

Interpolating between QFT and machine learning

Michael R. Douglas

CMSA, Harvard University

Costas Bachas Fest
ENS Summer Institute
June 26, 2024

Abstract

We discuss relations between QFT and probability theory, with an eye towards defining spaces of QFTs.

Costas and I wrote several papers together, on Dirichlet branes and on conformal interfaces. We also co-organized the Les Houches schools in 2001 and 2007. It will be a pleasure to revisit the topics we discussed and worked on over the years.

But these days I am working more on machine learning. Fortunately there is a highly developed interface between ML and field theory:

- Statistical physics and machine learning: Parisi, Mézard, Zecchina, Montanari, Monasson, Krzakala, Zdeborova, ...
- ML techniques for lattice FT: Cranmer, Shanahan, Urban, ...
- Exact RG and relations to information theory: Berman and Klingler, Cotler and Rezchikov, ...
- QFT-NN relations: Halverson, Maiti, Demertas, Schwartz, ...

So I will revisit topics discussed in “Calabi’s diastasis as interface entropy” 1311.2202 (also with Ilka Brunner and Leonardo Rastelli) in the light of what I have since learned about ML and statistics.

Costas and I wrote several papers together, on Dirichlet branes and on conformal interfaces. We also co-organized the Les Houches schools in 2001 and 2007. It will be a pleasure to revisit the topics we discussed and worked on over the years.

But these days I am working more on machine learning. Fortunately there is a highly developed interface between ML and field theory:

- Statistical physics and machine learning: Parisi, Mézard, Zecchina, Montanari, Monasson, Krzakala, Zdeborova, ...
- ML techniques for lattice FT: Cranmer, Shanahan, Urban, ...
- Exact RG and relations to information theory: Berman and Klinger, Cotler and Rezchikov, ...
- QFT-NN relations: Halverson, Maiti, Demertas, Schwartz, ...

So I will revisit topics discussed in “Calabi’s diastasis as interface entropy” 1311.2202 (also with Ilka Brunner and Leonardo Rastelli) in the light of what I have since learned about ML and statistics.

Costas and I wrote several papers together, on Dirichlet branes and on conformal interfaces. We also co-organized the Les Houches schools in 2001 and 2007. It will be a pleasure to revisit the topics we discussed and worked on over the years.

But these days I am working more on machine learning. Fortunately there is a highly developed interface between ML and field theory:

- Statistical physics and machine learning: Parisi, Mézard, Zecchina, Montanari, Monasson, Krzakala, Zdeborova, ...
- ML techniques for lattice FT: Cranmer, Shanahan, Urban, ...
- Exact RG and relations to information theory: Berman and Klinger, Cotler and Rezchikov, ...
- QFT-NN relations: Halverson, Maiti, Demertas, Schwartz, ...

So I will revisit topics discussed in “Calabi’s diastasis as interface entropy” 1311.2202 (also with Ilka Brunner and Leonardo Rastelli) in the light of what I have since learned about ML and statistics.

The main result of 1311.2202 was the following:

- Consider a family of $d = 2$ $(2, 2)$ SCFTs, with a moduli space \mathcal{M} of theories connected by (c, c) (complex structure) or (a, c) (Kähler) deformations.
- Consider theories T and T' in this family, associated to points in \mathcal{M} with coordinates t and t' .
- Then there is a superconformal interface between T and T' with boundary entropy g given by

$$2 \log g = K(t, \bar{t}) + K(t', \bar{t}') - K(t, \bar{t}') - K(t', \bar{t}) \quad (1)$$

where $K(t, \bar{t})$ is the Kähler potential on \mathcal{M} analytically continued to general t, \bar{t} .

The combination on the r.h.s. is the Calabi diastatic function (Calabi 1953). It is a function on $\mathcal{M} \times \mathcal{M}$ (the dependence on Kähler-Weyl transformations $K \rightarrow K + F(t) + \bar{F}(\bar{t})$ cancels out), zero if $t = t'$.

While interesting in its own right, our initial motivation for studying this was the idea that one could use such a conformal interface to define a natural distance between a pair of CFTs. Recall that there is a natural Riemannian metric on a moduli space \mathcal{M} of CFTs, the Zamolodchikov metric

$$g_{ij}^{(Zam)} = \langle \phi_i(0) \phi_j(1) \rangle \quad (2)$$

where $\langle \dots \rangle$ is the normalized correlation function and $\phi_i(z)$ are the marginal operators corresponding to tangent vectors in \mathcal{M} . Given a pair of points T, T' in the same connected component of \mathcal{M} , the length of the shortest path between them defines a distance $d(T, T')$.

But what if \mathcal{M} is not connected? Say T, T' are obtained from two distinct CY sigma models, not connected by varying moduli. Is there a natural definition of distance between T and T' (asked in 1005.2779) ?

Our idea was that a conformal interface between T and T' could be used to define

$$d(T, T') = \min \sqrt{\log g}. \quad (3)$$

While interesting in its own right, our initial motivation for studying this was the idea that one could use such a conformal interface to define a natural distance between a pair of CFTs. Recall that there is a natural Riemannian metric on a moduli space \mathcal{M} of CFTs, the Zamolodchikov metric

$$g_{ij}^{(Zam)} = \langle \phi_i(0) \phi_j(1) \rangle \quad (2)$$

where $\langle \dots \rangle$ is the normalized correlation function and $\phi_i(z)$ are the marginal operators corresponding to tangent vectors in \mathcal{M} . Given a pair of points T, T' in the same connected component of \mathcal{M} , the length of the shortest path between them defines a distance $d(T, T')$.

