A UNIFIED THEORY OF RANDOM PROJECTION FOR INFLUENCE FUNCTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Influence functions and related data attribution scores take the form of inverse-sensitive bilinear functionals $g^\top F^{-1} g'$, where $F \succeq 0$ is a curvature operator and g, g' are training and test gradients. In modern overparameterized models, forming or inverting $F \in \mathbb{R}^{d\times d}$ is prohibitive, motivating scalable influence computation via *random projection* with a sketch $P \in \mathbb{R}^{m\times d}$. This practice is commonly justified via the Johnson–Lindenstrauss (JL) lemma, which ensures approximate preservation of Euclidean geometry for a fixed dataset. However, preserving pairwise distances does not address how sketching behaves under inversion. Furthermore, there is no existing theory that explains how sketching interacts with other widely-used heuristics, such as ridge regularization (replacing F^{-1} with $(F+\lambda I)^{-1}$) and structured curvature approximations.

We develop a unified theory characterizing when projection provably preserves influence functions, with a focus on the required sketch size m. When $g, g' \in \mathrm{range}(F)$, we show that: (i) **Unregularized projection**: exact preservation holds if and only if P is injective on $\mathrm{range}(F)$, which necessitates $m \geq \mathrm{rank}(F)$; (ii) **Regularized projection**: ridge regularization fundamentally alters the sketching barrier, with approximation guarantees governed by the *effective dimension* of F at the regularization scale λ. This dependence is both sufficient and worst-case necessary, and can be substantially smaller than $\mathrm{rank}(F)$; and (iii) **Factorized influence**: for Kronecker-factored curvatures $F = A \otimes E$, the guarantees continue to hold for decoupled sketches $P = P_A \otimes P_E$, even though such sketches exhibit structured row correlations that violate canonical i.i.d. assumptions; the analysis further reveals an explicit computational–statistical trade-off inherent to factorized sketches. Beyond this range-restricted setting, we analyze **out-of-range test gradients** and quantify a sketch-induced *leakage* term that arises when test gradients have components in $\ker(F)$. This yields guarantees for influence queries on general, unseen test points.

Overall, this work develops a novel and rigorous theory that characterizes when projection provably preserves influence and provides principled, instance-adaptive guidance for choosing the sketch size m in practice.

1 INTRODUCTION

Data attribution aims to explain a trained model's behavior by tracing its predictions back to the training examples (Hammoudeh & Lowd, 2024; Deng et al., 2025). A classical tool is the *influence function* (Hampel, 1974; Koh & Liang, 2017), which measures how reweighting a training example changes the loss at a test point. In modern neural networks, computing influence involves extremely high-dimensional per-example gradients and ill-conditioned (often singular) curvature operators F. Consequently, scalable influence methods rely on *random projection*, which compresses gradients and curvature to a much smaller dimension before carrying out influence computations (Wojnowicz et al., 2016; Park et al., 2023; Choe et al., 2024; Hu et al., 2025). In these works, projection is often heuristically justified via the Johnson–Lindenstrauss (JL) lemma (Lindenstrauss & Johnson, 1984), since common sketches (Gaussian, Rademacher, and sparse JL) approximately preserve Euclidean geometry (Ailon & Chazelle, 2009; Kane & Nelson, 2014; Nelson & Nguyên, 2013; Cohen, 2016). However, influence depends on an *inverse-sensitive* bilinear form induced by F^{-1}, so JL-style arguments do not, on their own, guarantee that projection preserves influence. Furthermore, while recent empirical evidence suggests that the quality of projected influence is sensitive to the sketch

size and other hyperparameters, such as the regularization strength (Wang et al., 2025), a formal theoretical understanding of how projection affects influence functions is still lacking.

In this work, we develop a unified theoretical analysis of projection across three widely used influence-function variants in large-scale neural networks: (1) **Unregularized projection** (Wojnowicz et al., 2016; Park et al., 2023), which applies sketching directly to influence computations without explicit regularization; (2) **Regularized projection** (Zheng et al., 2024; Mlodozeniec et al., 2025), which combines sketching with ridge regularization to stabilize inverse curvature computations (Koh & Liang, 2017); and (3) **Kronecker-factored influence** (Choe et al., 2024; Hu et al., 2025), which applies factorized projection on top of structured curvature approximations such as K-FAC (Martens & Grosse, 2015; George et al., 2018).

Setup and Notation. Let g and g' denote training and test gradients with respect to the trained model parameters $\theta \in \mathbb{R}^d$, and let $F \succeq 0$ be a curvature matrix evaluated at θ, with $r := \operatorname{rank}(F)$. Typical choices of F include the generalized Gauss–Newton matrix (Bae et al., 2022; Mlodozeniec et al., 2025) and the empirical Fisher $\frac{1}{n}\sum_{i=1}^{n} g_i g_i^\top$ (Grosse et al., 2023; Kwon et al., 2024), both standard approximations to the Hessian. We study the inverse-sensitive bilinear form with a ridge parameter $\lambda \geq 0$, denoted as $\tau_\lambda(g, g') := g^\top (F + \lambda I_d)^{-1} g'$, where F^{-1} denotes either the matrix inverse or the Moore–Penrose pseudoinverse when F is singular. Unless otherwise stated, we let $P \in \mathbb{R}^{m \times d}$ denote a sketch whose rows are i.i.d. $1/\sqrt{m}$-scaled isotropic sub-Gaussian vectors (Vershynin, 2018, Chapter 2).[1] Such matrices are commonly referred to as *oblivious sketching matrices* and include Gaussian, Rademacher, and sparse JL transforms widely used in practice. The resulting projected (possibly regularized) influence is defined $\widetilde{\tau}_\lambda(g, g') := (Pg)^\top (PFP^\top + \lambda I_m)^{-1} (Pg')$.

Our Contributions. We present a sequence of results characterizing when projection provably preserves influence functions across a range of settings. Under the assumption $g, g' \in \operatorname{range}(F)$, we precisely delineate when projection *can* and *cannot* succeed without regularization, show how ridge regularization alters the required sketch size, and extend the analysis to Kronecker-factored curvature approximations. We then relax the assumption on g' and quantify an additional sketch-induced *leakage* term arising from components of the test gradient in $\ker(F)$, yielding guarantees for influence queries at general, unseen test points.

First, we ask whether sketching can preserve the unregularized influence $\tau_0(g, g')$. We show a dichotomy: unless the sketch is injective on $\operatorname{range}(F)$, uniform multiplicative approximation is impossible, in the sense that no bound of the form $|\tau_0(g, g') - \widetilde{\tau}_0(g, g')| \leq \varepsilon \tau_0(g, g')$ can hold for all g, g' and any $\varepsilon > 0$. Conversely, injectivity on $\operatorname{range}(F)$ guarantees *exact* preservation.

Main Result 1 (Unregularized projection, Theorem 2.1): Let $F \succeq 0$ with $r := \operatorname{rank}(F)$. For $\lambda = 0$, for all $g, g' \in \operatorname{range}(F)$, $\widetilde{\tau}_0(g, g') = \tau_0(g, g')$ *if and only if* P is injective on $\operatorname{range}(F)$. If P is not injective on $\operatorname{range}(F)$ (in particular if $m < r$), then for any constant factor, no uniform multiplicative approximation guarantee is possible over $g, g' \in \operatorname{range}(F) \setminus \{0\}$.

Theorem 2.1 shows that, without regularization, influence preservation requires m to scale on the order of r. In contrast, when ridge regularization is employed, we show that the required sketch size is no longer governed by r but instead by the *effective dimension* $d_\lambda(F) := \operatorname{tr}(F(F + \lambda I)^{-1})$, a classical notion in Bayesian model selection (Gull, 1989; MacKay, 1991). This quantity is always bounded above by r and can be substantially smaller when the spectrum of F decays quickly.

Main Result 2 (Regularized projection: Theorems 2.2 and 2.4): Fix $\lambda > 0$ and define $d_\lambda(F) = \operatorname{tr}(F(F + \lambda I)^{-1})$. If $m = \Omega\big((d_\lambda(F) + \log(1/\delta))/\varepsilon^2\big)$, then with probability at least $1 - \delta$, for all $g, g' \in \operatorname{range}(F)$,

$$\left|\widetilde{\tau}_\lambda(g, g') - \tau_\lambda(g, g')\right| \leq \varepsilon \sqrt{\tau_0(g, g)} \sqrt{\tau_0(g', g')}$$

Conversely, for Gaussian oblivious sketches, there exist $F \succeq 0$ such that if $m = o(d_\lambda(F)/\varepsilon^2)$, there exists some $g, g' \in \operatorname{range}(F)$ admits an $O(\varepsilon)$ error with constant probability.

[1] A mean-zero random variable X is *sub-Gaussian* with parameter σ^2 if $\mathbb{E}[\exp(tX)] \leq \exp(\sigma^2 t^2/2)$ for all $t \in \mathbb{R}$; a random vector is sub-Gaussian if all one-dimensional marginals are sub-Gaussian.

In large neural networks, influence computation hinges on curvature inversion, yet forming or inverting the full empirical Hessian or Fisher is infeasible. Consequently, practical pipelines adopt structured curvature approximations, most notably Kronecker-factored approximate curvature (K-FAC). This motivates us to develop a projection theory tailored to this setting.

As reviewed in Section 2.3, K-FAC models the curvature as $F = A \otimes E$ (in a layerwise manner), where A and E capture the empirical covariances of forward activations and backpropagated gradients, respectively. To exploit this structure, one natural idea is to enforce the sketch to share the same factorization $P = P_A \otimes P_E$, where P_A and P_E are oblivious sketching matrices (Choe et al., 2024). While this yields substantial computational savings, the Kronecker structure breaks the i.i.d. row assumption on P, rendering a direct adaptation of Theorems 2.1 and 2.2 inapplicable. We overcome this technical challenge through a fine-grained analysis and establish rigorous approximation guarantees.

Main Result 3 (Factorized influence, Theorems 2.5 and 2.6): Assume $F = A \otimes E \succeq 0$ and a Kronecker sketch $P = P_A \otimes P_E$ with factor sketch sizes m_A and m_E.

(i) **Unregularized barrier.** For $\lambda = 0$, exact invariance on $\mathrm{range}(F)$ holds if and only if P_A is injective on $\mathrm{range}(A)$ and P_E is injective on $\mathrm{range}(E)$, which in particular necessitates $m_A \geq \mathrm{rank}(A)$ and $m_E \geq \mathrm{rank}(E)$.

(ii) **Regularized approximation.** Let P_A and P_E each to be oblivious sketch. For $\lambda > 0$, letting $\lambda_E := \lambda/\|E\|_2$ and $\lambda_A := \lambda/\|A\|_2$, if $m_A = \Omega\big((d_{\lambda_E}(A) + \log(1/\delta))/\varepsilon^2\big)$ and $m_E = \Omega\big((d_{\lambda_A}(E) + \log(1/\delta))/\varepsilon^2\big)$, then with probability at least $1-\delta$, for all $g, g' \in \mathrm{range}(F)$,

$$|\widetilde{\tau}_\lambda(g, g') - \tau_\lambda(g, g')| \leq \varepsilon \sqrt{\tau_0(g, g)} \sqrt{\tau_0(g', g')}.$$

Finally, we note that all of the above guarantees are stated for gradients lying in $\mathrm{range}(F)$, which, when F is the empirical Fisher, includes all training gradients used for attribution. In practice, however, a test gradient g' may have a component in $\ker(F)$. We show that, in both the unregularized and regularized settings, these components do not affect the true (unsketched) influence, while sketching introduces an additional *leakage* term. We quantify this "out-of-range leakage" and show it decays at the usual $O(m^{-1/2})$ rate with explicit dependence on λ and the spectrum of F.

Main Result 4 (Projection leakage, Theorems 3.1 and 3.2): For a general $g' \in \mathbb{R}^d$, write $g' = g'_{/\!/} + g'_{\perp}$ with $g'_{/\!/} \in \mathrm{range}(F)$ and $g'_{\perp} \in \ker(F)$. We show that in this case, for either $\lambda = 0$ or $\lambda > 0$,

$$|\widetilde{\tau}_\lambda(g, g') - \tau_\lambda(g, g')| \leq |\widetilde{\tau}_\lambda(g, g'_{/\!/}) - \tau_\lambda(g, g'_{/\!/})| + |\widetilde{\tau}_\lambda(g, g'_{\perp})|,$$

with an additional leakage error $|\widetilde{\tau}_\lambda(g, g'_{\perp})|$ beyond Theorem 2.2. We then prove in Theorem 3.1 that for a collection of k test gradients $\{g'_j\}_{j=1}^k$, with sketch size $m = \Omega\big((r + \log(k/\delta))/\varepsilon^2\big)$,[a]

(i) **Unregularized**: $|\widetilde{\tau}_0(g, g'_{\perp})| \leq \varepsilon \|g\|_2 \|g'_{\perp}\|_2 / \lambda^{+}_{\min}(F)$.

(ii) **Regularized**: $|\widetilde{\tau}_\lambda(g, g'_{\perp})| \leq \varepsilon \|g\|_2 \|g'_{\perp}\|_2 (1/\lambda + 2\|F\|_2/\lambda^2)$.

Moreover, in Theorem 3.2, we show that a similar leakage guarantees extend to the factorized influence setting.

[a] Or alternatively linear in $k' = \dim\big(\mathrm{span}(\{g'_{j,\perp}\}_{j=1}^k)\big)$, which in practice is usually worse than $\log(k)$.

Taken together, we develop a unified theory for when projection can provably approximate influence-style data attribution scores of the form $g^\top (F + \lambda I)^{-1} g'$. Specifically, without regularization, projection preserves influence for all $g, g' \in \mathrm{range}(F)$ only when the sketch is injective on $\mathrm{range}(F)$, which essentially forces $m \geq \mathrm{rank}(F)$; otherwise, uniform multiplicative approximation is impossible. With regularization, the required sketch size is instead governed by the effective dimension $d_\lambda(F)$. We further extend these guarantees to Kronecker-factored (K-FAC-style) curvature and sketches. Finally, we quantify an additional sketch-induced leakage term that can appear when test gradients have components in $\ker(F)$. Overall, our results provide principled, instance-adaptive guidance for choosing m and clarify how projection interacts with regularization and structured curvature approximations.

1.1 Related Works

Influence functions were originally introduced as a classical tool in robust statistics (Hampel, 1974) and later adapted to machine learning by Koh & Liang (2017). Owing to their flexibility and generality, influence-based methods have since been widely applied to tasks such as data cleaning (Teso et al., 2021), model debugging (Guo et al., 2021), and subset selection (Hu et al., 2024), and have been extended to large-scale models, including large language models (Grosse et al., 2023) and diffusion models (Mlodozeniec et al., 2025). However, applying influence functions to modern neural networks poses significant computational challenges due to the need to invert a high-dimensional, often rank-deficient, curvature matrix F (Koh & Liang, 2017; Schioppa et al., 2022).

Several recent works propose scalable approximations based on random projection and related sketching techniques, where they typically project per-sample gradients into a lower-dimensional space before computing influence scores (Wojnowicz et al., 2016; Schioppa et al., 2022; Park et al., 2023), sometimes in combination with explicit regularization (Choe et al., 2024; Hu et al., 2025). Despite their empirical success, the theoretical guarantees underlying these methods remain limited, and their correctness is often justified heuristically.

Specifically, existing theoretical justifications for projection-based influence methods typically appeal to the Johnson–Lindenstrauss (JL) lemma (Lindenstrauss & Johnson, 1984) in the data attribution literature (Wojnowicz et al., 2016; Park et al., 2023; Deng et al., 2025). Given a finite set of vectors of size n in $\mathbb{R}^d$, the JL lemma guarantees that $m = O(\log(n)/\varepsilon^2)$ suffices to approximately preserve the pairwise distances between the n points up to a $(1 \pm \varepsilon)$ factor. While powerful, this guarantee is fundamentally misaligned with the structure of influence functions. Influence scores are not determined by Euclidean distances between gradients, but by *inverse-sensitive* bilinear forms $\tau_0(g, g') = g^\top F^{-1} g'$ involving the inverse (or pseudoinverse) of a second-order matrix F, and sketching changes the operator to be inverted. Thus, preserving $\|Pg\|_2$ (even uniformly over a finite set) does not directly control either the stability of matrix inversion after projection, nor the resulting bilinear form.

Consistent with this mismatch, empirical studies on hyperparameter sensitivity show that the quality of projected influence does not improve monotonically with the sketch size in certain scenarios (Park et al., 2023). More detailed ablation analyses further attribute this behavior to a coupled interaction between sketch size and regularization strength (Wang et al., 2025). Taken together, these observations underscore the need for a formal theoretical understanding of how projection interacts with the curvature operator, in order to guide the principled use of influence function methods in practice.

2 Projection-Based Influence Approximation

2.1 Unregularized Projection

In this section, we show that in the absence of regularization, projection alone encounters a fundamental barrier in the sketch size m. In particular, there is a sharp phase transition: when $m < r$, no multiplicative approximation guarantee is possible; whereas when $m \geq r$, a continuous sketch yields exact invariance with probability one.

Theorem 2.1 (Barrier of unregularized projection)**.** *The equality $\tau_0(g, g') = \widetilde{\tau}_0(g, g')$ holds for any $g, g' \in \mathrm{range}(F)$* iff *$P$ is injective on $\mathrm{range}(F)$, i.e. $\mathrm{rank}(PU) = \mathrm{rank}(F) = r$ where $F = U\Lambda U^\top$ is the compact eigendecomposition of F with $U \in \mathbb{R}^{d\times r}$ orthonormal and $\Lambda \in \mathbb{R}^{r\times r}$ positive definite. Subsequently, for any PSD $F \in \mathbb{R}^{d\times d}$ and **any** matrix $P \in \mathbb{R}^{m\times d}$, one cannot hope to obtain any multiplicative approximation of $\tau_0(g, g')$ via $\widetilde{\tau}_0(g, g')$ when $\mathrm{rank}(PU) < r$.*

The proof can be found in Section A. Theorem 2.1 shows that exact preservation of unregularized influence requires the sketch to be injective on $\mathrm{range}(F)$, forcing $m \geq r$. In overparameterized regimes where the high-dimensional per-sample gradients are likely to be in general position, one typically has $r \approx n$, and thus m must scale with the dataset size. In contrast, we will show that introducing ridge regularization ($\lambda > 0$) fundamentally changes this requirement, with the sketch size governed instead by the effective dimension $d_\lambda(F)$, which can be substantially smaller than r.

2.2 REGULARIZED PROJECTION

Unlike the unregularized case, in this section, we show that for the projected influence function, the extra damping term λI_d helps control the effective dimension by shrinking small eigenvalues of the curvature operator F, effectively reducing the Gaussian complexity governing the uniform concentration bound. In particular, we show that the sketch size requires *only* to scale with the *effective dimension* of F with $\lambda > 0$:

$$d_\lambda(F) := \operatorname{tr}\left(F(F+\lambda I_d)^{-1}\right) = \sum_{j=1}^{r} \frac{\lambda_j(F)}{\lambda_j(F)+\lambda} \le r.$$

In practice, as we shall observe in Section 4, the spectrum of F decays rapidly, and thus $d_\lambda \ll r \ll d$ for moderate λ. Hence, having the sketch size m to only scale with the effective dimension d_λ at scale λ rather than the ambient dimension d or the rank r of F makes the regularized projection approach feasible at scale. We now state the theorem and sketch the proof below.

Theorem 2.2 (Upper bound of regularized projection)**.** *Let $P \in \mathbb{R}^{m\times d}$ be a oblivious sketching matrix with rows $P_i^\top = \frac{1}{\sqrt{m}} W_i^\top$, where $\{W_i\}_{i=1}^m \sim W$ are i.i.d. sub-Gaussian random vectors in $\mathbb{R}^d$ satisfying $\mathbb{E}[W] = 0$ and $\mathbb{E}[WW^\top] = I_d$.[2] For any $\varepsilon, \delta \in (0,1)$, if the sketch size satisfies*

$$m = \Omega\left(\frac{d_\lambda(F) + \log(1/\delta)}{\varepsilon^2}\right),$$

then with probability at least $1-\delta$, the following bounds hold for all $g, g' \in \operatorname{range}(F)$:

$$|\widetilde{\tau}_\lambda(g,g') - \tau_\lambda(g,g')| \le \varepsilon \sqrt{\tau_0(g,g)}\sqrt{\tau_0(g',g')}.$$

Proof. Let $g, g' \in \operatorname{range}(F)$ and write $g = F^{1/2}y$ and $g' = F^{1/2}y'$. Using the push-through identity $A(A^\top A + \lambda I)^{-1} = (AA^\top + \lambda I)^{-1}A$ with $A = PF^{1/2}$, and defining $G := F^{1/2}P^\top PF^{1/2}$ yields

$$\begin{aligned}\widetilde{\tau}_\lambda(g,g') &= (Pg)^\top (PFP^\top + \lambda I)^{-1}(Pg') \\ &= y^\top F^{1/2}P^\top (PFP^\top + \lambda I)^{-1} PF^{1/2}y' = y^\top G(G+\lambda I)^{-1}y'.\end{aligned}$$

On the other hand, define $B := F^{1/2}(F+\lambda I)^{-1/2}$. Since F and $F+\lambda I$ are simultaneously diagonalizable (they share the eigenbasis of F), all matrix functions of these operators commute; in particular, $F^{1/2}$, $(F+\lambda I)^{-1/2}$, and $(F+\lambda I)^{-1}$ commute and $BB^\top = F^{1/2}(F+\lambda I)^{-1}F^{1/2} = F(F+\lambda I)^{-1}$. Hence, we have

$$\tau_\lambda(g,g') = g^\top (F+\lambda I)^{-1} g' = y^\top BB^\top y' = y^\top F(F+\lambda I)^{-1}y',$$

which gives $|\widetilde{\tau}_\lambda(g,g') - \tau_\lambda(g,g')| = |y^\top G(G+\lambda I)^{-1}y' - y^\top F(F+\lambda I)^{-1}y'|$. Thus, it suffices to control the spectrum of $F(F+\lambda I)^{-1} - G(G+\lambda I)^{-1}$.

