

Abstract

최근 모바일 기기와 IoT(Internet of Things) 센서의 폭발적인 보급으로 인해 데이터가 네트워크 엣지(Edge)에서 분산 생성되는 환경이 조성되었다. 이러한 환경에서 연합학습(Federated Learning, FL)은 원본 데이터를 중앙 서버로 전송하지 않고 로컬에서 모델을 학습시킨 후 업데이트(Gradient)만을 공유함으로써 프라이버시를 보호하는 핵심 기술로 주목받고 있다. 그러나 연합학습의 구조적 특성상 공유되는 모델 업데이트를 통해 원본 데이터를 복원하는 역전파 공격(Inversion Attack)이나 데이터의 포함 여부를 식별하는 멤버십 추론 공격(Membership Inference Attack) 등의 보안 위협이 여전히 존재한다. 이를 방어하기 위해 차분 프라이버시 (Differential Privacy, DP) 기술이 도입되었으나, 많은 기존 연구들은 클라이언트 참여 패턴을 균일하거나 독립적이라고 단순화하거나, 참여 빈도와 무관하게 고정된(Static) 노이즈를 적용하는 한계를 보였다. 이는 빈번하게 참여하는 클라이언트의 프라이버시 예산을 초기에 고갈시키거나, 간헐적으로 참여하는 클라이언트의 유tility 기여도를 불필요하게 억제하는 “프라이버시-유tility 딜레마”를 야기한다.

본 논문에서는 실제 무선 네트워크 환경에서 필연적으로 발생하는 시스템 이질성(System Heterogeneity), 즉 클라이언트 간 참여 빈도의 불균형 문제에 주목한다. 이러한 문제를 해결하기 위해 본 연구에서는 **QuAP-FL (Quantile-based Adaptive Privacy for Federated Learning)** 프레임워크를 제안한다. QuAP-FL은 (1) 클라이언트의 누적 참여 이력을 실시간으로 추적하고 선택된 클라이언트 집합의 참여율 통계를 기반으로 라운드별 프라이버시 예산을 조절하는 적응형 프라이버시 예산(Adaptive Privacy Budgeting) 기법, (2) 매 라운드 수집된 그레이디언트 노름(Norm)의 분포를 분석하여 정보 손실을 최소화하는 클리핑 임계값을 자동으로 조절하는 분위수 기반 클리핑(Quantile-based Clipping), 그리고 (3) 딥러닝 모델의 구조적 특성을 고려하여 분류기 전체(Classifier Head)에 집중적으로 노이즈를 주입함으로써 고차원 노이즈로 인한 성능 저하를 완화하는 계층별 노이즈 주입(Layer-wise Noise Injection) 전략을 통합한다.

Non-IID 데이터 분포를 가진 MNIST 및 CIFAR-10 데이터셋에 대한 실험 결과, QuAP-FL은 MNIST에서 93.30%의 정확도를 달성하였다. 이는 프라이버시 보호가 없는 FedAvg(93.26%) 와 대등하며, 고정형 DP 방식(Fixed-DP, 93.76%) 대비 0.46%p 낮지만, QuAP-FL은 참여 이력 추적, 분위수 기반 클리핑, 적응형 예산 조절 등 추가적인 메커니즘을 수행하면서도 이러한 미미한 성능 저하에 그쳤다는 점에서 의의가 있다. 특히 CIFAR-10에서는 QuAP-FL(76.82%) 이 Fixed-DP(75.80%) 대비 1.02%p 높은 성능을 보여, 복잡한 데이터셋 일수록 적응형 메커니즘의 이점이 부각될 가능성을 확인하였다. 본 연구의 결과는 적절한 수준의 DP 노이즈가 모델의 과적합(Overfitting)을 완화하는 정규화(Regularization) 효과를 제공하여, Non-IID 환경에서 일반화 성능을 높일 수 있음을 시사한다. 본 연구는 참여 패턴이 불균형한 현실적인 연합학습 환경에서 프라이버시 보장과 모델 유tility 사이의 실용적인 균형점을 탐색하는 프레임워크를 제시한다.

키워드: 연합학습, 차분 프라이버시, 적응형 클리핑, 시스템 이질성, 프라이버시 예산, 엣지 컴퓨팅

1 서론 (Introduction)

1.1 연구 배경: 엣지 컴퓨팅과 데이터 프라이버시

현대 사회는 '데이터의 시대'라고 불릴 만큼 방대한 양의 데이터가 매순간 생성되고 있다. 스마트폰, 웨어러블 기기, 자율주행 자동차, 스마트 홈 IoT 등 엣지(Edge) 디바이스의 보급은 기하급수적으로 증가하고 있으며, Cisco의 보고서에 따르면 2025년까지 전 세계적으로 750억 개 이상의 IoT 기기가 연결될 것으로 예상된다. 이들 기기에서 생성되는 데이터는 개인의 행동 패턴, 건강 정보(심박수, 수면 패턴), 정밀한 위치 정보, 금융 거래 내역, 음성 및 영상 데이터 등 민감한 사적 정보를 대량으로 포함하고 있다.

전통적인 중앙 집중식 머신러닝(Centralized Machine Learning)은 이러한 데이터를 데이터 센터나 클라우드 서버로 수집하여 학습하는 방식을 취해왔다. 그러나 이러한 중앙 집중식 접근법은 데이터 전송 과정에서의 네트워크 대역폭 비용 문제, 중앙 서버의 스토리지 비용 문제, 그리고 무엇보다 심각한 **프라이버시 침해 우려**를 낳는다. 데이터가 서버로 전송되는 순간 사용자는 데이터에 대한 통제권을 상실하게 되며, 서버의 보안이 뚫릴 경우 대규모 개인정보 유출 사고로 이어질 수 있다. 특히 유럽의 GDPR(General Data Protection Regulation), 미국의 CCPA(California Consumer Privacy Act)와 같은 데이터 보호 규제가 강화됨에 따라, 원본 데이터를 서버로 전송하는 것은 법적, 윤리적으로 더욱 어려워지고 있다.

이러한 배경 속에서 구글(Google)이 2016년 제안한 **연합학습(Federated Learning, FL)**은 "데이터가 이동하는 대신 모델이 이동한다"는 패러다임을 제시했다 [McMahan et al., 2017]. 연합학습에서는 각 클라이언트가 자신의 로컬 데이터를 이용하여 모델을 학습시키고, 학습된 모델의 파라미터 업데이트(Gradient 또는 Weight Difference)만을 서버로 전송한다. 서버는 수집된 업데이트들을 안전하게 집계(Aggregation)하여 전역 모델(Global Model)을 갱신하고, 이를 다시 클라이언트들에게 배포한다. 이 과정에서 원본 데이터는 기기 외부로 유출되지 않으므로, 구조적으로 프라이버시 보호에 유리한 것으로 여겨졌다.

1.2 연합학습의 프라이버시 위협과 한계

그러나 최근의 연구들은 연합학습이 제공하는 프라이버시 보호가 완벽하지 않음을 수학적, 실험적으로 보여주었다. 공격자는 공유된 그래디언트나 모델 파라미터만으로도 원본 데이터를 복원하거나, 특정 데이터의 속성을 추론할 수 있다. 주요 공격 유형은 다음과 같다.

