LASER: LineAr CompreSsion in WirEless DistRibuted Optimization

Ashok Vardhan Makkuva (EPFL)

Joint work with Marco Bondaschi, Thijs Vogels, Martin Jaggi, Hyeji Kim, Michael Gastpar

Outline

Outline

- Motivation

Outline

- Motivation

- Problem setting

Outline

- Motivation
- Problem setting
- Contribution: LASER

Outline

- Motivation
- Problem setting
- Contribution: LASER
- Future directions

Outline

- Motivation

Obligatory slide

ChatGPT

Obligatory slide

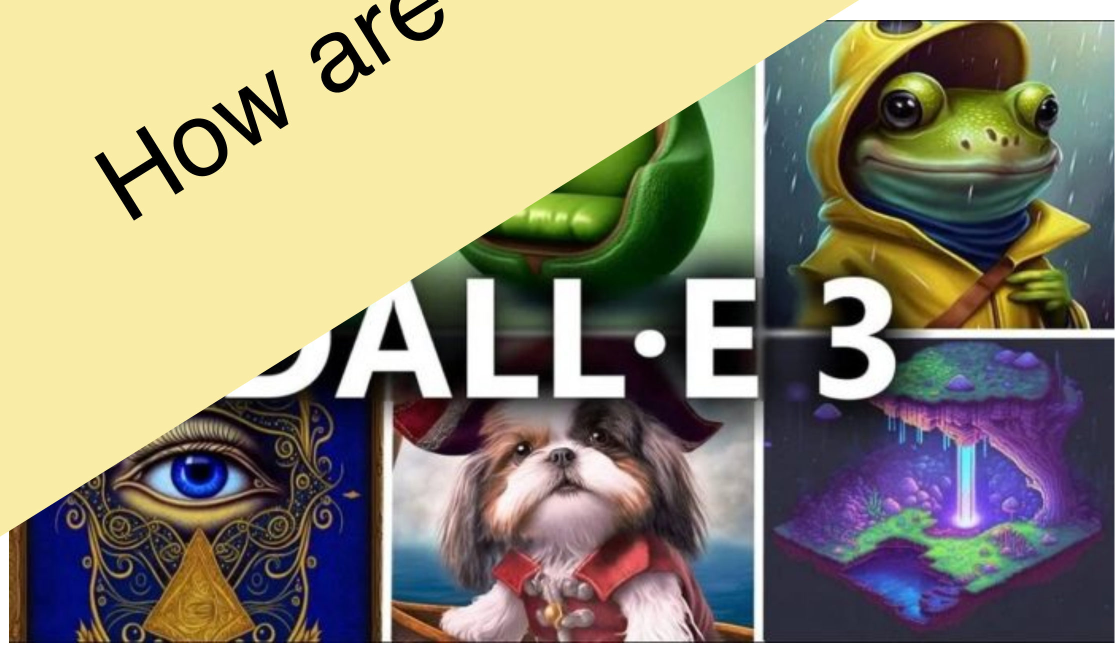
ChatGPT

GPT-4

Obligatory slide

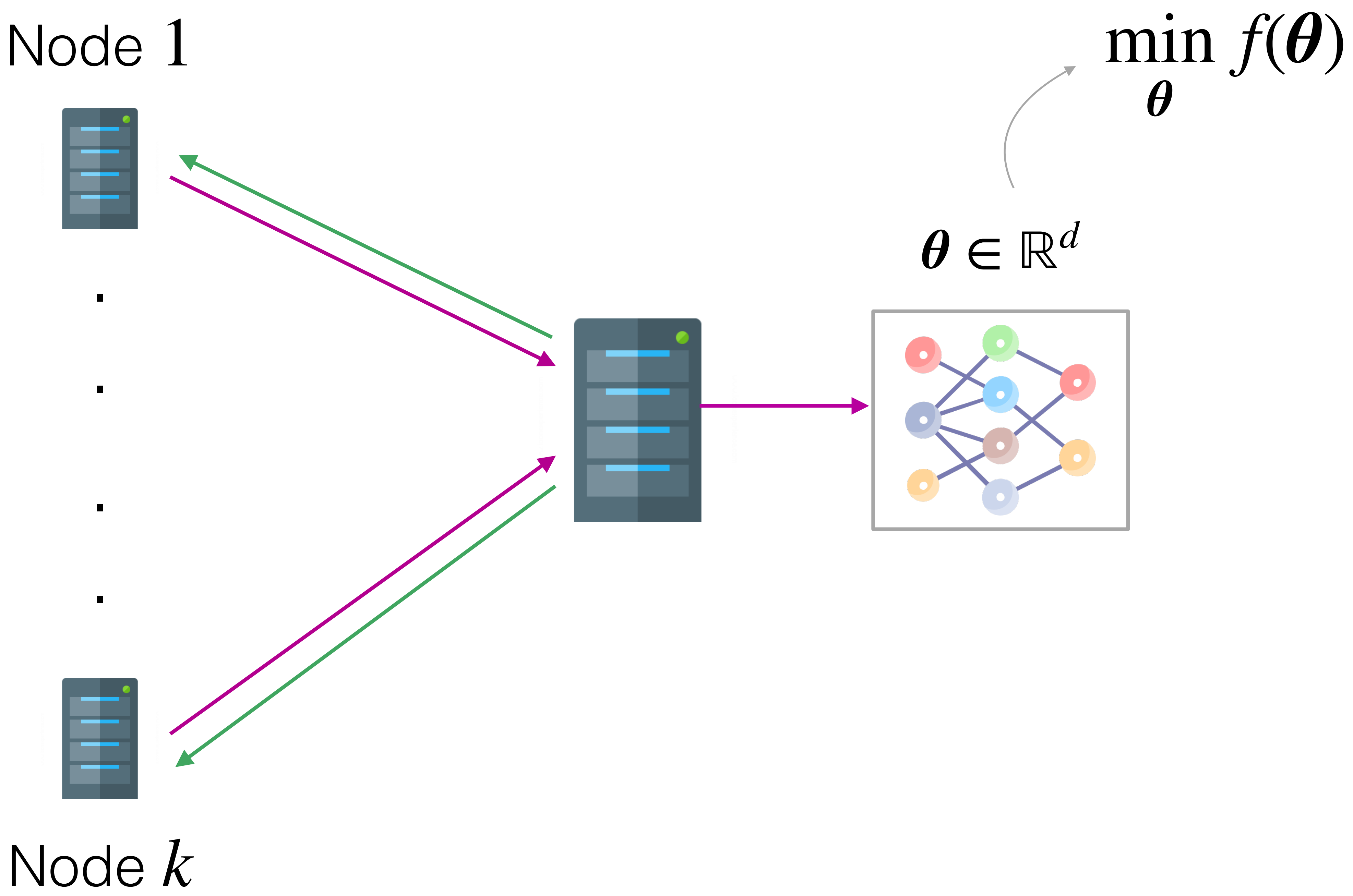
ChatGPT

How are they trained?

