

Which Optimizer Leaks Most? Optimizer Implicit Bias as a Training-Data Privacy Controller

Sagi Habani*
Ben-Gurion University
sagihab@post.bgu.ac.il

Yehonathan Refael*
Tel Aviv University
refaelkalim@mail.tau.ac.il

Ofir Lindenbaum
Bar Ilan University
ofir.lindenbaum@biu.ac.il

Abstract

Releasing a neural network checkpoint can leak information about its training data, and existing reconstruction results rely on gradient-descent training (Haim et al., 2022). We ask whether the choice of optimizer affects the ability to recover the training data. Extending work connecting Karush–Kuhn–Tucker conditions to training-data reconstruction, we show that the implicit biases of popular optimizers (SGD, Adam, Muon) leave very different fingerprints of the training data. We argue that the key distinction is the *weight-specificity* of these implicit biases, namely how much continuous information about the final weights the optimizer-dependent target of the stationarity conditions encodes. The limiting stationarity conditions for models trained with SGD and Muon retain enough weight-specific information to reconstruct training examples in practice. For Adam in the ℓ_∞ implicit-bias regime, the optimizer-dependent part of these conditions encodes only a discrete pattern of maximal coordinates and their signs, and at nondegenerate KKT points the conditions are too few to identify the training set to first order. Experiments on MNIST and CIFAR-10 support these claims: models trained with Muon can yield stronger reconstructions at much lower training cost than models trained with SGD, although this advantage depends on the training regime and architecture, and models trained with Adam appear consistently more resistant to the evaluated attacks.

1 INTRODUCTION

Over-parameterized networks can perform well while memorizing training examples (Zhang et al., 2017), and good generalization may require memorizing atypical examples (Feldman, 2020), raising a privacy concern: how much memorized data can be recovered? Recently, Haim et al. (2022); Buzaglo et al. (2023) demonstrated the ability to reconstruct a large fraction of the training images used to train a classifier, using only the parameters of the trained classifier (without query access or other data). They achieve this by optimizing the input to satisfy a stationarity condition that relates the weights of the neural network to the training data. Merely sharing a trained model (a checkpoint) may therefore leak its training data.

This attack exploits the *implicit bias* of gradient descent (GD): for a homogeneous network trained with an exponentially tailed loss, the parameter direction converges, under suitable conditions, to that of a KKT point of the maximum-margin problem $\min_{\theta} \frac{1}{2} \|\theta\|_2^2$, s.t. $y_i f(\mathbf{x}_i; \theta) \geq 1$ for all i (Lyu and Li, 2020; Ji and Telgarsky, 2020). Its stationarity condition expresses the weights as a nonnegative combination of the signed gradients $y_i \nabla_{\theta} f(\mathbf{x}_i; \theta)$, and Haim et al. (2022) invert it by solving for inputs $\mathbf{x}_i$ with the weights fixed: the optimizer’s bias gives the attacker a system of equations relating the public weights to the training data.

Other optimizers induce other margin geometries: under suitable assumptions for smooth homogeneous models, Muon Jordan et al. (2024) maximizes a max-spectral-norm margin and Adam (Kingma and Ba, 2015), with a negligible stability constant, an ℓ_∞ margin (Gronich and Vardi, 2026). Different optimizers therefore induce different stationarity equations; we analyze how these differences affect data reconstruction and whether the choice of optimizer can reduce this risk.

We generalize the objective of Haim et al. (2022) to the

*Equal contribution.



Figure 1: Training-data reconstructions from networks trained with Muon, SGD and Adam, without an image prior ($n = 500$). The top row is binary MNIST and the bottom row binary CIFAR-10. In each panel, odd rows show reconstructions and the row directly below shows their nearest training images under the one-to-one matching of Section 5. Muon reconstructions are visually closest to their matches, SGD recovers coarser structure, and Adam reconstructions are largely unrecognizable.

margin geometry of the chosen optimizer (Section 3). A direct extension replaces its Euclidean stationarity term with a condition on the subdifferential of the margin norm. When applied naively, however, a formulation based on a single leading singular-vector pair may fail when the top singular value is repeated (a property encouraged by Muon) or when the relative block scales it fixes differ from those realized at a KKT point. It also breaks the rescaling symmetry of the KKT system, so it can be lowered by sending the multipliers to zero rather than by aligning the candidates with the data (Lemma A.7). To avoid this problem, we derive a scale-invariant penalty that enforces equality in the dual-norm inequality and use the layerwise Euler identity to show that no cross-block coupling term is needed.

Our analysis (Section 4) identifies two properties that help explain why the attack works. The first is *weight-specificity* – the number of independent continuous functionals of θ that the optimizer-dependent stationarity target determines. Equation counting at a generic weight vector does not distinguish the three geometries (there the target set has codimension $\Theta(p)$ in all three, where p is the parameter count), but specificity does. At a KKT direction, where every

block is active (Corollary A.2), and with simple top singular values ($\rho_k = 1$), S has the values $S_{\ell_2} = p-1$, $S_{\text{spectral}} = \sum_k (m_k + n_k - 2)$, $S_{\ell_\infty} = 0$, where block k has dimensions $m_k \times n_k$, so each spectral block contributes a number of functionals proportional to its width. For the ℓ_∞ margin, the optimizer-dependent target depends only on the locations and signs of the coordinates of θ with maximum absolute value, regardless of the multiplicity of the maximum. Moreover, at nondegenerate KKT points, the stationarity constraints impose at most as many equations as there are on-margin examples, far fewer than the unknowns, so the training set is not identifiable from them to first order.

The second property is *reachability* – the speed of convergence to the limiting parameter direction of the optimizer. Combining the two, we find the practical leakage ordering $\text{Muon} \succ \text{SGD} \succ \text{Adam}$ in our main experiments, where $\succ$ denotes greater reconstruction leakage (Section 5). The Muon comparison depends on the training budget, sample size, and architecture (SGD becomes comparable to Muon at small sample sizes and matches or exceeds it in some architectures), while Adam remains the most resistant to the evaluated attacks; we do not, however, provide a formal

privacy guarantee.

Our contributions are: (i) a reconstruction attack that depends on the optimizer’s margin norm, recovers the objective of Haim et al. (2022) in the Euclidean case up to the multiplier scale, and corrects two failure modes of the naive extension (Section 3); (ii) a theory of *weight-specificity* under which Adam’s ℓ_∞ target has zero specificity at every multiplicity of the maximum and its nondegenerate KKT points fail regular first-order identifiability (Section 4); (iii) an analysis combining specificity and *reachability* that explains the ordering $\text{Muon} \succ \text{SGD} \succ \text{Adam}$ in our main experiments, with experiments showing that the Muon comparison depends on the training regime while Adam remains the most resistant to the evaluated attacks (Section 5).

2 RELATED WORK

Reconstruction among privacy attacks. Haim et al. (2022) reconstruct training samples from homogeneous classifiers by inverting the ℓ_2 -margin KKT conditions; follow-up work extends this to the multi-class setting, weight decay and general losses (Buzaglo et al., 2023), to transfer learning (Oz et al., 2024), and to membership inference (Golbari et al., 2025), and examines the specific reliance of reconstruction attacks on inductive image priors (Refael et al., 2026b). All use the ℓ_2 implicit bias of gradient descent; we extend the formulation to other margin geometries and study how the optimizer affects the attack.

Implicit bias, and the geometry of Adam and Muon. Adam has a GD-like implicit bias when its stability constant dominates (Wang et al., 2021) and an ℓ_∞ margin bias when it is absent (Zhang et al., 2024); spectral descent and Muon have been analyzed in the multiclass linear setting (Fan et al., 2025). Under suitable assumptions, Gronich and Vardi (2026) extend these results to smooth homogeneous networks, relating idealized Muon to normalized momentum steepest descent under the *max-spectral* norm and Adam to the *global* ℓ_∞ margin, and leave open whether KKT-based reconstruction extends to these optimizers. We derive attacks for both geometries: reconstruction succeeds for Muon, while our analysis identifies a first-order identifiability limitation of Adam’s stationarity constraints, consistent with its greater resistance in our experiments.

2.1 Preliminaries: notation and threat model

Notation. We write $[n] := \{1, \dots, n\}$, $(a)_+ := \max\{a, 0\}$, $\mathbf{e}_j$ for the j -th standard basis vector, and

S^{d-1} for the unit sphere in $\mathbb{R}^d$. We denote the convex, affine, and conic hulls of a set X by $\text{conv } X$, $\text{aff } X$, and $\text{cone } X$, respectively. For a matrix A , $\sigma_1(A) \geq \sigma_2(A) \geq \dots$ are its singular values, $A \succeq 0$ means that A is positive semidefinite, and $\{M \succeq 0 : \text{tr } M = 1\}$ is a spectrahedron. We write $S^\circ := S \setminus \{0\}$.

Parameters, blocks and norms. The network has parameters $\boldsymbol{\theta} = (W_1, \dots, W_K)$, where $W_k \in \mathbb{R}^{m_k \times n_k}$ and each weight matrix is a *block*. We identify $\boldsymbol{\theta}$ with the concatenated vector in $\mathbb{R}^p$, where $p = \sum_k m_k n_k$. We write θ_j for its j -th coordinate and $\mathbf{u}_{W_k}$ for the k -th block of any $\mathbf{u} \in \mathbb{R}^p$, and use the Euclidean inner product $\langle \mathbf{u}, \mathbf{v} \rangle = \sum_k \text{tr}(\mathbf{u}_{W_k}^\top \mathbf{v}_{W_k})$. For each block, we use the Frobenius, spectral, and nuclear norms $\|W\|_F$, $\|W\|_{\text{sp}} = \sigma_1(W)$, and $\|W\|_{\text{nuc}} = \sum_i \sigma_i(W)$, respectively. On the full parameter vector, we use $\|\boldsymbol{\theta}\|_1$, $\|\boldsymbol{\theta}\|_2 = (\sum_k \|W_k\|_F^2)^{1/2}$, $\|\boldsymbol{\theta}\|_\infty = \max_j |\theta_j|$, and the *max-spectral* norm $\|\boldsymbol{\theta}\|_{\text{msp}} = \max_k \|W_k\|_{\text{sp}}$ of Gronich and Vardi (2026). An unadorned $\|\cdot\|$ denotes the *margin norm* of the optimizer under discussion (Eq. (P)), and $\|\cdot\|_\star$ denotes its *dual norm*, $\|\mathbf{u}\|_\star = \sup\{\langle \mathbf{u}, \mathbf{v} \rangle : \|\mathbf{v}\| \leq 1\}$. These norms satisfy $\langle \mathbf{u}, \mathbf{v} \rangle \leq \|\mathbf{u}\|_\star \|\mathbf{v}\|$, which we refer to as the *generalized Cauchy–Schwarz inequality*.

Alignment, gauge, strata. We write $\partial g(\boldsymbol{\theta})$ for the subdifferential of a convex function g at $\boldsymbol{\theta}$. Where g is differentiable, this is the singleton $\{\nabla g(\boldsymbol{\theta})\}$. For a norm, $\partial\|\boldsymbol{\theta}\| = \{\mathbf{u} : \|\mathbf{u}\|_\star \leq 1, \langle \mathbf{u}, \boldsymbol{\theta} \rangle = \|\boldsymbol{\theta}\|\}$. We say that $\mathbf{u}$ is *dual-aligned* with $\boldsymbol{\theta}$ when equality holds in generalized Cauchy–Schwarz:

$$\begin{aligned} \langle \mathbf{u}, \boldsymbol{\theta} \rangle &= \|\mathbf{u}\|_\star \|\boldsymbol{\theta}\|, \\ \text{equivalently } \mathbf{u}/\|\mathbf{u}\|_\star &\in \partial\|\boldsymbol{\theta}\| \quad (\mathbf{u} \neq \mathbf{0}). \end{aligned} \tag{1}$$

A *gauge* is a group of transformations of the unknowns that preserves the constraint system and maps solutions to solutions. Since the constraints cannot distinguish such solutions, we make our objective invariant along gauge directions.

Finally, we divide parameter space into *strata* on which the structural type of $\partial\|\boldsymbol{\theta}\|$ is constant (Appendix F.1).

Blockwise homogeneity. Our structural results use homogeneity layer by layer.

Assumption 2.1 (Blockwise homogeneity). For each block k there is $d_k > 0$ such that $f(\mathbf{x}; W_1, \dots, \alpha W_k, \dots, W_K) = \alpha^{d_k} f(\mathbf{x}; \boldsymbol{\theta})$ for all $\alpha > 0$, and $f(\mathbf{x}; \cdot)$ is differentiable at the points considered. Then f is L -homogeneous with $L = \sum_k d_k$.

Optimizer implicit bias. Homogeneous networks trained with logistic or exponential loss converge in

direction to a KKT point of a maximum-margin problem; this implicit bias is the key phenomenon underlying the attacks we study. For gradient flow, this follows from Theorem F.1 in Appendix F.3. Under the assumptions of Gronich and Vardi (2026), Adam and Muon converge to KKT directions of (P), with the optimizer-dependent margin norm given in (2).

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \|\boldsymbol{\theta}\|^2 \quad \text{s.t.} \quad y_i \Phi(\mathbf{x}_i; \boldsymbol{\theta}) \geq 1 \quad \forall i \in [n]. \quad (\text{P})$$

Since every count below depends on how that norm aggregates across blocks, we fix it explicitly:

$$\begin{aligned} \|\boldsymbol{\theta}\|_{\text{SGD}} &= \left(\sum_k \|W_k\|_F^2 \right)^{1/2}, \\ \|\boldsymbol{\theta}\|_{\text{Muon}} &= \max_k \frac{\|W_k\|_{\text{sp}}}{c_k}, \\ \|\boldsymbol{\theta}\|_{\text{Adam}} &= \|\boldsymbol{\theta}\|_{\infty}. \end{aligned} \quad (2)$$

with duals ℓ_2 , $\sum_k c_k \|\cdot\|_{\text{nuc}}$ and ℓ_1 . For $c_k \equiv 1$, the Muon form is $\|\cdot\|_{\text{msp}}$ used by Gronich and Vardi (2026) (their Corollary 3.3); general $c_k > 0$ are the per-block weights induced by practical Muon’s per-block step sizes (their Theorem 3.6). All our statements for Muon hold for arbitrary c_k , and the objective we construct does not depend on them. The Adam ℓ_{∞} norm is global over all p coordinates rather than per tensor. This margin norm is the *only* property of the optimizer the attack uses, and the only axis along which our three optimizers differ in the analysis.

Threat model. Following Haim et al. (2022), the adversary observes the final parameters $\boldsymbol{\theta}$ and the architecture and knows the optimizer *family*, hence $\|\cdot\|$; it observes neither the training inputs nor their labels or count n , nor any intermediate iterate. In particular $\boldsymbol{\theta}$ is *fixed and given*: the attacker may not rescale it (Section 3). As is standard, the attacker posits m candidates (we use $m = 2n$ following Haim et al. (2022), as an upper bound on n suffices) with fixed balanced labels and recovers a subset corresponding to true samples.

The gradient-descent attack (Haim et al., 2022). For ℓ_2 , $\partial(\frac{1}{2}\|\boldsymbol{\theta}\|_2^2) = \{\boldsymbol{\theta}\}$, so stationarity reads $\boldsymbol{\theta} = \sum_i \lambda_i y_i \nabla_{\boldsymbol{\theta}} \Phi(\mathbf{x}_i; \boldsymbol{\theta})$ with $\lambda_i \geq 0$, and complementary slackness forces $\lambda_i = 0$ off the margin. With $\boldsymbol{\theta}$ known and the inputs unknown, Haim et al. (2022) minimize over candidate inputs and multipliers

$$\begin{aligned} L_{\text{rec}} &= \underbrace{\left\| \boldsymbol{\theta} - \sum_i \lambda_i y_i \nabla_{\boldsymbol{\theta}} \Phi(\mathbf{x}_i; \boldsymbol{\theta}) \right\|_2^2}_{L_{\text{stat}}} \\ &\quad + a_2 \underbrace{\sum_i (-\lambda_i)_+^2}_{L_{\lambda}} + a_3 L_{\text{prior}}, \end{aligned} \quad (3)$$

where L_{λ} enforces $\lambda_i \geq 0$ and L_{prior} is a weak input prior such as a pixel-range penalty. Only L_{stat} depends on the optimizer, and only through $\partial(\frac{1}{2}\|\boldsymbol{\theta}\|^2)$. That is the starting point of the next section.

