Varentropy: Overview, Computational Routes, and Structural Decomposition

Anatoly Vitold Stankyavichyus

✉ For correspondence: anatolyvitold.stankyavichyus@gmail.com

Abstract

Varentropy is the variance of information content, for a random variable or random vector with density f. Varentropy can be viewed in two complementary ways: as an intrinsic descriptor of distributional shape and as the variance of self-information, tied to concentration, typicality, and entropy estimation. This paper has two main aims. First, we give a concise overview of general tools for analyzing varentropy and organize several standard computational routes: direct integral formulas, cumulant and Rényi representations, a coarea representation of the law of information content, invariance under decreasing rearrangement, and simple sufficient conditions for finiteness. Second, for smooth parametric families we develop a structural decomposition of varentropy into tangent, nonlinear, score, and isoresidual components, and we derive necessary and sufficient criteria, stated directly in terms of the log-likelihood, score, and mixed derivatives, under which this decomposition simplifies. In particular, we identify an exact tangent regime characterized by closure under normalized density powers. A final section works out concrete examples, including Beta, Pareto-type, and Generalized Inverse Gaussian families, and highlights moving-support models as a particularly interesting direction beyond the present common-support framework.

Keywords: self-information, varentropy, entropy, risk management, heavy-tailed distributions.

I Introduction

Let X be an absolutely continuous random variable or random vector with density f relative to a reference measure μ. Its *information content* is the random variable

$$Y = -\log f(X),$$

its entropy is

$$h(X) = \mathbb{E}[-\log f(X)],$$

and its *varentropy* is

$$V_H(X) = \mathrm{Var}(-\log f(X)).$$

Following Song, one may regard varentropy as a function describing distributional shape rather than merely as an information-theoretic dispersion parameter [1]. That point of view is attractive for both conceptual and practical reasons. At the same time, the use of varentropy as a shape descriptor remains substantially less developed in the literature, and much less common in practice, than classical summaries such as kurtosis or tail index.

A classical way to summarize shape or tail behavior is through kurtosis or a tail index, but these summaries may be unavailable, unstable, or heavily model dependent. In heavy-tail practice there is no universally satisfactory automatic threshold rule for tail-index estimation, and different choices of the number of tail observations may lead to materially different conclusions [2].

In contrast with moment-based shape summaries, varentropy is available for a much broader class of continuous distributions [1]. It also exhibits a rigorous compatibility with more classical shape orderings. Di Crescenzo, Paolillo, and Suárez-Llorens showed that for symmetric unimodal random variables, kurtosis order implies varentropy order; combined with invariance under symmetric decreasing rearrangement, this yields the corresponding implication after rearrangement

as well [3]. In this sense, it shows that varentropy extends kurtosis-style shape comparisons into regimes where classical fourth-moment kurtosis may fail to exist while the information content remains square integrable.

Another attractive feature of varentropy is the availability of sharp universal bounds on broad geometric classes. In particular, Fradelizi, Li, and Madiman obtained sharp concentration inequalities for information content and corresponding varentropy bounds for s-concave measures, with log-concave laws as a distinguished special case [4]. From a practical standpoint, such results suggest one-sided diagnostic procedures: at the population level, a d-dimensional distribution with varentropy exceeding d cannot be log-concave.

A natural viewpoint for interpreting varentropy is through typicality and the asymptotic equipartition property. A particularly intriguing construction of asymptotic equipartition property based on Csiszar's generalized entropy and rectifiable measures are provided in [5], [6]. Informally, varentropy controls the second-order scale at which normalized self-information approaches entropy, and therefore how quickly long observed sequences become typical in the information-theoretic sense.

On the statistical side, Leonenko, Sun, and Taufer proposed a nearest-neighbor graph estimator for varentropy and established unbiasedness together with L^2-consistency [7]. This estimator can be viewed as an extention of the Kozachenko-Leonenko entropy estimator to varentropy.

Yet one important aspect remains underdeveloped: structural simplifications that permit relatively effortless derivations of varentropy. Even when a density is explicitly known, the computation of

$$V_H(X) = \mathrm{Var}(-\log f(X))$$

may be awkward. One can integrate directly, differentiate a cumulant-generating function, or analyze the induced law of $Y = -\log f(X)$, but the symbolic integrals involved are often unwieldy. The long-term objective of the present paper is to identify broad model classes for which varentropy can be analyzed through structural properties of the log-likelihood rather than through bespoke integration on a case-by-case basis.

More precisely, the new results of the paper center on a structural decomposition for parametric families $\{f_\theta\}$ of the form

$$V_H(\theta) = \nabla h(\theta)^\top I(\theta)^{-1} \nabla h(\theta) + \Delta_{\mathrm{nl}}(\theta) + \Delta_{\mathrm{score}}(\theta) + \Delta_{\mathrm{iso}}(\theta),$$

where $h(\theta) = \mathbb{E}_\theta[-\log f_\theta(X)]$ is the entropy functional and $I(\theta)$ is the Fisher information matrix. The first term is the tangent, Fisher-explainable component; the remaining terms measure nonlinear score effects, transverse score effects, and genuinely non-score-explainable fluctuations. Once the Fisher information is tractable, the remaining question can often be reduced to analytic on the original log-likelihood $\ell_\theta(x) = \log f_\theta(x)$, the score map $s_\theta(x)$, and their mixed x–θ derivatives. The novelty is not only the identity itself but the possibility of deciding whether it simplifies directly from the displayed form of the likelihood. Futhermore, whenever the standing assumptions hold, the decomposition yields a lower bound on varentropy. This lower bound may be viewed as a counterpart to the upper bound established in Corollary 4.4 of [4].

The present section also serves as a brief overview of the field. Section II collects the foundational material needed later: definitions, standing assumptions, the standard computational routes to varentropy, a coarea-based representation of the law of information content, and the invariance of information content under decreasing rearrangement. Section III develops the structural decomposition itself, Section IV derives practical simplification criteria, including derivative-based tests and the characterization of the exact tangent case by closure under normalized density powers, and Section V records worked examples together with a moving-support boundary case that marks the main direction in which the present common-support theory still needs to be extended. Throughout, we work under the standing assumption that the information content lies in L^2; subsection II-B records one easy-to-check sufficient criterion, while a sharper existence theory for varentropy is deferred to later work.

II Information Content and Standard Representations of Varentropy

II-A Information content, entropy, and varentropy

Let $(E, \mathcal{E}, \mu)$ be a σ-finite measure space, and let $f : E \to [0, \infty)$ be measurable with

$$\int_E f \, d\mu = 1.$$

We write

$$\rho(B) := \int_B f \, d\mu, \qquad B \in \mathcal{E},$$

for the induced probability measure, and we let $X \sim \rho$.

Definition II.1 (Information content, entropy, and varentropy)**.** The *information content* of X is the random variable

$$Y := -\log f(X),$$

with the convention $-\log 0 = +\infty$. Whenever $Y \in L^1(\rho)$ we define the entropy by

$$h(f) = h(X) := \mathbb{E}[Y] = -\int_E f \log f \, d\mu.$$

Whenever $Y \in L^2(\rho)$ we define the *varentropy* by

$$V_H(f) = V_H(X) := \mathrm{Var}(Y) = \mathbb{E}[(Y - h(X))^2].$$

For the moment we still take $Y \in L^2(\rho)$ as a standing hypothesis whenever varentropy is mentioned. The next subsection records one simple sufficient criterion for this integrability, but not an exhaustive characterization.

Proposition II.2 (Affine invariance)**.** *Let X be an absolutely continuous random vector in $\mathbb{R}^d$ with density f, and let $Z = AX + b$ where $A \in \mathbb{R}^{d\times d}$ is invertible and $b \in \mathbb{R}^d$. If g is the density of Z, then*

$$-\log g(Z) = -\log f(X) + \log|\det A|.$$

Consequently,

$$h(Z) = h(X) + \log|\det A|, \qquad V_H(Z) = V_H(X).$$

Proof. The density of Z is

$$g(z) = \frac{f(A^{-1}(z-b))}{|\det A|}.$$

Substituting $z = AX + b$ gives the identity for information content, and the formulas for entropy and varentropy follow immediately. □

II-B Simple sufficient conditions for finite varentropy

The later sections assume $Y \in L^2(\rho)$, so it is useful to record at least one workable sufficient criterion. For entropy itself, a convenient guide is given by the logarithmic-moment condition [8]: on $\mathbb{R}^d$, a finite first logarithmic moment controls the positive part of $-\log f(X)$, and if the density is essentially bounded then the negative part is controlled as well. Motivated by that observation, we record below a second-moment analogue for varentropy. The result is only sufficient, but its hypotheses are often easy to verify in concrete examples.

Proposition II.3 (A reference-density criterion)**.** *Let X be an absolutely continuous random vector in $\mathbb{R}^d$ with density f, and write $Y := -\log f(X)$. Assume that*

$$M := \|f\|_\infty < \infty.$$

Let g be another density on $\mathbb{R}^d$ such that

$$\mathbb{E}\left[(-\log g(X))^2\right] < \infty.$$

Then $Y \in L^2$, hence $h(X) \in \mathbb{R}$ and $V_H(X) < \infty$. More precisely,

$$\mathbb{E}[Y^2] \le 6\,\mathbb{E}\left[(-\log g(X))^2\right] + 4(\log M)^2 + \frac{8}{e^2}. \tag{1}$$

Proof. Let

$$A := \{x : f(x) \le g(x)\}, \qquad B := \{x : f(x) > g(x)\}.$$

Since

$$Y = -\log g(X) + \log\frac{g(X)}{f(X)},$$

we have

$$Y^2 \le 2(-\log g(X))^2 + 2\log^2\frac{g(X)}{f(X)}.$$

It therefore suffices to control the second term.

On A, write $t(x) := f(x)/g(x) \in [0,1]$. Then

$$\int_A f \log^2 \frac{g}{f}\, dx = \int_A g\, t \log^2 \frac{1}{t}\, dx \le \sup_{0<t\le 1} t \log^2 \frac{1}{t} = \frac{4}{e^2}.$$

On B one has $f \le M$ and $f > g$, hence $g < M$ and therefore

$$\log \frac{f}{g} \le \log \frac{M}{g} = \log M - \log g.$$

Using $(a+b)^2 \le 2a^2 + 2b^2$, we obtain

$$\log^2 \frac{f}{g} \le 2(\log M)^2 + 2(-\log g)^2,$$

and so

$$\int_B f \log^2 \frac{g}{f}\, dx = \int_B f \log^2 \frac{f}{g}\, dx \le 2(\log M)^2 + 2\,\mathbb{E}\big[(-\log g(X))^2\big].$$

Combining the estimates over A and B gives

$$\mathbb{E}\left[\log^2 \frac{g(X)}{f(X)}\right] \le \frac{4}{e^2} + 2(\log M)^2 + 2\,\mathbb{E}\big[(-\log g(X))^2\big].$$

Substituting this into the bound for Y^2 yields (1). Since ρ is a probability measure, $Y \in L^2$ implies $Y \in L^1$ by Cauchy–Schwarz, so both $h(X)$ and $V_H(X)$ are finite. □

Corollary II.4 (A logarithmic-moment criterion)**.** *Let X be an absolutely continuous random vector in $\mathbb{R}^d$ with density f. If*

$$\|f\|_\infty < \infty, \qquad \mathbb{E}\big[\log^2(1 + \|X\|)\big] < \infty,$$

then $Y = -\log f(X)$ belongs to L^2. In particular,

$$h(X) \in \mathbb{R}, \qquad V_H(X) < \infty.$$

Proof. Let

$$g_d(x) := c_d (1 + \|x\|)^{-(d+1)}, \qquad x \in \mathbb{R}^d,$$

where $c_d > 0$ normalizes g_d to integrate to one. Then

$$-\log g_d(X) = -\log c_d + (d+1)\log(1 + \|X\|),$$

so the assumed second logarithmic moment gives

$$\mathbb{E}\big[(-\log g_d(X))^2\big] < \infty.$$

Applying Proposition II.3 with $g = g_d$ proves the claim. □

These hypotheses are deliberately easy to check rather than sharp. In particular, the corollary shows that entropy is controlled by a first logarithmic moment, while varentropy is controlled by its second-moment analogue.

