OPT2025: 17th Annual Workshop on Optimization for Machine Learning

Convergence for Discrete Parameter Update Schemes

Paul Wilson [*] [†] PAUL@HELLAS.AI
Fabio Zanasi [†] F.ZANASI@UCL.AC.UK
George Constantinides G.CONSTANTINIDES@IMPERIAL.AC.UK

Abstract

Modern deep learning models require immense computational resources, motivating research into low-precision training. Quantised training addresses this by representing training components in low-bit integers, but typically relies on discretising real-valued updates. We introduce an alternative approach where the update rule itself is discrete, avoiding the quantisation of continuous updates by design. We establish convergence guarantees for a general class of such discrete schemes, and present a multinomial update rule as a concrete example, supported by empirical evaluation. This perspective opens new avenues for efficient training, particularly for models with inherently discrete structure.

1. Introduction

State-of-the-art deep learning models comprise hundreds of billions of parameters and require sizeable computational and memory resources. As models continue to scale, hardware constraints become a central challenge, motivating research into low-precision computation and memory-efficient designs. Among such approaches, quantised training [2, 6, 7] seeks to train models using low-precision numerical representations (e.g. 8-bit, 4-bit, or even 2-bit integers) for various components of training. Recent work has demonstrated that quantised training is feasible at scale [3, 5, 11], with stability issues addressed through a variety of techniques.

A common trait of these approaches is that updates are first computed in real values and then discretised via a quantisation function, usually based on rounding (e.g. round-to-nearest or stochastic rounding). In this paper, we propose an alternative route: instead of discretising continuous updates, we assume from the outset that the update function itself is discrete (integer-valued). Our broader motivation is to identify models where directly discrete updates offer greater efficiency than quantisation. We expect such cases to arise in particular when training non-real-valued systems such as arithmetic or Boolean circuits, as developed in [13, 14], and in binarised neural networks [12]. A more systematic exploration of such schemes is left for future work; in this paper, we focus on laying the mathematical foundations of our approach.

Perhaps the work closest in spirit to ours is BOLD [10]: it introduces a fully-discrete update scheme that completely eliminates the need for floating point latent values, yielding a significant savings in memory requirements during both training and inference. BOLD relies on a bespoke backpropagation framework tailored to boolean architectures, with convergence analysis specific to that setting. By contrast, our aim is to provide a general framework for discrete updates under rela-

[*] I thank Diptarko Roy for helpful discussions.
[†] The first two authors acknowledge support from ARIA's Safeguarded AI programme

tively weak assumptions; these encompass boolean-specific methods such as BOLD, while opening up a broader space of discrete training schemes. We return to the comparison in Section 3.

Our contribution is structured as follows. Section 2 introduces the assumptions underlying discrete update schemes and presents our main mathematical result: a convergence theorem. Section 3 provides a concrete example of such a scheme, which we then evaluate empirically in Section 4.

2. Discrete Gradient Updates and their Convergence

The standard stochastic gradient update may be written with learning rate sequence α_k, update function g and random variable ξ_k, as

$$w_{k+1} \leftarrow w_k - \alpha_k \cdot g(w_k, \xi_k).$$

In contrast, we seek an update that can be computed with simple, performant fixed-precision operations. We propose the following scheme.

Definition 1 (Discrete Stochastic Gradient Update)

$$w_{k+1} \leftarrow w_k - \bar{g}(w_k, \xi_k)$$

where $\bar{g} : \mathbb{R}^d \to S^d \subseteq \mathbb{Z}^d$ is a 'discrete' (i.e., integer-valued) function.

We now focus on showing convergence of such scheme. It is convenient to follow the framework of the work [1], which focuses on convergence of (standard) stochastic gradient update. In a nutshell, we 'discretise' the assumptions used in [1], and follow a similar proof strategy to demonstrate convergence of discrete gradient update. The first requirement is Lipschitz continuity of gradients of the objective function F (*cf.* Assumption 4.1, [1]), as follows.

