
A Quasi-Newton Method Based Vertical Federated Learning Framework for Logistic Regression

Kai Yang
WeBank & ShanghaiTech University
yangkai@shanghaitech.edu.cn

Tao Fan
WeBank
dylanfan@webank.com

Tianjian Chen
WeBank
tobychen@webank.com

Yuanming Shi
ShanghaiTech University
shiyu@shanghaitech.edu.cn

Qiang Yang
Hong Kong University of Science and Technology
yangqiang@hkust.edu.cn

Abstract

Data privacy and security becomes a major concern in building machine learning models from different data providers. Federated learning shows promise by leaving data at providers locally and exchanging encrypted information. This paper studies the vertical federated learning structure for logistic regression where the data sets at two parties have the same sample IDs but own disjoint subsets of features. Existing frameworks adopt the first-order stochastic gradient descent algorithm, which requires large number of communication rounds. To address the communication challenge, we propose a quasi-Newton method based vertical federated learning framework for logistic regression under the additively homomorphic encryption scheme. Our approach can considerably reduce the number of communication rounds with a little additional communication cost per round. Numerical results demonstrate the advantages of our approach over the first-order method.

1 Introduction

With the surge of artificial intelligence (AI) driven services including recommender system and natural language processing, data privacy and security have raised worldwide concerns [1]. More and more stringent requirements of data privacy and security become an emerging trend of laws and regulations from states across the world. A known example is the General Data Protection Regulation (GDPR) by the European Union [2]. Traditional AI service providers usually collect and transfer data instances from one party to another party. Then a machine learning model is trained at the cloud data center with the fused data set. However, it faces challenges of data breach and violation of data protection laws and regulations [3].

Recently, federated learning [1, 4, 5] is an emerging frontier field studying privacy-preserving collaborative machine learning while leaving data instances at their providers locally. A line of works [4, 5, 6] focus on the horizontal structure, in which each node has a subset of data instances with complete data attributes. There are also many researches studying the vertical federated learning structure where the data set is vertically partitioned and owned by different data providers. That is, each data provider holds a disjoint subset of attributes for all data instances. The target is to learn a machine learning model collaboratively without transferring any data from one data provider to another. In particular, [7] proposes a privacy-preserving tree-boosting system *SecureBoost* and [8] propose a logistic regression framework for vertically partitioned data.

Communication is one of the main bottlenecks in federated learning due to the much worse network conditions than the cloud computing center [4]. To address the communication challenge in horizontal federated learning, structured updates are considered in [4] to reduce the communication costs per round and an iterative model averaging algorithm is proposed in [5] to reduce the number of communication rounds. For vertical federated learning structure, [8] considers a two party (denoted by party A and party B) logistic regression problem and proposes a stochastic gradient descent (SGD) method based privacy-preserving framework. Due to the slow convergence of first-order algorithms, it requires a large number of communication rounds. This work shall propose a quasi-Newton method based vertical federated learning system with sub-sampled Hessian information to reduce the communication round.

Related Works Second-order Newton’s method is known to converge faster than first-order gradient based methods. To avoid the high cost of computing the inversion of Hessian matrix, a well recognized quasi-Newton method Limited-memory BFGS (L-BFGS) [9] algorithm is proposed by directly approximating inverse Hessian matrix. There are a number of works [10, 11, 12] focus on developing stochastic quasi-Newton algorithms for problems with large amounts of data. However, the inverse Hessian estimated by [10] may be not stable for small batch sizes and the algorithm in [12] requires computing the full gradient which would double the communication cost in each epoch compared with SGD. This paper develops a communication efficient vertical federated learning framework based on the stochastic quasi-Newton method proposed in [11].

