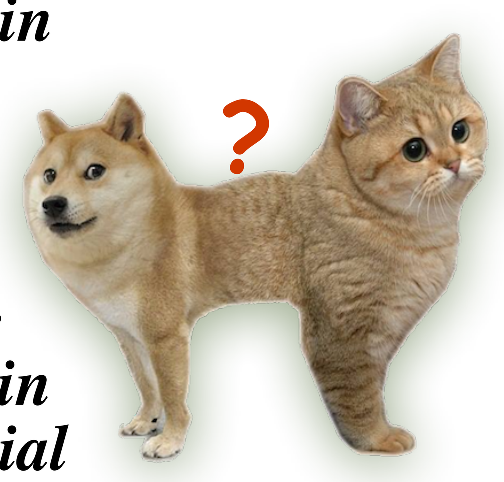
MC²RAM

In-SRAM Markov Chain Monte Carlo Sampling for Fast Bayesian Inference

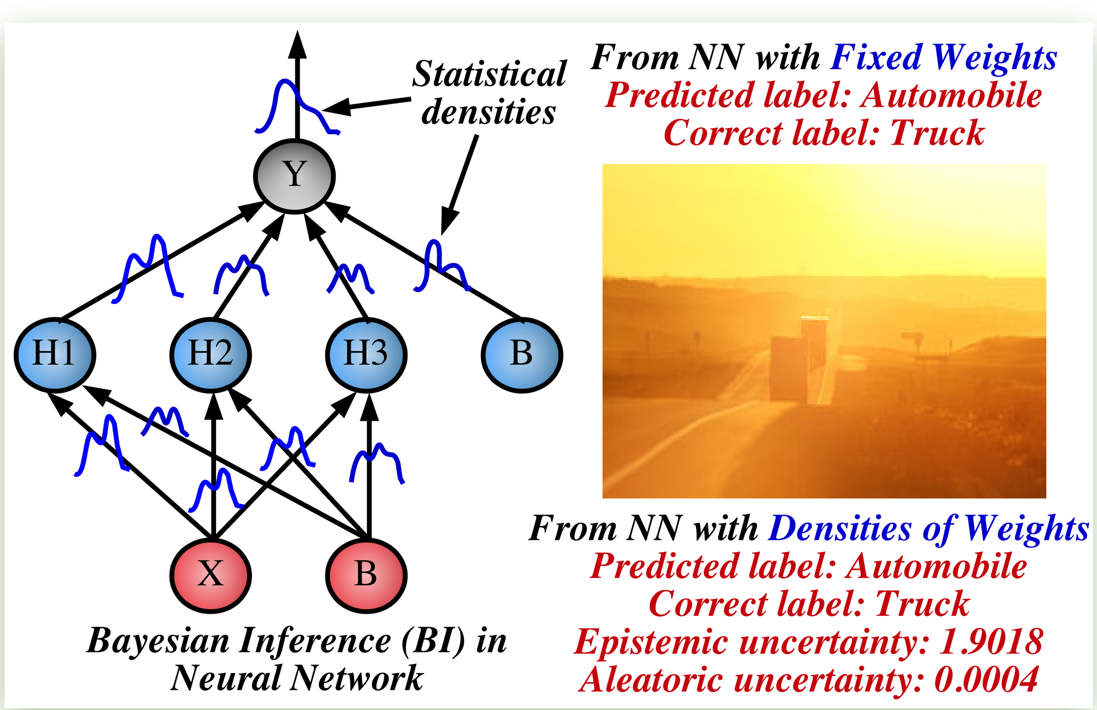
Priyesh Shukla and Amit Ranjan Trivedi

Department of Electrical and Computer Engineering, University of Illinois at Chicago, IL, USA

In Bayesian Inference (BI), the predictions over different model parameters are weighted by how much we believe in those parameter values given the data.



