DISTRIBUTIONALLY ROBUST OPTIMIZATION THROUGH RANDOM EMPIRICAL WEIGHTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose learning objectives that apply a risk functional to the random expected loss produced by a distribution over empirical reweightings. Under an unpenalized risk envelope representation, the objective equals a robust optimization problem over mixtures of the candidate measures. The construction inherits properties of the outer risk functional. A density specified through inverse KL divergence gives a Dirichlet family interpolating between the empirical expectation and the risk of a uniformly sampled observation. At the latter endpoint, the uniform vertex law is maximal in convex order among probability weight laws with the same mean. Gaussian perturbations continue beyond this boundary with weights that may be negative and still sum to one, allowing integrated losses outside the observed range for nonconstant losses. For a fixed outer risk respecting convex order, the induced risks are ordered along the path, with nested ambiguity sets under the envelope representation. For losses linear in the decision, a prescribed weight covariance recovers spherical linear shrinkage in a mean–standard deviation objective when the centered feature matrix has full column rank. For Dirichlet weights and conditional value-at-risk (CVaR), we develop a stochastic subgradient algorithm. On Mammographic Mass, selecting the weight distribution by cross-validation gives slightly higher mean AUC than ν-SVM at all five tested CVaR levels and lower mean classification error at the three higher levels.

1 INTRODUCTION

Empirical risk minimization evaluates a decision by averaging its losses on the observed sample. Distributionally robust optimization (DRO) instead minimizes the largest expected loss over a set of candidate distributions. An ambiguity set specifies which candidates are admissible. A probability law over those candidates can describe how uncertainty is distributed among them. We use such a law to construct a random expected loss, then apply a risk functional to that random variable. This separates the choice of how candidate distributions vary from the choice of how their expected losses enter the decision criterion.

Let $X_\theta(\xi)$ be the loss of decision θ on observation ξ. We condition on $n \geq 2$ indexed observations $\xi_1, \ldots, \xi_n$, with sample losses $x_i(\theta) = X_\theta(\xi_i)$ and empirical distribution $\widehat{P}_n = n^{-1} \sum_i \delta_{\xi_i}$. Our starting point is DRO with an ambiguity set defined by an f-divergence ball around $\widehat{P}_n$, restricting candidates to the same observations (Ben-Tal et al., 2013). Writing $\Delta_n = \{w \in \mathbb{R}^n_+ \mid \sum_i w_i = 1\}$ and $\mu_w = \sum_i w_i \delta_{\xi_i}$, the problem is

$$\min_\theta \sup_{\mu_w \in \mathcal{U}_\varepsilon} \mathbb{E}_{\mu_w}[X_\theta], \qquad \mathcal{U}_\varepsilon = \{\mu_w \mid w \in \Delta_n,\ D_f(\mu_w \| \widehat{P}_n) \leq \varepsilon\}. \tag{1}$$

Here D_f is the chosen divergence and $\varepsilon \geq 0$ its radius. A candidate distribution is therefore a reweighting of the observations. The divergence and radius determine which reweightings the adversary can use.

Instead of prescribing an ambiguity set, specify a probability law π on Δ_n and draw $W \sim \pi$. Each draw defines an empirical distribution μ_W. Its expected loss varies with W, even at a fixed decision. We consider

$$\min_\theta \ \rho(X_\theta), \qquad \rho(X_\theta) = \mathcal{R}(Y_\theta), \qquad Y_\theta = \mathbb{E}_{\mu_W}[X_\theta] = \sum_i W_i x_i(\theta). \tag{2}$$

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.

The inner expectation averages observations under one realized candidate. The *outer risk functional* $\mathcal{R}$ evaluates the resulting random loss under π. We specify π as part of the optimization model, rather than obtain it from a Bayesian posterior. Its specification can depend on the observed data.

The outer functional determines whether this randomization changes the learning problem. Let $u = \mathbf{1}/n$, where $\mathbf{1}$ is the vector of ones. If $\mathbb{E}W = u$, then $\mathbb{E}_\pi[Y_\theta] = \mathbb{E}_{\widehat{P}_n}[X_\theta]$. An outer expectation returns empirical risk minimization, while a nonlinear functional allows the variability of the candidate expected losses to matter. Conditional value-at-risk (CVaR) averages the upper tail of these losses (Rockafellar & Uryasev, 2000). The distribution π can consequently change the objective even when its mean weights remain fixed.

Our first result relates this construction back to DRO. Under an unpenalized envelope representation of $\mathcal{R}$, the objective equals the largest expected loss over an induced ambiguity set of mixtures. The construction changes how the set is specified while retaining a worst case interpretation. It also inherits properties of the outer functional, including convexity, positive homogeneity, and translation equivariance.

We then choose the distribution of weights in two ways. A density that depends on inverse KL divergence gives a Dirichlet family whose concentration interpolates between the empirical expectation and the outer risk of an empirical draw. At the latter endpoint, the weights follow the uniform vertex law, which is maximal in convex order among probability weights with mean u. Gaussian perturbations continue the family on the hyperplane where weights sum to one. They allow signed weights and, for nonconstant losses, integrated losses outside the observed range. The associated risks increase and their induced ambiguity sets expand under the stated order assumptions.

Prescribing the covariance of the weights gives a further way to choose the objective. With mean–standard deviation as the outer functional and losses linear in the decision, an appropriate covariance recovers spherical linear shrinkage of the empirical covariance. A bound for probability weights explains why this recovery can require signed weights at the specified mean and scale. Finally, induced maximizing weights supply subgradients for optimization. In classification, validation selects the weight law through one parameter. We compare the resulting classifier with empirical CVaR, which corresponds to a ν-SVM.

Related work. Smooth ambiguity and composite risk evaluate an outer functional on losses that depend on an uncertain model (Klibanoff et al., 2005; Dentcheva et al., 2017). Closely related formulations in optimization use a Bayesian posterior or a generalized Bayesian posterior over models (Wu et al., 2018; Bariletto & Ho, 2024; Javeed et al., 2023). Here the law is specified directly on empirical weights and can extend to signed candidate measures. Dirichlet weights also appear in the Bayesian bootstrap (Rubin, 1981). Their concentration order is established by Letac & Piccioni (2014). We use that result in a family that continues beyond the simplex and identify its induced ambiguity sets and maximizing weights. The covariance application recovers the spherical linear shrinkage form (Ledoit & Wolf, 2004) at a specified intensity. Risk envelopes and subgradients provide the established convex analysis used in the representation and algorithm (Artzner et al., 1999; Rockafellar & Wets, 1998).

2 From random weights to a robust objective

For a fixed loss, suppress θ and write $x_i = X(\xi_i)$ and $Y = \mathbb{E}_{\mu_W}[X]$. We assume the outer functional admits the representation

$$\mathcal{R}(Y) = \sup_{Q \in \mathcal{Q}} \mathbb{E}[QY], \qquad \mathbb{E}Q = 1. \tag{3}$$

Expectations without a subscript are over the specified law of the weights. The set $\mathcal{Q}$ is the *risk envelope*. For now, its elements are nonnegative, so each Q is a density that changes π to $Q\,d\pi$. The envelope determines the admissible changes in emphasis on candidates.

A change of law on candidates also determines weights on observations. Set $q_i = \mathbb{E}[QW_i]$ and $\overline{\mu}_Q = \sum_i q_i \delta_{\xi_i}$. Linearity gives

$$\mathbb{E}[Q\,\mathbb{E}_{\mu_W}[X]] = \sum_i \mathbb{E}[QW_i]x_i = \mathbb{E}_{\overline{\mu}_Q}[X]. \tag{4}$$

Under $Q\,d\pi$, the candidate W is random, whereas q is its mean. The measure $\overline{\mu}_Q$ thus averages the candidates using the probabilities assigned by the changed law. Its weight on observation i is the average of that observation's candidate weights. Maximizing over admissible densities produces a robust problem on the original sample.

Theorem 1 (Ambiguity set induced by probability mixtures)**.** *Suppose $W \in \Delta_n$ almost surely and* (3) *holds with integrable $Q \geq 0$. The induced ambiguity set and risk satisfy*

$$\mathcal{U} = \overline{\mathrm{conv}}\{\overline{\mu}_Q \mid Q \in \mathcal{Q}\}, \qquad \rho(X) = \sup_{\mu \in \mathcal{U}} \mathbb{E}_\mu[X]. \tag{5}$$

The representation holds for every sample loss vector, with closure in the finite vector of sample weights. Every member of $\mathcal{U}$ is a probability distribution on the observed sample.

The closed convex hull leaves the supremum unchanged. Once π and $\mathcal{R}$ are chosen, the ambiguity set is fixed. Its maximizing member can still depend on the loss. If Q^* attains the outer supremum, then $q^* = \mathbb{E}[Q^*W]$ gives the maximizing observation weights. Thus W describes one candidate, Q^* emphasizes candidates that are adverse for the current loss, and q^* combines them into a measure on observations. The resulting set in (5) is specified by these admissible mixtures, rather than by the divergence constraint in (1). Appendix A proves the representation and its extension to signed measures.

Inherited risk properties. The construction also gives a new risk functional with properties inherited from $\mathcal{R}$. The important structure is the linear map from sample losses to their weighted sum, together with the normalization of the weights. The following statement applies to any real weights whose sum is one, including the signed weights introduced in the next section. A functional is law invariant when it depends only on the distribution of its argument.

Theorem 2 (Inheritance of risk properties)**.** *Let $\sum_i W_i = 1$ almost surely, and suppose $\rho(X) = \mathcal{R}(\sum_i W_i X(\xi_i))$ is well defined and finite for every sample loss vector. Convexity, positive homogeneity, subadditivity, translation equivariance, and constant preservation each transfer from $\mathcal{R}$ to ρ. If $W \in \Delta_n$, monotonicity also transfers, and $\rho(X_\theta)$ is convex whenever $\mathcal{R}$ is convex and monotone and each $x_i(\theta)$ is convex. If the law of W is invariant under coordinate permutations and $\mathcal{R}$ is law invariant, then ρ is invariant under permutations of the sample losses.*

In particular, simplex weights preserve coherent risk measures (Artzner et al., 1999). Regular risk measures (Rockafellar & Uryasev, 2013) are also preserved when $\mathbb{E}W = u$ and the weight covariance is positive definite on the subspace of vectors whose entries sum to zero. This condition ensures that every nonconstant sample loss remains random after weighting. The proof and regularity argument are in Appendix A.

3 Interpolation and Extrapolation through the Weight Distribution

3.1 Designing the Distribution of Weights

The representation leaves us free to design π. One choice is to make its density depend on the divergence of a candidate from the empirical weights. A divergence ball uses a cutoff to declare candidates admissible. A density instead assigns relative probability throughout the simplex, with a parameter controlling how strongly it favors particular candidates.

Let U be the uniform probability measure on Δ_n. For $\beta \geq 0$, consider a density proportional to $\exp\{-\beta\,\mathrm{KL}(u\|w)\}$, where $\mathrm{KL}(u\|w) = \sum_i u_i \log(u_i/w_i)$ on the interior of the simplex. Candidates farther from u in this divergence receive smaller density. Increasing β strengthens this preference, while $\beta = 0$ gives U. The direction of KL is important because its logarithms give a product of powers of the coordinates.

Proposition 3 (A Dirichlet law from inverse KL divergence)**.** *For $\beta \in \mathbb{R}$, the density with respect to U given by*

$$\frac{d\pi_\beta}{dU}(w) \propto \exp\{-\beta\,\mathrm{KL}(u\|w)\} \quad \textit{defines} \quad \pi_\beta = \mathrm{Dir}\left(1 + \frac{\beta}{n}, \ldots, 1 + \frac{\beta}{n}\right) \tag{6}$$

if and only if $\beta > -n$. Its mean weights are u.

Indeed, the exponential is proportional to $\prod_i w_i^{\beta/n}$, which is the Dirichlet density in (6). The admissible negative range $-n < \beta < 0$ extends the design to densities that favor the boundary rather than the center. Writing $\tau = n + \beta$ therefore gives the full symmetric family $W_\tau \sim \mathrm{Dir}(\tau u)$ for $\tau > 0$. This is a choice of the law of the weights and is independent of the choice of outer risk. Appendix B gives the normalization calculation and the extension to a nonuniform reference.

3.2 Dirichlet interpolation

The total concentration τ controls how the candidate expected losses vary while preserving the mean weights. Define $Y_\tau = \mathbb{E}_{\mu_{W_\tau}}[X]$ and $\overline{X} = \mathbb{E}_{\widehat{P}_n}[X]$. Large concentration places weights near u, so the candidate expected losses are near $\overline{X}$. Small concentration places weights near vertices, where one observation receives all the mass. Let I be uniform on $\{1, \dots, n\}$ and e_i be the ith coordinate vector. Then x_I follows the empirical distribution of the sample losses.

Theorem 4 (Endpoints of the Dirichlet family)**.** *For every fixed sample loss,*

$$\begin{aligned} &W_\tau \Rightarrow e_I, \quad Y_\tau \Rightarrow x_I && (\tau \downarrow 0), \\ &W_\tau \to u, \quad Y_\tau \to \overline{X} \quad \textit{in } L^2 && (\tau \to \infty). \end{aligned} \tag{7}$$

If $\mathcal{R}$ preserves constants, is law invariant, and is continuous under weak convergence of uniformly bounded random variables, then

$$\lim_{\tau \to \infty} \mathcal{R}(Y_\tau) = \overline{X}, \qquad \lim_{\tau \downarrow 0} \mathcal{R}(Y_\tau) = \mathcal{R}(x_I). \tag{8}$$

The family connects averaging the observations with applying $\mathcal{R}$ directly to their empirical loss distribution. Between these endpoints, $\mathcal{R}$ evaluates the distribution of the weighted averages produced by W_τ, rather than a linear combination of the two endpoint objectives. We write $W_\infty = u$ and $W_0 = e_I$ for the limiting laws. The uniform law on the simplex occurs at $\tau = n$ and differs from both limits.

