
Gradient Energy Matching for Distributed Asynchronous Gradient Descent

Joeri R. Hermans
University of Liège
joeri.hermans@doct.uliege.be

Gilles Louppe
University of Liège
g.louppe@uliege.be

Abstract

Distributed asynchronous SGD has become widely used for deep learning in large-scale systems, but remains notorious for its instability when increasing the number of workers. In this work, we study the dynamics of distributed asynchronous SGD under the lens of Lagrangian mechanics. Using this description, we introduce the concept of energy to describe the optimization process and derive a sufficient condition ensuring its stability as long as the collective energy induced by the active workers remains below the energy of a target synchronous process. Making use of this criterion, we derive a stable distributed asynchronous optimization procedure, GEM, that estimates and maintains the energy of the asynchronous system below or equal to the energy of sequential SGD with momentum. Experimental results highlight the stability and speedup of GEM compared to existing schemes, even when scaling to one hundred asynchronous workers. Results also indicate better generalization compared to the targeted SGD with momentum.

1 Introduction

In deep learning, stochastic gradient descent (SGD) and its variants have become the optimization method of choice for most training problems. For large-scale systems, a popular variant is distributed asynchronous optimization [2, 3] based on a master-slave architecture. That is, a set of n workers w individually contribute updates $\Delta\theta_t$ asynchronously to a master node holding the central variable θ_t , under a global clock t , such that

$$\theta_{t+1} = \theta_t + \Delta\theta_t. \quad (1)$$

Due to the presence of asynchronous updates, i.e. without locks or synchronization barriers, an implicit queuing model emerges in the system [15], such that workers are updating θ_t with updates $\Delta\theta_t$ that are possibly based on a previous parameterization

$$s_t \triangleq \theta_{t-\tau_t} \quad (2)$$

of the central variable, where τ_t is the number updates that occurred between the time the worker responsible for the update at time t pulled (read) the central variable, and committed (wrote) its update. The term τ_t is traditionally called the staleness or the delay of the update at time t . Assuming a homogeneous infrastructure, the expected staleness [15] for a worker w under a simple queuing model \mathcal{Q} can be shown to be $\mathbb{E}_{\mathcal{Q}}[\tau_t] = n - 1$. In this setup, instability issues are common because updates that are committed are most likely based on past and outdated versions of the central variable, in particular as the number n of workers increases. To mitigate this effect, previous approaches [6, 9, 24] suggest to specify a projection function $\Psi(\cdot)$ that modifies $\Delta\theta_t$ such that the instability that arises from stale updates is mitigated. That is,

$$\theta_{t+1} = \theta_t + \Psi(\Delta\theta_t). \quad (3)$$

In particular, considering staleness to have a negative effect, λ -SOFTSYNC [24] and DYN SGD [9] make use of τ_t to weigh down an update $\Delta\theta_t$. While there is a significant amount of empirical

evidence that these methods are able to converge, there is also an equivalent amount of theoretical and experimental evidence that shows that parameter staleness can actually be beneficial, especially when the number of asynchronous workers is small [13, 15, 23]. Most notably, [15] identifies that asynchrony induces an implicit update momentum, which is beneficial if kept under control, but that can otherwise have a negative effect when being too strong. Clearly, these results suggest that approaches which use τ_t , or the number of workers n , are impairing the contribution of individual workers. Of course, this is not desired. As a result, asynchronous methods that are commonly used in practice are variants of DOWNPOUR [3] or HOGWILD! [17], where the number of asynchronous workers is typically restricted or the learning rate lowered to ensure stability while at the same time leveraging effects such as implicit momentum. As the number of workers increases, one might wonder what would be an effective way to measure how distributed asynchronous SGD adheres to the desired dynamics of a stable process. Answering this question requires a framework in which we are able to quantify this, along with a definition of stability and desired behavior.

Contributions Our contributions are as follows:

1. We formulate stochastic gradient descent in the context of Lagrangian mechanics and derive a sufficient condition for ensuring the stability of a distributed asynchronous system.
2. Building upon this framework, we propose a variant of distributed asynchronous SGD, GEM, that views the set of active workers as a whole and adjusts individual worker updates in order to match the dynamics of a target synchronous process.
 - Contrary to previous methods, this paradigm scales to a large number of asynchronous workers, as convergence is guaranteed by the target dynamics.
 - Speedup is attained by partitioning the effective mini-batch size over the workers, as the system is constructed such that they collectively adhere to the target dynamics.