Assumption 2 (Lipschitz-Continuous objective gradients) *The objective function $F : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and the gradient function of F, namely $\nabla F : \mathbb{R}^d \to \mathbb{R}^d$, is Lipschitz continuous with Lipschitz constant $L > 0$, i.e.*

$$\|\nabla F(w) - \nabla F(\bar{w})\|_2 \leq L \|w - \bar{w}\|_2$$

for all $w, \bar{w} \in \mathbb{R}^d$

This implies the following inequality, shown in [1, (4.3), Appendix B]).

$$F(w) \leq F(\bar{w}) + \nabla F(\bar{w})^T (w - \bar{w}) + \frac{1}{2} L \|w - \bar{w}\|_2^2 \tag{2.1}$$

The second assumption poses requirements on the update.

Assumption 3 (Update bound) *Let $\bar{g}$ be a discrete gradient update as defined in Definition 1, α_k a sequence of learning rates, and F an objective function satisfying Assumption 2.*

(a) The sequences of iterates w_k is contained in an open set over which F is bounded below by a scalar F_{inf} [1, Assumption 4.3 (a)]

(b) There exist scalars $\mu > 0$, $M \geq 0$, and $M_G > 0$ such that:

$$\nabla F(w_k)^T \mathbb{E}_{\xi_k}[\bar{g}(w_k, \xi_k)] \geq \alpha_k \mu \left\|\nabla F(w_k)\right\|_2^2 \tag{2.2}$$

$$\mathbb{E}_{\xi_k}\left[\left\|\bar{g}(w_k, \xi_k)\right\|_2^2\right] \leq \alpha_k M + \alpha_k^2 M_G \left\|\nabla F(w_k)\right\|_2^2 \tag{2.3}$$

Conditions (a) and (b) are similar to [1, Assumption 4.3], except in (b) we directly assume a bound on the second moment instead of deriving it from bounds on the 2-norm and variance. This will allow our proof of convergence to closely mirror the approach in [1, Theorem 4.8]. We are now ready to state our convergence result under these assumptions.

Proposition 4 *Let F be a (possibly non-convex) function, and fix $\alpha_k = \bar{\alpha}$ for all $k \in \mathbb{N}$ where*

$$0 < \bar{\alpha} \leq \frac{\mu}{LM_G}$$

Then we have

$$\begin{aligned}\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F(w_k)\right\|_2^2\right] &\leq \frac{2(F(w_1) - F_{\mathsf{inf}})}{K\mu\bar{\alpha}} + \frac{LM}{\mu} \\ &\xrightarrow{K\to\infty} \frac{LM}{\mu}\end{aligned}$$

Proof From Proposition 9 (Appendix A) and by assumption on $\bar{\alpha}$ we have

$$\begin{aligned}\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) &\leq -(\mu - \frac{1}{2}\bar{\alpha}LM_G)\bar{\alpha}\left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\bar{\alpha}LM && \text{(Proposition 9)} \\ &\leq -\frac{1}{2}\mu\bar{\alpha}\left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\bar{\alpha}LM && \text{(Assumption on } \bar{\alpha}\text{)}\end{aligned}$$

Taking the total expectation, we obtain

$$\mathbb{E}\left[F(w_{k+1})\right] - \mathbb{E}\left[F(w_k)\right] \leq -\frac{1}{2}\mu\bar{\alpha}\mathbb{E}\left[\left\|\nabla F(w_k)\right\|_2^2\right] + \frac{1}{2}\bar{\alpha}LM$$

Now telescoping the sum for $k \in \{1 \dots K\}$,

$$\begin{aligned}F_{\mathsf{inf}} - F(w_1) &\leq \mathbb{E}\left[F(w_{K+1})\right] - F(w_1) \\ &\leq -\sum_{k=1}^{K}\left(\frac{1}{2}\mu\bar{\alpha}\mathbb{E}\left[\left\|\nabla F(w_k)\right\|_2^2\right] + \frac{1}{2}\bar{\alpha}LM\right) \\ &= -\frac{1}{2}\mu\bar{\alpha}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(w_k)\right\|_2^2\right] + \frac{1}{2}\bar{\alpha}KLM\end{aligned}$$

