Improving Aggregated Forecasts of Probability*

Guanchun Wang, Sanjeev Kulkarni, H.Vincent Poor
Department of Electrical Engineering
Princeton University
{guanchun, kulkarni, poor}@princeton.edu

Daniel N. Osherson
Department of Psychology
Princeton University
osherson@princeton.edu

***Abstract*—The Coherent Approximation Principle (CAP) is a method for aggregating forecasts of probability from a group of judges by enforcing coherence with minimal adjustment. This paper explores two methods to further improve the forecasting accuracy within the CAP framework and proposes practical algorithms that implement them. These methods allow flexibility to add fixed constraints to the coherentization process and compensate for the psychological bias present in probability estimates from human judges. The algorithms were tested on a data set of nearly half a million probability estimates of events related to the 2008 U.S. presidential election (from about 16000 judges). The results show that both methods improve the stochastic accuracy of the aggregated forecasts compared to using simple CAP.**

I. Introduction

In many situations, a number of human judges may be asked to provide their subjective estimates of the chance that certain events will happen. Applications of this approach can be found in different fields such as data mining, economics, finance, geopolitics, meteorology and sports (see [1], [2] for surveys). In many cases, this will involve subjective estimates of the chance of both simple and complex events. Such forecasts are useful when judges have information about the likely co-occurrence of events, or the probability of one event conditional upon another.[1] Including complex events in queries to judges may thus potentially improve the accuracy of aggregate forecasts.

One challenge for this setup is how to enforce the probabilistic coherence of the aggregated results. Simple aggregation methods like linear averaging may not yield a coherent aggregate, because human judges often violate probability axioms (e.g., [3], [4]) and the linear average of incoherent forecasts is generally also incoherent. Moreover, even if all the judges are individually coherent, when the forecasts of a given judge concern only a subset of events (because of specialization), the averaged results may still be incoherent. To address the foregoing limitations, a generalization of linear averaging was discussed in [5] and [6], and is known as the Coherent Approximation Principle (CAP). CAP proposes a coherent forecast that is minimally different, in terms of squared deviation, from the judges' forecasts.

To fully test the effectiveness of CAP and various associated algorithms, it is of interest to make comparisons using a large data set. The 2008 U.S. presidential election provided an opportunity to elicit a very large pool of judgments. In the months prior to the election, we established a website to collect probability estimates of election related events. Each respondent was presented 28 questions concerning election outcomes involving 7 randomly selected states, and was asked to estimate the probability of each outcome. For example, a user might be asked questions about simple events such as "What is the probability that Obama wins Indiana?" and also questions about complex events like "What is the probability that Obama wins Vermont and McCain wins Texas?" or "What is the probability that McCain wins Florida supposing that McCain wins Maine?" The respondents provided estimates of these probabilities with numbers from zero to one hundred percent.

Nearly sixteen thousand respondents completed the survey. Having a large number of judges allows us to tap into the collective wisdom, but the size of the data set presents computational challenges. More subtly, it also raises the issue of how to reduce the systematic bias that a large group of human judges may have so that we can further improve the forecasting accuracy of the events. Hence our goal is to develop algorithms that can efficiently implement (a relaxed version of) CAP and at the same time allow such adjustments.

The remainder of the paper is organized as follows. In §2, we introduce notation. Then we review CAP and a scalable algorithm for its approximation. In §3, we propose two algorithms that solve CAP with hard constraints and allow bias-adjustment of the original

*This research was supported in part by the NSF Science & Technology Center grant CCF-0939370, in part by the U.S. Army Research Office under Grant W911NF-07-1-0185, and in part by the Qatar National Research Fund under Grant NPRP 08-522-2-211. Daniel Osherson acknowledges the Henry Luce Foundation.

[1] Example events might include "President Obama will be re-elected in 2012", "Obama will be re-elected if the U.S. unemployment rate drops below 8% by the end of 2011".

forecasts within CAP, respectively. Brier score ([7]) and some other popular measures of forecasting accuracy are introduced in §4, and are used to compare different aggregation methods. We conclude in §5 with a discussion of implications and extensions.