II-C Direct integral representation

The most immediate representation is obtained by direct expansion of the variance. Whenever $Y \in L^2$, one has

$$V_H(X) = \int_E f(x)\,[\log f(x)]^2\, d\mu(x) - \left(\int_E f(x)\, \log f(x)\, d\mu(x)\right)^2. \tag{2}$$

This formula is elementary but often computationally demanding. It is most useful when $f \log^2 f$ is explicitly integrable. However, it is rarely the case.

II-D The cumulant-generating and Rényi route

The second standard route is to work with the information-content random variable itself. For every t such that the integral below is finite,

$$M_Y(t) := \mathbb{E}[e^{tY}] = \int_E f(x)^{1-t}\, d\mu(x).$$

Hence the cumulant-generating function of Y is

$$K_Y(t) := \log M_Y(t) = \log \int_E f^{1-t}\, d\mu. \tag{3}$$

Whenever K_Y is twice differentiable at the origin,

$$h(X) = K_Y'(0), \qquad V_H(X) = K_Y''(0).$$

This representation is closely related to Rényi entropy. If

$$H_\alpha(f) := \frac{1}{1-\alpha} \log \int_E f^\alpha\, d\mu, \qquad \alpha \neq 1,$$

denotes the differential Rényi entropy of order α, then

$$K_Y(t) = t\, H_{1-t}(f), \qquad t \neq 0. \tag{4}$$

Thus varentropy may be recovered from the second-order behavior of the Rényi spectrum at $\alpha = 1$. This point of view goes back to Song's treatment of the Rényi information spectrum and its derivative-based shape functionals [1].

II-E A coarea representation for the law of information content

The previous formulas compute V_H without first identifying the law of $Y = -\log f(X)$. In some settings it is useful to do the opposite: compute the law of Y and then read off the first two moments. Under Sobolev regularity plus an additional structural hypothesis on the critical set, one can do this through the coarea formula, which is a standard tool of Geometric Measure Theory. To our knowledge, no coarea-based formula for information content Y specifically has been recorded in the literature. However, general applications in the context of the transformations of random variables have already been discussed [9].

Assume for this subsection that $E = \mathbb{R}^d$, that f is nonnegative, bounded, and belongs to $W^{1,1}_{\mathrm{loc}}(\mathbb{R}^d)$, and that we work with a precise representative of f in the sense of Malý, Swanson, and Ziemer [10]. Partition the domain into

$$R := \{x \in \mathbb{R}^d : |\nabla f(x)| > 0\}, \qquad P := \{x \in \mathbb{R}^d : |\nabla f(x)| = 0\}.$$

For $k \in \mathbb{R}$ define the regular level set

$$\Sigma_k^* := \{x \in \mathbb{R}^d : f(x) = e^{-k},\ |\nabla f(x)| > 0\}.$$

We also assume that, up to a Lebesgue-null set, the critical region P is the disjoint union of countably many plateaux $P_j \subset P$ on which $f \equiv e^{-k_j}$. Under this hypothesis the absolutely continuous part of the law of Y comes from the regular region R, while each plateau contributes an atom at the single information-content level k_j. The next two lemmas package the required coarea bookkeeping. In the sequel, $\mathcal{H}^d$ denotes d-dimensional Hausdorff measure. Note that, when d is a non-negative integer, d-dimensional Hausdorff and Lebesgue measures coincide, assuming, of course, that Hausdorff measure is defined with appropriate scaling factor.

Lemma II.5 (Superlevel-set identity)**.** *Let $u \in W^{1,1}_{\mathrm{loc}}(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$ be nonnegative and precisely represented, and write*

$$R_u := \{x \in \mathbb{R}^d : |\nabla u(x)| > 0\}, \qquad P_u := \{x \in \mathbb{R}^d : |\nabla u(x)| = 0\}.$$

Assume that, up to a Lebesgue-null set, P_u is the disjoint union of countably many plateaux $A_j \subset P_u$ on which $u \equiv t_j$. Define the superlevel function

$$\mu_u(t) := \lambda^d(\{x \in \mathbb{R}^d : u(x) > t\}), \qquad t > 0.$$

Then for every $0 < a < b < \operatorname{ess\,sup} u$ one has

$$\mu_u(a) - \mu_u(b) = \int_a^b \left(\int_{\{u=t\} \cap R_u} \frac{1}{|\nabla u(x)|}\, d\mathcal{H}^{d-1}(x) \right) dt + \sum_{t_j \in (a,b]} \lambda^d(A_j). \tag{5}$$

Consequently μ_u is absolutely continuous on every interval that avoids the plateau values $\{t_j\}$, and for Lebesgue-a.e. $t \notin \{t_j\}$,

$$-\mu_u'(t) = \int_{\{u=t\}\cap R_u} \frac{1}{|\nabla u(x)|}\, d\mathcal{H}^{d-1}(x). \tag{6}$$

Proof. Fix $0 < a < b < \operatorname{ess\,sup} u$. Up to a Lebesgue-null set,

$$\{a < u \le b\} = \big(\{a < u \le b\} \cap R_u\big) \dot\cup \bigcup_{t_j \in (a,b]} A_j.$$

Apply the Malý-Swanson-Ziemer coarea formula [10] on R_u with

$$g_{a,b}(x) := \frac{\mathbf{1}_{\{a<u(x)\le b\}\cap R_u}}{|\nabla u(x)|}.$$

Since $g_{a,b}(x)|\nabla u(x)| = \mathbf{1}_{\{a<u(x)\le b\}\cap R_u}$, this gives

$$\lambda^d\big(\{a < u \le b\} \cap R_u\big) = \int_a^b \left(\int_{\{u=t\}\cap R_u} \frac{1}{|\nabla u(x)|}\, d\mathcal{H}^{d-1}(x) \right) dt.$$

Adding the plateau contribution $\sum_{t_j\in(a,b]} \lambda^d(A_j)$ yields (5). The derivative formula (6) follows because the jump set $\{t_j\}$ is countable, so μ_u is absolutely continuous on compact subintervals of $(0, \operatorname{ess\,sup} u) \setminus \{t_j\}$. □

Lemma II.6 (Compositional coarea identity)**.** *Under the assumptions of Lemma II.5, let $g : [0, \operatorname{ess\,sup} u] \to \mathbb{R}$ be a bounded Borel function with $g(0) = 0$ and $\int_{\mathbb{R}^d} |g(u(x))|\, dx < \infty$. Then*

$$\int_{\mathbb{R}^d} g(u(x))\, dx = \int_0^{\operatorname{ess\,sup} u} g(t) \left(\int_{\{u=t\}\cap R_u} \frac{1}{|\nabla u(x)|}\, d\mathcal{H}^{d-1}(x) \right) dt + \sum_j g(t_j)\, \lambda^d(A_j). \tag{7}$$

Proof. By linearity it is enough to treat nonnegative g, and by monotone convergence it is then enough to prove the formula for simple functions. For an interval indicator $g = \mathbf{1}_{(a,b]}$ with $0 \le a < b \le \operatorname{ess\,sup} u$, the identity is exactly Equation (5). Finite linear combinations of such indicators yield all nonnegative simple functions, and an increasing approximation completes the proof. Signed bounded g follow by decomposing into positive and negative parts. □

Theorem II.7 (Coarea representation of the information-content law)**.** *Under the assumptions above, the law of $Y = -\log f(X)$ decomposes as*

$$\mathcal{L}(Y) = f_Y^{(ac)}(k)\, dk + \sum_j w_j\, \delta_{k_j}, \tag{8}$$

where

$$f_Y^{(ac)}(k) = e^{-2k} \int_{\Sigma_k^*} \frac{1}{|\nabla f(x)|}\, d\mathcal{H}^{d-1}(x) \tag{9}$$

for Lebesgue-a.e. k, and the atoms correspond to plateaux $P_j \subset P$ on which $f \equiv e^{-k_j}$, with mass

$$w_j = \int_{P_j} f(x)\, dx.$$

In one dimension this reduces to

$$f_Y^{(ac)}(k) = e^{-2k} \sum_{x_k \in \Sigma_k^*} \frac{1}{|f'(x_k)|}. \tag{10}$$

Proof. Let $\nu_Y := \mathcal{L}(Y)$ and fix a bounded Borel test function $\varphi : \mathbb{R} \to \mathbb{R}$. Define $g : [0, \operatorname{ess\,sup} f] \to \mathbb{R}$ by $g(0) := 0$ and

$$g(t) := t\, \varphi(-\log t), \qquad t > 0.$$

Then

$$\int_{\mathbb{R}} \varphi(k)\, \nu_Y(dk) = \int_{\mathbb{R}^d} \varphi(-\log f(x)) f(x)\, dx = \int_{\mathbb{R}^d} g(f(x))\, dx.$$

Applying Lemma II.6 with $u = f$ yields

$$\int_{\mathbb{R}} \varphi(k)\, \nu_Y(dk) = \int_0^{\operatorname{ess\,sup} f} \varphi(-\log t)\, t \left(\int_{\{f=t\}\cap R} \frac{1}{|\nabla f(x)|}\, d\mathcal{H}^{d-1}(x) \right) dt + \sum_j \varphi(k_j) e^{-k_j}\, \lambda^d(P_j).$$

In the first term, set $t = e^{-k}$, so $dt = -e^{-k}\,dk$ and $\{f = e^{-k}\} \cap R = \Sigma_k^*$. This gives

$$\int_{\mathbb{R}} \varphi(k)\, \nu_Y(dk) = \int_{\mathbb{R}} \varphi(k)\, e^{-2k} \left(\int_{\Sigma_k^*} \frac{1}{|\nabla f(x)|}\, d\mathcal{H}^{d-1}(x) \right) dk + \sum_j \varphi(k_j) w_j,$$

where $w_j = e^{-k_j} \lambda^d(P_j) = \int_{P_j} f(x)\,dx$. Since this identity holds for every bounded Borel φ, it identifies the law of Y as in Equations (8) and (9). In one dimension $\mathcal{H}^0$ is counting measure, so the absolutely continuous density reduces to equation (10). □

II-F Decreasing rearrangement and invariance of information content

The coarea representation is one route from f to the law of $Y = -\log f(X)$. A very different route is to reduce the problem to a one-dimensional monotone density by decreasing rearrangement. An excellent resource on the theory of rearrangement can be found in [11].

Definition II.8 (Decreasing rearrangement)**.** Let $(E, \mathcal{E}, \mu)$ be σ-finite and let $f : E \to [0, \infty]$ be measurable. Its distribution function is

$$\lambda_f(t) := \mu(\{x \in E : f(x) > t\}), \qquad t \geq 0.$$

The *decreasing rearrangement* of f is the function

$$f^*(s) := \inf\{t \geq 0 : \lambda_f(t) \leq s\}, \qquad s \geq 0.$$

Then f^* is the unique right-continuous nonincreasing function on $[0, \infty)$ such that

$$|\{s \geq 0 : f^*(s) > t\}| = \lambda_f(t), \qquad t \geq 0,$$

where $|\cdot|$ denotes Lebesgue measure.

Theorem II.9 (Information content is invariant under decreasing rearrangement)**.** *Let $(E, \mathcal{E}, \mu)$ be a σ-finite measure space, and let $f : E \to [0, \infty)$ be measurable with $\int_E f\,d\mu = 1$. Define the probability measures*

$$\rho(B) := \int_B f\,d\mu, \qquad \rho^*(A) := \int_A f^*(s)\,ds.$$

Then ρ^* is a probability measure. If

$$X \sim \rho, \qquad X^* \sim \rho^*,$$

then

$$f(X) \overset{d}{=} f^*(X^*).$$

Consequently,

$$-\log f(X) \overset{d}{=} -\log f^*(X^*),$$

with the convention $-\log 0 = +\infty$.

Proof. First,

$$\int_0^\infty f^*(s)\,ds = \int_0^\infty |\{s \geq 0 : f^*(s) > t\}|\,dt = \int_0^\infty \lambda_f(t)\,dt = \int_E f\,d\mu = 1,$$

so ρ^* is a probability measure.

