ALGORITHMS AND HARDNESS FOR CONCAVE EXPECTATIONAL FUNCTIONALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Statistical similarity and several related overlaps lack simple coordinatewise formulas on product distributions, unlike familiar quantities such as Kullback–Leibler divergence and Hellinger affinity. This distinction matters for the Bayes-optimal error of a naive-Bayes classifier: the error can be exponentially small, making additive estimates and standard bounds from coordinatewise quantities uninformative. We study the expectational functionals $F(P,Q) = \mathbf{E}_{x\sim Q}[f(P(x)/Q(x))]$ for every non-negative concave f with $f(0) = 0$ and $f \not\equiv 0$. The family includes statistical similarity $1-d_{\mathrm{TV}}(P,Q)$, which determines Bayes-optimal error, as well as log, harmonic, exponential, $\tanh$, $\arctan$, and square-root overlaps. For product distributions, given a rational lower bound on $f(1)$ and a polynomial-time evaluation oracle for f, we give a deterministic fully polynomial-time approximation scheme for the entire family. Its running time is polynomial in the logarithm of the reciprocal of the smallest nonzero marginal probability, and it approximates naive-Bayes error within a factor $1+\varepsilon$. On a simple naive-Bayes family, the lower bound from the factorizing Bhattacharyya coefficient misses the true error by an exponential factor. For Bayes nets of in-degree 2, deciding whether $F(P,Q) > 0$ is NP-complete for every member of the family, ruling out polynomial-time multiplicative approximation unless $\mathsf{P} = \mathsf{NP}$. The algorithm uses subhomogeneity and a positive lower bound on partial likelihood-ratio products; no growth condition on f at the origin is needed.

1 INTRODUCTION

The Bayes-optimal error of a binary classifier is the smallest error achievable from its features. With class priors α_0, α_1 and class-conditionals P_0, P_1, it equals $R^* = \sum_x \min(\alpha_0 P_0(x), \alpha_1 P_1(x))$ (Fukunaga, 1990, Eq. (3.7)). In a naive-Bayes model, the class-conditionals are products, yet R^* can be exponentially small in the number of features. An additive estimate can therefore obscure the error entirely. The computational question is whether R^* can be approximated *multiplicatively* from the marginal tables.

Familiar divergences such as Kullback–Leibler, χ^2, Hellinger and Rényi have coordinatewise formulas on products and reduce to one-dimensional computations (Bhattacharyya et al., 2023). Their usual bounds need not determine R^*: on a simple naive-Bayes family, the lower end of the Bhattacharyya bracket falls below the true error by a factor exponential in the number of features, while the Pinsker bound is vacuous (Theorem C.1). Total variation d_{TV} and its complement, statistical similarity $S_{\mathrm{stat}}(P,Q) := \sum_x \min(P(x), Q(x)) = 1 - d_{\mathrm{TV}}(P,Q)$, lack such formulas. An FPTAS for d_{TV} (Feng et al., 2024) does not give one for S_{stat}: when $S_{\mathrm{stat}} = 2^{-n}$, a constant relative error in d_{TV} can overwhelm S_{stat} (Theorem 2.9). This echoes the asymmetry between #DNF and #CNF (Karp et al., 1989).

This paper studies the complexity of multiplicatively approximating the whole family

$$F(P,Q) := \mathop{\mathbf{E}}_{x\sim Q}[f(P(x)/Q(x))] = \sum_{x:Q(x)>0} Q(x)\cdot f\!\left(\frac{P(x)}{Q(x)}\right), \tag{1.1}$$

where $f\colon [0,\infty) \to [0,\infty)$ is *admissible*: concave, with $f(0) = 0$ and $f \not\equiv 0$. Such f is automatically non-decreasing, and no condition on its growth at the origin is needed. The choice $f(r) = \min(r, 1)$ gives S_{stat}, and $f(r) = \min(r, \alpha_1/\alpha_0)$ gives $R^* = \alpha_0 F(P_0, P_1)$.

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

1.1 THE FUNCTIONALS

Table 1 lists examples of the family; Section D discusses each at length. *Statistical similarity and its variants.* Besides the Bayes error, S_{stat} is the maximum coupling probability $\max_{(X,Y)} \mathbf{Pr}[X = Y]$ (Lindvall, 2002) and the minimum sum of Type-I and Type-II errors in testing P against Q (Kontorovich & Avital, 2025; Lehmann & Romano, 2005); capping at c, $f(r) = \min(r, c)$ gives the Bayes risk under asymmetric 0–1 loss with cost ratio c. *Information-theoretic overlaps.* Log-overlap, $f(r) = \log(1 + r)$, determines the Jensen–Shannon divergence: $D_{\mathrm{KL}}(Q\|(P + Q)/2) = \log 2 - F(P, Q)$, hence $D_{\mathrm{JS}}(P\|Q) = \log 2 - (F(P, Q) + F(Q, P))/2$; harmonic overlap, $f(r) = r/(1 + r)$, gives the triangular discrimination $\Delta(P, Q) = \sum_x (P(x) - Q(x))^2/(P(x) + Q(x)) = 2 - 4F(P, Q)$, which satisfies $2d_{\mathrm{TV}}^2 \le \Delta \le 2d_{\mathrm{TV}}$. *Smooth and bounded overlaps.* Exponential, $\tanh$ and $\arctan$ overlaps round off the corner of $\min(r, 1)$ at $r = 1$, and their capped versions bound the influence of any one atom, as in robust statistics (Huber, 1964; Hampel et al., 1986); the square-root overlaps $\min(\sqrt{r}, 1)$ and $2\sqrt{r}/(1 + \sqrt{r})$ are Hellinger-style overlaps that, unlike the Bhattacharyya coefficient, do not factorize.

1.2 OUR RESULTS

We give complementary algorithmic and hardness results for the entire family.

A generic FPTAS on products (Theorem 4.1). For every admissible f, given a rational φ with $0 < \varphi \le f(1)$ and an evaluation oracle for f, we give a deterministic algorithm that approximates $F(P, Q)$ within a factor $1 + \varepsilon$ on product distributions over $[\ell]^n$. It uses $O(n\ell m \log m)$ arithmetic operations, where $m = O((n/\varepsilon)(n \log(1/\tau) + 1))$ and τ is the smallest nonzero marginal probability. The dependence on τ is logarithmic, so exponentially small marginals are allowed. The algorithm covers every functional in Table 1, including S_{stat}. For S_{stat} it complements the FPRAS of Feng et al. (2023) and the FPTAS of Feng et al. (2024) for d_{TV}; for the other functionals we know of no earlier FPTAS. In particular, it gives a deterministic FPTAS for the Bayes-optimal error R^* of naive-Bayes models with arbitrary per-feature marginals and priors.

Generic hardness on Bayes nets (Theorem 5.1). For every admissible f, deciding whether $F(P, Q) > 0$ is NP-complete for Bayes nets of in-degree 2; hence no polynomial-time algorithm approximates F within any multiplicative factor unless P = NP, and exact computation is #P-hard whenever $f(1)$ can be computed to $n + O(1)$ bits in polynomial time. One reduction serves the whole family because it produces $F(P, Q) = f(1)|\mathrm{Sol}(\phi)|/2^n$, which uses only $f(0) = 0$ and $f(1) > 0$. For S_{stat} this is the statement that deciding $S_{\mathrm{stat}}(P, Q) > 0$ is hard; the earlier result of Bhattacharyya et al. (2023) is about the other end, deciding $S_{\mathrm{stat}}(P, Q) < 1$, and says nothing about multiplicative approximation of S_{stat}.

Exact computation on products (Section G). Exact computation of $\min(r, c)$-functionals is #P-hard on products for every rational $c > 0$ (Theorem G.5), so for these the FPTAS is the best one can hope for. We also correct the natural route to hardness for the smooth members: it would go through the moments $\mathbf{E}_{x\sim Q}\left[(P/Q)^K\right]$ of the likelihood ratio, but these factorize on products and are computable exactly (Theorem G.12).

1.3 TECHNIQUES AND RELATED WORK

Our algorithm maintains a sparsified representation of partial likelihood-ratio products, as in the deterministic FPTAS of Feng et al. (2024) for d_{TV}. The analysis differs in two places. First, merging within a slice is controlled by subhomogeneity, $f(\lambda r) \le \lambda f(r)$ for $\lambda \ge 1$. This follows from concavity through the origin alone and requires no smoothness, so one analysis covers both piecewise-linear and smooth members. Second, near-zero ratios, where rounding would have unbounded relative error, never arise: every positive partial-product ratio is at least τ^n (Theorem A.2). A slice below τ^n therefore stays empty, and sparsification incurs no additive error, regardless of the growth of f at the origin. For d_{TV}, the slice adjacent to $r = 1$ does carry an additive error. Feng et al. (2024) charge it against the coordinatewise bound $d_{\mathrm{TV}}(P, Q) \ge \max_i d_{\mathrm{TV}}(p_i, q_i)$. No analogous single-coordinate lower bound exists for F (Theorem 2.9), and our analysis needs none.

Positive partial-product ratios range over $[\tau^n, \tau^{-n}]$, an interval of logarithmic width $2n\log(1/\tau)$. Covering it at resolution $1+\Theta(\varepsilon/n)$ makes the slice count quadratic in n. By comparison, Feng et al. (2024) approximate d_{TV} on products over $[q]^n$ in time $O((qn^2/\varepsilon)\log q \log(n/(\varepsilon d_{\mathrm{TV}}(P,Q))))$, where $\log(1/d_{\mathrm{TV}}(P,Q)) = O(\ell b_0)$ follows from $d_{\mathrm{TV}}(P,Q) \geq \max_i d_{\mathrm{TV}}(p_i, q_i)$. The width of our range costs an extra factor n. Our hardness reduction adapts that of Bhattacharyya et al. (2023) for d_{TV}, moving the randomness so that the hard end is $F = 0$ rather than $d_{\mathrm{TV}} = 0$. An FPRAS for d_{TV} on same-structure Bayes nets with efficient inference (Bhattacharyya et al., 2024) concerns a different objective. Work on f-divergences between Ising models (Feng & Fu, 2026) and d_{TV} between spin systems (Feng et al., 2025; Bhattacharyya et al., 2025) uses a different input model. Further related work is in Section E.

Outline. Section 2 sets notation and proves the universal lower bound. Section 3 discusses admissibility and lists the functionals. Section 4 gives the FPTAS and Section 5 the hardness result. The appendices hold the proofs for Sections 4 and 5, the naive-Bayes example, the motivation for each functional, further related work, the verification of admissibility, exact computation on products, and open problems.

2 PRELIMINARIES

We write $[n] := \{1, \ldots, n\}$, and $\log$ denotes the natural logarithm. For a finite set D, $\Delta(D)$ denotes the set of probability distributions on D.

Definition 2.1 (FPTAS)**.** A function $g\colon \{0,1\}^* \to \mathbb{R}_{\geq 0}$ admits a *fully polynomial-time approximation scheme* (FPTAS) if there is a deterministic algorithm $\mathcal{A}$ that, on every input $x \in \{0,1\}^*$ and $\varepsilon \in (0,1]$, outputs a value $\widehat{g}(x,\varepsilon)$ satisfying

$$\frac{g(x)}{1+\varepsilon} \leq \widehat{g}(x,\varepsilon) \leq (1+\varepsilon)g(x),$$

in time $\mathrm{poly}(|x|, 1/\varepsilon)$.

Definition 2.2 (Product distribution)**.** A *product distribution* P over $[\ell]^n$ is given by marginal distributions $p_1, \ldots, p_n \in \Delta([\ell])$, and assigns probability $P(y) = \prod_{i=1}^n p_i(y_i)$ to every $y \in [\ell]^n$. The encoding size is $O(n\ell)$ rational numbers.

Definition 2.3 (Bayes net)**.** A *Bayes net* on a directed acyclic graph (DAG) G over node set $[n]$ specifies a distribution over $[\ell]^n$ via conditional probability tables (CPTs): each node $i \in [n]$ has a CPT $\mathbf{Pr}[X_i = \cdot | X_{\Pi(i)} = \cdot]$, where $\Pi(i) \subseteq [n]$ are the parents of i in G. The induced joint is $P(x) = \prod_{i=1}^n \mathbf{Pr}[X_i = x_i | X_{\Pi(i)} = x_{\Pi(i)}]$. The *in-degree* of the Bayes net is $\max_i |\Pi(i)|$. The encoding size is $O(n\ell^{1+d})$ where d is the in-degree.

Product distributions are exactly Bayes nets of in-degree 0.

Remark 2.4 (Input model). Both distribution classes are *succinct*: a product over $[\ell]^n$ has an $O(n\ell)$-entry marginal table, and a Bayes net of in-degree d has $O(n\ell^{d+1})$-entry CPTs; probabilities are rationals of bit-complexity at most $b_0 = \mathrm{poly}\,(n)$. "Exact computation" of $F(P,Q)$ means an exact rational when f is piecewise rational, or arbitrary precision in time polynomial in the precision when f is smooth; "approximation" is multiplicative throughout (cf. Theorem 2.1). This model — standard in distributional-inference complexity (Bhattacharyya et al., 2023; Feng et al., 2024; Feng & Fu, 2026) — is complementary to (not competing with) sample access: even on $\{0,1\}^n$, $S_{\mathrm{stat}}(P,Q)$ can be as small as $\exp(-n)$, beyond the reach of estimators that average $\mathrm{poly}\,(n)$ samples, whose error is additive (cf. Canonne, 2022 for sample-complexity background). It is the natural setting for graphical-model inference (Koller & Friedman, 2009; Wainwright & Jordan, 2008), naive-Bayes classifiers, and structured generative models with discrete latent state.

Definition 2.5 (Expectational functional)**.** Let $f\colon [0,\infty) \to [0,\infty)$. The *$f$-expectational functional* of distribution pair $(P,Q) \in \Delta(D)^2$ is

$$F(P,Q) = F_f(P,Q) := \mathop{\mathbf{E}}_{x\sim Q}[f(P(x)/Q(x))] = \sum_{x\in D: Q(x)>0} Q(x)\cdot f(P(x)/Q(x))\,.$$

The same formula extends F to non-negative (sub-probability) measures P, Q on D, used as shorthand in motivating examples (e.g. scaled class-conditionals $\alpha_i P_i$).

Definition 2.6 (Admissible integrands)**.** We call $f\colon [0,\infty) \to [0,\infty)$ *admissible* if f is concave on $[0,\infty)$, $f(0) = 0$, and $f \not\equiv 0$.

Admissibility is the only hypothesis on f in the two main theorems (the FPTAS also takes a rational $\varphi \in (0, f(1)]$ and an evaluation oracle for f as input). Every admissible f is non-decreasing (Theorem 2.7(d)), so a larger per-atom likelihood ratio never lowers the contribution, and it has a slope $C_f := \lim_{r\downarrow 0} f(r)/r \in (0,\infty]$ at the origin (Theorem 2.7(a)). The slope is finite for every entry of Table 1 except the two square-root overlaps, such as $f(r) = \min(\sqrt{r}, 1)$, for which $f(r)/r \to \infty$ as $r \downarrow 0$; neither theorem depends on it.

Lemma 2.7 (Consequences of admissibility)**.** *Let f be admissible, and let $r \geq 0$ and $\lambda \geq 1$.* (a) *The limit $C_f := \lim_{u\downarrow 0} f(u)/u$ exists in $(0,\infty]$ and equals $\sup_{u>0} f(u)/u$; in particular $f(r) \leq C_f r$ when $C_f < \infty$.* (b) $f(\lambda r) \leq \lambda f(r)$ (subhomogeneity). (c) $f(1) > 0$*, and* $f(r) \geq \min(r,1) f(1) > 0$ *for every $r > 0$.* (d) *f is non-decreasing on $[0,\infty)$.*

Proof. We first prove that the chord-slope $s(r) := f(r)/r$ is non-increasing on $(0,\infty)$. Fix $0 < r_1 \leq r_2$ and put $\theta := r_1/r_2 \in (0,1]$, so that $r_1 = \theta r_2 + (1-\theta)\cdot 0$. Concavity together with $f(0) = 0$ gives

$$f(r_1) = f\left(\theta r_2 + (1-\theta)\cdot 0\right) \geq \theta f(r_2) + (1-\theta) f(0) = \theta f(r_2) = \frac{r_1}{r_2} f(r_2),$$

and dividing by $r_1 > 0$ yields $s(r_1) \geq s(r_2)$; hence s is non-increasing. *(d)* Suppose $f(a) > f(b)$ for some $0 \leq a < b$. For $r > b$ write $b = \theta a + (1-\theta) r$ with $\theta := (r-b)/(r-a) \in (0,1)$. Concavity gives $f(b) \geq \theta f(a) + (1-\theta) f(r)$, that is, $(1-\theta) f(r) \leq f(b) - \theta f(a)$. As $r \to \infty$, $\theta \to 1$, so the right-hand side tends to $f(b) - f(a) < 0$ and is negative for all large r; since $1 - \theta > 0$, this gives $f(r) < 0$, contradicting $f \geq 0$. Hence f is non-decreasing. *(c)* Suppose $f(1) = 0$. Then $0 \leq f(r) \leq f(1) = 0$ for $r \in [0,1]$ by (d), and for $r > 1$ monotonicity of s gives $f(r)/r \leq f(1)/1 = 0$, so $f(r) = 0$; hence $f \equiv 0$, which admissibility excludes. So $f(1) > 0$. For $r \in (0,1]$, monotonicity of s gives $f(r)/r \geq f(1)$, i.e., $f(r) \geq r f(1)$; for $r > 1$, (d) gives $f(r) \geq f(1)$. In both cases $f(r) \geq \min(r,1) f(1) > 0$. *(a)* Since s is non-increasing on $(0,\infty)$, its limit as $u \downarrow 0$ exists in $(-\infty,\infty]$ and equals $\sup_{u>0} s(u)$, which is at least $s(1) = f(1) > 0$ by (c). If $C_f < \infty$, then $s(r) \leq C_f$ for every $r > 0$, i.e., $f(r) \leq C_f r$; for $r = 0$, $f(0) = 0 = C_f \cdot 0$. *(b)* For $r > 0$ and $\lambda \geq 1$ we have $r \leq \lambda r$, so $s(\lambda r) \leq s(r)$, i.e., $f(\lambda r)/(\lambda r) \leq f(r)/r$, i.e., $f(\lambda r) \leq \lambda f(r)$; for $r = 0$, both sides equal 0. □

Lemma 2.8 (Universal lower bound)**.** *Let f be admissible, let P, Q be product distributions over $[\ell]^n$, and let τ be the smallest nonzero marginal probability across $\{p_i, q_i\}_{i=1}^n$. If $F(P,Q) > 0$, then $F(P,Q) \geq f(1) \cdot \tau^n$.*

Proof. Since $F(P,Q)$ is a sum of non-negative terms, $F(P,Q) > 0$ gives some z with $Q(z) > 0$ and $f(P(z)/Q(z)) > 0$; then $P(z) > 0$ as well, since otherwise $f(P(z)/Q(z)) = f(0) = 0$. Put $x^* := z$. Since $P(x^*) = \prod_i p_i(x_i^*) > 0$ and $Q(x^*) = \prod_i q_i(x_i^*) > 0$, every factor $p_i(x_i^*)$ and $q_i(x_i^*)$ is a nonzero marginal probability, hence at least τ, so $P(x^*), Q(x^*) \geq \tau^n$. With $r^* := P(x^*)/Q(x^*)$: if $r^* \leq 1$, applying Theorem 2.7(b) with $\lambda = 1/r^* \geq 1$ gives $f(1) \leq f(r^*)/r^*$, so $Q(x^*) f(r^*) \geq f(1) P(x^*) \geq f(1)\tau^n$; if $r^* > 1$, Theorem 2.7(d) gives $f(r^*) \geq f(1)$, so $Q(x^*) f(r^*) \geq f(1) Q(x^*) \geq f(1)\tau^n$. Either way $F(P,Q) \geq Q(x^*) f(r^*) \geq f(1)\tau^n$ (dropping the remaining terms of F, all non-negative since $f \geq 0$), with $f(1) > 0$ by Theorem 2.7(c). □

Remark 2.9 (The bound is attained). Take $\ell = 2$, identifying $[2]$ with $\{0,1\}$; let P be the point mass at 0^n and Q the uniform distribution on $\{0,1\}^n$, both products, so that $\tau = 1/2$. For $f(r) = \min(r,1)$, since $Q(x) = 2^{-n} > 0$ for every x and only $x = 0^n$ has $P(x) > 0$, we get $F(P,Q) = \sum_x \min(P(x), Q(x)) = \min(1, 2^{-n}) = 2^{-n} = f(1)\tau^n$. In contrast with d_{TV}, where a single coordinate already gives $d_{\mathrm{TV}}(P,Q) \geq \max_i d_{\mathrm{TV}}(p_i, q_i)$, no lower bound on F in terms of a single coordinate exists: here every marginal pair has $S_{\mathrm{stat}}(p_i, q_i) = 1/2$, while $S_{\mathrm{stat}}(P,Q) = 2^{-n}$, so any lower bound must decay exponentially in n. The same instance shows that a $(1+\varepsilon)$-approximation of $d_{\mathrm{TV}} = 1 - S_{\mathrm{stat}}$ says nothing multiplicative about S_{stat} unless its additive error $\varepsilon d_{\mathrm{TV}}(P,Q)$ falls below 2^{-n}: here $d_{\mathrm{TV}}(P,Q) = 1 - 2^{-n}$, and when $\varepsilon d_{\mathrm{TV}}(P,Q) \geq 2^{-n}$, that is, $(1+\varepsilon) d_{\mathrm{TV}}(P,Q) \geq 1$, the output 1 is a valid $(1+\varepsilon)$-approximation of $d_{\mathrm{TV}}(P,Q)$ and yields the estimate 0 for $S_{\mathrm{stat}}(P,Q) = 2^{-n}$.

Definition 2.10 (Ratio random variable, J-sparsification)**.** For distributions P, Q on D, let $R := P\|Q$ denote the random variable taking value $P(x)/Q(x)$ with probability $Q(x)$ (over x with $Q(x) > 0$). For ratios $R_i := P_i\|Q_i$, the *independent product* $R_1 \cdot R_2$ takes value $(P_1(x)/Q_1(x))(P_2(y)/Q_2(y))$ with probability $Q_1(x)Q_2(y)$. For a discrete random variable S supported on $[0, B]$ and a partition J of $[0, B]$ into measurable sets, the *J-sparsification* $\widetilde{S}$ takes value $\mathbf{E}[S|S \in J_k]$ with probability $\mathbf{Pr}[S \in J_k]$ for each part $J_k \in J$ with $\mathbf{Pr}[S \in J_k] > 0$ (and is undefined / contributes nothing on parts of zero probability) (Feng et al., 2024).

3 Admissible Functionals

This section unpacks admissibility (Theorem 2.6) and lists a broad family of admissible functionals; verifications appear in Section F.

3.1 Why Admissibility Matters

Concavity with $f(0) = 0$. The geometric content is the chord-slope inequality: $f(r)/r$ is non-increasing in $r > 0$, equivalently $f(\lambda r) \le \lambda f(r)$ for $\lambda \ge 1$ (Theorem 2.7(b)). This is the property that drives the FPTAS: when sparsification rounds a value r up to a slightly larger value λr, the functional value grows by at most a factor λ, allowing multiplicative error to be controlled. Concavity also expresses a *saturation* principle natural to overlap functionals: the marginal effect of additional likelihood-ratio mass diminishes.

Monotonicity. A non-negative concave function cannot decrease (Theorem 2.7(d)): as the per-atom likelihood ratio $r = P(x)/Q(x)$ grows, the contribution $f(r)$ does not decrease, so more relative mass under P means a larger overlap contribution. The FPTAS uses this to round values down as well as up.

Not identically zero. This excludes only the trivial $f \equiv 0$, and it gives $f(1) > 0$ (Theorem 2.7(c)), which is all the hardness reduction needs besides $f(0) = 0$.

No growth condition at the origin. The slope $C_f = \lim_{r\downarrow 0} f(r)/r = \sup_{r>0} f(r)/r \in (0, \infty]$ (Theorem 2.7(a)) may be infinite. The FPTAS never rounds a value near 0, because every positive partial-product ratio is at least τ^n (Theorem A.2). So the square-root overlaps $\min(\sqrt{r}, 1)$ and $2\sqrt{r}/(1 + \sqrt{r})$, which have $C_f = \infty$, are covered.

3.2 Representative Functionals

Table 1 lists representative functionals, with the slope C_f at the origin, the value $f(1)$ and the status of exact computation on products.

Table 1: Representative admissible functionals. The two main theorems apply to every row; the last column records the status of exact computation on products for rational parameters. Here C_f is the slope at the origin (Theorem 2.7(a)).

$f(r)$	Name	C_f	$f(1)$	Exact on products
$\min(r, 1)$	S_{stat}	1	1	#P-hard (Theorem G.2)
$\min(r, c)$, $c \in (0, 1]$	Capped S_{stat}	1	c	#P-hard (Theorem G.5)
$\min(r, c)$, $c > 1$	Truncated S_{stat}	1	1	#P-hard (Theorem G.5)
$\log(1 + r)$	Log-overlap	1	$\log 2$	open
$1 - e^{-r}$	Exp. overlap	1	$1 - 1/e$	open
$r/(1 + r)$	Harmonic overlap	1	$1/2$	open
$\tanh r$	tanh overlap	1	$\tanh 1$	open
$\arctan r$	arctan overlap	1	$\pi/4$	open
$\min(\log(1 + r), \beta)$, $\beta > 0$	Bounded log	1	$\min(\log 2, \beta)$	partial
$\min(1 - e^{-r}, c)$, $c \in (0, 1)$	Capped exp.	1	$\min(1 - 1/e, c)$	partial
$\min(\sqrt{r}, 1)$	Capped root	∞	1	open
$2\sqrt{r}/(1 + \sqrt{r})$	Root harmonic	∞	1	open
$\theta g_1(r) + (1 - \theta)g_2(r)$	Convex combs.	see text	$\theta g_1(1) + (1 - \theta)g_2(1)$	see text

The list is illustrative: pointwise minima and convex combinations of admissible functions are again admissible (Theorems F.5 and F.6). In the last column, "open" means that no #P-hardness reduction is known, while "partial" means that a capped functional coincides with its uncapped version on instances where the cap never fires (Theorem G.13). Every fixed row has the evaluation oracle and rational lower bound on $f(1)$ required by Theorem 4.1 (Section F); the convex-combination row does when its components do. In that row g_1, g_2 are admissible, $\theta \in [0,1]$, and $C_f = \theta C_{g_1} + (1-\theta)C_{g_2}$, with $0 \cdot \infty := 0$; exact computation is #P-hard when $\theta \in (0,1]$ is rational, F_{g_1} is #P-hard, and F_{g_2} is polynomial-time computable (Theorem G.8). What admissibility leaves out is instructive. Convex integrands such as r^p for $p > 1$ (the moments of the likelihood ratio, including $\chi^2 + 1$) and $r \log r$ (Kullback–Leibler) are not concave; but these, like the Hellinger and Rényi quantities built from r^α (admissible for $\alpha \in (0,1)$, convex for $\alpha > 1$), factorize over the coordinates of a product, into a product or a sum of n one-dimensional quantities, and so are computed to any desired precision in polynomial time; on products they pose no approximation problem (Bhattacharyya et al., 2023). The quantities that do not factorize are d_{TV} and the concave overlaps of Table 1, and these are what the two theorems address.

