Improving power posterior estimation of statistical evidence

Nial Friel, Merrilee Hurn and Jason Wyse

Department of Mathematical Sciences, University of Bath, UK

10 June 2013

Bayesian Model Choice

- Possible models $m_1, \ldots, m_l$ for data y.
- Posterior distribution given data y model m_i is

$$p(\theta_i|y, m_i) = \frac{p(y|\theta_i, m_i)p(\theta_i|m_i)}{p(y|m_i)}$$

  where θ_i are the parameters for model m_i.
- The evidence/marginal likelihood for data y given model m_i is the normalising constant of the posterior distribution within model m_i,

$$p(y|m_i) = \int_{\theta_i} p(y|\theta_i, m_i)p(\theta_i|m_i)\, d\theta_i.$$

- The marginal likelihood is often then used to calculate Bayes factors to compare two competing models, m_i and m_j,

$$BF_{ij} = \frac{p(y|m_i)}{p(y|m_j)} = \frac{p(m_i|y)}{p(m_j|y)}\frac{p(m_j)}{p(m_i)}.$$

Estimating the Marginal Likelihood

Estimation of the evidence is non-trivial for most statistical models:

$$p(y|m_i) = \int_{\theta_i} p(y|\theta_i, m_i)p(\theta_i|m_i)\, d\theta_i.$$

- Laplace's method (Tierney and Kadane, 1986)
- Chib's method (Chib 1995)
- bridge sampling (Meng and Wong, 1996)
- annealed importance sampling (Neal 2001)
- nested sampling (Skilling 2006)
- power posteriors (Friel and Pettitt, 2008)
- stepping stone sampler (Xie, Lewis, Fan, Kuo and Chen, 2011)

A recent review Friel and Wyse (2012)

Power posteriors - theory

Dropping the explicit conditioning on model m_i for notational simplicity, define the power posterior at inverse temperature t by

$$\begin{aligned} p_t(\theta|y) &\propto p(y|\theta)^t p(\theta), \ t \in [0,1] \\ \text{with } z(y|t) &= \int_\theta p(y|\theta)^t p(\theta) d\theta. \end{aligned}$$

Two extremes:

- $t = 0$: $p_0(\theta|y)$ is the prior and $z(y|0) = 1$ by assumption
- $t = 1$: $p_1(\theta|y)$ is the posterior and $z(y|1)$ is the evidence

The power posterior estimator for the evidence uses identity

$$\begin{aligned} \int_0^1 \mathbf{E}_{\theta|y,t} \log(p(y|\theta)) dt &= [\ \log(z(y|t))\ ]_0^1 \\ &= \log(z(y|1)) - \log(1) \end{aligned}$$

which is the log of the desired marginal likelihood.

Power posteriors - implementation and costs

- Discretise the inverse temperatures $0 = t_0 < t_1 < \ldots < t_n = 1$ (Friel and Pettitt recommend powered fraction $t_i = (i/n)^5$)
- For each t_i in turn, sample from $p(\theta|y, t_i)$ to estimate $\mathbf{E}_{\theta|y,t_i} \log(p(y|\theta))$.
- Approximate the integral using the trapezoidal rule

$$\log p(y) \approx \sum_{i=1}^{n} (t_i - t_{i-1}) \frac{(\mathbf{E}_{\theta|y,t_{i-1}} \log(p(y|\theta)) + \mathbf{E}_{\theta|y,t_i} \log(p(y|\theta)))}{2}$$

- Discretising t introduces extra approximation into the method
- The cost of estimating $\log p(y)$ is all in estimating the $\mathbf{E}_{\theta|y,t_i} \log(p(y|\theta))$, so an extra n times the cost of just sampling from the posterior
- How do we minimise the costs by minimising the error for a fixed number of t_i?

Power posteriors - implementation and costs

- Discretise the inverse temperatures $0 = t_0 < t_1 < \ldots < t_n = 1$ (Friel and Pettitt recommend powered fraction $t_i = (i/n)^5$)
- For each t_i in turn, sample from $p(\theta|y, t_i)$ to estimate $\mathbf{E}_{\theta|y,t_i} \log(p(y|\theta))$.
- Approximate the integral using the trapezoidal rule

$$\log p(y) \approx \sum_{i=1}^{n} (t_i - t_{i-1}) \frac{(\mathbf{E}_{\theta|y,t_{i-1}} \log(p(y|\theta)) + \mathbf{E}_{\theta|y,t_i} \log(p(y|\theta)))}{2}$$

- Discretising t introduces extra approximation into the method
- The cost of estimating $\log p(y)$ is all in estimating the $\mathbf{E}_{\theta|y,t_i} \log(p(y|\theta))$, so an extra n times the cost of just sampling from the posterior
- How do we minimise the costs by minimising the error for a fixed number of t_i?

Power posteriors - implementation and costs

- Discretise the inverse temperatures $0 = t_0 < t_1 < \ldots < t_n = 1$ (Friel and Pettitt recommend powered fraction $t_i = (i/n)^5$)
- For each t_i in turn, sample from $p(\theta|y, t_i)$ to estimate $\mathbf{E}_{\theta|y,t_i} \log(p(y|\theta))$.
- Approximate the integral using the trapezoidal rule

$$\log p(y) \approx \sum_{i=1}^{n} (t_i - t_{i-1}) \frac{(\mathbf{E}_{\theta|y,t_{i-1}} \log(p(y|\theta)) + \mathbf{E}_{\theta|y,t_i} \log(p(y|\theta)))}{2}$$

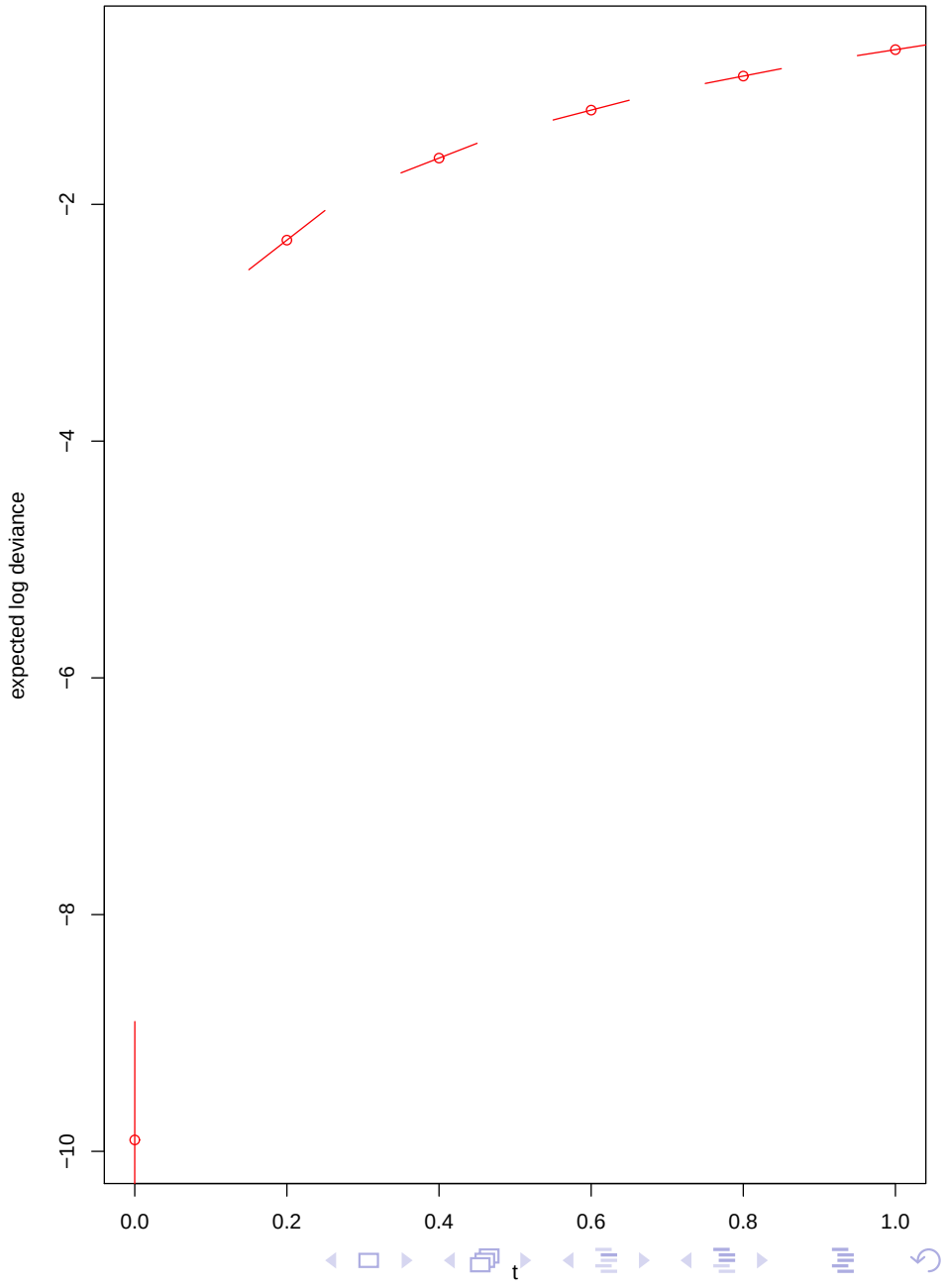
- Discretising t introduces extra approximation into the method
- The cost of estimating $\log p(y)$ is all in estimating the $\mathbf{E}_{\theta|y,t_i} \log(p(y|\theta))$, so an extra n times the cost of just sampling from the posterior
- How do we minimise the costs by minimising the error for a fixed number of t_i?

