On the Primal Feasibility in Dual Decomposition Methods Under Additive and Bounded Errors

Chathuranga Weeraddana

Department of Electronic & Telecommunication Engineering

University of Moratuwa

Workshop

5G Evolution & 6G

ICIIS 2021

A Joint Work with

- Hansi Abeynanda (SLIIT, Sri Lanka)
- G. H. J. Lanel (University of Sri Jayewardenepura, Sri Lanka)
- Carlo Fischione (KTH Royal Institute of Technology, Swden)

Main Area of Focus

Distributed Optimization

- distributed optimization, roots trace back to
  - Dangtiz & Wolfe, 1960
  - Tsitsiklis, 1984
- the most common algorithms
  - standard gradient + gossip-type averaging, Nedic & Ozdaglar, 2009
  - methods relying on problem structure
    * dual decomposition methods
    * primal decomposition methods
    * alternating direction method of multipliers (ADMM)
    * dual averaging methods
    * block-coordinate methods

Distributed Optimization

- distributed optimization, roots tracked back to
  - Dangtiz & Wolfe, 1960
  - Tsitsiklis, 1984
- the most common algorithms
  - standard gradient + gossip-type averaging, Nedic & Ozdaglar, 2009
  - methods relying on problem structure
    - dual decomposition methods 
    - primal decomposition methods
    - alternating direction method of multipliers (ADMM)
    - dual averaging methods
    - block-coordinate methods

Distributed Optimization in Future Wireless Systems

Distributed Optimization

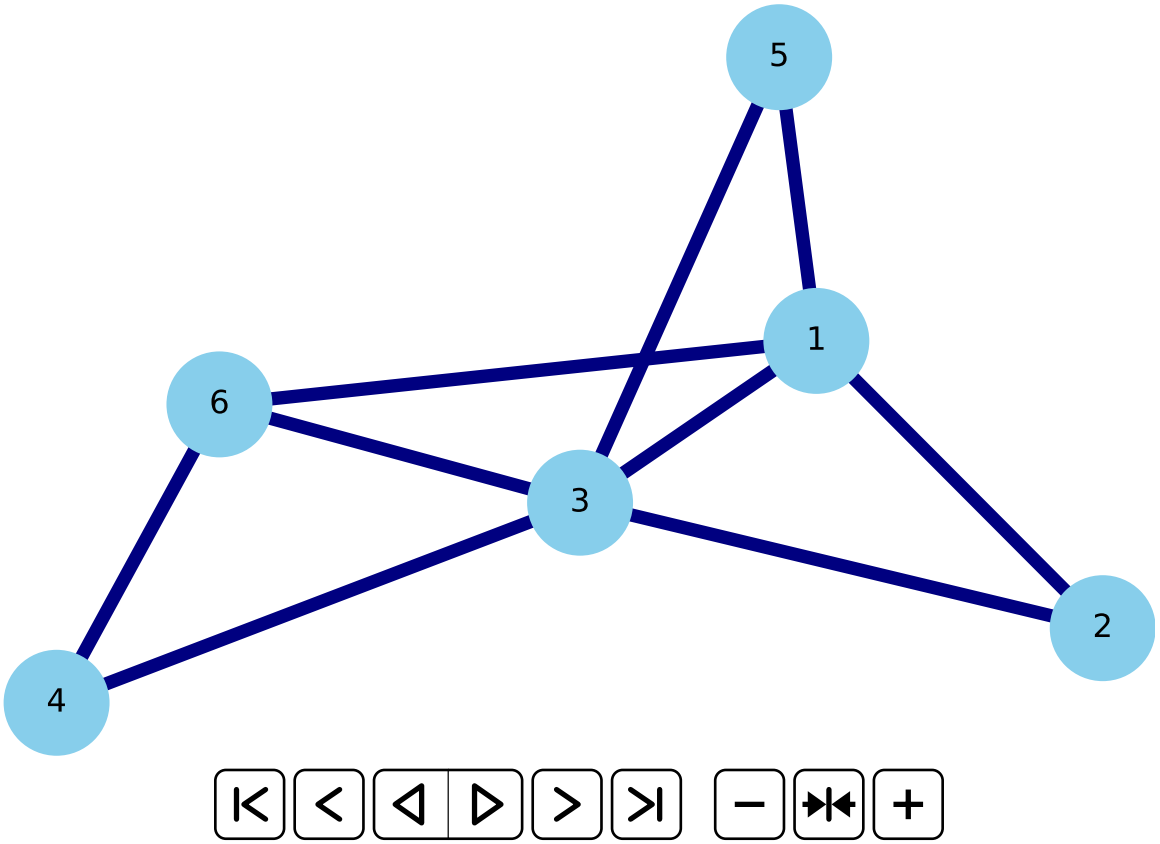
- already been studied for decades
  - decentralized radio resource management
  - distributed estimation in sensor networks
  - many others
- recently been an overwhelmingly appealing requirement
  - invasion of machine learning to wireless network designs
  - centralized, big data, toward massive, distributed "small" data
  - "collective network intelligence": intelligence is pushed at the edge
  - machine learning over wireless (Federated Learning, McMahn, 2016)

Distributed Optimization: Key Idea

Distributed Optimization: Main Steps

- setting
  - finitely many subsystems (e.g., hospitals)
  - they have a common goal (e.g., design a classifier to detect cancers)
- distributed solution method
  - subsystems make local decisions/computations
  - subsystems coordinate through a thin protocol
  - above steps are iteratively performed

Distributed Optimization: Main Steps



Convergence of the Algorithm?

Under Reasonable Conditions

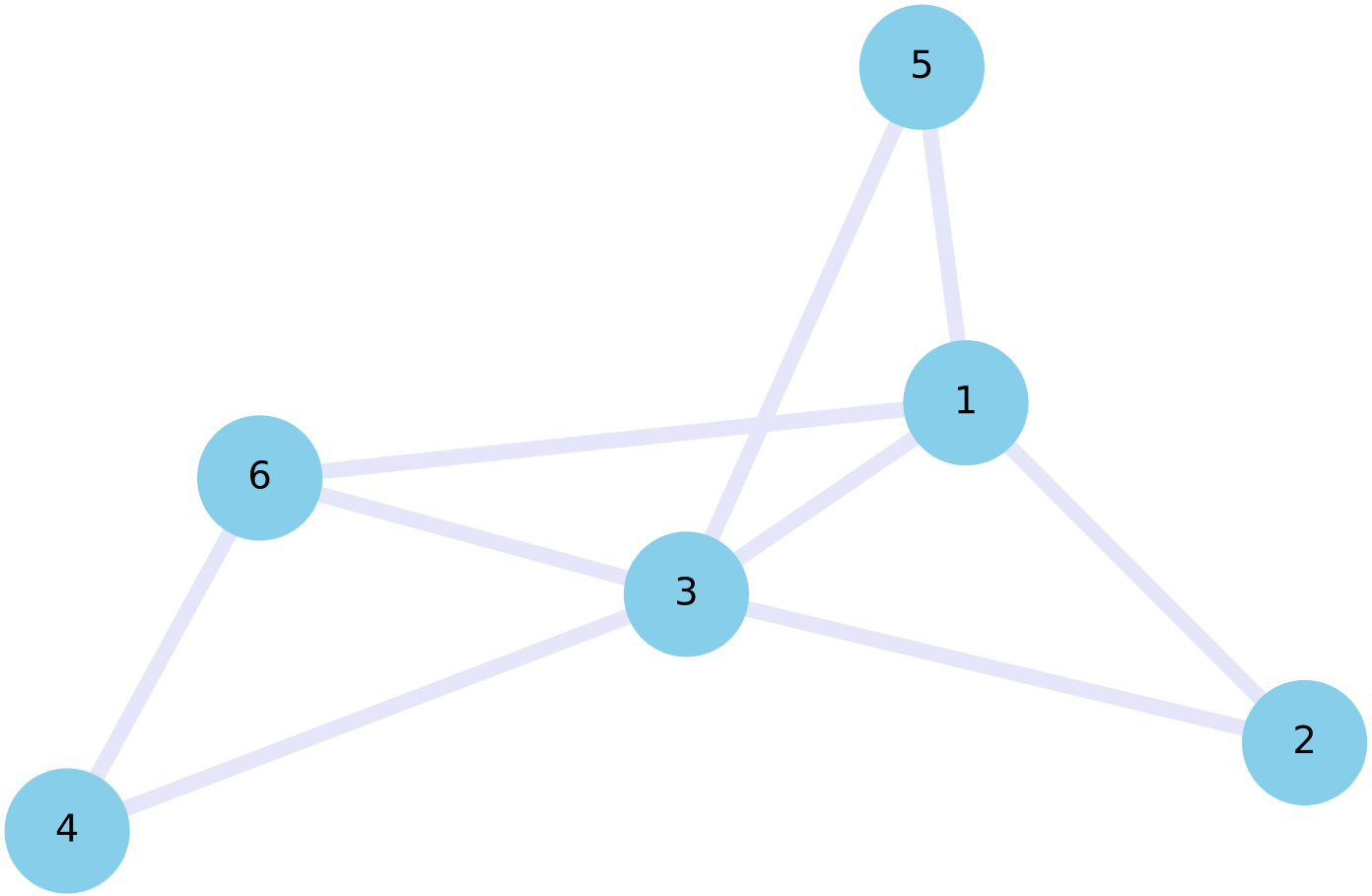
Convergence is Guaranteed!