Table 1: Notation summary

η	Learning rate
μ	Momentum term
t	Time at global (parameter server) clock.
θ_t	Central variable (central parameterization) at time t .
τ_t	Staleness of worker w responsible for update t .
s_t	The parameterization $\theta_{t-\tau_t}$ of the central variable used by worker w responsible for update t .
$\Delta\theta_t$	Update produced by worker w responsible for update t . In this work, $\Delta\theta_t$ is computed using SGD, i.e., $\Delta\theta_t = -\eta\nabla_{\theta}\mathcal{L}(s_t)$.
$ \Delta\theta_t $	Absolute value of an update $\Delta\theta_t$.
$\mathcal{E}_k(t+1)$	Kinetic energy of the proxy at time $t+1$.
$E_k(t+1)$	Kinetic energy of the central variable at time $t+1$.

2 Stability from energy

Taking inspiration from physics, the traditional explicit momentum SGD update rule can be re-expressed within the framework of Lagrangian mechanics, as shown in Equations 4 and 5, where D is Rayleigh’s dissipation function to describe the non-conservative friction force:

$$L \triangleq \frac{1}{2}\dot{\theta}^2 - \eta\mathcal{L}(\theta) \quad (4)$$

$$D \triangleq \frac{1}{2}(\mu\dot{\theta})^2 \quad (5)$$

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{\theta}} \right) - \frac{\partial L}{\partial \theta} = \frac{\partial D}{\partial \dot{\theta}} \quad (6)$$

Solving the Euler-Lagrange equation in terms of Equations 4 and 5 yields the dynamics

$$\ddot{\theta} = \mu\dot{\theta} - \eta\nabla_{\theta}\mathcal{L}(\theta), \quad (7)$$

which is equivalent to the traditional explicit momentum SGD formulation. Since the Euler-Lagrange equation can be used to derive the equations of motion using energy terms, we can similarly express these energies in an optimization setting. Using Equation 4 and discretizing the update process, we find that the kinetic energy E_k and the potential energy E_p are defined at time $t + 1$ as

$$E_k(t + 1) = \frac{1}{2} \dot{\theta}_{t+1}^2 = \frac{1}{2} (\theta_{t+1} - \theta_t)^2 \quad (8)$$

$$E_p(t + 1) = -\eta \mathcal{L}(\theta_{t+1}) \quad (9)$$

where the loss function \mathcal{L} describes the potential under a learning rate η .

Equations 8 and 9 only hold in a sequential (or synchronous) setting as the velocity term $\dot{\theta}_{t+1}$ is ill-defined in the asynchronous case. However, since asynchronous methods are concerned with the optimization of the central variable, a worker w can implicitly observe the velocity of θ_{t+1} , that is the parameter shift, between the moment it read the central variable and the moment it committed its update. In the asynchronous setting,

$$\theta_{t+1} = \theta_t + \Delta\theta_t \quad (10)$$

where $\Delta\theta_t$ was computed from $s_t = \theta_{t-\tau_t}$. Therefore, the parameter shift between the time worker w responsible for the update t read s_t and eventually committed $\Delta\theta_t$ is $\theta_{t+1} - \theta_{t-\tau_t}$. From this, we define the kinetic energy $E_k(t + 1)$ of the central variable as observed from w to produce θ_{t+1} as

$$E_k(t + 1) = \frac{1}{2} (\theta_{t+1} - \theta_{t-\tau_t})^2 = \frac{1}{2} \left(\sum_{i=0}^{\tau_t} \Delta\theta_{t-i} \right)^2. \quad (11)$$

Intuitively, our definition of the kinetic energy of the asynchronous system therefore corresponds to the collective kinetic energy induced by the last updates from every worker responsible for τ_t .

This definition allows us to measure the kinetic energy of the central variable. From this, we may now formulate a sufficient condition ensuring the stability and compliance of the asynchronous system. To accomplish this, we introduce the kinetic energy of a proxy, $\mathcal{E}_k(t + 1)$, for which we know, by assumption, that it eventually converges to a stationary solution θ^* . As a result, an asynchronous system can be said to be stable whenever

$$E_k(t + 1) \leq \mathcal{E}_k(t + 1), \quad (12)$$

and compliant when

$$E_k(t + 1) = \mathcal{E}_k(t + 1). \quad (13)$$

That is, stability and compliance are ensured whenever the kinetic energy collectively induced by the set of workers matches with the kinetic energy of a proxy known to converge. Therefore, equations 12 and 13 are sufficient to ensure convergence, albeit not necessary.

3 Gradient Energy Matching

We now present a new variant of distributed asynchronous SGD, Gradient Energy Matching (GEM), which aims at satisfying the compliance condition with respect to the dynamics of a target synchronous process, thereby ensuring its stability. In addition, we design GEM so as to minimize computation at the master node, thereby reducing the risk of bottleneck.