Key new observation

We can show that the gradient of the expected curve $\mathbf{E}_{\theta|y,t} \log(p(y|\theta))$ we want to integrate is given by the variance $\mathbf{Var}_{\theta|y,t} \log(p(y|\theta))$.

The same MCMC samples at fixed values of t lead to estimates of the mean and variance of $\log(p(y|\theta))$.

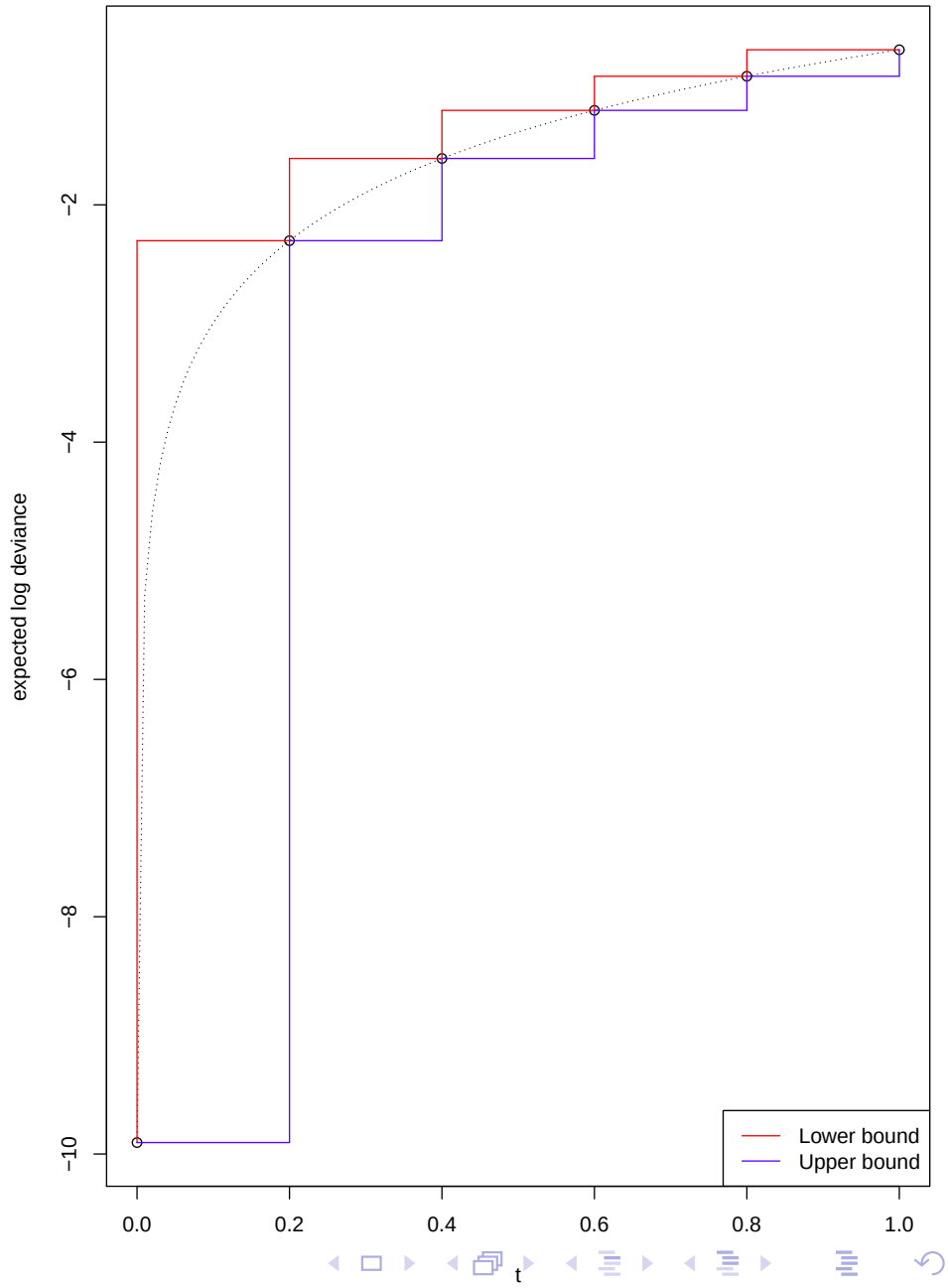
We can we use this information in two ways.



1. Choosing the interior t_i points

Rationale Since we now know that the curve is increasing, we have an upper and lower bound on the integral by considering the two bounding step functions. Aim to place the t_i to minimise the area between the two.

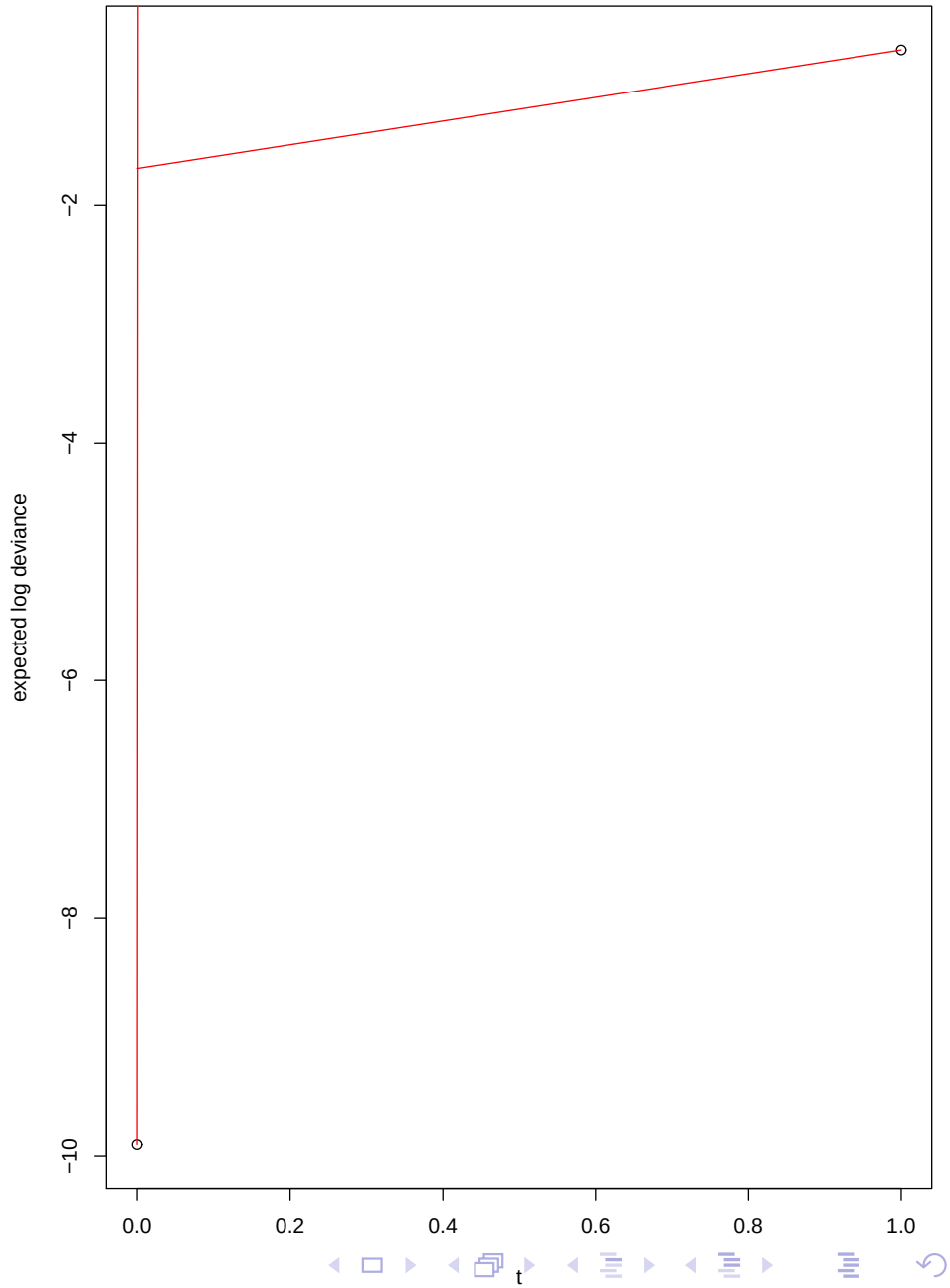
(More formally, minimising this area minimises the Kullback-Leibler distance between the true curve and the approximation.)