Let $B := F^{1/2}(F+\lambda I)^{-1/2}$, so that $B^\top B = (F+\lambda I)^{-1/2}F(F+\lambda I)^{-1/2}$ and $\|B\|_2^2 = \|B^\top B\|_2 \le 1$. Applying Theorem B.2 with $M = B$ and $m = \Omega\big(\varepsilon^{-2}\big(d_\lambda(F) + \log(1/\delta)\big)\big)$ yields $\|B^\top(P^\top P - I)B\|_2 \le \varepsilon/2$. Conjugating by $(F+\lambda I)^{1/2}$ and using $G = F^{1/2}P^\top PF^{1/2}$, this implies a PSD sandwich

$$\left(1 - \tfrac{\varepsilon}{2}\right)(F+\lambda I) \preceq (G+\lambda I) \preceq \left(1+\tfrac{\varepsilon}{2}\right)(F+\lambda I).$$

Inverting the sandwich gives $\|(G+\lambda I)^{-1} - (F+\lambda I)^{-1}\|_2 \le \frac{1}{\lambda}\cdot\frac{\varepsilon/2}{1-\varepsilon/2} \le \varepsilon/\lambda$. Finally, using the identity $A(A+\lambda I)^{-1} = I - \lambda(A+\lambda I)^{-1}$ for any PSD A, we get

$$\left\|F(F+\lambda I)^{-1} - G(G+\lambda I)^{-1}\right\|_2 = \lambda \left\|(G+\lambda I)^{-1} - (F+\lambda I)^{-1}\right\|_2 \le \varepsilon,$$

which is the desired operator control (formal details are in Theorem B.3). Therefore,

$$|\widetilde{\tau}_\lambda(g,g') - \tau_\lambda(g,g')| = \left| y^\top \left[G(G+\lambda I)^{-1} - F(F+\lambda I)^{-1}\right] y' \right| \le \varepsilon \|y\|_2 \|y'\|_2.$$

As $\|y\|_2^2 = \tau_0(g,g)$ and $\|y'\|_2^2 = \tau_0(g',g')$, we conclude the proof. □

[2] Since we only assume bounded sub-Gaussian norm on the random vectors W_i, the result applies to a wide range of random projection matrices, including Gaussian, Rademacher, and sparse JL transform.

Remark 2.3. *The core technical challenge in the proof of Theorem 2.2 is to bound* $\|F(F+\lambda I)^{-1} - G(G+\lambda I)^{-1}\|_2$*. A natural alternative (see Section B.2) is to invoke an oblivious subspace embedding (OSE) (Woodruff, 2014). For a fixed matrix* $A \in \mathbb{R}^{d\times r}$*,* $P \in \mathbb{R}^{m\times d}$ *is an* ε*-OSE for* $\mathrm{range}(A)$ *if*

$$-\varepsilon A^\top A \preceq A^\top (P^\top P - I)A \preceq \varepsilon A^\top A.$$

Instantiating $A = F^{1/2}$ *yields a sandwich* $(1-\varepsilon)F \preceq G \preceq (1+\varepsilon)F$*, which implies* $\|F(F+\lambda I)^{-1} - G(G+\lambda I)^{-1}\|_2 = O(\varepsilon)$ *by operator monotonicity of* $t \mapsto t/(t+\lambda)$*. However, OSE enforces uniform multiplicative accuracy over* $\mathrm{range}(F^{1/2})$*, so even directions with* $\lambda_j(F) \ll \lambda$ *must be preserved up to a* $(1\pm\varepsilon)$ *factor, leading to* $m = \Omega(r/\varepsilon^2)$ *(Woodruff, 2014, Theorems 2.3 and 6.10).*

Our proof instead exploits the weaker, λ*-dependent requirement: it suffices for* P *to be an approximate isometry on the whitened subspace* $B = F^{1/2}(F+\lambda I)^{-1/2}$*, i.e.,* $\|B^\top(P^\top P - I)B\|_2 \le O(\varepsilon)$*. This yields the* ***ridge-regularized sandwich*** $(1-\varepsilon)(F+\lambda I) \preceq G + \lambda I \preceq (1+\varepsilon)(F+\lambda I)$*. Importantly, this condition controls* $F+\lambda I$ *rather than* F *itself: in directions where* $\lambda_j(F) \ll \lambda$*, both* $F+\lambda I$ *and* $G+\lambda I$ *are dominated by* λ*, so even large relative errors in* F *have negligible impact on the inverse. Consequently, such low-eigenvalue directions need not be preserved multiplicatively, and the required sketch size is governed by the effective dimension at scale* λ*, yielding the sharper bound* $m = \Omega(d_\lambda(F)/\varepsilon^2)$*.*

We now complement Theorem 2.2 with a worst-case matching lower bound, showing that the effective dimension $d_\lambda(F)$ characterizes the tight dependence of m for oblivious sketching in regularized influence. Concretely, we show that for Gaussian oblivious sketches, if the sketch size is smaller than $\Theta(d_\lambda(F)/\varepsilon^2)$, then there exist problem instances on which the sketched influence incurs $\Omega(\varepsilon)$ error with constant probability.

Theorem 2.4 (Lower bound for regularized projection)**.** *Let* $P \in \mathbb{R}^{m\times d}$ *be a Gaussian oblivious sketch with rows i.i.d.* $\mathcal{N}(0, I_d)$*. There exists a family of* $F \in \mathbb{R}^{d\times d}$ *such that if* $m = o(d_\lambda(F)/\varepsilon^2)$*, then there exists* $g \in \mathrm{range}(F)$ *with* $|\widetilde{\tau}_\lambda(g,g) - \tau_\lambda(g,g)| = \Omega(\varepsilon)\tau_0(g,g)$ *with constant probability.*

Full details are in Section B. We see that Theorem 2.4 formalizes a worst-case limitation for this class of sketches: with Gaussian oblivious projections, one cannot uniformly beat the $d_\lambda(F)/\varepsilon^2$ scaling. Combined with the instance-adaptive upper bound in Theorem 2.2, this identifies $d_\lambda(F)$ as the fundamental complexity parameter governing regularized projection.

2.3 Factorized Influence

In many large-scale settings, explicitly forming or inverting the empirical Fisher/Hessian F is infeasible, and second-order methods instead rely on structured approximations. A common choice is a Kronecker factorization (e.g., K-FAC (Martens & Grosse, 2015; Grosse et al., 2023)), which models each layerwise block as $F \approx A \otimes E$ for smaller PSD factors $A \in \mathbb{R}^{d_A\times d_A}$ and $E \in \mathbb{R}^{d_E\times d_E}$, which are forward activation and backprop-gradient covariances, respectively.

This structure suggests a natural computational counterpart on the sketching side: use a *factorized sketch* $P = P_A \otimes P_E$, where $P_A \in \mathbb{R}^{m_A\times d_A}$ and $P_E \in \mathbb{R}^{m_E\times d_E}$ are respectively the standard oblivious sketching considered in Theorem 2.2.[3] The resulting sketch has ambient dimension $d := d_A d_E$ and sketch dimension $m := m_A m_E$, i.e., $P \in \mathbb{R}^{m\times d}$. Moreover, write a per-example layer gradient as a matrix $G \in \mathbb{R}^{d_E\times d_A}$ with $g = \mathrm{vec}(G) \in \mathbb{R}^d$. Then the projection can be computed without materializing the full $m \times d$ sketching matrix as $Pg = (P_A \otimes P_E)\,\mathrm{vec}(G) = \mathrm{vec}(P_E G P_A^\top)$. Consequently, the per-example cost reduces to two smaller multiplies $P_E G$ and $(P_E G)P_A^\top$, plus solving the resulting regularized system in sketch dimension m. Similarly, we can also form the sketched curvature efficiently: using the mixed-product identity of Kronecker products, $PFP^\top = (P_A \otimes P_E)(A \otimes E)(P_A \otimes P_E)^\top = (P_A A P_A^\top) \otimes (P_E E P_E^\top)$.

In the unregularized case ($\lambda = 0$), the exact invariance barrier becomes strictly more stringent under a Kronecker sketch: exact preservation on $\mathrm{range}(F)$ holds if and only if *both* factor sketches are injective on their respective ranges.

[3] Concretely, rows of P_A and P_E are i.i.d. isotropic sub-Gaussian random vectors with scaling $1/\sqrt{m_A}$ or $1/\sqrt{m_E}$.

Theorem 2.5 (Barrier of unregularized projection for factorized influence)**.** *Let $F = A \otimes E \succeq 0$ and $P = P_A \otimes P_E$ as above. Then $\widetilde{\tau}_0(g, g') = \tau_0(g, g')$ for all $g, g' \in \mathrm{range}(F)$ if and only if P_A is injective on $\mathrm{range}(A)$ and P_E is injective on $\mathrm{range}(E)$. In particular, this necessitates $m_A \geq \mathrm{rank}(A)$ and $m_E \geq \mathrm{rank}(E)$, hence $m = m_A m_E \geq \mathrm{rank}(A)\,\mathrm{rank}(E) = \mathrm{rank}(F)$.*

See Section C.1 for a proof. This motivates integrating regularization and consider

$$\begin{aligned}\widetilde{\tau}_\lambda(g, g') &= (Pg)^\top (PFP^\top + \lambda I_m)^{-1}(Pg') \\ &= \mathrm{vec}(P_E G P_A^\top)^\top \left((P_A A P_A^\top) \otimes (P_E E P_E^\top) + \lambda I_m\right)^{-1} \mathrm{vec}(P_E G' P_A^\top).\end{aligned}$$

However, factorization changes the sketching analysis: when $P = P_A \otimes P_E$, the matrix $P^\top P$ is no longer a standard i.i.d. sample covariance, so the covariance-type deviation driving the proof of Theorem 2.2 requires a dedicated argument. We now present the corresponding approximation guarantee for regularized projection under this factorized model. The key technical step is a factorized covariance deviation bound (Theorem C.1), proved in Section C.

Theorem 2.6 (Upper bound of regularized projection for factorized influence)**.** *Let $F = A \otimes E \succeq 0$ and $P = P_A \otimes P_E$ be as above, with the factors P_A, P_E denote the sketching matrix defined in Theorem 2.2. Assume $\lambda \leq \|A\|_2 \|E\|_2$, and define the rescaled regularization levels $\lambda_E := \lambda/\|E\|_2$ and $\lambda_A := \lambda/\|A\|_2$. For any $\varepsilon, \delta \in (0, 1)$, if the sketch sizes for P_A and P_E satisfy*

$$m_A = \Omega\left(\frac{d_{\lambda_E}(A) + \log(1/\delta)}{\varepsilon^2}\right), \qquad m_E = \Omega\left(\frac{d_{\lambda_A}(E) + \log(1/\delta)}{\varepsilon^2}\right),$$

then with probability at least $1 - \delta$, the following holds for all $g, g' \in \mathrm{range}(F)$:

$$|\widetilde{\tau}_\lambda(g, g') - \tau_\lambda(g, g')| \leq \varepsilon \sqrt{\tau_0(g, g)}\sqrt{\tau_0(g', g')}.$$

Proof. The proof follows the same template as Theorem 2.2. Let $B := F^{1/2}(F + \lambda I)^{-1/2}$ and $G := F^{1/2} P^\top P F^{1/2}$, and the key step is again to control the covariance-type deviation $\|B^\top (P^\top P - I) B\|_2$. We apply Theorem C.1 (proved in Section C) with parameters $\varepsilon_0 := \varepsilon/10$ and $\delta_0 := \delta/2$. Under the stated conditions on m_A and m_E, this yields that with probability at least $1 - 2\delta_0 = 1 - \delta$,

$$\|B^\top (P^\top P - I) B\|_2 \leq 2\varepsilon_0 + 3\varepsilon_0^2 \leq \varepsilon/2,$$

where the last inequality uses $\varepsilon \in (0, 1)$. On this event, the same PSD sandwich and resolvent perturbation argument used in Theorem B.3 implies $\|F(F + \lambda I)^{-1} - G(G + \lambda I)^{-1}\|_2 \leq \varepsilon$, which in turn gives the stated bilinear (and quadratic) influence error bounds. □

Remark 2.7. *Theorem 2.6 highlights a fundamental computational–statistical trade-off. While factorized sketches offer clear computational advantages over unfactorized ones, they incur a higher statistical cost in terms of the required sketch size. In particular, the total sketch size is $m = m_A m_E$, and achieving ε-approximation error requires $m = m_A m_E = \widetilde{\Omega}\big(\varepsilon^{-4}(d_{\lambda_E}(A) d_{\lambda_A}(E))\big)$. This exhibits a worse dependence on ε (from ε^{-2} to ε^{-4}) compared to the unfactorized sketch guarantee in Theorem 2.2. Informally, factorized sketches regularize each mode independently, replacing joint regularization with separable computation. As a result, they are most effective in regimes where the computational and memory savings dominate the statistical overhead.*

3 Influence with Out-of-Range Test Gradients

The analysis in Section 2 assumes that both arguments of the (regularized) influence bilinear form lie in $\mathrm{range}(F)$. This assumption is natural for training gradients: when F is instantiated as the (empirical) Fisher information matrix, $F = \frac{1}{n}\sum_{i=1}^n g_i g_i^\top$, every training gradient lies in $\mathrm{range}(F)$ by construction. In practice, however, we are often interested in the influence of an unseen test point z' with respect to a training point z, for which the corresponding test gradients g' need not lie in $\mathrm{range}(F)$.

We extend the above guarantees to this setting by explicitly characterizing the additional sketch-induced error arising from the component of g' orthogonal to $\mathrm{range}(F)$.

3.1 LEAKAGE OF PROJECTION

To make the source of this additional term explicit, we decompose $g' = g'_{/\!/} + g'_{\perp}$, where $g'_{/\!/} \in \mathrm{range}(F)$ and $g'_{\perp} \in \ker(F)$, such that the decomposition is orthogonal in the Euclidean inner product. Using linearity of $\tau_\lambda(\cdot,\cdot)$ and $\widetilde{\tau}_\lambda(\cdot,\cdot)$ in their second argument, we have $\tau_\lambda(g,g') = \tau_\lambda(g,g'_{/\!/}) + \tau_\lambda(g,g'_{\perp})$ and $\widetilde{\tau}_\lambda(g,g') = \widetilde{\tau}_\lambda(g,g'_{/\!/}) + \widetilde{\tau}_\lambda(g,g'_{\perp})$. Consequently,

$$|\widetilde{\tau}_\lambda(g,g') - \tau_\lambda(g,g')| = \left|\left(\widetilde{\tau}_\lambda(g,g'_{/\!/}) - \tau_\lambda(g,g'_{/\!/})\right) + \widetilde{\tau}_\lambda(g,g'_{\perp}) - \tau_\lambda(g,g'_{\perp})\right|.$$

Observe that the true (regularized) influence does not couple $\mathrm{range}(F)$ and $\ker(F)$, i.e., $\tau_\lambda(g,g'_{\perp}) = g^\top (F+\lambda I)^{-1} g'_{\perp} = 0$ for all $\lambda \geq 0$: indeed, F and $(F+\lambda I)^{-1}$ share the same eigenbasis, and since $g \in \mathrm{range}(F)$ and $g'_{\perp} \in \ker(F)$, hence $(F+\lambda I)^{-1} g$ and $g'_{\perp}$ lie in orthogonal subspaces. Hence,

$$|\widetilde{\tau}_\lambda(g,g') - \tau_\lambda(g,g')| \leq |\widetilde{\tau}_\lambda(g,g'_{/\!/}) - \tau_\lambda(g,g'_{/\!/})| + |\widetilde{\tau}_\lambda(g,g'_{\perp})|.$$

The first term can be bounded via Theorem 2.2; on the other hand, the remaining term is a purely sketch-induced artifact: the sketch can introduce a nonzero *leakage term* $\widetilde{\tau}_\lambda(g,g'_{\perp})$ due to mixing between $\mathrm{range}(F)$ and $\ker(F)$ under $P^\top P$. We now present a general bound on the leakage:

Theorem 3.1. *Let $\{g'_j\}_{j=1}^k \subset \mathbb{R}^d$, and for each j let $g'_{j,\perp} \coloneqq \Pi_{\ker(F)} g'_j$ denote the orthogonal projection of g'_j onto $\ker(F)$. Let $k' \coloneqq \dim\big(\mathrm{span}(\{\Pi_{\ker(F)} g'_j\}_{j=1}^k)\big)$. For any $\varepsilon,\delta \in (0,1)$, if*

$$m = \Omega\left(\frac{r + \min\left\{\log(k/\delta), k' + \log(1/\delta)\right\}}{\varepsilon^2}\right),$$

then with probability at least $1-\delta$, the following holds for all $j \in \{1,\dots,k\}$:

- ***Unregularized**: $|\widetilde{\tau}_0(g,g'_{j,\perp})| \leq \varepsilon \|g\|_2 \|g'_{j,\perp}\|_2 / \lambda^+_{\min}(F)$, where $\lambda^+_{\min}(F)$ denotes the smallest non-zero eigenvalue of F.*
- ***Regularized**: $|\widetilde{\tau}_\lambda(g,g'_{j,\perp})| \leq \varepsilon \|g\|_2 \|g'_{j,\perp}\|_2 \big(\frac{1}{\lambda} + \frac{2\|F\|_2}{\lambda^2}\big)$ for any $\lambda > 0$.*

Proof sketch. The proof is organized around a deterministic reduction: Theorem D.1 (in Section D) shows that both the unregularized and regularized leakage bounds follow as soon as two concentration conditions hold for the sketch P: (i) an operator-norm bound on $\mathrm{range}(F)$, $\|U^\top (P^\top P - I) U\|_2 \leq \varepsilon$ for an orthonormal basis U of $\mathrm{range}(F)$, and (ii) a cross-term bound between $\mathrm{range}(F)$ and the kernel direction(s), $\|U^\top (P^\top P - I) g'\|_2 \leq \varepsilon \|g'\|_2$. For a single test gradient $g'_{\perp}$, both conditions follow from applying Theorem B.2 to the $(r+1)$-dimensional subspace $\mathrm{span}(\mathrm{range}(F) \cup \{g'\})$, which yields the claimed $m = \Omega((r + \log(1/\delta))/\varepsilon^2)$ scaling. To obtain uniform control over multiple test gradients, we use two complementary arguments: a *subspace argument*, which applies the same concentration bound to $\mathrm{span}(\mathrm{range}(F) \cup \{g'_{j,\perp}\}_{j=1}^k)$ and yields the dependence on $k' = \dim \mathrm{span}(\{g'_{j,\perp}\})$ (Theorem D.3); or a *union-bound argument*, which establishes a fixed-g' tail bound and unions over k, yielding the $O(\log k)$ dependence (Theorem D.4). □

3.2 LEAKAGE OF FACTORIZED INFLUENCE

Theorem 3.1 is stated for oblivious sketches with i.i.d. rows. We now extend and prove an analogous leakage guarantee for factorized sketches $P = P_A \otimes P_E$ when F admits a Kronecker factorization.

