AIMS Mathematics, 11(8): 24331–24352.
DOI: 10.3934/math.2026982
Received: 28 April 2026
Revised: 16 July 2026
Accepted: 23 July 2026
Published: 10 August 2026

https://www.aimspress.com/journal/Math

Research article

Inference for maximum-entropy models with moment uncertainty sets: Duality and stability

Badr S. Alnssyan[1], Abdelaziz Alsubie[2], Sajad A. Sheikh[3,*] and Javid Gani Dar[4,*]

[1] Department of Management Information Systems, College of Business and Economics, Qassim University, Buraydah 51452, Saudi Arabia; Email: b.alnssyan@qu.edu.sa

[2] Department of Basic Sciences, College of Science and Theoretical Studies, Saudi Electronic University, Riyadh 11673, Saudi Arabia; Email: a.alsubie@seu.edu.sa

[3] Department of Mathematics, University of Kashmir, South Campus, Jammu and Kashmir 192101, India

[4] Department Of Applied Sciences, Symbiosis Institute of Technology, Symbiosis International (Deemed University), 412115, India

* **Correspondence:** Email: sajadsheikh@uok.edu.in, javid.dar@sitpune.edu.in.

Abstract: We study a robust formulation of the maximum-entropy principle in which moment information is specified through uncertainty sets that encode estimation error. The estimator selects the least-informative distribution consistent with a convex admissible set of moment vectors. We provide a self-contained duality theory for general compact convex uncertainty sets, including strong duality, existence and uniqueness of primal solutions, and an exponential-family characterization governed by a penalized log-partition dual objective. We derive quantitative stability bounds showing Lipschitz dependence of optimal dual parameters on the estimated moments and the uncertainty radius, and we translate these bounds into controls on Kullback–Leibler (KL) divergence, total variation distance, and errors of bounded forecasts. Data-driven uncertainty sets built from concentration inequalities yield finite-sample feasibility and oracle-type entropy guarantees. Applications to probabilistic forecasting and risk modeling are developed, together with explicit discrete examples illustrating entropy-driven shrinkage within uncertainty regions.

Keywords: maximum entropy; exponential families; moment constraints; support function; information projection; distributional robustness; stability; concentration inequalities; probabilistic forecasting and modeling; simulation; statistical model
Mathematics Subject Classification: 62F10, 62G20, 90C25, 94A17, 60E15

1. Introduction and motivation

The maximum-entropy principle selects a probability distribution by maximizing Shannon entropy subject to constraints that summarize available information [1–3]. When constraints are given by exact moment equalities, the solution lies in an exponential family, and the natural parameter is determined by the specified moments through a convex dual program [4–7]. In data-driven settings, moments are estimated from finite samples, sensor readings, or aggregated information across heterogeneous sources. Such estimates carry error, and equality constraints can propagate sampling fluctuations into substantial changes of the inferred distribution. Robust inference aims for estimators that remain well-posed under perturbations of empirical constraints and admit finite-sample guarantees; relaxations of exact moment matching in this spirit appear in generalized maximum-entropy estimation [8].

This work develops a robust maximum-entropy framework in which moment information is specified through an *uncertainty set* of admissible moment vectors. Given an empirical moment estimate and a confidence region, we maximize entropy over distributions whose moment vector belongs to the region. This formulation connects to robust and distributionally robust optimization under moment information [9–13], to transport-based ambiguity sets and their regularization interpretations [14–16], and to information projection theory for relative entropy minimization under linear constraints [4, 5, 7]. In contrast to worst-case expectation objectives, the present objective is entropy itself, yielding a canonical "least informative" distribution within the ambiguity set.

Applications motivate both the model and the analysis. In probabilistic forecasting, constraints encode calibration or feature matching, and entropy discourages unwarranted structure [17]. In risk modeling, moments of loss variables are estimated with error, and an entropy-maximizing distribution consistent with uncertainty bounds provides a conservative baseline for scenario generation and stress testing [11, 15, 18]. Stability bounds are valuable in rolling-window estimation and in sequential systems where constraints evolve with time.

Contributions. We introduce a robust maximum-entropy estimation under compact convex moment uncertainty sets and establish a complete dual characterization. We prove strong duality under a transparent relative-interior feasibility condition, show dual attainment, and obtain a Karush–Kuhn–Tucker (KKT) system that characterizes primal solutions as exponential-family distributions whose moment vector is a supporting point of the uncertainty set. We provide an equivalent reduced formulation in the moment space as the minimization of the convex conjugate of the log-partition function over the uncertainty set. We then develop stability bounds: Under mild local strong convexity of the log-partition function, the optimal dual parameter depends Lipschitz-continuously on the estimated moment center and on the uncertainty radius, and the induced distributions satisfy explicit bounds in Kullback–Leibler divergence, total variation distance, and bounded forecast error. We construct data-driven uncertainty sets using concentration inequalities for bounded features and derive finite-sample feasibility and oracle-type entropy guarantees. Applications to forecasting and risk modeling are presented, including explicit discrete examples showing shrinkage toward maximum-entropy moments within uncertainty sets.

Organization. Section 2 states the robust maximum-entropy problem, gives geometric and statistical interpretations of uncertainty sets, and discusses the scope of the assumptions. Section 3 collects

convex-analytic preliminaries and the Gibbs variational principle for relative entropy. Section 4 proves the existence and uniqueness of primal solutions. Section 5 derives the dual problem with intermediate conjugacy steps, proves strong duality with attainment, and develops KKT and moment-space characterizations. Section 6 establishes sensitivity bounds and explains their forecasting implications. Section 7 constructs finite-sample uncertainty sets and gives an implementation-oriented calibration guide. Section 8 develops applications, examples, and comparisons with neighboring entropy-learning approaches. Section 9 reports the numerical study of the uncertainty radius. Section 10 concludes.

2. Problem setup and assumptions

2.1. Topological and measure-theoretic setting

Let $(\mathcal{X}, d)$ be a Polish space with Borel σ-field $\mathcal{F}$ and let μ be a Borel probability measure on $(\mathcal{X}, \mathcal{F})$. We consider probability measures P on $(\mathcal{X}, \mathcal{F})$ that satisfy $P \ll \mu$ with density $p = \frac{dP}{d\mu}$. The relative entropy (KL divergence) of P from μ is

$$\mathrm{KL}(P\|\mu) := \int_{\mathcal{X}} \log\left(\frac{dP}{d\mu}\right) dP = \int_{\mathcal{X}} p(x) \log p(x)\, \mu(dx) \in [0, \infty].$$

The entropy of P relative to μ is $H_\mu(P) := -\mathrm{KL}(P\|\mu)$.

Let $\phi : \mathcal{X} \to \mathbb{R}^d$ be bounded and continuous. For $P \ll \mu$, define the moment vector directly through the coordinate variable x:

$$m(P) := \int_{\mathcal{X}} \phi(x)\, P(dx) \in \mathbb{R}^d.$$

2.2. Uncertainty sets and robust maximum entropy

Let $\mathcal{U} \subset \mathbb{R}^d$ be a nonempty compact convex set. In applications, $\mathcal{U}$ is centered at an empirical moment estimate and its size reflects estimation error.

The geometry becomes transparent after introducing the attainable moment body

$$\mathcal{M} := \{m(P) : P \ll \mu,\ \mathrm{KL}(P\|\mu) < \infty\} \subset \mathbb{R}^d.$$

Feasibility means that $\mathcal{U}$ intersects $\mathcal{M}$. A norm ball permits equal tolerance after the chosen scaling, while an ellipsoid assigns direction-dependent tolerances and can encode covariance among estimated moments. The optimizer selects a point $u_\star \in \mathcal{U} \cap \mathcal{M}$ with the smallest information cost $A^*(u_\star)$. As $\mathcal{U}$ expands, the selected point can move toward the reference moment $m(\mu) = \int \phi(x)\, \mu(dx)$. Once $m(\mu) \in \mathcal{U}$, the reference law μ is feasible and attains the minimum value zero. In the Bernoulli example of Proposition 8.1, this movement is the projection of $1/2$ onto an uncertainty interval.

Definition 2.1 (Robust maximum entropy under moment uncertainty)**.** Given (μ, ϕ) and $\mathcal{U}$, define the robust maximum-entropy problem

$$\text{maximize } H_\mu(P) \quad \text{over } P \ll \mu \quad \text{subject to } m(P) \in \mathcal{U}. \tag{2.1}$$

Equivalently,

$$\text{minimize } \mathrm{KL}(P\|\mu) \quad \text{over } P \ll \mu \quad \text{subject to } m(P) \in \mathcal{U}. \tag{2.2}$$

In the rest of the paper, we make the following two assumptions to develop our results.

Assumption 2.2 (Bounded features and finite-information feasibility)**.** There exists $R < \infty$ such that $\|\phi(x)\|_2 \le R$ for all $x \in \mathcal{X}$. The set $\mathcal{U} \subset \mathbb{R}^d$ is nonempty, compact, and convex. There exists a feasible probability measure $P^{\mathrm{f}} \ll \mu$ satisfying $m(P^{\mathrm{f}}) \in \mathcal{U}$ and $\mathrm{KL}(P^{\mathrm{f}}\|\mu) < \infty$.

