Bayesian Optimization for Fast Inference of Simulator-Based Models

Michael Gutmann

https://sites.google.com/site/michaelgutmann

University of Helsinki and Aalto University

December 2015

Program

Simulator-based models
- Recap of statistical inference
- What are simulator-based models?
- Pros and cons

Inference with simulator-based models
- Difficulties
- Classical solutions
- Critique

Speeding up the inference with Bayesian optimization
- Approach via statistical modeling
- Leveraging Bayesian optimization
- Application

Big picture of statistical inference

- Given: A statistical model which describes data $\mathbf{y} = (y_1, \ldots, y_n)$; the model has parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$
- Given: Observed data $\mathbf{y}^o$
- Possibly given: A (prior) probability density function (pdf) for $\boldsymbol{\theta}$, $p_{\boldsymbol{\theta}}$
- Wanted: Some probabilistic statement about $\boldsymbol{\theta}$
  - which value of $\boldsymbol{\theta}$ has generated $\mathbf{y}^o$ most likely?
  - what is the mean value of $\boldsymbol{\theta}$ given $\mathbf{y}^o$?
  - which interval contains θ_1 with probability 0.95 ?
  - ...

Family of pdfs as statistical model

- Let statistical model $\equiv$ family of pdfs $p_{\mathbf{y}|\boldsymbol{\theta}}$ indexed by $\boldsymbol{\theta}$
- Mechanics of maximum likelihood estimation:

$$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} p_{\mathbf{y}|\boldsymbol{\theta}}(\mathbf{y}^o|\boldsymbol{\theta}) \tag{1}$$

- Mechanics of Bayesian inference:

$$p_{\boldsymbol{\theta}|\mathbf{y}}(\boldsymbol{\theta}|\mathbf{y}^o) \propto p_{\mathbf{y}|\boldsymbol{\theta}}(\mathbf{y}^o|\boldsymbol{\theta}) \times p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \tag{2}$$

$$\text{posterior} \propto \text{likelihood function} \times \text{prior} \tag{3}$$

- Often written without subscripts ("function overloading")

$$p(\boldsymbol{\theta}|\mathbf{y}^o) \propto p(\mathbf{y}^o|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}) \tag{4}$$

Likelihood function

- $L(\boldsymbol{\theta}) = p(\mathbf{y}^o|\boldsymbol{\theta})$ indicates how likely/plausible different parameter values are for the observed data.
- For discrete random variables:

$$p(\mathbf{y}^o|\boldsymbol{\theta}) = \Pr(\mathbf{y} = \mathbf{y}^o|\boldsymbol{\theta}) \tag{5}$$

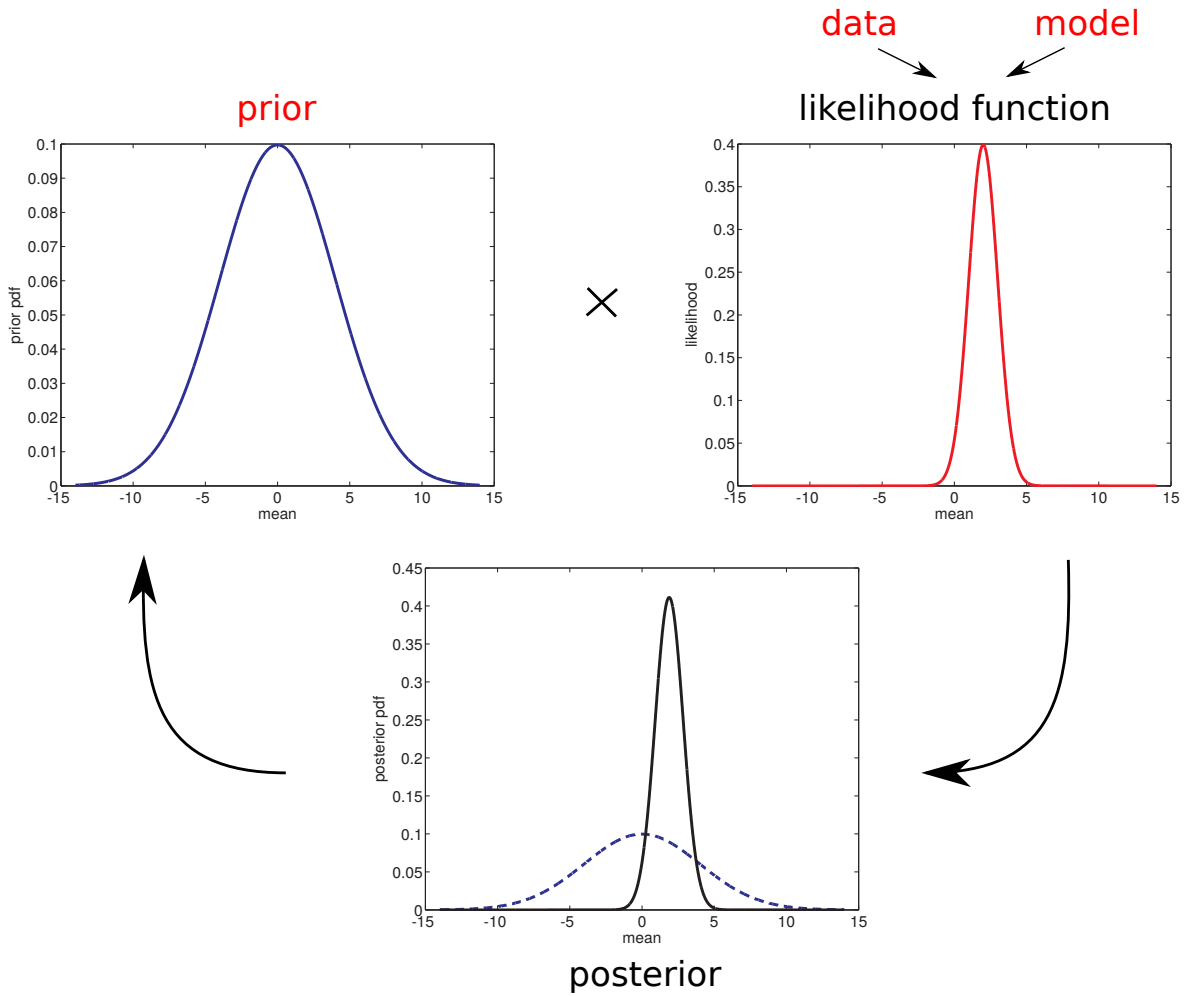
  Probability that sampling from the model with parameter value $\boldsymbol{\theta}$ yields data $\mathbf{y}$ which are equal to $\mathbf{y}^o$.
- For continuous random variables:

$$p(\mathbf{y}^o|\boldsymbol{\theta}) = \lim_{\epsilon \to 0} \frac{\Pr(\mathbf{y} \in B_\epsilon(\mathbf{y}^o)|\boldsymbol{\theta})}{\mathrm{Vol}(B_\epsilon(\mathbf{y}^o))} \tag{6}$$

  Proportional to the probability that sampling yields data in a small ball $B_\epsilon(\mathbf{y}^o)$ around $\mathbf{y}^o$.

