Sparse Multinomial Logistic Regression via Approximate Message Passing

Evan Byrne and Philip Schniter

The Ohio State University

(Supported by NSF CCF-1018368 and CCF-1218754)

Problem Statement

- Goal: Infer the D-ary label y_0 from "test" feature vector $\boldsymbol{a}_0 \in \mathbb{R}^N$ given training $\{y_m, \boldsymbol{a}_m\}_{m=1}^M$.
- Linear classification: Estimate weight matrix $\widehat{\boldsymbol{X}} \in \mathbb{R}^{N \times D}$, then predict $\widehat{y}_0 = \arg\max_d \, [\widehat{\boldsymbol{X}}^T \boldsymbol{a}_0]_d$.
- Feature selection: Determine which subset of N features is needed to accurately predict the label y_0.
- We're especially interested in the case $M \ll N$ (MVPA, text-mining, micro-array gene expression).
  - Possible if "true" $\boldsymbol{X}$ is K-row-sparse with $K \ll M$.

Multinomial Logistic Regression

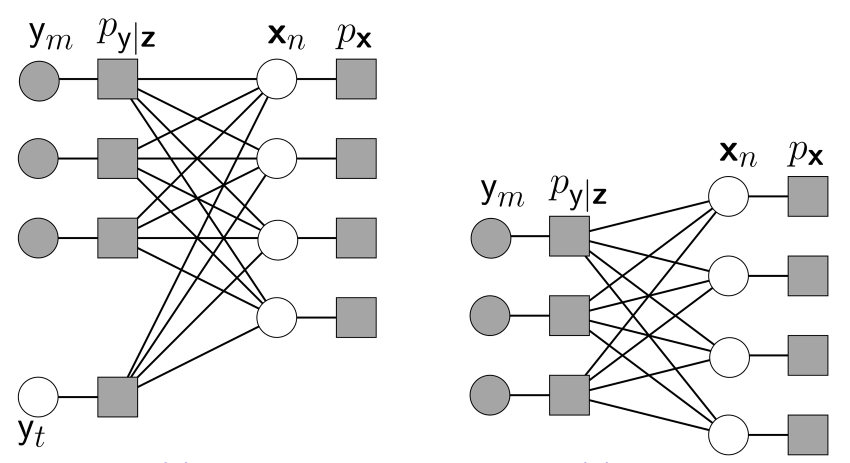
- One approach to designing $\boldsymbol{X}$ is Multinomial Logistic Regression (MLR).
- In MLR, we use the multinomial logistic likelihood:

$$p_{\mathsf{y}|\mathbf{z}}(y_m|\boldsymbol{z}_m) = \frac{\exp([\boldsymbol{z}_m]_{y_m})}{\sum_{d=1}^{D} \exp([\boldsymbol{z}_m]_d)}, \quad y_m \in \{1, \dots, D\} \quad \text{where} \quad \boldsymbol{z}_m \triangleq \boldsymbol{X}^T \boldsymbol{a}_m. \tag{1}$$

- Also, $\boldsymbol{X}$ is regularized through some prior $p_{\mathbf{X}}(\boldsymbol{X})$.
- Existing approaches to sparse MLR include SMLR [Krishnapuram Carin Figueiredo Hartemink 05], SBMLR [Cawley Talbot Girolami 07], and GLMNET [Friedman Hastie Tibshirani 10], which all employ a Laplacian prior for $p_{\mathbf{x}}$ and MAP estimation to find $\widehat{\boldsymbol{X}}$.

HyGAMP for MLR

- Assuming a separable likelihood $p_{\mathbf{y}\,|\,\mathbf{Z}}(\boldsymbol{y}\,|\,\boldsymbol{Z}) = \prod_m p_{\mathsf{y}|\mathbf{z}}(y_m\,|\,\boldsymbol{z}_m)$ and prior $p_{\mathbf{X}}(\boldsymbol{X}) = \prod_n p_{\mathbf{x}}(\boldsymbol{x}_n)$, $p_{\mathbf{y},\mathbf{X}}(\boldsymbol{y}, \boldsymbol{X})$ can be represented by the following factor graph:



(a) Full (b) Reduced

- Through message passing, we break one large inference problem into many smaller inference problems.
- Under large i.i.d. $\boldsymbol{A}$ and scalar z_m & x_n, we can apply generalized approximate message passing (GAMP) [Rangan 11]. Has been used for *binary* logistic regression [Ziniel Schniter Sederberg 15].
- However, our $\mathbf{z}_m$ & $\mathbf{x}_n$ are vector valued, so we instead apply hybrid GAMP (HyGAMP) [Rangan Fletcher Goyal Schniter 12].
  - MSA variant: computes MAP estimate of $\boldsymbol{X}$.
  - SPA variant: computes approximate marginal posteriors of y_0 and $\boldsymbol{X}$ $\Rightarrow$ approximately minimizes test-error rate!
  - Passes $O(M+N)$ messages in the form of D-dimensional Gaussian pdfs.

