Online Distributed Optimization with Efficient Communication via Temporal Similarity

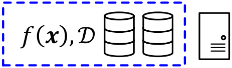
Juncheng Wang*, Ben Liang*, Min Dong†, Gary Boudreau‡, and Ali Afana‡

*Department of Electrical and Computer Engineering, University of Toronto, Canada
†Department of Electrical, Computer and Software Engineering, Ontario Tech University, Canada
‡Ericsson Canada Inc., Canada

May 18, 2023

IEEE INFOCOM 2023

Centralized Optimization (e.g., Supervised Learning)



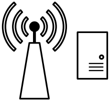
Central server

Global loss and data

$$\min_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} l\left(\boldsymbol{x}; \boldsymbol{u}^i, v^i\right)$$

- $\mathbf{x}$: model parameters
- $\mathbf{u}^i$: features of i-th data sample
- v^i: label of i-th data sample
- $\mathcal{D}$: set of all data samples
- $l(\cdot)$: per-sample loss function

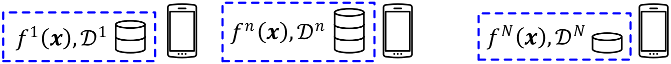
Distributed Optimization (e.g., Federated Learning)



Central server

$$\min_{\mathbf{x}} \; f(\mathbf{x}) = \sum_{n=1}^{N} w^n f^n(\mathbf{x}) = \frac{1}{|\mathcal{D}|} \sum_{n=1}^{N} \sum_{i=1}^{|\mathcal{D}^n|} l\left(\mathbf{x}; \mathbf{u}^{n,i}, v^{n,i}\right)$$

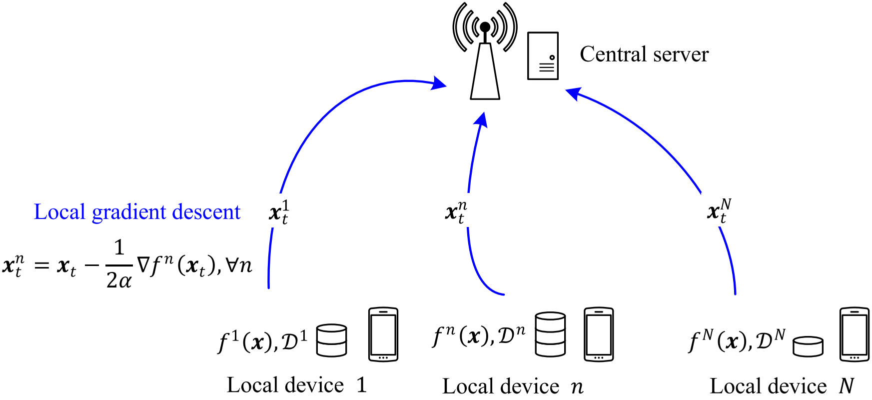
Local loss and data



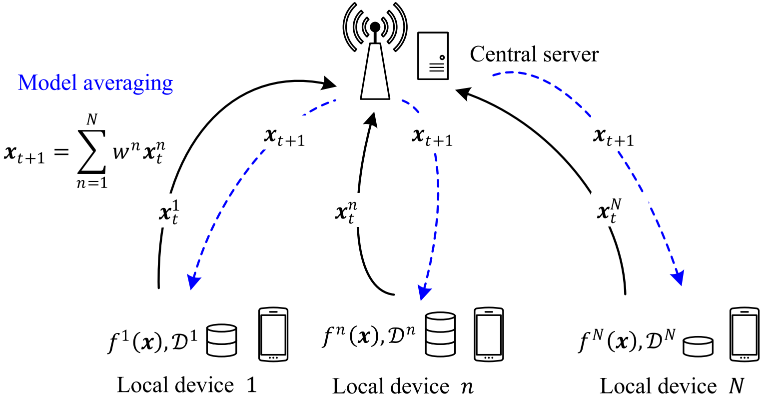
Local device 1   Local device n Local device N

- $f^n(\mathbf{x})$: local lost function of n-th device
- w^n: weight of n-th device

