Approximated Newton Algorithm for the Ising Model Inference Speeds Up Convergence, Performs Optimally and Avoids Over-fitting

Ulisse Ferrari

Institut de la Vision, Sorbonne Universités, UPMC

New Frontiers in Non-equilibrium Physics 2015

Outlook of the seminar

Model Inference:

Finding the probability distribution reproducing the data system statistics.

Model Inference:

Finding the probability distribution reproducing the data system statistics.

Useful for characterizing the behavior of systems of many, strongly correlated, units:

neurons, proteins, virus, species distribution, bird flocks

but...

Model Inference:

Finding the probability distribution reproducing the data system statistics.

Useful for characterizing the behavior of systems of many, strongly correlated, units:

neurons, proteins, virus, species distribution, bird flocks

but...

which distribution?

Model Inference:

Finding the probability distribution reproducing the data system statistics.

Useful for characterizing the behavior of systems of many, strongly correlated, units:

neurons, proteins, virus, species distribution, bird flocks

but...

which distribution?

Maximum Entropy (MaxEnt) Inference:

Search for the largest entropy distribution satisfying a set of constraints.

Example: pairwise Ising Model

Given binary units data-set of B configurations of N units:

$$\left\{ \{\sigma_i(b)\}_{i=1}^{N} \right\}_{b=1}^{B}$$

Find the MaxEnt model reproducing single and pairwise correlations:

$$\langle \sigma_i \rangle_{\text{MODEL}} = \langle \sigma_i \rangle_{\text{DATA}} \equiv \frac{1}{B} \sum_b \sigma_i(b)$$

$$\langle \sigma_i \sigma_j \rangle_{\text{MODEL}} = \langle \sigma_i \sigma_j \rangle_{\text{DATA}} \equiv \frac{1}{B} \sum_b \sigma_i(b) \sigma_j(b)$$

Example: pairwise Ising Model

Given binary units data-set of B configurations of N units:

$$\left\{\{\sigma_i(b)\}_{i=1}^N\right\}_{b=1}^B$$

Find the MaxEnt model reproducing single and pairwise correlations:

$$\langle\sigma_i\rangle_{\text{MODEL}} = \langle\sigma_i\rangle_{\text{DATA}} \equiv \frac{1}{B}\sum_b \sigma_i(b)$$

$$\langle\sigma_i\sigma_j\rangle_{\text{MODEL}} = \langle\sigma_i\sigma_j\rangle_{\text{DATA}} \equiv \frac{1}{B}\sum_b \sigma_i(b)\sigma_j(b)$$

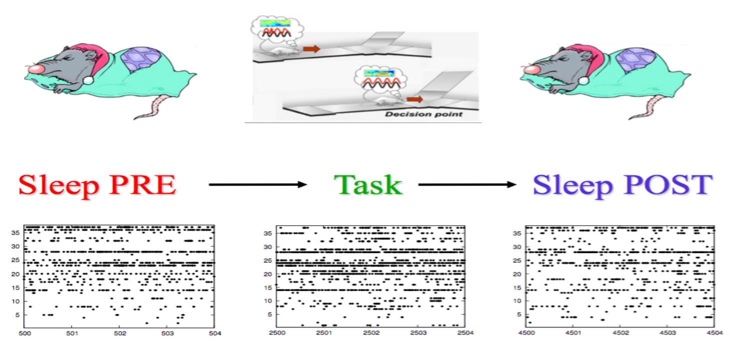
Finely tune the parameters $\{h, J\}$ of the pairwise Ising model:

$$P_{h,J}(\sigma) = \exp\left\{\sum_i h_i\sigma_i + \sum_{ij} J_{ij}\sigma_i\sigma_j\right\}/Z[h, J]$$

In vivo Pre-Frontal Cortex Recording:

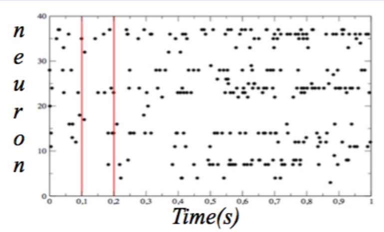
In vivo Pre-Frontal Cortex Recording:

97 experimental sessions of:



Peyrache et al. Nat. Neurosci. (2009)

Ising Model Inference

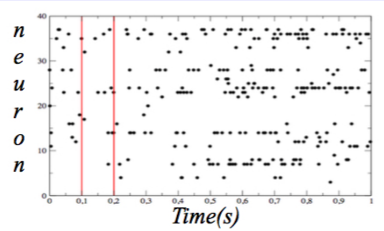


$\sigma_i(b) = 1$ if neuron i spiked during time-bin b

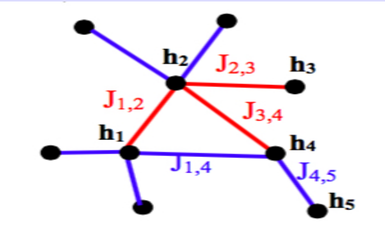
Ask to reproduce neurons firing rates and correlations.

Schneidman et al. Nature 2006; *Cocco, Monasson*, PRL (2011)

Ising Model Inference



$\Rightarrow$ $\Rightarrow$ $\Rightarrow$

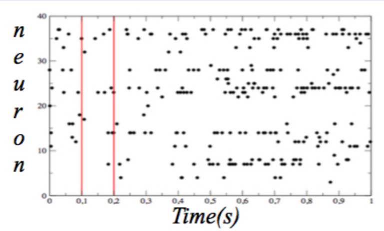


$\sigma_i(b) = 1$ if neuron i spiked during time-bin b

Ask to reproduce neurons firing rates and correlations.

Schneidman et al. Nature 2006; *Cocco, Monasson*,PRL (2011)

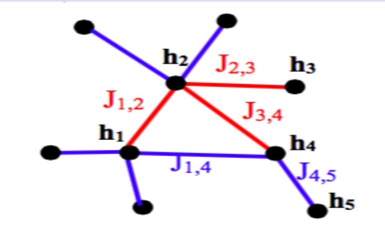
Ising Model Inference



$\Rightarrow$

$\Rightarrow$

$\Rightarrow$



$\sigma_i(b) = 1$ if neuron i spiked during time-bin b

Ask to reproduce neurons firing rates and correlations.