Assumption 2.3 (Relative-interior feasibility)**.** There exists $P^{\circ} \ll \mu$ with $\mathrm{KL}(P^{\circ}\|\mu) < \infty$ such that $m(P^{\circ}) \in \mathrm{ri}(\mathcal{U})$.

Remark 2.4 (Scope of the assumptions). The assumptions provide sufficient conditions for a concise weak-compactness and Fenchel-duality treatment. Bounded continuity of ϕ makes $P \mapsto m(P)$ weakly continuous and makes $A(\lambda)$ finite for every λ. Exponential-integrability and coercivity conditions support extensions to unbounded features. Finite-information feasibility ensures a finite primal value. Compactness makes the support function finite and covers the confidence balls and ellipsoids used in applications; closed unbounded sets can be treated when the dual objective is coercive. Relative-interior feasibility supplies a Lagrange multiplier and zero duality gap. The relative interior is taken inside the affine hull of $\mathcal{U}$, so lower-dimensional uncertainty sets and exact moment sets are included. Primal uniqueness follows from strict convexity of relative entropy. Uniqueness of the dual parameter additionally requires a minimal feature representation or positive covariance on the parameter region, as formalized by Assumption 6.1.

3. Preliminaries

3.1. Support functions and dual norms

For a nonempty compact convex set $\mathcal{U} \subset \mathbb{R}^d$, its support function is

$$\sigma_{\mathcal{U}}(y) := \sup_{u \in \mathcal{U}} y^{\top} u, \qquad y \in \mathbb{R}^d.$$

It is convex, finite everywhere, and positively homogeneous.

For a norm $\|\cdot\|$ on $\mathbb{R}^d$, its dual norm is $\|y\|_* := \sup_{\|x\| \le 1} y^{\top} x$. If $\mathcal{U} = \widehat{m} + \rho\,\mathbb{B}$ with $\mathbb{B} = \{u : \|u\| \le 1\}$, then

$$\sigma_{\mathcal{U}}(y) = y^{\top}\widehat{m} + \rho\|y\|_*. \tag{3.1}$$

3.2. Log-partition and exponential families

Define the log-partition function

$$A(\lambda) := \log\left(\int_{\mathcal{X}} e^{\lambda^{\top}\phi(x)}\,\mu(dx)\right), \qquad \lambda \in \mathbb{R}^d. \tag{3.2}$$

Under Assumption 2.2, $A(\lambda)$ is finite for all λ and is C^{∞} and convex. Let $Z(\lambda) := e^{A(\lambda)}$ and define

$$\frac{dP_{\lambda}}{d\mu}(x) := \exp(\lambda^{\top}\phi(x) - A(\lambda)). \tag{3.3}$$

Differentiation under the integral gives $\nabla A(\lambda) = m(P_{\lambda})$ and

$$\nabla^2 A(\lambda) = \int_{\mathcal{X}} (\phi(x) - m(P_{\lambda}))(\phi(x) - m(P_{\lambda}))^{\top} P_{\lambda}(dx).$$

3.3. Gibbs variational principle

Lemma 3.1 (Gibbs variational principle)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be bounded and measurable. Then*

$$\log\left(\int_{\mathcal{X}} e^{f(x)}\,\mu(dx)\right) = \sup_{P \ll \mu}\left\{\int_{\mathcal{X}} f(x)\,P(dx) - \mathrm{KL}(P\|\mu)\right\}. \tag{3.4}$$

The supremum is attained at the probability measure P_f with density

$$\frac{dP_f}{d\mu}(x) = \frac{e^{f(x)}}{\int e^f\,d\mu}.$$

Proof. Define $Z := \int e^f\,d\mu$ and $q := \frac{e^f}{Z}$. Then q is a density with respect to μ. For any $P \ll \mu$ with density p,

$$\mathrm{KL}(P\|P_f) = \int p \log\left(\frac{p}{q}\right) d\mu \geq 0.$$

Expand:

$$\int p \log p\,d\mu - \int p \log q\,d\mu \geq 0.$$

Since $\log q = f - \log Z$, this becomes

$$\mathrm{KL}(P\|\mu) - \int f\,dP + \log Z \geq 0.$$

Rearrange:

$$\int f\,dP - \mathrm{KL}(P\|\mu) \leq \log Z.$$

Equality holds at $P = P_f$ because then $\mathrm{KL}(P\|P_f) = 0$. This proves (3.4). □

4. Existence and uniqueness

4.1. A tightness lemma under bounded KL

Lemma 4.1 (Set probability bound under KL)**.** *Let μ be a probability measure and let $P \ll \mu$ satisfy $\mathrm{KL}(P\|\mu) \leq C < \infty$. For any measurable set B with $\mu(B) \in (0, 1)$,*

$$P(B) \leq \frac{C + \log 2}{\log(1/\mu(B))}. \tag{4.1}$$

Proof. Set $q := \mu(B) \in (0, 1)$ and choose $a := \log(1/q) > 0$. Applying the Gibbs variational inequality from Lemma 3.1 to the bounded function $f = a\mathbf{1}_B$ gives

$$aP(B) - \mathrm{KL}(P\|\mu) \leq \log \int_{\mathcal{X}} e^{a\mathbf{1}_B(x)}\,\mu(dx).$$

The integral on the right equals $1 - q + qe^a = 2 - q \leq 2$. Using $\mathrm{KL}(P\|\mu) \leq C$ therefore yields

$$aP(B) \leq C + \log 2.$$

Substitution of $a = \log(1/q)$ proves (4.1). □

4.2. Primal existence and uniqueness

Theorem 4.2 (Existence and uniqueness of the robust maximum-entropy solution)**.** *Assume 2.2. Then the primal problem* (2.2) *admits a unique minimizer.*

Proof. Let

$$\mathcal{F} := \{P \ll \mu : m(P) \in \mathcal{U}\}, \qquad v := \inf_{P \in \mathcal{F}} \mathrm{KL}(P\|\mu).$$

Assumption 2.2 gives $v \le \mathrm{KL}(P^{\mathrm{f}}\|\mu) < \infty$. Select a minimizing sequence (P_n) with $\mathrm{KL}(P_n\|\mu) \le v + 1$ for all sufficiently large n.

Fix $\eta > 0$. Tightness of μ gives a compact set K with $\mu(K^c)$ small enough that

$$\frac{v + 1 + \log 2}{\log(1/\mu(K^c))} < \eta.$$

When $\mu(K^c) = 0$, absolute continuity gives $P_n(K^c) = 0$. Otherwise, Lemma 4.1 gives $P_n(K^c) < \eta$. Thus the minimizing sequence is tight. Prokhorov's theorem supplies a weakly convergent subsequence, still denoted by (P_n), with limit $P_\star$.

Bounded continuity of ϕ gives $m(P_n) \to m(P_\star)$. Closedness of $\mathcal{U}$ yields $m(P_\star) \in \mathcal{U}$. Lower semicontinuity of relative entropy gives

$$\mathrm{KL}(P_\star\|\mu) \le \liminf_{n\to\infty} \mathrm{KL}(P_n\|\mu) = v.$$

The finite right-hand side also implies $P_\star \ll \mu$, so $P_\star$ is feasible and optimal.

For uniqueness, let P_1 and P_2 be distinct feasible measures with finite relative entropy. For $t \in (0, 1)$, strict convexity of $z \mapsto z \log z$ gives

$$\mathrm{KL}(tP_1 + (1 - t)P_2\|\mu) < t\,\mathrm{KL}(P_1\|\mu) + (1 - t)\,\mathrm{KL}(P_2\|\mu).$$

Convexity of $\mathcal{U}$ makes the mixture feasible. Hence two distinct minimizers cannot occur. □

5. Duality and characterization

5.1. Dual problem and strong duality

Theorem 5.1 (Strong duality with attainment)**.** *Assume 2.2 and 2.3. Define the primal value*

$$v_{\mathrm{P}} := \inf_{P:\, m(P)\in\mathcal{U}} \mathrm{KL}(P\|\mu)$$

and the convex dual objective $D(\lambda) := A(\lambda) + \sigma_{\mathcal{U}}(-\lambda)$*. Then*

$$v_{\mathrm{P}} = \sup_{\lambda\in\mathbb{R}^d}\{-A(\lambda) - \sigma_{\mathcal{U}}(-\lambda)\} = -\min_{\lambda\in\mathbb{R}^d} D(\lambda). \tag{5.1}$$

The minimum is attained. For every dual minimizer $\lambda_\star$*, the unique primal minimizer is* $P_{\lambda_\star}$ *defined by* (3.3)*.*

Proof. Introduce an auxiliary moment variable u and write the primal problem as

$$\inf_{P,u} \{\mathrm{KL}(P\|\mu) + I_{\mathcal{U}}(u) : m(P) - u = 0\},$$

where $I_{\mathcal{U}}$ is the convex indicator of $\mathcal{U}$. With multiplier $\lambda \in \mathbb{R}^d$, use the Lagrangian

$$L(P, u; \lambda) = \mathrm{KL}(P\|\mu) - \lambda^\top m(P) + \lambda^\top u + I_{\mathcal{U}}(u).$$

The infimum over P is evaluated by the Gibbs variational principle:

$$\inf_{P \ll \mu} \{\mathrm{KL}(P\|\mu) - \lambda^\top m(P)\} = -A(\lambda),$$

and equality occurs at $P = P_\lambda$. The infimum over u is

$$\inf_{u \in \mathcal{U}} \lambda^\top u = -\sigma_{\mathcal{U}}(-\lambda).$$

Consequently, the Lagrange dual is

$$\sup_{\lambda \in \mathbb{R}^d} \{-A(\lambda) - \sigma_{\mathcal{U}}(-\lambda)\},$$

which is equivalent to minimizing D.

