Communication-efficient distributed optimization algorithms

Laurent Condat

King Abdullah Univ. of Science and Technology (KAUST)
Saudi Arabia

Peter Richtárik

Hanoi, Mar. 2025

Foreword: the power of randomness

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \ \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{using the } \nabla f_i$$

with every f_i L-smooth and μ-strongly convex

lower bounds in Woodworth & Srebro [2016] on number of gradient calls:

- deterministic algorithms: $\Omega\left(n\sqrt{L/\mu}\log\epsilon^{-1}\right)$

KAUST

Foreword: the power of randomness

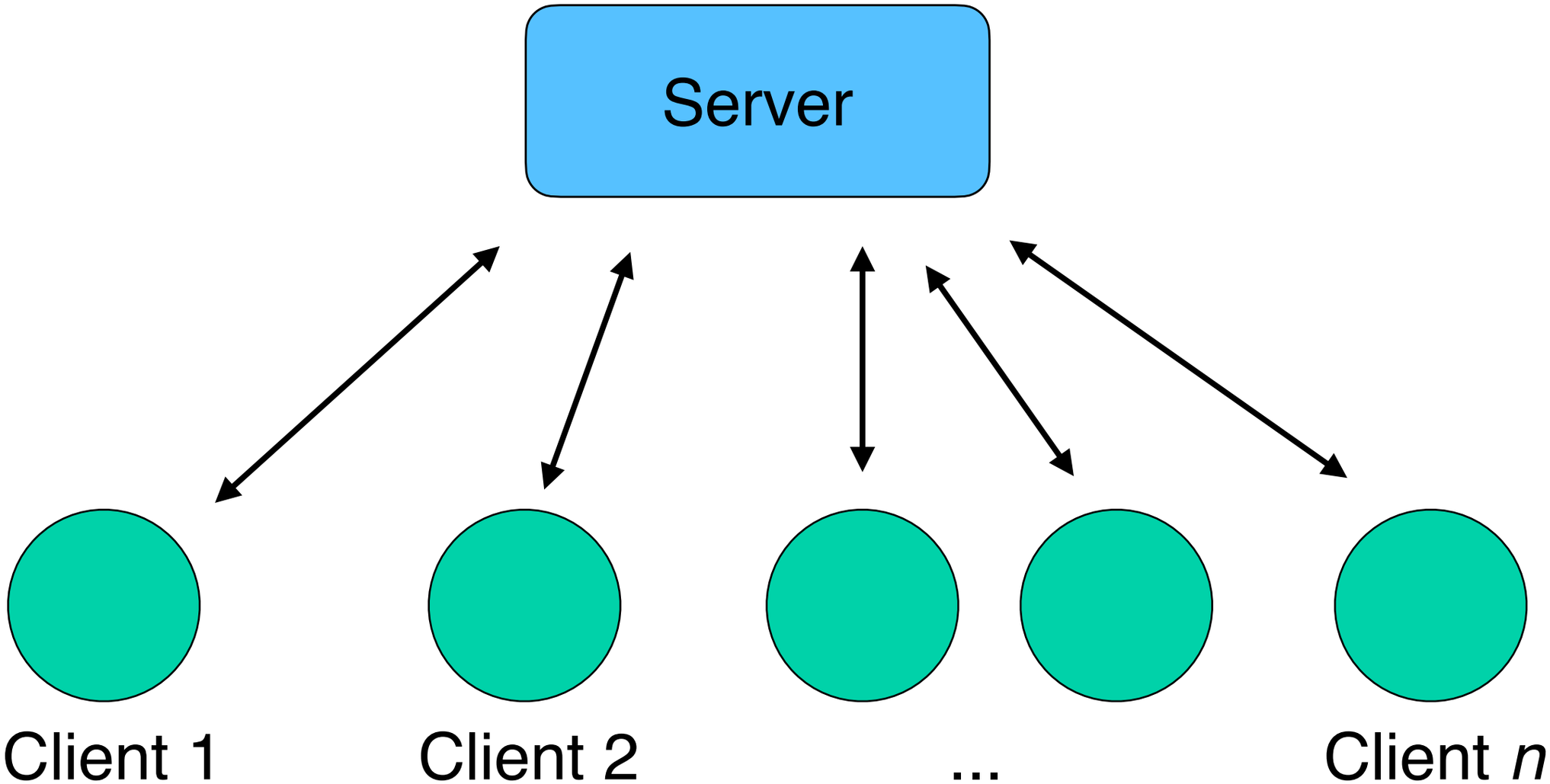
$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \ \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{using the } \nabla f_i$$

with every f_i L-smooth and μ-strongly convex

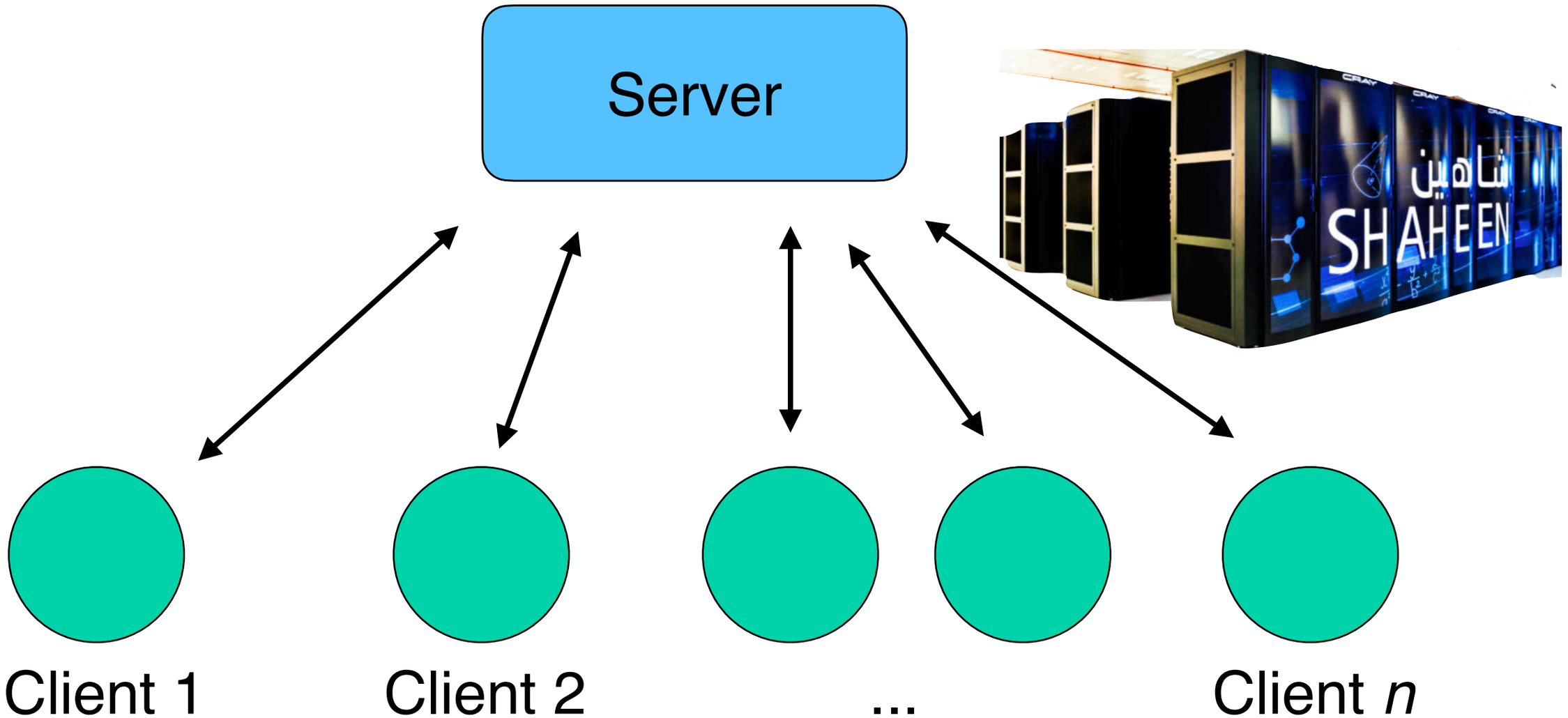
☞ lower bounds in Woodworth & Srebro [2016] on number of gradient calls:

- deterministic algorithms: $\Omega\big(n\sqrt{L/\mu}\log\epsilon^{-1}\big)$
- randomized algorithms: $\Omega\big((n+\sqrt{nL/\mu})\log\epsilon^{-1}\big)$

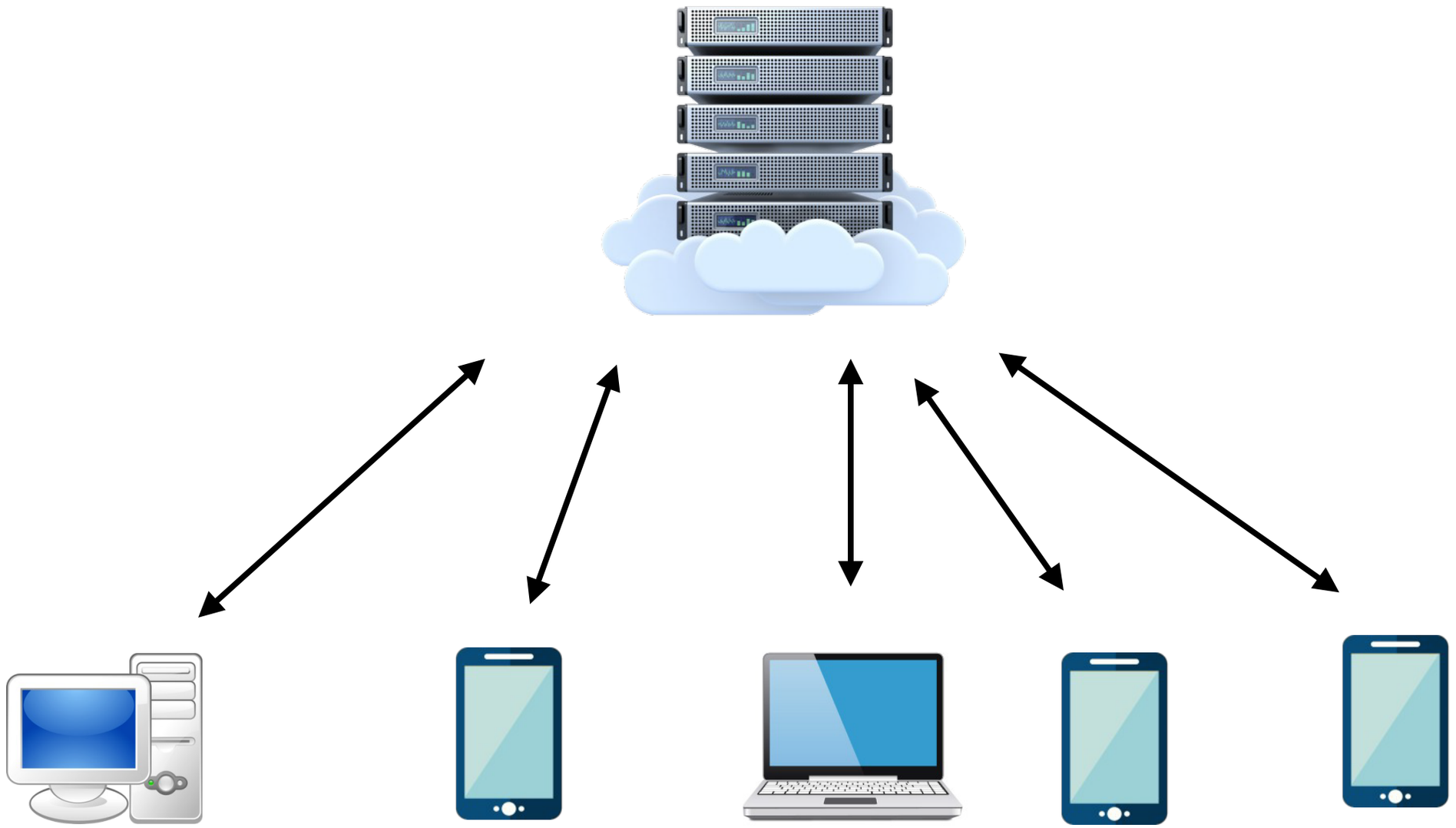
Distributed optimization



Distributed optimization

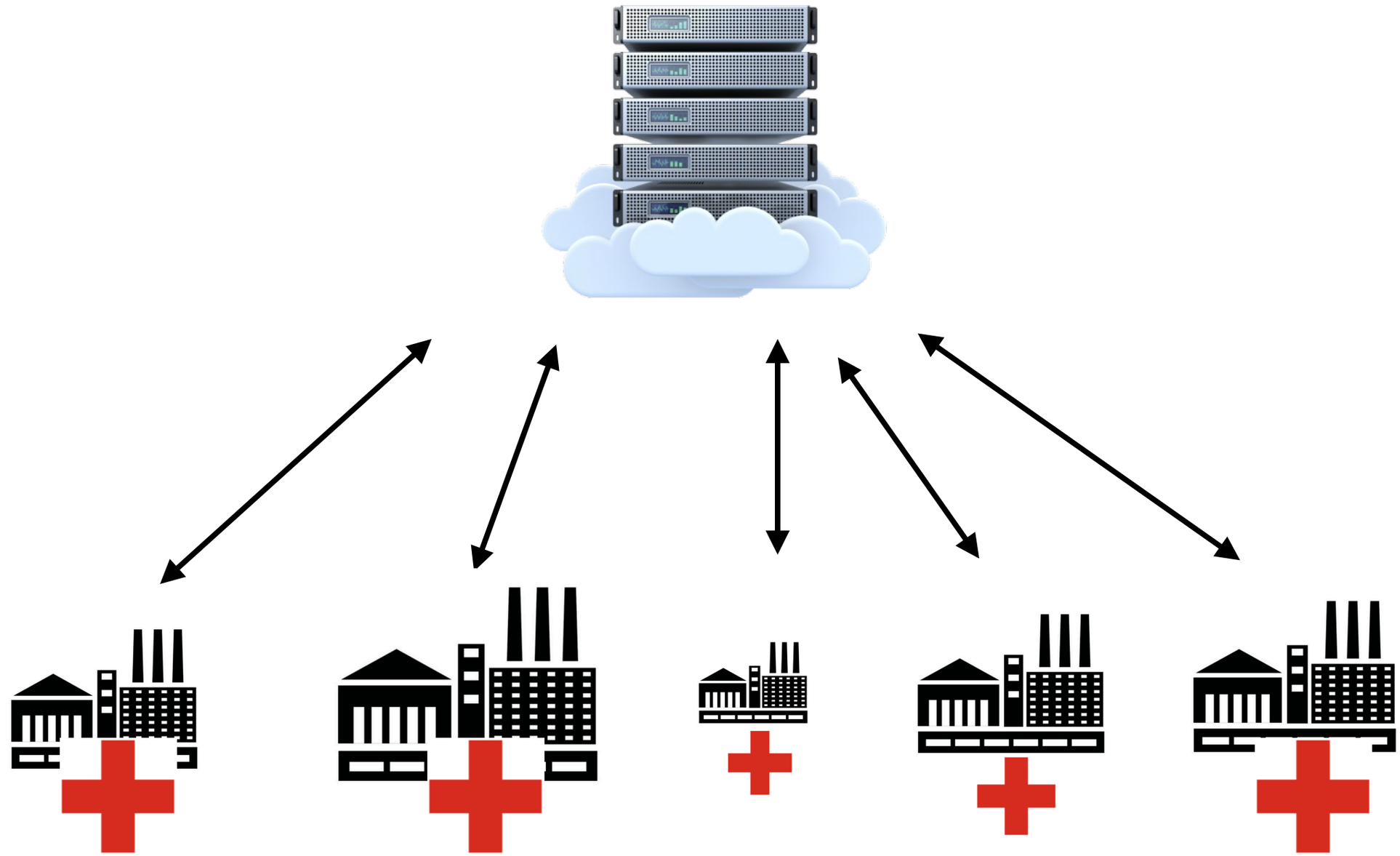


Federated learning

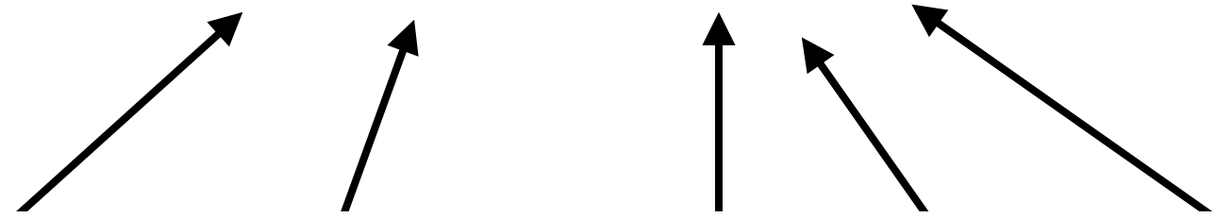


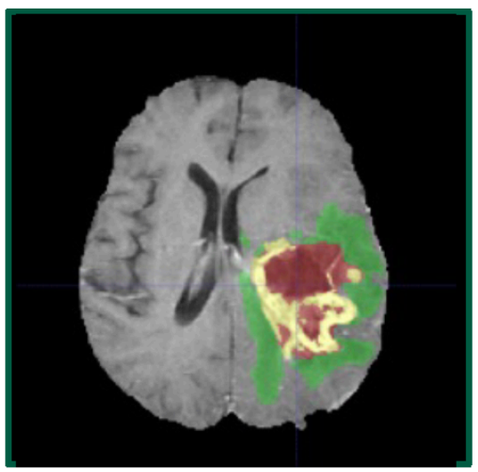
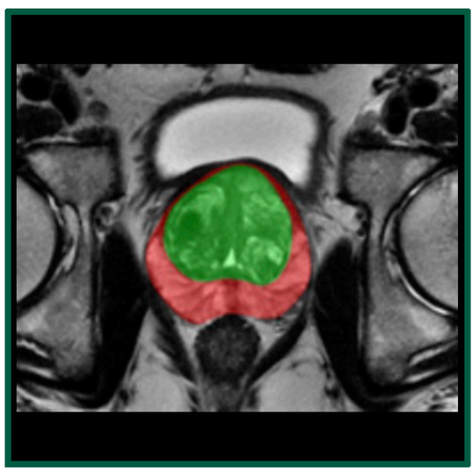
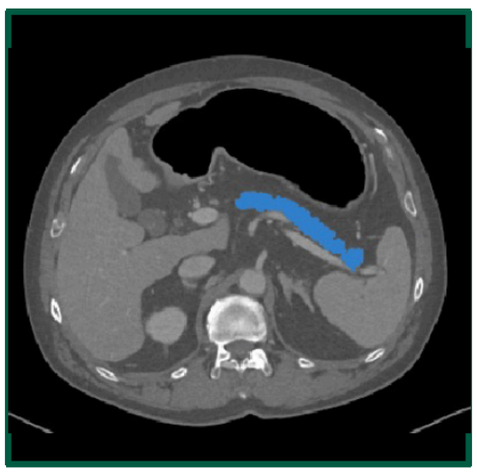
Federated learning

