

In-context Learning and Gradient Descent Revisited

Gilad Deutch*

Nadav Magar*

Tomer Bar Natan

Guy Dar

The Blavatnik School of Computer Science
Tel Aviv University

Abstract

In-context learning (ICL) has shown impressive results in few-shot learning tasks, yet its underlying mechanism is still not fully understood. A recent line of work suggests that ICL performs gradient descent (GD)-based optimization implicitly. While appealing, much of the research focuses on simplified settings, where the parameters of a shallow model are optimized. In this work, we revisit evidence for ICL-GD correspondence on realistic NLP tasks and models. We find gaps in evaluation, both in terms of problematic metrics and insufficient baselines. We show that surprisingly, even untrained models achieve comparable ICL-GD similarity scores despite not exhibiting ICL. Next, we explore a major discrepancy in the flow of information throughout the model between ICL and GD, which we term *Layer Causality*. We propose a simple GD-based optimization procedure that respects layer causality, and show it improves similarity scores significantly. Our code implementation is available at: <https://github.com/GilDe/ft-vs-icl>.

1 Introduction

In recent years, large language models have shown strong emergent in-context learning abilities (Brown et al., 2020; Wei et al., 2022) – where a pre-trained model’s performance significantly improves on a task by conditioning the language model on a small set of input-label pairs (demonstrations). Despite substantial research, the inner workings of ICL remain elusive. At face value, in-context learning and gradient descent-based finetuning have very little in common. Nevertheless, a series of recent studies discuss apparent similarities between ICL and gradient descent-based optimization, mostly in synthetic scenarios (von Oswald et al., 2023a,b; Akyürek et al., 2023; Ahn et al., 2023, *inter alia*). The claim this body of research aims to make is that

ICL can implement implicit GD, using in-context demonstrations as training examples. While most of the synthetic setups concern: (1) restricted transformers, (2) simplified regression tasks, and (3) direct training for ICL – the work of Dai et al. (2023) stands out in its ability to demonstrate an ostensible similarity between ICL and GD optimization in (1) full-fledged transformers, (2) for realistic NLP tasks, (3) naturally occurring in models trained only on causal text generation. We call the hypothesis that ICL mimics finetuning *on the model itself* – as is analyzed in Dai et al. (2023) – the *strong ICL-GD correspondence*. We will later discuss how this diverges from the ICL-GD correspondence other works consider.

In this paper, we make two main complementary contributions. We perform a careful re-analysis of the work of Dai et al. (2023) and show how seemingly mild problems in evaluation lead to a significant overestimation of similarity between the two procedures. Surprisingly, we find that untrained models achieve similarity scores *at least as good* as trained ones. This result provides strong evidence *against* the strong ICL-GD correspondence.¹

Secondly, in an attempt to relax the strong ICL-GD correspondence hypothesis, we suggest a rectified version of GD that we show aligns better with ICL. To do this, we first identify a core discrepancy in the flow of information throughout the model between in-context learning and vanilla gradient descent, which we call *Layer Causality*. In ICL, the information that influences the hidden state comes from the output of shallow layers (“earlier layers”) alone. In GD, however, the update to the weights of a layer depends on gradients, which come from all of the model layers including deeper (“later layers”). We showcase the importance of this simple

¹However, it should be noted that the similarity metrics assume a certain correspondence holds in every layer in a specific way. It does not preclude the possibility of a relaxed correspondence.

*Equal contribution

observation by suggesting a simple variant of GD that incorporates layer causality. This simple modification, *Layer Causal Gradient Descent* (LCGD), consistently improves upon vanilla gradient descent on the similarity metrics. Notably, it outperforms the trained transformer significantly in terms of both similarity metrics. In comparison to the untrained baselines, it significantly surpasses them in attention map similarity (SimAM Δ) and is consistently on the high end in terms of hidden state similarity (SimAOU). In spite of that, the scores are still low. This can be due to a suboptimal choice of hyperparameters but likely has to do with inherent problems in the strong ICL-GD correspondence hypothesis, even with the layer causal version we propose. We leave this for future work to explore.

Lastly, we dedicate a short discussion to the line of work on synthetic settings that builds on insights from von Oswald et al. (2023a). We observe terminology differences with Dai et al. (2023) that might cause confusion. “Gradient Descent” is used differently in both cases. While synthetic settings usually consider gradients of shallow implicit functions, Dai et al. (2023) consider complex gradients with respect to the model itself. In the synthetic setting, layer causality is often trivially satisfied.

Our contributions are the following:

- ▶ We discuss issues in the evaluation process of Dai et al. (2023) in terms of baselines and evaluation metrics. Notably, we demonstrate that untrained transformers perform as well as pretrained models.
- ▶ We highlight core problems with the hypothesis that GD approximates ICL in the naive sense. We study a layer-causal GD variant and demonstrate empirically that it is better at simulating ICL.
- ▶ Finally, we briefly survey works in synthetic settings and find that their ICL-GD correspondence is significantly different from the strong ICL-GD correspondence which we try to refute.

In summary, our work shows there’s little evidence for the strong ICL-GD correspondence in its current form. We show a non-trivial increase in the similarity metrics (especially in SimAM Δ) with a layer-causal variant. This might suggest that a weaker, more nuanced hypothesis might hold. However, we acknowledge there may be irrelevant causes for the increase.

All code for replicating our work is publicly available at: <https://github.com/GiilDe/ft-vs-icl>.

2 Preliminaries

In this work, we build on the benchmark proposed by Dai et al. (2023). We focus on its setting using the same datasets and examine the same similarity metrics to compare the behavior of ICL and finetuning. This section provides details on the benchmark they use. In the next section, we will address problems in the metrics described below.

