Federated Learning Over Noisy Channel

Xizixiang Wei

[C] Xizixiang Wei, Cong Shen, Federated Learning in the Presence of Communication Errors, in Proc. IEEE International Conference on Communications (ICC), June 2021.

[J] Xizixiang Wei and Cong Shen, Federated Learning over Noisy Channels: Convergence Analysis and Design Examples, IEEE Transactions on Cognitive Communications and Networking, vol. 8, no. 2, pp. 1253-1268, June 2022.

Background: Federated Learning

1

Machine learning enables powerful applications

2

Massive real-word data are generated at edge devices

3

We want to keep sensitive data at edge devices

Federated Learning (FL)

- A distributed machine learning paradigm
- Obtain a global model while keeping data locally

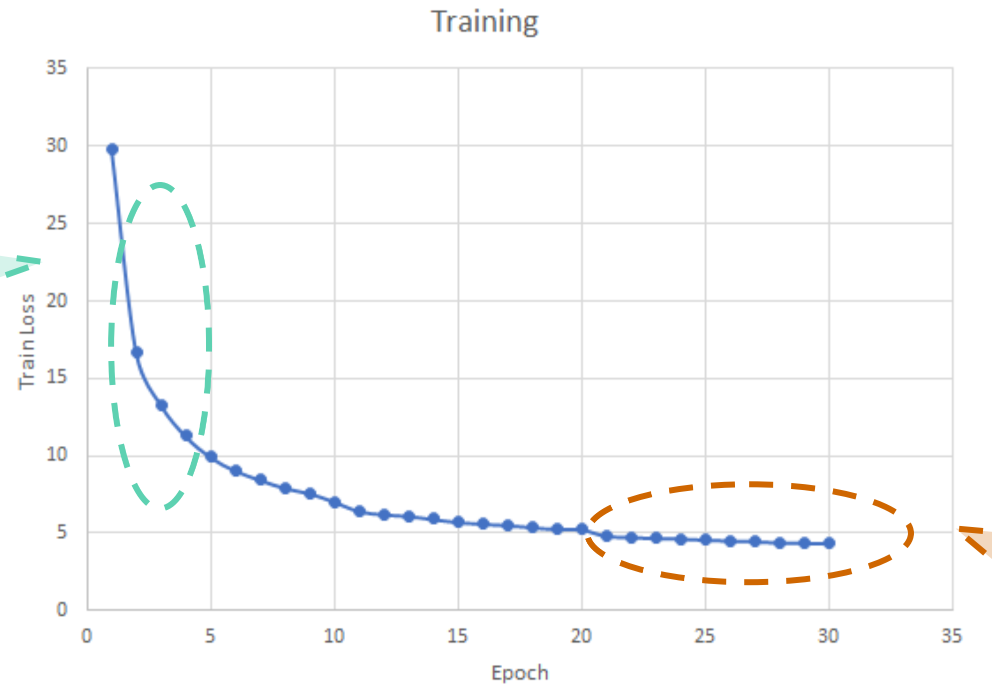
Motivation

Main difference from traditional communications

- **Different** quality-of-service (QoS) requirements over time

Early stage