97×3 couplings network sets ($97 \times$ {PRE, TASK , POST})

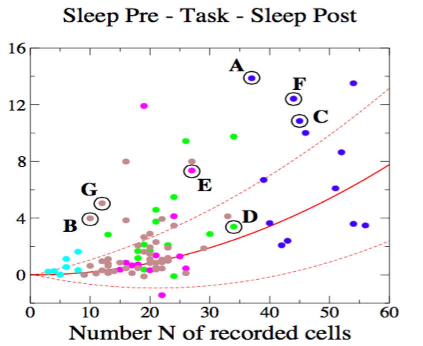
Schneidman et al. Nature 2006; *Cocco, Monasson*,PRL (2011)

Learning related coupling Adjustement

$$A = \sum_{i,j:J^{\text{TASK}},J^{\text{POST}}\neq 0} \text{sign}\left(J_{ij}^{\text{TASK}} - J_{ij}^{\text{PRE}}\right) \cdot \left(J_{ij}^{\text{POST}} - J_{ij}^{\text{PRE}}\right)$$

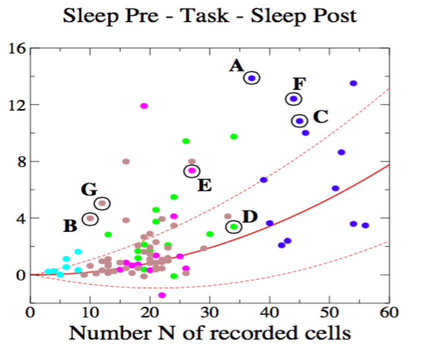
Learning related coupling Adjustement

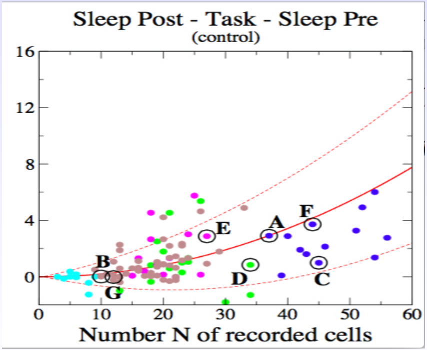
$$A = \sum_{i,j:J^{\text{TASK}},J^{\text{POST}}\neq 0} \text{sign}\left(J_{ij}^{\text{TASK}} - J_{ij}^{\text{PRE}}\right)\cdot\left(J_{ij}^{\text{POST}} - J_{ij}^{\text{PRE}}\right)$$



Learning related coupling Adjustement

$$A = \sum_{i,j:J^{\text{TASK}},J^{\text{POST}}\neq 0} \text{sign}\left(J_{ij}^{\text{TASK}} - J_{ij}^{\text{PRE}}\right) \cdot \left(J_{ij}^{\text{POST}} - J_{ij}^{\text{PRE}}\right)$$





General MaxEnt

Given a list of D observables to reproduce $\{\Sigma_a(\sigma)\}_{a=1}^{D}$

(generic functions of the system units)

Find the MaxEnt model parameters $\{X_a\}_{a=1}^{D}$

$$P_{\mathbf{X}}(\sigma) = \exp\left\{\sum_a X_a \Sigma_a(\sigma)\right\}/Z[X]$$

reproducing the observables averages:

$$\langle \Sigma_a \rangle_{\text{DATA}} \equiv P_a = Q_a[\mathbf{X}] \equiv \langle \Sigma_a \rangle_{\mathbf{X}}$$

Equivalent to log-likelihood maximization:

$$\mathbf{X}^* = \arg\max_{\mathbf{X}} \left[\log L[\,\mathbf{X}\,] \right] \equiv \arg\max_{\mathbf{X}} \left[\mathbf{X} \cdot \mathbf{P} - \log Z[\mathbf{X}] \right]$$

Equivalent to log-likelihood maximization:

$$\mathbf{X}^* = \arg\max_{\mathbf{X}} \left[\log \mathrm{L}[\,\mathbf{X}\,] \right] \equiv \arg\max_{\mathbf{X}} \left[\mathbf{X} \cdot \mathbf{P} - \log Z[\mathbf{X}] \right]$$

in fact:

$$\nabla_a \log \mathrm{L}[\,\mathbf{X}\,] = \frac{d}{dX_a} \left[\mathbf{X} \cdot \mathbf{P} - \log Z[\mathbf{X}] \right] = P_a - Q_a[\mathbf{X}]$$

Equivalent to log-likelihood maximization:

$$\mathbf{X}^* = \arg\max_{\mathbf{X}} \left[\text{logL}[\ \mathbf{X}\ ] \right] \equiv \arg\max_{\mathbf{X}} \left[\mathbf{X} \cdot \mathbf{P} - \log Z[\mathbf{X}] \right]$$

in fact:

$$\nabla_a \text{logL}[\ \mathbf{X}\ ] = \frac{d}{dX_a} \left[\mathbf{X} \cdot \mathbf{P} - \log Z[\mathbf{X}] \right] = P_a - Q_a[\mathbf{X}]$$

Cannot be solved analytically. Ackley, Hinton and Sejnowski (Vanilla Gradient):

$$\mathbf{X}_{t+1} = \mathbf{X}_t + \delta \mathbf{X}_t^{\text{VG}}; \qquad \delta \mathbf{X}_t^{\text{VG}} = \alpha(\mathbf{P} - \mathbf{Q}[\mathbf{X_t}])$$

Equivalent to log-likelihood maximization:

$$\mathbf{X}^* = \arg\max_{\mathbf{X}}\left[\mathrm{logL}[\ \mathbf{X}\ ]\right] \equiv \arg\max_{\mathbf{X}}\left[\mathbf{X}\cdot\mathbf{P} - \log Z[\mathbf{X}]\right]$$

in fact:

$$\nabla_a \mathrm{logL}[\ \mathbf{X}\ ] = \frac{d}{dX_a}\left[\mathbf{X}\cdot\mathbf{P} - \log Z[\mathbf{X}]\right] = P_a - Q_a[\mathbf{X}]$$

Cannot be solved analytically. Ackley, Hinton and Sejnowski (Vanilla Gradient):

$$\mathbf{X}_{t+1} = \mathbf{X}_t + \delta\mathbf{X}_t^{\mathrm{VG}}; \qquad \delta\mathbf{X}_t^{\mathrm{VG}} = \alpha(\mathbf{P} - \mathbf{Q}[\mathbf{X_t}])$$

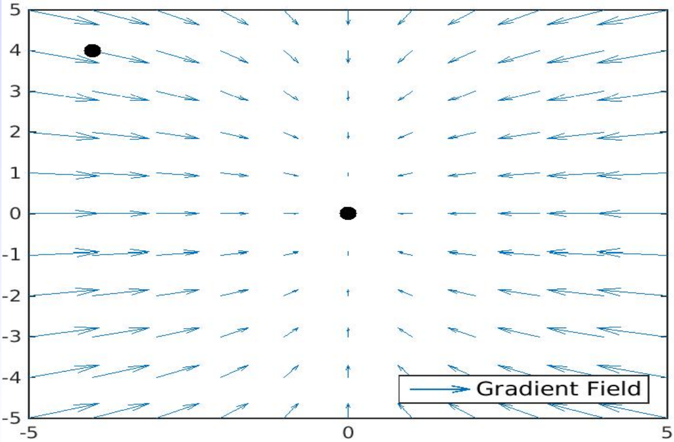
If $0 < P_a < 1$ for all $a = 1, \ldots D$, the problem is well posed:

$\mathbf{X}^*$ exists and is unique and the dynamics converges

(for infinitesimally small α)

