Catalog

Question 1

Answer:

To complete the convergence proof and show that there exists a learning rate γ such that

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla L(\theta_t)\|^2 \in O\left(\frac{\zeta^2}{T^{2/3}} + \frac{\rho^2\lambda}{T}\right),$$

Proof steps:

1. Derive the Descent Lemma for the Average Model:

We start by analyzing the average model $\theta_t = \frac{1}{n}\sum_{i=1}^{n}\theta_i^t$. The update rule for the local models is given by:

$$\theta_i^{t+1} = \theta_i^t - \gamma\nabla L_i(\theta_i^t) + \alpha \sum_{j\in N(i)} (\theta_j^t - \theta_i^t),$$

where α is the gossip rate, and $N(i)$ denotes the neighbors of node i in the graph G.

Averaging over all nodes, we get:

$$\theta_{t+1} = \theta_t - \gamma\frac{1}{n}\sum_{i=1}^{n} \nabla L_i(\theta_i^t),$$

since the gossip terms cancel out due to symmetry.

2. Express the Update in Terms of Gradient Differences:

Define $\delta_t = \frac{1}{n}\sum_{i=1}^{n}[\nabla L_i(\theta_i^t) - \nabla L_i(\theta_t)]$. Using this, we can write:

$$\theta_{t+1} = \theta_t - \gamma\nabla L(\theta_t) - \gamma\delta_t,$$

where $\nabla L(\theta_t) = \frac{1}{n}\sum_{i=1}^{n} \nabla L_i(\theta_t)$.

3. Apply the Smoothness of the Loss Function:

Using the Lipschitz smoothness of L, we have:

$$L(\theta_{t+1}) \leq L(\theta_t) + \langle \nabla L(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{\lambda}{2}\|\theta_{t+1} - \theta_t\|^2.$$

Substituting $\theta_{t+1} - \theta_t = -\gamma \nabla L(\theta_t) - \gamma \delta_t$, we obtain:

$$L(\theta_{t+1}) \leq L(\theta_t) - \gamma\|\nabla L(\theta_t)\|^2 - \gamma\langle \nabla L(\theta_t), \delta_t \rangle + \frac{\lambda\gamma^2}{2}\|\nabla L(\theta_t) + \delta_t\|^2.$$

4. Simplify and Bound the Error Terms:

Expanding the squared norm and rearranging terms, we get:

$$L(\theta_{t+1}) \leq L(\theta_t) - \left(\gamma - \frac{\lambda\gamma^2}{2}\right)\|\nabla L(\theta_t)\|^2 + (-\gamma + \lambda\gamma^2)\langle \nabla L(\theta_t), \delta_t \rangle + \frac{\lambda\gamma^2}{2}\|\delta_t\|^2.$$

Assuming γ is small enough, we approximate $\gamma - \frac{\lambda\gamma^2}{2} \approx \gamma$ and $-\gamma + \lambda\gamma^2 \approx -\gamma$, simplifying the inequality to:

$$L(\theta_{t+1}) \leq L(\theta_t) - \gamma\|\nabla L(\theta_t)\|^2 - \gamma\langle \nabla L(\theta_t), \delta_t \rangle + \frac{\lambda\gamma^2}{2}\|\delta_t\|^2.$$

5. Bound the Inner Product and Norm of δ_t:

Using the Cauchy-Schwarz inequality and the Lipschitz continuity of ∇L_i, we have:

$$\|\delta_t\| \leq \frac{\lambda}{n}\sum_{i=1}^{n} \|\theta_i^t - \theta_t\| = \lambda E_t,$$

where $E_t^2 = \frac{1}{n}\sum_{i=1}^{n}\|\theta_i^t - \theta_t\|^2$ represents the consensus error.

