Supplementary Information:
Pre-Registration Report: Facing Election

Mathias Rask[*] Stig Hebbelstrup[†] Robert Klemmensen[‡]

November 16, 2021

[*] Aarhus University, Political Science. Email: mathiasrask@ps.au.dk.

[†] Aarhus University, Political Science. Email: stighebbelstrup@ps.au.dk.

[‡] University of Southern Denmark, Political Science. Email: rkl@sam.sdu.dk.

A — PAC Bayesian Aggregation

As put forward in the main pre-registration report, we use a PAC-Bayesian approach as a supplementary approach to analyze the generalizability of our expectation that facial images to a certain extent can predict the electoral success of political candidates. PAC stands for **P**robably **A**pproximately **C**orrect (Valiant 1984) and is a learning framework to derive models that with high probability (the "probably" aspect) have low generalization error ("the approximately correct" aspect). As mentioned in the main manuscript for the pre-registration, we expect that our learning task is difficult due to a low signal-to-noise ratio combined with a fairly small sample size. Therefore, we risk using a unstable model to investigate whether our approach generalizes to the 2021 data. PAC-Bayesian learning is, however, suitable to better cope with instability and variance as it provides more freedom and control over the bias-variance trade-off.

PAC-learning is based on *randomized classifiers* defined by a distribution ρ over a hypothesis class $\mathcal{H}$. A randomized classifier is also called a *Gibbs classifier* (Germain et al. 2009) or a *stochastic classifier* (Germain et al. 2015), but the terms refer to the same idea; in each prediction round (i.e. for each image in the set of 2021 images), draw a hypothesis h from $\mathcal{H}$ according to the posterior probability ρ (defined over $\mathcal{H}$) and use h to predict the image. This is repeated for all images in the 2021 data set meaning that multiple models are used to generate the predictions, but with various probabilities defined by ρ. This is in contrast to the typical approach to prediction where one would use the single best model $\hat{h}^*_{\mathbb{S}}$ trained and validated on $\mathbb{S}^{\text{train}}$ and $\mathbb{S}^{\text{val}}$ respectively to predict all images in the 2021 data set.

We combine the PAC-Bayesian analysis with a ρ-weighted majority vote to reduce the impact of potential high variance in model performance in our prediction of the electoral success on the 2021 data. That is, we weight the predictions with the posterior probability ρ to assign more confidence to "good" hypotheses and less confidence to "bad" hypotheses. We refer to this approach as *PAC-Bayesian aggregation*. This procedure is generally *ensemble learning*; the aggregation of multiple classifiers h through a majority vote. Ensemble learning is commonly used in machine (Friedman et al. 2001) and deep learning (Xie et al. 2013) to reduce both bias and variance in the generalization. For instance, methods such as Bagging, Boosting, and Random Forests are well-known examples of techniques where the prediction is an output of a majority vote. Majority vote prediction has also been widely used in Bayesian approaches where the predictions are weighted by the posterior probability ρ defined over the hypothesis class $\mathcal{H}$ (Germain et al. 2009, 2015; Lacasse et al. 2006; Roy et al. 2011; Thiemann et al. 2017). Intuitively, the improvement in model performance by ensemble learning comes from the "cancellation of errors" effect: If the individual errors are uncorrelated and if the models are better than random guessing ($L(h) < 0.5$ or equivalently accuracy $> .5$ for $0/1$ loss), the errors cancel out yielding less bias and less variance in the predictions. Formally, for a single input (i.e. a single image), we define the ρ-weighted majority vote as:

$$\mathsf{MV}_\rho(X) = \operatorname{sign}\left(\sum_{h\in\mathcal{H}} \rho(h)h(X)\right)$$

which is then repeated for each input image X in the 2021 test set $\mathcal{T} = \{X_1, \ldots, X_n\}$ where n denotes the total number of samples in the 2021 data.