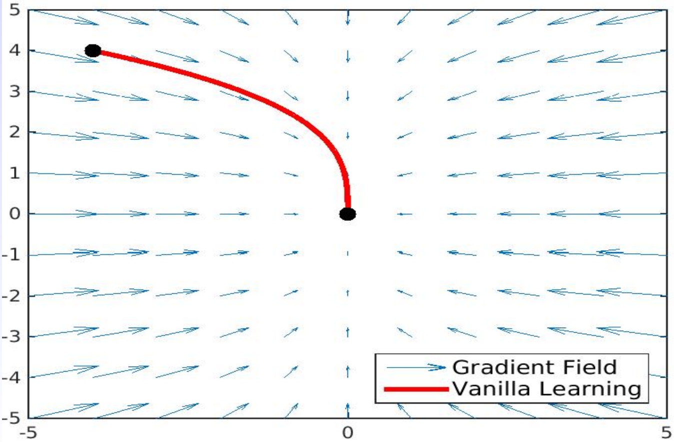
A 2-dimensional example:

$$\log L[u,v] = -\frac{a}{2}(u-u_\infty)^2 - \frac{b}{2}(v-v_\infty)^2$$



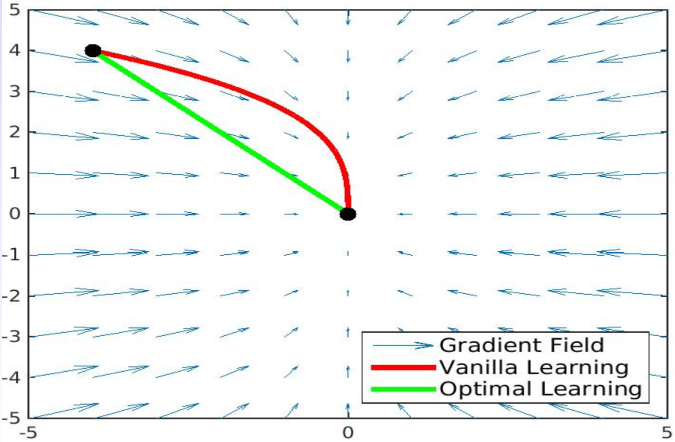
A 2-dimensional example:

$$\log L[u,v] = -\frac{a}{2}(u-u_\infty)^2 - \frac{b}{2}(v-v_\infty)^2$$



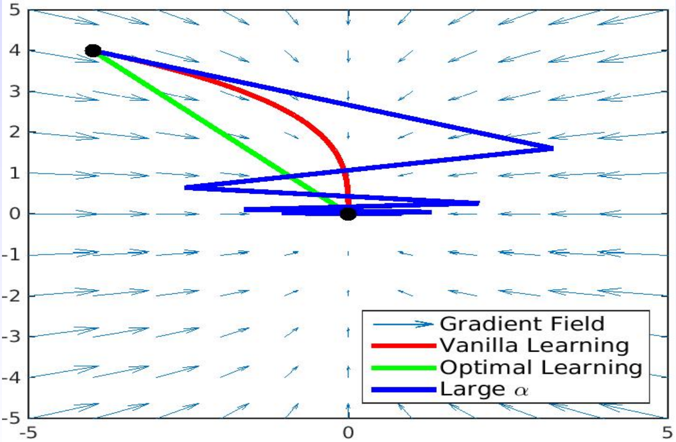
A 2-dimensional example:

$$\log L[u, v] = -\frac{a}{2}(u - u_\infty)^2 - \frac{b}{2}(v - v_\infty)^2$$



A 2-dimensional example:

$$\log L[u, v] = -\frac{a}{2}(u - u_\infty)^2 - \frac{b}{2}(v - v_\infty)^2$$



A 2-dimensional example:

$$\mathrm{logL}[u,v] = -\frac{a}{2}(u-u_\infty)^2 - \frac{b}{2}(v-v_\infty)^2$$

Vanilla Gradient:

$$\delta u_t^{\mathrm{VG}} \sim (1-\alpha\, a)^{-t} \Rightarrow \alpha < 2/a; \qquad \delta v_t^{\mathrm{VG}} \sim (1-\alpha\, b)^{-t} \Rightarrow \alpha < 2/b$$

A 2-dimensional example:

$$\mathrm{logL}[u,v] = -\frac{a}{2}(u-u_\infty)^2 - \frac{b}{2}(v-v_\infty)^2$$

Vanilla Gradient:

$$\delta u_t^{\mathrm{VG}} \sim (1-\alpha\, a)^{-t} \Rightarrow \alpha < 2/a; \qquad \delta v_t^{\mathrm{VG}} \sim (1-\alpha\, b)^{-t} \Rightarrow \alpha < 2/b$$

Newton Method:

$$\delta u_t^{\mathrm{VG}} \sim (1-\alpha)^{-t} \Rightarrow \alpha < 2; \qquad \delta v_t^{\mathrm{VG}} \sim (1-\alpha)^{-t} \Rightarrow \alpha < 2$$

A 2-dimensional example:

$$\text{logL}[u,v] = -\frac{a}{2}(u-u_\infty)^2 - \frac{b}{2}(v-v_\infty)^2$$

Vanilla Gradient:

$$\delta u_t^{\text{VG}} \sim (1-\alpha\, a)^{-t} \Rightarrow \alpha < 2/a; \qquad \delta v_t^{\text{VG}} \sim (1-\alpha\, b)^{-t} \Rightarrow \alpha < 2/b$$

Newton Method:

$$\delta u_t^{\text{VG}} \sim (1-\alpha)^{-t} \Rightarrow \alpha < 2; \qquad \delta v_t^{\text{VG}} \sim (1-\alpha)^{-t} \Rightarrow \alpha < 2$$

$\alpha = 1 \quad \Rightarrow \quad$ convergence in one step!

The same happens for the MaxEnt inference:

$$\mathrm{logL}[\mathbf{X} \approx \mathbf{X}^*] \approx \mathrm{logL}[\mathbf{X}^*] - \frac{1}{2} \sum_{ab} (X_a - X_a^*)\, \chi[\mathbf{X}^*]_{ab}\, (X_b - X_b^*)$$

$$\chi_{ab}[\mathbf{X}] \equiv -\frac{\partial^2 \mathrm{logL}[\mathbf{X}]}{\partial X_a \partial X_b} = \langle \Sigma_a \Sigma_b \rangle_{\mathbf{X}} - \langle \Sigma_a \rangle_{\mathbf{X}} \langle \Sigma_b \rangle_{\mathbf{X}}$$

The same happens for the MaxEnt inference:

$$\mathrm{logL}[\mathbf{X} \approx \mathbf{X}^*] \approx \mathrm{logL}[\mathbf{X}^*] - \frac{1}{2}\sum_{ab}(X_a - X_a^*)\,\chi[\mathbf{X}^*]_{ab}\,(X_b - X_b^*)$$

$$\chi_{ab}[\mathbf{X}] \equiv -\frac{\partial^2 \mathrm{logL}[\mathbf{X}]}{\partial X_a \partial X_b} = \langle \Sigma_a \Sigma_b \rangle_{\mathbf{X}} - \langle \Sigma_a \rangle_{\mathbf{X}} \langle \Sigma_b \rangle_{\mathbf{X}}$$

