CORNELL TECH | HOME OF THE JACOBS TECHNION-CORNELL INSTITUTE

[Fall 2025] ECE 5290/7290 and ORIE 5290
Distributed Optimization for Machine Learning and AI

Homework 3

Gradescope Due: October 20th at 11:59PM

Objective of This Assignment

The goal of this assignment is to deepen your understanding of distributed optimization methods used in machine learning. You will analyze the convergence of consensus averaging (synchronous and gossip), understand how spectral properties control rates, compare mini-batch and parallel SGD, and study communication–computation trade-offs in local SGD. A coding problem will help you visualize convergence behaviors and the impact of network topology.

Instruction of Homework Submission

This assignment includes both an analytical part and a coding part (Problem 5). We will use Gradescope to check the correctness of your code. Therefore, you will see two separate assignments on Gradescope.

(a) A starter `.py` file is provided for Problem 5. Do not change the function names, signatures, or filename.

(b) Upload the written PDF to **Homework 3** and the code to **Homework 3 Coding**.

Question 1: Consensus Averaging and Spectral Gap (20 points)

Let $\mathbf{x}(0) \in \mathbb{R}^N$ be scalar values held by N nodes on an undirected connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The synchronous (Jacobi) consensus update is

$$\mathbf{x}(t+1) = \mathbf{W}\,\mathbf{x}(t), \qquad \mathbf{W} = \mathbf{I} - \alpha \mathbf{L},$$

where $\mathbf{L}$ is the graph Laplacian matrix and $0 < \alpha < 1/\lambda_{\max}(\mathbf{L})$; $\bar{x} = \frac{1}{N}\mathbf{1}^\top \mathbf{x}(0)$ is the global average that the algorithm wants to achieve the consensus on. Define the consensus error as $\mathbf{z}(t) = \mathbf{x}(t) - \bar{x}\,\mathbf{1}$.

(a) **(5 points)** Show that $\mathbf{1}$ is an eigenvector of $\mathbf{W}$ with eigenvalue 1; further, $\mathbf{W}$ is symmetric and doubly stochastic for the chosen α.

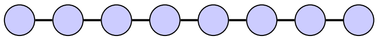
(b) **(5 points)** Let $1 = \lambda_1(\mathbf{W}) \geq \lambda_2(\mathbf{W}) \geq \cdots \geq \lambda_N(\mathbf{W}) > -1$. Prove the following

$$\|\mathbf{z}(t)\|_2 \leq \left|\lambda_2(\mathbf{W})\right|^t \|\mathbf{z}(0)\|_2.$$

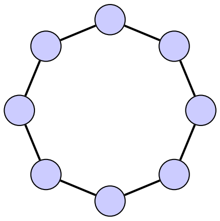
(c) **(10 points)** Consider three graphs with N nodes and $\mathbf{W} = \mathbf{I} - \alpha\mathbf{L}$:

- Path graph P_N with $\alpha = 1/\Delta$ (max degree $\Delta = 2$).
- Ring graph C_N with $\alpha = 1/\Delta$ (max degree $\Delta = 2$).
- Complete graph K_N with $\alpha = 1/N$.

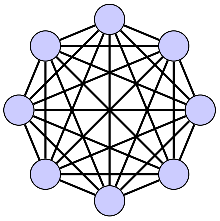
Recall $\lambda_2(\mathbf{L})$ for each graph from class and give a clean upper bound on $|\lambda_2(\mathbf{W})| = |1 - \alpha\lambda_2(\mathbf{L})|$. Which graph mixes fastest?



Path graph P_N.



Ring Graph - C_N



Complete Graph - C_N

Question 2: Randomized Gossip (Pairwise Averaging) (20 points)

At each step t, pick an edge $(i,j) \in \mathcal{E}$ uniformly at random; the two endpoints average their values:

$$x_i(t{+}1) = x_j(t{+}1) = \tfrac{1}{2}\big(x_i(t) + x_j(t)\big), \quad x_\ell(t{+}1) = x_\ell(t) \text{ for } \ell \notin \{i,j\}.$$

Define the disagreement potential $V(t) = \sum_{m=1}^{N} \big(x_m(t) - \bar{x}\big)^2$ with $\bar{x} = \frac{1}{N}\mathbf{1}^\top \mathbf{x}(0)$ as the global average.

(a) **(8 points)** Show that $V(t)$ is nonincreasing and derive

$$\mathbb{E}\big[V(t{+}1) \mid \mathbf{x}(t)\big] = V(t) - \frac{1}{2|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \big(x_i(t) - x_j(t)\big)^2.$$

(b) **(6 points)** Use $\sum_{(i,j)\in\mathcal{E}}(x_i - x_j)^2 = \mathbf{x}^\top \mathbf{L}\mathbf{x} \geq \lambda_2(\mathbf{L})\|\mathbf{x} - \bar{x}\mathbf{1}\|^2 = \lambda_2(\mathbf{L})\, V$ to prove

$$\mathbb{E}\big[V(t{+}1) \mid \mathbf{x}(t)\big] \;\leq\; \Big(1 - \frac{\lambda_2(\mathbf{L})}{2|\mathcal{E}|}\Big)\, V(t).$$

(c) **(6 points)** For the complete graph with uniform edge sampling, show

$$\mathbb{E}\big[V(t)\big] \;\le\; \Big(1 - \frac{1}{N}\Big)^{t} V(0).$$

(Hint: $\lambda_2(\mathbf{L}_{\mathsf{K}_N}) = N$ and $|\mathcal{E}| = \frac{N(N-1)}{2}$.)