3.3 The simplex boundary and Gaussian continuation

The vertex law gives an upper bound on variability for every distribution of probability weights with mean u. We compare integrable scalar random variables by *convex order*, writing $A \preceq_{\mathrm{cx}} B$ when they have the same mean and $\mathbb{E}\psi(A) \le \mathbb{E}\psi(B)$ for every convex ψ for which the expectations exist.

Proposition 5 (Boundary for probability weights)**.** *If $W \in \Delta_n$ almost surely and $\mathbb{E}W = u$, then for every sample loss X and convex function $\psi : \mathbb{R} \to \mathbb{R}$,*

$$\psi(\overline{X}) \le \mathbb{E}[\psi(\mathbb{E}_{\mu_W}[X])] \le \frac{1}{n} \sum_i \psi(x_i). \tag{9}$$

Thus $\overline{X} \preceq_{\mathrm{cx}} \mathbb{E}_{\mu_W}[X] \preceq_{\mathrm{cx}} x_I$. The constant weights u attain the lower bound and the uniform vertex law attains the upper bound.

Jensen's inequality gives the lower bound by averaging over W and the upper bound by applying it to each candidate's weighted average. Thus changing to another law on the simplex cannot continue beyond the vertex endpoint in this order. We instead allow negative weights while retaining their sum of one.

The candidate weights now lie on $\mathcal{H} = \{w \in \mathbb{R}^n \mid \mathbf{1}^\top w = 1\}$. The law π remains a probability distribution, but a realization μ_W can be a signed measure. We retain $\mathbb{E}_{\mu_W}[X] = \sum_i W_i x_i$ as notation for signed integration. The general mixture representation in Appendix A, Theorem 9, extends the robust representation to signed measures under its integrability condition. The inheritance theorem preserves convexity in the sample losses. Consequently affine sample losses give a convex objective in the decision whenever $\mathcal{R}$ is convex.

Let $\mathsf{T} = \{z \mid \mathbf{1}^\top z = 0\}$ and let $P_{\mathsf{T}} = I_n - \mathbf{1}\mathbf{1}^\top/n$ be its orthogonal projector. The covariance of e_I is $\Sigma_u = P_{\mathsf{T}}/n$. We choose the Gaussian scale to continue the Dirichlet covariance formula $\mathrm{Cov}(W_\tau) = \Sigma_u/(1+\tau)$ to $-1 < \tau < 0$. Define

$$W_\tau = e_I + Z_\tau, \qquad Z_\tau \sim \mathcal{N}(0, S_\tau), \qquad S_\tau = -\frac{\tau}{1+\tau}\Sigma_u, \qquad Z_\tau \perp I. \tag{10}$$

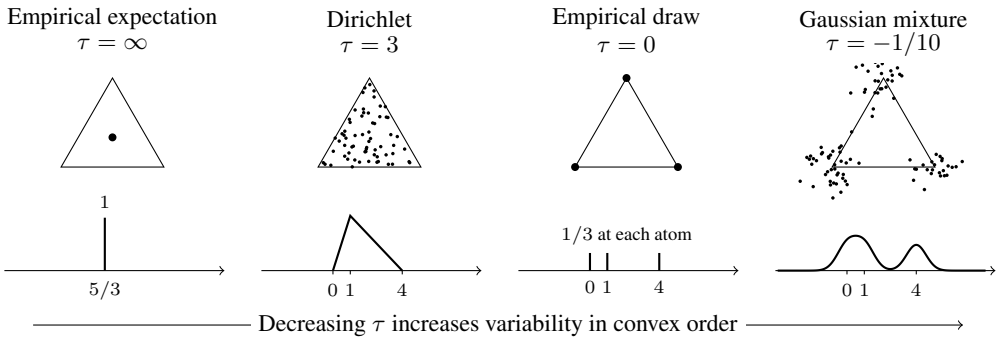


Figure 1: The weight distribution and the induced loss distribution along the path. The upper panels show illustrative draws on the plane $\sum_i W_i = 1$, with the simplex outlined, for $n = 3$. The lower panels show densities or point masses of Y_τ for losses $(0, 1, 4)$, all with mean $5/3$. The Dirichlet panel uses 64 uniform-simplex draws. The Gaussian panel uses 96 draws at $\tau = -1/10$, allowing signed weights and losses outside $[0, 4]$. Its unbounded support is shown through finite windows.

The perturbation lies in T, so its entries sum to zero and the total weight remains one. It is isotropic within this subspace, with equal variance in every unit direction. The resulting law is an equal mixture of Gaussians centered at the vertices. Its covariance within each component is S_τ, whereas the covariance of the whole mixture is $\Sigma_u + S_\tau = \Sigma_u/(1 + \tau)$.

For the integrated loss, write $\sigma_X^2 = n^{-1} \sum_i (x_i - \overline{X})^2$. The Gaussian construction gives

$$Y_\tau \overset{d}{=} x_I + \sqrt{-\frac{\tau}{1+\tau}}\, \sigma_X G, \qquad G \sim \mathcal{N}(0, 1), \qquad G \perp I. \tag{11}$$

Each observed loss is the center of a normal component whose variance is proportional to the empirical loss variance. Constant losses remain constant. For nonconstant losses, this Gaussian variation permits values outside the observed range while every candidate measure retains the same sample support. Appendix C, Proposition 10, gives the formal statement and proof.

3.4 One ordered family

The Dirichlet and Gaussian laws join at the vertex endpoint, giving

$$W_\infty = u, \qquad W_\tau \sim \mathrm{Dir}(\tau u)\ (\tau > 0), \qquad W_0 = e_I, \qquad W_\tau = e_I + Z_\tau\ (-1 < \tau < 0). \tag{12}$$

Throughout this family,

$$\mathbb{E} W_\tau = u, \qquad \mathrm{Cov}(W_\tau) = \frac{\Sigma_u}{1+\tau} \quad (-1 < \tau < \infty). \tag{13}$$

Figure 1 shows the weights and their integrated losses. The Dirichlet concentration order follows from Letac & Piccioni (2014, Theorem 1.1). Adding independent centered Gaussian noise continues that order beyond the vertex law. We state the comparison for the integrated losses, to which the outer risk is applied.

Theorem 6 (Order of the integrated losses and induced ambiguity sets)**.** *For every sample loss X and $-1 < \tau_2 < \tau_1 \le \infty$,*

$$Y_{\tau_1} \preceq_{\mathrm{cx}} Y_{\tau_2}. \tag{14}$$

If $\mathcal{R}$ preserves constants, is law invariant, is finite on these variables, and respects scalar convex order, then, with $\rho_\tau(X) = \mathcal{R}(Y_\tau)$,

$$\overline{X} = \rho_\infty(X) \le \rho_{\tau_1}(X) \le \rho_{\tau_2}(X). \tag{15}$$

If it also admits (3) *on each weight law, under the integrability condition of Theorem 9 in Appendix A, then*

$$\mathcal{U}_{\tau_1} \subseteq \mathcal{U}_{\tau_2}. \tag{16}$$

The induced ambiguity sets are nested as τ decreases. For nonnegative outer densities, they consist of probability mixtures when $\tau \geq 0$ and can contain signed measures when $\tau < 0$. CVaR and mean plus a nonnegative multiple of standard deviation both respect the required order. The risk inequality holds for every loss vector, which gives the set inclusion. Appendices B and C give the proofs, including the stronger vector comparison for the weights.

3.5 Induced weights for the Gaussian mixture

The robust representation lets us examine which observation weights the Gaussian construction produces in the worst case. A nonnegative density Q attaining the outer supremum at Y_τ is called a *risk identifier*. It changes both the probabilities of the mixture components and the average perturbation within those components. The resulting observation weights also form a subgradient with respect to the sample losses, supplying coefficients for optimization.

Theorem 7 (Induced weights for the Gaussian mixture)**.** *Fix $-1 < \tau < 0$ and a nonconstant sample loss X. Let $Q \geq 0$ be a risk identifier at Y_τ with $\mathbb{E}Q = 1$ and $\mathbb{E}[(1 + \|Z_\tau\|)Q] < \infty$. Set $x = (x_1, \dots, x_n)^\top$ and $p_i^Q = \mathbb{E}[Q\mathbf{1}_{\{I=i\}}]$. Then*

$$q = \mathbb{E}[QW_\tau] = p^Q + \mathbb{E}[QZ_\tau], \qquad p^Q \in \Delta_n, \quad \mathbb{E}[QZ_\tau] \in \mathsf{T}, \qquad q \in \partial_x \rho_\tau(X). \tag{17}$$

If $Q = h(Y_\tau)$, then

$$\begin{aligned} p_i^Q &= \frac{1}{n}\mathbb{E}[h(x_i + \sqrt{x^\top S_\tau x}\, G)], \\ q - p^Q &= \frac{x - \overline{X}\mathbf{1}}{n\sigma_X^2}\mathbb{E}[h(Y_\tau)(Y_\tau - x_I)], \qquad G \sim \mathcal{N}(0,1). \end{aligned} \tag{18}$$

For nondecreasing h, the scalar expectation in the second line is nonnegative.

Under the changed law $Q\, d\pi$, p_i^Q is the probability of selecting component i. Replacing every selected draw by its component center would produce the probability vector p^Q. The second term in (17) retains the mean perturbation. Although each Gaussian perturbation has mean zero under the original law, emphasizing draws with larger losses can change that mean. Its coordinates still sum to zero, so adding it to p^Q preserves the total weight. An observation receives a negative induced weight when its subtraction exceeds its component probability.

The condition $Q = h(Y_\tau)$ means that the change in emphasis depends on a candidate only through its integrated loss. If h is nondecreasing, candidates with larger losses receive greater emphasis, and (18) identifies the direction of the resulting displacement. It is parallel to $x - \overline{X}\mathbf{1}$, adding weight to losses above the empirical mean and subtracting from those below it. The covariance determines this direction, while the outer risk determines the component probabilities and the size of the displacement. Appendix C proves the result and gives the extension to signed identifiers.

For CVaR at level α, the changed law conditions the mixture on the upper fraction $1 - \alpha$ of its integrated loss distribution. Then p_i^Q is the proportion of selected draws from component i, and $\mathbb{E}[QZ_\tau]$ is their mean perturbation. The formulas in Appendix C express these quantities using normal tail probabilities and the mixture density.

4 Recovering covariance shrinkage

The preceding path changes the amount of variation. Choosing a general covariance of the weights also changes its directions. This distinction becomes explicit for the mean–standard deviation functional $\mathcal{R}_\kappa^{\mathrm{msd}}(Y) = \mathbb{E}Y + \kappa\, \mathrm{sd}(Y)$, with $\kappa \geq 0$. If $\mathbb{E}W = u$ and $S = \mathrm{Cov}(W)$, then

$$\rho^{\mathrm{msd}}(X) = \mathbb{E}_{\widehat{P}_n}[X] + \kappa\sqrt{x^\top S x}. \tag{19}$$

Thus, with the mean fixed, this objective depends on the law of W only through S. Choosing the covariance specifies how variation among the sample losses contributes to the risk. Here the outer envelope can contain signed elements Q. The mixture calculation remains valid, but $Q\, d\pi$ need not be a probability law. Appendix D gives the envelope representation and moment calculation.

Now take losses linear in the decision, $X_\theta(\xi) = a(\xi)^\top \theta$, where $a : \Xi \to \mathbb{R}^d$. Let A have rows $a_i^\top = a(\xi_i)^\top$, write $\bar{a} = n^{-1} \sum_i a_i$, and set

$$A_c = P_{\mathsf{T}} A, \qquad \widehat{C} = \frac{1}{n} A_c^\top A_c, \qquad \rho^{\mathrm{msd}}(X_\theta) = \bar{a}^\top \theta + \kappa \sqrt{\theta^\top A^\top S A \theta}. \tag{20}$$

The matrix A_c contains the centered observations, and $A^\top S A$ is the covariance of the integrated feature vector $\mathbb{E}_{\mu_W}[a]$. Vertex weights give $A^\top \Sigma_u A = \widehat{C}$, the empirical covariance. We ask whether another law can replace it by a structured estimate.

Empirical eigenvalues can have substantial estimation error. Spherical linear shrinkage balances this variation against bias by moving the eigenvalues toward their average while preserving the trace (Ledoit & Wolf, 2004). Its covariance is

$$C_\lambda = (1 - \lambda)\widehat{C} + \lambda \mu I_d, \qquad \mu = \frac{\mathrm{tr}(\widehat{C})}{d}, \qquad 0 \le \lambda \le 1. \tag{21}$$

We recover this form for a specified intensity λ by choosing S so that $A^\top S A = C_\lambda$.

Suppose A_c has full column rank, and let A_c^+ denote its Moore–Penrose pseudoinverse. Prescribe

$$S_\lambda = (1 - \lambda)\frac{P_{\mathsf{T}}}{n} + \lambda \mu (A_c^+)^\top A_c^+. \tag{22}$$

The first term retains a fraction $1 - \lambda$ of the vertex covariance, so it contributes $(1 - \lambda)\widehat{C}$ in feature space. The second maps an isotropic feature perturbation into the weights. For a feature perturbation $g \in \mathbb{R}^d$, the weight change $(A_c^+)^\top g$ has entries summing to zero and produces the feature change g, because $A_c^\top (A_c^+)^\top g = g$. Applying this map to a perturbation with covariance $\lambda \mu I_d$ gives the second term of S_λ. Thus the two terms produce the two parts of the desired shrinkage covariance.

