The maximum-entropy and minimax-redundancy distribution classes of sufficiently small codelength

10th Workshop on Information Theoretic Methods in Science and Engineering

Paris, France

September 11, 2017

David Bickel
University of Ottawa

uOttawa

Model := distribution class

- A *model* is a class of distributions
- Each of these is a model:
  - Family of data distributions (input to 2-part code)
  - Predictive distribution formed by integrating over such a family with respect to weights

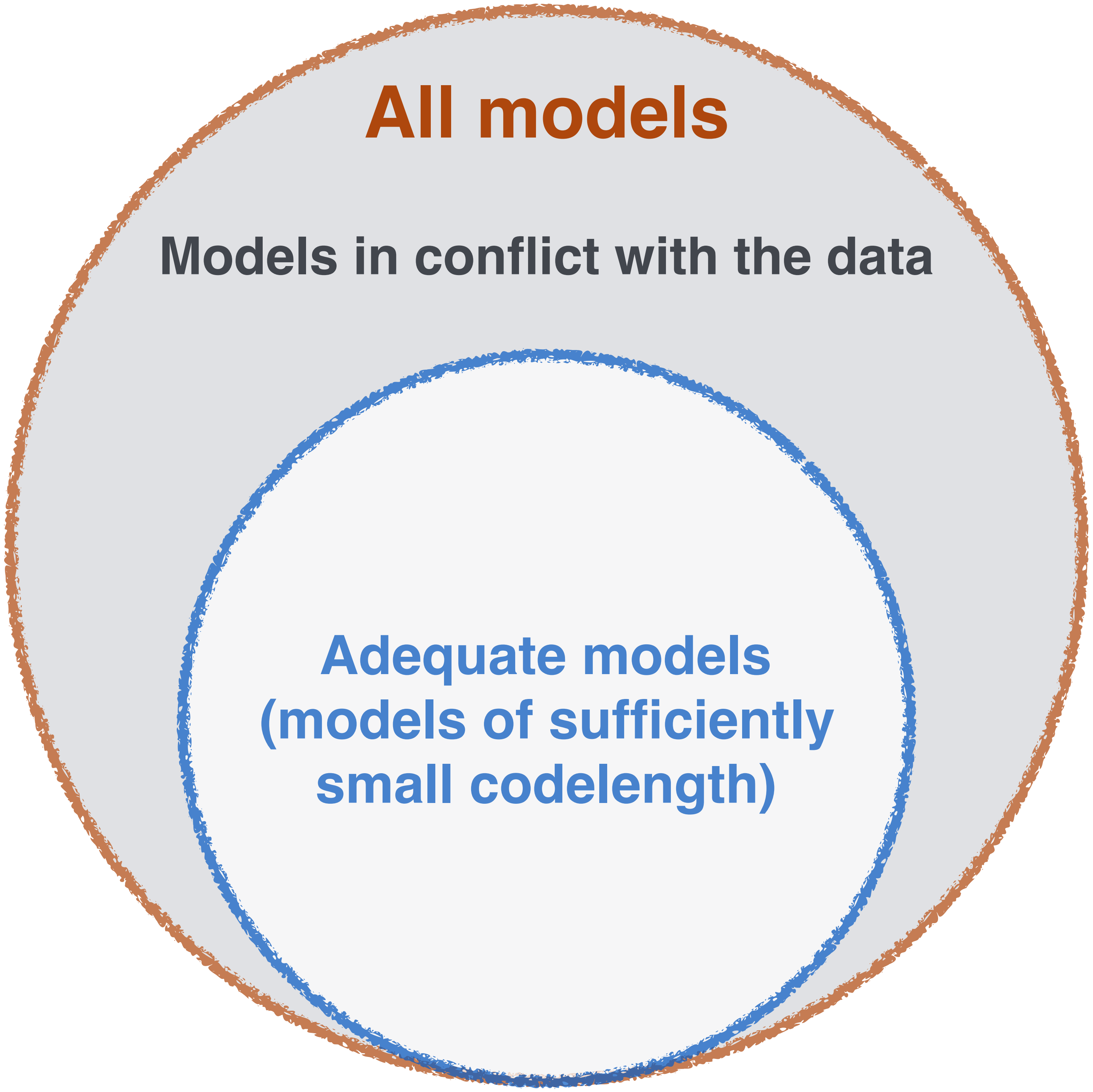
Adequate models have sufficient evidence