Theorem 3.2. *Let $A, E \succeq 0$ and $F \coloneqq A \otimes E$, with $P = P_A \otimes P_E$ be the same setting as Theorem 2.6, and let $r_A \coloneqq \mathrm{rank}(A)$, $r_E \coloneqq \mathrm{rank}(E)$, and $r \coloneqq \mathrm{rank}(F) = r_A r_E$. Let $\{g'_j\}_{j=1}^k$ be test gradients of the form $g'_j = a'_j \otimes e'_j$, and write $a'_j = a'_{j,/\!/} + a'_{j,\perp}$ with $a'_{j,/\!/} \in \mathrm{range}(A)$ and $a'_{j,\perp} \perp \mathrm{range}(A)$, and similarly $e'_j = e'_{j,/\!/} + e'_{j,\perp}$. Define $k_A \coloneqq \sum_{j=1}^k \mathbb{1}(a'_{j,\perp} \neq 0)$, $k_E \coloneqq \sum_{j=1}^k \mathbb{1}(e'_{j,\perp} \neq 0)$, and $k'_A \coloneqq \dim\big(\mathrm{span}(\{a'_{j,\perp}\}_{j=1}^k)\big)$, $k'_E \coloneqq \dim\big(\mathrm{span}(\{e'_{j,\perp}\}_{j=1}^k)\big)$. For any $\varepsilon,\delta \in (0,1)$, if*

$$m_A = \Omega\left(\frac{r_A + \min\{\log(\frac{k_A}{\delta}), k'_A + \log(\frac{1}{\delta})\}}{\varepsilon^2}\right), m_E = \Omega\left(\frac{r_E + \min\{\log(\frac{k_E}{\delta}), k'_E + \log(\frac{1}{\delta})\}}{\varepsilon^2}\right),$$

then with probability at least $1-\delta$, the following bounds hold simultaneously for all $j \in \{1,\dots,k\}$:

- ***Unregularized:*** $|\widetilde{\tau}_0(g, g'_{j,\perp})| \le \varepsilon \|g\|_2 \|g'_{j,\perp}\|_2 / \lambda^+_{\min}(F)$.
- ***Regularized:*** $|\widetilde{\tau}_\lambda(g, g'_{j,\perp})| \le \varepsilon \|g\|_2 \|g'_{j,\perp}\|_2 \left(\frac{1}{\lambda} + \frac{2\|F\|_2}{\lambda^2}\right)$ *for any* $\lambda > 0$,

Proof sketch. The factorized theorem is proved by following the same high-level template as Theorem 3.1: we first reduce the leakage bound to the two concentration conditions in Theorem D.1 (stability on $\mathrm{range}(F)$ and a cross-term bound between $\mathrm{range}(F)$ and $\ker(F)$). For a Kronecker sketch $P = P_A \otimes P_E$, the stability condition on $\mathrm{range}(F) = \mathrm{range}(A) \otimes \mathrm{range}(E)$ is obtained by controlling the factor-level subspace deviations $\|U_A^\top (P_A^\top P_A - I) U_A\|_2$ and $\|U_E^\top (P_E^\top P_E - I) U_E\|_2$ (with U_A, U_E bases of $\mathrm{range}(A), \mathrm{range}(E)$). For the cross-term condition, we expand $P^\top P - I$ into factor deviations and use Theorem E.2 (in Section E) to reduce $\|U^\top (P^\top P - I) g'_\perp\|_2$ to a small collection of factor-level "primitive" quantities such as $\|U_A^\top (P_A^\top P_A - I)(\cdot)\|_2$ and $\|U_E^\top (P_E^\top P_E - I)(\cdot)\|_2$. Finally, as in the proof of Theorem 3.1, these primitives are controlled via a *union-bound argument* (yielding the $O(\log k.)$ terms) or a *subspace argument* (yielding the $k'_\cdot$ terms). Plugging these bounds into Theorem D.1 yields the stated leakage guarantees; full details are in Section E. □

Remark 3.3. *Which argument is tighter depends on the geometry of the test gradients. When* $\{g'_j\}$ *are strongly correlated or effectively low-dimensional, one can have* $k' \ll k$*, in which case the subspace argument is preferable. In contrast, in high ambient dimension, moderately many generic test gradients are typically in general position, so* k' *rapidly grows to* $\min\{k, d\}$ *and in particular satisfies* $k' \approx k$ *once* $k \ll d$*. In this common regime, the union-bound argument yields the more practical scaling in* k*, requiring only an additional* $O(\log k)$ *sketch size to ensure uniform control.*

4 Experiment and Discussion

We empirically illustrate several implications of our theory. Throughout, we consider F to be the empirical Fisher, and P to be the sparse JL transform (Kane & Nelson, 2014), and we always report the results across 5 independent runs with different sampled P. Following the data attribution library `dattri` (Deng et al., 2024), we consider three dataset–model pairs: 1.) MNIST-10 + LR, 2.) MNIST-10 + MLP, and 3.) CIFAR-2 + ResNet9. Each setting uses 5000 training examples and 500 held-out test examples, so the empirical Fisher has rank at most $r \le 5000$.

Firstly, we show the effective dimension $d_\lambda(F) = \sum_{i=1}^r \lambda_i/(\lambda_i + \lambda)$ can be much smaller than $r = \mathrm{rank}(F)$. Specifically, Figure 1 plots the ordered eigenvalues $\{\lambda_i\}_{i=1}^r$ of F. The spectrum decays quickly, hence for moderate λ, the terms $\lambda_i/(\lambda_i + \lambda)$ become small for large i, and consequently $d_\lambda(F)$ can be far smaller than r.

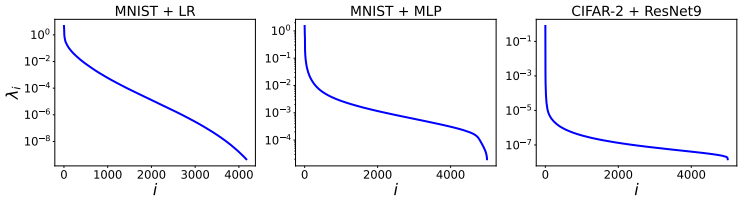

Figure 1: Ordered spectrum λ_i of the empirical Fisher F.

We next test the predictions of Theorems 2.2 and 3.1 by directly measuring the approximation error. Given $\lambda \ge 0$, we consider $\varepsilon_\lambda(g, g') = |\widetilde{\tau}_\lambda(g, g') - \tau_\lambda(g, g')| / \sqrt{\tau_0(g, g)}\sqrt{\tau_0(g', g')}$ for gradients g and g', which is the normalized error considered in Theorem 2.2.

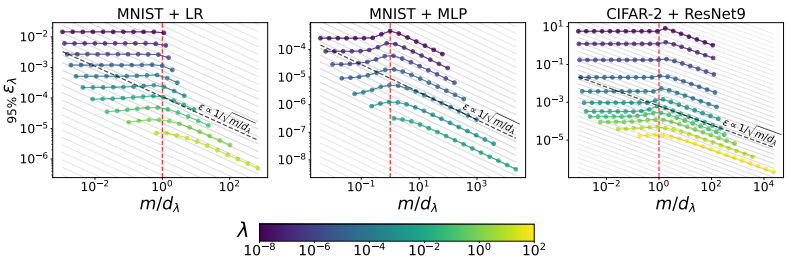

Figure 2: Approximation error versus normalized sketch size.

Figure 2 supports the scaling predicted by our theory. Each curve plots the 95$^{\text{th}}$ percentile of $\varepsilon_\lambda(g, g')$ against the normalized sketch size $m/d_\lambda(F)$. Once m is on the order of $d_\lambda(F)$, the error begins to decay in the manner suggested

by Theorem 2.2. Empirically, this indicates that (i) the hidden constant in the sketch-size requirement is modest and (ii) the additional leakage effect from Theorem 3.1 decreases quickly as m grows.

Faithfulness–Utility Tradeoff. A small approximation error does not necessarily imply strong downstream performance. In particular, optimizing ε_λ to be very small typically favors larger λ and larger sketch size m, because stronger regularization makes the influence computation less sensitive to sketching. As a result, the λ that minimizes ε_λ need not be the λ that maximizes downstream utility, especially when the curvature information in F is important for the task. We illustrate this using LDS (Park et al., 2023), a standard metric in data attribution. Figure 3 reports LDS over a range of sketch sizes and regularization strengths, and as we expect, the best-performing λ^* is typically intermediate.

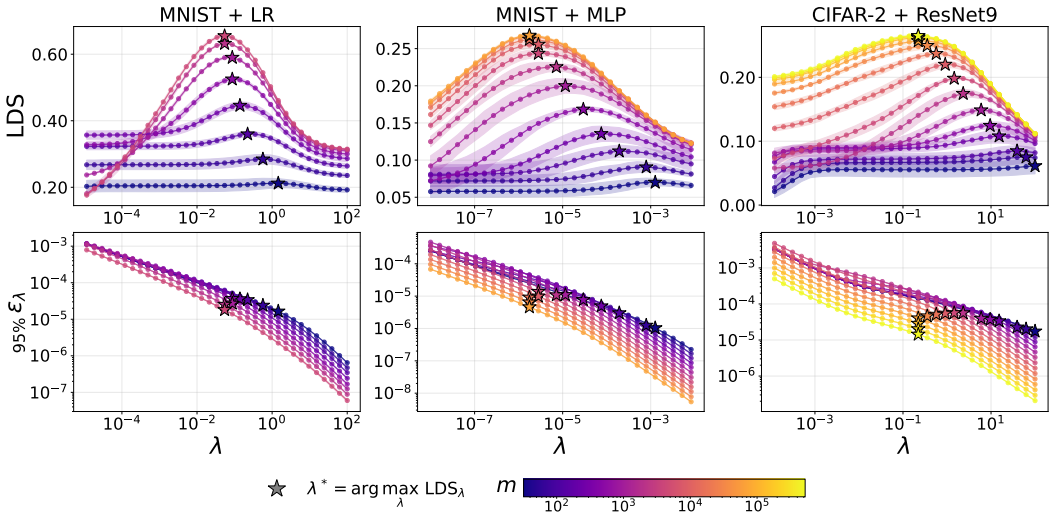

Figure 3: Approximation error and LDS versus λ.

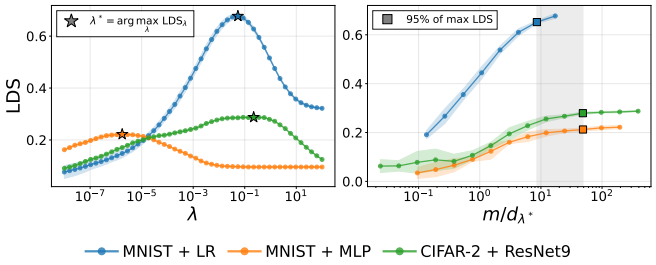

Figure 4: Left: selecting λ^* on a validation set using large m. Right: held-out test LDS versus $m/d_{\lambda^*}(F)$.

These observations suggest a simple two-stage procedure. First, using a small validation set and a sufficiently large sketch size m, sweep over λ and select λ^* that maximizes the downstream metric. Second, fix $\lambda = \lambda^*$ and increase m until $m \gtrsim C d_{\lambda^*}(F)$, which ensures that the influence estimates are faithful. Figure 4 illustrates this strategy for LDS: the square markers in the right panel (95^{th} percentile LDS) indicate how large m must be to approach the best attainable LDS. In our experiments, a constant $C \in (10, 100)$ is sufficient, making the dependence on $d_{\lambda^*}(F)$ operational.

5 Conclusion

In this work, we show that projection-based influence is governed by the interaction between the sketch and the curvature operator, and that conventional JL arguments, which only control Euclidean geometry, are inapplicable (Park et al., 2023; Schioppa, 2024; Hu et al., 2025). By precisely characterizing how projection interacts with common heuristics such as regularization and structured curvature approximations, our unified theory provides principled and actionable guidance for applying influence functions reliably at scale.

REFERENCES

Nir Ailon and Bernard Chazelle. The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on computing*, 39(1):302–322, 2009.

Juhan Bae, Nathan Hoyen Ng, Alston Lo, Marzyeh Ghassemi, and Roger Baker Grosse. If influence functions are the answer, then what is the question? In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=hzbguA9zMJ.

Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, et al. What is your data worth to gpt? llm-scale data valuation with influence functions. *arXiv preprint arXiv:2405.13954*, 2024.

Michael B Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*, pp. 278–287. SIAM, 2016.

Michael B Cohen, Jelani Nelson, and David P Woodruff. Optimal approximate matrix product in terms of stable rank. In *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*, pp. 11–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2016.

Junwei Deng, Ting-Wei Li, Shiyuan Zhang, Shixuan Liu, Yijun Pan, Hao Huang, Xinhe Wang, Pingbang Hu, Xingjian Zhang, and Jiaqi Ma. dattri: A library for efficient data attribution. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 136763–136781. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f732683302d91e47610b2416b4977a66-Paper-Datasets_and_Benchmarks_Track.pdf.

Junwei Deng, Yuzheng Hu, Pingbang Hu, Ting-Wei Li, Shixuan Liu, Jiachen T. Wang, Dan Ley, Qirun Dai, Benhao Huang, Jin Huang, Cathy Jiao, Hoang Anh Just, Yijun Pan, Jingyan Shen, Yiwen Tu, Weiyi Wang, Xinhe Wang, Shichang Zhang, Shiyuan Zhang, Ruoxi Jia, Himabindu Lakkaraju, Hao Peng, Weijing Tang, Chenyan Xiong, Jieyu Zhao, Hanghang Tong, Han Zhao, and Jiaqi W. Ma. A survey of data attribution: Methods, applications, and evaluation in the era of generative ai. *SSRN*, 2025. doi: 10.2139/ssrn.5451054. Available at SSRN: https://ssrn.com/abstract=5451054.

Sjoerd Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20(53):1–29, 2015.

Thomas George, César Laurent, Xavier Bouthillier, Nicolas Ballas, and Pascal Vincent. Fast approximate natural gradient descent in a kronecker factored eigenbasis. *Advances in neural information processing systems*, 31, 2018.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023.

Stephen F Gull. Developments in maximum entropy data analysis. In *Maximum Entropy and Bayesian Methods: Cambridge, England, 1988*, pp. 53–71. Springer, 1989.

Han Guo, Nazneen Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. Fastif: Scalable influence functions for efficient model interpretation and debugging. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10333–10350, 2021.

Zayd Hammoudeh and Daniel Lowd. Training data influence analysis and estimation: A survey. *Machine Learning*, 113(5):2351–2403, 2024.

Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

Pingbang Hu, Joseph Melkonian, Weijing Tang, Han Zhao, and Jiaqi W. Ma. Grass: Scalable data attribution with gradient sparsification and sparse projection. In *Advances in Neural Information Processing Systems*, 2025.

Yuzheng Hu, Pingbang Hu, Han Zhao, and Jiaqi Ma. Most influential subset selection: Challenges, promises, and beyond. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=qWi33pPecC.

Daniel M Kane and Jelani Nelson. Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM)*, 61(1):1–23, 2014.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. Datainf: Efficiently estimating data influence in loRA-tuned LLMs and diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=9m02ib92Wz.

W Johnson J Lindenstrauss and J Johnson. Extensions of lipschitz maps into a hilbert space. *Contemp. Math*, 26(189-206):2, 1984.

David MacKay. Bayesian model comparison and backprop nets. *Advances in neural information processing systems*, 4, 1991.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2408–2417, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/martens15.html.

Bruno Kacper Mlodozeniec, Runa Eschenhagen, Juhan Bae, Alexander Immer, David Krueger, and Richard E. Turner. Influence functions for scalable data attribution in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=esYrEndGsr.

Jelani Nelson and Huy L Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 ieee 54th annual symposium on foundations of computer science*, pp. 117–126. IEEE, 2013.

Raymond EAC Paley and Antoni Zygmund. On some series of functions,(3). In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 28, pp. 190–205. Cambridge University Press, 1932.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. In *International Conference on Machine Learning*, pp. 27074–27113. PMLR, 2023.

Andrea Schioppa. Efficient sketches for training data attribution and studying the loss landscape. *Advances in Neural Information Processing Systems*, 37:37692–37735, 2024.

Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. Scaling up influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8179–8186, 2022.

Stefano Teso, Andrea Bontempelli, Fausto Giunchiglia, and Andrea Passerini. Interactive label cleaning with example-based explanations. *Advances in Neural Information Processing Systems*, 34:12966–12977, 2021.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Weiyi Wang, Junwei Deng, Yuzheng Hu, Shiyuan Zhang, Xirui Jiang, Runting Zhang, Han Zhao, and Jiaqi W. Ma. Taming hyperparameter sensitivity in data attribution: Practical selection without costly retraining. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=qVDEM93mCP.

Mike Wojnowicz, Ben Cruz, Xuan Zhao, Brian Wallace, Matt Wolff, Jay Luan, and Caleb Crable. "influence sketching": Finding influential samples in large-scale regressions. In *2016 IEEE International Conference on Big Data (Big Data)*, pp. 3601–3612. IEEE, 2016.

David P Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Xiaosen Zheng, Tianyu Pang, Chao Du, Jing Jiang, and Min Lin. Intriguing properties of data attribution on diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vKViCoKGcB.

CONTENTS

A Proofs for Section 2.1 (Unregularized Projection)

In this section, we prove Theorem 2.1, which we first repeat the statement for convenience:

Theorem. *The equality $\tau_0(g, g') = \widetilde{\tau}_0(g, g')$ holds for any $g, g' \in \mathrm{range}(F)$* iff *$P$ is injective on $\mathrm{range}(F)$, i.e. $\mathrm{rank}(PU) = \mathrm{rank}(F) = r$ where $F = U\Lambda U^\top$ is the compact eigendecomposition of F with $U \in \mathbb{R}^{d\times r}$ orthonormal and $\Lambda \in \mathbb{R}^{r\times r}$ positive definite. Subsequently, for any PSD $F \in \mathbb{R}^{d\times d}$ and **any** matrix $P \in \mathbb{R}^{m\times d}$, one cannot hope to obtain any multiplicative approximation of $\tau_0(g, g')$ via $\widetilde{\tau}_0(g, g')$ when $\mathrm{rank}(PU) < r$.*

Proof. For the "if" direction, suppose $\mathrm{rank}(PU) = r$. Let $A := PU\Lambda^{1/2} \in \mathbb{R}^{m\times r}$ and it follows that A has full column rank. Then for any $g \in \mathrm{range}(U) = \mathrm{range}(F)$, write $g = Uz$ and $g'Uz'$ for some $z, z' \in \mathbb{R}^r$ and note $Pg = PUz = A\Lambda^{-1/2}z$ and similarly, $Pg' = A\Lambda^{-1/2}z'$, and $PFP^\top = AA^\top$. For full-column-rank A, $A^\top(AA^\top)^\dagger A = I_r$. Therefore

$$(Pg)^\top(PFP^\top)^\dagger(Pg') = z^\top\Lambda^{-1/2}A^\top(AA^\top)^\dagger A\Lambda^{-1/2}z' = z^\top\Lambda^{-1}z' = g^\top F^\dagger g'.$$

For the "only if" direction, suppose $\mathrm{rank}(PU) < r$. Then there exists a nonzero $z \in \mathbb{R}^r$ such that $PUz = 0$. Let $g = Uz \in \mathrm{range}(F)$ be the corresponding vector. Then, as $g^\top F^\dagger g = z^\top\Lambda^{-1}z > 0$, $(Pg)^\top(PFP^\top)^\dagger(Pg) = 0 \neq g^\top F^\dagger g > 0$, proving the result. □

B Proofs for Section 2.2 (Regularized Projection)

This section collects technical results used in Section 2.2 that are omitted in the main text.

B.1 Proof of Resolvent Perturbation Concentration for Regularized Projection

We prove the key operator-norm perturbation step used in the proof of Theorem 2.2.[4] The general idea is to use the concentration of the sample covariance (Theorem B.2) to control the resolvent-type map $A \mapsto A(A + \lambda I)^{-1}$ in operator norm, enabling the comparison of $F(F + \lambda I)^{-1}$ and $G(G + \lambda I)^{-1}$ in the proof of Theorem 2.2.

To prove Theorem B.2, the key input is a standard high-probability covariance estimation bound for sub-Gaussian vectors (Vershynin (2018, Exercise 9.2.5)), which we restate and prove as Theorem B.1.

Proposition B.1 (High-Probability Covariance Estimation)**.** *Let $\Sigma \succeq 0$ and let $X, X_1, \ldots, X_m \in \mathbb{R}^d$ be i.i.d. mean-zero sub-Gaussian random vectors with covariance $\Sigma = \mathbb{E}[XX^\top]$. Define the sample covariance*

$$\Sigma_m := \frac{1}{m}\sum_{i=1}^m X_i X_i^\top.$$

Then for any $u \geq 0$, with probability at least $1 - 2e^{-u}$,

$$\|\Sigma_m - \Sigma\|_2 \leq C\left(\sqrt{\frac{r(\Sigma) + u}{m}} + \frac{r(\Sigma) + u}{m}\right)\|\Sigma\|_2,$$

where $r(\Sigma) := \mathrm{tr}(\Sigma)/\|\Sigma\|_2$ is the stable rank of $\Sigma^{1/2}$ and $C > 0$ is a universal constant.

Proof. Write $X = \Sigma^{1/2}Z$, where Z is an isotropic, mean-zero, sub-Gaussian random vector, and similarly $X_i = \Sigma^{1/2}Z_i$ with i.i.d. copies $Z_1, \ldots, Z_m$. Let $A \in \mathbb{R}^{m\times d}$ be the matrix whose i-th row

[4] This can be viewed as a special case of approximate matrix multiplication for sub-Gaussian sketches; see Cohen et al. (2016, Theorem 1). Here, we state and prove the special case for clarity.

is $Z_i^\top$. As in the proof of Vershynin (2018, Theorem 9.2.4), define $T \coloneqq \Sigma^{1/2} S^{d-1}$ where S^{d-1} denotes the Euclidean unit sphere, then

$$\|\Sigma_m - \Sigma\|_2 = \frac{1}{m} \sup_{x \in T} \left| \|Ax\|_2^2 - m\|x\|_2^2 \right|.$$

Consider the stochastic process

$$Y_x \coloneqq \|Ax\|_2 - \sqrt{m}\|x\|_2, \qquad x \in T.$$

By Vershynin (2018, Theorem 9.1.3), $(Y_x)_{x \in T}$ has sub-Gaussian increments. Applying the high-probability Talagrand comparison inequality (Dirksen, 2015, Theorem 3.2), we obtain that with probability at least $1 - 2e^{-v^2}$,

$$\sup_{x \in T} |Y_x| \le C \left(\gamma(T) + v \operatorname{rad}(T) \right),$$

where $\operatorname{rad}(T) \coloneqq \sup_{x \in T} \|x\|_2$ denotes the *radius* of T, and $\gamma(T) \coloneqq \mathbb{E}\big[\sup_{x \in T} |\langle g, x \rangle| \big]$ denotes the *Gaussian complexity* of T, for $g \sim \mathcal{N}(0, I_d)$.

