(0, I_n)$ is independent of X. As before, $s_{n,p}^2 = \sigma^2 + \|\theta\|_2^2/p$, and the boundary is set to zero on rank-deficient designs. No upper bound is imposed on $s_{n,p}$.

Theorem 10. *Under the preceding model, if*

$$n \longrightarrow \infty, \qquad \frac{n}{p} \longrightarrow 0, \qquad \frac{\|\theta\|_\infty \sqrt{\log n}}{\sqrt{p}\, s_{n,p}} \longrightarrow 0,$$

then, in probability,

$$\sup_{\tau \in (0,1)} \left| s_{n,p} \sqrt{2 \log n}\, \bar{\lambda}(\tau) - \{\tau \wedge (1-\tau)\} \right| \longrightarrow 0.$$

Theorem 10 allows the signal variance to be either fixed or increasing. The key condition ($\|\theta\|_\infty \sqrt{\log n}/(\sqrt{p}\, s_{n,p}) \to 0$) prevents any single signal coordinate from dominating the response extremes relative to the total response scale. For positive sequences, $b_n \gg c_n$ means $b_n/c_n \to \infty$, and $b_n \ll c_n$ means $b_n/c_n \to 0$. For a nonzero signal, it is mathematically sufficient that $\|\theta\|_2^2 / \|\theta\|_\infty^2 \gg \log n$, which is a condition on the true parameters and not on the fitted model. This complements Remark 1, which previously showed that highly concentrated coefficients still yield the Gaussian boundary limit when their total signal variance tends to zero. Because this theorem recovers the asymptotic scale of the Gaussian case, the first extinct level again converges to the median by Theorem 2, and the critical penalty phases remain exactly as stated in Corollary 3.

The next model shows why the number of signal coordinates is relevant. Suppose that the ξ_{ij} are independent uniform signs. For fixed $r > 0$, let $\theta_j = r\sqrt{p/k_n}$ for $1 \le j \le k_n$, and let all other coefficients be zero, where $1 \le k_n \le p$. Thus the signal variance is r^2 for every k_n. For $0 < \alpha < \infty$, define

$$H_\alpha(t) = \frac{\sigma^2 t^2}{2} + \frac{\alpha}{2} \log\cosh\left(rt\sqrt{\frac{2}{\alpha}} \right), \qquad I_\alpha(x) = \sup_{t \ge 0} \{tx - H_\alpha(t)\}.$$

Let c_α be the unique positive solution of $I_\alpha(c_\alpha) = 1/2$, and set $c_0 = \sigma$ and $c_\infty = \sqrt{\sigma^2 + r^2}$. The proof below establishes existence and uniqueness, which specifies these constants without a further assumption.

Theorem 11. *Under the sign-feature model, if $n \to \infty$, $n/p \to 0$, and $k_n/\log n \to \alpha \in [0, \infty]$, then, in probability,*

$$\sup_{\tau \in (0,1)} \left| c_\alpha \sqrt{2 \log n}\, \bar{\lambda}(\tau) - \{\tau \wedge (1-\tau)\} \right| \longrightarrow 0.$$

The function $\alpha \mapsto c_\alpha$ is continuous and strictly increasing on $(0, \infty)$, with limits c_0 and c_∞ at its endpoints.

Theorem 11 separates the boundary behavior into three regimes for models with identical signal variance. If the signal is supported on few features ($k_n \ll \log n$), the noise variance alone determines the boundary scale. If the signal is spread out over many features ($k_n \gg \log n$), the full response variance determines the scale. In the critical transition regime where k_n is proportional to $\log n$, the boundary scale takes an intermediate value determined by the one-dimensional equation for c_α. This transition affects only the height of the boundary; its normalized triangular shape across quantile levels remains intact. This gives a precise extension of the one-coordinate example in Proposition 4. It does not assert that the same transition holds for every non-Gaussian feature distribution. The constants describe the population scale; computation of the exact boundary from X and y does not require r or σ.

S.2.1 Proofs for non-Gaussian signal effects

The proofs separate two questions. We first compare the coordinates of $a = (XX^\top)^{-1}y$ with the responses. We then determine which response extremes are produced by the signal. All probability bounds follow from the distributional and parameter conditions in Section S.2.

S.2.1.1 Coordinate approximation

The following bound applies without the coefficient condition in Theorem 10. It allows us to study response extremes before multiplication by the inverse Gram matrix. In its statement, set $a = 0$ on rank-deficient designs.

Lemma 1. *Under the independent sub-Gaussian model in Section S.2, if $n \to \infty$ and $n/p \to 0$, then*

$$\mathbb{P}\{\operatorname{rank}(X) = n\} \longrightarrow 1, \qquad \|a - y\|_\infty = O_{\mathbb{P}}\left(s_{n,p}\sqrt{\frac{n\log n}{p}}\right).$$

The error in Lemma 1 is negligible relative to $s_{n,p}\sqrt{\log n}$. An operator-norm bound alone would introduce an unnecessary factor $\sqrt{n}$ at this step.

Proof. Write $b = \|\theta\|_2/\sqrt{p}$ and $\rho = \sqrt{n/p}$. Constants below depend only on the uniform sub-Gaussian bound. The Gram-matrix concentration inequality for independent isotropic sub-Gaussian rows, applied to $\sqrt{p}\,X^\top$, gives

$$\mathbb{P}\{\|XX^\top - I_n\|_{\mathrm{op}} > C\rho\} \le 2e^{-cn}.$$

Here is a direct justification of the concentration step. For a unit vector $v \in \mathbb{R}^n$, the variables $Z_j = \sum_i v_i\xi_{ij}$ are independent, have variance one, and have uniformly bounded sub-Gaussian norms. Thus $Z_j^2 - 1$ have uniformly bounded sub-exponential norms. Bernstein's inequality gives

$$\mathbb{P}\{|v^\top(XX^\top - I_n)v| > t\} \le 2\exp\{-cp\min(t^2, t)\}.$$

A $1/4$-net of the unit sphere has at most 9^n points, and the operator norm of a symmetric matrix is at most twice its largest absolute quadratic form on this net. A union bound with t a sufficiently large multiple of $\sqrt{n/p}$ gives the claimed probability. These are the standard net and concentration inequalities described in Vershynin (2018); the distributions of the entries need not be identical. For large n, the resulting bound implies full row rank and

$$\|Q - I_n\|_{\mathrm{op}} \le C\rho, \qquad \|Q\|_{\mathrm{op}} \le 2.$$

If $b > 0$, the variables $x_i^\top\theta/b$ are independent, centered, and uniformly sub-Gaussian. Concentration of their squared sum gives

$$\mathbb{P}\{\|X\theta\|_2 > Cb\sqrt{n}\} \le 2e^{-cn}.$$

For $b = 0$, all signal terms below are zero.

Delete row i from X to obtain X_{-i}, and let P_{-i} be the orthogonal projection onto its row space. Consider the event determined by X_{-i} on which

$$\lambda_{\min}(X_{-i}X_{-i}^\top) \ge \frac{1}{2}, \qquad \|X_{-i}\theta\|_2 \le Cb\sqrt{n}.$$

The preceding Gram and norm events imply these two inequalities simultaneously for all i. On the event for row i,

$$\|P_{-i}\theta\|_2^2 = (X_{-i}\theta)^\top (X_{-i}X_{-i}^\top)^{-1} X_{-i}\theta \le Cb^2 n.$$

Conditionally on X_{-i}, the vector x_i is independent of $P_{-i}\theta$. Thus, for $u > 0$, the probability that this row event holds and $|x_i^\top P_{-i}\theta| > Cb\rho u$ is at most $2e^{-cu^2}$. A union bound with u a sufficiently large multiple of $\sqrt{\log n}$ gives

$$\max_{i\le n} |x_i^\top P_{-i}\theta| = O_{\mathbb{P}}(b\rho\sqrt{\log n}), \qquad \max_{i\le n} |x_i^\top \theta| = O_{\mathbb{P}}(b\sqrt{\log n}).$$

This conditioning uses only the event determined by the other rows, not an event that also constrains x_i.

The Schur complement identity on a full-row-rank design gives

$$(QX\theta)_i = Q_{ii}x_i^\top (I - P_{-i})\theta, \qquad (QX\theta - X\theta)_i = (Q_{ii} - 1)x_i^\top\theta - Q_{ii}x_i^\top P_{-i}\theta.$$

The last two bounds and the bounds on Q show that

$$\|QX\theta - X\theta\|_\infty = O_{\mathbb{P}}(b\rho\sqrt{\log n}).$$

Conditionally on X, the vector $(Q - I_n)\varepsilon$ is centered Gaussian. Each coordinate has variance at most $\|Q - I_n\|_{\mathrm{op}}^2$. Gaussian tail bounds and a union bound hence give

$$\|(Q - I_n)\varepsilon\|_\infty = O_{\mathbb{P}}(\rho\sqrt{\log n}).$$

Adding the signal and noise terms and using $b + \sigma \le \sqrt{2}s_{n,p}$ proves the stated rate. The rank-deficient event has probability tending to zero, and the convention there does not change the conclusion. This completes the proof. □

S.2.1.2 Proof of Theorem 10

The coefficient condition ensures the appropriate logarithmic tail probabilities, beyond convergence in distribution of a single response. Such tail probabilities are needed because the boundary depends on the largest coordinates over all observations.

Proof. Put $s = s_{n,p}$, $w_j = \theta_j/(\sqrt{p}\, s)$, $w_0 = \sigma/s$, and $h = \max_{1\le j\le p} |w_j|$. The normalized responses are independent and satisfy

$$Z_i := y_i/s = \sum_{j=1}^p w_j\xi_{ij} + w_0\varepsilon_i, \qquad \sum_{j=1}^p w_j^2 + w_0^2 = 1, \qquad h\sqrt{\log n} \longrightarrow 0.$$

For a centered, variance-one variable ξ with the stated sub-Gaussian bound, let $g(u) = \log \mathbb{E}e^{u\xi}$. There are constants $c, C > 0$, uniform over the entries of the design, such that for $|u| \le c$,

$$|g(u) - u^2/2| \le C|u|^3, \qquad |g'(u) - u| \le Cu^2, \qquad |g''(u) - 1| \le C|u|.$$

To justify the uniform remainder, the sub-Gaussian tail bound implies a uniform bound on $\mathbb{E}(|\xi|^j e^{c|\xi|})$ for $0 \le j \le 3$, and Jensen's inequality gives $\mathbb{E}e^{u\xi} \ge 1$. Differentiating $\log \mathbb{E}e^{u\xi}$ three times hence gives a uniformly bounded third derivative on $[-c, c]$. Taylor expansion at zero proves the inequalities.

Let $K_i(t) = \log \mathbb{E}e^{tZ_i}$. Independence and the Gaussian moment generating function give, whenever $h|t| \le c$,

$$K_i(t) = t^2/2 + O(h|t|^3), \qquad K_i'(t) = t + O(ht^2), \qquad K_i''(t) = 1 + O(h|t|).$$

The constants are uniform in i. If $h = 0$, these equalities hold with zero remainder. Fix $d > 0$ and let $x = d\sqrt{2\log n}$. Chernoff's inequality at $t = x$ gives

$$\log \mathbb{P}(Z_i \ge x) \le -x^2/2 + o(x^2)$$

uniformly in i.

For the lower bound, fix $\eta > 0$ and change probability by the density $e^{tZ_i - K_i(t)}$, where $t = (1+\eta)x$. Under this probability, the mean of Z_i is $K_i'(t) = (1+\eta)x + o(x)$ and its variance is $K_i''(t) = 1 + o(1)$. Chebyshev's inequality shows, uniformly in i, that the probability of $x \le Z_i \le (1+2\eta)x$ tends to one. Reversing the change of probability yields

$$\begin{aligned}\log \mathbb{P}(Z_i \ge x) &\ge K_i(t) - t(1+2\eta)x + o(1) \\ &= -\{1/2 + 2\eta + 3\eta^2/2\}x^2 + o(x^2).\end{aligned}$$

First let n tend to infinity and then let η tend to zero. Applying the same argument to $-Z_i$ proves, for either sign,

$$\sup_{i \le n} \left| \frac{\log \mathbb{P}(\pm Z_i \ge d\sqrt{2\log n})}{\log n} + d^2 \right| \longrightarrow 0.$$

For $d > 1$, a union bound makes both upper-exceedance probabilities tend to zero. For $0 < d < 1$, independence of the rows gives

$$\mathbb{P}\{\max_i(\pm Z_i) < d\sqrt{2\log n}\} \le \exp\{-n \min_i \mathbb{P}(\pm Z_i \ge d\sqrt{2\log n})\} \longrightarrow 0.$$

Thus, both signed response maxima divided by $s\sqrt{2\log n}$ tend to one in probability. Lemma 1 gives

$$\frac{A_+}{s\sqrt{2\log n}} \longrightarrow 1, \qquad \frac{A_-}{s\sqrt{2\log n}} \longrightarrow 1$$

in probability. With probability tending to one both maxima are positive. Theorem 2 then bounds the asserted supremum by

$$\max\left\{ \left| \frac{s\sqrt{2\log n}}{A_+} - 1 \right|, \left| \frac{s\sqrt{2\log n}}{A_-} - 1 \right| \right\},$$

which tends to zero. The same two signed maxima give the stated consequences for $\tau^\dagger$ and the penalty regimes. This completes the proof. □

S.2.1.3 Proof of Theorem 11

In this model the signal variance is fixed. The calculation below determines how the number of its nonzero coordinates changes the tail probabilities at the level relevant to the sample maximum.

Proof. Write $k = k_n$ and $u_n = \sqrt{2\log n}$. Each response has the distribution of

$$Y_n = \frac{r}{\sqrt{k}} \sum_{j=1}^{k} \xi_j + \sigma\varepsilon,$$

where the signs and the standard Gaussian variable are independent. If $k/\log n \to 0$, the absolute signal is bounded by $r\sqrt{k} = o(u_n)$. The two signed Gaussian noise maxima divided by u_n tend to σ. This proves the response-maxima assertion with $c_0 = \sigma$. If $k/\log n \to \infty$, the coefficient condition in Theorem 10 holds, because its left side is $(r/\sqrt{\sigma^2 + r^2})\sqrt{\log n/k}$. That theorem proves the assertion with $c_\infty = \sqrt{\sigma^2 + r^2}$.

Suppose next that $k/\log n \to \alpha \in (0, \infty)$. Define

$$H_n(t) = u_n^{-2} \log \mathbb{E} e^{t u_n Y_n} = \frac{\sigma^2 t^2}{2} + \frac{k}{u_n^2} \log\cosh\left(\frac{r t u_n}{\sqrt{k}} \right).$$

The functions H_n and their first two derivatives converge uniformly on each bounded interval to H_α and its corresponding derivatives. Also,

$$H_\alpha'(0) = 0, \qquad \sigma^2 \le H_\alpha''(t) = \sigma^2 + r^2 \operatorname{sech}^2\left(rt\sqrt{2/\alpha} \right) \le \sigma^2 + r^2.$$

For each $x > 0$, there is a unique $t_x > 0$ with $H_\alpha'(t_x) = x$. It is the maximizer defining $I_\alpha(x)$, and $t_x \le x/\sigma^2$. In particular, $I_\alpha(0) = 0$, I_α is continuous and strictly increasing on $(0, \infty)$, and

$$\frac{x^2}{2(\sigma^2 + r^2)} \le I_\alpha(x) \le \frac{x^2}{2\sigma^2}.$$

These facts prove existence and uniqueness of c_α.

For fixed $x > 0$, Chernoff's inequality gives

$$\limsup_n u_n^{-2} \log \mathbb{P}(Y_n \geq x u_n) \leq -I_\alpha(x).$$

For a matching lower bound, fix $\epsilon > 0$ and choose $t > 0$ such that $H'_\alpha(t) = x + \epsilon/2$. Under the probability with density $e^{t u_n Y_n - u_n^2 H_n(t)}$, the mean and variance of Y_n/u_n are $H'_n(t)$ and $H''_n(t)/u_n^2$. Hence the probability of $x \leq Y_n/u_n \leq x + \epsilon$ tends to one. It follows that

$$\liminf_n u_n^{-2} \log \mathbb{P}(Y_n \geq x u_n) \geq H_\alpha(t) - t(x + \epsilon).$$

Letting ϵ decrease to zero proves

$$\lim_n u_n^{-2} \log \mathbb{P}(Y_n \geq x u_n) = -I_\alpha(x).$$

The same formula holds for $-Y_n$ by symmetry. Because $u_n^2 = 2 \log n$, independent rows and the upper and lower maximum bounds used in the preceding proof give

$$\frac{\max_i y_i}{u_n} \longrightarrow c_\alpha, \qquad \frac{\max_i (-y_i)}{u_n} \longrightarrow c_\alpha$$

in probability. Lemma 1, now with $s_{n,p}^2 = \sigma^2 + r^2$, transfers these limits to A_+ and A_-. Theorem 2 proves the boundary limit, including the case $\alpha = 0$ established above.

It remains to establish the dependence on α. For $t > 0$, put $z = rt\sqrt{2/\alpha}$. Direct differentiation gives

$$\frac{\partial H_\alpha(t)}{\partial \alpha} = \frac{1}{4}\{2 \log \cosh z - z \tanh z\} > 0.$$

Indeed, the expression in braces is zero at $z = 0$ and has derivative $\tanh z - z \operatorname{sech}^2 z > 0$ for $z > 0$, because $\sinh z \cosh z > z$. If $0 < \alpha_1 < \alpha_2$, evaluation at the positive maximizer for $I_{\alpha_2}(x)$ gives

$$I_{\alpha_1}(x) > I_{\alpha_2}(x), \qquad x > 0.$$

Thus $c_{\alpha_1} < c_{\alpha_2}$. For finite positive α and $t > 0$,

$$\frac{\sigma^2 t^2}{2} < H_\alpha(t) < \frac{(\sigma^2 + r^2) t^2}{2}.$$

The first strict inequality evaluated at the maximizer of $I_\alpha(\sigma)$ shows $I_\alpha(\sigma) < 1/2$. The second evaluated at $t = 1/\sqrt{\sigma^2 + r^2}$ shows $I_\alpha(\sqrt{\sigma^2 + r^2}) > 1/2$. Hence $\sigma < c_\alpha < \sqrt{\sigma^2 + r^2}$.

Finally, $H_\alpha(t)$ converges locally uniformly to $\sigma^2 t^2/2$ as $\alpha \downarrow 0$ and to $(\sigma^2 + r^2)t^2/2$ as $\alpha \to \infty$. For the first limit, use $\log \cosh z \leq |z|$; for the second, use its expansion at zero. The bound $t_x \leq x/\sigma^2$ makes the suprema defining $I_\alpha(x)$ range over a common compact interval when x is bounded. This proves locally uniform convergence of the rate functions. The same argument at each finite positive α proves continuity. Their unique $1/2$ roots hence converge to σ and $\sqrt{\sigma^2 + r^2}$, respectively, and depend continuously on α. This completes the proof. □

S.3 Boundary screening without a coefficient bound in the search

Theorem 6 gives guarantees for a penalty choice that depends on a known bound on the true coefficients in a correctly specified uniform-noise model. Such coefficient bounds are rarely known. The following theorem analyzes a search over a fixed penalty grid that does not use the true coefficient norm. Clipping predictions to the known response bound keeps the validation losses bounded without a common upper bound on the fitted coefficients, and screening out penalties at or below the observed extinction boundary removes the extinct candidates without removing the accurate one. The guarantee allows approximation error and bounded, possibly heteroskedastic conditional response distributions under the stated population, sampling, and grid conditions.

Let (Z, Y) satisfy $|Y| \leq M$ and $\|\psi(Z)\|_2 \leq \kappa$, where $M > 0$ is known and ψ is fixed before sampling. The feature distribution is such that n independent feature rows have rank n almost

surely. Use $n \geq 2$ independent copies of (Z, Y) for training and $m \geq 1$ further independent copies for validation. Fix L levels $\tau_\ell \in (0,1)$ and measurable conditional quantiles q_ℓ. In this subsection, write

$$R_\ell(f) = \mathbb{E}_* \rho_{\tau_\ell}\{Y - f(Z)\}, \qquad r_{\min} = \min_{\ell \leq L} R_\ell(q_\ell), \qquad \pi_M(t) = \max\{-M, \min(M, t)\}.$$

For a fitted predictor, $\mathbb{E}_*$ conditions on the training and validation samples. Thus $R_\ell(\psi^\top b)$ equals the coefficient risk $R_\ell(b)$ in Section 4.

For $J_n = \lceil 3 \log_2 n \rceil$, use the deterministic grid

$$\Lambda_n = \{2^j/n : 0 \leq j \leq J_n\}, \qquad \Lambda_{n,\ell}^{\mathrm{scr}} = \{\lambda \in \Lambda_n : \lambda > \bar{\lambda}(\tau_\ell)\}.$$

For each level, fit the candidates in $\Lambda_{n,\ell}^{\mathrm{scr}}$ and evaluate the ordinary validation pinball loss of

$$f_{\ell,\lambda}(z) = \pi_M\{\psi(z)^\top \widehat{\beta}(\lambda, \tau_\ell)\}.$$

Write $\widehat{R}_\ell(f)$ for the mean validation loss. Choose the largest penalty among exact validation minimizers and denote its predictor by $\widehat{f}_\ell$. If the screened grid is empty, set $\widehat{f}_\ell = 0$. This defines the procedure for every training sample of full row rank. It uses M, but not a coefficient bound or a noise distribution.

To establish a theoretical risk bound, we fix a deterministic $B > 0$ and compare the selected model against an arbitrary deterministic reference set of coefficients $b_1, \ldots, b_L$ satisfying $\|b_\ell\|_2 \leq B$. We define their corresponding population approximation error

$$A_B = \max_{\ell \leq L}\{R_\ell(\psi^\top b_\ell) - R_\ell(q_\ell)\}.$$

Using the constants t_n and e_n from Section 4 for this arbitrary B and a fixed failure probability $\delta \in (0,1)$, we define the following analytical quantities:

$$h_n = (M + \kappa B)\sqrt{\frac{\log(4L/\delta)}{2n}}, \qquad \lambda_* = \frac{n(e_n + A_B)}{2B^2},$$

$$D_{n,m}^{\mathrm{scr}} = \frac{3}{2}(e_n + A_B) + 4M\sqrt{\frac{\log\{4L(J_n+1)/\delta\}}{2m}}.$$

For the calibration bounds, let $F_z(t) = \mathbb{P}(Y \leq t \mid Z = z)$ denote the true conditional distribution function, and let $L_F > 0$ be a fixed constant. A conditional density bound is required only for the additional calibration conclusion. The quantity λ_* is used only in the proof; the actual search procedure never needs to compute it. The following theorem demonstrates that under its stated conditions, boundary screening reliably preserves a candidate model whose excess population risk is controlled by the approximation and training errors.

Theorem 12. *Assume the preceding population and sampling conditions and*

$$r_{\min} > e_n + h_n, \qquad n^{-1} \leq \lambda_* \leq 2^{J_n}/n.$$

With probability at least $1 - \delta$, simultaneously for $\ell \leq L$,

$$\bar{\lambda}(\tau_\ell) < \lambda_*, \qquad \Lambda_{n,\ell}^{\mathrm{scr}} \neq \emptyset, \qquad R_\ell(\widehat{f}_\ell) - R_\ell(q_\ell) \leq D_{n,m}^{\mathrm{scr}}.$$

If $|F_z(t) - F_z(s)| \leq L_F|t - s|$ for almost every z and all $s, t \in \mathbb{R}$, then on the same event

$$\mathbb{E}_*\{F_Z(\widehat{f}_\ell(Z)) - \tau_\ell\}^2 \leq 2L_F D_{n,m}^{\mathrm{scr}}, \qquad |\mathbb{P}_*\{Y \leq \widehat{f}_\ell(Z)\} - \tau_\ell| \leq \sqrt{2L_F D_{n,m}^{\mathrm{scr}}}.$$

The theorem separates approximation error, training error, and validation error. No lower bound on a conditional density is needed for the calibration conclusion. The coefficient bound appears in the risk comparison, not in the search. Nonextinction concerns the fitted coefficients before clipping; it does not imply separation of clipped predictors. The result also does not assert a risk improvement over every unscreened grid.

Proof. Let P_n and P denote training and population means. Put $\ell_\ell(b) = \rho_{\tau_\ell}(Y - \psi(Z)^\top b)$, $A_\ell = P\ell_\ell(b_\ell) - R_\ell(q_\ell)$, and $h_{\ell,b} = \ell_\ell(b) - \ell_\ell(b_\ell)$. Then $0 \le A_\ell \le A_B$ and $Ph_{\ell,b} \ge -A_\ell$. The loss difference $h_{\ell,b}$ is a 1-Lipschitz function of $\psi(Z)^\top(b - b_\ell)$ and equals zero at $b = b_\ell$. Symmetrization and contraction give

$$\mathbb{E} \sup_{\|b\|_2 \le 3B} |(P_n - P)h_{\ell,b}| \le 4\mathbb{E} \sup_{\|v\|_2 \le 4B} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \psi(Z_i)^\top v \right| \le \frac{16B}{n} \mathbb{E} \left\| \sum_{i=1}^n \epsilon_i \psi(Z_i) \right\|_2 \le \frac{16\kappa B}{\sqrt{n}},$$

where the ϵ_i are independent uniform signs. These are the complexity inequalities in Bartlett & Mendelson (2002). Replacing one observation changes the supremum by at most $8\kappa B/n$. The bounded-differences inequality and a union bound give

$$\sup_{\ell \le L,\, \|b\|_2 \le 3B} |(P_n - P)h_{\ell,b}| \le e_n \tag{S.3}$$

with probability at least $1 - \delta/4$.

Also, $0 \le \ell_\ell(b_\ell) \le M + \kappa B$. The inequality of Hoeffding (1963) and a union bound give

$$P_n\ell_\ell(b_\ell) \ge R_\ell(q_\ell) + A_\ell - h_n \qquad (\ell \le L) \tag{S.4}$$

with probability at least $1 - \delta/4$.

Consider the intersection of (S.3) and (S.4). For $\gamma = \lambda/n$, define $J_{\ell,\gamma}(b) = P_n\ell_\ell(b) + \gamma \|b\|_2^2/2$. If $\|b\|_2 = 3B$ and $\lambda \ge \lambda_*$, then

$$J_{\ell,\gamma}(b) - J_{\ell,\gamma}(b_\ell) \ge -A_\ell - e_n + 4\gamma B^2 \ge e_n + A_B > 0.$$

Strict convexity and coercivity give a unique minimizer. If its norm were at least $3B$, the segment from b_ℓ to that minimizer would intersect the radius-$3B$ sphere at a point with objective at most $J_{\ell,\gamma}(b_\ell)$, contradicting the preceding inequality. Hence

$$\left\| \widehat{\beta}(\lambda, \tau_\ell) \right\|_2 < 3B \qquad (\lambda \ge \lambda_*,\ \ell \le L). \tag{S.5}$$

An interpolating b with $\|b\|_2 \le 3B$ would satisfy

$$-A_\ell \le Ph_{\ell,b} \le -P_n\ell_\ell(b_\ell) + e_n \le -R_\ell(q_\ell) - A_\ell + h_n + e_n < -A_\ell,$$

a contradiction. Thus (S.5) and Theorem 1 imply $\lambda_* > \bar{\lambda}(\tau_\ell)$ for every level.

The grid-range condition gives a deterministic grid point $\lambda^\circ \in \Lambda_n$ with $\lambda_* \le \lambda^\circ \le 2\lambda_*$. For example, choose the smallest $j \ge 0$ with $2^j/n \ge \lambda_*$. Every screened grid contains λ°. Empirical optimality and (S.3) give

$$\begin{aligned} R_\ell\{\psi^\top \widehat{\beta}(\lambda^\circ, \tau_\ell)\} - R_\ell(q_\ell) &\le A_\ell + e_n + \frac{\lambda^\circ B^2}{2n} \\ &\le \frac{3}{2}(e_n + A_B). \end{aligned}$$

For $|Y| \le M$, replacing a prediction by π_M of that prediction cannot increase its check loss. The same risk bound holds for f_{ℓ,λ°.

Conditional on training, there are at most $N = L(J_n + 1)$ candidate predictors, each with validation loss in $[0, 2M]$. A two-sided Hoeffding bound gives, with conditional probability at least $1 - \delta/2$,

$$\max_{\ell,\, \lambda \in \Lambda_{n,\ell}^{\mathrm{scr}}} |\widehat{R}_\ell(f_{\ell,\lambda}) - R_\ell(f_{\ell,\lambda})| \le 2M \sqrt{\frac{\log(4N/\delta)}{2m}}.$$

Compare the validation minimizer with f_{ℓ,λ°. The two deviations and the candidate risk bound prove the stated risk inequality. The total failure probability is at most δ.

Under the Lipschitz condition, F_z is continuous and $F_z(q_\ell(z)) = \tau_\ell$. Fix z and abbreviate $q = q_\ell(z)$ and $\tau = \tau_\ell$. The conditional excess risk is

$$\mathbb{E}\{\rho_\tau(Y - t) - \rho_\tau(Y - q) \mid Z = z\} = \int_q^t \{F_z(s) - \tau\}\, ds.$$

For $t \geq q$, put $u = F_z(t) - \tau$. Then $0 \leq u \leq L_F(t-q)$ and

$$\int_q^t \{F_z(s) - \tau\}\, ds \geq \int_{t-u/L_F}^t \{u - L_F(t-s)\}\, ds = \frac{u^2}{2L_F}.$$

For $t < q$, put $u = \tau - F_z(t)$. The same argument gives

$$\int_t^q \{\tau - F_z(s)\}\, ds \geq \int_t^{t+u/L_F} \{u - L_F(s-t)\}\, ds = \frac{u^2}{2L_F}.$$

Apply these inequalities at $t = \widehat{f}_\ell(z)$ and integrate over Z. Jensen's inequality gives the marginal calibration bound. This completes the proof. □

Under the conditional distribution condition, the comparison risk is positive for every requested level. Indeed, for a conditional τ-quantile q and $h = L_F$, continuity gives $F_z(q) = \tau$, and

$$\begin{aligned}\mathbb{E}\{\rho_\tau(Y-q) \mid Z = z\} &= (1-\tau)\int_{-\infty}^q F_z(s)\, ds + \tau \int_q^\infty \{1 - F_z(s)\}\, ds \\ &\geq \frac{(1-\tau)\tau^2 + \tau(1-\tau)^2}{2h} = \frac{\tau(1-\tau)}{2h}.\end{aligned}$$

The inequality follows by integrating $F_z(q-v) \geq \tau - hv$ for $0 \leq v \leq \tau/h$ and $1 - F_z(q+v) \geq 1 - \tau - hv$ for $0 \leq v \leq (1-\tau)/h$. Thus $r_{\min} \geq \min_\ell \tau_\ell(1-\tau_\ell)/(2L_F) > 0$; this follows from the density bound and requires no additional empirical condition.

The conditions include heteroskedastic models with nonuniform errors. In the trigonometric feature class of Section A.13, let $c_1, c_2 \in \mathbb{R}$ and

$$Y = c_1 \cos\zeta_0 + c_2 \sin\zeta_0 + (\sigma_0 + \sigma_1 \cos\zeta_0)U, \qquad 0 < |\sigma_1| < \sigma_0,$$

where U is independent of Z and has density $f_U(u) = 3(1-u^2)\mathbf{1}(|u| \leq 1)/4$. Writing u_ℓ for its τ_ℓ-quantile, the target coefficient on the first three features is $(\sigma_0 u_\ell, c_1 + \sigma_1 u_\ell, c_2)^\top$. One may take

$$B^2 = \sigma_0^2 + (|c_1| + |\sigma_1|)^2 + c_2^2, \quad M = \sqrt{c_1^2 + c_2^2} + \sigma_0 + |\sigma_1|, \quad L_F = \frac{3}{4(\sigma_0 - |\sigma_1|)}.$$

Here $A_B = 0$ and $r_{\min} = \sigma_0 \min_{\ell \leq L} \mathbb{E}\rho_{\tau_\ell}(U - u_\ell) > 0$. The nuisance features give full row rank by the argument in Section A.13. For fixed model constants, the theorem's population inequalities hold for all large n, its risk bound is of order $n^{-1/2} + \sqrt{\log\log n/m}$, and its calibration bound has the corresponding square-root order.

The theorem uses elementary complexity and validation bounds; it does not assert a new optimal learning rate. Its conclusion is that screening by the observed interpolation boundary preserves an accurate candidate without using the comparison coefficient norm in the procedure. The learning-rate results of Wang & Feng (2024) and the quantile-risk analysis of Steinwart & Christmann (2011) address different aspects of this comparison.

S.4 Convex optimality and interpolation geometry

This section establishes the foundational subgradient optimality conditions for the ridge-penalized check loss, along with the minimum-norm interpolation identities. These geometric properties serve as the building blocks for every exact certificate derived in the subsequent sections.

For a convex function h, ∂h denotes its subdifferential. The following lemma formally reduces the global optimality of the coefficients to a set of stationary conditions and coordinatewise subgradient inclusions, which we will later exploit to construct face certificates.

Lemma 2. *Fix $\tau \in (0,1)$ and $\lambda > 0$. A vector $\beta \in \mathbb{R}^p$ equals $\widehat{\beta}(\lambda,\tau)$ if and only if, with $r = y - X\beta$, there exists $g \in \mathbb{R}^n$ such that*

$$g_i \in \partial\rho_\tau(r_i), \qquad i \leq n, \qquad \lambda\beta = X^\top g,$$

where

$$\partial\rho_\tau(r_i) = \begin{cases}\{\tau\}, & r_i > 0,\\ [\tau - 1, \tau], & r_i = 0,\\ \{\tau - 1\}, & r_i < 0.\end{cases}$$

Proof. The criterion is proper, lower semicontinuous, and λ-strongly convex. The subgradient chain rule for $r = y - X\beta$ gives the necessary and sufficient condition

$$0 \in -X^\top g + \lambda\beta, \qquad g_i \in \partial\rho_\tau(r_i) \quad (i \le n).$$

The three cases of $\partial\rho_\tau$ give the equivalence. □

Next, we formalize the geometric fact that any set of fitted training values can be achieved by a unique coefficient vector of minimal Euclidean norm, and we characterize this minimal norm using the inverse Gram matrix Q.

Lemma 3. *Let $f \in \mathbb{R}^n$. Among all $\beta \in \mathbb{R}^p$ satisfying $X\beta = f$, the unique minimum-norm vector is*

$$\beta(f) = X^\top Q f, \quad \textit{and} \quad \|\beta(f)\|_2^2 = f^\top Q f.$$

Proof. Full row rank gives $K \succ 0$. If $X\beta = f$, then

$$X\{\beta - X^\top Q f\} = 0.$$

Write $\beta = X^\top Q f + v$ with $Xv = 0$. The cross term is zero:

$$(X^\top Q f)^\top v = f^\top Q X v = 0.$$

The norm decomposes as

$$\|\beta\|_2^2 = f^\top Q X X^\top Q f + \|v\|_2^2 = f^\top Q f + \|v\|_2^2 .$$

The unique minimizer is obtained at $v = 0$. □

The following proposition uses this minimum-norm property. It reduces the problem to an equivalent n-dimensional optimization over the fitted training values f. Once the optimal fitted values $\widehat{f}$ are found, the corresponding optimal coefficients are uniquely recovered via $X^\top Q\widehat{f}$.

Let

$$\widehat{f}(\lambda, \tau) = X\widehat{\beta}(\lambda, \tau).$$

Proposition 7. *$\widehat{f}(\lambda, \tau)$ is the unique minimizer, over $f \in \mathbb{R}^n$, of*

$$\sum_{i=1}^{n} \rho_\tau(y_i - f_i) + \frac{\lambda}{2} f^\top Q f,$$

and

$$\widehat{\beta}(\lambda, \tau) = X^\top Q\widehat{f}(\lambda, \tau).$$

Proof. For fixed $f = X\beta$, Lemma 3 gives

$$\min_{\beta : X\beta = f} \|\beta\|_2^2 = f^\top Q f$$

with unique minimizer $X^\top Q f$. Minimizing over f gives the stated criterion, which is strongly convex because $Q \succ 0$. Substitution of $\widehat{f}(\lambda, \tau)$ gives the coefficient identity. □

S.5 Prediction-space and dual geometry

We formulate the training problem in the prediction space and analyze the contraction properties of the K-metric projection. These identities give the sensitivity and monotonicity bounds used by the later stability certificates.

For $v \in \mathbb{R}^n$, write $\|v\|_K = (v^\top K v)^{1/2}$ and $\|v\|_Q = (v^\top Q v)^{1/2}$. For a nonempty closed convex set $C \subset \mathbb{R}^n$, let

$$P_C^K(z) = \arg\min_{h \in C} (h - z)^\top K (h - z).$$

The following lemma proves that the K-metric projection is non-expansive. In practice, this means that with K and C held fixed, distances between dual points cannot increase under projection in the K-norm, which controls the stability of the dual optimizer.

Lemma 4. *Let $C \subset \mathbb{R}^n$ be nonempty, closed, and convex. For $u, v \in \mathbb{R}^n$,*

$$\{P_C^K(u) - P_C^K(v)\}^\top K(u - v) \geq \left\|P_C^K(u) - P_C^K(v)\right\|_K^2,$$

and

$$\left\|P_C^K(u) - P_C^K(v)\right\|_K \leq \|u - v\|_K.$$

Proof. Put $p = P_C^K(u)$ and $q = P_C^K(v)$. The variational inequalities are

$$(h - p)^\top K(p - u) \geq 0, \qquad (h - q)^\top K(q - v) \geq 0 \qquad (h \in C).$$

Taking $h = q$ in the first inequality and $h = p$ in the second gives

$$(q - p)^\top K(p - u) + (p - q)^\top K(q - v) \geq 0.$$

With $d = p - q$, this is

$$d^\top K(u - v) - d^\top K d \geq 0,$$

which proves the first inequality. Cauchy's inequality in the K inner product gives

$$\|d\|_K^2 \leq \|d\|_K \|u - v\|_K.$$

The second inequality follows, with the case $d = 0$ included. □

The following proposition shows that the fitted predictions, the signs of all residuals, and the optimal objective value are all determined by one quantity, the projection residual.

Let $P_\tau^K = P_{\mathcal{B}_\tau}^K$, $g_{\lambda,\tau} = P_\tau^K(\lambda a)$, $\widehat{f}(\lambda, \tau) = X\widehat{\beta}(\lambda, \tau)$, and $r_{\lambda,\tau} = y - \widehat{f}(\lambda, \tau)$. Define

$$N_{\mathcal{B}_\tau}(g) = \{v \in \mathbb{R}^n : v^\top(h - g) \leq 0 \text{ for every } h \in \mathcal{B}_\tau\},$$

and put

$$L = \{i : g_{\lambda,\tau,i} = \tau - 1\}, \qquad U = \{i : g_{\lambda,\tau,i} = \tau\}, \qquad F = (L \cup U)^c,$$

where the complement is taken in $\{1, \ldots, n\}$.

Proposition 8.

$$\widehat{f}(\lambda, \tau) = \lambda^{-1} K g_{\lambda,\tau}, \qquad r_{\lambda,\tau} = \lambda^{-1} K(\lambda a - g_{\lambda,\tau}),$$

and

$$\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^n \rho_\tau(y_i - x_i^\top \beta) + \frac{\lambda}{2} \|\beta\|_2^2 \right\} = \frac{\lambda}{2} y^\top Q y - \frac{1}{2\lambda} \|\lambda a - g_{\lambda,\tau}\|_K^2.$$

Also,

$$K(\lambda a - g_{\lambda,\tau}) \in N_{\mathcal{B}_\tau}(g_{\lambda,\tau}).$$

Equivalently,

$$r_{\lambda,\tau,i} \leq 0 \ (i \in L), \qquad r_{\lambda,\tau,i} = 0 \ (i \in F), \qquad r_{\lambda,\tau,i} \geq 0 \ (i \in U).$$

Proof. Proposition 1 gives the fitted-value and residual identities and the dual criterion

$$g^\top y - \frac{1}{2\lambda} g^\top K g, \qquad g \in \mathcal{B}_\tau.$$

Because $y = Ka$,

$$g^\top y - \frac{1}{2\lambda} g^\top K g = \frac{\lambda}{2} a^\top K a - \frac{1}{2\lambda}(g - \lambda a)^\top K(g - \lambda a).$$

Strong duality gives the value formula at $g = g_{\lambda,\tau}$, and $a^\top K a = y^\top Q y$.

The projection variational inequality gives

$$(h - g_{\lambda,\tau})^\top K(g_{\lambda,\tau} - \lambda a) \geq 0 \qquad (h \in \mathcal{B}_\tau).$$

The box normal cone is nonpositive on L, zero on F, and nonnegative on U. Multiplication by λ^{-1} gives the residual signs. □

The following proposition relates prediction departure at a new covariate to the training residual.

For $x \in \mathbb{R}^p$, put $k_x = Xx$.

