

# Modelling behavioural data

Quentin Huys  
MA PhD MBBS MBPsS

Translational Neuromodeling Unit, ETH Zürich  
Psychiatrische Universitätsklinik Zürich

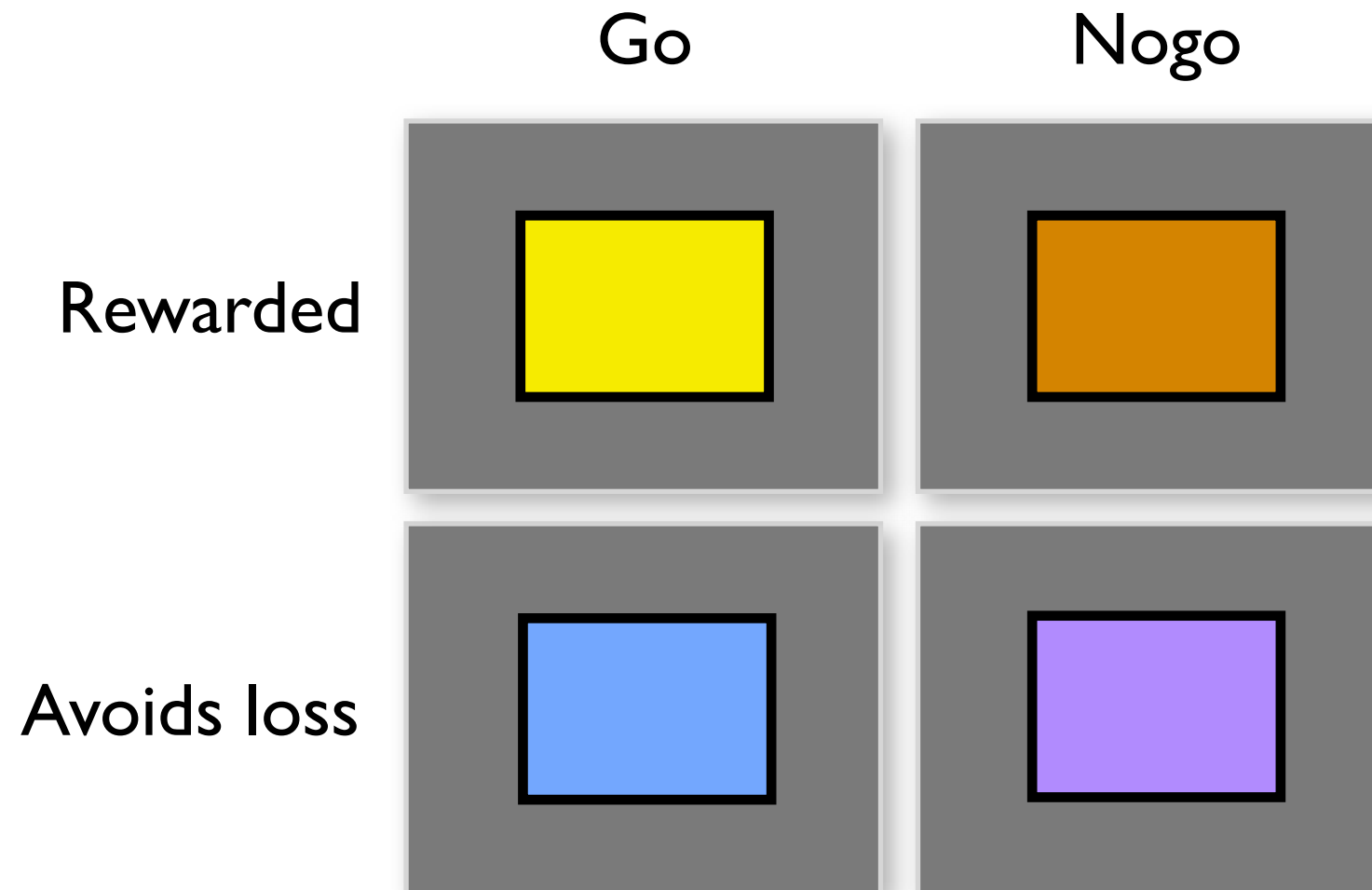
# Outline

---

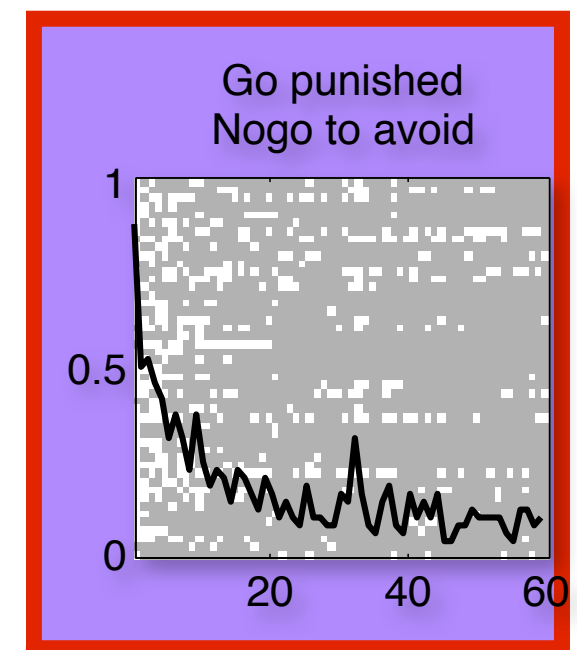
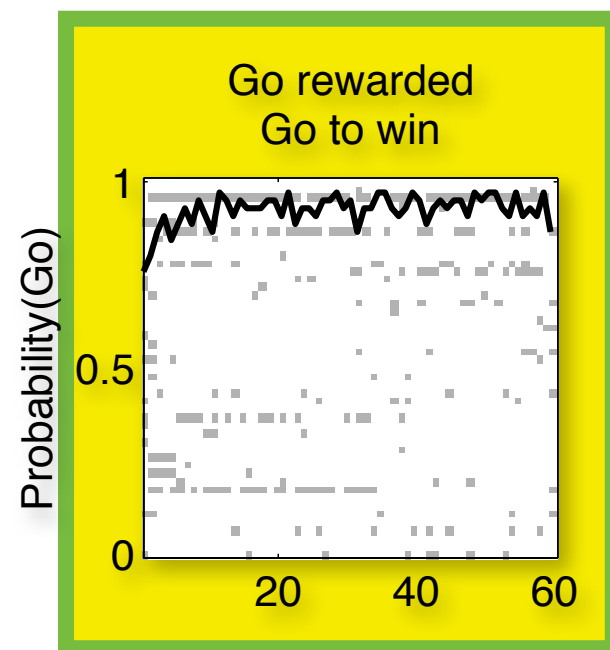
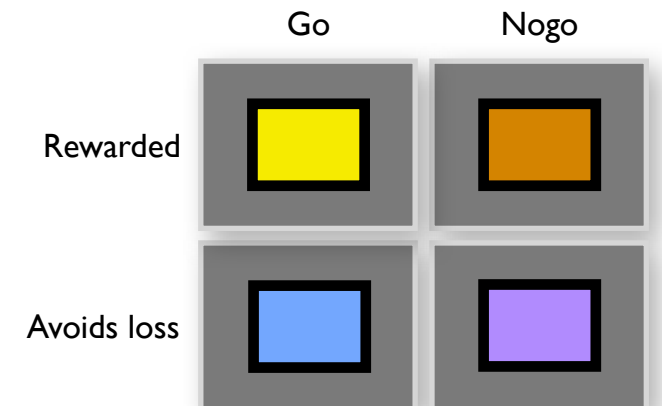
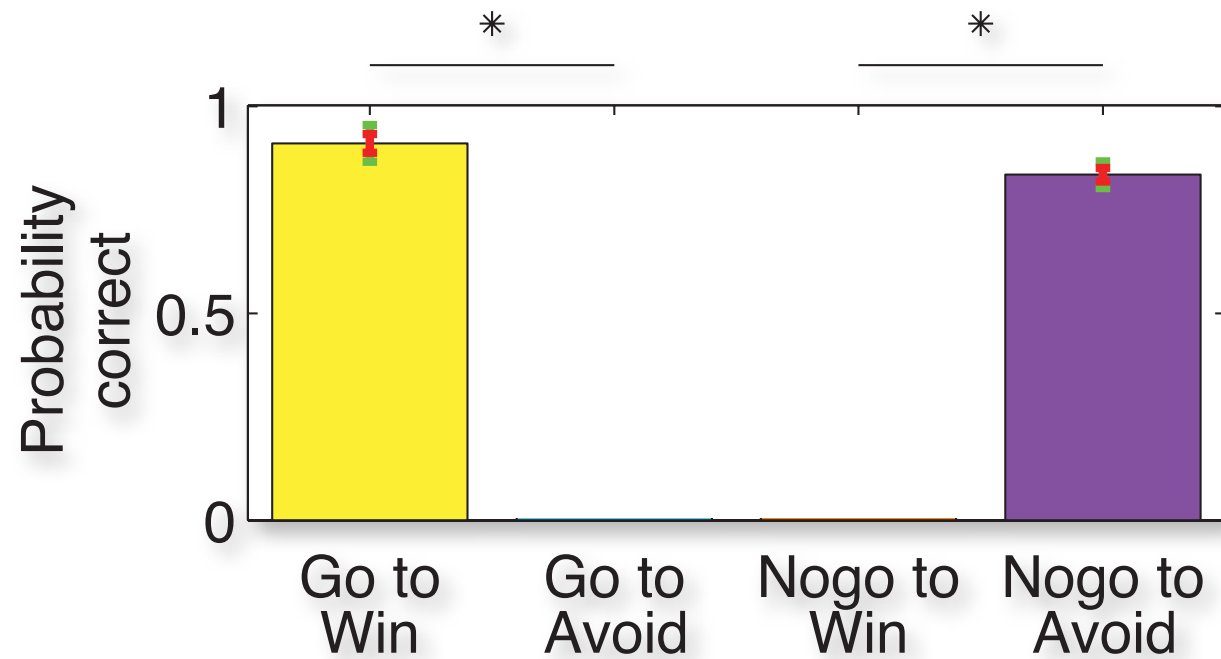
- ▶ An example task
- ▶ Why build models? What is a model
- ▶ Fitting models
- ▶ Validating & comparing models
- ▶ Model comparison issues in psychiatry

# Example task

---



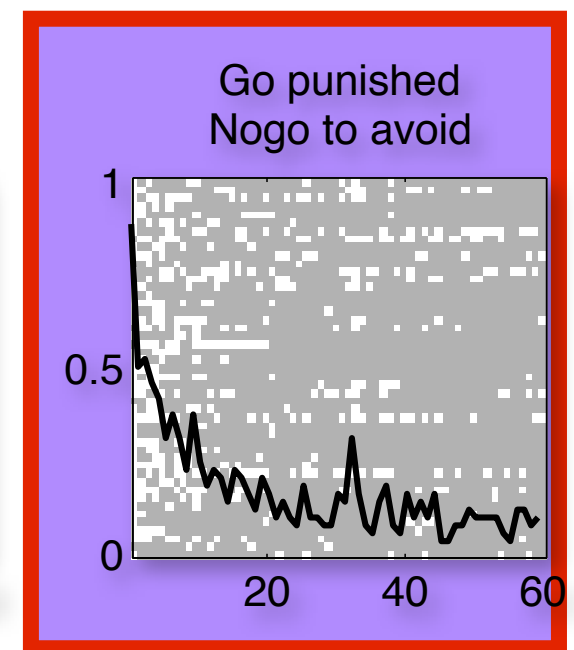
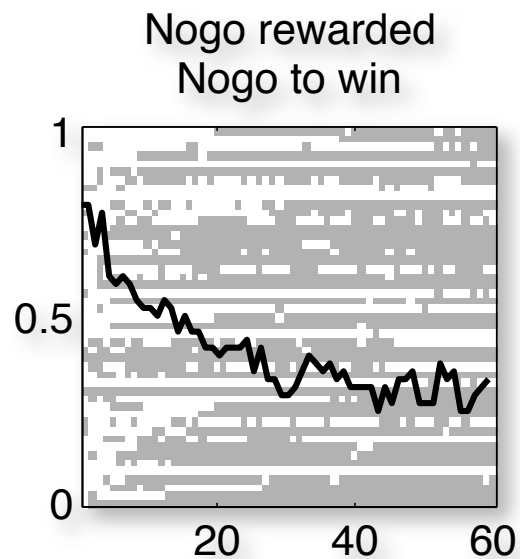
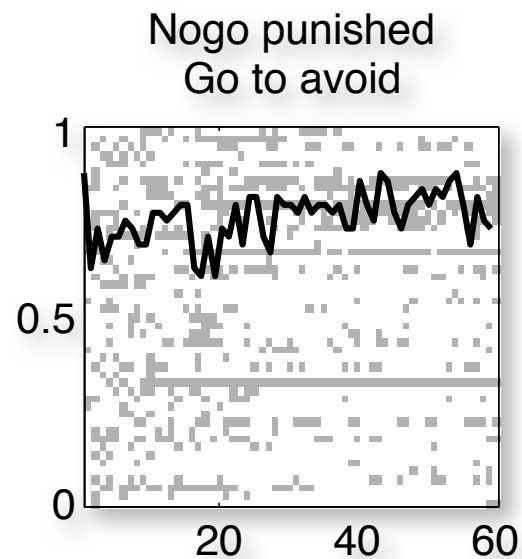
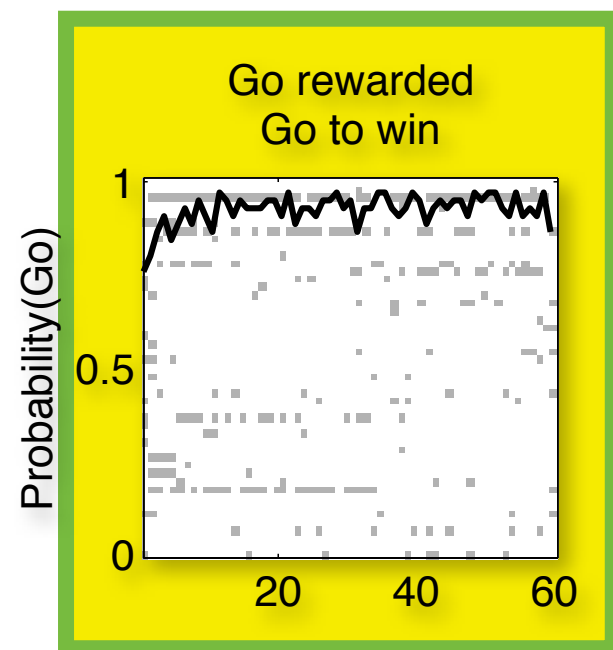
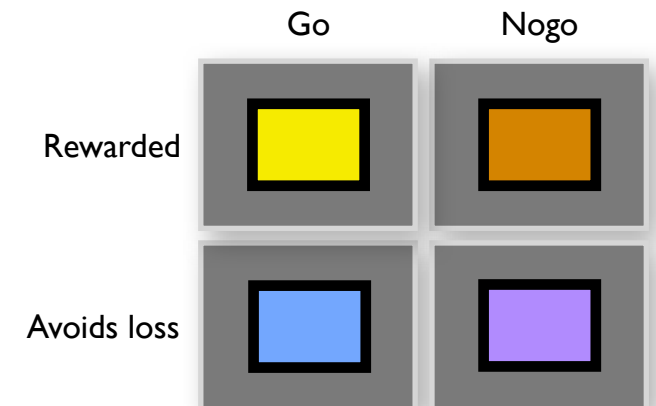
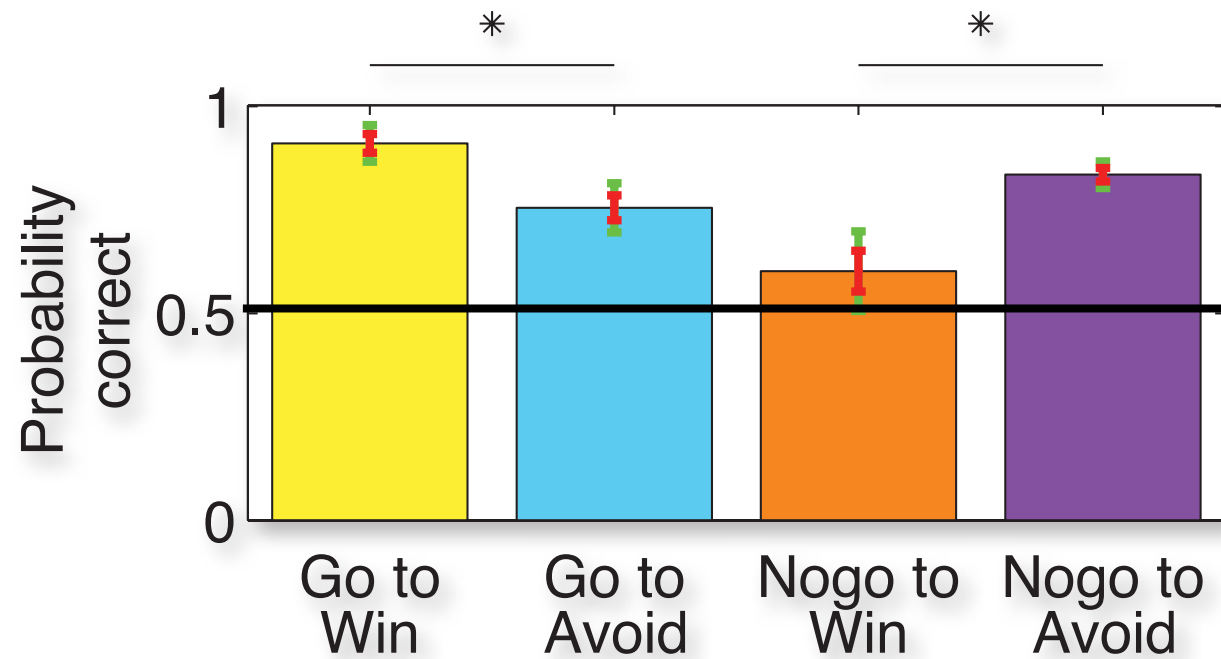
# Example task



Guitart-Masip, Huys et al. *Submitted*

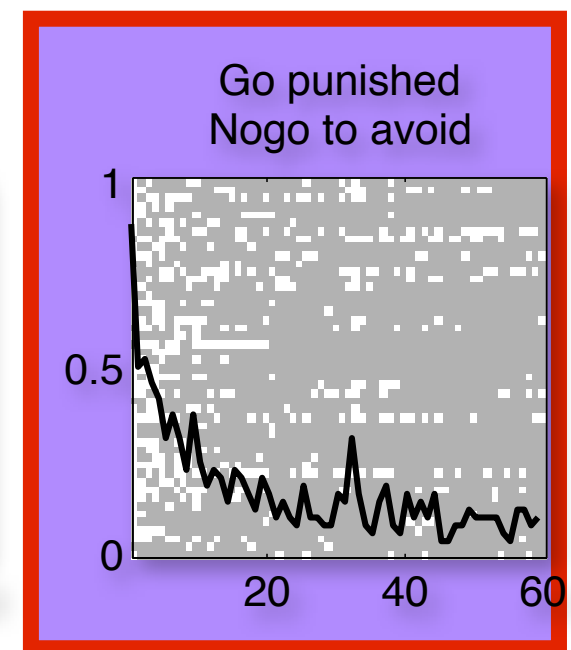
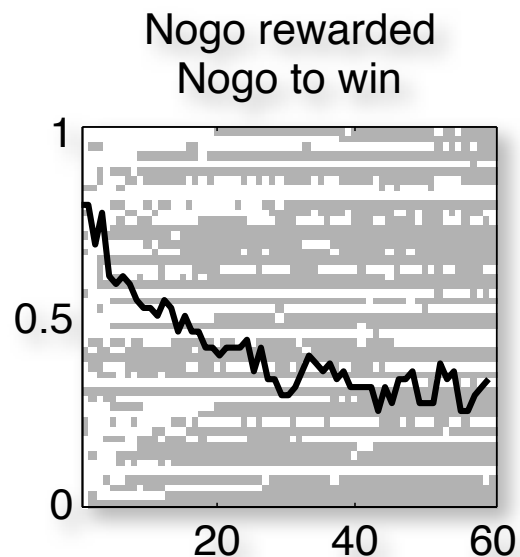
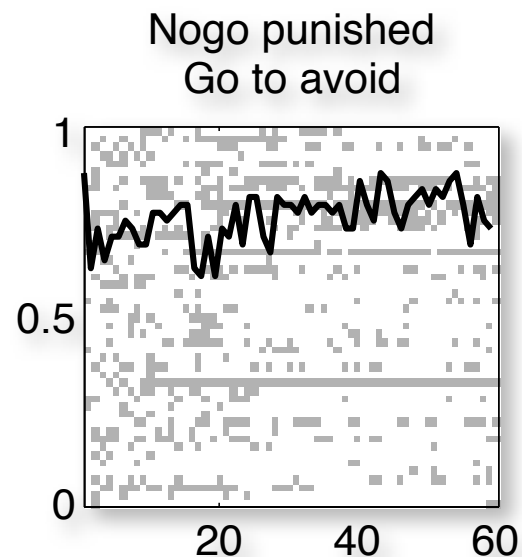
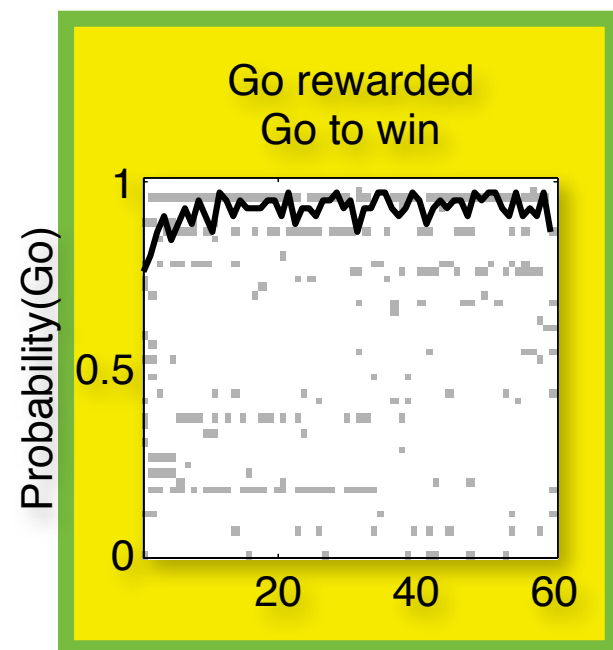
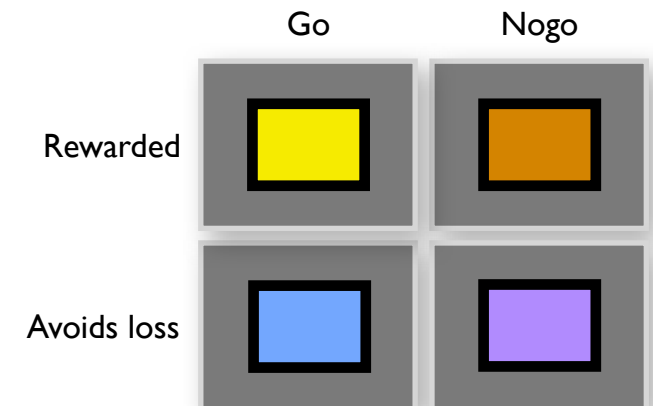
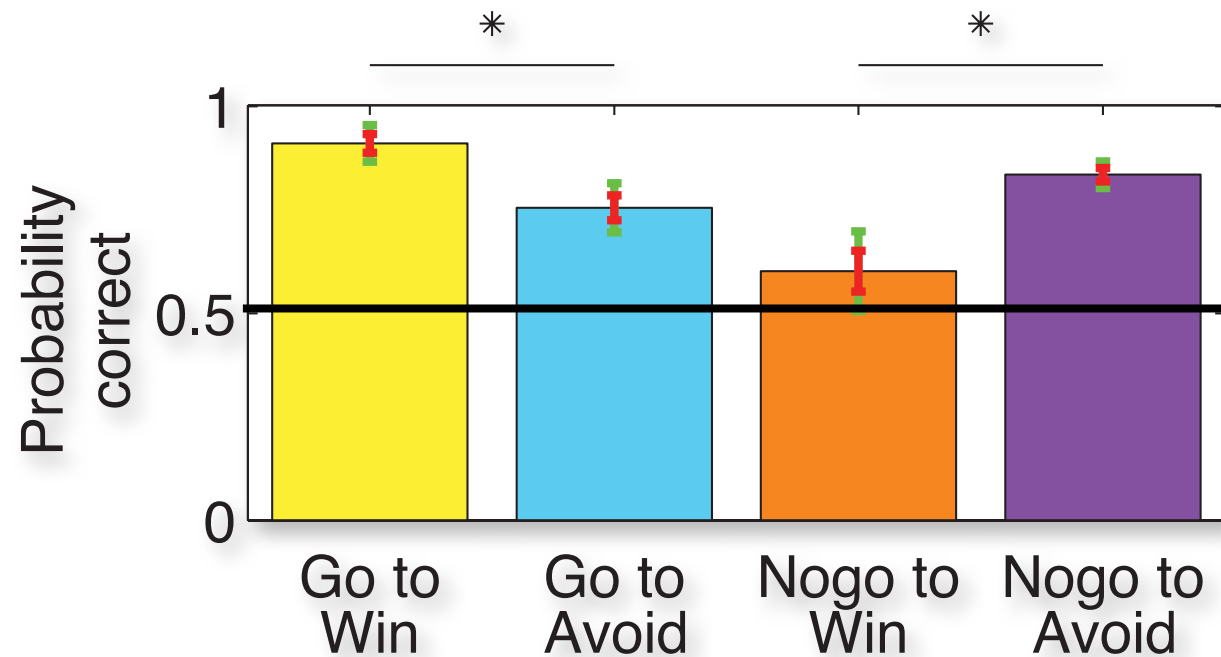


# Example task



Guitart-Masip, Huys et al. *Submitted*

# Example task



Think of it as four separate two-armed bandit tasks

Guitart-Masip, Huys et al. *Submitted*

# Analysing behaviour

---

## ► Standard approach:

- Decide which feature of the data you care about
- Run descriptive statistical tests, e.g. ANOVA

## ► Many strengths

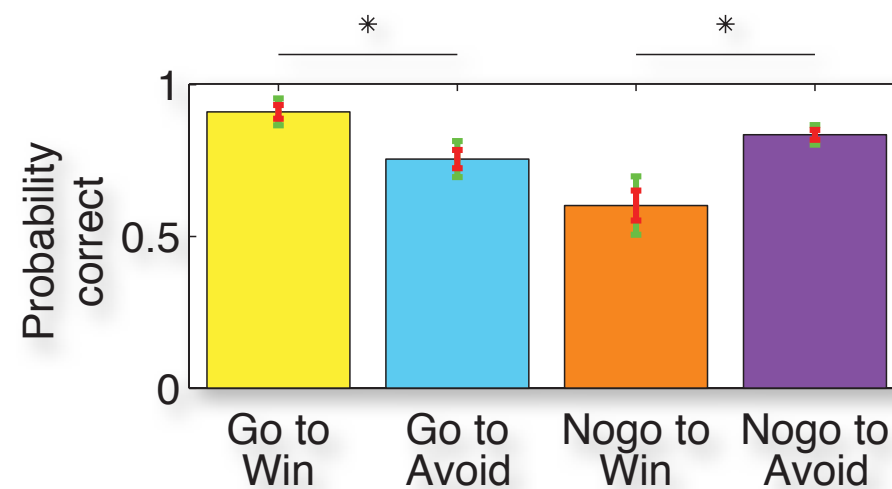
## ► Weakness

- Piecemeal, not holistic / global
- Descriptive, not generative
- No internal variables

# Analysing behaviour

## ► Standard approach:

- Decide which feature of the data you care about
- Run descriptive statistical tests, e.g. ANOVA



## ► Many strengths

## ► Weakness

- Piecemeal, not holistic / global
- Descriptive, not generative
- No internal variables

## ► Holistic

- Aim to model the process by which the data came about in its “entirety”

## ► Generative

- They can be run on the task to generate data as if a subject had done the task

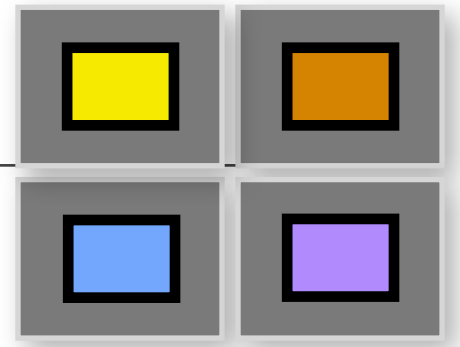
## ► Inference process

- Capture the inference process subjects have to make to perform the task.
- Do this in sufficient detail to replicate the data.

