Machine learning for CY

Harold Erbin*[1]

[1]Arnold Sommerfeld Center for Theoretical Physics, Ludwig–Maximilians–Universität München, Theresienstraße 37, 80333 München, Germany

20th July 2020

Contents

*harold.erbin@physik.lmu.de

1 Calabi–Yau manifold

A Calabi–Yau (CY) manifold CY_n is a n-dimensional Kähler (and thus complex) manifold X with vanishing first Chern class.

1.1 General properties

General introduction can be found in [1, ch. 9, 2, sec. 3] (see also [3] for a more advanced and mathematical approach). Useful papers or more specific reviews include [4].

Spaces can be distinguished at different levels, in particular at the levels of the topological, differential and complex structures. Each concept if finer than the previous one, hence many CY are topologically equivalent but different as complex manifolds. Depending on the context, X will denote either a topological manifold, and thus a family of complex manifolds, either a specific member of the family. Hence, the (differential or) complex structure can be understood as parametrizing the precise shape of the manifold.

1.1.1 Differential geometry

Different properties of a compact CY are:

- $\mathrm{SU}(n)$ holonomy;
- Kähler;
- vanishing first Chern class $c_1(X)$;
- existence of a unique Ricci-flat metric;
- nowhere vanishing n-form;
- trivial canonical bundle.

For $n = 1$ the only compact CY manifold is the torus T^2 (the complex plane $\mathbb{C}$ is the only non-compact CY); for $n = 2$ the compact CY are the 4-torus $T^4 = T^2 \times T^2$ and the K_3 surface (non-compact CY are $\mathbb{C}^2$ and $\mathbb{C} \times T^2$).

The classification for $n \geq 3$ is not known; it is not even clear if the number of CY is finite. The simplest family for $n \geq 3$ of CY is given by the complete intersection CY, to be discussed below and the ones to be studied first. Typically only simply connected CY – at the exception of the n-torus T^n – will be considered.

There are a number of interesting topological quantities that one wishes to compute: the Euler number, the Hodge numbers, the Chern classes and the intersection numbers. They are described below.

The exterior derivative on the space $\Omega^n(X, \mathbb{R})$ of n-forms on X is denoted by d_n (note that here the spaces are defined as real spaces, which means that the maximal value of n is $\dim_{\mathbb{R}} X = 2 \dim_{\mathbb{C}} X$). The nth de Rham cohomology $H^n(X, \mathbb{R})$ corresponds to the cohomology of d_n

$$H^n(X, \mathbb{R}) = \frac{\ker d_n}{\operatorname{Im} d_{n-1}} \tag{1.1}$$

(this is the space of closed but non-exact forms) and its dimension is called the nth Betti number

$$b_n = \dim H^n(X, \mathbb{R}). \tag{1.2}$$

Equivalently, b_n can be interpreted as the number of zero eigenvalues for the Laplacian acting on n-forms. From this, it follows that

$$b_0 = 1 \tag{1.3}$$

since the only zero eigenvalue of the Laplacian acting on functions (0-form) is the constant function.

On a complex manifold the space of n-forms can be decomposed into the spaces of harmonic (p,q)-forms $H^{p,q}(X)$

$$H^n(X,\mathbb{C}) = \bigoplus_{p+q=n} H^{p,q}(X). \tag{1.4}$$

The corresponding dimensions are called the Hodge numbers

$$h_{p,q} = \dim H^{p,q}(X). \tag{1.5}$$

The Hodge numbers are symmetric (because $H^{p,q}$ and $H^{q,p}$ are complex conjugate) and – less trivially – they are related by reflection

$$h_{p,q} = h_{q,p}, \qquad h_{p,q} = h_{n-p,n-q}. \tag{1.6}$$

The Hodge and Betti numbers are related as

$$b_n = \sum_{p=0}^{n} h_{p,n-p}. \tag{1.7}$$

While in $n = 1$ (complex) dimensions the Hodge number is a topological invariant, this is not the case in higher dimension [5, 6, sec. 4.3]. For $n = 2$, this means that one can find two surfaces X and X' which are homeomorphic (but not diffeomorphic); the Hodge numbers are equal if the homeomorphism preserves the orientation defined by the complex structure. However, the Hodge numbers don't change under small deformations of the complex structure (i.e. two Kähler manifolds in small neighbourhood have the same Hodge numbers).

A simply-connected manifold has a trivial fundamental group (first homotopy group) and therefore a vanishing first homology group, which implies

$$h_{1,0} = h_{0,1} = 0. \tag{1.8}$$

Moreover, a compact connected Kähler manifold has

$$h_{0,0} = 1 \tag{1.9}$$

(constant functions). For a CY, one finds the additional property

$$h_{p,0} = h_{n-p,0}. \tag{1.10}$$

Deformations of the Kähler and complex structure are respectively counted by the Hodge numbers $h_{1,1}$ and $h_{n-1,1}$ [7, sec. 15.7.2]. Parameters corresponding to both deformations span the *moduli space*. CY with different Hodge numbers are not homeomorphic, and as a consequence they are distinct topological manifolds. The Hodge number $h_{1,1}$ counts the number of harmonic $(1,1)$-forms, one of which being the Kähler form. It is related to the number of 2-cycles in the surface, and hence it tends to increase as one consider fibres (and not direct products) of manifolds. This is related to the fact that to a basis of $(1,1)$-forms e_α ($\alpha = 1, \ldots, h_{1,1}$, one can associated a dual basis of cycles C_2^α such that

$$\int_{C_2^\alpha} e_\beta = \delta^\alpha_\beta. \tag{1.11}$$

In particular, $\mathbb{C}P^n$ and any hypersurface inside admits the Kähler form as its unique $(1,1)$-form, leading to $h_{1,1} = 1$.

Given a manifold X with curvature 2-form R, the Chern polynomials $c_k(R)$ are closed $2k$-forms defined by the expansion of

$$c = \det\left(1 + \frac{\mathrm{i}}{2\pi} R\right) \tag{1.12}$$

in powers of the curvature

$$c = \sum_{k=0}^{n} c_k. \tag{1.13}$$

The integration of the c_k over the manifold leads to analytic invariants: they are left invariant under smooth deformations of the complex structure. However, as for the Hodge numbers, they are not topological invariants.

The Euler number (or characteristic) is a topological important computed by the alternating sum of Betti numbers

$$\chi = \sum_{n\geq 0} (-1)^n b_n. \tag{1.14}$$

For $n = 3$ it corresponds to the integral of the third Chern class c_3.

The pth cohomology of a line bundle $\mathcal{L}$ over X and its dimension are respectively denoted by $H^p(X, \mathcal{L})$ and $h^p(X, \mathcal{L})$.

1.1.2 Projective spaces

A complex projective space $\mathbb{C}P^n$ (also denoted as $\mathbb{P}^n$) is the space of all (complex) lines in $\mathbb{C}^{n+1}$. The simplest way to build this space is to start with coordinates[1] $Z = (Z^0, \dots, Z^n)$ on $\mathbb{C}^{n+1} - \{0\}$ (called the homogeneous coordinates) and to perform the identification

$$(Z^0, \dots, Z^n) \sim (\lambda Z^0, \dots, \lambda Z^n), \qquad \forall \lambda \in \mathbb{C}^*. \tag{1.15}$$

Indeed, rescaling all coordinates of a given point amounts to move along a line. A standard patch is obtained by fixing $Z^0 \neq 0$ and by defining $z^i = Z^i/Z^0$ for $i = 1, \dots, n$. The manifold can be covered by $n+1$ such patches by permuting the coordinate in the denominator. The isometry group of $\mathbb{C}P^n$ is the projective linear group

$$\mathrm{PGL}(n, \mathbb{C}) = \frac{\mathrm{GL}(n, \mathbb{C})}{\mathbb{C}^*}, \tag{1.16}$$

and elements of these groups are called projective (linear) transformation or homography. Since the dimension of $\mathrm{GL}(n, \mathbb{C})$ is n^2, the dimension of $\mathrm{PGL}(n, \mathbb{C})$ is $n^2 - 1$ due to the quotient by scale transformation.