For completeness, the conjugate calculation has the same form. The convex conjugate of $P \mapsto \mathrm{KL}(P\|\mu)$ evaluated at $x \mapsto \lambda^\top \phi(x)$ is $A(\lambda)$ by Lemma 3.1, and the conjugate of $I_{\mathcal{U}}$ at $-\lambda$ is $\sigma_{\mathcal{U}}(-\lambda)$. Assumption 2.3 places the image of a finite-information point in $\mathrm{ri}(\mathrm{dom}\, I_{\mathcal{U}})$. Fenchel–Rockafellar duality therefore gives equality of the primal and dual values and attainment of the dual supremum [19].

Let $\lambda_\star$ be a dual minimizer and let $P_\star$ be the primal minimizer from Theorem 4.2. Equality of the primal and dual values forces equality in both infimum calculations. Equality in the Gibbs variational principle gives $P_\star = P_{\lambda_\star}$. The primal solution is unique, so every dual minimizer generates the same probability law. □

5.2. KKT conditions and supporting moments

Theorem 5.2 (KKT characterization)**.** *Assume 2.2 and 2.3, and let $\lambda_\star$ be a dual minimizer. Then*

$$m(P_{\lambda_\star}) = \nabla A(\lambda_\star) \in \arg\max_{u \in \mathcal{U}} (-\lambda_\star)^\top u. \tag{5.2}$$

Equivalently,

$$\nabla A(\lambda_\star) \in \partial \sigma_{\mathcal{U}}(-\lambda_\star). \tag{5.3}$$

Proof. First-order optimality for $D(\lambda) = A(\lambda) + \sigma_{\mathcal{U}}(-\lambda)$ gives

$$0 \in \nabla A(\lambda_\star) - \partial \sigma_{\mathcal{U}}(-\lambda_\star),$$

where the minus sign follows from the linear map $\lambda \mapsto -\lambda$. This inclusion is exactly (5.3). For a compact convex set,

$$\partial \sigma_{\mathcal{U}}(y) = \arg\max_{u \in \mathcal{U}} y^\top u.$$

Taking $y = -\lambda_\star$ yields (5.2). The identity $m(P_{\lambda_\star}) = \nabla A(\lambda_\star)$ follows by differentiating the log-partition function. □

5.3. Reduced formulation in moment space

Define the convex conjugate of A:

$$A^*(u) := \sup_{\lambda\in\mathbb{R}^d}\{\lambda^\top u - A(\lambda)\}, \qquad u \in \mathbb{R}^d.$$

Theorem 5.3 (Moment-space reduction)**.** *Assume 2.2 and 2.3. Then*

$$v_{\mathrm{P}} = \min_{u\in\mathcal{U}} A^*(u). \tag{5.4}$$

If $\lambda_\star$ is any dual minimizer, then the minimizing moment is

$$u_\star = m(P_{\lambda_\star}) = \nabla A(\lambda_\star),$$

and $\lambda_\star \in \partial A^*(u_\star)$.

Proof. For every $P \ll \mu$, the Gibbs variational inequality gives

$$\lambda^\top m(P) - A(\lambda) \le \mathrm{KL}(P\|\mu) \quad \text{for all } \lambda.$$

Taking the supremum over λ yields $A^*(m(P)) \le \mathrm{KL}(P\|\mu)$. Hence

$$\inf_{u\in\mathcal{U}} A^*(u) \le v_{\mathrm{P}}.$$

Let $P_\star = P_{\lambda_\star}$ be the primal solution and set $u_\star = m(P_\star)$. The Fenchel equality for the differentiable pair (A, A^*) gives

$$A^*(u_\star) = \lambda_\star^\top u_\star - A(\lambda_\star).$$

The exponential-family identity gives the same expression for $\mathrm{KL}(P_\star\|\mu)$. Therefore

$$A^*(u_\star) = \mathrm{KL}(P_\star\|\mu) = v_{\mathrm{P}}.$$

Since $u_\star \in \mathcal{U}$, it attains the moment-space minimum. The relation $u_\star = \nabla A(\lambda_\star)$ is equivalent to $\lambda_\star \in \partial A^*(u_\star)$. □

5.4. Explicit dual penalties for common uncertainty sets

Corollary 5.4 (Norm-ball uncertainty)**.** *Let $\mathcal{U} = \widehat{m} + \rho\mathbb{B}$ with $\mathbb{B} = \{u : \|u\| \le 1\}$. Then the dual* (5.1) *becomes*

$$\min_{\lambda\in\mathbb{R}^d} \{A(\lambda) - \lambda^\top\widehat{m} + \rho\|\lambda\|_*\}. \tag{5.5}$$

Proof. Use (3.1) with $y = -\lambda$. □

Corollary 5.5 (Ellipsoidal uncertainty)**.** *Let $\mathcal{U} = \{\widehat{m} + W^{-1}v : \|v\|_2 \le \rho\}$ for an invertible matrix $W \in \mathbb{R}^{d\times d}$. Then*

$$\sigma_{\mathcal{U}}(-\lambda) = -\lambda^\top\widehat{m} + \rho\|W^{-\top}\lambda\|_2,$$

and the dual is $\min_\lambda\{A(\lambda) - \lambda^\top\widehat{m} + \rho\|W^{-\top}\lambda\|_2\}$.

Proof. For every $y \in \mathbb{R}^d$,

$$\begin{aligned}\sigma_{\mathcal{U}}(y) &= \sup_{\|v\|_2\le\rho} y^\top(\widehat{m} + W^{-1}v)\\ &= y^\top\widehat{m} + \sup_{\|v\|_2\le\rho}(W^{-\top}y)^\top v\\ &= y^\top\widehat{m} + \rho\|W^{-\top}y\|_2.\end{aligned}$$

Setting $y = -\lambda$ gives the stated expression. □

6. Stability and sensitivity

Throughout this section, we focus on norm-ball uncertainty sets and quantify dependence on the center $\widehat{m}$ and radius ρ.

6.1. Curvature bounds

Assumption 2.2 implies a global upper bound on the Hessian of A:

$$\nabla^2 A(\lambda) \preceq R^2 I_d \qquad \text{for all } \lambda \in \mathbb{R}^d. \tag{6.1}$$

For any $v \in \mathbb{R}^d$, the covariance representation gives

$$v^\top \nabla^2 A(\lambda) v = \int_{\mathcal{X}} \left[v^\top(\phi(x) - m(P_\lambda))\right]^2 P_\lambda(dx) \le \int_{\mathcal{X}} (v^\top \phi(x))^2 P_\lambda(dx) \le \|v\|_2^2 R^2.$$

A lower bound is local.

Assumption 6.1 (Local strong convexity)**.** There exist $\alpha > 0$ and a convex set $\Lambda \subset \mathbb{R}^d$ such that for all $\lambda \in \Lambda$,

$$\nabla^2 A(\lambda) \succeq \alpha I_d.$$

6.2. Optimality conditions for norm-ball uncertainty

Assume $\mathcal{U} = \widehat{m} + \rho\mathbb{B}$.

Theorem 6.2 (KKT for norm balls)**.** *Let $\mathcal{U} = \widehat{m} + \rho\mathbb{B}$, where $\mathbb{B}$ is the unit ball of a norm $\|\cdot\|$. Let $\lambda_\star$ minimize* (5.5)*. Then*

$$0 \in \nabla A(\lambda_\star) - \widehat{m} + \rho\, \partial\|\lambda_\star\|_*. \tag{6.2}$$

Moreover, $u_\star := \nabla A(\lambda_\star)$ satisfies $\|u_\star - \widehat{m}\| \le \rho$ and $u_\star \in \mathcal{U}$.

Proof. Differentiate $A(\lambda) - \lambda^\top \widehat{m}$ and take the subdifferential of $\rho\|\lambda\|_*$. This yields (6.2). Let $g \in \partial\|\lambda_\star\|_*$ satisfy $\nabla A(\lambda_\star) - \widehat{m} + \rho g = 0$. By the definition of the subdifferential of a norm,

$$\|g\| \le 1 \quad \text{and} \quad g^\top \lambda_\star = \|\lambda_\star\|_*.$$

Hence

$$u_\star - \widehat{m} = -\rho g,$$

so $\|u_\star - \widehat{m}\| = \rho\|g\| \le \rho$, which implies $u_\star \in \widehat{m} + \rho\mathbb{B} = \mathcal{U}$. □

For Euclidean uncertainty, the mismatch has an explicit form.

