ASYMPTOTICALLY OPTIMAL PARTITIONS FOR AGREEMENT OF MANY GAUSSIAN SAMPLES

ROY VAN RIJN

Abstract. We study the probability that m equicorrelated Gaussian observations all fall in the same cell of a partition into k equal-measure sets. For fixed $k \geq 3$ and correlation $0 < \rho < 1$, the largest agreement probability is asymptotic to $2H_m$ as $m \to \infty$, uniformly in the ambient dimension. Here H_m is the probability that all observations fall in a single halfspace of mass $1/k$, and $H_m = m^{-(1-\rho)/\rho+o(1)}$. Dividing one coordinate into k equal-mass intervals attains this bound asymptotically. The regular simplex has agreement exponent $2(k-1)/k$ times the halfspace exponent. This gives a counterexample to the many-observation simplex proposal of Isaksson and Mossel and yields exactly balanced low-influence voting rules with greater unanimous agreement than plurality for sufficiently many correlated tallies.

hexagon:2611.01282v1

1. The agreement problem

A partition of Gaussian space labels each observation by its cell. Given many observations with the same positive pairwise correlation, we ask for the largest probability that they all receive one label.

Fix $k \geq 3$, $0 < \rho < 1$, and $d \geq 1$. Write γ_d for standard Gaussian measure on $\mathbb{R}^d$. A partition $A = (A_1, \ldots, A_k)$ is *balanced* if $\gamma_d(A_i) = 1/k$ for each i. For jointly Gaussian observations $Y_1, \ldots, Y_m$ with mean zero, covariance I_d per observation, and $\mathrm{Cov}(Y_j, Y_\ell) = \rho I_d$ for $j \neq \ell$, set

$$M_m(A) = \sum_{i=1}^{k} \Pr(Y_1, \ldots, Y_m \in A_i).$$

Let H_m be the corresponding probability for a single halfspace of mass $1/k$. Let Q be the partition into the inverse images of the k quantile intervals of a standard normal under the first coordinate. The two end cells of Q are opposite halfspaces.

Theorem 1. *For fixed k and ρ as above, in every $d \geq 1$,*

$$2H_m \;\leq\; M_m(Q) \;\leq \sup_{A \text{ balanced}} M_m(A) \;=\; (2 + o(1))H_m.$$

Here $H_m = m^{-\alpha+o(1)}$, with $\alpha = (1-\rho)/\rho$, as $m \to \infty$. The error in the supremum is uniform over $d \geq 1$.

When $m = 2$, this is the agreement objective in the Standard Simplex Conjecture of Isaksson and Mossel [2, Conjecture 1.4]. Their Section 8 also proposes a many-observation extension. In the equicorrelated setting of that extension, the simplex can be beaten for every fixed $k \geq 3$ and $\rho \in (0, 1)$ when m is sufficiently large. For $d \geq k-1$, let T denote the partition whose cells are the winning regions of a centered regular simplex, extended cylindrically to $\mathbb{R}^d$.

Proposition 2. *With k and ρ fixed and $d \geq k-1$, the centered regular-simplex partition satisfies*

$$M_m(T) = m^{-\beta\alpha+o(1)}, \qquad \beta = \frac{2(k-1)}{k}.$$

2020 *Mathematics Subject Classification.* 60E15, 60G15, 60F05.

Key words and phrases. Gaussian noise stability, equicorrelated agreement, regular simplex, quantile slabs, low influences.

Together the two estimates give $M_m(Q)/M_m(T) = m^{(\beta-1)\alpha+o(1)}$. For example, when $k = 3$ and $\rho = 1/2$, the halfspace exponent is 1 and the simplex exponent is 4/3. The exponent gap comes from the need to satisfy several facet inequalities to make the conditional probability of a simplex cell close to one. In the upper bound for an arbitrary partition, a different geometric effect forces the constant 2. Conditional probabilities close to one for different labels require well-separated common Gaussian means. Three such sets cannot all have large Gaussian measure. The single-cell comparison of Isaksson and Mossel [2, Theorem 1.2] then controls the two remaining terms. The result gives the asymptotic value of the supremum. Identifying exact optimizers at finite m and classifying the families that attain the asymptotic value are separate questions.

