DIRICHLET FOLLOW-THE-LEADER CLOSES THE GAP IN SIMULTANEOUS MULTICLASS U-CALIBRATION

Pahan Dewasurendra
Johns Hopkins University

ABSTRACT

Can one forecaster attain the optimal regret rate for every bounded proper loss and also adapt to every smooth proper loss? Recent work answered this up to a dimension gap. Its self-concordant perturbation gives roughly $K^{5/4}\sqrt{T}$ worst-case regret and incurs an additional $\beta\sqrt{K}\log K$ for β-smooth losses. We close both gaps with a one-line forecaster. After observing class counts c_{t-1}, draw the next prediction from $\mathrm{Dir}(c_{t-1})$, on the face of classes seen so far. This is a fresh Bayesian bootstrap of the outcomes. The analysis rests on an exact identity: averaging any bounded proper loss under $\mathrm{Dir}(\alpha)$ equals a discrete derivative of its Dirichlet-averaged Bayes risk. The identity makes the be-the-perturbed-leader term telescope to a nonpositive Jensen gap. A one-count likelihood ratio then bounds stability by the inverse square root of that class's count. The resulting single, horizon-free algorithm satisfies

$$\sup_{\ell} \mathbb{E}\,\mathrm{Reg}_\ell \le 4\sqrt{S_T T} \le 4\sqrt{KT},$$

$$\mathbb{E}\,\mathrm{Reg}_\ell \le \tfrac{5}{2}\beta(1+\log T) \quad \text{for every } \beta\text{-smooth proper } \ell.$$

Here S_T is the number of observed classes. Known lower bounds show that both rates are optimal in their nontrivial regimes. The proof covers nondifferentiable losses and changes of the active simplex face.

1 INTRODUCTION

A probabilistic forecast is often consumed by an agent whose downstream utility is unknown to the forecaster. Proper losses capture this uncertainty: each proper loss corresponds to a way of valuing probabilistic predictions, and truthful prediction minimizes expected loss. U-calibration asks for one online sequence of forecasts with low regret under every bounded proper loss (Kleinberg et al., 2023). This requirement is substantially stronger than optimizing a fixed score such as Brier loss.

For K outcomes, Luo et al. (2024) established the optimal $\Theta(\sqrt{KT})$ worst-case pseudo-U-calibration rate. Their optimal forecaster does not adapt to easier scores. In particular, it can incur $\Omega(\sqrt{T})$ regret on squared loss even though Follow-the-Leader (FTL) has logarithmic regret. Very recent work showed that this conflict is not fundamental (Frongillo et al., 2026). A carefully shaped self-concordant perturbation simultaneously gives logarithmic regret on smooth scores and sublinear regret on all proper scores. Its bounds leave two linked gaps. The general rate is $\widetilde{O}(K^{5/4}\sqrt{T})$ rather than $O(\sqrt{KT})$, and the smooth rate contains an additive $O(\beta\sqrt{K}\log K)$ term. The paper explicitly asks whether simultaneous optimality is possible.

We answer this question affirmatively. Let $c_t = \sum_{s\le t} y_s$ be the vector of outcome counts. At time $t \ge 2$, independently sample

$$P_t \sim \mathrm{Dir}(c_{t-1}).$$

Zero count coordinates are fixed at zero, so the distribution lives on the face spanned by classes observed so far. Equivalently, assign independent Gamma weights to the observed outcomes and normalize. Thus the algorithm is exactly a fresh Bayesian bootstrap (Rubin, 1981). It requires no horizon, smoothness parameter, learning rate, or optimization oracle. The Bayesian-bootstrap draw itself is classical. Our claims concern its new adversarial proper-loss analysis and simultaneous regret guarantees.

The main proof is not a generic posterior-sampling argument. If f is the concave Bayes risk of a proper loss and $F(\alpha) = \mathbb{E}_{P\sim\mathrm{Dir}(\alpha)} f(P)$, we prove

$$\mathbb{E}_{P\sim\mathrm{Dir}(\alpha)} \ell(P, e_j) = nF(\alpha) - (n-1)F(\alpha - e_j), \qquad n = \sum_i \alpha_i. \tag{1}$$

This identity applies to arbitrary bounded proper losses, including polyhedral and V-shaped losses. It turns the perturbed-leader contribution into $T(F(c_T) - f(c_T/T)) \le 0$. The remaining stability term compares two Dirichlet laws whose parameters differ by one count. Their likelihood ratio is linear in one coordinate, which gives stability $O(1/\sqrt{m})$ when that coordinate has appeared m times. Summing separately within each class produces $O(\sum_i \sqrt{N_i}) = O(\sqrt{KT})$.