Federated learning



Federated learning



MICCAI 2019	MICCAI 2020	JAMIA 2021	MICCAI 2020	Nature Medicine 2021
				
Brain Tumor Segmentation KCL	Mammography Classification ACR/DASA/OSU/MGB/Stanford	Prostate Segmentation SUNY/UCLA/NIH	Pancreas Segmentation NTU/Nagoya	CXR/EMR Classification 20 Global Collaborators

FL for medical imaging [Holger Roth, Nvidia]

Optimization problem

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \ \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

Every function f_i is μ-strongly convex and L-smooth, for some $L \geq \mu > 0$

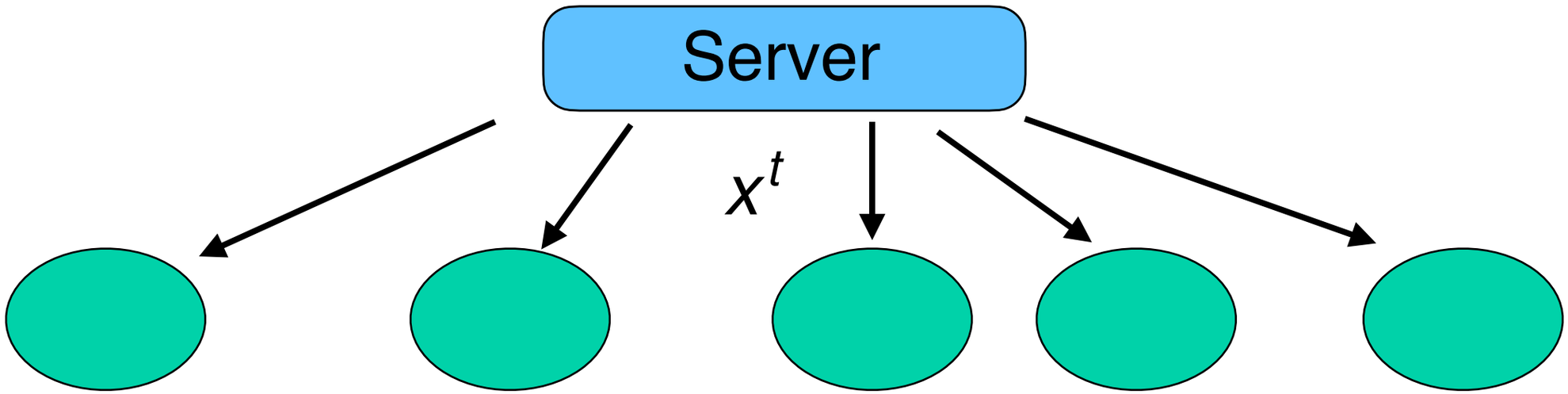
Optimization problem

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \; \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

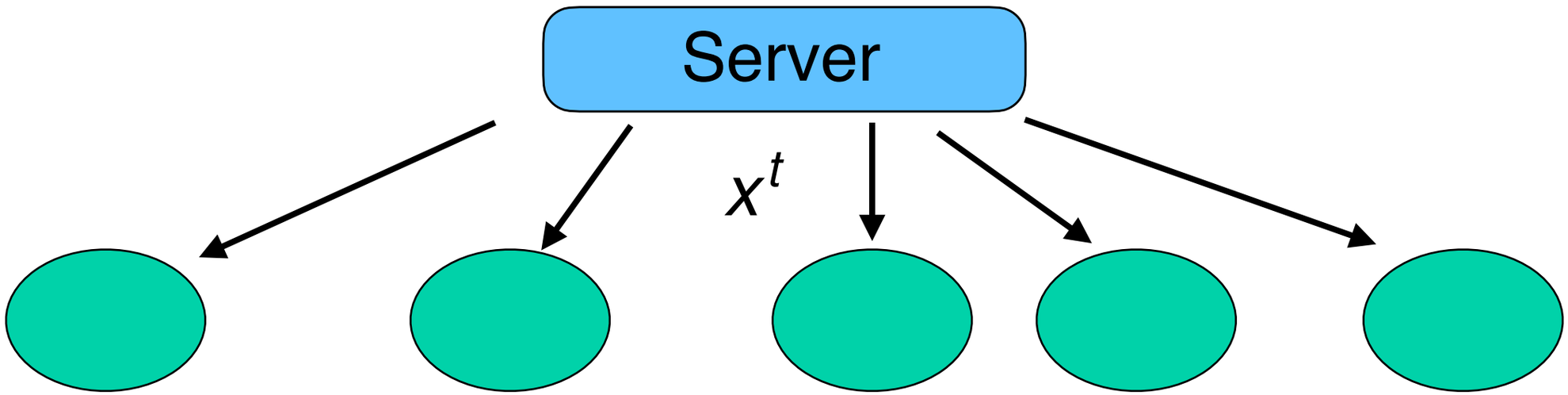
Every function f_i is μ-strongly convex and L-smooth, for some $L \geq \mu > 0$

