OPT2025: 17th Annual Workshop on Optimization for Machine Learning

On Optimizing Large Scale Multi-Class Logistic Regression

Yifan Kang YIFANK@CLEMSON.EDU
Yarui Cao YARUIC@CLEMSON.EDU
Kai Liu KAIL@CLEMSON.EDU
Clemson University, USA

Abstract

In this paper, we study multinomial logistic regression (MLR), a fundamental machine learning algorithm used for multi-class classification problems. We first analyze some favorable properties of the MLR objective function. By leveraging these properties, we design an optimization algorithm that operates in a feature-wise manner, which offers potential advantages in terms of computational efficiency and scalability. We also provide a convergence analysis for the proposed algorithms (both stochastic and cyclic versions). We establish theoretical guarantees that ensure the algorithm converges, thereby validating its effectiveness in optimizing the MLR model. To assess the practical performance of our algorithm, we compare our approach with a range of commonly used MLR algorithms. The experimental results demonstrate the efficiency of our algorithm.

1. Introduction

Multinomial logistic regression (MLR) is a classical yet powerful model for multi-class classification, where the probability of a categorical label is expressed via a softmax transformation of linear scores. Owing to its statistical interpretability, convexity, and wide applicability, MLR has been a cornerstone in supervised learning tasks [5, 11]. Despite its theoretical elegance, MLR poses significant algorithmic challenges in large-scale settings. In particular, when the number of classes C, features d, or training samples n is large, standard optimization approaches—such as batch gradient descent or Newton methods—suffer from poor scalability due to their high per-iteration cost and memory footprint [8, 16]. Moreover, incorporating structured regularization, such as sparsity, group penalties, or low-rank constraints, often leads to non-smooth and block-separable objective functions, further complicating the optimization landscape. To address these issues, recent works have explored stochastic and proximal optimization techniques for generalized linear models [7, 13, 14]. In this work, we propose a novel (stochastic) block-wise proximal gradient (BPG) algorithm tailored for regularized multinomial logistic regression. Our method updates one feature block either in a cyclic or stochastic manner, followed by a block-specific proximal operator which achieves low per-iteration complexity and accommodates a wide range of regularizers.

The contributions of this work are summarized as follows:

- We propose an efficient updating algorithm for MLR where we optimize in a blockwise manner which is guaranteed to converge for both convex and nonconvex regularization terms.
- Inspired by modern stochastic variance reduction techniques, we propose a stochastic version of blockwise update, which is faster than updating as a whole.

- We conduct experiments on real-world datasets, both medium- and large-scale, and show that our method is very competitive compared with SVRG, LBFGS and Newton-like methods.

2. Multiomial Logistic Regression

In real-world applications, multi-class classification is everywhere. Traditional methods to generalize from binary classification include One-vs-All (OvA) (also known as One-vs-Rest) and One-vs-One (OvO). However, those two options have their own shortcomings such as too many binary classifiers to be trained, classifiers may be inconsistent with each other. In comparison, multinomial logistic regression (*a.k.a.* 'softmax regression') is a single model which directly optimizes the multi-class likelihood, similar to the binary case, which is typically more consistent and theoretically justified than OvA or OvO, as it considers all classes jointly during training [10]. Assume we have K different classes, our goal is to learn $\mathbf{W} = \{\mathbf{w}_1; \mathbf{w}_2; \dots; \mathbf{w}_K\}$ with each $\mathbf{w}_k \in \mathbb{R}^d$, where a sample $\mathbf{x}$'s score to each class is $\exp(\mathbf{w}_k^T\mathbf{x})$. Therefore the probability to each class is defined as: $P(Y = k|\mathbf{x}) = \frac{\exp(\mathbf{w}_k^T\mathbf{x})}{\sum_{l=1}^K \exp(\mathbf{w}_l^T\mathbf{x})}$, where we will come to *shift-invariance*. Therefore, without loss of generality, we can set $\mathbf{w}_K = \mathbf{0}$. For the inference of test data $\mathbf{x}_{\text{test}}$, it will be assigned to class $k = \arg\max_{i\in[K]} \mathbf{w}_i^T\mathbf{x}_{\text{test}}$. Therefore, it boils down to finding optimal $\mathbf{w}_i$'s to fit the minimum negative log-likelihood function:

$$L(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{K-1}) = -\sum_{i=1}^n \log \frac{\exp(\mathbf{w}_{y_i}^T\mathbf{x}_i)}{1 + \sum_{l=1}^{K-1} \exp(\mathbf{w}_l^T\mathbf{x}_i)} = \sum_{i=1}^n \left[\log(1 + \sum_{l=1}^{K-1} \exp(\mathbf{w}_l^T\mathbf{x}_i)) - \mathbf{w}_{y_i}^T\mathbf{x}_i\right]. \tag{1}$$

Apparently:

$$\frac{\partial L}{\partial \mathbf{w}_k} = -\sum_{i=1}^n \left[\mathbf{1}_{y_i=k} - \frac{\exp(\mathbf{w}_k^T\mathbf{x}_i)}{1 + \sum_{l=1}^{K-1} \exp(\mathbf{w}_l^T\mathbf{x}_i)}\right]\mathbf{x}_i; \frac{\partial^2 L}{\partial \mathbf{w}_k^2} = \sum_{i=1}^n \frac{\exp(\mathbf{w}_k^T\mathbf{x}_i)}{1 + \sum_{l=1}^{K-1} \exp(\mathbf{w}_l^T\mathbf{x}_i)}\left[1 - \frac{\exp(\mathbf{w}_k^T\mathbf{x}_i)}{1 + \sum_{l=1}^{K-1} \exp(\mathbf{w}_l^T\mathbf{x}_i)}\right]\mathbf{x}_i\mathbf{x}_i^T. \tag{2}$$

If we denote $p_k := \frac{\exp(\mathbf{w}_k^T\mathbf{x}_i)}{1+\sum_{l=1}^{K-1} \exp(\mathbf{w}_l^T\mathbf{x}_i)}$, then $1 - \frac{\exp(\mathbf{w}_k^T\mathbf{x}_i)}{1+\sum_{l=1}^{K-1} \exp(\mathbf{w}_l^T\mathbf{x}_i)} = 1 - p_k$ and $\frac{\partial^2 L}{\partial \mathbf{w}_k^2} = \sum_{i=1} p_k(1 - p_k)\mathbf{x}_i\mathbf{x}_i^T$, thus we can bound Lipschitz gradient continuous constant by

$$\sigma_{\max}(\frac{\partial^2 L}{\partial \mathbf{w}_k^2}) = \max_{\|\mathbf{v}\|=1} \mathbf{v}^T \frac{\partial^2 L}{\partial \mathbf{w}_k^2}\mathbf{v} = \sum_{i=1} (\mathbf{v}^*)^T p_k(1 - p_k)\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}^* \le \frac{1}{4}\sum_{i=1} (\mathbf{v}^*)^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}^* = \frac{1}{4}(\mathbf{v}^*)^T\mathbf{X}\mathbf{X}^T\mathbf{v}^* \le \frac{1}{4}\|\mathbf{X}\|_2^2, \tag{3}$$

