DualFL: A Duality-based Federated Learning Algorithm with Communication Acceleration in the General Convex Regime

Jongho Park
Computer, Electrical and Mathematical Sciences and Engineering Division
King Abdullah University of Science and Technology (KAUST)
Thuwal 23955
jongho.park@kaust.edu.sa

Jinchao Xu
Computer, Electrical and Mathematical Sciences and Engineering Division
King Abdullah University of Science and Technology (KAUST)
Thuwal 23955
jinchao.xu@kaust.edu.sa

Abstract

We propose a novel training algorithm called DualFL (**Dual**ized **F**ederated **L**earning), for solving a distributed optimization problem in federated learning. Our approach is based on a specific dual formulation of the federated learning problem. DualFL achieves communication acceleration under various settings on smoothness and strong convexity of the problem. Moreover, it theoretically guarantees the use of inexact local solvers, preserving its optimal communication complexity even with inexact local solutions. DualFL is the first federated learning algorithm that achieves communication acceleration, even when the cost function is either nonsmooth or non-strongly convex. Numerical results demonstrate that the practical performance of DualFL is comparable to those of state-of-the-art federated learning algorithms, and it is robust with respect to hyperparameter tuning.

1 Introduction

This paper is devoted to a novel approach to design efficient training algorithms for *federated learning* [29]. Unlike standard machine learning approaches, federated learning encourages each client to have its own dataset and to update a local correction of a global model maintained by an orchestrating server via the local dataset and a local training algorithm. Recently, federated learning has been considered as an important research topic in the field of machine learning as data becomes increasingly decentralized and privacy of individual data is an utmost importance [16, 27].

In federated learning, it is assumed that communication costs dominate [29]. Hence, training algorithms for federated learning should be designed toward a direction that the amount of communication among the clients is reduced. For example, FedAvg [29], one of the most popular training algorithms for federated learning, improves its communication efficiency by adopting local training. Namely, multiple local gradient descent steps instead of a single step are performed in each client before communication among the clients. In recent years, various local training approaches have been considered to improve the communication efficiency of federated learning; e.g., operator splitting [37],

Preprint. Under review.

augmented Lagrangian [51], Douglas–Rachfold splitting [46], client-level momentum [48], and sharpness-aware minimization [38].

An important observation made in [17] is that data heterogeneity in federated learning can cause client drift, which in turn affects the convergence of federated learning algorithms. Indeed, it was observed in [18, Figure 3] that a large number of local gradient descent steps without shifting of the local gradient leads to solution nonconvergence. To address this issue, several gradient shift techniques that can compensate for client drift have been considered: `Scaffold` [17], `FedDyn` [1], `S-Local-SVRG` [13], `FedLin` [31], and `Scaffnew` [30]. These techniques achieve linear convergence rates of the training algorithms through carefully designed gradient shift techniques.

Recently, it was investigated in a pioneering work [30] that communication acceleration can be achieved by a federated learning algorithm if we use a tailored gradient shift scheme and a probabilistic approach for communication frequency. Specifically, it was shown that `Scaffnew` [30] achieves the optimal $\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon))$-communication complexity of distributed convex optimization [2] for the smooth strongly convex regime, where κ is the condition number of the problem and ϵ measures a target level of accuracy. Since then, several federated learning algorithms with communication acceleration have been considered; to name a few, `ProxSkip-VR` [28], `APDA-Inexact` [42], and `RandProx` [12]. One may refer to [28] for a historical survey on the theoretical progress of federated learning algorithms.

In this paper, we continue growing the list of federated learning algorithms with communication acceleration by proposing a novel algorithm called `DualFL` (**Dual**ized **F**ederated **L**earning). The key idea is to establish a certain duality [40] between a model federated learning problem and a composite optimization problem. By the nature of composite optimization problems [33], the dual problem can be solved efficiently by a forward-backward splitting algorithm with the optimal convergence rate [5, 11, 39]. By applying the predualization technique introduced in [22, 24] to an optimal forward-backward splitting method for the dual problem, we obtain our proposed algorithm `DualFL`. While each individual technique used in this paper is not new, a combination of the techniques yields the following desirable results:

- `DualFL` achieves the optimal $O(\sqrt{\kappa}\log(1/\epsilon))$-communication complexity in the smooth strongly convex regime.
- `DualFL` achieves communication acceleration even when the cost function is either nonsmooth or non-strongly convex.
- `DualFL` can adopt any optimization algorithm as its local solver, making it adaptable to each client's local problem.
- Communication acceleration of `DualFL` is guaranteed in a deterministic manner. That is, both the algorithm and its convergence analysis do not rely on stochastic arguments.

In particular, we would like to highlight that `DualFL` is the first federated learning algorithm that achieves communication acceleration for either nonsmooth or non-strongly convex problems.

The remainder of this paper is organized as follows. In Section 2, we state a model federated learning problem and several standard assumptions. We introduce the proposed `DualFL` and its convergence properties in Section 3. In Section 4, we introduce a regularization technique for `DualFL` to deal with non-strongly convex problems. We establish connections to existing federated learning algorithms in Section 5. In Section 6, we establish a duality relation between `DualFL` and a forward-backward splitting algorithm applied to a certain dual formulation. We present numerical results of `DualFL` in Section 7. Finally, we discuss limitations of this paper in Section 8.

2 Problem description

In this section, we present a standard mathematical model for federated learning and introduce several key assumptions that are used throughout the paper. In federated learning, it is assumed that each client possesses its own dataset, and that a local cost function is defined with respect to the dataset of each client. Hence, we consider the problem of minimizing the average of N cost functions stored on

N clients [16, 17, 28]:

$$\min_{\theta \in \Omega} \left\{ E(\theta) := \frac{1}{N} \sum_{j=1}^{N} f_j(\theta) \right\}, \tag{2.1}$$

where Ω is a parameter space and $f_j \colon \Omega \to \mathbb{R}$, $1 \leq j \leq N$, is a continuous and convex local cost function of the ith client. The local cost function f_j depends on the dataset of the jth client, but not on those of the other clients. We further assume that the cost function E is coercive, so that (2.1) admits a solution $\theta^* \in \Omega$ [4, Proposition 11.14]. Since problems of the form (2.1) arise in various applications in machine learning and statistics [43, 44], a number of algorithms have been developed to solve (2.1), e.g., stochastic gradient methods [6, 7, 21, 50]. In the following, We state several standard assumptions on each f_j in (2.1).

Assumption 2.1. Each f_j, $1 \leq j \leq N$, in (2.1) is μ-strongly convex for some $\mu > 0$. That is, we have

$$f_j(\theta) \geq f_j(\phi) + \langle \nabla f_j(\phi), \theta - \phi \rangle + \frac{\mu}{2} \|\theta - \phi\|^2, \quad \theta, \phi \in \Omega.$$

Assumption 2.2. Each f_j, $1 \leq j \leq N$, in (2.1) is L-smooth for some $L > 0$. That is, we have

$$f_j(\theta) \leq f_j(\phi) + \langle \nabla f_j(\phi), \theta - \phi \rangle + \frac{L}{2} \|\theta - \phi\|^2, \quad \theta, \phi \in \Omega.$$

We note that we do not need to make any similarity assumptions for f_j (cf. [38, Assumptions 2 and 3]). Under Assumption 2.1, the solution of the problem (2.1) is unique [11].

In what follows, an element of Ω^N is denoted by a bold symbol. For $\boldsymbol{\theta} \in \Omega^N$ and $1 \leq j \leq N$, we denote the ith component of $\boldsymbol{\theta}$ by θ_j, i.e., $\boldsymbol{\theta} = (\theta_j)_{j=1}^N$. We use the notation $A \lesssim B$ to represent that $A \leq CB$ for some constant $C > 0$ independent of the number of iterations n.

3 Main results

This section is devoted to the main findings of this paper: the proposed algorithm, called `DualFL`, and its convergence theorems. We now present `DualFL` in Algorithm 1 as follows.

