Phase transitions in Generalized Linear Models

Cargèse summer school

August 22, 2018

Léo Miolane

with Jean Barbier, Florent Krzakala, Nicolas Macris & Lenka Zdeborová

Generalized Linear Models

Definition

Statistical model

- One observes for $1 \leq \mu \leq m$

$$Y_\mu \sim P_{\mathrm{out}}\Big(\cdot \ \Big| \ \langle \mathbf{\Phi}_\mu, \mathbf{X}^* \rangle \Big)$$

- $\mathbf{X}^* \in \mathbb{R}^n$: signal vector of dimension n.
- $\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_m \in \mathbb{R}^n$: measurement vectors.
- P_{out}: transition kernel.

Goal: recover $\mathbf{X}^*$ from $\mathbf{Y}$ (and $\mathbf{\Phi}$).

Generalized Linear Models

Definition

Statistical model

- One observes for $1 \leq \mu \leq m$

$$Y_\mu \sim P_{\text{out}}\Big(\cdot \,\Big|\, \langle \mathbf{\Phi}_\mu, \mathbf{X}^* \rangle \Big)$$

- $\mathbf{X}^* \in \mathbb{R}^n$: signal vector of dimension n.
- $\mathbf{\Phi}_1, \dots, \mathbf{\Phi}_m \in \mathbb{R}^n$: measurement vectors.
- P_{out}: transition kernel.

Goal: recover $\mathbf{X}^*$ from $\mathbf{Y}$ (and $\mathbf{\Phi}$).

- When is it information-theoretically possible?
- When is it computationally tractable?

Examples

Some interesting particular cases

- Linear model:

$$\mathbf{Y} = \mathbf{\Phi} \mathbf{X}^*$$

- Phase retrieval:

$$\mathbf{Y} = \left| \mathbf{\Phi} \mathbf{X}^* \right|$$

- 1-bit CS (“Planted” perceptron):

$$\mathbf{Y} = \mathrm{sign}\left(\mathbf{\Phi} \mathbf{X}^* \right)$$

Examples

Some interesting particular cases

- Linear model:

$$\mathbf{Y} = \mathbf{\Phi}\mathbf{X}^* + \text{Noise}$$

- Phase retrieval:

$$\mathbf{Y} = \left|\mathbf{\Phi}\mathbf{X}^*\right| + \text{Noise}$$

- 1-bit CS (“Planted” perceptron):

$$\mathbf{Y} = \mathrm{sign}\left(\mathbf{\Phi}\mathbf{X}^* + \text{Noise}\right)$$

Examples

Some interesting particular cases

- Linear model:

$$\mathbf{Y} = \mathbf{\Phi}\mathbf{X}^* + \text{Noise}$$

- Phase retrieval:

$$\mathbf{Y} = \left|\mathbf{\Phi}\mathbf{X}^*\right| + \text{Noise}$$

- 1-bit CS ("Planted" perceptron):

$$\mathbf{Y} = \text{sign}\big(\mathbf{\Phi}\mathbf{X}^* + \text{Noise}\big)$$

- Logistic model:

$$Y_\mu = \begin{cases} +1 & \text{with probability} \quad \dfrac{1}{1+\exp(-\lambda\langle\mathbf{\Phi}_\mu, \mathbf{X}^*\rangle)} \\[2ex] -1 & \text{with probability} \quad \dfrac{1}{1+\exp(\lambda\langle\mathbf{\Phi}_\mu, \mathbf{X}^*\rangle)} \end{cases}$$

Setting

Assumptions

$$\mathbf{Y} \sim P_{\mathrm{out}}\Big(\cdot \,\Big|\, \mathbf{\Phi X}^* \Big)$$

- Asymptotic regime: $n \to \infty, \quad m/n \to \alpha > 0.$
- $\mathbf{X}^* = (X_1^*, \ldots, X_n^*) \overset{\text{i.i.d.}}{\sim} P_0, \qquad \mathbb{E}_{P_0} X^2 = \rho.$

Setting

Assumptions

$$\mathbf{Y} \sim P_{\text{out}}\Big(\cdot \;\Big|\; \mathbf{\Phi X}^* \Big)$$

- Asymptotic regime: $n \to \infty, \quad m/n \to \alpha > 0$.
- $\mathbf{X}^* = (X_1^*, \ldots, X_n^*) \overset{\text{i.i.d.}}{\sim} P_0, \qquad \mathbb{E}_{P_0} X^2 = \rho$.
- $\big(\Phi_{i,j}\big)$ are independent, $\begin{cases} \mathbb{E}\Phi_{i,j} = 0 \\ \mathbb{E}\Phi_{i,j}^2 = 1/n \\ \sup_{i,j} \mathbb{E}|\Phi_{i,j}|^3 \text{ remains bounded.} \end{cases}$
- $\mathbb{E}\big[|Y_\mu|^{2+\epsilon}\big]$ remains bounded, for some $\epsilon > 0$.