But what if \mathcal{M} is not connected? Say T, T' are obtained from two distinct CY sigma models, not connected by varying moduli. Is there a natural definition of distance between T and T' (asked in 1005.2779) ?

Our idea was that a conformal interface between T and T' could be used to define

$$d(T, T') = \min \sqrt{\log g}. \quad (3)$$

$$2 \log g = K(t, \bar{t}) + K(t', \bar{t}') - K(t, \bar{t}') - K(t', \bar{t}) \quad (4)$$

This works infinitesimally – one can show that the diastasis agrees with the Zamolodchikov metric to second order. But it does not work in general. A sensible distance must satisfy the triangle inequality,

$$d(x, y) \leq d(x, z) + d(z, y) \quad \forall x, y, z, \quad (5)$$

but the square root of the diastasis does not in general.

OK. There are other definitions given in 1005.2779, such as the following “quantum Gromov-Hausdorff distance.” Given theories T_1, T_2 , look at quantities like

$$d(T_1, T_2)^2 = \min_U \text{Tr} (U e^{-tH_1} U^\dagger - e^{-tH_2})^2 \quad (6)$$

where U is a unitary implementing duality equivalences. This can be generalized to higher genus diagrams with a reflection symmetry and U a topological interface.

$$2 \log g = K(t, \bar{t}) + K(t', \bar{t}') - K(t, \bar{t}') - K(t', \bar{t}) \quad (4)$$

This works infinitesimally – one can show that the diastasis agrees with the Zamolodchikov metric to second order. But it does not work in general. A sensible distance must satisfy the triangle inequality,

$$d(x, y) \leq d(x, z) + d(z, y) \quad \forall x, y, z, \quad (5)$$

but the square root of the diastasis does not in general.

OK. There are other definitions given in 1005.2779, such as the following “quantum Gromov-Hausdorff distance.” Given theories T_1 , T_2 , look at quantities like

$$d(T_1, T_2)^2 = \min_U \text{Tr} (U e^{-tH_1} U^\dagger - e^{-tH_2})^2 \quad (6)$$

where U is a unitary implementing duality equivalences. This can be generalized to higher genus diagrams with a reflection symmetry and U a topological interface.

But, we can also regard a QFT or CFT as a Euclidean path integral,

$$Z = \int [d\phi] e^{-S[\phi]}; \quad \langle \phi_1 \dots \phi_k \rangle = \frac{1}{Z} \int [d\phi] e^{-S[\phi]} \phi_1 \dots \phi_k. \quad (7)$$

The integrand is a probability measure over fields (for statistical field theories; more generally it can be complex). And there are many definitions of distances and related quantities for such measures. Why not use one of these?

Some families of probability distributions

- The “categorical” distributions $\mathcal{C}(P(1), \dots, P(k))$ with $k - 1$ parameters. Draw $x \in S$ from a finite set $1, 2, \dots, k$, and require $P(x) \geq 0 \forall x$ and $\sum_{x \in S} P(x) = 1$.
- The multivariate normal (Gaussian) distribution $\mathcal{N}(\vec{\mu}, \Sigma)$ on $\vec{x} \in \mathbb{R}^k$,

$$P_{\mu, \Sigma}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^k \det \Sigma}} \exp -\frac{1}{2}(\vec{x} - \vec{\mu}) \cdot \Sigma \cdot (\vec{x} - \vec{\mu}), \quad (8)$$

normalized so that $\int d\vec{x} P(x) = 1$. This is a family with parameters $\mu \in \mathbb{R}^k$ and symmetric positive definite $\Sigma \in \mathbb{R}^{k(k+1)/2}$.

What are natural distances between pairs of such distributions?
For example, is there a natural distance between a \mathcal{C} and a \mathcal{N} ?

Some families of probability distributions

- The “categorical” distributions $\mathcal{C}(P(1), \dots, P(k))$ with $k - 1$ parameters. Draw $x \in S$ from a finite set $1, 2, \dots, k$, and require $P(x) \geq 0 \forall x$ and $\sum_{x \in S} P(x) = 1$.
- The multivariate normal (Gaussian) distribution $\mathcal{N}(\vec{\mu}, \Sigma)$ on $\vec{x} \in \mathbb{R}^k$,

$$P_{\mu, \Sigma}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^k \det \Sigma}} \exp -\frac{1}{2}(\vec{x} - \vec{\mu}) \cdot \Sigma \cdot (\vec{x} - \vec{\mu}), \quad (8)$$

normalized so that $\int d\vec{x} P(x) = 1$. This is a family with parameters $\mu \in \mathbb{R}^k$ and symmetric positive definite $\Sigma \in \mathbb{R}^{k(k+1)/2}$.

What are natural distances between pairs of such distributions?
For example, is there a natural distance between a \mathcal{C} and a \mathcal{N} ?

Statistical and quantum systems

Canonical ensemble: $P(\phi) = (1/Z) \exp -E(\phi)$, Z for normalization.

For example, the generalized Ising or scalar field theory.

