Comment on "Bayesian estimates of free energies from nonequilibrium work data in the presence of instrument noise"

John D. Chodera,[1,*] David D. L. Minh,[2,†] and Michael R. Shirts[3,‡]

[1]*Department of Chemistry, Stanford University, Stanford, CA 94305, USA*
[2]*Laboratory of Chemical Physics, NIDDK, National Institutes of Health, Bethesda, Maryland 20892, USA*
[3]*Department of Chemical Engineering, University of Virginia, Charlottesville, VA 22904, USA*

(Dated: September 10, 2008)

The development of methods to estimate free energy differences from nonequilibrium work measurements remains an active area of research. Recently, Maragakis et al. presented a new Bayesian scheme for the analysis of single-molecule pulling data in the presence of instrument noise [1]. In principle, Bayesian methods provide a powerful tool for assessing the confidence in an estimate, thereby guiding experimentalists in deciding how much data need be collected. However, we believe that a word of caution is necessary, as this method can produce misleading estimates of the uncertainty when improperly applied.

Our concern lies with the posterior distribution for the free energy estimate ΔF_Λ given measured work values W and protocols Λ (forward or reverse measurements) in the absence of measurement error, which forms the core of proposed scheme [1]:

$$\begin{aligned} P(\Delta F_\Lambda | W_n, \Lambda) &\propto P(W, \Lambda | \Delta F_\Lambda)\, P(\Delta F_\Lambda) \\ &\propto \prod_{n=1}^{N} f(\beta W_n - \beta \Delta F_{\Lambda_n} + M_{\Lambda_n}) \quad (1) \end{aligned}$$

In this expression, β is the inverse temperature, $f(x) = (1+e^{-x})^{-1}$ is the standard logistic function, and $M_\Lambda = \ln(N_\Lambda / N_{\tilde{\Lambda}})$ is the log ratio of the number of forward and reverse work measurements. The method can be thought of as a Bayesian extension to the Bennett acceptance ratio (BAR) [2], and will henceforth be referred to as BBAR for brevity.

Though derivation of the Bayesian [1] and maximum likelihood [3, 4] variants differ slightly, both contain an essential step such as the one used to compute $P(W, \Lambda | \Delta F)$ from known quantities:

$$\frac{P(+W, \Lambda | \Delta F_\Lambda)}{P(-W, \tilde{\Lambda} | \Delta F_{\tilde{\Lambda}})} = \frac{P(+W | \Lambda, \Delta F_\Lambda)}{P(-W | \tilde{\Lambda}, \Delta F_{\tilde{\Lambda}})} \cdot \frac{P(\Lambda | \Delta F_\Lambda)}{P(\tilde{\Lambda} | \Delta F_{\tilde{\Lambda}})} \quad (2)$$

The Crooks fluctuation theorem [5] can be used to evaluate the first ratio, and the second ratio — the relative probability of forward measurements Λ over reverse measurements $\tilde{\Lambda}$ — is evaluated as $N_\Lambda / N_{\tilde{\Lambda}}$. Together with the assumption that $P(+W | \Lambda, \Delta F_\Lambda) + P(-W | \tilde{\Lambda}, \Delta F_{\tilde{\Lambda}})$ is constant — a necessary assumption in order to provide a prior on the work distribution — Eq. 1 is obtained.

We point out two features of the Bayesian posterior (Eq. 1). First, if either N_Λ or $N_{\tilde{\Lambda}}$ are zero — only work measurements along one protocol (but not its reverse) are available — the posterior is unnormalizable, and no estimate is produced, only a soft bound [1]. This need not be the case: A recent multistate extension of BAR [6] provides both an estimator and an estimate of the statistical uncertainty (through an asymptotic variance estimate) in this case as well. It may be possible to construct a modified posterior that also furnishes a well-defined estimate for this situation.

The second issue provides a caution for both BBAR and methods based on the likelihood approach lies in the separation of work measurements from their protocol to produce the ratio $P(\Lambda | \Delta F_\Lambda) / P(\tilde{\Lambda} | \Delta F_{\tilde{\Lambda}})$, evaluated as $N_\Lambda / N_{\tilde{\Lambda}}$. Most experiments are not conducted with a fixed *probability* of recording forward and reverse measurements, which would allow the actual number of samples in each direction to vary, but a fixed *number* of measurements in each direction. Because of this difference, Shirts et al. [3] were forced to employ a correction term [7] that reduced the variance by $-(N_\Lambda^{-1} + N_{\tilde{\Lambda}}^{-1})$ in order to produce an estimate of the uncertainty that agreed with the original derivation by Bennett [2].

Only much later was it appreciated that Geyer had earlier proposed a similar likelihood approach, framing the problem as "reverse logistic regression" [8]. Geyer's method suffered from the same problem, in that the uncorrected asymptotic variance of the likelihood function gives an overestimate of the statistical uncertainty, for reasons elegantly explained by Kong et al. [9]. While the estimators are asymptotically unbiased for both types of experiments (fixed-number and fixed-probability), there is a clear difference in the asymptotic variance of the estimator between these two types of experiments [9]. It is reasonable to ask, therefore, what practical impact this difference has on BBAR, and under what conditions we

*Electronic address: jchodera@stanford.edu; Author to whom correspondence should be addressed.
†Electronic address: daveminh@gmail.com
‡Electronic address: mrshirts@gmail.com

[1] There is also the problem that M_Λ may be undefined, but Maragakis et al. propose the modified definition $M_\Lambda = \ln(N_\Lambda + 1)/(N_{\tilde{\Lambda}} + 1)$ as a way to avoid this [1].

might find the posterior to be unreliable.

