Identifying Genetic Material with Compressed Sensing Techniques

Jamie Dougherty

8th October 2018

Acknowledgements

With thanks to

- The SRIM Scheme
- Dr Sergio Bacallado

Outline

Metagenomics

- Sequence DNA of mixture genomes in large groups of micro-organisms.

Metagenomics

- Sequence DNA of mixture genomes in large groups of micro-organisms.
- Conventional cultivation methods have trouble in these environments.

Metagenomics

- Sequence DNA of mixture genomes in large groups of micro-organisms.
- Conventional cultivation methods have trouble in these environments.
- Applications in medicine - treatment of infections.

Metagenomics

- Sequence DNA of mixture genomes in large groups of micro-organisms.
- Conventional cultivation methods have trouble in these environments.
- Applications in medicine - treatment of infections.
- Still not well understood - room for new methods.

Problem Spec

- Problem can be framed as the reconstruction of a noisy low-rank matrix with structured factors.
- Using prior information from factors can improve error. Still a challenge to derive algorithms that use this information.

Approximate Message Passing (AMP)

- Message passing algorithms allow efficient marginalising of distributions with a certain structure of dependencies.

Approximate Message Passing (AMP)

- Message passing algorithms allow efficient marginalising of distributions with a certain structure of dependencies.
- AMP does away with this structure. Just requires high number of dependencies.

Approximate Message Passing (AMP)

- Message passing algorithms allow efficient marginalising of distributions with a certain structure of dependencies.
- AMP does away with this structure. Just requires high number of dependencies.
- Some limit theorems are then used to derive a different iterative procedure (AMP).

Approximate Message Passing (AMP)

- Message passing algorithms allow efficient marginalising of distributions with a certain structure of dependencies.
- AMP does away with this structure. Just requires high number of dependencies.
- Some limit theorems are then used to derive a different iterative procedure (AMP).
- Some motivation comes from Statistical Physics (Replica and Cavity methods).

Approximate Message Passing (AMP)

- Message passing algorithms allow efficient marginalising of distributions with a certain structure of dependencies.
- AMP does away with this structure. Just requires high number of dependencies.
- Some limit theorems are then used to derive a different iterative procedure (AMP).
- Some motivation comes from Statistical Physics (Replica and Cavity methods).

Example

Problem: find sparse x given $y = Ax, \quad A \in \mathbb{R}^{N \times n}, \quad n << N$
AMP iterate:

$$x^{t+1} = \eta(x^t + A^T z^t)$$
$$z^t = y - Ax^t + \eta'(x^{t-1} + A^T z^{t-1})$$

Interesting Points

- Structure of the problem is exploited.
- Updates are computationally efficient.

Interesting Points

- Structure of the problem is exploited.
- Updates are computationally efficient.
- Engenders a separate set of equations know as State Evolution:
  1. Approximates error of iterative procedure.
  2. Can be used by algorithm to modify optimal mapping to target structure.

Interesting Points

- Structure of the problem is exploited.
- Updates are computationally efficient.
- Engenders a separate set of equations know as State Evolution:
  1. Approximates error of iterative procedure.
  2. Can be used by algorithm to modify optimal mapping to target structure.

Example with state evolution

Problem: find sparse x given $y = Ax, \quad A \in \mathbb{R}^{N \times n}, \quad n << N$
AMP update:

$$x^{t+1} = \eta(x^t + A^T z^t, \hat{\sigma}^t)$$
$$z^t = y - Ax^t + \eta'(x^{t-1} + A^T z^{t-1}, \hat{\sigma}^{t-1})$$

State evolution:

