Distributed Asynchronous Stochastic Dual Coordinate Ascent without Duality

Zhouyuan Huo  
zhouyuan.huo@mavs.uta.edu

Heng Huang  
heng@uta.edu

May 31, 2016

Abstract

In this paper, we propose new Distributed Asynchronous Dual-Free Coordinate Ascent method (Asy-df SDCA), and provide the proof of convergence rate for two cases: the individual loss is convex and the individual loss is non-convex but its expected loss is convex. Stochastic Dual Coordinate Ascent (SDCA) model is a popular method and often has better performances than stochastic gradient descent methods in solving regularized convex loss minimization problems. Dual-Free Stochastic Dual Coordinate Ascent method is a variation of SDCA, and can be applied to non-convex problem when its dual problem is meaningless. We extend Dual-Free Stochastic Dual Coordinate Ascent method to the distributed mode with considering the star network in this paper.

1 Introduction

We consider the following ℓ₂-norm regularized loss minimization problem:

\[
\min_{w \in \mathbb{R}^d} P(w) = \frac{1}{n} \sum_{i=1}^{n} \phi_i(w) + \frac{\lambda}{2} ||w||^2.
\]  (1)

Many optimization methods have been proposed to solve this problem including [6, 15, 2, 12, 13, 14, 16, 18]. Experimental results in [18] verify that SDCA method enjoys strong theoretical convergence guarantee properties and often has better performances than stochastic gradient descent (SGD) based methods. In [6], the paper points out that SDCA is a variation of SGD method, and its update is based on an unbiased estimate of gradient. Unlike most of the SGD methods which solve primal problem directly, as its name indicates, SDCA is derived by considering a dual problem of (1). However, the dual problem of \( \phi_i \) is meaningless sometimes. In [13], a variation of SDCA was proposed and applied to problems in which individual \( \phi_i \) is non-convex.

Recently, as the size of data and model grows larger and larger, many distributed optimization algorithms have been proposed to solve large-scale problems [9, 21, 22, 11, 8, 1]. There are mainly two architectures in distributed system: one is shared-memory architecture, and the other one is distributed-memory structure.
architecture. In this paper, we only consider distributed-memory architecture. In [19, 5, 17], distributed SDCA method was proposed with proved linear convergence when \( \phi_i \) is smooth and convex.

In this paper, we propose a Distributed Asynchronous Dual-Free Coordinate Ascent (Asy-df SDCA) method. The corresponding convergence analysis is provided on two different assumptions: one is that \( \phi_i \) is \( L \)-smooth and convex, and the other one is that \( \phi_i \) is \( L \)-smooth and non-convex, but the average of \( \phi_i \) is strongly convex.

## 2 Asynchronous Dual Free Stochastic Dual Coordinate Ascent Method

Details of our proposed Distributed Asynchronous Dual-Free Coordinate Ascent method (Asy-df SDCA) are described in Algorithms (1) and (2). Algorithm (1) presents the pseudo code of Asy-df SDCA on each worker node. \( \alpha_i \in \mathbb{R}^d, i \in \{1, \cdots, n\} \) denotes pseudo-dual vector for each sample, and they are maintained by workers. We assume datasets are evenly distributed in \( K \) workers, and there are \( n_k \) samples in worker \( k \). Algorithm (1) summarizes the pseudo code on server node. Parameter \( w \) is maintained in the server, and \( v_i \) represents update received from workers in each iteration.

### Algorithm 1 Asy-df SDCA (Worker \( k \))

Initialize \( \alpha_i^{0,0} \in \mathbb{R}^d, i \in \{1, \cdots, n_k\} \)

for \( s = 1, 2, \cdots, S \) do

for \( t = 1, 2, \cdots, n_k \) do

  Pull \( w^{s,t-\tau} \) from server.

