Partition Functions from Rao-Blackwellized Tempered Sampling

David Carlson* Patrick Stinson* Ari Pakman* Liam Paninski

Department of Statistics and Grossman Center for the Statistics of Mind, Columbia University

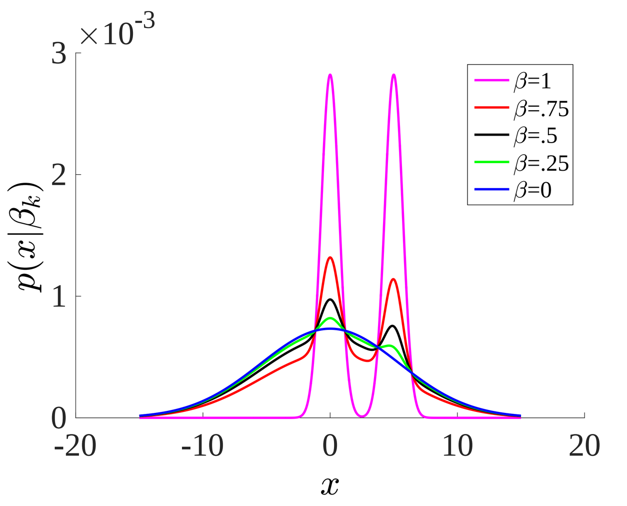
* Equal contribution, randomized order.

Abstract

We present Rao-Blackwellized Tempered Sampling (RTS), a new method to compute partition functions of complex and multimodal distributions. The method exploits the multinomial probability law of the inverse temperatures in simulated tempering, and provides estimates of the partition function in terms of a simple quotient of Rao-Blackwellized marginal inverse temperature probability estimates, which are updated while sampling. The method has interesting connections with several alternative popular methods. We empirically find that RTS provides more accurate estimates than Annealed Importance Sampling when calculating partition functions of large Restricted Boltzmann Machines (RBM); moreover, the method is sufficiently accurate to track training and validation log-likelihoods during learning of RBMs, at minimal computational cost.

Simulated Tempering

To sample from a multimodal unnormalized distribution $f(x)$, we introduce a normalized, easy-to-sample distribution $p_1(x)$ and construct an interpolating family parametrized by an inverse temperature sequence $\{0 = \beta_1 < \beta_2 < \ldots < \beta_K = 1\}$.



- The interpolating functions are

$$p(x|\beta_k) = f_k(x)/Z_k,$$

where $f_k(x) = f(x)^{\beta_k} p_1(x)^{1-\beta_k}$ and $Z_k = \int f_k(x)dx$.

- The inverse temperatures become random variables, with $p(\beta_k) = r_k$, and we define the joint distribution:

$$p(x, \beta_k) = p(x|\beta_k) r_k = f_k(x) r_k / Z_k\,.$$

- Z_k is unknown (except $Z_1 = 1$) – approximated with $\hat{Z}_k$:

$$q(x, \beta_k) \propto f_k(x) r_k / \hat{Z}_k\,.$$

- The marginal distribution over β_k is

$$q(\beta_k) \propto r_k Z_k / \hat{Z}_k\,.$$

Rao-Blackwellized Tempered Sampling (RTS)

Main idea: partition function obtained from the marginal distribution of β_k:

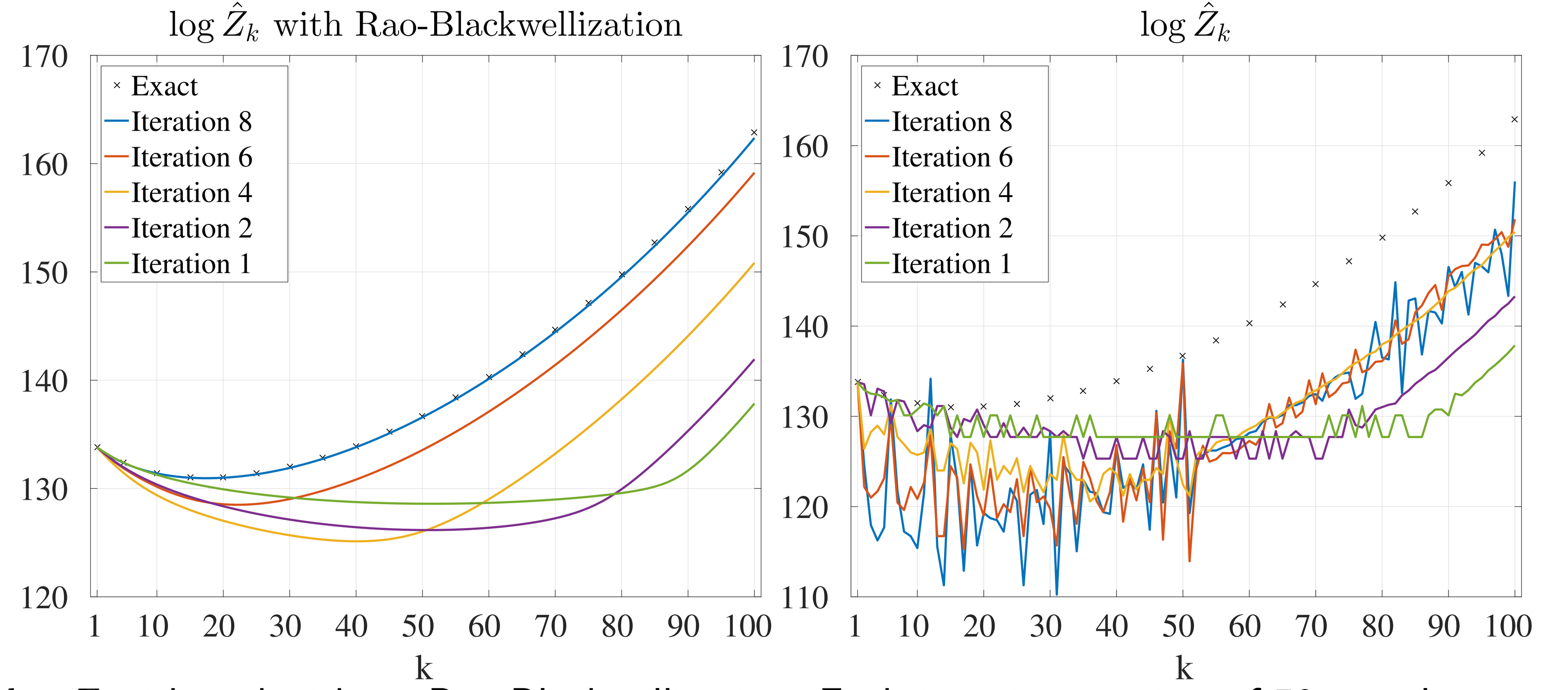
$$Z_k = \hat{Z}_k \frac{r_1 q(\beta_k)}{r_k q(\beta_1)}, \qquad k = 2, \ldots, K\,.$$

- Estimates of $q(\beta_k)$ from tempered samples $\{x^{(i)}, \beta_{k^{(i)}}\}$
  - Simple estimate: $c_k = \frac{1}{N}\sum_{i=1}^N \delta_{k,k^{(i)}}$
  - Rao-Blackwellized estimate: $c_k = \frac{1}{N}\sum_{i=1}^N q(\beta_{k^{(i)}}|x^{(i)})$

Plugging in the estimates yields the consistent estimator:

$$\hat{Z}_k^{\mathrm{RTS}} = \hat{Z}_k \frac{r_1 c_k}{r_k c_1}, \qquad k = 2, \ldots, K\,.$$

- Issue: Tempered sampling mixes poorly with inaccurate $\hat{Z}_k$
- Solution: Update the $\hat{Z}_k$ with small numbers of samples until stable



$\log Z$ with and without Rao-Blackwellization. Each iteration consists of 50 samples on 100 parallel chains in a small RBM.

Relationships to Other Tempered Approaches

- Define simplified notation: $\Delta_x = \log f(x) - \log p_1(x)$

Multistate Bennett Acceptance Ratio (MBAR):

- Based on the identity ($\alpha(x)$ is arbitrary)

$$\frac{Z_k}{Z_1} = \frac{\mathbb{E}_{p(x|\beta_1)}[\alpha(x) f_k(x)]}{\mathbb{E}_{p(x|\beta_k)}[\alpha(x) f_1(x)]},$$

- MBAR estimate is given by maximizing a log-likelihood:

$$L[\mathbf{Z}] = \frac{1}{N}\sum_{i=1}^N \log\left(\sum_{k=1}^K \frac{n_k}{N} \exp(-\log Z_k + \beta_k \Delta_{x_i})\right) + \sum_{r=1}^K \frac{n_r}{N} \log Z_r.$$

- n_k is number of samples at β_k and $N = \sum_{k=1}^K n_k$
- Solved by Newton-Raphson (high computational overhead)
- Surprise: If $\frac{n_k}{N}$ is replaced by its expectation $q(\beta_k)$, MBAR and RTS are **equivalent** estimators

Thermodynamic Integration (TI):

- Based on numerical integration
- For continuous $\beta \in [0, 1]$:

$$\frac{d}{d\beta}\log Z(\beta) = \int \frac{1}{Z(\beta)} \frac{d}{d\beta} f_\beta(x) dx = \mathbb{E}_{x|\beta}[\Delta_x],$$