Corollary 6.3 (Euclidean boundary condition)**.** *Assume $\mathcal{U} = \widehat{m} + \rho\mathbb{B}_2$ and let $\lambda_\star$ be the dual minimizer. If $\lambda_\star \neq 0$, then*

$$\nabla A(\lambda_\star) = \widehat{m} - \rho\, \frac{\lambda_\star}{\|\lambda_\star\|_2} \quad \textit{and} \quad \|\nabla A(\lambda_\star) - \widehat{m}\|_2 = \rho.$$

If $\lambda_\star = 0$, then $\nabla A(0) = \mathbb{E}_\mu[\phi] \in \mathcal{U}$ and the moment constraint is satisfied with slack.

Proof. For $\|\cdot\| = \|\cdot\|_2$, the dual norm equals $\|\cdot\|_2$ and $\partial\|\lambda\|_2 = \{\lambda/\|\lambda\|_2\}$ for $\lambda \neq 0$. Insert this into (6.2). □

6.3. Lipschitz dependence on the moment center

We state the result in general norms using dual norms for the Lipschitz constants.

Theorem 6.4 (Lipschitz sensitivity in $\widehat{m}$)**.** *Assume 6.1. Fix $\rho \geq 0$ and a norm $\|\cdot\|$ with dual $\|\cdot\|_*$. Let $\lambda_\star(\widehat{m})$ denote the unique minimizer of* (5.5) *for center $\widehat{m}$ and the given ρ. Suppose $\lambda_\star(\widehat{m}_1), \lambda_\star(\widehat{m}_2) \in \Lambda$. Then*

$$\|\lambda_\star(\widehat{m}_1) - \lambda_\star(\widehat{m}_2)\|_2 \leq \frac{1}{\alpha} \|\widehat{m}_1 - \widehat{m}_2\|_2. \tag{6.3}$$

Proof. Define $F_{\widehat{m}}(\lambda) := A(\lambda) - \lambda^\top \widehat{m} + \rho\|\lambda\|_*$. On Λ, A is α-strongly convex, hence $F_{\widehat{m}}$ is also α-strongly convex.

Let $\lambda_i := \lambda_\star(\widehat{m}_i)$ for $i = 1, 2$. Strong convexity yields, for any $\lambda, \lambda' \in \Lambda$ and any $s \in \partial F_{\widehat{m}_1}(\lambda)$,

$$F_{\widehat{m}_1}(\lambda') \geq F_{\widehat{m}_1}(\lambda) + s^\top(\lambda' - \lambda) + \frac{\alpha}{2}\|\lambda' - \lambda\|_2^2.$$

At $\lambda = \lambda_1$, first-order optimality gives $0 \in \partial F_{\widehat{m}_1}(\lambda_1)$, hence

$$F_{\widehat{m}_1}(\lambda_2) \geq F_{\widehat{m}_1}(\lambda_1) + \frac{\alpha}{2}\|\lambda_2 - \lambda_1\|_2^2. \tag{6.4}$$

Similarly,

$$F_{\widehat{m}_2}(\lambda_1) \geq F_{\widehat{m}_2}(\lambda_2) + \frac{\alpha}{2}\|\lambda_2 - \lambda_1\|_2^2. \tag{6.5}$$

Add (6.4) and (6.5). The terms $A(\lambda_i)$ and $\rho\|\lambda_i\|_*$ cancel, leaving

$$-(\widehat{m}_1)^\top \lambda_2 - (\widehat{m}_2)^\top \lambda_1 \geq -(\widehat{m}_1)^\top \lambda_1 - (\widehat{m}_2)^\top \lambda_2 + \alpha\|\lambda_2 - \lambda_1\|_2^2.$$

Rearrange:

$$(\lambda_2 - \lambda_1)^\top(\widehat{m}_2 - \widehat{m}_1) \geq \alpha\|\lambda_2 - \lambda_1\|_2^2.$$

Apply Cauchy–Schwarz:

$$\|\lambda_2 - \lambda_1\|_2 \, \|\widehat{m}_2 - \widehat{m}_1\|_2 \geq \alpha\|\lambda_2 - \lambda_1\|_2^2.$$

Division by $\|\lambda_2 - \lambda_1\|_2$ yields (6.3) when $\lambda_1 \neq \lambda_2$, and the bound holds trivially when $\lambda_1 = \lambda_2$. □

6.4. Lipschitz dependence on the radius

Theorem 6.5 (Sensitivity in ρ for Euclidean uncertainty)**.** *Assume 6.1. Fix a center $\widehat{m}$ and consider $\mathcal{U}_\rho = \widehat{m} + \rho\mathbb{B}_2$. Let $\lambda_\star(\rho)$ be the unique minimizer of $A(\lambda) - \widehat{m}^\top\lambda + \rho\|\lambda\|_2$. Suppose $\lambda_\star(\rho_1), \lambda_\star(\rho_2) \in \Lambda$. Then*

$$\|\lambda_\star(\rho_1) - \lambda_\star(\rho_2)\|_2 \leq \frac{1}{\alpha} |\rho_1 - \rho_2|. \tag{6.6}$$

Moreover, the map $\rho \mapsto \|\lambda_\star(\rho)\|_2$ is nonincreasing.

Proof. Write $F_\rho(\lambda) := A(\lambda) - \widehat{m}^\top\lambda + \rho\|\lambda\|_2$. The optimality condition at $\lambda_\star(\rho)$ is

$$0 = \nabla A(\lambda_\star(\rho)) - \widehat{m} + \rho\, s(\rho),$$

where $s(\rho) \in \partial\|\lambda_\star(\rho)\|_2$. For $\lambda \neq 0$, $s = \lambda/\|\lambda\|_2$, and for $\lambda = 0$, s can be any vector with Euclidean norm at most 1.

Assume $\rho_2 \geq \rho_1$; the opposite ordering follows by interchanging the indices. Let $\lambda_i := \lambda_\star(\rho_i)$ and select $s_i \in \partial\|\lambda_i\|_2$ satisfying

$$\nabla A(\lambda_i) - \widehat{m} + \rho_i s_i = 0, \qquad i = 1, 2. \tag{6.7}$$

Subtract (6.7):

$$\nabla A(\lambda_1) - \nabla A(\lambda_2) + \rho_1 s_1 - \rho_2 s_2 = 0.$$

Take inner product with $\lambda_1 - \lambda_2$:

$$(\lambda_1 - \lambda_2)^\top(\nabla A(\lambda_1) - \nabla A(\lambda_2)) = (\rho_2 s_2 - \rho_1 s_1)^\top(\lambda_1 - \lambda_2). \tag{6.8}$$

Strong convexity of A on Λ implies strong monotonicity of ∇A:

$$(\lambda_1 - \lambda_2)^\top(\nabla A(\lambda_1) - \nabla A(\lambda_2)) \geq \alpha\|\lambda_1 - \lambda_2\|_2^2. \tag{6.9}$$

For the right-hand side of (6.8), use the duality bound $|s^\top v| \leq \|s\|_2\|v\|_2 \leq \|v\|_2$ for any $s \in \partial\|\cdot\|_2$. Thus

$$|(\rho_2 s_2 - \rho_1 s_1)^\top(\lambda_1 - \lambda_2)| \leq (|\rho_2|\,\|s_2\|_2 + |\rho_1|\,\|s_1\|_2)\|\lambda_1 - \lambda_2\|_2 \leq (\rho_1 + \rho_2)\|\lambda_1 - \lambda_2\|_2.$$

A sharper bound uses

$$(\rho_2 s_2 - \rho_1 s_1)^\top(\lambda_1 - \lambda_2) = (\rho_2 - \rho_1)s_2^\top(\lambda_1 - \lambda_2) + \rho_1(s_2 - s_1)^\top(\lambda_1 - \lambda_2).$$

Monotonicity of the subdifferential of $\|\cdot\|_2$ gives $(s_2 - s_1)^\top(\lambda_2 - \lambda_1) \geq 0$, hence $\rho_1(s_2 - s_1)^\top(\lambda_1 - \lambda_2) \leq 0$. Therefore

$$(\rho_2 s_2 - \rho_1 s_1)^\top(\lambda_1 - \lambda_2) \leq (\rho_2 - \rho_1)s_2^\top(\lambda_1 - \lambda_2) \leq |\rho_2 - \rho_1|\,\|\lambda_1 - \lambda_2\|_2.$$

Combine with (6.8) and (6.9):

$$\alpha\|\lambda_1 - \lambda_2\|_2^2 \leq |\rho_2 - \rho_1|\,\|\lambda_1 - \lambda_2\|_2,$$

which yields (6.6).

