A Note on Diagonal Preconditioner in Large-scale Logistic Regression

Wei-Sheng Chin, Bo-Wen Yuan, Meng-Yuan Yang, and Chih-Jen Lin

Department of Computer Science, National Taiwan University, Taipei 10617, Taiwan

Preconditioning is a common technique to accelerate conjugate gradient methods. Although [2] has conducted some comparisons, they only show the total number of CG iterations needed to reach the following stopping condition.

$$\|\nabla f(\boldsymbol{w})\|_\infty \leq 0.001. \tag{1}$$

Their conclusion is that diagonal preconditioner may not be always useful. However, we find that (1) may be too tight in practice, so the effectiveness of diagonal preconditioning is still unclear if the training program stops earlier. To this end, we follow the settings in [2, Section 5] to redo their experiments with more details provided.

We consider the same six data sets and the same $C = 16$ used in [2]. In addition to the trust-region Newton method proposed in [2], we consider another setting that replaces trust-region technique with line search. For each data set, we check the relation between the accumulated CG iterations and the relative difference to the optimal function value

$$\frac{f - f^*}{f^*},$$

where f is the function value at a specific number of accumulated CG iterations and f^* is the optimal solution. For each data set, two horizontal lines are drawn to respectively indicate the stopping condition (1) and the default stopping condition (below) in LIBLINEAR [1] are satisfied.

$$\|\nabla f(\boldsymbol{w})\|_2 \leq \frac{\min(l^+, l^-)}{100l} \|\nabla f(\boldsymbol{w}_{\text{init}})\|_2, \tag{2}$$

where l^+ is the number of position instances, l^- is the number of negative instances, l is the number of total instances, and $\boldsymbol{w}_{\text{init}}$ stands for the initial solution. Note that following [2], we add a bias term in the model and the regularization term. That is,

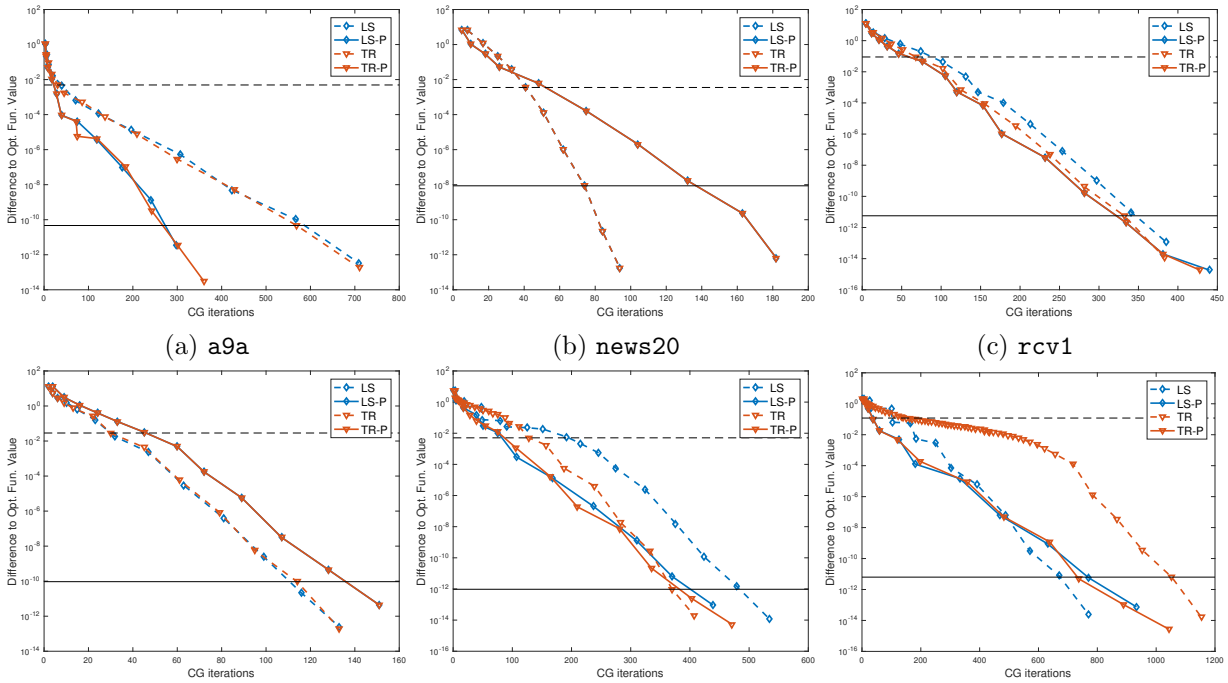
$$\boldsymbol{x}_i^T \leftarrow [\boldsymbol{x}_i^T, 1] \qquad \boldsymbol{w}^T \leftarrow [\boldsymbol{w}^T, b].$$

Our results are shown in Figure 1. We have the following observations.

- The stopping condition (1) used in [2] is much tighter than the one that is more often used in practice.
- In the early stage, PCG helps to more quickly or at least comparably reduce the function value, but it may not be better in a later stage.

In summary, our results are consistent with the conclusion in [2], which is based on results of using a strict stopping condition. However, in practice the Newton method stops much earlier. In such a situation our experiments show that PCG is useful or at least not harmful for training logistic regression. Programs used for this experiments are available at

`www.csie.ntu.edu.tw/~cjlin/papers/logistic`.



(a) `a9a` (b) `news20` (c) `rcv1`

(d) `real-sim` (e) `yahoo-japan` (f) `yahoo-korea`

Figure 1: Effects of diagonal preconditioning for truncated Newton methods. TR: trust-region technique is used. LS: line search is used. Note that "P" indicates the cases with preconditioning. Dashed horizontal line: (2) is satisfied. Solid horizontal line: (1) is satisfied.

References

[1] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: a library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[2] C.-J. Lin, R. C. Weng, and S. S. Keerthi. Trust region Newton method for large-scale logistic regression. *Journal of Machine Learning Research*, 9:627–650, 2008.