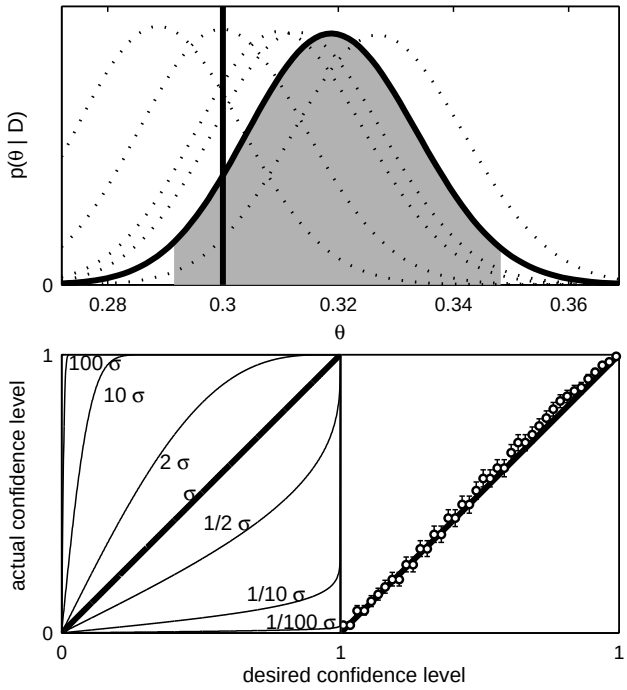

FIG. 1: **Testing the posterior for inference of a biased coin flip experiment.** *Top:* Posterior distribution for inferring the probability of heads, θ, for a biased coin from an sequence of $N = 1000$ coin flips (dark line) with 95% symmetric confidence interval about the mean (shaded area). The true probability of heads is 0.3 (vertical thick line). Posteriors from five different experiments are shown as dotted lines. *Bottom left:* Desired and actual confidence levels for an idealized normal posterior distribution that either overestimates (upper left curves) or underestimates (bottom right curves) the true posterior variance by different degrees. *Bottom right:* Desired and actual confidence levels for the Binomial-Beta posterior for the coin flip problem depicted in upper panel. Error bars show 95% confidence intervals estimates from 1000 independent experimental trials. For inference, we use a likelihood function such that the observed number of heads is $N_H|\theta \sim \mathrm{Binomial}(N_H, N, \theta)$ and conjugate Jeffreys prior [10, 11] $\theta \sim \mathrm{Beta}(1/2, 1/2)$ which produces posterior $\theta|N_H \sim \mathrm{Beta}(N_H + 1/2, N_T + 1/2)$ along with constraint $N_H + N_T = N$.

How can we test a Bayesian posterior distribution? One of the more powerful features of a Bayesian model is its ability to provide confidence intervals that correctly reflect the level of certainty that the true value will lie within in. For example, if the experiment were to be repeated many times, the true value of the parameter being estimated should fall within the confidence interval for a 95% confidence level 95% of the time. As an illustrative example, consider a biased coin where the probability of turning heads is θ. From an observed sample of N coin flips, we can estimate θ using a Binomial model for the number of coin flips that turn up heads and a conjugate Beta Jeffreys prior [10, 11]. Each time we run experiment, we get a different posterior estimate for θ, and a different confidence interval (Figure 1, top). If we run many trials and record what fraction of the time the true (unknown) value of θ falls within the confidence interval estimated from that trial, we can see if our model is correct. If correct, the observed confidence level should match the desired confidence level (Figure 1, bottom right). Deviation from parity means that the posterior is either two broad or too narrow, and that the statistical uncertainty is being either over- or underestimated (Figure 1, bottom left).

We tested the BBAR posterior in the same way on a two-state system with Gaussian work distributions (Figure 2). We compared the behavior with the asymptotic variance estimates of standard BAR [2, 3, 6], which effectively assumes a normal posterior, either with the correction for fixed-number (BAR-FN) or without the correction, which would be appropriate for fixed-probability experiments (BAR-FP). We find that, under a variety of conditions, BBAR significantly overestimates the uncertainty in experiments where a fixed number of forward and reverse work measurements are made, while BAR-FN accurately estimates the confidence intervals (Figure 2, left column). This behavior persists for both small (top) and large (bottom) numbers of samples. If instead the experiment is run where forward or reverse measurements are selected according to a fixed probability, we see that BBAR correctly estimates the true confidence intervals, as does BAR-FP (Figure 2, right column).