2. Halfspace and simplex posterior probabilities

Write $s = \sqrt{\rho}$, $\sigma = \sqrt{1-\rho}$, and $a = \sigma/s$. One can represent the observations as $Y_j = sX + \sigma Z_j$, with $X, Z_1, \dots, Z_m$ independent and standard Gaussian in $\mathbb{R}^d$. For each cell set

$$p_i(x) = \Pr(sx + \sigma Z_1 \in A_i), \qquad M_m(A) = \sum_{i=1}^{k} \mathbb{E}[p_i(X)^m].$$

In the displayed identity, the observations are conditionally independent given X. Gaussian smoothing of a cell indicator gives a continuous function p_i. We may choose Borel cell representatives since changing a cell on a Gaussian null set leaves every conditional probability unchanged.

Write Φ for the standard normal distribution function and $\overline{\Phi} = 1 - \Phi$. For the halfspace $\{y : y_1 > h\}$ of mass $1/k$, where $h = \Phi^{-1}(1-1/k)$, its conditional probability is $p_H(x) = \Phi((sx_1 - h)/\sigma)$. For $\epsilon \in (0, 1/2)$ put $t = \Phi^{-1}(1-\epsilon)$. Then

$$\Pr(p_H(X) > 1 - \epsilon) = \overline{\Phi}(at + h/s). \tag{1}$$

The Gaussian tail estimate $\log \overline{\Phi}(t) = -t^2/2 + o(t^2)$ gives $\Pr(p_H(X) > 1-\epsilon) = \epsilon^{\alpha+o(1)}$ as $\epsilon \downarrow 0$.

Lemma 3. *If P takes values in $[0,1]$ and $\Pr(P > 1-\epsilon) = \epsilon^{\eta+o(1)}$ as $\epsilon \downarrow 0$ for some $\eta > 0$, then $\mathbb{E}[P^m] = m^{-\eta+o(1)}$.*

Proof. On $\{P > 1 - 1/m\}$, the factor P^m is bounded below by $(1-1/m)^m$ for $m \geq 2$. For the reverse bound, split at $1 - C\log(m)/m$ with fixed $C > \eta$. Below this level the mth power is at most m^{-C}. The probability of exceeding the level is $m^{-\eta+o(1)}$ by the assumed tail. □

Since $H_m = \mathbb{E}[p_H(X)^m]$, Lemma 3 gives the halfspace power law of Theorem 1.

Proof of Proposition 2. Identify the active space with $E = \mathbf{1}^\perp \subset \mathbb{R}^k$ and write $e_1, \dots, e_k$ for the coordinate vectors of $\mathbb{R}^k$. The cell with label i is the intersection of the halfspaces with inward unit normals $n_{ij} = (e_i - e_j)/\sqrt{2}$, $j \neq i$. If $p_i(x) > 1 - \epsilon$, each containing halfspace has conditional probability greater than $1-\epsilon$. Thus $x \cdot n_{ij} > at$ for all $j \neq i$. The sum of the $k-1$ normals has length $\sqrt{k(k-1)/2}$. If u_i is its unit direction, summing the inequalities gives

$$x \cdot u_i > \sqrt{\beta}\, at, \qquad \Pr(p_i(X) > 1-\epsilon) \leq \overline{\Phi}(\sqrt{\beta}\, at). \tag{2}$$

For the converse let $t' = \Phi^{-1}(1 - \epsilon/(k-1))$. Each n_{ij} has inner product $1/\sqrt{\beta}$ with u_i. On the event

$$X \cdot u_i > \sqrt{\beta}(at' + 1), \qquad \left|X_E - (X \cdot u_i)u_i\right| \leq 1,$$

where X_E denotes the projection onto E, every margin exceeds at'. The union bound over the $k-1$ halfspaces gives $p_i(X) > 1 - \epsilon$. The two conditions are independent, and the second has a fixed positive probability in the $k-2$ dimensional space perpendicular to u_i within E. The Gaussian tail estimate gives $t'/t \to 1$. With (2) we conclude that

$$\Pr(p_i(X) > 1 - \epsilon) = \epsilon^{\beta\alpha+o(1)} \qquad (\epsilon \downarrow 0).$$