where we make use of the fact that for convex function (Log-Sum-Exp is convex [4]), Lipschitz gradient continuous constant is the largest singular value of its Hessian Matrix [2]; in the second equation, we assume $\mathbf{v}^*$ admits the supremum; the first inequality makes use of the simple fact that $p(1 - p) \le \frac{1}{4}$ for any positive scalar p; the last equation is from the fact that $\sum_{i=1} \mathbf{x}_i\mathbf{x}_i^T = \mathbf{X}\mathbf{X}^T$ while the last inequality we again make use of the definition of matrix spectral norm. We can also update $\mathbf{W}$ as a whole using gradient descent: $\mathbf{W}^+ = \mathbf{W} - t\nabla L(\mathbf{W})$. To ensure the objective is monotonically decreasing, stepsize t should not exceed $2/\|\mathbf{X}\|_2^2$. However, this approach may suffer from the following: 1. To calculate the spectral norm of $\mathbf{X}$ is computationally demanding. It is true that the stepsize $\frac{2}{\|\mathbf{X}\|_F^2}$ will also decrease the objective, however it is likely to be too small and conservative; 2. For large-scale $\mathbf{X}$, the spectral norm is large in expectation as well ($\propto (\sqrt{n}+\sqrt{d})$ [1]), admitting slow convergence. Inspired by the above, we propose an algorithm in feature-based blockwise update instead of class-based as Fig. 1 demonstrates. Our first observation

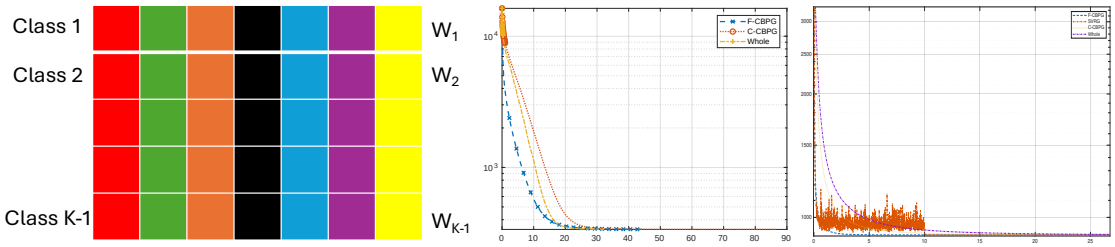

Figure 1: Left: Assume $\mathbf{W} \in \mathbb{R}^{(K-1)\times m}$ ($\mathbf{w}^i$ denotes the i-th column of $\mathbf{W}$ while $\mathbf{w}_i$ i-th row), where each row denotes one class. Our method is not to optimize row by row (class-wise), but column by column, namely feature-wise (various colors represent different blocks). Middle: We compare the feature-wise cyclic update (F-CBPG) vs class-wise (C-CBPG where $L_i = \|\mathbf{X}\|_2^2/4$) and matrix-wise ($L = \|\mathbf{X}\|_2^2/2$, see more details in Appendix E) on MNIST dataset (partial) where X-axis denotes time consumption in seconds and Y-axis the objective. Right: we compare with SVRG on SEGMENT [8] dataset and find it is slower than ours and shows fluctuations.

is that instead of updating as a whole, coordinate-wise or block-wise version can be faster, admitting monotone decrease with update. More formally, we have the following theorem (Chapter 11 in [3]):

Theorem 1 *If we denote* $\mathbf{W} = (\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^p)$, $\mathcal{U}_i(\mathbf{d}) = (\mathbf{0}, \ldots, \mathbf{0}, \underbrace{\mathbf{d}}_{i\text{-th block}}, \mathbf{0}, \ldots, \mathbf{0})$, $\nabla f(\mathbf{W}) = (\nabla_1 f(\mathbf{W}), \nabla_2 f(\mathbf{W}), \ldots, \nabla_p f(\mathbf{W}))$. *Suppose there exists* $L_i > 0$ *for which*

$$\|\nabla_i f(\mathbf{Y}) - \nabla_i f(\mathbf{Y} + \mathcal{U}_i(\mathbf{d}))\|_F \leq L_i \|\mathbf{d}\|_F,$$

then for

$$\min_{\mathbf{w}^1,\ldots,\mathbf{w}^p} \left\{ F(\mathbf{w}^1, \ldots, \mathbf{w}^p) = f(\mathbf{w}^1, \ldots, \mathbf{w}^p) + \sum_{j=1}^{p} g_j(\mathbf{w}^j) \right\} \tag{4}$$

updating $\mathbf{W}^+ = \mathbf{W} + \mathcal{U}_i \left(\text{prox}_{\frac{1}{L_i} g_i}(\mathbf{w}^i - \frac{1}{L_i} \nabla_i f(\mathbf{W})) - \mathbf{w}^i \right)$ *will admit sufficient decrease:*

$$F(\mathbf{W}) - F(\mathbf{W}^+) \geq \frac{1}{2L_i} \left\| L_i(\text{prox}_{\frac{1}{L_i} g_i}(\mathbf{w}^i - \frac{1}{L_i} \nabla_i f(\mathbf{W})) - \mathbf{w}^i) \right\|_2^2 = \frac{L_i}{2} \|\mathbf{W}^+ - \mathbf{W}\|_F^2. \tag{5}$$

One can also see if we add regularization term for MLR, such as $\|\mathbf{W}\|_1$, then $\forall i \in [d]$, $g_i = g = \|\cdot\|_1$, while for $\|\mathbf{W}\|_F^2$, we have $\forall i \in [d]$, $g_i = g = \|\cdot\|_2^2$. In addition, if we add nonnegative constraint, then g_i is a nonnegativity indicator function. In short, the above theorem is a very general framework which admits many regularized models.

We now turn to find the blockwise or coordinatewise Lipschitz gradient continuous constant for $\mathbf{W}$. One can find the Hessian (class-stacked parameters) [10] $\mathbf{H} \in \mathbb{R}^{d(K-1)\times d(K-1)}$ is:

$$\mathbf{H} = \sum_{i=1}^{n} \left(\text{diag}(\mathbf{p}_i) - \mathbf{p}_i \mathbf{p}_i^\top \right) \otimes \left(\mathbf{x}_i \mathbf{x}_i^\top \right).$$

Accordingly, Feature-by-feature (r,s) block $\mathbf{H}_{rs}$ is given by:

$$\mathbf{H}_{rs} = \sum_{i=1}^{n} \mathbf{x}_{ir} \, \mathbf{x}_{is} \left(\text{diag}(\mathbf{p}_i) - \mathbf{p}_i \mathbf{p}_i^\top \right) \in \mathbb{R}^{(K-1)\times(K-1)}.$$

