

SHARPER MAXIMUM-ENTROPY CERTIFICATES

ABSTRACT. We study the maximum entropy of a finite distribution subject to prescribed one-coordinate marginals, a quantity used in combination-loss analyses of matrix multiplication. From any additive approximation g to $\log_2 y$ for a feasible positive distribution y , we derive an exact convex-dual upper certificate whose excess over $H(y)$ is $D_{\text{KL},2}(y||q_g)$, together with a sharp residual-only bound that is quadratic near zero rather than linear. We also give an independent linear-programming certificate, handle marginal constraints that force zero coordinates, and prove the identity $H^{\max} - H(\rho) = D_{\text{KL},2}(\rho||p^*)$ for the unique maximum-entropy coupling p^* . A first-variation formula shows that the penalty gradient is $\log_2(\rho/p^*)$, so differentiation through the entropy solver is unnecessary. Finally, we give directed-rounding certificates and explain their monotone substitution in matrix-multiplication feasibility proofs.

1. INTRODUCTION

Let D be a finite set of triples and let ρ be a probability distribution on D . The maximum entropy among distributions on D having the same one-coordinate marginals as ρ is a standard finite-dimensional convex optimization problem. It also appears repeatedly in the combination-loss framework for upper bounds on the matrix multiplication exponent. In the recent optimization of Dupont et al. [2], a positive feasible distribution y , an additive function g , and a uniform residual estimate $|\log_2 y - g| \leq \varepsilon$ are used to certify the upper bound $H^{\max} \leq H(y) + 2\varepsilon$. The factor 2ε is obtained by applying a tangent inequality for entropy and then bounding an inner product by an ℓ_∞ - ℓ_1 estimate.

We replace that estimate in two complementary ways. The first is the ordinary log-partition dual of the marginal entropy problem. The same additive function g gives an upper bound whose exact excess over $H(y)$ is a Kullback–Leibler divergence. If $r = g - \log_2 y$ has range R , this excess is at most a sharp scalar function $\psi(R)$ satisfying $\psi(R) < R$ for $R > 0$ and $\psi(R) \leq (\ln 2)R^2/8$. Thus the uniform residual certificate improves from 2ε to $\psi(2\varepsilon)$, which is $O(\varepsilon^2)$ near zero. The second certificate retains the tangent argument but computes its exact worst-case value as a linear program over the marginal fiber. Taking the smaller of the two bounds is always valid.

We also resolve several technical points needed by a rigorous verifier. Marginal constraints may force some cells to have zero mass, so a positive feasible point need not exist on the originally declared support; we identify and remove precisely those cells. We prove that the entropy penalty is a KL divergence and derive its first variation without differentiating the maximum-entropy solution map. Finally, we state rational, directed-rounding versions of both certificates. These results strengthen the entropy component of the proof architecture in [2]; a new numerical value of the matrix multiplication exponent would additionally require the authors’ rational parameter file and an independent execution of the complete verifier, so no unsupported decimal improvement is asserted here.

2. ENTROPY FIBERS

Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be finite sets and let $D \subseteq \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ be nonempty. Write $\Delta(D)$ for the probability simplex on D , use $0 \log_2 0 = 0$, and define $H(p) = -\sum_{a \in D} p(a) \log_2 p(a)$. For distributions p, q with $q > 0$ on $\text{supp}(p)$, let $D_{\text{KL},2}(p||q) = \sum_a p(a) \log_2(p(a)/q(a))$. For $W \in \{X, Y, Z\}$, the W -marginal of p is denoted by p_W . Given $\rho \in \Delta(D)$, set

$$\mathcal{P}_D(\rho) = \{p \in \Delta(D) : p_X = \rho_X, p_Y = \rho_Y, p_Z = \rho_Z\}, \quad H_D^{\max}(\rho) = \max_{p \in \mathcal{P}_D(\rho)} H(p).$$

2020 *Mathematics Subject Classification.* 94A17, 90C25, 68Q25.

Key words and phrases. maximum entropy, convex duality, entropy certificate, matrix multiplication exponent, combination loss.

The maximum exists because the fiber is a nonempty compact polytope. A function $g : D \rightarrow \mathbb{R}$ is called additive if there are functions g_X, g_Y, g_Z such that $g(a) = g_X(a_X) + g_Y(a_Y) + g_Z(a_Z)$; constants are included by taking one of the three component functions constant.