2 Problem Statement

Consider a typical logistic regression problem with vertically partitioned data [8]. Let $\mathbf{X} \in \mathbb{R}^{n \times T}$ be the data set consisting of T data samples and each instance has n features. The class attribute information, i.e., the label of data, is given by $\mathbf{y} \in \{-1, +1\}^T$. The data set is vertically partitioned and distributed on two honest-but-curious private parties A (the **host** data provider with only features) and B (the **guest** data provider with features and labels). Let $\mathbf{X}^A \in \mathbb{R}^{n_A \times T}$ be the data set owned by party A and $\mathbf{X}^B \in \mathbb{R}^{n_B \times T}$ owned by party B. Each party owns a disjoint subset of data features over a common sample IDs with $\mathbf{X} = (\mathbf{X}^A, \mathbf{X}^B)$. In addition, only party B has access to the labels \mathbf{y} . The target of logistic regression is to train a linear model for classification by solving

$$\underset{\mathbf{w} \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{T} \sum_i^T l(\mathbf{w}; \mathbf{x}_i, y_i), \quad (1)$$

where \mathbf{w} is the model parameters, \mathbf{x}_i is the i -th data instance and y_i is the corresponding label. The negative log-likelihood loss function is given by $l(\mathbf{w}; \mathbf{x}_i, y_i) = \log(1 + \exp(y_i \mathbf{w}^\top \mathbf{x}_i))$. In this paper, we suppose that party A and party B hold the model parameters corresponding to their features respectively, which can be denoted as $\mathbf{w} = (\mathbf{w}^A, \mathbf{w}^B)$ where $\mathbf{w}^A \in \mathbb{R}^{n_A}$ and $\mathbf{w}^B \in \mathbb{R}^{n_B}$.

[8] proposes a stochastic gradient descent (SGD) based vertical logistic regression framework by computing gradients via exchanging encrypted intermediate values at each iteration. Specifically, party A and party B collaboratively compute the vertically partitioned encrypted gradient $\mathbf{g}^A \in \mathbb{R}^{n_A}$ and $\mathbf{g}^B \in \mathbb{R}^{n_B}$, which can be decrypted by the third party. To achieve secure computation without transferring data from one party to another, the additively homomorphic encryption is adopted. Additively homomorphic encryption schemes such as Paillier [13] allow any party can encrypt their data with a public key, while the private key for decryption is owned by the third party, i.e., the **coordinator**. With additively homomorphic encryption we can compute the additive of two encrypted numbers as well as the product of an unencrypted number and an encrypted one, which can be denoted as $\llbracket u \rrbracket + \llbracket v \rrbracket = \llbracket u + v \rrbracket$, $v \cdot \llbracket u \rrbracket = \llbracket vu \rrbracket$ by using $\llbracket \cdot \rrbracket$ as the encryption operation. Unfortunately, the loss function and its gradient cannot be computed directly with additively homomorphic encryption. To address this issue, we will adopt the Taylor approximation for the loss function is proposed in [8, 14] as

$$\text{Taylor loss: } l(\mathbf{w}; \mathbf{x}_i, y_i) \approx \log 2 - \frac{1}{2} y_i \mathbf{w}^\top \mathbf{x}_i + \frac{1}{8} (\mathbf{w}^\top \mathbf{x}_i)^2. \quad (2)$$

3 A Quasi-Newton Method Based Vertical Federated Learning Framework

In federated learning, the communication cost between different parties is much more expensive than it in the cloud computing center since the data providers are usually across distant data centers, across

different networks, or even in a wireless environment with limited bandwidth [5]. So it becomes one of the main bottlenecks for efficiently model training. For this reason, we develop a communication efficient vertical federated learning framework by incorporating second-order information [11] to reduce the communication rounds between parties, which is illustrated in Fig. 1.

The gradient and the Hessian of the Taylor loss in equation (2) with respect to the i -th data instance are respectively given by $\nabla l(\mathbf{w}; \mathbf{x}_i, y_i) \approx (\frac{1}{4}\mathbf{w}^\top \mathbf{x}_i - \frac{1}{2}y_i) \mathbf{x}_i$, $\nabla^2 l(\mathbf{w}; \mathbf{x}_i, y_i) \approx \frac{1}{4}\mathbf{x}_i \mathbf{x}_i^\top$. In the k -th iteration, classical L-BFGS algorithm uses the history information in last M iterations by differencing gradient and model parameters between every two consecutive iterations to obtain an estimated inverse Hessian matrix $\mathbf{H} \in \mathbb{R}^{n \times n}$. But it will lead to a unstable curvature estimation if we use mini-batch data instead of full data. To obtain a stable estimation of \mathbf{H} , we shall use the sub-sampled Hessian information as suggested by [11]. Moreover, the curvature information \mathbf{H} can be updated every L iterations to reduce the communication overhead as well as improve the stability of quasi-Newton algorithm. The details of computing the key ingredients for our system are introduced in the following part.

