

Recalibration of Predictive Models as Approximate Probabilistic Updates

Evan T. R. Rosenman
Data Science Initiative
Harvard University

Santiago Olivella
Department of Political Science
University of North Carolina, Chapel Hill

December 14, 2021

Abstract

The output of predictive models is routinely recalibrated by reconciling low-level predictions with known derived quantities defined at higher levels of aggregation. For example, models predicting turnout probabilities at the individual level in U.S. elections can be adjusted so that their aggregation matches the observed vote totals in each state, thus producing better calibrated predictions. In this research note, we provide theoretical grounding for one of the most commonly used recalibration strategies, known colloquially as the “logit shift”. Typically cast as a heuristic optimization problem (whereby an adjustment is found such that it minimizes the difference between aggregated predictions and the target totals), we show that the “logit shift” in fact offers a fast and accurate approximation to a principled, but often computationally impractical adjustment strategy: computing the posterior prediction probabilities, conditional on the target totals. After deriving analytical bounds on the quality of the approximation, we illustrate the accuracy of the approach using Monte Carlo simulations. The simulations also confirm analytical results regarding scenarios in which users of the simple logit shift can expect it to perform best — namely, when the aggregated targets are comprised of many individual predictions, and when the distribution of true probabilities is symmetric and tight around 0.5.

1 Problem Description

A common problem in predictive modeling is that of calibrating probabilities to observed totals. For example, an analyst may obtain individual-level scores $p_i \in (0, 1)$, $i = 1, \dots, N$, to estimate the probability that each of the N registered voters in a particular voting precinct will support the Democratic candidate in an upcoming election. After the election occurs, the analyst can observe the total number of Democratic votes, D , cast among the subset

$\mathcal{V} \subset \{1, \dots, N\}$ of registered voters who cast a ballot. But she cannot observe individual-level outcomes due to the secret ballot. In the absence of perfect prediction, the analyst will find that $\sum_{i \in \mathcal{V}} p_i \neq D$. She must then decide how to compute recalibrated scores, \tilde{p}_i , to better reflect the realized electoral outcome.

This practical problem has direct implications for public opinion research. For example, Ghitza and Gelman (2020) recalibrate their MRP estimates of voter support levels after an election to match county-level totals, while Schwenzfeier (2019) proposes using the magnitude of the calibration to estimate non-response bias in public opinion polls. The problem is also of great importance in campaign work. Campaigns frequently seek to target voters who are most likely to have supported their party in the prior presidential election. Estimates of prior party support may also serve as predictor variables in models estimating support in successive elections. Recalibrating the scores to match observed outcomes is thus a crucial step to improve the scores’ accuracy and bolster future electioneering.

A common heuristic solution to the recalibration problem is the use of the “uniform swing” (Butler, 1951) on the logit scale. This approach is simple: first, one defines the function

$$h(\alpha) = \sum_{i \in \mathcal{V}} \frac{1}{1 + \frac{1-p_i}{p_i} \alpha},$$

and, having observed a total D , one finds the α that satisfies the equation

$$h(\alpha) = D. \tag{1}$$

The function $h(\cdot)$ is monotonic in α , so Equation 1 can be solved in logarithmic time using binary search. The updated scores are then computed as

$$\tilde{p}_i = \frac{1}{1 + \frac{1-p_i}{p_i} \alpha}.$$

Solving Eq. 1 is equivalent to finding the set of probabilities \tilde{p}_i which sum to D and minimize the Kullback–Leibler divergence (Kullback and Leibler, 1951) with the distribution induced by the original scores p_i . Moreover, if the p_i are defined based on a logistic regression, then this update is equivalent to shifting the intercept in the model by $\log(\alpha)$. For more details on these characterizations, see the Appendix, Section A.

Examples of this simple recalibration strategy are given by Ghitza and Gelman (2013), Hanretty et al. (2016), and Ghitza and Gelman (2020). This procedure is also familiarly referred to by campaign workers as the “logit shift”.¹

In this research note, we provide analytical justification for the logit shift. First, we introduce an alternative procedure for score updating, which simply computes the updated scores as posterior probabilities, conditional on the target totals. In this procedure, we assume the original scores p_i capture a kind of prior Democratic support probability, while the updated scores \tilde{p}_i reflect the conditional Democratic voting probability given observed outcomes. Next, we show that this second, more principled approach is well approximated by the logit shift in large samples. We demonstrate this result analytically and illustrate it in a small simulation study. Finally, we discuss potential extensions to cases where a uniform swing is insufficient to capture observed electoral dynamics.