Accounting for these uncertainties in prediction is crucial for critical real-time decision making in settings like autonomous driving and surgical robots



Θ : weight (w), bias (b)

Bayes model: $P(\theta|D) \propto P(D|\theta).P(\theta)$

Prediction: $P(y|x,D) = \int P(y|x,\theta).P(\theta|D)d\theta$

(Intractable integral)

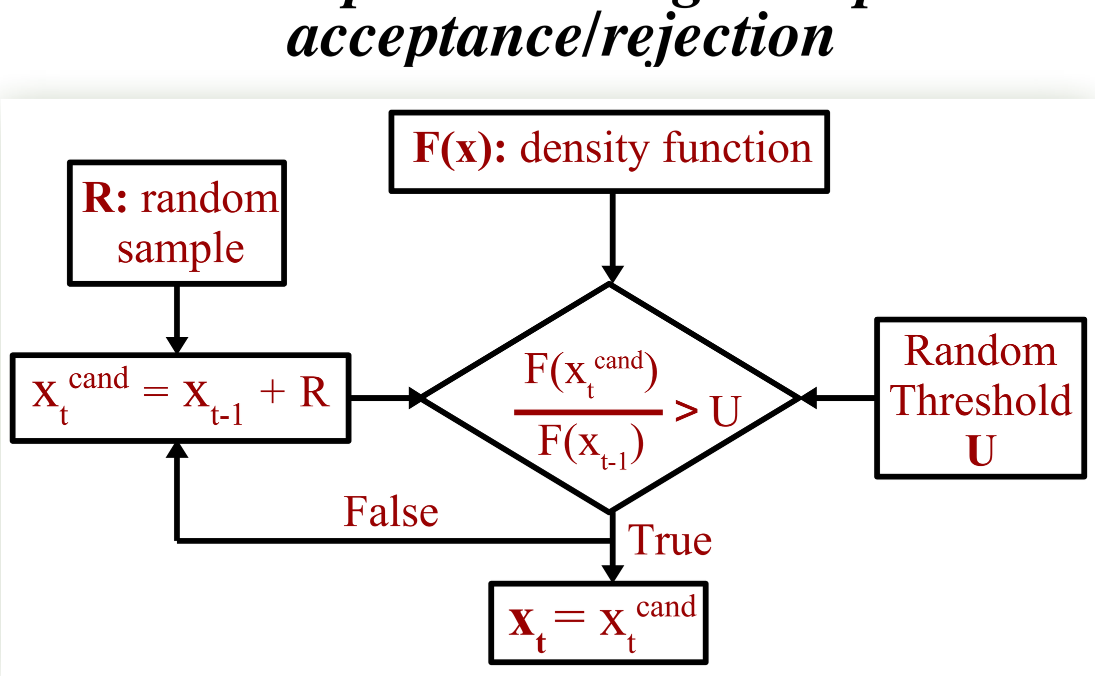
Using Monte Carlo approach to numerically compute these quantities