Setting

Assumptions

$$\mathbf{Y} \sim P_{\mathrm{out}}\Big(\cdot \;\Big|\; \mathbf{\Phi}\mathbf{X}^*\Big)$$

- Asymptotic regime: $n \to \infty, \quad m/n \to \alpha > 0$.
- $\mathbf{X}^* = (X_1^*, \dots, X_n^*) \overset{\text{i.i.d.}}{\sim} P_0, \qquad \mathbb{E}_{P_0} X^2 = \rho$.
- $\big(\Phi_{i,j}\big)$ are independent, $\begin{cases} \mathbb{E}\Phi_{i,j} = 0 \\ \mathbb{E}\Phi_{i,j}^2 = 1/n \\ \sup_{i,j} \mathbb{E}|\Phi_{i,j}|^3 \text{ remains bounded.} \end{cases}$
- $\mathbb{E}\big[|Y_\mu|^{2+\epsilon}\big]$ remains bounded, for some $\epsilon > 0$.
- $x \in \mathbb{R} \mapsto P_{\mathrm{out}}(\cdot \,|\, x)$ is continuous almost everywhere.

Setting

Assumptions

$$\mathbf{Y} \sim P_{\text{out}}\Big(\cdot \,\Big|\, \mathbf{\Phi X}^* \Big)$$

- Asymptotic regime: $n \to \infty, \quad m/n \to \alpha > 0$.
- $\mathbf{X}^* = (X_1^*, \dots, X_n^*) \overset{\text{i.i.d.}}{\sim} P_0, \qquad \mathbb{E}_{P_0} X^2 = \rho$.
- $\big(\Phi_{i,j}\big)$ are independent, $\begin{cases} \mathbb{E}\Phi_{i,j} = 0 \\ \mathbb{E}\Phi_{i,j}^2 = 1/n \\ \sup_{i,j} \mathbb{E}|\Phi_{i,j}|^3 \text{ remains bounded.} \end{cases}$
- $\mathbb{E}\big[|Y_\mu|^{2+\epsilon}\big]$ remains bounded, for some $\epsilon > 0$.
- $x \in \mathbb{R} \mapsto P_{\text{out}}(\cdot \,|\, x)$ is continuous almost everywhere.
- P_{out} has to be "regularized" by some (small) Gaussian noise: $\forall x \in \mathbb{R}$, "$P_{\text{out}}(\cdot \,|\, x) = \widetilde{P}_{\text{out}}(\cdot \,|\, x) + \mathcal{N}(0, \sigma^2)$", where $\sigma > 0$.
- If P_{out} takes values in $\mathbb{N}$, no need for regularization ($\sigma = 0$).

Setting

Assumptions

$$\mathbf{Y} \sim P_{\mathrm{out}}\Big(\cdot \, \Big| \, \mathbf{\Phi X}^* \Big)$$

- Asymptotic regime: $n \to \infty, \quad m/n \to \alpha > 0$.
- $\mathbf{X}^* = (X_1^*, \dots, X_n^*) \overset{\text{i.i.d.}}{\sim} P_0, \qquad \mathbb{E}_{P_0} X^2 = \rho$.
- $\big(\Phi_{i,j}\big)$ are independent, $\begin{cases} \mathbb{E}\Phi_{i,j} = 0 \\ \mathbb{E}\Phi_{i,j}^2 = 1/n \\ \sup_{i,j} \mathbb{E}|\Phi_{i,j}|^3 \text{ remains bounded.} \end{cases}$
- $\mathbb{E}\Big[|Y_\mu|^{2+\epsilon}\Big]$ remains bounded, for some $\epsilon > 0$.
- $x \in \mathbb{R} \mapsto P_{\mathrm{out}}(\cdot \,|\, x)$ is continuous almost everywhere.
- P_{out} has to be "regularized" by some (small) Gaussian noise:
  $\forall x \in \mathbb{R}$, "$P_{\mathrm{out}}(\cdot \,|\, x) = \widetilde{P}_{\mathrm{out}}(\cdot \,|\, x) + \mathcal{N}(0, \sigma^2)$", where $\sigma > 0$.
- If P_{out} takes values in $\mathbb{N}$, no need for regularization ($\sigma = 0$).

The statistician knows the model, i.e. P_0 and P_{out}.

Information-theoretic study

The mutual information

Posterior distribution $P(\mathbf{X}^*|\mathbf{Y}, \mathbf{\Phi})$:

$$P(\mathbf{x}|\mathbf{Y}, \mathbf{\Phi}) = \frac{1}{\mathcal{Z}_n} P_0^{\otimes n}(\mathbf{x}) \prod_{\mu=1}^{m} P_{\mathrm{out}}\big(Y_\mu \,\big|\langle \mathbf{\Phi}_\mu, \mathbf{x}\rangle\big)$$

where $\mathcal{Z}_n$ is the appropriate normalization.