Proposition 9.

$$x^\top \widehat{\beta}(\lambda,\tau) = k_x^\top Q \widehat{f}(\lambda,\tau), \qquad \left| x^\top \{\widehat{\beta}(\lambda,\tau) - \widehat{\beta}_0\} \right| \le (k_x^\top Q k_x)^{1/2} \{r_{\lambda,\tau}^\top Q r_{\lambda,\tau}\}^{1/2}.$$

Proof. Proposition 7 gives

$$\widehat{\beta}(\lambda,\tau) = X^\top Q \widehat{f}(\lambda,\tau).$$

Multiplication by x proves the first identity. Also,

$$\widehat{\beta}(\lambda,\tau) - \widehat{\beta}_0 = X^\top Q\{\widehat{f}(\lambda,\tau) - y\} = -X^\top Q r_{\lambda,\tau}.$$

Cauchy's inequality in the Q inner product gives the bound. □

The following proposition gives Lipschitz control of fitted values and coefficients under response perturbations and bounds the corresponding residual change.

Fix X, $\lambda > 0$, and $\tau \in (0,1)$. For $y, \widetilde{y} \in \mathbb{R}^n$, let $\widehat{f}_y, \widehat{\beta}_y, r_y$ and $\widehat{f}_{\widetilde{y}}, \widehat{\beta}_{\widetilde{y}}, r_{\widetilde{y}}$ be the fitted values, coefficients, and residuals obtained from the two response vectors.

Proposition 10.

$$\left\| \widehat{f}_y - \widehat{f}_{\widetilde{y}} \right\|_Q \le \|y - \widetilde{y}\|_Q, \qquad \left\| \widehat{\beta}_y - \widehat{\beta}_{\widetilde{y}} \right\|_2 \le \|y - \widetilde{y}\|_Q, \qquad \|r_y - r_{\widetilde{y}}\|_Q \le \|y - \widetilde{y}\|_Q.$$

Proof. Let $a = Qy$, $\widetilde{a} = Q\widetilde{y}$, $g = P_\tau^K(\lambda a)$, and $\widetilde{g} = P_\tau^K(\lambda \widetilde{a})$. Proposition 8 gives

$$\widehat{f}_y - \widehat{f}_{\widetilde{y}} = \lambda^{-1} K (g - \widetilde{g}).$$

Lemma 4 gives

$$\left\| \widehat{f}_y - \widehat{f}_{\widetilde{y}} \right\|_Q = \lambda^{-1} \|g - \widetilde{g}\|_K \le \lambda^{-1} \|\lambda(a - \widetilde{a})\|_K = \|y - \widetilde{y}\|_Q.$$

The coefficient identity follows because

$$\widehat{\beta}_y - \widehat{\beta}_{\widetilde{y}} = X^\top Q (\widehat{f}_y - \widehat{f}_{\widetilde{y}}), \qquad \left\| X^\top Q v \right\|_2^2 = v^\top Q X X^\top Q v = v^\top Q v.$$

For residuals, firm nonexpansiveness in Lemma 4 gives, with $P = P_\tau^K$,

$$\|(u - Pu) - (v - Pv)\|_K^2 = \|u - v\|_K^2 + \|Pu - Pv\|_K^2 - 2(u - v)^\top K (Pu - Pv) \le \|u - v\|_K^2.$$

Taking $u = \lambda Q y$ and $v = \lambda Q \widetilde{y}$ in

$$r_y - r_{\widetilde{y}} = \lambda^{-1} K \{(u - Pu) - (v - Pv)\}$$

proves the residual bound. □

The following proposition gives Lipschitz bounds for the scaled fitted values and coefficients as functions of the penalty and the quantile level.

Proposition 11. *Fix $y \in \mathbb{R}^n$. For $\lambda, \mu > 0$ and $\tau, \nu \in (0,1)$,*

$$\left\| \lambda \widehat{f}(\lambda,\tau) - \mu \widehat{f}(\mu,\tau) \right\|_Q \le |\lambda - \mu| \, \|y\|_Q, \qquad \left\| \lambda \widehat{\beta}(\lambda,\tau) - \mu \widehat{\beta}(\mu,\tau) \right\|_2 \le |\lambda - \mu| \, \|y\|_Q,$$

and

$$\left\| \lambda \{\widehat{f}(\lambda,\tau) - \widehat{f}(\lambda,\nu)\} \right\|_Q \le |\tau - \nu| \, \|\mathbf{1}_n\|_K, \qquad \left\| \lambda \{\widehat{\beta}(\lambda,\tau) - \widehat{\beta}(\lambda,\nu)\} \right\|_2 \le |\tau - \nu| \, \|\mathbf{1}_n\|_K.$$

Proof. For the penalty bounds, let $g_\lambda = P_\tau^K(\lambda a)$ and $g_\mu = P_\tau^K(\mu a)$. Proposition 8 gives

$$\left\|\lambda\widehat{f}(\lambda,\tau) - \mu\widehat{f}(\mu,\tau)\right\|_Q = \|g_\lambda - g_\mu\|_K .$$

Lemma 4 gives

$$\|g_\lambda - g_\mu\|_K \le |\lambda-\mu|\,\|a\|_K = |\lambda-\mu|\,\|y\|_Q .$$

The coefficient bound is identical because $\lambda\widehat{\beta}(\lambda,\tau) = X^\top g_\lambda$.

Let $\delta = \nu - \tau$. The identity $\mathcal{B}_\nu = \mathcal{B}_\tau + \delta\mathbf{1}_n$ gives

$$P_{\mathcal{B}_\nu}^K(z) = \delta\mathbf{1}_n + P_{\mathcal{B}_\tau}^K(z - \delta\mathbf{1}_n).$$

With $P = P_{\mathcal{B}_\tau}^K$ and $z = \lambda a$, the nonexpansiveness of $I - P$ proved in Proposition 10 gives

$$\left\|P_{\mathcal{B}_\nu}^K(z) - P_{\mathcal{B}_\tau}^K(z)\right\|_K = \|(I-P)z - (I-P)(z-\delta\mathbf{1}_n)\|_K \le |\delta|\,\|\mathbf{1}_n\|_K .$$

The two level bounds follow from $\lambda\widehat{f} = Kg_{\lambda,\tau}$ and $\lambda\widehat{\beta} = X^\top g_{\lambda,\tau}$. □

The following proposition shows that increasing the penalty cannot increase the fitted Q-norm and cannot decrease the check loss or residual Q-norm.

Fix $\tau \in (0,1)$. For $0 < \lambda \le \mu$, put $\widehat{f}_\lambda = \widehat{f}(\lambda,\tau)$ and $r_\lambda = y - \widehat{f}_\lambda$.

Proposition 12.

$$\sum_{i=1}^n \rho_\tau(y_i - \widehat{f}_{\lambda,i}) \le \sum_{i=1}^n \rho_\tau(y_i - \widehat{f}_{\mu,i}), \qquad \widehat{f}_\mu^\top Q\widehat{f}_\mu \le \widehat{f}_\lambda^\top Q\widehat{f}_\lambda, \qquad r_\lambda^\top Q r_\lambda \le r_\mu^\top Q r_\mu.$$

Proof. For $\lambda = \mu$ all comparisons are equalities. Suppose $\lambda < \mu$. Let

$$L(f) = \sum_{i=1}^n \rho_\tau(y_i - f_i), \qquad R(f) = \frac{1}{2} f^\top Q f.$$

Proposition 7 gives

$$L(\widehat{f}_\lambda) + \lambda R(\widehat{f}_\lambda) \le L(\widehat{f}_\mu) + \lambda R(\widehat{f}_\mu), \qquad L(\widehat{f}_\mu) + \mu R(\widehat{f}_\mu) \le L(\widehat{f}_\lambda) + \mu R(\widehat{f}_\lambda).$$

Adding the two inequalities gives

$$(\mu - \lambda)\{R(\widehat{f}_\mu) - R(\widehat{f}_\lambda)\} \le 0,$$

which proves the norm monotonicity. The first optimality inequality then gives $L(\widehat{f}_\lambda) \le L(\widehat{f}_\mu)$.

Let $C = \mathcal{B}_\tau$, $p_t = P_C^K(ta)$, and $D(t) = \|ta - p_t\|_K$. Because $0 \in C$ and C is convex, $(\lambda/\mu)p_\mu \in C$. This gives

$$D(\lambda) \le \|\lambda a - (\lambda/\mu)p_\mu\|_K = \frac{\lambda}{\mu} D(\mu).$$

By Proposition 8,

$$\|r_\lambda\|_Q = \lambda^{-1} D(\lambda), \qquad \|r_\mu\|_Q = \mu^{-1} D(\mu).$$

The residual monotonicity follows. □

S.6 Level-grid geometry and training-fit contrasts

This section translates the continuous extinction interval into practical consequences for finite grids of target quantile levels. We characterize which levels collapse simultaneously under a shared penalty, prove that any zero-sum linear contrast of these collapsed coefficients is strictly zero, and demonstrate a strong absorption property: even if we allow the specified two-component additive model with independent positive ridge penalties, the second component remains zero whenever the base penalty falls inside the standalone base fit's extinction interval.

The following corollary formalizes a key topological consequence of the extinction geometry: on any ordered grid of quantiles, the collapsed levels form either an empty set or a single contiguous block, consisting exactly of the grid levels between λA_+ and $1 - \lambda A_-$, including equality.

Let $T = \{\tau_1 < \cdots < \tau_m\} \subset (0, 1)$. For fixed $\lambda > 0$, define

$$\mathcal{I}_\lambda = \{k : \widehat{\beta}(\lambda, \tau_k) = \widehat{\beta}_0\}.$$

Corollary 7.

$$\mathcal{I}_\lambda = \{k : \lambda A_+ \le \tau_k \le 1 - \lambda A_-\}.$$

The set $\mathcal{I}_\lambda$ is empty or contiguous. Also,

$$\widehat{\beta}(\lambda, \tau_k) \neq \widehat{\beta}_0 \quad (k = 1, \dots, m) \qquad \Longleftrightarrow \qquad \lambda > \max_{1 \le k \le m} \bar{\lambda}(\tau_k).$$

Proof. Theorem 2 gives

$$\lambda \le \bar{\lambda}(\tau_k) \qquad \Longleftrightarrow \qquad \lambda A_+ \le \tau_k \le 1 - \lambda A_-.$$

Theorem 1 gives the coefficient identity. The grid intersects this interval in an empty or contiguous block; its complement gives the final equivalence. □

Next, we prove that when multiple quantile levels undergo extinction simultaneously, the fitted coefficient vectors retain no information about their separation. The following proposition shows that any zero-sum contrast across these coefficients equals zero, and their separation cannot be recovered by applying the same fixed linear transformation to each coefficient vector.

Let $T = \{\tau_1, \dots, \tau_m\} \subset (0, 1)$ with $m \ge 2$, and put

$$B_\lambda = \left[\widehat{\beta}(\lambda, \tau_1), \dots, \widehat{\beta}(\lambda, \tau_m)\right] \in \mathbb{R}^{p \times m}.$$

Proposition 13. *If $0 < \lambda \le \min_{1 \le k \le m} \bar{\lambda}(\tau_k)$, then*

$$B_\lambda = \widehat{\beta}_0 \mathbf{1}_m^\top.$$

For positive integers q, r, every $V \in \mathbb{R}^{q \times p}$, and every $C \in \mathbb{R}^{m \times r}$ satisfying $\mathbf{1}_m^\top C = 0$,

$$V B_\lambda C = 0.$$

Proof. Theorem 1 gives

$$\widehat{\beta}(\lambda, \tau_k) = \widehat{\beta}_0 \qquad (k = 1, \dots, m).$$

The matrix identity follows. Multiplication by C gives

$$B_\lambda C = \widehat{\beta}_0 \mathbf{1}_m^\top C = 0,$$

and multiplication by V preserves the zero matrix. □

Finally, we consider a more flexible two-component additive model, with a base quantile fit and an additive correction for a second quantile fit. The following proposition shows an absorption effect: if the base penalty is small enough to cause extinction, the base model fits the data exactly and the secondary correction term equals zero for every value of its own positive penalty.

Fix $\tau, \nu \in (0, 1)$ and $\lambda_\beta, \lambda_\gamma > 0$. Let

$$(\widehat{\beta}, \widehat{\gamma}) = \arg\min_{\beta, \gamma \in \mathbb{R}^p} \left\{ \sum_{i=1}^n \rho_\tau(y_i - x_i^\top \beta) + \sum_{i=1}^n \rho_\nu(y_i - x_i^\top(\beta + \gamma)) + \frac{\lambda_\beta}{2} \|\beta\|_2^2 + \frac{\lambda_\gamma}{2} \|\gamma\|_2^2 \right\}.$$

Proposition 14. *If $0 < \lambda_\beta \le \bar{\lambda}(\tau)$, then*

$$(\widehat{\beta}, \widehat{\gamma}) = (\widehat{\beta}_0, 0).$$

Proof. Strong convexity makes a stationary point the unique minimizer. At $(\widehat{\beta}_0, 0)$, both residual vectors are zero. Choose

$$g = \lambda_\beta a \in [\tau - 1, \tau]^n, \qquad h = 0 \in [\nu - 1, \nu]^n.$$

The stationarity equations are

$$-X^\top(g + h) + \lambda_\beta \widehat{\beta}_0 = 0, \qquad -X^\top h + \lambda_\gamma 0 = 0.$$

Because $\widehat{\beta}_0 = X^\top a$, these equations hold and establish optimality. □

S.7 WEIGHTED OBSERVATIONS AND QUADRATIC PENALTIES

This section generalizes the extinction identity to accommodate positive observation weights and arbitrary positive definite quadratic penalties. Under these broader conditions, the zero-residual interpolant becomes the one with the minimal norm in the penalty metric, while the coordinatewise extinction boundaries are rescaled by the individual observation weights.

As a geometric preliminary, the following lemma identifies the unique coefficient vector that minimizes the generalized quadratic penalty while maintaining a specific, arbitrary training fit.

Let $H = H^\top \in \mathbb{R}^{p\times p}$ satisfy $H \succ 0$. For $v \in \mathbb{R}^p$, write $\|v\|_H = (v^\top H v)^{1/2}$. Put

$$K_H = XH^{-1}X^\top, \qquad Q_H = K_H^{-1}.$$

Lemma 5. *For $f \in \mathbb{R}^n$, the unique minimizer of $\beta^\top H\beta$ subject to $X\beta = f$ is*

$$\beta_H(f) = H^{-1}X^\top Q_H f, \quad \textit{and} \quad \beta_H(f)^\top H\beta_H(f) = f^\top Q_H f.$$

Proof. For every $u \neq 0$,

$$u^\top K_H u = (X^\top u)^\top H^{-1}(X^\top u) > 0,$$

because $X^\top$ is one-to-one on $\mathbb{R}^n$ and $H^{-1} \succ 0$. The Lagrange equations for minimizing $\beta^\top H\beta$ subject to $X\beta = f$ give

$$H\beta = X^\top \mu, \qquad XH^{-1}X^\top \mu = f.$$

The equations give $\mu = Q_H f$ and $\beta = H^{-1}X^\top Q_H f$. Direct substitution gives the stated minimum value. □

Let $H = H^\top \in \mathbb{R}^{p\times p}$ satisfy $H \succ 0$, and let $w = (w_1, \ldots, w_n)^\top \in (0,\infty)^n$. Put

$$K_H = XH^{-1}X^\top, \qquad Q_H = K_H^{-1}, \qquad a_H = Q_H y, \qquad \widehat{\beta}_{H,0} = H^{-1}X^\top a_H.$$

Define

$$\widehat{\beta}_H(\lambda,\tau) = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^n w_i \rho_\tau(y_i - x_i^\top \beta) + \frac{\lambda}{2}\beta^\top H\beta.$$

Let

$$M_+ = \max_{i:a_{H,i}>0} \frac{a_{H,i}}{w_i}, \qquad M_- = \max_{i:a_{H,i}<0} \frac{|a_{H,i}|}{w_i},$$

with empty maxima equal to 0 and $c/0 = +\infty$ for $c > 0$.

Building on this geometry, the following theorem establishes that introducing heterogeneous observation weights and a generalized quadratic penalty preserves the extinction phenomenon. The coefficient vector still collapses to the generalized minimum-norm interpolant, with coordinate thresholds determined by the ratio $a_{H,i}/w_i$.

Theorem 13. *Under the preceding setup,*

$$\widehat{\beta}_H(\lambda,\tau) = \widehat{\beta}_{H,0} \qquad \Longleftrightarrow \qquad 0 < \lambda \le \frac{\tau}{M_+} \wedge \frac{1-\tau}{M_-}.$$

If $M_+M_- > 0$, this boundary is concave and piecewise linear in τ, with unique maximizer

$$\tau_H^\dagger = \frac{M_+}{M_+ + M_-}.$$

Proof. The weighted check-loss subgradient at zero residual is $\prod_{i=1}^n [w_i(\tau-1), w_i\tau]$. The Karush–Kuhn–Tucker (KKT) condition is

$$g_i \in w_i \partial\rho_\tau(y_i - x_i^\top\beta), \qquad \lambda H\beta = X^\top g.$$

At $\widehat{\beta}_{H,0}$ all residuals are zero and $H\widehat{\beta}_{H,0} = X^\top a_H$. Stationarity at $\widehat{\beta}_{H,0}$ is equivalent to

$$X^\top(g - \lambda a_H) = 0.$$

Injectivity of $X^\top$ gives $g = \lambda a_H$, hence

$$\lambda a_{H,i} \in [w_i(\tau - 1), w_i\tau] \qquad (i \le n).$$

The positive and negative coordinates give the two bounds; zero coordinates impose none. Taking their minimum gives

$$0 < \lambda \le \frac{\tau}{M_+} \wedge \frac{1-\tau}{M_-}.$$

For $M_+M_- > 0$, the increasing and decreasing terms meet at $M_+/(M_+ + M_-)$. This completes the proof. □

S.8 CENTERED QUADRATIC PENALTIES

In Bayesian settings and transfer learning, a quadratic penalty is often centered at a fixed nonzero reference coefficient. The extinction results extend to this penalty.

Let $H = H^\top \in \mathbb{R}^{p\times p}$ satisfy $H \succ 0$, let $b \in \mathbb{R}^p$, and let $w = (w_1, \dots, w_n)^\top \in (0,\infty)^n$. The exact certificate is obtained by replacing y by $y - Xb$ and by using the closest interpolant to b in the H-norm. Put

$$K_H = XH^{-1}X^\top, \qquad Q_H = K_H^{-1}, \qquad \nu_b = y - Xb, \qquad a_{H,b} = Q_H\nu_b,$$

and

$$\widehat{\beta}_{H,b,0} = b + H^{-1}X^\top a_{H,b}.$$

For $\tau \in (0,1)$, define

$$\bar{\lambda}_{H,b,w}(\tau) = \frac{\tau}{M_{H,b,w,+}} \wedge \frac{1-\tau}{M_{H,b,w,-}},$$

where

$$M_{H,b,w,+} = \max_{i:a_{H,b,i}>0} \frac{a_{H,b,i}}{w_i}, \qquad M_{H,b,w,-} = \max_{i:a_{H,b,i}<0} \frac{|a_{H,b,i}|}{w_i},$$

empty maxima are zero, and $c/0 = +\infty$ for $c > 0$. For $\lambda > 0$, set

$$\widehat{\beta}_{H,b}(\lambda,\tau) = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \sum_{i=1}^n w_i\rho_\tau(y_i - x_i^\top\beta) + \frac{\lambda}{2}(\beta - b)^\top H(\beta - b) \right\}.$$

The following theorem proves that centering the penalty at a reference vector b replaces the response vector by the residuals $y - Xb$. The form of the exact extinction interval is preserved, with its boundary recomputed from $y - Xb$. The coefficients now collapse to the interpolant closest to b in the H-norm, which differs from the minimum-norm interpolant when $b \neq 0$.

Theorem 14. *For every $\lambda > 0$ and $\tau \in (0,1)$,*

$$\widehat{\beta}_{H,b}(\lambda,\tau) = \widehat{\beta}_{H,b,0} \quad\Longleftrightarrow\quad 0 < \lambda \le \bar{\lambda}_{H,b,w}(\tau).$$

If $a_{H,b} = 0$, then $\bar{\lambda}_{H,b,w}(\tau) = +\infty$ for every $\tau \in (0,1)$.

Proof. Strong convexity gives uniqueness. Also,

$$X\widehat{\beta}_{H,b,0} = Xb + XH^{-1}X^\top Q_H(y - Xb) = y.$$

At $\widehat{\beta}_{H,b,0}$,

$$H(\widehat{\beta}_{H,b,0} - b) = X^\top a_{H,b}.$$

The weighted zero-residual subgradient set is

$$\prod_{i=1}^n [w_i(\tau - 1), w_i\tau].$$

The KKT condition requires $g \in \mathbb{R}^n$ satisfying

$$g_i \in [w_i(\tau - 1), w_i\tau], \qquad X^\top g = \lambda X^\top a_{H,b}.$$

Injectivity of $X^\top$ gives $g = \lambda a_{H,b}$ and

$$\lambda a_{H,b,i} \in [w_i(\tau - 1), w_i\tau] \qquad (i \le n).$$

Positive and negative coordinates give the stated caps; zero coordinates impose none. Their minimum gives the equivalence, including the infinite boundary when $a_{H,b} = 0$. □

The following corollary shows that an interpolating reference coefficient is the penalized solution at every positive penalty and every quantile level.

Corollary 8. *If $Xb = y$, then*

$$\widehat{\beta}_{H,b}(\lambda,\tau) = b \qquad (\lambda > 0,\ 0 < \tau < 1).$$

Proof. The condition $Xb = y$ gives $\nu_b = 0$, $a_{H,b} = 0$, and $\widehat{\beta}_{H,b,0} = b$. Theorem 14 gives the identity for every positive λ and every $\tau \in (0,1)$. □

The following corollary shows that all levels below a common centered boundary give the same prediction function determined by $\widehat{\beta}_{H,b,0}$.

Corollary 9. *Let $T \subset (0,1)$ be nonempty. If $0 < \lambda \le \inf_{\tau\in T} \bar{\lambda}_{H,b,w}(\tau)$, then*

$$x^\top \widehat{\beta}_{H,b}(\lambda,\tau) = x^\top \widehat{\beta}_{H,b,0} \qquad (x \in \mathbb{R}^p,\ \tau \in T).$$

Proof. For each $\tau \in T$, Theorem 14 gives

$$\widehat{\beta}_{H,b}(\lambda,\tau) = \widehat{\beta}_{H,b,0}.$$

Multiplication by x proves the claim. □

S.8.1 Centered weighted dual and face certificates

Centering changes the response entering the dual problem. Observation weights determine the side lengths of the dual box, while the quadratic penalty determines the projection metric. The following identities determine which residual coordinates can change as the penalty varies. We use the notation from Section S.8. For $G = G^\top \succ 0$ and $v \in \mathbb{R}^n$, write $\|v\|_G = (v^\top G v)^{1/2}$. Define the weighted zero-residual box

$$\mathcal{B}_{\tau,w} = \prod_{i=1}^{n} [w_i(\tau-1), w_i\tau]$$

and let

$$g_{H,b,w}(\lambda,\tau) = P^{K_H}_{\mathcal{B}_{\tau,w}}(\lambda a_{H,b}),$$

where the projection notation is defined in Section S.5. For $r \in \mathbb{R}^n$, write

$$L_{\tau,w}(r) = \sum_{i=1}^{n} w_i \rho_\tau(r_i),$$

and for $\beta \in \mathbb{R}^p$ write

$$\Psi_{H,b,w,\lambda,\tau}(\beta) = L_{\tau,w}(y - X\beta) + \frac{\lambda}{2}(\beta - b)^\top H(\beta - b).$$

The following proposition represents the centered weighted estimator by a K_H-metric projection onto the weighted subgradient box.

Proposition 15. *For every $\lambda > 0$ and $\tau \in (0,1)$,*

$$\widehat{\beta}_{H,b}(\lambda,\tau) = b + \lambda^{-1} H^{-1} X^\top g_{H,b,w}(\lambda,\tau),$$

$$y - X\widehat{\beta}_{H,b}(\lambda,\tau) = \lambda^{-1} K_H\{\lambda a_{H,b} - g_{H,b,w}(\lambda,\tau)\},$$

and

$$\min_{\beta\in\mathbb{R}^p} \Psi_{H,b,w,\lambda,\tau}(\beta) = \frac{\lambda}{2}\nu_b^\top Q_H \nu_b - \frac{1}{2\lambda}\left\|\lambda a_{H,b} - g_{H,b,w}(\lambda,\tau)\right\|^2_{K_H}.$$

Proof. Put $\nu = \nu_b$, $a = a_{H,b}$, $g = g_{H,b,w}(\lambda, \tau)$, and $u = X(\beta - b)$. Lemma 5, applied to $\beta - b$, gives the prediction-space criterion

$$\min_{u \in \mathbb{R}^n} \left\{ L_{\tau,w}(\nu - u) + \frac{\lambda}{2} u^\top Q_H u \right\},$$

with coefficient $b + H^{-1} X^\top Q_H u$. For each $r \in \mathbb{R}^n$,

$$L_{\tau,w}(r) = \max_{h \in \mathcal{B}_{\tau,w}} h^\top r.$$

The projection variational inequality gives

$$K_H(\lambda a - g) \in N_{\mathcal{B}_{\tau,w}}(g),$$

where $N_{\mathcal{B}_{\tau,w}}(g)$ is the normal cone. Set

$$r = \lambda^{-1} K_H(\lambda a - g), \qquad u = K_H a - r = \lambda^{-1} K_H g.$$

Then $\lambda r \in N_{\mathcal{B}_{\tau,w}}(g)$ and $g \in \partial L_{\tau,w}(\nu - u)$. The subgradient condition for the prediction-space criterion is

$$0 \in -g + \lambda Q_H u,$$

which holds by definition. Strong convexity gives the unique fit u and hence the coefficient and residual identities.

Because g maximizes $h^\top(\nu - u)$ over $h \in \mathcal{B}_{\tau,w}$,

$$L_{\tau,w}(\nu - u) = g^\top(\nu - u).$$

At the optimizer,

$$L_{\tau,w}(\nu - u) + \frac{\lambda}{2} u^\top Q_H u = g^\top \nu - g^\top u + \frac{\lambda}{2} u^\top Q_H u = g^\top \nu - \frac{1}{2\lambda} g^\top K_H g.$$

Using $\nu = K_H a$ gives

$$g^\top \nu - \frac{1}{2\lambda} g^\top K_H g = \frac{\lambda}{2} a^\top K_H a - \frac{1}{2\lambda} (\lambda a - g)^\top K_H (\lambda a - g).$$

Because $a^\top K_H a = \nu_b^\top Q_H \nu_b$, the value identity follows. This completes the proof. □

The following proposition shows that one K_H-projection distance simultaneously determines the coefficient departure, residual norm, and objective gap.

Let

$$d_{H,b,w,\lambda,\tau} = \lambda a_{H,b} - g_{H,b,w}(\lambda, \tau).$$

Proposition 16. *For every $\lambda > 0$ and $\tau \in (0, 1)$,*

$$\widehat{\beta}_{H,b}(\lambda, \tau) - \widehat{\beta}_{H,b,0} = -\lambda^{-1} H^{-1} X^\top d_{H,b,w,\lambda,\tau},$$

$$y - X\widehat{\beta}_{H,b}(\lambda, \tau) = \lambda^{-1} K_H d_{H,b,w,\lambda,\tau},$$

$$\{\widehat{\beta}_{H,b}(\lambda, \tau) - \widehat{\beta}_{H,b,0}\}^\top H \{\widehat{\beta}_{H,b}(\lambda, \tau) - \widehat{\beta}_{H,b,0}\} = \lambda^{-2} \left\| d_{H,b,w,\lambda,\tau} \right\|_{K_H}^2,$$

$$\{y - X\widehat{\beta}_{H,b}(\lambda, \tau)\}^\top Q_H \{y - X\widehat{\beta}_{H,b}(\lambda, \tau)\} = \lambda^{-2} \left\| d_{H,b,w,\lambda,\tau} \right\|_{K_H}^2,$$

and

$$\Psi_{H,b,w,\lambda,\tau}(\widehat{\beta}_{H,b,0}) - \Psi_{H,b,w,\lambda,\tau}(\widehat{\beta}_{H,b}(\lambda, \tau)) = \frac{1}{2\lambda} \left\| d_{H,b,w,\lambda,\tau} \right\|_{K_H}^2.$$

Proof. Proposition 15 gives

$$\widehat{\beta}_{H,b}(\lambda, \tau) = b + \lambda^{-1} H^{-1} X^\top (\lambda a_{H,b} - d_{H,b,w,\lambda,\tau}) = \widehat{\beta}_{H,b,0} - \lambda^{-1} H^{-1} X^\top d_{H,b,w,\lambda,\tau}.$$

The residual identity is the second identity in Proposition 15. The quadratic identity follows from

$$(H^{-1} X^\top d)^\top H (H^{-1} X^\top d) = d^\top X H^{-1} X^\top d = d^\top K_H d, \quad (K_H d)^\top Q_H (K_H d) = d^\top K_H d.$$

At $\widehat{\beta}_{H,b,0}$, the residual vector is zero and

$$\Psi_{H,b,w,\lambda,\tau}(\widehat{\beta}_{H,b,0}) = \frac{\lambda}{2} \nu_b^\top Q_H \nu_b.$$

The value identity in Proposition 15 gives the stated objective gap. This completes the proof. □

For index sets I, J, v_I denotes the coordinates of a vector v indexed by I, and M_{IJ} denotes the submatrix of M with rows in I and columns in J. A single row-index set in X_I selects those rows of X. The following theorem converts every valid lower and upper dual face into exact coefficient and residual formulas and recovers every realized dual face.

Fix $\tau \in (0,1)$. Let $L, U \subset \{1,\dots,n\}$ be disjoint, put $A = L \cup U$, put $F = A^c$, and assume $A \neq \emptyset$. Define

$$c_i = \begin{cases} w_i(\tau - 1), & i \in L, \\ w_i\tau, & i \in U. \end{cases}$$

If F is nonempty, put

$$S_A = K_{H,AA} - K_{H,AF}K_{H,FF}^{-1}K_{H,FA},$$

$$d_A(\lambda) = a_{H,b,A} - \lambda^{-1}c_A, \qquad d_F(\lambda) = -K_{H,FF}^{-1}K_{H,FA}d_A(\lambda),$$

$$g_A(\lambda) = c_A, \qquad g_F(\lambda) = \lambda\{a_{H,b,F} - d_F(\lambda)\},$$

and

$$r_A(\lambda) = S_A d_A(\lambda), \qquad r_F(\lambda) = 0.$$

If $F = \emptyset$, the formulas involving F are absent, and

$$S_A = K_{H,AA}, \qquad d_A(\lambda) = a_{H,b,A} - \lambda^{-1}c_A, \qquad g_A(\lambda) = c_A,$$

$$r_A(\lambda) = S_A d_A(\lambda).$$

Let $d(\lambda)$, $g(\lambda)$, and $r(\lambda)$ denote the vectors with these active coordinates and, when $F \neq \emptyset$, the free coordinates defined above. Let $I_{H,b,w,L,U}$ be the set of $\lambda > 0$ such that

$$w_i(\tau - 1) < g_i(\lambda) < w_i\tau \quad (i \in F),$$

$$r_i(\lambda) \leq 0 \quad (i \in L), \qquad r_i(\lambda) \geq 0 \quad (i \in U).$$

Theorem 15. *For every $\lambda \in I_{H,b,w,L,U}$,*

$$\widehat{\beta}_{H,b}(\lambda,\tau) = \widehat{\beta}_{H,b,0} - H^{-1}X^\top d(\lambda), \qquad y - X\widehat{\beta}_{H,b}(\lambda,\tau) = r(\lambda).$$

If $g_{H,b,w}(\lambda,\tau)$ has lower active set L, upper active set U, and free set F at a given λ, then $\lambda \in I_{H,b,w,L,U}$ and the formulas above hold.

Proof. Assume first that $\lambda \in I_{H,b,w,L,U}$. Put

$$\beta_\lambda = \widehat{\beta}_{H,b,0} - H^{-1}X^\top d(\lambda).$$

Because $X\widehat{\beta}_{H,b,0} = y$,

$$y - X\beta_\lambda = K_H d(\lambda) = r(\lambda),$$

where the last equality follows from the Schur-complement definitions. Also,

$$\lambda H(\beta_\lambda - b) = \lambda X^\top\{a_{H,b} - d(\lambda)\} = X^\top g(\lambda).$$

The defining box constraints and residual signs give $g_i(\lambda) \in w_i\partial\rho_\tau(r_i(\lambda))$. Thus KKT optimality gives $\beta_\lambda = \widehat{\beta}_{H,b}(\lambda,\tau)$.

For the converse, let $g = g_{H,b,w}(\lambda,\tau)$ have the stated active sets and set

$$d = a_{H,b} - \lambda^{-1}g.$$

Proposition 15 gives

$$\widehat{\beta}_{H,b}(\lambda,\tau) = \widehat{\beta}_{H,b,0} - H^{-1}X^\top d, \qquad r = y - X\widehat{\beta}_{H,b}(\lambda,\tau) = K_H d.$$

For $i \in F$, interior subgradient membership gives $r_i = 0$. The identity $g_A = c_A$ gives $d_A = a_{H,b,A} - \lambda^{-1}c_A$. If $F \neq \emptyset$, the equation $r_F = 0$ gives

$$K_{H,FA}d_A + K_{H,FF}d_F = 0,$$

which determines d_F. Substitution gives the remaining formulas, and the KKT signs give $\lambda \in I_{H,b,w,L,U}$. This completes the proof. □

The following corollary gives an explicit positive penalty radius on which one binding coordinate is the only nonzero residual coordinate.

Fix $\tau \in (0,1)$, let $\lambda_0 = \bar{\lambda}_{H,b,w}(\tau) < \infty$, and suppose the minimum defining $\bar{\lambda}_{H,b,w}(\tau)$ is attained at the unique index j. Put $a = a_{H,b}$ and $q_{k\ell} = e_k^\top Q_H e_\ell$. For $k \neq j$, define

$$v_k = a_k - \frac{q_{kj}}{q_{jj}} a_j, \qquad \eta_k = \begin{cases} \{w_k\tau - \lambda_0 a_k\}/v_k, & v_k > 0, \\ \{\lambda_0 a_k - w_k(\tau-1)\}/(-v_k), & v_k < 0, \\ +\infty, & v_k = 0, \end{cases}$$

and let $\eta_* = \min_{k \neq j} \eta_k$, with the empty minimum equal to $+\infty$.

Corollary 10. *For $0 < s < \eta_*$ and $\lambda = \lambda_0 + s$,*

$$\widehat{\beta}_{H,b}(\lambda,\tau) = \widehat{\beta}_{H,b,0} - \frac{s a_j}{\lambda q_{jj}} H^{-1} X^\top Q_H e_j, \quad y - X\widehat{\beta}_{H,b}(\lambda,\tau) = \frac{s a_j}{\lambda q_{jj}} e_j.$$

Proof. The unique binding condition gives $a_j \neq 0$, $q_{jj} > 0$, and

$$w_k(\tau-1) < \lambda_0 a_k < w_k \tau \qquad (k \neq j).$$

Put

$$t = \frac{s a_j}{\lambda q_{jj}}, \qquad d = t Q_H e_j, \qquad g = \lambda(a - d).$$

Then

$$K_H d = t e_j, \qquad \widehat{\beta}_{H,b,0} - H^{-1} X^\top d = \widehat{\beta}_{H,b,0} - \frac{s a_j}{\lambda q_{jj}} H^{-1} X^\top Q_H e_j.$$

For the binding coordinate,

$$g_j = \lambda(a_j - t q_{jj}) = \lambda_0 a_j.$$

If $a_j > 0$, then $g_j = w_j \tau$ and $t > 0$. If $a_j < 0$, then $g_j = w_j(\tau - 1)$ and $t < 0$. For $k \neq j$,

$$g_k = \lambda_0 a_k + s\left(a_k - \frac{q_{kj}}{q_{jj}} a_j\right) = \lambda_0 a_k + s v_k.$$

The condition $0 < s < \eta_*$ gives

$$w_k(\tau-1) < g_k < w_k \tau \qquad (k \neq j).$$

If $a_j > 0$, take $L = \emptyset$ and $U = \{j\}$. If $a_j < 0$, take $L = \{j\}$ and $U = \emptyset$. Then $K_H d = t e_j$ has the required active sign, and all other coordinates are free. Theorem 15 proves the coefficient and residual identities. This completes the proof. □

The following proposition shows that a fixed centered weighted face is valid on an interval and gives affine formulas in λ^{-1} on that interval.

Proposition 17. *Use the notation of Theorem 15. The set $I_{H,b,w,L,U}$ is empty or an interval in $(0,\infty)$. On $I_{H,b,w,L,U}$ there exist vectors $b_0, b_1 \in \mathbb{R}^p$, $s_0, s_1, h_0, h_1 \in \mathbb{R}^n$ such that*

$$\widehat{\beta}_{H,b}(\lambda,\tau) = b_0 + \lambda^{-1} b_1, \qquad y - X\widehat{\beta}_{H,b}(\lambda,\tau) = s_0 + \lambda^{-1} s_1, \qquad g_{H,b,w}(\lambda,\tau) = h_0 + \lambda h_1.$$

Also, $\lambda H\{\widehat{\beta}_{H,b}(\lambda,\tau) - b\}$ and $\lambda X \widehat{\beta}_{H,b}(\lambda,\tau)$ are affine functions of λ on $I_{H,b,w,L,U}$.

Proof. If $F \neq \emptyset$, Theorem 15 gives

$$d_A(\lambda) = a_{H,b,A} - \lambda^{-1} c_A,$$

$$d_F(\lambda) = -K_{H,FF}^{-1} K_{H,FA} a_{H,b,A} + \lambda^{-1} K_{H,FF}^{-1} K_{H,FA} c_A.$$

If $F = \emptyset$, only the formula for $d_A(\lambda)$ remains. In either case, $d(\lambda) = d_0 + \lambda^{-1} d_1$ for fixed vectors d_0, d_1. Theorem 15 gives

$$\widehat{\beta}_{H,b}(\lambda,\tau) = \widehat{\beta}_{H,b,0} - H^{-1} X^\top d(\lambda), \qquad y - X\widehat{\beta}_{H,b}(\lambda,\tau) = K_H d(\lambda),$$

which proves the affine forms for $\widehat{\beta}_{H,b}$ and the residual vector. Also,

$$g_{H,b,w}(\lambda,\tau) = \lambda\{a_{H,b} - d(\lambda)\} = \lambda(a_{H,b} - d_0) - d_1,$$

which proves the affine form for $g_{H,b,w}$.

The free-coordinate constraints have the form

$$w_i(\tau - 1) < h_{0,i} + \lambda h_{1,i} < w_i\tau.$$

The active residual sign constraints are

$$s_{0,i} + \lambda^{-1} s_{1,i} \le 0 \qquad \text{or} \qquad s_{0,i} + \lambda^{-1} s_{1,i} \ge 0.$$

After multiplying the latter by $\lambda > 0$, all constraints are linear in λ, giving an interval. The final affine assertions follow from

$$\lambda H\{\widehat{\beta}_{H,b}(\lambda,\tau) - b\} = X^\top g_{H,b,w}(\lambda,\tau), \quad \lambda X\widehat{\beta}_{H,b}(\lambda,\tau) = \lambda Xb + K_H g_{H,b,w}(\lambda,\tau).$$

This completes the proof. □

The following corollary covers the positive penalty path by finitely many face intervals with the affine formulas of Proposition 17.

Corollary 11. *Fix $\tau \in (0,1)$. The half-line $(0,\infty)$ is covered by finitely many intervals on which $\widehat{\beta}_{H,b}(\lambda,\tau)$ and $y - X\widehat{\beta}_{H,b}(\lambda,\tau)$ are affine functions of λ^{-1}, while $g_{H,b,w}(\lambda,\tau)$, $\lambda H\{\widehat{\beta}_{H,b}(\lambda,\tau) - b\}$, and $\lambda X\widehat{\beta}_{H,b}(\lambda,\tau)$ are affine functions of λ.*

Proof. There are finitely many disjoint pairs $L, U \subset \{1,\dots,n\}$. For a fixed $\lambda > 0$, take

$$L = \{i : g_{H,b,w,i}(\lambda,\tau) = w_i(\tau-1)\}, \qquad U = \{i : g_{H,b,w,i}(\lambda,\tau) = w_i\tau\}.$$

If $L \cup U = \emptyset$, then $g_{H,b,w}(\lambda,\tau) = \lambda a_{H,b}$, and Theorem 14 gives

$$\widehat{\beta}_{H,b}(\lambda,\tau) = \widehat{\beta}_{H,b,0}, \qquad y - X\widehat{\beta}_{H,b}(\lambda,\tau) = 0.$$

The stated affine forms hold on the interval

$$\{\lambda > 0 : w_i(\tau-1) < \lambda a_{H,b,i} < w_i\tau \text{ for every } i \le n\}.$$

If $L \cup U \neq \emptyset$, Theorem 15 and Proposition 17 give the same affine forms on $I_{H,b,w,L,U}$. The finite collection of these intervals covers $(0,\infty)$. □

The following proposition converts the minimum weighted-box slack into a deterministic response-perturbation radius that preserves extinction over T.