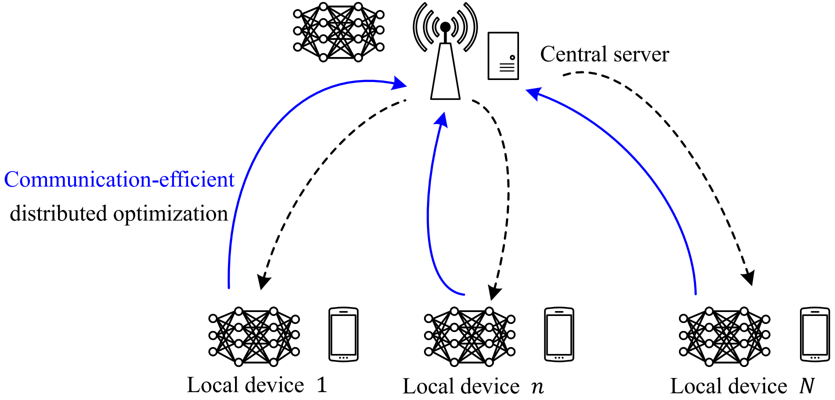
FL Algorithm



FL Algorithm



Concern #1: Communication Efficiency



Reduction of Communication Overhead

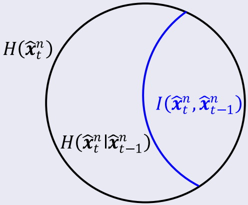
- Quantization:

E.g.,

$$\hat{x}_t^{n,i} = x_{\max}\,\mathrm{sign}(x_t^{n,i})\left\lfloor \frac{|x|}{x_{\max}}(2^b - 1) + \frac{1}{2} \right\rfloor$$

- b: quantization bit length
- $x_{\max}$: maximum decision value

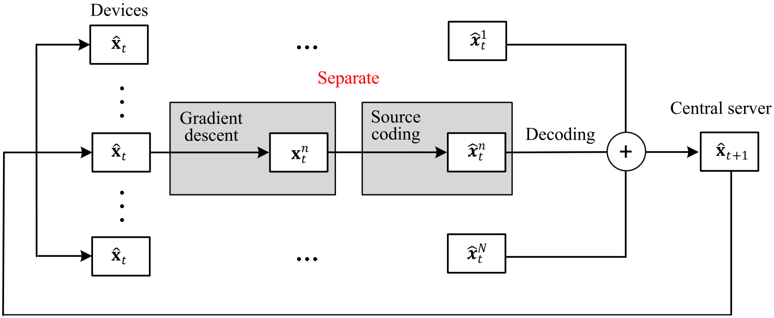
- Conditional entropy coding:



Mutual information $I(\hat{\mathbf{x}}_t^n, \hat{\mathbf{x}}_{t-1}^n)$ is high due to correlation in time.

Existing Works

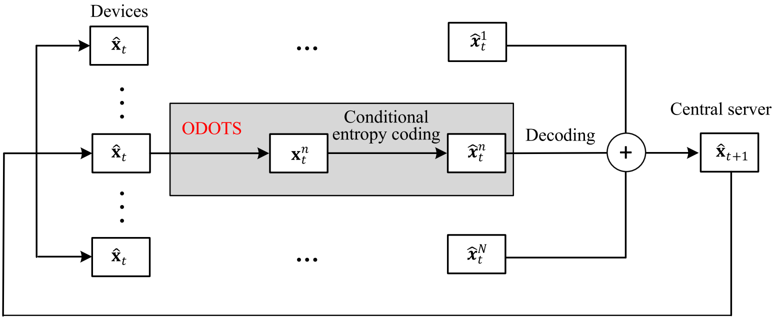
- Existing works separate the computation of $\{\mathbf{x}_t^n\}$ from their communication.



- But the best $\mathbf{x}_t^n$ for loss minimization (e.g., from gradient descent) usually is not the most efficient for transmission!

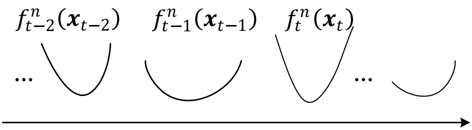
Our Approach

- We jointly consider loss minimization and communication efficiency when designing $\{\mathbf{x}_t^n\}$.