Machine learning model is rough

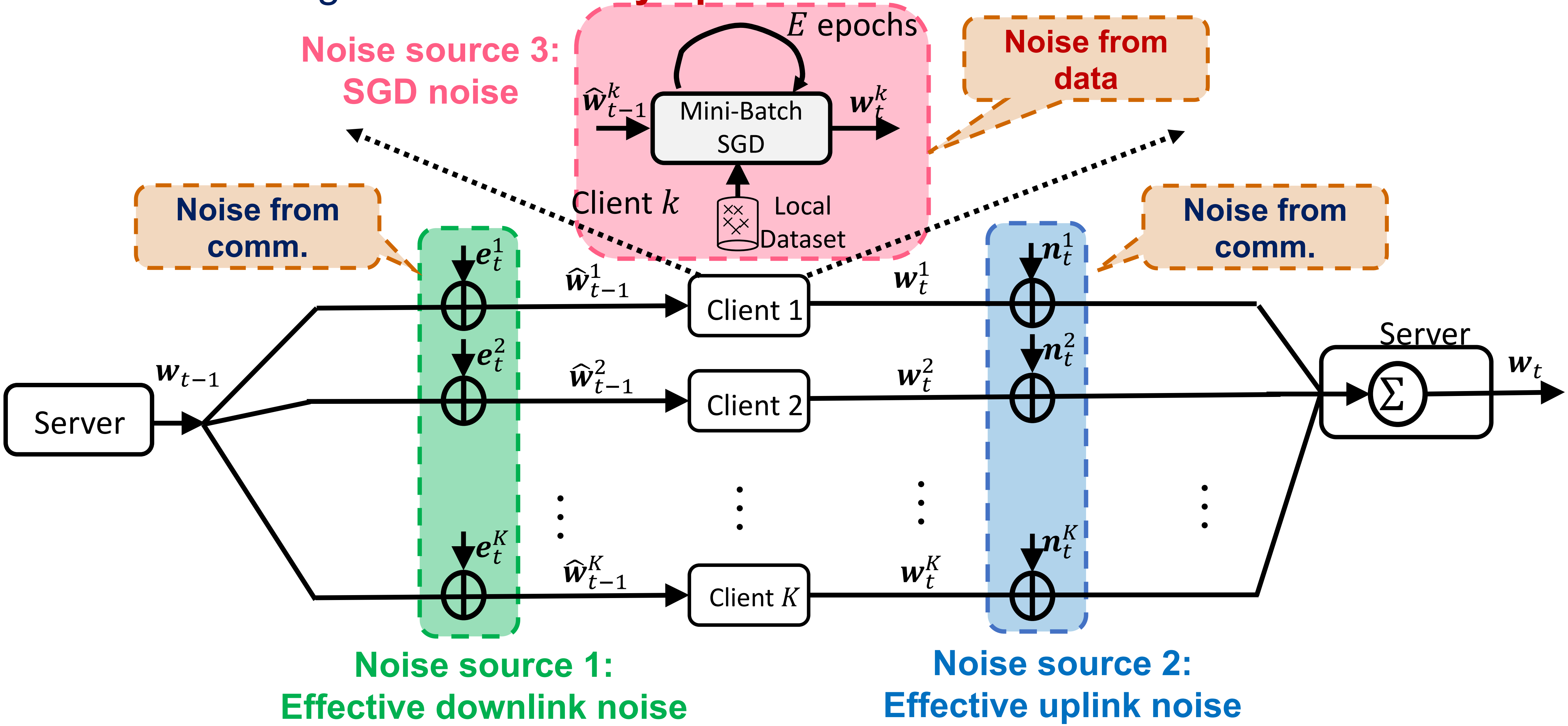


Final stage

Machine learning model is accurate

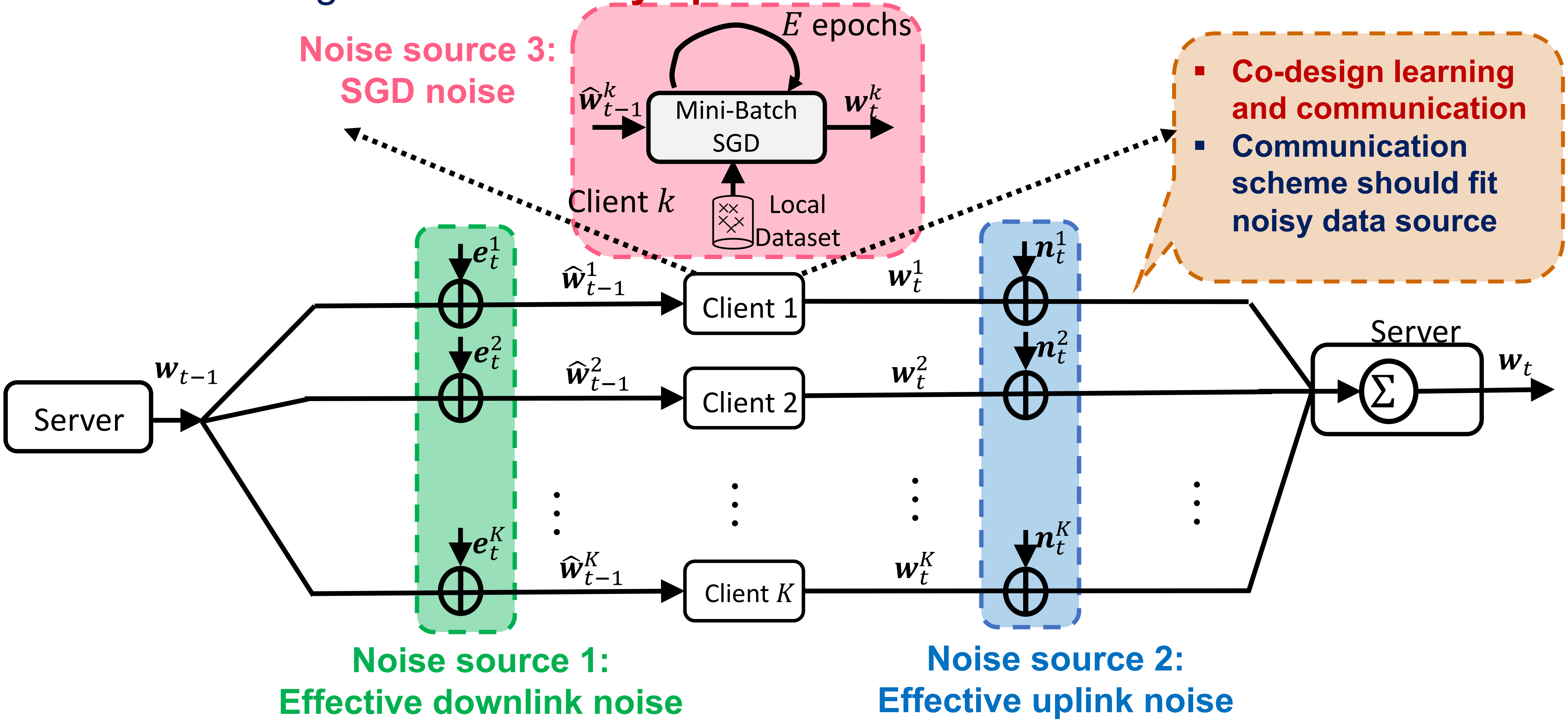
System model

Federated learning over **both noisy uplink and downlink** channels



System model

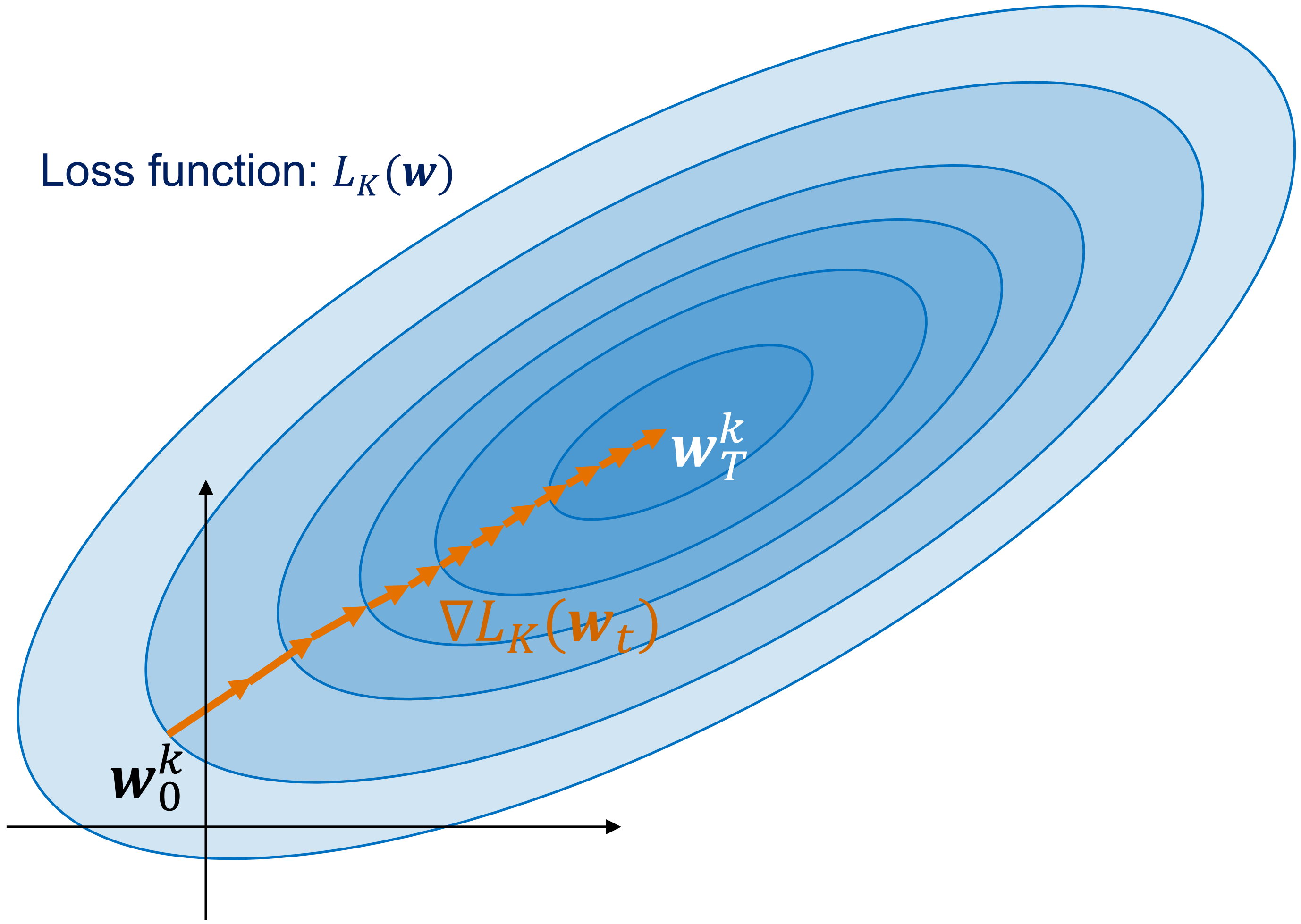
Federated learning over **both noisy uplink and downlink** channels



- **Co-design learning and communication**
- **Communication scheme should fit noisy data source**