Question 3: Mini-batch vs. Parallel SGD (15 points)

Consider empirical risk $F(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{x})$. Let K workers each draw independent mini-batches of size B (with replacement). Two **commonly used yet alternative** updates at iteration t:

1) **(Option 1: Single-node mini-batch)** One node samples a single mini-batch $\mathcal{B}$ of size KB:

$$g(\mathbf{x}(t)) = \frac{1}{KB}\sum_{i\in\mathcal{B}} \nabla f_i(\mathbf{x}(t)), \quad \text{update} \quad \mathbf{x}(t+1) = \mathbf{x}(t) - \eta\, g(\mathbf{x}(t)).$$

2) **(Option 2: Parallel averaging)** Each worker computes $g^{(k)}(\mathbf{x}(t)) = \frac{1}{B}\sum_{i\in\mathcal{B}_k} \nabla f_i(\mathbf{x}(t))$, then average:

$$\bar{g}(\mathbf{x}(t)) = \frac{1}{K}\sum_{k=1}^{K} g^{(k)}(\mathbf{x}(t)) \;\; \text{update} \quad \mathbf{x}(t+1) = \mathbf{x}(t) - \eta\, \bar{g}(\mathbf{x}(t)).$$

(a) **(5 points)** Show both estimators in Option 1 and Option 2 are unbiased for $\nabla F(\mathbf{x}(t))$.

(b) **(5 points)** Assuming the per-sample gradient variance is bounded by $\mathbb{E}\|\nabla f_i(\mathbf{x}) - \nabla F(\mathbf{x})\|^2 \leq \sigma^2$, prove

$$\mathrm{Var}(\bar{g}) \;=\; \frac{\sigma^2}{KB}, \qquad \mathrm{Var}(g) \;=\; \frac{\sigma^2}{KB}.$$

(c) **(5 points)** Discuss when parallel averaging and single-node mini-batch are *not* equivalent in practice.

Question 4: Local SGD - Communication/Computation Trade-offs (25 points)

Consider K workers collaboratively minimizing a global objective

$$F(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} F_k(\mathbf{x}) \quad \text{with} \quad F_k(\mathbf{x}) = \frac{1}{B}\sum_{i\in\mathcal{B}_k} f_i(\mathbf{x}),$$

where worker k has access to local data and computes stochastic gradients $g^{(k)}$. Each worker starts from a common model $\mathbf{x}(t)$, performs τ local SGD steps (with $\mathbf{x}_0^{(k)}(t) = \mathbf{x}(t)$):

$$\mathbf{x}_{s+1}^{(k)}(t) = \mathbf{x}_s^{(k)}(t) - \eta\, g_s^{(k)}(t), \qquad s = 0, 1, \ldots, \tau - 1,$$

and then all workers synchronize by averaging:

$$\mathbf{x}(t+1) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{x}_\tau^{(k)}(t).$$

Assume that each local function F_k is μ-strongly convex and L-smooth, and that the stochastic gradients have bounded variance σ^2.

(a) **(15 points) For ECE 7290 students:** (Sketch) Show that compared to fully synchronized mini-batch SGD (from **Question 3**), local SGD includes an additional *drift term* due to model divergence between averaging rounds. Derive a bound of the form

$$\mathbb{E}\|\mathbf{x}(t+1) - \mathbf{x}^\star\|^2 \;\le\; \rho^\tau\, \mathbb{E}\|\mathbf{x}(t) - \mathbf{x}^\star\|^2 \;+\; C_1 \frac{\eta\sigma^2}{\mu K} \;+\; C_2\, \eta^2 \tau \Gamma^2,$$

where Γ^2 captures the gradient dissimilarity (data heterogeneity) across nodes. Explain the dependence on τ. For homogeneous data ($\Gamma^2 \approx 0$), what τ do you recommend? For heterogeneous data (large Γ^2), how would you adjust τ?

(b) **(15 points) For ECE/ORIE 5290 students:** Assuming (a) holds, for homogeneous data ($\Gamma^2 \approx 0$), what τ do you recommend? For heterogeneous data (large Γ^2), how would you adjust τ?

(c) **(10 points)** Suppose the total training time is limited. Each local iteration costs c_{comp} time units for computation, and each synchronization costs c_{comm}. Qualitatively describe how to choose (η, τ) to balance runtime efficiency and convergence accuracy.

Question 5: Coding - Consensus and parallel SGD (20 points)

You will implement two small simulations.

(A) Consensus vs. Gossip (10 points) Generate $N = 20$ i.i.d. initial values in $[0, 1]$. Consider:

- Synchronous consensus $\mathbf{x}(t+1) = \mathbf{W}\mathbf{x}(t)$ on a ring with $\alpha = 1/2$.
- Randomized gossip: pick a random edge (i, j) on the ring and average the pair.

Plot 1: the disagreement $V(t) = \sum_i (x_i(t) - \bar{x})^2$ vs. iterations for both methods (same random seed).
Plot 2: sample trajectories of two nodes to illustrate smoothing.

(B) Local vs. Parallel SGD (10 points) Binary logistic regression on a synthetic dataset (will be posted by TA on Canvas), split evenly across $K = 4$ workers. Compare:

- Parallel (synchronous) mini-batch SGD with global batch size KB.
- Local SGD with the same local batch size B and averaging period $\tau \in \{1, 5, 20\}$.

Plot 3: training loss $F(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} F_k(\mathbf{x})$ vs. # of *communication rounds.* **Short analysis (5–7 sentences):** discuss the effect of τ on speed/accuracy and when local SGD matches parallel SGD.