Now rearranging,

$$\begin{aligned}
\frac{1}{2}\mu\bar{\alpha}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] &\leq F(w_1) - F_{\mathsf{inf}} + \frac{1}{2}\bar{\alpha}KLM \\
\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] &\leq \frac{2(F(w_1) - F_{\mathsf{inf}})}{\mu\bar{\alpha}} + \frac{KLM}{\mu} && \text{(Divide by } \tfrac{1}{2}\mu\alpha\text{)} \\
\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(w_k)\|_2^2\right] &\leq \frac{2(F(w_1) - F_{\mathsf{inf}})}{K\mu\bar{\alpha}} + \frac{LM}{\mu} && \text{(Divide by } K\text{, } \mathbb{E} \text{ Linearity)} \\
&\xrightarrow{K\to\infty} \frac{LM}{\mu}
\end{aligned}$$

■

Note the term $\frac{LM}{\mu}$ does not involve the learning rate $\bar{\alpha}$, in contrast to [1, Theorem 4.8]. This is the error bound resulting from the use of discrete weights, as common in other work on quantized learning [8–10].

3. Example: Multinomial

We now study a specific discrete gradient update $\bar{g}$ satisfying Assumption 3. Briefly, the idea is that $\bar{g}(w_k, \xi_k)$ samples from a 'zero-inflated' multinomial distribution whose probabilities are proportional to the gradient vector, while the 'zero inflation' incorporates a learning rate.

Definition 5 (Zero-inflated multinomial) *Let $n \in \mathbb{N}$ represent a number of trials, $0 \leq r \leq 1$ a probability, and $q \in \mathbb{R}^d$ be a discrete probability distribution. Write $y \sim \mathsf{ZIMultinomial}(n, r, q)$ to mean the random variable equal to $x_{1:d}$, where $x \sim \mathsf{Multinomial}(n, p)$, and*

$$\begin{aligned}
p_0 &= 1 - r \\
p_i &= rq_i \text{ for } i \in \{1..d\}
\end{aligned}$$

Note that $\sum_{i=0}^{d} p_i = p_0 + \sum_{i=1}^{d} p_i = (1-r) + r\sum_{i=1}^{d} q_i = 1$ and so p indeed defines a discrete probability distribution. The first and second moments of the zero-inflated multinomial are given in Proposition 10. We can now define a discrete update $\bar{g}$ using the ZIMultinomial.

Definition 6 (ZIM update) *Fix constants $0 \leq r \leq 1$ and $c > 0$. Define $\bar{g}(w_k, \xi_k) := x \odot \operatorname{sign}(\nabla F(w_k))$ where:*

$$\begin{aligned}
x &\sim \mathsf{ZIMultinomial}_{\xi_k}(n, r, q) \\
q &:= \frac{|\nabla F(w_k)| + c}{\sum_{i=1}^{d}(|\nabla F(w_k)| + c)} = \frac{|\nabla F(w_k)| + c}{\|\nabla F(w_k)\|_1 + cd}
\end{aligned}$$

The (non-zero-inflated) probabilities q are designed such that the expectation of $\bar{g}$ is proportional to the gradient. The dimension of the weights is denoted d, and c is a Laplace-smoothing-like factor ensuring that q is defined when $w_k = 0$. The ZIM update satisfies Assumption 3 as follows.

Proposition 7 *for ZIM update* $\bar{g}(w_k, \xi_k)$ *there exist scalars* $\mu = \frac{n}{\sqrt{d}L+cd}$, $M = n$, *and* $M_G = n^2 - n$ *such that:*

$$\nabla F(w_k)^T \mathbb{E}_{\xi_k}[\bar{g}(w_k, \xi_k)] \geq \alpha_k \mu \left\|\nabla F(w_k)\right\|_2^2$$

and

$$\mathbb{E}_{\xi_k}\left[\left\|\bar{g}(w_k, \xi_k)\right\|_2^2\right] \leq \alpha_k M + \alpha_k^2 M_G \left\|\nabla F(w_k)\right\|_2^2$$