Vanilla Gradient: $\delta X_t^{\mathrm{VG}} = \alpha\, \nabla \mathrm{logL}[\mathbf{X}_{t-1}]$

$$\delta X_t^{\mu} \equiv \sum_a V_a^{\mu} \delta X_{a,t} \sim (1 - \alpha\, \lambda_{\mu})^{-t}$$

The same happens for the MaxEnt inference:

$$\mathrm{logL}[\mathbf{X} \approx \mathbf{X}^*] \approx \mathrm{logL}[\mathbf{X}^*] - \frac{1}{2} \sum_{ab} (X_a - X_a^*)\, \chi[\mathbf{X}^*]_{ab}\, (X_b - X_b^*)$$

$$\chi_{ab}[\mathbf{X}] \equiv -\frac{\partial^2 \mathrm{logL}[\mathbf{X}]}{\partial X_a \partial X_b} = \langle \Sigma_a \Sigma_b \rangle_{\mathbf{X}} - \langle \Sigma_a \rangle_{\mathbf{X}} \langle \Sigma_b \rangle_{\mathbf{X}}$$

Vanilla Gradient: $\delta X_t^{\mathrm{VG}} = \alpha\, \nabla \mathrm{logL}[\mathbf{X}_{t-1}]$

$$\delta X_t^{\mu} \equiv \sum_a V_a^{\mu} \delta X_{a,t} \sim (1 - \alpha\, \lambda_{\mu})^{-t}$$

Newton Method[1]: $\delta X_t^{\mathrm{NM}} = \alpha\, \chi^{-1}[\mathbf{X}_{t-1}]\, \nabla \mathrm{logL}[\mathbf{X}_{t-1}]$

$$\delta X_t^{\mu} \equiv \sum_a V_a^{\mu} \delta X_{a,t} \sim (1 - \alpha)^{-t}$$

[1] (here equivalent to Amari98 Natural Gradient)

The same happens for the MaxEnt inference:

$$\mathrm{logL}[\mathbf{X} \approx \mathbf{X}^*] \approx \mathrm{logL}[\mathbf{X}^*] - \frac{1}{2}\sum_{ab}(X_a - X_a^*)\,\chi[\mathbf{X}^*]_{ab}\,(X_b - X_b^*)$$

$$\chi_{ab}[\mathbf{X}] \equiv -\frac{\partial^2 \mathrm{logL}[\mathbf{X}]}{\partial X_a \partial X_b} = \langle \Sigma_a \Sigma_b \rangle_{\mathbf{X}} - \langle \Sigma_a \rangle_{\mathbf{X}} \langle \Sigma_b \rangle_{\mathbf{X}}$$

Vanilla Gradient: $\delta X_t^{\mathrm{VG}} = \alpha\, \nabla \mathrm{logL}[\mathbf{X}_{t-1}]$

$$\delta X_t^{\mu} \equiv \sum_a V_a^{\mu} \delta X_{a,t} \sim (1 - \alpha\, \lambda_{\mu})^{-t}$$

Newton Method[1]: $\delta X_t^{\mathrm{NM}} = \alpha\, \chi^{-1}[\mathbf{X}_{t-1}]\, \nabla \mathrm{logL}[\mathbf{X}_{t-1}]$

$$\delta X_t^{\mu} \equiv \sum_a V_a^{\mu} \delta X_{a,t} \sim (1 - \alpha)^{-t}$$

VERY SLOW: expensive estimation & inversion of $\chi[\mathbf{X}]$

[1] (here equivalent to Amari98 Natural Gradient)

However, for the Ising model we can approximate:

$$\chi_{ab}[\mathbf{X}^*] \approx \overline{\chi}_{ab} \equiv \langle \Sigma_a \Sigma_b \rangle_{\mathrm{DATA}} - \langle \Sigma_a \rangle_{\mathrm{DATA}} \langle \Sigma_b \rangle_{\mathrm{DATA}}$$

However, for the Ising model we can approximate:

$$\chi_{ab}[\mathbf{X}^{*}] \approx \overline{\chi}_{ab} \equiv \langle \Sigma_a \Sigma_b \rangle_{\text{DATA}} - \langle \Sigma_a \rangle_{\text{DATA}} \langle \Sigma_b \rangle_{\text{DATA}}$$

Approximated Newton (AN) Method:

$$\delta X_t^{\text{AN}} = \alpha\, \overline{\chi}^{-1}\, \nabla \text{logL}[\mathbf{X}_{t-1}]$$

However, for the Ising model we can approximate:

$$\chi_{ab}[\mathbf{X}^*] \approx \overline{\chi}_{ab} \equiv \langle \Sigma_a \Sigma_b \rangle_{\mathrm{DATA}} - \langle \Sigma_a \rangle_{\mathrm{DATA}} \langle \Sigma_b \rangle_{\mathrm{DATA}}$$

Approximated Newton (AN) Method:

$$\delta X_t^{\mathrm{AN}} = \alpha \, \overline{\chi}^{-1} \, \nabla \mathrm{logL}[\mathbf{X}_{t-1}]$$

Remarks on $\chi[\mathbf{X}^*] \approx \overline{\chi}$:

- equivalent to say that an Ising distribution properly describes data.
- states that the model Fisher is close to the observables co-variance.

As the algorithm works iteratively, it requires an

early-stop condition

As the algorithm works iteratively, it requires an

early-stop condition

idea: stop the algorithm when

$\mathbf{Q}[\mathbf{X}]$ is statistically compatible with $\mathbf{P}$

using the $\mathbf{P}$-covariance $\overline{\chi}/B$

As the algorithm works iteratively, it requires an

early-stop condition

idea: stop the algorithm when

$\mathbf{Q}[\mathbf{X}]$ is statistically compatible with $\mathbf{P}$

using the $\mathbf{P}$-covariance $\overline{\chi}/B$

$$\epsilon\left(\mathbf{P}\, ,\, \mathbf{Q}[\mathbf{X}] \right) \equiv \frac{B}{2D} \sum_{ab} (P_a - Q_a) \left(\overline{\chi}^{-1} \right)_{ab} (P_b - Q_b)$$

quantifies the distance between $\mathbf{Q}[\mathbf{X}]$ and $\mathbf{P}$ in the $\overline{\chi}/B$ metric.

As the algorithm works iteratively, it requires an

early-stop condition

idea: stop the algorithm when

$\mathbf{Q}[\mathbf{X}]$ is statistically compatible with $\mathbf{P}$

using the $\mathbf{P}$-covariance $\overline{\chi}/B$

$$\epsilon\left(\mathbf{P}\,,\ \mathbf{Q}[\mathbf{X}]\right) \equiv \frac{B}{2D}\sum_{ab}(P_a - Q_a)\left(\overline{\chi}^{-1}\right)_{ab}(P_b - Q_b)$$

quantifies the distance between $\mathbf{Q}[\mathbf{X}]$ and $\mathbf{P}$ in the $\overline{\chi}/B$ metric.

