Numerical CY metrics from holomorphic networks

Michael R. Douglas, Subramanian Lakshminarasimhan and Yidi Qi

Stony Brook / CMSA

Geometria em Lisboa seminar, April 12, 2022

Abstract

We propose machine learning inspired methods for computing numerical Ricci-flat Kähler metrics, and compare them with previous work. `arXiv:2012.04797` and `arXiv:2105.03991`

ST Yau famously proved in 1978 that a Kähler manifold with zero first Chern class has a Ricci flat metric. We still have no closed form expressions for these metrics, but there has been a fair amount of work on computing numerical approximations and using them for physics applications: Headrick and Wiseman `hep-th/0506129`, Donaldson `math.DG/0512625`, Douglas Karp Lukic Reinbacher (DKLR) `hep-th/0606261` and `hep-th/0612075`, Braun, Ovrut *et al* `0712.3563, 0805.3689` , Anderson *et al* `0904.2186, 1004.4399, 1103.3041`, Headrick and Nassar `0908.2635`, Cui and Gray `1912.11068`, and others. Related work on other Kähler-Einstein metrics appears in Doran *et al* `hep-th/0703057`.

The tremendous advances in machine learning (ML) have brought new interest to this line of work. ML uses function approximation methods, especially feed forward networks (FFNs), which mitigate the "curse of dimensionality" in high dimensions. ML software is 1000s of times faster than general purpose numerical code. Many recent works are analyzing and using ML inspired methods for numerical PDE.

ST Yau famously proved in 1978 that a Kähler manifold with zero first Chern class has a Ricci flat metric. We still have no closed form expressions for these metrics, but there has been a fair amount of work on computing numerical approximations and using them for physics applications: Headrick and Wiseman `hep-th/0506129`, Donaldson `math.DG/0512625`, Douglas Karp Lukic Reinbacher (DKLR) `hep-th/0606261` and `hep-th/0612075`, Braun, Ovrut *et al* `0712.3563, 0805.3689` , Anderson *et al* `0904.2186, 1004.4399, 1103.3041`, Headrick and Nassar `0908.2635`, Cui and Gray `1912.11068`, and others. Related work on other Kähler-Einstein metrics appears in Doran *et al* `hep-th/0703057`.

The tremendous advances in machine learning (ML) have brought new interest to this line of work. ML uses function approximation methods, especially feed forward networks (FFNs), which mitigate the "curse of dimensionality" in high dimensions. ML software is 1000s of times faster than general purpose numerical code. Many recent works are analyzing and using ML inspired methods for numerical PDE.

In this talk I describe my work with Lakshminarasimhan and Qi, following a more mathematical presentation arXiv:2105.03991. Related work on ML for numerical methods includes Ashmore, He and Ovrut 1910.08605, Ashmore 2011.13929, and 2012.04656 by Anderson, Gerdes, Gray, Krippendorf, Raghuram and Ruehle.

In this talk I describe my work with Lakshminarasimhan and Qi, following a more mathematical presentation arXiv:2105.03991. Related work on ML for numerical methods includes Ashmore, He and Ovrut 1910.08605, Ashmore 2011.13929, and 2012.04656 by Anderson, Gerdes, Gray, Krippendorf, Raghuram and Ruehle.

Calabi-Yau manifolds: Kähler with $c_1(M) = 0 \leftrightarrow \exists g_{ij}$ with Ricci $= 0$.
Standard constructions: complete intersections in projective space, hypersurfaces in weighted projective space and in toric varieties.
Simplest examples: quintics in $\mathbb{CP}^4$ with $126 - 25$ complex moduli,

$$0 = f(Z^1, Z^2, Z^3, Z^4, Z^5) = \sum_{i=1}^{5} (Z^i)^5 + \psi Z^1 Z^2 Z^3 Z^4 Z^5 + \text{other degree 5.}$$

Keeping the one modulus ψ, we have the Dwork family with generic $\mathbb{Z}^4 \times S_5$ symmetry (and $\mathbb{Z}_2$ for ψ real).

A Calabi-Yau manifold has two preferred volume forms, the metric volume and the "holomorphic volume"

$$\begin{aligned} \text{vol}_\omega &\equiv \frac{1}{n!} \det \omega \qquad \text{where } \omega \equiv \partial\bar{\partial} K,\ n \equiv \dim_{\mathbb{C}} M \\ \text{vol}_\Omega &\equiv (-i)^n \mathcal{N}_\Omega \Omega^{(n,0)} \wedge \bar{\Omega}^{(0,n)} \end{aligned}$$

Their ratio $\eta \equiv \text{vol}_\omega / \text{vol}_\Omega$ is constant for the Ricci flat metric. By choice of normalization $\mathcal{N}_\Omega$ one can set this constant to $\eta = 1$.

Calabi-Yau manifolds: Kähler with $c_1(M) = 0 \leftrightarrow \exists g_{ij}$ with Ricci $= 0$.
Standard constructions: complete intersections in projective space, hypersurfaces in weighted projective space and in toric varieties.
Simplest examples: quintics in $\mathbb{CP}^4$ with $126 - 25$ complex moduli,

$$0 = f(Z^1, Z^2, Z^3, Z^4, Z^5) = \sum_{i=1}^{5} (Z^i)^5 + \psi Z^1 Z^2 Z^3 Z^4 Z^5 + \text{other degree 5.}$$

Keeping the one modulus ψ, we have the Dwork family with generic $\mathbb{Z}^4 \times S_5$ symmetry (and $\mathbb{Z}_2$ for ψ real).

A Calabi-Yau manifold has two preferred volume forms, the metric volume and the "holomorphic volume"

$$\begin{aligned} \text{vol}_\omega &\equiv \frac{1}{n!} \det \omega \qquad \text{where } \omega \equiv \partial\bar{\partial} K,\ n \equiv \dim_{\mathbb{C}} M \\ \text{vol}_\Omega &\equiv (-i)^n \mathcal{N}_\Omega \Omega^{(n,0)} \wedge \bar{\Omega}^{(0,n)} \end{aligned}$$

Their ratio $\eta \equiv \text{vol}_\omega / \text{vol}_\Omega$ is constant for the Ricci flat metric. By choice of normalization $\mathcal{N}_\Omega$ one can set this constant to $\eta = 1$.

So far as anyone knows there is no analytic expression for these Ricci flat metrics, though there may be one for K3 (Kachru Tripathy and Zimet 1810.10540, 2006.02435, Gaiotto *et al* 0907.3987). But numerics could suffice. Substituting $\omega = \omega_0 + \partial\bar{\partial}\phi$ into the $\eta = 1$ equation, one must solve a nonlinear PDE for a single function ϕ (the complex Monge-Ampere equation).

Numerical methods for PDE is a vast subject but there are two main types of method. One is local methods such as finite elements. These discretize space and represent functions in terms of their local values (on points, links, *etc.*).

The other is spectral methods, such as Fourier space methods. Here one expands the functions in an analytically simple basis: polynomials, exponentials, special functions, *etc.*.

All numerical methods are challenging for PDEs with gauge and coordinate freedom – unless carefully designed, small effects build up and lead to numerical instability. In CY metric work, besides the simplification of solving for a single function, these problems are mitigated by working with a fixed complex coordinate system.

So far as anyone knows there is no analytic expression for these Ricci flat metrics, though there may be one for K3 (Kachru Tripathy and Zimet 1810.10540, 2006.02435, Gaiotto *et al* 0907.3987). But numerics could suffice. Substituting $\omega = \omega_0 + \partial\bar{\partial}\phi$ into the $\eta = 1$ equation, one must solve a nonlinear PDE for a single function ϕ (the complex Monge-Ampere equation).

Numerical methods for PDE is a vast subject but there are two main types of method. One is local methods such as finite elements. These discretize space and represent functions in terms of their local values (on points, links, *etc.*).

The other is spectral methods, such as Fourier space methods. Here one expands the functions in an analytically simple basis: polynomials, exponentials, special functions, *etc.*.

All numerical methods are challenging for PDEs with gauge and coordinate freedom – unless carefully designed, small effects build up and lead to numerical instability. In CY metric work, besides the simplification of solving for a single function, these problems are mitigated by working with a fixed complex coordinate system.

So far as anyone knows there is no analytic expression for these Ricci flat metrics, though there may be one for K3 (Kachru Tripathy and Zimet 1810.10540, 2006.02435, Gaiotto *et al* 0907.3987). But numerics could suffice. Substituting $\omega = \omega_0 + \partial\bar{\partial}\phi$ into the $\eta = 1$ equation, one must solve a nonlinear PDE for a single function ϕ (the complex Monge-Ampere equation).

Numerical methods for PDE is a vast subject but there are two main types of method. One is local methods such as finite elements. These discretize space and represent functions in terms of their local values (on points, links, *etc.*).

The other is spectral methods, such as Fourier space methods. Here one expands the functions in an analytically simple basis: polynomials, exponentials, special functions, *etc.*.

All numerical methods are challenging for PDEs with gauge and coordinate freedom – unless carefully designed, small effects build up and lead to numerical instability. In CY metric work, besides the simplification of solving for a single function, these problems are mitigated by working with a fixed complex coordinate system.

So far as anyone knows there is no analytic expression for these Ricci flat metrics, though there may be one for K3 (Kachru Tripathy and Zimet 1810.10540, 2006.02435, Gaiotto *et al* 0907.3987). But numerics could suffice. Substituting $\omega = \omega_0 + \partial\bar{\partial}\phi$ into the $\eta = 1$ equation, one must solve a nonlinear PDE for a single function ϕ (the complex Monge-Ampere equation).

Numerical methods for PDE is a vast subject but there are two main types of method. One is local methods such as finite elements. These discretize space and represent functions in terms of their local values (on points, links, *etc.*).

The other is spectral methods, such as Fourier space methods. Here one expands the functions in an analytically simple basis: polynomials, exponentials, special functions, *etc.*.

All numerical methods are challenging for PDEs with gauge and coordinate freedom – unless carefully designed, small effects build up and lead to numerical instability. In CY metric work, besides the simplification of solving for a single function, these problems are mitigated by working with a fixed complex coordinate system.

Headrick and Wiseman `hep-th/0506129` found a metric on a Kummer surface M as follows. Recall that the Kummer surface is the blowup of $T^4/\mathbb{Z}_2$ at the 16 fixed points. For the maximally symmetric case, one can cover M with two types of patch – $T^4/\mathbb{Z}_2$ minus the neighborhood of the fixed points, and 16 copies of a deformed Eguchi-Hanson space (asymptotically $\mathbb{C}^2/\mathbb{Z}_2$). One then represents K using its values on a finite set of points p in each patch, and its derivatives as finite differences. HW then used a relaxation method (Gauss-Seidel) in which one iterates through points p and solves for $K(p)$ with the values of the neighbors fixed. This converges well.

Local methods suffer from the "curse of dimensionality" – to represent a function on length scales $1/k$ in D dimensions requires $\mathcal{O}(k^D)$ lattice points. 10^8 points would be the practical limit. In most numerical work, $D = 3$ is considered "high dimensional." $D = 4$ is pushing it and at $D = 6$ one cannot describe much local structure.
Also, one needs to find and program explicit coordinate patches and overlaps. While straightforward in principle, this is a lot of work.

Headrick and Wiseman `hep-th/0506129` found a metric on a Kummer surface M as follows. Recall that the Kummer surface is the blowup of $T^4/\mathbb{Z}_2$ at the 16 fixed points. For the maximally symmetric case, one can cover M with two types of patch – $T^4/\mathbb{Z}_2$ minus the neighborhood of the fixed points, and 16 copies of a deformed Eguchi-Hanson space (asymptotically $\mathbb{C}^2/\mathbb{Z}_2$). One then represents K using its values on a finite set of points p in each patch, and its derivatives as finite differences. HW then used a relaxation method (Gauss-Seidel) in which one iterates through points p and solves for $K(p)$ with the values of the neighbors fixed. This converges well.

Local methods suffer from the "curse of dimensionality" – to represent a function on length scales $1/k$ in D dimensions requires $\mathcal{O}(k^D)$ lattice points. 10^8 points would be the practical limit. In most numerical work, $D = 3$ is considered "high dimensional." $D = 4$ is pushing it and at $D = 6$ one cannot describe much local structure.
Also, one needs to find and program explicit coordinate patches and overlaps. While straightforward in principle, this is a lot of work.

Donaldson was inspired by HW's work to develop another method in `math/0512625`, based on the math he was doing (Kähler-Einstein metrics with $c_1 > 0$). This is a spectral method which takes advantage of many special features of the problem.

First, it is an embedding method, meaning that the manifold M is embedded in a simple higher dimensional ambient space. This has the advantage that one can get a large parameterized family of metrics, by varying either the embedding or the metric on the ambient space. Say we have $X : M \to \mathbb{R}^N$, then we could pull back the Euclidean metrics $h_{ij} = \text{const}$ on $\mathbb{R}^N$ to get a family of metrics on M,

$$ds^2 = h_{ij} dX^i dX^j.$$

In general, embeddings have the disadvantage that they require arbitrary choices. Thus they can develop bad behavior and singularities not related to the original problem.

Donaldson was inspired by HW's work to develop another method in `math/0512625`, based on the math he was doing (Kähler-Einstein metrics with $c_1 > 0$). This is a spectral method which takes advantage of many special features of the problem.

First, it is an embedding method, meaning that the manifold M is embedded in a simple higher dimensional ambient space. This has the advantage that one can get a large parameterized family of metrics, by varying either the embedding or the metric on the ambient space. Say we have $X : M \to \mathbb{R}^N$, then we could pull back the Euclidean metrics $h_{ij} = \text{const}$ on $\mathbb{R}^N$ to get a family of metrics on M,

$$ds^2 = h_{ij} dX^i dX^j.$$

In general, embeddings have the disadvantage that they require arbitrary choices. Thus they can develop bad behavior and singularities not related to the original problem.