## ► Parameters

- replace test statistics
- their meaning is explicit in the model
- their contribution to the data is assessed in a holistic manner

# A simple Rescorla-Wagner model



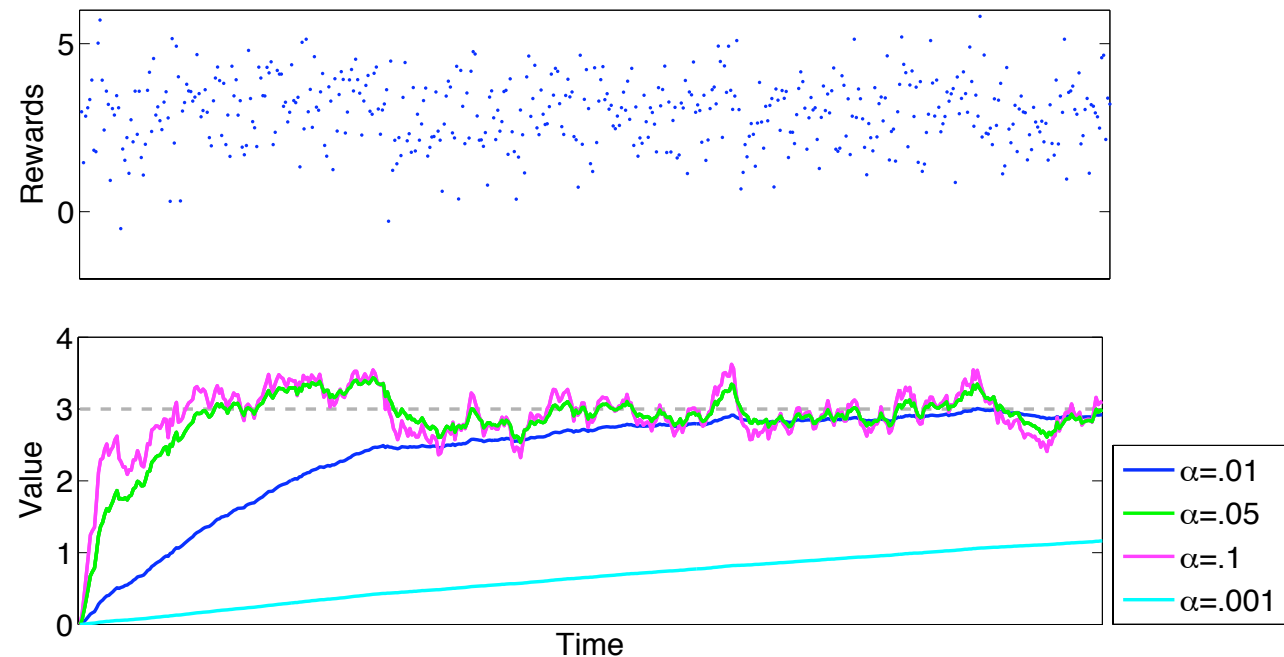
## ► Q values

$$Q_t(a_t, s_t) = Q_{t-1}(a_t, s_t) + \epsilon(r_t - Q_{t-1}(a_t, s_t))$$

$a_t$  action on trial  $t$ ; can be either 'go' or 'logo'

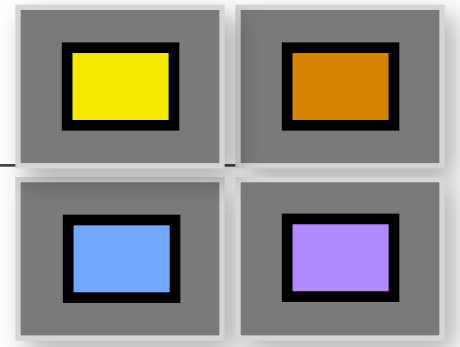
$s_t$  stimulus presented on trial  $t$

$\epsilon$  learning rate



## ► Key points:

- Q is the key part of the hypothesis
- formally states the learning process in quantitative detail
- formalizes internal quantities that are used in the task



## ► Q values

$$Q_t(a_t, s_t) = Q_{t-1}(a_t, s_t) + \epsilon(r_t - Q_{t-1}(a_t, s_t))$$

## ► Action probabilities: “softmax” of Q value

$$\begin{aligned} p(a_t | s_t, h_t, \beta) &= p(a_t | Q(a_t, s_t), \beta) \\ &= \frac{e^{\beta Q(a_t, s_t)}}{\sum_{a'} e^{\beta Q(a', s_t)}} \end{aligned}$$

## ► Features:

$$\begin{aligned} p(a_t | s_t) &\propto Q(a_t, s_t) \\ 0 &\leq p(a) \leq 1 \end{aligned}$$

## ► links learning process and observations

- choices, RTs, or any other data
- link function in GLMs
- many other forms

## ► Maximum likelihood (ML) parameters

$$\hat{\theta} = \operatorname{argmax}_{\theta} \mathcal{L}(\theta)$$

## ► where the likelihood of all choices is:

$$\begin{aligned} \mathcal{L}(\theta) &= \log p(\{a_t\}_{t=1}^T | \{s_t\}_{t=1}^T, \{r_t\}_{t=1}^T, \underbrace{\theta}_{\beta, \epsilon}) \\ &= \log p(\{a_t\}_{t=1}^T | \{Q(s_t, a_t; \epsilon)\}_{t=1}^T, \beta) \\ &= \log \prod_{t=1}^T p(a_t | Q(s_t, a_t; \epsilon), \beta) \\ &= \sum_{t=1}^T \log p(a_t | Q(s_t, a_t; \epsilon), \beta) \end{aligned}$$



# Fitting models II

---

- ▶ No closed form
- ▶ Use your favourite method
  - gradients
  - fminunc / fmincon...
- ▶ Gradients for RW model

$$\begin{aligned}\frac{d\mathcal{L}(\theta)}{d\theta} &= \frac{d}{d\theta} \sum_t \log p(a_t | \mathcal{Q}_t(a_t, s_t; \epsilon), \beta) \\ &= \sum_t \frac{d}{d\theta} \beta \mathcal{Q}_t(a_t, s_t; \epsilon) - \sum_{a'} p(a' | \mathcal{Q}_t(a', s_t; \epsilon), \beta) \frac{d}{d\theta} \beta \mathcal{Q}_t(a', s_t; \epsilon) \\ \frac{d\mathcal{Q}_t(a_t, s_t; \epsilon)}{d\epsilon} &= (1 - \epsilon) \frac{d\mathcal{Q}_{t-1}(a_t, s_t; \epsilon)}{d\epsilon} + (r_t - \mathcal{Q}_{t-1}(a_t, s_t; \epsilon))\end{aligned}$$

## ► Transform your variables

$$\beta = e^{\beta'}$$

$$\Rightarrow \beta' = \log(\beta)$$

$$\epsilon = \frac{1}{1 + e^{-\epsilon'}}$$

$$\Rightarrow \epsilon' = \log\left(\frac{\epsilon}{1 - \epsilon}\right)$$

$$\frac{d \log \mathcal{L}(\theta')}{d\theta'}$$

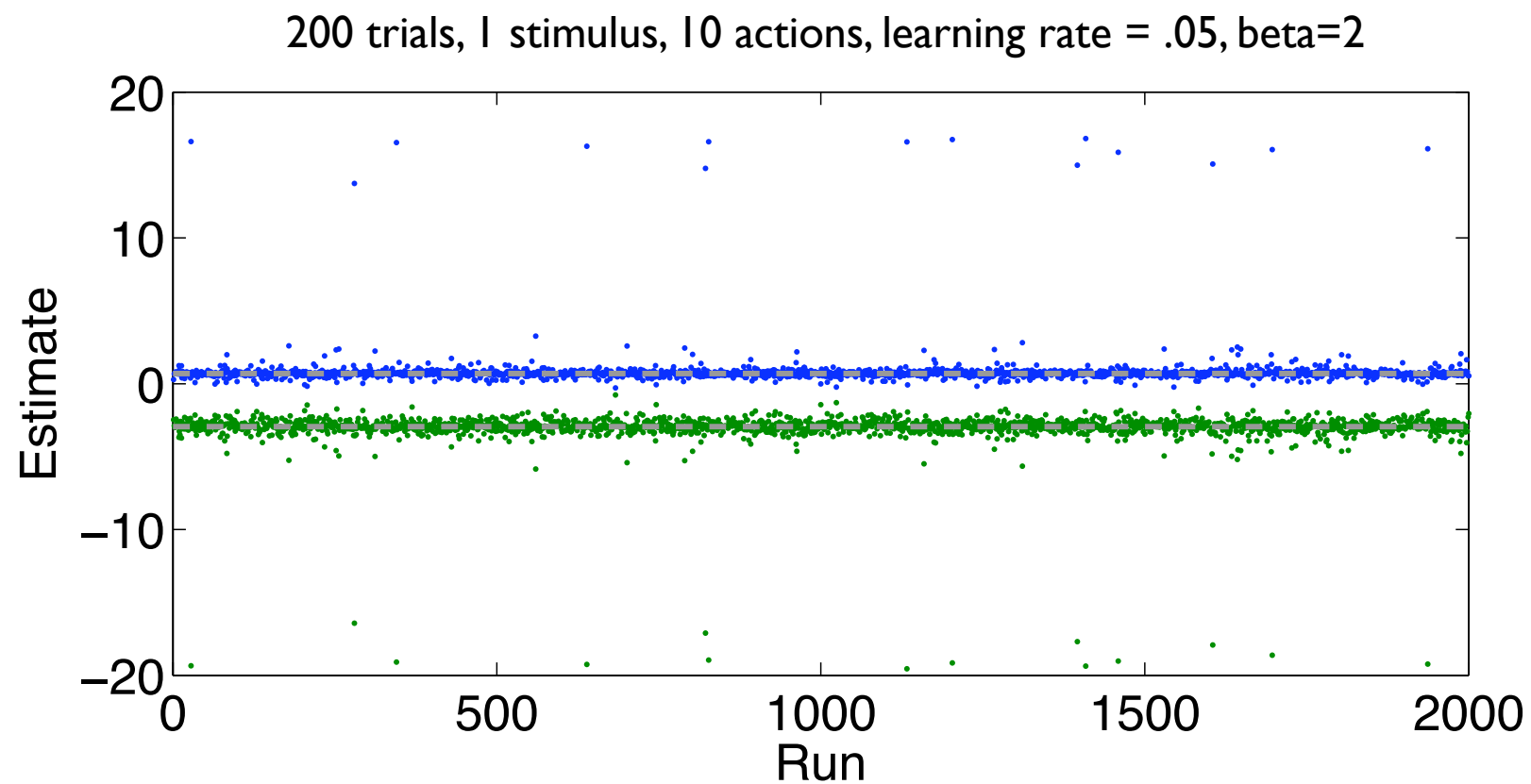
## ► Avoid over/underflow

$$y(a) = \beta \mathcal{Q}(a)$$

$$y_m = \max_a y(a)$$

$$p = \frac{e^{y(a)}}{\sum_b e^{y(b)}} = \frac{e^{y(a) - y_m}}{\sum_b e^{y(b) - y_m}}$$

# ML characteristics

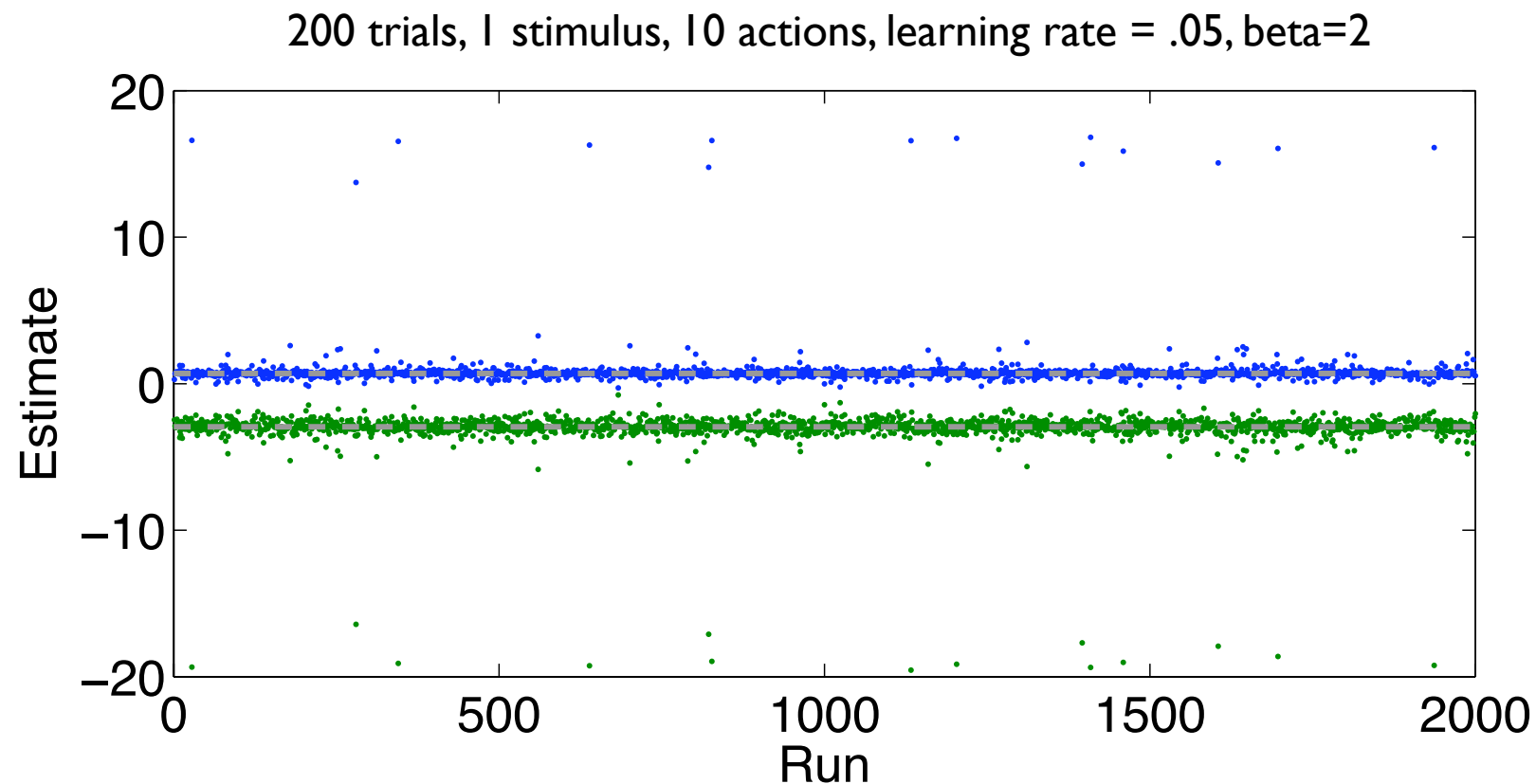


Daw 2010

$$\frac{d^2}{d\theta_i d\theta_j} \mathcal{L}(\theta)$$

# ML characteristics

- ▶ ML is asymptotically consistent, but variance high
  - 10-armed bandit, infer beta and epsilon

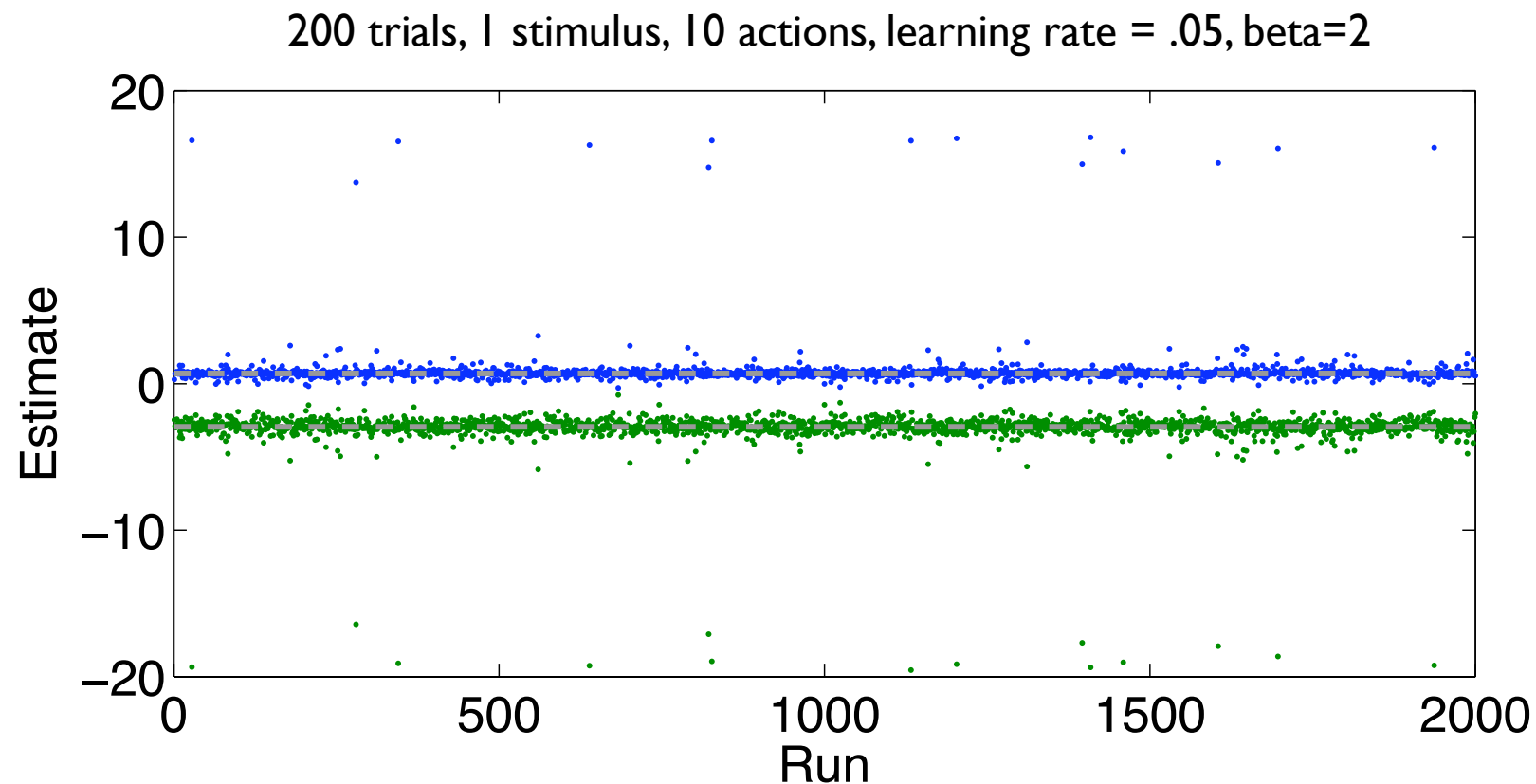


Daw 2010

- Hessian  $\frac{d^2}{d\theta_i d\theta_j} \mathcal{L}(\theta)$  can be used to derive confidence intervals and identify poorly constrained estimates

# ML characteristics

- ▶ ML is asymptotically consistent, but variance high
  - 10-armed bandit, infer beta and epsilon

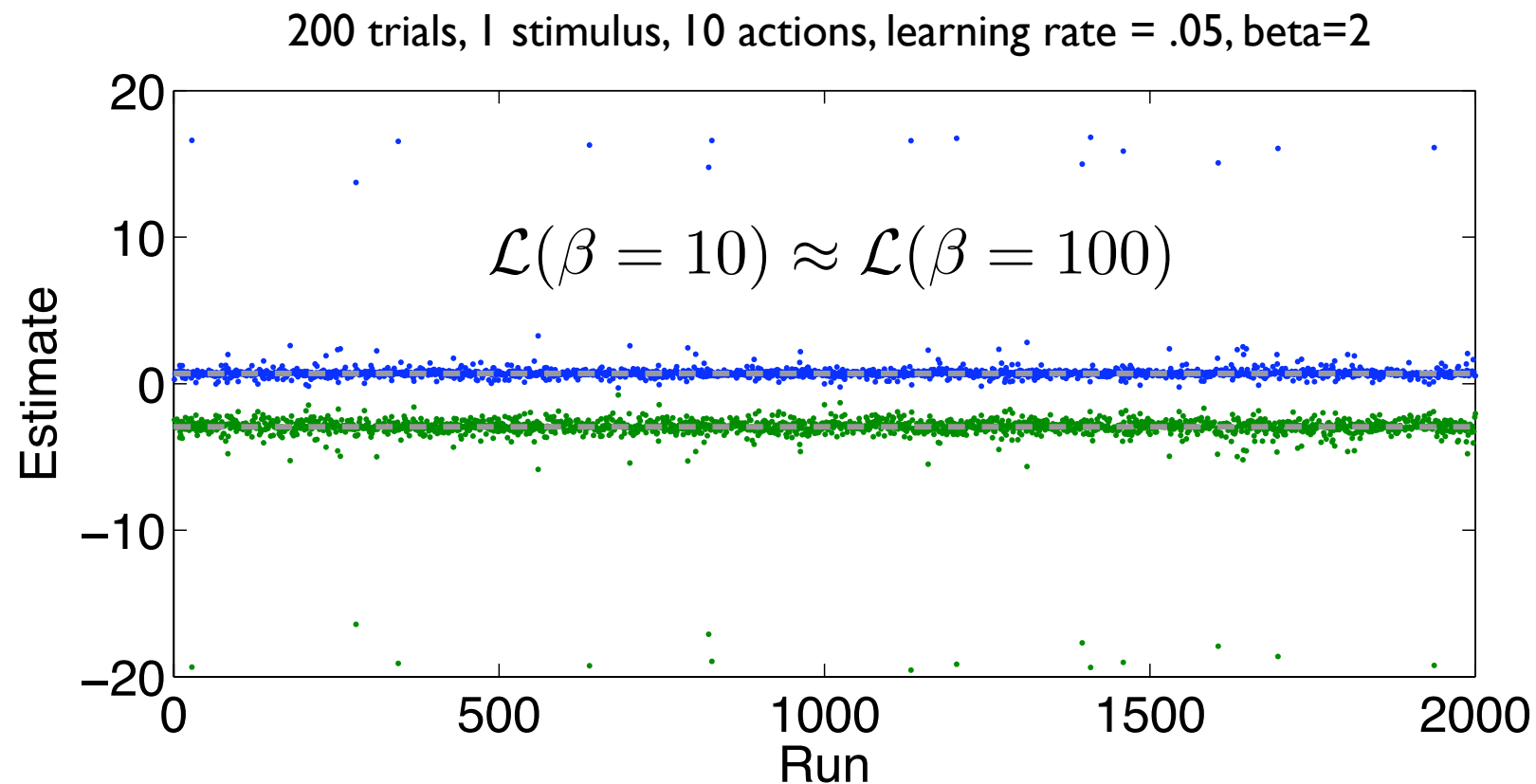


Daw 2010

- Hessian  $\frac{d^2}{d\theta_i d\theta_j} \mathcal{L}(\theta)$  can be used to derive confidence intervals and identify poorly constrained estimates
- ▶ ML can overfit... more later

# ML characteristics

- ▶ ML is asymptotically consistent, but variance high
  - 10-armed bandit, infer beta and epsilon



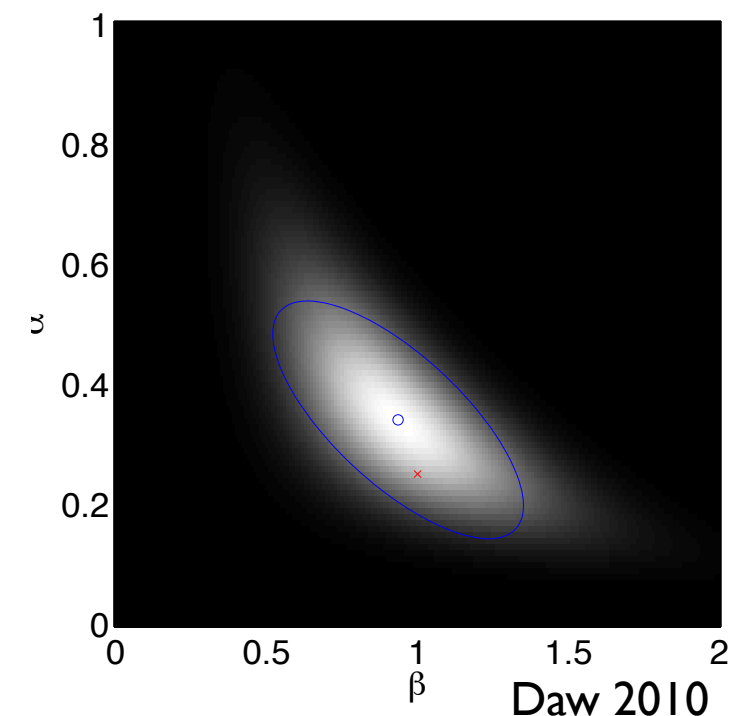
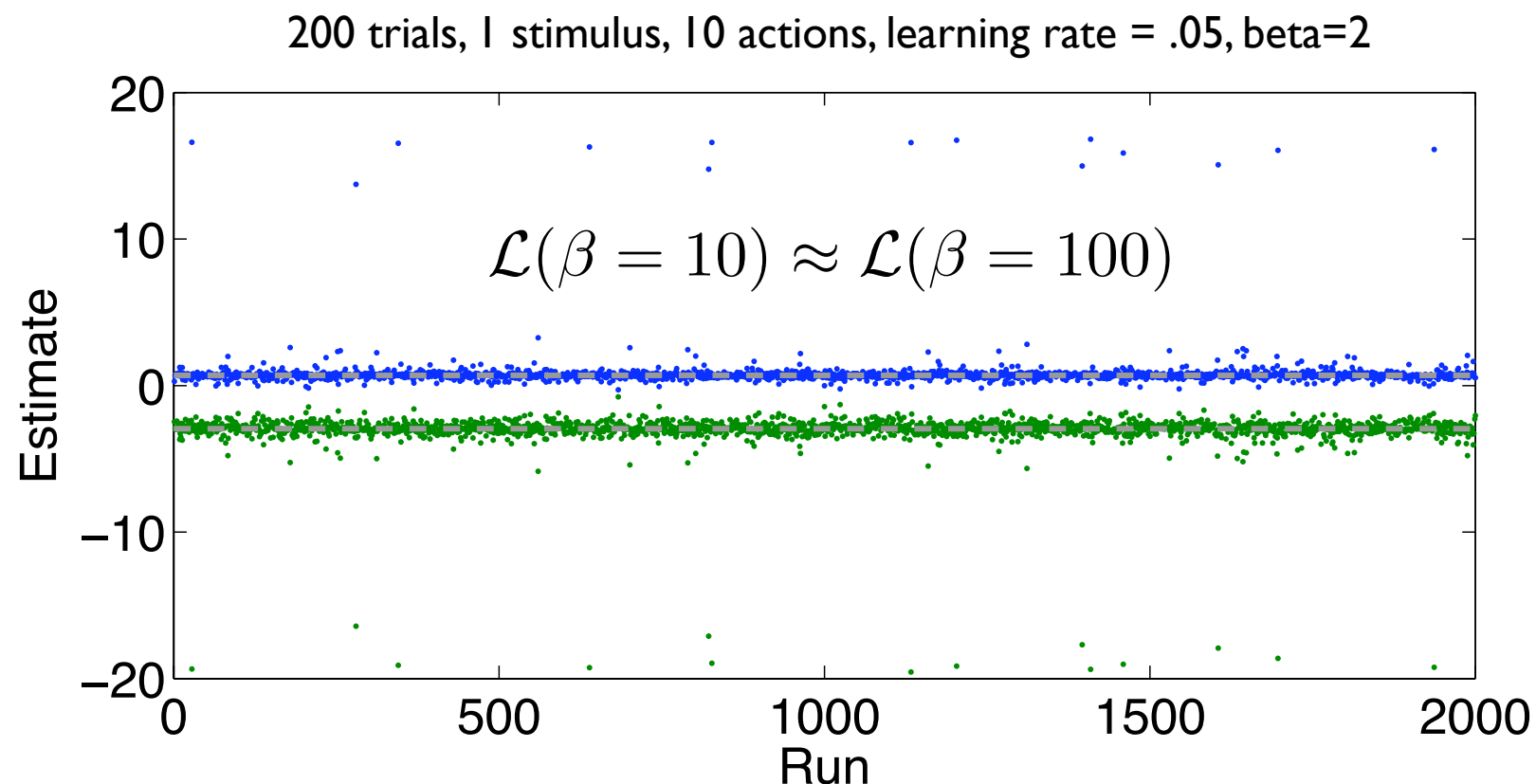
Daw 2010

- Hessian  $\frac{d^2}{d\theta_i d\theta_j} \mathcal{L}(\theta)$  can be used to derive confidence intervals and identify poorly constrained estimates
- ▶ ML can overfit... more later

# ML characteristics

- ▶ ML is asymptotically consistent, but variance high
  - 10-armed bandit, infer beta and epsilon

beta and epsilon can trade off

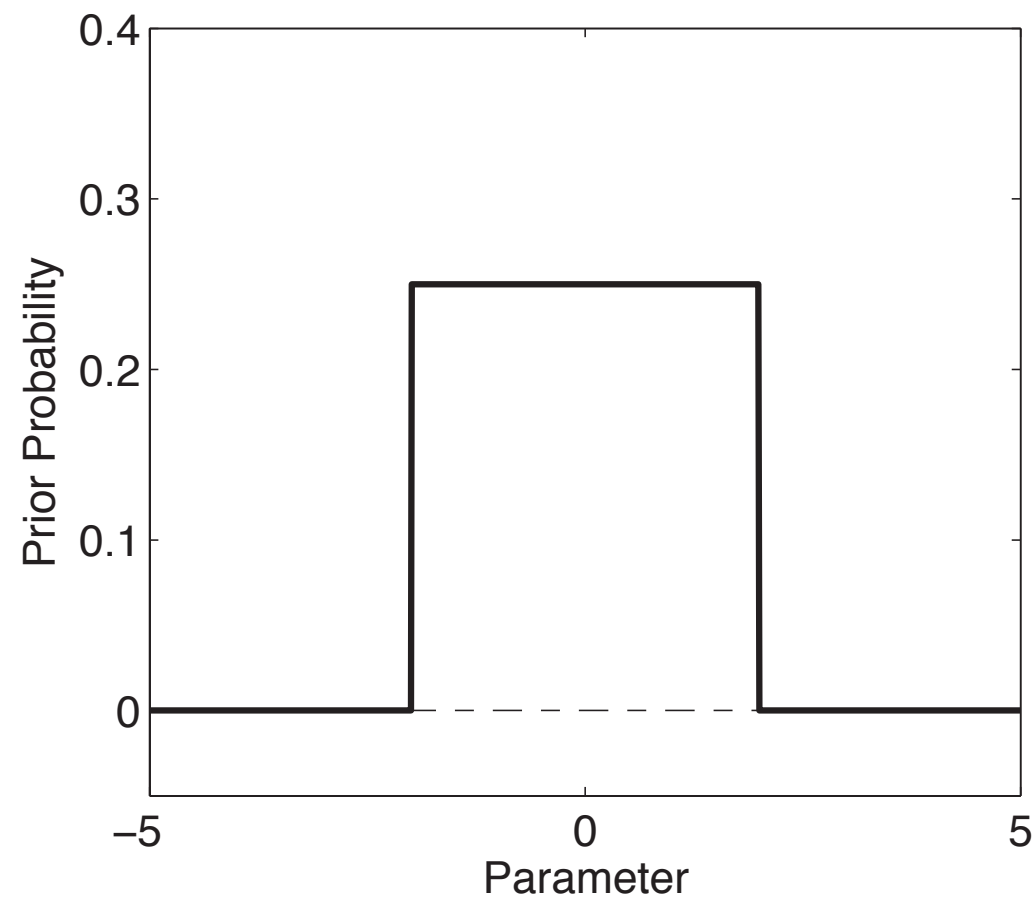


- Hessian  $\frac{d^2}{d\theta_i d\theta_j} \mathcal{L}(\theta)$  can be used to derive confidence intervals and identify poorly constrained estimates
- ▶ ML can overfit... more later