In light of these results, we make the following recommendations, until such time as additional theory work can produce a variant of BBAR appropriate for the fixed-number experiment. (1) Practitioners of simulation or experiment should incorporate a fixed probability of forward or reverse measurement into their experimental design, or post-process the data by picking a subset of data from the pool of measurements with fixed probability in estimating the error in the estimate (which can still be produced from the entire dataset without concern). (2) Theorists wishing to produce estimators based on BBAR should exercise caution that their experimental design corresponds to the fixed-probability, rather than the fixed-number, situation, unless a large upper bound on the uncertainty can be tolerated.

Acknowledgments

The authors thank Gavin E. Crooks (Lawrence Berkeley Natl. Lab.) and Paul Maragakis (D. E. Shaw Research) for enlightening discussions of their work. JDC acknowledges support through an NSF grant for Cyberinfrastructure (NSF CHE-0535616) through Vijay S. Pande.

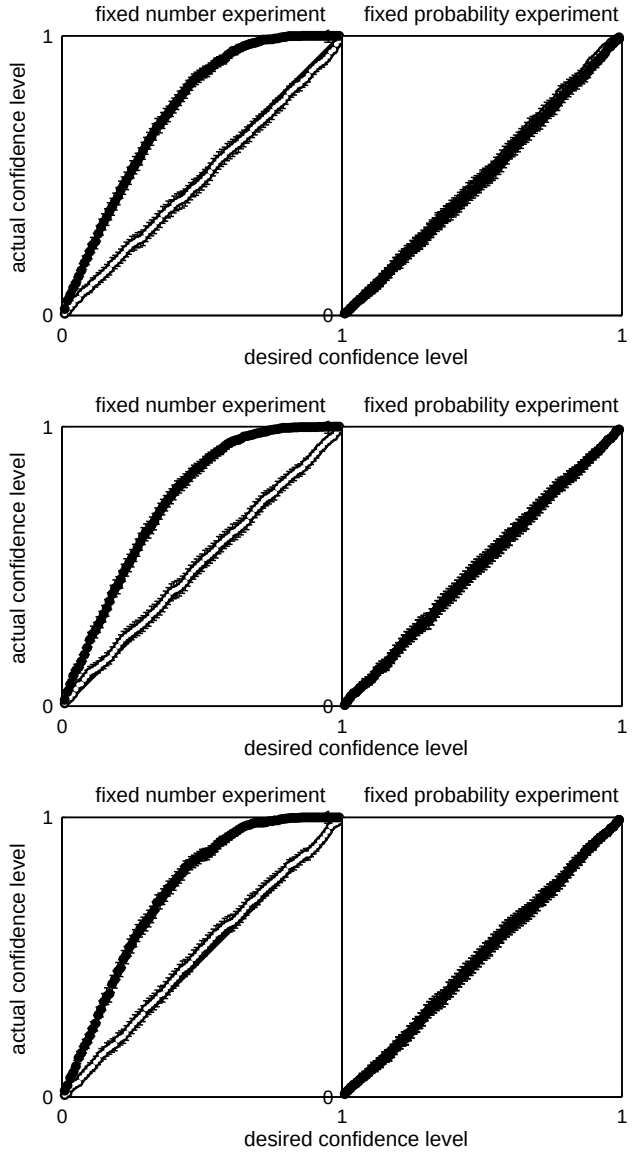

FIG. 2: **Testing the posterior for Bayesian BAR.** *Left:* Actual confidence levels for BBAR (filled circles) and BAR-FN (solid circles) evaluated for replications of an experiment containing a *fixed number* of 10 (top), 100 (middle), or 1000 (bottom) work measurements in each direction. *Right:* Actual confidence levels for *fixed probability* case comparing BBAR (filled circles) and BAR-FP (solid circles). Error bars show 95% confidence intervals estimated from 1000 independent experimental trials. Work measurements were obtained from a model with Gaussian forward work distribution of unit mean and variance; the reverse work distribution is fully determined from the forward distribution by the Crooks fluctuation theorem [5]. Tests with a system of displaced one-dimensional harmonic oscillators showed similar behavior.

[1] P. Maragakis, F. Ritort, C. Bustamante, M. Karplus, and G. E. Crooks, J. Chem. Phys. **129**, 024102 (2008).

[2] C. H. Bennett, J. Comput. Phys. **22**, 245 (1976).

[3] M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, Phys. Rev. Lett. **91**, 140601 (2003).
[4] P. Maragakis, M. Spichty, and M. Karplus, Phys. Rev. Lett. **96**, 100602 (2006).
[5] G. E. Crooks, Phys. Rev. E **60**, 2721 (1999).
[6] M. R. Shirts and J. D. Chodera, J. Chem. Phys. **in press**, (2008).
[7] J. A. Anderson, Biometrika **59**, 19 (1972).
[8] C. J. Geyer, Technical Report No. 568, School of Statistics, University of Minnesota, Minneapolis, Minnesota (unpublished).
[9] A. Kong, P. McCullagh, X.-L. Meng, D. Nicolae, and Z. Tan, J. R. Stat. Soc. B. **65**, 585 (2003).
[10] H. Jeffreys, Proc. Royal Soc. London **186**, 453 (1946).
[11] P. Goyal, AIP Conference Proceedings **803**, 366 (2005).