Algorithm 1 `DualFL`: **Dual**ized **F**ederated **L**earning

Given $\rho \geq 0$ and $\nu > 0$,
set $\theta^{(0)} = \theta_j^{(0)} = 0 \in \Omega$ ($1 \leq j \leq N$), $\boldsymbol{\zeta}^{(0)} = \boldsymbol{\zeta}^{(-1)} = \mathbf{0} \in \Omega^N$, and $t_0 = 1$.
for $n = 0, 1, 2, \dots$ **do**
for each client ($1 \leq j \leq N$) **in parallel do**

$$\theta_j^{(n+1)} \approx \operatorname*{arg\,min}_{\theta_j \in \Omega} \left\{ E^{n,j}(\theta_j) := f_j(\theta_j) - \nu \langle \zeta_j^{(n)}, \theta_j \rangle \right\} \tag{3.1}$$

end for

$$\theta^{(n+1)} = \frac{1}{N} \sum_{j=1}^{N} \theta_j^{(n+1)} \tag{3.2}$$

for each client ($1 \leq j \leq N$) **in parallel do**

$$\zeta_j^{(n+1)} = (1 + \beta_n) \left(\zeta_j^{(n)} + \theta^{(n+1)} - \theta_j^{(n+1)} \right) - \beta_n \left(\zeta_j^{(n-1)} + \theta^{(n)} - \theta_j^{(n)} \right), \tag{3.3}$$

where β_n is given by

$$t_{n+1} = \frac{1 - \rho t_n^2 + \sqrt{(1 - \rho t_n^2)^2 + 4 t_n^2}}{2}, \quad \beta_n = \frac{t_n - 1}{t_{n+1}} \frac{1 - t_{n+1}\rho}{1 - \rho}. \tag{3.4}$$

end for
end for

DualFL updates the server parameter from $\theta^{(n)}$ to $\theta^{(n+1)}$ by the following steps. First, each client computes its local solution $\theta_j^{(n+1)}$ by solving the local problem (3.1). Note that the local problem (3.1) is defined in terms of the local control variate $\zeta_j^{(n)}$. Then the server aggregates all the local solutions $\theta_j^{(n+1)}$ by averaging them to obtain a new server parameter $\theta^{(n+1)}$. After obtaining the new server parameter $\theta^{(n+1)}$, it is transferred to each client, and the local control variate is updated using (3.3). The overrelaxation parameter β_n in (3.3) can be obtained by a simple recursive formula (3.4), which relies on the hyperparameter ρ.

One feature of the proposed DualFL is its flexibility in choosing local solvers for the local problem (3.1). More precisely, the method allows for the adoption of any local solvers, making it adaptable to each local problem in a client. The same advantage was reported in several existing works such as [1, 46, 51]. Another notable feature of DualFL is its fully deterministic nature, in contrast to some existing federated learning algorithms that rely on randomness to achieve communication acceleration [12, 28, 30]. Specifically, DualFL does not rely on uncertainty to ensure communication acceleration, which enhances its reliability. Very recently, several federated learning algorithms that share the same advantage have been proposed; see, e.g., [42].

3.1 Inexact local solvers

In DualFL, local problems of the form (3.1) are typically solved inexactly using iterative algorithms. The resulting local solutions may deviate from the exact minimizers, and this discrepancy can affect the convergence behavior. Here, we present a certain inexactness assumption for local solvers that does not deteriorate the convergence properties of DualFL.

For a function $f\colon X \to \overline{\mathbb{R}}$ defined on a Euclidean space X, let $f^*\colon X \to \overline{\mathbb{R}}$ denote the Legendre–Fenchel conjugate of f, i.e.,

$$f^*(p) = \sup_{x \in X} \left\{ \langle p, x \rangle - f(x) \right\}, \quad p \in X.$$

The following proposition is readily deduced by the Fenchel–Rockafellar duality (see Appendix A).

Proposition 3.1. *Suppose that Assumption 2.1 holds. For a positive constant $\nu \in (0, \mu]$, if $\theta_j \in \Omega$ solves* (3.1)*, then $\xi_j = \nu(\zeta_j^{(n)} - \theta_j) \in \Omega$ solves*

$$\min_{\xi_j \in \Omega} \left\{ E_{\mathrm{d}}^{n,j}(\xi_j) := g_j^*(\xi_j) + \frac{1}{2\nu} \|\xi_j - \nu \zeta_j^{(n)}\|^2 \right\}, \tag{3.5}$$

where $g_j(\theta) = f_j(\theta) - \frac{\nu}{2}\|\theta\|^2$. Moreover, we have

$$E^{n,j}(\theta_j) + E_{\mathrm{d}}^{n,j}(\xi_j) = 0.$$

Thanks to Proposition 3.1, $\theta_j \in \Omega$ is a solution of Equation (3.1) if and only if the primal-dual gap $\Gamma^{n,j}(\theta_j)$ defined by

$$\Gamma^{n,j}(\theta_j) = E^{n,j}(\theta_j) + E_{\mathrm{d}}^{n,j}(\nu(\zeta_j^{(n)} - \theta_j)) \tag{3.6}$$

vanishes [8]. The primal-dual gap $\Gamma^{n,j}(\theta_j)$ can play a role of an implementable inexactness criterion since it is observable by simple arithmetic operations (see Section 2.1 of [3]). If the local problem (3.1) is solved by a convergent iterative algorithm such as gradient descent methods, then the primal-dual gap $\Gamma^{n,j}(\theta_j^{(n+1)})$ can be arbitrarily small with a sufficiently large number of inner iterations.

3.2 Convergence theorems

The following theorem states that DualFL is provably convergent in the nonsmooth strongly convex regime if each local problem is solved so accurately that the primal-dual gap becomes less than a certain value. Moreover, DualFL achieves communication acceleration in the sense that the squared solution error $\|\theta^{(n)} - \theta^*\|^2$ at the nth communication round is bounded by $\mathcal{O}(1/n^2)$, which is derived by momentum acceleration; see Section 6. As we are aware, DualFL is the first federated learning algorithm with communication acceleration that is convergent even if the cost function is nonsmooth. A proof sketch of Theorem 3.2 will be provided in Section 6; see Appendix B for the full proof.

Theorem 3.2. *Suppose that Assumption 2.1 holds. In addition, suppose that the number of local iterations for the jth client at the nth epoch of* `DualFL` *is large enough to satisfy*

$$\Gamma^{n,j}(\theta_j^{(n+1)}) \leq \frac{1}{N\nu(n+1)^{4+\gamma}} \tag{3.7}$$

for some $\gamma > 0$ ($1 \leq j \leq N$, $n \geq 0$). If we choose the hyperparameters ρ and ν in `DualFL` *such that $\rho = 0$ and $\nu \in (0, \mu]$, then the sequence $\{\theta^{(n)}\}$ generated by* `DualFL` *converges to the solution θ^* of* (2.1)*. Moreover, for $n \geq 0$, we have*

$$\|\theta^{(n)} - \theta^*\|^2 \lesssim \frac{1}{n^2}.$$

If we additionally assume that Assumption 2.2 holds, then we are able to obtain an improved convergence rate of `DualFL`. Under Assumptions 2.1 and 2.2, we define the condition number κ of the problem (2.1) as $\kappa = L/\mu$. If we choose the hyperparameters ρ and ν appropriately, then `DualFL` becomes linearly convergent with the rate $1 - 1/\sqrt{\kappa}$. Consequently, `DualFL` achieves the optimal $\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon))$-communication efficiency [2]. This observation is summarized in Theorem 3.3; see Section 6 for a proof sketch.

Theorem 3.3. *Suppose that Assumptions 2.1 and 2.2 hold. In addition, suppose that the number of local iterations for the jth client at the nth epoch of* `DualFL` *is large enough to satisfy*

$$\Gamma^{n,j}(\theta_j^{(n+1)}) \leq \frac{1}{N}\left(\frac{1-\sqrt{\rho}}{1+\gamma}\right)^n \tag{3.8}$$

for some $\gamma > 0$ ($1 \leq j \leq N$, $n \geq 0$). If we choose the hyperparameters ρ and ν in `DualFL` *such that $\rho \leq [0, \nu/L]$ and $\nu \leq (0, \mu]$, then the sequence $\{\theta^{(n)}\}$ generated by* `DualFL` *converges to the solution θ^* of* (2.1)*. Moreover, for $n \geq 0$, we have*

$$E(\theta^{(n)}) - E(\theta^*) \lesssim \|\theta^{(n)} - \theta^*\|^2 \lesssim (1-\sqrt{\rho})^n .$$

In particular, if we set $\rho = \kappa^{-1}$ and $\nu = \mu$ in `DualFL`*, then we have*

$$E(\theta^{(n)}) - E(\theta^*) \lesssim \|\theta^{(n)} - \theta^*\|^2 \lesssim \left(1 - \frac{1}{\sqrt{\kappa}}\right)^n,$$

where $\kappa = L/\mu$. Namely, `DualFL` *achieves the optimal $\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon))$-communication complexity of distributed convex optimization in the smooth strongly convex regime.*

Theorem 3.3 implies that `DualFL` is linearly convergent with an acceptable rate $1 - \sqrt{\rho}$ even if the hyperparameters were not chosen optimally. That is, the performance `DualFL` is robust with respect to a choice of the hyperparameters.