  Randomly select sample \( i \) from \( \{1, \cdots, n_k\} \);

  \( v_i^{s,t} = \nabla \phi_i(w^{s,t-\tau}) + \alpha_i^{s,t-\tau} \)

  Update \( \alpha_i^{s,t} = \alpha_i^{s,t-1} - \lambda \eta v_i^{s,t} \)

  Push \( v_i^{s,t} \) to server.

end for

end for

### Algorithm 2 Asy-df SDCA (Server)

Initialize \( w^{0,0} \in \mathbb{R}^d \).

for \( s = 1, 2, \cdots, S \) do

for \( t = 1, 2, \cdots, n \) do

  Receive \( v_i^{s,t} \) from worker.

  Update \( w_i^{s,t} = w_i^{s,t-1} - \eta v_i^{s,t} \)

end for

\( w^{s+1,0} = w^{s,n} \)

end for
3 Convergence Analysis

We provide convergence analysis of our proposed method on two different cases: (1) \( \phi_i \) is \( L \)-smooth and convex, and (2) \( \phi_i \) is \( L \)-smooth and non-convex, but the average of \( \phi_i \) is strongly convex.

3.1 Convex Case

For further analysis, in this section, we make the following assumptions for problem (1). All of them are common assumptions in the theoretical analysis of distributed methods and stochastic gradient method.

**Assumption 1** We assume the following conditions hold:

- \( \phi_i \) is \( L \)-smooth,
  \[ \| \nabla \phi(x) - \nabla \phi_i(y) \| \leq L \| x - y \|. \]  
  \( (2) \)

- \( \phi_i \) is convex,
  \[ \phi_i(x) \geq \phi_i(y) + \nabla \phi_i(y)^T (x - y). \]  
  \( (3) \)

- Time delay \( \tau \) is no larger than \( \Delta \).

Following the above assumptions, we know that our method is able to have linear convergence rate with the following theorem.

**Theorem 1** When the above assumptions satisfy, and let \( w^* \) be the minimizer of \( P(w) \), \( \alpha_i^* = -\nabla \phi_i(w^*) \). If \( \eta \leq \frac{1}{2L + n\lambda + 4L\Delta} \), then we have

\[
E \left[ \| w^{s,0} - w^* \|^2 + \frac{1}{2L} \sum_{i=1}^{n} \| \alpha_i^{s,0} - \alpha_i^* \|^2 \right] 
\leq e^{-\eta^s} \left[ \| w^{0,0} - w^* \|^2 + \frac{1}{2L} \sum_{i=1}^{n} \| \alpha_i^{0,0} - \alpha_i^* \|^2 \right].
\]  
(4)

3.2 Non-Convex Case

For further analysis, in this section, we make the following assumptions for problem (1).

**Assumption 2** We assume the following conditions holds:

- \( \phi_i \) is \( L \)-smooth,
  \[ \| \nabla \phi(x) - \nabla \phi_i(y) \| \leq L \| x - y \|. \]  
  \( (5) \)

- \( \phi_i \) is non-convex, and the average of \( \phi_i \) is \( \gamma \)-strongly convex,
  \[ \frac{1}{n} \sum_{i=1}^{n} \phi_i(x) \geq \frac{1}{n} \sum_{i=1}^{n} \phi_i(y) + \frac{\gamma}{2} \| x - y \|^2. \]  
  \( (6) \)
• Time delay $\tau$ is no larger than $\Delta$.

Following the above assumptions, we know that our method is able to have linear convergence rate with the following theorem.