Theorem 8 (Recovery of spherical shrinkage)**.** *Assume A_c has full column rank. The matrix S_λ in* (22) *is positive semidefinite, satisfies $S_\lambda \mathbf{1} = 0$, and obeys*

$$A^\top S_\lambda A = C_\lambda. \tag{23}$$

A law on $\mathcal{H}$ with mean u and covariance S_λ exists, and every such law yields

$$\rho_\lambda^{\mathrm{msd}}(X_\theta) = \bar{a}^\top \theta + \kappa \sqrt{\theta^\top C_\lambda \theta}. \tag{24}$$

For example, $\mathcal{N}(u, S_\lambda)$ has these moments. There also exists a different law for W_λ with the same mean and covariance whose integrated feature vector is

$$\mathbb{E}_{\mu_{W_\lambda}}[a] = \bar{a} + \sqrt{1 - \lambda}(a_I - \bar{a}) + \sqrt{\lambda \mu}\, G_d. \tag{25}$$

Here G_d is standard normal in $\mathbb{R}^d$ and independent of I. The empirical draw is contracted toward its mean, and independent isotropic variation supplies the spherical term. Appendix D, Theorem 12, constructs the corresponding weights. The rank assumption ensures that every feature direction can be obtained by changing the sample weights. With deficient rank, any $A^\top S A$ is confined to the directions spanned by the centered data, as detailed in (95) of that appendix.

The earlier simplex boundary also limits covariance. Probability weights with mean u satisfy

$$\mathrm{Cov}(W) \preceq \Sigma_u, \qquad \mathrm{Cov}(\mathbb{E}_{\mu_W}[a]) \preceq \widehat{C}. \tag{26}$$

This is the covariance consequence of the averaging restriction in Proposition 5. If $\widehat{C}$ has unequal eigenvalues and $\lambda > 0$, spherical shrinkage raises an eigenvalue below μ, violating the bound. A simplex law with mean u therefore cannot realize the specified C_λ at this scale, whereas the signed construction does. Rescaling the covariance and compensating through κ is a different comparison, discussed in Appendix D. Unlike the ordered path, shrinkage can increase some directional variances and decrease others.

5 Optimizing CVaR through the induced weights

The robust representation also supplies the coefficients needed for optimization. If Q^* attains the outer supremum at $Y = \mathbb{E}_{\mu_W}[X]$, the finiteness and integrability conditions in Appendix E give

$$q^* = \mathbb{E}[Q^* W] \in \partial_x \mathcal{R}(\mathbb{E}_{\mu_W}[X])\,. \tag{27}$$

Thus the maximizing sample weights form a subgradient with respect to the sample losses. This follows from the outer subgradient inequality and linearity (Rockafellar & Wets, 1998; Javeed et al., 2023). Computing Q^* still depends on the chosen risk. CVaR makes that choice explicit.

We use the convention that CVaR_α averages the upper fraction $1-\alpha$ of a loss distribution. For Dirichlet weights and nonconstant sample losses, $Y = \mathbb{E}_{\mu_{W_\tau}}[X_\theta]$ has a continuous law. At its α quantile C^*, $Q^* = \mathbf{1}_{\{Y>C^*\}}/(1-\alpha)$, so

$$q^* = \mathbb{E}[W_\tau \mid Y > C^*]. \tag{28}$$

The adversarial weights are the mean candidate weights among draws in the upper tail. To optimize the decision and threshold jointly, we use subgradients evaluated at their current values.

Let $x_i(\theta)$ and a regularizer $\Omega(\theta)$ be convex, and let Θ be a nonempty closed convex set. The Rockafellar–Uryasev representation (Rockafellar & Uryasev, 2000) gives

$$\min_{\theta\in\Theta,\, C\in\mathbb{R}} F(\theta, C), \qquad F(\theta, C) = C + \frac{\mathbb{E}[(Y-C)_+]}{1-\alpha} + \Omega(\theta), \qquad W_\tau \sim \mathrm{Dir}(\tau u). \tag{29}$$

At any current pair, choose $h_i \in \partial x_i(\theta)$ and $v \in \partial\Omega(\theta)$. The coefficients and joint subgradient are

$$q(\theta, C) = \frac{\mathbb{E}[W_\tau \mathbf{1}_{\{Y>C\}}]}{1-\alpha}, \qquad g_\theta = \sum_i q_i(\theta, C) h_i + v, \qquad g_C = 1 - \sum_i q_i(\theta, C). \tag{30}$$

For each realized weight vector, a subgradient of the weighted loss is the corresponding weighted sum of the h_i. The hinge selects draws above C. Since their weights sum to one, $\sum_i q_i = \Pr(Y > C)/(1-\alpha)$. At a general threshold, q need not be a probability vector. Its total mass tells the threshold update whether too many or too few draws lie above C.

Draw $W^{(1)}, \ldots, W^{(M)}$ independently from $\mathrm{Dir}(\tau u)$, set $Y_m = \mathbb{E}_{\mu_{W^{(m)}}}[X_\theta]$, and compute

$$\widehat{q} = \frac{1}{M(1-\alpha)} \sum_{m=1}^{M} W^{(m)} \mathbf{1}_{\{Y_m > C\}}, \qquad \widehat{g}_\theta = \sum_i \widehat{q}_i h_i + v, \qquad \widehat{g}_C = 1 - \sum_i \widehat{q}_i. \tag{31}$$

With fresh independent batches, these estimates are conditionally unbiased for $q(\theta, C)$ and the joint subgradient of F, under the assumptions of Proposition 16 in Appendix E. The expectation is over π with the observed sample fixed. Algorithm 1 uses the estimates to update the decision and threshold jointly, with Π_Θ denoting Euclidean projection onto Θ.

Algorithm 1 Stochastic subgradient method for Dirichlet CVaR

Require: $\theta_0 \in \Theta$, $\tau > 0$, $\alpha \in (0,1)$, batches M_0, M, steps η_t, threshold scale $s_C > 0$, iterations T, averaging start T_{av}

1: Draw M_0 independent weights from $\mathrm{Dir}(\tau u)$ and set C_0 to the empirical α quantile of their losses $\mathbb{E}_{\mu_W}[X_{\theta_0}]$.
2: **for** $t = 0, \ldots, T-1$ **do**
3: Evaluate $x_i(\theta_t)$, choose $h_i^t \in \partial x_i(\theta_t)$ and $v_t \in \partial\Omega(\theta_t)$.
4: Draw M fresh independent weights $W^{(m)} \sim \mathrm{Dir}(\tau u)$ and evaluate $Y_m = \mathbb{E}_{\mu_{W^{(m)}}}[X_{\theta_t}]$.
5: Compute $\widehat{q}_t$, $\widehat{g}_{\theta,t}$, and $\widehat{g}_{C,t}$ from the same batch using (31).
6: $\theta_{t+1} \leftarrow \Pi_\Theta(\theta_t - \eta_t \widehat{g}_{\theta,t})$, $\quad C_{t+1} \leftarrow C_t - \eta_t s_C \widehat{g}_{C,t}$.
7: **end for**
8: **return** the average of (θ_{t+1}, C_{t+1}) for $T_{\mathrm{av}} \le t < T$.

Appendix E gives tie rules, sampling safeguards, and convergence conditions. On the Gaussian segment, the scalar mixture in (11) instead gives the analytic threshold objective in (58) of Appendix C.

6 Classification with a selected weight law

The empirical CVaR endpoint connects the construction to a familiar classifier. For features a and label $y \in \{-1, 1\}$, use score $a^\top\theta + b$ and the loss $X_{\theta,b}(a, y) = -y(a^\top\theta + b)$. We fit

$$\min_{\theta,b} \quad \mathrm{CVaR}_\alpha\big(\mathbb{E}_{\mu_{W_\tau}}[X_{\theta,b}]\big) + \frac{\lambda_{\mathrm{reg}}}{2}\|\theta\|_2^2, \qquad \lambda_{\mathrm{reg}} = 0.05. \tag{32}$$

At $\tau = 0$, the risk is empirical CVaR. For nondegenerate solutions, this objective is equivalent up to positive scaling to a linear ν-SVM with $\nu = 1 - \alpha$, through the CVaR correspondence established by Gotoh & Takeda (2005). Appendix F.1 gives the scaling for our objective. Positive τ gives Dirichlet interpolation, and negative τ gives Gaussian extrapolation. The affine loss makes (32) convex on both segments. The scheme therefore adds τ as a hyperparameter. We tune it by cross-validation at fixed α and λ_{reg} and compare the resulting classifier with the empirical CVaR baseline.

We evaluate this procedure on Mammographic Mass, which has 961 observations, including 516 benign and 445 malignant cases (Elter, 2007). Age, shape, margin, and density are the predictors, with malignancy as the positive class. Imputation, scaling, and encoding are fitted within each training fold. Five stratified outer folds evaluate performance, and three stratified inner folds tune τ by mean balanced accuracy, weighting sensitivity and specificity equally. The grid contains six Dirichlet concentrations, empirical CVaR, and three Gaussian choices. We repeat the comparison separately at $\alpha \in \{0.10, 0.20, 0.30, 0.40, 0.50\}$ using the same outer partition.

The intercept is unpenalized. Dirichlet fits use Algorithm 1 with a fixed iteration budget. Empirical CVaR uses ν-SVM with radial scaling, and Gaussian fits use the analytic mixture objective. Tuning uses eligible configurations. The Gaussian and empirical solvers must terminate successfully and yield nondegenerate classifiers. Appendix F records the grid, feasibility checks, solver settings, and evaluation metrics. The selected model is fitted again on the complete outer training set before testing.

Table 1: Classification on Mammographic Mass. Entries are means $\pm$ sample standard deviations across five outer folds. Error and F1 are percentages. Inner cross-validation selects τ at each fixed α using mean balanced accuracy. The ν-SVM baseline fixes $\tau = 0$. Arrows indicate the preferred direction, and bold marks the better mean.

α	Method	Error ↓	AUC ↑	F1 ↑
0.10	ν-SVM	$\mathbf{20.08} \pm 2.28$	0.865 ± 0.011	79.03 ± 1.91
	CVaR with selected τ	20.19 ± 2.85	$\mathbf{0.866} \pm 0.011$	$\mathbf{79.22} \pm 2.52$
0.20	ν-SVM	$\mathbf{19.15} \pm 2.80$	0.866 ± 0.017	$\mathbf{80.68} \pm 2.39$
	CVaR with selected τ	19.35 ± 3.16	$\mathbf{0.867} \pm 0.015$	80.34 ± 2.85
0.30	ν-SVM	19.98 ± 2.42	0.863 ± 0.020	79.99 ± 2.06
	CVaR with selected τ	$\mathbf{19.15} \pm 3.27$	$\mathbf{0.868} \pm 0.016$	$\mathbf{80.55} \pm 2.99$
0.40	ν-SVM	20.08 ± 2.66	0.862 ± 0.021	$\mathbf{79.90} \pm 2.39$
	CVaR with selected τ	$\mathbf{19.67} \pm 3.05$	$\mathbf{0.866} \pm 0.018$	79.88 ± 2.68
0.50	ν-SVM	21.54 ± 1.75	0.862 ± 0.021	78.91 ± 1.49
	CVaR with selected τ	$\mathbf{19.56} \pm 2.93$	$\mathbf{0.869} \pm 0.016$	$\mathbf{80.00} \pm 2.51$

Table 1 compares CVaR with selected τ against the ν-SVM baseline. Mean AUC is slightly higher after selection at all five levels. Mean error is lower at $\alpha = 0.30$, 0.40, and 0.50, with the largest reduction at $\alpha = 0.50$, from 21.54% to 19.56%. At $\alpha = 0.10$ and 0.20, the baseline has slightly lower error. Mean F1 is higher after selection at three levels and lower at two. The classification benefit thus depends on the metric and fixed risk level. These results use one partition into five outer folds, with standard deviations summarizing variation across folds. Appendix F records the weight laws chosen by inner CV.

7 Conclusion

A distribution of empirical weights induces a robust objective through mixtures, with properties inherited from the outer risk. Designing the density through inverse KL divergence yields Dirichlet interpolation, while the simplex boundary motivates Gaussian continuation. Signed weights also permit recovery of shrinkage covariances that probability weights with the same mean cannot realize at the same scale. The induced maximizing weights supply subgradients for CVaR optimization. In classification, the resulting family contains an empirical ν-SVM formulation and adds one hyperparameter selected by cross-validation. On Mammographic Mass, selection gives slightly higher mean AUC at all five tested levels and lower mean classification error at the three higher levels, demonstrating the effectiveness of our approach.

8 AI DISCLOSURE

In this work, we used generative AI tools for: formulate mathematical claims, assist in the writing of proofs. We have not used generative AI tools for: generate synthetic data sets, help develop theoretical models or conceptual frameworks, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for: create or modify scientific figures or images, summarize or analyse existing literature, edit a research paper to improve readability. We have reviewed all AI-assisted work. We checked LLM-generated research ideas for potential plagiarism through a manual literature survey. LLM-generated code was verified and tested for correctness by the first author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999. doi: 10.1111/1467-9965.00068. URL https://doi.org/10.1111/1467-9965.00068.