Given two metric spaces Σ and K , the random variable is a map $\phi : \Sigma \rightarrow K$, and we use the Boltzmann weight

$$E = -J \sum_{i \neq j \in \Sigma} \frac{d^2(\phi(i), \phi(j))}{d^2(i, j)} + \sum_{i \in \Sigma} V(\phi(i)). \quad (9)$$

Here $J > 0$ (attractive/ferromagnetic) or $J < 0$ (repulsive/antiferromagnetic).

For quantum systems, $P \rightarrow \rho$ (the density matrix) and $\int \rightarrow \text{Tr}$.

Spin glass: $J(i, j)$ can depend on i and j and can have both signs.

Statistical and quantum systems

Canonical ensemble: $P(\phi) = (1/Z) \exp -E(\phi)$, Z for normalization.

For example, the generalized Ising or scalar field theory.

Given two metric spaces Σ and K , the random variable is a map $\phi : \Sigma \rightarrow K$, and we use the Boltzmann weight

$$E = -J \sum_{i \neq j \in \Sigma} \frac{d^2(\phi(i), \phi(j))}{d^2(i, j)} + \sum_{i \in \Sigma} V(\phi(i)). \quad (9)$$

Here $J > 0$ (attractive/ferromagnetic) or $J < 0$ (repulsive/antiferromagnetic).

For quantum systems, $P \rightarrow \rho$ (the density matrix) and $\int \rightarrow \text{Tr}$.

Spin glass: $J(i, j)$ can depend on i and j and can have both signs.

Statistical and quantum systems

Canonical ensemble: $P(\phi) = (1/Z) \exp -E(\phi)$, Z for normalization.

For example, the generalized Ising or scalar field theory.

Given two metric spaces Σ and K , the random variable is a map $\phi : \Sigma \rightarrow K$, and we use the Boltzmann weight

$$E = -J \sum_{i \neq j \in \Sigma} \frac{d^2(\phi(i), \phi(j))}{d^2(i, j)} + \sum_{i \in \Sigma} V(\phi(i)). \quad (9)$$

Here $J > 0$ (attractive/ferromagnetic) or $J < 0$ (repulsive/antiferromagnetic).

For quantum systems, $P \rightarrow \rho$ (the density matrix) and $\int \rightarrow \text{Tr}$.

Spin glass: $J(i, j)$ can depend on i and j and can have both signs.

Statistical and quantum systems

Canonical ensemble: $P(\phi) = (1/Z) \exp -E(\phi)$, Z for normalization.

For example, the generalized Ising or scalar field theory.

Given two metric spaces Σ and K , the random variable is a map $\phi : \Sigma \rightarrow K$, and we use the Boltzmann weight

$$E = -J \sum_{i \neq j \in \Sigma} \frac{d^2(\phi(i), \phi(j))}{d^2(i, j)} + \sum_{i \in \Sigma} V(\phi(i)). \quad (9)$$

Here $J > 0$ (attractive/ferromagnetic) or $J < 0$ (repulsive/antiferromagnetic).

For quantum systems, $P \rightarrow \rho$ (the density matrix) and $\int \rightarrow \text{Tr}$.

Spin glass: $J(i, j)$ can depend on i and j and can have both signs.

Distances and information theory

- Kullback-Leibler divergence (KL divergence or relative entropy)

$$D_{\text{KL}}(P||Q) = \int dx P(x) \log \frac{P(x)}{Q(x)}. \quad (10)$$

Like a distance, $D_{\text{KL}} \geq 0$ with equality only if $P = Q$, but it is not symmetric between P and Q . It is the extra information required to encode a sample from P given a code optimized for Q .

- Fisher information metric – given a family P_t with parameters t , this is a Riemannian metric whose value at an (arbitrary) point $t = 0$ is

$$g_{ij}(P_0) = \left. \partial_i \partial_j D_{\text{KL}}(P_0 || P_t) \right|_{t=0}. \quad (11)$$

- Wasserstein distances (more later).

Exponential families

These definitions make sense for general families $P_t(x)$, but in stat mech and QFT we usually use the parameterization

$$P_t(x) = P_0(x) e^{F(t) - \sum t^i O_i} \quad (12)$$

with $F(t) = -\log Z = -\log \int dx P_0(x) \exp - \sum t^i O_i$.

This is the maximal entropy distribution with the set of expectation values $\langle O_i \rangle$. So this parameterization is much used by everyone.

For example, in the categorical distribution, taking $O_i = \delta(i, x)$ then $P(i) = \langle O_i \rangle = e^{-t^i} / \sum_j e^{-t^j}$. This is the “softmax” function of ML.

In general, the relation $t^i \leftrightarrow m_i \equiv \langle O_i \rangle$ is a duality (involution) given by Legendre transform $\Gamma(m_i) = \inf_t \sum_i t^i m_i - F(t)$.

KL divergence $D_{\text{KL}}(P(0) || P(t)) = \langle e^{-S(t)} \rangle_{P(0)} = \langle \sum t^i O_i \rangle_{P(0)}$.

Fisher information metric $g_{ij}(t) = -\partial_i \partial_j F = \langle O^i O^j \rangle_c$.

In this sense Fisher \sim Zamolodchikov.