2.1 Datasets

Following Dai et al. (2023), we use six datasets for our experiment: **SST2** (Socher et al., 2013), **SST5** (Socher et al., 2013), **MR** (Pang and Lee, 2005) and **Subj** (Pang and Lee, 2004) are four datasets for sentiment classification; **AGNews** (Zhang et al., 2015) is a topic classification dataset; and **CB** (de Marnaffe et al., 2019) is used for natural language inference. Data statistics are provided in Table 3 (Appendix A).

2.2 Metric I: SimAOU Normalized

The first metric quantifies the similarity of two setups (finetuning and in-context learning) in terms of the attention output (AO) vector of each layer. More precisely, we quantify the similarity between the *changes* to the AO vector (changes being the difference from the AO vector in the zero-shot setup). Given a test prompt, let $h_S^{(l)}$ be the output representation of the last token at the l -th attention layer in setting S where $S \in \{\text{ZSL}, \text{ICL}, \text{FT}\}$ – zero-shot learning, in-context learning, and finetuning. The updates induced by ICL and finetuning are given by $h_{\text{ICL}}^{(l)} - h_{\text{ZSL}}^{(l)}$ and $h_{\text{FT}}^{(l)} - h_{\text{ZSL}}^{(l)}$, respectively. The *attention output update similarity* (SimAOU) is defined as the cosine similarity between these updates, averaged across all layers. A high SimAOU score indicates that ICL adjusts the attention output in the same direction as finetuning. As a baseline, they compare with random attention output updates: $h_{\text{rand}}^{(l)} - h_{\text{ZSL}}^{(l)}$ where $h_{\text{rand}}^{(l)}$ is sampled uniformly. We note that the authors used a slight variation of this, where $h_S^{(l)}$ is *normalized* before computing the difference. We call this metric SimAOU $_{\text{norm}}$ and would later show that this normalization can cause misleading results.

| | | SST2 | SST5 | MR | Subj | CB |
|-------------------|--------------------|------------------------|------------------------|------------------------|------------------------|------------------------|
| SimAOU | Trained | 0.05 \pm 0.01 | 0.04 \pm 0.02 | 0.17 \pm 0.03 | 0.06 \pm 0.01 | 0.11 \pm 0.01 |
| | Trained Embeddings | 0.11 \pm 0.02 | <u>0.06</u> \pm 0.00 | 0.24 \pm 0.00 | 0.20 \pm 0.00 | 0.01 \pm 0.00 |
| | No Training | 0.09 \pm 0.00 | 0.07 \pm 0.03 | 0.18 \pm 0.03 | 0.06 \pm 0.01 | 0.04 \pm 0.01 |
| SimAM $_{\Delta}$ | Trained | 0.15 \pm 0.02 | 0.31 \pm 0.02 | 0.14 \pm 0.05 | 0.25 \pm 0.07 | <u>0.25</u> \pm 0.01 |
| | Trained Embeddings | 0.09 \pm 0.02 | 0.03 \pm 0.00 | 0.18 \pm 0.02 | 0.16 \pm 0.02 | 0.05 \pm 0.10 |
| | No Training | 0.11 \pm 0.04 | 0.05 \pm 0.03 | 0.16 \pm 0.03 | 0.17 \pm 0.03 | 0.26 \pm 0.05 |

Table 1: SimAOU and SimAM comparison of vanilla GD for trained and untrained transformers. When the difference between the highest and second-highest score in a column is ≤ 0.01 , we underline both scores.

2.3 Metric II: SimAM

SimAM is used to measure the similarity between attention *maps* of ICL and finetuning. Given a test example, let $m_S^{(l,h)}$ represent the attention weights before softmax in the h -th head of the l -th layer for setting S . In ICL, we focus solely on the test examples’ token attention weights, excluding demonstration tokens so that the shapes of FT and ICL attention weights will be compatible. We calculate the cosine similarity between $m_{\text{ICL}}^{(l,h)}$ and $m_{\text{FT}}^{(l,h)}$ to obtain SimAM. Notice here we do not measure the similarity between *updates* but rather between the raw attention weights themselves. We will return to this shortly when we analyze the metric choices and biases they introduce into the benchmark.

3 Rethinking the Benchmark

3.1 SimAOU

In the original setting, Dai et al. (2023) have shown that random noise gets a minuscule score on this metric. However, we show that even two random update vectors of sufficient norm can achieve a high SimAOU score. Let $\mathbf{z} = h_{\text{ZSL}}^{(l)}$ be the *unnormalized* attention output in zero-shot. Assume $\mathbf{r}, \mathbf{r}' \sim \mathcal{N}(0, \sigma I)$ are random gaussian noise vectors with variance σ^2 . Now, choose σ such that $\|\mathbf{r}\|^2 = \|\mathbf{r}'\|^2 = 3\|\mathbf{z}\|^2$ holds² and set $\mathbf{z}_{\text{ICL}} = \mathbf{z} + \mathbf{r}$, $\mathbf{z}_{\text{FT}} = \mathbf{z} + \mathbf{r}'$. The random vectors are approximately uncorrelated with each other and with \mathbf{z} , that is $\mathbf{z}^T \mathbf{r} = \mathbf{r}^T \mathbf{r}' = \mathbf{z}^T \mathbf{r}' = 0$. By the Pythagorean theorem, $\|\mathbf{z}_{\text{ICL}}\|^2 = \|\mathbf{z} + \mathbf{r}\|^2 = \|\mathbf{z}\|^2 + \|\mathbf{r}\|^2 = 4\|\mathbf{z}\|^2 = \|\mathbf{z} + \mathbf{r}'\|^2 = \|\mathbf{z}_{\text{FT}}\|^2$. So, $\|\mathbf{z}_{\text{FT}}\| = \|\mathbf{z}_{\text{ICL}}\| = 2\|\mathbf{z}\|$. We get that

²This will make the computation cleaner, but other options such as $\|\mathbf{z}\| = \|\mathbf{r}\| = \|\mathbf{r}'\|$ are just as good, leading to slightly different similarity scores.