$$B\left(M;x\right)=\frac{f_{M}\left(x\right)}{f_{\neg M}\left(x\right)}=\frac{\int f_{M}\left(x|\theta\right)\pi_{M}\left(\theta\right)d\theta}{\int f_{M}\left(x|\theta\right)\pi_{\neg M}\left(\theta\right)d\theta}$$

$$w\left(M\right)=\log B\left(M;x\right)$$

$$\mathcal{M}\left(a\right)=\left\{M\in\mathcal{M}:w\left(M\right)>a\right\}$$

Assessing multiple models



Bayesian model assessment

Does the prior or parametric family conflict with the data?

- Yes, if it has relatively high codelength (low support)
- **Measures of support** satisfying the criteria of Bickel, *International Statistical Review* **81**, 188-206:
  - Likelihood ratio (limited to comparing two simple models)
  - Bayes factor
  - Penalized likelihood ratio

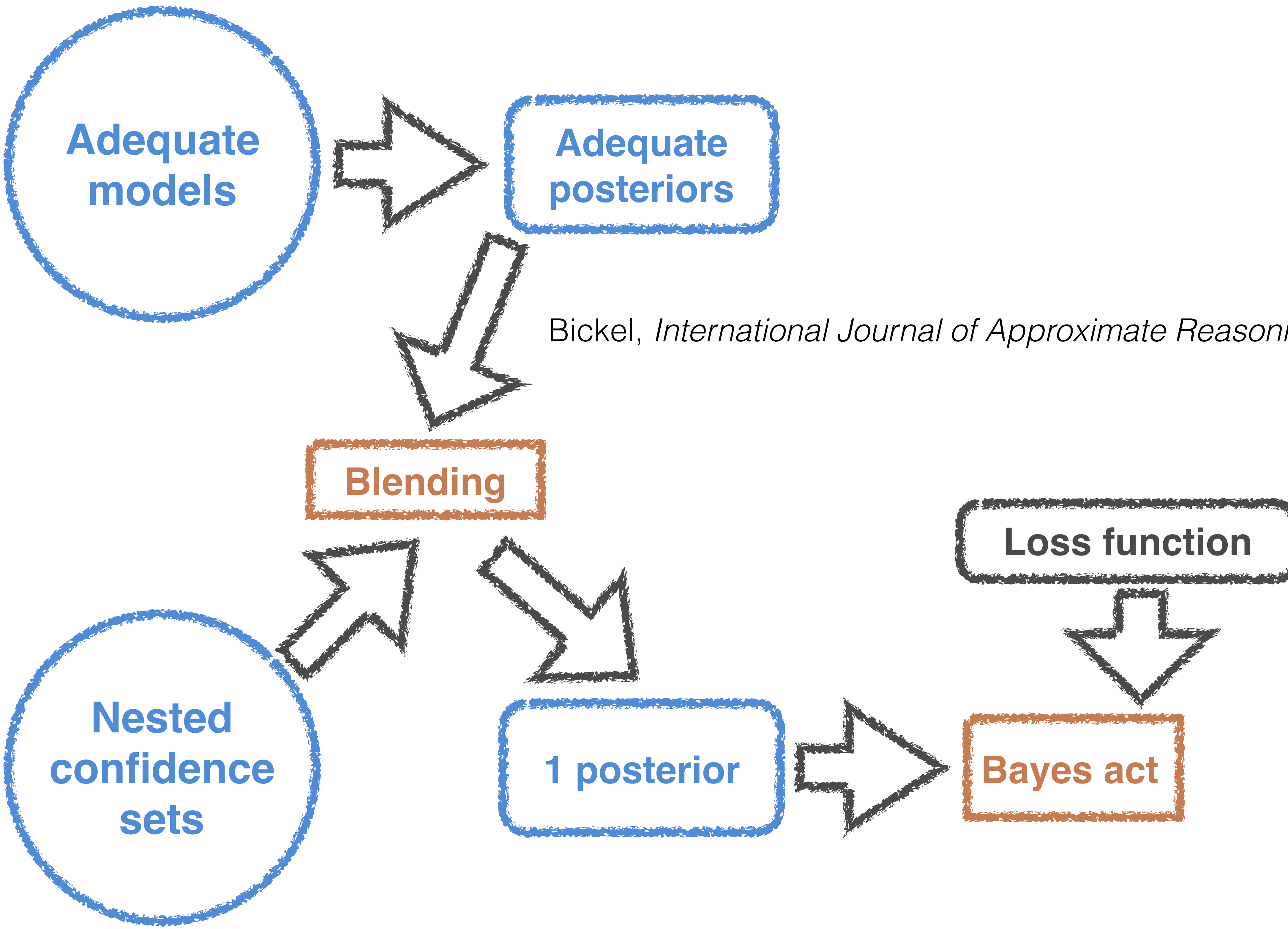
No

Use the model for decisions
(minimize posterior expected loss)
even if other models are adequate

Yes

Change the model
to an adequate model, a model of small codelength