3.3 Local iteration complexity

We analyze the local iteration complexity of `DualFL` under the conditions of Theorems 3.2 and 3.3. We recall that `DualFL` is compatible with any optimization algorithm as its local solver. Hence, we may assume that we use an optimal first-order optimization algorithm in the sense of Nemirovskii and Nesterov [32, 36]. That is, optimization algorithms of iteration complexity $\mathcal{O}(1/\epsilon)$ and $\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon))$ are considered in the cases corresponding to Theorems 3.2 and 3.3. Based on this setting, we have the following results regarding the local iteration complexity of `DualFL`. Both theorems can be derived straightforwardly by substituting ϵ in the iteration complexity of local solvers with the threshold values given in Theorems 3.2 and 3.3. Note that the number of outer iterations of `DualFL` to meet the target accuracy $\epsilon_{\text{out}} > 0$ is $\mathcal{O}(1/\sqrt{\epsilon_{\text{out}}})$ and $\mathcal{O}((1/\sqrt{\rho})\log(1/\epsilon_{\text{out}}))$ in the cases of Theorems 3.2 and 3.3, respectively.

Theorem 3.4. *Suppose that the assumptions given in Theorem 3.2 hold. If the local problem* (3.1) *is solved by an optimal first-order algorithm of iteration complexity $\mathcal{O}(1/\epsilon)$, then the number of inner iterations M_n at the nth epoch of* `DualFL` *satisfies*

$$M_n = \mathcal{O}\left(N(n+1)^{4+\gamma}\right) = \mathcal{O}\left(\frac{N}{\epsilon_{\text{out}}^{2+\frac{\gamma}{2}}}\right),$$

where $\epsilon_{\text{out}} > 0$ is the target accuracy of the outer iterations of `DualFL`.

Theorem 3.5. *Suppose that the assumptions given in Theorem 3.3 hold. If the local problem* (3.1) *is solved by an optimal first-order algorithm of iteration complexity* $\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon))$*, then the number of inner iterations* M_n *at the* n*th epoch of* `DualFL` *satisfies*

$$M_n = \mathcal{O}\left(\sqrt{\kappa}\left(\log(N(1+\gamma)^n) + n\sqrt{\rho}\right)\right) = \mathcal{O}\left(\sqrt{\kappa}\log\frac{N(1+\gamma)^n}{\epsilon_{\mathrm{out}}}\right),$$

where $\epsilon_{\mathrm{out}} > 0$ *is the target accuracy of the outer iterations of* `DualFL`*. In particular, if we set* $\gamma \to 0^+$*, then we have*

$$M_n = \mathcal{O}\left(\sqrt{\kappa}\log\frac{N}{\epsilon_{\mathrm{out}}}\right).$$

Similar to other state-of-the-art federated learning algorithms [12, 14, 30], the local iteration complexity of `DualFL` scales with $\sqrt{\kappa}$. This implies that `DualFL` is computationally efficient, not only in terms of communication complexity but also in terms of total complexity.

4 Extension to non-strongly convex problems

The convergence properties of the proposed `DualFL` presented in Section 3 rely on Assumption 2.1, which implies that the cost function E of (2.1) is μ-strongly convex for some $\mu > 0$. Although this assumption has been considered as a standard one in many existing works on federated learning algorithms [1, 17, 30, 42], it may not hold in practical settings and is often unrealistic. In this section, we deal with how to apply `DualFL` to non-strongly convex problems, i.e., when Assumption 2.1 does not hold. Throughout this section, we assume that each f_j, $1 \leq j \leq N$, in the model problem (2.1) is not strongly convex. In this case, (2.1) admits nonunique solutions in general. For a positive real number $\alpha > 0$, we consider the following ℓ^2-regularization [34] of (2.1):

$$\min_{\theta\in\Omega}\left\{E^\alpha(\theta) := \frac{1}{N}\sum_{j=1}^N f_j^\alpha(\theta)\right\}, \quad f_j^\alpha(\theta) = f_j(\theta) + \frac{\alpha}{2}\|\theta\|^2. \tag{4.1}$$

Then (4.1) satisfies Assumption 2.1 with $\mu = \alpha$. Hence, `DualFL` applied to (4.1) satisfy the convergence properties stated in Theorems 3.2 and 3.3. In particular, the sequence $\{\theta^{(n)}\}$ generated by `DualFL` applied to (4.1) converges to the unique solution $\theta^\alpha \in \Omega$ of (4.1). Invoking the epigraphical convergence theory from [41], we establish Theorem 4.1, which means that for sufficiently small α and large n, $\theta^{(n)}$ is a good approximation of a solution θ^* of (2.1). A detailed proof of Theorem 4.1 can be found in Appendix C.

Theorem 4.1. *In* `DualFL` *applied to the regularized problem* (4.1)*, suppose that the local problems are solved with sufficient accuracy so that* (3.7) *holds. If we choose* $\rho = 0$ *and* $\nu = \alpha$ *in* `DualFL`*, then the sequence* $\{\theta^{(n)}\}$ *generated by* `DualFL` *applied to* (4.1) *satisfies*

$$E(\theta^{(n)}) - E(\theta^*) \to 0 \quad \text{as } n \to \infty \text{ and } \alpha \to 0^+.$$

In the proof of Theorem 4.1, we use the fact that $E(\theta^\alpha) \to E(\theta^*)$ as $\alpha \to 0^+$ [41, Theorem 7.33]. Hence, by the coercivity of E, for any $\alpha_0 > 0$, we have $R_0 > 0$ such that

$$\{\theta^\alpha : \alpha \in (0, \alpha_0]\} \subset \{\theta : \|\theta\| \leq R_0\}. \tag{4.2}$$

Under Assumption 2.2, i.e., if E is smooth, we can show that `DualFL` achieves communication acceleration in the sense that the number of communication rounds to make the gradient error $\|\nabla E(\theta^{(n)})\|$ smaller than ϵ is $\mathcal{O}((1/\sqrt{\epsilon})\log(1/\epsilon))$, which agrees with the optimal estimate for first-order methods up to a logarithmic factor [20]. A proof of Theorem 4.2 can be found in Appendix C.

Theorem 4.2. *Suppose that Assumption 2.2 holds. In addition, in* `DualFL` *applied to the regularized problem* (4.1)*, suppose that the local problems are solved with sufficiently accuracy so that* (3.8) *holds. If we choose* $\rho = \alpha/(L+\alpha)$ *and* $\nu = \alpha$ *in* `DualFL`*, then, for* $n \geq 0$*, we have*

$$\|\nabla E(\theta^{(n)})\| \lesssim \left(1 - \sqrt{\frac{\alpha}{L+\alpha}}\right)^{\frac{n}{2}} + \alpha\|\theta^\alpha\|. \tag{4.3}$$

Moreover, if we choose $\alpha = \epsilon/(2R_0)$ *for some* $\epsilon \in (0, 2R_0\alpha_0]$*, where* α_0 *and* R_0 *were given in* (4.2)*, then the number of communication rounds* M_{comm} *to achieve* $\|\nabla E(\theta^{(n)})\| \leq \epsilon$ *satisfies*

$$M_{\mathrm{comm}} \leq \left(1 + 2\sqrt{1 + \frac{2LR_0}{\epsilon}}\right)\left(\log\frac{1}{\epsilon} + \text{constant}\right) = \mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\log\frac{1}{\epsilon}\right). \tag{4.4}$$

Table 1: Comparison between `DualFL` and other fifth-generation federated learning algorithms that achieve acceleration of communication complexity. The $\tilde{\mathcal{O}}$-notation neglects logarithmic factors.