SimAOU_{norm} equals:

$$\begin{aligned} & \frac{\frac{\mathbf{z}_{\text{ICL}}}{\|\mathbf{z}_{\text{ICL}}\|} - \frac{\mathbf{z}}{\|\mathbf{z}\|}}{\left\| \frac{\mathbf{z}_{\text{ICL}}}{\|\mathbf{z}_{\text{ICL}}\|} - \frac{\mathbf{z}}{\|\mathbf{z}\|} \right\|} \cdot \frac{\frac{\mathbf{z}_{\text{FT}}}{\|\mathbf{z}_{\text{FT}}\|} - \frac{\mathbf{z}}{\|\mathbf{z}\|}}{\left\| \frac{\mathbf{z}_{\text{FT}}}{\|\mathbf{z}_{\text{FT}}\|} - \frac{\mathbf{z}}{\|\mathbf{z}\|} \right\|}} = \\ & \frac{\frac{\mathbf{z} + \mathbf{r}}{2\|\mathbf{z}\|} - \frac{\mathbf{z}}{\|\mathbf{z}\|}}{\left\| \frac{\mathbf{z} + \mathbf{r}}{2\|\mathbf{z}\|} - \frac{\mathbf{z}}{\|\mathbf{z}\|} \right\|}} \cdot \frac{\frac{\mathbf{z} + \mathbf{r}'}{2\|\mathbf{z}\|} - \frac{\mathbf{z}}{\|\mathbf{z}\|}}{\left\| \frac{\mathbf{z} + \mathbf{r}'}{2\|\mathbf{z}\|} - \frac{\mathbf{z}}{\|\mathbf{z}\|} \right\|}} = \\ & \frac{\mathbf{r} - \mathbf{z}}{\|\mathbf{r} - \mathbf{z}\|} \cdot \frac{\mathbf{r}' - \mathbf{z}}{\|\mathbf{r}' - \mathbf{z}\|} = \frac{\|\mathbf{z}\|^2}{2\|\mathbf{z}\| \cdot 2\|\mathbf{z}\|} = \frac{1}{4} \end{aligned}$$

The problem our computation reveals is the fact that after normalization, \mathbf{z} terms don’t cancel out completely and interact with each other. This is a general problem not limited to random noise. We compare unnormalized SimAOU with SimAOU_{norm} in Table 2 and show it has a substantial impact on the similarity scores.

3.2 SimAM $_{\Delta}$

To better measure the similarity between the updates to the attention maps induced by ICL and FT, we suggest a modified metric, SimAM $_{\Delta}$. Specifically we compute the cosine similarity between $m_{\text{ICL}}^{(l,h)} - m_{\text{ZS}}^{(l,h)}$ and $m_{\text{FT}}^{(l,h)} - m_{\text{ZS}}^{(l,h)}$, the *update* vectors. The new metric is no longer sensitive to the magnitude of the update vector. In the original setting, the cosine similarity might be dominated by $m_{\text{ZS}}^{(l,h)}$ so a model drifting further during FT from $m_{\text{ZS}}^{(l,h)}$ will be penalized even if the update direction is more similar to ICL’s. Update size in general can be manipulated by adjusting the learning rate,³ and so should not be a core feature of the similarity metric.

3.3 Untrained Transformer Baseline

We have discussed problems with metrics. We now turn to baselines. We use untrained models

³Though effects are not guaranteed to change linearly with the change learning rate, as learning rate change can often have unpredictable effects.

as our baseline. In-context learning is an emergent property attained through pretraining (Brown et al., 2020), therefore any similarity between the “ICL”⁴ setup and the finetuning setup on untrained models cannot be attributed to a learned form of mesa-optimization (Hubinger et al., 2021). In Table 1, we compare the original model with two baselines: a completely untrained model (*No Training*) and a model where we kept the input and output embeddings (including positional embeddings) and layer norms (*Trained Embeddings*). We find that in terms of SimAOU the untrained baselines slightly exceed vanilla GD.

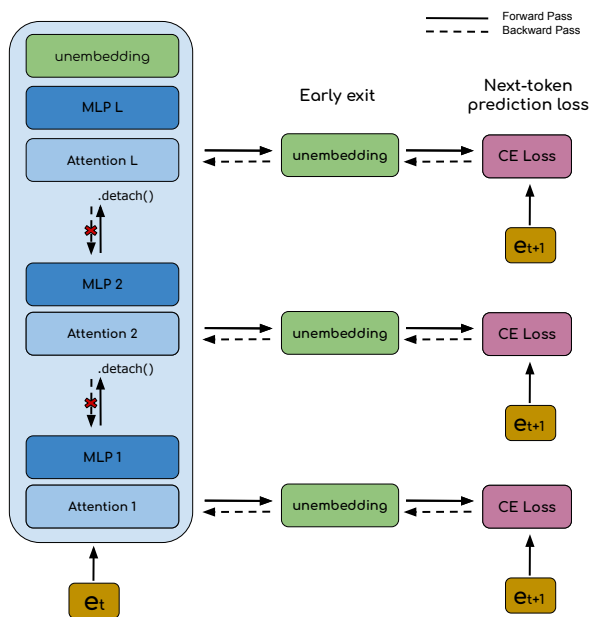


Figure 1: **Layer-causal GD**: The output of each layer is projected to the label space and used as an intermediate prediction. We compute the prediction loss of each intermediate layer sequentially.

4 Investigation into Layer Causality

4.1 Layer Causality

We characterize a core problem with the strong ICL-GD correspondence in the following statement.