3.1 Proxy

The definition of the proxy can be chosen arbitrarily, as it depends on the desired dynamics of the central variable. In this work, we target a proxy for which the dynamics of θ_t behaves similarly to what would be computed sequentially in SGD with an explicit momentum term μ under a static learning rate η . Traditionally, this is written as

$$\Delta\theta_t = \mu\Delta\theta_{t-1} - \eta\nabla_{\theta}\mathcal{L}(\theta_t), \quad (14)$$

with the corresponding kinetic energy

$$\mathcal{E}_k(t + 1) \triangleq \frac{1}{2} (\mu\Delta\theta_{t-1} - \eta\nabla_{\theta}\mathcal{L}(\theta_t))^2. \quad (15)$$

3.2 Energy matching

Consider the situation of a small number of asynchronous workers, as is typically the case for DOWNPOUR or HOGWILD!. Experimental results reported in the literature suggest that these methods usually converge to a stationary solution, which implies that the stability condition (Equation 12) is most likely satisfied. However, as the number of asynchronous workers increases, so does the kinetic energy of the central variable and there is a limit beyond which the system dynamics do no longer satisfy Equation 12. To address this, we introduce scaling factors π_t , whose purpose are to modify updates $\Delta\theta_t$ such that the compliance condition remains satisfied. As a result, the central variable energy (Equation 11) is extended to

$$E_k(t+1) \triangleq \frac{1}{2}(\theta_t - s_t + \pi_t \Delta\theta_t)^2, \quad (16)$$

Scaling factors π_t can be assigned in several ways, depending on the information available to each worker w and on the desired usage of the individual updates $\Delta\theta_t$. To prevent wasting computational resources, we aim at equally maximizing the effective contribution of all updates $\Delta\theta_t$ while achieving compliance with the proxy, i.e.,

$$\mathcal{E}_k(t+1) = E_k(t+1), \quad (17)$$

$$= \frac{1}{2}(\theta_t - s_t + \pi_t \Delta\theta_t)^2. \quad (18)$$

Directly solving for π_t yields

$$\pi_t = \frac{\sqrt{2}\sqrt{\mathcal{E}_k(t+1)} - |\theta_t - s_t|}{|\Delta\theta_t|}. \quad (19)$$

This mechanism constitutes the core of GEM, as it is through scaling factors π_t that the energy of the active workers is matched to the energy of the target proxy.

To show the reader how π_t adjusts the active workers to collectively adhere to the proxy, let us consider the quantity $|\theta_t - s_t|$. From [15], we know that under a homogeneity assumption the average staleness is $\mathbb{E}_Q[\tau] = n - 1$. In expectation, the behavior of the system is therefore similar to updates occurring in a round-robin fashion. In this case, Equation 11 reduces to

$$E_k(t+1) \triangleq \frac{1}{2} \left(\sum_{i=0}^{n-1} \Delta\pi_{t-i} \theta_{t-i} \right)^2, \quad (20)$$

$$\equiv \frac{1}{2} \left(\pi_t \Delta\theta_t + \sum_{i=1}^{n-1} \pi_{t-i} \Delta\theta_{t-i} \right)^2, \quad (21)$$

$$\equiv \frac{1}{2} (\pi_t \Delta\theta_t + |\theta_t - s_t|)^2, \quad (22)$$

which shows that to ensure compliance, the updates of all workers have to collectively be scaled.

3.3 Proxy estimation

While the previous derivation defines an assignment for π_t , it remains to locally estimate the desired kinetic energy $\mathcal{E}_k(t+1)$. This is particularly problematic since updates $\Delta\theta_t$ from the other workers are not available. As an approximation, we propose to aggregate local updates such that

$$\mathcal{E}_k(t+1) \triangleq \frac{1}{2} (\mu m_{t-\tau_t} - \eta \nabla_{\theta} \mathcal{L}(s_t))^2, \quad (23)$$

with the first moment m_t being defined as

$$m_t \triangleq \mu m_{t-\tau_t} - \eta \nabla_{\theta} \mathcal{L}(s_t). \quad (24)$$

Additionally, Appendix B examines the effects of scaling $\mathcal{E}_k(t+1)$ with an amplification factor to compensate for the staleness of the proxy due to the local approximation.

This concludes GEM, as summarized in Algorithm 1.

Algorithm 1 Worker procedure of Gradient Energy Matching.

Hyper-parameters: learning rate $\eta = 0.01$, momentum $\mu = 0.9$, number of asynchronous workers n , and $\epsilon = 10^{-16}$ to prevent division by zero.