Specifically, to update s-th block feature, we need bound the spectral norm of $\mathbf{H}_{ss} = \sum_{i=1}^{n} \mathbf{x}_{is}^2 \left(\text{diag}(\mathbf{p}_i) - \mathbf{p}_i\mathbf{p}_i^\top\right)$. Denote $\mathbf{J}_i = \text{diag}(\mathbf{p}_i) - \mathbf{p}_i\mathbf{p}_i^\top$, we have

$$\|\mathbf{H}_{ss}\|_2 \leq \sum_{i=1} \mathbf{x}_{is}^2 \|\mathbf{J}_i\|_2 = \sum_{i=1} \mathbf{x}_{is}^2 \lambda_{max}(\mathbf{J}_i) \tag{6}$$

given $\mathbf{H}_{ss}$ is Symmetric Positive Semi-Definite. For any probability vector $\mathbf{p}$, $\mathbf{J} = \text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$:

$$\lambda_{max}(\mathbf{J}) = \max_{\|\mathbf{v}\|=1} \text{Var}_p(\mathbf{v}) \leq \frac{(\max(\mathbf{v}) - \min(\mathbf{v}))^2}{4} \leq \frac{1}{2}, \tag{7}$$

where the last step uses $\|\mathbf{v}\| = 1$ and Popoviciu's inequality on variance [12] (equation attained when $\mathbf{v} = [\sqrt{2}/2, -\sqrt{2}/2, 0, \ldots, 0]$). Thus $\|\mathbf{H}_{ss}\|_2 \leq \frac{1}{2}\sum_{i=1} \mathbf{x}_{is}^2 = \frac{1}{2}\|\mathbf{x}^s\|_2^2$. By comparing with Theorem 1, we can conclude if the input data is a vector of size d, then $p = d$, $L_i = \frac{\|\mathbf{x}^i\|_2^2}{2}$ and each block $\mathbf{w}^i$ is a vector size of $K-1$ corresponds to each column of the matrix in Fig. 1. Apparently, different from calculating spectral norm of matrix $\mathbf{X}$ in Eq. (3), it is way more efficient to calculate $\|\mathbf{x}^s\|_2^2$ since $\mathbf{x}^s$ is a vector where $\|\mathbf{x}^s\|_2^2$ is equal to summation of each element's square. See more details in Appendix D and discussions on class-wise L_k in Appendix F.

3. Optimization and Convergence

In the previous section, we demonstrated that the objective function of multinomial logistic regression can be updated blockwise instead of matrix-wise, where sufficient decrease is guaranteed for each update, thereby ensuring that the objective is monotonically decreasing. However, it remains unclear which block should be selected at each step and whether the order of updates impacts convergence. In this section, we study the convergence properties under different update orderings.

3.1. Cyclic Block Proximal Gradient Method

In the cyclic block proximal gradient (CBPG) method we successively pick a block in a cyclic manner and perform a prox-grad step w.r.t. the chosen block. The t-th iteration is denoted as $\mathbf{W}_t = (\mathbf{w}_t^1, \mathbf{w}_t^2, \ldots, \mathbf{w}_t^d)$. Each iteration of the CBPG method involves d subiterations, and the by-products of these subiterations will be denoted by the following auxiliary subsequences:

$$\mathbf{W}_{t,0} = \mathbf{W}_t = (\mathbf{w}_t^1, \mathbf{w}_t^2, \ldots, \mathbf{w}_t^d), \mathbf{W}_{t,1} = (\mathbf{w}_{t+1}^1, \mathbf{w}_t^2, \ldots, \mathbf{w}_t^d), \cdots, \mathbf{W}_{t,d} = (\mathbf{w}_{t+1}^1, \mathbf{w}_{t+1}^2, \ldots, \mathbf{w}_{t+1}^d). \tag{8}$$

By following the notations above, one can easily see $\mathbf{W}_{t,i-1}^i = \mathbf{W}_t^i = \mathbf{w}_t^i, \mathbf{W}_{t,i}^i = \mathbf{W}_{t+1}^i = \mathbf{w}_{t+1}^i$. Our first observation is a direct consequence of the sufficient decrease property from Eq. (5): $F(\mathbf{W}_t) - F(\mathbf{W}_{t+1}) \geq \frac{L_{\min}}{2}\|\mathbf{W}_t - \mathbf{W}_{t+1}\|_F^2$, where $L_{\min} = \min_{i\in[d]} L_i$. This inequality says that each block update will have a sufficient decrease, then after one cycle it will be bounded by the right side of the above equation.

Assumption For any $\alpha > 0$, there exists $R_\alpha > 0$ such that $\max_{\mathbf{x},\mathbf{x}^*}\{\|\mathbf{x} - \mathbf{x}^*\| : F(\mathbf{x}) \leq \alpha\} \leq R_\alpha$.

Lemma 2 *Let $\{\mathbf{W}_t\}_{t\geq 0}$ be the sequence generated by the CBPG method described in Algorithm 1 for solving problem Eq. (4). Then for any $t \geq 0$,*

$$F(\mathbf{W}_t) - F(\mathbf{W}_{t+1}) \geq \frac{L_{\min}}{2d(L_f + L_{\max})^2 R^2}(F(\mathbf{W}_{t+1}) - F_{\text{opt}}), \tag{9}$$

where $R = R_{F(\mathbf{W}_0)}$, $L_{\max} = \max_{j=1,2,\ldots,d} L_j$ *and* $L_{\min} = \min_{j=1,2,\ldots,d} L_j$.

Theorem 3 ($\mathcal{O}(1/t)$ rate of convergence of CBPG) *Suppose the assumption holds. Let* $\{\mathbf{W}_t\}_{t\geq 0}$ *be the sequence generated by the CBPG method for solving problem Eq. (4). For any* $t \geq 2$,

$$F(\mathbf{W}^t) - F_{\text{opt}} \leq \max\left\{\left(\frac{1}{2}\right)^{(t-1)/2}(F(\mathbf{W}_0) - F_{\text{opt}}), \frac{8d(L_{\max} + L_{\text{f}})^2 R^2}{L_{\min}(t-1)}\right\}. \tag{10}$$

In addition, $F(\mathbf{W}^t) - F_{\text{opt}} \leq \epsilon$ *as long as* $t \geq 2$ *satisfies*

$$t \geq 1 + \max\left\{\frac{2}{\log 2}(\log(F(\mathbf{W}_0) - F_{\text{opt}}) + \log\frac{1}{\epsilon}), \frac{8d(L_{\max} + L_{\text{f}})^2 R^2}{L_{\min}\epsilon}\right\}. \tag{11}$$

Algorithm 1: Cyclic Block Proximal Gradient (CBPG) Method

Data: $\mathbf{X} \in \mathbb{R}^{n\times d}$; initial $\mathbf{W}_0 = [\mathbf{w}_0^1, \mathbf{w}_0^2, \ldots, \mathbf{w}_0^d]$; compute $L_i = \frac{\|\mathbf{x}^i\|_2^2}{2}$, set T and $t = 0$
for $t \leq T$ **do**
  $\mathbf{W}_{t,0} \leftarrow \mathbf{W}_t$;
  for $i = 1, \ldots, d$ **do**
    $\mathbf{W}_{t,i} \leftarrow \mathbf{W}_{t,i-1} + \mathcal{U}_i\left(\text{prox}_{\frac{1}{L_i}g_i}\left(\mathbf{W}^i_{t,i-1} - \frac{1}{L_i}\nabla_i f(\mathbf{W}_{t,i-1})\right) - \mathbf{W}^i_{t,i-1}\right)$;
  end
  $\mathbf{W}_{t+1} \leftarrow \mathbf{W}_{t,d}$;
end

3.2. The Randomized Block Proximal Gradient Method

We now analyze a version of the block proximal gradient method, in which at each iteration, a prox-grad step is performed at a randomly chosen block. Similar to CBPG, the randomized method is also monotonically decreasing while the difference is that in each loop, only a random block will be updated instead of every block. The main convergence result will now be stated.

