Expectations and the Exceptional Divisor

Exceptional Expectations

Benjamin Gerraty & Daniel Murfet

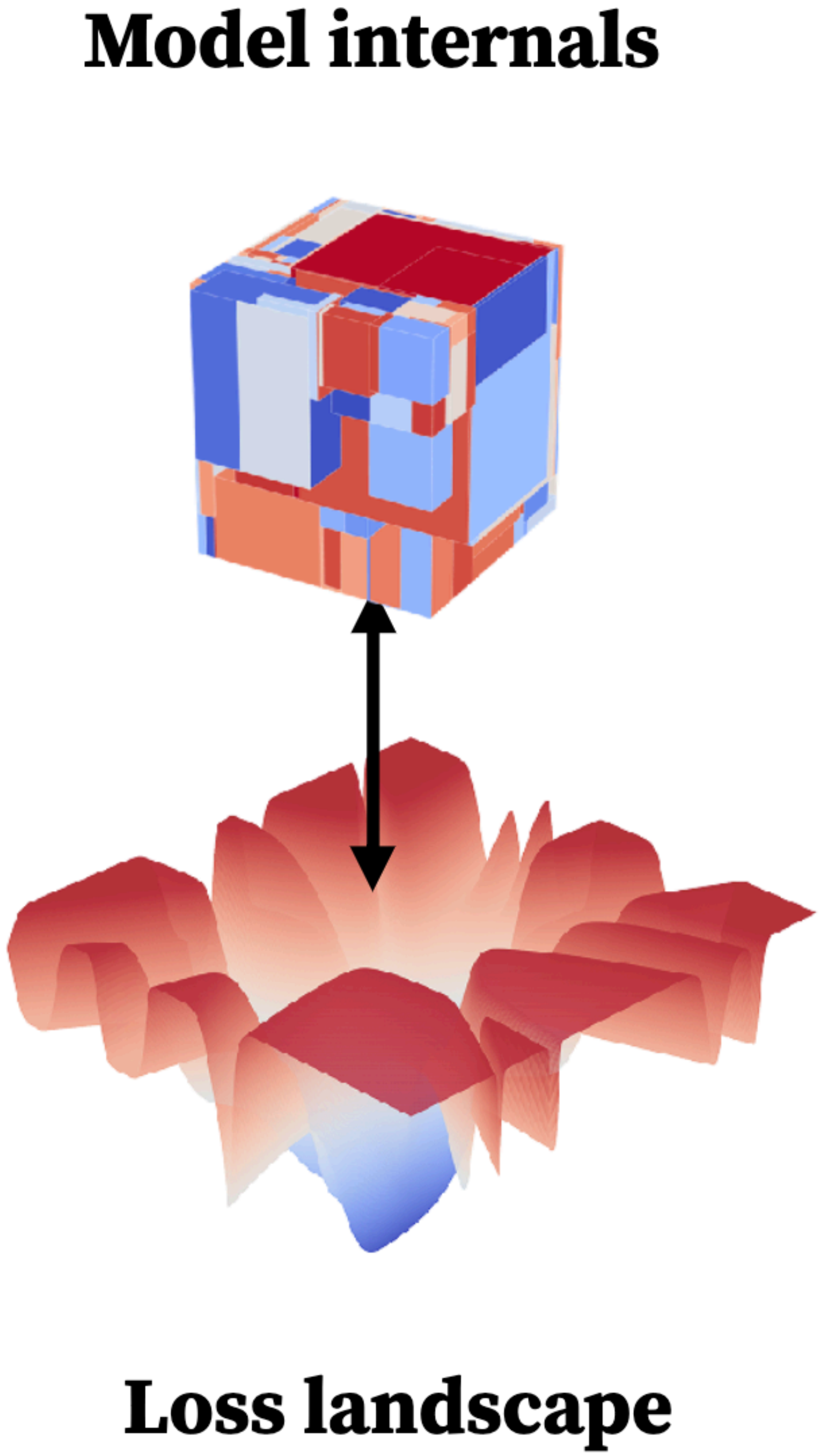
28/08/2025

THE UNIVERSITY OF MELBOURNE

Timaeus

Probing Geometry as Interpretability

- We propose that a conceptual picture where expectation values of observables $\phi(w)$ probe the geometry of the loss landscape $L(w)$
- According to Watanabe *"Knowledge to be discovered from examples corresponds to a singularity"*
- Structural Bayesianism posits that the *structure* of knowledge corresponds to the *structure* of the singularity
- Geometry of the loss landscape $L(w)$ corresponds to the internal structure of a model
- This week we have focused on a specific observable (susceptibilities $\chi = \frac{1}{n}\frac{\partial}{\partial h}\langle\phi\rangle_h\Big|_{h=0}$) detecting model internals such as induction circuits
- Another example is $\mathbb{E}_n[L_n(w)]$ for estimating λ
- Here we explicitly link these idea to geometry



Outline

1. Singular Learning Theory
2. Expectations
3. The Exceptional Divisor

1. Singular Learning Theory

Bayesian Picture

- Let $w \in W \subset \mathbb{R}^d$ be compact. We want to consider **neural networks** $\phi_w(x) : \mathbb{R}^M \to \mathbb{R}^N$
- Given a dataset $D_n = \{(x_i, y_i)\}_{i=1}^n$ assume there is a **true** distribution $q(x)$
- We approximate this distribution with a **model** $p(x \mid w)$
- $p(x \mid w) = \dfrac{1}{(\sqrt{2\pi})^N} \exp\left(-\dfrac{1}{2}\|y - \phi_w(x)\|^2\right)$ will be our model for the DLN setting
- We assume a **prior** probability distribution $\varphi(w)$ on the parameters $w \in W$
- We measure our model against the truth via the **KL divergence** $K(w) := \mathrm{KL}(q\|p) = \displaystyle\int q(x) \log \frac{p(x \mid w)}{q(x)} dx$

Asymptotic Expansion of Z_n

- The loss $L(w) := -\int q(x) \log p(x \mid w) dx$ differs from $K(w)$ by a constant entropy $S := -\int q(x) \log q(x) dx$ so $K(w)$ and $L(w)$ share the same geometry
- We will assume the model is realisable, meaning $W_0 = \{w \in W \mid K(w) = 0\} \neq \emptyset$
- Neural networks $\phi_w(x)$ can be viewed as a statistical mechanical system whose **Boltzmann weight** is the quenched ($K(w)$ not $K_n(w)$) posterior density $e^{-nK(w)}\varphi(w)$
- SLT study singular integrals such as the **partition function** $Z_n := \int_W e^{-nK(w)}\varphi(w)dw$
- SLT determines the large n behaviour of the partition function $Z_n = Cn^{-\lambda}(\log n)^{m-1} + O(n^{-\lambda-1})$ for $C \in \mathbb{R}$, $\lambda \in \mathbb{Q}$ and $m \in \mathbb{Z}^+$