Proposition 18. *Fix $\lambda > 0$ and a nonempty set $T \subset (0,1)$. Put*

$$m_{H,b,w}(T) = \inf_{\tau\in T} \min_{i\le n}\{w_i\tau - \lambda a_{H,b,i},\ \lambda a_{H,b,i} - w_i(\tau-1)\}.$$

Assume $m_{H,b,w}(T) \ge 0$. Let $\widetilde{y} \in \mathbb{R}^n$ and set $\widetilde{a}_{H,b} = Q_H(\widetilde{y} - Xb)$. If

$$\|Q_H(\widetilde{y} - y)\|_\infty \le \lambda^{-1} m_{H,b,w}(T),$$

then the estimator formed with response $\widetilde{y}$ and the same X, H, b, w satisfies

$$\widehat{\beta}^{\widetilde{y}}_{H,b}(\lambda,\tau) = b + H^{-1}X^\top \widetilde{a}_{H,b} \qquad (\tau \in T).$$

Proof. For every $\tau \in T$ and every $i \le n$, the definition of $m_{H,b,w}(T)$ gives

$$w_i(\tau-1) + m_{H,b,w}(T) \le \lambda a_{H,b,i} \le w_i\tau - m_{H,b,w}(T).$$

The perturbation bound gives

$$|\lambda(\widetilde{a}_{H,b,i} - a_{H,b,i})| \le m_{H,b,w}(T).$$

Combining the two inequalities gives

$$w_i(\tau-1) \le \lambda\widetilde{a}_{H,b,i} \le w_i\tau \qquad (i \le n,\ \tau \in T).$$

Theorem 14, applied with response $\widetilde{y}$, proves the stated identity. □

The following proposition shows that simultaneous extinction over an interval or finite grid depends only on its smallest and largest levels.

Let $0 < \tau_- \le \tau_+ < 1$ and $I = [\tau_-, \tau_+]$. Define

$$\Lambda_{H,b,w}(S) = \inf_{\tau \in S} \bar{\lambda}_{H,b,w}(\tau)$$

for nonempty $S \subset (0,1)$.

Proposition 19. *With the preceding definitions,*

$$\Lambda_{H,b,w}(I) = \frac{\tau_-}{M_{H,b,w,+}} \wedge \frac{1-\tau_+}{M_{H,b,w,-}}.$$

For every $\lambda > 0$,

$$\widehat{\beta}_{H,b}(\lambda, \tau) = \widehat{\beta}_{H,b,0} \quad (\tau \in I) \quad \iff \quad 0 < \lambda \le \Lambda_{H,b,w}(I).$$

If $T = \{\tau_1 < \cdots < \tau_m\} \subset (0,1)$, then

$$\Lambda_{H,b,w}(T) = \frac{\tau_1}{M_{H,b,w,+}} \wedge \frac{1-\tau_m}{M_{H,b,w,-}},$$

and

$$\widehat{\beta}_{H,b}(\lambda, \tau_k) = \widehat{\beta}_{H,b,0} \quad (k = 1, \ldots, m) \quad \iff \quad 0 < \lambda \le \Lambda_{H,b,w}(T).$$

Proof. Theorem 14 gives

$$\lambda \le \bar{\lambda}_{H,b,w}(\tau) \quad \iff \quad \lambda M_{H,b,w,+} \le \tau, \qquad \lambda M_{H,b,w,-} \le 1 - \tau,$$

using the zero-denominator convention in Section S.8. Requiring these two inequalities for all $\tau \in I$ is equivalent to

$$\lambda M_{H,b,w,+} \le \tau_-, \qquad \lambda M_{H,b,w,-} \le 1 - \tau_+.$$

This is exactly $0 < \lambda \le \Lambda_{H,b,w}(I)$. Theorem 14 gives the coefficient identity on I.

The same calculation at the extreme levels τ_1 and τ_m gives the finite-grid conclusions. □

The following proposition identifies the extinct grid block and shows that zero-sum contrasts supported on this block equal zero.

Let $T = \{\tau_1 < \cdots < \tau_m\} \subset (0,1)$ and let $\lambda > 0$. Define

$$\mathcal{K}_{H,b,w,\lambda}(T) = \{k : \widehat{\beta}_{H,b}(\lambda, \tau_k) = \widehat{\beta}_{H,b,0}\}.$$

Proposition 20. *If $M_{H,b,w,+} = M_{H,b,w,-} = 0$, then $\mathcal{K}_{H,b,w,\lambda}(T) = \{1, \ldots, m\}$. If $M_{H,b,w,+} + M_{H,b,w,-} > 0$, then*

$$\mathcal{K}_{H,b,w,\lambda}(T) = \{k : \lambda M_{H,b,w,+} \le \tau_k \le 1 - \lambda M_{H,b,w,-}\}.$$

For positive integers N, q, any $Z \in \mathbb{R}^{N \times p}$, and any $C \in \mathbb{R}^{m \times q}$ satisfying

$$C_{k\ell} = 0 \quad (k \notin \mathcal{K}_{H,b,w,\lambda}(T),\ \ell \le q), \qquad \sum_{k=1}^{m} C_{k\ell} = 0 \quad (\ell \le q),$$

we have

$$Z[\widehat{\beta}_{H,b}(\lambda, \tau_1), \ldots, \widehat{\beta}_{H,b}(\lambda, \tau_m)]C = 0.$$

Proof. If $M_{H,b,w,+} = M_{H,b,w,-} = 0$, then $a_{H,b} = 0$, and Theorem 14 gives $\bar{\lambda}_{H,b,w}(\tau) = +\infty$ for every $\tau \in (0,1)$. This proves the first claim.

Assume $M_{H,b,w,+} + M_{H,b,w,-} > 0$. By Theorem 14,

$$\widehat{\beta}_{H,b}(\lambda, \tau_k) = \widehat{\beta}_{H,b,0} \quad \iff \quad \lambda M_{H,b,w,+} \le \tau_k \le 1 - \lambda M_{H,b,w,-}.$$

This proves the formula for $\mathcal{K}_{H,b,w,\lambda}(T)$.

The columns indexed by $\mathcal{K}_{H,b,w,\lambda}(T)$ equal $\widehat{\beta}_{H,b,0}$; all others multiply zero rows of C. Thus

$$Z[\widehat{\beta}_{H,b}(\lambda, \tau_1), \ldots, \widehat{\beta}_{H,b}(\lambda, \tau_m)]C = Z\widehat{\beta}_{H,b,0}\mathbf{1}_m^\top C = 0.$$

This proves the contrast identity. □

S.9 CONVEX RESIDUAL LOSSES

In this section, we isolate the nonsmooth property of the check loss that produces extinction. We demonstrate that any arbitrary proper closed convex residual loss that is finite in a neighborhood of zero admits an identical extinction certificate based on its interval subdifferential at zero. An entire interval of small positive penalties is extinct when the ray $\{\lambda a : \lambda > 0\}$ stays inside the product of these subdifferentials near zero. This geometric perspective immediately reveals a stark contrast: while non-smooth losses can keep the fit at the minimum-norm interpolant over an entire continuous interval of penalties, a loss function that is differentiable at zero (such as squared error) cannot sustain extinction for more than a single, isolated penalty value when $a \neq 0$.

Let $\phi_i : \mathbb{R} \to (-\infty, \infty]$ be proper closed convex functions, finite on a neighborhood of 0, and define

$$\widehat{\beta}_\phi(\lambda) = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^n \phi_i(y_i - x_i^\top \beta) + \frac{\lambda}{2} \|\beta\|_2^2 \right\}, \qquad D = \prod_{i=1}^n \partial\phi_i(0).$$

The following proposition shows that, for a general convex residual loss, the collapse of coefficients to the minimum-norm interpolant is equivalent to the subgradient membership condition $\lambda a \in \prod_{i=1}^n \partial\phi_i(0)$.

Proposition 21. *For every $\lambda > 0$,*

$$\widehat{\beta}_\phi(\lambda) = \widehat{\beta}_0 \qquad \Longleftrightarrow \qquad \lambda a \in D.$$

If $\partial\phi_i(0) = [L_i, U_i]$, then extinction is equivalent to

$$\lambda a_i \in [L_i, U_i] \qquad (i \leq n).$$

Proof. Lower semicontinuity, coercivity, and strong convexity give a unique minimizer. At $\widehat{\beta}_0 = X^\top a$, the residual is zero and optimality requires

$$0 \in -X^\top g + \lambda X^\top a \qquad \text{for some } g \in D.$$

Injectivity of $X^\top$ makes this equivalent to $\lambda a \in D$. Interval subdifferentials give the coordinate condition. □

The following corollary shows that, when $a \neq 0$, losses differentiable at zero cannot produce extinction over an interval of positive penalties. Their subdifferential at zero is a single point, and the ray $\{\lambda a : \lambda > 0\}$ meets it at no more than one penalty. If $a = 0$ and every derivative at zero is zero, extinction holds for every positive penalty. For squared error and $a \neq 0$, no positive penalty causes extinction.

Corollary 12. *Assume the conditions of Proposition 21. If each ϕ_i is differentiable at zero, put $d_i = \phi_i'(0)$ and $d = (d_1, \ldots, d_n)^\top$. Then*

$$\widehat{\beta}_\phi(\lambda) = \widehat{\beta}_0 \qquad \Longleftrightarrow \qquad \lambda a = d.$$

If $a \neq 0$, the set of positive penalties satisfying this identity contains at most one point. If $a = 0$ and $d = 0$, the identity holds for every $\lambda > 0$.

Proof. Differentiability gives $D = \{d\}$. Proposition 21 gives the equivalence $\lambda a = d$. If the equality holds at two distinct positive penalties, subtracting the two vector equalities gives $a = 0$. The remaining cases follow from the same equivalence. □

S.10 ACTIVE-SET ALGEBRA

In this section, we derive fully explicit, closed-form formulas for the fitted coefficients, dual subgradients, and primal residuals on every face of the dual box. Indexing these formulas by their lower and upper active sets upgrades the local-departure characterization of Proposition 2 into exact, global optimizer certificates with computable validity radii.

We first formalize the exact geometric conditions required for optimality on any specific dual face. The following lemma simplifies the global KKT conditions, showing that, for a feasible dual point and its associated coefficient vector, optimality reduces to zero residuals on the free coordinates and nonpositive or nonnegative residuals on the lower or upper active coordinates, respectively.

Lemma 6. *Fix $\tau \in (0,1)$ and $\lambda > 0$. Let $g \in \mathcal{B}_\tau$, $\beta = \lambda^{-1} X^\top g$, and $r = y - X\beta$. Then $\beta = \widehat{\beta}(\lambda, \tau)$ if and only if*

$$r_i = 0 \quad \text{when } \tau - 1 < g_i < \tau, \qquad r_i \le 0 \quad \text{when } g_i = \tau - 1, \qquad r_i \ge 0 \quad \text{when } g_i = \tau.$$

Proof. By definition, $\lambda\beta = X^\top g$. If the stated signs hold, then $g_i \in \partial\rho_\tau(r_i)$ for every i, and Lemma 2 gives $\beta = \widehat{\beta}(\lambda, \tau)$. For the reverse implication, if $\beta = \widehat{\beta}(\lambda, \tau)$, then Lemma 2 gives a vector g' with $g'_i \in \partial\rho_\tau(r_i)$ for every i and $X^\top g' = \lambda\beta = X^\top g$; injectivity of $X^\top$, which follows from the rank-n condition, gives $g' = g$, and the three cases of $\partial\rho_\tau(r_i)$ are the stated signs. □

Using this sign-matching condition, the following theorem gives exact matrix formulas for the coefficients and residuals on every valid dual face. These formulas compute the fitted coefficient vector as it departs from the interpolant beyond the extinction boundary.

Fix $\tau \in (0,1)$. Let $L, U \subset \{1, \dots, n\}$ be disjoint, put $A = L \cup U$ and $F = A^c$, and assume $A \neq \emptyset$. Define

$$c_i = \begin{cases} \tau - 1, & i \in L, \\ \tau, & i \in U. \end{cases}$$

If $F \neq \emptyset$, put

$$S_A = K_{AA} - K_{AF} K_{FF}^{-1} K_{FA}.$$

For $\lambda > 0$, define

$$d_A(\lambda) = a_A - \lambda^{-1} c_A, \qquad d_F(\lambda) = -K_{FF}^{-1} K_{FA} d_A(\lambda),$$

$$g_A(\lambda) = c_A, \qquad g_F(\lambda) = \lambda\{a_F - d_F(\lambda)\}, \qquad r_F(\lambda) = 0, \qquad r_A(\lambda) = S_A d_A(\lambda).$$

If $F = \emptyset$, the formulas involving F are absent, and

$$S_A = K_{AA}, \qquad d_A(\lambda) = a_A - \lambda^{-1} c_A, \qquad g_A(\lambda) = c_A, \qquad r_A(\lambda) = S_A d_A(\lambda).$$

Let $d(\lambda)$, $g(\lambda)$, and $r(\lambda)$ denote the vectors with these active coordinates and, when $F \neq \emptyset$, the free coordinates defined above. Let $I_{L,U}$ be the set of $\lambda > 0$ such that

$$\tau - 1 < g_i(\lambda) < \tau \quad (i \in F), \qquad r_i(\lambda) \le 0 \quad (i \in L), \qquad r_i(\lambda) \ge 0 \quad (i \in U).$$

Theorem 16. *For every $\lambda \in I_{L,U}$,*

$$\widehat{\beta}(\lambda, \tau) = \widehat{\beta}_0 - X^\top d(\lambda), \qquad y - X\widehat{\beta}(\lambda, \tau) = r(\lambda).$$

If $g_{\lambda,\tau}$ has lower active set L, upper active set U, and free set F at a given λ, then $\lambda \in I_{L,U}$ and the formulas above hold.

Proof. Let $\lambda \in I_{L,U}$ and use the defined vectors d, g, and r. On the free coordinates,

$$K_{FF} d_F(\lambda) + K_{FA} d_A(\lambda) = 0.$$

Equivalently,

$$\{K(g(\lambda) - \lambda a)\}_F = 0,$$

because $g(\lambda) - \lambda a = -\lambda d(\lambda)$. The active coordinates satisfy $g_A(\lambda) = c_A$ by construction. The coefficient candidate is

$$\beta_\lambda = \lambda^{-1} X^\top g(\lambda) = X^\top\{a - d(\lambda)\} = \widehat{\beta}_0 - X^\top d(\lambda).$$

Its residual is $y - X\beta_\lambda = Kd(\lambda)$. Its blocks are 0 on F and $S_A d_A(\lambda)$ on A. The defining inequalities and Lemma 6 give $\beta_\lambda = \widehat{\beta}(\lambda, \tau)$.

Now take L, U, F from the actual projection $g_{\lambda,\tau}$. Put $d = a - \lambda^{-1} g_{\lambda,\tau}$. On A, the active constraints give $d_A = a_A - \lambda^{-1} c_A$. On F, the box-projection normal equations give $\{K(g_{\lambda,\tau} - \lambda a)\}_F = 0$, which is

$$K_{FF} d_F + K_{FA} d_A = 0.$$

For $F \neq \emptyset$, invertibility of K_{FF} gives d_F. The remaining formulas follow from

$$\widehat{\beta}(\lambda, \tau) = \lambda^{-1} X^\top g_{\lambda,\tau} = X^\top\{a - d\}, \qquad y - X\widehat{\beta}(\lambda, \tau) = Kd.$$

The KKT sign conditions give $\lambda \in I_{L,U}$. □

The following corollary gives an explicit positive radius on which a unique binding coordinate yields the one-coordinate departure formula.

Fix $\tau \in (0,1)$, let $\lambda_0 = \bar{\lambda}(\tau) < \infty$, and suppose the minimum defining $\bar{\lambda}(\tau)$ is attained at the unique index j. Write $q_{k\ell} = e_k^\top Q e_\ell$, put $c = \lambda_0 a_j$, and define

$$v_k = a_k - \frac{q_{kj}}{q_{jj}} a_j, \qquad k \neq j.$$

For $k \neq j$, define

$$\eta_k = \begin{cases} (\tau - \lambda_0 a_k)/v_k, & v_k > 0, \\ (\lambda_0 a_k - \tau + 1)/(-v_k), & v_k < 0, \\ +\infty, & v_k = 0, \end{cases} \qquad \eta_* = \min_{k \neq j} \eta_k,$$

where the minimum over an empty index set is $+\infty$.

Corollary 13. *Then $\eta_* > 0$. For every $0 < s < \eta_*$, with $\lambda = \lambda_0 + s$,*

$$\widehat{\beta}(\lambda, \tau) = \widehat{\beta}_0 - \frac{s a_j}{\lambda q_{jj}} X^\top Q e_j, \qquad y - X\widehat{\beta}(\lambda, \tau) = \frac{s a_j}{\lambda q_{jj}} e_j.$$

Proof. Unique binding gives

$$\lambda_0 a_k \in (\tau - 1, \tau) \qquad (k \neq j).$$

Because $Q \succ 0$, $q_{jj} > 0$. All finite η_k are positive, hence $\eta_* > 0$ also under the empty-index convention. Put

$$t = \frac{s a_j}{\lambda q_{jj}}, \qquad \beta_\lambda = \widehat{\beta}_0 - t X^\top Q e_j, \qquad g_\lambda = \lambda Q(y - t e_j).$$

Then

$$y - X\beta_\lambda = t e_j.$$

Also

$$g_{\lambda,j} = \lambda \left(a_j - \frac{s a_j}{\lambda q_{jj}} q_{jj} \right) = \lambda_0 a_j = c.$$

For $k \neq j$,

$$g_{\lambda,k} = \lambda_0 a_k + s \left(a_k - \frac{q_{kj}}{q_{jj}} a_j \right) = \lambda_0 a_k + s v_k.$$

The definition of η_* gives

$$g_{\lambda,k} \in (\tau - 1, \tau) \qquad (k \neq j)$$

for $0 < s < \eta_*$. If $a_j > 0$, then $c = \tau$ and $t > 0$; if $a_j < 0$, then $c = \tau - 1$ and $t < 0$. Lemma 6 gives both identities. □

The following proposition gives simultaneous exact certificates at every point of a finite inside-radius perturbation grid.

Proposition 22. *Fix $\tau \in (0,1)$, let $\lambda_0 = \bar{\lambda}(\tau) < \infty$, and suppose the minimum defining $\bar{\lambda}(\tau)$ is attained at the unique index j. Write $q_{k\ell} = e_k^\top Q e_\ell$. Let v_k, η_k, and η_* be as in Corollary 13. Let $\mathcal{H} \subset (0, \infty)$ be finite and suppose*

$$h < \eta_* \qquad (h \in \mathcal{H}).$$

For $h \in \mathcal{H}$, put

$$\lambda_h = \lambda_0 + h, \qquad t_h = \frac{h a_j}{\lambda_h q_{jj}},$$

$$\beta_h = \widehat{\beta}_0 - t_h X^\top Q e_j, \qquad g_h = \lambda_h Q(y - t_h e_j).$$

Then, for every $h \in \mathcal{H}$,

$$\widehat{\beta}(\lambda_h, \tau) = \beta_h, \qquad y - X\widehat{\beta}(\lambda_h, \tau) = t_h e_j.$$

Also,

$$g_{h,j} = \lambda_0 a_j, \qquad \tau - 1 < g_{h,k} < \tau \quad (k \neq j).$$

If $a_j > 0$, then $g_{h,j} = \tau$ and $t_h > 0$. If $a_j < 0$, then $g_{h,j} = \tau - 1$ and $t_h < 0$.

Proof. The unique binding condition gives $a_j \neq 0$ and

$$\lambda_0 a_k \in (\tau - 1, \tau) \qquad (k \neq j).$$

Positive definiteness of Q gives $q_{jj} > 0$. The identities $X\widehat{\beta}_0 = y$ and $XX^\top Q e_j = e_j$ give

$$y - X\beta_h = t_h e_j.$$

The stationarity vector satisfies

$$\lambda_h \beta_h = \lambda_h X^\top Q(y - t_h e_j) = X^\top g_h.$$

For the binding coordinate,

$$g_{h,j} = \lambda_h(a_j - t_h q_{jj}) = (\lambda_0 + h)a_j - h a_j = \lambda_0 a_j.$$

If $a_j > 0$, the definition of λ_0 gives $\lambda_0 a_j = \tau$ and $t_h > 0$. If $a_j < 0$, it gives $\lambda_0 a_j = \tau - 1$ and $t_h < 0$.

For $k \neq j$,

$$g_{h,k} = \lambda_h(a_k - t_h q_{kj}) = \lambda_0 a_k + h\left(a_k - \frac{q_{kj}}{q_{jj}} a_j\right) = \lambda_0 a_k + h v_k.$$

The definition of η_* and the condition $h < \eta_*$ give

$$\tau - 1 < g_{h,k} < \tau \qquad (k \neq j).$$

The residual $t_h e_j$ has the sign required by $g_{h,j}$. Lemma 6 gives $\beta_h = \widehat{\beta}(\lambda_h, \tau)$, proving both identities. □

The following corollary applies the same exact certificate to finitely many fixed-feature problems with different dimensions and quantile levels.

Let M be a positive integer. For $m = 1, \ldots, M$, let $X_m \in \mathbb{R}^{n_m \times p_m}$ have rank n_m, let $y_m \in \mathbb{R}^{n_m}$, and fix $\tau_m \in (0, 1)$. Let $e_{m,k}$ denote coordinate vector k of $\mathbb{R}^{n_m}$. Let

$$K_m = X_m X_m^\top, \qquad Q_m = K_m^{-1}, \qquad a_m = Q_m y_m, \qquad \widehat{\beta}_{m,0} = X_m^\top a_m.$$

For $i \leq n_m$, define

$$b_{m,i} = \begin{cases} \tau_m / a_{m,i}, & a_{m,i} > 0, \\ (1 - \tau_m)/|a_{m,i}|, & a_{m,i} < 0, \\ +\infty, & a_{m,i} = 0, \end{cases} \qquad \lambda_{m,0} = \min_{i \leq n_m} b_{m,i}.$$

Assume $\lambda_{m,0} < \infty$ and that the minimum is attained at the unique index j_m. Put $q_{m,k\ell} = e_{m,k}^\top Q_m e_{m,\ell}$ and, for $k \neq j_m$,

$$v_{m,k} = a_{m,k} - \frac{q_{m,kj_m}}{q_{m,j_m j_m}} a_{m,j_m}.$$

Define

$$\eta_{m,k} = \begin{cases} (\tau_m - \lambda_{m,0} a_{m,k})/v_{m,k}, & v_{m,k} > 0, \\ (\lambda_{m,0} a_{m,k} - \tau_m + 1)/(-v_{m,k}), & v_{m,k} < 0, \\ +\infty, & v_{m,k} = 0, \end{cases} \qquad \eta_m = \min_{k \neq j_m} \eta_{m,k},$$

where the minimum over an empty index set is $+\infty$. Let $h_1, \ldots, h_M$ be positive numbers satisfying

$$0 < h_m < \eta_m \qquad (m = 1, \ldots, M).$$

Put

$$\lambda_m = \lambda_{m,0} + h_m, \qquad t_m = \frac{h_m a_{m,j_m}}{\lambda_m q_{m,j_m j_m}},$$

$$\beta_m = \widehat{\beta}_{m,0} - t_m X_m^\top Q_m e_{m,j_m}, \qquad g_m = \lambda_m Q_m(y_m - t_m e_{m,j_m}).$$

Let $\widehat{\beta}_m(\lambda, \tau)$ denote the penalized check-loss estimator formed from (X_m, y_m).

Corollary 14. *Then, for $m = 1, \dots, M$,*

$$\widehat{\beta}_m(\lambda_m, \tau_m) = \beta_m, \qquad y_m - X_m \widehat{\beta}_m(\lambda_m, \tau_m) = t_m e_{m,j_m}.$$

Also,

$$g_{m,j_m} = \lambda_{m,0} a_{m,j_m}, \qquad \tau_m - 1 < g_{m,k} < \tau_m \quad (k \neq j_m).$$

Proof. For each m, apply Corollary 13 and Proposition 22 to (X_m, y_m, τ_m) with $\mathcal{H} = \{h_m\}$. Because $0 < h_m < \eta_m$, they give

$$\widehat{\beta}_m(\lambda_m, \tau_m) = \beta_m, \qquad y_m - X_m \widehat{\beta}_m(\lambda_m, \tau_m) = t_m e_{m,j_m},$$

and the identities and strict free-coordinate inequalities for g_m. □

Several coordinates can reach the boundary at the same penalty. The following proposition gives the residual under a sign condition that also permits some binding coordinates to have zero residuals.

Fix $\tau \in (0, 1)$, let $\lambda_0 = \bar{\lambda}(\tau) < \infty$, and define

$$J = \{i \leq n : \ b_i(\tau) = \lambda_0\}.$$

Proposition 23. *Suppose*

$$a_j (Q_{JJ}^{-1} a_J)_j \geq 0, \qquad j \in J.$$

Then there exists $\eta > 0$ such that, for $\lambda_0 < \lambda < \lambda_0 + \eta$, the vector defined by

$$r_J(\lambda) = \left(1 - \frac{\lambda_0}{\lambda}\right) Q_{JJ}^{-1} a_J, \qquad r_{J^c}(\lambda) = 0,$$

satisfies

$$\widehat{\beta}(\lambda, \tau) = \widehat{\beta}_0 - X^\top Q r(\lambda), \qquad y - X\widehat{\beta}(\lambda, \tau) = r(\lambda).$$

Proof. Set

$$\beta_\lambda = \widehat{\beta}_0 - X^\top Q r(\lambda), \qquad g_\lambda = \lambda Q(y - r(\lambda)).$$

Then $y - X\beta_\lambda = r(\lambda)$ and

$$\lambda \beta_\lambda = X^\top g_\lambda.$$

On J,

$$g_{\lambda,J} = \lambda\{a_J - Q_{JJ} r_J(\lambda)\} = \lambda_0 a_J.$$

Every a_j with $j \in J$ is nonzero. If $r_j(\lambda) > 0$, the sign condition gives $g_{\lambda,j} = \tau$; if $r_j(\lambda) < 0$, it gives $g_{\lambda,j} = \tau - 1$. If $r_j(\lambda) = 0$, either value belongs to $\partial\rho_\tau(0)$. Thus $g_{\lambda,j} \in \partial\rho_\tau\{r_j(\lambda)\}$ in all three cases.

For $k \notin J$,

$$g_{\lambda,k} \to \lambda_0 a_k \qquad \text{as } \lambda \downarrow \lambda_0.$$

The definition of J gives

$$\lambda_0 a_k \in (\tau - 1, \tau) \qquad (k \notin J).$$

Choose $\eta > 0$ small enough that

$$g_{\lambda,k} \in [\tau - 1, \tau] \qquad (k \notin J)$$

for $\lambda_0 < \lambda < \lambda_0 + \eta$. Lemma 6 proves the result. □

The following proposition shows that each nonempty face-validity set is an interval with affine formulas for the estimator, residual, and scaled dual quantities.

Proposition 24. *Fix $\tau \in (0, 1)$. For disjoint $L, U \subset \{1, \dots, n\}$, put $A = L \cup U$. If $A = \emptyset$, define*

$$I_{\emptyset,\emptyset} = \{\lambda > 0 : \tau - 1 < \lambda a_i < \tau \text{ for every } i \leq n\}.$$

If $A \neq \emptyset$, let $I_{L,U}$ be the set in Theorem 16. Then $I_{L,U}$ is empty or an interval in $(0, \infty)$. On $I_{L,U}$ there exist vectors $b_0, b_1 \in \mathbb{R}^p$, $s_0, s_1, h_0, h_1 \in \mathbb{R}^n$ such that

$$\widehat{\beta}(\lambda, \tau) = b_0 + \lambda^{-1} b_1, \qquad r_{\lambda,\tau} = s_0 + \lambda^{-1} s_1, \qquad g_{\lambda,\tau} = h_0 + \lambda h_1.$$

Also, $\lambda\widehat{\beta}(\lambda, \tau)$ and $\lambda\widehat{f}(\lambda, \tau)$ are affine functions of λ on $I_{L,U}$.

Proof. If $A = \emptyset$, the set $I_{\emptyset,\emptyset}$ is an intersection of linear inequalities in λ. On this set, Theorem 1 gives

$$\widehat{\beta}(\lambda, \tau) = \widehat{\beta}_0, \qquad r_{\lambda,\tau} = 0, \qquad g_{\lambda,\tau} = \lambda a,$$

and the stated affine forms hold.

Assume $A \neq \emptyset$ and use the notation of Theorem 16. If $F \neq \emptyset$,

$$d_A(\lambda) = a_A - \lambda^{-1} c_A, \qquad d_F(\lambda) = -K_{FF}^{-1} K_{FA} a_A + \lambda^{-1} K_{FF}^{-1} K_{FA} c_A.$$

If $F = \emptyset$, only the first formula remains. In either case, $d(\lambda) = d_0 + \lambda^{-1} d_1$ for fixed vectors d_0, d_1. Theorem 16 gives

$$\widehat{\beta}(\lambda, \tau) = \widehat{\beta}_0 - X^\top d(\lambda), \qquad r_{\lambda,\tau} = K d(\lambda),$$

which gives the affine forms in λ^{-1}. Also,

$$g_{\lambda,\tau} = \lambda\{a - d(\lambda)\} = \lambda(a - d_0) - d_1,$$

which is affine in λ. The identities $\lambda\widehat{\beta} = X^\top g_{\lambda,\tau}$ and $\lambda\widehat{f} = K g_{\lambda,\tau}$ prove the last assertion.

The free-coordinate constraints are

$$\tau - 1 < h_{0,i} + \lambda h_{1,i} < \tau.$$

The active residual sign constraints are

$$s_{0,i} + \lambda^{-1} s_{1,i} \leq 0 \qquad \text{or} \qquad s_{0,i} + \lambda^{-1} s_{1,i} \geq 0.$$

After multiplying the latter by $\lambda > 0$, all constraints are linear in λ, giving an interval. □

The following corollary shows that the positive penalty path has finitely many face intervals, with coefficients and residuals affine in λ^{-1} on each interval.

Corollary 15. *Fix $\tau \in (0, 1)$. The half-line $(0, \infty)$ is covered by finitely many intervals on which $\widehat{\beta}(\lambda, \tau)$ and $r_{\lambda,\tau}$ are affine functions of λ^{-1}, while $g_{\lambda,\tau}$, $\lambda\widehat{\beta}(\lambda, \tau)$, and $\lambda\widehat{f}(\lambda, \tau)$ are affine functions of λ.*

Proof. For each $\lambda > 0$, let $L = \{i : g_{\lambda,\tau,i} = \tau - 1\}$ and $U = \{i : g_{\lambda,\tau,i} = \tau\}$. If $L \cup U = \emptyset$, then $\lambda \in I_{\emptyset,\emptyset}$. If $L \cup U \neq \emptyset$, Theorem 16 gives $\lambda \in I_{L,U}$. There are finitely many such pairs, and Proposition 24 gives the affine formulas on each interval. □

S.10.1 Certified active-set numerical audit

This audit checks the one-coordinate formula in Corollary 13 only at perturbations strictly inside its computed validity radius. It contains 600 accepted unique-binding cases and 2400 inside-radius rows. Table S.1 reports machine-precision stationarity, subgradient membership, and residual identities, while Figure S.1 shows their dependence on the fraction of the certified radius. Outside-radius rows and solver comparisons are auxiliary diagnostics and are not part of the certificate.

Table S.1: Certified active-set interval diagnostics. Rows use accepted unique-binding cases and perturbations strictly inside the computed active-set radius. Outside-radius rows are diagnostic and are not used in this table.

Certified active-set check	Statistic	Value
Accepted unique-binding cases	total	600
Inside-radius rows	total	2400
Inside-radius KKT residual	max stationarity	2.220e-16
Inside-radius KKT residual	max subgradient	4.441e-16
Inside-radius residual identity	max error	2.970e-15
Inside-radius face slack	minimum	7.351e-06
Inside-radius active residual	minimum sign margin	2.729e-06

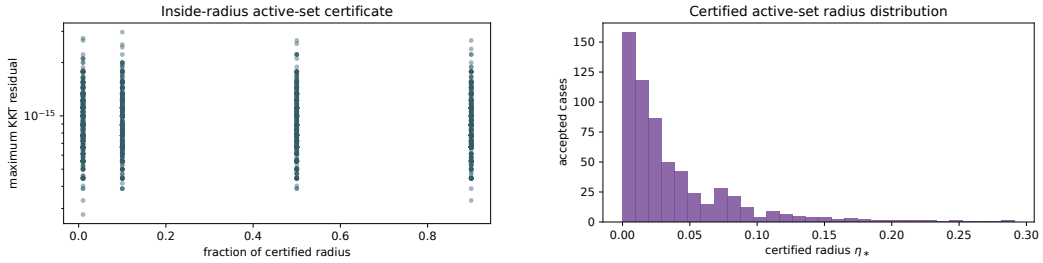

Figure S.1: Certified active-set diagnostics. Left: maximum analytic KKT residual for inside-radius perturbations as a fraction of the certified radius. Right: distribution of certified radii across 600 accepted unique-binding cases.

S.11 JOINT FACE REGIONS AND FINITE-GRID CERTIFICATES

Each face of the dual box has an explicit validity region in the (λ, τ) plane. This region is convex, but it may be empty or lower-dimensional. Within this stable region, the scaled coefficient vector $\lambda\widehat{\beta}(\lambda, \tau)$ behaves as an exactly affine function of (λ, τ). This linear structure allows us to use finite systems of linear inequalities to certify the optimality of a given face simultaneously over entire rectangular domains and finite grids.

Let $L, U \subset \{1, \dots, n\}$ be disjoint, put $A = L \cup U$, and suppose $A \neq \emptyset$. Let $\mathbf{1}_A \in \mathbb{R}^A$ be the vector of ones and let $\mathbf{1}_L \in \mathbb{R}^A$ have coordinate 1 on L and coordinate 0 on U. Put

$$c_A(\tau) = \tau \mathbf{1}_A - \mathbf{1}_L, \qquad d_A(\lambda, \tau) = a_A - \lambda^{-1} c_A(\tau), \qquad g_A(\lambda, \tau) = c_A(\tau).$$

For $F = A^c$, if $F \neq \emptyset$, set

$$S_A = K_{AA} - K_{AF} K_{FF}^{-1} K_{FA}, \quad d_F(\lambda, \tau) = -K_{FF}^{-1} K_{FA} d_A(\lambda, \tau), \quad g_F(\lambda, \tau) = \lambda\{a_F - d_F(\lambda, \tau)\},$$

and

$$r_F(\lambda, \tau) = 0, \qquad r_A(\lambda, \tau) = S_A d_A(\lambda, \tau).$$

If $F = \emptyset$, the formulas involving F are absent, and

$$S_A = K_{AA}, \qquad r_A(\lambda, \tau) = S_A d_A(\lambda, \tau).$$

Let $d(\lambda, \tau)$, $g(\lambda, \tau)$, and $r(\lambda, \tau)$ denote the vectors with these active coordinates and, when $F \neq \emptyset$, the free coordinates defined above. Let $\mathcal{R}_{L,U}$ be the set of $(\lambda, \tau) \in (0, \infty) \times (0, 1)$ satisfying

$$\tau - 1 < g_i(\lambda, \tau) < \tau \quad (i \in F), \qquad r_i(\lambda, \tau) \leq 0 \quad (i \in L), \qquad r_i(\lambda, \tau) \geq 0 \quad (i \in U).$$

Theorem 17. *For every $(\lambda, \tau) \in \mathcal{R}_{L,U}$,*

$$\widehat{\beta}(\lambda, \tau) = \widehat{\beta}_0 - X^\top d(\lambda, \tau), \qquad y - X\widehat{\beta}(\lambda, \tau) = r(\lambda, \tau).$$

If the dual projection $g_{\lambda,\tau}$ has lower active set L and upper active set U at a pair (λ, τ), then $(\lambda, \tau) \in \mathcal{R}_{L,U}$ and the same formulas hold.

Proof. Fix (λ, τ). The definitions of $c_A(\tau)$ and $\mathcal{R}_{L,U}$ are exactly the box boundaries and sign constraints in Theorem 16 at level τ. That theorem gives both the identities on $\mathcal{R}_{L,U}$ and the converse for the actual active sets. □

The following proposition gives explicit coefficient, residual, and subgradient formulas throughout each certified penalty–level face.

Proposition 25. *For every disjoint pair L, U with $L \cup U \neq \emptyset$, the set $\mathcal{R}_{L,U}$ in Theorem 17 is an intersection of finitely many open or closed half-planes in the (λ, τ) plane. On $\mathcal{R}_{L,U}$ there exist fixed vectors $b_0, b_1, b_2 \in \mathbb{R}^p$, $s_0, s_1, s_2 \in \mathbb{R}^n$, and $h_0, h_1, h_2 \in \mathbb{R}^n$ such that*

$$\widehat{\beta}(\lambda, \tau) = b_0 + \lambda^{-1} b_1 + \tau\lambda^{-1} b_2, \qquad r_{\lambda,\tau} = s_0 + \lambda^{-1} s_1 + \tau\lambda^{-1} s_2, \qquad g_{\lambda,\tau} = h_0 + \lambda h_1 + \tau h_2.$$

Also, $\lambda\widehat{\beta}(\lambda, \tau)$ and $\lambda\widehat{f}(\lambda, \tau)$ are affine functions of (λ, τ) on $\mathcal{R}_{L,U}$.

Proof. Use the notation of Theorem 17. If $F \neq \emptyset$, then

$$d_A(\lambda, \tau) = a_A + \lambda^{-1}\mathbf{1}_L - \tau\lambda^{-1}\mathbf{1}_A,$$

$$d_F(\lambda, \tau) = -K_{FF}^{-1}K_{FA}a_A - \lambda^{-1}K_{FF}^{-1}K_{FA}\mathbf{1}_L + \tau\lambda^{-1}K_{FF}^{-1}K_{FA}\mathbf{1}_A.$$

If $F = \emptyset$, only the first formula remains. In either case,

$$d(\lambda, \tau) = d_0 + \lambda^{-1}d_1 + \tau\lambda^{-1}d_2$$

for fixed vectors d_0, d_1, d_2. Theorem 17 gives

$$\widehat{\beta}(\lambda, \tau) = \widehat{\beta}_0 - X^\top d(\lambda, \tau), \qquad r_{\lambda,\tau} = Kd(\lambda, \tau),$$

which gives the stated forms for $\widehat{\beta}$ and $r_{\lambda,\tau}$. Also,

$$g_{\lambda,\tau} = \lambda\{a - d(\lambda, \tau)\}$$

on F, while $g_A(\lambda, \tau) = \tau\mathbf{1}_A - \mathbf{1}_L$. This gives the stated affine form for $g_{\lambda,\tau}$. Multiplying the coefficient and fitted-value identities by λ gives affine functions of (λ, τ).

The free-coordinate constraints are

$$\tau - 1 < h_{0,i} + \lambda h_{1,i} + \tau h_{2,i} < \tau.$$

These are strict affine inequalities in (λ, τ). For active coordinates,

$$r_i(\lambda, \tau) = s_{0,i} + \lambda^{-1}s_{1,i} + \tau\lambda^{-1}s_{2,i}.$$

Multiplying the residual sign constraints by $\lambda > 0$ gives affine inequalities in (λ, τ). Including $\lambda > 0$ and $0 < \tau < 1$ proves the half-plane statement. □

The following corollary reduces a continuum face certificate on a penalty–level rectangle to four corner checks.

Corollary 16. *Let $\mathcal{J} = [\lambda_-, \lambda_+] \times [\tau_-, \tau_+] \subset (0, \infty) \times (0, 1)$, with $\lambda_- < \lambda_+$ and $\tau_- < \tau_+$. In the notation of Proposition 25, to prove $\mathcal{J} \subset \mathcal{R}_{L,U}$ it is enough to check every defining affine inequality of $\mathcal{R}_{L,U}$ at the four corners of $\mathcal{J}$, with strict inequality at the four corners for the open constraints. Under this check, the identities in Theorem 17 hold for every $(\lambda, \tau) \in \mathcal{J}$.*

Proof. An affine function attains its extrema over a rectangle at corners. Thus both strict and non-strict affine inequalities at the four corners hold throughout $\mathcal{J}$, giving $\mathcal{J} \subset \mathcal{R}_{L,U}$. Theorem 17 gives the identities. □

The following proposition converts a computable stationarity residual into deterministic bounds on coefficient error and the objective gap.

Proposition 26. *Fix $\lambda > 0$ and $\tau \in (0, 1)$. Let $\beta \in \mathbb{R}^p$, put $r = y - X\beta$, and suppose there is $g \in \mathbb{R}^n$ such that*

$$g_i \in \partial\rho_\tau(r_i), \qquad i \leq n.$$

Set $s = \lambda\beta - X^\top g$. Then

$$\left\|\beta - \widehat{\beta}(\lambda, \tau)\right\|_2 \leq \lambda^{-1} \|s\|_2 ,$$

and

$$\sum_{i=1}^n \rho_\tau(r_i) + \frac{\lambda}{2}\|\beta\|_2^2 - \min_{\theta \in \mathbb{R}^p}\left\{\sum_{i=1}^n \rho_\tau(y_i - x_i^\top\theta) + \frac{\lambda}{2}\|\theta\|_2^2\right\} \leq \frac{1}{2\lambda}\|s\|_2^2 .$$

If $s = 0$, then $\beta = \widehat{\beta}(\lambda, \tau)$.