Proof Propositions 12 and 11 (Appendix B) give μ and α_k, M, M_G, respectively. ■

We can therefore specialise Proposition 4 to prove convergence of the ZIM update: Recall that

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F(w_k)\right\|_2^2\right] \leq \frac{2(F(w_1) - F_{\mathsf{inf}})}{K\mu\bar{\alpha}} + \frac{LM}{\mu} \xrightarrow{K\to\infty} \frac{LM}{\mu}$$

Then taking $r = \bar{\alpha}$ and substituting $\mu = \frac{n}{\sqrt{d}L+cd}$, $M = n$

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F(w_k)\right\|_2^2\right] \xrightarrow{K\to\infty} \frac{Ln}{\frac{n}{\sqrt{d}L+cd}} = L(\sqrt{d}L + cd)$$

Compare this to the discrete update used in BOLD [10], whose convergence bound is proportional to dL versus our $\sqrt{d}L^2$. Our bound is tighter when the Lipschitz constant L grows slowly relative to model dimension d. As a concrete example, take $L = 17$ from [4, Figure 3(a)] for the single layer MNIST network with 100 hidden units, then calculate our bound as 1.4×10^6 compared to 2.7×10^6 for BOLD.

4. Empirical Evaluation

We conclude with an empirical[1] demonstration of convergence for our approach, compared to Stochastic Gradient Descent (SGD). It is clear that both the simple convolutional and larger ResNet models converge. Although our method pays a $0.5\% - 1\%$ accuracy penalty (reflecting the 'noise floor' discussed in Section 2), it works 'out of the box' with existing architectures. We view these results as a proof of concept. Going forward, we expect that not only can significant improvements be made by modifying the architectures and update schemes used, but also that our method opens the door to the possibility of fully discrete learning systems.

References

[1] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning, 2018. URL https://arxiv.org/abs/1606.04838.

[2] Brian Chmiel, Maxim Fishman, Ron Banner, and Daniel Soudry. Fp4 all the way: Fully quantized training of llms, 2025. URL https://arxiv.org/abs/2505.19115.

1. experiment code: https://github.com/hellas-ai/neurips2025-convergence-for-discrete-parameter-updates

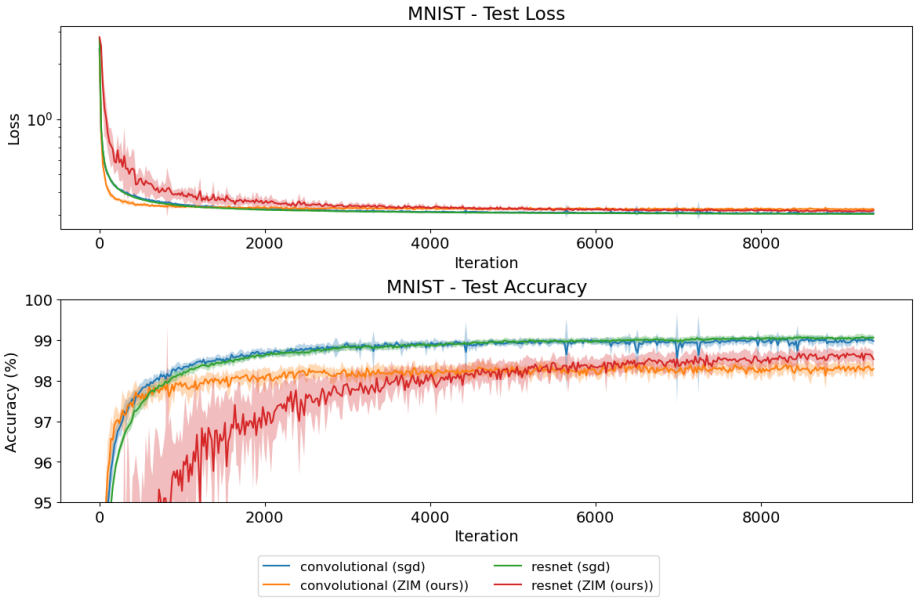

Figure 1: MNIST results: both convolutional and ResNet models converge over 10 epochs under our discrete update (ZIM), compared to SGD. Each curve is averaged over 10 runs; shaded regions show ± 1 std.