Sources of Inexactness

Inexact Steps of the Solution Method

- errors at local decisions
- errors at the coordination

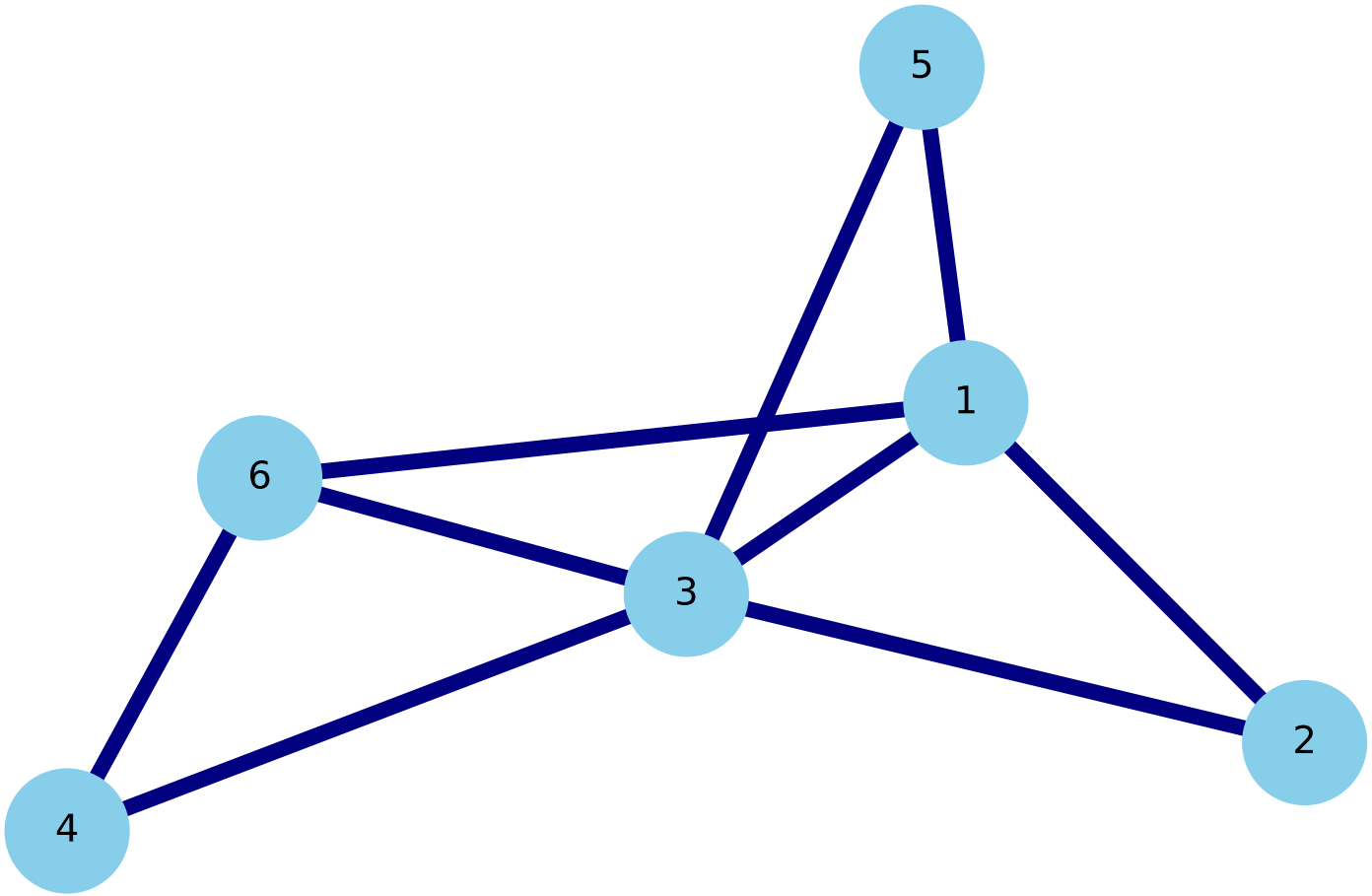
Errors at Local Decisions



Errors at Local Decisions

- approximation errors
- subproblem solver accuracy
- early termination of subproblem algorithms

Errors at the Coordination



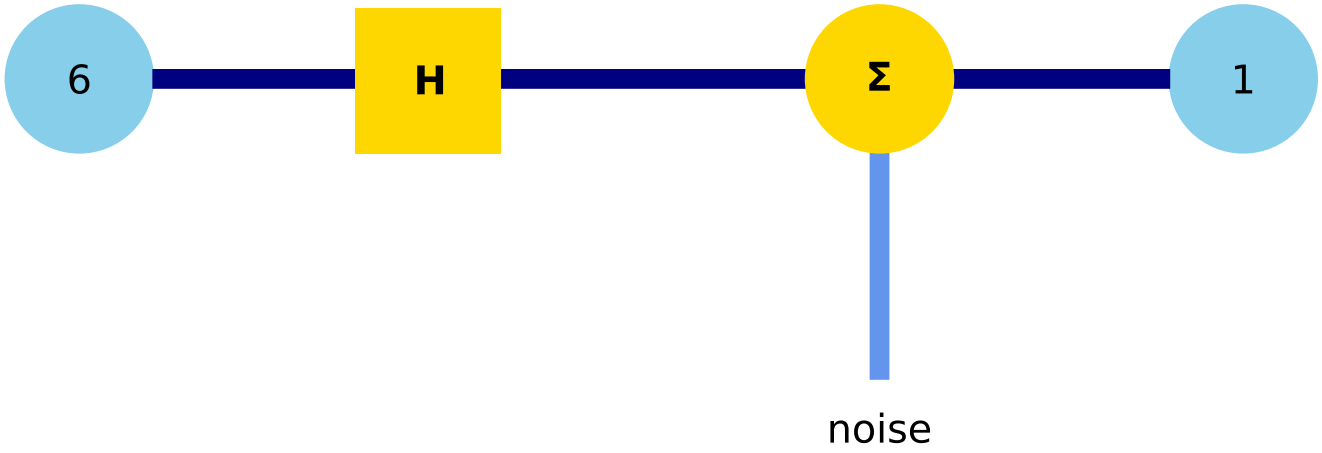
Errors at the Coordination



Errors at the Coordination



Errors at the Coordination



Errors at the Coordination

- quantization errors
- measurement errors
- noise induced by wireless channel
- errors due to fading

Convergence of Inexact Algorithms?

Convergence is Not Guaranteed
In General!

Research Directions

- convergence under inexact settings?
  - convergence to a neighborhood of the optimality?
- rate of convergence?

