A Training Manual for Bayesian Mechanics

Dalton A R Sakthivadivel

Stony Brook University; VERSES Research Lab

25th July 2022

Outline

1. Brief
2. Parameterisation / Bayesian Inference
3. Inference / Surprisal
4. Statistical Physics / Statistical Manifolds
5. Debrief

Slides can be found later at darsakthi.github.io/talks

For supplementary material, see /research and /blog

Based on 2204.05084, 2205.11543, 2206.12996 (+ others)


Ben Nevis Inn
WEATHER FORECASTING STONE
CONDITION
FORECAST
Stone is Wet
Rain
Stone is Dry
Not Raining
Shadow on Ground
Sunny
White on Top
Snowing
Can't see Stone
Foggy
Swinging Stone
Windy
Stone Jumping Up & Down
Earthquake
Stone Gone
Tornado

Coupled random dynamics

Let a random dynamical system consist of both dynamics on a state space (deterministic dynamics, F) and a probability space (noise, ω). One has a coupled RDS when the degrees of freedom of one or more RDSs interact with a given RDS (typically bidirectionally)

Ex: additive Wiener noise $\mathrm{d}X_t = F(X_t, t)\,\mathrm{d}t + \mathrm{d}W_t$. Suppose the state of some other system appears in F; this is coupled

In physics, coupled RDSs are a proxy for complexity: they are generically open, path-dependent, non-linear, and out-of-equilibrium; they fluctuate and exhibit signs of complexity, and do computations

Coupled random dynamics

Let a random dynamical system consist of both dynamics on a state space (deterministic dynamics, F) and a probability space (noise, ω). One has a coupled RDS when the degrees of freedom of one or more RDSs interact with a given RDS (typically bidirectionally)

Ex: additive Wiener noise $\mathrm{d}X_t = F(X_t, t)\,\mathrm{d}t + \mathrm{d}W_t$. Suppose the state of some other system appears in F; this is coupled

In physics, coupled RDSs are a proxy for complexity: they are generically open, path-dependent, non-linear, and out-of-equilibrium; they fluctuate and exhibit signs of complexity, and do computations

Coupled random dynamics

Let a random dynamical system consist of both dynamics on a state space (deterministic dynamics, F) and a probability space (noise, ω). One has a coupled RDS when the degrees of freedom of one or more RDSs interact with a given RDS (typically bidirectionally)

Ex: additive Wiener noise $\mathrm{d}X_t = F(X_t, t)\,\mathrm{d}t + \mathrm{d}W_t$. Suppose the state of some other system appears in F; this is coupled

In physics, coupled RDSs are a proxy for complexity: they are generically open, path-dependent, non-linear, and out-of-equilibrium; they fluctuate and exhibit signs of complexity, and do computations

Coupled random dynamics

Let a random dynamical system consist of both dynamics on a state space (deterministic dynamics, F) and a probability space (noise, ω). One has a coupled RDS when the degrees of freedom of one or more RDSs interact with a given RDS (typically bidirectionally)

Ex: additive Wiener noise $\mathrm{d}X_t = F(X_t, t)\,\mathrm{d}t + \mathrm{d}W_t$. Suppose the state of some other system appears in F; this is coupled

In physics, coupled RDSs are a proxy for complexity: they are generically open, path-dependent, non-linear, and out-of-equilibrium; they fluctuate and exhibit signs of complexity, and do computations

Why is the synchronisation of statistics interesting?

Two (related) problems:

- Hard to understand coupled random dynamical systems using analysis
- Hard to understand coupled random dynamical systems using physics

Two (hopefully related) solutions:

- Reduce complex systems to Bayesian estimators or controllers; avoid stochastic analysis, uses other tools
- Deduce a simpler physics for inferential objects and map it back to ‘material physics’

Why is the synchronisation of statistics interesting?