**Theorem 2** When the above assumptions satisfy, and let $w^*$ be the minimizer of $P(w)$, $\alpha_i^* = -\nabla \phi_i(w^*)$. If $\eta$ satisfies that:

$$
\left( \Delta^2 + \frac{2\gamma \Delta}{\gamma - \lambda} \right) \eta^2 + \left( \frac{2\Delta + 1}{\gamma - \lambda} + \frac{n}{2L^2} \right) \eta - \frac{1}{2L^2} \leq 0,
$$

(7)

then we have the final conclusion:

$$
\mathbb{E} \left[ \frac{1}{\gamma - \lambda} \|w^{s,0} - w^*\|^2 + \frac{1}{2L^2} \eta^2 + \frac{n}{2L^2} \sum_{i=1}^{n} \|\alpha_i^{s,0} - \alpha_i^*\|^2 \right] \\
\leq e^{-\eta \lambda s} \left[ \frac{1}{\gamma - \lambda} \|w^{0,0} - w^*\|^2 + \frac{1}{2L^2} \eta^2 + \frac{n}{2L^2} \sum_{i=1}^{n} \|\alpha_i^{0,0} - \alpha_i^*\|^2 \right].
$$

(8)

**References**

[1] Alekh Agarwal and John C Duchi. Distributed delayed stochastic optimization. In *Advances in Neural Information Processing Systems*, pages 873–881, 2011.

[2] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[3] Mingyi Hong. A distributed, asynchronous and incremental algorithm for non-convex optimization: An admm based approach. *arXiv preprint arXiv:1412.6058*, 2014.

[4] Zhouyuan Huo and Heng Huang. Asynchronous stochastic gradient descent with variance reduction for non-convex optimization. *arXiv preprint arXiv:1604.03584*, 2016.

[5] Martin Jaggi, Virginia Smith, Martin Takáč, Jonathan Terhorst, Sanjay Krishnan, Thomas Hofmann, and Michael I Jordan. Communication-efficient distributed dual coordinate ascent. In *Advances in Neural Information Processing Systems*, pages 3068–3076, 2014.

[6] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[7] John Langford, Alexander Smola, and Martin Zinkevich. Slow learners are fast. *arXiv preprint arXiv:0911.0491*, 2009.

[8] Mu Li, David G Andersen, Alex J Smola, and Kai Yu. Communication efficient distributed machine learning with the parameter server. In *Advances in Neural Information Processing Systems*, pages 19–27, 2014.
[9] Xiangru Lian, Yijun Huang, Yuncheng Li, and Ji Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. In Advances in Neural Information Processing Systems, pages 2719–2727, 2015.

[10] Ji Liu, Stephen J Wright, and Srikrishna Sridhar. An asynchronous parallel randomized kaczmarz algorithm. arXiv preprint arXiv:1401.4780, 2014.

[11] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In Advances in Neural Information Processing Systems, pages 693–701, 2011.

[12] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. arXiv preprint arXiv:1309.2388, 2013.

[13] Shai Shalev-Shwartz. Sdca without duality. arXiv preprint arXiv:1502.06177, 2015.

[14] Shai Shalev-Shwartz and Tong Zhang. Accelerated mini-batch stochastic dual coordinate ascent. In Advances in Neural Information Processing Systems, pages 378–385, 2013.

[15] 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.

[16] Martin Takáč, Avleen Bijral, Peter Richtárik, and Nathan Srebro. Mini-batch primal and dual methods for svms. arXiv preprint arXiv:1303.2314, 2013.

[17] Martin Takáč, Peter Richtárik, and Nathan Srebro. Distributed mini-batch sdca. arXiv preprint arXiv:1507.08322, 2015.

[18] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. SIAM Journal on Optimization, 24(4):2057–2075, 2014.

[19] Tianbao Yang. Trading computation for communication: Distributed stochastic dual coordinate ascent. In Advances in Neural Information Processing Systems, pages 629–637, 2013.

[20] Ruiliang Zhang and James Kwok. Asynchronous distributed admm for consensus optimization. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), pages 1701–1709, 2014.

[21] Ruiliang Zhang, Shuai Zheng, and James T Kwok. Fast distributed asynchronous sgd with variance reduction. arXiv preprint arXiv:1508.01633, 2015.