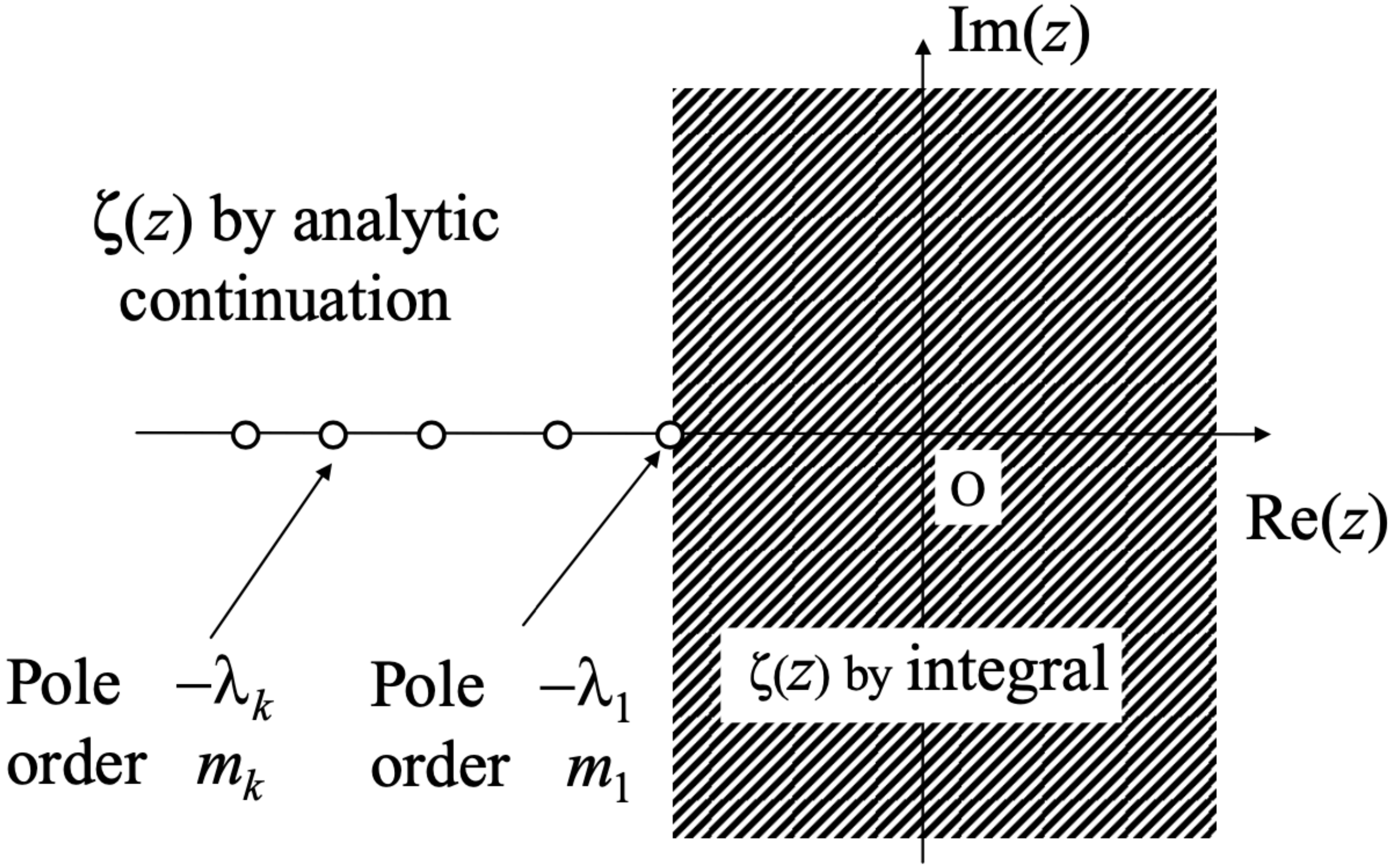
Machinery for the Calculation

- The **density of states** $v(t) := \int_W \delta(t - K(w))\varphi(w)dw$ tells us how many parameter configurations give $K(w) = t$ as $t \to 0^+$
- The **Mellin Transform** of $v(t)$ gives the **zeta function** $\mathscr{M}\{v(t)\}(z) = \zeta(z) := \int_W K(w)^z \varphi(w)dw$
- The zeta function has the Laurent series expansion $\zeta(z) = \sum_{\alpha>0} \sum_{i=1}^{d} \frac{d_{\alpha,i}}{(z-\lambda_i)^{m_i}} + P(z)$
- The largest pole of $\zeta(z)$ is the learning coefficient $\lambda = \lambda_1$ with multiplicity $m = m_1$
- The **Laplace Transform** of $v(t)$ gives the partition function $\mathscr{L}_n\{v(t)\} = Z_n$

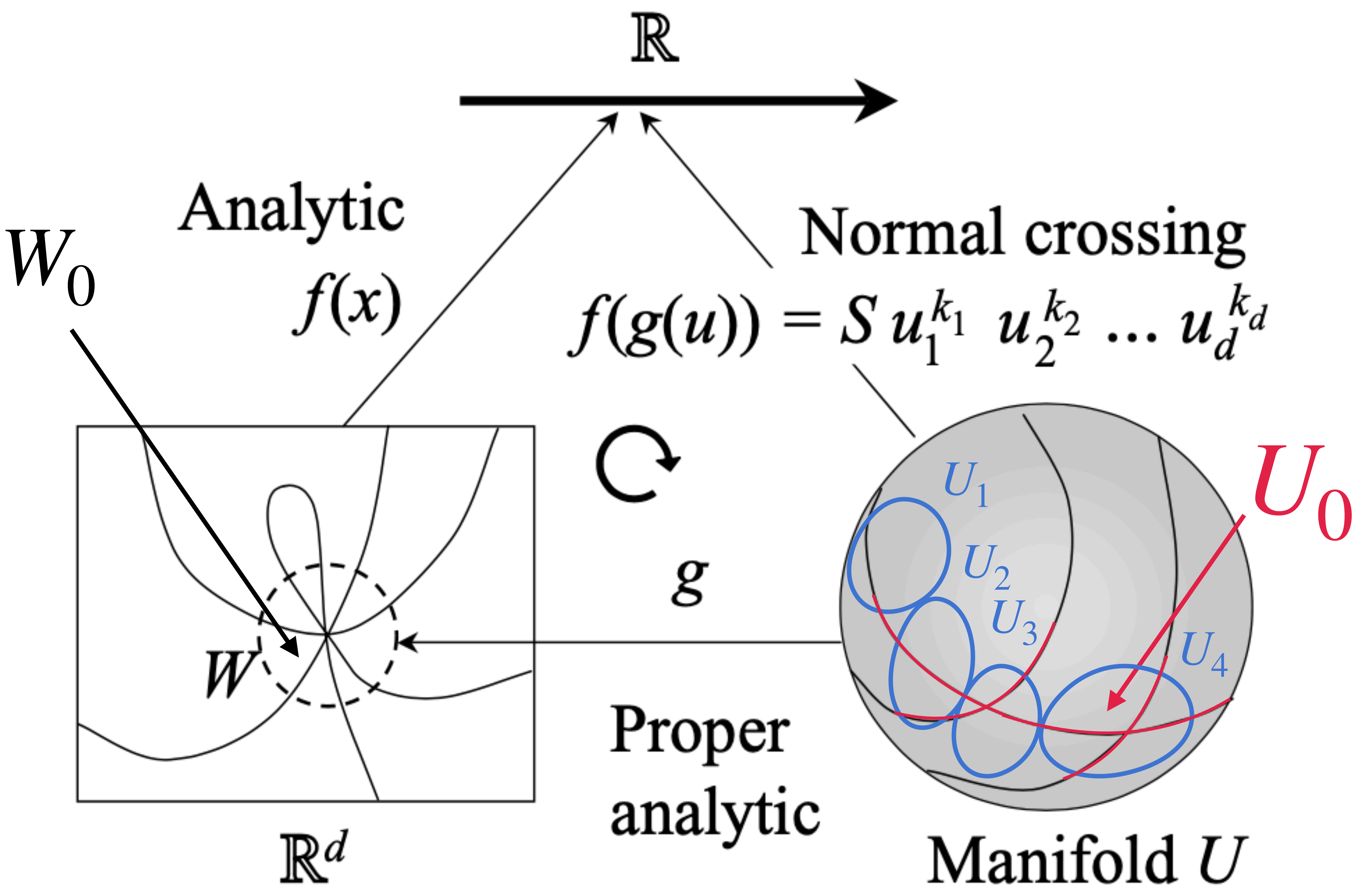
$$\zeta(z) \xrightarrow{\mathscr{M}^{-1}} v(t) \xrightarrow{\mathscr{L}} Z_n$$

$$K(a,b) = a^4 b^2 \qquad \lambda = \frac{1}{4} \qquad \int v(t)dt \sim t^{\lambda}$$

