Supplementary Information to Recalibration of Predicted Probabilities Using the "Logit Shift": Why does it work, and when can it be expected to work well?

Evan T. R. Rosenman[1], Cory McCartan[2], and Santiago Olivella[3]

[1]Data Science Initiative, Harvard University, Cambridge, MA 02138. Email: erosenm@fas.harvard.edu
[2]Department of Statistics, Harvard University, Cambridge, MA 02138. Email: cmccartan@g.harvard.edu
[3]Department of Political Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599. Email: olivella@unc.edu

1 Characterizations of the Logit Shift: Minimum Summed KL Divergences

First, we show that the logit shift minimizes the summed KL divergence from the original scores p_i to the updated scores $\tilde{p}_i$, subject to $\sum \tilde{p}_i = D$ constraint.

We denote a generic optimization variable as z_i, and define the minimum-KL-divergence optimization problem as

$$\begin{aligned} \text{minimize} \quad & \sum_{i \in \mathcal{V}} -z_i \log\left(\frac{p_i}{z_i}\right) - (1-z_i)\log\left(\frac{1-p_i}{1-z_i}\right) \\ \text{subject to} \quad & \sum_{i \in \mathcal{V}} z_i = D, \quad 0 \leq z_i \leq 1 \text{ for } i \in \mathcal{V}. \end{aligned} \tag{1}$$

The Lagrangian for Optimization Problem 1 is

$$L(\{z_i\}, \{\lambda_i\}, \{\nu_i\}, \gamma) = \sum_{i \in \mathcal{V}} -z_i \log\left(\frac{p_i}{z_i}\right) - (1-z_i)\log\left(\frac{1-p_i}{1-z_i}\right) + \sum_{i \in \mathcal{V}} \lambda_i (z_i - 1) - \nu_i \tilde{x}_i + \gamma\left(\sum_{i \in \mathcal{V}} z_i - D\right).$$

We make the standard assumptions that $0 < p_i < 1$ for all $i \in \mathcal{V}$ and $0 < D < |\mathcal{V}|$. Define the point $(\{z_i\}, \{\lambda_i\}, \{\nu_i\}, \gamma) = (\{\tilde{p}_i\}, \{0\}, \{0\}, \log(\alpha))$. We consider the Karush-Kuhn-Tucker (KKT) conditions at this point. For the Lagrangian gradient condition, observe:

$$\begin{aligned}\nabla L\left(\{\tilde{p}_i\}, \{0\}, \{0\}, \log(\alpha)\right) &= -\log\left(\frac{p_i/(1-p_i)}{\tilde{p}_i/(1-\tilde{p}_i)}\right) + \log(\alpha) \\ &= -\log\left(\frac{p_i/(1-p_i)}{p_i/(\alpha(1-p_i))}\right) + \log(\alpha) \\ &= 0\,,\end{aligned}$$

while the other four KKT conditions are automatically satisfied at this point. It follows that this point is dual optimal. Lastly, because the objective function is convex and there exist choices of z_i satisfying $0 < z_i < 1$ for $i \in \mathcal{V}$ and $\sum_{i\in\mathcal{V}} z_i = D$, strong duality is attained. Hence, our point is optimal and the $\tilde{p}_i$ are a solution to Optimization Problem 1. For background technical details, see Boyd et al. (2004).

2 Proof of Theorem 1

Define $g(\boldsymbol{v}, \boldsymbol{s})$ as the function

$$g(\boldsymbol{v}, \boldsymbol{s}) = \sum_{i\in\mathcal{V}} f(v_i, s_i)$$

where $\boldsymbol{v} = \{v_i\}_{i\in\mathcal{V}}$ and $\boldsymbol{s} = \{s_i\}_{i\in\mathcal{V}}$. Observe that $g(\boldsymbol{v}, \boldsymbol{s})$ is monotonically decreasing in every component of s. Denote $\boldsymbol{\alpha} = \{\alpha\}_{i\in\mathcal{V}}$, the vector repeating α a total of $|\mathcal{V}|$ times, and $\boldsymbol{\phi} = \{\phi_i\}_{i\in\mathcal{V}}$. Because

$$g(\boldsymbol{p}, \boldsymbol{\phi}) = D \qquad \text{and} \qquad g(\boldsymbol{p}, \boldsymbol{\alpha}) = D,$$

it follows immediately that α must lie between the largest and smallest value of ϕ_i across all choices of i.