Fix $t \geq 0$. Then

$$\mathbb{P}(f(X) > t) = \int_{\{f > t\}} f\,d\mu.$$

Decompose

$$f\,\mathbf{1}_{\{f>t\}} = (f - t)\mathbf{1}_{\{f>t\}} + t\mathbf{1}_{\{f>t\}},$$

and apply the layer-cake formula to $(f - t)\mathbf{1}_{\{f>t\}}$ to obtain

$$\int_{\{f>t\}} (f - t)\,d\mu = \int_t^\infty \lambda_f(v)\,dv.$$

Therefore

$$\mathbb{P}(f(X) > t) = t\,\lambda_f(t) + \int_t^\infty \lambda_f(v)\,dv. \tag{11}$$

The same calculation with f^* and Lebesgue measure yields

$$\mathbb{P}(f^*(X^*) > t) = t\,|\{f^* > t\}| + \int_t^\infty |\{f^* > v\}|\,dv.$$

Since f^* is the decreasing rearrangement of f,

$$|\{f^* > u\}| = \lambda_f(u), \qquad u \geq 0,$$

so by (11)

$$\mathbb{P}(f^*(X^*) > t) = \mathbb{P}(f(X) > t), \qquad t \geq 0.$$

Thus $f(X)$ and $f^*(X^*)$ have the same distribution. Composing with the Borel map $u \mapsto -\log u$ yields the conclusion. □

Corollary II.10 (Entropy and varentropy are rearrangement invariant)**.** *Whenever the relevant moments are finite,*

$$h(f) = h(f^*), \qquad V_H(f) = V_H(f^*).$$

In particular, any computation of entropy or varentropy may be reduced to the one-dimensional decreasing rearrangement.

Proof. By Theorem II.9, the information-content random variables $-\log f(X)$ and $-\log f^*(X^*)$ are equal in distribution. Their first and second moments therefore coincide whenever finite. □

II-G Monotone rearrangements and one-dimensional inversion

The rearrangement Theorem becomes especially practical when $f^* : [0,\infty) \to (0,\infty)$ is strictly decreasing. If $X^* \sim f^*(s)\,ds$ and $Y^* := -\log f^*(X^*)$, then the usual one-dimensional inversion gives

$$Y^* \leq y \iff X^* \leq (f^*)^{-1}(e^{-y}),$$

so for every y in the essential range of Y^*,

$$F_{Y^*}(y) = F_{X^*}\big((f^*)^{-1}(e^{-y})\big).$$

If moreover f^* is differentiable with $(f^*)'(s) \neq 0$, the ordinary change-of-variables formula yields

$$f_{Y^*}(y) = \frac{e^{-2y}}{\left|(f^*)'\big((f^*)^{-1}(e^{-y})\big)\right|}.$$

Thus, whenever the decreasing rearrangement is strictly monotone, the law of information content reduces to an explicit scalar inversion problem. By Theorem II.9, this one-dimensional formula reproduces the law of $-\log f(X)$ for the original density. It is the one-dimensional counterpart of Theorem II.7, with the geometry of the sample space compressed into the monotone rearrangement.

III Structural Decomposition of Varentropy in Parametric Families

We now pass from the density-level representations of Section 2 to the parametric setting that motivates the rest of the paper. The goal of this section is to show that varentropy admits a canonical decomposition into four nonnegative pieces:

(i) a tangent part explained by Fisher geometry,
(ii) a nonlinear effect already visible along the entropy direction,
(iii) an additional score-explainable part coming from score coordinates transverse to that direction, and
(iv) an iso-score remainder that is invisible even to the full score vector.

III-A Parametric setup and the entropy gradient identity

Let $\{f_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$ be a p-parameter family of densities on $(E, \mathcal{E}, \mu)$, where Θ is open. We assume throughout this section that:

(P1) the support of f_θ does not depend on θ;
(P2) for μ-a.e. x, the map $\theta \mapsto f_\theta(x)$ is continuously differentiable;
(P3) differentiation under the integral sign is justified for the quantities considered below;
(P4) the score vector belongs to $L^2(P_\theta)$ and the information content belongs to $L^2(P_\theta)$ for every θ under consideration;
(P5) the Fisher information matrix is positive definite.

Here P_θ denotes the probability law with density f_θ.

We write

$$\ell_\theta(x) := \log f_\theta(x), \qquad s_\theta(x) := \nabla_\theta \ell_\theta(x) \in \mathbb{R}^p,$$

and for $X \sim P_\theta$ we set

$$Y_\theta := -\ell_\theta(X), \qquad h(\theta) := \mathbb{E}_\theta[Y_\theta], \qquad V_H(\theta) := \mathrm{Var}_\theta(Y_\theta), \qquad S_\theta := s_\theta(X).$$

The Fisher information matrix is

$$I(\theta) := \mathbb{E}_\theta[S_\theta S_\theta^\top].$$

The first identity we need is that the gradient of entropy is the covariance between information content and the score.

Proposition III.1 (Entropy gradient identity)**.** *Under* (P1)–(P5)*,*

$$\mathbb{E}_\theta[S_\theta] = 0, \qquad \nabla h(\theta) = \mathbb{E}_\theta[Y_\theta S_\theta] = \mathrm{Cov}_\theta(Y_\theta, S_\theta).$$

Proof. The score identity $\mathbb{E}_\theta[S_\theta] = 0$ is standard and follows from $\int_E f_\theta \, d\mu = 1$. For the entropy gradient, write componentwise

$$h(\theta) = -\int_E f_\theta(x) \log f_\theta(x) \, d\mu(x).$$

Differentiating with respect to θ^j and using (P1)–(P4) gives

$$\begin{aligned} \partial_j h(\theta) &= -\int_E \partial_j f_\theta(x) \, \log f_\theta(x) \, d\mu(x) - \int_E \partial_j f_\theta(x) \, d\mu(x) \\ &= -\int_E f_\theta(x) \, s_{\theta,j}(x) \, \log f_\theta(x) \, d\mu(x), \end{aligned}$$

because $\int_E \partial_j f_\theta \, d\mu = \partial_j 1 = 0$. Since $Y_\theta = -\log f_\theta(X)$, this is

$$\partial_j h(\theta) = \mathbb{E}_\theta[Y_\theta S_{\theta,j}],$$

and stacking the components gives the stated vector identity. □

III-B Projection onto the score span

The finite-dimensional span of the score coordinates is the tangent space of the model inside $L^2(P_\theta)$. The first decomposition is simply the orthogonal projection of centered information content onto that tangent space.

Theorem III.2 (Tangent–residual decomposition)**.** *Define*

$$a_\theta := I(\theta)^{-1} \nabla h(\theta) \in \mathbb{R}^p.$$

Then there exists a unique random variable $r_\theta \in L^2(P_\theta)$ such that

$$Y_\theta - h(\theta) = a_\theta^\top S_\theta + r_\theta, \tag{12}$$

with

$$\mathbb{E}_\theta[r_\theta] = 0, \qquad \mathbb{E}_\theta[r_\theta S_\theta] = 0 \in \mathbb{R}^p.$$

Consequently,

$$V_H(\theta) = \underbrace{\nabla h(\theta)^\top I(\theta)^{-1} \nabla h(\theta)}_{=:\ \Delta_{\mathrm{lin}}(\theta)} + \underbrace{\mathrm{Var}_\theta(r_\theta)}_{=:\ \Delta_{\mathrm{res}}(\theta)}. \tag{13}$$

Proof. Set

$$r_\theta := Y_\theta - h(\theta) - a_\theta^\top S_\theta.$$

Since $\mathbb{E}_\theta[Y_\theta - h(\theta)] = 0$ and $\mathbb{E}_\theta[S_\theta] = 0$, we have $\mathbb{E}_\theta[r_\theta] = 0$. Moreover, by Proposition III.1,

$$\mathbb{E}_\theta[r_\theta S_\theta] = \mathbb{E}_\theta[(Y_\theta - h(\theta))S_\theta] - \mathbb{E}_\theta[S_\theta S_\theta^\top]a_\theta = \nabla h(\theta) - I(\theta)a_\theta = 0.$$

Thus r_θ is orthogonal to each score coordinate. Since $\mathbb{E}_\theta[r_\theta\, a_\theta^\top S_\theta] = a_\theta^\top \mathbb{E}_\theta[r_\theta S_\theta] = 0$, we obtain the Pythagorean identity

$$V_H(\theta) = \mathrm{Var}_\theta(a_\theta^\top S_\theta) + \mathrm{Var}_\theta(r_\theta).$$

Finally,

$$\mathrm{Var}_\theta(a_\theta^\top S_\theta) = a_\theta^\top I(\theta) a_\theta = \nabla h(\theta)^\top I(\theta)^{-1} \nabla h(\theta).$$

Uniqueness follows because any other decomposition with a score-orthogonal residual would yield a difference vector $b \in \mathbb{R}^p$ satisfying $I(\theta)b = 0$, and hence $b = 0$ by positive definiteness of $I(\theta)$. □

Corollary III.3 (Alignment coefficient)**.** *Whenever $V_H(\theta) > 0$, the ratio*

$$\rho_H^2(\theta) := \frac{\nabla h(\theta)^\top I(\theta)^{-1} \nabla h(\theta)}{V_H(\theta)} \in [0, 1] \tag{14}$$

measures the fraction of varentropy explained by the tangent score direction. In particular,

$$\Delta_{\mathrm{res}}(\theta) = V_H(\theta)\big(1 - \rho_H^2(\theta)\big).$$

Equality $\rho_H^2(\theta) = 1$ holds if and only if $r_\theta = 0$ almost surely.

Proof. The claim follows immediately from Theorem III.2 and the nonnegativity of both terms in (13). □

III-C Residual varentropy and the score sigma-field

The residual term in Theorem III.2 still contains two very different effects: one part is already explained by the full score vector but not by its linear projection $a_\theta^\top S_\theta$, while the other part is genuinely invisible even after conditioning on the score itself. The next Theorem separates these two contributions.

Theorem III.4 (Residual decomposition through the score)**.** *Let*

$$m_\theta(u) := \mathbb{E}_\theta[Y_\theta \mid S_\theta = u]$$

be a Borel version of the conditional mean of information content given the score. Then

$$r_\theta = \underbrace{\big(Y_\theta - m_\theta(S_\theta)\big)}_{=:\,\eta_\theta} + \underbrace{\big(m_\theta(S_\theta) - h(\theta) - a_\theta^\top S_\theta\big)}_{=:\,\zeta_\theta}, \tag{15}$$

and the two summands are orthogonal in $L^2(P_\theta)$. Consequently,

$$\Delta_{\mathrm{res}}(\theta) = \mathrm{Var}_\theta\big(m_\theta(S_\theta) - h(\theta) - a_\theta^\top S_\theta\big) + \mathbb{E}_\theta\big[\mathrm{Var}_\theta(Y_\theta \mid S_\theta)\big]. \tag{16}$$

Proof. By definition,

$$\mathbb{E}_\theta[\eta_\theta \mid S_\theta] = 0,$$

while ζ_θ is $\sigma(S_\theta)$-measurable. Therefore

$$\mathbb{E}_\theta[\eta_\theta \zeta_\theta] = \mathbb{E}_\theta\big[\zeta_\theta \mathbb{E}_\theta(\eta_\theta \mid S_\theta)\big] = 0.$$

Taking variances in (15) gives

$$\mathrm{Var}_\theta(r_\theta) = \mathrm{Var}_\theta(\eta_\theta) + \mathrm{Var}_\theta(\zeta_\theta).$$

Finally,

$$\mathrm{Var}_\theta(\eta_\theta) = \mathbb{E}_\theta\big[\mathrm{Var}_\theta(Y_\theta \mid S_\theta)\big]$$

by the conditional variance formula. □

Remark III.5. The second term in (16) is the genuinely non-score-explainable part of varentropy: it vanishes if and only if information content is measurable with respect to the full score vector. The first term is still score-explainable, but only nonlinearly.

III-D The full four-term decomposition

The score-explainable residual in Theorem III.4 itself contains two distinct pieces. One is already visible along the one-dimensional entropy direction

$$U_\theta := a_\theta^\top S_\theta,$$

while the other comes from score coordinates transverse to that direction. The next Theorem yields the full decomposition announced in the introduction.