For two *i.i.d* data-sets: $\epsilon\left(\mathbf{P}\,,\ \mathbf{P}'\right) \approx 1$

$\Rightarrow$ we stop the algorithm as soon as $\epsilon < 1$

APPROXIMATED NEWTON ALGORITHM:

1. Initialization:
   (a) Chose $\mathbf{X}_0$ and compute $\mathbf{Q}[\mathbf{X_0}]$ and $\epsilon_0 = \epsilon\Big(\mathbf{P} \,,\, \mathbf{Q}[\mathbf{X_0}] \Big)$
   (b) Then set $\alpha_0 = 1$ and $M = \min(\frac{2B}{\epsilon_0}, B)$ MCMC samplings

APPROXIMATED NEWTON ALGORITHM:

1. Initialization:
   (a) Chose $\mathbf{X}_0$ and compute $\mathbf{Q}[\mathbf{X_0}]$ and $\epsilon_0 = \epsilon\Big(\ \mathbf{P}\ ,\ \mathbf{Q}[\mathbf{X_0}]\ \Big)$
   (b) Then set $\alpha_0 = 1$ and $M = \min(\frac{2B}{\epsilon_0}, B)$ MCMC samplings
2. Iterate the following step:
   (a) update the $\mathbf{X_t}$
   (b) estimate $\mathbf{Q}[\mathbf{X_t}]$ with $M = \min(\frac{2B}{\epsilon_{t-1}}, B)$ MCMC samplings
   (c) compute $\epsilon_t = \epsilon\Big(\ \mathbf{P}\ ,\ \mathbf{Q}[\mathbf{X_t}]\ \Big)$,
      (d1) $\epsilon_t < \epsilon_{t-1}$: accept the update and increase α
      (d2) $\epsilon_t > \epsilon_{t-1}$: discard the update, lower α and re-estimate $\mathbf{Q}[\mathbf{X_t}]$.

APPROXIMATED NEWTON ALGORITHM:

1. Initialization:
   (a) Chose $\mathbf{X}_0$ and compute $\mathbf{Q}[\mathbf{X_0}]$ and $\epsilon_0 = \epsilon\Big(\mathbf{P} \,,\; \mathbf{Q}[\mathbf{X_0}] \Big)$
   (b) Then set $\alpha_0 = 1$ and $M = \min(\frac{2B}{\epsilon_0}, B)$ MCMC samplings
2. Iterate the following step:
   (a) update the $\mathbf{X_t}$
   (b) estimate $\mathbf{Q}[\mathbf{X_t}]$ with $M = \min(\frac{2B}{\epsilon_{t-1}}, B)$ MCMC samplings
   (c) compute $\epsilon_t = \epsilon\Big(\mathbf{P} \,,\; \mathbf{Q}[\mathbf{X_t}] \Big)$,
      (d1) $\epsilon_t < \epsilon_{t-1}$: accept the update and increase α
      (d2) $\epsilon_t > \epsilon_{t-1}$: discard the update, lower α and re-estimate $\mathbf{Q}[\mathbf{X_t}]$.
3. stop the algorithm when $\epsilon_t < 1$.

Rat retina ganglion cells

Two moving bars.

$2.1h$ of MEA recording

$B = 4.8 \cdot 10^5$ of $\Delta t = 16ms$

$N = 95$ cells

$D = 4560$ parameters to infer.

Rat retina ganglion cells

Two moving bars.

$2.1h$ of MEA recording

$B = 4.8 \cdot 10^5$ of $\Delta t = 16ms$

$N = 95$ cells

$D = 4560$ parameters to infer.

Convergence time from independent spins model with 8×3.4Ghz CPUs:

$$T^{\text{AN}} = 144 \pm 4s$$

$$T^{\text{VG}}(\alpha = 0.15) = 4.2 \cdot 10^4 s$$

Rat retina ganglion cells

Two moving bars.

$2.1h$ of MEA recording

$B = 4.8 \cdot 10^5$ of $\Delta t = 16ms$

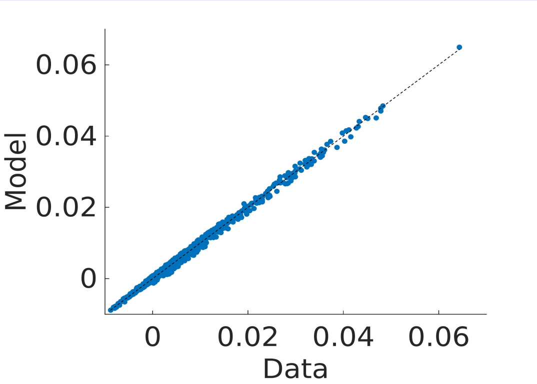
$N = 95$ cells

$D = 4560$ parameters to infer.

Convergence time from independent spins model with 8×3.4Ghz CPUs:

$T^{\text{AN}} = 144 \pm 4s$

$T^{\text{VG}}(\alpha = 0.15) = 4.2 \cdot 10^4 s$



$$c_{ij} = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle$$

Rat retina ganglion cells

Two moving bars.

$2.1h$ of MEA recording

$B = 4.8 \cdot 10^5$ of $\Delta t = 16ms$

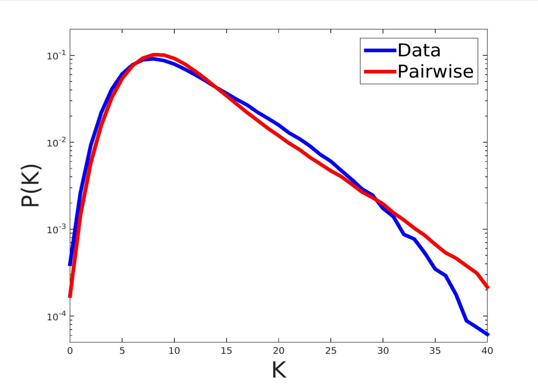
$N = 95$ cells

$D = 4560$ parameters to infer.

Convergence time from independent spins model with 8×3.4Ghz CPUs:

$T^{\text{AN}} = 144 \pm 4s$

$T^{\text{VG}}(\alpha = 0.15) = 4.2 \cdot 10^4 s$



$P(K) = \text{Prob}(\sum_i \sigma_i = K)$

$\mathbf{Q}[\mathbf{X}]$ is estimated through M MCMC measurements.

$\mathbf{Q}[\mathbf{X}] \;\Rightarrow\; \mathbf{Q}[\mathbf{X}]^{\mathrm{MC}}$ is random variable!

$\mathbf{Q}[\mathbf{X}]$ is estimated through M MCMC measurements.

$\mathbf{Q}[\mathbf{X}] \;\Rightarrow\; \mathbf{Q}[\mathbf{X}]^{\mathrm{MC}}$ is random variable!

$\nabla \log \mathrm{L}_{\mathbf{X}}^{\mathrm{MC}} = \mathbf{P} - \mathbf{Q}[\mathbf{X}]^{\mathrm{MC}} \rightarrow 0$ only on average,

$\mathbf{Q}[\mathbf{X}]$ is estimated through M MCMC measurements.

$\mathbf{Q}[\mathbf{X}] \;\Rightarrow\; \mathbf{Q}[\mathbf{X}]^{\mathsf{MC}}$ is random variable!