3 A NORM-PARAMETERIZED RECONSTRUCTION OBJECTIVE

We build the attack in three steps: the stationarity condition of (P) for a general margin norm (§3.1); removal of a non-informative degree of freedom, which yields an alignment condition and our loss (§3.2); and its instantiation for Muon (§3.3) and Adam (§3.4). Throughout, the trained $\boldsymbol{\theta}$ is fixed, the unknowns are $z := (\{\mathbf{x}_i\}_{i=1}^m, \{\lambda_i\}_{i=1}^m)$, and the residual is

$$r(z) := \sum_{i=1}^m \lambda_i y_i \nabla_{\boldsymbol{\theta}} \Phi(\mathbf{x}_i; \boldsymbol{\theta}). \quad (4)$$

3.1 Stationarity for a general norm

For (S)GD the margin norm is Euclidean, $\frac{1}{2}\|\boldsymbol{\theta}\|_2^2$ has gradient $\boldsymbol{\theta}$, and the stationarity term of (3) vanishes iff $\mathbf{r} = \boldsymbol{\theta}$. For Muon and Adam, $\frac{1}{2}\|\boldsymbol{\theta}\|^2$ is not differentiable, so we replace the gradient with the sub-differential; by the chain rule $\partial(\frac{1}{2}\|\boldsymbol{\theta}\|^2) = \|\boldsymbol{\theta}\| \partial\|\boldsymbol{\theta}\|$ (Lemma B.1), and the stationarity condition for (P) is

$$\mathbf{r}(z) \in \partial(\frac{1}{2}\|\boldsymbol{\theta}\|^2), \quad \lambda_i \geq 0. \quad (5)$$

To obtain a tractable loss, the following lemma characterizes this set-membership condition in terms of two scalar inequalities.

Lemma 3.1 (Membership rule). *For $\boldsymbol{\theta} \neq 0$, $\mathbf{r} \in \partial(\frac{1}{2}\|\boldsymbol{\theta}\|^2)$ if and only if $\|\mathbf{r}\|_{\star} \leq \|\boldsymbol{\theta}\|$ and $\langle \mathbf{r}, \boldsymbol{\theta} \rangle \geq \|\boldsymbol{\theta}\|^2$.*

Lemma 3.1 splits (5) into a *magnitude* and an *alignment* condition (generalizing (1)). Both $\|\mathbf{r}(z)\|_{\star}$ and $\langle \mathbf{r}(z), \boldsymbol{\theta} \rangle$ are computable and almost everywhere differentiable in z , and both inequalities hold with equality for feasible $\mathbf{r}$ (see the proof), so (5) can be enforced by penalizing $\|\mathbf{r}\|_{\star} = \|\boldsymbol{\theta}\|$ and $\langle \mathbf{r}, \boldsymbol{\theta} \rangle = \|\boldsymbol{\theta}\|^2$. As we show next, all information about $\{\mathbf{x}_i\}$ lies in the *alignment* requirement.

3.2 Two scale freedoms, and why only one may be discarded

The magnitude bound of Lemma 3.1 constrains the multipliers but says nothing about the candidate inputs, because the multipliers are themselves free unknowns.

Lemma 3.2 (The multiplier scale is a gauge). *The residual $\mathbf{r}(\cdot)$ is positively homogeneous of degree one in λ , and the constraint $\lambda \geq 0$ is preserved by $\lambda \mapsto$*

$c\lambda$ for $c > 0$. Hence for any z with $\mathbf{r}(z) \neq \mathbf{0}$ there is a unique $c > 0$ whose rescaled multipliers satisfy $\|\mathbf{r}\|_* = \|\boldsymbol{\theta}\|$ exactly. The magnitude bound therefore restricts no $\{\mathbf{x}_i\}$, and (5) is equivalent, modulo this one-dimensional gauge, to the cone membership

$$\begin{aligned} \mathbf{r}(z) \in C(\boldsymbol{\theta}) &:= \text{cone}(\partial(\tfrac{1}{2}\|\boldsymbol{\theta}\|^2)) \\ &= \{t\mathbf{k} : \mathbf{k} \in \partial(\tfrac{1}{2}\|\boldsymbol{\theta}\|^2), t \geq 0\}. \end{aligned} \quad (6)$$

Thus the magnitude of $\mathbf{r}$ is a scale convention, not a property of the data: everything the stationarity condition says about the training set is carried by the *direction* of $\mathbf{r}$, i.e. by the cone $C(\boldsymbol{\theta})$, the central object of Section 4. A reconstruction loss should therefore be invariant under $\lambda \mapsto c\lambda$; otherwise the attack can lower the loss merely by rescaling the multipliers (Lemma A.7). The second rescaling, $\boldsymbol{\theta} \mapsto c\boldsymbol{\theta}$, must be treated differently because the attacker cannot rescale the given $\boldsymbol{\theta}$ (Lemma A.1, Appendix A).

A layerwise Euler identity. Blockwise homogeneity ties every residual block to one shared scalar, for any candidate and not only at solutions.

Lemma 3.3 (Layerwise Euler identity). *Under Assumption 2.1, for every z and every block k ,*

$$\langle \mathbf{r}_{W_k}(z), W_k \rangle = d_k \Gamma(z), \quad \Gamma(z) := \sum_{i=1}^m \lambda_i y_i f(\mathbf{x}_i; \boldsymbol{\theta}). \quad (7)$$

If moreover $\mathbf{r}(z) \in C^\circ(\boldsymbol{\theta})$, then $\Gamma(z) = \langle \mathbf{r}(z), \boldsymbol{\theta} \rangle / L > 0$, hence $\mathbf{r}_{W_k}(z) \neq \mathbf{0}$ and $W_k \neq \mathbf{0}$ for every k ; for a solution of (5), $\Gamma(z) = \|\boldsymbol{\theta}\|^2 / L$.

For the max-type norms of Muon and Adam, this identity implies that every block is active at a KKT direction (Corollary A.2, Appendix A).

From cone membership to a loss. Membership in $C(\boldsymbol{\theta})$ is equivalent to tightness in the generalized Cauchy–Schwarz inequality (1), so we penalize the relative Cauchy–Schwarz gap block by block, with block norms $\|\cdot\|_{(k)}$ and their duals,

$$L_{\text{align}} = \sum_{k=1}^K \left(1 - \frac{\langle \mathbf{r}_{W_k}, W_k \rangle}{\|\mathbf{r}_{W_k}\|_{(k),*} \|W_k\|_{(k)}} \right)^2, \quad (8)$$

defined for $r_{W_k} \neq 0$ for every k .

Each term lies in $[0, 4]$ and vanishes iff $\mathbf{r}_{W_k}$ is dual-aligned with W_k , i.e. $\mathbf{r}_{W_k} \in C_k(\boldsymbol{\theta}) \setminus \{\mathbf{0}\}$, where $C_k(\boldsymbol{\theta}) := \text{cone} \partial \|W_k\|_{(k)}$. The numerator and denominator are positively homogeneous of degree one in each argument, so the loss is invariant under the gauge of Lemma 3.2 and requires no choice of singular vectors. By Lemma 3.3, $\mathbf{r}_{W_k} \neq \mathbf{0}$ in every block for any cone-feasible z , so this restriction discards no valid solution

(for numerical stability we add a small ϵ to the denominator). Plugging in the appropriate $\|\cdot\|_{(k)}$ covers all three attacks with one formulation.

Proposition 3.4 (Per-block alignment is exact). *Let Assumption 2.1 hold, let $\|\cdot\|$ be one of the norms in (2) (Muon with arbitrary $c_k > 0$), and let (5) have a solution at $\boldsymbol{\theta}$. Then*

$$\begin{aligned} \{z : \mathbf{r}(z) \in C^\circ(\boldsymbol{\theta})\} &= \{z : \mathbf{r}_{W_k}(z) \in C_k(\boldsymbol{\theta}) \setminus \{\mathbf{0}\} \ \forall k\} \\ &= \{z : L_{\text{align}}(z) = 0\}. \end{aligned}$$

In particular, the zero set of L_{align} does not depend on the Muon weights c_k .

Hence no cross-block coupling term is needed (Remark A.3, Appendix A).

To mount the attack we set $L_{\text{stat}} := L_{\text{align}}$ in (3), keeping L_λ and L_{prior} unchanged. For ℓ_2 , Proposition 3.4 gives $L_{\text{stat}} = 0$ iff $\mathbf{r} = c\boldsymbol{\theta}$ for a shared $c > 0$, so up to the gauge of Lemma 3.2 the Euclidean case recovers the objective of Haim et al. (2022).

3.3 The Muon (spectral) objective

Muon performs steepest descent under the spectral norm: for each block W , $\|\cdot\| = \|\cdot\|_{\text{sp}} = \sigma_1(\cdot)$, with dual $\|\cdot\|_{\text{nuc}}$. Its subdifferential is a spectrahedron rather than a point: if the top singular value has multiplicity ρ , with left and right singular subspaces spanned by the columns of $U_1 \in \mathbb{R}^{m \times \rho}$ and $V_1 \in \mathbb{R}^{n \times \rho}$, then

$$\partial \|W\|_{\text{sp}} = \{U_1 M V_1^\top : M \succeq 0, \text{tr } M = 1\}, \quad (9)$$

which reduces to the rank-one $\mathbf{p}\mathbf{q}^\top$ only when $\rho = 1$. A natural first instinct is to keep the template of (3) with a spectral target, minimizing the “single-target” loss $\|\mathbf{k}^* - \mathbf{r}_W\|_2^2$ for $\mathbf{k}^* = \nabla(\frac{1}{2}\|W\|_{\text{sp}}^2) = \sigma_1 \mathbf{p}\mathbf{q}^\top$. This breaks down for two reasons: *within* a block, a repeated top singular value makes the target basis-dependent, and *across* blocks, the relative block scales it fixes need not match those realized at a KKT point (Lemma A.7, Appendix A.8). The Muon instance of (8), in which c_k cancels, reads

$$L_{\text{align}}^{\text{Muon}} = \sum_k \left(1 - \frac{\langle \mathbf{r}_{W_k}, W_k \rangle}{\|\mathbf{r}_{W_k}\|_{\text{nuc}} \|W_k\|_{\text{sp}}} \right)^2. \quad (10)$$

Since every block attains the maximum at a KKT direction (Corollary A.2), the extra tightness condition of the max-type norm is vacuous (Lemma B.3) and (10) is exact for arbitrary per-block weights (Proposition 3.4): no balancing assumption is needed, and every layer carries information.

3.4 The Adam (ℓ_∞) objective

Adam with a negligible stability constant maximizes the *global* ℓ_∞ margin, so $\|\boldsymbol{\theta}\| = \|\boldsymbol{\theta}\|_\infty$ with dual $\|\cdot\|_1$.

With $A = \arg \max_j |\theta_j|$ and $s = |A|$, Hölder tightness holds iff $\mathbf{r}_j = 0$ for $j \notin A$ and $\mathbf{r}_j \theta_j \geq 0$ for $j \in A$, with $\mathbf{r} \neq 0$: the candidate gradient combination must be supported on the largest-magnitude weights with matching signs, and nothing else is asked of it. We use the single global term

$$(1 - \langle \mathbf{r}, \boldsymbol{\theta} \rangle / (\|\mathbf{r}\|_1 \|\boldsymbol{\theta}\|_\infty))^2 \quad (11)$$

(a per-tensor variant is discussed in Appendix A.11). Why this target is substantially less weight-specific is the subject of the next section.

4 HOW MUCH DO THE OPTIMIZERS LEAK? A SPECIFICITY ANALYSIS

The attack seeks z whose residual lies in $C^\circ(\boldsymbol{\theta})$, and succeeds to the extent that this *forces* the candidates to be the training set. Counting the scalar equations that $\mathbf{r} \in C(\boldsymbol{\theta})$ imposes at a generic $\boldsymbol{\theta}$ gives $\Theta(p)$ for all three geometries and cannot explain the ordering (Appendix A.4); we instead ask how much of $\boldsymbol{\theta}$ the optimizer-dependent target $C(\boldsymbol{\theta})$ encodes.

4.1 Weight-specificity and target-cone variability

Counting tells us how many equations there are, not how much they say about the weights. The constraint references the weights through the residual map $z \mapsto \mathbf{r}(z; \boldsymbol{\theta})$, fixed by the architecture and $\boldsymbol{\theta}$, and through the target $C(\boldsymbol{\theta})$, the only place where the optimizer enters.

Definition 4.1 (Weight-specificity). *For a margin norm $\|\cdot\|$ and a stratum Σ , the weight-specificity $S_\Sigma(\|\cdot\|)$ is the dimension of the image of $\boldsymbol{\theta} \mapsto C(\boldsymbol{\theta})$ restricted to the points of Σ at which all blocks are active, i.e. the number of independent continuous functionals of $\boldsymbol{\theta}$ that the stationarity target determines.*

Specificity is a property of the geometry and the stratum, not of a dataset: it counts the continuous degrees of freedom through which the target cone varies with the weights, and is not by itself an information-theoretic bound on input identifiability.

Lemma 4.2 (Specificity of the three geometries). *On the stratum with multiplicities ρ_k and arg-max set A , at points where all blocks are active,*

$$S = \begin{cases} p - 1, & \ell_2, \\ \sum_k S_k(\rho_k), & \text{spectral}, \\ 0, & \ell_\infty, \end{cases} \quad (12)$$

since C is the ray through $\boldsymbol{\theta}$, determined by $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2$ (ℓ_2); $C_k = \{U_{1,k} M V_{1,k}^\top : M \succeq 0\}$, determined by

$U_{1,k} V_{1,k}^\top$ (spectral); and $C = \{\mathbf{r} : \mathbf{r}_j = 0 \text{ for } j \notin A, \mathbf{r}_j \theta_j \geq 0 \text{ for } j \in A\}$ (ℓ_∞). Here $S_k(\rho) := \rho(m_k + n_k - \rho) - \rho(\rho + 1)/2$ is the dimension of the manifold of rank- ρ partial isometries in $\mathbb{R}^{m_k \times n_k}$; $S_k(1) = m_k + n_k - 2$. In the ℓ_∞ case C depends on $\boldsymbol{\theta} \neq \mathbf{0}$ only through the discrete data $(A, \text{sign } \theta|_A)$, for every multiplicity $s = |A|$.

4.2 Consequences, and why Adam resists

For SGD the cone is the ray through $\boldsymbol{\theta}$, so the constraint encodes the full direction $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2$: the candidate gradients must reproduce the entire weight vector up to one scalar. This is the maximally specific geometry, and the regime in which Haim et al. (2022) reconstruct. For **Muon**, C_k encodes the top singular structure $U_{1,k} V_{1,k}^\top$ of *every* block (all are active at a KKT direction), i.e. $S_k(\rho_k)$ numbers per layer, at least $\Theta(m_k + n_k)$ (Appendix A.9). That is far less than ℓ_2 , but the constraint stays *tight*: by Lemma A.4 many of each block's equations survive at any multiplicity in ordinary matrix-block regimes, and it is attached to a weight-specific template that grows with width, so Muon both constrains candidates sharply and carries a genuine data fingerprint. For **Adam**, with $A = \arg \max_l |\theta_l|$,

$$\partial(\tfrac{1}{2}\|\boldsymbol{\theta}\|_\infty^2) = \|\boldsymbol{\theta}\|_\infty \text{conv}\{\text{sign}(\theta_j) \mathbf{e}_j : j \in A\}, \quad (13)$$

an $(s-1)$ -simplex whose *interior directions are unconstrained*: the condition pins the support and admissible signs of $\mathbf{r}$ but never the relative magnitudes of $\boldsymbol{\theta}$. Hence $S = 0$ (cf. Remark A.8).