Computing Loss and Gradient at Party A&B Let $\mathcal{S} \subseteq \{1, \dots, T\}$ be the index set of the chosen mini-batch data instances. The corresponding loss and gradient are given by $\text{loss} = F(\mathbf{w}) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} l(\mathbf{w}; \mathbf{x}_i, y_i)$, $\mathbf{g} = \nabla F(\mathbf{w}) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \nabla l(\mathbf{w}; \mathbf{x}_i, y_i)$. By denoting $\mathbf{u}_A = \{\mathbf{u}_A[i] = \mathbf{w}^A^\top \mathbf{x}_i^A : i \in \mathcal{S}\}$, $\mathbf{u}_A^2 = \{\mathbf{u}_A^2[i] = (\mathbf{w}^A^\top \mathbf{x}_i^A)^2 : i \in \mathcal{S}\}$ for party A (similarly \mathbf{u}_B and \mathbf{u}_B^2 for party B) and $\mathbf{d} = \{d_i : i \in \mathcal{S}\}$, the encrypted loss and gradient can be computed by transmitting $\llbracket \mathbf{u}_A \rrbracket$ from party A to party B, and transmitting $\llbracket \mathbf{d} \rrbracket$ from B to A following

$$\llbracket \text{loss} \rrbracket \approx \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \llbracket \log 2 \rrbracket - \frac{1}{2} y_i (\llbracket \mathbf{u}_A[i] \rrbracket + \llbracket \mathbf{u}_B[i] \rrbracket) + \frac{1}{8} (\llbracket \mathbf{u}_A^2[i] \rrbracket + 2\mathbf{u}_B[i] \llbracket \mathbf{u}_A[i] \rrbracket + \llbracket \mathbf{u}_B^2[i] \rrbracket) \quad (3)$$

$$\llbracket \mathbf{g} \rrbracket \approx \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \llbracket d_i \rrbracket \mathbf{x}_i = \underbrace{\left(\sum_{i \in \mathcal{S}} \llbracket d_i \rrbracket \mathbf{x}_i^A \right)}_{\llbracket \mathbf{g}^A \rrbracket}, \underbrace{\left(\sum_{i \in \mathcal{S}} \llbracket d_i \rrbracket \mathbf{x}_i^B \right)}_{\llbracket \mathbf{g}^B \rrbracket}, \llbracket d_i \rrbracket = \frac{1}{4} (\llbracket \mathbf{u}_A[i] \rrbracket + \llbracket \mathbf{u}_B[i] \rrbracket + \llbracket -\frac{1}{2} y_i \rrbracket). \quad (4)$$

Computing Updates for Estimating Curvature Information at Party A&B To achieve cheap communication costs introduced additionally, the curvature information \mathbf{H} is updated every L iterations at the coordinator by collecting encrypted $\mathbf{v} = (\mathbf{v}^A, \mathbf{v}^B) \in \mathbb{R}^n$ from party A and B. Specifically, every L iterations we shall compute the difference of average model parameters as

$$\mathbf{s}_t = \bar{\mathbf{w}}_t - \bar{\mathbf{w}}_{t-1} = (\mathbf{s}_t^A, \mathbf{s}_t^B), \bar{\mathbf{w}}_t = \sum_{i=k-L+1}^k \mathbf{w}_i / L, \bar{\mathbf{w}}_{t-1} = \sum_{i=k-2L+1}^{k-L} \mathbf{w}_i / L \quad (5)$$

at party A and party B. Then the product of sub-sampled Hessian $\nabla^2 \hat{F}(\bar{\mathbf{w}}_t)$ and average model difference \mathbf{s}_t are given by