But in this problem, there is a canonical family of embeddings into $\mathbb{CP}^N$. We defined our quintic as a hypersurface in $\mathbb{CP}^4$, so this is one embedding. In the terms above, the coordinates Z^i are functions from M to $\mathbb{CP}^4$, the quotient of $\mathbb{C}^5$ by $Z^i \sim \lambda Z^i$. The family of metrics is then defined by pulling back the Fubini-Study metrics

$$K_h = \log \sum_{I,\bar{J}} h_{I,\bar{J}} Z^I \bar{Z}^{\bar{J}}.$$

This only gives us 25 real parameters. But we can generalize to higher N by replacing the Z's in this ansatz with homogeneous polynomials of degree k, $S^{IJ} \equiv Z^I Z^J$, and so on. After removing redundancies following from $f(Z) = 0$, this gives us $\mathcal{O}(k^6)$ real parameters and can approximate arbitrary Kähler metrics on M to arbitrary precision.

All this can be applied to general projective manifolds by reinterpreting the polynomials as a basis of sections s^I of a line bundle $\mathcal{L}^k$. If M has symmetries, one can impose these on $h_{I,\bar{J}}$. And since this is a canonical embedding, it is less likely to introduce bad behavior.

But in this problem, there is a canonical family of embeddings into $\mathbb{CP}^N$. We defined our quintic as a hypersurface in $\mathbb{CP}^4$, so this is one embedding. In the terms above, the coordinates Z^i are functions from M to $\mathbb{CP}^4$, the quotient of $\mathbb{C}^5$ by $Z^i \sim \lambda Z^i$. The family of metrics is then defined by pulling back the Fubini-Study metrics

$$K_h = \log \sum_{I,\bar{J}} h_{I,\bar{J}} Z^I \bar{Z}^{\bar{J}}.$$

This only gives us 25 real parameters. But we can generalize to higher N by replacing the Z's in this ansatz with homogeneous polynomials of degree k, $S^{IJ} \equiv Z^I Z^J$, and so on. After removing redundancies following from $f(Z) = 0$, this gives us $\mathcal{O}(k^6)$ real parameters and can approximate arbitrary Kähler metrics on M to arbitrary precision.

All this can be applied to general projective manifolds by reinterpreting the polynomials as a basis of sections s^I of a line bundle $\mathcal{L}^k$. If M has symmetries, one can impose these on $h_{I,\bar{J}}$. And since this is a canonical embedding, it is less likely to introduce bad behavior.

But in this problem, there is a canonical family of embeddings into $\mathbb{CP}^N$. We defined our quintic as a hypersurface in $\mathbb{CP}^4$, so this is one embedding. In the terms above, the coordinates Z^i are functions from M to $\mathbb{CP}^4$, the quotient of $\mathbb{C}^5$ by $Z^i \sim \lambda Z^i$. The family of metrics is then defined by pulling back the Fubini-Study metrics

$$K_h = \log \sum_{I,\bar{J}} h_{I,\bar{J}} Z^I \bar{Z}^{\bar{J}}.$$

This only gives us 25 real parameters. But we can generalize to higher N by replacing the Z's in this ansatz with homogeneous polynomials of degree k, $S^{IJ} \equiv Z^I Z^J$, and so on. After removing redundancies following from $f(Z) = 0$, this gives us $\mathcal{O}(k^6)$ real parameters and can approximate arbitrary Kähler metrics on M to arbitrary precision.

All this can be applied to general projective manifolds by reinterpreting the polynomials as a basis of sections s^I of a line bundle $\mathcal{L}^k$. If M has symmetries, one can impose these on $h_{I,\bar{J}}$. And since this is a canonical embedding, it is less likely to introduce bad behavior.

A second idea introduced by Donaldson was to approximate the Ricci flat metric by a different metric, the balanced metric. This is defined by the following, at first strange sounding prescription:

$$(h^{-1})^{I,\bar{J}} = \frac{1}{\text{vol } M} \int_M d\text{vol} \frac{S^I \bar{S}^{\bar{J}}}{\sum_{K,\bar{L}} h_{K,\bar{L}} S^K \bar{S}^{\bar{L}}}.$$

What does this mean? The integral defines a hermitian inner product on the space of sections, in other words on $\mathbb{C}^{N+1}$, and we can use it to define an orthonormal basis. The balanced metric is the one for which h is the identity matrix in an orthonormal basis.

This has some physical resonance as it is also the statement that the density of states ρ for a quantum Hall system on M in the magnetic field $F = \omega$ is constant, $\rho(z) = \sum ||S(z)||_h^2 = N + 1$. Klevtsov and I 0811.0367 and Eager, Gary and Roberts 1011.5231 tried to connect this to supersymmetric black holes and giant gravitons in AdS.

A second idea introduced by Donaldson was to approximate the Ricci flat metric by a different metric, the balanced metric. This is defined by the following, at first strange sounding prescription:

$$(h^{-1})^{I,\bar{J}} = \frac{1}{\text{vol } M} \int_M d\text{vol} \frac{S^I \bar{S}^{\bar{J}}}{\sum_{K,\bar{L}} h_{K,\bar{L}} S^K \bar{S}^{\bar{L}}}.$$

What does this mean? The integral defines a hermitian inner product on the space of sections, in other words on $\mathbb{C}^{N+1}$, and we can use it to define an orthonormal basis. The balanced metric is the one for which h is the identity matrix in an orthonormal basis.

This has some physical resonance as it is also the statement that the density of states ρ for a quantum Hall system on M in the magnetic field $F = \omega$ is constant, $\rho(z) = \sum ||S(z)||_h^2 = N + 1$. Klevtsov and I `0811.0367` and Eager, Gary and Roberts `1011.5231` tried to connect this to supersymmetric black holes and giant gravitons in AdS.

A primary motivation for the balanced metric (as I understand it) is that the problem of showing that it exists and is unique, is much simpler than understanding the best approximation to a Ricci flat, Kähler-Einstein or constant scalar curvature metric within a finite dimensional space of metrics. One can then show, using the Tian-Yau-Zelditch expansion

$$\rho(z) = N + 1 + R(z) + \mathcal{O}(N^{-1}),$$

that as $k \to \infty$, the balanced metrics converge to these other metrics.

A systematic way to formulate these problems is to postulate an energy (or "loss") functional $\mathcal{L}$ on the full space of metrics, whose minima are the metrics of interest. If $\mathcal{L}$ is convex, $\partial^2 \mathcal{L} > 0$, then it will either have a unique minimum or run away to infinity. The latter can be excluded by looking at one parameter restrictions (stability).
The balanced metrics are critical points of a convex functional of h,

$$\mathcal{L}_b = \frac{1}{\text{vol } M} \int_M d\text{vol}\, K_h - \log \det h,$$

A primary motivation for the balanced metric (as I understand it) is that the problem of showing that it exists and is unique, is much simpler than understanding the best approximation to a Ricci flat, Kähler-Einstein or constant scalar curvature metric within a finite dimensional space of metrics. One can then show, using the Tian-Yau-Zelditch expansion

$$\rho(z) = N + 1 + R(z) + \mathcal{O}(N^{-1}),$$

that as $k \to \infty$, the balanced metrics converge to these other metrics.

A systematic way to formulate these problems is to postulate an energy (or "loss") functional $\mathcal{L}$ on the full space of metrics, whose minima are the metrics of interest. If $\mathcal{L}$ is convex, $\partial^2 \mathcal{L} > 0$, then it will either have a unique minimum or run away to infinity. The latter can be excluded by looking at one parameter restrictions (stability).

The balanced metrics are critical points of a convex functional of h,

$$\mathcal{L}_b = \frac{1}{\text{vol } M} \int_M d\text{vol}\, K_h - \log \det h,$$

A primary motivation for the balanced metric (as I understand it) is that the problem of showing that it exists and is unique, is much simpler than understanding the best approximation to a Ricci flat, Kähler-Einstein or constant scalar curvature metric within a finite dimensional space of metrics. One can then show, using the Tian-Yau-Zelditch expansion

$$\rho(z) = N + 1 + R(z) + \mathcal{O}(N^{-1}),$$

that as $k \to \infty$, the balanced metrics converge to these other metrics.

A systematic way to formulate these problems is to postulate an energy (or "loss") functional $\mathcal{L}$ on the full space of metrics, whose minima are the metrics of interest. If $\mathcal{L}$ is convex, $\partial^2 \mathcal{L} > 0$, then it will either have a unique minimum or run away to infinity. The latter can be excluded by looking at one parameter restrictions (stability).
The balanced metrics are critical points of a convex functional of h,

$$\mathcal{L}_b = \frac{1}{\text{vol } M} \int_M d\text{vol}\, K_h - \log \det h,$$

Donaldson showed that the balanced definition we discussed earlier could be used to define an iteration,

$$h^{(n)}_{I,\bar{J}} \to h^{(n+1)}_{I,\bar{J}} \equiv \left(\int \frac{S^I \bar{S}^{\bar{J}}}{h^{(n)} S \bar{S}} \right)^{-1},$$

which converges on the balanced metric when it exists. As a numerical method, guaranteed convergence is a great advantage, although doing $(N+1)^2$ integrals is expensive.

Donaldson implemented this procedure on a K3 defined as a hypersurface in $\mathcal{O}_{\mathbb{CP}^2}(3)$, again doing the integrals by choosing explicit coordinate charts and a lattice discretization. He reports computations up to $k = 9$ done on his PC which produced a balanced metric with $|\eta - 1| \sim 1\% \sim 1/k^2$, as predicted by the TYZ argument.
There are many other ideas in this work, for example an approximation to the scalar Laplacian using h and not the explicit metric on M.

Donaldson showed that the balanced definition we discussed earlier could be used to define an iteration,

$$h^{(n)}_{I,\bar{J}} \to h^{(n+1)}_{I,\bar{J}} \equiv \left(\int \frac{S^I \bar{S}^{\bar{J}}}{h^{(n)} S \bar{S}} \right)^{-1},$$

which converges on the balanced metric when it exists. As a numerical method, guaranteed convergence is a great advantage, although doing $(N+1)^2$ integrals is expensive.

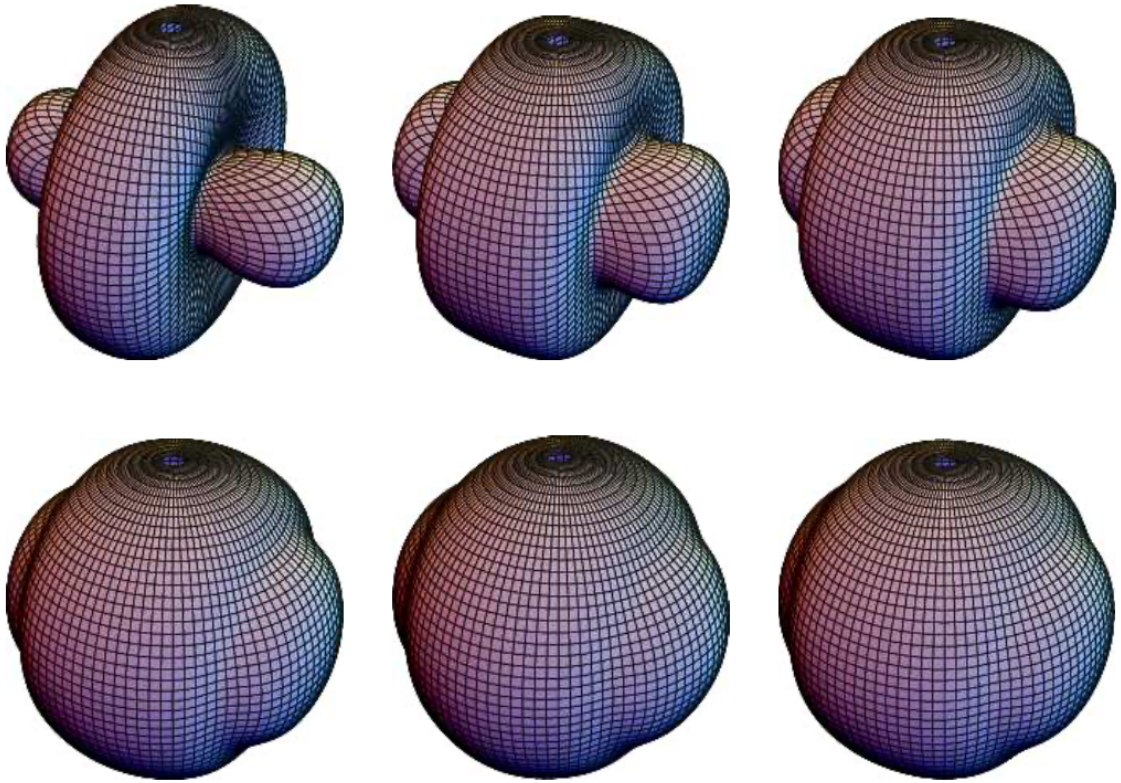
Donaldson implemented this procedure on a K3 defined as a hypersurface in $\mathcal{O}_{\mathbb{CP}^2}(3)$, again doing the integrals by choosing explicit coordinate charts and a lattice discretization. He reports computations up to $k = 9$ done on his PC which produced a balanced metric with $|\eta - 1| \sim 1\% \sim 1/k^2$, as predicted by the TYZ argument.
There are many other ideas in this work, for example an approximation to the scalar Laplacian using h and not the explicit metric on M.

Subsequent works:

- DKLR `hep-th/0612075` followed Donaldson but replaced the integration over coordinate patches with Monte Carlo,

$$\int_M d^D z\, f(z) \to \frac{1}{N} \sum_i f(z_i).$$

  This is easy to program and allowed treating the quintic threefold.
- `hep-th/0606261` treated hermitian Yang-MIlls, using a generalization of the balanced condition due to X Wang (2005). Mathematically, one embeds the vector bundle V twisted by $\mathcal{L}$ in a Grassmannian manifold. More concretely, $A = G^{-1}\partial G$ with $G = hS\bar{S}$. Donaldson's algorithm converges if V is stable.
- Braun, Ovrut *et al* `0712.3563` treated general complete intersections and quotients. In `0805.3689` the scalar Laplacian was studied.
- Anderson *et al* `1004.4399, 1103.3041` studied HYM in detail and reproduced variations of stability.



The rational curve $(Z_0 = z_0,\, Z_1 = -z_0,\, Z_2 = z_1,\, Z_3 = 0,\, Z_4 = -z_1)$ for the balanced metric on the Fermat quintic, from DKLR.