Layer Causality. *In ICL, the update to the output of the l -th attention layer is dependent only on the output of previous (lower) layers. In contrast, the update to the l -th attention output induced by finetuning is determined by the gradient of the entire model’s trainable parameters.*

⁴Formally speaking, it shouldn’t really qualify as ICL, as the model hasn’t attained this capability yet.

4.2 Design Choices

Motivated by this observation, we propose to use a layer causality-compatible finetuning method, where each layer is updated individually, instead of propagating information back to earlier layers. Then, we will explore how a layer-causal variant fares compared to full-blown vanilla gradient descent. There are many possible ways to design such an algorithm. In this work, we will define an instantiation of layer causality-compatible optimization, that we call *Layer-causal Gradient Descent* (LCGD). We make the decision based on the following guiding principles:

- ▷ **Minimal Changes**: We want to leave the procedure as close as possible to vanilla GD. The goal is to isolate the effect of layer causality on the modification we make as much as possible. Otherwise, other design decisions might come into play.
- ▷ **Simplicity**: We want the procedure to be interpretable and easy to reason about.
- ▷ **Plausibility** (Occam’s razor): We want to design a “plausible” procedure. A major part of what we call plausibility is layer causality. Plausibility in a broader sense may include any other aspect that one cannot expect a forward pass of the model to easily implement using a clear and simple mechanism.

These principles might conflict. We prioritize them in the following way: we want the procedure to be layer-causal (a special case of plausibility), but other than that, we will always favor the first and second principles. One example of where we favor simplicity over plausibility is when we choose to take the derivative of the entire layer on every step of the procedure (see below), including the softmax operation. This goes against plausibility because the derivative of softmax cannot be plausibly computed with a single attention layer.

4.3 Motivation: Short-circuited Transformers

A simple finetuning method that respects layer causality is by short-circuiting a model at any layer l , i.e. by removing all layers from $l + 1$ onwards. In a normal (not short-circuited) forward pass, the model outputs the next-token prediction by taking the final hidden state, applying a final layer norm operation to it, and multiplying by the output embedding matrix (a.k.a. the unembedding matrix).

| | SST2 | SST5 | MR | Subj | AGNews | CB | Average |
|-------------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|--------------|
| SimAOU _{norm} (GD) | 0.11 \pm 0.00 | 0.09 \pm 0.01 | 0.22 \pm 0.01 | 0.18 \pm 0.02 | 0.31 \pm 0.04 | 0.21 \pm 0.01 | 0.187 |
| SimAOU _{norm} (LCGD) | 0.22 \pm 0.01 | 0.11 \pm 0.00 | 0.33 \pm 0.01 | 0.35 \pm 0.01 | 0.33 \pm 0.01 | 0.34 \pm 0.00 | 0.279 |
| SimAOU (GD) | 0.05 \pm 0.01 | 0.04 \pm 0.02 | 0.17 \pm 0.03 | 0.06 \pm 0.01 | 0.18 \pm 0.03 | 0.11 \pm 0.01 | 0.102 |
| SimAOU (LCGD) | 0.13 \pm 0.01 | 0.11 \pm 0.01 | 0.21 \pm 0.03 | 0.18 \pm 0.00 | 0.13 \pm 0.01 | 0.24 \pm 0.01 | 0.167 |
| SimAM (GD) | 0.59 \pm 0.01 | 0.40 \pm 0.03 | 0.49 \pm 0.01 | 0.45 \pm 0.06 | 0.48 \pm 0.04 | 0.20 \pm 0.03 | 0.435 |
| SimAM (LCGD) | 0.58 \pm 0.01 | 0.39 \pm 0.03 | 0.30 \pm 0.00 | 0.27 \pm 0.01 | 0.12 \pm 0.00 | 0.04 \pm 0.01 | 0.283 |
| SimAM $_{\Delta}$ (GD) | 0.15 \pm 0.02 | 0.31 \pm 0.02 | 0.14 \pm 0.05 | 0.25 \pm 0.07 | 0.50 \pm 0.05 | 0.25 \pm 0.01 | 0.267 |
| SimAM $_{\Delta}$ (LCGD) | 0.30 \pm 0.02 | 0.33 \pm 0.01 | 0.26 \pm 0.00 | 0.32 \pm 0.01 | 0.43 \pm 0.02 | 0.38 \pm 0.01 | 0.336 |

Table 2: SimAOU and SimAM comparison of vanilla GD and layer-causal GD across six classification datasets. Layer causal GD achieves higher SimAOU across all tasks, yet its SimAM is significantly lower. SimAM $_{\Delta}$ is higher for layer causal GD, except for AGNews.

Analogously, in a model short-circuited at layer l , the next-token prediction is obtained by projecting the l -th hidden state on the unembedding matrix, after applying the final layer norm. This is justified by the *early exit* approach (Teerapittayanon et al., 2017; Din et al., 2023), where it has been observed that a short-circuited model is often sufficiently good at predicting the next token. Early exit is closely related to the residual stream hypothesis (nostalgebraist, 2020; Elhage et al., 2021; Geva et al., 2022; Dar et al., 2023), which stipulates that language models refine the next-token prediction throughout the layers – and so projecting internal layers into the vocabulary space gives the current prediction in every layer. We will refer to the combination of the final layer norm and the unembedding matrix as the unembedding projection head and denote it by the function $U(\cdot)$.

4.4 Algorithm

We now describe the LCGD finetuning procedure. In LCGD we project the output of each layer onto logits in the vocabulary space using the unembedding head $U(\cdot)$ and compute the cross-entropy loss of this prediction with respect to the one-hot embedding of the next token. Unlike vanilla finetuning, it does not violate the causal structure of the network, as it depends only on data available at this layer. To reiterate, $U(\cdot)$ normally takes the final hidden state of the model and projects it onto the logits over the vocabulary. We follow the early exit/residual stream approach and apply it on internal hidden states.