Exponential families

These definitions make sense for general families $P_t(x)$, but in stat mech and QFT we usually use the parameterization

$$P_t(x) = P_0(x) e^{F(t) - \sum t^i O_i} \quad (12)$$

with $F(t) = -\log Z = -\log \int dx P_0(x) \exp - \sum t^i O_i$.

This is the maximal entropy distribution with the set of expectation values $\langle O_i \rangle$. So this parameterization is much used by everyone.

For example, in the categorical distribution, taking $O_i = \delta(i, x)$ then $P(i) = \langle O_i \rangle = e^{-t^i} / \sum_j e^{-t^j}$. This is the “softmax” function of ML.

In general, the relation $t^i \leftrightarrow m_i \equiv \langle O_i \rangle$ is a duality (involution) given by Legendre transform $\Gamma(m_i) = \inf_t \sum_i t^i m_i - F(t)$.

KL divergence $D_{\text{KL}}(P(0) || P(t)) = \langle e^{-S(t)} \rangle_{P(0)} = \langle \sum t^i O_i \rangle_{P(0)}$.

Fisher information metric $g_{ij}(t) = -\partial_i \partial_j F = \langle O^i O^j \rangle_c$.

In this sense Fisher \sim Zamolodchikov.

Exponential families

These definitions make sense for general families $P_t(x)$, but in stat mech and QFT we usually use the parameterization

$$P_t(x) = P_0(x) e^{F(t) - \sum t^i O_i} \quad (12)$$

with $F(t) = -\log Z = -\log \int dx P_0(x) \exp - \sum t^i O_i$.

This is the maximal entropy distribution with the set of expectation values $\langle O_i \rangle$. So this parameterization is much used by everyone.

For example, in the categorical distribution, taking $O_i = \delta(i, x)$ then $P(i) = \langle O_i \rangle = e^{-t^i} / \sum_j e^{-t^j}$. This is the “softmax” function of ML.

In general, the relation $t^i \leftrightarrow m_i \equiv \langle O_i \rangle$ is a duality (involution) given by Legendre transform $\Gamma(m_i) = \inf_t \sum_i t^i m_i - F(t)$.

KL divergence $D_{\text{KL}}(P(0)||P(t)) = \langle e^{-S(t)} \rangle_{P(0)} = \langle \sum t^i O_i \rangle_{P(0)}$.

Fisher information metric $g_{ij}(t) = -\partial_i \partial_j F = \langle O^i O^j \rangle_c$.

In this sense Fisher \sim Zamolodchikov.

Exponential families

These definitions make sense for general families $P_t(x)$, but in stat mech and QFT we usually use the parameterization

$$P_t(x) = P_0(x) e^{F(t) - \sum t^i O_i} \quad (12)$$

with $F(t) = -\log Z = -\log \int dx P_0(x) \exp - \sum t^i O_i$.

This is the maximal entropy distribution with the set of expectation values $\langle O_i \rangle$. So this parameterization is much used by everyone.

For example, in the categorical distribution, taking $O_i = \delta(i, x)$ then $P(i) = \langle O_i \rangle = e^{-t^i} / \sum_j e^{-t^j}$. This is the “softmax” function of ML.

In general, the relation $t^i \leftrightarrow m_i \equiv \langle O_i \rangle$ is a duality (involution) given by Legendre transform $\Gamma(m_i) = \inf_t \sum_i t^i m_i - F(t)$.

KL divergence $D_{\text{KL}}(P(0)||P(t)) = \langle e^{-S(t)} \rangle_{P(0)} = \langle \sum t^i O_i \rangle_{P(0)}$.

Fisher information metric $g_{ij}(t) = -\partial_i \partial_j F = \langle O^i O^j \rangle_c$.

In this sense Fisher \sim Zamolodchikov.

In field theory one often uses the generating functional of connected correlation functions

$$F(J) = -\log \int d\phi \exp -S_0(\phi) + i \int dx J(x)\phi(x). \quad (13)$$

This is also an exponential family.

The exponential family is defined in terms of coordinates t^i . How does one say this geometrically?

One can add couplings, it makes sense to say $t_1^i + t_2^i$. The space of couplings has an affine structure which can be described by an affine connection ∇ . It is flat and its simplest description is $\Gamma_{jk}^i = 0$ in the t coordinates.

There is more structure: information geometry. For example, there is also an affine connection on the space of vevs (moments) m_i . In the t coordinates its components are $\Gamma_{jk}^i = g^{il} \langle O_l O_j O_k \rangle_c$. The dual generating function is the 1PI effective action $\Gamma(m)$.

See for example Floerchinger 2303.04081.

In field theory one often uses the generating functional of connected correlation functions

$$F(J) = -\log \int d\phi \exp -S_0(\phi) + i \int dx J(x)\phi(x). \quad (13)$$

This is also an exponential family.

The exponential family is defined in terms of coordinates t^i . How does one say this geometrically?

One can add couplings, it makes sense to say $t_1^i + t_2^i$. The space of couplings has an affine structure which can be described by an affine connection ∇ . It is flat and its simplest description is $\Gamma_{jk}^i = 0$ in the t coordinates.

There is more structure: information geometry. For example, there is also an affine connection on the space of vevs (moments) m_i . In the t coordinates its components are $\Gamma_{jk}^i = g^{il} \langle O_l O_j O_k \rangle_c$. The dual generating function is the 1PI effective action $\Gamma(m)$.

See for example Floerchinger 2303.04081.