Algorithm	Comm. acceleration			Local iter. complexity		Deterministic / Stochastic
	smooth	nonsmooth	smooth	smooth	nonsmooth	
	strongly convex		non-strongly convex	strongly convex		
`Scaffnew` [30]	Yes	N/A	N/A	$\tilde{\mathcal{O}}(\sqrt{\kappa})$	N/A	Stochastic
`APDA-Inexact` [42]	Yes	N/A	N/A	better	N/A	Deterministic
`5GCS` [14]	Yes	N/A	N/A	$\tilde{\mathcal{O}}(\sqrt{\kappa})$	N/A	Deterministic
`RandProx` [12]	Yes	N/A	No	$\tilde{\mathcal{O}}(\sqrt{\kappa})$	N/A	Stochastic
`DualFL`	Yes	**Yes**	**Yes**	$\tilde{\mathcal{O}}(\sqrt{\kappa})$	$\tilde{\mathcal{O}}(1/\epsilon^2)$	Deterministic

5 Comparison with existing algorithms and convergence theory

In this section, we discuss connections to existing federated learning algorithms. Based on the classification established in [28], `DualFL` can be classified as a fifth-generation federated learning algorithm, which achieves communication acceleration. In the smooth strongly convex regime, `DualFL` achieves the $\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon))$-communication complexity, which is comparable to other existing algorithms in the same generation such as `Scaffnew` [30], `APDA-Inexact` [42], and `RandProx` [12]. The optimal communication complexity of `DualFL` is achieved without relying on randomness; all the statements in the algorithm are deterministic. This feature is shared with some recent federated learning algorithms such as `APDA-Inexact` [42] and `5GCS` [14]. A distinct novelty of `DualFL` is its communication acceleration, even when the cost function is either nonsmooth or non-strongly convex. Among the existing fifth generation of federated learning algorithms, only `RandProx` has been proven to be convergent in the smooth non-strongly convex regime, with an $O(1/n)$-convergence rate of the energy error [12, Theorem 11]. However, this rate is the same of those of federated learning algorithms without communication acceleration such as `Scaffold` [17] and `FedDyn` [1]. In contrast, `DualFL` achieves the $O((1/\sqrt{\epsilon})\log(1/\epsilon))$-communication complexity with respect to the gradient error, which has not been achieved by the existing algorithms. Furthermore, not only communication acceleration but also convergence to a solution in the nonsmooth strongly convex regime have not been addressed by the existing fifth generation algorithms.

In the local problem (3.1) of `DualFL`, we minimize not only the local cost function $f_j(\theta_j)$ but also an additional term $-\nu\langle\zeta_j^{(n)}, \theta_j\rangle$. That is, $-\nu\zeta_j^{(n)}$ serves as a gradient shift to mitigate client drift and accelerate convergence. In this viewpoint, `DualFL` can be classified as a federated learning algorithm with gradient shift. This class includes other methods such as `Scaffold` [17], `FedDyn` [1], `S-Local-SVRG` [13], `FedLin` [31], and `Scaffnew` [30]. Meanwhile, `DualFL` belongs to the class of primal-dual methods for federated learning, e.g., `FedPD` [51], `FedDR` [46], `APDA-Inexact` [42], and `5GCS` [14]. While almost of the existing methods utilize a consensus reformulation of (2.1) (see [30, Equation (6)]), `DualFL` is based on a certain dual formulation of (2.1), as we will see in Section 6. More precisely, we will show that `DualFL` is obtained by applying predualization [22, 24] to an accelerated forward-backward splitting algorithm [5, 11, 39] for the dual problem. The dual problem has a particular structure that makes the forward-backward splitting algorithm equivalent to the prerelaxed nonlinear block Jacobi method [25], which belongs to a broad class of parallel subspace correction methods [47] for convex optimization [35, 45].

Table 1 provides an overview of the comparison between `DualFL` and other fifth-generation federated learning algorithms discussed above.

6 Mathematical theory

This section provides a mathematical theory for `DualFL`. We establish a duality relation between `DualFL` and an accelerated forward-backward splitting algorithm [5, 11, 39] applied to a certain dual formulation of the model problem (2.1). Utilizing this duality relation, we derive the convergence theorems presented in this paper, namely Theorems 3.2 and 3.3. Moreover, the duality relation provides a rationale for naming the proposed algorithm as `DualFL`. Throughout this section, we may assume that Assumption 2.1 holds, as `DualFL` for a non-strongly convex problem utilizes the strongly convex regularization (4.1).

We first introduce a dual formulation of the model federated learning problem (2.1) that is required for the convergence analysis. For a positive constant $\nu \in (0, \mu]$, the dual formulation of the problem (2.1) is given by

$$\min_{\boldsymbol{\xi} \in \Omega^N} \left\{ E_{\mathrm{d}}(\boldsymbol{\xi}) := \sum_{j=1}^{N} g_j^*(\xi_j) + \frac{1}{2N\nu} \left\| \sum_{j=1}^{N} \xi_j \right\|^2 \right\}, \tag{6.1}$$

where $g_j(\theta) = f_j(\theta) - \frac{\nu}{2}\|\theta\|^2$. A detailed derivation of (6.1) can be found in Appendix A. We note that problems of the form (6.1) have been applied in some limited cases in machine learning, such as support vector machines [15] and logistic regression [49]. Very recently, the dual problem (6.1) was utilized in federated learning in [14]. Let $\boldsymbol{\xi}^* \in \Omega^N$ denote a solution of (6.1). We have the following primal-dual relation between the primal solution θ^* and the dual solution $\boldsymbol{\xi}^*$:

$$\theta^* = -\frac{1}{N\nu} \sum_{j=1}^{N} \xi_j^*, \quad \xi_j^* = \nabla g_j(\theta^*). \tag{6.2}$$

For $\boldsymbol{\xi} \in \Omega^N$, let

$$F_{\mathrm{d}}(\boldsymbol{\xi}) = \frac{1}{2N\nu} \left\| \sum_{j=1}^{N} \xi_j \right\|^2, \quad G_{\mathrm{d}}(\boldsymbol{\xi}) = \sum_{j=1}^{N} g_j^*(\xi_j).$$

Then (6.1) is rewritten as the following composite optimization problem [33]:

$$\min_{\boldsymbol{\xi} \in \Omega^N} \left\{ E_{\mathrm{d}}(\boldsymbol{\xi}) := F_{\mathrm{d}}(\boldsymbol{\xi}) + G_{\mathrm{d}}(\boldsymbol{\xi}) \right\}. \tag{6.3}$$

By the Cauchy–Schwarz inequality, F_{d} is ν^{-1}-smooth. Moreover, under Assumptions 2.1 and 2.2, G_{d} is $(L-\nu)^{-1}$-strongly convex if $\nu \in (0, \mu]$. Since (6.3) is a composite optimization problem, forward-backward splitting algorithms are well-suited to solve it. Among several variants of forward-backward splitting algorithms, we focus on an inexact version of FISTA [5] proposed in [39], which accommodates strongly convex objectives and inexact proximal operations. Inexact FISTA with the fixed step size ν applied to (6.3) is summarized in Algorithm 2, in the form suitable for our purposes.