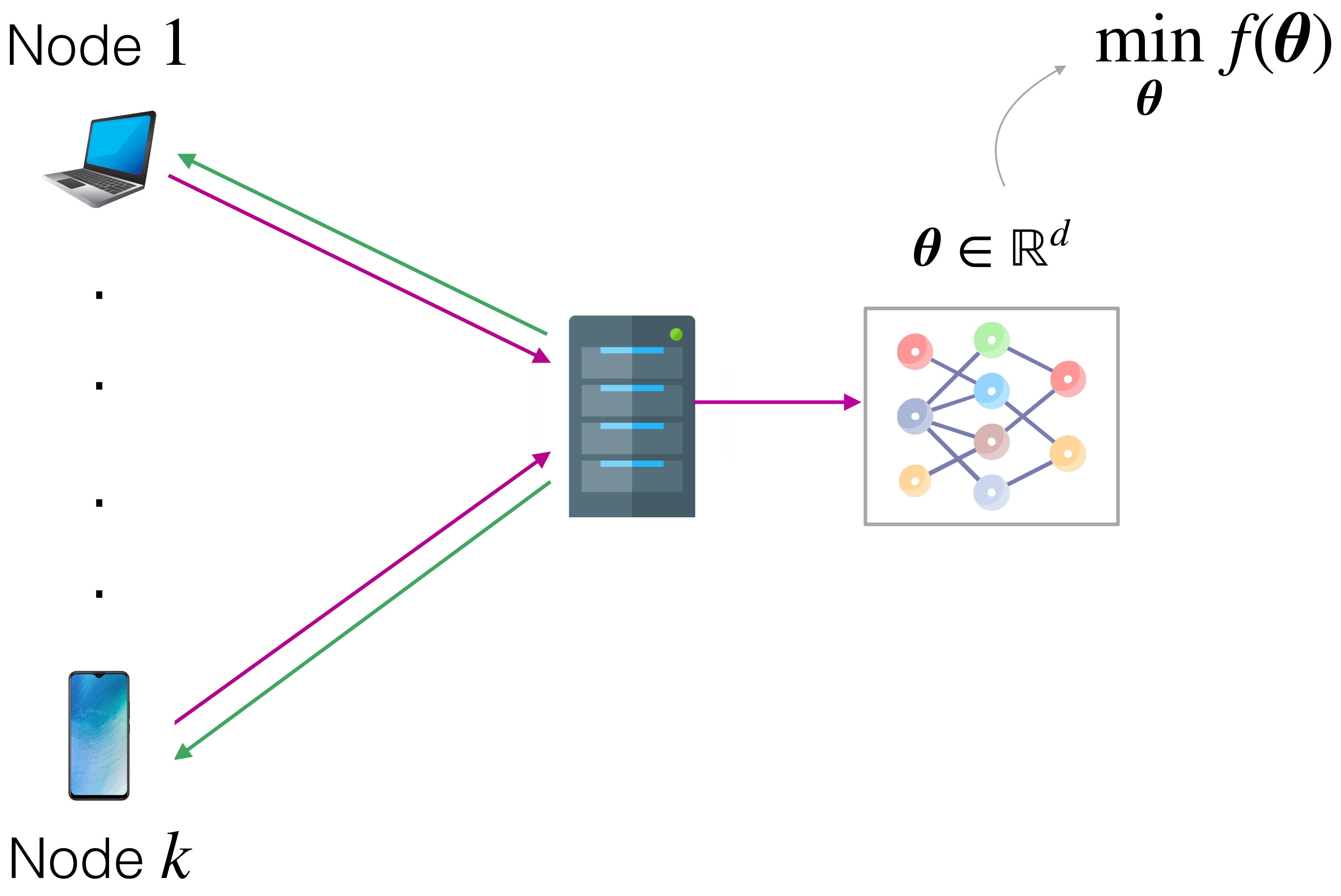


Distributed Optimization

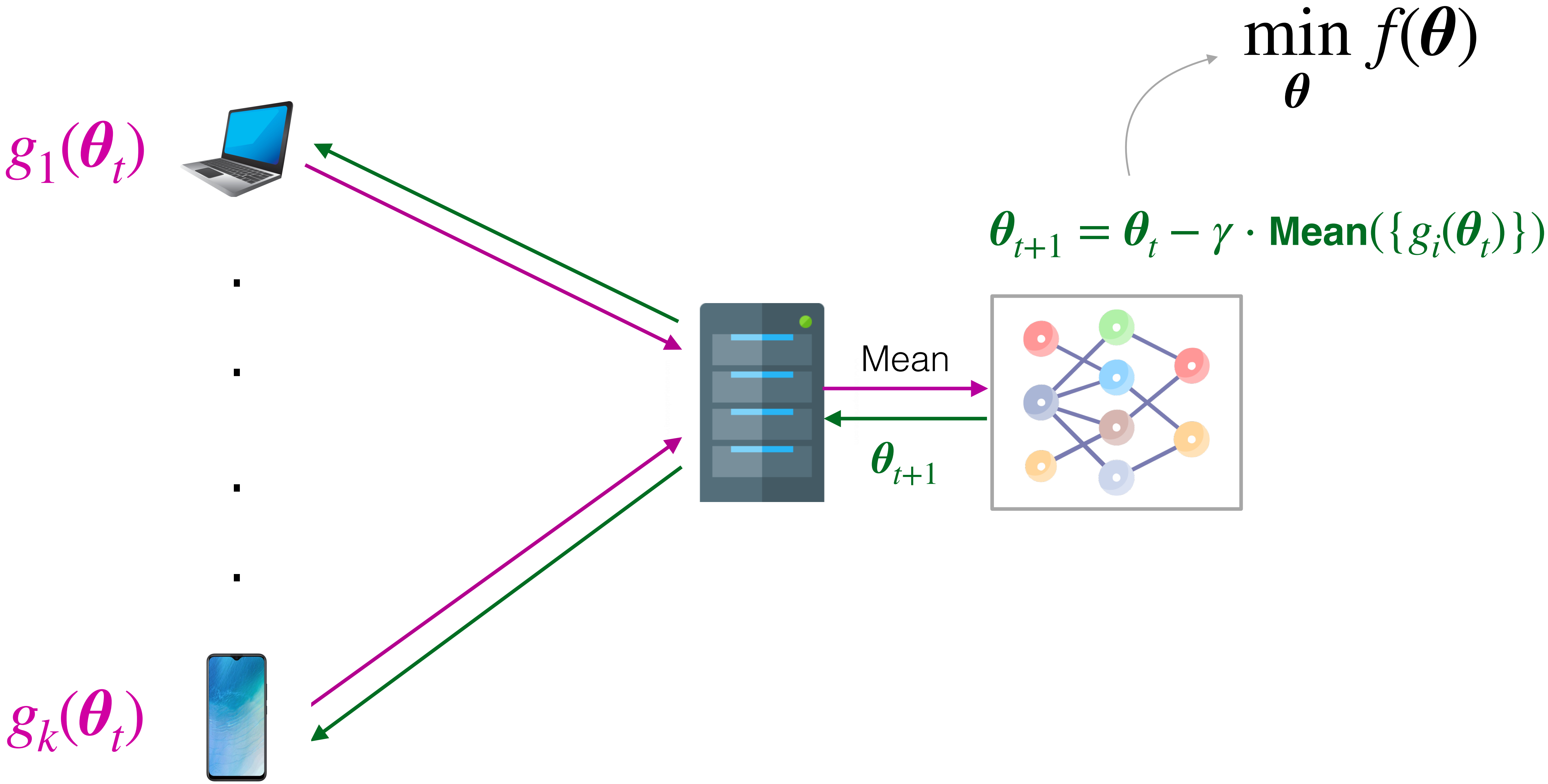
Distributed Optimization



Distributed Optimization



Distributed SGD



SGD: assumptions

SGD: assumptions

- Noiseless communication links

SGD: assumptions

- Noiseless communication links
  - Data center

SGD: assumptions

- Noiseless communication links
  - Data center
  - Federated learning

SGD: assumptions

- Noiseless communication links
  - Data center
  - Federated learning

- Federated learning

SGD: assumptions

- Noiseless communication links
  - Data center
  - Federated learning

- Federated learning
  - Noisy links: Error-correcting codes

SGD: assumptions

- Noiseless communication links
  - Data center
  - Federated learning

- Federated learning
  - Noisy links: Error-correcting codes
  - Low-latency: server should decode each client to compute mean

