Spectral and Temporal Denoising for Differentially Private Optimization

Hyeju Shin[1,5,†], Kyudan Jung[2,5,†], Seongwon Yun[3,5,†], Juyoung Yun[4,5,†∗]

[1]Electronics and Telecommunications Research Institute (ETRI), Republic of Korea
[2]Chung-ang University, Department of Mathematics, Republic of Korea
[3]Hanwha Life Insurance, Republic of Korea
[4]Stony Brook University, Department of Computer Science, United States
[5]OpenNN Lab, MODULABS, Republic of Korea

[†]Equal contribution

Abstract

This paper introduces the FFT-Enhanced Kalman Filter (FFTKF), a differentially private optimization method that addresses the challenge of preserving performance in DP-SGD, where added noise typically degrades model utility. FFTKF integrates frequency-domain noise shaping with Kalman filtering to enhance gradient quality while preserving (ε, δ)-DP guarantees. It employs a high-frequency shaping mask in the Fourier domain to concentrate differential privacy noise in less informative spectral components, preserving low-frequency gradient signals. A scalar-gain Kalman filter with finite-difference Hessian approximation further refines the denoised gradients. With a per-iteration complexity of $\mathcal{O}(d \log d)$, FFTKF demonstrates improved test accuracy over DP-SGD and DiSK across MNIST, CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets using CNNs, Wide ResNets, and Vision Transformers. Theoretical analysis confirms that FFTKF maintains equivalent privacy guarantees while achieving a tighter privacy-utility trade-off through reduced noise and controlled bias.

1 Introduction

Differential Privacy (DP) has become a foundational concept for safeguarding individual-level information in machine learning and data analysis, providing rigorous guarantees against information leakage from model outputs [1, 18, 25, 30]. Standard DP mechanisms, such as the Laplace and Gaussian mechanisms, achieve privacy by injecting calibrated noise into data or gradients. However, this noise injection often results in substantial degradation in model utility, particularly in high-dimensional or deep learning scenarios.

One of the central challenges in differentially private learning is improving the performance of DP-SGD, a widely used algorithm that combines stochastic gradient descent with Gaussian noise perturbation [1]. Due to the stochastic and high-magnitude nature of DP noise, especially under tight privacy budgets, the privatized gradients tend to exhibit poor signal-to-noise ratios. Consequently, effectively denoising DP-SGD updates to enhance performance while strictly preserving DP guarantees remains a formidable and open problem. Recent studies highlight that the noise introduced in DP-SGD can significantly impair convergence and model accuracy, particularly in complex tasks, making denoising a critical yet challenging objective [23, 26]. These works underscore the difficulty

---

[∗]Currently Under Review. Corresponding author: `daniel@open-nn.com`

of balancing noise reduction with the preservation of privacy constraints, as excessive denoising risks compromising the (ϵ, δ)-DP guarantees.

To address this, recent methods have introduced signal processing and state estimation tools to enhance gradient quality under DP. For example, the DiSK framework [32] incorporates Kalman filtering into DP optimization to recursively estimate cleaner gradients, resulting in improved convergence and test accuracy. Kalman filters are well-established tools for estimating latent states in dynamic systems based on noisy observations [19, 20]. Their application in DP optimization allows for smoother updates by exploiting temporal correlations across gradient steps, which is particularly effective in large-scale and deep models.

In parallel, frequency-domain techniques such as low-pass filtering have shown promise in separating useful signal components from high-frequency noise [9, 17, 22]. Leveraging the Fast Fourier Transform (FFT), recent signal processing approaches selectively attenuate noise by filtering out high-frequency components in a computationally efficient manner [6]. When adapted to machine learning, these spectral methods provide a structured way to shape and redistribute noise, preserving signal fidelity in lower-frequency domains [2, 26].

Building on these developments, we propose the *FFT-Enhanced Kalman Filter* (FFTKF), a novel optimization framework that combines spectral noise shaping via FFT with Kalman filtering for improved gradient estimation under differential privacy. Our method reshapes the injected DP noise to concentrate its energy in high-frequency spectral components, thereby preserving low-frequency gradient information that is critical for effective model training. The resulting privatized gradients are then denoised using a scalar-gain Kalman filter that leverages temporal structure to produce stable update directions.

Our contributions are as follows. First, we introduce a frequency-domain noise shaping strategy that integrates into the DP optimization pipeline while preserving the required (ϵ, δ)-DP guarantees. Second, we combine this with a simplified Kalman filter that enables efficient gradient tracking with a per-step complexity of $O(d \log d)$. Third, we empirically validate FFTKF on standard benchmarks— MNIST, CIFAR-10, CIFAR-100, and Tiny-ImageNet—across multiple model architectures including CNNs, Wide ResNets, and Vision Transformers. Our results show that FFTKF consistently improves test accuracy over DP-SGD and DiSK, particularly under tight privacy budgets. These findings demonstrate the effectiveness of combining spectral filtering and state estimation for denoising gradients in private optimization.