Information-theoretic study

The mutual information

Posterior distribution $P(\mathbf{X}^*|\mathbf{Y}, \mathbf{\Phi})$:

$$P(\mathbf{x}|\mathbf{Y}, \mathbf{\Phi}) = \frac{1}{\mathcal{Z}_n} P_0^{\otimes n}(\mathbf{x}) \prod_{\mu=1}^{m} P_{\text{out}}\big(Y_\mu \,\big|\langle \mathbf{\Phi}_\mu, \mathbf{x}\rangle\big)$$

where $\mathcal{Z}_n$ is the appropriate normalization. The free energy is

$$f_n = -\frac{1}{n}\mathbb{E}\log \mathcal{Z}_n = -\frac{1}{n}\mathbb{E}\left[\log \int_{\mathbf{x}\in\mathbb{R}^n} dP_0^{\otimes n}(\mathbf{x}) \prod_{\mu=1}^{m} P_{\text{out}}\big(Y_\mu \,\big|\langle \mathbf{\Phi}_\mu, \mathbf{x}\rangle\big)\right]$$

Information-theoretic study

The mutual information

Posterior distribution $P(\mathbf{X}^*|\mathbf{Y},\mathbf{\Phi})$:

$$P(\mathbf{x}|\mathbf{Y},\mathbf{\Phi}) = \frac{1}{\mathcal{Z}_n} P_0^{\otimes n}(\mathbf{x}) \prod_{\mu=1}^{m} P_{\text{out}}\big(Y_\mu \,\big|\langle \mathbf{\Phi}_\mu, \mathbf{x}\rangle\big)$$

where $\mathcal{Z}_n$ is the appropriate normalization. The free energy is

$$f_n = -\frac{1}{n}\mathbb{E}\log \mathcal{Z}_n = -\frac{1}{n}\mathbb{E}\left[\log \int_{\mathbf{x}\in\mathbb{R}^n} dP_0^{\otimes n}(\mathbf{x}) \prod_{\mu=1}^{m} P_{\text{out}}\big(Y_\mu \,\big|\langle \mathbf{\Phi}_\mu, \mathbf{x}\rangle\big)\right]$$

Equivalently, we are going to study the mutual information:

$$\frac{1}{n} I(\mathbf{X}^*;\mathbf{Y}|\mathbf{\Phi}) = f_n + \text{Constant} + o_n(1)\,.$$

Limiting expression for the mutual information

“Replica Symmetric” formula

Theorem

$$\frac{1}{n} I\Big(\mathbf{X}^*;\mathbf{Y}\Big|\mathbf{\Phi}\Big) \xrightarrow[n\to\infty]{} \inf_{q\in[0,\rho]} \sup_{r\geq 0} \Big\{ I_{P_0}(r) + \alpha\,\mathcal{I}_{P_{\text{out}}}(q) - \frac{r}{2}(\rho - q) \Big\}$$

Limiting expression for the mutual information

"Replica Symmetric" formula

Theorem

$$\frac{1}{n} I\Big(\mathbf{X}^*; \mathbf{Y} \Big| \mathbf{\Phi}\Big) \xrightarrow[n\to\infty]{} \inf_{q\in[0,\rho]} \sup_{r\geq 0} \Big\{ I_{P_0}(r) + \alpha\, \mathcal{I}_{P_{\mathrm{out}}}(q) - \frac{r}{2}(\rho - q) \Big\}$$

Example: Linear regression

- $\mathbf{Y} = \mathbf{\Phi}\mathbf{X}^* + \sigma \mathbf{Z}$.
- "Tanaka formula", proved by Barbier et al., 2016 and Reeves and Pfister, 2016.

Limiting expression for the mutual information

"Replica Symmetric" formula

Theorem

$$\frac{1}{n} I\Big(\mathbf{X}^*; \mathbf{Y} \Big| \mathbf{\Phi}\Big) \xrightarrow[n\to\infty]{} \inf_{q\in[0,\rho]} \sup_{r\geq 0} \Big\{ I_{P_0}(r) + \alpha\, \mathcal{I}_{P_{\mathrm{out}}}(q) - \frac{r}{2}(\rho - q) \Big\}$$

Example: 'planted' perceptron.

- $\mathbf{Y} = \mathrm{sign}(\mathbf{\Phi}\mathbf{X}^*)$, where $X_1^*, \dots, X_n^* \overset{\text{i.i.d.}}{\sim} \mathcal{U}(+1,-1)$.
- $S_n = \big\{\mathbf{x} \,\big|\, \forall \mu,\, \mathrm{sign}(\mathbf{\Phi}_\mu \mathbf{x}) = Y_\mu \big\}$
- $\frac{1}{n} I(\mathbf{X}^*;\mathbf{Y}|\mathbf{\Phi}) = \log 2 - \frac{1}{n}\mathbb{E}\big[\log \# S_n\big]$.
- Formula obtained by Gardner and Derrida, 1989.