Nicola Bariletto and Nhat Ho. Bayesian nonparametrics meets data-driven distributionally robust optimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 36438–36470. Curran Associates, Inc., 2024. doi: 10.52202/079017-1149. URL https://doi.org/10.52202/079017-1149.

Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013. doi: 10.1287/mnsc.1120.1641. URL https://doi.org/10.1287/mnsc.1120.1641.

Darinka Dentcheva, Spiridon Penev, and Andrzej Ruszczyński. Statistical estimation of composite risk functionals and risk optimization problems. *Annals of the Institute of Statistical Mathematics*, 69 (4):737–760, 2017. doi: 10.1007/s10463-016-0559-8. URL https://doi.org/10.1007/s10463-016-0559-8.

Matthias Elter. Mammographic Mass [dataset]. UCI Machine Learning Repository, 2007. URL https://doi.org/10.24432/C53K6Z.

Jun-ya Gotoh and Akiko Takeda. A linear classification model based on conditional geometric score. *Pacific Journal of Optimization*, 1(2):277–296, 2005. URL http://www.ybook.co.jp/online2/oppjo/vol1/p277.html.

Aurya S. Javeed, Drew P. Kouri, and Thomas M. Surowiec. A risk management perspective on statistical estimation and generalized variational inference. arXiv:2310.17376, 2023. URL https://arxiv.org/abs/2310.17376.

Peter Klibanoff, Massimo Marinacci, and Sujoy Mukerji. A smooth model of decision making under ambiguity. *Econometrica*, 73(6):1849–1892, 2005. doi: 10.1111/j.1468-0262.2005.00640.x. URL https://doi.org/10.1111/j.1468-0262.2005.00640.x.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4. URL https://doi.org/10.1016/S0047-259X(03)00096-4.

Gérard Letac and Mauro Piccioni. The Dirichlet curve of a probability in $\mathbb{R}^d$. arXiv:1405.4744, 2014. URL https://arxiv.org/abs/1405.4744.

Cheng Peng, Anton Malandii, and Stan Uryasev. Risk quadrangle and robust optimization based on extended φ-divergence. arXiv:2403.10987, 2024. URL https://arxiv.org/abs/2403.10987.

R. Tyrrell Rockafellar and Stan Uryasev. The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, 18 (1–2):33–53, 2013. doi: 10.1016/j.sorms.2013.03.001. URL https://doi.org/10.1016/j.sorms.2013.03.001.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2(3):21–41, 2000. doi: 10.21314/JOR.2000.038. URL https://doi.org/10.21314/JOR.2000.038.

R. Tyrrell Rockafellar and Roger J. B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 1998. ISBN 978-3-540-62772-2. doi: 10.1007/978-3-642-02431-3. URL https://doi.org/10.1007/978-3-642-02431-3.

Donald B. Rubin. The Bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981. doi: 10.1214/aos/1176345338. URL https://doi.org/10.1214/aos/1176345338.

Bernhard Schölkopf, Alex J. Smola, Robert C. Williamson, and Peter L. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000. doi: 10.1162/089976600300015565. URL https://doi.org/10.1162/089976600300015565.

Di Wu, Helin Zhu, and Enlu Zhou. A Bayesian risk approach to data-driven stochastic optimization: Formulations and asymptotics. *SIAM Journal on Optimization*, 28(2):1588–1612, 2018. doi: 10.1137/16M1101933. URL https://doi.org/10.1137/16M1101933.

A MIXTURE REPRESENTATION AND INHERITED RISK PROPERTIES

A.1 FINITE REFERENCE SPACE AND RISK CONVENTIONS

We regard observations as indexed outcomes, including when recorded values coincide. For a sample loss $X \in L^\infty(\widehat{P}_n)$, integration against μ_W is the finite sum

$$\mathbb{E}_{\mu_W}[X] = \sum_{i=1}^{n} W_i X(\xi_i). \tag{33}$$

It is independent of the representative modulo $\widehat{P}_n$ null sets and is measurable in W. Signed integration uses the same notation. Absolute integrability against each envelope element is specified in the following theorem.

We use the loss convention. Monotonicity means $Y \le Z \Rightarrow \mathcal{R}(Y) \le \mathcal{R}(Z)$, translation equivariance means $\mathcal{R}(Y + c) = \mathcal{R}(Y) + c$, and positive homogeneity means $\mathcal{R}(aY) = a\mathcal{R}(Y)$ for $a \ge 0$. Law invariance means dependence only on the law of the argument, and preservation of constants means $\mathcal{R}(c) = c$.

For comparison with the starting DRO formulation, let $P_w = \sum_i w_i \delta_{\xi_i}$ and let $p \in \operatorname{int} \Delta_n$. If f is lower semicontinuous and convex on $[0, \infty)$ with $f(1) = 0$, then

$$D_f(P_w \| P_p) = \sum_i p_i f(w_i/p_i), \qquad \min_\theta \sup_{w \in \Delta_n \,:\, D_f(P_w \| P_p) \le \varepsilon} \mathbb{E}_{P_w}[X_\theta]. \tag{34}$$

Here w_i/p_i is the density on the indexed sample. The choices $f(t) = t \log t - t + 1$ and $f(t) = (t-1)^2$ give KL and Pearson divergence, respectively. This supremum is a risk envelope on the observation space. In (3), the envelope instead acts on the law of W.

Theorem 9 (Envelope induced by mixtures)**.** *Let $W \sim \pi$ take values in $\mathcal{H}$, and suppose that $\mathcal{R}$ has representation* (3)*, with $\mathbb{E}[|Q| \sum_i |W_i|] < \infty$ for each $Q \in \mathcal{Q}$. Then $\overline{\mu}_Q = \sum_i q_i \delta_{\xi_i}$, where $q_i = \mathbb{E}[QW_i]$ and $\sum_i q_i = 1$. With closure in the finite vector of sample weights,*

$$\mathcal{U} = \overline{\operatorname{conv}}\{\overline{\mu}_Q : Q \in \mathcal{Q}\}, \qquad \rho(X) = \sup_{\mu \in \mathcal{U}} \mathbb{E}_\mu[X]. \tag{35}$$

If $W \in \Delta_n$ and every Q is nonnegative, $\mathcal{U}$ is an ambiguity set of probability distributions, obtained as mixtures of the candidate empirical distributions and limits of such mixtures.

A.2 PROOF OF THEOREM 9

The absolute integrability assumption makes $q_i = \mathbb{E}[QW_i]$ finite. Since $\sum_i W_i = 1$ and $\mathbb{E}Q = 1$, finite summation gives $\sum_i q_i = \mathbb{E}[Q \sum_i W_i] = 1$. Thus $\overline{\mu}_Q = \sum_i q_i \delta_{\xi_i}$ is a finite signed measure of total mass one. For any sample loss X,

$$\mathbb{E}[Q\, \mathbb{E}_{\mu_W}[X]] = \sum_i \mathbb{E}[QW_i] X(\xi_i) = \mathbb{E}_{\overline{\mu}_Q}[X]. \tag{36}$$

Taking the supremum over $Q \in \mathcal{Q}$ gives (5). The map $\mu \mapsto \mathbb{E}_\mu[X]$ is linear and continuous in the finite vector of sample weights. Its supremum is therefore unchanged by taking the closed convex hull.

For $Q \ge 0$, the measure $\overline{\mu}_Q$ averages the candidates μ_w under the probability law $Q(w)\pi(dw)$. If $W \in \Delta_n$ as well, each q_i is nonnegative, so this average is a probability mixture. Convex combinations and limits preserve that property. The signed and nonnegative cases consequently have the same algebraic representation, with the interpretation determined by their signs.

A.3 AN ILLUSTRATION WITH TWO CANDIDATE DISTRIBUTIONS

For a concrete illustration, take sample losses $(0, 1, 4)$ and a law that assigns equal probability to $(1/2, 1/3, 1/6)$ and $(1/6, 1/3, 1/2)$. The mean weights are u, and the candidate expected losses are 1 and $7/3$. Their expectation is the empirical mean $5/3$, whereas CVaR over the upper half equals $7/3$. Its maximizing density selects the second candidate, so the induced observation weights are $(1/6, 1/3, 1/2)$.

A.4 Proof of Theorem 2

For $t \in [0,1]$, linearity gives

$$\mathbb{E}_{\mu_W}[tX + (1-t)\widetilde{X}] = t\mathbb{E}_{\mu_W}[X] + (1-t)\mathbb{E}_{\mu_W}[\widetilde{X}].$$

Applying convexity of $\mathcal{R}$ gives convexity of ρ. The same linearity transfers positive homogeneity and subadditivity. The normalization of W gives $\mathbb{E}_{\mu_W}[X + c] = \mathbb{E}_{\mu_W}[X] + c$ and $\mathbb{E}_{\mu_W}[c] = c$, which transfer translation equivariance and constant preservation.

For $W \in \Delta_n$, nonnegative weights preserve $X \le \widetilde{X}$ on the sample under integration. Monotonicity of $\mathcal{R}$ therefore transfers to ρ in this case. Convex sample losses satisfy $X_{t\theta+(1-t)\widetilde{\theta}} \le tX_\theta + (1-t)X_{\widetilde{\theta}}$ at every observation. Monotonicity followed by convexity of ρ proves convexity in θ.

For a permutation σ, define $X^\sigma(\xi_i) = X(\xi_{\sigma(i)})$. If the law of W is invariant under coordinate permutations, then $\mathbb{E}_{\mu_W}[X^\sigma]$ and $\mathbb{E}_{\mu_W}[X]$ have the same distribution. Law invariance of $\mathcal{R}$ gives $\rho(X^\sigma) = \rho(X)$.

Coherence and regularity. The simplex case preserves the properties of coherent risk measures (Artzner et al., 1999). It also gives the usual continuity bound: with $d = \|X - \widetilde{X}\|_{L^\infty(\widehat{P}_n)}$, monotonicity and translation equivariance applied to $\widetilde{X} - d \le X \le \widetilde{X} + d$ yield $|\rho(X) - \rho(\widetilde{X})| \le d$.

For regularity in the sense of Rockafellar & Uryasev (2013), suppose that $\mathbb{E}W = u$ and $S = \mathrm{Cov}(W)$ is positive definite on $\mathsf{T} = \{z : \sum_i z_i = 0\}$. The map $x \mapsto \sum_i W_i x_i$ is continuous from $\mathbb{R}^n$ to $L^2(\pi)$, since its L^2 norm is bounded by $\|x\|_2(\mathbb{E}\|W\|_2^2)^{1/2}$. Composition therefore preserves lower semicontinuity, as well as the convexity and constant preservation already proved. For a nonconstant sample loss, its centered vector $x - \overline{X}\mathbf{1}$ is a nonzero element of T. Since $S\mathbf{1} = 0$,

$$\mathrm{Var}(\mathbb{E}_{\mu_W}[X]) = (x - \overline{X}\mathbf{1})^\top S(x - \overline{X}\mathbf{1}) > 0. \tag{37}$$

The random signed expectation is consequently nonconstant. Risk aversion of the outer regular risk gives $\rho(X) > \mathbb{E}[\mathbb{E}_{\mu_W}[X]] = \mathbb{E}_{\widehat{P}_n}[X]$. This proves regularity of ρ on the finite reference space. For the families in the main text, $S = \Sigma_u/(1+\tau)$ is positive definite on T at every finite $\tau > -1$.

B Proofs for the Dirichlet path

We prove the endpoint and contraction statements for an arbitrary $p \in \operatorname{int}\Delta_n$. The results in the main text follow by taking $p = u$. Write $W_\tau^{(p)} \sim \mathrm{Dir}(\tau p)$ and let $\Pr(I_p = i) = p_i$.

B.1 Proof of Proposition 3

For $u = \mathbf{1}/n$, the identity $\mathrm{KL}(u\|w) = n^{-1}\sum_i \log((1/n)/w_i)$ gives $\exp\{-\beta\,\mathrm{KL}(u\|w)\} \propto \prod_i w_i^{\beta/n}$. Because the uniform measure has constant density on the simplex, this is a Dirichlet density with each parameter equal to $1 + \beta/n$. Its integral is finite exactly when every parameter is positive, that is, $\beta > -n$. At $\beta = 0$ it is the uniform law. Values of the density on the boundary can be assigned arbitrarily, since that boundary has uniform measure zero.

For a nonuniform $p \in \operatorname{int}\Delta_n$, the same calculation gives parameters $\mathbf{1} + \beta p$, with normalizability condition $1 + \beta p_i > 0$ for all i, equivalently $\beta > -1/\max_i p_i$. The mean is $(\mathbf{1} + \beta p)/(n + \beta)$. For $\beta \ge 0$ it lies between u and p, approaching p as $\beta \to \infty$. For $p = u$ the mean stays fixed over the full admissible range.