$$\hat{\sigma}^t = \hat{\sigma}^t \eta'(A^T z^{t-1} + x^t, \hat{\sigma}^{t-1})$$

AMP for matrix factorisation

Problem

Find $\boldsymbol{x_0} \sim \mu_0$

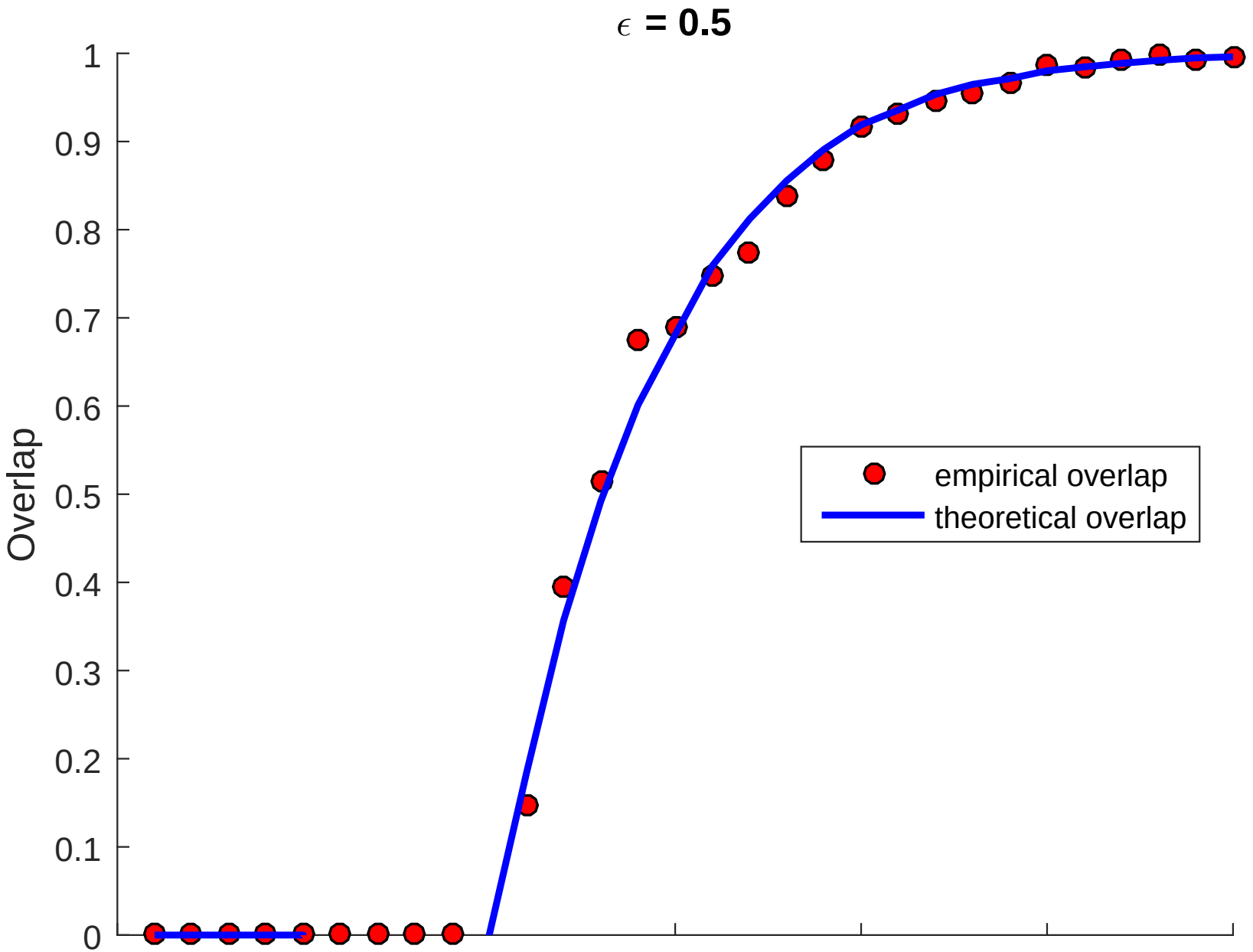
$$A = \frac{\lambda}{n} \boldsymbol{x}_0 \boldsymbol{x}_0^T + W$$

W distributed as a Gaussian orthogonal ensemble.