Proof. Let

$$F(\theta) = \sum_{i=1}^n \rho_\tau(y_i - x_i^\top\theta) + \frac{\lambda}{2}\|\theta\|_2^2 .$$

Because $s \in \partial F(\beta)$, $0 \in \partial F(\widehat{\beta}(\lambda,\tau))$, and F is λ-strongly convex,

$$s^\top\{\beta - \widehat{\beta}(\lambda,\tau)\} \geq \lambda \left\|\beta - \widehat{\beta}(\lambda,\tau)\right\|_2^2.$$

Cauchy's inequality proves the distance bound. Strong convexity also gives

$$F(\widehat{\beta}(\lambda,\tau)) \geq F(\beta) + s^\top\{\widehat{\beta}(\lambda,\tau) - \beta\} + \frac{\lambda}{2}\left\|\widehat{\beta}(\lambda,\tau) - \beta\right\|_2^2.$$

The right side is bounded below by $F(\beta) - \frac{1}{2\lambda}\|s\|_2^2$, which proves the value bound. If $s = 0$, the distance bound gives equality of the candidate and the minimizer. □

The following corollary gives a uniform finite-grid certificate in which zero stationarity residuals identify every optimizer and its extinct status exactly.

Corollary 17. *Let $\mathcal{G} \subset (0,\infty)\times(0,1)$ be nonempty and finite. For each $(\lambda,\tau) \in \mathcal{G}$, let $\beta_{\lambda,\tau} \in \mathbb{R}^p$, set $r_{\lambda,\tau} = y - X\beta_{\lambda,\tau}$, and suppose there is $g_{\lambda,\tau} \in \mathbb{R}^n$ satisfying*

$$g_{\lambda,\tau,i} \in \partial\rho_\tau(r_{\lambda,\tau,i}), \qquad i \leq n.$$

Put $s_{\lambda,\tau} = \lambda\beta_{\lambda,\tau} - X^\top g_{\lambda,\tau}$. Then

$$\max_{(\lambda,\tau)\in\mathcal{G}} \left\|\beta_{\lambda,\tau} - \widehat{\beta}(\lambda,\tau)\right\|_2 \leq \max_{(\lambda,\tau)\in\mathcal{G}} \lambda^{-1}\|s_{\lambda,\tau}\|_2.$$

If $s_{\lambda,\tau} = 0$ for every $(\lambda,\tau) \in \mathcal{G}$, then all candidates are exact optimizers. In that case, for any $E \subset \mathcal{G}$ satisfying

$$\beta_{\lambda,\tau} = \widehat{\beta}_0 \quad ((\lambda,\tau)\in E), \qquad \beta_{\lambda,\tau} \neq \widehat{\beta}_0 \quad ((\lambda,\tau)\in\mathcal{G}\setminus E),$$

one has

$$E = \{(\lambda,\tau)\in\mathcal{G} : \widehat{\beta}(\lambda,\tau) = \widehat{\beta}_0\}.$$

Proof. Apply Proposition 26 pointwise and take the maximum over $\mathcal{G}$. Zero stationarity residuals give $\beta_{\lambda,\tau} = \widehat{\beta}(\lambda,\tau)$. The two candidate identities and uniqueness then characterize E. □

S.12 Stability and fixed-feature maps

We bound the sensitivity of the extinction boundary, its maximizing quantile level, and the associated extinction certificate to worst-case perturbations of the interpolation vector, the responses, or the realized feature map, subject to the stated norm bounds. The boundary and maximizer bounds require both signed extremes of the interpolation vector to be bounded away from zero. By establishing these stability bounds, we guarantee that an extinction certificate computed for a specific data realization with a strictly positive margin inside the subgradient box remains mathematically valid across an explicit, computable neighborhood around that realization.

For $a \in \mathbb{R}^n$, write $A_\pm(a)$, $\bar{\lambda}_a(\tau)$, and $\tau^\dagger(a)$ for the quantities in Theorem 2.

Proposition 27. *Suppose*

$$A_+(a) \wedge A_-(a) \geq \alpha > 0, \qquad \|\widetilde{a} - a\|_\infty \leq \delta < \alpha.$$

Then

$$A_+(\widetilde{a}) \wedge A_-(\widetilde{a}) \geq \alpha - \delta, \qquad \sup_{\tau\in(0,1)} \left|\bar{\lambda}_{\widetilde{a}}(\tau) - \bar{\lambda}_a(\tau)\right| \leq \frac{\delta}{\alpha(\alpha-\delta)}.$$

Also,

$$\left|\tau^\dagger(\widetilde{a}) - \tau^\dagger(a)\right| \leq \frac{\delta}{A_+(a) + A_-(a) - 2\delta}.$$

Proof. The maps $a \mapsto A_+(a)$ and $a \mapsto A_-(a)$ are one-Lipschitz under $\|\cdot\|_\infty$. This gives

$$|A_\pm(\widetilde{a}) - A_\pm(a)| \leq \delta, \qquad A_\pm(\widetilde{a}) \geq \alpha - \delta.$$

For any $x \geq \alpha$ and $\widetilde{x} \geq \alpha - \delta$ with $|\widetilde{x} - x| \leq \delta$,

$$\left|\frac{1}{\widetilde{x}} - \frac{1}{x}\right| = \frac{|\widetilde{x} - x|}{\widetilde{x}x} \leq \frac{\delta}{\alpha(\alpha - \delta)}.$$

Multiplication by τ or $1 - \tau$ and taking the minimum preserve this uniform bound, giving the bound for $\bar{\lambda}$.

Let

$$x = A_+(a), \quad y = A_-(a), \quad \widetilde{x} = A_+(\widetilde{a}), \quad \widetilde{y} = A_-(\widetilde{a}).$$

Then

$$\left|\frac{\widetilde{x}}{\widetilde{x} + \widetilde{y}} - \frac{x}{x + y}\right| = \frac{|\,\widetilde{x}y - x\widetilde{y}\,|}{(\widetilde{x} + \widetilde{y})(x + y)} \leq \frac{\delta(x + y)}{(x + y - 2\delta)(x + y)}.$$

This proves the bound for $\tau^\dagger$. □

The following proposition quantifies the response perturbations that preserve extinction simultaneously over a prescribed set of levels.

Proposition 28. *Fix $\lambda > 0$, a nonempty set $T \subset (0, 1)$, and a response vector y. Set $a = Qy$ and*

$$m = \inf_{\tau \in T} \min_{i \leq n} \{\tau - \lambda a_i, 1 - \tau + \lambda a_i\}.$$

Suppose $m \geq 0$. If $\widetilde{y} \in \mathbb{R}^n$ and

$$\|Q(\widetilde{y} - y)\|_\infty \leq \lambda^{-1} m,$$

then, with $\widetilde{a} = Q\widetilde{y}$ and with $\widehat{\beta}_{\widetilde{y}}(\lambda, \tau)$ denoting the estimator formed from response $\widetilde{y}$,

$$\widehat{\beta}_{\widetilde{y}}(\lambda, \tau) = X^\top \widetilde{a} \qquad (\tau \in T).$$

Proof. For any $\tau \in T$ and any coordinate i, $\tau - 1 \leq \lambda a_i \leq \tau$ by the definition of $m \geq 0$. Also,

$$|\lambda(\widetilde{a}_i - a_i)| \leq \lambda \|Q(\widetilde{y} - y)\|_\infty \leq m.$$

The definition of m gives

$$\tau - 1 \leq \lambda \widetilde{a}_i \leq \tau \qquad (i \leq n,\ \tau \in T).$$

Theorem 1, applied with response vector $\widetilde{y}$, gives the stated identity. □

For a matrix M, $\|M\|_{\infty \to \infty} = \max_i \sum_j |M_{ij}|$ is the operator norm induced by the maximum coordinate norm. The following proposition separates the effects of response and Gram-matrix perturbations on the interpolation vector and hence on the extinction boundary.

Proposition 29. *Let K and $\widetilde{K}$ be positive definite $n \times n$ matrices, set $Q = K^{-1}$, $\Delta = \widetilde{K} - K$, $a = Qy$, and $\widetilde{a} = \widetilde{K}^{-1}\widetilde{y}$. If $\varepsilon = \|Q\Delta\|_{\infty \to \infty} < 1$, then*

$$\|\widetilde{a} - a\|_\infty \leq \frac{\|Q(\widetilde{y} - y)\|_\infty + \varepsilon \|a\|_\infty}{1 - \varepsilon}.$$

If $A_+(a) \wedge A_-(a) \geq \alpha > 0$ and the upper bound above is at most some $\delta < \alpha$, then

$$\sup_{\tau \in (0,1)} \left|\bar{\lambda}_{\widetilde{a}}(\tau) - \bar{\lambda}_a(\tau)\right| \leq \frac{\delta}{\alpha(\alpha - \delta)}.$$

Proof. The equations $Ka = y$ and $(K + \Delta)\widetilde{a} = \widetilde{y}$ give

$$(K + \Delta)(\widetilde{a} - a) = \widetilde{y} - y - \Delta a.$$

Multiplication by Q gives

$$\widetilde{a} - a = Q(\widetilde{y} - y) - Q\Delta a - Q\Delta(\widetilde{a} - a).$$

Taking the ℓ_∞ norm gives

$$\|\widetilde{a} - a\|_\infty \leq \|Q(\widetilde{y} - y)\|_\infty + \varepsilon \|a\|_\infty + \varepsilon \|\widetilde{a} - a\|_\infty.$$

Because $\varepsilon < 1$, rearranging proves the perturbation bound. The final inequality follows from Proposition 27. □

The following corollary states that all fixed-design identities apply conditionally to any realized feature map whose training matrix has full row rank.

Corollary 18. *Let $z_1, \dots, z_n$ be arbitrary covariate points, and let ψ map such points into $\mathbb{R}^p$. Set*

$$X = \begin{pmatrix} \psi(z_1)^\top \\ \vdots \\ \psi(z_n)^\top \end{pmatrix}.$$

On every realization for which $\operatorname{rank}(X) = n$, Theorems 1 and 2 apply with

$$K = XX^\top, \qquad a = K^{-1}y.$$

Proof. Fixing the features gives the problem in Section 2, whose deterministic result applies to X. □

S.13 SELECTED FIXED-FEATURE CERTIFICATES

We establish a deterministic extinction certificate computed from the realized feature matrix and response vector. We prove that whenever the normalized feature rows are nearly orthogonal—specifically, when their Gram matrix has distance less than one from the identity matrix in the ℓ_∞ operator norm—we can derive explicit, closed-form coordinate bounds on the interpolation vector. For target levels bounded away from zero and one, these bounds immediately yield a guaranteed sufficient penalty range over which every target quantile level is exactly extinct.

Let $x_i^\top$ be the realized fixed-feature rows, $K = XX^\top$, and $D = \operatorname{diag}(K_{11}, \dots, K_{nn})$. Assume D is positive definite and define

$$C = D^{-1/2} K D^{-1/2}, \qquad \varepsilon_\psi = \|C - I_n\|_{\infty\to\infty}.$$

When $\varepsilon_\psi < 1$, define

$$M_\psi = \frac{\left\|D^{-1/2}y\right\|_\infty}{1 - \varepsilon_\psi}.$$

Under the same condition, for $i \le n$, put

$$B_{\psi,i} = D_{ii}^{-1/2} M_\psi, \qquad B_\psi = \max_{i \le n} B_{\psi,i}.$$

The following proposition formalizes this geometric intuition, converting the near-orthogonality of the normalized feature rows and the response-dependent bounds into a common extinction certificate that holds simultaneously over all requested quantile levels under the stated common-penalty condition.

Proposition 30. *If $\varepsilon_\psi < 1$, then K is positive definite and*

$$|a_i| \le B_{\psi,i} \qquad (i \le n).$$

Let $T \subset (0,1)$ be nonempty and put

$$m_T = \inf_{\tau \in T} \{\tau \wedge (1-\tau)\}.$$

If $\lambda > 0$ and $\lambda B_\psi \le m_T$, then

$$\widehat{\beta}(\lambda, \tau) = \widehat{\beta}_0 \qquad (\tau \in T).$$

Proof. Write $C = I_n + E$. Because C is symmetric,

$$\|E\|_2 \le \|E\|_{\infty\to\infty} = \varepsilon_\psi < 1.$$

All eigenvalues of C are positive, and $K = D^{1/2} C D^{1/2}$ is positive definite.

The Neumann series gives

$$\left\|C^{-1}\right\|_{\infty\to\infty} \le \sum_{r=0}^{\infty} \|E\|_{\infty\to\infty}^r = \frac{1}{1 - \varepsilon_\psi}.$$

Because $a = K^{-1}y = D^{-1/2}C^{-1}D^{-1/2}y$, we have

$$|a_i| \le D_{ii}^{-1/2} \left\| C^{-1}D^{-1/2}y \right\|_\infty \le D_{ii}^{-1/2} M_\psi = B_{\psi,i}.$$

For $\tau \in T$, $m_T \le \tau \wedge (1-\tau)$ and $\lambda B_\psi \le m_T$ give

$$\tau - 1 \le -m_T \le \lambda a_i \le m_T \le \tau \qquad (i \le n).$$

Theorem 1 proves the coefficient identity for every $\tau \in T$. □

The following corollary shows that the realized-matrix certificate makes every zero-sum contrast across certified levels equal zero on any evaluation design.

Corollary 19. *Let the conditions of Proposition 30 hold. Let $T = \{\tau_1, \dots, \tau_m\} \subset (0,1)$, put $m_T = \inf_{\tau \in T}\{\tau \wedge (1-\tau)\}$, and assume $\lambda B_\psi \le m_T$. Then, for positive integers N, q, every $Z \in \mathbb{R}^{N \times p}$, and every $C_0 \in \mathbb{R}^{m \times q}$ satisfying $\mathbf{1}_m^\top C_0 = 0$,*

$$Z[\widehat{\beta}(\lambda, \tau_1), \dots, \widehat{\beta}(\lambda, \tau_m)]C_0 = 0.$$

Proof. Proposition 30 gives

$$\widehat{\beta}(\lambda, \tau_k) = \widehat{\beta}_0 \qquad (k = 1, \dots, m).$$

The stated identity follows after multiplication by Z and C_0. □

S.14 Exact diagnostics and level bands

In this section, we develop an array of exact geometric diagnostics. For a fixed positive penalty, the Euclidean distance of the coefficients from the interpolant, the Q-weighted Mahalanobis norm of the fitted residuals, and the exact objective gap between the interpolant and the optimum are all determined by one quantity: the K-metric distance from λa to $\mathcal{B}_\tau$. Using the boundary geometry, we then derive closed-form expressions for the length of the extinct level set, the integral of the boundary curve, and the penalty floor that must be exceeded to avoid extinction throughout any compact band of target quantiles.

Let $d_{\lambda,\tau} = \lambda a - g_{\lambda,\tau}$ and

$$\Phi_{\lambda,\tau}(\beta) = \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top \beta) + \frac{\lambda}{2} \|\beta\|_2^2 .$$

Proposition 31. *For every $\lambda > 0$ and $\tau \in (0,1)$,*

$$\widehat{\beta}(\lambda,\tau) - \widehat{\beta}_0 = -\lambda^{-1} X^\top d_{\lambda,\tau}, \qquad r_{\lambda,\tau} = y - X\widehat{\beta}(\lambda,\tau) = \lambda^{-1} K d_{\lambda,\tau},$$

$$\left\| \widehat{\beta}(\lambda,\tau) - \widehat{\beta}_0 \right\|_2^2 = r_{\lambda,\tau}^\top Q r_{\lambda,\tau} = \lambda^{-2} d_{\lambda,\tau}^\top K d_{\lambda,\tau},$$

and

$$\Phi_{\lambda,\tau}(\widehat{\beta}_0) - \Phi_{\lambda,\tau}(\widehat{\beta}(\lambda,\tau)) = \frac{1}{2\lambda} d_{\lambda,\tau}^\top K d_{\lambda,\tau} = \frac{\lambda}{2} \left\| \widehat{\beta}(\lambda,\tau) - \widehat{\beta}_0 \right\|_2^2 .$$

Proof. Proposition 1 gives

$$\widehat{\beta}(\lambda,\tau) = \lambda^{-1} X^\top g_{\lambda,\tau}, \qquad \widehat{\beta}_0 = X^\top a,$$

giving the coefficient and residual identities. The norm identities follow from

$$\left\| X^\top v \right\|_2^2 = v^\top X X^\top v = v^\top K v$$

with $v = \lambda^{-1} d_{\lambda,\tau}$, and from

$$r_{\lambda,\tau}^\top Q r_{\lambda,\tau} = \lambda^{-2} d_{\lambda,\tau}^\top K Q K d_{\lambda,\tau} = \lambda^{-2} d_{\lambda,\tau}^\top K d_{\lambda,\tau}.$$

At $\widehat{\beta}_0$, the residual vector is zero, and

$$\Phi_{\lambda,\tau}(\widehat{\beta}_0) = \frac{\lambda}{2} y^\top Q y.$$

Proposition 8 gives

$$\Phi_{\lambda,\tau}(\widehat{\beta}(\lambda,\tau)) = \frac{\lambda}{2} y^\top Q y - \frac{1}{2\lambda} d_{\lambda,\tau}^\top K d_{\lambda,\tau}.$$

The stated objective identity follows. □

The following corollary makes extinction equivalent to zero dual distance, zero residual, and zero objective gap.

Corollary 20. *For every $\lambda > 0$ and $\tau \in (0,1)$,*

$$\widehat{\beta}(\lambda,\tau) = \widehat{\beta}_0 \quad \Longleftrightarrow \quad r_{\lambda,\tau} = 0 \quad \Longleftrightarrow \quad d_{\lambda,\tau} = 0 \quad \Longleftrightarrow \quad \Phi_{\lambda,\tau}(\widehat{\beta}_0) = \Phi_{\lambda,\tau}(\widehat{\beta}(\lambda,\tau)).$$

Proof. All four statements are equivalent by Proposition 31 and positive definiteness of K and Q. □

The following proposition measures the fraction of the level interval that is extinct at a fixed penalty and gives the average boundary height.

Proposition 32. *For $\lambda > 0$, let*

$$E_\lambda = \{\tau \in (0,1) : \lambda \le \bar{\lambda}(\tau)\}.$$

If $A_+ = A_- = 0$, then $E_\lambda = (0,1)$. If $A_+ + A_- > 0$, then

$$|E_\lambda| = \max\{1 - \lambda(A_+ + A_-), 0\},$$

where $|\cdot|$ denotes interval length. In the same nonzero case,

$$\int_0^1 \bar{\lambda}(\tau)\, d\tau = \frac{1}{2(A_+ + A_-)}.$$

Proof. The all-zero case follows from Theorem 2. Suppose $A_+ + A_- > 0$. Theorem 2 gives

$$E_\lambda = [\lambda A_+, 1 - \lambda A_-] \cap (0,1)$$

when $0 < \lambda \le (A_+ + A_-)^{-1}$, including the one-sided cases. Its length is $1 - \lambda(A_+ + A_-)$. For larger λ,

$$\tau \ge \lambda A_+, \qquad \tau \le 1 - \lambda A_-$$

have no common point, giving the length formula.

For the integral, first consider $A_+ A_- > 0$ and put $S = A_+ + A_-$. Theorem 2 gives

$$\bar{\lambda}(\tau) = \frac{\tau}{A_+} \quad (0 < \tau \le A_+/S), \qquad \bar{\lambda}(\tau) = \frac{1-\tau}{A_-} \quad (A_+/S \le \tau < 1).$$

Direct integration gives

$$\int_0^1 \bar{\lambda}(\tau)\, d\tau = \frac{A_+}{2S^2} + \frac{A_-}{2S^2} = \frac{1}{2S}.$$

If $A_- = 0 < A_+$, integrate τ/A_+ to obtain $1/(2A_+)$. The case $A_+ = 0 < A_-$ follows by integrating $(1-\tau)/A_-$. □

The following proposition locates the largest extinction boundary on a compact level band by restricting the peak level to that band.

Proposition 33. *Let $0 < \tau_- \le \tau_+ < 1$ and $T = [\tau_-, \tau_+]$. If $A_+ = A_- = 0$, then $\Lambda_n(T) = +\infty$. If $A_+ A_- > 0$, put*

$$\tau_T = (\tau^\dagger \vee \tau_-) \wedge \tau_+.$$

Then

$$\Lambda_n(T) = \bar{\lambda}(\tau_T) = \frac{\tau_T}{A_+} \wedge \frac{1 - \tau_T}{A_-}.$$

If $A_- = 0 < A_+$, then

$$\Lambda_n(T) = \frac{\tau_+}{A_+}.$$

If $A_+ = 0 < A_-$, then

$$\Lambda_n(T) = \frac{1 - \tau_-}{A_-}.$$

Proof. The all-zero case follows from Theorem 2. Suppose $A_+ A_- > 0$. Theorem 2 gives

$$\bar{\lambda}(\tau) = \frac{\tau}{A_+} \quad (0 < \tau \le \tau^\dagger), \qquad \bar{\lambda}(\tau) = \frac{1 - \tau}{A_-} \quad (\tau^\dagger \le \tau < 1).$$

The function increases up to $\tau^\dagger$ and decreases thereafter, hence its maximum over $[\tau_-, \tau_+]$ is attained at τ_T.

If $A_- = 0 < A_+$, Theorem 2 gives $\bar{\lambda}(\tau) = \tau / A_+$, maximized at τ_+. If $A_+ = 0 < A_-$, then $\bar{\lambda}(\tau) = (1 - \tau)/A_-$ is maximized at τ_-. This completes the proof. □

S.15 COMMON-PENALTY LEVEL CERTIFICATES

This section identifies the critical penalty thresholds at or below which an entire interval—or finite grid—of target quantiles undergoes extinction simultaneously. We characterize these common boundaries and the zeros that this joint collapse imposes on every zero-sum linear contrast of the fitted coefficients supported on the extinct levels.

For a nonempty set $T \subset (0, 1)$, define

$$\Lambda_{\mathrm{all}}(T) = \inf_{\tau \in T} \bar{\lambda}(\tau).$$

For a finite ordered grid $T = \{\tau_1 < \cdots < \tau_m\}$ and $\lambda > 0$, define

$$\mathcal{K}_\lambda(T) = \{k \in \{1, \ldots, m\} : \widehat{\beta}(\lambda, \tau_k) = \widehat{\beta}_0\}.$$

The following proposition shows that the common extinction boundary of a closed interval of levels in $(0, 1)$, or of a finite ordered grid, is determined by its two extreme levels, which reduces the computation of Λ_{all} to two boundaries.

Proposition 34. *Let $0 < \tau_- \le \tau_+ < 1$ and $I = [\tau_-, \tau_+]$. With the convention $c/0 = +\infty$ for $c > 0$,*

$$\Lambda_{\mathrm{all}}(I) = \frac{\tau_-}{A_+} \wedge \frac{1 - \tau_+}{A_-}.$$

For every $\lambda > 0$,

$$\widehat{\beta}(\lambda, \tau) = \widehat{\beta}_0 \quad (\tau \in I) \quad \Longleftrightarrow \quad 0 < \lambda \le \Lambda_{\mathrm{all}}(I).$$

If $T = \{\tau_1 < \cdots < \tau_m\} \subset (0, 1)$, then

$$\Lambda_{\mathrm{all}}(T) = \frac{\tau_1}{A_+} \wedge \frac{1 - \tau_m}{A_-},$$

and

$$\widehat{\beta}(\lambda, \tau_k) = \widehat{\beta}_0 \quad (k = 1, \ldots, m) \quad \Longleftrightarrow \quad 0 < \lambda \le \Lambda_{\mathrm{all}}(T).$$

Proof. Theorem 2 gives

$$\lambda \le \bar{\lambda}(\tau) \quad \Longleftrightarrow \quad \lambda A_+ \le \tau, \qquad \lambda A_- \le 1-\tau,$$

with the zero-denominator convention stated in the theorem. Requiring these two inequalities for every $\tau \in I$ is equivalent to

$$\lambda A_+ \le \tau_-, \qquad \lambda A_- \le 1-\tau_+.$$

This is exactly $0 < \lambda \le \Lambda_{\mathrm{all}}(I)$. Theorem 1 gives the coefficient identity on I.

The same calculation at the extreme levels τ_1 and τ_m gives the finite-grid conclusions. □

The following proposition shows that extinct levels form a contiguous block on every ordered grid and that contrasts supported on this block equal zero.

Proposition 35. *Let $T = \{\tau_1 < \cdots < \tau_m\} \subset (0,1)$ and let $\lambda > 0$. If $A_+ = A_- = 0$, then $\mathcal{K}_\lambda(T) = \{1,\dots,m\}$. If $A_+ + A_- > 0$, then*

$$\mathcal{K}_\lambda(T) = \{k : \lambda A_+ \le \tau_k \le 1 - \lambda A_-\}.$$

In particular, $\mathcal{K}_\lambda(T)$ is empty or a contiguous block of $\{1,\dots,m\}$. For positive integers N, q, any $Z \in \mathbb{R}^{N\times p}$, and any $C \in \mathbb{R}^{m\times q}$ satisfying

$$C_{k\ell} = 0 \quad (k \notin \mathcal{K}_\lambda(T),\ \ell \le q), \qquad \sum_{k=1}^{m} C_{k\ell} = 0 \quad (\ell \le q),$$

we have

$$Z[\widehat{\beta}(\lambda,\tau_1),\dots,\widehat{\beta}(\lambda,\tau_m)]C = 0.$$

Proof. If $A_+ = A_- = 0$, then $a = 0$, and Theorem 2 gives $\bar{\lambda}(\tau) = +\infty$ for every $\tau \in (0,1)$. This proves the first claim.

Assume $A_+ + A_- > 0$. By Theorems 1 and 2,

$$\widehat{\beta}(\lambda,\tau_k) = \widehat{\beta}_0 \quad \Longleftrightarrow \quad \lambda A_+ \le \tau_k \le 1 - \lambda A_-.$$

The right side selects the intersection of the ordered grid with an interval, which is empty or contiguous.

The columns indexed by $\mathcal{K}_\lambda(T)$ equal $\widehat{\beta}_0$; all others multiply zero rows of C. Thus

$$Z[\widehat{\beta}(\lambda,\tau_1),\dots,\widehat{\beta}(\lambda,\tau_m)]C = Z\widehat{\beta}_0 \mathbf{1}_m^\top C = 0.$$

This proves the contrast identity. □

S.16 Simultaneous finite-level fits

A common heuristic to prevent quantile crossing is to estimate multiple levels simultaneously using a joint penalty or noncrossing constraints. We prove that joint estimation in the fixed-feature setting below does not prevent extinction under the stated conditions. We show that simultaneous fits coupled by any nonnegative penalty (provided it equals zero at the interpolant tuple) still collapse to the same minimum-norm interpolant when the interpolant tuple is feasible and each positive ridge penalty, divided by its positive loss weight, is at or below that level's extinction boundary. In fact, imposing finite noncrossing constraints preserves this collapse because the common interpolant already makes adjacent levels equal.

Proposition 36. *Let $\tau_1,\dots,\tau_m \in (0,1)$, $\alpha_k > 0$, and $\lambda_k > 0$. Let $\Omega \subset (\mathbb{R}^p)^m$ satisfy $(\widehat{\beta}_0,\dots,\widehat{\beta}_0) \in \Omega$. Let $J : \Omega \to [0,\infty]$ satisfy $J(\widehat{\beta}_0,\dots,\widehat{\beta}_0) = 0$. If*

$$0 < \lambda_k \le \alpha_k \bar{\lambda}(\tau_k), \qquad k = 1,\dots,m,$$

then the unique minimizer over Ω of

$$\sum_{k=1}^{m} \alpha_k \sum_{i=1}^{n} \rho_{\tau_k}(y_i - x_i^\top \beta_k) + \sum_{k=1}^{m} \frac{\lambda_k}{2} \|\beta_k\|_2^2 + J(\beta_1,\dots,\beta_m)$$

is $(\widehat{\beta}_0,\dots,\widehat{\beta}_0)$.

Proof. For each k, Theorem 1 applied with penalty λ_k/α_k gives the unique minimizer

$$\widehat{\beta}_0 = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \sum_{i=1}^{n} \rho_{\tau_k}(y_i - x_i^\top \beta) + \frac{\lambda_k}{2\alpha_k} \|\beta\|_2^2 \right\}.$$

Multiplication by $\alpha_k > 0$ preserves the minimizer. Hence, on Ω,

$$\sum_{k=1}^{m} \alpha_k \sum_{i=1}^{n} \rho_{\tau_k}(y_i - x_i^\top \beta_k) + \sum_{k=1}^{m} \frac{\lambda_k}{2} \|\beta_k\|_2^2$$

is uniquely minimized at $(\widehat{\beta}_0, \ldots, \widehat{\beta}_0)$. Adding $J \geq 0$, which equals zero there, preserves this minimizer. □

S.16.1 Finite-level monotonicity constraints

The next corollary applies Proposition 36 to monotone finite-level fits. Finite noncrossing constraints can certify equality of adjacent levels inside the extinction band; they do not create level separation there.

Let $\tau_1 < \cdots < \tau_m$ belong to $(0,1)$, let $\alpha_k > 0$ and $\lambda_k > 0$, let N be a positive integer, and let $Z \in \mathbb{R}^{N\times p}$. Define

$$\Omega_Z = \{(\beta_1, \ldots, \beta_m) \in (\mathbb{R}^p)^m : Z\beta_1 \leq Z\beta_2 \leq \cdots \leq Z\beta_m\},$$

with coordinatewise inequalities. Let $J_Z : \Omega_Z \to [0,\infty]$.

Corollary 21. *Suppose $J_Z(\widehat{\beta}_0, \ldots, \widehat{\beta}_0) = 0$. If*

$$0 < \lambda_k \leq \alpha_k \bar{\lambda}(\tau_k) \qquad (k = 1, \ldots, m),$$

then the unique minimizer over Ω_Z of

$$\sum_{k=1}^{m} \alpha_k \sum_{i=1}^{n} \rho_{\tau_k}(y_i - x_i^\top \beta_k) + \sum_{k=1}^{m} \frac{\lambda_k}{2} \|\beta_k\|_2^2 + J_Z(\beta_1, \ldots, \beta_m)$$

is $(\widehat{\beta}_0, \ldots, \widehat{\beta}_0)$. Also, for every positive integer q and every $C \in \mathbb{R}^{m\times q}$ satisfying $\mathbf{1}_m^\top C = 0$,

$$Z[\widehat{\beta}_0, \ldots, \widehat{\beta}_0]C = 0.$$

Proof. The tuple $(\widehat{\beta}_0, \ldots, \widehat{\beta}_0)$ belongs to Ω_Z because all adjacent finite-level predictions are equal. Proposition 36, applied with $\Omega = \Omega_Z$ and $J = J_Z$, proves the minimizer identity. The contrast identity follows from

$$Z[\widehat{\beta}_0, \ldots, \widehat{\beta}_0]C = Z\widehat{\beta}_0 \mathbf{1}_m^\top C = 0.$$

This completes the proof. □

S.17 Matrix-head quadratic coupling

In the fixed-feature model considered here, we couple several quantile heads through a shared quadratic penalty on the joint coefficient matrix. The following results identify the penalty regime in which every head interpolates its response column. When the coupled heads target the same response vector at different quantile levels, joint interpolation at the penalized optimum gives all heads the same coefficient vector.

Let $m \geq 1$. For $B \in \mathbb{R}^{p\times m}$, write b_ℓ for its ℓth column. For $Y \in \mathbb{R}^{n\times m}$, write y_ℓ for its ℓth column. Fix positive definite matrices $H = H^\top \in \mathbb{R}^{p\times p}$ and $M = M^\top \in \mathbb{R}^{m\times m}$, levels $\tau_\ell \in (0,1)$, and weights $\alpha_\ell > 0$. Put

$$K_H = XH^{-1}X^\top, \qquad Q_H = K_H^{-1}, \qquad B_{H,0} = H^{-1}X^\top Q_H Y, \qquad C_{H,M} = Q_H Y M.$$

For $\alpha > 0$, $\tau \in (0,1)$, and $c \in \mathbb{R}^n$, define

$$\kappa_{\alpha,\tau}(c) = \min_{i\leq n} \kappa_{\alpha,\tau,i}(c),$$

where

$$\kappa_{\alpha,\tau,i}(c) = \begin{cases} \alpha\tau/c_i, & c_i > 0, \\ \alpha(1-\tau)/|c_i|, & c_i < 0, \\ +\infty, & c_i = 0. \end{cases}$$

For $\lambda > 0$, define

$$\widehat{B}_{H,M}(\lambda) = \arg\min_{B \in \mathbb{R}^{p\times m}} \left\{ \sum_{\ell=1}^{m} \alpha_\ell \sum_{i=1}^{n} \rho_{\tau_\ell}(Y_{i\ell} - x_i^\top b_\ell) + \frac{\lambda}{2} \operatorname{tr}(B^\top HBM) \right\}.$$

The following theorem computes the exact, explicit penalty boundary at or below which the coupled heads cannot escape the interpolating matrix $B_{H,0}$.

Theorem 18. *For every $\lambda > 0$, the minimizer $\widehat{B}_{H,M}(\lambda)$ is unique and*

$$\widehat{B}_{H,M}(\lambda) = B_{H,0} \quad \Longleftrightarrow \quad 0 < \lambda \le \min_{1\le\ell\le m} \kappa_{\alpha_\ell,\tau_\ell}(C_{H,M,\cdot\ell}).$$

Proof. Because $H \succ 0$ and $M \succ 0$,

$$\operatorname{tr}(B^\top HBM) = \left\| H^{1/2} B M^{1/2} \right\|_F^2,$$

where $\|\cdot\|_F$ is the Frobenius norm. The positive definite quadratic makes the objective strongly convex, giving uniqueness.

With $r_{i\ell} = Y_{i\ell} - x_i^\top b_\ell$, convex optimality holds exactly when some $G \in \mathbb{R}^{n\times m}$ satisfies

$$G_{i\ell} \in \alpha_\ell \partial\rho_{\tau_\ell}(r_{i\ell}), \qquad \lambda HBM = X^\top G.$$

The matrix $B_{H,0}$ interpolates the response matrix because

$$XB_{H,0} = XH^{-1}X^\top Q_H Y = Y.$$

At $B_{H,0}$ the residuals are zero, and

$$HB_{H,0}M = X^\top Q_H Y M = X^\top C_{H,M}.$$

Stationarity at $B_{H,0}$ is equivalent to $X^\top(G - \lambda C_{H,M}) = 0$. Columnwise injectivity of $X^\top$ gives $G = \lambda C_{H,M}$. Hence

$$\lambda C_{H,M,i\ell} \in [\alpha_\ell(\tau_\ell - 1), \alpha_\ell\tau_\ell] \qquad (i \le n,\ \ell \le m).$$

Positive and negative coordinates give the stated caps; zero coordinates impose none. Taking the minimum over coordinates and heads proves the equivalence. □

The following corollary identifies the effective penalty of each head when the coupling matrix is diagonal.

Corollary 22. *Assume $M = \operatorname{diag}(m_1, \ldots, m_m)$ with $m_\ell > 0$. Put $a_{H,\ell} = Q_H y_\ell$. Then*

$$\widehat{B}_{H,M}(\lambda) = B_{H,0} \quad \Longleftrightarrow \quad 0 < \frac{\lambda m_\ell}{\alpha_\ell} \le \kappa_{1,\tau_\ell}(a_{H,\ell}) \qquad (\ell = 1, \ldots, m).$$

Proof. The diagonal form gives

$$C_{H,M,\cdot\ell} = m_\ell Q_H y_\ell = m_\ell a_{H,\ell}.$$

For $m_\ell > 0$,

$$\kappa_{\alpha_\ell,\tau_\ell}(m_\ell a_{H,\ell}) = \frac{\alpha_\ell}{m_\ell} \kappa_{1,\tau_\ell}(a_{H,\ell}).$$

Theorem 18 gives the claim. □

The following corollary proves that a shared response yields identical coefficient columns below the joint matrix-head boundary.

Corollary 23. *Assume $Y = y\mathbf{1}_m^\top$. Put*

$$a_H = Q_H y, \qquad \widehat{\beta}_{H,0} = H^{-1}X^\top a_H, \qquad s = M\mathbf{1}_m.$$

Then

$$\widehat{B}_{H,M}(\lambda) = \widehat{\beta}_{H,0}\mathbf{1}_m^\top \quad \Longleftrightarrow \quad 0 < \lambda \le \min_{1\le \ell\le m} \kappa_{\alpha_\ell,\tau_\ell}(s_\ell a_H).$$

On this interval, for positive integers N, q, every $Z \in \mathbb{R}^{N\times p}$, and every $D \in \mathbb{R}^{m\times q}$ satisfying $\mathbf{1}_m^\top D = 0$,

$$Z\widehat{B}_{H,M}(\lambda)D = 0.$$

Proof. The identity $Y = y\mathbf{1}_m^\top$ gives

$$B_{H,0} = H^{-1}X^\top Q_H y\mathbf{1}_m^\top = \widehat{\beta}_{H,0}\mathbf{1}_m^\top, \qquad C_{H,M} = Q_H y\mathbf{1}_m^\top M = a_H s^\top.$$

The equivalence follows from Theorem 18. On the stated interval,

$$Z\widehat{B}_{H,M}(\lambda)D = Z\widehat{\beta}_{H,0}\mathbf{1}_m^\top D = 0.$$

This completes the proof. □

The following proposition shows that validation predictions and check-loss scores cannot distinguish penalties inside the matrix-head extinction interval.

Proposition 37. *Use the notation of Theorem 18. Let N be a positive integer, let $Z^{\mathrm{val}} \in \mathbb{R}^{N\times p}$ and $Y^{\mathrm{val}} \in \mathbb{R}^{N\times m}$ be arbitrary, and write $(z_j^{\mathrm{val}})^\top$ for row j of Z^{val}. Put*

$$\Lambda_* = \min_{1\le \ell\le m} \kappa_{\alpha_\ell,\tau_\ell}(C_{H,M,\cdot\ell}).$$

For $0 < \lambda, \mu \le \Lambda_*$,

$$Z^{\mathrm{val}}\widehat{B}_{H,M}(\lambda) = Z^{\mathrm{val}}B_{H,0} = Z^{\mathrm{val}}\widehat{B}_{H,M}(\mu).$$

Writing $\widehat{b}_\ell(\lambda)$ for column ℓ of $\widehat{B}_{H,M}(\lambda)$, the finite validation score

$$V(\lambda) = \sum_{\ell=1}^{m}\alpha_\ell \sum_{j=1}^{N}\rho_{\tau_\ell}\{Y_{j\ell}^{\mathrm{val}} - (z_j^{\mathrm{val}})^\top\widehat{b}_\ell(\lambda)\}$$

satisfies $V(\lambda) = V(\mu)$.

Proof. Theorem 18 gives

$$\widehat{B}_{H,M}(\lambda) = B_{H,0} = \widehat{B}_{H,M}(\mu).$$

Multiplication by Z^{val} gives identical predictions, hence $V(\lambda) = V(\mu)$. This completes the proof. □

S.18 Face atlases and finite prediction maps

The subsequent results extend the fixed-feature KKT formulas from pointwise identities to finite, global certificates that cover entire regions of the penalty and level sets. For disjoint $L, U \subset \{1, \dots, n\}$, put $A = L \cup U$. If $A = \emptyset$, define

$$\mathcal{R}_{\emptyset,\emptyset} = \{(\lambda,\tau) \in (0,\infty)\times(0,1) : \tau - 1 < \lambda a_i < \tau \text{ for every } i \le n\}.$$

If $A \neq \emptyset$, let $\mathcal{R}_{L,U}$ be the set in Theorem 17.

The following proposition shows that these disjoint face regions cover the whole penalty-level domain.

Proposition 38. *The sets $\mathcal{R}_{L,U}$ satisfy*

$$(0,\infty)\times(0,1) = \bigcup_{L\cap U=\emptyset}\mathcal{R}_{L,U}.$$

For $(\lambda,\tau) \in \mathcal{R}_{\emptyset,\emptyset}$,

$$\widehat{\beta}(\lambda,\tau) = \widehat{\beta}_0, \qquad r_{\lambda,\tau} = 0, \qquad g_{\lambda,\tau} = \lambda a.$$

For $L \cup U \neq \emptyset$, the identities in Theorem 17 hold. Each $\mathcal{R}_{L,U}$ is an intersection of finitely many open or closed affine half-planes in (λ,τ).

Proof. The half-plane assertion for $L \cup U \neq \emptyset$ is Proposition 25. For $A = \emptyset$, the defining inequalities are

$$\lambda a_i - \tau < 0, \qquad \lambda a_i - \tau + 1 > 0, \qquad i \leq n,$$

which are strict affine inequalities. If $(\lambda, \tau) \in \mathcal{R}_{\emptyset,\emptyset}$, then λa lies in the interior of $\mathcal{B}_\tau$. Proposition 1 gives $g_{\lambda,\tau} = \lambda a$, and Theorem 1 gives the coefficient and residual identities.