4 Generic FPTAS

Theorem 4.1 (Generic FPTAS). *Let $f\colon [0,\infty) \to [0,\infty)$ be admissible, and suppose we are given a rational φ with $0 < \varphi \le f(1)$ and an oracle that, on a rational $v \ge 0$ and an integer $t \ge 1$, returns a rational within 2^{-t} of $f(v)$. There is a deterministic algorithm that, on input product distributions P, Q over $[\ell]^n$ whose marginals are rationals of bit-length at most b_0 and an accuracy parameter $\varepsilon \in (0,1]$, outputs $\widehat{F}$ with*

$$\frac{F(P,Q)}{1+\varepsilon} \le \widehat{F} \le (1+\varepsilon)F(P,Q).$$

It performs $O(n\ell m \log m)$ arithmetic operations and comparisons on rationals of bit-length $O(n\ell b_0 + m\log(2n/\varepsilon) + \beta)$, where β is the total bit-length of φ and ε, together with $O(m)$ operations on the oracle's answers, and at most m oracle calls with $t = O(nb_0 + \log(2/\varepsilon) + \log(1+1/\varphi))$, where $\tau \ge 2^{-b_0}$ is the smallest nonzero marginal probability and m, the number of geometric slices of the partition built by Algorithm 1, satisfies

$$m = O\left(\frac{n}{\varepsilon}\left(n\log\frac{1}{\tau} + 1\right)\right).$$

No growth condition at the origin is needed: the theorem covers every non-negative concave f with $f(0) = 0$ other than $f \equiv 0$ (see the paragraph after Theorem 2.6), including the square-root overlaps of Table 1, whose slope at the origin is infinite. For every entry of Table 1 with its parameters fixed, $f(1)$ is a positive constant, so with φ a fixed rational in $(0, f(1)]$ and ε rounded as below, the bit-lengths above are $O(n\ell b_0 + m\log(2n/\varepsilon))$ with $m = O(n^2\log(1/\tau)/\varepsilon + n/\varepsilon)$. The algorithm is therefore an FPTAS in the sense of Theorem 2.1 whenever the bit-length of φ is polynomial in the input length and the oracle runs in polynomial time; the bit-length of ε may be taken to be $O(\log(2/\varepsilon))$ by running with $2^{-\lceil \log_2(1/\varepsilon)\rceil} \in (\varepsilon/2, \varepsilon]$ in place of ε.

Algorithmic idea. The algorithm processes coordinates one at a time, maintaining a discrete random variable $\widetilde{Y}_i$ that approximates the partial product $R_1 \cdots R_i$ of marginal ratios. After each coordinate, we apply J-sparsification (Theorem 2.10): partition $[0, B]$ into a singleton $\{0\}$, a small interval $(0, \gamma]$, and geometric slices of ratio $1+\delta$, replacing the values within each slice by the conditional expectation. This keeps the sparsified support size at $O(m)$, while the value within each slice differs from the true value by at most a $(1+\delta)$ factor. In each direction, monotonicity (Theorem 2.7(d)) replaces one value by the other inflated by $1+\delta$, and subhomogeneity $f(\lambda r) \le \lambda f(r)$ for $\lambda \ge 1$ (Theorem 2.7(b)) turns that into a $(1+\delta)$ factor on the functional. Near 0 rounding would have unbounded relative error, but it never happens: every positive partial-product ratio is at least τ^n (Theorem A.2), so the near-zero slice $(0, \gamma]$ with $\gamma = \tau^n/2$ receives no atom, and the sparsification incurs no additive error at all. This is why no growth condition on f at the origin is needed. The same bound, with Theorem 2.7(c), gives $f(v) \ge \varphi\tau^n$ on every positive atom, which fixes the precision of the final evaluation. Sparse-partial-product representations of this general type were used by Feng et al. (2024) for d_{TV}; there the slice adjacent to $r = 1$ does carry an additive

error, charged against the coordinatewise lower bound $d_{\mathrm{TV}}(P,Q) \geq \max_i d_{\mathrm{TV}}(p_i, q_i)$, which has no analogue for F (Theorem 2.9).

4.1 Algorithm and Proof

Setup. Fix f and φ as in Theorem 4.1, product distributions P, Q over $[\ell]^n$, and $\varepsilon \in (0,1]$; write $\varepsilon_1 := \varepsilon/2$ for the accuracy of the sparsification stage. Let τ be the smallest nonzero marginal probability of P or Q, and define

$$B := \tau^{-n},$$
$$\delta := 2^{-k} \text{ for the least integer } k \geq 1 \text{ with } (1+2^{-k})^n \leq 1 + \varepsilon_1/2,$$
$$\gamma := \frac{1}{2B} = \frac{\tau^n}{2}.$$

Here B upper-bounds the per-atom ratios $P(x)/Q(x)$. Let $m := \lceil \log(B/\gamma)/\log(1+\delta) \rceil$ and define the partition

$$J := \{\{0\}, I_0, I_1, \ldots, I_m\}, \qquad I_0 := (0, \gamma], \quad I_j := (\gamma(1+\delta)^{j-1}, \gamma(1+\delta)^j] \text{ for } j \geq 1.$$

$I_1, \ldots, I_m$ are geometric slices of ratio $1+\delta$ covering $(\gamma, B]$; I_0 is the only part on which rounding could change a value by an unbounded factor, but since $\gamma < \tau^n$ and every positive value of Y_i is at least τ^n (Theorem A.2), no atom ever falls into I_0; $\{0\}$ holds atoms with $P(x) = 0$. Replacing I_m by $I_m \cap [0, B]$ makes J a partition of $[0, B]$ as in Theorem 2.10, and does not change the sparsification of any variable supported on $[0, B]$. Theorem A.1 in Section A shows that $(1+\delta)^n \leq 1 + \varepsilon_1/2$ and that m obeys the bound of Theorem 4.1.

Let $R_i := p_i \| q_i$, $Y_1 := R_1$, and $Y_{i+1} := \widetilde{Y}_i \cdot R_{i+1}$, where $\widetilde{Y}_i$ is the J-sparsification of Y_i. Let $Z_i := R_{i+1} \cdots R_n$ (with $Z_n := 1$); the global ratio satisfies $R = Y_1 \cdot Z_1$ and the iteration property $Y_{i+1} \cdot Z_{i+1} = \widetilde{Y}_i \cdot Z_i$ in law for $1 \leq i \leq n-1$ (both sides are $\widetilde{Y}_i \cdot R_{i+1} \cdots R_n$ with independent factors). Every Y_i is supported on $[0, B]$, so the J-sparsification is well-defined throughout: conditional means never exceed the maximum value, whence inductively $\max Y_{i+1} \leq \max \widetilde{Y}_i \cdot \max R_{i+1} \leq \tau^{-(i+1)} \leq B$. *Notation:* for a non-negative random variable S, write $F(S) := \mathbf{E}[f(S)]$; this extends Theorem 2.5 since $F(P,Q) = F(R) = \mathbf{E}[f(R)]$ for $R = P \| Q$.

Algorithm 1 Generic FPTAS for $F(P,Q)$.

Require: Marginal tables of product distributions P, Q over $[\ell]^n$; a rational φ and an evaluation oracle for f as in Theorem 4.1; accuracy $\varepsilon \in (0,1]$.
Ensure: $\widehat{F} \in [F(P,Q)/(1+\varepsilon), (1+\varepsilon)F(P,Q)]$.

for $i = 1, \ldots, n$ **do**
  if $\mathrm{supp}(p_i) \cap \mathrm{supp}(q_i) = \emptyset$ **then**
    return 0
Compute τ, B, δ, γ as in the setup, with $\varepsilon_1 = \varepsilon/2$
Compute the thresholds $\gamma(1+\delta)^j$ for $j = 0, 1, \ldots$ up to the first with $\gamma(1+\delta)^j \geq B$; this j is m
Build $J = \{\{0\}, (0, \gamma], (\gamma, \gamma(1+\delta)], \ldots, (\gamma(1+\delta)^{m-1}, B]\}$
$Y_1 \leftarrow p_1 \| q_1$
for $i = 1, \ldots, n$ **do**
  $\widetilde{Y}_i \leftarrow J$-sparsification of Y_i
  if $i < n$ **then**
    $Y_{i+1} \leftarrow \widetilde{Y}_i \cdot (p_{i+1} \| q_{i+1})$
$t \leftarrow \max\{1, \lceil \log_2(8/(\varepsilon\varphi\tau^n)) \rceil\}$
return $\widehat{F} := \sum_v \mathbf{Pr}[\widetilde{Y}_n = v] \cdot \widehat{f}(v)$, where $\widehat{f}(0) := 0$ and $\widehat{f}(v)$ is the oracle's answer on (v, t) for $v > 0$

Correctness. If any i has $\mathrm{supp}(p_i) \cap \mathrm{supp}(q_i) = \emptyset$, the algorithm returns 0 correctly: at every z with $Q(z) > 0$, $q_i(z_i) > 0$ forces $p_i(z_i) = 0$, so $P(z) = 0$ and $F(P,Q) = \sum_z Q(z) f(0) = 0$. Otherwise each pair has overlapping support, the joint atom $(x_1, \ldots, x_n)$ with $p_i(x_i), q_i(x_i) > 0$ has $P, Q > 0$, so $F(P,Q) > 0$ (its term $Q(x) f(P(x)/Q(x))$ is positive, since the ratio is positive and

$f > 0$ on $(0, \infty)$ by Theorem 2.7(c), while all other terms are non-negative), and the sparsification stage is controlled by the next two lemmas, applied with accuracy ε_1.

Lemma 4.2 (Upper bound). *If $F(P, Q) > 0$, then $F(\widetilde{Y}_n) \le (1 + \varepsilon_1) F(P, Q)$.*

Lemma 4.3 (Lower bound). *If $F(P, Q) > 0$, then $F(\widetilde{Y}_n) \ge F(P, Q)/(1 + \varepsilon_1)$.*

Both are proved in Section A. The proofs share one inflation step: on a slice I_j with $j \ge 1$ the rounded and true values differ by a factor at most $1 + \delta$, which monotonicity and subhomogeneity turn into a factor at most $1 + \delta$ on f, while on $\{0\} \cup I_0$ both values are 0, since I_0 receives no atom, so both contribute $f(0) = 0$. Over n rounds the factors multiply to $(1 + \delta)^n \le 1 + \varepsilon_1/2$. The final line of Algorithm 1 evaluates f with relative error at most $\varepsilon/8$ on every atom, which together with Theorems 4.2 and 4.3 gives the guarantee of Theorem 4.1; the running time and the bit-lengths are counted in Section A, which completes the proof of Theorem 4.1.

4.2 Applications

Theorem 4.1 instantiates uniformly to every entry of Table 1, including the two square-root overlaps, whose slope at the origin is infinite; for the convex-combinations row see Section F.13. The instance of most direct use is the Bayes-optimal error of a binary classifier with product class-conditionals: with priors $\alpha_0, \alpha_1 > 0$ and $c := \alpha_1/\alpha_0$,

$$R^* = \sum_x \min\left(\alpha_0 P_0(x), \alpha_1 P_1(x)\right) = \alpha_0 \sum_x \min\left(P_0(x), c\, P_1(x)\right) = \alpha_0\, F_{\min(r,c)}(P_0, P_1),$$

and $f(r) = \min(r, c)$ is admissible, with $f(1) = \min(1, c)$, which is a rational φ for rational priors, so Theorem 4.1 computes R^* to within a factor $1 + \varepsilon$ for naive-Bayes models with arbitrary per-feature marginals. Multiplicative accuracy is what the question calls for, since R^* is typically exponentially small in the number of features; Theorem C.1 shows that the bracket on R^* available from the divergences that tensorize can be exponentially wide, its lower end falling below R^* by a factor exponential in n.

5 Generic Hardness

Theorem 5.1 (Generic hardness). *Let f be admissible; in particular, f may be any entry of Table 1. Given Bayes-net distributions P, Q of in-degree at most 2, it is* NP*-complete to decide whether $F(P, Q) > 0$. Hence, unless* P $=$ NP*, no polynomial-time algorithm approximates $F(P, Q)$ within any multiplicative factor on Bayes nets of in-degree 2.*

The reduction adapts the construction of Bhattacharyya et al. (2023), who proved that deciding whether $d_{\mathrm{TV}}(P, Q) \ne 0$ is NP-complete for Bayes nets of in-degree at least 2. By Scheffé's identity $S_{\mathrm{stat}} = 1 - d_{\mathrm{TV}}$, that result says that deciding $S_{\mathrm{stat}}(P, Q) < 1$ is NP-hard, which concerns the upper end $S_{\mathrm{stat}} = 1$ rather than the lower end that multiplicative approximation of S_{stat} must resolve. What is needed is the other end, $F(P, Q) = 0$ against $F(P, Q) > 0$, and since multiplicative approximation does not transfer between a functional and its complement, such as S_{stat} and $d_{\mathrm{TV}} = 1 - S_{\mathrm{stat}}$ (Theorem 2.9), the construction is re-engineered so that the output gate pins Q rather than randomizing P.

Construction. Given a CNF formula ϕ on n variables, view it as a circuit of AND, OR and NOT gates of fan-in at most 2, with internal gates $Y_1, \ldots, Y_m$ in topological order and Y_m the output, and let G_ϕ be its DAG on $n + m$ nodes. Under both P and Q, each input is a uniform bit and each gate $Y_1, \ldots, Y_{m-1}$ is the deterministic function of its parents that the circuit prescribes; the output gate Y_m computes $\phi(x)$ under P, while under Q it is the constant 1. Then P and Q are Bayes nets on G_ϕ of in-degree at most 2 with probabilities in $\{0, 1/2, 1\}$, and

$$F(P, Q) = f(1) \cdot \frac{|\mathrm{Sol}(\phi)|}{2^n}, \qquad \mathrm{Sol}(\phi) := \{x \in \{0,1\}^n : \phi(x) = 1\}, \tag{5.1}$$

so $F(P, Q) > 0$ if and only if ϕ is satisfiable. The identity uses only $f(0) = 0$, and the equivalence only $f(1) > 0$, which is why one reduction serves the whole family. The full proof of Theorem 5.1, including Equation (5.1) and membership in NP, is in Section B.

5.1 APPLICATIONS

Theorem 5.1 yields, uniformly across Table 1, NP-hardness of multiplicative approximation of F on Bayes nets of in-degree 2. For S_{stat} it gives NP-completeness of deciding $S_{\mathrm{stat}}(P,Q) > 0$, and so inapproximability within every multiplicative factor, where the result of Bhattacharyya et al. (2023) concerns the opposite end $S_{\mathrm{stat}} = 1$. Since Equation (5.1) recovers the integer $|\mathrm{Sol}(\phi)| \in [0, 2^n]$ from $F(P,Q)$ and $f(1)$, exact computation of F on Bayes nets of in-degree 2 is #P-hard, in the sense of Theorem 2.4, for every f whose value $f(1)$ can be computed to additive error $f(1)2^{-n}/8$ (that is, to $n + 3 + \log_2(1/f(1))$ bits) in polynomial time, which includes every entry of Table 1: if $\widetilde{F}$ and $\widetilde{c}$ are within $f(1)2^{-n}/8$ of $F(P,Q)$ and $f(1)$, then $\widetilde{c} \geq 7f(1)/8$ and $|2^n\widetilde{F}/\widetilde{c} - |\mathrm{Sol}(\phi)|| \leq (f(1)/8 + f(1)/8)/(7f(1)/8) = 2/7 < 1/2$, so rounding recovers $|\mathrm{Sol}(\phi)|$. On a common DAG, the FPTAS-versus-hardness boundary therefore sits between in-degree 0 and in-degree 2. If P and Q may use different DAGs, deciding $F > 0$ is NP-complete already at in-degree 1 over a ternary alphabet (Theorem B.1, by reduction from 3-colourability); on a common DAG of in-degree 1, a forest, it is in polynomial time (Theorem B.2), so there only multiplicative approximation remains open (Section H). In contrast, additive approximation of S_{stat} is easy on every Bayes net: $S_{\mathrm{stat}}(P,Q) = \mathbf{E}_{x\sim Q}[\min(P(x)/Q(x), 1)]$ is the mean of a $[0,1]$-valued quantity that can be sampled and evaluated exactly, so an empirical mean over $\lceil \ln(2/\beta)/(2\varepsilon^2) \rceil$ samples of $x \sim Q$, drawn along a topological order, is within ε of it with probability at least $1 - \beta$ by Hoeffding's inequality (Hoeffding, 1963); the difficulty our results address is specific to multiplicative error.

6 CONCLUSION

We establish a tractability divide for concave expectational functionals. Given an evaluation oracle and a rational lower bound on $f(1)$, every admissible integrand has a deterministic FPTAS on products. On Bayes nets of in-degree 2, deciding whether its functional is positive is NP-complete. The algorithm uses monotonicity and subhomogeneity to control sparsification. Positive partial likelihood-ratio products stay above τ^n, so no growth condition at the origin is needed. The hardness reduction uses only $f(0) = 0$ and $f(1) > 0$.

Two boundaries remain open. On a common DAG of in-degree 1, the zero test is easy, but multiplicative approximation is unresolved; on different DAGs the zero test is already hard. On products, exact computation of the smooth overlaps is unresolved, although the capped similarities are #P-hard. Section H discusses both questions.

AI USE STATEMENT

The authors formulated the problem and developed the initial proof sketches. Claude (Anthropic) helped identify the conditions needed for the FPTAS and NP-hardness results, expand the sketches into complete proofs, find relevant literature, and improve the prose. Codex (OpenAI) assisted with reviewing the arguments, citations, and presentation of the ICLR submission. The authors reviewed the AI-assisted work and take responsibility for the paper and any errors.

Under the ICLR 2027 AI Policy for Authors, the required-disclosure tasks include developing the conceptual framework, formulating mathematical claims, providing proof ingredients, and writing proofs. The recommended-disclosure tasks include drafting, literature search and summary, and editing for readability. No experiments, data, research code, or translation were involved.

REPRODUCIBILITY STATEMENT

Every result in this paper is analytical, and there are no experiments, data or code. Admissibility of the integrand is defined in Theorem 2.6, the input model in Theorem 2.4, and each result carries a complete proof, either in the body or in the appendix with a pointer beside its statement. Each result imported from the literature is stated with its citation and is used only as stated.

REFERENCES

Milton Abramowitz and Irene A. Stegun (eds.). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards, Washington, DC, 1964.

Bengt Aspvall, Michael F. Plass, and Robert Endre Tarjan. A linear-time algorithm for testing the truth of certain quantified Boolean formulas. *Information Processing Letters*, 8(3):121–123, 1979. doi: 10.1016/0020-0190(79)90002-4.

Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, Dimitrios Myrisiotis, A. Pavan, and N. V. Vinodchandran. On approximating total variation distance. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023*, pp. 3479–3487. ijcai.org, 2023.

Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, Dimitrios Myrisiotis, A. Pavan, and N. V. Vinodchandran. Total variation distance meets probabilistic inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3776–3794. PMLR, 2024. URL https://proceedings.mlr.press/v235/bhattacharyya24a.html.

Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, Dimitrios Myrisiotis, A. Pavan, and N. V. Vinodchandran. Computational explorations of total variation distance. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/af9c083d5d6bf8b860dbe57051c65984-Paper-Conference.pdf.

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

Andrei A. Bulatov. The complexity of the counting constraint satisfaction problem. *Journal of the ACM*, 60(5):34:1–34:41, 2013. doi: 10.1145/2528400.

Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer, 1986.

Clément L. Canonne. Topics and techniques in distribution testing: A biased but representative sample. *Foundations and Trends in Communications and Information Theory*, 19(6):1032–1198, 2022. doi: 10.1561/0100000114.

Gregory F Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(2-3):393–405, 1990.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006. doi: 10.1002/047174882X.

Imre Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

Paul Dagum and Michael Luby. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 60(1):141–153, 1993.

René Descartes. *La Géométrie*. Jan Maire, Leyden, 1637. Appendix to *Discours de la méthode*; origin of the rule of signs.

Pierre A. Devijver. A multiclass, k-NN approach to Bayes risk estimation. *Pattern Recognition Letters*, 3(1):1–6, 1985. doi: 10.1016/0167-8655(85)90035-2.

Martin Dyer, Leslie Ann Goldberg, Catherine Greenhill, and Mark Jerrum. The relative complexity of approximate counting problems. *Algorithmica*, 38(3):471–500, 2004.

Weiming Feng and Yucheng Fu. On approximating the f-divergence between two Ising models. In *17th Innovations in Theoretical Computer Science Conference, ITCS 2026*, volume 362 of *LIPIcs*, pp. 59:1–59:23. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2026. doi: 10.4230/LIPIcs.ITCS.2026.59.

Weiming Feng, Heng Guo, Mark Jerrum, and Jiaheng Wang. A simple polynomial-time approximation algorithm for the total variation distance between two product distributions. *TheoretiCS*, 2:1–7, 2023. doi: 10.46298/theoretics.23.7.

Weiming Feng, Liqiang Liu, and Tianren Liu. On deterministically approximating total variation distance. In *Proceedings of the 2024 ACM-SIAM Symposium on Discrete Algorithms, SODA 2024, Alexandria, VA, USA, January 7-10, 2024*, pp. 1766–1791. SIAM, 2024.

Weiming Feng, Hongyang Liu, and Minji Yang. Approximating the total variation distance between spin systems. In Nika Haghtalab and Ankur Moitra (eds.), *The Thirty Eighth Annual Conference on Learning Theory, 30-4 July 2025, Lyon, France*, volume 291 of *Proceedings of Machine Learning Research*, pp. 1974–2025. PMLR, 2025. URL https://proceedings.mlr.press/v291/feng25a.html.

Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, second edition, 1990.

Keinosuke Fukunaga and Larry D. Hostetler. k-nearest-neighbor Bayes-risk estimation. *IEEE Transactions on Information Theory*, 21(3):285–293, 1975.

M. R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, 1986.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Russell Impagliazzo and Ramamohan Paturi. On the complexity of k-SAT. *Journal of Computer and System Sciences*, 62(2):367–375, 2001.

Takashi Ishida, Ikko Yamane, Nontawat Charoenphakdee, Gang Niu, and Masashi Sugiyama. Is the performance of my deep network too good to be true? A direct approach to estimating the Bayes error in binary classification. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=FZdJQgy05rz.

Richard M. Karp, Michael Luby, and Neal Madras. Monte-Carlo approximation algorithms for enumeration problems. *J. Algorithms*, 10(3):429–448, 1989.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: Principles and techniques*. MIT press, 2009.

Aryeh Kontorovich and Ariel Avital. Sharp bounds on aggregate expert error. In *The 36th International Conference on Algorithmic Learning Theory, 24-27 February 2025, Milan, Italy*, volume 272 of *Proceedings of Machine Learning Research*, pp. 653–663. PMLR, 2025. URL https://proceedings.mlr.press/v272/kontorovich25a.html.

Erich Leo Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses, Third Edition*. Springer texts in statistics. Springer, 2005.

Friedrich Liese and Igor Vajda. *Convex Statistical Distances*. Teubner, 1987.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

Torgny Lindvall. *Lectures on the Coupling Method*. Dover Publications, 2002. Reprint of the 1992 Wiley edition.

Jack Murtagh and Salil P. Vadhan. The complexity of computing the optimal composition of differential privacy. *Theory of Computing*, 14(8):1–35, 2018. doi: 10.4086/toc.2018.v014a008.

Frank Nielsen. Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means. *Pattern Recognition Letters*, 42:25–34, 2014.

Morteza Noshad, Li Xu, and Alfred Hero. Learning to benchmark: Determining best achievable misclassification error from training data. *arXiv preprint arXiv:1909.07192*, 2019. URL https://arxiv.org/abs/1909.07192.

Judea Pearl. *Probabilistic reasoning in intelligent systems - networks of plausible inference*. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann, 1988.

Dan Roth. On the hardness of approximate reasoning. *Artificial Intelligence*, 82(1-2):273–302, 1996.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 3rd edition, 1976.

Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

Volker Strassen. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965.

Ryan Theisen, Huan Wang, Lav R. Varshney, Caiming Xiong, and Richard Socher. Evaluating state-of-the-art classification models against bayes optimality. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9367–9377, 2021. URL https://proceedings.neurips.cc/paper/2021/file/4e0ccd2b894f717df5ebc12f4282ee70-Paper.pdf.

Flemming Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 46(4):1602–1609, 2000.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer series in statistics. Springer, 2009. ISBN 978-0-387-79051-0. doi: 10.1007/B13794. URL https://doi.org/10.1007/b13794.

Leslie G. Valiant. The complexity of computing the permanent. *Theoretical Computer Science*, 8 (2):189–201, 1979.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001.

Xiaoshen Wang. A simple proof of Descartes's rule of signs. *The American Mathematical Monthly*, 111(6):525–526, 2004.

A Proofs for Section 4

This appendix completes the proof of Theorem 4.1. Throughout, the notation is that of the setup in Section 4.1: $\varepsilon_1 = \varepsilon/2$, and $B = \tau^{-n}$, $\delta = 2^{-k}$, $\gamma = \tau^n/2$ and m are as defined there. We write $\delta_* := (1+\varepsilon_1/2)^{1/n} - 1$ for the real number with $(1+\delta_*)^n = 1 + \varepsilon_1/2$.

Lemma A.1 (The parameters)**.** *The parameters of Section 4.1 satisfy:* (a) $\delta_*/2 < \delta \le \delta_*$*, so* $(1+\delta)^n \le 1 + \varepsilon_1/2$*, and* $\varepsilon/(12n) \le \delta \le \varepsilon/(4n)$*;* (b) $2n/\varepsilon \le m = O\big(\frac{n}{\varepsilon}\left(n \log \frac{1}{\tau} + 1\right)\big)$.