SGD noise

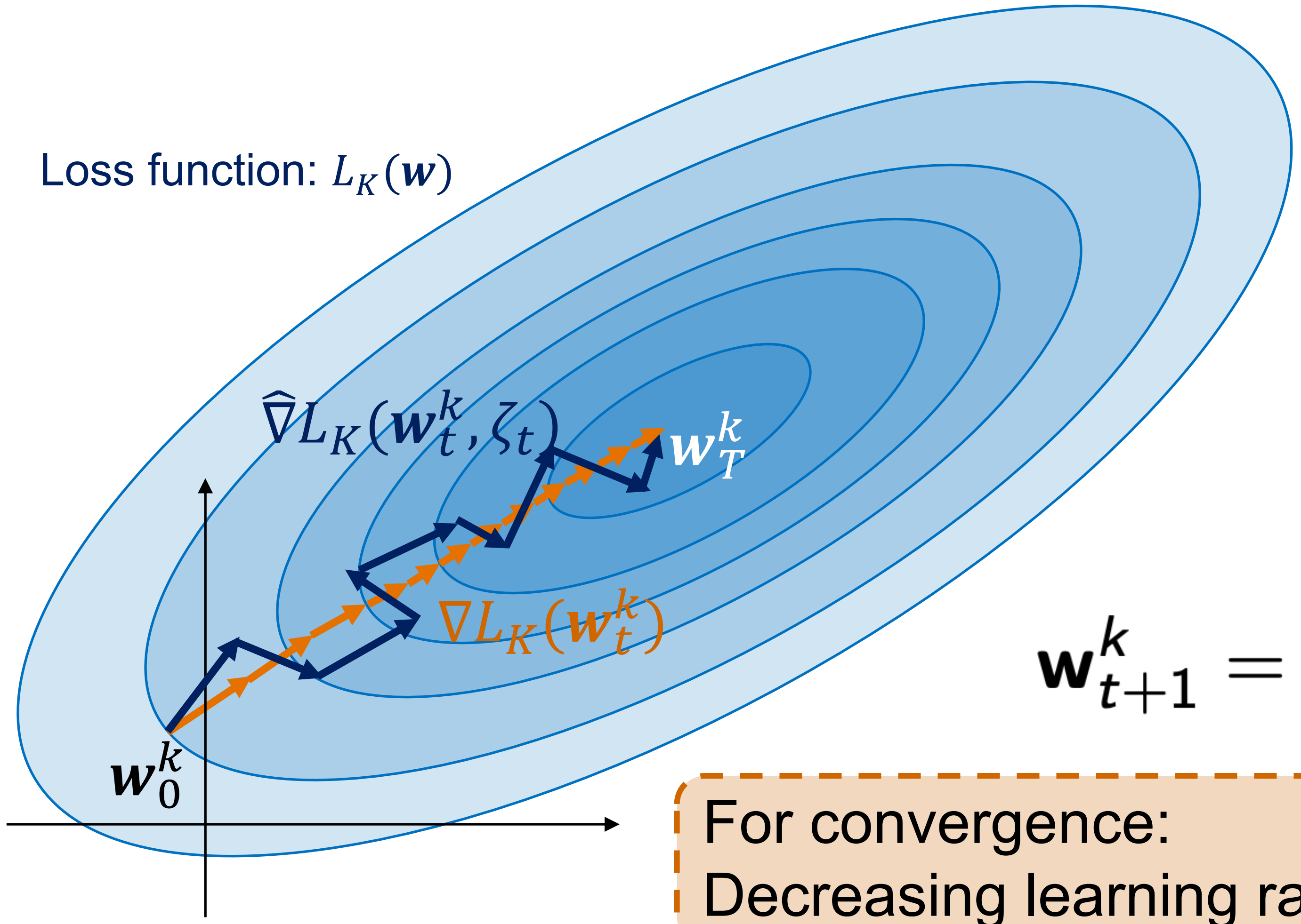
Gradient descent (GD)



SGD noise

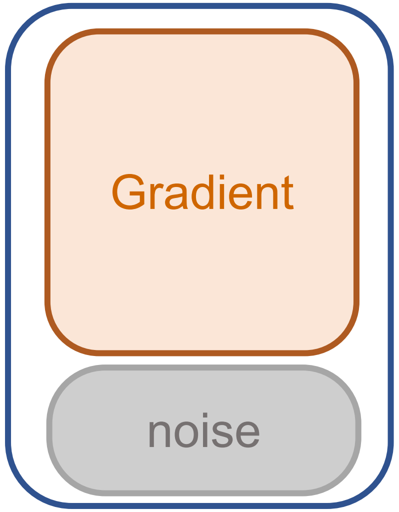
Gradient descent (GD)

Stochastic gradient descent (SGD)



$$\nabla L(\mathbf{w}_t^k) = \mathbb{E}[\hat{\nabla} L(\mathbf{w}_t^k, \zeta_t)]$$