For networks the residual map carries $\boldsymbol{\theta}$ -dependence of its own, but the polyhedral geometry has a second consequence that does extend: minimizing $\|\boldsymbol{\theta}\|_\infty$ under margin constraints drives $\boldsymbol{\theta}$ onto a large face of the ℓ_∞ ball, where the constraint has few normal equations. Under a first-order nondegeneracy condition (Assumption A.9), Theorem A.10 (Appendix A.10) shows that $\text{codim } C(\boldsymbol{\theta}) = p - s \leq n_I$, where n_I is the number of on-margin examples. Adam's stationarity system thus has at most n_I normal equations for $n_I(d+1)$ unknowns and, when $n_I d \geq 2$, fails regular first-order identifiability; under additional constant-rank and relative-interior conditions, the stationarity-consistent perturbations of the on-margin candidates form a C^1 manifold of dimension at least $n_I d$. The result concerns the idealized ℓ_∞ bias (Adam with a negligible stability constant) and this stationarity-only attack family only, so it is not a differential-privacy-style guarantee (Abadi et al., 2016).

4.3 Two axes, and the vulnerability ordering

Specificity concerns a *converged* model, but the attack inverts a condition that holds only in the limit, so a second axis matters: *reachability*, how quickly an optimizer’s parameter direction approaches the KKT direction of its own margin problem. Muon is high on both axes: its specificity is at least $\Theta(\text{width})$ in every Muon-trained layer, and per-block orthogonalization escapes the small-initialization stall that traps gradient descent. SGD has the *highest* specificity, $p-1$, but the *slowest* reachability (Haim et al., 2022 train $\sim 10^6$ iterations): vulnerable in principle, expensive in practice. Adam has fast reachability but zero continuous target specificity and, at nondegenerate KKT points, too few equations for regular first-order identification (Theorem A.10). Folding in cost, the *practical* ordering is

$$\boxed{\text{Muon} \succ \text{SGD} \succ \text{Adam}.}$$

The Muon order is a prediction from reachability, not a theorem: that gap is about training *cost* and should narrow wherever SGD has budget to converge, while the Adam gap reflects weaker continuous target dependence and regular underdetermination (summary in Table 3, Appendix B).

5 EXPERIMENTS

We evaluate the attack of Section 3 on networks trained with SGD, Adam, and Muon and test the predictions of Section 4. Implementation details are in Appendix D and full results in Appendix E.

Setup. Unless stated otherwise, we use $n = 500$ examples from binary MNIST LeCun et al. (2010) parity, binary CIFAR-10 Krizhevsky et al. (2009) vehicles versus animals, and 10-class CIFAR-10. Models are fully connected ReLU networks $d-1000-1000-C$ ($C = 1$ binary, $C = 10$ multiclass), with only a first-layer bias, absorbed into W_1 , so Assumption 2.1 holds. SGD and Adam update all layers; Muon applies five-step Newton–Schulz Muon (Jordan et al., 2024) to hidden-to-hidden matrices and SGD to the input and output layers. We use the binary SGD models of Haim et al. (2022). Models trained by Adam and Muon reach training loss $O(10^{-7})$ within 10^4 iterations, whereas to reach the same order of training loss, SGD requires 10^6 iterations (Table 4, Appendix D).

Attack and evaluation. The attack uses the Euclidean loss (3) for SGD, and the global ℓ_∞ alignment loss (§3.4) for Adam, and a smooth relaxation of (10) for Muon. It optimizes $m = 2n$ candidates with fixed balanced labels and their multipliers; the multiclass extension is described in Appendix D. Unless stated

otherwise, no image prior is used. Reconstructions are matched one-to-one to training images by the Hungarian algorithm on centered, normalized ℓ_2 distance (more details in Appendix D). We evaluate SSIM@30 as our primary leakage metric. All-pairs averages can favor Adam because its candidates are generic images; @30 measures whether individual samples are recovered. We report SSIM, PSNR, and ℓ_1 and ℓ_2 distances over all pairs and the best 30 pairs (@30); SSIM > 0.4 is considered visually recognizable.

5.1 Main Results

Muon-trained models are the most vulnerable.

Muon leads to the best reconstructions on every metric and task as visually illustrated for the binary classification setup in Figure 1. Its SSIM@30 is 0.431 on MNIST, 0.551 on binary CIFAR-10 and 0.534 on 10-class CIFAR-10, against 0.346, 0.229 and 0.383 for SGD (the 10-class gap is discussed in Appendix E). The gap is largest on binary CIFAR-10, where Muon more than doubles the SGD score. Since no prior is used, these reconstructions come from the stationarity condition alone, consistent with the fact that the spectral objective (10) inverts the Muon condition and that the per-layer spectral template of Lemma 4.2 is exploitable, with two orders of magnitude fewer training iterations than SGD. This is confirmed by the sorted SSIM curves without priors across both CIFAR-10 and MNIST (Figure 6).

Adam-trained models are the most resistant.

Adam has the lowest SSIM@30 on every task without a prior, at most 0.249, well below the 0.4 recognizability level, consistent with the analysis of §4.2 : Adam’s ℓ_∞ target carries no continuous information about the weight magnitudes, leaving the stationarity system too weakly constrained to identify individual samples.

Ablation Studies: The effect of image priors.

We add five regularizers to the attack objective: a range prior penalizing pixel values outside $[0, 1]$, total variation (TV), an ℓ_1 penalty favoring low-energy backgrounds, $\text{TV} + \ell_1$, and all three combined. As the only optimizer-independent component of the objective, the prior’s marginal effect measures how much of a reconstruction does not come from the weights (Table 1; full results in Tables 5–7), consistent with recent findings on how priors can dominate checkpoint reconstruction fidelity Refael et al. (2026b). Qualitative comparisons showing reconstructions with and without priors are provided in Figure 3 for MNIST and Figure 5 for CIFAR-10.

Priors do not change the leading optimizer: Muon is best under every prior (SSIM@30 between 0.431 and

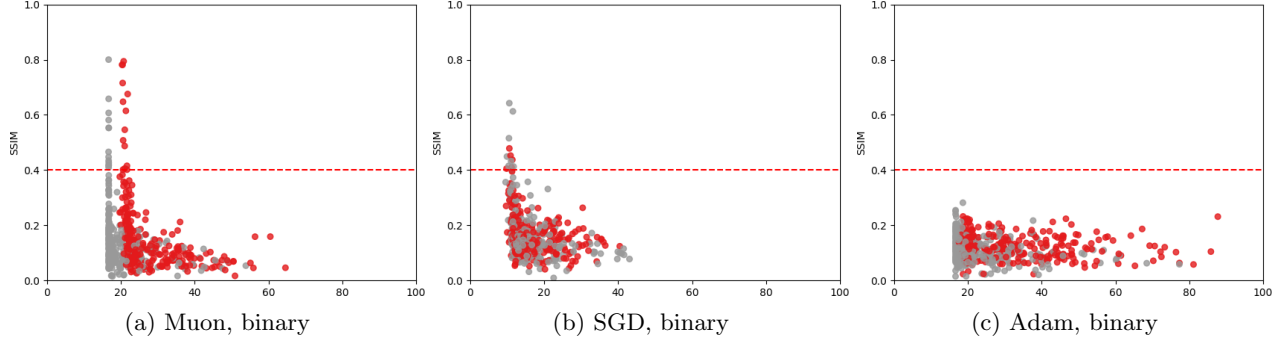


Figure 2: Per-sample SSIM versus training margin γ_i on CIFAR-10 ($n = 500$, range prior), for binary models. Each point is a training sample, colored by class; the dashed line marks $\text{SSIM} = 0.4$. Recognizable reconstructions occur near the minimal margin of their class. Muon produces most of them, and Adam essentially none.

Table 1: SSIM@30 ($\uparrow$) under different image priors ($n = 500$). Best per task and prior in **bold**, second best underlined. Muon is best in all 18 settings, and in all 144 entries of the full tables (Appendix E).

Prior	MNIST (binary)			CIFAR-10 (binary)			CIFAR-10 (10-class)		
	SGD	Adam	Muon	SGD	Adam	Muon	SGD	Adam	Muon
None	0.3463	0.2487	0.4310	0.2287	0.1379	0.5508	0.3827	0.2066	0.5339
Range	0.4057	0.2994	0.6371	0.2336	0.1367	0.6130	0.4515	0.1771	0.5315
TV	0.3768	0.2882	0.5808	0.2574	0.2115	0.6132	0.4332	0.3442	0.5481
ℓ_1	0.3614	0.1752	0.5302	0.2375	0.2112	0.5598	0.4141	0.2739	0.5371
TV+ ℓ_1	0.3977	0.2311	0.6211	0.2440	0.2369	0.6286	0.4343	0.3050	0.5890
All	0.4511	0.2655	0.6599	0.2536	0.2377	0.6189	0.4343	0.2603	0.5795

0.660), while Adam never exceeds 0.344. On MNIST, combining all priors raises SSIM@30 by 0.229 for Muon and 0.105 for SGD but only 0.017 for Adam: priors sharpen reconstructions when the stationarity condition already constrains them. On binary CIFAR-10, smoothness priors help Adam most (from 0.138 to 0.238 with all priors). We attribute this to smoothness priors making generic natural-image candidates more similar to the training set, rather than to identification of specific samples, and to SGD not having reached its limiting direction on this task (Table 4).

Effect of sample size and architecture. To remove the training-budget asymmetry, we retrain all models on binary CIFAR-10 to a training loss below 10^{-8} , varying $n \in \{50, 100, 500, 1000\}$ or the hidden widths, and attack them with the range prior (Table 2).

Across sample sizes, Muon is best on every metric and Adam has the lowest SSIM@30. Once SGD is trained to convergence, however, it becomes competitive at small n : at $n = 100$ its SSIM@30 is 0.623 against 0.698 for Muon, compared with 0.234 against 0.613 at $n = 500$. The Muon gap is therefore partly a gap in training cost, as predicted by the reachability argument of §4.3.

Adam is worst on every metric for every architecture,

Table 2: Effect of sample size (top, 1000–1000 hidden units) and architecture (bottom, $n = 500$) on binary CIFAR-10. All models are trained to a training loss below 10^{-8} ; the attack uses the range prior. Best per configuration in **bold**.

Configuration	Optimizer	SSIM $\uparrow$	SSIM@30 $\uparrow$	PSNR $\uparrow$	PSNR@30 $\uparrow$	$\ell_1 \downarrow$	ℓ_1 @30 $\downarrow$	$\ell_2 \downarrow$	ℓ_2 @30 $\downarrow$
$n = 50$	SGD	0.3024	0.4486	13.188	15.066	0.1586	0.1250	0.2191	0.1765
	Adam	0.1452	0.1809	11.733	12.663	0.1989	0.1808	0.2590	0.2327
	Muon	0.4070	0.5848	14.409	17.390	0.1353	0.0954	0.1904	0.1351
$n = 100$	SGD	0.2949	0.6232	13.499	18.765	0.1554	0.0828	0.2114	0.1153
	Adam	0.0976	0.1488	11.214	12.570	0.2087	0.1775	0.2750	0.2352
	Muon	0.3501	0.6984	14.257	21.465	0.1404	0.0639	0.1937	0.0845
$n = 500$	SGD	0.0994	0.2336	11.665	13.435	0.2027	0.1634	0.2611	0.2129
	Adam	0.1299	0.1367	12.334	13.441	0.1863	0.1624	0.2417	0.2128
	Muon	0.1921	0.613	12.8084	19.1901	0.175	0.0825	0.2289	0.1098
$n = 1000$	SGD	0.0942	0.3001	11.537	14.168	0.2059	0.1476	0.2650	0.1957
	Adam	0.0783	0.1202	11.150	11.956	0.2149	0.1968	0.2770	0.2525
	Muon	0.1442	0.4186	12.203	15.559	0.1895	0.1252	0.2454	0.1667
100–100	SGD	0.1132	0.1754	11.878	13.373	0.1975	0.1634	0.2547	0.2145
	Adam	0.0569	0.0836	10.795	12.129	0.2242	0.1912	0.2886	0.2475
	Muon	0.1195	0.1826	11.921	14.001	0.1957	0.1598	0.2535	0.2039
1000–500–100	SGD	0.1654	0.4170	12.577	15.990	0.1808	0.1197	0.2351	0.1587
	Adam	0.0927	0.1226	11.489	12.318	0.2065	0.1836	0.2664	0.2422
	Muon	0.1689	0.4234	12.832	16.182	0.1719	0.1170	0.2263	0.1552
1000–500–100–50	SGD	0.1617	0.3944	12.594	15.882	0.1804	0.1196	0.2346	0.1607
	Adam	0.1181	0.1546	11.944	12.736	0.1950	0.1741	0.2528	0.2308
	Muon	0.1896	0.4236	13.5685	17.962	0.1622	0.1085	0.2192	0.1396

with SSIM@30 below half that of the better optimizer. The Adam separation is thus robust and the Muon comparison regime-dependent, as the theory suggests: Adam’s weakness follows from the structure of the ℓ_∞ target cone, whereas the comparison between the ℓ_2 and spectral geometries depends on the residual map and the training budget.

Which Samples Are Reconstructed? By complementary slackness, only samples on the margin carry positive multipliers and enter the stationarity condition. Figure 2 tests this by plotting, for each training sample, the SSIM of its matched reconstruction against its margin γ_i . In both the binary and multiclass settings (Figure 4), recognizable reconstructions occur, for every optimizer, only for samples at or near the minimal margin, and samples well inside the margin are never recovered. Muon produces by far the most samples above the threshold, many concentrated at the minimal margin (a narrow vertical band); SGD produces a handful and Adam essentially none. Being

on the margin is necessary for recovery under every geometry, but it is sufficient only when the stationarity condition carries weight-specific information, which the ℓ_∞ condition does not.

6 DISCUSSION

Our results identify optimizer implicit bias as an overlooked axis of training-data privacy. Even with identical architectures, data, and interpolation behavior, optimizers leave distinct geometric fingerprints: Muon’s spectral bias tightly constrains candidate inputs and yields the strongest reconstructions, SGD leaks less, and Adam is most resistant. This theoretically predicted ordering—driven by the specificity and reachability of stationarity constraints—shows that optimizer choice can serve as a practical privacy control, not merely an optimization decision. Although our checkpoint-only, stationarity-based analysis provides no formal privacy guarantee, it demonstrates that evaluating model privacy independently of the training algorithm can overlook a consequential source of leakage.

Finally, while our study focuses on standard implementations of SGD, Adam, and Muon, modern training pipelines increasingly adopt memory-efficient, low-rank, and orthogonalized optimizers. Extending geometric implicit-bias analyses to methods such as AdaRankGrad Refael et al. (2025b), SUMO Refael et al. (2025a), Refael et al. (2026a), and Svirsky et al. (2026) presents a promising direction to understand how adaptive rank reduction and subspace orthogonalization govern training-data leakage.

AI use statement

In this work, we used generative AI tools for text polishing and obtaining suggestions on document organization and section structure. We have not used generative AI tools for research idea generation, code development, mathematical derivations, or experimental data generation/analysis, and all other disclosure categories are not applicable to this work. We have reviewed all AI-assisted work. Specifically, all AI-suggested structural edits and polished text were thoroughly examined, manually refined, and verified for technical accuracy and consistency by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *ACM SIGSAC*

Conference on Computer and Communications Security (CCS).

