For the second term,

$$\begin{aligned}2u^T v &= 2(\nabla_w \mathcal{L}(f(x+\Delta x), y) - \nabla_w \mathcal{L}(f(x), y))^T \\ &\quad (\nabla_w \mathcal{L}(f(x), y) - g^*) \\ &\approx 2\Delta x^T \nabla_x \nabla_w \mathcal{L}(f(x), y)(\nabla_w \mathcal{L}(f(x), y) - g^*) \\ &\quad (\because \text{ Taylor's first-order approximation with } \mu < \delta) \\ &= -\mu(\frac{\partial \mathcal{L}_{grad}}{\partial x})^T(\frac{\partial \mathcal{L}_{grad}}{\partial x}) \\ &= -\mu||\frac{\partial \mathcal{L}_{grad}}{\partial x}||_2^2 \\ &\quad (\because \text{ Recall that how } \Delta x \text{ is computed in the first step}).\end{aligned}$$

For the third term, $||v||_2^2 = ||\nabla_w \mathcal{L}(f(x), y) - g^*||_2^2 = \mathcal{L}_{grad}(x)$.

Then, summing up the three terms considered above will lead to the following inequality:

$$\begin{aligned}\mathcal{L}_{grad}(x+\Delta x) &\leq \mathcal{L}_{grad}(x) + (L^2\mu^2 - \mu)||\frac{\partial \mathcal{L}_{grad}}{\partial x}||_2^2 \\ &= \mathcal{L}_{grad}(x) - \frac{1}{4L^2}||\frac{\partial \mathcal{L}_{grad}}{\partial x}||_2^2 \\ &\quad (\because \text{ min. at } \mu = \frac{1}{2L^2}).\end{aligned}$$

For both $\mu = \frac{1}{2L^2}$ and $\mu < \delta$ to be met, $\frac{1}{2L^2} < \delta$ should be satisfied and this is why the premise that $L > \epsilon = \frac{1}{\sqrt{2\delta}}$ is required for the theorem. We empirically found that $\tilde{L}$, the estimated value for $L$, mostly turned out to be not small, thus meeting the premise in practice. We will derive the theorem for small $L$ to explain outliers in future work. □

## A2. Proof of Theorem 2

**Theorem.** (The second gradient inversion loss theorem). Suppose $||g_w(x_1) - g_w(x_2)|| \geq M||x_1 - x_2||\forall x_1, x_2$ for $M > 0$ and $\mathcal{L}_{grad}(x) = ||g_w(x) - g^*||_2^2$ is the gradient matching loss. Then, when gradient descent $\Delta x$ is applied with step size $\mu < \delta_1$ for some $\delta_1 > 0$, the following holds:

$$\mathcal{L}_{grad}(x+\Delta x) \geq \mathcal{L}_{grad}(x) - \frac{1}{4M^2}||\frac{\partial \mathcal{L}_{grad}^x}{\partial x}||_2^2,$$

$\mu < \delta_1$ satisfies $||\mu\Delta x|| < \delta$ such that $g_w(x + \mu\Delta x) = g_w(x) + \mu\nabla_x g_w(x)\Delta x$ holds approximately.

*Proof.* Similar to the proof of previous theorem, we can use the following equation again:

$$\begin{aligned}\mathcal{L}_{grad}(x+\Delta x) &= ||\nabla_w \mathcal{L}(f(x+\Delta x), y) - g^*||_2^2 \\ &= ||\nabla_w \mathcal{L}(f(x+\Delta x), y) - \nabla_w \mathcal{L}(f(x), y) + \\ &\qquad \nabla_w \mathcal{L}(f(x), y) - g^*||_2^2 \\ &= ||u||_2^2 + 2u^T v + ||v||_2^2\end{aligned}$$

, where $u = \nabla_w \mathcal{L}(f(x+\Delta x), y) - \nabla_w \mathcal{L}(f(x), y)$ and $v = \nabla_w \mathcal{L}(f(x), y) - g^*$.

For the first term, $||u||_2^2 = ||\nabla_w \mathcal{L}(f(x+\Delta x), y) - \nabla_w \mathcal{L}(f(x), y)||_2^2 \geq M^2||\Delta x||^2 = M^2\mu^2||\frac{\partial \mathcal{L}_{grad}}{\partial x}||_2^2$ due to the given condition.

