Gaussian Information Bottleneck and the Non-Perturbative Renormalization Group

Adam Kline and Stephanie Palmer

Presented by: Achint Kumar

Duke University

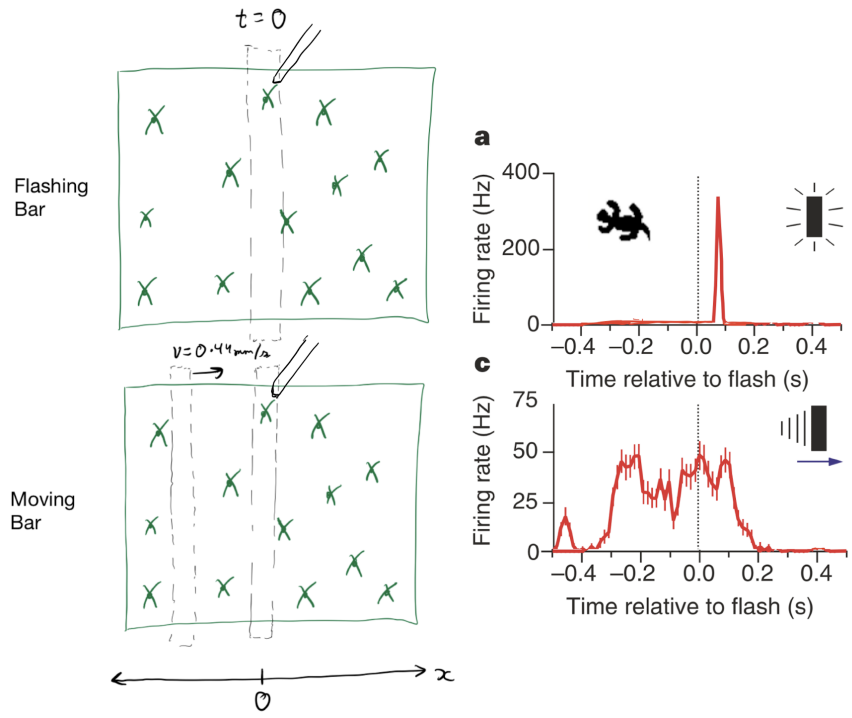
October 7, 2021

Desiderata

Desiderata

Motion is predicted by retinal ganglion cells

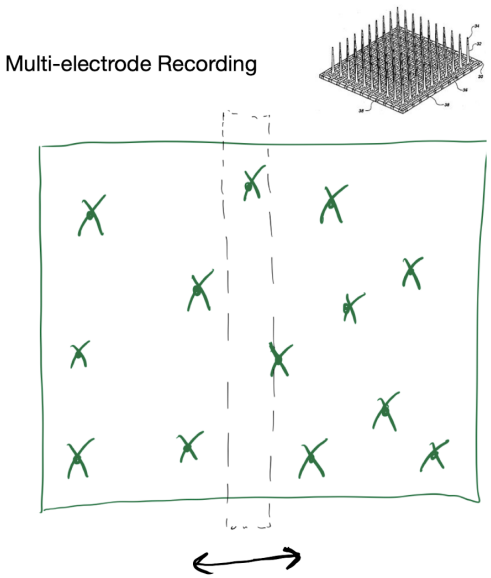
- A flashing bar and moving bar was presented to salamander retina and response of a ganglion cell was recorded.
- Retinal ganglion cells are able to predict motion of bar upto 1mm/s speed

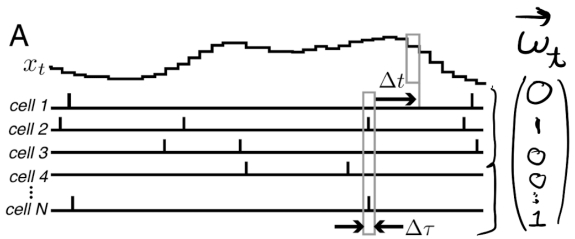


Credit: *Berry and Meister 1999.*

Recording from population of ganglion cells

- The bar performs a stochastic motion. We record from many($N \sim 50$) ganglion cells and create a binary vector $\vec{w}(t)$ representing the spike activity of neurons
- Question: How much information about bar's position is encoded in ganglion cells?





$$\frac{d^2x(t)}{dt^2} = -\omega^2 x(t) + \gamma v(t) + \eta(t)$$

Credit: *Palmer et.al.*, 2015

How much information about bar's position is encoded in ganglion cells?