The same distribution is automatically adapted to smooth losses. Its mean is empirical FTL and its squared radius is at most $1/t$. Centering removes the first-order smoothness term, leaving $O(\beta/t)$. A short semiconcavity argument also gives an explicit $2\beta H_T$ FTL bound under the one-sided smoothness definition used in prior work.

Our contributions are:

- We give one efficient, horizon-free forecaster with $4\sqrt{S_T T} \le 4\sqrt{KT}$ expected regret for every bounded proper loss and simultaneous $O(\beta \log T)$ regret for every bounded β-smooth proper loss.
- We prove the count-deletion identity in (1), including nondifferentiable Bayes risks, zero count coordinates, and the singular case $\alpha_j = 1$.
- We isolate two exact geometric facts behind simultaneous adaptation: one-count Dirichlet total variation controls nonsmooth loss, while centered covariance controls smooth loss.

We emphasize the quantifier. Our result controls $\sup_\ell \mathbb{E}\,\mathrm{Reg}_\ell$, called pseudo-U-calibration. It does not claim the stronger $\mathbb{E}\sup_\ell \mathrm{Reg}_\ell$ for the infinite class of all proper losses. This is the same expected-regret criterion studied by Frongillo et al. (2026).

2 Setting and Algorithm

Let $e_1, \dots, e_K$ denote the vertices of Δ_K. At round t, a forecaster chooses a possibly random $P_t \in \Delta_K$, then observes an outcome $y_t \in \{e_1, \dots, e_K\}$. We first state results for an oblivious outcome sequence. Appendix A gives the standard extension to a nonanticipating adaptive adversary when fresh randomness is used each round.

A loss $\ell : \Delta_K \times \{e_i\}_{i=1}^K \to [-1, 1]$ is *proper* if

$$p \in \arg\min_{q\in\Delta_K} \mathbb{E}_{Y\sim p} \ell(q, Y) \quad \text{for every } p \in \Delta_K.$$

For an outcome sequence, propriety makes its empirical distribution $q_T = T^{-1}\sum_t y_t$ a hindsight minimizer. Define

$$\mathrm{Reg}_\ell(T) = \mathbb{E}\sum_{t=1}^T \ell(P_t, y_t) - Tf(q_T), \qquad f(p) = \sum_i p_i \ell(p, e_i).$$

The expectation is over the forecaster. Let $\mathcal{L}$ be all proper losses with range in $[-1, 1]$ and let $\mathrm{PUCal}_T = \sup_{\ell\in\mathcal{L}} \mathrm{Reg}_\ell(T)$.

Following Frongillo et al. (2026), a differentiable loss is β-smooth when

$$\ell(q, y) - \ell(p, y) \le \langle \nabla_p \ell(p, y), q - p\rangle + \frac{\beta}{2}\|q - p\|_2^2 \tag{2}$$

for every $p, q \in \Delta_K$ and outcome y. Gradients and differentiability on the closed simplex are understood relative to its affine hull, including one-sided differentiability on boundary faces.

For a nonnegative parameter vector α, write $A = \{i : \alpha_i > 0\}$. We define $\mathrm{Dir}(\alpha)$ on the face Δ_A by drawing independent $G_i \sim \mathrm{Gamma}(\alpha_i, 1)$ for $i \in A$, setting $P_i = G_i/\sum_{k\in A} G_k$, and setting $P_i = 0$ outside A.

Algorithm 1 Dirichlet Follow-the-Leader

Initialize $c_0 = 0$ and predict $P_1 = (1/K, \dots, 1/K)$.
for $t = 1, 2, \dots, T$ **do**
  if $t \geq 2$ **then**
    Draw $P_t \sim \mathrm{Dir}(c_{t-1})$ independently of past draws.
  end if
  Observe y_t and set $c_t = c_{t-1} + y_t$.
end for

Theorem 1 (Simultaneously optimal regret)**.** *For every $K, T \geq 1$, let S_T be the number of distinct outcomes observed. Algorithm 1 satisfies*

$$\mathrm{PUCal}_T \leq 4\sqrt{S_T T} \leq 4\sqrt{\min\{K, T\}T}.$$

Simultaneously, for every β-smooth $\ell \in \mathcal{L}$,

$$\mathrm{Reg}_\ell(T) \leq \frac{5}{2}\beta H_T \leq \frac{5}{2}\beta(1 + \log T),$$

where $H_T = \sum_{t=1}^{T} 1/t$.