2 Related Works

Gradient-Based Optimization. Stochastic Gradient Descent (SGD) and its variants, such as Adam, form the backbone of optimization in machine learning by iteratively updating model parameters using gradient information [12, 24]. SGD relies on noisy gradient estimates computed from random mini-batches, offering computational efficiency but suffering from variance in updates. Adam enhances SGD by incorporating adaptive momentum and second-moment estimates, which accelerate convergence, particularly in non-convex settings. However, these methods lack inherent privacy protections, as gradients can leak sensitive information about training data. This limitation has spurred the development of privacy-preserving optimizers like DP-SGD, which integrate noise to safeguard data privacy [1].

Differential Privacy Optimization. Differential Privacy (DP) is a rigorous framework for protecting individual data privacy in machine learning by injecting calibrated noise into data or gradients, often at the cost of reduced model utility [1, 18, 25, 30]. DP-SGD, a widely adopted algorithm, achieves privacy by adding Gaussian noise to gradients during stochastic gradient descent, but the high-magnitude noise often degrades signal-to-noise ratios, impairing convergence [1]. Recent efforts address this trade-off through innovative techniques. For instance, Fan and Xiong proposed an adaptive noise adjustment method for real-time aggregate monitoring, dynamically balancing privacy and utility based on data characteristics [10]. Similarly, Zhang et al.'s DiSK framework integrates Kalman filtering to denoise privatized gradients, enhancing model performance while adhering to privacy constraints [32]. Adaptive clipping techniques further improve DP-SGD by dynamically tuning noise levels to preserve gradient information [28].
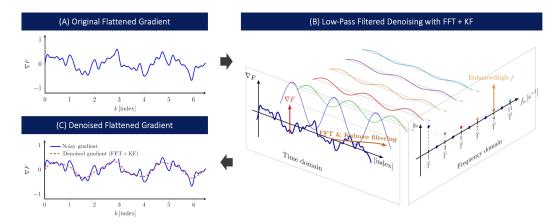
Figure 1: Visualization of the proposed frequency-domain gradient denoising process. (A) The original flattened gradient before privatization. (B) The FFT transforms the gradient into the frequency domain, where shaped Gaussian noise is added and attenuated by a high-frequency mask. Kalman filtering is then applied in the time domain to exploit temporal correlations. (C) The final denoised gradient (dashed line) shows reduced high-frequency perturbations while preserving the underlying signal structure.

Kalman Filter Optimization. Kalman filters are powerful tools for state estimation in noisy dynamic systems, leveraging temporal correlations to refine estimates from noisy observations [19, 20]. In the context of differential privacy, the DiSK framework employs a simplified Kalman filter to smooth noisy gradient estimates, improving convergence rates and reducing computational overhead in DP optimization [32]. This approach is particularly effective for large-scale models, where noise can significantly disrupt training. Recent research has also explored Kalman filters in federated learning, where they help balance privacy and accuracy in distributed environments by refining gradient updates across clients. These advancements highlight the versatility of Kalman-based methods in privacy-preserving machine learning.

Low-Pass Filtering Techniques. Low-pass filters are essential in signal processing for attenuating high-frequency noise while preserving critical signal components, making them valuable for data analysis [6]. Fourier transform-based low-pass filters, known for their computational efficiency, are widely used to separate noise from meaningful signals [3]. In differential privacy, adaptive low-pass filtering techniques optimize cutoff frequencies to maximize utility under strict privacy budgets, effectively denoising gradients in high-dimensional tasks [2, 5, 26]. These methods are particularly relevant for deep learning, where preserving low-frequency gradient information is crucial for maintaining model performance while ensuring robust privacy guarantees.

3 Methodology

Preliminaries. Let $\mathbb{R}^d$ denote the d-dimensional Euclidean space. A vector $z \in \mathbb{R}^d$ is written as $z = (z_0, \ldots, z_{d-1})^\top$, with ℓ_2-norm $\|z\|_2 = \sqrt{\sum_{i=0}^{d-1} z_i^2}$. The identity matrix in $\mathbb{R}^{d \times d}$ is I_d, and a diagonal matrix with entries $\varphi_0, \ldots, \varphi_{d-1}$ is $\mathrm{diag}(\varphi_0, \ldots, \varphi_{d-1})$. For a function $f : \mathbb{R}^d \to \mathbb{R}$, its gradient is ∇f, and Hessian is $\nabla^2 f$. The Hadamard product is $\odot$, and $\lfloor x \rfloor$ denotes the floor function. The Gaussian distribution with mean zero and covariance $\sigma^2 I_d$ is $\mathcal{N}(0, \sigma^2 I_d)$. We aim to minimize $F(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[f(x; \xi)]$, the expected loss over a data distribution $\mathcal{D}$. Parameters at iteration t are $x_t \in \mathbb{R}^d$, updated as $x_{t+1} = x_t - \eta \tilde{g}_t$, where $\eta > 0$ is the learning rate and $\tilde{g}_t$ is the estimated gradient. The true gradient $\nabla F(x_t)$ is approximated via a mini-batch $\mathcal{B}_t \subseteq \mathcal{D}$ of size B, giving $g_t = \frac{1}{B} \sum_{\xi \in \mathcal{B}_t} \nabla f(x_t; \xi)$. The parameter difference is $d_t = x_{t+1} - x_t$. In differential privacy, a mechanism satisfies (ε, δ)-DP if, for neighboring datasets D, D' differing in one entry, and any output set S, $\Pr[\mathcal{M}(D) \in S] \leq e^\varepsilon \Pr[\mathcal{M}(D') \in S] + \delta$. Gradients are clipped

as $\mathrm{clip}(v, C) = v \cdot \min(1, C/\|v\|_2)$ to bound their ℓ_2-norm by C, and noise $w_t \sim \mathcal{N}(0, \sigma_w^2 I_d)$ is added.