Example

$$p(\theta) = \frac{1}{\sqrt{2\pi \cdot 4^2}} \exp\left(-\frac{\theta^2}{2 \cdot 4^2}\right) \qquad y^o = 2 \qquad p(y|\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-\theta)^2}{2}\right)$$



Other specifications of statistical models

- Many statistical models are defined via a family of pdfs.
- Statistical models can be specified in other ways as well.
- *Here: models which are specified via some mechanism for generating data ("simulator-based models")*
- Example: Instead of

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-\theta)^2}{2}\right) \tag{7}$$

we could have specified the model via

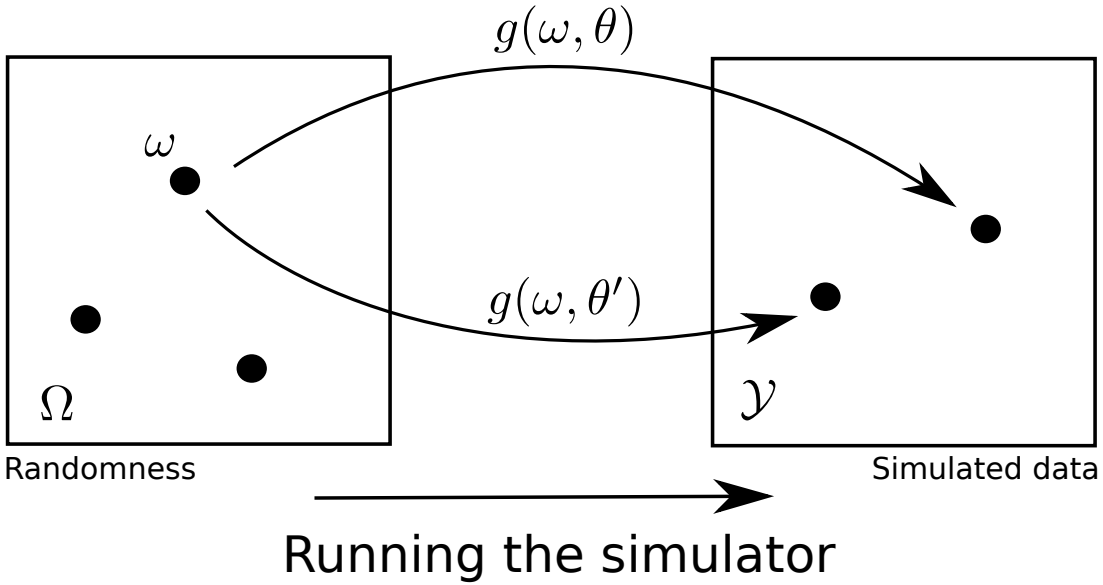
$$y = z + \theta \qquad z \sim \mathcal{N}(0,1) \tag{8}$$

Definition of simulator-based models

- Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space.
- A simulator-based model is a collection of (measurable) functions $g(., \boldsymbol{\theta})$ parametrized by $\boldsymbol{\theta}$,

$$\boldsymbol{\omega} \in \Omega \mapsto \mathbf{y} = g(\boldsymbol{\omega}, \boldsymbol{\theta}) \in \mathcal{Y} \tag{9}$$

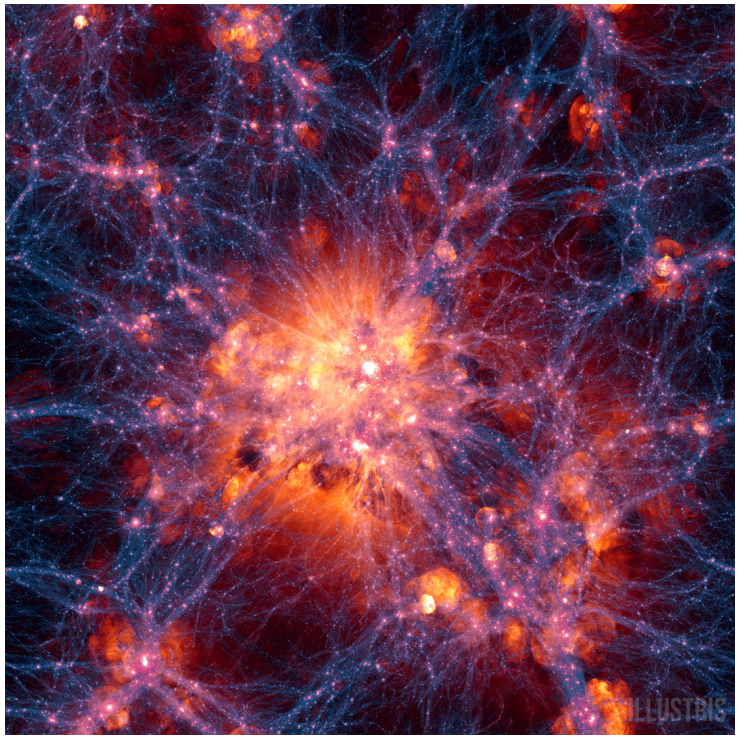
- For any fixed $\boldsymbol{\theta}$, $\mathbf{y}_{\boldsymbol{\theta}} = g(., \boldsymbol{\theta})$ is a random variable.



Examples of simulator-based models

Simulator-based models are used in:

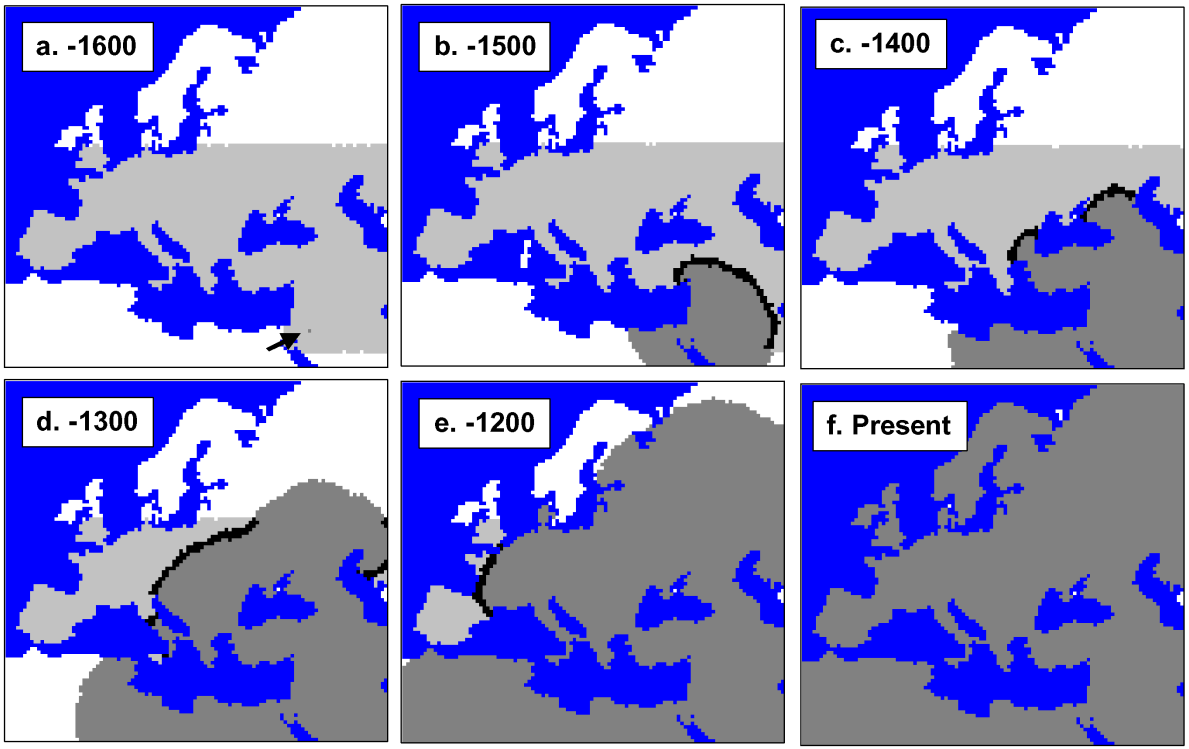
- Astrophysics:
  Simulating the formation of galaxies, stars, or planets
- Evolutionary biology:
  Simulating the evolution of life
- Health science:
  Simulating the spread of an infectious disease
- . . .