Theorem 4 ($\mathcal{O}(1/t)$ rate of convergence of RBPG) *Let* $\{\mathbf{W}_t\}_{t\geq 0}$ *be the sequence generated by the RBPG method for solving problem Eq. (4). For any* $t \geq 0$ *and denote* $\|\mathbf{W}\|_L^2 := \sum_{i=1}^{d} L_i \|\mathbf{w}^i\|_2^2$,

$$\mathbb{E}_{\xi_t}\left[F(\mathbf{W}_{t+1})\right] - F_{opt} \leq \frac{d}{d+t+1}\left(\frac{1}{2}\|\mathbf{W}_0 - \mathbf{W}_*\|_L^2 + F(\mathbf{W}_0) - F_{opt}\right). \tag{12}$$

4. Experiment

We evaluate our proposed algorithm for multinomial logistic regression across multiple benchmark datasets. Comparisons are made against the following baselines: **SVRG** [7], **L-BFGS** [9], **Trust Region** [17], and **Newton Conjugate Gradient (Newton-CG)** [8, 15]. As discussed earlier, although the stochastic/randomized method is non-deterministic, it still guarantees that the objective

Algorithm 2: Randomized Block Proximal Gradient (RBPG) Method

Data: $\mathbf{X} \in \mathbb{R}^{n \times d}$; initial $\mathbf{W}_0 = [\mathbf{w}_0^1, \mathbf{w}_0^2, \ldots, \mathbf{w}_0^d]$; compute $L_i = \frac{\|\mathbf{x}^i\|_2^2}{2}$, set T and $t = 0$

for $t \leq T$ **do**

Sample $i_t \in \{1, 2, \ldots, d\}$ uniformly at random or with probability $L_{i_t} / \sum_j L_j$;

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \mathcal{U}_{i_t} \left(\text{prox}_{\frac{1}{L_{i_t}} g_{i_t}} (\mathbf{w}_t^{i_t} - \frac{1}{L_{i_t}} \nabla_{i_t} f(\mathbf{W}_t)) - \mathbf{w}_t^{i_t} \right);$$

end

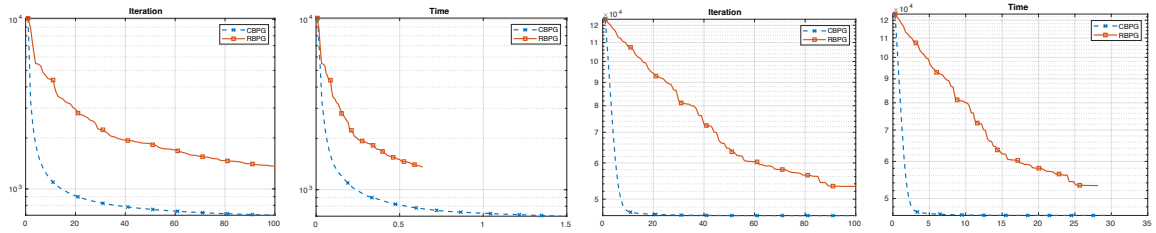

Figure 2: Comparison of our two methods on Segment and Poker datasets [8].

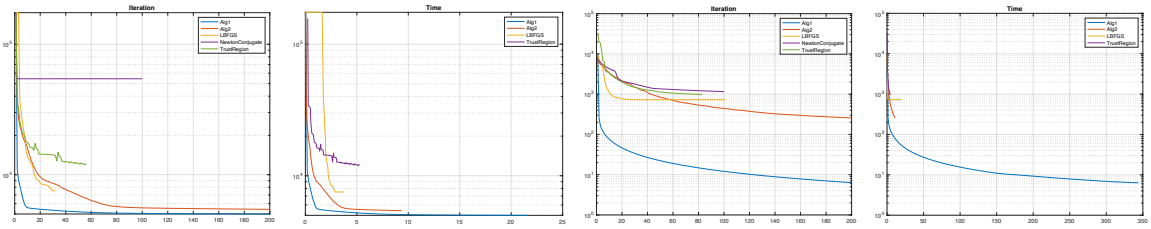

Figure 3: Comparison of our two methods with counterparts on DNA and Poker datasets [8].

function is monotonically decreasing. This differs from SVRG, where only the expected objective is guaranteed to decrease. Besides, SVRG requires strong convexity of the objective, which is not true for vanilla MLR. Thus we add $\|\mathbf{W}\|_F^2$ as a regularizer to satisfy the requirement. Fig. 2 shows the comparison on two real-world datasets: Segment and Poker. In Algorithm 1, each outer loop involves a full cycle of feature-wise updates, which is equivalent to d iterations in Algorithm 2. We observe that to achieve the same objective value, the ratio $\frac{\text{\# iterations of RBPG}}{\text{\# iterations of CBPG}} \approx d$. For example, in the third subfigure, to reach an objective value of 6×10^4, *RBPG* requires 60 iterations, while *CBPG* needs only about 6. A similar pattern can be seen in the first subfigure, with a ratio around 18. In terms of time consumption, although the theoretical complexity is similar, practical performance differs. This discrepancy arises mainly because *RBPG* frequently computes the objective function, which becomes costly when the dataset is large. We also compared our method with traditional optimization algorithms such as *L-BFGS* on medium-scale datasets. Our method consistently outperforms these baselines in both iteration count and wall-clock time. Fig. 3 presents results on the DNA and Poker datasets, which are representative of broader trends. We exclude the time performance of *Newton-CG* due to its significantly higher computational cost—primarily from computing the Hessian matrix—despite not being the worst in terms of iteration count.

References

[1] Zhidong D Bai and Yong Q Yin. Convergence to the semicircle law. *The Annals of Probability*, 16(2):863–875, 1988.

[2] Amir Beck. *Introduction to nonlinear optimization: Theory, algorithms, and applications with MATLAB*. SIAM, 2014.

[3] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[4] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[5] Trevor Hastie. The elements of statistical learning: data mining, inference, and prediction, 2009.

[6] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[7] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

[8] Chih-Jen Lin, Ruby C Weng, and S Sathiya Keerthi. Trust region newton method for large-scale logistic regression. *Journal of Machine Learning Research*, 9(4), 2008.