The inner product is bounded by:

$$|\langle \nabla L(\theta_t), \delta_t \rangle| \leq \|\nabla L(\theta_t)\|\|\delta_t\| \leq \|\nabla L(\theta_t)\|\lambda E_t.$$

6. Obtain the Descent Inequality:

Incorporating the bounds, the inequality becomes:

$$L(\theta_{t+1}) \leq L(\theta_t) - \gamma\|\nabla L(\theta_t)\|^2 + \gamma\|\nabla L(\theta_t)\|\lambda E_t + \frac{\lambda^3\gamma^2}{2}E_t^2.$$

7. Simplify the Error Terms Using Inequalities:

By applying the inequality $ab \leq \frac{a^2+b^2}{2}$ to the inner product term, we get:

$$\gamma \|\nabla L(\theta_t)\| \lambda E_t \leq \frac{\gamma}{2}(\|\nabla L(\theta_t)\|^2 + \lambda^2 E_t^2).$$

The inequality now becomes:

$$L(\theta_{t+1}) \leq L(\theta_t) - \frac{\gamma}{2}\|\nabla L(\theta_t)\|^2 + \left(\frac{\gamma\lambda^2}{2} + \frac{\lambda^3\gamma^2}{2}\right)E_t^2.$$

8. Establish a Recursion for the Consensus Error E_t^2:

Analyzing the update of the consensus error, we find:

$$E_{t+1}^2 \leq \rho^2 E_t^2 + 2\gamma^2\zeta^2,$$

where $\rho = 1 - \alpha\mu_G$ and μ_G is the algebraic connectivity of the graph G.

Solving this recursion yields:

$$E_t^2 \leq \frac{2\gamma^2\zeta^2}{1-\rho^2} = \frac{2\gamma^2\zeta^2}{2\alpha\mu_G} = \frac{\gamma^2\zeta^2}{\alpha\mu_G}.$$

9. Combine the Results to Bound the Gradient Norm:

Summing the descent inequality over $t = 1$ to T and rearranging terms, we get:

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L(\theta_t)\|^2 \leq \frac{2[L(\theta_0) - L^*]}{\gamma T} + \left(\frac{\gamma\lambda^2}{2} + \frac{\lambda^3\gamma^2}{2}\right)\frac{E_t^2}{\gamma}.$$

Substituting the bound on E_t^2, we have:

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L(\theta_t)\|^2 \leq \frac{2[L(\theta_0) - L^*]}{\gamma T} + \left(\frac{\gamma\lambda^2}{2} + \frac{\lambda^3\gamma^2}{2}\right)\frac{\gamma\zeta^2}{\alpha\mu_G}.$$

10. Choose γ to Balance the Terms:

To balance the terms, set $\gamma = cT^{-1/3}$ for some constant $c > 0$. This choice yields:

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L(\theta_t)\|^2 \leq O\left(\frac{1}{T^{2/3}}\right) + O\left(\frac{\gamma^2\lambda^2\zeta^2}{\alpha\mu_G}\right).$$

Since $\gamma^2 = O(T^{-2/3})$, the second term becomes $O\left(\frac{T^{-2/3}\lambda^2\zeta^2}{\alpha\mu_G}\right)$.

11. Express the Bound in Terms of ρ and Simplify:

Recognizing that $\alpha\mu_G = 1 - \rho$, we rewrite the bound:

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla L(\theta_t)\|^2 \leq O\left(\frac{\zeta^2}{T^{2/3}}\right) + O\left(\frac{(1-\rho)^2\lambda}{T}\right).$$

12. Conclude the Proof:

Thus, have shown that:

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla L(\theta_t)\|^2 \in O\left(\frac{\zeta^2}{T^{2/3}} + \frac{\rho^2\lambda}{T}\right),$$

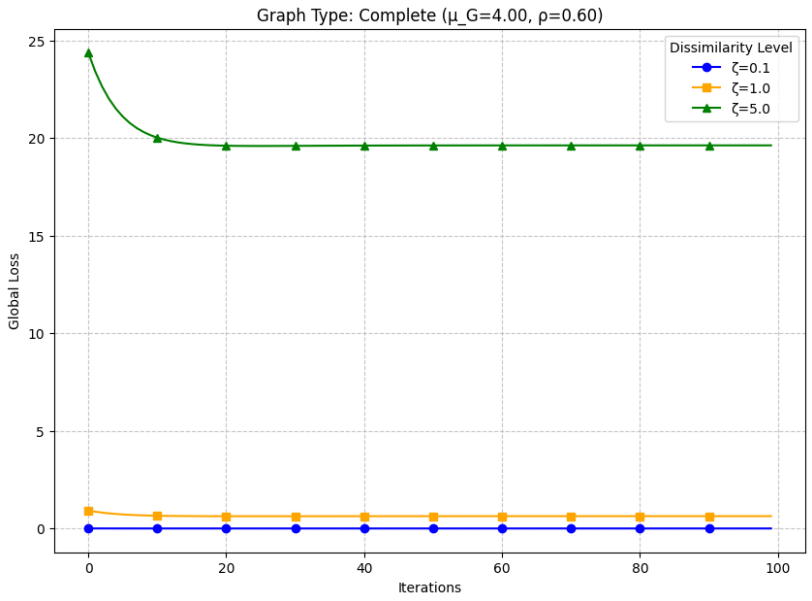
completing the convergence proof.