Dark matter density simulated by the Illustris collaboration
(Figure from `http://www.illustris-project.org`)

Examples (evolutionary biology)

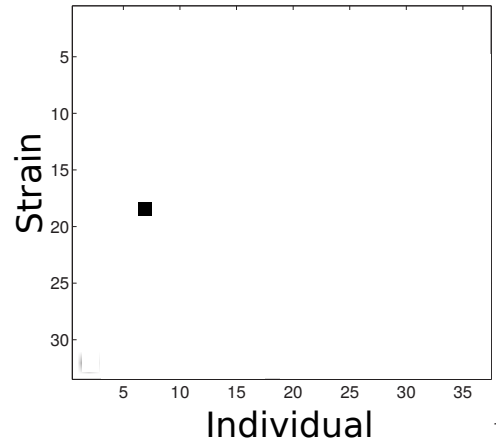
- Simulation of different hypothesized evolutionary scenarios
- Interaction between early modern humans (Homo sapiens) and their Neanderthal contemporaries in Europe



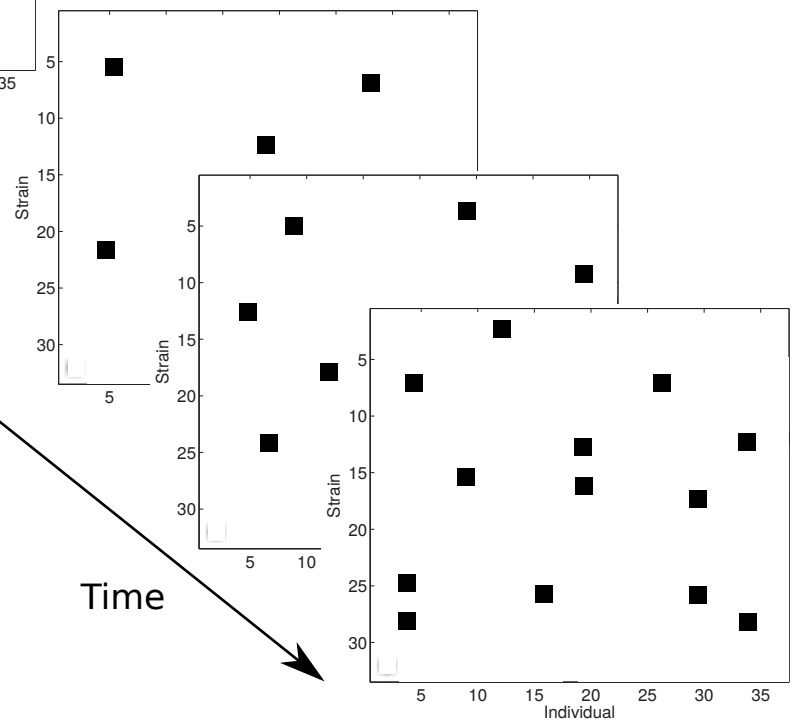
Immigration of Modern Humans into Europe from the Near East. Light gray: Neanderthal population. Dark: Homo sapiens. from (Currat and Excoffier, *Plos Biology*, 2004, 10.1371/journal.pbio.0020421). The numbers in the figures indicate generations. See also Pinhasi et al, The genetic history of Europeans, *Trends in Genetics*, 2012

Examples (health science)

- Simulation of bacterial transmission dynamics in child day care centers (Numminen et al, *Biometrics*, 2013)



$$
\begin{aligned}
\Pr(I_{is}^{t+h} = 0 | I_{is}^{t} = 1) &= h + o(h) \\
\Pr(I_{is}^{t+h} = 1 | I_{is'}^{t} = 0 \; \forall s') &= R_s(t)h + o(h) \\
\Pr(I_{is}^{t+h} = 1 | I_{is}^{t} = 0, \; \exists s' : I_{is'}^{t} = 1) &= \theta R_s(t)h + o(h) \\
R_s(t) &= \beta E_s(t) + \Lambda P_s
\end{aligned}
$$



Other names for simulator-based models

- Models specified via a data generating mechanism occur in multiple and diverse scientific fields.
- Different communities use different names for simulator-based models:
  - Generative models
  - Implicit models
  - Stochastic simulation models
  - Probabilistic programs

Advantages of simulator-based models

- Direct implementation of hypotheses of how the observed data were generated.
- Neat interface with scientific models (e.g. from physics or biology).
- Modeling by replicating the mechanisms of nature which produced the observed/measured data. ("Analysis by synthesis")
- Possibility to perform experiments in silico.

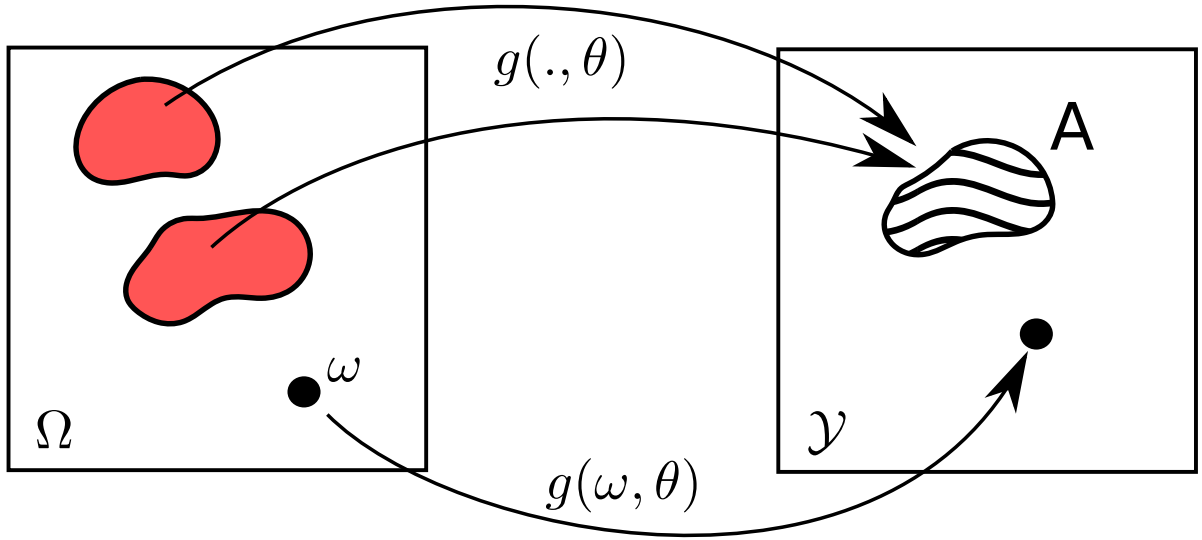
Disadvantages of simulator-based models

- Generally elude analytical treatment.
- Can be easily made more complicated than necessary.
- Statistical inference is difficult. Main reason:
  The simulator defines the model pdfs only implicitly.