For any (λ, τ), let

$$L = \{i : g_{\lambda,\tau,i} = \tau - 1\}, \qquad U = \{i : g_{\lambda,\tau,i} = \tau\}.$$

If $L \cup U \neq \emptyset$, the converse part of Theorem 17 gives $(\lambda, \tau) \in \mathcal{R}_{L,U}$. If $L \cup U = \emptyset$, the normal cone at the interior projection is $\{0\}$. Thus $g_{\lambda,\tau} = \lambda a$ and $(\lambda, \tau) \in \mathcal{R}_{\emptyset,\emptyset}$. This completes the proof. □

With the global domain partitioned into these stable tiles, the following proposition proves that tracing the coefficient map across any compact penalty interval involves crossing at most a finite number of active-set boundaries.

Fix $\tau \in (0, 1)$ and $0 < \lambda_- < \lambda_+$. For disjoint L, U, set

$$I_{L,U}(\tau) = \{\lambda \in [\lambda_-, \lambda_+] : (\lambda, \tau) \in \mathcal{R}_{L,U}\}.$$

Let $\mathcal{P}_\tau$ be the union of the boundaries, relative to $[\lambda_-, \lambda_+]$, of the nonempty intervals $I_{L,U}(\tau)$.

Proposition 39. *The set $\mathcal{P}_\tau$ is finite. On each connected component J of $[\lambda_-, \lambda_+] \setminus \mathcal{P}_\tau$, the actual lower and upper active sets are fixed, and there exist fixed vectors $b_0, b_1 \in \mathbb{R}^p$, $s_0, s_1, h_0, h_1 \in \mathbb{R}^n$ such that, for every $\lambda \in J$,*

$$\widehat{\beta}(\lambda, \tau) = b_0 + \lambda^{-1} b_1, \qquad r_{\lambda,\tau} = s_0 + \lambda^{-1} s_1, \qquad g_{\lambda,\tau} = h_0 + \lambda h_1.$$

Proof. By Proposition 38, each slice $I_{L,U}(\tau)$ is an interval or empty, with at most two relative boundary points. The finite number of pairs L, U gives finiteness of $\mathcal{P}_\tau$.

Let J be a connected component of $[\lambda_-, \lambda_+] \setminus \mathcal{P}_\tau$, and choose $\lambda_0 \in J$. Let L, U be the actual active sets at (λ_0, τ). Because $I_{L,U}(\tau)$ contains λ_0 and has no relative boundary in connected J, it contains J. Uniqueness of $g_{\lambda,\tau}$ fixes the active sets at L, U throughout J. If $L \cup U = \emptyset$, Proposition 38 gives the stated form with $b_1 = s_0 = s_1 = h_0 = 0$, $b_0 = \widehat{\beta}_0$, and $h_1 = a$. If $L \cup U \neq \emptyset$, Proposition 24 gives the stated form. This completes the proof. □

The following proposition combines face certificates at discrete levels into exact formulas for the coefficients, residuals, and predictions over a whole penalty interval, provided each chosen face remains valid throughout that interval.

Proposition 40. *Let $T = \{\tau_1, \ldots, \tau_m\} \subset (0, 1)$, let $0 < \lambda_- < \lambda_+$, let N be a positive integer, and let $Z \in \mathbb{R}^{N \times p}$. For each k, choose disjoint $L_k, U_k \subset \{1, \ldots, n\}$ and suppose*

$$(\lambda, \tau_k) \in \mathcal{R}_{L_k,U_k} \qquad (\lambda \in [\lambda_-, \lambda_+],\ k = 1, \ldots, m).$$

Then there exist matrices $B_0, B_1 \in \mathbb{R}^{p \times m}$ and $R_0, R_1 \in \mathbb{R}^{n \times m}$ such that, for every $\lambda \in [\lambda_-, \lambda_+]$,

$$[\widehat{\beta}(\lambda, \tau_1), \ldots, \widehat{\beta}(\lambda, \tau_m)] = B_0 + \lambda^{-1} B_1,$$

$$[r_{\lambda,\tau_1}, \ldots, r_{\lambda,\tau_m}] = R_0 + \lambda^{-1} R_1, \qquad Z[\widehat{\beta}(\lambda, \tau_1), \ldots, \widehat{\beta}(\lambda, \tau_m)] = ZB_0 + \lambda^{-1} ZB_1.$$

If $L_k = U_k = \emptyset$ for every k, then $B_1 = 0$ and, for every positive integer q and every $C \in \mathbb{R}^{m \times q}$ satisfying $\mathbf{1}_m^\top C = 0$,

$$Z[\widehat{\beta}(\lambda, \tau_1), \ldots, \widehat{\beta}(\lambda, \tau_m)]C = 0 \qquad (\lambda \in [\lambda_-, \lambda_+]).$$

The stated face condition follows by checking the defining affine inequalities at λ_- and λ_+, with strict inequality for the open constraints, after multiplying each residual sign inequality by the positive variable λ.

Proof. Fix k. If $L_k = U_k = \emptyset$, Proposition 38 gives

$$\widehat{\beta}(\lambda, \tau_k) = \widehat{\beta}_0, \qquad r_{\lambda,\tau_k} = 0$$

for every $\lambda \in [\lambda_-, \lambda_+]$. If $L_k \cup U_k \neq \emptyset$, Proposition 24 gives vectors $b_{0k}, b_{1k} \in \mathbb{R}^p$ and $s_{0k}, s_{1k} \in \mathbb{R}^n$ such that

$$\widehat{\beta}(\lambda, \tau_k) = b_{0k} + \lambda^{-1} b_{1k}, \qquad r_{\lambda,\tau_k} = s_{0k} + \lambda^{-1} s_{1k}$$

throughout the interval. Stacking these columnwise gives B_0, B_1, R_0, R_1. Multiplication by Z gives the prediction identity.

If all faces are empty, $B_0 = \widehat{\beta}_0 \mathbf{1}_m^\top$ and $B_1 = 0$. For $\mathbf{1}_m^\top C = 0$,

$$ZB_0C = Z\widehat{\beta}_0 \mathbf{1}_m^\top C = 0.$$

After multiplying residual sign inequalities by $\lambda > 0$, all constraints are affine in λ. Their extrema occur at the two ends of the interval, proving the final assertion. This completes the proof. □

Finally, the following corollary characterizes the geometric behavior of the predictions themselves: at a fixed positive penalty, they must be exactly piecewise affine across any compact interval of target quantile levels, and constant across any extinct interval cell.

Corollary 24. *Fix $\lambda > 0$, $0 < \tau_- < \tau_+ < 1$, let N be a positive integer, and let $Z \in \mathbb{R}^{N \times p}$. There is a finite set $\mathcal{D}_\lambda \subset [\tau_-, \tau_+]$ such that, on every connected component J of $[\tau_-, \tau_+] \setminus \mathcal{D}_\lambda$, there exist fixed vectors $p_0, p_1 \in \mathbb{R}^N$ satisfying*

$$Z\widehat{\beta}(\lambda, \tau) = p_0 + \tau p_1 \qquad (\tau \in J).$$

If $\lambda \leq \bar{\lambda}(\tau)$ for every $\tau \in J$, then

$$Z\widehat{\beta}(\lambda, \tau) = Z\widehat{\beta}_0 \qquad (\tau \in J).$$

Proof. For each disjoint pair L, U, define

$$J_{L,U}(\lambda) = \{\tau \in [\tau_-, \tau_+] : (\lambda, \tau) \in \mathcal{R}_{L,U}\}.$$

By Proposition 38, each $J_{L,U}(\lambda)$ is empty or an interval defined by affine inequalities in τ. Let $\mathcal{D}_\lambda$ be the finite union of their boundaries relative to $[\tau_-, \tau_+]$. On a connected component of $[\tau_-, \tau_+] \setminus \mathcal{D}_\lambda$, the same argument as in Proposition 39 fixes the actual face. Proposition 25, with λ fixed, gives

$$\widehat{\beta}(\lambda, \tau) = b_0 + \tau b_1$$

on that component, with the all-free case included by taking $b_1 = 0$. Multiplication by Z gives the first identity. The final identity follows from Theorem 1. □

S.19 Local perturbation calculus

In this section, we derive exact, closed-form differential calculus for the estimator, its fitted values, and its residuals with respect to continuous changes in the penalty, the quantile level, and the observed responses, valid where all defining inequalities of a fixed face are strict. Because these derivatives are computable from the active-set data, they give first-order expansions of the estimator along parameter perturbations that remain within that face.

Fix disjoint $L, U \subset \{1, \ldots, n\}$, put $A = L \cup U \neq \emptyset$ and $F = A^c$, and let S_A be the Schur complement defined in Theorem 17, with $S_A = K$ when $F = \emptyset$. The following proposition uses the Schur complement of the kernel matrix to give exact local derivatives of the fitted coefficients and residuals while the active set remains unchanged.

Let $(\lambda_0, \tau_0) \in \mathcal{R}_{L,U}$. Put $u_0 = \lambda_0^{-1}$, let $\mathbf{1}_A \in \mathbb{R}^A$ be the vector of ones, let $\mathbf{1}_L \in \mathbb{R}^A$ have coordinate 1 on L and coordinate 0 on U, and set

$$c_A(\tau_0) = \tau_0 \mathbf{1}_A - \mathbf{1}_L.$$

Define $\xi^{(u)}, \xi^{(\tau)} \in \mathbb{R}^n$ by

$$\xi_A^{(u)} = c_A(\tau_0), \qquad \xi_A^{(\tau)} = u_0 \mathbf{1}_A.$$

If $F \neq \emptyset$, also set

$$\xi_F^{(u)} = -K_{FF}^{-1} K_{FA} c_A(\tau_0), \qquad \xi_F^{(\tau)} = -u_0 K_{FF}^{-1} K_{FA} \mathbf{1}_A.$$

If $F = \emptyset$, the formulas involving F are absent.

Proposition 41. *Assume that all inequalities defining $\mathcal{R}_{L,U}$ are strict at (λ_0, τ_0). Then $\widehat{\beta}(1/u, \tau)$ and $r_{1/u,\tau}$ are differentiable at (u_0, τ_0), and*

$$\frac{\partial}{\partial u}\widehat{\beta}(1/u,\tau)\Big|_{(u_0,\tau_0)} = X^\top \xi^{(u)}, \qquad \frac{\partial}{\partial \tau}\widehat{\beta}(1/u,\tau)\Big|_{(u_0,\tau_0)} = X^\top \xi^{(\tau)}.$$

The residual derivatives satisfy

$$\left\{\frac{\partial}{\partial u} r_{1/u,\tau}\Big|_{(u_0,\tau_0)}\right\}_F = 0, \qquad \left\{\frac{\partial}{\partial u} r_{1/u,\tau}\Big|_{(u_0,\tau_0)}\right\}_A = -S_A c_A(\tau_0),$$

$$\left\{\frac{\partial}{\partial \tau} r_{1/u,\tau}\Big|_{(u_0,\tau_0)}\right\}_F = 0, \qquad \left\{\frac{\partial}{\partial \tau} r_{1/u,\tau}\Big|_{(u_0,\tau_0)}\right\}_A = -u_0 S_A \mathbf{1}_A.$$

The derivatives with respect to λ at (λ_0, τ_0) are $-\lambda_0^{-2}$ times the derivatives with respect to u.

Proof. Strictness and continuity preserve the face near (λ_0, τ_0). By Theorem 17, with $u = \lambda^{-1}$,

$$d_A(u,\tau) = a_A - u(\tau \mathbf{1}_A - \mathbf{1}_L).$$

If $F \neq \emptyset$, then

$$d_F(u,\tau) = -K_{FF}^{-1} K_{FA} d_A(u,\tau),$$

and the case $F = \emptyset$ has only the active block. Differentiation gives

$$\frac{\partial d_A}{\partial u} = -c_A(\tau_0), \qquad \frac{\partial d_F}{\partial u} = K_{FF}^{-1} K_{FA} c_A(\tau_0),$$

$$\frac{\partial d_A}{\partial \tau} = -u_0 \mathbf{1}_A, \qquad \frac{\partial d_F}{\partial \tau} = u_0 K_{FF}^{-1} K_{FA} \mathbf{1}_A.$$

The coefficient formula $\widehat{\beta}(1/u,\tau) = \widehat{\beta}_0 - X^\top d(u,\tau)$ gives the two coefficient derivatives. The residual formula has $r_F = 0$ and

$$r_A(u,\tau) = S_A d_A(u,\tau).$$

Differentiation gives the residual derivatives; $du/d\lambda = -\lambda^{-2}$ gives the final statement. This completes the proof. □

When the observed responses themselves are subject to perturbation, the impact smoothly decouples across the model's geometry. The following proposition separates this first-order response effect into a coefficient derivative in the span of the free feature rows, and a residual derivative supported on the active observation coordinates.

Fix $\lambda > 0$, $\tau \in (0,1)$, and a direction $\delta y \in \mathbb{R}^n$. Set $y(t) = y + t\delta y$, $a(t) = Qy(t)$, and let $\widehat{\beta}_t$ and r_t be the estimator and residual formed from response $y(t)$. Let L and U be the lower and upper sets of the actual face at $t = 0$. Put $A = L \cup U$, $F = A^c$, and $\delta a = Q\delta y$. Let S_A be the Schur complement defined in Theorem 17, with $S_A = K$ when $F = \emptyset$.

If $A \neq \emptyset$, define $\zeta \in \mathbb{R}^n$ by

$$\zeta_A = 0, \qquad \zeta_F = \delta a_F + K_{FF}^{-1} K_{FA} \delta a_A$$

when $F \neq \emptyset$, and by $\zeta = 0$ when $F = \emptyset$. Let N be a positive integer and let $Z \in \mathbb{R}^{N \times p}$.

Proposition 42. *Assume that all defining inequalities of the actual face at $t = 0$ are strict. If $A = \emptyset$, then*

$$\frac{d}{dt}\widehat{\beta}_t\Big|_{t=0} = X^\top \delta a, \qquad \frac{d}{dt} r_t\Big|_{t=0} = 0.$$

If $A \neq \emptyset$, then

$$\frac{d}{dt}\widehat{\beta}_t\Big|_{t=0} = X^\top \zeta.$$

The residual derivative then satisfies

$$\left\{ \frac{d}{dt} r_t \Big|_{t=0} \right\}_F = 0, \qquad \left\{ \frac{d}{dt} r_t \Big|_{t=0} \right\}_A = S_A \delta a_A,$$

and

$$\frac{d}{dt} Z\widehat{\beta}_t \Big|_{t=0} = ZX^\top \zeta$$

in the case $A \neq \emptyset$, and equals $ZX^\top \delta a$ in the case $A = \emptyset$.

Proof. Strictness and continuity preserve the face for $|t| < \varepsilon$, for some $\varepsilon > 0$. If $A = \emptyset$, $\lambda a(t) \in \mathrm{int}(\mathcal{B}_\tau)$, and Theorem 1 gives

$$\widehat{\beta}_t = X^\top a(t), \qquad r_t = 0,$$

which proves the two derivatives in the all-free case.

Assume $A \neq \emptyset$. Theorem 16 applied to $a(t)$ gives

$$d_A(t) = a_A(t) - \lambda^{-1} c_A, \qquad r_A(t) = S_A d_A(t), \qquad r_F(t) = 0,$$

where c_A is fixed by L, U, and τ. If $F \neq \emptyset$, then

$$d_F(t) = -K_{FF}^{-1} K_{FA} d_A(t).$$

Differentiating at $t = 0$ gives

$$\dot{d}_A(0) = \delta a_A, \qquad \dot{d}_F(0) = -K_{FF}^{-1} K_{FA} \delta a_A.$$

The coefficient identity

$$\widehat{\beta}_t = X^\top a(t) - X^\top d(t)$$

gives

$$\frac{d}{dt} \widehat{\beta}_t \Big|_{t=0} = X^\top \{\delta a - \dot{d}(0)\}.$$

The vector in braces has active block zero and free block $\delta a_F + K_{FF}^{-1} K_{FA} \delta a_A$. Differentiate $r_A(t) = S_A d_A(t)$ and $r_F(t) = 0$ for the residuals, and multiply by Z for the prediction derivative. This completes the proof. □

S.20 VALIDATION SELECTION OVER FINITE CANDIDATES

We bound the population risk for clipped check loss over a finite set of candidate models. Candidate models for the same training sample and quantile level inside the extinction interval produce identical validation predictions, and no criterion computed only from those predictions on the same validation sample can distinguish them. The validation observations are assumed to be independent draws from the target distribution and independent of training, while clipping bounds the loss; the finite grids and fitted prediction maps are chosen using training data only, and all inequalities involving the exact boundary $\bar{\lambda}(\tau)$ are deterministic consequences of the fixed training sample.

For $B > 0$ define the clipped check loss

$$\rho_\tau^B(u) = \rho_\tau(u) \wedge B.$$

The following proposition separates training from validation. Conditioning on the realized training sample, it gives, with the stated conditional probability, a uniform upper bound on the population risk for clipped check loss of an empirical validation minimizer relative to the best candidate at each level.

Let $\mathcal{Z}$ be a measurable space, let P be a distribution on $\mathcal{Z} \times \mathbb{R}$, and let $\mathcal{D}_{\mathrm{tr}}$ denote a training sample. Let $B > 0$, $\delta \in (0, 1)$, $m \geq 1$, $L \geq 1$, and let $M_1, \ldots, M_L$ be positive integers. Conditional on $\mathcal{D}_{\mathrm{tr}}$, fix levels $\tau_1, \ldots, \tau_L \in (0, 1)$ and measurable maps $f_{\ell k} : \mathcal{Z} \to \mathbb{R}$, $k = 1, \ldots, M_\ell$, with

$$N = \sum_{\ell=1}^{L} M_\ell.$$

Let $(Z_j^{\mathrm{val}}, Y_j^{\mathrm{val}})_{j=1}^m$ be independent draws from P and independent of $\mathcal{D}_{\mathrm{tr}}$. For (Z, Y) with distribution P and independent of $\mathcal{D}_{\mathrm{tr}}$, define

$$R_{\ell k} = \mathbb{E}\{\rho_{\tau_\ell}^B(Y - f_{\ell k}(Z)) \mid \mathcal{D}_{\mathrm{tr}}\}, \qquad \widehat{R}_{\ell k} = \frac{1}{m}\sum_{j=1}^m \rho_{\tau_\ell}^B(Y_j^{\mathrm{val}} - f_{\ell k}(Z_j^{\mathrm{val}})).$$

For each ℓ, let $\widehat{k}_\ell$ be a measurable element of

$$\arg\min_{1\le k\le M_\ell} \widehat{R}_{\ell k}.$$

Proposition 43. *Under this sampling scheme, with conditional probability at least* $1-\delta$*,*

$$R_{\ell\widehat{k}_\ell} \le \min_{1\le k\le M_\ell} R_{\ell k} + 2B\left\{\frac{\log(2N/\delta)}{2m}\right\}^{1/2} \qquad (\ell = 1, \dots, L).$$

Proof. Condition on $\mathcal{D}_{\mathrm{tr}}$ and on the finite collection of maps. For each pair (ℓ, k), the variables

$$W_{j\ell k} = \rho_{\tau_\ell}^B(Y_j^{\mathrm{val}} - f_{\ell k}(Z_j^{\mathrm{val}})), \qquad j = 1, \dots, m,$$

are independent, take values in $[0, B]$, and have conditional mean $R_{\ell k}$. Hoeffding's inequality (Hoeffding, 1963) gives, for every $t > 0$,

$$\mathbb{P}\{|\widehat{R}_{\ell k} - R_{\ell k}| > t \mid \mathcal{D}_{\mathrm{tr}}\} \le 2\exp(-2mt^2/B^2).$$

With

$$t = B\left\{\frac{\log(2N/\delta)}{2m}\right\}^{1/2},$$

a union bound gives

$$|\widehat{R}_{\ell k} - R_{\ell k}| \le t \qquad (\ell = 1, \dots, L,\ k = 1, \dots, M_\ell)$$

with conditional probability at least $1-\delta$. On this event, choose $k_\ell^* \in \arg\min_{1\le k\le M_\ell} R_{\ell k}$. The definition of $\widehat{k}_\ell$ gives

$$R_{\ell\widehat{k}_\ell} \le \widehat{R}_{\ell\widehat{k}_\ell} + t \le \widehat{R}_{\ell k_\ell^*} + t \le R_{\ell k_\ell^*} + 2t.$$

This is the stated bound. □

The following proposition proves that validation cannot select among penalties inside the extinction interval because their fitted predictions are identical.

Fix $\tau \in (0, 1)$ and an arbitrary validation array $(x_j^{\mathrm{val}}, y_j^{\mathrm{val}})_{j=1}^m \subset \mathbb{R}^p \times \mathbb{R}$. For $B > 0$ and $\lambda > 0$, put

$$\widehat{R}^{\mathrm{val}}(\lambda, \tau) = \frac{1}{m}\sum_{j=1}^m \rho_\tau^B(y_j^{\mathrm{val}} - (x_j^{\mathrm{val}})^\top \widehat{\beta}(\lambda, \tau)).$$

Proposition 44. *If* $0 < \lambda \le \bar{\lambda}(\tau)$ *and* $0 < \mu \le \bar{\lambda}(\tau)$*, then*

$$(x_j^{\mathrm{val}})^\top \widehat{\beta}(\lambda, \tau) = (x_j^{\mathrm{val}})^\top \widehat{\beta}(\mu, \tau) = (x_j^{\mathrm{val}})^\top \widehat{\beta}_0 \qquad (j = 1, \dots, m)$$

and

$$\widehat{R}^{\mathrm{val}}(\lambda, \tau) = \widehat{R}^{\mathrm{val}}(\mu, \tau).$$

If a finite penalty set $\Lambda \subset (0, \bar{\lambda}(\tau)]$ *is nonempty, every element of* Λ *minimizes* $\widehat{R}^{\mathrm{val}}(\lambda, \tau)$ *over* $\lambda \in \Lambda$*.*

Proof. Theorem 1 gives

$$\widehat{\beta}(\lambda, \tau) = \widehat{\beta}_0 \qquad \text{and} \qquad \widehat{\beta}(\mu, \tau) = \widehat{\beta}_0.$$

The predictions and hence $\widehat{R}^{\mathrm{val}}$ agree. If $\Lambda \subset (0, \bar{\lambda}(\tau)]$, every pair agrees, giving argmin set Λ. □

The following corollary separates deterministic removal of extinct candidates from the finite-sample risk bound for the remaining candidates.

Let $\mathcal{Z}$ be a measurable space, let P be a distribution on $\mathcal{Z} \times \mathbb{R}$, and let $\mathcal{D}_{\mathrm{tr}}$ denote a training sample. Let $B > 0$, $\delta \in (0,1)$, $m \geq 1$, and $L \geq 1$. Let $(Z_j^{\mathrm{val}}, Y_j^{\mathrm{val}})_{j=1}^m$ be independent draws from P and independent of $\mathcal{D}_{\mathrm{tr}}$. Let $\psi : \mathcal{Z} \to \mathbb{R}^p$ be selected using $\mathcal{D}_{\mathrm{tr}}$ alone, with jointly measurable evaluation, and let the training design X have full row rank. The levels $\tau_1, \dots, \tau_L \in (0,1)$ and finite penalty grids $\Lambda_\ell \subset (0,\infty)$ may also be selected using $\mathcal{D}_{\mathrm{tr}}$ alone. Put

$$\Lambda_\ell^+ = \{\lambda \in \Lambda_\ell : \lambda > \bar{\lambda}(\tau_\ell)\}, \qquad M_\ell^+ = |\Lambda_\ell^+|, \qquad \mathcal{L}^+ = \{\ell : M_\ell^+ \geq 1\}, \qquad N^+ = \sum_{\ell \in \mathcal{L}^+} M_\ell^+.$$

For $\lambda \in \Lambda_\ell$, define

$$\widehat{R}_\ell^{\mathrm{val}}(\lambda) = \frac{1}{m} \sum_{j=1}^m \rho_{\tau_\ell}^B \left(Y_j^{\mathrm{val}} - \psi(Z_j^{\mathrm{val}})^\top \widehat{\beta}(\lambda, \tau_\ell) \right).$$

For each $\ell \in \mathcal{L}^+$, enumerate

$$\Lambda_\ell^+ = \{\lambda_{\ell 1}, \dots, \lambda_{\ell M_\ell^+}\}$$

and define

$$f_{\ell k}(z) = \psi(z)^\top \widehat{\beta}(\lambda_{\ell k}, \tau_\ell), \qquad k = 1, \dots, M_\ell^+.$$

For (Z, Y) with distribution P and independent of $\mathcal{D}_{\mathrm{tr}}$, set

$$R_{\ell k} = \mathbb{E}\{\rho_{\tau_\ell}^B(Y - f_{\ell k}(Z)) \mid \mathcal{D}_{\mathrm{tr}}\}, \qquad \widehat{R}_{\ell k} = \frac{1}{m} \sum_{j=1}^m \rho_{\tau_\ell}^B(Y_j^{\mathrm{val}} - f_{\ell k}(Z_j^{\mathrm{val}})).$$

For each $\ell \in \mathcal{L}^+$, let $\widehat{k}_\ell$ be a measurable element of $\arg\min_{1 \leq k \leq M_\ell^+} \widehat{R}_{\ell k}$.

The kept class has its own validation comparison, conditional on every choice made from the training data.

Corollary 25. *Assume $N^+ \geq 1$. With conditional probability at least $1 - \delta$,*

$$R_{\ell \widehat{k}_\ell} \leq \min_{1 \leq k \leq M_\ell^+} R_{\ell k} + 2B \left\{ \frac{\log(2N^+/\delta)}{2m} \right\}^{1/2} \qquad (\ell \in \mathcal{L}^+).$$

For every fixed ℓ and every $\lambda, \mu \in \Lambda_\ell \setminus \Lambda_\ell^+$,

$$\psi(Z_j^{\mathrm{val}})^\top \widehat{\beta}(\lambda, \tau_\ell) = \psi(Z_j^{\mathrm{val}})^\top \widehat{\beta}(\mu, \tau_\ell) = \psi(Z_j^{\mathrm{val}})^\top \widehat{\beta}_0 \qquad (j = 1, \dots, m),$$

and

$$\widehat{R}_\ell^{\mathrm{val}}(\lambda) = \widehat{R}_\ell^{\mathrm{val}}(\mu).$$

The oracle comparison is over kept candidates only; excluding the interpolant can change the comparison class.

Proof. Conditional on $\mathcal{D}_{\mathrm{tr}}$, the maps $f_{\ell k}$ and grids are fixed, and the validation sample has distribution P^m. Proposition 43, applied to the finite collection indexed by $\ell \in \mathcal{L}^+$ and $1 \leq k \leq M_\ell^+$, gives the validation-risk bound.

If $\lambda, \mu \in \Lambda_\ell \setminus \Lambda_\ell^+$, then

$$0 < \lambda \leq \bar{\lambda}(\tau_\ell), \qquad 0 < \mu \leq \bar{\lambda}(\tau_\ell).$$

Proposition 44, applied with validation covariates $\psi(Z_j^{\mathrm{val}})$, gives the prediction identities and equality of the empirical clipped validation risks. This completes the proof. □

S.20.1 Removal of duplicate predictions without changing the risk benchmark

Removing every extinct candidate restricts validation to nonextinct fits, but it can remove a useful predictor at an individual level. A different reduction keeps one representative of the extinct candidates. It then preserves every candidate prediction function and the oracle risk over the original grid. This distinction separates a restriction on the fitted model from a reduction that only avoids repeated validation scores.

Fix a level τ and a nonempty finite grid Λ after observing the training data. Let $\Lambda^0 = \{\lambda \in \Lambda : \lambda \leq \bar{\lambda}(\tau)\}$. If Λ^0 is nonempty, choose $\lambda^0 = \max \Lambda^0$ and set

$$\Lambda^{\mathrm{red}} = (\Lambda \setminus \Lambda^0) \cup \{\lambda^0\}.$$

If Λ^0 is empty, set $\Lambda^{\mathrm{red}} = \Lambda$. Put $f_\lambda(z) = \psi(z)^\top \widehat{\beta}(\lambda, \tau)$ and let $R(\lambda)$ and $\widehat{R}(\lambda)$ denote its population and validation risks for the clipped loss ρ_τ^B, conditionally on the training data. Let $\widehat{\lambda}_{\mathrm{red}}$ minimize $\widehat{R}$ over Λ^{red}, with a measurable rule for ties.

Proposition 45. *Under the sampling conditions of Proposition 43,*

$$\{f_\lambda : \lambda \in \Lambda^{\mathrm{red}}\} = \{f_\lambda : \lambda \in \Lambda\}, \qquad \min_{\lambda \in \Lambda^{\mathrm{red}}} R(\lambda) = \min_{\lambda \in \Lambda} R(\lambda).$$

With conditional probability at least $1 - \delta$*,*

$$R(\widehat{\lambda}_{\mathrm{red}}) \leq \min_{\lambda \in \Lambda} R(\lambda) + 2B \left\{ \frac{\log(2|\Lambda^{\mathrm{red}}|/\delta)}{2m} \right\}^{1/2}.$$

The minimum validation risk and the set of minimizing prediction functions also remain unchanged. If ties are resolved by the largest penalty, the selected penalty is unchanged.

The risk benchmark is now the entire original grid, including its extinct candidates. The smaller count enters the validation bound because predictions that coincide are scored only once. The result does not require the minimum-norm interpolant to be a poor predictor, and it requires the same rule for ties when identifying the selected penalty.

Proof. For every $\lambda \in \Lambda^0$, Theorem 1 gives $f_\lambda(z) = \psi(z)^\top \widehat{\beta}_0$ for every z. The representative λ^0 keeps this function when $\Lambda^0 \neq \emptyset$; every nonextinct candidate remains. Thus the function sets, population and validation minima, and minimizing functions agree. Apply Proposition 43 with one level and $|\Lambda^{\mathrm{red}}|$ candidates, and replace its oracle minimum by the equal minimum over Λ. Because λ^0 is the largest extinct candidate and all nonextinct candidates remain, the largest-penalty tie rule is preserved. This completes the proof. □

The following example shows that the distinction is necessary. Take $n = 1$, $p = 2$, $X = (1, 0)$, $y = 1$, and $\tau = 1/2$. Then $a = 1$, $\widehat{\beta}_0 = (1, 0)^\top$ and $\bar{\lambda}(1/2) = 1/2$. For the grid $\Lambda = \{1/4, 1\}$, direct minimization gives

$$\widehat{\beta}(1/4, 1/2) = (1, 0)^\top, \qquad \widehat{\beta}(1, 1/2) = (1/2, 0)^\top.$$

Let a new covariate equal $(1, 0)^\top$ and let its response equal 1 with probability $3/4$ and 2 with probability $1/4$. This population can also generate the stated training observation with positive probability. For $B \geq 1$, clipping changes none of the losses, and

$$R(1/4) = \frac{1}{8}, \qquad R(1) = \frac{3}{8}.$$

Every possible validation observation increases the loss by $1/4$ when the extinct candidate is discarded. Thus complete exclusion need not preserve the original-grid oracle, even with unlimited validation data. Corollary 25 correctly uses the oracle among the kept candidates. Proposition 45 preserves the original-grid oracle by keeping one representative. Neither result asserts that avoiding extinction alone ensures calibration or consistent quantile estimation.

S.20.2 Validation with one interpolating candidate

We apply the reduction in Proposition 45 to the candidate fits from the Gaussian study in Section S.25 and the diabetes study in Section S.28. Each study has 120 repetitions and five levels. For each level, we keep the largest of the five extinct penalties and all 23 nonextinct penalties. The reduction leaves 24 of the original 28 candidates. Write $\widehat{R}(\lambda)$ for the original mean validation pinball loss at penalty λ. We use these scores and the same rule for ties: scores within $10^{-12}\max\{1, |\min_\lambda \widehat{R}(\lambda)|\}$ of the minimum are tied, and the largest penalty is selected.

In each study, this reduction removes 2400 of 16800 candidate evaluations and preserves all 600 selected candidate indices. Hence it also preserves their test pinball risks and calibration errors. The Gaussian study selects no interpolating candidate; the diabetes study selects one, which the reduction keeps. These results verify the prediction-preserving property of the reduction, including a case in which the interpolant is selected. They do not require a claim of lower risk than validation over the original grid. The script `scripts/check_representative_validation.py` performs this comparison using the existing fitted candidates without refitting them.

S.21 Validation risk under computation reinvestment

While evaluating additional penalty candidates can theoretically improve the best available prediction, the finite-candidate validation guarantee used here must account for the statistical cost of every candidate that could have been evaluated.

S.21.1 Reinvesting computations in additional penalties

Fix one level and a training-measurable original grid Λ. Let Λ^{red} keep one extinct representative as in Proposition 45. Use the saved computation to construct a finite set $A \subset (\bar{\lambda}(\tau), \infty)$ from training data only, and put $\Lambda^{\mathrm{aug}} = \Lambda^{\mathrm{red}} \cup A$. Suppose the candidate losses are bounded by H and $|\Lambda^{\mathrm{aug}}| \le J$. Here $R(\lambda)$ and $\widehat{R}(\lambda)$ denote the population and validation means of the ordinary pinball loss. Let $\widehat{\lambda}_{\mathrm{aug}}$ minimize $\widehat{R}$ over Λ^{aug}, with a measurable rule for ties. Define

$$\Delta = \min_{\lambda\in\Lambda} R(\lambda) - \min_{\lambda\in\Lambda^{\mathrm{aug}}} R(\lambda) \ge 0.$$

The inequality follows from preservation of all original prediction functions. Applying the same Hoeffding argument to ordinary bounded check loss gives, with conditional probability at least $1-\delta$,

$$R(\widehat{\lambda}_{\mathrm{aug}}) \le \min_{\lambda\in\Lambda} R(\lambda) - \Delta + 2H\sqrt{\frac{\log(2J/\delta)}{2m}}.$$

Additional candidates reduce the risk benchmark by Δ, while the selection penalty grows as the square root of the logarithm of the number of candidates in the predetermined collection. With conditional probability at least $1-\delta$, the selected augmented-grid predictor has lower population risk than the original-grid oracle whenever Δ exceeds the validation term. Duplicate removal alone has $\Delta = 0$. Theorem 6 provides a target-relative bound when the grids also satisfy its conditions. A computation budget limits the number of new candidates through their actual costs; equal candidate counts do not imply equal computation. If validation scores determine the evaluation order, fix a finite collection of all possible candidate penalties from training data first and count that entire collection in J. The uniform deviation bound then applies to every visited subset, including subsets chosen using validation scores, with its risk benchmark given by the best completed candidate. The original-grid guarantee applies when the reduced original grid has been completed. Counting only the adaptively visited candidates would not justify this guarantee; the bound counts the whole predetermined collection.

S.22 Risk and calibration after reinvesting computation

To evaluate reinvestment, we use the bounded trigonometric feature class introduced in Section A.13, with $b_1 = 1$, $b_2 = 0.5$, $\eta = 0.5$, $n = 80$, and $p = 257$. The noise is uniform on $[-\sigma, \sigma]$, with $\sigma = 0.5$ or 1. Each setting has 120 independent training samples, 12000 independent

validation observations, and 8192 independent test covariates per repetition. All methods use the levels $0.10, 0.25, 0.50, 0.75, 0.90$ and the same observations. We keep every draw without sign conditioning or rejection. The production seed is 202609081, with independent random streams indexed by setting and repetition. A separate pilot seed is 202609080.

The original grid has 17 summed-loss penalties $10^{-8+10j/16}$, $j = 0, \ldots, 16$. The original procedure evaluates every penalty in decreasing order and selects the smallest validation loss. Duplicate removal keeps the largest extinct penalty and all nonextinct penalties, giving exactly the same selected predictor. For reinvestment, first evaluate this reduced grid. Between consecutive nonextinct grid penalties, and between the extinction boundary and the smallest nonextinct grid penalty, place candidates at the quarter, half, and three-quarter points on the logarithmic scale. These locations are determined from training data only. Next evaluate the unused candidate nearest to the current validation minimizer on the logarithmic scale, resolving equal distances by the smaller penalty. Recompute the minimizer after each completed evaluation. Validation ties within $10^{-12} \max\{1, |\min \widehat{R}|\}$ are resolved by the largest penalty. The independent test observations are never used to construct, order, or select candidates. The finite-candidate bound in Section S.21.1 counts every possible candidate in the predetermined collection, including candidates that the budget does not permit evaluating.

All three procedures use the same cached $K = XX^\top$, Cholesky factorization, and interpolation vector, with initial dual point $\mathrm{proj}_{\mathcal{B}_\tau}(\lambda a)$, where $\mathrm{proj}_{\mathcal{B}_\tau}$ is Euclidean projection, obtained by clipping each coordinate to $[\tau - 1, \tau]$. Thus the comparison credits the original procedure with an exact interpolation initial point. Fits use L-BFGS-B and a subsequent active-set solve. A primal–dual gap verifies each completed candidate before validation. The largest relative gap, normalized by one plus the primal objective, is 3.141×10^{-12} over all 46874 completed candidate evaluations. The driver is `scripts/run_budget_reinvestment.py`; the independent validator checks all selected indices, costs, summaries, and saved prediction calculations.

The recorded matrix-arithmetic cost charges $n^2p + n^3/6 + n^2$ computational units for the full dense Gram product, the Cholesky factorization, and the two required triangular solves, where one unit represents a matrix multiply-add in the stated cost model. Each solver matrix-vector product costs n^2; each active-set factorization and solve is charged by its actual submatrix dimension; coefficient recovery and the primal–dual certificate each cost np; and validation prediction costs mp, with $m = 12000$. The boundary scan is charged n units, and attempted computations for an unfinished candidate remain in the total. The original full-grid cost sets the cap separately for each repetition and level. The reinvested procedure uses 98.21% and 97.65% of this cap on average at the two noise levels. The remaining amount cannot complete the next charged operation. Duplicate removal alone uses 44.62% and 48.28% of the original arithmetic budget on average for $\sigma = 0.5$ and 1, respectively, and preserves every selected penalty and its risk. On average about 10 of the 17 original penalties are extinct, and this proportion determines the size of the saving.

Matrix operations do not include all elapsed-time costs of the solver. We also recorded the completion time of every candidate using one numerical-library thread. The common Gram calculation and solve are timed once and charged equally to each procedure. Subsequent timing includes boundary computation, optimization, candidate ordering, optimality checks, and validation prediction and scoring; common feature generation and final test evaluation are excluded. At the elapsed time used by the original full grid, we select only among the completed candidates of the reinvested procedure. The mean original times are 0.02154 and 0.02256 seconds per level, and duplicate removal alone takes 0.01567 and 0.01715 seconds. The arithmetic-budget reinvestment takes 0.04168 and 0.03976 seconds per level, which is 1.94 and 1.76 times the original elapsed time. One of the 1200 elapsed-time checkpoints precedes completion of the reduced initial grid; it is kept in the results. These timings describe this implementation and machine.

The recorded matrix-arithmetic budget does not measure elapsed time.

For each test covariate, we integrate the conditional uniform response distribution exactly. Averaging these conditional risks over the 8192 test covariates estimates population pinball risk. Calibration is the absolute difference between the mean conditional coverage probability and the requested level. Thus the calibration comparison includes covariate Monte Carlo error but no additional simulation of coverage indicators. We first average each metric across levels within each repetition; Table S.2 reports the mean and standard deviation across the 120 repetitions. The known Bayes risks averaged over these levels equal 0.08050 and 0.16100, respectively.

Table S.2: Population-risk estimates and calibration under equal arithmetic and elapsed-time budgets. Numbers in parentheses represent standard deviations.

σ	Selection	Pinball risk	Calibration error	New candidates
0.5	Original grid	0.08580 (0.00229)	0.0589 (0.0177)	0.00
0.5	Reinvestment: arithmetic	0.08501 (0.00230)	0.0508 (0.0162)	8.14
0.5	Reinvestment: elapsed time	0.08545 (0.00233)	0.0583 (0.0161)	1.36
1.0	Original grid	0.16886 (0.00370)	0.0501 (0.0173)	0.00
1.0	Reinvestment: arithmetic	0.16818 (0.00369)	0.0463 (0.0171)	7.66
1.0	Reinvestment: elapsed time	0.16854 (0.00366)	0.0492 (0.0171)	1.34

The paired risk improvements under the arithmetic budget are 0.000791 and 0.000678, with Monte Carlo standard errors 0.000030 and 0.000040. The paired calibration improvements are 0.008129 and 0.003784, with standard errors 0.000662 and 0.000361. At the elapsed-time checkpoints, the paired risk improvements are 0.000352 and 0.000317, with standard errors 0.000032 and 0.000030, and the calibration improvements are 0.000622 and 0.000835, with standard errors 0.000411 and 0.000193. An approximate 95% Monte Carlo interval for the first elapsed-time calibration improvement is $[-0.000184, 0.001427]$, which does not establish a strict calibration gain. Under equal elapsed time, reinvestment adds about 1.35 candidates per level, and its mean risk reductions of 0.41% and 0.19% are smaller than the reductions of 0.92% and 0.40% under the arithmetic budget.

The arithmetic-budget search and the elapsed-time checkpoints each compare against the original full grid, not against the weak-penalty baseline. The original grid is deliberately wide and contains many extinct candidates; a grid that already excludes or consolidates them has less computation to recover. The experiment assesses the benefit of reinvesting duplicate evaluations in this stated design. It does not assert a universal risk gain for every target class, original grid, validation sample, or computing environment. Theorem 6 provides a sufficient population guarantee for its specified penalty range; these finite-sample experiments use the registered practical grid and do not claim to satisfy its conservative sample-size inequality.