Theorem III.6 (Full structural decomposition of varentropy)**.** *Let*

$$U_\theta := a_\theta^\top S_\theta, \qquad q_\theta(t) := \mathbb{E}_\theta[Y_\theta \mid U_\theta = t], \qquad m_\theta(u) := \mathbb{E}_\theta[Y_\theta \mid S_\theta = u].$$

Then

$$Y_\theta - h(\theta) = U_\theta + \big(q_\theta(U_\theta) - h(\theta) - U_\theta\big) + \big(m_\theta(S_\theta) - q_\theta(U_\theta)\big) + \big(Y_\theta - m_\theta(S_\theta)\big), \tag{17}$$

where U_θ represents linear entropy direction, $\big(q_\theta(U_\theta) - h(\theta) - U_\theta\big)$ is nonlinear entropy-direction effect, $\big(m_\theta(S_\theta) - q_\theta(U_\theta)\big)$ is transverse score effect, and $\big(Y_\theta - m_\theta(S_\theta)\big)$ is iso-score noise.
The four summands are pairwise orthogonal in $L^2(P_\theta)$. Consequently,

$$V_H(\theta) = \Delta_{\text{lin}}(\theta) + \Delta_{\text{nl}}(\theta) + \Delta_{\text{score}}(\theta) + \Delta_{\text{iso}}(\theta), \tag{18}$$

where

$$\Delta_{\text{lin}}(\theta) := \operatorname{Var}_\theta(U_\theta) = \nabla h(\theta)^\top I(\theta)^{-1} \nabla h(\theta), \tag{19}$$

$$\Delta_{\text{nl}}(\theta) := \operatorname{Var}_\theta\big(q_\theta(U_\theta) - h(\theta) - U_\theta\big), \tag{20}$$

$$\Delta_{\text{score}}(\theta) := \operatorname{Var}_\theta\big(m_\theta(S_\theta) - q_\theta(U_\theta)\big), \tag{21}$$

$$\Delta_{\text{iso}}(\theta) := \mathbb{E}_\theta\big[\operatorname{Var}_\theta(Y_\theta \mid S_\theta)\big]. \tag{22}$$

Proof. The decomposition (17) is an identity obtained by adding and subtracting the two conditional expectations $q_\theta(U_\theta)$ and $m_\theta(S_\theta)$. Pairwise orthogonality follows from repeated applications of the tower property:

- U_θ is orthogonal to $q_\theta(U_\theta) - h(\theta) - U_\theta$ because

$$\mathbb{E}_\theta\big[U_\theta q_\theta(U_\theta)\big] = \mathbb{E}_\theta[U_\theta Y_\theta] = a_\theta^\top \mathbb{E}_\theta[S_\theta Y_\theta] = a_\theta^\top \nabla h(\theta) = \operatorname{Var}_\theta(U_\theta),$$

  while $\mathbb{E}_\theta[U_\theta] = 0$.
- $m_\theta(S_\theta) - q_\theta(U_\theta)$ has zero conditional expectation given U_θ, so it is orthogonal to every U_θ-measurable term, in particular to the first two summands.
- $Y_\theta - m_\theta(S_\theta)$ has zero conditional expectation given S_θ, and hence is orthogonal to all $\sigma(S_\theta)$-measurable terms, including the first three summands.

Taking variances yields (18); the identification of Δ_{iso} with a conditional variance is again the standard law of total variance. □

Corollary III.7 (Score-explainable varentropy)**.** *The part of varentropy explained by the full score vector is*

$$\operatorname{Var}_\theta\big(\mathbb{E}_\theta[Y_\theta \mid S_\theta]\big) = \Delta_{\text{lin}}(\theta) + \Delta_{\text{nl}}(\theta) + \Delta_{\text{score}}(\theta), \tag{23}$$

and therefore

$$\Delta_{\text{lin}}(\theta) \le \operatorname{Var}_\theta\big(\mathbb{E}_\theta[Y_\theta \mid S_\theta]\big) \le V_H(\theta). \tag{24}$$

The gap between the rightmost two quantities is exactly $\Delta_{\text{iso}}(\theta)$.

Proof. By the law of total variance,

$$V_H(\theta) = \operatorname{Var}_\theta\big(\mathbb{E}_\theta[Y_\theta \mid S_\theta]\big) + \mathbb{E}_\theta\big[\operatorname{Var}_\theta(Y_\theta \mid S_\theta)\big].$$

Comparing with (18) gives (23). The inequalities are immediate because each Δ-term is nonnegative. □

Corollary III.8 (One-parameter specialization)**.** *If $p = 1$, then $I(\theta)$ is a scalar and*

$$\Delta_{\text{lin}}(\theta) = \frac{h'(\theta)^2}{I(\theta)}.$$

If in addition $h'(\theta) \neq 0$, then U_θ determines the scalar score S_θ and therefore

$$\Delta_{\text{score}}(\theta) = 0.$$

Thus, away from entropy-stationary points, the scalar decomposition reduces to

$$V_H(\theta) = \frac{h'(\theta)^2}{I(\theta)} + \Delta_{\text{nl}}(\theta) + \Delta_{\text{iso}}(\theta). \tag{25}$$

At points where $h'(\theta) = 0$, one has $\Delta_{\text{lin}}(\theta) = 0$ and $U_\theta \equiv 0$, so the transverse score term may survive:

$$V_H(\theta) = \Delta_{\text{score}}(\theta) + \Delta_{\text{iso}}(\theta). \tag{26}$$

Proof. When $p = 1$, $a_\theta = h'(\theta)/I(\theta)$ and hence

$$\Delta_{\text{lin}}(\theta) = a_\theta^2 I(\theta) = \frac{h'(\theta)^2}{I(\theta)}.$$

If $h'(\theta) \neq 0$, then $U_\theta = a_\theta S_\theta$ is a nonzero scalar multiple of the score itself, so $\sigma(U_\theta) = \sigma(S_\theta)$ and $m_\theta(S_\theta) = q_\theta(U_\theta)$ almost surely. Therefore $\Delta_{\text{score}}(\theta) = 0$. The remaining statements follow from Theorem III.6. □

IV Derivative-Based Criteria for Simplification

We now convert the abstract vanishing conditions from Section 3 into pointwise criteria on the original log-likelihood. Throughout this section we strengthen the regularity assumptions as follows: the common support of the family is an open, connected set $M \subset \mathbb{R}^d$, the map $(x, \theta) \mapsto \ell_\theta(x)$ is C^2 on $M \times \Theta$, and for the fixed parameter value under discussion we write

$$J_\theta(x) := D_x s_\theta(x) = \left[\partial_{x_i} \partial_{\theta_j} \ell_\theta(x)\right]_{j,i} \in \mathbb{R}^{p \times d}$$

for the mixed x–θ derivative matrix. Thus $J_\theta(x)$ is the Jacobian of the score map $x \mapsto s_\theta(x)$, and if

$$v_\theta(x) := a_\theta^\top s_\theta(x), \qquad U_\theta := v_\theta(X),$$

then

$$\nabla_x v_\theta(x) = J_\theta(x)^\top a_\theta.$$

The criteria below are exact once one adds the natural connected-fibre hypotheses that rule out global topological obstructions to factorization through the score map or through the one-dimensional entropy direction. Operationally, the weaker collapse

$$V_H(\theta) = \Delta_{\text{lin}}(\theta) + \Delta_{\text{nl}}(\theta)$$

should indeed be read as a first-order PDE / foliation problem in the data variable x: with $v_\theta(x) = a_\theta^\top s_\theta(x)$, one asks whether $-\ell_\theta$ factors through the scalar field v_θ. Equivalently, one asks whether $d_x \ell_\theta$ is everywhere colinear with $d_x v_\theta$, or, in geometric language, whether ℓ_θ is constant along the characteristic distribution $\ker D_x v_\theta$.

IV-A A deterministic factorization lemma

Lemma IV.1 (Factorization through a smooth map)**.** *Let $M \subset \mathbb{R}^d$ be open and connected, let $T : M \to \mathbb{R}^k$ be a C^1 map of constant rank, and let $\phi : M \to \mathbb{R}$ be C^1. Assume that every fibre $T^{-1}(y)$ is connected. Then the following are equivalent:*

(i) *there exists a Borel function $g : T(M) \to \mathbb{R}$ such that $\phi = g \circ T$ on M;*

(ii) *for every $x \in M$ and every $v \in \ker DT(x)$,*

$$D\phi(x)v = 0;$$

(iii) *for every $x \in M$,*

$$\nabla\phi(x) \in \text{Row}(DT(x)).$$

If $k = 1$, these are also equivalent to the existence of a scalar field $\lambda : M \to \mathbb{R}$ such that

$$\nabla\phi(x) = \lambda(x)\, \nabla T(x), \qquad x \in M.$$

Proof. The implication (i)⇒(ii) is immediate: if $\phi = g \circ T$, then ϕ is constant on each fibre of T, so every directional derivative tangent to a fibre — equivalently, every derivative along $v \in \ker DT(x)$ — must vanish.

The equivalence (ii)⇔(iii) is elementary linear algebra: for any matrix A, the annihilator of $\ker A$ is exactly the row space of A. Thus (ii) says precisely that $\nabla\phi(x)$ belongs to the row space of $DT(x)$ for every x.

It remains to show (ii)⇒(i). Fix $x_0 \in M$ and let $r = \operatorname{rank} DT$. By the constant-rank Theorem there are local coordinates around x_0 in which T becomes the projection onto the first r coordinates. In those coordinates the vectors tangent to the fibres are exactly the coordinate vectors in the remaining $d - r$ directions. Condition (ii) therefore forces ϕ to be locally independent of the fibre coordinates, so ϕ is locally a function of T. Since the fibres are connected, these local representatives agree along each fibre, hence ϕ is globally constant on each fibre. One may therefore define $g(y)$ to be the common value of ϕ on the fibre $T^{-1}(y)$, which gives $\phi = g \circ T$ on M. When $k = 1$, the row space of $DT(x)$ is one-dimensional and is generated by $\nabla T(x)$, so (iii) is equivalent to the existence of the scalar multiplier $\lambda(x)$. □

Remark IV.2. Without the connected-fibre hypothesis, the derivative conditions in Lemma IV.1 still characterize local factorization through T, but one can lose a globally single-valued representation on $T(M)$. The connected-fibre assumption is exactly what turns the local criterion into a global one.

IV-B When the iso-score term vanishes

Theorem IV.3 (Necessary and sufficient conditions for $\Delta_{\mathrm{iso}}(\theta) = 0$)**.** *Fix $\theta \in \Theta$. Then the following are equivalent:*

(i) $\Delta_{\mathrm{iso}}(\theta) = 0$*;*

(ii) *there exists a Borel function $\psi_\theta : s_\theta(M) \to \mathbb{R}$ such that*

$$-\ell_\theta(x) = \psi_\theta\big(s_\theta(x)\big) \qquad \text{for } P_\theta\text{-a.e. } x.$$

Assume in addition that the score map $s_\theta : M \to \mathbb{R}^p$ has constant rank and connected fibres. Then (i) *and* (ii) *are also equivalent to each of*

the following pointwise conditions:

(iii) *for every $x \in M$ and every $v \in \ker J_\theta(x)$,*

$$\nabla_x \ell_\theta(x)^\top v = 0;$$

(iv) *for every $x \in M$,*

$$\nabla_x \ell_\theta(x) \in \operatorname{Row}(J_\theta(x)).$$

Equivalently, the information content is score-determined if and only if the x-gradient of ℓ_θ annihilates every direction along which the full score vector stays constant, or, what is the same, if and only if that gradient lies in the row space of the mixed derivative matrix $J_\theta(x) = \big[\partial_{x_i}\partial_{\theta_j}\ell_\theta(x)\big]_{j,i}$.

Proof. By (22),

$$\Delta_{\mathrm{iso}}(\theta) = \mathbb{E}_\theta\big[\operatorname{Var}_\theta(Y_\theta \mid S_\theta)\big].$$

Since the conditional variance is nonnegative, $\Delta_{\mathrm{iso}}(\theta) = 0$ if and only if $\operatorname{Var}_\theta(Y_\theta \mid S_\theta) = 0$ almost surely, which is equivalent to Y_θ being $\sigma(S_\theta)$-measurable. Because $Y_\theta = -\ell_\theta(X)$ and $S_\theta = s_\theta(X)$, this is equivalent to the existence of a Borel function ψ_θ such that $-\ell_\theta(x) = \psi_\theta(s_\theta(x))$ for P_θ-a.e. x. This proves (i)⇔(ii).