$t = 0.01$ $\quad$ $t = 0.1$ $\quad$ $t = 1$ $\quad$ $t = 10$



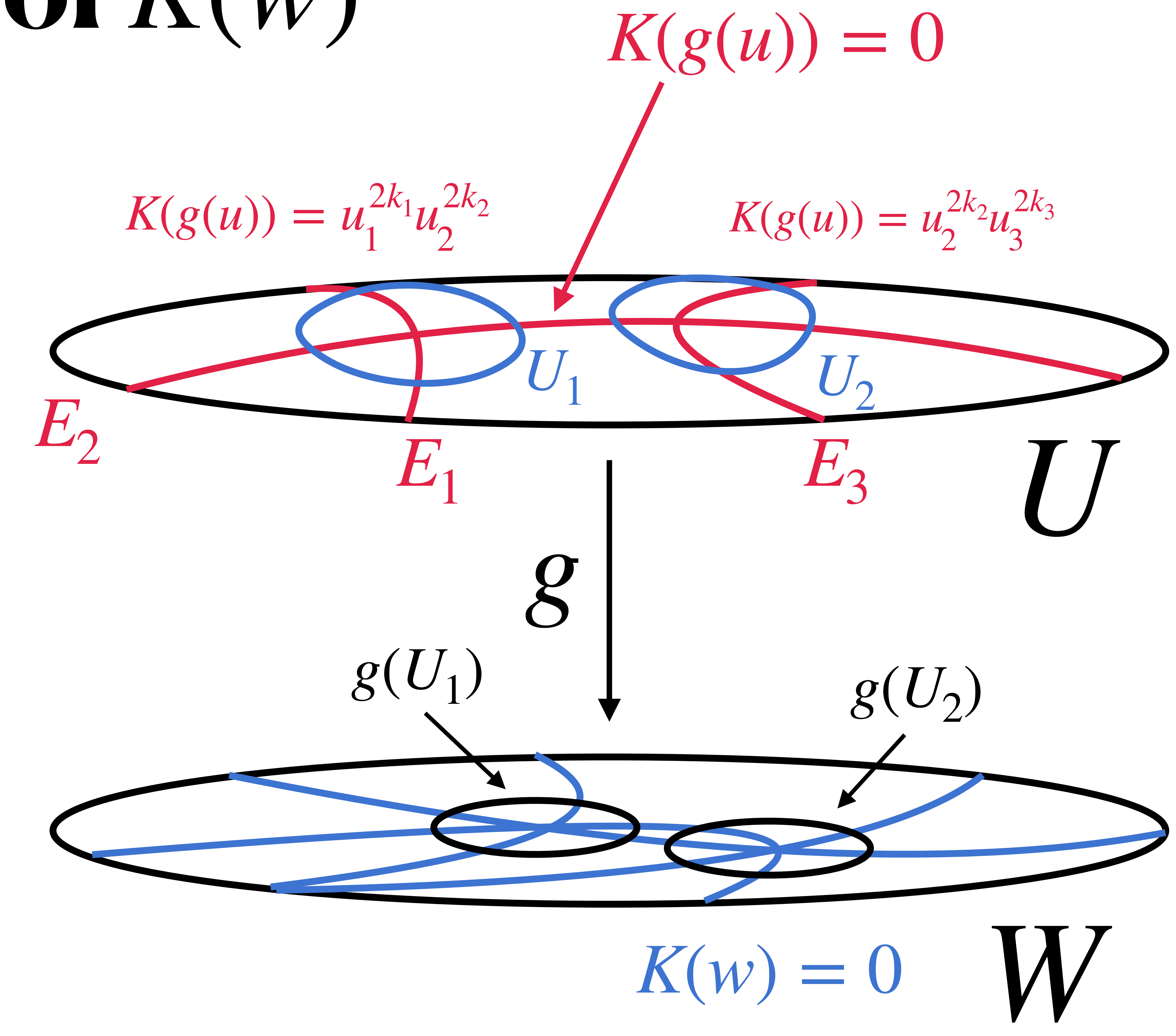
Resolution of Singularities



- To extract the largest pole of $\zeta(z)$ we need to perform a resolution of singularities
- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a real analytic map such that $f(0) = 0$ (f is not a constant function)
- Then there exists a triple (W, U, g) where:
- W is an open set in $\mathbb{R}^d$, U is a real analytic d-dimensional manifold and g is a real analytic (proper) map $g : U \to W$
- Let $U_0 = \{u \in U \mid f(g(u)) = 0\}$, for any point $P \in U_0$ there is a local coordinates $u = (u_1, \ldots, u_d)$ where P is the origin
- Since $W \subseteq \mathbb{R}^d$ is compact, we can cover U with a finite number of charts $U = \cup_{\alpha \in \Lambda} U_\alpha$
- Locally on each chart, $f(g(u)) = S u_1^{k_1} \cdots u_d^{k_d}$ with Jacobian $|g'(u)| = b(u)\,|u_1^{h_1} \cdots u_d^{h_d}|$ for $S = \pm 1$ and $b(u) \neq 0$ is real analytic

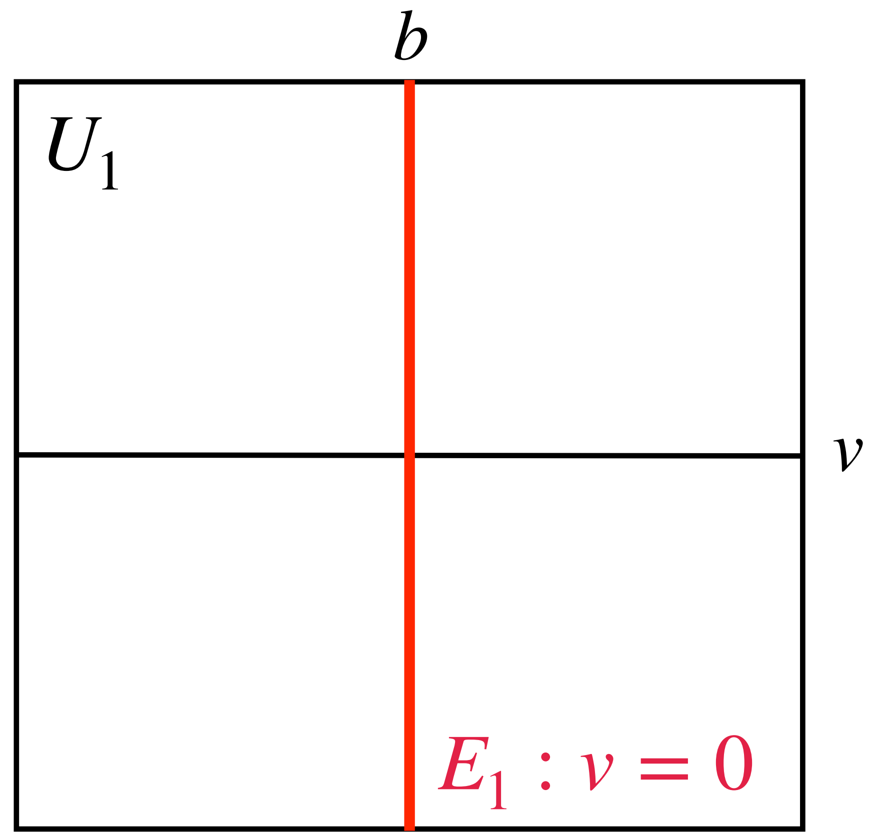
Resolution of $K(w)$

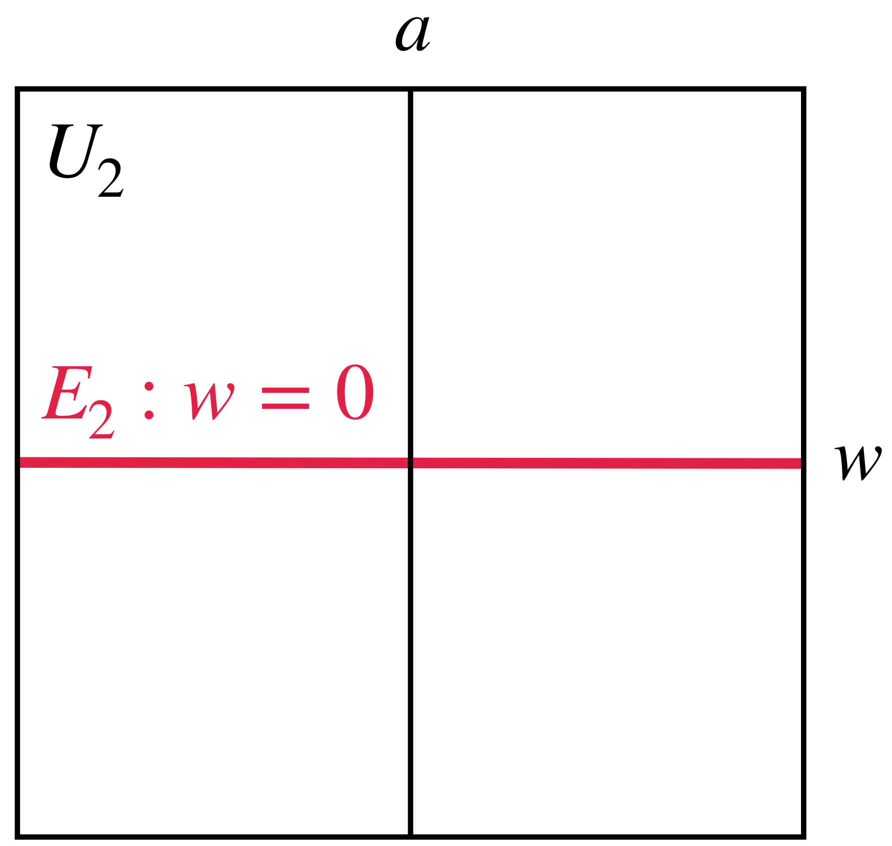
- We now apply resolution to $K(w) : \mathbb{R}^d \to \mathbb{R}$, giving us a triple (W, U, g)
- Cover the manifold U with finitely many local charts $\cup_{\alpha \in \Lambda} U_\alpha$
- Locally on each chart U_α, our KL-divergence is $K(g(u)) = u_1^{2k_1} \cdots u_d^{2k_d}$
- $\zeta(z) = \int_W K(w)^z \varphi(w) dw$ becomes $\zeta(z) = \sum_\alpha \int_{U_\alpha} u_1^{2k_1 z} \cdots u_d^{2k_d z} b(u) | u_1^{h_1} \cdots u_d^{h_d} | du_1 \cdots du_d$
- Then integrating, the largest pole is $\lambda = \min_\alpha \min_{1 \le i \le d} \left(\frac{h_i + 1}{2k_i} \right)$ with multiplicity $m \in \mathbb{Z}^+$