Answer: We need to find mutual information between position at time t' and spike activity vector at time t.

$$I(w_t, x_{t'}) = \sum_{w_t, x_{t'}} P(w_t, x_{t'}) \log \left(\frac{P(w_t, x_{t'})}{P(w_t) P(x_{t'})} \right)$$

$$= \sum_{w_t, x_{t'}} P(w_t) P(x_{t'}|w_t) \log \left(\frac{P(x_{t'}|w_t)}{P(x_{t'})} \right)$$

To estimate $I(w_t, x_{t'})$ we do the following:

1. $P(x_{t'})$ is univariate distribution and can calculated analytically by solving Fokker-Planck equation or estimated by frequency analysis
2. $P(w_t)$ is a N dimensional distribution. To prevent undersampling we use $N \leq 7$ and then again use frequency analysis to estimate it
3. $P(x_{t'}|w_t)$ is univariate distribution and is estimated by frequency analysis

How much information about bar's position is encoded in ganglion cells?

Answer: We need to find mutual information between position at time t' and spike activity vector at time t.

$$I(w_t, x_{t'}) = \sum_{w_t, x_{t'}} P(w_t, x_{t'}) \log \left(\frac{P(w_t, x_{t'})}{P(w_t) P(x_{t'})} \right)$$

$$= \sum_{w_t, x_{t'}} P(w_t) P(x_{t'}|w_t) \log \left(\frac{P(x_{t'}|w_t)}{P(x_{t'})} \right)$$

To estimate $I(w_t, x_{t'})$ we do the following:

1. $P(x_{t'})$ is univariate distribution and can calculated analytically by solving Fokker-Planck equation or estimated by frequency analysis
2. $P(w_t)$ is a N dimensional distribution. To prevent undersampling we use $N \leq 7$ and then again use frequency analysis to estimate it
3. $P(x_{t'}|w_t)$ is univariate distribution and is estimated by frequency analysis

How much information about bar's position is encoded in ganglion cells?

Answer: We need to find mutual information between position at time t' and spike activity vector at time t.

$$I(w_t, x_{t'}) = \sum_{w_t, x_{t'}} P(w_t, x_{t'}) \log \left(\frac{P(w_t, x_{t'})}{P(w_t)P(x_{t'})} \right)$$

$$= \sum_{w_t, x_{t'}} P(w_t)P(x_{t'}|w_t) \log \left(\frac{P(x_{t'}|w_t)}{P(x_{t'})} \right)$$

To estimate $I(w_t, x_{t'})$ we do the following:

1. $P(x_{t'})$ is univariate distribution and can calculated analytically by solving Fokker-Planck equation or estimated by frequency analysis
2. $P(w_t)$ is a N dimensional distribution. To prevent undersampling we use $N \leq 7$ and then again use frequency analysis to estimate it
3. $P(x_{t'}|w_t)$ is univariate distribution and is estimated by frequency analysis

How much information about bar's position is encoded in ganglion cells?

Answer: We need to find mutual information between position at time t' and spike activity vector at time t.

$$I(w_t, x_{t'}) = \sum_{w_t, x_{t'}} P(w_t, x_{t'}) \log \left(\frac{P(w_t, x_{t'})}{P(w_t) P(x_{t'})} \right)$$

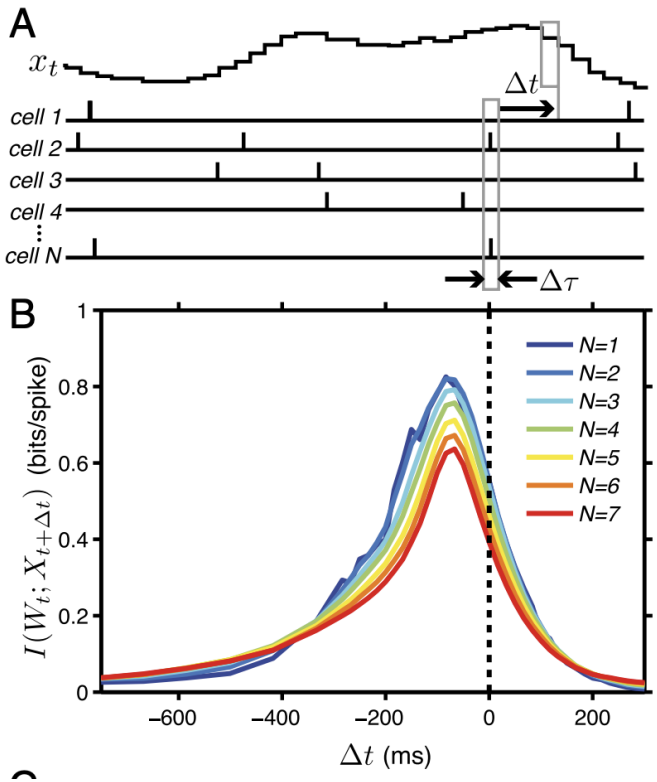
$$= \sum_{w_t, x_{t'}} P(w_t) P(x_{t'}|w_t) \log \left(\frac{P(x_{t'}|w_t)}{P(x_{t'})} \right)$$