In field theory one often uses the generating functional of connected correlation functions

$$F(J) = -\log \int d\phi \exp -S_0(\phi) + i \int dx J(x)\phi(x). \quad (13)$$

This is also an exponential family.

The exponential family is defined in terms of coordinates t^i . How does one say this geometrically?

One can add couplings, it makes sense to say $t_1^i + t_2^i$. The space of couplings has an affine structure which can be described by an affine connection ∇ . It is flat and its simplest description is $\Gamma_{jk}^i = 0$ in the t coordinates.

There is more structure: information geometry. For example, there is also an affine connection on the space of vevs (moments) m_i . In the t coordinates its components are $\Gamma_{jk}^i = g^{il} \langle O_l O_j O_k \rangle_c$. The dual generating function is the 1PI effective action $\Gamma(m)$.

See for example Floerchinger 2303.04081.

In field theory one often uses the generating functional of connected correlation functions

$$F(J) = -\log \int d\phi \exp -S_0(\phi) + i \int dx J(x)\phi(x). \quad (13)$$

This is also an exponential family.

The exponential family is defined in terms of coordinates t^i . How does one say this geometrically?

One can add couplings, it makes sense to say $t_1^i + t_2^i$. The space of couplings has an affine structure which can be described by an affine connection ∇ . It is flat and its simplest description is $\Gamma_{jk}^i = 0$ in the t coordinates.

There is more structure: information geometry. For example, there is also an affine connection on the space of vevs (moments) m_i . In the t coordinates its components are $\Gamma_{jk}^i = g^{il} \langle O_l O_j O_k \rangle_c$. The dual generating function is the 1PI effective action $\Gamma(m)$.

See for example Floerchinger 2303.04081.

Optimal transport

M É M O I R E

SUR LA

THÉORIE DES DÉBLAIS ET DES REMBLAIS.

Par M. M O N G E.

LORSQU'ON doit transporter des terres d'un lieu dans un autre, on a coutume de donner le nom de *Déblai* au volume des terres que l'on doit transporter, & le nom de *Remblai* à l'espace qu'elles doivent occuper après le transport.

Le prix du transport d'une molécule étant, toutes choses d'ailleurs égales, proportionnel à son poids & à l'espace qu'on lui fait parcourir, & par conséquent le prix du transport total devant être proportionnel à la somme des produits des molécules multipliées chacune par l'espace parcouru, il s'en suit que le déblai & le remblai étant donnés de figure & de position, il n'est pas indifférent que telle molécule du déblai soit transportée dans tel ou tel autre endroit du remblai, mais qu'il y a une certaine distribution à faire des molécules du premier dans le second, d'après laquelle la somme de ces produits sera la moindre possible, & le prix du transport total fera un *minimum*.

Consider two distributions $p_1(\vec{x})$ and $p_2(\vec{y})$ of matter in space. Suppose the cost of moving one “molecule” of matter from x to y is the distance $d(x, y) = |x - y|$. The “earth mover’s distance” (Monge 1781) between them is the minimal cost to turn p_1 into p_2 by movement of molecules of matter.

Monge-Kantorovich-Wasserstein distances

This problem was generalized by (1) considering the p 'th power of distances and (2) instead of transporting all matter at a point x to y , allow splitting it up. This can be expressed in terms of a “coupling” $\pi(x, y)$ such that $p_1 = \int dy \pi$ and $p_2 = \int dx \pi$. Then the p -Wasserstein distance between p_1 and p_2 is

$$W_p(p_1, p_2) = \left(\inf_{\pi} \int dx dy \pi(x, y) d(x, y)^p \right)^{1/p}. \quad (14)$$

This makes sense even if p_1 and p_2 are distributions over different spaces, so we could use it to define a distance between (for example) the \mathcal{C} and \mathcal{N} distributions. But it depends on postulating a distance $d(x, y)$, so it does not give a unique definition.

Probability distributions are not universal unless we take a limit:

- Sum of independent variables $x = \sum_i x_i$ – normal distribution (central limit theorem).
- Sum of noncommuting (free) independent variables – semicircle distribution (as in RMT; free probability of Voiculescu).
- IR limit under renormalization group – conformal field theories.
- Continuum limit – more general statistical/quantum field theories; flows from UV to IR fixed point.

Renormalization and the RG are (of course) unavoidable elements of the questions we began the talk with. This often spoils positivity, e.g. $(x - y)^2$: requires subtractions.

Interesting recent progress on treating these in frameworks of probability theory and information theory:

- RG flow as optimal transport (Cotler and Rezchikov 2202.11737).
- RG flow as inverse Bayesian inference (Berman *et al*)
- Rigorous stochastic quantization (Hairer *et al*, Gubinelli and Hofmanova, ...).

Probability distributions are not universal unless we take a limit:

- Sum of independent variables $x = \sum_i x_i$ – normal distribution (central limit theorem).
- Sum of noncommuting (free) independent variables – semicircle distribution (as in RMT; free probability of Voiculescu).
- IR limit under renormalization group – conformal field theories.
- Continuum limit – more general statistical/quantum field theories; flows from UV to IR fixed point.

Renormalization and the RG are (of course) unavoidable elements of the questions we began the talk with. This often spoils positivity, e.g. $(x - y)^2$: requires subtractions.