$\nabla \mathrm{log} \mathrm{L}_{\mathbf{X}}^{\mathsf{MC}} = \mathbf{P} - \mathbf{Q}[\mathbf{X}]^{\mathsf{MC}} \rightarrow 0$ only on average,

Change of Framework:

$$\mathbf{X}_t \rightarrow P_t(\mathbf{X})$$

$\mathbf{X}$, rather than converge to a fixed point,

approaches a stationary $P_\infty(\mathbf{X})$

$\mathbf{Q}[\mathbf{X}]$ is estimated through M MCMC measurements.

$\mathbf{Q}[\mathbf{X}] \;\Rightarrow\; \mathbf{Q}[\mathbf{X}]^{\mathrm{MC}}$ is random variable!

$\nabla \mathrm{logL}_{\mathbf{X}}^{\mathrm{MC}} = \mathbf{P} - \mathbf{Q}[\mathbf{X}]^{\mathrm{MC}} \to 0$ only on average,

<u>Change of Framework:</u>

$$\mathbf{X}_t \to P_t(\mathbf{X})$$

$\mathbf{X}$, rather than converge to a fixed point,

approaches a stationary $P_\infty(\mathbf{X})$

<u>Master Equation:</u>

$$P_{t+1}(\mathbf{X}') = \int D\mathbf{X}\; P_t(\mathbf{X})\; W_{\mathbf{X}\to\mathbf{X}'}(\alpha)$$

For $M \gg 1$ and $\mathbf{X} \approx \mathbf{X}^*$:

$$\mathrm{logL}[\mathbf{X}] \simeq \mathrm{logL}[\mathbf{X}^*] - \frac{1}{2} \sum_{ab} (X_a - X_a^*)\, \chi[\mathbf{X}^*]_{ab}\, (X_b - X_b^*)$$

For $M \gg 1$ and $\mathbf{X} \approx \mathbf{X}^*$:

$$\mathrm{logL}[\mathbf{X}] \simeq \mathrm{logL}[\mathbf{X}^*] - \frac{1}{2}\sum_{ab}(X_a - X_a^*)\, \chi[\mathbf{X}^*]_{ab}\, (X_b - X_b^*)$$

1. $\langle \nabla_a \mathrm{logL}_{\mathbf{X}}^{\mathrm{MC}} \rangle = \sum_b \chi[\mathbf{X}^*]_{ab}\, (X_b^* - X_b) \approx \sum_b \overline{\chi}_{ab}(X_b^* - X_b)$

For $M \gg 1$ and $\mathbf{X} \approx \mathbf{X}^*$:

$$\mathrm{logL}[\mathbf{X}] \simeq \mathrm{logL}[\mathbf{X}^*] - \tfrac{1}{2} \textstyle\sum_{ab}(X_a - X_a^*)\, \chi[\mathbf{X}^*]_{ab}\, (X_b - X_b^*)$$

1. $\langle \nabla_a \mathrm{logL}_{\mathbf{X}}^{\mathrm{MC}} \rangle = \sum_b \chi[\mathbf{X}^*]_{ab}\, (X_b^* - X_b) \approx \sum_b \overline{\chi}_{ab}(X_b^* - X_b)$
2. $\left\langle \nabla_a \mathrm{logL}_{\mathbf{X}}^{\mathrm{MC}}\ \nabla_b \mathrm{logL}_{\mathbf{X}}^{\mathrm{MC}} \right\rangle_c = \chi[\mathbf{X}]_{ab}/M \simeq \chi[\mathbf{X}^*]_{ab}/M \approx \overline{\chi}_{ab}/M$

For $M \gg 1$ and $\mathbf{X} \approx \mathbf{X}^*$:

$$\mathrm{logL}[\mathbf{X}] \simeq \mathrm{logL}[\mathbf{X}^*] - \frac{1}{2} \sum_{ab} (X_a - X_a^*)\, \chi[\mathbf{X}^*]_{ab}\, (X_b - X_b^*)$$

1. $\langle \nabla_a \mathrm{logL}_{\mathbf{X}}^{\mathrm{MC}} \rangle = \sum_b \chi[\mathbf{X}^*]_{ab}\, (X_b^* - X_b) \approx \sum_b \overline{\chi}_{ab} (X_b^* - X_b)$
2. $\left\langle \nabla_a \mathrm{logL}_{\mathbf{X}}^{\mathrm{MC}}\ \nabla_b \mathrm{logL}_{\mathbf{X}}^{\mathrm{MC}} \right\rangle_c = \chi[\mathbf{X}]_{ab}/M \simeq \chi[\mathbf{X}^*]_{ab}/M \approx \overline{\chi}_{ab}/M$

a normal approximation gives:

$$P(\nabla \mathrm{logL}_{\mathbf{X}}^{\mathrm{MC}}) \simeq \mathcal{N}\Big[\ \overline{\chi} \cdot (\mathbf{X}^* - \mathbf{X})\ ;\ \overline{\chi}/M\ \Big](\nabla \mathrm{logL}_{\mathbf{X}}^{\mathrm{MC}})$$

For $M \gg 1$ and $\mathbf{X} \approx \mathbf{X}^*$:

$$\text{logL}[\mathbf{X}] \simeq \text{logL}[\mathbf{X}^*] - \frac{1}{2} \sum_{ab} (X_a - X_a^*)\ \chi[\mathbf{X}^*]_{ab}\ (X_b - X_b^*)$$

1. $\langle \nabla_a \text{logL}_{\mathbf{X}}^{\text{MC}} \rangle = \sum_b \chi[\mathbf{X}^*]_{ab}\ (X_b^* - X_b) \approx \sum_b \overline{\chi}_{ab} (X_b^* - X_b)$
2. $\left\langle\ \nabla_a \text{logL}_{\mathbf{X}}^{\text{MC}}\ \nabla_b \text{logL}_{\mathbf{X}}^{\text{MC}}\ \right\rangle_c = \chi[\mathbf{X}]_{ab}/M \simeq \chi[\mathbf{X}^*]_{ab}/M \approx \overline{\chi}_{ab}/M$

a normal approximation gives:

$$P(\nabla \text{logL}_{\mathbf{X}}^{\text{MC}}) \simeq \mathcal{N}\Big[\ \overline{\chi} \cdot (\mathbf{X}^* - \mathbf{X})\ ;\ \overline{\chi}/M\ \Big](\nabla \text{logL}_{\mathbf{X}}^{\text{MC}})$$

- $W^{\text{VG}}_{\mathbf{X}\to\mathbf{X}'}(\alpha) = \text{Prob}\left(\nabla \text{logL}_{\mathbf{X}}^{\text{MC}} = \frac{\mathbf{X}' - \mathbf{X}}{\alpha}\right)$
- $W^{\text{AN}}_{\mathbf{X}\to\mathbf{X}'}(\alpha) = \text{Prob}\left(\nabla \text{logL}_{\mathbf{X}}^{\text{MC}} = \overline{\chi} \cdot \frac{\mathbf{X}' - \mathbf{X}}{\alpha}\right)$