$$\kappa := \frac{L}{\mu}$$

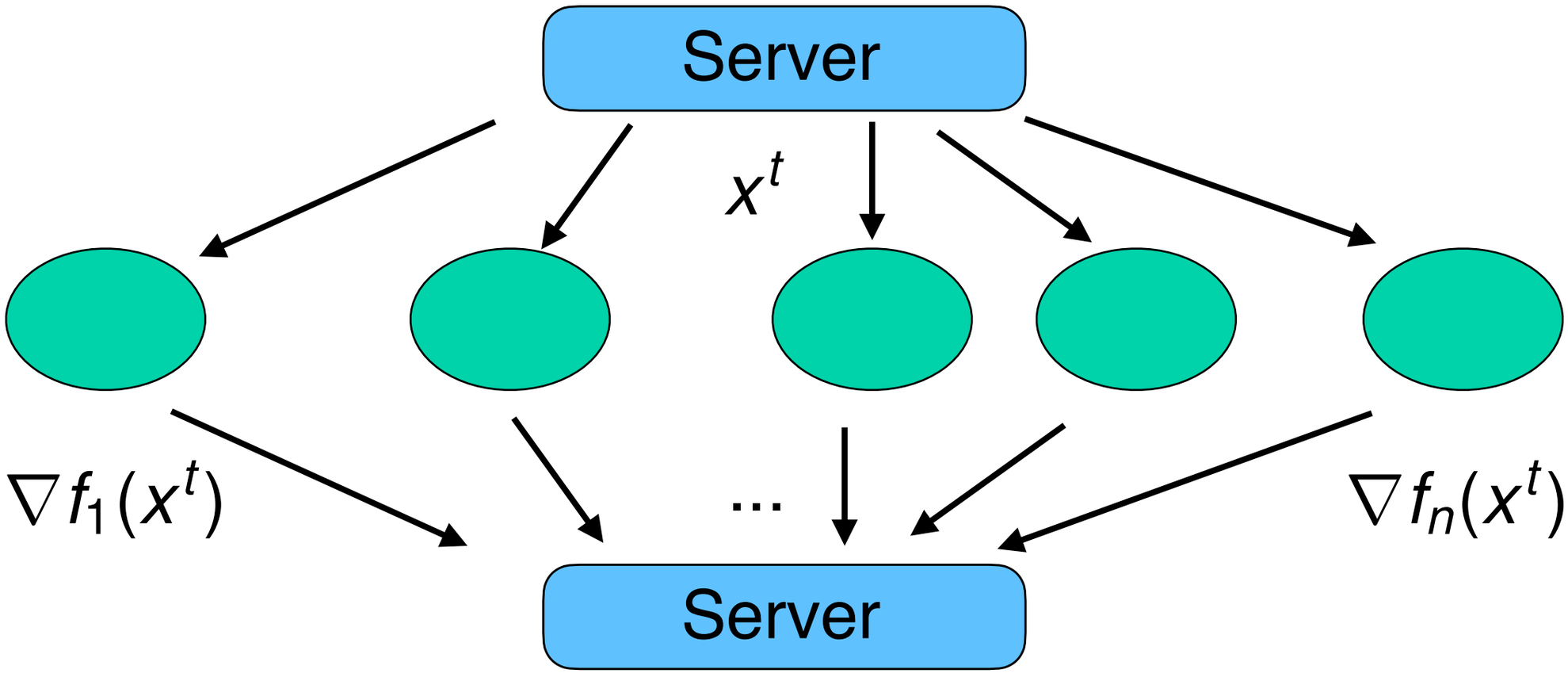
Distributed GD



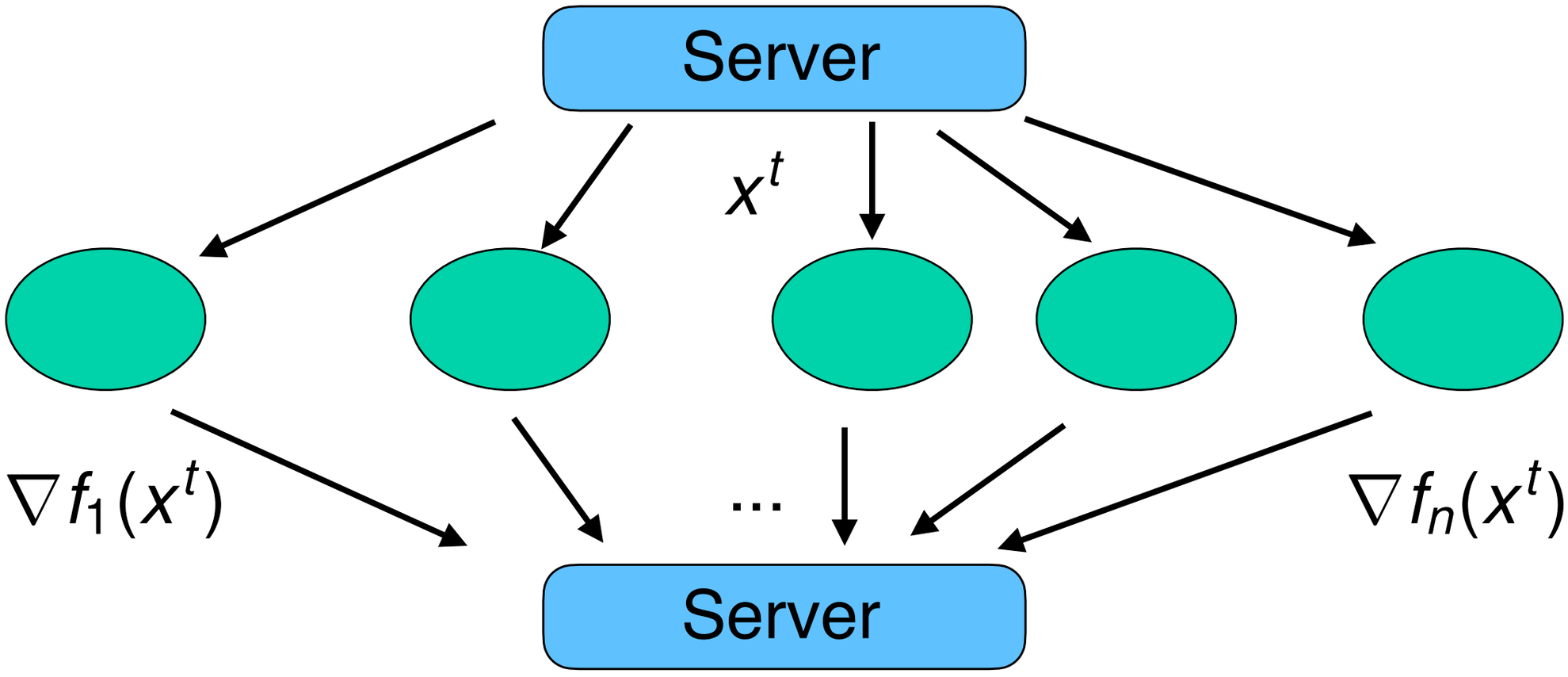
Distributed GD



Distributed GD

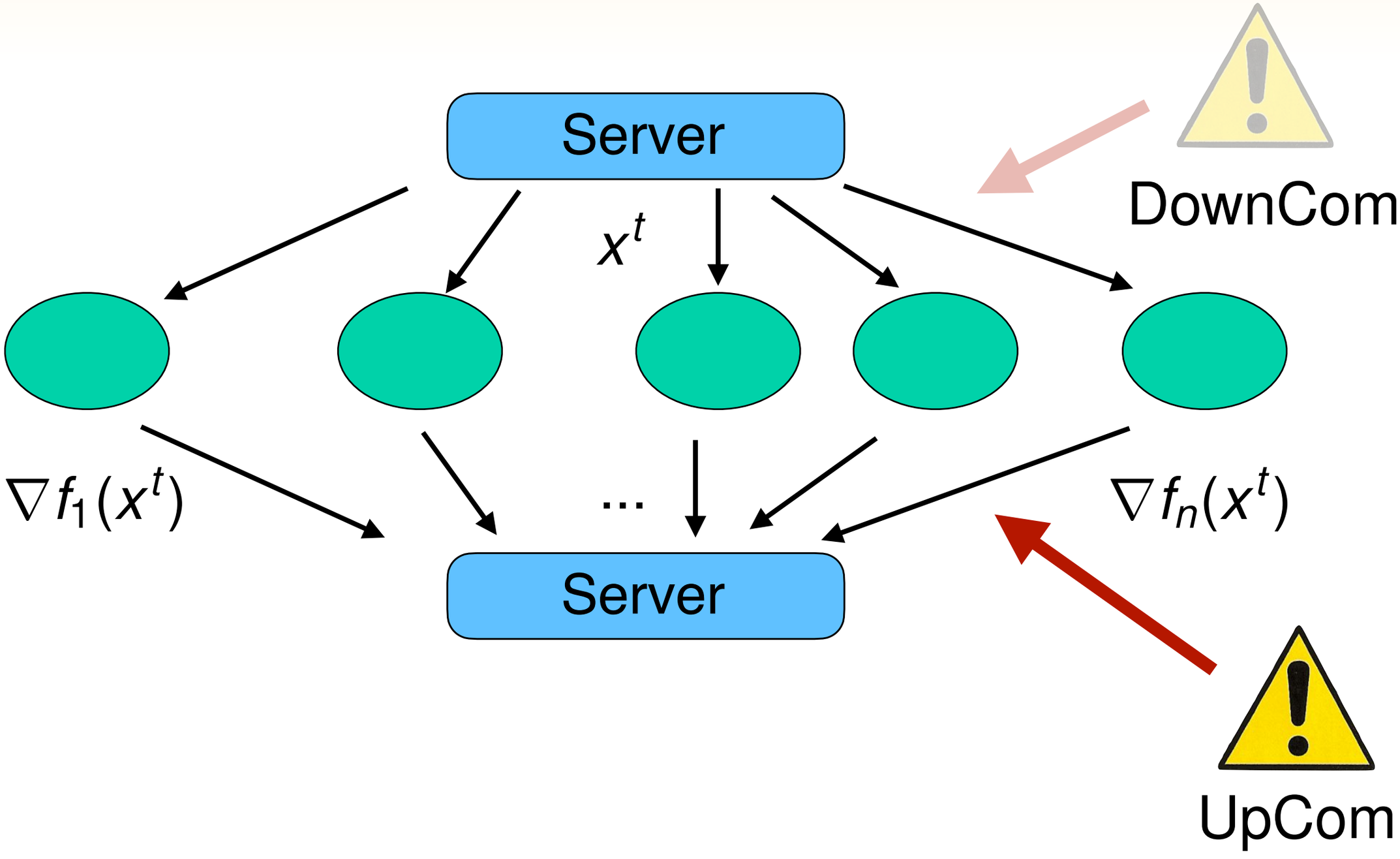


Distributed GD

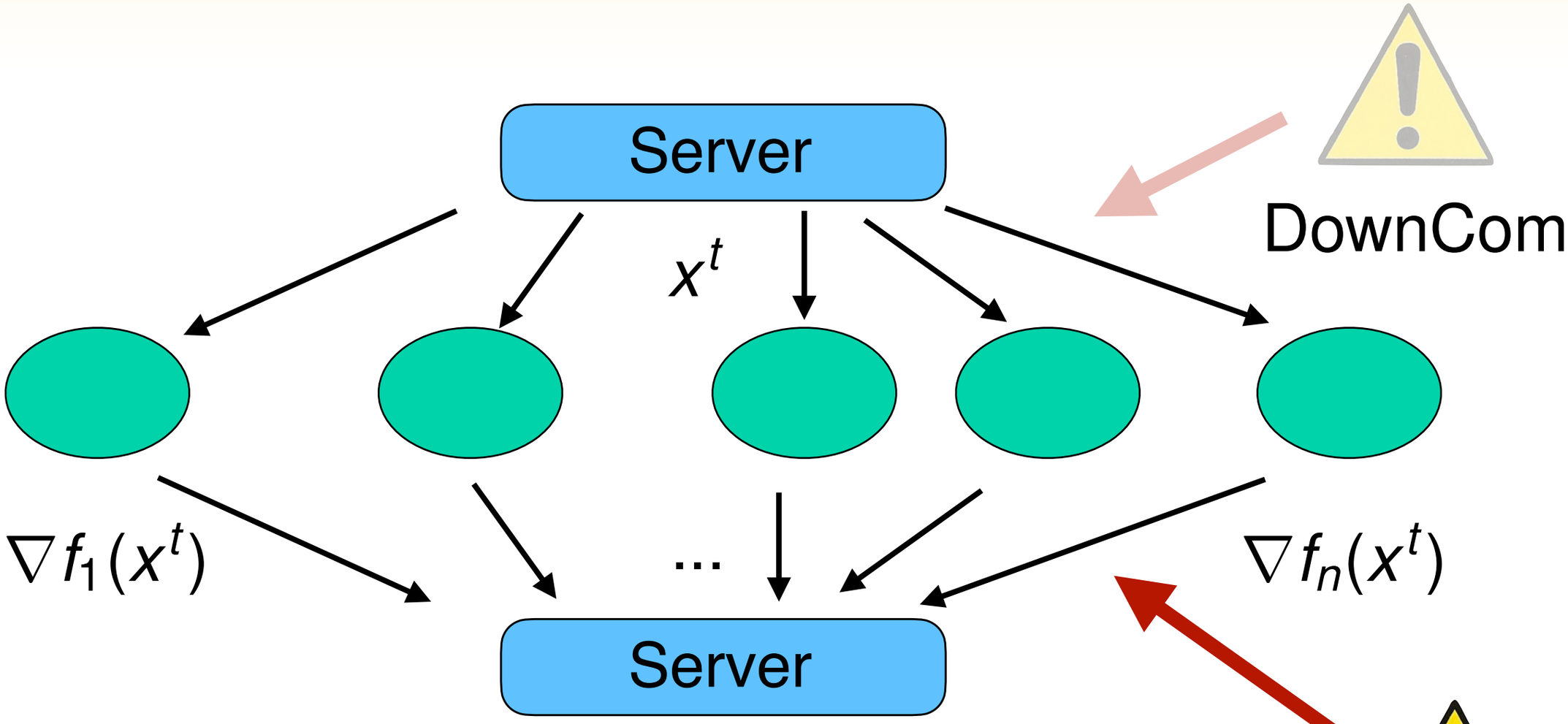


$$x^{t+1} := x^t - \frac{\gamma}{n} \sum_{i=1}^{n} \nabla f_i(x^t)$$

Distributed GD

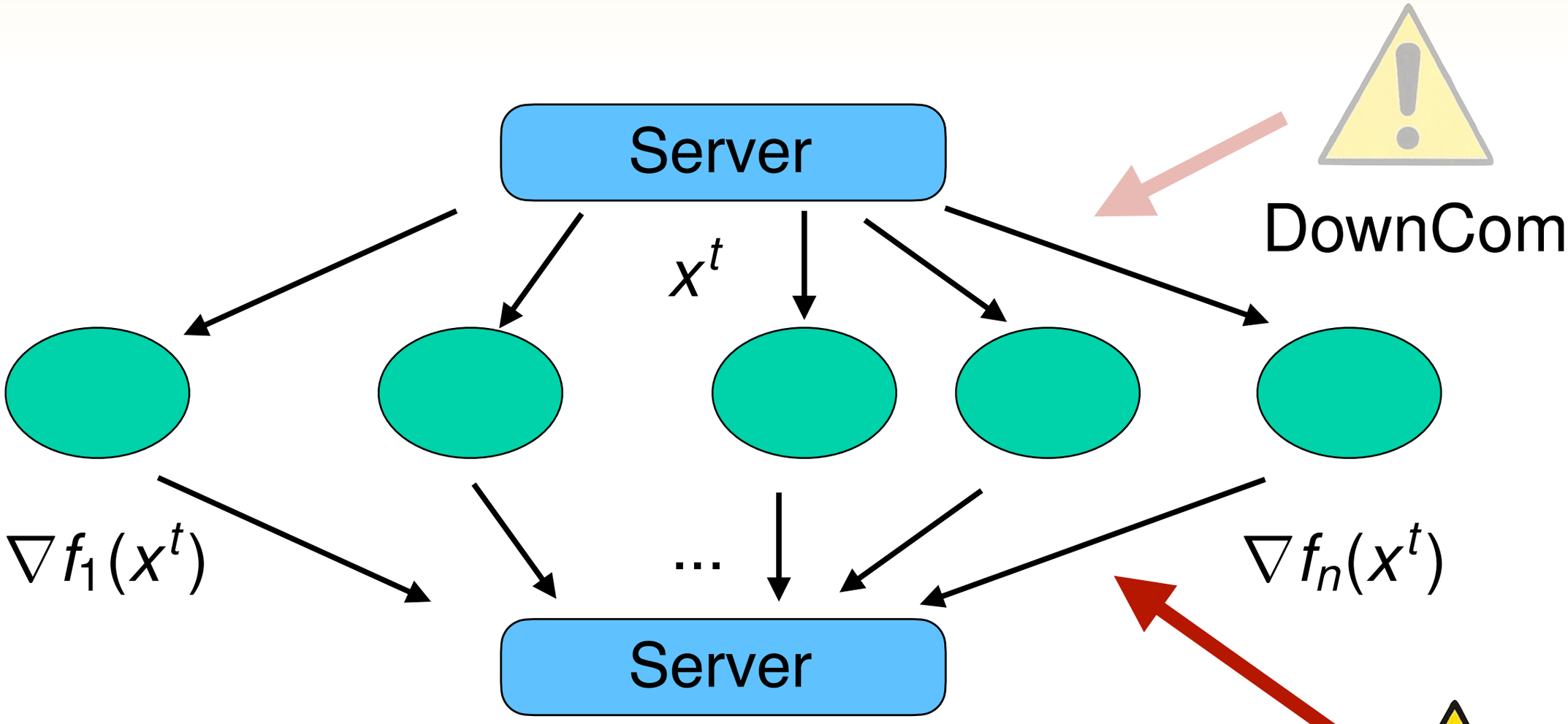


Distributed GD



complexity = #com. rounds $\times$ #reals per round to reach ϵ-accuracy

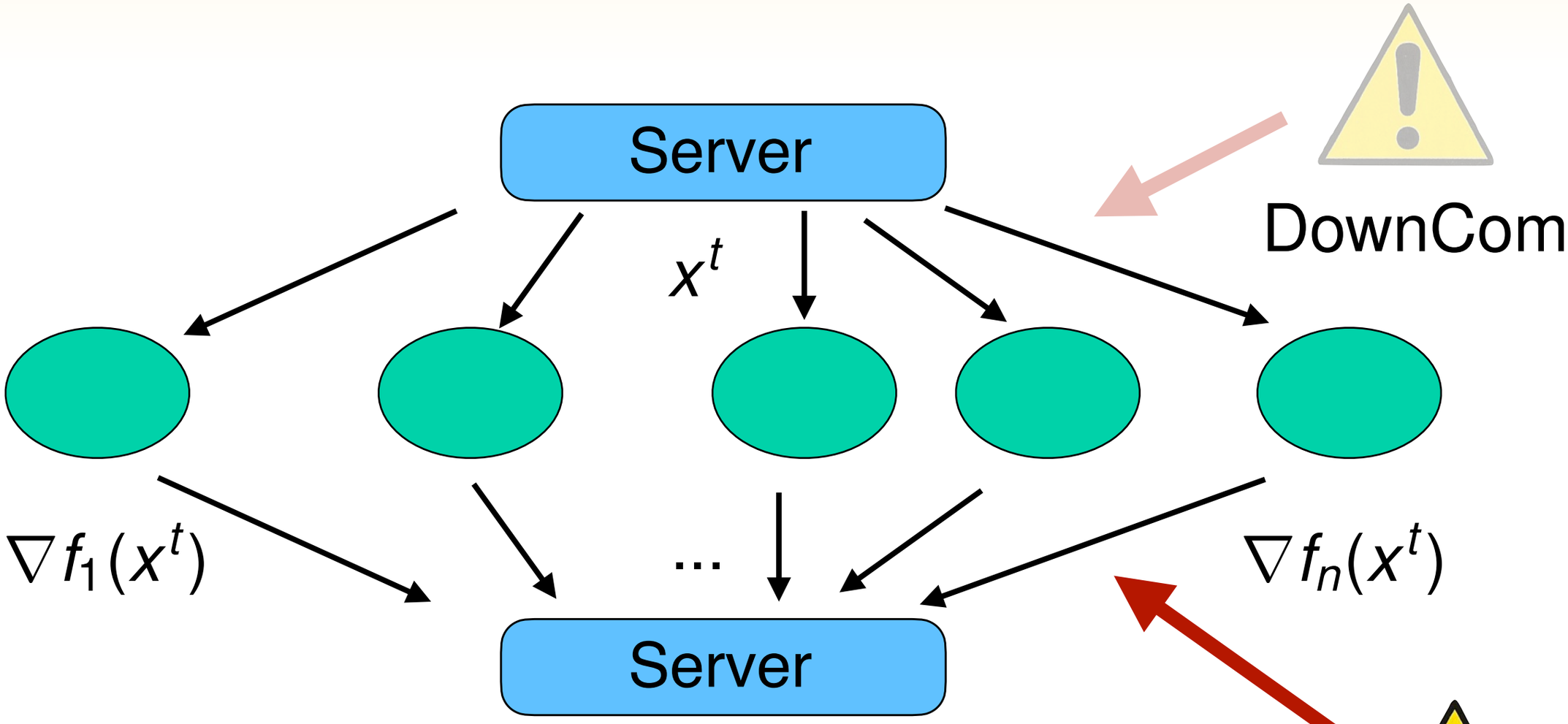
Distributed GD



complexity = #com. rounds $\times$ #reals per round to reach ϵ-accuracy

GD: $\mathcal{O}\left(d\kappa \log\left(\epsilon^{-1}\right)\right)$

Distributed GD



complexity = #com. rounds $\times$ #reals per round to reach ϵ-accuracy

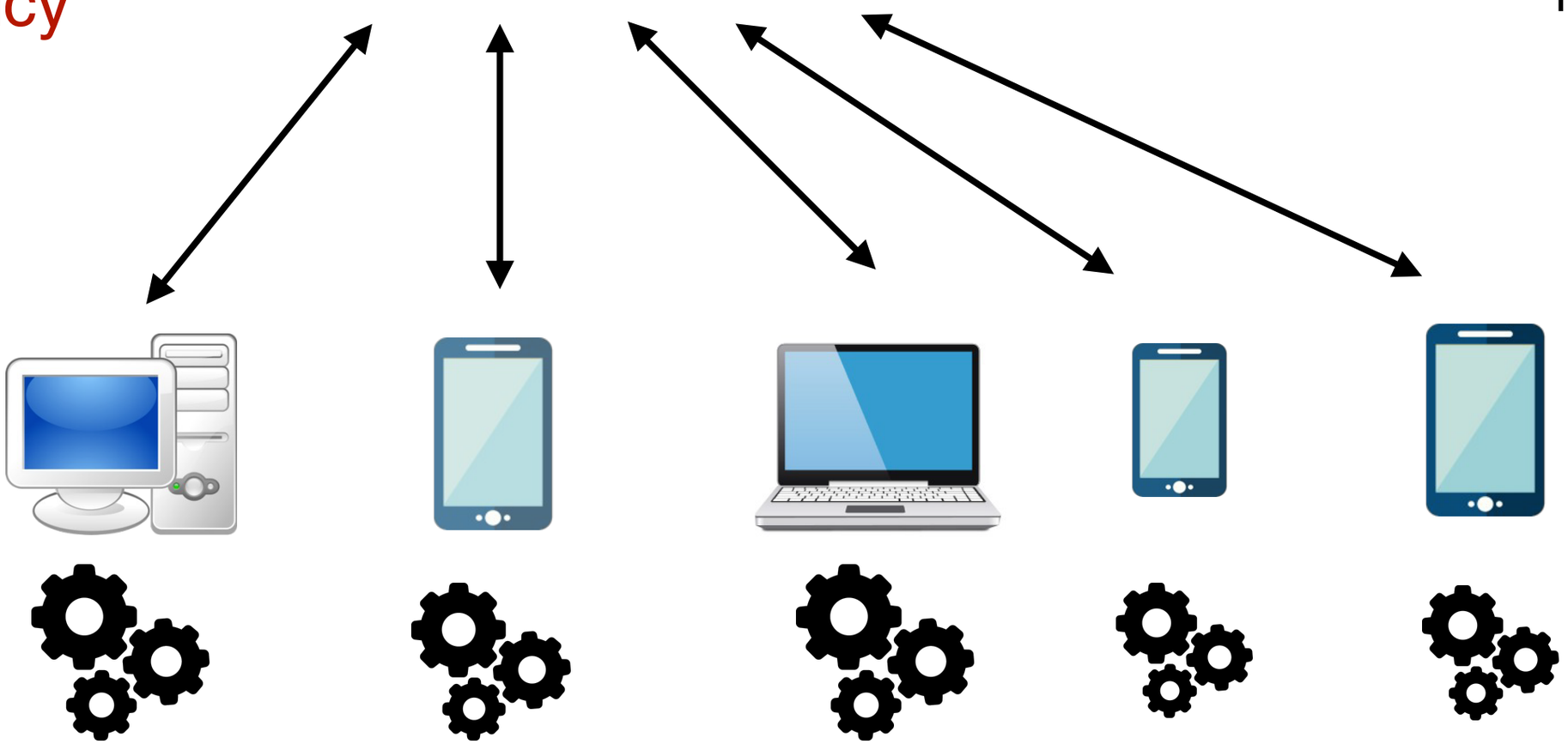
GD: $\mathcal{O}\left(d\kappa \log\left(\epsilon^{-1}\right)\right)$ AGD: $\mathcal{O}\left(d\sqrt{\kappa} \log\left(\epsilon^{-1}\right)\right)$

Local training

decrease frequency

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \; f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

condition number κ

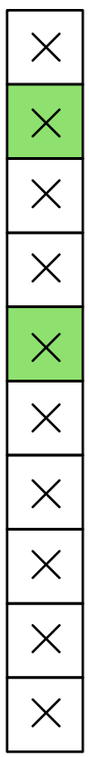


Compression

compress vectors

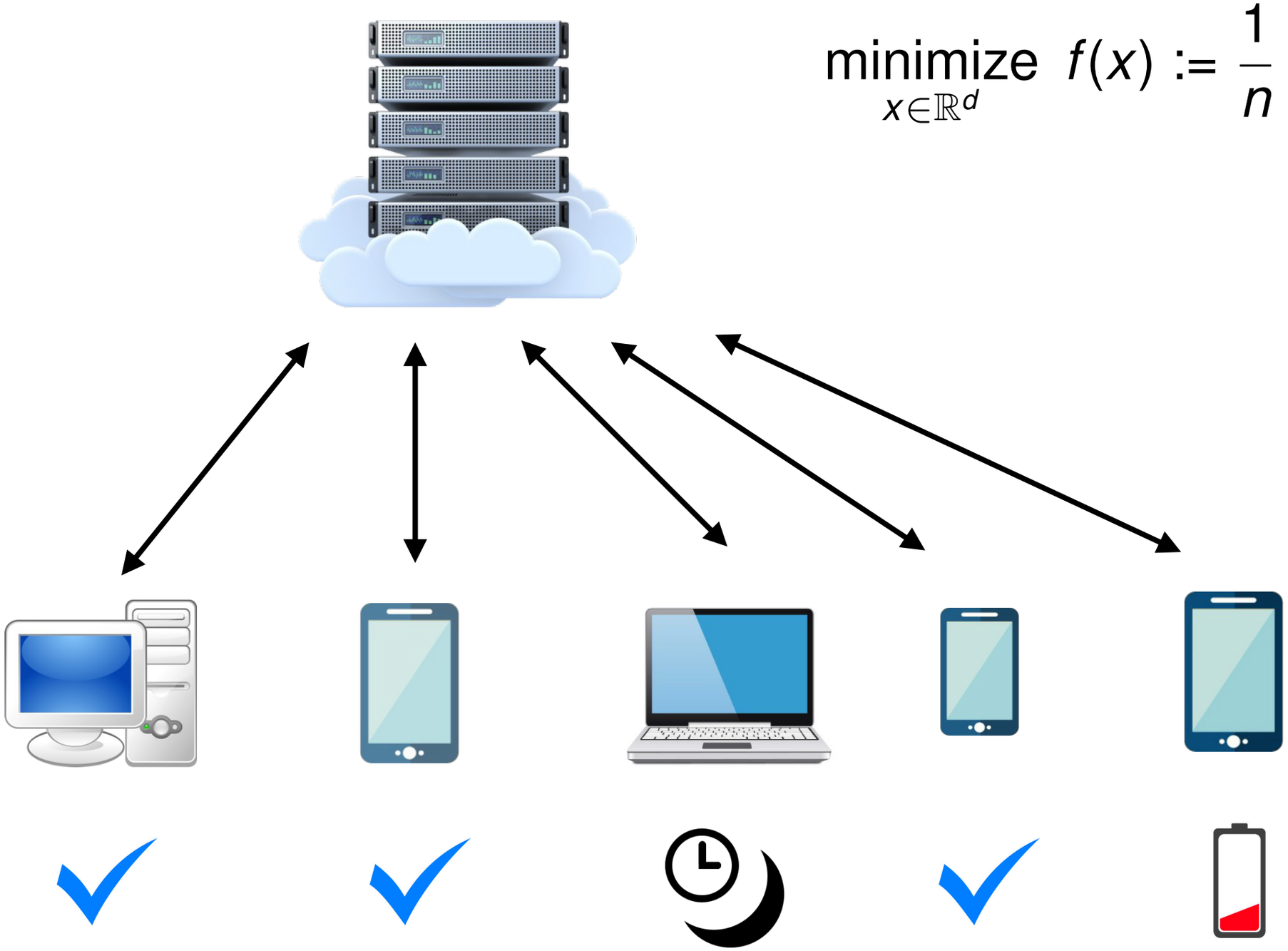
$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\ f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

e.g.