B.2 Proof of Theorem 4

For nonnegative integers $k_1, \dots, k_n$, set $k_0 = \sum_i k_i$. The Dirichlet moment formula is

$$\mathbb{E}\prod_{i=1}^n \big(W_{\tau,i}^{(p)}\big)^{k_i} = \frac{\prod_{i=1}^n (\tau p_i)_{k_i}}{(\tau)_{k_0}}, \tag{38}$$

where $(a)_k = \Gamma(a+k)/\Gamma(a)$. For $k \geq 1$, $(\tau c)_k = \tau c(k-1)! + O(\tau^2)$ as $\tau \downarrow 0$. Hence the moment in (38) converges to p_j if $k_j > 0$ is the only positive exponent, and to zero if at least two exponents are positive. These are exactly the monomial moments of e_{I_p}. Polynomials are dense in the continuous functions on the compact simplex, so

$$W_\tau^{(p)} \Rightarrow e_{I_p} \qquad (\tau \downarrow 0). \tag{39}$$

The Dirichlet mean and covariance are

$$\mathbb{E}W_\tau^{(p)} = p, \qquad \mathrm{Cov}(W_\tau^{(p)}) = \frac{\mathrm{diag}(p) - pp^\top}{\tau + 1}. \tag{40}$$

Therefore

$$\mathbb{E}\|W_\tau^{(p)} - p\|_2^2 = \mathrm{tr}\Big(\mathrm{Cov}(W_\tau^{(p)})\Big) \longrightarrow 0 \qquad (\tau \to \infty). \tag{41}$$

The affine map $w \mapsto \mathbb{E}_{\mu_w}[X]$ carries these limits to x_{I_p} and $\mathbb{E}_{\mu_p}[X]$. All scalar losses remain in the compact interval $[\min_i x_i, \max_i x_i]$. Law invariance, weak continuity on uniformly bounded variables, and constant preservation therefore yield the two risk limits. Setting $p = u$ proves Theorem 4.

B.3 Dirichlet segment of Theorem 6

We apply the Dirichlet-mean theorem of Letac & Piccioni (2014, Theorem 1.1). For a probability measure ν on $\mathbb{R}^n$ with finite first moment, let P_t be a Dirichlet random probability measure with parameter $t\nu$. Their theorem states that the random mean $\int x\, P_t(dx)$ decreases in convex order as $t > 0$ increases.

Take $\nu = \sum_i p_i \delta_{e_i}$. The masses of P_t on these vertices have distribution $\mathrm{Dir}(tp)$, so its random mean has distribution $W_t^{(p)}$. The finite support satisfies the moment assumption. Therefore

$$W_t^{(p)} \preceq_{\mathrm{cx}} W_s^{(p)} \qquad (0 < s < t). \tag{42}$$

This is also the simplex example in Section 2 of Letac & Piccioni (2014).

The two endpoints follow directly from Jensen's inequality. For every convex Φ,

$$\Phi(p) = \Phi(\mathbb{E}W_\tau^{(p)}) \leq \mathbb{E}\Phi(W_\tau^{(p)}) \leq \sum_i p_i \Phi(e_i) = \mathbb{E}\Phi(e_{I_p}). \tag{43}$$

The second inequality uses $w = \sum_i w_i e_i$ and $\mathbb{E}W_\tau^{(p)} = p$. Setting $p = u$ gives the stronger vector comparison

$$W_{\tau_1} \preceq_{\mathrm{cx}} W_{\tau_2} \qquad (0 \leq \tau_2 < \tau_1 \leq \infty). \tag{44}$$

Applying it to the convex function $w \mapsto \psi(\mathbb{E}_{\mu_w}[X])$ gives the corresponding scalar order. The support-function argument in the next appendix then proves induced-envelope inclusion.

B.4 Proof of Proposition 5

For $Y = \sum_i W_i x_i$, the mean condition gives $\mathbb{E}Y = \overline{X}$. Jensen's inequality gives $\psi(\overline{X}) \leq \mathbb{E}\psi(Y)$. Applying the same inequality to each realized convex combination gives $\psi(\sum_i W_i x_i) \leq \sum_i W_i \psi(x_i)$. Taking expectations yields the upper bound in (9). The constant weights u and the vertex law e_I attain the corresponding bounds.

C Proofs for the hyperplane extension

Proposition 10 (Gaussian loss distribution)**.** *For $-1 < \tau < 0$, the weights in* (10) *belong to $\mathcal{H}$, with mean u and covariance $\Sigma_u/(1+\tau)$. For any sample loss X, the random variable $Y_\tau = \mathbb{E}_{\mu_{W_\tau}}[X]$ has the representation in distribution*

$$Y_\tau \overset{d}{=} x_I + \sqrt{-\frac{\tau}{1+\tau}}\,\sigma_X G, \qquad G \sim \mathcal{N}(0,1), \qquad G \perp I. \tag{45}$$

Thus it is an equal mixture of normal distributions with means x_i and common variance $-\tau\sigma_X^2/(1+\tau)$.

C.1 Proof of Proposition 10

The Gaussian covariance satisfies $S_\tau \mathbf{1} = 0$, so $\sum_i Z_{\tau,i} = 0$ almost surely. Hence $W_\tau \in \mathcal{H}$. Independence and centering give $\mathbb{E}W_\tau = u$ and $\mathrm{Cov}(W_\tau) = \Sigma_u + S_\tau = \Sigma_u/(1+\tau)$. Conditional on $I = i$, the signed expectation is $Y_\tau = x_i + \sum_j Z_{\tau,j} x_j$. Its Gaussian term has mean zero and variance

$$x^\top S_\tau x = -\frac{\tau}{1+\tau}\frac{1}{n}\sum_i (x_i - \overline{X})^2 = -\frac{\tau}{1+\tau}\sigma_X^2.$$

This term is independent of I, proving (11) and the mixture description.

C.2 Proof of Theorem 6

For the induced measures, write $\mathcal{Q}_\tau$ for the outer envelope on the law of W_τ and assume at each parameter under consideration that

$$\mathbb{E}\left[|Q| \sum_i |W_{\tau,i}|\right] < \infty \qquad \text{for every } Q \in \mathcal{Q}_\tau. \tag{46}$$

This makes the coefficients of each $\overline{\mu}_{\tau,Q}$ finite and justifies the signed-mixture identity (36).

Couple the Gaussian segment using a centered Gaussian process $(B_t)_{t\geq 0}$ on T, independent of $V = e_I$, with independent increments and $B_t \sim \mathcal{N}(0, t\Sigma_u)$. Then

$$W_\tau = V + B_{-\tau/(1+\tau)} \qquad (-1 < \tau < 0). \tag{47}$$

Independence gives

$$\mathrm{Cov}(W_\tau) = \Sigma_u - \frac{\tau}{1+\tau}\Sigma_u = \frac{\Sigma_u}{1+\tau}.$$

Together with the Dirichlet and vertex covariances, this proves (13).

For $0 \leq \tau_2 < \tau_1 \leq \infty$, Appendix B gives $W_{\tau_1} \preceq_{\mathrm{cx}} W_{\tau_2}$. For $-1 < \tau_2 < \tau_1 \leq 0$, independent Gaussian increments give

$$W_{\tau_2} = W_{\tau_1} + B_{-\tau_2/(1+\tau_2)} - B_{-\tau_1/(1+\tau_1)}. \tag{48}$$

The increment is independent of W_{τ_1} and centered. Therefore $\mathbb{E}[W_{\tau_2} \mid W_{\tau_1}] = W_{\tau_1}$, and conditional Jensen gives the convex order, including the case $\tau_1 = 0$. For a pair across the two segments, transitivity gives $W_{\tau_1} \preceq_{\mathrm{cx}} V \preceq_{\mathrm{cx}} W_{\tau_2}$. Applying the vector order to the convex function $w \mapsto \psi(\mathbb{E}_{\mu_w}[X])$ proves the scalar order in (14).

Monotonicity of $\mathcal{R}$ under scalar convex order then gives $\rho_{\tau_1}(X) \leq \rho_{\tau_2}(X)$. At $\tau = \infty$, $\mu_{W_\infty} = \widehat{P}_n$, so constant preservation gives $\rho_\infty(X) = \mathbb{E}_{\widehat{P}_n}[X]$.

Under (3), the mixture calculation gives

$$\rho_\tau(X) = h_{\mathcal{U}_\tau}(X), \qquad h_K(X) = \sup_{\mu \in K} \mathbb{E}_\mu[X]. \tag{49}$$

On the finite reference sample, this is the support function of the set of measure coefficients. The risk ordering compares these support functions for every observed loss vector. The characterization of inclusion for nonempty closed convex sets therefore gives $\mathcal{U}_{\tau_1} \subseteq \mathcal{U}_{\tau_2}$ (Rockafellar & Wets, 1998). For nonnegative Q and $\tau \geq 0$, the induced coefficients are nonnegative. In every case their sum is one.

For the examples used in the main text, scalar convex order directly implies the required risk comparison. If $Y \preceq_{\mathrm{cx}} \widetilde{Y}$, then $\mathbb{E}[(Y-t)_+] \leq \mathbb{E}[(\widetilde{Y}-t)_+]$ for every t. Taking infima in the CVaR threshold representation preserves the inequality. If both variables have finite second moments, their equal means and the convex function $y \mapsto y^2$ give $\mathrm{Var}(Y) \leq \mathrm{Var}(\widetilde{Y})$. Consequently mean plus κ times standard deviation respects this order for every $\kappa \geq 0$.

The scalar convex-order assumption has a separate role from homogeneity, constant preservation, and law invariance. For example, with $\gamma > 0$, the functional $\mathcal{R}(Y) = \mathbb{E}Y - \gamma\,\mathrm{sd}(Y)$ has these last three properties. Along the family it equals

$$\mathcal{R}(Y_\tau) = \mathbb{E}_{\widehat{P}_n}[X] - \frac{\gamma\sigma_X}{\sqrt{1+\tau}},$$

which decreases as τ decreases when $\sigma_X > 0$.

C.3 Regularity conditions and proof of Theorem 7

Assume

$$\mathbb{E}[(1 + \|Z_\tau\|)|Q|] < \infty. \tag{50}$$

Since $W_\tau = e_I + Z_\tau$, $q = \mathbb{E}[Qe_I] + \mathbb{E}[QZ_\tau]$. For $Q \geq 0$ with $\mathbb{E}Q = 1$, $p^Q = \mathbb{E}[Qe_I]$ is a probability vector. Also $\sum_i Z_{\tau,i} = 0$, so the displacement $\mathbb{E}[QZ_\tau]$ has zero sum. The joint law on (I, Z_τ) with density Q assigns probability p_i^Q to $I = i$.

For a second loss $\widetilde{X}$ with values $\widetilde{x}_i$, the outer subgradient inequality gives

$$\rho_\tau(\widetilde{X}) \geq \rho_\tau(X) + \mathbb{E}\Big[Q\mathbb{E}_{\mu_{W_\tau}}[\widetilde{X} - X]\Big] \tag{51}$$

$$= \rho_\tau(X) + \sum_i q_i(\widetilde{x}_i - x_i). \tag{52}$$

This is the subgradient inequality in the observed values and proves the last statement in (17).

If $Q = h(Y_\tau)$, conditional Gaussian regression gives

$$\mathbb{E}[Z_\tau \mid I, Y_\tau] = \frac{S_\tau x}{x^\top S_\tau x}(Y_\tau - x_I) = \frac{x - \overline{X}\mathbf{1}}{n\sigma_X^2}(Y_\tau - x_I). \tag{53}$$

The second equality uses $x^\top S_\tau x = -\tau\sigma_X^2/(1+\tau)$ and $S_\tau x = -\tau(x - \overline{X}\mathbf{1})/[n(1+\tau)]$. Multiplying by $h(Y_\tau)$ and taking expectations gives the displacement in (18). Conditional on $I = i$, the residual $Y_\tau - x_i$ is centered normal with variance $x^\top S_\tau x$, which gives the formula for p_i^Q.

For the smooth formula, assume in addition that h is locally absolutely continuous and $\mathbb{E}|h'(Y_\tau)| < \infty$. Together with (50), these conditions allow Gaussian integration by parts in every component:

$$\mathbb{E}[Z_\tau h(Y_\tau)] = S_\tau x\, \mathbb{E}[h'(Y_\tau)]. \tag{54}$$

For any nondecreasing h, the coefficient in the regression formula is nonnegative. Conditional on $I = i$, let $\zeta = Y_\tau - x_i$ and let ζ' be an independent copy. Since $\mathbb{E}[\zeta \mid I = i] = 0$,

$$\mathbb{E}[\zeta h(x_i + \zeta) \mid I = i] = \frac{1}{2}\mathbb{E}[(\zeta - \zeta')(h(x_i + \zeta) - h(x_i + \zeta')) \mid I = i] \geq 0.$$

Averaging over i gives $\mathbb{E}[h(Y_\tau)(Y_\tau - x_I)] \geq 0$.

The identities also hold for signed identifiers with $\mathbb{E}Q = 1$ and the same integrability conditions. Then p^Q has total mass one and may have signed entries. The mean–standard deviation example below uses this extension.