Decisions under the entropy-maximizing model of sufficient evidence



Bickel, *International Journal of Approximate Reasoning* **66**, 53-72

Posterior distributions of the parameter

- Set $\dot{\mathcal{P}}$ of adequate posterior distributions on $(\Theta, \mathcal{A})$
  - Subjective interpretation: interval levels of belief
  - Objective interpretation: physical constraints
- Set $\ddot{\mathcal{P}}$ of confidence distributions
  - Confidence distribution $\ddot{P}$ on $(\Theta, \mathcal{A})$

Information theory

- Information divergence of P with respect to Q on $(\Theta, \mathcal{A})$:

  $P \ll Q \implies I(P||Q) = \int dP \log\left(\frac{dP}{dQ}\right)$

  $P \not\ll Q \implies I(P||Q) = \infty$

- Inferential gain of Q relative to $\ddot{P}$ given $\dot{P}$:

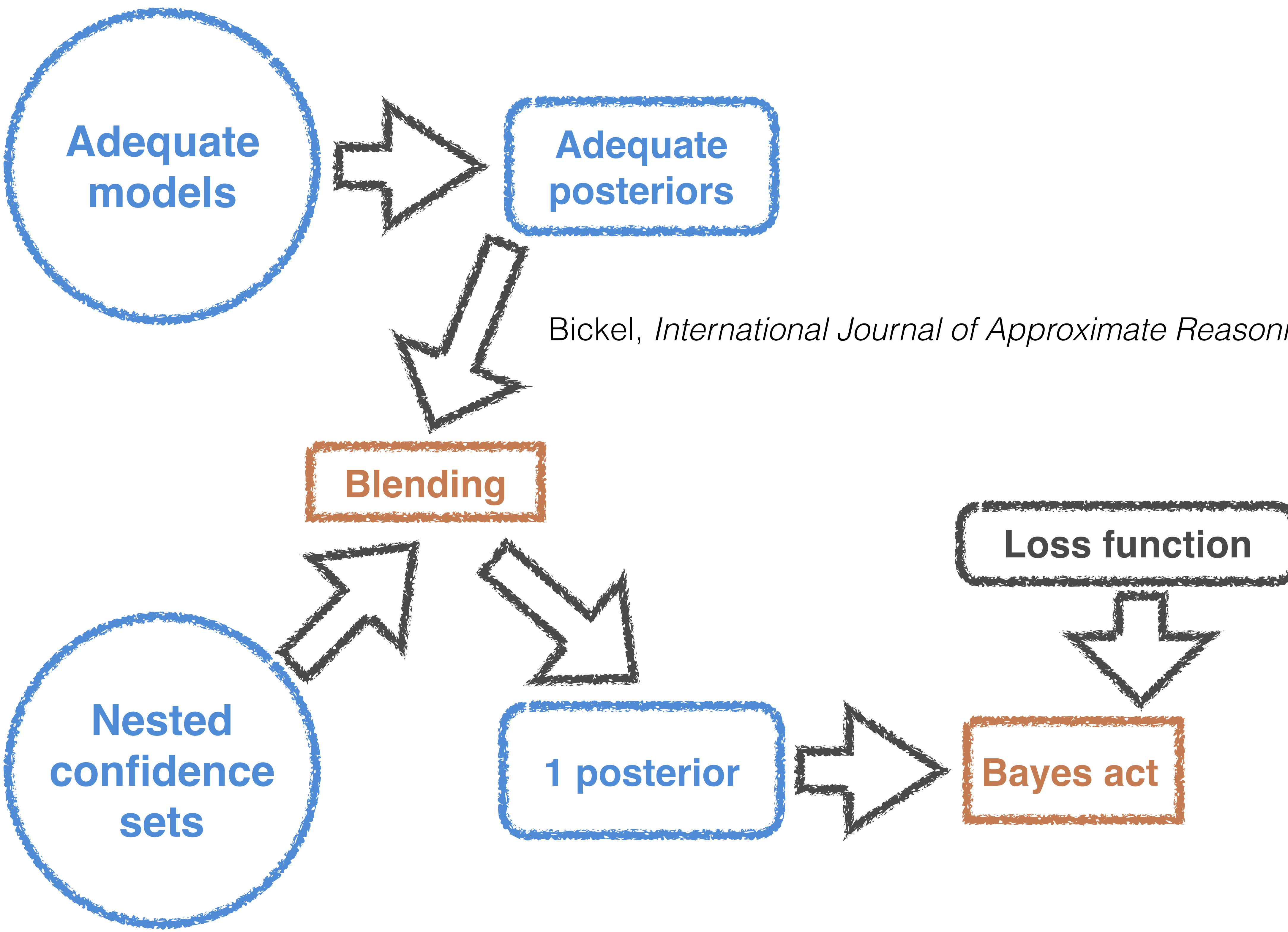
  $I\left(\dot{P}||\ddot{P} \rightsquigarrow Q\right) = I\left(\dot{P}||\ddot{P}\right) - I\left(\dot{P}||Q\right)$