¹The term “logit swing” is also commonly used.

2 Recalibration as a posterior update

To motivate the posterior update approach, we introduce some additional notation. We define each voter’s choice as a binary variable $W_i \in \{0, 1\}$, where $W_i = 1$ signifies a Democratic vote and $W_i = 0$ signifies a Republican vote (we suppose a two-candidate election for simplicity). The W_i are modeled as independent Bernoulli random variables, where $W_i \sim \text{Bern}(p_i)$. In other words, the $p_i = P(W_i = 1)$ can be thought of as the prior, unconditional probability of casting a Democratic vote. In this model, it is straightforward to approach score recalibration by defining a new set of updated scores, $\{p_i^*\}$, using the following conditional probability (which automatically sum to D over voters i):

$$\begin{aligned} p_i^* &= \mathbb{P}\left(W_i = 1 \mid \sum_{j \in \mathcal{V}} W_j = D\right) \\ &= \frac{\mathbb{P}\left(W_i = 1, \sum_{j \in \mathcal{V}} W_j = D\right)}{\mathbb{P}\left(\sum_{j \in \mathcal{V}} W_j = D\right)} \\ &= p_i \times \xi_i \end{aligned} \tag{2}$$

where $\xi_i = \frac{\mathbb{P}(\sum_{j \neq i} W_j = D-1)}{\mathbb{P}(\sum_j W_j = D)}$ is a ratio of two Poisson-Binomial probabilities — that is, probabilities over the sum of independent *but not identically distributed* Bernoulli random variables (Chen and Liu, 1997). Explicit computation of the p_i^* is quite challenging, as efficient computation of Poisson-Binomial probabilities is extremely computationally demanding at even moderate sample sizes, despite substantial recent advances in the literature (Olivella and Shiraito, 2017; Junge, 2020). To compute the p_i^* , we would need to compute one unique Poisson-Binomial probability per unit in the population. Hence, if the number of actual voters $|\mathcal{V}|$ were even modestly large, it would be computationally infeasible to obtain these exact posterior probabilities.

3 Logit shift approximates the correct posterior

3.1 Preliminaries

In this section, we show analytically why the logit shift is a good approximation to the general posterior update in Eq. 2. To do so, we begin by defining two terms, the ratio $\phi_i = \frac{\mathbb{P}(\sum_{j \neq i} W_j = D)}{\mathbb{P}(\sum_{j \neq i} W_j = D-1)}$, and the function $f(x, s) = \frac{1}{1 + \frac{1-x}{x}(s)}$.

Simple substitution, along with a useful recursive property of the Poisson-Binomial dis-

tribution,² makes it clear that

$$\begin{aligned}
\sum_i f(p_i, \phi_i) &= \sum_i \frac{1}{1 + \frac{1-p_i}{p_i} \phi_i} \\
&= \sum_i \frac{1}{1 + \frac{1-p_i}{p_i} \frac{\mathbb{P}(\sum_{j \neq i} W_j = D)}{\mathbb{P}(\sum_{j \neq i} W_j = D-1)}} \\
&= \sum_i \frac{p_i \times \mathbb{P}(\sum_{j \neq i} W_j = D-1)}{p_i \times \mathbb{P}(\sum_{j \neq i} W_j = D-1) + (1-p_i) \times \mathbb{P}(\sum_{j \neq i} W_j = D)} \quad (3) \\
&= \sum_i \frac{\mathbb{P}(W_i = 1, \sum_i W_i = D)}{\mathbb{P}(\sum_i W_i = D)} \\
&= \sum_i p_i^* \\
&= D
\end{aligned}$$

In words, Eq. 3 shows that ϕ_i is precisely the “shift” that turns each p_i into the desired, recalibrated posterior probability p_i^* . The logit shift, however, uses a constant α to approximate the vector of recalibrating shifts $\{\phi_i\}_{i \in \mathcal{V}}$. What remains, therefore, is to show that the value of α that solves Eq. 1 is a very good approximation of ϕ_i for all values of i .