It is useful to encode the marginals linearly. Let $A : \mathbb{R}^D \rightarrow \mathbb{R}^X \oplus \mathbb{R}^Y \oplus \mathbb{R}^Z$ send p to (p_X, p_Y, p_Z) . Its adjoint is

$$(A^*u)(a) = u_X(a_X) + u_Y(a_Y) + u_Z(a_Z),$$

so the space of additive functions is $\mathcal{A}(D) = \text{im } A^*$. Define the effective support of the fiber by

$$D_\rho = \bigcup_{p \in \mathcal{P}_D(\rho)} \text{supp}(p).$$

Every feasible distribution vanishes on $D \setminus D_\rho$, and averaging one feasible witness for each point of D_ρ produces a member of the fiber that is positive on all of D_ρ . Consequently, replacing D by D_ρ changes neither the fiber nor its maximum entropy.

Theorem 2.1 (Structure, duality, and the entropy penalty). *After replacing D by D_ρ , there is a unique maximizer $p^* \in \mathcal{P}_D(\rho)$, it satisfies $p^*(a) > 0$ for every $a \in D$, and $\log_2 p^*$ is additive. Moreover,*

$$(1) \quad H_D^{\max}(\rho) = \min_{g \in \mathcal{A}(D)} \left\{ \log_2 \sum_{a \in D} 2^{g(a)} - \sum_{a \in D} \rho(a)g(a) \right\},$$

and the minimum is attained by $g = \log_2 p^*$. Finally,

$$(2) \quad H_D^{\max}(\rho) - H(\rho) = D_{\text{KL},2}(\rho \| p^*).$$

Proof. Strict concavity of entropy gives uniqueness. Let y be a feasible distribution positive on D . If a maximizer p^* vanished at some a , then the right derivative of $H((1-t)p^* + ty)$ at $t = 0$ would be $+\infty$, because the term at a contains $-ty(a) \log_2(ty(a))$, whereas the changes at coordinates where p^* is positive are only $O(t)$. This contradicts maximality, so $p^* > 0$.

The feasible two-sided directions at p^* are the vectors in $\ker A$. The derivative of entropy at p^* is the vector $-\log_2 p^* - (\log_2 e)\mathbf{1}$, and its inner product with every vector in $\ker A$ is zero. Hence it belongs to $(\ker A)^\perp = \text{im } A^*$, proving that $\log_2 p^*$ is additive.

For additive g , put $Z_g = \sum_a 2^{g(a)}$ and $q_g(a) = 2^{g(a)}/Z_g$. If $p \in \mathcal{P}_D(\rho)$, additivity gives $\sum_a p(a)g(a) = \sum_a \rho(a)g(a)$. Nonnegativity of $D_{\text{KL},2}(p \| q_g)$ therefore yields, in one rearrangement,

$$H(p) \leq \log_2 Z_g - \sum_a \rho(a)g(a).$$

Taking the maximum over p and then the infimum over g proves one direction of (1). For $g = \log_2 p^*$ one has $Z_g = 1$, and the expectation of g under ρ equals its expectation under p^* because g is additive. The resulting dual value is $-\sum_a p^*(a) \log_2 p^*(a) = H(p^*)$, proving equality and attainment.

The same marginal-expectation identity gives $\sum_a \rho(a) \log_2 p^*(a) = \sum_a p^*(a) \log_2 p^*(a) = -H(p^*)$. Thus $D_{\text{KL},2}(\rho \| p^*) = -H(\rho) + H(p^*)$, which is (2). $\square$

3. SHARPER CERTIFICATES

Fix a positive $y \in \mathcal{P}_D(\rho)$ and an additive function g . Let

$$r(a) = g(a) - \log_2 y(a), \quad R = \text{osc}_D(r) := \max_{a \in D} r(a) - \min_{a \in D} r(a),$$

and let $q_g(a) = 2^{g(a)} / \sum_b 2^{g(b)}$. The next theorem gives two upper certificates. The first is exponential but smooth; the second is a linear program.