Partial participation

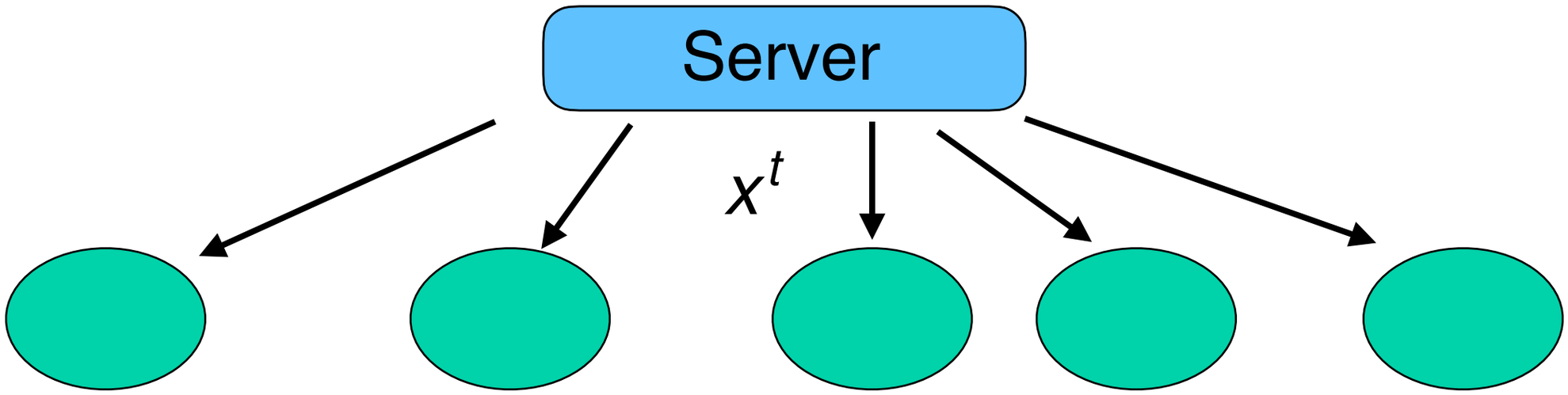
$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \; f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$



Laurent Condat

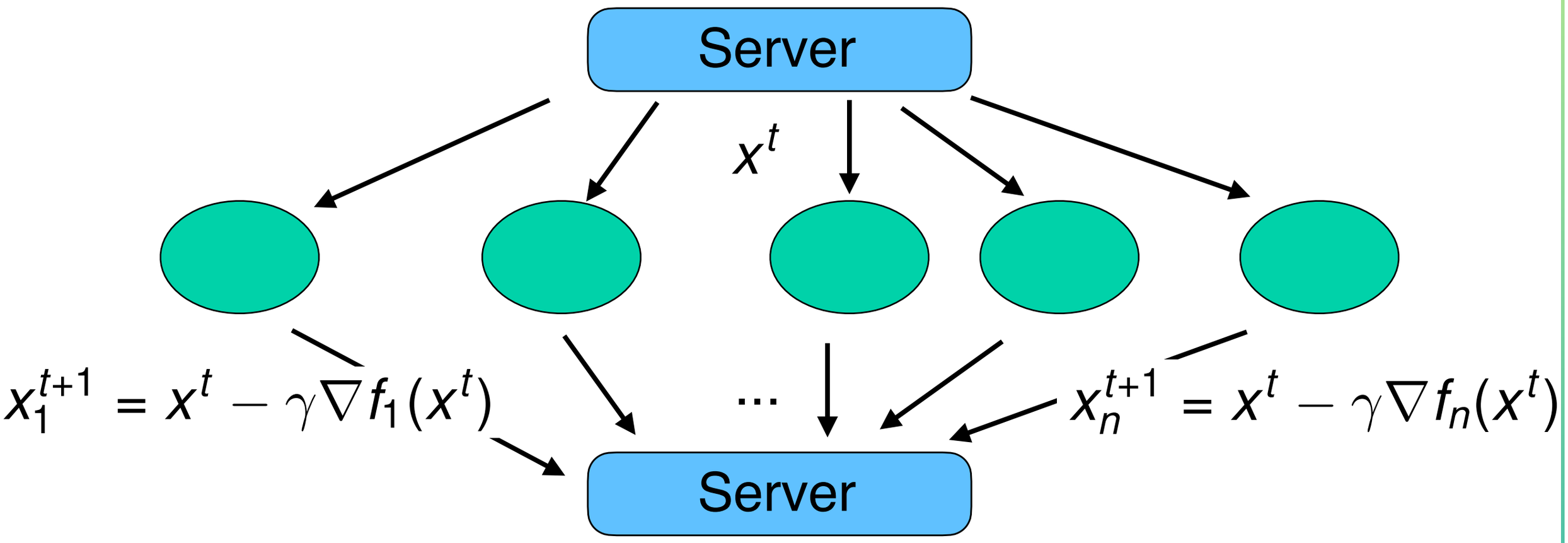
1) Local Training

Distributed GD



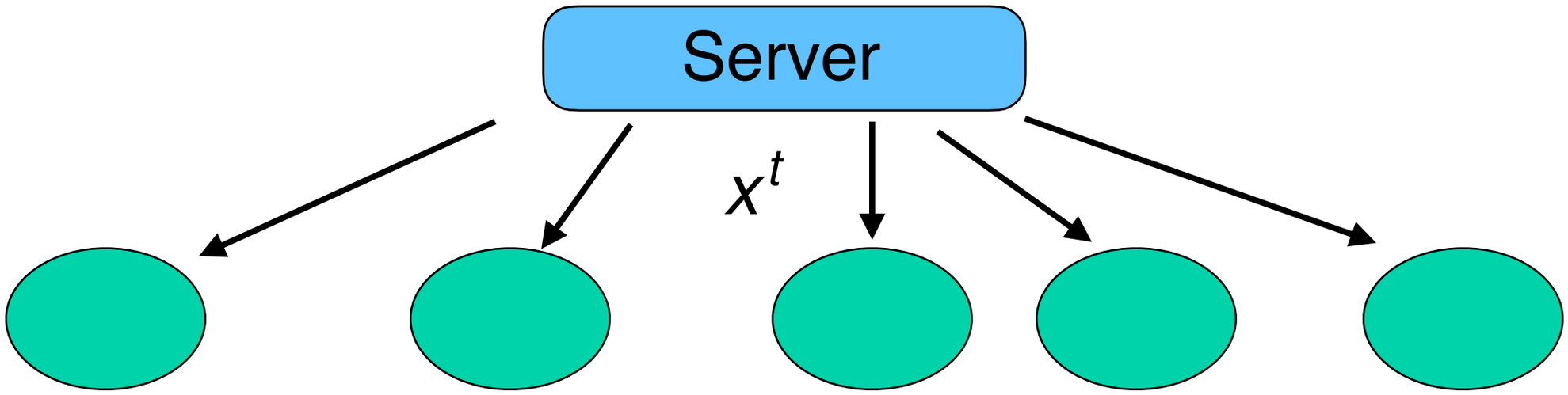
$$x_1^{t+1} = x^t - \gamma \nabla f_1(x^t) \qquad \cdots \qquad x_n^{t+1} = x^t - \gamma \nabla f_n(x^t)$$

Distributed GD



$$x^{t+1} := \frac{1}{n} \sum_{i=1}^{n} x_i^{t+1}$$

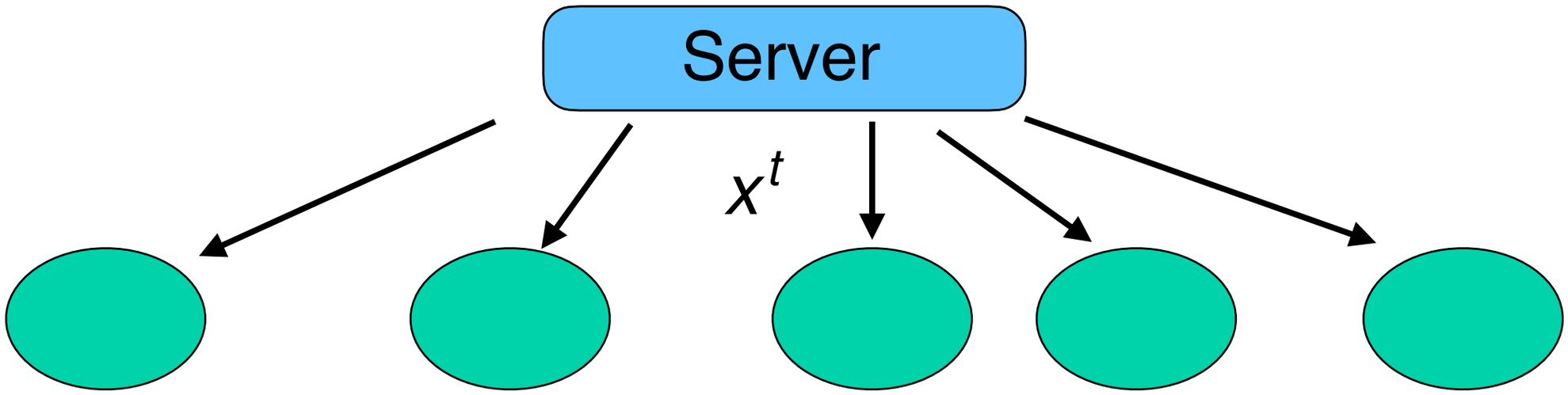
Distributed **Local** GD = FedAvg



$$x_1^{t+1} = x^t - \gamma \nabla f_1(x^t) \qquad \cdots \qquad x_n^{t+1} = x^t - \gamma \nabla f_n(x^t)$$

$$x_1^{t+2} = x_1^{t+1} - \gamma \nabla f_1(x_1^{t+1}) \qquad x_n^{t+2} = x_n^{t+1} - \gamma \nabla f_n(x_n^{t+1})$$

Distributed **Local** GD = FedAvg



$$x_1^{t+1} = x^t - \gamma \nabla f_1(x^t) \qquad \cdots \qquad x_n^{t+1} = x^t - \gamma \nabla f_n(x^t)$$

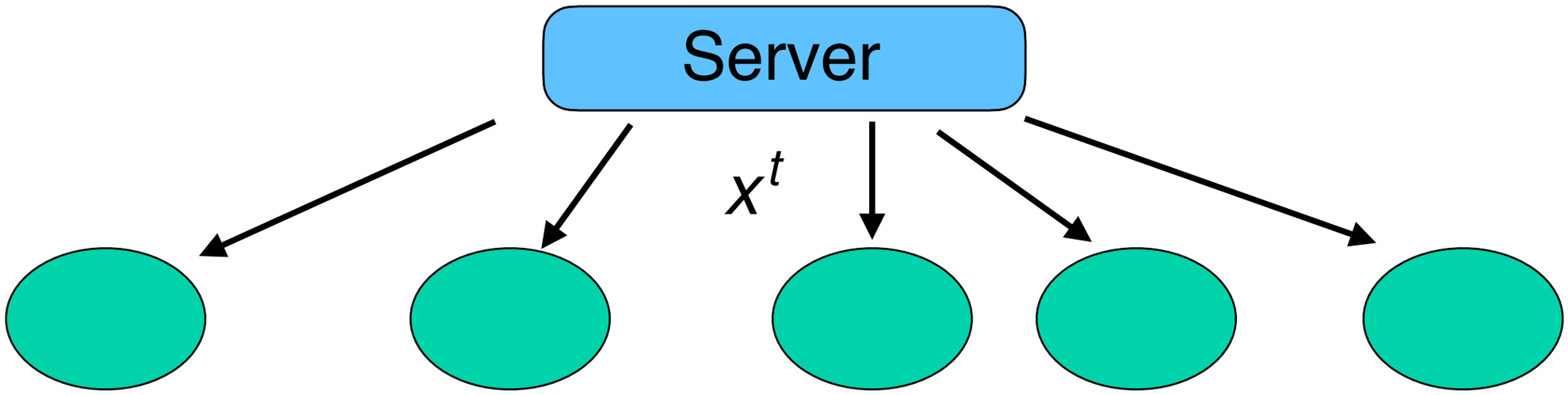
$$x_1^{t+2} = x_1^{t+1} - \gamma \nabla f_1(x_1^{t+1}) \qquad x_n^{t+2} = x_n^{t+1} - \gamma \nabla f_n(x_n^{t+1})$$

$$\cdots \qquad\qquad\qquad\qquad \cdots$$

$$x_1^{t+H} = x_1^{t+H-1} - \gamma \nabla f_1(x_1^{t+H-1}) \qquad x_n^{t+H} = x_n^{t+H-1} - \gamma \nabla f_n(x_n^{t+H-1})$$

$$H \geq 1$$

Distributed **Local** GD = FedAvg

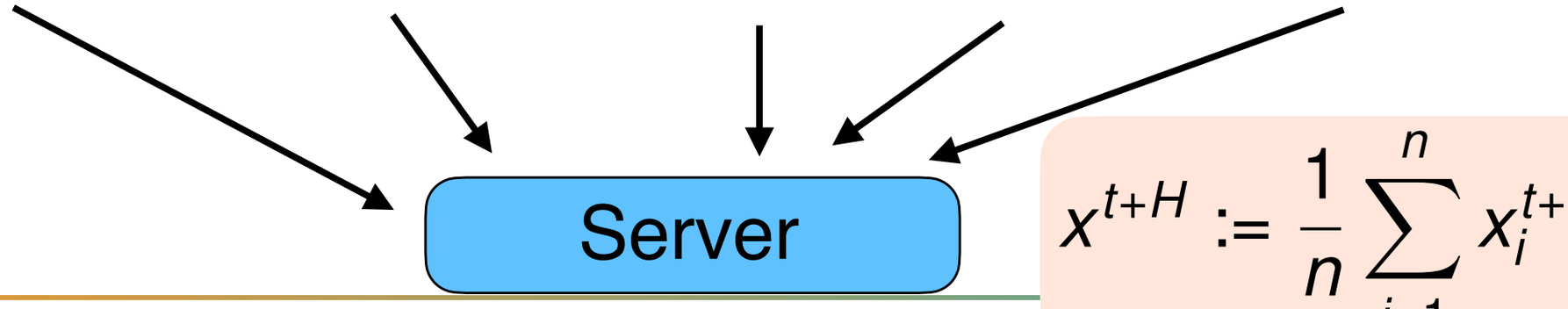


$$x_1^{t+1} = x^t - \gamma \nabla f_1(x^t) \qquad \cdots \qquad x_n^{t+1} = x^t - \gamma \nabla f_n(x^t)$$

$$x_1^{t+2} = x_1^{t+1} - \gamma \nabla f_1(x_1^{t+1}) \qquad x_n^{t+2} = x_n^{t+1} - \gamma \nabla f_n(x_n^{t+1})$$

$$\cdots \qquad\qquad\qquad\qquad \cdots$$

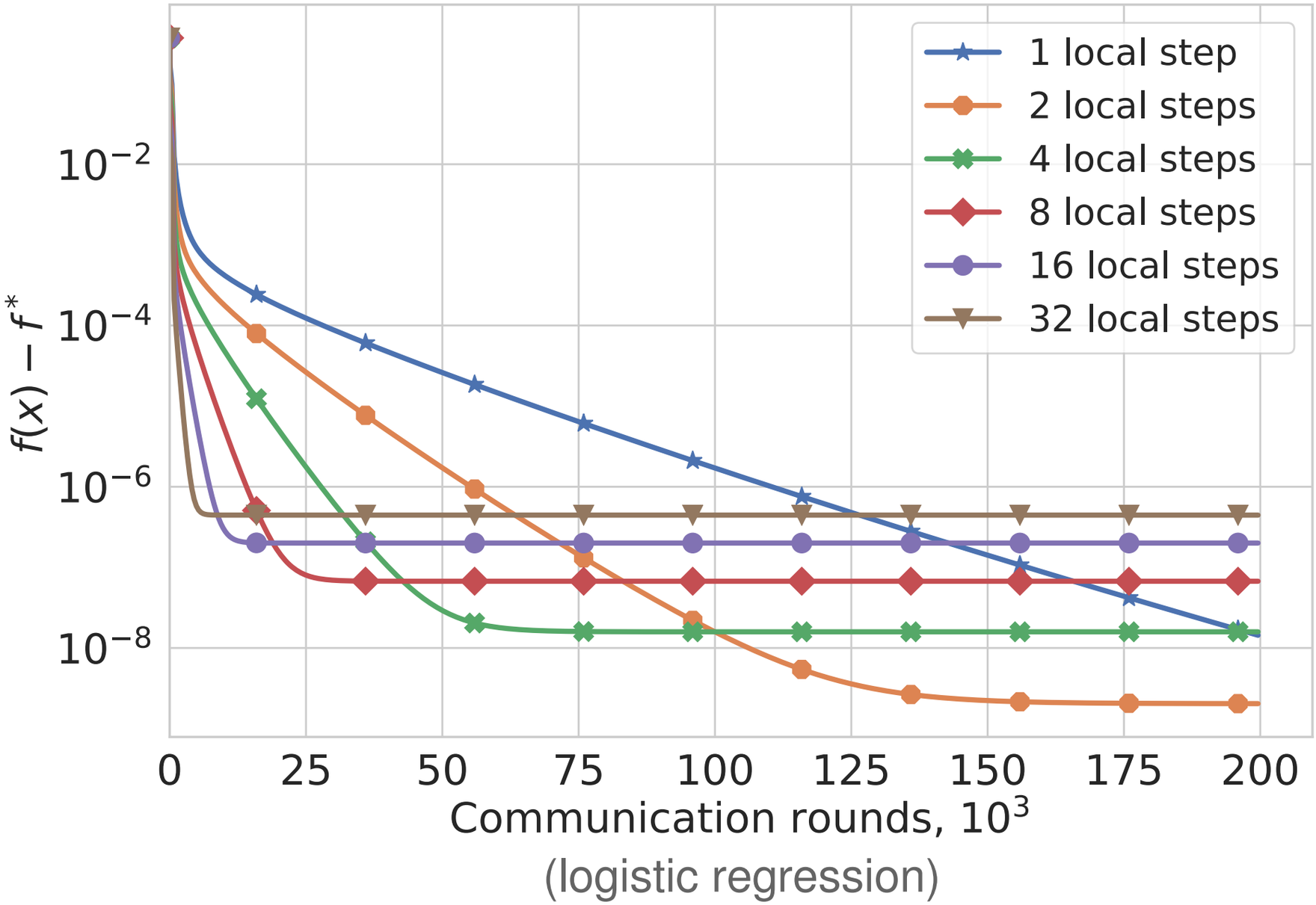
$$x_1^{t+H} = x_1^{t+H-1} - \gamma \nabla f_1(x_1^{t+H-1}) \qquad x_n^{t+H} = x_n^{t+H-1} - \gamma \nabla f_n(x_n^{t+H-1})$$