Technicalities: Dual Decomposition in the Considered Problem Setting

Problem Formulation

Problem Formulation

$$\begin{array}{ll} \underset{\mathbf{z}}{\text{minimize}} & \sum_{i=1}^{m} f_i(\mathbf{z}) \\ \text{subject to} & \mathbf{z} \in \mathcal{Y} \end{array} \tag{1}$$

- global consensus optimization problems
- m : number of subsystems (SS)
- $f_i : \mathbb{R}^n \to \mathbb{R}$: cost function at subsystem i
- common goal is to solve (1), i.e., compute an optimal solution $\mathbf{z}^\star$
- subsystems want to solve (1) in a distributed manner
  - we apply dual decomposition methods

An Equivalent Problem (Decomposition)



- the public variable is $\mathbf{z}$
- $\mathbf{y}_i \in \mathbb{R}^n$: a local versions of the public variable $\mathbf{z}$
- i.e., $\mathbf{y}_1 = \mathbf{y}_2 = \cdots = \mathbf{y}_m = \mathbf{z}$

The Equivalent Problem Formulation

$$\begin{array}{ll} \underset{\mathbf{y}_1,\ldots,\mathbf{y}_m}{\text{minimize}} & f(\mathbf{y_1},\ldots,\mathbf{y_m}) = \sum_{i=1}^{m} f_i(\mathbf{y}_i) \\ \text{subject to} & \mathbf{y}_i \in \mathcal{Y},\ i = 1,\ldots,m \\ & \mathbf{y}_i = \mathbf{y}_{i+1},\ i = 1,\ldots,m-1 \end{array} \tag{2}$$

- $\mathbf{y}_i \in \mathbb{R}^n$: a local versions of the public variable $\mathbf{z}$
- dualize the last set of constraints
  - introduce Lagrange multiplier $\boldsymbol{\lambda}_i \in \mathbb{R}^n,\ i = 1,\ldots,m-1$

The Dual Function and the Dual Problem

- Lagrangian is given by

$$L(\mathbf{y}_1, \ldots, \mathbf{y}_m, \boldsymbol{\lambda}) = \sum_{i=1}^{m} f_i(\mathbf{y}_i) + \sum_{i=1}^{m-1} \boldsymbol{\lambda}_i^{\mathsf{T}}(\mathbf{y}_i - \mathbf{y}_{i+1}) \tag{3}$$

- dual function is given by

$$g(\boldsymbol{\lambda}) = \inf_{\mathbf{y}_i \in \mathcal{Y},\ i=1,\ldots,m} \left[\sum_{i=1}^{m} f_i(\mathbf{y}_i) + \sum_{i=1}^{m-1} \boldsymbol{\lambda}_i^{\mathsf{T}}(\mathbf{y}_i - \mathbf{y}_{i+1}) \right] \tag{4}$$

$$= \sum_{i=1}^{m} \inf_{\mathbf{y}_i \in \mathcal{Y}} \left[f_i(\mathbf{y}_i) + (\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_{i-1})^{\mathsf{T}} \mathbf{y}_i \right]. \tag{5}$$

Solving the Problem

Computations to be Conducted

- solving the problem: perform in an iterative manner
  - primal variable update, i.e., $\mathbf{y}_i$s
  - dual variable update, i.e., $\boldsymbol{\lambda}_i$s

Primal Variable Update (Exact)

Primal Variable Update

- given $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{i-1}$
- the following problem is to be solved

$$\begin{array}{ll} \underset{\mathbf{y}}{\text{minimize}} & \sum_{i=1}^{m} f_i(\mathbf{y}_i) + \sum_{i=1}^{m-1} \boldsymbol{\lambda}_i^{\mathsf{T}}(\mathbf{y}_i - \mathbf{y}_{i+1}) \\ \text{subject to} & \mathbf{y}_i \in \mathcal{Y}, \qquad i \in \{1, \ldots, m\} \end{array} \tag{6}$$

- note that the Lagrangian minimization splits or decomposes
- thus it can be performed completely independently, e.g., in parallel

Dual Variable Update (Exact)

Dual Variable Update

- given $\mathbf{y}_1, \ldots, \mathbf{y}_m$

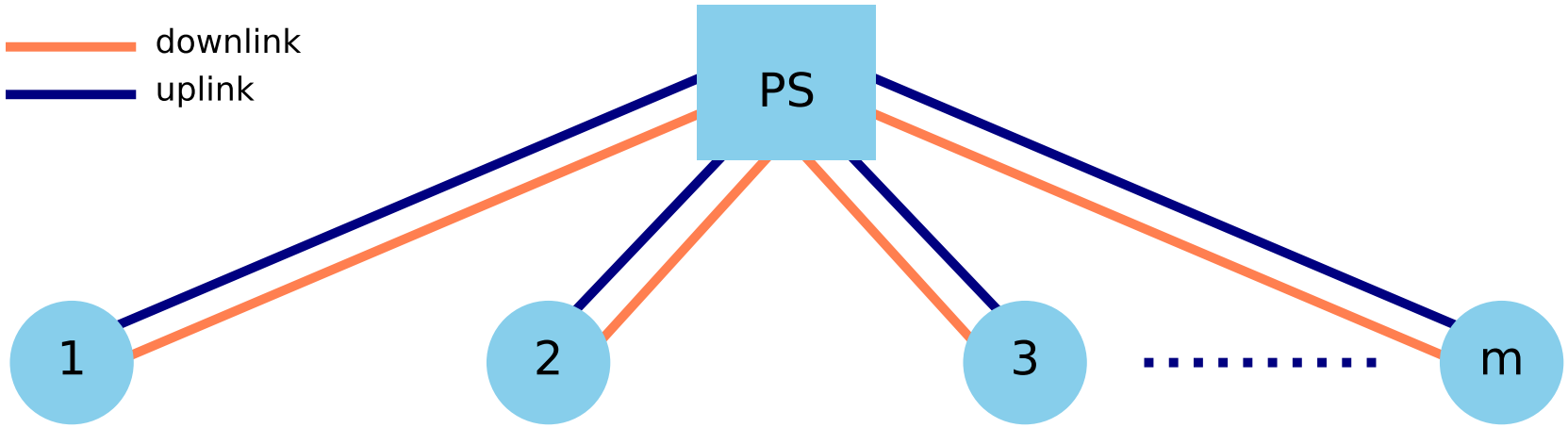
$$\boldsymbol{\lambda} := \boldsymbol{\lambda} + \gamma \mathbf{d} \tag{7}$$

- $\mathbf{d} \in \mathbb{R}^{n(m-1)}$ is a subgradient of the dual function at $\boldsymbol{\lambda}$
- more specifically, we have

$$\begin{pmatrix} \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2 \\ \boldsymbol{\lambda}_3 \\ \vdots \\ \vdots \\ \boldsymbol{\lambda}_{m-1} \end{pmatrix}_{n(m-1)\times 1} := \begin{pmatrix} \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2 \\ \boldsymbol{\lambda}_3 \\ \vdots \\ \vdots \\ \boldsymbol{\lambda}_{m-1} \end{pmatrix} + \gamma \begin{pmatrix} \mathbf{y}_1 - \mathbf{y}_2 \\ \mathbf{y}_2 - \mathbf{y}_3 \\ \mathbf{y}_3 - \mathbf{y}_4 \\ \vdots \\ \vdots \\ \mathbf{y}_{m-1} - \mathbf{y}_m \end{pmatrix}$$