Theorem 3.1 (Dual and residual certificates). *With the notation above, define*

$$(3) \quad \mathfrak{D}(y, g) = \log_2 \sum_{a \in D} y(a) 2^{r(a)} - \sum_{a \in D} y(a) r(a) = D_{\text{KL},2}(y \| q_g),$$

$$(4) \quad \Gamma_D(y) = \max_{p \in \mathcal{P}_D(\rho)} \sum_{a \in D} r(a) (p(a) - y(a)).$$

Then $\Gamma_D(y)$ is independent of the chosen additive function g , and

$$(5) \quad H(y) \leq H_D^{\max}(\rho) \leq H(y) + \min\{\mathfrak{D}(y, g), \Gamma_D(y)\}.$$

Furthermore, $0 \leq \Gamma_D(y) \leq R$, and

$$(6) \quad H(y) + \Gamma_D(y) = \min_u \left\{ \sum_{W \in \{X, Y, Z\}} \sum_w \rho_W(w) u_W(w) : \sum_W u_W(a_W) \geq -\log_2 y(a) \text{ for all } a \in D \right\}.$$

$\Gamma_D(y) = 0$ if and only if $y = p^*$. Moreover, $\mathfrak{D}(y, g) = 0$ if and only if $y = q_g$; this implies $y = p^*$, and when $y = p^*$ the choice $g = \log_2 y$ makes $\mathfrak{D}(y, g) = 0$.

Proof. The dual bound (1) gives $H_D^{\max}(\rho) \leq \log_2 \sum_a 2^{g(a)} - \sum_a y(a) g(a)$, since y and ρ have the same marginals. Substituting $g = \log_2 y + r$ shows that this upper bound equals $H(y) + \mathfrak{D}(y, g)$; the equality with $D_{\text{KL},2}(y \| q_g)$ follows directly from the definition of q_g .

Concavity of entropy at the positive point y gives $H(p) - H(y) \leq \langle -\log_2 y - (\log_2 e) \mathbf{1}, p - y \rangle$. The constant term vanishes, and $\langle g, p - y \rangle = 0$ by additivity and equality of marginals, so the right-hand side is $\langle r, p - y \rangle$. Maximizing over the fiber proves the second upper bound in (5). It also shows that $\Gamma_D(y)$ does not depend on g . Since y is feasible, $\Gamma_D(y) \geq 0$, while for every feasible p one has $\mathbb{E}_p r - \mathbb{E}_y r \leq \max r - \min r = R$.

For (6), put $c(a) = -\log_2 y(a)$. Then $H(y) + \Gamma_D(y) = \max\{\langle c, p \rangle : p \geq 0, Ap = A\rho\}$. Ordinary finite-dimensional linear-programming duality gives the displayed minimum, because the adjoint constraint is $A^*u \geq c$. Finally, $\Gamma_D(y) = 0$ is precisely the first-order optimality condition for maximizing the differentiable concave function H over the fiber, hence is equivalent to $y = p^*$. Also $\mathfrak{D}(y, g) = 0$ means $y = q_g$; a feasible distribution with additive logarithm satisfies the same first-order condition and is therefore p^* . Conversely, when $y = p^*$, taking $g = \log_2 y$ gives $q_g = y$ and zero dual gap. $\square$

The dual gap has a sharp bound depending only on the residual range. Define $\psi(0) = 0$, and for $R > 0$ let, with $s = R \ln 2$,

$$(7) \quad \psi(R) = \frac{1}{\ln 2} \left[\ln \left(\frac{e^s - 1}{s} \right) - 1 + \frac{s}{e^s - 1} \right].$$

Lemma 3.2 (Sharp residual-range bound). *If a real random variable V is supported in an interval of length R , then*

$$(8) \quad \log_2 \mathbb{E}[2^V] - \mathbb{E}[V] \leq \psi(R) \leq \min \left\{ R, \frac{\ln 2}{8} R^2 \right\}.$$

The first inequality is sharp among all probability laws supported in an interval of length R .

Proof. The expression is invariant under adding a constant to V , so for $R > 0$ write $V = RU$ with $0 \leq U \leq 1$, set $p = \mathbb{E}U$, and put $s = R \ln 2$. Convexity of the exponential on $[0, 1]$ gives $e^{sU} \leq 1 + U(e^s - 1)$, hence

$$\ln \mathbb{E} e^{sU} - s \mathbb{E}U \leq f_s(p) := \ln(1 + p(e^s - 1)) - sp.$$

The unique maximizer is $p_s = (e^s - 1 - s)/(s(e^s - 1))$, and substituting it gives the numerator in (7). Equality is attained by a Bernoulli random variable supported on $\{0, 1\}$ with mean p_s , proving sharpness.