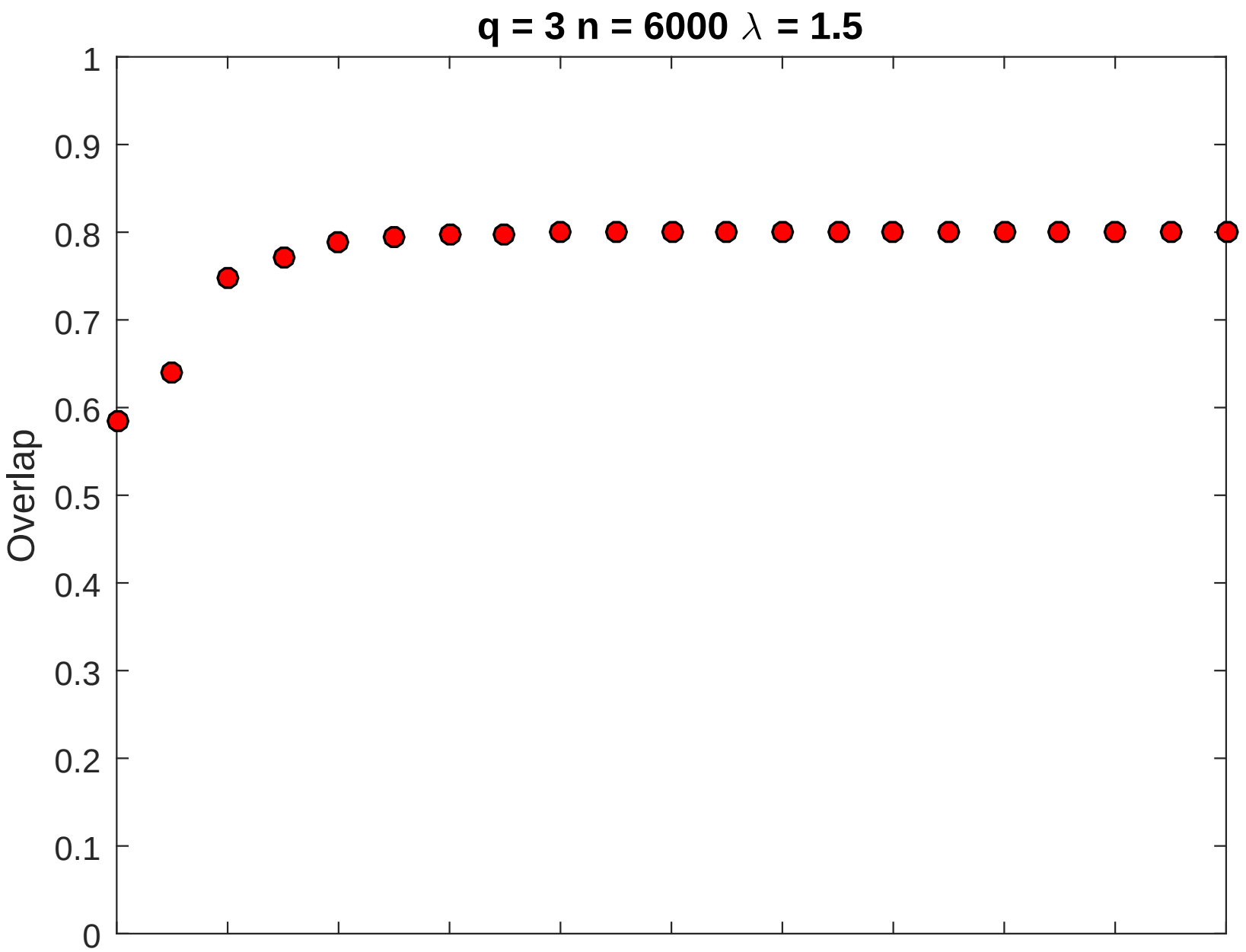
AMP Solution

$$\boldsymbol{x}^{t+1} = A f_t(\boldsymbol{x}^t) - \boldsymbol{b_t} f_{t-1}(\boldsymbol{x}^{t-1})$$

Results in rank one case



How iterates improve over time



Problems

What happens when we try and apply these techniques to our Metagenomic problem?

- Algorithm becomes more complex.