To estimate $I(w_t, x_{t'})$ we do the following:

1. $P(x_{t'})$ is univariate distribution and can calculated analytically by solving Fokker-Planck equation or estimated by frequency analysis
2. $P(w_t)$ is a N dimensional distribution. To prevent undersampling we use $N \leq 7$ and then again use frequency analysis to estimate it
3. $P(x_{t'}|w_t)$ is univariate distribution and is estimated by frequency analysis

Mutual information result

1. Retina is most informative about the position of the object at $\Delta t = -80$ms because of latency in response.
2. Neural responses carry information about the position that extends far into the past and into the future.
3. Notice bits/spike goes down slightly with increasing N. This indicates redundant coding.



Credit: *Palmer et.al.*, 2015

Let's focus on future prediction: Information Bottleneck

In Information Bottleneck framework, we assume brain performs a trade-off between maximally predicting the future while minimally representing the the past.

$$X_{\text{past}} \xrightarrow{p(z|x_{past})} Z \xrightarrow{p(x_{future}|z)} X_{\text{future}}$$

The following objective function is minimized,

$$\min_{p(z|x_{past})} \mathcal{L} = I(X_{\text{past}}, Z) - \beta I(Z, X_{\text{future}})$$

The parameter β sets the trade-off between compression (reducing the information that we keep about the past, $I(X_{\text{past}}, Z)$ and prediction [increasing the information that we keep about the future, $I(Z, X_{\text{future}})$]

Let's focus on future prediction: Information Bottleneck

In Information Bottleneck framework, we assume brain performs a trade-off between maximally predicting the future while minimally representing the the past.

$$X_{\text{past}} \xrightarrow{p(z|x_{past})} Z \xrightarrow{p(x_{future}|z)} X_{\text{future}}$$

The following objective function is minimized,

$$\min_{p(z|x_{past})} \mathcal{L} = I(X_{\text{past}}, Z) - \beta I(Z, X_{\text{future}})$$

The parameter β sets the trade-off between compression (reducing the information that we keep about the past, $I(X_{\text{past}}, Z)$ and prediction [increasing the information that we keep about the future, $I(Z, X_{\text{future}})$]

Solving the Information Bottleneck problem

Objective function is:

$$\min_{p(z|x_{past})} \mathcal{L} = I(X_{\text{past}}, Z) - \beta I(Z, X_{\text{future}})$$

Since, $p(z|x_{past})$ must be normalized, we instead consider the add a Lagrange multiplier to the objective function.

$$\min_{p(z|x_{past})} \mathcal{L} = I(X_{\text{past}}, Z) - \beta I(Z, X_{\text{future}}) - \sum_{x_{past}, z} \lambda(x_{past})(p(z|x_{past}) - 1)$$

We perform, $\frac{\delta \mathcal{L}}{\delta p(z|x_{past})} = 0$ to get,

$$p(z|x_{past}) = \frac{1}{\mathcal{Z}_\beta(x_{past})} \exp\left[-\beta D_{KL}(p(x_{future}|x_{past})||p(x_{future}|z))\right]$$

Note: We don't know $p(x_{future}|x_{past})$ or $p(x_{future}|z)$

Solving the Information Bottleneck problem

Objective function is:

$$\min_{p(z|x_{past})} \mathcal{L} = I(X_{\text{past}}, Z) - \beta I(Z, X_{\text{future}})$$

Since, $p(z|x_{past})$ must be normalized, we instead consider the add a Lagrange multiplier to the objective function.