For the second term,

$$\begin{aligned}2u^T v &= 2(\nabla_w \mathcal{L}(f(x+\Delta x), y) - \nabla_w \mathcal{L}(f(x), y))^T \\ &\quad (\nabla_w \mathcal{L}(f(x), y) - g^*) \\ &\approx 2\Delta x^T \nabla_x \nabla_w \mathcal{L}(f(x), y)(\nabla_w \mathcal{L}(f(x), y) - g^*) \\ &\quad (\because \text{ Taylor's first-order approximation with } \mu < \delta) \\ &= -\mu(\frac{\partial \mathcal{L}_{grad}}{\partial x})^T(\frac{\partial \mathcal{L}_{grad}}{\partial x}) \\ &= -\mu||\frac{\partial \mathcal{L}_{grad}}{\partial x}||_2^2 \\ &\quad (\because \text{ Recall that how } \Delta x \text{ is computed in the first step}).\end{aligned}$$

For the third term, $||v||_2^2 = ||\nabla_w \mathcal{L}(f(x), y) - g^*||_2^2 = \mathcal{L}_{grad}(x)$.

Then, summing up the three terms considered above will lead to the following inequality:

$$\begin{aligned}\mathcal{L}_{grad}(x+\Delta x) &\geq \mathcal{L}_{grad}(x) + (M^2\mu^2 - \mu)||\frac{\partial \mathcal{L}_{grad}}{\partial x}||_2^2 \\ &\geq \mathcal{L}_{grad}(x) - \frac{1}{4M^2}||\frac{\partial \mathcal{L}_{grad}}{\partial x}||_2^2 \\ &\quad (\because \text{ min. at } \mu = \frac{1}{2M^2}).\end{aligned}$$

Unlike the case for Theorem 1, the inequality above holds for any $\mu$, thus no restriction on $\mu$. □

## A3. Proof of Theorem 3

**Theorem.** (Hessian at ground truth for L2 distance). Suppose $\mathcal{L}_{grad}$ is L2 distance, then the hessian at ground truth is like the following:

$$H_{\text{L2}}(x^*) = J(x^*)^T J(x^*),$$

where $J(x^*)$ is the Jacobian of gradient with respect to input variable (i.e., $J(x^*) = \nabla_{x=x^*} g_w(x)$).

*Proof.* Note that $\mathcal{L}_{grad}(x) = \frac{1}{2}||\nabla_w \mathcal{L}(f(x), y) - g^*||_2^2$. Then, $\frac{\partial \mathcal{L}_{grad}(x)}{\partial x} = J(x)^T(\nabla_w \mathcal{L}(f(x), y) - g^*)$. Then, the hessian would be $\frac{\partial}{\partial x}\frac{\partial L_{grad}(x)}{\partial x} = \frac{\partial J(x)^T}{\partial x}(\nabla_w \mathcal{L}(f(x), y) - g^*) + J(x)^T J(x)$ (by Product Rule). Note that $\nabla_w \mathcal{L}(f(x^*), y) = g^*$. Thus, replacing $x$ with $x^*$ cancels the former term, thus $H_{\text{L2}}(x^*) = \frac{\partial}{\partial x}\frac{\partial \mathcal{L}_{grad}(x)}{\partial x}|_{x=x^*} = J(x^*)^T J(x^*)$ holds. □

## A4. Proof of Theorem 4

**Theorem.** (Hessian at ground truth for cosine distance). Suppose $\mathcal{L}_{grad}$ is cosine distance, then the hessian at ground truth is like the following:

$$H_{\cos}(x^*) = \frac{1}{||g^*||_2^2} J(x^*)^T (I - \frac{g^*}{||g^*||_2} \frac{g^{*T}}{||g^*||_2}) J(x^*),$$

$I$ is the identity matrix.

*Proof.* Note that $\mathcal{L}_{grad}(x) = 1 - \frac{g^*}{||g^*||_2^2}^T \frac{\nabla_w \mathcal{L}(f(x),y)}{||\nabla_w \mathcal{L}(f(x),y)||_2}$.

Let $v$ denote $\frac{\nabla_w \mathcal{L}(f(x),y)}{||\nabla_w \mathcal{L}(f(x),y)||_2}$ and $h$ denote $\nabla_w \mathcal{L}(f(x), y)$.

Then, $\frac{\partial L_{grad}(x)}{\partial x} = \frac{\partial h}{\partial x} \frac{\partial v}{\partial h} \frac{\partial h}{\partial v}$ (by Chain Rule) $= -J(x)^T \frac{1}{||h||} (I - \frac{1}{||h||^2} hh^T) \frac{g^*}{||g^*||_2}$.

Then, hessian $\frac{\partial}{\partial x} \frac{\partial L_{grad}(x)}{\partial x} = \frac{\partial}{\partial x}(-J(x)^T \frac{1}{||h||_2})(I - \frac{1}{||h||_2^2} hh^T) \frac{g^*}{||g^*||_2} - J(x)^T \frac{1}{||h||_2} \frac{\partial}{\partial x}(I - \frac{1}{||h||_2^2} hh^T) \frac{g^*}{||g^*||_2}$.