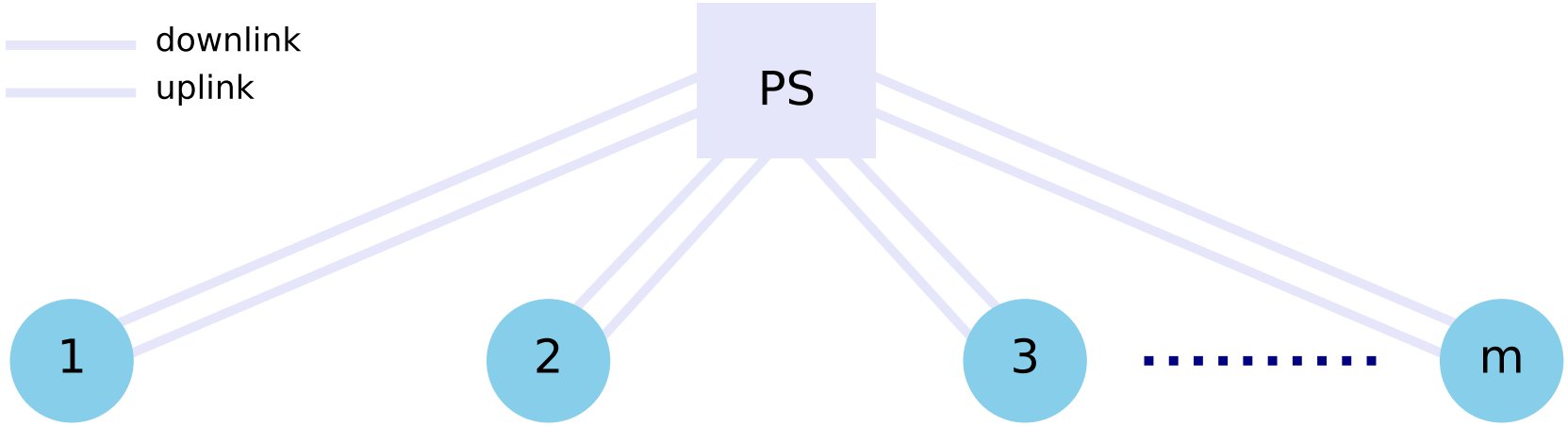
Solving the Problem in a Distributed Manner
A Federated Setting (Exact)

A Federated Setting



- a machine learning setting where SSs (clients) collaborate
- PS is called the parameter server (a central server)
- the coordination is done by PS

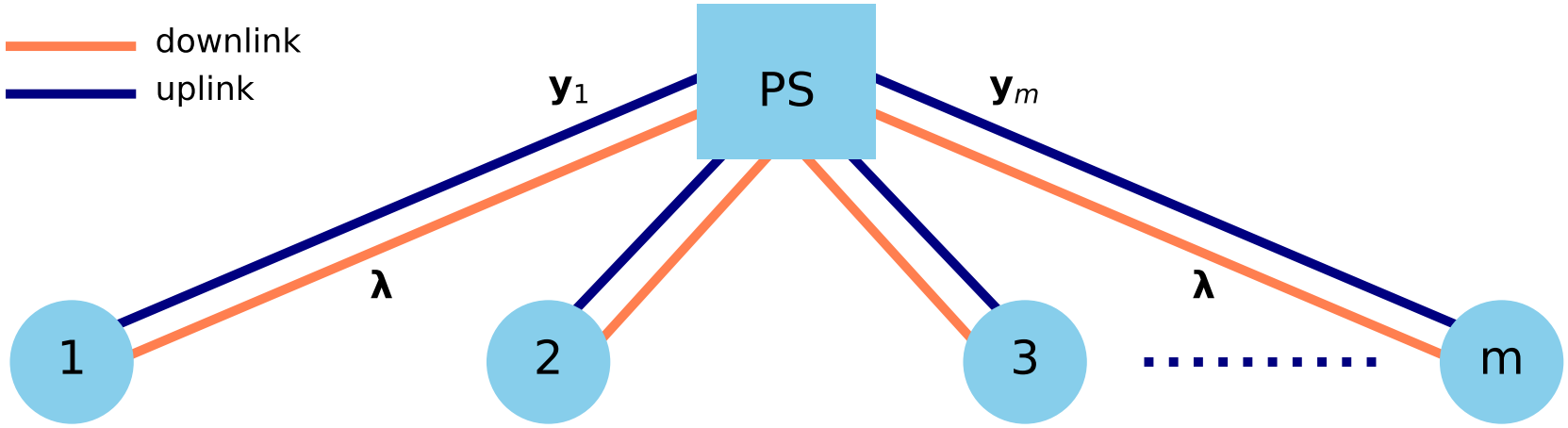
Local Decisions at Subsystem i



- recall.. the problem of Lagrangian minimization decomposes
- thus, given $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda}_{i-1}$, SS i perform its own primal variable update:

$$\mathbf{y}_i := \underset{\mathbf{v}_i \in \mathcal{Y}}{\arg\min} \quad f_i(\mathbf{v}_i) + (\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_{i-1})^{\mathsf{T}} \mathbf{v}_i \tag{8}$$

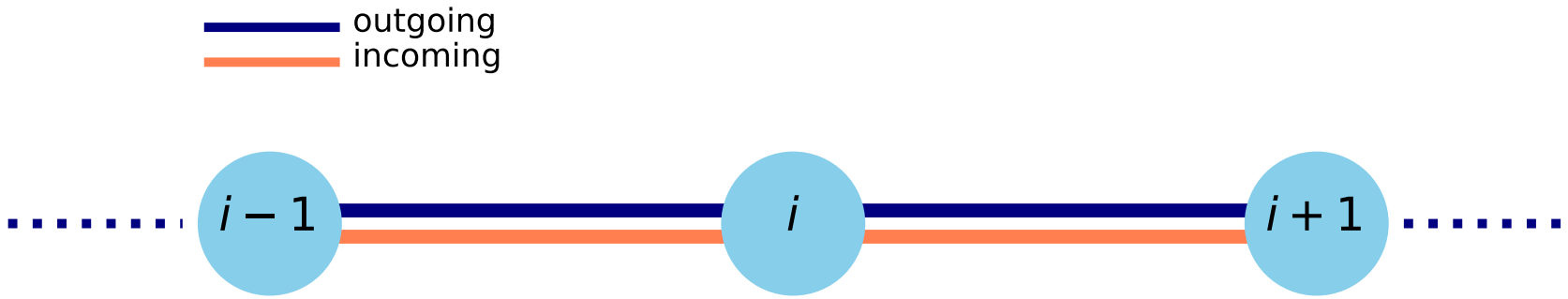
Subsystem Coordination



- subsystem i transmits $\mathbf{y}_i$ to PS (uplink)
- PS updates $\boldsymbol{\lambda}$ by using $\mathbf{y}_i$s
- PS broadcasts $\boldsymbol{\lambda}$ to all the subsystems (downlink)

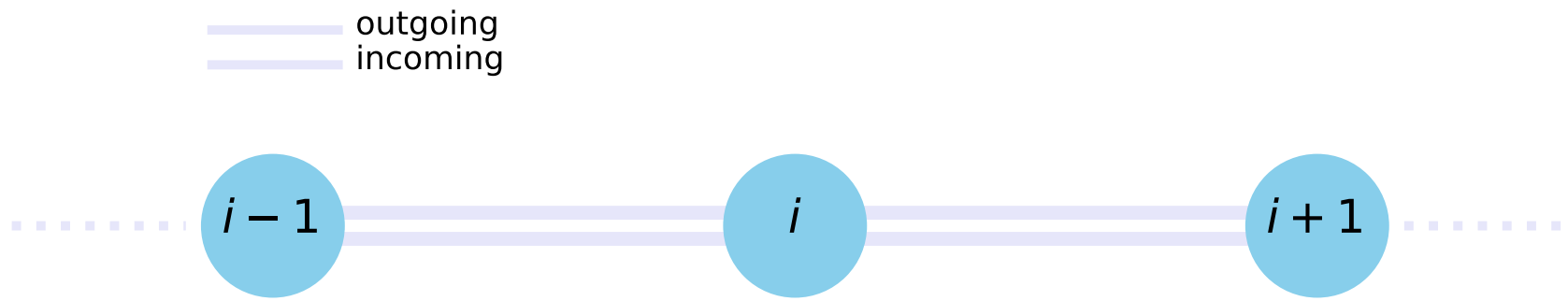
Solving the Problem in a Distributed Manner
A Setting with Local Communications (Exact)

A Fully Distributed Setting



- a machine learning setting where no central authorities
- SS i communicate with its neighbors $i-1$ and $i+1$