$$x^{t+H} := \frac{1}{n} \sum_{i=1}^{n} x_i^{t+H}$$

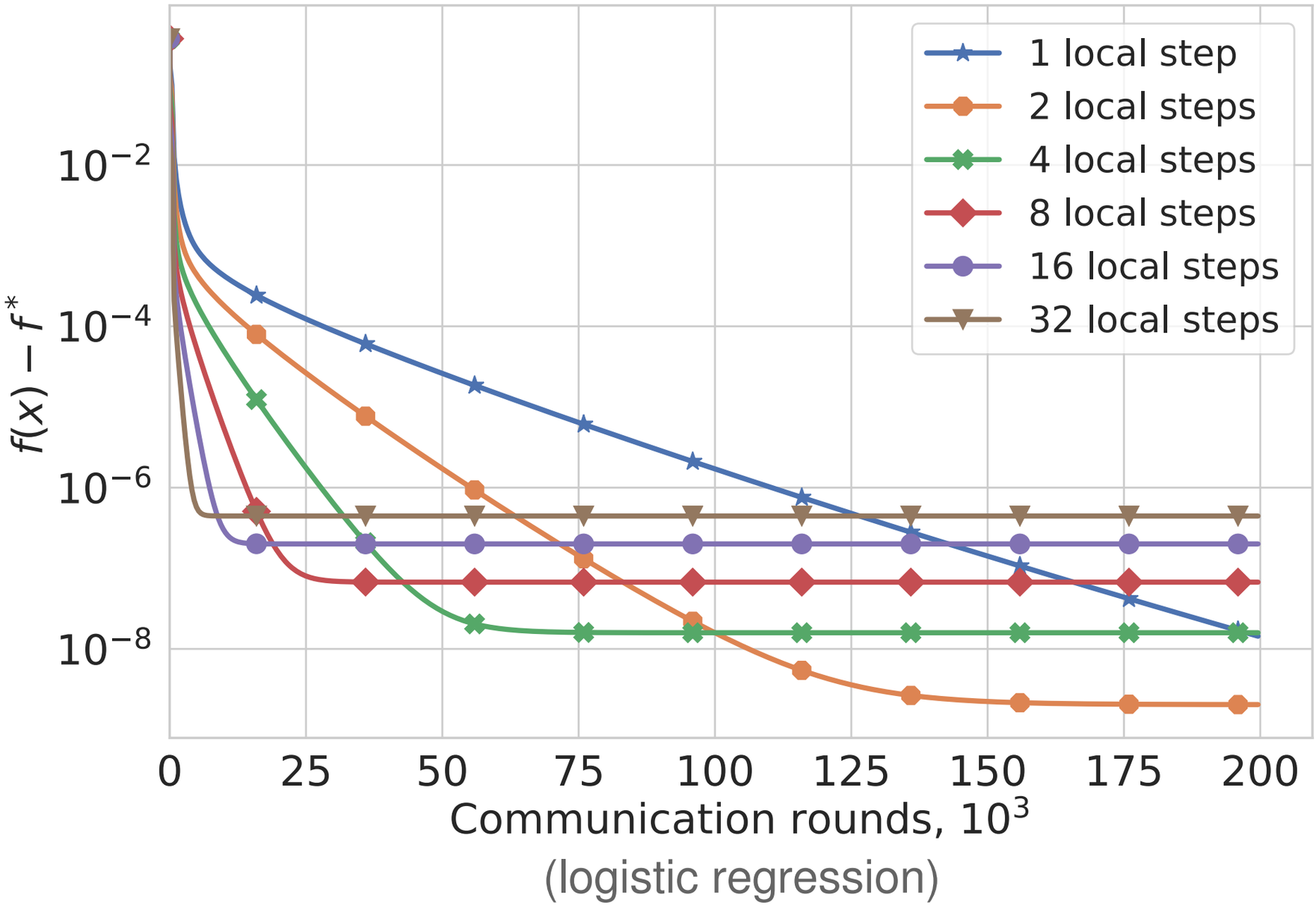
KAUST

Distributed **Local** GD = FedAvg

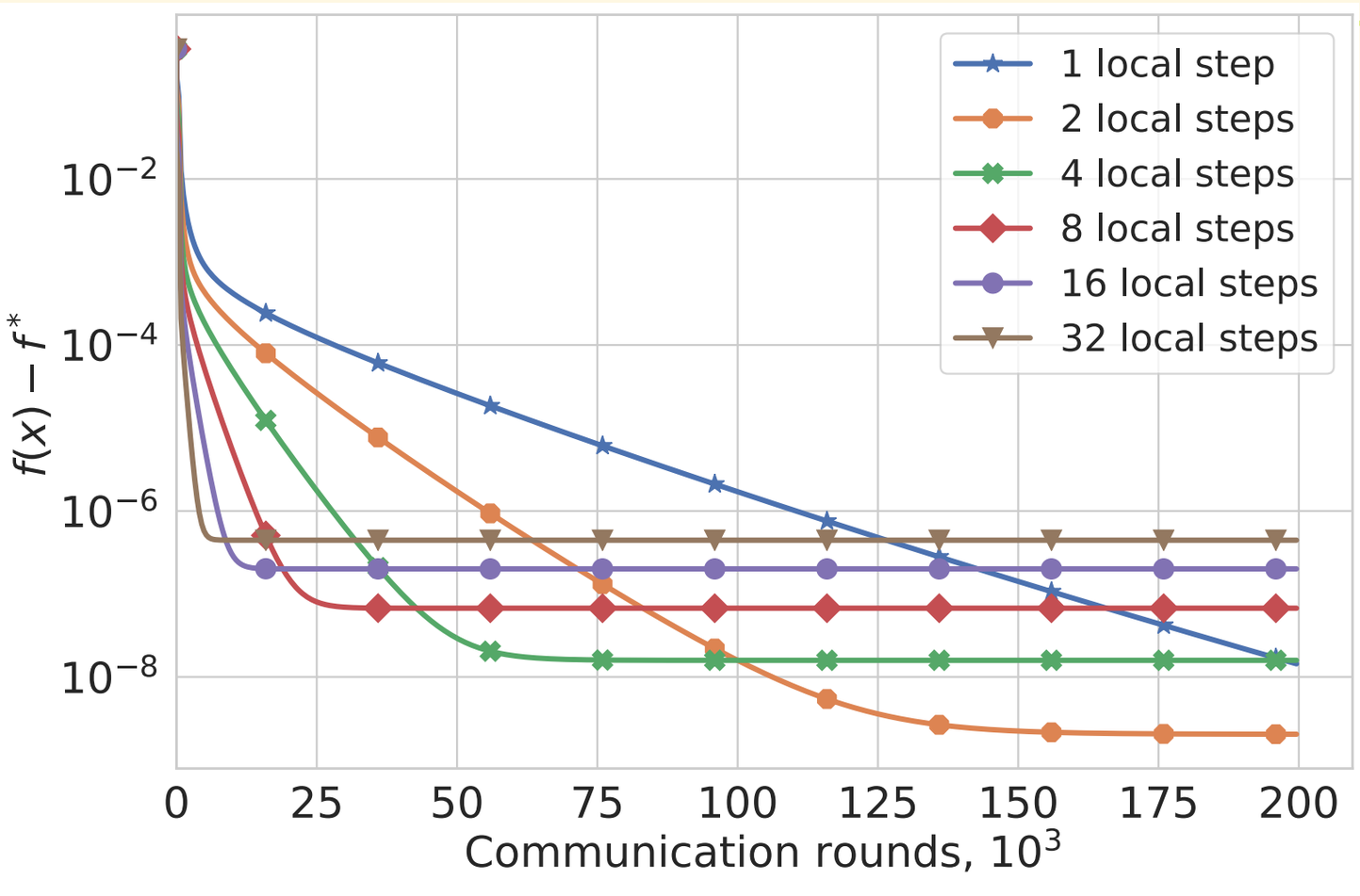


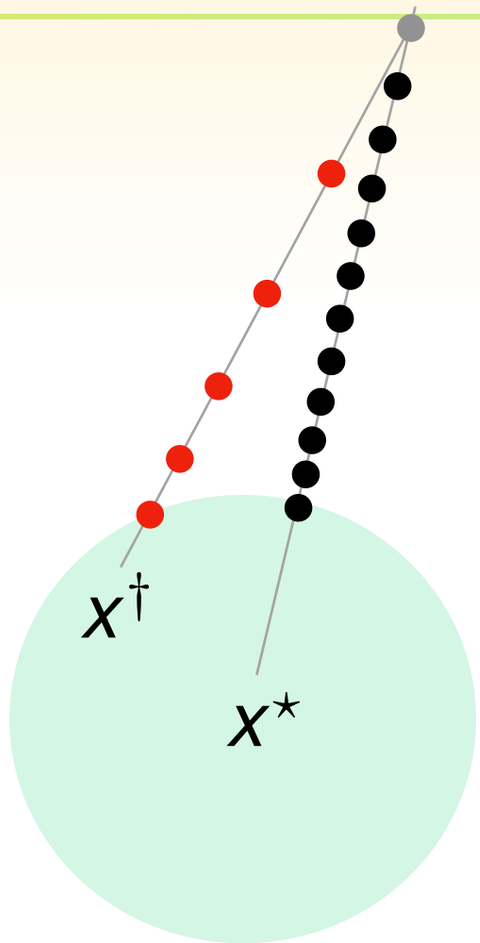
(logistic regression)

Local GD: analysis



Malinovsky, Kovalev, Gasanov, Condat, Richtárik, “From local SGD to local fixed point methods for federated learning,” ICML 2020





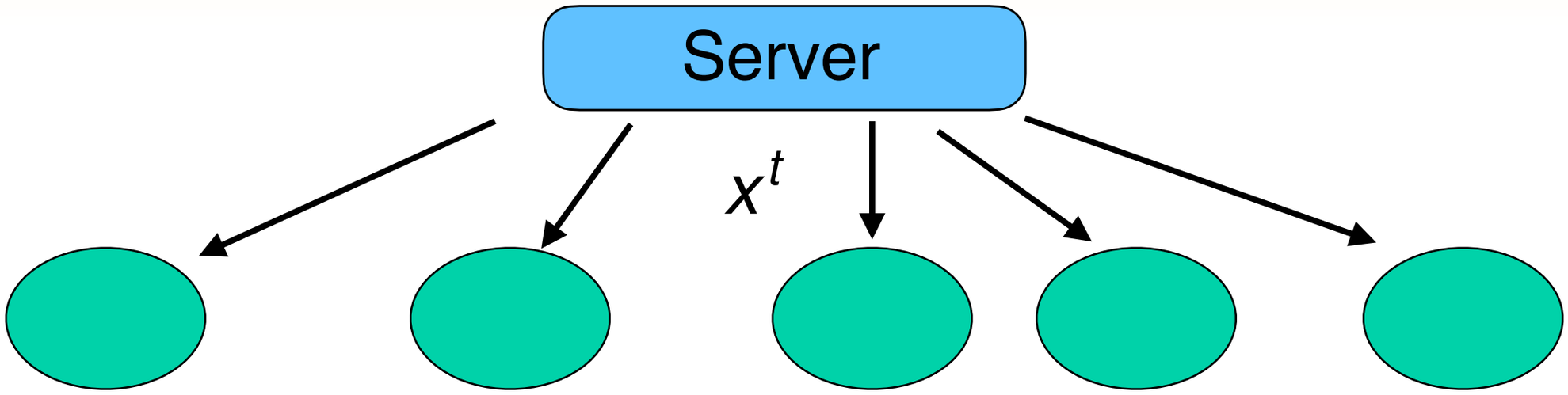
Theorem 2.11 (linear convergence) With $\gamma \in (0, \frac{2}{L+\mu}]$, $(x^{rH})_{r\geq 0}$ converges linearly to $x^\dagger$ with rate ξ^H and

$$\|x^\dagger - x^\star\| \leq S,$$

where

$$\xi = 1 - \gamma\mu, \quad S = \frac{\xi}{1-\xi} \frac{1-\xi^{H-1}}{1-\xi^H} \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x^\star)\|.$$

Scaffnew: **Variance-reduced** local GD



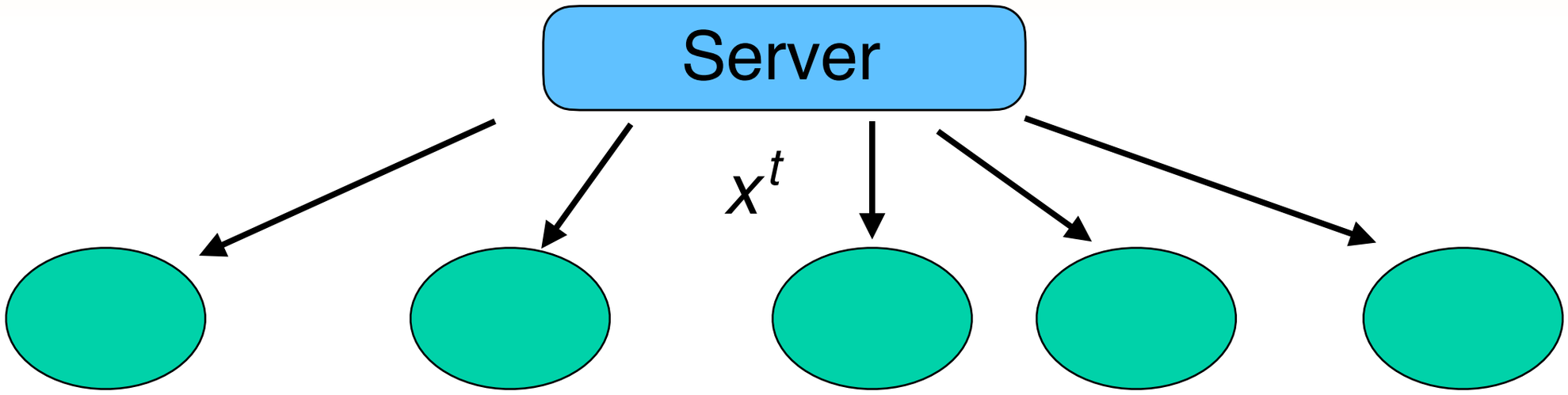
$$\hat{x}_1^t = x_1^t - \gamma \nabla f_1(x_1^t) + \gamma h_1^t \qquad \cdots \qquad \hat{x}_n^t = x_n^t - \gamma \nabla f_n(x_n^t) + \gamma h_n^t$$

Probabilistic Com: $x_i^{t+1} := \begin{cases} \frac{1}{n}\sum_{j=1}^n \hat{x}_j^t & \text{with probability } p \\ \hat{x}_i^t & \text{with probability } 1-p \end{cases}$

Mishchenko, Malinovsky, Stich, Richtárik, “ProxSkip: Yes! Local Gradient Steps Provably Lead to Communication Acceleration!,” ICML 2022

$p = \frac{1}{\sqrt{\kappa}}$ ☞ TotalCom $\mathcal{O}\left(d\sqrt{\kappa}\log\epsilon^{-1}\right)$

Scaffnew: **Variance-reduced** local GD



$$\hat{x}_1^t = x_1^t - \gamma \nabla f_1(x_1^t) + \gamma h_1^t \qquad \cdots \qquad \hat{x}_n^t = x_n^t - \gamma \nabla f_n(x_n^t) + \gamma h_n^t$$

