Cromwell's principle idealized under the theory of large deviations

Seminar, Statistics and Probability Research Group, University of Ottawa

Ottawa, Ontario

April 27, 2018

David Bickel
University of Ottawa

uOttawa

Cromwell’s principle

- *Cromwell's principle*: P(model is wrong) > 0, i.e., the prior probability that the assumptions are incorrect is greater than 0
  —D. Lindley (2013), *Understanding Uncertainty*
    - Relevance to Bayesian model checking: Diagnostics often reveal that prior distributions require revision—impossible under Bayes's theorem if P(model) = 1
- *Idealized Cromwell's principle*: P(model) vanishes
    - Large deviations version of the infinitesimal probability due to A. Hájek and M. Smithson (2012), *Synthese* **187**, 33–48

D. R. Bickel (2018), *Statistics*, http://bit.ly/2vkE23Q

Empirical distribution converges to max. entropy

$$p^{x}\left(\mu_{0}\right)=\operatorname{Prob}_{\bar{X}^{\text{prior}}\sim\mathrm{N}\left(\mu_{0},\sigma^{2}/n+\sigma_{0}^{2}\right)}\left(\left(\bar{X}^{\text{prior}}-\mu_{0}\right)^{2}\geq\left(\bar{x}-\mu_{0}\right)^{2}\right)$$

$$\Gamma\left(\alpha;x\right)=\operatorname{clco}\left\{\mathrm{N}\left(\mu_{0},\sigma_{0}^{2}\right):p^{x}\left(\mu_{0}\right)\geq\alpha,\mu_{0}\in\mathbb{R}\right\}$$

$$\inf_{P'\in\Gamma}D\left(P'||P\right)=\inf_{P'\in\operatorname{int}\Gamma}D\left(P'||P\right)<\infty$$