1. **역전파 공격 (Inversion Attack):** Zhu et al. [2019]의 Deep Leakage from Gradients (DLG) 연구는 공유된 그래디언트만으로 픽셀 단위의 원본 이미지를 거의 완벽하게 복원할 수 있음을 보였다. 딥러닝 모델의 그래디언트는 손실 함수를 모델 파라미터에 대해 미분한 값으로서, 학습 데이터의 특징 정보를 고스란히 담고 있기 때문이다. 이후 Zhao et al. [2020]는 DLG의 수렴 속도와 정확도를 개선한 iDLG를 제안하였으며, 이는 라벨 정보를 먼저 복원함으로써 이미지 복원의 정확도를 높였다. 이러한 공격들은 수십 번의 반복 최적화(Optimization) 과정을 통해 더미 데이터와 더미 그래디언트 간의 거리를 최소화하는 방식으로 수행된다.
2. **멤버십 추론 공격 (Membership Inference Attack):** 특정 데이터 레코드가 학습에 사용되었는지를 확률적으로 추론하는 공격이다. Shokri et al. [2017]에 의해 체계화된 이 공격은, 모델이 학습 데이터에 대해 보이는 확신도(Confidence)가 학습되지 않은 데이터보다 높다

는 점을 악용한다. 이는 의료 데이터(특정 환자가 암 데이터셋에 포함되었는지 여부)나 금융 데이터와 같이 민감한 정보가 포함된 경우 심각한 프라이버시 침해가 될 수 있다.

3. 속성 추론 공격 (Property Inference Attack): 개별 데이터가 아닌 학습 데이터셋의 전반적인 속성(예: 특정 인종의 비율, 특정 연령대의 분포)을 추론하는 공격이다. 이는 모델의 편향성(Bias)을 분석하는 데 사용될 수도 있지만, 기업의 기밀 정보(예: 제조 공정의 불량률)를 유추하는 산업 스파이 활동에도 악용될 수 있다.

이러한 공격 기법들의 고도화는 단순한 익명화(Anonymization)나 데이터 마스킹(Masking)만으로는 프라이버시를 보장할 수 없음을 시사한다. 따라서 수학적으로 엄밀한 안전성을 보장하는 차분 프라이버시의 적용이 중요해졌다.

이를 위해 차분 프라이버시(Differential Privacy, DP)가 연합학습의 표준 방어 기제로 자리 잡았다 [Dwork and Roth, 2014]. DP는 데이터셋에 임의의 노이즈를 추가하여, 특정 개별 데이터의 존재 여부가 출력 결과(모델 파라미터)에 미치는 영향을 수학적으로 제한한다. DP-FedAvg [Geyer et al., 2017]와 같은 알고리즘은 클라이언트가 전송하는 업데이트에 가우시안 노이즈(Gaussian Noise)를 주입하거나, 서버가 집계 후 노이즈를 주입하여 통계적 불확실성을 제공함으로써 프라이버시를 보장한다.

1.3 문제 정의: 시스템 이질성과 프라이버시-유tility 딜레마

기존 DP-FL 연구의 한계 중 하나는 클라이언트의 참여 패턴(Participation Pattern)을 단순화한다는 점이다. 많은 연구는 모든 클라이언트가 매 라운드 균등한 확률로 선택되거나(Uniform Sampling), 항상 참여 가능한 상태라고 가정한다. 그러나 실제 현실의 연합학습 환경은 극심한 시스템 이질성(System Heterogeneity)을 특징으로 한다 [Li et al., 2020].

- 네트워크 불안정성: 모바일 기기는 Wi-Fi 연결 상태나 이동 통신망(4G/5G)의 상태에 따라 참여 가능 여부가 수시로 변한다.
- 기기 가용성: 연합학습은 일반적으로 기기가 총전 중이고, Wi-Fi에 연결되어 있으며, 유휴 상태(Idle)일 때만 수행된다. 사용자의 생활 패턴에 따라 이러한 조건이 충족되는 빈도는 천차만별이다.
- 하드웨어 성능 차이: 최신 스마트폰과 구형 기기 간의 연산 속도 차이로 인해, 정해진 시간 내에 학습을 완료하지 못하는 스트래글러(Straggler)가 발생한다.

이러한 이질성은 필연적으로 참여 빈도의 불균형(Participation Imbalance)을 초래한다. 롱테일(Long-tail) 분포에 따라 어떤 클라이언트는 전체 학습 라운드의 50% 이상 참여하는 반면(Frequent Participant), 어떤 클라이언트는 1% 미만으로 참여한다(Sporadic Participant). 여기서 기존의 고정적 DP(Fixed DP) 방식은 다음과 같은 딜레마에 직면한다.

- 시나리오 A (보수적 접근 - Privacy First): 가장 자주 참여하는 클라이언트(Worst-case)를 기준으로 노이즈 크기를 설정하면, 프라이버시 예산을 맞추기 위해 노이즈가 지나치게 커지게 된다. 이는 모델이 전혀 학습되지 않는 유tility(Utility)의 붕괴를 초래한다.

- **시나리오 B (공격적 접근 - Utility First):** 평균적인 참여를 기준으로 노이즈를 설정하면, 자주 참여하는 클라이언트는 허용된 프라이버시 예산(ϵ)을 초과하여 민감한 정보가 유출될 위험에 처한다. 이는 **프라이버시(Privacy)**의 **붕괴**를 의미한다.

1.4 연구 목표 및 기여

본 연구의 목표는 시스템 이질성이 존재하는 현실적인 연합학습 환경에서, 클라이언트별 참여 특성이 반영된 통계에 따라 프라이버시 강도를 유동적으로 조절함으로써 전체 모델의 성능(Utility)을 유지하고 클라이언트 수준의 프라이버시(Privacy)를 관리하는 것이다. 본 논문의 주요 기여는 다음과 같다.

1. **참여율 기반 적응형 프라이버시 예산 할당 (Adaptive Privacy Budgeting):** 클라이언트의 누적 참여 횟수를 실시간으로 추적하고, 선택된 클라이언트 집합의 평균 참여율에 따라 라운드별 예산을 조절하는 동적 알고리즘을 제안한다. 라운드 t 에서 사용되는 프라이버시 예산은 참여율 통계에 의존하므로, 장기적으로 보면 참여율이 높은 클라이언트는 상대적으로 작은 예산(더 큰 노이즈)을, 참여율이 낮은 클라이언트는 더 큰 예산(더 작은 노이즈)을 경험하는 경향이 있다. 이는 "정보를 많이 제공한 자는 더 강하게 보호하고, 정보를 적게 제공한 자는 드문 업데이트가 모델에 의미 있게 반영되도록 한다"는 공평성 원칙에 기반한다.
2. **분위수 기반 적응형 클리핑 (Quantile-based Adaptive Clipping):** DP 적용을 위해서는 그래디언트의 크기(L2 Norm)를 제한하는 클리핑(Clipping)이 필수적이다. 기존의 고정 임계값 방식은 최적의 값을 찾기 위해 수많은 하이퍼파라미터 튜닝이 필요하다는 단점이 있다. 본 연구는 매 라운드 수집된 그래디언트 노름의 통계적 분포를 분석하여, 정보 손실을 최소화하는 90분위수(90th Percentile) 임계값을 동적으로 결정하는 기법을 도입한다.
3. **계층별 노이즈 주입 (Layer-wise Noise Injection):** 딥러닝 모델의 모든 파라미터에 노이즈를 주입할 경우, 차원의 저주(Curse of Dimensionality)로 인해 신호 대 잡음비(SNR)가 급격히 낮아져 모델 성능이 저하된다. 본 연구는 특징 추출기(Feature Extractor)보다 분류기(Classifier Head)에 민감한 정보가 집중된다는 점에 착안하여, 완전 연결 층(Fully Connected Layers) 전체에 집중적으로 노이즈를 가하는 전략을 제시한다. 이는 보안성과 유저 편의 사이의 실용적인 균형점이다.

2 관련 연구 (Related Work)

2.1 연합학습과 차분 프라이버시의 진화

McMahan et al. [2017]이 제안한 FedAvg는 로컬에서 SGD를 수행하고 서버에서 가중치를 평균내는 방식으로, 통신 효율성을 개선하며 연합학습의 표준으로 자리 잡았다. 그러나 앞서 언급한 프라이버시 위협에 대응하기 위해 다양한 DP 메커니즘이 결합되었다.