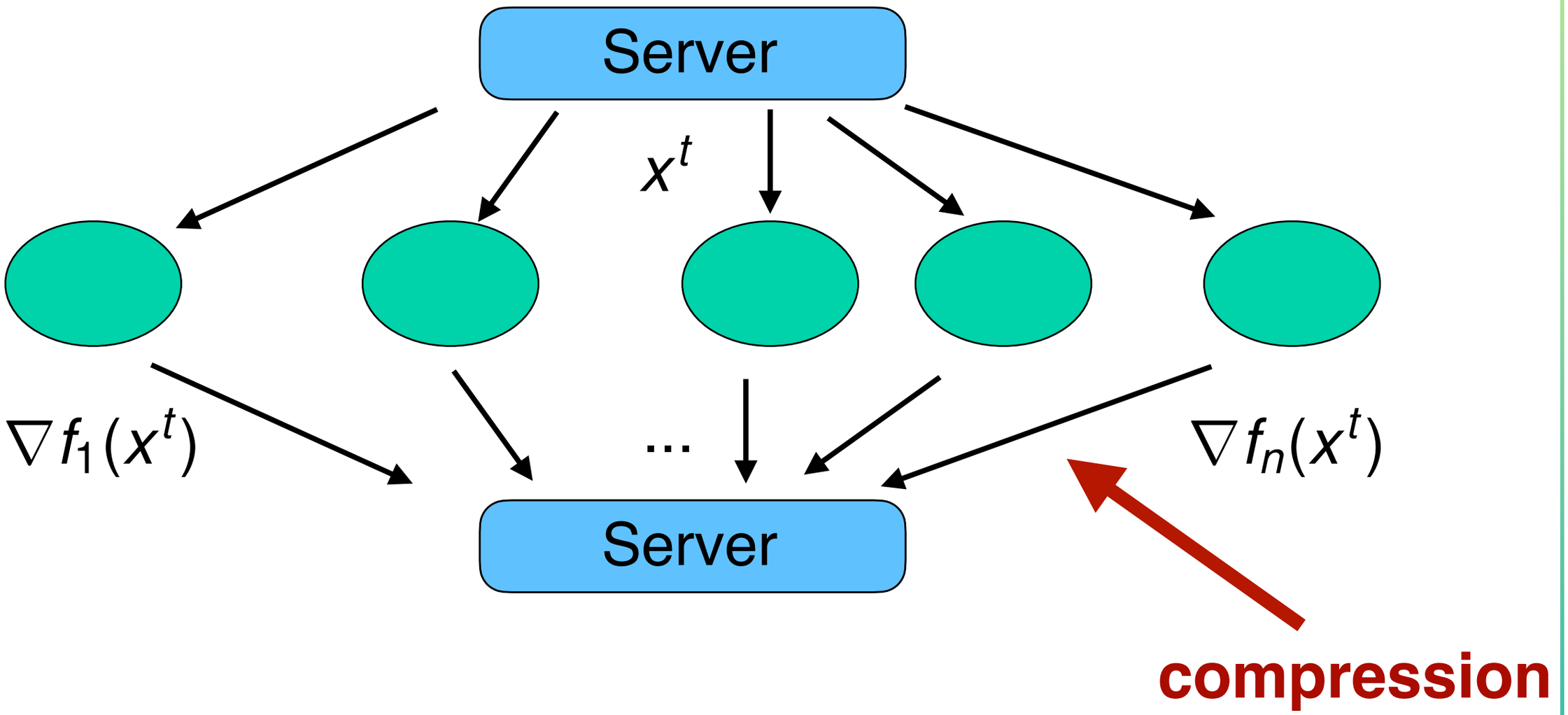
Probabilistic Com: $x_i^{t+1} := \begin{cases} \frac{1}{n}\sum_{j=1}^n \hat{x}_j^t & \text{with probability } p \\ \hat{x}_i^t & \text{with probability } 1-p \end{cases}$

Mishchenko, Malinovsky, Stich, Richtárik, “ProxSkip: Yes! Local Gradient Steps Provably Lead to Communication Acceleration!,” ICML 2022

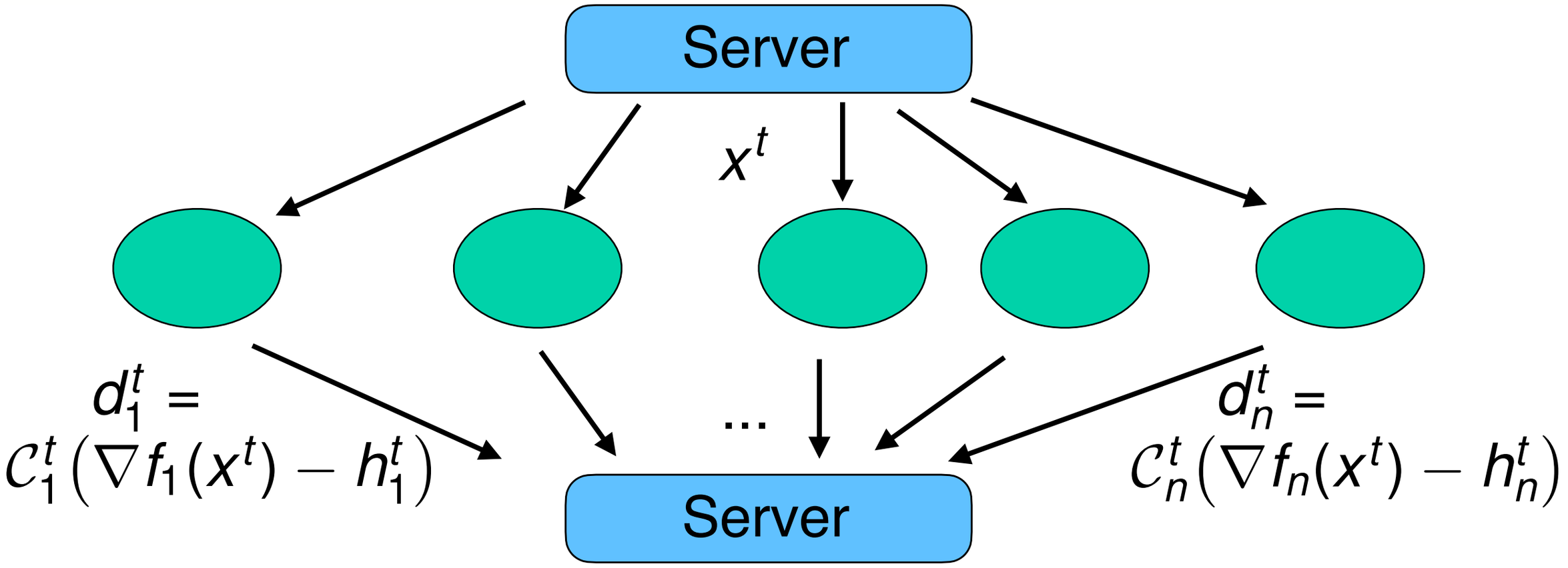
Condat and Richtárik, “RandProx: Primal-Dual Optimization Algorithms with Randomized Proximal Updates,” ICLR 2023

2) Compression

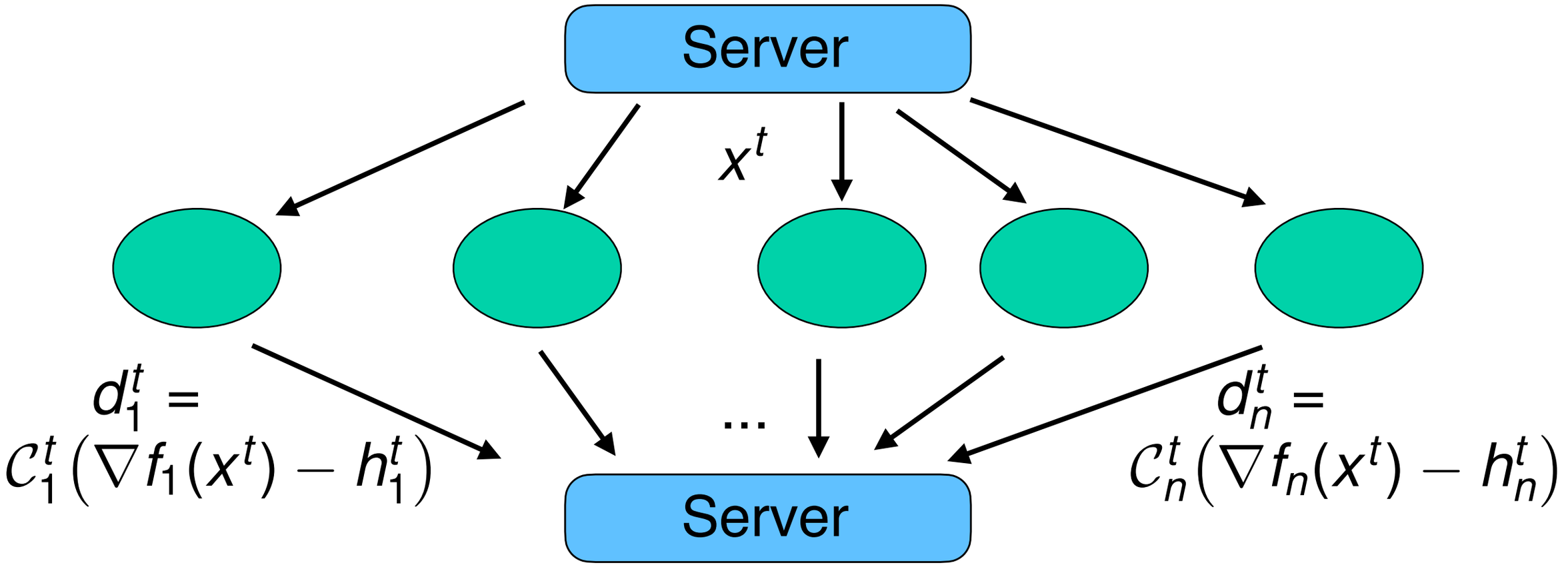
Distributed GD



Distributed GD with compression

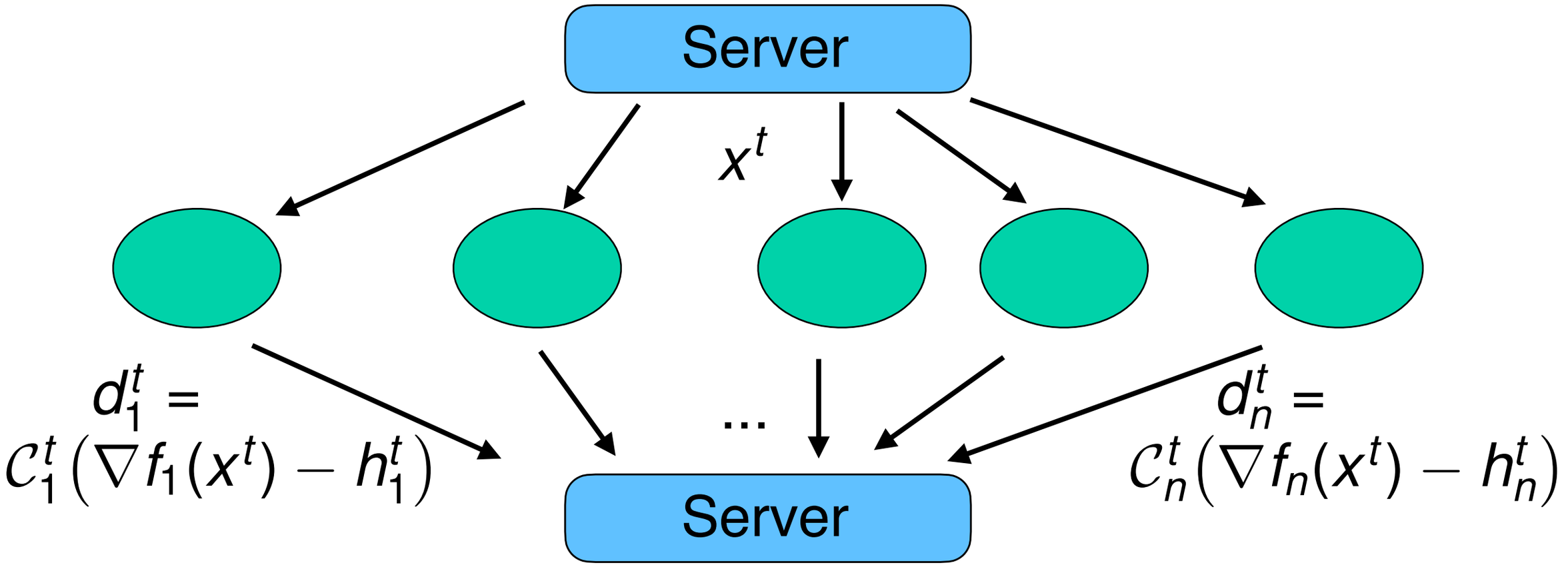


DIANA



$$x^{t+1} := x^t - \gamma g^t \quad \text{with} \quad g^t = h^t + \frac{1}{n}\sum_{i=1}^{n} d_i^t$$

DIANA

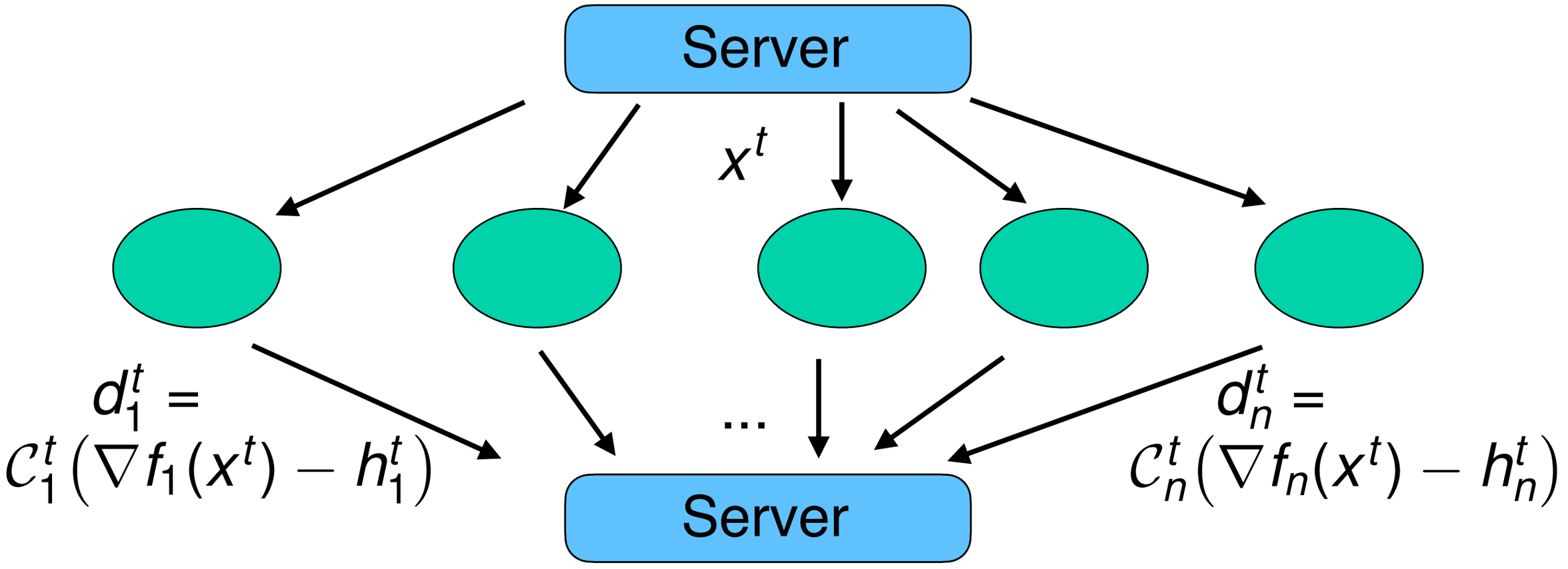


$$x^{t+1} := x^t - \gamma g^t \quad \text{with} \quad g^t = h^t + \frac{1}{n}\sum_{i=1}^{n} d_i^t$$

update of the control variates: $h_i^{t+1} := h_i^t + \lambda d_i^t$

Mishchenko et al., “Distributed Learning with Compressed Gradient Differences,” 2019, published in 2024

DIANA

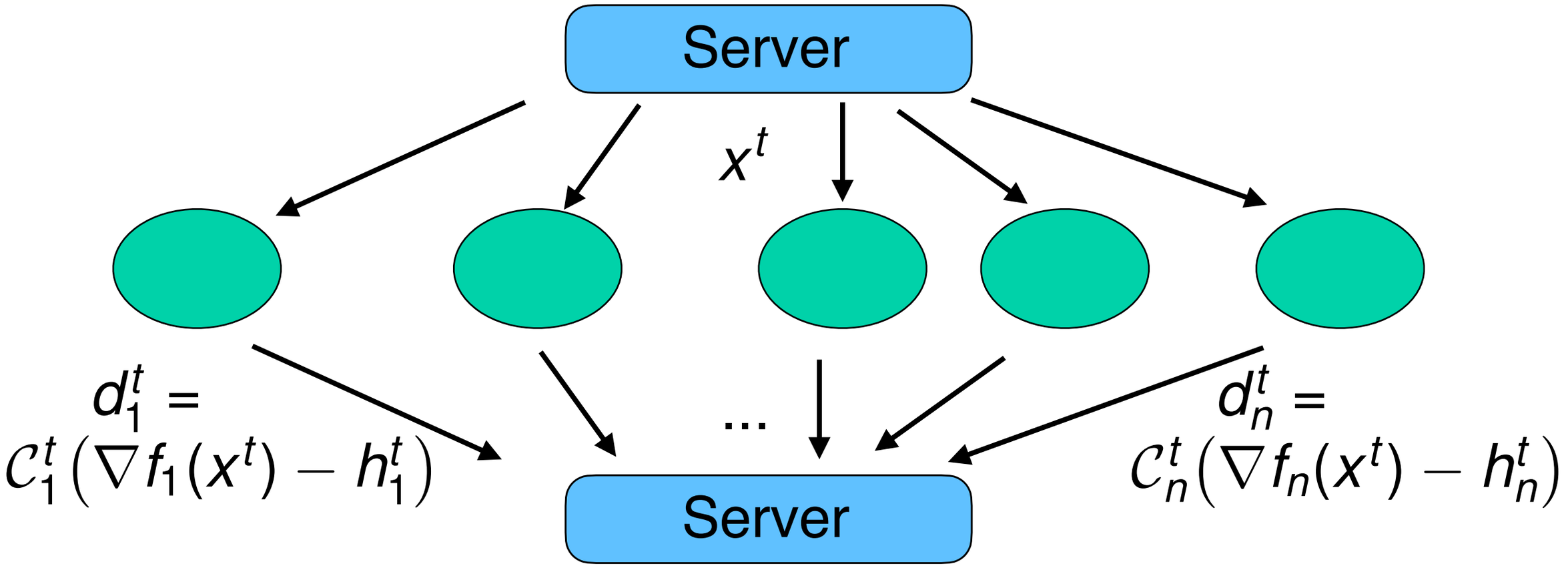


$$x^{t+1} := x^t - \gamma g^t \quad \text{with} \quad g^t = h^t + \frac{1}{n}\sum_{i=1}^n d_i^t$$

update of the control variates: $h_i^{t+1} := h_i^t + \lambda d_i^t$

Condat and Richtárik, “MURANA: A Generic Framework for Stochastic Variance-Reduced Optimization,” MSML 2022