$$\min_{p(z|x_{past})} \mathcal{L} = I(X_{\text{past}}, Z) - \beta I(Z, X_{\text{future}}) - \sum_{x_{past}, z} \lambda(x_{past})(p(z|x_{past}) - 1)$$

We perform, $\frac{\delta \mathcal{L}}{\delta p(z|x_{past})} = 0$ to get,

$$p(z|x_{past}) = \frac{1}{\mathcal{Z}_\beta(x_{past})} \exp\left[-\beta D_{KL}(p(x_{future}|x_{past})||p(x_{future}|z))\right]$$

Note: We don't know $p(x_{future}|x_{past})$ or $p(x_{future}|z)$

Solving the Information Bottleneck problem: Blahut-Arimoto algorithm

We have the following set of equations:

$$p(z|x_{past}) = \frac{p(z)}{\mathcal{Z}_\beta(x_{past})} \exp\left[-\beta D_{KL}(p(x_{future}|x_{past})||p(x_{future}|z))\right] \tag{1}$$

$$p(z) = \sum_{x_{past}} p(z|x_{past})p(x_{past}) \tag{2}$$

$$p(x_{future}|z) = \frac{1}{p(z)} \sum_{x_{past}} p(x_{future}|x_{past})p(z|x_{past})p(x_{past}) \tag{3}$$

These can be solved iteratively using Blahut-Arimoto algorithm. In gaussian information bottleneck framework(coming soon!) these can be solved analytically.

Retinal population saturate the predictive bound

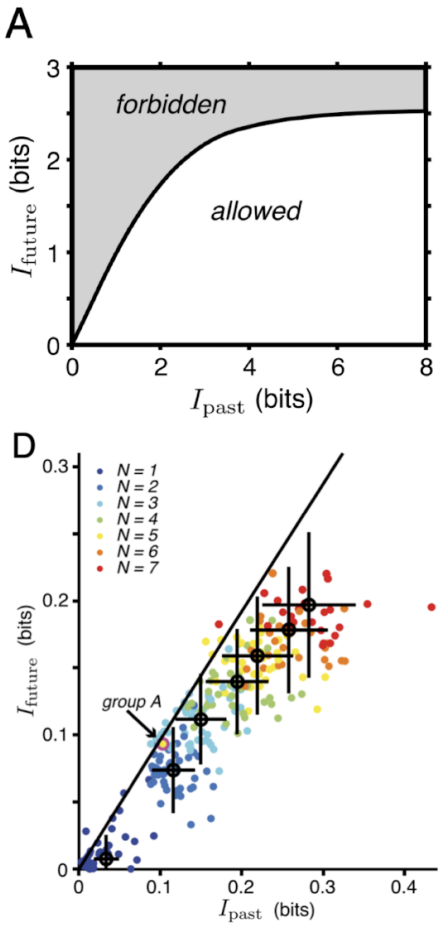
Recall, IB objective function is:

$$\mathcal{L} = I_{past} - \beta I_{future}$$

where,

$$I_{past} \triangleq I(X_{\text{past}}, Z)$$
$$I_{future} \triangleq I(Z, X_{\text{future}})$$

1. The ganglion cells maximally encode information about the future



Credit: *Palmer et.al.*, 2015

Gaussian Information Bottleneck(GIB)

$$X_{\text{past}} \xrightarrow{p(z|x_{past})} Z \xrightarrow{p(x_{future}|z)} X_{\text{future}}$$

From now on, in accordance with the paper I will use $X = X_{\text{past}}$,$\tilde{X} = Z$ and $Y = X_{\text{future}}$ In GIB framework, we assume that p(x,y) is jointly Gaussian. I will assume mean=0 while the covariance matrix looks like:

$$\Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY} & \Sigma_Y \end{pmatrix}$$

We must have,

$$\tilde{X} = AX + \xi$$

where ξ is a gaussian white noise($\sim \mathcal{N}(0, \Sigma_\xi)$)

GIB solution

We have,

$$X \xrightarrow{p(\tilde{x}|x)} \tilde{X} \xrightarrow{p(y|\tilde{x})} Y$$

The objective function is,

$$\mathcal{L} = I(X, \tilde{X}) - \beta I(\tilde{X}, Y)$$

The compression transformation is,

$$\tilde{X} = AX + \xi$$

For any β, the exact solution turns out to be,

$$\Sigma_\xi = \mathrm{I}$$

$$A(\beta) = \mathrm{diag}(\alpha_i(\beta))V^T$$

The matrix V represents the set of eigenvectors of $\Sigma_X^{-1}\Sigma_{X|Y}$, and $\alpha_i(\beta)$ is a complicated function of beta, eigenvalue of $\Sigma_X^{-1}\Sigma_{X|Y}$, etc.