Since $T = \Sigma^{1/2} S^{d-1}$, we have $\operatorname{rad}(T) = \|\Sigma\|_2^{1/2}$. Moreover,

$$\gamma(T) = \mathbb{E}\big[\|\Sigma^{1/2} g\|_2 \big] \le \sqrt{\mathbb{E}[g^\top \Sigma g]} = \sqrt{\mathbb{E}[\operatorname{tr}(\Sigma g g^\top)]} = \sqrt{\operatorname{tr}(\Sigma)},$$

where the inequality follows from Jensen's inequality. Setting $u = v^2$ and recalling that $\operatorname{tr}(\Sigma) = r(\Sigma)\|\Sigma\|_2$, we conclude that, with probability at least $1 - 2e^{-u}$,

$$\sup_{x \in T} |Y_x| \le C \|\Sigma\|_2^{1/2} (\sqrt{r(\Sigma)} + \sqrt{u}).$$

Fix $x \in T$ and write $a \coloneqq \|Ax\|_2$ and $b \coloneqq \sqrt{m}\|x\|_2$. Then $b \le \sqrt{m}\|\Sigma\|_2^{1/2}$ and

$$|a^2 - b^2| \le |a - b|(|a - b| + 2b).$$

Using the bound above on $|a - b|$ and the fact that $b \ge 0$, we obtain

$$\sup_{x \in T} |a^2 - b^2| \le C \|\Sigma\|_2 (\sqrt{r(\Sigma)} + \sqrt{u}) \left(\sqrt{r(\Sigma)} + \sqrt{u} + \sqrt{m} \right).$$

Dividing by m yields

$$\|\Sigma_m - \Sigma\|_2 \le C \|\Sigma\|_2 \left(\frac{r(\Sigma) + u}{m} + \sqrt{\frac{r(\Sigma) + u}{m}} \right),$$

where we used $(\sqrt{r(\Sigma)} + \sqrt{u})^2 \lesssim r(\Sigma) + u$. This completes the proof. □

We now prove the concentration of sample covariance formally.

Lemma B.2. *Let $P \in \mathbb{R}^{m \times d}$ be a sketching matrix whose rows are given by $P_i^\top = \frac{1}{\sqrt{m}} W_i^\top$, where $\{W_i\}_{i=1}^m \sim W$ are i.i.d. sub-Gaussian random vectors in $\mathbb{R}^d$ satisfying $\mathbb{E}[W] = 0$ and $\mathbb{E}[WW^\top] = I_d$. Let $M \in \mathbb{R}^{d \times s}$ be a matrix and define $\Sigma \coloneqq M^\top M$. For any $\varepsilon, \delta \in (0, 1)$, if*

$$m = \Omega \left(\frac{r(\Sigma) + \log(1/\delta)}{\varepsilon^2} \right),$$

where $r(\Sigma) = \operatorname{tr}(\Sigma)/\|\Sigma\|_2$ is the stable rank of $\Sigma^{1/2}$, then with probability at least $1 - \delta$,

$$\|M^\top (P^\top P - I_d) M\|_2 \le \varepsilon \|M\|_2^2.$$

Proof. The rows of P satisfy $P_i^\top = \frac{1}{\sqrt{m}} X_i^\top$, where $\{X_i\}_{i=1}^m$ are i.i.d. isotropic sub-Gaussian vectors. Observe that

$$M^\top P^\top P M = M^\top \left(\frac{1}{m} \sum_{i=1}^m X_i X_i^\top \right) M = \frac{1}{m} \sum_{i=1}^m (M^\top X_i)(M^\top X_i)^\top =: \Sigma_m.$$

Define $Y_i \coloneqq M^\top X_i$. Then $\{Y_i\}_{i=1}^m$ are i.i.d. mean-zero sub-Gaussian vectors with covariance

$$\mathbb{E}[YY^\top] = M^\top \mathbb{E}[XX^\top]M = M^\top M = \Sigma.$$

Applying Vershynin (2018, Exercise 9.2.5, Theorem B.1) yields that, with probability at least $1 - 2e^{-u}$,

$$\|\Sigma_m - \Sigma\|_2 \le C\left(\sqrt{\frac{r(\Sigma)+u}{m}} + \frac{r(\Sigma)+u}{m}\right)\|\Sigma\|_2,$$

Choosing $m \ge (r(\Sigma)+u)/\varepsilon^2$ ensures $\sqrt{(r(\Sigma)+u)/m} \le \varepsilon$ and $(r(\Sigma)+u)/m \le \varepsilon^2 < \varepsilon$ for $\varepsilon < 1$. Since $\|\Sigma\|_2 = \|M^\top M\|_2 = \|M\|_2^2$, we conclude that

$$\|M^\top P^\top PM - M^\top M\|_2 = \|M^\top (P^\top P - I_d)M\|_2 \le \varepsilon \|M\|_2^2.$$

Setting $u = \Theta(\log(1/\delta))$ completes the proof. □

We can now state and prove the concentration of resolvent perturbation for regularized projection as follows:

Lemma B.3. *Let $F \succeq 0$ and $\lambda > 0$, and define $G = F^{1/2}P^\top PF^{1/2}$. Then for any $\varepsilon, \delta \in (0,1)$, if $m = \Omega(\varepsilon^{-2}(d_\lambda(F) + \log(1/\delta)))$, with probability at least $1-\delta$,*

$$\|F(F+\lambda I)^{-1} - G(G+\lambda I)^{-1}\|_2 \le \varepsilon.$$

Proof. Applying Theorem B.2 with $M = B = F^{1/2}(F+\lambda I)^{-1/2}$, for any $\delta, \epsilon > 0$, if $m = \Omega(\epsilon^{-2}(r(B^\top B) + \log(1/\delta)))$ then with probability at least $1-\delta$,

$$\|B^\top (P^\top P - I_d)B\|_2 \le \epsilon \|B\|_2^2.$$

We first note that if $\|B\|_2^2 = 0$, then the bound is trivial. Assuming $\|B\|_2 > 0$. Then we see that $\|B\|_2^2 = \|B^\top B\|_2 = \|F(F+\lambda I)^{-1}\|_2 \le 1$ since the eigenvalues of $F(F+\lambda I)^{-1}$ equal $\lambda_i(F)/(\lambda_i(F)+\lambda)$. Now, pick $\epsilon \coloneqq \min(1, \varepsilon/2\|B\|_2)$, and note that $\|B\|_F^2 = \mathrm{tr}(F(F+\lambda I)^{-1}) = d_\lambda(F)$, we have

$$r(B^\top B) = \frac{\mathrm{tr}(B^\top B)}{\|B^\top B\|_2} = \frac{\|B\|_F^2}{\|B\|_2^2} = \frac{d_\lambda(F)}{\|B\|_2^2}.$$

After substitution, with $\|B\|_2^2 \le 1$, we conclude that if

$$m = \Omega\left(\epsilon^{-2}\left(\frac{d_\lambda(F)}{\|B\|_2^2} + \log(1/\delta)\right)\right) = \Omega\left(\varepsilon^{-2}\left(d_\lambda(F) + \log(1/\delta)\right)\right),$$

we have $\|B^\top (P^\top P - I_d)B\|_2 \le \epsilon\|B\|_2^2 \le \varepsilon/2$. This implies

$$-\frac{\varepsilon}{2}I \preceq B^\top (P^\top P - I)B \preceq \frac{\varepsilon}{2}I \implies B^\top B - \frac{\varepsilon}{2}I \preceq B^\top P^\top PB \preceq B^\top B + \frac{\varepsilon}{2}I.$$

With

$$B^\top B = (F+\lambda I)^{-1/2}F(F+\lambda I)^{-1/2},$$
$$B^\top P^\top PB = (F+\lambda I)^{-1/2}\underbrace{F^{1/2}P^\top PF^{1/2}}_{G}(F+\lambda I)^{-1/2},$$

we can conjugate by $(F+\lambda I)^{1/2}$, which yields

$$\left(1-\frac{\varepsilon}{2}\right)F - \frac{\varepsilon}{2}\lambda I \preceq G \preceq \left(1+\frac{\varepsilon}{2}\right)F + \frac{\varepsilon}{2}\lambda I.$$

Adding λI gives

$$\left(1-\frac{\varepsilon}{2}\right)(F+\lambda I) \preceq G + \lambda I \preceq \left(1+\frac{\varepsilon}{2}\right)(F+\lambda I).$$

Define $S \coloneqq (F+\lambda I)^{-1/2}(G+\lambda I)(F+\lambda I)^{-1/2}$. Conjugating the above by $(F+\lambda I)^{-1/2}$ yields

$$\left(1-\frac{\varepsilon}{2}\right)I \preceq S \preceq \left(1+\frac{\varepsilon}{2}\right)I.$$

Hence, $S \succ 0$ and $\|S - I\|_2 \le \varepsilon/2$ and $\|S^{-1}\|_2 \le \frac{1}{1-\varepsilon/2}$. From the definition of S,

$$(G + \lambda I)^{-1} = (F + \lambda I)^{-1/2} S^{-1} (F + \lambda I)^{-1/2},$$

hence

$$(G + \lambda I)^{-1} - (F + \lambda I)^{-1} = (F + \lambda I)^{-1/2} (S^{-1} - I)(F + \lambda I)^{-1/2},$$

giving

$$\|(G + \lambda I)^{-1} - (F + \lambda I)^{-1}\|_2 \le \|(F + \lambda I)^{-1}\|_2 \|S^{-1} - I\|_2.$$

From the identity $S^{-1} - I = S^{-1}(I - S)$, we have

$$\|S^{-1} - I\|_2 \le \|S^{-1}\|_2 \|S - I\|_2 \le \frac{\varepsilon/2}{1 - \varepsilon/2}.$$

With $\|(F + \lambda I)^{-1}\|_2 \le 1/\lambda$, we have

$$\|(G + \lambda I)^{-1} - (F + \lambda I)^{-1}\|_2 \le \frac{1}{\lambda} \frac{\varepsilon/2}{1 - \varepsilon/2}.$$

From the identity $A(A + \lambda I)^{-1} = I - \lambda(A + \lambda I)^{-1}$ for any PSD A, we have

$$F(F + \lambda I)^{-1} - G(G + \lambda I)^{-1} = \lambda \left((G + \lambda I)^{-1} - (F + \lambda I)^{-1} \right),$$

and hence

$$\|F(F + \lambda I)^{-1} - G(G + \lambda I)^{-1}\|_2 \le \lambda \frac{1}{\lambda} \frac{\varepsilon/2}{1 - \varepsilon/2} = \frac{\varepsilon/2}{1 - \varepsilon/2}.$$

Finally, note that $\frac{\varepsilon/2}{1-\varepsilon/2} \le \varepsilon$ for any $\varepsilon \in (0, 1)$, this proves the result. □

B.2 OSE-Based Alternative Analysis

We record a self-contained proof of the OSE-based alternative analysis sketched in discussion following Theorem 2.2. Let $A \in \mathbb{R}^{d \times r}$ be a fixed matrix. A random matrix $P \in \mathbb{R}^{m \times d}$ is an *oblivious subspace embedding (OSE)* for $\mathrm{range}(A)$ with distortion $\varepsilon \in (0, 1)$ if, with high probability,

$$(1 - \varepsilon)\|Ax\|_2^2 \le \|PAx\|_2^2 \le (1 + \varepsilon)\|Ax\|_2^2, \qquad \forall x \in \mathbb{R}^r.$$

Equivalently,

$$-\varepsilon A^\top A \preceq A^\top (P^\top P - I_d) A \preceq \varepsilon A^\top A.$$

It is well known that standard oblivious sketches (Gaussian, Rademacher, SJLT) satisfy this property provided $m = \Omega\big(\varepsilon^{-2} \mathrm{rank}(A)\big)$ (Woodruff, 2014, Theorems 2.3 and 6.10). In our case, we apply the OSE framework with $A = F^{1/2}$. It is straightforward to see that $\mathrm{rank}(A) = \mathrm{rank}(F) = r$, so achieving an ε-OSE for $\mathrm{range}(A)$ requires $m = \Omega(r/\varepsilon^2)$.

Define $G \coloneqq F^{1/2} P^\top P F^{1/2}$. The OSE condition gives

$$(1 - \varepsilon)F \preceq G \preceq (1 + \varepsilon)F.$$

Consider $f(t) \coloneqq \frac{t}{t+\lambda}$ for $t \ge 0$. Since $t \mapsto (t + \lambda)^{-1}$ is operator monotone decreasing on $[0, \infty)$, it follows that $f(t) = 1 - \lambda(t + \lambda)^{-1}$ is operator monotone increasing.

Applying f to the sandwich gives

$$f\big((1 - \varepsilon)F\big) \preceq f(G) \preceq f\big((1 + \varepsilon)F\big),$$

or

$$(1 - \varepsilon)F\big((1 - \varepsilon)F + \lambda I\big)^{-1} \preceq G(G + \lambda I)^{-1} \preceq (1 + \varepsilon)F\big((1 + \varepsilon)F + \lambda I\big)^{-1}.$$

Since F commutes with any function of itself, the resulting operator-norm deviation reduces to a scalar supremum. For example,

$$\big\|f\big((1 + \varepsilon)F\big) - f(F)\big\|_2 = \sup_{t \ge 0} \left| \frac{(1 + \varepsilon)t}{(1 + \varepsilon)t + \lambda} - \frac{t}{t + \lambda} \right| = \sup_{t \ge 0} \frac{\varepsilon \lambda t}{\big((1 + \varepsilon)t + \lambda\big)(t + \lambda)}.$$

The same bound holds with $(1 + \varepsilon)$ replaced by $(1 - \varepsilon)$. A short calculus argument shows the supremum is at most ε; hence

$$\big\|f(G) - f(F)\big\|_2 = \big\|G(G + \lambda I)^{-1} - F(F + \lambda I)^{-1}\big\|_2 \le O(\varepsilon).$$

Combining the above operator control with the argument in the proof of Theorem 2.2 yields the same bilinear and quadratic influence error bounds. The key difference is the sample complexity: the OSE route fundamentally scales with r, whereas our main analysis scales with the effective dimension $d_\lambda(F)$.

B.3 Proof of Anti-Concentration of Gaussian Sample Covariance

Next, we prove the worst-case lower bound (Theorem 2.4). The proof consists of two main components:

1. An anti-concentration result for the sample covariance of Gaussian matrices, which shows that deviations of order $\sqrt{k/m}$ occur with constant probability (Theorem B.4).
2. A carefully constructed hard instance F for which such deviations translate directly into a large error in the regularized quadratic form.

We note that since the proof of Theorem B.4 contains many technical computation, we defer them for a cleaner presentation after the main proof.

Lemma B.4. *Let $W \in \mathbb{R}^{m\times k}$ have rows $w_1, \dots, w_m \sim \mathcal{N}(0, I_k)$ i.i.d., and define $S \coloneqq \frac{1}{m} W^\top W$. Then for all $m, k \geq 1$,*

$$\Pr\left(\|S - I_k\|_2 \geq \frac{1}{2}\sqrt{\frac{k}{m}} \right) \geq \frac{3}{80}.$$

Proof. Define

$$A \coloneqq S - I_k = \frac{1}{m}\sum_{i=1}^{m} X_i, \qquad X_i \coloneqq w_i w_i^\top - I_k.$$

Then $\mathbb{E}[X_i] = 0$ and $X_1, \dots, X_m$ are independent. Let $g \coloneqq \|A\|_F^2 \geq 0$. Expanding, we have

$$g = \left\| \frac{1}{m}\sum_{i=1}^{m} X_i \right\|_F^2 = \frac{1}{m^2}\sum_{i,j=1}^{m} \langle X_i, X_j \rangle.$$

Since $\mathbb{E}[\langle X_i, X_j\rangle] = 0$ for $i \neq j$ from independence and $\mathbb{E}[X_i] = 0$,

$$\mathbb{E}[g] = \frac{1}{m^2}\sum_{i=1}^{m} \mathbb{E}[\|X_i\|_F^2] = \frac{1}{m}\mathbb{E}[\|X_1\|_F^2].$$

A direct computation gives $\mathbb{E}[\|X_1\|_F^2] = k(k+1)$, hence

$$\mathbb{E}[g] = \frac{k(k+1)}{m}.$$

On the other hand, as $g = \frac{1}{m^2}\sum_{i,j} Y_{ij}$ where $Y_{ij} \coloneqq \langle X_i, X_j\rangle$, we have

$$\mathbb{E}[g^2] = \frac{1}{m^4}\sum_{i,j,p,q} \mathbb{E}[Y_{ij}Y_{pq}].$$

By independence and centering, only overlapping index patterns contribute, and one obtains

$$\mathbb{E}[g^2] = \frac{1}{m^4}\Big(ma + m(m-1)\mu^2 + 2m(m-1)b \Big),$$

where $\mu \coloneqq \mathbb{E}[\|X_1\|_F^2]$, $a \coloneqq \mathbb{E}[\|X_1\|_F^4]$, and $b \coloneqq \mathbb{E}[\langle X, Y\rangle^2]$, and $X \coloneqq ww^\top - I_k$, $Y \coloneqq uu^\top - I_k$ with $w \perp u$ i.i.d. $\mathcal{N}(0, I_k)$. Moreover, a moment calculations yield

$$\mu = k(k+1), \qquad a = k^4 + 10k^3 + 25k^2 + 24k, \qquad b = 2k^2 + 2k.$$

Substituting these expressions into the above formula for $\mathbb{E}[g^2]$ and simplifying gives the explicit comparison

$$\frac{(\mathbb{E}[g])^2}{\mathbb{E}[g^2]} \geq \frac{1}{15},$$

uniformly for all $m, k \geq 1$. Equivalently, $\mathbb{E}[g^2] \leq 15(\mathbb{E}[g])^2$. Then, by Paley–Zygmund (Paley & Zygmund, 1932), for any $\theta \in (0, 1)$,

$$\Pr(g \geq \theta\mathbb{E}[g]) \geq (1-\theta)^2 \frac{(\mathbb{E}[g])^2}{\mathbb{E}[g^2]} \geq \frac{(1-\theta)^2}{15}.$$

Taking $\theta = 1/4$ yields

$$\Pr\left(g \geq \frac{1}{4}\mathbb{E}[g]\right) \geq \frac{(3/4)^2}{15} = \frac{3}{80}.$$

On this event,

$$g = \|A\|_F^2 \geq \frac{1}{4} \cdot \frac{k(k+1)}{m}.$$

Using $\|A\|_F^2 \leq k\|A\|_2^2$, we obtain

$$\|A\|_2^2 \geq \frac{1}{k}\|A\|_F^2 \geq \frac{1}{k} \cdot \frac{1}{4} \cdot \frac{k(k+1)}{m} = \frac{k+1}{4m} \geq \frac{k}{4m}.$$

Hence, with probability at least $3/80$,

$$\|A\|_2 = \|S - I_k\|_2 \geq \frac{1}{2}\sqrt{\frac{k}{m}}.$$

□

We now provide the routine calculations used in the proof of Theorem B.4. In particular, we compute moments of Gaussian rank-one matrices and enumerate the index patterns in $\mathbb{E}[\|A\|_F^4]$.