Assume now that s_θ has constant rank and connected fibres. We first show that (ii) implies (iii). Because $f_\theta > 0$ on the open support M, statement (ii) is equivalent to the existence of a Borel function ψ_θ such that

$$-\ell_\theta(x) = \psi_\theta\big(s_\theta(x)\big) \qquad \text{for Lebesgue-a.e. } x \in M.$$

Fix $x_0 \in M$ and let $r = \operatorname{rank} J_\theta(x_0)$; by constant rank, this value is independent of x_0. By the constant-rank theorem, after shrinking neighbourhoods there exist open sets $\Omega \subset M$ and $W \subset \mathbb{R}^p$ together with C^1 coordinate maps

$$\Phi : \Omega \to U \times V \subset \mathbb{R}^r \times \mathbb{R}^{d-r}, \qquad \Psi : W \to \widetilde{W} \subset \mathbb{R}^r \times \mathbb{R}^{p-r},$$

such that $x_0 \in \Omega$, $s_\theta(\Omega) \subset W$, and

$$(\Psi \circ s_\theta \circ \Phi^{-1})(u, z) = (u, 0).$$

Writing

$$F := -\ell_\theta \circ \Phi^{-1}, \qquad \widetilde{\psi}_\theta(u) := \psi_\theta\big(\Psi^{-1}(u,0)\big),$$

we obtain

$$F(u,z) = \widetilde{\psi}_\theta(u) \qquad \text{for Lebesgue-a.e. } (u,z) \in U \times V.$$

By Fubini, for almost every $u \in U$ the function $z \mapsto F(u,z)$ equals the constant $\widetilde{\psi}_\theta(u)$ for almost every $z \in V$. Since F is continuous, $F(u,\cdot)$ is then constant on V for almost every u. Let $A \subset U$ be the full-measure set of such u. Because A is dense in U, for any $u \in U$ and any $z_1, z_2 \in V$ we may choose $u_n \in A$ with $u_n \to u$; continuity gives

$$F(u,z_1) = \lim_{n\to\infty} F(u_n,z_1) = \lim_{n\to\infty} F(u_n,z_2) = F(u,z_2).$$

Thus F is locally independent of the fibre coordinates. Equivalently, (iii) holds on Ω. Because x_0 was arbitrary, (iii) holds on all of M.

The equivalence (iii)⇔(iv) is the same linear-algebra statement as in Lemma IV.1: the annihilator of $\ker J_\theta(x)$ is $\mathrm{Row}(J_\theta(x))$.

Conversely, if (iii) or (iv) holds, then Lemma IV.1 applied with $T = s_\theta$ and $\phi = -\ell_\theta$ yields a pointwise factorization

$$-\ell_\theta(x) = g_\theta\big(s_\theta(x)\big), \qquad x \in M,$$

for some Borel function $g_\theta : s_\theta(M) \to \mathbb{R}$. This implies (ii). Hence (i)–(iv) are equivalent. □

IV-C Reduction to the entropy score direction

Theorem IV.4 (Necessary and sufficient conditions for $V_H(\theta) = \Delta_{\mathrm{lin}}(\theta) + \Delta_{\mathrm{nl}}(\theta)$)**.** *Fix $\theta \in \Theta$ and define the scalar score coordinate*

$$v_\theta(x) := a_\theta^\top s_\theta(x), \qquad U_\theta = v_\theta(X).$$

Then the following are equivalent:

(i) $\Delta_{\mathrm{score}}(\theta) = 0$ *and* $\Delta_{\mathrm{iso}}(\theta) = 0$*;*

(ii) *there exists a Borel function $\varphi_\theta : v_\theta(M) \to \mathbb{R}$ such that*

$$-\ell_\theta(x) = \varphi_\theta\big(v_\theta(x)\big) \qquad \textit{for } P_\theta\textit{-a.e. } x.$$

Equivalently, the full four-term decomposition collapses to

$$V_H(\theta) = \Delta_{\mathrm{lin}}(\theta) + \Delta_{\mathrm{nl}}(\theta)$$

if and only if the information content is already a deterministic function of the one-dimensional entropy direction $a_\theta^\top s_\theta(x)$.

Assume in addition that the map $v_\theta : M \to \mathbb{R}$ has constant rank and connected fibres. Then (i) *and* (ii) *are also equivalent to each of the following pointwise conditions:*

(iii) *for every $x \in M$ and every $v \in \ker D_x v_\theta(x)$,*

$$\nabla_x \ell_\theta(x)^\top v = 0;$$

(iv) *there exists a scalar field $\lambda_\theta : M \to \mathbb{R}$ such that*

$$\nabla_x \ell_\theta(x) = \lambda_\theta(x)\, J_\theta(x)^\top a_\theta, \qquad x \in M.$$

Proof. From Theorem III.6,

$$\Delta_{\mathrm{score}}(\theta) = \mathrm{Var}_\theta\big(m_\theta(S_\theta) - q_\theta(U_\theta)\big), \qquad \Delta_{\mathrm{iso}}(\theta) = \mathbb{E}_\theta\big[\mathrm{Var}_\theta(Y_\theta \mid S_\theta)\big].$$

Hence (i) holds if and only if both

$$Y_\theta = m_\theta(S_\theta) \qquad \text{and} \qquad m_\theta(S_\theta) = q_\theta(U_\theta)$$

almost surely, which is equivalent to Y_θ being $\sigma(U_\theta)$-measurable. Since $Y_\theta = -\ell_\theta(X)$ and $U_\theta = v_\theta(X)$, this is exactly (ii).

Assume now that v_θ has constant rank and connected fibres. Because $f_\theta > 0$ on the open support M, statement (ii) is equivalent to the existence of a Borel function φ_θ such that

$$-\ell_\theta(x) = \varphi_\theta\big(v_\theta(x)\big) \qquad \text{for Lebesgue-a.e. } x \in M.$$

Fix $x_0 \in M$ and let $r = \operatorname{rank} D_x v_\theta(x_0) \in \{0, 1\}$; by constant rank, this value is independent of x_0. By the constant-rank theorem, after shrinking neighbourhoods there exists a C^1 coordinate map

$$\Phi : \Omega \to U \times V \subset \mathbb{R}^r \times \mathbb{R}^{d-r}$$

with $x_0 \in \Omega$ such that

$$(v_\theta \circ \Phi^{-1})(u, z) = \widetilde{v}(u)$$

for some C^1 map $\widetilde{v} : U \to \mathbb{R}$. Writing

$$F := -\ell_\theta \circ \Phi^{-1}, \qquad \widetilde{\varphi}_\theta(u) := \varphi_\theta\big(\widetilde{v}(u)\big),$$

we obtain

$$F(u, z) = \widetilde{\varphi}_\theta(u) \qquad \text{for Lebesgue-a.e. } (u, z) \in U \times V.$$

By Fubini, for almost every $u \in U$ the function $z \mapsto F(u, z)$ equals the constant $\widetilde{\varphi}_\theta(u)$ for almost every $z \in V$. Since F is continuous, $F(u, \cdot)$ is then constant on V for almost every u. Let $A \subset U$ be the full-measure set of such u. Because A is dense in U, for any $u \in U$ and any $z_1, z_2 \in V$ we may choose $u_n \in A$ with $u_n \to u$; continuity gives

$$F(u, z_1) = \lim_{n\to\infty} F(u_n, z_1) = \lim_{n\to\infty} F(u_n, z_2) = F(u, z_2).$$

Thus F is locally independent of the fibre coordinates. Equivalently, (iii) holds on Ω. Because x_0 was arbitrary, (iii) holds on all of M.

The scalar part of Lemma IV.1, applied with $T = v_\theta$ and $\phi = -\ell_\theta$, shows that (iii) and (iv) are equivalent. The same lemma then yields a pointwise factorization

$$-\ell_\theta(x) = g_\theta\big(v_\theta(x)\big), \qquad x \in M,$$

for some Borel function $g_\theta : v_\theta(M) \to \mathbb{R}$. This implies (ii). Hence (i)–(iv) are equivalent. □

IV-D Exact tangent explainability and pure score explainability

Theorem IV.5 (Necessary and sufficient conditions for $V_H(\theta) = \Delta_{\text{lin}}(\theta)$)**.** *Fix $\theta \in \Theta$. Then the following are equivalent:*

(i) $V_H(\theta) = \Delta_{\text{lin}}(\theta)$*;*

(ii)

$$-\ell_\theta(x) = h(\theta) + a_\theta^\top s_\theta(x) \qquad \textit{for } P_\theta\textit{-a.e. } x.$$

Under the standing C^2 and connected-support assumptions of this section, these are also equivalent to the pointwise mixed-derivative identity

$$\nabla_x \ell_\theta(x) + J_\theta(x)^\top a_\theta = 0, \qquad x \in M. \tag{27}$$

Equivalently,

$$\partial_{x_i} \ell_\theta(x) + \sum_{j=1}^{p} a_{\theta,j}\, \partial_{x_i} \partial_{\theta_j} \ell_\theta(x) = 0, \qquad i = 1, \dots, d.$$

Proof. By Theorem III.6,

$$V_H(\theta) = \Delta_{\text{lin}}(\theta)$$

if and only if the three nonnegative remainder terms $\Delta_{\text{nl}}(\theta)$, $\Delta_{\text{score}}(\theta)$, and $\Delta_{\text{iso}}(\theta)$ all vanish. By (17), this is equivalent to

$$Y_\theta - h(\theta) = U_\theta = a_\theta^\top S_\theta \qquad \text{almost surely,}$$

which is exactly statement (ii).

Now assume (ii). The function

$$F_\theta(x) := \ell_\theta(x) + a_\theta^\top s_\theta(x) + h(\theta)$$

is continuous on the open support M and vanishes P_θ-almost surely. Because $f_\theta > 0$ on M, any point at which F_θ were nonzero would have an open neighborhood of positive P_θ-probability on which F_θ kept the same sign, a contradiction. Hence $F_\theta \equiv 0$ on M, and differentiating with respect to x gives (27).

Conversely, if (27) holds, then the gradient of F_θ vanishes identically on the connected set M. Therefore F_θ is constant on M. Taking P_θ-expectations and using $\mathbb{E}_\theta[S_\theta] = 0$ shows that this constant must be zero, so statement (ii) follows. □

Corollary IV.6 (Necessary and sufficient conditions for $V_H(\theta) = \Delta_{\text{score}}(\theta)$)**.** *Fix $\theta \in \Theta$. Then the following are equivalent:*

(i) $V_H(\theta) = \Delta_{\text{score}}(\theta)$*;*
(ii) $\nabla h(\theta) = 0$ *and* $\Delta_{\text{iso}}(\theta) = 0$*;*
(iii) $\nabla h(\theta) = 0$ *and there exists a Borel function $\psi_\theta : s_\theta(M) \to \mathbb{R}$ such that*

$$-\ell_\theta(x) = \psi_\theta\big(s_\theta(x)\big) \qquad \text{for } P_\theta\text{-a.e. } x.$$

If, in addition, the score map s_θ has constant rank and connected fibres, then these are also equivalent to the derivative condition

$$\nabla h(\theta) = 0, \qquad \nabla_x \ell_\theta(x) \in \mathrm{Row}(J_\theta(x)) \quad \text{for every } x \in M.$$

Proof. If $\nabla h(\theta) = 0$, then $a_\theta = I(\theta)^{-1}\nabla h(\theta) = 0$, hence $U_\theta \equiv 0$. Therefore

$$\Delta_{\text{lin}}(\theta) = 0, \qquad \Delta_{\text{nl}}(\theta) = \mathrm{Var}_\theta\big(q_\theta(0) - h(\theta)\big) = 0.$$

So in this case the four-term decomposition reduces to

$$V_H(\theta) = \Delta_{\text{score}}(\theta) + \Delta_{\text{iso}}(\theta).$$

Thus (i) holds if and only if $\nabla h(\theta) = 0$ and $\Delta_{\text{iso}}(\theta) = 0$, proving (i)⇔(ii). The equivalence (ii)⇔(iii) and the final derivative criterion now follow immediately from Theorem IV.3. □

Remark IV.7. In a one-parameter family, Corollary III.8 already showed that $\Delta_{\text{score}}(\theta) = 0$ whenever $h'(\theta) \neq 0$. The criterion in Corollary IV.6 therefore says that a genuinely nonzero transverse score term can only occur at entropy-stationary parameter values, where the entropy direction itself collapses.

IV-E Characterization by normalized density powers

For a positive density g and an exponent $\beta > 0$ such that $g^\beta \in L^1(M)$, define its normalized density-power transform (also called the escort transform) by

$$\mathcal{E}_\beta[g](x) := \frac{g(x)^\beta}{\int_M g(y)^\beta \, dy}, \qquad x \in M.$$

The exact tangent identity turns out to be equivalent to local closure of the family under these transforms along the entropy vector field $a_\theta = I(\theta)^{-1}\nabla h(\theta)$.