Limiting expression for the mutual information

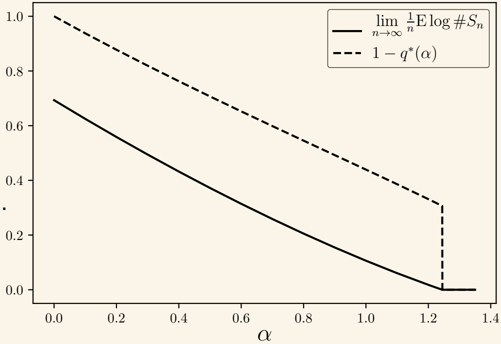
“Replica Symmetric” formula

Theorem

$$\frac{1}{n} I\Big(\mathbf{X}^*; \mathbf{Y} \Big| \mathbf{\Phi}\Big) \xrightarrow[n\to\infty]{} \inf_{q\in[0,\rho]} \sup_{r\geq 0} \left\{ I_{P_0}(r) + \alpha\, \mathcal{I}_{P_{\mathrm{out}}}(q) - \frac{r}{2}(\rho - q) \right\}$$

Example: ‘planted’ perceptron.

- $\mathbf{Y} = \mathrm{sign}(\mathbf{\Phi}\mathbf{X}^*)$, where $X_1^*, \dots, X_n^* \overset{\text{i.i.d.}}{\sim} \mathcal{U}(+1,-1)$.
- $S_n = \left\{\mathbf{x} \,\middle|\, \forall\mu,\ \mathrm{sign}(\mathbf{\Phi}_\mu \mathbf{x}) = Y_\mu \right\}$
- $\frac{1}{n} I(\mathbf{X}^*;\mathbf{Y}|\mathbf{\Phi}) = \log 2 - \frac{1}{n}\mathbb{E}\big[\log \# S_n\big]$.
- Formula obtained by Gardner and Derrida, 1989.



Two scalar inference channels

Explanation of the formula

Recall: $\mathbf{Y} \sim P_{\mathrm{out}}\big(\,\cdot\,\big|\,\mathbf{\Phi}\mathbf{X}^*\big)$

$$\frac{1}{n} I(\mathbf{X}^*; \mathbf{Y}|\mathbf{\Phi}) \xrightarrow[n\to\infty]{} \inf_{q\in[0,\rho]} \sup_{r\geq 0} \Big\{ I_{P_0}(r) + \alpha\, \mathcal{I}_{P_{\mathrm{out}}}(q) - \frac{r}{2}(\rho - q) \Big\}$$

Additive Gaussian channel

$$I_{P_0}(r) = I(X^*; \sqrt{r}X^* + Z)$$

where $X^* \sim P_0$ and $Z \sim \mathcal{N}(0,1)$.

Two scalar inference channels

Explanation of the formula

Recall: $\mathbf{Y} \sim P_{\mathrm{out}}\big(\,\cdot\,\big|\,\mathbf{\Phi}\mathbf{X}^*\big)$

$$\frac{1}{n} I(\mathbf{X}^*; \mathbf{Y}|\mathbf{\Phi}) \xrightarrow[n\to\infty]{} \inf_{q\in[0,\rho]} \sup_{r\geq 0} \Big\{ I_{P_0}(r) + \alpha\, \mathcal{I}_{P_{\mathrm{out}}}(q) - \frac{r}{2}(\rho - q) \Big\}$$

Additive Gaussian channel

$$I_{P_0}(r) = I(X^*; \sqrt{r} X^* + Z)$$

where $X^* \sim P_0$ and $Z \sim \mathcal{N}(0,1)$.

Non-linear Gaussian retrieval

$$\mathcal{I}_{P_{\mathrm{out}}}(q) = I(W^*; Y^{(q)}|V)$$

where $V, W^* \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and

$$Y^{(q)} \sim P_{\mathrm{out}}\big(\,\cdot\,\big|\,\sqrt{q}V + \sqrt{\rho - q}W^*\big)$$

Proof technique

The interpolation method

In the spirit of Talagrand's interpolation scheme for the perceptron.

	$t=0$	$0<t<1$	$t=1$
Observations	$\mathbf{Y} \sim P_{\rm out}(\cdot \vert \mathbf{\Phi}\mathbf{X}^*)$	$\begin{cases} \mathbf{Y} \sim P_{\rm out}(\,\cdot\,\vert \mathbf{S}_t) \\ \mathbf{Y}' = \sqrt{rt}\,\mathbf{X}^* + \mathbf{Z} \end{cases}$ $\mathbf{S}_t = \sqrt{1-t}\,\mathbf{\Phi}\mathbf{X}^* + \sqrt{t}\big(\sqrt{q}\,\mathbf{V} + \sqrt{\rho-q}\,\mathbf{W}^*\big)$	$\begin{cases} \mathbf{Y} \sim P_{\rm out}(\cdot \vert \sqrt{q}\,\mathbf{V} + \sqrt{\rho-q}\,\mathbf{W}^*) \\ \mathbf{Y}' = \sqrt{r}\,\mathbf{X}^* + \mathbf{Z} \end{cases}$
Mutual information	$\frac{1}{n} I\big(\mathbf{X}^*; \mathbf{Y} \big\vert \mathbf{\Phi}\big)$	$f_n(t)$	$I_{P_0}(r) + \frac{m}{n}\mathcal{I}_{P_{\rm out}}(q)$

Proof technique

The interpolation method

In the spirit of Talagrand's interpolation scheme for the perceptron.