stochastic gradient = gradient + noise



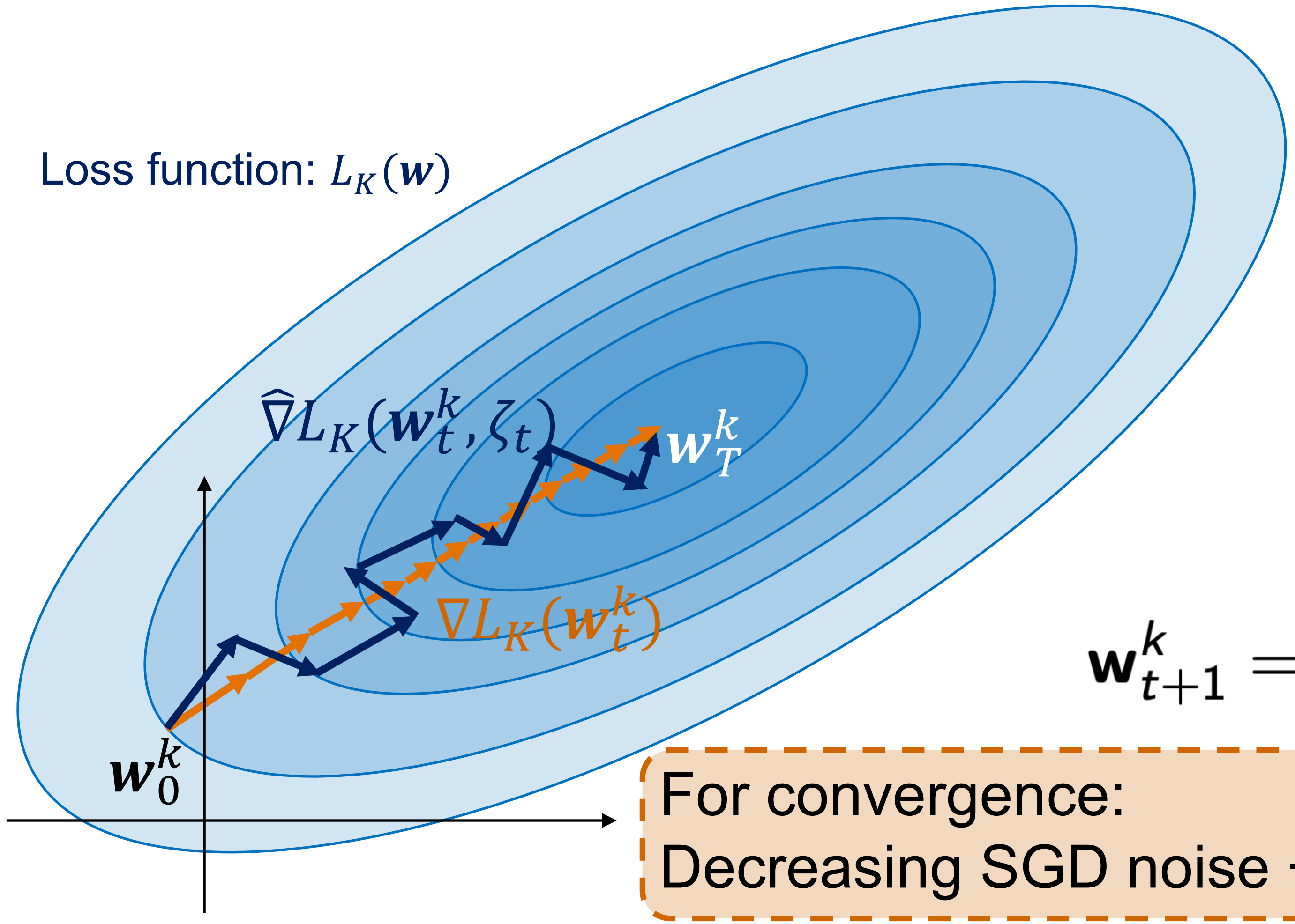
$$\mathbf{w}_{t+1}^k = \mathbf{w}_t^k - \boxed{\eta_t} \hat{\nabla} L(\mathbf{w}_t^k, \zeta_t)$$

For convergence:
Decreasing learning rate → Decreasing effective SGD noise

SGD noise

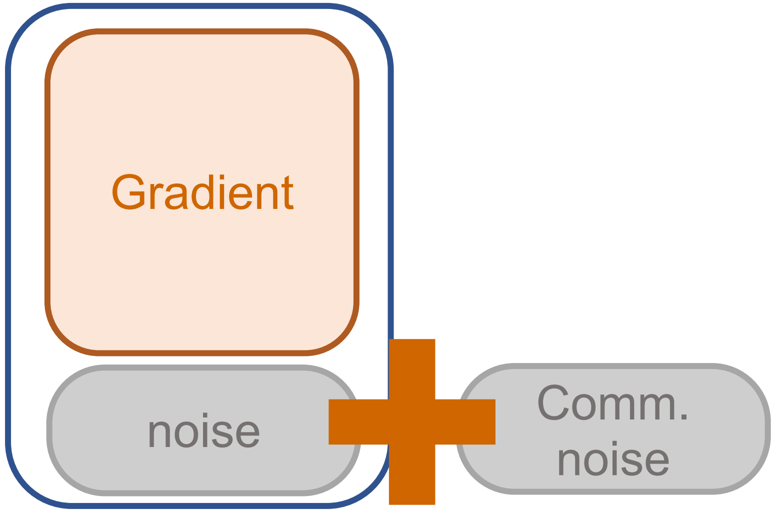
Gradient descent (GD)

Stochastic gradient descent (SGD)



$$\nabla L(\mathbf{w}_t^k) = \mathbb{E}[\hat{\nabla} L(\mathbf{w}_t^k, \zeta_t)]$$

stochastic gradient = gradient + noise



$$\mathbf{w}_{t+1}^k = \mathbf{w}_t^k - \eta_t \hat{\nabla} L(\mathbf{w}_t^k, \zeta_t) + \boxed{\mathbf{n}_{t+1}^k}$$

For convergence:

Decreasing SGD noise + **Decreasing effective comm. noise**

Convergence over noisy channel

For L-smooth, μ-strongly convex and bounded-gradient loss function

- Effective SNR control policy

Effective DL noise power

$$\sigma_t^2 \sim \mathcal{O}\left(\frac{1}{t^2}\right)$$

+

Effective UL noise power

$$\zeta_t^2 \sim \mathcal{O}\left(\frac{1}{t^2}\right)$$

FL tasks with non-IID datasets and partial/full clients participation converge at rate $\mathcal{O}\left(\frac{1}{t}\right)$.

Channel noise should not dominate the SGD noise.

Model differential for UL

Model differential

$$\boldsymbol{x}_t^k = \boldsymbol{w}_t^k - \boldsymbol{w}_{t-1}$$

Uplink comm.