Headrick and Nassar 0908.2635 returned to the Ricci flat metric problem, but instead of using Donaldson's algorithm to find the balanced metric, they proposed to minimize an energy function which measures the distance of the metric to Ricci flatness. Since a Calabi-Yau metric satisfies $\eta = 1$, there are many candidates:

$$\begin{aligned} \mathcal{L}_{\eta,p} &\equiv ||\eta - 1||_p \qquad p = 1 \text{ (MAPE)}, p = 2 \text{ (RMSE)}, p = \infty \text{ (MAX)} \\ \mathcal{L}_R &\equiv ||\partial \log \eta||_2 \qquad \text{(related to the mean Ricci scalar)} \end{aligned}$$

where $||f||_p$ is the p-norm $\left(\int_M d\mu\, |f|^p \right)^{1/p}$ for some measure $d\mu$. Another interesting choice is the KL divergence $\int_M \mathrm{vol}_\Omega \log \eta$.

These $\mathcal{L}$'s are convex when considered as functionals on the space of Kähler metrics and after restricting to the Fubini-Study metrics, as long as we do the continuum integral over M. It is less clear if we do Monte Carlo, but should hold if the number of sample points is much greater than the number of parameters. HN carried this procedure out for the Fermat K3 and Dwork quintics. Their large discrete symmetry groups enable going to large k with relatively few parameters, about 100 for Fermat K3 with $k = 17$, achieving very high accuracy (MSE $\sim 10^{-16}$).

Headrick and Nassar 0908.2635 returned to the Ricci flat metric problem, but instead of using Donaldson's algorithm to find the balanced metric, they proposed to minimize an energy function which measures the distance of the metric to Ricci flatness. Since a Calabi-Yau metric satisfies $\eta = 1$, there are many candidates:

$$\begin{aligned} \mathcal{L}_{\eta,p} &\equiv ||\eta - 1||_p \qquad p = 1 \text{ (MAPE)}, p = 2 \text{ (RMSE)}, p = \infty \text{ (MAX)} \\ \mathcal{L}_R &\equiv ||\partial \log \eta||_2 \qquad \text{(related to the mean Ricci scalar)} \end{aligned}$$

where $||f||_p$ is the p-norm $\left(\int_M d\mu \, |f|^p \right)^{1/p}$ for some measure $d\mu$. Another interesting choice is the KL divergence $\int_M \text{vol}_\Omega \log \eta$. These $\mathcal{L}$'s are convex when considered as functionals on the space of Kähler metrics and after restricting to the Fubini-Study metrics, as long as we do the continuum integral over M. It is less clear if we do Monte Carlo, but should hold if the number of sample points is much greater than the number of parameters. HN carried this procedure out for the Fermat K3 and Dwork quintics. Their large discrete symmetry groups enable going to large k with relatively few parameters, about 100 for Fermat K3 with $k = 17$, achieving very high accuracy (MSE $\sim 10^{-16}$).

If one does not know whether a solution exists, or wants to prove this rigorously, then the mathematical properties of the balanced metric are a great advantage. But if one knows the solution exists and wants the most accurate approximation to it, then combining the embedding method with optimization is the more straightforward approach.

One can show that the approximation error decreases exponentially in the order k of the polynomials. The Ricci flat metric is known to be analytic (C^∞) and thus its coefficients in Fourier space fall off exponentially. The same is true for this basis (it is a spectral basis for the Laplacian on $\mathbb{CP}^{n+1}$).

Also like a Fourier basis, one expects that k'th order polynomials can represent structure on length scales down to $1/k$, but not on shorter scales. In the CY problem one can vary the length scales by varying the complex structure – for a hypersurface, the defining function f. For example by tuning $\psi \to -5$ above, one approaches a conifold (ODP) singularity. In this limit, a three-cycle becomes small and the accuracy becomes low, as found by HN and by Cui and Gray. 10 12 11 68

If one does not know whether a solution exists, or wants to prove this rigorously, then the mathematical properties of the balanced metric are a great advantage. But if one knows the solution exists and wants the most accurate approximation to it, then combining the embedding method with optimization is the more straightforward approach.

One can show that the approximation error decreases exponentially in the order k of the polynomials. The Ricci flat metric is known to be analytic (C^∞) and thus its coefficients in Fourier space fall off exponentially. The same is true for this basis (it is a spectral basis for the Laplacian on $\mathbb{CP}^{n+1}$).

Also like a Fourier basis, one expects that k'th order polynomials can represent structure on length scales down to $1/k$, but not on shorter scales. In the CY problem one can vary the length scales by varying the complex structure – for a hypersurface, the defining function f. For example by tuning $\psi \to -5$ above, one approaches a conifold (ODP) singularity. In this limit, a three-cycle becomes small and the accuracy becomes low, as found by HN and by Cui and Gray. 1912.11068

If one does not know whether a solution exists, or wants to prove this rigorously, then the mathematical properties of the balanced metric are a great advantage. But if one knows the solution exists and wants the most accurate approximation to it, then combining the embedding method with optimization is the more straightforward approach.

One can show that the approximation error decreases exponentially in the order k of the polynomials. The Ricci flat metric is known to be analytic (C^∞) and thus its coefficients in Fourier space fall off exponentially. The same is true for this basis (it is a spectral basis for the Laplacian on $\mathbb{CP}^{n+1}$).

Also like a Fourier basis, one expects that k'th order polynomials can represent structure on length scales down to $1/k$, but not on shorter scales. In the CY problem one can vary the length scales by varying the complex structure – for a hypersurface, the defining function f. For example by tuning $\psi \to -5$ above, one approaches a conifold (ODP) singularity. In this limit, a three-cycle becomes small and the accuracy becomes low, as found by HN and by Cui and Gray 1912.11068.

$$K_h = \log \sum_{I,\bar{J}} h_{I,\bar{J}} s^I \bar{s}^{\bar{J}}.$$

One might increase k to represent this smaller scale structure. However, the number of coefficients in h grows as $k^{\dim_{\mathbb{R}} M} = k^6$. To deal with this, almost all previous works on this problem only considered highly symmetric CY's such as the Fermat/Dwork quintics. In this case one can restrict to coefficients which respect the symmetry and get up to $k \sim 15$, as we discussed.

Braun *et al* 0805.3689 considered a random quintic with no symmetry at all, and computed the balanced metric for $k = 8$. They got about 5% MAPE. The method we will describe does 100 times better on similar problems.

Ashmore, He and Ovrut 1910.08605 developed an ML improvement on Donaldson's algorithm, which extrapolates the results at degree k to larger degrees k'. They report results for the Fermat quintic – it would be interesting to see this for cases with less symmetry.

$$K_h = \log \sum_{I,\bar{J}} h_{I,\bar{J}} s^I \bar{s}^{\bar{J}}.$$

One might increase k to represent this smaller scale structure. However, the number of coefficients in h grows as $k^{\dim_{\mathbb{R}} M} = k^6$. To deal with this, almost all previous works on this problem only considered highly symmetric CY's such as the Fermat/Dwork quintics. In this case one can restrict to coefficients which respect the symmetry and get up to $k \sim 15$, as we discussed.

Braun *et al* 0805.3689 considered a random quintic with no symmetry at all, and computed the balanced metric for $k = 8$. They got about 5% MAPE. The method we will describe does 100 times better on similar problems.

Ashmore, He and Ovrut 1910.08605 developed an ML improvement on Donaldson's algorithm, which extrapolates the results at degree k to larger degrees k'. They report results for the Fermat quintic – it would be interesting to see this for cases with less symmetry.

$$K_h = \log \sum_{I,\bar{J}} h_{I,\bar{J}} s^I \bar{s}^{\bar{J}}.$$

One might increase k to represent this smaller scale structure. However, the number of coefficients in h grows as $k^{\dim_{\mathbb{R}} M} = k^6$. To deal with this, almost all previous works on this problem only considered highly symmetric CY's such as the Fermat/Dwork quintics. In this case one can restrict to coefficients which respect the symmetry and get up to $k \sim 15$, as we discussed.

Braun *et al* 0805.3689 considered a random quintic with no symmetry at all, and computed the balanced metric for $k = 8$. They got about 5% MAPE. The method we will describe does 100 times better on similar problems.

Ashmore, He and Ovrut 1910.08605 developed an ML improvement on Donaldson's algorithm, which extrapolates the results at degree k to larger degrees k'. They report results for the Fermat quintic – it would be interesting to see this for cases with less symmetry.

The growth

$$N_{parameters} \sim k^{\dim M}$$

is the "curse of dimensionality," named by Bellman in the 50's. It is a fundamental problem with all computational work in high dimensions and there is a vast literature on methods to mitigate it.

It is also a fundamental problem in machine learning. Standard tasks are to classify images with millions of pixels, or documents with tens of thousands of words. The success of deep learning and other modern ML techniques at such tasks has forced a reexamination of the accepted principles of statistics.

Researchers are now adapting ML techniques to solve PDE's and other classic problems of computational science, and finding that the curse of dimensionality can be mitigated. Some well known works on the subject include Carleo and Troyer `1606.02318` (published in *Science*) on quantum many-body problems, and Han *et al* (PNAS 2017) on solving Black-Scholes and related equations.

The growth

$$N_{parameters} \sim k^{\dim M}$$

is the "curse of dimensionality," named by Bellman in the 50's. It is a fundamental problem with all computational work in high dimensions and there is a vast literature on methods to mitigate it.

It is also a fundamental problem in machine learning. Standard tasks are to classify images with millions of pixels, or documents with tens of thousands of words. The success of deep learning and other modern ML techniques at such tasks has forced a reexamination of the accepted principles of statistics.

Researchers are now adapting ML techniques to solve PDE's and other classic problems of computational science, and finding that the curse of dimensionality can be mitigated. Some well known works on the subject include Carleo and Troyer 1606.02318 (published in *Science*) on quantum many-body problems, and Han *et al* (PNAS 2017) on solving Black-Scholes and related equations.

The growth

$$N_{parameters} \sim k^{\dim M}$$

is the "curse of dimensionality," named by Bellman in the 50's. It is a fundamental problem with all computational work in high dimensions and there is a vast literature on methods to mitigate it.

It is also a fundamental problem in machine learning. Standard tasks are to classify images with millions of pixels, or documents with tens of thousands of words. The success of deep learning and other modern ML techniques at such tasks has forced a reexamination of the accepted principles of statistics.

Researchers are now adapting ML techniques to solve PDE's and other classic problems of computational science, and finding that the curse of dimensionality can be mitigated. Some well known works on the subject include Carleo and Troyer `1606.02318` (published in *Science*) on quantum many-body problems, and Han *et al* (PNAS 2017) on solving Black-Scholes and related equations.

Feedforward networks

A feed-forward network (FFN, also called MLP for multilayer perceptron) is a nonlinear map $F[\vec{W}]$ from a vector space $\mathcal{X}$ to another vector space $\mathcal{Y}$ with parameters $\vec{W}$. It is built by composing a sequence of maps which alternate between two types, general linear maps W and fixed nonlinear transformations θ, as in

$$F[\vec{W}] = W^{(d)} \circ \theta|_{V_{d-1}} \circ W^{(d-1)} \circ \ldots \circ \theta|_{V_2} \circ W^{(2)} \circ \theta|_{V_1} \circ W^{(1)}.$$

Each linear map $W^{(i)}$ has as its range a new vector space V_i, so

$$\begin{aligned} W^{(1)} &\in \mathrm{Hom}(\mathcal{X}, \mathrm{V}_1), \\ \mathrm{W}^{(2)} &\in \mathrm{Hom}(\mathrm{V}_1, \mathrm{V}_2), \\ &\vdots \\ \mathrm{W}^{(\mathrm{d})} &\in \mathrm{Hom}(\mathrm{V}_{\mathrm{d}-1}, \mathcal{Y}) \end{aligned}$$

The combination $\theta \circ \mathrm{W}$ is called a layer, with the final layer $\mathrm{W}^{(\mathrm{d})}$ being an exception in not having θ. The number of layers d is the depth.

$$F[\vec{W}] = W^{(d)} \circ \theta|_{V_{d-1}} \circ W^{(d-1)} \circ \ldots \circ \theta|_{V_2} \circ W^{(2)} \circ \theta|_{V_1} \circ W^{(1)}.$$

Generally one allows the W's to be arbitrary linear transformations, so the parameters consist of the list of $W^{(i)}$. To define the θ's, we start with the one dimensional case $\theta|_{\mathbb{K}} : \mathbb{K} \to \mathbb{K}$, which is called the **activation function**. This could be any function; two popular choices for $\mathbb{K} = \mathbb{R}$ are $\theta(x) = \tanh x$, and the "rectified linear unit" or ReLU function

$$\theta_{\mathrm{ReLU}}(x) = \begin{cases} x, x \geq 0 \\ 0, x < 0 \end{cases} .$$

To define θ_V for a general vector space V, we pick a basis e_i for V and apply $\theta|_{\mathbb{K}}$ componentwise,

$$\theta_V \left(\sum_i c_i \, e_i \right) = \sum_i \theta_{\mathbb{K}}(c_i) \, e_i .$$

Since every θ is both prefixed and postfixed by a general linear transformation, the space of maps is independent of the basis.

It has been shown that feed-forward networks can approximate arbitrary real valued functions. This is the case even for $\mathrm{d} = 2$ (Cybenko1989), but in this case one can need an exponentially large number of units, as would be the case for simpler methods of interpolation (the "curse of dimensionality"). By using more layers, one can gain many advantages – complicated functions can be represented with many fewer units, and local optimization techniques are much more effective. How exactly this works is under intensive study. A few references:

- Some observations on partial differential equations in Barron and multi-layer spaces, E and Wojtowytsch, `arXiv:2012.01484`
- Error bounds for approximations with deep ReLU networks, Yarotsky, `arXiv:1610.01145`
- Approximation and Estimation for High-Dimensional Deep Learning Networks, Barron and Klusowski, `arXiv:1809.03090`

Holomorphic and bihomogeneous networks

The CY metric computations represent the Kähler potential as the pullback of a Fubini-Study Kähler potential by an embedding $s : M \hookrightarrow \mathbb{CP}^N$,

$$K_h = \log \sum_{I,\bar{J}} h_{I,\bar{J}} s^I \bar{s}^{\bar{J}}.$$

We can regard this as the log of a homogeneous polynomial on M, and replace it with a feed forward network with the same homogeneity property.