3 Proof of Theorem 2

The log-concavity of the Poisson-Binomial distribution is a well-established result (see e.g. Wang, 1993). Hence, for any choice of i, we have

$$\mathbb{P}\left(\sum_{j\neq i} W_j = D-2\right)\mathbb{P}\left(\sum_{j\neq i} W_j = D\right) \leq \mathbb{P}\left(\sum_{j\neq i} W_j = D-1\right)^2 \tag{2}$$

Multiplying both sides of the equality by p_i, and adding the same quantity to both sides, we obtain an updated inequality

$$p_i\mathbb{P}\left(\sum_{j\neq i} W_j = D-2\right)\mathbb{P}\left(\sum_{j\neq i} W_j = D\right) + (1-p_i)\mathbb{P}\left(\sum_{j\neq i} W_j = D-1\right)\mathbb{P}\left(\sum_{j\neq i} W_j = D\right) \leq$$
$$p_i\mathbb{P}\left(\sum_{j\neq i} W_j = D-1\right)^2 + (1-p_i)\mathbb{P}\left(\sum_{j\neq i} W_j = D-1\right)\mathbb{P}\left(\sum_{j\neq i} W_j = D\right).$$

Collecting terms, we get

$$\mathbb{P}\left(\sum_{j\neq i} W_j = D\right)\left(p_i\mathbb{P}\left(\sum_{j\neq i} W_j = D-2\right) + (1-p_i)\mathbb{P}\left(\sum_{j\neq i} W_j = D-1\right)\right) \leq$$
$$\mathbb{P}\left(\sum_{j\neq i} W_j = D-1\right)\left(p_i\mathbb{P}\left(\sum_{j\neq i} W_j = D-1\right) + (1-p_i)p\left(\sum_{j\neq i} W_j = D\right)\right) \tag{3}$$

The terms in parentheses can be collapsed into a single Poisson-Binomial probability, making use of the recursion defined in Footnote 5. Subbing these expressions into Inequality 3, we obtain

$$\mathbb{P}\left(\sum_{j\neq i} W_j = D\right)\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j = D-1\right) \leq \mathbb{P}\left(\sum_{j\neq i} W_j = D-1\right)\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j = D\right),$$

which yields the upper bound in Theorem 2.

The proof of the lower bound proceeds by incrementing D by 1 in Inequality 2 and following the same set of steps.

4 Proof of Theorem 3

Under the bounds defined in Inequality 7 in the main text, we obtain the following approximation bounds for α_i:

$$\frac{|\alpha-\phi_i|}{\phi_i} \leq \frac{\frac{\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j=D\right)}{\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j=D-1\right)} - \frac{\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j=D+1\right)}{\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j=D\right)}}{\frac{\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j=D+1\right)}{\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j=D\right)}}. \tag{4}$$

It is a well-established result that, for large N, the Poisson-Binomial behaves approximately as a Normal random variable with the same mean and variance (see e.g. Siripraparat and Neammanee, 2021), namely

$$\mu = \sum_j p_j \qquad \text{and} \qquad \sigma^2 = \sum_j p_j(1-p_j).$$

Denote as $\psi(d)$ the density of a Normal distribution $\mathcal{N}(\mu, \sigma^2)$ with this mean and variance, evaluated at d. Siripraparat and Neammanee (2021) show that over all possible choices of d, the largest deviation between $\psi(d)$ and $\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j = d\right)$ is bounded above by C_1/σ^2 for a constant $C_1 > 0$. Hence

$$\begin{aligned}\frac{|\alpha-\phi_i|}{\phi_i} &\leq \frac{\psi(D)^2}{\psi(D-1)\psi(D+1)} - 1 + \mathcal{O}\left(\frac{1}{\sigma^2}\right) \\ &= \exp\left(\frac{1}{\sigma^2}\right) - 1 + \mathcal{O}\left(\frac{1}{\sigma^2}\right) = \mathcal{O}\left(\frac{1}{\sigma^2}\right)\end{aligned} \tag{5}$$

This follows from a series expansion of the fraction in (4).