- Subset $\dot{\mathcal{P}}\left(\ddot{P}\right) = \left\{\dot{P} \in \dot{\mathcal{P}} : I\left(\dot{P}||\ddot{P}\right) < \infty\right\}$ of Bayes posteriors

Game theory

- The inferential gain of Q is $I\left(\dot{P}||\ddot{P} \rightsquigarrow Q\right) = I\left(\dot{P}||\ddot{P}\right) - I\left(\dot{P}||Q\right)$.
- The <u>blended posterior distribution</u> $\hat{P}$ is defined by this game:
  $\inf_{\dot{P}\in\dot{\mathcal{P}}(\ddot{P})} I\left(\dot{P}||\ddot{P} \rightsquigarrow \hat{P}\right) = \sup_{Q\in\mathcal{P}} \inf_{\dot{P}\in\dot{\mathcal{P}}(\ddot{P})} I\left(\dot{P}||\ddot{P} \rightsquigarrow Q\right)$.
- If $I\left(\dot{P}||\ddot{P}\right) < \infty$ for some $\dot{P} \in \dot{\mathcal{P}}$ and if $\dot{\mathcal{P}}\left(\ddot{P}\right)$ is convex,
  then $I\left(\hat{P}||\ddot{P}\right) = \inf_{\dot{P}\in\dot{\mathcal{P}}(\ddot{P})} I\left(\dot{P}||\ddot{P}\right)$.
  - F. Topsøe, *Kybernetika* 15 (1979), 8-27; P. Harremoës and F. Topsøe, *Entropy* 3 (2001), 191-226; F. Topsøe, in *Entropy, Search, Complexity* (Springer, 2007), 179-207