Algorithm 2 Inexact FISTA for the dual problem (6.3)

Given $\rho \geq 0$, $\nu > 0$, and $\{\delta_n\}_{n=0}^{\infty}$,
set $\boldsymbol{\xi}^{(0)} = \boldsymbol{\eta}^{(0)} = \mathbf{0} \in \Omega^N$, and $t_0 = 1$.
for $n = 0, 1, 2, \dots$ **do**

$$\boldsymbol{\xi}^{(n+1)} \approx \operatorname*{arg\,min}_{\boldsymbol{\xi} \in \Omega^N} \left\{ E_{\mathrm{d}}^n(\boldsymbol{\xi}) := \langle \nabla F_{\mathrm{d}}(\boldsymbol{\eta}^{(n)}), \boldsymbol{\xi} - \boldsymbol{\eta}^{(n)} \rangle + \frac{1}{2\nu} \|\boldsymbol{\xi} - \boldsymbol{\eta}^{(n)}\|^2 + G_{\mathrm{d}}(\boldsymbol{\xi}) \right\} \tag{6.4}$$

such that $E_{\mathrm{d}}^n(\boldsymbol{\xi}^{(n+1)}) - \min E_{\mathrm{d}}^n \leq \delta_n$.

$$\boldsymbol{\eta}^{(n+1)} = (1+\beta_n)\boldsymbol{\xi}^{(n+1)} - \beta_n \boldsymbol{\xi}^{(n)}, \tag{6.5}$$

where β_n is given by (3.4).
end for

An important observation is that there exists a duality relation between the sequences generated by Algorithm 2 and those generated by `DualFL`. In `DualFL`, we define two auxiliary sequences $\{\boldsymbol{\xi}^{(n)}\}$ and $\{\boldsymbol{\eta}^{(n)}\}$ as follows:

$$\xi_j^{(n+1)} = \nu(\zeta_j^{(n)} - \theta_j^{(n+1)}), \quad \xi_j^{(0)} = 0, \tag{6.6a}$$

$$\eta_j^{(n+1)} = \nu(\zeta_j^{(n+1)} - (1+\beta_n)\theta^{(n+1)} + \beta_n \theta^{(n)}), \quad \eta_j^{(0)} = 0, \tag{6.6b}$$

for $n \geq 0$ and $1 \leq j \leq N$. The following lemma summarizes the duality relation between `DualFL` and Algorithm 2; the sequences $\{\boldsymbol{\xi}^{(n)}\}$ and $\{\boldsymbol{\eta}^{(n)}\}$ defined in (6.6) agree with those generated by Algorithm 2. A proof of Lemma 6.1 is provided in Appendix B.

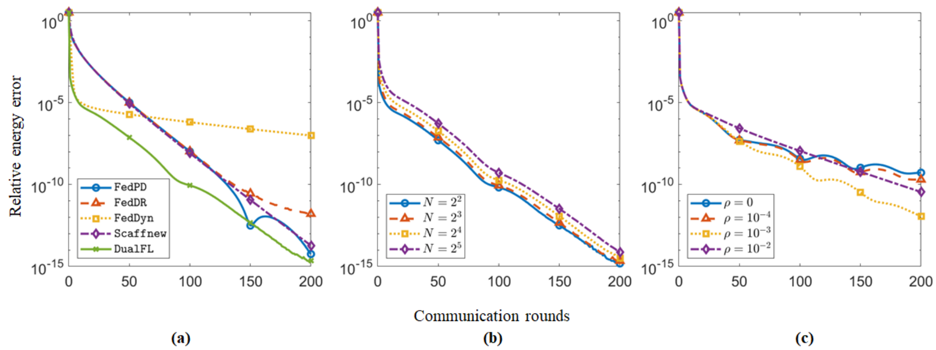

Figure 1: Relative energy error $\frac{E(\theta)-E(\theta^*)}{E(\theta^*)}$ with respect to the number of communication rounds in various training algorithms for multinomial logistic regression on the MNIST training dataset. **(a)** Comparison of `DualFL` with benchmark algorithms. **(c)** Convergence of `DualFL` when the number of clients N changes. **(c)** Convergence of `DualFL` when the value of the hyperparameter ρ changes.

Lemma 6.1. *Suppose that Assumption 2.1 hold. In addition, suppose that the number of local iterations for the jth client at the nth epoch of* `DualFL` *is large enough to satisfy*

$$\Gamma^{n,j}(\theta_j^{(n+1)}) \leq \frac{\delta_n}{N}$$

for some $\delta_n > 0$ ($1 \leq j \leq N$, $n \geq 0$). Then the sequences $\{\boldsymbol{\xi}^{(n)}\}$ and $\{\boldsymbol{\eta}^{(n)}\}$ defined in (6.6) *agree with those generated by Algorithm 2 for the dual problem* (6.3)*.*

Lemma 6.1 implies that `DualFL` is a predualization [22, 24] of Algorithm 2. Namely, `DualFL` can be constructed by transforming the dual sequence $\{\boldsymbol{\xi}^{(n)}\}$ generated by Algorithm 2 into the primal sequence $\{\theta^{(n)}\}$ by leveraging the primal-dual relation (6.2). Finally, the main convergence theorems for `DualFL`, Theorems 3.2 and 3.3, can be derived by combining the optimal convergence properties of Algorithm 2 proven in [39, Corollaries 3.3 and 3.4] and the duality relation presented in Lemma 6.1. A detailed derivation is provided in Appendix B.

7 Numerical experiments

In this section, we present numerical results that demonstrate the performance of `DualFL`. As benchmarks, we choose the following recent federated learning algorithms: `FedPD` [51], `FedDR` [46], `FedDyn` [1], and `Scaffnew` [30]. To test the performance of the algorithms, we use multinomial logistic regression on the MNIST training dataset [23]. The full details, including the computing resources and the choice of hyperparameters, are provided in Appendix D.

Numerical results are presented in Figure 1. Figure 1(a) displays the convergence behavior of the benchmark algorithms, along with `DualFL`, when $N = 2^3$. While the linear convergence rate of `DualFL` appears to be similar to those of `FedPD`, `FedDR`, and `Scaffnew`, the energy curve of `DualFL` is consistently lower than those of the other algorithms because `DualFL` achieves faster energy decay in the first several iterations, similar to `FedDyn`. That is, the `DualFL` loss decays as fast as `FedDyn` in the first several iterations, and then the linear decay rate of `DualFL` becomes similar to those of `FedPD`, `FedDR`, and `Scaffnew`. Figure 1(b) verifies that the convergence rate of `DualFL` does not deteriorate even if the number of clients N becomes large. That is, `DualFL` is robust to a large number of clients. Finally, Figure 1(c) illustrates the convergence behavior of `DualFL` under the condition where ν is fixed by μ, and the value of ρ are varied. It can be seen that even when ρ is chosen far from its tuned value, the convergence rate of `DualFL` does not deteriorate significantly. This verifies the robustness of `DualFL` with respect to hyperparameter tuning.

8 Limitations and future works

A major limitation of this paper is that all the results are based on the convex setting. Although this limitation is also present in many recent works on federated learning algorithms [12, 30, 42], the nonconvex setting should be considered in future research to cover a wider range of practical machine learning tasks.

While our primary focus in this paper is on the communication efficiency of training algorithms, we acknowledge that there are other crucial aspects of federated learning, such as client sampling and communication compression. We expect that our results can be extended to incorporate client sampling by carefully following existing works, such as [14], on federated learning algorithms with client sampling. On the other hand, since communication compression can be modeled by stochastic gradients [13], we consider extending our results for stochastic gradients as a future work.

Acknowledgments and Disclosure of Funding

This work was supported by the KAUST Baseline Research Fund.

References

[1] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021.

[2] Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[3] Mathieu Barré, Adrien B. Taylor, and Francis Bach. Principled analyses and design of first-order methods with inexact proximal operators. *Mathematical Programming*, pages 1–46, 2022.

[4] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, 2011.

[5] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[6] Dimitri P. Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129(2):163–195, 2011.

[7] Dimitri P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *arXiv preprint arXiv:1507.01030*, 2015.

[8] Silvia Bonettini, Simone Rebegoldi, and Valeria Ruggiero. Inertial variable metric techniques for the inexact forward–backward algorithm. *SIAM Journal on Scientific Computing*, 40(5):A3180–A3210, 2018.

[9] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[10] Luca Calatroni and Antonin Chambolle. Backtracking strategies for accelerated descent methods with smooth composite objectives. *SIAM Journal on Optimization*, 29(3):1772–1798, 2019.

[11] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

[12] Laurent Condat and Peter Richtárik. RandProx: Primal-dual optimization algorithms with randomized proximal updates. In *The Eleventh International Conference on Learning Representations*, 2023.

[13] Eduard Gorbunov, Filip Hanzely, and Peter Richtarik. Local SGD: Unified theory and new efficient methods. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3556–3564. PMLR, 2021.

[14] Michał Grudzień, Grigory Malinovsky, and Peter Richtárik. Can 5th generation local training methods support client sampling? Yes! *arXiv preprint arXiv:2212.14370*, 2022.

[15] C. Hsieh, K. Chang, Lin C., S. Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear SVM. In *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pages 408–415. Omnipress, 2008.