Imposing $P_{t+1}(\mathbf{X}) = P_t(\mathbf{X})$

- $P_\infty^{\mathrm{VG}}(\mathbf{X}) = \mathcal{N}\left[\mathbf{X}^*; \frac{\alpha}{M}\left(2\delta - \alpha\overline{\chi}\right)^{-1}\right](\mathbf{X})$
- $P_\infty^{\mathrm{AN}}(\mathbf{X}) = \mathcal{N}\left[\mathbf{X}^*; \frac{\alpha}{M(2-\alpha)}\overline{\chi}^{-1}\right](\mathbf{X})$

Imposing $P_{t+1}(\mathbf{X}) = P_t(\mathbf{X})$

- $P_\infty^{\mathrm{VG}}(\mathbf{X}) = \mathcal{N}\left[\mathbf{X}^*; \frac{\alpha}{M}\left(2\delta - \alpha\overline{\chi}\right)^{-1}\right](\mathbf{X}), \qquad \alpha\lambda_\mu < 2$
- $P_\infty^{\mathrm{AN}}(\mathbf{X}) = \mathcal{N}\left[\mathbf{X}^*; \frac{\alpha}{M(2-\alpha)}\overline{\chi}^{-1}\right](\mathbf{X}), \qquad \alpha < 2$

Imposing $P_{t+1}(\mathbf{X}) = P_t(\mathbf{X})$

- $P_\infty^{\mathrm{VG}}(\mathbf{X}) = \mathcal{N}\left[\mathbf{X}^*; \frac{\alpha}{M}\left(2\delta - \alpha\overline{\chi}\right)^{-1}\right](\mathbf{X}),$ $\qquad \alpha\lambda_\mu < 2$
- $P_\infty^{\mathrm{AN}}(\mathbf{X}) = \mathcal{N}\left[\mathbf{X}^*; \frac{\alpha}{M(2-\alpha)}\overline{\chi}^{-1}\right](\mathbf{X}),$ $\qquad \alpha < 2$

Which self-consistently defines $\mathbf{X} \approx \mathbf{X}^*$

Imposing $P_{t+1}(\mathbf{X}) = P_t(\mathbf{X})$

- $P_\infty^{\mathrm{VG}}(\mathbf{X}) = \mathcal{N}\left[\mathbf{X}^*; \frac{\alpha}{M}\left(2\delta - \alpha\overline{\chi}\right)^{-1}\right](\mathbf{X}), \qquad \alpha\lambda_\mu < 2$
- $P_\infty^{\mathrm{AN}}(\mathbf{X}) = \mathcal{N}\left[\mathbf{X}^*; \frac{\alpha}{M(2-\alpha)}\overline{\chi}^{-1}\right](\mathbf{X}), \qquad \alpha < 2$

Which self-consistently defines $\mathbf{X} \approx \mathbf{X}^*$

From $P(\nabla \mathrm{logL}_{\mathbf{X}}^{\mathrm{MC}}) = P(\mathbf{P} - \mathbf{Q}[\mathbf{X}]^{\mathrm{MC}})$

- $P_\infty^{\mathrm{VG}}(\mathbf{Q}^{\mathrm{MC}}) = \mathcal{N}\left[\mathbf{P}; \frac{2}{M}\overline{\chi}\left(2\delta - \alpha\overline{\chi}\right)^{-1}\right](\mathbf{Q}^{\mathrm{MC}})$
- $P_\infty^{\mathrm{AN}}(\mathbf{Q}^{\mathrm{MC}}) = \mathcal{N}\left[\mathbf{P}; \frac{2}{M(2-\alpha)}\overline{\chi}\right](\mathbf{Q}^{\mathrm{MC}})$

Imposing $P_{t+1}(\mathbf{X}) = P_t(\mathbf{X})$

- $P^{\mathrm{VG}}_{\infty}(\mathbf{X}) = \mathcal{N}\left[\mathbf{X}^*; \frac{\alpha}{M}\left(2\delta - \alpha\overline{\chi}\right)^{-1}\right](\mathbf{X}), \qquad \alpha\lambda_\mu < 2$
- $P^{\mathrm{AN}}_{\infty}(\mathbf{X}) = \mathcal{N}\left[\mathbf{X}^*; \frac{\alpha}{M(2-\alpha)}\overline{\chi}^{-1}\right](\mathbf{X}), \qquad \alpha < 2$

Which self-consistently defines $\mathbf{X} \approx \mathbf{X}^*$

From $P(\nabla \mathrm{logL}^{\mathrm{MC}}_{\mathbf{X}}) = P(\mathbf{P} - \mathbf{Q}[\mathbf{X}]^{\mathrm{MC}})$

- $P^{\mathrm{VG}}_{\infty}(\mathbf{Q}^{\mathrm{MC}}) = \mathcal{N}\left[\mathbf{P}; \frac{2}{M}\overline{\chi}\left(2\delta - \alpha\overline{\chi}\right)^{-1}\right](\mathbf{Q}^{\mathrm{MC}})$
- $P^{\mathrm{AN}}_{\infty}(\mathbf{Q}^{\mathrm{MC}}) = \mathcal{N}\left[\mathbf{P}; \frac{2}{M(2-\alpha)}\overline{\chi}\right](\mathbf{Q}^{\mathrm{MC}})$

Which is better? How to set the parameters?

Algorithm Vs Empirical distributions

An experiment provides

empirical estimates of $\mathbf{Q}^{\mathbf{EMP}}$:

$$P^{\text{EMP}}(\mathbf{Q}^{\mathbf{EMP}}) \simeq \mathcal{N}\left[\mathbf{P}^{\text{TRUE}}, \chi^{\text{EMP}}\right]$$

- $\mathbf{P}^{\text{TRUE}}$: result from infinitely long experiment
- χ^{EMP} expected co-variance for B measurements

Algorithm Vs Empirical distributions

An experiment provides

empirical estimates of $\mathbf{Q}^{\mathbf{EMP}}$:

$$P^{\text{EMP}}(\mathbf{Q}^{\mathbf{EMP}}) \simeq \mathcal{N}\left[\mathbf{P}^{\text{TRUE}}, \chi^{\text{EMP}}\right]$$

An inference algorithm provides

numerical estimates of $\mathbf{Q}^{\mathbf{MC}}$:

$$P^{\text{ALG}}_{\mathbf{P}}(\mathbf{Q}^{\mathbf{MC}}) \simeq \mathcal{N}\left[\mathbf{P}, \chi^{\text{ALG}}\right]$$

- $\mathbf{P}^{\text{TRUE}}$: result from infinitely long experiment
- χ^{EMP} expected co-variance for B measurements
- $\mathbf{P}$ one-shot sampling of P^{EMP}

Algorithm Vs Empirical distributions

An experiment provides empirical estimates of $\mathbf{Q}^{\mathbf{EMP}}$:

$$P^{\mathsf{EMP}}(\mathbf{Q}^{\mathbf{EMP}}) \simeq \mathcal{N}\left[\mathbf{P}^{\mathsf{TRUE}}, \chi^{\mathsf{EMP}}\right]$$

An inference algorithm provides numerical estimates of $\mathbf{Q}^{\mathbf{MC}}$:

$$P^{\mathsf{ALG}}_{\mathbf{P}}(\mathbf{Q}^{\mathbf{MC}}) \simeq \mathcal{N}\left[\mathbf{P}, \chi^{\mathsf{ALG}}\right]$$

- $\mathbf{P}^{\mathsf{TRUE}}$: result from infinitely long experiment
- χ^{EMP} expected co-variance for B measurements
- $\mathbf{P}$ one-shot sampling of P^{EMP}

An optimal inference algorithm should provide:

P^{ALG} **as close as possible** to P^{EMP}.