II. The scalable approach of CAP

A. Coherent Approximation Principle

Let $\mathbf{\Omega}$ be a finite *outcome space* so that subsets of $\mathbf{\Omega}$ are *events*. A forecast is defined to be a mapping from a set of events to estimated probability values, i.e., $f : \mathcal{E} \to [0,1]^n$, where $\mathcal{E} = \{E_1, \dots, E_n\}$ is a collection of events. Also we let $\mathbf{1}_E : \mathbf{\Omega} \to \{0,1\}$ denote the indicator function of an event E. We distinguish two types of events: simple events, and complex events formed from simple events using basic logic and conditioning. As subjective probability estimates of human judges are often incoherent, it is common to have incoherence within a single judge and among a panel of judges. To compensate for the inadequacy of linear averaging when incoherence is present, the Coherent Approximation Principle was proposed in [6] with the following definition of probabilistic coherence.

Definition 1. *A forecast f over a set of events $\mathcal{E}$ is probabilistically coherent if and only if it conforms to some probability distribution on $\mathbf{\Omega}$, i.e., there exists a probability distribution f' on $\mathbf{\Omega}$ such that $f(E) = f'(E)$ for all $E \in \mathcal{E}$.*

With a panel of judges each evaluating a (potentially different) set of events, CAP achieves coherence with minimal modification of the original judgments, which can be mathematically formulated as the following optimization problem:

$$\begin{aligned} &\min \sum_{i=1}^{m} \sum_{E \in \mathcal{E}_i} (f(E) - f_i(E))^2 \\ &\textit{s.t. } f \text{ is coherent.} \end{aligned} \tag{1}$$

Here we assume a panel of m judges, where $\mathcal{E}_i$ denotes the set of events evaluated by judge i; the forecasts $\{f_i\}_{i=1}^m$ are the input data and f is the output of (1), which is a coherent aggregate forecast for the events in $\mathcal{E} = \vee_{i=1}^m \mathcal{E}_i$.

B. A Scalable Approach

Although CAP can be framed as a constrained optimization problem with $|\mathcal{E}|$ optimization variables, it can be computationally infeasible to solve using standard techniques when there is a large number of judges forecasting a very large set of events (e.g., our election data set). In addition, the nonconvexity introduced by the ratio equality constraints from conditional probabilities can lead to the existence of local minima. In [8], the concept of local coherence was introduced, motivated by the fact that the logical complexity of events that human judges can assess is usually bounded, typically not going beyond a combination of three simple events or their negations. Hence, the global coherence constraint can be well approximated by sets of local coherence constraints, which in turn allows the optimization problem to be solved using the Successive Orthogonal Projection (SOP) algorithm (see [9] for related material). Below, we provide the definition of local coherence and the formulation of the optimization program.

Definition 2. *Let $f : \mathcal{E} \to [0,1]$ be a forecast and let $\mathcal{F}$ be a subset of $\mathcal{E}$. We say that f is* locally coherent *with respect to the subset $\mathcal{F}$, if and only if f restricted to $\mathcal{F}$ is probabilistically coherent, i.e., there exists a probability distribution g on Ω such that $g(E) = f(E)$ for all $E \in \mathcal{F}$.*

We illustrate via Table I. We see that f is not locally coherent with respect to $\mathcal{F}_1 = \{E_1, E_2, E_1 \wedge E_2\}$ because $f(E_1)+f(E_2)-f(E_1 \wedge E_2) > 1$, while f is locally coherent with respect to $\mathcal{F}_2 = \{E_1, E_2, E_1 \vee E_2\}$ and $\mathcal{F}_3 = \{E_2, E_1 \wedge E_2, E_1 \mid E_2\}$. Note that f is not globally coherent (namely, coherent with respect to $\mathcal{E}$) in this example, as global coherence requires that f be locally coherent with respect to all $\mathcal{F} \subseteq \mathcal{E}$.