Let the *detached* hidden states after the ℓ -th at-

tention layer at token i be denoted:

$$\hat{h}_i^\ell = \text{Attn} \left(W_V \text{SG}(X^\ell), W_K \text{SG}(X^\ell), \text{SG}(\mathbf{q}_i^\ell) \right)$$

where $\text{SG}(\cdot)$ stands for the “stop gradient” operation (also called `.detach()` in PyTorch) which does not affect the forward pass, but in the backward pass it does not back-propagate the gradient to its input, meaning it is treated as a constant. Let the tokens of the model be represented by a list of one-hot vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_T$. For each token, we define the objective function:⁵

$$\mathcal{L} = \sum_{\ell=1}^L \text{CE} \left(U(\hat{h}_i^\ell), \mathbf{e}_{i+1} \right)$$

U is taken to be frozen as well. CE is cross-entropy loss. We optimize by taking steps with respect to the gradient $\nabla_W \mathcal{L}$, one token at a time, where the “stop gradient” operator makes sure each layer is updated independently.

4.5 Experimental Setup

We use the same GPT-like pre-trained language models used by Dai et al. (2023) with 1.3B implemented in fairseq.⁶ We test vanilla and layer-causal GD in terms of their similarity to ICL with the four variants we discussed above (SimAOU, SimAOU_{norm}, SimAM, SimAM $_{\Delta}$). For reliable results, we average across 3 different seeds. This whole project’s computation took the equivalent of 12 hours on a single Tesla V100 GPU. Table 2

⁵Notice that we are allowed to take the sum of the cross-entropies across all layers in parallel, as the updates to the weight matrices will take effect only when processing the next token.

⁶<https://github.com/facebookresearch/fairseq>

shows both variants of SimAOU and SimAM for both methods.

Overall, with the exception of **AGNews**, layer-causal GD is significantly more aligned with ICL in terms of the modified similarity metrics and the normalized variant of SimAOU. However, it is important to note that the modified metrics are low for both variants. In comparison to untrained transformers, LCGD is much better in terms of SimAM $_{\Delta}$ and is mostly better by some small margin in terms of SimAOU.

Comparison with Untrained Baselines Combining Tables 1 & 2, we see that LCGD is competitive with respect to all three contenders, showing high-end scores consistently across the board, while it is not always the highest in terms of SimAOU. In terms of SimAM $_{\Delta}$ is *significantly* better than any of the other baselines across all datasets explored. There remains work to be done to show this advantage is indeed due to structural superiority and not rudimentary features, such as its ability to impact layers more strongly (as the gradient norm of updates in LCGD is larger – see Appendix C), which could have accumulating effects across layers and timesteps. Even if this is the case, it is important to understand the implications of this observation on other variants as well. We leave it for future research to work out the correct interpretation of the results in this section.

4.6 Additional Experiments

In Appendix B, we perform a more fine-grained comparison of LCGD and vanilla GD. First, we try to assess how similar the two variants are in the latent space, the intuition being that the layer-causal variant can be a simple approximation to vanilla GD. We find that this similarity is in fact relatively low, around 0.1 more or less in terms of cosine similarity, across datasets (this is shown in Figure 2). Then, we perform a layerwise analysis of the way the similarity scores change. The results are shown in Figure 3. We see a non-trivial variability in the similarity across layers, which seems to suggest a non-uniform behavior across layers. Curiously, we see that LCGD is not better in all layers. In the case of SimAOU, we see a small advantage for LCGD across virtually all layers, but the dynamics of SimAM $_{\Delta}$ are more complicated, suggesting deeper analysis is required to fully understand the advantage of LCGD over GD (see Appendix for more details on the additional experiments).

5 Conflation of Terms in ICL-GD Correspondence

Works rooted in the work of von Oswald et al. (2023a) usually have a common structure: The model is given training examples of the form $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$, where it holds that $y_i = f_{\theta}(\mathbf{x}_i)$ for some latent parameter vector θ .⁷ The model is also fed a test query \mathbf{x}_{test} . It is trained to output the value $y_{\text{test}} = f_{\theta}(\mathbf{x}_{\text{test}})$. The function f_{θ} is always a shallow function, usually a linear model $f_{\theta}(\mathbf{x}) = \theta^{\top} \mathbf{x}$, or a kernel regression problem. This distinction is important since the gradient of such functions has a simple closed form. This is in stark contrast to Dai et al. (2023), where the gradient is unwieldily complicated. Another difference is that the gradient in Dai et al. (2023) is computed with respect to the transformer itself, not a subsidiary function f_{θ} . In these crucial aspects, the gradients discussed are extremely different. The strong ICL-GD correspondence explored in Dai et al. (2023) is different than the one that the ICL-shallow GD correspondence von Oswald et al. (2023a) considered – the use of the term “Gradient Descent” in these two cases is incompatible. In Appendix D, we go over a subset of these works to demonstrate what kinds of shallow GD they rely on.

6 Discussion

In this work, we provide different perspectives on the ICL-GD correspondence. We show evidence against it but also show that it might be fixed. We find that previous work does not justify the strong ICL-GD correspondence, and instead discusses a weaker notion of a shallow GD. This should also apply to layer-causal GD, as it is designed as a modification of the strong ICL-GD correspondence. Still, we see it outperforms untrained transformers in terms of attention map similarity (and fares well in terms of hidden state similarity). This can be due to irrelevant causes (see limitations below). However, it is worth noting that the layer-causal variant can be justified by its similarity to the kernel regression and functional GD variants that have been addressed in the literature on synthetic settings (Cheng et al., 2024). Future work can use the (corrected) similarity metrics suggested in Dai et al. (2023) to gauge the similarity of shallow GD methods to ICL.