Proof. (a) The map $x \mapsto (1+x)^n$ is increasing on $[0,\infty)$, so $(1+2^{-k})^n \le 1 + \varepsilon_1/2$ holds if and only if $2^{-k} \le \delta_*$. Since $\delta_* > 0$ and $2^{-k} \to 0$, such k exist, and δ is 2^{-k} for the least one, so $\delta \le \delta_*$. Moreover $\delta_* \le (1+\varepsilon_1/2) - 1 = \varepsilon_1/2 \le 1/4 < 2^{-1}$, using $(1+\varepsilon_1/2)^{1/n} \le 1 + \varepsilon_1/2$ for $n \ge 1$, so $k = 1$ does not qualify and $k \ge 2$; by minimality $2^{-(k-1)} > \delta_*$, so $\delta = 2^{-(k-1)}/2 > \delta_*/2$, and $\delta \le 1/4$. For the lower bound on δ, write $\delta_* = e^y - 1$ with $y := \log(1+\varepsilon_1/2)/n$; then $e^y - 1 \ge y$, and $\log(1+x) \ge x/(1+x)$ for $x \ge 0$ gives $y \ge (\varepsilon_1/2)/(n(1+\varepsilon_1/2))$. With $\varepsilon_1/2 = \varepsilon/4 \le 1/4$ this is at least $(\varepsilon/4)/(5n/4) = \varepsilon/(5n)$, so $\delta > \delta_*/2 \ge \varepsilon/(10n) \ge \varepsilon/(12n)$. For the upper bound, $1 + n\delta \le (1+\delta)^n \le 1 + \varepsilon_1/2 = 1 + \varepsilon/4$ by Bernoulli's inequality, so $\delta \le \varepsilon/(4n)$.

(b) Since $B/\gamma = 2\tau^{-2n}$, we have $\log(B/\gamma) = 2n\log(1/\tau)+\log 2$. By (a), $\log(1+\delta) \geq \delta/(1+\delta) \geq \delta/2 \geq \varepsilon/(24n)$, so

$$m = \left\lceil \frac{\log(B/\gamma)}{\log(1+\delta)} \right\rceil \leq 1 + \frac{24n}{\varepsilon}\left(2n\log\frac{1}{\tau} + \log 2\right) = O\left(\frac{n}{\varepsilon}\left(n\log\frac{1}{\tau}+1\right)\right),$$

where the leading 1 is absorbed because $n/\varepsilon \geq 1$. Conversely, $\log(1+\delta) \leq \delta \leq \varepsilon/(4n)$ by (a) and $\log(B/\gamma) \geq \log 2$ since $\tau \leq 1$, so $m \geq \log 2 \cdot 4n/\varepsilon \geq 2n/\varepsilon$. □

Lemma A.2 (Positive values)**.** *For every $i \in [n]$, every value $v > 0$ taken by Y_i or $\widetilde{Y}_i$ satisfies $\tau^i \leq v \leq \tau^{-i}$.*

Proof. First, for every $j \in [n]$, every positive value of $R_j = p_j \| q_j$ is $p_j(x)/q_j(x)$ for some x with $p_j(x), q_j(x) > 0$; both are nonzero marginal probabilities, so both lie in $[\tau, 1]$ by the definition of τ, and the value lies in $[\tau, \tau^{-1}]$. We now argue by induction on i. Since $Y_1 = R_1$, the claim holds for Y_1. Suppose every positive value of Y_i lies in $[\tau^i, \tau^{-i}]$. Such a value is at most $\tau^{-i} \leq \tau^{-n} = B \leq \gamma(1+\delta)^m$, using $\tau \leq 1$, $i \leq n$ and the choice of m, so it falls into one of the parts $I_0, \ldots, I_m$, all of which lie in $(0, \infty)$; the part $\{0\}$ holds only the value 0 and has conditional mean 0. Hence a positive value of $\widetilde{Y}_i$ is $\mathbf{E}[Y_i \mid Y_i \in I_j]$ for some $j \geq 0$ with $\mathbf{Pr}[Y_i \in I_j] > 0$, a convex combination with positive weights of values of Y_i in $[\tau^i, \tau^{-i}]$, so it lies in $[\tau^i, \tau^{-i}]$. Finally, a value of $Y_{i+1} = \widetilde{Y}_i \cdot R_{i+1}$ is a product vr of a value v of $\widetilde{Y}_i$ and a value r of R_{i+1}; both are non-negative, so $vr > 0$ forces $v > 0$ and $r > 0$, and then $v \in [\tau^i, \tau^{-i}]$ and $r \in [\tau, \tau^{-1}]$ give $vr \in [\tau^{i+1}, \tau^{-(i+1)}]$. □

A.1 Proof of the Upper Bound

A coupling. Both proofs compare $\widetilde{Y}_i$ with Y_i pointwise, so we put them on one probability space. Realize Y_i and Z_i as independent random variables; they are built from the disjoint sets of factors $R_1, \ldots, R_i$ (with the sparsifications of the earlier rounds) and $R_{i+1}, \ldots, R_n$, so $Y_i Z_i$ has the law of the independent product $Y_i \cdot Z_i$. Let $g_i(y) := \mathbf{E}[Y_i \mid Y_i \in J_k]$ for the part $J_k \in J$ containing y, and put $\widetilde{Y}_i := g_i(Y_i)$; it takes the value $\mathbf{E}[Y_i \mid Y_i \in J_k]$ with probability $\mathbf{Pr}[Y_i \in J_k]$, so it has the law of the J-sparsification of Y_i (Theorem 2.10), and as a function of Y_i it is independent of Z_i, so $\mathbf{E}[f(\widetilde{Y}_i Z_i)] = F(\widetilde{Y}_i \cdot Z_i)$. The parts $\{0\}, I_0, \ldots, I_m \cap [0, B]$ form a partition of $[0, B]$, which contains the values of Y_i (Setup), so every value of Y_i lies in exactly one part.

Proof of Theorem 4.2. We claim the per-coordinate inflation bound

$$F(\widetilde{Y}_i \cdot Z_i) \leq (1+\delta) F(Y_i \cdot Z_i) \qquad \text{for every } i \in [n]. \tag{A.1}$$

Splitting the expectation $F(\widetilde{Y}_i \cdot Z_i) = \mathbf{E}[f(\widetilde{Y}_i Z_i)]$ by the part of J containing Y_i (and using that $\widetilde{Y}_i$ takes a constant value on each part, under the coupling above):

$$F(\widetilde{Y}_i \cdot Z_i) = \mathbf{E}\Big[f(\widetilde{Y}_i Z_i)\mathbf{1}[Y_i \in \{0\} \cup I_0]\Big] + \sum_{j \geq 1} \mathbf{E}\Big[f(\widetilde{Y}_i Z_i)\mathbf{1}[Y_i \in I_j]\Big].$$

The part $\{0\} \cup I_0$. By Theorem A.2, every positive value of Y_i is at least $\tau^i \geq \tau^n > \tau^n/2 = \gamma$, using $\tau \leq 1$ and $i \leq n$, so Y_i never lies in $I_0 = (0, \gamma]$. On the event $Y_i \in \{0\}$, the value $\widetilde{Y}_i$ is the conditional mean over $\{0\}$, which is 0, so $f(\widetilde{Y}_i Z_i) = f(0) = 0$. This part contributes 0.

A part I_j, $j \geq 1$. On this event, $Y_i \in (\gamma(1+\delta)^{j-1}, \gamma(1+\delta)^j]$, and $\widetilde{Y}_i$, a conditional mean over I_j, lies in the same interval; so $\widetilde{Y}_i \leq \gamma(1+\delta)^j < (1+\delta)Y_i$. By Theorem 2.7(d) and (b), $f(\widetilde{Y}_i Z_i) \leq f((1+\delta)Y_i Z_i) \leq (1+\delta) f(Y_i Z_i)$. Summing over $j \geq 1$ and using $f \geq 0$:

$$\sum_{j \geq 1} \mathbf{E}[f(\widetilde{Y}_i Z_i)\mathbf{1}[Y_i \in I_j]] \leq (1+\delta) \sum_{j \geq 1} \mathbf{E}[f(Y_i Z_i)\mathbf{1}[Y_i \in I_j]] \leq (1+\delta) F(Y_i Z_i).$$

Combining the two parts gives Equation (A.1).

Iteration. $Y_1 Z_1 = R_1 \cdots R_n = P \| Q$, so $F(Y_1 Z_1) = F(P, Q)$. Since $Y_{i+1} Z_{i+1} = \widetilde{Y}_i Z_i$ in law, Equation (A.1) gives $a_{i+1} \le (1+\delta) a_i$ for $a_i := F(Y_i Z_i)$, $1 \le i \le n-1$, so $a_n \le (1+\delta)^{n-1} a_1$ by induction. Applying Equation (A.1) once more at $i = n$, where $Z_n = 1$, and then Theorem A.1(a),

$$F(\widetilde{Y}_n) \le (1+\delta) a_n \le (1+\delta)^n F(P, Q) \le (1 + \varepsilon_1/2) F(P, Q) \le (1 + \varepsilon_1) F(P, Q). \qquad \square$$

A.2 Proof of the Lower Bound

Proof of Theorem 4.3. We claim the reverse inflation bound

$$F(Y_i \cdot Z_i) \le (1+\delta) F(\widetilde{Y}_i \cdot Z_i) \qquad \text{for every } i \in [n]. \tag{A.2}$$

Split $F(Y_i Z_i)$ by the part of J containing Y_i, under the same coupling as before. As in the proof of Theorem 4.2, Y_i never lies in I_0, and on the event $Y_i \in \{0\}$ we have $f(Y_i Z_i) = f(0) = 0$, so this part contributes 0. On the event $Y_i \in I_j$ with $j \ge 1$, both Y_i and $\widetilde{Y}_i$ lie in $(\gamma(1+\delta)^{j-1}, \gamma(1+\delta)^j]$ (the conditional mean of values in this interval lies in it), so $Y_i \le \gamma(1+\delta)^j < (1+\delta)\widetilde{Y}_i$. Since $Z_i \ge 0$, Theorem 2.7(d) gives $f(Y_i Z_i) \le f((1+\delta)\widetilde{Y}_i Z_i)$, and Theorem 2.7(b) with $\lambda = 1+\delta$ gives $f((1+\delta)\widetilde{Y}_i Z_i) \le (1+\delta) f(\widetilde{Y}_i Z_i)$. Summing over $j \ge 1$,

$$\sum_{j \ge 1} \mathbf{E}[f(Y_i Z_i)\, \mathbf{1}[Y_i \in I_j]] \le (1+\delta) \sum_{j \ge 1} \mathbf{E}[f(\widetilde{Y}_i Z_i)\, \mathbf{1}[Y_i \in I_j]] \le (1+\delta) F(\widetilde{Y}_i Z_i),$$

where the last step uses $f \ge 0$ to add back the term of $F(\widetilde{Y}_i Z_i)$ on $\{0\} \cup I_0$; adding the first part gives Equation (A.2).

With $a_i := F(Y_i Z_i)$ as before and $a_{n+1} := F(\widetilde{Y}_n)$, Equation (A.2), $Y_{i+1} Z_{i+1} = \widetilde{Y}_i Z_i$ in law and $Z_n = 1$ give $a_i \le (1+\delta) a_{i+1}$ for $1 \le i \le n$, so by downward induction $F(P, Q) = a_1 \le (1+\delta)^n a_{n+1}$. By Theorem A.1(a),

$$F(\widetilde{Y}_n) \ge \frac{F(P, Q)}{(1+\delta)^n} \ge \frac{F(P, Q)}{1 + \varepsilon_1/2} \ge \frac{F(P, Q)}{1 + \varepsilon_1}. \qquad \square$$

A.3 The Final Evaluation and the Running Time

The final evaluation. Suppose $F(P, Q) > 0$, so that the algorithm reaches its last line. By Theorem A.2, every positive value v of $\widetilde{Y}_n$ satisfies $v \ge \tau^n$, so $f(v) \ge \min(v, 1) f(1) \ge \tau^n \varphi$ by Theorem 2.7(c) and $\tau \le 1$. The oracle is called with $t = \max\{1, \lceil \log_2(8/(\varepsilon \varphi \tau^n)) \rceil\}$, a positive integer as the oracle requires, so its answer satisfies $|\widehat{f}(v) - f(v)| \le 2^{-t} \le (\varepsilon/8) \varphi \tau^n \le (\varepsilon/8) f(v)$; for $v = 0$, $\widehat{f}(0) = 0 = f(0)$. Summing over the values of $\widetilde{Y}_n$ with their probabilities, which are non-negative, $(1 - \varepsilon/8) F(\widetilde{Y}_n) \le \widehat{F} \le (1 + \varepsilon/8) F(\widetilde{Y}_n)$. Combining with Theorems 4.2 and 4.3 at $\varepsilon_1 = \varepsilon/2$:

$$\widehat{F} \le (1 + \varepsilon/8)(1 + \varepsilon/2) F(P, Q) = \left(1 + \tfrac{5\varepsilon}{8} + \tfrac{\varepsilon^2}{16}\right) F(P, Q) \le (1 + \varepsilon) F(P, Q),$$

since $\varepsilon^2/16 \le 3\varepsilon/8$ for $\varepsilon \le 1$; and

$$\widehat{F} \ge \frac{(1 - \varepsilon/8)}{1 + \varepsilon/2} F(P, Q) \ge \frac{F(P, Q)}{1 + \varepsilon},$$

since $(1 - \varepsilon/8)(1 + \varepsilon) = 1 + 7\varepsilon/8 - \varepsilon^2/8 \ge 1 + \varepsilon/2$ for $\varepsilon \le 3$. Together with the case $F(P, Q) = 0$ treated in Section 4.1, this proves the approximation guarantee of Theorem 4.1.

Exact representation. The algorithm stores each atom a of Y_i and of $\widetilde{Y}_i$ as a pair of rationals $(\pi(a), \kappa(a))$, where $\kappa(a) = \mathbf{Pr}[\cdot = a]$ is its probability and $\pi(a) := \kappa(a) \cdot v(a)$ is its probability times its value $v(a)$. For Y_1, the atom from x with $q_1(x) > 0$ has $(\pi, \kappa) = (p_1(x), q_1(x))$. Merging the atoms b of one part of J into a single atom a gives $\kappa(a) = \sum_b \kappa(b)$ and value $v(a) = \sum_b \kappa(b) v(b) / \sum_b \kappa(b)$, so $\pi(a) = \sum_b \pi(b)$. Multiplying an atom a of $\widetilde{Y}_i$ by the outcome x of R_{i+1} with $q_{i+1}(x) > 0$ gives the atom with $\kappa = \kappa(a) q_{i+1}(x)$ and value $v(a) p_{i+1}(x) / q_{i+1}(x)$, so $\pi = \pi(a) p_{i+1}(x)$. By induction, each atom a of Y_i or $\widetilde{Y}_i$ corresponds to a set of tuples

$(x_1, \ldots, x_i)$ with $q_k(x_k) > 0$ for every k, these sets being disjoint for different atoms (merging takes the union of the sets of the merged atoms, and multiplying by x extends each tuple by x), and $\kappa(a)$ and $\pi(a)$ are the sums of $\prod_{k\le i} q_k(x_k)$ and of $\prod_{k\le i} p_k(x_k)$ over the tuples of a. Hence $0 \le \pi(a) \le \sum_{x_1,\ldots,x_i} \prod_{k\le i} p_k(x_k) = 1$, and likewise $0 \le \kappa(a) \le 1$. Let d_k be the least common multiple of the denominators of the 2ℓ entries of p_k and q_k; each entry has bit-length at most b_0, so $d_k \le 2^{2\ell b_0}$. Then every κ and every π is an integer multiple of $1/(d_1 \cdots d_i)$ in $[0, 1]$; the algorithm stores it as an integer numerator over the common denominator $d_1 \cdots d_i$, so that merging is integer addition and multiplying by $p_{i+1}(x)$ or $q_{i+1}(x)$ is one integer multiplication by the precomputed integer $p_{i+1}(x)d_{i+1}$ or $q_{i+1}(x)d_{i+1}$, with the common denominator becoming $d_1 \cdots d_{i+1}$, and every numerator and denominator is at most $d_1 \cdots d_n \le 2^{2n\ell b_0}$, of bit-length $O(n\ell b_0)$. The value of an atom is $v(a) = \pi(a)/\kappa(a)$, and no rounding occurs anywhere in the sparsification stage.

The partition. The thresholds are $\gamma(1+\delta)^j = \tau^n(2^k+1)^j/2^{1+kj}$ for $0 \le j \le m$, rationals of bit-length $O(nb_0 + kj) = O(nb_0 + m\log(2n/\varepsilon))$, using $\tau \ge 2^{-b_0}$ and $k = O(\log(2n/\varepsilon))$ from $\delta = 2^{-k} \ge \varepsilon/(12n)$. They are computed with m multiplications, and m itself is found along the way: since $(1+\delta)^j \ge B/\gamma$ holds if and only if $j \ge \log(B/\gamma)/\log(1+\delta)$, m is the least $j \ge 0$ with $\gamma(1+\delta)^j \ge B$, which costs one comparison per threshold. An atom with $\pi(a) = 0$ goes to $\{0\}$; otherwise, its part is found by binary search over the thresholds, each comparison $v(a) \le \gamma(1+\delta)^j$ being the comparison $\pi(a) \le \gamma(1+\delta)^j\kappa(a)$ of rationals of the bit-lengths above. The integer k is found by testing $k = 1, 2, \ldots$ with the exact comparison $(2^k+1)^n \le 2^{kn}(1+\varepsilon/4)$, which stops at $k = O(\log(2n/\varepsilon))$; each test takes $O(\log n)$ multiplications by repeated squaring, on numbers of bit-length $O(kn+\beta)$. The integer t is read off the bit-lengths of the numerator and denominator of $8/(\varepsilon\varphi\tau^n)$ with one comparison (for a rational a/b, $\lfloor \log_2(a/b) \rfloor$ is one of two consecutive integers determined by the bit-lengths of a and b). Computing τ takes $O(n\ell)$ comparisons, and $B = \tau^{-n}$ and γ take $O(n)$ operations on rationals of bit-length $O(nb_0)$. All of this preprocessing is within the bound below because $m \ge 2n/\varepsilon$ by Theorem A.1(b).

Counting operations. The support test takes $O(n\ell)$ comparisons. $\widetilde{Y}_{i-1}$ has at most $|J| = m+2$ atoms, so Y_i has at most $(m+2)\ell$ atoms, and forming it takes $O(m\ell)$ multiplications. Sparsifying Y_i takes $O(\log m)$ comparisons per atom for the binary search and two additions per atom for the sums $\sum_b \pi(b)$ and $\sum_b \kappa(b)$, so $O(m\ell \log m)$ operations. Over n rounds this is $O(n\ell m \log m)$ operations on rationals of bit-length $O(n\ell b_0 + m\log(2n/\varepsilon))$, and with the preprocessing, whose $O(n\ell + n\log(2n/\varepsilon)\log n + m)$ operations are $O(n\ell m \log m)$ since $m \ge 2n/\varepsilon$, the total is $O(n\ell m \log m)$ operations on rationals of bit-length $O(n\ell b_0 + m\log(2n/\varepsilon) + \beta)$. The final line makes at most m oracle calls (one per part $I_1, \ldots, I_m$ holding an atom; none for $\{0\}$, and none for I_0, which holds no atom), with $t \le 4 + n\log_2(1/\tau) + \log_2(1/\varepsilon) + \max(0, \log_2(1/\varphi)) = O(nb_0 + \log(2/\varepsilon) + \log(1+1/\varphi))$, and $O(m)$ further operations on the oracle's answers to form $\widehat{F}$. With the bound on m from Theorem A.1(b), this is the running time stated in Theorem 4.1, which completes its proof. □

B Proof of Theorem 5.1

Construction. Let ϕ be a CNF formula on n variables $X_1, \ldots, X_n$. We may assume that ϕ has at least one clause and no empty clause: otherwise ϕ is trivially satisfiable or trivially unsatisfiable, and the reduction outputs a fixed yes-instance ($P = Q$ uniform on $\{0,1\}$, with $F = f(1) > 0$) or a fixed no-instance (P and Q point masses at 0 and 1, with $F = f(0) = 0$). View ϕ as a Boolean circuit with AND, OR, NOT gates of fan-in at most 2 (negative literals $\neg X_i$ become NOT gates with parent X_i, and multi-literal clauses and the outer conjunction are converted to balanced binary trees of fan-in-2 gates, with $\mathrm{poly}(|\phi|)$ blow-up), with internal gates $Y_1, \ldots, Y_m$ ordered topologically with Y_m the output (m polynomial in $|\phi|$; if ϕ is the single positive literal X_i, we add the gate $X_i \wedge X_i$, so that the output is always a gate and $m \ge 1$). Let G_ϕ be the underlying DAG on $n+m$ nodes, with edges going from each input or gate to the gates using it; its in-degree is at most 2.

Define two Bayes nets P, Q on G_ϕ over $\{0,1\}^{n+m}$ via the following CPTs:

- Each input X_i is a uniform random bit (under both P and Q): $\mathbf{Pr}[X_i = 0] = \mathbf{Pr}[X_i = 1] = 1/2$.

- For each internal gate Y_i with $1 \le i \le m-1$, the CPT is the deterministic function of its parents given by the corresponding gate (under both P and Q): e.g., for an AND gate with parents A, B, $\mathbf{Pr}[Y_i = a \wedge b | A = a, B = b] = 1$.
- For the output gate Y_m: under P, the CPT is the deterministic function of the gate's parents (the actual circuit output, $\phi(x)$); under Q, $Y_m \equiv 1$.

P and Q differ only at Y_m. Both are valid Bayes nets on G_ϕ with in-degree at most 2 and rational probabilities in $\{0, 1/2, 1\}$, so the reduction runs in polynomial time.

Support analysis. For an input $x \in \{0,1\}^n$, every gate $Y_1, \ldots, Y_m$ has a deterministic CPT under both P and Q, so its value is determined by x; let $g_P(x), g_Q(x) \in \{0,1\}^m$ denote the resulting gate-value strings under P, Q respectively. By construction:

- For $i = 1, \ldots, m-1$: $g_P(x)_i = g_Q(x)_i$ (both equal the gate-output, computed deterministically from the same parents with the same CPT).
- For $i = m$: $g_P(x)_m \in \{0,1\}$ equals $\phi(x)$, the formula's output on x, while $g_Q(x)_m = 1$ always.

Hence $g_P(x) = g_Q(x)$ if and only if $\phi(x) = 1$, i.e., if and only if $x \in \mathrm{Sol}(\phi)$.

Computing $F(P, Q)$. For an atom $z = (x, y) \in \{0,1\}^{n+m}$, the contribution $Q(z) f(P(z)/Q(z))$ to $F(P,Q) = \sum_z Q(z) f(P(z)/Q(z))$ vanishes unless $Q(z) > 0$ (i.e., $y = g_Q(x)$, the unique gate-string with positive Q-probability for input x). We split into two cases:

Case 1: $x \notin \mathrm{Sol}(\phi)$. Then $g_P(x) \neq g_Q(x)$, so $P(x, g_Q(x)) = 0$ (under P the only gate-string of positive probability for input x is $g_P(x)$, since every gate CPT is deterministic); the contribution is $Q(z) f(P(z)/Q(z)) = Q(z) f(0) = 0$.

Case 2: $x \in \mathrm{Sol}(\phi)$. Then $g_P(x) = g_Q(x)$, so $z = (x, g_Q(x))$ has positive mass under both. The n inputs are uniform (each contributing a factor $\frac{1}{2}$) and each of the m deterministic gates contributes a factor 1, so $P(z) = 2^{-n}$ and likewise $Q(z) = 2^{-n}$; the per-atom ratio is $P(z)/Q(z) = 1$ and the contribution is $Q(z) f(1) = f(1)/2^n$. Summing over the 2^n inputs (those in $\mathrm{Sol}(\phi)$ contribute $f(1)/2^n$ each, the rest 0) gives Equation (5.1).

Combining the cases. By Theorem 2.7(c), $f(1) > 0$, since f is admissible. The identity uses only $f(0) = 0$, and this step uses only $f(1) > 0$. Hence $F(P,Q) > 0$ if and only if $|\mathrm{Sol}(\phi)| > 0$, i.e., if and only if ϕ is satisfiable.

NP-completeness. *Membership:* $f > 0$ on $(0, \infty)$ by Theorem 2.7(c), so $F(P,Q) > 0$ if and only if there exists z with $P(z), Q(z) > 0$ (forward, because a term with $P(z) = 0$ is $Q(z) f(0) = 0$; backward, because all terms of F are non-negative and this one is positive). Such a z is a witness of polynomial size, and $P(z)$ and $Q(z)$ are computed in polynomial time as products of CPT entries along a topological order. *Hardness:* the reduction above runs in polynomial time and ϕ is satisfiable if and only if $F(P,Q) > 0$; since CNF-SAT is NP-complete, deciding $F(P,Q) > 0$ is NP-hard. Finally, an algorithm that outputs a value within any multiplicative factor of $F(P,Q)$ outputs 0 exactly when $F(P,Q) = 0$, so it decides $F(P,Q) > 0$; this gives the inapproximability statement. □

B.1 In-Degree One

Theorem 5.1 places P and Q on a common DAG of in-degree 2. This subsection settles the zero test at in-degree 1: it is hard when P and Q may use different DAGs, and easy on a common DAG or over a binary alphabet. Throughout, f is admissible, so that, as in the membership part of the proof above, $F(P,Q) > 0$ if and only if some z has $P(z) > 0$ and $Q(z) > 0$. For a Bayes net P, $P(z) > 0$ if and only if every CPT entry $P(z_v \mid z_{\Pi(v)})$ along z is positive, a constraint on z_v alone when v is a root and on the pair $(z_v, z_{\Pi(v)})$ when v has one parent.