- Battash, B., Wolf, L., and Lindenbaum, O. (2024). Revisiting the noise model of stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 4780–4788. PMLR.
- Buzaglo, G., Haim, N., Yehudai, G., Vardi, G., Oz, Y., Nikankin, Y., and Irani, M. (2023). Deconstructing data reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.
- Fan, C., Schmidt, M., and Thrampoulidis, C. (2025). Implicit bias of spectral descent and muon on multiclass separable data.
- Feldman, V. (2020). Does learning require memorization? a short tale about a long tail. In *ACM Symposium on Theory of Computing (STOC)*.
- Fredrikson, M., Jha, S., and Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*.
- Golbari, Y., Wasserman, N., Vardi, G., and Irani, M. (2025). Impmia: Leveraging implicit bias for membership inference attack.
- Gronich, E. and Vardi, G. (2026). The implicit bias of adam and muon on smooth homogeneous neural networks. In *International Conference on Machine Learning (ICML)*.
- Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. (2018). Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning (ICML)*.
- Haim, N., Vardi, G., Yehudai, G., Shamir, O., and Irani, M. (2022). Reconstructing training data from trained neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification.
- Ji, Z. and Telgarsky, M. (2020). Directional convergence and alignment in deep learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17176–17186. Curran Associates, Inc.
- Jordan, K. et al. (2024). Muon: An optimizer for hidden layers in neural networks.
- Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

- Krizhevsky, A., Nair, V., and Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.
- LeCun, Y., Cortes, C., and Burges, C. (2010). Mnist handwritten digit database. *ATT Labs [Online]*, 2.
- Lyu, K. and Li, J. (2020). Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*.
- Oz, Y., Yehudai, G., Vardi, G., Antebi, I., Irani, M., and Haim, N. (2024). Reconstructing training data from real-world models trained with transfer learning.
- Refael, Y., Arbel, I., Lindenbaum, O., and Tirer, T. (2026a). Lorenza: Enhancing generalization in low-rank gradient LLM training via efficient zeroth-order adaptive SAM. *Transactions on Machine Learning Research*.
- Refael, Y., Smorodinsky, G., Lindenbaum, O., and Safran, I. (2026b). No prior, no leakage: Revisiting reconstruction attacks in trained neural networks. In *The Fourteenth International Conference on Learning Representations*.
- Refael, Y., Smorodinsky, G., Tirer, T., and Lindenbaum, O. (2025a). SUMO: Subspace-aware moment-orthogonalization for accelerating memory-efficient LLM training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Refael, Y., Svirsky, J., Shustin, B., Huleihel, W., and Lindenbaum, O. (2025b). Adarankgrad: Adaptive gradient rank and moments for memory-efficient LLMs training and fine-tuning. In *The Thirteenth International Conference on Learning Representations*.
- Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (S&P)*.
- Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. (2018). The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research (JMLR)*.
- Svirsky, J., Refael, Y., and Lindenbaum, O. (2026). Train less, infer faster: Efficient model finetuning and compression via structured sparsity. In *The 29th International Conference on Artificial Intelligence and Statistics*.
- Vardi, G. (2023). On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6).
- Watson, G. A. (1992). Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45.
- Wang, B., Meng, Q., Chen, W., and Liu, T.-Y. (2021). The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *International Conference on Machine Learning (ICML)*.
- Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*.
- Zhang, C., Zou, D., and Cao, Y. (2024). The implicit bias of adam on separable data. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Zhu, L., Liu, Z., and Han, S. (2019). Deep leakage from gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Supplementary Materials

A ADDITIONAL ANALYSIS OF RECONSTRUCTION OBJECTIVES AND PRIVACY LEAKAGE

A.1 Scale covariance in the trained weights (from §3.2)

Besides the multiplier gauge of Lemma 3.2, there is another rescaling invariance, on the weights, that should be treated differently.

Lemma A.1 (Scale covariance in θ). *If $z = (\{\mathbf{x}_i\}, \{\lambda_i\})$ satisfies (5) at θ , then for every $c > 0$ the pair $(\{\mathbf{x}_i\}, \{c^{2-L}\lambda_i\})$ satisfies it at $c\theta$.*

In line with Haim et al. (2022), Lemma A.1 justifies disregarding the norm of θ in the homogeneous stationarity equation: every θ on the ray remains a valid stationary point up to a rescaling of the multipliers. Importantly, this invariance does not preserve primal feasibility or complementary slackness under the rigid margin constraint “ ≥ 1 ” in (P).

A.2 Every block is active (from §3.2)

For the max-type norms governing Muon and Adam, the Euler identity (Lemma 3.3) enforces a direct structural constraint.

Corollary A.2 (Every block is active at a KKT direction). *Let Assumption 2.1 hold and let (5) have a solution at θ . (a) For the Muon norm of (2), every block attains the maximum: $\|W_k\|_{\text{sp}}/c_k = \|\theta\|$ for all k . (b) For the Adam norm, every block contains a globally maximal coordinate: $A \cap W_k \neq \emptyset$ and $\|W_k\|_\infty = \|\theta\|_\infty$ for all k , where $A = \arg \max_j |\theta_j|$.*

A.3 Why per-block alignment loses nothing (from §3.2)

At first glance, optimizing a per-block loss appears to relax the global condition. Under an ℓ_2 norm, for example, it permits $\mathbf{r}_{W_k} = t_k W_k$ using a distinct scalar $t_k > 0$ for each block. For max-type norms, it seemingly overlooks which specific blocks actually attain the maximum. However, the Euler identity proves that at the exact weights an attacker encounters in a real scenario, this per-block approach remains completely exact (Proposition 3.4).

Remark A.3 (No coupling term is needed). One might add $L_{\text{couple}} = \text{Var}_k [\log(\|\mathbf{r}_{W_k}\|_{(k),*}/\|W_k\|_{(k)})]$ to restore the relative block scales that a per-block loss seems to drop. On the zero set of L_{align} , (7) gives $\|\mathbf{r}_{W_k}\|_{(k),*}\|W_k\|_{(k)} = \langle \mathbf{r}_{W_k}, W_k \rangle = d_k \Gamma(z)$, so $L_{\text{couple}} = \text{Var}_k [\log(d_k/\|W_k\|_{(k)}^2)]$ is a function of θ alone. It therefore cannot discriminate between candidates satisfying $L_{\text{align}} = 0$, and we do not use it.

A.4 Equation counting at a generic weight vector does not separate the geometries

The constraint $\mathbf{r} \in C(\theta)$ imposes $\text{codim } C(\theta)$ scalar equalities, plus sign constraints. By Lemma B.4 and Corollary A.2, at a KKT direction on the stratum of θ ,

$$\text{codim } C(\theta) = \begin{cases} p-1, & \ell_2 \text{ } (C = \text{ray through } \theta), \\ \sum_k (m_k n_k - \rho_k(\rho_k + 1)/2), & \text{spectral } (C = \prod_k \{U_{1,k} M V_{1,k}^\top : M \succeq 0\}), \\ p-s, & \ell_\infty \text{ } (C = \text{simplicial cone on } A, s = |A|). \end{cases} \quad (14)$$

At a generic θ ($\rho_k = 1$, $s = 1$) these are $p-1$, $p-K$ and $p-1$, all $\Theta(p)$ in the usual nontrivial matrix-block regime, so counting at a generic weight vector explains no ordering. But KKT points are not generic, and the geometries respond very differently to degeneracy. For the spectral norm, higher multiplicity costs few equations:

Lemma A.4 (The spectral constraint cannot collapse). *For every θ and all multiplicities, with $r_k = \min(m_k, n_k)$,*

$$\sum_k (m_k n_k - \rho_k (\rho_k + 1)/2) \geq \frac{1}{2} (p - \sum_k r_k).$$

For ℓ_∞ , in contrast, $p - s$ can be as small as 0, and §4.2 shows that at the KKT points Adam reaches it is at most the number of on-margin examples. The relevant comparison is with the unknowns. Each on-margin candidate contributes $d+1$ of them (input and multiplier), and the true data can satisfy the regular first-order identifiability criterion only if $\text{codim } C(\theta) \geq n_I(d+1) - 1$, where n_I is the number of on-margin examples (Lemma C.2). This is necessary, not sufficient: off-margin candidates contribute nothing ($\lambda_i = 0$), the prior and $\lambda \geq 0$ add restrictions, and recovering a *subset* of the points needs no unique global solution. The ℓ_2 and spectral geometries meet it with room to spare in the over-parameterized regime $p \gg nd$, and it is exactly what fails for ℓ_∞ .

A.5 Target-cone sufficiency for linear-in-parameter models (from §4.1)

When the residual map does not depend on θ , the checkpoint affects the stationarity-feasible set only through the target cone. This occurs for models linear in their parameters, $f(\mathbf{x}; \theta) = \langle \theta, \phi(\mathbf{x}) \rangle$ for a fixed feature map ϕ (random-feature, kernel and linearized-network models), where $\mathbf{r}(z) = \sum_i \lambda_i y_i \phi(\mathbf{x}_i)$.

Lemma A.5 (Target-cone sufficiency for linear-in-parameter models). *Let $f(\mathbf{x}; \theta) = \langle \theta, \phi(\mathbf{x}) \rangle$. If $C(\theta) = C(\theta')$, then the stationarity-feasible sets*

$$\{z : \mathbf{r}(z) \in C^\circ(\theta)\} \quad \text{and} \quad \{z : \mathbf{r}(z) \in C^\circ(\theta')\}$$

coincide, as do the zero sets of the corresponding global alignment losses. Hence, within the stationarity-only formulation, the checkpoint affects feasibility only through $C(\theta)$. On each stratum these cones vary in an S -dimensional continuous family. This invariance does not alone imply non-identifiability of the inputs, which also depends on the residual map ϕ and on the codimension of the cone.

For such models $C(\theta)$ is a sufficient statistic of the weights for this stationarity-only formulation. For networks, $\mathbf{r}(z; \theta) = \sum_i \lambda_i y_i \nabla_\theta f(\mathbf{x}_i; \theta)$ depends on θ as well, so two networks with equal cones but different weights have different residual maps and $C(\theta)$ is no longer sufficient. Specificity then measures only the part of the constraint that the optimizer controls; the network-level regular first-order result is Theorem A.10. Attackers with extra information (checkpoints, query access, a strong generative prior) are outside the scope of either.

A.6 Stratum invariance under ℓ_∞ (from §4.2)

Proposition A.6 (Stratum invariance under ℓ_∞). *Let $f(\mathbf{x}; \theta) = \langle \theta, \phi(\mathbf{x}) \rangle$. The stationarity-feasible set for the Adam geometry depends on θ only through $(A, \text{sign } \theta|_A)$. Consequently, checkpoints sharing this arg-max/sign pattern induce exactly the same stationarity equations and the same feasible candidate datasets. Thus the target contributes no continuous dependence on the weight magnitudes within a fixed stratum. This invariance does not by itself imply non-identifiability of the inputs, since the equations $\mathbf{r}(z) \in C(\theta)$ may still constrain or identify z through the feature map ϕ .*

For these models Adam’s target cone transmits an $O(p)$ -bit combinatorial arg-max/sign pattern but no continuous information about the weight magnitudes within a fixed stratum. This makes the target less weight-specific, although the resulting stationarity equations may still constrain the inputs through the feature map.

A.7 Discussion of Theorem A.10 (from §4.2)

On Assumption A.9: The span condition says that θ is first-order isolated given its active set: no direction moves θ while preserving all tied maxima and all active margins to first order. For linear-in-parameter models it holds whenever the ℓ_∞ max-margin linear program has a unique solution, since a unique LP solution is a vertex (Remark B.5). It is checkable on a trained model through its consequence $s \geq p - n_I$.

Thus, under Assumption A.9, Adam’s stationarity system has at most n_I normal equations for $n_I(d+1)$ unknowns and fails regular first-order identifiability. Under the additional constant-rank and relative-interior hypotheses of part (iii), this first-order deficiency becomes an actual positive-dimensional family of stationarity-consistent candidates, whose input projection has dimension at least $n_I(d-1)$. Without those additional hypotheses, the

rank count alone does not exclude an isolated solution. These conclusions concern the stationarity-only attack, not the full KKT system or other reconstruction attacks. The ℓ_2 and spectral constraints have many more normal equations in ordinary over-parameterized matrix-block regimes (Lemma A.4). Three caveats remain. The results concern the idealized ℓ_∞ bias, i.e. Adam with a negligible stability constant. The network statement rests on Assumption A.9, a property of the trained model, and actual local ambiguity additionally rests on the hypotheses of Theorem A.10 (iii). Finally, they cover *this* attack family only, so they are not a differential-privacy-style guarantee (Abadi et al., 2016).

Specificity and Theorem A.10 explain why Adam is less exposed to this stationarity-only attack, but they do not show that ℓ_2 or spectral geometry *succeeds*. Regular first-order well-posedness in the attacker’s variables is formalized in Appendix C as a rank condition on the Jacobian of the constrained residual at z^* . Of the resulting Conjecture C.4, the Adam failure of regular first-order identifiability follows from Theorem A.10 (Proposition C.3); for the positive half we prove only the necessary condition $\text{codim } C(\theta) \geq n_I(d+1) - 1$, which the ℓ_2 and spectral geometries satisfy in the stated over-parameterized regime.

A.8 Set-valued target and collapse (from §3.3)

Degeneracy within blocks: spectral-margin maximization drives blocks toward repeated top singular values, and then $\mathbf{p}, \mathbf{q}$ are defined only up to rotation within the corresponding top singular subspaces U_1, V_1 . Thus the single-target construction is basis-dependent and is not selected by the KKT condition.

Degeneracy across blocks: by Corollary A.2 every block attains the maximum at a KKT direction, so $\partial(\frac{1}{2}\|\theta\|^2)$ is set-valued even when every $\rho_k = 1$. It contains a $(K-1)$ -simplex of relative block weights, and the element that solutions actually realize is singled out only through the Euler identity, with relative block scales that depend on the weights c_k and the degrees d_k .

Lemma A.7 (Set-valued target and collapse). *Let Assumption 2.1 hold and let (5) have a solution at θ for the Muon norm with weights c_k . Then:*

- (a) *Each term of (10) is invariant under $\mathbf{r}_{W_k} \mapsto a_k \mathbf{r}_{W_k}$ and $W_k \mapsto b_k W_k$ for all $a_k, b_k > 0$;*
- (b) *$\partial(\frac{1}{2}\|\theta\|^2) = \{(\|\theta\| \mu_k c_k^{-1} G_k)_k : \mu \in \Delta_K, G_k \in \partial\|W_k\|_{\text{sp}}\}$, a convex set of dimension $(K-1) + \sum_k (\rho_k(\rho_k + 1)/2 - 1)$, where Δ_K is the probability simplex. Every solution of (5) has $\mathbf{r}_{W_k} = (d_k \sigma_1(W_k)/(L c_k^2)) G_k$ with $G_k \in \partial\|W_k\|_{\text{sp}}$. Hence if all $\rho_k = 1$, solutions realize a single point $\mathbf{k}^\circ$, which is a positive multiple of the naive target $\mathbf{k}^*$ iff d_k/c_k^2 is constant in k ; if some $\rho_k > 1$, the KKT condition selects no unique point target. Although $U_{1,k} V_{1,k}^\top / \rho_k$ is a canonical basis-independent choice, imposing it would add a restriction not implied by stationarity;*
- (c) *for any fixed target $\mathbf{k}$ and direction $\mathbf{s}$, on the ray $\mathbf{r} = t\mathbf{s}$, $t \geq 0$, the fixed-target loss $\|\mathbf{k} - t\mathbf{s}\|^2$ is minimized at $t^* = (\langle \mathbf{k}, \mathbf{s} \rangle)_+ / \|\mathbf{s}\|^2$. So if $\langle \mathbf{k}, \mathbf{s} \rangle \leq 0$ the minimizer is $\mathbf{r} = \mathbf{0}$, and otherwise the minimum is $\|\mathbf{k}\|^2 \sin^2 \angle(\mathbf{k}, \mathbf{s})$, zero iff $\mathbf{s}$ is a positive multiple of $\mathbf{k}$. Each term of (10) is constant along the ray.*

Part (a) shows that the alignment loss is blind to both per-block scales, so by Proposition 3.4 its zero set is the exact solution set, up to the gauge, for any c_k , which the attacker need not know. Part (b) identifies what the naive loss gets wrong. Within a block with $\rho_k > 1$ it must select one element of a spectrahedron, even though stationarity does not specify which element the true residual realizes. Across blocks its relative scales are correct only when d_k/c_k^2 is constant; otherwise $L_{\text{single}} = 0$ has no solution at all, even for the true data. Part (c) is a landscape property shared by every fixed-target loss, including the Euclidean (3): when the residual is negatively correlated with the target, the cheapest descent is to send the multipliers to zero, switching the candidates off rather than aligning them with data. The alignment loss is constant along multiplier rays and has no such incentive. Being restricted to one ray, (c) is a mechanism rather than a proof of failure on the full domain.