The pth cohomology of $\mathbb{C}P^n$ is

$$H^p(\mathbb{C}P^n) = \delta_{p,0}\, \mathbb{C}. \tag{1.17}$$

The tautological line bundle is denoted as $O_{\mathbb{P}^n}(-1)$: given a point $x \in \mathbb{C}P^n$, the fibre at that point corresponds to the line in $\mathbb{C}^{n+1}$ [8]. The tautological bundle is the dual of the hyperplane bundle $O_{\mathbb{P}^n}(1)$ (the index $\mathbb{P}^n$ is omitted when the space under consideration is clear). From this space one can consider the bundle $O_{\mathbb{P}^n}(a) \sim O_{\mathbb{P}^n}(1)^{\otimes a}$. The interpretation of these spaces is the following: a section $p_a(Z^0, \dots, Z^n) \in H^0\big(\mathbb{P}^n, O_{\mathbb{P}^n}(a)\big)$ corresponds to an homogeneous polynomial of degree a in the $n+1$ variables $(Z^0, \dots, Z^n)$. Hence the dimension of such a bundle is the number of monomials of order a in $(n+1)$ variables

$$h^0\big(\mathbb{P}^n, O_{\mathbb{P}^n}(a)\big) \equiv \dim H^0\big(\mathbb{P}^n, O_{\mathbb{P}^n}(a)\big) = \binom{n+a}{n} = \frac{(n+a)!}{n!\, a!}. \tag{1.18}$$

[1] Remember that coordinate indices are denoted at the top of the symbol, and the coordinate vector by the same symbol with any index.

More generally, the dimension of $H^p\big(\mathbb{P}^n, O_{\mathbb{P}^n}(a)\big)$ is given by the Bott formula [8, app. B]

$$h^p\big(\mathbb{P}^n, O_{\mathbb{P}^n}(a)\big) = \begin{cases} \binom{n+a}{n} & p = 0, a \geq 0 \\ \binom{-a-1}{-a-n-1} & p = n, a < -n \\ 0 & \text{otherwise} \end{cases} \tag{1.19}$$

Given a product of projective spaces

$$\mathcal{A} = \mathbb{C}P^{n_1} \times \cdots \times \mathbb{C}P^{n_m}, \tag{1.20}$$

a line bundle $\mathcal{L}$ will be a product of line bundles for each individual spaces

$$\mathcal{L} = O_{\mathcal{A}}(a^1, \ldots, a^m) = O_{\mathbb{P}^{n_1}}(a^1) \times \cdots \times O_{\mathbb{P}^{n_m}}(a^m). \tag{1.21}$$

The cohomology of the line bundle is given by the Künneth formula

$$H^p(\mathcal{A}, \mathcal{L}) = \bigoplus_{k_1 + \cdots k_m = p} H^{k_1}\big(\mathbb{P}^{n_1}, O_{\mathbb{P}^{n_1}}(a^1)\big) \times \cdots \times H^{k_m}\big(\mathbb{P}^{n_m}, O_{\mathbb{P}^{n_m}}(a^m)\big), \tag{1.22}$$

In particular, for $a^r \geq 0$, the only non-vanishing cohomology is $p = 0$ and has dimension

$$h^0\big(\mathcal{A}, \mathcal{L}\big) = \prod_{r=1}^{m} h^0\big(\mathbb{P}^{n_r}, O_{\mathbb{P}^{n_r}}(a^r)\big) = \prod_{r=1}^{m} \binom{n_r + a_r}{n_r}. \tag{1.23}$$

1.2 Physics

1.2.1 Superstring compactification – CY 3-fold

The following properties are of relevance to determine the low-energy physics [3, sec. 0.8]:

1. $\chi/2$ counts the number of fermion generations (preferred: $\chi = 6$ or a multiple thereof);
2. $h_{1,1}$ and $h_{2,1}$ count the number of generations and mirror generations (preferred: at least 2 for one of them, but not much bigger);
3. $\dim H^1(X_n, \operatorname{End} TX_n)$ counts the number of light neutral matter fields;
4. $\pi_1(X_n)$ leads to a breaking of the E_6 gauge group to a group closer to the standard model (preferred: at least $\mathbb{Z}_6$);
5. compute the Yukawa couplings;
6. compute the "normalisation matrices" (metric of the kinetic terms);
7. compute quantum and string corrections.

HE: this is a very rough list, to be improved later. ⇐ 1

The triple intersection number are defined as

$$\kappa_{\alpha\beta\gamma} = \int_X e_\alpha \wedge e_\beta \wedge e_\gamma. \tag{1.24}$$

They can be used to distinguish between different manifold (even when the Hodge numbers agree).

1.3 Complete intersection

A complete intersection Calabi–Yau n-fold (CICY$_n$) X is a n-dimensional Calabi–Yau manifold defined as the intersection of k hypersurfaces in a product of m projective spaces

$$\boxed{\mathcal{A} = \mathbb{C}P^{n_1} \times \cdots \times \mathbb{C}P^{n_m}} \tag{1.25}$$

called the ambient space. Fundamental papers on the topic are [9–11].

The motivation for considering such spaces is that they are compact and thus CY constructed from them will also be compact; on the other hand a theorem ensures that no Kähler submanifold of $\mathbb{C}^n$ can be compact [10] (this is different from the real cases, for example the 2-sphere S^2 can be embedded in $\mathbb{R}^3$).

1.3.1 General construction

Single equation A codimension 1 subspace X of $\mathbb{C}P^n$ ($\dim X = n-1$) can be obtained as a holomorphic homogeneous equation of degree a in the embedding coordinates $(Z^0, \dots, Z^n)$

$$p_a(Z^0, \dots, Z^n) = 0, \qquad p_a(\lambda Z^0, \dots, \lambda Z^n) = \lambda^a \, P(Z^0, \dots, Z^n). \tag{1.26}$$

The general form of such an equation is

$$p_{i_1 \cdots i_a} Z^{i_1} \cdots Z^{i_a} = 0 \tag{1.27}$$

where $p_{i_1 \cdots i_1}$ are the coefficients of the different terms. The equation must be homogeneous in order to be well-defined in $\mathbb{C}P^n$. Such a homogeneous equation of order a in $\mathbb{C}P^n$ is denoted by the equivalent notations

$$X = \mathbb{P}^n[a] = [\mathbb{P}^n \mid a] = [n \mid a]. \tag{1.28}$$

The advantage of the first two notations is to be valid also when considering intersection in more general spaces, as we will do in later sections.

Remember that X denotes a family of complex manifolds, one for each (inequivalent) choice of coefficients. Such a family of manifolds is also called a *configuration*, and as explained at the beginning we will use the symbol X both for the configuration and for any specific member, when there is no confusion. The question of the classification is discussed below.

A configuration X defines a CY manifold is the condition

$$n + 1 = a \tag{1.29}$$

holds. This can be found by considering a Fermat polynomial and computing the Ricci curvature using the Fubini–Study metric for the ambient space.