Two (related) problems:

- Hard to understand coupled random dynamical systems using analysis
- Hard to understand coupled random dynamical systems using physics

Two (hopefully related) solutions:

- Reduce complex systems to Bayesian estimators or controllers; avoid stochastic analysis, uses other tools
- Deduce a simpler physics for inferential objects and map it back to 'material physics'

Relation to other fields

Bayesian mechanics draws on constructions in cybernetics, as well as the free energy principle and active inference

Connects to the physics of non-equilibria via stochastic thermodynamics and the principle of maximum calibre

Augments methods in stochastic analysis like path-wise analysis and large deviations principles

Connections to supersymmetry suggest a well-axiomatised physics of complexity grounded in topology

So far nothing in applied category theory

Relation to other fields

Bayesian mechanics draws on constructions in cybernetics, as well as the free energy principle and active inference

Connects to the physics of non-equilibria via stochastic thermodynamics and the principle of maximum calibre

Augments methods in stochastic analysis like path-wise analysis and large deviations principles

Connections to supersymmetry suggest a well-axiomatised physics of complexity grounded in topology

So far nothing in applied category theory

Relation to other fields

Bayesian mechanics draws on constructions in cybernetics, as well as the free energy principle and active inference

Connects to the physics of non-equilibria via stochastic thermodynamics and the principle of maximum calibre

Augments methods in stochastic analysis like path-wise analysis and large deviations principles

Connections to supersymmetry suggest a well-axiomatised physics of complexity grounded in topology

So far nothing *explicitly* **in applied category theory**

Synchronisation

There are two ingredients to synchronisation between coupled random dynamical systems:

- separation
- estimation

These define inference

Inference and estimation

There exists some map

$$\text{stone} \xrightarrow{\text{forecast}} \text{weather},$$
$$\text{forecast}(\text{stone}) \mapsto \text{weather}$$

such that the stone and the weather are coupled

This relation is conditional on the boundary of the stone interacting with weather states

Extracting ‘weather’ from ‘stone’ requires knowing that map, and, that the map is an injection

We call mapping ‘stone’ to ‘weather’ estimation

Inference and estimation

There exists some map

$$\text{stone} \xrightarrow{\text{forecast}} \text{weather},$$
$$\text{forecast(stone)} \mapsto \text{weather}$$

such that the stone and the weather are coupled

This relation is conditional on the boundary of the stone interacting with weather states

Extracting ‘weather’ from ‘stone’ requires knowing that map, and, that the map is an injection

We call mapping ‘stone’ to ‘weather’ estimation

Inference and estimation

There exists some map

$$\text{stone} \xrightarrow{\text{forecast}} \text{weather},$$
$$\text{forecast}(\text{stone}) \mapsto \text{weather}$$

such that the stone and the weather are coupled

This relation is conditional on the boundary of the stone interacting with weather states

Extracting ‘weather’ from ‘stone’ requires knowing that map, and, that the map is an injection

We call mapping ‘stone’ to ‘weather’ estimation

Inference and estimation

There exists some map

$$\text{stone} \xrightarrow{\text{forecast}} \text{weather},$$
$$\text{forecast}(\text{stone}) \mapsto \text{weather}$$

such that the stone and the weather are coupled

This relation is conditional on the boundary of the stone interacting with weather states

Extracting ‘weather’ from ‘stone’ requires knowing that map, and, that the map is an injection

We call mapping ‘stone’ to ‘weather’ estimation

Inference and estimation

Let states of the stone be denoted by μ (resp weather states by η) and let σ be the forecast. Let b be the boundary ('Markov blanket') between the two systems. System states are related by this common interface

Then there exists a map $\sigma : \mu(b) \mapsto \eta(b)$

μ estimates η by being the preimage of σ

η can be inferred* from μ by σ

Inference and estimation

Let states of the stone be denoted by μ (resp weather states by η) and let σ be the forecast. Let b be the boundary ('Markov blanket') between the two systems. System states are related by this common interface