For monotonicity of $\|\lambda_\star(\rho)\|_2$, take $\rho_2 > \rho_1$. Optimality gives

$$F_{\rho_1}(\lambda_\star(\rho_1)) \leq F_{\rho_1}(\lambda_\star(\rho_2)), \qquad F_{\rho_2}(\lambda_\star(\rho_2)) \leq F_{\rho_2}(\lambda_\star(\rho_1)).$$

Add the inequalities and cancel common terms:

$$\rho_1\|\lambda_\star(\rho_1)\|_2 + \rho_2\|\lambda_\star(\rho_2)\|_2 \leq \rho_1\|\lambda_\star(\rho_2)\|_2 + \rho_2\|\lambda_\star(\rho_1)\|_2,$$

which rearranges to $(\rho_2 - \rho_1)\|\lambda_\star(\rho_2)\|_2 \leq (\rho_2 - \rho_1)\|\lambda_\star(\rho_1)\|_2$ and yields $\|\lambda_\star(\rho_2)\|_2 \leq \|\lambda_\star(\rho_1)\|_2$. □

6.5. Distributional consequences

Lemma 6.6 (Bregman form of exponential-family KL)**.** *For any $\lambda, \lambda' \in \mathbb{R}^d$,*

$$\mathrm{KL}(P_\lambda\|P_{\lambda'}) = A(\lambda') - A(\lambda) - (\lambda' - \lambda)^\top\nabla A(\lambda). \tag{6.10}$$

Proof. By (3.3),

$$\log \frac{dP_\lambda}{dP_{\lambda'}}(x) = (\lambda - \lambda')^\top \phi(x) - \big(A(\lambda) - A(\lambda')\big).$$

Integrate under P_λ:

$$\begin{aligned} \mathrm{KL}(P_\lambda \| P_{\lambda'}) &= (\lambda - \lambda')^\top \int_{\mathcal{X}} \phi(x)\, P_\lambda(dx) - A(\lambda) + A(\lambda') \\ &= (\lambda - \lambda')^\top \nabla A(\lambda) - A(\lambda) + A(\lambda'). \end{aligned}$$

This is (6.10). □

Theorem 6.7 (Quadratic KL and total variation bounds)**.** *Assume 6.1 and let $\lambda, \lambda' \in \Lambda$. Then*

$$\frac{\alpha}{2}\|\lambda - \lambda'\|_2^2 \le \mathrm{KL}(P_\lambda \| P_{\lambda'}) \le \frac{R^2}{2}\|\lambda - \lambda'\|_2^2. \tag{6.11}$$

Consequently,

$$\mathrm{TV}(P_\lambda, P_{\lambda'}) \le \frac{R}{2}\|\lambda - \lambda'\|_2. \tag{6.12}$$

Proof. Using Lemma 6.6 and the integral remainder in Taylor's theorem,

$$A(\lambda') = A(\lambda) + (\lambda' - \lambda)^\top \nabla A(\lambda) + \int_0^1 (1-t)(\lambda' - \lambda)^\top \nabla^2 A(\lambda + t(\lambda' - \lambda))(\lambda' - \lambda)\, dt.$$

Substitute into (6.10) to obtain

$$\mathrm{KL}(P_\lambda \| P_{\lambda'}) = \int_0^1 (1-t)(\lambda' - \lambda)^\top \nabla^2 A(\lambda + t(\lambda' - \lambda))(\lambda' - \lambda)\, dt.$$

Since Λ is convex, $\lambda + t(\lambda' - \lambda) \in \Lambda$ for $t \in [0,1]$. Use $\nabla^2 A \succeq \alpha I_d$ on Λ to get

$$\mathrm{KL}(P_\lambda \| P_{\lambda'}) \ge \int_0^1 (1-t)\alpha \|\lambda' - \lambda\|_2^2\, dt = \frac{\alpha}{2}\|\lambda' - \lambda\|_2^2.$$

Use the global bound (6.1) to get the upper bound:

$$\mathrm{KL}(P_\lambda \| P_{\lambda'}) \le \int_0^1 (1-t)R^2 \|\lambda' - \lambda\|_2^2\, dt = \frac{R^2}{2}\|\lambda' - \lambda\|_2^2.$$

Pinsker's inequality implies $\mathrm{TV}(P_\lambda, P_{\lambda'}) \le \sqrt{\frac{1}{2}\,\mathrm{KL}(P_\lambda \| P_{\lambda'})}$ [20], and the upper KL bound yields (6.12). □

Corollary 6.8 (Forecast error under moment perturbations)**.** *Assume 6.1 and let $\widehat{m}_1, \widehat{m}_2$ be two centers with corresponding Euclidean-uncertainty solutions $\lambda_i = \lambda_\star(\widehat{m}_i)$ in Λ. Then for any bounded measurable $g : \mathcal{X} \to \mathbb{R}$,*

$$\left|\mathbb{E}_{P_{\lambda_1}}[g] - \mathbb{E}_{P_{\lambda_2}}[g]\right| \le \frac{R}{\alpha}\, \|g\|_\infty\, \|\widehat{m}_1 - \widehat{m}_2\|_2.$$

Proof. By (6.12) and Theorem 6.4,

$$\mathrm{TV}(P_{\lambda_1}, P_{\lambda_2}) \le \frac{R}{2}\|\lambda_1 - \lambda_2\|_2 \le \frac{R}{2\alpha}\|\widehat{m}_1 - \widehat{m}_2\|_2.$$

Then $|\mathbb{E}_P[g] - \mathbb{E}_Q[g]| \le 2\|g\|_\infty\, \mathrm{TV}(P, Q)$ gives the claim. □

Operational meaning of the stability bounds. Suppose two training windows produce moment centers separated by ε in Euclidean norm. Corollary 6.8 guarantees that every bounded forecast functional changes by at most $(R/\alpha)\|g\|_\infty\varepsilon$. For an event forecast $g = \mathbf{1}_B$, this directly controls the change in predicted event probability. The factor R measures feature scale, while $1/\alpha$ measures local conditioning of the exponential family. Standardizing features and avoiding nearly redundant coordinates improves the numerical value of this guarantee. The same argument converts a concentration rate for empirical moments into a generalization rate for bounded predictive summaries.

7. Finite-sample inference

7.1. Data model and empirical moments

Let $X_1, \dots, X_n$ be i.i.d. samples from an unknown distribution $P_0 \ll \mu$ with $\mathrm{KL}(P_0\|\mu) < \infty$. Define the population moment vector by $m_0 := \int_{\mathcal{X}} \phi(x)\, P_0(dx)$ and the empirical moments by

$$\widehat{m}_n := \frac{1}{n}\sum_{i=1}^{n} \phi(X_i).$$

We construct uncertainty sets $\mathcal{U}_n$ such that $m_0 \in \mathcal{U}_n$ with high probability.

7.2. Practical construction of data-driven uncertainty sets

A reproducible construction proceeds through four choices. First, select features whose scale and scientific meaning are stable; standardized coordinates make a common radius interpretable. Second, choose the geometry. Coordinate-wise bounds lead naturally to an ℓ_∞ box, while an estimated covariance matrix $\widehat{\Sigma}$ motivates an ellipsoid of the form

$$\mathcal{U}_n^{\mathrm{ell}} = \left\{u : (u - \widehat{m}_n)^\top (\widehat{\Sigma} + \eta I)^{-1}(u - \widehat{m}_n) \le c_{n,\delta}\right\},$$

with a small ridge $\eta > 0$ when the covariance estimate is ill-conditioned. Third, calibrate $c_{n,\delta}$ or a norm radius from a concentration inequality, an empirical-Bernstein bound, a valid bootstrap, an empirical-likelihood confidence region for the moment parameter [21], or a robust mean procedure. Fourth, intersect the resulting region with known structural restrictions and with the closed convex hull of the feature range whenever this hull is available. The radius records the uncertainty of the moment estimator; feature clipping or robust centers are appropriate when heavy tails or contamination affect the empirical moments.

7.3. Coordinate-wise confidence sets

Assume component-wise boundedness.

Assumption 7.1 (Coordinate bounds)**.** For each $j \in \{1, \dots, d\}$, there exist finite $a_j < b_j$ such that $a_j \le \phi_j(x) \le b_j$ for P_0-almost every $x \in \mathcal{X}$.

Theorem 7.2 (High-probability ℓ_∞ uncertainty set)**.** *Assume 7.1. Fix $\delta \in (0, 1)$. Define*

$$r_{n,\delta} := \max_{1\le j\le d}(b_j - a_j)\sqrt{\frac{\log(2d/\delta)}{2n}}.$$

Then with probability at least $1-\delta$,

$$\|\widehat{m}_n - m_0\|_\infty \le r_{n,\delta}.$$

Therefore the uncertainty set

$$\mathcal{U}_n^{(\infty)} := \left\{u \in \mathbb{R}^d : \ \|u - \widehat{m}_n\|_\infty \le r_{n,\delta}\right\}$$

satisfies $\mathbb{P}(m_0 \in \mathcal{U}_n^{(\infty)}) \ge 1-\delta$.

Proof. For fixed j, Hoeffding's inequality [22] gives

$$\mathbb{P}\Big(|(\widehat{m}_n)_j - (m_0)_j| \ge t\Big) \le 2\exp\left(-\frac{2nt^2}{(b_j-a_j)^2}\right).$$

Set $t = (b_j - a_j)\sqrt{\frac{\log(2d/\delta)}{2n}}$ to obtain a tail probability at most δ/d. Apply a union bound across j to obtain the ℓ_∞ claim. □

7.4. Robust maximum-entropy estimator and oracle entropy guarantee

Let $\widehat{P}_n$ be the unique minimizer of (2.2) with $\mathcal{U} = \mathcal{U}_n^{(\infty)}$ or any other data-driven compact convex set.