[3] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan,

Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

[4] Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George J. Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks, 2023. URL https://arxiv.org/abs/1906.04893.

[5] Maxim Fishman, Brian Chmiel, Ron Banner, and Daniel Soudry. Scaling fp8 training to trillion-token llms, 2025. URL https://arxiv.org/abs/2409.12517.

[6] Yunhui Guo. A survey on methods and theories of quantized neural networks, 2018. URL https://arxiv.org/abs/1808.04752.

[7] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized neural networks: Training neural networks with low precision weights and activations, 2016. URL https://arxiv.org/abs/1609.07061.

[8] Hao Li, Soham De, Zheng Xu, Christoph Studer, Hanan Samet, and Tom Goldstein. Training quantized nets: A deeper understanding, 2017. URL https://arxiv.org/abs/1706.02379.

[9] Zheng Li and Christopher De Sa. *Dimension-free bounds for low-precision training*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[10] Van Minh Nguyen, Cristian Ocampo, Aymen Askri, Louis Leconte, and Ba-Hien Tran. Bold: Boolean logic deep learning, 2024. URL https://arxiv.org/abs/2405.16339.

[11] Houwen Peng, Kan Wu, Yixuan Wei, Guoshuai Zhao, Yuxiang Yang, Ze Liu, Yifan Xiong, Ziyue Yang, Bolin Ni, Jingcheng Hu, Ruihang Li, Miaosen Zhang, Chen Li, Jia Ning, Ruizhe Wang, Zheng Zhang, Shuguang Liu, Joe Chau, Han Hu, and Peng Cheng. Fp8-lm: Training fp8 large language models, 2023. URL https://arxiv.org/abs/2310.18313.

[12] Haotong Qin, Ruihao Gong, Xianglong Liu, Xiao Bai, Jingkuan Song, and Nicu Sebe. Binary neural networks: A survey. *Pattern Recognition*, 105:107281, September 2020. ISSN 0031-3203. doi: 10.1016/j.patcog.2020.107281. URL http://dx.doi.org/10.1016/j.patcog.2020.107281.

[13] Paul W. Wilson and Fabio Zanasi. Reverse derivative ascent: A categorical approach to learning boolean circuits. In David I. Spivak and Jamie Vicary, editors, *Proceedings of the 3rd Annual International Applied Category Theory Conference 2020, ACT 2020, Cambridge, USA, 6-10th July 2020*, volume 333 of *EPTCS*, pages 247–260, 2020. doi: 10.4204/EPTCS.333.17. URL https://doi.org/10.4204/EPTCS.333.17.

[14] Paul W. Wilson and Fabio Zanasi. An axiomatic approach to differentiation of polynomial circuits. *J. Log. Algebraic Methods Program.*, 135:100892, 2023. doi: 10.1016/J.JLAMP.2023.100892. URL https://doi.org/10.1016/j.jlamp.2023.100892.

Appendix A. Lemmas for Section 2

We will now give a proposition analogous to [1, Lemma 4.2] under Assumption 3.

Proposition 8 *Let $\bar{g}$ be a discrete update satisfying Assumption 3. Then we have*

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\nabla F(w_k)^T \mathbb{E}_{\xi_k}[\bar{g}(w_k, \xi_k)] + \frac{1}{2} L \mathbb{E}_{\xi_k}\left[\|\bar{g}(w_k, \xi_k)\|_2^2\right]$$

Proof Following [1, Lemma 4.2], we have

$$\begin{aligned} F(w_{k+1}) - F(w_k) &\leq \nabla F(w_k)^T (w_{k+1} - w_k) + \frac{1}{2} L \|w_{k+1} - w_k\|_2^2 && \text{By (2.1)} \\ &\leq -\nabla F(w_k)^T \bar{g}(w_k, \xi_k) + \frac{1}{2} L \|\bar{g}(w_k, \xi_k)\|_2^2 && \text{Def. 1} \end{aligned}$$