The bound by R follows immediately from $\log_2 \mathbb{E} 2^V \leq \max V$ and $\mathbb{E}V \geq \min V$. For the quadratic bound, fix p and write $h_p(s) = \ln(1 - p + pe^s) - ps$. Then $h_p(0) = h'_p(0) = 0$ and

$h_p''(s) = q_s(1 - q_s) \leq 1/4$, where $q_s = pe^s/(1 - p + pe^s)$. Integrating twice yields $h_p(s) \leq s^2/8$; division by $\ln 2$ proves the last inequality in (8). $\square$

Corollary 3.3 (Uniform residual certificate). *Suppose $|\log_2 y(a) - g(a)| \leq \varepsilon$ for every $a \in D$. Then*

$$(9) \quad H_D^{\max}(\rho) \leq H(y) + \mathfrak{D}(y, g) \leq H(y) + \psi(2\varepsilon) \leq H(y) + \min \left\{ 2\varepsilon, \frac{\ln 2}{2} \varepsilon^2 \right\}.$$

For every $\varepsilon > 0$, the exact residual-only term satisfies $\psi(2\varepsilon) < 2\varepsilon$.

Proof. The residual range is at most 2ε , so Theorem 3.1 and Lemma 3.2 apply. Strict inequality follows because the maximizing Bernoulli law in Lemma 3.2 has positive mass at both endpoints and its value is strictly smaller than the interval length. $\square$

Remark 3.4. The exact quantity $\mathfrak{D}(y, g)$ is never larger than the residual-only bound $\psi(R)$ and can be substantially smaller because it uses the actual weights $y(a)$ and residuals $r(a)$. The LP quantity $\Gamma_D(y)$ uses the geometry of the marginal fiber and may be smaller still. There is no universal ordering between $\mathfrak{D}(y, g)$ and $\Gamma_D(y)$, so a verifier should compute both when practical.

4. DIFFERENTIATION AND EXACT VERIFICATION

The dual problem also yields a stable way to compute p^* and shows why no derivative of the entropy solver is needed. Write $b = A\rho$ and, for a potential vector u , define

$$F_b(u) = \log_2 \sum_{a \in D} 2^{(A^*u)(a)} - \langle u, b \rangle, \quad q_u(a) = \frac{2^{(A^*u)(a)}}{\sum_c 2^{(A^*u)(c)}}.$$

Direct differentiation gives

$$(10) \quad \nabla F_b(u) = Aq_u - b, \quad \nabla^2 F_b(u) = (\ln 2)A(\text{diag}(q_u) - q_u q_u^T)A^*.$$

The Hessian is positive semidefinite. Its quadratic form vanishes exactly when A^*v is constant on D , which is the gauge freedom that leaves q_u unchanged. On any fixed complementary gauge slice the Hessian is positive definite. Since a positive distribution maps under A to the relative interior of the marginal polytope $A\Delta(D)$, the inverse function theorem applied on such a slice shows that p^* depends smoothly on the marginals throughout this relative interior.

Theorem 4.1 (First variation of the penalty). *Let $\rho(t)$ be a C^1 path in the positive simplex $\Delta^\circ(D)$, and let $p^*(t)$ be the maximum-entropy distribution with the same marginals as $\rho(t)$. Then*

$$(11) \quad \frac{d}{dt} (H(p^*(t)) - H(\rho(t))) = \sum_{a \in D} \dot{\rho}(t, a) \log_2 \frac{\rho(t, a)}{p^*(t, a)}.$$

Thus the gradient of the entropy penalty on the simplex tangent space is represented by $\log_2(\rho/p^*)$.

Proof. The smoothness discussion above permits differentiation. The vector $\nabla H(p^*) = -\log_2 p^* - (\log_2 e)\mathbf{1}$ is additive by Theorem 2.1. Since $p^*(t)$ and $\rho(t)$ have identical marginals for every t , their derivatives also have identical marginals; hence the inner products of $\nabla H(p^*)$ with p^* and with $\dot{\rho}$ agree. Subtracting the derivative of $H(\rho)$ and canceling the constant terms gives (11) in one line. $\square$

Equation (11) is an envelope identity: an implementation needs the value of p^* but not the Jacobian of the map $\rho \mapsto p^*$. In particular, a Newton or Newton–CG solve of the convex dual, using (10) after gauge fixing, can replace differentiation through a finite number of Sinkhorn iterations.

For a computer-assisted proof, all comparisons must be directed in the safe direction. The following two formats reduce verification to rational inequalities plus certified enclosures of elementary transcendental functions.

