Theoretical Limits of Streaming Inference and Mini-Batch Message-Passing Algorithms

Andre Manoel
Neurospin, CEA
Université Paris-Saclay, Bât. 145
F-91191 Gif-sur-Yvette, France
Email: andremanoel@gmail.com

Eric W. Tramel
Team DYOGENE
Inria Paris
75012 Paris, France

Thibault Lesieur & Lenka Zdeborová
IPhT, CNRS, CEA
Université Paris-Saclay
F-91191 Gif-sur-Yvette, France

Florent Krzakala
LPS, CNRS UMR-8500
ENS, PSL, & Sorbonne Universités
Université Pierre & Marie Curie
75005 Paris, France

Over the past decade, the incredible growth of data aggregation has outpaced our ability to effectively process the data we acquire. Often, due to memory constraints, one must consider *online* or *streaming* methods, processing only small fractions of a dataset at once. This is especially true in the context of machine learning. For example, stochastic gradient descent [1] has been applied in many contexts, famously in deep learning [2], where massive datasets frustrate the exact computation of parameter gradients.

In this work, we are interested in the case of streaming computation, which bridges the gap between offline and online algorithms. We denote online algorithms as those which process only a single data point at once, and offline algorithms as those which consider all available data at once. There exists a natural trade-off between these two settings: while the memory requirements are optimized by online approaches, accuracy is optimized by offline ones. Specifically, we attempt to answer the following question. Given a dataset $\mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M\}$, assuming the data points are generated by a known process, what is the optimal manner in which to divide the dataset, $\mathcal{Y} = \{\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_B\}$, so as to maximize accuracy while minimizing the number of data points, $m = M/B > 1$, in each of these so-called mini-batches? How do we best quantify this trade-off? What are the best algorithms to perform learning in the streaming setting? The answers to these questions have a direct and practical impact on the challenges data scientists face daily.

We answer these questions in a quantitative and insightful manner in two different contexts: in supervised learning, via binary-weights logistic [3], [4] and sparse-weights linear regression (SLR) [5], [6]; and in unsupervised learning, via Gaussian mixture models. By building on the analysis of these models in the offline setting accomplished with the methodology of statistical physics [3], [4], [7]–[13], we show the existence of interesting phase transitions appearing for these streaming inference problems. Their characterization provides information about the learning error that is achievable both in terms of infomation-theoretic limits and computational tractability. We adapt the approximate message passing (AMP) algorithm to the mini-batch setting (Mini-AMP). In the streaming setting, the theory we develop here characterize the performance of this Mini-AMP algorithm. We analyze in detail how mini-batch learning interpolates between the purely online and the offline case. Our quantitative analysis provides a basis for an optimal choice of the mini-batch size.

Denoting the set of N parameters to be learned as the vector $\boldsymbol{x}$, we consider the streaming problem within a Bayesian framework [14]–[16]. Given a mini-batch, the posterior on $\boldsymbol{x}$ is dependent upon the posterior given the mini-batch processed before it,

$$P(\boldsymbol{x}|\underbrace{\mathcal{Y}_k, \ldots, \mathcal{Y}_1}_{\mathcal{D}_k}) = \frac{1}{\mathcal{Z}(\mathcal{Y}_k)} \prod_{\boldsymbol{y}\in\mathcal{Y}_k} P(\boldsymbol{y}|\boldsymbol{x})\, P(\boldsymbol{x}|\mathcal{D}_{k-1}), \quad (1)$$

where k denotes the mini-batch index. The posterior at the first mini-batch is given by $P(\boldsymbol{x}|\mathcal{D}_1) = \frac{1}{\mathcal{Z}(\mathcal{Y}_1)} \prod_{\boldsymbol{y}\in\mathcal{Y}_1} P(\boldsymbol{y}|\boldsymbol{x}) P_0(\boldsymbol{x})$, where $P_0(\boldsymbol{x})$ is a prior distribution on the unknown parameters. The functions $\mathcal{Z}(\cdot)$ are normalizations parameterized by the mini-batch data. For SLR, where $\boldsymbol{y} = F\boldsymbol{x} + \mathcal{N}(0, \Delta\mathbb{I}_M)$ for a matrix $F \in \mathbb{R}^{M\times N}$, we write the likelihood $P(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{F}_\mu \cdot \boldsymbol{x}, \Delta\mathbb{I}_M)$. For the perceptron, where $\boldsymbol{y} = \text{sign}(F\boldsymbol{x} + \mathcal{N}(0, \Delta\mathbb{I}_M))$, we use a probit likelihood, $P(y_\mu|z_\mu \triangleq \boldsymbol{F}_\mu \cdot \boldsymbol{x}) = \frac{1}{2}\,\text{erfc}(-\frac{y_\mu z_\mu}{\sqrt{2\Delta}})$. For the online case $m = 1$, the above definition is easily factorized and has been studied under the monikers online Bayesian learning [14] and assumed density filtering [17].