Theorem IV.8 (Characterization by normalized density powers)**.** *Let $\Theta_0 \subset \Theta$ be open and assume that the map $\theta \mapsto a_\theta = I(\theta)^{-1}\nabla h(\theta)$ is C^1 on Θ_0. Let Φ_t denote the local flow of the vector field a. Then the following are equivalent:*

(i) $V_H(\vartheta) = \Delta_{\text{lin}}(\vartheta)$ *for every $\vartheta \in \Theta_0$;*
(ii) *for every $\theta \in \Theta_0$ and every t such that $\Phi_t(\theta) \in \Theta_0$,*

$$f_{\Phi_t(\theta)}(x) = \mathcal{E}_{e^{-t}}[f_\theta](x) = \frac{f_\theta(x)^{e^{-t}}}{\int_M f_\theta(y)^{e^{-t}} \, dy}, \qquad x \in M. \tag{28}$$

Equivalently, the family satisfies $V_H = \Delta_{\text{lin}}$ exactly on Θ_0 if and only if each entropy-flow line is generated by normalized density-power transforms.

Proof. Assume (i) and fix $\theta \in \Theta_0$. Write $\theta_t := \Phi_t(\theta)$ whenever the flow is defined inside Θ_0. By Theorem IV.5, the hypothesis $V_H(\theta_t) = \Delta_{\text{lin}}(\theta_t)$ means that for every such t and every $x \in M$,

$$a_{\theta_t}^\top \nabla_\vartheta \ell_\vartheta(x)\Big|_{\vartheta=\theta_t} = -\ell_{\theta_t}(x) - h(\theta_t).$$

Because $\dot\theta_t = a_{\theta_t}$, this is the scalar ODE

$$\frac{d}{dt}\ell_{\theta_t}(x) = -\ell_{\theta_t}(x) - h(\theta_t).$$

Solving it gives

$$\ell_{\theta_t}(x) = e^{-t}\ell_\theta(x) - e^{-t}\int_0^t e^u h(\theta_u)\, du.$$

Hence

$$f_{\theta_t}(x) = c_t(\theta)\, f_\theta(x)^{e^{-t}}, \qquad c_t(\theta) := \exp\!\left(-e^{-t}\int_0^t e^u h(\theta_u)\,du\right),$$

where $c_t(\theta)$ does not depend on x. Since f_{θ_t} is a probability density, normalization identifies

$$c_t(\theta)^{-1} = \int_M f_\theta(y)^{e^{-t}}\,dy,$$

and (28) follows.

Conversely, assume (ii). Fix $\theta \in \Theta_0$ and write

$$c_t(\theta) := \log \int_M f_\theta(y)^{e^{-t}}\,dy.$$

Taking logarithms in (28) yields

$$\ell_{\Phi_t(\theta)}(x) = e^{-t}\ell_\theta(x) - c_t(\theta).$$

Choose any $x_0 \in M$. Because $(x,\vartheta) \mapsto \ell_\vartheta(x)$ is C^1 and $t \mapsto \Phi_t(\theta)$ is differentiable, the scalar function

$$c_t(\theta) = e^{-t}\ell_\theta(x_0) - \ell_{\Phi_t(\theta)}(x_0)$$

is differentiable at $t = 0$. Differentiating the previous display at $t = 0$ gives

$$a_\theta^\top \nabla_\theta \ell_\theta(x) = -\ell_\theta(x) - c_0'(\theta).$$

Taking P_θ-expectations and using $\mathbb{E}_\theta[S_\theta] = 0$ yields

$$0 = \mathbb{E}_\theta[a_\theta^\top S_\theta] = -\mathbb{E}_\theta[\ell_\theta(X)] - c_0'(\theta) = h(\theta) - c_0'(\theta).$$

Thus $c_0'(\theta) = h(\theta)$, and therefore

$$a_\theta^\top s_\theta(x) = -\ell_\theta(x) - h(\theta), \qquad x \in M.$$

By Theorem IV.5, this is equivalent to $V_H(\theta) = \Delta_{\mathrm{lin}}(\theta)$. Since θ was arbitrary, (i) follows. □

Corollary IV.9 (Local normal form)**.** *Assume the hypotheses of Theorem IV.8 and suppose that $V_H(\vartheta) = \Delta_{\mathrm{lin}}(\vartheta)$ throughout Θ_0. Fix $\theta_0 \in \Theta_0$ with $a_{\theta_0} \neq 0$. Then there exist local coordinates (t,ζ) on a neighborhood $U \subset \Theta_0$ of θ_0 and positive seed densities g_ζ such that*

$$f_{t,\zeta}(x) = \frac{g_\zeta(x)^{e^{-t}}}{\int_M g_\zeta(y)^{e^{-t}}\,dy}, \qquad (t,\zeta) \in U,\ x \in M. \tag{29}$$

Conversely, any local family of the form (29) satisfies $V_H(t,\zeta) = \Delta_{\mathrm{lin}}(t,\zeta)$ throughout U; in these coordinates, the entropy vector field is ∂_t.

Proof. Because $a_{\theta_0} \neq 0$, the flow-box Theorem gives local coordinates (t,ζ) on a neighborhood U of θ_0 in which the vector field a becomes ∂_t. Let $g_\zeta := f_{0,\zeta}$. Applying Theorem IV.8 along the flow lines $u \mapsto (u,\zeta)$ gives exactly (29).

Conversely, suppose (29) holds and write

$$Z(t,\zeta) := \int_M g_\zeta(y)^{e^{-t}}\,dy, \qquad \ell_{t,\zeta}(x) = e^{-t}\log g_\zeta(x) - \log Z(t,\zeta).$$

Differentiating in t gives

$$\partial_t \ell_{t,\zeta}(x) = -e^{-t}\log g_\zeta(x) - \partial_t \log Z(t,\zeta).$$

Because $\mathbb{E}_{t,\zeta}[s_t(X)] = 0$, taking expectations yields

$$\partial_t \log Z(t,\zeta) = -e^{-t}\mathbb{E}_{t,\zeta}[\log g_\zeta(X)].$$

On the other hand,

$$h(t,\zeta) = -\mathbb{E}_{t,\zeta}[\ell_{t,\zeta}(X)] = -e^{-t}\mathbb{E}_{t,\zeta}[\log g_\zeta(X)] + \log Z(t,\zeta) = \partial_t \log Z(t,\zeta) + \log Z(t,\zeta).$$

Therefore

$$\partial_t \ell_{t,\zeta}(x) = -e^{-t}\log g_\zeta(x) - h(t,\zeta) + \log Z(t,\zeta) = -\ell_{t,\zeta}(x) - h(t,\zeta).$$

Thus the t-score satisfies

$$s_t(X) = \partial_t \ell_{t,\zeta}(X) = Y_{t,\zeta} - h(t,\zeta).$$

For every coordinate $\theta_j \in \{t, \zeta_1, \dots, \zeta_{p-1}\}$,

$$\partial_{\theta_j} h(t,\zeta) = \mathbb{E}_{t,\zeta}\Big[(Y_{t,\zeta} - h(t,\zeta)) s_{\theta_j}(X)\Big] = \mathbb{E}_{t,\zeta}\Big[s_t(X) s_{\theta_j}(X)\Big] = I_{tj}(t,\zeta).$$

Hence $\nabla h(t,\zeta) = I(t,\zeta) e_t$, where $e_t = (1,0,\dots,0)^\top$ is the coordinate vector in the t-direction. Therefore $a_{(t,\zeta)} = I(t,\zeta)^{-1} \nabla h(t,\zeta) = e_t$, i.e. the entropy vector field is ∂_t. Applying Theorem IV.5 gives $V_H(t,\zeta) = \Delta_{\mathrm{lin}}(t,\zeta)$ on U. □

IV-F Broad classes closed under normalized density powers

The characterization by normalized density powers makes the exact tangent class practically useful only if one can recognize large model classes that are automatically closed under the transform $g \mapsto \mathcal{E}_\beta[g]$. The next two results show that this covers both a genuinely heavy-tail family (Student-t with all standard parameters) and a broad portion of classical exponential-family models.

Proposition IV.10 (The full multivariate Student-t family is closed under normalized density powers)**.** *For $\mu \in \mathbb{R}^d$, positive definite $\Sigma \in \mathbb{R}^{d\times d}$, and $\nu > 0$, let*

$$f_{\mu,\Sigma,\nu}(x) = c_{d,\nu} |\Sigma|^{-1/2} \left(1 + \frac{1}{\nu} q_{\mu,\Sigma}(x)\right)^{-(\nu+d)/2}, \qquad q_{\mu,\Sigma}(x) := (x-\mu)^\top \Sigma^{-1} (x-\mu).$$

For $\beta > 0$ such that

$$\nu_\beta := \beta(\nu + d) - d > 0, \qquad \Sigma_\beta := \frac{\nu}{\nu_\beta}\Sigma,$$

one has

$$\mathcal{E}_\beta[f_{\mu,\Sigma,\nu}] = f_{\mu,\Sigma_\beta,\nu_\beta}. \tag{30}$$

Consequently, the full location-scatter-degrees-of-freedom Student-t family is locally closed under normalized density powers and therefore satisfies

$$V_H(\mu,\Sigma,\nu) = \Delta_{\mathrm{lin}}(\mu,\Sigma,\nu)$$

throughout its parameter domain.

Proof. A direct calculation gives

$$f_{\mu,\Sigma,\nu}(x)^\beta \propto |\Sigma|^{-\beta/2} \left(1 + \frac{1}{\nu} q_{\mu,\Sigma}(x)\right)^{-\beta(\nu+d)/2}.$$

Because $\nu_\beta + d = \beta(\nu + d)$ and

$$q_{\mu,\Sigma_\beta}(x) = (x-\mu)^\top \Sigma_\beta^{-1} (x-\mu) = \frac{\nu_\beta}{\nu} q_{\mu,\Sigma}(x),$$

one has

$$1 + \frac{1}{\nu_\beta} q_{\mu,\Sigma_\beta}(x) = 1 + \frac{1}{\nu} q_{\mu,\Sigma}(x).$$

Hence

$$f_{\mu,\Sigma,\nu}(x)^\beta \propto f_{\mu,\Sigma_\beta,\nu_\beta}(x),$$

and normalization yields (30). For β near 1 one has $\nu_\beta > 0$, so the closure is local around every parameter point. Moreover, the parameter update

$$\Gamma_t(\mu,\Sigma,\nu) := \left(\mu, \frac{\nu}{e^{-t}(\nu+d) - d}\Sigma,\ e^{-t}(\nu+d) - d\right)$$

satisfies $\Gamma_0 = \mathrm{id}$ and $\Gamma_{t+s} = \Gamma_t \circ \Gamma_s$ whenever both sides are defined. Therefore Γ_t is a local flow, and (30) rewrites as

$$f_{\Gamma_t(\mu,\Sigma,\nu)}(x) = \frac{f_{\mu,\Sigma,\nu}(x)^{e^{-t}}}{\int_M f_{\mu,\Sigma,\nu}(y)^{e^{-t}}\, dy}.$$

Let

$$b_{\mu,\Sigma,\nu} := \frac{d}{dt}\Gamma_t(\mu,\Sigma,\nu)\Big|_{t=0}, \qquad c_t(\mu,\Sigma,\nu) := \log \int_{\mathbb{R}^d} f_{\mu,\Sigma,\nu}(y)^{e^{-t}}\, dy.$$

Taking logarithms in the last display and differentiating at $t = 0$ gives

$$\ell_{\mu,\Sigma,\nu}(x) + b_{\mu,\Sigma,\nu}^\top \nabla_{(\mu,\Sigma,\nu)} \ell_{\mu,\Sigma,\nu}(x) = -c_0'(\mu,\Sigma,\nu).$$

The right-hand side is independent of x, so Proposition IV.13 yields $V_H(\mu,\Sigma,\nu) = \Delta_{\mathrm{lin}}(\mu,\Sigma,\nu)$. □

Proposition IV.11 (Exact tangent criterion for exponential families)**.** *Consider a regular exponential family written in the form*

$$f_\eta(x) = \exp\big(\eta^\top T(x) - A(\eta) + c(x)\big), \qquad \eta \in \mathcal{H} \subset \mathbb{R}^m,$$

on a common support. Then the following are equivalent:

(i) $V_H(\eta_0) = \Delta_{\mathrm{lin}}(\eta_0)$ *for some parameter value* $\eta_0 \in \mathcal{H}$*;*
(ii) $V_H(\eta) = \Delta_{\mathrm{lin}}(\eta)$ *for every* $\eta \in \mathcal{H}$*;*
(iii) *there exist* $b_0 \in \mathbb{R}$ *and* $b \in \mathbb{R}^m$ *such that*

$$c(x) = b_0 + b^\top T(x) \tag{31}$$

on the common support.