For $K \leq T$, the $\sqrt{KT}$ worst-case rate is matched by the multiclass V-shaped lower bound of Luo et al. (2024). Squared loss gives an $\Omega(\log T)$ lower bound for the smooth subclass. Hence both worst-case rates in Theorem 1 are optimal up to universal constants in their nontrivial regimes. The support-sensitive refinement is an additional instance-dependent guarantee.

3 A Dirichlet identity for every proper loss

Write $\ell_p = (\ell(p, e_1), \dots, \ell(p, e_K))$. Propriety gives

$$f(q) \leq \langle q, \ell_p \rangle, \qquad f(p) = \langle p, \ell_p \rangle. \tag{3}$$

Thus f is concave and ℓ_p is a supergradient representation. Since $\|\ell_p\|_\infty \leq 1$, applying (3) in both directions also gives

$$|f(p) - f(q)| \leq \|p - q\|_1. \tag{4}$$

Lemma 2 (Count deletion)**.** *Let $\alpha \geq 0$, let $n = \sum_i \alpha_i$, and suppose $\alpha_j \geq 1$. For $F(\alpha) = \mathbb{E}_{P \sim \mathrm{Dir}(\alpha)} f(P)$,*

$$\mathbb{E}_{P \sim \mathrm{Dir}(\alpha)} \ell(P, e_j) = nF(\alpha) - (n-1)F(\alpha - e_j).$$

The statement uses the active-face convention above. When $n = 1$, the second term is zero.

Proof. First suppose $\alpha_j > 1$ and work relative to the active face. A concave function is differentiable almost everywhere on the relative interior. At each such p, (3) makes the tangent component of ℓ_p the unique supergradient of f. Therefore

$$\ell(p, e_j) = f(p) + D_{e_j - p} f(p) \tag{5}$$

almost everywhere.

Let ρ_α be the Dirichlet density and set $v(p) = e_j - p$. A direct calculation on the active face gives

$$\mathrm{div}\, v + D_v \log \rho_\alpha = \frac{\alpha_j - 1}{p_j} - (n-1).$$

The boundary flux vanishes. At a face $p_i = 0$ with $i \neq j$, the normal component $v_i = -p_i$ cancels the possible density singularity. At $p_j = 0$, the assumption $\alpha_j > 1$ makes the flux vanish. Integration by parts and (5) now yield

$$\mathbb{E}_\alpha D_v f(P) = (n-1)F(\alpha) - (\alpha_j - 1)\mathbb{E}_\alpha \frac{f(P)}{P_j},$$

$$(\alpha_j - 1)\mathbb{E}_\alpha \frac{f(P)}{P_j} = (n-1)F(\alpha - e_j),$$

where the second equality is the ratio of Dirichlet normalizers. Combining the two displays proves the claim for $\alpha_j > 1$.

If $\alpha_j = 1$, apply the proved identity to $\alpha + \varepsilon e_j$ and let $\varepsilon \downarrow 0$. The first Dirichlet law converges in total variation, which is enough for the possibly discontinuous bounded function $\ell(\cdot, e_j)$. After deleting one count, the law with parameter ε in coordinate j converges weakly to the lower-dimensional face. Equation (4) makes f continuous, so its expectations converge. The case $n = 1$ is immediate from $\ell(e_j, e_j) = f(e_j)$. □

The distinction between total variation and weak convergence in the last paragraph matters. A bounded proper loss can jump across a decision boundary, while its Bayes risk remains continuous.