To tackle the streaming problem for $m > 1$, we turn to AMP in order to perform approximate inference of $P(\boldsymbol{x}|\mathcal{D}_k)$. We also use AMP for low-rank matrix factorization [18], [19] to investigate the case $P(U, V^{(k)}|\mathcal{D}_k) \propto \prod_{Y\in\mathcal{Y}_k}\prod_{ij} P(Y_{ij}|W_{ij} \triangleq \boldsymbol{U}_i \cdot \boldsymbol{V}_j^{(k)}) \times P(U, V^{(k-1)}|\mathcal{D}_{k-1})$, which is especially pertinent to both streaming clustering and recommender systems. Using AMP, we find the minimum mean-square-error (MMSE) estimate of $\boldsymbol{x}$ under an approximation of these posteriors.

We study the behavior of Mini-AMP for known $P_0(\boldsymbol{x})$ in the thermodynamic limit, $N \to \infty$, using state evolution and the replica free energy, both of which we adapt to the streaming setting in a novel way. These two approaches give us an understanding of the limiting performance of Mini-AMP. Analyzing Mini-AMP for both SLR and perceptron learning via these techniques, we recover a set of phase transitions over the space of relative batch sizes versus the total number of data points, shown in Figs. 1 and 2, respectively. One might intuitively assume that processing more batches is always advantageous. Indeed, for early batches, we observe an exponential decay in MSE. Surprisingly, the decay ceases to be exponential as the batch size exceeds a certain critical point, and instead we see an abrupt decline in the MSE after a few batches have been processed, behavior typical of first-order phase transitions. Using the replica free energy, we demonstrate the occurance of this phase transition as local high-MSE solutions disspear at this critical batch size, causing a precipitous jump to favorable low-MSE solutions.

Besides showing a close correspondence between random realizations of synthetic problems with finite N to our $N \to \infty$ predictions for SLR, perceptron learning, and low-rank matrix factorization, we also applied our approach to streaming clustering of real-world datasets. Since the underlying generating distributions are, depending on perspective, unknown or non-existent, we cannot exactly predict the behavior of Mini-AMP. However, as shown in Fig. 3, Mini-AMP compares favorably to state-of-the-art mini-batch K-means [20].

This research was funded by the ERC under the EU's 7th Framework Programme (FP/2007-2013/ERC Grant Agreement 307087-SPARCS).

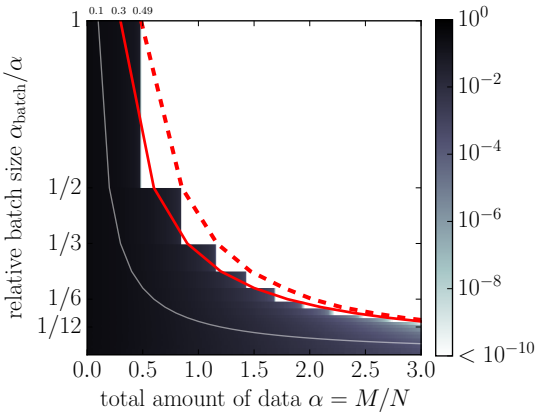

Fig. 1: MSE obtained in the SLR problem using many different batch sizes. We use a transformed y axis $\frac{\alpha_{\text{batch}}}{\alpha} = \frac{1}{k}$. Solid/dashed red lines follows the impossible-hard and hard-easy transition respectively; the white line follows the error for a fixed batch size $\alpha_{\text{batch}} = 0.1$. Here $\Delta = 0$ and $\rho = 0.3$.

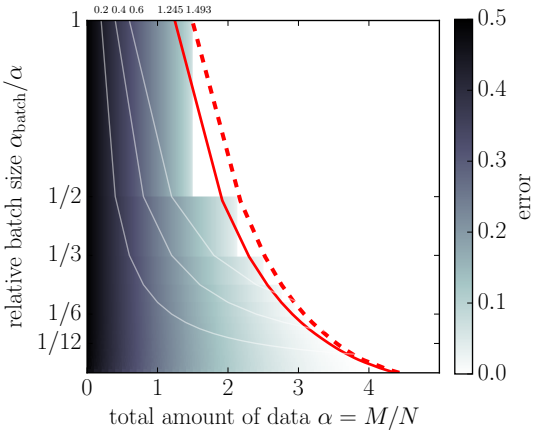

Fig. 2: Error obtained in the Ising perceptron for many different batch sizes; white lines follow the error for fixed batch sizes, solid and dashed red lines give static (impossible-hard) and dynamic (hard-easy) transition respectively. Here $\Delta = 10^{-8}$.