$$\int G(x).F(x)dx \approx (1/T).\sum_t G(x_{F(x)})$$

G(x)≡Mixture of Gaussians which is proportional to Exponent computed as

$$E_j(t) = E_j(t-1) + (R/\sigma_j^2).R + 2.(R/\sigma_j^2)$$

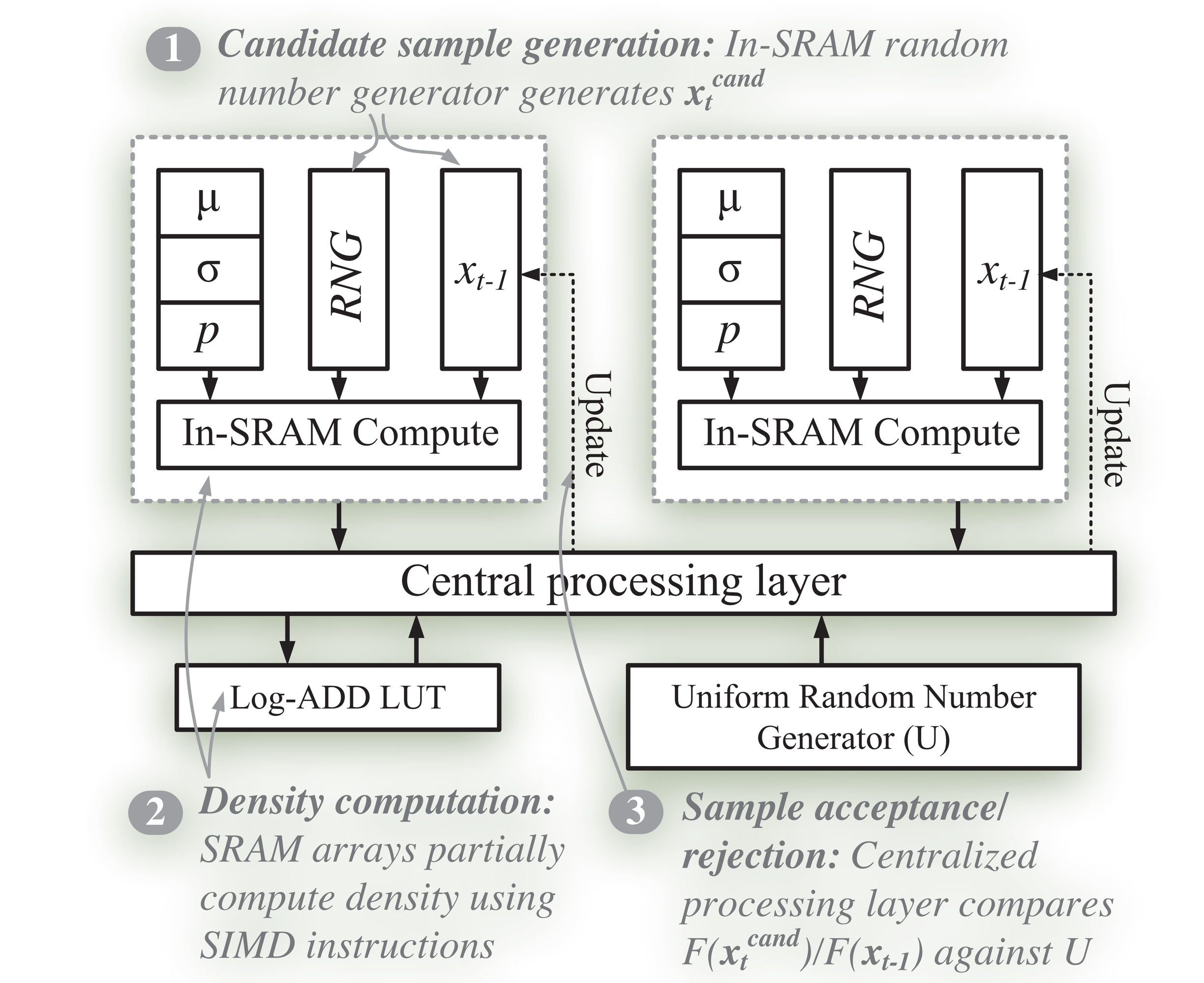
Metropolis-Hastings sample acceptance/rejection