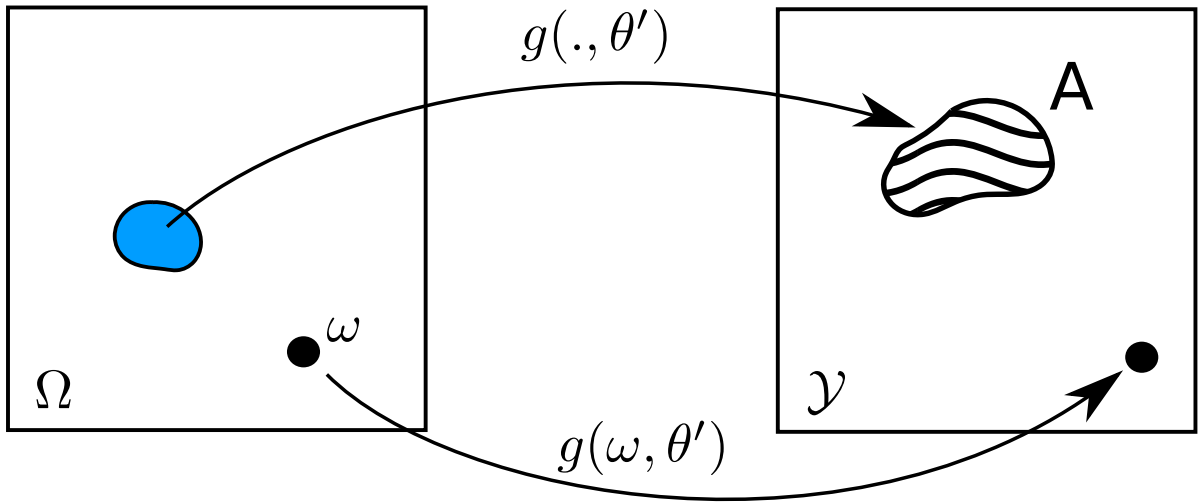
Implicit definition of the model pdfs

$$\Pr\left(y \in A \mid \theta\right) = \mathcal{P}\left(\{\omega : g(\omega, \theta) \in A\}\right)$$

Parameter value θ



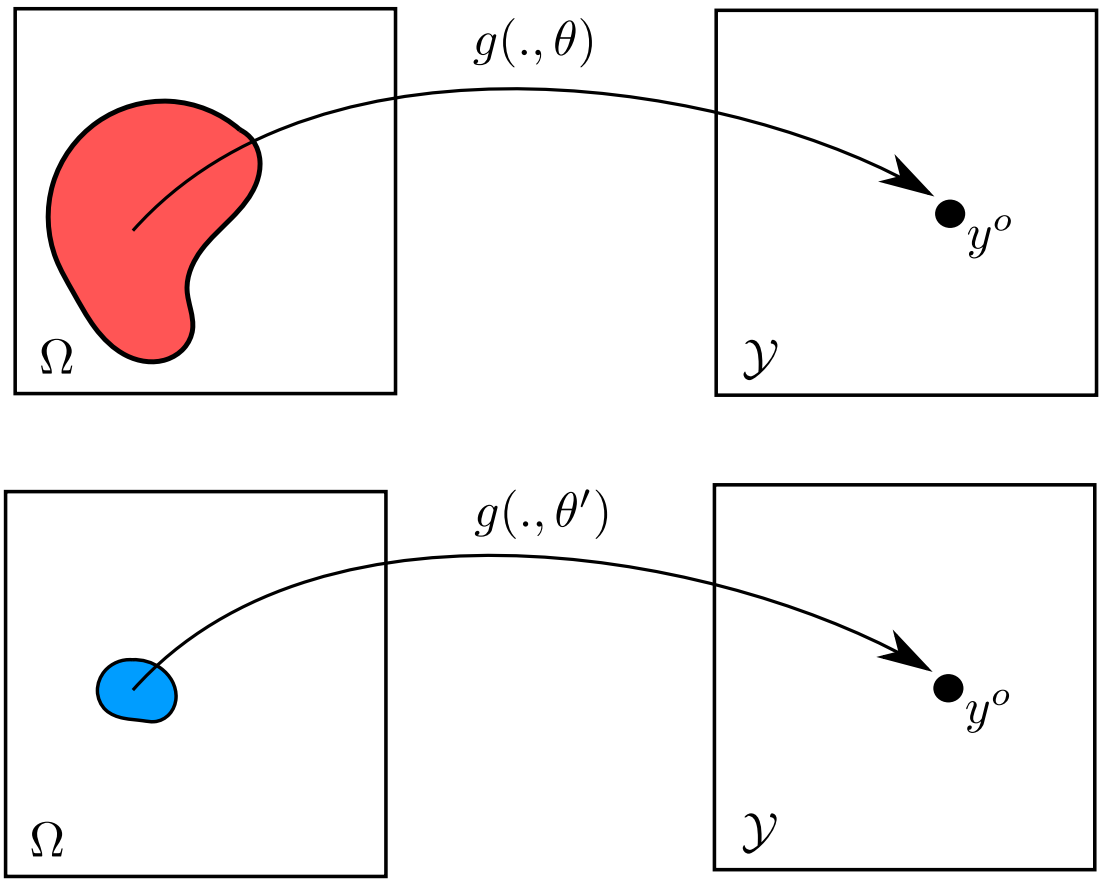
Parameter value θ'



Implicit definition of the likelihood function

- For discrete random variables:

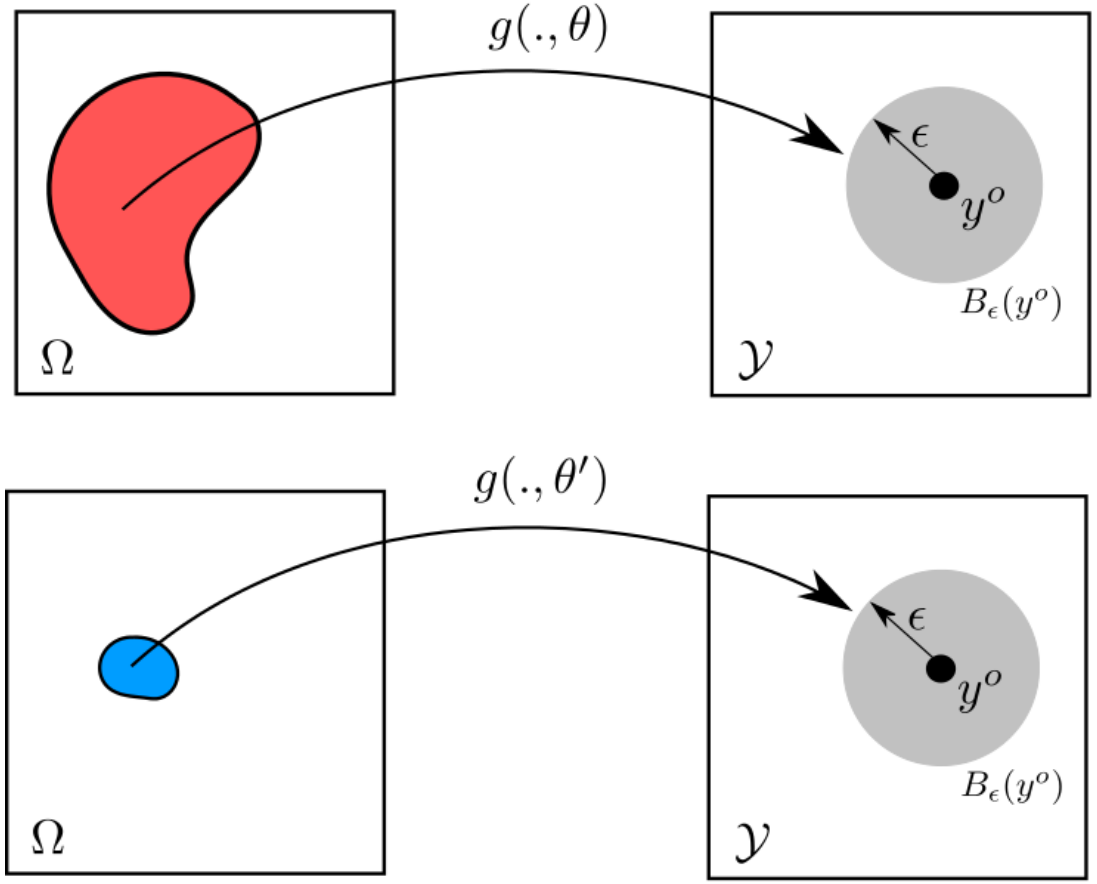
$$L(\theta) = \Pr\left(y = y^o \mid \theta\right) = \mathcal{P}\left(\{\omega : g(\omega, \theta) = y^o\}\right)$$