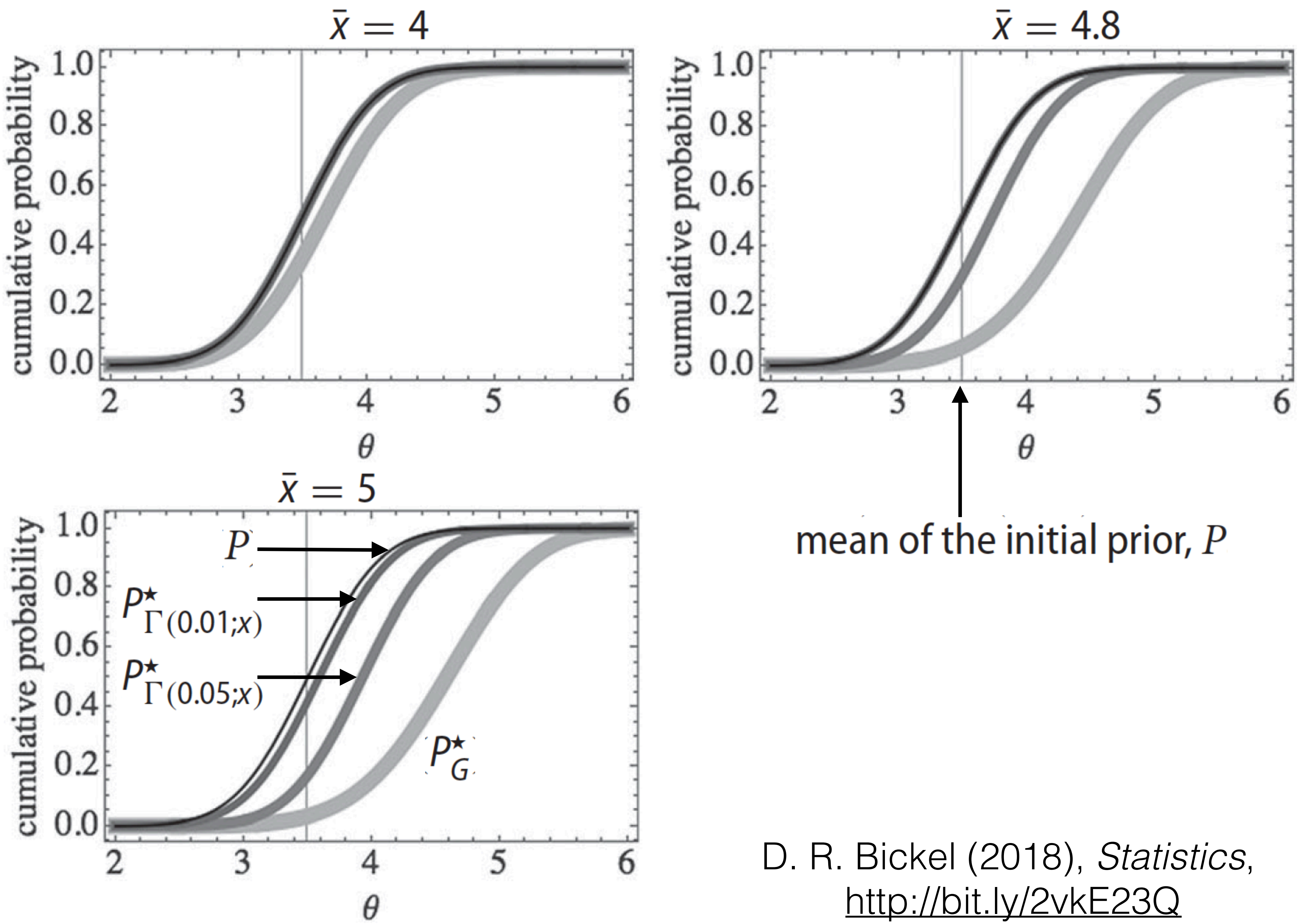
$$\Downarrow$$

Weak convergence to maximum entropy distribution

$$D\left(P_{\Gamma}^{\star}||P\right)=\inf_{P'\in\Gamma}D\left(P'||P\right)$$

D. R. Bickel (2018), *Statistics*, http://bit.ly/2vkE23Q

Normal-normal CDFs



D. R. Bickel (2018), *Statistics*,
http://bit.ly/2vkE23Q

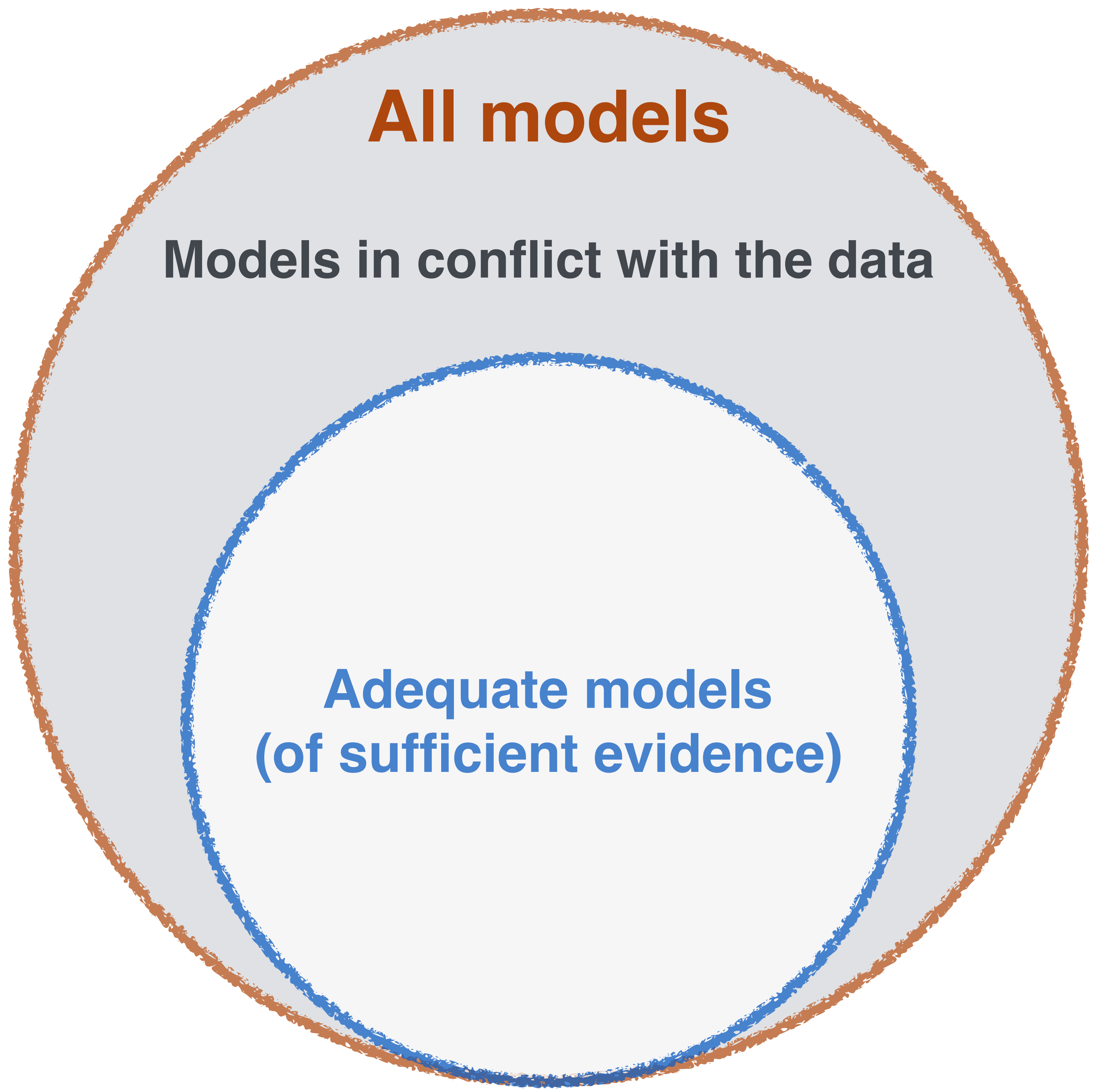
Adequate models have sufficient evidence

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

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

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

Assessing multiple models



Bayesian model assessment

Does the prior or parametric family conflict with the data?

- Yes, if it has relatively low evidence
- **Measures of evidence** 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 high evidence

Loss function example

$$\ell(\theta, \delta) = \begin{cases} 0 & \text{if } \delta = \theta \in \{0, 1\} \\ \ell_{\mathrm{I}} & \text{if } \delta = 1, \theta = 0 \\ \ell_{\mathrm{II}} & \text{if } \delta = 0, \theta = 1 \end{cases}$$

What act is appropriate?