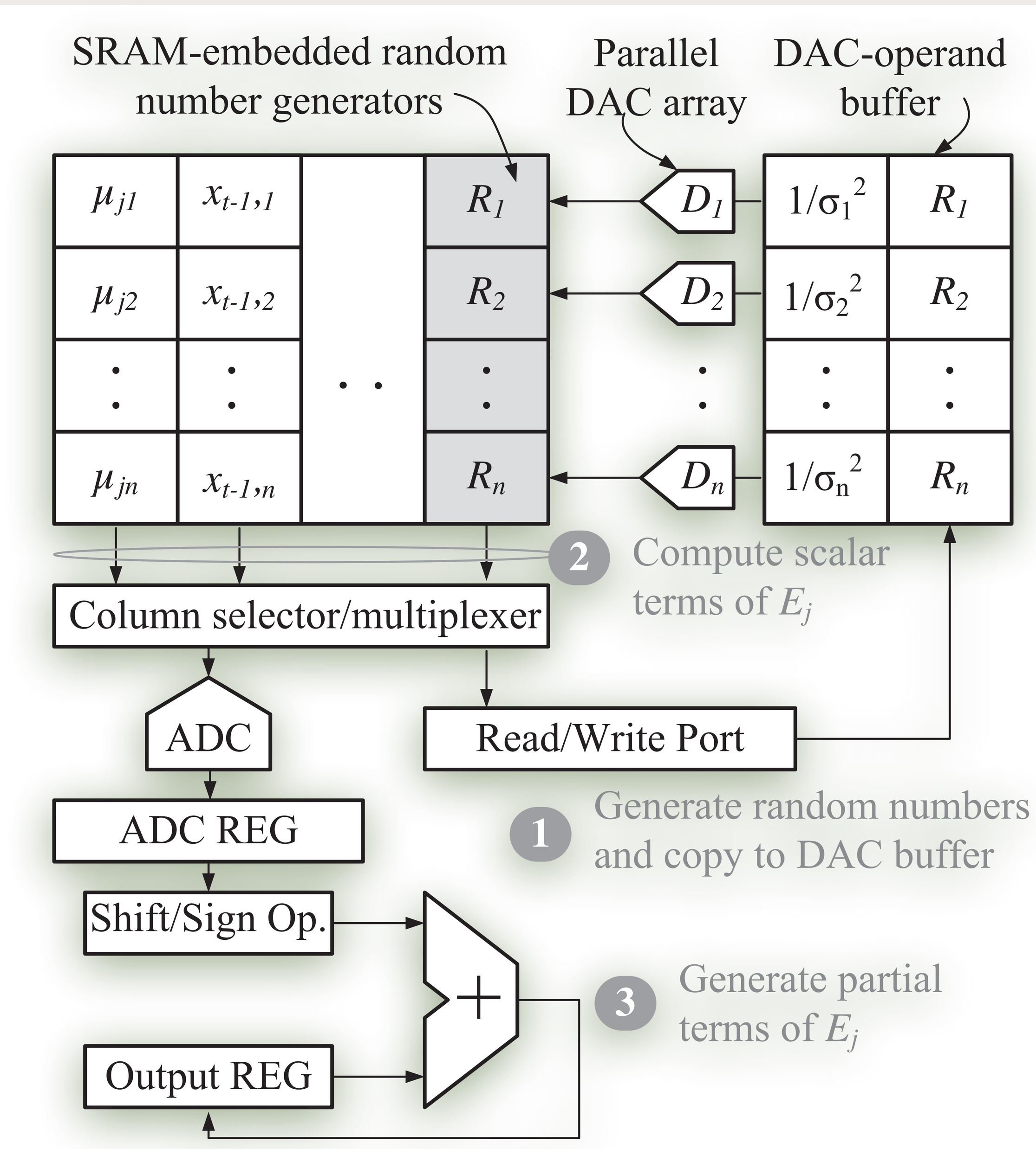
The posterior probability distribution of weights over data for a given model is a mixture of gaussian (GMM) components.

Bivariate GMM posterior of weights is sampled by MC²RAM using Metropolis-Hastings (MH) MCMC sampling criteria.