Then there exists a map $\sigma : \mu(b) \mapsto \eta(b)$

μ estimates η by being the preimage of σ

η can be inferred* from μ by σ

Inference and estimation

Let states of the stone be denoted by μ (resp weather states by η) and let σ be the forecast. Let b be the boundary ('Markov blanket') between the two systems. System states are related by this common interface

Then there exists a map $\sigma : \mu(b) \mapsto \eta(b)$

μ estimates η by being the preimage of σ

η can be inferred* from μ by σ

Inference and estimation

Let states of the stone be denoted by μ (resp weather states by η) and let σ be the forecast. Let b be the boundary ('Markov blanket') between the two systems. System states are related by this common interface

Then there exists a map $\sigma : \mu(b) \mapsto \eta(b)$

μ estimates η by being the preimage of σ

η can be inferred* from μ by σ

Inference and estimation

Let states of the stone be denoted by μ (resp weather states by η) and let σ be the forecast. Let b be the boundary ('Markov blanket') between the two systems. System states are related by this common interface

Then there exists a map $\sigma : \mu(b) \mapsto \eta(b)$

μ estimates η by being the preimage of σ

η can be inferred* from μ by σ

*More like read off: no uncertainty introduced yet

Instant inference: just add noise

Suppose μ (resp η) is a state of a random dynamical system with probability measure $p(\mu)\,\mathrm{d}\mu$ (resp $p(\eta)\,\mathrm{d}\eta$)

Moreover suppose the mean $\hat{\mu}$ is a sufficient statistic for $p(\mu)\,\mathrm{d}\mu$ (and likewise for η)

Now define conditional expectations, and a map $\sigma : \hat{\mu}(b) \to \hat{\eta}(b)$

One system estimates the parameters of the probability density over states of the other:

Instant inference: just add noise

Suppose μ (resp η) is a state of a random dynamical system with probability measure $p(\mu)\,\mathrm{d}\mu$ (resp $p(\eta)\,\mathrm{d}\eta$)

Moreover suppose the mean $\hat{\mu}$ is a sufficient statistic for $p(\mu)\,\mathrm{d}\mu$ (and likewise for η)

Now define conditional expectations, and a map $\sigma : \hat{\mu}(b) \to \hat{\eta}(b)$

One system estimates the parameters of the probability density over states of the other:

Instant inference: just add noise

Suppose μ (resp η) is a state of a random dynamical system with probability measure $p(\mu)\,\mathrm{d}\mu$ (resp $p(\eta)\,\mathrm{d}\eta$)

Moreover suppose the mean $\hat{\mu}$ is a sufficient statistic for $p(\mu)\,\mathrm{d}\mu$ (and likewise for η)

Now define conditional expectations, and a map $\sigma : \hat{\mu}(b) \to \hat{\eta}(b)$

One system estimates the parameters of the probability density over states of the other:

Instant inference: just add noise

Suppose μ (resp η) is a state of a random dynamical system with probability measure $p(\mu)\,\mathrm{d}\mu$ (resp $p(\eta)\,\mathrm{d}\eta$)

Moreover suppose the mean $\hat{\mu}$ is a sufficient statistic for $p(\mu)\,\mathrm{d}\mu$ (and likewise for η)

Now define conditional expectations, and a map $\sigma : \hat{\mu}(b) \to \hat{\eta}(b)$

One system estimates the parameters of the probability density over states of the other:

Instant inference: just add noise

Suppose μ (resp η) is a state of a random dynamical system with probability measure $p(\mu)\,\mathrm{d}\mu$ (resp $p(\eta)\,\mathrm{d}\eta$)

Moreover suppose the mean $\hat{\mu}$ is a sufficient statistic for $p(\mu)\,\mathrm{d}\mu$ (and likewise for η)

Now define conditional expectations, and a map $\sigma : \hat{\mu}(b) \to \hat{\eta}(b)$

One system estimates the parameters of the probability density over states of the other: Bayesian inference