SGD: assumptions

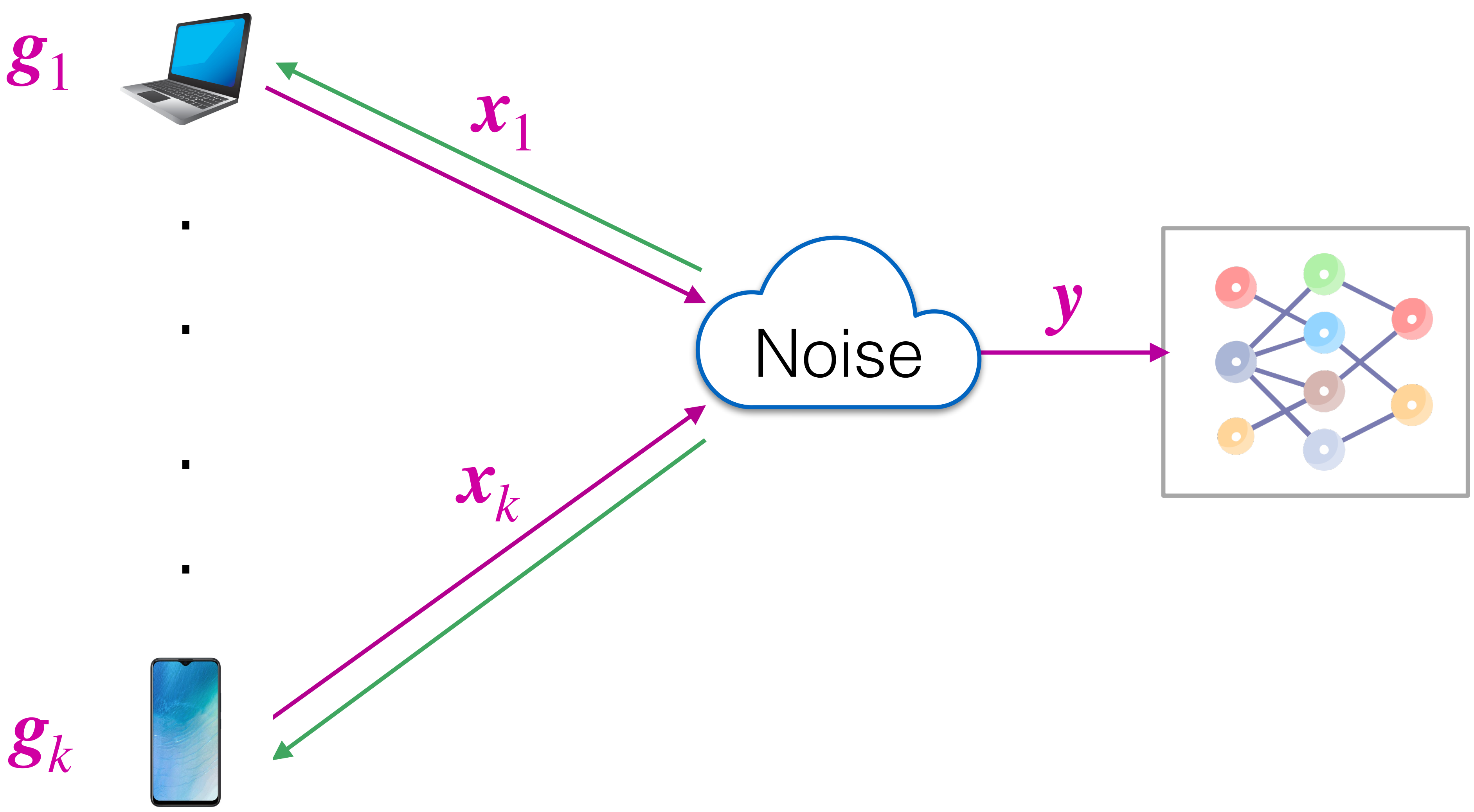
- Noiseless communication links
  - Data center
  - Federated learning

- Federated learning
  - Noisy links: Error-correcting codes
  - Low-latency: server should decode each client to compute mean

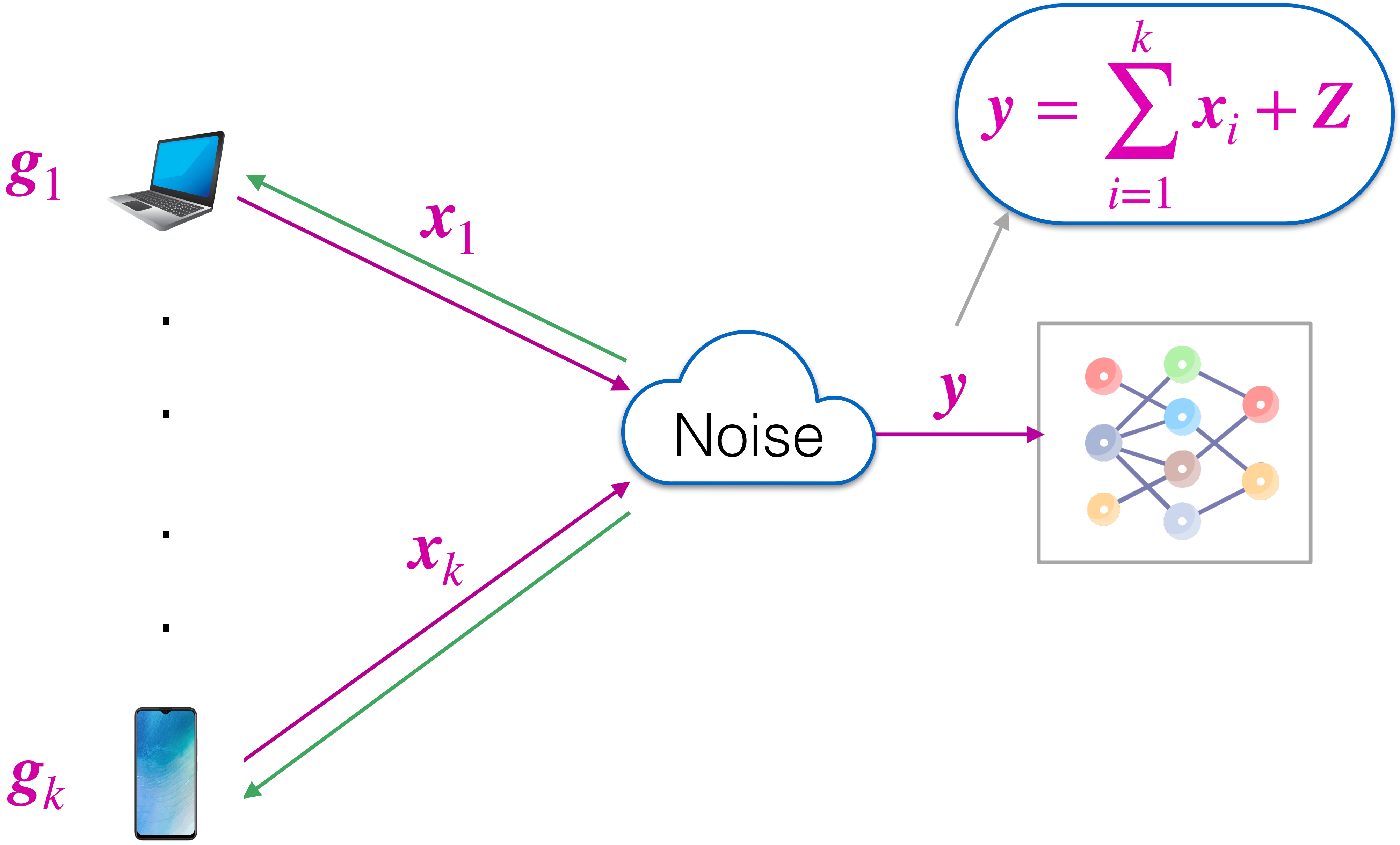
- Can we tame the noise directly?