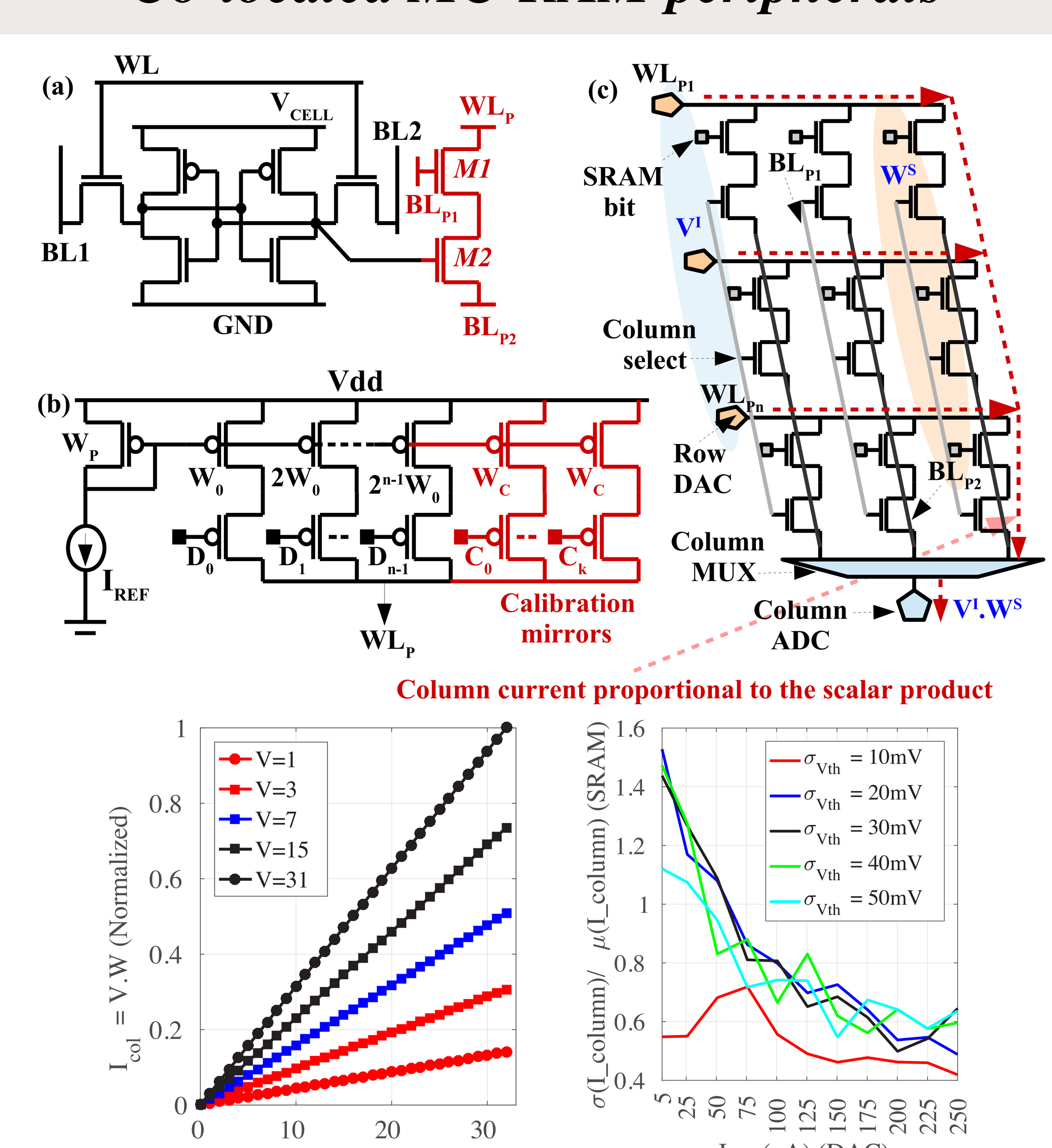
A key framework to accelerate MCMC based sampling for BI