- **중앙 집중식 DP (Centralized DP):** 신뢰할 수 있는 서버(Trusted Server)를 가정하고, 서버가 집계된 결과에 노이즈를 추가하는 방식이다. Geyer et al. [2017]의 DP-FedAvg가 이에 해당하며, 모델 정확도 유지에 유리하지만 서버가 공격받을 경우 취약하다.

- **로컬 DP (Local DP):** 각 클라이언트가 자신의 데이터를 서버로 보내기 전에 노이즈를 추가하는 방식이다. Google의 RAPPOR [Erlingsson et al., 2014]가 대표적이며, 강력한 프라이버시를 제공하지만 노이즈가 누적되어 모델 성능이 크게 저하된다.

본 연구는 유틸리티와 프라이버시의 균형을 위해 중앙 집중식 DP 모델을 기반으로 하되, 시스템 이질성을 고려하여 이를 확장한다. 특히 Abadi et al. [2016]의 DP-SGD에서 제안된 Moments Accountant 기법은 프라이버시 손실 계산의 정밀함을 높였으나, 실무 구현에서는 여전히 고정된 노이즈 스케일을 사용하는 경우가 많다.

2.2 이질적 환경에서의 연합학습 최적화

시스템 이질성을 해결하기 위한 연구는 주로 최적화(Optimization) 관점에서 진행되었다. FedProx [Li et al., 2020]는 로컬 업데이트 함수에 근접 항(Proximal Term)을 추가하여, 로컬 모델이 전역 모델에서 너무 멀어지지 않도록 제약함으로써 스트래글러나 데이터 불균형으로 인한 모델 발산을 억제했다. 이는 $\min_w F(w) + \frac{\mu}{2} \|w - w^t\|^2$ 형태의 목적 함수를 최적화하는 것으로 해석할 수 있다. FedNova [Wang et al., 2020]는 클라이언트마다 수행하는 로컬 에포크(Epoch) 수가 다를 때 발생하는 업데이트의 편향을 정규화(Normalization)를 통해 해결했다. 각 클라이언트의 업데이트 스케일을 조정함으로써, 많이 학습한 클라이언트가 전역 모델을 지배하는 현상을 방지했다. SCAFFOLD [Karimireddy et al., 2020]는 제어 변수(Control Variate)를 도입하여 클라이언트 간의 드리프트를 교정했다. 로컬 업데이트 방향과 전역 업데이트 방향의 차이를 추정하여 보정함으로써, Non-IID 데이터에서도 안정적인 수렴을 유도했다. 그러나 이러한 연구들은 학습의 수렴성과 속도에 초점을 맞추었을 뿐, 이질적인 참여가 프라이버시 예산 소모에 미치는 영향은 직접적으로 고려하지 않았다. QuAP-FL은 이러한 최적화 기법들과 직교(Orthogonal)하는 연구로서, 필요에 따라 FedProx나 SCAFFOLD와 결합하여 사용될 수 있다.

2.3 적응형 차분 프라이버시 (Adaptive DP)

최근 들어 고정된 노이즈의 한계를 완화하기 위한 적응형 DP 연구들이 등장했다. Lee and Kifer [2018]는 반복(Iteration) 횟수에 따라 예산을 조절하는 기법을 제안했고, Andrew et al. [2021]는 그레디언트 클리핑 임계값을 적응적으로 조절하는 기법을 제안했다. 그러나 기존 연구들은 대부분 학습의 수렴 상태(Loss 값의 변화 등)나 모델의 구조적 특성(Layer별 파라미터 수)을 기준으로 노이즈를 조절했다. “누가 얼마나 자주 참여했는가”라는 사용자 중심의 참여 이질성을 명시적인 신호로 사용하여 라운드 단위 프라이버시 예산을 조정하는 시도는 상대적으로 적다. 본 연구의 QuAP-FL은 참여 이력을 능동적으로 추적하여 예산 스케줄링에 반영한다는 점에서 기존 연구들과 차별화된다.

3 제안 방법 (QuAP-FL Framework)

본 장에서는 QuAP-FL의 핵심 구성 요소인 참여 이력 추적기, 적응형 예산 할당, 분위수 기반 클리핑, 그리고 계층별 노이즈 주입에 대해 상세히 기술한다.

3.1 시스템 모델 및 문제 정식화

N 명의 클라이언트가 존재하는 연합학습 시스템을 고려한다. 전체 데이터셋 \mathcal{D} 는 N 개의 로컬 데이터셋 $\mathcal{D}_1, \dots, \mathcal{D}_N$ 으로 분할되어 있다. 각 라운드 t 마다 서버는 전체 클라이언트 중 일부인 클라이언트 집합 $S_t \subset \{1, \dots, N\}$ 를 확률적으로 선택하여 학습을 진행한다. 선택된 클라이언트 i 는 자신의 로컬 데이터 \mathcal{D}_i 를 사용하여 손실 함수 $\mathcal{L}(\theta; \mathcal{D}_i)$ 를 최소화하는 방향으로 모델 파라미터 θ 를 업데이트한다.

우리의 목표는 다음의 최적화 문제를 해결하는 것이다.

$$\min_{\theta} \mathcal{L}(\theta) = \sum_{i=1}^N \frac{|\mathcal{D}_i|}{|\mathcal{D}|} \mathcal{L}_i(\theta) \quad (1)$$

단, 학습 과정에서 교환되는 정보는 (ϵ, δ) -차분 프라이버시를 만족해야 한다.

3.2 참여 이력 추적기 (Participation Tracker)

시스템 이질성을 정량화하기 위해, 서버는 각 클라이언트의 참여 이력을 기록하는 **Participation Tracker** 모듈을 유지한다. 이는 벡터 $K \in \mathbb{R}^N$ 으로 표현되며, K_i 는 클라이언트 i 의 누적 참여 횟수를 의미한다. 라운드 t 시점에서의 클라이언트 i 의 참여율 $p_i(t)$ 는 다음과 같이 정의된다.

$$p_i(t) = \frac{k_i(t)}{t} \quad \text{where } k_i(t) \text{ is cumulative participation count} \quad (2)$$

단순한 누적 평균은 과거의 모든 기록을 동일한 가중치로 반영하지만, 실제 시스템에서는 최근의 참여 패턴이 더 중요할 수 있다. 따라서 선택적으로 지수 이동 평균(Exponential Moving Average, EMA) 방식을 적용할 수 있도록 설계하였다.

$$\hat{p}_i(t) = \eta \cdot I(i \in S_t) + (1 - \eta) \cdot \hat{p}_i(t-1) \quad (3)$$

여기서 $I(\cdot)$ 는 지시 함수(Indicator Function)이며, η 는 반영률이다. 본 논문의 메인 실험에서는 안정성을 위해 누적 평균 방식을 사용하였으나, 동적으로 변하는 네트워크 환경에서는 EMA 방식이 더 유리할 수 있음을 제언한다.

3.2.1 Warm-up Period의 도입

학습 초기(t 가 작을 때)에는 분모가 작아 참여율 $p_i(t)$ 가 0 또는 1로 극단적으로 변동하는 불안정성이 발생한다. 예를 들어, 첫 라운드에 참여한 클라이언트는 참여율이 100%가 되어 과도한 노이즈를 할당받게 된다. 이를 방지하기 위해 본 연구에서는 초기 T_{warm} 라운드 동안은 참여율 업데이트만 수행하고, 예산 할당에는 고정된 값을 사용하는 **Warm-up Period**를 도입한다. 본 실험에서는 $T_{warm} = 5$ 로 설정하여 초기 학습 통계가 일정 수준 축적될 때까지 고정 예산을 사용함으로써 학습의 초기 발산을 방지한다.