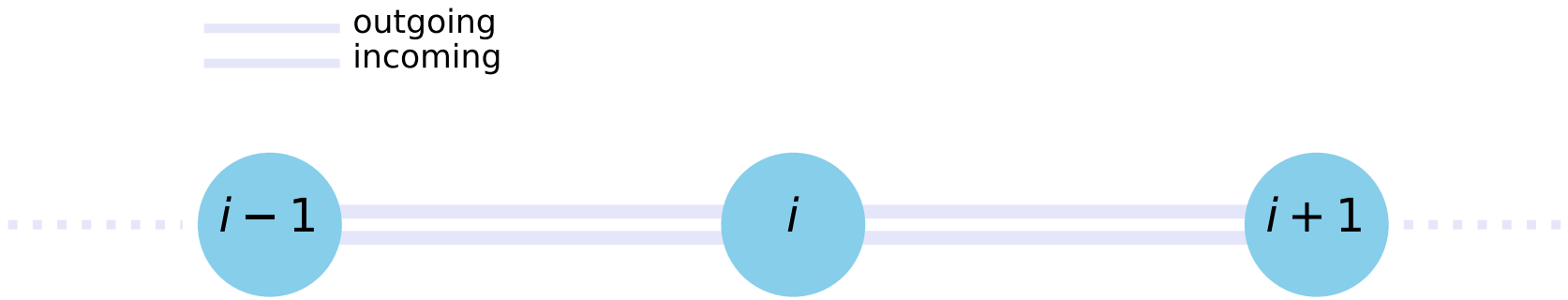
Local Decisions at Subsystem i



- given $\boldsymbol{\lambda}_i$, $\boldsymbol{\lambda}_{i-1}$, $\mathbf{y}_{i-1}$, $\mathbf{y}_{i+1}$
- perform the own primal variable update:

$$\mathbf{y}_i := \underset{\mathbf{v}_i \in \mathcal{Y}}{\arg\min} \quad f_i(\mathbf{v}_i) + (\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_{i-1})^{\mathsf{T}} \mathbf{v}_i \tag{9}$$

Local Decisions at Subsystem i

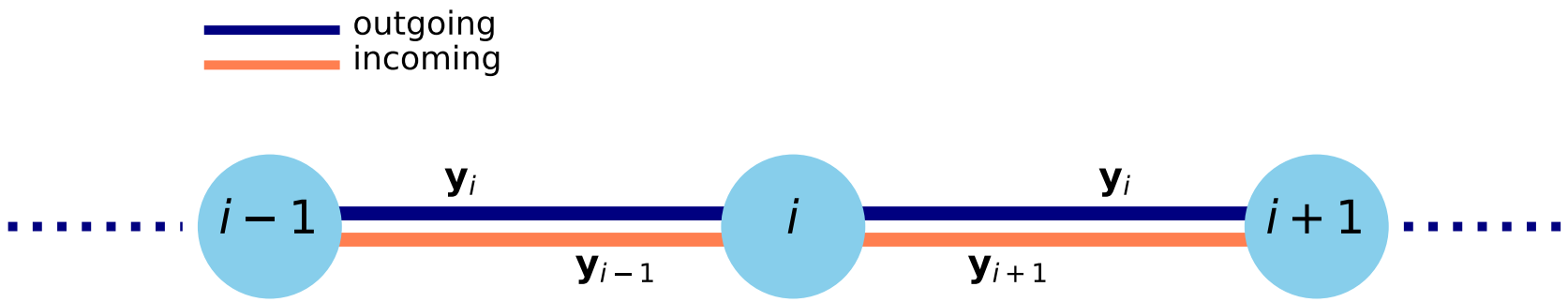


- given $\boldsymbol{\lambda}_i$, $\boldsymbol{\lambda}_{i-1}$, $\mathbf{y}_{i-1}$, $\mathbf{y}_{i+1}$

- perform the dual variable update:

$$\begin{pmatrix} \boldsymbol{\lambda}_{i-1} \\ \boldsymbol{\lambda}_i \end{pmatrix}_{2n\times 1} := \begin{pmatrix} \boldsymbol{\lambda}_{i-1} \\ \boldsymbol{\lambda}_i \end{pmatrix} + \gamma \begin{pmatrix} \mathbf{y}_{i-1} - \mathbf{y}_i \\ \mathbf{y}_i - \mathbf{y}_{i+1} \end{pmatrix}$$

Subsystem Coordination



- subsystem i transmits $\mathbf{y}_i$ to subsystems $i-1$ and $i+1$
- subsystem i receives $\mathbf{y}_{i-1}$ and $\mathbf{y}_{i+1}$ from subsystems $i-1$ and $i+1$

Solving the Problem in a Distributed Manner
A Federated Setting (Inexact)

Local Decisions at Subsystem i

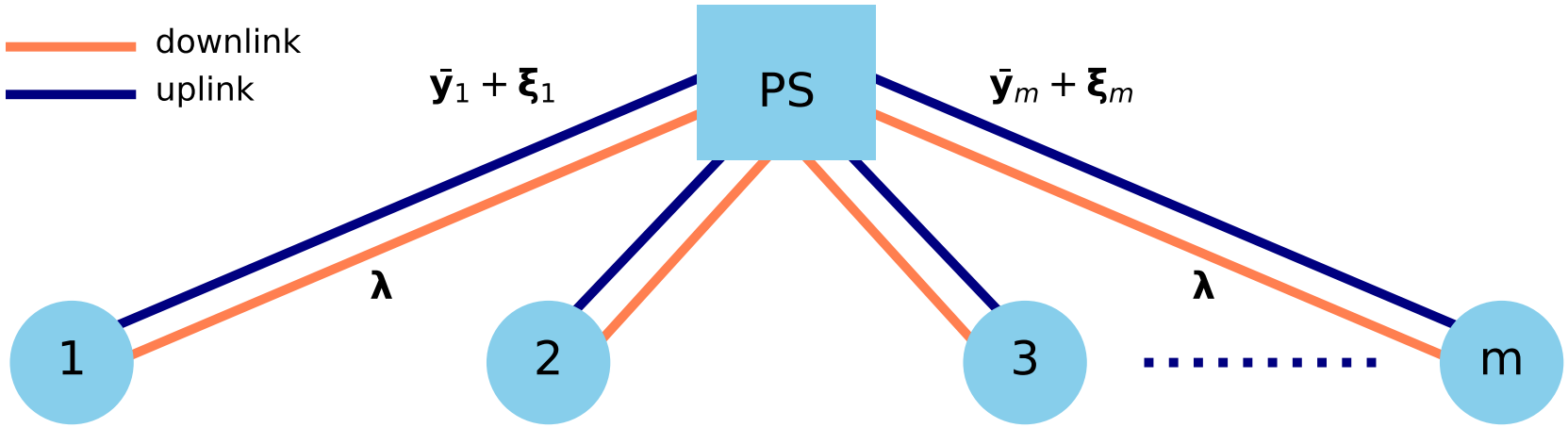
- given $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda}_{i-1}$
- perform the own primal variable update:

$$\bar{\mathbf{y}}_i := \left[\underset{\mathbf{v}_i \in \mathcal{Y}}{\arg\min} \quad f_i(\mathbf{v}_i) + (\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_{i-1})^{\mathsf{T}} \mathbf{v}_i \right] + \boldsymbol{\epsilon}_i \tag{10}$$

$$= \mathbf{y}_i + \boldsymbol{\epsilon}_i \tag{11}$$

- assumptions can be imposed on $\boldsymbol{\epsilon}_i$, e.g., $\|\boldsymbol{\epsilon}_i\|$ is bounded

Subsystem Coordination



- subsystem i transmits $\bar{\mathbf{y}}_i$ to PS (uplink)
- PS receives $\hat{\mathbf{y}}_i := \bar{\mathbf{y}}_i + \boldsymbol{\xi}_i = (\mathbf{y}_i + \boldsymbol{\epsilon}_i) + \boldsymbol{\xi}_i = \mathbf{y}_i + (\boldsymbol{\epsilon}_i + \boldsymbol{\xi}_i)$
- assumptions can be imposed on $\boldsymbol{\xi}_i$, e.g., $\|\boldsymbol{\xi}_i\|$ is bounded
- PS update and broadcasts $\boldsymbol{\lambda}$ to all the subsystems (downlink)

Solving the Problem in a Distributed Manner
A Setting with Local Communications (Inexact)

