Sketched Clustering via Hybrid Approximate Message Passing

Evan Byrne, Rémi Gribonval, and Philip Schniter

The Ohio State University

(Supported by NSF Grant 1716388 and MIT Lincoln Labs)

Traditionally, clustering problems involve the task of identifying distinct groups of objects from a dataset. However, in modern applications, datasets are often large and complex, making traditional clustering algorithms computationally expensive. To address these challenges, the paper introduces Sketched Clustering via Hybrid Approximate Message Passing (SHyGAMP), a method that aims to make clustering more efficient and scalable.

### Sketched Clustering

#### Description of SHyGAMP

- **SHyGAMP** approximates sum-product loopy belief propagation on factor graphs of the form.

  - **Hybrid Approximate Message Passing**

- **SHyGAMP** iteratively passes messages back and forth between the $p_y$ and $p_{\mathbf{y}}$ nodes until convergence.

- **SHyGAMP** is invariant to the order in which messages are passed.

- **SHyGAMP** is capable of handling large datasets by using a sketching technique.

#### The SHyGAMP Algorithm

- **Require:** frequency matrix $W$, sketch pdfs $p_y$ and $p_{\mathbf{y}}$ from (9,10), initializations $P_0(0), Q_0(0), \alpha$.

  - **Ensure:** $\alpha < 0$, $\rho(W) = 0$.

  - **Repeat**

    1. $t = t + 1$

    2. $m = m + 1$

    3. $y = y + 1$

    4. $x = x + 1$

    5. $w = w + 1$

    6. $t = t + 1$

    7. $m = m + 1$

    8. $y = y + 1$

    9. $x = x + 1$

    10. $w = w + 1$

- **Until Terminated**

#### Computation of SHyGAMP Non-linear Steps

- The key technical challenge in applying SHyGAMP to sketched clustering is computing Lines 6-7 of the SHyGAMP algorithm when $p_y$ has the form of (6).

- We have developed a method based on approximating $p_{\mathbf{y}} | x = \mathbf{m}$ with a Generalized von Mises distribution and evaluating the necessary integrals with the Laplace Approximation.

### Parameter Tuning

- **Our Gaussian Mixture model (4) requires properly selecting $c_y$ and $\tau$ in (6).**

- **Currently, we assume $\tau$ is invariant to $m$.**

- **Allowing $\tau$ to vary with $m$ increases the generalizability of the model, but it is more difficult to learn.**

### References


- W. Rangan and D. Slepian, "Compressed Bayesian Inference via Generalized Approximate Message Passing."