A.9 Remarks on Lemma 4.2 (from §4.1)

Since S_k is concave in ρ , $S_k(\rho_k) \geq \min\{S_k(1), S_k(r_k)\}$, and $S_k(r_k) \geq r_k(r_k - 1)/2$. In particular $S_k(\rho) \geq m_k + n_k - 2$ for every ρ whenever $r_k \geq 4$: higher multiplicity never erases Muon’s template. Specificity thus separates the continuous target-cone variability of the geometries where generic counting could not. It is also consistent with §3: the magnitudes $\sigma_1(W_k)$ are absent from (12) because the cone is scale-free, the same fact that makes the block scales of a fixed target a nuisance (Lemma A.7(b)). A loss cannot recover continuous target information the constraint does not contain, although the fixed cone can still constrain inputs through

the residual map.

Remark A.8 (Linear-in-parameter models). For $f(\mathbf{x}; \boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \phi(\mathbf{x}) \rangle$ the residual map does not depend on $\boldsymbol{\theta}$, so the checkpoint affects the stationarity-feasible set only through $C(\boldsymbol{\theta})$ (Lemma A.5); under the Adam geometry this set depends on $\boldsymbol{\theta}$ only through $(A, \text{sign } \boldsymbol{\theta}|_A)$ (Proposition A.6). Neither statement alone implies non-identifiability of the inputs (Appendix A).

A.10 Collapse of the ℓ_∞ constraint (from §4.2)

For networks the residual map carries $\boldsymbol{\theta}$ -dependence of its own. The polyhedral geometry, however, has a second consequence that does extend. Minimizing $\|\boldsymbol{\theta}\|_\infty$ under margin constraints drives $\boldsymbol{\theta}$ onto a large face of the ℓ_∞ ball (for $a^\top \boldsymbol{\theta} \geq 1$ with all $a_j \neq 0$, the solution $\text{sign}(a)/\|a\|_1$ has *every* coordinate maximal), and on such a face the constraint has few normal equations. We make this precise under a first-order nondegeneracy condition. Let z^* be the true solution, $I = \{i : \lambda_i^* > 0\}$ and $n_I = |I|$.

Assumption A.9 (Nondegenerate ℓ_∞ KKT point). Strict complementarity holds (the examples with $\lambda_i^* > 0$ are exactly those on the margin), and the vectors $\{\mathbf{e}_j : j \in A\} \cup \{\nabla_{\boldsymbol{\theta}} f(\mathbf{x}_i^*; \boldsymbol{\theta}) : i \in I\}$ span $\mathbb{R}^p$.

Theorem A.10 (Collapse of the ℓ_∞ constraint). *Under Assumption A.9, with $\mathbf{r}$ continuously differentiable in z near z^* :*

- (i) $s \geq p - n_I$, hence $\text{codim } C(\boldsymbol{\theta}) = p - s \leq n_I$;
- (ii) let $z_I = (\{\mathbf{x}_i\}, \{\lambda_i\})_{i \in I} \in \mathbb{R}^{n_I(d+1)}$, the remaining candidates being held at z^* , and $\Phi(z_I) := P_N \mathbf{r}(z)$ with P_N the projection onto the coordinates outside A . Then $\text{rank } D\Phi(z_I^*) \leq n_I$; if $n_I d \geq 2$, this is less than $n_I(d+1) - 1$, so z^* is not regularly first-order identifiable (Definition C.1);
- (iii) if moreover $D\Phi$ has constant rank near z_I^* and $\mathbf{r}_j(z^*) \neq 0$ for $j \in A$, the set of z_I near z_I^* with $\lambda_i > 0$ and $\mathbf{r}(z) \in C^\circ(\boldsymbol{\theta})$ is a C^1 manifold through z_I^* of dimension at least $n_I d$. Arbitrarily close to z_I^* , its projection onto the input coordinates contains C^1 manifolds of dimension at least $n_I(d-1)$.

See Appendix A.7 for a discussion of the assumption and of the result.

A.11 Per-tensor form of the Adam objective (from §3.4)

Besides the single global term, we also evaluate the *per-tensor* form (8) with $\|\cdot\|_{(k)}$ the ℓ_∞ norm on block k . For general $\boldsymbol{\theta}$ the per-tensor form is a strict relaxation, since it permits a nonzero residual in tensors containing no globally maximal coordinate. At a KKT direction every tensor contains one (Corollary A.2(b)), and the two forms have the same zero set (Proposition 3.4).

B PROOFS

Throughout, $\|\cdot\|$ is the optimizer’s implicit-bias (margin) norm, $\|\cdot\|_*$ its dual, and $h(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|^2$. We collect two elementary lemmas (§B.1), then prove the membership test (§B.2), the scale statements and the Euler identity (§B.3), exactness of per-block alignment (§B.4), the set-valued-target lemma (§B.5), the block decomposition (§B.6), and the specificity and collapse results (§B.7).

Table 3: Summary. The cone $C(\boldsymbol{\theta})$ is the optimizer-dependent part of the constraint and, for linear-in-parameter models, is the part through which the checkpoint affects stationarity feasibility (Lemma A.5). Codimensions are at a KKT direction on the stratum of $\boldsymbol{\theta}$; at a generic $\boldsymbol{\theta}$ all three are $\Theta(p)$ in the usual nontrivial matrix-block regime and separate nothing. $S_k(\rho) = \rho(m_k + n_k - \rho) - \rho(\rho + 1)/2$, $r_k = \min(m_k, n_k)$.

optimizer	cone $C(\boldsymbol{\theta})$	codim at KKT stratum	specificity S	reachability
SGD	ray through $\boldsymbol{\theta}$	$p - 1$	$p - 1$	slow
Muon	$\prod_k \{U_{1,k} M V_{1,k}^\top : M \succeq 0\}$	$\geq \frac{1}{2}(p - \sum_k r_k)$	$\sum_k S_k(\rho_k)$	fast
Adam	simplicial cone on A , signs of $\boldsymbol{\theta} _A$	$p - s \leq n_I$ (Thm. A.10)	0 (combinatorial only)	fast

B.1 Two helper lemmas

Lemma B.1 (Subdifferential of a squared norm). *For any norm $\|\cdot\|$ with dual $\|\cdot\|_*$ and any θ ,*

$$\partial h(\theta) = \|\theta\| \partial \|\theta\|, \quad \partial \|\theta\| = \{\mathbf{u} : \|\mathbf{u}\|_* \leq 1, \langle \mathbf{u}, \theta \rangle = \|\theta\|\}.$$

Moreover $\partial \|\cdot\|$ is 0-homogeneous, so $\partial \|c\theta\| = \partial \|\theta\|$ and $\partial h(c\theta) = c \partial h(\theta)$ for every $c > 0$.

Proof. Write $h = g \circ N$ with $N = \|\cdot\|$ and $g(t) = \frac{1}{2}t^2$. The outer function g is convex, differentiable and nondecreasing on $[0, \infty) \supseteq \text{range}(N)$, so the convex chain rule gives $\partial h(\theta) = g'(N(\theta)) \partial N(\theta) = N(\theta) \partial N(\theta)$. The displayed form of ∂N is standard, from $N(\theta) = \sup_{\|\mathbf{u}\|_* \leq 1} \langle \mathbf{u}, \theta \rangle$. For $c > 0$, $\mathbf{u} \in \partial \|c\theta\|$ iff $\|\mathbf{u}\|_* \leq 1$ and $\langle \mathbf{u}, c\theta \rangle = \|c\theta\| = c\|\theta\|$, i.e. iff $\mathbf{u} \in \partial \|\theta\|$. Hence $\partial h(c\theta) = c\|\theta\| \partial \|\theta\| = c \partial h(\theta)$. $\square$

Lemma B.2 (Gradient homogeneity). *If $f(\mathbf{x}; \cdot)$ is L -homogeneous and differentiable in θ , then $\nabla_{\theta} f(\mathbf{x}; \alpha\theta) = \alpha^{L-1} \nabla_{\theta} f(\mathbf{x}; \theta)$ for all $\alpha > 0$.*

Proof. Differentiate $f(\mathbf{x}; \alpha\theta) = \alpha^L f(\mathbf{x}; \theta)$ with respect to θ : the left-hand side is $\alpha \nabla_{\theta} f(\mathbf{x}; \alpha\theta)$ and the right-hand side $\alpha^L \nabla_{\theta} f(\mathbf{x}; \theta)$; divide by $\alpha > 0$. For ReLU networks the identity holds wherever the gradient exists, and everywhere for the smooth surrogate used in the attack. $\square$

B.2 Membership test (Lemma 3.1)

Proof. ($\Leftarrow$) The two hypotheses and the generalized Cauchy-Schwarz inequality $\langle \mathbf{r}, \theta \rangle \leq \|\mathbf{r}\|_* \|\theta\|$ chain into $\|\theta\|^2 \leq \langle \mathbf{r}, \theta \rangle \leq \|\mathbf{r}\|_* \|\theta\| \leq \|\theta\|^2$, which forces all three to be equalities. Thus $\|\mathbf{r}\|_* = \|\theta\|$ and $\langle \mathbf{r}, \theta \rangle = \|\theta\|^2$, so $\mathbf{v} := \mathbf{r}/\|\theta\|$ satisfies $\|\mathbf{v}\|_* = 1$ and $\langle \mathbf{v}, \theta \rangle = \|\theta\|$, that is $\mathbf{v} \in \partial \|\theta\|$ and $\mathbf{r} = \|\theta\| \mathbf{v} \in \|\theta\| \partial \|\theta\|$ by Lemma B.1. ($\Rightarrow$) is immediate from the same two defining properties of $\partial \|\theta\|$. $\square$

B.3 Scale statements and the Euler identity

Proof of Lemma 3.2. The residual $\mathbf{r}(z) = \sum_i \lambda_i y_i \nabla_{\theta} f(\mathbf{x}_i; \theta)$ is linear in λ , so replacing $\lambda \mapsto c\lambda$ with $c > 0$ gives $\mathbf{r} \mapsto c\mathbf{r}$ and preserves $\lambda \geq 0$; if $\mathbf{r} \neq \mathbf{0}$, the choice $c := \|\theta\|/\|\mathbf{r}\|_*$ achieves $\|c\mathbf{r}\|_* = \|\theta\|$, and it is the only one. For the equivalence, fix the inputs. ($\Rightarrow$) If $\mathbf{r} \in \partial h(\theta)$ then $\mathbf{r} \in C(\theta)$, and $\mathbf{r} \neq \mathbf{0}$ because $\langle \mathbf{r}, \theta \rangle = \|\theta\|^2 > 0$ by Lemma 3.1. ($\Leftarrow$) If $\mathbf{r} \in C^o(\theta)$ then $\mathbf{r} = t\mathbf{k}$ with $t > 0$ and $\mathbf{k} \in \partial h(\theta)$; replacing λ by $\lambda/t \geq 0$ turns the residual into $\mathbf{k}$, a solution of (5). No homogeneity of f is used. $\square$

Proof of Lemma A.1. Fix $c > 0$ and set $\lambda'_i = c^{2-L} \lambda_i \geq 0$. By Lemma B.2 with $\alpha = c$,

$$\mathbf{r}' = \sum_i \lambda'_i y_i \nabla_{\theta} f(\mathbf{x}_i; c\theta) = \sum_i c^{2-L} \lambda_i y_i c^{L-1} \nabla_{\theta} f(\mathbf{x}_i; \theta) = c\mathbf{r}.$$

By hypothesis $\mathbf{r} \in \partial h(\theta)$, and by Lemma B.1 $\partial h(c\theta) = c \partial h(\theta)$; hence $\mathbf{r}' \in \partial h(c\theta)$, which together with $\lambda'_i \geq 0$ is (5) at $c\theta$. This proves covariance of stationarity; primal feasibility and complementary slackness for the fixed unit margin are not asserted. $\square$

Proof of Lemma 3.3. Fix $\mathbf{x}$ and k . Differentiating $\alpha \mapsto f(\mathbf{x}; W_1, \dots, \alpha W_k, \dots, W_K) = \alpha^{d_k} f(\mathbf{x}; \theta)$ at $\alpha = 1$ gives Euler's identity $\langle \nabla_{W_k} f(\mathbf{x}; \theta), W_k \rangle = d_k f(\mathbf{x}; \theta)$. Multiplying by $\lambda_i y_i$ at $\mathbf{x} = \mathbf{x}_i$ and summing over i gives (7). Summing (7) over k gives $\langle \mathbf{r}(z), \theta \rangle = L \Gamma(z)$. If $\mathbf{r}(z) \in C^o(\theta)$, write $\mathbf{r} = t\mathbf{k}$ with $t > 0$ and $\mathbf{k} \in \partial h(\theta)$; then $\langle \mathbf{r}, \theta \rangle = t\|\theta\|^2 > 0$ by Lemma 3.1, so $\Gamma(z) > 0$ and $\langle \mathbf{r}_{W_k}, W_k \rangle = d_k \Gamma(z) > 0$ for every k , which forces $\mathbf{r}_{W_k} \neq \mathbf{0}$ and $W_k \neq \mathbf{0}$. For a solution of (5), $t = 1$ and $\Gamma = \|\theta\|^2/L$. $\square$

Proof of Corollary A.2. Let z solve (5); by Lemma 3.3, $\mathbf{r}_{W_k}(z) \neq \mathbf{0}$ for all k . (a) Lemma B.3(b) requires $\mathbf{r}_{W_k} = \mathbf{0}$ for every block outside $\arg \max_j \|W_j\|_{\text{sp}}/c_j$, so every block is in the arg-max. (b) By Lemma B.4(ii), every element of $C(\theta)$ is supported on A ; since $\mathbf{r}_{W_k} \neq \mathbf{0}$, block k meets A , so $\|W_k\|_{\infty} = \|\theta\|_{\infty}$. $\square$

B.4 Exactness of per-block alignment

Proof of Proposition 3.4. The second equality holds for any θ . A summand of (8) vanishes iff $\langle \mathbf{r}_{W_k}, W_k \rangle = \|\mathbf{r}_{W_k}\|_{(k),*} \|W_k\|_{(k)}$ with $\mathbf{r}_{W_k} \neq \mathbf{0}$, i.e. by (1) applied to the block norm, iff $\mathbf{r}_{W_k}/\|\mathbf{r}_{W_k}\|_{(k),*} \in \partial\|W_k\|_{(k)}$, i.e. iff $\mathbf{r}_{W_k} \in C_k(\theta) \setminus \{\mathbf{0}\}$. For Muon, $\partial(\|\cdot\|_{\text{sp}}/c_k) = c_k^{-1}\partial\|\cdot\|_{\text{sp}}$, so C_k and the summand do not depend on c_k .

Max-type norms (Muon, Adam). Write $\|\theta\| = \max_k N_k(W_k)$ with $N_k = \|\cdot\|_{(k)}$. By Corollary A.2 all blocks attain the maximum, so the subdifferential of a pointwise maximum of convex functions is the convex hull of the active subdifferentials: $\partial\|\theta\| = \text{conv} \bigcup_k \{\iota_k \mathbf{g} : \mathbf{g} \in \partial N_k(W_k)\}$, with ι_k the embedding of block k . Since $\|\theta\| > 0$, $C(\theta) = \text{cone} \partial\|\theta\| = \{\sum_k t_k \iota_k \mathbf{g}_k : t_k \geq 0, \mathbf{g}_k \in \partial N_k(W_k)\} = \prod_k C_k(\theta)$. Hence if $\mathbf{r} \in C^\circ(\theta)$ then $\mathbf{r}_{W_k} \in C_k(\theta)$ for all k , and $\mathbf{r}_{W_k} \neq \mathbf{0}$ by Lemma 3.3. Conversely, if $\mathbf{r}_{W_k} \in C_k(\theta) \setminus \{\mathbf{0}\}$ for all k , then $\mathbf{r} \in \prod_k C_k(\theta) = C(\theta)$ and $\mathbf{r} \neq \mathbf{0}$.