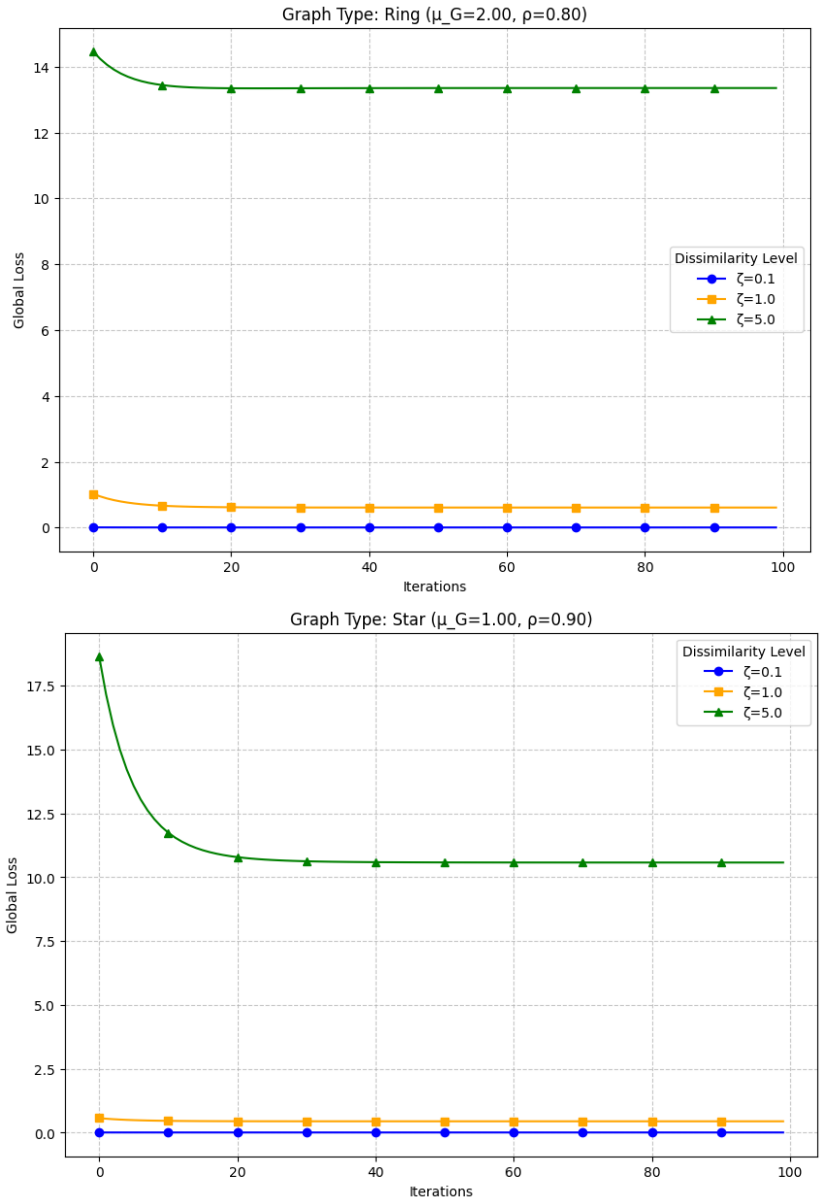
Question 2

Answer:

Check out the link please: Empirically check the dependence of the error in (1) on (a) and (b)

Results:





Analysis:

(a) algebraic connectivity (μ_G)