GIB solution

We have,

$$X \xrightarrow{p(\tilde{x}|x)} \tilde{X} \xrightarrow{p(y|\tilde{x})} Y$$

The objective function is,

$$\mathcal{L} = I(X, \tilde{X}) - \beta I(\tilde{X}, Y)$$

The compression transformation is,

$$\tilde{X} = AX + \xi$$

For any β, the exact solution turns out to be,

$$\Sigma_\xi = \mathrm{I}$$

$$A(\beta) = \mathrm{diag}(\alpha_i(\beta))V^T$$

The matrix V represents the set of eigenvectors of $\Sigma_X^{-1}\Sigma_{X|Y}$, and $\alpha_i(\beta)$ is a complicated function of beta, eigenvalue of $\Sigma_X^{-1}\Sigma_{X|Y}$, etc.

Reparameterization of GIB

The solution to GIB is not unique. Let's say for some β we have IB optimal solutions (A, Σ_ξ). Then,

$$X \xrightarrow[\beta]{(A,\Sigma_\xi)} \tilde{X} \to Y$$

Let's imagine having two latent variables instead.

$$X \xrightarrow[\beta_1]{(A_1,\Sigma_{\xi_1})} \tilde{X}_1 \xrightarrow[\beta_2]{(A_2,\Sigma_{\xi_2})} \tilde{X}_2 \to Y$$

It turns out,

$$\beta = \frac{\beta_2 \beta_1}{\beta_2 + \beta_1 - 1}$$

$$A = A_2 A_1$$

$$\Sigma_\xi = A_2 A_2^T + \mathrm{I}$$

Reparameterization of GIB

The solution to GIB is not unique. Let's say for some β we have IB optimal solutions (A, Σ_ξ). Then,

$$X \xrightarrow[\beta]{(A,\Sigma_\xi)} \tilde{X} \to Y$$

Let's imagine having two latent variables instead.

$$X \xrightarrow[\beta_1]{(A_1,\Sigma_{\xi_1})} \tilde{X}_1 \xrightarrow[\beta_2]{(A_2,\Sigma_{\xi_2})} \tilde{X}_2 \to Y$$

It turns out,

$$\beta = \frac{\beta_2 \beta_1}{\beta_2 + \beta_1 - 1}$$
$$A = A_2 A_1$$
$$\Sigma_\xi = A_2 A_2^T + \mathrm{I}$$

Semi-group structure of GIB

We have the following composition law:

$$\beta = \beta_2 \circ \beta_1 = \frac{\beta_2 \beta_1}{\beta_2 + \beta_1 - 1}$$

Direct computations show that the composition operator satisfies closure and associativity, and thus furnishes the space in which β values live, that is $\mathbb{R} > 1$.

$\beta = \infty$ is the identity element.

β's form with a semi-group structure because there is no inverse. That is, there is no β' such that $\beta' \circ \beta = \mathrm{I}(\infty)$.

Semi-group structure of GIB

We have the following composition law:

$$\beta = \beta_2 \circ \beta_1 = \frac{\beta_2 \beta_1}{\beta_2 + \beta_1 - 1}$$

Direct computations show that the composition operator satisfies closure and associativity, and thus furnishes the space in which β values live, that is $\mathbb{R} > 1$.
$\beta = \infty$ is the identity element.
β's form with a semi-group structure because there is no inverse. That is, there is no β' such that $\beta' \circ \beta = \mathrm{I}(\infty)$.