ℓ_2 . Here $C(\theta) = \{t\theta : t \geq 0\}$ and $C_k(\theta) = \{tW_k : t \geq 0\}$. Let z^* solve (5), i.e. $\mathbf{r}(z^*) = \theta$; by Lemma 3.3, $\|W_k\|_F^2 = \langle \mathbf{r}_{W_k}(z^*), W_k \rangle = d_k \Gamma^*$ with $\Gamma^* = \|\theta\|^2/L > 0$. ($\subseteq$) If $\mathbf{r} = t\theta$ with $t > 0$, then $\mathbf{r}_{W_k} = tW_k \neq \mathbf{0}$. ($\supseteq$) If $\mathbf{r}_{W_k} = t_k W_k$ with $t_k > 0$ for all k , then (7) gives $t_k \|W_k\|_F^2 = d_k \Gamma(z)$, hence $t_k = \Gamma(z)/\Gamma^*$ for every k , and $\mathbf{r} = t\theta$ with $t = \Gamma(z)/\Gamma^* > 0$. $\square$

Verification of Remark A.3. On the zero set of L_{align} , every block is tight, so $\|\mathbf{r}_{W_k}\|_{(k),*} \|W_k\|_{(k)} = \langle \mathbf{r}_{W_k}, W_k \rangle = d_k \Gamma(z)$ by (7). Hence $\log(\|\mathbf{r}_{W_k}\|_{(k),*} \|W_k\|_{(k)}) = \log \Gamma(z) + \log(d_k/\|W_k\|_{(k)}^2)$, and the variance over k removes the k -independent term $\log \Gamma(z)$. $\square$

B.5 Set-valued target and collapse

Proof of Lemma A.7. (a) Numerator and denominator are positively 1-homogeneous in each argument: $\langle a_k \mathbf{r}_{W_k}, b_k W_k \rangle = a_k b_k \langle \mathbf{r}_{W_k}, W_k \rangle$, $\|a_k \mathbf{r}_{W_k}\|_{\text{nuc}} = a_k \|\mathbf{r}_{W_k}\|_{\text{nuc}}$ and $\|b_k W_k\|_{\text{sp}} = b_k \|W_k\|_{\text{sp}}$, so the ratio is unchanged.

(b) By Corollary A.2 all blocks are active, and as in the proof of Proposition 3.4, $\partial\|\theta\| = \{(\mu_k c_k^{-1} G_k)_k : \mu \in \Delta_K, G_k \in \partial\|W_k\|_{\text{sp}}\}$; multiplying by $\|\theta\|$ gives $\partial h(\theta)$. On the relative interior ($\mu_k > 0$) the parametrization is injective, since $\|G_k\|_{\text{nuc}} = \text{tr} M = 1$ recovers $\mu_k = c_k \|\mathbf{u}_{W_k}\|_{\text{nuc}}/\|\theta\|$ from an element $\mathbf{u}$ and then G_k . Each $\partial\|W_k\|_{\text{sp}}$ is a spectrahedron of dimension $\rho_k(\rho_k + 1)/2 - 1$ (Lemma B.4), which gives the stated dimension. For a solution, $\mathbf{r}_{W_k} = \|\theta\| \mu_k c_k^{-1} G_k$, and since $\langle G_k, W_k \rangle = \sigma_1(W_k) = c_k \|\theta\|$, Lemma 3.3 gives $\|\theta\|^2 \mu_k = \langle \mathbf{r}_{W_k}, W_k \rangle = d_k \|\theta\|^2/L$, i.e. $\mu_k = d_k/L$. Hence $\mathbf{r}_{W_k} = \|\theta\| d_k G_k/(L c_k) = d_k \sigma_1(W_k) G_k/(L c_k^2)$, using $\|\theta\| = \sigma_1(W_k)/c_k$. If $\rho_k = 1$, $G_k = \mathbf{p}_k \mathbf{q}_k^\top$ is unique, so $\mathbf{k}_{W_k}^\circ = (d_k/(L c_k^2)) \mathbf{k}_{W_k}^*$, which is a common positive multiple of $\mathbf{k}^*$ iff d_k/c_k^2 is constant. If $\rho_k > 1$, G_k ranges over a spectrahedron of positive dimension, so stationarity does not select a unique point. The barycentric choice $G_k = U_{1,k} V_{1,k}^\top/\rho_k$ is basis-independent, but selecting it adds a condition not implied by stationarity.

(c) $\|\mathbf{k} - t\mathbf{s}\|^2 = \|\mathbf{k}\|^2 - 2t\langle \mathbf{k}, \mathbf{s} \rangle + t^2 \|\mathbf{s}\|^2$ is a convex quadratic with unconstrained minimizer $\langle \mathbf{k}, \mathbf{s} \rangle/\|\mathbf{s}\|^2$; projecting onto $t \geq 0$ gives $t^* = (\langle \mathbf{k}, \mathbf{s} \rangle)_+/\|\mathbf{s}\|^2$. If $\langle \mathbf{k}, \mathbf{s} \rangle \leq 0$ then $t^* = 0$. Otherwise the minimum is $\|\mathbf{k}\|^2(1 - \cos^2 \angle(\mathbf{k}, \mathbf{s}))$, which vanishes iff $\mathbf{s}$ is a positive multiple of $\mathbf{k}$. Each term of (10) is invariant under $\mathbf{r} \mapsto t\mathbf{r}$, $t > 0$, by (a). $\square$

B.6 Per-block versus global tightness for Muon

Lemma B.3 (Block decomposition of tightness). *With $\|\theta\| = \max_k \|W_k\|_{\text{sp}}/c_k$ and $\|\cdot\|_* = \sum_k c_k \|\cdot\|_{\text{nuc}}$, global tightness $\langle \mathbf{r}, \theta \rangle = \|\mathbf{r}\|_* \|\theta\|$ holds if and only if (a) $\langle \mathbf{r}_{W_k}, W_k \rangle = \|\mathbf{r}_{W_k}\|_{\text{nuc}} \|W_k\|_{\text{sp}}$ for every k with $\mathbf{r}_{W_k} \neq \mathbf{0}$, and (b) $\mathbf{r}_{W_k} = \mathbf{0}$ unless $k \in \arg \max_j \|W_j\|_{\text{sp}}/c_j$.*

Proof. $\langle \mathbf{r}, \theta \rangle = \sum_k \langle \mathbf{r}_{W_k}, W_k \rangle \leq \sum_k \|\mathbf{r}_{W_k}\|_{\text{nuc}} \|W_k\|_{\text{sp}} = \sum_k (c_k \|\mathbf{r}_{W_k}\|_{\text{nuc}}) (\|W_k\|_{\text{sp}}/c_k) \leq (\sum_k c_k \|\mathbf{r}_{W_k}\|_{\text{nuc}}) \|\theta\|$. Equality in the first inequality is (a); equality in the second requires $c_k \|\mathbf{r}_{W_k}\|_{\text{nuc}} (\|\theta\| - \|W_k\|_{\text{sp}}/c_k) = 0$ for all k , which is (b). $\square$

Condition (a) is what (10) enforces. Condition (b) would be a severe additional requirement if some block failed to attain the maximum, but by Corollary A.2 none does at a KKT direction of a blockwise-homogeneous network, so (b) is vacuous there and (10) is exact (Proposition 3.4).

B.7 Specificity and collapse

Lemma B.4 (Exact KKT targets). *For $\boldsymbol{\theta} \neq \mathbf{0}$: (i) ℓ_2 : $\partial(\frac{1}{2}\|\boldsymbol{\theta}\|_2^2) = \{\boldsymbol{\theta}\}$; (ii) ℓ_∞ : $\partial(\frac{1}{2}\|\boldsymbol{\theta}\|_\infty^2) = \|\boldsymbol{\theta}\|_\infty \text{conv}\{\text{sign}(\theta_j)\mathbf{e}_j : j \in A\}$ with $A = \arg \max_l |\theta_l|$, a simplex of dimension $|A| - 1$, whose cone is $\{\mathbf{r} : \mathbf{r}_j = 0 \text{ for } j \notin A, \mathbf{r}_j \theta_j \geq 0 \text{ for } j \in A\}$; (iii) spectral, for $W \neq 0$ whose top singular value has multiplicity ρ : $\partial(\frac{1}{2}\|W\|_{\text{sp}}^2) = \sigma_1\{U_1 M V_1^\top : M \succeq 0, \text{tr } M = 1\}$, a spectrahedron of dimension $\rho(\rho + 1)/2 - 1$, which for $\rho = 1$ is the singleton $\{\sigma_1 \mathbf{p} \mathbf{q}^\top\}$.*

Proof. By Lemma B.1, $\partial(\frac{1}{2}\|\cdot\|^2) = \|\cdot\| \partial\|\cdot\|$ in each case. (i) $\|\cdot\|_2$ is differentiable at $\boldsymbol{\theta} \neq 0$ with $\partial\|\boldsymbol{\theta}\|_2 = \{\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2\}$. (ii) $\|\boldsymbol{\theta}\|_\infty = \max_{j, \epsilon \in \{\pm 1\}} \epsilon \theta_j$ is a maximum of linear functions, so Danskin's theorem gives $\partial\|\boldsymbol{\theta}\|_\infty = \text{conv}\{\text{sign}(\theta_j)\mathbf{e}_j : j \in A\}$; nonnegative combinations of these vectors are exactly the vectors supported on A whose nonzero coordinates have matching signs, equivalently $\mathbf{r}_j \theta_j \geq 0$ on A . (iii) is the classical spectral-norm subdifferential (Watson, 1992); the map $M \mapsto U_1 M V_1^\top$ is injective, so the dimension is that of $\{M \succeq 0, \text{tr } M = 1\}$. For $\rho = 1$, Danskin's theorem applied to $\|W\|_{\text{sp}} = \max_{\|\mathbf{u}\|=\|\mathbf{v}\|=1} \mathbf{u}^\top W \mathbf{v}$ gives $\partial\|W\|_{\text{sp}} = \{\mathbf{p} \mathbf{q}^\top\}$. $\square$

Proof of Lemma A.4. Since $\rho_k \leq r_k$, $m_k n_k - \rho_k(\rho_k + 1)/2 \geq m_k n_k - r_k(r_k + 1)/2$, and $m_k n_k - r_k(r_k + 1)/2 - \frac{1}{2}(m_k n_k - r_k) = \frac{1}{2}(m_k n_k - r_k^2) \geq 0$. Sum over k . $\square$

Proof of Lemma A.5. For $f(\mathbf{x}; \boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \phi(\mathbf{x}) \rangle$, $\nabla_{\boldsymbol{\theta}} f(\mathbf{x}; \boldsymbol{\theta}) = \phi(\mathbf{x})$, so $\mathbf{r}(z) = \sum_i \lambda_i y_i \phi(\mathbf{x}_i)$ does not depend on $\boldsymbol{\theta}$. The stationarity-feasible set $\{z : \mathbf{r}(z) \in C^\circ(\boldsymbol{\theta})\}$ therefore depends on $\boldsymbol{\theta}$ only through $C(\boldsymbol{\theta})$, and the global alignment loss vanishes iff $\mathbf{r}(z) \in C^\circ(\boldsymbol{\theta})$ by (1). Hence $C(\boldsymbol{\theta}) = C(\boldsymbol{\theta}')$ implies that the feasible sets and zero sets coincide. Thus the dependence of this stationarity-only formulation on the checkpoint factors through $C(\boldsymbol{\theta})$, whose image has dimension S on the stratum by Definition 4.1. No input non-identifiability conclusion follows from this factorization alone. $\square$

Proof of Lemma 4.2. ℓ_2 . By Lemma B.4(i), $C(\boldsymbol{\theta}) = \{t\boldsymbol{\theta} : t \geq 0\}$, which is determined exactly by the direction $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2 \in S^{p-1}$, and distinct directions give distinct cones; hence $S = p - 1$.

Spectral. With all blocks active, $C(\boldsymbol{\theta}) = \prod_k C_k$ with $C_k = \{U_{1,k} M V_{1,k}^\top : M \succeq 0\}$ (proof of Proposition 3.4 and Lemma B.4(iii)); $\sigma_1(W_k)$ does not enter. C_k determines the partial isometry $P_k := U_{1,k} V_{1,k}^\top$: the elements of C_k have singular values equal to the eigenvalues of M , so up to positive scale P_k is the unique element of rank ρ_k with all nonzero singular values equal ($M = cI$). Conversely, P_k determines C_k , because any two factorizations of P_k differ by $(U_1, V_1) \mapsto (U_1 Q, V_1 Q)$ with Q orthogonal, which leaves $\{U_1 M V_1^\top : M \succeq 0\}$ unchanged. The image of $\boldsymbol{\theta} \mapsto C(\boldsymbol{\theta})$ on the stratum is therefore the image of $\boldsymbol{\theta} \mapsto (P_k)_k$. This image is open in the product of the manifolds $\mathcal{P}_k$ of rank- ρ_k partial isometries: given P'_k near P_k , write $P'_k = Q_k P_k Q'_k{}^\top$ with orthogonal Q_k, Q'_k near the identity, and replace W_k by $Q_k W_k Q'_k{}^\top$, which keeps all singular values, hence the stratum and activity of every block. Finally, $\dim \mathcal{P}_k = S_k(\rho_k)$: the polar decomposition $B = P(B^\top B)^{1/2}$ identifies the manifold of rank- ρ matrices, of dimension $\rho(m + n - \rho)$, with pairs of a rank- ρ partial isometry and a positive definite operator on its ρ -dimensional row space, of dimension $\rho(\rho + 1)/2$. For $\rho = 1$, $S_k = m_k + n_k - 2$.

ℓ_∞ . By Lemma B.4(ii), $C(\boldsymbol{\theta})$ is a function of $(A, \text{sign } \boldsymbol{\theta}|_A)$ alone, which is discrete data; no continuous functional of $\boldsymbol{\theta}$ is determined, so $S = 0$, for every $|A| = s$. $\square$

Proof of the multiplicity bound in Appendix A.9. $S_k(\rho) = \rho(m + n) - \frac{3}{2}\rho^2 - \frac{1}{2}\rho$ is concave, so on $[1, r]$ it is at least $\min\{S_k(1), S_k(r)\}$. With $r = n \leq m$, $S_k(r) = nm - n(n + 1)/2 \geq n(n - 1)/2$, and $S_k(r) - (m + n - 2) = m(n - 1) - n(n + 3)/2 + 2 \geq n(n - 1) - n(n + 3)/2 + 2 = (n - 1)(n - 4)/2 \geq 0$ for $n \geq 4$. $\square$

Proof of Proposition A.6. By Lemma B.4(ii), $C(\boldsymbol{\theta})$ depends only on $(A, \text{sign } \boldsymbol{\theta}|_A)$. The conclusion about stationarity-feasible sets follows from Lemma A.5. The final qualification is necessary: even a cone that is constant on a stratum can impose equations on $\mathbf{r}(z)$ that identify z for a particular feature map. $\square$

Remark B.5 (Nondegeneracy for linear-in-parameter models). For $f = \langle \boldsymbol{\theta}, \phi(\mathbf{x}) \rangle$, the ℓ_∞ max-margin problem is the linear program $\min t$ s.t. $t - \theta_j \geq 0$, $t + \theta_j \geq 0$, $y_i \langle \boldsymbol{\theta}, \phi(\mathbf{x}_i) \rangle \geq 1$, in the variables $(\boldsymbol{\theta}, t) \in \mathbb{R}^{p+1}$. If its solution is unique it is a vertex, so the normals of the active constraints, namely $(-\text{sign}(\theta_j)\mathbf{e}_j, 1)$ for $j \in A$ and $(y_i \phi(\mathbf{x}_i), 0)$ for the active margins, span $\mathbb{R}^{p+1}$, and their $\boldsymbol{\theta}$ -components span $\mathbb{R}^p$. Under strict complementarity the active margins are exactly I , which is the span condition of Assumption A.9.

Proof of Theorem A.10. (i) The spanning set of Assumption A.9 has at most $s + n_I$ elements, so $p \leq s + n_I$. By Lemma B.4(ii), $C(\theta)$ is a full-dimensional cone in $\text{span}\{\mathbf{e}_j : j \in A\}$, so $\text{codim } C(\theta) = p - s$.