Interesting recent progress on treating these in frameworks of probability theory and information theory:

- RG flow as optimal transport (Cotler and Rezchikov 2202.11737).
- RG flow as inverse Bayesian inference (Berman *et al*)
- Rigorous stochastic quantization (Hairer *et al*, Gubinelli and Hofmanova, ...).

Exact RG and diffusion in theory space

Polchinski equation: define a cutoff theory by cutting off the quadratic part of the action,

$$Z_\Lambda[J] = \int d\phi \exp -\frac{1}{2} \int d^D p \frac{\phi(p)\phi(-p)}{K_\Lambda(p^2)(p^2 + m^2)} - S_{\text{int}}[\phi] - \int J\phi, \quad (15)$$

then varying Λ can be compensated by varying the interaction $S_{\text{int}}[\phi]$ as

$$-\Lambda \frac{\partial}{\partial \Lambda} e^{-S_{\text{int}}[\phi]} = \frac{1}{2} \int d^D p \frac{\Lambda}{p^2 + m^2} \frac{\partial K_\Lambda(p^2)}{\partial \Lambda} \frac{\delta^2}{\delta\phi(p)\delta\phi(-p)} e^{-S_{\text{int}}[\phi]}. \quad (16)$$

This is a diffusion equation on the (unnormalized) probability distribution. It becomes a diffusion-convection equation on the normalized distribution $e^{-S_{\text{int}}}/Z$.

It can be generalized to the Wegner-Morris exact RG

$$-\Lambda \frac{\partial}{\partial \Lambda} P_\Lambda[\phi] = \int d^D x \frac{\delta}{\delta\phi(x)} (\Psi[\phi, x] P_\Lambda[\phi]). \quad (17)$$

Exact RG and diffusion in theory space

Polchinski equation: define a cutoff theory by cutting off the quadratic part of the action,

$$Z_\Lambda[J] = \int d\phi \exp -\frac{1}{2} \int d^D p \frac{\phi(p)\phi(-p)}{K_\Lambda(p^2)(p^2 + m^2)} - S_{\text{int}}[\phi] - \int J\phi, \quad (15)$$

then varying Λ can be compensated by varying the interaction $S_{\text{int}}[\phi]$ as

$$-\Lambda \frac{\partial}{\partial \Lambda} e^{-S_{\text{int}}[\phi]} = \frac{1}{2} \int d^D p \frac{\Lambda}{p^2 + m^2} \frac{\partial K_\Lambda(p^2)}{\partial \Lambda} \frac{\delta^2}{\delta\phi(p)\delta\phi(-p)} e^{-S_{\text{int}}[\phi]}. \quad (16)$$

This is a diffusion equation on the (unnormalized) probability distribution. It becomes a diffusion-convection equation on the normalized distribution $e^{-S_{\text{int}}}/Z$.

It can be generalized to the Wegner-Morris exact RG

$$-\Lambda \frac{\partial}{\partial \Lambda} P_\Lambda[\phi] = \int d^D x \frac{\delta}{\delta\phi(x)} (\Psi[\phi, x] P_\Lambda[\phi]). \quad (17)$$

RG and optimal transport (Cotler and Rezhchikov)

The Wegner-Morris RG flow can be thought of as a field reparameterization

$$\phi \rightarrow \phi + \frac{\delta\Lambda}{\Lambda} \Psi[\phi, \mathbf{x}] \quad (18)$$

and the special case of varying the cutoff is

$$\Psi[\phi, \mathbf{x}] = -\frac{1}{2} \int d^d y \dot{C}_\Lambda(\mathbf{x} - \mathbf{y}) \frac{\delta \Sigma_\Lambda[\phi]}{\delta \phi(\mathbf{y})} \quad (19)$$

with \dot{C} the variation of the propagator and $\Sigma_\Lambda = S[\phi] - 2\hat{S}[\phi]$ depends on a “seed” or reference action. Recall $P(\phi) = e^{-S[\phi]}/Z$ and define $P^{\text{ref}}(\phi) = e^{-2\hat{S}[\phi]}/\hat{Z}$ analogously. Then Cotler and Rezhchikov show that

$$-\Lambda \frac{\partial}{\partial \Lambda} P_\Lambda[\phi] = -\nabla_{\mathcal{W}_2} \text{D}_{\text{KL}}(P_\Lambda[\phi] \parallel P_\Lambda^{\text{ref}}[\phi]) . \quad (20)$$

where \mathcal{W}_2 (the metric in the gradient) is a Wasserstein metric defined using \dot{C} as the measure of distance. Intuitively the flow measures how much the reference action must be modified to become $P[\phi]$.

Statistics and theory of machine learning

A primary task in statistics is to, given a dataset and a family of probability distributions, find the distribution in the family which best fits the data. This is called model estimation or “training” in ML.

A very common procedure is stochastic gradient descent (SGD). In each step, one takes a sample from the dataset and varies the model parameters by the gradient of an objective function which measures the fit. For function fitting this could be least squares, while for a probability distribution one often takes relative entropy. For a dataset $\{x_i\}$ this looks like

$$\dot{t}^i = -G^{ij} \partial_j \sum_i \log P_t(x_i) \quad (21)$$

Ideally G^{ij} is the Fisher metric at P_t , though this is not usual in practice. One would like to know, for a given class of models and dataset, how many samples are needed to learn the distribution. This is known as sample complexity.

