Question 1: Answer the following questions about SGD (2 x 6 = 12 marks)

a. Conditions on the learning rate ηt\eta_tηt for convergence of SGD:

The learning rate η_t must satisfy two conditions for convergence:

$$\eta_t \to 0 \quad \text{as} \quad t \to \infty$$
$$\sum_{t=1}^{\infty} \eta_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty$$

b. Does θ_t converge in SGD?

Yes, θ_t converges to a local or global minimum if the learning rate is properly decayed and the loss function is convex. In non-convex cases, θ_t may converge to a local minimum.

c. Impact of condition number of A_i on momentum-based SGD:

As the condition number (ratio of largest to smallest eigenvalue) of A_i increases, SGD with momentum **becomes faster** compared to standard SGD. This is because momentum helps accelerate convergence in the direction of the gradient, and a higher condition number indicates a more ill-conditioned problem, which benefits from the smoothing effect of momentum.

d. What is adaptive gradient update?

Adaptive gradient update methods (e.g., AdaGrad, RMSprop) adjust the learning rate based on the history of gradients, giving smaller updates for frequently occurring features. It can be used with both stochastic and non-stochastic optimization.

e. Can stochastic averages gradient descent be used with minibatch updates?

Yes, stochastic average gradient descent (SAG) can be used with minibatches by averaging the gradients computed over multiple samples within the batch to update the parameters.

f. What is internal covariate shift in Batch Normalization?

Internal covariate shift refers to the change in the distribution of network activations during training, which can slow down learning. Batch normalization addresses this by normalizing the activations across mini-batches to stabilize the training process.

Question 2: Answer the following questions about ADMM (2 x 4 = 8 marks)

a. Purpose of primal and dual residuals in ADMM:

Primal residual measures the consistency between two primal variables x and z, ensuring they agree. Dual residual measures the change in dual variables, helping ensure convergence and stabilization in the updates.

b. Main advantage of ADMM over dual decomposition:

ADMM combines the benefits of dual decomposition with augmented Lagrangian methods, which allows for better convergence in non-smooth problems and enables distributed optimization.

c. ADMM-based consensus for IoT networks:

Yes, ADMM can be used for decentralized consensus by having each IoT device update its local variable based on local data, while the ADMM algorithm ensures that these local variables converge to a global consensus through communication with neighboring devices.

d. Purpose of parameter ρ in the augmented Lagrangian:

The parameter ρ controls the penalty for violating the equality constraint. If ρ=0, the method becomes the regular Lagrangian method, losing its augmented property, which can lead to slower convergence or even divergence.