Chi-square moments. Let $r \sim \chi_k^2$. For any $n \in \mathbb{N}_+$,

$$\mathbb{E}[r^n] = \prod_{i=0}^{n-1}(k+2i).$$

In particular,

$$\mathbb{E}[r] = k, \quad \mathbb{E}[r^2] = k(k+2), \quad \mathbb{E}[r^3] = k(k+2)(k+4), \quad \mathbb{E}[r^4] = k(k+2)(k+4)(k+6).$$

Moments of $X = ww^\top - I_k$. Let $w \sim \mathcal{N}(0, I_k)$ and define $X \coloneqq ww^\top - I_k$. Write $r \coloneqq \|w\|_2^2 \sim \chi_k^2$. We compute $\mu \coloneqq \mathbb{E}[\|X\|_F^2]$ and $a \coloneqq \mathbb{E}[\|X\|_F^4]$. First,

$$\begin{aligned}\|X\|_F^2 &= \operatorname{tr}(X^\top X) = \operatorname{tr}(X^2) = \operatorname{tr}\left((ww^\top - I_k)^2\right) \\ &= \operatorname{tr}(ww^\top ww^\top) - 2\operatorname{tr}(ww^\top) + \operatorname{tr}(I_k).\end{aligned}$$

By trace cyclicity, $\operatorname{tr}(ww^\top ww^\top) = \operatorname{tr}(w(w^\top w)w^\top) = (w^\top w)\operatorname{tr}(ww^\top) = r^2$, while $\operatorname{tr}(ww^\top) = r$ and $\operatorname{tr}(I_k) = k$. Hence

$$\|X\|_F^2 = r^2 - 2r + k.$$

Taking expectation and using the moments above gives

$$\mu = \mathbb{E}[r^2 - 2r + k] = k(k+2) - 2k + k = k(k+1).$$

Moreover,

$$a = \mathbb{E}(r^2 - 2r + k)^2 = \mathbb{E}\left[r^4 - 4r^3 + (4+2k)r^2 - 4kr + k^2\right].$$

Substituting $\mathbb{E}[r], \ldots, \mathbb{E}[r^4]$ yields

$$a = k^4 + 10k^3 + 25k^2 + 24k.$$

The mixed term $b = \mathbb{E}[\langle X, Y\rangle^2]$. Let $w, u \sim \mathcal{N}(0, I_k)$ be independent, and define $X \coloneqq ww^\top - I_k$ and $Y \coloneqq uu^\top - I_k$. Set $r \coloneqq \|w\|_2^2$, $s \coloneqq \|u\|_2^2$, and $t \coloneqq w^\top u$. A direct expansion gives

$$\langle X, Y\rangle = \operatorname{tr}\left((ww^\top - I_k)(uu^\top - I_k)\right) = t^2 - r - s + k,$$

since $\operatorname{tr}(ww^\top uu^\top) = \operatorname{tr}(w(w^\top u)u^\top) = (w^\top u)^2 = t^2$. Therefore,

$$b = \mathbb{E}[(t^2 - r - s + k)^2] = \mathbb{E}[t^4] + \mathbb{E}[(r+s-k)^2] - 2\mathbb{E}\left[t^2(r+s-k)\right].$$

To evaluate these terms, write $t = \sum_{\ell=1}^k Z_\ell$ with $Z_\ell \coloneqq w_\ell u_\ell$. Then $\mathbb{E}[Z_\ell] = 0$, $\mathbb{E}[Z_\ell^2] = 1$, and $\mathbb{E}[Z_\ell^4] = 9$, and hence

$$\mathbb{E}[t^4] = \sum_{\ell=1}^k \mathbb{E}[Z_\ell^4] + 6\sum_{1 \leq i < j \leq k} \mathbb{E}[Z_i^2]\mathbb{E}[Z_j^2] = 9k + 6\binom{k}{2} = 3k^2 + 6k.$$

Next, since $r, s \sim \chi^2_k$ are independent, we have $\mathbb{E}[r] = \mathbb{E}[s] = k$ and $\mathrm{Var}[r] = \mathrm{Var}[s] = 2k$, so

$$\mathbb{E}[(r+s-k)^2] = \mathrm{Var}[r+s-k] + (\mathbb{E}[r+s-k])^2 = 4k + k^2.$$

Finally, conditioning on w gives $t \mid w \sim \mathcal{N}(0, \|w\|_2^2) = \mathcal{N}(0, r)$, so $\mathbb{E}[t^2 \mid w] = r$ and hence $\mathbb{E}[t^2] = \mathbb{E}[r] = k$. Moreover,

$$\mathbb{E}[t^2 r] = \mathbb{E}\big[r\mathbb{E}[t^2 \mid w]\big] = \mathbb{E}[r^2] = k(k+2).$$

By symmetry, $\mathbb{E}[t^2(r+s-k)] = 2\mathbb{E}[t^2 r] - k\mathbb{E}[t^2] = k^2 + 4k$, so altogether

$$b = (3k^2 + 6k) + (k^2 + 4k) - 2(k^2 + 4k) = 2k^2 + 2k.$$

Enumerating index patterns in $\mathbb{E}\|A\|_F^4$. Let $A = \frac{1}{m}\sum_{i=1}^m X_i$ with $X_i = w_i w_i^\top - I_k$ i.i.d. and mean-zero, and set $Z = \|A\|_F^2$. With $Y_{ij} \coloneqq \langle X_i, X_j\rangle$, we have

$$Z = \frac{1}{m^2}\sum_{i,j=1}^m Y_{ij}, \qquad Z^2 = \frac{1}{m^4}\sum_{i,j,p,q=1}^m Y_{ij}Y_{pq}, \qquad \mathbb{E}[Z^2] = \frac{1}{m^4}\sum_{i,j,p,q}\mathbb{E}[Y_{ij}Y_{pq}].$$

The expectation $\mathbb{E}[Y_{ij}Y_{pq}]$ is zero unless $\{i,j\} \cap \{p,q\} \neq \varnothing$. Indeed, if $\{i,j\} \cap \{p,q\} = \varnothing$, then the two factors depend on disjoint sets of independent random variables. Moreover, for $i \neq j$, $\mathbb{E}[Y_{ij}] = \mathbb{E}[\langle X_i, X_j\rangle] = \langle \mathbb{E}[X_i], \mathbb{E}[X_j]\rangle = 0$, so such disjoint products vanish. The only contributing configurations are:

(T1) $(i,j) = (p,q)$, contributing $\mathbb{E}[Y_{ij}^2]$;

(T2) $(i,j) = (q,p)$, contributing $\mathbb{E}[Y_{ij}Y_{ji}] = \mathbb{E}[Y_{ij}^2]$ since $Y_{ij} = Y_{ji}$;

(T3) $i = j$ and $p = q$ with $i \neq p$, contributing $\mathbb{E}[Y_{ii}]\mathbb{E}[Y_{pp}] = \mu^2$.

Counting multiplicities, type Item (T1) gives $\sum_{i,j}\mathbb{E}[Y_{ij}^2] = ma + m(m-1)b$, where $a = \mathbb{E}[\|X_1\|_F^4]$ (since $Y_{11} = \langle X_1, X_1\rangle = \|X_1\|_F^2$) and $b = \mathbb{E}[\langle X, Y\rangle^2]$ for independent copies X, Y. Type Item (T2) contributes another $m(m-1)b$, and type Item (T3) contributes $m(m-1)\mu^2$. Hence

$$\mathbb{E}[Z^2] = \frac{1}{m^4}\Big(ma + m(m-1)\mu^2 + 2m(m-1)b\Big).$$

Using $\mu = k(k+1)$, $a = k^4 + 10k^3 + 25k^2 + 24k$, and $b = 2k^2 + 2k$, one checks that for all $m, k \geq 1$,

$$\mathbb{E}[Z^2] \leq \frac{15k^2(k+1)^2}{m^2}.$$

B.4 Proof of Worst-Case Lower Bound

We restate Theorem 2.4 below for convenience:

Theorem. *Let $P \in \mathbb{R}^{m\times d}$ be a Gaussian oblivious sketch with rows i.i.d. $\mathcal{N}(0, I_d)$. There exists a family of matrices $F \in \mathbb{R}^{d\times d}$ such that if $m = o(d_\lambda(F)/\varepsilon^2)$, then with constant probability, there exists $g \in \mathrm{range}(F)$ such that*

$$|\widetilde{\tau}_\lambda(g,g) - \tau_\lambda(g,g)| = \Omega(\varepsilon)\tau_0(g,g).$$

Proof. Fix integers $k \leq r = \mathrm{rank}(F) \leq d$ and define

$$F = \mathrm{diag}(\underbrace{\lambda, \ldots, \lambda}_{k}, \underbrace{\eta\lambda, \ldots, \eta\lambda}_{r-k}, \underbrace{0, \ldots, 0}_{d-r}),$$

where $\eta > 0$ will be chosen sufficiently small (as a function of ε and fixed constants only). Then

$$d_\lambda(F) = \sum_{i=1}^d \frac{\lambda_i(F)}{\lambda_i(F) + \lambda} = \frac{k\lambda}{\lambda + \lambda} + \frac{(r-k)\eta\lambda}{\eta\lambda + \lambda} = \frac{k}{2} + \frac{\eta}{1+\eta}(r-k) = \Theta(k) \quad \text{for } \eta \ll 1.$$

Let $P = \frac{1}{\sqrt{m}} W$ where $W \in \mathbb{R}^{m \times d}$ has i.i.d. $\mathcal{N}(0,1)$ entries, and partition

$$P = (P_L, P_S, P_Z)$$

according to the blocks of F, i.e. $P_L \in \mathbb{R}^{m \times k}$, $P_S \in \mathbb{R}^{m \times (r-k)}$. Choose

$$g = F^{1/2} y, \qquad y = \begin{pmatrix} y_L \\ 0 \\ 0 \end{pmatrix}, \quad \|y_L\|_2 = 1,$$

for some y. We see that $g_L = \sqrt{\lambda} y_L$. Now, we see that

$$\tau_\lambda(g,g) = g^\top (F + \lambda I)^{-1} g = y^\top F^{1/2} (F + \lambda I)^{-1} F^{1/2} y = y_L^\top \frac{\lambda}{\lambda + \lambda} I_k y_L = \frac{1}{2},$$

and

$$\tau_0(g,g) = g^\top F^\dagger g = y^\top y = \|y_L\|_2^2 = 1.$$

On the other hand, the sketched quantity equals

$$\widetilde{\tau}_\lambda(g,g) = g^\top P^\top (PFP^\top + \lambda I)^{-1} P g.$$

Since $g = F^{1/2} y$ and $F = \lambda \operatorname{diag}(I_k, \eta I_{r-k}, 0)$, we have $Pg = \sqrt{\lambda} P_L y_L$ and

$$PFP^\top + \lambda I = \lambda \big(P_L P_L^\top + \eta P_S P_S^\top + I \big).$$

Therefore

$$\widetilde{\tau}_\lambda(g,g) = y_L^\top P_L^\top \big(P_L P_L^\top + \eta P_S P_S^\top + I \big)^{-1} P_L y_L.$$

Decomposing the error, write

$$\big| \widetilde{\tau}_\lambda(g,g) - \tau_\lambda(g,g) \big| \geq T_1 - T_2,$$

where

$$T_1 \coloneqq \left| y_L^\top P_L^\top (P_L P_L^\top + I)^{-1} P_L y_L - \frac{1}{2} \right|,$$

$$T_2 \coloneqq \left| y_L^\top P_L^\top \left[(P_L P_L^\top + I)^{-1} - (P_L P_L^\top + \eta P_S P_S^\top + I)^{-1} \right] P_L y_L \right|.$$

Lower-Bounding T_1. Let $M \coloneqq P_L^\top P_L \in \mathbb{R}^{k \times k}$. Using the push-through identity

$$P_L^\top (P_L P_L^\top + I)^{-1} P_L = M(M+I)^{-1},$$

we have

$$\left| y_L^\top P_L^\top (P_L P_L^\top + I)^{-1} P_L y_L - \frac{1}{2} \right| = \left| y_L^\top M (M+I)^{-1} y_L - \frac{1}{2} \right|.$$

Let $\lambda_1, \ldots, \lambda_k$ be the eigenvalues of M and choose y_L to be a unit eigenvector corresponding to an eigenvalue $\lambda_\star$. Then

$$y_L^\top M (M+I)^{-1} y_L = \frac{\lambda_\star}{\lambda_\star + 1},$$

and hence

$$\left| y_L^\top M (M+I)^{-1} y_L - \frac{1}{2} \right| = \left| \frac{\lambda_\star}{\lambda_\star + 1} - \frac{1}{2} \right| = \left| \frac{\lambda_\star - 1}{2(\lambda_\star + 1)} \right| = \frac{|\lambda_\star - 1|}{2(\lambda_\star + 1)}.$$

Using $\lambda_\star + 1 \leq |\lambda_\star - 1| + 2$, we obtain

$$\left| y_L^\top M (M+I)^{-1} y_L - \frac{1}{2} \right| \geq \frac{|\lambda_\star - 1|}{2|\lambda_\star - 1| + 4} \geq \min \left\{ \frac{|\lambda_\star - 1|}{8}, \frac{1}{4} \right\}.$$

Now observe that $\|M - I\|_2 = \max_i |\lambda_i - 1|$, so if $\|M - I\|_2 \geq t$, then there exists $\lambda_\star$ with $|\lambda_\star - 1| \geq t$ and the above choice of y_L yields

$$\left| y_L^\top M (M+I)^{-1} y_L - \frac{1}{2} \right| \geq \min \left\{ \frac{t}{8}, \frac{1}{4} \right\}.$$

Since $P_L = \frac{1}{\sqrt{m}} W_L$ with $W_L \in \mathbb{R}^{m\times k}$ i.i.d. Gaussian rows, we have

$$M = P_L^\top P_L = \frac{1}{m} W_L^\top W_L.$$

Applying Theorem B.4 to W_L gives

$$\Pr\left(\|M - I_k\|_2 \geq \frac{1}{2}\sqrt{\frac{k}{m}} \right) \geq \frac{3}{80}.$$

On this event we may take $t = \frac{1}{2}\sqrt{k/m}$ above, giving

$$\left| y_L^\top M(M+I)^{-1} y_L - \frac{1}{2} \right| \geq \min\left\{ \frac{1}{16}\sqrt{\frac{k}{m}}, \frac{1}{4} \right\},$$

with probability at least $3/80$.

Upper-Bounding T_2. Let $A \coloneqq P_L P_L^\top + I \succ 0$ and $B \coloneqq A + \eta P_S P_S^\top \succ 0$. By Woodbury matrix identity,

$$B^{-1} = (A + \eta P_S P_S^\top)^{-1} = A^{-1} - A^{-1} P_S (\eta^{-1} I + P_S^\top A^{-1} P_S)^{-1} P_S^\top A^{-1}.$$

Hence

$$A^{-1} - B^{-1} = A^{-1} P_S (\eta^{-1} I + P_S^\top A^{-1} P_S)^{-1} P_S^\top A^{-1}.$$

Using $|y_L^\top (\cdot) y_L| \leq \|\cdot\|_2$, we obtain

$$T_2 \leq \left\| P_L^\top (A^{-1} - B^{-1}) P_L \right\|_2 = \left\| P_L^\top A^{-1} P_S \left(\eta^{-1} I + P_S^\top A^{-1} P_S \right)^{-1} P_S^\top A^{-1} P_L \right\|_2.$$

Define

$$X \coloneqq A^{-1/2} P_L, \qquad C \coloneqq A^{-1/2} P_S.$$

Then $P_L^\top A^{-1} P_S = X^\top C$ and $P_S^\top A^{-1} P_S = C^\top C$, so

$$T_2 \leq \left\| X^\top C \left(\eta^{-1} I + C^\top C \right)^{-1} C^\top X \right\|_2 \leq \|X\|_2^2 \cdot \left\| C \left(\eta^{-1} I + C^\top C \right)^{-1} C^\top \right\|_2.$$

We claim $\|X\|_2 \leq 1$. Indeed,

$$X^\top X = P_L^\top A^{-1} P_L = P_L^\top (P_L P_L^\top + I)^{-1} P_L = M(M+I)^{-1} \preceq I,$$

where $M = P_L^\top P_L \succeq 0$, and the last inequality holds since the eigenvalues of $M(M+I)^{-1}$ are $\lambda/(\lambda+1) \in [0,1)$. Therefore $\|X\|_2^2 \leq 1$, and hence

$$T_2 \leq \left\| C \left(\eta^{-1} I + C^\top C \right)^{-1} C^\top \right\|_2.$$

Next, diagonalize $C^\top C$ and let $\sigma_{\max}^2 = \|C\|_2^2$ be its largest eigenvalue. The nonzero eigenvalues of $C(\eta^{-1} I + C^\top C)^{-1} C^\top$ are

$$\frac{\sigma_i^2}{\eta^{-1} + \sigma_i^2} = \frac{\eta \sigma_i^2}{1 + \eta \sigma_i^2},$$

so

$$\left\| C \left(\eta^{-1} I + C^\top C \right)^{-1} C^\top \right\|_2 = \frac{\eta \|C\|_2^2}{1 + \eta \|C\|_2^2} \leq \eta \|C\|_2^2.$$

Finally, since $A \succeq I$, we have $\|C\|_2 = \|A^{-1/2} P_S\|_2 \leq \|P_S\|_2$, and thus

$$T_2 \leq \frac{\eta \|P_S\|_2^2}{1 + \eta \|P_S\|_2^2} \leq \eta \|P_S\|_2^2.$$

It remains to control $\|P_S\|_2$. Since $P_S = \frac{1}{\sqrt{m}} W_S$ is Gaussian, standard spectral norm bounds imply that for any $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$\|P_S\|_2 \leq 1 + \sqrt{\frac{r-k}{m}} + \sqrt{\frac{\log(2/\delta)}{m}}.$$

In particular, if $r - k \leq m$ and $m \geq \log(2/\delta)$, then on this event $\|P_S\|_2 \leq 3$ and hence

$$T_2 \leq 9\eta.$$

Choosing Parameters. Fix $\delta := \frac{1}{160}$ and assume $r - k \le m$ and $m \ge \log(2/\delta)$. This is possible by choosing r appropriately, e.g., $r = k + m$ or $r = 2k$ when $m \le k$. Then the above bound on T_2 holds with probability at least $1 - \delta$. By Theorem B.4, the lower bound on T_1 holds with probability at least $3/80$. By the union bound, both events hold simultaneously with probability at least $3/80 - 1/160 = 1/32$. On this intersection event, using the bound from the T_1 part,

$$|\widetilde{\tau}_\lambda(g,g) - \tau_\lambda(g,g)| \ge T_1 - T_2 \ge \min\left\{\frac{1}{16}\sqrt{\frac{k}{m}}, \frac{1}{4}\right\} - 9\eta.$$

We work in the nontrivial regime $\sqrt{k/m} \le 4$, so the minimum equals $\frac{1}{16}\sqrt{k/m}$. Now choose

$$\eta := \frac{\varepsilon}{288}.$$

If $m \le k/\varepsilon^2$, then $\sqrt{k/m} \ge \varepsilon$, and hence

$$\min\left\{\frac{1}{16}\sqrt{\frac{k}{m}}, \frac{1}{4}\right\} - 9\eta \ge \frac{1}{16}\varepsilon - \frac{9}{288}\varepsilon = \frac{1}{32}\varepsilon.$$

Therefore, with probability at least $1/32$,

$$|\widetilde{\tau}_\lambda(g,g) - \tau_\lambda(g,g)| \ge \frac{1}{32}\varepsilon.$$

Recalling that $\tau_0(g,g) = 1$ and that $d_\lambda(F) = \Theta(k)$ for $\eta \ll 1$, this shows that whenever $m = o(d_\lambda(F)/\varepsilon^2)$, with constant probability there exists $g \in \mathrm{range}(F)$ such that

$$|\widetilde{\tau}_\lambda(g,g) - \tau_\lambda(g,g)| = \Omega(\varepsilon)\tau_0(g,g),$$

as claimed. □

C Proofs for Section 2.3 (Factorized Influence)

C.1 Proof of the Barrier of Unregularized Factorized Influence

We first record the factorized counterpart of the sharp barrier for exact preservation (Theorem 2.1) discussion in Section 2.3, i.e., Theorem 2.5. While Theorem 2.1 characterizes exact invariance for general sketches, the factorized sketch $P = P_A \otimes P_E$ admits a more explicit, factor-level injectivity condition. We restate Theorem 2.5 and prove it below:

Theorem. *Let $A \succeq 0 \in \mathbb{R}^{d_A \times d_A}$, $E \succeq 0 \in \mathbb{R}^{d_E \times d_E}$, and $F := A \otimes E \succeq 0 \in \mathbb{R}^{(d_A d_E)\times(d_A d_E)}$. Let $r_A := \mathrm{rank}(A)$, $r_E := \mathrm{rank}(E)$, and $r := \mathrm{rank}(F) = r_A r_E$. Fix $P_A \in \mathbb{R}^{m_A \times d_A}$, $P_E \in \mathbb{R}^{m_E \times d_E}$, and define $P := P_A \otimes P_E \in \mathbb{R}^{(m_A m_E)\times(d_A d_E)}$. Then the following are equivalent:*

- *(i) For all $g, g' \in \mathrm{range}(F)$, we have $\widetilde{\tau}_0(g,g') = \tau_0(g,g')$.*
- *(ii) P is injective on $\mathrm{range}(F)$, i.e., $\mathrm{rank}(PU) = r$ for* any *orthonormal basis $U \in \mathbb{R}^{(d_A d_E)\times r}$ of $\mathrm{range}(F)$.*
- *(iii) P_A is injective on $\mathrm{range}(A)$ and P_E is injective on $\mathrm{range}(E)$. Equivalently, for orthonormal bases $U_A \in \mathbb{R}^{d_A \times r_A}$ of $\mathrm{range}(A)$ and $U_E \in \mathbb{R}^{d_E \times r_E}$ of $\mathrm{range}(E)$, we have $\mathrm{rank}(P_A U_A) = r_A$ and $\mathrm{rank}(P_E U_E) = r_E$.*

In particular, $m_A \ge r_A$ and $m_E \ge r_E$ are necessary, hence $m = m_A m_E \ge r_A r_E = r$.

Proof. The equivalence between (i) and (ii) is exactly Theorem 2.1. It remains to relate (ii) and (iii) in the factorized setting. Let U_A and U_E be orthonormal bases of $\mathrm{range}(A)$ and $\mathrm{range}(E)$, respectively. Then $U := U_A \otimes U_E$ is an orthonormal basis of $\mathrm{range}(F)$. Using the mixed-product identity,

$$PU = (P_A \otimes P_E)(U_A \otimes U_E) = (P_A U_A) \otimes (P_E U_E).$$

Moreover, $\mathrm{rank}(X \otimes Y) = \mathrm{rank}(X)\,\mathrm{rank}(Y)$ for any matrices X, Y. Therefore,

$$\mathrm{rank}(PU) = \mathrm{rank}(P_A U_A)\,\mathrm{rank}(P_E U_E).$$

Since $\operatorname{rank}(P_A U_A) \le r_A$ and $\operatorname{rank}(P_E U_E) \le r_E$, we have $\operatorname{rank}(PU) = r_A r_E$ if and only if $\operatorname{rank}(P_A U_A) = r_A$ and $\operatorname{rank}(P_E U_E) = r_E$, which is equivalent to injectivity of P_A on $\operatorname{range}(A)$ and P_E on $\operatorname{range}(E)$.

