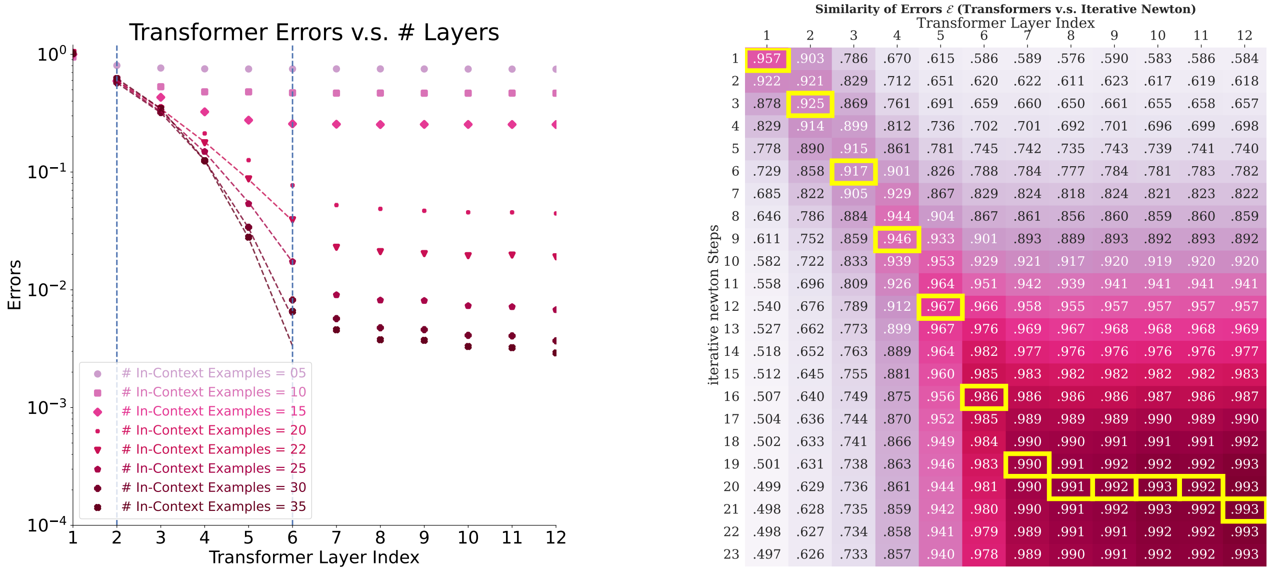
NeurIPS 2024 Submission #1395 Rebuttal – New Experiments

1 Noisy Linear Regression

We repeat the same experiments on noisy linear regression tasks with $y = \boldsymbol{w}^\top \boldsymbol{x} + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with noise level $\sigma = 0.1$. As shown in Figure 1a, Transformers still show superlinear convergence on noisy linear regression tasks. Since the predictor is $\hat{\boldsymbol{w}} = \left(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}\right)^\dagger \boldsymbol{X}^\top \boldsymbol{y}$ for some λ, the iterative newton's method is applied to $\boldsymbol{S} = \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}$. Iterative Newton's method still keeps the same superlinear convergence rates. As shown in Figure 1b, Transformers and Iternative Newton's rates match linearly, as in the noiseless linear regression tasks.



	1	2	3	4	5	6	7	8	9	10	11	12
1	.957	.903	.786	.670	.615	.586	.589	.576	.590	.583	.586	.584
2	.922	.921	.829	.712	.651	.620	.622	.611	.623	.617	.619	.618
3	.878	.925	.869	.761	.691	.659	.660	.650	.661	.655	.658	.657
4	.829	.914	.899	.812	.736	.702	.701	.692	.701	.696	.699	.698
5	.778	.890	.915	.861	.781	.745	.742	.735	.743	.739	.741	.740
6	.729	.858	.917	.901	.826	.788	.784	.777	.784	.781	.783	.782
7	.685	.822	.905	.929	.867	.829	.824	.818	.824	.821	.823	.822
8	.646	.786	.884	.944	.904	.867	.861	.856	.860	.859	.860	.859
9	.611	.752	.859	.946	.933	.901	.893	.889	.893	.892	.893	.892
10	.582	.722	.833	.939	.953	.929	.921	.917	.920	.919	.920	.920
11	.558	.696	.809	.926	.964	.951	.942	.939	.941	.941	.941	.941
12	.540	.676	.789	.912	.967	.966	.958	.955	.957	.957	.957	.957
13	.527	.662	.773	.899	.967	.976	.969	.967	.968	.968	.968	.969
14	.518	.652	.763	.889	.964	.982	.977	.976	.976	.976	.976	.977
15	.512	.645	.755	.881	.960	.985	.983	.982	.982	.982	.982	.983
16	.507	.640	.749	.875	.956	.986	.986	.986	.986	.987	.986	.987
17	.504	.636	.744	.870	.952	.985	.989	.989	.989	.990	.989	.990
18	.502	.633	.741	.866	.949	.984	.990	.990	.991	.991	.991	.992
19	.501	.631	.738	.863	.946	.983	.990	.991	.992	.992	.992	.993
20	.499	.629	.736	.861	.944	.981	.990	.991	.992	.993	.992	.993
21	.498	.628	.735	.859	.942	.980	.990	.991	.992	.993	.992	.993
22	.498	.627	.734	.858	.941	.979	.989	.991	.991	.992	.992	.993
23	.497	.626	.733	.857	.940	.978	.989	.990	.991	.992	.992	.993