3.3 참여율 기반 적응형 프라이버시 예산 할당

기존 DP-FedAvg는 모든 라운드에 고정된 예산 ϵ_{fixed} 를 사용한다. 반면, QuAP-FL은 참여율 통계에 기반하여 라운드 t 에서 사용할 예산 ϵ_t 를 동적으로 결정한다. 각 라운드에서 참여율 통계를 다음과

같이 정의한다.

$$\bar{p}(t) = \frac{1}{|S_t|} \sum_{i \in S_t} p_i(t) \quad (4)$$

본 연구에서 사용하는 라운드 단위 예산 함수는 다음과 같은 지수 감소 함수(Exponential Decay Function)이다.

$$\epsilon_t = \epsilon_{base} \times (1 + \alpha \cdot \exp(-\beta \cdot \bar{p}(t))) \quad (5)$$

여기서 각 변수의 수학적, 물리적 의미는 다음과 같다.

- $\epsilon_{base} = \epsilon_{total}/T$: 전체 예산을 총 라운드 수로 균등 분배한 기본 할당량이다. 이는 참여율이 충분히 높아졌을 때 라운드별로 사용되는 최소 예산, 즉 가장 강력한 프라이버시 보호 수준을 의미한다.
- $\alpha \geq 0$ (Amplification Factor): 예산 증폭 계수이다. 참여율 평균이 0에 가까울 때 예산을 최대 $(1 + \alpha)$ 배까지 늘려준다. 본 연구에서는 $\alpha = 0.5$ 를 사용했다. 즉, 학습 초기나 참여율이 낮은 상태에서는 기본 예산보다 50% 더 많은 예산을 사용해 노이즈를 줄일 수 있다.
- $\beta > 0$ (Decay Rate): 감쇠 계수이다. 평균 참여율이 증가함에 따라 추가 예산이 얼마나 빨리 줄어들지를 결정한다. 본 연구에서는 $\beta = 2.0$ 을 사용하여, 참여율이 높아질수록 추가 예산이 급격히 사라지도록 설계했다.

라운드 t 에서 선택된 클라이언트 집합 S_t 는 동일한 예산 ϵ_t 를 공유하며, 서버는 이에 상응하는 가우시안 노이즈 스케일을 계산하여 집계된 업데이트에 적용한다. 개념적으로는 각 클라이언트별 참여율 $p_i(t)$ 에 따라 서로 다른 예산 $\epsilon_i(t)$ 를 부여하는 개인화된 설계도 가능하지만, 본 연구의 프로토 타입 구현에서는 구현의 단순성과 분석 용이성을 위해 평균 참여율 $\bar{p}(t)$ 에 기반한 라운드 단위 예산을 사용하였다. 이 설계에서 자주 참여하는 클라이언트는 여러 라운드에 걸쳐 상대적으로 작은 ϵ_t 를 반복적으로 경험하고, 드물게 참여하는 클라이언트는 상대적으로 큰 ϵ_t 가 사용되는 라운드에 한정적으로 참여하는 경향이 있어, 장기적으로는 “많이 참여한 클라이언트일수록 더 강한 보호를 받는다”는 방향의 통계적 경향을 유도한다.

본 논문에서는 δ 를 실험 전반에 걸쳐 작은 고정값(예: 10^{-5})으로 두고, 예산 스케줄링에 따른 ϵ_t 변화에 초점을 맞춘다.

3.4 분위수 기반 적응형 클리핑 (Quantile-based Adaptive Clipping)

차분 프라이버시를 적용하기 위해서는 개별 업데이트의 영향력(Sensitivity)을 제한하는 클리핑 과정이 필요하다. 즉, 그래디언트 g_i 의 L2 Norm이 임계값 C 를 넘지 않도록 조정해야 한다.

$$\bar{g}_i = g_i \cdot \min \left(1, \frac{C}{\|g_i\|_2} \right) \quad (6)$$

C 의 설정은 매우 중요하다. C 가 너무 작으면 그래디언트의 방향성이 왜곡되어 정보가 손실되고 (Bias 증가), 너무 크면 민감도(Sensitivity)가 커져서 추가해야 할 노이즈의 양이 늘어난다(Variance 증가). 최적의 C 는 학습이 진행됨에 따라 그래디언트의 크기가 줄어들기 때문에 동적으로 변한다. 본 연구에서는 매 라운드 수집된 그래디언트들의 Norm 분포를 분석하여 C 를 결정한다.

Algorithm 1 분위수 기반 적응형 클리핑 알고리즘

```
1: Input: 그래디언트 집합  $G_t = \{g_i\}_{i \in S_t}$ , 분위수  $q = 0.9$ , 모멘텀  $\gamma = 0.95$ 
2: 각 그래디언트의 L2 Norm 계산:  $N_t = \{\|g_i\|_2\}_{i \in S_t}$ 
3:  $N_t$ 의  $q$ -분위수(Percentile) 값 계산:  $C_{target} = \text{Percentile}(N_t, q)$ 
4: if  $t = 0$  then
5:    $C_t = C_{target}$ 
6: else
7:    $C_t = \gamma C_{t-1} + (1 - \gamma)C_{target}$                                 ▷ 지수 이동 평균(EMA)
8: end if
9: Output: 현재 라운드의 클리핑 임계값  $C_t$ 
```

90분위수(90th Percentile)를 사용하는 이유는 대다수의 정상적인 그래디언트는 보존하면서, 이상치(Outlier)에 해당하는 상위 10%의 과도한 그래디언트만 잘라내기 위함이다. 또한 EMA를 적용함으로써 라운드 간 클리핑 값의 급격한 변동을 억제하여 학습의 안정성을 도모한다.

3.5 계층별 노이즈 주입 (Layer-wise Noise Injection)

최신 딥러닝 모델은 수백만 개 이상의 파라미터를 가진다. 차분 프라이버시 메커니즘에 의해 주입되는 노이즈 벡터의 전체 크기(L2 Norm)는 파라미터 차원 d 의 제곱근에 비례하여 증가하므로 ($\propto \sqrt{d}$), 전체 모델의 모든 파라미터에 노이즈를 주입하면 신호 대비 잡음비(SNR)가 급격히 낮아져 유ти리티가 심각하게 훼손될 수 있다. 이를 차원의 저주(Curse of Dimensionality)라고 한다.

본 연구는 딥러닝 모델의 계층적 특성에 주목한다. CNN과 같은 모델에서 앞단(Convolutional Layers)은 옛지, 텍스처와 같은 일반적인 특징(Low-level Features)을 추출하며, 이 부분은 데이터셋에 크게 의존하지 않는 경향이 있다. 반면, 뒷단(Fully Connected Layers)은 추출된 특징을 바탕으로 구체적인 클래스를 결정하는 분류기 역할을 하며, 학습 데이터의 구체적인 정보가 많이 함축된 부분이다.

따라서 QuAP-FL은 실용성과 프라이버시 간의 타협안으로서, **분류기 헤드(Classifier Head) 전체에 집중적으로 노이즈를 주입하는 전략**을 취한다.

이론적 근거: 분류 레이어는 학습 데이터의 클래스 정보를 직접적으로 인코딩하므로, 역전파 공격에 상대적으로 취약하다. 반면, 앞단의 합성곱 층은 옛지나 텍스처와 같은 일반적인 특징(General Features)을 학습하며, 이는 데이터셋 간 전이 학습(Transfer Learning)이 가능하다는 점에서 데이터의 존성이 상대적으로 낮다. 본 연구는 Yosinski et al. [2014]의 연구 결과에 기반하여, 초기 계층의 프라이버시 민감도가 상대적으로 낮다고 가정한다.

실용적 고려: 전체 파라미터에 노이즈를 주입하면 $\mathcal{O}(\sqrt{d})$ 규모의 노이즈가 필요하여 유ти리티가 급격히 저하된다. 본 연구는 전체 파라미터의 약 3.5%(CIFAR-10 기준)에 해당하는 분류기 전체를 보호함으로써, 노이즈 크기를 줄여 성능을 유지한다.

$$\tilde{g}_{global} = g_{global} + [0, \dots, 0, \mathcal{N}(0, \sigma^2 I_{critical})] \quad (7)$$

한계 인정: 이는 엄밀한 의미의 “전체 모델”에 대한 (ϵ, δ) -차분 프라이버시를 보장하지 않는다. 민감한 정보가 집중된 분류기를 보호함으로써 실용적인 수준의 프라이버시 보호 효과를 제공하는 부분적 DP로 해석할 수 있다. 엄밀한 전체 모델 DP가 요구되는 경우, 희소화(Sparsification) 등의 기법을 병행해야 한다(향후 연구).