The dimensional necessity $m_A \ge r_A$, $m_E \ge r_E$ follows immediately from $\operatorname{rank}(P_A U_A) \le \min\{m_A, r_A\}$ and $\operatorname{rank}(P_E U_E) \le \min\{m_E, r_E\}$. □

C.2 Proof of Factorized Resolvent Perturbation Concentration for Regularized Projection

This section proves the key technical lemma used in the factorized influence analysis in the main text (Theorem 2.6). The main technical challenges relative to the i.i.d. sketching setting are that, for a Kronecker sketch $P = P_A \otimes P_E$, the matrix $P^\top P$ decomposes into a sum of Kronecker-structured error terms rather than a single sample covariance, and P no longer satisfies the i.i.d. assumptions.

In the following, we prove the factorized version of Theorem B.2:

Theorem C.1 (Factorized covariance deviation for K-FAC)**.** *Let $F = A \otimes E \succeq 0$ and $P = P_A \otimes P_E$ be as above, and fix $\varepsilon, \delta \in (0,1)$. Assuming $\lambda \le \|A\|_2 \|E\|_2$, and define the rescaled regularization levels $\lambda_E := \lambda / \|E\|_2$ and $\lambda_A := \lambda / \|A\|_2$. If*

$$m_A = \Omega\left(\frac{d_{\lambda_E}(A) + \log(1/\delta)}{\varepsilon^2}\right), \qquad m_E = \Omega\left(\frac{d_{\lambda_A}(E) + \log(1/\delta)}{\varepsilon^2}\right),$$

then with probability at least $1 - 2\delta$,

$$\|B^\top (P^\top P - I) B\|_2 \le 2\varepsilon + 3\varepsilon^2.$$

Proof. Write

$$\Delta_A := P_A^\top P_A - I_{d_A}, \qquad \Delta_E := P_E^\top P_E - I_{d_E}.$$

Using $(X \otimes Y)^\top (X \otimes Y) = (X^\top X) \otimes (Y^\top Y)$, we have

$$P^\top P - I_{d_A d_E} = (P_A^\top P_A) \otimes (P_E^\top P_E) - I_{d_A} \otimes I_{d_E} = \Delta_A \otimes I_{d_E} + I_{d_A} \otimes \Delta_E + \Delta_A \otimes \Delta_E.$$

Therefore, by the triangle inequality,

$$\|B^\top (P^\top P - I) B\|_2 \le T_1 + T_2 + T_3, \tag{1}$$

where

$$T_1 := \left\|B^\top (\Delta_A \otimes I_{d_E}) B\right\|_2, \quad T_2 := \left\|B^\top (I_{d_A} \otimes \Delta_E) B\right\|_2, \quad T_3 := \left\|B^\top (\Delta_A \otimes \Delta_E) B\right\|_2.$$

Bounding T_1. Let $A = U_A \Lambda_A U_A^\top$ and $E = U_E \Lambda_E U_E^\top$ be eigendecompositions with $\Lambda_A = \operatorname{diag}(\{\alpha_i\}_{i=1}^{d_A})$, $\Lambda_E = \operatorname{diag}(\{\gamma_j\}_{j=1}^{d_E})$, and U_A, U_E orthonormal. Then $F = A \otimes E$ is diagonalized by $U := U_A \otimes U_E$, and

$$B = F^{1/2}(F + \lambda I)^{-1/2} = U D U^\top,$$

where D is diagonal with entries β_{ij} such that

$$\beta_{ij} := \sqrt{\frac{\alpha_i \gamma_j}{\alpha_i \gamma_j + \lambda}}, \qquad (i,j) \in [d_A] \times [d_E].$$

Define $\widetilde{\Delta}_A := U_A^\top \Delta_A U_A$. Then using the basic identity $(X \otimes Y)(Z \otimes W) = (XZ) \otimes (YW)$,

$$\begin{aligned} T_1 &= \|U D U^\top (\Delta_A \otimes I_{d_E}) U D U^\top\|_2 \\ &= \|D (U_A^\top \otimes U_E^\top)(\Delta_A \otimes I_{d_E})(U_A \otimes U_E) D\|_2 \\ &= \|D (U_A^\top \Delta_A U_A \otimes I_{d_E}) D\|_2 = \|D (\widetilde{\Delta}_A \otimes I_{d_E}) D\|_2. \end{aligned}$$

The matrix $D(\widetilde{\Delta}_A \otimes I) D$ is not itself a Kronecker product, but it becomes block diagonal after a permutation of coordinates. Let $\Pi \in \{0,1\}^{(d_A d_E) \times (d_A d_E)}$ be the canonical commutation matrix satisfying

$$\Pi (X \otimes Y) \Pi^\top = Y \otimes X \qquad \text{for all conformable } X, Y.$$

Since Π is orthogonal, $\|M\|_2 = \|\Pi M \Pi^\top\|_2$ for any M. Thus,

$$T_1 = \|\Pi D(\widetilde{\Delta}_A \otimes I_{d_E}) D \Pi^\top\|_2 = \|D_\Pi (I_{d_E} \otimes \widetilde{\Delta}_A) D_\Pi\|_2,$$

where $D_\Pi \coloneqq \Pi D \Pi^\top$ remains diagonal. The matrix $D_\Pi(I_{d_E} \otimes \widetilde{\Delta}_A)D_\Pi$ is block diagonal with d_E blocks; the j-th block (corresponding to the j-th eigenvalue γ_j) equals

$$D^{(j)} \widetilde{\Delta}_A D^{(j)}, \qquad D^{(j)} \coloneqq \operatorname{diag}(\{\beta_{ij}\}_{i=1}^{d_A}) = \operatorname{diag}\left(\left\{ \sqrt{\frac{\alpha_i \gamma_j}{\alpha_i \gamma_j + \lambda}} \right\}_{i=1}^{d_A} \right).$$

Hence,

$$T_1 = \max_{j \in [d_E]} \|D^{(j)} \widetilde{\Delta}_A D^{(j)}\|_2. \tag{2}$$

We now compare each $D^{(j)}$ to a single dominating diagonal depending only on A. Since $\gamma_j \le \|E\|_2$ and $\alpha_i \ge 0$,[5]

$$\frac{\alpha_i \gamma_j}{\alpha_i \gamma_j + \lambda} \le \frac{\alpha_i}{\alpha_i + \lambda/\gamma_j} \le \frac{\alpha_i}{\alpha_i + \lambda/\|E\|_2} = \frac{\alpha_i}{\alpha_i + \lambda_E}.$$

Define

$$D_A^{\max} \coloneqq \operatorname{diag}\left(\left\{ \sqrt{\frac{\alpha_i}{\alpha_i + \lambda_E}} \right\}_{i=1}^{d_A} \right).$$

Then for each j there exists a diagonal contraction $S^{(j)}$ such that

$$D^{(j)} = S^{(j)} D_A^{\max} = D_A^{\max} S^{(j)}, \qquad \|S^{(j)}\|_2 \le 1,$$

and therefore,

$$\|D^{(j)} \widetilde{\Delta}_A D^{(j)}\|_2 = \|S^{(j)} D_A^{\max} \widetilde{\Delta}_A D_A^{\max} S^{(j)}\|_2 \le \|D_A^{\max} \widetilde{\Delta}_A D_A^{\max}\|_2.$$

Combining with Eq.(2) yields

$$T_1 \le \|D_A^{\max} \widetilde{\Delta}_A D_A^{\max}\|_2.$$

Finally, note that

$$D_A^{\max} \widetilde{\Delta}_A D_A^{\max} = (U_A D_A^{\max})^\top (P_A^\top P_A - I_{d_A})(U_A D_A^{\max}).$$

Let $M_A \coloneqq U_A D_A^{\max}$. Applying Theorem B.2 to M_A and sketching P_A (with failure probability δ) gives that when

$$m_A = \Omega\left(\frac{r(M_A^\top M_A) + \log(1/\delta)}{\varepsilon^2} \right),$$

we have $T_1 \le \varepsilon$ with probability at least $1 - \delta$. It remains to identify $r(M_A^\top M_A)$. Since $M_A^\top M_A = (D_A^{\max})^2$ is diagonal with spectral norm at most 1,

$$r(M_A^\top M_A) = \frac{\operatorname{tr}((D_A^{\max})^2)}{\|(D_A^{\max})^2\|_2} = \operatorname{tr}((D_A^{\max})^2) = \sum_{i=1}^{d_A} \frac{\alpha_i}{\alpha_i + \lambda_E} = d_{\lambda_E}(A).$$

Thus, under the stated condition on m_A, with probability at least $1 - \delta$,

$$T_1 \le \varepsilon. \tag{3}$$

Bounding T_2. The bound for T_2 is identical by symmetry (and is in fact simpler because $I_{d_A} \otimes \widetilde{\Delta}_E$ is already block diagonal in the A-first ordering). Specifically, define $\widetilde{\Delta}_E \coloneqq U_E^\top \Delta_E U_E$ and

$$D_E^{\max} \coloneqq \operatorname{diag}\left(\left\{ \sqrt{\frac{\gamma_j}{\gamma_j + \lambda_A}} \right\}_{j=1}^{d_E} \right), \qquad M_E \coloneqq U_E D_E^{\max}.$$

Applying Theorem B.2 to M_E and P_E yields that, when

$$m_E = \Omega\left(\frac{d_{\lambda_A}(E) + \log(1/\delta)}{\varepsilon^2} \right),$$

we have with probability at least $1 - \delta$,

$$T_2 \le \varepsilon. \tag{4}$$

[5]Note that the inequality holds trivially when $\gamma_j = 0$.

Bounding T_3. We show that the diagonal D is dominated by a Kronecker product of the dominating diagonals $D_A^{\max}$ and $D_E^{\max}$, up to a universal constant, provided $\lambda \le \|A\|_2\|E\|_2$. For each (i,j),

$$\beta_{ij}^2 = \frac{\alpha_i\gamma_j}{\alpha_i\gamma_j + \lambda}.$$

We claim that

$$\frac{\alpha_i\gamma_j}{\alpha_i\gamma_j + \lambda} \le 3 \cdot \frac{\alpha_i}{\alpha_i + \lambda_E} \cdot \frac{\gamma_j}{\gamma_j + \lambda_A}. \tag{5}$$

Indeed, Eq.(5) is equivalent (after taking reciprocals of positive quantities) to

$$(\alpha_i + \lambda_E)(\gamma_j + \lambda_A) \le 3(\alpha_i\gamma_j + \lambda).$$

Expanding the left-hand side gives

$$(\alpha_i + \lambda_E)(\gamma_j + \lambda_A) = \alpha_i\gamma_j + \alpha_i\lambda_A + \gamma_j\lambda_E + \lambda_A\lambda_E.$$

Using $\alpha_i \le \|A\|_2$, $\gamma_j \le \|E\|_2$, and the definitions $\lambda_A = \lambda/\|A\|_2$, $\lambda_E = \lambda/\|E\|_2$, we obtain

$$\alpha_i\lambda_A \le \lambda, \qquad \gamma_j\lambda_E \le \lambda, \qquad \lambda_A\lambda_E = \frac{\lambda^2}{\|A\|_2\|E\|_2} \le \lambda,$$

where the last inequality uses the assumption $\lambda \le \|A\|_2\|E\|_2$. Therefore,

$$(\alpha_i + \lambda_E)(\gamma_j + \lambda_A) \le \alpha_i\gamma_j + 3\lambda \le 3(\alpha_i\gamma_j + \lambda),$$

which proves Eq.(5). Taking square-roots yields

$$\beta_{ij} \le \sqrt{3}\sqrt{\frac{\alpha_i}{\alpha_i + \lambda_E}}\sqrt{\frac{\gamma_j}{\gamma_j + \lambda_A}}.$$

Therefore, there exists a diagonal contraction S such that

$$D = \sqrt{3}S(D_A^{\max} \otimes D_E^{\max}) = \sqrt{3}(D_A^{\max} \otimes D_E^{\max})S, \qquad \|S\|_2 \le 1,$$

and therefore

$$\begin{aligned} T_3 &= 3\|S(D_A^{\max} \otimes D_E^{\max})(\widetilde{\Delta}_A \otimes \widetilde{\Delta}_E)(D_A^{\max} \otimes D_E^{\max})S\|_2 \\ &\le 3\|(D_A^{\max}\widetilde{\Delta}_A D_A^{\max}) \otimes (D_E^{\max}\widetilde{\Delta}_E D_E^{\max})\|_2. \end{aligned}$$

Since $\|X \otimes Y\|_2 = \|X\|_2\|Y\|_2$, this becomes

$$T_3 \le 3\|D_A^{\max}\widetilde{\Delta}_A D_A^{\max}\|_2\|D_E^{\max}\widetilde{\Delta}_E D_E^{\max}\|_2.$$

On the event where both Eq.(3) and Eq.(4) hold, we obtain

$$T_3 \le 3\varepsilon^2. \tag{6}$$

Putting together. By Eqs.(1), (3), (4) and (6), on the intersection of the two concentration events (one for P_A, one for P_E),

$$\|B^\top(P^\top P - I)B\|_2 \le \varepsilon + \varepsilon + 3\varepsilon^2 = 2\varepsilon + 3\varepsilon^2.$$

The two concentration events each fail with probability at most δ, so by a union bound, the intersection holds with probability at least $1 - 2\delta$. This completes the proof. □

C.3 Note on Proof of Theorem 2.6

We note that while Theorem C.1 is stated with failure probability 2δ and deviation level $2\varepsilon + 3\varepsilon^2$, in the proof of Theorem 2.6, we require it to be with failure probability δ and deviation level ε. This is only for notational convenience: given target parameters (ε, δ), one may apply the theorem with $\epsilon := \varepsilon/10$ and $\eta := \delta/2$, which yields probability at least $1 - 2\eta = 1 - \delta$ and deviation at most $2\epsilon + 3\epsilon^2 \le \varepsilon/2 \le \varepsilon$ for $\varepsilon \in (0, 1)$.

D PROOFS FOR SECTION 3.1 (LEAKAGE OF PROJECTION)

In this section, we prove Theorem 3.1, which we first repeat the statement for convenience:

Theorem. *Let $\{g'_j\}_{j=1}^k \subset \mathbb{R}^d$, and for each j let $g'_{j,\perp}$ denote the orthogonal projection of g'_j onto* $\ker(F)$. *Let $k' = \dim(\mathrm{span}(\{g'_{j,\perp}\}_{j=1}^k))$. For any $\varepsilon, \delta \in (0,1)$, if*

$$m = \Omega\left(\frac{r + \min\left(\log(k/\delta), k' + \log(1/\delta)\right)}{\varepsilon^2}\right),$$

then with probability at least $1-\delta$, the following holds for all $j \in \{1, \dots, k\}$:

- ***Unregularized**: For $T_j \coloneqq (Pg)^\top (PFP^\top)^\dagger (Pg'_{j,\perp})$, we have*

$$|T_j| \le \varepsilon \frac{\|g\|_2 \|g'_{j,\perp}\|_2}{\lambda^+_{\min}(F)},$$

  where $\lambda^+_{\min}(F)$ denotes the smallest non-zero eigenvalue of F.

- ***Regularized**: For $T_{\lambda,j} \coloneqq (Pg)^\top (PFP^\top + \lambda I)^{-1} (Pg'_{j,\perp})$, we have*

$$|T_{\lambda,j}| \le \varepsilon \|g\|_2 \|g'_{j,\perp}\|_2 \left(\frac{1}{\lambda} + \frac{2\|F\|_2}{\lambda^2}\right)$$

D.1 PROOF PLAN FOR THEOREM 3.1

The main organizing step is a deterministic reduction: Theorem D.1 shows that *both* the regularized and unregularized leakage bounds follow once the sketch P satisfies two concentration conditions with respect to an orthonormal basis U of $\mathrm{range}(F)$: (i) subspace stability on $\mathrm{range}(F)$, $\|U^\top (P^\top P - I_d) U\|_2 \le \varepsilon$, and (ii) cross-term control between $\mathrm{range}(F)$ and the kernel direction(s), $\|U^\top (P^\top P - I_d) g'_{j,\perp}\|_2 \le \varepsilon \|g'_{j,\perp}\|_2$ for each j. Indeed, this is shown formally in Theorem D.1.

Lemma D.1. *Let $\{g'_j\}_{j=1}^k \subset \mathbb{R}^d$, and for each j let $g'_{j,\perp}$ denote the orthogonal projection of g'_j onto* $\ker(F)$. *Fix a realization of P, and let $U \in \mathbb{R}^{d \times r}$ be an orthonormal basis for* $\mathrm{range}(F)$. *Assume that for some $\varepsilon \in (0,1)$, the following two inequalities hold:*

(i) $\|U^\top (P^\top P - I_d) U\|_2 \le \varepsilon$,

(ii) $\|U^\top (P^\top P - I_d) g'_{j,\perp}\|_2 \le \varepsilon \|g'_{j,\perp}\|_2$ *for every $j \in \{1, \dots, k\}$.*

Then for any fixed $g \in \mathrm{range}(F)$, the following bounds hold simultaneously for all $j \in \{1, \dots, k\}$*:*

$$\left|(Pg)^\top (PFP^\top)^\dagger (Pg'_{j,\perp})\right| \le \varepsilon \frac{1+\varepsilon}{(1-\varepsilon)^2} \cdot \frac{\|g\|_2 \|g'_{j,\perp}\|_2}{\lambda^+_{\min}(F)}.$$

and

$$\left|(Pg)^\top (PFP^\top + \lambda I)^{-1} (Pg'_{j,\perp})\right| \le \varepsilon \|g\|_2 \|g'_{j,\perp}\|_2 \left(\frac{1}{\lambda} + \frac{2\|F\|_2}{\lambda^2}\right).$$

Proof. We prove the unregulairzed case first.

Unregularized Case. Fix j. Using $\mathrm{range}(PFP^\top) = \mathrm{range}(PU)$, let

$$\Pi_{PU} \coloneqq PU(U^\top P^\top PU)^{-1}(PU)^\top$$

denote the orthogonal projector onto $\mathrm{range}(PU)$. Then

$$(Pg)^\top (PFP^\top)^\dagger (Pg'_{j,\perp}) = g^\top P^\top (PFP^\top)^\dagger \Pi_{PU} P g'_{j,\perp},$$

and hence

$$\left|(Pg)^\top (PFP^\top)^\dagger (Pg'_{j,\perp})\right| \le \|(PFP^\top)^\dagger\|_2 \|Pg\|_2 \|\Pi_{PU} P g'_{j,\perp}\|_2.$$

We bound the three terms on the right-hand side.

First, we bound $\|(PFP^\top)^\dagger\|_2$. Write the compact eigendecomposition $F = U\Sigma U^\top$, where $\Sigma \succ 0$ is diagonal and $\|\Sigma^{-1}\|_2 = 1/\lambda^+_{\min}(F)$. Since $PFP^\top = (PU)\Sigma(PU)^\top$, we have

$$\|(PFP^\top)^\dagger\|_2 = \|(PU)^\dagger\|_2^2\|\Sigma^{-1}\|_2 = \frac{1}{\sigma_{\min}(PU)^2}\cdot\frac{1}{\lambda^+_{\min}(F)}.$$

Moreover, assumption (i) implies that all eigenvalues of $U^\top P^\top PU = (PU)^\top(PU)$ lie in $[1-\varepsilon, 1+\varepsilon]$, hence $\sigma_{\min}(PU)^2 \geq 1-\varepsilon$ and

$$\|(PFP^\top)^\dagger\|_2 \leq \frac{1}{(1-\varepsilon)\lambda^+_{\min}(F)}.$$

To bound $\|Pg\|_2$, write $g = Uh$, we have $\|Pg\|_2 \leq \sqrt{1+\varepsilon}\|g\|_2$ since

$$\|Pg\|_2^2 = h^\top(U^\top P^\top PU)h \leq (1+\varepsilon)\|h\|_2^2 = (1+\varepsilon)\|g\|_2^2.$$

Finally, to bound $\|\Pi_{PU}Pg'_{j,\perp}\|_2$, with $U^\top g'_{j,\perp} = 0$, we have $U^\top P^\top Pg'_{j,\perp} = U^\top(P^\top P - I_d)g'_{j,\perp}$, hence

$$\begin{aligned}\|\Pi_{PU}Pg'_{j,\perp}\|_2 &= \left\|PU(U^\top P^\top PU)^{-1}U^\top P^\top Pg'_{j,\perp}\right\|_2 \\ &\leq \|PU\|_2\|(U^\top P^\top PU)^{-1}\|_2\|U^\top(P^\top P - I_d)g'_{j,\perp}\|_2.\end{aligned}$$

By assumption (i), $\|PU\|_2 = \sigma_{\max}(PU) \leq \sqrt{1+\varepsilon}$ and $\|(U^\top P^\top PU)^{-1}\|_2 \leq 1/(1-\varepsilon)$. By assumption (ii), $\|U^\top(P^\top P - I_d)g'_{j,\perp}\|_2 \leq \varepsilon\|g'_{j,\perp}\|_2$. Therefore,

$$\|\Pi_{PU}Pg'_{j,\perp}\|_2 \leq \varepsilon\frac{\sqrt{1+\varepsilon}}{1-\varepsilon}\|g'_{j,\perp}\|_2.$$

Combining the bounds yields

$$|(Pg)^\top(PFP^\top)^\dagger(Pg'_{j,\perp})| \leq \varepsilon\frac{1+\varepsilon}{(1-\varepsilon)^2}\cdot\frac{\|g\|_2\|g'_{j,\perp}\|_2}{\lambda^+_{\min}(F)}.$$

Regularized Case. Fix $g \in \mathrm{range}(F)$ and j. Write $g = Uh$. Set $V \coloneqq PF^{1/2}$, so that $PFP^\top = VV^\top$. By the Woodbury identity,

$$(VV^\top + \lambda I)^{-1} = \frac{1}{\lambda}I - \frac{1}{\lambda^2}V\Big(I + \frac{1}{\lambda}V^\top V\Big)^{-1}V^\top.$$

Therefore, writing $T_{\lambda,j} = (Pg)^\top(VV^\top + \lambda I)^{-1}(Pg'_{j,\perp})$, we have the decomposition $T_{\lambda,j} = T^{(1)}_{\lambda,j} - T^{(2)}_{\lambda,j}$ where

$$T^{(1)}_{\lambda,j} = \frac{1}{\lambda}g^\top P^\top Pg'_{j,\perp}, \qquad T^{(2)}_{\lambda,j} = \frac{1}{\lambda^2}g^\top P^\top PF^{1/2}\Big(I + \frac{1}{\lambda}V^\top V\Big)^{-1}F^{1/2}P^\top Pg'_{j,\perp}.$$

Since $g^\top g'_{j,\perp} = 0$,

$$|T^{(1)}_{\lambda,j}| = \frac{1}{\lambda}\big|g^\top(P^\top P - I_d)g'_{j,\perp}\big| = \frac{1}{\lambda}\big|h^\top U^\top(P^\top P - I_d)g'_{j,\perp}\big| \leq \frac{\varepsilon}{\lambda}\|g\|_2\|g'_{j,\perp}\|_2,$$

using Cauchy–Schwarz and assumption (ii).