A.1. PAC-Bayes-λ bound

To select the distribution ρ over the hypothesis set $\mathcal{H}$, we use the approach set out by Thiemann et al. (2017). The points and explanations made below are a compact summary of the details found in their paper. Based on the PAC-Bayes-kl inequality proposed by Seeger (2002), Thiemann et al. (2017) derive a PAC-Bayesian bound they coin a *PAC-Bayes-λ bound* or a *PAC-Bayes-λ inequality* that is convex in the posterior distribution ρ and with distribution-free assumptions about both the prior and posterior distributions, π and ρ. The bound uses refined Pinsker's inequality (Cover 1999) to obtain a bound that, in addition to the posterior ρ, involves an additional trade-off parameter λ:

$$\mathbb{E}_\rho \left[L(h)\right] \leq \frac{\mathbb{E}_\rho \left[\hat{L}(h,S)\right]}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}\left(\rho \| \pi\right) + \ln \frac{2\sqrt{n}}{\delta}}{\lambda\left(1-\frac{\lambda}{2}\right)n} \tag{1}$$

Importantly, the bound in (1) is convex in ρ for a fixed λ and convex in λ for fixed ρ. Consequently, the bound can be minimized through alternating minimization of the two parameters, ρ and λ. The bound holds for any (prior) distribution π over $\mathcal{H}$ which is independent of the training data S and *for all* (posterior) distributions ρ over $\mathcal{H}$ and $\lambda \in (0,2)$ simultaneously. Although the bound holds for any prior, researchers often use an uniform distribution such that $\pi(h) = \frac{1}{m}$ such that the confidence is equally distributed across the m hypotheses *before* training the models. Consequently, for e.g. $m = 100$, each hypothesis is assigned a weight proportional to 0.01 such that the prior satisfies $\sum_{h\in\mathcal{H}} \pi(h) \leq 1$. The δ in the nominator of the second term on the right-hand side of the inequality controls the probability by which the bound yields generalization guarantees. In its current form in (1), the bound is read as: for $\delta \in (0,1)$ the bound holds with probability greater than $1-\delta$. We use $\delta = 0.05$ when minimizing the bound.

Since the bound is convex in ρ for a fixed λ and convex in λ for fixed ρ, it implies that that we can find the optimal ρ and λ[1] by minimizing the bound w.r.t. ρ and λ. By re-arranging the inequality, fixing λ, and since $\mathbb{E}_\rho \left[\hat{L}(h,S)\right]$ is linear in ρ and $\mathrm{KL}\left(\rho \| \pi\right)$ is convex in ρ, the optimal posterior distribution ρ is obtained by

[1] λ is of secondary interest to us as we are primarily interested in how to weight the ensemble of m classifiers through ρ.

$$\rho(h) = \frac{\pi(h) e^{-\lambda n \hat{L}(h,S)}}{\mathbb{E}_\pi \left[e^{-\lambda n \hat{L}(h,S)} \right]} \tag{2}$$

Similarly, for a fixed ρ the bound in (1) is convex λ. Therefore, minimizing w.r.t. to λ yields:

$$\lambda = \frac{2}{\sqrt{\frac{2n\mathbb{E}_\rho\left[\hat{L}(h,S)\right]}{\left(\mathrm{KL}(\rho\|\pi)+\ln\frac{2\sqrt{n}}{\delta}\right)}+1}+1} \tag{3}$$

By iteratively using the equations in (2) and (3) to update the values of ρ and λ, the bound is minimized until a specified convergence criteria is met. We terminate the minimization when the change in the bound is less than $10\mathrm{e}{-16}$.

A.2. Defining the hypothesis space

As noted by Thiemann et al. (2017), computation of the denominator in (2) is prohibitive if the hypothesis set $\mathcal{H}$ is infinite[2]. However, the authors present a straightforward approach to recover a finite hypothesis set. Specifically, the authors train m models on r randomly drawn samples from S with a total of n points and validated on the remaining $n - r$ points. This means that the hypothesis set $\mathcal{H}$ has cardinality $\mid m \mid = 100$ (if using 100 hypotheeis), which obviously is finite. This approach is reminiscent of cross-validation, but the approach suggested by Thiemann et al. (2017) does not require the m different validation sets to non-overlapping and does not require no overlaps between the training and validation set.

