BEYOND SQUARE ROOTS: A MEMORY-EFFICIENT EXPLICIT FACTORIZATION FOR MULTI-EPOCH PRIVATE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Correlated-noise mechanisms are among the most promising approaches for improving the utility of differentially private model training, but rigorous guarantees require explicit, analyzable factorizations, and practical deployment requires memory efficiency. Recent works have developed banded inverse factorizations, which address both requirements by exploiting a banded structure in the correlation matrix. The bandwidth controls the size of the noise buffer used to correlate noise across iterations, and thus governs the tradeoff between utility and memory cost. At the two ends of this tradeoff, DP-λCGD achieves high memory efficiency by using only a one-step noise buffer, but this limits its utility gains, while the banded inverse square root (BISR) factorization exploits larger correlation windows and is asymptotically optimal for large bandwidths but performs poorly at low bandwidths. To bridge this gap, we introduce γ-BIFR, a unified generalization of both factorizations. In the low-memory, low-bandwidth regime, γ-BIFR improves RMSE, amplified RMSE, and private training performance, while yielding tighter theoretical guarantees for multi-participation error in multi-epoch training.

1 INTRODUCTION

As readily available public data become increasingly exploited, private and sensitive data, such as medical records, financial information, and personal communications, represent an important source of information for machine learning, making strong privacy safeguards essential for their responsible use. Differential Privacy (Dwork et al., 2006) has become the gold standard as a formal notion of privacy. By varying the privacy parameters, one can choose a trade-off between privacy and utility that is appropriate for the problem at hand. In machine learning, differential privacy is typically achieved through variants of differentially private stochastic gradient descent (DP-SGD) (Rajkumar & Agarwal, 2012; Abadi et al., 2016), which clip per-sample gradients to bound their influence on the model and add *independent* Gaussian noise to ensure privacy.

One promising way to improve the utility of DP-SGD is to introduce correlation in the noise across iterations. By partially canceling noise added in previous steps, such methods can improve the utility of the resulting model. This linear correlation structure is typically described by a *correlation matrix*, and the resulting approach is known as the matrix factorization mechanism (Li et al., 2015). Originally developed for linear queries (Li et al., 2015) and continual counting (Dwork et al., 2010; Fichtenberger et al., 2023; Andersson & Pagh, 2023; Andersson et al., 2025), this method has since been extended to machine learning (Denisov et al., 2022), including settings with multi-epoch training (Choquette-Choo et al., 2023b;a; Kalinin & Lampert, 2024).

To introduce correlations in the noise across iterations, one must have access to previously generated noise, for example by storing it. This creates an additional memory and computational burden for correlated-noise methods. Recent work addresses this issue by imposing structural restrictions on the inverse of the correlation matrix, referred to as the strategy matrix, thereby reducing the memory requirement to storing only a small buffer of noise vectors (Dvijotham et al., 2024; McKenna, 2025; Andersson & Pagh, 2025; Henzinger et al., 2025). One recent approach imposes a banded structure on the correlation matrix directly, so that only a number of noise vectors equal to the bandwidth minus one must be stored. The resulting factorizations are referred to as banded inverse methods and include

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the optimization-based BandInvMF (Kalinin et al., 2026b), the explicit Banded Inverse Square Root (BISR) (Kalinin et al., 2026b), and its application to training with learning-rate scheduling (Kalinin & Andersson, 2026). The main advantage of the banded inverse approach is that it cancels only recently added noise, precisely where the model is most likely to benefit from such cancellation. However, BISR performs well only for relatively large bandwidths. BandInvMF addresses this limitation, but it does not provide theoretical guarantees and performs poorly under amplification by subsampling.

A recent work by Kalinin et al. (2026a) introduced DP-λCGD, a memory-free and time-efficient factorization that correlates the noise only with one previous iteration. By exploiting the pseudorandom nature of computer-generated noise, the method avoids storing past noise vectors and instead regenerates them when needed. This idea can be applied efficiently to any banded inverse factorization, since such methods rely only on fresh noise. Its main limitation, however, is utility: the proposed method performs worse than alternative memory-efficient approaches that introduce correlations across a larger number of iterations, such as Buffered Linear Toeplitz (BLT) (McMahan et al., 2024).

In this paper, we strictly generalize DP-λCGD to larger bandwidths by introducing generalized Banded Inverse Fractional Root factorization (γ-BIFR), which takes an arbitrary matrix power γ and recovers BISR when $\gamma = 1/2$. Intuitively, γ controls how aggressively previously added noise is canceled: when $\gamma = 0$, the method corresponds to DP-SGD with no noise cancellation, while when $\gamma = 1$, it fully cancels the noise from the previous iteration. Empirically, we observe that the optimal choice of γ depends strongly on the available bandwidth. In the small-bandwidth regime, the best-performing value of γ is far from $1/2$ and typically much closer to 1, achieving substantially lower error than BISR under the same bandwidth constraint.

Contributions.

- We propose a new explicit γ-BIFR factorization for the low-bandwidth regime, generalizing DP-λCGD to larger bandwidths while preserving memory-free noise regeneration and time efficiency; see Section 3.
- We prove Root Mean Squared Error (RMSE) guarantees in the multi-participation (multi-epoch) setting; see Theorem 2.
- We analyze amplified RMSE and show that tuning γ achieves lower error than existing methods; see Sections B and 3.1.
- We empirically validate the method on CIFAR-10, CIFAR-100 and IMDb, demonstrating accuracy improvements over existing factorization approaches; see Section 4.

2 Background

In this paper, we adopt differential privacy as our formal privacy definition. More specifically, we report our privacy guarantees in the form of approximate differential privacy[1].

Definition 1 (Approximate Differential Privacy (Dwork, 2006))**.** A randomized algorithm M satisfies (ε, δ)-differential privacy if, for every pair of datasets D and D' differing in at most one element, and for every measurable subset S of the output space of M, the following holds:

$$\Pr[M(D) \in S] \leq e^{\varepsilon} \Pr[M(D') \in S] + \delta. \tag{1}$$

The standard approach to ensuring differential privacy during the training of a machine learning model is Differentially Private Stochastic Gradient Descent (DP-SGD) (Abadi et al., 2016). Formally, at each step, we sample a batch B_i, compute the per-example gradients $g_{i,j}$, and clip them using a predefined clipping norm $\zeta > 0$ to ensure that their ℓ_2 norm is bounded. We then average the clipped gradients within the batch to obtain an aggregated gradient $\hat{g}_i$. To ensure privacy, we add appropriately scaled Gaussian noise $z_i \in \mathbb{R}^d$ with variance $\sigma^2_{\varepsilon,\delta}$, guaranteeing (ε, δ)-differential privacy. The computation of the noise level $\sigma_{\varepsilon,\delta}$ can be based on the Gaussian mechanism (Dwork et al., 2006; Balle & Wang, 2018), or can be made more sophisticated by accounting for the randomness introduced by batch sampling (subsampling) (Abadi et al., 2016; Koskela et al., 2020).

The performance of DP-SGD can be substantially improved, both in the centralized setting (with subsampling) and in the federated learning setting (without subsampling), by using the matrix

[1]The matrix mechanism, as a specific instance of the Gaussian mechanism, naturally satisfies a stronger privacy guarantee, namely Gaussian differential privacy (Dong et al., 2022).

factorization mechanism (Li et al., 2015; Denisov et al., 2022). Rather than adding independent Gaussian noise z_i at each iteration, this method introduces noise that is linearly correlated across iterations. The correlation coefficients are collected in a lower triangular matrix, commonly denoted by C^{-1} and referred to as the "correlation matrix". Its inverse, C, is referred to as the "strategy" matrix and is used to determine the noise level required to maintain differential privacy. Given the correlation matrix C^{-1} and an appropriately chosen noise level $\sigma_{\varepsilon,\delta}$, one can train a differentially private model in much the same way as with DP-SGD, but instead by adding correlated noise of the form $\sum_{j=1}^{i}(C^{-1})_{i,j}z_j$. The intuition behind this method is that, when the goal is to continually estimate partial sums of vectors rather than the vectors themselves, one can add more noise to the individual vectors while subtracting parts of the previously added noise in such a way that the total variance is reduced, and can in fact be asymptotically smaller (Dwork et al., 2010).

The design of the strategy matrix C turns out to be subtle, with multiple works proposing different structures, or "factorizations" (see, e.g., the survey by Pillutla et al., 2025). The most common design is based on the Root Mean Squared Error (RMSE) (Denisov et al., 2022; Choquette-Choo et al., 2023b; Kalinin & Lampert, 2024; McKenna, 2025). Specifically, let us concatenate all aggregated clipped gradients into a matrix $\hat{G}$ and all noise vectors into a matrix Z. Privacy is then ensured by adding correlated noise of the form $\hat{G} + C^{-1}Z$. However, to compute the intermediate models, the gradients must be summed over time. For this purpose, consider the lower triangular matrix $E \in \mathbb{R}^{n\times n}$, referred to as the prefix-sum matrix:

$$E = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}. \tag{2}$$

The partial sums of the gradients are then given by $E(\hat{G} + C^{-1}Z) = E\hat{G} + EC^{-1}Z$, which is equivalent to adding correlated noise to the intermediate models with correlation matrix EC^{-1}. The RMSE objective then suggests minimizing

$$\sqrt{\mathbb{E}_Z\|EC^{-1}Z\|_F^2/n} = \sqrt{d}\|EC^{-1}\|_F \cdot \sigma_{\varepsilon,\delta}(C)\zeta/\sqrt{n}, \tag{3}$$

which corresponds to the average variance introduced into the partial sums of the gradients. In the non-amplified setting, the noise level $\sigma_{\varepsilon,\delta}(C)$ is given by the product of the Gaussian mechanism noise multiplier for sensitivity 1, denoted by $\sigma_{\varepsilon,\delta}$ (can be computed tightly numerically (Balle & Wang, 2018)), and the global sensitivity of the product $C\hat{G}$, namely

$$\mathrm{sens}(C) = \sup_{G\sim G'} \|CG - CG'\|_F, \tag{4}$$

where the streams G and G' differ in the participation of a single data point or user. To compute this sensitivity, it is sufficient to consider non-adaptive streams, meaning that one only needs to account for the direct participation of data points in the gradient computation and may ignore their indirect influence through the model weights in future iterations. For a formal proof of this statement, we refer to Theorem 2.1 in Denisov et al. (2022). To compute the sensitivity, one must first specify the participation pattern, namely the iterations in which a data point is allowed to participate. In the non-amplified setting, we adopt the notion of so-called *b-min separation* (Choquette-Choo et al., 2023a), whereby the time gap between consecutive participations is at least b. One may further restrict the maximum number of participations to be at most k, although it is typically assumed that k is as large as the separation parameter allows. For lower triangular Toeplitz matrices with positive, decreasing coefficients along the sub-diagonals, the sensitivity takes a particularly simple form, given by the following theorem.

Theorem 1 (Sensitivity for lower triangular Toeplitz matrices under b-min-separation; Theorem 2 from Kalinin & Lampert (2024))**.** *Let C be a lower triangular Toeplitz matrix with decreasing non-negative entries $c_0 \geq c_1 \geq c_2 \geq \cdots \geq c_{n-1} \geq 0$ on the diagonals. Then the* sensitivity *of C in the setting of* b-min-separation *is*

$$\mathrm{sens}_{k,b}(C) = \left\| \sum_{j=0}^{k-1} C_{[\cdot,\, jb]} \right\|_2 = \sqrt{\sum_{i=0}^{n-1} \left[\sum_{j=0}^{\min(k-1,\lfloor i/b \rfloor)} c_{i-jb} \right]^2}. \tag{5}$$

Since the factor $\sqrt{d}\zeta$ in Equation (3) is independent of the correlation matrix C^{-1}, we omit it and, assuming $E = BC$, define the RMSE, with or without amplification, as

$$\mathrm{RMSE}_{\varepsilon,\delta}(B, C) := \|B\|_F \cdot \sigma_{\varepsilon,\delta}(C)/\sqrt{n} \overset{\text{(no amp.)}}{=} \|B\|_F \cdot \mathrm{sens}(C)\sigma_{\varepsilon,\delta}/\sqrt{n}, \tag{6}$$

where the second equality holds in the absence of amplification. For our theoretical analysis in the non-amplified setting, we set $\sigma_{\varepsilon,\delta} = 1$ to simplify the expressions and remain consistent with previous work, and define

$$\mathcal{E}(B, C) := \|B\|_F \cdot \mathrm{sens}(C)/\sqrt{n}. \tag{7}$$

Given an explicit way of computing the sensitivity and the RMSE objective Equation (7), one can obtain a matrix C via gradient-based optimization (Choquette-Choo et al., 2023a). There are, however, notable limitations to this approach. First, the factorization obtained through gradient-based methods is essentially a black box and provides limited theoretical insight and guarantees. Second, in order to correlate the noise across iterations, one must store the noise in a buffer, which requires a significant memory overhead compared with DP-SGD. Several works have addressed these limitations in the multi-participation setting, including Banded Inverse Square Root (BISR) (Kalinin et al., 2026b) and DP-λCGD (Kalinin et al., 2026a).

Both methods belong to the class of *banded inverse* factorizations, in which the correlation matrix C^{-1} has bandwidth p: only the main diagonal and the first $p-1$ subdiagonals can be nonzero. Consequently, the noise added at a given iteration depends only on noise generated during the most recent $p-1$ iterations, reducing the memory requirement. BISR constructs these coefficients from the inverse square root of the prefix-sum matrix $E^{-1/2}$. DP-λCGD instead uses the smallest nontrivial bandwidth $p = 2$, with C^{-1} having 1 on the main diagonal and $-\lambda$ on the first subdiagonal, so that each update correlates its noise only with the immediately preceding iteration.

BISR provides the greatest benefit when larger bandwidth, and hence more memory, is available. DP-λCGD is instead designed for the low-memory regime, achieving strong performance with the smallest nontrivial bandwidth, but at the cost of reduced utility from its very limited correlation window. In the next section, we introduce a unified family that interpolates between these regimes through a tunable fractional power.

3 γ-Banded Inverse Fractional Root (γ-BIFR)

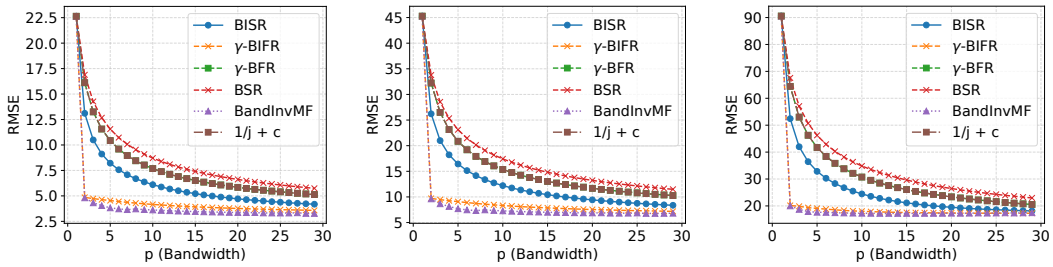
Our goal is to obtain an explicit banded-inverse factorization that performs well across a broad range of bandwidths, including the low-memory regime where BISR is less effective. The key observation is that the inverse square-root construction underlying BISR fixes the matrix exponent to $1/2$, although this choice need not be optimal when the bandwidth is small.

We replace the fixed square root by a tunable fractional power $\gamma \in [0, 1]$. Varying γ changes how aggressively previously added noise is canceled, allowing the factorization to adapt to the available bandwidth. The resulting family recovers BISR at $\gamma = 1/2$ and, at bandwidth $p = 2$, includes DP-λCGD as a special case for $\gamma = \lambda$. We now formalize this construction, which we call γ-BIFR .

Definition 2 (γ-BIFR : γ-Banded Inverse Fractional Root)**.** Fix $0 \le \gamma \le 1$ and an integer bandwidth parameter $p \ge 1$. We define the strategy γ-BIFR matrix $C_\gamma^{(p)}$ through its inverse by

$$\left(C_\gamma^{(p)}\right)^{-1} := \begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \\ (\tilde{c}_\gamma)_1 & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ (\tilde{c}_\gamma)_{p-1} & (\tilde{c}_\gamma)_{p-2} & \cdots & 1 & \cdots & 0 \\ 0 & (\tilde{c}_\gamma)_{p-1} & \cdots & (\tilde{c}_\gamma)_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (\tilde{c}_\gamma)_{p-1} & \cdots & 1 \end{pmatrix}, \qquad \text{for} \quad (\tilde{c}_\gamma)_i := (-1)^i \binom{\gamma}{i}, \tag{8}$$

where $\binom{\gamma}{i} = \prod_{j=1}^{i} \frac{\gamma+1-j}{j}$ is a generalized binomial coefficient. That is, $\left(C_\gamma^{(p)}\right)^{-1}$ is the lower triangular Toeplitz matrix with first p subdiagonals given by $1, (\tilde{c}_\gamma)_1, \ldots, (\tilde{c}_\gamma)_{p-1}$, and all remaining entries equal to 0.