Next, we bound $T^{(2)}_{\lambda,j}$. Note that $V^\top V = F^{1/2}P^\top PF^{1/2} \succeq 0$, so $I + \frac{1}{\lambda}V^\top V \succeq I$, which implies $\big\|(I + \frac{1}{\lambda}V^\top V)^{-1}\big\|_2 \leq 1$. Moreover, since $\mathrm{range}(F^{1/2}) = \mathrm{range}(F)$, we have $F^{1/2} = \Pi_F F^{1/2} = F^{1/2}\Pi_F$, and hence we may insert Π_F on both sides of each $F^{1/2}$ factor. Using sub-multiplicativity and $\|F^{1/2}\|_2^2 = \|F\|_2$, we obtain

$$|T^{(2)}_{\lambda,j}| \leq \frac{1}{\lambda^2}\|\Pi_F P^\top Pg\|_2\|F\|_2\|\Pi_F P^\top Pg'_{j,\perp}\|_2.$$

Since $\Pi_F = UU^\top$ and $g = Uh$,

$$\|\Pi_F P^\top Pg\|_2 = \|U(U^\top P^\top PU)h\|_2 \leq \|U^\top P^\top PU\|_2\|g\|_2 \leq (1+\varepsilon)\|g\|_2,$$

where we used $U^\top P^\top PU = I_r + U^\top(P^\top P - I_d)U$ and assumption (i). Moreover, since $U^\top g'_{j,\perp} = 0$,

$$\|\Pi_F P^\top P g'_{j,\perp}\|_2 = \|U^\top P^\top P g'_{j,\perp}\|_2 = \|U^\top(P^\top P - I_d)g'_{j,\perp}\|_2 \le \varepsilon\|g'_{j,\perp}\|_2,$$

by assumption (ii). Hence

$$|T^{(2)}_{\lambda,j}| \le \frac{\|F\|_2}{\lambda^2}(1+\varepsilon)\varepsilon\|g\|_2\|g'_{j,\perp}\|_2.$$

Combining the two pieces and using $\varepsilon \le 1$ gives

$$|T_{\lambda,j}| \le \varepsilon\|g\|_2\|g'_{j,\perp}\|_2\left(\frac{1}{\lambda} + \frac{2\|F\|_2}{\lambda^2}\right).$$

□

Thus, the remaining work in the proof is to verify these two conditions in the single-gradient and multi-gradient regimes. We break the proof into the following cases:

1. For a single kernel component $g'_\perp \in \ker(F)$:
   - Theorem D.2: prove the bound for unregularized and regularized case.
2. Extend both cases to $\{g'_{j,\perp}\}_{j=1}^k \subseteq \ker(F)$ with $k' = \dim(\mathrm{span}(\{g'_{j,\perp}\}_{j=1}^k))$:
   - Theorem D.3: subspace argument with $m = \Omega\left(\frac{r+k'+\log(1/\delta)}{\varepsilon^2}\right)$.
   - Theorem D.4: union-bound argument with $m = \Omega\left(\frac{r+\log(k/\delta)}{\varepsilon^2}\right)$.

We now start the proof.

D.2 Proof of Single Test Gradient Leakage

Proposition D.2. *Assume $g \in \mathrm{range}(F)$ and let $g' \in \mathbb{R}^d$. For any $\varepsilon, \delta \in (0,1)$, if $m = \Omega(\varepsilon^{-2}(r + \log(1/\delta)))$, then with probability at least $1-\delta$,*

1. ***Unregularized**: Let $T := (Pg)^\top(PFP^\top)^\dagger(Pg'_\perp)$, then*

$$|T| \le \varepsilon\frac{\|g\|_2\|g'_\perp\|_2}{\lambda^+_{\min}(F)},$$

   where $\lambda^+_{\min}(F)$ denotes the smallest non-zero eigenvalue of F.

2. ***Regularized**: Let $T_\lambda := (Pg)^\top(PFP^\top + \lambda I)^{-1}(Pg'_\perp)$, then*

$$|T_\lambda| \le \varepsilon\|g\|_2\|g'_\perp\|_2\left(\frac{1}{\lambda} + \frac{2\|F\|_2}{\lambda^2}\right).$$

Proof. From Theorem D.1, it suffices to verify conditions (i) and (ii).

Let $S := \mathrm{span}(\mathrm{range}(F) \cup \{g'_\perp\})$, so $\dim(S) = r+1$, and let $W \in \mathbb{R}^{d\times(r+1)}$ be an orthonormal basis for S. Fix any $\eta \in (0,1)$ and define the event

$$\mathcal{E}(\eta) := \left\{\left\|W^\top(P^\top P - I_d)W\right\|_2 \le \eta\right\}.$$

On $\mathcal{E}(\eta)$, for any orthonormal basis $U \in \mathbb{R}^{d\times r}$ of $\mathrm{range}(F) \subseteq S$ there exists $R \in \mathbb{R}^{(r+1)\times r}$ with $R^\top R = I_r$ such that $U = WR$. Thus,

$$\|U^\top(P^\top P - I_d)U\|_2 = \|R^\top W^\top(P^\top P - I_d)WR\|_2 \le \eta.$$

Moreover, since $g'_\perp \in S$, we have $g'_\perp = WW^\top g'_\perp$ and $\|W^\top g'_\perp\|_2 = \|g'_\perp\|_2$, and hence

$$\|U^\top(P^\top P - I_d)g'_\perp\|_2 = \|R^\top W^\top(P^\top P - I_d)WW^\top g'_\perp\|_2 \le \eta\|g'_\perp\|_2.$$

Therefore, on $\mathcal{E}(\eta)$ the assumptions of Theorem D.1 hold with parameter η.

Unregularized. By Theorem B.2 applied to S, if $m = \Omega\big(((r+1)+\log(1/\delta))/\eta^2\big)$, then $\mathbb{P}(\mathcal{E}(\eta)) \geq 1-\delta$. Taking $\eta = \varepsilon/4$ and applying Theorem D.1 yields

$$|T| \leq \frac{\varepsilon}{4} \cdot \frac{1+\varepsilon/4}{(1-\varepsilon/4)^2} \cdot \frac{\|g\|_2 \|g'_\perp\|_2}{\lambda^+_{\min}(F)}.$$

As in the previous argument, $\frac{1+\varepsilon/4}{(1-\varepsilon/4)^2} \leq \frac{20}{9}$, hence the prefactor is $\leq \varepsilon$.

Regularized. Taking $\eta = \varepsilon$ and applying Theorem D.1 gives

$$|T_\lambda| \leq \varepsilon \|g\|_2 \|g'_\perp\|_2 \left(\frac{1}{\lambda} + \frac{2\|F\|_2}{\lambda^2}\right).$$

□

This result shows that, in the unregularized case, the kernel leakage term decays at rate $O(m^{-1/2})$, with constants that depend on the smallest non-zero eigenvalue of F. On the other hand, in the regularized case, the kernel leakage term also decays at rate $O(m^{-1/2})$, but with constants depending on $\|F\|_2$ and the regularization parameter λ. This dependence reflects the sensitivity of the pseudoinverse to near-degeneracies in the spectrum of F.

D.3 Proof of Multiple Test Gradients Leakage

Having established the deterministic reduction in Theorem D.1, we now show how to enforce its two assumptions uniformly over multiple test gradients. First, we observe that the previous analysis naturally generalizes by considering the subspace spanned by all test gradients.

Proposition D.3. *Let $\{g'_j\}_{j=1}^k \subset \mathbb{R}^d$, and for each j let $g'_{j,\perp}$ denote the orthogonal projection of g'_j onto $\ker(F)$. Let $k' = \dim\big(\mathrm{span}(\{g'_{j,\perp}\}_{j=1}^k)\big)$. For any $\varepsilon, \delta \in (0,1)$, if*

$$m = \Omega\left(\frac{r + k' + \log(1/\delta)}{\varepsilon^2}\right),$$

then with probability at least $1-\delta$, the leakage bounds in Theorem D.2 hold simultaneously for all $j \in \{1, \ldots, k\}$.

Proof. Let $S \coloneqq \mathrm{span}(\mathrm{range}(F) \cup \{g'_{j,\perp}\}_{j=1}^k)$, so that $\dim(S) = r + k'$, and let $W \in \mathbb{R}^{d \times (r+k')}$ be an orthonormal basis for S. By Theorem B.2, with probability at least $1-\delta$,

$$\big\|W^\top (P^\top P - I_d) W\big\|_2 \leq \varepsilon,$$

provided that $m = \Omega\big(\varepsilon^{-2}(r + k' + \log(1/\delta))\big)$. On this event, for all $x, y \in S$,

$$\big|x^\top (P^\top P - I_d) y\big| = \big|(W^\top x)^\top W^\top (P^\top P - I_d) W (W^\top y)\big| \leq \varepsilon \|x\|_2 \|y\|_2.$$

Now let $U \in \mathbb{R}^{d \times r}$ be an orthonormal basis for $\mathrm{range}(F)$. Since $\mathrm{range}(F) \subseteq S$, the columns of U are contained in S, and hence

$$\big\|U^\top (P^\top P - I_d) U\big\|_2 \leq \big\|W^\top (P^\top P - I_d) W\big\|_2 \leq \varepsilon.$$

Moreover, for each j, using that U has orthonormal columns and $\mathrm{range}(F) \subseteq S$, we have

$$\begin{aligned}\big\|U^\top (P^\top P - I_d) g'_{j,\perp}\big\|_2 &= \sup_{\substack{a \in \mathbb{R}^r \\ \|a\|_2 = 1}} \big|a^\top U^\top (P^\top P - I_d) g'_{j,\perp}\big| \\ &= \sup_{\substack{x \in \mathrm{range}(F) \\ \|x\|_2 = 1}} \big|x^\top (P^\top P - I_d) g'_{j,\perp}\big| \leq \varepsilon \|g'_{j,\perp}\|_2,\end{aligned}$$

where the last inequality applies the bilinear bound above with $x \in \mathrm{range}(F) \subseteq S$ and $y = g'_{j,\perp} \in S$. Thus, the assumptions of Theorem D.1 hold simultaneously for all j, and the corollary follows by applying Theorem D.1. □

While Theorem D.3 is effective when the test gradients are low-dimensional, as $g'_j \in \mathbb{R}^d$ lies in high dimension, it is almost certain that k' will be large, and most likely $k' \approx k$. In this case, by directly controlling the concentration of the bilinear form, we can obtain a bound that scales only logarithmically with the *number* of test gradients.

Proposition D.4. *Let $\{g'_j\}_{j=1}^k \subset \mathbb{R}^d$, and for each j let $g'_{j,\perp}$ denote the orthogonal projection of g'_j onto $\ker(F)$. For any $\varepsilon, \delta \in (0,1)$, if*

$$m = \Omega\left(\frac{r + \log(k/\delta)}{\varepsilon^2}\right),$$

then with probability at least $1-\delta$, the leakage bounds in Theorem D.2 hold for all $j \in \{1, \dots, k\}$.

Proof. Let $U \in \mathbb{R}^{d\times r}$ be an orthonormal basis for $\mathrm{range}(F)$. We will verify the two assumptions of Theorem D.1 uniformly over $\{g'_{j,\perp}\}_{j=1}^k$. Since Theorem D.1 incurs a benign factor $\frac{1+\varepsilon}{(1-\varepsilon)^2}$ in the unregularized case, we will run the concentration argument below with accuracy parameter $\varepsilon/4$; the resulting constant-factor strengthening is absorbed by the $\Omega(\cdot)$ sample complexity.

Controlling $\|U^\top(P^\top P - I_d)U\|_2$. By Theorem B.2 applied to the r-dimensional subspace $\mathrm{range}(F)$, with probability at least $1-\delta/2$,

$$\|U^\top(P^\top P - I_d)U\|_2 \le \varepsilon,$$

provided that $m = \Omega\big(\varepsilon^{-2}(r + \log(2/\delta))\big)$.

Controlling $\|U^\top(P^\top P - I_d)g'_{j,\perp}\|_2$ for all j. Fix $g'_\perp \in \ker(F)$ with $\|g'_\perp\|_2 = 1$. Note that

$$\|U^\top(P^\top P - I_d)g'_\perp\|_2 = \sup_{a\in S^{r-1}} |(Ua)^\top(P^\top P - I_d)g'_\perp| = \sup_{x\in US^{r-1}} |x^\top(P^\top P - I_d)g'_\perp|,$$

where $US^{r-1} = \{Ua \colon a \in \mathbb{R}^r, \|a\|_2 = 1\}$ is the unit sphere in $\mathrm{range}(F)$.

By the polarization identity $x^\top y = \frac{1}{4}(\|x+y\|_2^2 - \|x-y\|_2^2)$, the bilinear form can be written as:

$$x^\top(P^\top P - I_d)g'_\perp = \frac{1}{4}\left(\|P(x+g'_\perp)\|_2^2 - \|x+g'_\perp\|_2^2\right) - \frac{1}{4}\left(\|P(x-g'_\perp)\|_2^2 - \|x-g'_\perp\|_2^2\right).$$

To bound this uniformly over $x \in US^{r-1}$, define the set $T = T_+ \cup T_-$, where $T_\pm = \{x \pm g'_\perp \colon x \in US^{r-1}\}$. It follows that

$$\sup_{x\in US^{r-1}} |x^\top(P^\top P - I_d)g'_\perp| \le \frac{1}{2}\sup_{z\in T} |\|Pz\|_2^2 - \|z\|_2^2|.$$

Define the sub-Gaussian stochastic process $Y_z = \|Pz\|_2 - \|z\|_2$ for $z \in T$, similar to the proof of Theorem B.1. Applying the Talagrand comparison inequality (Dirksen, 2015, Theorem 3.2), with probability at least $1 - 2e^{-u}$,

$$\sup_{z\in T} |\|Pz\|_2 - \|z\|_2| \le \frac{C}{\sqrt{m}}(\gamma(T) + \sqrt{u}\cdot \mathrm{rad}(T)).$$

We analyze the radius and Gaussian complexity of T:

- $\mathrm{rad}(T)$: For any $z \in T$, $z = x \pm g'_\perp$. Since $x \perp g'_\perp$ as $x \in \mathrm{range}(F)$ and $g'_\perp \in \ker(F)$, the Pythagorean theorem gives $\|z\|_2^2 = \|x\|_2^2 + \|g'_\perp\|_2^2 = 1 + 1 = 2$, giving $\mathrm{rad}(T) = \sqrt{2} = O(1)$.
- $\gamma(T)$: By definition, $\gamma(T) = \mathbb{E}[\sup_{z\in T} |\langle h, z\rangle|]$ for $h \sim \mathcal{N}(0, I_d)$. For $z = x \pm g'_\perp$, we have $\langle h, x \pm g'_\perp\rangle = \langle h, x\rangle \pm \langle h, g'_\perp\rangle$. Thus,

  $$\gamma(T) \le \mathbb{E}\left[\sup_{x\in US^{r-1}} |\langle h, x\rangle|\right] + \mathbb{E}[|\langle h, g'_\perp\rangle|].$$

  The first term is the Gaussian complexity of the unit sphere in an r-dimensional subspace, which is bounded by $\sqrt{r}$. The second term is $\mathbb{E}[Z]$ for $Z \sim \mathcal{N}(0,1)$, which is $\sqrt{2/\pi}$. Overall, $\gamma(T) \le \sqrt{r} + \sqrt{2/\pi} \lesssim \sqrt{r}$.

With again $|a^2 - b^2| \le |a - b|(|a - b| + 2b)$ with $a = \|Pz\|_2$ and $b = \|z\|_2 = \sqrt{2}$, we have

$$\sup_{z \in T} |\|Pz\|_2^2 - \|z\|_2^2| \le \frac{C}{\sqrt{m}}(\gamma(T) + \sqrt{u}\,\mathrm{rad}(T)) \left(\frac{C}{\sqrt{m}}(\gamma(T) + \sqrt{u}\,\mathrm{rad}(T)) + 2\,\mathrm{rad}(T) \right).$$

Distributing the terms and substituting $\mathrm{rad}(T) = \sqrt{2}$ and $\gamma(T) \le \sqrt{r} + 1$, we have

$$\sup_{z \in T} |\|Pz\|_2^2 - \|z\|_2^2| \le C \left(\frac{r+u}{m} + \sqrt{\frac{r+u}{m}} \right).$$

Setting $u = \log(4k/\delta)$ ensures that $2e^{-u} = \delta/(2k)$. Hence, for a fixed $g'_\perp$, we have $\|U^\top (P^\top P - I_d) g'_\perp\|_2 \le \varepsilon$ with probability at least $1 - \delta/(2k)$, provided that $m = \Omega((r + \log(k/\delta))/\varepsilon^2)$. By a union bound over $j \in \{1, \dots, k\}$, the bound holds simultaneously for all k test gradients with probability at least $1 - \delta/2$.

Finally, taking a union bound over the two failure events (the subspace event and the k bilinear events), the same argument (with ε replaced by $\varepsilon/4$) yields that with probability at least $1 - \delta$, $\left\|U^\top (P^\top P - I_d) U\right\|_2 \le \varepsilon/4$ and

$$\left\|U^\top (P^\top P - I_d) g'_{j,\perp}\right\|_2 \le (\varepsilon/4)\|g'_{j,\perp}\|_2 \quad \text{for all } j \in \{1, \dots, k\}.$$

On this event, we apply Theorem D.1 with parameter $\varepsilon/4$. The regularized leakage bound then holds with prefactor $\varepsilon/4 \le \varepsilon$. For the unregularized leakage bound, we obtain

$$\left|(Pg)^\top (PFP^\top)^\dagger (Pg'_{j,\perp})\right| \le \frac{\varepsilon}{4} \cdot \frac{1 + \varepsilon/4}{(1 - \varepsilon/4)^2} \cdot \frac{\|g\|_2 \|g'_{j,\perp}\|_2}{\lambda^+_{\min}(F)} \le \varepsilon \cdot \frac{\|g\|_2 \|g'_{j,\perp}\|_2}{\lambda^+_{\min}(F)},$$

using $\frac{1+\varepsilon/4}{(1-\varepsilon/4)^2} \le \frac{20}{9}$ as in Theorem D.2. This completes the proof. □

E Proofs for Section 3.2 (Leakage of Factorized Influence)

In this subsection, we extend the leakage analysis in Section D to the factorized influence setting. We consider curvature matrices of the form

$$F = A \otimes E \in \mathbb{R}^{(d_A d_E) \times (d_A d_E)},$$

where $A \succeq 0$ and $E \succeq 0$. We analyze a factorized sketch

$$P = P_A \otimes P_E, \qquad P_A \in \mathbb{R}^{m_A \times d_A}, P_E \in \mathbb{R}^{m_E \times d_E},$$

so that $P \in \mathbb{R}^{(m_A m_E) \times (d_A d_E)} = \mathbb{R}^{m \times d}$ with $m = m_A m_E$ and $d = d_A d_E$. Throughout, we assume P_A and P_E are both oblivious sketches as defined in Theorem 2.2. We will show that the only new work needed is to bound the cross-term quantity $\|U^\top (P^\top P - I) g'_\perp\|_2$ (for kernel components $g'_\perp \in \ker(F)$) appearing in Theorem D.1 via factor-level primitive bounds.