Proposition B.1 (Hardness on different DAGs). *Deciding whether $F(P,Q) > 0$, for Bayes nets P and Q over $[3]^N$ each of in-degree at most 1, on possibly different DAGs, is NP-complete. Hence,*

unless $\mathsf{P} = \mathsf{NP}$*, no polynomial-time algorithm approximates* $F(P, Q)$ *within any multiplicative factor on such pairs.*

Proof. Membership is as in Theorem 5.1. For hardness we reduce from 3-colourability of graphs without loops, which is NP-complete (Garey & Johnson, 1979, Problem GT4). Given $G = (V, E)$, orient each edge e arbitrarily as (u, v) and create two nodes, $X_{e,1}$ for the endpoint u and $X_{e,2}$ for the endpoint v; for each isolated vertex v create one node X_v. For each vertex v fix an order $X_1^{(v)}, \ldots, X_{d_v}^{(v)}$ ($d_v \geq 1$) on the nodes created for it. Every node takes values in $[3]$, and $N = 2|E|$ plus the number of isolated vertices. In P, each $X_1^{(v)}$ is a root with the uniform distribution on $[3]$, and for $j \geq 2$ the parent of $X_j^{(v)}$ is $X_{j-1}^{(v)}$, with the CPT that copies the parent's value; its DAG is a disjoint union of paths, one per vertex. In Q, for each edge e, $X_{e,1}$ is a root with the uniform distribution on $[3]$ and $X_{e,2}$ has parent $X_{e,1}$, with the CPT uniform on the two values different from the parent's; each X_v is a root with the uniform distribution; its DAG is a matching together with isolated nodes. Both nets have in-degree at most 1 and CPT entries in $\{0, 1/3, 1/2, 1\}$, and they are built in time linear in $|V| + |E|$. Now $P(z) > 0$ exactly when z is constant on the nodes of each vertex, and $Q(z) > 0$ exactly when $z_{X_{e,1}} \neq z_{X_{e,2}}$ for every edge e. If z has both properties, colouring each vertex by the common value of its nodes is a proper 3-colouring, since the two endpoints of every edge e receive the values $z_{X_{e,1}} \neq z_{X_{e,2}}$. Conversely, a proper 3-colouring c gives the z that assigns $c(v)$ to every node of v, which has both properties. So $F(P, Q) > 0$ if and only if G is 3-colourable. The inapproximability follows as in Theorem 5.1: an algorithm within any multiplicative factor outputs 0 exactly when $F(P, Q) = 0$. □

Proposition B.2 (Easy cases at in-degree one)**.** *Deciding whether* $F(P, Q) > 0$ *is in polynomial time* (a) *for Bayes nets* P *and* Q *over* $[\ell]^N$ *on a common DAG of in-degree at most* 1*, and* (b) *for Bayes nets* P *and* Q *over* $\{0, 1\}^N$ *each of in-degree at most* 1*, on possibly different DAGs.*

Proof. In both cases we decide whether the constraints of P and of Q described above can be satisfied by one z. *(a)* On a common DAG, a forest, every constraint of P or Q is on a single node or on a node and its parent. For each node v, taken in an order in which every node comes after its children, let $A_v \subseteq [\ell]$ be the set of values a that the constraints on v alone allow and such that every child w of v has some $b \in A_w$ with (b, a) allowed by the constraints on the pair (z_w, z_v) in both nets. By induction from the leaves, A_v is exactly the set of values of z_v that extend to an assignment of the subtree rooted at v satisfying every constraint inside it: the constraints of that subtree are those on v, those on the pairs (w, v) for the children w, and those inside the subtrees of the children, which share no variable. Hence the constraints are satisfiable if and only if $A_r \neq \emptyset$ for every root r, and computing all A_v takes $O(N\ell^2)$ lookups of CPT entries. *(b)* Over $\{0, 1\}$ every constraint is a relation on at most two Boolean variables, and every such relation is the set of satisfying assignments of the conjunction of the clauses, each on the same at most two variables, that exclude its forbidden assignments one at a time. So the assignments z with $P(z), Q(z) > 0$ are the satisfying assignments of a 2-CNF formula with at most 4 clauses per CPT, hence at most $8N$ clauses, whose satisfiability is decided in linear time (Aspvall et al., 1979). □

So on a common DAG of in-degree 1 any hardness of multiplicative approximation must come from an approximation gap rather than from a zero test, and the ternary alphabet of Theorem B.1 cannot be reduced to a binary one unless $\mathsf{P} = \mathsf{NP}$.

C A Worked Example: Bayes Error of a Naive-Bayes Model

This appendix proves Theorem C.1, which quantifies how far the bounds on the Bayes-optimal error that come from divergences computable on products can be from the error itself.

The model. Fix $p \in (1/2, 1)$ and an odd integer $n = 2k + 1$. A naive-Bayes model has a class label $C \in \{0, 1\}$ with prior $\mathbf{Pr}[C = 0] = \mathbf{Pr}[C = 1] = 1/2$ and features $X_1, \ldots, X_n \in \{0, 1\}$ that are independent given C; its Bayes net is the star with center C. We take the class-conditionals $P_0 := \mathrm{Bern}(p)^{\otimes n}$ and $P_1 := \mathrm{Bern}(1-p)^{\otimes n}$, so each feature points to class 0 with probability p. The Bayes-optimal error is $R^* = S_{\mathrm{stat}}(\frac{1}{2}P_0, \frac{1}{2}P_1) = \frac{1}{2}\sum_x \min(P_0(x), P_1(x))$, and the Bhattacharyya coefficient is $\mathrm{BC} := \sum_x \sqrt{P_0(x)P_1(x)}$.

Proposition C.1 (Divergence brackets on the Bayes error are exponentially wide)**.** *In the model above, with* $\rho := \sqrt{(1-p)/p} \in (0,1)$*:*

(a) $R^* = \mathbf{Pr}[\mathrm{Bin}(n,p) \le k]$ *and* $\mathrm{BC} = (2\sqrt{p(1-p)})^n$*;*

(b) $\dfrac{\rho}{\sqrt{2(n+1)}}\,\mathrm{BC} \le R^* \le \dfrac{\sqrt{p(1-p)}}{(2p-1)\sqrt{n+2}}\,\mathrm{BC}$*;*

(c) the upper bound $R^* \le \frac{1}{2}\mathrm{BC}$ *holds, and exceeds* R^* *by a factor at least* $(2p-1)\sqrt{n+2}/(2\sqrt{p(1-p)})$*;*

(d) the lower bound $R^* \ge \frac{1}{2}\left(1-\sqrt{1-\mathrm{BC}^2}\right)$ *holds, and falls below* R^* *by a factor at least* $\rho/(\mathrm{BC}\sqrt{2(n+1)})$*;*

(e) $D_{\mathrm{KL}}(P_0\|P_1) = n(2p-1)\log\frac{p}{1-p}$*, so the Pinsker lower bound* $R^* \ge \frac{1}{2}\left(1-\sqrt{D_{\mathrm{KL}}(P_0\|P_1)/2}\right)$ *is negative once* $n(2p-1)\log\frac{p}{1-p} > 2$*.*

Since $\mathrm{BC} = (4p(1-p))^{n/2}$ and $4p(1-p) < 1$ for $p \neq 1/2$, the factor in (d) grows exponentially in n, and the two ends of the bracket in (c) and (d) differ by a factor that also grows exponentially. The coefficient BC, like D_{KL}, factorizes over the features and is therefore available in closed form, to any precision in polynomial time, on any pair of products, but the bracket (c)–(d) it yields, which holds for every pair of distributions with the same coefficient, locates R^* only up to a factor exponential in n, while Theorem 4.1 returns R^* within a factor $1+\varepsilon$. In this symmetric model R^* is also available directly as the binomial tail in (a), and the example is there to exhibit the gap rather than to need the algorithm: once the features have different marginals, or the priors differ, the error is $\alpha_0 F_{\min(r,c)}(P_0,P_1)$, with priors α_0,α_1 and $c=\alpha_1/\alpha_0$ as in Section 4.2, with no closed form, and Theorem 4.1 still applies. The upper end of the bracket is loose only by a factor $\Theta(\sqrt{n})$: by (b), $\frac{1}{2}\mathrm{BC}/R^* \le \sqrt{n+1}/(\sqrt{2}\rho)$, and by (c) it is at least $(2p-1)\sqrt{n+2}/(2\sqrt{p(1-p)})$; the exponential looseness is on the lower side, (d).

Proof. (a) Write w for the number of ones in $x \in \{0,1\}^n$. Then $P_0(x) = p^w(1-p)^{n-w}$ and $P_1(x) = (1-p)^w p^{n-w}$, so $P_0(x)/P_1(x) = (p/(1-p))^{2w-n}$, which is at least 1 exactly when $2w \ge n$, that is, $w \ge k+1$. Hence $\min(P_0(x),P_1(x))$ equals $P_0(x)$ when $w \le k$ and $P_1(x)$ when $w \ge k+1$, and

$$\sum_x \min(P_0(x),P_1(x)) = \sum_{w=0}^{k}\binom{n}{w}p^w(1-p)^{n-w} + \sum_{w=k+1}^{n}\binom{n}{w}(1-p)^w p^{n-w}.$$

Substituting $w' := n-w$ in the second sum turns it into $\sum_{w'=0}^{k}\binom{n}{w'}p^{w'}(1-p)^{n-w'}$, since $\binom{n}{w} = \binom{n}{n-w}$ and $w \ge k+1$ if and only if $w' \le n-k-1 = k$. So both sums equal $\mathbf{Pr}[\mathrm{Bin}(n,p) \le k]$, and $R^* = \frac{1}{2}\cdot 2\,\mathbf{Pr}[\mathrm{Bin}(n,p) \le k]$. For BC, the product structure gives $\sum_x \sqrt{P_0(x)P_1(x)} = \prod_{i=1}^{n}\left(\sqrt{p(1-p)}+\sqrt{(1-p)p}\right) = (2\sqrt{p(1-p)})^n$.

Two binomial estimates. For $j \ge 1$ let $a_j := \binom{2j}{j}4^{-j} = \prod_{i=1}^{j}\frac{2i-1}{2i}$. For $i \ge 2$, $(2i-1)^2 = 4i^2-4i+1 > 2i(2i-2)$, so $\frac{2i-1}{2i} > \frac{2i-2}{2i-1}$; and for $i \ge 1$, $(2i-1)(2i+1) = 4i^2-1 < (2i)^2$, so $\frac{2i-1}{2i} < \frac{2i}{2i+1}$. Using the first inequality on one of the two factors of a_j^2 for each $i \ge 2$,

$$a_j^2 = \frac{1}{2}\prod_{i=2}^{j}\frac{2i-1}{2i}\cdot\prod_{i=1}^{j}\frac{2i-1}{2i} \ge \frac{1}{2}\prod_{i=2}^{j}\frac{2i-2}{2i-1}\cdot\prod_{i=1}^{j}\frac{2i-1}{2i} = \frac{1}{2}\cdot\frac{1}{2j},$$

since the last product telescopes: its numerators are $2,4,\ldots,2j-2$ and $1,3,\ldots,2j-1$, and its denominators are $3,5,\ldots,2j-1$ and $2,4,\ldots,2j$, leaving $1/(2j)$. Using the second inequality on one factor for each $i \ge 1$, $a_j^2 \le \prod_{i=1}^{j}\frac{2i-1}{2i}\cdot\frac{2i}{2i+1} = \frac{1}{2j+1}$. Hence

$$\frac{4^j}{2\sqrt{j}} \le \binom{2j}{j} \le \frac{4^j}{\sqrt{2j+1}} \qquad (j \ge 1). \tag{C.1}$$

(b) Let $b_w := \binom{n}{w} p^w (1-p)^{n-w}$, so $R^* = \sum_{w=0}^{k} b_w$. The top term is $b_k = \binom{2k+1}{k} p^k (1-p)^{k+1}$, and $\binom{2k+1}{k} = \frac{1}{2}\binom{2k+2}{k+1}$ because $\binom{2k+2}{k+1} = \binom{2k+1}{k} + \binom{2k+1}{k+1}$ and the two terms are equal. Also $p^k(1-p)^{k+1} = (p(1-p))^{n/2}\sqrt{(1-p)/p} = \rho\,(p(1-p))^{n/2}$, since $k = (n-1)/2$ and $k+1 = (n+1)/2$. By Equation (C.1) with $j = k+1$ and $2(k+1) = n+1$,

$$\frac{2^n}{\sqrt{2(n+1)}} = \frac{1}{2} \cdot \frac{4^{k+1}}{2\sqrt{k+1}} \le \binom{2k+1}{k} \le \frac{1}{2} \cdot \frac{4^{k+1}}{\sqrt{2k+3}} = \frac{2^n}{\sqrt{n+2}},$$

so, with $2^n (p(1-p))^{n/2} = \mathrm{BC}$,

$$\frac{\rho\,\mathrm{BC}}{\sqrt{2(n+1)}} \le b_k \le \frac{\rho\,\mathrm{BC}}{\sqrt{n+2}}.$$

The lower bound of (b) follows from $R^* \ge b_k$. For the upper bound, for $1 \le w \le k$,

$$\frac{b_{w-1}}{b_w} = \frac{w}{n-w+1} \cdot \frac{1-p}{p} \le \frac{1-p}{p} = \rho^2,$$

because $w \le k < k+2 \le n-w+1$. So $b_{k-j} \le \rho^{2j} b_k$ for $0 \le j \le k$, and $R^* \le b_k \sum_{j \ge 0} \rho^{2j} = b_k/(1-\rho^2)$. Since $1-\rho^2 = (2p-1)/p$ and $\rho\,p/(2p-1) = \sqrt{p(1-p)}/(2p-1)$, this gives the upper bound of (b).

(c) For non-negative a, b, $\min(a,b) \le \sqrt{ab}$, so $R^* = \frac{1}{2}\sum_x \min(P_0(x), P_1(x)) \le \frac{1}{2}\mathrm{BC}$. By the upper bound of (b), $\frac{1}{2}\mathrm{BC}/R^* \ge \frac{1}{2} \cdot (2p-1)\sqrt{n+2}/\sqrt{p(1-p)}$.

(d) The inequality $\sum_x \min(P_0(x), P_1(x)) \ge 1 - \sqrt{1-\mathrm{BC}^2}$ is Le Cam's (Tsybakov, 2009, Lemma 2.3): with the squared Hellinger distance $H^2 := \sum_x (\sqrt{P_0(x)} - \sqrt{P_1(x)})^2 = 2(1-\mathrm{BC})$, the bound $d_{\mathrm{TV}}(P_0, P_1) \le H\sqrt{1 - H^2/4}$ reads $d_{\mathrm{TV}}(P_0, P_1) \le \sqrt{1-\mathrm{BC}^2}$, since $H^2(1 - H^2/4) = 2(1-\mathrm{BC})(1+\mathrm{BC})/2 = 1 - \mathrm{BC}^2$, and $\sum_x \min(P_0(x), P_1(x)) = 1 - d_{\mathrm{TV}}(P_0, P_1)$. Dividing by 2 gives the stated bound on R^*. For $y \in [0,1]$, $\sqrt{1-y} \ge 1-y$, so $1 - \sqrt{1-y} \le y$; with $y = \mathrm{BC}^2 \in [0,1]$, the bound is at most $\frac{1}{2}\mathrm{BC}^2$. By the lower bound of (b), R^* exceeds it by a factor at least $\frac{\rho\,\mathrm{BC}}{\sqrt{2(n+1)}} \cdot \frac{2}{\mathrm{BC}^2} \ge \frac{\rho}{\mathrm{BC}\sqrt{2(n+1)}}$.

(e) Both distributions are products, so $D_{\mathrm{KL}}(P_0 \| P_1) = n\, D_{\mathrm{KL}}(\mathrm{Bern}(p) \| \mathrm{Bern}(1-p)) = n\left(p \log \frac{p}{1-p} + (1-p)\log\frac{1-p}{p}\right) = n(2p-1)\log\frac{p}{1-p}$. Pinsker's inequality $d_{\mathrm{TV}}(P_0, P_1) \le \sqrt{D_{\mathrm{KL}}(P_0\|P_1)/2}$ (Tsybakov, 2009, Lemma 2.5) and $R^* = \frac{1}{2}(1 - d_{\mathrm{TV}}(P_0, P_1))$ give the stated lower bound, whose right-hand side is negative exactly when $D_{\mathrm{KL}}(P_0\|P_1) > 2$. □

Remark C.2 (The sharpest analytic bounds). For conditionally independent binary features, Kontorovich & Avital (2025) give the sharpest analytic bounds on R^* we know of. In the model above, which is their symmetric case with sensitivity and specificity p and bias $1/2$ (with the labels of the two classes exchanged), their Theorem 4 gives the lower bound $R^* \ge \frac{1}{2}\,2^n (p(1-p))^{n/2} \exp(-\frac{1}{2}\|w\|_2)$ with $w_i = \log\frac{p}{1-p}$ for every i, that is, $\frac{1}{2}\mathrm{BC}\,(\frac{p}{1-p})^{-\sqrt{n}/2}$. By Theorem C.1(b), R^* exceeds this bound by a factor at least

$$\frac{\rho\,\mathrm{BC}}{\sqrt{2(n+1)}} \cdot \frac{2}{\mathrm{BC}} \left(\frac{p}{1-p}\right)^{\sqrt{n}/2} = \frac{\sqrt{2}}{\sqrt{n+1}} \left(\frac{p}{1-p}\right)^{(\sqrt{n}-1)/2},$$

using $\rho = (p/(1-p))^{-1/2}$; this grows like $\exp(\Theta(\sqrt{n}))$, faster than every polynomial in n, while their upper bound in this case is $\frac{1}{2}\mathrm{BC}$ and overshoots by the factor of Theorem C.1(c). So even the best closed-form bracket leaves a gap that grows without bound, where Theorem 4.1 returns R^* within a factor $1+\varepsilon$, for features over any finite alphabet.

D Motivation for the Functionals

This appendix develops the motivation for each functional in Table 1: the operational meanings of the associated $F(P, Q)$, the application domains where these quantities arise, and the connections

to classical statistical and information-theoretic literature. Section 1.1 provides a high-level survey; this appendix gives an extended discussion. The verification that each f is admissible (Theorem 2.6) is in Section F; this appendix focuses entirely on *why* these functionals matter. Each subsection follows the same structure: a *Functional* paragraph stating the explicit formula, followed by a *Why it matters* paragraph explaining the operational meaning, applications, and connections to classical literature.

D.1 Statistical Similarity ($f(r) = \min(r, 1)$)

Functional. $F(P, Q) = S_{\mathrm{stat}}(P, Q) = \sum_{x \in D} \min(P(x), Q(x))$, the statistical similarity.

Why it matters. S_{stat} is the canonical concave overlap between probability distributions, with multiple equivalent operational interpretations:

- *Maximum coupling probability*: $S_{\mathrm{stat}}(P, Q) = \max_{(X,Y): X \sim P, Y \sim Q} \mathbf{Pr}[X = Y]$, where the max ranges over all couplings of (X, Y) (Strassen, 1965; Lindvall, 2002). The optimal coupling places mass $\min(P(x), Q(x))$ on the diagonal at each x.
- *Indistinguishability*: $S_{\mathrm{stat}}(P, Q) = 1 - d_{\mathrm{TV}}(P, Q)$, where d_{TV} is total variation distance. In one-sample hypothesis testing of P vs. Q given a single observation, the maximum power of any test minus its size is exactly d_{TV} (Tsybakov, 2009; Cam, 1986), so S_{stat} measures *indistinguishability*.
- *Minimum sum of testing errors*: For testing $H_0 : P$ vs. $H_1 : Q$ on a single observation with deterministic decision rules, the minimum sum of Type-I and Type-II errors over all tests is exactly $S_{\mathrm{stat}}(P, Q)$ (Kontorovich & Avital, 2025; Lehmann & Romano, 2005).
- *Bayes-optimal classification error*: For class-conditionals P_0, P_1 with priors α_0, α_1, the Bayes-optimal classification error is $R^* = S_{\mathrm{stat}}(\alpha_0 P_0, \alpha_1 P_1)$ (Fukunaga, 1990, Eq. (3.7)), providing the operational meaning of S_{stat} as the unavoidable error rate of classification under the input distribution model.

The integrand $\min(r, 1)$ has a clean structural interpretation: it grows linearly with the per-atom mass on P when $P(x) \leq Q(x)$ (the "P-light" atoms) and saturates at $Q(x)$ when $P(x) > Q(x)$ (the "P-heavy" atoms), formalizing S_{stat} as the total "shared mass."

D.2 Capped Statistical Similarity ($f(r) = \min(r, c)$, $c \in (0, 1]$)

Functional. $S_{\mathrm{stat}\,c}(P, Q) = \sum_{x \in D} \min(P(x), c \cdot Q(x))$, capped statistical similarity.

Why it matters. Capped similarity has two operational interpretations:

- *Asymmetric Bayes risk*: $S_{\mathrm{stat}\,c}$ governs Bayes risk under *asymmetric* 0-1 loss with cost ratio c. When class-0 misclassifications cost 1 and class-1 misclassifications cost c, the minimum achievable risk on class-conditionals P_0, P_1 with priors α_0, α_1 is $S_{\mathrm{stat}\,c}(\alpha_0 P_0, \alpha_1 P_1)$. The cost ratio c encodes the relative seriousness of the two error types, a standard model in cost-sensitive classification (Fukunaga, 1990, Ch. 3).
- *Bounded-influence aggregation*: Capped overlaps arise in robust estimation, where bounded-influence aggregations limit the contribution of any single observation to the test statistic (Huber, 1964; Hampel et al., 1986; Lehmann & Romano, 2005). Each atom contributes at most $c \cdot Q(x)$, which prevents a single high-likelihood-ratio atom from dominating the aggregate.

D.3 Truncated Similarity ($f(r) = \min(r, c)$, $c > 1$)

Functional. $F(P, Q) = \sum_x \min(P(x), c \cdot Q(x))$, generalizing S_{stat} by allowing per-atom contributions to exceed $Q(x)$ up to a factor c. As $c \to \infty$, $F(P, Q) \to \sum_{x: Q(x) > 0} P(x) \leq 1$ (with equality precisely when $P \ll Q$); for c near 1, F is close to S_{stat}.

Why it matters. Truncated similarity is a natural variant of S_{stat} in which the structural role is to enlarge the integrand's range above the S_{stat}-cap of 1 while retaining concavity, monotonicity, and the linear-near-zero property. The parameter $c > 1$ models settings where P-mass exceeding Q-mass by a moderate factor is still informative (rather than saturated as in S_{stat}); $c = 2$, for example, weights P-heavy atoms up to twice the Q-mass before saturating. Such truncations of the likelihood ratio appear in robust hypothesis testing, where bounding the per-atom influence is a standard technique (Huber, 1964; Hampel et al., 1986). We include this admissible functional for completeness of the family.

D.4 Log-Overlap ($f(r) = \log(1 + r)$)

Functional. $F(P, Q) = \sum_{x:Q(x)>0} Q(x) \log(1 + P(x)/Q(x))$.

Why it matters. The Jensen–Shannon divergence is $D_{\mathrm{JS}}(P\|Q) = \frac{1}{2}D_{\mathrm{KL}}(P\|M) + \frac{1}{2}D_{\mathrm{KL}}(Q\|M)$ where $M = (P + Q)/2$. A direct computation gives

$$D_{\mathrm{KL}}(Q\|M) = \sum_x Q(x) \log \frac{Q(x)}{(P(x) + Q(x))/2} = \sum_x Q(x) \log \frac{2Q(x)}{P(x) + Q(x)} = \log 2 - F(P, Q).$$

Thus $F(P, Q) = \log 2 - D_{\mathrm{KL}}(Q\|M)$, and symmetrically $D_{\mathrm{KL}}(P\|M) = \log 2 - F(Q, P)$, so that $D_{\mathrm{JS}}(P\|Q) = \log 2 - \frac{1}{2}\big(F(P, Q) + F(Q, P)\big)$. Since $F \le \log 2$ by Jensen's inequality, two runs of the FPTAS of Theorem 4.1, one on (P, Q) and one on (Q, P), give D_{JS} on products to additive error $\varepsilon \log 2$; D_{KL} itself is a convex Csiszár divergence (Csiszár, 1967), outside our (concave) family. The JS divergence (Lin, 1991) is a bounded, symmetrized version of KL. Unlike $\min(r, 1)$, the integrand $\log(1 + r)$ is smooth at $r = 1$, and it grows sub-linearly in large likelihood ratios, a soft analogue of the cap in capped similarity.

D.5 Exponential Overlap ($f(r) = 1 - e^{-r}$)

Functional. $F(P, Q) = 1 - \mathbf{E}_{x\sim Q}\big[e^{-P(x)/Q(x)}\big] = 1 - \sum_{x:Q(x)>0} Q(x) e^{-P(x)/Q(x)}$, a bounded smoothed measure of distributional overlap, valued in $[0, 1)$.

Why it matters. The exponential damping e^{-r} provides a sharper saturation than the polynomial damping in harmonic overlap $r/(1 + r)$: for large r, $1 - e^{-r}$ approaches 1 with exponentially small gap, while $r/(1 + r)$ approaches 1 with $1/r$ gap.

D.6 Harmonic / Triangular Overlap ($f(r) = r/(1 + r)$)

Functional. $F(P, Q) = \sum_{x:Q(x)>0} Q(x) \cdot \frac{P(x)/Q(x)}{1+P(x)/Q(x)} = \sum_x \frac{P(x)Q(x)}{P(x)+Q(x)}$ (treating $0/0 = 0$). Geometrically, each atom contributes the harmonic mean of $P(x)$ and $Q(x)$ divided by 2.

Why it matters. The triangular discrimination is $\Delta(P, Q) := \sum_x (P(x) - Q(x))^2/(P(x) + Q(x))$ (Liese & Vajda, 1987) (with the convention $0/0 = 0$), a known f-divergence. A direct computation gives $\Delta(P, Q) = 2 - 4F(P, Q)$ for any probability distributions P, Q (since $(P-Q)^2 + 4PQ = (P+Q)^2$, so $\sum_x[(P-Q)^2+4PQ]/(P+Q) = \sum_x(P+Q) = 2$). Multiplicative estimation of F thus yields additive estimation of Δ. The standard relations $2d_{\mathrm{TV}}(P, Q)^2 \le \Delta(P, Q) \le 2d_{\mathrm{TV}}(P, Q)$ (Topsøe, 2000; Sason & Verdú, 2016) (the upper bound from $(P - Q)^2/(P + Q) \le |P - Q|$; the lower from Cauchy–Schwarz) make Δ a useful (squared-sandwich) proxy for d_{TV}. Beyond this connection, the harmonic overlap is the prototypical smooth bounded overlap with $f(r) = r/(1 + r) \to 1$ as $r \to \infty$ and linear near zero, in the same structural family as exponential and $\tanh$ overlaps. The polynomial saturation rate (gap $\sim 1/r$ to the upper bound 1) is gentler than the exponential saturation of $1 - e^{-r}$, making harmonic overlap more sensitive to large likelihood ratios while still bounded.

D.7 Hyperbolic-Tangent Overlap ($f(r) = \tanh r$)

Functional. $F(P, Q) = \sum_x Q(x) \tanh(P(x)/Q(x))$, saturating smoothly at 1 as the per-atom likelihood ratio grows.

Why it matters. The saturation of $\tanh r$ at 1 as $r \to \infty$ bounds the contribution of each atom by $Q(x)$, as the cap of S_{stat} does, but with a smooth rather than piecewise-linear cap. In this it resembles Huber's clipped score $\min(\max(r, -k), k)$ (Huber, 1964), of which it is a smooth analogue, and bounded influence of a single observation is the organizing principle of robust statistics (Hampel et al., 1986). The transformation $r \mapsto \tanh r$ maps $[0, \infty)$ onto $[0, 1)$, which is convenient for normalized comparisons.

D.8 Inverse-Tangent Overlap ($f(r) = \arctan r$)

Functional. $F(P, Q) = \sum_x Q(x) \arctan(P(x)/Q(x))$, saturating smoothly at $\pi/2$ as the per-atom likelihood ratio grows. Rescaling by $2/\pi$ gives a $[0, 1)$-valued analogue of S_{stat}.

Why it matters. The two functions $\tanh$ and $\arctan$ share the qualitative saturation profile but differ quantitatively: $\arctan$ saturates more slowly ($\arctan r \approx \pi/2 - 1/r$ for large r) than $\tanh$ ($\tanh r \approx 1 - 2e^{-2r}$), placing greater weight on intermediate likelihood ratios. This makes $\arctan$ overlap more discriminating for moderate r, while $\tanh$ overlap is more aggressively bounded.