Multiple equations Additional equations decrease further the dimension of the resulting subspace (think to intersecting lines in $3d$: the result is a point) and one can thus start with a bigger projective space: a $(n-1)$-dimensional surface X can be obtained by considering k polynomial equations in $\mathbb{C}P^{n-1+k}$, each homogeneous of degree $(a_1, \dots, a_k)$

$$\begin{cases} p_{a_1}(Z_0, \dots, Z_n) = 0, \\ \quad \vdots \\ p_{a_k}(Z_0, \dots, Z_n) = 0. \end{cases} \tag{1.30}$$

The resulting surface is denoted by

$$X = \mathbb{P}^{n-1+k}[a_1, \ldots, a_k] = \left[\begin{array}{c|ccc} \mathbb{P}^{n-1+k} & a_1 & \cdots & a_k \end{array}\right] = \left[\begin{array}{c|ccc} n-1+k & a_1 & \cdots & a_k \end{array}\right] \tag{1.31}$$

It should be noted that linear equation (degree 1) simply reduces the dimension of the space since one can always solve for one of the coordinate in terms of the other:

$$\left[\begin{array}{c|cccc} \mathbb{P}^{n-1+k} & a_1 & \cdots & a_{k-1} & 1 \end{array}\right] = \left[\begin{array}{c|ccc} \mathbb{P}^{n-2+k} & a_1 & \cdots & a_{k-1} \end{array}\right] \tag{1.32}$$

The number of coefficients for a single monomial of order a and with $n+1$ variables describing X is given by (1.18) and equals C^n_{n+a}. However, not all the terms are independent as one can act with $\mathrm{PGL}(n, \mathbb{C})$ to redefine the coordinates and change n^2-1 coefficients. Additionally the overall scale is not fixed and one coefficient can be set to unity as a consequence. Hence the total number of free coefficients is $C^n_{n+a} - n^2$.

This counting hints to another interpretation of the polynomial equations describing X. Let's consider $\mathcal{A}$ together with a sum of line bundles $O(a_1) \oplus \cdots \oplus O(a_k)$. Then the the space X can be specified as the locus where a section $p \in H^0\big(\mathcal{A}, O(a_1) \oplus \cdots \oplus O(a_k)\big)$ vanishes:

$$X = p^{-1}(0). \tag{1.33}$$

Multiple equations in product spaces Finally, it is also possible to consider a product of projective spaces $\mathbb{C}P^{n_1} \times \cdots \mathbb{C}P^{n_m}$ and to consider k polynomial equations in all the coordinates (i.e. with crossed-terms). This method can be used to define the CICYs: the vanishing of the first Chern class will lead to constraint on the degree of homogeneity. Such a CICY_n is completely specified by a $m \times k$ configuration matrix, often denoted by the same symbol X as the space

$$\boxed{X = \left[\begin{array}{c|ccc} \mathbb{P}^{n_1} & a_1^1 & \cdots & a_k^1 \\ \vdots & \ddots & \vdots & \\ \mathbb{P}^{n_m} & a_1^m & \cdots & a_k^m \end{array}\right] = \left[\begin{array}{c|ccc} \boldsymbol{n} & \boldsymbol{a} \end{array}\right], \qquad a^r_\alpha \in \mathbb{N},} \tag{1.34}$$

where $\boldsymbol{n}$ and $\boldsymbol{a}$ denote respectively the vector of projective space dimensions and the matrix of degrees. The numbers a^r_α for fixed α are called the multi-degree of the equation α. The projective spaces and constraints are respectively indexed by

$$r = 1, \ldots, m, \qquad \alpha = 1, \ldots, k. \tag{1.35}$$

Sometimes one writes $\mathbb{C}P^{n_r}_{z_r}$ when one wants to specify both the space and its coordinates, and also the Euler and Hodge numbers can be written at the top or bottom right of the matrix. The entries obey the constraints

$$\boxed{\dim_{\mathbb{C}} X = n = \sum_{r=1}^{m} n_r - k,} \tag{1.36a}$$

$$\boxed{\forall r: \quad n_r + 1 = \sum_{\alpha=1}^{k} a^r_\alpha.} \tag{1.36b}$$

The first condition simply states that the dimension of the product of projective spaces (the sum of all n_r) minus the number of equations is equal to the dimension of the CY. The second set of constraints allows to compute the dimension n_r of the $\mathbb{C}P^{n_r}$ and as a consequence the first column of the matrix is redundant; nonetheless, it is often indicated

for clarity. This constraint is necessary in order to ensure the vanishing of the first Chern class (see [10, sec. 1] for a simple proof).

Concretely, denoting by Z_r the homogeneous coordinates of $\mathbb{C}P^{n_r}$, we are led to a system of k equations

$$\begin{cases} p^{(1)}_{a^1_1,\dots,a^m_1}(Z_1,\dots,Z_m)=0, \\ \quad\vdots \\ p^{(k)}_{a^1_k,\dots,a^m_k}(Z_1,\dots,Z_m)=0, \end{cases} \tag{1.37}$$

with the following scaling property for each equation

$$p^{(\alpha)}_{a^1_\alpha,\dots,a^m_\alpha}(\lambda\, Z_1,\dots,\lambda\, Z_m)=\left(\prod_{r=1}^{m}\lambda^{a^r_\alpha}\right)\, p_{a^1_\alpha,\dots,a^m_\alpha}(Z_1,\dots,Z_m). \tag{1.38}$$

The precise meaning of complete intersection is that the form

$$\Theta=\mathrm{d}p^{(1)}\wedge\cdots\wedge\mathrm{d}p^{(k)} \tag{1.39}$$

is nowhere vanishing on X.

To give a simple example, the configuration

$$X=\left[\begin{array}{c|ccc}\mathbb{P}^3_x & 3 & 0 & 1\\ \mathbb{P}^3_y & 0 & 3 & 1\end{array}\right] \tag{1.40}$$

corresponds to the system of polynomial equations

$$f_{abc}X^aX^bX^c=0,\qquad g_{\alpha\beta\gamma}Y^\alpha Y^\beta Y^\gamma=0,\qquad h_{a\alpha}X^aY^\alpha=0, \tag{1.41}$$

with $a,\alpha=0,1,2,3$ (remember that homogeneous coordinates are used in the equations).

The interest of the matrix notation, which forgets about the coefficients of the polynomials to keep only the powers, is that many properties of the manifold depend only on the latter and not on the former.

As for the case of a single projective space, it is possible to introduce a sum of line bundles

$$\mathcal{L}=\mathcal{L}_1\oplus\cdots\oplus\mathcal{L}_k,\qquad \mathcal{L}_\alpha=O_{\mathcal{A}}(a^1_\alpha,\dots,a^m_\alpha)=O_{\mathbb{P}^{n_1}}(a^1_\alpha)\times\cdots\times O_{\mathbb{P}^{n_m}}(a^m_\alpha). \tag{1.42}$$

As an aside, note that a space with

$$\sum_{\alpha=1}^{k}a^r_\alpha\leq n_r+1 \tag{1.43}$$

are called almost ample if there at least one strict inequality. Moreover, it is ample if the inequalities are strict for all r. Ample spaces in $d=2$ and $d=3$ are respectively called del Pezzo and Fano surfaces.

1.3.2 Classification

Then comes the question of classifying the CICY [11, 12].

First, there is the question of whether the solutions to the equations are regular or singular (two parallel lines have no intersection). Bertini's theorem ensures that, for a generic choice of coefficients, the spaces X are regular (one condition is that the gradient of the equations don't vanish in the locus of solutions), i.e. that it is smooth and of the

expected dimension [3, p. 29]. It was checked specifically in [9] that every configuration is not empty.

