Bayesian Inference for RRAM

April 22, 2019

1 Introduction

After we write a value into a RRAM unit, the conductance of the unit may fluctuate in short-term time due to a non-ideal relaxation factor. Besides, there also exists noise in reading operation, which makes it difficult to analyze the relaxation mechanism. Here, we use a hierarchical Bayesian model to account the uncertainty in the writing and reading operation of a RRAM. This model is based on the practical operating (writing and reading) procedure and Lin's previous analysis. We hope to build an effective and reasonable tool to analyze the characteristics of relaxation by performing inference on the Bayesian model.

2 Model

We assume that we write M times and each time we have i.i.d. samples from N RRAM units.

1. $w_m, m = 1, ..., M$ are the values to be written. We regard the last verify-pass values during mapping procedure as the written values. W is the collection of $w_1...w_M$.

2. a_m^n is a one-hot vector that denotes the state of relaxation on the n-th R-RAM unit. Based on previous analysis, at any given time, the probability of relaxation is in the same distribution family in different conductance levels. We assume that there are K states ($k = 0$ denotes that RRAM unit does not show relaxation phenomenon, $k = 1$ and $k = 2$ denotes that the conductance of the RRAM unit shifts up and down due to relaxation, respectively) and $a_{m,k}^n = 1$ denotes that the k-th state is activated. $a_m^n | w_m \sim \mathcal{C}(a_m^n | f_\pi(w_m))$ where $f_\pi(w_m)$ is a function of w_m with parameters π, and it outputs a K dimensional simplex,and $\mathcal{C}$ denotes the categorical distribution. A is the collection of all a.

3. v_m^n is the actual conductance of the RRAM unit, i.e., the conductance after relaxing at a given time. Based on previous analysis, the distribution form of relaxation is the exponential distribution. So, $v_m | w_m, a_{m,0} = 1 \sim \mathcal{E}(v_m | w_m, a_{m,0} = 1, g_{\lambda_1}(w_m))$, where $g_{\lambda_1}(w_m)$ is a function of w_m with

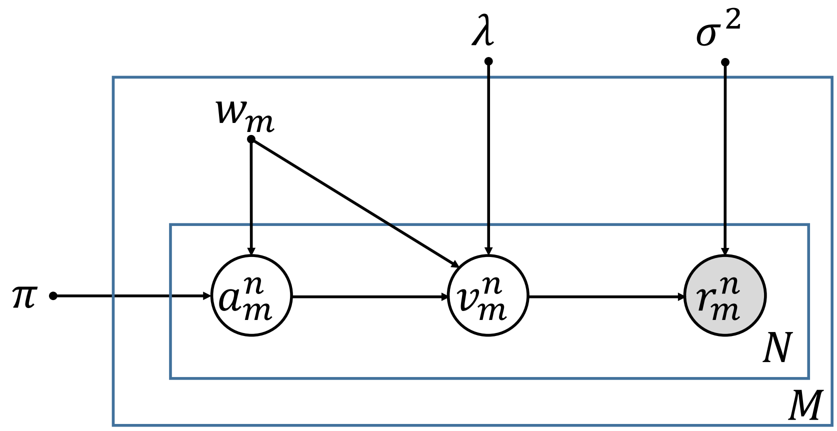

Figure 1: Graphical model illustration of the hierarchical Bayesian model.

parameters λ_1 and it outputs a positive value. λ is the collection of λ and V is the collection of all v.

4. r_m^n denotes the value read from the n-th RRAM. Considering that the reading procedure will introduce a Gaussian noise, we get $r_m | v_m \sim \mathcal{N}(v_m, \sigma^2)$. R is the collection of all r.

Figure 1 shows the graphical model. Formally, the conditional likelihood is:

$$
\begin{aligned}
&p(A, V, R | W, \pi, \lambda, \sigma^2) \\
=&p(A|W, \pi) p(V|A, W, \lambda) p(R|V, \sigma^2) \\
=&\prod_{m=1}^{M} \prod_{n=1}^{N} p(a_m^n | f_\pi(w_m)) p(v_m^n | a_m^n, g_\lambda(w_m)) p(r_m^n | v_m^n, \sigma^2)
\end{aligned}
$$

Variational inference We use a variational posterior to approximate the true posterior.

$$q(A, V | R, W, \theta) = q(V|R, W, \theta) q(A|V, R, W)$$

$$q(V|R, W, \theta) = \prod_{m=1}^{M} \prod_{n=1}^{N} q(v_m^n | h_{\theta_1}(r_m^n, w_m), h_{\theta_2}(r_m^n, w_m))$$

$$q(A|V, R, W) = p(A|V, W) = \prod_{m=1}^{M} \prod_{n=1}^{N} \frac{p(a_m^n | f_\pi(w_m)) p(v_m^n | a_m^n, g_\lambda(w_m))}{\sum_{a_m^n} p(a_m^n | f_\pi(w_m)) p(v_m^n | a_m^n, g_\lambda(w_m))}$$

Approximate maximum conditional likelihood learning We optimize an variational lower-bound of the log conditional likelihood as follows:

$$
\begin{aligned}
&\log p(A, V, R|W, \pi, \lambda, \sigma^2) \\
\geq &\mathbb{E}_{q(A,V|R,W,\theta)} \log \frac{p(A, V, R|W, \pi, \lambda, \sigma^2)}{q(A, V|R, W, \theta)} \\
= &\sum_n \sum_m \mathbb{E}_{q(a_m^n, v_m^n|r_m^n, w_m, \theta)} \log \frac{p(a_m^n|f_\pi(w_m))p(v_m^n|a_m^n, g_\lambda(w_m))p(r_m^n|v_m^n, \sigma^2)}{q(v_m^n|h_{\theta_1}(r_m^n, w_m), h_{\theta_2}(r_m^n, w_m)) \frac{p(a_m^n|f_\pi(w_m))p(v_m^n|a_m^n, g_\lambda(w_m))}{\sum_{a_m^n} p(a_m^n|f_\pi(w_m))p(v_m^n|a_m^n, g_\lambda(w_m))}} \\
= &\sum_n \sum_m \mathbb{E}_{q(a_m^n, v_m^n|r_m^n, w_m, \theta)} \log \frac{\left[\sum_{a_m^n} p(a_m^n|f_\pi(w_m))p(v_m^n|a_m^n, g_\lambda(w_m))\right] p(r_m^n|v_m^n, \sigma^2)}{q(v_m^n|h_{\theta_1}(r_m^n, w_m), h_{\theta_2}(r_m^n, w_m))} \\
= &\sum_n \sum_m \mathbb{E}_{q(v_m^n|r_m^n, w_m, \theta)} \log \frac{\left[\sum_{a_m^n} p(a_m^n|f_\pi(w_m))p(v_m^n|a_m^n, g_\lambda(w_m))\right] p(r_m^n|v_m^n, \sigma^2)}{q(v_m^n|h_{\theta_1}(r_m^n, w_m), h_{\theta_2}(r_m^n, w_m))}
\end{aligned}
$$

We directly optimize that bound to update all parameters using stochastic gradient variational Bayes (SGVB) estimate.

Inference on test data The probability distribution of the actual conductance given r and w is :

$$q(v|r, w, \theta) = \mathcal{N}(v|h_{\theta_1}(r, w), h_{\theta_2}(r, w))$$

The probability distribution of the relaxation given r and w is :

$$q(a|r, w, \theta) = \int q(a, v|r, w, \theta)dv = \int q(v|r, w, \theta)p(a|v, w)dv$$

which can be estimated using Monte Carlo approximation.

3 Experiment