Whenever these conditions hold, the family is locally closed under normalized density powers and obeys

$$\mathcal{E}_\beta[f_\eta] = f_{\eta_\beta}, \qquad \eta_\beta := \beta(\eta + b) - b, \tag{32}$$

for all β *sufficiently close to* 1*.*

Proof. Assume (i). For the parameter value η_0, write $a_0 := I(\eta_0)^{-1}\nabla h(\eta_0)$. Since

$$\ell_{\eta_0}(x) = \eta_0^\top T(x) - A(\eta_0) + c(x), \qquad s_{\eta_0}(x) = T(x) - \nabla A(\eta_0),$$

Theorem IV.5 gives

$$-\eta_0^\top T(x) + A(\eta_0) - c(x) = h(\eta_0) + a_0^\top T(x) - a_0^\top \nabla A(\eta_0).$$

Therefore

$$c(x) = A(\eta_0) - h(\eta_0) + a_0^\top \nabla A(\eta_0) - \big(\eta_0 + a_0\big)^\top T(x),$$

which is exactly (31). Thus (i) implies (iii).

Now assume (iii). Then

$$\ell_\eta(x) = (\eta + b)^\top T(x) - \widetilde{A}(\eta), \qquad \widetilde{A}(\eta) := A(\eta) - b_0,$$

so the information content is

$$Y_\eta = -\ell_\eta(X) = \widetilde{A}(\eta) - (\eta + b)^\top T(X).$$

After centering,

$$Y_\eta - h(\eta) = -(\eta + b)^\top \big(T(X) - \mathbb{E}_\eta[T(X)]\big) = -(\eta + b)^\top s_\eta(X).$$

Multiplying by $s_\eta(X)$ and taking expectations gives

$$\nabla h(\eta) = \mathbb{E}_\eta\big[(Y_\eta - h(\eta))s_\eta(X)\big] = -I(\eta)(\eta + b),$$

so $a_\eta = I(\eta)^{-1}\nabla h(\eta) = -(\eta + b)$. Therefore Theorem IV.5 gives $V_H(\eta) = \Delta_{\mathrm{lin}}(\eta)$ for every η, proving (iii)⇒(ii). The implication (ii)⇒(i) is immediate.

Finally, under (iii),

$$f_\eta(x)^\beta \propto \exp\big(\beta(\eta + b)^\top T(x)\big),$$

so after normalization one obtains exactly (32). Since $\eta_\beta \to \eta$ as $\beta \to 1$, this gives local closure under normalized density powers. □

Remark IV.12 (Practical inspection rule for exponential families)*.* Proposition IV.11 says that, for exact tangent explainability, every x-dependent term in $\log f_\eta(x)$ must carry a free coefficient. A genuinely frozen carrier term blocks the identity $V_H = \Delta_{\mathrm{lin}}$. Thus Gaussian mean-precision families, Beta and Dirichlet models, and Gamma shape-scale families lie inside the exact tangent class, while Poisson-type models with carrier $-\log(x!)$ lie outside it.

IV-G Practical verification criteria

The characterization by normalized density powers provides the structural classification of exact tangent families, but it is not the most efficient way to check a concrete model. In practice, one usually does *not* construct the entropy flow explicitly. Instead, one looks for an x-independent coefficient vector multiplying the score, or equivalently the mixed derivatives, and only then invokes the structural theory to interpret the result.

Proposition IV.13 (Direct verification criteria for the exact tangent identity)**.** *Fix $\theta \in \Theta$. Then the following are equivalent:*

(i) $V_H(\theta) = \Delta_{\mathrm{lin}}(\theta)$*;*

(ii) *there exists a vector $b_\theta \in \mathbb{R}^p$ such that*

$$\ell_\theta(x) + b_\theta^\top \nabla_\theta \ell_\theta(x) \tag{33}$$

is independent of $x \in M$;

(iii) *there exists a vector $b_\theta \in \mathbb{R}^p$ such that*

$$\nabla_x \ell_\theta(x) + J_\theta(x)^\top b_\theta = 0, \qquad x \in M. \tag{34}$$

Whenever these conditions hold, the vector b_θ is unique and equals

$$b_\theta = a_\theta = I(\theta)^{-1} \nabla h(\theta).$$

Proof. The equivalence (i)⇔(ii) is immediate from Theorem IV.5: if (i) holds, take $b_\theta = a_\theta$. Conversely, if (ii) holds, then for some constant c_θ one has

$$\ell_\theta(x) + b_\theta^\top s_\theta(x) = c_\theta, \qquad x \in M.$$

Taking P_θ-expectations and using $\mathbb{E}_\theta[S_\theta] = 0$ gives $c_\theta = -h(\theta)$, so

$$Y_\theta - h(\theta) = b_\theta^\top S_\theta.$$

Multiplying by S_θ and taking expectations yields the entropy-score covariance identity

$$\nabla h(\theta) = \mathbb{E}_\theta\big[(Y_\theta - h(\theta))S_\theta\big] = \mathbb{E}_\theta\big[(b_\theta^\top S_\theta)S_\theta\big] = I(\theta) b_\theta,$$

whence $b_\theta = a_\theta$. Therefore (ii) yields the score-linear identity in Theorem IV.5, and hence (i) follows.

The implication (ii)⇒(iii) follows by differentiating (33) with respect to x. Conversely, if (34) holds, then the function

$$F_\theta(x) := \ell_\theta(x) + b_\theta^\top s_\theta(x)$$

has zero x-gradient on the connected set M, so it is constant on M, proving (ii). □

Corollary IV.14 (Coefficient-matching criterion)**.** *Assume that, for a fixed parameter value θ, the log-likelihood admits the representation*

$$\ell_\theta(x) = c(\theta) + \sum_{r=1}^{m} \beta_r(\theta)\, \phi_r(x), \qquad x \in M, \tag{35}$$

where $1, \phi_1, \dots, \phi_m$ are linearly independent on M. Let $\beta(\theta) = (\beta_1(\theta), \dots, \beta_m(\theta))^\top$ and let $D_\theta \beta(\theta)$ denote its $m \times p$ Jacobian. Then

$$V_H(\theta) = \Delta_{\mathrm{lin}}(\theta) \iff \exists\, b_\theta \in \mathbb{R}^p \text{ such that } D_\theta \beta(\theta) b_\theta = -\beta(\theta).$$

Whenever this holds, one necessarily has $b_\theta = a_\theta$. In particular, if $D_\theta \beta(\theta)$ has full row rank m, then $V_H(\theta) = \Delta_{\mathrm{lin}}(\theta)$ automatically.

Proof. By (35),

$$\ell_\theta(x) + b_\theta^\top \nabla_\theta \ell_\theta(x) = c(\theta) + b_\theta^\top \nabla c(\theta) + \sum_{r=1}^{m} \big(\beta_r(\theta) + \nabla \beta_r(\theta)^\top b_\theta\big) \phi_r(x).$$

Because $1, \phi_1, \dots, \phi_m$ are linearly independent on M, this expression is independent of x if and only if

$$\beta_r(\theta) + \nabla \beta_r(\theta)^\top b_\theta = 0, \qquad r = 1, \dots, m,$$

which is exactly $D_\theta \beta(\theta) b_\theta = -\beta(\theta)$. The conclusion now follows from Proposition IV.13. □

Corollary IV.15 (One-parameter scalar-observation ratio test)**.** *Assume $p = 1$ and that the support $M \subset \mathbb{R}$ is an open interval. Suppose that $\partial_x \partial_\theta \ell_\theta(x) \neq 0$ for all $x \in M$. Then*

$$V_H(\theta) = \Delta_{\mathrm{lin}}(\theta) \iff -\frac{\partial_x \ell_\theta(x)}{\partial_x \partial_\theta \ell_\theta(x)}$$

is independent of $x \in M$. When this holds, the common value is $a_\theta = I(\theta)^{-1} h'(\theta)$.

Proof. When $p = 1$, the vector b_θ in Proposition IV.13 is a scalar. The differential condition (34) therefore becomes

$$\partial_x \ell_\theta(x) + b_\theta \, \partial_x \partial_\theta \ell_\theta(x) = 0, \qquad x \in M.$$

Because the denominator is everywhere nonzero by assumption, this is equivalent to saying that

$$-\frac{\partial_x \ell_\theta(x)}{\partial_x \partial_\theta \ell_\theta(x)} \equiv b_\theta$$

is constant in x. The final identification $b_\theta = a_\theta$ comes from Proposition IV.13. □

V Examples and Moving-Support Boundary Cases

The general results of Sections III and IV become most informative when they can be recognized directly on familiar model classes. The three examples below are chosen to illustrate the verification routes developed earlier. None of them requires an explicit integral calculation of the entropy gradient.

We find the Pareto and generalized inverse Gaussian (GIG) examples particularly interesting. The fixed-threshold Pareto model shows the limits of the present common-support framework and points toward future work on parameter-dependent support. The GIG example, on the other hand, highlights the utility of the structural approach in a setting where direct symbolic integration for entropy and varentropy is cumbersome.

V-A Examples

Proposition V.1 (Beta family)**.** *For $a, b > 0$, let*

$$f_{a,b}(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, \qquad 0 < x < 1.$$

Then the exact tangent identity

$$V_H(a,b) = \Delta_{\mathrm{lin}}(a,b)$$

holds throughout $(0,\infty)^2$. Moreover,

$$V_H(a,b) = (a-1)^2 \psi_1(a) + (b-1)^2 \psi_1(b) - (a+b-2)^2 \psi_1(a+b),$$

where ψ_1 denotes the trigamma function.

Proof. The log-density is

$$\ell_{a,b}(x) = -\log B(a,b) + (a-1)\log x + (b-1)\log(1-x),$$

so the Beta family is a regular exponential family with common support, sufficient statistic

$$T(x) = \big(\log x, \log(1-x)\big),$$

and carrier term $c(x) \equiv 0$. Proposition IV.11 therefore gives

$$V_H(a,b) = \Delta_{\mathrm{lin}}(a,b)$$

for every (a,b).

The score coordinates are

$$S_a = \log X - \psi(a) + \psi(a+b), \qquad S_b = \log(1-X) - \psi(b) + \psi(a+b),$$

and the Fisher information matrix is the Hessian of $\log B(a,b)$,

$$I(a,b) = \begin{pmatrix} \psi_1(a) - \psi_1(a+b) & -\psi_1(a+b) \\ -\psi_1(a+b) & \psi_1(b) - \psi_1(a+b) \end{pmatrix}.$$

Because $c(x) \equiv 0$, the proof of Proposition IV.11 yields

$$Y_{a,b} - h(a,b) = -(a-1)S_a - (b-1)S_b.$$

Hence

$$V_H(a,b) = \begin{pmatrix} a-1 & b-1 \end{pmatrix} I(a,b) \begin{pmatrix} a-1 \\ b-1 \end{pmatrix},$$

which expands to the stated trigamma formula. □

Proposition V.2 (Pareto tail-index family with fixed threshold)**.** *Fix $x_m > 0$ and consider the one-parameter Pareto family*

$$f_\alpha(x) = \alpha x_m^\alpha x^{-(\alpha+1)} \mathbf{1}_{(x_m,\infty)}(x), \qquad \alpha > 0.$$

Then

$$V_H(\alpha) = \Delta_{\text{lin}}(\alpha)$$

for every $\alpha > 0$, and in fact

$$V_H(\alpha) = \left(\frac{\alpha+1}{\alpha}\right)^2.$$

Proof. On the common support $M = (x_m, \infty)$,

$$\ell_\alpha(x) = \log\alpha + \alpha \log x_m - (\alpha+1)\log x.$$

Therefore

$$\partial_x \ell_\alpha(x) = -\frac{\alpha+1}{x}, \qquad \partial_x \partial_\alpha \ell_\alpha(x) = -\frac{1}{x},$$

so the ratio in Corollary IV.15 is the constant

$$-\frac{\partial_x \ell_\alpha(x)}{\partial_x \partial_\alpha \ell_\alpha(x)} = -(\alpha+1).$$

Hence $V_H(\alpha) = \Delta_{\text{lin}}(\alpha)$.