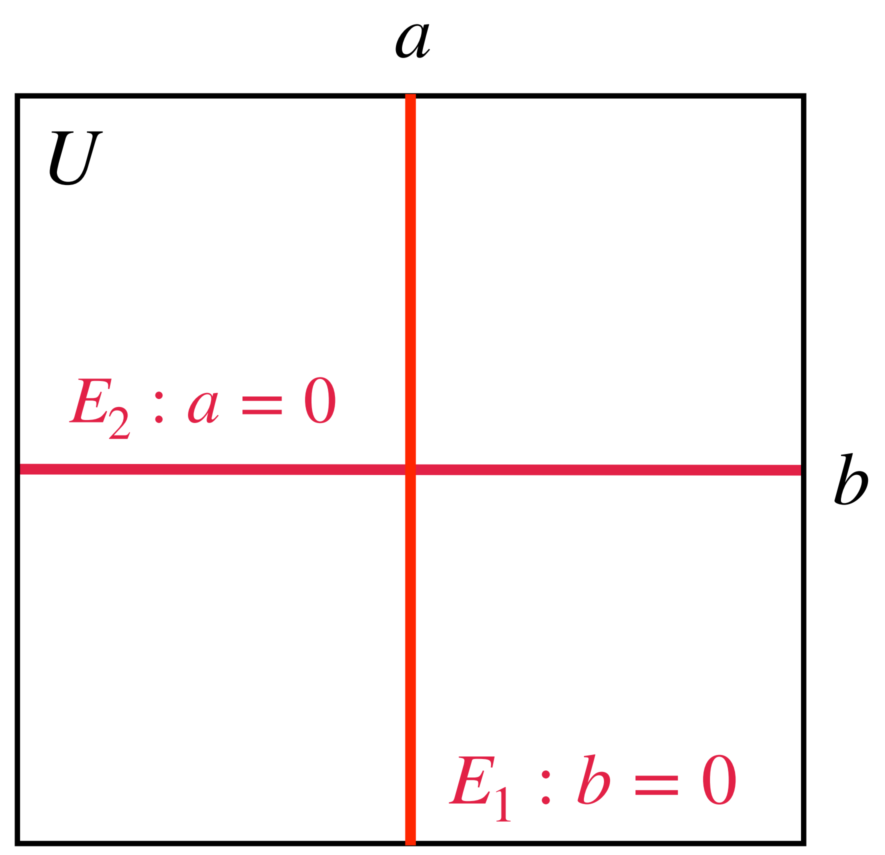


Interlude: Blowup Example

- Consider a simple example $K_1(a,b) = a^2 + b^2$. We want the largest pole of $\zeta(z) = \int\int K_1(a,b)^z \varphi(a,b) da db$
- we take the blowup on two charts U_1, U_2
- On U_1 we take $\begin{cases} a = v \\ b = bv \end{cases}$ giving $K_1(v,b) = v^2(1+b^2)$ with Jacobian $|v|$
- On U_1 we have the zeta function $\zeta_1(z) = \int\int v^{2z}(1+b^2)^z |v| \varphi(v,b) dv db$, integrating over v gives the largest pole $\lambda = 1$
- On U_2 take $\begin{cases} a = wa \\ b = w \end{cases}$ to obtain the same pole $\lambda = 1$
- Another example $K_2(a,b) = a^2 b^2$ is already normal crossing on a manifold U







Summary

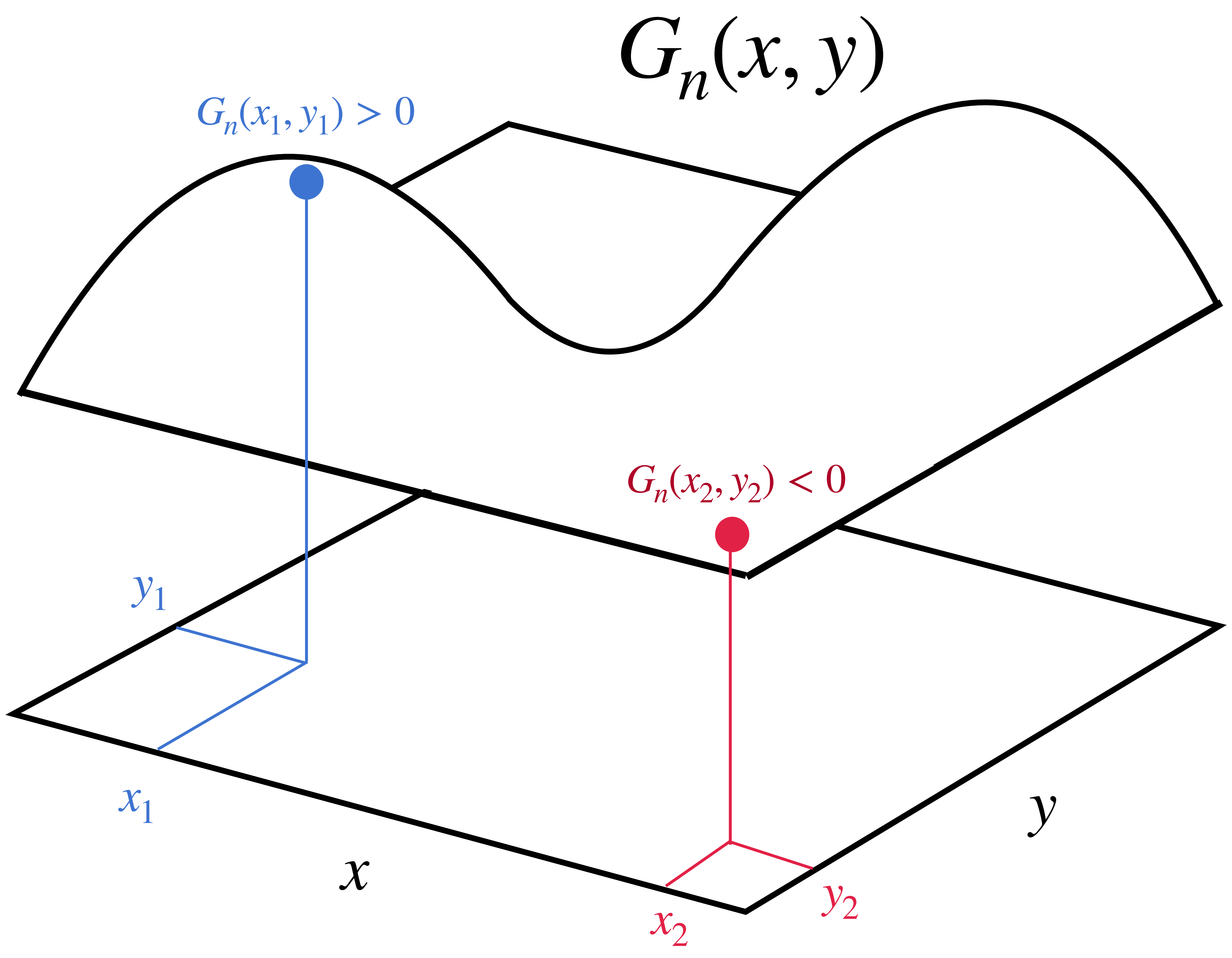
To Calculate $Z_n = Cn^{-\lambda}(\log n)^{m-1} + O(n^{-\lambda-1})$

1. Start with the zeta function $\zeta(z) = \int_W K(w)^z \varphi(w) dw$
2. Perform a resolution of singularities at $K(w) = 0$ giving the standard form $\zeta(z) = \sum_\alpha \int_{U_\alpha} u_1^{2k_1 z} \cdots u_d^{2k_d z} b(u) \,|\, u_1^{h_1} \cdots u_d^{h_d} \,|\, du_1 \cdots du_d$
3. Integrate and identify the largest pole $\lambda = \min_\alpha \min_{1 \le i \le d} \left(\frac{h_i + 1}{2k_i} \right) \in \mathbb{Q}$ with multiplicity $m \in \mathbb{Z}^+$
4. Inverse Mellin Transform to arrive at density of states $v(t) \sim \frac{a^{-\lambda}}{(m-1)!} t^{-\lambda-1} \left(\log \frac{a}{t} \right)^{m-1}$ for some $a \in \mathbb{R}$
5. Laplace Transform to obtain leading order of asymptotic expansion of $Z_n \sim Cn^{-\lambda}(\log n)^{m-1}$ for some $C \in \mathbb{R}$