3.1 Fast Fourier Transform

This section briefly reviews the discrete Fast Fourier transform (FFT) and the algorithmic considerations that motivate its use for gradient denoising.

Recall that the *discrete Fourier transform* (DFT) of a real-valued vector $z = (z_0, \ldots, z_{d-1})^\top \in \mathbb{R}^d$ is the complex vector $\hat{z} = \mathcal{F}(z) \in \mathbb{C}^d$, with components

$$\hat{z}_k = \sum_{n=0}^{d-1} z_n \, e^{-2\pi i k n / d}, \quad \text{for } k = 0, \ldots, d-1,$$

and its inverse is given by

$$z_n = \frac{1}{d} \sum_{k=0}^{d-1} \hat{z}_k \, e^{2\pi i k n / d}, \quad \text{for } n = 0, \ldots, d-1.$$

With this normalization, the Fourier transform $\mathcal{F}$ is unitary, i.e.,

$$\mathcal{F}^{-1}(\mathcal{F}(z)) = z,$$

and Parseval's identity holds:

$$\|z\|_2^2 = \frac{1}{d} \|\hat{z}\|_2^2,$$

where $\hat{z} = \mathcal{F}(z)$.

Consequently, adding Gaussian noise in the Fourier domain preserves the ℓ_2-sensitivity required for (ε, δ)-DP due to the invariance of the ℓ_2-norm under unitary transformations.

Low/high–frequency split. Fix a pivot index $k_0 = \lfloor \lambda d \rfloor$ for some $\lambda \in (0, 1)$. Frequencies $k < k_0$ are called *low-frequency components*, and $k \geq k_0$ *high-frequency components*. This separation reflects the empirical observation that most signal information, especially in gradient vectors of smooth loss landscapes, is concentrated in the lower spectral range, while the high-frequency components often contain stochastic noise.

Spectral filtering. A diagonal mask $\Phi = \mathrm{diag}(\varphi_0, \ldots, \varphi_{d-1})$ defines a linear filter

$$\mathcal{G}_\Phi(z) = \mathcal{F}^{-1}\left(\Phi \, \hat{z}\right) = \frac{1}{d} \sum_{k=0}^{d-1} \varphi_k \hat{z}_k \, e^{2\pi i k n / d},$$

where $\hat{z} = \mathcal{F}(z)$.

By the convolution theorem, this operation in the frequency domain is equivalent to convolution in the time domain and can be evaluated in $O(d \log d)$ time via the FFT algorithm, which significantly improves efficiency compared to the naive $O(d^2)$ convolution [4, 11, 29].

High-frequency shaping mask. To enhance denoising while maintaining DP, we use a smooth mask function

$$\varphi_k = \begin{cases} 1, & k < k_0, \\ 1 - \rho & k \geq k_0, \end{cases}$$

where $\rho \in (0, 1)$ controls the magnitude of suppression. This *step-wise attenuation* suppresses higher-frequency components beyond a cutoff index k_0, thereby reducing the influence of DP noise concentrated in those frequencies. Unlike sharp cutoffs, this mask gently dampens high-frequency content while preserving the low-frequency structure of gradients, offering a balance between denoising and signal fidelity [30, 33].

Remarks. Let $\Phi_\rho = \mathrm{diag}(\varphi_0, \ldots, \varphi_{d-1})$, then the filtered version of a privatized gradient $g = \nabla f + w$ is

$$\hat{g} = \mathcal{G}_{\Phi_\rho}(g) = \mathcal{F}^{-1}(\Phi_\rho \mathcal{F}(\})).$$

4

When $w \sim \mathcal{N}(0, \sigma^2 I)$, the transformed noise $\hat{w} := \mathcal{F}^{-1}(\Phi_\rho \mathcal{F}(\sqsupseteq))$ is still zero-mean but now has reduced energy in the low-frequency components:

$$\mathbb{E}[\|\hat{w}_{<k_0}\|_2^2] \ll \mathbb{E}[\|\hat{w}\|_2^2],$$

facilitating more accurate recovery of the gradient signal after filtering.

This FFT recap underpins our *FFT-Enhanced Kalman Filter* in Sec. 3.4, where we combine spectral noise shaping with a scalar-gain Kalman predictor to denoise privatized gradients efficiently, achieving both computational and privacy-preserving benefits.

3.2 Gradient Dynamics with High-Frequency Differential Privacy

To explain our proposed idea of using the FFT-Enhanced Kalman Filter for denoising gradients, we first establish a dynamic system for the gradients. This system consists of a *system update* equation and an *observation* equation. The system update of the gradient dynamics is derived via Taylor expansion of ∇F around x_{t-1}, allowing for a second-order approximation of the gradient evolution at step t:

$$\nabla F(x_t) = \nabla F(x_{t-1}) + \boldsymbol{H}_t \cdot (x_t - x_{t-1}) + v_t, \tag{1}$$

where $\boldsymbol{H}_t := \nabla^2 F(x_{t-1}) \in \mathbb{R}^{d \times d}$ is the local Hessian, and v_t represents the Taylor expansion remainder:

$$v_t = \frac{1}{2} \int_0^1 \nabla^3 F\big((1-z)x_{t-1} + zx_t\big)[x_t - x_{t-1}]^{\otimes 2} \, dz.$$

Here, $(\cdot)^{\otimes 2}$ denotes the second-order tensor product, capturing the nonlinearity of the third derivative in the remainder [21, 32].