D.9 Bounded Log-Overlap ($f(r) = \min(\log(1 + r), \beta)$, $\beta > 0$)

Functional. $F(P, Q) = \sum_x Q(x) \min(\log(1 + P(x)/Q(x)), \beta)$, capping the integrand at β (so each atom contributes at most $\beta \cdot Q(x)$) to limit the influence of high-likelihood-ratio atoms.

Why it matters. Bounded log-overlap caps a smoothed log-likelihood-ratio statistic at β, controlling the influence of high-likelihood-ratio atoms on the aggregate measure. The use of $\log(1 + r)$ rather than $\log r$ is a smoothing convention that keeps $f(0) = 0$ (a hard requirement of our framework); the unsmoothed truncation $\min(\log r, \beta)$ of the log-likelihood ratio tends to $-\infty$ as $r \to 0$, so it does not satisfy $f(0) = 0$ and falls outside our framework. The cap parameter β limits each atom's contribution to at most $\beta \cdot Q(x)$; increasing β recovers the uncapped log-overlap pointwise.

D.10 Capped Exponential Overlap ($f(r) = \min(1 - e^{-r}, c)$, $c \in (0, 1)$)

Functional. $F(P, Q) = \sum_x Q(x) \min(1 - e^{-P(x)/Q(x)}, c)$, a bounded version of the exponential overlap of Section D.5 obtained by capping the integrand at c (so each atom contributes at most $c \cdot Q(x)$).

Why it matters. The cap $c \in (0, 1)$ provides a hard upper bound on per-atom contributions, on top of the exponential damping of $1 - e^{-r}$. The integrand equals $1 - e^{-r}$ for $r \leq \log(1/(1 - c))$ and the constant c above that threshold: the exponential shape governs moderate likelihood ratios, and the cap sets a strict ceiling on large ones.

D.11 Convex Combinations $f = \theta g_1 + (1 - \theta) g_2$, $\theta \in [0, 1]$

Functional. By linearity of expectation, $F_f(P, Q) = \theta F_{g_1}(P, Q) + (1 - \theta) F_{g_2}(P, Q)$.

Why it matters. Convex combinations are operationally meaningful as hybrid statistics that interpolate between qualitatively distinct overlap regimes. A weighted blend of S_{stat} and exponential overlap, for example, combines the piecewise-linear robustness profile of S_{stat} (each atom contributes at most $Q(x)$, but the contribution is non-differentiable at $r = 1$) with the smooth saturation of exp-overlap (each atom's contribution approaches $Q(x)$ exponentially, with no corner), yielding a statistic that keeps part of S_{stat}'s interpretation as "shared mass" while softening, though not removing, its corner at the diagonal: the one-sided derivatives at $r = 1$ become $\theta + (1 - \theta)/e$ and $(1 - \theta)/e$. More generally, blending any two members of Table 1 produces a hybrid whose qualitative behavior is controlled by the mixing weight θ. The technical closure properties of admissible functions under convex combinations and pointwise minimum, and their consequence that the framework's algorithmic guarantees lift uniformly to the entire convex hull, are deferred to Section F.13.

E Other Related Work

This appendix expands on Section 1.3 with discussion of related literature beyond the most directly comparable prior work.

Bayes-net inference complexity. The complexity of probabilistic inference on Bayes nets is a foundational topic: Cooper (1990) proved that exact inference is NP-hard, Roth (1996) extended this to #P-hardness for counting-style queries, and approximate inference is NP-hard to multiplicatively approximate within any factor (Dagum & Luby, 1993). These classical results concern marginal probabilities and conditional probabilities of single events. In the comparison-of-distributions setting, Bhattacharyya et al. (2024) reduced relative approximation of d_{TV} between same-structure Bayes nets to probabilistic inference, yielding an FPRAS when inference is efficient, including on bounded-treewidth graphs. Our hardness result for F applies to common graphs of in-degree 2 without an efficient-inference assumption. The reduction in our Theorem 5.1 (and that of Bhattacharyya et al. (2023) for the opposite end of d_{TV}) uses a SAT encoding rather than inference-complexity machinery.

Approximating d_{TV} and the dual side. The complexity of approximating total variation distance between high-dimensional structured distributions has been studied extensively. Bhattacharyya et al. (2023) initiated the systematic study, proving that exact computation of d_{TV} on products is #P-complete, giving an FPTAS for d_{TV} on products when one distribution is uniform, and proving that deciding whether $d_{\mathrm{TV}}(P, Q) \neq 0$ is NP-complete on Bayes nets of in-degree at least 2, so that multiplicative approximation of d_{TV} there is NP-hard. Feng et al. (2023) gave a randomized FPRAS for d_{TV} on products via a coupling-and-rejection-sampling argument. Feng et al. (2024) subsequently derandomized this to a deterministic FPTAS via slice-decomposition of the ratio random variable. The two settings are structurally distinct: their merging error is charged to the integrand of d_{TV} itself, which vanishes at $r = 1$, while for the concave family the delicate region would be $r \to 0$, where f vanishes; positive partial-product ratios never fall below τ^n, so in our analysis a near-zero slice $(0, \tau^n/2]$ stays empty and no additive error arises, and the bound $f(v) \geq \varphi\tau^n$ on each positive atom fixes the precision of the final evaluation. We use a sparse-partial-product representation of the same general type, adapted to the geometry of the concave problem.

The slice-decomposition technique in detail. In more detail: Feng et al. (2024) maintain a sparse representation of the joint distribution of partial-product likelihood ratios $R_1 \cdot R_2 \cdots R_i$ (which a priori has support of size ℓ^i), by merging values that fall into the same interval and replacing the merged values by their conditional mean. Below 1 their intervals are $[a_t, a_{t+1})$ with $a_t = 1 - (1 + \varepsilon_s)^{-t}$, so that each interval's width is at most $\varepsilon_s(1 - r)$ for the values r it holds; the error of merging is then at most an ε_s fraction of $\mathbf{E}[\max(1 - R, 0)] = d_{\mathrm{TV}}$, so it is charged to the integrand itself, and above 1 the same is done with $1 - 1/r$; only the interval adjacent to $r = 1$ is bounded additively. The concave-family analysis we develop here turns on a different mechanism. Rounding near $r = 0$ would have unbounded relative error, since $f(r) \to 0$ there, but positive partial-product ratios are at least τ^n (Theorem A.2). The chord-slope inequality $f(\lambda r) \leq \lambda f(r)$ for $\lambda \geq 1$ controls rounding up within a geometric slice, and monotonicity together with the same inequality controls rounding down. Our analysis therefore has no additive terms at all, and the error of the final evaluation is made multiplicative by the same bound τ^n on positive values; the additive error of the slice adjacent to $r = 1$ in Feng et al. (2024) is charged instead against d_{TV} being bounded below by $\max_i d_{\mathrm{TV}}(p_i, q_i)$, a single coordinate's distance.

f-divergence estimation in structured-distribution settings. Feng & Fu (2026) studied approximation of convex f-divergences ($f(1) = 0$) between Ising models specified by interaction matrices and external fields. Their results use sampling and approximate counting under model-specific assumptions; our concave functionals are computed from product marginals or Bayes-net conditional probability tables. On the hardness side, Bhattacharyya et al. (2025) proved that relative approximation of d_{TV} between arbitrary Ising models is impossible in randomized polynomial time unless NP $\subseteq$ RP. That work also gave a deterministic polynomial-time equivalence test for mixtures of product distributions.

Bayes-risk and pattern-recognition perspectives. The connection between S_{stat} and Bayes-optimal classification error is textbook material (Fukunaga, 1990, Eq. (3.7)), and k-nearest-neighbour estimators of the Bayes risk from samples go back to Fukunaga & Hostetler (1975) and Devijver (1985): for class-conditionals P_0, P_1 with priors α_0, α_1, the Bayes-optimal error is $R^* = S_{\mathrm{stat}}(\alpha_0 P_0, \alpha_1 P_1)$, providing operational meaning to S_{stat} as a similarity. The same identity makes exact Bayes-risk computation hard already for naive-Bayes models: with equal priors $R^* = \frac{1}{2} S_{\mathrm{stat}}(P_0, P_1)$, so the #P-hardness of exact S_{stat} on products (Bhattacharyya et al., 2023, Theorem 4) transfers to R^* for product class-conditionals, and Theorem 5.1 shows that on Bayes-net class-conditionals of in-degree 2 even deciding $R^* > 0$ is NP-complete. Theorem 4.1 is what remains available on products: a $(1+\varepsilon)$-approximation of R^*. Other bounds on Bayes error include generalized Bhattacharyya and Chernoff bounds (Nielsen, 2014). More recent work has revisited Bayes-error estimation: Noshad et al. (2019) used an ensemble of sample-based ball estimators; Theisen et al. (2021) evaluated state-of-the-art classifiers against Bayes optimality; Ishida et al. (2023) proposed a direct estimator for deep-network benchmarks. These works use samples or learned representations rather than exact algorithms on succinctly represented distributions.

Approximation of related similarity measures. The triangular discrimination $\Delta(P, Q) = \sum_x (P - Q)^2/(P + Q)$ is related to harmonic overlap via $\Delta = 2 - 4F$ (Section D.6), so the FPTAS for F gives Δ on products to additive error 2ε (as $F \leq 1/2$); on Bayes nets of in-degree 2 no FPTAS for Δ exists unless P = NP, since $\Delta(P, Q) = 0$ exactly when $P = Q$, and deciding $P \neq Q$ there is NP-complete (Bhattacharyya et al., 2023). The Bhattacharyya coefficient $\mathrm{BC}(P, Q) = \sum_x \sqrt{P(x)Q(x)} = F_{\sqrt{r}}(P, Q)$ and the squared Hellinger distance $H^2(P, Q) = \sum_x (\sqrt{P(x)} - \sqrt{Q(x)})^2 = 2(1 - \mathrm{BC}(P, Q))$ behave differently from the overlaps of Table 1 on products, where they factorize and are computed to any precision. On Bayes nets they are covered by the hardness side: $f(r) = \sqrt{r}$ is admissible, so by Theorem 5.1 deciding $\mathrm{BC}(P, Q) > 0$ is NP-complete on Bayes nets of in-degree 2; and since $H^2(P, Q) = 0$ exactly when $P = Q$, that is, when $d_{\mathrm{TV}}(P, Q) = 0$, the result of Bhattacharyya et al. (2023) shows that multiplicative approximation of H^2 on the same class is NP-hard. The same holds for $\mathbf{E}_{x\sim Q}[(P/Q)^\alpha]$, the argument of the Rényi divergence D_α for $\alpha \in (0, 1)$, whose integrand r^α is admissible.

Connection to randomized algorithms for #P-hard problems and counting complexity. The non-transfer between S_{stat} and d_{TV} — S_{stat} admits a deterministic FPTAS on products (Theorem 4.1), and so does $d_{\mathrm{TV}} = 1 - S_{\mathrm{stat}}$ (Feng et al., 2024), yet neither algorithm transfers to the other via Scheffé's identity $S_{\mathrm{stat}} + d_{\mathrm{TV}} = 1$ due to multiplicative-error blowup when one side is small — parallels the classical #DNF-vs-#CNF asymmetry (Karp et al., 1989): #DNF admits an FPRAS, but #CNF is NP-hard to multiplicatively approximate. Both phenomena reflect the same underlying issue: multiplicative approximation is fragile under additive transformations $f \mapsto C - f$ when the value of interest is small relative to C. The classical counting complexity literature provides extensive context: the class #P of counting problems associated to NP relations (Valiant, 1979), the subclass #BIS (counting independent sets in bipartite graphs) believed to be intermediate (Dyer et al., 2004), and the dichotomies for counting CSPs (Bulatov, 2013) all sit beneath these approximation questions. The full implications for the design of approximation schemes across the concave/convex divide — for instance, identifying which other concave/convex pairs admit asymmetric tractability — are an interesting direction for future work.

F Verification That the Functionals Are Admissible

This appendix verifies that each entry of Table 1 is admissible (Theorem 2.6), computes its slope C_f at the origin and $f(1)$, and confirms that the framework of Sections 4 and 5 applies. The operational meaning, applications, and connections to classical literature for each functional are in Section D.

F.1 Preliminaries

We collect the definitions, conventions, and standard analytic results used throughout this appendix, so that it is self-contained.

Admissibility and constants. Recall from Theorem 2.6 that $f\colon [0,\infty) \to [0,\infty)$ is admissible if it is concave, $f(0) = 0$ and $f \not\equiv 0$. The induced functional is $F_f(P,Q) := \mathbf{E}_{x\sim Q}[f(P/Q)]$. Every admissible f is non-decreasing and has a slope $C_f := \lim_{r\downarrow 0} f(r)/r = \sup_{r>0} f(r)/r \in (0,\infty]$ at the origin (Theorem 2.7(a), (d)). For each entry below we verify concavity and $f(0) = 0$, verify that f is non-decreasing, which with $f(0) = 0$ gives $f \geq 0$, and evaluate $f(1) > 0$, which gives $f \not\equiv 0$; we also compute C_f for Table 1.

Standard analytic facts. The verifications use the following classical results, which we state once and invoke by number.

Fact F.1 (First-derivative monotonicity test (Rudin, 1976, Thm. 5.11))**.** *If f is continuous on $[0,\infty)$ and differentiable on $(0,\infty)$ with $f'(r) \geq 0$ for all $r > 0$, then f is non-decreasing on $[0,\infty)$. (For $0 \leq a < b$, the mean value theorem gives $f(b) - f(a) = f'(\xi)(b-a) \geq 0$ for some $\xi \in (a,b)$.) In particular $f'(r) > 0$ on $(0,\infty)$ implies f non-decreasing.*

Fact F.2 (Second-derivative concavity test (Boyd & Vandenberghe, 2004, §3.1.4))**.** *If f is continuous on $[0,\infty)$ and twice differentiable on $(0,\infty)$ with $f''(r) \leq 0$ for all $r > 0$, then f is concave on $[0,\infty)$. (Then f' is non-increasing, and a function with non-increasing derivative is concave.)*

Fact F.3 (Slope at the origin)**.** *If $f(0) = 0$ and f is differentiable from the right at 0, then*

$$C_f = \lim_{r\downarrow 0} \frac{f(r)}{r} = \lim_{r\downarrow 0} \frac{f(r) - f(0)}{r - 0} = f'(0^+).$$

If moreover f has a convergent power series $f(r) = c_1 r + c_2 r^2 + \cdots$ near 0 (with $c_0 = f(0) = 0$), then $f(r)/r = c_1 + c_2 r + \cdots \to c_1$, so $C_f = c_1 = f'(0^+)$ is exactly the coefficient of r.

Fact F.4 (Standard Maclaurin expansions (Abramowitz & Stegun, 1964, Ch. 4))**.** *The following expansions hold near $r = 0$, with the stated radii of convergence R:*

$$\begin{aligned}
&\log(1+r) = r - \tfrac{r^2}{2} + \tfrac{r^3}{3} - \cdots \quad (R=1), && 1 - e^{-r} = r - \tfrac{r^2}{2} + \tfrac{r^3}{6} - \cdots \quad (R=\infty),\\
&\frac{r}{1+r} = r - r^2 + r^3 - \cdots \quad (R=1), && \arctan r = r - \tfrac{r^3}{3} + \tfrac{r^5}{5} - \cdots \quad (R=1),\\
&\tanh r = r - \tfrac{r^3}{3} + \tfrac{2r^5}{15} - \cdots \quad (R=\tfrac{\pi}{2}).
\end{aligned}$$

In every case the coefficient of r is 1.

Methodology. For each functional we verify monotonicity by exhibiting $f'(r) \geq 0$ and invoking Theorem F.1 (or, for piecewise functions, Theorem F.5 below); concavity by exhibiting $f''(r) \leq 0$ and invoking Theorem F.2 together with $f(0) = 0$ (or Theorem F.5); and we compute $C_f = f'(0^+)$ via Theorem F.3, reading the linear coefficient from Theorem F.4 where convenient. We then evaluate $f(1)$ and instantiate Theorems 4.1 and 5.1. The smooth f below are infinitely differentiable on an open interval containing $[0,\infty)$ ($(-1,\infty)$ for $\log(1+r)$ and $r/(1+r)$, and $\mathbb{R}$ for $1-e^{-r}$, $\tanh r$ and $\arctan r$), hence continuous on $[0,\infty)$, twice differentiable on $(0,\infty)$ and right-differentiable at 0, as Theorems F.1 to F.3 require. For Theorem 4.1, every entry of Section F.3–Section F.12 is admissible: it is non-negative (non-decreasing with $f(0) = 0$) and concave with $f(0) = 0$, as verified below, and $f \not\equiv 0$ since $f(1) > 0$. The parameters c and β are rational throughout. For φ we take $f(1)$ when it is rational (1, c and $1/2$), and otherwise $\varphi = 1/2$, which is at most $f(1)$ because $\log 2 > 1/2$ (as $e < 4$), $1 - 1/e > 1/2$ (as $e > 2$), $\tanh 1 = (e^2-1)/(e^2+1) > 1/2$ (as $e^2 > 3$) and $\pi/4 > 1/2$; for the two capped rows we take $\varphi = \min(1/2, \beta)$ and $\min(1/2, c)$, at most $f(1) = \min(\log 2, \beta)$ and $\min(1 - 1/e, c)$ respectively. Each f is built from rational constants, $\min$, and $\log$, $\exp$, $\tanh$ or $\arctan$, each of which can be evaluated at a rational to within 2^{-t} in time polynomial in t and the bit-length of the argument, which provides the oracle; for $\min(g, c)$, returning $\min(\widetilde{g}, c)$ with $|\widetilde{g} - g(v)| \leq 2^{-t}$ is within 2^{-t} of $f(v)$, since $|\min(a,c) - \min(a',c)| \leq |a - a'|$.

A note on positive scaling. We record two scaling closures for later reference; both are proved in full.

Input scaling. Fix $\lambda > 0$ and set $g(r) := f(\lambda r)$, where f is admissible. Then g is admissible, with slope λC_f: *monotone:* for $0 \leq r_1 \leq r_2$ we have $\lambda r_1 \leq \lambda r_2$, so $g(r_1) = f(\lambda r_1) \leq f(\lambda r_2) = g(r_2)$ since f is non-decreasing; *concave, with $g(0) = 0$:* for $\theta \in [0,1]$ and $r, s \geq 0$, $g(\theta r + (1-\theta)s) =$

$f(\theta(\lambda r) + (1-\theta)(\lambda s)) \geq \theta f(\lambda r) + (1-\theta) f(\lambda s) = \theta g(r) + (1-\theta) g(s)$ by concavity of f at the points $\lambda r, \lambda s$, and $g(0) = f(0) = 0$; *slope:* substituting $s = \lambda r$ (so $r = s/\lambda$ and $r \downarrow 0 \iff s \downarrow 0$),

$$C_g = \lim_{r \downarrow 0} \frac{f(\lambda r)}{r} = \lim_{s \downarrow 0} \frac{f(s)}{s/\lambda} = \lambda \lim_{s \downarrow 0} \frac{f(s)}{s} = \lambda C_f \in (0, \infty].$$

Finally $g(1) = f(\lambda) > 0$ by Theorem 2.7(c) applied to f, so $g \not\equiv 0$.

Output scaling. Fix $\mu > 0$ and set $h(r) := \mu f(r)$. Then h is non-decreasing (a positive multiple of a non-decreasing function), concave (a positive multiple of a concave function: $h(\theta r + (1-\theta)s) = \mu f(\theta r + (1-\theta)s) \geq \mu(\theta f(r) + (1-\theta) f(s)) = \theta h(r) + (1-\theta) h(s)$), with $h(0) = \mu f(0) = 0$ and $C_h = \lim_{r \downarrow 0} \mu f(r)/r = \mu C_f \in (0, \infty]$. Moreover $h(1) = \mu f(1) > 0$, so h is admissible, and by linearity of expectation $F_h(P, Q) = \mathbf{E}_{x \sim Q}[\mu f(P/Q)] = \mu \, \mathbf{E}_{x \sim Q}[f(P/Q)] = \mu F_f(P, Q)$. So the framework is closed under both input and output positive scaling.

F.2 A Closure Lemma for Pointwise Minima

Several entries below are pointwise minima of two functions (the capped and truncated similarities, the bounded-log and capped-exponential overlaps). The following lemma, proved in full, justifies the recurring claims that such minima are admissible; we invoke it by reference thereafter.

Lemma F.5 (Pointwise minimum preserves admissibility)**.** *Let $g_1, g_2 \colon [0, \infty) \to \mathbb{R}$ and set $f := \min(g_1, g_2)$, i.e., $f(r) = \min(g_1(r), g_2(r))$ for each r.*

(i) *If g_1 and g_2 are non-decreasing, then f is non-decreasing.*

(ii) *If g_1 and g_2 are concave on $[0, \infty)$, then f is concave on $[0, \infty)$.*

(iii) *If g_1, g_2 are admissible, then f is admissible, with $C_f = \min(C_{g_1}, C_{g_2})$.*

Proof. (i) Fix $0 \leq r_1 \leq r_2$. Write $m_1 := f(r_1) = \min(g_1(r_1), g_2(r_1))$. Since g_1 is non-decreasing, $m_1 \leq g_1(r_1) \leq g_1(r_2)$; since g_2 is non-decreasing, $m_1 \leq g_2(r_1) \leq g_2(r_2)$. Thus m_1 is a lower bound for both $g_1(r_2)$ and $g_2(r_2)$, hence for their minimum: $m_1 \leq \min(g_1(r_2), g_2(r_2)) = f(r_2)$. As $r_1 \leq r_2$ were arbitrary, f is non-decreasing.

(ii) Fix $r, s \geq 0$ and $\theta \in [0, 1]$, and set $w := \theta r + (1-\theta) s \in [0, \infty)$. For each $j \in \{1, 2\}$, concavity of g_j gives

$$g_j(w) \geq \theta \, g_j(r) + (1-\theta) \, g_j(s) \geq \theta \min(g_1(r), g_2(r)) + (1-\theta) \min(g_1(s), g_2(s)),$$

where the second inequality uses $g_j(r) \geq \min(g_1(r), g_2(r)) = f(r)$ and $g_j(s) \geq \min(g_1(s), g_2(s)) = f(s)$. The right-hand side, namely $\theta f(r) + (1-\theta) f(s)$, is independent of j, so it is a lower bound for both $g_1(w)$ and $g_2(w)$, hence for $f(w) = \min(g_1(w), g_2(w))$. Therefore $f(w) \geq \theta f(r) + (1-\theta) f(s)$, proving f concave.

(iii) Concavity of f follows from (ii), and $f(0) = \min(g_1(0), g_2(0)) = \min(0, 0) = 0$. Since g_1, g_2 take values in $[0, \infty)$, so does f, and $f(1) = \min(g_1(1), g_2(1)) > 0$ by Theorem 2.7(c) applied to g_1 and g_2, so $f \not\equiv 0$ and f is admissible. For the slope, note that for any reals a, b and any $c > 0$ we have $\min(a, b)/c = \min(a/c, b/c)$ (dividing by a positive number preserves order, hence preserves the minimum). Taking $a = g_1(r)$, $b = g_2(r)$, $c = r > 0$,

$$\frac{f(r)}{r} = \frac{\min(g_1(r), g_2(r))}{r} = \min\left(\frac{g_1(r)}{r}, \frac{g_2(r)}{r} \right).$$

By Theorem 2.7(a), $g_j(r)/r \to C_{g_j} \in (0, \infty]$ as $r \downarrow 0$. The map $(u, v) \mapsto \min(u, v)$ is continuous on $(-\infty, \infty]^2$, so the limit of the minimum equals the minimum of the limits:

$$C_f = \lim_{r \downarrow 0} \frac{f(r)}{r} = \lim_{r \downarrow 0} \min\left(\frac{g_1(r)}{r}, \frac{g_2(r)}{r} \right) = \min\left(\lim_{r \downarrow 0} \frac{g_1(r)}{r}, \lim_{r \downarrow 0} \frac{g_2(r)}{r} \right) = \min(C_{g_1}, C_{g_2}).$$

□

F.3 Statistical Similarity ($f(r) = \min(r, 1)$)

- **Monotone:** $r \mapsto r$ is non-decreasing and $r \mapsto 1$ is non-decreasing (constant), so $f = \min(r, 1)$ is non-decreasing by Theorem F.5(i).
- **Concave, $f(0) = 0$:** $r \mapsto r$ is concave (it is affine) and $r \mapsto 1$ is concave (constant), so f is concave by Theorem F.5(ii); and $f(0) = \min(0, 1) = 0$. (Here Theorem F.5(iii) does not apply directly because the constant 1 violates $g(0) = 0$, so we use parts (i)–(ii) and compute the slope directly.)
- **Slope:** for $r \in (0, 1]$ we have $r \leq 1$, hence $\min(r, 1) = r$ and $f(r)/r = 1$; therefore $C_f = \lim_{r \downarrow 0} f(r)/r = 1 \in (0, \infty)$.

$f(1) = \min(1, 1) = 1$. Framework instantiation: $C_f = f(1) = 1$; FPTAS via Theorem 4.1; NP-hardness via Theorem 5.1.

F.4 Capped Statistical Similarity ($f(r) = \min(r, c)$, $c \in (0, 1]$)

- **Monotone:** $r \mapsto r$ and the constant $r \mapsto c$ are both non-decreasing, so $f = \min(r, c)$ is non-decreasing by Theorem F.5(i).
- **Concave, $f(0) = 0$:** $r \mapsto r$ is affine (hence concave) and $r \mapsto c$ is constant (hence concave), so f is concave by Theorem F.5(ii); and $f(0) = \min(0, c) = 0$ since $c > 0$.
- **Slope:** since $c > 0$, for $r \in (0, c]$ we have $r \leq c$, hence $\min(r, c) = r$ and $f(r)/r = 1$; therefore $C_f = \lim_{r \downarrow 0} f(r)/r = 1 \in (0, \infty)$.

$f(1) = \min(1, c) = c$ since $c \leq 1$. Framework instantiation: $C_f = 1$, $f(1) = c$; FPTAS via Theorem 4.1; NP-hardness via Theorem 5.1.

F.5 Truncated Similarity ($f(r) = \min(r, c)$, $c > 1$)

For $c > 1$, $f(r) = \min(r, c)$ equals r on $[0, c]$ and the constant c on $[c, \infty)$.

- **Monotone:** $r \mapsto r$ and the constant $r \mapsto c$ are non-decreasing, so f is non-decreasing by Theorem F.5(i).
- **Concave, $f(0) = 0$:** $r \mapsto r$ is affine and $r \mapsto c$ is constant, both concave, so f is concave by Theorem F.5(ii); and $f(0) = \min(0, c) = 0$.
- **Slope:** since $c > 1 > 0$, for $r \in (0, c]$ we have $\min(r, c) = r$ and $f(r)/r = 1$; therefore $C_f = \lim_{r \downarrow 0} f(r)/r = 1 \in (0, \infty)$.