By replacing $x$ with $x^*$, the former term is canceled out because $(I - \frac{1}{||h||_2^2} hh^T) \frac{g^*}{||g^*||_2} = (I - \frac{1}{||g^*||_2^2} g^* g^{*T}) \frac{g^*}{||g^*||_2} = 0$.

Then, the latter term is only left, thus hessian would be $-J(x)^T \frac{1}{||h||_2} \frac{\partial}{\partial x}(I - \frac{1}{||h||_2^2} hh^T) \frac{g^*}{||g^*||_2} = 2J(x)^T \frac{1}{||h||_2^2} \frac{\partial}{\partial x}(I - \frac{1}{||h||_2^2} hh^T) J(x) \frac{h}{||h||_2}^T \frac{g^*}{||g^*||_2}$. Using the substitution $x = x^*$ (then, $h = g^*$), hessian would be like the following: $2J(x^*)^T \frac{1}{||g^*||_2^2} \frac{\partial}{\partial x}(I - \frac{1}{||g^*||_2^2} g^* g^{*T}) J(x^*) \frac{g^*}{||g^*||_2}^T \frac{g^*}{||g^*||_2} = 2J(x^*)^T \frac{1}{||g^*||_2^2} \frac{\partial}{\partial x}(I - \frac{1}{||g^*||_2^2} g^* g^{*T}) J(x^*)$, thus the theorem holds. □

## A5. The correlation between $L$ and $M$ (key variables in section 4.1) to the maximum and the minimum eigenvalues of the Hessian.

By Taylor's second-order approximation, $\mathcal{L}_{grad}(x^* + \Delta x) = \mathcal{L}_{grad}(x^*) + \Delta x^T \nabla_{x=x^*} \mathcal{L}_{grad}(x) + \frac{1}{2} \Delta x^T H(x^*) \Delta x$. For $\mathcal{L}_{grad}$ being either L2 or cosine distance, $\mathcal{L}_{grad}(x^*) = 0$ and $\nabla_{x=x^*} \mathcal{L}_{grad}(x) = \mathbf{0}$ hold (see the proofs in A3. and A4.). Then, we can rewrite $\mathcal{L}_{grad}(x^* + \Delta x) = \frac{1}{2} \Delta x^T H(x^*) \Delta x$ and the following holds.

$$\frac{||\nabla_w \mathcal{L}(f(x^* + \Delta x), y) - \nabla_w \mathcal{L}(f(x^*), y)||_2}{||\Delta x||_2}$$

$$= \frac{\sqrt{\mathcal{L}_{grad}(x^* + \Delta x)}}{||\Delta x||_2}$$

$$= \frac{\sqrt{\mathcal{L}_{grad}(x^* + \Delta x) - \mathcal{L}_{grad}(x^*)}}{||\Delta x||_2} \; (\because \mathcal{L}_{grad}(x^*) = 0)$$

$$= \frac{\sqrt{\frac{1}{2} \Delta x^T H(x^*) \Delta x}}{||\Delta x||_2}$$

, and

$\frac{\sqrt{\lambda_{min}}}{\sqrt{2}} \leq \frac{\sqrt{\frac{1}{2} \Delta x^T H(x^*) \Delta x}}{||\Delta x||_2} \leq \frac{\sqrt{\lambda_{max}}}{\sqrt{2}}$, where $\lambda_{min}$ and $\lambda_{max}$ are minimum and maximum eigenvalues of $H(x^*)$. Therefore, $\frac{\sqrt{\lambda_{max}}}{\sqrt{2}}$ and $\frac{\sqrt{\lambda_{min}}}{\sqrt{2}}$ provides the lower and upper bounds of bi-Lipschitz constants $L$ and $M$ in Theorem 1 and 2 respectively. Note that $\frac{\sqrt{\lambda_{max}}}{\sqrt{2}}$ and $\frac{\sqrt{\lambda_{min}}}{\sqrt{2}}$ are exactly $L$ and $M$ respectively for the special case when $x_2 = x^*$, which is of our interest.

# Algorithms

In Algorithm 1, we describe how the maximum eigenvalue of Hessian is computed using the pseudocode.

# Limitations and Future Work

Our work is focused on pure gradient matching loss for fundamental analysis, without batch statistics matching loss. However, state-of-the-art method currently uses batch statistics matching, thus a theoretical approach on optimizing batch statistics matching is required to craft more advanced proxy for the vulnerability under state-of-the-art gradient inversion attacks.