(a) $k = 1$ (b) $k = 4$ (c) $k = 16$

Figure 1: Comparison of the proposed γ-BIFR and γ-BFR factorizations with the explicit BSR (Kalinin & Lampert, 2024) and BISR factorizations, as well as the zeroth-order banded factorization $1/j + c$, where c is optimized to minimize the RMSE. All plots use $n = 1024$ iterations for different numbers of participations k and no amplification by subsampling. We observe that γ-BIFR is consistently much closer to the optimal banded inverse factorization BandInvMF, substantially improving the error compared with BISR. In contrast, for small bandwidths, γ-BFR provides little improvement over the $1/j + c$ factorization.

The correlation coefficients $\tilde{c}_\gamma$ can be computed efficiently using the simple linear recurrence

$$(\tilde{c}_\gamma)_0 = 1, \qquad (\tilde{c}_\gamma)_j = (\tilde{c}_\gamma)_{j-1} \cdot \frac{j - 1 - \gamma}{j} \quad \text{for } j \geq 1.$$

We refer to Lemma 5 in the appendix for a proof of the form of the coefficients $\tilde{c}_\gamma$, as well as for tight bounds on their values. In Figure 1, we show the behavior of various factorizations in the low-memory (small-p) regime, demonstrating that the proposed γ-BIFR achieves almost optimal RMSE values while remaining an explicit factorization that is easy to construct and analyze.

We theoretically bound the RMSE of the proposed γ-BIFR in the following theorem.

Theorem 2. *For any bandwidth $2 \leq p \leq n$ and number of participations (epochs) $1 \leq k \leq \frac{n}{b}$, γ-BIFR with $\gamma = \gamma^\star$ where*

$$\gamma^\star = \max\left\{\frac{1}{2}, 1 - \frac{\log(1 + p \log p/\sqrt{n})}{\log p}\right\}$$

attains an error bound of

$$\mathcal{E}\left(B_\gamma^{(p)}, C_\gamma^{(p)}\right) = O\left(\sqrt{k}\log p + \sqrt{\frac{nk}{b}} + \sqrt{\frac{nk\log(1 + p/\sqrt{n})}{p}} + \sqrt{\frac{kp\log p}{b}}\right).$$

The proof of Theorem 2 relies on the following lemma, which may also be of independent interest. It provides a tight, convenient characterization, up to constant factors, of the sensitivity of any Toeplitz matrix C with positive, nonincreasing diagonals in terms of standard operator norms.

Lemma 1 (Multi-participation sensitivity via operator norms)**.** *Let $C \in \mathbb{R}^{n\times n}$ be a lower-triangular Toeplitz matrix with nonnegative, nonincreasing entries $c_0 \geq c_1 \geq c_2 \geq \cdots \geq c_{n-1} \geq 0$ on the diagonals. Then the* sensitivity *of C in the setting of b-*min-separation *satisfies the following bounds:*

$$\frac{k}{2}\|C\|_{1\to 2}^2 + \frac{k}{3b}\|C\|_{1\to 1}^2 \leq \mathrm{sens}_{k,b}(C)^2 \leq k\|C\|_{1\to 2}^2 + \frac{k}{b}\|C\|_{1\to 1}^2,$$

where the lower bound requires $n = kb$, while the upper bound holds for any $k \leq \frac{n}{b}$. The norms $\|\cdot\|_{1\to 2}$ and $\|\cdot\|_{1\to 1}$ denote the maximum column ℓ_2 and ℓ_1 norms, respectively.

Comparison with BISR and DP-λCGD. The bound in Theorem 2 coincides with the RMSE guarantee of BISR, except that the $\log p$ in the term $\sqrt{nk\log(p)/p}$ is replaced by $\log(1 + p/\sqrt{n}) \leq 2\log p$. In particular, γ-BIFR is never worse than BISR, and exactly recovers it for $\gamma = 1/2$. This guarantee is asymptotically optimal for sufficiently large bandwidth: when $k = \lceil n/b \rceil$, the optimal

Algorithm 1 Differentially Private Model Training with γ-BIFR

Require: Model initialization $\theta_0 \in \mathbb{R}^d$, dataset $\mathcal{D}$, batch size B, clip norm ζ, learning rate $\eta > 0$, loss $\ell(\theta, d)$, parameter $\gamma \in [0, 1)$, b separation, bandwidth $p \geq 1$, target privacy level ε, δ, number of iterations n, pseudorandom noise generator Gen, random seed GSt_0.

Let $(\tilde{c}_\gamma)_0 = (c_\gamma^{(p)})_0 = 1$
$(\tilde{c}_\gamma)_j \leftarrow (\tilde{c}_\gamma)_{j-1} \cdot \frac{j-1-\gamma}{j}$ for $j = 1, \dots p-1$
$(c_\gamma^{(p)})_j \leftarrow -\sum_{i=1}^{\min(j,p-1)} (\tilde{c}_\gamma)_i (c_\gamma^{(p)})_{j-i}$ for $j = 1, \dots n-1$
$C_\gamma^{(p)} \leftarrow \mathrm{LTT}(c_\gamma^{(p)})$ ▷ Construct the strategy matrix of γ-BIFR
$\sigma \leftarrow \mathrm{Subsampling_Accountant}(C_\gamma^{(p)}, n, b, B, |\mathcal{D}|, \varepsilon, \delta)$ ▷ Compute noise multiplier
for $i = 1$ to n **do**
  Sample batch S_i in accordance with the sampling scheme
  $g_j \leftarrow \nabla_\theta \ell(\theta_{i-1}, d_j)$ for $j = 1, \dots, |S_i|$
  $\tilde{g}_j \leftarrow \min\left(1, \frac{\zeta}{\|g_j\|}\right) g_j$ for $j = 1, \dots, |S_i|$ ▷ per-example clipping
  $x_i \leftarrow \sum_{j=1}^{|S_i|} \tilde{g}_j$ ▷ aggregate clipped gradients
  $\mathrm{Gen} \leftarrow \mathrm{Gen.set_state}(\mathrm{GSt}_{\max(i-p,0)})$ ▷ Initialize PRNG's state
  $\hat{x}_i \leftarrow x_i$
  for $j = \max(1, p-i+1)$ to $p-1$ **do**
    $Z_{i-p+j} \sim \mathrm{Gen}()$ ▷ regenerate noise
    $\hat{x}_i \leftarrow \hat{x}_i + \zeta\sigma(\tilde{c}_\gamma)_{p-j} Z_{i-p+j}$
    $\mathrm{GSt}_{i-p+j} \leftarrow \mathrm{Gen.get_state}()$
  end for
  $Z_i \sim \mathrm{Gen}()$ ▷ generate fresh noise
  $\mathrm{GSt}_i \leftarrow \mathrm{Gen.get_state}()$
  $\hat{x}_i \leftarrow \hat{x}_i + \zeta\sigma Z_i$
  $\theta_i \leftarrow \theta_{i-1} - \frac{\eta}{B}\hat{x}_i$ ▷ model update
end for
return θ_n

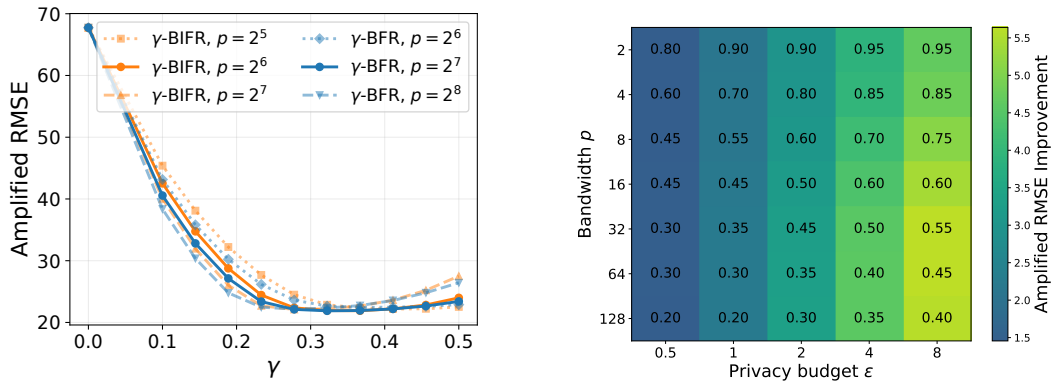
RMSE over all factorizations is $\Omega(k + \sqrt{k}\log n)$, and BISR attains this rate with an appropriate choice of $p = O(b \log b)$.

At small bandwidths, however, the square-root choice $\gamma = 1/2$ can be substantially suboptimal. In particular, for $p = 2$ and $k = \lceil n/b \rceil$, the BISR bound scales as $O(k + \sqrt{nk})$, whereas DP-λCGD, which is designed specifically for bandwidth $p = 2$, achieves $\Theta(k + \sqrt{k}n^{1/4})$. The gap is especially pronounced when the number of participations k is small.

Theorem 2 extends this favorable low-bandwidth behavior beyond $p = 2$. In particular, for $p \leq \sqrt{n}$, an appropriate choice of γ matches the asymptotic scaling of DP-λCGD while allowing larger bandwidths. Thus, γ-BIFR retains the large-bandwidth behavior of BISR while extending the favorable low-bandwidth behavior of DP-λCGD.

Noise Regeneration. As a factorization in the class of banded inverse methods, γ-BIFR can efficiently correlate noise without storing it in a buffer by using pseudorandomness. Most noise generators, including cryptographically secure ones, produce noise sequentially and deterministically from a potentially physically random (system-random) seed. Thus, one can store the state of the generator, for example, a 128-bit state in the commonly used Philox4x32-10 generator in PyTorch, and use it to *regenerate* the noise from the last $p-1$ iterations, assuming bandwidth p. Therefore, the proposed method can be implemented with negligible memory overhead for noise storage compared with DP-SGD.

Noise regeneration, however, incurs additional computation time. Regenerating noise from all previous iterations would introduce substantial overhead, making non-banded-inverse methods such as BandMF, BSR, and BLT impractical in this setting. In the work of Kalinin et al. (2026a), it was



(a) Amplified RMSE vs γ

(b) Optimal γ (text) and Amplified RMSE (colors)

Figure 2: (a) Balls-in-Bins amplified RMSE for the proposed γ-BIFR and γ-BFR methods with $n = 2048$, $k = 8$, $\varepsilon = 1$, and $\delta = 10^{-5}$. (b) Optimal values of γ for γ-BIFR are shown in the cells, together with the corresponding optimal amplified RMSE improvement over DP-SGD ($\gamma = 0$), indicated by color. The results are reported for CIFAR-10/100 training setting with batch size 512, $k = 20$ epochs, and $\delta = 10^{-5}$.

shown that 16 noise vectors can be regenerated without significant time overhead relative to DP-SGD. Moreover, this limit can be pushed even further by increasing the batch size.

Banded factorization. We also compare against the following banded alternative.

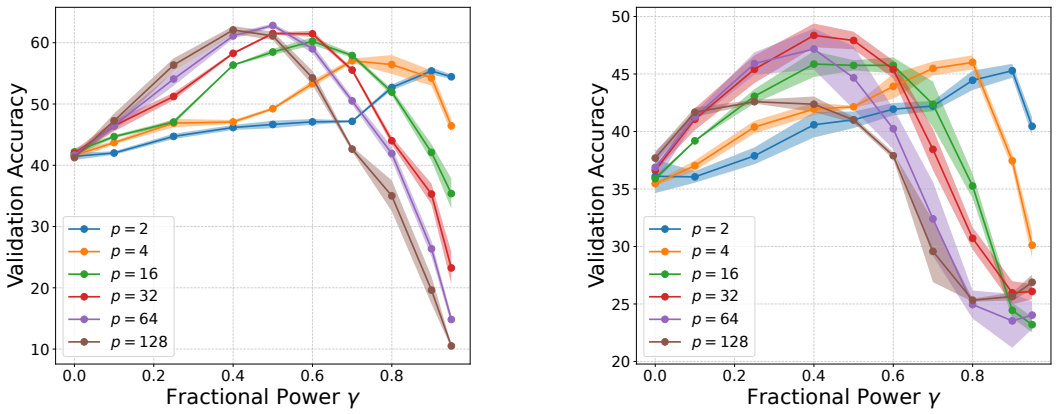
Definition 3 (γ-BFR: γ-Banded Fractional Root)**.** Fix $0 \leq \gamma < 1$ and an integer bandwidth parameter $p \geq 1$. The γ-BFR strategy matrix $C^{(p)}_{\gamma,\mathrm{BFR}}$ is the bandwidth-p truncation of E^{γ}, retaining the main diagonal and the first $p-1$ subdiagonals.

For non-amplified RMSE, γ-BFR scales poorly with bandwidth and does not improve over the explicit banded factorization of McKenna (2025), whose coefficients have the form $1/j + c$ for a constant c (see Appendix J therein). Nevertheless, γ-BFR is useful in the amplified setting without memory restrictions, where it improves RMSE under b-sep subsampling; see Section B.

3.1 Amplification by Subsampling

Matrix mechanisms can also be analyzed together with amplification by subsampling, meaning that the privacy guarantees can be improved by accounting for randomness in batch formation. Several sampling schemes are available for matrix mechanisms, including "Cyclic Poisson" (Choquette-Choo et al., 2023a), "Poisson" subsampling (Choquette-Choo et al., 2024b), Balls-in-Bins (Choquette-Choo et al., 2025; Schuchardt & Kalinin, 2026), and the novel b-min-separated Balls-in-Bins (Dong & Ganesh, 2026). Each sampling scheme has its own limitations. "Cyclic Poisson" provides limited amplification and is applicable only to banded methods. "Poisson" accounting based on conditional composition is extremely loose. Balls-in-Bins, which is the most practical method for banded-inverse matrices and is available in the JAX_privacy library (McKenna et al., 2026), provides, however, limited amplification because batch formation involves less randomness than in Poisson subsampling. Finally, the analysis of b-min-separated Balls-in-Bins was designed for banded matrices and is less applicable to banded-inverse matrices.

In this subsection, we analyze the amplified RMSE under Balls-in-Bins amplification by subsampling, as it provides the strongest amplification for banded-inverse methods. Specifically, instead of computing the multi-participation sensitivity $\mathrm{sens}_{k,b}(C)$, we directly plug the matrix into the accountant and scale the expected error by the resulting noise multiplier; see Equation (6). First, we observe that the optimal value of γ for minimizing the amplified RMSE is smaller than $1/2$ when arbitrary bandwidth is allowed; see Figure 2. In the left subfigure (a), we also observe that, near the optimal value, the objective is not very sensitive to the choice of γ. Moreover, the alternative γ-BFR factorization, being banded, achieves almost identical performance, when there is no restriction on the bandwidth.



(a) privacy budget $\varepsilon = 8$.

(b) privacy budget $\varepsilon = 1$.

Figure 3: Validation accuracy on CIFAR-10 for different values of γ using Balls-in-Bins amplified γ-BIFR, shown for different bandwidths p, with $k = 10$ epochs and batch size $B = 128$, $\delta = 10^{-5}$. For small bandwidths, choosing γ closer to 1 rather than the BISR choice $\gamma = 1/2$ can significantly improve performance. However, for larger bandwidths, γ lower than $1/2$ could be the best choice.

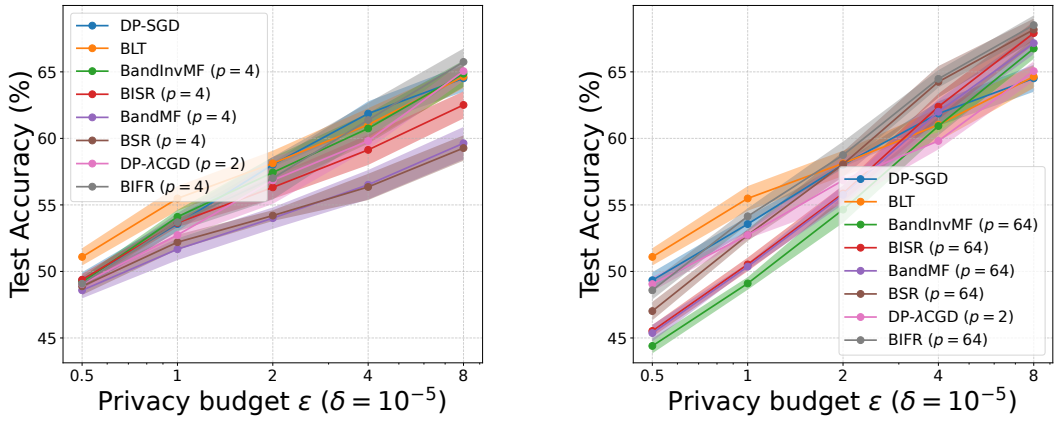
In the right subfigure (b), we evaluate the optimal value of γ (shown as text) and the improvement in amplified RMSE over DP-SGD ($\gamma = 0$) for the corresponding privacy level (shown as colors). We show that for small bandwidth values, the optimal value of γ is much larger than $1/2$, whereas for large bandwidth values, it is lower. The optimal RMSE is insensitive to the bandwidth when optimized over γ for small ε. However, for large values of ε, much larger values of γ are preferred, with $p = 64$ being optimal in the considered practical setting.