2. Expectations

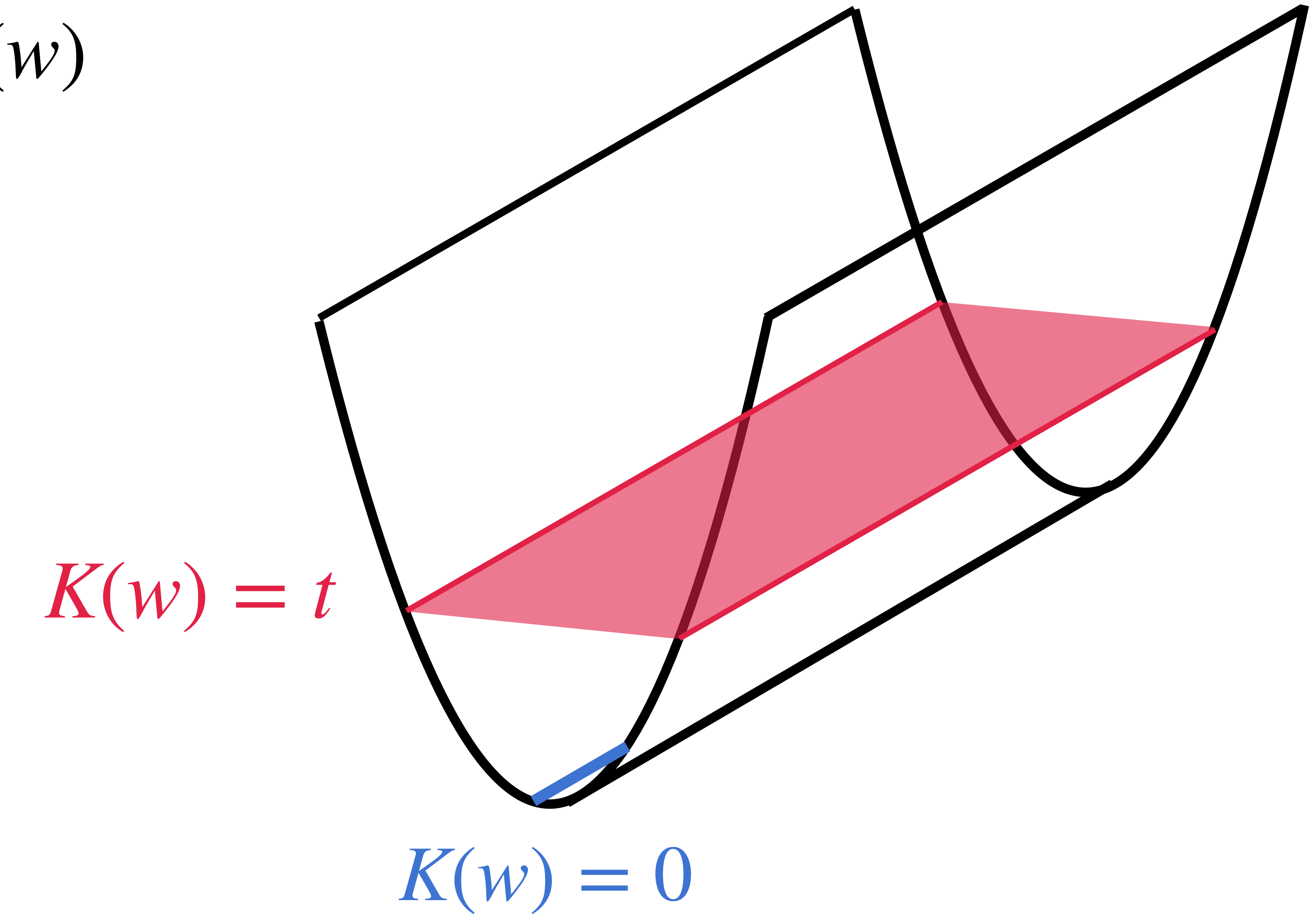
Correlation functions

- Many quantities of interest can be written as expectation values $\mathbb{E}_n[\phi] = \int_W \phi(w)e^{-nK(w)}\varphi(w)dw$
- Consider a linear network with two-layers with parameter $w = (A, B)$. The function computed is $\phi_w(x) : \mathbb{R}^M \to \mathbb{R}^N, \quad \phi_w(x) = BAx$
- We are going to focus on a type of **correlation function**
- Defining **centred networks** via $\bar{\phi}_w(x) = (BA - B_0A_0)x$, we consider the observable $f(w; x, y) = \langle \bar{\phi}_w(x), \bar{\phi}_w(y) \rangle$
- As a function of inputs vectors $x, y \in \mathbb{R}^M$, we define a **correlation function** $G_n(x, y) := \mathbb{E}_n[f(w; x, y)]$



Extending the Machinery

- We need SLT to handle expectation values $\mathbb{E}_n[f]$ where $f \neq K(w)$
- We define the **correlation of states** as $v(t, x, y) = \int_W \langle \phi_w(x), \phi_w(y) \rangle \delta(t - K(w)) \varphi(w) dw$
- We define the **correlation zeta function** as $J(z, x, y) = \int_W \langle \phi_w(x), \phi_w(y) \rangle K(w)^z \varphi(w) dw$
- As before, the **Mellin Transform** of the correlation of states $v(t, x, y)$ is the correlation zeta function $J(z, x, y)$, and the **Laplace transform** of the correlation of states (divided by Z_n) is the correlation function $G_n(x, y)$
- Two analytic functions $f(w), g(w)$ are **contact equivalent** if there exists $\alpha_1, \alpha_2 > 0$ such that $\alpha_1\, g(w) \leq f(w) \leq \alpha_2\, g(w)$



$$\begin{tikzcd} J(z, x, y) & \\ v(t, x, y) \arrow[u, "\mathscr{M}"] \arrow[r, "\mathscr{L}"] & G_n(x, y) \end{tikzcd}$$

Calculation Sketch

1. Start with the KL divergence $K(w)$ with network $\phi_w : \mathbb{R}^M \to \mathbb{R}^N$, $\phi_w(x) = BAx$ learning $\phi_0(x) = B_0 A_0 x$
2. $K(w)$ is **contact equivalent** to $\|BA - B_0A_0\|^2$ since we assume $(\Sigma^x)_{ij} = \int x_i x_j q(x) dx$ has positive eigenvalues.
3. The zeta function $\zeta(z) = \int_W \langle \overline{\phi}_w(x), \overline{\phi}_w(x) \rangle \|BA - B_0A_0\|^{2z} \varphi(w) dw$ where $w = (A, B)$ has the same pole structure
4. Change variables a few times in $K(w)$ and $\langle \overline{\phi}_w(x), \overline{\phi}_w(y) \rangle$, then simultaneously resolve the singularities of both polynomials for fixed $x, y \in \mathbb{R}^M$ via an induction on dimension of the hidden layer
5. Integrate to extract the largest pole from $J(z, x, y)$, then inverse Mellin transform to the correlation of states $v(t, x, y)$ then Laplace transform to get $G_n(x, y)$

Main Result 1

- The correlation function $G_n(x, y)$'s asymptotic expansion has the leading terms $G_n(x, y) \sim \sum_{i \in \mathcal{I}} n^{-i} \left\langle x, \mathcal{O}_i(w_0) y \right\rangle$ where $\mathcal{I} = \{1, \frac{3}{2}, 2, \frac{5}{2}, 3\}$ and for each i we have a symmetric operator $\mathcal{O}_i(w_0) : \mathbb{R}^M \to \mathbb{R}^M$ which depends on the true parameter
- The matrix entries of the symmetric operators $\mathcal{O}_i(w_0)$ are surface integrals capturing the geometry of the loss landscape around the set W_0
- We can use the sign of the leading term of $G_n(x, y) \sim n^{-1} \left\langle x, \mathcal{O}_1(w_0) y \right\rangle$ to determine if two datapoints are positively or negatively correlated
- As we vary $x, y \in \mathbb{R}^M$, we can cancel off the leading term n^{-1}. While keeping $x \perp y$ with respect to $\mathcal{O}_1(w_0)$, we can vary within this subspace to cancel off $n^{-3/2}$, and so on.
- This gives the conceptual picture where the correlation function probes the strata of the exceptional divisor

Toy Model of Cancellation

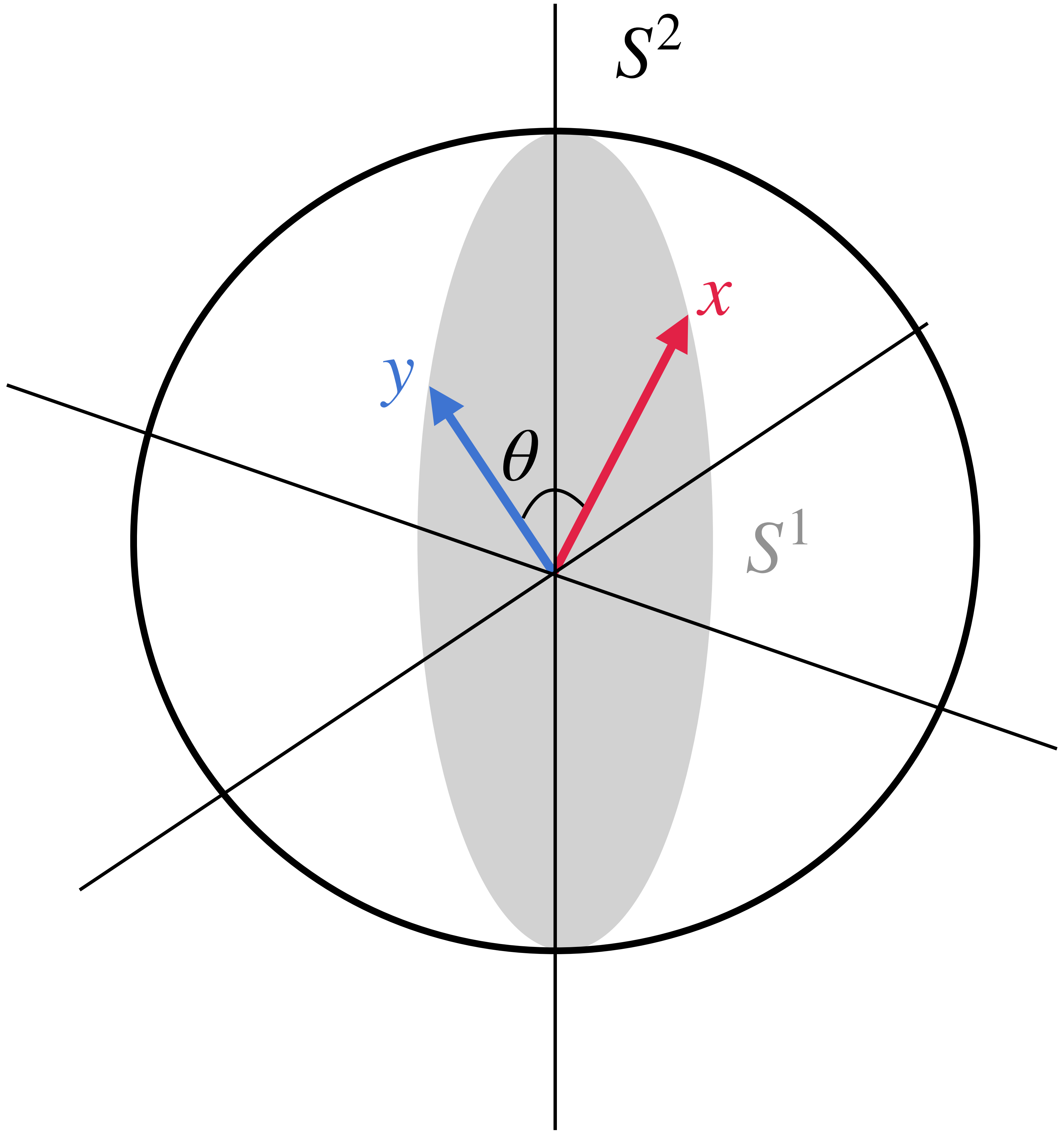
- We want to illustrate the leading order exponent in $G_n(x, y)$ changes discontinuously as function of x, y
- We consider a toy model $\phi_w(x) = BAx$ where $B \in \mathbb{R}$ and $A \in M_{(1,3)}(\mathbb{R})$ and $x \in \mathbb{R}^3$ with $B_0A_0 = (1 \quad 0 \quad 0)$
- We consider the zeta function $J(z, x, y) = \int_W \langle \bar{\phi}_w(x), \bar{\phi}_w(y) \rangle \|BA - B_0A_0\|^{2z} \varphi(w)dw$ and find largest pole at $-5/2$
- We show $J(z, x, y) = \langle x, \begin{pmatrix} c_1^{(2)}(z, \varepsilon) + 4c_5^{(2)}(z, \varepsilon) & 2c_4^{(2)}(z, \varepsilon) & 2c_4^{(2)}(z, \varepsilon) \\ 2c_4^{(2)}(z, \varepsilon) & J_{22}(z, \varepsilon) & 0 \\ 2c_4^{(2)}(z, \varepsilon) & 0 & J_{33}(z, \varepsilon) \end{pmatrix} y \rangle$ with $\mathcal{O}_{-5/2}(w_0) := \lim_{z \to -5/2} (z + 5/2) J(z, x, y)$
- Set $\varepsilon = 0.1$ as the integration region and each $c_i^{(p)}(z, \varepsilon)$ is a meromorphic function,