Local Decisions at Subsystem i

- given $\boldsymbol{\lambda}_i$, $\boldsymbol{\lambda}_{i-1}$, and quantized vectors $Q(\mathbf{y}_{i-1})$, $Q(\mathbf{y}_{i+1})$
- perform the own primal variable update:

$$\bar{\mathbf{y}}_i := \left[\underset{\mathbf{v}_i \in \mathcal{Y}}{\arg\min} \quad f_i(\mathbf{v}_i) + (\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_{i-1})^{\mathsf{T}} \mathbf{v}_i \right] + \boldsymbol{\epsilon}_i \tag{12}$$

$$= \mathbf{y}_i + \boldsymbol{\epsilon}_i \tag{13}$$

- assumptions can be imposed on $\boldsymbol{\epsilon}_i$, e.g., $\|\boldsymbol{\epsilon}_i\|$ is bounded
- for simplicity suppose, $\boldsymbol{\epsilon}_i = 0$ for all i

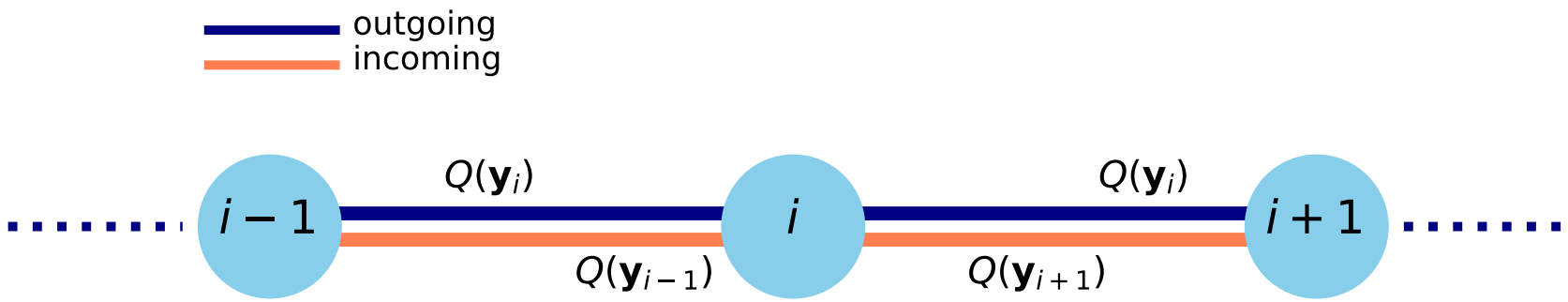
Local Decisions at Subsystem i

- perform the dual variable update:

$$\begin{pmatrix} \boldsymbol{\lambda}_{i-1} \\ \boldsymbol{\lambda}_i \end{pmatrix}_{2n\times 1} := \begin{pmatrix} \boldsymbol{\lambda}_{i-1} \\ \boldsymbol{\lambda}_i \end{pmatrix} + \gamma \begin{pmatrix} Q(\bar{\mathbf{y}}_{i-1}) - Q(\bar{\mathbf{y}}_i) \\ Q(\bar{\mathbf{y}}_i) - Q(\bar{\mathbf{y}}_{i+1}) \end{pmatrix}$$

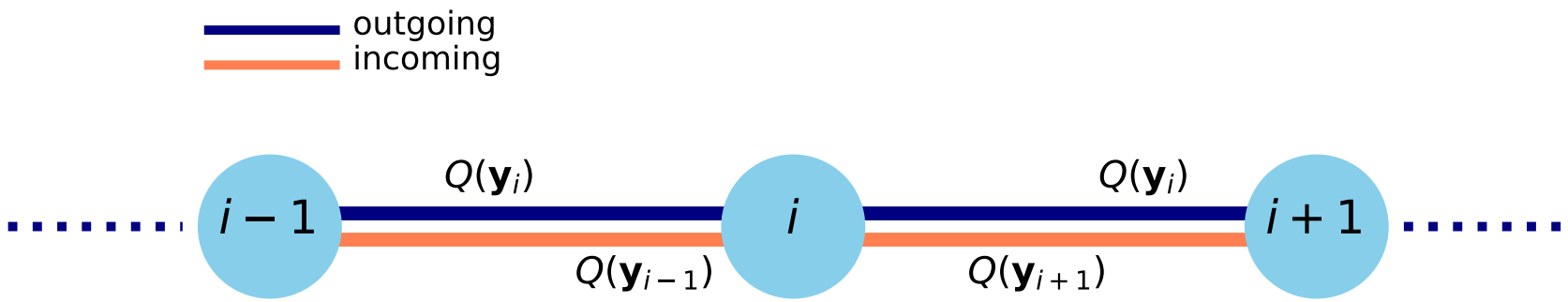
$$= \begin{pmatrix} \boldsymbol{\lambda}_{i-1} \\ \boldsymbol{\lambda}_i \end{pmatrix} + \gamma \begin{pmatrix} Q(\mathbf{y}_{i-1}) - Q(\mathbf{y}_i) \\ Q(\mathbf{y}_i) - Q(\mathbf{y}_{i+1}) \end{pmatrix}$$

Subsystem Coordination



- subsystem i quantizes $\mathbf{y}_i$ to yield $Q(\mathbf{y}_i)$
- quantization distortion $\boldsymbol{\xi}_i = Q(\mathbf{y}_i) - \mathbf{y}_i$
  - assumptions can be imposed on $\boldsymbol{\xi}_i$, e.g., $\|\boldsymbol{\xi}_i\|$ is bounded

Subsystem Coordination



- subsystem i transmits $Q(\mathbf{y}_i)$ to subsystems $i-1$ and $i+1$
- subsystem i receives $Q(\mathbf{y}_{i-1})$, $Q(\mathbf{y}_{i+1})$ from subsystems $i-1$, $i+1$

Main Assumptions

General

Assumption 1: Closedness

The set $\mathcal{Y}$ and the functions f_is are *closed*.

Assumption 2: Absolute Deterministic Distortion

The distortion $\mathbf{r}_i$ associated with $\mathbf{y}_i$, $i = 1, \ldots, m$, $k \in \mathbb{Z}_+^0$ is bounded by $\varepsilon_i \in \mathbb{R}$, i.e.,

$$\|\mathbf{r}_i\| \leq \varepsilon_i, \quad i = 1, \ldots, m, \quad k \in \mathbb{Z}_+^0.$$

Assumption 3: Strong Duality

The optimal values $p^\star$ and $d^\star$ of primal problem (2) and the dual problem, respectively, are attained. Moreover, strong duality between holds, i.e.,

$$p^\star = f(\mathbf{y}^\star) = d^\star. \tag{14}$$

Specific

Assumption I: Strongly Convex Objectives at SSs

The functions f_is in problem (2) are strongly convex with constant $\mu_i > 0$, $i = 1, \ldots, m$.

Assumption II: Gradient Lipschitz Continuous Objectives at SSs

The set $\mathcal{Y}$ in problem (2) equals $\mathbb{R}^n$. Moreover, f_is are differentiable and the gradients ∇f_is are Lipschitz continuous on $\mathbb{R}^n$ with constant $L_i > 0$, $i = 1, \ldots, m$.