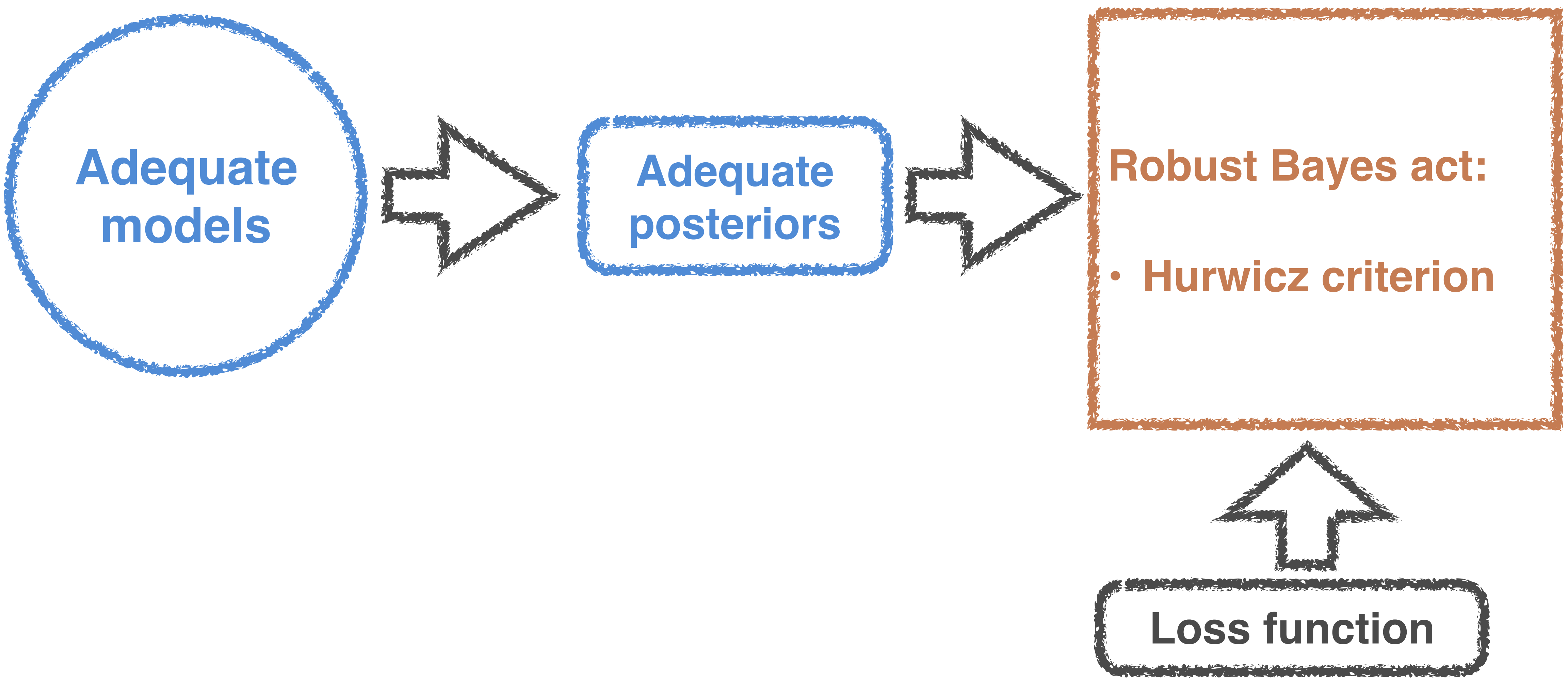
- Given **multiple priors** of sufficient evidence, agreement with the data
- With or without **confidence sets**

Hurwicz criterion

$$\delta(a,\kappa) = \arg\inf_{\delta\in\mathcal{D}} \left(\kappa \sup_{M\in\mathcal{M}(a)} E_M\left(\ell\left(\vartheta,\delta\right)\right) + (1-\kappa) \inf_{M\in\mathcal{M}(a)} E_M\left(\ell\left(\vartheta,\delta\right)\right) \right)$$