Global model recovery

$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} + \frac{1}{k}\sum_{k=1}^{K} \boldsymbol{x}_t^k = \frac{1}{k}\sum_{k=1}^{K} \boldsymbol{w}_t^k$$

Model differential for UL

Model differential

$$\boldsymbol{x}_t^k = \boldsymbol{w}_t^k - \boldsymbol{w}_{t-1}$$

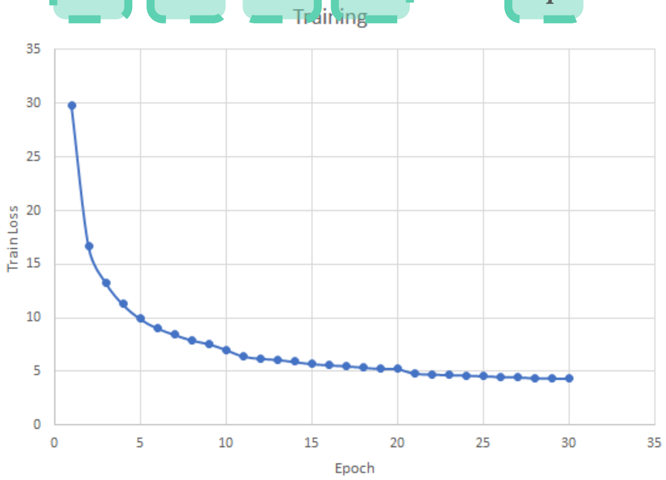
Uplink comm.

Global model recovery

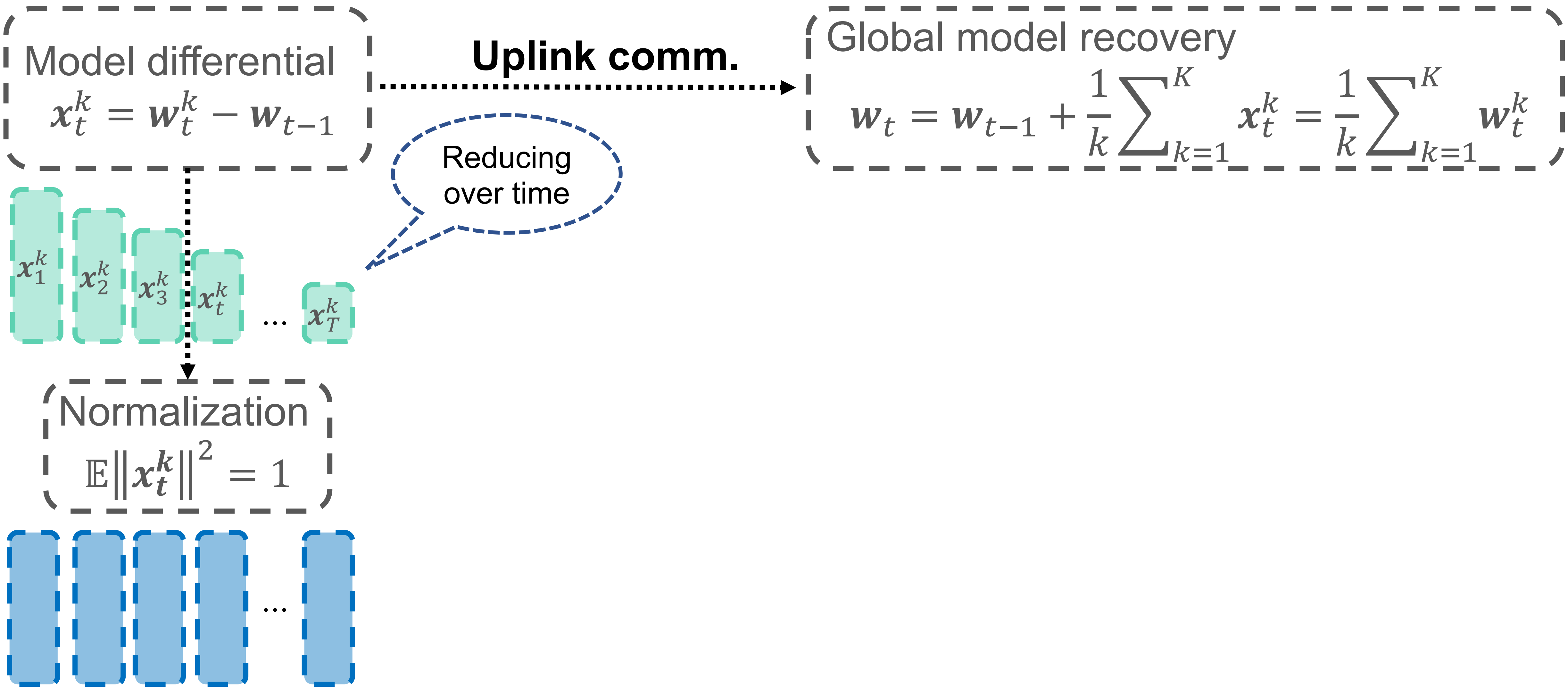
$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} + \frac{1}{k}\sum_{k=1}^{K} \boldsymbol{x}_t^k = \frac{1}{k}\sum_{k=1}^{K} \boldsymbol{w}_t^k$$