[22] Shen-Yi Zhao and Wu-Jun Li. Fast asynchronous parallel stochastic gradient descent: A lock-free approach with convergence guarantee. 2016.
A Proof of Theorem \[\text{\textbf{1}}\]

Lemma 3 Assuming that each $\phi_i$ is $L$-smooth and convex, for every $w$, we have:

$$
\frac{1}{n} \sum_{i=1}^{n} \|\nabla \phi_i (w) - \nabla \phi_i (w^*)\|^2 \leq 2L \left( P(w) - P(w^*) - \frac{\lambda}{2} \|w - w^*\|^2 \right) \tag{9}
$$

This lemma was proved in [13].

Proof 1 (Proof of Theorem \[\text{\textbf{1}}\]). As per Lemma 3 we know that:

$$
\mathbb{E}[\|\nabla \phi_i (w^{t-\tau}) + \alpha_i^*\|^2] \leq \mathbb{E}[\|\nabla \phi_i (w^{t-\tau}) - \nabla \phi_i (w^{t-1}) + \nabla \phi_i (w^{t-1}) - \nabla \phi_i (w^*)\|^2]
\leq 2\mathbb{E}[\|\nabla \phi_i (w^{t-\tau}) - \nabla \phi_i (w^{t-1})\|^2] + 2\mathbb{E}[\|\nabla \phi_i (w^{t-1}) - \nabla \phi_i (w^*)\|^2]
\leq 4L \left( P(w^{t-\tau}) - P(w^{t-1}) - \frac{\lambda}{2} \mathbb{E}[\|w^{t-\tau} - w^{t-1}\|^2] \right)
+ 4L \left( P(w^{t-1}) - P(w^*) - \frac{\lambda}{2} \mathbb{E}[\|w^{t-1} - w^*\|^2] \right)
\leq 4L(P(w^{t-\tau}) - P(w^*)) - 2\lambda L \mathbb{E}[\|w^{t-1} - w^*\|^2] \tag{10}
$$

$$
\mathbb{E}[(w^{t-1} - w^*)^T v_i^t] = \mathbb{E}[(w^{t-1} - w^{t-\tau} + w^{t-\tau} - w^*)^T v_i^t]
= \eta \sum_{j=t-\tau+1}^{t-1} \mathbb{E}[(v_{j, i}^t)^T v_i^t] + (w^{t-\tau} - w^*)^T \nabla P(w^{t-\tau})
\geq \eta \sum_{j=t-\tau+1}^{t-1} \mathbb{E}[(v_{j, i}^t)^T v_i^t] + P(w^{t-\tau}) - P(w^*) \tag{11}
$$

where the final inequality follows from convexity of $P(w)$.

Let’s define $C_t = c_a A_t + c_b B_t$, and take expectation over $i$:

$$
\mathbb{E}[C_t] = \mathbb{E} \left[ c_a (1 - \eta) A_{t-1} + c_a \eta \lambda \|\nabla \phi_i (w^{t-\tau}) + \alpha_i^*\|^2 - c_a \eta \lambda (1 - \beta) \|v_i^t\|^2 \right]
+ c_b B_{t-1} - 2c_b \eta (w^{t-1} - w^*)^T v_i^t + c_b \eta^2 \|v_i^t\|^2
\leq \mathbb{E} \left[ c_a (1 - \eta \lambda) A_{t-1} + c_a \eta \lambda \left( 4L(P(w^{t-\tau}) - P(w^*)) - 2\lambda L \mathbb{E}[\|w^{t-1} - w^*\|^2] \right)
- c_a \eta \lambda (1 - \beta) \|v_i^t\|^2 + c_b B_{t-1} + c_b \eta^2 \|v_i^t\|^2
- 2c_b \eta \left( \eta \sum_{j=t-\tau+1}^{t-1} \mathbb{E}[(v_{j, i}^t)^T v_i^t] + P(w^{t-\tau}) - P(w^*) \right) \right]
\leq c_a (1 - \eta \lambda) \mathbb{E}[A_{t-1}] + (c_b - 2c_a \eta \lambda) \mathbb{E}[B_{t-1}]
+ (c_b \eta^2 - c_a \eta \lambda (1 - \beta) + c_b \Delta \eta^2) \mathbb{E}[\|v_i^t\|^2]
+ (4c_a \eta \lambda - 2c_b \eta) \left( P(w^{t-\tau}) - P(w^*) \right) + c_b \eta^2 \sum_{j=t-\tau+1}^{t-1} \mathbb{E}[\|v_{j, i}^t\|^2] \tag{12}
$$
Summing over $E[C_t]$, we have:

$$
\sum_{t=1}^{n} E[C_t] \leq \sum_{t=1}^{n} \left( c_a (1 - \eta \lambda) E[A_{t-1}] + (c_b - 2c_a \eta L \lambda^2) E[B_{t-1}] \right) \\
+ \sum_{t=1}^{n} (4c_a \eta \lambda L - 2c_b \eta) \left( P(w^{t-1}) - P(w^*) \right) \\
+ \sum_{t=1}^{n} (c_b \eta^2 - c_a \eta \lambda (1 - \beta) + 2c_b \Delta \eta^2) E[\|v_t^i\|^2]
$$

(13)

We denote

$$
c_b - 2c_a \eta L \lambda^2 = c_b (1 - \eta \lambda)
$$

(14)

$$
c_b \eta^2 - c_a \eta \lambda (1 - \beta) + 2c_b \Delta \eta^2 \leq 0.
$$

(15)

Therefore, if $c_b = 2c_a \lambda L$ and $\eta \leq \frac{1}{2L_n + \eta L + 4L \Delta}$, we have:

$$
\sum_{t=1}^{n} E[C_t] \leq (1 - \eta \lambda) \sum_{t=1}^{n} E[C_t - 1] \\
\leq \sum_{t=2}^{n} E[C_{t-1}] + (1 - \eta \lambda) E[C_0]
$$

(16)

Thus

$$
E[C_n] \leq (1 - \eta \lambda) E[C_0]
$$

(17)

Because $E[C_n] = E[C_{s+1,0}]$ and $E[C_0] = E[C_{s,0}]$, we have:

$$
E[C_{s,0}] \leq (1 - \eta \lambda) E[C_{s-1,0}] \\
\leq (1 - \eta \lambda)^s C_{0,0} \\
\leq e^{-\eta \lambda s} C_{0,0}
$$

(18)

Let $c_a = \frac{1}{2L \lambda}$ and $c_b = 1$, then we have the final conclusion:

$$
E \left[ \|w^{s,0} - w^*\|^2 + \frac{1}{2L} \sum_{i=1}^{n} \|\alpha_i^{s,0} - \alpha_i^*\|^2 \right] \\
\leq e^{-\eta \lambda s} \left[ \|w^{0,0} - w^*\|^2 + \frac{1}{2L} \sum_{i=1}^{n} \|\alpha_i^{0,0} - \alpha_i^*\|^2 \right]
$$

(19)

**B Proof of Theorem 2**

Proof 2 (Proof of Theorem 2) Let $w^*$ be the minimizer of $P(w)$ and let $\alpha_i^* = -\nabla \phi_i(w^*)$.

$$
u_i^{s,t} = \nabla \phi_i(w_i^{s,t-1}) + \alpha_i^{s,t-1}
$$

(20)
\[ v_i^{s,t} = \nabla \phi_s(w_i^{s,t-\tau}) + \alpha_i^{s,t-\tau} \] (21)

Because \( \alpha_i \) will be not updated in the process from \( t - \tau \) to \( t - 1 \), so \( \alpha_i^{s,t-\tau} = \alpha_i^{s,t-1} \).