Taking expectations of both sides and noting that $F(w_k)$ does not depend on ξ_k, we obtain the result:

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\nabla F(w_k)^T \mathbb{E}_{\xi_k}[\bar{g}(w_k, \xi_k)] + \frac{1}{2} L \mathbb{E}_{\xi_k}\left[\|\bar{g}(w_k, \xi_k)\|_2^2\right] \quad \text{(Linearity of } \mathbb{E}\text{)}$$

■

Similarly, we can give a proposition analogous to [1, Lemma 4.4]. The sole difference is the final term: theirs contains an α_k^2, whereas ours only has an α_k.

Proposition 9

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -(\mu - \frac{1}{2}\alpha_k L M_G)\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k L M$$

Proof

$$\begin{aligned} \mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) &\leq -\nabla F(w_k)^T \mathbb{E}_{\xi_k}[\bar{g}(w_k, \xi_k)] + \frac{1}{2} L \mathbb{E}_{\xi_k}\left[\|g(w_k, \xi_k)\|_2^2\right] && \text{(Proposition 8)} \\ &\leq -\alpha_k \mu \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k L M + \frac{1}{2}\alpha_k^2 L M_G \|\nabla F(w_k)\|_2^2 \\ & && \text{(Assumption 3)} \\ &\leq -(\mu - \frac{1}{2}\alpha_k L M_G)\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k L M \end{aligned}$$

■

Appendix B. Lemmas for Section 3

Proposition 10 (First and Second Moments of ZIMultinomial) *If* $y \sim \mathsf{ZIMultinomial}(n, r, q)$ *then...*

$$\mathbb{E}[y] = nrq_i$$
$$\mathbb{E}\left[\|y\|_2^2\right] = nr + r^2(n^2 - n)\sum_{i=1}^{d} q_i^2 \qquad (\textstyle\sum_i q_i = 1)$$

Proof The first moment calculation is straightforward: $\mathbb{E}[y_i] = \mathbb{E}[x_i] = nrq_i$. The second is as follows:

$$\begin{aligned}
\mathbb{E}\left[\|y\|_2^2\right] &= \mathbb{E}\left[\sum_{i=1}^{d} y_i^2\right] && \text{(Definition of } \|\cdot\|_2^2\text{)} \\
&= \sum_{i=1}^{d} \mathbb{E}\left[y_i^2\right] && \text{(Linearity of } \mathbb{E}[\cdot]\text{)} \\
&= \sum_{i=1}^{d} \mathbb{E}\left[x_i^2\right] && (y_i = x_i\text{, underlying multinomial counts}) \\
&= \sum_{i=1}^{d} \left(np_i + n(n-1)p_i^2\right) && (x_i \sim \text{Binomial}(n, p_i)) \\
&= n\sum_{i=1}^{d} p_i \;+\; n(n-1)\sum_{i=1}^{d} p_i^2 && \text{(Algebra)} \\
&= nr\sum_{i=1}^{d} q_i \;+\; n(n-1)r^2\sum_{i=1}^{d} q_i^2 && (p_i = rq_i) \\
&= nr + r^2(n^2 - n)\sum_{i=1}^{d} q_i^2 && (\textstyle\sum_i q_i = 1)
\end{aligned}$$

■

Proposition 11 (First and Second Moments of ZIM update) *The first and second moments of the ZIM update* $\bar{g}(w_k, \xi_k)$ *are*

$$\mathbb{E}[\bar{g}(w_k, \xi_k)] = \frac{nr}{\|\nabla F(w_k)\|_1 + cd}(\nabla F(w_k) + c\ \text{sign}(\nabla F(w_k)))$$
$$\mathbb{E}\left[\|\bar{g}(w_k, \xi_k)\|_2^2\right] = nr + r^2(n^2 - n)\sum_{i=1}^{d} q_i^2$$