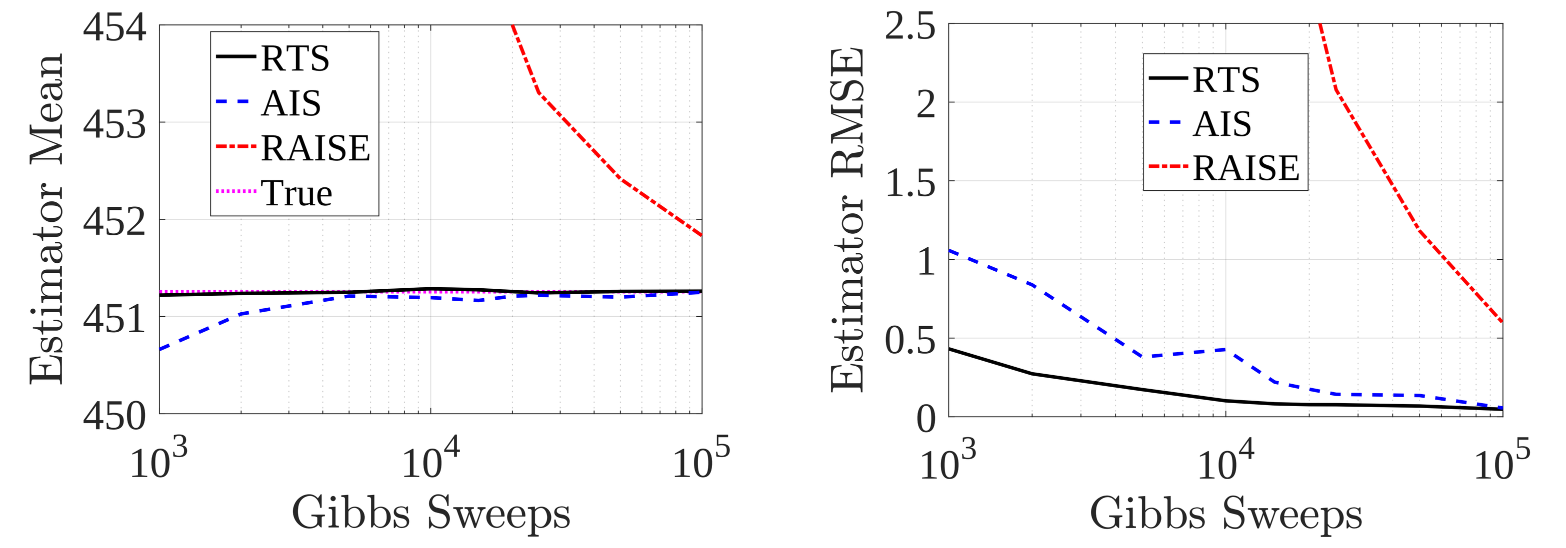
- Discrete temperatures used in practice (discretization error)
- Our Rao-Blackwellization strategy can also be applied to TI:

$$\frac{d}{d\beta}\log Z(\beta)\bigg|_{\beta=\beta_k} \simeq \Sigma_{i=1}^N \frac{q(\beta_k|x_i)\Delta_{x_i}}{\Sigma_{j=1}^N q(\beta_k|x_j)}.$$

- Surprise: In the continuous limit, Rao-Blackwellized TI and RTS are **equivalent** estimators
- RTS has no discretization error

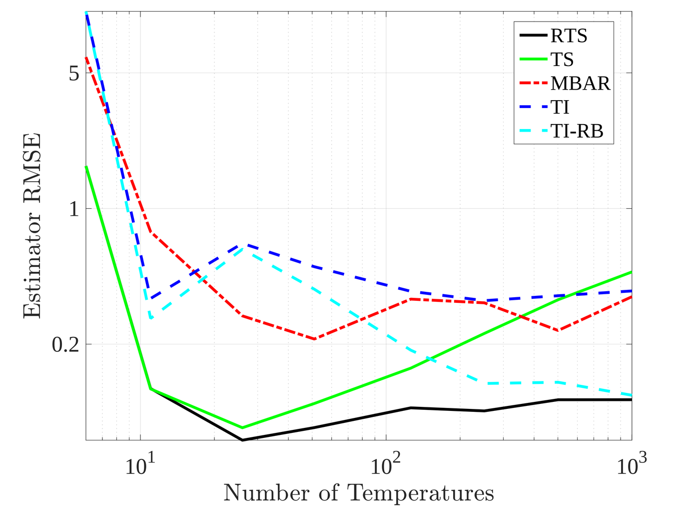
Example: RBM Partition Function

Mean and Root Mean Squared Error of Z_K in an RBM with 500 hidden units. RTS is compared to Annealed Importance Sampling (AIS) and Reverse Annealed Importance Sampling (RAISE). Gibbs sweeps are performed over 100 parallel chains.



Robustness to Number of Temperatures

Performance of various estimators as a function of the number of temperatures. RTS is comparatively robust to the number of temperatures



Tracking Partition Functions While Learning

Since RTS requires a relatively low number of samples and the parameters are slowly changing, we are able to track the value of train- and validation-set likelihoods during RBM training at minimal additional cost.