EF-BV



$$x^{t+1} := x^t - \gamma g^t \quad \text{with} \quad g^t = h^t + \frac{\nu}{n}\sum_{i=1}^{n} d_i^t$$

update of the control variates: $h_i^{t+1} := h_i^t + \lambda d_i^t$

Condat, Yi, Richtárik, “EF-BV: A unified theory of error feedback and variance reduction for biased and unbiased compression in distributed optimization,” NeurIPS 2022

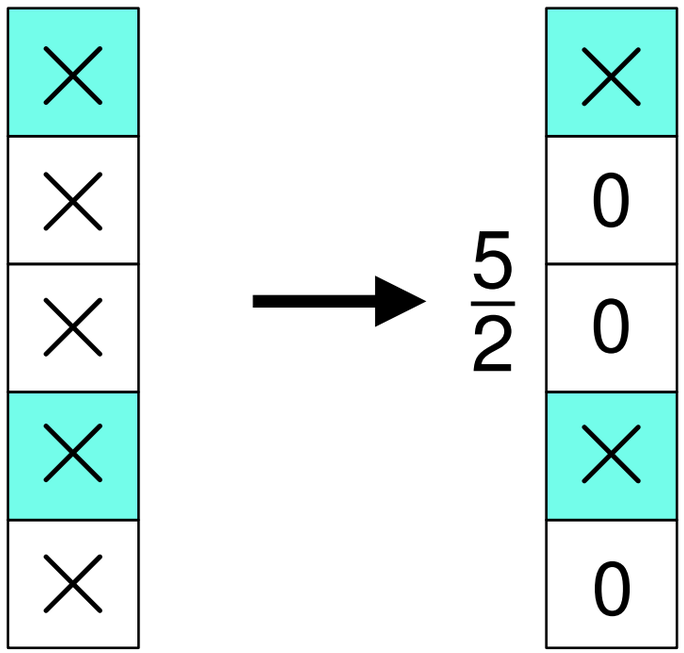
Unbiased random compression

For every $\omega \geq 0$, $\mathbb{U}(\omega)$ is the set of compression operators $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ such that, for every $x \in \mathbb{R}^d$,

- $\mathbb{E}[\mathcal{C}(x)] = x$
- $\mathbb{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq \omega \|x\|^2$

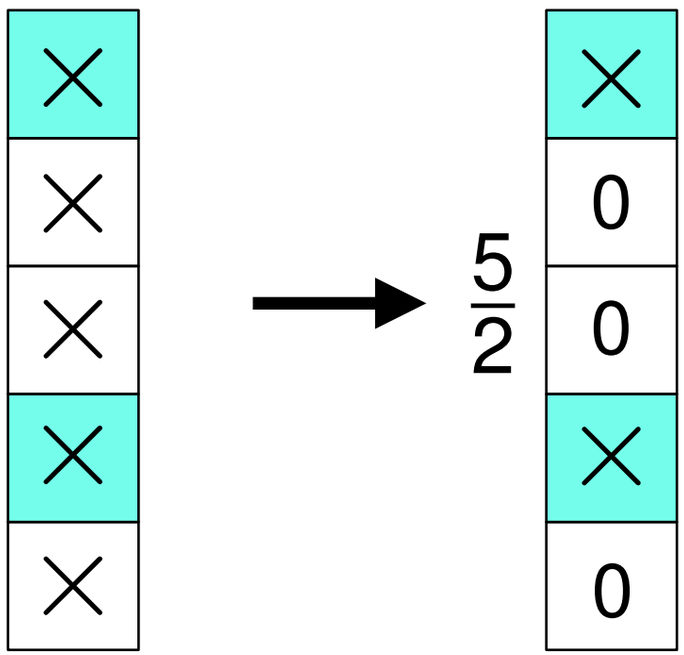
Unbiased random compression

- `rand-`k: k elements out of d chosen unif. at random and scaled by $\frac{d}{k}$, other ones set to 0.



Unbiased random compression

- `rand-`k: k elements out of d chosen unif. at random and scaled by $\frac{d}{k}$, other ones set to 0.



$$\texttt{rand-}k \in \mathbb{U}\left(\frac{d}{k} - 1\right)$$

Unbiased random compression

- `rand-`k: k elements out of d chosen unif. at random and scaled by $\frac{d}{k}$, other ones set to 0.

- quantization

$$\mathcal{C}(1.2) = \begin{cases} 1 \text{ with probability } \frac{4}{5} \\ 2 \text{ with probability } \frac{1}{5} \end{cases}$$

Unbiased random compression

- `rand-`k: k elements out of d chosen unif. at random and scaled by $\frac{d}{k}$, other ones set to 0.

- quantization

$$\mathcal{C}(1.2) = \begin{cases} 1 \text{ with probability } \frac{4}{5} \\ 2 \text{ with probability } \frac{1}{5} \end{cases}$$

$$\mathcal{C} \in \mathbb{U}\left(\frac{1}{8}\right)$$

Unbiased random compression

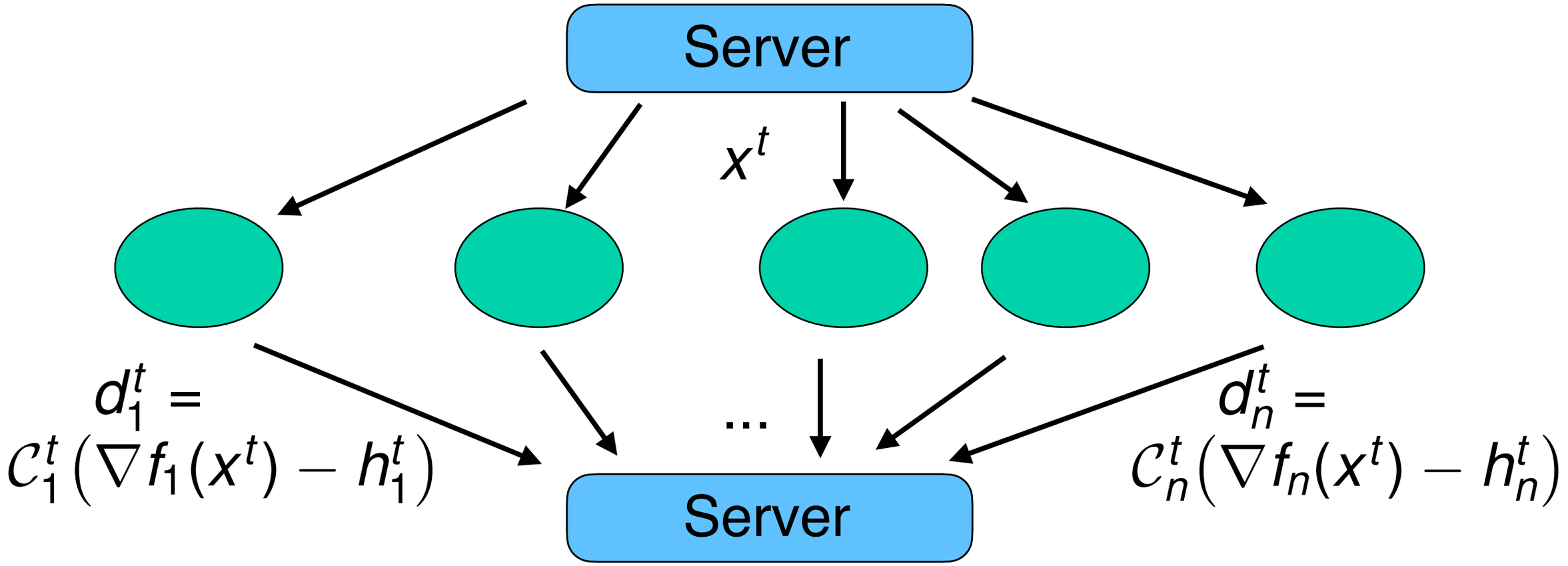
- `rand-`k: k elements out of d chosen unif. at random and scaled by $\frac{d}{k}$, other ones set to 0.

- quantization

$$\mathcal{C}(1.2) = \begin{cases} 1 \text{ with probability } \frac{4}{5} \\ 2 \text{ with probability } \frac{1}{5} \end{cases}$$

Albasyoni, Safaryan, Condat, Richtárik, “Optimal Gradient Compression for Distributed and Federated Learning,” 2020

Distributed GD with compression



DIANA with independent rand-1 compressors

☞ Uplink communication complexity (UpCom):

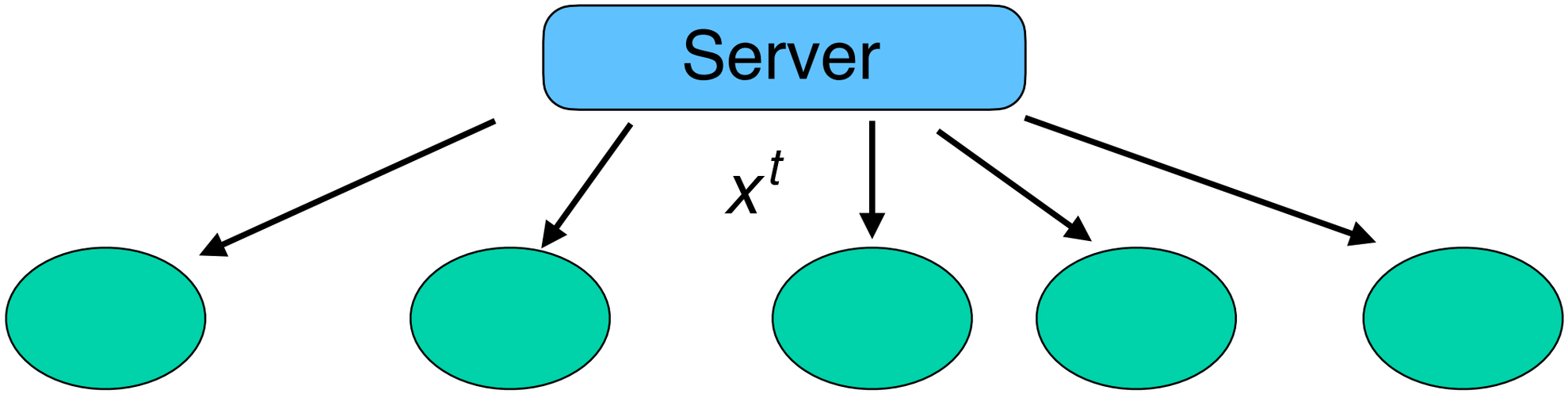
$$\mathcal{O}\left(\left(\frac{d\kappa}{n} + \kappa + d\right)\log \epsilon^{-1}\right)$$

Local Training

+

Compression

CompressedScaffnew

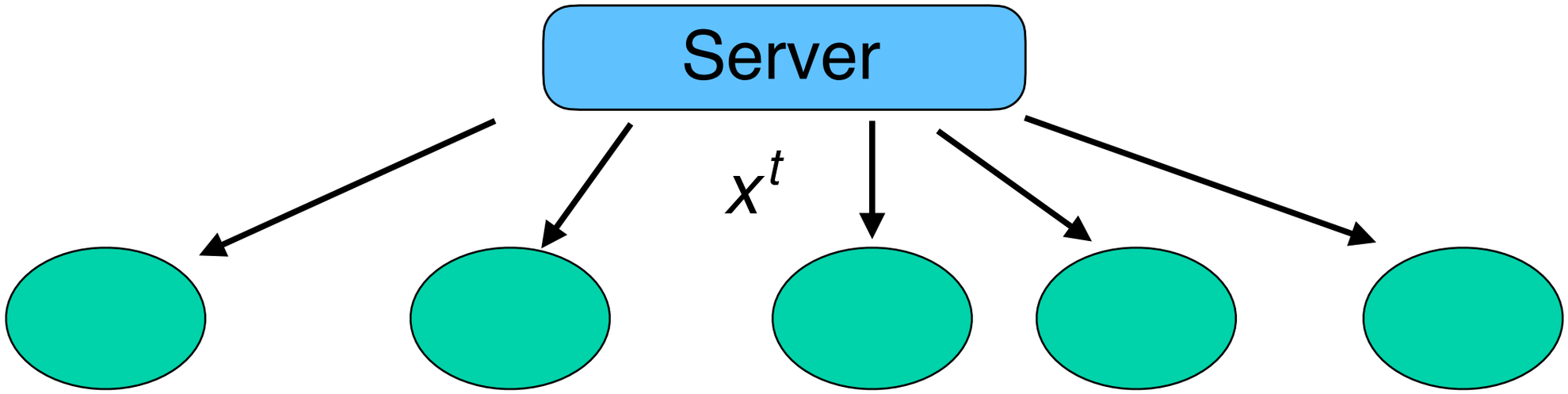


$$\hat{x}_1^t = x_1^t - \gamma \nabla f_1(x_1^t) + \gamma h_1^t \qquad \cdots \qquad \hat{x}_n^t = x_n^t - \gamma \nabla f_n(x_n^t) + \gamma h_n^t$$

$$x_i^{t+1} := \begin{cases} \frac{1}{n} \sum_{j=1}^n \mathcal{C}_i^t(\hat{x}_j^t) & \text{with probability } p \\ \hat{x}_i^t & \text{with probability } 1-p \end{cases}$$

Condat, Agarsky, Richtárik, “Provably Doubly Accelerated Federated Learning: The First Theoretically Successful Combination of Local Training and Compressed Communication,” 2022

CompressedScaffnew



$$\hat{x}_1^t = x_1^t - \gamma \nabla f_1(x_1^t) + \gamma h_1^t \qquad \cdots \qquad \hat{x}_n^t = x_n^t - \gamma \nabla f_n(x_n^t) + \gamma h_n^t$$

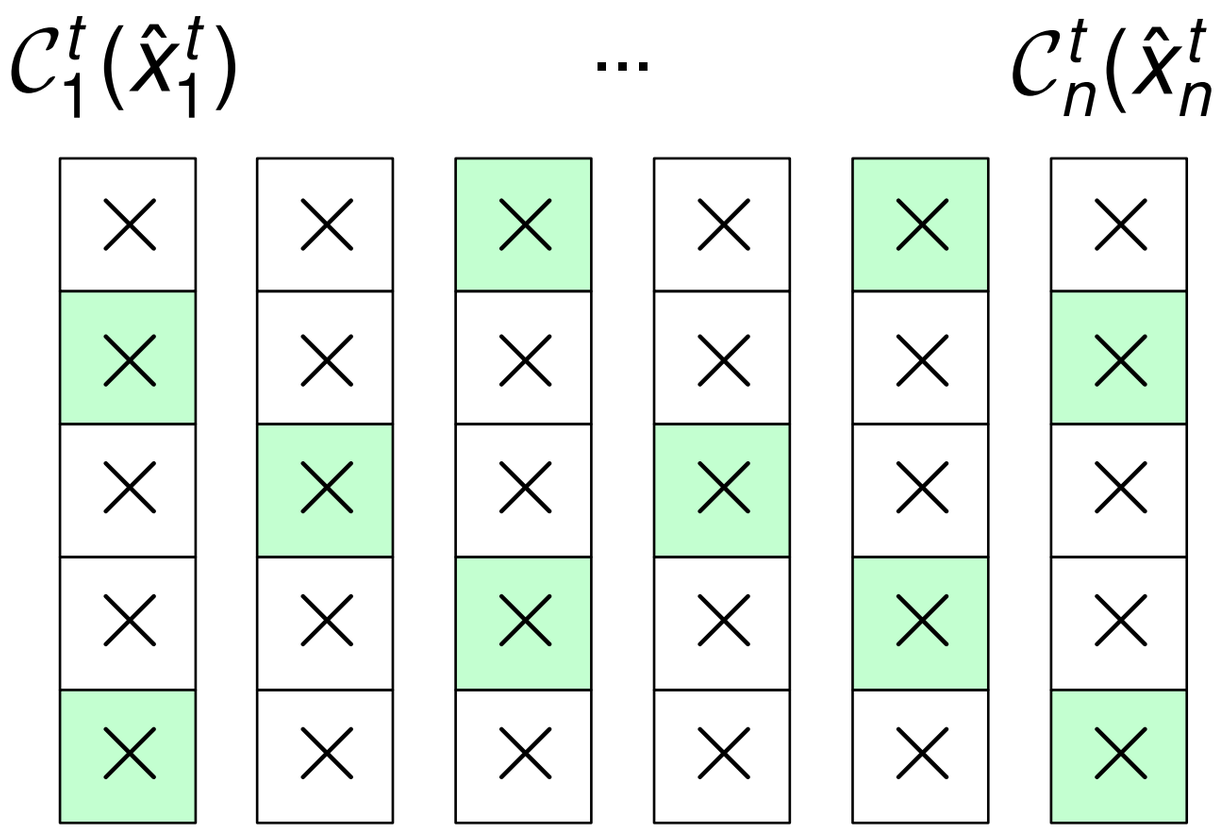
$$x_i^{t+1} := \begin{cases} \frac{1}{n} \sum_{j=1}^n \mathcal{C}_i^t(\hat{x}_j^t) & \text{with probability } p \\ \hat{x}_i^t & \text{with probability } 1-p \end{cases}$$