$\mathcal{E}$	E_1	E_2	$E_1 \wedge E_2$	$E_1 \vee E_2$	$E_1 \mid E_2$
f	0.8	0.5	0.2	0.9	0.4

TABLE I
Local coherence example

By relaxing global coherence to local coherence with respect to a collection of sets $\{\mathcal{F}_l\}_{l=1}^L$, the optimization problem (1) can be modified to

$$\begin{aligned} \min \quad & \textstyle\sum_{i=1}^{m} \sum_{E \in \mathcal{E}_i} (f(E) - f_i(E))^2 \\ \textit{s.t.} \quad & f \text{ is locally coherent w.r.t. } \mathcal{F}_l \quad \forall l = 1, \dots, L. \end{aligned} \tag{2}$$

We can consider solving this optimization problem as finding the projection onto the intersection of the spaces formed by the L sets of local coherence constraints so that the SOP algorithm fits naturally into the judgment aggregation framework.

The computational advantage of this iterative algorithm is that CAP can now be decomposed into subproblems which require the computation and update of only a small number of variables determined by the local coherence set $\mathcal{F}_l$. So the tradeoff between complexity and speed depends on the selection of the local coherence sets $\{\mathcal{F}_l\}_{l=1}^L$, which is a design choice. Note that when each set includes only one event, the problem degenerates to linear averaging and, on the other hand, when all events are grouped into one single set, the problem becomes the same as requiring global coherence. Fortunately, we can often approximate global coherence using a collection of local coherence sets in which only a small number of events are involved, because complex events are often formed in view of the limited logical capacity of human judges and therefore involve a small number of simple events.

It is also shown in [8] that, regardless of the eventual outcome, the scalable approach guarantees stepwise improvement in stochastic accuracy (excluding forecasts of conditional events) measured by Brier score (also called quadratic penalty), which is defined as follows:

$$BS(f) = \frac{1}{|\mathcal{E}|} \sum_{E \in \mathcal{E}} (\mathbf{1}_E - f(E))^2. \tag{3}$$

So we choose $\{\mathcal{F}_l\}_{l=1}^L$ based on each complex event plus the corresponding simple events to achieve efficient and full coverage of $\mathcal{E}$.

III. Enhancing CAP for a Large Set of Forecasts

The scalable CAP algorithm solves the computational complexity issue. However, when the number of judges and events is large, the reliability of the original estimates can vary considerably among judges and the group forecasts might demonstrate systematic bias. In [10], we addressed this issue by allowing the weighting of judges' forecasts by their credibility (i.e., assigning greater weight to potentially more credible judges). Here, we explore further methods that can improve the coherent aggregated forecasts.

A. CAP with Fixed Constraints (CAP-FC)

In some situations, outside the group of judges, there might be some expert whose specialized knowledge entitles him or her to offer very authoritative and potentially very accurate judgments involving a subset of the events in play. This knowledge might include knowing that one estimate is equal to a constant or within a range, or even more generally, there may be some very credible relationships between estimates of different events. In the context of CAP, we can list these judgments as additional fixed constraints to the original coherent requirements. So we can rewrite (2) in the following form:

$$\begin{array}{ll} \min & \sum_{i=1}^m \sum_{E \in \mathcal{E}_i} (f(E) - f_i(E))^2 \\ \textit{s.t.} & f \text{is locally coherent to } \mathcal{F}_l \quad l = 1, \ldots, L. \\ & h_k(\mathcal{E}_k) = 0, k = 1, \ldots, M. \\ & g_j(\mathcal{E}_j) \leq 0, j = 1, \ldots, N. \end{array} \tag{4}$$

Here h_k represents the kth of the M fixed equality constraints the expert presents and $\mathcal{E}_k$ denotes the set of events involved in h_k. Similarly, g_j represents the jth of the N fixed inequality constraints and $\mathcal{E}_j$ denotes the corresponding set of events.