Desiderata

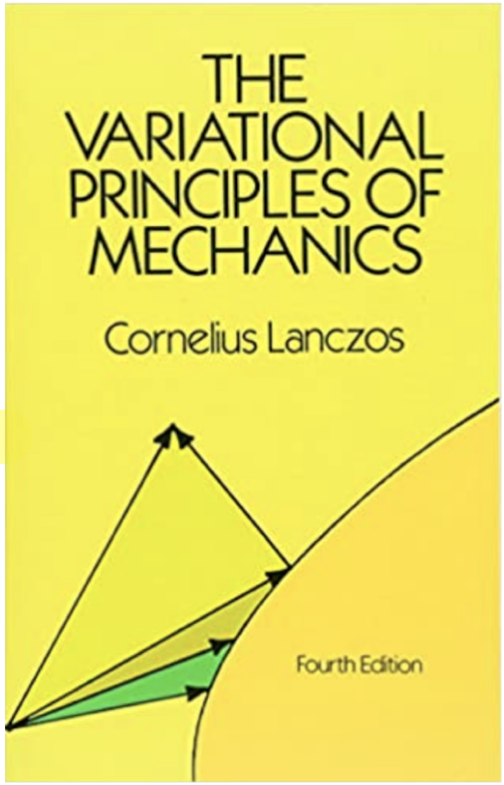
Renormalization Group: Introduction

CHAPTER VIII

THE PARTIAL DIFFERENTIAL EQUATION OF HAMILTON-JACOBI

Put off thy shoes from off thy feet, for the place whereon thou standest is holy ground. EXODUS III, 5

Introduction. We have done considerable mountain climbing. Now we are in the rarefied atmosphere of theories of excessive beauty and we are nearing a high plateau on which geometry, optics, mechanics, and wave mechanics meet on common ground. Only concentrated thinking, and a considerable amount of re-creation, will reveal the full beauty of our subject in which the last word has not yet been spoken. We start with the integration theory of Jacobi and continue with Hamilton's own investigations in the realm of geometrical optics and mechanics. The combination of these two approaches leads to de Broglie's and Schroedinger's great discoveries, and we come to the end of our journey.



Some mysteries: Why is nature simple?

1. Why simple models of neurons(like LIF or Izhikevich neurons) are able to reproduce so much of complexity of real neurons?
2. Why Wilson-Cowan equations are able to describe whole population of neurons using scalar function with scalar coupling?
3. Controversial: Deep learning algorithms employ a generalized RG-like scheme for feature learning and data compression

Renormalization group perspective

The degrees of freedom depend on length/time scale we are probing. The aforementioned models are an *effective theory* that represents the effective degrees of freedom at the scale we're probing.The evolution of parameters as we eliminate the uninteresting degrees of freedom is what we mean by renormalization.

Some mysteries: Why is nature simple?

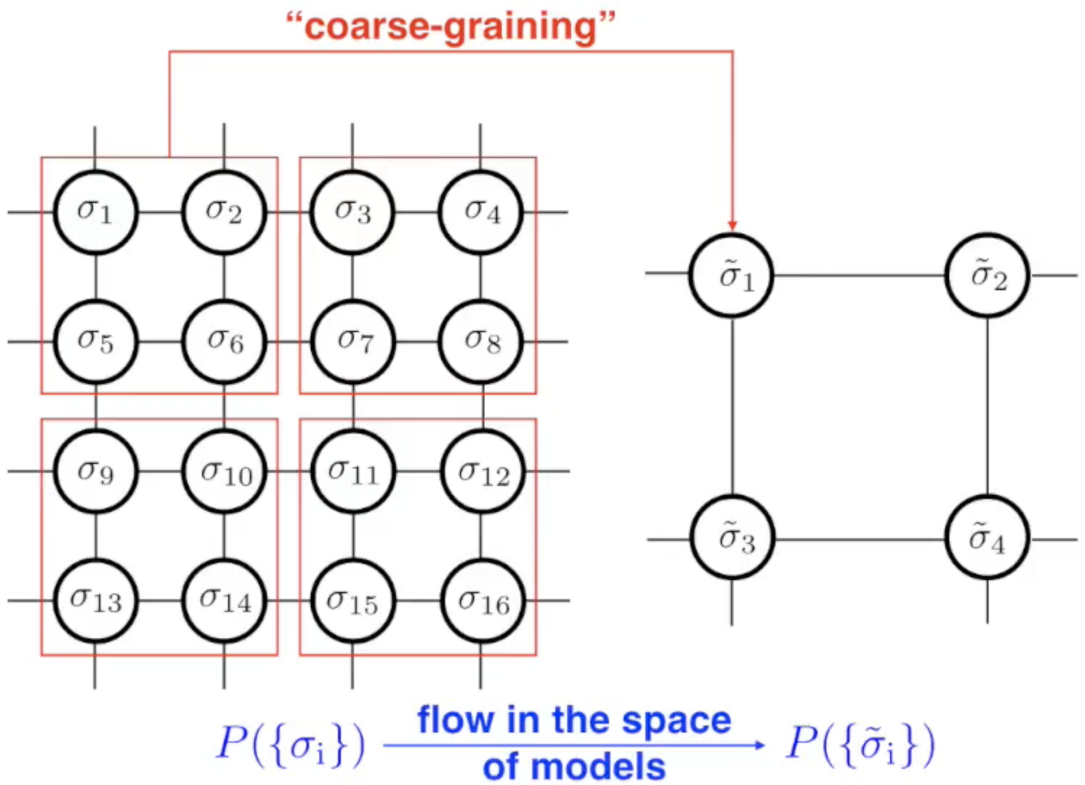
1. Why simple models of neurons(like LIF or Izhikevich neurons) are able to reproduce so much of complexity of real neurons?
2. Why Wilson-Cowan equations are able to describe whole population of neurons using scalar function with scalar coupling?
3. Controversial: Deep learning algorithms employ a generalized RG-like scheme for feature learning and data compression