Algorithm Summary

- HyGAMP iteratively passes messages back and forth between the $p_{\mathsf{y}|\mathbf{z}}$ and $p_{\mathbf{x}}$ nodes until convergence.
- The algorithm can be divided into "linear" and "non-linear" steps.

Linear steps:

- Involve $N+M$ matrix inversions of size $D \times D$.
- Identical for SPA and MSA variants of HyGAMP.

Non-linear steps:

- At each node n and m, HyGAMP approximates the posterior distributions as:

$$p_{\mathbf{x}|\mathbf{r}}(\boldsymbol{x}_n\,|\,\widehat{\boldsymbol{r}}_n; \boldsymbol{Q}_n^{\mathbf{r}}) \propto p_{\mathbf{x}}(\boldsymbol{x}_n)\mathcal{N}(\boldsymbol{x}_n; \widehat{\boldsymbol{r}}_n, \boldsymbol{Q}_n^{\mathbf{r}}) \tag{2}$$

$$p_{\mathbf{z}|\mathsf{y},\mathbf{p}}(\boldsymbol{z}_m\,|\,y_m, \widehat{\boldsymbol{p}}_m; \boldsymbol{Q}_m^{\mathbf{p}}) \propto p_{\mathsf{y}|\mathbf{z}}(y_m\,|\,\boldsymbol{z}_m)\mathcal{N}(\boldsymbol{z}_m; \widehat{\boldsymbol{p}}_m, \boldsymbol{Q}_m^{\mathbf{p}}), \tag{3}$$

for $\widehat{\boldsymbol{p}}_m$, $\boldsymbol{Q}_m^{\mathbf{p}}$, $\widehat{\boldsymbol{r}}_n$, $\boldsymbol{Q}_m^{\mathbf{r}}$ calculated in the linear steps.

- SPA variant: computes the means ($\widehat{\boldsymbol{x}}_n$ and $\widehat{\boldsymbol{z}}_m$) and covariances ($\boldsymbol{Q}_n^{\mathbf{x}}$ and $\boldsymbol{Q}_m^{\mathbf{z}}$) of above posteriors.
- MSA variant: computes the modes ($\widehat{\boldsymbol{x}}_n$ and $\widehat{\boldsymbol{z}}_m$) and inverse Hessians of above log posteriors.
- For MLR likelihood and most sparsity-inducing priors, there are no closed-form solutions. Need approximations like numerical integration, importance sampling, Newtons method, minorize-maximization.
- Typically, to enforce sparsity, we use a Bernoulli-Gaussian prior in SPA-HyGAMP and a Laplacian prior in MSA-HyGAMP.

Simplified HyGAMP (SHyGAMP) for MLR

- Unfortunately, HyGAMP is not computationally competitive due to
  - expensive linear steps (e.g., matrix inversion)
  - expensive non-linear steps (e.g., iterative algorithms)
  - numerical instabilities
- Our Solution: Assume all matrices $\boldsymbol{Q}$ are diagonal.
  - trivializes the linear steps (i.e., no matrix inversion)
  - drastically simplifies the non-linear steps
  - enables use of existing GAMPmatlab software framework [Rangan, Schniter, Parker, Ziniel, et al.]

SPA SHyGAMP

Non-linear $\boldsymbol{z}_m$ steps:

- Simplified posterior mean/variance computations:

$$\widehat{z}_{md} = C_m^{-1} \int_{\mathbb{R}^D} z_d \, p_{\mathsf{y}|\mathbf{z}}(y_m|\boldsymbol{z}) \prod_{k=1}^{D} \mathcal{N}(z_k; \widehat{p}_{mk}, q_{mk}^{\mathbf{p}}) \, \mathrm{d}\boldsymbol{z} \tag{4}$$

$$q_{md}^{\mathbf{z}} = C_m^{-1} \int_{\mathbb{R}^D} z_d^2 \, p_{\mathsf{y}|\mathbf{z}}(y_m|\boldsymbol{z}) \prod_{k=1}^{D} \mathcal{N}(z_k; \widehat{p}_{mk}, q_{mk}^{\mathbf{p}}) \, \mathrm{d}\boldsymbol{z} - \widehat{z}_{md}^2 \tag{5}$$