⁷ f_{θ} can be stochastic.

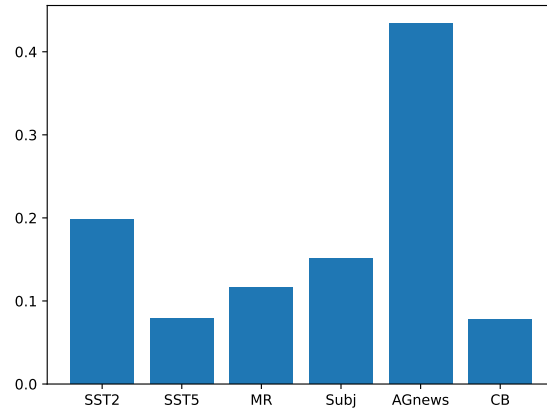


Figure 2: α averaged over all layers for each task. Computed for one seed per task.

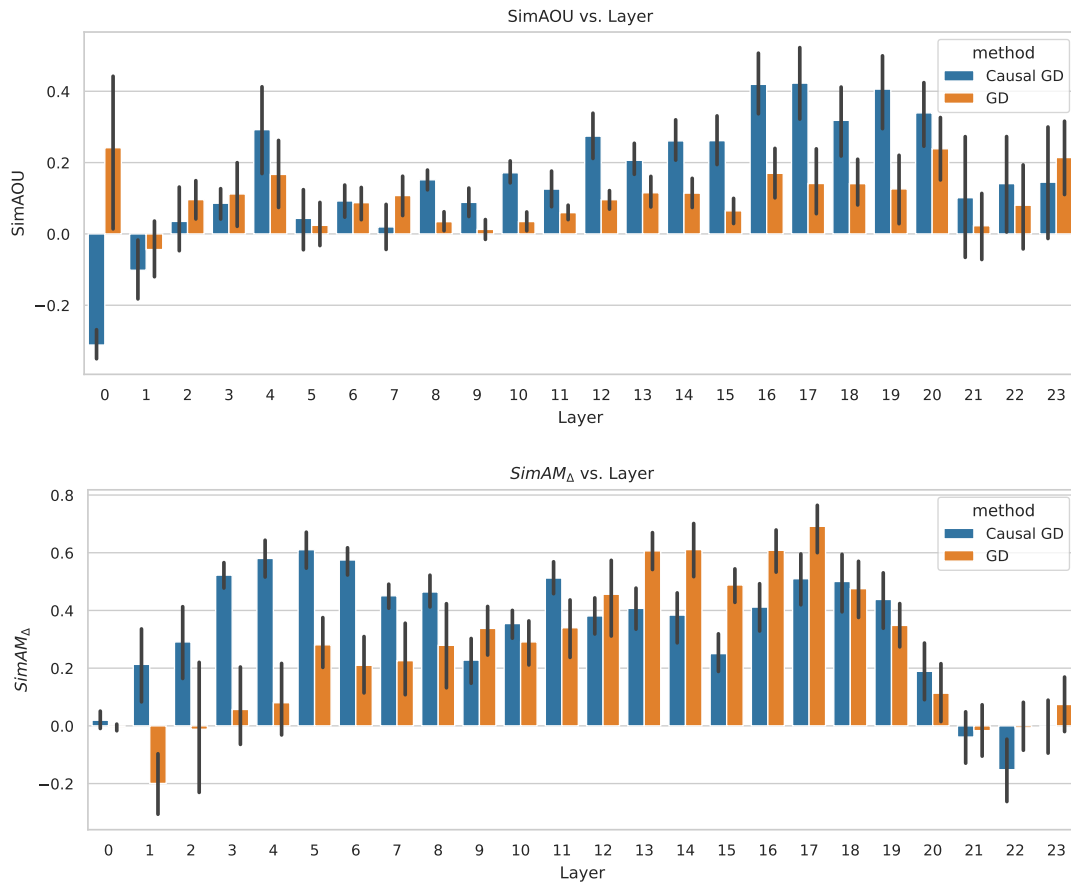


Figure 3: Similarity computed per layer aggregated across tasks and seeds. Error bar is presented. Blue bars represent layer causal GD and orange is used for vanilla GD. *Top*: SimAOU of each layer's update vector. *Bottom*: SimAM_Δ of each layer's update vector.

7 Limitations

- ▷ *Similarity Metrics*: The similarity metrics we use only consider a very specific correspondence between ICL and GD, where each layer applies GD to the model. However, it is possible that the exact mechanism is different (e.g. not all layers do GD).
- ▷ *Datasets*: We use the same datasets used in the original paper by Dai et al. (2023) to make sure we do not introduce factors that benefit our method inadvertently. The dataset collection needs to be diversified. Four out of six datasets are sentiment classification datasets. One of the other tasks, CB, is very small, contributing to instability. Similarly, we consider a specific model in all our experiments. To make a more general claim, other models should be tested too.
- ▷ *LCGD*: We propose a specific instantiation of layer-causal gradient descent. Better instances may exist. While the results for LCGD are (mildly) encouraging, we were unable to rule out the intervention of different secondary effects in score improvement. Despite our best efforts, we suspect such effects might have taken place. One immediate direction for future work is doing hyperparameter search to understand whether there’s an impact of different learning rates on the similarity scores.

8 Related Work

Many works consider synthetic settings (Akyürek et al., 2023; von Oswald et al., 2023a,b; Ahn et al., 2023; Cheng et al., 2024). They are mostly concerned with ICL implementing GD of a shallow model, mostly variants of linear models or kernel regression.