Noise

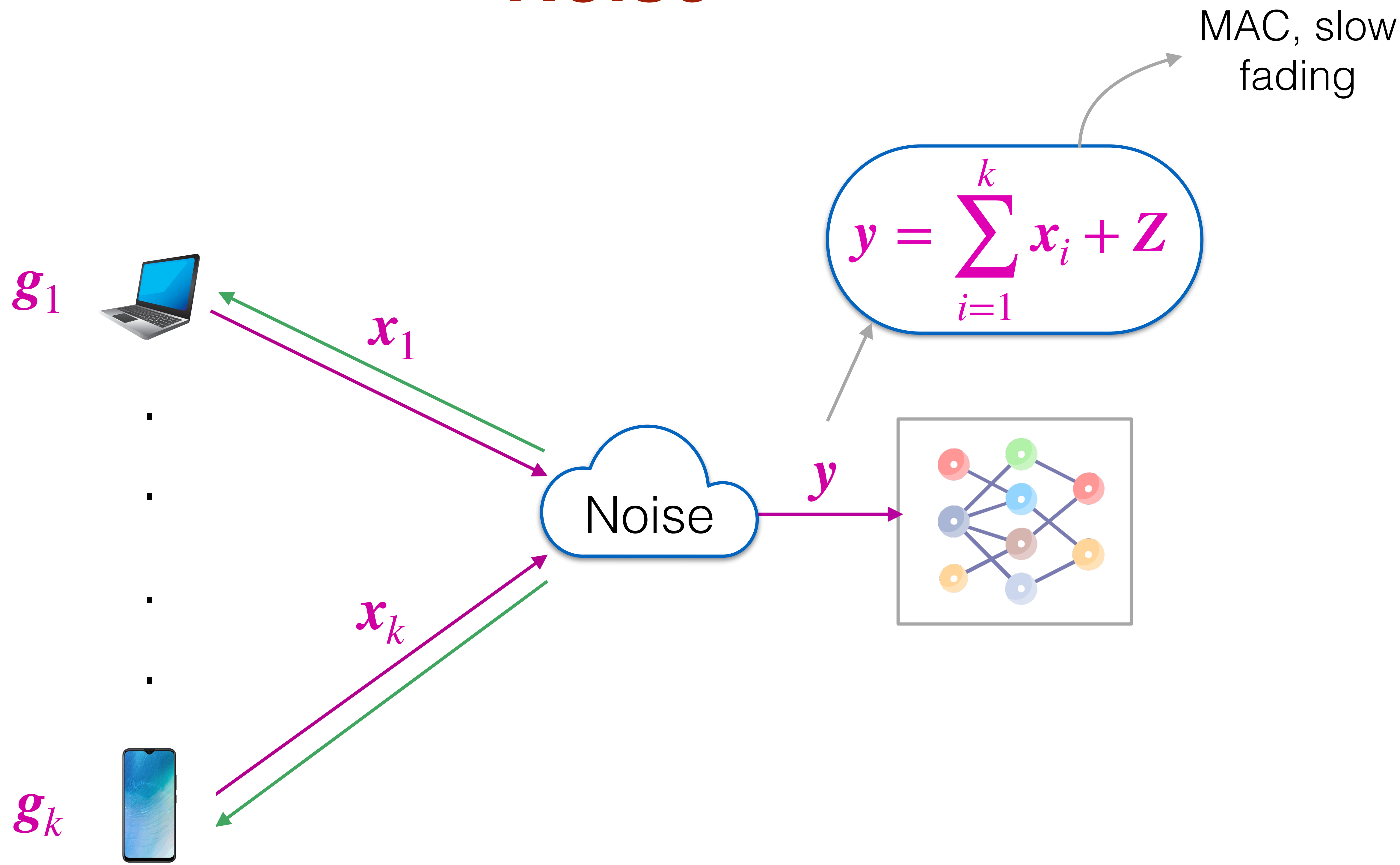
Noise



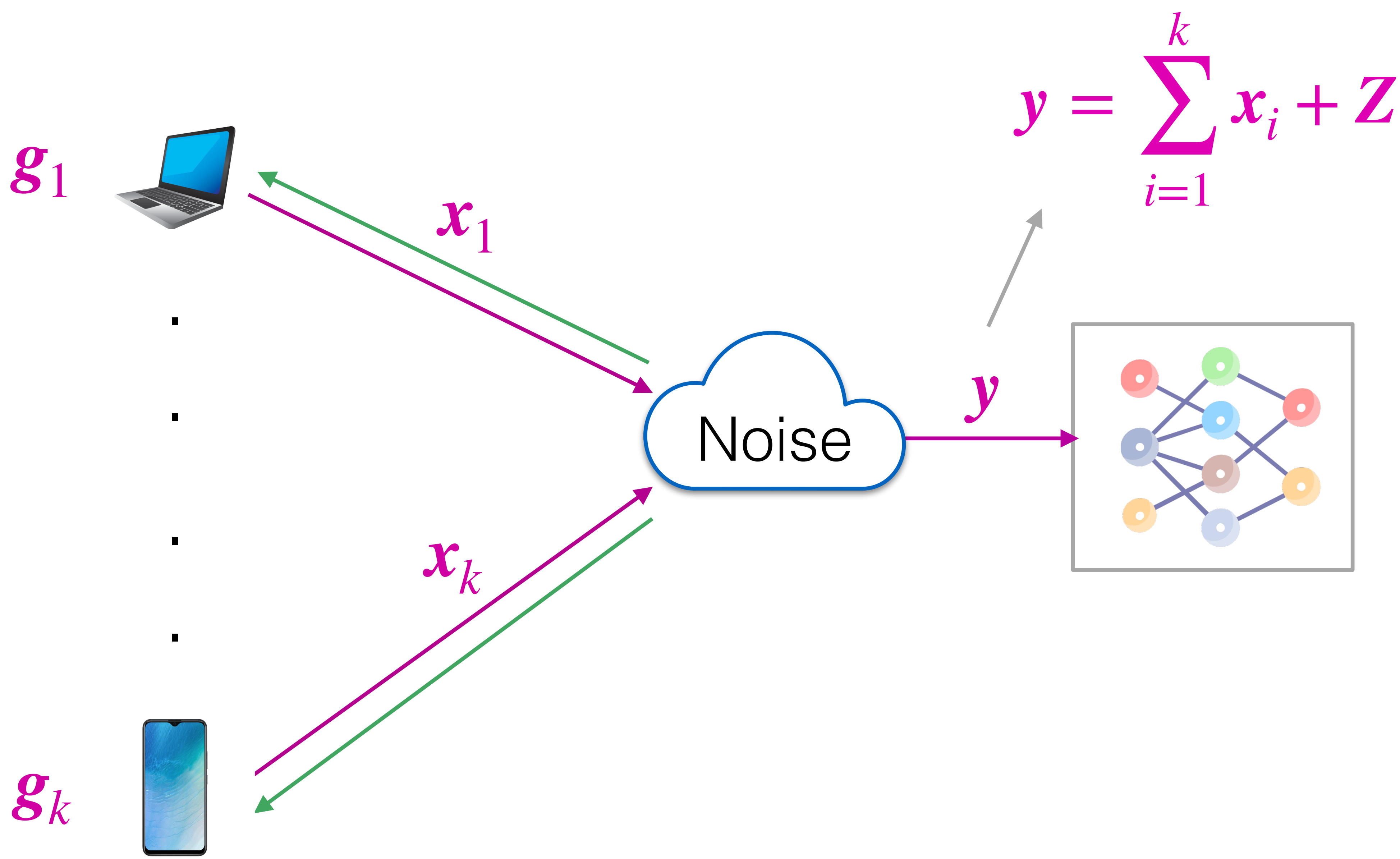
Noise



Noise

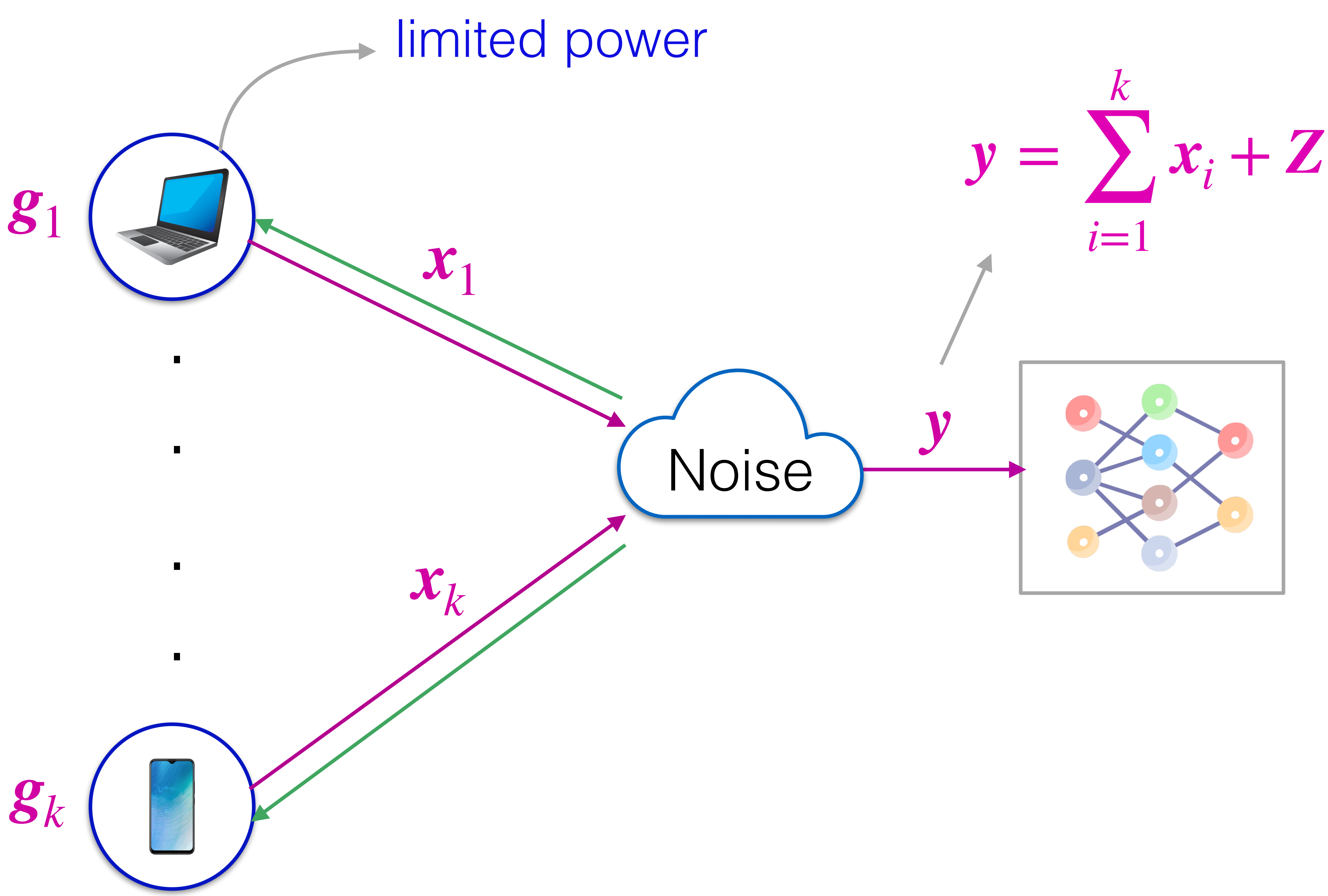


Noise

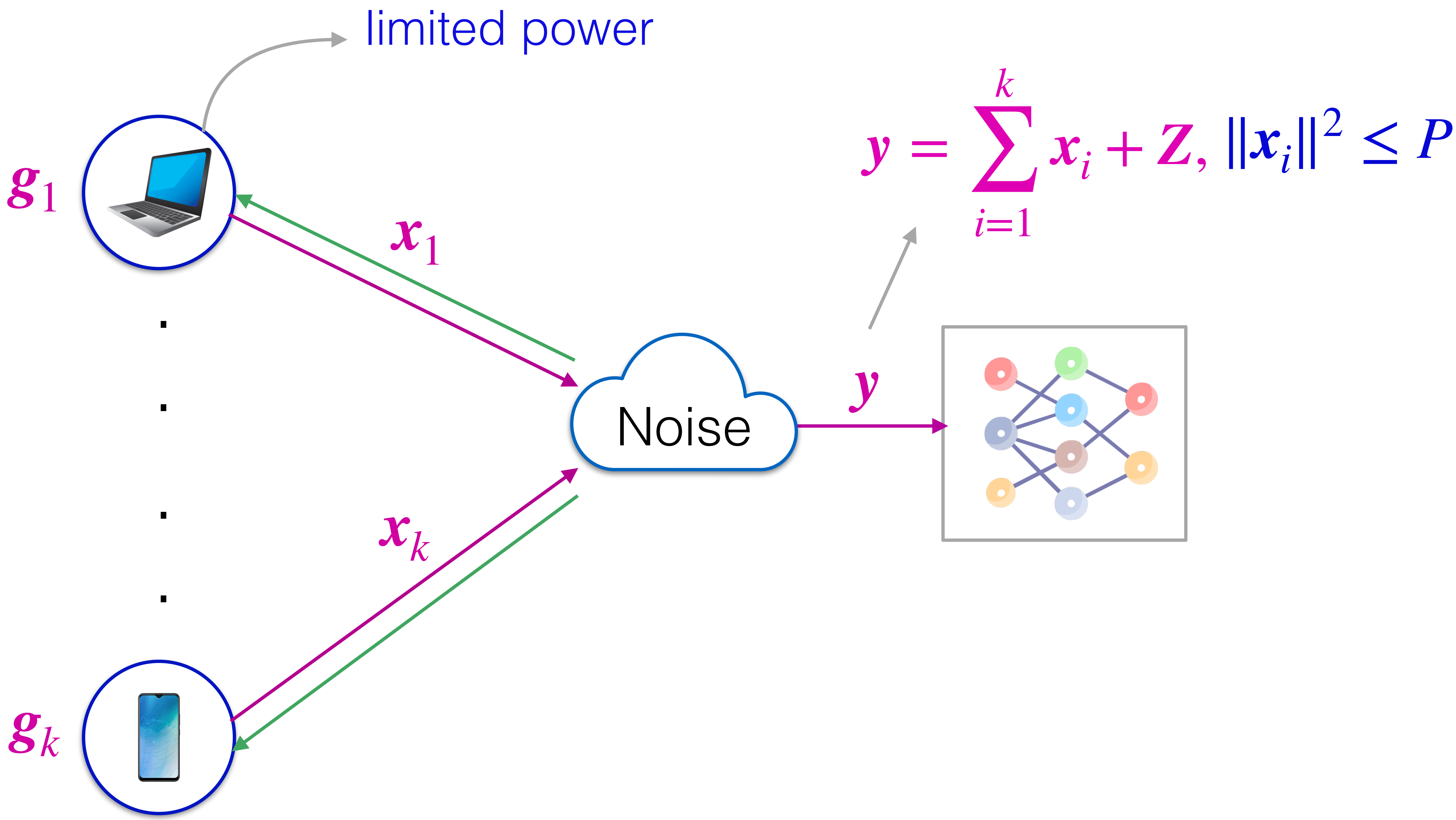


$\boldsymbol{g}_1$
$\boldsymbol{x}_1$
$\boldsymbol{x}_k$
$\boldsymbol{g}_k$
Noise
$\boldsymbol{y}$
$\boldsymbol{y} = \sum_{i=1}^{k} \boldsymbol{x}_i + \boldsymbol{Z}$