[9] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

[10] Kevin P Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.

[11] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[12] Tiberiu Popoviciu. Sur les équations algébriques ayant toutes leurs racines réelles. *Mathematica*, 9(129-145):20, 1935.

[13] Peter Richtarik and Martin Takavc. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1): 1–38, 2014.

[14] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. *The Journal of Machine Learning Research*, 14(1):567–599, 2013.

[15] Rui Wang, Naihua Xiu, and Chao Zhang. Greedy projected gradient-newton method for sparse logistic regression. *IEEE transactions on neural networks and learning systems*, 31(2):527–538, 2019.

[16] Kai Yang, Tao Fan, Tianjian Chen, Yuanming Shi, and Qiang Yang. A quasi-newton method based vertical federated learning framework for logistic regression. *arXiv preprint arXiv:1912.00513*, 2019.

[17] Nayyar A Zaidi and Geoffrey I Webb. A fast trust-region newton method for softmax logistic regression. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 705–713. SIAM, 2017.

The convergence proof in this section largely follows [3], and we leave the sampling importance version of RBPG to the future work.

Appendix A. Lemma 2

Lemma 5 *Let $\{\mathbf{W}_t\}_{t\geq 0}$ be the sequence generated by the CBPG method described in Algorithm 1 for solving problem Eq. (4). Then for any $t \geq 0$,*

$$F(\mathbf{W}_t) - F(\mathbf{W}_{t+1}) \geq \frac{L_{\min}}{2d(L_f + L_{\max})^2 R^2}(F(\mathbf{W}_{t+1}) - F_{\text{opt}}), \tag{13}$$

where $R = R_{F(\mathbf{W}_0)}$, $L_{\max} = \max_{j=1,2,\ldots,d} L_j$ and $L_{\min} = \min_{j=1,2,\ldots,d} L_j$.

Proof By the definition of CBPG method, for any $t \geq 0$ and $j \in \{1, 2, \ldots, d\}$,

$$\mathbf{W}_{t,j}^j = \text{prox}_{\frac{1}{L_j} g_j}\left(\mathbf{W}_{t,j-1}^j - \frac{1}{L_j}\nabla_j f(\mathbf{W}_{t,j-1})\right). \tag{14}$$

By making use of second prox theorem [3], for any $\mathbf{y}$:

$$g_j(\mathbf{y}) \geq g_j(\mathbf{W}_{t,j}^j) + L_j\left\langle \mathbf{W}_{t,j-1}^j - \frac{1}{L_j}\nabla_j f(\mathbf{W}_{t,j-1}) - \mathbf{W}_{t,j}^j,\ \mathbf{y} - \mathbf{W}_{t,j}^j\right\rangle. \tag{15}$$

By the definition in Eq. (8), we have

$$g_j(\mathbf{y}) \geq g_j(\mathbf{W}_{t+1}^j) + L_j\left\langle \mathbf{W}_t^j - \frac{1}{L_j}\nabla_j f(\mathbf{W}_{t,j-1}) - \mathbf{W}_{t+1}^j,\ \mathbf{y} - \mathbf{W}_{t+1}^j\right\rangle. \tag{16}$$

Since the above holds for any $\mathbf{y}$, if we set $\mathbf{y} = \mathbf{W}_*^j$, then

$$g_j(\mathbf{W}_*^j) \geq g_j(\mathbf{W}_{t+1}^j) + L_j\left\langle \mathbf{W}_t^j - \frac{1}{L_j}\nabla_j f(\mathbf{W}_{t,j-1}) - \mathbf{W}_{t+1}^j,\ \mathbf{W}_*^j - \mathbf{W}_{t+1}^j\right\rangle. \tag{17}$$

By summing the above over $j = \{1, 2, \ldots, d\}$ yields

$$g(\mathbf{W}_*) \geq g(\mathbf{W}_{t+1}) + \sum_{j=1}^d L_j\left\langle \mathbf{W}_t^j - \frac{1}{L_j}\nabla_j f(\mathbf{W}_{t,j-1}) - \mathbf{W}_{t+1}^j,\ \mathbf{W}_*^j - \mathbf{W}_{t+1}^j\right\rangle. \tag{18}$$

By making use the convexity property of function f:

$$\begin{aligned} &F(\mathbf{W}_{t+1}) - F(\mathbf{W}_*) = f(\mathbf{W}_{t+1}) - f(\mathbf{W}_*) + g(\mathbf{W}_{t+1}) - g(\mathbf{W}_*) \\ &\leq \langle \nabla f(\mathbf{W}_{t+1}), \mathbf{W}_{t+1} - \mathbf{W}_*\rangle + g(\mathbf{W}_{t+1}) - g(\mathbf{W}_*) \\ &= \sum_{j=1}^d \left\langle \nabla_j f(\mathbf{W}_{t+1}), \mathbf{W}_{t+1}^j - \mathbf{W}_*^j\right\rangle + g(\mathbf{W}_{t+1}) - g(\mathbf{W}_*), \end{aligned} \tag{19}$$

which together with Eq. (18) implies:

$$\begin{aligned} &F(\mathbf{W}_{t+1}) - F(\mathbf{W}_*) \leq \sum_{j=1}^d \left\langle \nabla_j f(\mathbf{W}_{t+1}), \mathbf{W}_{t+1}^j - \mathbf{W}_*^j\right\rangle + \sum_{j=1}^d L_j\left\langle \mathbf{W}_t^j - \frac{1}{L_j}\nabla_j f(\mathbf{W}_{t,j-1}) - \mathbf{W}_{t+1}^j,\ \mathbf{W}_{t+1}^j - \mathbf{W}_*^j\right\rangle \\ &= \sum_{j=1}^d \left\langle \nabla_j f(\mathbf{W}_{t+1}) - \nabla_j f(\mathbf{W}_{t,j-1}) + L_j(\mathbf{W}_t^j - \mathbf{W}_{t+1}^j),\ \mathbf{W}_{t+1}^j - \mathbf{W}_*^j\right\rangle. \end{aligned} \tag{20}$$

By using Cauchy–Schwarz and triangle inequalities:

$$
\begin{aligned}
&F(\mathbf{W}_{t+1}) - F(\mathbf{W}_*) \leq \sum_{j=1}^{d} \Big(\|\nabla_j f(\mathbf{W}_{t+1}) - \nabla_j f(\mathbf{W}_{t,j-1})\| + L_j \|\mathbf{W}_t^j - \mathbf{W}_{t+1}^j\| \Big) \cdot \|\mathbf{W}_{t+1}^j - \mathbf{W}_*^j\| \\
\leq &\sum_{j=1}^{d} \Big(\|\nabla f(\mathbf{W}_{t+1}) - \nabla f(\mathbf{W}_{t,j-1})\| + L_j \|\mathbf{W}_t^j - \mathbf{W}_{t+1}^j\| \Big) \cdot \|\mathbf{W}_{t+1}^j - \mathbf{W}_*^j\| \\
\leq &\sum_{j=1}^{d} \left(L_f \|\mathbf{W}_{t+1} - \mathbf{W}_{t,j-1}\| + L_{\max} \|\mathbf{W}_t - \mathbf{W}_{t+1}\| \right) \cdot \|\mathbf{W}_{t+1}^j - \mathbf{W}_*^j\| \\
\leq &(L_f + L_{\max}) \|\mathbf{W}_{t+1} - \mathbf{W}_t\| \sum_{j=1}^{d} \|\mathbf{W}_{t+1}^j - \mathbf{W}_*^j\|.
\end{aligned}
\tag{21}
$$