$$c_i^{(2)}(z, \varepsilon) = \int_{1-\varepsilon}^{1+\varepsilon} da_{11} \left(\frac{(\varepsilon/a_{11})^{2z+5} g_i(z, \varepsilon/a_{11})}{2\left(z+5/2\right)} - \frac{(\varepsilon/a_{11})^{2z+6} g_i'(z, \varepsilon/a_{11})}{4\left(z+5/2\right)(z+3)} + \frac{(\varepsilon/a_{11})^{2z+7} g_i''(z, \varepsilon/a_{11})}{8\left(z+5/2\right)(z+3)(z+7/2)} - \frac{\int_0^{\varepsilon/a_{11}} v^{2z+7} g_i'''(z, v) dv}{8\left(z+5/2\right)(z+3)(z+7/2)} \right)$$

- Each $g_i(z, v)$ is a holomorphic function

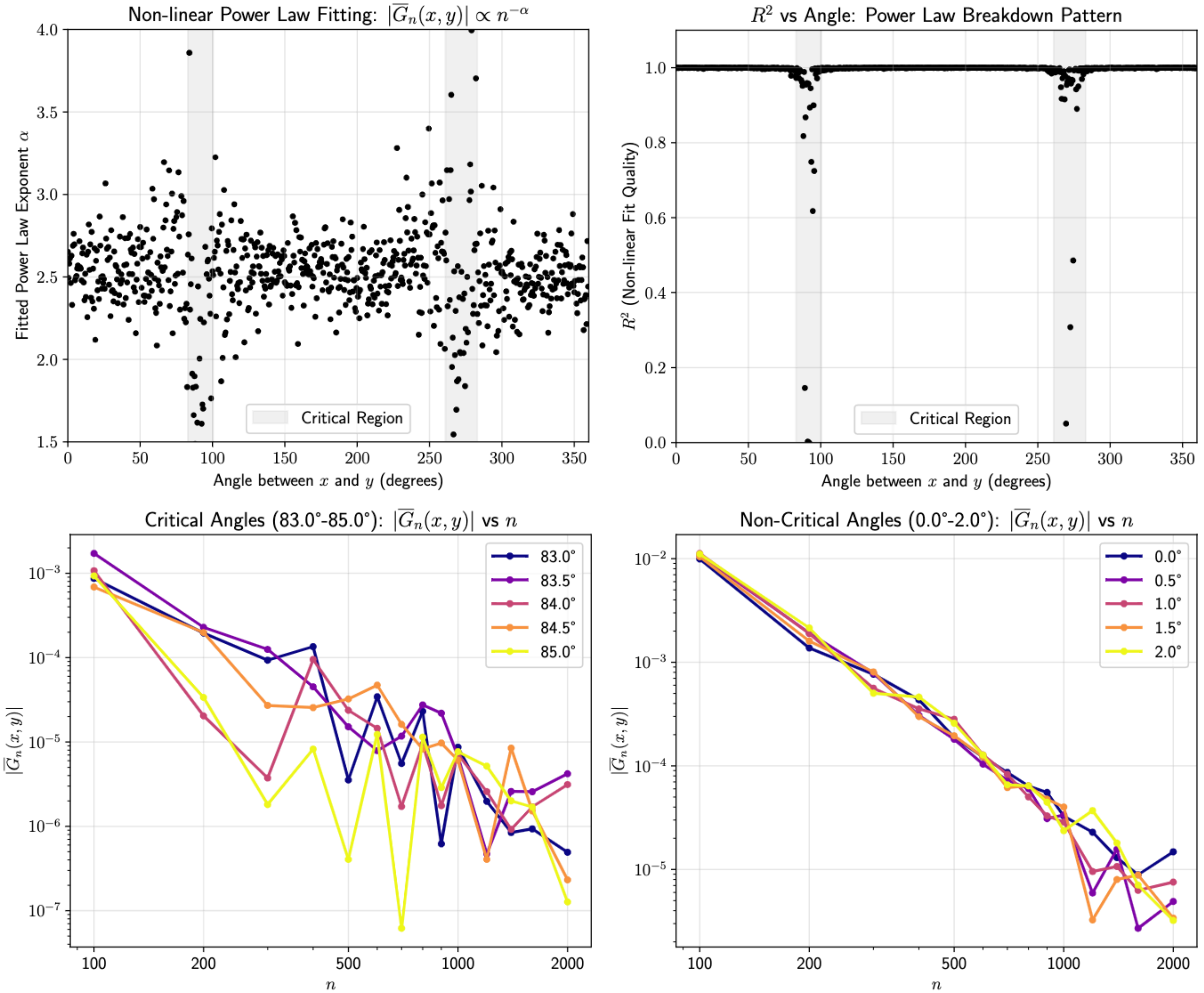
$$g_1(z, v) = \int_0^{\varepsilon v/a_{11}} df_{11} \int_0^{\varepsilon v/a_{11}} df_{12} (1 + f_{11}^2 + f_{12}^2)^z$$

Experimental Setting



- Theoretically we find that $\overline{G}_n(x,y) := \int_W \langle \overline{\phi}_w(x), \overline{\phi}_w(x) \rangle e^{-nK(w)} \varphi(w) dw$ behaves as $\overline{G}_n(x,y) \sim n^{-5/2}$
- $-5/2$ the largest pole of $J(z,x,y)$
- We will fix a data point $x = (10,10,10)$ and sweep y varying the angle between x and y in a circle S^1 in $\mathbb{R}^3$
- We expect when the leading term cancels, the power law behaviour of $\overline{G}_n(x,y)$ will break and lower order terms dominate, making extracting exponent difficult experimentally
- Theoretically we find when $\mathcal{O}_1(w_0) = 0$ that $\mathcal{O}_2(w_0) = 0.241059 \neq 0$

- **Top Left:** Leading order terms exponent α against angle between $|\overline{G}_n(x,y)| \sim n^{-5/2}$
- **Top Right**: Highlights power law breakdown where $x \perp y$ in leading order term
- **Bottom Left**: Power in low R^2 (grey breakdown) region
- **Bottom Right**: Power law in high R^2 region