$$\mathbf{v}_t = \nabla^2 \hat{F}(\bar{\mathbf{w}}_t) \mathbf{s}_t, \text{ where } \nabla^2 \hat{F}(\bar{\mathbf{w}}_t) = \frac{1}{|\mathcal{S}_H|} \sum_{i \in \mathcal{S}_H} \nabla^2 l(\bar{\mathbf{w}}_t; \mathbf{x}_i, y_i) = \frac{1}{|\mathcal{S}_H|} \sum_{i \in \mathcal{S}_H} \mathbf{x}_i \mathbf{x}_i^\top. \quad (6)$$

The sub-sampled Hessian is calculated with respect to a randomly chosen subset of data \mathcal{S}_H . Under additively homomorphic encryption, $\llbracket \mathbf{v}_t \rrbracket$ can be computed following

$$\llbracket \mathbf{v}_t \rrbracket = \frac{1}{|\mathcal{S}_H|} \sum_{i \in \mathcal{S}_H} \llbracket h_i \rrbracket \mathbf{x}_i = (\llbracket \mathbf{v}_t^A \rrbracket, \llbracket \mathbf{v}_t^B \rrbracket) = \left(\frac{1}{|\mathcal{S}_H|} \sum_{i \in \mathcal{S}_H} \llbracket h_i \rrbracket \mathbf{x}_i^A, \frac{1}{|\mathcal{S}_H|} \sum_{i \in \mathcal{S}_H} \llbracket h_i \rrbracket \mathbf{x}_i^B \right), \quad (7)$$

where $h_i = \Delta \bar{u}_i^A + \Delta \bar{u}_i^B = \mathbf{s}_t^A^\top \mathbf{x}_i^A + \mathbf{s}_t^B^\top \mathbf{x}_i^B$. By transmitting $\llbracket \Delta \bar{u}_A \rrbracket = \{\llbracket \Delta \bar{u}_i^A \rrbracket : i \in \mathcal{S}_H\}$ from party A to party B, and transmitting $\llbracket \mathbf{h} \rrbracket = \{\llbracket h_i \rrbracket : i \in \mathcal{S}_H\}$ from B to A, the corresponding components $\llbracket \mathbf{v}_t^A \rrbracket$ can be computed at party A and $\llbracket \mathbf{v}_t^B \rrbracket$ is computed at party B privately.

Computing Descent Direction at the Coordinator After collecting the encrypted loss $\llbracket \text{loss} \rrbracket$, gradient $\llbracket \mathbf{g} \rrbracket$, and $\llbracket \mathbf{v} \rrbracket$ from party A&B, the coordinator should determine a descent direction $\tilde{\mathbf{g}}$ for updating \mathbf{w}^A and \mathbf{w}^B , i.e., $\mathbf{w} \leftarrow \mathbf{w} - \tilde{\mathbf{g}} = (\mathbf{w}^A - \tilde{\mathbf{g}}^A, \mathbf{w}^B - \tilde{\mathbf{g}}^B)$. Given an estimated \mathbf{H} , the descent direction is given by $\tilde{\mathbf{g}} = \eta \mathbf{H} \mathbf{g}$ where $\eta > 0$ is the learning rate. Every L iterations, \mathbf{v} and \mathbf{s} are stored in two queues with length M . \mathbf{H} is determined by successively computing

$$\mathbf{H} \leftarrow (\mathbf{I} - \rho_j \mathbf{s}_j \mathbf{s}_j^\top) \mathbf{H} (\mathbf{I} - \rho_j \mathbf{v}_j \mathbf{v}_j^\top) + \rho_j \mathbf{s}_j \mathbf{s}_j^\top, \rho_j = 1/(\mathbf{v}_j^\top \mathbf{s}_j), \forall j = t - M + 1, \dots, t \quad (8)$$

from the initial point $\mathbf{H} = (\mathbf{v}_t^\top \mathbf{s}_t / \mathbf{v}_t^\top \mathbf{v}_t) \mathbf{I}$. It should be noted that \mathbf{s}_t can be computed locally at the coordinator as $\mathbf{s}_t = \sum_{i=k-L+1}^k \tilde{\mathbf{g}}_i / L - \sum_{i=k-2L+1}^{k-L} \tilde{\mathbf{g}}_i / L$ without any additional transmissions. The overall quasi-Newton method based vertical federated learning framework is illustrated in Fig. 1. The source code will be released in an upcoming version of the FATE framework [15].