We compare the amplified RMSE of different factorizations in Table 8 in the appendix. We show that, in the high-privacy regime, the proposed γ-BIFR matches or improves upon the amplified RMSE of competing methods. We also note that zero-order methods, which require optimizing only a single scalar parameter, achieve the best performance, outperforming the gradient-based optimization of BandMF with respect to the Balls-in-Bins accountant described in (Choquette-Choo et al., 2025). We emphasize that optimizing a factorization through the Balls-in-Bins analysis is a challenging optimization problem and offers no guarantees of convergence to a globally optimal factorization. Together with the $1/j + c$ factorization, which was proposed but whose amplified RMSE was never evaluated, our factorizations are the first to demonstrate improved amplified RMSE performance. This metric also better reflects practical performance, as it accounts for the amount of noise added to the model. For the evaluation, we used Google DeepMind's JAX_privacy implementation (McKenna et al., 2026) of various factorizations and amplification schemes.

4 Experiments

In this section, we evaluate the performance of the proposed γ-BIFR on machine learning tasks. We begin by studying how performance depends on the value of γ. Specifically, we consider a simple setting of training a CNN model on the CIFAR-10 dataset with batch size 128 over 10 epochs at various privacy levels. In Figure 3, we plot the validation accuracy as a function of γ. The results show that, for small bandwidth p, the optimal value of γ differs significantly from $1/2$, and thus outperforms the BISR factorization in the small-bandwidth regime. Moreover, the accuracy achieved with bandwidth $p > 2$ is substantially higher than that with bandwidth $p = 2$, thereby outperforming DP-λCGD.

We compare the test accuracy of γ-BIFR with prior methods under two memory regimes: a high-memory setting with bandwidth $p = 64$ for all banded methods and a low-memory setting with $p = 4$. We evaluate image classification on CIFAR-10 and CIFAR-100 using the 3-block VGG model from prior work (Kairouz et al., 2021), and sentiment analysis on IMDb by fine-tuning BERT-Base (Devlin et al., 2019). The results for CIFAR-10 are shown in Figure 4, while the results for CIFAR-100 and IMDb are shown in Figure 5 and Figure 6, respectively, in the appendix. The CIFAR-10/100



(a) CIFAR-10/VGG. low memory

(b) CIFAR-10/VGG. high memory

Figure 4: Test accuracy on CIFAR-10 in the low- ($p = 4$) and high-memory ($p = 64$) regimes for different factorizations across privacy budgets ε. The full experimental setup, including hyperparameters and a table of accuracy values (Tables 4 and 5), is provided in Section A of the appendix.

experiments are run for 20 epochs, whereas the IMDb experiments are run for 10 epochs. In all cases, we use a batch size of 512, clipping norm 10, privacy budgets $\varepsilon \in \{0.5, 1, 2, 4, 8\}$, and $\delta = 10^{-5}$. We report final test accuracy with error bars computed over three repeats.

For the baselines, we compare against BLT (McMahan et al., 2024), BandInvMF and BISR (Kalinin et al., 2026b), BandMF (McKenna, 2025), and BSR (Kalinin & Lampert, 2024). We also compare against standard DP-SGD, for which we use the Opacus implementation (Yousefpour et al., 2021) with Poisson subsampling and the PRV accountant (Gopi et al., 2021). For the matrix-factorization baselines, we use Balls-in-Bins amplification with 500k Monte Carlo samples, following the accountant of Choquette-Choo et al. (2025). For BLT, we use four buffers and we also include the DP-λCGD comparison (Kalinin et al., 2026a). The results on CIFAR-10 and CIFAR-100 show that γ-BIFR improves over the alternative methods in the low-privacy regime and performs on par with the banded-inverse methods, with only BLT achieving better performance in the high-privacy regime. The results on IMDb show comparable performance across the different methods.

Conclusion and Future Directions

We proposed a new factorization method that strictly generalizes both BISR and DP-λCGD. We showed theoretically that it achieves RMSE at least as good as the better of the two methods, and we demonstrated numerically that it improves RMSE, amplified RMSE, and accuracy in CIFAR-10 and CIFAR-100 training. The proposed factorization can also be combined efficiently with noise regeneration, allowing it to run with negligible overhead compared with DP-SGD.

Matrix factorization techniques would benefit greatly from more accurate and efficient amplification accounting. As future work, it is important to combine banded-inverse factorizations with Poisson subsampling in a computationally efficient manner.

AI use statement

In this work, we used generative AI tools to discover the statement of Lemma 1 and the general idea underlying its proof. In particular, this suggestion was made by GPT Codex 5.5 for the upper bound and GPT 6 Astra for the lower bound. We have not used generative AI tools to generate the proofs of technical statements; the detailed proof of Lemma 1, as well as all other technical proofs, was developed by the authors. Additionally, we used generative AI tools for editing and basic coding support. We have reviewed all AI-assisted work. In particular, the AI-suggested statement and proof idea were independently checked and developed by the authors, and any AI-assisted code

was reviewed and verified for correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM Computer and Communications Security Conference (CCS)*, 2016.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

Joel Daniel Andersson and Rasmus Pagh. A smooth binary mechanism for efficient private continual observation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Joel Daniel Andersson and Rasmus Pagh. Streaming private continual counting via binning. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2025.

Joel Daniel Andersson, Rasmus Pagh, Teresa Anna Steiner, and Sahel Torkamani. Count on your elders: Laplace vs Gaussian noise. In *Foundations of Responsible Computing (FORC)*, 2025.

Borja Balle and Yu-Xiang Wang. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning (ICML)*, 2018.

Adrian Rivera Cardoso and Ryan Rogers. Differentially private histograms under continual observation: Streaming selection into the unknown. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Christopher A Choquette-Choo, Arun Ganesh, Ryan McKenna, H Brendan McMahan, John Rush, Abhradeep Guha Thakurta, and Zheng Xu. (Amplified) banded matrix factorization: A unified approach to private training. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023a.

Christopher A Choquette-Choo, H Brendan McMahan, Keith Rush, and Abhradeep Thakurta. Multi-epoch matrix factorization mechanisms for private machine learning. In *International Conference on Machine Learning (ICML)*, 2023b.

Christopher A Choquette-Choo, Krishnamurthy Dvijotham, Krishna Pillutla, Arun Ganesh, Thomas Steinke, and Abhradeep Thakurta. Correlated noise provably beats independent noise for differentially private learning. In *International Conference on Learning Representations (ICLR)*, 2024a.

Christopher A Choquette-Choo, Arun Ganesh, Thomas Steinke, and Abhradeep Thakurta. Privacy amplification for matrix mechanisms. In *International Conference on Learning Representations (ICLR)*, 2024b.

Christopher A Choquette-Choo, Arun Ganesh, Saminul Haque, Thomas Steinke, and Abhradeep Thakurta. Near exact privacy amplification for matrix mechanisms. In *International Conference on Learning Representations (ICLR)*, 2025.

Sergey Denisov, H Brendan McMahan, John Rush, Adam Smith, and Abhradeep Guha Thakurta. Improved differential privacy for SGD via optimal private linear operators on adaptive streams. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American chapter of the association for computational linguistics (NAACL)*, 2019.

Andy Dong and Arun Ganesh. Privacy amplification for BandMF via b-min-sep subsampling, 2026. arXiv preprint arXiv:2602.09338.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2022.

Krishnamurthy Dvijotham, H. Brendan McMahan, Krishna Pillutla, Thomas Steinke, and Abhradeep Thakurta. Efficient and near-optimal noise generation for streaming differential privacy. In *Symposium on Foundations of Computer Science (FOCS)*, 2024.

Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming (ICALP)*, 2006.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, 2006.

Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N Rothblum. Differential privacy under continual observation. In *Symposium on Theory of Computing (STOC)*, 2010.

Hendrik Fichtenberger, Monika Henzinger, and Jalaj Upadhyay. Constant matters: Fine-grained error bound on differentially private continual observation. In *International Conference on Machine Learning (ICML)*. PMLR, 2023.

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical Composition of Differential Privacy. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Monika Henzinger, Nikita P. Kalinin, and Jalaj Upadhyay. Binned group algebra factorization for differentially private continual counting, 2025. arXiv:2504.04398.

James Honaker. Efficient use of differentially private binary trees. In *Theory and Practice of Differential Privacy (TPDP)*, 2015.

Peter Kairouz, Brendan McMahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. In *International Conference on Machine Learning (ICML)*, 2021.

Nikita P. Kalinin and Joel Daniel Andersson. Learning rate scheduling with matrix factorization for private training. In *Foundations of Responsible Computing (FORC)*, 2026.

Nikita P. Kalinin and Christoph H. Lampert. Banded square root matrix factorization for differentially private model training. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Nikita P Kalinin, Ryan McKenna, Rasmus Pagh, and Christoph H Lampert. DP-λCGD: Efficient noise correlation for differentially private model training, 2026a. arXiv preprint arXiv:2601.22334.

Nikita P Kalinin, Ryan McKenna, Jalaj Upadhyay, and Christoph H Lampert. Back to square roots: An optimal bound on the matrix factorization error for multi-epoch differentially private SGD. In *International Conference on Learning Representations (ICLR)*, 2026b.

Antti Koskela, Joonas Jälkö, and Antti Honkela. Computing tight differential privacy guarantees using FFT. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Chao Li, Gerome Miklau, Michael Hay, Andrew McGregor, and Vibhor Rastogi. The matrix mechanism: optimizing linear counting queries under differential privacy. *International Conference on Very Large Data Bases (VLDB)*, 2015.

Ryan McKenna. Scaling up the banded matrix factorization mechanism for differentially private ML. In *International Conference on Learning Representations (ICLR)*, 2025.

Ryan McKenna, Galen Andrew, Borja Balle, Vadym Doroshenko, Arun Ganesh, Weiwei Kong, Alex Kurakin, Brendan McMahan, and Mikhail Pravilov. JAX-Privacy: A library for differentially private machine learning, 2026. arXiv preprint arXiv:2602.17861.

H Brendan McMahan, Zheng Xu, and Yanxiang Zhang. A hassle-free algorithm for private learning in practice: Don't use tree aggregation, use BLTs. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Krishna Pillutla, Jalaj Upadhyay, Christopher A Choquette-Choo, Krishnamurthy Dvijotham, Arun Ganesh, Monika Henzinger, Jonathan Katz, Ryan McKenna, H Brendan McMahan, Keith Rush, et al. Correlated noise mechanisms for differentially private learning, 2025. arXiv preprint arXiv:2506.08201.

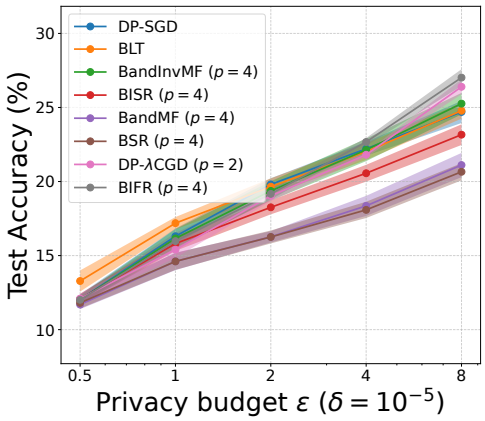
Wahbeh Qardaji, Weining Yang, and Ninghui Li. Understanding hierarchical methods for differentially private histograms. *Proc. VLDB Endow.*, 2013.

Arun Rajkumar and Shivani Agarwal. A differentially private stochastic gradient descent algorithm for multiparty classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

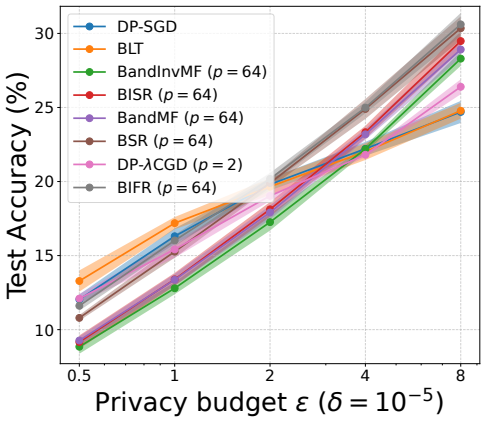
Jan Schuchardt and Nikita Kalinin. Sampling-free privacy accounting for matrix mechanisms under random allocation, 2026. arXiv preprint arXiv:2601.21636.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch, 2021.

A Full Experimental Details

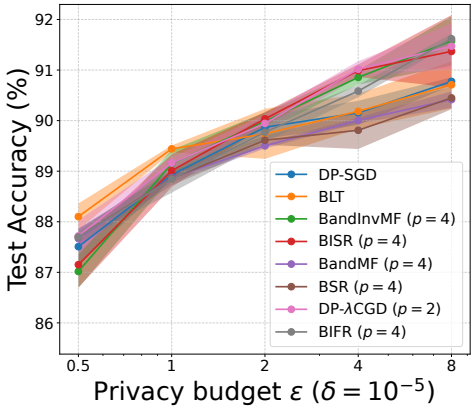


(a) CIFAR-100/VGG. low memory

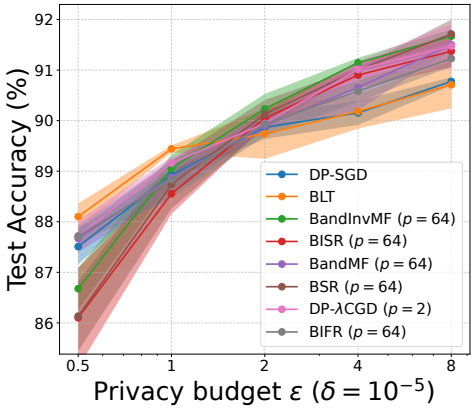


(b) CIFAR-100/VGG. high memory

Figure 5: Test accuracy on CIFAR-100 in the low- ($p = 4$) and high-memory ($p = 64$) regimes for different factorizations across privacy budgets ε. The full experimental setup, including hyperparameters and a table of accuracy values (Tables 2 and 3), is provided in Section A of the appendix.



(a) IMDb/BERT-Base. low memory



(b) IMDb/BERT-Base. high memory

Figure 6: Test accuracy of BERT-Base LoRA (rank 4) fine-tuning on IMDb dataset in the low- ($p = 4$) and high-memory ($p = 64$) regimes for different factorizations across privacy budgets ε. To reduce the variance in the training we additionally used momentum parameter 0.95 to smooth the trajectory. The resulting accuracies (Tables 6 and 7) are not sensitive enough to the method to draw decisive conclusions.

Hyperparameter selection For Bayesian optimization, we separate 10% of the training split for validation. This gives $45{,}000$ training and $5{,}000$ validation examples for CIFAR-10/100, and $22{,}500$ training and $2{,}500$ validation examples for IMDb. For each method, dataset, privacy budget, and seed, we tune only the learning rate with Optuna (Akiba et al., 2019), using 25 trials over the log-uniform range $[10^{-3}, 0.3]$. After selecting the learning rate, we retrain on the full training set and evaluate the model on the test set. We report test accuracy for all tasks.

We fix the batch size 512, clipping norm 10, bandwidth for all methods, and also the γ for γ-BIFR . We select the optimal γ-BIFR γ from a pre-computed (using 200k MC samples) amplified-RMSE grid. For CIFAR-10, these are the optimal values shown in Figure 2b. For fine-tuning BERT-Base on IMDb, we additionally use a momentum parameter of 0.95 to smooth the training trajectory. Table 1 states the exact γ-BIFR γ-values used in the experiments.

Table 1: γ-BIFR γ values used in the fixed-p experiments. The $p = 2$ rows correspond to the DP-λCGD comparison. These values are chosen based on the amplified RMSE.

Dataset	p	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 8$
CIFAR-10/100	2	0.80	0.90	0.90	0.95	0.95
CIFAR-10/100	4	0.60	0.70	0.80	0.85	0.85
CIFAR-10/100	64	0.30	0.30	0.35	0.40	0.45
IMDb	2	0.70	0.80	0.85	0.90	0.90
IMDb	4	0.50	0.55	0.70	0.75	0.80
IMDb	64	0.15	0.20	0.30	0.35	0.40

We report the accuracies in Tables 6 and 7. For the γ selection, we also compared tuning γ for the momentum workload, thereby accounting for the fact that momentum is used. However, we did not observe any significant improvement, and thus kept the same γ as in the no-momentum setting in the plots.

Compute resources We ran each training run as one Slurm job on one full GPU node. Each node consists of four AMD MI250X GPU modules, one 64-core AMD CPU, 512 GiB CPU memory, and 512 GiB total GPU memory. Individual CIFAR-10/VGG runs took around 1–2 node-hours, and individual IMDb/BERT-Base runs took 4.5–6 node hours including hyperparameter selection. The reported 480 runs used approximately 5k GPU hours. Additionally, for development, preliminary sweeps, failed jobs, and debugging, we estimate that we used an extra 5–10k GPU hours.

A.1 Private training experiment values

In this appendix, we report the final test accuracy values used in Figures 4 to 6. The CIs are 95% t-intervals computed over three repeats.

Table 2: Final test accuracy for CIFAR-100/VGG, low memory. Entries are mean accuracy in percent with 95% confidence intervals over three repeats.