Statistics and theory of machine learning

A primary task in statistics is to, given a dataset and a family of probability distributions, find the distribution in the family which best fits the data. This is called model estimation or “training” in ML.

A very common procedure is stochastic gradient descent (SGD). In each step, one takes a sample from the dataset and varies the model parameters by the gradient of an objective function which measures the fit. For function fitting this could be least squares, while for a probability distribution one often takes relative entropy. For a dataset $\{x_i\}$ this looks like

$$\dot{t}^i = -G^{ij} \partial_j \sum_i \log P_t(x_i) \quad (21)$$

Ideally G^{ij} is the Fisher metric at P_t , though this is not usual in practice. One would like to know, for a given class of models and dataset, how many samples are needed to learn the distribution. This is known as sample complexity.

Statistics and theory of machine learning

A primary task in statistics is to, given a dataset and a family of probability distributions, find the distribution in the family which best fits the data. This is called model estimation or “training” in ML.

A very common procedure is stochastic gradient descent (SGD). In each step, one takes a sample from the dataset and varies the model parameters by the gradient of an objective function which measures the fit. For function fitting this could be least squares, while for a probability distribution one often takes relative entropy. For a dataset $\{x_i\}$ this looks like

$$\dot{t}^i = -G^{ij} \partial_j \sum_i \log P_t(x_i) \quad (21)$$

Ideally G^{ij} is the Fisher metric at P_t , though this is not usual in practice. One would like to know, for a given class of models and dataset, how many samples are needed to learn the distribution. This is known as sample complexity.

The purely theoretical problem of this type is model recovery, also called the teacher-student problem. One generates data using model M_{true} , usually chosen randomly from some parameterized set $P(t)$. One then starts with a second randomly initialized model M_0 and, by training on data generated by M_{true} , tries to reproduce it.

One can show that the gradient descent rule approximates the flow

$$\partial_t P(t) = -\nabla D_{\text{KL}}(M_{\text{true}} || P(t)); \quad i.e. \quad \dot{t}^i = -G^{ij} \partial_j D_{\text{KL}}(M_{\text{true}} || P(t)). \quad (22)$$

This can also be thought of as a continuous form of the Bayesian inference rule

$$P_{\text{post}}(t|x) = \frac{P(x|t)}{P(x)} P_{\text{prior}}(t). \quad (23)$$

Berman *et al* 2305.10491 suggest thinking of this as an inverse of the RG. The RG takes a theory and produces a simpler theory which does not describe the high energy observables. As we saw earlier it is mathematically a diffusion (or convection-diffusion) equation. By contrast Bayesian inference incorporates measurements and in examples looks like an inverse convection-diffusion equation.

The purely theoretical problem of this type is model recovery, also called the teacher-student problem. One generates data using model M_{true} , usually chosen randomly from some parameterized set $P(t)$. One then starts with a second randomly initialized model M_0 and, by training on data generated by M_{true} , tries to reproduce it.

One can show that the gradient descent rule approximates the flow

$$\partial_t P(t) = -\nabla D_{\text{KL}}(M_{\text{true}}||P(t)); \quad i.e. \quad \dot{t}^i = -G^{ij} \partial_j D_{\text{KL}}(M_{\text{true}}||P(t)). \quad (22)$$

This can also be thought of as a continuous form of the Bayesian inference rule

$$P_{\text{post}}(t|x) = \frac{P(x|t)}{P(x)} P_{\text{prior}}(t). \quad (23)$$

Berman *et al* 2305.10491 suggest thinking of this as an inverse of the RG. The RG takes a theory and produces a simpler theory which does not describe the high energy observables. As we saw earlier it is mathematically a diffusion (or convection-diffusion) equation. By contrast Bayesian inference incorporates measurements and in examples looks like an inverse convection-diffusion equation.


The purely theoretical problem of this type is model recovery, also called the teacher-student problem. One generates data using model M_{true} , usually chosen randomly from some parameterized set $P(t)$. One then starts with a second randomly initialized model M_0 and, by training on data generated by M_{true} , tries to reproduce it.

One can show that the gradient descent rule approximates the flow

$$\partial_t P(t) = -\nabla D_{\text{KL}}(M_{\text{true}} || P(t)); \quad i.e. \quad \dot{t}^i = -G^{ij} \partial_j D_{\text{KL}}(M_{\text{true}} || P(t)). \quad (22)$$

This can also be thought of as a continuous form of the Bayesian inference rule

$$P_{\text{post}}(t|x) = \frac{P(x|t)}{P(x)} P_{\text{prior}}(t). \quad (23)$$

Berman *et al* 2305.10491 suggest thinking of this as an inverse of the RG. The RG takes a theory and produces a simpler theory which does not describe the high energy observables. As we saw earlier it is mathematically a diffusion (or convection-diffusion) equation. By contrast Bayesian inference incorporates measurements and in examples looks like an inverse convection-diffusion equation. 

So what can we say about distances or divergences defined by integrating these flows? When are they finite or infinite?

For the learning problem, the distance corresponds to learning time (sample complexity). If $P(t)$ is close enough to the target model, then learning is fast: If $D_{\text{KL}}(M||P(t)) = |t|^2$ then $\dot{t}^i = -2t^i$ and time $T = \log d_{\text{init}}/d_{\text{final}}$.