- Investigated approaches based on numerical integration, importance sampling, Taylor series approximation
- Proposed novel Gaussian-mixture approximation with improved accuracy-runtime tradeoff

Non-linear $\boldsymbol{x}_n$ steps:

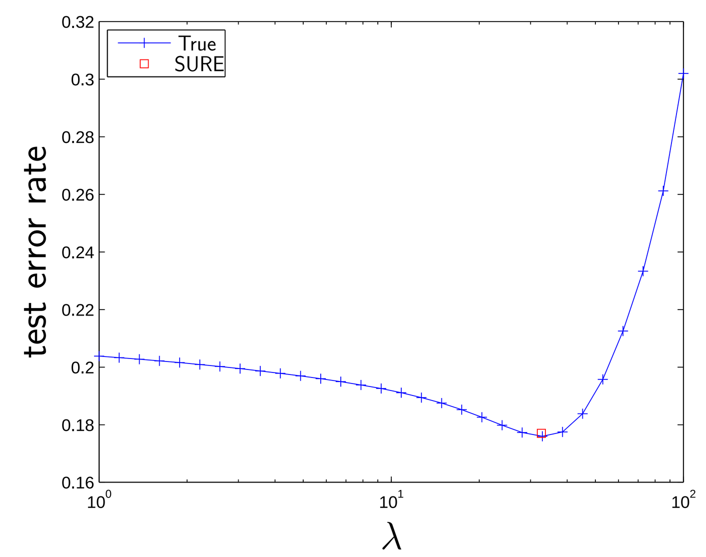
- Choosing separable prior allows further decoupling into D scalar inference problems
- Example: i.i.d. BG: $p_{\mathsf{x}}(x_{nd}) \triangleq \beta\mathcal{N}(x_{nd}; 0, \sigma_x^2) + (1-\beta)\delta(x_{nd}) \; \forall n, d$
- Parameters σ_x^2 and β can be tuned online via EM [Vila, Schniter 13].

MSA SHyGAMP

- Non-linear $\boldsymbol{z}_m$ steps: solved via component-wise Newton's method.
- Non-linear $\boldsymbol{x}_n$ steps: choose ℓ_1 regularization, solve via soft-thresholding.
- λ tuned online via variation on SURE procedure [Mousavi Maleki Baraniuk 13].

SURE tuning procedure

- Idea: at each GAMP iteration, choose λ to minimize the SURE of the thresholder.
- Challenge: objective function is highly non-smooth.
- Our Solution: approximate empirical data by GM distribution $\Rightarrow$ smooth objective function. Minimize using conventional techniques (e.g., gradient descent, bisection search).



- $N = 30000$
- $M = 300$
- $K = 25$
- $D = 4$
- Bayes Error Rate $= .1$
- average of 12 trials

Test error rate vs λ for fixed-λ MSA SHyGAMP, with final error and λ of SURE-tuned MSA SHyGAMP superimposed

Selected References

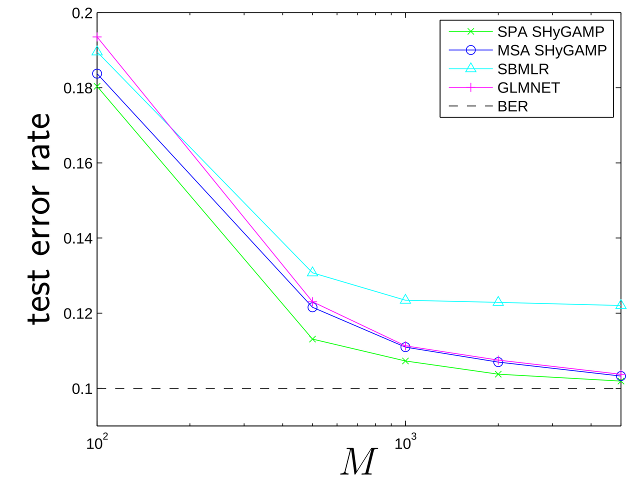
- E. M. Byrne, "Sparse multinomial logistic regression via approximate message passing," Master's thesis, The Ohio State University, July 2015.
- B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, pp. 957-968, June 2005.
- G. C. Cawley, N. L. C. Talbot, and M. Girolami, "Sparse multinomial logistic regression via Bayesian L1 regularisation," in *Proc. Neural Inform. Process. Syst. Conf.*, pp. 209-216, 2007.
- J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Statst. Softw.*, vol. 33, pp. 1-22, Jan. 2010.
- S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inform. Thy.*, (Saint Petersburg, Russia), pp. 2168-2172, Aug. 2011. (full version at *arXiv:*1010.5141)
- J. Ziniel, P. Schniter, and P. Sederberg, "Binary classification and feature selection via generalized approximate message passing," *IEEE Trans. Signal Process.*, vol. 63, no. 8, pp. 2020-2032, 2015.
- S. Rangan, A. K. Fletcher, V. K. Goyal, and P. Schniter, "Hybrid generalized approximate message passing with applications to structured sparsity," in *Proc. IEEE Int. Symp. Inform. Thy.*, pp. 1236-1240, July 2012. (full version at *arXiv:*1111.2581).
- S. Rangan, P. Schniter, J. T. Parker, J. Ziniel, J. Vila, M. Borgerding, and et al. GAMPmatlab. `https://sourceforge.net/projects/gampmatlab/`.
- J. P. Vila and P. Schniter, "Expectation-maximization Gaussian-mixture approximate message passing," *IEEE Trans. Signal Process.*, vol. 61, pp. 4658-4672, Oct. 2013.
- A. Mousavi, A. Maleki, and R. G. Baraniuk, "Parameterless optimal approximate message passing," Oct. 2013. (full version at *arXiv:*1311.0035)