For the explicit value, note that the score is

$$S_\alpha(X) = \partial_\alpha \ell_\alpha(X) = \frac{1}{\alpha} - \log\left(\frac{X}{x_m}\right).$$

Since $\log(X/x_m) \sim \text{Exp}(\alpha)$, one has

$$I(\alpha) = \text{Var}_\alpha(S_\alpha) = \text{Var}\left(\log\left(\frac{X}{x_m}\right)\right) = \frac{1}{\alpha^2}.$$

With $a_\alpha = -(\alpha+1)$ from the ratio test,

$$V_H(\alpha) = a_\alpha^2 I(\alpha) = \frac{(\alpha+1)^2}{\alpha^2}.$$

□

Remark V.3 (Moving-support Pareto-type models)*.* Proposition V.2 treats the common-support subfamily with fixed threshold x_m. If the lower endpoint itself is unknown, as in the usual two- parameter Pareto family, then assumption (P1) fails. In that regime the entropy gradient identity from Section 3 acquires boundary terms and the present four-term decomposition does not apply.

This is, in our view, the most structurally interesting direction for further development of the theory. What is needed is a revised entropy-gradient identity and a corresponding varentropy decomposition with the correct endpoint contributions built in. Accomplishing this would bring threshold, truncation, and support-shift models into the same structural picture and would substantially enlarge the practical scope of the theory.

Proposition V.4 (Generalized inverse Gaussian family)**.** *For $p \in \mathbb{R}$ and $a, b > 0$, let*

$$f_{p,a,b}(x) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})}\, x^{p-1} \exp\left(-\frac{ax + b/x}{2}\right), \qquad x > 0,$$

where K_p is the modified Bessel function of the second kind. Then the standing assumptions of Section 3 hold locally on $\mathbb{R} \times (0,\infty)^2$. Moreover, for every $\beta > 0$,

$$\mathcal{E}_\beta[f_{p,a,b}] = f_{1+\beta(p-1),\,\beta a,\,\beta b},$$

so the family satisfies

$$V_H(p,a,b) = \Delta_{\mathrm{lin}}(p,a,b)$$

throughout $\mathbb{R} \times (0,\infty)^2$. Writing $t := \sqrt{ab}$ and

$$L := (p-1)\partial_p + t\partial_t,$$

one further has

$$h(p,a,b) = \frac{1}{2}\log\left(\frac{b}{a}\right) + \log\big(2K_p(t)\big) - L\log K_p(t)$$

and

$$V_H(p,a,b) = \big(L^2 - L\big)\log K_p(t).$$

Proof. The support is the fixed set $(0,\infty)$. On every compact parameter set $K \subset \mathbb{R}\times(0,\infty)^2$, the parameters a and b are bounded below by a positive constant, so the density and its first parameter derivatives are dominated on K by an integrable envelope of the form

$$C_K\big(1 + |\log x| + x + x^{-1}\big)e^{-c_K(x+x^{-1})/2}, \qquad x > 0.$$

This justifies differentiation under the integral sign and implies that both the score and the information content belong to $L^2(P_{p,a,b})$. To verify positive definiteness of the Fisher information matrix, it is enough to show that

$$\mathrm{Cov}_{p,a,b}\big(\log X, X, X^{-1}\big)$$

is nonsingular. Suppose

$$u_1 \log X + u_2 X + u_3 X^{-1}$$

has zero variance under $P_{p,a,b}$. Since $f_{p,a,b} > 0$ on $(0,\infty)$, this quantity is $P_{p,a,b}$-a.s. constant, hence constant on $(0,\infty)$ by continuity. Differentiating with respect to x gives

$$\frac{u_1}{x} + u_2 - \frac{u_3}{x^2} = 0 \qquad \text{for all } x > 0,$$

so multiplying by x^2 yields

$$u_2 x^2 + u_1 x - u_3 = 0 \qquad \text{for all } x > 0.$$

Therefore $u_1 = u_2 = u_3 = 0$. Thus $\mathrm{Cov}_{p,a,b}(\log X, X, X^{-1})$ is positive definite. Since the score coordinates differ from $(\log X, X, X^{-1})$ only by additive constants and nonzero scalar factors, the Fisher information matrix is positive definite as well. Thus the standing assumptions of Section 3 hold locally throughout $\mathbb{R} \times (0,\infty)^2$.

For the normalized density powers,

$$f_{p,a,b}(x)^\beta \propto x^{\beta(p-1)} \exp\left(-\frac{\beta a\, x + \beta b/x}{2}\right),$$

which is again a generalized inverse Gaussian density with parameters

$$p_\beta := 1 + \beta(p-1), \qquad a_\beta := \beta a, \qquad b_\beta := \beta b.$$

Hence

$$\mathcal{E}_\beta[f_{p,a,b}] = f_{p_\beta,a_\beta,b_\beta},$$

which is consistent with Theorem IV.8.

Now write

$$\ell_{p,a,b}(x) = c(p,a,b) + (p-1)\log x - \frac{a}{2}x - \frac{b}{2}x^{-1},$$

where

$$c(p,a,b) = \frac{p}{2}\log\left(\frac{a}{b}\right) - \log 2 - \log K_p(t).$$

Set

$$b_{p,a,b} := (1-p, -a, -b).$$

A direct coefficient check gives

$$\ell_{p,a,b}(x) + b_{p,a,b}^{\top} \nabla_{(p,a,b)} \ell_{p,a,b}(x) = c(p,a,b) + b_{p,a,b}^{\top} \nabla c(p,a,b),$$

which is independent of x. Proposition IV.13 therefore yields

$$V_H(p,a,b) = \Delta_{\mathrm{lin}}(p,a,b) \qquad \text{and} \qquad b_{p,a,b} = a_{p,a,b}.$$

Taking $P_{p,a,b}$-expectations in the displayed identity above and using $\mathbb{E}_{p,a,b}[S_{p,a,b}] = 0$ gives

$$h(p,a,b) = -c(p,a,b) - b_{p,a,b}^{\top} \nabla c(p,a,b).$$

Introduce the first-order differential operator

$$D := (p-1)\partial_p + a\partial_a + b\partial_b.$$

Since $b_{p,a,b}^{\top} \nabla = -D$ and $A := -c = \log 2 + \frac{p}{2}\log(b/a) + \log K_p(t)$, the entropy identity becomes

$$h = A - DA.$$

Now D acts on functions of (p,t) as $L = (p-1)\partial_p + t\partial_t$, while

$$D\Big(\frac{p}{2}\log\Big(\frac{b}{a}\Big)\Big) = \frac{p-1}{2}\log\Big(\frac{b}{a}\Big).$$

Therefore

$$h(p,a,b) = \frac{1}{2}\log\Big(\frac{b}{a}\Big) + \log\big(2K_p(t)\big) - L\log K_p(t).$$

Because the coefficient functions $(p-1, -a/2, -b/2)$ are affine in (p,a,b), the parameter Hessian of $\ell_{p,a,b}(x)$ is simply $\nabla^2_{(p,a,b)} c(p,a,b)$, which is independent of x. Hence

$$I(p,a,b) = -\mathbb{E}_{p,a,b}\big[\nabla^2_{(p,a,b)} \ell_{p,a,b}(X)\big] = -\nabla^2_{(p,a,b)} c(p,a,b) = \nabla^2_{(p,a,b)} A(p,a,b).$$

Using the exact tangent identity,

$$V_H(p,a,b) = b_{p,a,b}^{\top} I(p,a,b)\, b_{p,a,b} = b_{p,a,b}^{\top} \nabla^2_{(p,a,b)} A(p,a,b)\, b_{p,a,b}.$$

Since $D = (p-1)\partial_p + a\partial_a + b\partial_b = -b_{p,a,b}^{\top}\nabla$, one has

$$D^2 A = DA + b_{p,a,b}^{\top} \nabla^2_{(p,a,b)} A\, b_{p,a,b},$$

and therefore

$$V_H = D^2 A - DA.$$

Applying this to $A = \log 2 + \frac{p}{2}\log(b/a) + \log K_p(t)$, the contribution of $\frac{p}{2}\log(b/a)$ cancels between D^2A and DA, while D reduces to L on functions of (p,t). This yields

$$V_H(p,a,b) = \big(L^2 - L\big)\log K_p(t),$$

as claimed.

Note that in the GIG case it is easy to check the result using conventional methods of deriving varentropy, since the moment generating function exists and has closed-form solution. Indeed, set

$$Z(\beta) := \int_0^\infty f_{p,a,b}(x)^\beta \, dx.$$

Using the standard Bessel-K integral

$$\int_0^\infty x^{q-1} \exp\Big(-\frac{ux + v/x}{2}\Big)\, dx = 2\Big(\frac{v}{u}\Big)^{q/2} K_q\big(\sqrt{uv}\big), \qquad u, v > 0,$$

with $q = 1 + \beta(p-1)$, $u = \beta a$, and $v = \beta b$, one finds

$$Z(\beta) = 2^{1-\beta}\Big(\frac{b}{a}\Big)^{(1-\beta)/2} \frac{K_{1+\beta(p-1)}(\beta t)}{K_p(t)^\beta}.$$

Note that the cumulant generating function of Y is

$$K_Y(m) = \log \mathbb{E}[e^{mY}] = \log Z(1-m).$$

Hence $K_Y'(0) = -(\log Z)'(1) = h(p,a,b)$ and $K_Y''(0) = (\log Z)''(1) = V_H(p,a,b)$, which recovers the stated formulas. □

References

[1] K.-S. Song, "Rényi information, loglikelihood and an intrinsic distribution measure," *Journal of Statistical Planning and Inference*, vol. 93, no. 1-2, pp. 51–69, 2001. DOI: 10.1016/S0378-3758(00)00169-5

[2] J. Nair, A. Wierman, and B. Zwart, *The Fundamentals of Heavy Tails: Properties, Emergence, and Estimation*. Cambridge: Cambridge University Press, 2022.

[3] A. Di Crescenzo, L. Paolillo, and A. Suárez-Llorens, "Stochastic comparisons, differential entropy and varentropy for distributions induced by probability density functions," *Metrika*, vol. 88, no. 1, pp. 43–59, 2025. DOI: 10.1007/s00184-024-00947-3

[4] M. Fradelizi, J. Li, and M. Madiman, "Concentration of information content for convex measures," *Electronic Journal of Probability*, vol. 25, pp. 1–22, 2020. DOI: 10.1214/20-EJP416

[5] J. P. Vigneaux, "Typicality for stratified measures," *IEEE Transactions on Information Theory*, vol. 69, no. 11, pp. 6922–6940, 2023. DOI: 10.1109/TIT.2023.3297322

[6] J. P. Vigneaux, "On the entropy of rectifiable and stratified measures," in *Geometric Science of Information*, Cham: Springer Nature Switzerland, 2023, pp. 338–346.

[7] N. Leonenko, Y. Sun, and E. Taufer, "Varentropy estimation via nearest neighbor graphs," *arXiv preprint arXiv:2402.09374*, 2024. [Online]. Available: https://arxiv.org/abs/2402.09374

[8] O. Rioul, "Information theoretic proofs of entropy power inequalities," *IEEE Transactions on Information Theory*, vol. 57, pp. 33–55, 1 2011. DOI: 10.1109/TIT.2010.2090193 [Online]. Available: https://ieeexplore.ieee.org/document/5673809

[9] L. Negro, "Sample distribution theory using coarea formula," *Communications in Statistics - Theory and Methods*, vol. 53, no. 5, pp. 1864–1889, 2024. DOI: 10.1080/03610926.2022.2116284

[10] J. Malý, D. Swanson, and W. P. Ziemer, "The co-area formula for sobolev mappings," *Transactions of the American Mathematical Society*, vol. 355, no. 2, pp. 477–492, 2003. DOI: 10.1090/S0002-9947-02-03091-X

[11] A. Baernstein II, D. Drasin, and R. Laugesen, *Symmetrization in Analysis*. Cambridge: Cambridge University Press, 2019.