Every choice of coefficients defines a smooth manifold. Two manifolds described by the same configuration matrix but different polynomials are equivalent as real manifold (they are diffeomorphic) but they are different as complex manifolds. Hence, a given configuration matrix describes a deformation class, in which different complex manifolds are related by varying the complex structure; the latter is parametrized by $h_{n-1,1}$ complex parameters. The space of such manifolds (i.e. a configuration) has a single connected component [3].

Varying the coefficients of the polynomial equations leads to a modification of the complex structure, and thus one may try to count the number of different coefficients to derive $h_{n-1,1}$. However, this is not quite correct [10, 13]: not all complex structure deformations are polynomial, which means that not all CY of a given topological class can be described as a CICY, and in such cases $h_{n-1,1}$ will be higher than the number of polynomial coefficients. Conversely, there are instances where the number of polynomial coefficients is higher than $h_{n-1,1}$.

There is a small caveat when one says that X and X' equivalent representations: in general, this means that the families described by X and X' are not isomorphic but equivalent up to deformations. This relates to the comment above that polynomial deformations do not generically describe all possible deformations of the complex structure. A configuration is called *minimal* if it contains CY which do not appear in any lower-dimensional configuration.

There is a huge redundancy in the description and not all configuration matrices lead to independent (topological) manifolds, and this for two reasons:

1. the configuration is obviously independent under permutations of the rows and columns (change the order of the $\mathbb{C}P^n$ and constraints);
2. different sets of equations (for different projective spaces) can lead to the same manifold, and such instances can be found in two ways:
   (a) through ineffective splittings and contractions
   (b) by using identities between hypersurfaces.

Both are discussed below, but beforehand we describe how the size of the configuration matrix can be bounded.

The CICY$_n$ have been classified for $n \leq 4$ [10–12, 14]. Note that the datasets contain duplicates.

Bounding the matrices The relation (1.32) generalizes to the case of several projective spaces: if a polynomial is linear in the coordinates of one space and independent of the other spaces, then one can directly solve for this constraint and reduce the overall dimension. As a consequence, it is sufficient to consider matrices satisfying the condition

$$\boxed{\forall \alpha : \quad \sum_{r=1}^{m} a_\alpha^r \geq 2.} \tag{1.44}$$

This constraint limits severely the number of relevant matrices, and in fact make them finite in number [10, sec. 1]. Summing the Chern constraints (1.36b) over α and using the inequality leads to

$$\sum_{r=1}^{m}(n_r+1) = \sum_{r=1}^{m}\sum_{\alpha=1}^{k} a_\alpha^r \geq \sum_{\alpha=1}^{k} 2 = 2k. \tag{1.45}$$

Then using (1.36a) gives

$$2k \leq \sum_{r=1}^{m} n_r + m = n + k + m \tag{1.46}$$

and thus

$$\boxed{n + m \geq k.} \tag{1.47}$$

Let's denote respectively by f and F the numbers of $\mathbb{C}P^1$ and $\mathbb{C}P^n$ with $n \neq 1$, with the obvious relation

$$m = f + F, \tag{1.48}$$

and define the *over-dimension* by

$$\boxed{N_{\text{over}} \equiv \sum_{r=1}^{F} (n_i - 1) = n + k - f - F} \tag{1.49}$$

by rewriting (1.36a) as

$$n = \sum_{r=1}^{m} n_r - k = \sum_{r=1}^{F} n_r + f - k. \tag{1.50}$$

Using the inequality (1.47) leads to

$$N_{\text{over}} = n + k - f - F \leq n - f - F + n + m = 2n. \tag{1.51}$$

Finally, we have that $n_r \geq 2$ for the sum in N_{over}, and as a consequence every term of the sum contributes at least as 1, which leads to the following inequality on F:

$$\boxed{F \leq N_{\text{over}} \leq 2n.} \tag{1.52}$$

Hence there is a maximum of $2n$ projective space which are not $\mathbb{C}P^1$.

Another bound can be obtained for f:

$$\boxed{f \leq 3n.} \tag{1.53}$$

This uses the fact that a quadratic constraint acting on a single $\mathbb{C}P^1$ should also involve other variables, otherwise the matrix has a block diagonal form

$$\left[\begin{array}{c|cc} \mathbb{P}^1 & 2 & 0 \\ M & 0 & \boldsymbol{A} \end{array}\right]. \tag{1.54}$$

Indeed, the space $[1 \mid 2]$ corresponds to two points in $\mathbb{C}P^1$, and thus the above manifold corresponds to two copies of $[M \mid \boldsymbol{A}]$. Moreover, M can only be a CICY of lower dimension and the result space is a product and hence trivial (for $n = 3$, it can be a torus or a K_3 surface).

The number of projective spaces is then also bounded as

$$\boxed{m = f + F \leq 5n,} \tag{1.55}$$

Finally, it will also be convenient to define the excess number N_{ex} as [10, p. 501]

$$\boxed{N_{\text{ex}} = \sum_{r=1}^{m} (n_r + 1) - 2k = \sum_{r=1}^{m} n_r + m - 2k = \sum_{r=1}^{F} n_r + f + m - 2k = N_{\text{over}} + 2(m - k).} \tag{1.56}$$

This number vanishes when all columns sum to 2, which is the lowest value avoiding trivial constraints. A larger N_{ex} implies that there are many inequivalent configuration matrices for this ambient space.

HE: add bound on matrix coefficient ⇐ 2

Product spaces A first peculiar class corresponds to product spaces. In this case, the configuration matrix is block diagonal of the form

$$X = \left[\begin{array}{c|cc} \mathcal{A}_1 & \boldsymbol{M}_1 & 0 \\ \mathcal{A}_2 & 0 & \boldsymbol{M}_2 \end{array}\right]. \tag{1.57}$$

In this case, all the the properties of X descend from the two CY spaces $X_1 = [\mathcal{A}_1 \mid \boldsymbol{M}_1]$ and $X_2 = [\mathcal{A}_2 \mid \boldsymbol{M}_2]$. In particular, one has

$$\chi(X) = \chi(X_1)\chi(X_2). \tag{1.58}$$

Splitting and contractions Splitting and contraction are two inverse operations which consists in connecting different singularities by considering singular cases and blowing up or deforming the singularities. We first start by discussing these processes for CICY 3-folds, before commenting on higher-dimensional cases. This section is heavily based [10, sec. 3] for the 3-fold, see [12, 15] for the 4-folds.

Consider the following example

$$X = \left[\begin{array}{c|cc} \mathbb{P}^1_x & 1 & 1 \\ \mathbb{P}^2_y & 3 & 0 \\ \mathbb{P}^2_z & 0 & 3 \end{array}\right], \tag{1.59}$$

and write the system of equations as

$$X^0 P_1(Y) + X^1 P_2(Y) = 0, \qquad X^0 Q_1(Z) + X^1 Q_2(Z) = 0, \tag{1.60}$$

where the P_i and Q_i are cubic polynomials. Because $(X^0, X^1) = (0,0) \notin \mathbb{P}^1$, the above system of equations admits a solution if the determinant of

$$\Delta(Y,Z) = P_1(Y)Q_2(Z) - P_2(Y)Q_1(Z) \tag{1.61}$$

vanishes. This leads to a polynomial equation of degree $(3,3)$ and one could be tempted to identify the space as

$$X' = \left[\begin{array}{c|c} \mathbb{P}^2 & 3 \\ \mathbb{P}^2 & 3 \end{array}\right], \tag{1.62}$$