Main Results:
Convergence in Dual Domain

Corollary 1

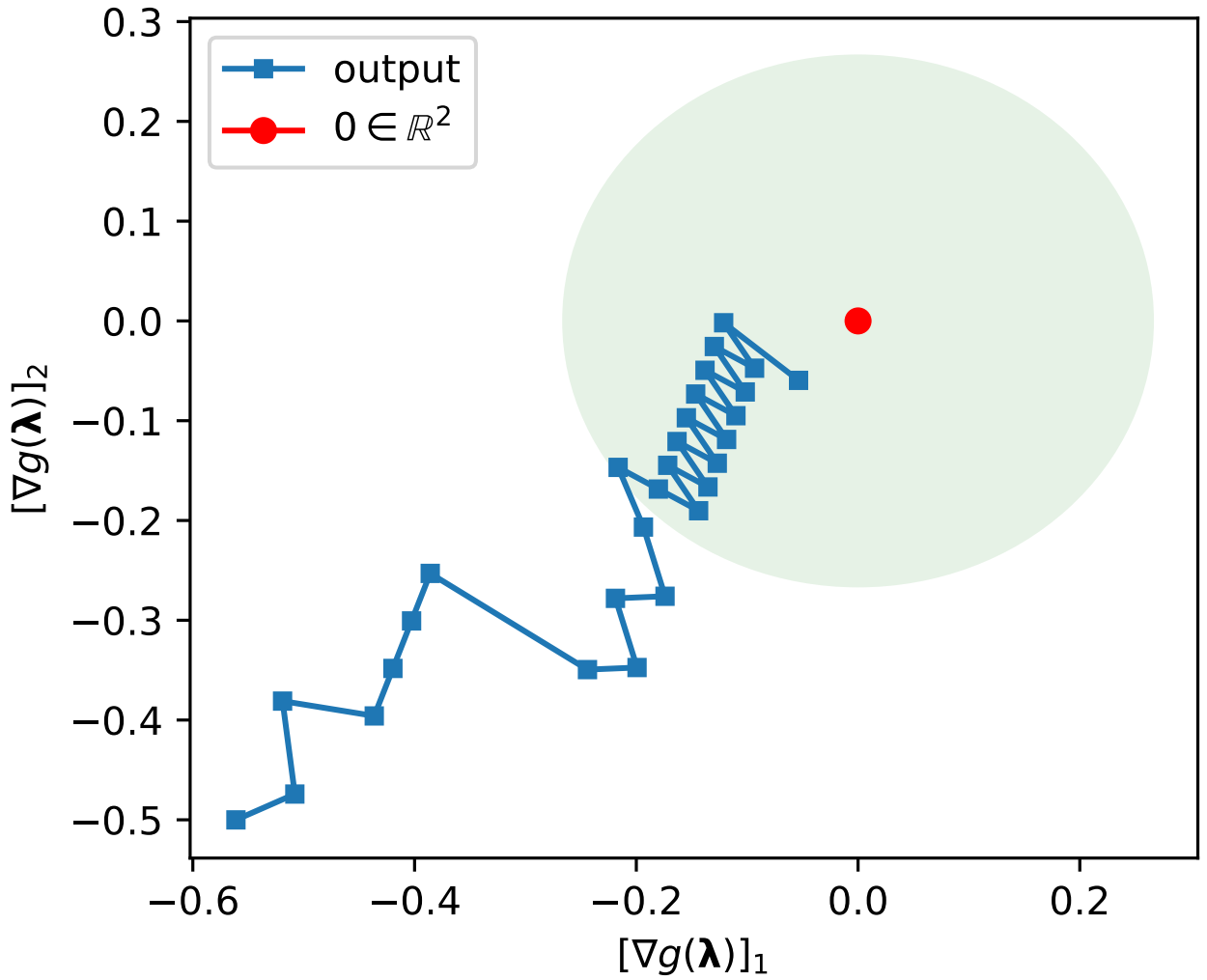
Suppose Assumption 1, Assumptions 2, and Assumption I hold. Let $\{\boldsymbol{\lambda}^{(k)}\}$ be the sequence of Lagrange multipliers generated. Moreover, suppose γ_k satisfy the nonsummable step size rule with $0 < \gamma_k \leq 1/L_h$, where $L_h = (1/\mu)\left(2 + 2\cos(\pi/m)\right)$ and $\mu = \min_i \ \mu_i$.Then

1. $\limsup_k \min_{i \in \{0,\ldots,k\}} \|\nabla g(\boldsymbol{\lambda}^{(i)})\| \leq \epsilon$.
2. for $\gamma_k = \gamma/(k+1)^p$, where $0 < \gamma \leq 1/L_h$ and $0 \leq p \leq 1$,

$$\min_{i \in \{0,\ldots,k\}} \|\nabla g(\boldsymbol{\lambda}^{(i)})\| = \begin{cases} O\left(\frac{1}{\sqrt{k^{1-p}}}\right) + \epsilon, & p \in [0,1) \\ O\left(\frac{1}{\sqrt{\log k}}\right) + \epsilon, & p = 1, \end{cases}$$

and the best convergence rate is of the order $O(1/\sqrt{k})$, which is achieved when $p = 0$, i.e., the fixed step size.

A Geometric View, $m = 3$, $n = 1$, Quantized with 5 bits



Main Results: Convergence in Primal Domain (Not Necessarily Feasible)

Proposition 1

Suppose Assumption 1, Assumptions 2, Assumption 3, and Assumption I hold. Let $\{\boldsymbol{\lambda}^{(k)}\}$ be the sequence of Lagrange multipliers and $\{\mathbf{y}^{(k)}\}$ be the corresponding sequence of primal variables generated. Moreover, suppose that the functions f_i, $i = 1, \ldots, m$ are differentiable. Let γ_k satisfy the nonsummable step size rule with $0 < \gamma_k \leq 1/L_h$. If the distance to the dual optimal solution $\|\boldsymbol{\lambda}^{(k)} - \boldsymbol{\lambda}^\star\|$ is uniformly bounded by some scalar D, then

1. $\limsup_k \min_{i \in \{0,\ldots,k\}} \|\mathbf{y}^{(i)} - \mathbf{y}^\star\| \leq \rho_1(\epsilon)$, where $\rho_1(\epsilon) \to 0$ as $\epsilon \to 0$..
2. $\limsup_k \min_{i \in \{0,\ldots,k\}} \left(f(\mathbf{y}^{(i)}) - f(\mathbf{y}^\star)\right) \leq \rho_2(\epsilon)$, where $\rho_2(\epsilon) \to 0$ as $\epsilon \to 0$.

Main Results: Convergence in Primal Domain (A Feasible Point)

Proposition 2

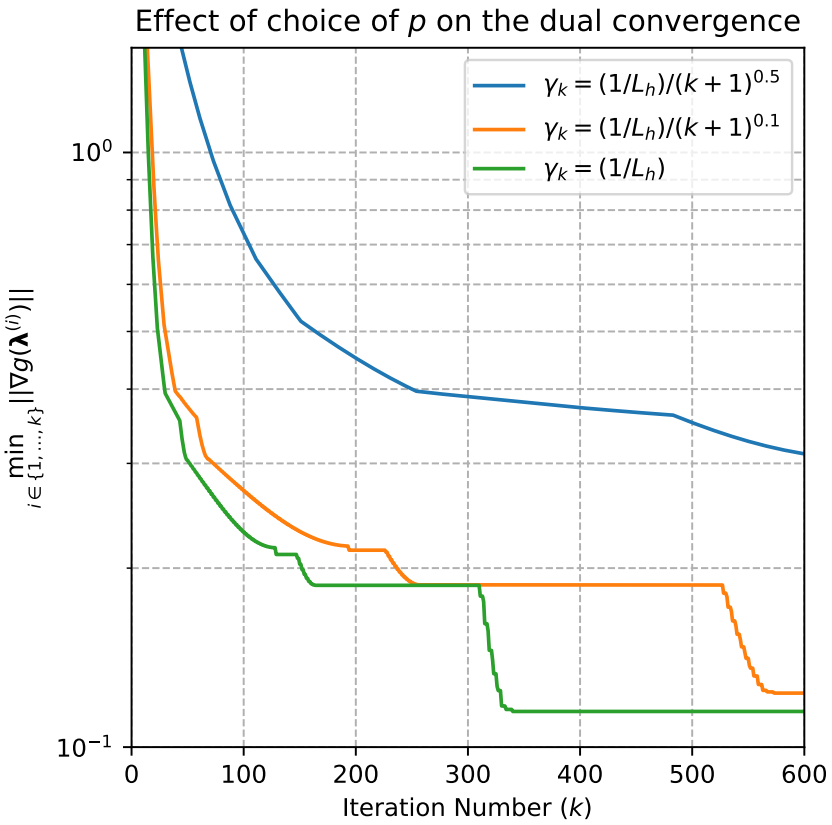
Suppose Assumption 1, Assumptions 2, Assumption 3, and Assumption I hold. Let $\{\boldsymbol{\lambda}^{(k)}\}$ be the sequence of Lagrange multipliers and $\{\tilde{\mathbf{y}}^{(k)}\}$ be the corresponding sequence of primal variables, and $\tilde{\mathbf{y}}^{(k)} = (1/m)\ (\mathbf{1}_{m\times m} \otimes \mathbf{I}_n)\ \mathbf{y}^{(k)}$ be the resulting feasible points. Moreover, suppose that the functions f_i, $i = 1, \ldots, m$ are differentiable and the distance to the dual optimal solution $\|\boldsymbol{\lambda}^{(k)} - \boldsymbol{\lambda}^{\star}\|$ is uniformly bounded by some D. Let γ_k satisfy the nonsummable step size rule with $0 < \gamma_k \leq 1/L_h$. Then