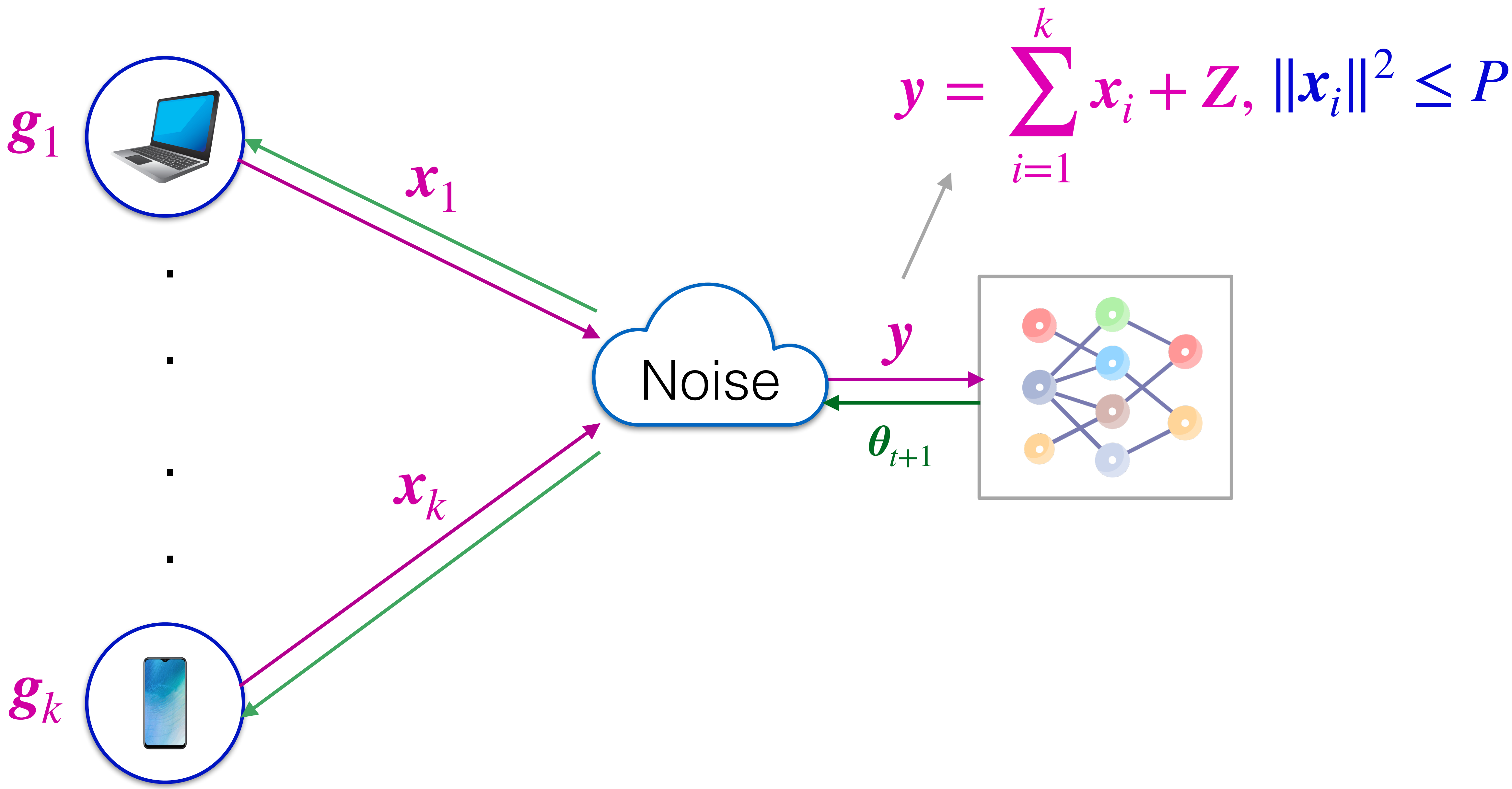
Noise



Noise + Power constraint



Wireless distributed optimization

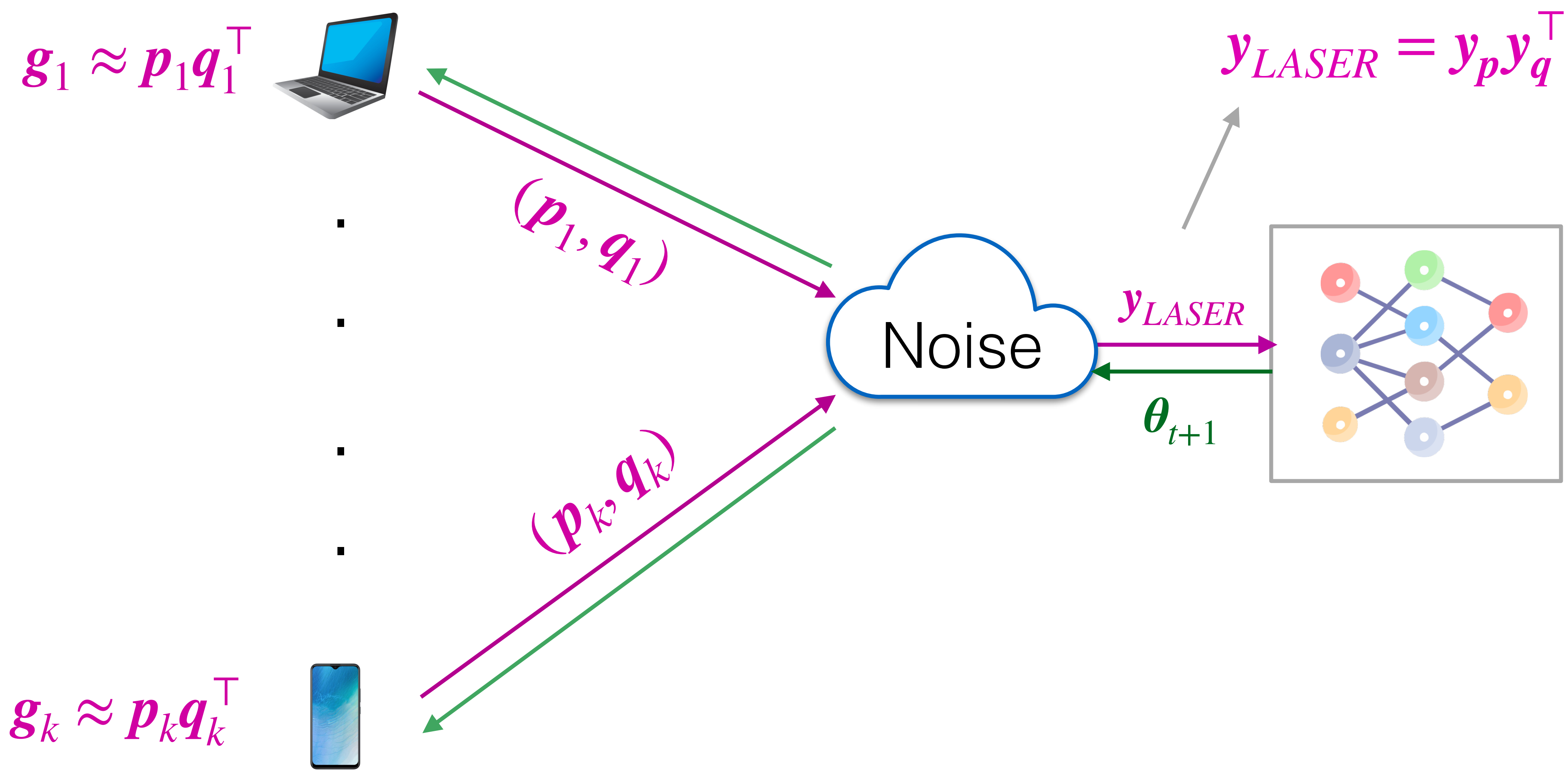


Main question

Can we design reliable and efficient training algorithms for wireless distributed optimization?