To do so, we establish a couple of facts: that the value of α is bounded by the range of $\{\phi_i\}_{i \in \mathcal{V}}$, and that each ϕ_i in turn has well-defined bounds:

Theorem 1. *The value of α which solves Equation 1 satisfies:*

$$\min_i \frac{\mathbb{P}(\sum_{j \neq i} W_j = D)}{\mathbb{P}(\sum_{j \neq i} W_j = D-1)} \leq \alpha \leq \max_i \frac{\mathbb{P}(\sum_{j \neq i} W_j = D)}{\mathbb{P}(\sum_{j \neq i} W_j = D-1)}.$$

Proof. The proof can be found in the Appendix, Section B. □

Theorem 2. *For any choice of $i \in \mathcal{V}$, we have*

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

Proof. The proof can be found in the Appendix, Section C. □

²Namely,

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

3.2 Main Results

The bounds from Theorem 2 apply regardless of the choice of i , so we can combine the two theorems to find that

$$\begin{aligned} \frac{\mathbb{P}\left(\sum_{j \in \mathcal{V}} W_j = D + 1\right)}{\mathbb{P}\left(\sum_{j \in \mathcal{V}} W_j = D\right)} &\leq \min_i \frac{\mathbb{P}\left(\sum_{j \neq i} W_j = D\right)}{\mathbb{P}\left(\sum_{j \neq i} W_j = D - 1\right)} \leq \alpha \\ &\leq \max_i \frac{\mathbb{P}\left(\sum_{j \neq i} W_j = D\right)}{\mathbb{P}\left(\sum_{j \neq i} W_j = D - 1\right)} \leq \frac{\mathbb{P}\left(\sum_{j \in \mathcal{V}} W_j = D\right)}{\mathbb{P}\left(\sum_{j \in \mathcal{V}} W_j = D - 1\right)}. \end{aligned} \tag{4}$$

This is useful, because we can now use the outer bounds in Eq. 4 to obtain a bound on the approximation error when estimating recalibrated scores p_i^* (obtained from the posterior update approach) via \tilde{p}_i (obtained from the logit shift):

Theorem 3. *For large sample sizes, we obtain*

$$\tilde{p}_i = p_i^* + \mathcal{O}\left(\frac{1}{\sum_{j \in \mathcal{V}} p_j(1 - p_j)}\right).$$

Proof. The proof can be found in the Appendix, Section D. □

Theorem 3 states that the error in using the logit shift approach to approximate the posterior recalibration update depends on the precision of the Poisson-Binomial distribution over sums of binary outcomes being aggregated (votes for the Democratic candidate, in our running example). As the variance of Poisson-Binomial deviates is maximal when all underlying probabilities are equal to 0.5, it follows that, in our running example, the approximation will perform best when voters in \mathcal{V} are equally likely to vote for either party. As voters become more heterogenous, or as their support becomes more lopsided (or both, as would be the case in heavily polarized electorates), the quality of the approximation is expected to suffer. Fortunately, the binding bounds in Eq. 4 are extremely tight for large enough samples, so that even in the worst case-scenarios, the approximation can be expected to perform well. We now briefly illustrate our analytical results with a small Monte-Carlo simulation.

4 Simulations

We conduct a brief simulation study to empirically demonstrate the efficacy of this approach. We simulate using 1,000 units, a sample size at which computations of the true p_i^* are possible. We draw the initial probabilities p_i according to the six distributions discussed in Biscarri et al. (2018). We then consider the cases in which the observed D is either 20% above or 20% below the expectation, $\sum_i p_i$.

We compute the true probabilities using Biscarri’s algorithm as implemented in the *PoissonBinomial* package (Junge, 2020), and compare it against the estimates obtained using our heuristic method. We report the RMSE as well as the proportion of variance in the true p_i^* that is *not* explained by our method. Results are given in Table 4. Across all settings, our approximations perform extremely well.

p_i Setting	Sampling Distribution	Observed D	RMSE	$1 - R^2$
Uniform	Uniform(0, 1)	-20%	0.00022	5.58×10^{-7}
Uniform	Uniform(0, 1)	20%	0.00021	5.46×10^{-7}
Close to Zero	Beta(0.1, 3)	-20%	0.00047	3.41×10^{-5}
Close to Zero	Beta(0.1, 3)	20%	0.00039	1.85×10^{-5}
Close to One	Beta(3, 0.1)	-20%	0.00036	1.22×10^{-6}
Close to One	Beta(3, 0.1)	20%	-	-
Extremal	$0.5 * \text{Beta}(0.1, 3) + 0.5 * \text{Beta}(3, 0.1)$	-20%	0.00048	1.14×10^{-6}
Extremal	$0.5 * \text{Beta}(0.1, 3) + 0.5 * \text{Beta}(3, 0.1)$	20%	0.00048	1.12×10^{-6}
Central	Beta(3, 3)	-20%	0.00015	6.86×10^{-7}
Central	Beta(3, 3)	20%	0.00015	6.86×10^{-7}
Bimodal	$0.5 * \text{Beta}(3, 10) + 0.5 * \text{Beta}(10, 3)$	-20%	0.00023	6.92×10^{-7}
Bimodal	$0.5 * \text{Beta}(3, 10) + 0.5 * \text{Beta}(10, 3)$	20%	0.00023	6.86×10^{-7}