$f(1) = \min(1, c) = 1$ since $c > 1$. Framework instantiation: $C_f = f(1) = 1$; FPTAS via Theorem 4.1; NP-hardness via Theorem 5.1.

F.6 Log-Overlap ($f(r) = \log(1 + r)$)

- **Monotone:** $f(r) = \log(1 + r)$ gives $f'(r) = \frac{1}{1+r}$ (chain rule), which is > 0 for all $r \in [0, \infty)$; by Theorem F.1, f is non-decreasing (indeed strictly increasing).
- **Concave, $f(0) = 0$:** differentiating again, $f''(r) = -\frac{1}{(1+r)^2} < 0$ on $[0, \infty)$; by Theorem F.2, f is concave (strictly). And $f(0) = \log 1 = 0$.
- **Slope:** by Theorem F.3, $C_f = f'(0^+) = \frac{1}{1+0} = 1$. Equivalently, the Maclaurin expansion of $\log(1 + r)$ (Theorem F.4) has linear coefficient 1, so $f(r)/r \to 1$. Thus $C_f = 1 \in (0, \infty)$.

$f(1) = \log 2 \approx 0.693$. Framework instantiation: $C_f = 1$, $f(1) = \log 2 = O(1)$; FPTAS via Theorem 4.1; NP-hardness via Theorem 5.1.

F.7 Exponential Overlap ($f(r) = 1 - e^{-r}$)

- **Monotone:** $f(r) = 1 - e^{-r}$ gives $f'(r) = e^{-r}$ (chain rule), which is > 0 for all r; by Theorem F.1, f is non-decreasing.

- **Concave,** $f(0) = 0$**:** $f''(r) = -e^{-r} < 0$ on $[0, \infty)$; by Theorem F.2, f is strictly concave. And $f(0) = 1 - e^0 = 1 - 1 = 0$.
- **Slope:** by Theorem F.3, $C_f = f'(0^+) = e^0 = 1$. Equivalently, the Maclaurin expansion of $1 - e^{-r}$ (Theorem F.4) has linear coefficient 1, so $f(r)/r \to 1$. Thus $C_f = 1 \in (0, \infty)$.

$f(1) = 1 - 1/e \approx 0.632$. Framework instantiation: $C_f = 1$, $f(1) = 1 - 1/e = O(1)$; FPTAS via Theorem 4.1; NP-hardness via Theorem 5.1.

F.8 Harmonic / Triangular Overlap ($f(r) = r/(1 + r)$)

- **Monotone:** writing $f(r) = \frac{r}{1+r} = 1 - \frac{1}{1+r}$, the quotient/chain rule gives $f'(r) = \frac{1}{(1+r)^2} > 0$ on $[0, \infty)$; by Theorem F.1, f is non-decreasing (strictly).
- **Concave,** $f(0) = 0$**:** differentiating $f'(r) = (1 + r)^{-2}$ gives $f''(r) = -2(1 + r)^{-3} = -\frac{2}{(1+r)^3} < 0$ on $[0, \infty)$; by Theorem F.2, f is strictly concave. And $f(0) = 0/(1+0) = 0$.
- **Slope:** by Theorem F.3, $C_f = f'(0^+) = \frac{1}{(1+0)^2} = 1$. Equivalently, the Maclaurin expansion of $r/(1 + r)$ (Theorem F.4) has linear coefficient 1, so $f(r)/r \to 1$. Thus $C_f = 1 \in (0, \infty)$.

$f(1) = \frac{1}{1+1} = 1/2$. Framework instantiation: $C_f = 1$, $f(1) = 1/2 = O(1)$; FPTAS via Theorem 4.1; NP-hardness via Theorem 5.1.

F.9 Hyperbolic-Tangent Overlap ($f(r) = \tanh r$)

- **Monotone:** $f(r) = \tanh r$ gives $f'(r) = \operatorname{sech}^2 r = 1 - \tanh^2 r$, which is > 0 for all r (since $|\tanh r| < 1$); by Theorem F.1, f is non-decreasing.
- **Concave,** $f(0) = 0$**:** differentiating $f'(r) = \operatorname{sech}^2 r$ by the chain rule, and using $\frac{d}{dr}\operatorname{sech} r = -\operatorname{sech} r \tanh r$,

$$f''(r) = \frac{d}{dr}\big(\operatorname{sech}^2 r\big) = 2\operatorname{sech} r \cdot \big(-\operatorname{sech} r \tanh r\big) = -2\operatorname{sech}^2 r \cdot \tanh r.$$

  For $r \geq 0$ we have $\tanh r \geq 0$ (with equality only at $r = 0$) and $\operatorname{sech}^2 r > 0$, so $f''(r) \leq 0$ on $[0, \infty)$; by Theorem F.2, f is concave. And $f(0) = \tanh 0 = 0$.
- **Slope:** by Theorem F.3, $C_f = f'(0^+) = \operatorname{sech}^2 0 = 1$. Equivalently, the Maclaurin expansion of $\tanh r$ (Theorem F.4) has linear coefficient 1, so $f(r)/r \to 1$. Thus $C_f = 1 \in (0, \infty)$.

$f(1) = \tanh 1 = (e - e^{-1})/(e + e^{-1}) \approx 0.7616$. Framework instantiation: $C_f = 1$, $f(1) = \tanh 1 = O(1)$; FPTAS via Theorem 4.1; NP-hardness via Theorem 5.1.

F.10 Inverse-Tangent Overlap ($f(r) = \arctan r$)

- **Monotone:** $f(r) = \arctan r$ gives $f'(r) = \frac{1}{1+r^2} > 0$ for all r; by Theorem F.1, f is non-decreasing.
- **Concave,** $f(0) = 0$**:** differentiating $f'(r) = (1 + r^2)^{-1}$ by the chain rule, $f''(r) = -(1 + r^2)^{-2} \cdot 2r = -\frac{2r}{(1+r^2)^2}$, which is ≤ 0 on $[0, \infty)$ (numerator $-2r \leq 0$, denominator > 0); by Theorem F.2, f is concave. And $f(0) = \arctan 0 = 0$.
- **Slope:** by Theorem F.3, $C_f = f'(0^+) = \frac{1}{1+0} = 1$. Equivalently, the Maclaurin expansion of $\arctan r$ (Theorem F.4) has linear coefficient 1, so $f(r)/r \to 1$. Thus $C_f = 1 \in (0, \infty)$.

$f(1) = \arctan 1 = \pi/4 \approx 0.7854$. Framework instantiation: $C_f = 1$, $f(1) = \pi/4 = O(1)$; FPTAS via Theorem 4.1; NP-hardness via Theorem 5.1.

F.11 Bounded Log-Overlap ($f(r) = \min(\log(1 + r), \beta)$, $\beta > 0$)

- **Monotone:** $r \mapsto \log(1+r)$ is non-decreasing (shown in Section F.6) and $r \mapsto \beta$ is constant, hence non-decreasing; so $f = \min(\log(1 + r), \beta)$ is non-decreasing by Theorem F.5(i).

- **Concave,** $f(0) = 0$**:** $r \mapsto \log(1 + r)$ is concave (shown in Section F.6) and $r \mapsto \beta$ is constant, hence concave; so f is concave by Theorem F.5(ii). And $f(0) = \min(\log 1, \beta) = \min(0, \beta) = 0$ since $\beta > 0$.
- **Slope:** the inequality $\log(1 + r) < \beta$ is equivalent to $1 + r < e^\beta$, i.e., $r < e^\beta - 1$, and $e^\beta - 1 > 0$ because $\beta > 0$. Hence on the neighborhood $(0, e^\beta - 1)$ of 0 we have $f(r) = \log(1 + r)$, so $f(r)/r = \log(1 + r)/r \to 1$ as $r \downarrow 0$ (from Section F.6). Thus $C_f = 1 \in (0, \infty)$.

$f(1) = \min(\log 2, \beta)$, positive for any $\beta > 0$. Framework instantiation: $C_f = 1$, $f(1) = \min(\log 2, \beta)$; FPTAS via Theorem 4.1; NP-hardness via Theorem 5.1.

F.12 Capped Exponential Overlap ($f(r) = \min(1 - e^{-r}, c)$, $c \in (0, 1)$)

- **Monotone:** $r \mapsto 1 - e^{-r}$ is non-decreasing (shown in Section F.7) and $r \mapsto c$ is constant, hence non-decreasing; so $f = \min(1 - e^{-r}, c)$ is non-decreasing by Theorem F.5(i).
- **Concave,** $f(0) = 0$**:** $r \mapsto 1 - e^{-r}$ is concave (shown in Section F.7) and $r \mapsto c$ is constant, hence concave; so f is concave by Theorem F.5(ii). And $f(0) = \min(1 - e^0, c) = \min(0, c) = 0$.
- **Slope:** the inequality $1 - e^{-r} < c$ is equivalent to $e^{-r} > 1 - c$, i.e., $-r > \log(1 - c)$, i.e., $r < -\log(1 - c) = \log \frac{1}{1-c}$; since $c \in (0, 1)$ we have $1 - c \in (0, 1)$ and hence $\log \frac{1}{1-c} > 0$. Thus on the neighborhood $(0, \log \frac{1}{1-c})$ of 0 we have $f(r) = 1 - e^{-r}$, so $f(r)/r \to 1$ as $r \downarrow 0$ (from Section F.7). Thus $C_f = 1 \in (0, \infty)$.

$f(1) = \min(1 - 1/e, c)$, positive whenever $c > 0$. Framework instantiation: $C_f = 1$, $f(1) = \min(1 - 1/e, c)$; FPTAS via Theorem 4.1; NP-hardness via Theorem 5.1.

F.13 Closure under Convex Combinations and Pointwise Minimum

Admissible functions enjoy two closure properties, which lift the explicit Table 1 to a much richer class. The pointwise-minimum closure is exactly Theorem F.5(iii), already proved. We state and prove the convex-combination closure in full.

Lemma F.6 (Convex combination preserves admissibility)**.** *Let g_1, g_2 be admissible, and let $\alpha \in [0, 1]$. Then $f := \alpha g_1 + (1 - \alpha) g_2$ is admissible, with $f(1) = \alpha g_1(1) + (1 - \alpha) g_2(1) > 0$; and $C_f = \alpha C_{g_1} + (1 - \alpha) C_{g_2} \in (0, \infty]$, with the convention $0 \cdot \infty := 0$.*

Proof. Monotone. Fix $0 \le r_1 \le r_2$. Since each g_j is non-decreasing, $g_j(r_1) \le g_j(r_2)$. As $\alpha \ge 0$ and $1 - \alpha \ge 0$, multiplying and adding preserves the inequalities:

$$f(r_1) = \alpha g_1(r_1) + (1 - \alpha) g_2(r_1) \le \alpha g_1(r_2) + (1 - \alpha) g_2(r_2) = f(r_2).$$

Concave. Fix $r, s \ge 0$ and $\theta \in [0, 1]$, with $w := \theta r + (1 - \theta) s$. Concavity of each g_j gives $g_j(w) \ge \theta g_j(r) + (1 - \theta) g_j(s)$. Forming the convex combination with non-negative weights $\alpha, 1 - \alpha$,

$$f(w) = \alpha g_1(w) + (1 - \alpha) g_2(w) \ge \alpha \big[\theta g_1(r) + (1 - \theta) g_1(s)\big] + (1 - \alpha) \big[\theta g_2(r) + (1 - \theta) g_2(s)\big],$$

which rearranges to $\theta \big[\alpha g_1(r) + (1 - \alpha) g_2(r)\big] + (1 - \theta) \big[\alpha g_1(s) + (1 - \alpha) g_2(s)\big] = \theta f(r) + (1 - \theta) f(s)$, so f is concave. And $f(0) = \alpha g_1(0) + (1 - \alpha) g_2(0) = \alpha \cdot 0 + (1 - \alpha) \cdot 0 = 0$. Moreover $f \ge 0$, since $g_1, g_2 \ge 0$ (Theorem 2.6) and $\alpha, 1 - \alpha \ge 0$.

Slope. Suppose first that $C_{g_1}, C_{g_2} < \infty$. For $r > 0$, $f(r)/r = \alpha\, g_1(r)/r + (1 - \alpha)\, g_2(r)/r$. Taking $r \downarrow 0$ and using $g_j(r)/r \to C_{g_j}$ (the limits exist by Theorem 2.7(a) and are finite by assumption), the limit of a finite linear combination is the linear combination of the limits:

$$C_f = \lim_{r \downarrow 0} \frac{f(r)}{r} = \alpha C_{g_1} + (1 - \alpha) C_{g_2}.$$

This is finite (both $C_{g_j} < \infty$). It is positive: if $\alpha \in (0, 1)$ then both summands are positive; if $\alpha = 0$ it equals $C_{g_2} > 0$; if $\alpha = 1$ it equals $C_{g_1} > 0$. Hence $C_f \in (0, \infty)$. Suppose next that $C_{g_1} = \infty$. If $\alpha > 0$, then $f(r)/r \ge \alpha\, g_1(r)/r \to \infty$, using $g_2 \ge 0$, so $C_f = \infty = \alpha C_{g_1} + (1 - \alpha) C_{g_2}$; if $\alpha = 0$,

then $f = g_2$ and $C_f = C_{g_2} = 0 \cdot \infty + C_{g_2}$. The case $C_{g_2} = \infty$ is symmetric, with $1 - \alpha$ in place of α.

Finally, since g_1 and g_2 are admissible, Theorem 2.7(c) gives $g_1(1), g_2(1) > 0$, so $f(1) = \alpha g_1(1) + (1 - \alpha) g_2(1) > 0$ (a convex combination of two positive numbers, hence positive for every $\alpha \in [0, 1]$), so $f \not\equiv 0$ and f is admissible. □

Pointwise minimum. By Theorem F.5(iii), if g_1, g_2 are admissible then so is $f = \min(g_1, g_2)$, with $C_f = \min(C_{g_1}, C_{g_2})$. (The bounded log-overlap and capped exponential overlap are *not* instances of this with two admissible inputs, since their cap is a constant violating $g(0) = 0$; they are handled in Sections F.11 and F.12 via Theorem F.5(i)–(ii). Genuine instances arise, e.g., from $\min(\log(1 + r), \tanh r)$, where both inputs are admissible.)

Implications. By Theorems F.5 and F.6, the family of admissible functions is closed under convex combinations and under pointwise minimum, so the entries of Table 1 generate infinitely many further members. Every such f is admissible, so by Theorem 5.1 deciding $F(P, Q) > 0$ is NP-complete on Bayes nets of in-degree 2, with no computability assumption on f. The same hypotheses give, by Theorem 4.1, an FPTAS on products whenever f can be evaluated to within 2^{-t} in time polynomial in t and in the bit-length of its argument and a rational $\varphi \in (0, f(1)]$ is known; $f(1) > 0$ by Theorem 2.7(c).

F.14 Square-Root Overlaps: Infinite Slope at the Origin

The two functionals below are admissible with $C_f = \infty$, since $f(r)/r \to \infty$ as $r \downarrow 0$; neither theorem needs a finite slope, so Theorems 4.1 and 5.1 apply to them. For Theorem 4.1, $f(1) = 1$ for both, so $\varphi = 1$. The oracle sets $s := t+2$, $N := \lfloor v\, 4^s \rfloor$ and $u := \lfloor \sqrt{N} \rfloor / 2^s$ (an integer square root), so that $|u - \sqrt{v}| \le 2^{1-s} = 2^{-t-1}$, and returns $\min(u, 1)$, respectively $2u/(1 + u)$; these are within 2^{-t} of $f(v)$, because $\min(\cdot, 1)$ is 1-Lipschitz and $u \mapsto 2u/(1 + u)$ has derivative $2/(1 + u)^2 \le 2$ on $[0, \infty)$. Unlike the Bhattacharyya coefficient $\sum_x \sqrt{P(x)Q(x)} = F_{\sqrt{r}}(P, Q)$, $F_{\min(\sqrt{r},1)}$ does not factorize over the coordinates of a product: for $p = (1/2, 1/2)$ and $q = (1/8, 7/8)$, $F(p, q) = 1/8 + \sqrt{7}/4$, whose square $29/64 + \sqrt{7}/16$ is irrational, while $F(p \otimes p, q \otimes q) = 1/64 + 2 \cdot 7/64 + (49/64)(4/7) = 43/64$.

Capped root overlap, $f(r) = \min(\sqrt{r}, 1)$. The map $r \mapsto \sqrt{r}$ is non-decreasing, and it is concave on $[0, \infty)$: for $r, s \ge 0$ and $\theta \in [0, 1]$, $(\theta\sqrt{r} + (1 - \theta)\sqrt{s})^2 = \theta^2 r + (1 - \theta)^2 s + 2\theta(1 - \theta)\sqrt{rs} \le \theta^2 r + (1 - \theta)^2 s + \theta(1 - \theta)(r + s) = \theta r + (1 - \theta)s$, using $2\sqrt{rs} \le r + s$, and taking square roots gives $\theta\sqrt{r} + (1 - \theta)\sqrt{s} \le \sqrt{\theta r + (1 - \theta)s}$. The constant 1 is non-decreasing and concave, so f is non-decreasing and concave by Theorem F.5(i)–(ii), with $f(0) = \min(0, 1) = 0$. Finally $f \not\equiv 0$, since $f(1) = 1$. The slope is infinite because for $r \in (0, 1]$ we have $f(r) = \sqrt{r}$ and $f(r)/r = r^{-1/2} \to \infty$ as $r \downarrow 0$.

Root-harmonic overlap, $f(r) = 2\sqrt{r}/(1 + \sqrt{r})$. We verify admissibility and show that the slope at the origin is infinite. Write $u = \sqrt{r} = r^{1/2}$, so $f = \frac{2u}{1+u} = 2 - \frac{2}{1+u}$. The function f is continuous on $[0, \infty)$, since $r \mapsto \sqrt{r}$ is continuous there and $1 + \sqrt{r} \ge 1 > 0$; this is the continuity hypothesis of Theorems F.1 and F.2, which here cannot be read off from differentiability, because $f'(r) \to \infty$ as $r \downarrow 0$. Differentiating with the chain rule ($\frac{du}{dr} = \frac{1}{2} r^{-1/2}$),

$$f'(r) = -2 \cdot \big(-(1 + r^{1/2})^{-2} \big) \cdot \tfrac{1}{2} r^{-1/2} = \frac{r^{-1/2}}{(1 + r^{1/2})^2} = \frac{1}{\sqrt{r}\,(1 + \sqrt{r})^2} > 0 \quad (r > 0),$$

so f is non-decreasing by Theorem F.1, and $f(0) = 2 \cdot 0/(1+0) = 0$. Differentiating once more, with $f'(r) = r^{-1/2}(1 + r^{1/2})^{-2}$ and the product rule (and $\frac{d}{dr}(1 + r^{1/2})^{-2} = -2(1 + r^{1/2})^{-3} \cdot \frac{1}{2} r^{-1/2} = -r^{-1/2}(1 + r^{1/2})^{-3}$),

$$f''(r) = -\tfrac{1}{2} r^{-3/2}(1 + r^{1/2})^{-2} \;-\; r^{-1}(1 + r^{1/2})^{-3},$$

which is a sum of two strictly negative terms for $r > 0$, so $f''(r) < 0$ and f is concave by Theorem F.2. *The slope is infinite:*

$$\frac{f(r)}{r} = \frac{2\sqrt{r}/(1+\sqrt{r})}{r} = \frac{2}{\sqrt{r}\,(1+\sqrt{r})} \xrightarrow[r\downarrow 0]{} \frac{2}{0^+ \cdot 1} = +\infty,$$

so $C_f = +\infty$. Still $f(1) = 1$, so $f \not\equiv 0$. The associated functional $F(P,Q) = \sum_x 2\sqrt{P(x)Q(x)}/(1+\sqrt{P(x)/Q(x)})$ (using $Q(x)\sqrt{P(x)/Q(x)} = \sqrt{P(x)Q(x)}$) is closely related to the Bhattacharyya coefficient $\sum_x \sqrt{P(x)Q(x)}$ and to Hellinger distance, and it is sensitive to small likelihood ratios in a way that linear-near-zero f are not: for Theorem 4.1 this does not matter, since the near-zero slice receives no atom (Theorem A.2).

F.15 Non-Examples

For completeness we record natural functionals that are *not* admissible, illustrating the operational content of each requirement through the families of statistics it excludes.

Not concave: power r^p, $p > 1$. For $f(r) = r^p$, $f'(r) = pr^{p-1}$ and $f''(r) = p(p-1)r^{p-2}$, which for $p > 1$ satisfies $f''(r) > 0$ on $(0,\infty)$ (each factor $p, p-1, r^{p-2}$ is positive); hence f is strictly convex, not concave, so it is not admissible. The associated functional $F(P,Q) = \mathbf{E}_{x\sim Q}[(P/Q)^p]$ is the p-th moment of the likelihood ratio under Q. For $p = 2$ and $P \ll Q$ (that is, $P(x) = 0$ whenever $Q(x) = 0$; otherwise $\chi^2(P,Q) = +\infty$ while F is finite) it equals $\chi^2(P,Q) + 1$: indeed, with every sum over x with $Q(x) > 0$,

$$\begin{aligned}
\mathop{\mathbf{E}}_{x\sim Q}\big[(P/Q)^2\big] &= \sum_x Q(x)\frac{P(x)^2}{Q(x)^2} = \sum_x \frac{P(x)^2}{Q(x)},\\
\chi^2(P,Q) &= \sum_x \frac{(P(x)-Q(x))^2}{Q(x)} = \sum_x \frac{P(x)^2}{Q(x)} - 2\sum_x P(x) + \sum_x Q(x)\\
&= \sum_x \frac{P(x)^2}{Q(x)} - 2 + 1 = \sum_x \frac{P(x)^2}{Q(x)} - 1,
\end{aligned}$$

using $\sum_{x:Q(x)>0} P(x) = 1$ (by $P \ll Q$) and $\sum_{x:Q(x)>0} Q(x) = 1$; so $\mathbf{E}_{x\sim Q}\big[(P/Q)^2\big] = \chi^2(P,Q) + 1$. Convex f exhibits *reversed* subhomogeneity: for $\lambda \geq 1$, $f(\lambda r) = (\lambda r)^p = \lambda^p r^p = \lambda^p f(r) \geq \lambda f(r)$, where $\lambda^p \geq \lambda$ holds because $\lambda \geq 1$ and $p > 1$ give $(p-1)\log\lambda \geq 0$, i.e., $\lambda^{p-1} \geq 1$. This reversal breaks the upper-bound argument of Theorem 4.2, which relies on $f(\lambda r) \leq \lambda f(r)$. On products these moments pose no computational problem: since the Q-support is $\prod_i \mathrm{supp}(q_i)$, for every P, Q and every integer $p \geq 1$ they factorize as $\mathbf{E}_{x\sim Q}[(P/Q)^p] = \prod_{i=1}^n \sum_{y:q_i(y)>0} p_i(y)^p q_i(y)^{1-p}$, a product of n sums of ℓ rationals each, which is computed exactly in polynomial time.

Not concave: indicator $f(r) = \mathbf{1}[r \geq 1]$. This f is non-decreasing: for $r_1 \leq r_2$, if $r_1 \geq 1$ then $r_2 \geq 1$ so both values are 1; otherwise $f(r_1) = 0 \leq f(r_2)$. It is bounded ($f \in \{0,1\}$). But the jump at $r = 1$ violates concavity. Concretely, take the points 0 and 2 and the weight $\theta = \frac{3}{4}$: then $\theta\cdot 0 + (1-\theta)\cdot 2 = \frac{1}{4}\cdot 2 = \frac{1}{2}$, and concavity would require

$$f\big(\tfrac{1}{2}\big) \;\geq\; \theta f(0) + (1-\theta) f(2) = \tfrac{3}{4}\cdot 0 + \tfrac{1}{4}\cdot 1 = \tfrac{1}{4}.$$

But $f(\frac{1}{2}) = \mathbf{1}[\frac{1}{2} \geq 1] = 0 < \frac{1}{4}$, so the concavity inequality fails and f is not admissible. (Equivalently, the chord from $(0,0)$ to $(2,1)$ has value $\frac{1}{4}$ at $r = \frac{1}{2}$, strictly above $f(\frac{1}{2}) = 0$.) The associated functional $F(P,Q) = \mathbf{Pr}_Q[P/Q \geq 1]$ counts the Q-mass of atoms at which P is at least as likely as Q — a natural quantity in some testing contexts, but not amenable to our slice-based analysis, which requires f to be insensitive (up to a factor $1+\delta$) to small multiplicative changes in r. The discontinuity at $r = 1$ precludes any such bound there.

Not concave, and $f(0) \neq 0$: S_{stat}-deficit, $f(r) = 1 - \min(r,1) = \max(0, 1-r)$. First, the two forms agree: for $r \leq 1$, $\min(r,1) = r$, so $1 - \min(r,1) = 1 - r = \max(0, 1-r)$ (as $1 - r \geq 0$);

for $r > 1$, $\min(r, 1) = 1$, so $1 - \min(r, 1) = 0 = \max(0, 1 - r)$ (as $1 - r < 0$). ***It is not non-decreasing:*** on $[0, 1]$, $f(r) = 1 - r$ is strictly decreasing (e.g. $f(0) = 1 > 0 = f(1)$), so f is not non-decreasing. ***It fails admissibility on two counts:*** $f = \max(0, 1 - r)$ is the pointwise maximum of the two affine functions $r \mapsto 0$ and $r \mapsto 1 - r$, hence convex (not concave, and not affine); and $f(0) = \max(0, 1) = 1 \neq 0$. The associated functional is the total variation distance:

$$F(P, Q) = \mathop{\mathbf{E}}_{x\sim Q}[1 - \min(P/Q, 1)] = 1 - \mathop{\mathbf{E}}_{x\sim Q}[\min(P/Q, 1)]$$
$$= 1 - \sum_x \min(P(x), Q(x)) = 1 - S_{\mathrm{stat}}(P, Q) = d_{\mathrm{TV}}(P, Q),$$

where $\mathbf{E}_{x\sim Q}[\min(P/Q, 1)] = \sum_x Q(x) \min(P(x)/Q(x), 1) = \sum_x \min(P(x), Q(x)) = S_{\mathrm{stat}}(P, Q)$, and the last equality uses the identity $S_{\mathrm{stat}} + d_{\mathrm{TV}} = 1$. Total variation admits a deterministic FPTAS on product distributions via the techniques of Feng et al. (2024) but not via our concave framework. This multi-axis failure is the structural manifestation of the concave/convex asymmetry discussed in Section 6.