Implicit definition of the likelihood function

- For continuous random variables: $L(\boldsymbol{\theta}) = \lim_{\epsilon \to 0} L_\epsilon(\boldsymbol{\theta})$

$$L_\epsilon(\theta) = \frac{\Pr(y \in B_\epsilon(y^o) \mid \theta)}{V_\epsilon} = \frac{\mathcal{P}(\{\omega : g(\omega, \theta) \in B_\epsilon(y^o)\})}{V_\epsilon}$$



Intractability of statistical inference

- To compute the likelihood function, we need to compute the probability that the simulator generates data close to $\mathbf{y}^o$,

$$\Pr(\mathbf{y} = \mathbf{y}^o|\boldsymbol{\theta}) \quad \text{or} \quad \Pr(\mathbf{y} \in B_\epsilon(\mathbf{y}^o)|\boldsymbol{\theta})$$

- No analytical expression available.
- Likelihood function generally well defined but incomputable
- Exact inference generally impossible

Towards approximate inference

- It is possible to empirically test whether simulated data equal $\mathbf{y}^o$ or are in $B_\epsilon(\mathbf{y}^o)$.
- This property provides a means to perform approximate inference for simulator-based models.
- Principle:
  1. Sample $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$
  2. Sample $\mathbf{y}|\boldsymbol{\theta}$ by running the simulator
  3. Accept $\boldsymbol{\theta}$ if $\mathbf{y} \in B_\epsilon(\mathbf{y}^o)$
- The accepted samples follow a posterior proportional to $p(\boldsymbol{\theta})L_\epsilon(\boldsymbol{\theta})$.

Towards approximate inference ("proof")

- Before the accept/reject step, $\boldsymbol{\theta}$ has pdf $p(\boldsymbol{\theta})$
- Associate with each $\boldsymbol{\theta}$ a binary random variable z with success probability $r_{\boldsymbol{\theta}}$,

$$r_{\boldsymbol{\theta}} = \Pr(z = 1|\boldsymbol{\theta}) = \Pr(\mathbf{y} \in B_\epsilon(\mathbf{y}^o)|\boldsymbol{\theta}) \propto L_\epsilon(\boldsymbol{\theta})$$

- $z = 1$ if $\mathbf{y} \in B_\epsilon(\mathbf{y}^o)$, $z = 0$ otherwise.
- Joint distribution of $(\boldsymbol{\theta}, z)$ is $p(\boldsymbol{\theta}) r_{\boldsymbol{\theta}}^z (1 - r_{\boldsymbol{\theta}})^{(z-1)}$.
- Retaining the parameter values which pass the accept-step corresponds to conditioning on $z = 1$.
- The pdf of the accepted samples is thus proportional to $p(\boldsymbol{\theta}) r_{\boldsymbol{\theta}} \propto p(\boldsymbol{\theta}) L_\epsilon(\boldsymbol{\theta})$.

A further approximation

- The conditional acceptance probability $\Pr(\mathbf{y} \in B_\epsilon(\mathbf{y}^o)|\boldsymbol{\theta})$ becomes vanishingly small when the amount of data increases.
- Instead of full data, work with lower dimensional summary statistics $\mathbf{t}_{\boldsymbol{\theta}}$ and $\mathbf{t}^o$,

$$\mathbf{t}_{\boldsymbol{\theta}} = T(\mathbf{y}_{\boldsymbol{\theta}}) \qquad \mathbf{t}^o = T(\mathbf{y}^o). \tag{10}$$

- This yields the rejection algorithm for approximate Bayesian computation (ABC).
- Choosing the summary statistics is generally difficult. Interesting research area but not the topic of this presentation.

For recent work on this topic see e.g.
Gutmann, Dutta, Kaski, and Corander
Statistical Inference of Intractable Generative Models via Classification
arXiv:1407.4981

Rejection ABC algorithm

- Algorithm consists in iterating:
  1. Sample $\boldsymbol{\theta} \sim p_{\boldsymbol{\theta}}$
  2. Sample $\mathbf{y}|\boldsymbol{\theta}$ by running the simulator
  3. Compute the discrepancy $\Delta = d(T(\mathbf{y}^o), T(\mathbf{y}))$ (d may or may not be a metric)
  4. Retain $\boldsymbol{\theta}$ if $\Delta \leq \epsilon$
- Produces samples $\boldsymbol{\theta}$ proportional to $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})\tilde{L}_\epsilon(\boldsymbol{\theta})$

$$\tilde{L}_\epsilon(\boldsymbol{\theta}) \propto \Pr\big(\underbrace{d(T(\mathbf{y}^o), T(\mathbf{y}))}_{\Delta} \leq \epsilon \mid \boldsymbol{\theta}\big) \tag{11}$$

Other ABC algorithms

- Rejection ABC forms the basis for more advanced ABC algorithms
  - Markov Chain Monte Carlo ABC (Marjoram, *PNAS*, 2002)
  - Population Monte Carlo ABC (Sisson et al, *PNAS*, 2007)
- Difference to rejection ABC: $\boldsymbol{\theta}$ is drawn from an adaptively constructed proposal distribution rather than from the prior.
- Improves the computational efficiency.
- Likelihood approximation (rejection step) is the same.

The likelihood approximation

$$\tilde{L}_\epsilon(\boldsymbol{\theta}) \propto \Pr(\Delta \leq \epsilon \mid \boldsymbol{\theta})$$

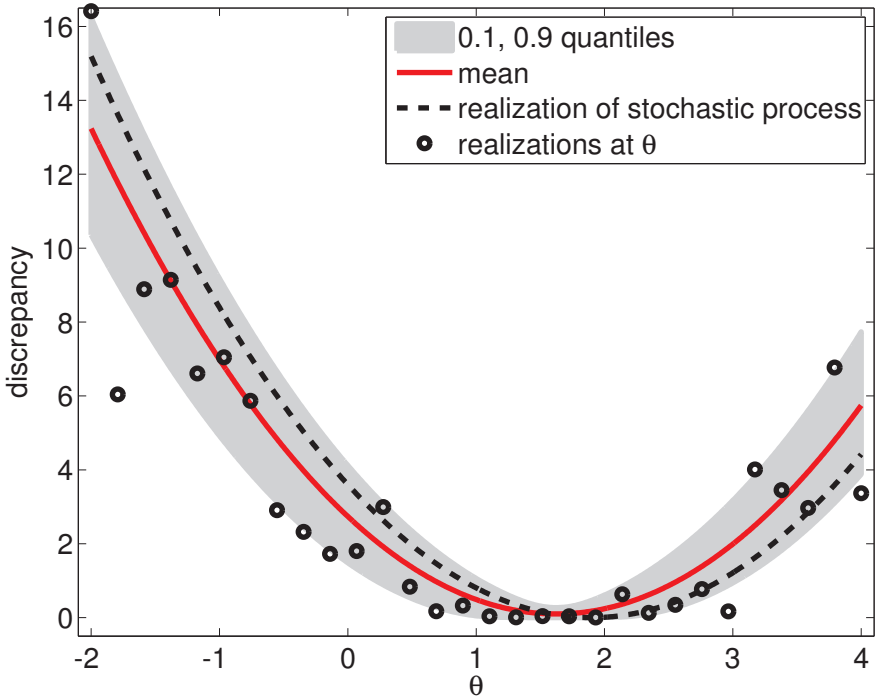
- Inference of the mean θ of a Gaussian of variance one.
- Discrepancy Δ:

