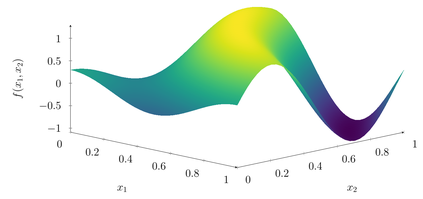
Optimization, Generalization and Differential Privacy Bounds for Gradient Descent on Kolmogorov–Arnold Networks

Puyu Wang [1], Junyu Zhou [2], Philipp Liznerski [1] and Marius Kloft [1]

[1] RPTU Kaiserslautern-Landau, Germany [2] Catholic University of Eichstätt-Ingolstadt, Germany

Kolmogorov–Arnold Representation: Intuition

Question. Can such a multivariate function be represented *exactly* as a finite sum of compositions of *univariate* functions and addition?



Example on $[0,1]^2$: $f(x_1,x_2)=\sin(2\pi x_1 x_2)+0.3\cos(2\pi(x_1+x_2))$.

At first glance: this surface does not look like a "sum of 1D pieces".

However, Kolmogorov–Arnold representation theorem says *yes*, even though the structure is not evident !

Kolmogorov–Arnold representation theorem (Informal statement). For any continuous function $f:[0,1]^d\to\mathbb{R}$, there exist continuous *univariate* functions $\phi_{q,p}:\mathbb{R}\to\mathbb{R}$ and $\Phi_q:\mathbb{R}\to\mathbb{R}$ such that

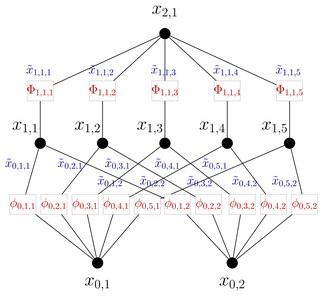
$$f(x_1,\dots,x_d)=\sum_{q=1}^{2d+1}\Phi_q\Big(\sum_{p=1}^{d}\phi_{q,p}(x_p)\Big).$$

In other words, any continuous multivariate function can be represented exactly as a finite sum of compositions of univariate functions and addition.

Special case ($d=2$). There exist univariate functions $\phi_{q,1},\phi_{q,2}$ and Φ_q such that

$$f(x_1,x_2)=\sum_{q=1}^{5}\Phi_q\big(\phi_{q,1}(x_1)+\phi_{q,2}(x_2)\big).$$

KAN Architecture & Standard Formulation



KAN idea: parameterize the Kolmogorov–Arnold representation using learnable univariate edge functions.

For $d=2$, a width $2d+1=5$ construction takes the form

$$f(x_{01},x_{02})=\sum_{q=1}^{5}\Phi_{1,1,q}\big(\phi_{0,q,1}(x_{01})+\phi_{0,q,2}(x_{02})\big).$$

Each hidden node corresponds to one index q, and each edge carries a univariate function.

Standard KAN layer. Let $\mathbf{x}_0=\mathbf{x}\in\mathbb{R}^{n_0}$. For each layer $\ell=0,\dots,L-1$, a KAN maps $\mathbf{x}_\ell\in\mathbb{R}^{n_\ell}$ to $\mathbf{x}_{\ell+1}\in\mathbb{R}^{n_{\ell+1}}$ by

$$x_{\ell+1,j}=\sum_{i=1}^{n_\ell}\phi_{\ell,j,i}(x_{\ell,i}),\qquad j=1,\dots,n_{\ell+1}.$$

Here each edge carries a learnable univariate function $\phi_{\ell,j,i}:\mathbb{R}\to\mathbb{R}$.

Why Theory for KANs?

KANs are empirically promising, but theoretical guarantees for training dynamics, generalization, and privacy remain limited.

We study GD/DP-GD for KANs:

- **Optimization:** when does GD reduce the training loss?
- **Generalization:** when do GD iterates perform well on test data?
- **Differential privacy:** what utility can private GD guarantee?

Key scaling question: How do m, T, and n trade off?

Learning Setup for Two-layer KANs

Two-layer KAN with B-splines. For $\mathbf{x}_0=\mathbf{x}\in\mathbb{R}^d$ and hidden width m,

$$f_\Theta(\mathbf{x})=\frac{1}{\sqrt{m}}\sum_{j=1}^{m}\sum_{k=1}^{p}c_{j,k}\,b_k(x_{1,j}),$$

where

$$x_{1,j}=\sigma\Big(\frac{1}{\sqrt{d}}\sum_{i=1}^{d}\sum_{k=1}^{p}a_{i,j,k}\,b_k(x_{0,i})\Big),\qquad j\in[m].$$

Here $\Theta=(\mathbf{a},\mathbf{c})$, and σ is bounded, Lipschitz, and smooth.