Potential infinities come from local minima and/or saddle points. But in practice one does a noisy gradient flow (by randomly sampling the dataset) and this often fixes the problem.

Does RG flow go finite distance or infinite distance? The argument we just gave seems to say finite distance.

But in the Cotler-Rezchikov RG equation, the structure seems to lie in the \mathcal{W}_2 metric, which is not even universal. It is not obvious (to me) why there are fixed points.

So what can we say about distances or divergences defined by integrating these flows? When are they finite or infinite?

For the learning problem, the distance corresponds to learning time (sample complexity). If $P(t)$ is close enough to the target model, then learning is fast: If $D_{\text{KL}}(M||P(t)) = |t|^2$ then $\dot{t}^i = -2t^i$ and time $T = \log d_{\text{init}}/d_{\text{final}}$.

Potential infinities come from local minima and/or saddle points. But in practice one does a noisy gradient flow (by randomly sampling the dataset) and this often fixes the problem.

Does RG flow go finite distance or infinite distance? The argument we just gave seems to say finite distance.

But in the Cotler-Rezchikov RG equation, the structure seems to lie in the \mathcal{W}_2 metric, which is not even universal. It is not obvious (to me) why there are fixed points.

So what can we say about distances or divergences defined by integrating these flows? When are they finite or infinite?

For the learning problem, the distance corresponds to learning time (sample complexity). If $P(t)$ is close enough to the target model, then learning is fast: If $D_{\text{KL}}(M||P(t)) = |t|^2$ then $\dot{t}^i = -2t^i$ and time $T = \log d_{\text{init}}/d_{\text{final}}$.

Potential infinities come from local minima and/or saddle points. But in practice one does a noisy gradient flow (by randomly sampling the dataset) and this often fixes the problem.

Does RG flow go finite distance or infinite distance? The argument we just gave seems to say finite distance.

But in the Cotler-Rezchikov RG equation, the structure seems to lie in the \mathcal{W}_2 metric, which is not even universal. It is not obvious (to me) why there are fixed points.

So what can we say about distances or divergences defined by integrating these flows? When are they finite or infinite?

For the learning problem, the distance corresponds to learning time (sample complexity). If $P(t)$ is close enough to the target model, then learning is fast: If $D_{\text{KL}}(M||P(t)) = |t|^2$ then $\dot{t}^i = -2t^i$ and time $T = \log d_{\text{init}}/d_{\text{final}}$.

Potential infinities come from local minima and/or saddle points. But in practice one does a noisy gradient flow (by randomly sampling the dataset) and this often fixes the problem.

Does RG flow go finite distance or infinite distance? The argument we just gave seems to say finite distance.

But in the Cotler-Rezchikov RG equation, the structure seems to lie in the \mathcal{W}_2 metric, which is not even universal. It is not obvious (to me) why there are fixed points.

Can we find a field theory version of the Wasserstein distance?

$$W_2(p_1, p_2)^2 = \inf_{\pi} \int d\phi_1 d\phi_2 (\phi_1 - \phi_2)^2 \pi(\phi_1, \phi_2) \quad (24)$$

where $\pi(\phi_1, \phi_2)$ is such that $e^{-S_1[\phi_1]}/Z_1 = \int d\phi_2 \pi$ and $e^{-S_2[\phi_2]}/Z_2 = \int d\phi_1 \pi$.

The coupling π is a field theory depending on both fields $\phi_{1,2}$. Simplest example is the direct product FT. We could perturb around this to get others. The normalization condition is the standard one.

If $T_1 \cong T_2$ then π should be the identity $\delta(\phi_1 - \phi_2)e^{-S_1}$. More generally replace δ by an interface.

Then we face the problem of defining $(\phi_1 - \phi_2)^2$ for operators whose coincidence limit is singular. Maybe better to look at a simpler observable about π , such as its central charge if it is a CFT, or boundary entropy for an interface. This would bring us back to the definition that Bachas, Brunner, Rastelli and I studied.

Put a boundary operator on the interface to represent $(\phi_1 - \phi_2)^2$?

Can we find a field theory version of the Wasserstein distance?

$$W_2(p_1, p_2)^2 = \inf_{\pi} \int d\phi_1 d\phi_2 (\phi_1 - \phi_2)^2 \pi(\phi_1, \phi_2) \quad (24)$$

where $\pi(\phi_1, \phi_2)$ is such that $e^{-S_1[\phi_1]}/Z_1 = \int d\phi_2 \pi$ and $e^{-S_2[\phi_2]}/Z_2 = \int d\phi_1 \pi$.

The coupling π is a field theory depending on both fields $\phi_{1,2}$. Simplest example is the direct product FT. We could perturb around this to get others. The normalization condition is the standard one.

If $T_1 \cong T_2$ then π should be the identity $\delta(\phi_1 - \phi_2)e^{-S_1}$. More generally replace δ by an interface.

Then we face the problem of defining $(\phi_1 - \phi_2)^2$ for operators whose coincidence limit is singular. Maybe better to look at a simpler observable about π , such as its central charge if it is a CFT, or boundary entropy for an interface. This would bring us back to the definition that Bachas, Brunner, Rastelli and I studied.

Put a boundary operator on the interface to represent $(\phi_1 - \phi_2)^2$?

Best wishes, Costas !!!