The observed gradient g_t is a noisy, privatized estimate of the true gradient:

$$g_t = \frac{1}{B} \sum_{\xi \in \mathcal{B}_t} \text{clip}(\nabla f(x_t, \xi), C) + w_t = C_t \nabla F(x_t) + w_t', \tag{2}$$

where w_t' contains both DP noise and subsampling noise. The matrix C_t is the effective observation operator determined by the clipping factor and the sampling mechanism. In the ideal full-batch, unclipped setting, $C_t = I_d$, but more generally it is a contraction mapping satisfying $\|C_t\|_2 \leq 1$.

Combining the update and observation equations, we obtain the system:

$$\nabla F(x_t) = \nabla F(x_{t-1}) + \boldsymbol{H}_t(x_t - x_{t-1}) + v_t, \qquad \text{(System update)}$$

$$g_t = C_t \nabla F(x_t) + w_t'. \qquad \text{(Observation)}$$

To enforce differential privacy while retaining useful structure, we inject noise in the *frequency domain*:

$$g_t = C_t \nabla F(x_t) + \tilde{w}_t, \qquad \tilde{w}_t := \mathcal{F}^{-1}\big(\Phi_\rho \odot \mathcal{F}(w_t)\big), \tag{3}$$

where $w_t \sim \mathcal{N}(0, \sigma_w^2 I_d)$ is isotropic Gaussian noise. The mask $\Phi_\rho \in \mathbb{R}^d$ is diagonal and satisfies

$$(\Phi_\rho)_k = \begin{cases} 1, & 0 \leq k < k_0, \\ 1 - \rho e^{-\alpha(k-k_0)}, & k_0 \leq k < d, \end{cases}$$

for some pivot frequency $k_0 = \lfloor \lambda d \rfloor$, with shaping strength $\rho \in (0, 1)$ and decay rate $\alpha > 0$. This ensures that the privacy-preserving noise $\tilde{w}_t$ is spectrally shaped to occupy primarily high-frequency components, which contribute less to gradient descent. This leads to improved recoverability of the informative low-frequency gradient content.

3.3 Frequency-Domain Denoising

To recover the low-frequency content of the privatized gradient, we apply the inverse of the noise shaping operation:

$$\mathcal{G}_\rho(z) := \mathcal{F}^{-1}\big(\Phi_\rho \odot \mathcal{F}(z)\big), \qquad z \in \mathbb{R}^d.$$

This filtering step yields the estimate $\hat{g}_t = \mathcal{G}_\rho(g_t)$. Since $\mathcal{G}_\rho$ is a linear operator with spectral mask Φ_ρ, this operation has complexity $O(d \log d)$ and does not distort the signal beyond a known attenuation factor. That is, the covariance of $\hat{g}_t$ is a spectrally reweighted version of $\text{Cov}(g_t)$, which we exploit in the Kalman update below.

3.4 FFT-Enhanced Kalman Filter

We adopt the scalar-gain Kalman filtering approximation introduced in [32], which simplifies the covariance matrices to scalar multiples of the identity. Specifically, we let $P_t = p_t I_d$, $K_t = \kappa I_d$, and estimate the Hessian action using a finite-difference formula with hyperparameter $\gamma > 0$.

Prediction Step. Given $\tilde{g}_{t-1}$, we predict the next gradient by using a first-order approximation based on finite differences:

$$\tilde{g}_{t|t-1} = \tilde{g}_{t-1} + \frac{1}{B} \sum_{\xi \in \mathcal{B}_t} \frac{\nabla f(x_t + \gamma d_{t-1}; \xi) - \nabla f(x_t; \xi)}{\gamma}, \tag{P}$$

where $d_{t-1} := x_t - x_{t-1} = -\eta \tilde{g}_{t-1}$. This approximates the action of the local Hessian without explicitly computing second-order derivatives.

Correction Step. The predicted gradient is then corrected using the filtered observation $\widehat{g}_t$:

$$\tilde{g}_t = (1 - \kappa) \tilde{g}_{t-1} + \kappa \widehat{g}_t, \tag{C}$$

where $\kappa \in (0, 1)$ is the Kalman gain that balances the reliance on the prediction versus the new (denoised) observation. This form ensures that the update direction incorporates temporal consistency across iterations while attenuating the influence of high-frequency noise.

Together, Eqs. (P) and (C) constitute a computationally lightweight Kalman filtering mechanism enhanced by frequency-domain denoising. It achieves $O(d \log d)$ per-step complexity and enhances DP optimization without sacrificing privacy guarantees.

Algorithm 1 FFT-Enhanced Kalman Filter Optimizer (FFTKF)

Require: initial point x_0, base optimiser Opt, learning rate η, gain κ, FD parameter γ, high–frequency ratio ρ, clipping bound C, noise scale σ_w.