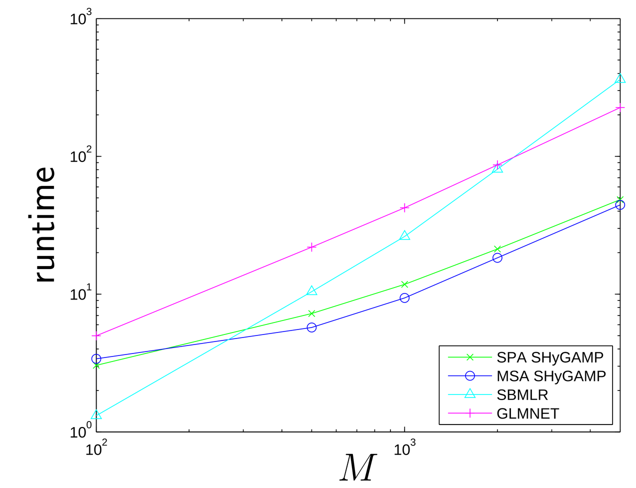
Classification Performance on Synthetic Data

Data generation model:

- features $\boldsymbol{a}_m \,|\, (y_m = d) \sim \mathcal{N}(\boldsymbol{\mu}_d, \sigma_a^2 \boldsymbol{I}_N)$
- feature means $\{\boldsymbol{\mu}_d\}_{d=1}^D$ orthonormal with K non-zero entries
- balanced training labels

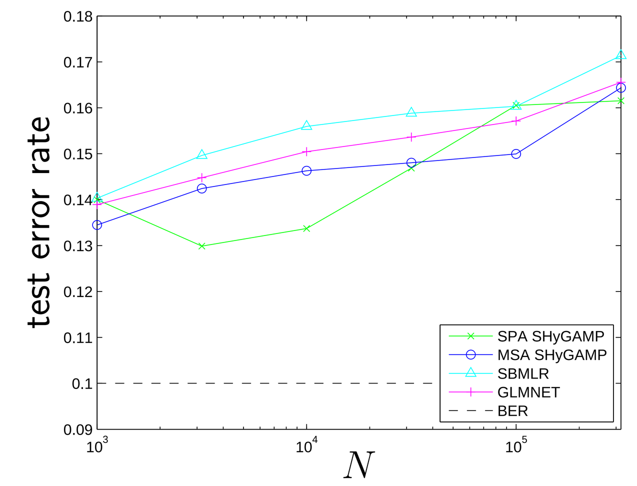
Average classification error and runtime vs M for fixed $D = 4$, $N = 10000$, $K = 10$, 12 trials:

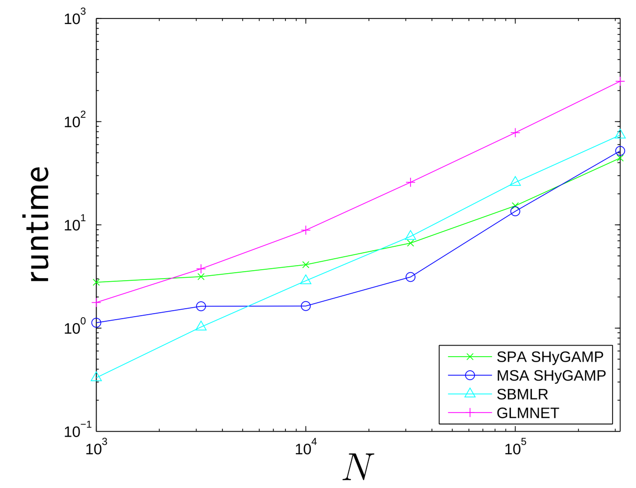




SPA-SHyGAMP wins in error. MSA-SHyGAMP beats SBMLR and GLMNET in both error and runtime (for large M).