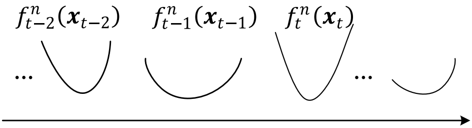
Online Distributed Optimization with Temporal Similarity (ODOTS)

Concern #2: Time-Varying Lost Functions



- $f_t^n(\mathbf{x}) = \frac{1}{|\mathcal{D}_t^n|} \sum_{i \in \mathcal{D}_t^n} l(\mathbf{x}; \mathbf{u}_t^{n,i}, v_t^{n,i})$: loss caused by time-varying local data.
- w_t^n: time-varying weight on device n.
- $f_t(\mathbf{x}_t) = \sum_{n=1}^{N} w_t^n f_t^n(\mathbf{x}_t)$: time-varying global loss function.

Concern #2: Time-Varying Lost Functions



- $f_t^n(\mathbf{x}) = \frac{1}{|\mathcal{D}_t^n|} \sum_{i \in \mathcal{D}_t^n} l(\mathbf{x}; \mathbf{u}_t^{n,i}, v_t^{n,i})$: loss caused by time-varying local data.
- w_t^n: time-varying weight on device n.
- $f_t(\mathbf{x}_t) = \sum_{n=1}^N w_t^n f_t^n(\mathbf{x}_t)$: time-varying global loss function.

- Need to make a sequence of decisions $\{\mathbf{x}_t\}$ without future information:

$$\min_{\{\mathbf{x}_t\}} \quad \frac{1}{T} \sum_{t=1}^{T} f_t(\mathbf{x}_t) \quad \approx \quad \min_{\{\mathbf{x}_t\}} \quad f_T(\mathbf{x}_T)$$

- Decisions are coupled over time by the communication-efficiency requirement.

Online Problem Formulation

$$\textbf{P1}: \quad \min_{\{\mathbf{x}_t^n \in \mathcal{X}\}} \quad \underbrace{\sum_{t=1}^{T} f_t(\hat{\mathbf{x}}_t)}_{\text{computation}}$$

$$\text{s.t.} \quad \underbrace{\frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} g_t^n(\mathbf{x}_t^n) \leq 0}_{\text{communication}}.$$

- $\hat{\mathbf{x}}_t \in \mathbb{R}^d$: global decision aggregated from quantized and coded versions of local decisions $\mathbf{x}_t^n \in \mathbb{R}^d$
- $g_t^n(\mathbf{x}) = \|\mathbf{x} - \hat{\mathbf{x}}_{t-1}^n\|^2 - \epsilon$: long-term decision dissimilarity function
  - It controls the communication overhead.
- $\mathcal{X} = \{\mathbf{x} : -x_{\max}\mathbf{1} \preceq \mathbf{x} \preceq x_{\max}\mathbf{1}\}$: short-term constraints

Online Distributed Optimization with Temporal Similarity (ODOTS)

ODOTS Algorithm Components

- Tunable virtual queue:

$$Q_{t+1}^n = \max\left\{0, (1-\gamma^2)Q_t^n + \gamma\eta g_t^n(\mathbf{x}_t^n)\right\}$$

  - Does not require the Slater's condition $g_t^n(\tilde{\mathbf{x}}) < 0$ for its upper bound.
  - But its stability does not exactly guarantee constraint satisfaction.

ODOTS Algorithm Components

- Tunable virtual queue:

$$Q_{t+1}^n = \max\left\{0, (1-\gamma^2)Q_t^n + \gamma\eta g_t^n(\mathbf{x}_t^n)\right\}$$

  - Does not require the Slater's condition $g_t^n(\tilde{\mathbf{x}}) < 0$ for its upper bound.
  - But its stability does not exactly guarantee constraint satisfaction.

- Modified family of Lyapunov drift functions for arbitrary $U \geq 0$:

$$\Theta_t^n = \frac{1}{2\gamma}(Q_{t+1}^n - U)^2 - \frac{1}{2\gamma}(Q_t^n - U)^2.$$

ODOTS Algorithm Components

- Tunable virtual queue:

$$Q_{t+1}^n = \max\left\{0, (1-\gamma^2)Q_t^n + \gamma\eta g_t^n(\mathbf{x}_t^n)\right\}$$

  - Does not require the Slater's condition $g_t^n(\tilde{\mathbf{x}}) < 0$ for its upper bound.
  - But its stability does not exactly guarantee constraint satisfaction.