# Plots on Other Datasets

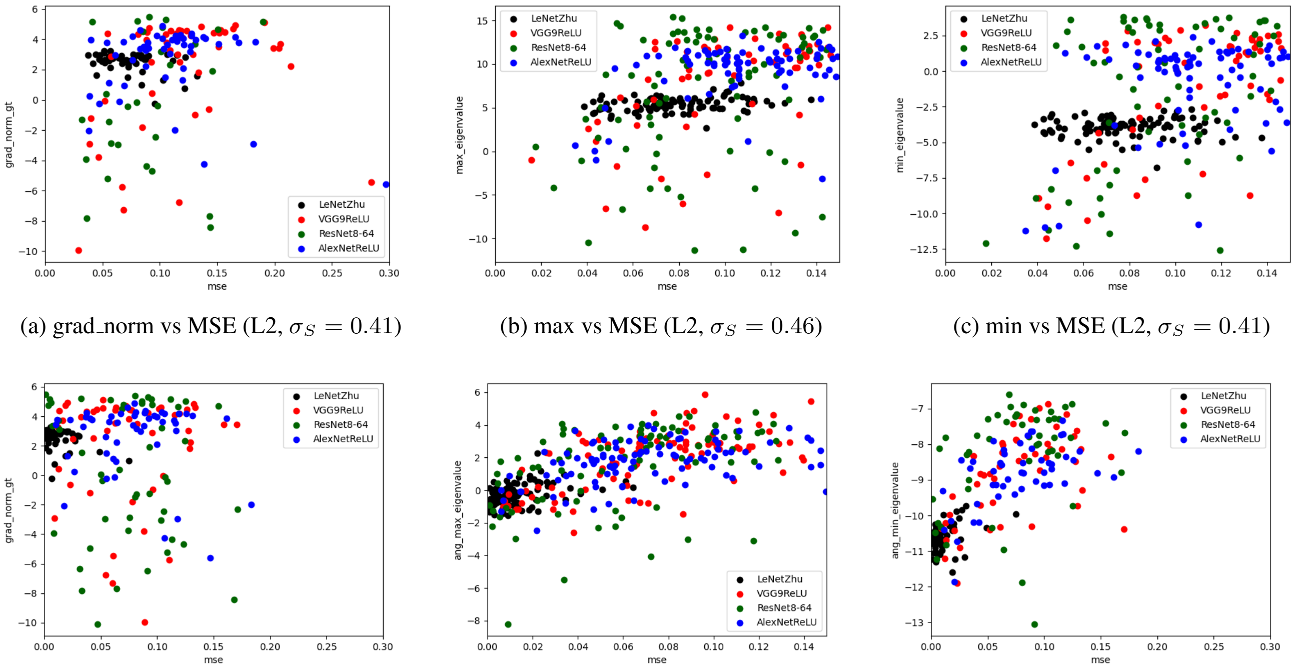
We included plot results for the qualitative comparison between gradient norm and our proposed measures for datasets CIFAR100 (in Figure 4), ImageNette (in Figure 5), and ImageWoof (in Figure 6).

# Understanding LAVP

In this section, we investigate whether LAVP works as expected from our theory. To simulate the local optimization scenario, the initialized image in the attack is sampled from $x_i = x^* + 0.1 sign(\mathcal{N})$, where $\mathcal{N}$ is the normal distribution and $sign$ is the sign function. This initialization scheme always ensures initial reconstruction error as MSE with $0.1$, thus simulating local optimization. In Table 4, we present the final gradient matching loss average with its standard deviation in this setup. For each sample, we run the attack algorithm five times and SGD optimizer is used for optimization. Note that LAVP-L2 is aligned with $\mathcal{L}_{grad}^{final}$ (L2) and LAVP-cos is aligned with $\mathcal{L}_{grad}^{final}$ (cos) while gradient norm does not correlate with final loss for any kind of loss function.

| Values / Image index | Image 1 | Image 2 | Image 3 |
|---|---|---|---|
| $\mathcal{L}_{grad}^{final}$ (L2) | 307.75 ± 15.45 | **0.001 ± 0.00** | 526.42 ± 50.32 |
| $\mathcal{L}_{grad}^{final}$ (cos) | 0.053 ± 0.02 | 0.032 ± 0.00 | **0.014 ± 0.00** |
| LAVP-L2(max) | 4.70E05 | **1.81** | 2.19E06 |
| LAVP-L2(min) | 1.64 | **3.70E-06** | 3.71 |
| LAVP-cos(ang_max) | 4.53 | 0.99 | **0.08** |
| LAVP-cos(ang_min) | 1.20E-03 | 4.00E-04 | **8.00E-05** |
| gradient norm | 24.64 | **0.07** | 42.02 |

Table 4: **Final gradient matching losses ($\mathcal{L}_{grad}^{final}$ (L2 distance), $\mathcal{L}_{grad}^{final}$ (cosine distance)), LAVPs for L2, LAVPs for cosine, and gradient norm for three different images. ) / LPIPS(↓)).** For each value, the smallest value among three images is marked as **bold** and the largest value among them is marked as underlined.