	$t=0$	$0<t<1$	$t=1$
Observations	$\mathbf{Y} \sim P_{\mathrm{out}}(\cdot \vert \mathbf{\Phi X}^*)$	$\begin{cases} \mathbf{Y} \sim P_{\mathrm{out}}(\,\cdot\,\vert \mathbf{S}_t) \\ \mathbf{Y}' = \sqrt{rt}\mathbf{X}^* + \mathbf{Z} \end{cases}$ $\mathbf{S}_t = \sqrt{1-t}\mathbf{\Phi X}^* + \sqrt{t}\big(\sqrt{q}\mathbf{V} + \sqrt{\rho-q}\mathbf{W}^*\big)$	$\begin{cases} \mathbf{Y} \sim P_{\mathrm{out}}(\cdot \vert \sqrt{q}\mathbf{V} + \sqrt{\rho-q}\mathbf{W}^*) \\ \mathbf{Y}' = \sqrt{r}\mathbf{X}^* + \mathbf{Z} \end{cases}$
Mutual information	$\frac{1}{n} I\big(\mathbf{X}^*; \mathbf{Y} \big\vert \mathbf{\Phi}\big)$	$f_n(t)$	$I_{P_0}(r) + \frac{m}{n}\mathcal{I}_{P_{\mathrm{out}}}(q)$

Goal: show that $f_n'(t) \simeq \frac{r}{2}(\rho - q)$.

Interpolation method

Derivative of the interpolating mutual information

- $$f_n'(t) = \mathbb{E}\left[\Big(\frac{1}{n}\sum_{i=1}^{n} x_i^{(t)} X_i^* - q\Big)\Big(\text{Bounded term}\Big)\right] + \frac{r}{2}(\rho - q)$$

  where $\mathbf{x}^{(t)} \sim P\big(\mathbf{X}^* \big| \text{Observations at time } t\big)$.

Interpolation method

Derivative of the interpolating mutual information

- $f_n'(t) = \mathbb{E}\left[\Big(\frac{1}{n}\sum_{i=1}^{n} x_i^{(t)} X_i^* - q\Big)\Big(\text{Bounded term}\Big)\right] + \frac{r}{2}(\rho - q)$

  where $\mathbf{x}^{(t)} \sim P\big(\mathbf{X}^* \big| \text{Observations at time } t\big)$.

- We have to show that the overlap

$$\frac{1}{n}\sum_{i=1}^{n} x_i^{(t)} X_i^*$$

  concentrates around some value, and then choose q to be equal to this value.

Interpolation method

Derivative of the interpolating mutual information

- $f_n'(t) = \mathbb{E}\left[\Big(\frac{1}{n}\sum_{i=1}^{n} x_i^{(t)} X_i^* - q\Big)\Big(\text{Bounded term}\Big)\right] + \frac{r}{2}(\rho - q)$

  where $\mathbf{x}^{(t)} \sim P\big(\mathbf{X}^* \big| \text{Observations at time } t\big)$.

- We have to show that the overlap

$$\frac{1}{n}\sum_{i=1}^{n} x_i^{(t)} X_i^*$$

  concentrates around some value, and then choose q to be equal to this value.

- In the case of Bayes-optimal inference problems this is true under mild assumptions: Montanari, 2008, Korada and Macris, 2010.

- More details about the techniques in Jean Barbier's talk on Saturday.

Limit of the overlap

Minimal Mean Squared Error

Theorem

For almost all $\alpha > 0$, the infimum of the "Mutual Information formula" admits a unique minimizer $q_*(\alpha)$ and

$$\left|\frac{1}{n}\sum_{i=1}^{n} x_i X_i^*\right| \xrightarrow[n\to\infty]{} q_*(\alpha), \quad \text{in probability,}$$

where $\mathbf{x} \sim P(\mathbf{X}^* = \cdot \,|\, \mathbf{\Phi}, \mathbf{Y})$ independently of everything else.

One deduces:

$$\mathrm{MMSE}_n(\alpha) := \frac{1}{n^2}\mathbb{E}\left\|\mathbf{X}^*\mathbf{X}^{*\intercal} - \mathbb{E}\left[\mathbf{X}^*\mathbf{X}^{*\intercal}|\mathbf{\Phi}, \mathbf{Y}\right]\right\|^2 \xrightarrow[n\to\infty]{} \rho^2 - q_*(\alpha)^2$$

Algorithmic analysis

Generalized Approximate Message Passing (GAMP)

- Precursors in physics: Mezard, 1989, Kabashima, 2008.
- Generalization of AMP (Donoho et al., 2009) introduced by Rangan, 2011. Iterative algorithm: produces estimates $\widehat{\mathbf{x}}^0, \dots, \widehat{\mathbf{x}}^t$.