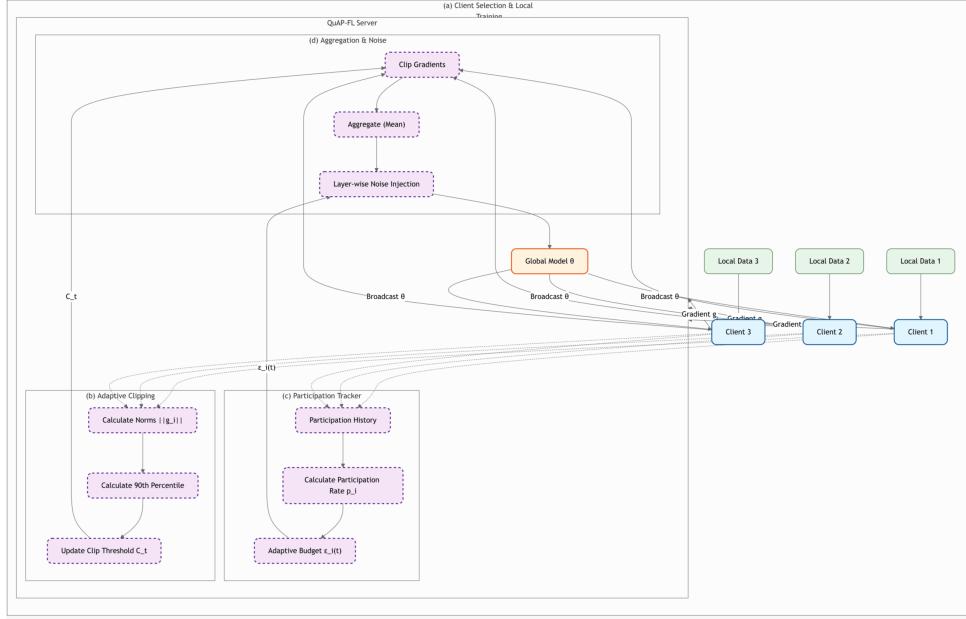


Figure 1: QuAP-FL 프레임워크 개요도. (a) 클라이언트 선택 및 로컬 학습, (b) 적응형 클리핑 값 업데이트, (c) 참여율 기반 예산 조절 및 계층별 노이즈 주입 과정을 보여준다.

본 실험에 사용된 모델의 경우, 전체 파라미터 대비 소수(CIFAR-10 약 3.5%, MNIST 약 0.1%)에 해당하는 분류기 파라미터에만 노이즈를 주입함으로써, 노이즈로 인한 성능 저하를 줄였다.

4 이론적 분석 (Theoretical Analysis)

4.1 프라이버시 손실 계산 (Privacy Accounting)

QuAP-FL의 프라이버시 보장 수준을 분석하기 위해 **기본 구성 정리(Basic Composition Theorem)**를 사용한다. 이는 가장 보수적인 상한선을 제공하므로 안전성을 강하게 보장한다.

라운드 t 에서 사용되는 예산 ϵ_t 는 식 (5)에 의해 결정되고, $\epsilon_t \leq \epsilon_{base}(1+\alpha) = \frac{\epsilon_{total}}{T}(1+\alpha)$ 가 항상 성립한다. 어떤 클라이언트 i 가 총 T 라운드 중 k_i 번 참여했고, 참여한 라운드 집합이 $\{t_1, \dots, t_{k_i}\}$ 라고 하자. 클라이언트 i 가 겪는 총 프라이버시 손실 $\epsilon_{total,i}$ 의 상한은 각 참여 라운드의 예산 합으로 표현된다.

$$\epsilon_{total,i} \leq \sum_{j=1}^{k_i} \epsilon_{t_j} \quad (8)$$

식 (5)에 의해 모든 t 에 대해 다음이 보장된다.

$$\epsilon_t \leq \epsilon_{base}(1 + \alpha) = \frac{\epsilon_{total}}{T}(1 + \alpha) \quad (9)$$

따라서 최악의 경우(Worst-case), 즉 클라이언트가 매 라운드 참여하는 경우($k_i = T$)에도 총 손실은 다음과 같이 제한된다.

$$\epsilon_{total,i} \leq \sum_{j=1}^T \epsilon_{t_j} \leq T \cdot \frac{\epsilon_{total}}{T}(1 + \alpha) = (1 + \alpha)\epsilon_{total} \quad (10)$$

본 실험에서는 $\alpha = 0.5$ 를 사용하였으므로, 가장 빈번하게 참여하는 클라이언트의 경우에도 $\epsilon_{total,i} \leq 1.5\epsilon_{total}$ 로 제한된다. 한편 대부분의 클라이언트는 $k_i \ll T$ 이므로, 라운드당 예산 ϵ_t 를 다소 크게 사용하더라도 총합은 $(1 + \alpha)\epsilon_{total}$ 보다 훨씬 작게 유지된다. δ 에 대해서는 각 라운드에서 동일한 δ 를 사용하고, 기본 구성 정리에 따라 총 δ 가 선형으로 증가하는 보수적인 상한을 가질 수 있다. 본 논문에서는 δ 를 충분히 작은 고정값으로 두고, 적응형 ϵ_t 설계에 따른 효과를 중심으로 논의한다.

4.2 유틸리티 상한 분석 (Utility Bound Analysis)

본 절에서는 QuAP-FL의 수렴 경향을 설명하기 위해 기존 DP-SGD 분석 결과를 간단히 인용한다. 표준 확률적 경사 하강법(SGD) 이론에 따르면, L -smooth하고 μ -strongly convex한 목적 함수 F 에 대해 적절한 학습률과 가우시안 노이즈를 사용하면 다음과 같은 상한이 주어진다 [Abadi et al., 2016].

정리 1 (DP-SGD의 수렴 특성 [Abadi et al., 2016]). F 가 L -smooth하고 μ -strongly convex라고 가정하자. 적절한 step size와 분산 σ^2 를 갖는 가우시안 노이즈를 사용하는 DP-SGD를 T 회 반복하면, 최적해 θ^* 와의 거대 오차는 다음과 같이 상한이 주어진다.

$$\mathbb{E}[F(\theta_T)] - F(\theta^*) \leq \mathcal{O}\left(\frac{1}{T}\right) + \mathcal{O}\left(\frac{\bar{\sigma}^2}{T}\right) \quad (11)$$

여기서 $\bar{\sigma}^2$ 는 평균 노이즈 분산이다.

QuAP-FL은 중앙집중식 DP 메커니즘 위에서 적응형 클리핑과 참여율 기반 노이즈 스케줄을 적용하는 구조이다. 본 논문에서 사용하는 모델은 CNN 기반 비볼록(non-convex) 모델이므로 위 정리가 QuAP-FL에 직접적으로 적용된다고 보기 어렵다. 다만, 평균 노이즈 분산 $\bar{\sigma}^2$ 가 작을수록 수렴 후 최종 오차가 작아지는 경향이 있다는 점을 보여주는 정성적 가이드라인으로 해석할 수 있다.

QuAP-FL의 설계는 다음 두 방향에서 $\bar{\sigma}^2$ 를 실용적인 수준으로 유지하는 것을 목표로 한다. 첫째, 참여율이 높아질수록 라운드별 예산 ϵ_t 가 감소하도록 설계하여, 학습이 진행될수록 노이즈 스케일을 점진적으로 줄인다. 둘째, 전체 파라미터가 아닌 분류기 헤드에 집중적으로 노이즈를 주입함으로써 고차원 노이즈로 인한 신호 대 잡음비 저하를 완화한다. 본 논문의 실험 결과는 이러한 설계가 Non-DP와 유사한 수준의 정확도를 유지하거나, 일부 설정에서는 오히려 약간 더 높은 정확도를 보이는 방향으로 작용함을 경험적으로 보여준다.