$$\Delta_\theta = (\hat{\mu}^o - \hat{\mu}_\theta)^2,$$

$$\hat{\mu}^o = \frac{1}{n}\sum_{i=1}^{n} y_i^o,$$

$$\hat{\mu}_\theta = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

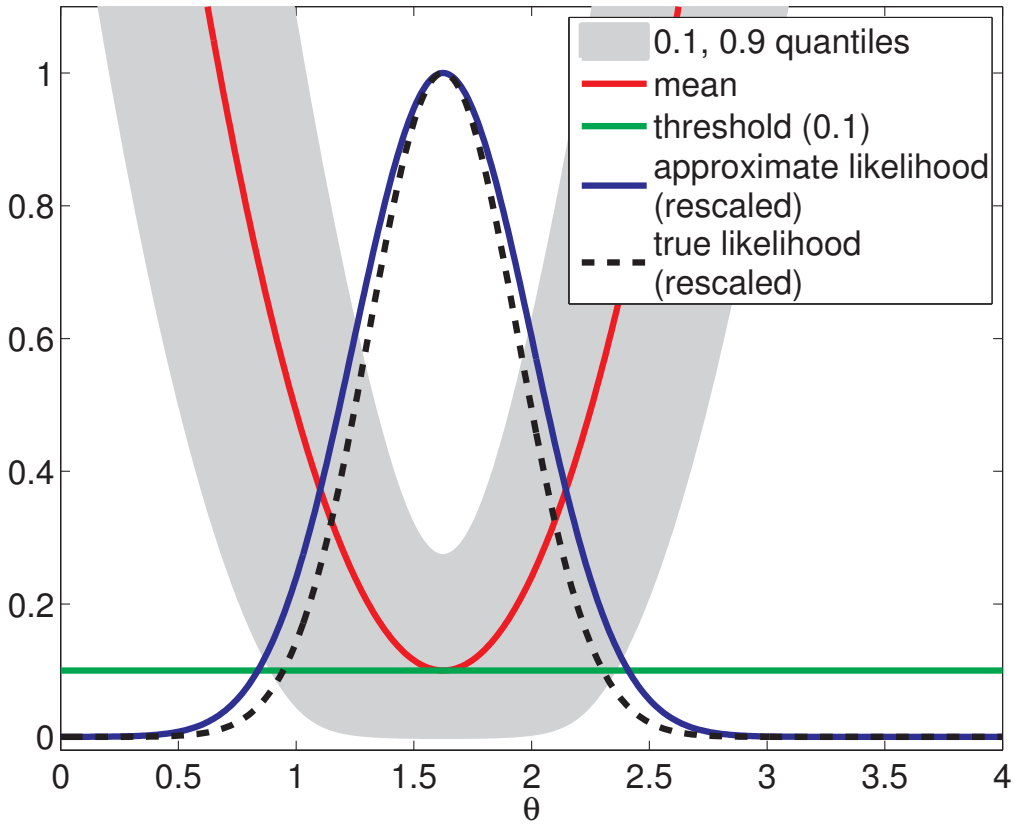
$$y_i \sim \mathcal{N}(\theta, 1)$$



The discrepancy is a random variable.

The likelihood approximation

Probability that Δ_θ is below some threshold ϵ approximates the likelihood function.



Problems with small ϵ

- The conditional probability for $\Delta \leq \epsilon$ can be computed in closed form

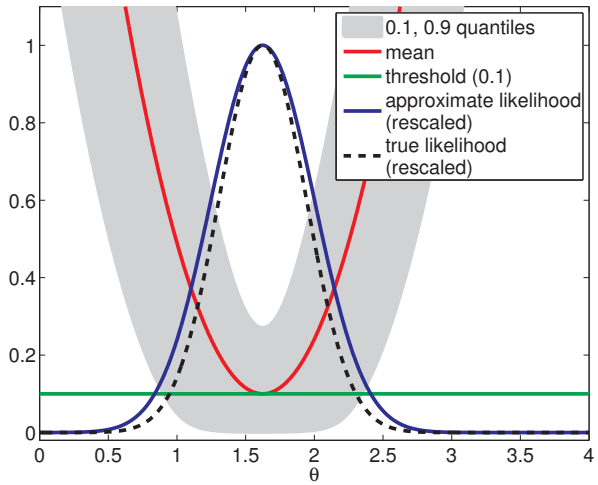
$$\Pr(\Delta \leq \epsilon|\theta) = \Phi\left(\sqrt{n}(\hat{\mu}^o - \theta) + \sqrt{n\epsilon}\right) - \Phi\left(\sqrt{n}(\hat{\mu}^o - \theta) - \sqrt{n\epsilon}\right)$$

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) \mathrm{d}u$$

- For $n\epsilon$ small: $\tilde{L}_\epsilon(\theta) \propto \Pr(\Delta \leq \epsilon|\theta) \propto \sqrt{\epsilon} L(\theta)$

- For small ϵ good approximation of the likelihood function. But: $\Pr(\Delta \leq \epsilon|\theta) \approx 0$

  Very few samples will be accepted, whatever the proposal distribution



An alternative approach via statistical modeling

- The approximate likelihood function $\tilde{L}_\epsilon(\boldsymbol{\theta})$ is determined by the conditional distribution of the (univariate) discrepancy Δ

$$\tilde{L}_\epsilon(\boldsymbol{\theta}) \propto \Pr(\Delta \leq \epsilon \mid \boldsymbol{\theta})$$

- If we knew the distribution of Δ we could compute $\tilde{L}_\epsilon(\boldsymbol{\theta})$.
- Distribution is unknown but we can learn a model from data and approximate $\tilde{L}_\epsilon(\boldsymbol{\theta})$ by $\hat{L}_\epsilon(\boldsymbol{\theta})$,

$$\hat{L}_\epsilon(\boldsymbol{\theta}) \propto \widehat{\Pr}(\Delta \leq \epsilon \mid \boldsymbol{\theta}) \tag{12}$$

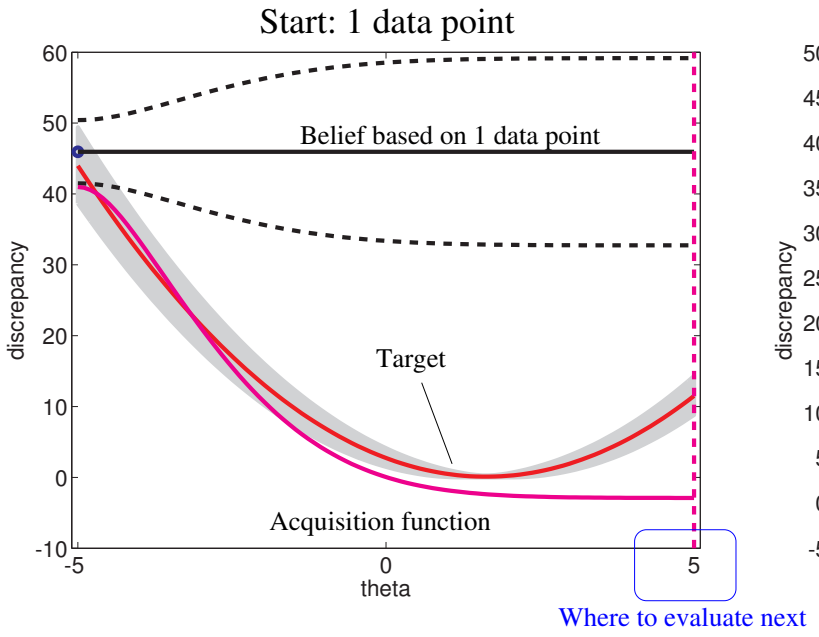
- $\widehat{\Pr}$ is the probability under the model.