Decisions under the entropy-maximizing Bayesian model of sufficient evidence



Bickel, *International Journal of Approximate Reasoning* **66**, 53-72

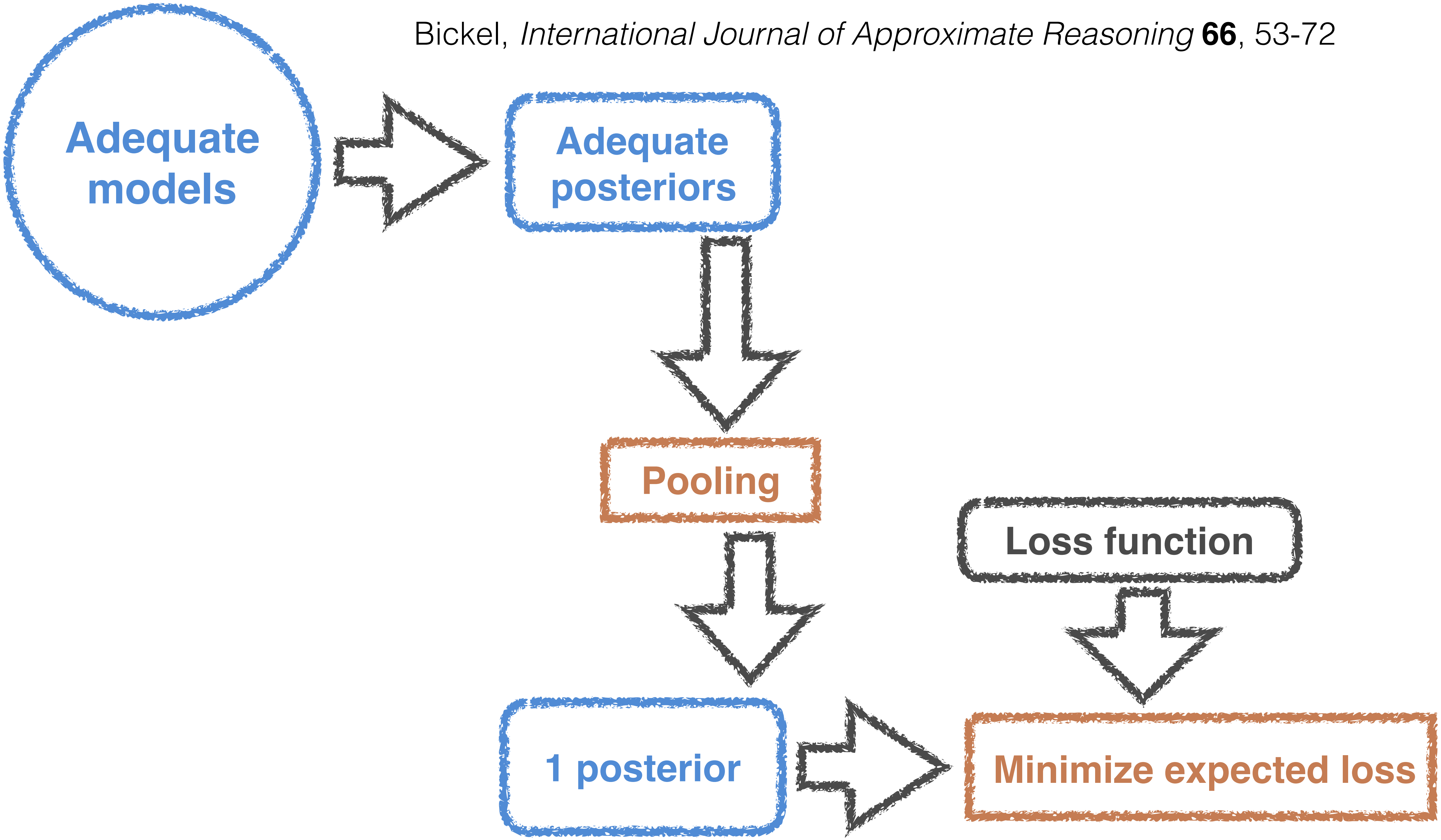
Pooling methods

- Good derived a harmonic mean of p-values from Bayes' rule
- Examples of subjectively weighting each expert's distribution:
    - Minimizing a weighted sum of divergences from the distributions being combined yields a linear combination of the distributions
        - Any linearly combined marginal distribution is the same whether marginalization or combination is carried out first
    - Weighted multiplicative combination of the distributions
        - Externally Bayesian updating

uOttawa

Decisions under minimax-redundancy pooling of models of small codelengths

Bickel, *International Journal of Approximate Reasoning* **66**, 53-72



The three players

- **Combiner** is you, the statistician who combines the candidate distributions
  - Goal 1: minimize error; Goal 2: beat Chooser

- **Chooser** is the imaginary statistician who instead chooses a single candidate distribution
  - Goal 1: minimize error; Goal 2: beat Combiner

- **Nature** picks the true distribution among those that are plausible in order to help Chooser

Information game

- Information divergence and inferential gain:

$$D\left(P||Q\right) = \int dP\left(\xi\right) \log \frac{dP(\xi)}{dQ(\xi)}$$

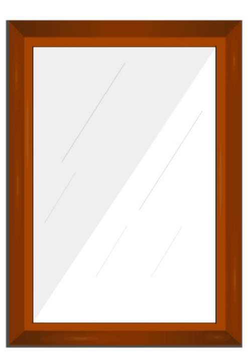
$$D\left(P'||P'' \to Q\right) = D\left(P'||P''\right) - D\left(P'||Q\right)$$