Choosing the interior t_i points - iterative approximation

Start with $t_0 = 0$ and $t_n = 1$. All we have is estimates of the function and its derivative at these two points.

Site the next t at the point where the two tangents meet.

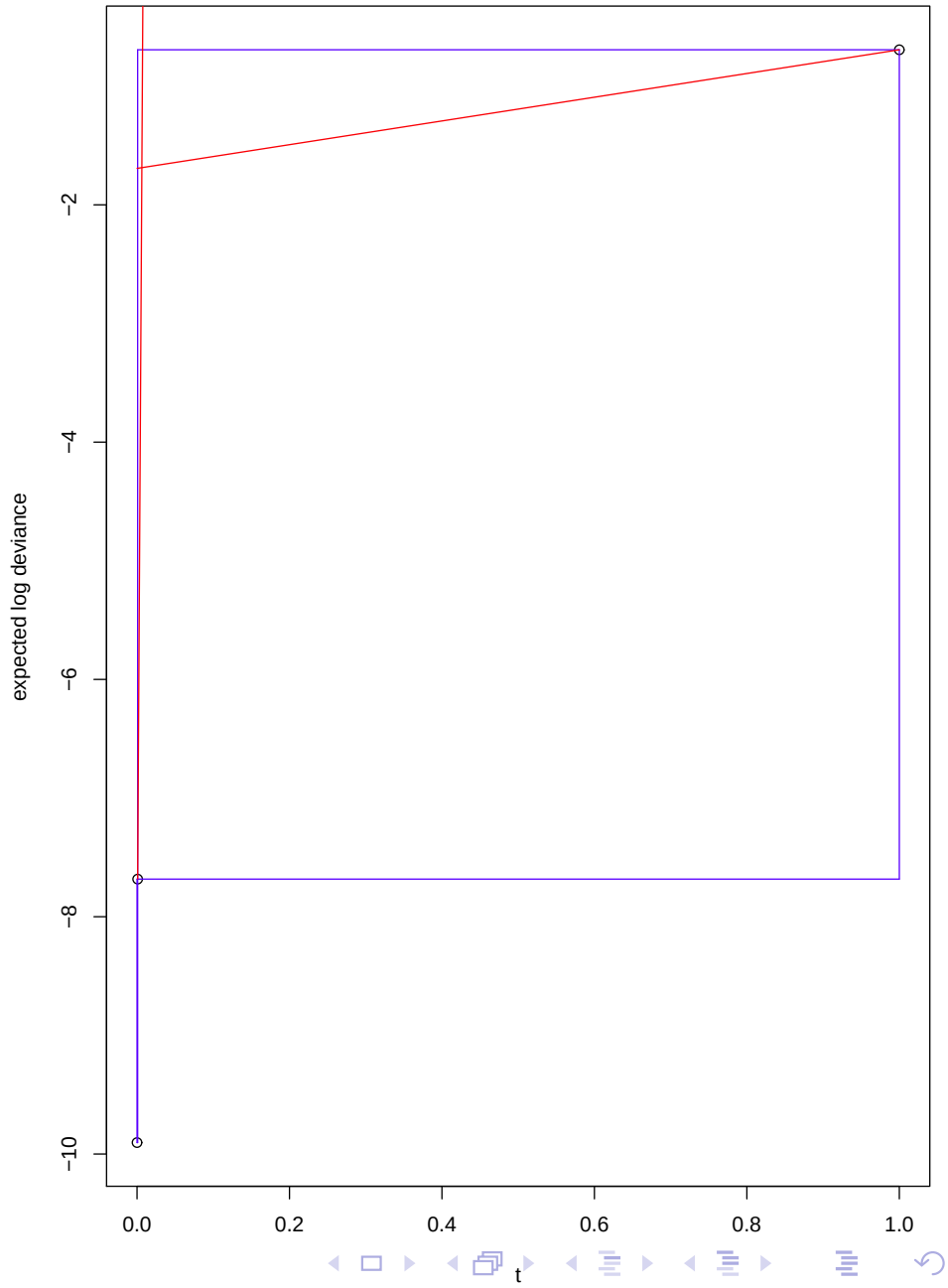


Choosing the interior t_i points - iterative approximation

Start with $t_0 = 0$ and $t_n = 1$. All we have is estimates of the function and its derivative at these two points.

Site the next t at the point where the two tangents meet.

We now have three evaluations and two rectangular contributions to the difference area. Subdivide the larger area using the intersection of tangents.