(ii) $\text{aff } C(\theta) = \text{span}\{\mathbf{e}_j : j \in A\}$, so P_N is the coordinate projection onto the $p - s$ coordinates outside A , and Φ takes values in $\mathbb{R}^{p-s}$. Hence $\text{rank } D\Phi(z_I^*) \leq p - s \leq n_I$. If $n_I d \geq 2$ then $n_I < n_I(d + 1) - 1 = \dim z_I - \dim \mathcal{G}$, so the rank condition of Definition C.1 fails.

(iii) By the constant-rank theorem, $\Phi^{-1}(\mathbf{0})$ is, near z_I^* , a C^1 manifold $\mathcal{M}$ of dimension $n_I(d + 1) - \text{rank } D\Phi \geq n_I d$, and it contains z_I^* because $\mathbf{r}(z^*) \in C(\theta)$. On $\mathcal{M}$ the coordinates of $\mathbf{r}$ outside A vanish. The coordinates $\mathbf{r}_j$, $j \in A$, are continuous and nonzero with sign $\text{sign } \theta_j$ at z^* , hence keep that sign nearby, and $\lambda_i > 0$ persists nearby; so $\mathbf{r}(z) \in C^\circ(\theta)$ on a neighbourhood of z_I^* in $\mathcal{M}$. For the projection π onto the input coordinates, $\ker(D\pi|_{T\mathcal{M}})$ lies in the n_I multiplier coordinates, so $\text{rank } D(\pi|_{\mathcal{M}}) \geq n_I d - n_I$ everywhere. Its maximal value is attained on an open set accumulating at z_I^* , and there the constant-rank theorem makes the image a C^1 manifold of dimension at least $n_I(d - 1)$. $\square$

Remark B.6 (Codimensions in (14)). The cone C has dimension 1 for ℓ_2 ; $\sum_k \rho_k(\rho_k + 1)/2$ for the spectral norm (a product of per-block spectrahedral cones, all blocks active), which is K when all $\rho_k = 1$; and s for ℓ_∞ . The codimensions in (14) follow.

C LOCAL IDENTIFIABILITY

Specificity measures the continuous variability of the optimizer-dependent target cone (Lemma A.5). The complementary question, whether the true data is regularly isolated among stationarity-consistent candidates, is one of well-posedness in the attacker's own variables. Only candidates with positive multipliers enter the local residual analysis, so we restrict to them.

Definition C.1 (Regular first-order identifiability). *Let z^* be the true solution, $I = \{i : \lambda_i^* > 0\}$, $n_I = |I|$, and $z_I = (\{\mathbf{x}_i\}, \{\lambda_i\})_{i \in I} \in \mathbb{R}^{n_I(d+1)}$, the remaining candidates being held at z^* . Let $P_{N(\theta)}$ be the orthogonal projection onto the orthogonal complement of $\text{aff } C(\theta)$ and $J = \partial_{z_I} [P_{N(\theta)} \mathbf{r}(z)]$ at z^* . Call z^* regularly first-order identifiable if $\text{rank } J = \dim z_I - \dim \mathcal{G} = n_I(d + 1) - 1$, where $\mathcal{G}$ is the one-dimensional multiplier gauge $\lambda \mapsto c\lambda$ of the cone formulation (6). Discrete symmetries (permutations of candidates) do not affect the rank.*

The multiplier gauge always lies in $\ker J$: along $(\mathbf{0}, \lambda_I^*)$ the derivative of $P_{N(\theta)} \mathbf{r}$ is $P_{N(\theta)} \mathbf{r}(z^*) = \mathbf{0}$. Maximal rank rules out every infinitesimal ambiguity except this gauge and is a sufficient regularity condition for local isolation modulo gauge. Its failure does not, by itself, prove set-theoretic non-identifiability, because a rank-deficient nonlinear equation can still have an isolated zero.

Lemma C.2 (A necessary condition). $\text{rank } J \leq \text{codim } C(\theta)$. Hence regular first-order identifiability requires $\text{codim } C(\theta) \geq n_I(d + 1) - 1$.

Proof. $C(\theta)$ is a cone containing $\mathbf{0}$, so $\text{aff } C(\theta) = \text{span } C(\theta)$ has dimension $\dim C(\theta)$, and $P_{N(\theta)} \mathbf{r}$ takes values in a space of dimension $p - \dim C(\theta) = \text{codim } C(\theta)$. $\square$

By (14) and Lemma A.4, the necessary condition holds for ℓ_2 and for the spectral norm whenever $p - 1$, respectively $\frac{1}{2}(p - \sum_k r_k)$, is at least $n_I(d + 1) - 1$, which is the over-parameterized regime. For ℓ_∞ it fails at nondegenerate KKT points:

Proposition C.3 (Failure of regular first-order identifiability). *Under Assumption A.9 and $n_I d \geq 2$, z^* is not regularly first-order identifiable for ℓ_∞ , at any width. Under the additional hypotheses of Theorem A.10(iii), it belongs to a positive-dimensional local family of stationarity-consistent candidates.*

Proof. Theorem A.10(ii)–(iii). $\square$

Conjecture C.4. For ℓ_2 , and for the spectral norm, z^* is regularly first-order identifiable for sufficiently over-parameterized blockwise-homogeneous f .

We prove Adam's failure of regular first-order identifiability (Proposition C.3) and, under the additional hypotheses of Theorem A.10(iii), actual positive-dimensional local ambiguity. For the positive ℓ_2 and spectral direction we prove only the necessary condition of Lemma C.2. The conjectured positive half says that high specificity

and high codimension are not merely necessary but sufficient for regular first-order well-posedness. It is directly measurable, requiring one singular value decomposition of J at a recovered solution.

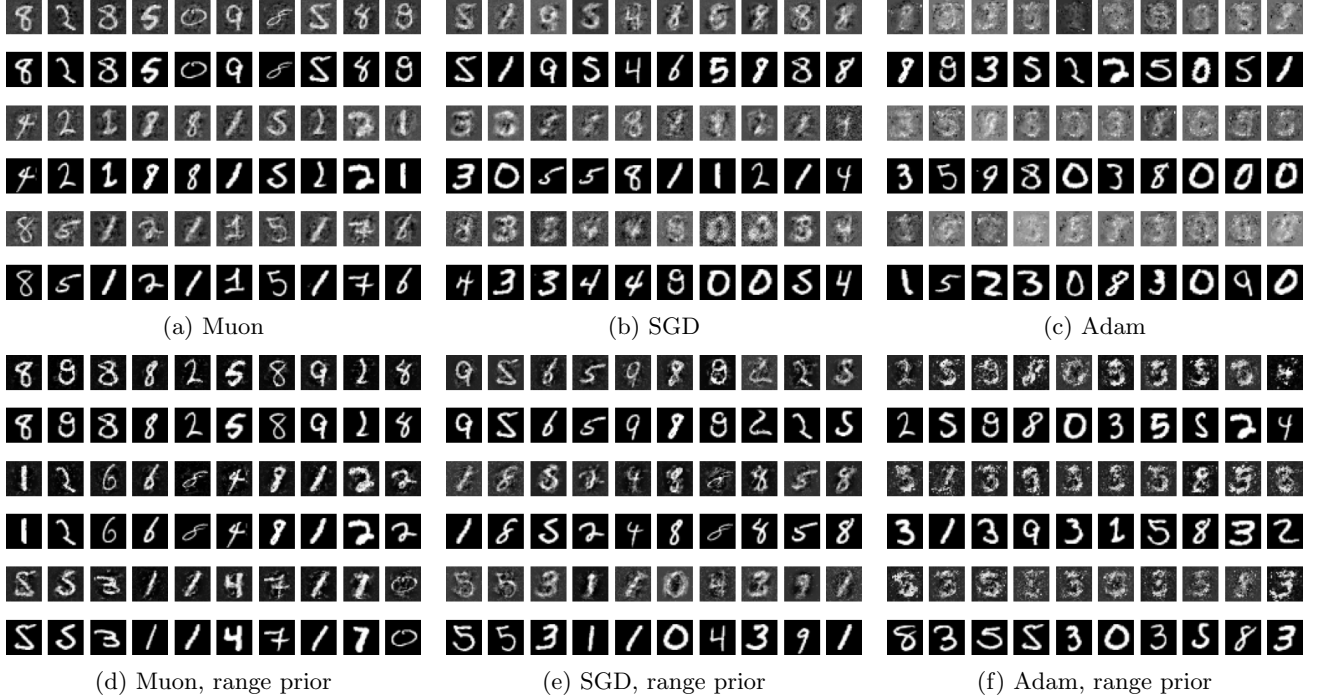


Figure 3: Reconstructions on binary MNIST ($n = 500$) without a prior (top) and with the range prior (bottom). In each panel, odd rows show reconstructions and the row below shows the matched training images. Muon recovers sharp digit structure even without a prior.

D EXPERIMENTAL DETAILS

Multiclass residual. For multiclass models the residual uses one multiplier per sample and competing class, $\sum_i \sum_{j \neq y_i} \lambda_{ij} \nabla_{\theta} [\Phi_{y_i}(\mathbf{x}_i; \theta) - \Phi_j(\mathbf{x}_i; \theta)]$ (Buzaglo et al., 2023).

Test data. For 10-class CIFAR-10 we evaluate on a class-balanced test subset of 1,000 images (100 per class).

Initialization. Following Haim et al. (2022), the first layer is initialized with variance 10^{-4} and the remaining layers with Kaiming normal initialization (He et al., 2015).

Trained models & hardware. Table 4 summarizes the trained model configurations used in Section 5. All networks were trained using PyTorch on a single NVIDIA GeForce GTX 1080 Ti GPU. Training required fewer than 2 minutes for 10,000 iterations over $N = 500$ training samples with learning rate 10^{-2} . For the binary classification setup, 250 samples per class were used; for the 10-class multiclass setup, 50 samples per class were used. Code and reproduction scripts will be made publicly available.

Muon geometry. In our setup, Muon is applied exclusively to internal 2D weight matrices (e.g., hidden-to-hidden layer W_2), whereas standard SGD (Battash et al., 2024) is used for both the input layer (W_1) and the final output layer (W_L). Consequently, the overall implicit bias is spectral norm, with nuclear dual norm on the internal blocks and Euclidean (L_2) on the first and last layers.

Data reconstruction attack implementation. Data reconstruction was executed by solving the KKT stationarity conditions corresponding to SGD, Adam, and Muon. Specifically, for layers optimized using SGD, the stationarity loss is defined by the stationarity term in Equation 3; for layers utilizing the Adam optimizer, the stationarity loss follows Equation 11; and for Muon optimizer layers, it is given by Equation 10.

Table 4: Trained models ($n = 500$). All models interpolate the training set. Because the network is homogeneous, an unnormalized margin depends on the parameter scale and is not comparable across optimizers.

Task	Optimizer	Test acc. (%)	Test loss	Min. margin
MNIST (binary)	SGD	90.0	0.790	9.21
	Adam	91.0	0.753	16.64
	Muon	92.0	0.853	16.65
CIFAR-10 (binary)	SGD	77.0	2.25	9.56
	Adam	79.0	3.41	16.64
	Muon	79.0	2.57	16.63
CIFAR-10 (10-class)	SGD	28.0	7.97	13.75
	Adam	26.0	13.5	14.11
	Muon	29.0	7.09	15.46

To facilitate gradient-based optimization during reconstruction without altering forward behavior, we adopted a modified backward pass for the ReLU activation function $y = \max(0, x)$ following Haim et al. (2022). Specifically, while the forward pass remains exact ReLU, the backward gradient is modified as:

$$\frac{\partial \widetilde{L}}{\partial x} = \frac{\partial L}{\partial y} \sigma(\alpha x), \quad \text{where } \sigma(z) = \frac{1}{1 + e^{-z}} \quad (15)$$

Here, α acts as a hyperparameter interpolating between a scaled linear backward pass function ($\alpha \rightarrow 0$) and the standard ReLU derivative ($\alpha \rightarrow \infty$).

Dual variables (Lagrange multipliers associated with stationarity conditions) and reconstructed input samples x_i were optimized. For the attack optimization (applied across all three attacks), we followed Haim et al. (2022) using SGD with a momentum of 0.9 over 100,000 iterations.

Hyperparameter search. We selected reconstruction hyperparameters through a random search maximizing SSIM@30:

- **Reconstruction learning rate:** Log-uniform sampling in $[10^{-4}, 10^{-2}]$.
- **Initialization scale (σ_x):** Log-uniform sampling in $[10^{-3}, 1]$ for the initial scale of x_i .
- **Modified ReLU parameter (α):** Uniform sampling in $[10, 500]$.

Post-processing & matching. To evaluate fidelity, reconstructed samples were matched one-to-one to ground-truth training samples using the Hungarian algorithm on centered, normalized ℓ_2 distance. The best-matched instances (such as the top 30 pairs) were selected for qualitative and quantitative reporting.

E ADDITIONAL EXPERIMENTAL RESULTS

To complement the qualitative visual comparisons presented in the main text, Tables 5–7 report the complete quantitative reconstruction results across all experimental configurations. We evaluate performance across both binary and multiclass tasks on MNIST and CIFAR-10 under multiple prior regimes: without prior constraints, under range prior constraints $[0, 1]$, and with additional regularized variants including Total Variation (TV) and L_1 sparsity priors. Reconstruction fidelity is evaluated using standard pixel-level and structural metrics, specifically Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), L_1 distance, and L_2 distance.

Across all datasets, and prior configurations, Muon consistently achieves the superior value in every single metric entry, outperforming both standard SGD and Adam. Remarkably, even in the completely unconstrained

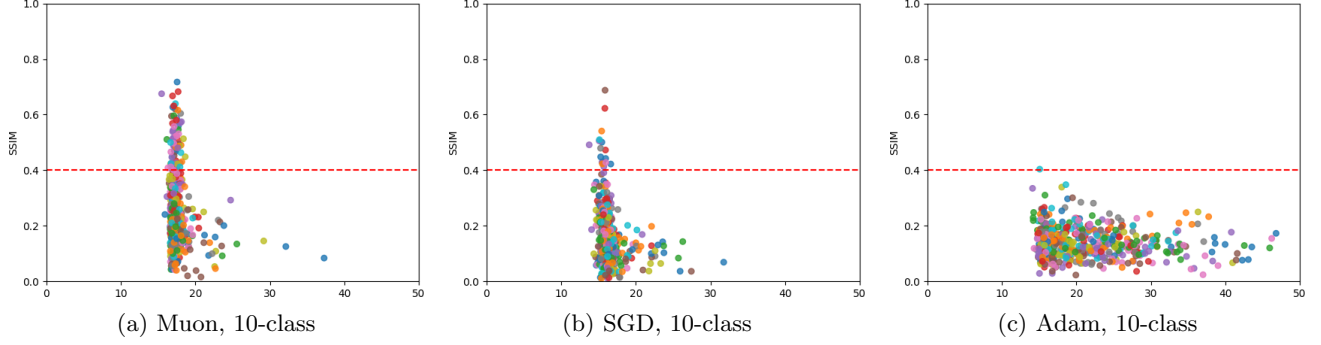


Figure 4: Per-sample SSIM versus training margin γ_i on CIFAR-10 ($n = 500$, range prior), for multiclass models. Each point is a training sample, colored by class; the dashed line marks $\text{SSIM} = 0.4$. Recognizable reconstructions occur only near the minimal margin. Muon produces the most of them and Adam essentially none.

setting (without prior), Muon retains significantly higher structural fidelity, highlighting the high specificity of reconstructions driven by its implicit bias.

Multiclass training dynamics. While the overall performance ordering (Muon > SGD > Adam) is largely maintained in the 10-class setting, the quantitative gap between Muon and SGD narrows compared to the binary regime. Specifically, Muon exhibits a higher concentration of samples exceeding the 0.4 recognizability threshold compared to SGD, as illustrated in Figure 4.

Comparing across sample sizes (Table 2). Evaluating reconstruction fidelity across varying dataset sizes n reveals that $\text{SSIM}@30$ peaks at $n = 100$ and decreases substantially for $n \geq 500$. Although selecting the top 30 instances from a larger candidate pool ($m = 2n$) represents a more selective subset, the observed drop in performance indicates that the underlying inverse problem becomes fundamentally harder to invert accurately as the scale n increases.