- Modified family of Lyapunov drift functions for arbitrary $U \geq 0$:

$$\Theta_t^n = \frac{1}{2\gamma}(Q_{t+1}^n - U)^2 - \frac{1}{2\gamma}(Q_t^n - U)^2.$$

- Minimize an upper bound of the drift plus penalty plus violation

$$\underbrace{\Theta_t^n}_{\text{drift}} + \underbrace{\langle \nabla f_t^n(\hat{\mathbf{x}}_t), \mathbf{x} - \hat{\mathbf{x}}_t \rangle + \alpha \|\mathbf{x} - \hat{\mathbf{x}}_t\|^2}_{\text{penalty on loss}} + \underbrace{U\eta g_t^n(\mathbf{x})}_{\text{violation}}.$$

  - Unlike the drift-plus-penalty algorithm, the penalty here is not exactly the optimization objective.

ODOTS: Device's Algorithm

1: Update local decision $\mathbf{x}_t^n$ by solving per-slot problem

$$\mathbf{P2}^n: \quad \min_{\mathbf{x}\in\mathcal{X}} \overbrace{\underbrace{\langle \nabla f_t^n(\hat{\mathbf{x}}_t), \mathbf{x} - \hat{\mathbf{x}}_t \rangle + \alpha \|\mathbf{x} - \hat{\mathbf{x}}_t\|^2}_{\text{loss}} + \underbrace{\eta Q_t^n g_t^n(\mathbf{x})}_{\text{violation}}}^{\text{jointly optimize computation and communication}}.$$

- Minimizes upper bound of drift plus penalty plus violation
- Solution: $\mathbf{x}_t^n = \left[\underbrace{\frac{\alpha}{\alpha + \eta Q_t^n}}_{\text{scaling}} \left(\underbrace{\frac{\eta Q_t^n}{\alpha} \hat{\mathbf{x}}_{t-1}^n}_{\text{regularization}} + \underbrace{\hat{\mathbf{x}}_t - \frac{1}{2\alpha} \nabla f_t^n(\hat{\mathbf{x}}_t)}_{\text{local gradient descent}} \right) \right]_{-x_{\max}\mathbf{1}}^{x_{\max}\mathbf{1}}$
- Independent of U!

ODOTS: Device's Algorithm

1: Update local decision $\mathbf{x}_t^n$ by solving per-slot problem

$$\mathbf{P2}^n: \quad \min_{\mathbf{x}\in\mathcal{X}} \quad \overbrace{\underbrace{\langle \nabla f_t^n(\hat{\mathbf{x}}_t), \mathbf{x} - \hat{\mathbf{x}}_t \rangle + \alpha \|\mathbf{x} - \hat{\mathbf{x}}_t\|^2}_{\text{loss}} + \underbrace{\eta Q_t^n g_t^n(\mathbf{x})}_{\text{violation}}}^{\text{jointly optimize computation and communication}}.$$

- Minimizes upper bound of drift plus penalty plus violation
- Solution: $\mathbf{x}_t^n = \left[\underbrace{\frac{\alpha}{\alpha + \eta Q_t^n}}_{\text{scaling}} \left(\underbrace{\frac{\eta Q_t^n}{\alpha} \hat{\mathbf{x}}_{t-1}^n}_{\text{regularization}} + \underbrace{\hat{\mathbf{x}}_t - \frac{1}{2\alpha} \nabla f_t^n(\hat{\mathbf{x}}_t)}_{\text{local gradient descent}} \right) \right]_{-x_{\max}\mathbf{1}}^{x_{\max}\mathbf{1}}$
- Independent of U!

2: Update local tunable virtual queue Q_t^n.

3: Update quantized local decision $\hat{\mathbf{x}}_t^n$.

4: Transmit $\hat{\mathbf{x}}_t^n$ via conditional entropy coding.

ODOTS: Server's Algorithm

1: Receive noisy local decisions $\hat{\mathbf{x}}_t^n$.