Since we want F_W to be homogeneous under rescaling the inputs $Z^I \to \lambda Z^I$, the activation functions must be homogeneous. The natural choice is

$$\theta(z) = z^2.$$

Holomorphic and bihomogeneous networks

The CY metric computations represent the Kähler potential as the pullback of a Fubini-Study Kähler potential by an embedding $s : M \hookrightarrow \mathbb{CP}^N$,

$$K_h = \log \sum_{I,\bar{J}} h_{I,\bar{J}} s^I \bar{s}^{\bar{J}}.$$

We can regard this as the log of a homogeneous polynomial on M, and replace it with a feed forward network with the same homogeneity property.

Since we want F_W to be homogeneous under rescaling the inputs $Z^I \to \lambda Z^I$, the activation functions must be homogeneous. The natural choice is

$$\theta(z) = z^2.$$

$$F[\vec{W}] = W^{(d)} \circ \theta|_{V_{d-1}} \circ W^{(d-1)} \circ \ldots \circ \theta|_{V_2} \circ W^{(2)} \circ \theta|_{V_1} \circ W^{(1)}.$$

In **holomorphic networks**, we take the inputs to be the homogeneous coordinates Z^I and the W's to be complex. Then the output $\vec{F}_W$ is a vector of sections of $\mathcal{L}^k$ with $k = 2^{\ell-1}$, a nonlinearly parameterized subspace of $H^0(\mathcal{L}^k)$. We then take

$$K_W = \log |\vec{F}_W(Z)|^2.$$

In **bihomogeneous networks**, $\mathbb{K} \equiv \mathbb{R}$, and we take the inputs to be the real and imaginary parts of the combinations $Z^I \bar{Z}^{\bar{J}}$. These are bihomogeneous under the rescaling $(Z^I, \bar{Z}^{\bar{J}} \to (\lambda_1 Z^I, \lambda_2 \bar{Z}^{\bar{J}})$. The ℓ'th layer activations are bihomogeneous with degree $(2^\ell, 2^\ell)$. We use a one dimensional output $\mathcal{Y} \equiv \mathbb{R}$, and

$$K_W = \log F_W(Z\bar{Z}).$$

$$F[\vec{W}] = W^{(d)} \circ \theta|_{V_{d-1}} \circ W^{(d-1)} \circ \ldots \circ \theta|_{V_2} \circ W^{(2)} \circ \theta|_{V_1} \circ W^{(1)}.$$

In **holomorphic networks**, we take the inputs to be the homogeneous coordinates Z^I and the W's to be complex. Then the output $\vec{F}_W$ is a vector of sections of $\mathcal{L}^k$ with $k = 2^{\ell-1}$, a nonlinearly parameterized subspace of $H^0(\mathcal{L}^k)$. We then take

$$K_W = \log |\vec{F}_W(Z)|^2.$$

In **bihomogeneous networks**, $\mathbb{K} \equiv \mathbb{R}$, and we take the inputs to be the real and imaginary parts of the combinations $Z^I \bar{Z}^{\bar{J}}$. These are bihomogeneous under the rescaling $(Z^I, \bar{Z}^{\bar{J}} \to (\lambda_1 Z^I, \lambda_2 \bar{Z}^{\bar{J}})$. The ℓ'th layer activations are bihomogeneous with degree $(2^\ell, 2^\ell)$. We use a one dimensional output $\mathcal{Y} \equiv \mathbb{R}$, and

$$K_W = \log F_W(Z\bar{Z}).$$

Using computational techniques we discuss shortly, we can find good approximate Ricci flat metrics using these networks. And to anticipate a bit, the bihomogeneous networks work much better than the holomorphic networks.

The problem with the holomorphic networks is that $\dim \mathrm{H}^0(\mathcal{L}^k) \sim k^{\dim M}$, and to obtain nondegenerate Fubini-Study metrics one needs to span it. But in our computations $\vec{F}$ is a relatively low dimension subspace.

By contrast, using the bihomogeneous networks, it is easy to embed the degree (k, k) Kähler potentials as a subset of the $(2k, 2k)$ Kähler potentials. Given an intermediate result $F_k = (\sum_i |Z^i|^2)^k$, one can take $F_{2k} = \theta(F_k)$.

One can then add additional terms to F_{2k} to improve the approximation, for example $F' = |W \cdot Z|^{4k}$, a function with a sharp peak at $Z \propto W$ of width $1/k$.

Using computational techniques we discuss shortly, we can find good approximate Ricci flat metrics using these networks. And to anticipate a bit, the bihomogeneous networks work much better than the holomorphic networks.

The problem with the holomorphic networks is that $\dim \mathrm{H}^0(\mathcal{L}^k) \sim k^{\dim M}$, and to obtain nondegenerate Fubini-Study metrics one needs to span it. But in our computations $\vec{F}$ is a relatively low dimension subspace.

By contrast, using the bihomogeneous networks, it is easy to embed the degree (k, k) Kähler potentials as a subset of the $(2k, 2k)$ Kähler potentials. Given an intermediate result $F_k = (\sum_i |Z^i|^2)^k$, one can take $F_{2k} = \theta(F_k)$.

One can then add additional terms to F_{2k} to improve the approximation, for example $F' = |W \cdot Z|^{4k}$, a function with a sharp peak at $Z \propto W$ of width $1/k$.

Using computational techniques we discuss shortly, we can find good approximate Ricci flat metrics using these networks. And to anticipate a bit, the bihomogeneous networks work much better than the holomorphic networks.

The problem with the holomorphic networks is that $\dim H^0(\mathcal{L}^k) \sim k^{\dim M}$, and to obtain nondegenerate Fubini-Study metrics one needs to span it. But in our computations $\vec{F}$ is a relatively low dimension subspace.

By contrast, using the bihomogeneous networks, it is easy to embed the degree (k, k) Kähler potentials as a subset of the $(2k, 2k)$ Kähler potentials. Given an intermediate result $F_k = (\sum_i |Z^i|^2)^k$, one can take $F_{2k} = \theta(F_k)$.

One can then add additional terms to F_{2k} to improve the approximation, for example $F' = |W \cdot Z|^{4k}$, a function with a sharp peak at $Z \propto W$ of width $1/k$.

Using computational techniques we discuss shortly, we can find good approximate Ricci flat metrics using these networks. And to anticipate a bit, the bihomogeneous networks work much better than the holomorphic networks.

The problem with the holomorphic networks is that $\dim \mathrm{H}^0(\mathcal{L}^k) \sim k^{\dim M}$, and to obtain nondegenerate Fubini-Study metrics one needs to span it. But in our computations $\vec{F}$ is a relatively low dimension subspace.

By contrast, using the bihomogeneous networks, it is easy to embed the degree (k, k) Kähler potentials as a subset of the $(2k, 2k)$ Kähler potentials. Given an intermediate result $F_k = (\sum_i |Z^i|^2)^k$, one can take $F_{2k} = \theta(F_k)$.

One can then add additional terms to F_{2k} to improve the approximation, for example $F' = |W \cdot Z|^{4k}$, a function with a sharp peak at $Z \propto W$ of width $1/k$.

Here are bihomogeneous networks with depth 2 and 3:

$$K^{b}_{\mathcal{L};2;D_1} = \log \sum_{1\leq I\leq D_1} W^{(2)}_{I} \Big(\sum_{i} W^{(1),I}_{i,\bar{j}} s^{i}\bar{s}^{\bar{j}}\Big)^2$$

$$K^{b}_{\mathcal{L};\vec{2};D_1,D_2} = \log \sum_{1\leq I\leq D_2} W^{(3)}_{I} \left(\sum_{1\leq J\leq D_1} W^{(2),I}_{J} \Big(\sum_{i} W^{(1),J}_{i,\bar{j}} s^{i}\bar{s}^{\bar{j}}\Big)^2 \right)^2 .$$

Let $\mathrm{BiH}[D, D_1, \ldots, D_\ell]$ be the set of Kähler potentials which can be realized by a bihomogeneous network with layer widths D, D_1, *etc.*. Such a network has

$$\dim_{\mathbb{R}} \mathrm{BiH}[D, D_1, \ldots, D_\ell] = DD_1 + D_1D_2 + \ldots + D_{\ell-1}D_\ell + D_\ell$$

parameters. Take all $D_i \sim D = d^2$, then this is $\sim \ell d^4 \ll 2^{d\ell}$ parameters. So we can represent length scales $1/k$ using many fewer parameters than the Fubini-Study metrics.

The nice thing about this is that it can be easily implemented using standard ML software such as TensorFlow/Keras. Let us briefly review supervised learning and compare it with our problem.
In supervised learning, we are given a dataset of input-output pairs $(\vec{x}_i, y_i)$, a class of models $y = f_W(x)$, and an objective function which evaluates their performance, for example least squares error

$$\mathcal{L} = \sum_i |y_i - f_W(\vec{x}_i)|^2.$$

One then optimizes $\mathcal{L}$ as a function of the weights, usually by gradient descent or stochastic gradient descent (SGD):

$$\vec{W}_{t+1} = \vec{W}_t - \eta \frac{\partial}{\partial \vec{W}} \mathcal{L}[\vec{W}]$$

In SGD one uses $\mathcal{L}$ with the sum restricted to "minibatches," random subsets of the dataset. This adds a noise term to the gradient, which helps get out of local minima.

Comparing with our problem, the main difference is that our "dataset" is the manifold M. By sampling points $\mathrm{Z_i} \in \mathrm{M}$ and following the definitions above, we can minimize one of the energy functionals given earlier, defined using Monte Carlo evaluation of the integral over M.

$$\int_{\mathrm{M}} \mathrm{d}\mu\, \mathrm{f} \to \frac{1}{\mathrm{N}} \sum_{\mathrm{i}} \mathrm{f}(\mathrm{Z_i})$$

The correspondence is

$$\vec{\mathrm{x}}_{\mathrm{i}} = \mathrm{Z}^{\mathrm{I}} \bar{\mathrm{Z}}^{\bar{\mathrm{J}}}|_{\mathrm{Z}=\mathrm{Z_i}}; \qquad \mathrm{y_i} = \Omega \wedge \bar{\Omega}|_{\mathrm{Z}=\mathrm{Z_i}}; \qquad \mathrm{f_w} = \det \partial\bar{\partial} \log \mathrm{F_W}[\mathrm{x_i}].$$

We just need to add code to sample $\mathrm{Z_i}$ and to compute $\Omega \wedge \bar{\Omega}$, and a layer to compute the volume form $\mathrm{F} \to \det \partial\bar{\partial}\mathrm{F}$.

ML software makes it easy to define feed forward networks,

```
class twolayers(tf.keras.Model):

    def __init__(self, n_units):
        super(twolayers, self).__init__()
        self.bihomogeneous = bnn.Bihomogeneous()
        self.layer1 = bnn.Dense(25, n_units[0], activation=tf.square)
        self.layer2 = bnn.Dense(n_units[0], n_units[1], activation=tf.square)
        self.layer3 = bnn.Dense(n_units[1], 1)

    def call(self, inputs):
        x = self.bihomogeneous(inputs)
        x = self.layer1(x)
        x = self.layer2(x)
        x = self.layer3(x)
        x = tf.math.log(x)
        return x
```

It also makes gradient descent easy, computing the derivatives $\partial \mathrm{f}_{\mathrm{W}}/\partial \mathrm{W}$ using backpropagation. Also provided are "housekeeping" tasks such as initialization, saving coefficients, *etc.*

We implemented this, and we have begun to study geometry with it (special Lagrangian torus fibrations, more general Einstein metrics) but this is work in progress. Here I will describe some implementation details, and the following points:

1. Implementation and numerical details.
2. Choice of hyperparameters (network width and depth, optimizers and learning schedule) to get accurate results.
3. Questions which are a focus of current ML research: the optimization landscape and the role of overparameterization.
4. The difference of expressivity of a network with a polynomial number of parameters versus an exponential number of parameters.

The last question, which we formulated in the course of our research, is very general. It could be asked about numerical methods for many PDEs and perhaps can be phrased as a question in computational complexity theory. Let us return to it.

Implementation details

Our code is at `http://github.com/yidiq7/MLGeometry`.
Two points which might not be obvious:

- One wants to avoid divisions by small numbers. This can happen when one goes to inhomogeneous coordinates in which one $Z^a = 1$, and it can also happen in the formula $\Omega = \prod dZ/\partial f/\partial Z^b$. To avoid this we provide two levels of coordinate patches, the first U_a in which $|Z^a| \geq |Z^b| \forall b$, and the second $U_{a,c}$ in which $|\partial f/\partial Z^c| \geq |\partial f/\partial Z^b| | \forall c$. The code automatically assigns each point Z_i to the correct patch.
- Gradient descent does not efficiently find the minimum of $\mathcal{L}$. One can do better by using the Adam adaptive optimization method, but this is still a first order method and does not converge quickly. We thus do the optimization in two stages – once Adam has reached the neighborhood of an optimum, we continue with the second order L-BFGS method.

Even so, the energy function $\mathcal{L}$ is not convex and it is not clear one is finding a global minimum. While this is a general problem in ML, the continuum Ricci flat Kähler problem is better behaved: if $\mathrm{F}(\eta)$ is convex, then the energy function $\int_{\mathrm{M}} \mathrm{F}(\eta)$ is convex.
However this will generally not be the case for feedforward networks. Even for the simplest network with a linear activation function,

$$\mathrm{F}_{\mathrm{W}} = \mathrm{W}_1 \mathrm{W}_2 \mathrm{x}$$

the energy function is not strictly convex. In general it is more complicated.

It is also not generally true if the integrals are done by sampling. Consider a finite number N_{p} of samples $(\mathrm{x}_{\mathrm{i}}, \mathrm{y}_{\mathrm{i}})$. There will be some number of parameters $\mathrm{P}_{\mathrm{int}}$ which given x_{i}, suffices to fit (interpolate) any generic prescribed values of the y_{i}. For $\mathrm{P} > \mathrm{P}_{\mathrm{int}}$ the model is overparameterized, and again the energy function will not be strictly convex. Metrics with no symmetry will often require large P.
Let us come back to these points after discussing some results.

Even so, the energy function $\mathcal{L}$ is not convex and it is not clear one is finding a global minimum. While this is a general problem in ML, the continuum Ricci flat Kähler problem is better behaved: if $\mathrm{F}(\eta)$ is convex, then the energy function $\int_{\mathrm{M}} \mathrm{F}(\eta)$ is convex.
However this will generally not be the case for feedforward networks. Even for the simplest network with a linear activation function,

$$\mathrm{F}_{\mathrm{W}} = \mathrm{W}_1 \mathrm{W}_2 \mathrm{x}$$

the energy function is not strictly convex. In general it is more complicated.