At each iteration, the communication costs of SGD are $3|\mathcal{S}|$ encrypted numbers between party A and party B, and $2n$ encrypted numbers between party A&B and the coordinator. With our quasi-Newton framework, the communication costs become $3|\mathcal{S}| + 2|\mathcal{S}_H|/L$ encrypted numbers between party A and party B, and $(2 + 1/L)n$ encrypted numbers between party A&B and the coordinator. By choosing $|\mathcal{S}_H| \leq |\mathcal{S}|$, the presented quasi-Newton method introduces no more than $1/L$ additional communication costs at per communication round compared with [8].

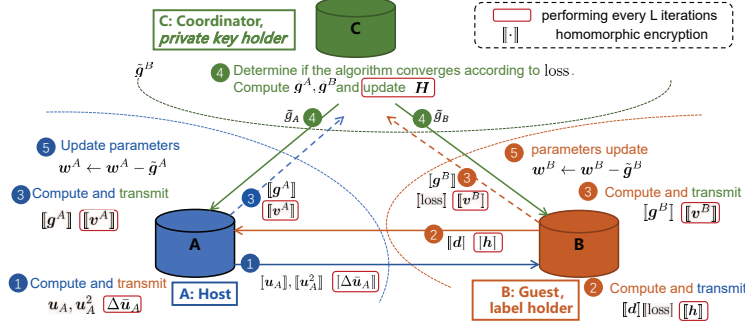


Figure 1: A Quasi-Newton Framework for Vertical Federated Learning

4 Experiments and Conclusion

We conduct numerical experiments on two credit scoring data sets to test the advantages of our system over the mini-batch SGD method based system in [8]. **Credit 1** [16]: It consists of 30000 data instances and each instance has $n = 25$ attributes; 2) **Credit 2** [17]: It contains 150000 data instances and each with 10 attributes. By splitting each data set into two parts vertically, each party holds a subset of features and party B also holds the labels. We randomly choose 80% data instances as the training set and the remaining 20% as the test set. We choose $\mathcal{S}_H = \mathcal{S}$ and $L = 4$ in all simulations and each algorithm stops when the loss between two consecutive epochs is less than 10^{-5} . The number of epochs, the training loss and the area under the curve (AUC) of the receiver operating characteristics (ROC) curve on the test set are shown in table 1. Numerical results demonstrate that the proposed system requires less communication overhead than the first-order SGD based framework.

Table 1: Numerical Results on Two Public Data Sets

| Batch Size | Method | Credit 1 | | | Credit 2 | | |
|------------|----------|----------|----------|--------|----------|----------|--------|
| | | Epochs | Loss | AUC | Epochs | Loss | AUC |
| 1000 | SGD | 12 | 0.496218 | 0.7224 | 12 | 0.314555 | 0.7033 |
| | Proposed | 3 | 0.496600 | 0.7222 | 4 | 0.314643 | 0.7061 |
| 3000 | SGD | 18 | 0.496194 | 0.7219 | 14 | 0.314648 | 0.6982 |
| | Proposed | 12 | 0.496317 | 0.7225 | 6 | 0.314490 | 0.7077 |

In this paper, we consider the communication challenges in vertical federated learning problem with two data providers for learning a logistic regression model collaboratively. We propose to use a quasi-Newton method to reduce the number of communication rounds. With the additively homomorphic encryption scheme, two data providers compute an encrypted gradient by exchanging encrypted intermediate values, and an additional vector every L iterations for updating the curvature information. Numerical experiment demonstrate that our method considerably reduces the number of communication rounds with a little additional communication cost per round.