2: Update noisy global decision $\hat{\mathbf{x}}_{t+1} = \sum_{n=1}^{N} w_t^n \hat{\mathbf{x}}_t^n$.

3: Broadcast $\hat{\mathbf{x}}_{t+1}$ to all devices.

Bounding Analysis

- Performance Bound
- Constraint Violation Bound

Assumptions on **P1** and Its Properties

Assumptions

The local loss function $f_t^n(\mathbf{x})$ is convex, *i.e.,*

$$f_t^n(\mathbf{y}) \geq f_t^n(\mathbf{x}) + \langle \nabla f_t^n(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \ \forall n, \ \forall t.$$

The local loss function $f_t^n(\mathbf{x})$ has bounded gradient $\nabla f_t^n(\mathbf{x})$: $\exists D > 0$, *s.t.,*

$$\| f_t^n(\mathbf{x}) \| \leq D, \ \forall \mathbf{x} \in \mathbb{R}^d, \ \forall n, \forall t.$$

Assumptions on **P1** and Its Properties

Assumptions

The local loss function $f_t^n(\mathbf{x})$ is convex, *i.e.*,

$$f_t^n(\mathbf{y}) \geq f_t^n(\mathbf{x}) + \langle \nabla f_t^n(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \ \forall n, \ \forall t.$$

The local loss function $f_t^n(\mathbf{x})$ has bounded gradient $\nabla f_t^n(\mathbf{x})$: $\exists D > 0$, *s.t.*,

$$\|f_t^n(\mathbf{x})\| \leq D, \ \forall \mathbf{x} \in \mathbb{R}^d, \ \forall n, \forall t.$$

Lemma 1

P1 satisfies the following:

$$\begin{aligned}
&\text{Bounded feasible set} : \|\mathbf{x} - \mathbf{y}\| \leq R, \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}, \\
&\text{Bounded communication error} : \|\hat{\mathbf{x}}_t - \mathbf{x}_t\| \leq \delta, \ \forall t, \\
&\text{Bounded constraint function} : |g_t^n(\mathbf{x})| \leq G, \ \forall \mathbf{x} \in \mathcal{X}, \forall n, \forall t.
\end{aligned}$$

where $R = 2\sqrt{d} x_{\max}$, $\delta = \frac{R}{4(2^b - 1)}$, and $G = \max\{\epsilon, R^2 + \delta^2 - \epsilon\}$.

Tunable Virtual Queue & Modified Lyapunov Drift

Lemma 2

The tunable virtual queue is upper bounded (without Slater's condition):

$$Q_t^n \leq \frac{\eta G}{\gamma}, \ \forall n, \forall t.$$

Tunable Virtual Queue & Modified Lyapunov Drift

Lemma 2

The tunable virtual queue is upper bounded (without Slater's condition):

$$Q_t^n \leq \frac{\eta G}{\gamma}, \ \forall n, \forall t.$$

Lemma 3

The modified Lyapunov drift is upper bounded:

$$\Theta_t^n \leq \underbrace{\eta Q_t^n g_t^n(\mathbf{x}_t^n)}_{\text{violation in } \mathbf{P2}^n} - \underbrace{U\eta g_t^n(\mathbf{x}_t^n)}_{\text{"plus violation"}} + \underbrace{2\gamma\eta^2 G^2 + \frac{\gamma}{2}U^2}_{\text{constants}}, \ \forall n, \forall t.$$

Bound on Per-Slot Local Loss and Constraint Violation

Lemma 4

The per-slot local loss and constraint violation is upper bounded:

$$f_t^n(\hat{\mathbf{x}}_t) + U\eta g_t^n(\mathbf{x}_t^n) \leq f_t^n(\mathbf{x}_t^{\text{ctr}}) + \frac{D^2}{4\alpha} + 2\gamma\eta^2 G^2 + \frac{\gamma}{2}U^2 - \Theta_t^n$$
$$+ \alpha\big(\phi_t + \psi_t^n + \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + 2R(\|\hat{\mathbf{x}}_t - \mathbf{x}_t\| + \pi_t)\big), \ \forall n, \forall t.$$