<pre>1: procedure $\Psi_{\text{GEM}}(\Delta\theta)$ 2: $m \leftarrow \mu m + \Delta\theta$ 3: $\pi \leftarrow \frac{ m - \theta - s }{ \Delta\theta_t + \epsilon}$ 4: return $\pi \odot \Delta\theta$ 5: end procedure</pre>	<pre>1: procedure WORKER 2: $m \leftarrow 0$ 3: $s \leftarrow \text{PULL}()$ 4: while not converged do 5: $\Delta\theta \leftarrow -\eta \nabla_{\theta} \mathcal{L}(s)$ 6: $\theta \leftarrow \text{PULL}()$ 7: $\text{COMMIT}(\Psi_{\text{GEM}}(\Delta\theta))$ 8: $s \leftarrow \theta$ 9: end while 10: end procedure</pre>
--	---

4 Experiments

We evaluate GEM against DOWNPOUR [3] and an adaptive staleness technique which rescales updates $\Delta\theta_t$ by $1/(\tau_t + 1)$, similar to what is typically applied to stabilize stale gradients in λ -SOFTSYNC [24] or DYN SGD [9]. We carry out experiments on MNIST [14] and ImageNet [19] to illustrate the effectiveness and robustness of our method. All algorithms are implemented in PyTorch [18] using the distributed sub-module. The source code ¹ we provide is to the best of our knowledge the first PyTorch-native parameter server implementation. Experiments described in this section are carried out on general purpose CPU machines without using GPGPU accelerators.

4.1 Illustrative experiments

Setup For all experiments described below we train a convolutional network (see Appendix A for specifications) on MNIST using GEM, DOWNPOUR or the adaptive staleness variant on a computer cluster as an MPI job. A fixed learning rate $\eta = 0.05$, momentum term $\mu = 0.9$ and worker batch-size $b = 64$ are applied unless stated otherwise. Additionally, every worker contributes a single epoch of training data. All results in this section is the average of 3 consecutive runs.

Stability Figure 1 illustrates the training loss of GEM compared to DOWNPOUR and adaptive staleness. We observe that GEM remains stable and adheres to the target dynamics as designed, even when increasing the number of asynchronous workers. This shows that GEM is able to cope with the induced staleness. Remarkably, GEM remains stable even when scaling to 100 asynchronous workers, as shown in Figure 2. By contrast, DOWNPOUR faces strong convergence issues when increasing the number of workers while adaptive staleness shows impaired efficiency as updates are rescaled proportionally to the staleness, which is often too aggressive.

Update rescaling Figure 1a shows the behavior of GEM for an increasing number of workers and a fixed worker batch-size b . When applying an asynchronous method, one would expect that the number of updates per worker to reach a given loss value would decrease since more updates are committed in total to the master node. However, since GEM tries to make the workers collectively adhere to the dynamics that are expressed by the proxy, this behavior should not be expected. Indeed, Figure 1a shows loss profiles that are similar whereas the number of workers varies from 5 to 30. As designed, inspecting the median per-parameter value of π , shown in Figure 3, reveals that the scaling factors steadily decrease whenever additional workers are added. Notwithstanding, we also do observe improvement as we increase the number of workers as it indirectly augments the effective batch size B . Furthermore, we observe from Figure 3 the adaptivity of GEM as the median π changes with time. While this behavior may appear as a limitation, better efficiency can be achieved by simply considering proxies of different energies, as further discussed in Appendix B.

Wall-clock time speedup In distributed synchronous optimization, wall-clock speedup is typically achieved by setting a desired effective batch-size B , and splitting the computation of the batch