Method	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 8$
DP-SGD	12.04 ± 1.04	16.31 ± 1.60	$\mathbf{19.81 \pm 1.26}$	22.21 ± 1.43	24.70 ± 2.27
BLT	$\mathbf{13.28 \pm 2.17}$	$\mathbf{17.19 \pm 1.25}$	19.65 ± 1.72	22.02 ± 1.72	24.78 ± 1.62
BandInvMF ($p = 4$)	12.02 ± 0.84	16.16 ± 1.90	19.37 ± 1.97	22.14 ± 2.31	25.26 ± 2.21
BISR ($p = 4$)	12.06 ± 1.11	15.81 ± 1.50	18.26 ± 1.45	20.56 ± 1.73	23.16 ± 2.12
BandMF ($p = 4$)	11.69 ± 0.94	14.61 ± 1.80	16.27 ± 1.16	18.36 ± 2.06	21.11 ± 2.49
BSR ($p = 4$)	11.81 ± 1.16	14.61 ± 1.74	16.26 ± 1.33	18.09 ± 1.67	20.66 ± 1.76
DP-λCGD ($p = 2$)	12.10 ± 0.94	15.40 ± 1.70	19.03 ± 1.06	21.81 ± 0.52	26.40 ± 1.53
γ-BIFR ($p = 4$)	12.01 ± 1.02	15.98 ± 1.65	19.17 ± 1.15	$\mathbf{22.67 \pm 0.81}$	$\mathbf{27.02 \pm 1.73}$

Table 3: Final test accuracy for CIFAR-100/VGG, high memory. Entries are mean accuracy in percent with 95% confidence intervals over three repeats.

Method	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 8$
DP-SGD	12.04 ± 1.04	16.31 ± 1.60	19.81 ± 1.26	22.21 ± 1.43	24.70 ± 2.27
BLT	$\mathbf{13.28 \pm 2.17}$	$\mathbf{17.19 \pm 1.25}$	19.65 ± 1.72	22.02 ± 1.72	24.78 ± 1.62
BandInvMF ($p = 64$)	8.85 ± 1.44	12.79 ± 1.24	17.26 ± 1.72	22.19 ± 0.67	28.30 ± 1.48
BISR ($p = 64$)	9.16 ± 1.36	13.39 ± 1.33	18.12 ± 1.51	23.33 ± 1.31	29.48 ± 1.94
BandMF ($p = 64$)	9.27 ± 0.85	13.37 ± 1.19	17.91 ± 1.41	23.19 ± 1.11	28.91 ± 1.67
BSR ($p = 64$)	10.79 ± 0.58	15.27 ± 1.27	19.90 ± 0.94	24.90 ± 2.04	30.35 ± 2.42
DP-λCGD ($p = 2$)	12.10 ± 0.94	15.40 ± 1.70	19.03 ± 1.06	21.81 ± 0.52	26.40 ± 1.53
γ-BIFR ($p = 64$)	11.61 ± 0.99	16.02 ± 1.17	$\mathbf{20.18 \pm 1.24}$	$\mathbf{24.99 \pm 1.16}$	$\mathbf{30.61 \pm 2.35}$

Table 4: Final test accuracy for CIFAR-10/VGG, low memory. Entries are mean accuracy in percent with 95% confidence intervals over three repeats.

Method	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 8$
DP-SGD	49.34 ± 1.83	53.56 ± 1.69	58.03 ± 0.55	$\mathbf{61.86 \pm 2.80}$	64.52 ± 3.09
BLT	$\mathbf{51.09 \pm 1.97}$	$\mathbf{55.49 \pm 2.82}$	$\mathbf{58.12 \pm 2.89}$	61.03 ± 3.78	64.64 ± 2.59
BandInvMF ($p = 4$)	49.18 ± 1.17	54.10 ± 1.58	57.44 ± 3.52	60.75 ± 3.48	64.86 ± 3.02
BISR ($p = 4$)	49.38 ± 1.35	53.65 ± 0.95	56.32 ± 2.66	59.14 ± 3.45	62.51 ± 3.09
BandMF ($p = 4$)	48.58 ± 1.85	51.67 ± 2.49	54.01 ± 2.45	56.52 ± 3.37	59.64 ± 3.71
BSR ($p = 4$)	48.88 ± 1.98	52.20 ± 1.71	54.20 ± 1.23	56.35 ± 3.01	59.27 ± 2.90
DP-λCGD ($p = 2$)	49.05 ± 1.64	52.74 ± 0.86	56.85 ± 5.39	59.81 ± 1.94	65.05 ± 2.24
γ-BIFR ($p = 4$)	49.07 ± 1.26	53.70 ± 2.10	57.01 ± 4.87	61.38 ± 3.94	$\mathbf{65.75 \pm 3.08}$

Table 5: Final test accuracy for CIFAR-10/VGG, high memory. Entries are mean accuracy in percent with 95% confidence intervals over three repeats.

Method	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 8$
DP-SGD	49.34 ± 1.83	53.56 ± 1.69	58.03 ± 0.55	61.86 ± 2.80	64.52 ± 3.09
BLT	$\mathbf{51.09 \pm 1.97}$	$\mathbf{55.49 \pm 2.82}$	58.12 ± 2.89	61.03 ± 3.78	64.64 ± 2.59
BandInvMF ($p = 64$)	44.40 ± 1.70	49.10 ± 1.44	54.65 ± 3.03	60.93 ± 2.14	66.75 ± 2.35
BISR ($p = 64$)	45.53 ± 1.35	50.53 ± 1.55	55.86 ± 1.84	62.38 ± 2.97	67.91 ± 2.61
BandMF ($p = 64$)	45.37 ± 1.90	50.36 ± 0.75	55.75 ± 2.18	62.00 ± 2.82	67.16 ± 2.53
BSR ($p = 64$)	47.02 ± 2.11	52.73 ± 1.33	58.05 ± 2.88	64.25 ± 3.60	68.18 ± 2.50
DP-λCGD ($p = 2$)	49.05 ± 1.64	52.74 ± 0.86	56.85 ± 5.39	59.81 ± 1.94	65.05 ± 2.24
γ-BIFR ($p = 64$)	48.58 ± 2.11	54.14 ± 1.84	$\mathbf{58.75 \pm 3.08}$	$\mathbf{64.47 \pm 1.99}$	$\mathbf{68.51 \pm 2.21}$

Table 6: Final test accuracy for IMDb/BERT LoRA rank 4, momentum 0.95, low memory. For γ-BIFR, we report results for both γ settings: one tuned for the momentum workload and one tuned for the prefix-sum workload. Entries are mean accuracy in percent with 95% confidence intervals over 3 repeats.

Method	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 8$
DP-SGD	87.51 ± 1.07	88.93 ± 0.26	89.87 ± 0.70	90.15 ± 0.73	90.77 ± 0.25
BLT	$\mathbf{88.10 \pm 0.80}$	$\mathbf{89.44 \pm 0.21}$	89.74 ± 1.51	90.18 ± 1.04	90.71 ± 1.42
BandInvMF ($p = 4$)	87.02 ± 1.00	89.15 ± 0.82	89.98 ± 0.54	90.85 ± 0.43	91.57 ± 1.48
BISR ($p = 4$)	87.15 ± 1.37	89.01 ± 0.99	90.04 ± 0.41	90.99 ± 0.38	91.37 ± 2.19
BandMF ($p = 4$)	87.67 ± 0.57	88.87 ± 0.27	89.50 ± 0.17	90.01 ± 0.18	90.42 ± 0.49
BSR ($p = 4$)	87.70 ± 0.49	88.86 ± 0.42	89.61 ± 0.28	89.81 ± 1.13	90.45 ± 0.68
DP-λCGD ($p = 2$)	87.72 ± 0.94	89.16 ± 0.35	89.94 ± 0.53	$\mathbf{91.02 \pm 0.48}$	91.47 ± 1.35
γ-BIFR ($p = 4$)	87.66 ± 1.10	89.01 ± 1.00	$\mathbf{90.23 \pm 0.42}$	90.45 ± 1.45	91.00 ± 2.21
γ-BIFR ($p = 4$, w/ moment.)	87.70 ± 0.33	88.84 ± 0.82	89.77 ± 0.61	90.58 ± 0.37	$\mathbf{91.62 \pm 0.30}$

Table 7: Final test accuracy for IMDb/BERT LoRA rank 4, momentum 0.95, high memory. For γ-BIFR, we report results for both γ settings: one tuned for the momentum workload and one tuned for the prefix-sum workload. Entries are mean accuracy in percent with 95% confidence intervals over 3 repeats.

Method	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 8$
DP-SGD	87.51 ± 1.07	88.93 ± 0.26	89.87 ± 0.70	90.15 ± 0.73	90.77 ± 0.25
BLT	$\mathbf{88.10 \pm 0.80}$	$\mathbf{89.44 \pm 0.21}$	89.74 ± 1.51	90.18 ± 1.04	90.71 ± 1.42
BandInvMF ($p = 64$)	86.68 ± 1.26	89.06 ± 0.82	$\mathbf{90.23 \pm 0.90}$	$\mathbf{91.14 \pm 0.30}$	91.67 ± 0.42
BISR ($p = 64$)	86.10 ± 3.00	88.56 ± 1.23	90.05 ± 0.46	90.90 ± 0.81	91.37 ± 0.97
BandMF ($p = 64$)	87.67 ± 0.82	88.88 ± 1.03	89.90 ± 0.26	90.65 ± 1.37	91.50 ± 0.29
BSR ($p = 64$)	86.13 ± 2.10	88.74 ± 1.45	90.13 ± 0.85	91.00 ± 0.15	$\mathbf{91.71 \pm 0.88}$
DP-λCGD ($p = 2$)	87.72 ± 0.94	89.16 ± 0.35	89.94 ± 0.53	91.02 ± 0.48	91.47 ± 1.35
γ-BIFR ($p = 64$)	87.63 ± 0.66	88.87 ± 0.73	90.03 ± 0.10	90.56 ± 1.16	91.25 ± 0.99
γ-BIFR ($p = 64$, w/ moment.)	87.71 ± 0.48	88.83 ± 1.10	89.93 ± 0.74	90.58 ± 0.63	91.23 ± 0.49

B AMPLIFIED RMSE ANALYSIS

Table 8: Comparison of amplified RMSE under the **Balls-in-Bins accountant**, with non-amplified RMSE shown in the last column, for different factorization methods with $n = 2048$ and $k = 8$. For each value, we optimize the bandwidth p over powers of 2, as well as γ, c, or ν, where applicable. To get a tight error estimate, we use 2×10^6 samples for the Balls-in-Bins accountant. The proposed γ-banded and banded-inverse fraction-root factorizations outperform the other factorizations in the high-privacy regime, including BandMF w/ Opt., which is optimized for amplified RMSE. In the low-privacy regime, BandMF w/ Opt. yields the lowest error, while BandMF achieves the best non-amplified RMSE.

Method	ε=1	ε=2	ε=4	ε=8	ε=8 (Non-Amp)
BISR	22.99	13.23	7.60	4.74	6.75
γ-BIFR (ours)	**22.41**	12.65	7.57	4.74	6.69
BandInvMF	28.79	15.43	8.42	4.81	6.55
DP-λCGD	25.15	16.35	10.90	7.24	9.68
BandMF	25.35	14.59	8.42	4.83	**6.35**
BandMF w/ Opt.	22.80	**12.60**	**7.44**	**4.53**	-
BandMF + Heuristic	24.71	13.34	7.57	4.82	-
BSR	23.24	13.16	7.82	4.55	6.57
γ-BFR (ours)	22.65	**12.62**	**7.44**	**4.55**	6.38
$1/j + c$	22.54	12.71	7.50	4.62	6.54
BLT	29.26	15.50	8.48	4.87	6.65
DP-νFTRL	22.72	12.83	7.61	4.82	6.90
k-ary Tree	48.18	25.88	14.16	8.15	12.53

We compare different factorizations in terms of amplified RMSE. For amplification, we primarily use the Balls-in-Bins scheme, which is the most efficient and general scheme (Choquette-Choo et al., 2025); see Table 8. For banded methods, in this section, we also consider cyclic Poisson amplification (Choquette-Choo et al., 2023a), a strict generalization of amplified DP-SGD, and the new b-min-sep sampling scheme (Dong & Ganesh, 2026); see Table 9.

We divide the factorizations in Table 8 into three groups. The first group, banded inverse methods, consists of Banded Inverse Square Root (BISR), the proposed γ-BIFR , Banded Inverse Matrix Factorization (BandInvMF) (Kalinin et al., 2026b), and DP-λCGD (Kalinin et al., 2026a). The second group, banded methods, includes BandMF (McKenna, 2025); its version optimized for Balls-in-Bins accounting (BandMF w/ Opt.) (Choquette-Choo et al., 2025); its version with a heuristic (see Appendix H of McKenna (2025)); Banded Square Root (BSR) (Kalinin & Lampert, 2024); γ-BFR, a banded version of the proposed factorization; and $1/j + c$, a single-parameter factorization with $c > 0$, proposed in Appendix J of (McKenna, 2025). The third group consists of non-banded factorizations, including BLT (McMahan et al., 2024) and DP-νFTRL (Choquette-Choo et al., 2024a); for the latter, we optimize over the choice of ν to achieve the best amplified RMSE.

We also compare against a k-ary tree mechanism, which predates matrix-factorization mechanisms and generalizes the binary tree mechanism to an arbitrary branching factor (Qardaji et al., 2013; Cardoso & Rogers, 2022; Andersson et al., 2025). The main challenge is that, when written naively as an encoding matrix, it is not lower triangular. One can make the encoding matrix lower triangular by a QL decomposition without affecting the error, since the column inner products are preserved, as suggested in Denisov et al. (2022). However, the resulting entries may become negative, preventing tight accounting in multi-participation settings, both with and without amplification. Fortunately, standard k-ary tree mechanisms based on unweighted aggregation of nodes can be expressed as lower-triangular factorizations with entries in $\{0, 1\}$. By contrast, tree-based mechanisms based on more advanced aggregation techniques (such as leveraging *subtraction* of nodes (Honaker, 2015; Andersson et al., 2025)) achieve lower RMSE in the single-participation setting, but are not known to admit factorizations with a positive and lower-triangular encoding matrix. As a result, they are beyond the scope of current accounting methods.

Table 9: Comparison of amplified RMSE under **Poisson and b-min-sep accountants** for different banded factorization methods, with $n = 2048$. For each setting, we optimize the bandwidth p over powers of 2, as well as γ or c, where applicable. Poisson accounting is exact, whereas b-min-sep accounting requires Monte Carlo sampling; for computational reasons, we use 2×10^5 samples.Each entry reports the mean $\pm$ standard deviation over five runs. The lowest mean RMSE for each (ε, k) setting is shown in bold.

		Number of participations (k)				
Method	ε	1	2	4	8	16
γ-BFR	0.5	6.17 ± 0.12	11.33 ± 0.22	$\mathbf{21.10 \pm 0.42}$	41.34 ± 0.78	$\mathbf{80.94 \pm 1.30}$
	1.0	$\mathbf{3.63 \pm 0.08}$	6.58 ± 0.07	11.58 ± 0.27	$\mathbf{22.09 \pm 0.36}$	$\mathbf{43.57 \pm 0.97}$
	2.0	$\mathbf{2.39 \pm 0.04}$	$\mathbf{3.90 \pm 0.04}$	$\mathbf{6.65 \pm 0.15}$	$\mathbf{12.15 \pm 0.25}$	$\mathbf{23.81 \pm 0.42}$
	4.0	$\mathbf{1.66 \pm 0.01}$	$\mathbf{2.42 \pm 0.06}$	$\mathbf{3.95 \pm 0.11}$	$\mathbf{7.56 \pm 0.10}$	$\mathbf{13.52 \pm 0.24}$
	8.0	$\mathbf{1.21 \pm 0.02}$	$\mathbf{1.65 \pm 0.04}$	$\mathbf{2.48 \pm 0.04}$	$\mathbf{4.57 \pm 0.07}$	$\mathbf{7.45 \pm 0.13}$
BSR	0.5	6.88 ± 0.14	11.78 ± 0.26	21.76 ± 0.38	41.33 ± 0.61	81.37 ± 0.89
	1.0	4.49 ± 0.09	7.16 ± 0.13	12.72 ± 0.25	22.78 ± 0.56	44.20 ± 0.92
	2.0	3.12 ± 0.07	4.44 ± 0.10	7.24 ± 0.17	12.95 ± 0.26	24.07 ± 0.42
	4.0	1.89 ± 0.03	2.78 ± 0.04	4.32 ± 0.12	7.89 ± 0.06	13.55 ± 0.25
	8.0	1.24 ± 0.02	1.75 ± 0.04	2.60 ± 0.05	$\mathbf{4.57 \pm 0.07}$	7.58 ± 0.17
BandMF	0.5	7.47 ± 0.17	12.50 ± 0.24	22.45 ± 0.38	42.38 ± 0.87	81.89 ± 1.20
	1.0	5.06 ± 0.13	7.76 ± 0.17	13.75 ± 0.19	23.63 ± 0.52	44.72 ± 0.75
	2.0	2.88 ± 0.06	4.93 ± 0.09	8.03 ± 0.19	13.73 ± 0.28	25.64 ± 0.53
	4.0	1.78 ± 0.02	2.85 ± 0.05	4.77 ± 0.14	8.23 ± 0.15	14.39 ± 0.25
	8.0	$\mathbf{1.22 \pm 0.01}$	1.84 ± 0.04	2.86 ± 0.04	4.85 ± 0.06	8.47 ± 0.13
$1/j + c$	0.5	$\mathbf{5.89 \pm 0.11}$	$\mathbf{11.25 \pm 0.10}$	$\mathbf{21.12 \pm 0.38}$	$\mathbf{41.19 \pm 0.77}$	81.21 ± 1.14
	1.0	4.11 ± 0.11	$\mathbf{6.29 \pm 0.15}$	$\mathbf{11.54 \pm 0.22}$	22.18 ± 0.27	44.61 ± 0.50
	2.0	2.61 ± 0.05	$\mathbf{3.89 \pm 0.07}$	$\mathbf{6.66 \pm 0.15}$	12.21 ± 0.25	23.93 ± 0.42
	4.0	1.92 ± 0.04	2.56 ± 0.04	$\mathbf{3.98 \pm 0.12}$	7.59 ± 0.12	13.53 ± 0.25
	8.0	1.39 ± 0.03	1.71 ± 0.03	$\mathbf{2.49 \pm 0.05}$	4.68 ± 0.08	7.48 ± 0.13
BandMF Pois.	0.5	8.58	13.49	23.31	43.06	82.63
	1.0	6.11	8.79	14.01	24.45	45.44
	2.0	4.51	6.07	8.98	14.61	25.81
	4.0	3.32	4.29	6.01	9.19	15.33
	8.0	1.97	2.81	4.06	5.95	9.48

Many of the methods in Table 8 have never been considered in the multi-participation setting with amplification, making this the first thorough comparison in the amplified setting.