Theorem 7.3 (Feasibility and entropy oracle inequality)**.** *Assume 2.2 and let $\mathcal{U}_n$ be any random compact convex set such that $m_0 \in \mathcal{U}_n$ on an event $\mathcal{E}$. On $\mathcal{E}$, P_0 is feasible for the robust program with $\mathcal{U} = \mathcal{U}_n$, and the robust estimator satisfies*

$$\mathrm{KL}(\widehat{P}_n\|\mu) \le \mathrm{KL}(P_0\|\mu). \tag{7.1}$$

If Assumption 2.3 holds for $\mathcal{U}_n$ on $\mathcal{E}$, then on $\mathcal{E}$ there exists $\widehat{\lambda}_n$ minimizing the dual objective $A(\lambda) + \sigma_{\mathcal{U}_n}(-\lambda)$, and $\widehat{P}_n = P_{\widehat{\lambda}_n}$.

Proof. On $\mathcal{E}$, $m(P_0) = m_0 \in \mathcal{U}_n$, hence P_0 is feasible. Since $\widehat{P}_n$ minimizes $\mathrm{KL}(\cdot\|\mu)$ over the feasible set, (7.1) follows. If Assumption 2.3 holds, Theorem 5.1 applies conditionally on $\mathcal{E}$, giving dual attainment and the exponential-family representation. □

7.5. A finite-sample comparison to an oracle robust target

The next result compares the data-driven solution to an oracle solution that uses the same radius but is centered at the population moment m_0. This is a well-defined theoretical target.

Theorem 7.4 (Finite-sample stability around the oracle center)**.** *Assume 6.1. Fix $\rho \ge 0$ and let $\mathcal{U}_{\widehat{m}} := \widehat{m} + \rho\mathbb{B}_2$. Let $\lambda_\star(\widehat{m})$ be the unique minimizer of $A(\lambda) - \widehat{m}^\top\lambda + \rho\|\lambda\|_2$, and define $P^{\mathrm{rob}}(\widehat{m}) := P_{\lambda_\star(\widehat{m})}$. Suppose $\lambda_\star(\widehat{m}_n), \lambda_\star(m_0) \in \Lambda$. Then*

$$\mathrm{TV}(P^{\mathrm{rob}}(\widehat{m}_n), P^{\mathrm{rob}}(m_0)) \le \frac{R}{2\alpha}\,\|\widehat{m}_n - m_0\|_2,$$

and for any bounded measurable g,

$$\left|\mathbb{E}_{P^{\mathrm{rob}}(\widehat{m}_n)}[g] - \mathbb{E}_{P^{\mathrm{rob}}(m_0)}[g]\right| \le \frac{R}{\alpha}\,\|g\|_\infty\,\|\widehat{m}_n - m_0\|_2.$$

Proof. Apply Theorem 6.4 with $\widehat{m}_1 = \widehat{m}_n$ and $\widehat{m}_2 = m_0$ to obtain

$$\|\lambda_\star(\widehat{m}_n) - \lambda_\star(m_0)\|_2 \le \frac{1}{\alpha}\|\widehat{m}_n - m_0\|_2.$$

Then apply Theorem 6.7 to bound the total variation distance, and use $|\mathbb{E}_P[g] - \mathbb{E}_Q[g]| \le 2\|g\|_\infty \mathrm{TV}(P, Q)$. □

Corollary 7.5 (High-probability forecast stability)**.** *Under the assumptions of Theorems 7.2 and 7.4, with probability at least* $1 - \delta$*,*

$$\mathrm{TV}(P^{\mathrm{rob}}(\widehat{m}_n), P^{\mathrm{rob}}(m_0)) \le \frac{R\sqrt{d}}{2\alpha}\, r_{n,\delta},$$

and every bounded measurable g satisfies

$$\left|\mathbb{E}_{P^{\mathrm{rob}}(\widehat{m}_n)}[g] - \mathbb{E}_{P^{\mathrm{rob}}(m_0)}[g]\right| \le \frac{R\sqrt{d}}{\alpha}\,\|g\|_\infty r_{n,\delta}.$$

Thus the bounded-forecast discrepancy is of order $\sqrt{d\log(d/\delta)/n}$ *when the coordinate ranges remain uniformly bounded.*

Proof. The event in Theorem 7.2 implies $\|\widehat{m}_n - m_0\|_2 \le \sqrt{d}\, r_{n,\delta}$. Substitute this inequality into Theorem 7.4. □

8. Applications and illustrative examples

This section translates the abstract dual and stability theory into operational settings. The first two subsections explain how the robust maximum-entropy estimator can be deployed in probabilistic forecasting and in risk modeling. The last two subsections provide transparent finite-dimensional examples in which the entropy-maximizing distribution can be described explicitly. The common theme is that the uncertainty set acts as a statistically meaningful tolerance region for moments, while the entropy objective selects the least-committal model inside that region.

8.1. Probabilistic forecasting with uncertainty-aware feature matching

Suppose that a forecaster wishes to construct a predictive law for a target variable Y using a feature map $\phi(Y)$ that summarizes calibration, dispersion, or selected covariate-conditioned averages. In practice, these features are estimated from finite samples, rolling windows, or heterogeneous data feeds, so the relevant constraint is rarely an exact equality. The robust maximum-entropy program replaces exact feature matching by the requirement $m(P) \in \mathcal{U}$, where $\mathcal{U}$ is chosen from concentration bounds, bootstrap regions, or expert-specified tolerances.

For norm-ball uncertainty sets, the dual representation (5.5) shows that the forecasting problem is equivalent to fitting an exponential-family model with a support-function penalty. The penalty $\rho\|\lambda\|_*$ shrinks the dual parameter toward zero as uncertainty grows, which in turn pulls the predictive distribution toward the reference law μ. This mechanism is valuable in rolling or online settings: When the incoming feature estimates fluctuate because of sampling noise, the stability results of Section 6

imply that the predictive distribution changes in a controlled way rather than reacting erratically to small perturbations.

A practical forecasting workflow takes the following form:

1. choose the feature map ϕ to encode calibration or structural information that should be retained in the predictive law;
2. construct a data-driven uncertainty set $\mathcal{U}_n$ around the empirical feature vector $\widehat{m}_n$;
3. solve the dual problem to obtain $\widehat{\lambda}_n$ and the predictive distribution $P_{\widehat{\lambda}_n}$;
4. update the predictive law as new data arrive, using the sensitivity bounds to monitor forecast drift.

Proper scoring rules motivate this construction because entropy produces a canonical low-commitment baseline under partial information [17]. In applications where feature estimates are unstable or partially pooled across data sources, the uncertainty-set formulation aligns the inferential constraints with the actual statistical precision of the estimated summaries.

8.2. Risk modeling and scenario generation under moment ambiguity

In risk management, one typically estimates quantities such as $\mathbb{E}[L]$, $\mathbb{E}[L^2]$, or more general feature expectations $\mathbb{E}[\psi(L)]$ for a loss variable L. These quantities are then used to generate scenarios, benchmark stress distributions, or initialize downstream optimization procedures. Exact moment matching can overfit the estimated summaries, especially when the available sample is short, heavy filtering is used, or the loss process is nonstationary across time.

The robust maximum-entropy estimator provides a disciplined baseline distribution for scenario generation. The uncertainty region $\mathcal{U}$ captures estimation error in the moments, while entropy maximization spreads mass as evenly as possible subject to those admissible summaries. This is especially useful before evaluating functionals such as value-at-risk or conditional value-at-risk, since the scenario law itself reflects uncertainty in the estimated moments rather than treating those moments as exact inputs [18]. The resulting distribution can be interpreted as a conservative reference model inside the ambiguity set.

The theory developed earlier also supplies operational guarantees. Theorem 7.3 ensures that whenever the population moment vector lies inside the data-driven uncertainty set, the robust estimator incurs no larger relative entropy than the true data-generating law viewed as a feasible comparator. The sensitivity bounds in Theorems 6.4 and 6.7 provide quantitative control of scenario drift across adjacent estimation windows. These properties are relevant in distributionally robust optimization, where moment uncertainty sets already play a central role [10–13]. Within that broader literature, the present contribution identifies the maximum-entropy element of the ambiguity set and characterizes how it changes with the statistical inputs.

8.3. Relation to minimum-error entropy and exponential-cost learning

Minimum-error entropy (MEE) methods place entropy on the prediction residual and optimize a kernel estimate of a residual-entropy criterion. Recent analysis by Jiang et al. developed MEE learning under heavy-tailed noise and derived robustness through the residual distribution itself [23]. The present estimator places Shannon relative entropy on the candidate outcome law and introduces robustness through a confidence region for selected moments. The mechanisms therefore respond to different statistical objects. MEE is directly tailored to non-Gaussian and heavy-tailed residuals.

Moment uncertainty controls sampling and specification error in feature expectations. Heavy-tail protection in the present framework enters through bounded or clipped features, robust moment centers, and confidence sets justified for heavy-tailed data. A Euclidean or coordinate-wise radius around an ordinary sample mean alone carries the robustness properties of that mean estimator.