Therefore,

$$
\begin{aligned}
&(F(\mathbf{W}_{t+1}) - F(\mathbf{W}_*))^2 \leq (L_f + L_{\max})^2 \|\mathbf{W}_{t+1} - \mathbf{W}_t\|^2 \left(\sum_{j=1}^{d} \|\mathbf{W}_{t+1}^j - \mathbf{W}_*^j\| \right)^2 \\
\leq &d(L_f + L_{\max})^2 \|\mathbf{W}_{t+1} - \mathbf{W}_t\|^2 \sum_{j=1}^{d} \|\mathbf{W}_{t+1}^j - \mathbf{W}_*^j\|^2 \\
= &d(L_f + L_{\max})^2 \|\mathbf{W}_{t+1} - \mathbf{W}_t\|^2 \cdot \|\mathbf{W}_{t+1} - \mathbf{W}_*\|^2 \leq d(L_f + L_{\max})^2 R_{F(\mathbf{W}_0)}^2 \|\mathbf{W}_{t+1} - \mathbf{W}_t\|^2,
\end{aligned}
\tag{22}
$$

where the last inequality follows by the monotonicity of the sequence of function values and assumption. Combining all above and define $R = R_{F(\mathbf{W}_0)}$, we obtain that:

$$
F(\mathbf{W}_t) - F(\mathbf{W}_{t+1}) \geq \frac{L_{\min}}{2} \|\mathbf{W}_{t+1} - \mathbf{W}_t\|^2 \geq \frac{L_{\min}}{2d(L_f + L_{\max})^2 R^2} \left(F(\mathbf{W}_{t+1}) - F(\mathbf{W}_*) \right)^2. \tag{23}
$$

■

Appendix B. Theorem 3

We first introduce the following lemma.

Lemma 6 *Let $\{a_n\}_{n \geq 0}$ be a nonnegative sequence of real numbers satisfying*

$$
a_n - a_{n+1} \geq \frac{1}{\gamma} a_{n+1}^2, \quad \textit{for all } n = 0, 1, 2, \ldots
$$

for some $\gamma > 0$. Then for any $n \geq 2$,

$$
a_n \leq \max \left\{ \left(\frac{1}{2} \right)^{(n-1)/2} a_0, \ \frac{4\gamma}{n-1} \right\}. \tag{24}
$$

In addition, for any $\varepsilon > 0$, if $n \geq 2$ satisfies

$$n \geq \max\left\{\frac{2\log 2}{\log a_0 + \log(1/\varepsilon)},\ \frac{4\gamma}{\varepsilon}\right\} + 1,$$

then $a_n \leq \varepsilon$.

Proof Let $n \geq 2$. If $a_n = 0$, then the bound Eq. (24) holds trivially. We can thus assume $a_n > 0$, which implies $a_1, a_2, \ldots, a_{n-1} > 0$. For any $k \in \{0, 1, \ldots, n-1\}$, from the assumption we have:

$$\frac{1}{a_{k+1}} - \frac{1}{a_k} = \frac{a_k - a_{k+1}}{a_k a_{k+1}} \geq \frac{1}{\gamma}\frac{a_{k+1}}{a_k} \tag{25}$$

Now, for each k, consider two options:

(i) $\dfrac{a_{k+1}}{a_k} \leq \dfrac{1}{2}$;

(ii) $\dfrac{a_{k+1}}{a_k} > \dfrac{1}{2}$.

Under option (ii), using Eq. (25), we obtain:

$$\frac{1}{a_{k+1}} - \frac{1}{a_k} \geq \frac{1}{2\gamma}$$

Assume n is an even positive integer. If there are at least $\frac{n}{2}$ indices $k \in \{0, \ldots, n-1\}$ for which (ii) holds, then summing the inequality above:

$$\frac{1}{a_n} - \frac{1}{a_0} \geq \frac{n}{2} \cdot \frac{1}{2\gamma} = \frac{n}{4\gamma} \quad \Rightarrow \quad a_n \leq \frac{4\gamma}{n}$$

Otherwise, option (i) occurs at least $\frac{n}{2}$ times. Each such step satisfies $a_{k+1} \leq \frac{1}{2}a_k$, so we have:

$$a_n \leq \left(\frac{1}{2}\right)^{n/2} a_0$$

Combining both cases, we have:

$$a_n \leq \max\left\{\left(\frac{1}{2}\right)^{n/2} a_0,\ \frac{4\gamma}{n}\right\}, \quad \text{for even } n. \tag{26}$$

If $n \geq 3$ is positive odd integer, then

$$a_n \leq a_{n-1} \leq \max\left\{\left(\frac{1}{2}\right)^{(n-1)/2} a_0,\ \frac{4\gamma}{n-1}\right\}. \tag{27}$$

Since the right-hand side of Eq. (27) is larger than the right-hand side of Eq. (26), the claimed result follows.

To ensure $a_n \leq \varepsilon$, it suffices that:

$$\left(\frac{1}{2}\right)^{(n-1)/2} a_0 \leq \varepsilon, \quad \frac{4\gamma}{n-1} \leq \varepsilon.$$

These are equivalent to:

$$n \geq \frac{2}{\log(2)} \left(\log(a_0) + \log(1/\varepsilon)\right) + 1, \quad n \geq \frac{4\gamma}{\varepsilon} + 1.$$

Hence, it suffices to choose:

$$n \geq \max\left\{\frac{2}{\log(2)} \left(\log(a_0) + \log(1/\varepsilon)\right), \ \frac{4\gamma}{\varepsilon}\right\} + 1.$$

■

We now turn back to the theorem part. Denote $a_n = F(\mathbf{W}_t) - F_{\text{opt}}$, by invoking Lemma 6, where $\gamma = \frac{2d(L_f + L_{\max})^2 R^2}{L_{\min}}$, the desired result follows immediately.