Since the fixed constraints do not introduce any new variables not covered by the local coherence constraints (i.e., $\mathcal{E}_i, \mathcal{E}_j \subset \mathcal{E}$), we can modify the original algorithm to solve the new optimization problem by enforcing the fixed constraints to be satisfied at the end of each iteration. Assuming h_k and g_j are well defined convex functions, the Alternating Projection theorem guarantees convergence. We summarize the algorithm as follows:

```
Input:    Forecasts $\{f_i\}_{i=1}^m$ and collections of events $\{\mathcal{E}_i\}_{i=1}^m$
Step 1:   Compute $t_i$ and $w_i$ for all judges.
Step 2:   Design $\{\mathcal{F}_l\}_{l=1}^L$ for $l = 1, \cdots, L$.
Step 3:   Let $f_0 = \hat{f}$
            for $t = 1, \cdots, T$
              for $l = 1, \cdots, L$
                $f_t :=$ argmin $\sum_{i=1}^m \sum_{E \in \mathcal{E}_i} (f(E) - f_i(E))^2$
                s.t. $f$ is locally coherent w.r.t. $\mathcal{F}_l$.
              $f_t :=$ argmin $\sum_{i=1}^m \sum_{E \in \mathcal{E}_k \cup \mathcal{E}_j} (f(E) - f_i(E))^2$
              s.t. $h_k = 0$ and $g_j \leq 0$.
Output:   $f_T$
```

TABLE II
The CAP-FC algorithm

In this way, the computation savings are preserved while the fixed constraints are also enforced.

B. CAP with Bias Adjustment (CAP-BA)

Individuals are known to often perceive probabilities differently from the true probabilities. Many phenomena such as favorite-longshot bias in horse race betting, insurance buying and demand for lotto are well documented in the psychology literature. When we compare the estimates for simple events in the election data with those from poll-based aggregators, it is noticeable that judges often assign probabilities smaller than the poll-based ones when these probabilities are relatively small, while doing the opposite for large probabilities.

Based on such empirical evidence, Kahneman and Tversky developed *Prospect Theory* as a psychologically realistic alternative to expected utility theory ([11] and [12]). In expected utility, gambles that yield risky outcomes x_i with probabilities p_i are valued according to $\sum p_i u(x_i)$, where $u(x)$ is the *utility* of outcome x. In prospect theory they are valued by $\sum \pi(p_i) v(x_i - r)$, where $v(x - r)$ is a value function with reference point r and $\pi(p)$ is a function that weights probabilities nonlinearly, overweighting small probabilities and underweighting larger probabilities as in Figure 1. Such a curve resembles a logit function.

In order to compensate for the bias in the probability estimates provided by human judges, we can transform the estimates using a compensatory sigmoid function σ. Since the bias is induced at the individual level, the transformation should be done before coherentization. Therefore we can revise (2) to the following:

$$\begin{array}{ll} \min & \sum_{i=1}^m \sum_{E \in \mathcal{E}_i} (\sigma(f(E)) - \sigma(f_i(E)))^2 \\ \textit{s.t.} & \sigma(f) \text{is locally coherent to } \mathcal{F}_l \quad l = 1, \ldots, L. \end{array} \tag{5}$$

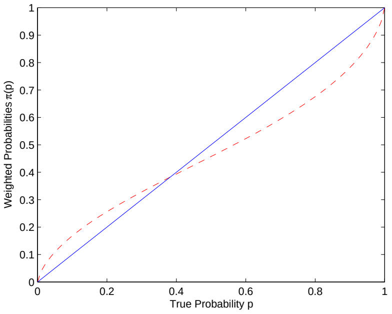

Fig. 1. Weighted probabilities vs. true probabilities

IV. Experimental Results

The data set consists of forecasts from only those judges who completed the questions and provided non-dogmatic (i.e., not 0 or 1) probability estimates for most events. Each participant was given an independent survey based on a randomly chosen set of 7 (out of 50) states. Each respondent was presented with 28 questions. All the questions related to the likelihood that a particular candidate would win a state, but involved negations, conjunctions, disjunctions, and conditionals, along with elementary events. Up to three states could figure in a complex event, e.g., "McCain wins Indiana given that Obama wins Illinois and Obama wins Ohio." Negations were formed by switching candidates (e.g., "McCain wins Texas" was considered to be the complement of "Obama wins Texas"). The respondent could enter an estimate by moving a slider and then pressing the button underneath the question to record his or her answer. Higher chances were reflected by numbers closer to 100 percent; lower chances by numbers closer to 0 percent. Some of the events were of the form X AND Y. The respondents were instructed that these occur only when both X and Y occur. Other events were of the form X OR Y, which occur only when one or both of X and Y occur. The respondent would also encounter events of the form X SUPPOSING Y. In such cases, he or she was told to assume that Y occurs and then give an estimate of the chances that X also occurs based on this assumption.