We also compare alternative amplification schemes for banded methods in Table 9, including cyclic Poisson amplification for BandMF (BandMF w/ Pois.). We note that the amplification effect depends only on the bandwidth and not on the coefficients, making BandMF automatically well suited to this amplification scheme. We also compare against the novel b-min-sep amplification scheme (Dong & Ganesh, 2026), which leads to nontrivial differences depending on the coefficients. We evaluate it for BandMF (BandMF w/ b-min-sep), BSR, γ-BFR, a banded version of our proposed factorization, and $1/j + c$. We observe that the parametrized factorizations achieve the lowest amplified RMSE, significantly lower than the corresponding values under Balls-in-Bins subsampling.

C PROOFS

This section is structured as follows. First, we prove the intermediate result of Lemma 1, which relates the sensitivity $\mathrm{sens}_{k,b}(\cdot)$ to the standard operator norms $\|\cdot\|_{1\to1}$ and $\|\cdot\|_{1\to2}$. This allows

for bounding the error $\mathcal{E}(B_\gamma^{(p)}, C_\gamma^{(p)})$ purely in terms of $\|B_\gamma^{(p)}\|_F$, $\|C_\gamma^{(p)}\|_{1\to1}$ and $\|C_\gamma^{(p)}\|_{1\to2}$. We proceed to compute tight bounds on the coefficients of $B_\gamma^{(p)}$ and $C_\gamma^{(p)}$, from which we derive bounds on the corresponding matrix norms, and consequently the error (Corollary 2). Lastly, we compute the error for different choices of $\gamma \in [1/2, 1)$, which ultimately yields Theorem 2.

Throughout $O(\cdot)$ only hides absolute constants independent of n, p, k, b and γ. We also note that $\log(\cdot)$ always refers to the natural logarithm with base e. We will also use the convention that any constant divided by zero is ∞, especially in the context of taking minimum of two expressions.

C.1 Bounding the Sensitivity via Operator Norms

Before proving Lemma 1, we briefly recount the operator norms $\|\cdot\|_{1\to1}$ and $\|\cdot\|_{1\to2}$ that feature in it. For any real-valued matrix M, $\|M\|_{1\to1}$ and $\|M\|_{1\to2}$ correspond to the largest ℓ_1 respectively ℓ_2 norm of any column in M. In the particular case of lower-triangular Toeplitz matrices, the largest such norms are always achieved for the first column.

Proof of Lemma 1. Throughout the proof we zero-index the columns of C, and for $j \in \{0, \dots, n-1\}$ let $C_j := C_{[\cdot,\, j]}$ denote the j-th column of C. By Theorem 1,

$$\mathrm{sens}_{k,b}(C)^2 = \left\| \sum_{j=0}^{k-1} C_{jb} \right\|_2^2 = \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} \langle C_{ib},\, C_{jb}\rangle \le k\langle C_0, C_0\rangle + 2 \sum_{0\le i<j\le k-1} \langle C_{ib}, C_{jb}\rangle. \tag{9}$$

where the inequality uses that $\langle C_j, C_j\rangle = \sum_{r=0}^{n-1-j} c_r^2 \le \sum_{r=0}^{n-1} c_r^2 = \langle C_0, C_0\rangle$. The remainder of the proof deals with the cross-term sum.

Note that for any of the terms $\langle C_{ib}, C_{jb}\rangle$ (where $i < j$), we have that

$$\langle C_{ib}, C_{jb}\rangle = \sum_{\ell=jb}^{n-1} c_{\ell-ib} c_{\ell-jb} = \sum_{r=0}^{n-1-jb} c_{r+(j-i)b} c_r \tag{10}$$

$$\le \sum_{r=0}^{n-1-(j-i)b} c_{r+(j-i)b} c_r = \sum_{\ell=(j-i)b}^{n-1} c_\ell c_{\ell-(j-i)b} = \langle C_0, C_{(j-i)b}\rangle, \tag{11}$$

where the inequality uses that expanding the summation range of non-negative summands can only increase the value of the sum. Hence, we can upper-bound the cross-term sum as

$$2 \sum_{0\le i<j\le k-1} \langle C_{ib}, C_{jb}\rangle \le 2 \sum_{m=1}^{k-1} (k-m)\langle C_0, C_{mb}\rangle \le 2k \sum_{m=1}^{k-1} \langle C_0, C_{mb}\rangle. \tag{12}$$

To upper bound $\langle C_0, C_{mb}\rangle$, we use an averaging argument. First, note that the map $q \mapsto \langle C_0, C_q\rangle = \sum_{r=0}^{n-1-q} c_r c_{r+q}$ is *decreasing* in q, as $0 \le c_{r+q+1} \le c_{r+q}$ (each summand decreases in value), and the summation range decreases with q. In particular, this implies that for each $m \ge 1$, $\langle C_0, C_{mb}\rangle$ is the minimum taken over the window $h \in [(m-1)b+1, mb]$, and so is bounded by the average:

$$\langle C_0, C_{mb}\rangle \le \frac{1}{b} \sum_{h=(m-1)b+1}^{mb} \langle C_0, C_h\rangle. \tag{13}$$

Continuing from Equation (12) and applying Equation (13):

$$2k \sum_{m=1}^{k-1} \langle C_0, C_{mb}\rangle \le \frac{2k}{b} \sum_{m=1}^{k-1} \sum_{h=(m-1)b+1}^{mb} \langle C_0, C_h\rangle = \frac{2k}{b} \sum_{h=1}^{(k-1)b} \langle C_0, C_h\rangle \le \frac{2k}{b} \sum_{h=1}^{n-1} \langle C_0, C_h\rangle \tag{14}$$

where the first equality uses that the windows $\{[(m-1)b+1, mb] : 1 \le m \le k-1\}$ exactly tile $[1, (k-1)b]$, and the last inequality that $(k-1)b \le n-1$.

To finish the proof, we will argue that $\sum_{h=1}^{n-1}\langle C_0, C_h\rangle = \frac{1}{2}(\|C_0\|_1^2 - \|C_0\|_2^2)$. Indeed,

$$\sum_{h=0}^{n-1}\langle C_0, C_h\rangle = \sum_{h=0}^{n-1}\sum_{r=0}^{n-1-h} c_r c_{r+h} = \sum_{0\le r\le r'\le n-1} c_r c_{r'} = \frac{1}{2}\left(\|C_0\|_1^2 + \|C_0\|_2^2\right), \tag{15}$$

where the last step uses that every term $c_r c_{r'}$ appears twice in $\|C_0\|_1^2$, except for c_r^2 which only appears once, which we correct by adding $\|C_0\|_2^2 = \sum_r c_r^2$, after which dividing by 2 yields the equality. Hence,

$$\sum_{h=1}^{n-1}\langle C_0, C_h\rangle = -\langle C_0, C_0\rangle + \sum_{h=0}^{n-1}\langle C_0, C_h\rangle = \frac{1}{2}(\|C_0\|_1^2 - \|C_0\|_2^2). \tag{16}$$

Putting everything together,

$$\begin{aligned}
\mathrm{sens}_{k,b}(C)^2 &\le k\|C_0\|_2^2 + \frac{k}{b}\left(\|C_0\|_1^2 - \|C_0\|_2^2\right) && (17)\\
&= k\left(1-\frac{1}{b}\right)\|C_0\|_2^2 + \frac{k}{b}\|C_0\|_1^2 && (18)\\
&\le k\|C_0\|_2^2 + \frac{k}{b}\|C_0\|_1^2, && (19)
\end{aligned}$$

where the final statement follows from identifying $\|C_0\|_2^2 = \|C\|_{1\to 2}^2$ and $\|C_0\|_1^2 = \|C\|_{1\to 1}^2$, as each of the norms correspond to largest column norm measured in either ℓ_2 or ℓ_1.

Lower Bound. First, we rewrite the matrix inequality $\mathrm{sens}_{k,b}(C)^2 \ge \frac{k}{2}\|C\|_{1\to 2}^2 + \frac{k}{3b}\|C\|_{1\to 1}^2$ in terms of sums of the subdiagonal elements c_j. In particular, let us split $[0, n-1]$ into k segments of the form $\ell + jb$ for $\ell \in [0, b-1]$ and $j \in [0, k-1]$:

$$\mathrm{sens}_{k,b}(C)^2 = \sum_{\ell=0}^{b-1}\sum_{j=0}^{k-1}\left(\sum_{i=0}^{j} c_{\ell+ib}\right)^2 \ge \frac{k}{2}\sum_{\ell=0}^{b-1}\sum_{j=0}^{k-1} c_{\ell+jb}^2 + \frac{k}{3b}\left(\sum_{\ell=0}^{b-1}\sum_{j=0}^{k-1} c_{\ell+jb}\right)^2. \tag{20}$$

It is then sufficient to prove the following inequality for any fixed value of ℓ:

$$\sum_{j=0}^{k-1}\left(\sum_{i=0}^{j} c_{\ell+ib}\right)^2 \ge \frac{k}{2}\sum_{j=0}^{k-1} c_{\ell+jb}^2 + \frac{k}{3}\left(\sum_{j=0}^{k-1} c_{\ell+jb}\right)^2. \tag{21}$$

Indeed, the first two sums are obtained simply by summing over ℓ; for the third sum, we need to use Cauchy–Schwarz:

$$\frac{k}{3}\sum_{\ell=0}^{b-1}\left(\sum_{j=0}^{k-1} c_{\ell+jb}\right)^2 \ge \frac{k}{3b}\left(\sum_{\ell=0}^{b-1}\sum_{j=0}^{k-1} c_{\ell+jb}\right)^2. \tag{22}$$

Denoting $x_\ell = \sum_{j=0}^{k-1} c_{\ell+jb}$, this turns into the simple inequality $b\sum_{\ell=0}^{b-1} x_\ell^2 \ge (\sum_{\ell=0}^{b-1} x_\ell)^2$. It remains to prove Inequality 21. For this, we will use the monotonicity of $c_{\ell+ib}$. Let us denote $d_i = c_{\ell+ib} - c_{\ell+(i+1)b} \ge 0$; now we can rewrite the inequality in terms of d_i, with $d_{k-1} = c_{\ell+(k-1)b}$. In particular,

$$\begin{aligned}
\sum_{j=0}^{k-1}\left(\sum_{i=0}^{j} c_{\ell+ib}\right)^2 &= \sum_{j=0}^{k-1}\left(\sum_{i=0}^{j}\sum_{r=i}^{k-1} d_r\right)^2 = \sum_{j=0}^{k-1}\left(\sum_{i=0}^{k-1} d_i \min(j+1, i+1)\right)^2\\
&= \sum_{i=0}^{k-1}\sum_{r=0}^{k-1} d_i d_r \sum_{j=0}^{k-1}\min(j+1, i+1)\min(j+1, r+1).
\end{aligned} \tag{23}$$

The second sum becomes

$$\frac{k}{2}\sum_{j=0}^{k-1} c_{\ell+jb}^2 = \frac{k}{2}\sum_{j=0}^{k-1}\left(\sum_{i=j}^{k-1} d_i\right)^2 = \frac{k}{2}\sum_{j=0}^{k-1}\sum_{i=j}^{k-1}\sum_{r=j}^{k-1} d_i d_r = \frac{k}{2}\sum_{i=0}^{k-1}\sum_{r=0}^{k-1} d_i d_r \min(i+1, r+1). \tag{24}$$

And the third sum gives

$$\frac{k}{3}\left(\sum_{j=0}^{k-1} c_{\ell+jb}\right)^2 = \frac{k}{3}\left(\sum_{j=0}^{k-1}\sum_{i=j}^{k-1} d_i\right)^2 = \frac{k}{3}\left(\sum_{i=0}^{k-1} d_i(i+1)\right)^2 = \frac{k}{3}\sum_{i=0}^{k-1}\sum_{r=0}^{k-1} d_i d_r (i+1)(r+1). \tag{25}$$

Then it suffices to prove, for all $i, r \in [0, k-1]$, that

$$\sum_{j=0}^{k-1} \min(j+1, i+1)\min(j+1, r+1) \geq \frac{k}{2}\min(i+1, r+1) + \frac{k}{3}(i+1)(r+1). \tag{26}$$

Then, multiplying it by the non-negative quantity $d_i d_r$ and summing over i, r proves Inequality 21. Note that Inequality 26 is symmetric in the variables i, r. Thus, without loss of generality, we may assume that $i \leq r$, giving

$$\begin{aligned}\sum_{j=0}^{k-1} \min(j+1, i+1)\min(j+1, r+1) &= \sum_{j=0}^{i}(j+1)^2 + \sum_{j=i+1}^{r}(j+1)(i+1) \\ &\quad + \sum_{j=r+1}^{k-1}(i+1)(r+1) \\ &= \frac{(i+1)(i+2)(2i+3)}{6} + \frac{(i+1)(r-i)(i+r+3)}{2} \\ &\quad + (i+1)(r+1)(k-1-r) \\ &= (i+1)\left[k(r+1) - \frac{r(r+1)}{2} - \frac{i(i+2)}{6}\right].\end{aligned} \tag{27}$$

Then, dividing both sides of Inequality 26 by the positive quantity $i+1$, we obtain the following equivalent inequality:

$$k(r+1) - \frac{r(r+1)}{2} - \frac{i(i+2)}{6} \geq \frac{k}{2} + \frac{k}{3}(r+1). \tag{28}$$

Subtracting the right-hand side, it suffices to show that

$$\frac{k}{6} + \frac{2kr}{3} - \frac{r(r+1)}{2} - \frac{i(i+2)}{6} \geq 0. \tag{29}$$

Now we use the assumption that $i \leq r$:

$$\frac{k}{6} + \frac{2kr}{3} - \frac{r(r+1)}{2} - \frac{i(i+2)}{6} \geq \frac{k(1+4r)}{6} - \frac{r(4r+5)}{6} = \frac{(k-1-r)(1+4r)+1}{6} \geq 0, \tag{30}$$

where in the last inequality we used that $0 \leq r \leq k-1$, concluding the proof. □

C.2 Coefficient Bounds

In this section we give bounds on coefficient sequences closely related to the γ-BIFR factorization. In particular, we recall from Definition 2 that $C_\gamma^{(p)}$ is defined through its inverse where

$$\left(C_\gamma^{(p)}\right)^{-1} := \begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \\ (\tilde{c}_\gamma)_1 & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ (\tilde{c}_\gamma)_{p-1} & (\tilde{c}_\gamma)_{p-2} & \cdots & 1 & \cdots & 0 \\ 0 & (\tilde{c}_\gamma)_{p-1} & \cdots & (\tilde{c}_\gamma)_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (\tilde{c}_\gamma)_{p-1} & \cdots & 1 \end{pmatrix}, \qquad \text{for} \quad (\tilde{c}_\gamma)_i := (-1)^i \binom{\gamma}{i},$$

and $\binom{\gamma}{i} = \prod_{j=1}^{i} \frac{\gamma+1-j}{j}$ is a generalized binomial coefficient. For the corresponding subdiagonals of the Toeplitz matrix $C_\gamma^{(p)}$ we use the notation $((c_\gamma^{(p)})_j)_{j\geq 0}$, which we bound in Lemma 7. Additionally, observe that $B_\gamma^{(p)} C_\gamma^{(p)} = E$, and so $B_\gamma^{(p)}$ equals $E(C_\gamma^{(p)})^{-1}$. Since E is the prefix-sum matrix, the corresponding entries of $B_\gamma^{(p)}$ are prefix sums on $((\tilde{c}_\gamma)_j)_{j\geq 0}$, and we bound these in Corollary 1.

In what follows Lemma 4 and 5 produce bounds on the sequences $((c_\gamma)_j)_{j\geq 0}$ and $((\tilde{c}_\gamma)_j)_{j\geq 0}$; Corollary 1 bounds prefix sums on $((\tilde{c}_\gamma)_j)_{j\geq 0}$ and Lemma 6 and 7 bound $((c_\gamma^{(p)})_j)_{j\geq 0}$.