[16] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira,

Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[17] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143. PMLR, 2020.

[18] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtarik. Tighter theory for local SGD on identical and heterogeneous data. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4519–4529. PMLR, 2020.

[19] Donghwan Kim and Jeffrey A. Fessler. Adaptive restart of the optimized gradient method for convex optimization. *Journal of Optimization Theory and Applications*, 178(1):240–263, 2018.

[20] Donghwan Kim and Jeffrey A. Fessler. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *Journal of Optimization Theory and Applications*, 188(1):192–219, 2021.

[21] Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical Programming*, 171(1):167–215, 2018.

[22] Andreas Langer and Fernando Gaspoz. Overlapping domain decomposition methods for total variation denoising. *SIAM Journal on Numerical Analysis*, 57(3):1411–1444, 2019.

[23] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[24] Chang-Ock Lee and Changmin Nam. Primal domain decomposition methods for the total variation minimization, based on dual decomposition. *SIAM Journal on Scientific Computing*, 39(2):B403–B423, 2017.

[25] Chang-Ock Lee and Jongho Park. Fast nonoverlapping block Jacobi method for the dual Rudin–Osher–Fatemi model. *SIAM Journal on Imaging Sciences*, 12(4):2009–2034, 2019.

[26] Chang-Ock Lee and Jongho Park. A dual-primal finite element tearing and interconnecting method for nonlinear variational inequalities utilizing linear local problems. *International Journal for Numerical Methods in Engineering*, 122(22):6455–6475, 2021.

[27] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[28] Grigory Malinovsky, Kai Yi, and Peter Richtárik. Variance reduced ProxSkip: Algorithm, theory and application to federated learning. In *Advances in Neural Information Processing Systems*, 2022.

[29] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 2017.

[30] Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtarik. ProxSkip: Yes! Local gradient steps provably lead to communication acceleration! Finally! In *Proceedings of the 39th International Conference on Machine Learning*, pages 15750–15769. PMLR, 2022.

[31] Aritra Mitra, Rayana Jaafar, George J. Pappas, and Hamed Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. In *Advances in Neural Information Processing Systems*, volume 34, pages 14606–14619. Curran Associates, Inc., 2021.

[32] Arkaddii S. Nemirovskii and Yu. E. Nesterov. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 25(2):21–30, 1985.

[33] Yu. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[34] Yurii Nesterov. How to make the gradients small. *Optima*, 88:10–11, 2012.

[35] Jongho Park. Additive Schwarz methods for convex optimization as gradient methods. *SIAM Journal on Numerical Analysis*, 58(3):1495–1530, 2020.

[36] Jongho Park. Fast gradient methods for uniformly convex and weakly smooth problems. *Advances in Computational Mathematics*, 48:Paper No. 34, 2022.

[37] Reese Pathak and Martin J. Wainwright. FedSplit: an algorithmic framework for fast federated optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 7057–7066. Curran Associates, Inc., 2020.

[38] Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18250–18280. PMLR, 2022.

[39] Simone Rebegoldi and Luca Calatroni. Scaled, inexact, and adaptive generalized FISTA for strongly convex optimization. *SIAM Journal on Optimization*, 32(3):2428–2459, 2022.

[40] Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, New Jersey, 2015.

[41] Ralph Tyrrell Rockafellar and Roger Jean-Baptiste Wets. *Variational Analysis*, volume 317. Springer, Berlin, 2009.

[42] Abdurakhmon Sadiev, Dmitry Kovalev, and Peter Richtárik. Communication acceleration of local gradient methods via an accelerated primal-dual algorithm with an inexact prox. In *Advances in Neural Information Processing Systems*, 2022.

[43] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, Cambridge, 2014.

[44] Vladimir Spokoiny. Parametric estimation. Finite sample theory. *The Annals of Statistics*, 40(6):2877–2909, 2012.

[45] Xue-Cheng Tai and Jinchao Xu. Global and uniform convergence of subspace correction methods for some convex optimization problems. *Mathematics of Computation*, 71(237):105–124, 2002.

[46] Quoc Tran-Dinh, Nhan Pham, Dzung T. Phan, and Lam M. Nguyen. FedDR – randomized Douglas–Rachford splitting algorithms for nonconvex federated composite optimization. In *Advances in Neural Information Processing Systems*, 2021.

[47] Jinchao Xu. Iterative methods by space decomposition and subspace correction. *SIAM Review*, 34(4):581–613, 1992.

[48] Jing Xu, Sen Wang, Liwei Wang, and Andrew Chi-Chih Yao. FedCM: Federated learning with client-level momentum. *arXiv preprint arXiv:2106.10874*, 2021.

[49] Hsiang-Fu Yu, Fang-Lan Huang, and Chih-Jen Lin. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1):41–75, 2011.

[50] Junyu Zhang and Lin Xiao. A composite randomized incremental gradient method. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7454–7462. PMLR, 2019.

[51] Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. FedPD: A federated learning framework with adaptivity to non-iid data. *IEEE Transactions on Signal Processing*, 69:6055–6070, 2021.

A Fenchel–Rockafellar duality

In this appendix, we present key features of the Fenchel–Rockafellar duality for the sake of completeness; see also [11, 40]. For a proper function $F\colon X \to \overline{\mathbb{R}}$ defined on a Euclidean space X, the effective domain $\operatorname{dom} F$ of F is defined by

$$\operatorname{dom} F = \{x \in X : F(x) < \infty\} .$$

Recall that the Legendre–Fenchel conjugate of F is denoted by $F^*\colon X \to \overline{\mathbb{R}}$, i.e.,

$$F^*(p) = \sup_{x \in X} \{\langle p, x\rangle - F(x)\} , \quad p \in X.$$

One may refer to [40] for the elementary properties of the Legendre–Fenchel conjugate. In Proposition A.1, we summarize the notion of the Fenchel–Rockafellar duality [40, Corollary 31.2.1], which plays an important role in the convergence analysis of `DualFL`.

Proposition A.1 (Fenchel–Rockafellar duality). *Let X and Y be Euclidean spaces. Consider the minimization problem*

$$\min_{x \in X} \{F(x) + G(Kx)\} , \tag{A.1}$$

where $K\colon X \to Y$ is a linear operator and $F\colon X \to \overline{\mathbb{R}}$ and $G\colon Y \to \overline{\mathbb{R}}$ are proper, convex, and lower semicontinuous functions. If there exists $x_0 \in X$ such that x_0 is in the relative interior of $\operatorname{dom} F$ *and Kx_0 is in the relative interior of* $\operatorname{dom} G$*, then the following relation holds:*

$$\min_{x \in X} \{F(x) + G(Kx)\} = - \min_{y \in Y} \left\{F^*(-K^{\mathrm{T}} y) + G^*(y)\right\} .$$

Moreover, the primal solution $x^* \in X$ and the dual solution $y^* \in Y$ satisfy

$$-K^{\mathrm{T}} y^* \in \partial F(x^*), \quad Kx^* \in \partial G(y^*). \tag{A.2}$$

Leveraging the Fenchel–Rockafellar duality, we are able to derive the dual formulation (6.1) from the model federated learning problem (2.1). For a positive constant ν, the problem (2.1) can be rewritten as follows:

$$\min_{\theta \in \Omega} \left\{ \frac{1}{N} \sum_{j=1}^{N} g_j(\theta) + \frac{\nu}{2} \|\theta\|^2 \right\} , \tag{A.3}$$

where $g_j(\theta) = f_j(\theta) - \frac{\nu}{2}\|\theta\|^2$. Under Assumption 2.1, each g_j is convex if $\nu \in (0, \mu]$. In (A.1), if we set

$$X = \Omega, \quad Y = \Omega^N, \quad K = \begin{bmatrix} I \\ \vdots \\ I \end{bmatrix}, \quad F(\theta) = \frac{N\nu}{2}\|\theta\|^2, \quad G(\boldsymbol{\xi}) = \sum_{j=1}^{N} g_j(\xi_j),$$

for $\theta \in \Omega$ and $\boldsymbol{\xi} \in \Omega^N$, then we obtain (A.3). Here, I is the identity matrix on Ω. By the definition of the Legendre–Fenchel conjugate, we readily get

$$F^*(\theta) = \frac{1}{2N\nu}\|\theta\|^2, \quad G^*(\boldsymbol{\xi}) = \sum_{j=1}^{N} g_j^*(\xi_j).$$

Hence, invoking Proposition A.1 yields the dual problem (6.1). Invoking (A.2), we obtain the primal-dual relation (6.2) between the primal solution θ^* and the dual solution $\boldsymbol{\xi}^*$.