¹<https://github.com/montefiore-ai/gradient-energy-matching>

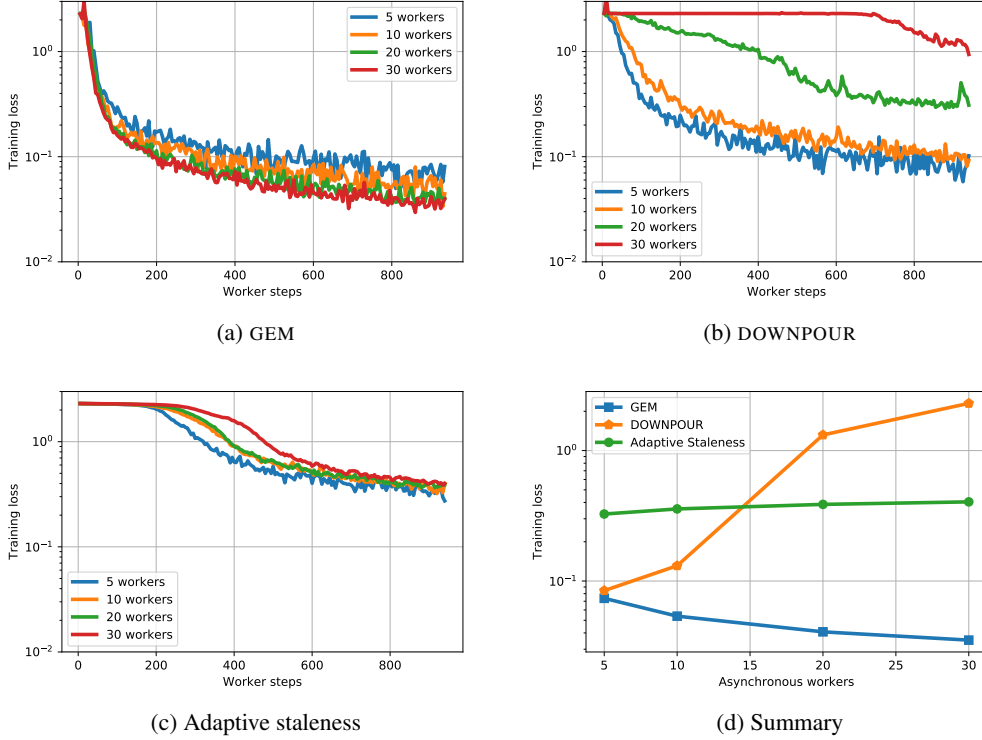


Figure 1: (a) Training loss of GEM for different number of workers. We observe that the optimizer adheres to the specified target dynamics, and that the variance of the loss becomes smaller when the number of workers increases. (b) DOWNPOUR on the other hand quickly shows suboptimal or even divergent behavior when the number of workers is increased. (c) The adaptive staleness technique is slower since the updates are scaled proportionally to their staleness τ . (d) Final achieved training loss given the number of workers, which highlights the stability of GEM.

over n workers to produce updates $\Delta\theta_t$ computed individually over $b = B/n$ samples. In the asynchronous setting, the same strategy can be applied for GEM to speed up the training procedure, as shown in Figure 4 for a fixed effective batch-size $B = 256$, a number of workers n between 1 and 16 and a corresponding worker batch size b between 256 and 16. As expected, and without loss in accuracy, we observe that wall-clock speedup is inversely proportional to b , depending on the efficiency of the implementation. However, we also observe larger variability as n increases due to the increased stochasticity of the updates.

Generalization A common argument against large mini-batch SGD is the poor generalization performance that is typically associated with the resulting model due to easily obtainable sharp minima [1, 7, 8, 10]. Nevertheless, [5, 11, 21] show that test accuracy is not significantly impacted when using larger learning rates to emulate the regularizing effects of smaller batches. Similarly, experimental results for GEM do not show such behavior. Even when training with 100 asynchronous workers, as done in Figure 2 and which corresponds to $B = nb = 100 \times 64$, GEM obtains a final test accuracy of 99.38%. By contrast, the targeted momentum SGD procedure starts to overfit, as shown in Figure 5. A possible intuitive explanation for this behavior is most in-line with the arguments presented in [1, 10], i.e., that asynchrony (staleness) regularizes the central variable such that it is not attracted to sharper minima, therefore preferring wider basins.

4.2 Resilience against hyper-parameter misspecification

These experiments use the same setup as in the previous MNIST experiments, with the exception of the learning rate, which is intentionally set to a large value $\eta = 0.1$. To show that our method is able to adhere to the proxy dynamics, we compare GEM against DOWNPOUR. From Figure 6, we

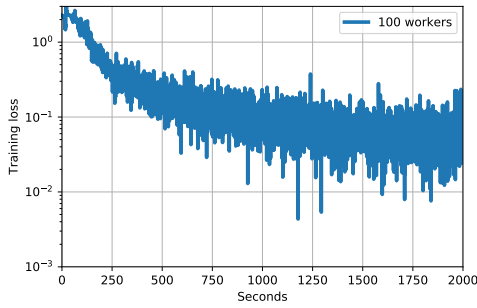


Figure 2: Training loss at the master node for 100 asynchronous workers. Even at that scale, GEM does not show signs of instability.

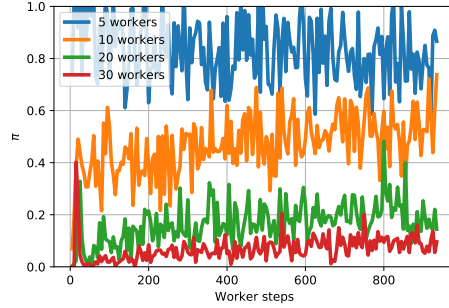


Figure 3: Change over time of the median value of the scaling factors π .

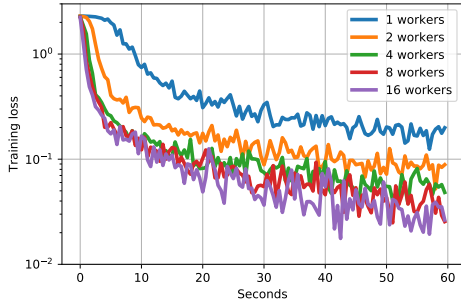


Figure 4: In GEM, wall-clock speedup is obtained by splitting the effective batch-size over n workers, as is done in synchronous methods.