Separation

These statistics are conditional on the state of the boundary, b

(i.e., $\hat{\mu}(b)$, $\hat{\eta}(b)$ are conditional expectations, conditioned on some event b)

We think of these systems as interacting via shared b states. . .
. . . but we also want them to be separable

Separation

These statistics are conditional on the state of the boundary, b

(i.e., $\hat{\mu}(b)$, $\hat{\eta}(b)$ are conditional expectations, conditioned on some event b)

We think of these systems as interacting via shared b states. . .
. . . but we also want them to be separable

Separation

These statistics are conditional on the state of the boundary, b

(i.e., $\hat{\mu}(b)$, $\hat{\eta}(b)$ are conditional expectations, conditioned on some event b)

We think of these systems as interacting via shared b states...
...but we also want them to be separable

Separation

These statistics are conditional on the state of the boundary, b

(i.e., $\hat{\mu}(b)$, $\hat{\eta}(b)$ are conditional expectations, conditioned on some event b)

We think of these systems as interacting via shared b states...
...but we also want them to be separable

Separation

Q: What makes two random dynamical system different?

A: They are statistically distinguishable: observed samples, statistical properties, etc, are different $\iff$ non-mixing, distinct systems (see arXiv:2207.07620 for a sketch)

Q: How do we model systems which couple but remain statistically distinct?

A: The parameters $\hat{\mu}$ and $\hat{\eta}$ must be independent conditioned on the state b

Separation

Q: What makes two random dynamical system different?

A: They are statistically distinguishable: observed samples, statistical properties, etc, are different $\iff$ non-mixing, distinct systems (see arXiv:2207.07620 for a sketch)

Q: How do we model systems which couple but remain statistically distinct?

A: The parameters $\hat{\mu}$ and $\hat{\eta}$ must be independent conditioned on the state b

Separation

Q: What makes two random dynamical system different?

A: They are statistically distinguishable: observed samples, statistical properties, etc, are different $\iff$ non-mixing, distinct systems (see arXiv:2207.07620 for a sketch)

Q: How do we model systems which couple but remain statistically distinct?

A: The parameters $\hat{\mu}$ and $\hat{\eta}$ must be independent conditioned on the state b

Separation

Q: What makes two random dynamical system different?

A: They are statistically distinguishable: observed samples, statistical properties, etc, are different $\iff$ non-mixing, distinct systems (see arXiv:2207.07620 for a sketch)

Q: How do we model systems which couple but remain statistically distinct?

A: The parameters $\hat{\mu}$ and $\hat{\eta}$ must be independent conditioned on the state b (this is a sort of closure condition)

Separation and boundaries

Ref: Lemma 4.3, arXiv:2204.11900

Consider the following factorisation of our set of systems:

$$\hat{\mu}_b \xrightarrow{\boldsymbol{\mu}^{-1}} b \xrightarrow{\boldsymbol{\eta}} \hat{\eta}_b$$

where each such intermediate map is linear in b

If we write this sequence of evaluations as the tensor-Hom adjunct

$$[b, \hat{\eta}_b] \otimes ([\hat{\mu}_b, b] \otimes \hat{\mu}_b) \to \hat{\eta}_b \cong [b, \hat{\eta}_b] \otimes [\hat{\mu}_b, b] \to [\hat{\mu}_b, \hat{\eta}_b]$$

we recover a function $\sigma = (\boldsymbol{\eta} \circ \boldsymbol{\mu}^{-1})(b)$ we can do calculations with

Separation and boundaries

Ref: Lemma 4.3, arXiv:2204.11900

Consider the following factorisation of our set of systems:

$$\hat{\mu}_b \xrightarrow{\boldsymbol{\mu}^{-1}} b \xrightarrow{\boldsymbol{\eta}} \hat{\eta}_b$$

where each such intermediate map is linear in b

If we write this sequence of evaluations as the tensor-Hom adjunct

$$[b, \hat{\eta}_b] \otimes ([\hat{\mu}_b, b] \otimes \hat{\mu}_b) \to \hat{\eta}_b \cong [b, \hat{\eta}_b] \otimes [\hat{\mu}_b, b] \to [\hat{\mu}_b, \hat{\eta}_b]$$

we recover a function $\sigma = (\boldsymbol{\eta} \circ \boldsymbol{\mu}^{-1})(b)$ we can do calculations with