3. Exceptional Divisor

Opening

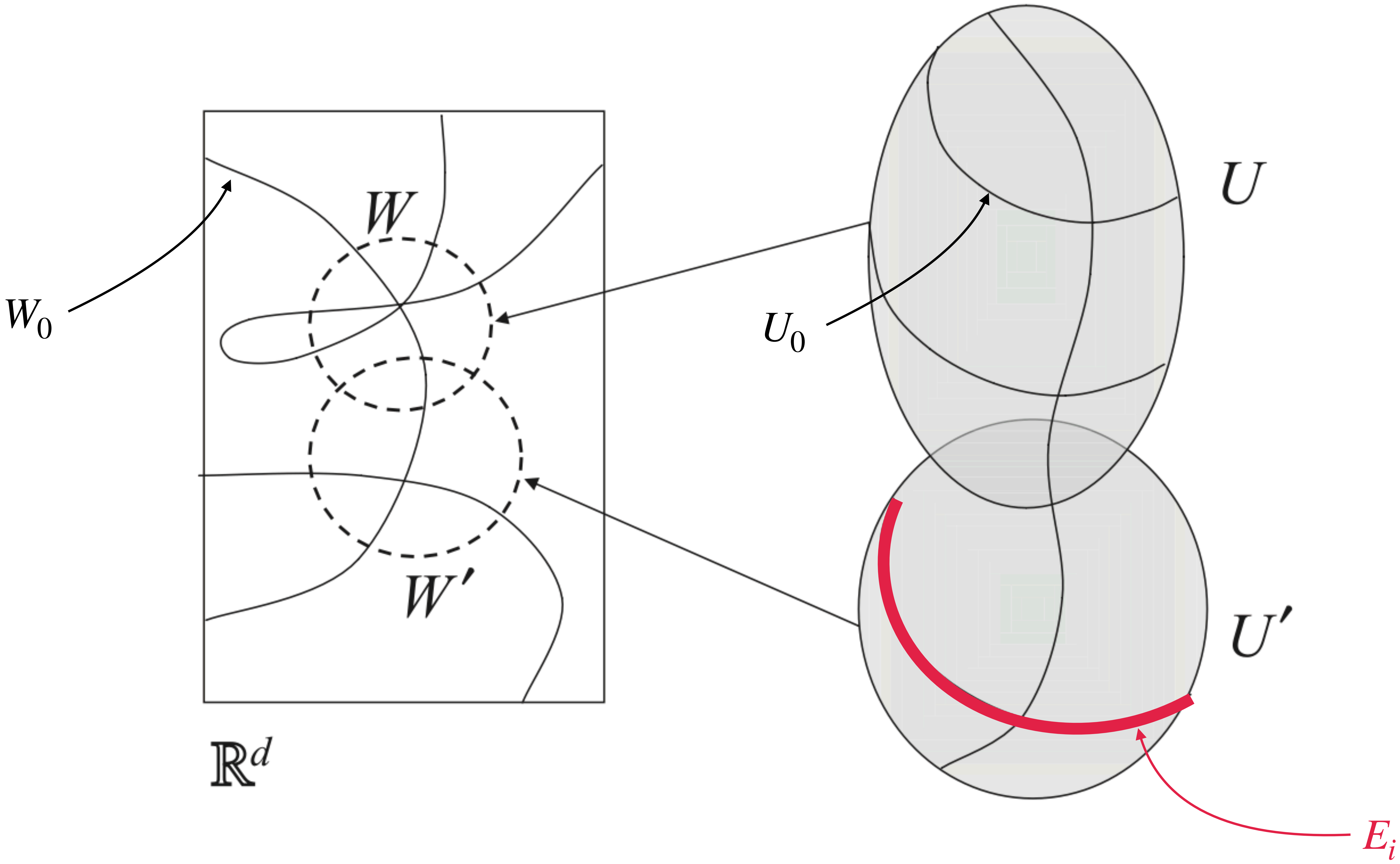
- Structural Bayesianism says internal structure in learning machines should correspond to internal structure in singularities. **But what is that?**
- Empirically, we study internal structure in models using “response matrices” whose entries are susceptibilities. These are covariances computed against the Bayesian posterior. **What does this have to do with geometry?**
- Expectation values against the Bayesian posterior “probe” the exceptional divisor of the resolution of singularities
- This object has some discrete structure (e.g. a natural stratification, the dual complex) and some continuous structure (the strata are manifolds)
- This structure “controls” the patterns in susceptibilities and so it is natural to relate it to internal computational structure in models and changes in such structure

Primes

- In algebraic geometry there is a generalisation of the prime decomposition $n = p_1^{a_1} \cdots p_k^{a_k}$ of an integer to a *primary decomposition* of an ideal $\mathfrak{a} = \cap_{i=1}^k \mathfrak{q}_\mathfrak{i}$ (see Atiyah & Macdonald)
- In a noetherian topological space X, every nonempty closed subset Y can be expressed as a finite union $Y = Y_1 \cup \cdots \cup Y_k$ of irreducible closed subsets. This expression is unique if we require $Y_i \not\supseteq Y_j$ for all i, j. The Y_i are called the *irreducible components* of Y and $Y = \cup_{i=1}^k Y_i$ as the *irreducible decomposition*
- You can think of the Y_i as the "prime factors" of Y
- What are the prime factors of the loss landscape?

Throat clearing

- $K : W \to \mathbb{R}$ is going to be the KL divergence of the model $p(x \mid w)$ to the truth $q(x)$
- Under the usual fundamental conditions, K is analytic and can be complexified $K_{\mathbb{C}}$ to a holomorphic function, resolution of singularities of $K, K_{\mathbb{C}}$ can be done "in parallel" (i.e. centers of blowups are defined by real equations)
- Avoid working directly with irreducible decompositions of real analytic functions
- It is convenient to work both with differential geometry of real points and surface integrals, and also with meromorphic differential forms / Leray residues over $\mathbb{C}$, I'll focus on the former in this talk but the latter leads to more powerful techniques

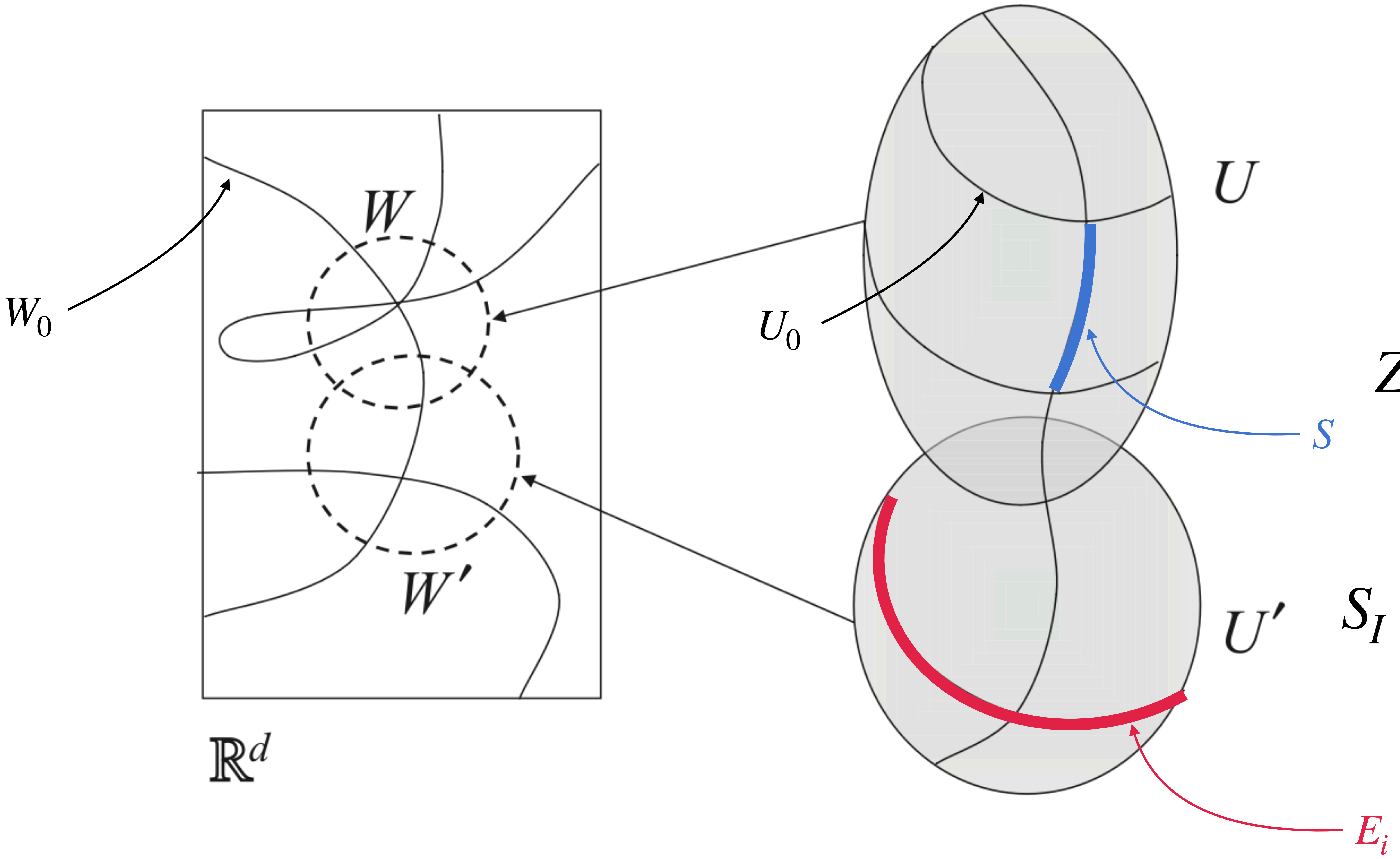


Exceptional divisor $E = U_0 = g^{-1}(W_0)$

Irreducible decomposition $E = \bigcup_i E_i$

Expectation values

- Let $f : W \to \mathbb{R}$ be a suitable function (e.g. smooth)
- $\mathbb{E}_n[f] = \dfrac{1}{Z_n}\displaystyle\int_W f(w)\exp(-nK(w))\varphi(w)dw = \dfrac{1}{Z_n}\mathbb{N}_n[f]$
- Both Z_n and $\mathbb{N}_n[f]$ have an asymptotic expansion, if we have $Z_n \sim Cn^{-\lambda}(\log n)^{m-1}$ and $\mathbb{N}_n[f] \sim C(f)n^{-\mu}(\log n)^{l-1}$ then $\mathbb{E}_n[f] \sim \frac{C(f)}{C}n^{-\lambda+\mu}(\log n)^{l-m}$
- To compute the asymptotic expansion of $\mathbb{N}_n[f]$ we monomialise $K(w)$, $|\det g'|\varphi$ as usual but *not* $f(w)$
- Choose partition of unity subordinated to the natural stratification of the exceptional divisor
- Use tubular neighborhoods of the irreducible components (and for these to be "neat" we rely on the resolution arranging transverse intersections with boundary) to pullback integral to (the total space of) the normal bundle, use symmetric algebra of the conormal bundle to "avoid" coordinate choices (i.e. w_α, w_β)
- The integral $\mathbb{N}_n[f]$ can be "localised" to the strata S using Taylor series expansions of f in the normal directions to this submanifold, reducing the integral to "surface" integrals along S with the pushforward from the normal bundle contributing "fiber moments" $\displaystyle\int_{\mathbb{R}} u^\gamma e^{-nu^{2k}}du$