References

- [1] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):12, 2019.
- [2] Jan Philipp Albrecht. How the GDPR will change the world. *Eur. Data Prot. L. Rev.*, 2:287, 2016.
- [3] Wiki. Data breach — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Data_breach&oldid=912247856, 2019.
- [4] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.
- [5] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282, 2017.
- [6] Kai Yang, Tao Jiang, Yuanming Shi, and Zhi Ding. Federated learning via over-the-air computation. *arXiv preprint arXiv:1812.11750*, 2018.
- [7] Kewei Cheng, Tao Fan, Yilun Jin, Yang Liu, Tianjian Chen, and Qiang Yang. Secureboost: A lossless federated learning framework. *arXiv preprint arXiv:1901.08755*, 2019.
- [8] Stephen Hardy, Wilko Henecka, Hamish Ivey-Law, Richard Nock, Giorgio Patrini, Guillaume Smith, and Brian Thorne. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. *arXiv preprint arXiv:1711.10677*, 2017.
- [9] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.
- [10] Nicol N Schraudolph, Jin Yu, and Simon Günter. A stochastic quasi-newton method for online convex optimization. In *Artificial intelligence and statistics*, pages 436–443, 2007.
- [11] Richard H Byrd, Samantha L Hansen, Jorge Nocedal, and Yoram Singer. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.
- [12] Philipp Moritz, Robert Nishihara, and Michael Jordan. A linearly-convergent stochastic L-BFGS algorithm. In *Artificial Intelligence and Statistics*, pages 249–258, 2016.
- [13] Pascal Paillier. Public-key cryptosystems based on composite degree residuosity classes. In *International Conference on the Theory and Applications of Cryptographic Techniques*, pages 223–238. Springer, 1999.
- [14] Yoshinori Aono, Takuya Hayashi, Le Trieu Phong, and Lihua Wang. Scalable and secure logistic regression via homomorphic encryption. In *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*, pages 142–144. ACM, 2016.
- [15] WeBank. FATE: An industrial grade federated learning framework. <https://fate.fedai.org>, 2018.
- [16] UCI Machine Learning Repository. default of credit card clients data set. <https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>, 2017.
- [17] Give me some credit. Give me some credit. <https://www.kaggle.com/c/GiveMeSomeCredit/data>, 2011.

Appendix A

We provide details of the proposed vertical federated learning framework in Algorithm 1.

Algorithm 1: A Quasi-Newton Framework for Vertical Federated Learning

Input : w_0^A, w_0^B, M, L
Output : w^A, w^B

```

1 Set  $t = 0, H = I$ 
2 for each round  $k = 1, \dots$ , do
3   Choose a minibatch  $\mathcal{S}$ 
4   if  $\text{mod}(k, L) \neq 0$  then
5     Party A&B: compute  $\llbracket \text{loss} \rrbracket, \llbracket g \rrbracket$  as equation (3) (4)
6     Coordinator:  $w_{k+1} = w_k - \tilde{g}_k$  where  $\tilde{g}_k = \eta H g$ 
7   else
8      $t \leftarrow t + 1$ 
9     Party A&B: Choose a minibatch  $\mathcal{S}_H$ 
10    compute  $\llbracket \text{loss} \rrbracket, \llbracket g \rrbracket, \llbracket v_t \rrbracket$  as equation (3) (4) (6)
11    Coordinator:  $w_{k+1} = w_k - \tilde{g}_k$  where  $\tilde{g}_k = \eta H g$ 
12     $s_t = \sum_{i=k-L+1}^k \tilde{g}_i / L - \sum_{i=k-2L+1}^{k-L} \tilde{g}_i / L$ 
13    if  $t > 1$  then
14       $H \leftarrow (s_t^T v_t) / (v_t^T v_t) I, \tilde{m} = \min\{M, t\}$ 
15      for  $j = t - \tilde{m} + 1, \dots, t$  do
16         $\rho_j = 1 / (v_j^T s_j)$ 
17         $H \leftarrow (I - \rho_j s_j v_j^T) H (I - \rho_j v_j s_j^T) + \rho_j s_j s_j^T$ 
18      end
19    end
20     $\tilde{w}_t = 0$ 
21  end
22 end
```