It is also not generally true if the integrals are done by sampling. Consider a finite number N_{p} of samples $(\mathrm{x}_{\mathrm{i}}, \mathrm{y}_{\mathrm{i}})$. There will be some number of parameters $\mathrm{P}_{\mathrm{int}}$ which given x_{i}, suffices to fit (interpolate) any generic prescribed values of the y_{i}. For $\mathrm{P} > \mathrm{P}_{\mathrm{int}}$ the model is overparameterized, and again the energy function will not be strictly convex. Metrics with no symmetry will often require large P.
Let us come back to these points after discussing some results.

We studied CY hypersurfaces with varying amounts of symmetry:

1. The Dwork quintics with maximal symmetry: $\mathrm{f} = \mathrm{f}_1$ below with $\phi = 0$.
2. A two parameter family with less symmetry,

$$\mathrm{f}_1 = z_0^5 + z_1^5 + z_2^5 + z_3^5 + z_4^5 + \psi z_0 z_1 z_2 z_3 z_4 + \phi(z_3 z_4^4 + z_3^2 z_4^3 + z_3^3 z_4^2 + z_3^4 z_4)$$

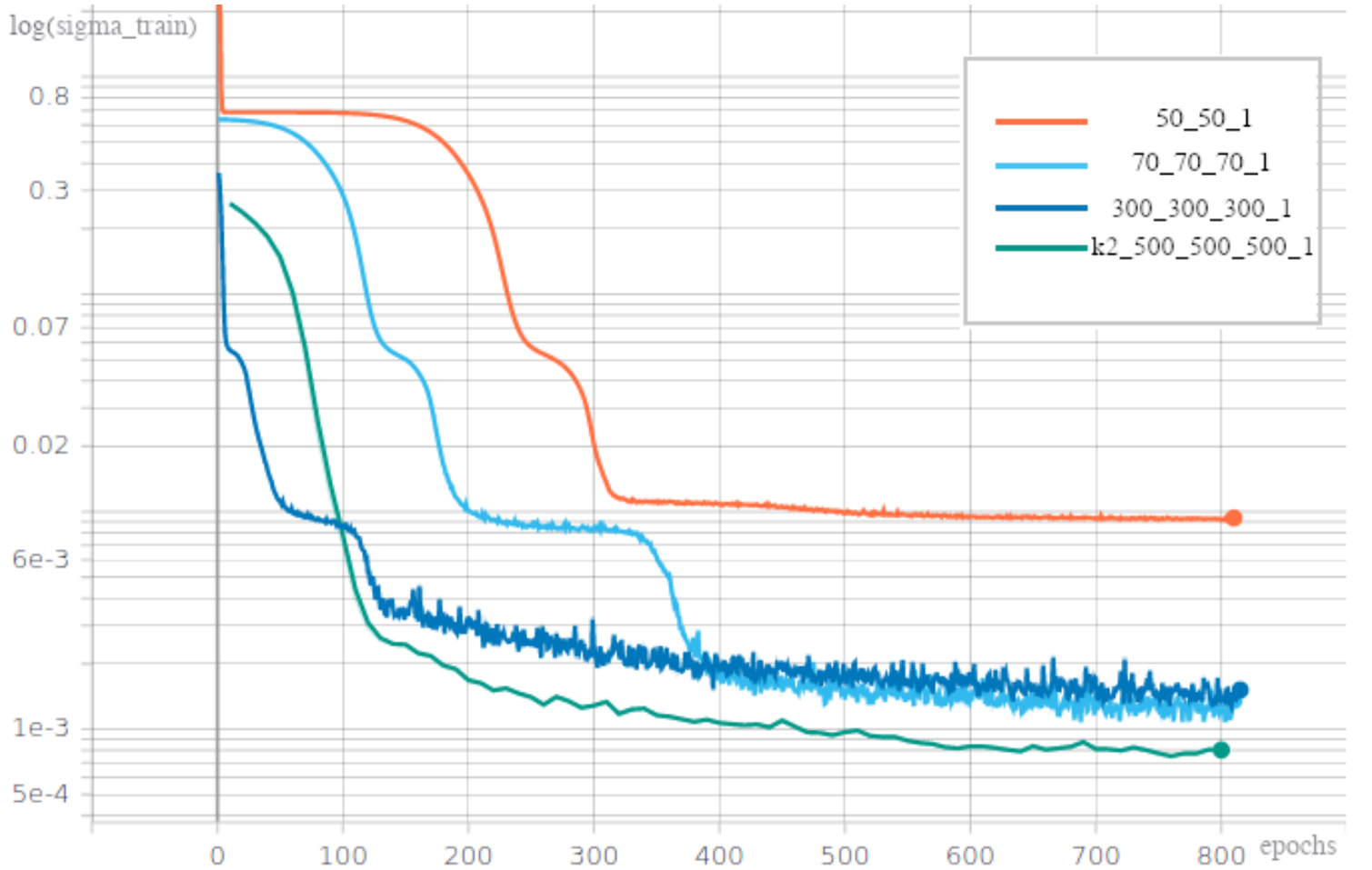
3. A two parameter family which generically has no symmetry,

$$\begin{aligned}\mathrm{f}_2 = \mathrm{f}_1|_{\phi=0} + \alpha\Big(& z_2 z_0^4 + z_0 z_4 z_1^3 + z_0 z_2 z_3 z_4^2 + \\ & z_3^2 z_1^3 + z_4 z_1^2 z_2^2 + z_0 z_1 z_2 z_3^2 + \\ & z_2 z_4 z_3^3 + z_0 z_1^4 + z_0 z_4^2 z_2^2 + z_4^3 z_1^2 + \\ & z_0 z_2 z_3^3 + z_3 z_4 z_0^3 + z_1^3 z_4^2 + \\ & z_0 z_2 z_4 z_1^2 + z_1^2 z_3^3 + z_1 z_4^4 + z_1 z_2 z_0^3 + \\ & z_2^2 z_4^3 + z_4 z_2^4 + z_1 z_3^4 \Big).\end{aligned}$$

We began by comparing various options and their effects on accuracy, speed and reliability (sometimes a run would work for some initializations and not others), with the following results:

- Holomorphic networks did not work reliably.
- MSE and MAPE loss functions worked equally well in training. While MAX did not work so well, for the 3 layer networks, it was helpful to add $0.1 * \text{MAX}$ in the early stage of training, to prevent getting stuck in a bad local minimum.
- Testing and training errors are comparable for the smaller networks, but larger networks sometimes overfit.
- MSE error is roughly the square of the MAPE error.
- MAX error is often larger than MAPE and had different hyperparameter dependence.
- ℓ_2 regularization did not help.
- 64 bit networks did not do better than 32 bit.
- The speed of Adam convergence (measured by number of epochs, not compute time) improved with both depth and width of the networks. Still a second pass of L-BFGS was helpful.

Figure: The training curves for the Dwork quintic with $\psi = 0.5$, trained with Adam optimizer and MAPE loss. The data for k2_500_500_500_1 was recorded every 10 epochs.



Experiments and observations

After narrowing down the hyperparameter choices and implementing L-BFGS, we studied the accuracy of the method as a function of the most significant parameters.
For the geometry of the CY, we want to distinguish dependence on

- The shortest length scale $\sim$ distance in moduli space to a singular CY, and
- The "complexity" of the CY, both lack of symmetry (thus requiring more parameters) and perhaps other factors.

For the model (network), we want to distinguish dependence on

- The depth $d = \ell + 1$ of the network: degree $k = 2^d$, and
- The total number of parameters.

The relation between distance to a singular CY and shortest length scale is a bit imprecise, and the actual distance in the Weil-Petersson moduli space metric, though well defined, is not easy to compute. Thus, we used a simpler proxy for the distance,

$$\sin\theta(f) \propto d(f) \equiv \min_{Z\in M} \frac{|\partial_i f(Z)|_H}{|Z|^{n-1}\,||f||_H},$$

where $||f||_H = \langle |f|^2 \rangle$ under the Gaussian measure $\exp -Z^\dagger H Z$.

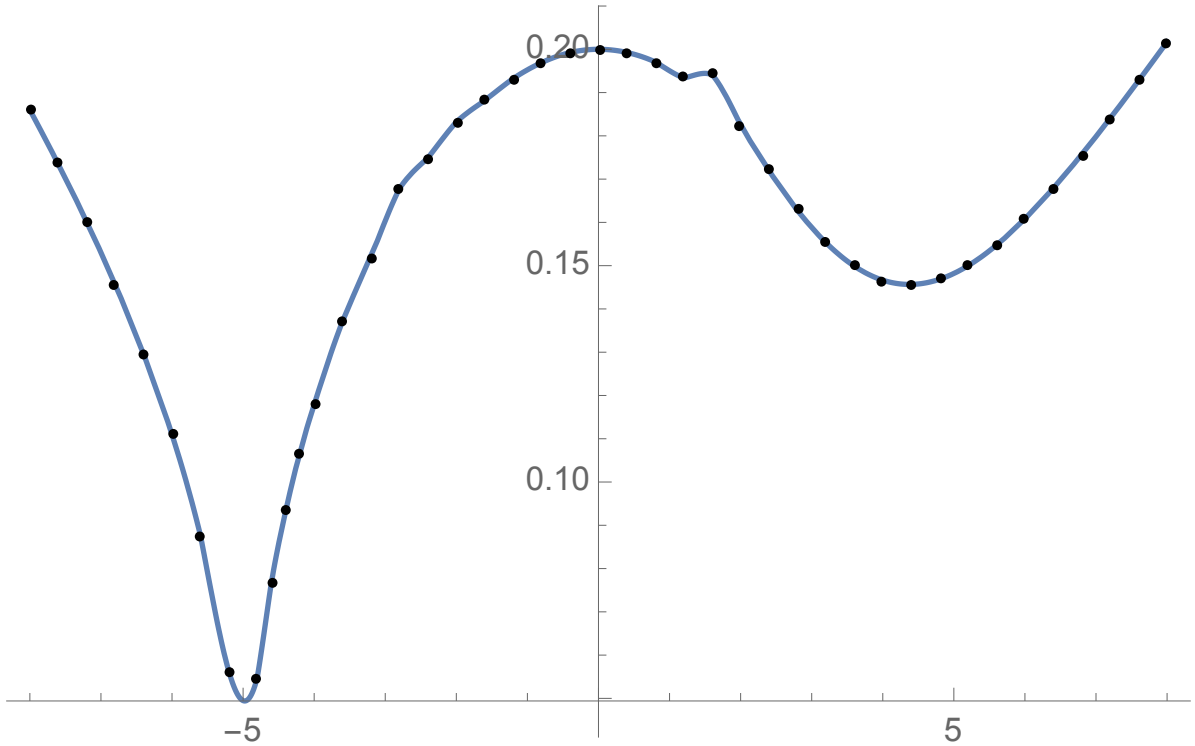
The idea (inspired by a similar question in Blum *et al*'s *Complexity and Real Computation*) is that for a given $Z \in M$, the equations $f(Z) = 0$ and $\{\partial_i f(Z) = 0\}$ define two linear subspaces of $\mathbb{C}^{126}$, the space of coefficients of f. The ratio is the shortest distance in the Fubini-Study metric on $\mathbb{CP}^{125}$ (derived from the metric H on $\mathbb{CP}^4$) between points in these two subspaces.

The relation between distance to a singular CY and shortest length scale is a bit imprecise, and the actual distance in the Weil-Petersson moduli space metric, though well defined, is not easy to compute. Thus, we used a simpler proxy for the distance,

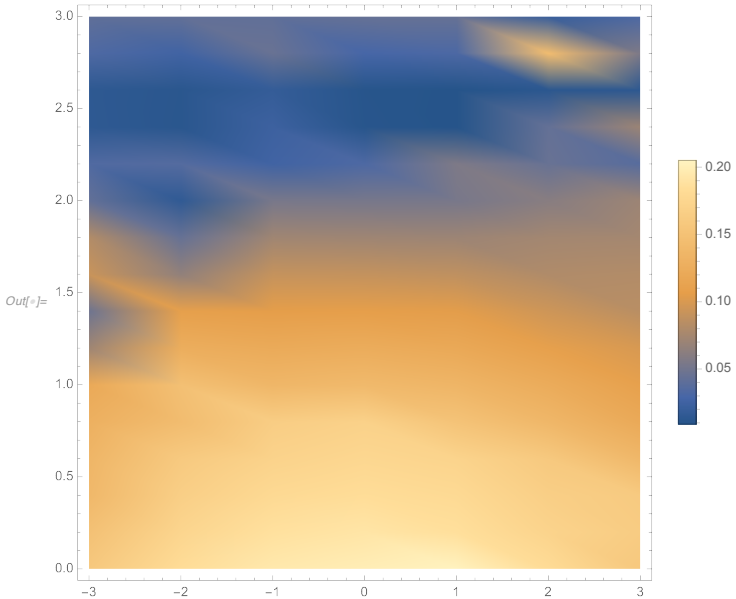
$$\sin\theta(f) \propto d(f) \equiv \min_{Z\in M} \frac{|\partial_i f(Z)|_H}{|Z|^{n-1}\,||f||_H},$$

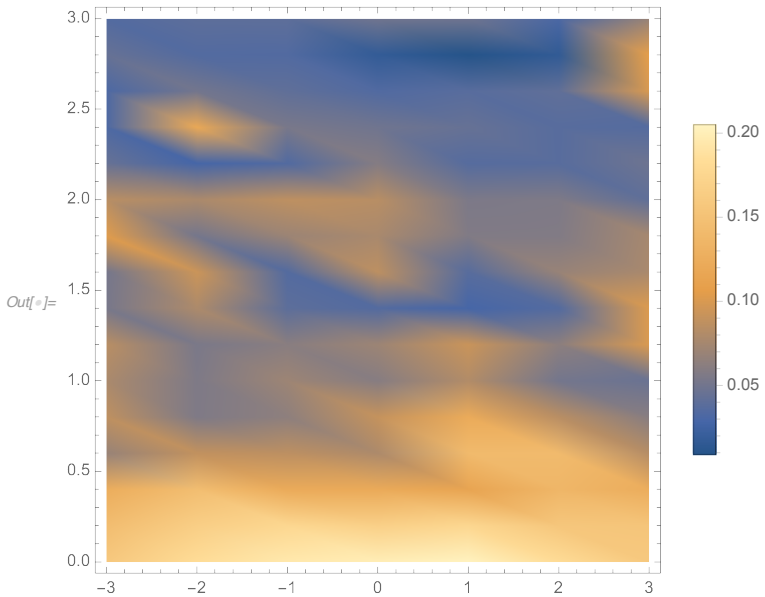
where $||f||_H = \langle |f|^2 \rangle$ under the Gaussian measure $\exp -Z^\dagger H Z$.