Theorem. *Let $A, E \succeq 0$ and $F := A \otimes E$, and let $P = P_A \otimes P_E$ with $P_A \in \mathbb{R}^{m_A \times d_A}$ and $P_E \in \mathbb{R}^{m_E \times d_E}$. Let $r_A := \mathrm{rank}(A)$, $r_E := \mathrm{rank}(E)$, and $r := \mathrm{rank}(F) = r_A r_E$.*

Let $\{g'_j\}_{j=1}^k \subset \mathbb{R}^{d_A d_E}$ be test gradients of the form $g'_j = a'_j \otimes e'_j$. For each j, define the kernel component $g'_{j,\perp} := \Pi_{\ker(F)} g'_j$. Write $a'_j = a'_{j,/\!/} + a'_{j,\perp}$ with $a'_{j,/\!/} \in \mathrm{range}(A)$ and $a'_{j,\perp} \perp \mathrm{range}(A)$, and similarly $e'_j = e'_{j,/\!/} + e'_{j,\perp}$. Define $k_A := \sum_{j=1}^k \mathbb{1}(a'_{j,\perp} \neq 0)$, $k_E := \sum_{j=1}^k \mathbb{1}(e'_{j,\perp} \neq 0)$, and $k'_A := \dim\big(\mathrm{span}(\{a'_{j,\perp}\}_{j=1}^k)\big)$, $k'_E := \dim\big(\mathrm{span}(\{e'_{j,\perp}\}_{j=1}^k)\big)$. For any $\varepsilon, \delta \in (0,1)$, if

$$m_A = \Omega\left(\frac{r_A + \min\{\log(\frac{k_A}{\delta}), k'_A + \log(\frac{1}{\delta})\}}{\varepsilon^2} \right), m_E = \Omega\left(\frac{r_E + \min\{\log(\frac{k_E}{\delta}), k'_E + \log(\frac{1}{\delta})\}}{\varepsilon^2} \right),$$

then with probability at least $1 - \delta$, the following bounds hold simultaneously for all $j \in \{1, \dots, k\}$:

- ***Unregularized:*** $|\widetilde{\tau}_0(g, g'_{j,\perp})| \le \varepsilon \|g\|_2 \|g'_{j,\perp}\|_2 / \lambda^+_{\min}(F)$.
- ***Regularized:*** $|\widetilde{\tau}_\lambda(g, g'_{j,\perp})| \le \varepsilon \|g\|_2 \|g'_{j,\perp}\|_2 \big(\frac{1}{\lambda} + \frac{2\|F\|_2}{\lambda^2}\big)$ *for any* $\lambda > 0$,

Setup and Notation. We fix orthonormal bases $U_A \in \mathbb{R}^{d_A \times r_A}$ and $U_E \in \mathbb{R}^{d_E \times r_E}$ for $\mathrm{range}(A)$ and $\mathrm{range}(E)$, respectively, and write $U := U_A \otimes U_E$ for the induced orthonormal basis of $\mathrm{range}(F) = \mathrm{range}(A) \otimes \mathrm{range}(E)$ (so $r = \mathrm{rank}(F) = r_A r_E$).

For factorized test gradients $g' = a' \otimes e'$, we decompose $a' = a'_{/\!/} + a'_{\perp}$ with $a'_{/\!/} \in \mathrm{range}(A)$ and $a'_{\perp} \perp \mathrm{range}(A)$, and similarly $e' = e'_{/\!/} + e'_{\perp}$. The orthogonal projection of g' onto $\ker(F)$ is

$$g'_{\perp} = a'_{/\!/} \otimes e'_{\perp} + a'_{\perp} \otimes e'_{/\!/} + a'_{\perp} \otimes e'_{\perp}. \tag{7}$$

In particular, $g'_{\perp} \in \ker(F)$, so (as in the i.i.d. case) it suffices to analyze leakage terms with kernel components $g'_{\perp} \in \ker(F)$.

E.1 Proof Plan for Theorem 3.2

The factorized proof follows the same structure as the i.i.d. sketch case in Section D:

1. **Deterministic reduction to two concentration conditions**. By Theorem D.1, it is enough to verify (i) subspace concentration on $\mathrm{range}(F)$, i.e., $\|U^\top (P^\top P - I) U\|_2 \le \varepsilon$, and (ii) cross-term concentration $\|U^\top (P^\top P - I) g'_{j,\perp}\|_2 \le \varepsilon \|g'_{j,\perp}\|_2$ for the relevant kernel components $\{g'_{j,\perp}\}_{j=1}^k$.
2. **Stability on** $\mathrm{range}(F) = \mathrm{range}(A) \otimes \mathrm{range}(E)$. We control $\|U^\top (P^\top P - I) U\|_2$ by bounding the corresponding factor-level deviations on $\mathrm{range}(A)$ and $\mathrm{range}(E)$.
3. **Cross-term via Kronecker reduction with primitive bounds**. We expand $P^\top P - I$ into factor sketch deviations and use Theorem E.2 to reduce the cross-term $\|U^\top (P^\top P - I) g'_{j,\perp}\|_2$ to a collection of factor-level primitive quantities. These primitives are then controlled uniformly over the k test gradients using either a union bound over the nonzero out-of-range factor components (yielding the logarithmic dependence on k_A, k_E) or a subspace argument on their spans (yielding the k'_A, k'_E dependence); see Theorem E.3.
4. **Conclusion**. Plugging the primitive bounds into Theorem E.2 and then into Theorem D.1 yields Theorem 3.2.

The single-gradient proofs in Theorem D.2 (and the uniform extensions in Theorem D.3) depend on P only through two inequalities in Theorem D.1. In the factorized influence setting, the only additional step is to control the cross-term $\|U^\top (P^\top P - I) g'_{\perp}\|_2$ for $g'_{\perp} \in \ker(F)$ from factor-level primitive quantities. Define the factor sketch deviations

$$\Delta_A := P_A^\top P_A - I_{d_A}, \qquad \Delta_E := P_E^\top P_E - I_{d_E}.$$

A direct expansion shows

$$P^\top P - I_{d_A d_E} = \Delta_A \otimes I_{d_E} + I_{d_A} \otimes \Delta_E + \Delta_A \otimes \Delta_E. \tag{8}$$

The same expansion also makes the stability condition in Theorem D.1 explicit.

Lemma E.1. *Assume $\|U_A^\top \Delta_A U_A\|_2 \le \varepsilon$ and $\|U_E^\top \Delta_E U_E\|_2 \le \varepsilon$ for some $\varepsilon \in (0,1)$. Then with $U = U_A \otimes U_E$,*

$$\|U^\top (P^\top P - I) U\|_2 \le \|U_A^\top \Delta_A U_A\|_2 + \|U_E^\top \Delta_E U_E\|_2 + \|U_A^\top \Delta_A U_A\|_2 \|U_E^\top \Delta_E U_E\|_2 \le 3\varepsilon.$$

Proof. Using Eq.(8) and $U^\top = (U_A \otimes U_E)^\top = U_A^\top \otimes U_E^\top$, we have

$$U^\top (P^\top P - I) U = (U_A^\top \Delta_A U_A) \otimes I_{r_E} + I_{r_A} \otimes (U_E^\top \Delta_E U_E) + (U_A^\top \Delta_A U_A) \otimes (U_E^\top \Delta_E U_E).$$

Taking operator norms and using $\|X \otimes Y\|_2 = \|X\|_2 \|Y\|_2$ gives the claim. □

Lemma E.2. *Fix P_A, P_E (hence P), and let U_A, U_E be orthonormal bases for* $\mathrm{range}(A)$ *and* $\mathrm{range}(E)$*, and $U := U_A \otimes U_E$. Let $g' = a' \otimes e'$, decompose $a' = a'_{/\!/} + a'_{\perp}$ and $e' = e'_{/\!/} + e'_{\perp}$, and let $g'_{\perp}$ be the orthogonal projection of g' onto* $\ker(F)$ *given by Eq.*(7)*. Define $\Delta_A := P_A^\top P_A - I_{d_A}$ and $\Delta_E := P_E^\top P_E - I_{d_E}$.*

Then

$$\begin{aligned}\|U^\top(P^\top P - I)g'_\perp\|_2 \le{}& \|U_A^\top \Delta_A a'_\perp\|_2\|e'_{/\!/}\|_2 + \|a'_{/\!/}\|_2\|U_E^\top \Delta_E e'_\perp\|_2 \\ &+ \|U_A^\top \Delta_A a'_{/\!/}\|_2\|U_E^\top \Delta_E e'_\perp\|_2 + \|U_A^\top \Delta_A a'_\perp\|_2\|U_E^\top \Delta_E e'_{/\!/}\|_2 \\ &+ \|U_A^\top \Delta_A a'_\perp\|_2\|U_E^\top \Delta_E e'_\perp\|_2.\end{aligned} \tag{9}$$

In particular, if for some $\varepsilon \in (0,1)$,

$$\|U_A^\top \Delta_A x\|_2 \le \varepsilon\|x\|_2 \text{ for } x \in \{a'_{/\!/}, a'_\perp\}, \qquad \|U_E^\top \Delta_E y\|_2 \le \varepsilon\|y\|_2 \text{ for } y \in \{e'_{/\!/}, e'_\perp\}, \tag{10}$$

then

$$\|U^\top(P^\top P - I)g'_\perp\|_2 \le (2\varepsilon + 3\varepsilon^2)\big(\|a'_{/\!/}\|_2\|e'_\perp\|_2 + \|a'_\perp\|_2\|e'_{/\!/}\|_2 + \|a'_\perp\|_2\|e'_\perp\|_2\big) \le 5\sqrt{3}\varepsilon\|g'_\perp\|_2. \tag{11}$$

Proof. Start from the decompositions Eqs.(7) and (8):

$$(P^\top P - I)g'_\perp = (\Delta_A \otimes I + I \otimes \Delta_E + \Delta_A \otimes \Delta_E)\big(a'_{/\!/} \otimes e'_\perp + a'_\perp \otimes e'_{/\!/} + a'_\perp \otimes e'_\perp\big).$$

Expanding gives nine Kronecker products. Applying $U^\top = U_A^\top \otimes U_E^\top$ yields the explicit expansion

$$\begin{aligned}&U^\top(P^\top P - I)g'_\perp \\ &= \underbrace{\big(U_A^\top \Delta_A a'_{/\!/}\big) \otimes \big(U_E^\top e'_\perp\big)}_{(\Delta_A\otimes I)(a'_{/\!/}\otimes e'_\perp)} + \underbrace{\big(U_A^\top \Delta_A a'_\perp\big) \otimes \big(U_E^\top e'_{/\!/}\big)}_{(\Delta_A\otimes I)(a'_\perp\otimes e'_{/\!/})} + \underbrace{\big(U_A^\top \Delta_A a'_\perp\big) \otimes \big(U_E^\top e'_\perp\big)}_{(\Delta_A\otimes I)(a'_\perp\otimes e'_\perp)} \qquad (12)\\ &\quad + \underbrace{\big(U_A^\top a'_{/\!/}\big) \otimes \big(U_E^\top \Delta_E e'_\perp\big)}_{(I\otimes\Delta_E)(a'_{/\!/}\otimes e'_\perp)} + \underbrace{\big(U_A^\top a'_\perp\big) \otimes \big(U_E^\top \Delta_E e'_{/\!/}\big)}_{(I\otimes\Delta_E)(a'_\perp\otimes e'_{/\!/})} + \underbrace{\big(U_A^\top a'_\perp\big) \otimes \big(U_E^\top \Delta_E e'_\perp\big)}_{(I\otimes\Delta_E)(a'_\perp\otimes e'_\perp)} \qquad (13)\\ &\quad + \underbrace{\big(U_A^\top \Delta_A a'_{/\!/}\big) \otimes \big(U_E^\top \Delta_E e'_\perp\big)}_{(\Delta_A\otimes\Delta_E)(a'_{/\!/}\otimes e'_\perp)} + \underbrace{\big(U_A^\top \Delta_A a'_\perp\big) \otimes \big(U_E^\top \Delta_E e'_{/\!/}\big)}_{(\Delta_A\otimes\Delta_E)(a'_\perp\otimes e'_{/\!/})} + \underbrace{\big(U_A^\top \Delta_A a'_\perp\big) \otimes \big(U_E^\top \Delta_E e'_\perp\big)}_{(\Delta_A\otimes\Delta_E)(a'_\perp\otimes e'_\perp)}. \qquad (14)\end{aligned}$$

Since $U_A^\top a'_\perp = 0$ and $U_E^\top e'_\perp = 0$, four terms vanish, leaving the five nonzero contributions

$$\begin{aligned}&U^\top(P^\top P - I)g'_\perp \qquad (15)\\ &\quad = \big(U_A^\top \Delta_A a'_\perp\big) \otimes \big(U_E^\top e'_{/\!/}\big) + \big(U_A^\top a'_{/\!/}\big) \otimes \big(U_E^\top \Delta_E e'_\perp\big) + \big(U_A^\top \Delta_A a'_{/\!/}\big) \otimes \big(U_E^\top \Delta_E e'_\perp\big) \qquad (16)\\ &\qquad + \big(U_A^\top \Delta_A a'_\perp\big) \otimes \big(U_E^\top \Delta_E e'_{/\!/}\big) + \big(U_A^\top \Delta_A a'_\perp\big) \otimes \big(U_E^\top \Delta_E e'_\perp\big). \qquad (17)\end{aligned}$$

Taking Euclidean norms and using $\|u \otimes v\|_2 = \|u\|_2\|v\|_2$, together with $\|U_A^\top a'_{/\!/}\|_2 = \|a'_{/\!/}\|_2$ and $\|U_E^\top e'_{/\!/}\|_2 = \|e'_{/\!/}\|_2$, yields Eq.(9).

Under Eq. (10), the first two (single-factor) terms in Eq. (9) are bounded by $\varepsilon\|a'_\perp\|_2\|e'_{/\!/}\|_2$ and $\varepsilon\|a'_{/\!/}\|_2\|e'_\perp\|_2$, respectively. The last three (product) terms are each bounded by $\varepsilon^2\|a'_\cdot\|_2\|e'_\cdot\|_2$, where $a'_\cdot$ can be either $a'_{/\!/}$ or $a'_\perp$, same for $e'_\cdot$. Summing and regrouping gives the first inequality in Eq.(11).

For the second inequality, note that the three summands in Eq.(7) are pairwise orthogonal (since $a'_{/\!/} \perp a'_\perp$ and $e'_{/\!/} \perp e'_\perp$), so

$$\|g'_\perp\|_2^2 = \|a'_{/\!/}\|_2^2\|e'_\perp\|_2^2 + \|a'_\perp\|_2^2\|e'_{/\!/}\|_2^2 + \|a'_\perp\|_2^2\|e'_\perp\|_2^2.$$

By Cauchy–Schwarz,

$$\|a'_{/\!/}\|_2\|e'_\perp\|_2 + \|a'_\perp\|_2\|e'_{/\!/}\|_2 + \|a'_\perp\|_2\|e'_\perp\|_2 \le \sqrt{3}\|g'_\perp\|_2.$$

Since $\varepsilon \le 1$, we have $2\varepsilon + 3\varepsilon^2 \le 5\varepsilon$, yielding the second inequality in Eq.(11). □

Theorem E.2 shows that to apply Theorem D.1 in the factorized influence setting, it suffices to control factor-level deviations $\|U_A^\top \Delta_A(\cdot)\|_2$ and $\|U_E^\top \Delta_E(\cdot)\|_2$ on the relevant vectors. Once these are controlled with parameter ε, the cross-term condition $\|U^\top(P^\top P - I)g'_\perp\|_2 \le \widetilde{\varepsilon}\|g'_\perp\|_2$ holds with $\widetilde{\varepsilon} = O(\varepsilon)$.

E.2 Proof of Concentration of Factor-Level Primitives

We now show how to obtain the factor-level bounds Eq. (10) with high probability from the same concentration tools used in Section D.3. The key point is that the K-FAC structure allows us to control the relevant quantities by augmenting and controlling U_A and U_E separately, rather than working in dimension $d_A d_E$ directly.

Proposition E.3. *Let $\{g'_j\}_{j=1}^k$ with $g'_j = a'_j \otimes e'_j$, and let $g'_{j,\perp}$ be the projection onto $\ker(F)$. Define U_A, U_E as above and write $a'_j = a'_{j,/\!/} + a'_{j,\perp}$ and $e'_j = e'_{j,/\!/} + e'_{j,\perp}$. Denote*

$$k_A := \sum_{j=1}^{k} \mathbb{1}(a'_{j,\perp} \neq 0), \qquad k_E := \sum_{j=1}^{k} \mathbb{1}(e'_{j,\perp} \neq 0),$$

and also,

$$k'_A := \dim\big(\mathrm{span}(\{a'_{j,\perp}\}_{j=1}^k)\big), \qquad k'_E := \dim\big(\mathrm{span}(\{e'_{j,\perp}\}_{j=1}^k)\big).$$

Assume P_A and P_E satisfy the same sketch assumptions as in Theorem B.2 (independently across factors). Then, for any $\varepsilon, \delta \in (0,1)$, if

$$m_A = \Omega\left(\frac{r_A + \min\{\log(k_A/\delta), k'_A + \log(1/\delta)\}}{\varepsilon^2}\right)$$

and

$$m_E = \Omega\left(\frac{r_E + \min\{\log(k_E/\delta), k'_E + \log(1/\delta)\}}{\varepsilon^2}\right),$$

then with probability at least $1-\delta$, the following bounds hold simultaneously for all $j \in \{1, \dots, k\}$:

$$\|U_A^\top (P_A^\top P_A - I) a'_{j,/\!/}\|_2 \le \varepsilon \|a'_{j,/\!/}\|_2, \qquad \|U_A^\top (P_A^\top P_A - I) a'_{j,\perp}\|_2 \le \varepsilon \|a'_{j,\perp}\|_2,$$

and

$$\|U_E^\top (P_E^\top P_E - I) e'_{j,/\!/}\|_2 \le \varepsilon \|e'_{j,/\!/}\|_2, \qquad \|U_E^\top (P_E^\top P_E - I) e'_{j,\perp}\|_2 \le \varepsilon \|e'_{j,\perp}\|_2.$$

Consequently, the cross-term condition

$$\|U^\top (P^\top P - I) g'_{j,\perp}\|_2 \le 5\sqrt{3}\varepsilon \|g'_{j,\perp}\|_2$$

holds for all j simultaneously.[6]

Proof. We prove the A-factor bounds; the E-factor bounds are identical. Firstly, by Theorem B.2 applied to the r_A-dimensional subspace $\mathrm{range}(A)$, with probability at least $1 - \delta/4$,

$$\|U_A^\top (P_A^\top P_A - I) U_A\|_2 \le \varepsilon,$$

provided $m_A = \Omega(\varepsilon^{-2}(r_A + \log(4/\delta)))$. Next, to control $\|U_A^\top (P_A^\top P_A - I) a'_{j,\perp}\|_2$ uniformly over j, we use either:

(i) a union bound over the k_A nonzero vectors $\{a'_{j,\perp}\}$, giving a $\log k_A$ dependence, or

(ii) a subspace argument on $\mathrm{span}(\mathrm{range}(A) \cup \{a'_{j,\perp}\}_{j=1}^k)$, giving a dependence on k'_A. These two routes yield the stated $\min\{\log k_A, k'_A\}$ dependence.

Concretely, route (i) follows exactly as in Theorem D.4: for a fixed unit vector $v \perp \mathrm{range}(A)$, $\|U_A^\top (P_A^\top P_A - I) v\|_2 \le \varepsilon$ holds with probability at least $1 - \delta/(4\max\{k_A, 1\})$ provided

$$m_A = \Omega\left(\frac{r_A + \log(4\max\{k_A, 1\}/\delta)}{\varepsilon^2}\right)$$

and a union bound over the nonzero $a'_{j,\perp}$ gives the desired uniform control.

[6] Equivalently, one can run the primitive bounds Eq. (10) with accuracy $\varepsilon/(5\sqrt{3})$ to obtain a cross-term tolerance of ε; this only changes m_A, m_E by constant factors in the $\Omega(\cdot)$ conditions.

On the other hand, route (ii) is obtained by applying Theorem B.2 to the $(r_A + k'_A)$-dimensional subspace $\mathrm{span}(\mathrm{range}(A) \cup \{a'_{j,\perp}\}_{j=1}^{k})$, which yields the same uniform bound with

$$m_A = \Omega\left(\frac{r_A + k'_A + \log(4/\delta)}{\varepsilon^2}\right).$$

For $a'_{j,/\!/} \in \mathrm{range}(A)$, the desired inequality follows deterministically from the operator-norm event:

$$\begin{aligned}\|U_A^\top(P_A^\top P_A - I)a'_{j,/\!/}\|_2 &= \|U_A^\top(P_A^\top P_A - I)U_A(U_A^\top a'_{j,/\!/})\|_2 \\ &\le \|U_A^\top(P_A^\top P_A - I)U_A\|_2 \cdot \|a'_{j,/\!/}\|_2 \le \varepsilon\|a'_{j,/\!/}\|_2.\end{aligned}$$

Repeating the above argument for the E-factor and union bounding the A and E events gives the four primitive inequalities simultaneously for all j. The claimed cross-term bound then follows by Theorem E.2. □

On the event in Theorem E.3, Theorem E.2 gives the cross-term condition required by Theorem D.1. The stability condition on $\mathrm{range}(F)$ follows from the factor operator-norm events via Theorem E.1. Thus Theorem D.1 applies and yields the stated unregularized and regularized leakage bounds for $g'_{j,\perp}$, uniformly over j.