A. Forecasting Accuracy Measures

Rival aggregation methods were compared in terms of their respective stochastic accuracies. For this purpose, we relied on the Brier score (defined in Equation (3)) along with the following accuracy measures.

- *Good score*: Like the Brier score, the Good score (GS) is a proper scoring rule ([13]), which means the subjective expected penalty is minimized if the judge honestly reports his or her belief of the event. Good score is defined as $-\frac{1}{|\mathcal{E}|}\sum_{E\in\mathcal{E}} \ln|1-\mathbf{1}_E - f(E)|$, where $\mathcal{E}$ denotes the set of all events excluding the conditional events whose conditioning events turn out to be false. GS can take unbounded values when judges are categorically wrong. Here, for the sake of our numerical analysis, we limit the upper bound to 5, since the Good score of an event is 4.6 if a judge is 99% from the truth.
- *Correlation*: We consider the probability estimate as a predictor for the outcome and compute the correlation between it and the true outcome. Note that this is a reward measure, and hence a higher value means greater accuracy, in contrast to the Brier score.
- *Slope*: The slope of a forecast is the average probability of events that come true minus the average of those that do not. Mathematically, it is defined as $\frac{1}{m_T}\sum_{E\in\mathcal{E}:\mathbf{1}_E=1} f(E) - \frac{1}{|\mathcal{E}|-m_T}\sum_{E\in\mathcal{E}:\mathbf{1}_E=0} f(E)$, where m_T denotes the number of true events in $\mathcal{E}$. The slope is also a reward measure.

As usual, conditional events enter the computation of these forecasting accuracy measures only if their conditioning events are true.

B. Aggregation Methods

We now compare the respective stochastic accuracies of the aggregation methods discussed above along with *raw*, i.e., the unprocessed forecasts. Brief explanations of the methods are as follows.

- *Linear*: Replace every estimate for a given event with the unweighted linear average of all the estimates of that event
- *Simple CAP*: Apply the scalable CAP algorithm to eliminate incoherence and replace the original forecasts with the coherentized ones.
- *CAP-FC*: Incorporate fixed constraints into CAP and make sure the aggregated forecast satisfy those constraints. In this experiment, the fixed constraints arose from lopsided prior probabilities. Specifically, we set the probability that Obama wins in S_i equal to 1 for $S_i \in \{$California, New York, Illinois, New Jersey, Massachusetts, Maryland, Connecticut, Hawaii, Rhode Island, Delaware, Vermont$\}$. These states are chosen because they have historically been lopsided blue states and the average estimates for Obama winning in these states are above 0.9.
- *CAP-BA*: Adjust individual estimates for probability bias and then coherentize for aggregated forecasts. In our experiment, we choose $\sigma(f(E)) = \frac{1}{(1+e^{B/2})e^{-Bf(E)}}$, where B is a tuning parameter we set at 10.[2]

[2] A range of different values of B gives very similar results.

C. Comparison Results

Table III summarizes the results, which show nearly[3] uniform improvement in all four accuracy measures (i.e., BS, GS, correlation and slope) from raw to simple linear, to simple (scalable) CAP and, finally, to CAP-BA and CAP-FC. Other than confirming the findings of [6] about *CAP* outperforming *Raw* and *Linear* in terms of Brier score and Slope, we also observe the following:

- CAP-FC outperforms simple CAP with respect to all accuracy measures, which verifies our hypothesis that a good choice of fixed constraints help improve forecasting accuracy;
- CAP-BA also outperforms simple CAP with respect to all accuracy measures, which means the judges did overweight the small probabilities and underweight the large probabilities, hence the improvement from the non-linear transformation .
- CAP-FC, in this case, yields greater forecast accuracy than CAP-BA, which shows the advantage of leveraging credible belief and expertise in selecting events and setting the constraints. The current version of CAP-BA adjusts all individual forecasts uniformly regardless of the complexity of the events; this might not produce the optimal results. In contrast, CAP-FC offers greater flexibility and control to the aggregator.