1: $\tilde{g}_{-1} \leftarrow 0$, $d_{-1} \leftarrow 0$
2: **for** $t = 0, 1, \ldots, T - 1$ **do**
3:     Sample mini-batch $\mathcal{B}_t$
4:     Compute shaped DP gradient

$$g_t \leftarrow \frac{1}{B} \sum_{\xi \in \mathcal{B}_t} \text{clip}\Big(\nabla f(x_t; \xi), C\Big) + \tilde{w}_t, \quad \tilde{w}_t \sim \mathcal{N}\big(0, \sigma_w^2 I_d\big)$$

5:     $\widehat{g}_t \leftarrow \mathcal{G}_\rho(g_t)$ ▷ FFT denoising
6:     $\tilde{g}_{t|t-1} \leftarrow$ Eq. (P)
7:     $\tilde{g}_t \leftarrow$ Eq. (C)
8:     $x_{t+1} \leftarrow \text{Opt}\big(x_t, \eta, \tilde{g}_t\big)$
9:     $d_t \leftarrow x_{t+1} - x_t$
10: **end for**

3.5 Additional Discussion

The high-frequency shaping in Eq. (3) intentionally pushes privacy noise into spectral regions that matter least for optimization. Because the Kalman filter relies on low-frequency temporal correlations captured by Eqs. (P)–(C), the FFT step removes most of the injected disturbance before the gain κ is applied, resulting in a provably lower steady-state covariance.

Let $\Sigma_w = \sigma_w^2 I_d$ be the covariance of the original DP noise w_t, then the shaped noise $\tilde{w}_t = \mathcal{F}^{-1}(\Phi_\rho \odot \mathcal{F}(w_t))$ has covariance

$$\Sigma_{\tilde{w}} = \mathcal{F}^{-1} \cdot \Phi_\rho^2 \cdot \mathcal{F} \cdot \Sigma_w \cdot \mathcal{F}^{-1} \cdot \Phi_\rho^2 \cdot \mathcal{F},$$

whose low-frequency principal components are suppressed relative to Σ_w. Hence, the Kalman filter receives observations with diminished low-frequency noise variance, resulting in lower mean-square estimation error.

Crucially, FFTKF inherits the $O(d)$ *memory* and $O(d)$ *algebraic* complexities of the simplified DiSK variant while adding only two in-place FFTs per iteration.

3.5.1 Connections to DISK

Scalar–gain Kalman simplification. Our FFT-Enhanced Kalman Filter (*FFTKF*) inherits the scalar–gain reduction of DISK [32], wherein both the state covariance P_t and the Kalman gain K_t are isotropic:

$$P_t = p_t I_d, \quad K_t = \kappa I_d.$$

This diagonal simplification ensures that all matrix-vector operations reduce to scalar multiples of vector additions, preserving an $\mathcal{O}(d)$ runtime and storage profile. The Hessian-vector product $\boldsymbol{H}_t d_{t-1}$ is approximated with a single finite-difference query:

$$\boldsymbol{H}_t d_{t-1} \approx \frac{\nabla F(x_t + \gamma d_{t-1}) - \nabla F(x_t)}{\gamma},$$

eliminating the need for Hessian storage or inversion.

FFT-based noise shaping. While DISK performs time-domain exponential smoothing, FFTKF additionally *reshapes* the injected DP noise to concentrate its energy in the high-frequency spectrum:

$$\tilde{w}_t := \mathcal{F}^{-1}\big(\Phi_\rho \odot \mathcal{F}(w_t)\big), \quad (\Phi_\rho)_k = \begin{cases} 1, & k < k_0, \\ 1 - \rho e^{-\alpha(k-k_0)}, & k \geq k_0, \end{cases}$$

with pivot index $k_0 = \lfloor \lambda d \rfloor$. The mask Φ_ρ acts as a soft high-pass filter for the noise, minimizing the effect of noise on low-frequency directions where the Kalman filter's predictive prior is most accurate. This filtering can be viewed as a dual to the temporal smoothing in DiSK, but operating in the spectral domain.

Computational footprint. Compared with DPSGD, FFTKF requires:

1. one additional forward pass per iteration to compute the finite-difference directional gradient, and

2. two in-place 1-D FFTs: a forward transform $\mathcal{F}$ and its inverse $\mathcal{F}^{-1}$.

Both operations scale as $O(d \log d)$, while the state vector $\tilde{g}_t$ and difference direction d_t are stored as $O(d)$ vectors. Thus, FFTKF matches the memory profile of DiSK [32] but enables more precise noise attenuation with marginal overhead.