Before we get going, however, we first need two brief helper lemmas.

Lemma 2. *Let $s > 0$, $m \geq 1$ an integer and define $H_m^{(s)} := \sum_{r=1}^{m} r^{-s}$. Then*

$$H_m^{(s)} \leq 1 + \max\{1, m^{1-s}\} \cdot \min\Big\{\log m,\ \frac{1}{|1-s|}\Big\}.$$

Moreover, if $s < 1$, then $H_m^{(s)} \leq m^{1-s}/(1-s)$.

Proof. Since the summands are non-increasing,

$$H_m^{(s)} \leq 1 + I, \qquad I := \int_1^m x^{-s}\, dx = \int_0^{\log m} e^{(1-s)t}\, dt.$$

The integrand is at most $\max\{1, m^{1-s}\}$, so $I \leq \max\{1, m^{1-s}\} \log m$. For $s \neq 1$, $I = (m^{1-s} - 1)/(1-s)$. If $s > 1$, this gives $I = (1 - m^{1-s})/(s-1) \leq 1/(s-1)$. If $s < 1$, it gives $I \leq m^{1-s}/(1-s)$. This proves the first bound. For the second statement, if $s < 1$ then

$$1 + I = \frac{m^{1-s} - s}{1-s} \leq \frac{m^{1-s}}{1-s}. \qquad \square$$

Lemma 3 (Gamma facts)**.** *Let $z > 0$ and Γ denote the Gamma function. The following facts hold:*

1. $\Gamma(z+1) = z\Gamma(z)$.
2. *If $z \in (0,2)$, then $\Gamma(z) \geq 1/2$.*
3. *If $z \in (1,2)$, then $\Gamma(z) \leq 1$.*

Lemma 4 (Bounds for $(c_\gamma)_j$)**.** *Let $0 < \gamma < 1$ and define $c_\gamma = \big((c_\gamma)_j\big)_{j\geq 0}$ by*

$$(1-x)^{-\gamma} = \sum_{j=0}^{\infty} (c_\gamma)_j\, x^j, \qquad |x| < 1. \tag{31}$$

Then, for every $j \geq 0$,

$$(c_\gamma)_j = \binom{\gamma + j - 1}{j} = \frac{\Gamma(\gamma + j)}{\Gamma(\gamma)\,\Gamma(j+1)}. \tag{32}$$

Moreover, for every integer $j \geq 1$,

$$\frac{1}{\Gamma(\gamma)\,(j+1)^{1-\gamma}} \ <\ (c_\gamma)_j \ <\ \frac{1}{\Gamma(\gamma)\, j^{1-\gamma}}. \tag{33}$$

Proof. The binomial series gives

$$(c_\gamma)_j = (-1)^j \binom{-\gamma}{j} = \binom{\gamma + j - 1}{j} = \frac{\Gamma(\gamma + j)}{\Gamma(\gamma)\,\Gamma(j+1)}. \tag{34}$$

For the bounds, rewrite

$$(c_\gamma)_j = \frac{1}{\Gamma(\gamma)}\, \frac{\Gamma(j+\gamma)}{\Gamma(j+1)}. \tag{35}$$

By Gautschi's inequality (applied with $x = j$ and $s = \gamma \in (0,1)$),

$$(j+1)^{\gamma-1} < \frac{\Gamma(j+\gamma)}{\Gamma(j+1)} < j^{\gamma-1}. \tag{36}$$

Multiplying by $1/\Gamma(\gamma)$ yields Equation (33). $\square$

Lemma 5 (Bounds for $(\tilde{c}_\gamma)_j$)**.** *Let $0 < \gamma < 1$ and write*

$$(1-x)^\gamma = \sum_{j=0}^{\infty} (\tilde{c}_\gamma)_j \, x^j, \qquad |x| < 1. \tag{37}$$

Then $(\tilde{c}_\gamma)_0 = 1$ and, for every $j \geq 1$,

$$(\tilde{c}_\gamma)_j = (-1)^j \binom{\gamma}{j} = -\frac{\gamma\,\Gamma(j-\gamma)}{\Gamma(1-\gamma)\,\Gamma(j+1)}. \tag{38}$$

Moreover, for every $j \geq 2$,

$$\frac{\gamma}{\Gamma(1-\gamma)}\,\frac{1}{j^{\gamma+1}} \;<\; |(\tilde{c}_\gamma)_j| \;<\; \frac{\gamma}{\Gamma(1-\gamma)}\,\frac{1}{j\,(j-1)^\gamma}. \tag{39}$$

Proof. The binomial series for $(1-x)^\gamma$ gives

$$(1-x)^\gamma = \sum_{j=0}^{\infty} (-1)^j \binom{\gamma}{j} x^j, \qquad |x| < 1, \tag{40}$$

so $(\tilde{c}_\gamma)_0 = 1$ and $(\tilde{c}_\gamma)_j = (-1)^j \binom{\gamma}{j}$ for $j \geq 1$.

For $j \geq 1$, start from the binomial-coefficient definition

$$\binom{\gamma}{j} = \frac{\gamma(\gamma-1)\cdots(\gamma-j+1)}{j!}. \tag{41}$$

Factoring out $-\gamma$ and rewriting the remaining product via Gamma functions gives

$$(-1)^j \binom{\gamma}{j} = -\frac{\gamma(1-\gamma)(2-\gamma)\cdots(j-1-\gamma)}{j!} = -\frac{\gamma}{\Gamma(j+1)}\,\frac{\Gamma(j-\gamma)}{\Gamma(1-\gamma)} = -\frac{\gamma\,\Gamma(j-\gamma)}{\Gamma(1-\gamma)\,\Gamma(j+1)}. \tag{42}$$

This yields the claimed closed form for $(\tilde{c}_\gamma)_j$.

For $j \geq 2$, rewrite

$$|(\tilde{c}_\gamma)_j| = \frac{\gamma}{\Gamma(1-\gamma)}\,\frac{\Gamma(j-\gamma)}{\Gamma(j+1)} = \frac{\gamma}{\Gamma(1-\gamma)}\,\frac{1}{j}\,\frac{\Gamma(j-\gamma)}{\Gamma(j)}. \tag{43}$$

Apply Gautschi's inequality to $\Gamma(j-\gamma)/\Gamma(j)$ with $x = j-1$ and $s = 1-\gamma \in (0,1)$:

$$\frac{\gamma}{\Gamma(1-\gamma)}\,\frac{1}{j^{\gamma+1}} \;<\; |(\tilde{c}_\gamma)_j| \;<\; \frac{\gamma}{\Gamma(1-\gamma)}\,\frac{1}{j\,(j-1)^\gamma}, \tag{44}$$

which is exactly the stated two-sided bound. □

Corollary 1 (Prefix sums of $\tilde{c}_\gamma$)**.** *Let $0 < \gamma < 1$ and let $\tilde{c}_\gamma = \big((\tilde{c}_\gamma)_j\big)_{j\geq 0}$ be defined by $(1-x)^\gamma = \sum_{j=0}^{\infty} (\tilde{c}_\gamma)_j x^j$ for $|x| < 1$. Define the prefix sums*

$$s_{\gamma,j} := \sum_{i=0}^{j} (\tilde{c}_\gamma)_i \qquad (j \geq 0). \tag{45}$$

Then $s_{\gamma,0} = 1$ and for every $j \geq 0$,

$$s_{\gamma,j} = (c_{1-\gamma})_j. \tag{46}$$

In particular, for every integer $j \geq 1$,

$$\frac{1}{\Gamma(1-\gamma)\,(j+1)^\gamma} \;<\; \sum_{i=0}^{j} (\tilde{c}_\gamma)_i \;<\; \frac{1}{\Gamma(1-\gamma)\,j^\gamma}. \tag{47}$$

Proof. Let $S_\gamma(x) := \sum_{j=0}^{\infty} s_{\gamma,j} x^j$ be the generating function of the prefix sums. By definition of $s_{\gamma,j}$,

$$S_\gamma(x) = \sum_{j=0}^{\infty} \left(\sum_{i=0}^{j} (\tilde{c}_\gamma)_i \right) x^j = \left(\sum_{i=0}^{\infty} (\tilde{c}_\gamma)_i x^i \right) \left(\sum_{k=0}^{\infty} x^k \right), \tag{48}$$

so for $|x| < 1$,

$$S_\gamma(x) = (1-x)^\gamma \cdot (1-x)^{-1} = (1-x)^{-1+\gamma} = (1-x)^{-(1-\gamma)}. \tag{49}$$

Thus $s_{\gamma,j}$ is exactly the j-th coefficient in the expansion of $(1-x)^{-(1-\gamma)}$, i.e. $s_{\gamma,j} = (c_{1-\gamma})_j$, proving Equation (46). The bounds Equation (47) follow from Lemma 4 applied with γ replaced by $1-\gamma$. □

Bounding $((c_\gamma^{(p)})_j)_{j\geq 0}$ is more complicated. As a first step, we prove that the entries are monotonically decreasing.

Lemma 6 (Monotonicity of $c_\gamma^{(p)}$)**.** *Let $0 < \gamma < 1$ and let the coefficients $c_\gamma^{(p)} = \left((c_\gamma^{(p)})_j\right)_{j\geq 0}$ denote the diagonal values of $C_\gamma^{(p)}$ from Definition 2. Then*

$$0 \leq (c_\gamma^{(p)})_{j+1} \leq (c_\gamma^{(p)})_j \qquad (j \geq 0). \tag{50}$$

Proof. The coefficients satisfy the recursion

$$(c_\gamma^{(p)})_j = - \sum_{k=1}^{\min\{p-1,j\}} (\tilde{c}_\gamma)_k \, (c_\gamma^{(p)})_{j-k}. \tag{51}$$

Since $(\tilde{c}_\gamma)_k \leq 0$ for $k \geq 1$, the right-hand side is a nonnegative linear combination of earlier coefficients. With $(c_\gamma^{(p)})_0 = 1$, it follows by induction that $(c_\gamma^{(p)})_j \geq 0$ for all $j \geq 0$.

For monotonicity, define

$$\Delta_j := (c_\gamma^{(p)})_j - (c_\gamma^{(p)})_{j+1} \qquad (j \geq 0). \tag{52}$$

If $p = 1$, then the sum in Equation (51) is empty and $(c_\gamma^{(p)})_j = 0$ for $j \geq 1$, so monotonicity is immediate. Assume $p \geq 2$. Using Equation (51) with $j = 1$ gives $(c_\gamma^{(p)})_1 = -(\tilde{c}_\gamma)_1 (c_\gamma^{(p)})_0 = \gamma$, hence $\Delta_0 = 1 - \gamma > 0$.

Now apply Equation (51) at indices $j+1$ and $j+2$ and subtract:

$$\Delta_{j+1} = \sum_{k=1}^{\min\{p-1,j+1\}} \left(-(\tilde{c}_\gamma)_k\right) \Delta_{j+1-k} \; + \begin{cases} (\tilde{c}_\gamma)_{j+2}, & j+2 \leq p-1, \\ 0, & j+2 > p-1. \end{cases} \tag{53}$$

Since $-(\tilde{c}_\gamma)_k \geq 0$ for all $k \geq 1$, the right-hand side is a sum of nonnegative terms whenever $\Delta_0, \ldots, \Delta_j \geq 0$ *and* $j+2 > p-1$. When $j+2 \leq p-1$, the right-hand side is however not obviously positive. For this case, we instead write

$$\Delta_{j+1} = (c_\gamma^{(p)})_{j+1} - (c_\gamma^{(p)})_{j+2} = (c_\gamma)_{j+1} - (c_\gamma)_{j+2} \tag{54}$$

$$> \frac{1}{\Gamma(\gamma)(j+2)^{1-\gamma}} - \frac{1}{\Gamma(\gamma)(j+2)^{1-\gamma}} = 0, \tag{55}$$

where we have used $(c_\gamma^{(p)})_\ell$ equals $(c_\gamma)_\ell$ for $\ell \leq p-1$, and the last step bounds on $(c_\gamma)_\ell$ are from Equation (33).

Therefore, by induction, $\Delta_j \geq 0$ for all $j \geq 0$, i.e., $(c_\gamma^{(p)})_{j+1} \leq (c_\gamma^{(p)})_j$. □

Our bound on $((c_\gamma^{(p)})_j)_{j\geq 0}$ is based on showing a geometric decay, with rate parametrized in p and γ, for coefficients with index greater than the bandwidth p. It is stated and proved next.

Lemma 7 (Geometric tail bound for $c_\gamma^{(p)}$)**.** *Let $0 < \gamma < 1$ and $p \geq 2$, and define*

$$\beta_{\gamma,p} := \min\left\{ \frac{1-\gamma}{2p} \log\Big(\frac{2p}{p+1}\Big),\ \frac{1-\gamma}{8\,p^\gamma\,(p-1)^{1-\gamma}} \right\}. \tag{56}$$

Then, for every $j \geq p$,

$$(c_\gamma^{(p)})_j \leq (c_\gamma)_p \left(1 - \beta_{\gamma,p}\right)^{j-p}. \tag{57}$$

Moreover, $\beta_{\gamma,p} = \Theta((1-\gamma)/p)$ and in particular satisfies the two-sided bound

$$\frac{1-\gamma}{8p} \leq \beta_{\gamma,p} \leq \frac{1-\gamma}{4p}. \tag{58}$$

Proof. We first establish the two-sided bound Equation (58) on $\beta_{\gamma,p}$. Denote the two terms in Equation (56) by

$$T_1 := \frac{1-\gamma}{2p} \log\Big(\frac{2p}{p+1}\Big), \qquad T_2 := \frac{1-\gamma}{8p^\gamma (p-1)^{1-\gamma}}. \tag{59}$$

For T_1: since $p \mapsto 2p/(p+1) = 2 - 2/(p+1)$ is increasing and equals $4/3$ at $p = 2$, $\log(2p/(p+1)) \geq \log(4/3) > 1/4$ for $p \geq 2$, hence $T_1 \geq (1-\gamma)/(8p)$. For T_2: rewrite

$$T_2 = \frac{1-\gamma}{8p(1-1/p)^{1-\gamma}}. \tag{60}$$

Since the base $(1-1/p) \in (0,1)$ and the exponent $(1-\gamma) \in (0,1)$, we have the elementary bounds

$$\frac{p-1}{p} = (1-1/p)^1 \leq (1-1/p)^{1-\gamma} \leq 1. \tag{61}$$

The upper bound on $(1-1/p)^{1-\gamma}$ gives $T_2 \geq (1-\gamma)/(8p)$, and the lower bound gives $T_2 \leq (1-\gamma)/(8(p-1)) \leq (1-\gamma)/(4p)$ for $p \geq 2$. Combining,

$$\frac{1-\gamma}{8p} \leq \min\{T_1, T_2\} = \beta_{\gamma,p} \leq T_2 \leq \frac{1-\gamma}{4p}, \tag{62}$$

proving Equation (58). In particular, $\beta_{\gamma,p} \leq 1/(4p) \leq 1/8$.

We proceed to prove the main statement. Recall the recursion from Equation (51), restated next:

$$(c_\gamma^{(p)})_j = - \sum_{k=1}^{\min\{p-1,j\}} (\tilde{c}_\gamma)_k\, (c_\gamma^{(p)})_{j-k} \qquad (j \geq 0). \tag{63}$$

Since $(\tilde{c}_\gamma)_k \leq 0$ for $k \geq 1$ and $(c_\gamma^{(p)})_0 = 1$, Equation (51) implies $(c_\gamma^{(p)})_j \geq 0$ for all j by induction.

Step 1. $(c_\gamma^{(p)})_j \leq (c_\gamma)_j$ for all $j \geq 0$. The exact coefficients satisfy, for $j \geq 1$,

$$(c_\gamma)_j = -\sum_{k=1}^{j} (\tilde{c}_\gamma)_k\, (c_\gamma)_{j-k}. \tag{64}$$

For $0 \leq j \leq p-1$ the two recursions coincide, hence $(c_\gamma^{(p)})_j = (c_\gamma)_j$. Fix $j \geq p$ and assume $(c_\gamma^{(p)})_m \leq (c_\gamma)_m$ for all $m < j$. Then, using $-(\tilde{c}_\gamma)_k \geq 0$ and $(c_\gamma)_m \geq 0$,

$$(c_\gamma^{(p)})_j = \sum_{k=1}^{p-1} \big(-(\tilde{c}_\gamma)_k\big)(c_\gamma^{(p)})_{j-k} \leq \sum_{k=1}^{p-1} \big(-(\tilde{c}_\gamma)_k\big)(c_\gamma)_{j-k} \leq \sum_{k=1}^{j} \big(-(\tilde{c}_\gamma)_k\big)(c_\gamma)_{j-k} = (c_\gamma)_j. \tag{65}$$

Step 2. Base range $p \leq j \leq 2p$. The ratio

$$\frac{(c_\gamma)_{j+1}}{(c_\gamma)_j} = \frac{j+\gamma}{j+1} \tag{66}$$

is increasing in j, hence $(c_\gamma)_j$ is log-convex. Therefore, for $r = 0, 1, \ldots, p$,

$$(c_\gamma)_{p+r} \le (c_\gamma)_p^{1-r/p}\,(c_\gamma)_{2p}^{r/p}. \tag{67}$$

Thus it suffices to prove the anchor inequality

$$(c_\gamma)_{2p} \le (c_\gamma)_p\,(1-\beta_{\gamma,p})^p, \tag{68}$$

because inserting Equation (68) into Equation (67) yields

$$(c_\gamma)_{p+r} \le (c_\gamma)_p^{1-r/p}\Big((c_\gamma)_p(1-\beta_{\gamma,p})^p\Big)^{r/p} = (c_\gamma)_p(1-\beta_{\gamma,p})^r, \tag{69}$$

which is exactly the desired bound for $p \le p+r \le 2p$.