(a) grad_norm vs MSE (L2, $\sigma_S = 0.41$)

(b) max vs MSE (L2, $\sigma_S = 0.46$)

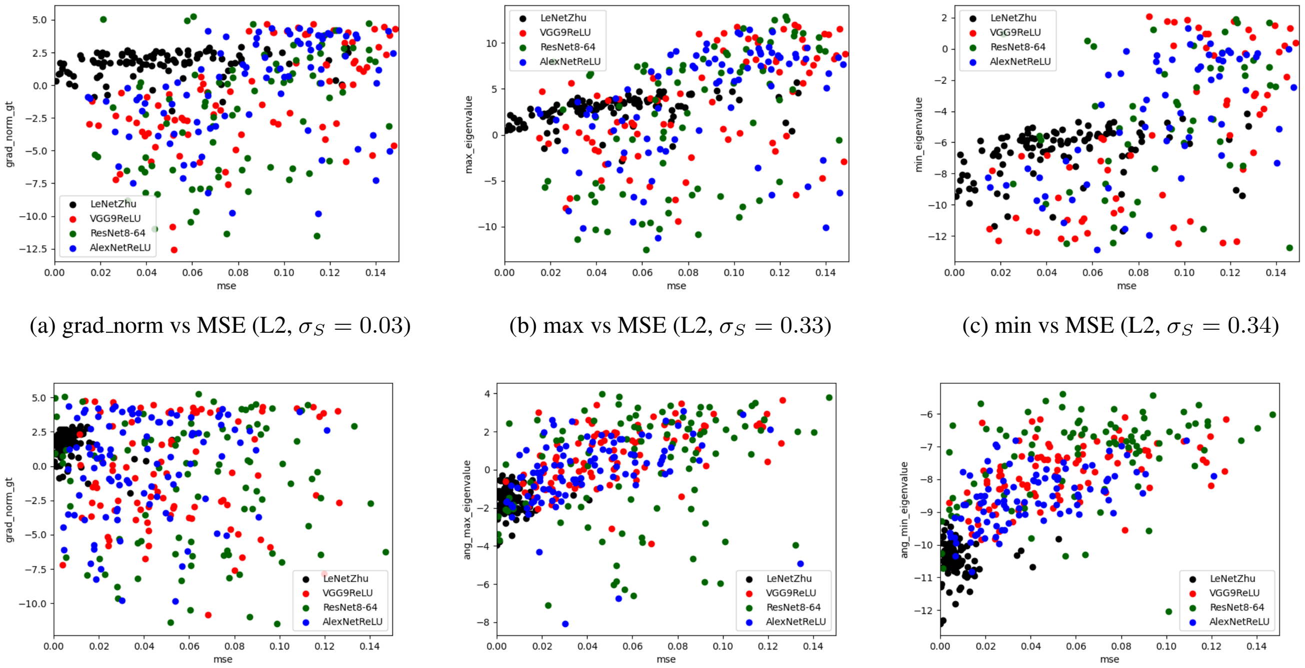
(c) min vs MSE (L2, $\sigma_S = 0.41$)

(d) grad_norm vs MSE (CS, $\sigma_S = 0.1$)

(e) ang_max vs MSE (CS, $\sigma_S = 0.67$)

(f) ang_min vs MSE (CS, $\sigma_S = 0.69$))

Figure 4: **Comparison of gradient norm, maximum and minimum eigenvalues of Hessian in terms of the correlation with MSE of reconstructed samples over several architectures on CIFAR100 test samples.**



(a) grad_norm vs MSE (L2, $\sigma_S = 0.03$)

(b) max vs MSE (L2, $\sigma_S = 0.33$)

(c) min vs MSE (L2, $\sigma_S = 0.34$)

(d) grad_norm vs MSE (CS, $\sigma_S = -0.13$)

(e) ang_max vs MSE (CS, $\sigma_S = 0.57$)

(f) ang_min vs MSE (CS, $\sigma_S = 0.75$))

Figure 5: **Comparison of gradient norm, maximum and minimum eigenvalues of Hessian in terms of the correlation with MSE of reconstructed samples over several architectures on ImageNette test samples.**