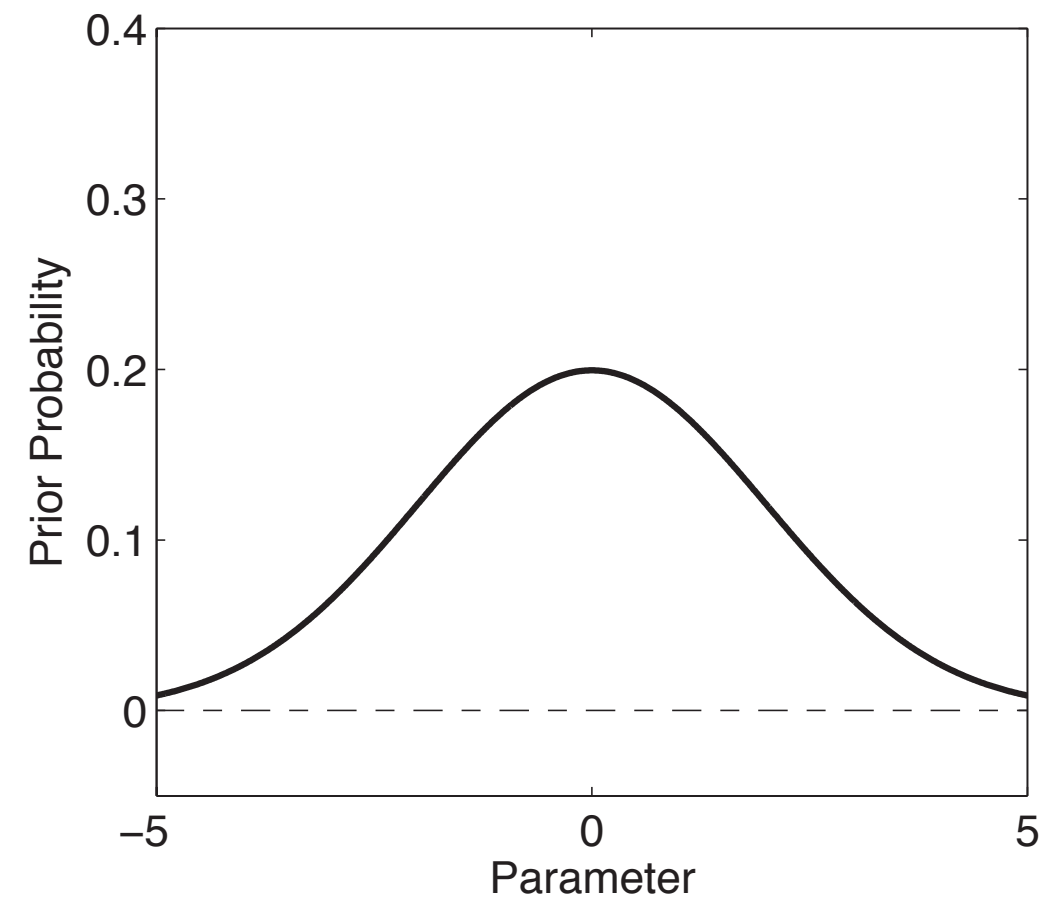
# Priors

---

Not so smooth



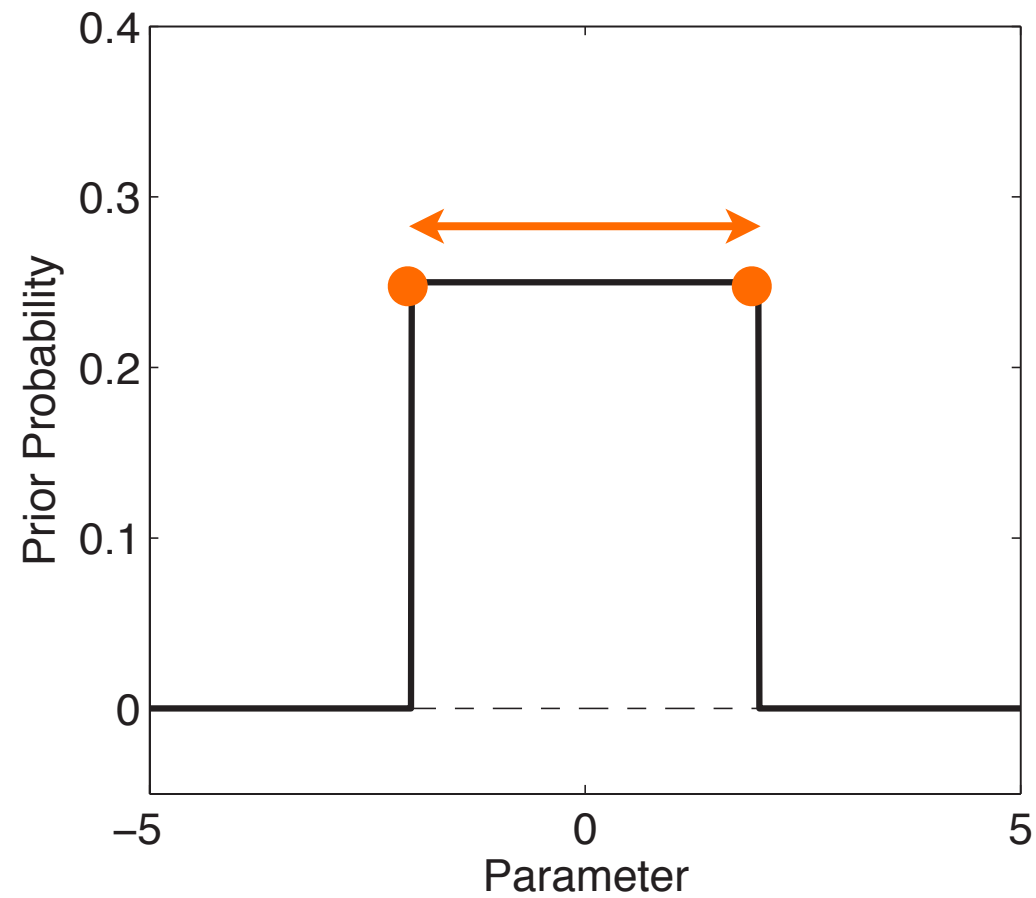
Smooth



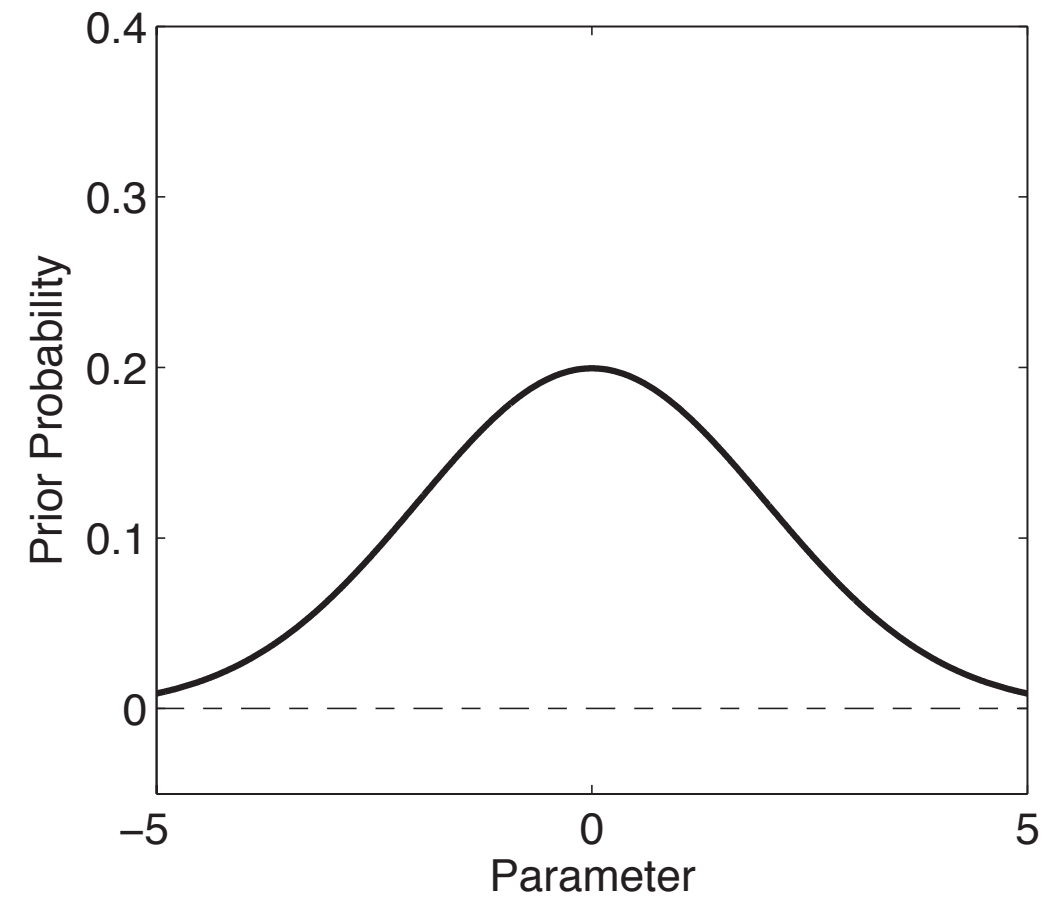


# Priors

Not so smooth

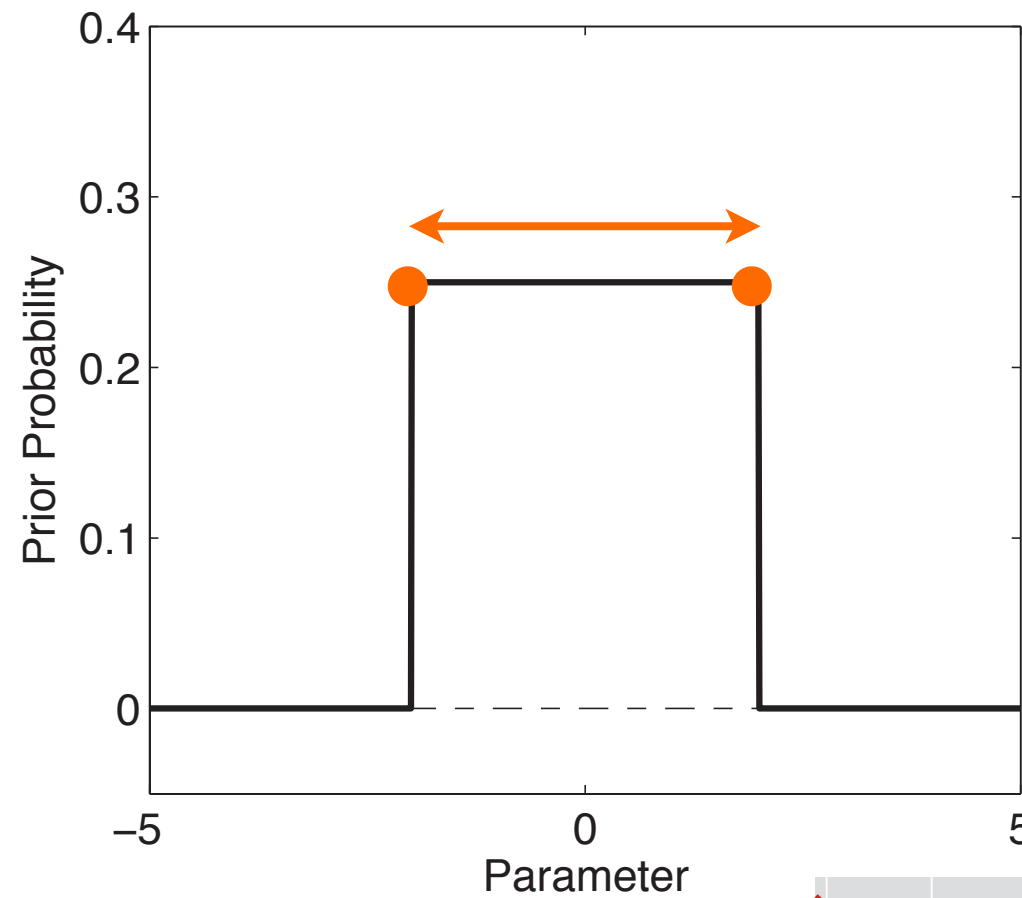


Smooth

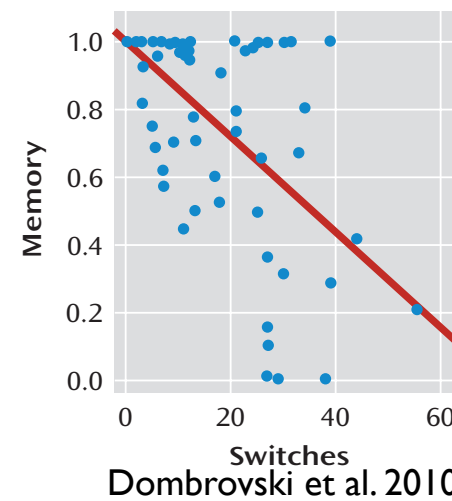
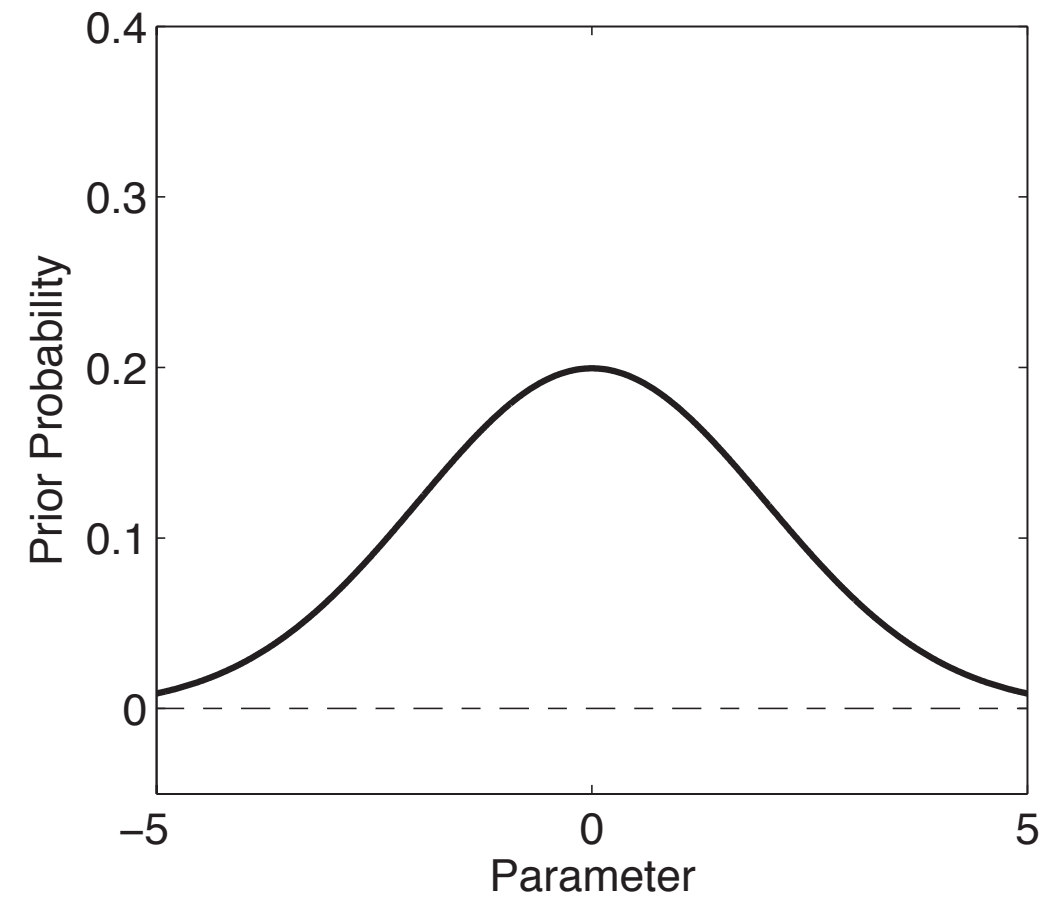


# Priors

## Not so smooth



## Smooth



# Maximum a posteriori estimate

---

$$\mathcal{P}(\theta) = p(\theta|a_{1...T}) = \frac{p(a_{1...T}|\theta)p(\theta)}{\int d\theta p(\theta|a_{1...T})p(\theta)}$$

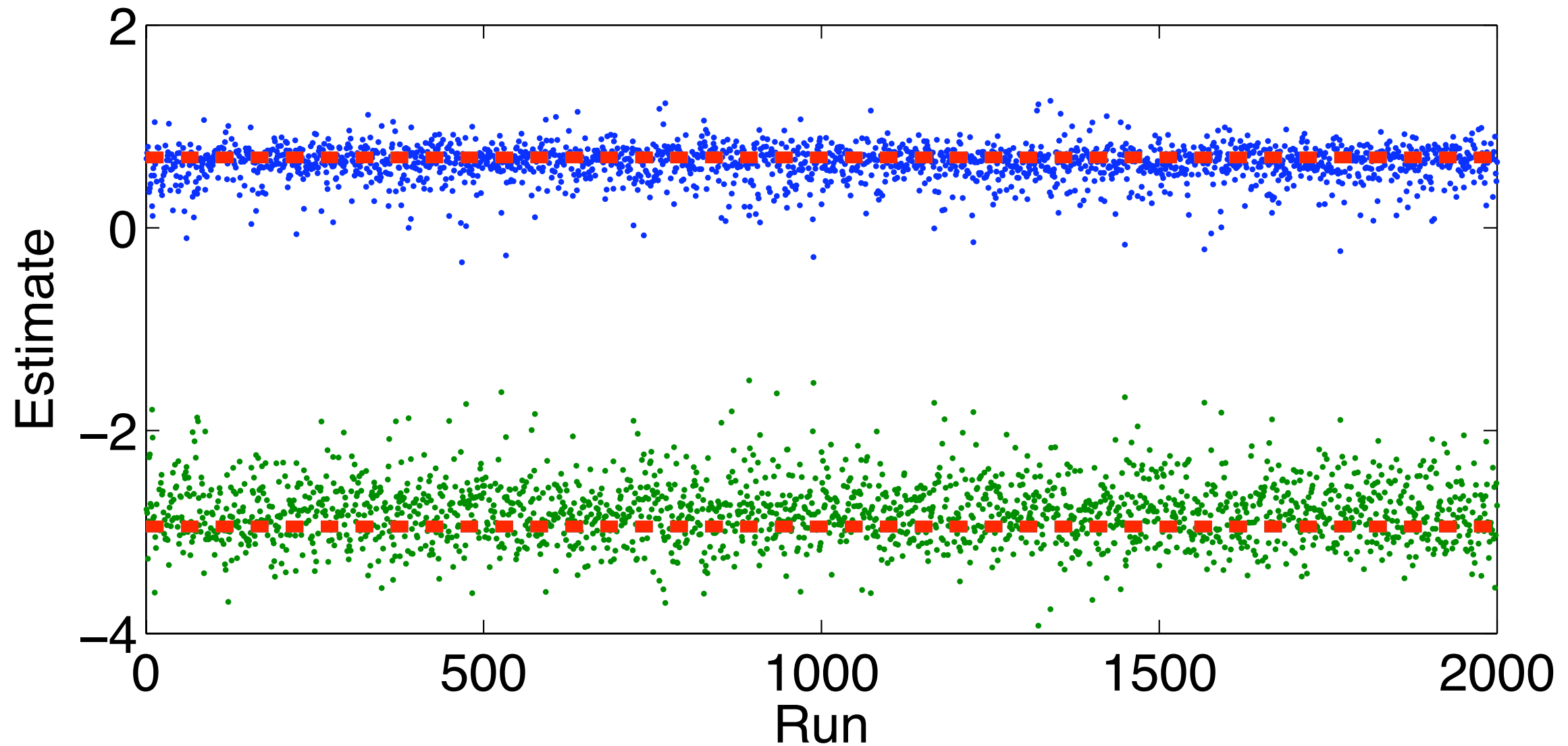
$$\log \mathcal{P}(\theta) = \sum_{t=1}^T \log p(a_t|\theta) + \log p(\theta) + \text{const.}$$

$$\frac{\log \mathcal{P}(\theta)}{d\alpha} = \frac{\log \mathcal{L}(\theta)}{d\alpha} + \frac{d p(\theta)}{d\theta}$$

- If likelihood is strong, prior will have little effect
  - mainly has influence on poorly constrained parameters
  - if a parameter is strongly constrained to be outside the typical range of the prior, then it will win over the prior

# Maximum a posteriori estimate

---



200 trials, 1 stimulus, 10 actions, learning rate = .05, beta=2  
 $m_{\text{beta}}=0$ ,  $m_{\text{eps}}=-3$ ,  $n=1$

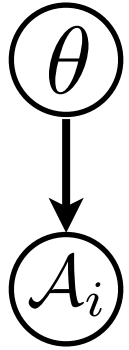
# But

---

What prior parameters should I use?

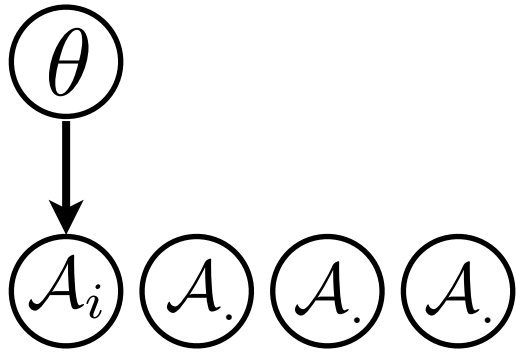
# ML characteristics: group data

---



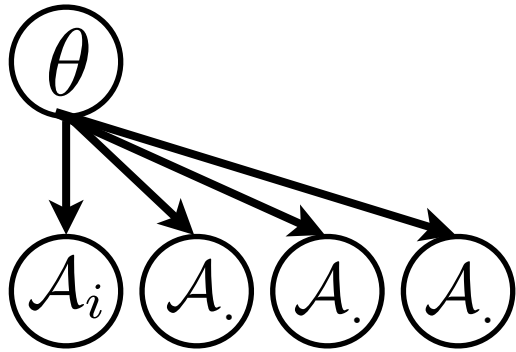
# ML characteristics: group data

---



# ML characteristics: group data

---

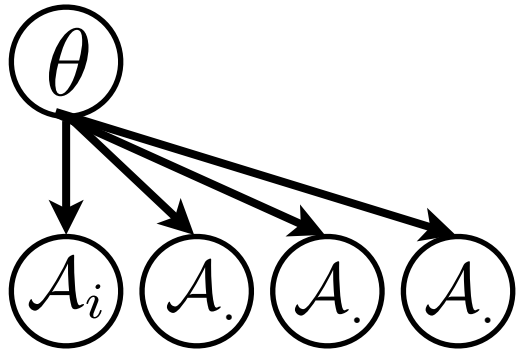


- Fixed effect
  - conflates within- and between- subject variability



# ML characteristics: group data

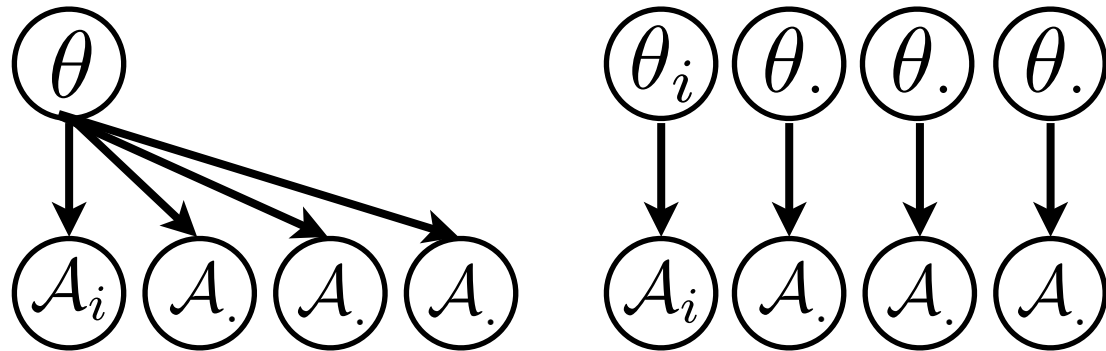
---



- ▶ **Fixed effect**
  - conflates within- and between- subject variability
- ▶ **Average behaviour**
  - disregards between-subject variability
  - need to adapt model

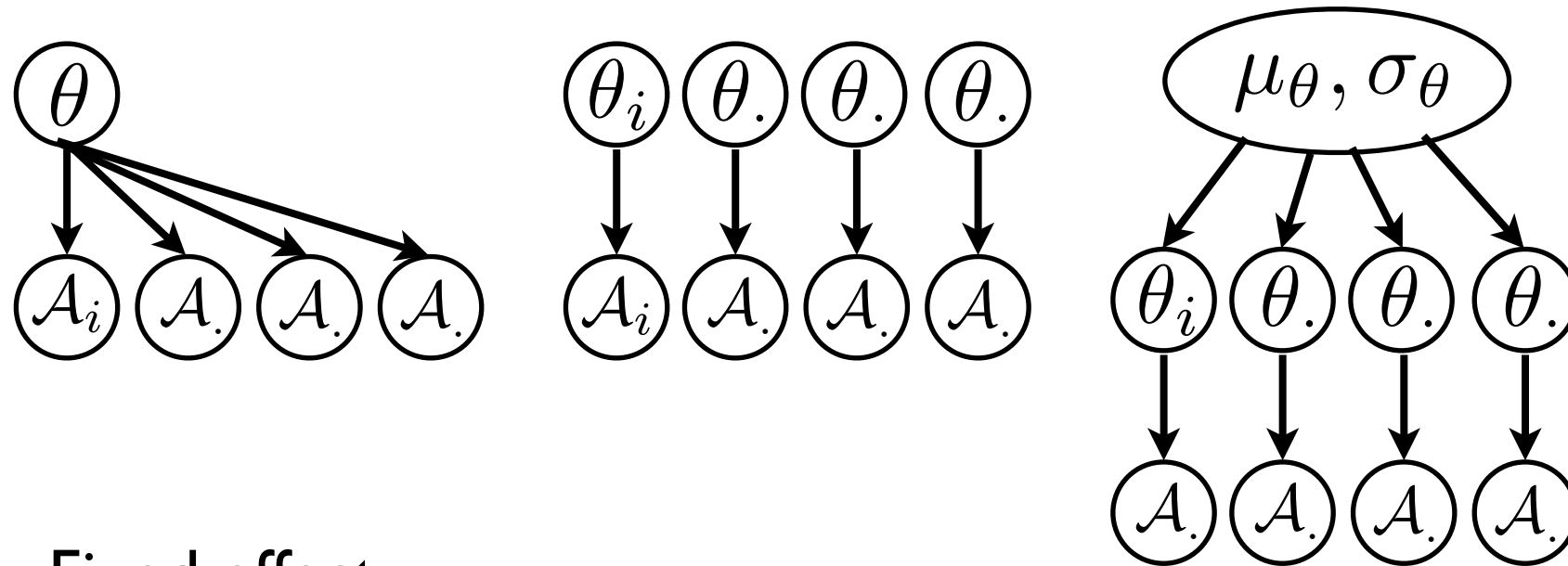
# ML characteristics: group data

---



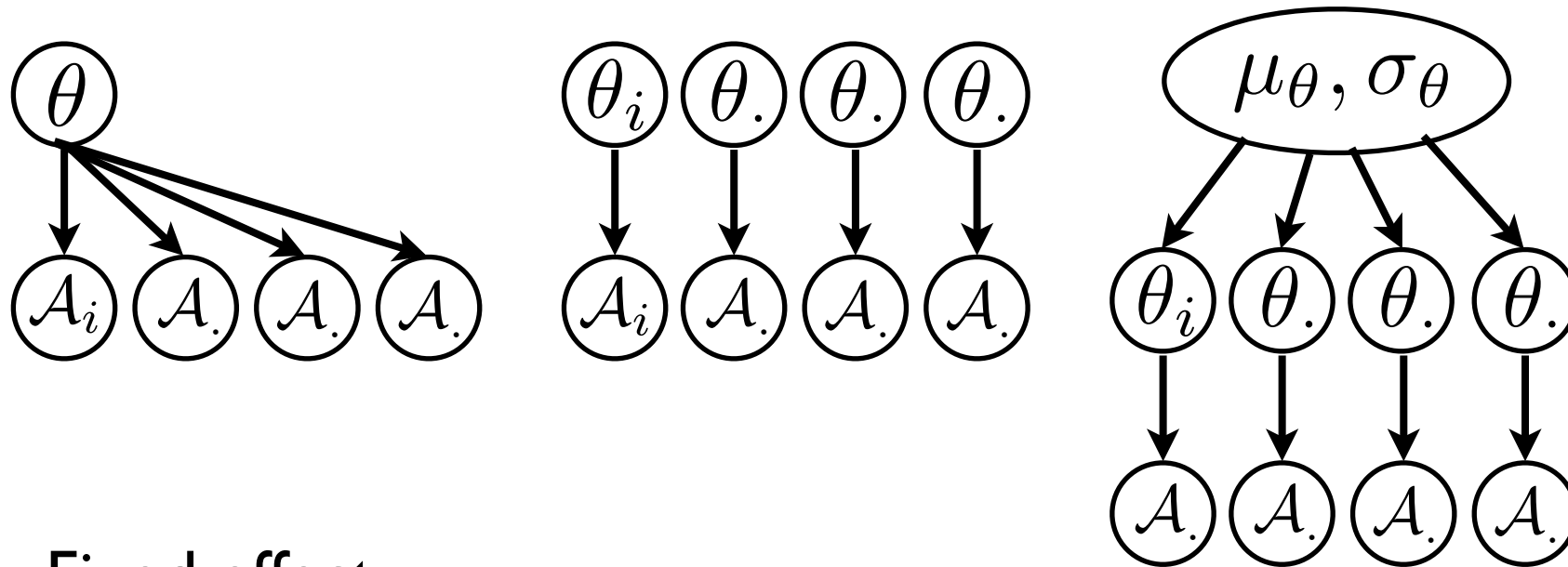
- ▶ **Fixed effect**
  - conflates within- and between- subject variability
- ▶ **Average behaviour**
  - disregards between-subject variability
  - need to adapt model
- ▶ **Summary statistic**
  - treat parameters as random variable, one for each subject
  - overestimates group variance as ML estimates noisy

# ML characteristics: group data



- ▶ **Fixed effect**
  - conflates within- and between- subject variability
- ▶ **Average behaviour**
  - disregards between-subject variability
  - need to adapt model
- ▶ **Summary statistic**
  - treat parameters as random variable, one for each subject
  - overestimates group variance as ML estimates noisy
- ▶ **Random effects**
  - prior mean = group mean

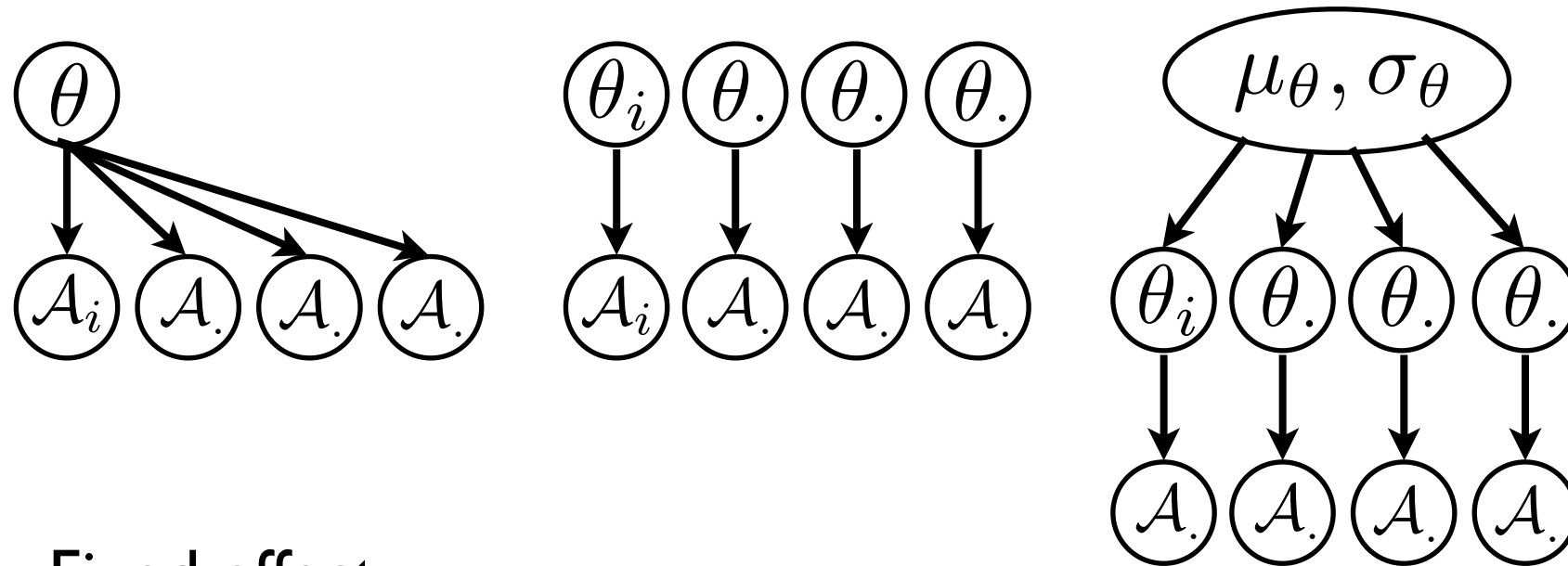
# ML characteristics: group data



- **Fixed effect**
  - conflates within- and between- subject variability
- **Average behaviour**
  - disregards between-subject variability
  - need to adapt model
- **Summary statistic**
  - treat parameters as random variable, one for each subject
  - overestimates group variance as ML estimates noisy
- **Random effects**
  - prior mean = group mean

$$p(\mathcal{A}_i | \mu_\theta, \sigma_\theta) = \int d\theta_i p(\mathcal{A}_i | \theta_i) p(\theta_i | \mu_\theta, \sigma_\theta)$$

# ML characteristics: group data



## ► Fixed effect

- conflates within- and between- subject variability

## ► Average behaviour

- disregards between-subject variability
- need to adapt model

## ► Summary statistic

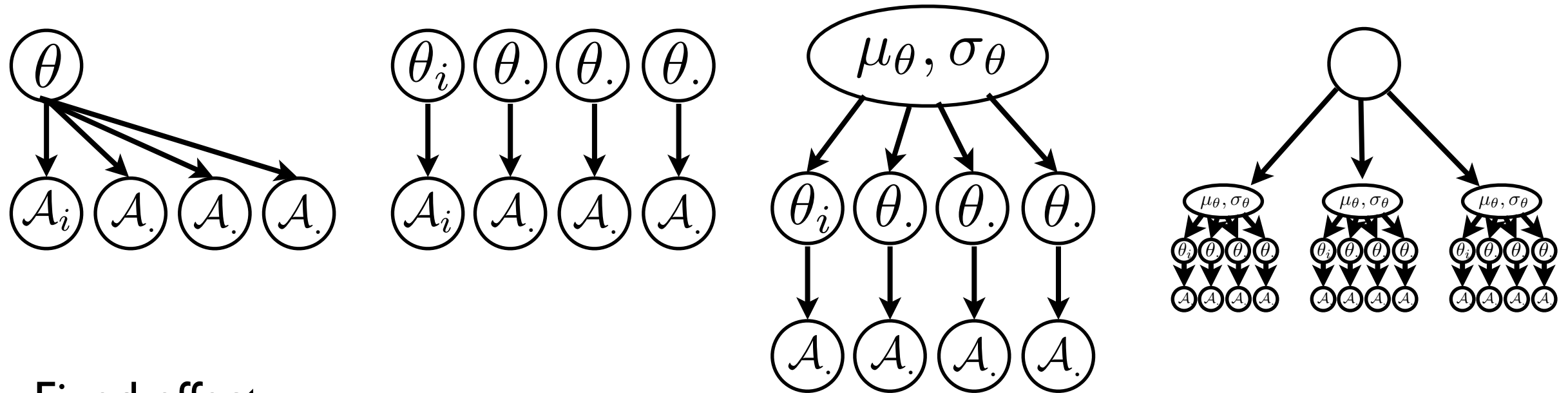
- treat parameters as random variable, one for each subject
- overestimates group variance as ML estimates noisy