- Utility paid to Statistician 2 (Combiner or Chooser):

$$U\left(\dot{P}, P_1, P_2\right) = \left(-D\left(\dot{P}||P_2\right), D\left(\dot{P}||P_1 \to P_2\right)\right)$$

- Reduction to game of Combiner v. Nature-Chooser Coalition:

$$P^{+} = \arg\inf_{Q} \sup_{P'} D\left(P'||Q\right)$$

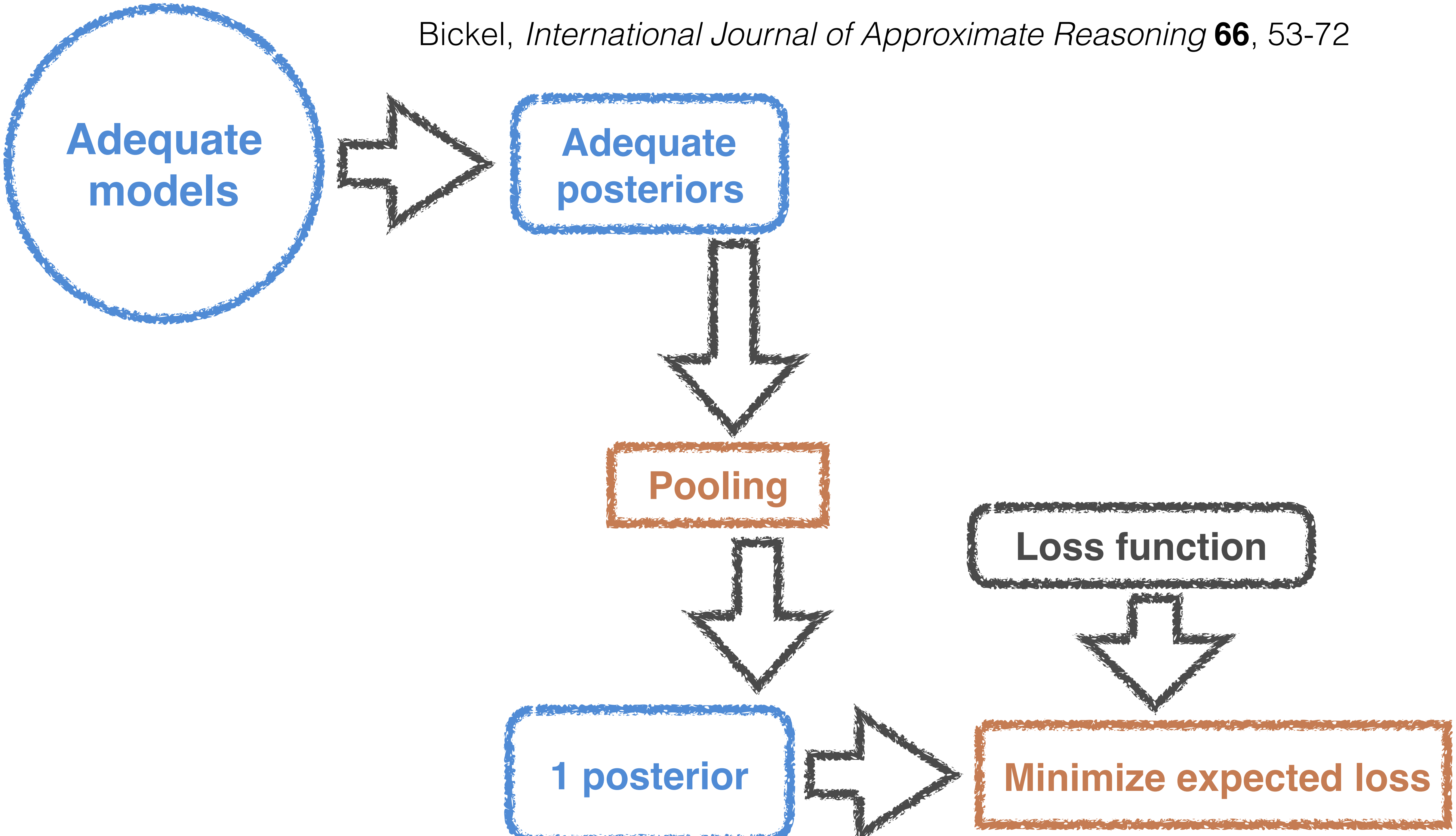


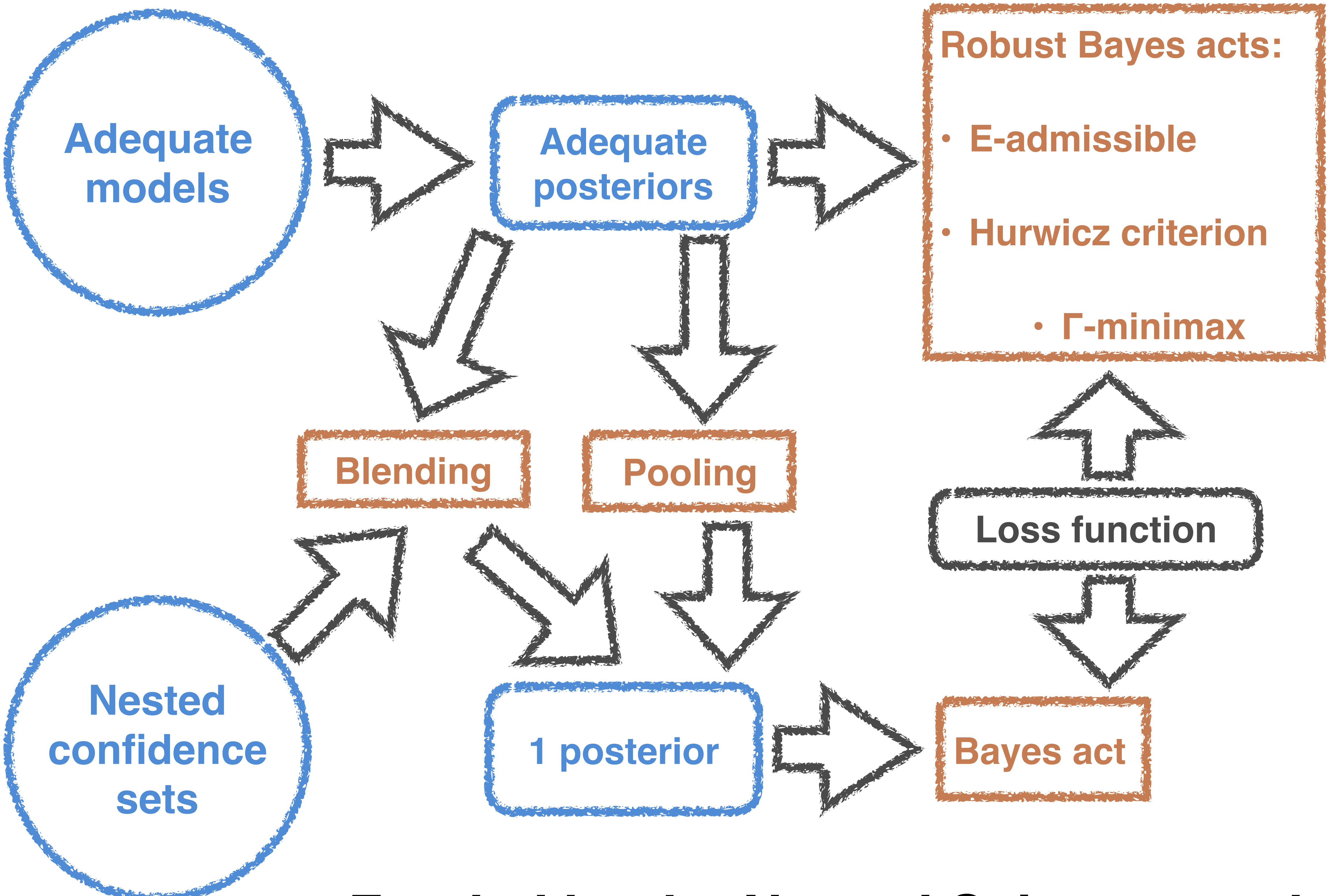
Q

P'

Decisions under minimax-redundancy pooling of models of small codelengths

Bickel, *International Journal of Approximate Reasoning* **66**, 53-72





Funded by the Natural Sciences and Engineering Research Council of Canada