Lemma 3 gives the exponent for one cell, and summing over the fixed k preserves it. □

3. Only two cells can contribute at the halfspace scale

For a balanced partition and $0 < \epsilon < 1/2$ let

$$D_i = \{x : p_i(x) > 1 - \epsilon\}, \qquad t = \Phi^{-1}(1 - \epsilon).$$

We first bound the measure of the third largest of the D_i. The geometry of the conditional laws supplies a bound uniform over all ambient dimensions.

Lemma 4. *For distinct labels i, j, all $x \in D_i$ and $y \in D_j$ satisfy $|x - y| \geq 2at$. For three distinct labels i, j, ℓ,*

$$\big(\gamma_d(D_i)\gamma_d(D_j)\gamma_d(D_\ell)\big)^{1/3} \leq \exp(-2a^2t^2/3).$$

Proof. Under the Gaussian law with mean sx and covariance $\sigma^2 I_d$, the set A_i has measure greater than $1 - \epsilon$. Under the law with mean sy and the same covariance it has measure less than ϵ, since A_j has measure greater than $1 - \epsilon$. The test using A_i therefore has sum of errors less than 2ϵ. The smallest possible sum of errors is the integral of the pointwise minimum of the two densities. For $x \neq y$ the two densities cross on the plane through the midpoint of their means, perpendicular to $x - y$, so projection gives the minimum as $2\overline{\Phi}(s|x - y|/(2\sigma))$. The same formula gives one when $x = y$. It follows that $s|x - y|/(2\sigma) \geq t$.

Now take x_1, x_2, x_3 in the three respective sets and put $z = (x_1 + x_2 + x_3)/3$. The separation just proved gives

$$\frac{1}{3}\sum_{\nu=1}^{3} |x_\nu|^2 - |z|^2 = \frac{1}{9}\sum_{\mu<\nu} |x_\mu - x_\nu|^2 \geq \frac{(2at)^2}{3}.$$

Write φ_d for the standard Gaussian density. The geometric mean of the three restricted densities $\varphi_d 1_{D_i}$, $\varphi_d 1_{D_j}$, $\varphi_d 1_{D_\ell}$, evaluated at their respective points, is bounded by $\exp(-2a^2t^2/3)\varphi_d(z)$. If any point lies outside its assigned set, the geometric mean is zero and the same bound holds. The three-function Prékopa–Leindler inequality [3, Theorem 3] applied with weights $1/3$ gives the asserted product bound. □

Proof of Theorem 1. By Lemma 4, at most two labels satisfy $\gamma_d(D_i) > \exp(-2a^2t^2/3)$. Choose two labels covering all those that do. The Exchangeable Gaussians Theorem of Isaksson and Mossel [2, Theorem 1.2] bounds the entire term $\Pr(Y_1, \ldots, Y_m \in A_i)$ by H_m for each of the two chosen labels. Its hypotheses apply to our jointly Gaussian observations with common positive cross covariance and to the repeated set A_i of mass $1/k$. The identical-cell case also follows from Borell's earlier Gaussian rearrangement theorem [1, Theorem 3.1].

For every other label, integrate the mth power of p_i separately on D_i and its complement. This gives the dimension-independent finite-m bound

$$M_m(A) \;\leq\; 2H_m + (k - 2)\big((1 - \epsilon)^m + \exp(-2a^2t^2/3)\big). \tag{3}$$

Choose $\epsilon = C\log(m)/m$ for any fixed $C > \alpha$. The first term inside the parentheses is at most m^{-C}. Since $t^2 = 2\log(1/\epsilon) + o(\log(1/\epsilon))$, the second term is $m^{-4\alpha/3+o(1)}$. Both are $o(H_m)$ by the halfspace estimate (1) and Lemma 3. The two end cells of the quantile partition each contribute H_m, so $M_m(Q) \geq 2H_m$. This gives both the asymptotic supremum and uniformity in dimension. □