Problems

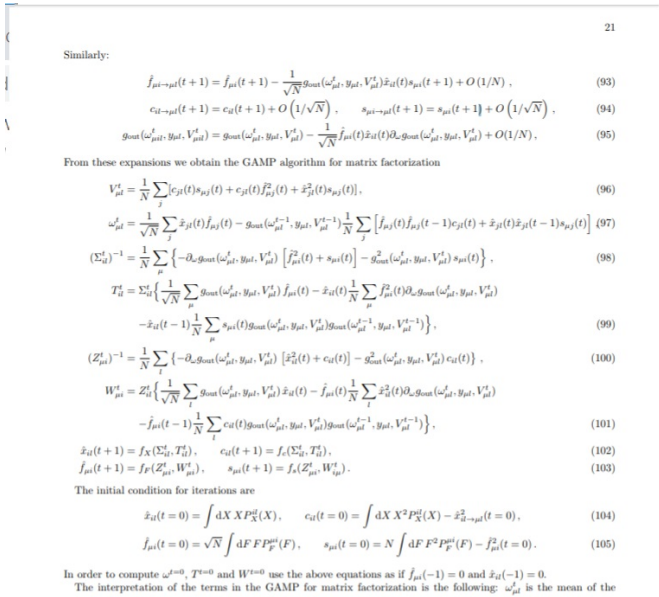
What happens when we try and apply these techniques to our Metagenomic problem?

- Algorithm becomes more complex.
- Much of the literature requires assumptions that don't apply.

Problems

What happens when we try and apply these techniques to our Metagenomic problem?

- Algorithm becomes more complex.
- Much of the literature requires assumptions that don't apply.



21

Similarly:

$$\hat{f}_{\mu i\to\mu t}(t+1) = \hat{f}_{\mu i}(t+1) - \frac{1}{\sqrt{N}} g_{\rm out}(\omega^t_{\mu t}, y_{\mu t}, V^t_{\mu t})\hat{x}_{it}(t) s_{\mu i}(t+1) + O(1/N)\,, \tag{93}$$

$$c_{it\to\mu t}(t+1) = c_{it}(t+1) + O\left(1/\sqrt{N}\right), \qquad s_{\mu i\to\mu t}(t+1) = s_{\mu i}(t+1) + O\left(1/\sqrt{N}\right), \tag{94}$$

$$g_{\rm out}(\omega^t_{\mu i t}, y_{\mu t}, V^t_{\mu i t}) = g_{\rm out}(\omega^t_{\mu t}, y_{\mu t}, V^t_{\mu t}) - \frac{1}{\sqrt{N}} \hat{f}_{\mu i}(t)\hat{x}_{it}(t)\partial_\omega g_{\rm out}(\omega^t_{\mu t}, y_{\mu t}, V^t_{\mu t}) + O(1/N)\,, \tag{95}$$

From these expansions we obtain the GAMP algorithm for matrix factorization

$$V^t_{\mu t} = \frac{1}{N}\sum_j \left[c_{jt}(t) s_{\mu j}(t) + c_{jt}(t)\hat{f}^2_{\mu j}(t) + \hat{x}^2_{jt}(t) s_{\mu j}(t)\right], \tag{96}$$

$$\omega^t_{\mu t} = \frac{1}{\sqrt{N}}\sum_j \hat{x}_{jt}(t)\hat{f}_{\mu j}(t) - g_{\rm out}(\omega^{t-1}_{\mu t}, y_{\mu t}, V^{t-1}_{\mu t})\frac{1}{N}\sum_j \left[\hat{f}_{\mu j}(t)\hat{f}_{\mu j}(t-1)c_{jt}(t) + \hat{x}_{jt}(t)\hat{x}_{jt}(t-1)s_{\mu j}(t)\right] \tag{97}$$

$$(\Sigma^t_{it})^{-1} = \frac{1}{N}\sum_\mu \left\{-\partial_\omega g_{\rm out}(\omega^t_{\mu t}, y_{\mu t}, V^t_{\mu t})\left[\hat{f}^2_{\mu i}(t) + s_{\mu i}(t)\right] - g^2_{\rm out}(\omega^t_{\mu t}, y_{\mu t}, V^t_{\mu t})\, s_{\mu i}(t)\right\}, \tag{98}$$

$$T^t_{it} = \Sigma^t_{it}\Big\{\frac{1}{\sqrt{N}}\sum_\mu g_{\rm out}(\omega^t_{\mu t}, y_{\mu t}, V^t_{\mu t})\,\hat{f}_{\mu i}(t) - \hat{x}_{it}(t)\frac{1}{N}\sum_\mu \hat{f}^2_{\mu i}(t)\partial_\omega g_{\rm out}(\omega^t_{\mu t}, y_{\mu t}, V^t_{\mu t})$$
$$-\hat{x}_{it}(t-1)\frac{1}{N}\sum_\mu s_{\mu i}(t) g_{\rm out}(\omega^t_{\mu t}, y_{\mu t}, V^t_{\mu t}) g_{\rm out}(\omega^{t-1}_{\mu t}, y_{\mu t}, V^{t-1}_{\mu t})\Big\}, \tag{99}$$