Privacy guarantee. Since the FFT operation is orthonormal, it preserves the ℓ_2 norm:

$$\|\tilde{w}_t\|_2 = \|\Phi_\rho \odot \mathcal{F}(w_t)\|_2 \leq \|w_t\|_2.$$

Thus, FFT-based reshaping does not increase the sensitivity of the privatized quantity. The overall privacy budget (ε, δ) remains exactly that of DPSGD and DISK, guaranteed by

4 Theoretical Analysis: Privacy-Utility Trade-off

Let the FFT operator be $\mathcal{F} : \mathbb{R}^d \to \mathbb{C}^d$ with inverse $\mathcal{F}^{-1}$, as introduced in Section 3.1. Fix a pivot index $k_0 = \lfloor \lambda d \rfloor$ ($\lambda \in (0,1)$) and a high–frequency attenuation ratio $\rho \in (0,1)$. Define the diagonal spectral mask

$$\Phi_\rho = \mathrm{diag}\big(\underbrace{1, \ldots, 1}_{k_0}, \underbrace{1 - \rho, \ldots, 1 - \rho}_{d-k_0}\big),$$

and the deterministic post-processing map $P(g) = \mathcal{F}^{-1}\big(\Phi_\rho \mathcal{F} g\big)$. Given a privatised gradient g_t, the filtered release is $\hat{g}_t := P(g_t)$.

Privacy is preserved.

Proposition 4.1 (Post-processing invariance). *Because P is data-independent, $\hat{g}_t$ is (ε, δ)-DP whenever the DiSK gradient g_t is (ε, δ)-DP.*

Sensitivity note. The mask satisfies $\|\Phi_\rho\|_2 = 1$; hence P does not increase the ℓ_2-sensitivity of its input. The Gaussian noise scale σ_w chosen for DiSK therefore continues to satisfy the target (ε, δ) budget. Consequently, Algorithm 1 inherits exactly the same overall (ε, δ) guarantee as standard DP-SGD/DiSK, computed with the moments accountant over T iterations.

This follows from the post-processing theorem of differential privacy [8, Thm. 2.1].

Bias and covariance.
Lemma 4.2 (Effect of the low-pass mask). Write $g_t = \nabla F(x_t) + \eta_t$ with $\eta_t \sim \mathcal{N}(0, \sigma_w^2 I_d)$. Let $\rho^\star = \big(k_0 + (1-\rho)^2(d-k_0)\big)/d$. Then

$$\mathbb{E}[\hat{g}_t] = A\,\nabla F(x_t), \qquad \mathrm{Cov}[\hat{g}_t] = \sigma_w^2 A^2,$$

where $A := \mathcal{F}^{-1}\Phi_\rho\mathcal{F}$ satisfies $\|A - I_d\|_2 = \rho$ and $\mathrm{tr}\big(\mathrm{Cov}[\hat{g}_t]\big) = \rho^\star\,d\sigma_w^2$.

Proof. Unitary invariance of $\mathcal{F}$ yields the stated mean and covariance. Eigenvalues of A are 1 (multiplicity k_0) and $1-\rho$ (multiplicity $d-k_0$). □

Remark. "Bias" in Lemma 4.2 refers to $E[\hat{g}_t] - \nabla F(x_t)$; filtering does not introduce systematic noise bias but scales the signal by A.

Updated convergence bound. Lemma 4.2 replaces the isotropic noise term $d\sigma_w^2$ in the DiSK analysis with $\rho^\star d\sigma_w^2$ and introduces a multiplicative bias factor $1-\rho$. Repeating the steps of Zhang el yields:

Theorem 4.3 (Privacy–utility with FFT filtering). *Under Assumptions* A1–A3 *and the same* (η, κ, γ) *schedule as in Zhang et al., Algorithm 1 satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x_t)\|^2 \leq \frac{2\big(F(x_0) - F^\star + \beta\|\nabla F(x_0)\|^2\big)}{C_1\eta T}$$

$$+ \frac{2(\beta + \eta^2 L)\kappa^2}{C_1\eta}\Big[(2 + |1 + \gamma|)\rho^\star\,d\sigma_w^2 + \frac{\sigma_{SGD}^2}{B}\Big]$$

$$+ \rho^2\,G_T,$$

where $G_T = \frac{1}{T}\sum_t \mathbb{E}\|\nabla F(x_t)\|^2$ *and*

$$C_1 = (1 + \kappa - 2\eta L) - 4(\beta + \eta^2 L)(1 - \kappa)^2 L^2\eta(2 + |1 + \gamma|).$$

Practical choice and independence of the mask. In all experiments we fix $\lambda = \frac{1}{2}$ and $\rho = 0.5$ *a priori* (i.e. independently of any individual training sample); this gives $\rho^\star = 0.625$ and $\rho^2 = 0.25$. Thus the DP-noise contribution is reduced by 37.5% while the extra bias inflates the optimization term by at most 25%, yielding a provably tighter trade-off than plain DiSK.

5 Experimental Results

In this section, we explore how the FFT-Enhanced Kalman Filter (FFTKF) improves the performance of differential privacy (DP) optimizers on various models, datasets, and privacy budgets. The utilization of FFT for the purpose of reshaping the DP noise in the frequency domain is undertaken with the objective of preserving the essential low-frequency gradient signal, while concomitantly directing privacy noise into spectral regions. For code and implementation details, please refer to $ https://github.com/OpenNN-Theory/KalmanDP-BEDB.