Separation and boundaries

Ref: Lemma 4.3, arXiv:2204.11900

Consider the following factorisation of our set of systems:

$$\hat{\mu}_b \xrightarrow{\boldsymbol{\mu}^{-1}} b \xrightarrow{\boldsymbol{\eta}} \hat{\eta}_b$$

where each such intermediate map is linear in b

If we write this sequence of evaluations as the tensor-Hom adjunct

$$[b, \hat{\eta}_b] \otimes ([\hat{\mu}_b, b] \otimes \hat{\mu}_b) \to \hat{\eta}_b \cong [b, \hat{\eta}_b] \otimes [\hat{\mu}_b, b] \to [\hat{\mu}_b, \hat{\eta}_b]$$

we recover a function $\sigma = (\boldsymbol{\eta} \circ \boldsymbol{\mu}^{-1})(b)$ we can do calculations with

Putting the pieces in place

We have

- two systems separated by a boundary
- a coupling, allowing us to relate the systems across that boundary
- a relationship between the statistics of the systems that induces inference

How can we use this to model two systems?

Parameterisation

Suppose $\hat{\eta}_b$ is a sufficient statistic. Then there exists a density $q(\eta; \hat{\eta}_b)$. By synchronisation this is equivalent to $q(\eta; \sigma(\hat{\mu}_b))$.

Variational inference: when (on average) the system obtains the state expected for a given blanket, it models its environment.

Mathematically: $q(\eta|\mu)$ where $\mu = \hat{\mu}_b$ yields $q(\eta)$ by $\sigma(\hat{\mu}_b) = \hat{\eta}_b$.

Not guaranteed that $q(\eta|\mu = \hat{\mu}_b) = p(\eta)$—depends on how useful the parameter is

Parameterisation

Suppose $\hat{\eta}_b$ is a sufficient statistic. Then there exists a density $q(\eta; \hat{\eta}_b)$. By synchronisation this is equivalent to $q(\eta; \sigma(\hat{\mu}_b))$.

Variational inference: when (on average) the system obtains the state expected for a given blanket, it models its environment.

Mathematically: $q(\eta|\mu)$ where $\mu = \hat{\mu}_b$ yields $q(\eta)$ by $\sigma(\hat{\mu}_b) = \hat{\eta}_b$.

Not guaranteed that $q(\eta|\mu = \hat{\mu}_b) = p(\eta)$—depends on how useful the parameter is

Parameterisation

Suppose $\hat{\eta}_b$ is a sufficient statistic. Then there exists a density $q(\eta; \hat{\eta}_b)$. By synchronisation this is equivalent to $q(\eta; \sigma(\hat{\mu}_b))$.

Variational inference: when (on average) the system obtains the state expected for a given blanket, it models its environment.

Mathematically: $q(\eta|\mu)$ where $\mu = \hat{\mu}_b$ yields $q(\eta)$ by $\sigma(\hat{\mu}_b) = \hat{\eta}_b$.

Not guaranteed that $q(\eta|\mu = \hat{\mu}_b) = p(\eta)$—depends on how useful the parameter is

Parameterisation

Suppose $\hat{\eta}_b$ is a sufficient statistic. Then there exists a density $q(\eta; \hat{\eta}_b)$. By synchronisation this is equivalent to $q(\eta; \sigma(\hat{\mu}_b))$.

Variational inference: when (on average) the system obtains the state expected for a given blanket, it models its environment.

Mathematically: $q(\eta|\mu)$ where $\mu = \hat{\mu}_b$ yields $q(\eta)$ by $\sigma(\hat{\mu}_b) = \hat{\eta}_b$.