$$(Z^t_{\mu i})^{-1} = \frac{1}{N}\sum_i \left\{-\partial_\omega g_{\rm out}(\omega^t_{\mu t}, y_{\mu t}, V^t_{\mu t})\left[\hat{x}^2_{it}(t) + c_{it}(t)\right] - g^2_{\rm out}(\omega^t_{\mu t}, y_{\mu t}, V^t_{\mu t})\, c_{it}(t)\right\}, \tag{100}$$

$$W^t_{\mu i} = Z^t_{ij}\Big\{\frac{1}{\sqrt{N}}\sum_i g_{\rm out}(\omega^t_{\mu t}, y_{\mu t}, V^t_{\mu t})\,\hat{x}_{it}(t) - \hat{f}_{\mu i}(t)\frac{1}{N}\sum_i \hat{x}^2_{it}(t)\partial_\omega g_{\rm out}(\omega^t_{\mu t}, y_{\mu t}, V^t_{\mu t})$$
$$-\hat{f}_{\mu i}(t-1)\frac{1}{N}\sum_i c_{it}(t) g_{\rm out}(\omega^t_{\mu t}, y_{\mu t}, V^t_{\mu t}) g_{\rm out}(\omega^{t-1}_{\mu t}, y_{\mu t}, V^{t-1}_{\mu t})\Big\}, \tag{101}$$

$$\hat{x}_{it}(t+1) = f_X(\Sigma^t_{it}, T^t_{it}), \qquad c_{it}(t+1) = f_c(\Sigma^t_{it}, T^t_{it}), \tag{102}$$

$$\hat{f}_{\mu i}(t+1) = f_F(Z^t_{\mu i}, W^t_{\mu i}), \qquad s_{\mu i}(t+1) = f_s(Z^t_{\mu i}, W^t_{i\mu}). \tag{103}$$

The initial condition for iterations are

$$\hat{x}_{it}(t=0) = \int {\rm d}X\, X P^0_X(X), \qquad c_{it}(t=0) = \int {\rm d}X\, X^2 P^0_X(X) - \hat{x}^2_{it\to\mu t}(t=0), \tag{104}$$

$$\hat{f}_{\mu i}(t=0) = \sqrt{N}\int {\rm d}F\, F P^{\rm pri}_F(F), \qquad s_{\mu i}(t=0) = N\int {\rm d}F\, F^2 P^{\rm pri}_F(F) - \hat{f}^2_{\mu i}(t=0). \tag{105}$$

In order to compute $\omega^{t=0}$, $T^{t=0}$ and $W^{t=0}$ use the above equations as if $\hat{f}_{\mu i}(-1) = 0$ and $\hat{x}_{it}(-1) = 0$.
The interpretation of the terms in the GAMP for matrix factorization is the following: $\omega^t_{\mu t}$ is the mean of the

Solutions?

How do we resolve some of these problems?

Solutions?

How do we resolve some of these problems?

- Don't try for Bayes optimal.

Solutions?

How do we resolve some of these problems?

- Don't try for Bayes optimal.
- Choose non-linearities based on intuition.

Solutions?

How do we resolve some of these problems?

- Don't try for Bayes optimal.
- Choose non-linearities based on intuition.
- Lot's of experimentation!

Solutions?

How do we resolve some of these problems?

- Don't try for Bayes optimal.
- Choose non-linearities based on intuition.
- Lot's of experimentation!

Don't have working algorithm, more testing needed.

Summary

Summary

- Metagenomic problem seems tractable.

Summary

- Metagenomic problem seems tractable.
- AMP is very efficient when used correctly.

Summary

- Metagenomic problem seems tractable.
- AMP is very efficient when used correctly.
- However it is difficult to use correctly in this context.

Summary

- Metagenomic problem seems tractable.
- AMP is very efficient when used correctly.
- However it is difficult to use correctly in this context.
- Better intuitions and more testing could generate a solution.