Appendix C. Theorem 4

Proof For the sake of convenience in later analysis, we define

$$\begin{aligned} &G_L^i(\mathbf{W}) = L(\mathbf{W}^i - \text{prox}_{\frac{1}{L}g_i}(\mathbf{W}^i - \frac{1}{L}\nabla_i f(\mathbf{W}))); \\ &\widetilde{G}(\mathbf{W}_t) = \left(G_{L_1}^1(\mathbf{W}_t), G_{L_2}^2(\mathbf{W}_t), \ldots, G_{L_d}^d(\mathbf{W}_t)\right), \|\mathbf{W}\|_{L,*}^2 = \sum_{i=1}^d \frac{1}{L_i}\|\mathbf{w}^i\|_2^2. \end{aligned} \tag{28}$$

For any $n \geq 0, r_n := \|\mathbf{W}_n - \mathbf{W}_*\|_L$, then for any $t \geq 0$:

$$\begin{aligned} r_{t+1}^2 =& \|\mathbf{W}_{t+1} - \mathbf{W}_*\|_L^2 = \left\|\mathbf{W}_t - \frac{1}{L_{i_t}}\mathcal{U}_{i_t}\left[G_{L_{i_t}}^{i_t}(\mathbf{W}_t)\right] - \mathbf{W}_*\right\|_L^2 \\ =& \|\mathbf{W}_t - \mathbf{W}_*\|_L^2 - \frac{2}{L_{i_t}} L_{i_t}\left\langle G_{L_{i_t}}^{i_t}(\mathbf{W}_t), \mathbf{W}_t^{i_t} - \mathbf{W}_*^{i_t}\right\rangle + \frac{1}{L_{i_t}}\|G_{L_{i_t}}^{i_t}(\mathbf{W}_t)\|^2 \\ =& r_t^2 - 2\left\langle G_{L_{i_t}}^{i_t}(\mathbf{W}_t), \mathbf{W}_t^{i_t} - \mathbf{W}_*^{i_t}\right\rangle + \frac{1}{L_{i_t}}\left\|G_{L_{i_t}}^{i_t}(\mathbf{W}_t)\right\|^2. \end{aligned}$$

Taking expectation w.r.t. i_t, we obtain

$$\begin{aligned} \mathbb{E}_{i_t}\left(\frac{1}{2}r_{t+1}^2\right) &= \frac{1}{2}r_t^2 - \frac{1}{d}\sum_{i=1}^d \langle G_{L_i}^i(\mathbf{W}_t), \mathbf{W}_t^i - \mathbf{W}_*^i\rangle + \frac{1}{2d}\sum_{i=1}^d \frac{1}{L_i}\|G_{L_i}^i(\mathbf{W}_t)\|^2 \\ &= \frac{1}{2}r_t^2 - \frac{1}{d}\left\langle \widetilde{G}(\mathbf{W}_t), \mathbf{W}_t - \mathbf{W}_*\right\rangle + \frac{1}{2d}\|\widetilde{G}(\mathbf{W}_t)\|_{L,*}^2. \end{aligned} \tag{29}$$

By the block descent lemma:

$$f(\mathbf{W}_{t+1}) = f\left(\mathbf{W}_t - \frac{1}{L_{i_t}}\mathcal{U}_{i_t}\left[G_{L_{i_t}}^{i_t}(\mathbf{W}_t)\right]\right) \leq f(\mathbf{W}_t) - \frac{1}{L_{i_t}}\left\langle \nabla_{i_t} f(\mathbf{W}_t), G_{L_{i_t}}^{i_t}(\mathbf{W}_t)\right\rangle + \frac{1}{2L_{i_t}}\|G_{L_{i_t}}^{i_t}(\mathbf{W}_t)\|^2. \tag{30}$$

Hence,

$$F(\mathbf{W}_{t+1}) \leq f(\mathbf{W}_t) - \frac{1}{L_{i_t}} \left\langle \nabla_{i_t} f(\mathbf{W}_t), G^{i_t}_{L_{i_t}}(\mathbf{W}_t) \right\rangle + \frac{1}{2L_{i_t}} \|G^{i_t}_{L_{i_t}}(\mathbf{W}_t)\|^2 + g(\mathbf{W}_{t+1}). \quad (31)$$

Taking expectation of both sides of the last inequality w.r.t. i_t, we obtain

$$\mathbb{E}_{i_t}[F(\mathbf{W}_{t+1})] \leq f(\mathbf{W}_t) - \frac{1}{d} \sum_{i=1}^{d} \frac{1}{L_i} \left\langle \nabla_i f(\mathbf{W}_t), G^{i_t}_{L_{i_t}}(\mathbf{W}_t) \right\rangle + \frac{1}{2d} \|\widetilde{G}(\mathbf{W}_t)\|^2_{L,*} + \mathbb{E}_{i_t}[g(\mathbf{W}_{t+1})]. \quad (32)$$

Since $\mathbf{W}^{i_t}_{t+1} = \text{prox}_{\frac{1}{L_{i_t}} g_{i_t}}(\mathbf{W}^{i_t}_t - \frac{1}{L_{i_t}} \nabla_{i_t} f(\mathbf{W}_t))$, by second prox theorem again:

$$g_{i_t}(\mathbf{W}^{i_t}_*) \geq g_{i_t}\left(\mathbf{W}^{i_t}_t - \frac{1}{L_{i_t}} G^{i_t}_{L_{i_t}}(\mathbf{W}_t)\right) + \left\langle -\nabla_{i_t} f(\mathbf{W}_t) + G^{i_t}_{L_{i_t}}(\mathbf{W}_t),\ \mathbf{W}^{i_t}_* - \mathbf{W}^{i_t}_t + \frac{1}{L_{i_t}} G^{i_t}_{L_{i_t}}(\mathbf{W}_t) \right\rangle. \quad (33)$$

Note that:

$$\mathbb{E}_{i_t}[g_{i_t}(\mathbf{W}^{i_t}_*)] = \frac{1}{d} g(\mathbf{W}_*) \mathbb{E}_{i_t}[g(\mathbf{W}_{t+1})] = \frac{d-1}{d} g(\mathbf{W}_t) + \frac{1}{d} \sum_{j=1}^{d} g_j \left(\mathbf{W}^j_t - \frac{1}{L_j} G^j_{L_j}(\mathbf{W}_t)\right). \quad (34)$$

By plugging the above into expectation of Eq. (33):

$$\begin{aligned} &\mathbb{E}_{i_t}[g(\mathbf{W}_{t+1})] - \frac{1}{d} \sum_{j=1}^{d} \frac{1}{L_j} \left\langle \nabla_j f(\mathbf{W}_t), G^j_{L_j}(\mathbf{W}_t) \right\rangle \\ &\leq \frac{1}{d} g(\mathbf{W}_*) + \frac{d-1}{d} g(\mathbf{W}_t) + \frac{1}{d} \left\langle \nabla f(\mathbf{W}_t) - \widetilde{G}(\mathbf{W}_t), \mathbf{W}_* - \mathbf{W}_t \right\rangle - \frac{1}{d} \|\widetilde{G}(\mathbf{W}_t)\|^2_{L,*}. \end{aligned} \quad (35)$$

Plugging the above into Eq. (32), we have

$$\mathbb{E}_{i_t}(F(\mathbf{W}_{t+1})) \leq f(\mathbf{W}_t) - \frac{1}{2d} \|\widetilde{G}(\mathbf{W}_t)\|^2_{L,*} + \frac{1}{d} g(\mathbf{W}_*) + \frac{1}{d} \left\langle \nabla f(\mathbf{W}_t) - \widetilde{G}(\mathbf{W}_t),\ \mathbf{W}_* - \mathbf{W}_t \right\rangle + \frac{d-1}{d} g(\mathbf{W}_t). \quad (36)$$