Algorithmic analysis

Generalized Approximate Message Passing (GAMP)

- Precursors in physics: Mezard, 1989, Kabashima, 2008.
- Generalization of AMP (Donoho et al., 2009) introduced by Rangan, 2011. Iterative algorithm: produces estimates $\widehat{\mathbf{x}}^0, \dots, \widehat{\mathbf{x}}^t$.
- Its performance can be rigorously tracked:

State evolution, Javanmard and Montanari, 2013

$$\frac{1}{n^2}\mathbb{E}\left\|\mathbf{X}^*\mathbf{X}^{*\intercal} - \widehat{\mathbf{x}}^t\widehat{\mathbf{x}}^{t\intercal}\right\|^2 \xrightarrow[n\to\infty]{} \rho^2 - (q^t)^2$$

where q^t is given by the recursion $(q^0 = 0)$:

$$\begin{cases} q^{t+1} &= \rho - 2I'_{P_0}(r^t) \\ r^t &= -2\alpha\,\mathcal{I}'_{P_{\mathrm{out}}}(q^t) \end{cases}$$

- GAMP converges to a stationary point $(q^{\mathrm{alg}}, r^{\mathrm{alg}})$ of the MI formula and if $q^{\mathrm{alg}} = q_*(\alpha)$, then GAMP achieves the MMSE!

Algorithmic analysis

Generalized Approximate Message Passing (GAMP)

- Precursors in physics: Mezard, 1989, Kabashima, 2008.
- Generalization of AMP (Donoho et al., 2009) introduced by Rangan, 2011. Iterative algorithm: produces estimates $\hat{\mathbf{x}}^0, \ldots, \hat{\mathbf{x}}^t$.
- Its performance can be rigorously tracked:

State evolution, Javanmard and Montanari, 2013

$$\frac{1}{n^2}\mathbb{E}\left\|\mathbf{X}^*\mathbf{X}^{*\intercal} - \hat{\mathbf{x}}^t\hat{\mathbf{x}}^{t\intercal}\right\|^2 \xrightarrow[n\to\infty]{} \rho^2 - (q^t)^2$$

where q^t is given by the recursion $(q^0 = 0)$:

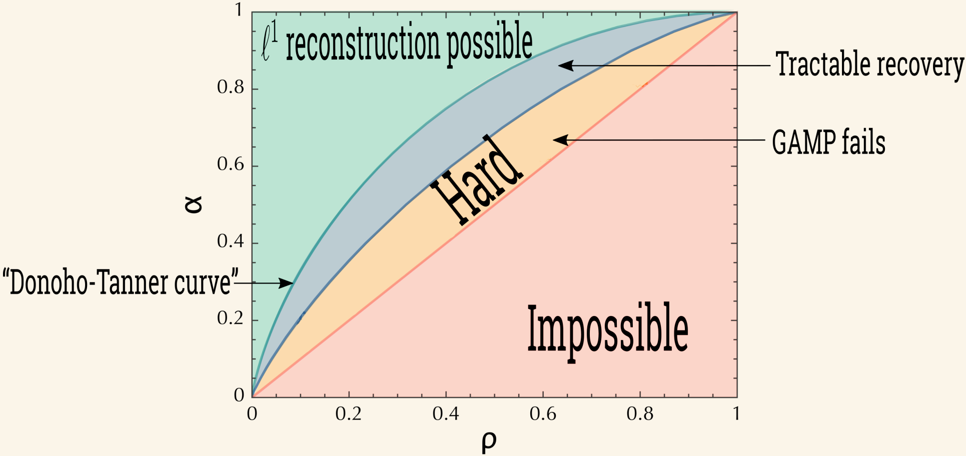
$$\begin{cases} q^{t+1} & = \ \rho - 2I'_{P_0}(r^t) \\ r^t & = \ -2\alpha\,\mathcal{I}'_{P_{\mathrm{out}}}(q^t) \end{cases}$$

- GAMP converges to a stationary point $(q^{\mathrm{alg}}, r^{\mathrm{alg}})$ of the MI formula and if $q^{\mathrm{alg}} = q_*(\alpha)$, then GAMP achieves the MMSE!
- **Main belief**: GAMP is optimal among all polynomial-time algorithms.

Phase diagrams: warm-up

Linear model

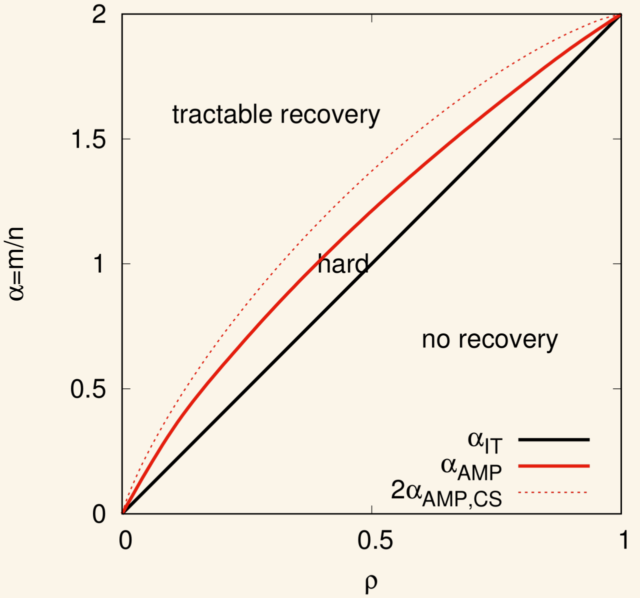
$$\mathbf{Y} = \mathbf{\Phi}\mathbf{X}^*, \qquad P_0 = \rho\mathcal{N}(0,1) + (1-\rho)\delta_0$$