S.23 Boundary screening under nonuniform heteroskedastic noise

We examine whether the fixed-grid boundary screening of Theorem 12 reduces computation without changing the predictions selected by full-grid validation. Boundary screening removes 1204 of 12000 candidate fits and selects the same prediction as full-grid validation in all 600 comparisons. The response distribution is nonuniform and heteroskedastic, and the search uses no coefficient bound.

We use the bounded trigonometric features of Section A.13 with $n = 80$, $p = 257$, and $\eta = 0.5$. The conditional mean is $\cos\zeta_0 + 0.5\sin\zeta_0$, the conditional noise scale is $0.75 + 0.25\cos\zeta_0$, and the independent error has density $3(1-u^2)/4$ on $[-1, 1]$. Thus the response follows the nonuniform example of Section S.3. We conduct 120 independent repetitions, each with 12000 independent validation observations and 8192 independent test covariates. The five levels are $0.10, 0.25, 0.50, 0.75, 0.90$, and the procedure uses the known response bound $M = \sqrt{1.25} + 1$.

All three procedures use the same 20 penalties $2^j/n$, $j = 0, \ldots, \lceil 3\log_2 n \rceil$, the same ridge solver, and predictions clipped to $[-M, M]$. Full-grid validation compares every candidate. The second procedure keeps the largest extinct penalty and every nonextinct penalty. Boundary screening excludes every extinct penalty. Each procedure selects the largest penalty among exact validation minimizers. Every candidate is fitted once to form a common set of predictions. Interpolating candidates use the exact minimum-norm coefficient vector, which makes duplicate validation scores identical. A primal–dual gap verifies each fit. All 120 training samples have full row rank, all fits meet the gap tolerance, and no screened candidate set is empty. The generation seed is 202609191, and the driver is `scripts/run_nonuniform_screened_validation.py`.

For each test covariate, we integrate the conditional response distribution to compute pinball risk and coverage. The conditional distribution function of the error is $F_U(u) = 1/2 + 3u/4 - u^3/4$ on $[-1, 1]$, with values zero and one outside this interval. Its integrated positive part is $\mathbb{E}(t - U)_+ =$

$3/16 + t/2 + 3t^2/8 - t^4/16$ for $|t| \leq 1$, zero for $t < -1$, and t for $t > 1$. These identities give the conditional pinball risk without simulating additional test responses. The calibration error is the absolute difference between mean conditional coverage and the requested level. Table S.3 first averages each metric over the five levels within a repetition and then reports its mean and standard deviation over repetitions.

Table S.3: Prediction and computation under nonuniform heteroskedastic noise. Numbers in parentheses represent standard deviations.

Selection	Pinball risk	Calibration error	Candidates
Full grid	0.09875 (0.00229)	0.0501 (0.0159)	20.00 (0.00)
One representative	0.09875 (0.00229)	0.0501 (0.0159)	18.99 (0.26)
Boundary screening	0.09875 (0.00229)	0.0501 (0.0159)	17.99 (0.26)

Keeping the largest extinct penalty as the representative under the stated tie rule removes 606 duplicate fits and preserves all 600 full-grid selections, as Proposition 45 guarantees. Complete screening removes 1204 fits, a 10.03% reduction, and preserves every selection in this study, although it has no general preservation guarantee. Both reductions leave prediction loss and calibration unchanged. Their mean recorded matrix-operation costs are 4.48% and 8.89% below that of the full grid. This count includes the common Gram calculation, factorization, initial solution, candidate fitting, prediction, and boundary scans. It excludes scalar clipping and scoring overhead and is not a measurement of elapsed-time improvement. The maximum relative primal–dual gap is 8.69×10^{-13}. The study evaluates the specified procedure at this sample size; it does not establish the sufficient sample-size inequality in Theorem 12.

S.24 Gaussian-boundary scaling under the random-feature model

This study evaluates the finite-sample scaling in Theorem 4. We set $\theta = 0$ and $\sigma = 1$ in model (2) and use four cells with $p = n^2$. In this case, $XX^\top$ has the same distribution as W/p for a Wishart matrix W with p degrees of freedom and identity scale. Each repetition generates this distribution exactly by Bartlett factorization and solves for $a = (XX^\top)^{-1}y$. The study records the exact uniform distance from the scaled boundary to $\tau \wedge (1-\tau)$, the error in its peak, the error $|\tau^\dagger - 1/2|$, and the errors at levels 0.10 and 0.90.

Table S.4: Gaussian-boundary scaling over 120 independent repetitions per cell. Numbers in parentheses represent standard deviations. Endpoint errors use $\tau = 0.10$ and $\tau = 0.90$.

n	p	n/p	Uniform error	Peak error	$\tau^\dagger$ error	Left error	Right error
50	2,500	0.02000	0.172 (0.096)	0.115	0.058	0.025	0.030
200	40,000	0.00500	0.140 (0.057)	0.099	0.041	0.023	0.022
800	640,000	0.00125	0.109 (0.047)	0.085	0.024	0.018	0.018
3200	10,240,000	0.00031	0.095 (0.037)	0.070	0.025	0.015	0.015

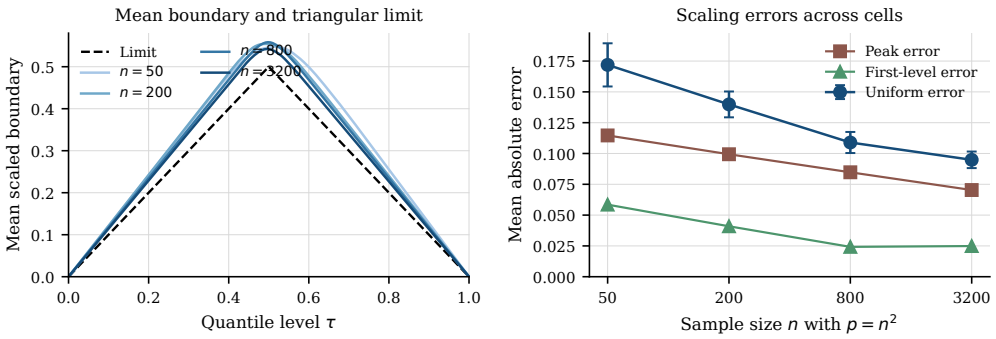

Figure S.2: Gaussian-boundary scaling over 120 repetitions per cell. Left: mean scaled boundaries. Right: mean errors, with two Monte Carlo standard errors for the uniform error.

Table S.4 and Figure S.2 show that the mean uniform error decreases from 0.172 at $n = 50$ to 0.095 at $n = 3200$, a reduction of 44.8%. The mean peak error decreases by 38.6%, and the mean first-extinct-level error decreases from 0.058 to 0.025. The mean endpoint errors decrease from 0.025 and 0.030 to 0.015 at both endpoints. These results give finite-sample evidence for the limits in Theorem 4 under the zero-signal Gaussian model. They do not assert finite-sample equality with the limiting triangle. The script `scripts/run_gaussian_boundary_scaling.py` generates 120 independent repetitions per cell with seed 20260909, and its validator checks every recorded boundary identity and summary value.

S.24.1 Dimension ratios and signal size

We evaluate the separate signal and noise contributions in Theorem 8 and the constant-feature extension in Corollary 6. We use $n = 80, 320, 1280$, $n/p = 0.25, 0.50, 0.80$, and signal variance $\|\theta\|_2^2/p$ equal to 0, 1, or $\sqrt{n}$, with $\sigma = 1$. For each setting, we consider independent Gaussian features and a penalized constant first feature with zero conditional-mean coefficient. A further Gaussian design has $p = n$ and uses only normalization by the observed length. The complete study has 63 settings and 120 independent repetitions per setting. All generated designs are kept.

The experiments sample the joint distribution of the Gram matrix and the signal exactly. For $p > n$, let $L \in \mathbb{R}^{n\times n}$ be lower triangular with positive diagonal, with $LL^\top$ having the Wishart distribution with $p-1$ degrees of freedom and identity scale. Let h be the first n coordinates of an independent uniform unit vector in $\mathbb{R}^{p-1}$. For signal variance t^2 with $t \geq 0$, generate

$$K = \frac{LL^\top + uu^\top}{p}, \qquad y = tLh + \varepsilon,$$

where $u \sim N(0, I_n)$ for the Gaussian design, $u = \mathbf{1}_n$ for the constant-feature design, and $\varepsilon \sim N(0, I_n)$ is independent. To verify this representation, write $GG^\top = LL^\top$ for a standard Gaussian $n \times (p-1)$ matrix G. Its right orthogonal invariance implies that $L^{-1}G$ is uniform among matrices with orthonormal rows and independent of L. Multiplication by a fixed unit signal direction gives the distribution of h. Thus tLh has the required joint distribution with K for a nonrandom signal satisfying $\theta_1 = 0$ and $\|\theta\|_2^2/p = t^2$. For $p = n$, we use p degrees of freedom, a unit vector in $\mathbb{R}^p$, and $K = LL^\top/p$. Triangular solves and the Sherman–Morrison identity compute $a = K^{-1}y$ without an approximation to its distribution.

For the two normalizations, define

$$D_r = \sup_{\tau\in(0,1)} \left| r_{n,p}\sqrt{2\log n}\,\bar{\lambda}(\tau) - \{\tau \wedge (1-\tau)\} \right|,$$

$$D_v = \sup_{\tau\in(0,1)} \left| v_{n,p}\sqrt{2\log n}\,\bar{\lambda}(\tau) - \{\tau \wedge (1-\tau)\} \right|.$$

Both errors are evaluated exactly at the knots 0, $\tau^\dagger$, $1/2$, and 1, where the absolute difference of the two piecewise affine functions attains its maximum. Tables S.5–S.7 report every setting. The entries marked by a dash concern the square design, for which no population scale is asserted.

Across all 18 proportional-dimension settings with $n = 1280$, the mean ratio $r_{n,p}^2/v_{n,p}^2$ lies between 0.995 and 1.003. For unit signal variance and Gaussian features, the corresponding ratios using marginal response variance are 1.855, 4.978, and 64.921 at dimension ratios 0.25, 0.50, and 0.80. These values show why the signal and noise corrections are needed. The mean D_v at $n = 1280$ lies between 0.099 and 0.112 over all proportional-dimension settings. For the square Gaussian design with unit signal variance, the mean D_r decreases from 0.173 to 0.123 and 0.102 as n increases. Figure S.3 shows this length and shape comparison. The finite-sample error remains positive because the extreme-value limit converges slowly.

The script `scripts/run_gaussian_ratio_boundary.py` uses seed 20260905 and produces 7560 records. Each record contains the signed maxima, interpolation-vector norm, normalization errors, and linear-system residual norms. The independent validator recomputes the normalization errors and every mean, standard deviation, and Monte Carlo standard error. Within each repetition, the signal sizes and the two feature designs share the Gaussian base matrix and noise; repetitions are independent within each setting.

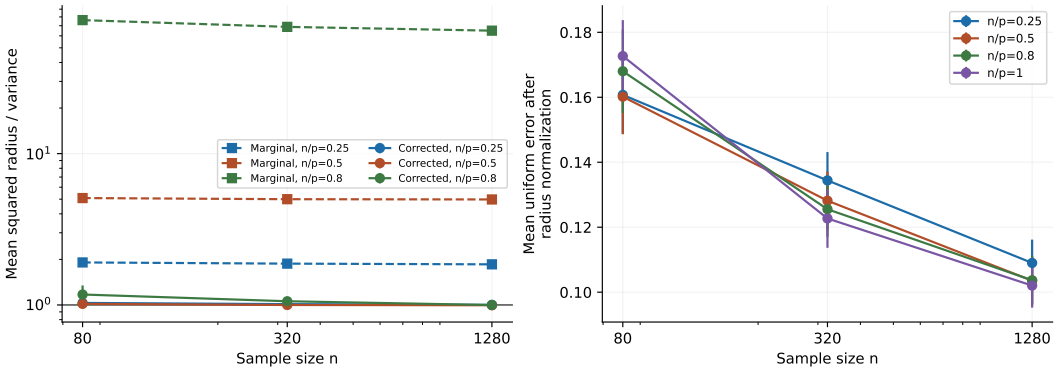

Figure S.3: Gaussian boundary normalization with unit signal variance. Error bars represent two Monte Carlo standard errors.

Table S.5: Gaussian boundary normalization with signal variance 0. Numbers in parentheses represent standard deviations.

n	n/p	Design	$r_{n,p}^2/v_{n,p}^2$	D_v	D_r
80	0.25	Constant	1.026 (0.263)	0.179 (0.094)	0.165 (0.067)
80	0.25	Gaussian	1.024 (0.264)	0.182 (0.095)	0.167 (0.067)
80	0.50	Constant	1.033 (0.358)	0.180 (0.105)	0.150 (0.058)
80	0.50	Gaussian	1.031 (0.357)	0.182 (0.110)	0.151 (0.060)
80	0.80	Constant	1.184 (1.221)	0.274 (0.222)	0.153 (0.057)
80	0.80	Gaussian	1.163 (0.967)	0.275 (0.217)	0.169 (0.067)
80	1.00	Gaussian	–	–	0.173 (0.061)
320	0.25	Constant	1.013 (0.144)	0.137 (0.060)	0.134 (0.047)
320	0.25	Gaussian	1.013 (0.144)	0.137 (0.060)	0.135 (0.047)
320	0.50	Constant	1.002 (0.188)	0.136 (0.068)	0.125 (0.050)
320	0.50	Gaussian	1.001 (0.187)	0.139 (0.069)	0.128 (0.051)
320	0.80	Constant	1.061 (0.375)	0.153 (0.088)	0.129 (0.045)
320	0.80	Gaussian	1.058 (0.375)	0.152 (0.089)	0.127 (0.045)
320	1.00	Gaussian	–	–	0.123 (0.049)
1280	0.25	Constant	1.001 (0.060)	0.105 (0.043)	0.105 (0.041)
1280	0.25	Gaussian	1.001 (0.061)	0.104 (0.043)	0.105 (0.040)
1280	0.50	Constant	1.003 (0.097)	0.105 (0.044)	0.104 (0.034)
1280	0.50	Gaussian	1.002 (0.097)	0.105 (0.044)	0.104 (0.034)
1280	0.80	Constant	1.000 (0.184)	0.110 (0.061)	0.101 (0.040)
1280	0.80	Gaussian	1.000 (0.183)	0.111 (0.060)	0.101 (0.041)
1280	1.00	Gaussian	–	–	0.102 (0.037)

Table S.6: Gaussian boundary normalization with signal variance 1. Numbers in parentheses represent standard deviations.

n	n/p	Design	$r^2_{n,p}/v^2_{n,p}$	D_v	D_r
80	0.25	Constant	1.032 (0.226)	0.167 (0.079)	0.158 (0.063)
80	0.25	Gaussian	1.031 (0.225)	0.169 (0.082)	0.161 (0.066)
80	0.50	Constant	1.014 (0.319)	0.181 (0.103)	0.151 (0.058)
80	0.50	Gaussian	1.017 (0.314)	0.187 (0.107)	0.160 (0.063)
80	0.80	Constant	1.202 (1.179)	0.260 (0.205)	0.154 (0.059)
80	0.80	Gaussian	1.172 (0.949)	0.266 (0.199)	0.168 (0.070)
80	1.00	Gaussian	–	–	0.173 (0.061)
320	0.25	Constant	1.013 (0.119)	0.134 (0.052)	0.134 (0.047)
320	0.25	Gaussian	1.013 (0.120)	0.134 (0.053)	0.134 (0.048)
320	0.50	Constant	0.999 (0.165)	0.136 (0.065)	0.126 (0.048)
320	0.50	Gaussian	1.000 (0.164)	0.137 (0.066)	0.128 (0.049)
320	0.80	Constant	1.060 (0.362)	0.150 (0.087)	0.129 (0.043)
320	0.80	Gaussian	1.058 (0.361)	0.147 (0.088)	0.126 (0.046)
320	1.00	Gaussian	–	–	0.123 (0.049)
1280	0.25	Constant	1.001 (0.054)	0.110 (0.041)	0.109 (0.039)
1280	0.25	Gaussian	1.002 (0.054)	0.110 (0.041)	0.109 (0.039)
1280	0.50	Constant	0.996 (0.080)	0.106 (0.045)	0.104 (0.041)
1280	0.50	Gaussian	0.996 (0.080)	0.106 (0.045)	0.104 (0.040)
1280	0.80	Constant	0.998 (0.183)	0.112 (0.062)	0.103 (0.041)
1280	0.80	Gaussian	0.999 (0.181)	0.112 (0.061)	0.104 (0.041)
1280	1.00	Gaussian	–	–	0.102 (0.037)

Table S.7: Gaussian boundary normalization with signal variance $\sqrt{n}$. Numbers in parentheses represent standard deviations.

n	n/p	Design	$r^2_{n,p}/v^2_{n,p}$	D_v	D_r
80	0.25	Constant	1.023 (0.178)	0.168 (0.072)	0.165 (0.059)
80	0.25	Gaussian	1.023 (0.179)	0.171 (0.073)	0.168 (0.060)
80	0.50	Constant	0.998 (0.245)	0.170 (0.088)	0.154 (0.060)
80	0.50	Gaussian	1.006 (0.244)	0.173 (0.091)	0.160 (0.065)
80	0.80	Constant	1.204 (0.946)	0.220 (0.150)	0.153 (0.063)
80	0.80	Gaussian	1.168 (0.810)	0.234 (0.160)	0.161 (0.066)
80	1.00	Gaussian	–	–	0.174 (0.060)
320	0.25	Constant	1.009 (0.086)	0.120 (0.048)	0.120 (0.047)
320	0.25	Gaussian	1.009 (0.087)	0.121 (0.048)	0.121 (0.048)
320	0.50	Constant	0.997 (0.107)	0.135 (0.053)	0.130 (0.045)
320	0.50	Gaussian	0.998 (0.108)	0.137 (0.053)	0.133 (0.046)
320	0.80	Constant	1.039 (0.273)	0.135 (0.081)	0.123 (0.048)
320	0.80	Gaussian	1.039 (0.271)	0.133 (0.081)	0.122 (0.049)
320	1.00	Gaussian	–	–	0.122 (0.049)
1280	0.25	Constant	1.000 (0.040)	0.102 (0.039)	0.102 (0.037)
1280	0.25	Gaussian	1.001 (0.040)	0.102 (0.039)	0.102 (0.038)
1280	0.50	Constant	0.995 (0.050)	0.099 (0.038)	0.097 (0.038)
1280	0.50	Gaussian	0.995 (0.050)	0.099 (0.038)	0.097 (0.037)
1280	0.80	Constant	1.000 (0.133)	0.104 (0.052)	0.099 (0.038)
1280	0.80	Gaussian	1.000 (0.131)	0.104 (0.052)	0.100 (0.036)
1280	1.00	Gaussian	–	–	0.102 (0.037)

S.24.2 GAUSSIAN DESIGNS WITH A SMALL EXCESS NUMBER OF FEATURES

We examine the extension in Theorem 8 when n/p approaches one. Let $d = p - n$. We use $n = 80, 320, 1280$ and compare $d = \lceil \sqrt{n} \rceil$ with $d = 8$. The entries of X are independent $N(0, 1/p)$ variables, and $y = X\theta + \varepsilon$ with independent $\varepsilon \sim N(0, I_n)$. The deterministic signal variance $\|\theta\|_2^2 / p$ equals 0, 1, or $(p/d)^2$. In the third case, the signal and noise contributions to $v_{n,p}^2$ are equal. The design has no constant feature. Each of the 18 settings uses 120 independent repetitions, and all generated designs are kept.

We sample the Gram matrix and signal jointly by taking $LL^\top$ to have the Wishart distribution with p degrees of freedom and identity scale. If h contains the first n coordinates of an independent uniform unit vector in $\mathbb{R}^p$ and $t^2 = \|\theta\|_2^2 / p$, then $K = LL^\top / p$ and $y = tLh + \varepsilon$ have the required joint distribution. This follows from the right orthogonal invariance used in Section S.24.1. Triangular solves give $a = K^{-1}y$. The three signal sizes share the same Gram matrix, spherical direction, and noise within each repetition. We use D_r and D_v defined above and let D_s be the same uniform error with $s_{n,p}$ replacing $v_{n,p}$. For $d = 8$, $v_{n,p}$ is a comparison scale; Theorem 8 does not assert concentration around it.

Tables S.8 and S.9 report all settings. The mean D_v is below the mean D_s in each of the 18 settings. For $d = \lceil \sqrt{n} \rceil$ and $n = 1280$, the mean ratios $r_{n,p}^2 / v_{n,p}^2$ are 1.079, 1.079, and 1.052 for the three signal variances. The corresponding mean errors D_v are 0.177, 0.176, and 0.138, compared with D_s values 0.497, 0.496, and 0.435. These comparisons show the gain from accounting separately for the inverse Gram factors in the signal and noise terms.

For zero signal and fixed $d = 8$, the standard deviation of $r_{n,p}^2 / v_{n,p}^2$ is 2.805 at $n = 1280$, while its mean D_r is 0.102. Figure S.4 separates this variation in length from the convergence of the normalized shape. The fixed-gap length comparison concerns Proposition 5 under zero signal; the other signal settings evaluate the shape assertion in Theorem 8. The script `scripts/run_gaussian_near_square.py` uses seed 202609051 and produces 2160 records. The recorded summaries include standard deviations, Monte Carlo standard errors, and length-ratio quantiles.

Table S.8: Gaussian boundary normalization with $p - n = \lceil \sqrt{n} \rceil$. Numbers in parentheses represent standard deviations.

n	$\|\theta\|_2^2/p$	$r_{n,p}^2/v_{n,p}^2$	D_v	D_s	D_r
80	0	1.660 (2.318)	0.354 (0.347)	0.478 (0.012)	0.170 (0.061)
80	1	1.652 (2.337)	0.345 (0.310)	0.470 (0.016)	0.171 (0.064)
80	$(p/d)^2$	1.362 (1.550)	0.268 (0.203)	0.363 (0.052)	0.175 (0.072)
320	0	1.089 (0.746)	0.245 (0.202)	0.492 (0.003)	0.131 (0.046)
320	1	1.085 (0.747)	0.244 (0.200)	0.489 (0.004)	0.129 (0.047)
320	$(p/d)^2$	1.035 (0.589)	0.194 (0.136)	0.403 (0.026)	0.125 (0.049)
1280	0	1.079 (0.616)	0.177 (0.130)	0.497 (0.001)	0.103 (0.039)
1280	1	1.079 (0.613)	0.176 (0.128)	0.496 (0.001)	0.102 (0.039)
1280	$(p/d)^2$	1.052 (0.404)	0.138 (0.093)	0.435 (0.013)	0.097 (0.039)

Table S.9: Gaussian boundary normalization with $p - n = 8$. Numbers in parentheses represent standard deviations.

n	$\|\theta\|_2^2/p$	$r_{n,p}^2/v_{n,p}^2$	D_v	D_s	D_r
80	0	1.671 (2.873)	0.426 (0.435)	0.480 (0.013)	0.175 (0.065)
80	1	1.654 (2.827)	0.440 (0.545)	0.471 (0.022)	0.176 (0.062)
80	$(p/d)^2$	1.340 (1.454)	0.286 (0.196)	0.369 (0.052)	0.174 (0.065)
320	0	2.335 (9.449)	0.342 (0.268)	0.497 (0.001)	0.131 (0.051)
320	1	2.331 (9.374)	0.340 (0.263)	0.496 (0.002)	0.130 (0.052)
320	$(p/d)^2$	1.707 (3.729)	0.223 (0.146)	0.436 (0.022)	0.123 (0.051)
1280	0	1.861 (2.805)	0.329 (0.329)	0.500 (0.000)	0.102 (0.040)
1280	1	1.861 (2.807)	0.329 (0.330)	0.500 (0.000)	0.102 (0.039)
1280	$(p/d)^2$	1.494 (1.769)	0.250 (0.180)	0.466 (0.013)	0.101 (0.036)

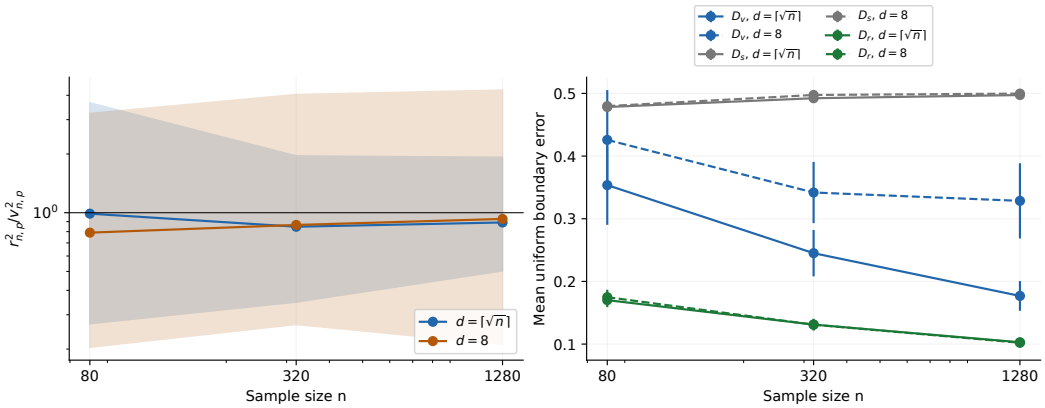

Figure S.4: Gaussian boundaries with zero signal and a small excess dimension. Left: medians and 10th–90th percentile bands of the squared length ratio; right: mean boundary errors with two Monte Carlo standard errors.

S.24.3 A CONSTANT FEATURE IN NEARLY SQUARE DESIGNS

Corollary 6 permits the number of features to approach the sample size when their difference tends to infinity. We examine whether its separate signal and noise contributions describe the boundary more accurately than the marginal response variance.

We use the constant-feature model in the corollary with $n = 80, 320, 1280$, $d = p - n = \lceil\sqrt{n}\rceil$, $\sigma = 1$, and $\|\theta_{-1}\|_2^2/p$ equal to 0, 1, or $(p/d)^2$. The last setting makes the signal and noise contributions to $v_{n,p}^2$ equal. The true conditional-mean coefficient of the penalized constant is zero. Each of the nine settings uses 120 independent repetitions. We generate the joint Gaussian Gram matrix and signal with $p - 1$ Bartlett degrees of freedom and an independent spherical direction, then add the constant column by the rank-one inverse identity. All repetitions are kept. The script `run_gaussian_constant_near_square.py` uses seed 202609071.

Table S.10 uses the errors D_v, D_s, D_r defined in Section S.24.2. The corrected scale has smaller mean uniform error than the marginal response scale in all nine settings. For unit signal variance, the mean D_v decreases from 0.336 to 0.169, while the mean D_s increases from 0.470 to 0.496. The mean squared-length ratio decreases from 2.057 to 1.189. These results support the distinct signal and noise factors in the constant-feature boundary scale.

The proof compares $a = (XX^\top)^{-1}y$ with $a_B = (ZZ^\top)^{-1}y$. For unit signal variance, the means of $\|a - a_B\|_2/(\sqrt{n}\, v_{n,p})$ and $\|a - a_B\|_\infty/\{v_{n,p}\sqrt{2\log n}\}$ decrease from 0.330 to 0.153 and from 0.296 to 0.140, respectively. This comparison measures the perturbation controlled by the corollary. Recorded means, standard deviations, Monte Carlo standard errors, and linear-system residuals accompany the numerical results.

Table S.10: Gaussian boundary normalization with a penalized constant and $p - n = \lceil\sqrt{n}\rceil$. Numbers in parentheses represent standard deviations.

n	$\|\theta_{-1}\|_2^2/p$	$r_{n,p}^2/v_{n,p}^2$	D_v	D_s	D_r
80	0	2.079 (4.930)	0.330 (0.278)	0.479 (0.010)	0.156 (0.071)
80	1	2.057 (4.828)	0.336 (0.277)	0.470 (0.015)	0.159 (0.067)
80	$(p/d)^2$	1.551 (2.316)	0.275 (0.204)	0.360 (0.054)	0.170 (0.063)
320	0	1.394 (1.449)	0.215 (0.152)	0.493 (0.002)	0.125 (0.047)
320	1	1.394 (1.431)	0.215 (0.153)	0.490 (0.003)	0.126 (0.048)
320	$(p/d)^2$	1.263 (0.851)	0.186 (0.126)	0.409 (0.027)	0.130 (0.051)
1280	0	1.188 (0.731)	0.169 (0.120)	0.497 (0.001)	0.104 (0.038)
1280	1	1.189 (0.732)	0.169 (0.120)	0.496 (0.001)	0.104 (0.038)
1280	$(p/d)^2$	1.119 (0.472)	0.134 (0.099)	0.435 (0.014)	0.098 (0.036)

S.25 PREDICTIVE COMPARISON UNDER A GAUSSIAN QUANTILE MODEL

This section evaluates the finite-sample behavior of boundary screening in a model with known conditional quantiles. The deterministic boundary checks appear in Section S.26. For each split, the first coordinate of every covariate vector equals $1/\sqrt{p}$, and the remaining coordinates are independent $N(0, 1/p)$ variables. We set $\theta_1 = 0$, rescale θ_{-1} to have Euclidean norm $0.75\sqrt{p}$, and generate $Y = x^\top\theta + \varepsilon$ with $\varepsilon \sim N(0, 1)$. The conditional quantile at level τ is $x^\top\theta + \Phi^{-1}(\tau)$, which belongs to the fitted linear class through the first coordinate. Training draws are kept when X has full row rank and a has both positive and negative coordinates.

The study uses 120 independent splits with $(n, p) = (100, 120)$, 1000 validation observations, 4000 test observations, and levels 0.10, 0.25, 0.50, 0.75, and 0.90. Each level has 28 candidate penalties, from $0.25\bar{\lambda}(\tau)$ to $1024\bar{\lambda}(\tau)$. The weak-penalty baseline uses $0.50\bar{\lambda}(\tau)$. Unfiltered validation minimizes the validation pinball loss over all 28 candidates, and boundary-screened validation minimizes the same loss over the 23 candidates with factor above one. We report test pinball risk, mean squared error from the known conditional quantile, absolute calibration error, and the mean test-sample range over the five fitted levels.

Table S.11: Predictive comparison in the Gaussian quantile model. Numbers in parentheses represent standard deviations across 120 independent splits.

Method	Test pinball	Quantile MSE	Calibration error	Level range
Weak penalty	0.989 (0.136)	6.100 (1.746)	0.265 (0.004)	4.4×10^{-15} (1.5×10^{-15})
Unfiltered	0.436 (0.009)	1.017 (0.048)	0.206 (0.007)	0.807 (0.087)
Screened	0.436 (0.009)	1.017 (0.048)	0.206 (0.007)	0.807 (0.087)

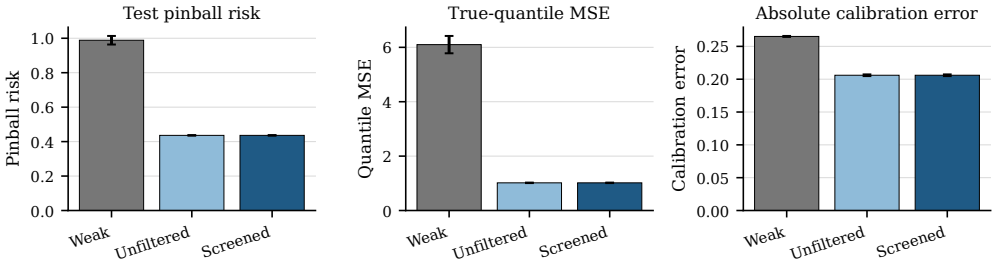

Figure S.5: Cost of collapse in the Gaussian predictive study over 120 independent samples: a penalty at half the realized boundary, validation over all 28 candidates, and validation over the 23 screened candidates. Bars are means, error bars are two Monte Carlo standard errors, and lower values are better.

Table S.11 shows that the weak penalty gives mean level range 4.422×10^{-15}, while boundary-screened validation gives mean level range 0.807. The screened selector reduces mean test pinball risk by 55.9%, true-quantile mean squared error by 83.3%, and absolute calibration error by 22.3% relative to the weak penalty. Each split has a positive improvement in all three measures after averaging over the five levels. Unfiltered and screened validation choose the same candidate in all 600 split-level comparisons. Screening excludes 3000 loss-blind candidates from the 16800 candidate rows, and none of the 600 screened selections occurs at the upper endpoint of the expanded grid. The script `scripts/run_predictive_comparison.py` and its validator generate and check these results with seed 20260901.

The weak-penalty baseline is half the realized boundary of each level and is extinct by construction. It measures the cost of collapse bounded in Corollary 4, not the performance of a penalty chosen in practice. Screened and unfiltered validation select the same model, and the selected penalties are between 24 and 512 times the realized boundary, with median 128. Because this design penalizes the constant feature $1/\sqrt{p}$, the validated fits remain under-dispersed: their mean coverage is 0.409 at level 0.1 and 0.590 at level 0.9. In 120 independent samples from the same model, the intercept-only training quantile has mean test pinball risk 0.350, true-quantile mean squared error

0.647, and absolute calibration error 0.031, and it is more accurate than the validated fits of this all-penalized class. For this reason, the studies in Section 5 use an unpenalized intercept. The script `scripts/check_reference_baselines.py` computes the reference values.

S.26 Additional boundary diagnostics

We compare the boundary rule with validation over an unfiltered penalty grid. We also compare a common weak penalty below all requested-level boundaries with a common penalty above them. The first simulation draws X_{ij} independently from $N(0, 1/p)$ with $(n, p) = (30, 90)$, rescales $\theta \sim N(0, I_p)$ to $\|\theta\|_2 = 1$, sets $y = X\theta + 0.5\varepsilon$ with $\varepsilon \sim N(0, I_n)$, and keeps the draws for which $a = (XX^\top)^{-1}y$ has both positive and negative coordinates. Before this sign restriction, conditional on θ, the design and response follow model (2) with $\sigma = 0.5$. The accepted draws assess the fixed-design identities, not the unconditional Gaussian limit. Across 120 independent repetitions and five quantile levels, Table S.12 reports machine-precision equality with the minimum-norm interpolant for the weak-penalty baseline and nonzero coefficient separation for every candidate kept after boundary screening. The local-departure rows use the same design family and 120 accepted unique-binding cases; they are directional checks of Proposition 2.

Supplementary Table S.1 reports a separate certified active-set study. The certificate uses Corollary 14: for each accepted unique-binding case, the analytic one-coordinate candidate is checked at perturbations strictly inside the computed active-set radius. The study has 600 accepted cases and 2400 inside-radius rows. The maximum candidate Karush–Kuhn–Tucker (KKT) stationarity residual is 2.220×10^{-16}, the maximum candidate subgradient violation is 4.441×10^{-16}, and the maximum residual-identity error is 2.970×10^{-15}. Outside-radius rows and solver comparisons are auxiliary checks and do not define the certificate.

The real-data comparisons use the diabetes data distributed with scikit-learn (Efron et al., 2004; Pedregosa et al., 2011), 640 random Fourier features (Rahimi & Recht, 2007), and train-only preprocessing. The first comparison uses 120 random train–test splits of the same cohort and shows that a weak penalty gives zero prediction spread across levels, while the fit above every boundary has positive spread in each split. A separate 120-repetition train–validation–test comparison selects from 28 multiples of the realized level-specific boundary. Relative to the weak penalty, screened validation reduces mean test pinball risk by 17.4% and absolute calibration error by 22.1%; both paired improvements are positive in all 120 splits. Screening removes 3000 loss-blind candidates, agrees with unfiltered validation in 599 of 600 selections, and reaches no upper grid endpoint. Supplementary Section S.28 gives the design, Table S.17 reports the numerical values, and Figure S.9 summarizes the comparison. Variation across these splits measures sensitivity to data division and random-feature generation within this cohort. The scripts and `results/summary/numerical_manifest.json` record the settings, row counts, and artifact hashes.

The empirical calibration metric averages absolute errors across levels. Corollary 4 bounds this mean under collapse. To evaluate the population criteria, we use a separate Gaussian model whose conditional quantiles are known. The feature construction includes a penalized constant, as in Corollary 6. Across 120 independently generated training–validation–test samples, Figure S.5 compares the weak-penalty baseline, validation over the unfiltered grid, and validation after boundary screening. Relative to the weak penalty, boundary-screened validation reduces mean test pinball risk by 55.9%, true-quantile mean squared error by 83.3%, and absolute calibration error by 22.3%. All three splitwise improvements are positive after averaging over the five fitted levels. Screening removes 3000 of 16800 candidates, and its 600 split-level selections equal those of unfiltered validation. Thus the rule removes the loss-blind part of the grid without changing the selected predictor in this experiment. Supplementary Section S.25 gives the model, the full penalty grid, and the numerical standard deviations.

Figure S.1 reports the certified active-set study: the left panel plots the analytic KKT residuals against the fraction of the certified radius, and the right panel records the certified radii for the 600 accepted cases. Figure S.6 shows one realized triangular boundary, with the shaded interval matching the set of levels satisfying $\lambda \le \bar{\lambda}(\tau)$. Its maximizer is $\tau^\dagger \approx 0.58$ for this draw. Supplementary Section S.24 reports 120-repetition scaling in four cells with $p = n^2$: the mean uniform error decreases from 0.172 to 0.095, and the mean first-extinct-level error decreases from 0.058 to 0.025.

Supplementary Section S.27 reports a denser finite-grid audit of the same deterministic boundary. The audit uses 120 fixed-feature designs, 199 quantile levels, and six penalty factors scaled by the maximum boundary value, giving 143280 level rows. Outside the predeclared boundary-tie region, the theorem-predicted extinction indicator matches the numerical fit indicator in every checked row. This is a deterministic finite-grid agreement check, not a tuning rule or a population validation study.

Supplementary Section S.28 reports aspect-ratio stress cells and repeated diabetes/random-feature splits. Tables S.15–S.16 and Figure S.8 show the same boundary separation in all four feature dimensions and all 120 real-data splits. Table S.17 and Figure S.9 give the independent train–validation–test comparison of predictive loss, calibration, and level separation.

For the five levels in the predictive studies, (4) bounds the mean absolute calibration error of a common predictor below by 0.26. Tables S.11 and S.17 report weak-penalty means of 0.265 and 0.268, respectively. These values are close to the lower bound imposed by collapse across levels.

Keeping one extinct representative and removing its duplicates removes 2400 duplicate candidates in each predictive study and preserves all 1200 original selections; Section S.20.2 gives the comparison. The Gaussian studies in Sections S.24.1 and S.24.2 verify the boundary scale across dimension ratios and signal sizes. Section S.24.3 verifies the constant-feature extension in nearly square designs; the corrected scale has a smaller mean boundary error than the marginal response scale in all nine settings.

Table S.12: Boundary screening and the weak-penalty baseline. The extinction rows summarize 120 independent designs, five quantile levels, and two sides of the boundary. The local-departure rows summarize 120 accepted unique-binding cases. The real-data rows use one fixed diabetes split and one random-feature seed.

Check	Statistic	Value
Weak-penalty baseline	max $\|\widehat{\beta} - \widehat{\beta}_0\|_2$	3.022e-15
Weak-penalty baseline	max $\|y - X\widehat{\beta}\|_2$	4.221e-15
Kept after boundary screening	min $\|\widehat{\beta} - \widehat{\beta}_0\|_2$	2.244e-01
Local departure	max relative formula error	7.256e-02
Local departure	min cosine with formula direction	0.997378
Weak-penalty real-data baseline	test spread	0.000e+00
Real-data fit after boundary screening	test spread	9.035e-01

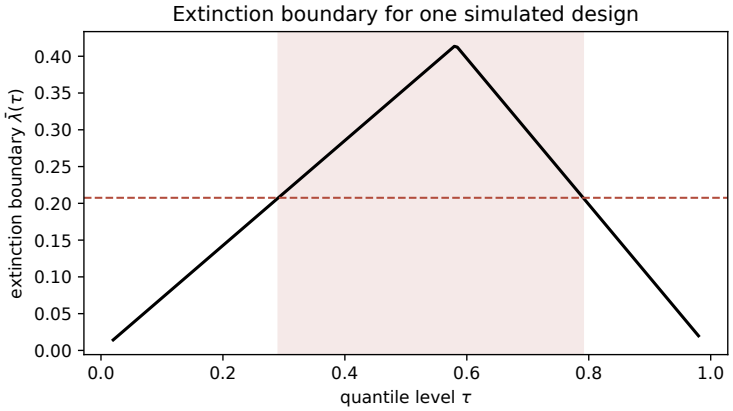

Figure S.6: A simulated extinction boundary. The shaded interval is the set of levels for which a fixed penalty lies below the boundary.

S.27 DENSE THRESHOLD-GRID NUMERICAL AUDIT

The dense threshold-grid artifact is a finite-grid check of Theorem 2 and Proposition 3. It uses 120 independent fixed-feature designs with $(n, p) = (30, 90)$ drawn as in Section S.26, 199 levels in $[0.005, 0.995]$, and penalty factors 0.25, 0.50, 0.90, 1.00, 1.10, and 1.50 multiplied by the maximum boundary value. The script is `scripts/run_dense_threshold_grid.py`; the artifact manifest records settings, row counts, script hashes, and output hashes.