Algebraic connectivity is a measure of how well connected a graph is. Higher algebraic connectivity means that the nodes in the network are better connected, leading to faster and more robust convergence in distributed optimization algorithms. In my case:

- **Complete Graph**: High algebraic connectivity ($\mu_G = 4.0$) leads to rapid convergence with a smaller error, especially when gradient dissimilarity (ζ) is low. This suggests that the nodes in a highly connected network can more effectively share information, allowing local models to converge more quickly.
- **Ring Graph**: Moderate connectivity ($\mu_G = 2.0$) results in a slightly slower convergence, with higher error compared to the complete graph, especially as ζ

increases.

- **Star Graph**: Low connectivity ($\mu_G = 1.0$) causes the slowest convergence and the highest error. This is because nodes on the periphery rely on a single central node for communication, slowing down the consensus-building process.

(b) gradient dissimilarity (ζ)

Gradient dissimilarity represents the difference between the local objectives of each client. Higher dissimilarity implies that each node's local optimum is farther from the others, which makes it harder for the algorithm to find a consensus. Here's how different values of ζ affect the error:

- **Low** $\zeta = 0.1$: With low gradient dissimilarity, the local objectives are relatively close to each other, which allows the algorithm to converge quickly across all graph types. The error remains low, as seen in the Complete, Ring, and Star graphs.
- **Medium** $\zeta = 1.0$: With moderate dissimilarity, the convergence is slower, and the final error is higher across all graphs. The increased gradient differences make it harder for the models to agree on a common solution.
- **High** $\zeta = 5.0$: High gradient dissimilarity leads to poor convergence, with the error remaining significantly higher, particularly in the less connected graphs (Ring and Star). This indicates that when local objectives are highly divergent, the network struggles to reach consensus, especially in topologies with low algebraic connectivity.

Summary of Empirical Observations

- **Error and Algebraic Connectivity**: Higher algebraic connectivity (μ_G) generally reduces error by facilitating faster information sharing among nodes. Complete graphs consistently outperform other topologies in convergence speed and error minimization.
- **Error and Gradient Dissimilarity**: Higher gradient dissimilarity (ζ) increases error and slows convergence, as it becomes harder for the nodes to align on a global objective. This effect is amplified in networks with lower connectivity.

The empirical results indicate that for robust convergence in Distributed Gradient Descent, both high algebraic connectivity and low gradient dissimilarity are beneficial.

Question 3

Answer:

Part 1

To find a sufficient condition on the weights w_{ij} such that the consensus error decreases, need to analyze how the disagreement measure Γ evolves with the update rule:

$$\theta_{t+1/2}^{(i)} = \sum_{j=1}^{n} w_{ij}\theta_t^{(j)}.$$

Definitions:

- **Average Model:**

$$\theta_t = \frac{1}{n}\sum_{i=1}^{n} \theta_t^{(i)}.$$

- **Consensus Error:**

$$\Gamma(\theta_t^{(1)}, \dots, \theta_t^{(n)}) = \frac{1}{n}\sum_{i=1}^{n} \left\|\theta_t^{(i)} - \theta_t\right\|^2.$$

Analysis:

1. **Express the Deviations from the Average:**

Define $\delta_t^{(i)} = \theta_t^{(i)} - \theta_t$.

2. **Update of Deviations:**

After the gossip step:

$$\delta_{t+1/2}^{(i)} = \theta_{t+1/2}^{(i)} - \theta_{t+1/2}.$$

To proceed, need to determine $\theta_{t+1/2}$.

3. **Compute the New Average $\theta_{t+1/2}$:**

$$\theta_{t+1/2} = \frac{1}{n}\sum_{i=1}^{n} \theta_{t+1/2}^{(i)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\theta_t^{(j)} = \sum_{j=1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n} w_{ij}\right)\theta_t^{(j)}.$$

4. **Assuming $\sum_{i=1}^{n} w_{ij} = 1$ (Symmetric and Row-Stochastic):**

To ensure $\theta_{t+1/2} = \theta_t$, we require that:

$$\sum_{i=1}^{n} w_{ij} = 1 \text{ for all } j.$$

Since $w_{ij} = w_{ji}$, this implies W is doubly stochastic (rows and columns sum to 1).