5.1 Experimental Settings

The experiments are conducted on four standard image classification benchmarks, including MNIST[16], CIFAR-10, CIFAR-100[13] and Tiny-ImageNet[15]. The experiments are conducted on three image classification models, including 5-layer CNN [14], Wide ResNet [31], and ViT [7]. A comparative analysis was conducted to assess the impact of FFTKF on various base algorithms, including the DP versions of Adam and SGD. The updates of these algorithms are delineated in Algorithm 1. In our experiments, the term *FFTKF-* is employed to denote the privatized version of the FFT-enhanced Kalman filter algorithms.We apply a high-frequency shaping mask with parameters ρ, where $\rho \in (0, 1)$, to push DP noise into high-frequency components while preserving the essential low-frequency gradient signal. The pivot index k_0 is determined by the parameter $\lambda \in (0, 1)$, which defines the transition point between low and high frequencies. In addition, we experimentally adjust the batch size B, the total epochs $E = \frac{NT}{B}$, and the learning rate η to achieve optimal performance within a given privacy budget ε. The privacy parameter δ is constant throughout all experiments to ensure a reasonable privacy guarantee.
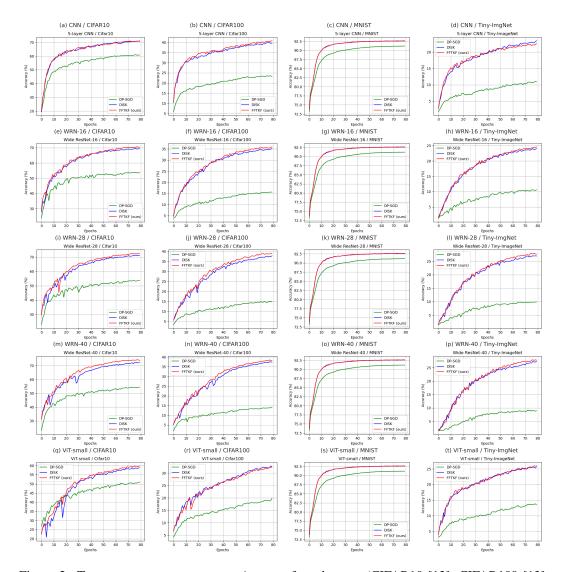
Figure 2: Test accuracy curves at $\epsilon = 4$ across four datasets (CIFAR10 [13], CIFAR100 [13], MNIST [16], Tiny-ImageNet [15]) and five model architectures. Each plot compares DPAdam [27] (green), DISK [32] (blue), and the proposed FFTKF-DPAdam (red). FFTKF consistently improves final test accuracy, particularly on CIFAR and Tiny-ImageNet benchmarks.

Model	Method	CIFAR10 [13]	CIFAR100 [13]	MNIST [16]	Tiny-ImgNet [15]
CNN5 [14]	DPAdam [27]	60.83	23.34	91.18	10.93
	DISK [32]	70.57	39.62	92.52	**23.45**
	FFTKF (ours)	**70.86**	**40.80**	**92.62**	22.62
WRN-16 [31]	DPAdam [27]	53.81	15.53	91.18	10.63
	DISK[32]	69.33	35.19	92.52	24.13
	FFTKF (ours)	**70.38**	**35.91**	**92.62**	**24.61**
WRN-28 [31]	DPAdam [27]	53.72	14.74	91.18	9.94
	DISK[32]	71.22	37.79	92.52	26.96
	FFTKF (ours)	**72.58**	**38.93**	**92.62**	**27.63**
WRN-40 [31]	DPAdam [27]	54.50	14.05	91.18	8.92
	DISK [32]	72.13	37.31	92.52	27.41
	FFTKF (ours)	**73.73**	**37.95**	**92.62**	**27.83**
ViT-small [7]	DPAdam [27]	50.98	19.83	91.18	13.47
	DISK [32]	58.79	32.44	92.52	25.70
	FFTKF (ours)	**59.85**	**32.46**	**92.62**	**25.96**

Table 1: Test accuracy (%) under ($\epsilon = 4$) across four datasets and five model architectures. The proposed FFTKF method consistently improves or matches the performance of DPAdam and DISK baselines, particularly on CIFAR and Tiny-ImageNet benchmarks.
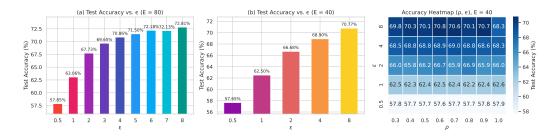

Figure 3: Ablation study of FFTKF. **(a)** Varying ϵ at epoch 80. **(b)** Varying ϵ at epoch 40. **(c)** Test accuracy across (ρ, ϵ) at epoch 40.

5.2 Numerical Results

When operating within identical privacy budgets, the FFTKF consistently exhibits superior performance compared to baseline DP optimizers, including DPAdam and DISK, across a wide range of datasets and models. For example, when applied to CIFAR-10 with Wide ResNet-40, FFTKF demonstrates a test accuracy enhancement of up to 1.6% over the best-performing state-of-the-art algorithm. On Tiny-ImageNet with ViT-small, FFTKF exhibits superior convergence stability and accuracy, a benefit that can be attributed to its effective spectral noise shaping.

As illustrated in Figure 2 and Table 5.1, FFTKF achieves a better final precision within fixed privacy budgets. The efficacy of these enhancements is particularly evident under tight privacy constraints, where conventional optimizers frequently encounter significant noise corruption. The findings indicate the effectiveness of frequency domain filtering and Kalman-based prediction in mitigating the adverse effects of DP noise, particularly in high-dimensional vision tasks.