(the fact that several columns have been contracted explains the name). But, this is not correct, as can be seen from the fact that both manifolds have different Euler characteristics ($\chi = 0$ and $\chi = -168$). The reason is that $\Delta(Y,Z)$ is not the most general polynomial described by this matrix (37 coefficients against 10^2, since a cubic has 10 independent coefficients), but also that it does not describe a regular space. Indeed, its gradient $\mathrm{d}\Delta$ can vanish in the solution set of $\Delta = 0$, and thus it does not describe a complete intersection. Here, this happens when each polynomial vanishes independently

$$P_i(Y) = 0, \qquad Q_i(Z) = 0. \tag{1.63}$$

These are four equations in a four-dimensional space, and the resulting solutions are points, described by the space

$$X^{\#} = \left[\begin{array}{c|cccc} \mathbb{P}^2 & 3 & 3 & 0 & 0 \\ \mathbb{P}^2 & 0 & 0 & 3 & 3 \end{array}\right] \tag{1.64}$$

These singular points can be removed by deforming the polynomial equation by a small generic polynomial of the same degree

$$\Delta \to \Delta + \epsilon\delta' \tag{1.65}$$

This has the effect of changing the Euler characteristics

$$\chi = \chi' + 2\nu, \tag{1.66}$$

where ν is the number of singular points (in the example there are 3^4 such points). The inverse process, called a splitting, involves removing the singularity by blowing up the singularities using $\mathbb{C}P^1$ spaces: this introduces new columns, which explains the name. If there are no singular points, $\nu = 0$, then both Euler numbers are equal $\chi = \chi'$ and one can show that both configurations are in fact equivalent (and thus describe the same family of manifolds);[2] such splittings and contractions are called *ineffective*. Conversely, a process for which $\chi \neq \chi'$ is called effective.

An example of an ineffective splitting is

$$X = \left[\begin{array}{c|ccc} \mathbb{P}^1 & 1 & 1 & 0 \\ \mathbb{P}^1 & 1 & 1 & 0 \\ \mathbb{P}^2 & 1 & 1 & 1 \\ \mathbb{P}^2 & 0 & 0 & 3 \end{array}\right] = \left[\begin{array}{c|cc} \mathbb{P}^1 & 2 & 0 \\ \mathbb{P}^2 & 2 & 1 \\ \mathbb{P}^2 & 0 & 3 \end{array}\right] \tag{1.67}$$

because the set of singular points is empty

$$\left[\begin{array}{c|ccccc} \mathbb{P}^1 & 1 & 1 & 1 & 1 & 0 \\ \mathbb{P}^1 & 1 & 1 & 1 & 1 & 1 \\ \mathbb{P}^2 & 0 & 0 & 0 & 0 & 3 \end{array}\right] = \emptyset. \tag{1.68}$$

The reason is that there are 4 equations for a 3-dimensional space $\mathbb{P}^1 \times \mathbb{P}^2$, and the system is generically over-constrained.

In general, one can use the contraction (left to right) and splitting (right to left) to relate the following configurations

$$\left[\begin{array}{c|ccc} \mathbb{P}^1 & 1 & 1 & 0 \\ M & \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{A} \end{array}\right] = \left[\begin{array}{c|cc} M & \boldsymbol{a}+\boldsymbol{b} & \boldsymbol{A} \end{array}\right], \tag{1.69}$$

where a and b are arbitrary row vectors, A a matrix, and M a product of projective spaces. This can be generalized to an arbitrary number of $\mathbb{C}P^1$

$$\left[\begin{array}{c|ccccc} \mathbb{P}^1 & 1 & \cdots & 1 & 0 \\ M & \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_N & \boldsymbol{A} \end{array}\right] = \left[\begin{array}{c|cc} M & \sum_{i=1}^{N} \boldsymbol{a}_i & \boldsymbol{A} \end{array}\right], \tag{1.70}$$

The condition above to distinguish between effective and ineffective splitting works only for CICY 3-folds. The reason is that the singular locus is zero-dimensional and hence a set of points. However, for a 4-fold, the locus can be 1-dimensional and hence have the topology of a torus, which has a vanishing Euler number [12, sec. 4]. In this case, the Euler number can be invariant even if the spaces are topologically inequivalent (they will be distinguished by other quantities). A necessary and sufficient condition for a $\mathbb{P}^1$ splitting described in [12, sec. 4] is that the volume of the singular manifold vanishes (in which case it is empty). For a $\mathbb{P}^n$ splitting they obtain a condition on a set of (non-CY) 3-folds splitting.

Identities Redundancies happen when two different systems of equations in two different ambient spaces describe in fact the same space. In this paragraph, we describe some of these identities that lead to redundancies in the configuration matrices. Since the identities become very wild as the dimension increases, we will generically be interested identities involving small factors.

[2] Remember the caveat indicated above: while the configuration are equivalent, it is possible that they are not isomorphic and that one describes more manifolds.

Due to the constraints (1.36b), it is sufficient to study identities involving only ample spaces, that is configuration satisfying (1.43)

$$\sum_{\alpha=1}^{k} a_\alpha^r \leq n_r + 1. \tag{1.71}$$

Identities are classified according to the dimension of the corresponding space. They are first given for the space itself, and then the rule for mapping the rest of the configuration matrix is given; the latter can be derived by finding a change of variables (in practice, it was found by an inspection of the dataset) or by following a chain of contractions and splittings.

There are three identities involving 1-fold, which correspond to equivalent ways to write $\mathbb{C}P^1$

$$\left[\begin{array}{c|c} \mathbb{P}^1 & 1 \\ \mathbb{P}^1 & 1 \end{array}\right] = \mathbb{P}^1 \Longrightarrow \left[\begin{array}{c|cc} \mathbb{P}^1 & 1 & \boldsymbol{a} \\ \mathbb{P}^1 & 1 & \boldsymbol{b} \\ M & 0 & \boldsymbol{A} \end{array}\right] = \left[\begin{array}{c|c} \mathbb{P}^1 & \boldsymbol{a}+\boldsymbol{b} \\ M & \boldsymbol{A} \end{array}\right], \tag{1.72a}$$

$$\left[\begin{array}{c|c} \mathbb{P}^1 & 2 \\ \mathbb{P}^1 & 1 \end{array}\right] = \mathbb{P}^1 \Longrightarrow \left[\begin{array}{c|cc} \mathbb{P}^1 & 2 & 0 \\ \mathbb{P}^1 & 1 & \boldsymbol{a} \\ M & 0 & \boldsymbol{A} \end{array}\right] = \left[\begin{array}{c|c} \mathbb{P}^1 & 2\boldsymbol{a} \\ M & \boldsymbol{A} \end{array}\right], \tag{1.72b}$$

$$\left[\begin{array}{c|c} \mathbb{P}^2 & 2 \end{array}\right] = \mathbb{P}^1 \Longrightarrow \left[\begin{array}{c|cc} \mathbb{P}^2 & 2 & \boldsymbol{a} \\ M & 0 & \boldsymbol{A} \end{array}\right] = \left[\begin{array}{c|c} \mathbb{P}^1 & 2\boldsymbol{a} \\ M & \boldsymbol{A} \end{array}\right], \tag{1.72c}$$

The first identity means that a bilinear constraint on $\mathbb{C}P^1 \times \mathbb{C}P^1$ reduces to $\mathbb{C}P^1$.