To prove Equation (68), note from Lemma 4 that

$$(c_\gamma)_{2p} < \frac{1}{\Gamma(\gamma)\,(2p)^{1-\gamma}}, \qquad (c_\gamma)_p > \frac{1}{\Gamma(\gamma)\,(p+1)^{1-\gamma}},$$

hence

$$\frac{(c_\gamma)_{2p}}{(c_\gamma)_p} < \left(\frac{p+1}{2p}\right)^{1-\gamma}. \tag{70}$$

Next, since $\beta_{\gamma,p} \le \frac{1}{2}$, the inequality $\log(1-x) \ge -2x$ for $x \in [0, \frac{1}{2}]$ implies

$$(1-\beta_{\gamma,p})^p = \exp\big(p\log(1-\beta_{\gamma,p})\big) \ge \exp(-2p\beta_{\gamma,p}). \tag{71}$$

Because $\beta_{\gamma,p} \le \frac{1-\gamma}{2p}\log\big(\frac{2p}{p+1}\big)$, we obtain

$$\exp(-2p\beta_{\gamma,p}) \ge \exp\left(-(1-\gamma)\log\Big(\frac{2p}{p+1}\Big)\right) = \left(\frac{p+1}{2p}\right)^{1-\gamma}. \tag{72}$$

Combining with Equation (70) yields

$$(1-\beta_{\gamma,p})^p \ge \frac{(c_\gamma)_{2p}}{(c_\gamma)_p}, \tag{73}$$

which is exactly Equation (68). Consequently, for $r = 0, 1, \ldots, p$,

$$(c_\gamma)_{p+r} \le (c_\gamma)_p(1-\beta_{\gamma,p})^r. \tag{74}$$

Together with Step 1, this gives for all $p \le j \le 2p$,

$$(c_\gamma^{(p)})_j \le (c_\gamma)_j \le (c_\gamma)_p(1-\beta_{\gamma,p})^{j-p}. \tag{75}$$

Step 3. Induction for all $j \ge 2p+1$. Assume that for some $j \ge 2p+1$,

$$(c_\gamma^{(p)})_m \le (c_\gamma)_p(1-\beta_{\gamma,p})^{\,m-p} \qquad \text{for all } m = p, p+1, \ldots, j-1. \tag{76}$$

Then $j-k \ge p$ for $k = 1, \ldots, p-1$, and Equation (51) gives

$$(c_\gamma^{(p)})_j = \sum_{k=1}^{p-1}\big(-(\tilde{c}_\gamma)_k\big)(c_\gamma^{(p)})_{j-k} \le (c_\gamma)_p(1-\beta_{\gamma,p})^{\,j-p}\sum_{k=1}^{p-1}\big(-(\tilde{c}_\gamma)_k\big)(1-\beta_{\gamma,p})^{-k}. \tag{77}$$

To finish the inductive step, and thereby the full proof, we will show that

$$\sum_{k=1}^{p-1}\big(-(\tilde{c}_\gamma)_k\big)(1-\beta_{\gamma,p})^{-k} \le 1. \tag{78}$$

To this end, write $(1-\beta_{\gamma,p})^{-k} = \exp(-k\log(1-\beta_{\gamma,p}))$. Firstly, $-\log(1-x) \le 2x$ on $[0, \frac{1}{2}]$, and as $\beta_{\gamma,p} \le \frac{1}{4p} \le \frac{1}{2}$, we note that

$$-k\log(1-\beta_{\gamma,p}) \le -(p-1)\log(1-\beta_{\gamma,p}) \le 2(p-1)\beta_{\gamma,p} \le \frac{2(p-1)}{4p} \le \frac{1}{2}. \tag{79}$$

where $k \le p-1$ was used at the first step. Secondly, since for $x \in [0,1]$ we have the standard inequality $e^x \le 1+2x$, we write

$$\exp(-k\log(1-\beta_{\gamma,p})) \le 1 + 2(-k\log(1-\beta_{\gamma,p})) \le 1 + 4k\beta_{\gamma,p}. \tag{80}$$

Substituting this bound into Equation (78):

$$\sum_{k=1}^{p-1}\big(-(\tilde{c}_\gamma)_k\big)(1-\beta_{\gamma,p})^{-k} \le \underbrace{\sum_{k=1}^{p-1}\big(-(\tilde{c}_\gamma)_k\big)}_{1-(c_{1-\gamma})_{p-1}} + 4\beta_{\gamma,p}\sum_{k=1}^{p-1} k\big(-(\tilde{c}_\gamma)_k\big). \tag{81}$$

We proceed to bound the second sum using Lemma 5: $-(\tilde{c}_\gamma)_1 = \gamma$, and for $k \ge 2$,

$$k(-\tilde{c}_\gamma)_k < \frac{\gamma}{\Gamma(1-\gamma)(k-1)^\gamma}. \tag{82}$$

By Lemma 2,

$$\sum_{k=2}^{p-1} k(-(\tilde{c}_\gamma)_k) < \frac{\gamma}{\Gamma(1-\gamma)}\sum_{j=1}^{p-2} j^{-\gamma} \le \frac{\gamma(p-1)^{1-\gamma}}{(1-\gamma)\Gamma(1-\gamma)}, \tag{83}$$

hence

$$\sum_{k=1}^{p-1} k(-(\tilde{c}_\gamma)_k) \le \gamma + \frac{\gamma(p-1)^{1-\gamma}}{(1-\gamma)\Gamma(1-\gamma)} \le \frac{2\gamma(p-1)^{1-\gamma}}{(1-\gamma)\Gamma(1-\gamma)}, \tag{84}$$

where the last step absorbs the γ term: by Lemma 3, $(1-\gamma)\Gamma(1-\gamma) = \Gamma(2-\gamma) \le 1$ for $\gamma \in (0,1)$, so $\gamma \le \gamma(p-1)^{1-\gamma}/((1-\gamma)\Gamma(1-\gamma))$ for $p \ge 2$.

Plugging Equation (84) into Equation (81)

$$\sum_{k=1}^{p-1}(-(\tilde{c}_\gamma)_k)(1-\beta_{\gamma,p})^{-k} \le 1 - (c_{1-\gamma})_{p-1} + \frac{8\gamma\beta_{\gamma,p}(p-1)^{1-\gamma}}{(1-\gamma)\Gamma(1-\gamma)}. \tag{85}$$

By Lemma 4, $(c_{1-\gamma})_{p-1} \ge 1/(\Gamma(1-\gamma)p^\gamma)$. Taking the second branch of the minimum in Equation (56), we get that

$$\frac{8\beta_{\gamma,p}\gamma(p-1)^{1-\gamma}}{(1-\gamma)\Gamma(1-\gamma)} \le \frac{\gamma}{\Gamma(1-\gamma)p^\gamma} \le \frac{1}{\Gamma(1-\gamma)p^\gamma} \le (c_{1-\gamma})_{p-1}. \tag{86}$$

It follows that the right-hand side of Equation (85) is at most 1, establishing Equation (78) and hence the inductive step. The lemma follows. □

C.3 Matrix Norm Bounds

In this section we derive matrix norm bounds for the factors in the γ-BIFR factorization $B_\gamma^{(p)} C_\gamma^{(p)} = E$. We begin with a bound on the Frobenius norm of $B_\gamma^{(p)}$.

Lemma 8 (Frobenius norm bound for $B_\gamma^{(p)}$)**.** *Let $0 < \gamma < 1$ and $n \ge p \ge 2$, and let $B_\gamma^{(p)}$ be as in Definition 2. Then*

$$\|B_\gamma^{(p)}\|_F^2 \;\le\; n + \frac{n}{\Gamma(1-\gamma)^2}\,H_{p-1}^{(2\gamma)} + \frac{(n-p)(n-p+1)}{2\,\Gamma(1-\gamma)^2}\cdot\frac{1}{(p-1)^{2\gamma}}, \tag{87}$$

where $H_m^{(s)} := \sum_{r=1}^{m} r^{-s}$. In particular, the following asymptotic bound holds for $\gamma \in [1/2, 1)$:

$$\frac{1}{\sqrt{n}}\|B_\gamma^{(p)}\|_F = O\big(1 + (1-\gamma)\sqrt{\Lambda_\gamma} + (1-\gamma)\sqrt{n}\,p^{-\gamma}\big), \tag{88}$$

where $\Lambda_\gamma = \min\{\log(ep), 1/|2\gamma-1|\} \ge 1$.

Proof. By construction, $B_\gamma^{(p)} = E(C_\gamma^{(p)})^{-1}$ is a lower-triangular Toeplitz matrix. Writing $r = i - j \in \{0, 1, \ldots, n-1\}$, the Toeplitz coefficient on the r-th subdiagonal equals the partial sum

$$\sum_{k=0}^{\min\{p-1,r\}} (\tilde{c}_\gamma)_k, \tag{89}$$

and this subdiagonal contains $n - r$ entries. Therefore,

$$\|B_\gamma^{(p)}\|_F^2 = \sum_{r=0}^{n-1}(n-r)\left(\sum_{k=0}^{\min\{p-1,r\}}(\tilde{c}_\gamma)_k\right)^2. \tag{90}$$

For $r = 0$ the inner sum equals $(\tilde{c}_\gamma)_0 = 1$, giving the contribution n. For $1 \le r \le p-1$, Corollary 1 yields

$$\sum_{k=0}^{r}(\tilde{c}_\gamma)_k = (c_{1-\gamma})_r, \tag{91}$$

and by Equation (47) we have

$$0 < \sum_{k=0}^{r}(\tilde{c}_\gamma)_k < \frac{1}{\Gamma(1-\gamma)\, r^\gamma}. \tag{92}$$

For $r \ge p$ the inner sum is constant and equals $\sum_{k=0}^{p-1}(\tilde{c}_\gamma)_k$, which is bounded by $\left(\Gamma(1-\gamma)\,(p-1)^\gamma\right)^{-1}$ in the same way. Hence

$$\|B_\gamma^{(p)}\|_F^2 \le n + \frac{1}{\Gamma(1-\gamma)^2}\sum_{r=1}^{p-1}\frac{n-r}{r^{2\gamma}} + \frac{1}{\Gamma(1-\gamma)^2}\left(\sum_{r=p}^{n-1}(n-r)\right)\frac{1}{(p-1)^{2\gamma}}. \tag{93}$$

Using $\sum_{r=p}^{n-1}(n-r) = \sum_{m=1}^{n-p} m = \frac{(n-p)(n-p+1)}{2}$ gives

$$\|B_\gamma^{(p)}\|_F^2 \le n + \frac{1}{\Gamma(1-\gamma)^2}\sum_{r=1}^{p-1}\frac{n-r}{r^{2\gamma}} + \frac{(n-p)(n-p+1)}{2\,\Gamma(1-\gamma)^2}\cdot\frac{1}{(p-1)^{2\gamma}}. \tag{94}$$

Finally, since $n - r \le n$ for $1 \le r \le p-1$,

$$\sum_{r=1}^{p-1}\frac{n-r}{r^{2\gamma}} \le n\sum_{r=1}^{p-1}\frac{1}{r^{2\gamma}} = n\, H_{p-1}^{(2\gamma)}. \tag{95}$$

Substituting this bound completes the proof of Equation (87).

Asymptotic bounds. To derive the asymptotic bound, we first rewrite Equation (87). By the definition of the Gamma function, $\Gamma(1-\gamma) = \Gamma(2-\gamma)/(1-\gamma)$, and that $\Gamma(z) \ge 1/2$ for $z \in (1,2)$ (Lemma 3), we have that $1/\Gamma(1-\gamma)^2 \le 4(1-\gamma)^2$. Secondly,

$$\frac{(n-p)(n-p+1)}{(p-1)^{2\gamma}} \le \frac{n^2}{(p/2)^{2\gamma}} \le 4n^2p^{-2\gamma} \tag{96}$$

where the first step used $p - 1 \ge p/2$ for $p \ge 2$ and the last step that $2^{2\gamma} \le 2^2 = 4$. Combining these observations, we arrive at the coarser bound:

$$\|B_\gamma^{(p)}\|_F^2 \le n + 4(1-\gamma)^2 n H_{p-1}^{(2\gamma)} + 16(1-\gamma)^2 n^2 p^{-2\gamma}\,. \tag{97}$$

Now, using Lemma 2 when $\gamma \in [1/2, 1)$, we get that

$$H_{p-1}^{(2\gamma)} \le 1 + \min\{\log(p-1), 1/(2\gamma-1)\} \le 1 + \Lambda_\gamma,$$

and so, bounding the right-hand side of Equation (97):

$$\begin{aligned}
\|B_\gamma^{(p)}\|_F^2 &\le n + 4(1-\gamma)^2(1+\Lambda_\gamma)n + 16(1-\gamma)^2 n^2 p^{-2\gamma} && (98)\\
&= n\left(1 + 4(1-\gamma)^2(1+\Lambda_\gamma) + 16(1-\gamma)^2 np^{-2\gamma}\right) && (99)\\
&= O\left(n\left(1 + (1-\gamma)^2\Lambda_\gamma + (1-\gamma)^2 np^{-2\gamma}\right)\right). && (100)
\end{aligned}$$

To get the final statement, take the square root for each bound, use that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for positive a, b, and divide by $\sqrt{n}$. □

We proceed to use our geometric tail bound on $((c_\gamma^{(p)})_j)_{j\geq 0}$ to derive a bound on $\|C_\gamma^{(p)}\|_{1\to 2}$.

Lemma 9 (Bound on $1 \to 2$ norm of $C_\gamma^{(p)}$)**.** *Let $0 < \gamma < 1$ and $n \geq p \geq 2$, and let $C_\gamma^{(p)}$ be as in Definition 2. Additionally let $\beta_{\gamma,p}$ be as in Lemma 7. Then*

$$\|C_\gamma^{(p)}\|_{1\to 2}^2 \;\leq\; 1 + \frac{1}{\Gamma(\gamma)^2}\, H_{p-1}^{(2-2\gamma)} + \frac{1}{\Gamma(\gamma)^2\, p^{2-2\gamma}} \cdot \frac{1}{\beta_{\gamma,p}\,(2-\beta_{\gamma,p})}, \tag{101}$$

where $H_m^{(s)} := \sum_{j=1}^m j^{-s}$. In particular, the following asymptotic bound holds for $\gamma \in [1/2, 1)$:

$$\|C_\gamma^{(p)}\|_{1\to 2} = O\big(p^{\gamma-\frac{1}{2}}(\sqrt{\Lambda_\gamma} + 1/\sqrt{1-\gamma})\big), \tag{102}$$

where $\Lambda_\gamma = \min\{\log(ep), 1/|2\gamma - 1|\} \geq 1$.

Proof. Since $C_\gamma^{(p)}$ is lower-triangular Toeplitz and $(c_\gamma^{(p)})_j \geq 0$, the column ℓ_2-norms are nonincreasing from left to right; hence

$$\|C_\gamma^{(p)}\|_{1\to 2}^2 = \max_{1\leq j\leq n} \|C_\gamma^{(p)} e_j\|_2^2 = \|C_\gamma^{(p)} e_1\|_2^2 = \sum_{r=0}^{n-1} (c_\gamma^{(p)})_r^2. \tag{103}$$

Split the sum at p:

$$\sum_{r=0}^{n-1} (c_\gamma^{(p)})_r^2 = \sum_{r=0}^{p-1} (c_\gamma^{(p)})_r^2 + \sum_{r=p}^{n-1} (c_\gamma^{(p)})_r^2. \tag{104}$$

For $0 \leq r \leq p-1$ we have $(c_\gamma^{(p)})_r = (c_\gamma)_r$, and by Lemma 4, $(c_\gamma)_0 = 1$ and $(c_\gamma)_r \leq (\Gamma(\gamma) r^{1-\gamma})^{-1}$ for $r \geq 1$, hence

$$\sum_{r=0}^{p-1} (c_\gamma^{(p)})_r^2 \leq 1 + \frac{1}{\Gamma(\gamma)^2} \sum_{r=1}^{p-1} \frac{1}{r^{2-2\gamma}} = 1 + \frac{1}{\Gamma(\gamma)^2}\, H_{p-1}^{(2-2\gamma)}. \tag{105}$$

For $r \geq p$, Lemma 7 gives

$$(c_\gamma^{(p)})_r \leq (c_\gamma)_p (1 - \beta_{\gamma,p})^{\,r-p}, \tag{106}$$

so using $(c_\gamma)_p \leq (\Gamma(\gamma) p^{1-\gamma})^{-1}$ from Lemma 4,

$$\sum_{r=p}^{n-1} (c_\gamma^{(p)})_r^2 \leq (c_\gamma)_p^2 \sum_{m=0}^{\infty} (1-\beta_{\gamma,p})^{2m} = \frac{(c_\gamma)_p^2}{1-(1-\beta_{\gamma,p})^2} \leq \frac{1}{\Gamma(\gamma)^2\, p^{2-2\gamma}} \cdot \frac{1}{\beta_{\gamma,p}(2-\beta_{\gamma,p})}. \tag{107}$$

Combining the two parts yields the bound in Equation (101).