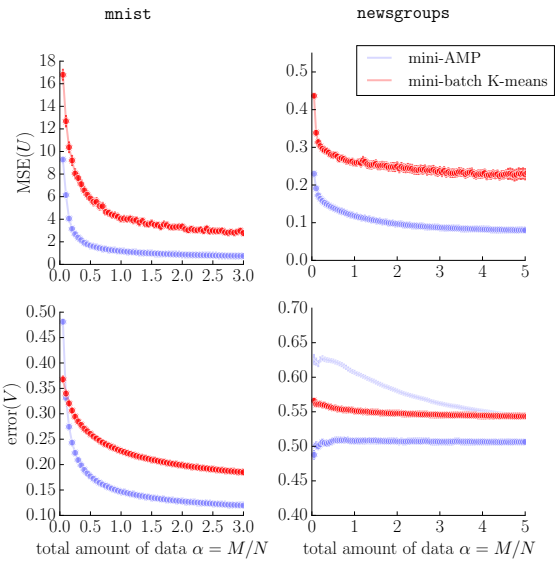

Fig. 3: Applying the mixture of Gaussians model on real data. *Left:* mean-squared error in U (centroids) and 0-1 loss in V (labels) using a batch size of $\alpha_{\text{batch}} = 0.05$ over the MNIST dataset, for clustering digits of size $N = 784$ in $K = 3$ different classes (0, 1 and 2). Blue/red circles give the performance of Mini-AMP and of the mini-batch K-means algorithms respectively. *Right:* same as left figure but over the 20 newsgroups dataset, for topic modelling using $N = 1000$ words and $K = 3$ top-level hierarchies (comp, rec and sci). Light/dark blue lines in the bottom figure give the results for the 1st and 2nd passes respectively. In both cases results are averages over 100 different orders of presentation.

References

[1] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. Conf. on Comp. Statistics*, 2010, pp. 177–186.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[3] E. Gardner, "The space of interactions in neural network models," *J. Phys. A: Math. Gen.*, vol. 21, no. 1, p. 257, 1988.

[4] H. Sompolinsky, N. Tishby, and H. S. Seung, "Learning from examples in large neural networks," *Phys. Rev. Lett.s*, vol. 65, no. 13, p. 1683.

[5] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[6] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[7] H. Nishimori and K. Y. M. Wong, "Statistical mechanics of image restoration and error-correcting codes," *Phys. Rev. E*, vol. 60, no. 1, pp. 132–144, Jul. 1999.

[8] T. L. H. Watkin, A. Rau, and M. Biehl, "The statistical mechanics of learning a rule," *Rev. Mod. Phys.*, vol. 65, no. 2, pp. 499–556, Apr. 1993.

[9] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, Nov. 2009.

[10] M. Mézard and A. Montanari, *Information, Physics, and Computation*, 1st ed. Oxford University Press, 2009.

[11] J. Ziniel, P. Schniter, and P. Sederberg, "Binary linear classification and feature selection via generalized approximate message passing," in *2014 48th Ann. Conf. on Information Sciences and Systems*, pp. 1–6.

[12] L. Zdeborová and F. Krzakala, "Statistical physics of inference: Thresholds and algorithms," *Advances in Physics*, vol. 65, no. 5, pp. 453–552, Sep. 2016, arXiv: 1511.02476.

[13] T. Lesieur, C. De Bacco, J. Banks, F. Krzakala, C. Moore, and L. Zdeborová, "Phase transitions and optimal algorithms in high-dimensional Gaussian mixture clustering," *arXiv:1610.02918 [cond-mat, stat]*.

[14] M. Opper, "A Bayesian Approach to Online Learning," in *On-line Learning in Neural Networks*, D. Saad, Ed. Cambridge, pp. 363–378.

[15] T. P. Minka, "Expectation Propagation for Approximate Bayesian Inference," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.

[16] T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan, "Streaming Variational Bayes," in *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013, pp. 1727–1735.

[17] P. S. Maybeck, *Stochastic Models, Estimation, and Control*. Academic Press, Aug. 1982.

[18] S. Rangan and A. K. Fletcher, "Iterative estimation of constrained rank-one matrices in noise," in *2012 IEEE International Symposium on Information Theory Proceedings (ISIT)*, Jul. 2012, pp. 1246–1250.

[19] T. Lesieur, F. Krzakala, and L. Zdeborová, "MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing*, pp. 680–687.

[20] D. Sculley, "Web-scale K-means Clustering," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 1177–1178.