Table 1: Approximation error, as measured by RMSE and $1 - R^2$, using our heuristic method vs. the true Poisson-Binomial probabilities, under various settings. No results are reported in the row in which $1.2 \times \sum_i p_i$ would exceed the sample size of 1,000.

5 Discussion

In this paper, we have considered the problem of updating voter scores to match observed vote totals from an election. We have shown that the relatively simple “logit shift” algorithm is a very good approximation to computing the true conditional probability. This is an especially useful insight for campaign analysts and researchers alike, because the logit shift is significantly more efficient computationally than the calculation of the exact posterior recalibration update.

It is worth being explicit about the limitations of this approach. Under the posterior update model, we treat the original scores p_i as a prior over Democratic vote probability. In turn, the updated scores p_i^* deviate from the initial scores only by assuming the observed vote tallies deviate from the expectation of $\sum_i p_i$ due to random error. Crucially, the updated probabilities retain the same ordering as the prior probabilities, which implies the original scoring model must discriminate positive and negative (but unobservable, in the case of voting) individual cases well. It is also important to note that the realization of \mathcal{V} over which values of D are defined can have an impact on the quality of the approximation: the approximation will be better when the number of Democratic votes D tallies the choices of voters in very competitive districts than when it tallies votes in landslide ones, and choosing a level of aggregation with too few voters in it could render the error bounds too loose. In most practical instances, however, the logit shift can be expected to perform very well.

Hence, this approach represents a useful – albeit crude – method of updating individual-level scores to incorporate information from a completed election. More complex insights about the electorate, such as the marked underperformance of Democrats among Hispanics voters in the 2020 election, cannot be directly incorporated by computing the posterior probabilities (or their approximation via the logit shift). Methods based on ecological inference (e.g. King et al., 2004) would be necessary to capture this structure. Such methods represent a promising potential extension of the insights provided in this manuscript.

References

- Biscarri, W., Zhao, S. D., and Brunner, R. J. (2018). A simple and fast method for computing the poisson binomial distribution function. *Computational Statistics & Data Analysis*, 122:92–100.
- Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.
- Butler, D. E. (1951). Appendix to the british general election of 1950. ed. h. g. nichols.
- Chen, S. X. and Liu, J. S. (1997). Statistical applications of the poisson-binomial and conditional bernoulli distributions. *Statistica Sinica*, pages 875–892.
- Ghitza, Y. and Gelman, A. (2013). Deep interactions with mrp: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, 57(3):762–776.
- Ghitza, Y. and Gelman, A. (2020). Voter registration databases and mrp: Toward the use of large-scale databases in public opinion research. *Political Analysis*, 28(4):507–531.
- Hanretty, C., Lauderdale, B., and Vivyan, N. (2016). Combining national and constituency polling for forecasting. *Electoral Studies*, 41:239–243.
- Junge, F. (2020). Package ‘poissonbinomial’. *Computational Statistics & Data Analysis*, 59:41–51.
- King, G., Tanner, M. A., and Rosen, O. (2004). *Ecological inference: New methodological strategies*. Cambridge University Press.
- Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.
- Olivella, S. and Shiraito, Y. (2017). poisbinom: A faster implementation of the poisson-binomial distribution. r package version 1.0. 1.
- Schwenzfeier, M. (2019). Which non-responders drive non-response bias? In *PolMeth XXXVI*, Cambridge, MA.
- Siripraparat, T. and Neammanee, K. (2021). A local limit theorem for poisson binomial random variable. *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.