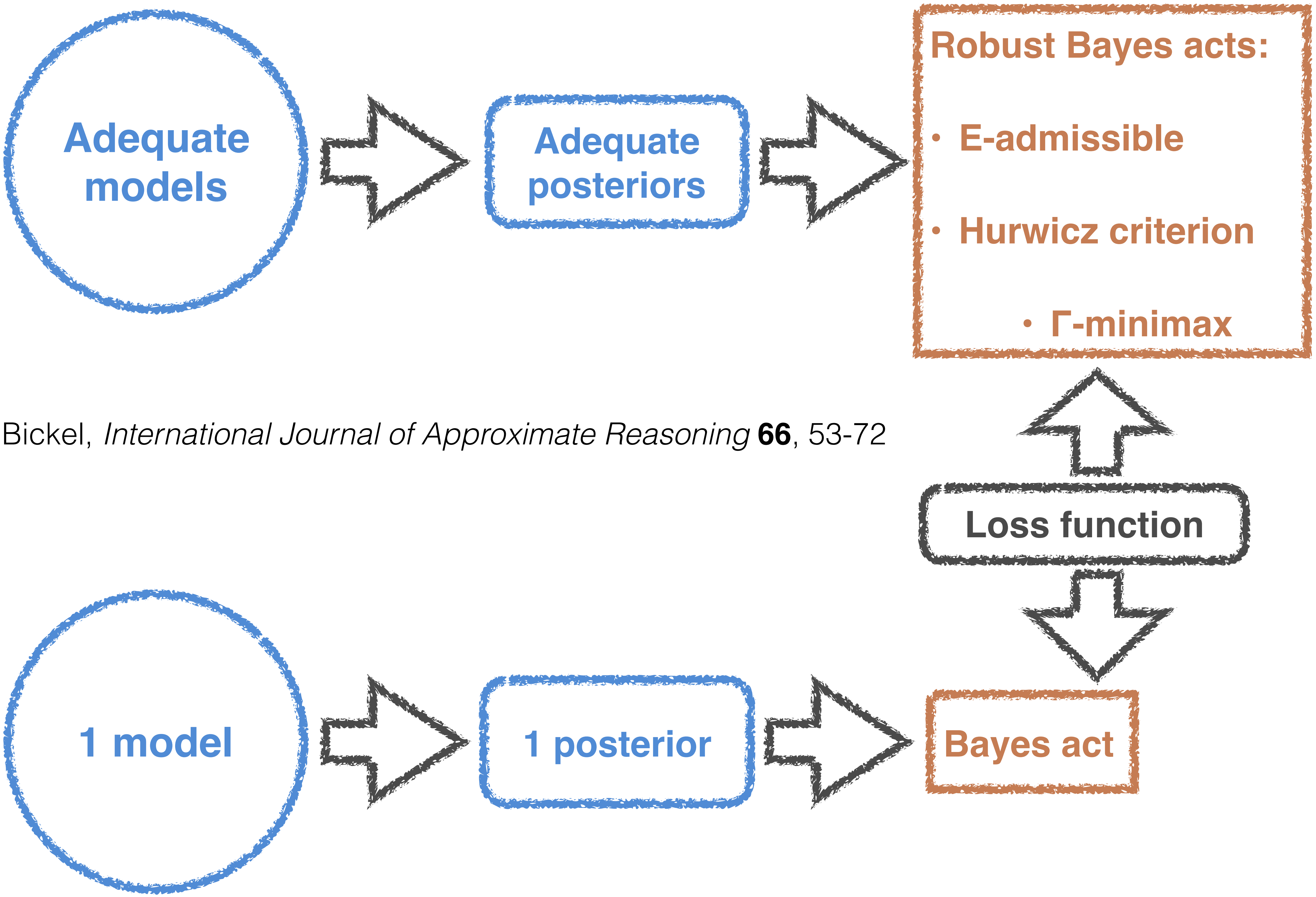
$$\ell(\theta,\delta) = \begin{cases} 0 & \text{if } \delta=\theta\in\{0,1\} \\ \ell_{\mathrm{I}} & \text{if } \delta=1, \theta=0 \\ \ell_{\mathrm{II}} & \text{if } \delta=0, \theta=1 \end{cases}$$

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



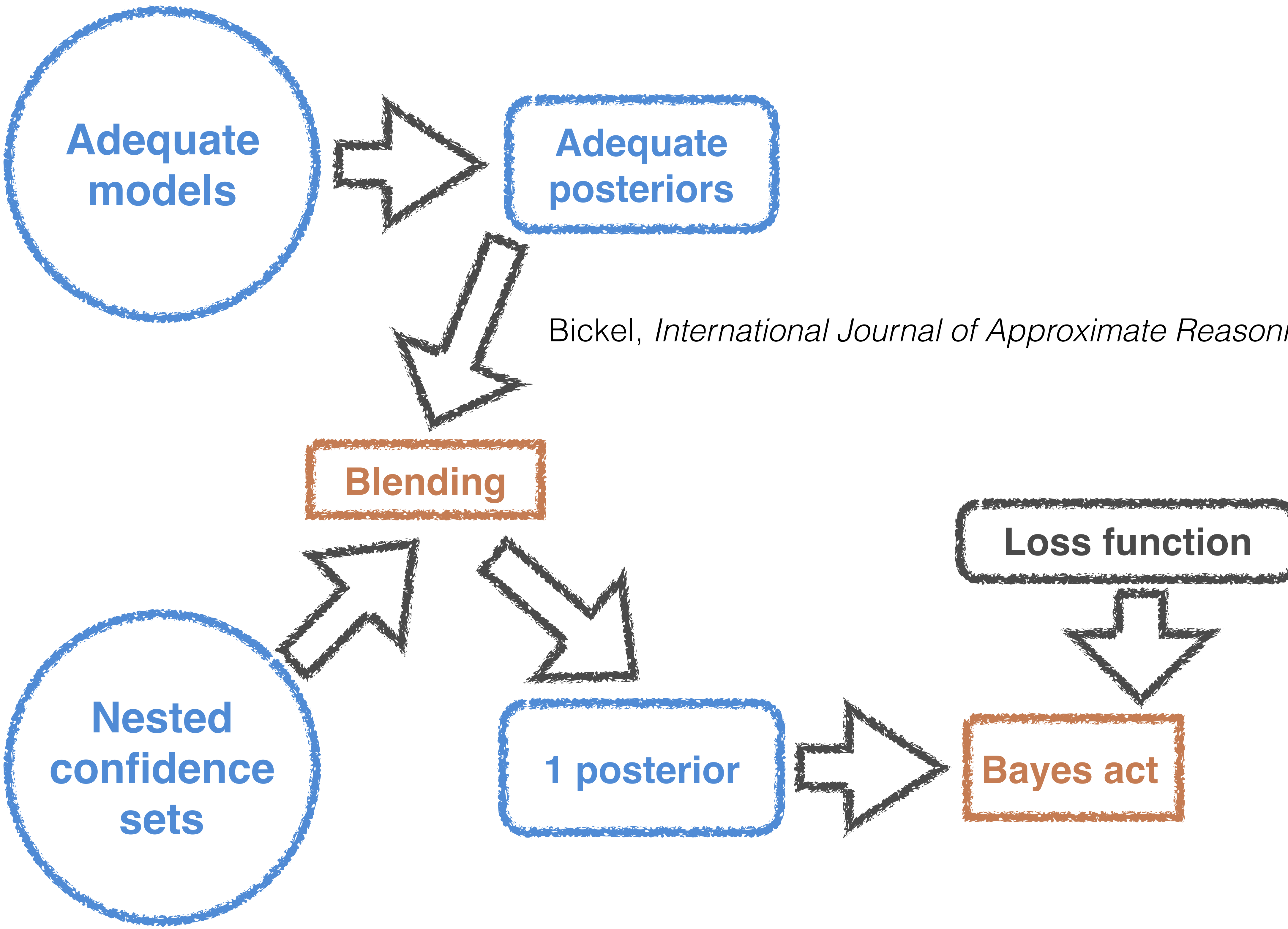
$$\delta\left(a,\kappa\right)=\arg\inf_{\delta\in\mathcal{D}}\left(\kappa\sup_{M\in\mathcal{M}(a)}E_M\left(\ell\left(\vartheta,\delta\right)\right)+\left(1-\kappa\right)\inf_{M\in\mathcal{M}(a)}E_M\left(\ell\left(\vartheta,\delta\right)\right)\right)$$