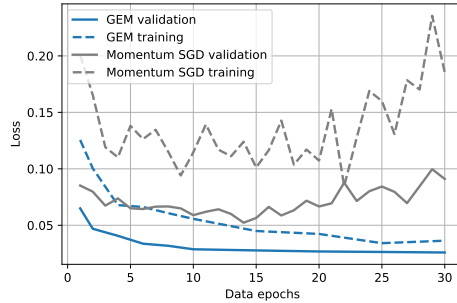


Figure 5: Training and validation losses for GEM and the proxy. Contrary to momentum SGD, GEM does not show signs of overfitting.

observe that DOWNPOUR is not able to cope with the increased learning rate, while GEM remains stable since the equivalent synchronous process converges as well under these settings.

4.3 ImageNet

We train AlexNet [12] on the ImageNet data [19] with an effective batch size of $B = 1024$ and $b = B/n$. GEM hyper-parameters are set to their default values, as specified in Algorithm 1. No learning rate scheduling is applied. All experiments ran for 24 hours with 8 and 16 asynchronous workers, each using 2 CPU cores due to cluster restrictions. A single core computes the gradients, while the other handles data-prefetching, data-processing, and logging. The results are summarized in Figure 7. The figure shows that GEM is able to effectively handle the staleness on a realistic use-case. Despite an identical effective batch size, we observe a significant improvement with regards to worker update efficiency, both in wall-clock time and the training loss. The reason for this effect remains to be studied. Figures 9 and 10 in supplementary materials show how individual tensors of an update are adjusted over time.

5 Related work

Many techniques [2, 6, 9, 15, 16, 20, 24] have been proposed to stabilize distributed asynchronous SGD based on the perception that staleness usually has a negative effect on the optimization process. In [15] however, the main inspiration behind this work, staleness is shown to be equivalent to an asynchrony-induced implicit momentum, which thereby explains its negative effect when being too strong but also highlights its potential benefits when kept under control. Similarly, in our framework, implicit momentum translates to added kinetic energy of the central variable. Close to our work, the authors of YELLOWFIN [23] build on top of [15] by proposing an adaptive scheme that tunes momentum dynamically. They show in an asynchronous setting how individual workers can adjust

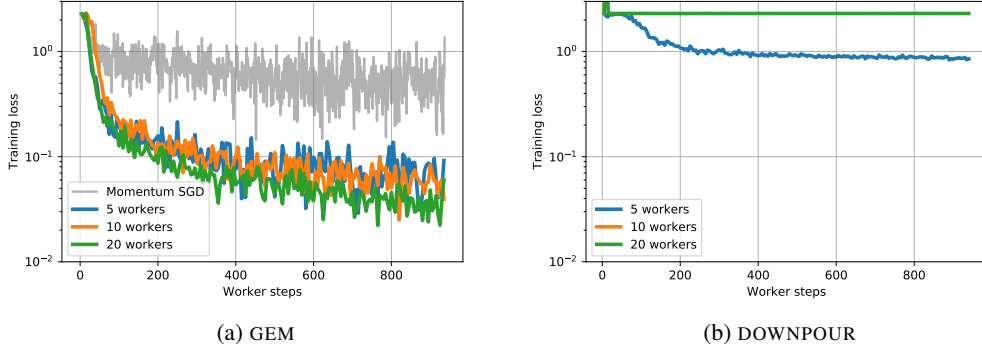


Figure 6: Training curves for GEM and DOWNPOUR when the learning rate is misspecified to $\eta = 0.1$. DOWNPOUR clearly shows suboptimal behavior while GEM and the target synchronous momentum SGD proxy converge without issues.

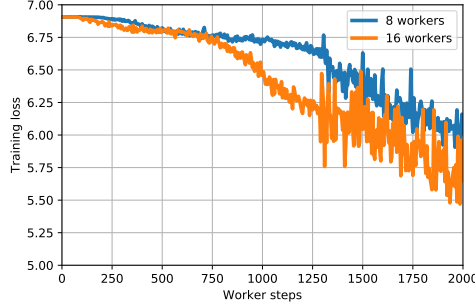


Figure 7: ImageNet experiments. All runs have a fixed time-budget of 24 hours. Training loss with respect to the worker steps. We observe a significant improvement in update efficiency when doubling the number of workers, while halving b .

momentum terms to collectively fit a target momentum provided by the optimizer. In particular, this is achieved by incorporating a negative feedback loop that attempts to estimate the momentum term μ from the observed staleness. We note the opportunity to extend this strategy in our framework, by using YELLOWFIN to estimate the optimal momentum term of the proxy, and have GEM acting as a mechanism to tune individual workers, thereby removing the original feed-back loop.

6 Conclusions

This work introduces Gradient Energy Matching, GEM, a novel algorithm for distributed asynchronous SGD. The core principle of GEM is to consider the set of active workers as a whole and to make them collectively adhere to the dynamics of a target synchronous process. More specifically, the estimated energy of the asynchronous system is matched with the energy of the target proxy by rescaling the gradient updates before sending them to the master node. Under the assumption that this proxy converges, the proposed method ensures the stability of the collective asynchronous system, even when scaling to a high number of workers. Wall-clock time speedup is attained by partitioning the effective mini-batch size across workers, while maintaining its stability and generalizability.