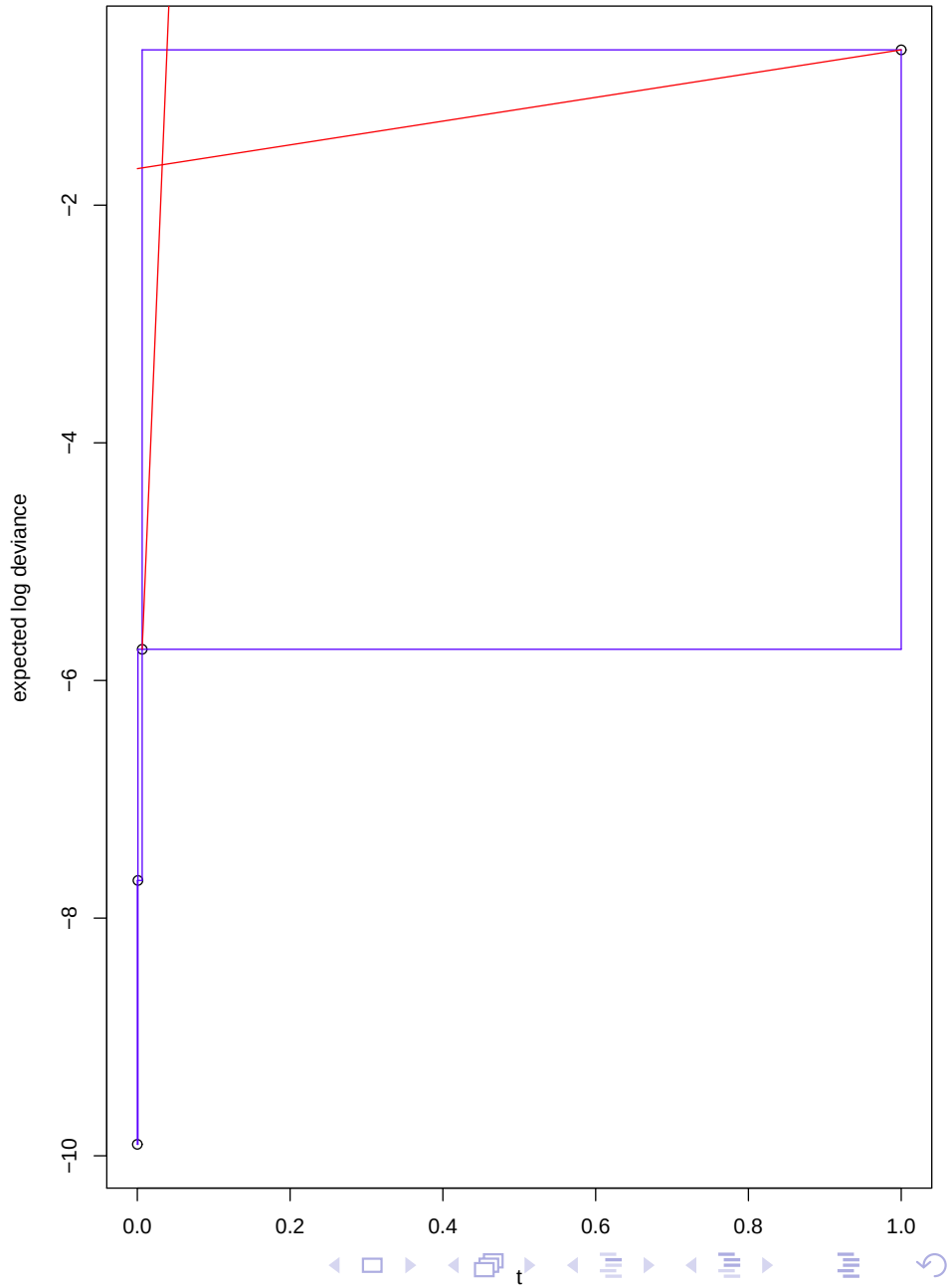


Choosing the interior t_i points - iterative approximation

Start with $t_0 = 0$ and $t_n = 1$. All we have is estimates of the function and its derivative at these two points.

Site the next t at the point where the two tangents meet.

We now have four evaluations and three rectangular contributions to the difference area. Subdivide the larger area using the intersection of tangents.

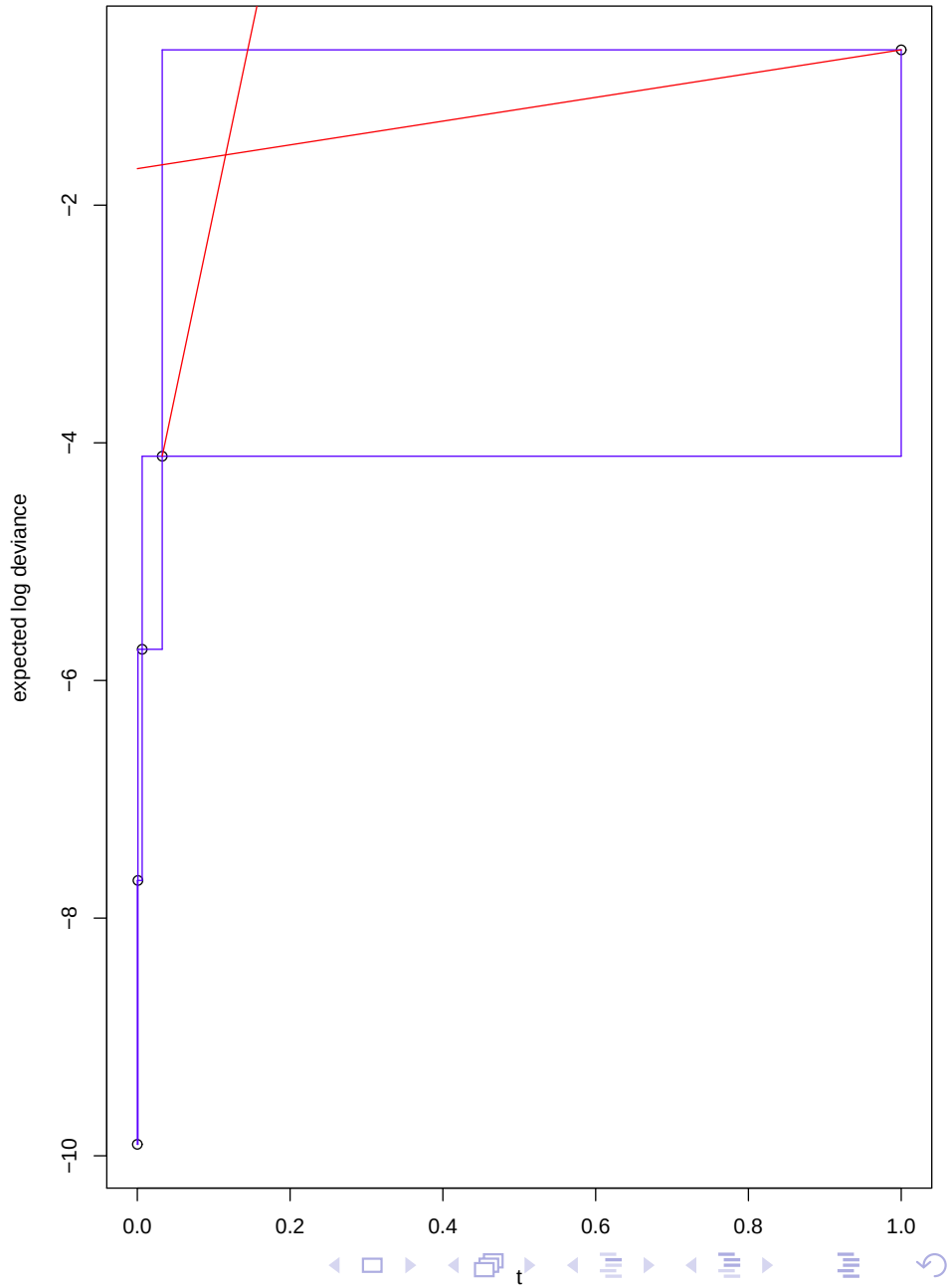


Choosing the interior t_i points - iterative approximation

Start with $t_0 = 0$ and $t_n = 1$. All we have is estimates of the function and its derivative at these two points.

Site the next t at the point where the two tangents meet.

We now have five evaluations and four rectangular contributions to the difference area. Subdivide the larger area using the intersection of tangents.