The idea (inspired by a similar question in Blum *et al*'s *Complexity and Real Computation*) is that for a given $Z \in M$, the equations $f(Z) = 0$ and $\{\partial_i f(Z) = 0\}$ define two linear subspaces of $\mathbb{C}^{126}$, the space of coefficients of f. The ratio is the shortest distance in the Fubini-Study metric on $\mathbb{CP}^{125}$ (derived from the metric H on $\mathbb{CP}^4$) between points in these two subspaces.



Distance to the discriminant locus for the Dwork quintics. Besides the conifold point at $\psi = -5$, there is a local minimum near $\psi = 5$, which fits with the feature seen in the plot of curvature versus ψ in Cui and Gray `1912.11068`. This is the point on the positive real axis closest to the conifold point, perhaps reached by following a path like $\psi = 5e^{i\theta}$.





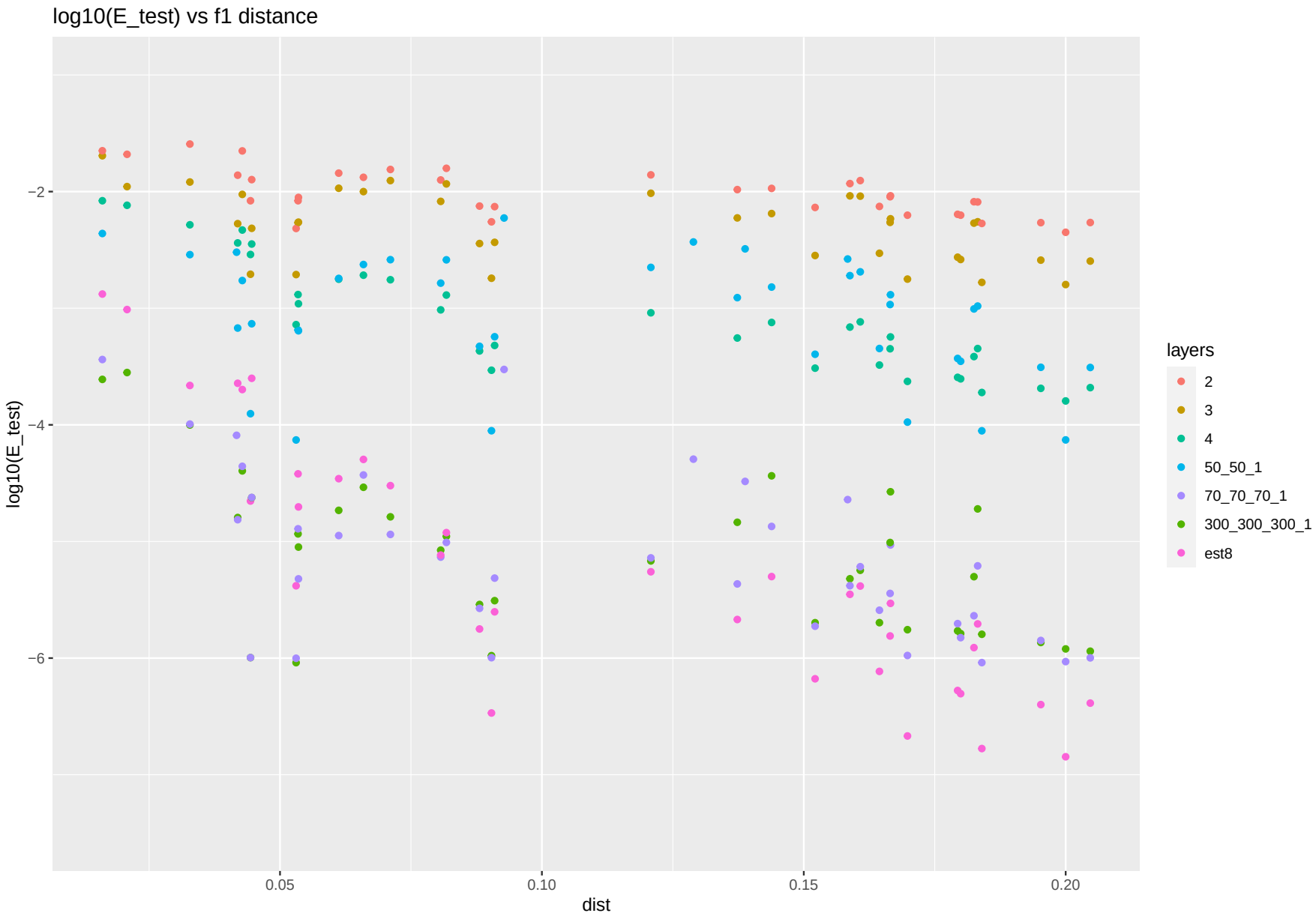
Distances to the discriminant locus as a function of ψ, ϕ in $f_1(Z)$, and as a function of ψ, α in $f_2(Z)$.

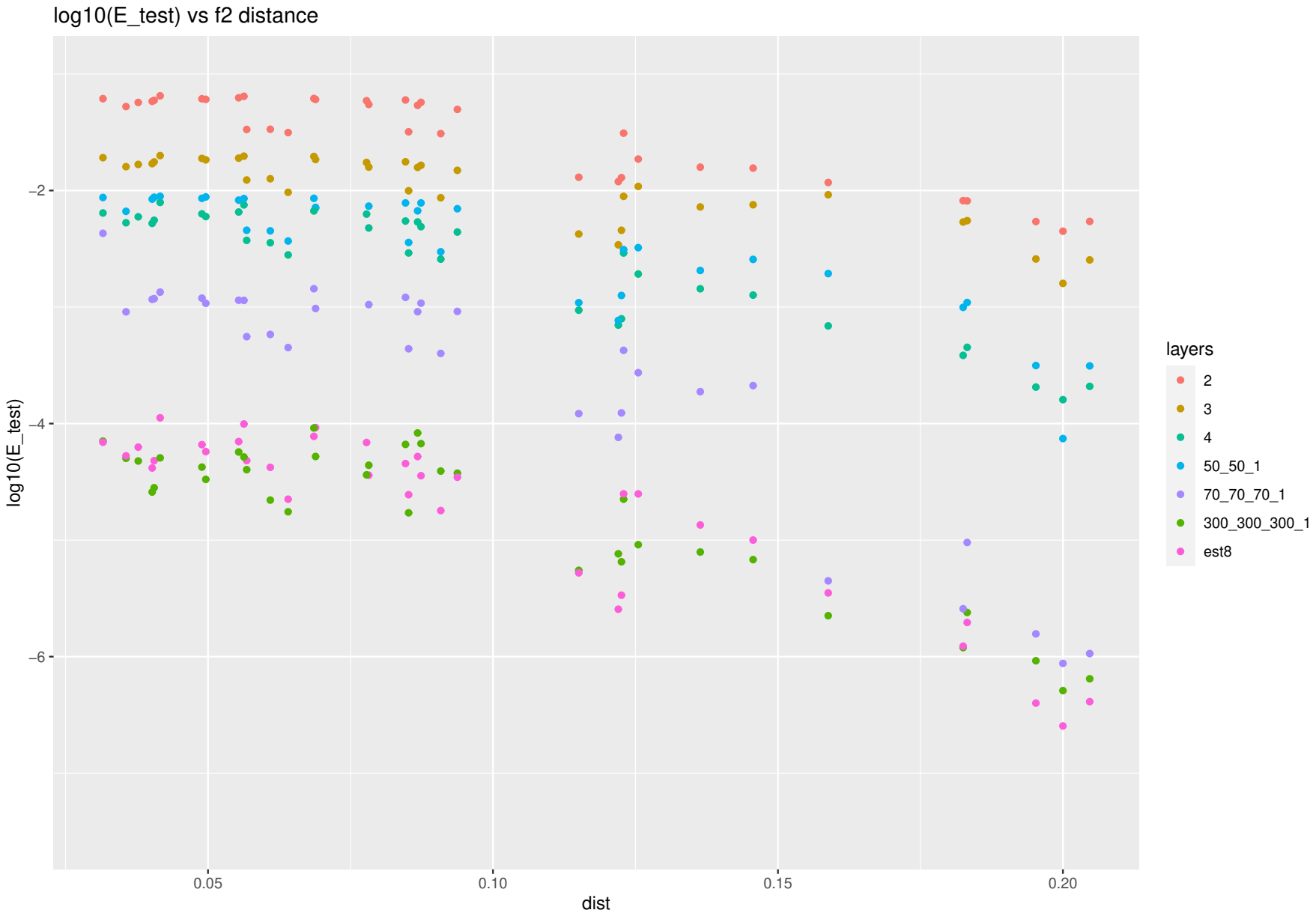
Our examples cover a variety of distances. The "complexity" of the CYs is harder to make precise, but f_2 with no symmetry would seem more complex than f_1.
Another useful data point is to know the maximal attainable accuracy for each CY within the space of Fubini-Study metrics. This would be a lot of computation to get directly. Instead we followed an observation of Headrick and Nassar `0908.2635`. As explained by Donaldson in this context, the error ϵ of the best polynomial approximation to a given smooth function will decrease faster than any power of the degree. This is analogous to the statement that the Fourier transform of a smooth function will decrease faster than $k^{-\nu}$ for any ν.
This does not immediately imply that $\epsilon \propto C^{-k}$, but Headrick and Nassar found this to be so in several examples over a wide range of k. Granting this, we can compute the error for $k = 2, 3, 4$ and extrapolate.

On the next slides, we plot results for a variety of networks, and the extrapolated $k = 8$ best possible accuracy, against distance to the singularity.

Our examples cover a variety of distances. The "complexity" of the CYs is harder to make precise, but f_2 with no symmetry would seem more complex than f_1.
Another useful data point is to know the maximal attainable accuracy for each CY within the space of Fubini-Study metrics. This would be a lot of computation to get directly. Instead we followed an observation of Headrick and Nassar `0908.2635`. As explained by Donaldson in this context, the error ϵ of the best polynomial approximation to a given smooth function will decrease faster than any power of the degree. This is analogous to the statement that the Fourier transform of a smooth function will decrease faster than $k^{-\nu}$ for any ν.
This does not immediately imply that $\epsilon \propto C^{-k}$, but Headrick and Nassar found this to be so in several examples over a wide range of k. Granting this, we can compute the error for $k = 2, 3, 4$ and extrapolate.

On the next slides, we plot results for a variety of networks, and the extrapolated $k = 8$ best possible accuracy, against distance to the singularity.


log10(E_test) vs f1 distance
log10(E_test)
−2
−4
−6
0.05
0.10
0.15
0.20
dist
layers
2
3
4
50_50_1
70_70_70_1
300_300_300_1
est8


log10(E_test) vs f2 distance
log10(E_test)
−2
−4
−6
0.05
0.10
0.15
0.20
dist
layers
2
3
4
50_50_1
70_70_70_1
300_300_300_1
est8

What do we take from this?

- The degree k is the hyperparameter with the largest effect.
- There is also a dependence $\log \epsilon \sim d(f)$.
- To distinguish k from the number of parameters, we studied two four layer networks, one with three layers of width 70 (or 70_70_70_1) and another 300_300_300_1. Both have $k = 8$, but the first has 11620 real parameters and the second has 187800 parameters, while the FS metric has 245025 parameters.
  Looking at f_1, both achieve roughly the optimal accuracy.
  But looking at f_2, this is only so for larger values of distance d. For more singular CYs, only the 300_300_300_1 network can match the optimal accuracy.

So, we have evidence that the accuracy is not controlled just by k and the distance to the singularity, but also by the total number of parameters, and some sort of "complexity" measure. Since the 300_300_300_1 network has almost as many parameters as the $k = 8$ FS metric, its accuracy is not surprising. There is still a memory savings, as the GPU works with arrays of size $N_{batch} \times$ width.

These four layer bihomogeneous networks give a pretty reliable 10^{-5} MSE, which is probably good enough for applications to geometry and string theory. It would be interesting to know what accuracy is required to guarantee the existence of a continuum solution.

One might be able to get similar results with FS $k = 8$. An example was done in Braun *et al* 0805.3689 but for the balanced metric.

We tried five layer networks with $k = 16$, as well as feeding in a complete basis of $k = 2$ or $k = 4$ sections as inputs, to get $k = 16$. Some runs got a factor of 10 better accuracy, but this was not reliable. Improving the accuracy may be an interesting challenge, and we are thinking of setting up a leaderboard on our Github site.

These four layer bihomogeneous networks give a pretty reliable 10^{-5} MSE, which is probably good enough for applications to geometry and string theory. It would be interesting to know what accuracy is required to guarantee the existence of a continuum solution.

One might be able to get similar results with FS $k = 8$. An example was done in Braun *et al* `0805.3689` but for the balanced metric.

We tried five layer networks with $k = 16$, as well as feeding in a complete basis of $k = 2$ or $k = 4$ sections as inputs, to get $k = 16$. Some runs got a factor of 10 better accuracy, but this was not reliable. Improving the accuracy may be an interesting challenge, and we are thinking of setting up a leaderboard on our Github site.

Our first geometry project (with Yidi Qi) is to study closed geodesics on the quintic threefolds. For example, a geodesic in $\Sigma \subset M$ defined as the fixed points of a discrete symmetry will lift to a geodesic on M. On the Fermat quintic we can take $Z_4 = Z_5 = 0$ to get a quintic curve in $\mathbb{CP}^2$, and then use a minimal length noncontractable cycle.