Decisions under robust Bayes rules applied to models of sufficient evidence



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

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 $\dot{P}$:

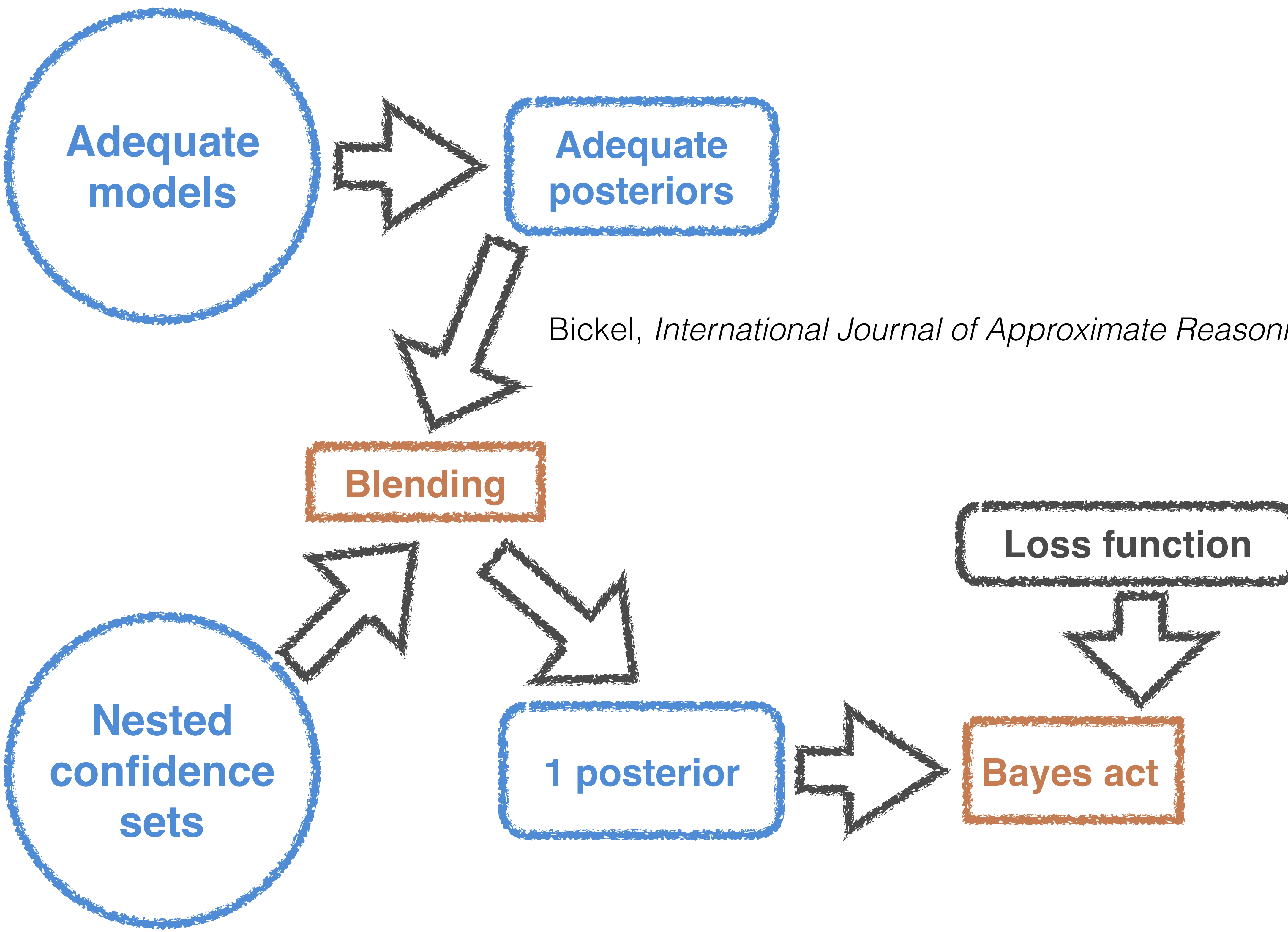
  $$I\left(\dot{P}||\ddot{P} \rightsquigarrow Q\right) = I\left(\dot{P}||\ddot{P}\right) - I\left(\dot{P}||Q\right)$$

- 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 blended posterior distribution $\hat{P}$ 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