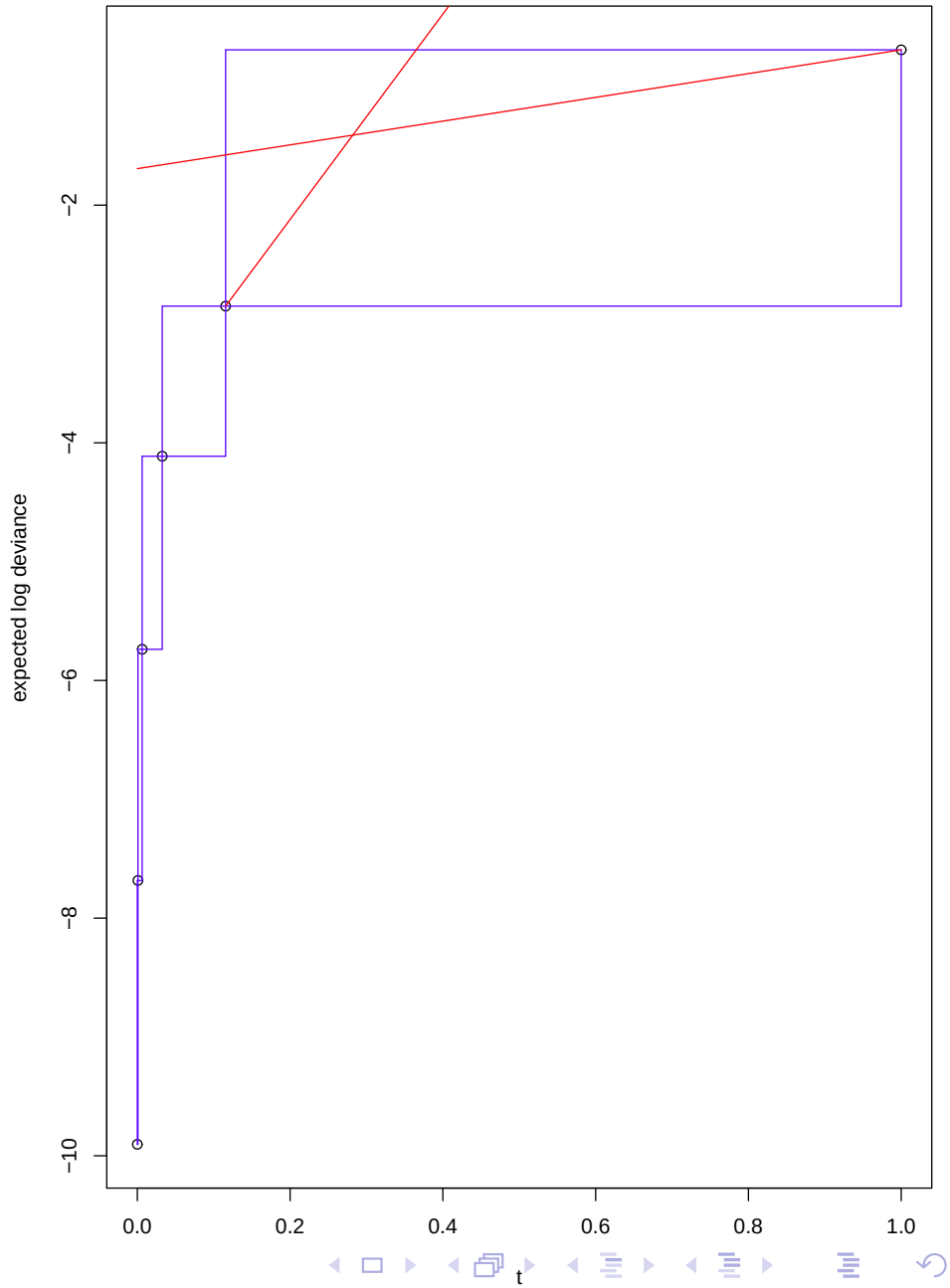


Choosing the interior t_i points - iterative approximation

Start with $t_0 = 0$ and $t_n = 1$. All we have is estimates of the function and its derivative at these two points.

Site the next t at the point where the two tangents meet.

We now have six evaluations and five rectangular contributions to the difference area. Subdivide the larger area using the intersection of tangents.

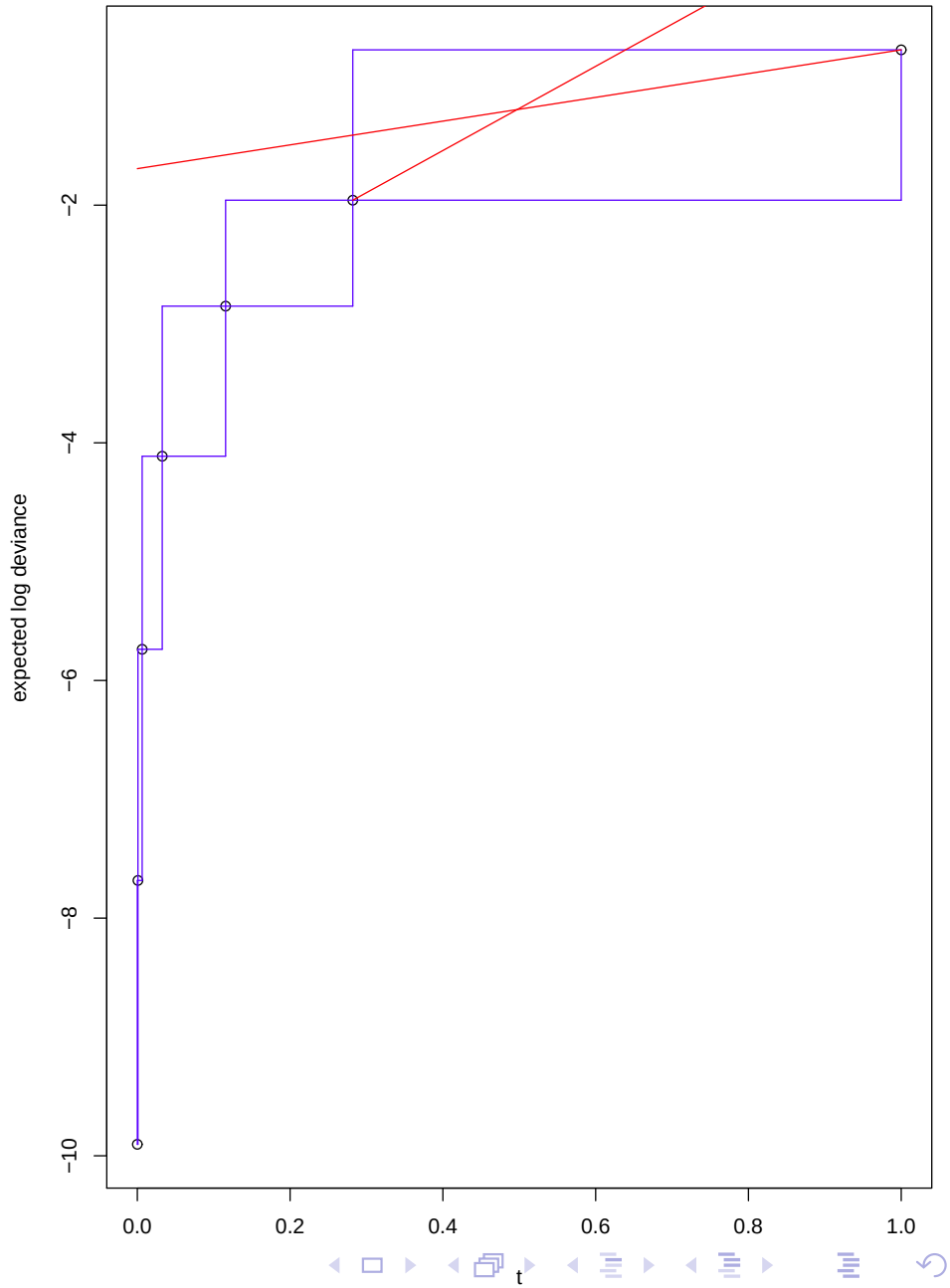


Choosing the interior t_i points - iterative approximation

Start with $t_0 = 0$ and $t_n = 1$. All we have is estimates of the function and its derivative at these two points.

Site the next t at the point where the two tangents meet.

We now have seven evaluations and six rectangular contributions to the difference area. Subdivide the larger area using the intersection of tangents.