Unlike these works, Dai et al. (2023), which we are heavily influenced by, study large GPT transformers on structured language classification tasks. Gradient Descent in Dai et al. (2023) is with respect to the transformer itself, which is also a significant departure. Panigrahi et al. (2023) show how a transformer can implement the backward pass of another (smaller) transformer in its forward pass. As far as we know, there is no indication that this process is happening in real-world models.

Recently, new works have emerged (Todd et al., 2023; Hendel et al., 2023) suggesting a different approach to interpreting ICL as an algorithm that

compresses training demonstrations into a function/task vector that steers the model to perform the task. Other perspectives of ICL include induction heads (Olsson et al., 2022) and Bayesian inference (Xie et al., 2022).

The work of Shen et al. (2023) points to another discrepancy between *full-batch* GD and ICL. They show that vanilla full-batch GD and ICL cannot be reconciled due to ICL’s sensitivity to the order of the demonstrations, while full-batch GD is invariant to it. However, this discrepancy can be mitigated easily by applying GD sequentially, as was done in the work of Dai et al. (2023) that we compare to.

Layer causal GD is similar to Bengio et al. (2006), where a similar idea was proposed to accelerate training by finding a good starting point using a greedy layer-wise approach.

9 Conclusions

Inspired by recent work, we explore the relationship between in-context learning and gradient descent-based finetuning in practical settings. We show problems with the strong version of the ICL-GD correspondence. We correct the similarity metrics used in prior work and propose alternatives. Furthermore, we show that a simple baseline of untrained models has higher similarity scores compared to trained models. Our work suggests considering the possibility that only a weak version of ICL-GD holds. We rely on layer causality to further justify this view. We study a potential workaround to this problem (LCGD) that does not violate layer causality and get mixed results. The study of LCGD is not comprehensive enough to make a definite statement for or against layer-causal GD mesa-optimizers. We note a potential connection to kernel regression and functional GD, that come up in works on synthetic setups that uphold the weak ICL-GD correspondence. We leave for future work to elucidate the nature of these connections, as well as propose better layer-causal variants.

References

- Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. 2023. [Transformers learn to implement preconditioned gradient descent for in-context learning](#).
- Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. [What learn-](#)

- ing algorithm is in-context learning? investigations with linear models.
- Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. 2006. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Xiang Cheng, Yuxin Chen, and Suvrit Sra. 2024. Transformers implement functional gradient descent to learn non-linear functions in context.
- Damai Dai, Yuntao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers.
- Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. Analyzing transformers in embedding space.
- Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse.
- Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. 2023. Jump to conclusions: Short-cutting transformers with linear transformations.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2021/framework/index.html>.
- Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space.
- Roe Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors.
- Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2021. Risks from learned optimization in advanced machine learning systems.
- nostalgebraist. 2020. interpreting gpt: the logit lens. <https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, T. J. Henighan, Benjamin Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, John Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom B. Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Christopher Olah. 2022. In-context learning and induction heads. *ArXiv*, abs/2209.11895.
- Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, page 271–es, USA. Association for Computational Linguistics.
- Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, page 115–124, USA. Association for Computational Linguistics.
- Abhishek Panigrahi, Sadhika Malladi, Mengzhou Xia, and Sanjeev Arora. 2023. Trainable transformer in transformer. *ArXiv*, abs/2307.01189.
- Lingfeng Shen, Aayush Mishra, and Daniel Khashabi. 2023. Do pretrained transformers really learn in-context by gradient descent?
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- Surat Teerapittayanon, Bradley McDanel, and H. T. Kung. 2017. Branchynet: Fast inference via early exiting from deep neural networks.
- Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2023. Function vectors in large language models.
- Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023a. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR.

Johannes von Oswald, Eyvind Niklasson, Maximilian Schlegel, Seijin Kobayashi, Nicolas Zucchet, Nino Scherrer, Nolan Miller, Mark Sandler, Blaise Agüera y Arcas, Max Vladymyrov, Razvan Pascanu, and João Sacramento. 2023b. [Uncovering mesa-optimization algorithms in transformers.](#)

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. [Emergent abilities of large language models.](#)

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. [An explanation of in-context learning as implicit bayesian inference.](#)

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. [Character-level convolutional networks for text classification.](#) In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

A Data Statistics

| | # Train | # Validation | Avg. # of Tokens |
|---------------|---------|--------------|------------------|
| SST2 | 67,349 | 1,821 | 55.43 |
| SST5 | 8,544 | 2,210 | 102.95 |
| MR | 8,530 | 1,066 | 113.39 |
| Subj | 8,000 | 2,000 | 129.23 |
| AGNews | 120,000 | 7,600 | 237.72 |
| CB | 250 | 250 | 295.80 |

Table 3: Data statistics of all the datasets in the benchmark

B Deeper Analysis of Layer Causality

B.1 Does Layer Causal Gradient Descent Approximate Gradient Descent?

A natural question that might arise is how similar GD is to the suggested layer causal method. Due to their relatively similar scores, one might conjecture that layer causal GD is a low-resource approximation for GD. We can gauge how similar the two update vectors are to each other using a variant of the attention map metric: $\text{SimAM}_{\Delta}^{\text{GD, LCGD}} = \text{CosSim} \left(m_{\text{LCGD}}^{(l,h)} - m_{\text{ZS}}^{(l,h)}, m_{\text{GD}}^{(l,h)} - m_{\text{ZS}}^{(l,h)} \right)$. This way we can measure how much of the score is attributable to the similarity between the update vectors. We will denote the metric by α .

We take one seed per task and compute the average α over the layers, for each task. Counter to our expectations, Figure 2 shows that for most datasets, $\alpha \approx 0.1 - 0.2$, which is very low. This shows that the updates are *not* very correlated, and most of the score of either of the procedures cannot be attributed to a common direction in space.