## ► Random effects

- prior mean = group mean

$$p(\mathcal{A}_i | \mu_\theta, \sigma_\theta) = \int d\theta_i p(\mathcal{A}_i | \theta_i) p(\theta_i | \underbrace{\mu_\theta, \sigma_\theta}_{\zeta})$$

# ML characteristics: group data



## ► Fixed effect

- conflates within- and between- subject variability

## ► Average behaviour

- disregards between-subject variability
- need to adapt model

## ► Summary statistic

- treat parameters as random variable, one for each subject
- overestimates group variance as ML estimates noisy

## ► Random effects

- prior mean = group mean

$$p(\mathcal{A}_i | \mu_\theta, \sigma_\theta) = \int d\theta_i p(\mathcal{A}_i | \theta_i) p(\theta_i | \underbrace{\mu_\theta, \sigma_\theta}_{\zeta})$$

# Estimating the hyperparameters

---

## ► MAP

$$\log \mathcal{P}(\theta) = \mathcal{L}(\theta) + \log \underbrace{p(\theta)}_{=p(\theta|\zeta)} + \text{const.}$$

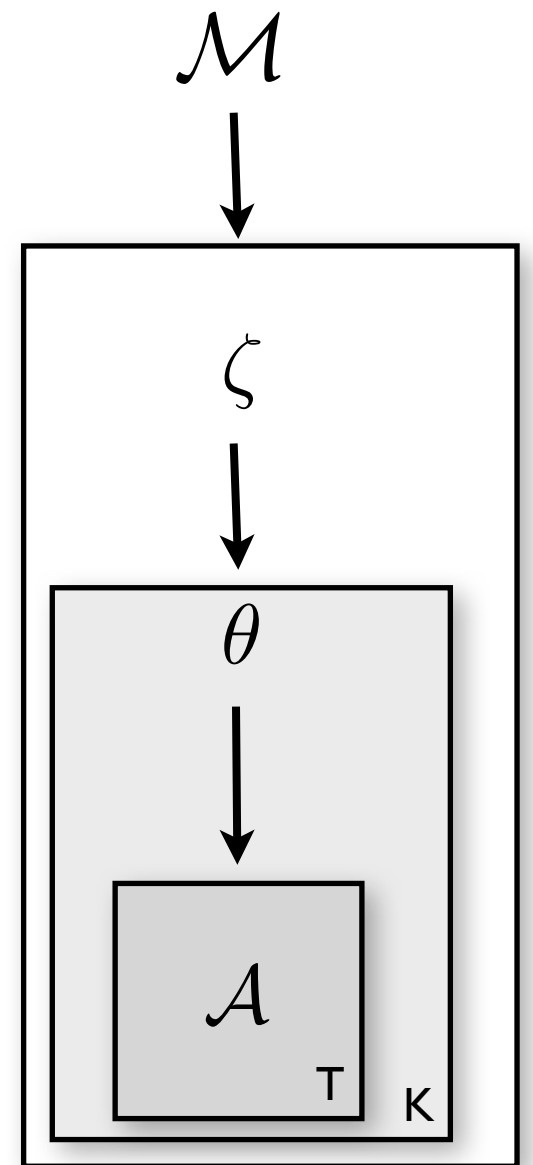
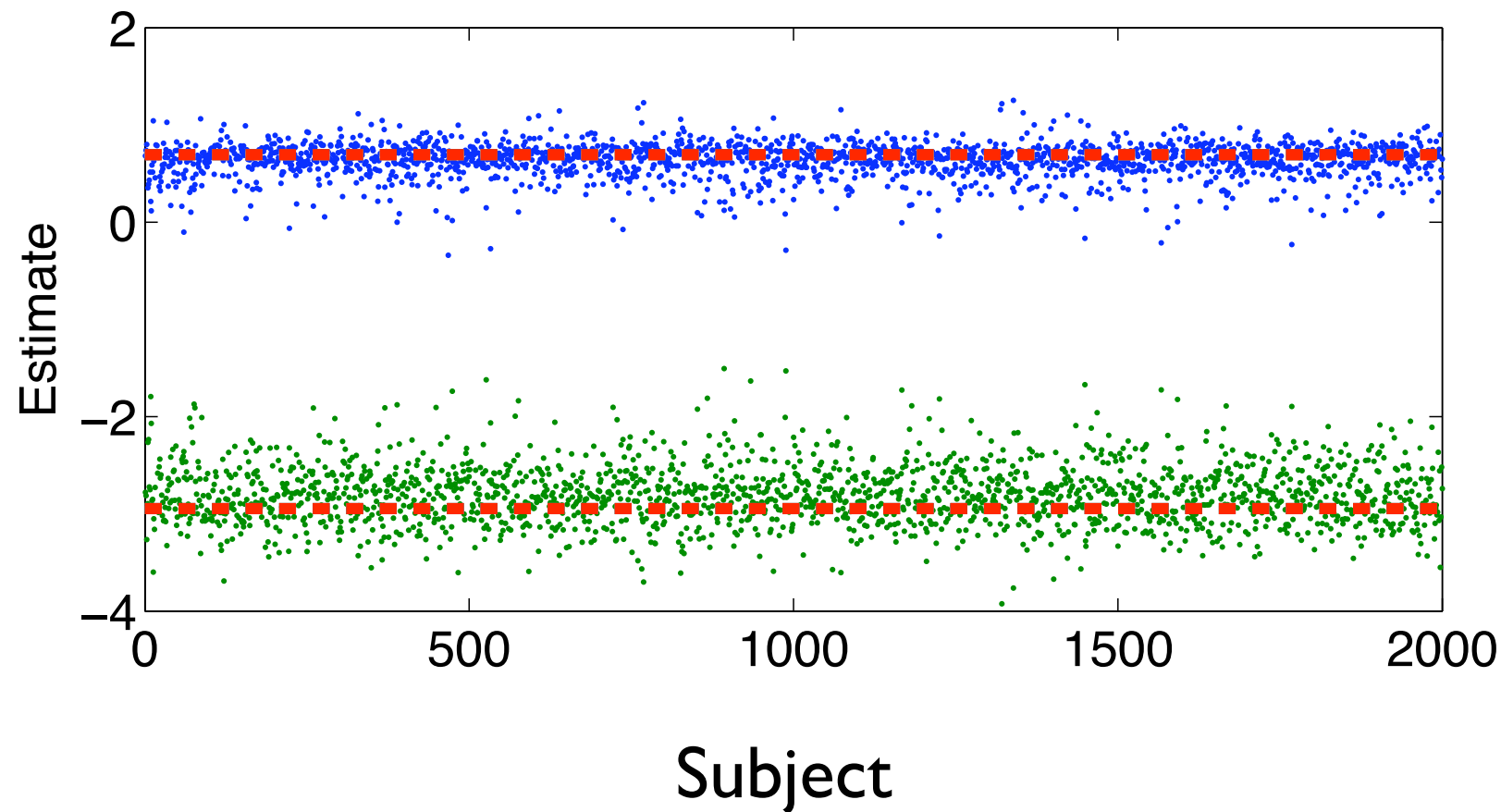
## ► Empirical Bayes: set them to ML estimate

$$\hat{\zeta} = \underset{\zeta}{\operatorname{argmax}} p(\mathcal{A}|\zeta)$$

## ► where we use all the actions by all the $k$ subjects

$$\mathcal{A} = \{a_{1\dots T}^k\}_{k=1}^K$$

# ML estimate of top-level parameters



$$\hat{\zeta} = \underset{\zeta}{\operatorname{argmax}} p(\mathcal{A}|\zeta)$$



# Estimating the hyperparameters

---

- Effectively we now want to do gradient ascent on:

$$\frac{d}{d\zeta} p(\mathcal{A}|\zeta)$$

- But this contains an integral over individual parameters:

$$p(\mathcal{A}|\zeta) = \int d\theta p(\mathcal{A}|\theta) p(\theta|\zeta)$$

- So we need to:

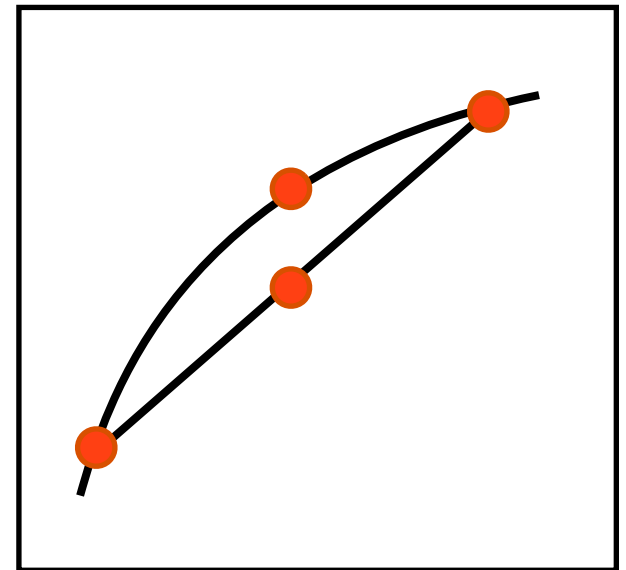
$$\begin{aligned}\hat{\zeta} &= \operatorname{argmax}_{\zeta} p(\mathcal{A}|\zeta) \\ &= \operatorname{argmax}_{\zeta} \int d\theta p(\mathcal{A}|\theta) p(\theta|\zeta)\end{aligned}$$

# Expectation Maximisation

$$\begin{aligned}\log p(\mathcal{A}|\zeta) &= \log \int d\theta p(\mathcal{A}, \theta|\zeta) \\ &= \log \int d\theta q(\theta) \frac{p(\mathcal{A}, \theta|\zeta)}{q(\theta)} \\ &\geq \int d\theta q(\theta) \log \frac{p(\mathcal{A}, \theta|\zeta)}{q(\theta)}\end{aligned}$$

$$k^{\text{th}} \text{ E step: } q^{(k+1)}(\theta) \leftarrow p(\theta|\mathcal{A}, \zeta^{(k)})$$

$$k^{\text{th}} \text{ M step: } \zeta^{(k+1)} \leftarrow \operatorname{argmax}_{\zeta} \int d\theta q(\theta) \log p(\mathcal{A}, \theta|\zeta)$$



Jensen's inequality

► There are other approaches


- Monte Carlo
- Analytical conjugate priors
- Variational Bayes

► Iterate between

- Estimating MAP parameters given prior parameters
- Estimating prior parameters from MAP parameters

# EM with Laplace approximation

- ▶ **E step:**  $q^{(k+1)}(\theta) \leftarrow p(\theta|\mathcal{A}, \zeta^{(k)})$ 
  - only need sufficient statistics to perform M step
  - Approximate  $p(\theta|\mathcal{A}, \zeta^{(k)}) \sim \mathcal{N}(\mathbf{m}_k, \mathbf{S}_k)$
  - and hence:

$$\begin{aligned} \text{E step: } q_k(\theta) &= \mathcal{N}(\mathbf{m}_k, \mathbf{S}_k) \\ \mathbf{m}_k &\leftarrow \underset{\theta}{\operatorname{argmax}} p(\mathbf{a}_k|\theta)p(\theta|\zeta^{(i)}) \\ \mathbf{S}_k^{-1} &\leftarrow \left. \frac{\partial^2 p(\mathbf{a}^k|\theta)p(\theta|\zeta^{(i)})}{\partial \theta^2} \right|_{\theta=\mathbf{m}_k} \end{aligned}$$


matlab: `[m,L,,,S]=fminunc(...)`

Just what we had before: MAP inference given some prior parameters

# EM with Laplace approximation

## ► Next update the prior

Prior mean = mean of MAP estimates

M step:

$$\zeta_{\mu}^{(i+1)} = \frac{1}{K} \sum_k \mathbf{m}_k$$
$$\zeta_{\nu^2}^{(i+1)} = \frac{1}{N} \sum_i \left[ (\mathbf{m}_k)^2 + \mathbf{S}_k \right] - (\zeta_{\mu}^{(i+1)})^2$$

Prior variance depends on inverse Hessian  $\mathbf{S}$  and variance of MAP estimates

Take uncertainty of estimates into account

## ► And now iterate until convergence

# Hierarchical / random effects models

---

## ► Advantages

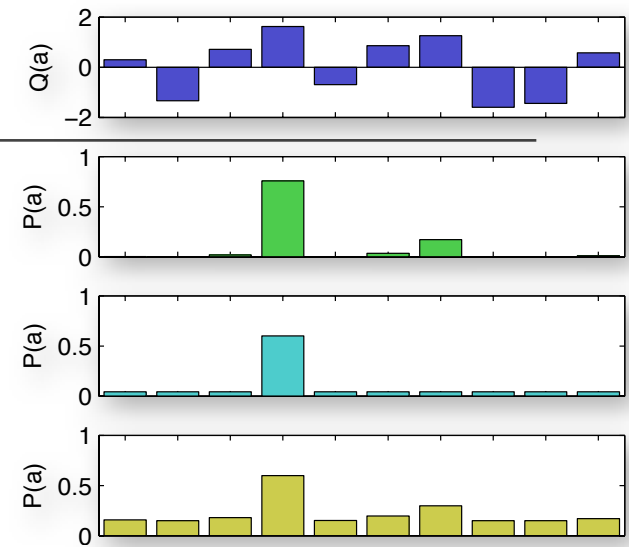
- Accurate group-level mean and variance
- Outliers due to weak likelihood are regularized
- Strong outliers are not
- Useful for model selection

## ► Disadvantages

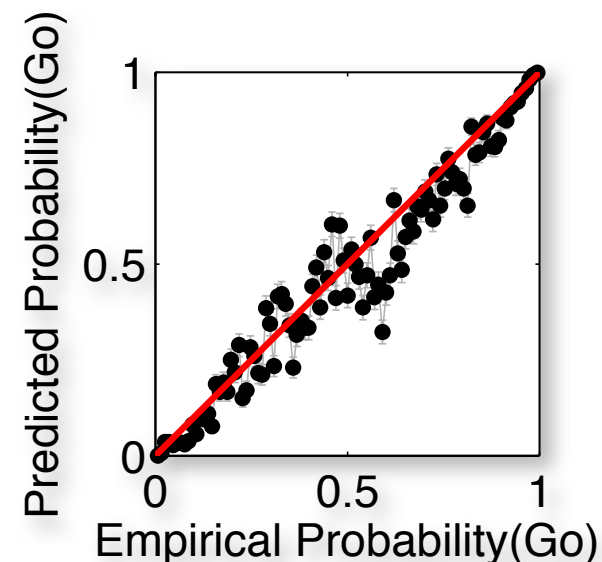
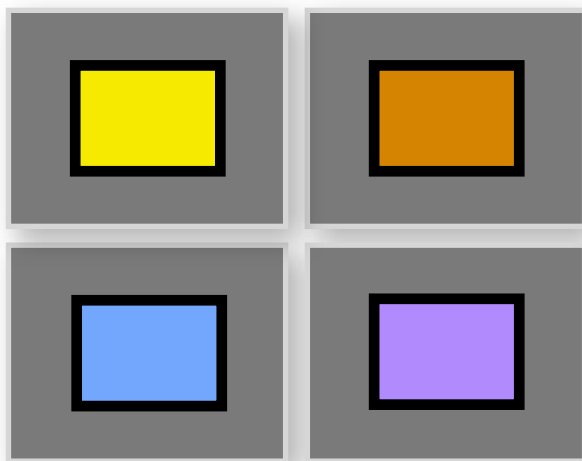
- Individual estimates  $\theta_i$  depend on other data, i.e. on  $\mathcal{A}_{j \neq i}$  and therefore need to be careful in interpreting these as summary statistics
- Error bars on group parameters (especially group variance) are difficult to obtain
- More involved; less transparent

# Link functions

- ▶ **Sigmoid**  $p(a|s) = \frac{e^{\beta Q(a,s)}}{\sum_{a'} e^{\beta Q(a',s)}}$
- ▶  **$\epsilon$ -greedy**  $p(a|s) = \begin{cases} c & \text{if } a = \operatorname{argmax}_a Q(a, s) \\ \frac{1-c}{|a|-1} & \text{else} \end{cases}$
- ▶ **irreducible noise**  $p(a|s) = \frac{1-g}{2} + g \frac{e^{\beta Q(a,s)}}{\sum_{a'} e^{\beta Q(a',s)}}$



- ▶ **critical sanity check I: reasonable link function?**



- ▶ **other link functions for other observations**

# Model comparison

---

- ▶ A fit by itself is not meaningful
- ▶ Generative test
  - qualitative
- ▶ Comparisons
  - vs random
  - vs other model -> test specific hypotheses and isolate particular effects in a generative setting

# Model fit: likelihood

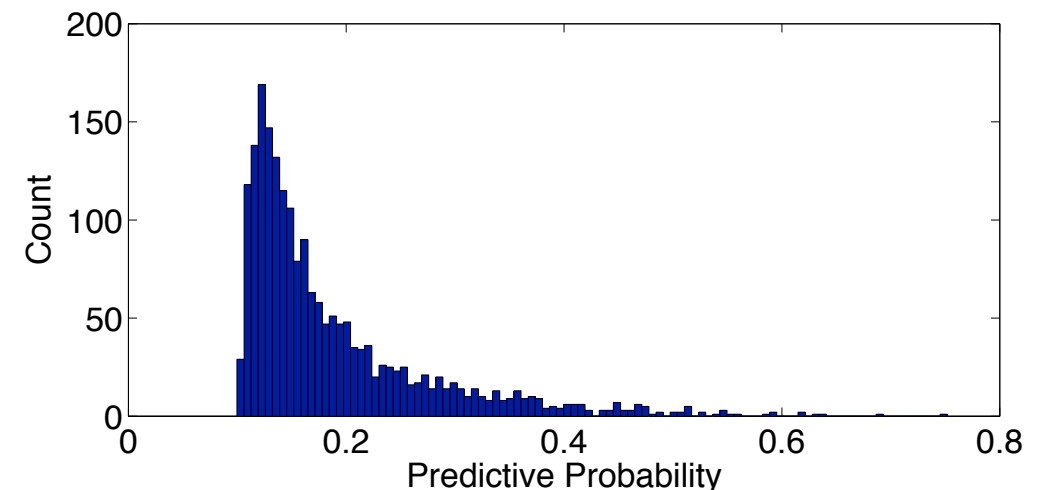
## ► How well does the model do?

- choice probabilities:

$$\begin{aligned}\mathbb{E}p(\text{correct}) &= e^{\mathcal{L}(\hat{\theta})/K/T} \\ &= e^{\log p(\mathcal{A}|\theta)/K/T} \\ &= \left( \prod_{k,t=1}^{K,T} p(a_{k,t}|\theta_k) \right)^{\frac{1}{KT}}\end{aligned}$$

“Predictive probabilities”

- typically around 0.65-0.75 for 2-way choice
- for 10-armed bandit example
- pseudo- $r^2$ :  $1 - L/R$
- better than chance?



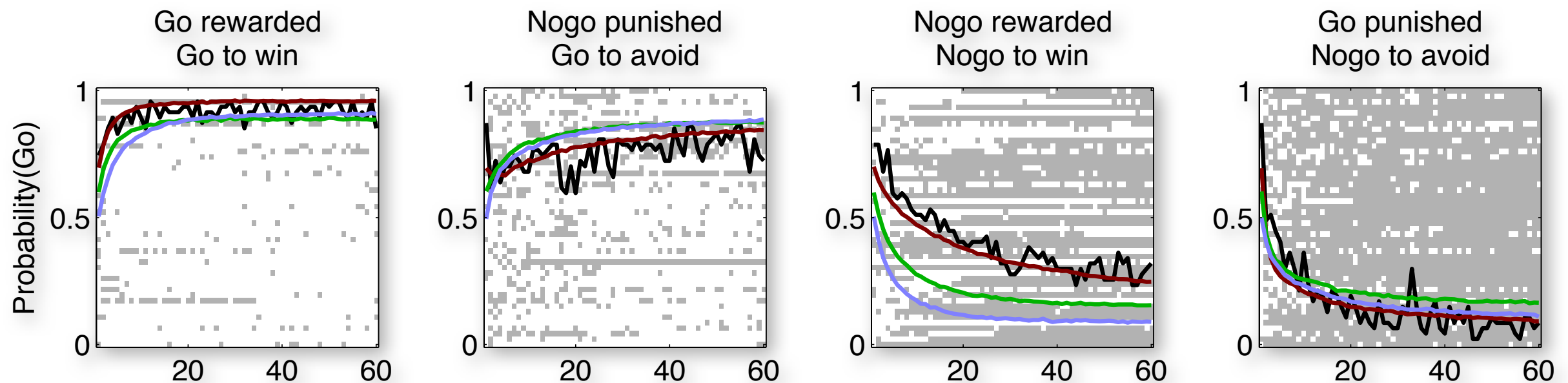
$$\mathbb{E}[N_k(\text{correct})] = \mathbb{E}[p_k(\text{correct})]T$$

$$p_{bin}(\mathbb{E}[N_k(\text{correct})] | N_k d, p_0 = 0.5) < 1 - \alpha$$



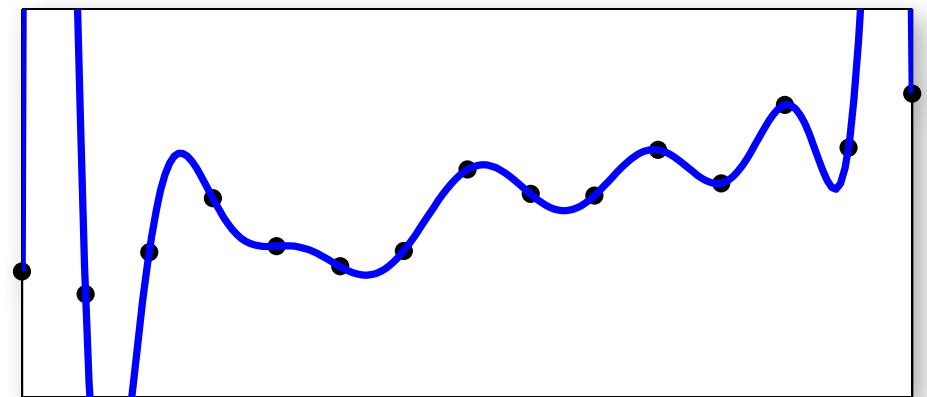
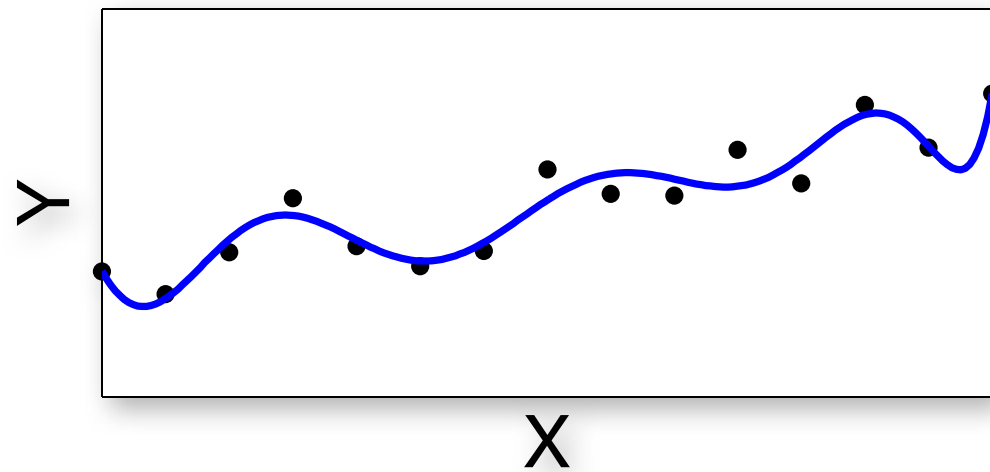
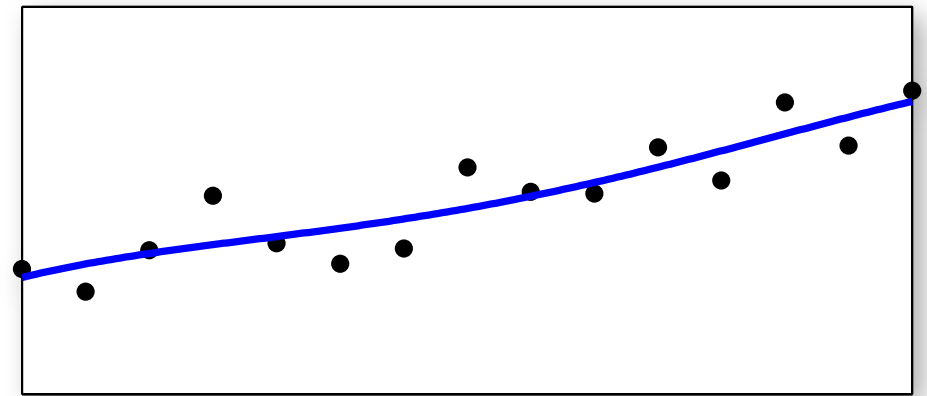
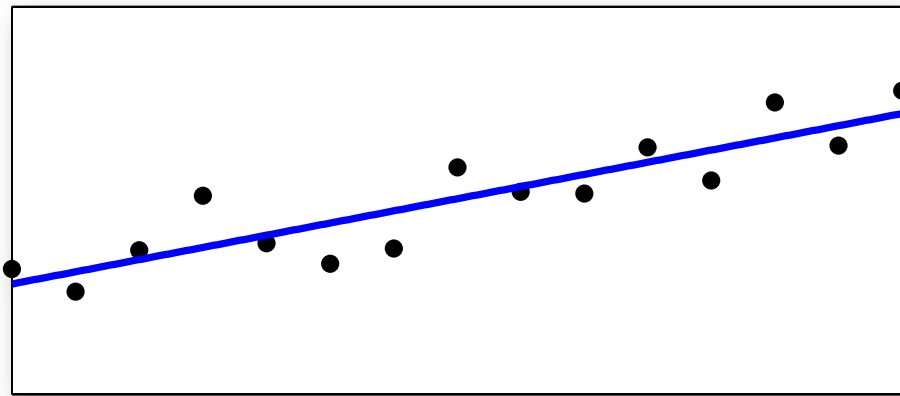
# Generative test

- Model: probability(actions)
  - simply draw from this distribution, and see what happens

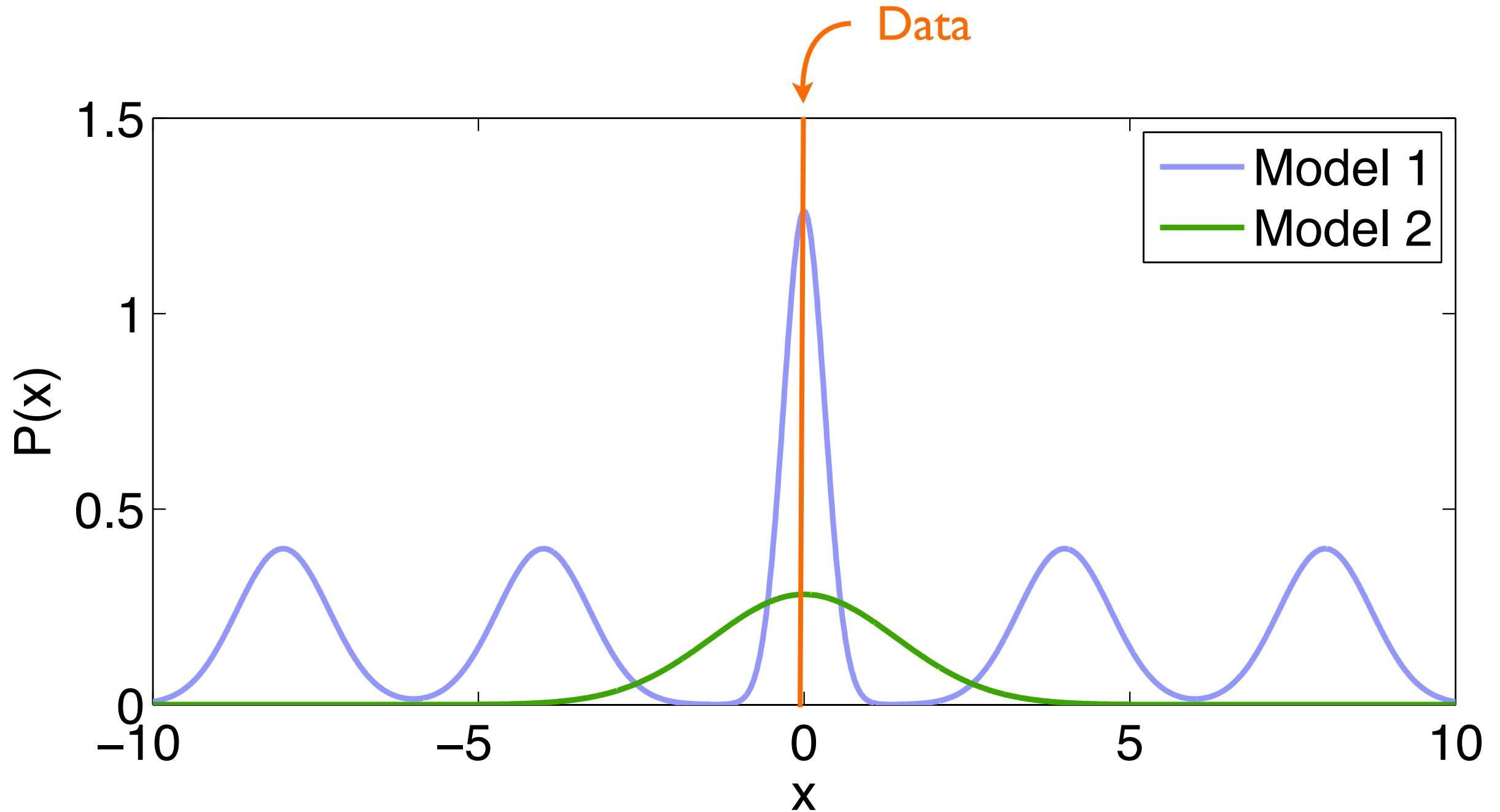


- Critical sanity test: is the model meaningful?
- Caveat: overfitting

# Overfitting



# Model comparison



# Model comparison

---

- ▶ Averaged over its parameter settings, how well does the model fit the data?

$$p(\mathcal{A}|\mathcal{M}) = \int d\theta p(\mathcal{A}|\theta) p(\theta|\mathcal{M})$$

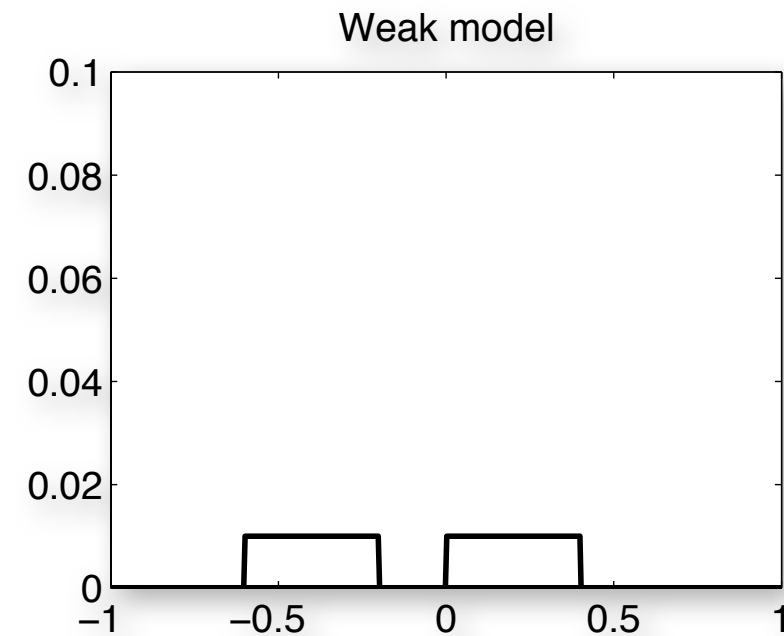
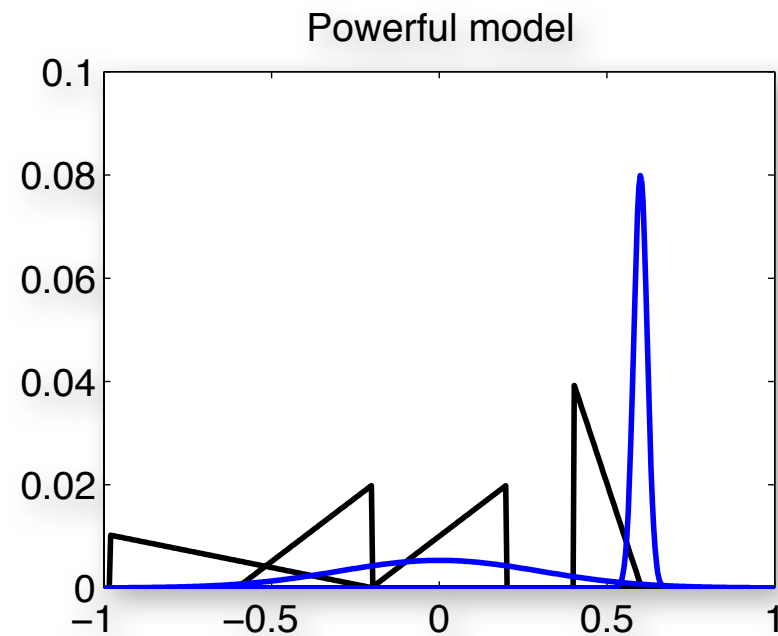
- ▶ Model comparison: Bayes factors

$$BF = \frac{p(\mathcal{M}_1|\mathcal{A})}{p(\mathcal{M}_0|\mathcal{A})} = \frac{p(\mathcal{A}|\mathcal{M}_1) p(\mathcal{M}_1)}{p(\mathcal{A}|\mathcal{M}_2) p(\mathcal{M}_2)}$$