4. Exactly balanced voting rules

Consider n independent voters. For each voter, draw a common ballot uniformly from $\{1, \ldots, k\}$. For each of the m copies independently, keep the common ballot with probability $s = \sqrt{\rho}$. Otherwise draw a new uniform ballot independently. The ballot profile in each copy is uniform. For a rule $f : \{1, \ldots, k\}^n \to \{1, \ldots, k\}$, its influence at a voter is the probability that independently resampling that vote in a uniform profile changes f. Exact balance means every label is chosen with probability $1/k$ on a uniform profile.

Theorem 5. *Fix $k \geq 3$ and $\rho \in (0,1)$. For every sufficiently large fixed m there are exactly balanced deterministic rules f_n whose maximum voter influence tends to zero as $n \to \infty$. Their limiting probability of unanimous agreement across the m copies is strictly greater than for plurality in the same model.*

Proof. Write $b = 1/k$ and for each copy center the count of votes for candidate 1 at nb and divide by $\sqrt{nb(1-b)}$. For a single voter the indicator of a vote for candidate 1 has variance $b(1-b)$. Its conditional mean given the common ballot B is $s1_{\{B=1\}} + (1-s)b$. Distinct copies are conditionally independent given B, so the cross covariance is $s^2 b(1-b) = \rho b(1-b)$. The multivariate central limit theorem, at fixed m, gives the one-dimensional equicorrelated Gaussian model for the normalized counts across copies.

Apply the k Gaussian quantile intervals to these counts to get a rule g_n. By convergence at the thresholds, its label masses tend to $1/k$ and its all-copy agreement tends to $M_m(Q)$. A resampled vote moves the normalized count by at most $(nb(1-b))^{-1/2}$. It changes the label only within that distance of a threshold. For each fixed positive neighborhood size the limit probability is its Gaussian measure. Letting that size decrease to zero shows that the maximum voter influence tends to zero.

To get exact balance, change labels on a minimal set of profiles by assigning surplus profiles to deficient labels. The target number k^{n-1} of profiles per label is integral. The fraction δ_n of profiles changed equals half the sum of the absolute discrepancies of the label masses from $1/k$ and tends to zero. For a given voter, the original profile and the independently resampled profile both have uniform marginals. The influence increases by at most $2\delta_n$. For m copies the probability of any change from g_n is at most $m\delta_n$, so the agreement limit of the exactly balanced rule f_n remains $M_m(Q)$.

For plurality use the centered count vector in each copy. Its one-voter covariance is k^{-1} times the orthogonal projection onto $\mathbf{1}^\perp$, and the cross-copy covariance is ρ times the same matrix, by conditional independence as above. After normalization, the multivariate central limit theorem gives m standard Gaussians on $\mathbf{1}^\perp$ with cross covariance ρI. The largest coordinate is the simplex label and the Gaussian tie planes have measure zero. Thus plurality agreement converges to $M_m(T)$ for any rule for breaking ties. Plurality itself can be balanced at every n by selecting the candidate of the first voter whose ballot names a tied winner. This choice commutes with permutations of the candidates. By Theorem 1 and Proposition 2, $M_m(Q) > M_m(T)$ for sufficiently large m. For each such fixed m, the two limits in n give the strict comparison. □

AI assistance and review. ChatGPT (OpenAI) assisted with mathematical arguments, literature checks, exposition, proof review and verification code.

References

[1] C. Borell, *Geometric bounds on the Ornstein–Uhlenbeck velocity process*, Z. Wahrscheinlichkeitstheorie verw. Gebiete **70** (1985), 1–13. doi:10.1007/BF00532234.

[2] M. Isaksson and E. Mossel, *Maximally stable Gaussian partitions with discrete applications*, Israel J. Math. **189** (2012), 347–396. doi:10.1007/s11856-011-0181-7. See also arXiv:0903.3362v3.

[3] A. Prékopa, *On logarithmic concave measures and functions*, Acta Sci. Math. (Szeged) **34** (1973), 335–343. Journal archive.