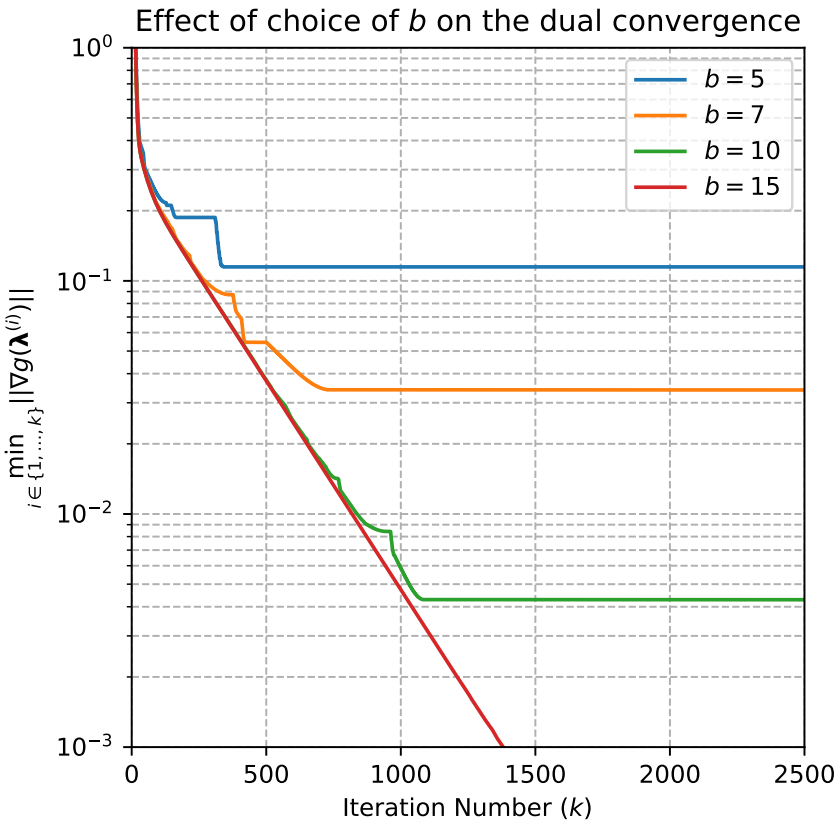
$$\limsup_{k} \min_{i\in\{0,\ldots,k\}} \|\tilde{\mathbf{y}}^{(i)} - \mathbf{y}^{\star}\| \leq \rho_3(\epsilon),$$

where $\rho_3(\epsilon) \to 0$ as $\epsilon \to 0$.

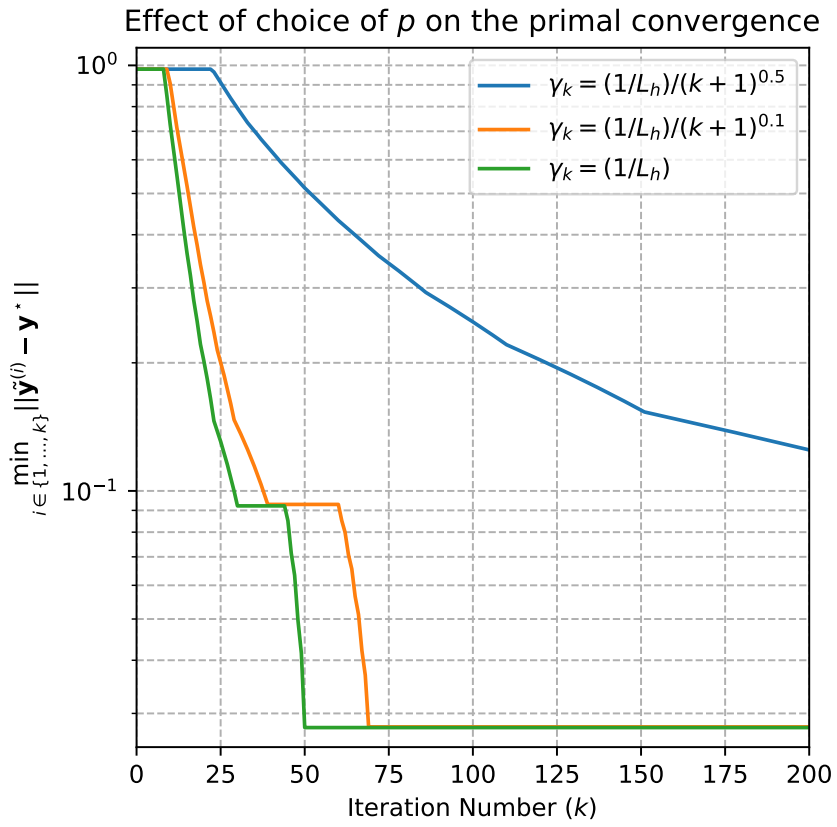
Numerical Experiments

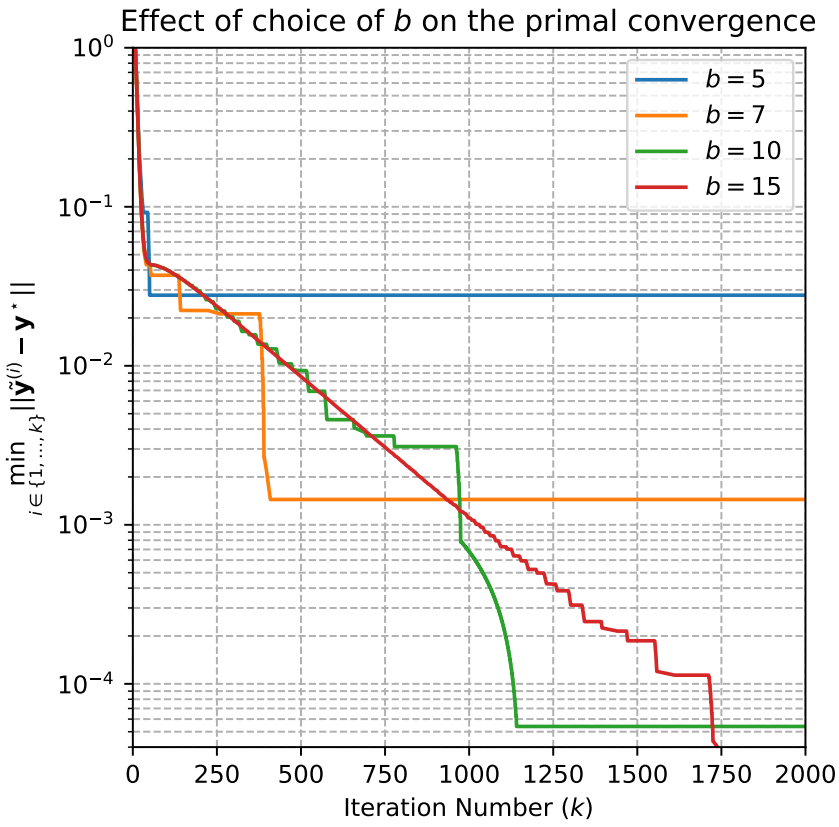
Effect of p and Quantization (b) on Dual Convergence





Effect of p and Quantization (b) on Primal Convergence





Conclusive Remarks

- benefits of distributed machine learning (ML) algorithms
  - to pervade every stratum of society
  - an appeal to investigation of how
    - the existing/future wireless systems & the learning processes?
    - many imperfect settings
    - what are the consequences/limitations of those?
  - ML over wireless mobile telephony systems
  - designing wireless systems for ML
  - an overwhelmingly appealing and timely area of research

Thank You