Proposition 4.2 (Directed-rounding certificates). *Let $b = A\rho$ have rational coordinates.*

- (1) Suppose u has rational coordinates and $L^+ \in \mathbb{Q}$ satisfies $L^+ \geq \log_2 \sum_{a \in D} 2^{(A^*u)(a)}$. Then

$$H_D^{\max}(\rho) \leq L^+ - \langle u, b \rangle.$$

- (2) Suppose $y \in \mathcal{P}_D(\rho)$ is positive and rational, rational numbers c_a^+ satisfy $c_a^+ \geq -\log_2 y(a)$, and a rational potential vector v satisfies $(A^*v)(a) \geq c_a^+$ for every a . Then

$$H_D^{\max}(\rho) \leq \langle v, b \rangle.$$

Proof. The first statement is the dual bound (1) with an upward-rounded log-partition value. For the second, concavity gives the cross-entropy inequality $H(p) \leq -\sum_a p(a) \log_2 y(a)$ for every feasible p . The assumed componentwise majorant and additivity then give $H(p) \leq \langle A^*v, p \rangle = \langle v, Ap \rangle = \langle v, b \rangle$. $\square$

For the first format, a verifier may upper-bound each $2^{(A^*u)(a)}$, sum the rational upper bounds, and then upper-bound the final logarithm. For the second, it need only lower-bound each $\log_2 y(a)$ and check the additive majorant by rational arithmetic. Neither format assumes that the numerical maximum-entropy iterate is exact, and the first does not require a positive feasible distribution at all. If the declared support contains forced-zero cells, one may either perform the effective-support reduction from Section 2 or retain the larger support and accept a potentially looser but still valid dual bound.

5. CONSEQUENCES FOR COMBINATION-LOSS BOUNDS

The entropy penalty used in combination-loss analysis is exactly the left-hand side of (2). In the notation of [2], each occurrence of $H_D^{\max}(\rho)$ is replaced by a certified upper bound, and the resulting penalty is subtracted from a retained-exponent expression. The same data used there—a feasible positive y , additive multipliers defining g , and a residual radius ε —therefore certifies not merely $H(y) + 2\varepsilon$, but the smaller quantity

$$U_{\text{new}} = H(y) + \min\{\mathfrak{D}(y, g), \Gamma_D(y)\} \leq H(y) + \psi(2\varepsilon) < H(y) + 2\varepsilon$$

whenever $\varepsilon > 0$. The exponential certificate can be evaluated directly from the multipliers, while the residual certificate is an LP on the finite support. Both admit the rational forms of Proposition 4.2.

This substitution is monotone. Any expression assembled from constants, addition, nonnegative scalar multiplication, and minima, with each maximum-entropy term occurring only with a negative sign, cannot decrease when its entropy upper bounds are lowered. Hence a parameter point feasible under the older 2ε certificates remains feasible, with weakly larger proof margin, under the new certificates. In a feasibility inequality of the form $E(x) + M(x)\Omega \geq C$ with $M(x) > 0$, the exponent attached to a fixed parameter point is $\Omega(x) = (C - E(x))/M(x)$; if the strengthened certificates increase the certified retained exponent from E_{old} to E_{new} , then the certified value improves by exactly $(E_{\text{new}} - E_{\text{old}})/M(x)$.

For numerical search, one may fix a target $\bar{\omega}$, maximize $E(x) + \bar{\omega}M(x) - C$, and lower $\bar{\omega}$ by continuation; hard minima are handled by epigraph variables $t \leq z_i(x)$. A proof-aligned implementation solves the entropy subproblems in their convex duals, uses Theorem 4.1 for gradients, and finishes with Proposition 4.2. The strengthening proved here is unconditional, but any further decimal bound on ω still requires independent verification of the complete rational solution and every non-entropy term.

REFERENCES

- [1] J. Alman, R. Duan, V. Vassilevska Williams, Y. Xu, Z. Xu, and R. Zhou, *More asymmetry yields faster matrix multiplication*, in Proceedings of the 2025 Annual ACM–SIAM Symposium on Discrete Algorithms, SIAM, 2025, pp. 2005–2039.
- [2] E. Dupont, M. Eisenberger, B. Kozlovskii, A. Mehrabian, F. J. R. Ruiz, A. See, R. Zhou, J. Alman, V. Vassilevska Williams, and M. Balog, *Improving the matrix multiplication exponent with modern optimization and AlphaEvolve*, arXiv:2608.16884v1, 2026.