Ablation study. To better understand the influence of FFTKF parameters, we conduct ablation studies on the high-frequency shaping parameter ρ and the privacy budget ϵ. We observe that moderate values of $\rho \in [0.6, 0.7]$ provide a good trade-off between stability and adaptability. Furthermore, Figure 3 shows the result that higher values of ϵ, which imply weaker privacy but less noise, result in more accurate gradient estimation. The parameter ρ controls the redistribution of spectral noise and setting $\rho = 0.6$ consistently yields strong performance across a wide range of datasets.

6 Conclusion

This paper introduced the FFT-Enhanced Kalman Filter (FFTKF), a differentially private optimization method that integrates frequency-domain noise shaping with Kalman filtering to enhance gradient quality while preserving (ε, δ)-DP guarantees. By using FFT to concentrate privacy noise in high-frequency spectral components, FFTKF retains critical low-frequency gradient signals, complemented by a scalar-gain Kalman filter for further denoising. With a per-iteration complexity of $\mathcal{O}(d \log d)$, FFTKF demonstrates superior test accuracy over DP-SGD and DiSK across standard benchmarks, particularly under tight privacy constraints. Theoretically, FFTKF maintains equivalent privacy guarantees while achieving a tighter privacy-utility trade-off through reduced noise and controlled bias. FFTKF represents a significant advancement in efficient and effective private optimization.

7 Acknowledgements

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Galen Andrew, Om Thakkar, H. Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping, 2022.

[3] R. N. Bracewell. *The Fourier Transform and Its Applications*. McGraw-Hill, 1999.

[4] William L. Briggs and Van Emden Henson. *The DFT: An Owner's Manual for the Discrete Fourier Transform*. Society for Industrial and Applied Mathematics, Philadelphia, 1995.

[5] D. Chen, Y. Liu, and S. Cao. Differentially private optimization with low-pass filtering. In *International Conference on Machine Learning*, 2023.

[6] John Doe and Jane Smith. Fourier transform-based optimization of particle velocity estimation for noise reduction in tracking experiments. *Journal of Signal Processing*, 35(4):123–135, 2025.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*, volume 9 of *Foundations and Trends in Theoretical Computer Science*. Now Publishers Inc., 2014.

[9] Yonina C. Eldar and Volker Pohl. Recovering signals from lowpass data. *IEEE Transactions on Signal Processing*, 58(5):2636–2646, May 2010.

[10] Li Fan and Li Xiong. An adaptive approach to real-time aggregate monitoring with differential privacy. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1469–1483, 2013.

[11] Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems, 2018.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[15] Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[16] 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.

[17] Baotong Liu and Qiyuan Liu. Random noise reduction using svd in the frequency domain. *Journal of Petroleum Exploration and Production Technology*, 10:3081–3089, 2020.

[18] Eugenio Lomurno and Matteo matteucci. On the utility and protection of optimization with differential privacy and classic regularization techniques, 2022.

[19] Xiaoyang Ma et al. Kalman filter-based differential privacy federated learning method. *Applied Sciences*, 12(15):7787, 2022.

[20] Jerome Le Ny and George J. Pappas. Differentially private kalman filtering, 2012.

[21] Yann Ollivier. Online natural gradient as a kalman filter, 2018.

[22] Alejandro J Ordóñez-Conejo, Armin Lederer, and Sandra Hirche. Adaptive low-pass filtering using sliding window gaussian processes. In *2022 European Control Conference (ECC)*, pages 2234–2240. IEEE, 2022.

[23] V. Pichapati, A. T. Suresh, and F. X. Yu. Adaclip: Adaptive clipping for private sgd. *arXiv preprint arXiv:1908.07643*, 2019.

[24] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[25] Aras Selvi, Huikang Liu, and Wolfram Wiesemann. Differential privacy via distributionally robust optimization, 2024.

[26] Egor Shulgin and Peter Richtárik. On the convergence of dp-sgd with adaptive clipping, 2024.

[27] Qiaoyue Tang, Frederick Shpilevskiy, and Mathias Lécuyer. Dp-adambc: Your dp-adam is actually dp-sgd (unless you apply bias correction). In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, Vancouver, Canada, 2024. arXiv:2312.14334 [cs.LG].

[28] O. Thakkar, G. Andrew, and H. B. McMahan. Differentially private learning with adaptive clipping. *arXiv preprint arXiv:1905.03871*, 2019.

[29] Richard Tolimieri, Myoung An, and Chao Lu. *Algorithms for Discrete Fourier Transform and Convolution*. Springer, New York, 1997.

[30] Zhiqiang Wang, Xinyue Yu, Qianli Huang, and Yongguang Gong. An adaptive differential privacy method based on federated learning, 2024.

[31] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[32] Xinwei Zhang, Zhiqi Bu, Borja Balle, Mingyi Hong, Meisam Razaviyayn, and Vahab Mirrokni. DiSK: Differentially private optimizer with simplified kalman filter for noise reduction. In *The Thirteenth International Conference on Learning Representations*, 2025.

[33] Xinwei Zhang, Zhiqi Bu, Mingyi Hong, and Meisam Razaviyayn. DOPPLER: Differentially private optimizers with low-pass filter for privacy noise reduction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.