A Characterizations of the Logit Shift

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

Define the minimum-KL-divergence optimization problem as

$$\begin{aligned}
& \text{minimize} && \sum_{i \in \mathcal{V}} -\tilde{x}_i \log\left(\frac{p_i}{\tilde{x}_i}\right) - (1 - \tilde{x}_i) \log\left(\frac{1 - p_i}{1 - \tilde{x}_i}\right) \\
& \text{subject to} && \sum_{i \in \mathcal{V}} \tilde{x}_i = D, \quad 0 \leq \tilde{x}_i \leq 1 \text{ for } i \in \mathcal{V}.
\end{aligned} \tag{5}$$

The Lagrangian for Optimization Problem 5 is

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

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

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

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 x_i satisfying $0 < x_i < 1$ for $i \in \mathcal{V}$ and $\sum_{i \in \mathcal{V}} x_i = D$, strong duality is attained. Hence, our point is optimal and the \tilde{p}_i are a solution to Optimization Problem 5. For background technical details, see Boyd et al. (2004).

Next, we show that this procedure is equivalent to an intercept shift if the p_i are defined based on a logistic regression, i.e.

$$p_i = \frac{\exp(\beta_0 + X_i \beta)}{1 + \exp(\beta_0 + X_i \beta)}.$$

where $\beta_0 \in \mathbb{R}$ is the intercept, $X_i \in \mathbb{R}^p$ is the covariate vector for unit i , and $\beta \in \mathbb{R}^p$ is the coefficient vector. Plugging this expression into the definition of \tilde{p}_i , we obtain

$$\begin{aligned}
\tilde{p}_i &= \frac{1}{1 + \exp(-\beta_0 - X_i \beta) \alpha} \\
&= \frac{1}{1 + \exp(-(\beta_0 - \log(\alpha)) - X_i \beta)} \\
&= \frac{\exp(\beta_0 - \log(\alpha) + X_i \beta)}{1 + \exp(\beta_0 - \log(\alpha) + X_i \beta)},
\end{aligned}$$

where on the last line we see that \tilde{p}_i is formulated as a logistic regression with an identical coefficient vector and its intercept shifted by $-\log(\alpha)$.

B Proof of Theorem 1

For a fixed choice of \mathbf{x} , observe that $g(\mathbf{x}, \mathbf{s})$ is monotonically decreasing in every component of \mathbf{s} . Denote $\boldsymbol{\alpha} = \{\alpha\}_{i \in \mathcal{V}}$, the vector repeating α a total of $|\mathcal{V}|$ times. Because

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

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

C 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

$$\mathbb{P} \left(\sum_{j \neq i} W_j = D - 2 \right) \mathbb{P} \left(\sum_{j \neq i} W_j = D \right) \leq \mathbb{P} \left(\sum_{j \neq i} W_j = D - 1 \right)^2 \quad (6)$$

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

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

Collecting terms, we get

$$\begin{aligned} \mathbb{P} \left(\sum_{j \neq i} W_j = D \right) \left(p_i \mathbb{P} \left(\sum_{j \neq i} W_j = D - 2 \right) + (1 - p_i) \mathbb{P} \left(\sum_{j \neq i} W_j = D - 1 \right) \right) &\leq \\ \mathbb{P} \left(\sum_{j \neq i} W_j = D - 1 \right) \left(p_i \mathbb{P} \left(\sum_{j \neq i} W_j = D - 1 \right) + (1 - p_i) \mathbb{P} \left(\sum_{j \neq i} W_j = D \right) \right) \end{aligned} \quad (7)$$

The terms in parentheses can be collapsed into a single Poisson Binomial probability, making use of the recursion defined in Footnote 1. Subbing these expressions into Inequality 7, 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 6 and following the same set of steps.

D Proof of Theorem 3

Under the bounds defined in Inequality 4 in the main text, we obtain the following approximation bounds for α_i :

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

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

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

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 $\mathbb{P}(\sum_{j \in \mathcal{V}} W_j = d)$ is bounded above by C_1/σ^2 for a constant $C_1 > 0$. Hence

$$\begin{aligned} \frac{|\alpha - \phi_i|}{\phi_i} &\leq \frac{\psi(D)^2}{\psi(D-1)\psi(D+1)} - 1 + \mathcal{O}\left(\frac{1}{\sigma^2}\right) \\ &= \exp\left(\frac{1}{\sigma^2}\right) - 1 + \mathcal{O}\left(\frac{1}{\sigma^2}\right) = \mathcal{O}\left(\frac{1}{\sigma^2}\right) \end{aligned} \tag{8}$$

Lastly, we observe

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

where the last line follows by plugging in the bound on $(\alpha - \phi_i)/\phi_i$ from Inequality 8 and observing

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