Reducing over time

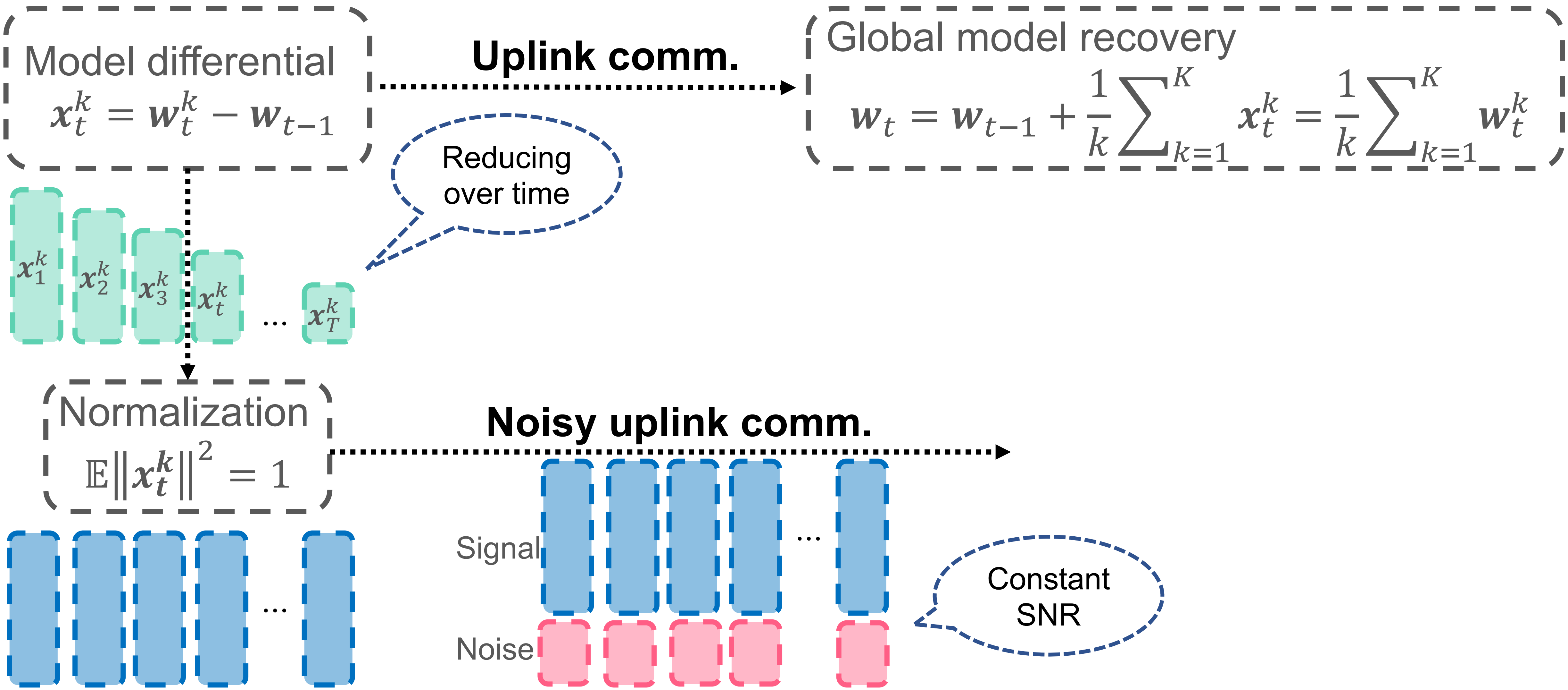
$\boldsymbol{x}_1^k$ $\boldsymbol{x}_2^k$ $\boldsymbol{x}_3^k$ $\boldsymbol{x}_t^k$ … $\boldsymbol{x}_T^k$