4 Worst-case analysis

For analysis, after observing round t, draw a ghost prediction $Q_{t+1} \sim \mathrm{Dir}(c_t)$. For any fixed loss,

$$\begin{aligned} \mathrm{Reg}_\ell(T) = &\sum_{t=1}^{T} \mathbb{E}[\ell(P_t, y_t) - \ell(Q_{t+1}, y_t)] \\ &+ \sum_{t=1}^{T} \mathbb{E}\ell(Q_{t+1}, y_t) - T f(q_T). \end{aligned} \tag{6}$$

If $y_t = e_j$, Lemma 2 gives

$$\mathbb{E}\ell(Q_{t+1}, y_t) = tF(c_t) - (t-1)F(c_{t-1}).$$

The second line of (6) therefore telescopes to

$$T\{F(c_T) - f(q_T)\} \le 0, \tag{7}$$

because $\mathbb{E}Q_{T+1} = q_T$ and f is concave.

It remains to control stability. The next elementary estimate is where the coordinate adaptivity enters.

Lemma 3 (One-count stability)**.** *Let $c \ge 0$, $n = \sum_i c_i$, and $c_j = m > 0$. Then*

$$\mathrm{TV}(\mathrm{Dir}(c), \mathrm{Dir}(c + e_j)) \le \frac{1}{2}\sqrt{\frac{n-m}{m(n+1)}} \le \frac{1}{2\sqrt{m}}.$$

Proof. On their common active face, the likelihood ratio is

$$\frac{d\,\mathrm{Dir}(c + e_j)}{d\,\mathrm{Dir}(c)}(p) = \frac{np_j}{m}.$$

Consequently,

$$\mathrm{TV}(\mathrm{Dir}(c), \mathrm{Dir}(c + e_j)) = \frac{1}{2}\mathbb{E}_{P \sim \mathrm{Dir}(c)}\left|\frac{nP_j}{m} - 1\right|.$$

Cauchy–Schwarz and $\mathrm{Var}(P_j) = m(n-m)/(n^2(n+1))$ prove both inequalities. When $b = n - m > 0$, direct integration at the likelihood-ratio crossing $p_j = m/n$ also gives the exact expression

$$\mathrm{TV} = \frac{(m/n)^m(1 - m/n)^b}{mB(m, b)}.$$

For $b = 0$, both laws are the same point mass and the distance is zero. □

If class j has appeared $m > 0$ times before round t, the range $[-1, 1]$ and Lemma 3 bound the corresponding stability term by $2\,\mathrm{TV} \le 1/\sqrt{m}$. If it is the first appearance, the supports differ and

the trivial bound is 2. Let $N_i = c_{T,i}$ and $S = |\{i : N_i > 0\}|$. Equations (6) and (7) give

$$\begin{aligned}\mathrm{Reg}_\ell(T) &\le 2S + \sum_{i=1}^{K}\sum_{m=1}^{N_i-1}\frac{1}{\sqrt{m}} \\ &\le 2S + 2\sum_{i:N_i>0}\sqrt{N_i} \\ &\le 2S + 2\sqrt{ST} \le 4\sqrt{ST},\end{aligned}$$

where $S \le T$ implies $S \le \sqrt{ST}$. The bound is independent of ℓ, so taking the supremum proves the first part of Theorem 1.

5 Smooth-loss adaptation

For $t \ge 2$, Dirichlet moments give

$$\mathbb{E}P_t = q_{t-1}, \qquad \mathbb{E}\|P_t - q_{t-1}\|_2^2 = \frac{1 - \|q_{t-1}\|_2^2}{t} \le \frac{1}{t}. \tag{8}$$

Set $q_0 = P_1$ for convenience. Applying (2), taking expectations, and using centering yields

$$\mathbb{E}[\ell(P_t, y_t) - \ell(q_{t-1}, y_t)] \le \frac{\beta}{2t}, \qquad t \ge 2. \tag{9}$$

For completeness, we prove the FTL estimate under precisely the one-sided definition (2). This avoids assuming that (2) implies full gradient Lipschitzness.

Lemma 4 (Smooth proper-loss FTL)**.** *For every differentiable proper loss satisfying* (2)*, empirical FTL satisfies*

$$\sum_{t=1}^{T}\ell(q_{t-1}, y_t) - Tf(q_T) \le 2\beta H_T.$$

Proof. We first establish an endpoint condition from propriety and boundary differentiability. Fix $i \ne j$ and set $q_\varepsilon = (1-\varepsilon)e_j + \varepsilon e_i$. Properness at q_ε, comparing the truthful prediction with e_j, says

$$(1-\varepsilon)\ell(q_\varepsilon, e_j) + \varepsilon\ell(q_\varepsilon, e_i) \le (1-\varepsilon)\ell(e_j, e_j) + \varepsilon\ell(e_j, e_i).$$

Expanding at zero gives $D_{e_i - e_j}\ell(e_j, e_j) \le 0$. Propriety at e_j also says that e_j globally minimizes $\ell(\cdot, e_j)$, so the reverse inequality holds. Thus every tangent derivative of $\ell(\cdot, e_j)$ at e_j is zero.