In an epoch \( s \), we use \( w^t \) to denote \( w_i^{s,t} \), \( \alpha_i^t \) to denote \( \alpha_i^{s,t} \), \( A_t, B_t, C_t \) to denote \( A_{s,t}, B_{s,t}, C_{s,t} \).

\[ A_t = \frac{1}{n} \sum_{i=1}^{n} ||\alpha_i^t - \alpha_i^*||^2 \] (22)

\[ B_t = ||w^t - w^*||^2 \] (23)

Let \( \beta = \eta \lambda n \), so in iteration \( t \), \( \alpha_i^t = (1 - \beta)\alpha_i^{t-1} + \beta(-\nabla \phi_i(w_i^{t-\tau})) \).

\[ A_t - A_{t-1} = \frac{1}{n} ||\alpha_i^t - \alpha_i^*||^2 - \frac{1}{n} ||\alpha_i^{t-1} - \alpha_i^*||^2 \]
\[ = \frac{1}{n} ||(1 - \beta)(\alpha_i^{t-1} - \alpha_i^*) + \beta(-\nabla \phi_i(w_i^{t-\tau}) - \alpha_i^*)||^2 - \frac{1}{n} ||\alpha_i^{t-1} - \alpha_i^*||^2 \]
\[ = \frac{1}{n} \left( (1 - \beta)||\alpha_i^{t-1} - \alpha_i^*||^2 + \beta||-\nabla \phi_i(w_i^{t-\tau}) - \alpha_i^*||^2 \right. \]
\[ - \beta(1 - \beta)||\alpha_i^{t-1} - \nabla \phi_i(w_i^{t-\tau})||^2 - ||\alpha_i^{t-1} - \alpha_i^*||^2 \right) \]
\[ = \frac{\beta}{n} \left( -||\alpha_i^{t-1} - \alpha_i^*||^2 + ||\nabla \phi_i(w_i^{t-\tau}) + \alpha_i^*||^2 - (1 - \beta)||v_i^t||^2 \right) \]
\[ = \eta \lambda \left( -||\alpha_i^{t-1} - \alpha_i^*||^2 + ||\nabla \phi_i(w_i^{t-\tau}) + \alpha_i^*||^2 - (1 - \beta)||v_i^t||^2 \right) \] (24)

In addition,

\[ B_t - B_{t-1} = ||w^t - w^*||^2 - ||w_i^{t-1} - w^*||^2 \]
\[ = ||w_i^{t-1} - \eta v_i^t - w^*||^2 - ||w_i^{t-1} - w^*||^2 \]
\[ = -2\eta(w_i^{t-1} - w^*)^T v_i^t + \eta^2||v_i^t||^2 \] (25)

\[ ||\nabla \phi_i(w_i^{t-\tau}) + \alpha_i^*||^2 = ||\nabla \phi_i(w_i^{t-\tau}) - \nabla \phi_i(w_i^{t-1}) + \nabla \phi_i(w_i^{t-1}) - \nabla \phi_i(w^*)||^2 \]
\[ \leq 2||\nabla \phi_i(w_i^{t-\tau}) - \nabla \phi_i(w_i^{t-1})||^2 + 2||\nabla \phi_i(w_i^{t-1}) - \nabla \phi_i(w^*)||^2 \]
\[ \leq 2L^2||w_i^{t-\tau} - w_i^{t-1}||^2 + 2L^2||w_i^{t-1} - w^*||^2 \]
\[ \leq 2L^2\Delta \sum_{j=t-\tau+1}^{t-1} ||w_i^j - w_i^{j-1}||^2 + 2L^2||w_i^{t-1} - w^*||^2 \]
\[ \leq 2\Delta \eta^2 L^2 \sum_{j=t-\tau+1}^{t-1} ||v_i^j||^2 + 2L^2||w_i^{t-1} - w^*||^2 \] (26)
where the first and fourth inequality follows from that $\| \sum_{i=1}^{n} a_i \|^2 \leq n \sum_{i=1}^{n} \| a_i \|^2$. The second inequality follows that $\phi_i$ is $L$-smooth. $\Delta$ is the upper bound of time delay.