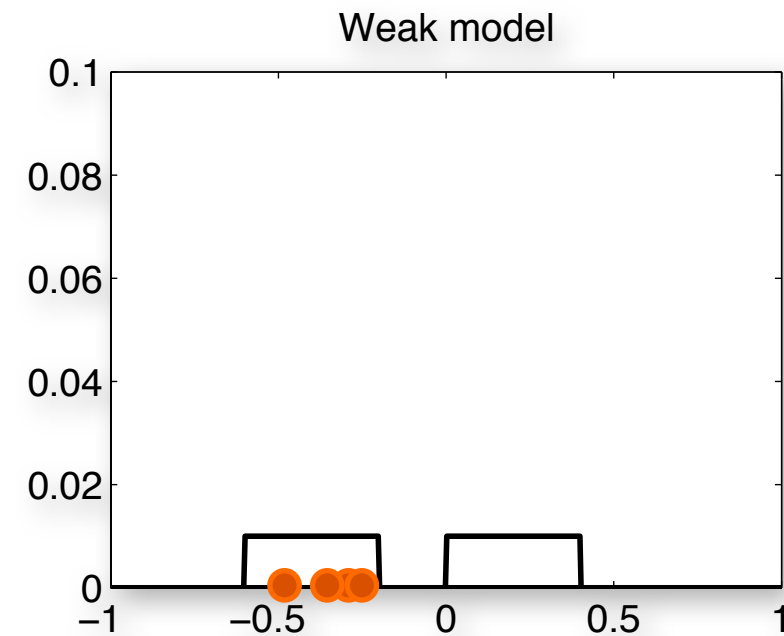
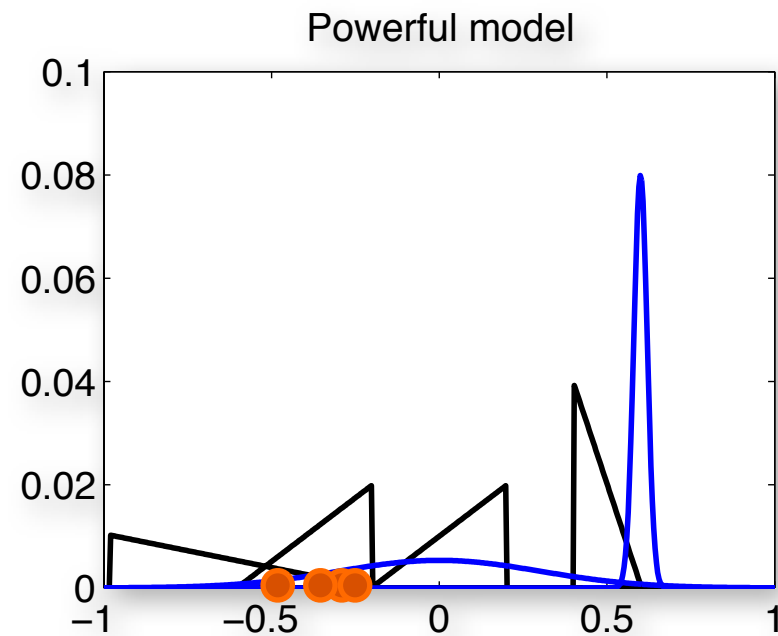
- ▶ Problem:

- integral rarely solvable
- approximation: Laplace, sampling, variational...

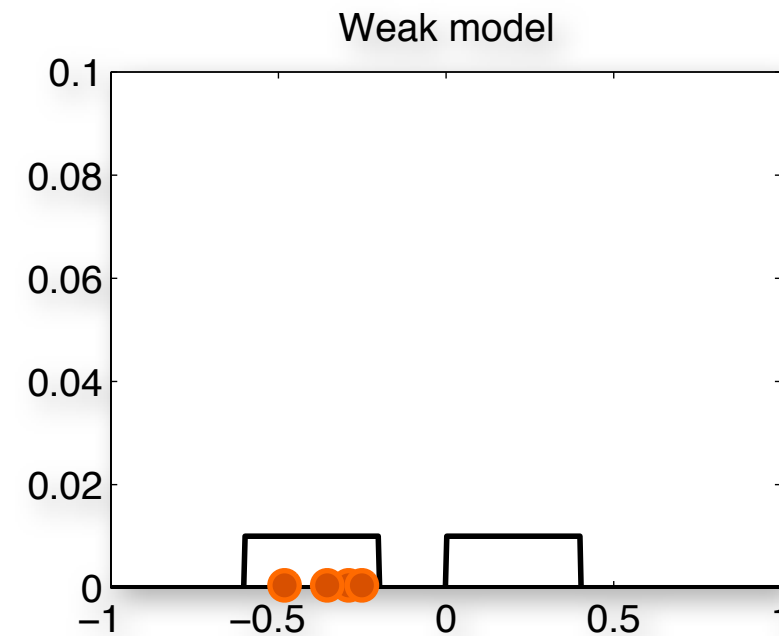
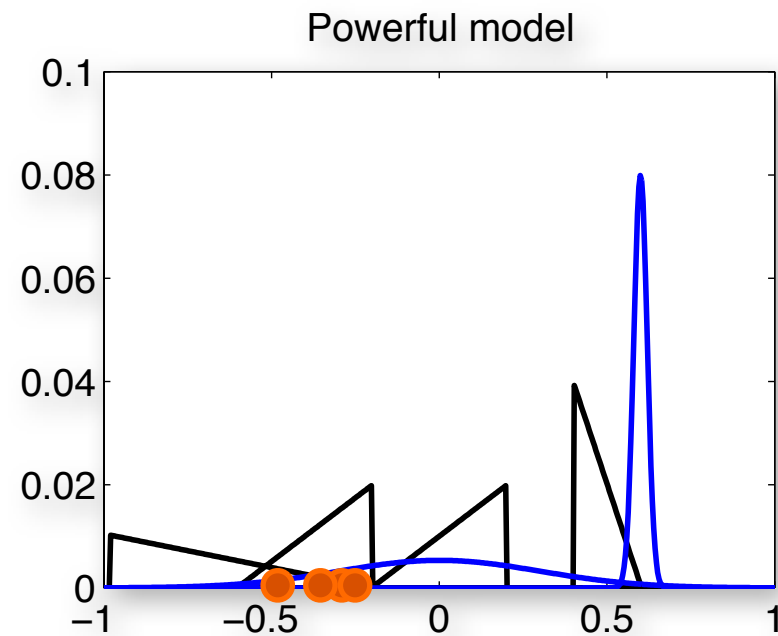
# Why integrals? The God Almighty test



# Why integrals? The God Almighty test



# Why integrals? The God Almighty test



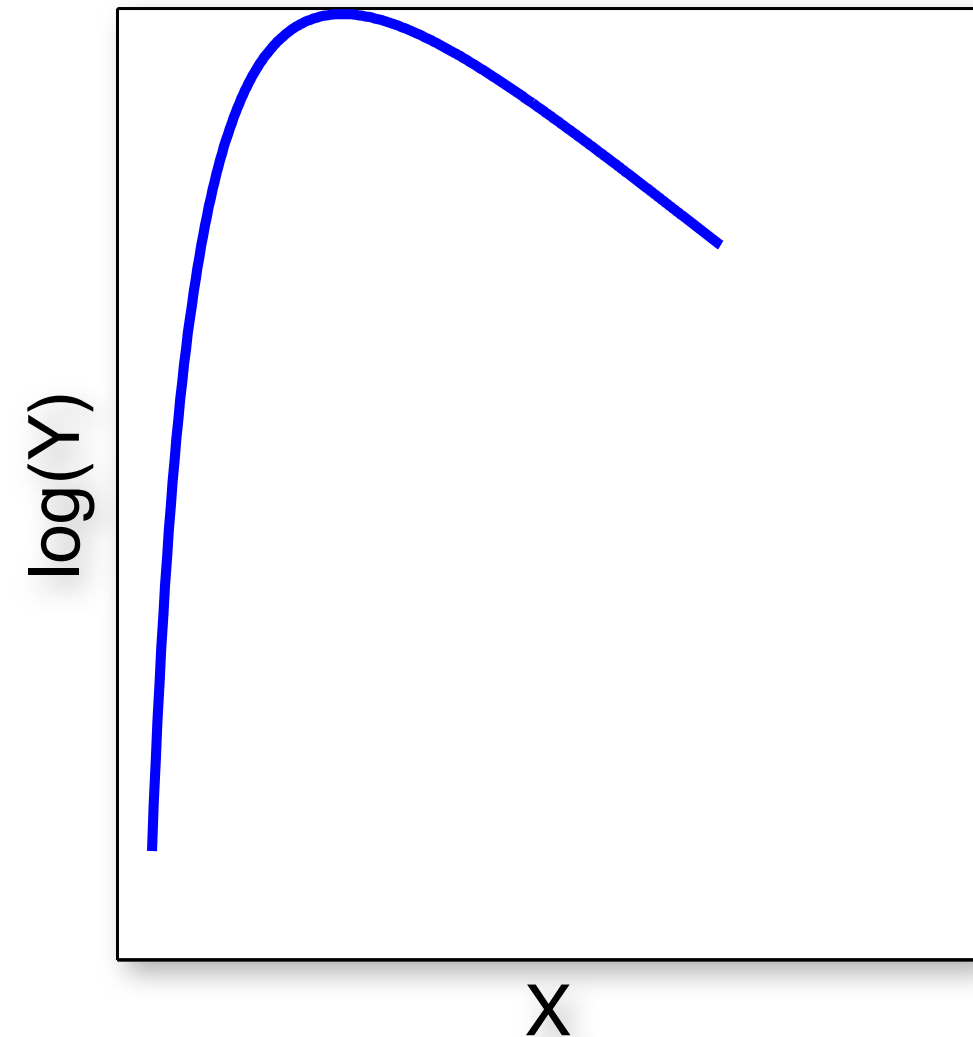
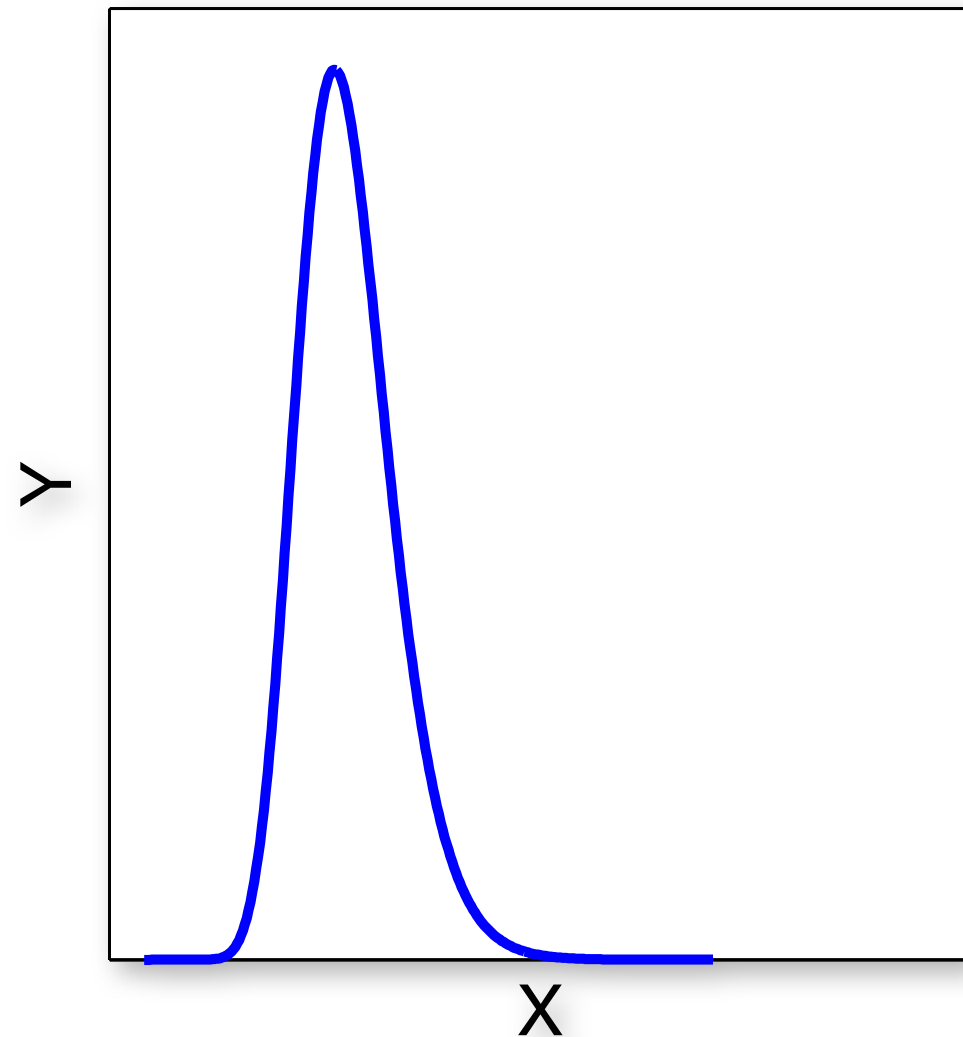
$$\frac{1}{N} (p(X|\theta_1) + p(X|\theta_2) + \dots)$$

These two factors fight it out  
Model complexity vs model fit

# Bayesian Information Criterion

---

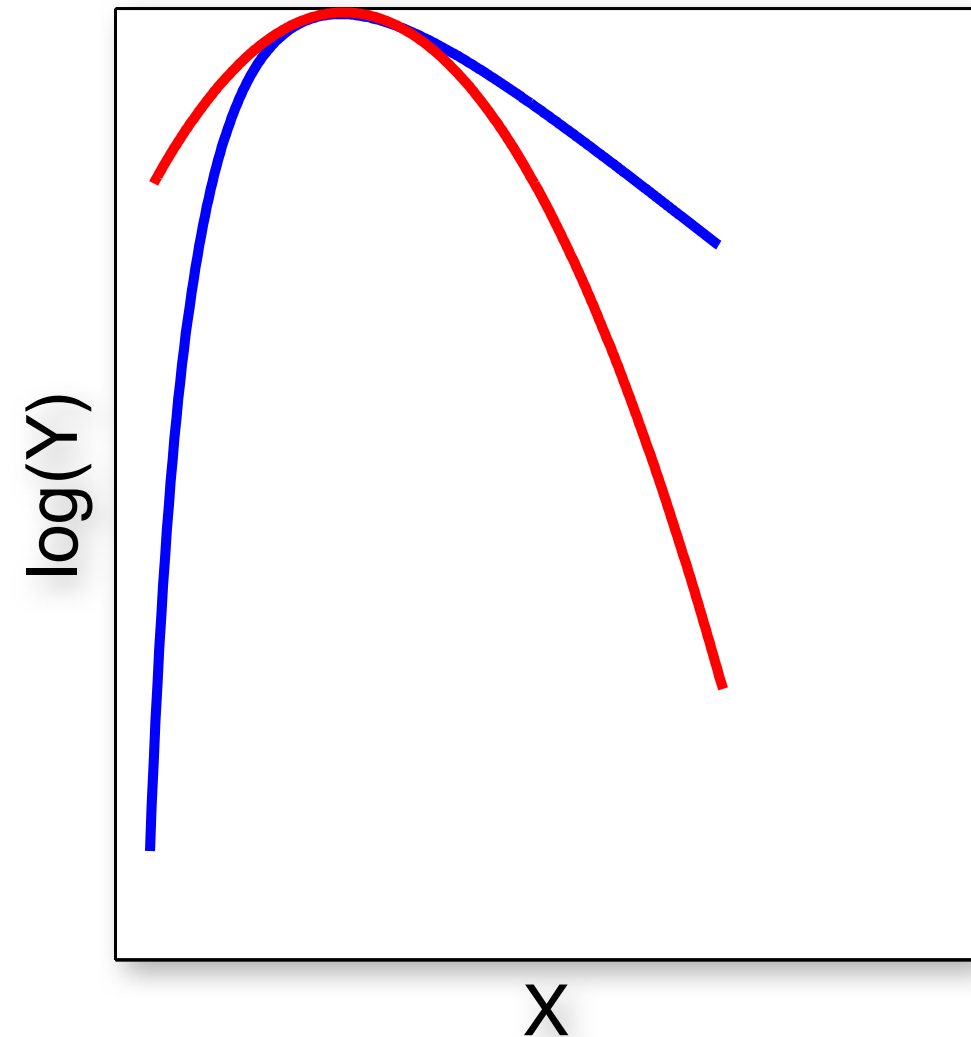
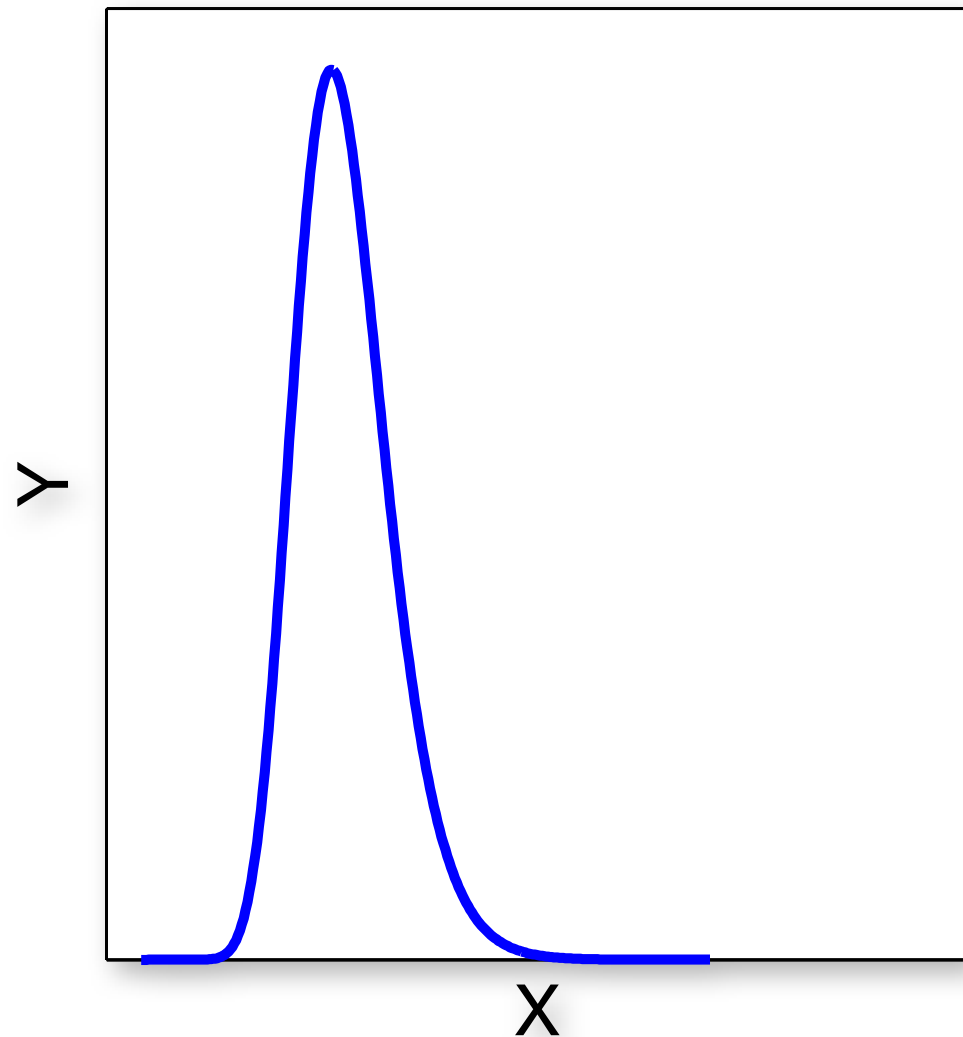
## ► Laplace's approximation (saddle-point method)





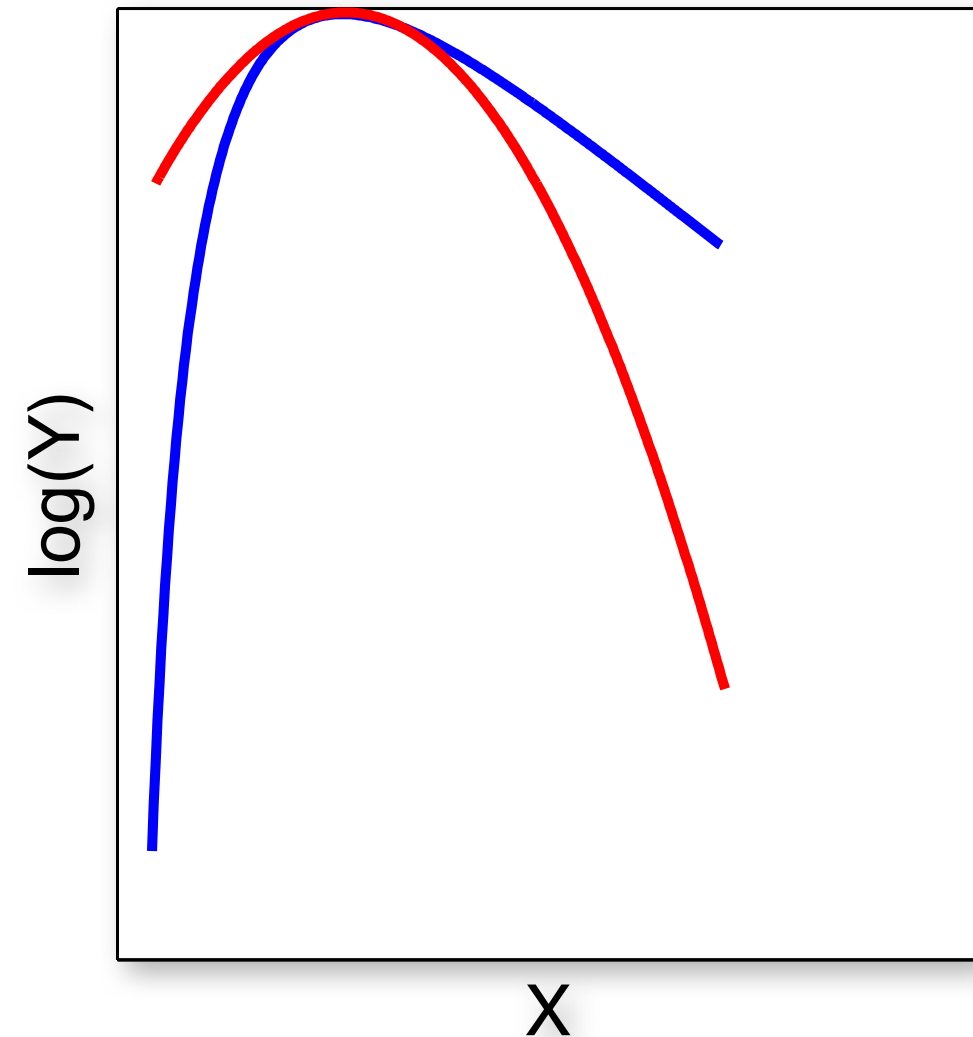
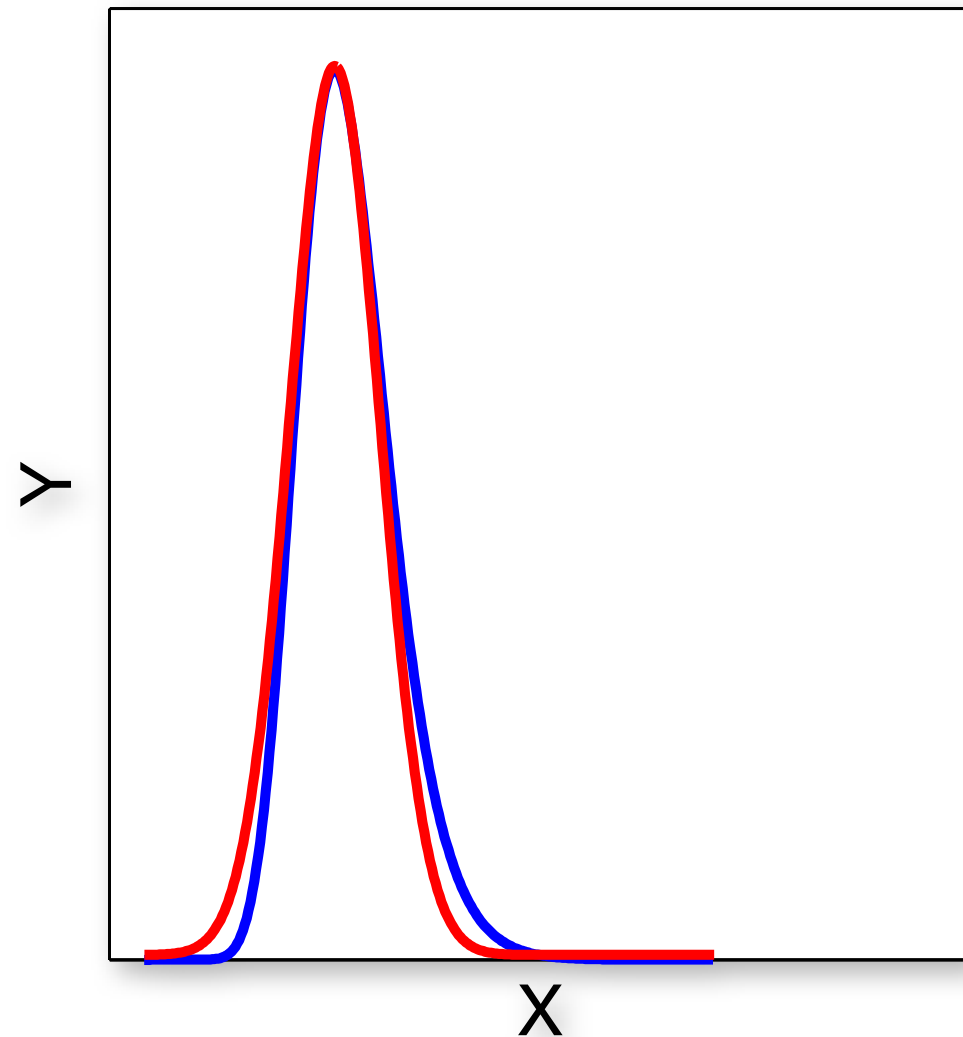
# Bayesian Information Criterion

## ► Laplace's approximation (saddle-point method)



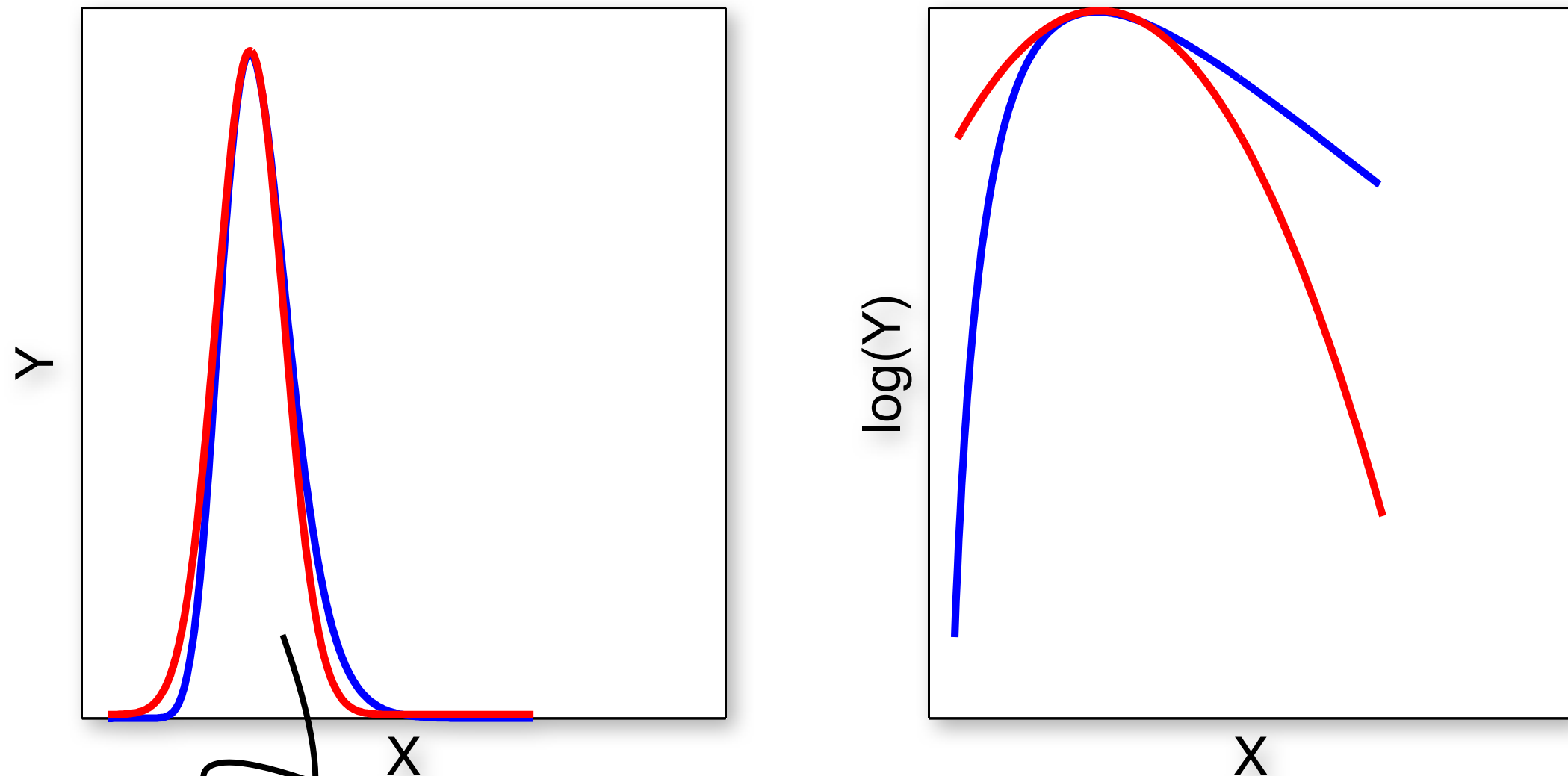
# Bayesian Information Criterion

## ► Laplace's approximation (saddle-point method)



# Bayesian Information Criterion

## ► Laplace's approximation (saddle-point method)



Just a Gaussian

$$\int dx f(x) \approx f^*(x_0) \sqrt{2\pi\sigma^2}$$

# Bayesian Information Criterion: one subject

---


$$p(\mathcal{A}|\mathcal{M}) = \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) p(\theta|\mathcal{M})$$

# Bayesian Information Criterion: one subject

---

$$p(\mathcal{A}|\mathcal{M}) = \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) p(\theta|\mathcal{M})$$