	Raw	Linear	CAP	CAP-BA	CAP-FC
Brier score	0.105	0.085	0.072	0.061	0.057
Good score	0.347	0.306	0.278	0.243	0.233
Correlation	0.763	0.833	0.879	0.901	0.912
Slope	0.560	0.560	0.586	0.628	0.635

TABLE III
FORECASTING ACCURACY COMPARISON RESULTS

V. Conclusions

In this paper, we have introduced two computationally efficient algorithms that can significantly improve the accuracy of a large set of forecasts under the CAP framework. By allowing additional fixed constraints to be included into the coherentization process, the aggregator enjoys greater flexibility and control to incorporate useful information. This is helpful because in some cases the hard constraints the forecasts need to satisfy might not be enforced by the coherence requirements. Additionally, by performing a non-linear transformation of the original estimates, inherent probability biases can be mitigated so that higher accuracy can be achieved after the coherent aggregation. The election data set provides strong empirical evidence of the effectiveness of these two aggregation methods.

Interesting issues for future study include exploring general criteria for choosing robust fixed constraints and further investigating the mapping between true probabilities and perceived probabilities so as to achieve better bias adjustment. In addition, we can extend the model to allow hard constraints involving conditional events, in which case the lack of convexity might become an issue for solving the optimization problem.

References

[1] R. T. Clemen, “Combining forecasts: A review and annotated bibliography,” *International Journal of Forecasting*, vol. 5, no. 4, pp. 559–583, 1989.

[2] M. G. Morgan and M. Henrion, *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge, UK: Cambridge University Press, 1990.

[3] A. Tversky and D. Kahneman, “The conjunction fallacy: A misunderstanding about conjunction?” *Cognitive Science*, vol. 90, no. 4, pp. 293–315, 1983.

[4] K. Tentori, N. Bonini, and D. Osherson, “Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment,” *Psychological Review*, vol. 28, no. 3, pp. 467–477, 2004.

[5] R. Batsell, L. Brenner, D. Osherson, M. Y. Vardi, and S. Tsavachidis, “Eliminating incoherence from subjective estimates of chance,” in *Proc. 8th Internat. Conf. Principles of Knowledge Representation and Reasoning (KR 2002)*, Toulouse, France, 2002.

[6] D. N. Osherson and M. Y. Vardi, “Aggregating disparate estimates of chance,” *Games and Economic Behavior*, vol. 56, no. 1, pp. 148–173, 2006.

[7] G. W. Brier, “Verification of forecasts expressed in terms of probability,” *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.

[8] J. B. Predd, D. N. Osherson, S. R. Kulkarni, and H. V. Poor, “Aggregating probabilistic forecasts from incoherent and abstaining experts,” *Decision Analysis*, vol. 5, no. 4, pp. 177–189, 2008.

[9] Y. Censor and S. A. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, New York, USA, 1997.

[10] G. Wang, S. R. Kulkarni, and H. V. Poor, “Aggregating disparate judgments using a coherence penalty,” in *Proceedings of CISS 2009*, March 2009, pp. 23–27.

[11] D. Kahneman and A. Tversky, “Prospect theory: An analysis of decision under risk,” *Econometrica*, vol. 47, no. 2, pp. 263–91, March 1979.

[12] A. Tversky and D. Kahneman, “Advances in prospect theory: Cumulative representation of uncertainty,” *Journal of Risk and Uncertainty*, vol. 5, no. 4, pp. 297–323, October 1992.

[13] J. B. Predd, R. Seiringer, E. H. Lieb, D. N. Osherson, H. V. Poor, and S. R. Kulkarni, “Probabilistic coherence and proper scoring rules,” *IEEE Trans. Inf. Theory.*, vol. 55, no. 10, pp. 4786–4792, 2009.

[3] The only exception is that *simple linear averaging* reports the same slope as *raw*