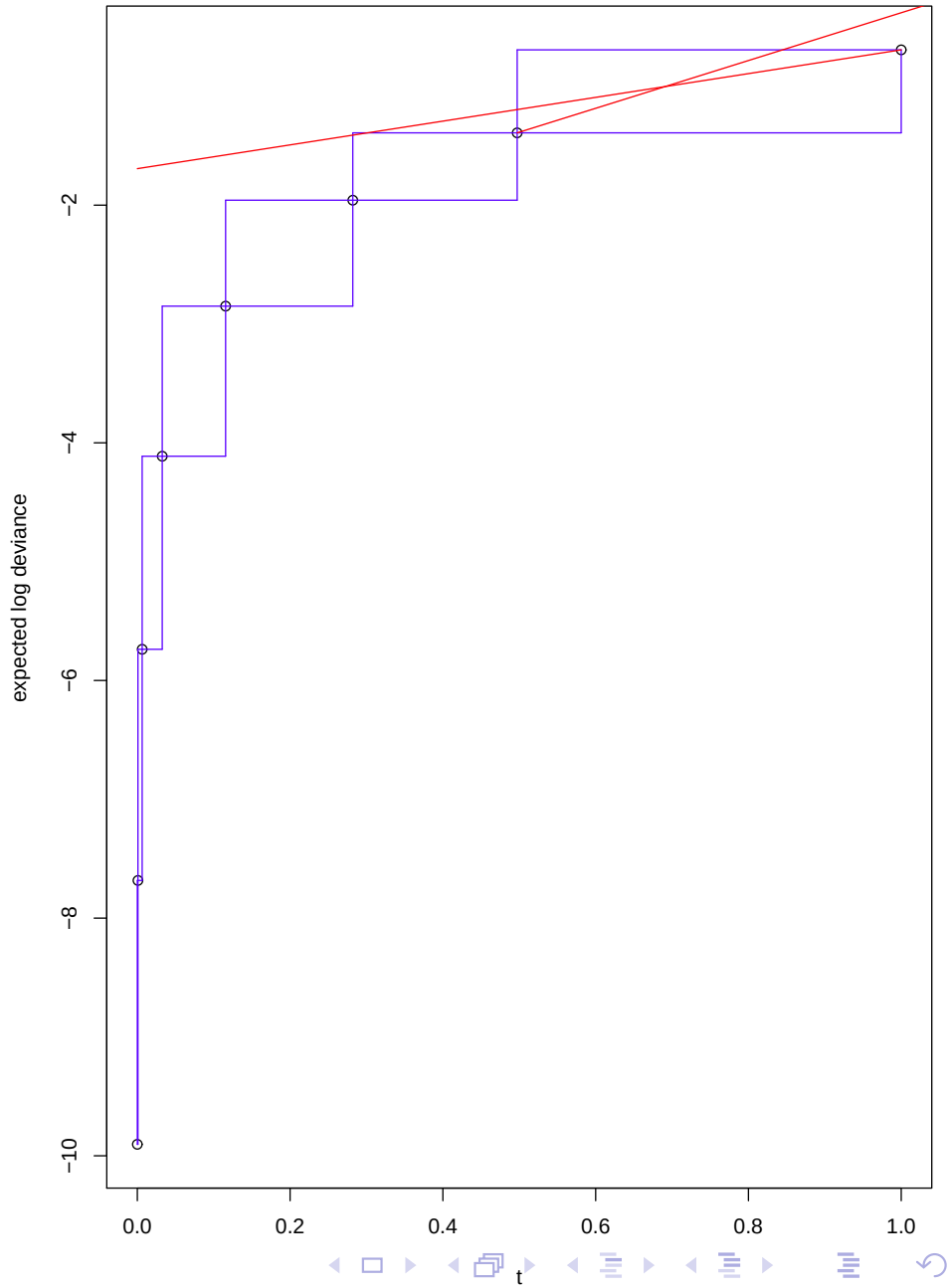


Choosing the interior t_i points - iterative approximation

Start with $t_0 = 0$ and $t_n = 1$. All we have is estimates of the function and its derivative at these two points.

Site the next t at the point where the two tangents meet.

We now have eight evaluations and seven rectangular contributions to the difference area. Subdivide the larger area using the intersection of tangents.



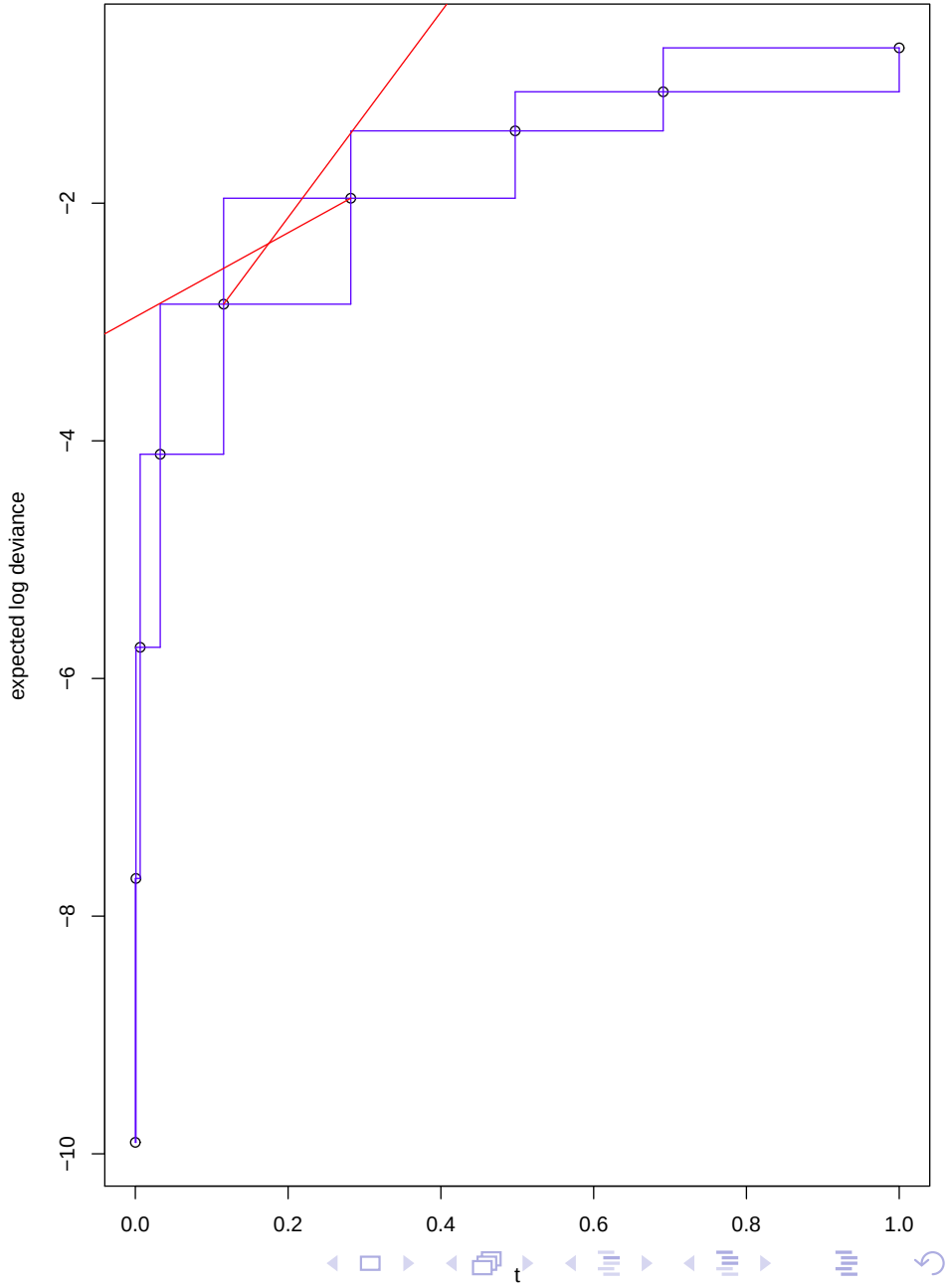
Choosing the interior t_i points - iterative approximation

Start with $t_0 = 0$ and $t_n = 1$. All we have is estimates of the function and its derivative at these two points.

Site the next t at the point where the two tangents meet.

We now have nine evaluations and eight rectangular contributions to the difference area. Subdivide the larger area using the intersection of tangents.

And so on...



2. The modified trapezium rule

We can also use the gradient information to improve the numerical integration:

When integrating a function f between points a and b

$$\int_a^b f(x)dx = (b-a)\left[\frac{f(b)+f(a)}{2}\right] - \frac{(b-a)^3}{12}f''(c)$$

where c is some point in $[a, b]$. The first term is the usual trapezium rule and the second can be approximated using

$$f''(c) \approx \frac{f'(b)-f'(a)}{b-a}$$

$$\text{so } \int_a^b f(x)dx \approx (b-a)\left[\frac{f(b)+f(a)}{2}\right] - \frac{(b-a)^2}{12}\left[f'(b)-f'(a)\right]$$

and, unusually, we have gradient information cheaply available via the variance terms.

Pima Indian Example

Data: diabetes incidence and possible disease indicators for $n = 532$ Pima Indian women aged over 20. Seven possible disease indicators: number of pregnancies (`NP`), plasma glucose concentration (`PGC`), diastolic blood pressure (`BP`), triceps skin fold thickness (`TST`), body mass index (`BMI`), diabetes pedigree function (`DP`) and age (`AGE`), (all covariates standardised).