There are 7 identities involving 2-folds. Several identities are given for 3-folds and more, but there are no generic classification. We refer to [10, sec. 4, app.] for a list, and we only report the generalization of the rule that a linear constraint decreases the overall dimension

$$\left[\begin{array}{c|c} \mathbb{P}^1 & 1 \\ \mathbb{P}^n & 1 \end{array}\right] = \mathbb{P}^1 \times \mathbb{P}^{n-1} \Longrightarrow \left[\begin{array}{c|cc} \mathbb{P}^1 & 1 & \boldsymbol{a} \\ \mathbb{P}^n & 1 & n\boldsymbol{b} \\ M & 0 & \boldsymbol{A} \end{array}\right] = \left[\begin{array}{c|c} \mathbb{P}^1 & \boldsymbol{a}+\boldsymbol{b} \\ \mathbb{P}^{n-1} & \boldsymbol{b} \\ M & \boldsymbol{A} \end{array}\right]. \tag{1.73}$$

Since not all identities commute with splitting, which was a key step in generating the configurations, only the identities which do commute are considered in [12] to reduce the list of 4-folds.

Favourable representation Defining the sets

$$D_\alpha = \{r \mid a_\alpha^r > 0\}, \tag{1.74}$$

a configuration is called favourable in [9, sec. 4, 3, p. 66] if the constraints can be ordered such that

$$\alpha' < \alpha \quad \Longrightarrow \quad \begin{cases} D_{\alpha'} \cap D_\alpha = \emptyset \\ D_{\alpha'} \subset D_\alpha \end{cases} \tag{1.75}$$

The condition can be rephrased as either $a_\alpha^r a_{\alpha'}^r = 0$ or either $a_{\alpha'}^r \leq a_\alpha^r a_{\alpha'}^r$ for all r [11, sec. 2]. A CY 3-fold in such a configuration is 1) simply connected and 2) has its second Betti number at least equal to the number of $\mathbb{C}P^n$, i.e. $b_2 \geq m$ if the configuration is minimal.

However, this definition is not the same as in [14] because all favourable configurations in their dataset have $h_{1,1} = b_2 = m$, while [9, p. 107] gives examples that have $b_2 > m$ (e.g. the manifold with $h_{1,1} = 6$ and $\chi = -96$ is described by different configuration matrices in both papers). For this reason, we call the previous condition "weakly-favourable".

A *favourable* representation, as defined in [14, sec. 2], is a configuration for which all the divisors (or equivalently the Picard group) of X descend from the ambient space $\mathcal{A}$. A

sufficient (but not necessary) condition is that the line bundle cohomologies $H^p(X, \mathcal{L})$ are non-vanishing only for $p = 0$ (note that the cohomologies are over X and not over $\mathcal{A}$, so the Künneth formula does not apply directly). The technology of exact sequences allow to compute the cohomology of this space from $H^p(\mathcal{A}, \mathcal{L})$ [11, sec. 3.3], and one finds

$$h^0\big(X, \mathcal{L}_\alpha|_X\big) = \dim H^0\big(\mathcal{A}, \mathcal{L}_\alpha\big) - \dim H^0(\mathcal{A}) = \prod_{r=1}^{m} \binom{n_r + a_r}{a_r} - 1, \tag{1.76a}$$

which can be found from (1.23) and using an exact sequence. Then the dimension $h^i(\mathcal{A}, \mathcal{L})$ is simply the sum of all the $h^i(\mathcal{A}, \mathcal{L}_\alpha)$.

If the configuration is (non-weakly) favourable, then the first Hodge number equals the number of projective spaces [15, sec. 2.2]

$$h_{1,1} = m. \tag{1.77}$$

A *Kähler favourable* representation is a matrix for which the Kähler and Mori cones directly descend from the ambient space $\mathcal{A}$. In some cases, the configuration can be made favourable with respect to a different ambient space $\mathcal{A}'$ (for example a product of del Pezzo surfaces).

Fibrations A manifold X is the fibration of a fibre $\mathcal{A}_1[\mathcal{F}]$ over a base $\mathcal{A}_2[\mathcal{B}]$ with twist $\mathcal{T}$ if the configuration matrix can be written as

$$X = \left[\begin{array}{c|cc} \mathcal{A}_1 & 0 & \mathcal{F} \\ \mathcal{A}_2 & \mathcal{B} & \mathcal{T} \end{array}\right]. \tag{1.78}$$

The condition (1.36b) always imply that $\mathcal{A}_1[\mathcal{F}]$ is a CICY. Hence if $\mathcal{A}_1[\mathcal{F}]$ is 1-dimensional, it is a genus-1 curve (topologically equivalent to the torus), if it is 2-dimensional it is a K_3 surface.

In general, a CICY admits multiple fibrations, and the ones that can be seen by a direct inspection of the configuration matrix are called "obvious fibrations". The favourable representation is the one where most fibrations are obvious. However, a characterization relying only the matrix will necessarily be incomplete since it depends on the specific algebraic description chosen.

Conventions In order to partially fix the column and row orders, the matrix components are written in a lexicographic order (starting at 0).

1.3.3 1-fold

The 2-torus T^2 is equivalent to

$$T^2 = \mathbb{P}^2[3]. \tag{1.79}$$

All the torus are homeomorphic. The identities between Hodge numbers imply:

$$h_{00} = h_{11} = h_{01} = h_{10} = 1. \tag{1.80}$$

This is possible because the torus is not simply connected.

1.3.4 2-fold

The K_3 is the simplest example of a CICY manifold (and, as a consequence, all simply connected CY_2 are also CICY). The configuration matrix of K_3 is

$$K_3 = \mathbb{P}^3[4], \tag{1.81}$$

leading to quartic polynomial equations. The simplest such equation is

$$\sum_{I=0}^{3}(Z^I)^4 = 0. \tag{1.82}$$

All K_3 surfaces are homeomorphic.

The only non-trivial Hodge number is

$$h_{1,1} = 20 \tag{1.83}$$

such that $\chi = 24$. Note that no explicit K_3 metric is known.

1.3.5 3-fold

According to the different constraints, the relevant Hodge numbers are $h_{1,1}$ and $h_{2,1}$. Moreover we find that

$$h_{2,0} = h_{1,0} = 0 \tag{1.84}$$

using the relation (1.10) for CYs and (1.8), such that the second Betti number (1.7) equals $h_{1,1}$

$$b_2 = h_{1,1}. \tag{1.85}$$

The Euler number (1.14) of a CY_3 reads

$$\chi = 2(h_{11} - h_{21}). \tag{1.86}$$

The numbers f and F of $\mathbb{C}P^1$ and $\mathbb{C}P^{n\neq 1}$ can be shown to satisfy the following bounds [10, sec. 1, 9]

$$f \leq 9, \qquad F \leq 6, \tag{1.87}$$

which directly implies a bound on the total number of $\mathbb{C}P^n$ and on the number of constraints

$$m \leq 15, \qquad k \leq 18. \tag{1.88}$$

The inequality $f \leq 9$ follows by requiring that the configuration does not describe a product space and that there are no linear constraints.

For a CICY defined by a single constraint, one has [11, p. 114]

$$h_{2,1} = \prod_{r=1}^{m}\binom{2n_r+1}{n_r} - 1 - \sum_{r=1}^{m} n_r(n_r+2). \tag{1.89}$$

For additional references on computing the topological properties, see [16, app. C], and most particularly [17, app.].