Table S.13: Dense threshold-grid deterministic audit. The table summarizes 143280 rows from 120 fixed-feature designs, 199 quantile levels, and six penalty factors scaled by the maximum boundary value. Mismatches are counted only outside the predeclared near-boundary rows.

Penalty factor	extinct rows	checked rows	mismatches	min nonextinct distance
0.25	18000	23880	0	3.848e-05
0.50	12000	23880	0	3.146e-05
0.90	2400	23880	0	4.078e-05
1.00	0	23880	0	1.112e-05
1.10	0	23880	0	1.183e-01
1.50	0	23880	0	4.482e-01

Table S.14: Dense threshold-grid numerical tolerances for the finite-grid audit. These diagnostics check the optimization rows used to compare the deterministic boundary with numerical fits.

Diagnostic	Statistic	Value
Extinct rows	max $\|\widehat{\beta} - \widehat{\beta}_0\|_2$	1.517e-15
Extinct rows	max $\|y - X\widehat{\beta}\|_2$	3.568e-15
Solver diagnostics	max stationarity	1.110e-16
Solver diagnostics	max subgradient violation	0.000e+00
Solver diagnostics	max projected-gradient residual	1.263e-06

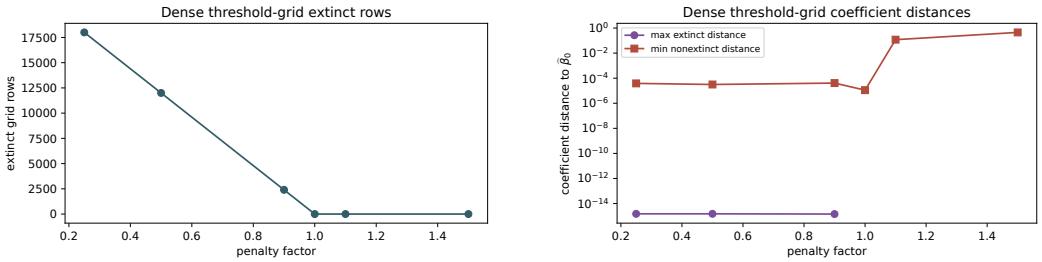

Figure S.7: Dense threshold-grid diagnostics. Left: number of extinct grid rows by penalty factor. Right: coefficient distances to the interpolant for extinct and nonextinct rows.

Table S.13 reports 143280 level rows and no mismatch outside the predeclared boundary-tie region. Table S.14 shows that the maximum coefficient distance on theorem-extinct rows is 1.517×10^{-15} and the maximum residual norm on those rows is 3.568×10^{-15}. The minimum coefficient distance on nonextinct rows is 1.112×10^{-5}. The maximum stationarity residual in the KKT ledger is 1.110×10^{-16}, the maximum projected-gradient residual is 1.263×10^{-6}, and the maximum subgradient residual under the numerical sign tolerance is zero. These rows audit the deterministic finite-grid boundary only.

S.28 ASPECT-RATIO AND REPEATED REAL-DATA COMPARISONS

This section checks the deterministic extinction identity across several fixed-feature dimensions and evaluates level separation, predictive loss, and calibration over repeated diabetes/random-feature

splits. The aspect-ratio cells use $(n, p) = (25, 50), (30, 90), (40, 160), (60, 300)$, 120 independent designs per cell, five levels, and four penalty factors scaled by $\bar{\lambda}(\tau)$. In every cell, X_{ij} is drawn independently from $N(0, 1/p)$, θ is an independent Gaussian vector rescaled to $\|\theta\|_2 = 1$, and $y = X\theta + 0.5\varepsilon$ with $\varepsilon \sim N(0, I_n)$. As in the primary simulation, we keep full-row-rank designs for which a has both positive and negative coordinates. These comparisons check the fixed-design identities.

The deterministic real-data comparison uses 120 random splits of the diabetes data, 160 training observations, 100 test observations, and 640 random Fourier features per split. Covariates and responses are standardized using only the training observations. For a standardized covariate $z \in \mathbb{R}^{10}$, the feature map is $\psi(z) = \sqrt{2/640}\cos(W^\top z + b)$, where the entries of $W \in \mathbb{R}^{10\times 640}$ are independent $N(0, 1)$ variables and the entries of $b \in \mathbb{R}^{640}$ are independent uniform variables on $[0, 2\pi]$. The same realized W and b map the test observations. The lower penalty is $0.75 \min_\tau \bar{\lambda}(\tau)$ and the upper penalty is $1.25 \max_\tau \bar{\lambda}(\tau)$ over the five fitted levels. The test spread is the maximum absolute difference between any two level predictions over all test observations.

The predictive comparison uses an independent seed and 120 random splits with 160 training observations, 100 validation observations, and 100 test observations. For each split and level, the penalty factors are 0.25, 0.50, 0.75, 0.90, 1.00, 1.05, 1.10, 1.25, 1.50, 2, 3, 4, 6, 8, 12, 16, 24, 32, 48, 64, 96, 128, 192, 256, 384, 512, 768, and 1024 times $\bar{\lambda}(\tau)$. The weak-penalty baseline uses factor 0.50. Unfiltered validation minimizes validation pinball loss over all 28 factors, while screened validation minimizes it over the factors greater than one. Ties are resolved in favor of the larger factor. Test pinball loss and empirical marginal calibration error are first computed at each level and then averaged over the five levels within a split. The scripts `scripts/run_aspect_realdata_stress.py` and `scripts/run_repeated_realdata_utility.py` generate the two real-data comparisons; the artifact manifest records their settings, row counts, script hashes, and output hashes.

Table S.15: Aspect-ratio deterministic stress audit. Each cell uses 120 independent fixed-feature designs, five levels, and four penalty factors scaled by the level-specific boundary.

Aspect cell	rows	extinct rows	max extinct distance	min nonextinct distance
25 : 50	2400	1200	4.300e-15	2.126e-01
30 : 90	2400	1200	2.998e-15	2.553e-01
40 : 160	2400	1200	2.813e-15	2.607e-01
60 : 300	2400	1200	3.605e-15	3.476e-01

Table S.16: Repeated fixed-feature real-data deterministic audit. The rows use train-only preprocessing and independent random-feature maps for each split. They audit level collapse below the grid thresholds and do not compare predictive performance.

Repeated diabetes/random-feature audit	Statistic	Value
Splits	total	120
Training observations	per split	160
Test observations	per split	100
Random features	per split	640
Below all grid thresholds	max test prediction spread	0.000e+00
Above all grid thresholds	median test prediction spread	1.080e+00
Above all grid thresholds	min test prediction spread	5.244e-01

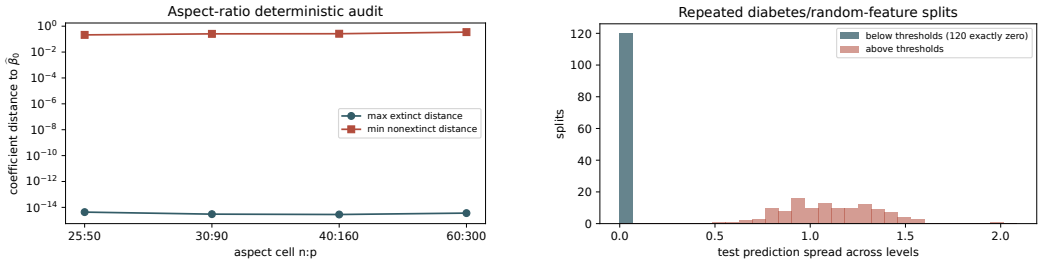

Figure S.8: Aspect-ratio and repeated real-data comparisons. Left: coefficient distances to $\widehat{\beta}_0$ in the aspect-ratio stress cells. Right: test prediction spread across quantile levels in repeated diabetes/random-feature splits.

Table S.15 shows exact agreement between the theorem-predicted extinction indicator and the numerical fits across all 9600 aspect-ratio rows. The maximum coefficient distance on extinct rows is 4.300×10^{-15}, the maximum training residual norm on extinct rows is 5.018×10^{-15}, and the minimum coefficient distance on nonextinct rows is 2.126×10^{-1}. Table S.16 reports that the maximum test prediction spread below all grid thresholds is zero across the 120 splits, while the median and minimum spreads above all grid thresholds are 1.080 and 5.244×10^{-1}. These comparisons show the advantage of using the realized boundary: it separates loss-blind fits at machine precision from fits whose coefficients differ materially, and it gives positive level spread in every repeated real-data split. They do not compare predictive risk.

Table S.17: Repeated diabetes predictive comparison. Numbers in parentheses represent standard deviations across 120 random train–validation–test splits of the same cohort.

Method	Test pinball	Calibration error	Level range
Weak penalty	0.412 (0.033)	0.268 (0.007)	6.5×10^{-16} (7.7×10^{-17})
Unfiltered	0.341 (0.031)	0.209 (0.018)	0.632 (0.071)
Screened	0.341 (0.031)	0.209 (0.018)	0.632 (0.071)

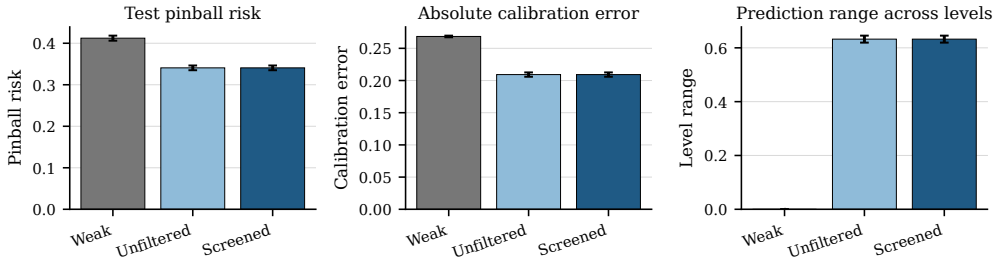

Figure S.9: Repeated diabetes predictive comparison across 120 splits. Bars are means, error bars are two Monte Carlo standard errors, and lower values are better in the first two panels.

Table S.17 shows that screened validation reduces mean test pinball risk from 0.412 to 0.341, a reduction of 17.4%, and reduces mean absolute calibration error from 0.268 to 0.209, a reduction of 22.1%. Both paired improvements are positive in all 120 splits. The mean prediction range across levels increases from 6.517×10^{-16} to 0.632. Screening excludes 3000 of 16800 candidates and agrees with unfiltered validation in 599 of 600 split-level selections. The two validation rules have equal mean calibration error, and their mean test pinball risks differ by 1.273×10^{-7}. No screened selection occurs at the upper grid endpoint. These results show that removing the loss-blind candidates preserves the predictive performance of the full grid and improves on a penalty at half the boundary in these repeated splits. On the same 120 splits, the intercept-only training quantile has mean test pinball risk 0.279 and absolute calibration error 0.040, and unpenalized linear

quantile regression on the ten covariates has 0.206 and 0.042. The random-feature class with every coefficient penalized is less accurate than both references on this cohort; the study illustrates the cost of collapse and does not propose a competitive predictor.

S.29 Designs and complete results of the numerical studies in Section 5

Both studies fit ridge-penalized quantile regression with an unpenalized intercept by solving its dual problem. For $\lambda > 0$ and $\tau \in (0,1)$, the dual vector g minimizes $g^\top Kg/2 - \lambda y^\top g$ over $g \in [\tau-1, \tau]^n$ with $\mathbf{1}_n^\top g = 0$. The optimality conditions of Section S.30 give the coefficients $\widehat{\beta} = X^\top g/\lambda$, and the intercept is the common value of $y_i - x_i^\top \widehat{\beta}$ over the coordinates with $\tau - 1 < g_i < \tau$, computed as their median, or the τ-quantile of these values over all coordinates when no such coordinate exists. Extinct fits use the closed form $(\alpha_Q, X^\top a_c)$ of Corollary 27. Other fits use accelerated projected gradient steps with gradient-based restarts; the projection onto the box intersected with the hyperplane is computed exactly from the sorted breakpoints of its piecewise linear equation. For $n \le 600$, a final active-set step solves the optimality equations on the free coordinates and checks the signs on the bound coordinates. Every fit is certified by its relative primal–dual gap $(P - D)/(1 + |P|)$, where P is the primal objective at the fitted pair and $D = y^\top g - g^\top Kg/(2\lambda)$ is the dual objective. The solver is `scripts/intercept_qr.py`.

Coverage study. The coverage study simulates model (3) with $\mu_0 = 1$, $\sigma = 1$, $\|\theta\|_2^2/p = 1$, and $p = n^2$ for $n \in \{100, 300, 1000\}$, with 120 independent repetitions for each n and production seed 202609231; a three-repetition pilot with seed 202609230 checked only the solver accuracy and running time. The pair $(K, X\theta)$ is generated exactly without forming X. With L lower triangular, L_{ii}^2 independent χ^2_{p-i+1} variables, and independent standard normal entries below the diagonal, $LL^\top$ has the Wishart distribution with p degrees of freedom and identity scale, and we put $K = LL^\top/p$. Let h contain the first n coordinates of an independent uniform unit vector in $\mathbb{R}^p$. Right orthogonal invariance of the Gaussian design gives $X\theta = (\|\theta\|_2/\sqrt{p})Lh$ jointly with K, as in Section S.24.1. For a fitted pair $(\widehat{\alpha}, \widehat{\beta})$ with $\widehat{\beta} = X^\top g/\lambda$, the test residual $Y_* - \widehat{\alpha} - X_*^\top \widehat{\beta}$ is Gaussian with mean $\mu_0 - \widehat{\alpha}$ and variance $\omega^2 = \sigma^2 + \left\|\theta - \widehat{\beta}\right\|_2^2/p$, and

$$\frac{\left\|\theta - \widehat{\beta}\right\|_2^2}{p} = \frac{\|\theta\|_2^2}{p} - \frac{2(X\theta)^\top g}{\lambda p} + \frac{g^\top Kg}{\lambda^2 p}.$$

The coverage $\Phi\{(\widehat{\alpha} - \mu_0)/\omega\}$ and the pinball risk $m\{\tau - \Phi(-m/\omega)\} + \omega\phi(m/\omega)$ with $m = \mu_0 - \widehat{\alpha}$ are thus exact given the training sample, where ϕ is the standard normal density. The normalized penalties are $c \in \{0.02, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.75, 1, 1.5, 2, 3, 5, 10, 20, 50, 100\}$, and $\lambda = c/\{s_{n,p}\sqrt{2\log n}\}$ with $s_{n,p} = \sqrt{2}$. The driver is `scripts/run_intercept_calibration.py`, and `scripts/validate_intercept_calibration.py` checks every extinction flag against the recorded boundary, the closed-form intercept, the equality of coverage across extinct levels, the duality gaps, and the summaries.

Table S.18 reports the mean coverage and the proportion of extinct repetitions at eight of the nineteen penalties, and Table S.19 reports the mean and standard deviation of the normalized boundary $s_{n,p}\sqrt{2\log n}\,\bar{\lambda}_c(\tau)$. All 34200 fits have relative duality gap at most 3.2×10^{-11}. In the 9858 extinct rows, the intercept equals α_Q, and all extinct levels of the same sample and penalty have identical coverage. The largest Monte Carlo standard error of a mean coverage is 0.0045. At $c = 0.02$, every level is extinct in every repetition, and the standard deviation of the common coverage over repetitions is 0.042, 0.024, and 0.013 for $n = 100$, 300, and 1000.

Learned and random features. The learned and random feature study uses a simulated heteroskedastic model and four public regression data sets. In the simulated model, z is uniform on $[-1, 1]^{10}$ and $y = \sin(\pi z_1 z_2) + 2(z_3 - 1/2)^2 + z_4 + (0.3 + 0.4|z_5|)\varepsilon$ with $\varepsilon \sim N(0,1)$ independent of z. California housing (20640 observations and 8 covariates) and diabetes (442 observations and 10 covariates) are distributed with scikit-learn, and puma8NH (8192 observations and 8 covariates)

Table S.18: Mean coverage in the coverage study, with the proportion of extinct repetitions in parentheses, over 120 repetitions for each n.

n	c	$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$
100	0.05	0.500 (1.00)	0.500 (1.00)	0.500 (1.00)	0.500 (1.00)	0.500 (1.00)
100	0.1	0.500 (0.93)	0.500 (1.00)	0.500 (1.00)	0.500 (1.00)	0.501 (0.87)
100	0.2	0.489 (0.00)	0.500 (1.00)	0.500 (1.00)	0.500 (0.99)	0.513 (0.00)
100	0.3	0.462 (0.00)	0.500 (0.55)	0.500 (1.00)	0.501 (0.48)	0.539 (0.00)
100	0.5	0.406 (0.00)	0.489 (0.00)	0.501 (0.81)	0.513 (0.00)	0.595 (0.00)
100	1	0.308 (0.00)	0.435 (0.00)	0.501 (0.00)	0.567 (0.00)	0.692 (0.00)
100	5	0.152 (0.00)	0.300 (0.00)	0.501 (0.00)	0.701 (0.00)	0.851 (0.00)
100	100	0.107 (0.00)	0.253 (0.00)	0.500 (0.00)	0.744 (0.00)	0.891 (0.00)
300	0.05	0.503 (1.00)	0.503 (1.00)	0.503 (1.00)	0.503 (1.00)	0.503 (1.00)
300	0.1	0.503 (0.92)	0.503 (1.00)	0.503 (1.00)	0.503 (1.00)	0.503 (0.91)
300	0.2	0.495 (0.00)	0.503 (1.00)	0.503 (1.00)	0.503 (0.99)	0.511 (0.00)
300	0.3	0.472 (0.00)	0.503 (0.45)	0.503 (1.00)	0.503 (0.47)	0.534 (0.00)
300	0.5	0.420 (0.00)	0.495 (0.00)	0.503 (0.83)	0.511 (0.00)	0.586 (0.00)
300	1	0.322 (0.00)	0.446 (0.00)	0.503 (0.00)	0.560 (0.00)	0.683 (0.00)
300	5	0.151 (0.00)	0.303 (0.00)	0.503 (0.00)	0.701 (0.00)	0.853 (0.00)
300	100	0.104 (0.00)	0.255 (0.00)	0.502 (0.00)	0.750 (0.00)	0.900 (0.00)
1000	0.05	0.502 (1.00)	0.502 (1.00)	0.502 (1.00)	0.502 (1.00)	0.502 (1.00)
1000	0.1	0.502 (0.88)	0.502 (1.00)	0.502 (1.00)	0.502 (1.00)	0.502 (0.90)
1000	0.2	0.497 (0.00)	0.502 (1.00)	0.502 (1.00)	0.502 (0.99)	0.507 (0.00)
1000	0.3	0.479 (0.00)	0.502 (0.33)	0.502 (1.00)	0.503 (0.40)	0.526 (0.00)
1000	0.5	0.431 (0.00)	0.497 (0.00)	0.502 (0.78)	0.507 (0.00)	0.573 (0.00)
1000	1	0.337 (0.00)	0.455 (0.00)	0.502 (0.00)	0.549 (0.00)	0.668 (0.00)
1000	5	0.158 (0.00)	0.309 (0.00)	0.501 (0.00)	0.695 (0.00)	0.846 (0.00)
1000	100	0.104 (0.00)	0.254 (0.00)	0.501 (0.00)	0.749 (0.00)	0.899 (0.00)

Table S.19: Mean normalized extinction boundary in the coverage study. Numbers in parentheses represent standard deviations over 120 repetitions.

n	$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$
100	0.125 (0.018)	0.312 (0.044)	0.563 (0.072)	0.306 (0.052)	0.122 (0.021)
300	0.120 (0.014)	0.299 (0.034)	0.553 (0.058)	0.298 (0.037)	0.119 (0.015)
1000	0.114 (0.012)	0.285 (0.029)	0.541 (0.050)	0.291 (0.030)	0.116 (0.012)

and house_16H (22784 observations and 16 covariates) are OpenML data sets 225 and 574. For house_16H, we remove the 52 observations with zero price and model the logarithm of the price. In each of 100 repetitions per data set, we draw disjoint training, validation, and test samples of sizes 128, 128, and 1000, or 186 test observations for diabetes, and we standardize the covariates and the response with the training means and standard deviations. Pooled Monte Carlo standard errors combine the within-data-set variances with the fixed pooling weights.

The network has two hidden ReLU layers of width 256 and five linear heads for the levels $0.1, 0.25, 0.5, 0.75, 0.9$. From the default PyTorch initialization, it is trained by 2000 full-batch Adam steps with learning rate 10^{-3} on the sum over levels of the averaged pinball losses plus $(\gamma/2)$ times the squared norm of all trunk parameters and head weights, with $\gamma = 10^{-4}$ and unpenalized head intercepts. Multiplying this objective by n gives (1) with $\lambda = n\gamma$ and $\Omega(\vartheta) = n\gamma \|\vartheta\|_2^2/2$. The learned features are the outputs of the last hidden layer. The random features are $\sqrt{2/2048}\cos(W^\top z + b)$, where $W \in \mathbb{R}^{d\times 2048}$ has independent $N(0, h^{-2})$ entries, h is the median pairwise distance between the training covariates, and b has independent coordinates uniform on $[0, 2\pi]$. For each feature map, level, and weight decay γ in the grid $\{10^{-6}, 10^{-5.5}, \ldots, 10^{1}\}$ of 15 values, we refit the head exactly with $\lambda = n\gamma$ and an unpenalized intercept. Extinct candidates use the closed form $(\alpha_Q, X^\top a_c)$. For the timing comparison, the dual solver also solves them, starting at the zero vector for the smallest weight decay and at the solution for the previous weight decay otherwise. Nonextinct candidates use the dual solver, started at the projection of λa_c, clipped to $[\tau - 1, \tau]^n$, onto the feasible set for the smallest nonextinct weight decay, and at the solution for the previous nonextinct weight decay otherwise. Validation selects, for each level, the weight decay with the smallest validation pinball loss, with ties resolved in favor of the largest weight decay. The boundary-aware search replaces all extinct candidates of a level by one closed-form fit. The reference predictors are the training τ-quantile of the response and unpenalized linear quantile regression on the standardized covariates, computed by the HiGHS solver in scikit-learn.

A pilot with two repetitions on the simulated model and California housing (seed 202609239) checked only the running time and the solver accuracy, and the production seed is 202609232. A second arm repeats the design with training weight decay $\gamma = 10^{-2}$ on the simulated model, California housing, and puma8NH (seed 202609234). The driver is `scripts/run_learned_heads.py`, and `scripts/validate_learned_heads.py` checks the counts, every extinction flag against the recorded boundary, the agreement of the solver with the closed form on extinct candidates, the duality gaps, the equality of the two selections, and the recomputation of the summaries and tables from the raw records.

Tables S.20–S.22 report the complete results of the two arms. The learned feature matrix has rank n in 500 of the 500 splits of the first arm and in 300 of the 300 splits of the second. All 120000 candidate fits have relative duality gap at most 2.2×10^{-10}, and on the extinct candidates the dual solver agrees with the closed form to within 7.9×10^{-11} in maximum norm. The boundary-aware search selects the same weight decay as the full search in every split, feature map, and level. In the first arm, it avoids 55% of the solver fits for learned features and 4.9% for random features, which account for 8.4% and 0.2% of the solver time. The trained heads interpolate the training responses only approximately. Across data sets, the median over splits of the largest absolute training residual ranges from 0.070 to 0.118 on the standardized response scale, and the mean spread of the five trained predictions at a test point ranges from 0.061 to 0.076. The trained heads have a higher mean pinball loss than the training quantiles of the response on puma8NH, house_16H, and diabetes and a lower one on the simulated model and California housing. The exact refit at the training weight decay 10^{-4} is extinct in 99.8% of the pairs of split and level, and its mean coverage ranges from 0.492 to 0.521 across data sets and levels. The refits with validated weight decay are extinct in 0.2% of these pairs, and the median ratio of the selected weight decay to the boundary is 141. Over the splits, random features with a validated weight decay have a lower mean pinball loss than the refitted learned heads on California housing, house_16H, and diabetes, by 0.010 to 0.022, and the two agree within two Monte Carlo standard errors of the paired difference on the simulated model and puma8NH. Table S.21 shows that the coverage of the exact refits moves toward the nominal levels only well above the boundary, for random as well as learned features. In the second arm, with training weight decay 10^{-2}, the refit at the training weight decay is extinct in no pair of split and level, yet its mean coverage ranges only from 0.492 to 0.532, and the trained heads have mean coverage from 0.481 to 0.546. Exact refits on these learned features keep mean coverage at the

levels 0.1 and 0.9 between 0.497 and 0.515 up to 1000 times the boundary of each level, and the refits with validated weight decay reach pooled coverages 0.162 and 0.825 at the levels 0.1 and 0.9.

Table S.20: Learned and random feature study with training weight decay 10^{-4}: median boundaries $\bar{\lambda}_c(\tau)/n$ and the shares of solver fits and solver time on the learned features avoided by the boundary-aware search (top), and mean test coverage by level and mean pinball loss over the five levels (bottom).

	Learned boundary		Random boundary		Solver fits	Solver time
Data	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.1$	$\tau = 0.5$	avoided	saved
Simulated	6.8×10^{-3}	3.2×10^{-2}	4.2×10^{-6}	1.8×10^{-5}	58%	11%
California	2.0×10^{-3}	9.0×10^{-3}	5.1×10^{-8}	2.2×10^{-7}	50%	5%
puma8NH	5.1×10^{-3}	2.4×10^{-2}	6.8×10^{-7}	2.9×10^{-6}	57%	13%
house_16H	3.2×10^{-3}	1.5×10^{-2}	5.0×10^{-7}	2.0×10^{-6}	54%	6%
Diabetes	3.7×10^{-3}	1.7×10^{-2}	2.1×10^{-7}	8.9×10^{-7}	55%	11%

Method	$\tau = 0.1$	0.25	0.5	0.75	0.9	Pinball
Simulated						
trained heads	0.483	0.490	0.496	0.505	0.512	0.197
refit, training decay	0.497	0.497	0.497	0.497	0.497	0.201
refit, validated decay	0.194	0.420	0.500	0.577	0.767	0.159
random, validated decay	0.187	0.362	0.498	0.637	0.816	0.160
linear QR	0.136	0.273	0.501	0.731	0.866	0.189
intercept only	0.096	0.248	0.498	0.739	0.898	0.282
California						
trained heads	0.505	0.514	0.521	0.529	0.538	0.252
refit, training decay	0.521	0.521	0.521	0.521	0.521	0.257
refit, validated decay	0.163	0.348	0.518	0.661	0.822	0.184
random, validated decay	0.170	0.305	0.505	0.703	0.842	0.162
linear QR	0.134	0.266	0.498	0.735	0.872	0.191
intercept only	0.100	0.247	0.496	0.742	0.898	0.276
puma8NH						
trained heads	0.482	0.488	0.493	0.497	0.502	0.334
refit, training decay	0.492	0.492	0.492	0.492	0.492	0.338
refit, validated decay	0.157	0.344	0.495	0.662	0.842	0.227
random, validated decay	0.142	0.281	0.500	0.714	0.860	0.227
linear QR	0.127	0.262	0.496	0.730	0.871	0.231
intercept only	0.103	0.249	0.495	0.747	0.903	0.282
house_16H						
trained heads	0.501	0.508	0.514	0.518	0.524	0.296
refit, training decay	0.514	0.514	0.514	0.514	0.514	0.300
refit, validated decay	0.152	0.350	0.510	0.629	0.847	0.214
random, validated decay	0.154	0.311	0.501	0.692	0.837	0.204
linear QR	0.168	0.288	0.508	0.719	0.843	0.247
intercept only	0.137	0.246	0.495	0.751	0.899	0.277
Diabetes						
trained heads	0.496	0.502	0.507	0.511	0.518	0.355
refit, training decay	0.507	0.507	0.507	0.507	0.507	0.360
refit, validated decay	0.163	0.336	0.512	0.674	0.848	0.227
random, validated decay	0.144	0.279	0.502	0.723	0.867	0.205
linear QR	0.139	0.269	0.498	0.731	0.864	0.213
intercept only	0.107	0.256	0.505	0.751	0.900	0.280

Table S.21: Mean test coverage of the exact refits by the ratio of the weight decay γ to the boundary $\bar{\lambda}_c(\tau)/n$ of the level, with training weight decay 10^{-4}, pooled over the data sets and splits.

	Learned features		Random features	
$\gamma/\{\bar{\lambda}_c(\tau)/n\}$	$\tau = 0.1$	$\tau = 0.9$	$\tau = 0.1$	$\tau = 0.9$
≤ 1 (extinct)	0.506	0.506	0.498	0.495
$(1, 3]$	0.506	0.506	0.494	0.502
$(3, 10]$	0.505	0.508	0.480	0.523
$(10, 30]$	0.491	0.520	0.444	0.564
$(30, 100]$	0.410	0.596	0.392	0.621
$(100, 300]$	0.236	0.758	0.330	0.688
$(300, 10^3]$	0.140	0.865	0.271	0.750
$> 10^3$	0.114	0.890	0.134	0.877

Table S.22: Second arm of the learned and random feature study, with training weight decay 10^{-2}, in the format of Table S.20.

	Learned boundary		Random boundary		Solver fits	Solver time
Data	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.1$	$\tau = 0.5$	avoided	saved
Simulated	2.4×10^{-5}	1.1×10^{-4}	3.8×10^{-6}	1.6×10^{-5}	26%	0%
California	3.9×10^{-7}	1.6×10^{-6}	4.7×10^{-8}	2.0×10^{-7}	4%	0%
puma8NH	3.9×10^{-5}	1.8×10^{-4}	6.6×10^{-7}	2.8×10^{-6}	28%	1%

Method	$\tau = 0.1$	0.25	0.5	0.75	0.9	Pinball
Simulated						
trained heads	0.489	0.496	0.504	0.511	0.518	0.187
refit, training decay	0.501	0.502	0.503	0.504	0.506	0.190
refit, validated decay	0.156	0.399	0.505	0.607	0.831	0.150
random, validated decay	0.186	0.368	0.501	0.641	0.817	0.160
linear QR	0.141	0.279	0.505	0.734	0.870	0.189
intercept only	0.105	0.254	0.499	0.748	0.902	0.281
California						
trained heads	0.488	0.506	0.518	0.530	0.546	0.241
refit, training decay	0.503	0.513	0.518	0.523	0.532	0.246
refit, validated decay	0.181	0.362	0.518	0.601	0.800	0.189
random, validated decay	0.163	0.310	0.501	0.692	0.839	0.164
linear QR	0.134	0.268	0.500	0.724	0.872	0.184
intercept only	0.101	0.252	0.501	0.747	0.900	0.273
puma8NH						
trained heads	0.481	0.488	0.496	0.502	0.510	0.314
refit, training decay	0.492	0.494	0.495	0.496	0.498	0.319
refit, validated decay	0.149	0.332	0.492	0.667	0.844	0.220
random, validated decay	0.144	0.284	0.502	0.721	0.858	0.226
linear QR	0.127	0.269	0.506	0.734	0.872	0.230
intercept only	0.099	0.247	0.493	0.744	0.904	0.281

S.30 UNPENALIZED AFFINE TERMS

This section adds unpenalized affine terms, an intercept in particular, and proves the extinction identity after the affine directions are projected out in the Q metric.

Let $1 \le d \le n$ and let $Z \in \mathbb{R}^{n\times d}$ have rank d. Write $z_i^\top$ for row i of Z. Define

$$P_Z^Q = Z(Z^\top QZ)^{-1}Z^\top Q, \qquad M_Z^Q = I_n - P_Z^Q.$$

For $y \in \mathbb{R}^n$, put

$$\theta_Z = (Z^\top QZ)^{-1}Z^\top Qy, \qquad \widetilde{y}_Z = M_Z^Q y, \qquad a_Z = Q\widetilde{y}_Z, \qquad \widehat{\beta}_{Z,0} = X^\top a_Z.$$

The following lemma identifies the minimum-norm interpolating pair after the unpenalized affine directions have been removed.

Lemma 7. *Among all $(\theta,\beta) \in \mathbb{R}^d \times \mathbb{R}^p$ satisfying $Z\theta + X\beta = y$, the unique pair with minimum $\|\beta\|_2$ is $(\theta_Z, \widehat{\beta}_{Z,0})$, and*

$$\left\|\widehat{\beta}_{Z,0}\right\|_2^2 = \widetilde{y}_Z^\top Q\widetilde{y}_Z.$$

Proof. For fixed θ, Lemma 3 with $f = y - Z\theta$ gives

$$\beta(\theta) = X^\top Q(y - Z\theta), \qquad \|\beta(\theta)\|_2^2 = (y - Z\theta)^\top Q(y - Z\theta).$$

Because $Z^\top QZ \succ 0$, the unique minimum satisfies

$$Z^\top Q(y - Z\theta) = 0,$$

and its unique solution is θ_Z. Substitution gives $\beta(\theta_Z) = \widehat{\beta}_{Z,0}$ and the norm identity. □

For $\tau \in (0,1)$ and $\lambda > 0$, define

$$\mathcal{M}_Z(\lambda,\tau) = \arg\min_{\theta\in\mathbb{R}^d,\,\beta\in\mathbb{R}^p} \left\{ \sum_{i=1}^n \rho_\tau(y_i - z_i^\top\theta - x_i^\top\beta) + \frac{\lambda}{2}\|\beta\|_2^2 \right\}.$$

For $i \le n$, let

$$b_{Z,i}(\tau) = \begin{cases} \tau/a_{Z,i}, & a_{Z,i} > 0, \\ (1-\tau)/|a_{Z,i}|, & a_{Z,i} < 0, \\ +\infty, & a_{Z,i} = 0, \end{cases} \qquad \bar{\lambda}_Z(\tau) = \min_{i\le n} b_{Z,i}(\tau).$$

The following theorem shows that unpenalized affine terms change the interpolation vector through a Q-orthogonal residualization while preserving the coordinatewise form of the extinction certificate.

Theorem 19. *For every $\tau \in (0,1)$ and $\lambda > 0$,*

$$(\theta_Z, \widehat{\beta}_{Z,0}) \in \mathcal{M}_Z(\lambda,\tau)$$

if and only if $\lambda a_Z \in [\tau - 1, \tau]^n$. Equivalently,

$$(\theta_Z, \widehat{\beta}_{Z,0}) \in \mathcal{M}_Z(\lambda,\tau) \qquad \Longleftrightarrow \qquad 0 < \lambda \le \bar{\lambda}_Z(\tau).$$

If these conditions hold, then

$$\mathcal{M}_Z(\lambda,\tau) = \{(\theta_Z, \widehat{\beta}_{Z,0})\}.$$

Proof. For $r_i = y_i - z_i^\top\theta - x_i^\top\beta$, convex optimality holds exactly when some $g \in \mathbb{R}^n$ satisfies

$$g_i \in \partial\rho_\tau(r_i), \qquad X^\top g = \lambda\beta, \qquad Z^\top g = 0.$$

At $(\theta_Z, \widehat{\beta}_{Z,0})$, the residual is zero because

$$Z\theta_Z + X\widehat{\beta}_{Z,0} = P_Z^Q y + KQ\widetilde{y}_Z = P_Z^Q y + M_Z^Q y = y.$$

The stationarity equation in β becomes

$$X^\top g = \lambda X^\top a_Z.$$

Injectivity of $X^\top$ gives $g = \lambda a_Z$. Stationarity in θ follows from

$$Z^\top a_Z = Z^\top Q M_Z^Q y = 0.$$

The KKT conditions at $(\theta_Z, \widehat{\beta}_{Z,0})$ are equivalent to $\lambda a_Z \in [\tau - 1, \tau]^n$, which gives the coordinate formula for $\bar{\lambda}_Z(\tau)$.

Convexity of the loss and strict convexity of $\|\beta\|_2^2$ make every minimizer share the coefficient $\widehat{\beta}_{Z,0}$. Its loss must be zero, giving

$$Z\theta = y - X\widehat{\beta}_{Z,0} = Z\theta_Z.$$

The rank condition on Z gives $\theta = \theta_Z$. This completes the proof. □

The following corollary gives the boundary of the residualized affine model. When a_Z has both positive and negative coordinates, the boundary is triangular and its peak is determined by the two signed extremes.

Put

$$A_{Z,+} = \max_{i \le n} a_{Z,i}^+, \qquad A_{Z,-} = \max_{i \le n} (-a_{Z,i})^+.$$

Corollary 26. *For every $\tau \in (0,1)$,*

$$\bar{\lambda}_Z(\tau) = \frac{\tau}{A_{Z,+}} \wedge \frac{1-\tau}{A_{Z,-}},$$

with $c/0 = +\infty$ for $c > 0$. If $A_{Z,+}A_{Z,-} > 0$, then $\bar{\lambda}_Z$ is concave and piecewise linear, with unique maximizer

$$\tau_Z^\dagger = \frac{A_{Z,+}}{A_{Z,+} + A_{Z,-}}, \qquad \bar{\lambda}_Z(\tau_Z^\dagger) = \frac{1}{A_{Z,+} + A_{Z,-}}.$$

Proof. Empty inner minima are interpreted as $+\infty$. The definition of $\bar{\lambda}_Z$ gives

$$\bar{\lambda}_Z(\tau) = \min\left\{ \min_{i: a_{Z,i} > 0} \frac{\tau}{a_{Z,i}}, \min_{i: a_{Z,i} < 0} \frac{1-\tau}{|a_{Z,i}|} \right\} = \frac{\tau}{A_{Z,+}} \wedge \frac{1-\tau}{A_{Z,-}}.$$

For $A_{Z,+}A_{Z,-} > 0$, the increasing and decreasing terms meet only at

$$\tau_Z^\dagger = \frac{A_{Z,+}}{A_{Z,+} + A_{Z,-}},$$

and the common value is $(A_{Z,+} + A_{Z,-})^{-1}$. This completes the proof. □

The following corollary gives the exact extinction certificate after a Q-weighted intercept has been removed.

Let

$$\alpha_Q = \frac{\mathbf{1}_n^\top Q y}{\mathbf{1}_n^\top Q \mathbf{1}_n}, \qquad \widetilde{y}_c = y - \alpha_Q \mathbf{1}_n, \qquad a_c = Q\widetilde{y}_c, \qquad \widehat{\beta}_{c,0} = X^\top a_c.$$

For $\tau \in (0,1)$ and $\lambda > 0$, define

$$\mathcal{M}_c(\lambda, \tau) = \arg \min_{\alpha \in \mathbb{R},\, \beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^n \rho_\tau(y_i - \alpha - x_i^\top \beta) + \frac{\lambda}{2} \|\beta\|_2^2 \right\}.$$

Corollary 27. *For every $\lambda > 0$ and $\tau \in (0,1)$,*

$$(\alpha_Q, \widehat{\beta}_{c,0}) \in \mathcal{M}_c(\lambda, \tau)$$

if and only if

$$\lambda a_c \in [\tau - 1, \tau]^n.$$

If this condition holds, then

$$\mathcal{M}_c(\lambda, \tau) = \{(\alpha_Q, \widehat{\beta}_{c,0})\}.$$

Proof. Apply Theorem 19 with $Z = \mathbf{1}_n$. □

The following proposition extends the level-separation dichotomy of Theorem 3 to an unpenalized intercept. Let $\bar{\lambda}_c(\tau)$ be defined as $\bar{\lambda}(\tau)$ with a replaced by a_c.

Proposition 46. *Let $\lambda > 0$ and let $\tau, \tau' \in (0,1)$ with $\tau \neq \tau'$. A pair (α, β) belongs to $\mathcal{M}_c(\lambda,\tau) \cap \mathcal{M}_c(\lambda,\tau')$ if and only if $\lambda \leq \bar{\lambda}_c(\tau) \wedge \bar{\lambda}_c(\tau')$ and $(\alpha,\beta) = (\alpha_Q, \widehat{\beta}_{c,0})$.*

Proof. The reverse implication follows from Corollary 27 at both levels. For the forward implication, write $r = y - \alpha\mathbf{1}_n - X\beta$. At a level τ, the objective is convex in (α,β), and the subgradient chain rule for r, as in the proof of Lemma 2, shows that $(\alpha,\beta) \in \mathcal{M}_c(\lambda,\tau)$ exactly when some $g \in \mathbb{R}^n$ satisfies $g_i \in \partial\rho_\tau(r_i)$ for every i, $\mathbf{1}_n^\top g = 0$, and $X^\top g = \lambda\beta$. Applying this at both levels gives vectors g and g' with $X^\top g = \lambda\beta = X^\top g'$, $\mathbf{1}_n^\top g = \mathbf{1}_n^\top g' = 0$, $g_i \in \partial\rho_\tau(r_i)$, and $g'_i \in \partial\rho_{\tau'}(r_i)$. Injectivity of $X^\top$ gives $g = g'$. If $r_i > 0$, then $g_i = \tau$ and $g_i = \tau'$. If $r_i < 0$, then $g_i = \tau - 1$ and $g_i = \tau' - 1$. Both cases contradict $\tau \neq \tau'$, and hence $r = 0$. Thus $X\beta = y - \alpha\mathbf{1}_n$ and $\beta = \lambda^{-1}X^\top g$, which give $g = \lambda Q(y - \alpha\mathbf{1}_n)$. The condition $\mathbf{1}_n^\top g = 0$ gives $\alpha = \alpha_Q$, and hence $g = \lambda a_c$ and $\beta = \widehat{\beta}_{c,0}$. Because g is a subgradient vector at both levels, $\lambda a_c \in [\tau - 1, \tau]^n \cap [\tau' - 1, \tau']^n$, which is equivalent to $\lambda \leq \bar{\lambda}_c(\tau) \wedge \bar{\lambda}_c(\tau')$ by the coordinatewise definition of $\bar{\lambda}_c$. This completes the proof. □

S.31 Weighted affine extinction

This section establishes an extinction identity with observation weights, a positive definite quadratic penalty, and unpenalized affine terms.