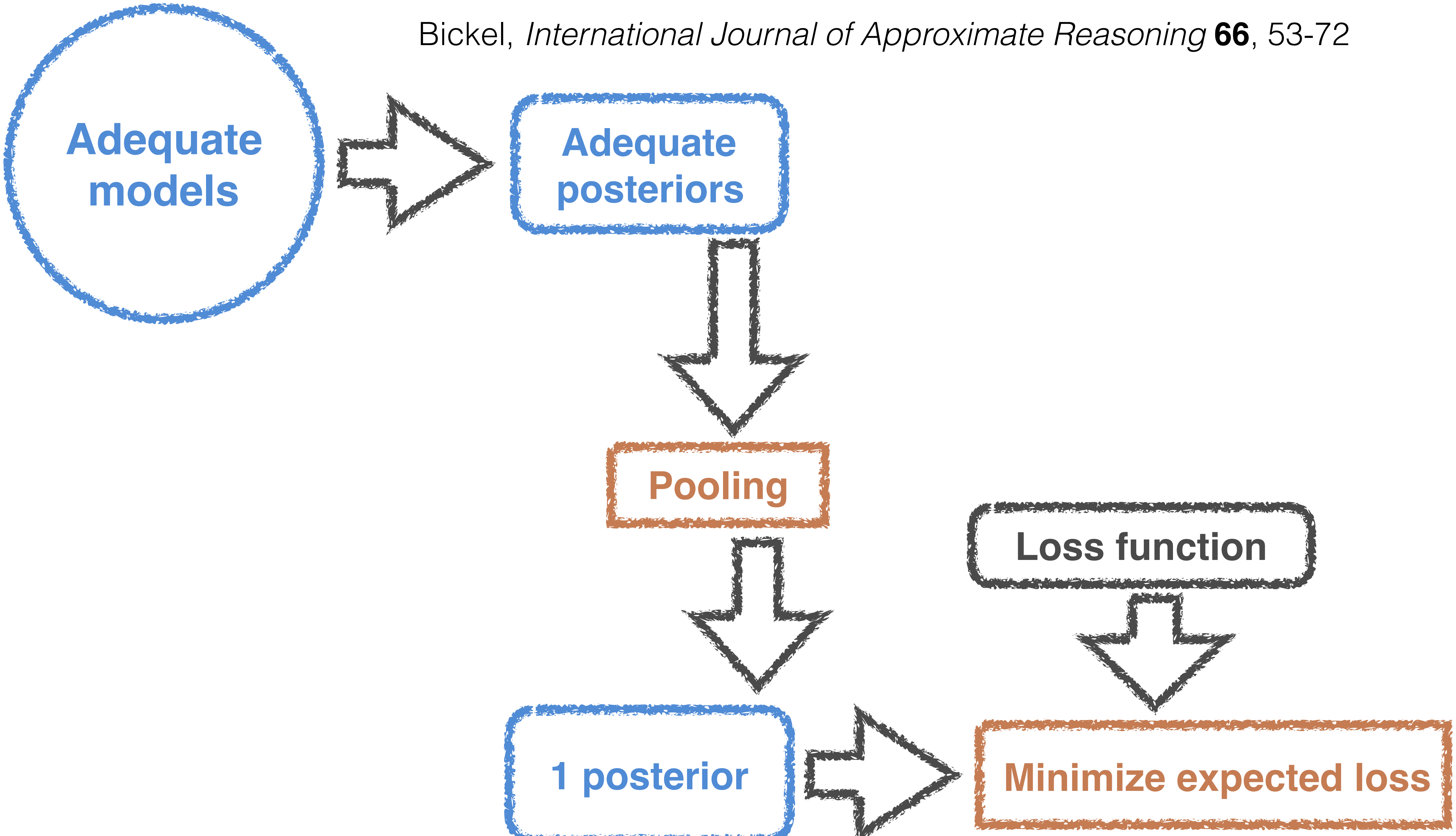
Pooling methods

- 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

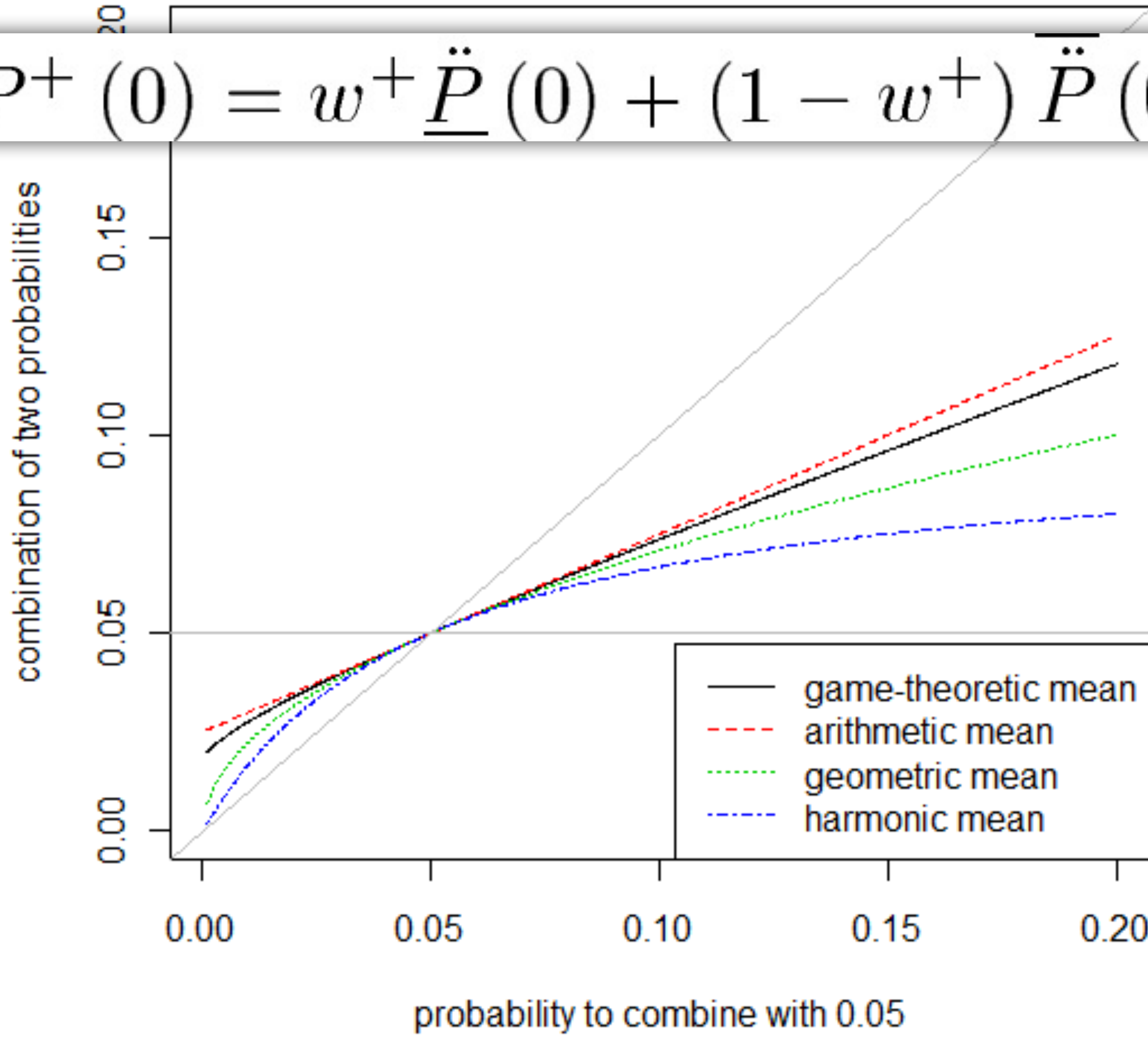
Decisions under minimax-redundancy pooling of models of small codelengths

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



Conflicting probabilities

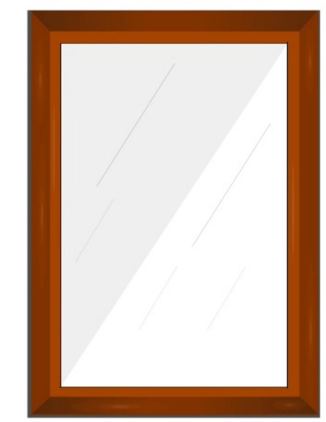
$$P^{+}(0) = w^{+}\underline{\ddot{P}}(0) + (1 - w^{+})\,\overline{\ddot{P}}(0)$$



uOttawa

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:

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

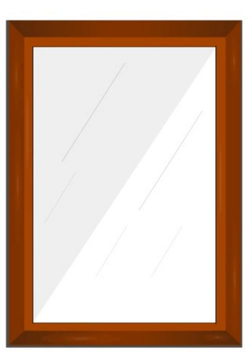
$$D\left(P'||P'' \to Q\right) = D\left(P'||P''\right) - D\left(P'||Q\right)$$

- 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:

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

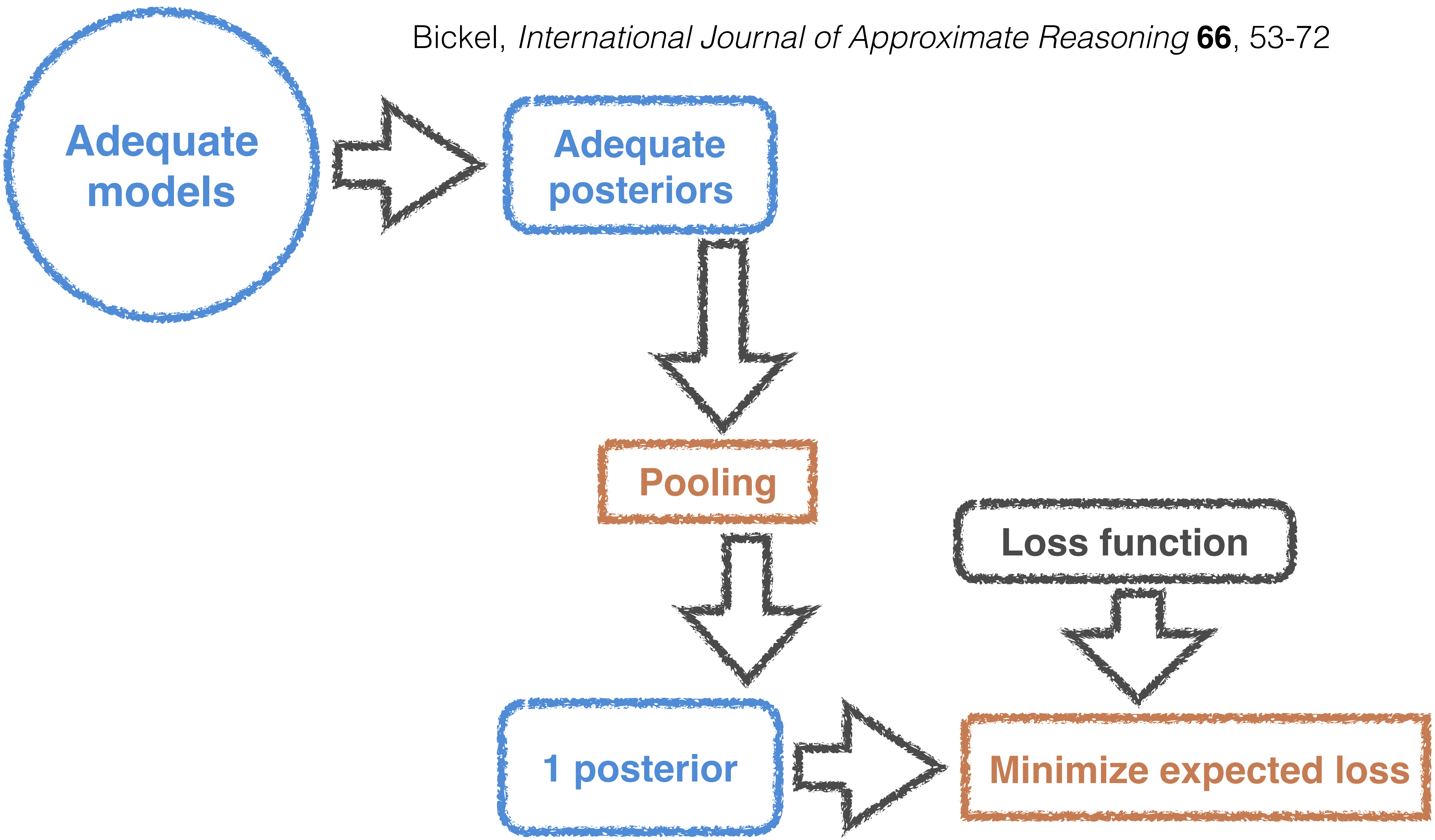


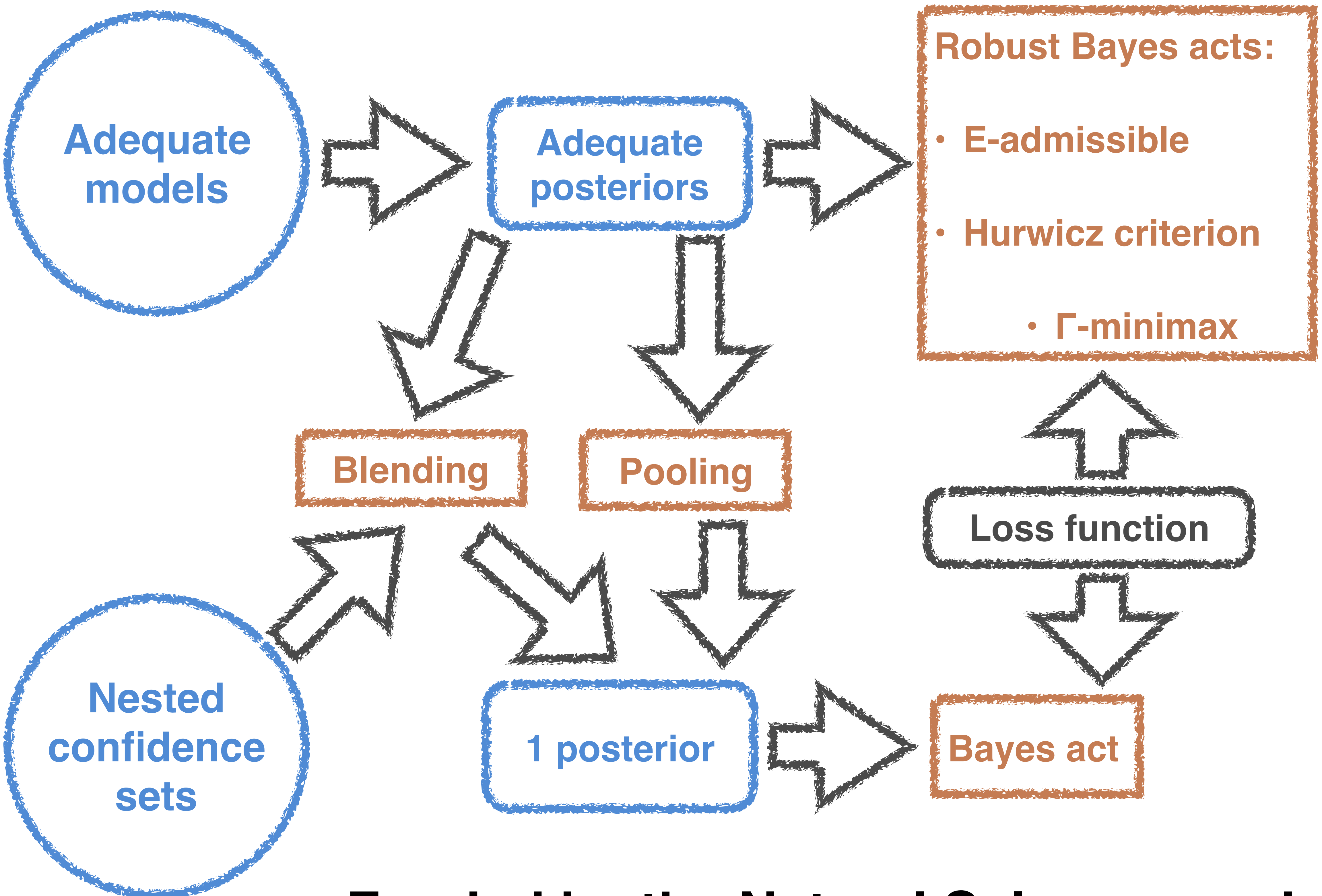
Q

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