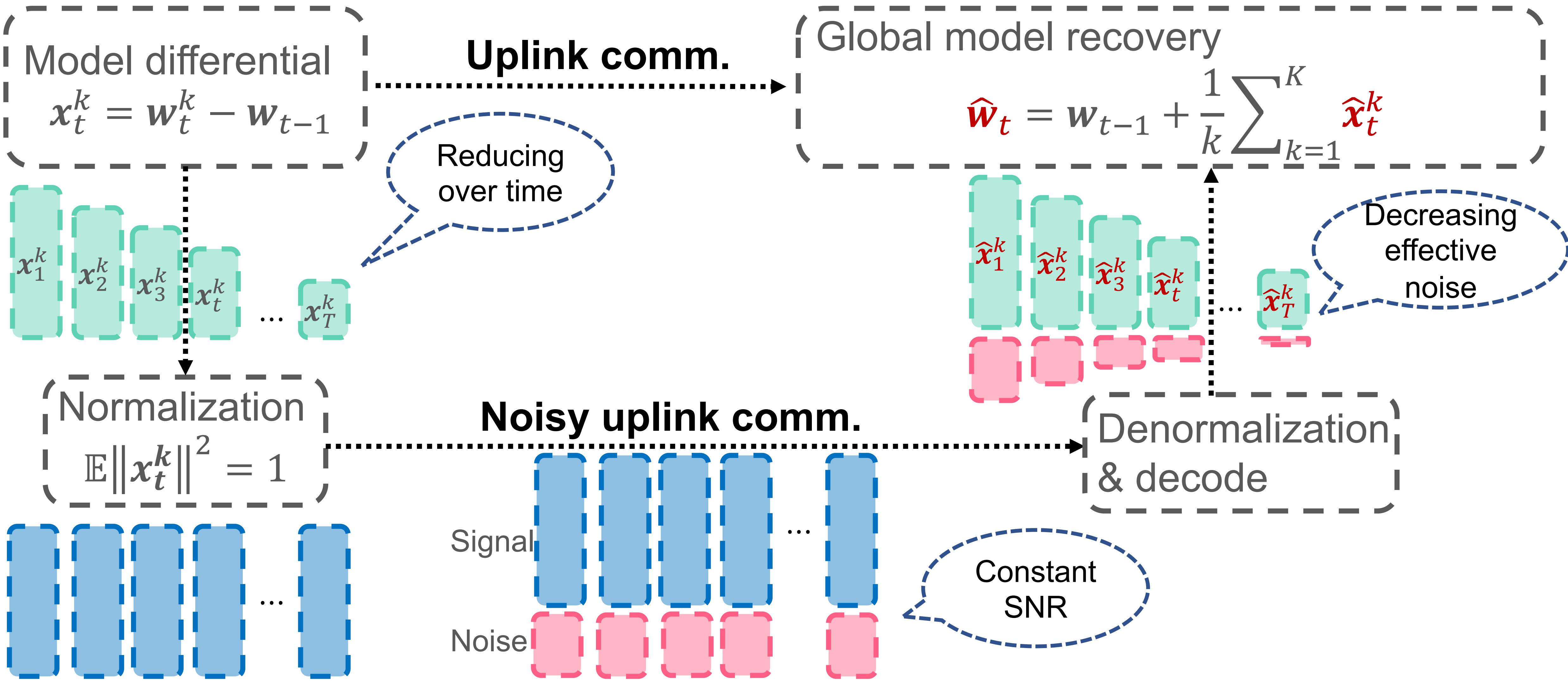
Model differential for UL



Model differential for UL



Model differential for Uplink



Convergence over noisy channel

For L-smooth, μ-strongly convex and bounded-gradient loss function

- Effective SNR control policy for **uplink model differential**

Effective DL noise power

$$\sigma_t^2 \sim \mathcal{O}\left(\frac{1}{t^2}\right)$$

+

Effective UL noise power

$$\zeta_t^2 \sim \mathcal{O}(1)$$

FL tasks with non-IID datasets and partial clients participation converge at rate $\mathcal{O}\left(\frac{1}{t}\right)$.

We cannot adopt model differential for downlink due to partial participation.

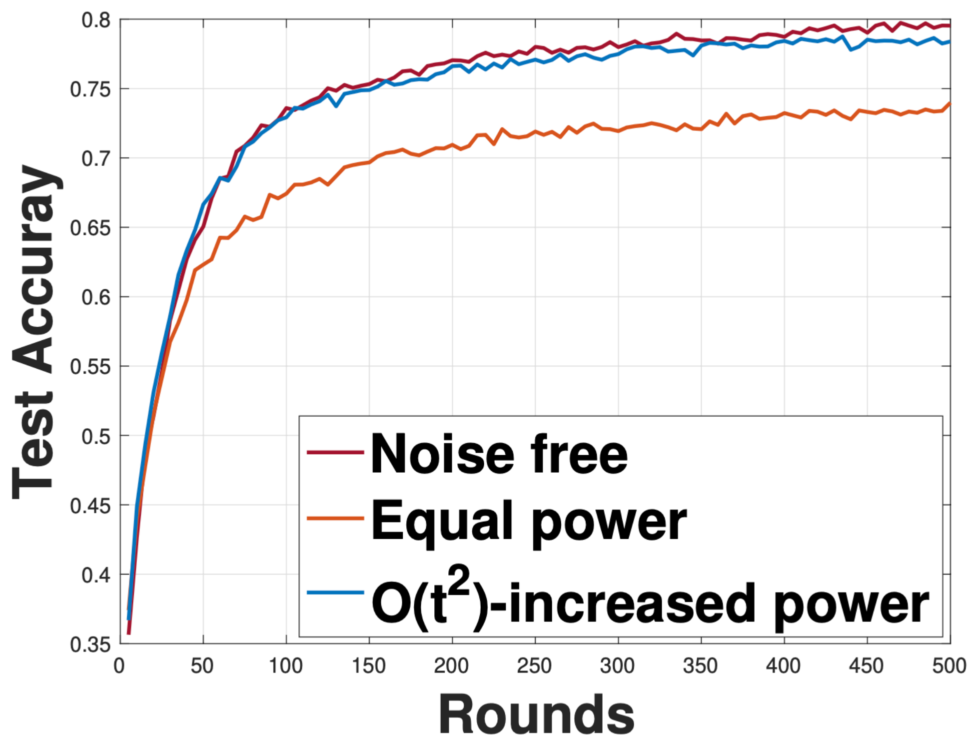
Experiment

- Noise free (ideal)

Under same budget

- Equal power allocation
- $\mathcal{O}(t^2)$-increased power allocation

CIFAR-10 IID



CIFAR-10 non-IID