B Analysis for strongly convex problems

In this appendix, we provide the missing proofs in Section 6 to complete the convergence analysis of the proposed `DualFL` in the strongly convex regime. We first state the convergence theorems of Algorithm 2, which are essential ingredients for the convergence analysis of `DualFL`. Recall that, if Assumption 2.1 is valid and $\nu \in (0, \mu]$, then F_{d} in (6.3) is ν^{-1}-smooth. Hence, we have the following convergence theorem of Algorithm 2 under Assumption 2.1 [39, Corollary 3.3].

Proposition B.1. *Suppose that Assumption 2.1 holds. In addition, suppose that the error sequence $\{\delta_n\}$ in Algorithm 2 is given by*

$$\delta_n = \frac{b_n}{(n+1)^2}, \quad n \geq 0,$$

where $\{b_n\}$ satisfies $\sum_{n=0}^{\infty} \sqrt{b_n} < \infty$. If we choose the hyperparameters ρ and ν in Algorithm 2 such that $\rho = 0$ and $\nu \in (0, \mu]$, then we have

$$E_{\mathrm{d}}(\boldsymbol{\xi}^{(n)}) - E_{\mathrm{d}}(\boldsymbol{\xi}^*) \lesssim \frac{1}{n^2}, \quad n \geq 0.$$

If we further assume that Assumption 2.2 holds, then G_{d} in (6.3) is $(L-\nu)^{-1}$-strongly convex. In this case, we have the following improved convergence theorem for Algorithm 2 [39, Corollary 3.4].

Proposition B.2. *Suppose that Assumptions 2.1 and 2.2 hold. In addition, suppose that the error sequence $\{\delta_n\}$ in Algorithm 2 is given by*

$$\delta_n = a^n, \quad n \geq 0,$$

where $a \in [0, 1-\sqrt{\rho})$. If we choose the hyperparameters ρ and ν in Algorithm 2 such that $\rho \in (0, \nu/L]$ and $\nu \in (0, \mu]$, then we have

$$E_{\mathrm{d}}(\boldsymbol{\xi}^{(n)}) - E_{\mathrm{d}}(\boldsymbol{\xi}^*) \lesssim (1-\sqrt{\rho})^n, \quad n \geq 0.$$

The dual problem (6.1) has a particular structure that allows Algorithm 2 to be viewed as a parallel subspace correction method for (6.1) [35, 45, 47]. That is, the proximal problem (6.4) can be decomposed into N independent subproblems, each defined in terms of ξ_j for $1 \leq j \leq N$. Specifically, Lemma B.3 shows that Algorithm 2 is equivalent to the prerelaxed block Jacobi method, which was introduced in [25].

Lemma B.3. *In Algorithm 2, suppose that $\tilde{\boldsymbol{\xi}}^{(n+1)} \in \Omega^N$ satisfies*

$$\tilde{\xi}_j^{(n+1)} \approx \operatorname*{arg\,min}_{\xi_j \in \Omega} \left\{ \tilde{E}_d^{n,j}(\xi_j) := g_j^*(\xi_j) + \frac{1}{2\nu} \left\| \xi_j - \eta_j^{(n)} + \frac{1}{N} \sum_{i=1}^N \eta_i^{(n)} \right\|^2 \right\}$$

such that $\tilde{E}_d^{n,j}(\tilde{\xi}_j^{(n+1)}) - \min \tilde{E}_d^{n,j} \leq \delta_n / N$ for $1 \leq j \leq N$. Then $\tilde{\boldsymbol{\xi}}^{(n+1)}$ solves the proximal problem (6.4) *such that $E_{\mathrm{d}}^n(\tilde{\boldsymbol{\xi}}^{(n+1)}) - \min E_{\mathrm{d}}^n \leq \delta_n$.*

Proof of Lemma B.3. By direct calculation, we get

$$\sum_{j=1}^N \tilde{E}_{\mathrm{d}}^{n,j}(\xi_j) = E_{\mathrm{d}}^n(\boldsymbol{\xi}) + \text{constant}$$

for any $\boldsymbol{\xi} \in \Omega^N$, which completes the proof. □

Using Lemma B.3, we can prove Lemma 6.1, which establishes the duality relation between Algorithm 2 and `DualFL`, as follows.

Proof of Lemma 6.1. It suffices to show that the sequences $\{\boldsymbol{\xi}^{(n)}\}$ and $\{\boldsymbol{\eta}^{(n)}\}$ defined in (6.6) satisfy (6.4) and (6.5). We first observe that

$$\sum_{i=1}^N \zeta_i^{(n)} = 0, \quad n \geq 0, \tag{B.1}$$

which can be easily derived by mathematical induction with (3.2) and (3.3). Now, we take any $n \geq 0$ and $1 \leq j \leq N$. By direct calculation, we obtain

$$\begin{aligned}
\sum_{i=1}^N \eta_i^{(n)} &\overset{\text{(6.6b)}}{=} \nu \sum_{i=1}^N \zeta_i^{(n)} - \nu N(1+\beta_n)\theta^{(n)} + \nu N \beta_n \theta^{(n-1)} \\
&\overset{\text{(B.1)}}{=} -N\nu(1+\beta_n)\theta^{(n)} + N\nu\beta_n\theta^{(n-1)} \\
&\overset{\text{(6.6b)}}{=} N\eta_j^{(n)} - N\nu\zeta^{(n)}.
\end{aligned}$$

Hence, we get

$$\nu\zeta_j^{(n)} = \eta_j^{(n)} - \frac{1}{N}\sum_{i=1}^{N}\eta_i^{(n)}. \tag{B.2}$$

Combining (3.5), (B.2), and Lemma B.3 implies that $\{\boldsymbol{\xi}^{(n)}\}$ and $\{\boldsymbol{\eta}^{(n)}\}$ satisfy (6.4). On the other hand, we obtain by direct calculation that

$$\begin{aligned}(1+\beta_n)\xi_j^{(n+1)} - \beta_n\xi_j^{(n)} &\overset{(6.6a)}{=} \nu(1+\beta_n)(\zeta_j^{(n)} - \theta_j^{(n+1)}) - \nu\beta_n(\zeta_j^{(n-1)} - \theta_j^{(n)}) \\ &\overset{(3.3)}{=} \nu\zeta_j^{(n+1)} - \nu(1+\beta_n)\theta^{(n+1)} - \nu\beta_n\theta^{(n)} \\ &\overset{(6.6b)}{=} \eta_j^{(n+1)},\end{aligned}$$

which implies that $\{\boldsymbol{\xi}^{(n)}\}$ and $\{\boldsymbol{\eta}^{(n)}\}$ satisfy (6.5). This completes the proof. □

Finally, we present the proof of our main convergence theorems for `DualFL`, Theorems 3.2 and 3.3.

Proof of Theorems 3.2 and 3.3. Thanks to Lemma 6.1, the sequence $\{\boldsymbol{\xi}^{(n)}\}$ defined in (6.6a) satisfies the convergence properties given in Propositions B.1 and B.2, i.e.,

$$E_{\mathrm{d}}(\boldsymbol{\xi}^{(n)}) - E_{\mathrm{d}}(\boldsymbol{\xi}^*) \lesssim \begin{cases} \dfrac{1}{n^2}, & \text{in the case of Theorem 3.2,} \\ (1-\sqrt{\rho})^n, & \text{in the case of Theorem 3.3.} \end{cases} \tag{B.3}$$

Next, we derive an estimate for the primal norm error $\|\theta^{(n)} - \theta^*\|$ by a similar argument as in of [26, Corollary 1]. Note that the dual cost function E_{d} given in (6.1) is $\frac{1}{N\nu}$-strongly convex relative to a seminorm $|\boldsymbol{\xi}| = \|\sum_{j=1}^{N}\xi_j\|$. Hence, by (6.2), (6.6a), and (B.1), we obtain

$$\|\theta^{(n)} - \theta^*\|^2 = \frac{1}{N^2\nu^2}\left\|\sum_{j=1}^{N}\left(\xi_j^{(n)} - \xi_j^*\right)\right\|^2 \leq \frac{2}{N\nu}\left(E_{\mathrm{d}}(\boldsymbol{\xi}^{(n)}) - E_{\mathrm{d}}(\boldsymbol{\xi}^*)\right). \tag{B.4}$$

Meanwhile, it is clear that

$$E(\theta^{(n)}) - E(\theta^*) \leq \frac{L}{2}\|\theta^{(n)} - \theta^*\|^2 \tag{B.5}$$

under Assumption 2.2. Combining (B.3), (B.4), and (B.5) completes the proof. □

C Analysis for non-strongly convex problems

This appendix is devoted to the complete proofs of Theorem 4.1 and Theorem 4.2, the convergence theorems of `DualFL` in the non-strongly convex regime. We first present the proof of Theorem 4.1, which is based on the epigraphical convergence theory developed in [41].