5. **Express Updated Deviations:**

$$\delta_{t+1/2}^{(i)} = \sum_{j=1}^{n} w_{ij}\theta_t^{(j)} - \theta_{t+1/2} = \sum_{j=1}^{n} w_{ij}\delta_t^{(j)}.$$

6. Compute the New Consensus Error:

$$\Gamma(\theta_{t+1/2}^{(1)}, \dots, \theta_{t+1/2}^{(n)}) = \frac{1}{n}\sum_{i=1}^{n} \left\| \delta_{t+1/2}^{(i)} \right\|^2.$$

7. Express in Matrix Form:

Let $\delta_t = [\delta_t^{(1)}; \dots; \delta_t^{(n)}]$ and W be the weight matrix. Then:

$$\delta_{t+1/2} = W\delta_t.$$

Therefore:

$$\Gamma_{t+1/2} = \frac{1}{n}\delta_t^{\top} W^{\top} W \delta_t.$$

8. Eigenvalue Analysis:

Since W is symmetric and doubly stochastic, its eigenvalues λ_i satisfy:

$\lambda_1 = 1$ (associated with the eigenvector $\mathbf{1}$)

All other eigenvalues λ_i satisfy $|\lambda_i| \leq 1$.

The consensus error can be bounded as:

$$\Gamma_{t+1/2} \leq \lambda_2^2 \Gamma_t,$$

where λ_2 is the second-largest eigenvalue of W in magnitude.

Sufficient Condition:

A sufficient condition is that the second-largest eigenvalue λ_2 satisfies:

$$\lambda_2^2 \leq c,$$

where $c \in [0,1)$. This ensures that:

$$\Gamma_{t+1/2} \leq c\Gamma_t.$$

Conclusion:

A sufficient condition is that the second-largest eigenvalue λ_2 of the symmetric weight matrix W satisfies $\lambda_2^2 \leq c \in [0,1)$. This ensures the consensus error decreases:

$$\Gamma(\theta_{t+1/2}^{(1)}, \dots, \theta_{t+1/2}^{(n)}) \leq c \cdot \Gamma(\theta_t^{(1)}, \dots, \theta_t^{(n)}).$$

Part 2

To maintain $\theta_{t+1/2} = \theta_t$ (i.e., the average remains unchanged after the gossip step), require the weight matrix W to be **row stochastic**:

$$\sum_{j=1}^{n} w_{ij} = 1 \text{ for all } i.$$

Explanation:

- **Row Stochasticity** ensures that:

$$\theta_{t+1/2}^{(i)} = \sum_{j=1}^{n} w_{ij}\theta_t^{(j)} \implies \theta_{t+1/2} = \sum_{i=1}^{n}\left(\frac{1}{n}\sum_{j=1}^{n} w_{ij}\theta_t^{(j)}\right) = \theta_t.$$

- **Symmetry** ($w_{ij} = w_{ji}$) combined with row stochasticity implies column stochasticity, making W **doubly stochastic**.

Conclusion:

To ensure $\theta_{t+1/2} = \theta_t$, we require the weights w_{ij} to satisfy:

- $w_{ij} = w_{ji}$ (symmetry)
- $\sum_{j=1}^{n} w_{ij} = 1$ for all i (row stochasticity)

This condition ensures the average model remains unchanged after the gossip step.

Part 3

The convergence result in (1) depends on the rate at which the consensus error Γ decreases. In the original analysis, the rate is determined by $\rho = 1 - \alpha\mu_G$, where μ_G is the algebraic connectivity of the graph G and α is the gossip rate.

With the generalized gossip rule, the convergence rate is governed by the second-largest eigenvalue λ_2 of the weight matrix W. Specifically, the consensus error decreases as:

$$\Gamma_{t+1/2} \leq \lambda_2^2 \Gamma_t.$$

Impact on Convergence Result:

- The term $\rho^2\lambda/T$ in the convergence bound becomes $\lambda_2^2\lambda/T$.
- The convergence rate depends on λ_2^2 instead of ρ^2.
- A smaller λ_2 (i.e., larger spectral gap) leads to faster convergence.