Not concave and not non-negative: KL-style integrand, $f(r) = r \log r$ (with $0 \log 0 = 0$). Here $f'(r) = \log r + 1$ and $f''(r) = 1/r$. ***It is not non-decreasing:*** $f'(r) = \log r + 1 = 0$ at $r = 1/e$, with $f'(r) < 0$ for $r \in (0, 1/e)$ and $f'(r) > 0$ for $r > 1/e$; so f is strictly decreasing on $(0, 1/e)$ and is not non-decreasing. ***It is not concave:*** $f''(r) = 1/r > 0$ on $(0, \infty)$, so f is strictly convex, not concave. (Indeed $f < 0$ on $(0, 1)$, so f is not even $[0, \infty)$-valued.) For $P \ll Q$ the associated functional is exactly the Kullback–Leibler divergence, with sums over x with $Q(x) > 0$ and terms with $P(x) = 0$ vanishing by $0 \log 0 = 0$:

$$F(P, Q) = \mathop{\mathbf{E}}_{x\sim Q}[(P/Q) \log(P/Q)] = \sum_{x:Q(x)>0} Q(x) \frac{P(x)}{Q(x)} \log \frac{P(x)}{Q(x)}$$
$$= \sum_{x:Q(x)>0} P(x) \log \frac{P(x)}{Q(x)} = D_{\mathrm{KL}}(P\|Q)$$

(Cover & Thomas, 2006); without $P \ll Q$, $D_{\mathrm{KL}}(P\|Q) = +\infty$ while $F(P, Q)$ is finite. Although $r \log r$ is not admissible, KL on product distributions $P = \prod_i p_i$, $Q = \prod_i q_i$ over $[\ell]^n$ is polynomial-time computable via additivity. To see this, assume $P \ll Q$, equivalently $\mathrm{supp}(p_i) \subseteq \mathrm{supp}(q_i)$ for all i, sum over $x \in \mathrm{supp}(P) = \prod_i \mathrm{supp}(p_i)$, where every $p_i(x_i), q_i(x_i) > 0$ and all logarithms are finite, factor the log of the product and exchange the order of summation:

$$D_{\mathrm{KL}}(P\|Q) = \sum_{x\in\mathrm{supp}(P)} P(x) \log \frac{\prod_i p_i(x_i)}{\prod_i q_i(x_i)} = \sum_{x\in\mathrm{supp}(P)} P(x) \sum_{i=1}^n \log \frac{p_i(x_i)}{q_i(x_i)}$$
$$= \sum_{i=1}^n \sum_{x\in\mathrm{supp}(P)} P(x) \log \frac{p_i(x_i)}{q_i(x_i)}.$$

For each fixed i, the summand depends on x only through the coordinate x_i, and marginalizing the product distribution over the other coordinates gives, for $a \in \mathrm{supp}(p_i)$, $\sum_{x\in\mathrm{supp}(P):x_i=a} P(x) = \sum_{x:x_i=a} P(x) = p_i(a)$, since P vanishes off $\mathrm{supp}(P)$; hence

$$\sum_{x\in\mathrm{supp}(P)} P(x) \log \frac{p_i(x_i)}{q_i(x_i)} = \sum_{a\in\mathrm{supp}(p_i)} p_i(a) \log \frac{p_i(a)}{q_i(a)} = D_{\mathrm{KL}}(p_i\|q_i),$$

so $D_{\mathrm{KL}}(P\|Q) = \sum_{i=1}^n D_{\mathrm{KL}}(p_i\|q_i)$, a sum of n terms each needing $O(\ell)$ arithmetic operations and logarithm evaluations, hence computable to within 2^{-t} in time polynomial in n, ℓ, b_0 and t. (As usual this is finite under $\mathrm{supp}(p_i) \subseteq \mathrm{supp}(q_i)$, with the conventions $0 \log(0/q) = 0$ for every $q \geq 0$ and $a \log(a/0) = +\infty$ for $a > 0$.) The framework's inapplicability here thus reflects a qualitative structural difference (convex, non-monotone integrand) rather than computational difficulty.

G Hardness of Exact Computation for Functionals

This appendix complements Theorem 4.1 from the algorithmic side: for the FPTAS guarantee on product distributions to be substantive, exact computation of F on the same input class should be

intractable. The starting point is that exact S_{stat} on product distributions over $\{0,1\}^n$ with $\mathrm{poly}(n)$-bit rational marginals is #P-hard (Bhattacharyya et al., 2023). We establish #P-hardness for the piecewise-linear members of Table 1 (capped and truncated similarity, convex combinations) via direct Turing reductions, explain why a natural moment-interpolation idea for the smooth members (log-, exp-, harmonic-, $\tanh$-, $\arctan$-overlap) cannot work, since the likelihood-ratio moments $M_K(P,Q) := \mathbf{E}_{x\sim Q}\big[(P/Q)^K\big]$ it would draw hardness from are computable exactly on products (so smooth-side hardness on products remains open), and discuss bounded variants and overall status.

G.1 Preliminaries

We collect the notation, conventions, and external results used throughout this appendix, so that it is self-contained.

Notation and conventions. Throughout, P, Q are product distributions over $\{0,1\}^n$ (or $[\ell]^n$) with $\mathrm{poly}(n)$-bit rational marginals $\{p_i\}, \{q_i\}$, and all functionals are evaluated on the Q-support $\{x : Q(x) > 0\}$ with the convention that atoms of Q-measure zero contribute nothing to a Q-expectation. We write:

- $F_f(P,Q) := \mathbf{E}_{x\sim Q}[f(P/Q)] = \sum_{x:Q(x)>0} Q(x)\, f\big(P(x)/Q(x)\big)$, the functional induced by f;
- $M_k(P,Q) := \mathbf{E}_{x\sim Q}\big[(P/Q)^k\big] = \sum_{x:Q(x)>0} P(x)^k/Q(x)^{k-1}$, the k-th likelihood-ratio moment (so $M_1 = \sum_{x:Q(x)>0} P(x) \le 1$, with equality iff $P \ll Q$);
- $\tau := \min\{p_i(a), q_i(a) : 1 \le i \le n,\ a,\ \text{the value is nonzero}\}$, the smallest nonzero marginal probability, and $B := 1/\tau^n$, so that $P(x)/Q(x) \le B$ for all x in the Q-support (as $P(x) \le 1$ and $Q(x) \ge \tau^n$);
- $S_{\mathrm{stat}\,c}(P,Q) := \mathbf{E}_{x\sim Q}[\min(P/Q, c)] = \sum_x \min(P(x), c\,Q(x))$ for a cap $c > 0$, and $S_{\mathrm{stat}} := S_{\mathrm{stat}\,1}$.

Complexity background. We use #P and poly-time Turing (Cook) reducibility in the standard sense. For functions g, h, write $g \le_{\mathrm{T}}^{\mathrm{p}} h$ to mean that g is computable in polynomial time by an algorithm with oracle access to h (a poly-time Turing reduction of g to h).

Fact G.1 (#P and Turing reductions (Valiant, 1979))**.** *#P is the class of functions counting the accepting paths of a poly-time nondeterministic Turing machine. A function h is #P-hard iff $g \le_{\mathrm{T}}^{\mathrm{p}} h$ for every $g \in$ #P. The relation $\le_{\mathrm{T}}^{\mathrm{p}}$ is transitive, and #P-hardness transfers along it: if g is #P-hard and $g \le_{\mathrm{T}}^{\mathrm{p}} h$, then h is #P-hard.*

Fact G.2 (Exact S_{stat} is #P-hard on products (Bhattacharyya et al., 2023, Theorem 4))**.** *Computing $d_{\mathrm{TV}}(P,Q)$ exactly, given product distributions P, Q over $\{0,1\}^n$ with rational marginals, is #P-complete. Since $S_{\mathrm{stat}} = 1 - d_{\mathrm{TV}}$, one oracle call to exact S_{stat} and one subtraction compute d_{TV}, so computing $S_{\mathrm{stat}}(P,Q) = \sum_x \min(P(x), Q(x))$ exactly is #P-hard.*

Analytic background. For the smooth-functional strategy we use two classical facts.

Fact G.3 (Coefficient bound inside the disk of convergence)**.** *Let $f(r) = \sum_{k\ge 0} a_k r^k$ have radius of convergence $R \in (0,\infty]$. For every r_0 with $0 < r_0 < R$, the series $\sum_{k\ge 0} |a_k|\, r_0^k$ converges; consequently $A := \sup_{k\ge 0} |a_k|\, r_0^k < \infty$, i.e., $|a_k| \le A\, r_0^{-k}$ for all k. Moreover, for $|r| < R$ the series converges absolutely, so expectation and summation may be interchanged: $\mathbf{E}_{x\sim Q}[f(\lambda P/Q)] = \sum_{k\ge 0} a_k \lambda^k M_k$ whenever $\lambda B < R$.*

Fact G.4 (Descartes' rule of signs (Descartes, 1637; Wang, 2004))**.** *Let $p(t) = \sum_{j=1}^m c_j t^{e_j}$ be a real polynomial with m nonzero terms and exponents $e_1 < \cdots < e_m$. Then the number of positive real roots of p, counted with multiplicity, is at most the number of sign changes in the coefficient sequence $(c_1, \ldots, c_m)$, which is at most $m-1$. In particular, p has at most $m-1$ distinct positive real roots.*

G.2 Piecewise-Linear Functionals: Direct Reductions

G.2.1 Capped and Truncated Statistical Similarity ($f(r) = \min(r, c)$, $c > 0$)

Theorem G.5 (Hardness of capped similarity)**.** *For every fixed rational $c > 0$, exact computation of $S_{\mathrm{stat}c}(P, Q) = \mathbf{E}_{x\sim Q}[\min(P/Q, c)]$ on product distributions P, Q over $\{0,1\}^n$ is #P-hard. This covers the capped similarities ($c \le 1$) and the truncated similarities ($c > 1$) of Table 1.*

Proof. We give a poly-time Turing reduction from exact S_{stat} that makes a single oracle call. We treat $c \in (0, 1]$ first and $c > 1$ at the end. First, on the Q-support, $Q(x)\min(P(x)/Q(x), c) = \min(P(x), c\,Q(x))$, so

$$S_{\mathrm{stat}c}(P, Q) = \mathop{\mathbf{E}}_{x\sim Q}[\min(P/Q, c)] = \sum_x \min(P(x), c\,Q(x))$$

(atoms with $Q(x) = 0$ contribute $\min(P(x), 0) = 0$ and may be included harmlessly).

Given a S_{stat}-instance (P, Q) on $\{0,1\}^n$, construct (P', Q') on $\{0,1\}^{n+1}$ by prepending an auxiliary coordinate X_0 with marginals $P_0 = (c, 1-c)$ and $Q_0 = (1, 0)$, and set $P' := P_0 \otimes P$, $Q' := Q_0 \otimes Q$. These are valid product distributions: P_0 is a probability vector since $c \in (0, 1]$ gives $c, 1-c \ge 0$ with $c + (1-c) = 1$, and Q_0 is the point mass at 0.

Split the sum defining $S_{\mathrm{stat}c}(P', Q')$ over the two values of X_0. Using $P'(0, x) = c\,P(x)$, $Q'(0, x) = Q(x)$, $P'(1, x) = (1-c)P(x)$, and $Q'(1, x) = 0$, for each $x \in \{0,1\}^n$:

$$\min\big(P'(0,x),\, c\,Q'(0,x)\big) = \min\big(c\,P(x),\, c\,Q(x)\big) = c\cdot\min(P(x), Q(x)),$$
$$\min\big(P'(1,x),\, c\,Q'(1,x)\big) = \min\big((1-c)P(x),\, 0\big) = 0 \quad (\text{since } (1-c)P(x) \ge 0).$$

Therefore, splitting over $X_0 \in \{0,1\}$,

$$\begin{aligned} S_{\mathrm{stat}c}(P', Q') &= \sum_x \min\big(P'(0,x), c\,Q'(0,x)\big) + \sum_x \min\big(P'(1,x), c\,Q'(1,x)\big) \\ &= c\sum_x \min(P(x), Q(x)) + 0 = c\cdot S_{\mathrm{stat}}(P, Q). \end{aligned}$$

Since c is a fixed rational of constant bit-complexity, the map $(P, Q) \mapsto (P', Q')$ and the subsequent division by c are both computable in polynomial time. Hence, computing $S_{\mathrm{stat}}(P, Q)$ requires one oracle call to $S_{\mathrm{stat}c}$ and one division, i.e., $S_{\mathrm{stat}} \le_{\mathrm{T}}^{\mathrm{p}} S_{\mathrm{stat}c}$.

For $c > 1$, prepend instead an auxiliary coordinate with marginals $P_0 = (1, 0)$ and $Q_0 = (1/c, 1 - 1/c)$, which are probability vectors since $c > 1$ gives $1/c, 1 - 1/c \in (0, 1)$, and set $P' := P_0 \otimes P$, $Q' := Q_0 \otimes Q$. Now $P'(0, x) = P(x)$, $Q'(0, x) = Q(x)/c$, $P'(1, x) = 0$ and $Q'(1, x) = (1 - 1/c)Q(x)$, so for each $x \in \{0,1\}^n$

$$\min\left(P'(0,x),\, c\,Q'(0,x)\right) = \min\left(P(x),\, Q(x)\right),$$
$$\min\left(P'(1,x),\, c\,Q'(1,x)\right) = \min\left(0,\, (c-1)Q(x)\right) = 0 \quad (\text{since } (c-1)Q(x) \ge 0),$$

and summing over $X_0 \in \{0,1\}$ gives $S_{\mathrm{stat}c}(P', Q') = S_{\mathrm{stat}}(P, Q)$, with no division needed. In both cases, since S_{stat} is #P-hard (Theorem G.2), Theorem G.1 gives that $S_{\mathrm{stat}c}$ is #P-hard. □

Remark G.6 (Edge cases). The case $c = 1$ recovers S_{stat} itself; the proof simplifies since the auxiliary coordinate becomes degenerate. For $c \to 0$ we have $S_{\mathrm{stat}c} \le c$, so the quantity itself becomes small; exact computation remains #P-hard for any rational c of polynomial bit-length, since the division by c remains polynomial-time.

Remark G.7 (Differential privacy). The complement of $S_{\mathrm{stat}c}$ is the hockey-stick divergence: since $P(x) - \min(P(x), c\,Q(x)) = \max(P(x) - c\,Q(x), 0)$, we have $\sum_x \max(P(x) - c\,Q(x), 0) = \sum_x P(x) - S_{\mathrm{stat}c}(P, Q)$. Murtagh & Vadhan (2018, Theorem 1.5) characterize the optimal composition of differentially private algorithms through this divergence, with $c = e^{\epsilon_g}$, between products of two-point distributions, and prove that computing the optimal composition is #P-complete (Murtagh & Vadhan, 2018, Theorem 1.6); their approximation algorithm has additive error in ϵ_g (Murtagh & Vadhan, 2018, Theorem 1.7). Theorem G.5 is the corresponding statement for $S_{\mathrm{stat}c}$ itself, for every rational $c > 0$, and Theorem 4.1 approximates $S_{\mathrm{stat}c}(P, Q)$ within a factor $1 + \varepsilon$ on products; this gives no multiplicative guarantee for the hockey-stick divergence itself (cf. Theorem 2.9).

G.2.2 Convex Combinations

Proposition G.8 (Hardness of convex combinations). *Let $f = \alpha g + (1-\alpha)h$ for rational $\alpha \in (0, 1]$, where g is admissible and exact computation of F_g on products is #P-hard, while F_h is polynomial-time computable. Then exact computation of F_f on products is #P-hard.*

Proof. By linearity of expectation, since $f(r) = \alpha g(r) + (1-\alpha)h(r)$ pointwise,

$$\begin{aligned} F_f(P,Q) &= \mathop{\mathbf{E}}_{x\sim Q}[\alpha g(P/Q) + (1-\alpha)h(P/Q)] \\ &= \alpha \mathop{\mathbf{E}}_{x\sim Q}[g(P/Q)] + (1-\alpha)\mathop{\mathbf{E}}_{x\sim Q}[h(P/Q)] = \alpha F_g(P,Q) + (1-\alpha)F_h(P,Q). \end{aligned}$$

Given the input (P, Q), a single call to an exact F_f-oracle returns $F_f(P, Q)$, and $F_h(P, Q)$ is computed in poly-time by hypothesis; then $\alpha F_g(P, Q) = F_f(P, Q) - (1-\alpha)F_h(P, Q)$ is obtained by subtraction, and $F_g(P, Q) = \alpha^{-1}\big(F_f(P, Q) - (1-\alpha)F_h(P, Q)\big)$ by dividing by α (a fixed nonzero rational of constant bit-complexity, so the division is poly-time). Hence $F_g \le_{\mathrm{T}}^{\mathrm{p}} F_f$, and by Theorem G.1 the #P-hardness of F_g transfers to F_f. (Here F_f and F_h are exact rationals, as they are for piecewise-rational g and h such as the instance below; for smooth h, the same argument with both values requested to precision $\alpha 2^{-t-2}$ gives F_g to within 2^{-t-1}, and for $g = \min(r, c)$ with $c = c_1/c_2$ in lowest terms this determines F_g exactly: with D_i the product of the denominators of the entries of p_i and q_i and $D := c_2 \prod_i D_i$, an integer of $O(n\ell b_0 + \log c_2)$ bits, every $P(x)$ and every $c\,Q(x)$ is an integer multiple of $1/D$, hence so is $F_g(P, Q) = \sum_x \min(P(x), c\,Q(x))$, and taking $t := \lceil \log_2 D \rceil + 1$ makes the error less than $1/(2D)$, so rounding D times the estimate to the nearest integer recovers $D \cdot F_g(P, Q)$.) □

Remark G.9 (A concrete instance). Theorem G.8 can be applied with $g(r) = \min(r, c)$ for a fixed rational $c > 0$ (S_{stat} at $c = 1$ by Theorem G.2, and $S_{\mathrm{stat}c}$ in general by Theorem G.5) and h any function $[0, \infty) \to [0, \infty)$ whose functional F_h is computable exactly in polynomial time on products. For instance, taking $h(r) = r$ gives, by factorization of the per-coordinate sum,

$$F_h(P,Q) = \mathop{\mathbf{E}}_{x\sim Q}[P/Q] = \sum_{x:Q(x)>0} P(x) = \sum_{x:\forall i,\, q_i(x_i)>0} \prod_i p_i(x_i) = \prod_i \sum_{x_i:q_i(x_i)>0} p_i(x_i),$$

which is polynomial-time computable for any product distribution (a product of n sums of $\le \ell$ terms) and equals 1 when $P \ll Q$, that is, when $\mathrm{supp}(p_i) \subseteq \mathrm{supp}(q_i)$ for every i. Hence, for every rational $\alpha \in (0, 1]$, exact computation of F_f for $f(r) = \alpha \min(r, 1) + (1-\alpha)r$ is #P-hard on products. This f is admissible: it is a combination with non-negative weights of non-decreasing concave functions vanishing at 0, and $f(r) = r$ on $[0, 1]$, so $C_f = 1$.

G.3 Smooth Functionals: Moment-Interpolation Strategy

For the smooth concave functionals

$$f \in \{\log(1+r),\ 1-e^{-r},\ r/(1+r),\ \tanh r,\ \arctan r\},$$

we are not aware of #P-hardness reductions to or from S_{stat} on product distributions, and smooth-side hardness on products remains open. This subsection records a natural moment-interpolation idea and explains why it cannot settle the question. The idea would transfer hardness from the moments of the likelihood ratio under Q to F_f, but these moments are computable exactly in polynomial time on products (Theorem G.12), so there is no hardness to transfer. We keep the analytic pieces (Sections G.3.1 to G.3.4), which show how exact values of F_f on rescaled instances encode the moments, and Section G.3.5 says why the route stops there.

G.3.1 The Moment-Interpolation Strategy

For each smooth f above, the analytic Taylor expansion at $r = 0$ has form $f(r) = \sum_{k\ge1} a_k r^k$ with $a_1 = C_f \in (0, \infty)$ and the remaining coefficients $\{a_k\}_{k\ge2}$ explicit (for $\log(1+r)$: $a_k = (-1)^{k-1}/k$; for $1-e^{-r}$: $a_k = (-1)^{k-1}/k!$; for $\arctan r$: $a_{2k+1} = (-1)^k/(2k+1)$ and $a_{2k} = 0$; for $\tanh r$: explicit via Bernoulli numbers; for $r/(1+r)$: $a_k = (-1)^{k-1}$). Note that the radii of convergence vary: $1-e^{-r}$ is entire (radius ∞); $\tanh r$ has radius $\pi/2$; $\log(1+r)$, $\arctan r$, and $r/(1+r)$ have radius 1.

Idea. For rational scalings $\lambda > 0$, define $F_\lambda(P, Q) := \mathbf{E}_{x\sim Q}[f(\lambda P/Q)]$. Expanding f in its Taylor series and interchanging expectation and summation — valid whenever $\lambda B < R$, by Theorem G.3 — gives

$$F_\lambda(P, Q) = \sum_{k\geq 1} a_k \lambda^k M_k,$$

where $M_k = M_k(P, Q) = \mathbf{E}_{x\sim Q}\big[(P/Q)^k\big]$ is the k-th likelihood-ratio moment defined in the Preliminaries (Section G.1). Multiple queries to an exact F_λ-oracle at $\lambda_0, \lambda_1, \ldots$ should determine the moments M_k via linear algebra, after truncating the tail of the series.

The moments encode rich structure: $M_1 = \sum_{x:Q(x)>0} P(x)$, equal to 1 when $P \ll Q$ (the absolute-continuity case relevant for our reduction); when $P \ll Q$, $M_2 = \chi^2(P, Q) + 1$ where $\chi^2(P, Q) = \sum_{x:Q(x)>0} P(x)^2/Q(x) - 1$ is the chi-squared divergence (Tsybakov, 2009); higher moments encode finer ratio statistics (M_k is the k-th moment of P/Q under Q, equivalently, when $P \ll Q$, the $(k-1)$-th moment of P/Q under P).

A polynomial-time recovery of $M_2 = \chi^2 + 1$ from finitely many F_λ values via the linear system would reduce exact computation of M_2 to exact computation of F on products (for log-overlap, exponential overlap and harmonic overlap, which have $a_1 = 1$ and nonzero a_2; for $\tanh$ and $\arctan$ overlaps, $a_2 = 0$ but $a_3 \neq 0$, and the analogous recovery targets M_3). Such a reduction would give hardness of F only if the moments were hard, and they are not: by Theorem G.12, every M_K with K a positive integer bounded by a polynomial in the input size is computable exactly in polynomial time on products. The strategy therefore cannot yield #P-hardness of F, whatever the conditioning of the linear system.

G.3.2 Auxiliary-Coordinate Construction

Given a target moment M_K on a product instance (P, Q) over $\{0, 1\}^n$ and a smooth f as above, fix an analyticity radius $r_0 \in (0, R)$ (where R is the radius of convergence of the Taylor series of f), as in Theorem G.10. The construction extends (P, Q) to multiple instances (P'_i, Q'_i) on $\{0, 1\}^{n+1}$ by prepending an auxiliary first coordinate that scales the likelihood ratio by λ_i. (One instance per unknown of the truncated system: for a truncation order $T \geq K$, one distinct scaling λ_i for each $k \leq T$ with $a_k \neq 0$, the index set to which Theorem G.11 is applied.) Specifically, for rational $\lambda_i \in (0, \min(1, r_0/(2B))]$ where $B = 1/\tau^n$ (so λ_i is both a valid probability and within the convergence guarantee of Theorem G.10), set:

$$P'_i := (\lambda_i, 1-\lambda_i) \otimes P, \qquad Q'_i := (1, 0) \otimes Q,$$

i.e., $P'_i(0, x) = \lambda_i P(x)$, $P'_i(1, x) = (1-\lambda_i)P(x)$, $Q'_i(0, x) = Q(x)$, $Q'_i(1, x) = 0$ for each $x \in \{0, 1\}^n$. The Q'_i-support is $\{(0, x) : Q(x) > 0\}$, on which the per-atom ratio is $P'_i(0, x)/Q'_i(0, x) = \lambda_i P(x)/Q(x)$. Hence

$$F_f(P'_i, Q'_i) = \mathop{\mathbf{E}}_{x\sim Q}[f(\lambda_i P/Q)] = \sum_{k\geq 1} a_k \lambda_i^k M_k. \tag{G.1}$$

G.3.3 Truncation Bound

Lemma G.10 (Truncation tail)**.** *Let $f(r) = \sum_{k\geq 1} a_k r^k$ for all $r \in [0, R)$, where $R \in (0, \infty]$ is the radius of convergence, and fix any r_0 with $0 < r_0 < R$. Then there is a finite constant $A = A(r_0)$ such that $|a_k| \leq A\, r_0^{-k}$ for all $k \geq 1$. Let (P, Q) be product distributions over $\{0, 1\}^n$ with τ as in Section G.1, and $B := 1/\tau^n \geq 1$. For any truncation order $T \geq 1$ and any rational scaling λ with $0 < \lambda \leq \min(1, r_0/(2B))$,*

$$\left| F_f(P', Q') - \sum_{k=1}^{T} a_k \lambda^k M_k \right| \leq A \cdot 2^{-T},$$

where (P', Q') is the auxiliary construction of Section G.3.2 with scaling λ.

Proof. Existence of A. By Theorem G.3, since $r_0 < R$ there is a finite constant $A = \sup_{k\geq 1} |a_k| r_0^k$ with $|a_k| \leq A\, r_0^{-k}$ for all $k \geq 1$ (the point r_0 lies strictly inside the disk of convergence, so

$\sum_{k\geq 1} |a_k| r_0^k < \infty$ and its terms are bounded). (For the entire $1 - e^{-r}$ any finite $r_0 > 0$ may be used; for $\tanh r$ any $r_0 < \pi/2$.)

Moment bound. On the Q-support, $0 \leq P(x)/Q(x) \leq B$, so for every $k \geq 1$, $(P(x)/Q(x))^k \leq B^{k-1}\,(P(x)/Q(x))$; taking $\mathbf{E}_{x\sim Q}[\cdot]$ gives $M_k \leq B^{k-1} M_1$. Moreover $M_1 = \sum_{x:Q(x)>0} P(x) \leq \sum_x P(x) = 1$, and $B \geq 1$ (since $\tau \leq 1$), so $M_k \leq B^{k-1} \leq B^k$.