Smith, Everhart, Dickson, Knowler and Johannes (1988)

Likelihood observed incidence $y = (y_1, \ldots, y_n)$ with d covariates

$$p(y|\theta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}, \qquad \log\left(\frac{p_i}{1-p_i}\right) = \theta^T x_i$$

- p_i the probability of incidence for person i
- covariates $x_i = (1, x_{i1}, \ldots, x_{id})^T$
- parameters $\theta = (\theta_0, \theta_1, \ldots, \theta_d)^T$

Prior for θ, independent Gaussian with mean zero and non-informative precision of $\tau = 0.01$

$$p(\theta) \propto \exp\left\{-\frac{\tau}{2}\theta^T\theta\right\}.$$

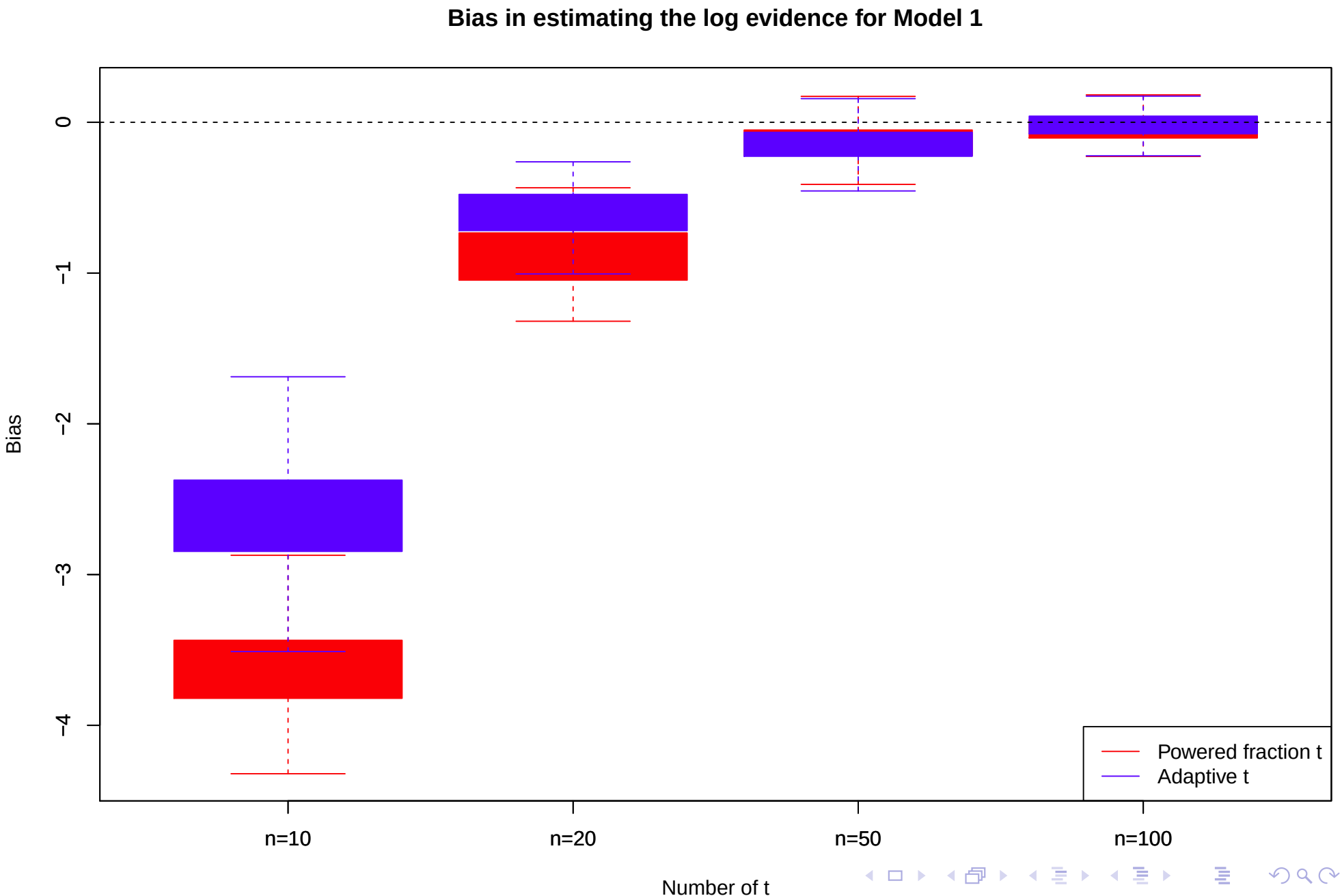
Friel and Wyse (2012)

Model choice Friel and Wyse's long RJMCMC run identifies the two highest posterior probability models as:

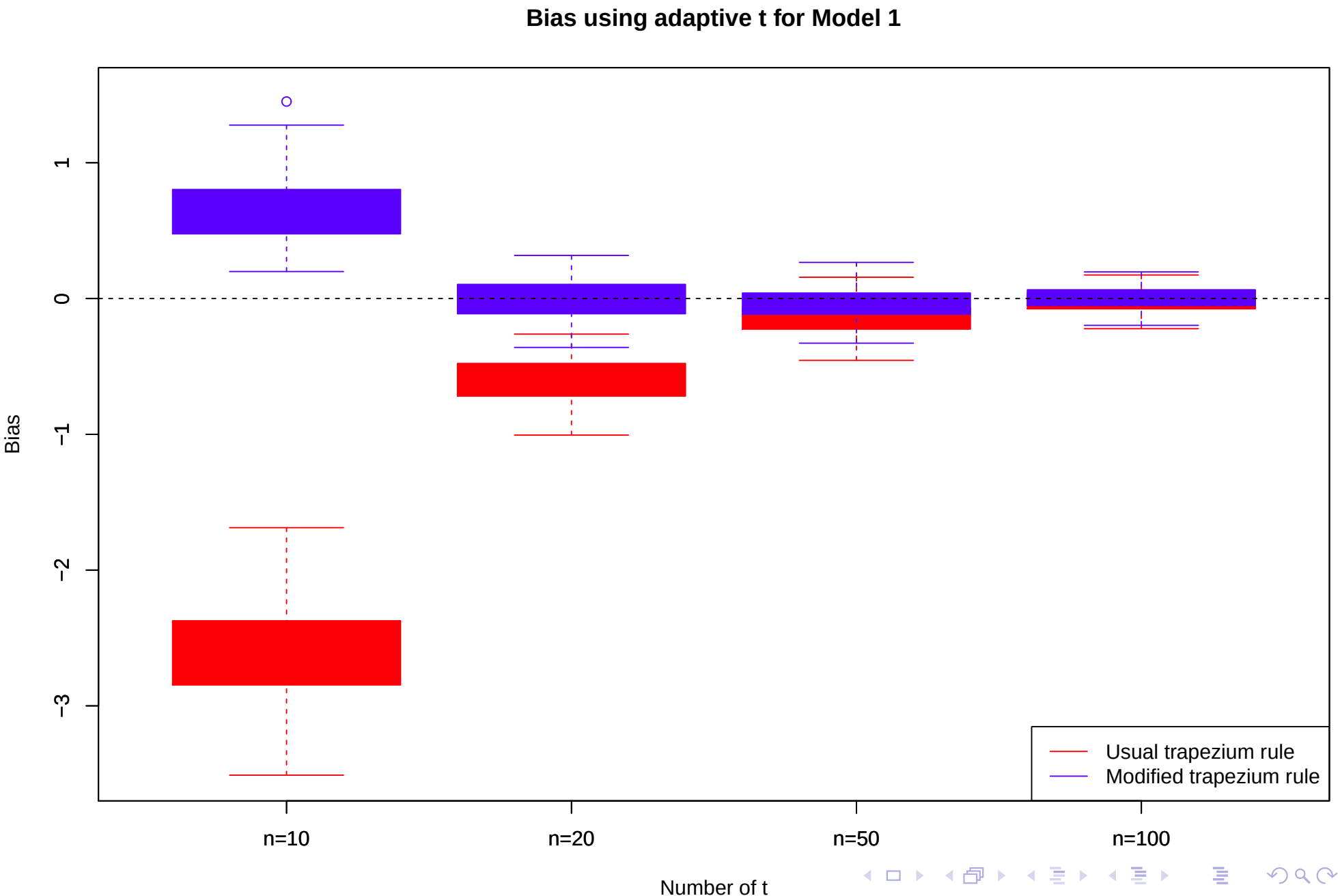
Model 1: `logit(p)= 1+NP+PGC+BMI+DP`
Model 2: `logit(p)= 1+NP+PGC+BMI+DP+AGE`

Experiments 100 estimates of the evidence for each model using either $n = 10, 20, 50$ or 100 with 10000 MCMC iterations at each t_i. Benchmark evidence from a very long run ($n = 2000$, 20000 MCMC iterations).