Asymptotic bound. For the asymptotic statement, note that for $\gamma \in (0,1)$, $\Gamma(\gamma) \geq 1/2$. Additionally using that $\frac{1-\gamma}{8p} \leq \beta_{\gamma,p} \leq \frac{1-\gamma}{4p} \leq 1$, we can arrive at a coarser version of Equation (101)

$$\|C_\gamma^{(p)}\|_{1\to 2}^2 \leq 1 + 4 H_{p-1}^{(2-2\gamma)} + \frac{32 p^{2\gamma-1}}{1-\gamma}. \tag{108}$$

To arrive at Equation (102), we lastly invoke Lemma 2 under the assumption of $\gamma \in [1/2, 1)$ for

$$H_{p-1}^{(2-2\gamma)} \leq 1 + (p-1)^{2\gamma-1} \min\{\log(p-1), 1/(2\gamma-1)\} \leq 1 + p^{2\gamma-1}\Lambda_\gamma, \tag{109}$$

hence, inserting the result into Equation (108) and taking a square-root,

$$\|C_\gamma^{(p)}\|_{1\to 2} \leq \sqrt{5 + 4p^{2\gamma-1}\Lambda_\gamma + \frac{32p^{2\gamma-1}}{1-\gamma}} = O(p^{\gamma-1/2}(\sqrt{\Lambda_\gamma} + 1/\sqrt{1-\gamma})), \tag{110}$$

where the additive constant can be absorbed as $p^{2\gamma-1} \geq 1$ and $1/(1-\gamma) \geq 1$. This completes the proof. □

The bound on $\|C_\gamma^{(p)}\|_{1\to 1}$ is derived analogously to Lemma 9, and is given next.

Lemma 10 (Bound on $1 \to 1$ norm of $C_\gamma^{(p)}$)**.** *Let $0 < \gamma < 1$ and $n \geq p \geq 2$, and let $C_\gamma^{(p)}$ be as in Definition 2. Additionally let $\beta_{\gamma,p}$ be as in Lemma 7. Then*

$$\|C_\gamma^{(p)}\|_{1\to 1} \;\leq\; 1 + \frac{1}{\Gamma(\gamma)}\, H_{p-1}^{(1-\gamma)} + \frac{1}{\Gamma(\gamma)\, p^{1-\gamma}\beta_{\gamma,p}}, \tag{111}$$

where $H_m^{(s)} := \sum_{j=1}^{m} j^{-s}$. In particular, the following asymptotic bound holds for $\gamma \in (0,1)$:

$$\|C_\gamma^{(p)}\|_{1\to 1} = O\!\left(\frac{p^\gamma}{\gamma(1-\gamma)}\right). \tag{112}$$

Proof. Since $C_\gamma^{(p)}$ is lower-triangular Toeplitz and $(c_\gamma^{(p)})_j \geq 0$, the column ℓ_1-norms are nonincreasing from left to right; hence

$$\|C_\gamma^{(p)}\|_{1\to 1} = \max_{1\leq j\leq n} \|C_\gamma^{(p)} e_j\|_1 = \|C_\gamma^{(p)} e_1\|_1 = \sum_{r=0}^{n-1} (c_\gamma^{(p)})_r. \tag{113}$$

Split the sum at p:

$$\sum_{r=0}^{n-1} (c_\gamma^{(p)})_r = \sum_{r=0}^{p-1} (c_\gamma^{(p)})_r + \sum_{r=p}^{n-1} (c_\gamma^{(p)})_r. \tag{114}$$

For $0 \leq r \leq p-1$ we have $(c_\gamma^{(p)})_r = (c_\gamma)_r$, and by Lemma 4, $(c_\gamma)_0 = 1$ and $(c_\gamma)_r \leq (\Gamma(\gamma) r^{1-\gamma})^{-1}$ for $r \geq 1$, hence

$$\sum_{r=0}^{p-1} (c_\gamma^{(p)})_r \leq 1 + \frac{1}{\Gamma(\gamma)} \sum_{r=1}^{p-1} \frac{1}{r^{1-\gamma}} = 1 + \frac{1}{\Gamma(\gamma)}\, H_{p-1}^{(1-\gamma)}. \tag{115}$$

For $r \geq p$, Lemma 7 gives

$$(c_\gamma^{(p)})_r \leq (c_\gamma)_p (1 - \beta_{\gamma,p})^{\,r-p}, \tag{116}$$

so using $(c_\gamma)_p \leq (\Gamma(\gamma) p^{1-\gamma})^{-1}$ from Lemma 4,

$$\sum_{r=p}^{n-1} (c_\gamma^{(p)})_r \leq (c_\gamma)_p \sum_{m=0}^{\infty} (1-\beta_{\gamma,p})^m = \frac{(c_\gamma)_p}{1-(1-\beta_{\gamma,p})} \leq \frac{1}{\Gamma(\gamma)\, p^{1-\gamma}\, \beta_{\gamma,p}}. \tag{117}$$

Combining the two parts yields the bound in Equation (111).

Asymptotic bounds. For the asymptotic statement, note that for $\gamma \in (0,1)$, $\Gamma(\gamma) \geq 1/2$. Additionally using that $\beta_{\gamma,p} \geq \frac{1-\gamma}{8p}$, we can simplify Equation (111) further

$$\|C_\gamma^{(p)}\|_{1\to 1} \leq 1 + 2H_{p-1}^{(1-\gamma)} + \frac{16p^\gamma}{1-\gamma} \leq 3 + \frac{2p^\gamma}{\gamma} + \frac{16p^\gamma}{1-\gamma}. \tag{118}$$

where the second step uses Lemma 2 for $H_{p-1}^{(1-\gamma)} \leq 1 + p^\gamma/\gamma$. The asymptotic statement follows from

$$3 + \frac{2p^\gamma}{\gamma} + \frac{16p^\gamma}{1-\gamma} = \frac{3\gamma(1-\gamma) + (2(1-\gamma)+16\gamma)p^\gamma}{\gamma(1-\gamma)} \leq \frac{3+18p^\gamma}{\gamma(1-\gamma)} = O\left(\frac{p^\gamma}{\gamma(1-\gamma)}\right), \tag{119}$$

completing the proof. □

C.4 Error Bounds

Finally, having derived the requisite matrix norm bounds, we move on to error bounds.

Corollary 2. *Let $\gamma \in [1/2, 1)$, $n \geq p \geq 2$ and $1 \leq k \leq \frac{n}{b}$. Then*

$$\mathcal{E}(B_\gamma^{(p)}, C_\gamma^{(p)}) = O\Big(\sqrt{k}\, p^{\gamma-\frac{1}{2}}\big(\Lambda_\gamma + \tau^{-1/2}\big) + \sqrt{\tfrac{nk}{p}}\big(\sqrt{\tau} + \tau\sqrt{\Lambda_\gamma}\big) + \sqrt{\tfrac{nk}{b}} + \sqrt{\tfrac{k}{b}}\, p^\gamma\big(\tau^{-1} + \sqrt{\Lambda_\gamma}\big)\Big).$$

where $\tau = 1-\gamma$ and $\Lambda_\gamma = \min\{\log(ep), 1/|2\gamma - 1|\} \geq 1$.

Proof. Statement follows from $\mathcal{E}(B_\gamma^{(p)}, C_\gamma^{(p)}) = \frac{1}{\sqrt{n}}\|B_\gamma^{(p)}\|_F \cdot \mathrm{sens}_{k,b}(C_\gamma^{(p)})$, recalling that

$$\mathrm{sens}_{k,b}(C_\gamma^{(p)}) \le \sqrt{k}\,\|C_\gamma^{(p)}\|_{1\to 2} + \sqrt{\frac{k}{b}}\,\|C_\gamma^{(p)}\|_{1\to 1}$$

from Lemma 1 (together with Lemma 6 for the monotonicity of $C_\gamma^{(p)}$'s Toeplitz coefficients), then invoking the bounds Lemma 8, Lemma 9 and Lemma 10 for each of the matrix norms involved, then carrying out the algebra.

To this end, using the requisite matrix norm bounds for $\gamma \in [1/2, 1)$, we have that

$$\frac{1}{\sqrt{n}}\|B_\gamma^{(p)}\|_F = O\left(1 + \tau(\sqrt{\Lambda_\gamma} + \sqrt{n}p^{-\gamma})\right)$$
$$\mathrm{sens}_{k,b}(C_\gamma^{(p)}) = O\left(\sqrt{k}p^{\gamma-1/2}\left(\sqrt{\Lambda_\gamma} + \tau^{-1/2}\right) + \sqrt{\frac{k}{b}}\frac{p^\gamma}{\tau}\right).$$

Multiplying the terms through and using that $\tau \le 1$, and $\sqrt{\Lambda_\gamma} \le \Lambda_\gamma$ gives the stated bound. □

While the expression is hard to parse, Corollary 2 captures the error of γ-BIFR in the most important regime of γ. For the choice of $\gamma = 1/2$, it recovers the error of BISR (Kalinin et al., 2026b), as expected, and for choices of γ closer to 1, it achieves an error guarantee comparable to DP-λCGD. Theorem 2 traces out this trade-off curve more explicitly.

For the remainder of this section, let

$$u := \frac{p}{\sqrt{n}}, \qquad W := \frac{p\log p}{\sqrt{n}} = u\log p, \qquad \tau^\star := \frac{\log(1+W)}{\log p}, \qquad \gamma^\star := \max\Big\{\tfrac{1}{2},\, 1-\tau^\star\Big\}.$$

We will need the following set of basic inequalities to prove Theorem 2.

Lemma 11 (Basic inequalities)**.** *Let $u, W > 0$, $L \ge \log 2$ and $p \ge 2$.*

(a) $W \le (1+W)\log(1+W)$.

(b) $\log(1+uL)/L \le 3\log(1+u)$.

(c) $u/(1+uL) \le 3\log(1+u)$.

(d) $\log\big(1 + (p^{1/4}-1)/\log p\big) \ge c_1 \log p$ *for an absolute constant* $c_1 > 0$.

Proof. **(a)** The function $(1+W)\log(1+W) - W$ vanishes at 0 and has derivative $\log(1+W) \ge 0$.

For (b) and (c) we use $\log(1+u) \ge u/2$ for $u \le 1$. **(b)** If $u \le 1$, then:

$$\log(1+uL)/L \le u \le 2\log(1+u). \tag{120}$$

If $u > 1$: then $1 + uL \le (1+u)(1+L)$, and so

$$\log(1+uL)/L \le \log(1+u)/\log 2 + 1 \le 2\log(1+u)/\log 2 \le 3\log(1+u). \tag{121}$$

(c) If $u \le 1$, then:

$$u/(1+uL) \le u \le 2\log(1+u). \tag{122}$$

If $u > 1$:

$$u/(1+uL) \le 1/L \le 1/\log 2 \le \log(1+u)/(\log 2)^2 \le 3\log(1+u). \tag{123}$$

(d) Divided by $\log p$, the left-hand side is continuous and positive on $[2, \infty)$. It tends to $\frac{1}{4}$ as $p \to \infty$. Hence its infimum over $[2, \infty)$ is a positive absolute constant. □

Proof of Theorem 2. Write $L := \log p$, $\Lambda := \Lambda_{\gamma^\star}$, $\tau := 1 - \gamma^\star = \min\{\tau^\star, \frac{1}{2}\}$ and $Q := \sqrt{(nk/p)\log(1+u)}$ for the third term of the claimed bound. Applying Corollary 2 with $\gamma = \gamma^\star$ gives

$$\mathcal{E}\big(B_\gamma^{(p)}, C_\gamma^{(p)}\big) = O\big(T_1 + T_2 + \sqrt{nk/b} + T_3\big), \tag{124}$$

where

$$T_1 := \sqrt{k}\, p^{\gamma^\star - \frac{1}{2}}\big(\Lambda + \tau^{-1/2}\big), \qquad T_2 := \sqrt{\tfrac{nk}{p}}\big(\sqrt{\tau} + \tau\sqrt{\Lambda}\big), \qquad T_3 := \sqrt{\tfrac{k}{b}}\, p^{\gamma^\star}\big(\tau^{-1} + \sqrt{\Lambda}\big).$$

Throughout we use $pL = W\sqrt{n}$ and $p^{1-\tau^\star} = p/(1+W)$. It remains to bound T_1, T_2 and T_3, for which we distinguish two cases.

Case 1: $\tau^\star \le \frac{1}{4}$**.** Then $\tau = \tau^\star$ and $\gamma^\star = 1 - \tau \ge \frac{3}{4}$, so $\Lambda \le 1/(2\gamma^\star - 1) \le 2$. Since $\tau^{-1/2} \ge 2 \ge \Lambda$ and $\tau^{-1} \ge 4 \ge \sqrt{\Lambda}$, we have that

$$T_1 \le 2\sqrt{k}\, p^{\gamma^\star - \frac{1}{2}} \tau^{-1/2}, \qquad T_3 \le 2\sqrt{k/b}\, p^{\gamma^\star}/\tau. \tag{125}$$

Since $\tau\sqrt{\Lambda} \le \sqrt{2}\,\tau \le \sqrt{2}\,\sqrt{\tau}$, we also have that $T_2 \le 3\sqrt{\tau nk/p}$. We proceed to show the corresponding bounds.

- Firstly, by Lemma 11(a) and then (c),

$$\frac{p^{2\gamma^\star - 1}}{\tau} = \frac{pL}{(1+W)^2 \log(1+W)} = \frac{\sqrt{n}\, W}{(1+W)^2 \log(1+W)} \tag{126}$$

$$\le \frac{\sqrt{n}}{1+W} = \frac{n}{p} \cdot \frac{u}{1+uL} \le 3\,\frac{n}{p}\log(1+u), \tag{127}$$

  hence $T_1^2 \le 12\, Q^2$.

- Secondly, by Lemma 11(b), $\tau = \log(1+uL)/L \le 3\log(1+u)$, hence $T_2^2 \le 27\, Q^2$.

- Lastly, by Lemma 11(a),

$$p^{\gamma^\star}/\tau = pL/\big((1+W)\log(1+W)\big) = \sqrt{n}\, W/\big((1+W)\log(1+W)\big) \le \sqrt{n}, \tag{128}$$

  hence $T_3 \le 2\sqrt{nk/b}$.

Case 2: $\tau^\star > \frac{1}{4}$**.** The assumption is equivalent to $1 + uL > p^{1/4}$, i.e. $u > (p^{1/4} - 1)/L$, so by Lemma 11(d)

$$\log(1+u) \ge c_1 L. \tag{129}$$

In particular $\sqrt{nk/p} = O(Q)$ and $\sqrt{nkL/p} = O(Q)$. Since $1 - \tau^\star < \frac{3}{4}$, we have $\gamma^\star \in [\frac{1}{2}, \frac{3}{4})$, so $\tau \in (\frac{1}{4}, \frac{1}{2}]$. Moreover, we also have that $\Lambda \le \log(ep) = 1 + L \le 3L$. Let $z := p/(1+W)^2$. Then

$$p^{\gamma^\star - \frac{1}{2}} = \max\big\{1,\, p^{\frac{1}{2} - \tau^\star}\big\} = \max\{1, \sqrt{z}\}, \qquad \sqrt{z} \le \frac{\sqrt{p}}{W} = \frac{\sqrt{n/p}}{L}, \tag{130}$$

and so

$$p^{\gamma^\star} = \sqrt{p}\, p^{\gamma^\star - \frac{1}{2}} \le \sqrt{p} + \frac{\sqrt{n}}{L}. \tag{131}$$

We are now ready to bound the terms for this case.

- Firstly,

$$T_1 \le \sqrt{k}\,(1 + \sqrt{z})(3L + 2) \le 6\sqrt{k}\, L\,(1 + \sqrt{z}) \tag{132}$$

$$\le 6\sqrt{k}\, L + 6\sqrt{nk/p} = O\big(\sqrt{k}\log p + Q\big), \tag{133}$$

  using $2 \le 3L$.

- Secondly, $T_2 \le \sqrt{nk/p}\,\big(1 + \frac{1}{2}\sqrt{3L}\big) = O\big(\sqrt{nkL/p}\big) = O(Q)$.

- Lastly, $T_3 \le \sqrt{k/b}\,\big(\sqrt{p} + \sqrt{n}/L\big)\big(4 + \sqrt{3L}\big) = O\big(\sqrt{kp\log p/b} + \sqrt{nk/b}\big)$, using that $4 \le 5\sqrt{L}$ and $\sqrt{L} \le 2L$.

Across both cases (i.e., for all values of $\gamma^\star \in [1/2, 1)$), T_1, T_2 and T_3 can each be bounded (up to constants) by terms from the bound in Theorem 2. This completes the proof. □