A Mixture-of-Gaussian Density Computation

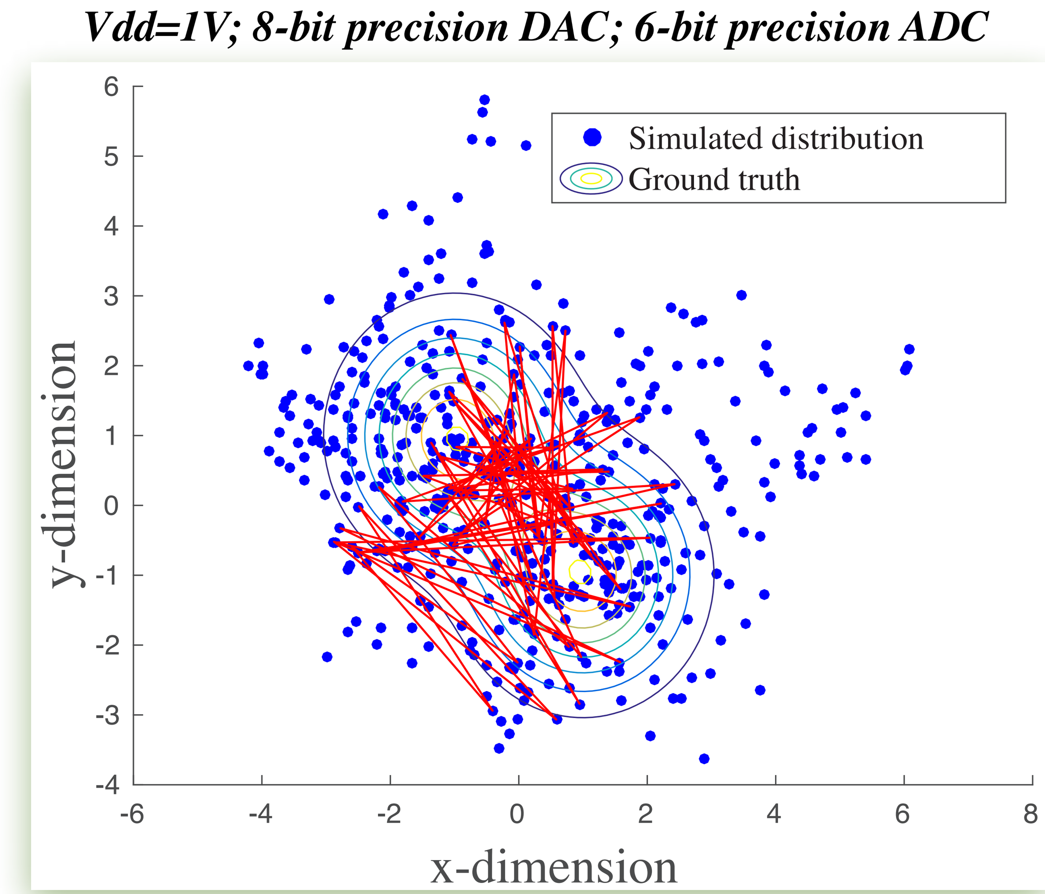


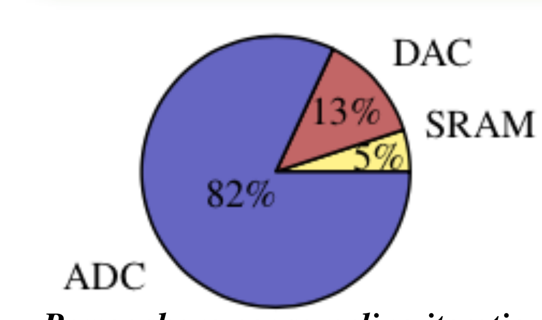
Co-located MC²RAM peripherals



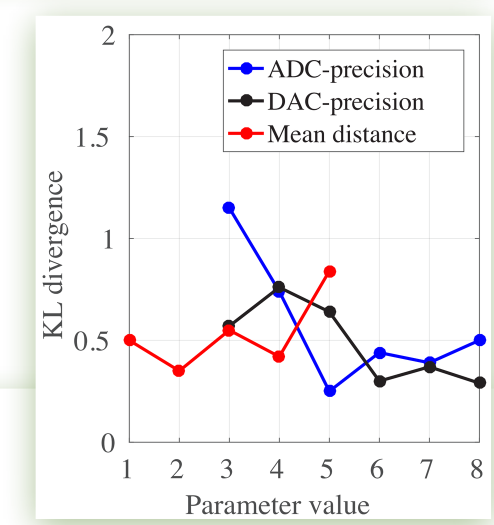
MCMC-MH based samples observed over ground truth posterior