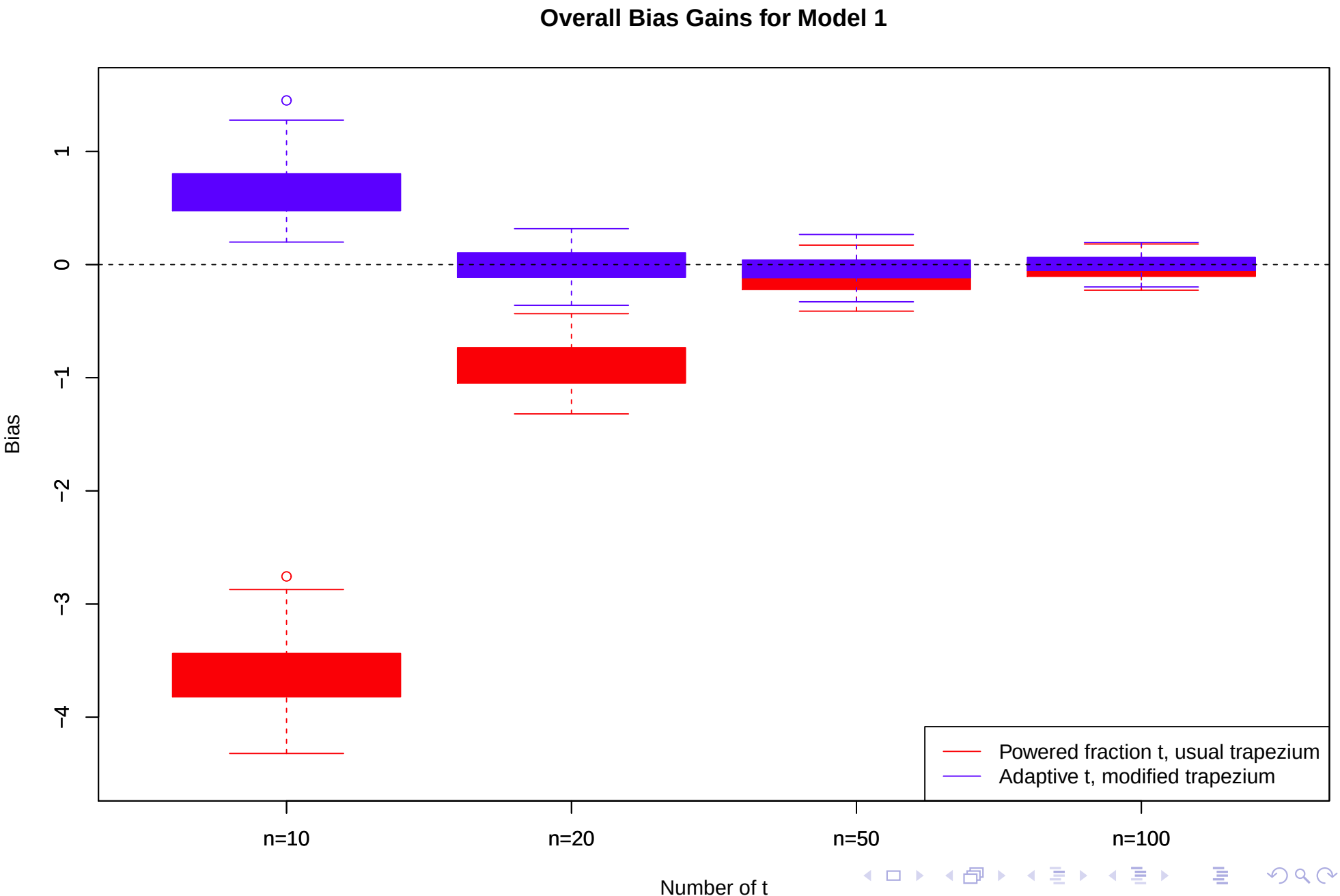
1. Effect of tuning the t_i



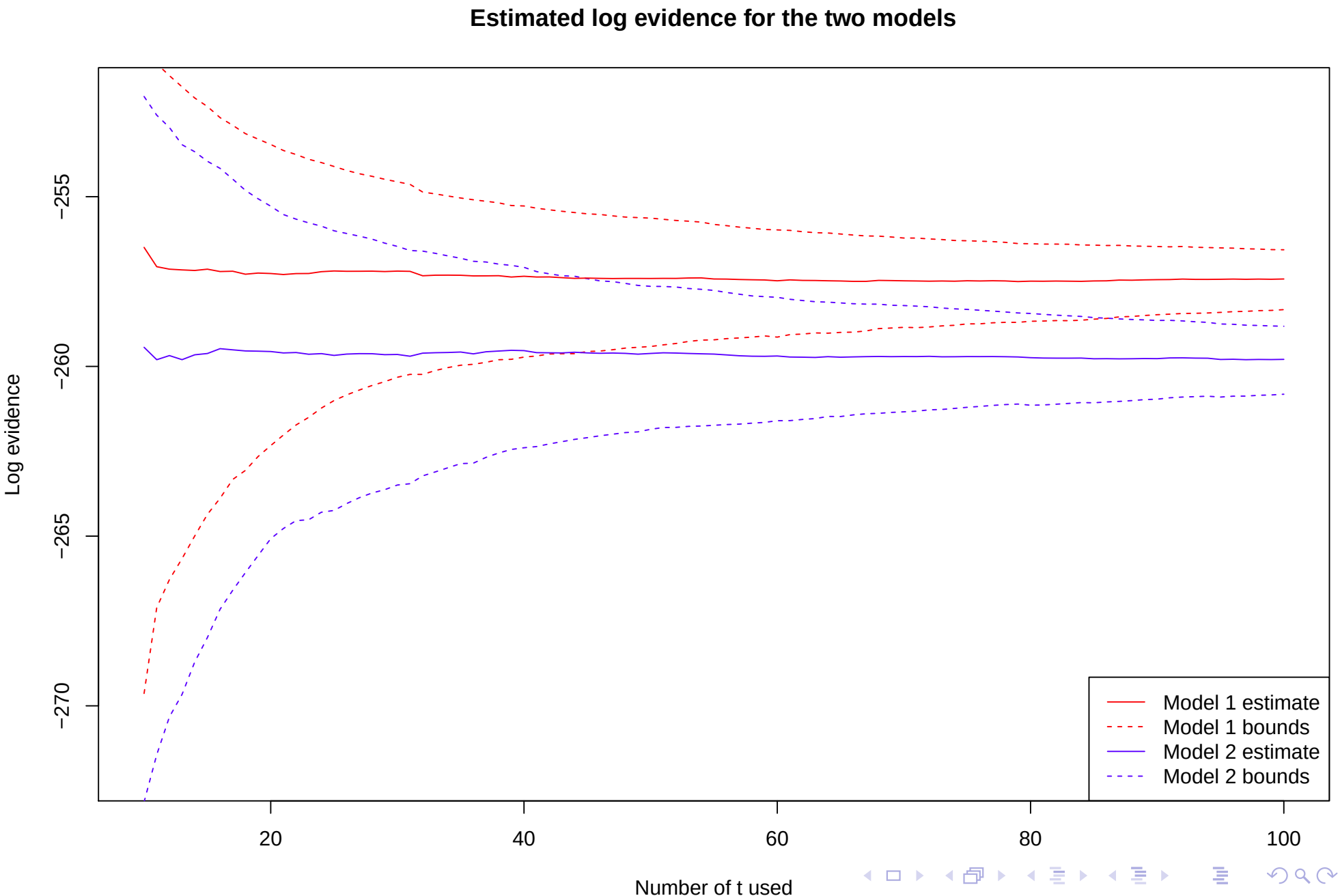
2. Effect of improving the trapezium rule



Combined effect of the two modifications



Estimated log marginal likelihoods + discretisation bounds



Discussion

- Estimating the log marginal likelihood via Power Posteriors is relatively straight-forward but computationally costly.
- To minimise the cost, we want to use as few t values as possible.
- How to choose those t? Simple algorithm very cheaply approximates minimising the gap between an upper and a lower estimated discretisation bound on the log evidence.
- We can also very cheaply improve on the trapezium rule for the numerical integration.
- The adaptive t placement allows us to use just as many t as it takes for the estimated discretisation bounds not to overlap when comparing a pair of models.