A direction for future theoretical work would be the study of the effects of asynchrony on the generalization performance in connection with [4] and how GEM differs from regular momentum. This could help explain the unexpected improvement observed in the ImageNet experiments. A further addition to our method would be to adjust the way the proxy is currently estimated, i.e., one could compute the proxy globally instead of locally. This would allow all workers to contribute informa-

tion to produce better estimates of the (true) proxy. As a result, the scaling factors that would be applied to all worker updates would in turn be more accurate.

Acknowledgements

Joeri Hermans acknowledges the financial support from F.R.S-FNRS for his FRIA PhD scholarship. We are also grateful to David Colignon for his technical help.

References

- [1] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *arXiv preprint arXiv:1611.01838*, 2016.
- [2] Trishul M Chilimbi, Yutaka Suzue, Johnson Apacible, and Karthik Kalyanaraman. Project adam: Building an efficient and scalable deep learning training system. In *OSDI*, volume 14, pages 571–582, 2014.
- [3] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Andrew Senior, Paul Tucker, Ke Yang, Quoc V Le, et al. Large scale distributed deep networks. In *Advances in neural information processing systems*, pages 1223–1231, 2012.
- [4] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. *arXiv e-prints*, 1703.04933, March 2017. URL <https://arxiv.org/abs/1703.04933>.
- [5] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.
- [6] J. Hermans, G. Spanakis, and R. Möckel. Accumulated Gradient Normalization. *ArXiv e-prints*, October 2017.
- [7] Sepp Hochreiter and Jürgen Schmidhuber. Simplifying neural nets by discovering flat minima. In *Advances in neural information processing systems*, pages 529–536, 1995.
- [8] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.
- [9] Jiawei Jiang, Bin Cui, Ce Zhang, and Lele Yu. Heterogeneity-aware distributed parameter servers. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 463–478. ACM, 2017.
- [10] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.
- [11] Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014.
- [12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [13] Thorsten Kurth, Jian Zhang, Nadathur Satish, Evan Racah, Ioannis Mitliagkas, Md Mostofa Ali Patwary, Tareq Malas, Narayanan Sundaram, Wahid Bhimji, Mikhail Smorkalov, et al. Deep learning at 15pf: supervised and semi-supervised classification for scientific data. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, page 7. ACM, 2017.
- [14] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. <http://yann.lecun.com/exdb/mnist/>, 2010. URL <http://yann.lecun.com/exdb/mnist/>.

- [15] I. Mitliagkas, C. Zhang, S. Hadjis, and C. Ré. Asynchrony begets Momentum, with an Application to Deep Learning. *ArXiv e-prints*, May 2016.
- [16] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1928–1937, 2016.
- [17] F. Niu, B. Recht, C. Re, and S. J. Wright. HOGWILD!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent. *ArXiv e-prints*, June 2011.
- [18] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch, 2017.
- [19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.
- [20] Tom Schaul and Yann LeCun. Adaptive learning rates and parallelization for stochastic, sparse, non-smooth gradients. *arXiv preprint arXiv:1301.3764*, 2013.
- [21] Samuel L Smith, Pieter-Jan Kindermans, and Quoc V Le. Don’t decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*, 2017.
- [22] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=2627435.2670313>.
- [23] J. Zhang, I. Mitliagkas, and C. Ré. YellowFin and the Art of Momentum Tuning. *ArXiv e-prints*, June 2017.
- [24] W. Zhang, S. Gupta, X. Lian, and J. Liu. Staleness-aware Async-SGD for Distributed Deep Learning. *ArXiv e-prints*, November 2015.

A Model specification

The convolutional network in the MNIST experiments is fairly straightforward. It consists of 2 convolutional layers, each having a kernel size of 5 with no padding, with stride and dilation set to 1. The first convolutional layer has 1 input channel, and 10 output channels, whereas the second convolutional layer has 20 output channels, both with the appropriate max-pooling operations. We apply dropout [22] using the default settings provided by PyTorch [18] on the final convolutional layer. Finally, we flatten the filters, and pass them through 3 fully connected layers with ReLU activations followed by a logarithmic softmax.