Not guaranteed that $q(\eta|\mu = \hat{\mu}_b) = p(\eta)$—depends on how useful the parameter is

Parameterisation

This can be written as a maximum entropy problem

Under the constraint that $q = p$ we have

$$-\int q \log q + \int q \log p = 0 \tag{1}$$

for the entropy of q.

Solution is $p = q$.

Parameterisation

This can be written as a maximum entropy problem

Under the constraint that $q = p$ we have

$$-\int q \log q + \int q \log p = 0 \tag{1}$$

for the entropy of q.

Solution is $p = q$.

Parameterisation

This can be written as a maximum entropy problem

Under the constraint that $q = p$ we have

$$-\int q \log q + \int q \log p = 0 \tag{1}$$

for the entropy of q.

Solution is $p = q$.

Parameterisation

Now use the fact that $\sigma(\hat{\mu}_b)$ is a sufficient statistic for q. Under the constraint that η minimises its distance from $\sigma(\hat{\mu}_b)$, we have

$$-\int q \log q - \lambda \int [\eta - \sigma(\hat{\mu}_b)]^2 q = 0$$

Solution becomes

$$-\log q = \lambda[\eta - \sigma(\hat{\mu}_b)]^2 \quad (2)$$

which is a Gaussian with variance $\lambda[\eta - \sigma(\hat{\mu}_b)]^2$.

Parameterisation

Now use the fact that $\sigma(\hat{\mu}_b)$ is a sufficient statistic for q. Under the constraint that η minimises its distance from $\sigma(\hat{\mu}_b)$, we have

$$-\int q \log q - \lambda \int [\eta - \sigma(\hat{\mu}_b)]^2 q = 0$$

Solution becomes

$$-\log q = \lambda[\eta - \sigma(\hat{\mu}_b)]^2 \quad (2)$$

which is a Gaussian with variance $\lambda[\eta - \sigma(\hat{\mu}_b)]^2$.

Parameterisation

Now use the fact that $\sigma(\hat{\mu}_b)$ is a sufficient statistic for q. Under the constraint that η minimises its distance from $\sigma(\hat{\mu}_b)$, we have

$$-\int q \log q - \lambda \int [\eta - \sigma(\hat{\mu}_b)]^2 q = 0$$

Solution becomes

$$-\log q = \lambda[\eta - \sigma(\hat{\mu}_b)]^2 \qquad (2)$$

which is a Gaussian with variance $\lambda[\eta - \sigma(\hat{\mu}_b)]^2$.

Some observations on the last

Eq (1) is a generalised free energy functional

Eq (2) says a system minimises surprisal when it does inference (i.e. when we have μ such that $\eta = \sigma(\hat{\mu}_b)$

We call this mode-matching: system fits a Gaussian density q to p by matching the mean of q to the most likely state of p (Laplace approximation)

Some observations on the last

Eq (1) is a generalised free energy functional

Eq (2) says a system minimises surprisal when it does inference (i.e. when we have μ such that $\eta = \sigma(\hat{\mu}_b)$)

We call this mode-matching: system fits a Gaussian density q to p by matching the mean of q to the most likely state of p (Laplace approximation)

Some observations on the last

Eq (1) is a generalised free energy functional

Eq (2) says a system minimises surprisal when it does inference (i.e. when we have μ such that $\eta = \sigma(\hat{\mu}_b)$)

We call this mode-matching: system fits a Gaussian density q to p by matching the mean of q to the most likely state of p (Laplace approximation)

Why are these interesting observations?

Recovers the free energy principle and active inference

Phrases coupled RDSs as surprisal minimisers or mode-matchers

Why are these interesting observations?

Recovers the free energy principle and active inference

Phrases coupled RDSs as surprisal minimisers or mode-matchers

A cybernetical example

Imagine a system with a preferred $\hat{\mu}_b$. There are two ways to minimise surprisal:

- Update preferences $\hat{\mu}_b$
- Change the system to which it is coupled such that $\sigma^{-1}(\hat{\eta}_b) = \hat{\mu}_b$

If a system embodies its environment, then action is changing what is embodied: this is a control theory

Open question

Can this be extended to planning and the control of future states?