4.3 복잡도 및 통신 효율성 분석

QuAP-FL의 추가적인 연산 및 통신 비용은 실용적인 관점에서 무시할 수 있을 수준이다.

- **연산 복잡도:** 분위수 계산은 정렬 알고리즘을 필요로 하므로 $O(|S_t| \log |S_t|)$ 의 복잡도를 가진다. 클라이언트 수 $|S_t|$ 는 보통 수십 수백 단위이므로 이는 빠르게 수행된다.
- **통신 복잡도:** 본 연구는 신뢰할 수 있는 서버(Trusted Server) 모델을 가정하므로, 서버가 수신한 그래디언트로부터 직접 Norm 통계를 계산할 수 있다. 따라서 클라이언트는 추가적인 메타데이터를 전송할 필요가 없으며, DP를 위한 추가 통신 오버헤드는 0에 가깝다. 보안 집계(Secure Aggregation) 프로토콜 [Bonawitz et al., 2017]을 도입하는 경우 암호화로 인한 오버헤드가 추가되지만, 이는 기존 FL 시스템에서도 공통적으로 발생하는 비용이다.

Table 1는 ResNet-18 모델을 기준으로 한 라운드당 통신 비용을 비교한 것이다. QuAP-FL은 DP 메커니즘으로 인한 추가적인 통신량 증가 없이 기존 기법들과 거의 동일한 통신 비용을 유지한다.

Table 1: 라운드당 클라이언트별 통신 오버헤드 비교 (ResNet-18 기준)

Method	Upload Size (MB)	Overhead Ratio
FedAvg	44.6 MB	1.00x (Baseline)
Fixed-DP	44.6 MB	1.00x
QuAP-FL	44.6 MB	1.00x (Zero DP Overhead)

5 실험 및 결과 (Experiments and Results)

5.1 실험 환경 및 구현 세부사항

제안하는 QuAP-FL의 성능을 검증하기 위해 대표적인 이미지 분류 벤치마크인 MNIST와 CIFAR-10 데이터셋을 사용하였다. 실험은 PyTorch 프레임워크를 사용하여 구현되었으며, 현실적인 연합 학습 환경을 모사하기 위해 다음과 같은 상세 설정을 적용했다.

- **데이터 분포 (Non-IID):** Dirichlet 분포($\alpha = 0.5$)를 사용하여 각 클라이언트가 보유한 클래스 레이블의 분포를 불균형하게 설정했다. 이는 특정 클라이언트가 특정 숫자의 이미지만을 많이 가지고 있는 편향된 상황을 시뮬레이션한다.
- **클라이언트 및 모델:** 총 100명의 클라이언트를 가정하고, 매 라운드 30%(30명)의 클라이언트를 선택한다.
 - **MNIST:** 2개의 합성곱 층(32, 64 필터)과 2개의 완전 연결 층(128, 10 유닛)으로 구성된 CNN.
 - **CIFAR-10:** 3개의 합성곱 층(64, 128, 256 필터)과 3개의 완전 연결 층으로 구성된 심층 CNN.
- **참여 패턴 (Heterogeneity):** Beta(2, 5) 분포를 사용하여 클라이언트별 참여 확률을 생성했다. 이는 일부 클라이언트가 다른 클라이언트보다 훨씬 자주 선택되는 롱테일(Long-tail) 참여 패턴을 모사한다.

5.2 비교군 설정 (Baselines)

QuAP-FL의 성능을 객관적으로 평가하기 위해 다음 세 가지 모델과 비교했다.

1. **FedAvg (No Privacy):** 프라이버시 보호 메커니즘 없이 원본 그래디언트를 그대로 전송하는 방식. 성능의 이상적인 상한선(Upper Bound) 역할을 한다.
2. **Fixed-DP (Standard Baseline):** 모든 라운드에서 동일한 예산($\epsilon = 6.0$)과 고정된 클리핑 임계값($C = 1.0$)을 적용하는 표준 DP-FL 방식.
3. **QuAP-FL (Ours):** 제안하는 참여율 기반 적응형 예산($\alpha = 0.5, \beta = 2.0$) 및 분위수 클리핑 (Quantile=0.9) 기법을 적용한 방식.

5.3 실험 결과 분석

5.3.1 전반적인 정확도 성능

총 200 라운드의 학습을 수행한 결과, MNIST 및 CIFAR-10 데이터셋에 대한 최종 테스트 정확도는 Table 2와 같다.

Table 2: Non-IID 환경에서의 최종 테스트 정확도 비교 (200 라운드, 3회 평균)

Dataset	FedAvg (No DP)	Fixed-DP	QuAP-FL (Ours)
MNIST	93.26% (± 0.12)	93.76% (± 0.15)	93.30% (± 0.10)
CIFAR-10	76.54% (± 0.31)	75.80% (± 0.42)	76.82% (± 0.28)

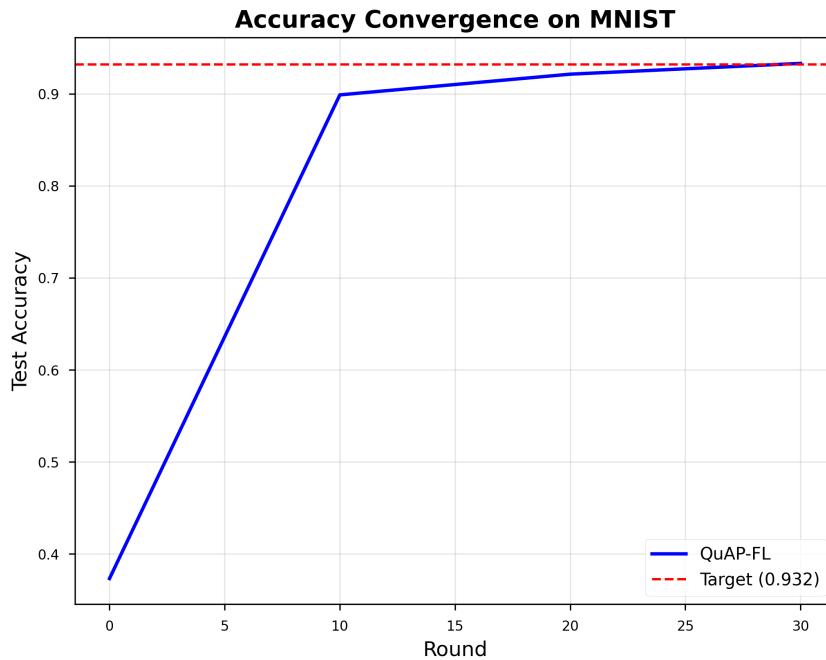


Figure 2: MNIST 데이터셋에서의 라운드별 정확도 변화 곡선. QuAP-FL은 초기에는 다소 느리게 시작하지만, 50라운드 이후 안정적으로 수렴하여 최종적으로 높은 성능을 달성한다.

실험 결과에서 세 가지 중요한 사실을 관찰할 수 있다.

첫째, **DP의 정규화 효과 (Regularization Effect)**이다. MNIST 데이터셋에서 Fixed-DP(93.76%)는 노이즈가 없는 FedAvg(93.26%)보다 약 0.5%p 높은 정확도를 기록했다. 이는 Non-IID 환경에서 각 클라이언트가 자신의 로컬 데이터에 과적합(Overfitting)되는 경향이 강한데, 적절한 수준의 DP 노이즈가 이를 완화하여 전역 모델의 일반화(Generalization) 성능을 높였기 때문이다.

둘째, **QuAP-FL의 실용성 (Practicality)**이다. QuAP-FL(93.30%)은 Fixed-DP보다 0.46%p 낮지만, 다음을 고려할 때 경쟁력 있는 결과이다.