where

- $\mathbf{x}_t^{\text{ctr}} \in \arg\min\{f_t(\mathbf{x}) | g_t^n(\mathbf{x}) \leq 0, \forall n\}$ is the centralized per-slot optimizer;
- $\phi_t = \|\mathbf{x}_t^{\text{ctr}} - \mathbf{x}_t\|^2 - \|\mathbf{x}_{t+1}^{\text{ctr}} - \mathbf{x}_{t+1}\|^2$
  $\psi_t^n = \|\mathbf{x}_t^{\text{ctr}} - \mathbf{x}_{t+1}\|^2 - \|\mathbf{x}_t^{\text{ctr}} - \mathbf{x}_t^n\|^2$
  $\pi_t = \|\mathbf{x}_t^{\text{ctr}} - \mathbf{x}_{t+1}^{\text{ctr}}\|$
  represent how dynamic **P1** is.

Bound on Performance Gap to $\{\mathbf{x}_t^{\text{ctr}}\}$

Theorem 1

$$\sum_{t=1}^{T} \left(f_t(\hat{\mathbf{x}}_t) - f_t(\mathbf{x}_t^{\text{ctr}})\right) \leq \frac{D^2 T}{4\alpha} + 2\gamma\eta^2 G^2 T + \frac{\eta^2 G^2 \Omega_T}{2\gamma^3} + \alpha\left(R^2 + \Lambda_{2,T} + 2R(\Lambda_T + \Pi_T)\right)$$

where we use these accumulated variation measures:

$\Pi_T = \sum_{t=1}^{T} \pi_t$
$\Omega_T = \sum_{t=1}^{T} \sum_{n=1}^{N} (w_{t+1}^n - w_t^n)$
$\Lambda_T = \sum_{t=1}^{T} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|$
$\Lambda_{2,T} = \sum_{t=1}^{T} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2$

Bound on Performance Gap to $\{\mathbf{x}_t^{\text{ctr}}\}$

Theorem 1

$$\sum_{t=1}^{T} \left(f_t(\hat{\mathbf{x}}_t) - f_t(\mathbf{x}_t^{\text{ctr}}) \right) \leq \frac{D^2 T}{4\alpha} + 2\gamma\eta^2 G^2 T + \frac{\eta^2 G^2 \Omega_T}{2\gamma^3} + \alpha\left(R^2 + \Lambda_{2,T} + 2R(\Lambda_T + \Pi_T) \right)$$

where we use these accumulated variation measures:

$\Pi_T = \sum_{t=1}^{T} \pi_t$

$\Omega_T = \sum_{t=1}^{T} \sum_{n=1}^{N} (w_{t+1}^n - w_t^n)$

$\Lambda_T = \sum_{t=1}^{T} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|$

$\Lambda_{2,T} = \sum_{t=1}^{T} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2$

- Proved by setting $U = 0$ in modified Lyapunov drift.

Bound on Constraint Violation

Theorem 2

$$\frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} g_t^n(\mathbf{x}_t^n) \leq \Big(\frac{2\gamma^2 T + 2}{\gamma\eta^2}\Big)^{\frac{1}{2}} \Big(\frac{D^2 T}{4\alpha} + 2\gamma\eta^2 G^2 T + D(R+\delta)T + \alpha\big(R^2(1+\Xi_T) + \Lambda_{2,T} + 2R(\Lambda_T + \Pi_T)\big)\Big)^{\frac{1}{2}}$$

where $\Xi_T \triangleq \sum_{t=1}^{T}\sum_{n=1}^{N}\left(w_t^n - \frac{1}{N}\right)$ is the accumulated weight imbalance.

Bound on Constraint Violation

Theorem 2

$$\frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} g_t^n(\mathbf{x}_t^n) \leq \Big(\frac{2\gamma^2 T + 2}{\gamma\eta^2}\Big)^{\frac{1}{2}} \Big(\frac{D^2 T}{4\alpha} + 2\gamma\eta^2 G^2 T + D(R+\delta)T + \alpha\big(R^2(1+\Xi_T) + \Lambda_{2,T} + 2R(\Lambda_T + \Pi_T)\big)\Big)^{\frac{1}{2}}$$

where $\Xi_T \triangleq \sum_{t=1}^{T}\sum_{n=1}^{N}\left(w_t^n - \frac{1}{N}\right)$ is the accumulated weight imbalance.