Conclusion:

The convergence result changes by replacing ρ^2 with λ_2^2 in the bound. The new convergence rate becomes:

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla L(\theta_t)\|^2 \in O\left(\frac{\zeta^2}{T^{2/3}} + \frac{\lambda_2^2 \lambda}{T}\right).$$

This shows that the convergence now depends on the square of the second-largest eigenvalue λ_2 of the weight matrix W.

Part 4

To design a convex programming problem that yields the smallest value for c (i.e., minimize λ_2^2) subject to the given constraints, can formulate the following semidefinite program (SDP).

Objective:

Minimize λ (representing λ_2)

Constraints:

1. Symmetry:

$$W = W^\top$$

2. Non-Negativity:

$$w_{ij} \geq 0 \text{ for all } i, j$$

3. Weight Structure:

$$w_{ij} = 0 \text{ if } (i, j) \notin E$$

4. Row Stochasticity:

$$\sum_{j=1}^{n} w_{ij} = 1 \text{ for all } i$$

5. Spectral Constraint:

For all vectors x orthogonal to $\mathbf{1}$ (the all-ones vector):

$$x^\top W x \leq \lambda x^\top x$$

This ensures that the second-largest eigenvalue $\lambda_2 \leq \lambda$.

Semidefinite Representation:

The spectral constraint can be expressed as:

$$W - \lambda\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right) \preccurlyeq 0,$$

where I is the identity matrix and $\mathbf{1}\mathbf{1}^\top/n$ projects onto the space spanned by $\mathbf{1}$.

Complete SDP Formulation:

$$\min_{\lambda, W} \lambda$$

Subject to:

$W = W^\top$

$W\mathbf{1} = \mathbf{1}$

$W \geq 0$ (element-wise)

$w_{ij} = 0$ if $(i,j) \notin E$

$$W - \lambda\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right) \preccurlyeq 0$$

Explanation:

- **Objective:** Minimize the upper bound λ on the second-largest eigenvalue λ_2.
- **Constraints:** Ensure W satisfies the properties of a symmetric, non-negative, row-stochastic matrix with the given sparsity pattern.
- **Spectral Constraint:** The matrix $W - \lambda\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)$ being negative semidefinite ($\preceq 0$) ensures that all eigenvalues of W other than the largest ($\lambda_1 = 1$) are less than or equal to λ.

Convexity:

- The feasible set is convex due to linear equality and inequality constraints.
- The spectral constraint defines a convex set in W and λ.
- The objective function is linear in λ.

Conclusion:

The convex programming problem to minimize c (i.e., λ_2^2) is:

$$\min_{\lambda, W} \lambda$$

Subject to:

$W = W^\top$

$W\mathbf{1} = \mathbf{1}$

$W \geq 0$ (element-wise)

$w_{ij} = 0$ if $(i,j) \notin E$

$$W - \lambda\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\top}\right) \preccurlyeq 0$$

Solving this semidefinite program yields the weight matrix W that minimizes the convergence rate $c = \lambda^2$, ensuring the fastest possible decrease in consensus error under the given constraints.

Question 4

Answer:

To prove the given bound, analyze the convergence of the Distributed Stochastic Gradient Descent (DSGD) method under the assumptions:

- Server-based architecture with synchronization.
- Each client i has stochastic gradients with variance $\sigma_i^2 = \sigma^2$.
- The global loss function L is λ-Lipschitz smooth.

To show that by choosing an appropriate constant learning rate γ, the expected average squared norm of the gradient over T iterations satisfies:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \|\nabla L(\theta_t)\|^2\right] \leq C\sqrt{\frac{\sigma^2}{nT}},$$

for some constant $C > 0$.