LASER

LASER



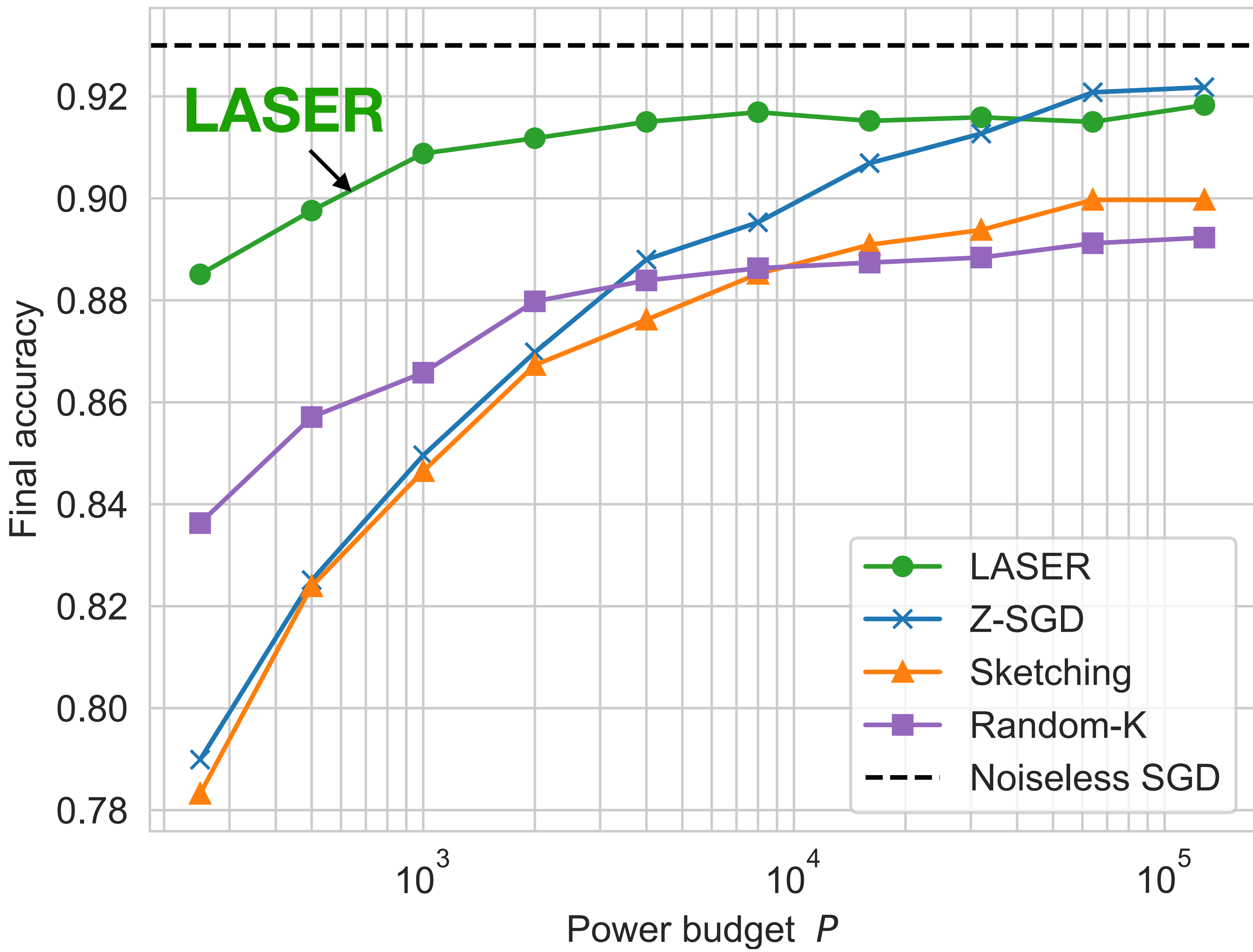
$g_1 \approx p_1 q_1^\top$
(p_1, q_1)
$g_k \approx p_k q_k^\top$
(p_k, q_k)
Noise
y_{LASER}
θ_{t+1}
$y_{LASER} = y_p y_q^\top$

Results

- CIFAR-10, ResNet-18 (11M params), 16 nodes

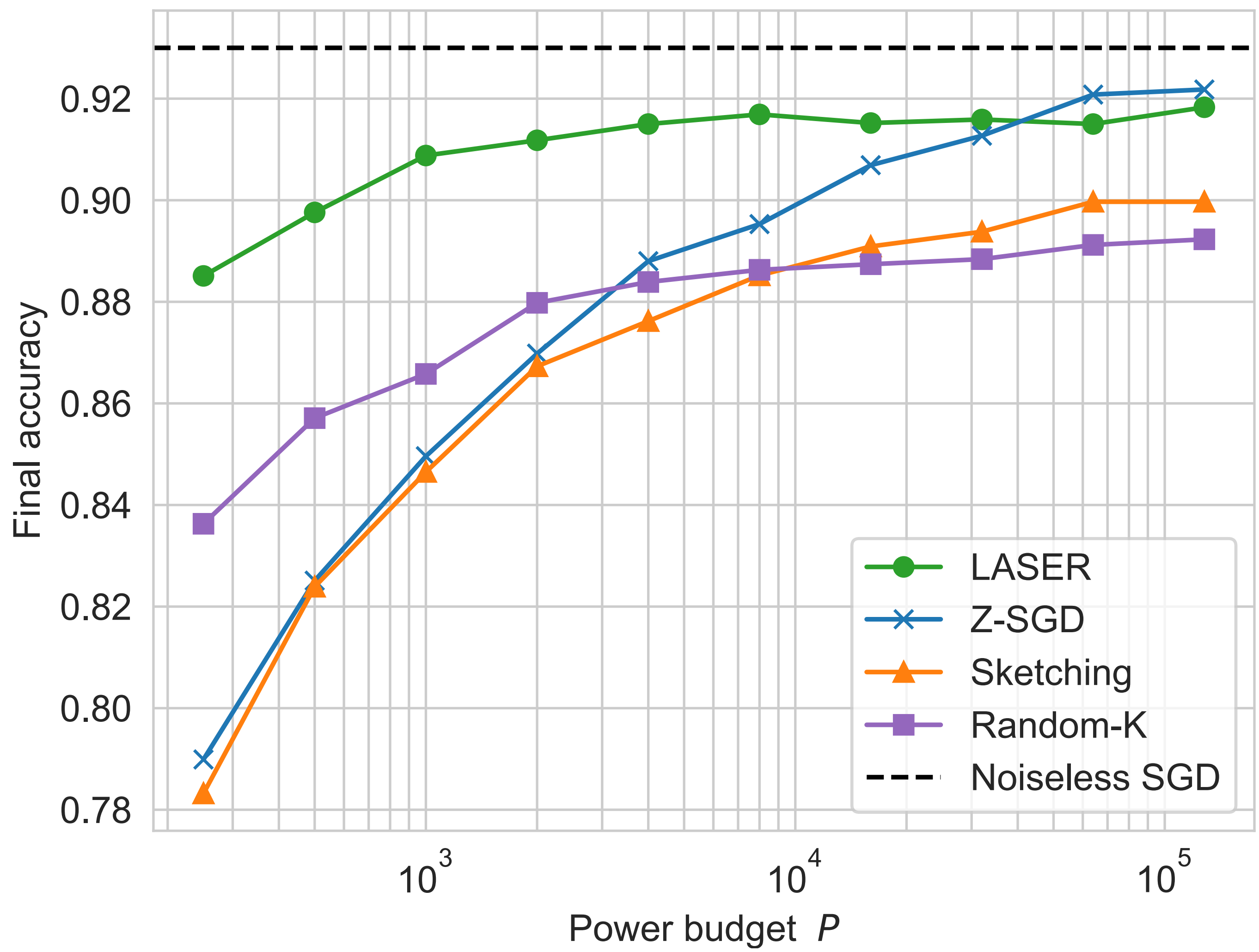
- WikiText-103, GPT-2 (123M params), 4 nodes