Proof of Theorem 4.1. It is clear that E^α decreases to E as $\alpha \to 0^+$. Hence, by [41, Proposition 7.4], E^α epi-converges to E. Since E is coercive, we conclude by [41, Theorem 7.33] that

$$E(\theta^\alpha) \to E(\theta^*) \quad \text{as } \alpha \to 0^+. \tag{C.1}$$

On the other hand, Theorem 3.2 implies that $\theta^{(n)} \to \theta^\alpha$ as $n \to \infty$. As E is continuous, we have

$$E(\theta^{(n)}) \to E(\theta^\alpha) \quad \text{as } n \to \infty. \tag{C.2}$$

Combining (C.1) and (C.2) yields

$$E(\theta^{(n)}) - E(\theta^*) \to 0 \quad \text{as } n \to \infty \text{ and } \alpha \to 0^+,$$

which is our desired result. □

Next, we provide the proof of Theorem 4.2, which states the communication complexity of `DualFL` in the smooth non-strongly convex regime.

Table 2: Description of the hyperparameters appearing in the benchmark algorithms FedPD, FedDR, FedDyn, and Scaffnew, and the proposed DualFL. We use the notation for each hyperparameter as given in the original paper. The value of each hyperparameter is determined using a grid search.

Algorithm	Hyper-param.	Description	Grid	Value
FedPD [51]	η	Local penalty parameter	$\{10^{-m} : m \in \mathbb{Z}_{\geq 0}\}$	10^{-4}
FedDR [46]	η	Local penalty parameter	$\{10^{-m} : m \in \mathbb{Z}_{\geq 0}\}$	10^{-4}
	α	Overrelaxation parameter	$\{1, 2\}$	1
FedDyn [1]	α	Local regularization parameter	$\{10^{m} : m \in \mathbb{Z}_{\geq 0}\}$	10^{3}
Scaffnew [30]	γ	Learning rate	$\{10^{-m} : m \in \mathbb{Z}_{\geq 0}\}$	10^{-5}
	p	Communication probability	$\{0.01, 0.05, 0.1, 0.5, 1\}$	0.1
DualFL	ρ	Momentum parameter	$\{m \times 10^{-3} : m \in \mathbb{Z}_{\geq 0}\}$	3×10^{-3}
	ν	Parameter to establish duality	$\{\mu\}$	μ

Proof of Theorem 4.2. Since θ^α minimizes E^α, we get

$$\nabla E^\alpha(\theta^\alpha) = \nabla E(\theta^\alpha) + \alpha\theta^\alpha = 0. \tag{C.3}$$

By the triangle inequality, Assumption 2.2, (C.3), and Theorem 3.3, we obtain

$$\begin{aligned}\|\nabla E(\theta^{(n)})\| &\leq \|\nabla E(\theta^{(n)}) - \nabla E(\theta^\alpha)\| + \|\nabla E(\theta^\alpha)\| \\ &\leq L\|\theta^{(n)} - \theta^\alpha\| + \alpha\|\theta^\alpha\| \\ &\lesssim \left(1 - \sqrt{\frac{\alpha}{L+\alpha}}\right)^{\frac{n}{2}} + \alpha\|\theta^\alpha\|,\end{aligned}$$

which proves (4.3).

Next, we proceed similarly as in [36, Theorem 3.3]. Let $\epsilon \in (0, 2R_0\alpha_0]$ and $\alpha = \epsilon/(2R_0)$, so that we have $\alpha \leq \alpha_0$ and $\|\theta^\alpha\| \leq R_0$ by (4.2). Then we obtain

$$\|\nabla E(\theta^{(n)})\| \leq C\left(1 - \sqrt{\frac{\epsilon}{\epsilon + 2LR_0}}\right)^{\frac{n}{2}} + \frac{\epsilon}{2} \leq C\left(1 + \sqrt{\frac{\epsilon}{\epsilon + 2LR_0}}\right)^{-\frac{n}{2}} + \frac{\epsilon}{2},$$

where C is a positive constant independent of n. Consequently, M_{comm} is determined by the following equation:

$$C\left(1 + \sqrt{\frac{\epsilon}{\epsilon + 2LR_0}}\right)^{-\frac{M_{\mathrm{comm}}}{2}} = \frac{\epsilon}{2}.$$

It follows that

$$M_{\mathrm{comm}} = \frac{2\log\frac{2C}{\epsilon}}{\log\left(1 + \sqrt{\frac{\epsilon}{\epsilon+2LR_0}}\right)} \leq \left(1 + 2\sqrt{1 + \frac{2LR_0}{\epsilon}}\right)\left(\log\frac{1}{\epsilon} + \log 2C\right),$$

where we used an elementary inequality [36, Equation (3.5)]

$$\log\left(1 + \frac{1}{t}\right) \geq \frac{2}{2t+1}, \quad t > 0.$$

This proves (4.4). □

D Experiment details

In this appendix, we present the full details of the numerical experiments conducted in Section 7. All the algorithms were programmed using MATLAB R2022b and performed on a desktop equipped with AMD Ryzen 5 5600X CPU (3.7GHz, 6C), 40GB RAM, NVIDIA GeForce GTX 1660 SUPER GPU with 6GB GDDR6 memory, and the operating system Windows 10 Pro.

The multinomial logistic regression problem is stated as

$$\min_{\theta=(w,b)\in\mathbb{R}^{d\times k}\times\mathbb{R}^k} \left\{ \frac{1}{n}\sum_{j=1}^n \log\left(\sum_{l=1}^k e^{(w_l\cdot x_j + b_l) - (w_{y_j}\cdot x_j + b_{y_j})}\right) + \frac{\mu}{2}\|\theta\|^2 \right\}, \tag{D.1}$$

where $\{(x_j, y_j)\}_{j=1}^n \subset \mathbb{R}^d \times \{1, \dots, k\}$ is a labeled dataset. In (D.1), we set the regularization parameter $\mu = 10^{-2}$. We use the MNIST training dataset [23]; we have $k = 10$, $n = 60,000$, and $d = 784$. We assume that the dataset is evenly distributed to N clients to form $f_1, \dots, f_N$, so that (D.1) is expressed in the form (2.1). A reference solution $\theta^* \in \mathbb{R}^{(d+1)\times k}$ of (D.1) is obtained by a sufficient number of damped Newton iterations [9].

As we mentioned in Section 7, we choose the following recent federated learning algorithms as benchmarks: `FedPD` [51], `FedDR` [46], `FedDyn` [1], and `Scaffnew` [30]. All the hyperparameters appearing in these algorithms are tuned by a grid search; see Table 2 for details of the tuned hyperparameters. To solve the local problems encountered in these algorithms, we employ the optimized gradient method with adaptive restart (`AOGM`) proposed in [19], with the stop criterion in which the algorithm terminates when the relative energy difference becomes less than 10^{-12}. In each iteration of `AOGM`, the step size is determined using the full backtracking scheme introduced in [10]. Finally, the hyperparameters of `DualFL` are chosen as $\rho = 3 \times 10^{-3}$ and $\nu = \mu$ unless otherwise stated, where the value of ρ is obtained by a grid search.

Remark D.1. While we also conducted experiments with several primal federated learning algorithms such as `FedAvg` [29], `FedCM` [48], and `FedSAM` [38], which do not rely on duality in their mechanisms, we do not present their results as their performances were not competitive compared to other methods.