What is the optimal χ^{ALG} value?

Kullback-Leibler distance between P^{EMP} and $P^{\text{ALG}}_{\mathbf{P}}$:

$$D_{KL}\left(P^{\text{EMP}}(\cdot)||P^{\text{ALG}}_{\mathbf{P}}(\cdot)\right)$$

Kullback-Leibler distance between P^{EMP} and $P^{\mathsf{ALG}}_{\mathbf{P}}$:

$$D_{KL}\left(P^{\mathsf{EMP}}(\cdot)||P^{\mathsf{ALG}}_{\mathbf{P}}(\cdot)\right)$$

$$\chi^{\mathsf{OPT}} = \arg\min_{\chi^{\mathsf{ALG}}} \int \mathbf{DP}\; P^{\mathsf{EMP}}(\mathbf{P})\; D_{KL}\left(P^{\mathsf{EMP}}(\cdot)||P^{\mathsf{ALG}}_{\mathbf{P}}(\cdot)\right)$$

Kullback-Leibler distance between P^{EMP} and $P^{\text{ALG}}_{\mathbf{P}}$:

$$D_{KL}\left(P^{\text{EMP}}(\cdot)||P^{\text{ALG}}_{\mathbf{P}}(\cdot)\right)$$

$$\chi^{\text{OPT}} = \arg\min_{\chi^{\text{ALG}}} \int \mathbf{DP}\; P^{\text{EMP}}(\mathbf{P})\; D_{KL}\left(P^{\text{EMP}}(\cdot)||P^{\text{ALG}}_{\mathbf{P}}(\cdot)\right)$$

The <u>solution</u> and its <u>approximation</u> are:

$$\chi^{\text{OPT}} = 2\chi^{\text{EMP}} \approx 2\overline{\chi}/B$$

Kullback-Leibler distance between P^{EMP} and $P^{\mathrm{ALG}}_{\mathbf{P}}$:

$$D_{KL}\left(P^{\mathrm{EMP}}(\cdot)||P^{\mathrm{ALG}}_{\mathbf{P}}(\cdot)\right)$$

$$\chi^{\mathrm{OPT}} = \arg\min_{\chi^{\mathrm{ALG}}} \int \mathbf{DP}\ P^{\mathrm{EMP}}(\mathbf{P})\ D_{KL}\left(P^{\mathrm{EMP}}(\cdot)||P^{\mathrm{ALG}}_{\mathbf{P}}(\cdot)\right)$$

The <u>solution</u> and its <u>approximation</u> are:

$$\chi^{\mathrm{OPT}} = 2\chi^{\mathrm{EMP}} \approx 2\overline{\chi}/B$$

to compare with:

$$\chi^{\mathrm{VG}} = \tfrac{2}{M}\overline{\chi}\,(2\delta - \alpha\overline{\chi})^{-1}, \qquad \chi^{\mathrm{AN}} = \tfrac{2}{M(2-\alpha)}\overline{\chi}$$

Kullback-Leibler distance between P^{EMP} and $P^{\text{ALG}}_{\mathbf{P}}$:

$$D_{KL}\left(P^{\text{EMP}}(\cdot)||P^{\text{ALG}}_{\mathbf{P}}(\cdot)\right)$$

$$\chi^{\text{OPT}} = \arg\min_{\chi^{\text{ALG}}} \int \mathbf{DP}\; P^{\text{EMP}}(\mathbf{P})\; D_{KL}\left(P^{\text{EMP}}(\cdot)||P^{\text{ALG}}_{\mathbf{P}}(\cdot)\right)$$

The solution and its approximation are:

$$\chi^{\text{OPT}} = 2\chi^{\text{EMP}} \approx 2\overline{\chi}/B$$

to compare with:

$$\chi^{\text{VG}} = \frac{2}{M}\overline{\chi}\,(2\delta - \alpha\overline{\chi})^{-1}, \qquad \chi^{\text{AN}} = \frac{2}{M(2-\alpha)}\overline{\chi}$$

AN with $M(2-\alpha) = B$ reaches the optimum!

VG underfits $\lambda_\mu \gg (2 - B/M)/\alpha$ and overfits $\lambda_\mu \ll (2 - B/M)/\alpha$

Synthetic data: Theory Vs Simulations

Bethe Lattice Ising Model

$$N = 10,\ c = 4$$

$$J_{ij} = \pm 0.53,$$
$$h_i = -0.14 - 2\sum_j J_{ij}$$

100 independent estimations

of $\mathbf{P}$ and $\overline{\chi}$

through 2^{16} sampling of P^{EMP}

Inference with $M = B$

Synthetic data: Theory Vs Simulations

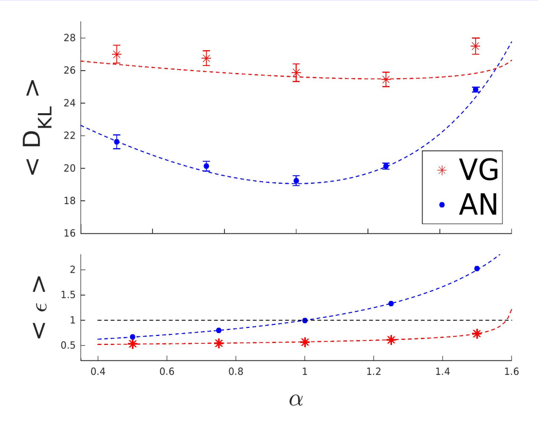
Bethe Lattice Ising Model

$$N = 10,\ c = 4$$

$$J_{ij} = \pm 0.53,$$
$$h_i = -0.14 - 2\sum_j J_{ij}$$

100 independent estimations of $\mathbf{P}$ and $\overline{\chi}$ through 2^{16} sampling of P^{EMP}

Inference with $M = B$



Conclusions:

- MaxEnt models are useful to describe multi-units systems
- The AN learning is faster than the VG algorithm.
- Within the large B approximation is possible to completely characterize the long time behavior
- The AN with $\alpha = 1$ and $M = B$ is optimal against overfitting.

Perspectives:

- Improve the gaussian approximations
- Test the algorithm to non-pairwise models
- Generalize the class of model distributions beyond MaxEnt
- Include hidden variables and the RBM framework

THANKS

Collaborators for P.F. Cortex work:

- Francesco Battaglia
- Simona Cocco
- Remi Monasson
- Gaia Tavoni

Founding

- EU-FP7 FET OPEN project Enlightenment 284801
- Human Brain Project (HBP CLAP)

arXiv:1507.04254