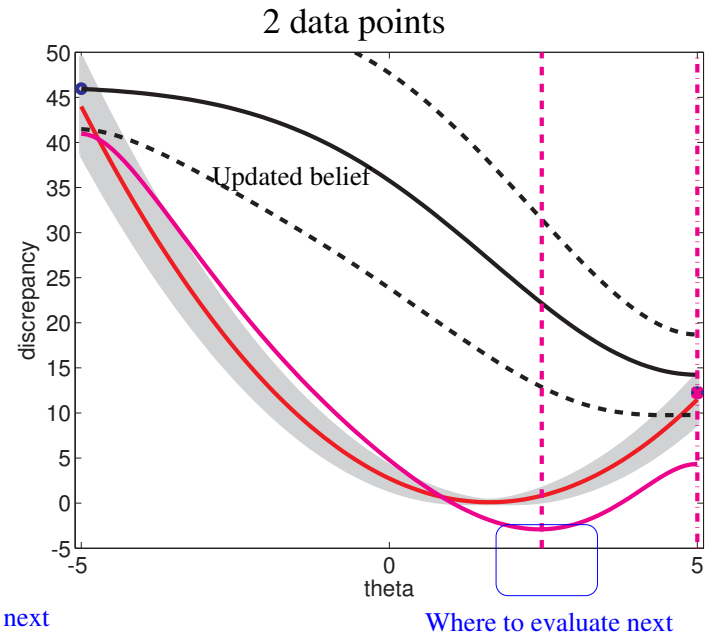
(Gutmann and Corander, *Journal of Machine Learning Research*, in press, 2015)

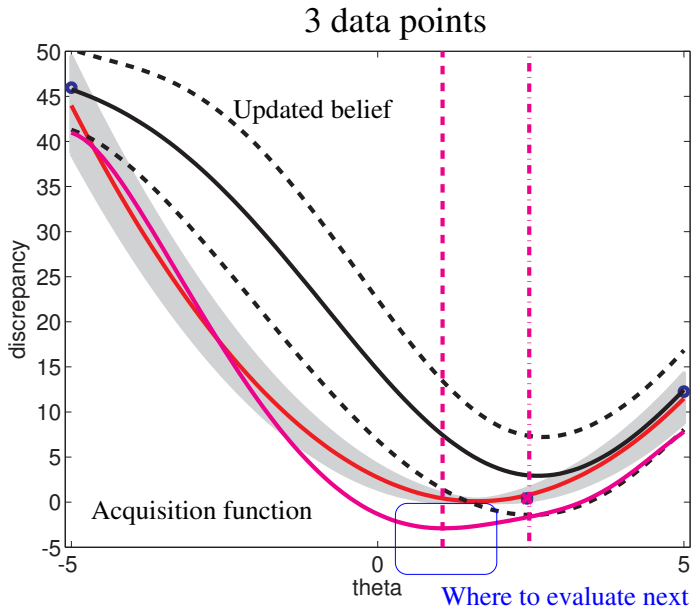
Learning a model of the discrepancy

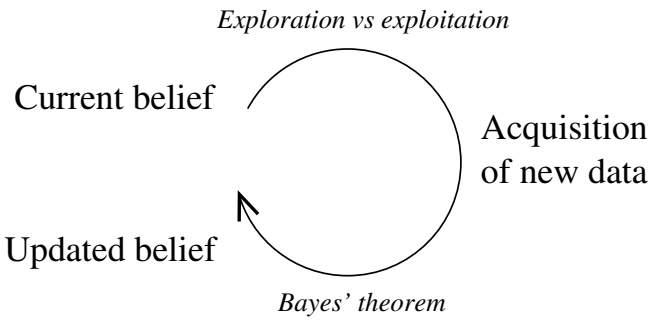
- Data are tuples $(\boldsymbol{\theta}_i, \Delta_i)$.
- Δ_i is obtained by comparing observed data $\mathbf{y}^o$ with simulated data $\mathbf{y}_i$ (generated with parameter value $\boldsymbol{\theta}_i$).
- $\boldsymbol{\theta}_i$ could be obtained by sampling from the prior or from one of the adaptively constructed proposal distributions.
- Here: use Bayesian optimization to quickly identify regions in the parameter space where Δ tends to be small.

General idea of Bayesian optimization









Vanilla implementation

- Assume (log) discrepancy follows a Gaussian process model.
- Assume a squared exponential covariance function $\mathrm{cov}(\Delta_{\boldsymbol{\theta}}, \Delta_{\boldsymbol{\theta}'}) = k(\boldsymbol{\theta}, \boldsymbol{\theta}')$,

$$k(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sigma_f^2 \exp\left(\sum_j \frac{1}{\lambda_j^2} (\theta_j - \theta_j')^2 \right). \tag{13}$$

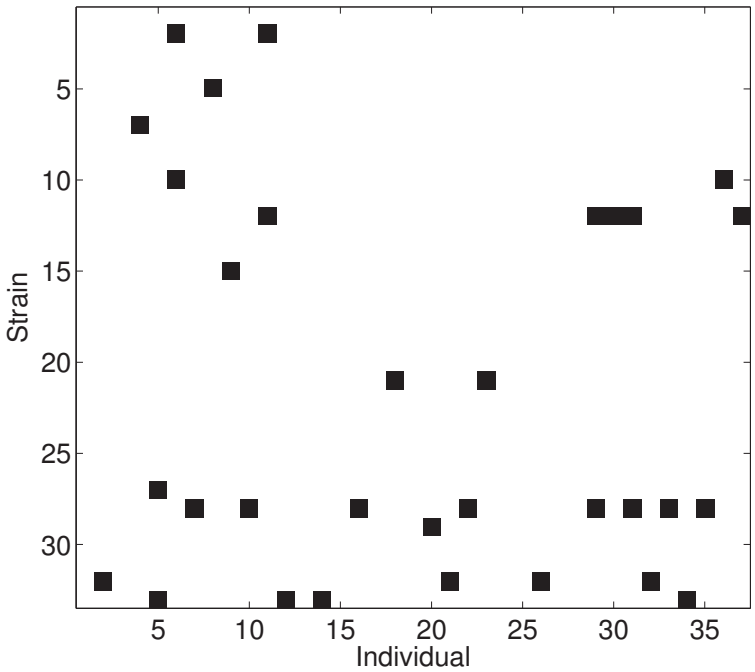
- Use lower confidence bound acquisition function (e.g. Cox and John, 1992; Srinivas et al, 2012)

$$\mathcal{A}_t(\boldsymbol{\theta}) = \underbrace{\mu_t(\boldsymbol{\theta})}_{\text{post mean}} - \sqrt{\underbrace{\eta_t^2}_{\text{weight}} \underbrace{v_t(\boldsymbol{\theta})}_{\text{post var}}} \tag{14}$$

- Possibly use stochastic acquisition rule: sample from Gaussian centered at $\mathrm{argmin}_{\boldsymbol{\theta}} \mathcal{A}_t(\boldsymbol{\theta})$ while respecting boundaries.

Application to epidemiology of infectious diseases

Data: Colonization states of sampled attendees of 29 child day care centers (DCCs) in Oslo greater area.



Example data from a DCC. Each square indicates an attendee colonized with a strain of the bacterium *Streptococcus pneumoniae*.