Proof

$$\begin{aligned}
\mathbb{E}_{\xi_k}[\bar{g}(w_k,\xi_k)] &= \mathbb{E}_{\xi_k}[\,x \odot \operatorname{sign}(\nabla F(w_k))] && \text{(Definition of } \bar{g}) \\
&= \operatorname{sign}(\nabla F(w_k)) \odot \mathbb{E}_{\xi_k}[x] && (\operatorname{sign}(\nabla F(w_k)) \text{ is deterministic}) \\
&= \operatorname{sign}(\nabla F(w_k)) \odot (nr\,q) && (\mathbb{E}[x] = nrq \text{ (Prop. 4.2)}) \\
&= nr\ \operatorname{sign}(\nabla F(w_k)) \odot \frac{|\nabla F(w_k)| + c}{\|\nabla F(w_k)\|_1 + cd} && \text{(Definition of } q) \\
&= \frac{nr}{\|\nabla F(w_k)\|_1 + cd}\big(\nabla F(w_k) + c\,\operatorname{sign}(\nabla F(w_k))\big). && (|u| \odot \operatorname{sign}(u) = u)
\end{aligned}$$

And

$$\begin{aligned}
\mathbb{E}_{\xi_k}\big[\|\bar{g}(w_k,\xi_k)\|_2^2\big] &= \mathbb{E}_{\xi_k}\left[\sum_{i=1}^{d} \bar{g}_i^2\right] && \text{(Definition of } \|\cdot\|_2^2) \\
&= \sum_{i=1}^{d} \mathbb{E}_{\xi_k}\Big[\big(x_i\ \operatorname{sign}(\nabla F(w_k)_i)\big)^2\Big] && (\bar{g}_i = x_i\ \operatorname{sign}(\nabla F(w_k))_i) \\
&= \sum_{i=1}^{d} \mathbb{E}_{\xi_k}\big[x_i^2\big] && (\operatorname{sign}(u)^2 = 1) \\
&= \sum_{i=1}^{d} \big(np_i + n(n-1)p_i^2\big) && (x_i \sim \text{Binomial}(n,p_i)) \\
&= n\sum_{i=1}^{d} p_i + n(n-1)\sum_{i=1}^{d} p_i^2 && \text{(Algebra)} \\
&= nr + r^2(n^2-n)\sum_{i=1}^{d} q_i^2 && (p_i = r\,q_i,\ \textstyle\sum_i q_i = 1)
\end{aligned}$$

■

Proposition 12 *There exists a μ such that*

$$\nabla F(w_k)^T \mathbb{E}_{\xi_k}[\bar{g}(w_k,\xi_k)] \geq \alpha_k \mu \|\nabla F(w_k)\|_2^2$$

with $\alpha_k = r$ and $\mu = \frac{n}{\sqrt{d}L + cd}$.

Proof By Lipschitz-continuity and Cauchy-Schwarz, we have $\|\nabla F(w_k)\|_1 \leq \sqrt{d}L$. Deriving,

$$\begin{aligned}
\nabla F(w_k)^T \mathbb{E}_{\xi_k}[\bar{g}(w_k,\xi_k)] &= \nabla F(w_k)^T \frac{nr}{\|\nabla F(w_k)\|_1 + cd}\big(\nabla F(w_k) + c\,\operatorname{sign}(\nabla F(w_k))\big) \\
&= \frac{nr}{\|\nabla F(w_k)\|_1 + cd}\big(\|\nabla F(w_k)\|_2^2 + c\,\|\nabla F(w_k)\|_1\big) \\
&\geq \frac{nr}{\|\nabla F(w_k)\|_1 + cd}\,\|\nabla F(w_k)\|_2^2 && (c\,\|\nabla F(w_k)\|_1 \geq 0) \\
&\geq r\frac{n}{\sqrt{d}L + cd}\,\|\nabla F(w_k)\|_2^2 && \text{(Lipschitz Continuity, Cauchy-Schwarz)} \\
&= \alpha_k \mu \|\nabla F(w_k)\|_2^2
\end{aligned}$$

■