Fix $p \in \Delta_K$, write $d = \|e_j - p\|_2$, and define $\phi(s) = \ell((1-s)p + se_j, e_j)$. Equation (2) is equivalent to concavity of $h(s) = \phi(s) - (\beta d^2/2)s^2$. Since $\phi'(1) = 0$, monotonicity of the derivative of h gives, for $s < 1$,

$$\phi'(s) - \beta d^2 s = h'(s) \ge h'(1) = -\beta d^2.$$

Hence $\phi'(s) \ge -\beta d^2(1-s)$. The FTL update after outcome e_j is $q_t = (1-1/t)q_{t-1} + e_j/t$. Integrating from 0 to $1/t$ gives

$$\ell(q_{t-1}, e_j) - \ell(q_t, e_j) \le \frac{\beta\|e_j - q_{t-1}\|_2^2}{t} \le \frac{2\beta}{t}.$$

The standard be-the-leader inequality $\sum_t \ell(q_t, y_t) \le Tf(q_T)$ and summation prove the result. □

Adding (9) to Lemma 4 gives at most $2\beta H_T + (\beta/2)(H_T - 1) \le (5\beta/2)H_T$, completing Theorem 1.

6 Related work and interpretation

U-calibration. Kleinberg et al. (2023) introduced U-calibration for binary outcomes and connected it to unknown downstream agents. Luo et al. (2024) obtained the optimal $\Theta(\sqrt{KT})$ multiclass pseudo-U-calibration rate, proved the V-shaped lower bound, and identified loss classes on

which FTL is logarithmic. Their worst-case-optimal perturbed leader is not adaptive to smooth losses. Frongillo et al. (2026) introduced a self-concordant prediction-space perturbation that adapts to smooth and log-barrier-smooth losses. For the classes considered here, its bounds are approximately $K^{5/4}\sqrt{T}$ and $\beta \log T + \beta\sqrt{K}\log K$. Our result removes these dimension gaps. Calibeating, omniprediction, and swap-regret results study related universal downstream guarantees with different benchmarks (Roth & Shi, 2024; Chen et al., 2026).

Perturbation and posterior sampling. Follow-the-Perturbed-Leader is classical (Kalai & Vempala, 2005; Cesa-Bianchi & Lugosi, 2006). Fresh perturbations also yield the standard nonanticipating adaptive-adversary extension (Hutter & Poland, 2005). Algorithm 1 can instead be viewed as posterior sampling for a categorical model under the limiting zero-mass Dirichlet prior, or as Rubin's Bayesian bootstrap. Dirichlet-weighted loss minimization appears in the loss-likelihood bootstrap (Lyddon et al., 2019), and normalized random weights have also been maintained in online Bayesian bagging (Lee & Clyde, 2004). Dirichlet Bayes mixtures are classical in universal coding under log loss (Watanabe et al., 2013), while size-biased Dirichlet integral identities are classical probability machinery (Last, 2020). None of these works gives adversarial regret for an unknown proper loss. Our contribution is the loss-uniform count-deletion specialization, its telescoping use, and the simultaneously optimal guarantees, not the bootstrap draw itself.

Why Dirichlet geometry fits both regimes. The update $c \mapsto c + e_j$ has likelihood ratio np_j/c_j, so its total variation automatically scales with the number of previous appearances of the updated class. This is the right geometry for discontinuous proper losses. At the same time,

$$\mathrm{Cov}(P_t) = \frac{\mathrm{diag}(q_{t-1}) - q_{t-1}q_{t-1}^\top}{t},$$

which has the centered $O(1/t)$ radius needed for smooth losses. The exact identity in Lemma 2 is what lets these two local facts control global regret without a separate perturbed-leader path-length penalty.

7 Limitations and Conclusion

Our theorem resolves simultaneous optimality for bounded proper losses and the smooth subclass under the expected-regret criterion. It does not control actual U-calibration $\mathbb{E}\sup_{\ell\in\mathcal{L}}\mathrm{Reg}_\ell$ over the entire infinite class. Moving the supremum inside expectation requires complexity control or a different argument. We also do not address the broader class of losses smooth relative to a log barrier, where Euclidean Dirichlet covariance need not be the right measure.