C.4 Homogeneous risks on the Gaussian segment

For mean–standard deviation,

$$\mathcal{R}_\kappa^{\mathrm{msd}}(Y_\tau) = \mathbb{E}_{\widehat{P}_n}[X] + \frac{\kappa\sigma_X}{\sqrt{1+\tau}}. \tag{55}$$

For nonconstant X and finite $\tau > -1$, an identifier and its induced coefficients are

$$Q_{\tau,\kappa}^* = 1 + \kappa\sqrt{1+\tau}\frac{Y_\tau - \overline{X}}{\sigma_X}, \qquad q_{\tau,\kappa}^*(X) = u + \frac{\kappa\Sigma_u x}{\sqrt{1+\tau}\,\sigma_X}. \tag{56}$$

For CVaR, let Φ and φ denote the standard normal distribution function and density. Put

$$\delta_\tau^2 = x^\top S_\tau x = -\frac{\tau}{1+\tau}\sigma_X^2. \tag{57}$$

For $-1 < \tau < 0$ and nonconstant X, the threshold representation is

$$\mathrm{CVaR}_\alpha(Y_\tau) = \min_{t\in\mathbb{R}}\left\{t + \frac{1}{n(1-\alpha)}\sum_i\left[(x_i - t)\Phi\left(\frac{x_i - t}{\delta_\tau}\right) + \delta_\tau\varphi\left(\frac{x_i - t}{\delta_\tau}\right)\right]\right\}. \tag{58}$$

Let $t_{\alpha,\tau}$ be the minimizing quantile, set $z_{\tau,i} = (t_{\alpha,\tau} - x_i)/\delta_\tau$ and $\bar{\Phi} = 1 - \Phi$, and write the mixture density as

$$f_{Y_\tau}(t) = \frac{1}{n\delta_\tau} \sum_i \varphi\left(\frac{t - x_i}{\delta_\tau}\right). \tag{59}$$

The induced coefficients are

$$\frac{1}{n(1-\alpha)} \begin{pmatrix} \bar{\Phi}(z_{\tau,1}) \\ \vdots \\ \bar{\Phi}(z_{\tau,n}) \end{pmatrix} + \frac{f_{Y_\tau}(t_{\alpha,\tau})}{1-\alpha} S_\tau x. \tag{60}$$

The first vector sums to one because the tail probability is $1 - \alpha$. The Gaussian term follows from

$$\mathbb{E}[Z_\tau \mathbf{1}_{\{Y_\tau > t_{\alpha,\tau}\}} \mid I = i] = \frac{S_\tau x}{\delta_\tau} \varphi(z_{\tau,i}).$$

D Proofs and additional results for covariance shrinkage

Proposition 11 (Dependence on the covariance of the weights)**.** *Let W take values in $\mathcal{H}$, with $\mathbb{E}W = u$ and covariance S. For any loss X, write $x = (X(\xi_1), \ldots, X(\xi_n))^\top$. Then*

$$\rho^{\mathrm{msd}}(X) = \mathcal{R}_\kappa^{\mathrm{msd}}(\mathbb{E}_{\mu_W}[X]) = \mathbb{E}_{\widehat{P}_n}[X] + \kappa\sqrt{x^\top S x}. \tag{61}$$

Thus any two weight laws with mean u and covariance S give the same mean–standard deviation objective.

D.1 Proof of Proposition 11

For $Y = \mathbb{E}_{\mu_W}[X] = \sum_i W_i x_i$, finite summation and the assumed moments give $\mathbb{E}Y = \sum_i u_i x_i = \mathbb{E}_{\widehat{P}_n}[X]$ and $\mathrm{Var}(Y) = x^\top S x$. Substitution into $\mathcal{R}_\kappa^{\mathrm{msd}}(Y) = \mathbb{E}Y + \kappa\,\mathrm{sd}(Y)$ proves (19). The expression is the same for every law with these moments.

D.2 Proof of Theorem 8

Both terms in (22) are positive semidefinite. Full column rank gives $A_c^+ = (A_c^\top A_c)^{-1} A_c^\top$ and $A_c^+ \mathbf{1} = 0$. Consequently $S_\lambda \mathbf{1} = 0$. Since $A = A_c + \mathbf{1}\bar{a}^\top$,

$$A^\top S_\lambda A = A_c^\top S_\lambda A_c \tag{62}$$

$$= (1-\lambda)\frac{1}{n} A_c^\top A_c + \lambda\mu A_c^\top (A_c^+)^\top A_c^+ A_c \tag{63}$$

$$= (1-\lambda)\widehat{C} + \lambda\mu I_d = C_\lambda. \tag{64}$$

The second term uses $A_c^+ A_c = I_d$. A Gaussian vector with mean u and covariance S_λ has total mass one almost surely, because its sum has mean one and variance $\mathbf{1}^\top S_\lambda \mathbf{1} = 0$. It therefore gives a law on $\mathcal{H}$ with the prescribed moments. Substituting the covariance identity into (20) proves (24).

For the minimum-norm interpretation, $(A_c^+)^\top g = A_c(A_c^\top A_c)^{-1} g$ solves $A_c^\top z = g$ and lies in $\mathrm{range}(A_c)$. Every other solution differs by a vector in $\ker(A_c^\top) = \mathrm{range}(A_c)^\perp$, so the Pythagorean theorem gives a larger Euclidean norm. The change is in T because $\mathrm{range}(A_c) \subseteq \mathsf{T}$.

D.3 An explicit distribution with the prescribed covariance

The following alternative to the Gaussian law separates a contracted empirical draw from an independent isotropic perturbation.

Theorem 12 (Random weights realizing spherical shrinkage)**.** *Assume that A_c has full column rank. Let I be uniform on $\{1, \ldots, n\}$ and let $G \sim \mathcal{N}(0, I_d)$ be independent of I. Define*

$$W_\lambda = u + \sqrt{1-\lambda}(e_I - u) + \sqrt{\lambda\mu}(A_c^+)^\top G. \tag{65}$$

Then $W_\lambda \in \mathcal{H}$, $\mathbb{E}W_\lambda = u$, and $\mathrm{Cov}(W_\lambda) = S_\lambda$. Moreover,

$$\mathbb{E}_{\mu_{W_\lambda}}[a] = \bar{a} + \sqrt{1-\lambda}(a_I - \bar{a}) + \sqrt{\lambda\mu}\, G, \qquad \mathrm{Cov}\Big(\mathbb{E}_{\mu_{W_\lambda}}[a]\Big) = C_\lambda. \tag{66}$$

Proof. Both random terms are centered, and independence gives covariance $(1-\lambda)\Sigma_u + \lambda\mu(A_c^+)^\top A_c^+ = S_\lambda$. The identity $A_c^+\mathbf{1} = 0$ gives $\sum_i W_{\lambda,i} = 1$. The first expectation is $\sum_i W_{\lambda,i} a_i$. Using $A_c^+ A = I_d$ in its Gaussian term gives (66), and independence gives its covariance. □

The first random term contracts $a_I - \bar{a}$ by $\sqrt{1-\lambda}$, and the second adds independent isotropic variation with covariance $\lambda\mu I_d$. The resulting covariance is C_λ.

Proposition 13 (Covariance ceiling for simplex weights)**.** *Let $W \in \Delta_n$ almost surely with $\mathbb{E}W = u$. Then*

$$\mathrm{Cov}(W) \preceq \Sigma_u, \qquad \mathrm{Cov}(\mathbb{E}_{\mu_W}[a]) \preceq \widehat{C}. \tag{67}$$

The vertex variable e_I attains both bounds. Equality in the first requires vertex-valued weights.

D.4 Proof of Proposition 13

We prove the more general bound

$$\mathrm{Cov}(W) \preceq \mathrm{diag}(p) - pp^\top \tag{68}$$

for any $p \in \Delta_n$ and any simplex-valued W with $\mathbb{E}W = p$. Conditional on W, draw a categorical vertex $V = e_I$ with $\Pr(I = i \mid W) = W_i$. Then

$$\mathbb{E}[V \mid W] = W, \qquad \mathrm{Cov}(V \mid W) = \mathrm{diag}(W) - WW^\top. \tag{69}$$

The marginal probabilities of V are p, so $\mathrm{Cov}(V) = \mathrm{diag}(p) - pp^\top$. The law of total covariance gives

$$\mathrm{diag}(p) - pp^\top = \mathrm{Cov}(W) + \mathbb{E}\big[\mathrm{diag}(W) - WW^\top\big]. \tag{70}$$

The final term is positive semidefinite, proving (68). The categorical vertex with probabilities p attains equality. Conversely, equality in (70) implies

$$\mathbb{E}\left[1 - \|W\|_2^2\right] = \mathbb{E}\,\mathrm{tr}\big(\mathrm{diag}(W) - WW^\top\big) = 0. \tag{71}$$

For $W \in \Delta_n$, the nonnegative quantity $1 - \|W\|_2^2$ vanishes only at a vertex. Hence equality requires W to be vertex valued almost surely. Setting $p = u$ gives $\mathrm{Cov}(W) \preceq \Sigma_u$. Multiplication by $A^\top$ and A, together with $A^\top \Sigma_u A = \widehat{C}$, gives the second bound in (26).

If $\widehat{C}$ has unequal eigenvalues, choose a unit eigenvector v with eigenvalue $c < \mu$. For $\lambda > 0$,

$$v^\top C_\lambda v = (1-\lambda)c + \lambda\mu > c = v^\top \widehat{C} v.$$

This violates the covariance ceiling and proves the obstruction stated after Proposition 13.

D.5 Spectral form of the shrinkage covariance

Let

$$A_c = UDV_A^\top, \qquad D = \mathrm{diag}(\sqrt{nc_1}, \ldots, \sqrt{nc_d}) \tag{72}$$

be a thin singular value decomposition. The columns of U lie in T, so $P_{\mathsf{T}} U = U$, and the pseudoinverse is $A_c^+ = V_A D^{-1} U^\top$. The columns of U are also eigenvectors of S_λ. Equation (22) gives

$$S_\lambda U_j = \frac{1}{n}\left((1-\lambda) + \lambda\frac{\mu}{c_j}\right) U_j, \tag{73}$$

so directions with smaller empirical eigenvalues require larger variance in the corresponding weight direction. For use below, set

$$c_j^\lambda = (1-\lambda)c_j + \lambda\mu. \tag{74}$$

The projector $P_{A_c} = A_c A_c^+$ onto $\mathrm{range}(A_c)$ gives a second decomposition:

$$S_\lambda = (A_c^+)^\top C_\lambda A_c^+ + \frac{1-\lambda}{n}(P_{\mathsf{T}} - P_{A_c}). \tag{75}$$

The first term is supported on $\mathrm{range}(A_c)$ and maps to C_λ under multiplication by $A_c^\top$ and A_c. The second retains a fraction $1-\lambda$ of the vertex variance in $\mathsf{T} \cap \ker(A_c^\top)$, which is invisible to every linear loss formed from the rows of A.

The following signed dual representation produces the mean–standard-deviation expression in (19). Its dual variables have expectation one and may take either sign (Peng et al., 2024). Define

$$\mathcal{Q}_\kappa^{(2)} = \left\{ Q \in L^2(\pi) : \mathbb{E}Q = 1, \quad \mathbb{E}[(Q-1)^2] \le \kappa^2 \right\}. \tag{76}$$

Proposition 14 (Quadratic representation and induced sample weights)**.** *Let $W \in \mathcal{H}$ have law π, mean u, and covariance S. Then*

$$\left\{ \mathbb{E}[QW] : Q \in \mathcal{Q}_\kappa^{(2)} \right\} = \left\{ u + \delta : \delta \in \mathrm{range}(S), \quad \delta^\top S^+ \delta \le \kappa^2 \right\}. \tag{77}$$

Its support function is (19)*, and its unique maximizing weight when $x^\top S x > 0$ is* (81)*.*

D.6 Proof of Proposition 14

Write $\widetilde{W} = W - u$ and let $L_0^2(\pi)$ denote the centered square-integrable random variables under π. Define the bounded linear map $\mathcal{A} : L_0^2(\pi) \to \mathbb{R}^n$ by $\mathcal{A}Z = \mathbb{E}[Z\widetilde{W}]$. For $b \in \mathbb{R}^n$, its adjoint is $\mathcal{A}^* b = b^\top \widetilde{W}$, and therefore

$$\mathcal{A}\mathcal{A}^* b = Sb. \tag{78}$$

Because the range of $\mathcal{A}$ is finite dimensional and therefore closed,

$$\mathrm{range}(\mathcal{A}) = \mathrm{range}(\mathcal{A}\mathcal{A}^*) = \mathrm{range}(S). \tag{79}$$

Thus every vector of the form $\mathcal{A}Z$ lies in $\mathrm{range}(S)$. Writing $Q = 1 + Z$ in (76), the centered part of the induced envelope is the image under $\mathcal{A}$ of the radius-κ ball in $L_0^2(\pi)$.

Fix $\delta \in \mathrm{range}(S)$ and define $Z_\delta = (S^+\delta)^\top \widetilde{W}$. Equation (78) gives

$$\mathcal{A}Z_\delta = SS^+\delta = \delta, \qquad \mathbb{E}Z_\delta^2 = \delta^\top S^+ \delta. \tag{80}$$

Moreover, $\mathcal{A}$ has finite rank, so $\mathrm{range}(\mathcal{A}^*)$ is closed and equals $(\ker \mathcal{A})^\perp$. In particular, $Z_\delta \in (\ker \mathcal{A})^\perp$. Every other Z satisfying $\mathcal{A}Z = \delta$ is Z_δ plus an element of $\ker \mathcal{A}$ and consequently has weakly larger L^2 norm. Thus δ is attainable with $\mathbb{E}Z^2 \le \kappa^2$ exactly when $\delta^\top S^+ \delta \le \kappa^2$, proving (77).

The supremum of the signed expectation $\mathbb{E}_{\mu_q}[X]$ over this ellipsoid is $\mathbb{E}_{\widehat{P}_n}[X] + \kappa\sqrt{x^\top S x}$. When $x^\top S x > 0$, equality in Cauchy–Schwarz gives the unique exposed point

$$q_S^*(X) = u + \delta^* = u + \kappa \frac{Sx}{\sqrt{x^\top S x}}. \tag{81}$$

This proves the final statement of Proposition 14.

D.7 Effect of shrinkage on the induced weights

The coefficients in (81) maximize the signed expectation over the ellipsoid in (77). Substituting S_λ gives a comparison with the unshrunk objective at the same decision.