$p(\mathcal{A}|\theta) p(\theta|\mathcal{M})$   
is propto Gaussian




# Bayesian Information Criterion: one subject

---

$$\begin{aligned} p(\mathcal{A}|\mathcal{M}) &= \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) p(\theta|\mathcal{M}) \\ &\approx p(\mathcal{A}|\theta^{ML}, \mathcal{M}) p(\theta^{ML}|\mathcal{M}) \times \sqrt{(2\pi)^N |\Sigma|} \end{aligned}$$

$p(\mathcal{A}|\theta) p(\theta|\mathcal{M})$   
is propto Gaussian



# Bayesian Information Criterion: one subject

---

$$\begin{aligned} p(\mathcal{A}|\mathcal{M}) &= \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) p(\theta|\mathcal{M}) \\ &\approx p(\mathcal{A}|\theta^{ML}, \mathcal{M}) p(\theta^{ML}|\mathcal{M}) \times \sqrt{(2\pi)^N |\Sigma|} \end{aligned}$$

$p(\mathcal{A}|\theta) p(\theta|\mathcal{M})$   
is propto Gaussian

$p(\theta|\mathcal{M}) = \text{const.}$   
Model doesn't prefer particular

# Bayesian Information Criterion: one subject

---

$$\begin{aligned} p(\mathcal{A}|\mathcal{M}) &= \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) p(\theta|\mathcal{M}) \\ &\approx p(\mathcal{A}|\theta^{ML}, \mathcal{M}) p(\theta^{ML}|\mathcal{M}) \times \sqrt{(2\pi)^N |\Sigma|} \\ \log p(\mathcal{A}|\mathcal{M}) &\approx \log p(\mathcal{A}|\theta^{ML}, \mathcal{M}) + \frac{1}{2} \log(|\Sigma|) + \frac{N}{2} \log(2\pi) \end{aligned}$$

$p(\mathcal{A}|\theta) p(\theta|\mathcal{M})$   
is propto Gaussian

$p(\theta|\mathcal{M}) = \text{const.}$   
Model doesn't prefer particular



# Bayesian Information Criterion: one subject

$$\begin{aligned} p(\mathcal{A}|\mathcal{M}) &= \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) p(\theta|\mathcal{M}) \\ &\approx p(\mathcal{A}|\theta^{ML}, \mathcal{M}) p(\theta^{ML}|\mathcal{M}) \times \sqrt{(2\pi)^N |\Sigma|} \\ \log p(\mathcal{A}|\mathcal{M}) &\approx \log p(\mathcal{A}|\theta^{ML}, \mathcal{M}) + \frac{1}{2} \log(|\Sigma|) + \frac{N}{2} \log(2\pi) \end{aligned}$$

$p(\mathcal{A}|\theta) p(\theta|\mathcal{M})$   
is propto Gaussian

$p(\theta|\mathcal{M}) = \text{const.}$   
Model doesn't prefer particular

# Bayesian Information Criterion: one subject

$$\begin{aligned} p(\mathcal{A}|\mathcal{M}) &= \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) p(\theta|\mathcal{M}) \\ &\approx p(\mathcal{A}|\theta^{ML}, \mathcal{M}) p(\theta^{ML}|\mathcal{M}) \times \sqrt{(2\pi)^N |\Sigma|} \\ \log p(\mathcal{A}|\mathcal{M}) &\approx \log p(\mathcal{A}|\theta^{ML}, \mathcal{M}) + \frac{1}{2} \log(|\Sigma|) + \frac{N}{2} \log(2\pi) \end{aligned}$$

$p(\mathcal{A}|\theta) p(\theta|\mathcal{M})$   
is propto Gaussian

$p(\theta|\mathcal{M}) = \text{const.}$   
Model doesn't prefer particular

$$\begin{aligned} \Sigma_{ii} \propto \frac{1}{T} &\Rightarrow \frac{1}{2} \log(|\Sigma|) \approx -\frac{N}{2} \log(T) && \text{Bayesian Information Criterion (BIC)} \\ &\approx -N && \text{Akaike Information Criterion (AIC)} \end{aligned}$$

# Bayesian Information Criterion: one subject

$$\begin{aligned} p(\mathcal{A}|\mathcal{M}) &= \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) p(\theta|\mathcal{M}) \\ &\approx p(\mathcal{A}|\theta^{ML}, \mathcal{M}) p(\theta^{ML}|\mathcal{M}) \times \sqrt{(2\pi)^N |\Sigma|} \\ \log p(\mathcal{A}|\mathcal{M}) &\approx \log p(\mathcal{A}|\theta^{ML}, \mathcal{M}) + \frac{1}{2} \log(|\Sigma|) + \frac{N}{2} \log(2\pi) \end{aligned}$$

$p(\mathcal{A}|\theta) p(\theta|\mathcal{M})$   
is propto Gaussian

$p(\theta|\mathcal{M}) = \text{const.}$   
Model doesn't prefer particular

$$\begin{aligned} \Sigma_{ii} \propto \frac{1}{T} \Rightarrow \frac{1}{2} \log(|\Sigma|) &\approx -\frac{N}{2} \log(T) && \text{Bayesian Information Criterion (BIC)} \\ &\approx -N && \text{Akaike Information Criterion (AIC)} \end{aligned}$$

Model fit vs Model complexity

# Group data

---

## ▶ Multiple subjects

## ▶ Multiple models

- do they use the same model? If not parameters are not comparable
- which model best accounts for all of them?

## ▶ Multiple groups

- difference in models?
- difference in parameters?
- $2^k$  possible model comparisons

## ▶ Multiple parameters

- $2^k$  possible correlations with any one psychometric measure

# Group data - approaches

---

## ► Summary statistic

- Treat individual model comparison measure as summary statistics, do ANOVA or t-test

## ► Fixed effect analysis

- Subject data independent

$$\begin{aligned}\log p(\mathcal{A}|\mathcal{M}) &= \sum_i \log p(\mathcal{A}_i|\mathcal{M}) \\ &= \sum_i \log \int d\theta_i p(\mathcal{A}_i|\theta_i) p(\theta_i|\mathcal{M}_i) \approx -\frac{1}{2} \sum_i \text{BIC}_i\end{aligned}$$

## ► Random effects analyses

- Hierarchical prior on group parameters

$$p(\mathcal{A}|\mathcal{M}) = \int d\zeta \int d\theta p(\mathcal{A}|\theta) p(\theta|\zeta) p(\zeta|\mathcal{M})$$

- Hierarchical prior on models

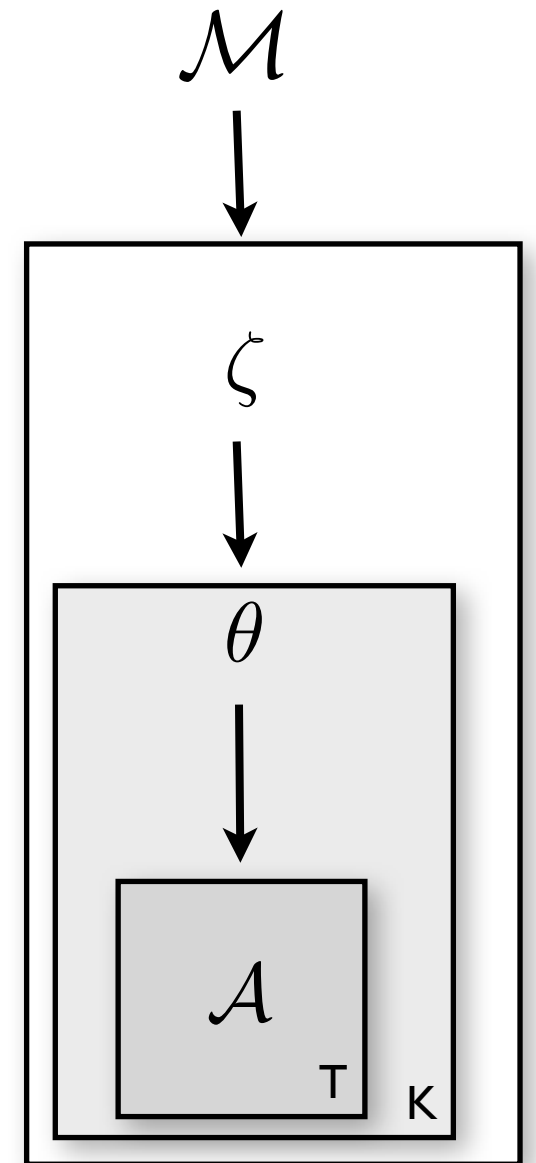
$$p(\mathcal{A}, \mathcal{M}_k, r|\alpha) = p(\mathcal{A}|\mathcal{M}_k) p(\mathcal{M}_k|r) p(r|\alpha)$$

# Group-level likelihood

► Contains two integrals:

- subject parameters
- prior parameters

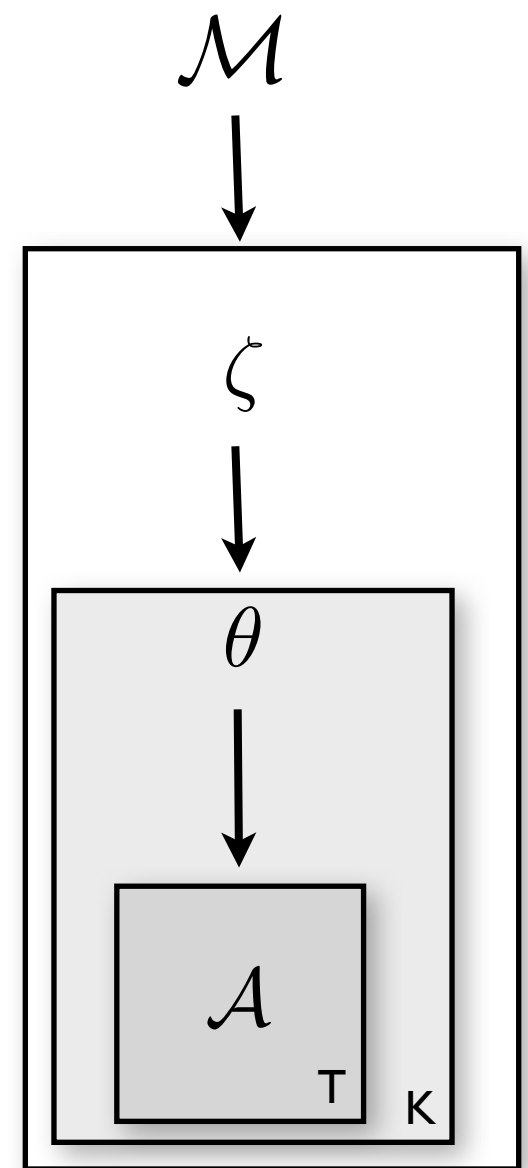
$$p(\mathcal{A}|\mathcal{M}) = \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) \int d\zeta p(\theta|\zeta) p(\zeta|\mathcal{M})$$



# Evaluating $p(\mathcal{A}|\mathcal{M})$

---

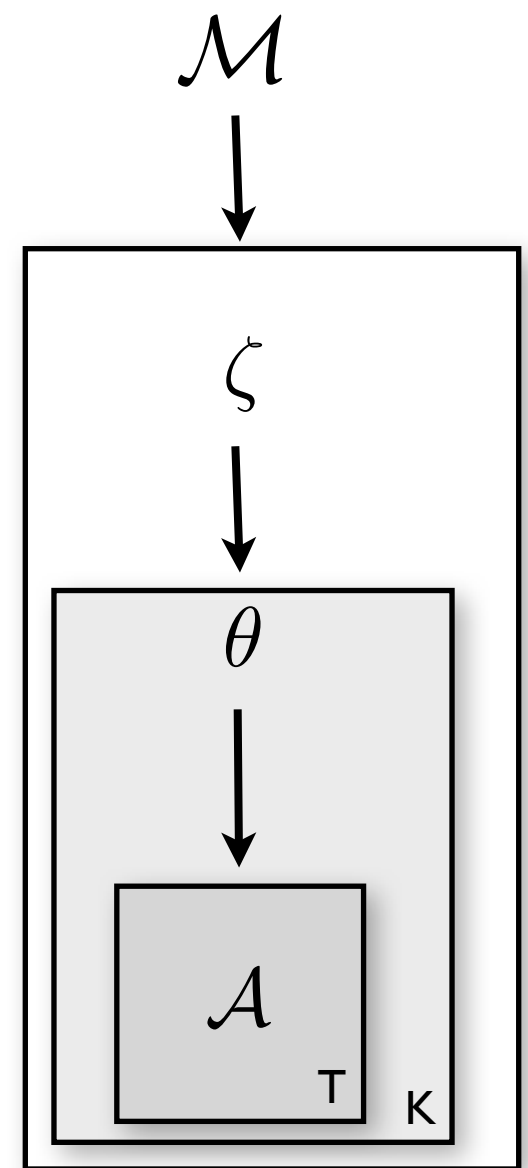
$$p(\mathcal{A}|\mathcal{M}) = \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) \int d\zeta p(\theta|\zeta) p(\zeta|\mathcal{M})$$



# Evaluating $p(A|M)$

- Two integrals
  - tricky

$$p(\mathcal{A}|\mathcal{M}) = \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) \int d\zeta p(\theta|\zeta) p(\zeta|\mathcal{M})$$





# Evaluating $p(\mathcal{A}|\mathcal{M})$

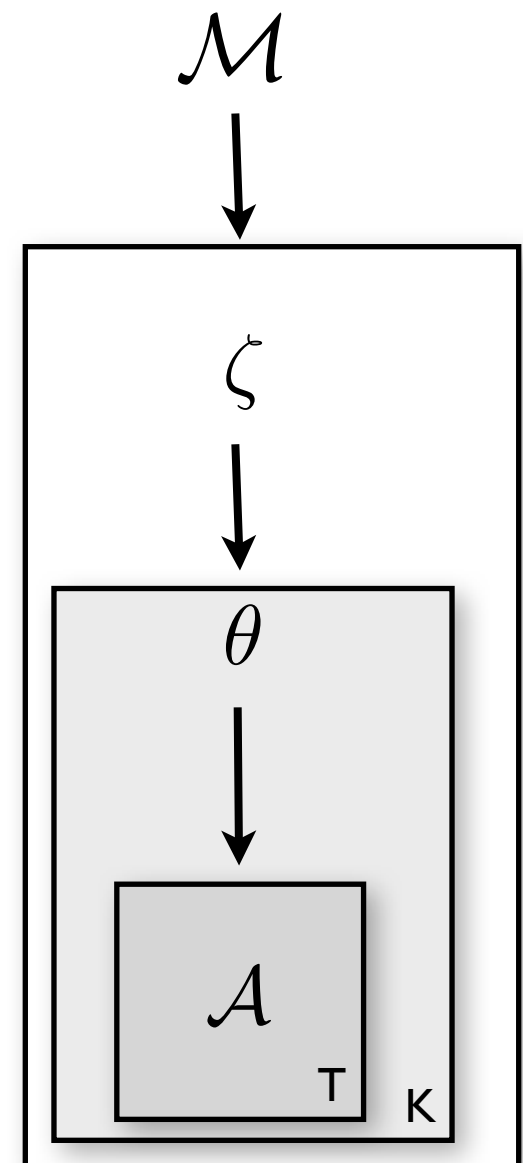
► Two integrals

- tricky

$$p(\mathcal{A}|\mathcal{M}) = \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) \int d\zeta p(\theta|\zeta) p(\zeta|\mathcal{M})$$

► Step by step: approximating levels separately

- Top level first:



# Evaluating $p(\mathcal{A}|\mathcal{M})$

## ► Two integrals

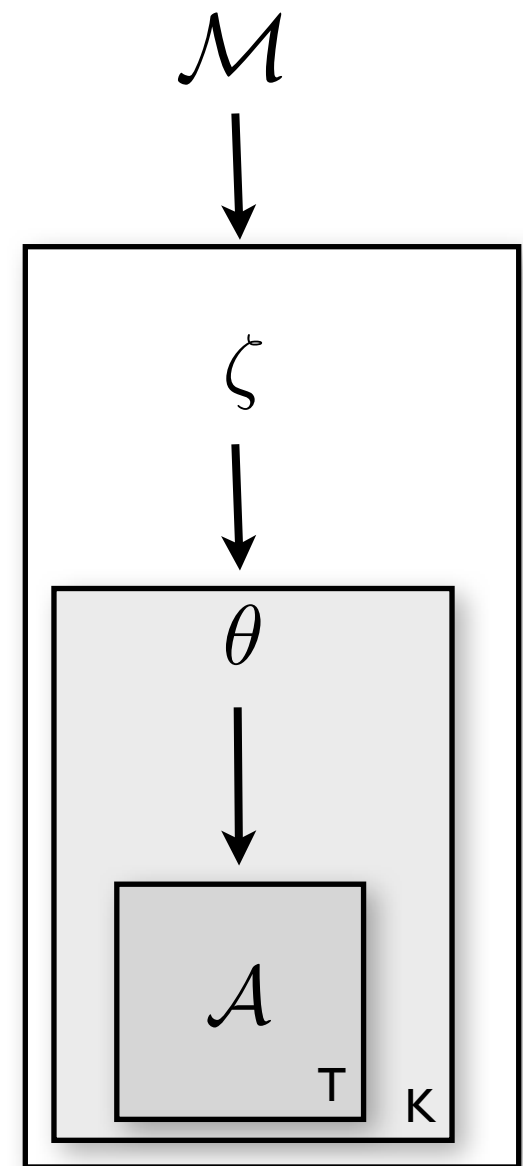
- tricky

$$p(\mathcal{A}|\mathcal{M}) = \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) \int d\zeta p(\theta|\zeta) p(\zeta|\mathcal{M})$$

## ► Step by step: approximating levels separately

- Top level first:

$$p(\mathcal{A}|\mathcal{M}) = \int d\zeta p(\mathcal{A}|\zeta, \mathcal{M}) p(\zeta|\mathcal{M})$$



# Evaluating $p(\mathcal{A}|\mathcal{M})$

---

- ▶ **Two integrals**
  - tricky
$$p(\mathcal{A}|\mathcal{M}) = \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) \int d\zeta p(\theta|\zeta) p(\zeta|\mathcal{M})$$
- ▶ **Step by step: approximating levels separately**
  - Approximate at the top level
  - less action

$$p(\mathcal{A}|\mathcal{M}) = \int d\zeta p(\mathcal{A}|\zeta, \mathcal{M}) p(\zeta|\mathcal{M})$$

# Evaluating $p(\mathcal{A}|\mathcal{M})$

---

► **Two integrals**


- tricky

$$p(\mathcal{A}|\mathcal{M}) = \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) \int d\zeta p(\theta|\zeta) p(\zeta|\mathcal{M})$$

► **Step by step: approximating levels separately**

- Approximate at the top level
- less action

$$p(\mathcal{A}|\zeta, \mathcal{M}) p(\zeta|\mathcal{M})$$

$$p(\mathcal{A}|\mathcal{M}) = \int d\zeta p(\mathcal{A}|\zeta, \mathcal{M}) p(\zeta|\mathcal{M})$$


# Evaluating $p(\mathcal{A}|\mathcal{M})$

## ► Two integrals

- tricky

$$p(\mathcal{A}|\mathcal{M}) = \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) \int d\zeta p(\theta|\zeta) p(\zeta|\mathcal{M})$$

## ► Step by step: approximating levels separately

- Approximate at the top level
- less action

$$\begin{aligned} p(\mathcal{A}|\mathcal{M}) &= \int d\zeta p(\mathcal{A}|\zeta, \mathcal{M}) p(\zeta|\mathcal{M}) \\ &\approx p(\mathcal{A}|\zeta^{ML}, \mathcal{M}) p(\zeta^{ML}|\mathcal{M}) \times \sqrt{(2\pi)^N |\Sigma|} \end{aligned}$$

$p(\mathcal{A}|\zeta, \mathcal{M}) p(\zeta|\mathcal{M})$   
is propto Gaussian

# Evaluating $p(\mathcal{A}|\mathcal{M})$

## ► Two integrals

- tricky

$$p(\mathcal{A}|\mathcal{M}) = \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) \int d\zeta p(\theta|\zeta) p(\zeta|\mathcal{M})$$

## ► Step by step: approximating levels separately

- Approximate at the top level
- less action

$$\begin{aligned} p(\mathcal{A}|\mathcal{M}) &= \int d\zeta p(\mathcal{A}|\zeta, \mathcal{M}) p(\zeta|\mathcal{M}) \\ &\approx p(\mathcal{A}|\zeta^{ML}, \mathcal{M}) p(\zeta^{ML}|\mathcal{M}) \times \sqrt{(2\pi)^N |\Sigma|} \end{aligned}$$

$p(\mathcal{A}|\zeta, \mathcal{M}) p(\zeta|\mathcal{M})$   
is propto Gaussian

Model doesn't prefer particular  $\zeta$

# Evaluating $p(\mathcal{A}|\mathcal{M})$

## ► Two integrals

- tricky

$$p(\mathcal{A}|\mathcal{M}) = \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) \int d\zeta p(\theta|\zeta) p(\zeta|\mathcal{M})$$

## ► Step by step: approximating levels separately

- Approximate at the top level
- less action

$$\begin{aligned} p(\mathcal{A}|\mathcal{M}) &= \int d\zeta p(\mathcal{A}|\zeta, \mathcal{M}) p(\zeta|\mathcal{M}) \\ &\approx p(\mathcal{A}|\zeta^{ML}, \mathcal{M}) p(\zeta^{ML}|\mathcal{M}) \times \sqrt{(2\pi)^N |\Sigma|} \\ \log p(\mathcal{A}|\mathcal{M}) &\approx \log p(\mathcal{A}|\zeta^{ML}, \mathcal{M}) + \frac{1}{2} \log(|\Sigma|) + \frac{N}{2} \log(2\pi) \end{aligned}$$

$p(\mathcal{A}|\zeta, \mathcal{M}) p(\zeta|\mathcal{M})$   
is propto Gaussian

Model doesn't prefer particular  $\zeta$

# Evaluating $p(\mathcal{A}|\mathcal{M})$

## ► Two integrals

- tricky

$$p(\mathcal{A}|\mathcal{M}) = \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) \int d\zeta p(\theta|\zeta) p(\zeta|\mathcal{M})$$

## ► Step by step: approximating levels separately

- Approximate at the top level
- less action

$p(\mathcal{A}|\zeta, \mathcal{M}) p(\zeta|\mathcal{M})$   
is propto Gaussian

$$p(\mathcal{A}|\mathcal{M}) = \int d\zeta p(\mathcal{A}|\zeta, \mathcal{M}) p(\zeta|\mathcal{M})$$
$$\approx p(\mathcal{A}|\zeta^{ML}, \mathcal{M}) p(\zeta^{ML}|\mathcal{M}) \times \sqrt{(2\pi)^N |\Sigma|}$$

Model doesn't prefer  
particular  $\zeta$

$$\log p(\mathcal{A}|\mathcal{M}) \approx \log p(\mathcal{A}|\zeta^{ML}, \mathcal{M}) + \frac{1}{2} \log(|\Sigma|) + \frac{N}{2} \log(2\pi)$$



# Evaluating $p(\mathcal{A}|\mathcal{M})$

## ► Two integrals

- tricky

$$p(\mathcal{A}|\mathcal{M}) = \int d\theta p(\mathcal{A}|\theta, \mathcal{M}) \int d\zeta p(\theta|\zeta) p(\zeta|\mathcal{M})$$

## ► Step by step: approximating levels separately

- Approximate at the top level
- less action

$$\begin{aligned} p(\mathcal{A}|\mathcal{M}) &= \int d\zeta p(\mathcal{A}|\zeta, \mathcal{M}) p(\zeta|\mathcal{M}) \\ &\approx p(\mathcal{A}|\zeta^{ML}, \mathcal{M}) p(\zeta^{ML}|\mathcal{M}) \times \sqrt{(2\pi)^N |\Sigma|} \\ \log p(\mathcal{A}|\mathcal{M}) &\approx \log p(\mathcal{A}|\zeta^{ML}, \mathcal{M}) + \frac{1}{2} \log(|\Sigma|) + \frac{N}{2} \log(2\pi) \end{aligned}$$

*$p(\mathcal{A}|\zeta, \mathcal{M}) p(\zeta|\mathcal{M})$  is propto Gaussian*

*Model doesn't prefer particular  $\zeta$*

just as before, top-level BIC

# Approximating level I

---

## ► Still leaves the first level:

- Approximate integral by sampling, e.g. importance sampling for few dimensions ( $< 10$ )

$$\begin{aligned}\log p(\mathcal{A}|\zeta^{ML}, \mathcal{M}) &= \log \int d\theta p(\mathcal{A}|\theta) p(\theta|\zeta^{ML}) \\ &\approx \log \frac{1}{B} \sum_{b=1}^B p(\mathcal{A}|\theta^b) \\ \theta^b &\sim p(\theta|\zeta^{ML})\end{aligned}$$

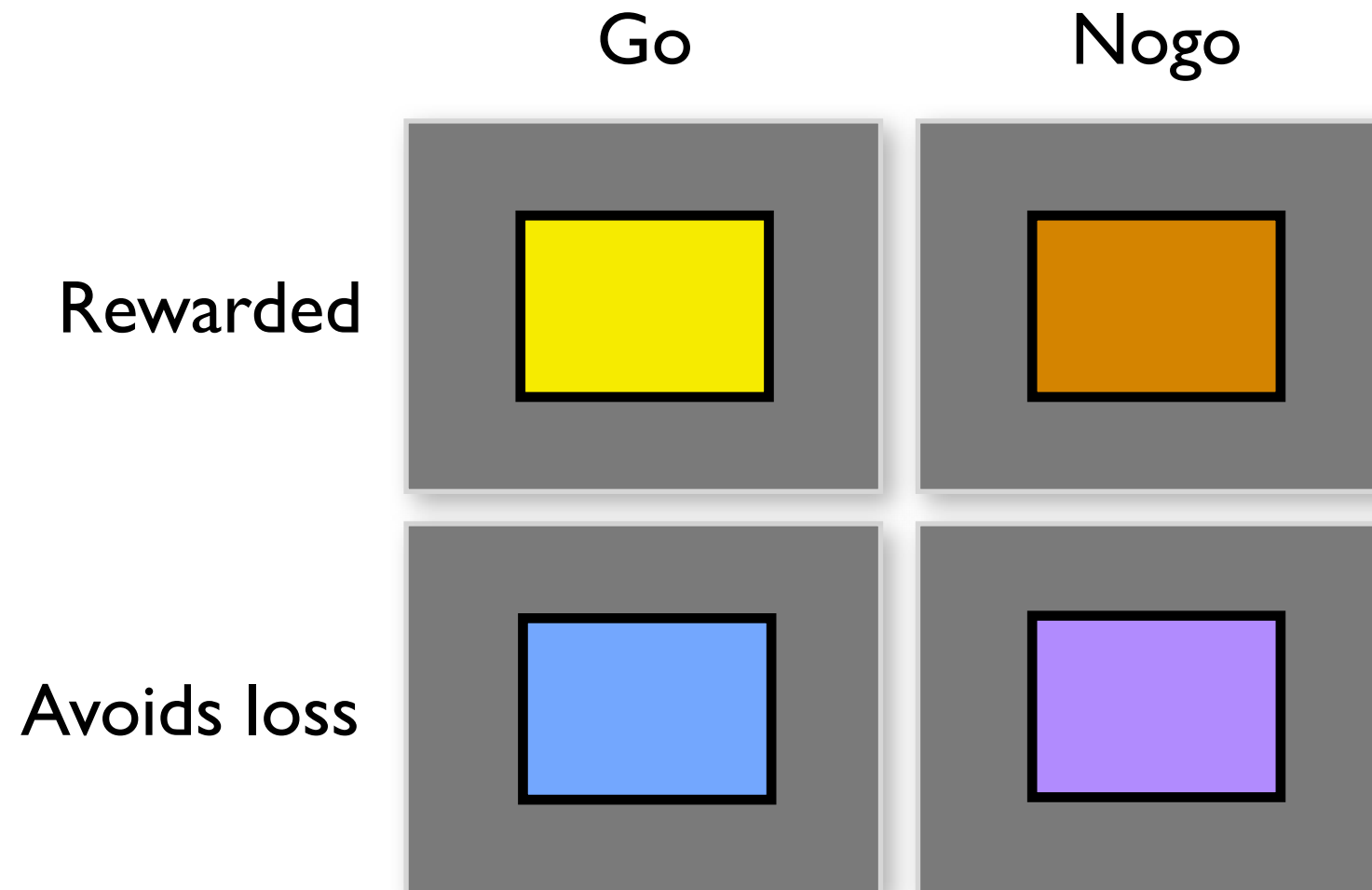
# Group-level BIC

---

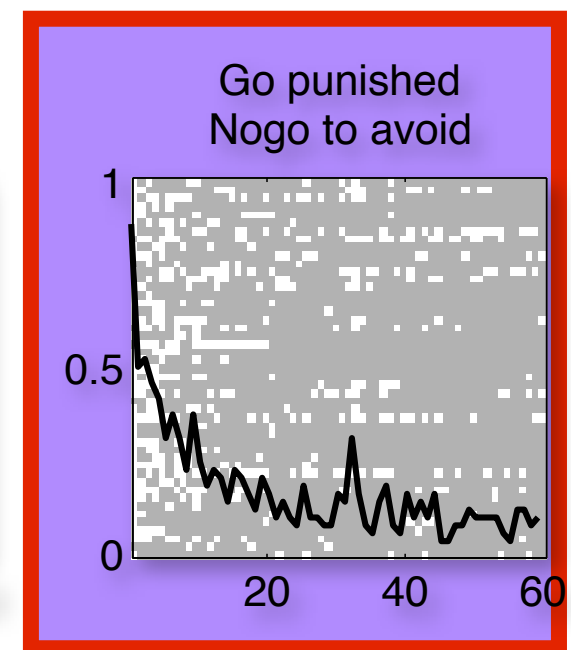
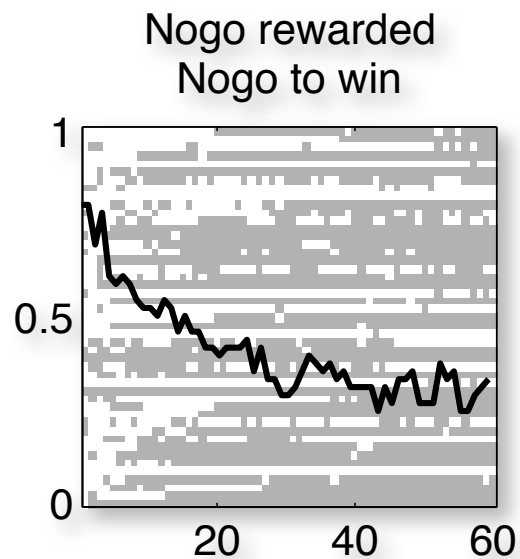
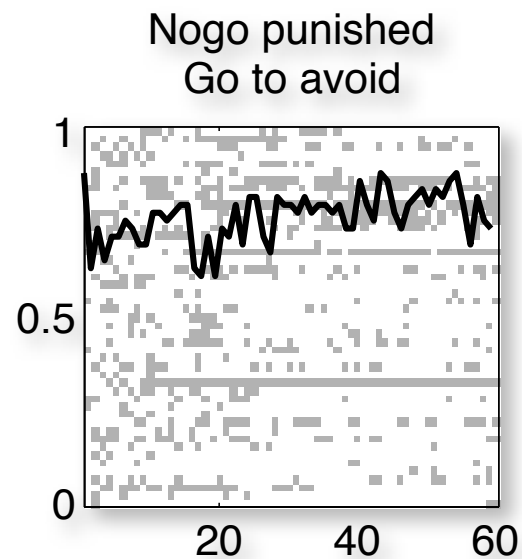
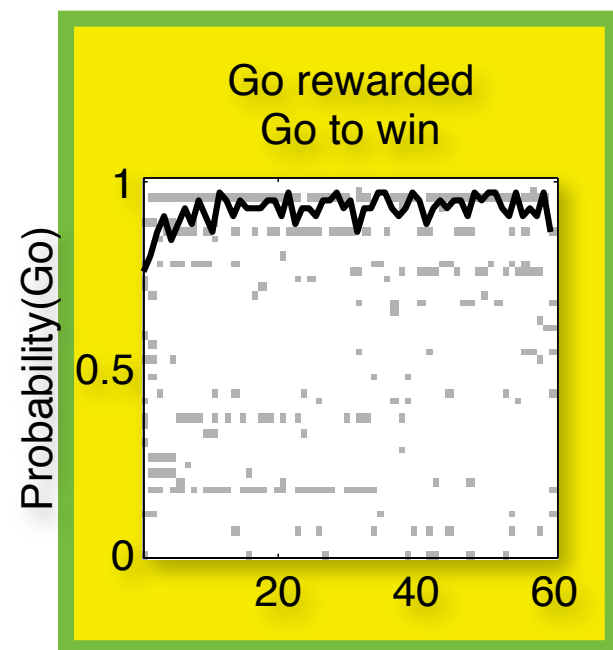
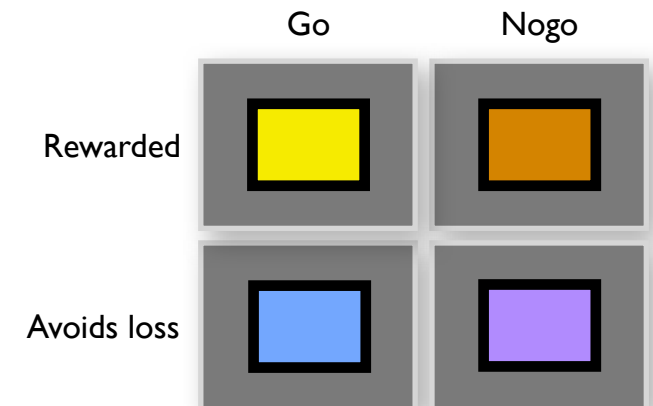
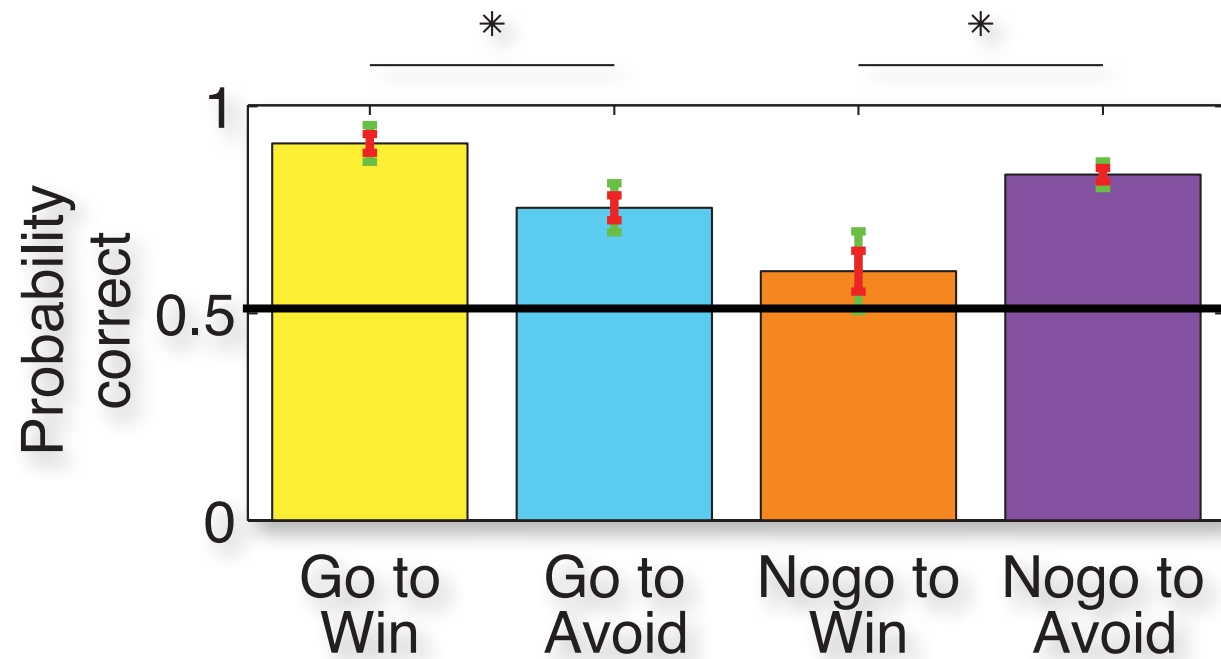
$$\begin{aligned}\log p(\mathcal{A}|\mathcal{M}) &= \int d\zeta \, p(\mathcal{A}|\zeta) \, p(\zeta|\mathcal{M}) \\ &\approx -\frac{1}{2} \text{BIC}_{\text{int}} \\ &= \log \hat{p}(\mathcal{A}|\hat{\zeta}^{ML}) - \frac{1}{2} |\mathcal{M}| \log(|\mathcal{A}|)\end{aligned}$$