$$E_I = \bigcap_{i\in I} E_i$$

$$Z_k = \bigcup_{|I|\geq r-k} E_I$$

$$S_I = E_I \setminus \bigcup_{j\notin I} E_j \supseteq S$$

Exceptional divisor $E = U_0 = g^{-1}(W_0)$

Irreducible decomposition $E = \bigcup_i E_i = Z_d \supseteq Z_{d-1} \supseteq \cdots \supseteq Z_0$

Strata as Factors

- The end result is $\mathbb{N}_n(f) = \sum_S \sum_\gamma \sum_{\alpha+\beta=\gamma} \left\{ \int_S (D_\perp^\alpha(f) J_\beta) dS \right\} n^{-\mu(\gamma)} (\log n)^{m(\gamma)-1}$ where J_β is part of the Taylor series expansion of the map $NS \supseteq V \to U \to W$ and $D_\perp^\alpha(f)$ denotes the "normal" Taylor series expansion of f
- This is a pretty standard idea (e.g. see Denef-Loeser "Motivic Igusa zeta functions")
- Think of the coefficient as $\langle$strata, observable$\rangle$
- Since f can vanish to some order along S the leading term here may have $\alpha \neq 0$
- **As the function f varies in a family** the variation in $\mathbb{E}_n[f]$ is controlled by the variation in the integrals $\int_S D_\perp^\alpha(f) J_\beta dS$, that is, by how the variation affects the strata integral
- In this sense expectation values against the Bayesian posterior "probe" the exceptional divisor of the resolution of singularities
- **Ben's part of the talk:** by varying f continuously if we change the decay exponent, we have detected some cancellation of these strata integrals, possibly an interesting signal about the geometry (derived from observation)

Strata as Factors

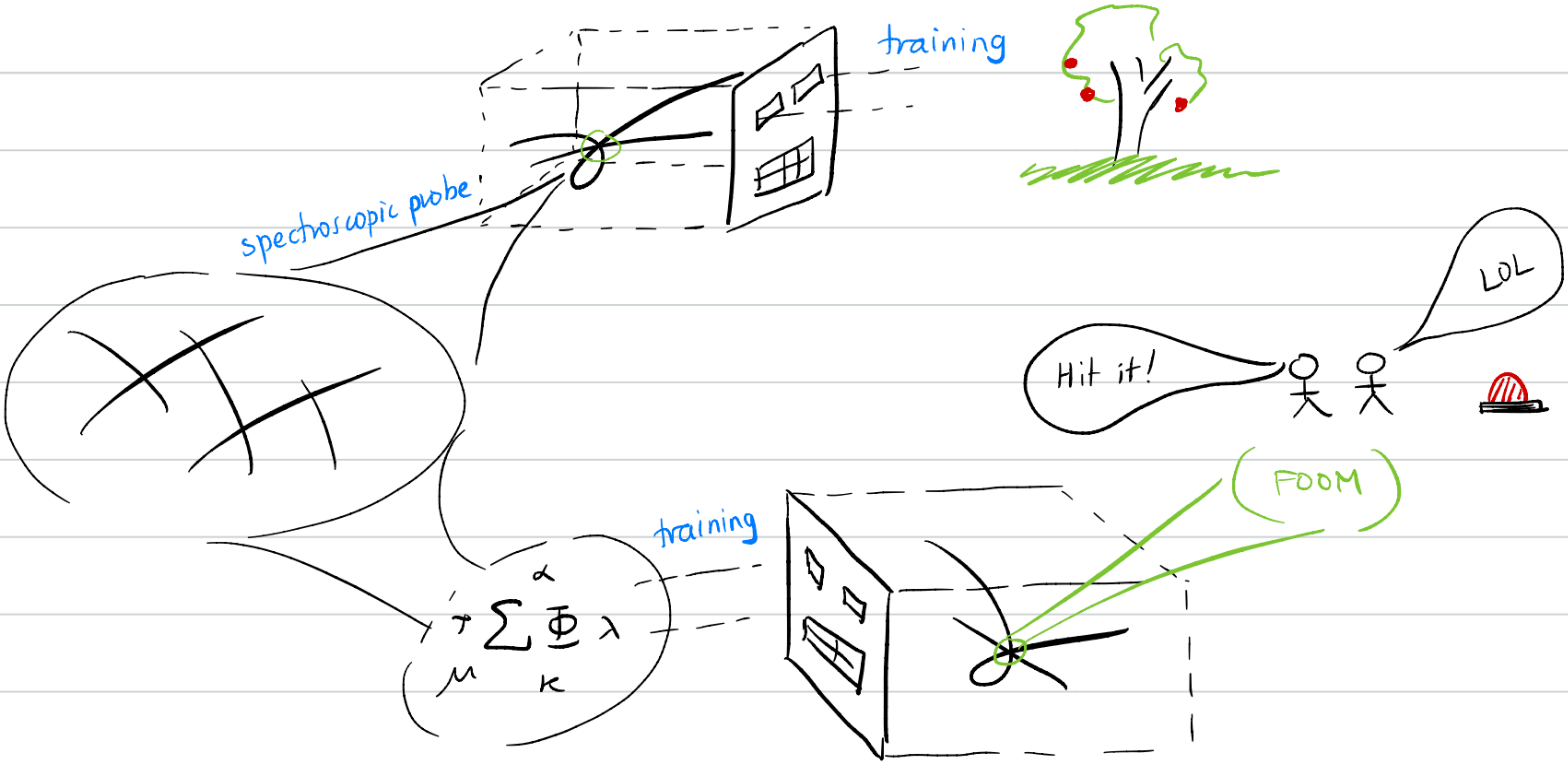
- Correlation functions / covariances (and thus susceptibilities) are defined in terms of expectation values $\mathbb{E}_n[\phi_1\phi_2]$, usually at fixed n
- When we vary ϕ_1, ϕ_2 (e.g. over tokens or components) this affects the expectation values through the variation of the strata integrals $\int_S D_\perp^\alpha(\phi_1\phi_2)J_\beta dS$
- **Proposal**: the relation between internal structure in singular models (e.g. found empirically by patterns in susceptibilities) and internal structure in a singularity involves the discrete "combinatorics" of the configuration of the exceptional divisor and the continuous geometry of the strata, and how these interact with integrals.

McKay Correspondence

- There is an intuition that irreducible components of the exceptional divisor are like "subatomic units" that make up the singularity
- Take a finite subgroup $G \subsetneq SL(2,\mathbb{C})$ that does not contain a pseudo-reflection (e.g. the subgroup C_n, a cyclic group of order $n+1$ generated by $\mathrm{diag}(e^{\frac{2\pi i}{n+1}}, e^{-\frac{2\pi i}{n+1}})$)
- Then Klein showed that $R = k[\![x,y]\!]^G$ is a hypersurface singularity $k[\![x,y,z]\!]/(f)$ of ADE type (e.g. isomorphic to $k[\![x,y,z]\!]/(x^{n+1}+y^2+z^2)$ if $G = C_n$)
- **McKay correspondence:** there is a one-to-one correspondence between (nontrivial) irreducible representations of G, irreducible components of the exceptional divisor of the minimal resolution of the hypersurface singularity, indecomposable MCM R-modules and indecomposable matrix factorisations of f
- For an accessible treatment, see "The McKay correspondence" by Graham Leuschke

Conclusion

- Necessary to address the role of stochasticity and interaction between $\xi_n(u)$, irreducible decomposition and observables
- Understand how variations in the true distribution affect the resolution and individual strata ("discrete" changes are adding/removing components, "continuous" changes to strata), e.g. equisingularity
- Towards categorical SLT: the set of natural observables (i.e. "things that couple to components of the exceptional divisor") can probably be clarified with categories of boundary conditions / matrix factorisations
- Study the computational meaning of the exceptional divisor (e.g. in "Programs as Singularities" setting, graphical models, etc)



Metauni talk and notes