(a) Transformers have superlinear convergence rate

(b) Transformers match Iterative Newton's rate

Figure 1: Empirical Results on Noisy Linear Regression

2 Effects of Hidden Dimension

We first restate the main theorem here, with text highlighted in red related to the hidden dimension.

Theorem 5.1. *There exist Transformer weights such that on any set of in-context examples* $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ *and test point* $\boldsymbol{x}_{\text{test}}$*, the Transformer predicts on* $\boldsymbol{x}_{\text{test}}$ *using* $\boldsymbol{x}_{\text{test}}^\top \hat{\boldsymbol{w}}_k^{\text{Newton}}$*. Here* $\hat{\boldsymbol{w}}_k^{\text{Newton}}$ *are the Iterative Newton updates given by* $\hat{\boldsymbol{w}}_k^{\text{Newton}} = \boldsymbol{M}_k \boldsymbol{X}^\top \boldsymbol{y}$ *where* $\boldsymbol{M}_j$ *is updated as* $\boldsymbol{M}_j = 2\boldsymbol{M}_{j-1} - \boldsymbol{M}_{j-1}\boldsymbol{S}\boldsymbol{M}_{j-1}, 1 \le j \le k, \boldsymbol{M}_0 = \alpha \boldsymbol{S}$*, for some* $\alpha > 0$ *and* $\boldsymbol{S} = \boldsymbol{X}^\top \boldsymbol{X}$*. The dimensionality of the hidden layers is* $\mathcal{O}(d)$*, and the number of layers is* $k + 8$*.*

Now we ablate 12-layer 1-head Transformers with various hidden sizes on $d = 20$ problems. As shown in Figure 2, we observe that Transformers can mimic OLS solution when the hidden size is 32 or 64, but fail with smaller sizes. This resonates with our theoretical results on $\mathcal{O}(d)$ hidden dimension, and in this case, Theorem 5.1 ensures a construction of transformers to implement Iternative Newton's method.

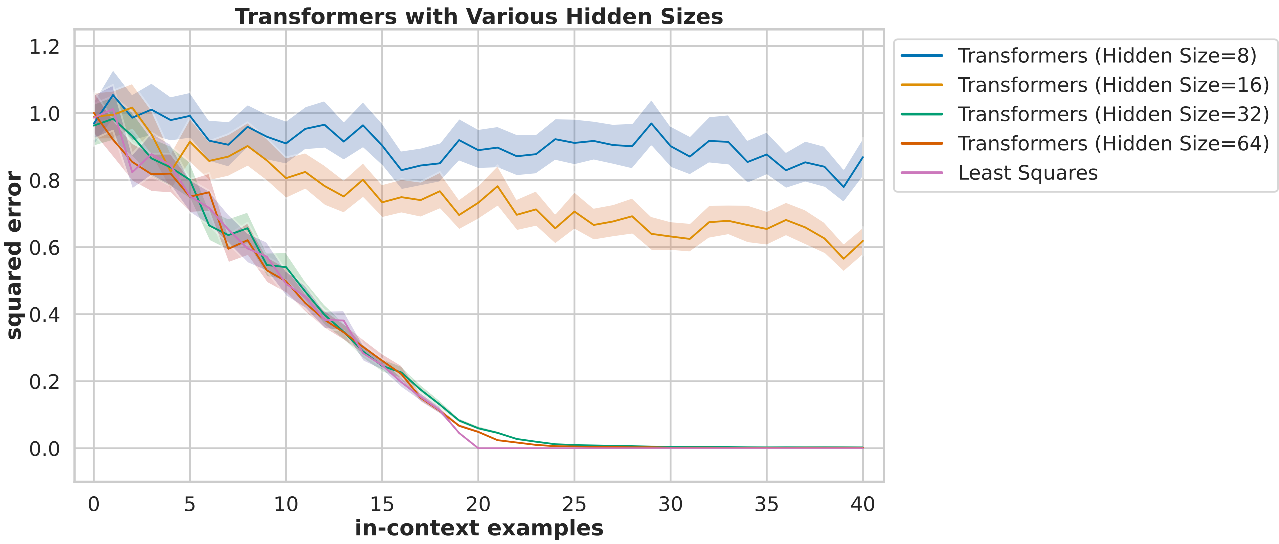


Figure 2: Ablation on Transformer's Hidden Size