Sun, Zhou, and Huang studied learning with an exponential cost and obtained fast statistical rates under moment and capacity conditions [24]. The exponential-family expression in this paper has a distinct variational origin: $A(\lambda)$ is the cumulant-generating normalizer induced by Shannon relative entropy and linear moment constraints. The dual objective $A(\lambda) - \widehat{m}^\top \lambda + \rho\|\lambda\|_*$ is a penalized log-partition problem. It generally does not coincide with empirical minimization of an exponential residual loss. Comparable fast excess-risk rates would require a prediction class, a loss, and complexity or margin conditions. The present fixed-dimensional concentration argument yields the explicit root-n forecast rate in Corollary 7.5. These approaches can be combined by using robust residual-based features or robust moment estimators inside the uncertainty-set construction, followed by the maximum-entropy projection developed here.

8.4. Worked example: Bernoulli model with uncertain mean

Let $\mathcal{X} = \{0, 1\}$, let μ be uniform on $\mathcal{X}$, and let $\phi(x) = x$. Any admissible law P is determined by $p := P(X = 1) \in [0, 1]$, and the moment map is simply $m(P) = p$. If the estimated mean is $\widehat{p}$ with uncertainty radius ρ, then the admissible moment set is the interval $\mathcal{U} = [\widehat{p} - \rho, \widehat{p} + \rho]$, intersected with $[0, 1]$.

Proposition 8.1 (Entropy-driven shrinkage for Bernoulli)**.** *Let* $\mathcal{U} = [\widehat{p}-\rho, \widehat{p}+\rho]$. *The robust maximum-entropy solution is Bernoulli with parameter*

$$p_\star = \Pi_{\mathcal{U}\cap[0,1]}(1/2),$$

the Euclidean projection of $1/2$ *onto the feasible interval.*

Proof. The Bernoulli entropy $h(p) = -p \log p - (1 - p) \log(1 - p)$ is strictly concave on $(0, 1)$ and attains its unique maximum over $[0, 1]$ at $p = 1/2$. Maximizing a strictly concave function over the closed interval $\mathcal{U} \cap [0, 1]$ therefore returns $1/2$ whenever that value is feasible. When $1/2$ lies outside the feasible interval, strict concavity implies that the maximizer is the endpoint closest to $1/2$, which is exactly the Euclidean projection rule. □

This example makes the geometry of the method explicit. The entropy objective always prefers the most diffuse Bernoulli law, namely $p = 1/2$, and the uncertainty interval restricts how far the optimizer may move toward that diffuse point. As the uncertainty radius grows, the feasible set expands and the optimizer shrinks toward $1/2$. As the radius contracts, the optimizer tracks the estimated mean more closely. In this sense, the uncertainty region controls the strength of entropy-driven regularization.

8.5. Worked example: Categorical distribution with uncertain class frequencies

Let $\mathcal{X} = \{1, \ldots, k\}$ with reference measure μ uniform on $\mathcal{X}$, and let $\phi(x) = e_x \in \mathbb{R}^k$ denote the one-hot encoding. Then $m(P)$ is the probability vector $\pi = (\pi_1, \ldots, \pi_k)$ itself. Given empirical frequencies

$\widehat{\pi}$, consider the admissible set

$$\mathcal{U} := \left\{ \pi \in \mathbb{R}^k : \ \|\pi - \widehat{\pi}\|_\infty \le \rho, \ \sum_{i=1}^{k} \pi_i = 1, \ \pi_i \ge 0 \right\}.$$

The robust maximum-entropy problem becomes the search for the highest-entropy point inside a capped simplex polytope.

The structure is easy to interpret. If the uniform distribution $(1/k, \dots, 1/k)$ lies in $\mathcal{U}$, then it is the optimizer because it uniquely maximizes entropy over the full simplex. Otherwise the optimizer is the point in $\mathcal{U}$ that is closest, in the entropy sense, to the uniform law. The dual characterization identifies this optimizer through a penalized softmax representation, with the support function of $\mathcal{U}$ encoding the active faces of the polytope. This example is directly relevant for uncertainty-aware smoothing of empirical class frequencies, coarse multinomial forecasting, and scenario generation for discretized outcome spaces.

8.6. Interpretive summary of the application layer

Across these examples, the same pattern emerges. The reference measure μ supplies a baseline shape, the feature map ϕ specifies which summaries must be respected, and the uncertainty set $\mathcal{U}$ determines the admissible tolerance around estimated information. The primal problem then chooses the least-structured distribution compatible with those ingredients. The dual problem reveals the associated regularization, and the stability theory quantifies the robustness of the resulting law. This combination makes the framework suitable both for principled statistical modeling and for operational decision pipelines in which empirical constraints are inherently noisy.

9. Numerical study: Estimation error and calibration of ρ

We use the Bernoulli model from Proposition 8.1, where the robust estimator has the closed form

$$p_\rho = \Pi_{[\widehat{p}-\rho, \widehat{p}+\rho] \cap [0,1]}(1/2), \qquad \lambda_\rho = \log \frac{p_\rho}{1 - p_\rho}.$$

The true success probability is $p_0 = 0.70$, the sample size is $n = 100$, and each curve averages 20,000 Monte Carlo replications using a fixed random seed. The clean experiment uses i.i.d. Bernoulli(p_0) observations. The contaminated experiment independently replaces each observation by 1 with probability 0.15, producing an observed mean of 0.745 and a moment-bias scale of 0.045. For each radius we record coverage of p_0 by the uncertainty interval, mean absolute error of p_ρ, predictive divergence KL(Bern(p_0)||Bern(p_ρ)), and $|\lambda_\rho|$. For selected numerical values from Monte Carlo study, one may refer to Table 1.

Figure 1(a) shows the expected monotone increase of coverage with ρ. Panels (b) and (c) display the regularization tradeoff. Under clean sampling, exact matching has mean absolute error 0.0366 and mean predictive KL 0.0052. At $\rho = 0.045$, entropy shrinkage moves the estimator toward the reference value 1/2, giving mean absolute error 0.0529 and mean KL 0.0091. Under contamination, the same radius absorbs the leading moment bias: Mean absolute error decreases from 0.0526 to 0.0353, and mean KL decreases from 0.0111 to 0.0047. Thus a radius matched to plausible

moment misspecification can improve predictive fidelity, while an oversized radius produces additional shrinkage.

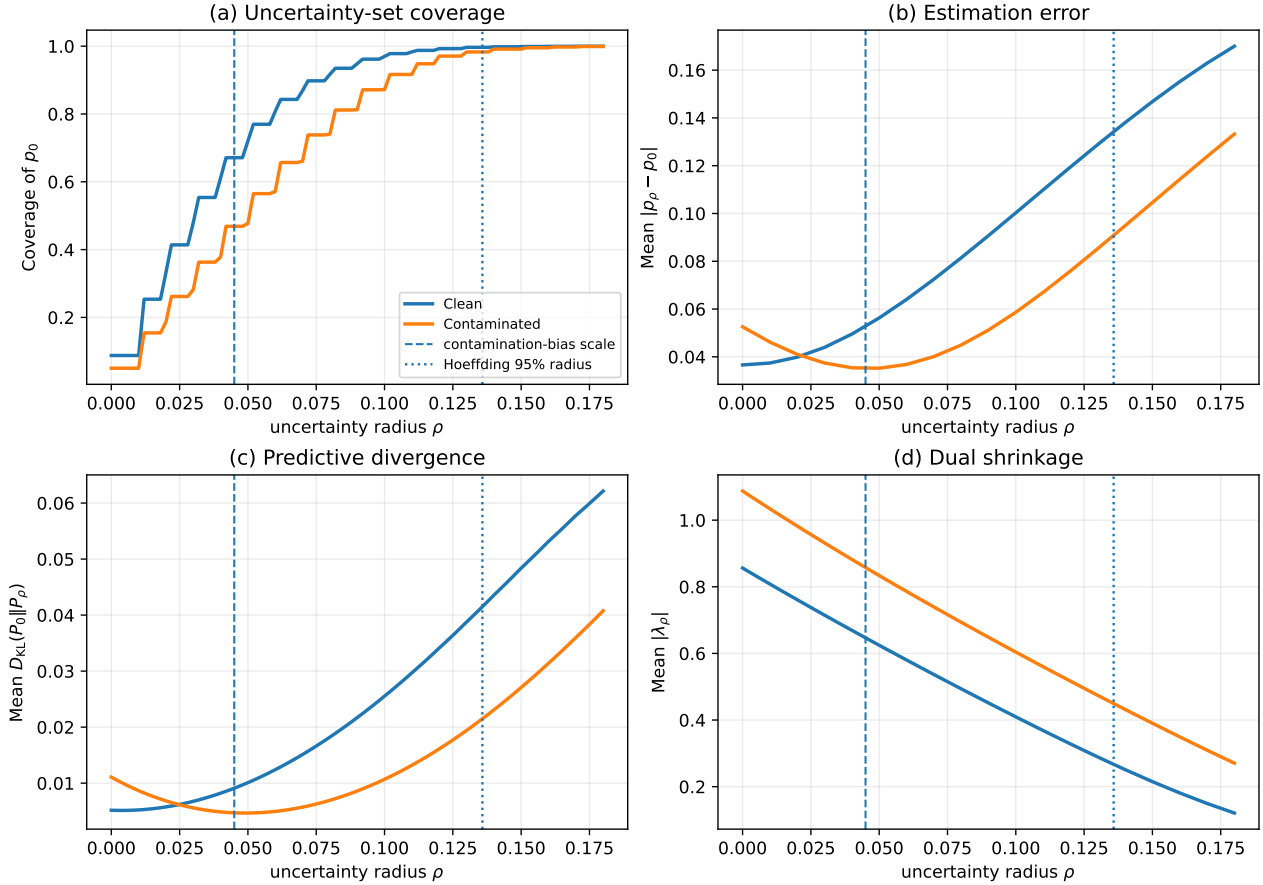

Figure 1. Monte Carlo relationship between the moment-uncertainty radius and inferential behavior. The dashed line is the contamination-bias scale 0.045; the dotted line is the coordinate-wise Hoeffding radius $\sqrt{\log(40)/(2n)} = 0.136$ for $\delta = 0.05$.