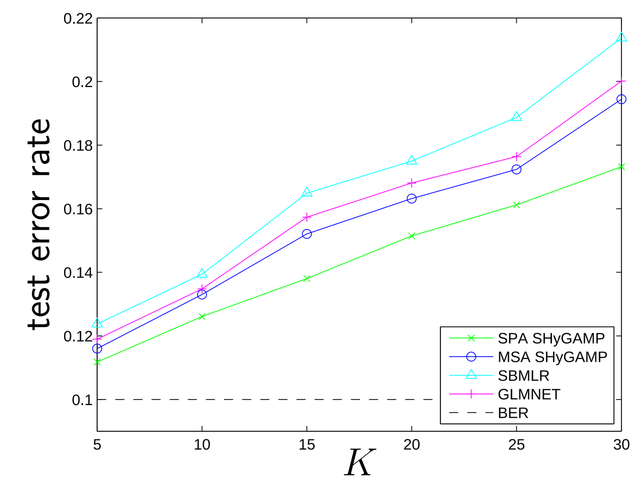
Average classification error and runtime vs N for fixed $D = 4$, $M = 200$, $K = 10$, 12 trials:

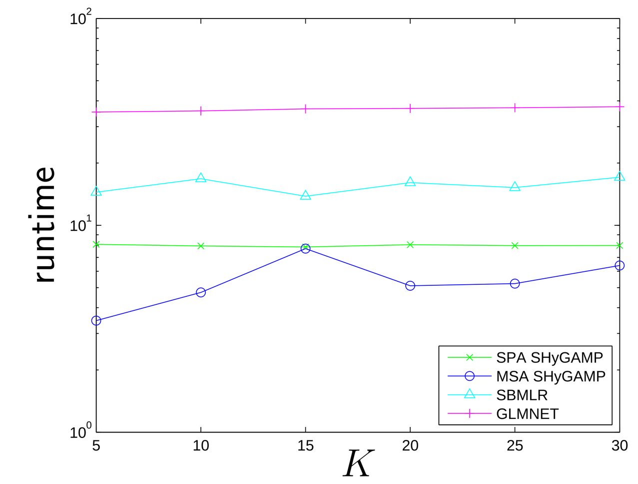




MSA-SHyGAMP beats SBMLR and GLMNET in both error and runtime (for large N).

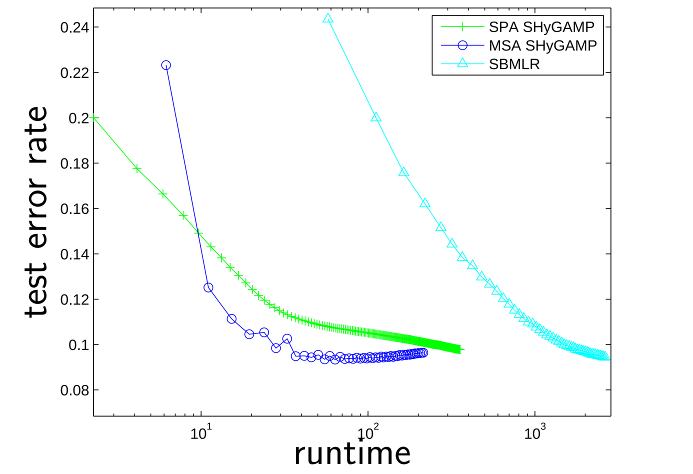
Average classification error and runtime vs K for fixed $D = 4$, $M = 300$, $N = 30000$, 12 trials:





SPA-SHyGAMP wins overall in error. MSA-SHyGAMP beats SBMLR and GLMNET in both error and runtime.

Classification Performance on RCV1 Dataset



- Features are word frequency; labels are document subject (e.g., business).
- $D = 25$, $M_{\text{train}} = 14147$, $N = 47236$, and $M_{\text{test}} = 469571$.
- Shown is test error rate vs training time for auto-tuned algorithms.

Both MSA and SPA SHyGAMP converge to the final error rate faster than SBMLR.