Corollary 15 (Effect of shrinkage on the induced weights)**.** *Assume that A_c has full column rank, $\kappa > 0$, and $\theta \neq 0$. Let $q_\lambda^*(\theta) = q_{S_\lambda}^*(X_\theta)$ and let $q_0^*(\theta)$ denote the induced weights for the unshrunk covariance. Then*

$$\begin{aligned} q_\lambda^*(\theta) - u = {} & \frac{\sqrt{\theta^\top \widehat{C}\theta}}{\sqrt{\theta^\top C_\lambda \theta}} \big(q_0^*(\theta) - u \big) \\ & + \frac{\kappa\lambda}{\sqrt{\theta^\top C_\lambda \theta}} \left[\mu (A_c^+)^\top - \frac{A_c}{n} \right] \theta. \end{aligned} \tag{82}$$

If $\widehat{C}\theta = c\theta$ for an eigenvalue $c > 0$, this reduces to

$$q_\lambda^*(\theta) - u = \sqrt{\frac{(1-\lambda)c + \lambda\mu}{c}} \big(q_0^*(\theta) - u \big). \tag{83}$$

The first term in (82) adjusts the normalization by the standard deviation. The second redistributes the weights according to the empirical covariance. For a decision along one eigendirection, (83) shows that shrinkage increases the departure from uniform weights when $c < \mu$ and reduces it when $c > \mu$. The comparison holds at any fixed nonzero decision, including a minimizer of the unshrunk objective.

Proof and spectral form. For the linear loss X_θ, its observed vector is $A\theta$. Use $S_\lambda A = S_\lambda A_c$ and $A_c^+ A_c = I_d$ to obtain

$$S_\lambda A_c\theta = \frac{1-\lambda}{n} A_c\theta + \lambda\mu(A_c^+)^\top\theta. \tag{84}$$

Since $\theta^\top C_\lambda\theta > 0$ for $\theta \neq 0$, substituting (84) into (81) gives

$$q_\lambda^*(\theta) - u = \frac{\kappa}{\sqrt{\theta^\top C_\lambda\theta}} \left[\frac{A_c}{n} + \lambda\left(\mu(A_c^+)^\top - \frac{A_c}{n}\right)\right]\theta.$$

At $\lambda = 0$ this is $q_0^*(\theta) - u = \kappa A_c\theta/[n\sqrt{\theta^\top\widehat{C}\theta}]$. Substitution proves (82). If $\widehat{C}\theta = c\theta$, then $(A_c^+)^\top\theta = A_c(A_c^\top A_c)^{-1}\theta = A_c\theta/(nc)$ and $\theta^\top C_\lambda\theta = ((1-\lambda)c + \lambda\mu)\|\theta\|^2$. These identities give (83).

With the decomposition (72) and $y = V_A^\top\theta$,

$$\frac{1}{n}A_c\theta = \frac{1}{\sqrt{n}}\sum_{j=1}^{d}\sqrt{c_j}\, y_j U_j, \tag{85}$$

$$\mu(A_c^+)^\top\theta = \frac{1}{\sqrt{n}}\sum_{j=1}^{d}\frac{\mu}{\sqrt{c_j}} y_j U_j. \tag{86}$$

Substitution into (84) yields

$$S_\lambda A_c\theta = \frac{1}{\sqrt{n}}\sum_{j=1}^{d}\frac{c_j^\lambda}{\sqrt{c_j}} y_j U_j. \tag{87}$$

Since $\theta^\top C_\lambda\theta = \sum_j c_j^\lambda y_j^2$, applying (81) to $X = X_\theta$ gives

$$q_\lambda^*(\theta) - u = \frac{\kappa}{\sqrt{n}} \frac{\displaystyle\sum_{j=1}^{d}\frac{c_j^\lambda}{\sqrt{c_j}} y_j U_j}{\sqrt{\displaystyle\sum_{j=1}^{d} c_j^\lambda y_j^2}}. \tag{88}$$

Relative to the unshrunk covariance, the change before normalization is

$$\left[\mu(A_c^+)^\top - \frac{1}{n}A_c\right]\theta = \frac{1}{\sqrt{n}}\sum_{j=1}^{d}\frac{\mu - c_j}{\sqrt{c_j}} y_j U_j. \tag{89}$$

The factor $(\mu - c_j)/\sqrt{c_j}$ is positive for eigenvalues below the average and negative for eigenvalues above it. For $c_j < \mu$, its magnitude increases as c_j decreases. Equation (88) then divides all components by the same standard deviation, which depends on the full decision vector.

D.8 General covariance targets and additive covariance

The main text uses the spherical target μI_d. The covariance identity extends to any positive semidefinite target C_0. Under the same assumption that A_c has full column rank, set

$$S_{\lambda,C_0} = (1-\lambda)\frac{P_{\mathsf{T}}}{n} + \lambda(A_c^+)^\top C_0 A_c^+, \qquad C_{\lambda,C_0} = (1-\lambda)\widehat{C} + \lambda C_0. \tag{90}$$

Then $S_{\lambda,C_0} \succeq 0$, $S_{\lambda,C_0}\mathbf{1} = 0$, and

$$A^\top S_{\lambda,C_0} A = C_{\lambda,C_0}. \tag{91}$$

These identities follow from the same calculation as (64), with μI_d replaced by C_0.

A sufficiently small positive multiple of S_{λ,C_0} is attainable by probability weights. If $S_{\lambda,C_0} = 0$, the constant weight $W = u$ realizes it. Otherwise, let $m = \operatorname{rank}(S_{\lambda,C_0}) \geq 1$ and write $S_{\lambda,C_0} = \sum_{j=1}^{m}\eta_j b_j b_j^\top$, where $b_j \in \mathsf{T}$. For small enough $c > 0$, all points

$$u \pm \sqrt{cm\eta_j}\, b_j, \qquad j = 1, \ldots, m, \tag{92}$$

belong to Δ_n. Assigning probability $1/(2m)$ to each of these points gives mean u and covariance cS_{λ,C_0}. For $\kappa > 0$, using this law with covariance cS_{λ,C_0} and outer coefficient $\kappa/\sqrt{c}$ gives the same standard deviation penalty as covariance S_{λ,C_0} and coefficient κ in (19). Thus the simplex obstruction concerns the covariance at its prescribed scale and mean, not every possible reparameterization of the objective.

A different construction adds covariance to the vertex distribution rather than shrinking the empirical covariance. Define

$$D_{\lambda,C_0} = \lambda(A_c^+)^\top C_0 A_c^+, \qquad W_{\lambda,C_0}^{\mathrm{add}} = V + Z_{\lambda,C_0}, \qquad Z_{\lambda,C_0} \sim \mathcal{N}(0, D_{\lambda,C_0}), \tag{93}$$

with $V = e_I$ for uniform I and Z_{λ,C_0} independent of V. Full column rank gives

$$\mathrm{Cov}\left(\mathbb{E}_{\mu_{W_{\lambda,C_0}^{\mathrm{add}}}}[a]\right) = \widehat{C} + \lambda C_0. \tag{94}$$

Moreover, $\mathbb{E}[W_{\lambda,C_0}^{\mathrm{add}} \mid V] = V$, so conditional Jensen's inequality gives $V \preceq_{\mathrm{cx}} W_{\lambda,C_0}^{\mathrm{add}}$. This construction therefore increases every risk functional that is monotone with respect to convex order.

Every covariance S of a random weight vector in $\mathcal{H}$ satisfies $S\mathbf{1} = 0$ and hence $S = P_{\mathsf{T}} S P_{\mathsf{T}}$. Consequently,

$$\mathrm{range}(A^\top S A) \subseteq \mathrm{range}(A_c^\top), \qquad \mathrm{rank}(A^\top S A) \le \min\{d, n-1\}. \tag{95}$$

This range condition describes the covariance directions available from weights on $\mathcal{H}$ when the centered data matrix is rank deficient.

E Details for the Dirichlet CVaR stochastic subgradient

E.1 General induced subgradients

The induced coefficient formula is a standard adjoint identity (Rockafellar & Wets, 1998; Javeed et al., 2023). Let $W \in \mathcal{H}$ have any law, put $Y = \mathbb{E}_{\mu_W}[X]$, and assume the resulting risk is finite for every observed loss vector x. If Q^* is a risk identifier at Y and $\mathbb{E}[|Q^*| \sum_i |W_i|] < \infty$, then

$$q^* = \mathbb{E}[Q^* W] \in \partial_x \mathcal{R}(\mathbb{E}_{\mu_W}[X]). \tag{96}$$

Indeed, for another loss $\widetilde{X}$ with values $\widetilde{x}_i$,

$$\mathcal{R}\Big(\mathbb{E}_{\mu_W}[\widetilde{X}]\Big) \ge \mathcal{R}(Y) + \mathbb{E}\Big[Q^* \mathbb{E}_{\mu_W}[\widetilde{X} - X]\Big] \tag{97}$$

$$= \mathcal{R}(Y) + \sum_i \mathbb{E}[Q^* W_i](\widetilde{x}_i - x_i). \tag{98}$$

This proves (96) by the finite-dimensional subgradient definition. When W and Q^* are nonnegative and normalized, the coefficients form a probability vector.

Proposition 16 (Conditionally unbiased subgradient estimates)**.** *Let $\mathcal{G}_t$ contain the iterates and randomness observed before iteration t. Suppose that the finite losses and chosen finite subgradients at (θ_t, C_t) are $\mathcal{G}_t$-measurable, and the iteration batch is independent of $\mathcal{G}_t$ with iid $\mathrm{Dir}(\tau u)$ draws. Assume $\mathbb{E}[\max_i \|h_i^t\| + \|v_t\|] < \infty$. Then the estimates using $\mathbf{1}_{\{Y>C\}}$ in (31) satisfy*

$$\begin{aligned} \mathbb{E}[\widehat{q}_t \mid \mathcal{G}_t] &= q(\theta_t, C_t), \\ \mathbb{E}[(\widehat{g}_{\theta,t}, \widehat{g}_{C,t}) \mid \mathcal{G}_t] &= (g_\theta(\theta_t, C_t), g_C(\theta_t, C_t)) \in \partial F(\theta_t, C_t). \end{aligned} \tag{99}$$

E.2 Proof of Proposition 16

First consider the population subgradient, allowing a general base $p \in \mathrm{int}\,\Delta_n$ and $W \sim \mathrm{Dir}(\tau p)$. Use the objective (29) with this weight law, put $Y = \mathbb{E}_{\mu_W}[X_\theta]$, and choose a measurable hinge selector $\zeta(W) \in \partial(Y - C)_+$. Thus $\zeta = 1$ above the threshold, $\zeta = 0$ below it, and any value in $[0, 1]$ is allowed at a tie. Define

$$q(\theta, C) = \frac{\mathbb{E}[\zeta(W) W]}{1 - \alpha}, \qquad g_C(\theta, C) = 1 - \frac{\mathbb{E}[\zeta(W)]}{1 - \alpha}. \tag{100}$$

For a fixed W, nonnegative sample weights preserve convexity, and $\sum_i W_i h_i$ is a subgradient of $\mathbb{E}_{\mu_W}[X_\theta]$. The hinge chain rule and $\zeta(W) \geq 0$ therefore give the joint subgradient

$$\left(\frac{\zeta(W)}{1-\alpha} \sum_i W_i h_i + v,\ 1 - \frac{\zeta(W)}{1-\alpha} \right).$$

Taking expectations in its subgradient inequality yields

$$\left(\sum_i q_i(\theta, C) h_i + v,\ g_C(\theta, C) \right) \in \partial F(\theta, C). \tag{101}$$

At a fixed iterate, the weights and selector are bounded. The finite sample losses and chosen finite subgradients make these expectations finite.

For the main proposition, set $p = u$ and $\zeta(W) = \mathbf{1}_{\{\mathbb{E}_{\mu_W}[X_{\theta_t}] > C_t\}}$. Conditional on $\mathcal{G}_t$, the iterate, h_i^t, and v_t are fixed. Independence and identical distribution of the fresh draws give

$$\mathbb{E}[\widehat{q}_t \mid \mathcal{G}_t] = \frac{1}{M(1-\alpha)} \sum_{m=1}^{M} \mathbb{E}[W^{(m)} \mathbf{1}_{\{Y_m > C_t\}} \mid \mathcal{G}_t] \tag{102}$$

$$= q(\theta_t, C_t). \tag{103}$$

The estimates satisfy $\|\widehat{g}_{\theta,t}\| \leq (1-\alpha)^{-1} \max_i \|h_i^t\| + \|v_t\|$ and $|\widehat{g}_{C,t}| \leq \max\{1, \alpha/(1-\alpha)\}$. Under the integrability assumed in the proposition, conditional expectation commutes with the finite sums defining the estimates. This proves both equalities in (99), while (101) proves subgradient membership.

For nonconstant observed losses, the affine map $w \mapsto \mathbb{E}_{\mu_w}[X_\theta]$ has level sets of zero relative volume in the simplex. Since the Dirichlet law has a density on its interior, Y has a continuous distribution. At a minimizing threshold C^*, $\Pr(Y > C^*) = 1 - \alpha$, so its scalar CVaR identifier is $Q^* = \mathbf{1}_{\{Y > C^*\}}/(1-\alpha)$. Equation (96) then gives (28).

For every finite batch at a general threshold,

$$\sum_i \widehat{q}_i = 1 - \widehat{g}_C. \tag{104}$$

This identity checks that the decision coefficients and threshold subgradient use the same selected draws and normalization.