Risks and training. For a nonnegative convex, smooth, self-bounded loss ℓ,

$$\mathcal{L}_S(\Theta)=\frac{1}{n}\sum_{i=1}^{n}\ell\left(y_i f_\Theta(\mathbf{x}_i)\right),\qquad \mathcal{L}(\Theta)=\mathbb{E}_{(\mathbf{x},y)\sim P}\left[\ell\left(y f_\Theta(\mathbf{x})\right)\right].$$

We train by gradient descent (GD):

$$\Theta^{(t+1)}=\Theta^{(t)}-\eta\nabla\mathcal{L}_S(\Theta^{(t)}).$$

We prove upper bounds on optimization risk, population risk, and DP utility, leading to scaling insights for width m, iterations T, and sample size n.

Main Results for GD

NTK separability. Assume the data are linearly separable with margin $\gamma>0$ in the tangent feature space at initialization, i.e., there exists a unit vector Θ_0,

$$y_i\left\langle\nabla_\Theta f_{\Theta^{(0)}}(\mathbf{x}_i),\Theta_0\right\rangle\ge\gamma,\qquad i\in[n].$$

Optimization. If $m\ge\mathrm{polylog}(n)$, then with high probability over initialization, GD achieves

$$\mathcal{L}_S(\Theta^{(T)})=\widetilde{O}\Big(\frac{1}{\gamma^2\eta T}\Big).$$

Generalization. If $m\ge\mathrm{polylog}(n)$ and $T\gtrsim n$, then

$$\mathbb{E}_S\left[\mathcal{L}(\Theta^{(T)})\right]=\widetilde{O}\Big(\frac{1}{\gamma^4 n}\Big).$$

Polylogarithmic width already suffices
for $\widetilde{O}(1/T)$ optimization and fast $\widetilde{O}(1/n)$ population risk.

Differentially Private KANs

Differential privacy. For neighboring datasets S,S' differing in one example, an algorithm A is (ϵ,δ)-DP if for any event E,

$$\mathbb{P}(\mathsf{A}(S)\in E)\le e^\epsilon\,\mathbb{P}(\mathsf{A}(S')\in E)+\delta.$$

DP-GD. To achieve (ϵ,δ)-DP, we add Gaussian noise to the gradient and project the iterates back to a neighborhood of initialization:

$$\widetilde{\Theta}^{(k+1)}=\Pi_\Omega\Big(\widetilde{\Theta}^{(k)}-\eta\big(\nabla\mathcal{L}_S(\widetilde{\Theta}^{(k)})+\mathbf{b}^{(k)}\big)\Big),\qquad \mathbf{b}^{(k)}\sim\mathcal{N}(0,\sigma^2 I).$$

Here σ^2 is calibrated according to n,T,ϵ,δ.

Private utility. If $m\asymp\mathrm{polylog}(n)$ and $\eta T\asymp\frac{\gamma^2 n\epsilon}{\sqrt{d}}$, then the (ϵ,δ)-DP variant satisfies

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}_{S,\mathsf{A}}\left[\mathcal{L}(\widetilde{\Theta}^{(k)})\right]=\widetilde{O}\Big(\frac{\sqrt{d}}{\gamma^4 n\epsilon}\Big).$$

- **Polylogarithmic width is necessary** to achieve $\widetilde{O}\left(\frac{\sqrt{d}}{n\epsilon}\right)$ population risk rate
- **DP-GD attains near-optimal private utility** up to logarithmic factors, despite the nonconvex KAN parameterization

Practical Implications & Experiments

Our theory gives explicit guidance for choosing the width m and the number of GD iterations T for both GD and DP-GD.

- **Training and test accuracy saturate beyond a critical width.**
- **Training accuracy may continue to improve with iterations.**
- **Test accuracy saturates with increasing iterations.**
- **Private utility is maximized in an admissible regime.**

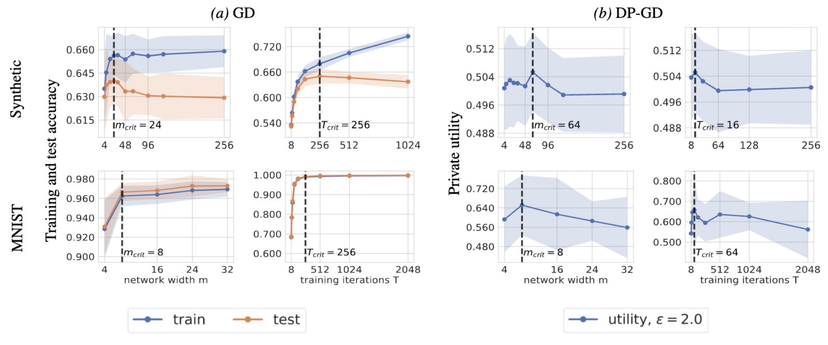


Figure. Effect of width m and iterations T on GD accuracy and DP-GD utility. Dashed lines indicate empirical change points: accuracy saturates beyond a critical width/training horizon, while DP utility degrades when privacy noise accumulates. These trends match the theory-predicted admissible regimes for choosing m and T.