Renormalization group perspective

The degrees of freedom depend on length/time scale we are probing. The aforementioned models are an *effective theory* that represents the effective degrees of freedom at the scale we're probing.The evolution of parameters as we eliminate the uninteresting degrees of freedom is what we mean by renormalization.

A concrete example

- In Ising model(or say Hopfield model), one can imagine combining blocks of neurons together by some rule. This would lead to new degrees of freedom and new couplings between the neurons.
- Notice that the coarse-graining process doesn't have an inverse. It is a semi-group(like IB).



Non-perturbative renormalization group

1. Suppose we are given a partition function,

$$\mathcal{Z} = \sum_i e^{-\beta E_i} \rightarrow \int \mathcal{D}\phi e^{-S(\phi)+\Delta S[\phi]+J\phi}$$

Here $\Delta S[\phi]$ is a regulator to ensure convergence of the integral and J is the source term(like external field).

2. Now, we consider a probabilistic mapping $\phi \xrightarrow{p(\tilde{\phi}|\phi)} \tilde{\phi}$. The map follows a gaussian distribution, $p(\tilde{\phi}|\phi) \sim \mathcal{N}(A\phi, \Sigma)$.
3. The paper gives expressions for flow of A, Σ, J and parameters defining the regulator by this mapping. .

Non-perturbative renormalization group

1. Suppose we are given a partition function,

$$\mathcal{Z} = \sum_i e^{-\beta E_i} \rightarrow \int \mathcal{D}\phi e^{-S(\phi)+\Delta S[\phi]+J\phi}$$

   Here $\Delta S[\phi]$ is a regulator to ensure convergence of the integral and J is the source term(like external field).

2. Now, we consider a probabilistic mapping $\phi \xrightarrow{p(\tilde{\phi}|\phi)} \tilde{\phi}$. The map follows a gaussian distribution, $p(\tilde{\phi}|\phi) \sim \mathcal{N}(A\phi, \Sigma)$.
3. The paper gives expressions for flow of A, Σ, J and parameters defining the regulator by this mapping. .

Non-perturbative renormalization group

1. Suppose we are given a partition function,

$$\mathcal{Z} = \sum_i e^{-\beta E_i} \rightarrow \int \mathcal{D}\phi e^{-S(\phi)+\Delta S[\phi]+J\phi}$$

   Here $\Delta S[\phi]$ is a regulator to ensure convergence of the integral and J is the source term(like external field).

2. Now, we consider a probabilistic mapping $\phi \xrightarrow{p(\tilde{\phi}|\phi)} \tilde{\phi}$. The map follows a gaussian distribution, $p(\tilde{\phi}|\phi) \sim \mathcal{N}(A\phi, \Sigma)$.
3. The paper gives expressions for flow of A, Σ, J and parameters defining the regulator by this mapping. .

Desiderata

Comparision between GIB and NPRG

1. In GIB, the goal is to find $p(\tilde{x}|x)$ for some β.
2. In NPRG, we are given $p(\tilde{\phi}|\phi)$ and the goal is to find the flow equations of parameters
3. In the paper, the authors highlighted the semi-group structure and the probabilistic mapping in both frameworks. It remains unclear how these connections can be brought to actual use in solving physics/deep learning problems.