# Example task

---

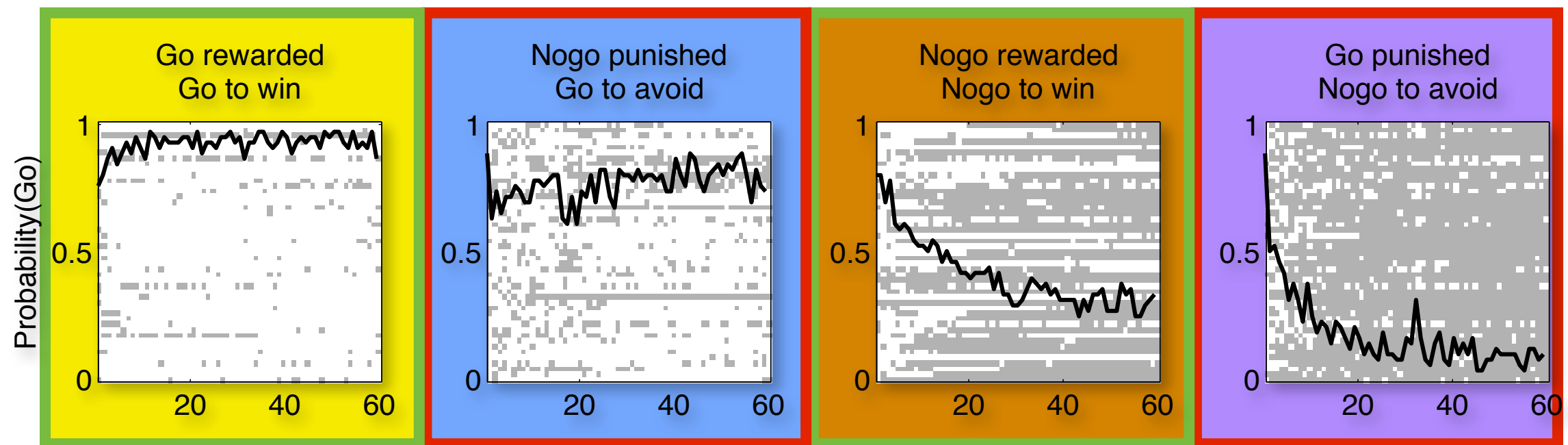
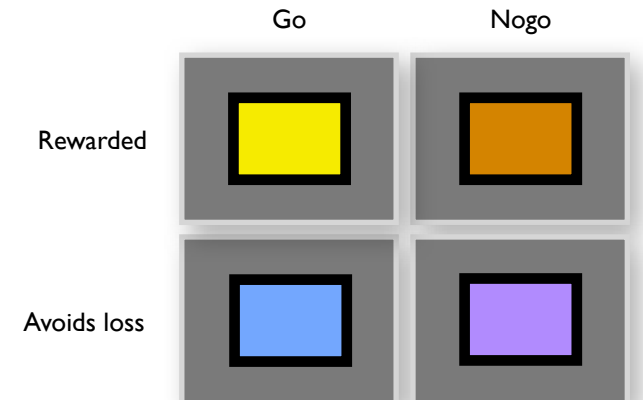


# Example task



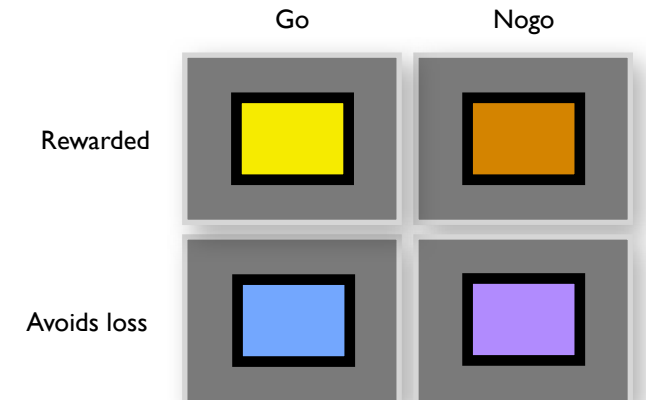
Guitart-Masip, Huys et al. *Submitted*

# Model validation: generating data

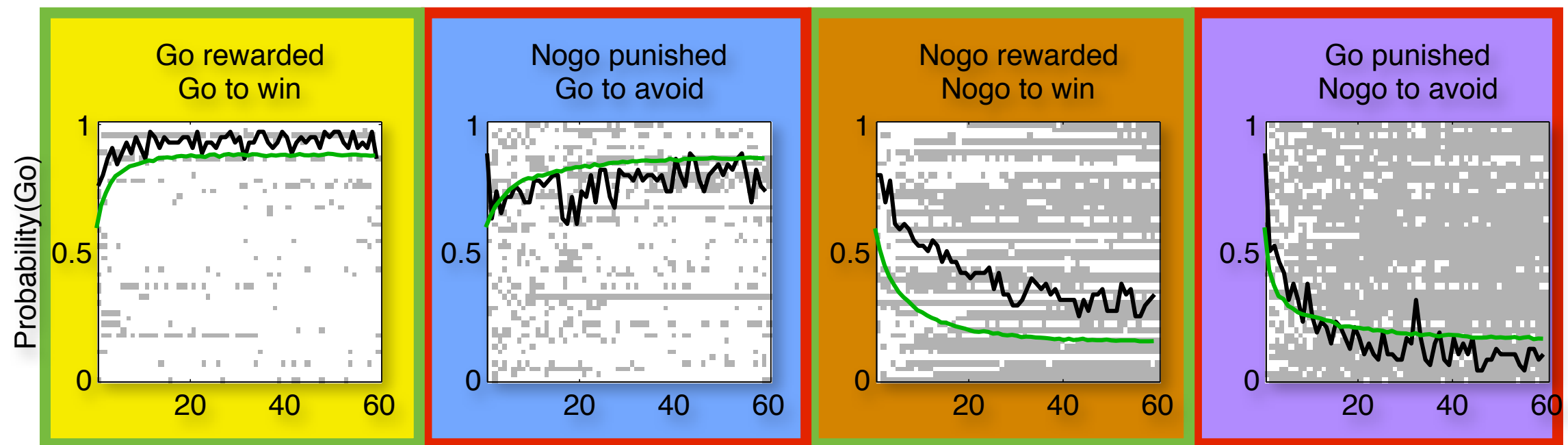


Guitart-Masip et al. 2011, Guitart-Masip, Huys et al. *Submitted*

# Model validation: generating data

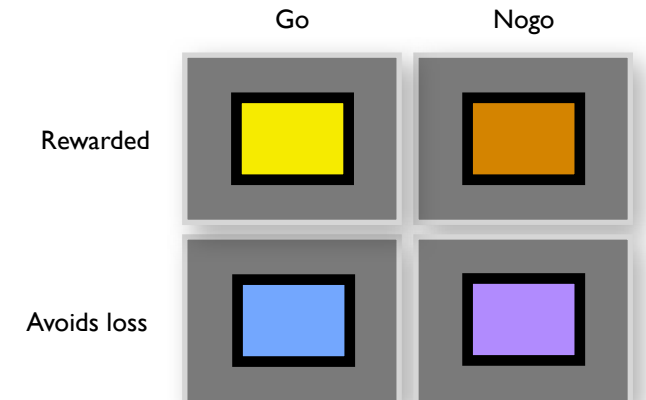


$$p(\text{go}|s_t) \propto Q_t(\text{go}|s_t) + \text{bias}(\text{go})$$

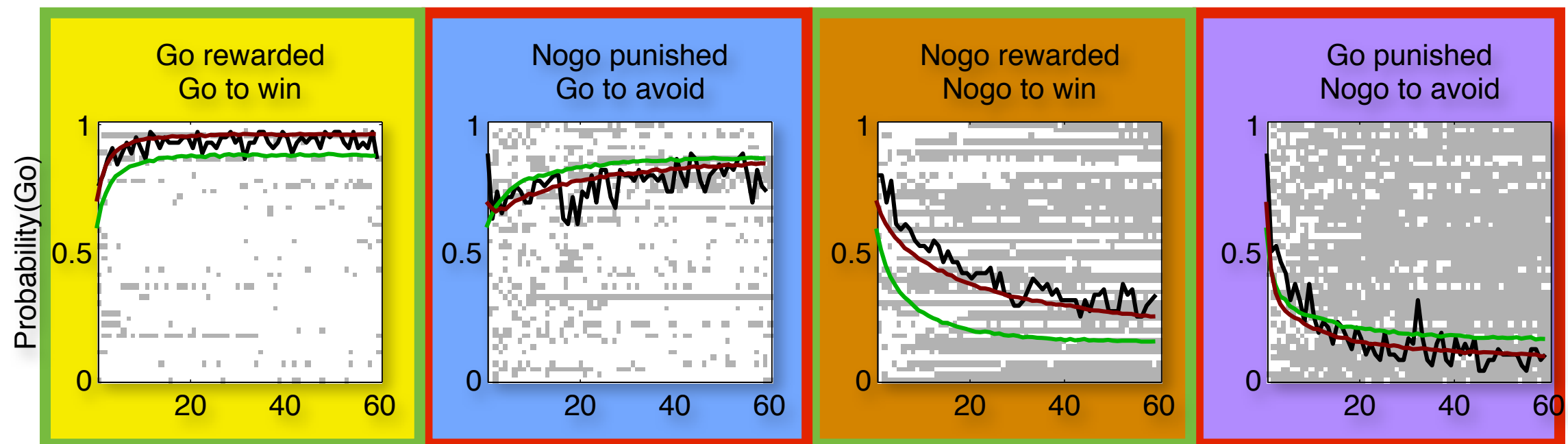


Guitart-Masip et al. 2011, Guitart-Masip, Huys et al. *Submitted*

# Model validation: generating data



$$p(\text{go}|s_t) \propto Q_t(\text{go}|s_t) + \text{bias}(\text{go}) + \mathcal{V}_t(s_t)$$
$$\mathcal{V}_t(s_t) = \mathcal{V}_{t-1}(s_t) + \epsilon(r_t - \mathcal{V}_{t-1}(s_t))$$

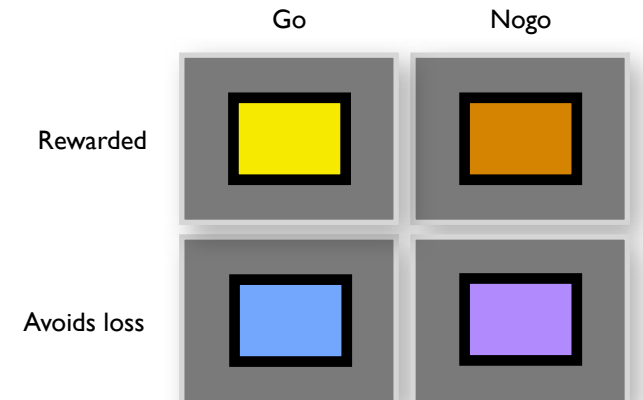


Guitart-Masip et al. 2011, Guitart-Masip, Huys et al. *Submitted*

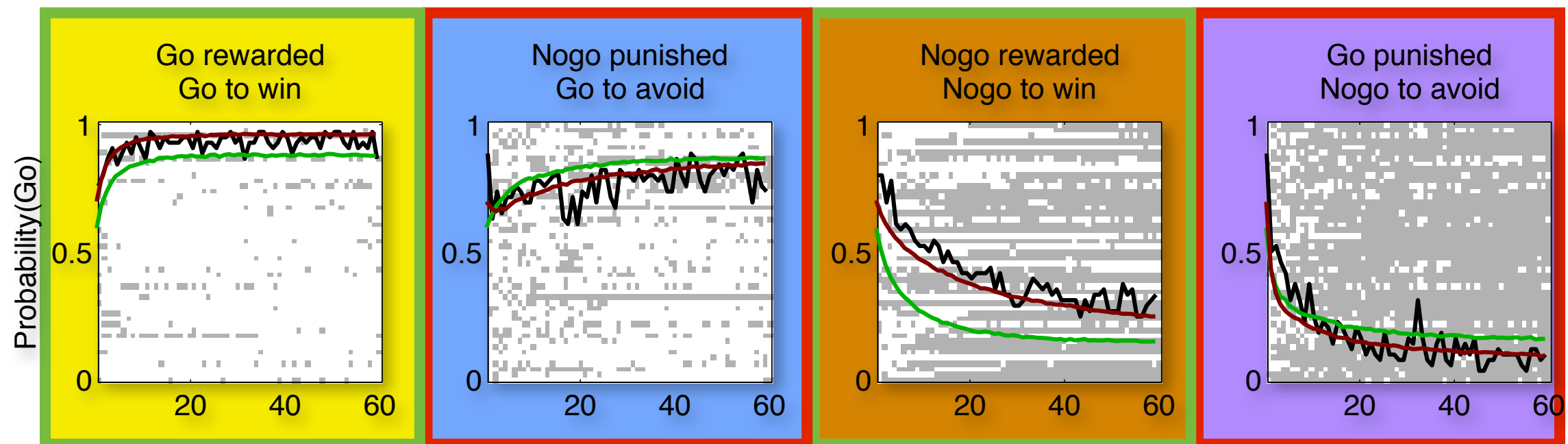


# Model validation: generating data

$$P(\text{go}) \propto \text{value of stimulus}$$

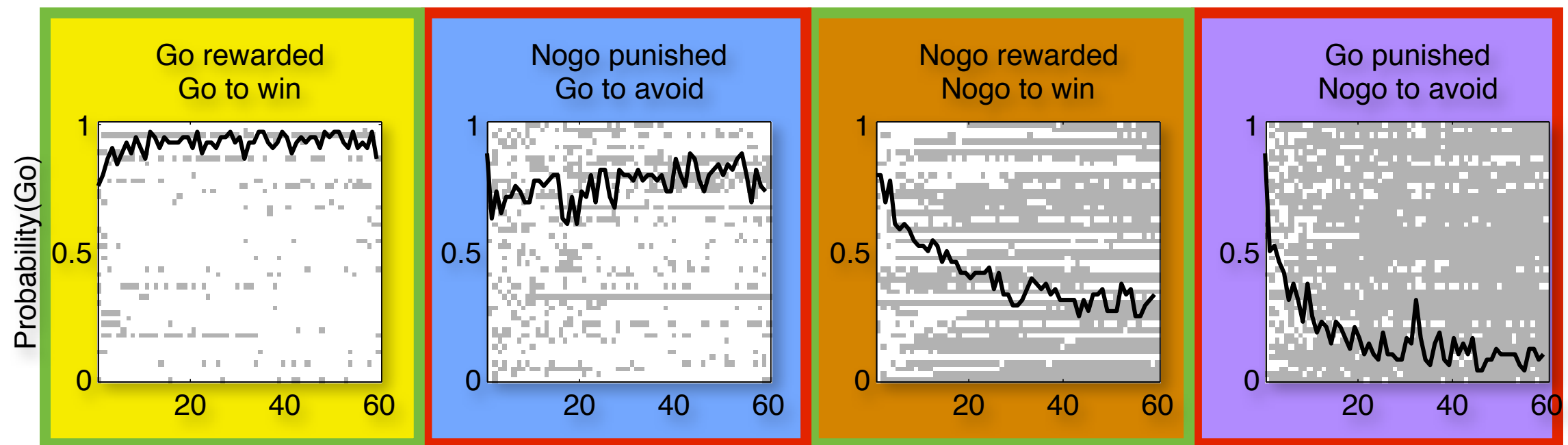
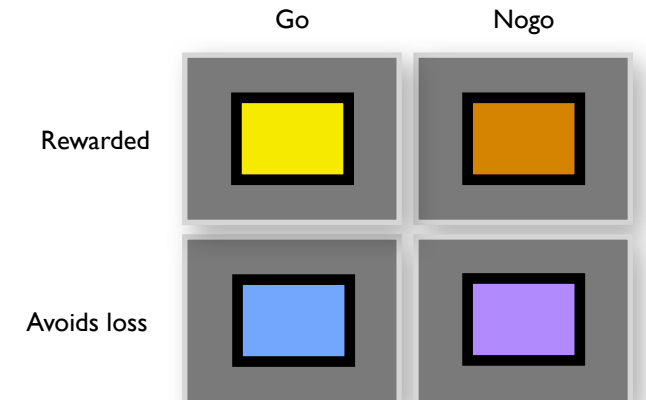
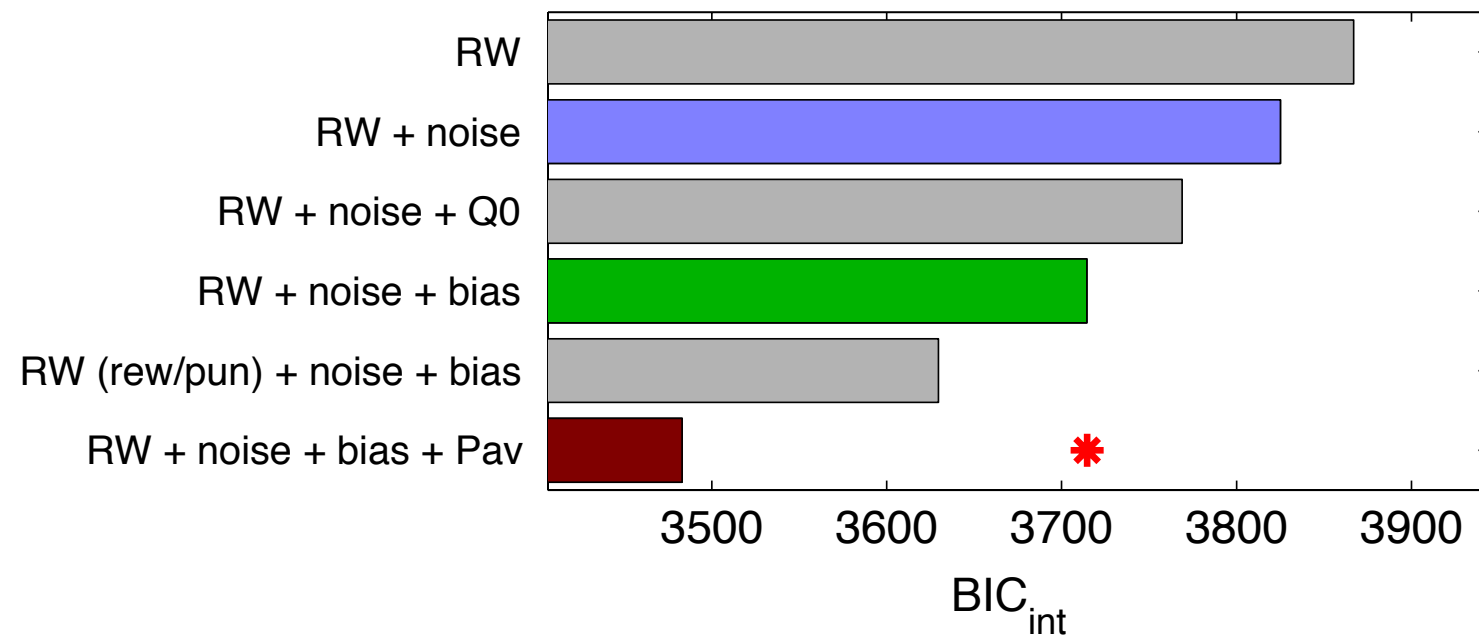


$$p(\text{go}|s_t) \propto Q_t(\text{go}|s_t) + \text{bias}(\text{go}) + \mathcal{V}_t(s_t)$$
$$\mathcal{V}_t(s_t) = \mathcal{V}_{t-1}(s_t) + \epsilon(r_t - \mathcal{V}_{t-1}(s_t))$$

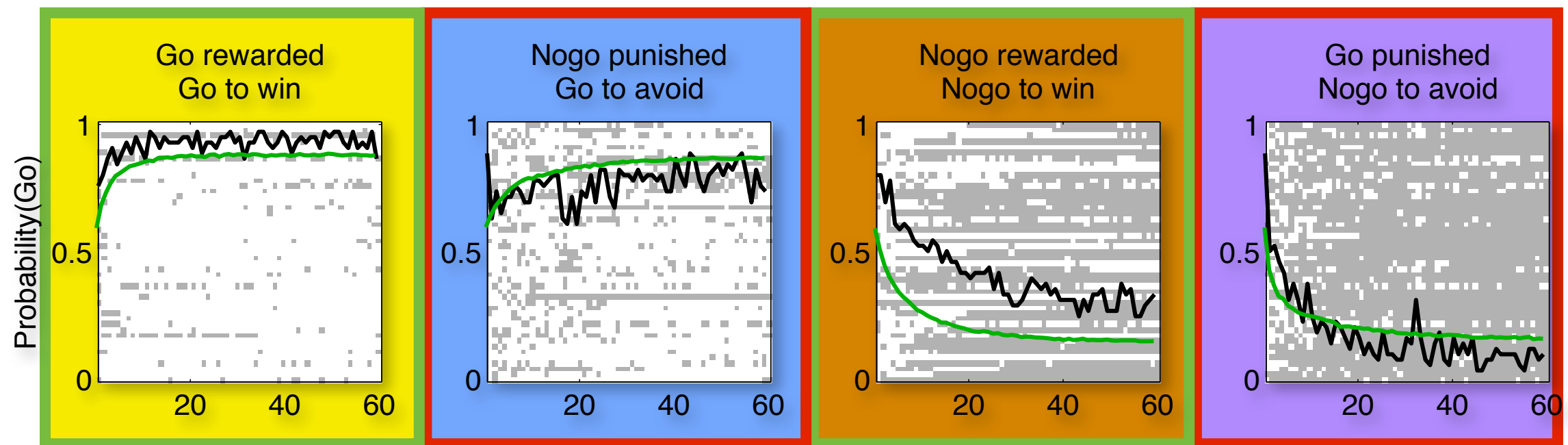
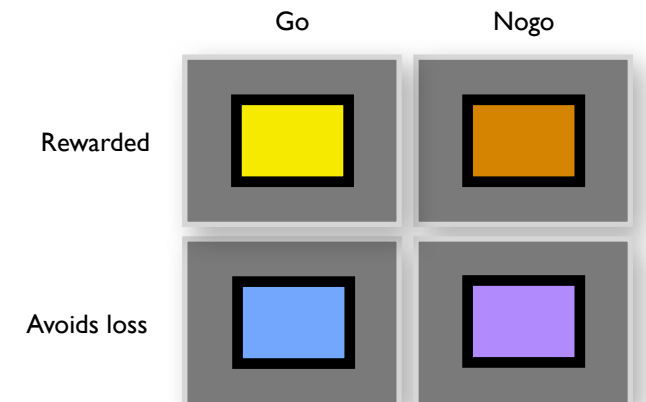
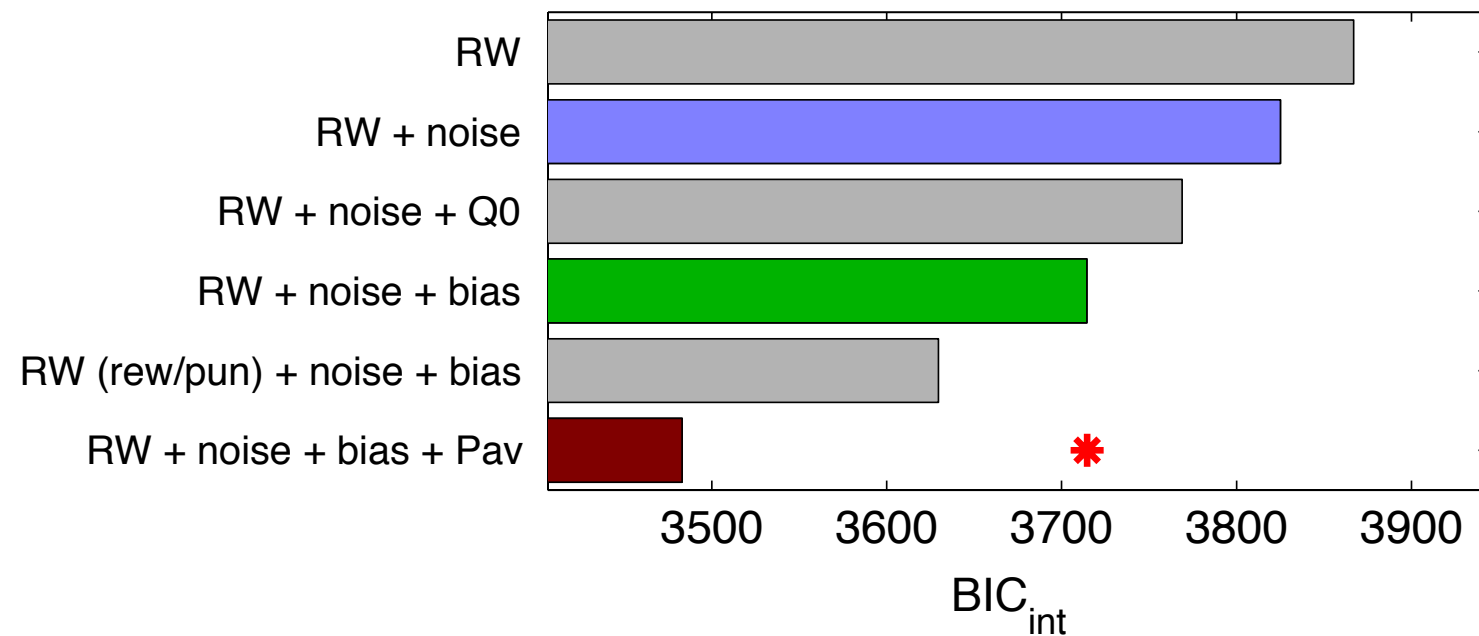


Guitart-Masip et al. 2011, Guitart-Masip, Huys et al. *Submitted*

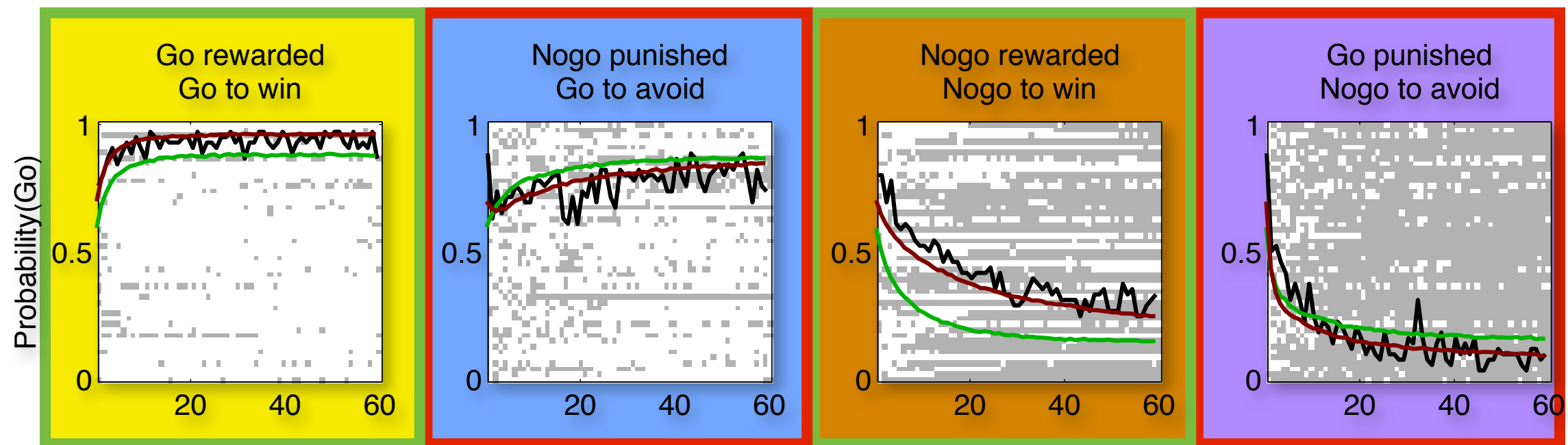
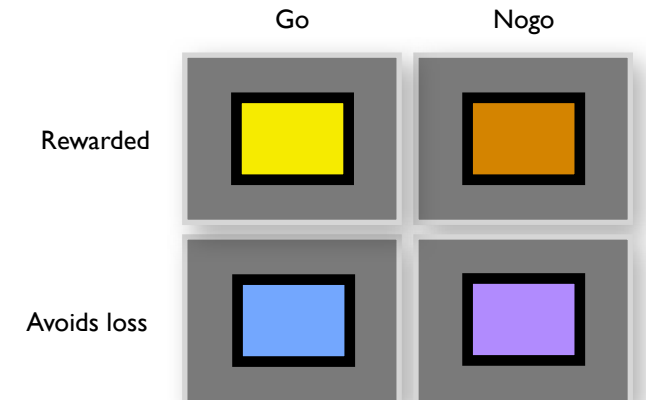
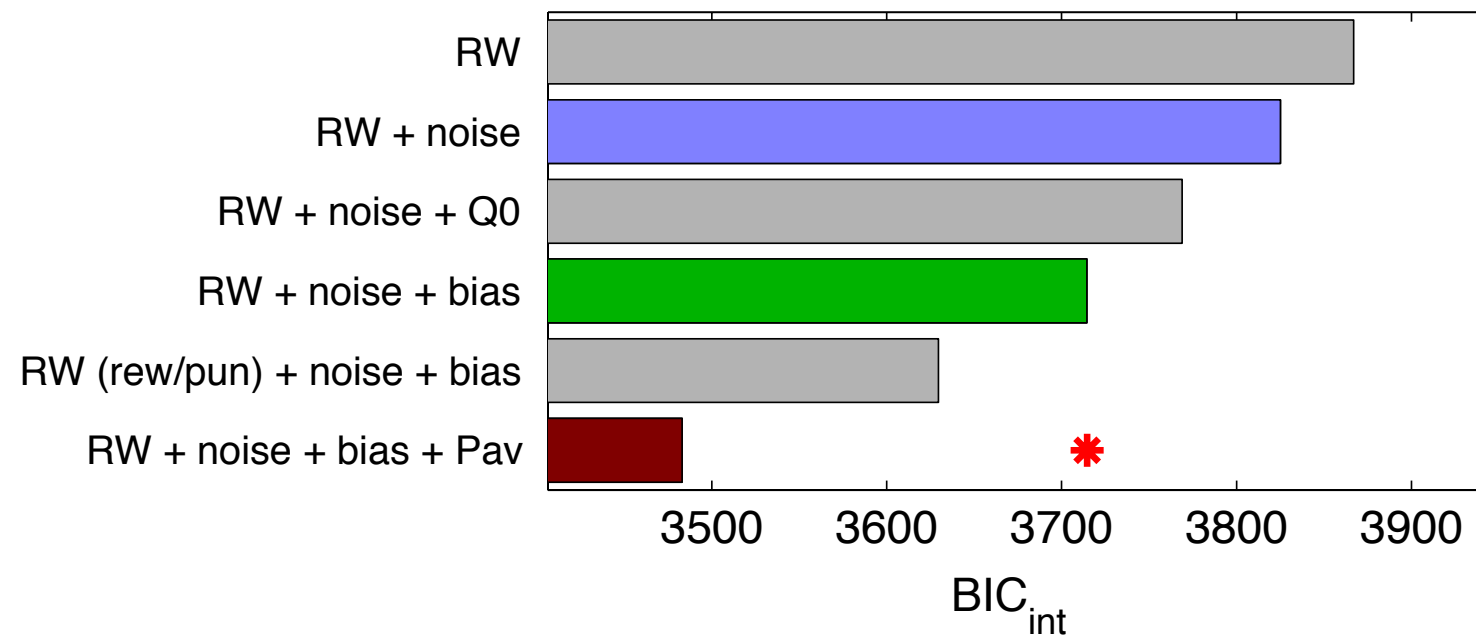
# Model comparison: overfitting?