HE: give formulas for the Hodge numbers, or at least explain how to compute ⇐ 3

HE: explain how to compute the other topological quantities ⇐ 4

Examples The simplest CICY (ID 7890) is called the *quintic* and defined by

$$\mathbb{P}^4[5], \tag{1.90}$$

and a representative polynomial is

$$\sum_{I=0}^{4}(Z^I)^5 = 0. \tag{1.91}$$

Its Hodge numbers are $h_{1,1} = 1$ and $h_{2,1} = 101$. Note that it is the worst outlier of the dataset: it has the biggest $h_{2,1}$ (the next biggest one being only 89).[3]

The above equation defines just one CY as a differential manifold. Other topologically equivalent manifolds can be obtained by modifying the monomials appearing in the equation. There are $C^4_{4+5} = 126$ degree 5 monomials in 5 variables, from which $5^2 - 1$ can be removed by acting with $\mathrm{PGL}(5, \mathbb{C})$, and one additional from the freedom of changing the overall scale of the equation. This leads to a total of $101 = h_{2,1}$ possible coefficients: this means that all inequivalent complex structures can be obtained by polynomial deformations in this case.

There four other CICY which can be defined from a single projective space (ID 7889, 7878, 7879, 7861):

$$\mathbb{P}^5[4,2], \qquad \mathbb{P}^5[3,3], \qquad \mathbb{P}^6[2,2,3], \qquad \mathbb{P}^7[2,2,2,2]. \tag{1.92}$$

They are also outliers: they all have $h_{1,1} = 1$ and $h_{2,1} = 89, 73, 73, 65$. In particular, $\mathbb{P}^5[4,2]$ is the only CICY with $h_{2,1} = 89$ (the next $h_{2,1}$ below 101).

There are 44 CICY defined as hypersurfaces of 2 projective spaces. Several of them are outliers, for example the only four CICY with $h_{2,1} = 86$ (the next $h_{2,1}$ below 89). One simple example (ID 7887) is

$$\left[\begin{array}{c|c} \mathbb{P}^1 & 2 \\ \mathbb{P}^3 & 4 \end{array}\right]. \tag{1.93}$$

Datasets General properties of the datasets:

- number of configurations: 7890
- number of product spaces: 22
- $a^r_\alpha \in [0, 5]$
- $h_{11} \in [0, 19]$, 18 distinct values
- $h_{21} \in [0, 101]$, 65 distinct values
- $\chi \in [-200, 0]$, 70 distinct values
- number of Hodge number combinations: 266
- number of CICY with at least one obvious genus-1 fibration: 7837
- average number of fibrations per CICY: 9.85

A method to compute all useful properties have been presented in [14], but not all results have been released in the dataset (for example the intersection numbers). The motivation for finding a new representation of the matrices is that many properties cannot be computed with standard tools if the matrix is not favourable.

Two different datasets have been released:

- CDLS (Candelas–Dale–Lutken–Schimmrigk) [10, 11]
  - configuration matrices: 12×15
  - number of favourable matrices (excluding product spaces): 4874
  - number of non-favourable matrices (excluding product spaces): 2994

[3] By outliers, we mean a CY with an extremely low and/or high values of the Hodge numbers, since there are very few such cases. However, these CY admit a favourable description and are thus not outliers from the point of view of the linear regression to predict h_{11}.

  - number of different ambient spaces: 235

- AGGL (Anderson–Gao–Gray–Lee) [14]
  - configuration matrices: 15×18
  - number of favourable matrices (excluding product spaces): 7820
  - number of non-favourable matrices (excluding product spaces): 48
  - number of Kähler favourable matrices: 4874
  - number of Kähler favourable matrices with respect to two almost del Pezzo surfaces: 83
  - number of different ambient spaces: 126

Note that the size of the configuration matrices in the AGGL set saturates the bound of 15×18.

There is a large redundancy in the above list, see in particular and [10, 14, 16]:

- in [10] it is noted that the identities not already used while generating the dataset reduces it by circa 20%;
- in [16, app. C.3] a smaller number of 435 redundant CICY is indicated.

1.3.6 4-fold

According to the different constraints, the relevant Hodge numbers are: $h_{1,1}$, $h_{2,1}$, $h_{3,1}$ and $h_{2,2}$. In fact, there is an additional relation between the Hodge numbers:

$$h_{2,2} = 2\,(22 + 2h_{1,1} + 2h_{3,1} - h_{2,1}). \tag{1.94}$$

The numbers $h_{1,1}$ and $h_{3,1}$ count respectively the possible deformations of the Kähler and complex structure.

The Euler number (1.14) of a CY_4 reads

$$\chi = 6\,(8 + h_{1,1} + h_{3,1} - h_{2,1}). \tag{1.95}$$

The formulas for the Chern classes and Euler number are given in [12, sec. 3, 15, sec. 2.1]. The Hodge numbers are computed in [15, sec. 2.2], where the implications of working with a favourable representation are also discussed.

The bounds on the number of $\mathbb{C}P^1$ and $\mathbb{C}P^{n>1}$ are

$$f \leq 12, \qquad F \leq 8. \tag{1.96}$$

Datasets General properties of the datasets [14]:

- number of configurations: 921 497
- number of product spaces: 15 813
- $h_{11} \in [1, 24]$, $h_{21} \in [0, 33]$, $h_{31} \in [20, 426]$, $h_{22} \in [204, 1752]$
- $\chi \in [0, 2610]$ (and $\chi \geq 288$ for non-product spaces)
- number of different χ: 206
- number of different ambient spaces: 660

- lower bound on topologically distinct manifolds: 36 779
- configuration matrices: 16×20
- number of CICY with at least one obvious genus-1 fibration: 921 420
- average number of fibrations per CICY: 54.6

1.4 Generalized complete intersection

These manifolds generalize the CICY and were proposed in [18]. The basic idea is to allow negative integers in the configuration matrices: indeed, which, after solving for part of the polynomial equations, become also polynomial equations.

1.5 Weighted projective spaces

1.6 Toric geometries

2 Machine learning

Our philosophy is to use machine learning to recover as much information as possible from a subset of the CICY data. In particular, we don't assume any knowledge of mathematical formulas. The idea is that if we were working with a dataset for which the general formulas are not known, then we should be able to recover them (or at least the machine should).

If it is necessary to write an algorithm to check some formulas, the best place to start with are [**Constantin:2018:FormulaeLineBundle**, 8–11, 14, 19].

2.1 Summary of questions

In the previous section we have described the structure of (CI)CY and discussed their properties. Here is a list of the interesting questions:

1. CY
   (a) find mirror pairs
   (b) study fibrations
   (c) compute the Hodge numbers
2. CICY
   (a) find duplicates in the list
   (b) find the favourable representation (unsupervised?)
   (c) (unsupervised) representation learning (PCA, autoencoder...): can the machine comes up with topological quantities to characterize the space in a more efficient way?
   (d) generate configuration matrices, check algorithmically whether they are correct
   (e) check if a network train on the 3-folds can be useful for the 4-folds
   (f) find identities for n-folds [10, sec. 4]
3. CICY_3
   (a) check if χ is divisible by 6
   (b) check how the network performs on the non-favourable representation
   (c) study redundancies (see [10, 14] and the `redun` column of the CDLS dataset)
4. CICY_4
   (a) find the relation between the Hodge numbers
   (b) study redundancies [12, p. 12]
5. CICY_5 (these spaces could be interesting for studying compactifications to $d = 0$ and $d = 1$)
   - build the dataset (see [20] for some discussion)

2.2 Feature engineering

Here is a list of the possible interesting features (some are already in the dataset, the other are generated):

- `is_prod`: if the CY is a direct product;
- `favour`: if the CY is a direct product;

2.3 Hodge numbers

In this section we want to predict the Hodge numbers.