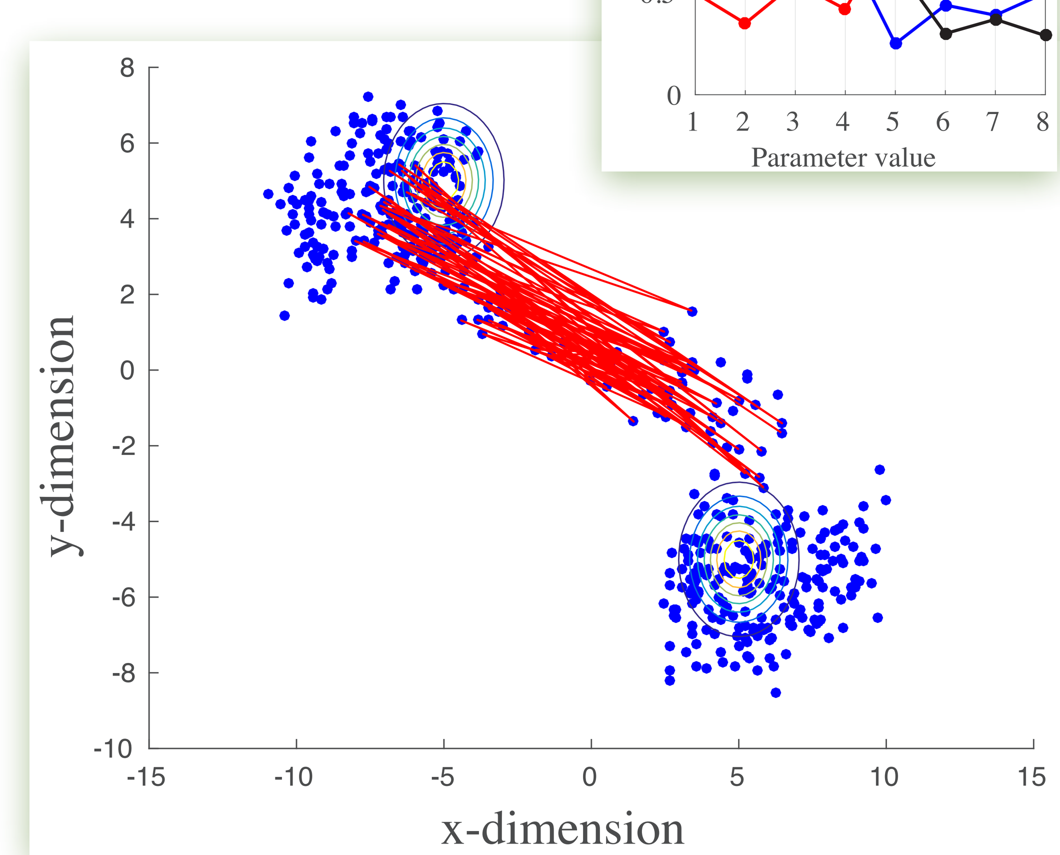
Vdd=1V; 8-bit precision DAC; 6-bit precision ADC





Power share per sampling iteration





- ***KL divergence figures illustrate the degree of deviation of sampled distribution from ground truth posterior***
- ***MCMC sampling is imprecision tolerant to SRAM peripherals (DAC/ADC)***
- ***Power share is greatly impacted by peripheral components in SRAM***
- ***Sampling in higher dimensions call out for efficient and high degree of parallel operations using multiple SRAM banks***

To Investigate…

Can peripherals be more efficient?

Is Metropolis-Hastings good enough for BI?

Can we efficiently accelerate sampling for 500-dimensional random variables?

Tapeout complexities?

Key References…

Blundell et. al, ICML 2015
Cai et. al, ASPLOS 2018
Zhang et. al, JSSC 2015
Kyle Dorman's Blogposts