A cybernetical example

Imagine a system with a preferred $\hat{\mu}_b$. There are two ways to minimise surprisal:

- Update preferences $\hat{\mu}_b$
- Change the system to which it is coupled such that $\sigma^{-1}(\hat{\eta}_b) = \hat{\mu}_b$

If a system embodies its environment, then action is changing what is embodied: this is a control theory

Open question

Can this be extended to planning and the control of future states?

A cybernetical example

Imagine a system with a preferred $\hat{\mu}_b$. There are two ways to minimise surprisal:

- Update preferences $\hat{\mu}_b$
- Change the system to which it is coupled such that $\sigma^{-1}(\hat{\eta}_b) = \hat{\mu}_b$

If a system embodies its environment, then action is changing what is embodied: this is a control theory

Open question

Can this be extended to planning and the control of future states?

A cybernetical example

Imagine a system with a preferred $\hat{\mu}_b$. There are two ways to minimise surprisal:

- Update preferences $\hat{\mu}_b$
- Change the system to which it is coupled such that $\sigma^{-1}(\hat{\eta}_b) = \hat{\mu}_b$

If a system embodies its environment, then action is changing what is embodied: this is a control theory

Open question

Can this be extended to planning and the control of future states?

Gradient ascents

Due to our Laplace assumption we can determine that a gradient descent on surprisal is a gradient ascent on probability

$\equiv$ the least surprising state is the most likely state

This allows us to write coupled random dynamics as a particular gradient flow on surprisal

Reference: Markov blankets, information geometry and stochastic thermodynamics and related papers like arXiv:2106.13830, arXiv:2205.07793

This recovers quite a lot of physics by representing the statistics of fluctuations as dynamics on surprisal

Gradient ascents

Due to our Laplace assumption we can determine that a gradient descent on surprisal is a gradient ascent on probability

$\equiv$ the least surprising state is the most likely state

This allows us to write coupled random dynamics as a particular gradient flow on surprisal

Reference: Markov blankets, information geometry and stochastic thermodynamics and related papers like arXiv:2106.13830, arXiv:2205.07793

This recovers quite a lot of physics by representing the statistics of fluctuations as dynamics on surprisal

Gradient ascents

Due to our Laplace assumption we can determine that a gradient descent on surprisal is a gradient ascent on probability

$\equiv$ the least surprising state is the most likely state

This allows us to write coupled random dynamics as a particular gradient flow on surprisal

Reference: Markov blankets, information geometry and stochastic thermodynamics and related papers like arXiv:2106.13830, arXiv:2205.07793

This recovers quite a lot of physics by representing the statistics of fluctuations as dynamics on surprisal

Gradient ascents

Due to our Laplace assumption we can determine that a gradient descent on surprisal is a gradient ascent on probability

$\equiv$ the least surprising state is the most likely state

This allows us to write coupled random dynamics as a particular gradient flow on surprisal

Reference: Markov blankets, information geometry and stochastic thermodynamics and related papers like arXiv:2106.13830, arXiv:2205.07793

This recovers quite a lot of physics by representing the statistics of fluctuations as dynamics on surprisal

Gradient ascents

Due to our Laplace assumption we can determine that a gradient descent on surprisal is a gradient ascent on probability

$\equiv$ the least surprising state is the most likely state

This allows us to write coupled random dynamics as a particular gradient flow on surprisal

Reference: Markov blankets, information geometry and stochastic thermodynamics and related papers like arXiv:2106.13830, arXiv:2205.07793

This recovers quite a lot of physics by representing the statistics of fluctuations as dynamics on surprisal

Statistical properties of physical dynamics

A physical state μ now carries a belief parameterised by that state, with the optimal belief lying at $\hat{\mu}_b$