Contrary to [21], we are learning both Hodge numbers and not only h_{11} for CICY$_3$. This is in agreement with our philosophy of not assuming knowledge of mathematical formulas (and in particular of the "simple combinatorial formula" for the Euler characteristics). Moreover, learning h_{11} is fairly simple since its range is limited and this does not provide a good test for how good machine learning can be applied to more general cases. Already for CICY$_4$ one has 4 Hodge numbers, and even if we assume the most complicated one to be fixed by some a priori knowledge of the Euler characteristics, the remaining three are complicated (more than h_{21}). For this reason it is useful to learn also h_{21}.

We tackle the prediction of the Hodge numbers as a regression problem. This also follows from our general philosophy: for a more general problem where the complete set is not known, it is also likely that the Hodge number range will be unknown. In particular it is not consistent to assume as in [21] that any "large set" will contain the highest Hodge number. For example, only 15 out of the 7890 (0.19%) CICY$_3$ have $h_{11} = 19$.

The other reason is that a classification leads to a proliferation of useless classes when considering several Hodge numbers. Assuming that the maximal value of each Hodge number is known, but not the allowed values (for example there is now CICY$_3$ with $h_{11} = 18$), then one finds a huge numbers of classes. For CICY$_3$, 2040 combinations are possible but only 266 are realised in practice. Assuming the knowledge of the latter classes would also go against the general philosophy, and assuming nothing renders the classification problem harder.

The last argument for seeing this as a regression is that the Hodge numbers have very different ranges. It is then useful to normalise the data (shifting by the mean value and dividing by the variance), which practically turns the Hodge numbers to real numbers, for which regression is naturally adapted. Then regression can naturally go beyond the maximal value found in the training data and is thus expected to generalize better.

3 Comments

Questions:

1. The study for $d = 3$ is performed with the "favourable" representation. How does the neural network trained with this set behaves with the older representation?

   Another way to say this is to find if the network learned the invariance under permutations of the rows and columns.

2. If outliers are excluded while learning, what gives the neural network predictions for them?

3. Is the favourable representation of a CICY unique up to permutations of row and columns? (In order to find the duplicates in the list.)

Acknowledgements

The work of H.E. is conducted under a Carl Friedrich von Siemens Research Fellowship of the Alexander von Humboldt Foundation for postdoctoral researchers.

References

[1] K. Becker, M. Becker and J. H. Schwarz. *String Theory and M-Theory: A Modern Introduction*. 1st edition. Cambridge University Press, Dec. 2006.

[2] L. B. Anderson and M. Karkheiran. 'TASI Lectures on Geometric Tools for String Compactifications' (Apr. 2018). arXiv: 1804.08792.

[3] T. Hübsch. *Calabi-Yau Manifolds: A Bestiary For Physicists*. English. Wspc, Mar. 1992.

[4] R. Davies. 'The Expanding Zoo of Calabi-Yau Threefolds'. *Advances in High Energy Physics* 2011 (2011), pp. 1–18.
DOI: 10.1155/2011/901898. arXiv: 1103.3156.

[5] F. Campana. *Une Remarque Sur Les Nombres de Hodge Des Variétés Complexes Projectives*. July 2004.
URL: http://www.iecl.univ-lorraine.fr/~Pierre-Yves.Gaillard/DIVERS/hodgenumbers.pdf.

[6] D. McDuff and D. Salamon. *Introduction to Symplectic Topology*. English. 3rd edition. Oxford University Press, July 2017.

[7] M. B. Green, J. H. Schwarz and E. Witten. *Superstring Theory: Loop Amplitudes, Anomalies and Phenomenology*. English. Vol. 2. Cambridge University Press, July 1988.

[8] L. B. Anderson, J. Gray, A. Lukas and B. Ovrut. 'Vacuum Varieties, Holomorphic Bundles and Complex Structure Stabilization in Heterotic Theories'. en. *Journal of High Energy Physics* 2013.7 (July 2013), p. 17.
DOI: 10.1007/JHEP07(2013)017. arXiv: 1304.2704.

[9] P. Green and T. Hübsch. 'Calabi-Yau Manifolds as Complete Intersections in Products of Complex Projective Spaces'. en. *Communications in Mathematical Physics* 109.1 (Mar. 1987), pp. 99–108.
DOI: 10.1007/BF01205673.

[10] P. Candelas, A. M. Dale, C. A. Lütken and R. Schimmrigk. 'Complete Intersection Calabi-Yau Manifolds'. *Nuclear Physics B* 298.3 (Mar. 1988), pp. 493–525.
DOI: 10.1016/0550-3213(88)90352-5.

[11] P. S. Green, T. Hübsch and C. A. Lütken. 'All the Hodge Numbers for All Calabi-Yau Complete Intersections'. en. *Classical and Quantum Gravity* 6.2 (1989), p. 105.
DOI: 10.1088/0264-9381/6/2/006.

[12] J. Gray, A. S. Haupt and A. Lukas. 'All Complete Intersection Calabi-Yau Four-Folds'. *Journal of High Energy Physics* 2013.7 (July 2013).
DOI: 10.1007/JHEP07(2013)070. arXiv: 1303.1832.

[13] P. Green and T. Hübsch. 'Polynomial Deformations and Cohomology of Calabi-Yau Manifolds'. en. *Communications in Mathematical Physics* 113.3 (Sept. 1987), pp. 505–528.
DOI: 10.1007/BF01221257.

[14] L. B. Anderson, X. Gao, J. Gray and S.-J. Lee. 'Fibrations in CICY Threefolds'. *Journal of High Energy Physics* 2017.10 (Oct. 2017).
DOI: 10.1007/JHEP10(2017)077. arXiv: 1708.07907.

[15] J. Gray, A. S. Haupt and A. Lukas. 'Topological Invariants and Fibration Structure of Complete Intersection Calabi-Yau Four-Folds'. *Journal of High Energy Physics* 2014.9 (Sept. 2014).
DOI: 10.1007/JHEP09(2014)093. arXiv: 1405.2073.

[16] L. B. Anderson, Y.-H. He and A. Lukas. 'Monad Bundles in Heterotic String Compactifications'. *Journal of High Energy Physics* 2008.07 (July 2008), pp. 104–104.
DOI: 10.1088/1126-6708/2008/07/104. arXiv: 0805.2875.

[17] L. B. Anderson, Y.-H. He and A. Lukas. 'Heterotic Compactification, An Algorithmic Approach'. *Journal of High Energy Physics* 2007.07 (July 2007), pp. 049–049.
DOI: 10.1088/1126-6708/2007/07/049. arXiv: hep-th/0702210.

[18] L. B. Anderson, F. Apruzzi, X. Gao, J. Gray and S.-J. Lee. 'A New Construction of Calabi-Yau Manifolds: Generalized CICYs'. *Nuclear Physics B* 906 (May 2016), pp. 441–496.
DOI: 10.1016/j.nuclphysb.2016.03.016. arXiv: 1507.03235.

[19] R. Blumenhagen, B. Jurke, T. Rahn and H. Roschy. 'Cohomology of Line Bundles: A Computational Algorithm'. *Journal of Mathematical Physics* 51.10 (Oct. 2010), p. 103525.
DOI: 10.1063/1.3501132. arXiv: 1003.5217.

[20] A. S. Haupt, A. Lukas and K. S. Stelle. 'M-Theory on Calabi-Yau Five-Folds'. *Journal of High Energy Physics* 2009.05 (May 2009), pp. 069–069.
DOI: 10.1088/1126-6708/2009/05/069. arXiv: 0810.2685.

[21] K. Bull, Y.-H. He, V. Jejjala and C. Mishra. 'Machine Learning CICY Threefolds'. *Physics Letters B* 785 (Oct. 2018), pp. 65–72.
DOI: 10.1016/j.physletb.2018.08.008. arXiv: 1806.03121.