We don't compress differences, so we need

$$\hat{x}_1^t = \cdots = \hat{x}_n^t \;\Rightarrow\; \mathcal{C}_i^t \equiv \mathrm{Id}$$

CompressedScaffnew

correlated rand-k compressors



- s communicated values per coordinate
- $k = \frac{sd}{c}$ communicated values per active client

CompressedScaffnew

Uplink communication complexity (UpCom):

- GD: $\tilde{\mathcal{O}}\left(d\kappa\right)$
- Scaffnew: $\tilde{\mathcal{O}}\left(d\sqrt{\kappa}\right)$
- DIANA: $\tilde{\mathcal{O}}\left(\frac{d\kappa}{n}+\kappa+d\right)$
- **CompressedScaffnew:** $\tilde{\mathcal{O}}\left(\frac{d\sqrt{\kappa}}{\sqrt{n}}+\sqrt{d\kappa}+d\right)$

TAMUNA

parameters: stepsizes $\gamma > 0$, $\eta > 0$;
number of participating clients $c \in \{2, \dots, n\}$
sparsity index $s \in \{2, \dots, n\}$ for compression
for $r = 0, 1, \dots$ (rounds) **do**
  choose a subset $\Omega^r \in [n]$ of size c
  choose the number of local steps L^r
  for clients $i \in \Omega^r$, in parallel, **do**
    $x^{r,0} := \bar{x}^r$
    for $l = 0, \dots, L^r$ **do**
      $x_i^{r,l} := x_i - \gamma \nabla f_i^t(x_i^{r,l}) + \gamma h_i^r$
    end for
    send $v_i^r := \mathcal{C}_i^r\left(x_i^{r,L^r}\right)$ to server // UpCom
  end for
  at server: $\bar{x}^{r+1} := \frac{1}{s} \sum_{i \in \Omega^r} v_i^r$ // model update
  $\bar{x}^{r+1}$ is sent to clients $i \in \Omega^r \cup \Omega^{r+1}$ // DownCom
  for clients $i \in \Omega^r$, in parallel, **do** // update of
    $h_i^{r+1} := h_i^r + \frac{\eta}{\gamma}\left(\mathcal{C}_i^r(\bar{x}^{r+1}) - v_i^r\right)$ control variates
  end for
  for clients $i \notin \Omega^r$, in parallel, **do**
    $h_i^{r+1} := h_i^r$ // idle clients
  end for
end for

Condat, Agarský, Malinovsky, Richtárik, "TAMUNA: Doubly accelerated federated learning with local training, compression, and partial participation," 2023

TAMUNA

parameters: stepsizes $\gamma > 0$, $\eta > 0$;
number of participating clients $c \in \{2, \dots, n\}$
sparsity index $s \in \{2, \dots, n\}$ for compression
for $r = 0, 1, \dots$ (rounds) **do**
  choose a subset $\Omega^r \in [n]$ of size c
  choose the number of local steps L^r
  for clients $i \in \Omega^r$, in parallel, **do**
    $x^{r,0} := \bar{x}^r$
    for $l = 0, \dots, L^r$ **do**
      $x_i^{r,l} := x_i - \gamma \nabla f_i^t(x_i^{r,l}) + \gamma h_i^r$
    end for
    send $v_i^r := \mathcal{C}_i^r\left(x_i^{r,L^r}\right)$ to server
end for
at server: $\bar{x}^{r+1} := \frac{1}{s} \sum_{i \in \Omega^r} v_i^r$
$\bar{x}^{r+1}$ is sent to clients $i \in \Omega^r \cup \Omega^{r+1}$
for clients $i \in \Omega^r$, in parallel, **do**
  $h_i^{r+1} := h_i^r + \frac{\eta}{\gamma}\left(\mathcal{C}_i^r(\bar{x}^{r+1}) - v_i^r\right)$
end for

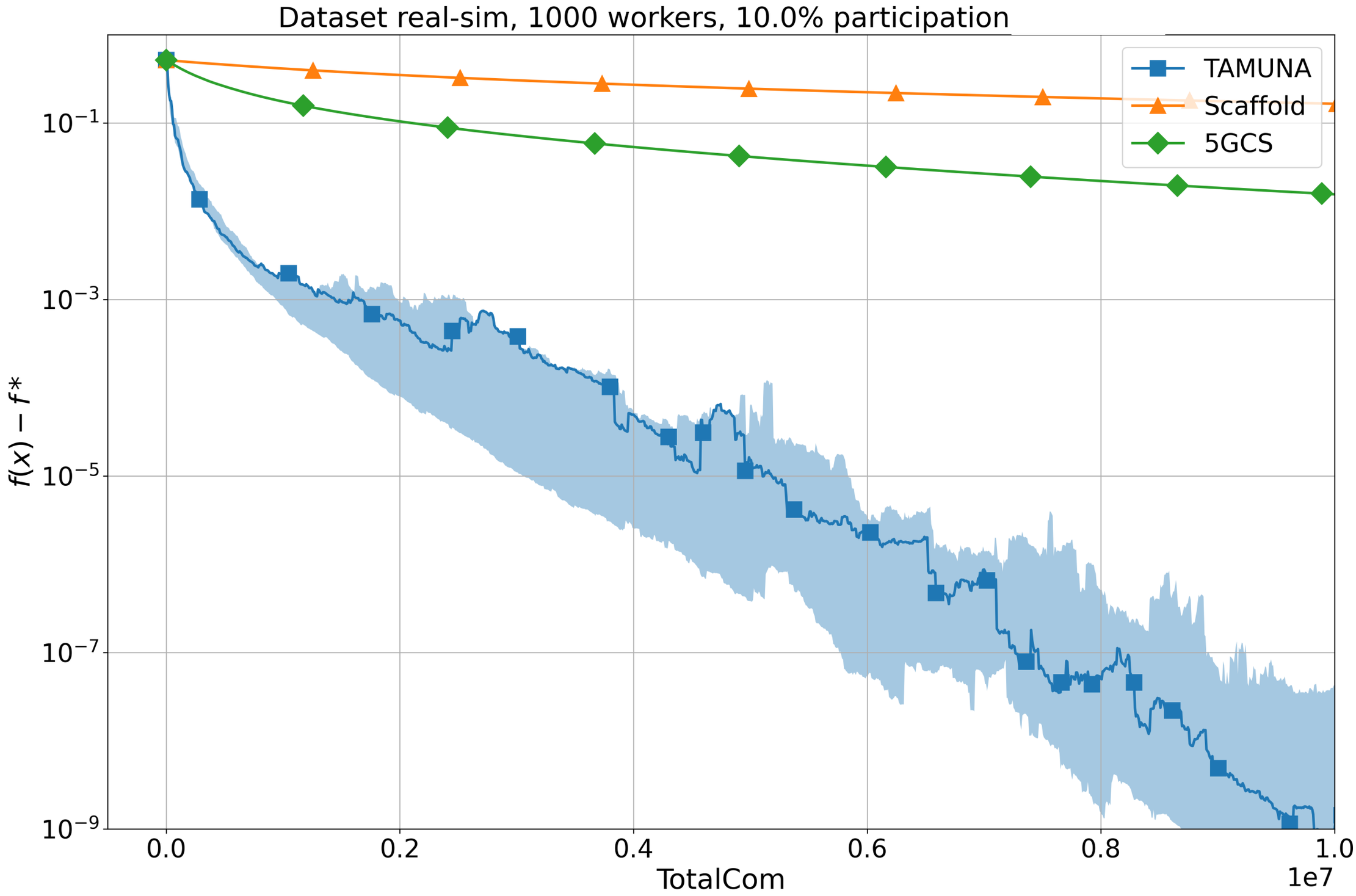
Condat, Agarský, Malinovsky, Richtárik, "TAMUNA: Doubly accelerated federated learning with local training, compression, and partial participation," 2023

$$\gamma \approx \frac{1}{L},\ s \approx \max\left(2, \frac{c}{d}\right),$$
$$L^r \approx \max\left(\sqrt{\frac{s\kappa}{n}}, 1\right),$$

☞ UpCom $\tilde{\mathcal{O}}\left(\sqrt{d\kappa}\sqrt{\frac{n}{c}} + d\sqrt{\kappa}\frac{\sqrt{n}}{c} + d\frac{n}{c}\right)$

KAUST

Experiment: logistic regression



Local Training

+

Compression

LoCoDL

Condat, Maranjyan, and Richtárik, “LoCoDL: Communication-Efficient Distributed Learning with Local Training and Compression,” ICLR 2025

Optimization problem

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \ \frac{1}{n} \sum_{i=1}^{n} f_i(x) + g(x)$$

All f_i and g are μ-strongly convex and L-smooth, for some $L \geq \mu > 0$

Client i makes calls to ∇f_i and ∇g

Reformulation

$$\underset{x_1,\ldots,x_n,y\in\mathbb{R}^d}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n} f_i(x_i) + g(y) \quad \text{s.t.} \quad x_1 = \cdots = x_n = y$$

Optimality conditions:

- $x_1 = \cdots = x_n = y$
- $0 = \nabla f_i(x_i) - u_i, \ \forall i \in [n]$
- $0 = \nabla g(y) - v$
- $0 = u_1 + \cdots + u_n + nv$

LoCoDL

input: stepsizes $\gamma > 0$, $\chi > 0$, $\rho > 0$; probability $p \in (0,1]$; $\omega \geq 0$; initial estimates $x_1^0, \ldots, x_n^0, y^0 \in \mathbb{R}^d$ and control variates $u_1^0, \ldots, u_n^0, v^0 \in \mathbb{R}^d$ such that $\frac{1}{n}\sum_{i=1}^n u_i^0 + v^0 = 0$.

for $t = 0, 1, \ldots$ **do**

  for $i = 1, \ldots, n$, at clients in parallel, **do**

    $\hat{x}_i^t := x_i^t - \gamma \nabla f_i(x_i^t) + \gamma u_i^t$

    $\hat{y}^t := y^t - \gamma \nabla g(y^t) + \gamma v^t$ // identical copies at the clients

    flip a coin $\theta^t \in \{0,1\}$ with $\mathrm{Prob}(\theta^t = 1) = p$

    if $\theta^t = 1$ **then**

      $d_i^t := \mathcal{C}_i^t\left(\hat{x}_i^t - \hat{y}^t\right)$

      send d_i^t to the server

      at server: aggregate $\bar{d}^t := \frac{1}{2n}\sum_{j=1}^n d_j^t$ and send $\bar{d}^t$ to all clients

      $x_i^{t+1} := (1-\rho)\hat{x}_i^t + \rho(\hat{y}^t + \bar{d}^t)$

      $u_i^{t+1} := u_i^t + \frac{p\chi}{\gamma(1+2\omega)}\left(\bar{d}^t - d_i^t\right)$

      $y^{t+1} := \hat{y}^t + \rho \bar{d}^t$

      $v^{t+1} := v^t + \frac{p\chi}{\gamma(1+2\omega)}\bar{d}^t$

    else

      $x_i^{t+1} := \hat{x}_i^t$, $y^{t+1} = \hat{y}^t$, $u_i^{t+1} := u_i^t$, $v^{t+1} := v^t$

    end if

  end for

end for

Linear convergence

Theorem (linear convergence of LoCoDL). In LoCoDL, suppose that $\gamma \in (0, \frac{2}{L})$, $2\rho - \rho^2(1 + \omega_{\text{av}}) - \chi \geq 0$. For every $t \geq 0$, define the Lyapunov function

$$\Psi^t := \frac{1}{\gamma}\left(\sum_{i=1}^{n} \left\|x_i^t - x^\star\right\|^2 + n\left\|y^t - x^\star\right\|^2\right) + \frac{\gamma(1+2\omega)}{p^2\chi}\left(\sum_{i=1}^{n} \left\|u_i^t - u_i^\star\right\|^2 + n\left\|v^t - v^\star\right\|^2\right),$$

where $v^\star := \nabla g(x^\star)$ and $u_i^\star := \nabla f_i(x^\star)$. Then LoCoDL converges linearly: for every $t \geq 0$,

$$\mathbb{E}\left[\Psi^t\right] \leq \tau^t \Psi^0, \quad \text{where} \quad \tau := \max\left((1-\gamma\mu)^2, (1-\gamma L)^2, 1 - \frac{p^2\chi}{1+2\omega}\right) < 1.$$

(+ almost sure convergence of all variables)

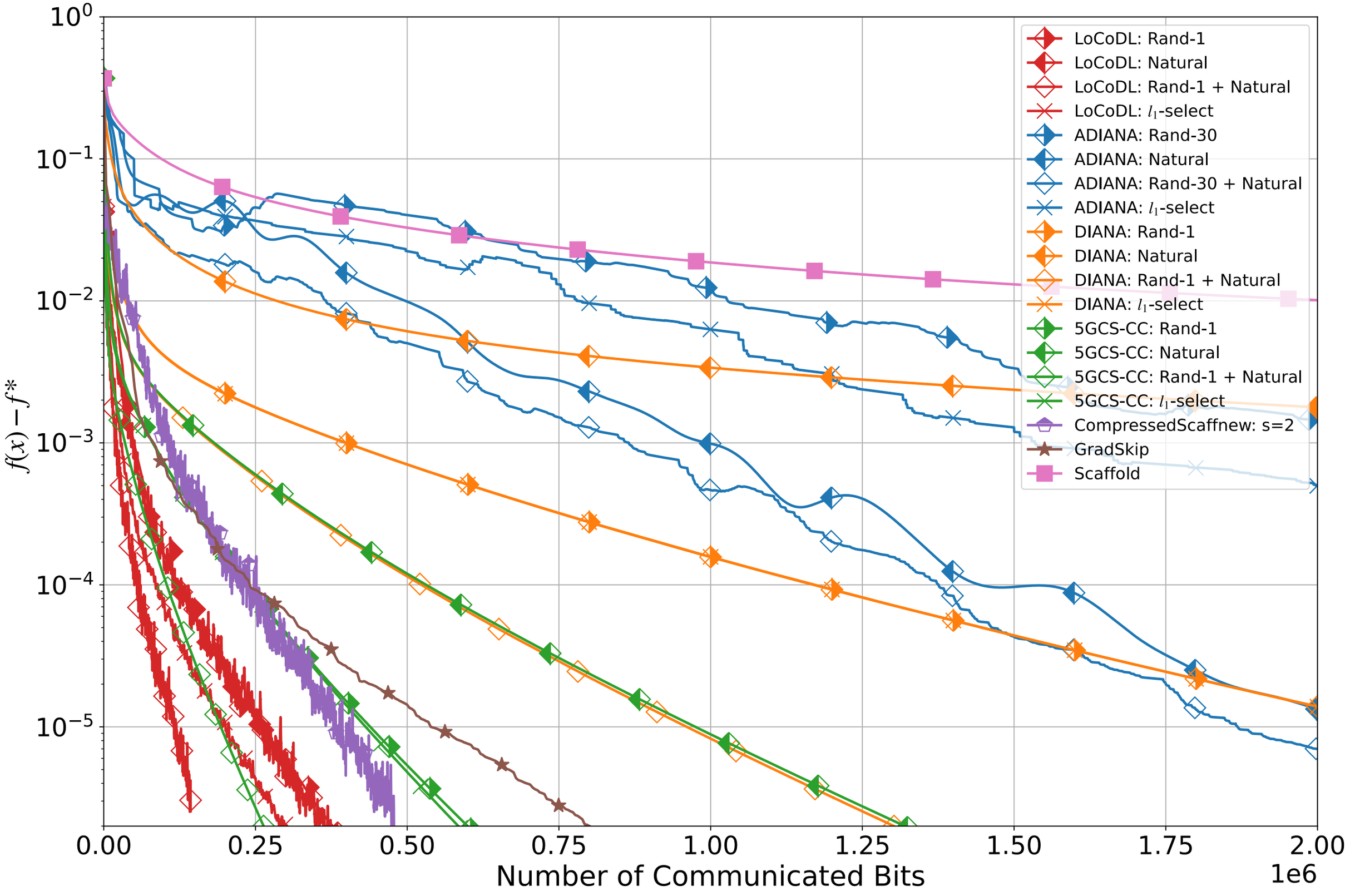
Linear convergence

Corollary. In LoCoDL, suppose that the compressors $\mathcal{C}_i^t$ are independent `rand`-k compressors with $k = \lceil \frac{d}{n} \rceil$, that $\gamma = \Theta(\frac{1}{L})$, $\chi = \rho = \frac{n}{n-1+d/k}$, and

$$p = \min\left(\sqrt{\frac{dk(n-1)+d^2}{nk^2\kappa}}, 1\right).$$

Then the uplink communication complexity in number of reals of LoCoDL is

$$\mathcal{O}\left(\left(\sqrt{d\kappa} + \frac{d\sqrt{\kappa}}{\sqrt{n}} + d\right)\log\epsilon^{-1}\right).$$



Logistic regression with the ‘a5a’ dataset of LibSVM. $d = 122$, $n = 288$

Conclusion

Local training & **compression** are efficient **communication acceleration** mechanisms that can be **combined**.

Future work

- bidirectional compression
- stochastic gradients
- nonconvex functions

…