There are physics arguments (Gao and Douglas 2011) that many such geodesics exist, because strings embedded into such geodesics are the "winding states" of the nonlinear sigma model with target space M. Naively, the physics arguments suggest that these should be "stable" in the sense that the second variation operator V should be non-negative definite. However, Bourguignon (1976) showed that this is never the case for a Ricci-flat Kähler manifold in $d > 1$, because the sum of the eigenvalues of V is zero. From the physics point of view this is a bit paradoxical.

A potential resolution is that the negative eigenvalue is associated to an oscillatory "breather" trajectory along which the string shrinks from the geodesic down to a point and then expands again.

Our first geometry project (with Yidi Qi) is to study closed geodesics on the quintic threefolds. For example, a geodesic in $\Sigma \subset M$ defined as the fixed points of a discrete symmetry will lift to a geodesic on M. On the Fermat quintic we can take $Z_4 = Z_5 = 0$ to get a quintic curve in $\mathbb{CP}^2$, and then use a minimal length noncontractable cycle.

There are physics arguments (Gao and Douglas 2011) that many such geodesics exist, because strings embedded into such geodesics are the "winding states" of the nonlinear sigma model with target space M. Naively, the physics arguments suggest that these should be "stable" in the sense that the second variation operator V should be non-negative definite. However, Bourguignon (1976) showed that this is never the case for a Ricci-flat Kähler manifold in $d > 1$, because the sum of the eigenvalues of V is zero. From the physics point of view this is a bit paradoxical.

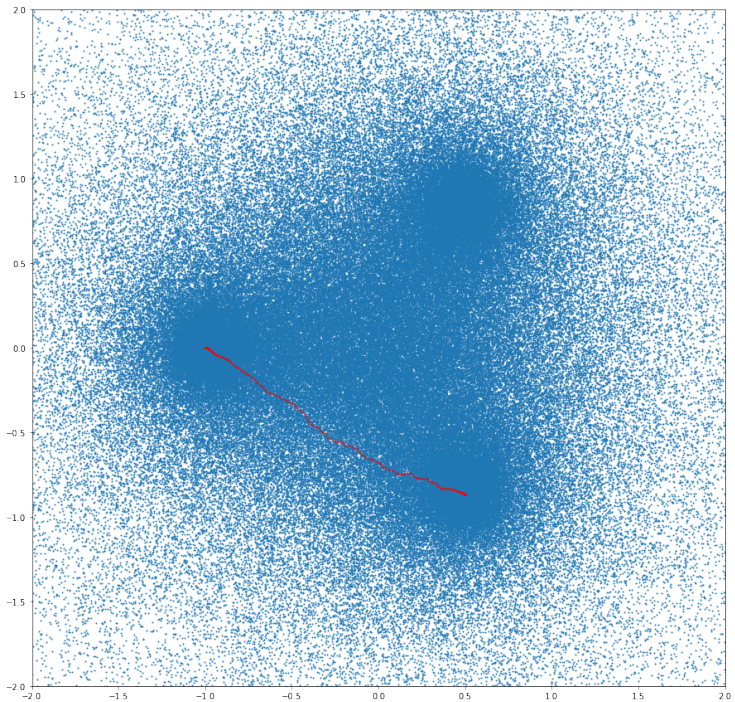
A potential resolution is that the negative eigenvalue is associated to an oscillatory "breather" trajectory along which the string shrinks from the geodesic down to a point and then expands again.

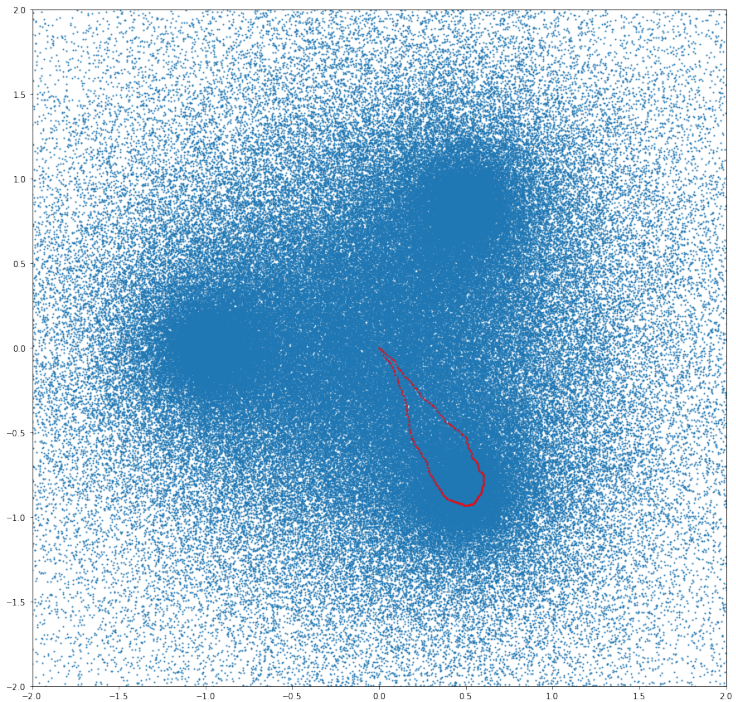
Our first geometry project (with Yidi Qi) is to study closed geodesics on the quintic threefolds. For example, a geodesic in $\Sigma \subset M$ defined as the fixed points of a discrete symmetry will lift to a geodesic on M. On the Fermat quintic we can take $Z_4 = Z_5 = 0$ to get a quintic curve in $\mathbb{CP}^2$, and then use a minimal length noncontractable cycle.

There are physics arguments (Gao and Douglas 2011) that many such geodesics exist, because strings embedded into such geodesics are the "winding states" of the nonlinear sigma model with target space M. Naively, the physics arguments suggest that these should be "stable" in the sense that the second variation operator V should be non-negative definite. However, Bourguignon (1976) showed that this is never the case for a Ricci-flat Kähler manifold in $d > 1$, because the sum of the eigenvalues of V is zero. From the physics point of view this is a bit paradoxical.

A potential resolution is that the negative eigenvalue is associated to an oscillatory "breather" trajectory along which the string shrinks from the geodesic down to a point and then expands again.

To see if this is the case we are numerically generating closed geodesics and will then simulate the conjectured dynamics. An approximate geodesic can be found by sampling points x_i and finding the shortest path through the metric graph with distances $d(x_i, x_j)$. This can be improved by gradient descent. So far we have closed geodesics on T^2 and K3.





Theoretical questions

As discussed earlier, the Ricci flat metric on M can be approximated by a degree k FS metric to $o(k^{-\nu})$ for any ν, indeed this would be the case for any smooth function. In other words, the log error goes as $\log \epsilon \sim -k^{\alpha}$ for some α. This is nice but we need $O(k^{\dim_R M})$, here $O(k^6)$ coefficients to do it.

Can we do the same with a series of depth d fixed width D networks? This has $k = 2^{d-1}$ so if we could, we would need only $O(\log k)$ coefficients to do it. Even if D grows as a power of d, we still have $O((\log k)^n)$ coefficients for some n. This difference between a power and a log is a complexity theory question.

Theoretical questions

As discussed earlier, the Ricci flat metric on M can be approximated by a degree k FS metric to $o(k^{-\nu})$ for any ν, indeed this would be the case for any smooth function. In other words, the log error goes as $\log \epsilon \sim -k^{\alpha}$ for some α. This is nice but we need $O(k^{\dim_R M})$, here $O(k^6)$ coefficients to do it.
Can we do the same with a series of depth d fixed width D networks? This has $k = 2^{d-1}$ so if we could, we would need only $O(\log k)$ coefficients to do it. Even if D grows as a power of d, we still have $O((\log k)^n)$ coefficients for some n. This difference between a power and a log is a complexity theory question.

It seems to me that there exist smooth functions which cannot be well approximated with $O((\log k)^n)$ coefficients. Perhaps one can prove this with a covering argument (a lower bound on the number of balls required to cover the space of functions).

A heuristic argument is to look at a simple class of networks which can approximate general metrics and then ask whether there are ways to reduce the number of parameters by sharing intermediate results. Start with a two layer network $\ell = 1$ with inputs of degree $(k/2, k/2)$. Since it has no intermediate results, one expects it to need as many parameters as the FS metric. Comparing the parameters at degree $(k/2, k/2)$ with (k, k), the minimal width (in n complex dimensions) is

$$\binom{k+n}{k}^2 = D_1 \binom{\frac{k}{2}+n}{\frac{k}{2}}^2; \qquad D_1 \sim \begin{cases} 2^{2n} & \text{for } n \ll k, \\ \left(\frac{n}{k}\right)^k & \text{for } n \gg k \end{cases}.$$

It seems to me that there exist smooth functions which cannot be well approximated with $O((\log k)^n)$ coefficients. Perhaps one can prove this with a covering argument (a lower bound on the number of balls required to cover the space of functions).

A heuristic argument is to look at a simple class of networks which can approximate general metrics and then ask whether there are ways to reduce the number of parameters by sharing intermediate results. Start with a two layer network $\ell = 1$ with inputs of degree $(k/2, k/2)$. Since it has no intermediate results, one expects it to need as many parameters as the FS metric. Comparing the parameters at degree $(k/2, k/2)$ with (k, k), the minimal width (in n complex dimensions) is

$$\binom{k+n}{k}^2 = D_1 \binom{\frac{k}{2}+n}{\frac{k}{2}}^2; \qquad D_1 \sim \begin{cases} 2^{2n} & \text{for } n \ll k, \\ \left(\frac{n}{k}\right)^k & \text{for } n \gg k \end{cases}.$$

So, for $n = 3 \ll k$, one only needs $D_1 = 2^6$. On the other hand, one needs each of these inputs to be an independently adjustable degree $(k/2, k/2)$ function. If we produce each of these with an independent network, we do not save any parameters.

If we continue this all the way to the first layer, we would need an exponentially large number of inputs. Of course this is highly redundant as there are only $(n+2)^2$ independent functions of degree $(1,1)$. In fact by replacing all of the independent networks in layers up to $d/2$ with the computation of the general function of degree $(2^{d/2}, 2^{d/2})$ one saves parameters. The final network would have $\sim 2\sqrt{N_p} \sim k^n$ parameters, many fewer than k^{2n} but still a power law.

The Ricci flat metric might be simpler, but I don't see a good reason to think so. If our computations had found a network which had fewer parameters than the FS metrics but always achieved comparable accuracy, that would be evidence. We did not find that so far.

So, for $n = 3 \ll k$, one only needs $D_1 = 2^6$. On the other hand, one needs each of these inputs to be an independently adjustable degree $(k/2, k/2)$ function. If we produce each of these with an independent network, we do not save any parameters.

If we continue this all the way to the first layer, we would need an exponentially large number of inputs. Of course this is highly redundant as there are only $(n+2)^2$ independent functions of degree $(1,1)$. In fact by replacing all of the independent networks in layers up to $d/2$ with the computation of the general function of degree $(2^{d/2}, 2^{d/2})$ one saves parameters. The final network would have $\sim 2\sqrt{N_p} \sim k^n$ parameters, many fewer than k^{2n} but still a power law.

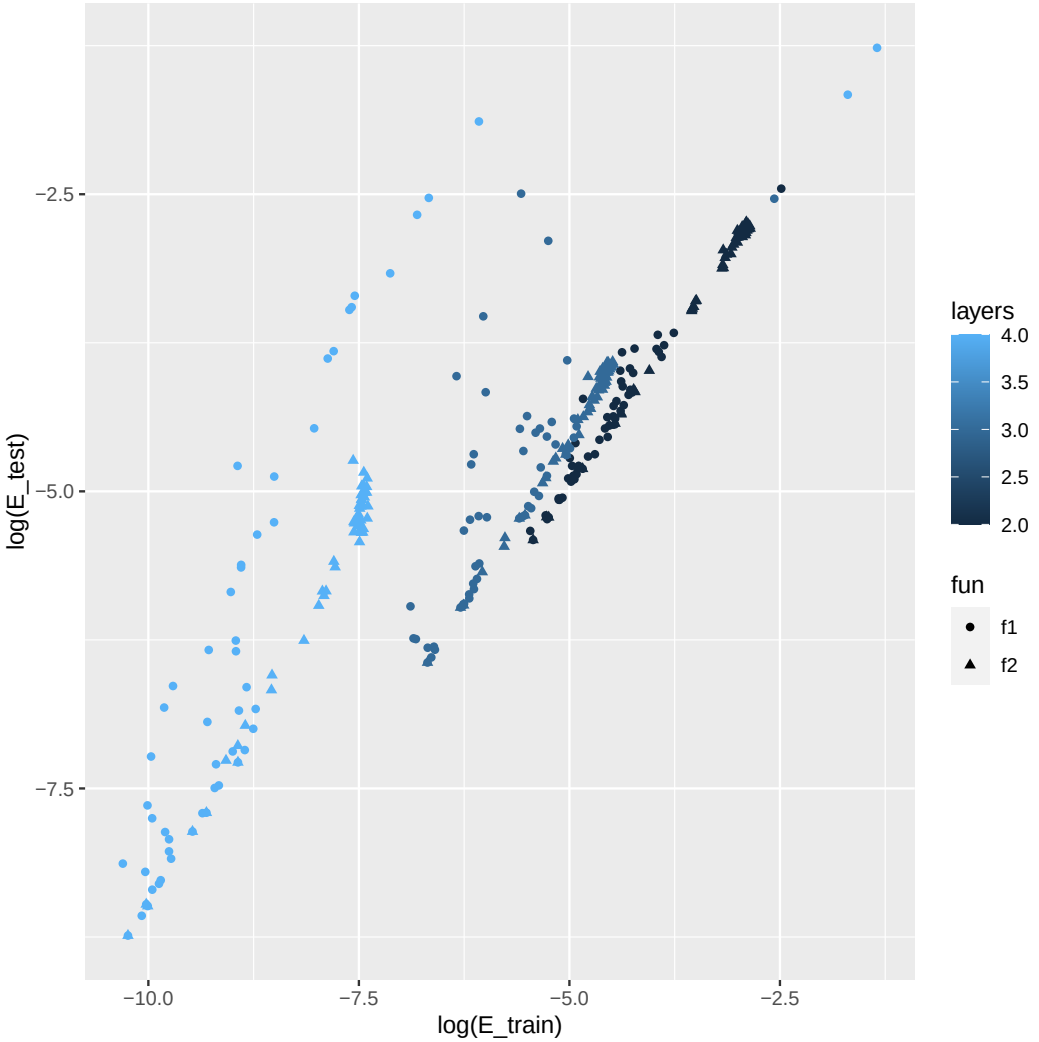
The Ricci flat metric might be simpler, but I don't see a good reason to think so. If our computations had found a network which had fewer parameters than the FS metrics but always achieved comparable accuracy, that would be evidence. We did not find that so far.