1. Preliminaries

1.1. Notations and Assumptions

- **Global Loss Function**: $L(\theta) = \frac{1}{n}\sum_{i=1}^{n} L_i(\theta)$, where L_i is the local loss at client i.
- **Smoothness**: Each L_i is λ-smooth, so L is λ-smooth.
- **Stochastic Gradients**: At iteration t, each client computes a stochastic gradient $\nabla L_i(\theta_t; \xi_t^i)$ such that:$\mathbb{E}_{\xi_t^i}\left[\nabla L_i(\theta_t; \xi_t^i)\right] = \nabla L_i(\theta_t), \mathbb{E}_{\xi_t^i}\left[\left\|\nabla L_i(\theta_t; \xi_t^i) - \nabla L_i(\theta_t)\right\|^2\right] \leq \sigma^2.$

1.2. DSGD Update Rule

At each iteration t, the server aggregates the stochastic gradients from all clients:

1. Clients Compute Stochastic Gradients:

$$g_t^i = \nabla L_i(\theta_t; \xi_t^i).$$

2. Server Aggregates Gradients and Updates Model:

$$g_t = \frac{1}{n}\sum_{i=1}^{n} g_t^i,$$

$$\theta_{t+1} = \theta_t - \gamma g_t.$$

2. Convergence Analysis

2.1. Smoothness Property

Since L is λ-smooth, we have:

$$L(\theta_{t+1}) \leq L(\theta_t) + \langle \nabla L(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{\lambda}{2}\|\theta_{t+1} - \theta_t\|^2.$$

2.2. Substituting the Update Rule

Substitute $\theta_{t+1} - \theta_t = -\gamma g_t$:

$$L(\theta_{t+1}) \leq L(\theta_t) - \gamma\langle \nabla L(\theta_t), g_t \rangle + \frac{\lambda\gamma^2}{2}\|g_t\|^2.$$

2.3. Taking Expectations

Take expectations over the randomness (denoted by $\mathbb{E}_t$) conditioned on θ_t:

1. Expected Gradient Estimate:

$$\mathbb{E}_t[g_t] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_t[g_t^i] = \frac{1}{n}\sum_{i=1}^{n} \nabla L_i(\theta_t) = \nabla L(\theta_t).$$

2. Variance of g_t:

   a) Since each g_t^i has variance σ^2 and is independent across clients:$\mathbb{E}_t[\|g_t - \nabla L(\theta_t)\|^2] = \frac{\sigma^2}{n}$.

   b) Therefore, the second moment:$\mathbb{E}_t[\|g_t\|^2] = \|\nabla L(\theta_t)\|^2 + \frac{\sigma^2}{n}$.

3. Expected Descent Inequality:

$$\mathbb{E}_t[L(\theta_{t+1})] \leq L(\theta_t) - \gamma\|\nabla L(\theta_t)\|^2 + \frac{\lambda\gamma^2}{2}\left(\|\nabla L(\theta_t)\|^2 + \frac{\sigma^2}{n}\right).$$

2.4. Rearranging Terms

Group like terms:

$$\mathbb{E}_t[L(\theta_{t+1})] \leq L(\theta_t) - \left(\gamma - \frac{\lambda\gamma^2}{2}\right)\|\nabla L(\theta_t)\|^2 + \frac{\lambda\gamma^2\sigma^2}{2n}.$$

2.5. Choosing the Learning Rate γ

To ensure that $\gamma \leq \frac{1}{\lambda}$, we can set $\gamma = \frac{1}{\lambda}$, which yields:

$$\gamma - \frac{\lambda\gamma^2}{2} = \gamma - \frac{\lambda(\gamma)^2}{2} = \gamma\left(1 - \frac{\lambda\gamma}{2}\right) = \gamma\left(1 - \frac{1}{2}\right) = \frac{\gamma}{2}.$$

Thus, the inequality simplifies to:

$$\mathbb{E}_t[L(\theta_{t+1})] \leq L(\theta_t) - \frac{\gamma}{2}\|\nabla L(\theta_t)\|^2 + \frac{\lambda\gamma^2\sigma^2}{2n}.$$

2.6. Telescoping Sum over T Iterations

Sum both sides over $t = 1$ to T:

$$\sum_{t=1}^{T} \frac{\gamma}{2}\mathbb{E}[\|\nabla L(\theta_t)\|^2] \leq L(\theta_1) - \mathbb{E}[L(\theta_{T+1})] + \sum_{t=1}^{T} \frac{\lambda\gamma^2\sigma^2}{2n}.$$

Assuming $L(\theta_{T+1}) \geq L^*$ (the minimal loss), we have:

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla L(\theta_t)\|^2] \leq \frac{2}{\gamma}(L(\theta_1) - L^*) + \frac{\lambda\gamma\sigma^2 T}{n}.$$

2.7. Dividing by T and Optimizing γ

Divide both sides by T:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla L(\theta_t)\|^2] \leq \frac{2(L(\theta_1) - L^*)}{\gamma T} + \frac{\lambda\gamma\sigma^2}{n}.$$

To minimize the right-hand side with respect to γ, we set:

$$\frac{2(L(\theta_1) - L^*)}{\gamma T} = \frac{\lambda\gamma\sigma^2}{n},$$

which leads to:

$$\gamma^2 = \frac{2n(L(\theta_1) - L^*)}{\lambda\sigma^2 T}.$$

Thus, the optimal γ is:

$$\gamma = \sqrt{\frac{2n(L(\theta_1) - L^*)}{\lambda\sigma^2 T}}.$$

2.8. Substituting γ Back into the Inequality

Plug γ back into the inequality:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla L(\theta_t)\|^2] \leq \frac{2(L(\theta_1)-L^*)}{\gamma T}+\frac{\lambda\gamma\sigma^2}{n}=2\sqrt{\frac{\lambda\sigma^2(L(\theta_1)-L^*)}{nT}}.$$

2.9. Conclusion

Therefore, have shown:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla L(\theta_t)\|^2\right] \leq 2\sqrt{\frac{\lambda\sigma^2(L(\theta_1)-L^*)}{nT}}.$$

Since λ and $L(\theta_1)-L^*$ are constants, can conclude:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla L(\theta_t)\|^2\right] \in O\left(\sqrt{\frac{\sigma^2}{nT}}\right).$$

Question 5

Answer:

1. Comparison with Local SGD (τ = 1)

1.1. Convergence Rate of Local SGD with $\tau = 1$

In the Local SGD method with $\tau = 1$, each client performs one local update per communication round. The convergence rate for Local SGD in this case is known to be:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla L(\theta_t)\|^2\right] \in O\left(\frac{\sigma^2}{nT}\right).$$

1.2. Convergence Rate of DSGD

From **Question 4**, have established for DSGD:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla L(\theta_t)\|^2\right] \in O\left(\sqrt{\frac{\sigma^2}{nT}}\right).$$

2. Which Convergence Rate is Tighter?

Comparing the two rates:

- **Local SGD with** $\tau = 1$: Convergence rate is $O\left(\frac{\sigma^2}{nT}\right)$.

- **DSGD**: Convergence rate is $O\left(\sqrt{\frac{\sigma^2}{nT}}\right)$.

Observation:

- The convergence rate of Local SGD with $\tau = 1$ is **tighter** (i.e., faster convergence) than that of DSGD.
- The rate $O\left(\frac{\sigma^2}{nT}\right)$ decays linearly with T, while $O\left(\sqrt{\frac{\sigma^2}{nT}}\right)$ decays at a slower rate (square root).

3. How Can We Fix the Gap?

3.1. Tightening the Analysis for DSGD

The gap arises due to the conservative bounds used in the convergence analysis of DSGD. To fix the gap, we can:

1. Refine the Variance Bound:

Instead of using a loose bound on the variance, we can perform a more precise variance decomposition.

Consider the fact that the variance of the averaged gradient decreases with n, allowing us to bound the variance term more tightly.

2. Use Strong Convexity:

If the loss function L is strongly convex, we can leverage this property to obtain faster convergence rates.

Strong convexity allows us to bound the distance to the optimum, which can lead to linear convergence rates.

3. Employ Variance Reduction Techniques:

Methods like SVRG (Stochastic Variance Reduced Gradient) or SAGA can reduce the variance of the stochastic gradients, leading to improved convergence rates.

Incorporating these techniques into DSGD can help match the convergence rate of Local SGD.

3.2. Alternative Learning Rate Selection

In the analysis of DSGD, if choose a diminishing learning rate $\gamma_t = \Theta\left(\frac{1}{\sqrt{t}}\right)$, the convergence rate can improve. Specifically, may achieve:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \|\nabla L(\theta_t)\|^2\right] \in O\left(\frac{\sigma^2}{nT}\right).$$

However, this requires careful tuning of the learning rate schedule.