$$
(w^{t-1} - w^*)^T v^t_i = (w^{t-1} - w^{t-\tau} + w^{t-\tau} - w^*)^T v^t_i \\
= \underbrace{\left( w^{t-1} - w^{t-\tau} \right)^T v^t_i + (w^{t-\tau} - w^*)^T v^t_i}_{T_1}
$$

(27)

Because $E[v^t_i] = \nabla P(w^{t-\tau})$, we have:

$$
E[T_2] = E[(w^{t-\tau} - w^*)^T v^t_i] \\
= (w^{t-\tau} - w^*)^T \nabla P(w^{t-\tau}) \\
\geq \gamma \| w^{t-\tau} - w^* \|^2 \\
\geq \gamma \frac{1}{2} \| w^{t-1} - w^* \|^2 - \| w^{t-\tau} - w^{t-1} \|^2 \\
\geq \frac{\gamma}{2} \| w^{t-1} - w^* \|^2 - \Delta \gamma \eta^2 \sum_{j=t-\tau+1}^{t-1} \| v^j_{ij} \|^2
$$

(28)

where the first inequality follows from the strong convexity of $P(w)$.

$$
(w^* - w^{t-\tau})^T \nabla P(w^{t-\tau}) \geq P(w^{t-\tau}) - P(w^*) + \frac{\gamma}{2} \| w^{t-\tau} - w^* \|^2
$$

(29)

$$
P(w^{t-\tau}) - P(w^*) \geq \frac{\gamma}{2} \| w^{t-\tau} - w^* \|^2
$$

(30)

The second inequality follows from the inequality:

$$
\| w^{t-1} - w^* \|^2 = \| w^{t-1} - w^{t-\tau} + w^{t-\tau} - w^* \|^2 \\
\leq 2 \| w^{t-1} - w^{t-\tau} \|^2 + 2 \| w^{t-\tau} - w^* \|^2
$$

(31)

$$
T_1 = (w^{t-1} - w^{t-\tau})^T v^t_i \\
= \left( \sum_{j=t-\tau+1}^{t-1} (w^j - w^{j-1}) \right)^T v^t_i \\
= \eta \sum_{j=t-\tau+1}^{t-1} (v^j_{ij})^T v^t_i
$$

(32)
We define $C_t = c_a A_t + c_b B_t$, and take expectation over $i$,

$$
E[C_t] = E \left[ c_a (A_{t-1} + \eta \lambda (-\|\alpha_{t-1}^i - \alpha_t^i\|^2 + \|\nabla \phi_i(w^{t-\tau}) + \alpha_t^i\|^2 - (1 - \beta)\|v_t^i\|^2)) \right. \\
+ c_b (B_{t-1} - \eta (w^{t-1} - w^*)^T v_t^i + \eta^2 \|v_t^i\|^2) \\
= E \left[ c_a (1 - \eta \lambda) A_{t-1} + c_a \eta \lambda \|\nabla \phi_i(w^{t-\tau}) + \alpha_t^i\|^2 - c_a \eta \lambda (1 - \beta)\|v_t^i\|^2 \right. \\
+ c_b B_{t-1} - 2c_b \eta (w^{t-1} - w^*)^T v_t^i + c_b \eta^2 \|v_t^i\|^2 \\
\leq E \left[ c_a (1 - \eta \lambda) A_{t-1} + c_a \eta \lambda \left( 2\Delta \eta^2 L^2 \sum_{j=t-\tau+1}^{t-1} \|v_j^i\|^2 + 2L^2 \|w^{t-1} - w^*\|^2 \right) \right. \\
- c_a \eta \lambda (1 - \beta)\|v_t^i\|^2 + c_b B_{t-1} + c_b \eta^2 \|v_t^i\|^2 \\
- 2c_b \eta \left( \frac{\gamma^2}{2} \|w^{t-1} - w^*\|^2 - \Delta \gamma \eta^2 \sum_{j=t-\tau+1}^{t-1} \|v_j^i\|^2 + \eta \sum_{j=t-\tau+1}^{t-1} (v_j^i)^T v_j^i \right) \\
\leq c_a (1 - \eta \lambda) E[A_{t-1}] + (2c_a \eta \lambda L^2 + c_b - c_b \eta \gamma) E[B_{t-1}] \\
+ (-c_a \eta \lambda (1 - \beta) + c_b \eta^2 + c_b \Delta \eta^2) E[\|v_t^i\|^2] \\
\left. + (2c_a \Delta \lambda L^2 \eta^3 + 2c_b \Delta \gamma \eta^3 + c_b \eta^2) \sum_{j=t-\tau+1}^{t-1} E[\|v_j^i\|^2] \right] \\
(33)
$$