Computational complexity. Let q be the number of reconstruction candidates, p the number of model parameters, and d the input dimension. Each attack iteration performs work equivalent to q differentiable parameter-gradient evaluations plus $O(p)$ residual aggregation. The SGD and Adam alignment terms require $O(p)$ additional work, while Muon’s dual-norm term additionally requires blockwise singular-value computations. Excluding model activations and automatic-differentiation workspace, the explicit attack state requires $O(qd + p)$ memory. We study the empirical dependence on training-set size in Section 5 and Table 2.

F FURTHER MATERIAL

F.1 Notation details (from §2.1)

For nonzero $\mathbf{u}$ and $\boldsymbol{\theta}$, alignment means that $\mathbf{u}$ is a positive multiple of $\boldsymbol{\theta}$ in the ℓ_2 geometry. For a single spectral-norm block, it means that $\mathbf{u} = U_1 M V_1^\top$ for some positive semidefinite M , where U_1 and V_1 span the corresponding top singular subspaces of $\boldsymbol{\theta}$. For the max-spectral norm, aligned blocks can be nonzero only at blocks attaining the maximum spectral norm. For ℓ_∞ , $\mathbf{u}$ is supported on the largest-magnitude coordinates of $\boldsymbol{\theta}$, with matching signs wherever $\mathbf{u}$ is nonzero.

On strata: For ℓ_∞ , this type is determined by the set A of coordinates attaining the maximum absolute value and their signs. For the spectral norm, it is determined by the multiplicity ρ of the top singular value; for the max-spectral norm, it also depends on which blocks attain the maximum. By “on the generic stratum,” we mean away from ties and other degeneracies: top singular values are simple, the ℓ_∞ maximum is unique, and the relevant active sets are locally constant. Unless stated otherwise, dimension counts are taken on this stratum.

On Assumption 2.1: A first-layer bias is absorbed by appending a constant 1 to the input, so that $[W_1 \mathbf{b}_1]$ is a single block (and Muon is understood to act on it). Differentiability holds at almost every $\boldsymbol{\theta}$ for ReLU networks, as used by Haim et al. (2022), and everywhere for the smooth homogeneous models covered by the guarantees



Figure 5: Reconstructions on binary CIFAR-10 ($n = 500$), in the layout of Figure 3. Muon reconstructions best preserve object shape and color.

of Gronich and Vardi (2026).

On the generalized Cauchy–Schwarz inequality: The three geometries considered here give familiar instances of this inequality: Cauchy–Schwarz for the (ℓ_2, ℓ_2) pair, Hölder for (ℓ_∞, ℓ_1) , and a bound obtained from von Neumann’s trace inequality for the max-spectral norm and its dual, the sum-nuclear norm $\sum_k \|\mathbf{u}_{W_k}\|_{\text{nuc}}$.

F.2 Other privacy attacks (from §2)

Other privacy attacks differ in their goals and requirements. Membership inference determines whether an example belongs to the training set (Shokri et al., 2017), model inversion can recover class-representative inputs (Fredrikson et al., 2015), and gradient leakage reconstructs data from shared gradients (Zhu et al., 2019). KKT-based reconstruction attacks use only the final weights of the trained model. Differentially private training (Abadi et al., 2016) provides a formal guarantee limiting the influence of individual training examples, potentially at a utility cost. Optimizer choice provides another way to affect reconstruction leakage, but does not provide the same guarantee.

F.3 Implicit bias of gradient flow (from §2.1)

Theorem F.1 (paraphrased version of Lyu and Li (2020), Ji and Telgarsky (2020)). *Let $\Phi(\boldsymbol{\theta}; \mathbf{x})$ be a normalized homogeneous¹ ReLU neural network. Consider minimizing the logistic ($z \mapsto \log(1+e^{-z})$) or the exponential ($z \mapsto e^{-z}$) loss using gradient flow (which is a continuous-time analog of gradient descent) over a binary classification set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \{-1, 1\}$. Assume that there is a time t_0 where $L(\boldsymbol{\theta}(t_0)) < \frac{1}{n}$. Then, gradient flow converges in direction² to a first order stationary point (KKT point (Cortes and Vapnik, 1995)) of the maximum-margin problem (P).*

A KKT point of (P) is characterized by the following set of conditions:

¹A network $\Phi(\boldsymbol{\theta}; \mathbf{x})$ is called *homogeneous* if there exists $c > 0$ such that for every $b > 0$, $\boldsymbol{\theta}$ and $\mathbf{x}$, it holds that $\Phi(b \cdot \boldsymbol{\theta}; \mathbf{x}) = b^c \Phi(\boldsymbol{\theta}; \mathbf{x})$ Haim et al. (2022).

²We say that gradient flow *converges in direction* to $\hat{\boldsymbol{\theta}}$ if $\lim_{t \rightarrow \infty} \frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|} = \frac{\hat{\boldsymbol{\theta}}}{\|\hat{\boldsymbol{\theta}}\|}$.

Table 5: Reconstruction on binary MNIST ($n = 500$). Best per prior in **bold**.

Prior	Optimizer	SSIM $\uparrow$	SSIM@30 $\uparrow$	PSNR $\uparrow$	PSNR@30 $\uparrow$	$\ell_1 \downarrow$	ℓ_1 @30 $\downarrow$	$\ell_2 \downarrow$	ℓ_2 @30 $\downarrow$
None	SGD	0.1661	0.3463	10.857	14.867	0.2059	0.1260	0.2865	0.1806
	Adam	0.1527	0.2487	10.345	10.903	0.2077	0.1926	0.3039	0.2850
	Muon	0.2350	0.4310	11.495	16.997	0.1718	0.0935	0.2662	0.1413
Range	SGD	0.2007	0.4057	11.186	15.464	0.1700	0.1008	0.2759	0.1686
	Adam	0.1898	0.2994	10.640	11.933	0.1958	0.1415	0.2938	0.2531
	Muon	0.3242	0.6371	12.115	20.154	0.1339	0.0517	0.2479	0.0982
TV	SGD	0.1811	0.3768	11.042	14.948	0.1964	0.1209	0.2805	0.1789
	Adam	0.1739	0.2882	10.593	11.270	0.1986	0.1801	0.2954	0.2732
	Muon	0.3288	0.5808	11.781	17.699	0.1550	0.0783	0.2576	0.1303
ℓ_1	SGD	0.1688	0.3614	10.894	14.845	0.2007	0.1227	0.2853	0.1810
	Adam	0.1445	0.1752	10.340	11.985	0.1829	0.1433	0.3041	0.2516
	Muon	0.2849	0.5302	11.610	17.291	0.1563	0.0819	0.2627	0.1366
TV+ ℓ_1	SGD	0.1848	0.3977	11.086	15.423	0.1908	0.1120	0.2791	0.1694
	Adam	0.1606	0.2311	10.538	12.028	0.1825	0.1487	0.2972	0.2504
	Muon	0.3299	0.6211	11.831	18.002	0.1489	0.0728	0.2561	0.1259
All	SGD	0.2289	0.4511	11.345	15.657	0.1641	0.0961	0.2709	0.1649
	Adam	0.1758	0.2655	10.834	13.006	0.1663	0.1215	0.2873	0.2237
	Muon	0.3400	0.6599	11.996	20.126	0.1361	0.0520	0.2513	0.0986

- **Stationarity:** $\theta = \sum_{i=1}^n \lambda_i \nabla_{\theta} [y_i \Phi(\theta; \mathbf{x}_i)]$
- **Primal feasibility:** $y_i \Phi(\theta; \mathbf{x}_i) \geq 1, \quad \forall i \in [n]$
- **Dual feasibility:** $\lambda_i \geq 0, \quad \forall i \in [n]$.
- **Comp. slackness:** $\lambda_i = 0$ if $y_i \Phi(\theta; \mathbf{x}_i) \neq 1, \quad \forall i \in [n]$.

F.4 Implicit-bias background (from §2)

The reconstruction attack relies on characterizing which interpolating solution an optimizer selects. For separable linear models, GD converges in direction to the ℓ_2 maximum-margin solution (Soudry et al., 2018). Related results establish convergence towards KKT directions of margin problems for homogeneous networks (Lyu and Li, 2020; Ji and Telgarsky, 2020), and connect steepest descent under other norms to the corresponding margin geometry (Gunasekar et al., 2018); see Vardi (2023) for a survey. These results allow us to study optimizer choice through its associated margin geometry.

Table 6: Reconstruction on binary CIFAR-10 ($n = 500$). Best per prior in **bold**.

Prior	Optimizer	SSIM $\uparrow$	SSIM@30 $\uparrow$	PSNR $\uparrow$	PSNR@30 $\uparrow$	$\ell_1 \downarrow$	ℓ_1 @30 $\downarrow$	$\ell_2 \downarrow$	ℓ_2 @30 $\downarrow$
None	SGD	0.0986	0.2287	11.671	13.523	0.2027	0.1623	0.2609	0.2108
	Adam	0.1365	0.1379	12.422	12.902	0.1849	0.1743	0.2393	0.2264
	Muon	0.1771	0.5508	12.567	17.802	0.1807	0.0974	0.2353	0.1288
Range	SGD	0.0994	0.2336	11.665	13.435	0.2027	0.1634	0.2611	0.2129
	Adam	0.1299	0.1367	12.334	13.441	0.1863	0.1624	0.2417	0.2128
	Muon	0.1921	0.6130	12.808	19.190	0.1750	0.0825	0.2289	0.1098
TV	SGD	0.1106	0.2574	11.857	14.099	0.1975	0.1517	0.2554	0.1973
	Adam	0.0958	0.2115	11.2531	13.5265	0.2122	0.1789	0.2656	0.2001
	Muon	0.2310	0.6132	13.202	18.974	0.1672	0.0852	0.2187	0.1125
ℓ_1	SGD	0.0984	0.2375	11.668	13.995	0.2026	0.1528	0.2610	0.1996
	Adam	0.0977	0.2112	10.2658	13.2568	0.2471	0.1793	0.2787	0.2323
	Muon	0.1933	0.5598	12.798	18.109	0.1750	0.0936	0.2291	0.1243
TV+ ℓ_1	SGD	0.1113	0.2440	11.872	13.748	0.1973	0.1581	0.2549	0.2054
	Adam	0.1218	0.2369	11.2451	13.1178	0.2284	0.1955	0.2854	0.2323
	Muon	0.2305	0.6286	13.216	19.185	0.1667	0.0832	0.2184	0.1098
All	SGD	0.1122	0.2536	11.865	14.194	0.1975	0.1502	0.2551	0.1951
	Adam	0.1827	0.2377	12.727	14.462	0.1768	0.1407	0.2310	0.1892
	Muon	0.2196	0.6189	13.138	18.848	0.1683	0.0852	0.2203	0.1142

Table 7: Reconstruction on 10-class CIFAR-10 ($n = 500$). Best per prior in **bold**.

Prior	Optimizer	SSIM $\uparrow$	SSIM@30 $\uparrow$	PSNR $\uparrow$	PSNR@30 $\uparrow$	$\ell_1 \downarrow$	ℓ_1 @30 $\downarrow$	$\ell_2 \downarrow$	ℓ_2 @30 $\downarrow$
None	SGD	0.2031	0.3827	11.264	13.435	0.1912	0.1657	0.2535	0.1906
	Adam	0.1530	0.2066	12.254	12.740	0.1856	0.1758	0.2440	0.2307
	Muon	0.2446	0.5339	13.284	18.826	0.1651	0.0884	0.2167	0.1145
Range	SGD	0.2157	0.4515	12.999	16.445	0.1722	0.1158	0.2239	0.1506
	Adam	0.1423	0.1771	12.250	13.238	0.1874	0.1653	0.2441	0.2178
	Muon	0.2445	0.5315	13.278	18.987	0.1651	0.0863	0.2168	0.1124
TV	SGD	0.2112	0.4332	12.977	16.134	0.1729	0.1198	0.2245	0.1561
	Adam	0.2310	0.3442	13.520	15.774	0.1603	0.1224	0.2109	0.1627
	Muon	0.2761	0.5481	13.880	19.199	0.1537	0.0848	0.2023	0.1097
ℓ_1	SGD	0.2081	0.4141	12.952	16.344	0.1731	0.1147	0.2251	0.1523
	Adam	0.1382	0.2739	11.940	14.062	0.1913	0.1491	0.2529	0.1981
	Muon	0.2529	0.5371	13.436	18.686	0.1615	0.0900	0.2129	0.1163
TV+ ℓ_1	SGD	0.2110	0.4343	13.009	16.085	0.1721	0.1197	0.2236	0.1569
	Adam	0.1946	0.3050	12.742	14.106	0.1756	0.1499	0.2306	0.1971
	Muon	0.2806	0.5890	13.791	19.408	0.1547	0.0821	0.2044	0.1071
All	SGD	0.2110	0.4343	13.009	16.085	0.1721	0.1197	0.2236	0.1569
	Adam	0.1598	0.2603	12.491	13.821	0.1809	0.1571	0.2374	0.2037
	Muon	0.2814	0.5795	13.780	19.279	0.1549	0.0837	0.2046	0.1087

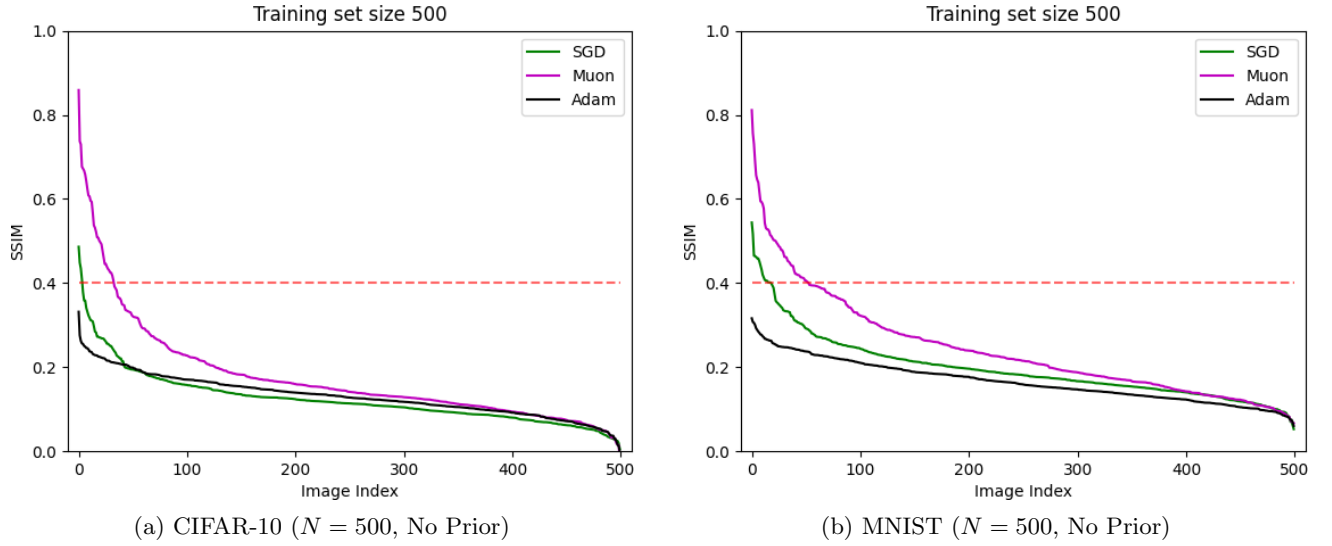


Figure 6: Sorted image-level SSIM values (from highest to lowest) for reconstructed samples on CIFAR-10 and MNIST without prior constraints ($n = 500$). The dashed horizontal reference line at $\text{SSIM} = 0.4$ indicates a threshold for acceptable visual reconstruction quality. Reconstructions under Muon consistently achieve higher SSIM scores, resulting in a substantially larger fraction of images reconstructed above 0.4 compared to SGD and Adam ($\text{Muon} > \text{SGD} > \text{Adam}$).