Let $1 \leq d \leq n$, let $H = H^\top \in \mathbb{R}^{p\times p}$ satisfy $H \succ 0$, let $w = (w_1,\dots,w_n)^\top \in (0,\infty)^n$, and let $Z \in \mathbb{R}^{n\times d}$ have rank d. Write $z_i^\top$ for row i of Z. Define

$$K_H = XH^{-1}X^\top, \qquad Q_H = K_H^{-1},$$
$$P_{Z,H} = Z(Z^\top Q_H Z)^{-1}Z^\top Q_H, \qquad M_{Z,H} = I_n - P_{Z,H}.$$

For $y \in \mathbb{R}^n$, put

$$\theta_{Z,H} = (Z^\top Q_H Z)^{-1}Z^\top Q_H y, \quad \widetilde{y}_{Z,H} = M_{Z,H}y, \quad a_{Z,H} = Q_H\widetilde{y}_{Z,H}, \quad \widehat{\beta}_{Z,H,0} = H^{-1}X^\top a_{Z,H}.$$

For $\tau \in (0,1)$ and $\lambda > 0$, let

$$\mathcal{M}_{Z,H,w}(\lambda,\tau) = \arg\min_{\theta\in\mathbb{R}^d,\,\beta\in\mathbb{R}^p}\left\{\sum_{i=1}^n w_i\rho_\tau(y_i - z_i^\top\theta - x_i^\top\beta) + \frac{\lambda}{2}\beta^\top H\beta\right\}.$$

Define

$$M_{Z,H,w,+} = \max_{i:a_{Z,H,i}>0}\frac{a_{Z,H,i}}{w_i}, \qquad M_{Z,H,w,-} = \max_{i:a_{Z,H,i}<0}\frac{|a_{Z,H,i}|}{w_i},$$

with empty maxima equal to 0 and $c/0 = +\infty$ for $c > 0$.

The following theorem gives the exact coordinatewise extinction boundary for the weighted affine model with a general quadratic penalty.

Theorem 20. *For every $\tau \in (0,1)$ and $\lambda > 0$,*

$$(\theta_{Z,H}, \widehat{\beta}_{Z,H,0}) \in \mathcal{M}_{Z,H,w}(\lambda,\tau)$$

if and only if

$$\lambda a_{Z,H,i} \in [w_i(\tau-1), w_i\tau] \qquad (i = 1,\dots,n).$$

Equivalently,

$$(\theta_{Z,H}, \widehat{\beta}_{Z,H,0}) \in \mathcal{M}_{Z,H,w}(\lambda,\tau) \iff 0 < \lambda \leq \frac{\tau}{M_{Z,H,w,+}} \wedge \frac{1-\tau}{M_{Z,H,w,-}}.$$

If these conditions hold, then

$$\mathcal{M}_{Z,H,w}(\lambda,\tau) = \{(\theta_{Z,H}, \widehat{\beta}_{Z,H,0})\}.$$

Proof. Lemma 5 and $\mathrm{rank}(Z) = d$ give $K_H \succ 0$ and $Z^\top Q_H Z \succ 0$. By definition,

$$Z\theta_{Z,H} = P_{Z,H}y, \qquad \widetilde{y}_{Z,H} = y - Z\theta_{Z,H}, \qquad Z^\top Q_H \widetilde{y}_{Z,H} = 0.$$

Also,

$$X\widehat{\beta}_{Z,H,0} = XH^{-1}X^\top a_{Z,H} = K_H Q_H \widetilde{y}_{Z,H} = \widetilde{y}_{Z,H}.$$

Thus $Z\theta_{Z,H} + X\widehat{\beta}_{Z,H,0} = y$.

The KKT conditions for the weighted affine criterion are necessary and sufficient: for $r_i = y_i - z_i^\top \theta - x_i^\top \beta$, there is $g \in \mathbb{R}^n$ such that

$$g_i \in w_i \partial \rho_\tau(r_i), \qquad X^\top g = \lambda H\beta, \qquad Z^\top g = 0.$$

At this zero-residual pair,

$$H\widehat{\beta}_{Z,H,0} = X^\top a_{Z,H}.$$

The stationarity equation in β is

$$X^\top g = \lambda X^\top a_{Z,H}.$$

Injectivity of $X^\top$ gives $g = \lambda a_{Z,H}$. Stationarity in θ holds because

$$Z^\top a_{Z,H} = Z^\top Q_H \widetilde{y}_{Z,H} = 0.$$

The KKT conditions at the zero-residual pair are exactly

$$\lambda a_{Z,H,i} \in [w_i(\tau - 1), w_i\tau] \qquad (i = 1, \ldots, n).$$

For $a_{Z,H,i} > 0$, this is $\lambda \le w_i\tau / a_{Z,H,i}$. For $a_{Z,H,i} < 0$, it is $\lambda \le w_i(1-\tau)/|a_{Z,H,i}|$. Coordinates with $a_{Z,H,i} = 0$ impose no upper bound. Taking the minimum over coordinates gives the stated boundary.

Convexity of the loss and strict convexity of $\beta^\top H\beta$ make every minimizer share the coefficient $\widehat{\beta}_{Z,H,0}$. Its check loss must be zero, hence

$$Z\theta + X\widehat{\beta}_{Z,H,0} = y = Z\theta_{Z,H} + X\widehat{\beta}_{Z,H,0}.$$

The rank condition on Z gives $\theta = \theta_{Z,H}$. This completes the proof. □

The following corollary gives the boundary under observation weights and a quadratic penalty. When both weighted signed extremes of $a_{Z,H}$ are positive, the boundary is triangular with slopes determined by those extremes.

Corollary 28. *For every $\tau \in (0,1)$, define*

$$\bar{\lambda}_{Z,H,w}(\tau) = \frac{\tau}{M_{Z,H,w,+}} \wedge \frac{1-\tau}{M_{Z,H,w,-}}.$$

If $M_{Z,H,w,+}M_{Z,H,w,-} > 0$, then $\bar{\lambda}_{Z,H,w}$ is concave and piecewise linear, with unique maximizer

$$\tau^\dagger_{Z,H,w} = \frac{M_{Z,H,w,+}}{M_{Z,H,w,+} + M_{Z,H,w,-}}, \qquad \bar{\lambda}_{Z,H,w}(\tau^\dagger_{Z,H,w}) = \frac{1}{M_{Z,H,w,+} + M_{Z,H,w,-}}.$$

Proof. For $M_{Z,H,w,+}M_{Z,H,w,-} > 0$, the increasing and decreasing terms defining $\bar{\lambda}_{Z,H,w}$ meet only at

$$\tau^\dagger_{Z,H,w} = \frac{M_{Z,H,w,+}}{M_{Z,H,w,+} + M_{Z,H,w,-}},$$

and the common value is $(M_{Z,H,w,+} + M_{Z,H,w,-})^{-1}$. This completes the proof. □

S.32 ISOTROPIC–ANISOTROPIC BOUNDARIES

The extinction boundary and its maximizing level are stable under Gram-matrix perturbations when both signed extremes of the interpolation vector stay away from zero. We also give an explicit margin under which an extinction certificate remains valid for the perturbed penalty.

Let $H = H^\top \in \mathbb{R}^{p\times p}$ satisfy $H \succ 0$, and put

$$K_H = XH^{-1}X^\top, \qquad Q_H = K_H^{-1}, \qquad a_H = Q_H y.$$

Define

$$A_{H,+} = \max_{i\le n} a_{H,i}^+, \qquad A_{H,-} = \max_{i\le n}(-a_{H,i})^+,$$

and, for $\tau \in (0,1)$,

$$\bar\lambda_H(\tau) = \frac{\tau}{A_{H,+}} \wedge \frac{1-\tau}{A_{H,-}},$$

with $c/0 = +\infty$ for $c > 0$. Let

$$\Delta_H = K_H - K, \qquad \varepsilon_H = \|Q\Delta_H\|_{\infty\to\infty}.$$

The following proposition quantifies how changing the quadratic penalty changes the extinction boundary and its maximizer.

Proposition 47. *Assume $\varepsilon_H < 1$ and put*

$$\delta_H = \frac{\varepsilon_H \|a\|_\infty}{1-\varepsilon_H}.$$

Then

$$\|a_H - a\|_\infty \le \delta_H.$$

If $\alpha = A_+ \wedge A_- > \delta_H$, then

$$\sup_{\tau\in(0,1)} |\bar\lambda_H(\tau) - \bar\lambda(\tau)| \le \frac{\delta_H}{\alpha(\alpha-\delta_H)}.$$

If $A_+A_- > 0$ and $\alpha > \delta_H$, then, with

$$\tau_H^\dagger = \frac{A_{H,+}}{A_{H,+}+A_{H,-}},$$

we have

$$|\tau_H^\dagger - \tau^\dagger| \le \frac{\delta_H}{A_+ + A_- - 2\delta_H}.$$

Proof. The equations $Ka = y$ and $(K+\Delta_H)a_H = y$ give

$$a_H - a = -Q\Delta_H a_H = -Q\Delta_H a - Q\Delta_H(a_H - a).$$

Taking ℓ_∞ norms yields

$$\|a_H - a\|_\infty \le \varepsilon_H \|a\|_\infty + \varepsilon_H \|a_H - a\|_\infty.$$

Because $\varepsilon_H < 1$, rearranging gives the first bound.

The boundary and maximizer bounds are Proposition 27 applied to $\widetilde a = a_H$. □

The following proposition gives a margin condition under which an isotropic extinction certificate remains valid for the anisotropic penalty.

Proposition 48. *Let $T \subset (0,1)$ be nonempty, let $\lambda > 0$, and keep the notation of Proposition 47. Define*

$$m_{\lambda,T} = \inf_{\tau\in T} \min_{i\le n}\{\tau - \lambda a_i,\ 1-\tau+\lambda a_i\}.$$

Assume $\varepsilon_H < 1$ and

$$m_{\lambda,T} \ge \lambda\delta_H.$$

Let $\widehat\beta_H(\lambda,\tau)$ denote the estimator with penalty $(\lambda/2)\beta^\top H\beta$ and unit observation weights, and let

$$\widehat\beta_{H,0} = H^{-1}X^\top a_H.$$

Then

$$\widehat\beta_H(\lambda,\tau) = \widehat\beta_{H,0} \qquad (\tau\in T).$$

Proof. Proposition 47 gives

$$|a_{H,i} - a_i| \le \delta_H \qquad (i \le n).$$

For every $\tau \in T$ and $i \le n$,

$$\lambda a_{H,i} \le \lambda a_i + \lambda\delta_H \le \tau - m_{\lambda,T} + \lambda\delta_H \le \tau,$$

and

$$\lambda a_{H,i} \ge \lambda a_i - \lambda\delta_H \ge \tau - 1 + m_{\lambda,T} - \lambda\delta_H \ge \tau - 1.$$

Theorem 13 with $w_i = 1$ gives the stated identity. □

S.33 General-rank interpolation certificates

The full-row-rank condition makes the zero-residual subgradient certificate coordinatewise. When $X^\top$ has a nontrivial null space, the same stationarity equation allows a null-space adjustment of the zero-residual subgradient. The extinction boundary becomes the largest multiple of the minimum-norm interpolant that remains in the polytope $X^\top \mathcal{B}_\tau$.

Let $X \in \mathbb{R}^{n\times p}$ have arbitrary rank and assume $y \in \mathrm{range}(X)$. Put $K = XX^\top$, let K^+ denote the Moore–Penrose inverse of K, and define

$$\widehat{\beta}^\dagger_X = X^\top K^+ y.$$

For $\tau \in (0,1)$ and $\lambda > 0$, let

$$\widehat{\beta}_X(\lambda,\tau) = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \sum_{i=1}^n \rho_\tau(y_i - x_i^\top\beta) + \frac{\lambda}{2}\|\beta\|_2^2 \right\}.$$

The following lemma identifies the minimum-norm interpolant and replaces coordinatewise membership by a subgradient feasibility condition at arbitrary rank.

Lemma 8. *The vector $\widehat{\beta}^\dagger_X$ is the unique minimum-norm vector satisfying $X\beta = y$. Also,*

$$\widehat{\beta}_X(\lambda,\tau) = \widehat{\beta}^\dagger_X \quad\Longleftrightarrow\quad \exists g \in \mathcal{B}_\tau \text{ such that } X^\top g = \lambda\widehat{\beta}^\dagger_X.$$

Proof. Because $y \in \mathrm{range}(X) = \mathrm{range}(K)$,

$$X\widehat{\beta}^\dagger_X = XX^\top K^+ y = KK^+ y = y.$$

If $X\beta = y$, write $\beta = \widehat{\beta}^\dagger_X + v$. Then $Xv = 0$, and

$$(\widehat{\beta}^\dagger_X)^\top v = y^\top K^+ Xv = 0.$$

This gives

$$\|\beta\|_2^2 = \left\|\widehat{\beta}^\dagger_X\right\|_2^2 + \|v\|_2^2,$$

which proves the minimum-norm claim.

The criterion is closed, proper, and λ-strongly convex. With $r = y - X\beta$, KKT optimality is equivalent to existence of $g \in \mathbb{R}^n$ satisfying

$$g_i \in \partial\rho_\tau(r_i), \qquad X^\top g = \lambda\beta.$$

At $\widehat{\beta}^\dagger_X$, zero residual gives $g \in \mathcal{B}_\tau$, and stationarity gives $X^\top g = \lambda\widehat{\beta}^\dagger_X$. □

Define

$$\bar{\lambda}^\dagger_X(\tau) = \sup\{t > 0 : t\widehat{\beta}^\dagger_X \in X^\top\mathcal{B}_\tau\},$$

where the supremum is $+\infty$ when the set is unbounded.

The following theorem replaces coordinatewise thresholds by radial membership in $X^\top\mathcal{B}_\tau$, which gives the exact extinction boundary when X is rank deficient.

Theorem 21. *For every $\tau \in (0,1)$ and $\lambda > 0$,*

$$\widehat{\beta}_X(\lambda,\tau) = \widehat{\beta}^\dagger_X \quad \Longleftrightarrow \quad 0 < \lambda \le \bar{\lambda}^\dagger_X(\tau).$$

If $\widehat{\beta}^\dagger_X = 0$, then $\bar{\lambda}^\dagger_X(\tau) = +\infty$. If $\widehat{\beta}^\dagger_X \neq 0$, then $0 < \bar{\lambda}^\dagger_X(\tau) < \infty$.

Proof. Lemma 8 gives

$$\widehat{\beta}_X(\lambda,\tau) = \widehat{\beta}^\dagger_X \quad \Longleftrightarrow \quad \lambda\widehat{\beta}^\dagger_X \in X^\top \mathcal{B}_\tau.$$

Put $C = X^\top \mathcal{B}_\tau$, a compact convex set containing 0. For $\widehat{\beta}^\dagger_X = 0$, every $t > 0$ is feasible and $\bar{\lambda}^\dagger_X(\tau) = +\infty$.

For $\widehat{\beta}^\dagger_X \neq 0$, the box $\mathcal{B}_\tau$ contains a ball around zero. Its image under $X^\top$ contains a relative neighborhood in $\mathrm{range}(X^\top)$, hence some $\varepsilon > 0$ satisfies

$$t\widehat{\beta}^\dagger_X \in C \qquad (0 \le t \le \varepsilon).$$

The set $S_0 = \{t \ge 0 : t\widehat{\beta}^\dagger_X \in C\}$ is compact and contains $[0,\varepsilon]$. For $t \in S_0$ and $0 \le s \le t$ with $t > 0$,

$$s\widehat{\beta}^\dagger_X = (s/t)t\widehat{\beta}^\dagger_X \in C$$

by convexity of C and $0 \in C$; if $t = 0$, then $s = 0 \in S_0$. Thus $S_0 = [0, \bar{\lambda}^\dagger_X(\tau)]$, giving the equivalence. □

For $s \in \mathbb{R}^n$, define

$$H_\tau(s) = \sum_{i=1}^n \{\tau s_i^+ + (1-\tau)(-s_i)^+\}.$$

This is the support function of $\mathcal{B}_\tau$.

The following proposition expresses the general-rank boundary as the smallest support-function ratio over directions with a positive interpolant contrast.

Proposition 49. *If $\widehat{\beta}^\dagger_X \neq 0$, then*

$$\bar{\lambda}^\dagger_X(\tau) = \inf_{v \in \mathbb{R}^p:\ v^\top \widehat{\beta}^\dagger_X > 0} \frac{H_\tau(Xv)}{v^\top \widehat{\beta}^\dagger_X}.$$

Proof. Let $C = X^\top \mathcal{B}_\tau$. For $v \in \mathbb{R}^p$,

$$\sup_{c \in C} v^\top c = \sup_{g \in \mathcal{B}_\tau} g^\top Xv = H_\tau(Xv).$$

If $t\widehat{\beta}^\dagger_X \in C$, then

$$tv^\top \widehat{\beta}^\dagger_X \le H_\tau(Xv) \qquad (v \in \mathbb{R}^p).$$

For every v with $v^\top \widehat{\beta}^\dagger_X > 0$, this gives

$$t \le \frac{H_\tau(Xv)}{v^\top \widehat{\beta}^\dagger_X}.$$

Taking the supremum over feasible t proves the upper bound.

For $t > \bar{\lambda}^\dagger_X(\tau)$, separation of $t\widehat{\beta}^\dagger_X$ from compact convex C gives $v \in \mathbb{R}^p$ with

$$tv^\top \widehat{\beta}^\dagger_X > \sup_{c \in C} v^\top c = H_\tau(Xv).$$

Because $0 \in C$, we have $v^\top \widehat{\beta}^\dagger_X > 0$, hence

$$\frac{H_\tau(Xv)}{v^\top \widehat{\beta}^\dagger_X} < t.$$

Taking $t \downarrow \bar{\lambda}^\dagger_X(\tau)$ proves the lower bound. □

The following corollary verifies that the ray–polytope boundary reduces exactly to the coordinate-wise boundary under full row rank.

Corollary 29. *If* $\operatorname{rank}(X) = n$*, then* $K^+ = Q$*,* $\widehat{\beta}^\dagger_X = \widehat{\beta}_0$*, and*

$$\bar{\lambda}^\dagger_X(\tau) = \bar{\lambda}(\tau) \qquad (0 < \tau < 1).$$

Proof. Full row rank gives $K^+ = Q$, $\widehat{\beta}^\dagger_X = X^\top Q y = \widehat{\beta}_0$, and injectivity of $X^\top$. Thus

$$t\widehat{\beta}_0 \in X^\top \mathcal{B}_\tau \quad \Longleftrightarrow \quad tX^\top a = X^\top g \text{ for some } g \in \mathcal{B}_\tau \quad \Longleftrightarrow \quad ta \in \mathcal{B}_\tau.$$

The supremum over positive t satisfying the last condition is $\bar{\lambda}(\tau)$ by Theorem 1. □

The following corollary shows that every certified level has the same coefficient and prediction function in the arbitrary-rank model.

Corollary 30. *Let* $T \subset (0,1)$ *be nonempty. If* $0 < \lambda \le \inf_{\tau \in T} \bar{\lambda}^\dagger_X(\tau)$*, then*

$$x^\top \widehat{\beta}_X(\lambda, \tau) = x^\top \widehat{\beta}^\dagger_X \qquad (x \in \mathbb{R}^p,\ \tau \in T).$$

Proof. For each $\tau \in T$, Theorem 21 gives

$$\widehat{\beta}_X(\lambda, \tau) = \widehat{\beta}^\dagger_X.$$

Multiplication by x proves the stated identity. □

The following proposition turns feasibility of the zero-residual subgradient equations into an exact finite-grid extinction certificate.

Proposition 50. *Let* $\mathcal{G} \subset (0,\infty) \times (0,1)$ *be finite. For each* $(\lambda, \tau) \in \mathcal{G}$*, suppose there exists* $g_{\lambda,\tau} \in \mathbb{R}^n$ *satisfying*

$$\tau - 1 \le g_{\lambda,\tau,i} \le \tau \quad (i = 1, \dots, n), \qquad X^\top g_{\lambda,\tau} = \lambda \widehat{\beta}^\dagger_X.$$

Then

$$\widehat{\beta}_X(\lambda, \tau) = \widehat{\beta}^\dagger_X \qquad ((\lambda, \tau) \in \mathcal{G}).$$

If the last identity holds for every $(\lambda, \tau) \in \mathcal{G}$*, then such vectors* $g_{\lambda,\tau}$ *exist.*

Proof. The inequalities give $g_{\lambda,\tau} \in \mathcal{B}_\tau$. Lemma 8 gives both implications at each grid point. □

S.33.1 Centered weighted general-rank certificate

The centered weighted certificate has the same ray-polytope form without the full-row-rank condition. The coordinatewise formula is recovered when X has full row rank.

Let $X \in \mathbb{R}^{n \times p}$ have arbitrary rank. Let $H = H^\top \in \mathbb{R}^{p \times p}$ satisfy $H \succ 0$, let $b \in \mathbb{R}^p$, let $w = (w_1, \dots, w_n)^\top \in (0,\infty)^n$, and assume $\nu_b = y - Xb \in \operatorname{range}(X)$. Put

$$K_H = XH^{-1}X^\top, \qquad \widehat{\beta}^\dagger_{H,b,X} = b + H^{-1}X^\top K_H^+ \nu_b,$$

and

$$\zeta_{H,b,X} = H(\widehat{\beta}^\dagger_{H,b,X} - b) = X^\top K_H^+ \nu_b.$$

For $\tau \in (0,1)$ and $\lambda > 0$, define

$$\widehat{\beta}_{H,b,X}(\lambda, \tau) = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^n w_i \rho_\tau(y_i - x_i^\top \beta) + \frac{\lambda}{2}(\beta - b)^\top H (\beta - b) \right\}.$$

The following lemma identifies the centered weighted interpolant and its subgradient feasibility condition at arbitrary rank.

Lemma 9. *The vector $\widehat{\beta}^{\dagger}_{H,b,X}$ is the unique minimizer of $(\beta - b)^\top H(\beta - b)$ over $\{\beta : X\beta = y\}$. Also,*

$$\widehat{\beta}_{H,b,X}(\lambda, \tau) = \widehat{\beta}^{\dagger}_{H,b,X} \quad \Longleftrightarrow \quad \exists g \in \mathcal{B}_{\tau,w} \text{ such that } X^\top g = \lambda \zeta_{H,b,X}.$$

Proof. Because $\nu_b \in \mathrm{range}(X) = \mathrm{range}(K_H)$,

$$X\widehat{\beta}^{\dagger}_{H,b,X} = Xb + K_H K_H^{+} \nu_b = y.$$

If $X\beta = y$, set $v = \beta - \widehat{\beta}^{\dagger}_{H,b,X}$. Then $Xv = 0$ and

$$(\widehat{\beta}^{\dagger}_{H,b,X} - b)^\top Hv = \nu_b^\top K_H^{+} Xv = 0.$$

This gives

$$(\beta - b)^\top H(\beta - b) = (\widehat{\beta}^{\dagger}_{H,b,X} - b)^\top H(\widehat{\beta}^{\dagger}_{H,b,X} - b) + v^\top Hv,$$

which proves the first claim.

The criterion is closed, proper, and strongly convex. With $r = y - X\beta$, KKT optimality is equivalent to existence of $g \in \mathbb{R}^n$ satisfying

$$g_i \in w_i \partial \rho_\tau(r_i), \qquad X^\top g = \lambda H(\beta - b).$$

At $\widehat{\beta}^{\dagger}_{H,b,X}$, zero residual gives $g \in \mathcal{B}_{\tau,w}$, and stationarity gives $X^\top g = \lambda \zeta_{H,b,X}$. □

Define

$$\bar{\lambda}^{\dagger}_{H,b,w,X}(\tau) = \sup\{t > 0 : t\zeta_{H,b,X} \in X^\top \mathcal{B}_{\tau,w}\},$$

where the supremum is $+\infty$ when the set is unbounded.

The following theorem shows that the same ray–polytope geometry applies to weighted check loss with a centered quadratic penalty at arbitrary rank.

Theorem 22. *For every $\tau \in (0, 1)$ and $\lambda > 0$,*

$$\widehat{\beta}_{H,b,X}(\lambda, \tau) = \widehat{\beta}^{\dagger}_{H,b,X} \quad \Longleftrightarrow \quad 0 < \lambda \le \bar{\lambda}^{\dagger}_{H,b,w,X}(\tau).$$

If $\zeta_{H,b,X} = 0$, then $\bar{\lambda}^{\dagger}_{H,b,w,X}(\tau) = +\infty$. If $\zeta_{H,b,X} \neq 0$, then $0 < \bar{\lambda}^{\dagger}_{H,b,w,X}(\tau) < \infty$.

Proof. Lemma 9 gives

$$\widehat{\beta}_{H,b,X}(\lambda, \tau) = \widehat{\beta}^{\dagger}_{H,b,X} \quad \Longleftrightarrow \quad \lambda\zeta_{H,b,X} \in X^\top \mathcal{B}_{\tau,w}.$$

Put $C = X^\top \mathcal{B}_{\tau,w}$, a compact convex set containing 0. For $\zeta_{H,b,X} = 0$, every $t > 0$ is feasible, giving the infinite boundary.

For $\zeta_{H,b,X} \neq 0$, the box $\mathcal{B}_{\tau,w}$ contains a ball around zero. Its image contains a relative neighborhood in $\mathrm{range}(X^\top)$, hence some $\varepsilon > 0$ satisfies

$$t\zeta_{H,b,X} \in C \qquad (0 \le t \le \varepsilon).$$

Let $S_0 = \{t \ge 0 : t\zeta_{H,b,X} \in C\}$. This set is compact, contains $[0, \varepsilon]$, and is an interval: if $t \in S_0$ and $0 \le s \le t$, then $s\zeta_{H,b,X} \in C$ by convexity of C and $0 \in C$. Thus $S_0 = [0, \bar{\lambda}^{\dagger}_{H,b,w,X}(\tau)]$, giving the equivalence. □

For $s \in \mathbb{R}^n$, define

$$H_{\tau,w}(s) = \sum_{i=1}^{n} w_i\{\tau s_i^{+} + (1 - \tau)(-s_i)^{+}\}.$$

This is the support function of $\mathcal{B}_{\tau,w}$.

The following proposition gives the corresponding support-function formula, with observation weights entering coordinatewise.

Proposition 51. *If $\zeta_{H,b,X} \neq 0$, then*

$$\bar{\lambda}^{\dagger}_{H,b,w,X}(\tau) = \inf_{v\in\mathbb{R}^p:\ v^\top \zeta_{H,b,X}>0} \frac{H_{\tau,w}(Xv)}{v^\top \zeta_{H,b,X}}.$$

Proof. Let $C = X^\top \mathcal{B}_{\tau,w}$. For $v \in \mathbb{R}^p$,

$$\sup_{c\in C} v^\top c = \sup_{g\in\mathcal{B}_{\tau,w}} g^\top Xv = H_{\tau,w}(Xv).$$

If $t\zeta_{H,b,X} \in C$, then

$$tv^\top \zeta_{H,b,X} \leq H_{\tau,w}(Xv) \qquad (v \in \mathbb{R}^p).$$

Taking the supremum over feasible t proves the upper bound.

For $t > \bar{\lambda}^{\dagger}_{H,b,w,X}(\tau)$, separation of $t\zeta_{H,b,X}$ from C gives $v \in \mathbb{R}^p$ with

$$tv^\top \zeta_{H,b,X} > \sup_{c\in C} v^\top c = H_{\tau,w}(Xv).$$

Because $0 \in C$, we have $v^\top \zeta_{H,b,X} > 0$, hence

$$\frac{H_{\tau,w}(Xv)}{v^\top \zeta_{H,b,X}} < t.$$

Taking $t \downarrow \bar{\lambda}^{\dagger}_{H,b,w,X}(\tau)$ proves the lower bound. □

The following proposition computes the centered weighted boundary by a finite linear program over the ray length and subgradient vector.

Proposition 52. *If $\zeta_{H,b,X} \neq 0$, then, for every $\tau \in (0,1)$,*

$$\bar{\lambda}^{\dagger}_{H,b,w,X}(\tau) = \max\{t \geq 0 : \exists g \in \mathbb{R}^n,\ X^\top g = t\zeta_{H,b,X},\ w_i(\tau-1) \leq g_i \leq w_i\tau\ (i \leq n)\}.$$

The maximum is attained.

Proof. The coordinate inequalities are exactly $g \in \mathcal{B}_{\tau,w}$. The equality constraint is $t\zeta_{H,b,X} \in X^\top \mathcal{B}_{\tau,w}$. Theorem 22 identifies the feasible values as $[0, \bar{\lambda}^{\dagger}_{H,b,w,X}(\tau)]$, with a finite positive upper bound when $\zeta_{H,b,X} \neq 0$. Thus the maximum is attained and equals the boundary. □

For $u \in \mathbb{R}^n$, write $\|u\|_1 = \sum_{i=1}^n |u_i|$. The following proposition gives the dual linear program and proves that an optimizer exists.

Proposition 53. *If $\zeta_{H,b,X} \neq 0$, then, for every $\tau \in (0,1)$,*

$$\bar{\lambda}^{\dagger}_{H,b,w,X}(\tau) = \min_{\substack{v\in\mathbb{R}^p,\ \alpha,\gamma\in\mathbb{R}^n \\ Xv=\alpha-\gamma,\ \alpha\geq 0,\ \gamma\geq 0 \\ v^\top\zeta_{H,b,X}=1}} \left\{\tau w^\top \alpha + (1-\tau) w^\top \gamma\right\}.$$

The minimum is attained.

Proof. For $s \in \mathbb{R}^n$,

$$H_{\tau,w}(s) = \min_{\alpha,\gamma\in\mathbb{R}^n:\alpha-\gamma=s,\ \alpha\geq 0,\ \gamma\geq 0} \{\tau w^\top \alpha + (1-\tau) w^\top \gamma\}.$$

The minimum is attained at $(\alpha_i, \gamma_i) = (s_i, 0)$ for $s_i \geq 0$ and $(0, -s_i)$ for $s_i < 0$.

Proposition 51 and positive homogeneity of $H_{\tau,w}$ give

$$\bar{\lambda}^{\dagger}_{H,b,w,X}(\tau) = \inf_{v\in\mathbb{R}^p:\ v^\top\zeta_{H,b,X}=1} H_{\tau,w}(Xv).$$

Substitution of the coordinatewise linear representation gives the stated optimization problem.

For attainment, take a minimizing sequence (v_m) with $v_m^\top \zeta_{H,b,X} = 1$. Decompose $v_m = u_m + z_m$, where $u_m \in \mathrm{range}(X^\top)$ and $z_m \in \ker(X)$. Because $\zeta_{H,b,X} \in \mathrm{range}(X^\top)$,

$$u_m^\top \zeta_{H,b,X} = 1, \qquad Xu_m = Xv_m.$$

Let $c_{\tau,w} = \min_i w_i\{\tau \wedge (1-\tau)\} > 0$. Then

$$H_{\tau,w}(Xu_m) \geq c_{\tau,w} \left\|Xu_m\right\|_1 .$$

Thus (Xu_m) is bounded. The restriction of X to $\mathrm{range}(X^\top)$ has a bounded inverse, giving bounded (u_m). A convergent subsequence attains the infimum; choose α, γ coordinatewise from its limit Xu. □

The following corollary certifies the boundary exactly when feasible primal and dual variables have the same objective value.

Corollary 31. *Fix $\tau \in (0,1)$ and suppose $\zeta_{H,b,X} \neq 0$. Suppose $t \geq 0$, $g \in \mathbb{R}^n$, $v \in \mathbb{R}^p$, and $\alpha, \gamma \in \mathbb{R}^n$ satisfy*

$$X^\top g = t\zeta_{H,b,X}, \qquad w_i(\tau-1) \leq g_i \leq w_i\tau \quad (i \leq n),$$
$$Xv = \alpha - \gamma, \qquad \alpha \geq 0, \qquad \gamma \geq 0, \qquad v^\top \zeta_{H,b,X} = 1, \qquad t = \tau w^\top \alpha + (1-\tau) w^\top \gamma.$$

Then

$$t = \bar{\lambda}^{\dagger}_{H,b,w,X}(\tau).$$

Proof. For any primal feasible pair (t, g) and any dual feasible triple (v, α, γ),

$$t = tv^\top \zeta_{H,b,X} = v^\top X^\top g = g^\top(\alpha - \gamma).$$

Because $g_i \leq w_i\tau$ and $g_i \geq w_i(\tau-1)$,

$$g^\top(\alpha-\gamma) \leq \tau w^\top \alpha + (1-\tau) w^\top \gamma.$$

Equality identifies the common value as $\bar{\lambda}^{\dagger}_{H,b,w,X}(\tau)$ by Propositions 52 and 53. □

The following corollary applies the weighted subgradient feasibility condition simultaneously over any finite penalty–level grid.

Corollary 32. *Let $\mathcal{G} \subset (0,\infty)\times(0,1)$ be finite. For each $(\lambda,\tau) \in \mathcal{G}$, suppose there exists $g_{\lambda,\tau} \in \mathbb{R}^n$ satisfying*

$$w_i(\tau-1) \leq g_{\lambda,\tau,i} \leq w_i\tau \quad (i = 1,\dots,n), \qquad X^\top g_{\lambda,\tau} = \lambda \zeta_{H,b,X}.$$

Then

$$\widehat{\beta}_{H,b,X}(\lambda,\tau) = \widehat{\beta}^{\dagger}_{H,b,X} \qquad ((\lambda,\tau) \in \mathcal{G}).$$

If the last identity holds for every $(\lambda,\tau) \in \mathcal{G}$, then such vectors $g_{\lambda,\tau}$ exist.

Proof. The inequalities give $g_{\lambda,\tau} \in \mathcal{B}_{\tau,w}$. Lemma 9 gives both implications at each grid point. □

The following corollary verifies that the arbitrary-rank centered boundary reduces to the earlier coordinate formula when X has full row rank.

Corollary 33. *If $\mathrm{rank}(X) = n$, then*

$$\widehat{\beta}^{\dagger}_{H,b,X} = \widehat{\beta}_{H,b,0}, \qquad \bar{\lambda}^{\dagger}_{H,b,w,X}(\tau) = \bar{\lambda}_{H,b,w}(\tau) \qquad (0 < \tau < 1).$$

Proof. Full row rank gives $K_H^+ = Q_H$, hence

$$\widehat{\beta}^{\dagger}_{H,b,X} = b + H^{-1}X^\top Q_H \nu_b = \widehat{\beta}_{H,b,0}.$$

Injectivity of $X^\top$ gives, with $a = a_{H,b}$,

$$t\zeta_{H,b,X} \in X^\top \mathcal{B}_{\tau,w} \iff X^\top(ta) = X^\top g \text{ for some } g \in \mathcal{B}_{\tau,w} \iff ta \in \mathcal{B}_{\tau,w}.$$

The supremum over positive t satisfying the last condition is $\bar{\lambda}_{H,b,w}(\tau)$ by Theorem 14. This completes the proof. □

S.34 DUAL-POLYTOPE BOUNDARY COMPUTATION

This section computes the general-rank extinction boundary exactly by linear programming. In this section X may have arbitrary rank, $y \in \mathrm{range}(X)$, and $\widehat{\beta}_X^\dagger$ and $\bar{\lambda}_X^\dagger(\tau)$ are the quantities in Section S.33. Assume $\widehat{\beta}_X^\dagger \neq 0$.

Proposition 54. *For every $\tau \in (0,1)$,*

$$\bar{\lambda}_X^\dagger(\tau) = \max\{t \geq 0 : \exists g \in \mathbb{R}^n,\ X^\top g = t\widehat{\beta}_X^\dagger,\ \tau - 1 \leq g_i \leq \tau\ (i \leq n)\}.$$

The maximum is attained.

Proof. The constraint $\tau - 1 \leq g_i \leq \tau$ is the condition $g \in \mathcal{B}_\tau$. The equality constraint is

$$t\widehat{\beta}_X^\dagger \in X^\top \mathcal{B}_\tau.$$

Theorem 21 identifies the feasible values as $[0, \bar{\lambda}_X^\dagger(\tau)]$, with a finite positive upper bound when $\widehat{\beta}_X^\dagger \neq 0$. Thus the maximum is attained and equals the boundary. □

The following proposition gives the dual linear program for the unweighted arbitrary-rank boundary and proves attainment.

Proposition 55. *For every $\tau \in (0,1)$,*

$$\bar{\lambda}_X^\dagger(\tau) = \min_{\substack{v \in \mathbb{R}^p,\ \alpha,\gamma \in \mathbb{R}^n \\ Xv = \alpha - \gamma,\ \alpha \geq 0,\ \gamma \geq 0 \\ v^\top \widehat{\beta}_X^\dagger = 1}} \left\{\tau \mathbf{1}_n^\top \alpha + (1-\tau)\mathbf{1}_n^\top \gamma\right\}.$$

The minimum is attained.

Proof. For $s \in \mathbb{R}^n$,

$$H_\tau(s) = \min_{\alpha,\gamma \in \mathbb{R}^n : \alpha - \gamma = s,\ \alpha \geq 0,\ \gamma \geq 0} \{\tau \mathbf{1}_n^\top \alpha + (1-\tau)\mathbf{1}_n^\top \gamma\}.$$

The minimum is attained at $(\alpha_i, \gamma_i) = (s_i, 0)$ for $s_i \geq 0$ and $(0, -s_i)$ for $s_i < 0$.

Proposition 49 and positive homogeneity of H_τ give

$$\bar{\lambda}_X^\dagger(\tau) = \inf_{v \in \mathbb{R}^p :\ v^\top \widehat{\beta}_X^\dagger = 1} H_\tau(Xv).$$

Substituting the coordinatewise linear representation of H_τ gives the stated optimization problem.

For attainment, take a minimizing sequence (v_m) with $v_m^\top \widehat{\beta}_X^\dagger = 1$. Decompose $v_m = u_m + w_m$, where $u_m \in \mathrm{range}(X^\top)$ and $w_m \in \ker(X)$. Because $\widehat{\beta}_X^\dagger \in \mathrm{range}(X^\top)$,

$$u_m^\top \widehat{\beta}_X^\dagger = 1, \qquad Xu_m = Xv_m.$$

Let $c_\tau = \tau \wedge (1-\tau) > 0$. Then

$$H_\tau(Xu_m) \geq c_\tau \left\|Xu_m\right\|_1.$$

Thus (Xu_m) is bounded. The restriction of X to $\mathrm{range}(X^\top)$ has a bounded inverse, giving bounded (u_m). A convergent subsequence attains the infimum; choose α, γ coordinatewise from its limit Xu. □

The following corollary turns equality of feasible primal and dual objectives into a checkable exact boundary certificate.

Corollary 34. *Fix $\tau \in (0,1)$. Suppose $t \geq 0$, $g \in \mathbb{R}^n$, $v \in \mathbb{R}^p$, and $\alpha, \gamma \in \mathbb{R}^n$ satisfy*

$$X^\top g = t\widehat{\beta}_X^\dagger, \qquad \tau - 1 \leq g_i \leq \tau \quad (i \leq n),$$

$$Xv = \alpha - \gamma, \qquad \alpha \geq 0, \qquad \gamma \geq 0, \qquad v^\top \widehat{\beta}_X^\dagger = 1, \qquad t = \tau \mathbf{1}_n^\top \alpha + (1-\tau)\mathbf{1}_n^\top \gamma.$$

Then

$$t = \bar{\lambda}_X^\dagger(\tau).$$

Proof. For any primal feasible pair (t, g) and any dual feasible triple (v, α, γ),

$$t = tv^\top \widehat{\beta}_X^\dagger = v^\top X^\top g = g^\top(\alpha - \gamma).$$

Because $g_i \le \tau$ and $g_i \ge \tau - 1$,

$$g^\top(\alpha - \gamma) \le \tau \mathbf{1}_n^\top \alpha + (1 - \tau)\mathbf{1}_n^\top \gamma.$$

Equality identifies the common value as $\bar{\lambda}_X^\dagger(\tau)$ by Propositions 54 and 55. □

SUPPLEMENTARY REFERENCES

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002. URL `https://jmlr.org/papers/v3/bartlett02a.html`.

Kenneth R. Davidson and Stanislaw J. Szarek. Local operator theory, random matrices and Banach spaces. In William B. Johnson and Joram Lindenstrauss (eds.), *Handbook of the Geometry of Banach Spaces, Vol. I*, pp. 317–366. North-Holland, Amsterdam, 2001.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004. doi: 10.1214/009053604000000067.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022. doi: 10.1214/21-AOS2133.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. John Wiley and Sons, 1982. ISBN 9780471094425. doi: 10.1002/9780470316559.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pp. 1177–1184. Curran Associates, Inc., 2007.

Ingo Steinwart and Andreas Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 2011. doi: 10.3150/10-BEJ267.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, 2018.

Caixing Wang and Xingdong Feng. Optimal kernel quantile learning with random features. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50419–50452. PMLR, 2024.