Phase diagram from Krzakala et al., 2012

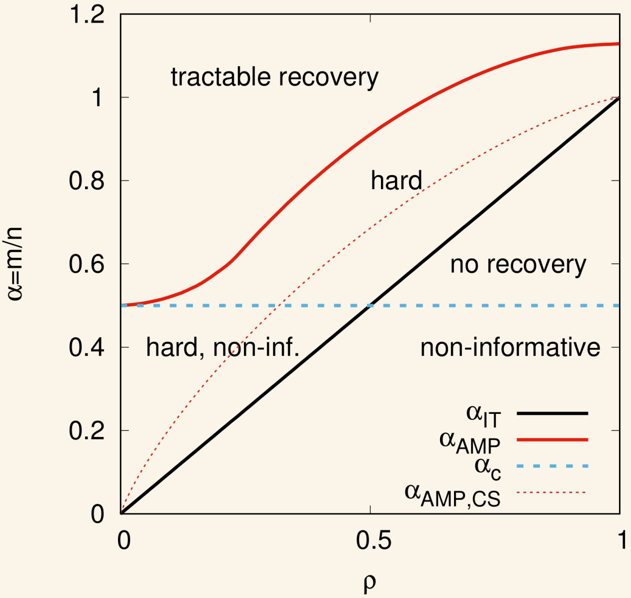
ReLu channel

$$\mathbf{Y} = \text{ReLu}(\mathbf{\Phi X}^*), \qquad \text{ReLu}(x) = x\mathbf{1}(x \geq 0), \qquad P_0 = \rho\mathcal{N}(0,1) + (1-\rho)\delta_0$$



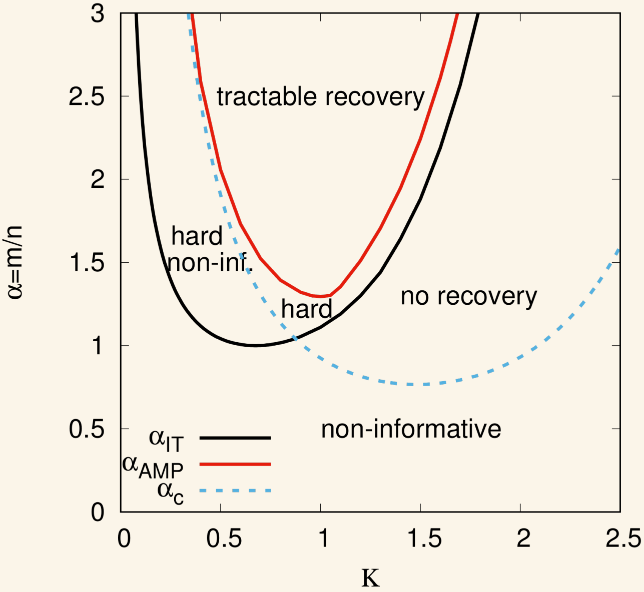
Absolute value channel

$$\mathbf{Y} = |\mathbf{\Phi}\mathbf{X}^*|, \qquad P_0 = \rho\mathcal{N}(0,1) + (1-\rho)\delta_0$$



Symmetric door channel

$$\mathbf{Y} = \mathbf{1}\big(\mathbf{\Phi X}^* \in [-K, K]\big), \qquad P_0 = \tfrac{1}{2}\delta_{-1} + \tfrac{1}{2}\delta_{+1}$$



A learning problem

A different point of view

- The points $\left\{(\mathbf{\Phi}_1, Y_1), \ldots, (\mathbf{\Phi}_m, Y_m)\right\}$ can be seen as data generated by some relation $\mathbf{Y} \sim P_{\text{out}}(\cdot|\mathbf{\Phi}\mathbf{X}^*)$.
- **Question**: How difficult is it to learn this relation?

A learning problem

A different point of view

- The points $\left\{(\mathbf{\Phi}_1, Y_1), \ldots, (\mathbf{\Phi}_m, Y_m)\right\}$ can be seen as data generated by some relation $\mathbf{Y} \sim P_{\text{out}}(\cdot|\mathbf{\Phi}\mathbf{X}^*)$.
- **Question**: How difficult is it to learn this relation?
- What is the optimal generalization error

$$\mathcal{E}_n^{\text{gen}} = \min_{\widehat{\theta}} \mathbb{E}\left[\left(Y^{(\text{new})} - \widehat{\theta}(\mathbf{\Phi}^{(\text{new})}; \mathbf{Y}, \mathbf{\Phi})\right)^2\right]$$

where $Y^{(\text{new})} \sim P_{\text{out}}\left(\cdot \,\middle|\, \langle \mathbf{\Phi}^{(\text{new})}, \mathbf{X}^* \rangle\right)$ is a new sample.