Combining it with convexity property $\langle \nabla f(\mathbf{W}_t),\ \mathbf{W}_* - \mathbf{W}_t \rangle \leq f(\mathbf{W}_*) - f(\mathbf{W}_t)$, we obtain:

$$\mathbb{E}_{i_t}(F(\mathbf{W}_{t+1})) \leq \frac{d-1}{d} F(\mathbf{W}_t) + \frac{1}{d} F(\mathbf{W}_*) - \frac{1}{2d} \|\widetilde{G}(\mathbf{W}_t)\|^2_{L,*} - \frac{1}{d} \left\langle \widetilde{G}(\mathbf{W}_t),\ \mathbf{W}_* - \mathbf{W}_t \right\rangle. \quad (37)$$

The above inequality, combined with Eq. 29, yields the relation:

$$\mathbb{E}_{i_t}\left(\frac{1}{2} r^2_{t+1}\right) \leq \frac{1}{2} r^2_t + \frac{d-1}{d} F(\mathbf{W}_t) + \frac{1}{d} F(\mathbf{W}_*) - \mathbb{E}_{i_t}(F(\mathbf{W}_{t+1})), \quad (38)$$

which can be rearranged as

$$\mathbb{E}_{i_t}\left(\frac{1}{2} r^2_{t+1} + F(\mathbf{W}_{t+1}) - F_{\text{opt}}\right) \leq \left(\frac{1}{2} r^2_t + F(\mathbf{W}_t) - F_{\text{opt}}\right) - \frac{1}{d}\left(F(\mathbf{W}_t) - F_{\text{opt}}\right). \quad (39)$$

Taking expectation over $\xi_{t-1} = \{i_0, i_1, \ldots, i_{t-1}\}$ (which is a multivariate random variable) of both sides we obtain

$$\mathbb{E}_{\xi_t}\left(\frac{1}{2} r^2_{t+1} + F(\mathbf{W}_{t+1}) - F_{\text{opt}}\right) \leq \mathbb{E}_{\xi_{t-1}}\left(\frac{1}{2} r^2_t + F(\mathbf{W}_t) - F_{\text{opt}}\right) - \frac{1}{d}\left(\mathbb{E}_{\xi_{t-1}}(F(\mathbf{W}_t)) - F_{\text{opt}}\right). \quad (40)$$

Therefore, we can conclude that:

$$\mathbb{E}_{\xi_t}\left(F(\mathbf{W}_{t+1})\right) - F_{\text{opt}} \leq \mathbb{E}_{\xi_t}\left(\frac{1}{2}r_{t+1}^2 + F(\mathbf{W}_{t+1}) - F_{\text{opt}}\right) \tag{41}$$

$$\leq \frac{1}{2}r_0^2 + F(\mathbf{W}_0) - F_{\text{opt}} - \frac{1}{d}\sum_{j=0}^{t}\left(\mathbb{E}_{\xi_{j-1}}\left(F(\mathbf{W}_j)\right) - F_{\text{opt}}\right). \tag{42}$$

Since the objective is always decreasing (each block will decrease), we have:

$$\mathbb{E}_{\xi_t}(F(\mathbf{W}_{t+1})) - F_{\text{opt}} \leq \frac{1}{2}r_0^2 + F(\mathbf{W}_0) - F_{\text{opt}} - \frac{t+1}{d}\left(\mathbb{E}_{\xi_t}(F(\mathbf{W}_{t+1})) - F_{\text{opt}}\right), \tag{43}$$

the desired result follows immediately from the above after simple rearrangement. ■

Appendix D. Bound on $\lambda_{max}(\mathbf{J})$

We now turn to bound $\lambda_{max}(\mathbf{J})$. Since $\mathbf{J}$ is PSD, by definition:

$$\lambda_{max}(\mathbf{J}) = \max_{\|\mathbf{v}\|=1} \mathbf{v}^\top \mathbf{J}\mathbf{v} = \mathbf{v}^\top \operatorname{diag}(\mathbf{p})\,\mathbf{v} - (\mathbf{p}^\top\mathbf{v})^2 = \sum_{i=1}^{d}\mathbf{p}_i\mathbf{v}_i^2 - \Big(\sum_{i=1}^{d}\mathbf{p}_i\mathbf{v}_i\Big)^2 = \operatorname{Var}_{\mathbf{p}}(\mathbf{v}).$$

Given the constraint $\|\mathbf{v}\| = 1$, and according the Popoviciu's inequality on variances, we obtain the desired result in Eq. (7).
In fact, we can also prove the above conclusion from another perspective:

Gershgorin circle theorem Every eigenvalue of $\mathbf{J}$ lies within at least one of the Gershgorin discs [6] $D(\mathbf{p}_i(1-\mathbf{p}_i), \mathbf{p}_i(1-\mathbf{p}_i))$, therefore $\lambda \leq 2\mathbf{p}_i(1-\mathbf{p}_i) = 2 * \frac{1}{4} = \frac{1}{2}$.

Appendix E. Bound on L for whole (matrix-wise) update

Given the fact that

$$\mathbf{H} \;=\; \sum_{i=1}^{n}\big(\operatorname{diag}(\mathbf{p}_i) - \mathbf{p}_i\mathbf{p}_i^\top\big) \otimes \big(\mathbf{x}_i\mathbf{x}_i^\top\big),$$

we have $\|\mathbf{H}\|_2 \leq \frac{1}{2}\|\sum_{i=1}^{n}\big(\mathbf{x}_i\mathbf{x}_i^\top\big)\|_2 = \frac{1}{2}\|\mathbf{X}\mathbf{X}^\top\|_2 = \frac{1}{2}\|\mathbf{X}\|_2^2$, where the first inequality we make use of the fact of Eq. (7).

Appendix F. Class-wise L_k

For sake of convenience, we denote $\tilde{\mathbf{H}}$ as the permutation of Hessian, which is given by:

$$\tilde{\mathbf{H}} \;=\; \sum_{i=1}^{n}\big(\mathbf{x}_i\mathbf{x}_i^\top\big) \otimes \big(\operatorname{diag}(\mathbf{p}_i) - \mathbf{p}_i\mathbf{p}_i^\top\big),$$

we now consider L_k for class-wise update. Simple manipulation gives

$$\|\tilde{\mathbf{H}}(k,k)\|_2 \;=\; \|\sum_{i=1}^{n}\big(\mathbf{x}_i\mathbf{x}_i^\top\big)\mathbf{p}_{i,k}(1-\mathbf{p}_{i,k})\|_2 \leq \frac{1}{4}\|\mathbf{X}\mathbf{X}^\top\|_2 = \frac{1}{4}\|\mathbf{X}\|_2^2,$$

which results in same conclusion as Eq. (3).