B.2 Layerwise Analysis

Until now, the metrics reported are averaged across all layers. However, it is interesting to look at similarity patterns across layers. In Figure 3, we show the SimAOU and SimAM $_{\Delta}$ scores averaged across all tasks and seeds for each layer. Interesting patterns emerge in the plots. First, we notice that LCGD outperforms vanilla GD in terms of SimAOU (except for layers 1, 3, and the last layer). In the second plot, we have a more complicated case. In the first half of the model, SimAM $_{\Delta}$ is greater for the causal variant (except for layer 9). However, for all layers 12-17, vanilla GD is substantially greater than layer causal. Beginning from layer 18, both scores decrease more or less together.

With this discrepancy between the metrics, it is worth discussing their different roles. SimAOU captures the similarity to ICL’s *hidden states*. They have a direct effect on the model’s prediction. Attention logits on the other hand only modulate the coefficients that determine the hidden state. The hidden state mediates their interactions with the rest of the model. They have no direct effect on the prediction, conditioned on the hidden state. On the other hand, attention maps can provide us insight into the way attention has shifted as a response to the parameter update. The higher this metric is, the better it replicates the way ICL attends to its input. While not directly affecting the output, it focuses on what “interests” ICL.

Finally, it is important to remember that our GD variant was selected intentionally due to its simplicity. Mild modifications might make it a better contender. Moreover, the setting we consider is limited to the one chosen by Dai et al. (2023), including reusing the same hyperparameters for both methods. It is possible that tuning the hyperparameters for our variant would have yielded better results. All in all, we can state rather confidently that even this simple baseline *performs on par with vanilla GD* across multiple benchmarks, and in some cases outperforms it. Furthermore, it has appealing features, such as being low resource, simple, and causally plausible.

C Gradient Norm in LCGD

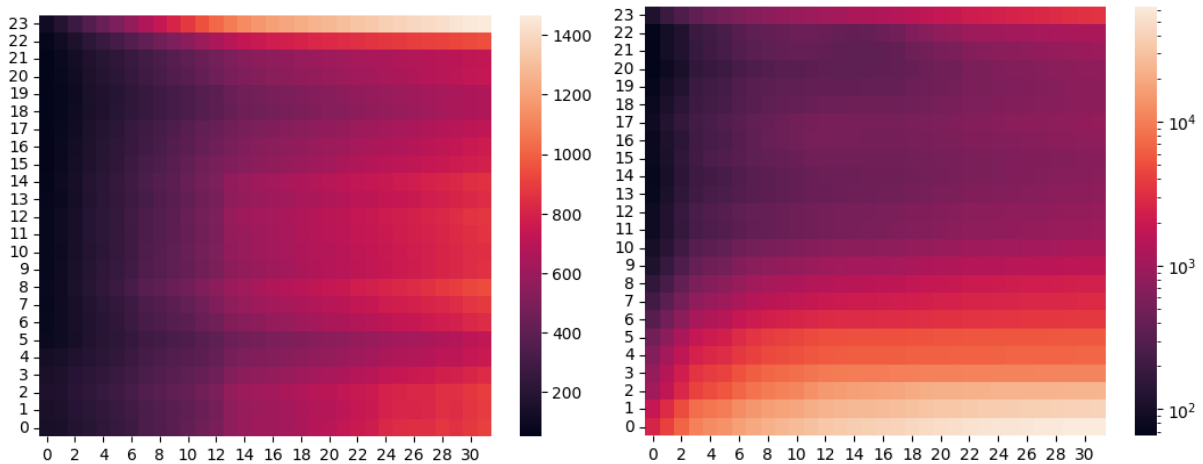


Figure 4: Heatmap of ℓ_2 norms of the gradients computed during finetuning on the **Subj** task. Note the different scales of magnitude. **Horizontal Axis**: Training demonstration index. **Vertical Axis**: Layer index in ascending order (from input to network output). **Left**: Vanilla GD. **Right**: LCGD (norm magnitude in *logarithmic scale*).

D Overview of Select Works in the Synthetic Line of Work

- [von Oswald et al. \(2023a\)](#) study linear transformers with data of the form $f_\theta(\mathbf{x}) = \theta^\top \mathbf{x}$. They found a variant of GD (w.r.t. f_θ) that they called GD^{++} that seems to be implemented by ICL.
- [Ahn et al. \(2023\)](#) discuss the same linear data scenario. They conclude the optimality of a preconditioned variant of GD/ GD^{++} under different assumptions.
- [von Oswald et al. \(2023b\)](#) study auto-regressive linear transformers. The function under consideration adds stochasticity to the model: $f_W(\mathbf{x}) = W\mathbf{x} + \epsilon$ with W being a matrix instead of a vector, and the input of each demonstration being the previous demonstration. They uncover an intriguing algorithm performed by the transformer, combining preconditioning and GD.
- [Cheng et al. \(2024\)](#) discuss transformers with non-linear attention of the form $\mathcal{K}(\mathbf{u}, \mathbf{v})$ where \mathcal{K} is a kernel function. The data in their case comes from a generalized Gaussian process. They consider the empirical quadratic loss objective:

$$\mathcal{L}(f) = \sum_{i=1}^N (f_\theta(\mathbf{x}_i) - y_i)^2$$

This objective function is more complicated than in other cases described here, as f_θ is no longer linear. However, they show optimality of gradient descent in *function space*, which turns out to take on a simple form: $\nabla_f \mathcal{L}(f) = \sum_{i=1}^N (y_i - f(\mathbf{x}_i)) \mathcal{K}(\mathbf{x}_i, \cdot)$. This is in line with the intuition that detached forms of GD are the ones that we should consider, the same intuition as in the construction of layer-causal GD.