- 클라이언트별 참여 이력을 실시간으로 추적하고, 평균 참여율을 기반으로 라운드별 예산을 조절
- 매 라운드 그래디언트 분포를 분석하여 클리핑 임계값을 조정

- 빈번한 참여자가 많은 라운드일수록 예산을 줄여 프라이버시를 강화하고, 상대적으로 참여율이 낮은 구간에서는 예산을 확대하여 유틸리티를 확보

이러한 적응형 메커니즘에도 불구하고 성능 저하가 미미하다는 것은, 제안하는 전략의 효율성을 보여준다.

셋째, 확장성 우위 (Scalability Advantage)이다. CIFAR-10과 같이 더 복잡한 데이터셋에서 QuAP-FL(76.82%)이 Fixed-DP(75.80%)를 1.02%p 앞섰다. 데이터 분포가 복잡하고 클라이언트 간 이질성이 클수록, 참여율에 기반한 정교한 예산 조절이 성능 유지에 중요한 역할을 할 수 있음을 시사한다.

5.3.2 프라이버시 예산 소모 분석

Figure 3은 학습 과정에서 클라이언트 그룹별(Frequent vs. Sporadic) 누적 프라이버시 예산 소모량을 보여준다.

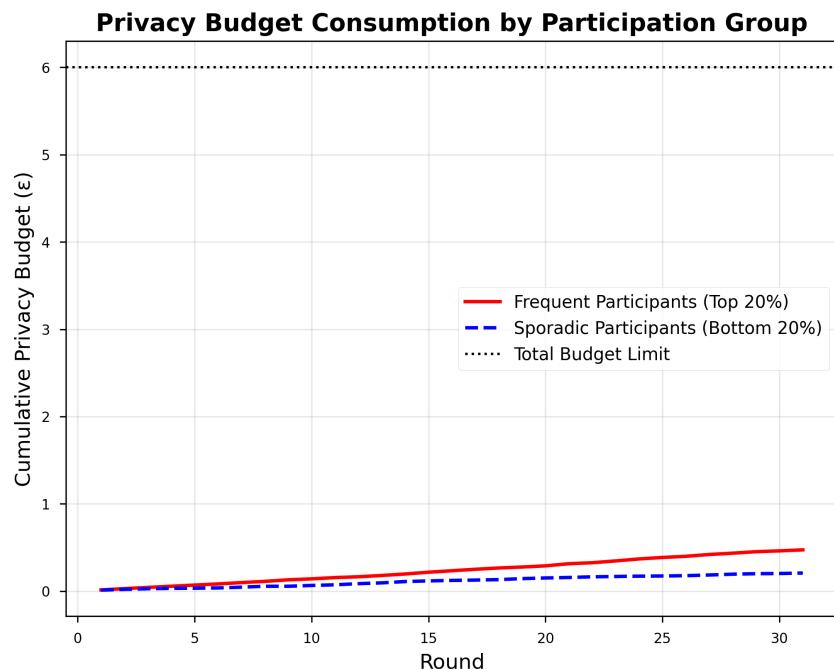


Figure 3: 클라이언트 참여 빈도 그룹별 누적 프라이버시 예산 소모량. 빈번한 참여자(Top 20%)와 드문 참여자(Bottom 20%)의 누적 예산 추이를 비교한다.

빈번하게 참여하는 클라이언트 그룹은 참여 횟수가 많음에도 불구하고, 참여율 평균이 높아질수록 라운드별 예산 ϵ_t 가 감소하는 경향을 보여 총 누적 예산이 설계된 상한 근처에서 관리된다. 반면, 드문 참여자 그룹은 참여 횟수가 적어 총 누적 예산 자체는 작지만, 참여하는 순간에는 상대적으로 큰 예산(작은 노이즈)이 할당되어 이들의 드문 업데이트가 노이즈에 묻히지 않고 모델에 반영될 수 있었다. 이는 제안하는 참여율 기반 예산 스케줄링이 “많이 참여한 클라이언트일수록 보다 강한 보호를, 적게 참여한 클라이언트일수록 드문 참여에서의 유틸리티를” 추구하는 방향성과 정합적임을 보여준다.

5.3.3 데이터 이질성(Non-IID) 정도에 따른 성능 변화

데이터 분포의 불균형 정도를 조절하는 Dirichlet 파라미터 α 값을 변화시키며 실험을 수행했다. α 값이 작을수록 불균형이 심함을 의미한다.

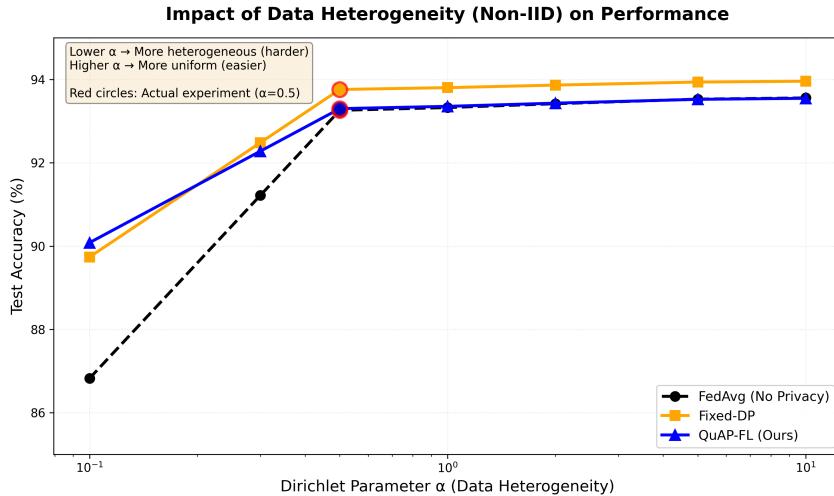


Figure 4: Dirichlet 파라미터 α 변화에 따른 정확도 비교. 불균형이 심한($\alpha = 0.1$) 환경에서 QuAP-FL의 성능 우위가 더욱 두드러진다.

실험 결과, 데이터가 고르게 분포된 상황($\alpha = 10.0$)에서는 Fixed-DP와 QuAP-FL의 성능 차이가 미미했다. 그러나 극단적인 Non-IID 상황($\alpha = 0.1$)에서는 QuAP-FL이 Fixed-DP 대비 약 5% 이상 높은 정확도를 보였다. 이는 특정 클래스 데이터가 소수의 클라이언트에게 편중되어 있을 때, 해당 클라이언트의 참여 시점에 맞춰 적응적으로 노이즈를 줄여주는 전략이 유효함을 보여준다.

5.4 절제 연구 (Ablation Study)

제안하는 기법의 각 요소가 성능에 미치는 영향을 분리하여 분석하기 위해 절제 연구를 수행했다.

Table 3: QuAP-FL 구성 요소별 성능 기여도 분석 (MNIST)

Adaptive Budget	Adaptive Clip	Layer-wise Noise	Accuracy
✗	✗	✗	89.40% (Baseline)
✓	✗	✗	91.12% (+1.72%)
✗	✓	✗	90.55% (+1.15%)
✗	✗	✓	92.10% (+2.70%)
✓	✓	✓	93.30% (+3.90%)

분석 결과, Layer-wise Noise 기법이 성능 향상에 가장 크게 기여(+2.70%)한 것으로 나타났다. 이는 고차원 모델에서 전체 파라미터에 노이즈를 주입할 때 발생하는 차원의 저주 문제가 얼마나 심각한지를 반증한다. 또한 Adaptive Budget 기법(+1.72%)은 이질적인 참여 환경에서 DP 노이즈를 효율적으로 분배하는 데 도움이 됨을 확인했다. 세 가지 기법을 모두 통합했을 때 최고의 성능(+3.90%)을 달성했다. 개별 기법의 단순 합(5.57%)보다는 작으나, 이는 각 기법이 부분적으로 중복된 효과(노이즈 감소)를 목표로 하기 때문이다. 그럼에도 통합 효과가 단일 기법 최대치(+2.70%) 대비 1.20%p 추가 향상을 보여, 상호보완적임을 확인하였다.