The result shows that the previous tradeoff was geometric rather than information-theoretic. A Bayesian-bootstrap sample of empirical FTL has exactly the count stability needed by nonsmooth scores and exactly the centering needed by smooth scores. The count-deletion identity makes those two properties compatible in one short analysis.

References

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Yurong Chen, Zhiyi Huang, Michael I. Jordan, and Haipeng Luo. Calibeating made simple. In *Conference on Learning Theory*, pp. 1373–1398. PMLR, 2026.

Rafael Frongillo, Haipeng Luo, Nishant A. Mehta, and Jon Schneider. Toward simultaneously optimal regret in u-calibration. In *Conference on Learning Theory*, 2026. arXiv:2606.18527.

Marcus Hutter and Jan Poland. Adaptive online prediction by following the perturbed leader. *Journal of Machine Learning Research*, 6(22):639–660, 2005.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Bobby Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-calibration: Forecasting for an unknown agent. In *Conference on Learning Theory*, pp. 5143–5145. PMLR, 2023.

Günter Last. An integral characterization of the dirichlet process. *Journal of Theoretical Probability*, 33(2):918–930, 2020.

Herbert K. H. Lee and Merlise A. Clyde. Lossless online bayesian bagging. *Journal of Machine Learning Research*, 5:143–151, 2004.

Haipeng Luo, Spandan Senapati, and Vatsal Sharan. Optimal multiclass u-calibration error and beyond. *Advances in Neural Information Processing Systems*, 37:7521–7551, 2024.

Simon P. Lyddon, Chris C. Holmes, and Stephen G. Walker. General bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478, 2019.

Aaron Roth and Mirah Shi. Forecasting for swap regret for all downstream agents. In *ACM Conference on Economics and Computation*, pp. 466–488, 2024.

Donald B. Rubin. The bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981.

Kazuho Watanabe, Teemu Roos, and Petri Myllymäki. Achievability of asymptotic minimax regret in online and batch prediction. In *Asian Conference on Machine Learning*, pp. 181–196. PMLR, 2013.

A Adaptive outcomes

The main text assumes a fixed outcome sequence. The same expected bounds hold for a nonanticipating adaptive adversary that may depend on past predictions but not on the fresh draw P_t. Condition on the history before the round and on the selected outcome. Given the current count vector, P_t is a fresh Dirichlet draw, so the one-step identity and stability bounds hold conditionally. Summing conditional expectations gives the same telescoping count potential along the realized count path. This is the usual fresh-randomization reduction for perturbed leader (Hutter & Poland, 2005). No guarantee is possible against an adversary that observes the current randomized forecast before choosing the same-round outcome under this protocol.

B Additional boundary details for Lemma 2

We record a standard approximation that makes the nonsmooth integration by parts fully formal. Restrict f to the relative interior of the active face. It is concave and Lipschitz by (4). Convolve it on compact subsets with a smooth mollifier, apply integration by parts to the smooth approximants, and then exhaust the face. The directional derivatives of a concave Lipschitz function converge almost everywhere and are uniformly bounded in tangent directions. Dominated convergence applies to every term when $\alpha_j > 1$. The boundary flux estimate in the main proof is uniform under this exhaustion.

For $\alpha_j = 1$, write $\alpha^{(\varepsilon)} = \alpha + \varepsilon e_j$. On the common active face, the densities of $\mathrm{Dir}(\alpha^{(\varepsilon)})$ converge in L^1 to that of $\mathrm{Dir}(\alpha)$, hence in total variation. After deleting e_j, the marginal coordinate P_j has a Beta distribution with first parameter ε and therefore converges to zero in probability. Conditional proportions of the remaining coordinates have distribution $\mathrm{Dir}(\alpha_{-j})$. This proves weak convergence to the deleted face. Boundedness handles the left side and continuity of f handles the right side.

C Sanity checks

A supplementary script evaluates the identity for Brier, spherical, binary V-shaped, and multiclass polyhedral proper losses. It also checks the likelihood-ratio total-variation bound and exhaustively enumerates all binary outcome sequences through $T = 8$. The smooth Brier identities are also checked in closed form. These calculations are not used by any proof.