So, for $n = 3 \ll k$, one only needs $D_1 = 2^6$. On the other hand, one needs each of these inputs to be an independently adjustable degree $(k/2, k/2)$ function. If we produce each of these with an independent network, we do not save any parameters.

If we continue this all the way to the first layer, we would need an exponentially large number of inputs. Of course this is highly redundant as there are only $(n+2)^2$ independent functions of degree $(1,1)$. In fact by replacing all of the independent networks in layers up to $d/2$ with the computation of the general function of degree $(2^{d/2}, 2^{d/2})$ one saves parameters. The final network would have $\sim 2\sqrt{N_p} \sim k^n$ parameters, many fewer than k^{2n} but still a power law.

The Ricci flat metric might be simpler, but I don't see a good reason to think so. If our computations had found a network which had fewer parameters than the FS metrics but always achieved comparable accuracy, that would be evidence. We did not find that so far.

Another issue is that we are using fewer data points than parameters.



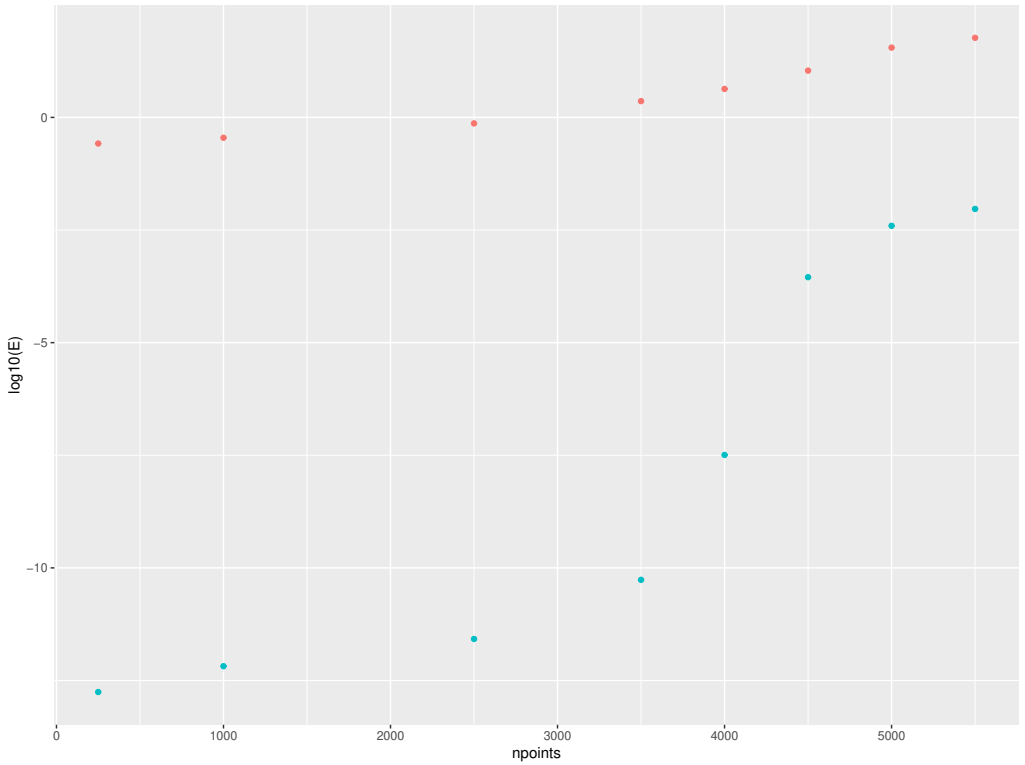
Our 300_300_300_1 network has 187800 parameters, while our Monte Carlo integrals used fewer points, between 20000 and 100000. Since each point provides one constraint on $\det \omega$, this is the "overparameterized" regime which can fit any function. This shows up in overfitting, $E\_train \ll E\_test$.

In statistics and numerical analysis textbooks, one is warned not to use models with so many parameters, as they will overfit the data. In statistics, real world data has noise which should not be fit. Numerical errors can also lead to problems.
However in deep learning one often successfully uses models with more parameters than data points. A well known work of Zhang *et al* `1611.03530` made the contradiction even sharper, by showing that a standard image classification model (CIFAR-10) could learn to associate randomly chosen labels with images. According to the dogma of statistics, such a model could not encode any prior knowledge about images, so it could not generalize to correctly classify images it had not seen.
A related observation is that a model that can fit every observation has training error zero. Thus there can be no relation between the training error and the testing error, so no reason to expect generalization.

The resolution of this paradox is not completely clear, but many researchers believe that the choice of initialization and optimization procedure used in deep learning leads to a preferred subset of minima of the error function, which have some sort of "implicit regularization." Regularization means the device of adding a term to the error function which favors small weights, for example in ℓ_2 regularization one adds the sum of the squares of the weights. The idea is that randomness in the initialization leads to an effective regularization term — this has been shown for random feature networks in Mei and Montanari `1908.05355`.

The overparameterized regime has advantages. Optimization seems to be easier in this regime. One sometimes finds better generalization than the traditional few parameter regime. This is part of a phenomenon called "double descent" in which generalization is worst right at the boundary of the overparameterized regime $P = N_{points}$, and improves in both directions.

We face similar issues and are looking for much more accurate computations. In general $N_{params} \geq N_{points}$ works better than one would think. Our models can also fit randomly shuffled vol_Ω values for $P = N_{params} > 2N_{points}$ (it looks like $\log \epsilon \sim 1/(1 + e^{N_p/P})$. Optimization is easier for larger P.



Log MSE error for the randomized $k = 4$ FS model. Blue points are training error, red testing error.

Tensor networks

The activation function $\theta(z) = z^2$ can be generalized to z^p and further generalized to a multilinear product

$$\theta_{\otimes} : \mathcal{L}^{k_1} \times \mathcal{L}^{k_2} \times \ldots \times \mathcal{L}^{k_n} \to \mathcal{L}^{k_1} \otimes \mathcal{L}^{k_2} \otimes \ldots \otimes \mathcal{L}^{k_n} \equiv \mathcal{L}^{k_1+k_2+\ldots+k_n}.$$

This operation can be represented graphically as a vertex with n incoming edges from the subnetworks generating the inputs and a single outgoing edge. These FFN's are tree graphs.

One can also generalize the weights W from linear transformations to multilinear tensors. One version of this, used in quantum many-body physics, is the matrix product state (MPS). This is the vector $\mathrm{Tr}_L(T^d) \in \otimes^d V$ parameterized by a tensor $T \in \mathrm{Hom}(\mathrm{Hom}(L, L), V)$.

An analogous construction of a Kähler potential:

$$K^T \equiv \log \mathrm{Tr}\,(T_1 \cdot Z)(T_1 \cdot Z)^\dagger (T_2 \cdot Z)(T_2 \cdot Z)^\dagger \ldots (T_d \cdot Z)(T_d \cdot Z)^\dagger.$$

Tensor networks

The activation function $\theta(z) = z^2$ can be generalized to z^p and further generalized to a multilinear product

$$\theta_{\otimes} : \mathcal{L}^{k_1} \times \mathcal{L}^{k_2} \times \ldots \times \mathcal{L}^{k_n} \to \mathcal{L}^{k_1} \otimes \mathcal{L}^{k_2} \otimes \ldots \otimes \mathcal{L}^{k_n} \equiv \mathcal{L}^{k_1+k_2+\ldots+k_n}.$$

This operation can be represented graphically as a vertex with n incoming edges from the subnetworks generating the inputs and a single outgoing edge. These FFN's are tree graphs.

One can also generalize the weights W from linear transformations to multilinear tensors. One version of this, used in quantum many-body physics, is the matrix product state (MPS). This is the vector $\mathrm{Tr}_L\,(T^d) \in \otimes^d V$ parameterized by a tensor $T \in \mathrm{Hom}(\mathrm{Hom}(L, L), V)$.

An analogous construction of a Kähler potential:

$$K^T \equiv \log \mathrm{Tr}\,(T_1 \cdot Z)(T_1 \cdot Z)^\dagger (T_2 \cdot Z)(T_2 \cdot Z)^\dagger \ldots (T_d \cdot Z)(T_d \cdot Z)^\dagger.$$

Tensor networks

The activation function $\theta(z) = z^2$ can be generalized to z^p and further generalized to a multilinear product

$$\theta_{\otimes} : \mathcal{L}^{k_1} \times \mathcal{L}^{k_2} \times \ldots \times \mathcal{L}^{k_n} \to \mathcal{L}^{k_1} \otimes \mathcal{L}^{k_2} \otimes \ldots \otimes \mathcal{L}^{k_n} \equiv \mathcal{L}^{k_1+k_2+\ldots+k_n}.$$

This operation can be represented graphically as a vertex with n incoming edges from the subnetworks generating the inputs and a single outgoing edge. These FFN's are tree graphs.

One can also generalize the weights W from linear transformations to multilinear tensors. One version of this, used in quantum many-body physics, is the matrix product state (MPS). This is the vector $\mathrm{Tr}_L(T^d) \in \otimes^d V$ parameterized by a tensor $T \in \mathrm{Hom}(\mathrm{Hom}(L, L), V)$.

An analogous construction of a Kähler potential:

$$K^T \equiv \log \mathrm{Tr}\,(T_1 \cdot Z)(T_1 \cdot Z)^\dagger (T_2 \cdot Z)(T_2 \cdot Z)^\dagger \ldots (T_d \cdot Z)(T_d \cdot Z)^\dagger.$$

Hermitian Yang-Mills

In DKLR `hep-th/0606261`, Anderson *et al* `1004.4399` and subsequent works, a numerical approach to hermitian Yang-Mills was developed. Given a rank r vector bundle E over M, we consider $E(k) \equiv E \otimes \mathcal{L}^k$ such that M is embedded into $\mathrm{Gr}(r, N)$ by its holomorphic sections s_I^a. We can then define a family of metrics on $E(k)$ parameterized by a hermitian matrix $h^{I\bar{J}}$ as

$$(G^{-1})^{a\bar{b}} = \sum_{I,\bar{J}} h^{I\bar{J}}\, s_I^a\, \bar{s}_{\bar{J}}^{\bar{b}}.$$

The HYM equations are then

$$c \cdot \mathbf{1} = \omega^{i\bar{j}} F_{i\bar{j}} = \omega^{i\bar{j}} \bar{\partial}_{\bar{j}} \left(G^{-1} \partial_i G \right).$$

X. Wang (2005) proposed a corresponding balanced embedding

$$\frac{1}{\mathcal{N}} \int_{g\cdot M} s \left(s^\dagger s \right)^{-1} s^\dagger = \frac{r}{N} \mathbf{1}.$$

To define a FFN for this problem we need a nonlinear map from sections of $E(k)$ to those of $E(k')$. I don't see any direct analog of $\theta_{\mathbb{K}}$. Rather, we probably need to define a combined network which also describes a subset of sections of $\mathcal{L}^n$ for various n. We can then use the bilinear tensor product $E(k) \otimes \mathcal{L}^n \to E(k+n)$.

The bihomogeneous version of this would take as its basic variables $r \times r$ hermitian matrices which represent sections of $E(k) \otimes \bar{E}(k)$. The inputs would be $s_I^a \bar{s}_J^{\bar{b}}$, and the operation $\theta_{\otimes}$ for each layer would be separately linear in the outputs of the previous layer and the $\mathcal{L}^n$ network.
One could call this a ' network module" – the $\mathbb{K}^{D_i}$'s are replaced by $\mathbb{K}$-modules, and the nonlinear operations are replaced by $\mathbb{K}$ actions.

$$(G^{-1})^{a\bar{b}} = \sum_K \left(\sum_{I,\bar{J}} W_K^{(1),I\bar{J}} \, s_I^a \, \bar{s}_J^{\bar{b}} \right) \left(\sum_i W_{K,i\bar{j}}^{(2)} z^i \bar{z}^{\bar{j}} \right)$$

To define a FFN for this problem we need a nonlinear map from sections of $E(k)$ to those of $E(k')$. I don't see any direct analog of $\theta_{\mathbb{K}}$. Rather, we probably need to define a combined network which also describes a subset of sections of $\mathcal{L}^n$ for various n. We can then use the bilinear tensor product $E(k) \otimes \mathcal{L}^n \to E(k+n)$.

The bihomogeneous version of this would take as its basic variables $r \times r$ hermitian matrices which represent sections of $E(k) \otimes \bar{E}(k)$. The inputs would be $s_I^a \bar{s}_J^{\bar{b}}$, and the operation $\theta_{\otimes}$ for each layer would be separately linear in the outputs of the previous layer and the $\mathcal{L}^n$ network.
One could call this a ' network module" – the $\mathbb{K}^{D_i}$'s are replaced by $\mathbb{K}$-modules, and the nonlinear operations are replaced by $\mathbb{K}$ actions.

$$(G^{-1})^{a\bar{b}} = \sum_K \left(\sum_{I,\bar{J}} W_K^{(1),I\bar{J}} \, s_I^a \, \bar{s}_{\bar{J}}^{\bar{b}} \right) \left(\sum_i W_{K,i\bar{j}}^{(2)} Z^i \bar{Z}^{\bar{j}} \right)$$

Summary and conclusions

- We have written a Tensorflow/Keras package that can find numerical approximations to Ricci flat Kähler metrics to around 0.1% accuracy on the quintic hypersurface in $\mathbb{CP}^4$.
- Generalization to hypersurfaces in toric varieties, and to other scalar computations such as the spectrum of the Laplacian, is straightforward. We briefly discussed the HYM equations and network modules.
- Some results of broader interest for ML inspired numerical methods: dependence of accuracy on depth versus on total number of parameters; methods can work in the overparameterized regime.
- Tangentially related: very interesting work of Wigderson and collaborators applying geometric invariant theory and stability to problems in computational complexity theory, see `arXiv:1910.12375` and Wigderson's review *Operator scaling: theory, applications and connections*.