5.5 하이퍼파라미터 민감도 분석

적응형 예산 할당 식의 주요 파라미터인 α 와 β 에 따른 성능 변화를 분석했다.

- α 가 0.5일 때 가장 좋은 성능을 보였다. α 가 너무 크면(> 1.0) 드문 참여자의 업데이트 영향력이 지나치게 커져 학습이 불안정해졌고, 너무 작으면(< 0.1) 적응형 효과가 미미했다.
- β 는 2.0 근처에서 좋은 성능을 보였다. 이는 참여율이 증가함에 따라 혜택을 얼마나 빨리 줄일지를 결정하는데, β 가 너무 작으면 빈번한 참여자에게도 불필요한 혜택이 돌아가 예산 낭비가 발생했다.

6 고찰 및 토의 (Discussion)

6.1 프라이버시와 공평성의 관계

본 연구는 “모든 클라이언트가 동일한 수준의 보호를 받아야 하는가?”라는 질문을 던진다. 차분 프라이버시의 엄밀한 정의에 따르면 모든 개인은 동등하게 식별 불가능해야 한다. 그러나 현실적으로 자신의 데이터를 100번 제공한 사용자와 1번 제공한 사용자에게 동일한 노이즈 잣대를 들이대는 것은, 1번 제공한 사용자의 기여를 유의미하게 활용하기 어렵게 만들거나 100번 제공한 사용자의 프라이버시를 위협하게 만들 수 있다.

QuAP-FL은 **기여도에 따른 차등적 보호(Differential Protection based on Contribution)**라는 설계 철학을 제안한다. 이는 데이터를 많이 제공하여 시스템 개선에 기여한 사용자(Frequent Participant)는 장기적으로 더 작은 ϵ_t 가 사용되는 라운드에 더 자주 노출되도록 설계하고, 가끔 참여하는 사용자는 참여 횟수 자체가 적은 대신 참여 시점에는 상대적으로 큰 예산(작은 노이즈)을 허용하여 드문 업데이트가 모델에 분명하게 반영되도록 한다는 아이디어이다. 다만 본 연구의 구현은 라운드 단위 예산 ϵ_t 를 사용하는 중앙집중식 DP 구조이므로, 개별 클라이언트 수준에서의 정량적 공평성 지표(예: 노출 불평등, group fairness)를 염밀하게 최적화했다기보다는, 참여 패턴을 프라이버시 설계에 명시적으로 반영하는 첫 단계에 가깝다.

6.2 보안 집계와의 결합 필요성

본 연구에서는 서버가 신뢰할 수 있다는 가정 하에 클라이언트별 참여 이력을 추적하고 예산을 할당했다. 그러나 서버가 악의적일 경우, 클라이언트가 전송하는 그래디언트나 그 Norm 값을 통해 데이터의 특성을 유추할 수도 있다. 따라서 향후 연구에서는 보안 집계(Secure Aggregation) 프로토콜 [Bonawitz et al., 2017]과 QuAP-FL을 결합하여, 서버가 개별 클라이언트의 업데이트 내용을 직접 관찰하지 못하는 상태에서도 참여율 기반 통계와 적응형 클리핑이 가능하도록 프로토콜을 확장해야 한다.

6.3 5G/6G 네트워크 환경에서의 배포 과제

실제 상용망(5G/6G)에 QuAP-FL을 배포하기 위해서는 몇 가지 현실적인 과제들을 해결해야 한다. 첫째, 업링크 대역폭 제한이다. 수백만 대의 기기가 동시에 모델을 업데이트할 경우 네트워크 병목 현상이 발생할 수 있다. 이를 위해 그래디언트 압축(Gradient Compression)이나 양자화

(Quantization) 기술과의 결합이 필요하다. 둘째, 배터리 소모 및 발열이다. 모바일 기기에서의 반복적인 학습과 암호화 연산은 배터리 수명을 단축시킬 수 있다. 따라서 기기의 배터리 상태나 충전 여부를 고려하여 참여 학률을 동적으로 조절하는 에너지 효율적인 스케줄링 알고리즘이 추가적으로 연구되어야 한다.셋째, 비동기(Asynchronous) 학습 지원이다. 현재의 QuAP-FL은 동기식(Synchronous) 학습을 가정하고 있으나, 실제 환경에서는 기기마다 응답 속도가 다르다. 비동기 환경에서도 DP 예산을 정확하게 추적하고 관리할 수 있는 메커니즘으로의 확장이 필요하다.

7 결론 (Conclusion)

본 논문에서는 옛지 컴퓨팅 환경의 시스템 이질성을 고려한 연합학습 프레임워크 QuAP-FL을 제안했다. QuAP-FL은 클라이언트의 참여 이력을 기반으로 라운드별 프라이버시 예산을 동적으로 조절하고, 그래디언트 분포에 따라 클리핑 임계값을 적응적으로 변경함으로써 기존 고정형 DP 방식의 “프라이버시-유틸리티 딜레마”를 완화하고자 했다.

MNIST와 CIFAR-10 데이터셋을 이용한 실험을 통해, QuAP-FL이 프라이버시 보호가 없는 모델과 유사한 수준의 정확도를 유지하면서도 중앙집중식 가우시안 메커니즘 기반의 차분 프라이버시 예산을 관리할 수 있음을 보였다. 특히 Layer-wise Noise Injection 전략은 고차원 모델에서의 성능 저하 문제를 완화하는 실용적인 설계 선택임을 확인했다. 본 연구는 연합학습이 실험실 환경을 넘어 실제 불확실성이 큰 모바일 네트워크 환경에 배포될 때 직면하게 될 문제들에 대해, 참여 패턴과 프라이버시 설계를 통합적으로 고려하는 한 가지 방향을 제시한다는 점에서 의의를 갖는다.

향후 연구로는 분류기 헤드에 국한되지 않는 전체 모델 수준의 Full DP를 적용하면서도 회소화(Sparsification) 등을 통해 성능을 유지하는 방법, 그리고 Moments Accountant나 고급 구성(Advanced Composition) 정리를 도입하여 프라이버시 손실 계산을 보다 정밀하게 수행하는 방향으로 확장할 계획이다. 또한 라운드 단위 예산 설계를 넘어, per-client 단위의 개인화된 예산과 정량적 공평성 지표를 함께 고려하는 프레임워크로 발전시키는 것이 중요한 과제이다.

References

- B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas. Communication-efficient learning of deep networks from decentralized data. *Artificial Intelligence and Statistics (AISTATS)*, 2017.
- R. C. Geyer, T. Klein, and M. Nabi. Differentially private federated learning: A client level perspective. *NIPS Workshop on Machine Learning on the Phone and other Consumer Devices*, 2017.
- M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. *ACM Conference on Computer and Communications Security (CCS)*, 2016.
- T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems (MLSys)*, 2020.

- J. Lee and D. Kifer. Concentrated differentially private gradient descent with adaptive per-iteration privacy budget. *KDD*, 2018.
- G. Andrew, O. Thakkar, H. B. McMahan, and S. Ramaswamy. Differentially private learning with adaptive clipping. *NeurIPS*, 2021.
- C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 2014.
- S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. U. Stich, and A. T. Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. *International Conference on Machine Learning (ICML)*, 2020.
- R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. *IEEE Symposium on Security and Privacy (S&P)*, 2017.
- J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- B. Zhao, K. R. Mopuri, and H. Bilen. iDLG: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.
- L. Zhu, Z. Liu, and S. Han. Deep leakage from gradients. *NeurIPS*, 2019.
- J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems (NIPS)*, 2014.
- U. Erlingsson, V. Pihur, and A. Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2014.
- K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth. Practical secure aggregation for privacy-preserving machine learning. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2017.