Table 1. Selected numerical values from the Monte Carlo study. Coverage refers to the event $|\widehat{p} - p_0| \le \rho$.

Radius choice	Sampling law	ρ	Coverage	Mean absolute error	Mean KL
Exact matching	Clean	0.000	0.088	0.0366	0.0052
Exact matching	Contaminated	0.000	0.050	0.0526	0.0111
Contamination-bias scale	Clean	0.045	0.671	0.0529	0.0091
Contamination-bias scale	Contaminated	0.045	0.469	0.0353	0.0047
Empirical 95% clean radius	Clean	0.090	0.962	0.0907	0.0217
Empirical 95% clean radius	Contaminated	0.090	0.871	0.0512	0.0086
Hoeffding 95% radius	Clean	0.136	0.997	0.1342	0.0415
Hoeffding 95% radius	Contaminated	0.136	0.983	0.0909	0.0215

The Hoeffding radius 0.136 provides coverage above 0.98 in both experiments and is conservative for this one-dimensional setting. Its larger regularization produces mean KL values 0.0415 and 0.0215 under clean and contaminated sampling, respectively. Panel (d) verifies the monotone decrease of the average dual magnitude as ρ grows, in agreement with Theorem 6.5. The numerical evidence supports a calibration principle: Choose ρ from a credible bound for moment-estimation and moment-specification error, and report sensitivity over nearby radii. Confidence coverage, predictive divergence, and dual shrinkage provide complementary diagnostics.

10. Conclusion and future work

We developed a robust maximum-entropy framework in which moment information is imposed through compact convex uncertainty sets rather than exact equalities. This formulation preserves the central variational interpretation of maximum entropy while incorporating statistical uncertainty in the constraints themselves. Under bounded features and a relative-interior feasibility condition, we established primal existence and uniqueness, derived a complete dual characterization with attainment, and showed that the optimizer remains an exponential-family distribution whose parameter is determined by a penalized log-partition problem. The support function of the uncertainty set enters the dual objective in a natural way and makes the geometry of the robustification transparent.

A second contribution is the quantitative stability theory. Under local strong convexity of the log-partition function, the optimal dual parameter varies Lipschitz-continuously with the empirical moment center and with the uncertainty radius. These parameter bounds translate into explicit controls on KL divergence, total variation distance, and bounded forecast errors. On the inferential side, data-driven uncertainty sets constructed from concentration inequalities yield high-probability feasibility and an entropy oracle inequality relative to the true data-generating law whenever the latter falls inside the admissible region. The application section illustrates how these abstract guarantees become meaningful in forecasting, risk modeling, and simple discrete examples where the entropy-shrinkage mechanism can be seen directly.

Several extensions are natural and mathematically interesting. One direction is to move beyond bounded features and treat unbounded or sub-exponential observables under coercivity or Orlicz-type integrability conditions. A second direction is to study high-dimensional regimes in which the number of moment constraints grows with the sample size and additional structural regularization becomes necessary. A third direction is to combine the present moment-uncertainty formalism with alternative reference divergences, sequential updating rules, or dependent data models. It would also be valuable to investigate uncertainty sets learned from resampling, empirical likelihood, or conformal calibration, and to analyze computational schemes for large-scale feature maps. These topics would broaden the scope of robust maximum entropy while preserving the central principle developed here: When information is uncertain, the admissible constraint set should reflect that uncertainty, and the selected distribution should remain maximally noncommittal within the statistically justified region.

Author contributions

B.S.A.: Validation, resources, writing—review and editing; A.A.: Validation, project administra tion, writing—review and editing; S.A.S.: Conceptualization, methodology, formal analysis, writing–original draft; J.G.D.: Validation, supervision, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The Researchers would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University (`www.qu.edu.sa`) for financial support (QU-APC-2026).

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.*, **27** (1948), 379–423, 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x
2. E. T. Jaynes, Information theory and statistical mechanics, *Phys. Rev.*, **106** (1957), 620–630. https://doi.org/10.1103/PhysRev.106.620
3. J. E. Shore, R. W. Johnson, Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy, *IEEE T. Inform. Theor.*, **26** (1980), 26–37. https://doi.org/10.1109/TIT.1980.1056144
4. I. Csiszár, I-divergence geometry of probability distributions and minimization problems, *Ann. Prob.*, **3** (1975),146–158.
5. I. Csiszár, Information projections revisited, *IEEE T. Inform. Theor.*, **49** (2003), 1474–1490. https://doi.org/10.1109/TIT.2003.810633
6. J. M. Borwein, A. S. Lewis, Duality relationships for entropy-like minimization problems, *SIAM J. Optim.*, **1** (1991), 191–205. https://doi.org/10.1137/0801014
7. C. Léonard, Minimization of entropy functionals, *J. Math. Anal. Appl.*, **346** (2008), 183–204. https://doi.org/10.1016/j.jmaa.2008.04.048
8. T. Sutter, D. Sutter, P. M. Esfahani, J. Lygeros, Generalized maximum entropy estimation, *J. Mach. Learn. Res.*, **20** (2019), 1–29.
9. A. Ben-Tal, L. El Ghaoui, A. Nemirovski, *Robust Optimization*, Princeton University Press, 2009.

10. D. Bertsimas, I. Popescu, Optimal inequalities in probability theory: A convex optimization approach, *SIAM J. Optim.*, **15** (2005), 780–804. https://doi.org/10.1137/S1052623401399903
11. E. Delage, Y. Ye, Distributionally robust optimization under moment uncertainty with application to data-driven problems, *Oper. Res.*, **58** (2010), 595–612. https://doi.org/10.1287/opre.1090.0741
12. H. Rahimian, S. Mehrotra, Frameworks and results in distributionally robust optimization, *Open J. Math. Optim.*, **3** (2022), 1–85. https://doi.org/10.5802/ojmo.15
13. D. Kuhn, S. Shafiee, W. Wiesemann, Distributionally robust optimization, *Acta Numer.*, **34** (2025), 579–804. https://doi.org/10.1017/S0962492924000084
14. P. M. Esfahani, D. Kuhn, Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations, *Math. Prog.*, **171** (2018), 115–166. https://doi.org/10.1007/s10107-017-1172-1
15. J. Blanchet, K. R. A. Murthy, Quantifying distributional model risk via optimal transport, *Math. Oper. Res.*, **44** (2019), 565–600. https://doi.org/10.1287/moor.2018.0936
16. S. Shafieezadeh-Abadeh, D. Kuhn, P. M. Esfahani, Regularization via mass transportation, *J. Mach. Learn. Res.*, **20** (2019), 1–68.
17. T. Gneiting, A. E. Raftery, Strictly proper scoring rules, prediction, and estimation, *J. Am. Stat. Assoc.*, **102** (2007), 359–378. https://doi.org/10.1198/016214506000001437
18. R. T. Rockafellar, S. Uryasev, Optimization of conditional value-at-risk, *J. Risk*, **2** (2000), 21–41. https://doi.org/10.21314/JOR.2000.038
19. R. T. Rockafellar, *Convex analysis*, Princeton University Press, 1970.
20. C. L. Canonne, A short note on an inequality between KL and TV, *arXiv preprint*, 2022, arXiv:2202.07198.
21. A. B. Owen, *Empirical likelihood*, Chapman & Hall/CRC, 2001.
22. W. Hoeffding, Probability inequalities for sums of bounded random variables, *J. Am. Stat. Assoc.*, **58** (1963), 13–30.
23. S. Jiang, H. Xu, Q. Sun, S. Huang, Robust learning of minimum error entropy under heavy-tailed noise, *J. Comput. Appl. Math.*, **487** (2026), 117686. https://doi.org/10.1016/j.cam.2026.117686
24. Q. Sun, Y. Zhou, S. Huang, Fast rates of exponential cost function, *Stat. Papers*, **66** (2025), 1–19. https://doi.org/10.1007/s00362-025-01672-3

AIMS Press