Absolute convergence and tail. For every x with $Q(x) > 0$, the ratio $r_x := \lambda P(x)/Q(x)$ satisfies $0 \leq r_x \leq \lambda B \leq r_0/2 < R$, so $f(r_x) = \sum_{k\geq 1} a_k r_x^k$; since the sum over x is finite, it may be exchanged with these convergent series, which gives (G.1): $F_f(P', Q') = \sum_{k\geq 1} a_k \lambda^k M_k$. Combining the two bounds and using $\lambda \leq r_0/(2B)$ (so $\lambda B/r_0 \leq 1/2$), this series converges absolutely:

$$\sum_{k\geq 1} |a_k|\, \lambda^k M_k \;\leq\; \sum_{k\geq 1} A\, r_0^{-k}\, \lambda^k\, B^k \;=\; \sum_{k\geq 1} A\,(\lambda B/r_0)^k \;\leq\; \sum_{k\geq 1} A\, 2^{-k} \;=\; A \;<\; \infty.$$

The truncation error is the tail of this absolutely convergent series, so by the triangle inequality and the same termwise bound,

$$\left| F_f(P', Q') - \sum_{k=1}^{T} a_k \lambda^k M_k \right| = \left| \sum_{k>T} a_k \lambda^k M_k \right| \leq \sum_{k>T} |a_k|\, \lambda^k M_k$$
$$\leq \sum_{k>T} A\,(\lambda B/r_0)^k \leq \sum_{k>T} A\, 2^{-k} = A \cdot 2^{-T},$$

where the final equality is the geometric-series identity $\sum_{k>T} 2^{-k} = 2^{-T}$. □

G.3.4 Vandermonde Recovery

Lemma G.11 (Vandermonde inversion)**.** *Let $S \subseteq \mathbb{Z}_{\geq 1}$ be a finite index set with $|S| = s$, let $\{a_k\}_{k\in S} \subseteq \mathbb{Q} \setminus \{0\}$, and let $\lambda_1 < \cdots < \lambda_s$ be distinct positive rationals. The $s \times s$ matrix*

$$M_{ik} := a_k \lambda_i^k \qquad (i \in \{1, \ldots, s\},\ k \in S)$$

is invertible.

Proof. Factor $M_{ik} = \lambda_i \cdot \lambda_i^{k-1} \cdot a_k$, so $M = \mathrm{diag}(\lambda_1, \ldots, \lambda_s) \cdot V \cdot \mathrm{diag}(a_k)_{k\in S}$, where V is the $s \times s$ matrix with entries $V_{ik} = \lambda_i^{k-1}$ (column exponents the distinct non-negative integers $\{k - 1 : k \in S\}$). Then $\det M = \big(\prod_i \lambda_i\big) \cdot \det V \cdot \big(\prod_{k\in S} a_k\big)$; since the λ_i are positive and the a_k are nonzero, it suffices to show $\det V \neq 0$.

Suppose, for contradiction, that $\det V = 0$. Then the columns of V are linearly dependent: there is a nonzero coefficient vector $(c_k)_{k\in S}$ with

$$\sum_{k\in S} c_k\, \lambda_i^{k-1} = 0 \qquad \text{for every } i = 1, \ldots, s.$$

Thus the function $q(t) := \sum_{k\in S} c_k\, t^{\,k-1}$ vanishes at the s distinct positive points $\lambda_1 < \cdots < \lambda_s$. The monomials $\{t^{k-1} : k \in S\}$ are linearly independent functions and not all c_k vanish, so q is not identically zero. Let $S' := \{k \in S : c_k \neq 0\}$ and $m := |S'|$, so that $1 \leq m \leq s$ and $q(t) = \sum_{k\in S'} c_k\, t^{\,k-1}$ is a real polynomial with exactly m nonzero terms and distinct exponents. By Theorem G.4, q has at most $m - 1 \leq s - 1$ positive real zeros counted with multiplicity, hence at most $s - 1$ *distinct* positive zeros. This contradicts q vanishing at the s distinct positive points $\lambda_1 < \cdots < \lambda_s$. Therefore $\det V \neq 0$, and so $\det M \neq 0$. □

G.3.5 Why the Route Stops Here

Theorems G.10 and G.11 show that the exact values of F_f on the rescaled instances satisfy a non-singular linear system in the moments M_k ($k \leq T$, $a_k \neq 0$), each equation up to an additive error at most $A\, 2^{-T}$; how this error propagates to the recovered moments depends on the conditioning of the system. We do not analyse the numerical conditioning of the resulting linear system: by Theorem G.12 the moments it would recover are computable exactly on products in any case, so no bound on that conditioning could turn this route into a hardness proof. A direct counting encoding for smooth integrands, in the spirit of Bhattacharyya et al. (2023) for S_{stat}, would bypass the moments altogether; whether one exists is open.

G.3.6 The Moments Are Easy on Products

Proposition G.12 (The moments factorize)**.** *For product distributions P, Q over $[\ell]^n$ and every integer $K \geq 1$,*

$$M_K(P,Q) = \prod_{i=1}^{n} \sum_{y \in [\ell]: q_i(y) > 0} \frac{p_i(y)^K}{q_i(y)^{K-1}},$$

so $M_K(P,Q)$ is computed exactly with $O(n\ell K)$ arithmetic operations on rationals of $O(n\ell K b_0 + n \log \ell)$ bits, in time polynomial in n, ℓ, K and b_0; in particular $\chi^2(P,Q) = M_2(P,Q) - 1$ is computed exactly in polynomial time when $P \ll Q$.

Proof. The Q-support is $\{x : q_i(x_i) > 0 \text{ for every } i\} = \prod_{i=1}^{n} S_i$ with $S_i := \{y : q_i(y) > 0\}$, and on it $P(x)^K/Q(x)^{K-1} = \prod_{i=1}^{n} p_i(x_i)^K/q_i(x_i)^{K-1}$. Hence

$$M_K(P,Q) = \sum_{x \in S_1 \times \cdots \times S_n} \prod_{i=1}^{n} \frac{p_i(x_i)^K}{q_i(x_i)^{K-1}} = \prod_{i=1}^{n} \sum_{y \in S_i} \frac{p_i(y)^K}{q_i(y)^{K-1}},$$

by expanding the product of sums, each of whose $\prod_i |S_i|$ terms is one term of the left-hand sum. Each factor is a sum of at most ℓ rationals, each formed with $O(K)$ multiplications and one division. Writing $p_i(y) = a/b$ and $q_i(y) = c/d$ with numerators and denominators of at most b_0 bits, each term $a^K d^{K-1}/(b^K c^{K-1})$ has numerator and denominator of at most $(2K-1)b_0$ bits; a sum of at most ℓ such terms over the product of their denominators has $O(\ell K b_0 + \log \ell)$ bits, and the product of the n factors has $O(n\ell K b_0 + n \log \ell)$ bits. For χ^2, the identity $\chi^2(P,Q) = M_2(P,Q) - 1$ when $P \ll Q$ is shown in Section F.15. □

Hence the moment-interpolation route of this subsection, which would at best reduce the computation of M_K to that of F_f, cannot establish hardness of F_f on products; the same factorization (multiplicative for χ^2, Hellinger and Rényi, additive for Kullback–Leibler) is why χ^2, Kullback–Leibler and the Hellinger and Rényi quantities are not approximation problems on products at all (Bhattacharyya et al., 2023).

G.4 Bounded Variants: A Partial Reduction

The bounded variants $f(r) = \min(\log(1+r), \beta)$ and $f(r) = \min(1 - e^{-r}, c)$ are intermediate between S_{stat}-style (piecewise linear with a cap on the integrand) and the smooth functionals (log-overlap, exp-overlap). For fixed β (resp. c), they admit a partial reduction.

Proposition G.13 (Inactive cap)**.** *Let $f(r) = \min(g(r), \beta)$ where g is admissible. For any product instance (P,Q) with $g(P(x)/Q(x)) \leq \beta$ for every x with $Q(x) > 0$ (i.e., the cap never fires), $F_f(P,Q) = F_g(P,Q)$.*

Proof. By hypothesis $g(P(x)/Q(x)) \leq \beta$ at every x, so $\min\big(g(P(x)/Q(x)), \beta\big) = g(P(x)/Q(x))$ for every x in the Q-support. The two integrands therefore agree pointwise on the support, and taking expectations under Q,

$$F_f(P,Q) = \mathop{\mathbf{E}}_{x \sim Q}[\min(g(P/Q), \beta)] = \mathop{\mathbf{E}}_{x \sim Q}[g(P/Q)] = F_g(P,Q). \qquad \square$$

Quantitative version. On products the largest likelihood ratio factorizes: the Q-support is $S_1 \times \cdots \times S_n$ with $S_i := \{y : q_i(y) > 0\}$, so $r_{\max} := \max_{x:Q(x)>0} P(x)/Q(x) = \prod_{i=1}^{n} \max_{y \in S_i} p_i(y)/q_i(y)$, and since g is non-decreasing (Theorem 2.7(d)) the cap never fires if and only if $g(r_{\max}) \leq \beta$; $r_{\max}$ is computed with $O(n\ell)$ rational operations. The condition is mathematical; an approximate evaluation oracle for g alone does not necessarily decide equality with β. Since $r_{\max} \leq 1/\tau^n$, a sufficient condition is $g(1/\tau^n) \leq \beta$; for $g = \log(1+r)$ it reads $\tau \geq (e^\beta - 1)^{-1/n}$, and for $g = 1 - e^{-r}$ and $c \in (0,1)$ it reads $\tau \geq (-\log(1-c))^{-1/n}$. The sufficient condition is restrictive (a marginal with two nonzero entries has smallest entry at most $1/2$, so it fails once 2^n exceeds $e^\beta - 1$, respectively $-\log(1-c)$), but the exact class $r_{\max} \leq g^{-1}(\beta)$ is not; for instance it contains every instance with $P = Q$ when $g(1) \leq \beta$.

This partial reduction therefore says: on instances where the cap does not fire, $F_f = F_g$, so any hardness of F_g on that instance class transfers to F_f; since hardness of F_g on products is open for every smooth g, this gives no unconditional hardness at present. A full reduction for fixed β would need to additionally handle the partially-saturating regime, where some atoms have $g(r) > \beta$ and others have $g(r) \leq \beta$; the saturated atoms contribute a constant $\beta \cdot Q(x)$ regardless of r, while the non-saturated atoms contribute $g(r) \cdot Q(x)$, and combining these in a hardness reduction requires distinguishing the two regimes from the oracle's output.

G.5 Summary

The functionals in Table 1 have the following exact-computation status on product distributions over $\{0,1\}^n$ with $\mathrm{poly}(n)$-bit rational marginals:

- S_{stat} ($f(r) = \min(r, 1)$): #P-hard (Bhattacharyya et al., 2023).
- Capped and truncated similarity S_{stat_c} ($f(r) = \min(r, c)$, any rational $c > 0$): #P-hard (Theorem G.5).
- Convex combinations $\alpha g + (1-\alpha)h$ with rational $\alpha \in (0, 1]$, g in the previous list, and F_h computable exactly in polynomial time on products (e.g. $h(r) = r$): #P-hard (Theorem G.8).
- Smooth concave functionals (log-, exp-, harmonic-, $\tanh$-, $\arctan$-overlap) hardness on products is open; the moment-interpolation strategy of Section G.3 cannot settle it, since the moments are computable exactly on products (Theorem G.12), and we are not aware of an alternative reduction.
- Square-root overlaps of Section F.14: hardness on products is open. The moment-interpolation strategy does not apply as stated, since neither has a power series in r at 0; for $2\sqrt{r}/(1+\sqrt{r}) = \sum_{k\geq 1} 2(-1)^{k-1} r^{k/2}$ ($r < 1$) the analogous route would recover the quantities $\sum_x P(x)^{k/2} Q(x)^{1-k/2}$, which factorize over the coordinates by the proof of Theorem G.12.
- Bounded variants ($\min(g, \beta)$ with smooth g): a partial reduction (Theorem G.13) shows that on instances where the cap does not fire, $F_{\min(g,\beta)} = F_g$, so any hardness of F_g on that class would transfer; since that is open, so is the exact complexity of these variants.

The takeaway is that multiplicative FPTAS guarantees of Theorem 4.1 hold uniformly on products for the entire family, while #P-hardness of exact computation is established unconditionally for three subclasses (S_{stat} itself, the capped and truncated variants S_{stat_c} for every rational $c > 0$, and convex combinations $\alpha g + (1-\alpha)h$ of these with any h whose F_h is computable exactly in polynomial time on products); for the smooth concave functionals, hardness on products remains open. We do not know of a poly-time exact algorithm for the smooth functionals or the bounded variants on products either, so the FPTAS results remain non-trivial in any case. For the multiplicative-approximation question on Bayes nets of in-degree 2, Theorem 5.1 resolves the entire family via a uniform NP-hardness reduction, independent of the open questions above.

H Open Problems and Future Work

This appendix collects open problems and directions for future work that extend the framework of Sections 4 and 5.

H.1 The Treewidth Threshold

Our hardness applies at in-degree 2, with P and Q on a common DAG, while Theorem 4.1 applies at in-degree 0 (products). In between, the answer depends on whether P and Q share a DAG. On different DAGs of in-degree 1, deciding $F > 0$ is already NP-complete (Theorem B.1). On a common DAG of in-degree 1, a forest, deciding $F > 0$ is in polynomial time (Theorem B.2), so any hardness there must come from an approximation gap rather than a zero test, and whether F admits an FPTAS on such pairs is open.

What extends to trees. The universal lower bound (Theorem 2.8) extends to tree-structured Bayes nets verbatim, with τ reinterpreted as the minimum nonzero CPT entry across all conditionals: if $F(P,Q) > 0$, there exists $z \in [\ell]^n$ with $P(z), Q(z) > 0$, and choosing $x^* = z$ gives $\mathbf{Pr}_P[X_i = x_i^* | X_{\Pi(i)} = x^*_{\Pi(i)}] \geq \tau$ and $\mathbf{Pr}_Q[\cdot] \geq \tau$ for every i. Hence $P(x^*), Q(x^*) \geq \tau^n$ and $F(P,Q) \geq f(1)\tau^n$ as before. NP-membership of "$F(P,Q) > 0$" also extends: the same poly-size witness is any atom z with $P(z), Q(z) > 0$, with $P(z), Q(z)$ verifiable in poly-time via CPT-entry lookups along a topological order.

What does not immediately extend. The algorithm of Section 4.1 relies on the factorization of the global ratio $R = R_1 \cdot R_2 \cdots R_n$ into independent per-coordinate ratios, which holds for products but not trees. On a tree, the conditional ratio at child i depends on the parent value, so the global ratio is a product of *coordinate-conditional* ratios. An adaptation via junction-tree-style message passing (Pearl, 1988) — maintaining, for each node, a sparsified distribution over per-subtree partial ratios conditioned on the value of the node's separator — seems plausible but is not immediate. The width of the messages must be controlled: each junction-tree message is over ℓ^w values where w is the width of the tree decomposition, so for trees ($w = 1$) the width is just ℓ.

Open question. Is there an FPTAS for F when P and Q are Bayes nets on a *common* DAG of in-degree 1, a forest, for every polynomial-time computable admissible f? On different DAGs of in-degree 1, no approximation within any multiplicative factor exists unless P = NP (Theorem B.1), while on a common forest the zero test is easy (Theorem B.2), so a negative answer would need an approximation gap.

H.2 Mixtures of Products

Mixtures $P = \sum_k w_k P^{(k)}$ where each $P^{(k)}$ is a product distribution arise frequently: Gaussian mixtures (after quantization), latent-variable models, hidden Markov chains with finite state space (after marginalizing the latent state), and naive-Bayes models with class-conditional independence given the class label.

Why the framework does not directly cover mixtures. The ratio $P(x)/Q(x) = \sum_k w_k P^{(k)}(x) / \sum_k w'_k Q^{(k)}(x)$ does not factorize across coordinates: the mixture introduces a sum at the top level, and ratios of sums are not products of ratios. Two natural strategies:

- *Latent-variable expansion.* Treat the latent mixture index k as a hidden coordinate. The resulting distribution on $\{1, \ldots, K\} \times [\ell]^n$ (where K is the number of mixture components) is a Bayes net with a single latent root and product children given the root. Estimating F on the original $[\ell]^n$ marginal requires marginalizing the latent, which generally has no closed form.
- *Component-wise computation.* For mixtures with matching weights $w_k = w'_k$, S_{stat} is jointly concave in (P,Q) (equivalently, d_{TV} is jointly convex), giving $S_{\mathrm{stat}}(\sum_k w_k P^{(k)}, \sum_k w_k Q^{(k)}) \geq \sum_k w_k S_{\mathrm{stat}}(P^{(k)}, Q^{(k)})$ by Jensen. Each $S_{\mathrm{stat}}(P^{(k)}, Q^{(k)})$ is between products, hence FPTAS-approximable by Theorem 4.1, so the RHS gives a poly-time approximable lower bound on the LHS. But this is not a multiplicative approximation of the LHS, since the gap between the LHS and RHS can be unbounded relative to the LHS. For mismatched weights ($w_k \neq w'_k$), the same argument with $m_k := \min(w_k, w'_k)$ gives the weaker bound $S_{\mathrm{stat}}(\sum_k w_k P^{(k)}, \sum_k w'_k Q^{(k)}) \geq \sum_k m_k S_{\mathrm{stat}}(P^{(k)}, Q^{(k)})$, since $\min$ is non-decreasing in each argument and $\min(a+b, c+d) \geq \min(a,c) + \min(b,d)$; a multiplicative approximation would require coupling the mixture indices explicitly.

Open question. Is S_{stat} between mixtures of two product distributions (each with K components) NP-hard to multiplicatively approximate when K is polynomially bounded? When K is constant? Tractability of mixtures of products under various structural assumptions is an open problem.

H.3 Lower Bounds and Tightness

Algorithmic tightness. For S_{stat}, Theorem 4.1 uses $O(n\ell m \log m)$ arithmetic operations with $m = O(n^2 \log(1/\tau)/\varepsilon + n/\varepsilon)$ slices. The n^3 exponent comes from: n coordinates (outer loop), and a slice count quadratic in n, because the geometric slices must cover the range $[\gamma, B]$ of ratio $2^{\Theta(n \log(1/\tau))}$ at resolution $1 + \Theta(\varepsilon/n)$. Matching $\widetilde{\Omega}(n^2)$ or $\widetilde{\Omega}(n^3)$ unconditional lower bounds, or fine-grained conditional lower bounds (from SETH or related hypotheses), would clarify whether the cubic dependence is optimal. We conjecture that the dependence on n can be reduced to $\widetilde{O}(n^2)$ via a more careful merging schedule across coordinates, but have not verified this.

For a fixed f, $f(1) > 0$ is a constant and a fixed rational $\varphi \in (0, f(1)]$ can be built into the algorithm, so Theorem 4.1 already runs in time polynomial in n, ℓ, b_0 and $1/\varepsilon$. A separate question arises only when f is part of the input, uniformly over a family: whether the $\log(1 + 1/\varphi)$ term in the oracle precision, which reflects how small $f(1)$ may be, is necessary.

Hardness tightness. Theorem 5.1 shows NP-hardness on Bayes nets of in-degree ≥ 2 (with the reduction producing in-degree-exactly-2 instances). The same construction extends to:

- **#P-hardness of exact computation**, via the exact identity $F(P, Q) = f(1) \cdot |\mathrm{Sol}(\phi)|/2^n$ in the proof of Theorem 5.1, which would yield exact $|\mathrm{Sol}(\phi)|$ from an exact-F oracle.
- **ETH-based exponential lower bound**: under the Exponential Time Hypothesis (Impagliazzo & Paturi, 2001), no $2^{o(\nu)}$-time algorithm decides satisfiability of 3-CNF formulas on ν variables. Our reduction maps such a formula to a pair of Bayes nets whose parentless nodes are exactly the ν inputs, in time polynomial in the formula, so under ETH no algorithm approximates F within any multiplicative factor in time $2^{o(\nu)} \cdot \mathrm{poly}(N)$ on Bayes nets of size N with ν parentless nodes.
- **NP-hardness of any constant-factor multiplicative approximation**: an α-approximation for any constant $\alpha \geq 1$ distinguishes $F = 0$ from $F > 0$, hence solves SAT. This rules out any constant-factor approximation under $\mathsf{P} \neq \mathsf{NP}$ — strictly stronger than APX-hardness, which rules out only a PTAS (and is compatible with the existence of constant-factor approximations at sufficiently large constants).

Additive error. Our reduction yields $F = 0$ for unsatisfiable inputs and $F = f(1)|\mathrm{Sol}(\phi)|/2^n > 0$ for satisfiable ones, which is as small as $f(1)/2^n$ when ϕ has a single satisfying assignment. For bounded f no reduction can do better under additive error: as noted in Section 5.1, $F(P, Q) = \mathbf{E}_{x \sim Q}[f(P(x)/Q(x))]$ is the mean of a bounded quantity that can be sampled and evaluated exactly on any Bayes net, so it has an additive randomized polynomial-time approximation, and NP-hardness of an additive-gap version would imply $\mathsf{NP} \subseteq \mathsf{BPP}$. This includes S_{stat} and every bounded entry of Table 1, and also log-overlap: although $\log(1+r)$ is unbounded, $Q(x) \geq \tau^n$ whenever $Q(x) > 0$ (τ the smallest nonzero CPT entry), so $\log(1+P(x)/Q(x)) \leq n \log(1/\tau) + \log 2$, the sampled quantity has a range polynomial in the input size, and Hoeffding's inequality still applies. For general unbounded f the additive question is open.

H.4 Identifying Further #P-Hard f with FPTASes on Both Sides

The pair $(S_{\mathrm{stat}}, d_{\mathrm{TV}}) = (\min(r, 1), \frac{1}{2}|r - 1|)$ exhibits the property that both functionals admit FPTASes on products, S_{stat} via our Theorem 4.1 and d_{TV} via Feng et al. (2024), although neither transfers to the other. Their integrands sum to the affine function $\min(r, 1) + \frac{1}{2}|r-1| = (1+r)/2$, so $F_{\min(r,1)} + F_{\frac{1}{2}|r-1|} = (1 + P(\{Q > 0\}))/2$, which is 1 when $P \ll Q$; this matches $S_{\mathrm{stat}} + d_{\mathrm{TV}} = 1$, and in general d_{TV} exceeds $F_{\frac{1}{2}|r-1|}(P, Q)$ by $\frac{1}{2}P(\{Q = 0\})$. This is despite each being individually #P-hard (Bhattacharyya et al., 2023).

Open question. Identify further pairs (F_f, F_g) of #P-hard functionals for which $F_f + F_g$ has a simple closed form (e.g., constant) and both F_f, F_g individually admit FPTASes on products. The natural candidates $f(r) = 1 - e^{-r}$ and $f(r) = \tanh r$ do not exhibit the same cancellation difficulty: by Jensen's inequality and $\mathbf{E}_{x \sim Q}[P/Q] \leq 1$, $\mathbf{E}_{x \sim Q}\left[e^{-P/Q}\right] \geq e^{-1}$ and $\mathbf{E}_{x \sim Q}[1 - \tanh(P/Q)] \geq 1 - \tanh 1$, so each complementary functional is bounded below by a constant, and a $(1 + \varepsilon')$-

approximation of F_f yields one of $1 - F_f$ with relative error $O(\varepsilon')$. The structural question is when multiplicative tractability transfers between related functionals; a characterization is open.

H.5 Undirected Graphical Models

We treat directed graphical models (Bayes nets), where each atom probability factorizes as $P(x) = \prod_i \mathbf{Pr}_P[X_i = x_i \mid X_{\Pi(i)} = x_{\Pi(i)}]$ via locally normalized CPTs. For undirected graphical models (MRFs / Gibbs / Ising distributions), $P(x) = \frac{1}{Z_P} \prod_c \psi_c^P(x_c)$ where ψ_c^P are clique potentials and $Z_P = \sum_x \prod_c \psi_c^P(x_c)$ is the partition function. Evaluating $P(x)$ at a single point already requires knowing Z_P, which is #P-hard in general but admits FPTASes in tractable regimes (high-temperature / correlation-decay, bounded treewidth).

Why F_f on MRFs is harder than on Bayes nets. For two MRFs P, Q on the same graph, the per-atom ratio is $P(x)/Q(x) = (Z_Q/Z_P) \cdot \prod_c (\psi_c^P(x_c)/\psi_c^Q(x_c))$, with the partition-function ratio Z_Q/Z_P appearing as an unknown global rescaling. A direct extension of our likelihood-ratio algorithm would thus need to control Z_Q/Z_P, a separate (and in general delicate) approximation problem. Feng & Fu (2026) study f-divergences between Ising models, while Feng et al. (2025) studies d_{TV} between spin systems; their algorithms use sampling and approximate counting under model-specific conditions.

Open question. In regimes where Z_P, Z_Q are FPTAS-approximable (e.g., high-temperature / correlation-decay regimes for ferromagnetic Ising models, or bounded-treewidth MRFs), is there an FPTAS for $F_f(P, Q)$ for every admissible f? A natural starting point is bounded-treewidth MRFs, where junction-tree message passing computes $P(x)$ exactly and reduces F_f to a structured functional over factorized messages.

H.6 Multi-Distribution Generalizations and K-Class Bayes Risk

The class-conditional Bayes-risk identity $R^* = S_{\mathrm{stat}}(\alpha_0 P_0, \alpha_1 P_1)$ that we leverage (Section D) is intrinsically binary. The natural multi-class analogue is the K-class Bayes-optimal classification error

$$R_K^* = 1 - \sum_x \max_k \alpha_k P_k(x) = \sum_x \Big(\sum_j \alpha_j P_j(x) - \max_k \alpha_k P_k(x) \Big)$$
$$= \sum_x g_K\big(\alpha_0 P_0(x), \ldots, \alpha_{K-1} P_{K-1}(x)\big),$$

where $\alpha_0, \ldots, \alpha_{K-1}$ are class priors with $\sum_k \alpha_k = 1$, $P_0, \ldots, P_{K-1}$ are class-conditional distributions, and $g_K(a_0, \ldots, a_{K-1}) = \sum_j a_j - \max_k a_k$. The integrand g_K is a K-ary functional that does not fit the binary $F_f(P, Q) = \mathbf{E}_Q[f(P/Q)]$ template; for $K = 2$ we recover $g_2(a, b) = \min(a, b)$ and $R_2^* = S_{\mathrm{stat}}(\alpha_0 P_0, \alpha_1 P_1)$ (Section D).

What generalizes. The slice-decomposition apparatus extends in principle to K-ary integrands: instead of tracking a one-dimensional partial-product ratio $R_1 \cdots R_i \in [0, B]$, one tracks a K-tuple of partial products $(\alpha_0 P_0(x_1, \ldots, x_i), \ldots, \alpha_{K-1} P_{K-1}(x_1, \ldots, x_i))$, equivalently a vector in the $(K-1)$-simplex paired with the total mass. Sparsification by joint slices is well-defined; the question is whether the support stays controllable as coordinates accumulate.

Open question. Is there an FPTAS for R_K^* on K-tuples of product distributions over $[\ell]^n$, with running time polynomial in $n, \ell, 1/\varepsilon, K$, and $\log(1/\tau)$? More broadly, what concavity / monotonicity / lower-bound conditions on a K-ary integrand $g\colon \mathbb{R}_{\geq 0}^K \to \mathbb{R}_{\geq 0}$ suffice for an FPTAS for $\sum_x g(\alpha_0 P_0(x), \ldots, \alpha_{K-1} P_{K-1}(x))$ on product class-conditionals? The Bayes-risk integrand g_K above is concave (as the difference of a linear functional and a convex one) and symmetric in its arguments; identifying the right generalization of admissibility is itself part of the question.