The finite construction of $\mathcal{H}$ has implications for the specific definitions of the equations in (1), (2), and (3). For a finite hypothesis set $\mathcal{H}$, the bound and optimal ρ and λ become:

[2] Note that the bound also holds for uncountable infinite hypothesis sets.

$$\mathbb{E}_{\rho}\left[L(h)\right] \leq \frac{\mathbb{E}_{\rho}\left[\hat{L}^{\texttt{val}}(h,S)\right]}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}\left(\rho\|\pi\right) + \ln\frac{2\sqrt{n-r}}{\delta}}{\lambda\left(1-\frac{\lambda}{2}\right)(n-r)} \tag{4}$$

$$\rho(h) = \frac{\pi(h)e^{-\lambda(n-r)\left(\hat{L}^{\texttt{val}}(h,s)-\hat{L}^{\texttt{val}}_{\texttt{min}}\right)}}{\sum_{h'}\pi(h')e^{-\lambda(n-r)\left(\hat{L}^{\texttt{val}}(h,S)-\hat{L}^{\texttt{val}}_{\texttt{min}}\right)}} \tag{5}$$

$$\lambda = \frac{2}{\sqrt{\frac{2(n-r)\mathbb{E}_{\rho}\left[\hat{L}^{\texttt{val}}(h,S)\right]}{\left(\mathrm{KL}(\rho\|\pi)+\ln\frac{2\sqrt{n-r}}{\delta}\right)}+1}+1} \tag{6}$$

A.3. Implementation of PAC-Bayesian Aggregation

To overcome potential small-sample issues in our learning task, we thus train m classifiers and weight their predictions by minimizing the PAC-Bayes-λ bound to obtain the optimal weights ρ. For each classifier m, we randomly split our sample S into S^{train} and S^{val} with r points in S^{train} and $n-r$ points in S^{val} where n is the total number samples in S. We used oversampling for S^{train} to reduce the class imbalance. See XX for details about our sampling strategy.

For each m, we then trained a CNN classifier using the architecture set out in section XX. To show the robustness of our approach, we randomly selected the hyperparameters for each classifier from the following search spaces:

Hyperparameter	Search space
`transfer learning`	$\in$ {`feature extractor, fine-tune`}
`learning rate`	$\in \{1\mathrm{e}{-}04^{*}, 1\mathrm{e}{-}05, 1\mathrm{e}{-}06\}$
`batch size`	$\in \{16, 32, 64\}$
`dropout`	$\in \{0.4, 0.5, 0.6\}$

* Only for female candidates

with a total of $3 \cdot 3 \cdot 3 \cdot 2 = 54$ model combinations for women and 36 for men. After fitting the models, we then computed the weighting of the classifiers (i.e. ρ) by alternating minimization of the bound in (4). We then apply the weighting scheme computed by minimizing the bound to predict the electoral success on the 2021 local elections.

References

Cover, Thomas M (1999). *Elements of information theory*. John Wiley & Sons.

Friedman, Jerome, Trevor Hastie, Robert Tibshirani, et al. (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York.

Germain, Pascal et al. (2009). “PAC-Bayesian learning of linear classifiers”. In: *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 353–360.

Germain, Pascal et al. (2015). “Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm”. In: *arXiv preprint arXiv:1503.08329*.

Lacasse, Alexandre et al. (2006). “PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier”. In: *NIPS*, pp. 769–776.

Roy, Jean-Francis, François Laviolette, and Mario Marchand (2011). “From PAC-Bayes bounds to quadratic programs for majority votes”. In: *ICML*.

Seeger, Matthias (2002). “PAC-Bayesian generalisation error bounds for Gaussian process classification”. In: *Journal of machine learning research* 3.Oct, pp. 233–269.

Thiemann, Niklas et al. (2017). “A strongly quasiconvex PAC-Bayesian bound”. In: *International Conference on Algorithmic Learning Theory*. PMLR, pp. 466–492.

Valiant, Leslie G (1984). “A theory of the learnable”. In: *Communications of the ACM* 27.11, pp. 1134–1142.

Xie, Jingjing, Bing Xu, and Zhang Chuang (2013). “Horizontal and vertical ensemble with deep representation for classification”. In: *arXiv preprint arXiv:1306.2759*.