Lastly, we observe

$$\begin{aligned}\tilde{p}_i &= \frac{1}{1+\frac{1-p_i}{p_i}\alpha} \\ &= \frac{1}{1+\frac{1-p_i}{p_i}\phi_i\left(1+\frac{\alpha-\phi_i}{\phi_i}\right)} \\ &= \frac{1}{1+\frac{1-p_i}{p_i}\phi_i} - \frac{(1-p_i)p_i\phi_i}{(p_i+\phi_i-p_i\phi_i)^2}\left(\frac{\alpha-\phi_i}{\phi_i}\right) + \mathcal{O}\left(\left(\frac{\alpha-\phi_i}{\phi_i}\right)^2\right) \\ &= p_i^\star + \mathcal{O}\left(\frac{1}{\sigma^2}\right),\end{aligned}$$

where the last line follows by plugging in the bound on $(\alpha-\phi_i)/\phi_i$ from Inequality 5 and observing

$$\left|\frac{(1-p_i)p_i\phi_i}{(p_i+\phi_i-p_i\phi_i)^2}\right| \leq \frac{1}{4} \qquad \text{for } 0 < p_i < 1, 0 < \phi_i.$$

5 Proof of Theorem 4

Fix a higher-level aggregation unit $B \in \mathcal{B}$. Then each person i residing in B also resides in some aggregation unit $A_i \in \mathcal{A}$, with A_i contained in B, and belongs to one of a set of mutually exclusive population groups (e.g., racial groups), $G_i \in \mathcal{G}$.

For a lower-level unit $A \in \mathcal{A}$, define

$$s_g(A) = \frac{\sum_{i:A_i=A} \mathbf{1}\{G_i = g\}}{|\{i : A_i = A\}|}$$

to be the share of the population of A that belongs to group $g \in \mathcal{G}$. The vector of these group shares $\boldsymbol{s}(A)$ necessarily sums to 1.

Then define a function $m(\boldsymbol{s}(A)) = 1 - \max_g s_g(A)$, which, given a vector of group shares for a unit A, computes the proportion of people who are in the minority in unit A. For example, $m((0,1,0)) = 0$ and $m((0.4, 0.3, 0.3)) = 0.6$. Notice that since the maximum function is convex, m is concave.

Let $I \in \{i : i \text{ resides in } B\}$ be the random variable that is uniformly distributed across all of the people residing in B. Then $\boldsymbol{s}(A_I)$ are the group shares for the unit containing a randomly-selected person, and $\mathbb{E}[\boldsymbol{s}(A_I)]$ is the vector of group shares for the overall unit B (this follows naturally from the definition of $s_g(A)$).

Similarly, $\mathbb{E}[m(\boldsymbol{s}(A_I))]$ is the overall proportion of people comprising a minority within their $\mathcal{A}$ unit. We can of course write the proportion of people comprising a minority within the enclosing unit B as $m(\mathbb{E}[\boldsymbol{s}(A_I)])$. The claim of the theorem is then that $\mathbb{E}[m(\boldsymbol{s}(A_I))] \leq m(\mathbb{E}[\boldsymbol{s}(A_I)])$. But this follows immediately from Jensen's inequality, since m is concave.

The inequality will be strict as long as there is at least one unit A where the majority population differs from the overall majority population, since then the support of $m(\boldsymbol{s}(A_I))$ will span a region where m is strictly concave.

References

Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Siripraparat, T. and Neammanee, K. (2021). A local limit theorem for Poisson Binomial random variables. *Sci. Asia. https://doi.org/10.2306/scienceasia1513-1874.2021*, 6.

Wang, Y. H. (1993). On the number of successes in independent trials. *Statistica Sinica*, pages 295–312.