where the first equality follows from $E[(\|\alpha_{t-1}^i - \alpha_t^i\|^2)] = A_{t-1}$.

Summing over $E[C_t]$ from $t = 1$ to $n$, we obtain:

$$
\sum_{t=1}^{n} E[C_t] \leq \sum_{t=1}^{n} \left( c_a (1 - \eta \lambda) E[A_{t-1}] + (2c_a \eta \lambda L^2 + c_b - c_b \eta \gamma) E[B_{t-1}] \right) \\
+ \sum_{t=1}^{n} (2c_a \Delta \lambda L^2 \eta^3 + 2c_b \Delta \gamma \eta^3 + 2c_b \Delta \eta^2 + c_b \eta^2 - c_a \eta \lambda (1 - \beta)) E[\|v_t^i\|^2] \\
(34)
$$

If two following inequalities hold

$$
2c_a \eta \lambda L^2 + c_b - c_b \eta \gamma = c_b (1 - \eta \lambda) \quad (35)
$$

$$
2c_a \Delta \lambda L^2 \eta^3 + 2c_b \Delta \gamma \eta^3 + 2c_b \Delta \eta + c_b \eta - c_a \lambda (1 - n \eta) \leq 0 \quad (36)
$$

then

$$
\sum_{t=1}^{n} E[C_t] \leq (1 - \eta \lambda) \sum_{t=1}^{n} E[C_{t-1}] \\
\leq \sum_{t=2}^{n} E[C_{t-1}] + (1 - \eta \lambda) E[C_0] \quad (37)
$$
Because $E[C_n] = E[C_{s+1,0}]$ and $E[C_0] = E[C_{s,0}]$, thus,

$$E[C_{s,0}] \leq (1 - \eta \lambda)E[C_{s-1,0}] \leq (1 - \eta \lambda)^s C_{0,0} \leq e^{-\eta \lambda s} C_{0,0} \quad (38)$$

We set $c_a = \frac{1}{2\lambda L^2}$ and $c_b = \frac{1}{\gamma - \lambda}$. If $\eta = 0$, the left-side value in (36) is $-c_a \lambda$, and it is smaller than 0. Thus there exists a value $\eta > 0$ to make the inequality (36) hold. Thus, we have the final conclusion. If $\eta$ satisfies the following inequality:

$$\left( \Delta^2 + \frac{2\gamma \Delta^2}{\gamma - \lambda} \right) \eta^2 + \left( \frac{2\Delta + 1}{\gamma - \lambda} + \frac{n}{2L^2} \right) \eta - \frac{1}{2L^2} \leq 0, \quad (39)$$

then

$$E \left[ \frac{1}{\gamma - \lambda} \| w^{s,0} - w^* \|^2 + \frac{1}{2\lambda L^2 n} \sum_{i=1}^n \| \alpha^{s,0}_i - \alpha^*_i \|^2 \right] \leq e^{-\eta \lambda s} \left[ \frac{1}{\gamma - \lambda} \| w^{0,0} - w^* \|^2 + \frac{1}{2\lambda L^2 n} \sum_{i=1}^n \| \alpha^{0,0}_i - \alpha^*_i \|^2 \right]. \quad (40)$$