Model (Numminen et al, Biometrics, 2013)

- Latent continuous time Markov chain for the transmissions inside a DCC

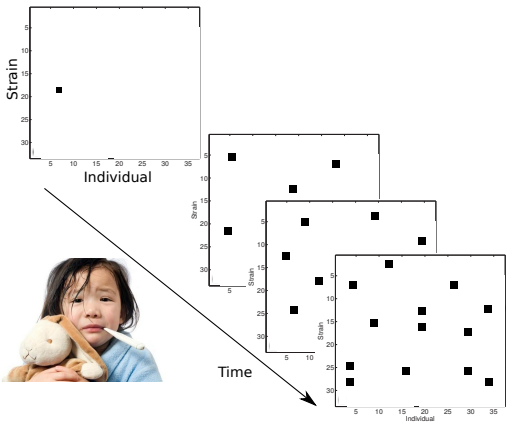
$$\Pr(I_{is}^{t+h} = 0 | I_{is}^{t} = 1) = h + o(h) \quad (15)$$

$$\Pr(I_{is}^{t+h} = 1 | I_{is'}^{t} = 0 \ \forall s') = R_s(t)h + o(h) \quad (16)$$

$$\Pr(I_{is}^{t+h} = 1 | I_{is}^{t} = 0, \ \exists s' : I_{is'}^{t} = 1) = \theta R_s(t)h + o(h) \quad (17)$$

$$R_s(t) = \beta E_s(t) + \Lambda P_s \quad (18)$$

- Observation model: Cross-sectional sampling at random time.



Parameters:

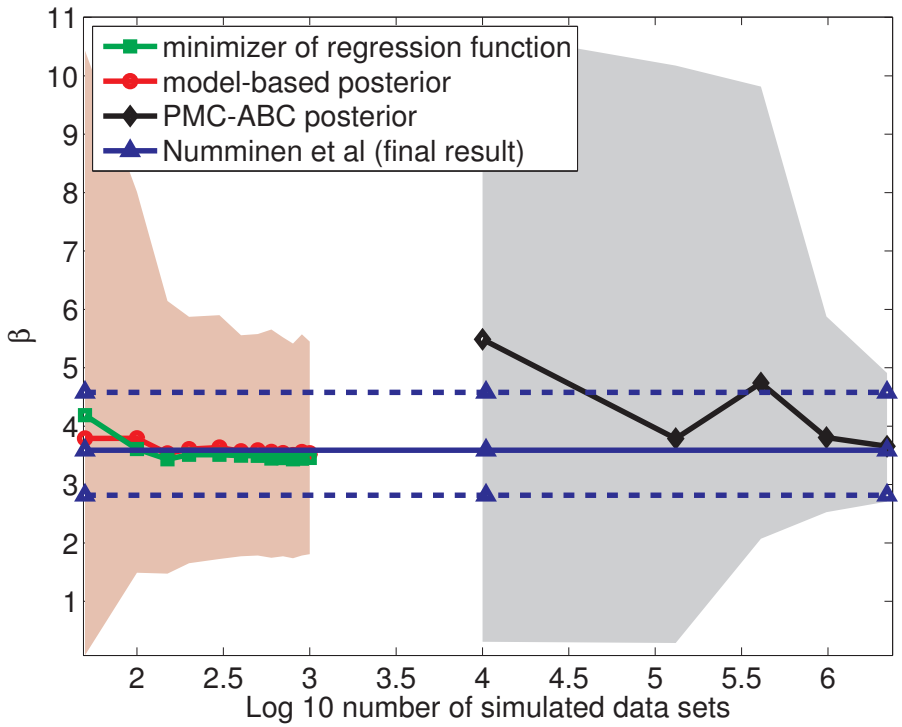
- β: rate of infections within a DCC
- Λ: rate of infections from outside
- θ: competition between the strains

Inference results

- Comparison of the model-based approach with a population Monte Carlo (PMC) ABC approach.
- Roughly equal results using 1000 times fewer simulations.
- The minimizer of the regression function under the model does not involve choosing a threshold ϵ.

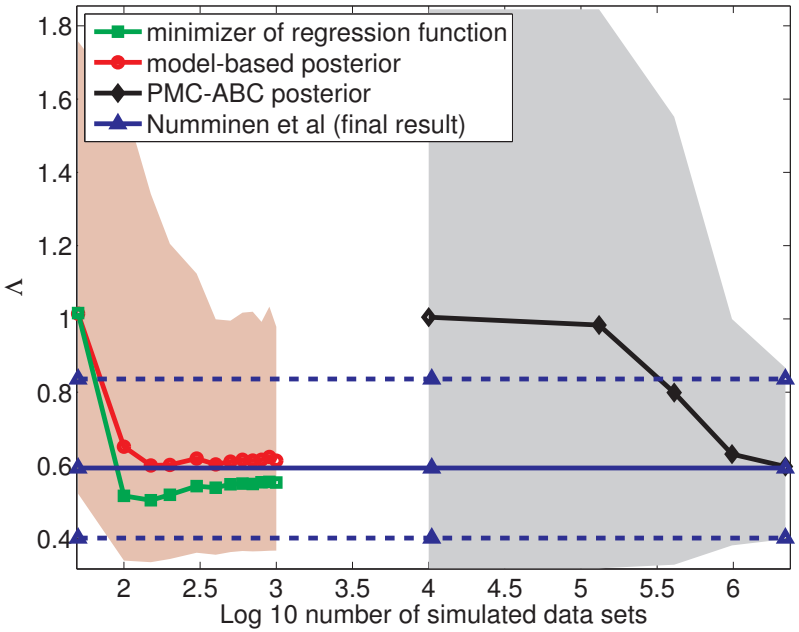
Posterior means: solid lines with markers,

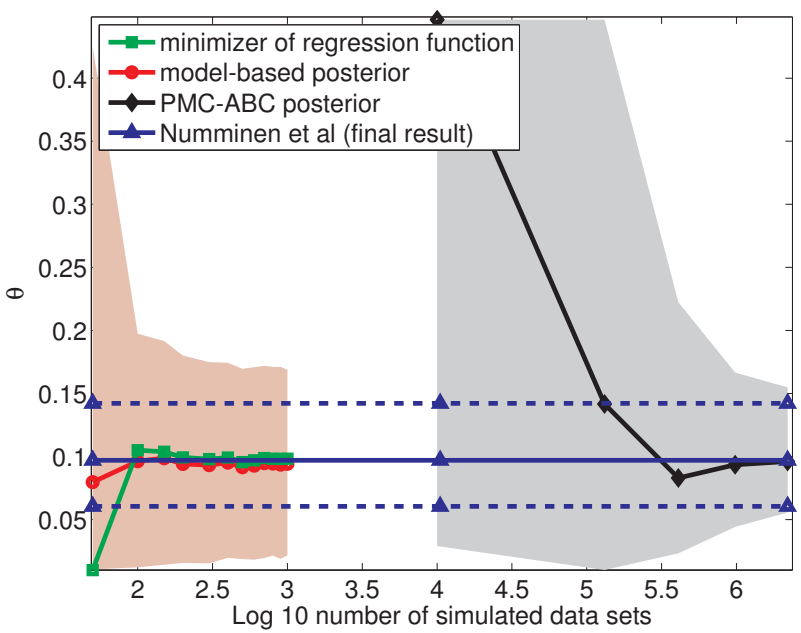
credibility intervals: shaded areas or dashed lines.



Application to epidemiology of infectious diseases

- Comparison of the model-based approach with a population Monte Carlo (PMC) ABC approach.





Posterior means are shown as solid lines with markers, credibility intervals as shaded areas or dashed lines.

Summary

- Introduction to inference with simulator-based models
- Introduction to approximate Bayesian computation
- Principle of ABC: Find parameter values which yield simulated data resembling the observed data.
- ABC is computationally very costly.
- Showed how Bayesian optimization can be used to increase the efficiency of the inference by several orders of magnitude.

Reference:
M.U. Gutmann and J. Corander
Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models, *Journal of Machine Learning Research*, in press, 2015
`http://arxiv.org/abs/1501.03291`