Changes in $\hat{\mu}_b$ change the belief held: Bayesian updating

As in all of statistical physics we have changed from physical dynamics to statistical properties of those dynamics

Statistical properties of physical dynamics

A physical state μ now carries a belief parameterised by that state, with the optimal belief lying at $\hat{\mu}_b$

Changes in $\hat{\mu}_b$ change the belief held: Bayesian updating

As in all of statistical physics we have changed from physical dynamics to statistical properties of those dynamics

Statistical properties of physical dynamics

A physical state μ now carries a belief parameterised by that state, with the optimal belief lying at $\hat{\mu}_b$

Changes in $\hat{\mu}_b$ change the belief held: Bayesian updating

As in all of statistical physics we have changed from physical dynamics to statistical properties of those dynamics

Motion on statistical manifolds

By changing $\hat{\mu}_b$ we change the belief embodied by the system: motion in a space of probability densities

This is called a statistical manifold ($S(-)$). The study of statistical manifolds is information geometry

Since the systems are coupled we have a *dual information geometry* where points in one statistical manifold can be written in terms of points in the other

Let f be a function on $S(\mu)$. Derivations on this manifold are constructed so as to yield $\nabla f(\mu) = \nabla f(\sigma^{-1}(\eta))\nabla\sigma^{-1}(\eta)$

Leads to an intuition that there is something lens-y going on

Motion on statistical manifolds

By changing $\hat{\mu}_b$ we change the belief embodied by the system: motion in a space of probability densities

This is called a statistical manifold ($S(-)$). The study of statistical manifolds is information geometry

Since the systems are coupled we have a *dual information geometry* where points in one statistical manifold can be written in terms of points in the other

Let f be a function on $S(\mu)$. Derivations on this manifold are constructed so as to yield $\nabla f(\mu) = \nabla f(\sigma^{-1}(\eta))\nabla\sigma^{-1}(\eta)$

Leads to an intuition that there is something lens-y going on

Motion on statistical manifolds

By changing $\hat{\mu}_b$ we change the belief embodied by the system: motion in a space of probability densities

This is called a statistical manifold ($S(-)$). The study of statistical manifolds is information geometry

Since the systems are coupled we have a *dual information geometry* where points in one statistical manifold can be written in terms of points in the other

Let f be a function on $S(\mu)$. Derivations on this manifold are constructed so as to yield $\nabla f(\mu) = \nabla f(\sigma^{-1}(\eta))\nabla\sigma^{-1}(\eta)$

Leads to an intuition that there is something lens-y going on

Motion on statistical manifolds

By changing $\hat{\mu}_b$ we change the belief embodied by the system: motion in a space of probability densities

This is called a statistical manifold ($S(-)$). The study of statistical manifolds is information geometry

Since the systems are coupled we have a *dual information geometry* where points in one statistical manifold can be written in terms of points in the other

Let f be a function on $S(\mu)$. Derivations on this manifold are constructed so as to yield $\nabla f(\mu) = \nabla f(\sigma^{-1}(\eta))\nabla\sigma^{-1}(\eta)$

Leads to an intuition that there is something lens-y going on

Motion on statistical manifolds

By changing $\hat{\mu}_b$ we change the belief embodied by the system: motion in a space of probability densities

This is called a statistical manifold ($S(-)$). The study of statistical manifolds is information geometry

Since the systems are coupled we have a *dual information geometry* where points in one statistical manifold can be written in terms of points in the other

Let f be a function on $S(\mu)$. Derivations on this manifold are constructed so as to yield $\nabla f(\mu) = \nabla f(\sigma^{-1}(\eta))\nabla\sigma^{-1}(\eta)$

Leads to an intuition that there is something lens-y going on

Concluding remarks

- Things that are coupled do statistical estimation
- Things that are coupled minimise surprisal when those estimates are ‘right’
- This means things that are coupled do inference
- But not always well
- This inference-from-coupling admits action
- It also recovers some physics