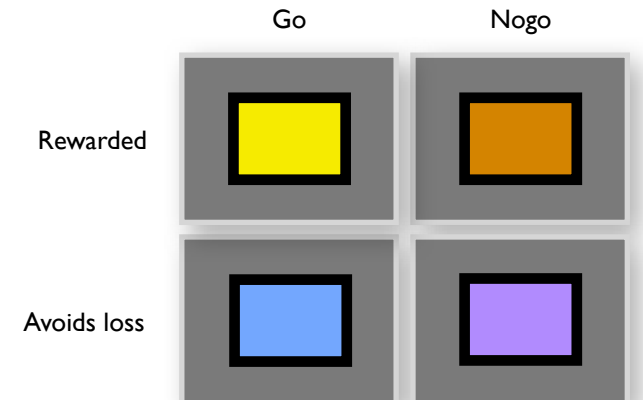
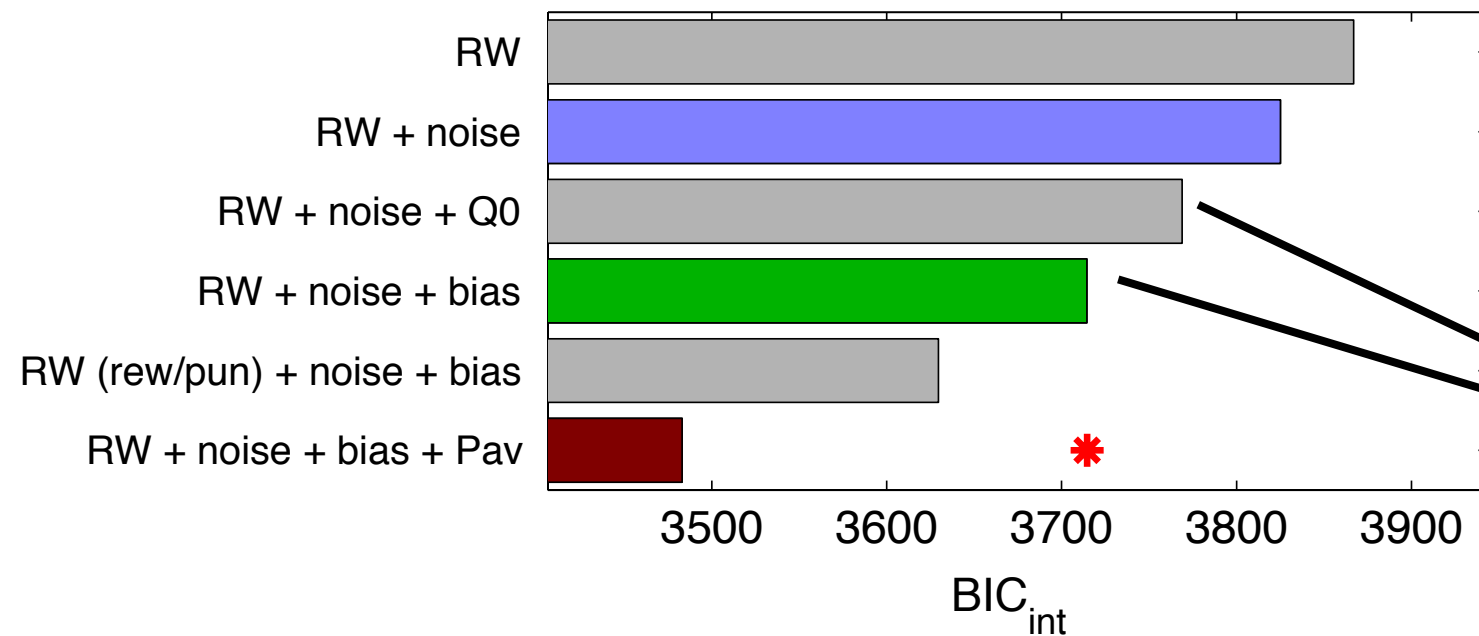
# Model comparison: overfitting?



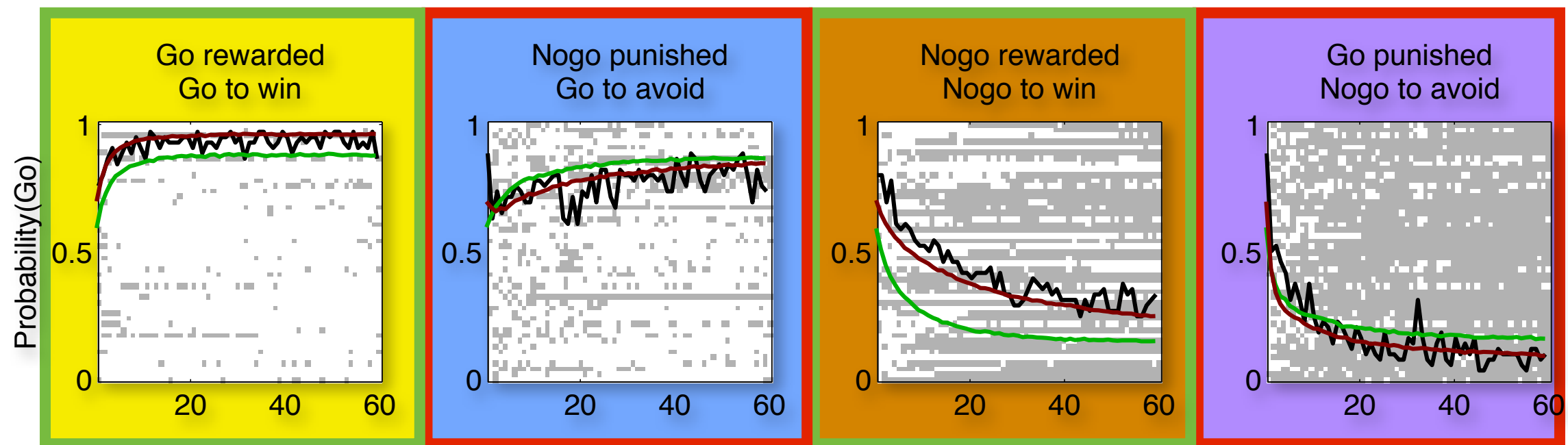
# Model comparison: overfitting?



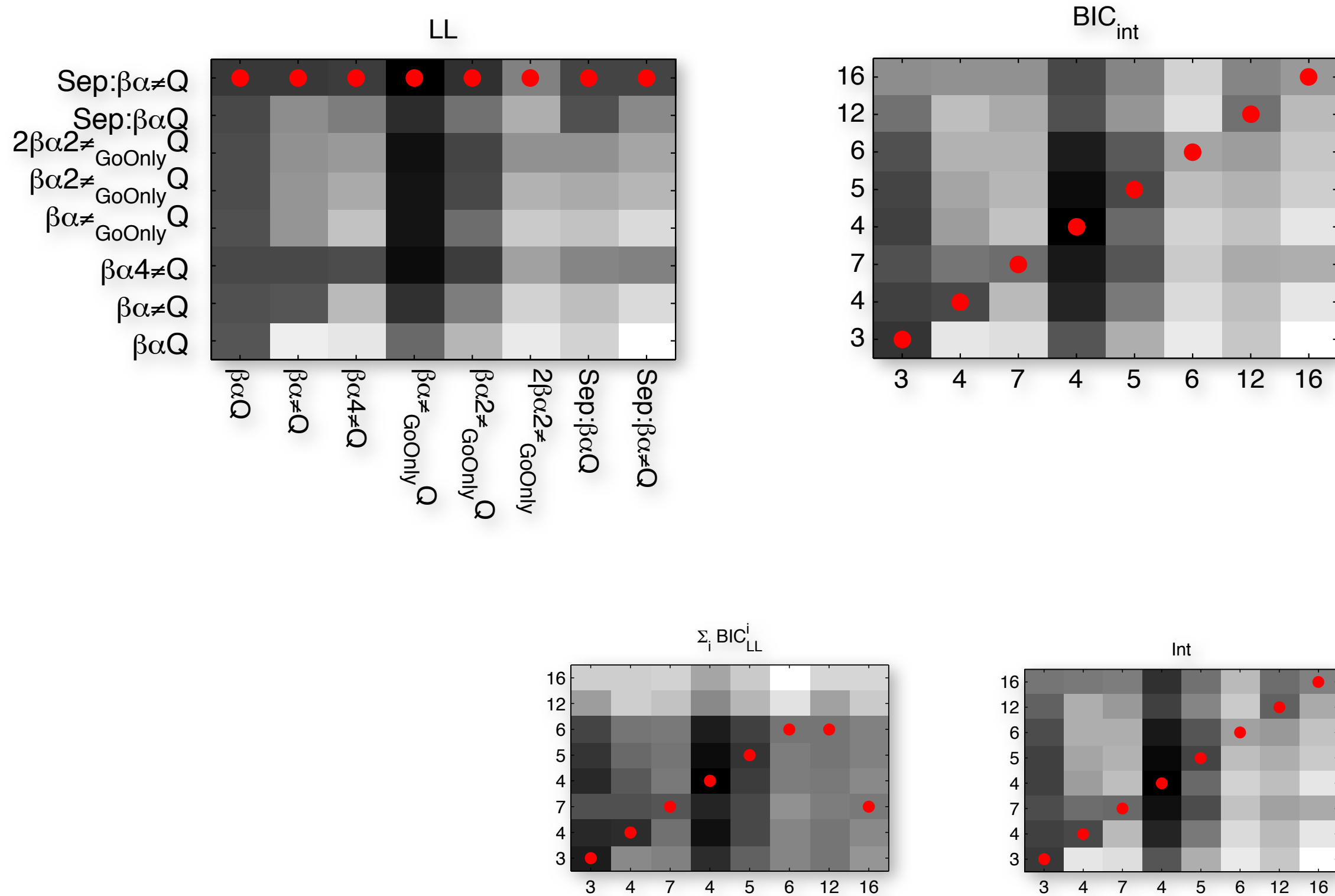
# Model comparison: overfitting?



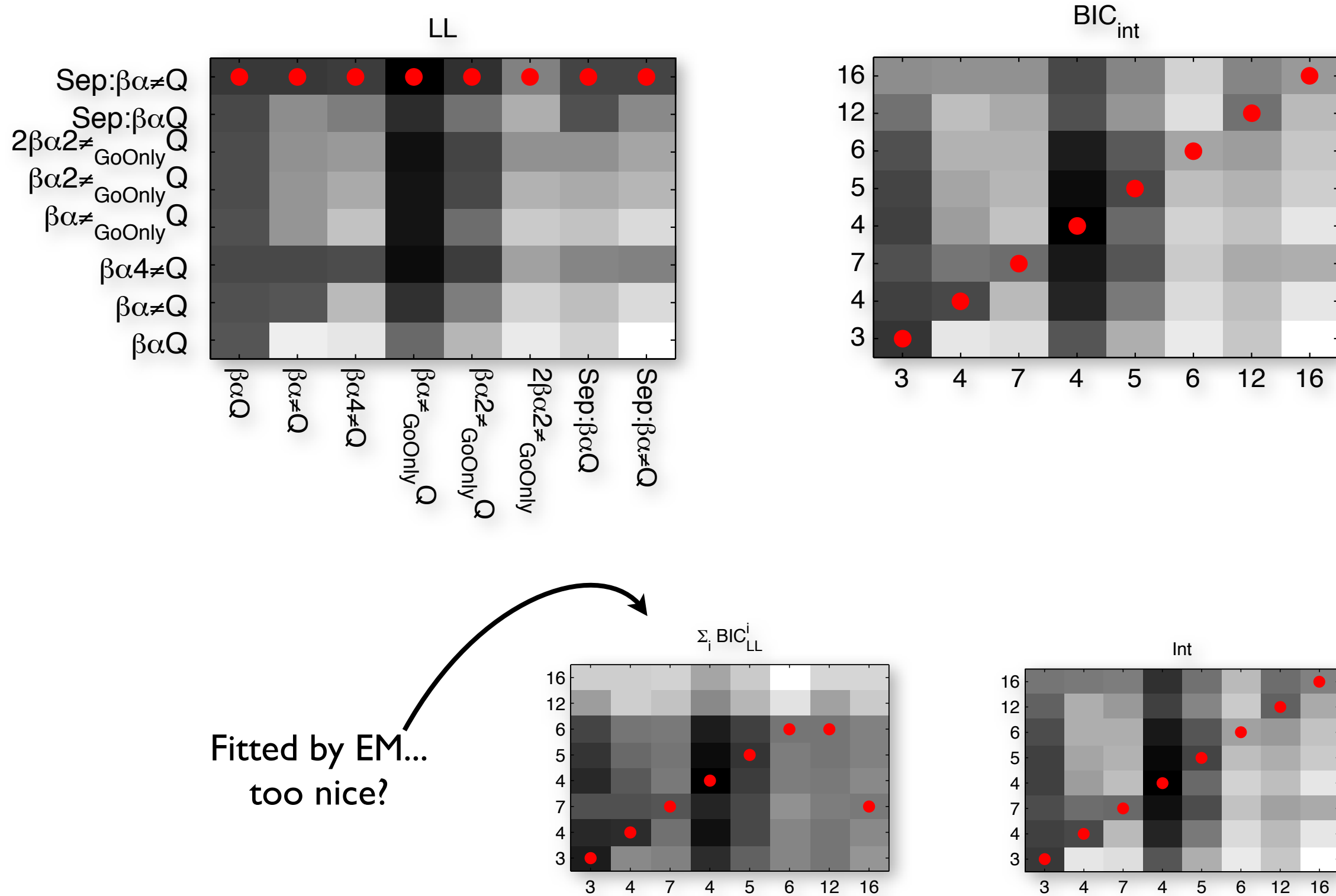
Note: same number of parameters



# How does it do?



# How does it do?



# Top-level Laplacian approximation

---

## ► Estimating the top-level determinant

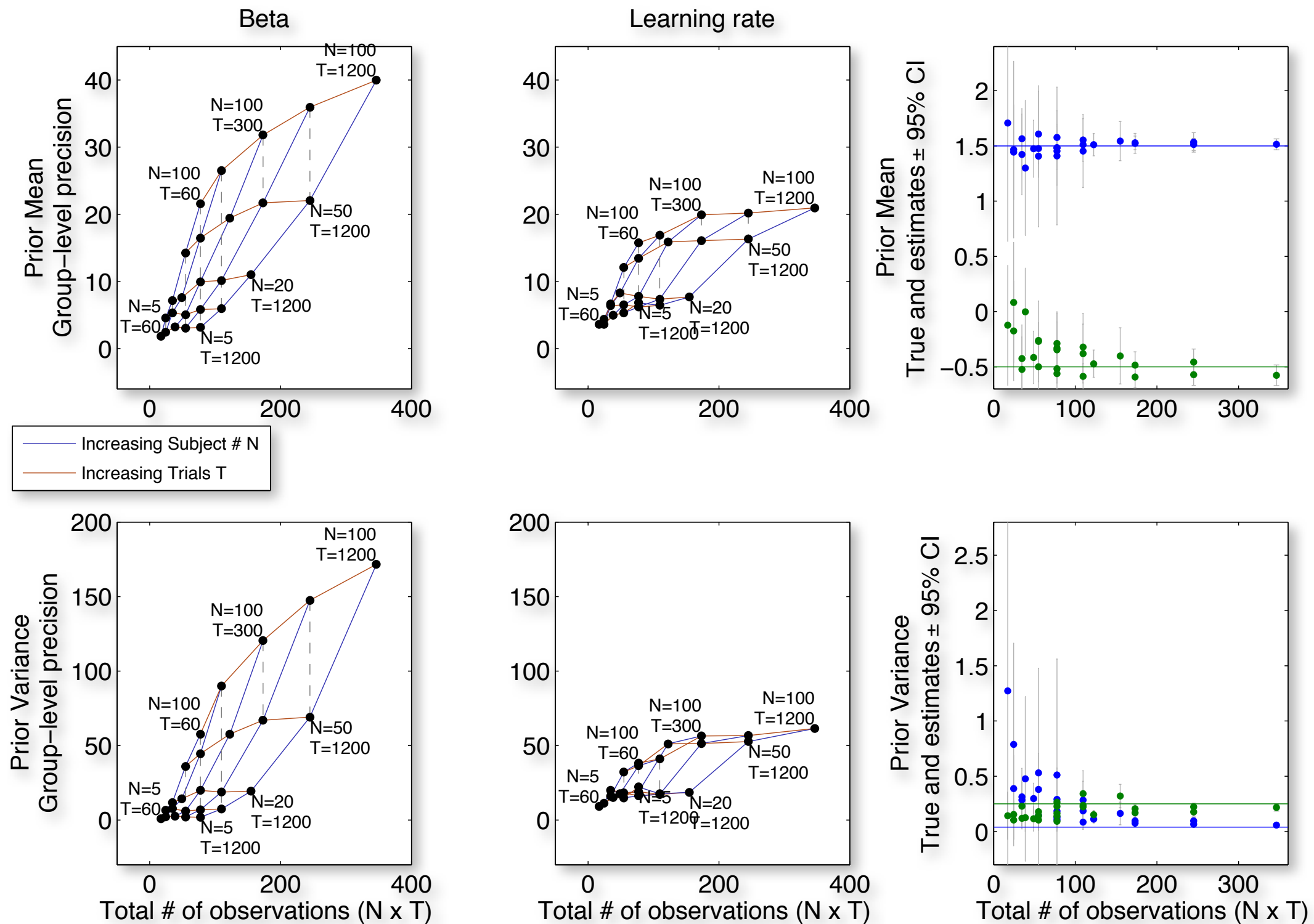
- using 2nd order finite differences

$$\left. \frac{d^2}{dh_{ij}^2} p(\mathcal{A}|\boldsymbol{\zeta}) \right|_{\boldsymbol{\zeta}=\hat{\boldsymbol{\zeta}}^{ML}} \approx \frac{1}{\delta^2} \left[ p(\mathcal{A}|\hat{\boldsymbol{\zeta}}^{ML} + \delta \mathbf{e}_i) - 2p(\mathcal{A}|\hat{\boldsymbol{\zeta}}^{ML}) + p(\mathcal{A}|\hat{\boldsymbol{\zeta}}^{ML} - \delta \mathbf{e}_j) \right]$$

- the shifted likelihoods can be evaluated by shifting the samples.

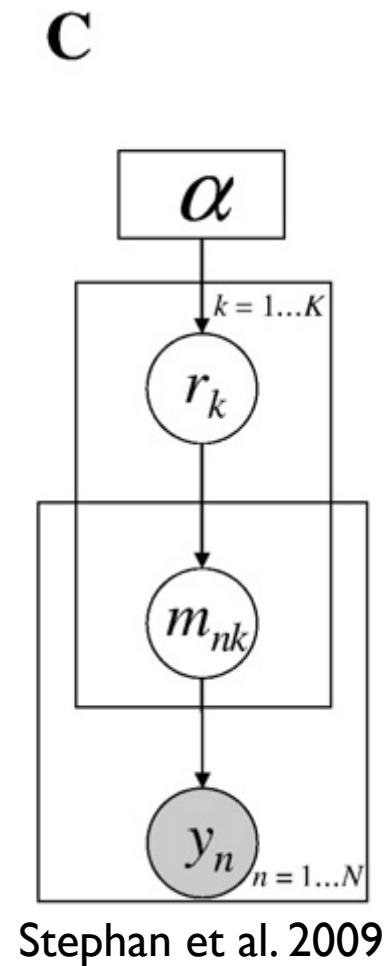
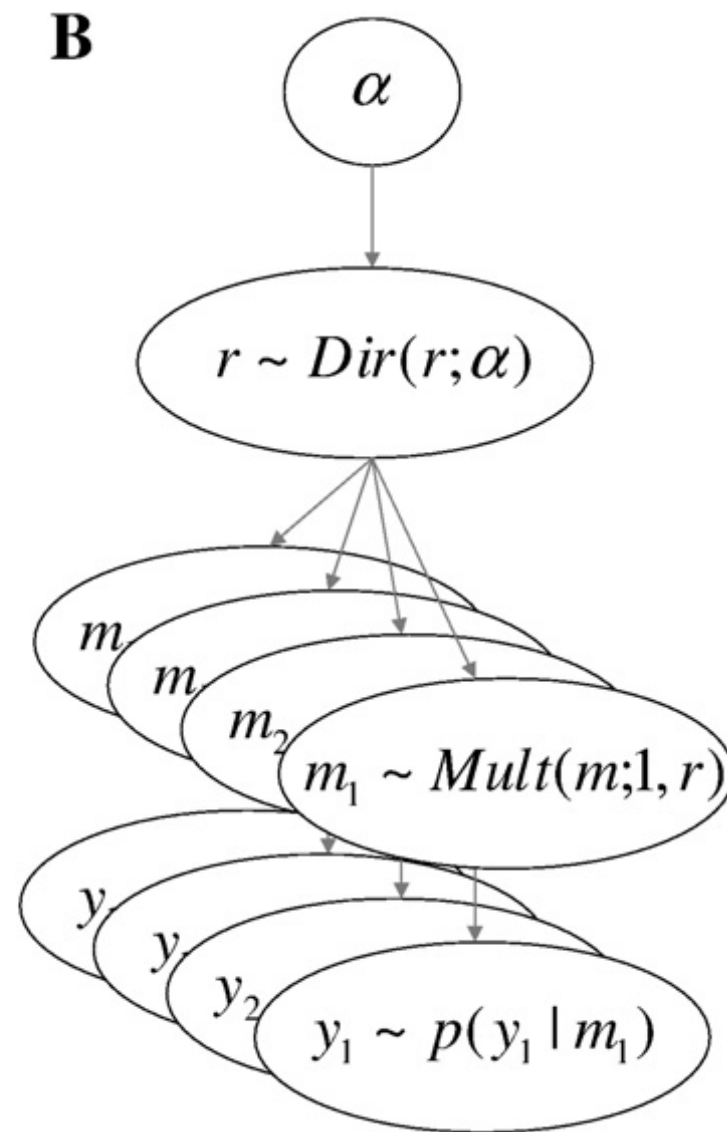


# Group level errors



# Posterior distribution on models

## ► Generative model for models



# Bayesian model selection - equations

---

- ▶ Write down joint distribution of generative model
- ▶ Variational approximations lead to set of very simple update equations
  - start with flat prior over model probabilities

$$\alpha = \alpha_0$$

- then update

$$u_k^i = \left( \int d\theta_i p(\mathcal{A}_i, \theta_i | \mathcal{M}_k) \right) \exp \left( \Psi(\alpha_k) - \Psi \left( \sum_k \alpha_k \right) \right)$$
$$\alpha_k \leftarrow \alpha_{0,k} + \sum_i \frac{u_k^i}{\sum_k u_k^i}$$

# Group Model selection

---

Integrate out your parameters

# Questions in psychiatry I: regression

## ► Parametric relationship with other variables $\psi$

- do standard second level analyses
- can use Hessians to determine weights

$$\begin{aligned}\text{E step: } q_k(\theta) &= \mathcal{N}(\mathbf{m}_k, \mathbf{S}_k) \\ \mathbf{m}_k &\leftarrow \underset{\theta}{\operatorname{argmax}} p(\mathbf{a}_k|\theta)p(\theta|\zeta^{(i)})\end{aligned}$$

$$\mathbf{S}_k^{-1} \leftarrow \frac{\partial^2 p(\mathbf{a}^k|\theta)p(\theta|\zeta^{(i)})}{\partial \theta^2} \Big|_{\theta=\mathbf{m}_k}$$

- better: compare two models

$$\begin{aligned}\text{Model 1: } & \prod_i p(\mathcal{A}_i|\theta_i) p(\theta_i|\mu_0, \sigma) \\ \text{i.e. } & \theta_i \sim \mathcal{N}(\mu_0, \sigma)\end{aligned}$$

$$\begin{aligned}\text{Model 2: } & \prod_i p(\mathcal{A}_i|\theta_i) p(\theta_i|\mu_0, c, \sigma, \psi_i) \\ \text{i.e. } & \theta_i \sim \mathcal{N}(\mu_0 + c\psi_i, \sigma)\end{aligned}$$

- ▶ Standard regression analysis:

$$\mathbf{m}_i = \mathbf{C}\mathbf{r}_i + \Sigma^{1/2}\boldsymbol{\eta} \quad \forall i$$

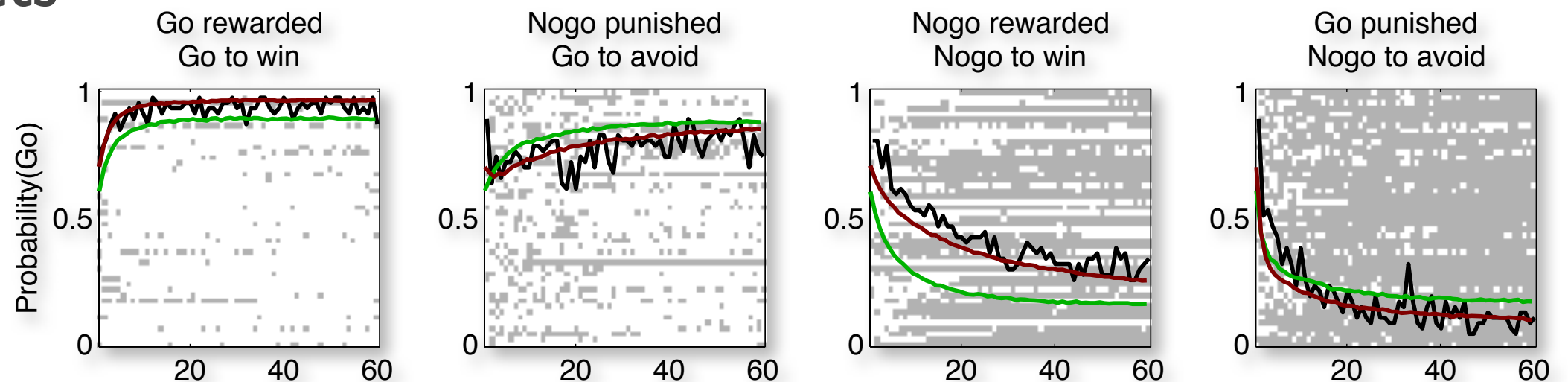
- ▶ Including uncertainty about each subject's inferred parameters

$$\mathbf{m}_i = \mathbf{C}\mathbf{r}_i + (\Sigma^{1/2} + \mathbf{S}_i^{1/2})\boldsymbol{\eta} \quad \forall i$$

- ▶ Careful: Finite difference estimates  $\mathbf{S}$  can be noisy!
  - regularize...

# Questions in psychiatry II: group differences

- ▶ Do groups differ in terms of parameter(s)?
- ▶ **Cannot** compare parameters across different models
  - even very similar parameters can account for different effects



- ▶ For models with  $k$  parameters, there are  $2^k$  possible comparisons
  - multiple comparisons?
  - posterior over models (Stephan et al. 2009)

# Group differences in parameters

---

- ▶ Are two groups similar in parameter  $x$ ?
- ▶ ANOVA: compare likelihood of two means to likelihood of one global mean. Take degrees of freedom into account.
- ▶ But: this tries to account for the parameters with one or two groups, not for the data
- ▶ Compare models with separate or joint parameter & prior:

Model 1	$\varepsilon$	$\beta_1, \beta_2$
Model 2	$\varepsilon$	$\beta$



# Questions in psychiatry III: Classification

---

- ▶ Who belongs to which of two groups?
- ▶ How many groups are there?

# Model comparison again

---

## ► What is ‘significant’?

$$BF = \frac{p(\mathcal{A}|\mathcal{M}_1)}{p(\mathcal{A}|\mathcal{M}_2)}$$
$$p(\Lambda < \eta)$$

$\log_{10}(B_{10})$	$B_{10}$	Evidence against $H_0$
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
>2	>100	Decisive

Kaas and Raftery 95

## ► “Spread of effect” in group comparisons

- Better model does not mean a behavioural effect is concentrated in one parameter
- Obvious raw differences spread between parameters

# Behavioural data modelling

---

## ► Are no panacea

- statistics about specific aspects of decision machinery
- only account for part of the variance

## ► Model needs to match experiment

- ensure subjects actually do the task the way you wrote it in the model
- model comparison

## ► Model = Quantitative hypothesis

- strong test
- need to compare **models**, not **parameters**
- includes all consequences of a hypothesis for choice

# Modelling in psychiatry

---

- ▶ Hypothesis testing
  - otherwise untestable hypotheses
  - internal processes
- ▶ Limited by data quality
  - Look for strong behaviours, not noisy
- ▶ “Holistic” testing of hypotheses
- ▶ Marr’s levels
  - physical
  - algorithm
  - computational