E.3 A FRACTIONAL RULE FOR TIES

The strict indicator is sufficient whenever Y is continuous. When one or more sampled losses equal C, let

$$N_+ = \sum_{m=1}^{M} \mathbf{1}_{\{Y_m > C\}}, \qquad N_0 = \sum_{m=1}^{M} \mathbf{1}_{\{Y_m = C\}}. \tag{105}$$

If $N_0 > 0$, set

$$\lambda_{\text{tie}} = \text{clip}\left(\frac{M(1-\alpha) - N_+}{N_0}, 0, 1 \right), \qquad \zeta_m = \mathbf{1}_{\{Y_m > C\}} + \lambda_{\text{tie}} \mathbf{1}_{\{Y_m = C\}}. \tag{106}$$

If $N_0 = 0$, take $\lambda_{\text{tie}} = 0$. This chooses an admissible hinge subgradient and makes the selected batch tail mass as close as possible to $M(1-\alpha)$. In particular, if all samples are tied, then $\lambda_{\text{tie}} = 1 - \alpha$ and $\widehat{g}_C = 0$. Because this selector depends on the whole batch, its expected coefficients can differ from the strict-indicator coefficients in Proposition 16. For example, if the loss is constant and equal to C, the strict rule gives $\widehat{q} = 0$ and $\widehat{g}_C = 1$, whereas the fractional rule gives $\widehat{q} = M^{-1} \sum_m W^{(m)}$ and $\widehat{g}_C = 0$. Both are valid batch subgradients. Taking expectations in the batch subgradient inequality shows that the fractional rule also has a population subgradient as its conditional mean.

E.4 Stable Dirichlet sampling and blockwise accumulation

A Dirichlet vector can be sampled by normalizing independent Gamma variables, but direct Gamma draws can underflow when τp_i is small. The implementation uses mutually independent H_{mi} and U_{mi} across all indices in the following exact identity for Gamma shapes:

$$\begin{aligned} H_{mi} &\sim \mathrm{Gamma}(\tau p_i + 1, 1), \qquad U_{mi} \sim \mathrm{Unif}(0,1), \\ \log G_{mi} &= \log H_{mi} + \frac{\log U_{mi}}{\tau p_i}, \qquad W_{mi} = \frac{\exp(\log G_{mi} - d_m)}{\sum_{j=1}^{n} \exp(\log G_{mj} - d_m)}, \\ d_m &= \max_j \log G_{mj}. \end{aligned} \tag{107}$$

Here $G_{mi} \sim \mathrm{Gamma}(\tau p_i, 1)$ exactly. The shift within each row keeps its largest finite log weight at zero. Provided at least one log weight is finite, the normalizing denominator is nonzero in floating-point arithmetic. The required Monte Carlo statistics are additive, so the rows can be processed in blocks $\mathcal{B}$ of size at most B. With $a_m = \mathbf{1}_{\{Y_m > C\}}$ and $e_m = \mathbf{1}_{\{Y_m = C\}}$, accumulate

$$\begin{aligned} (s_H, s_H^{(2)}) &\leftarrow (s_H, s_H^{(2)}) + \left(\sum_{m\in\mathcal{B}} (Y_m - C)_+, \sum_{m\in\mathcal{B}} (Y_m - C)_+^2 \right), \\ (N_+, N_0) &\leftarrow (N_+, N_0) + \left(\sum_{m\in\mathcal{B}} a_m, \sum_{m\in\mathcal{B}} e_m \right), \\ (s_q^+, s_q^0) &\leftarrow (s_q^+, s_q^0) + \left(\sum_{m\in\mathcal{B}} W^{(m)} a_m, \sum_{m\in\mathcal{B}} W^{(m)} e_m \right). \end{aligned} \tag{108}$$

After the final block, compute λ_{tie} from (106) and set $s_q = s_q^+ + \lambda_{\mathrm{tie}} s_q^0$ and $N_{\mathrm{sel}} = N_+ + \lambda_{\mathrm{tie}} N_0$. Then $\widehat{q} = s_q/[M(1-\alpha)]$ and $\widehat{g}_C = 1 - N_{\mathrm{sel}}/[M(1-\alpha)]$. The simultaneously accumulated value is

$$\widehat{F}_{\mathrm{RU}} = C + \frac{s_H}{M(1-\alpha)}, \qquad \widehat{\mathrm{se}}(\widehat{F}_{\mathrm{RU}}) = \frac{1}{1-\alpha} \sqrt{\frac{\widehat{\mathrm{Var}}((Y-C)_+)}{M}}. \tag{109}$$

These calculations give the same batch estimator as forming the full $M \times n$ matrix, up to floating-point reduction order.

E.5 Initialization and implementation safeguards

Given an initial model θ_0, draw an independent pilot batch $W_0^{(1)}, \ldots, W_0^{(M_0)}$ and take

$$C_0 = \mathrm{Quantile}_\alpha \left(\{ \mathbb{E}_{\mu_{W_0^{(m)}}} [X_{\theta_0}] \}_{m=1}^{M_0} \right). \tag{110}$$

This initializes the threshold at θ_0. All subsequent iterations update C together with θ. Independent random streams are used for the pilot, training, and diagnostic batches.

For a binary linear model, let $y_i \in \{-1, 1\}$ and use the margin loss $x_i(\theta, b) = -y_i(a_i^\top \theta + b)$, where a_i is the feature vector. Suppose that the intercept b is unpenalized, and let $m \le 1/2$ be the probability mass of the minority class under the base distribution. As b tends to infinity in the direction that favors the majority class, $x_i(\theta, b)/|b|$ tends to 1 on the minority class and to -1 on the majority class. If U_m is the total Dirichlet weight on the minority class, the weighted loss divided by $|b|$ therefore tends to $2U_m - 1$. Under uniform empirical weights, Dirichlet aggregation gives $U_m \sim \mathrm{Beta}(\tau m, \tau(1-m))$. Positive homogeneity of CVaR then shows that strict coercivity in this direction requires

$$\mathrm{CVaR}_\alpha(2U_m - 1) > 0. \tag{111}$$

On imbalanced data this condition can fail. When the intercept is explicitly excluded from the penalty, the implementation evaluates the condition before optimization and warns when necessary. An optional projection keeps the model iterates bounded and defines a constrained optimization problem. The feasibility check assesses the objective with an unrestricted intercept.

E.6 POPULATION STOCHASTIC APPROXIMATION

Let $\mathcal{G}_t$ be the sigma-field generated by the iterates and all randomness observed before iteration t. Because Algorithm 1 draws a new independent batch,

$$\mathbb{E}[(\widehat{g}_{\theta,t}, \widehat{g}_{C,t}) \mid \mathcal{G}_t] \in \partial F(\theta_t, C_t). \tag{112}$$

Thus each update estimates a subgradient of the population objective (29). In the convex setting, the standard assumptions are a nonempty solution set, bounded second moments, bounded iterates (intrinsically or by projection), and model steps satisfying

$$\sum_{t=0}^{\infty} \eta_t = \infty, \qquad \sum_{t=0}^{\infty} \eta_t^2 < \infty. \tag{113}$$

The threshold step in Algorithm 1 is $s_C \eta_t$ for a fixed $s_C > 0$, so it satisfies the same summability conditions. Under these conditions, standard stochastic-subgradient convergence results apply. The implementation averages both the model and threshold iterates after a chosen start time.

F EXPERIMENTAL DETAILS

F.1 NORMALIZATION OF THE EMPIRICAL CVAR BASELINE

The CVaR–ν-SVM correspondence is established by Gotoh & Takeda (2005). For our convention, $\nu = 1 - \alpha$. To match the regularization in (32), scale the classifier coefficients, intercept, margin threshold $r = -C$, and slack variables by $\nu\lambda_{\mathrm{reg}}$. The usual ν-SVM objective (Schölkopf et al., 2000) then equals $\nu^2\lambda_{\mathrm{reg}}$ times our empirical CVaR objective. The correspondence concerns attained, nondegenerate solutions. At an optimum with $\theta \neq 0$, minimization along the positive ray gives a negative objective value, which requires $r > 0$. The conventional nonnegative margin threshold therefore does not change such a solution.

A solver may return coefficients normalized by its margin parameter. For a returned direction (θ_0, b_0) with $\|\theta_0\| > 0$, let $R_0 = \mathrm{CVaR}_\alpha(\{-y_i(a_i^\top\theta_0 + b_0)\}_{i=1}^n)$. Along $(\theta, b) = \omega(\theta_0, b_0), \omega \geq 0$, positive homogeneity gives $\omega R_0 + \lambda_{\mathrm{reg}}\omega^2\|\theta_0\|^2/2$. Its minimizing scale is

$$\omega^* = \max\left\{0, -\frac{R_0}{\lambda_{\mathrm{reg}}\|\theta_0\|^2}\right\}. \tag{114}$$

A positive scale preserves predictions and AUC. A zero scale is a degenerate classifier and is handled by the eligibility criteria below. The same ray calculation applies to any positively homogeneous risk, including the Gaussian objective used for initialization.

F.2 DATA PROCESSING AND NESTED CROSS-VALIDATION

The Mammographic Mass data contain age and three attributes describing the mass: shape, margin, and density (Elter, 2007). We use these four predictors, following the dataset's designation of BI-RADS assessment as non-predictive. Missing values are imputed, retaining all 961 observations. Age and ordinal density receive median imputation and standardization. Shape and margin receive most-frequent imputation and one-hot encoding. An unseen category at evaluation receives an all-zero encoding in its categorical block. Every imputer, scaler, and encoder is fitted on the training portion of the current split, including during inner CV.

The outer partition uses shuffled stratified five-fold CV with seed 20260909. For each fixed α and outer fold $k \in \{1, \ldots, 5\}$, the three-fold inner partition is shuffled and stratified with seed $20260909 + 10^4\alpha + 100k + 1000$. All candidates in that search share these splits. Mean inner balanced accuracy determines the selected τ. Balanced accuracy is the mean of sensitivity and specificity, so the tuning criterion weights the two classes equally. The selected model is fitted on the complete outer training set before evaluation on the held-out fold. The five α values define separate comparisons, each with its own five outer evaluations.

F.3 CANDIDATE GRID AND HYPERPARAMETER TUNING

The Dirichlet grid is $\tau \in \{0.10, 0.25, 0.50, 1, 2, 4\}$, supplemented by empirical CVaR at $\tau = 0$ and Gaussian values $\tau \in \{-0.0197039506, -0.1735537190, -0.36\}$. The first two negative values are rounded here from $1.01^{-2} - 1$ and $1.10^{-2} - 1$. The regularization coefficient remains $\lambda_{\mathrm{reg}} = 0.05$ throughout. The numerical tie rule chooses the largest τ among scores within $10^{-8} + 10^{-5}|s_*|$ of the best inner score s_*.

For raw and Gaussian candidates, eligibility requires successful solver termination and a nondegenerate classifier in each inner fit. The Gaussian check requires both $\|\theta\|_2 \geq 10^{-3}$ and training-score standard deviation at least 10^{-4}. This prevents an arbitrary residual direction near the zero classifier from determining its validation score. Dirichlet candidates use the fixed iteration budget specified below. All positive concentrations in the grid satisfy the intercept condition (111) on every training split.

Across the 25 outer fits, inner CV chooses Dirichlet weights 15 times, Gaussian weights 8 times, and empirical CVaR twice. At $\alpha = 0.10$, all five selections are Dirichlet. The Dirichlet/Gaussian/empirical CVaR counts are $(1, 2, 2)$ at $\alpha = 0.20$, $(2, 3, 0)$ at $\alpha = 0.30$, $(3, 2, 0)$ at $\alpha = 0.40$, and $(4, 1, 0)$ at $\alpha = 0.50$. Both interpolation and extrapolation are thus selected. The performance comparison in Table 1 evaluates the complete tuning procedure. Isolating either family's contribution would require a separate comparison.

F.4 Optimization settings

Each Dirichlet fit uses 4,000 iterations of Algorithm 1, with 128 independent weight draws per update, $\eta_t = (t+10)^{-0.6}$ for $t = 0, \ldots, 3999$, and threshold step multiplier $s_C = 4$. We average the final 2,000 updated iterates of the coefficients, intercept, and threshold. A linear SVM with $C = 1$ initializes the classifier, and an independent batch of 256 draws initializes the CVaR threshold by an empirical quantile. Normalized Gamma sampling in logarithmic coordinates generates the weights, using batches of 16 draws and coordinate blocks of size 2,048. The iterates are unconstrained. The random seed for each Dirichlet fit is $20260909 + 10^4\alpha + 100k$. Training and diagnostic draws use separate random streams.

Empirical CVaR uses a linear ν-SVM with $\nu = 1 - \alpha$, solver tolerance 10^{-7}, and positive radial scaling to minimize (32) along its fitted direction. For Gaussian weights, the normal-mixture formula supplies the threshold objective and its gradient analytically. L-BFGS-B jointly updates the coefficients, intercept, and threshold, with a 4,000-iteration limit and function and gradient tolerances of 10^{-7}. Initialization uses a linear SVM with $C = 1$ followed by radial scaling for the Gaussian objective. The initialization seed is 20260909. All selected Gaussian outer refits terminated successfully and passed the nondegeneracy check.

F.5 Evaluation metrics

Predictions use the sign of $a^\top\theta + b$, with a zero score assigned to the malignant class. Classification error is the fraction of incorrect predictions. AUC is the area under the receiver operating characteristic curve, computed from the continuous classifier scores. F1 is the harmonic mean of precision and recall for the malignant class. Precision is the fraction of predicted malignant cases that are malignant, and recall is the fraction of malignant cases correctly classified.

Table 1 reports the mean and sample standard deviation across the five held-out folds at each fixed α. Error and F1 are expressed as percentages.