B High-energy proxies

Depending on the definition of the proxy, the efficiency of the optimization procedure can be increased by adjusting its energy by a constant factor κ . More specifically, to increase the permissible energy in GEM proportionally to κ , we simply redefine the proxy energy to

$$\mathcal{E}_k(t+1) \triangleq \frac{\kappa}{2} (\mu m_{t-\tau_t} + \Delta\theta_t)^2. \quad (25)$$

From this definition it is clear that κ amplifies the energy of the proxy, therefore allowing for larger worker contributions as the range of the proxy energy is extended. As π_t depends on the proxy, we can directly include κ in π_t as

$$\pi_t = \frac{\sqrt{\kappa}|m| - |\theta - s|}{|\Delta\theta_t| + \epsilon}. \quad (26)$$

Since gradually increasing κ does not have a significant effect on m (because of $\sqrt{\kappa}$), we modify Equation 26 for the purpose of this study to

$$\pi_t = \frac{\kappa|m| - |\theta - s|}{|\Delta\theta_t| + \epsilon}. \quad (27)$$

This effect can be observed from Figure 8, where the training loss improves as κ increases. However, we would like to note that increasing the proxy energy can have adverse effects on the convergence of the training as the energy levels expressed by the proxy is reaching the divergence limit.

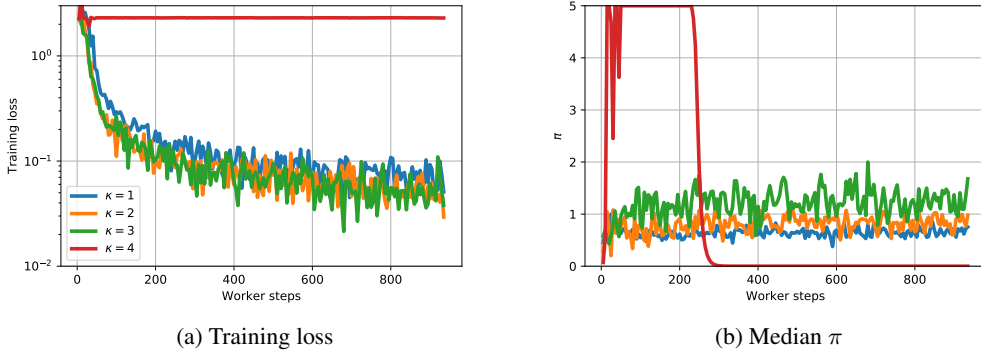


Figure 8: Effect of κ on the convergence of rate the central variable in GEM. Increasing the energy shows better convergence rates. However, there exists a certain limit as discussed earlier in Section 2 when GEM starts to show suboptimal and even divergent behavior (for $\kappa = 4$). This indicates that the energy expressed by the proxy is approaching the unknown divergence limit.

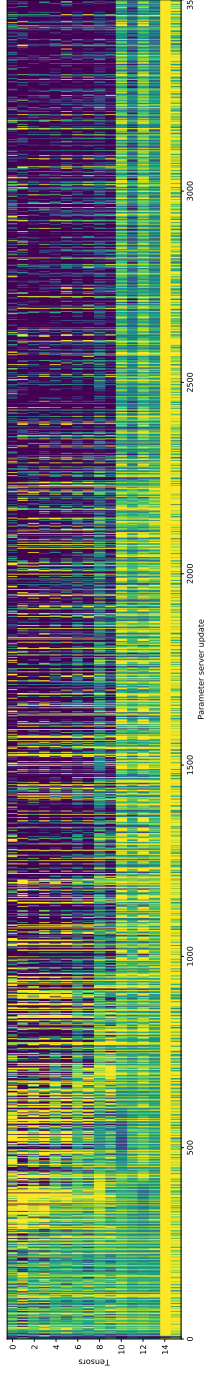


Figure 9: ImageNet experiment, for 8 asynchronous workers. The figure shows how individual AlexNet tensors are adjusted by GEM. Individual cells show the median scaling term π_t of that particular tensor. Brighter colors indicate large scaling values.

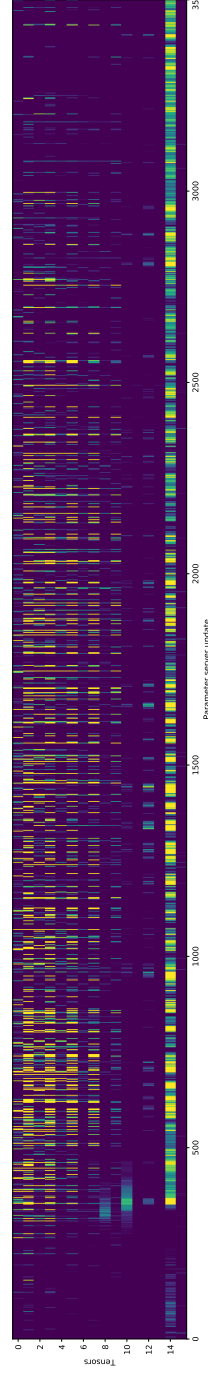


Figure 10: ImageNet experiment, for 16 asynchronous workers. Contrary to Figure 9, median values for π_t are significantly smaller. This is expected as this configuration is obtained by the double amount of workers. Nevertheless, it remains to be investigated.