CIFAR-10



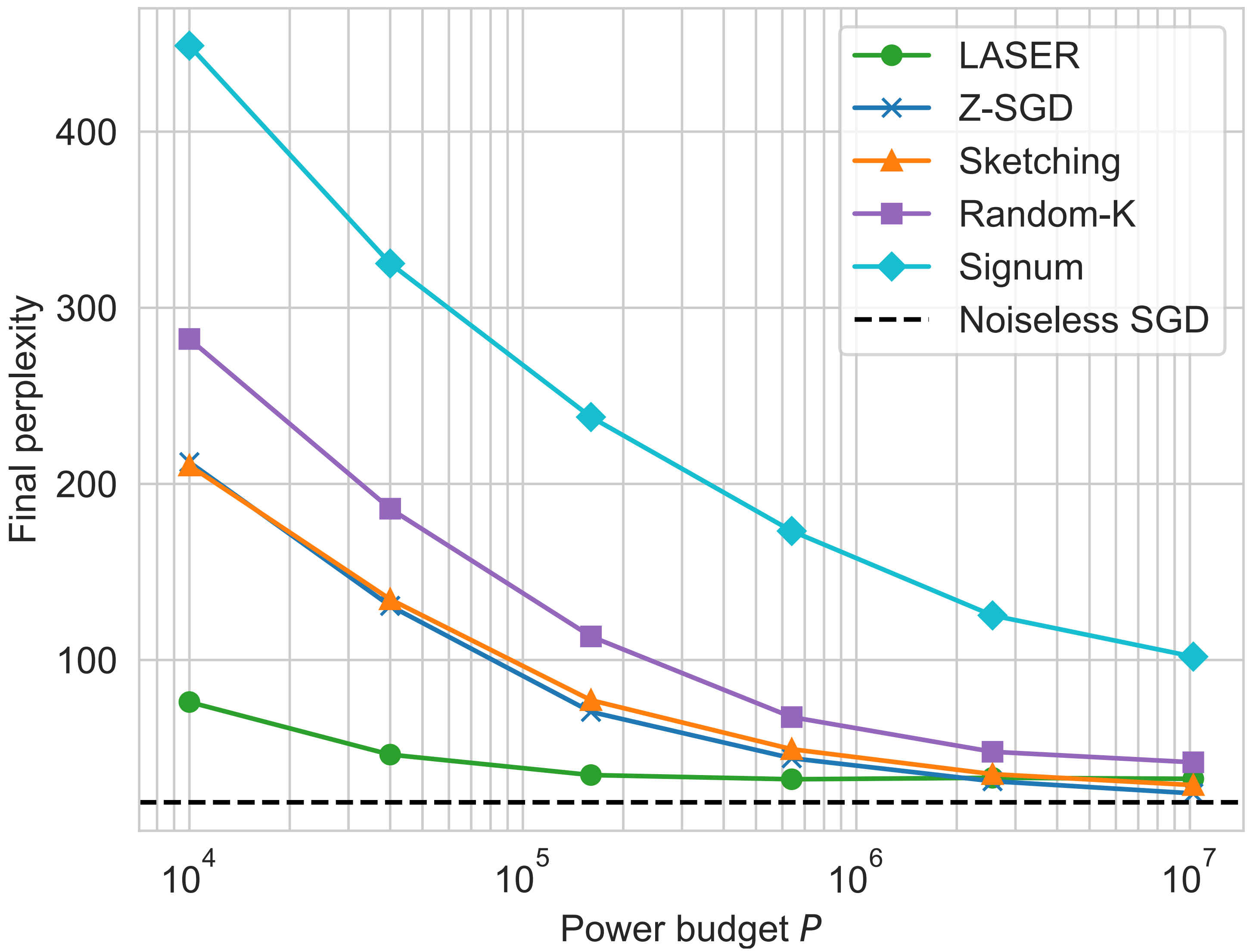
LASER
0.92
0.90
0.88
0.86
0.84
0.82
0.80
0.78
Final accuracy
10^3
10^4
10^5
Power budget P
LASER
Z-SGD
Sketching
Random-K
Noiseless SGD

CIFAR-10



Target	Power required		Reduction
	LASER	Z-SGD	
88%	250	4000	16×
89%	500	8000	16×
90%	1000	16000	16×
91%	2000	32000	16×

GPT language modeling



Target	**Power required**		**Reduction**
	Z-SGD	LASER	
80	160 K	10 K	16×
50	640 K	40 K	16×
40	2560 K	160 K	16×
35	2560 K	160 K	16×

Communication cost

Algorithm	Data sent per iteration	
Z-SGD	496 MB	(1×)
Signum	15 MB	(33×)
Random-K	99 MB	(5×)
Sketching	99 MB	(5×)
A-DSGD	n/a	n/a
LASER	**3 MB**	(165×)

Main theoretical result

Main theoretical result

- Under some standard assumptions, with $f_* = \min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$:
  - f is quasi-convex :

$$\mathbb{E} f(\boldsymbol{\theta}_{out}) - f_* = \tilde{O}\left(\frac{1 + \lambda_{LASER}}{T}\right)$$

  - f is non-convex :

$$\mathbb{E} \|\nabla f(\boldsymbol{\theta}_{out})\|^2 = \tilde{O}\left(\sqrt{\frac{1 + \lambda_{LASER}}{T}}\right)$$

Main theoretical result

- Under some standard assumptions, with $f_* = \min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$:
  - f is quasi-convex :

$$\mathbb{E}f(\boldsymbol{\theta}_{out}) - f_* = \tilde{O}\left(\frac{1 + \lambda_{LASER}}{T}\right)$$

  - f is non-convex :

$$\mathbb{E}\|\nabla f(\boldsymbol{\theta}_{out})\|^2 = \tilde{O}\left(\sqrt{\frac{1 + \lambda_{LASER}}{T}}\right)$$

$$\boxed{\lambda_{LASER} = \frac{4}{m \cdot SNR}\left(1 + \frac{1}{n \cdot SNR}\right)}$$

Main theoretical result

- Under some standard assumptions, with $f_* = \min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$:
  - f is quasi-convex :

$$\mathbb{E} f(\boldsymbol{\theta}_{out}) - f_* = \tilde{O}\left(\frac{1 + \lambda_{LASER}}{T}\right)$$

  - f is non-convex :

$$\mathbb{E}\|\nabla f(\boldsymbol{\theta}_{out})\|^2 = \tilde{O}\left(\sqrt{\frac{1 + \lambda_{LASER}}{T}}\right)$$

$$\boxed{\lambda_{LASER} \leq O\left(\frac{1}{m}\right) \lambda_{SGD}}$$

Conclusion

- Leverage channel and gradient structure: LASER

Conclusion

- Leverage channel and gradient structure: LASER
- Future steps: downlink noise, heterogenous nodes

La Fin

Thank you!

On the academic job market!


Any
Question