- Proved by setting $U = \frac{\gamma\eta}{\gamma^2 T+1}\max\{0, \frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} g_t^n(\mathbf{x}_t^n)\}$ in modified Lyapunov drift.

Sublinear Performance Gap and Constraint Violation

Corollary

Time-invariant equal weights: $w_t^n = \frac{1}{N}, \forall n, \forall t$.
Let $\max\{\Pi_T, \Xi_T, \Lambda_{2,T}, \Lambda_T\} = O(T^{\mu})$.

$$\sum_{t=1}^{T} \left(f_t(\hat{\mathbf{x}}_t) - f_t(\mathbf{x}_t^{\text{ctr}}) \right) = \mathcal{O}\left(T^{\frac{1+\mu}{2}} \right),$$

$$\frac{1}{N} \sum_{t=1}^{T} \sum_{n=1}^{N} g_t^n(\mathbf{x}_t^n) = \mathcal{O}\left(T^{\frac{3}{4}} \right).$$

Sublinear Performance Gap and Constraint Violation

Corollary

Time-invariant equal weights: $w_t^n = \frac{1}{N}, \forall n, \forall t$.
Let $\max\{\Pi_T, \Xi_T, \Lambda_{2,T}, \Lambda_T\} = O(T^{\mu})$.

$$\sum_{t=1}^{T} \left(f_t(\hat{\mathbf{x}}_t) - f_t(\mathbf{x}_t^{\text{ctr}})\right) = \mathcal{O}\left(T^{\frac{1+\mu}{2}}\right),$$

$$\frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} g_t^n(\mathbf{x}_t^n) = \mathcal{O}\left(T^{\frac{3}{4}}\right).$$

Time-varying weights: $\Omega_T = \mathcal{O}(T^{\nu})$.

$$\sum_{t=1}^{T} \left(f_t(\hat{\mathbf{x}}_t) - f_t(\mathbf{x}_t^{\text{ctr}})\right) = \mathcal{O}\left(\max\{T^{\frac{1+\mu}{2}}, T^{\frac{3+\nu}{4}}\}\right),$$

$$\frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} g_t^n(\mathbf{x}_t^n) = \mathcal{O}\left(\max\{T^{\frac{3+\mu}{4}}, T^{\frac{7+\nu}{8}}\}\right).$$

Example Application in Communication-Efficient Federated Learning

Experimental Setup and Benchmarks

- Simulated online federated learning environment:
  - Image classification on MNIST dataset.
  - $N = 10$ devices, each holding data samples of one digit only.
  - In each slot, each device processes $|\mathcal{D}_t^n| = 20$ non-i.i.d. data samples.

Experimental Setup and Benchmarks

- Simulated online federated learning environment:
  - Image classification on MNIST dataset.
  - $N = 10$ devices, each holding data samples of one digit only.
  - In each slot, each device processes $|\mathcal{D}_t^n| = 20$ non-i.i.d. data samples.
- Performance benchmarks
  - Error-free FL: performance upper bound.
  - Primal-dual GD: current best for distributed constrained online optimization.
  - QFL-CE: quantized FL with the same conditional entropy coding as ODOTS.

Performance Metrics

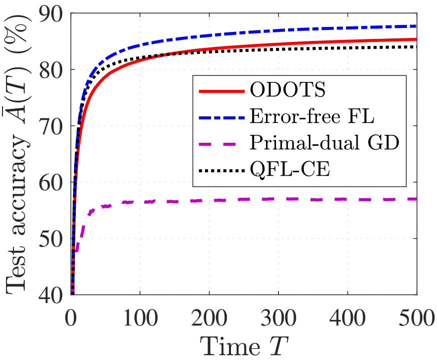
- Time-averaged test accuracy:

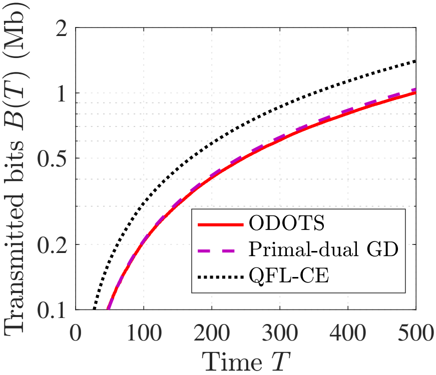
$$\bar{A}(T) = \frac{1}{|\mathcal{E}|T} \sum_{t=1}^{T} \sum_{i=1}^{|\mathcal{E}|} 1\left\{ \arg\max_j \left\{ \frac{\exp(\langle \hat{\mathbf{x}}_t[j], \mathbf{u}^i \rangle)}{\sum_{k=1}^{V} \exp(\langle \hat{\mathbf{x}}_t[k], \mathbf{u}^i \rangle)} \right\} = v^i \right\}.$$

- Total transmitted bits:

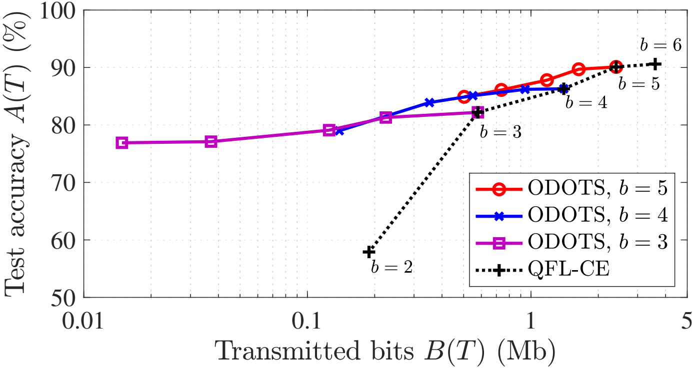
$$B(T) = \sum_{t=1}^{T} \sum_{n=1}^{N} H(\hat{\mathbf{x}}_t^n | \hat{\mathbf{x}}_{t-1}^n).$$

Convex Loss: Logistic Regression



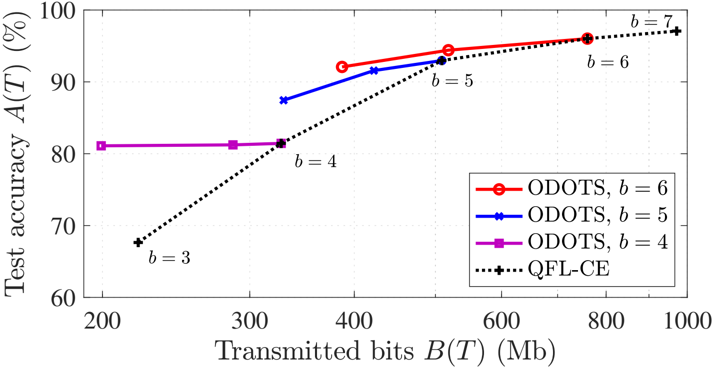


Test Accuracy vs. Transmitted Bits



- *b*: quantization bit length
- ODOTS with varying ϵ (average decision dis-similarity constraint)

Non-Convex Loss: Neural Network



- b: quantization bit length
- ODOTS with varying ϵ (average decision dis-similarity constraint)

Conclusions

- Communication-efficient online distributed optimization
  - Time-varying loss functions and weights
  - Temporal decision similarity through conditional entropy coding

Conclusions

- Communication-efficient online distributed optimization
  - Time-varying loss functions and weights
  - Temporal decision similarity through conditional entropy coding
- Online Distributed Optimization with Temporal Similarity (ODOTS)
  - Jointly considers loss minimization and communication efficiency;
  - Uses tunable virtual queue with modified Lyapunov drift analysis;

Conclusions

- Communication-efficient online distributed optimization
  - Time-varying loss functions and weights
  - Temporal decision similarity through conditional entropy coding
- Online Distributed Optimization with Temporal Similarity (ODOTS)
  - Jointly considers loss minimization and communication efficiency;
  - Uses tunable virtual queue with modified Lyapunov drift analysis;
  - Provides performance bounds on both computation and communication;
  - Outperforms current best alternatives especially under low quantization bit lengths.