A learning problem

A different point of view

- The points $\left\{(\mathbf{\Phi}_1, Y_1), \ldots, (\mathbf{\Phi}_m, Y_m)\right\}$ can be seen as data generated by some relation $\mathbf{Y} \sim P_{\text{out}}(\cdot|\mathbf{\Phi}\mathbf{X}^*)$.
- **Question**: How difficult is it to learn this relation?
- What is the optimal generalization error

$$\mathcal{E}_n^{\text{gen}} = \min_{\widehat{\theta}} \mathbb{E}\left[\left(Y^{(\text{new})} - \widehat{\theta}(\mathbf{\Phi}^{(\text{new})}; \mathbf{Y}, \mathbf{\Phi})\right)^2\right]$$

where $Y^{(\text{new})} \sim P_{\text{out}}\left(\cdot \ \middle| \ \langle \mathbf{\Phi}^{(\text{new})}, \mathbf{X}^* \rangle\right)$ is a new sample.

Theorem

$$\mathcal{E}_n^{\text{gen}} \xrightarrow[n \to \infty]{} E(q_*(\alpha))$$

where

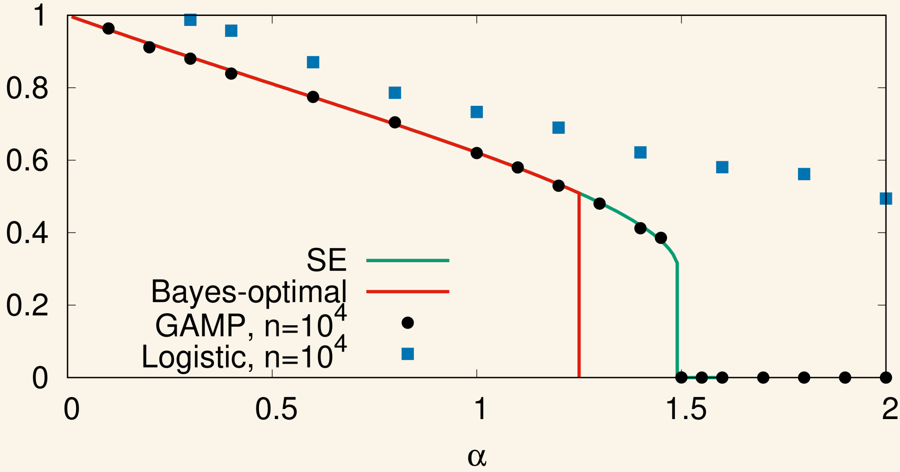
$$E(q) = \mathbb{E}\left[\left(Y^{(q)} - \mathbb{E}[Y^{(q)}|V]\right)^2\right]$$

Recall the second scalar channel: $Y^{(q)} \sim P_{\text{out}}(\cdot \,|\, \sqrt{q}V + \sqrt{\rho - q}W)$, $V, W \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$.

Classification: the perceptron

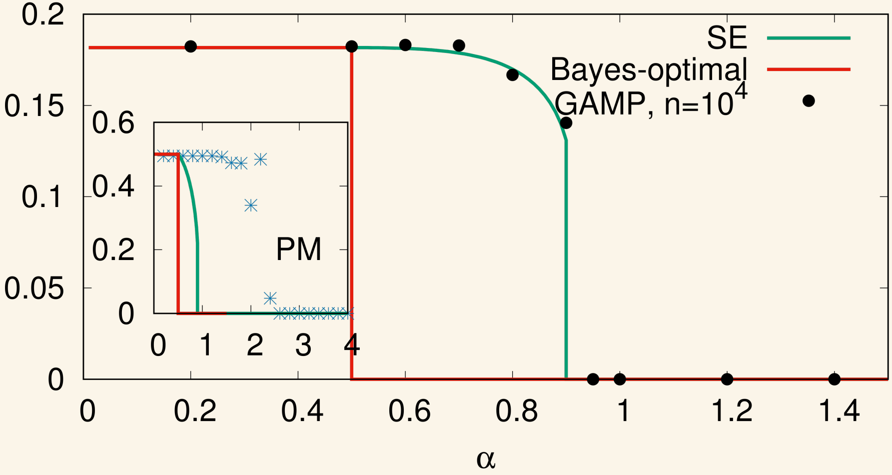
$$\mathbf{Y} = \mathsf{sign}\big(\mathbf{\Phi}\mathbf{X}^*\big), \qquad P_0 = \tfrac{1}{2}\delta_{-1} + \tfrac{1}{2}\delta_{+1}$$

Computed by Györgyi, 1990 and also Seung et al., 1992:



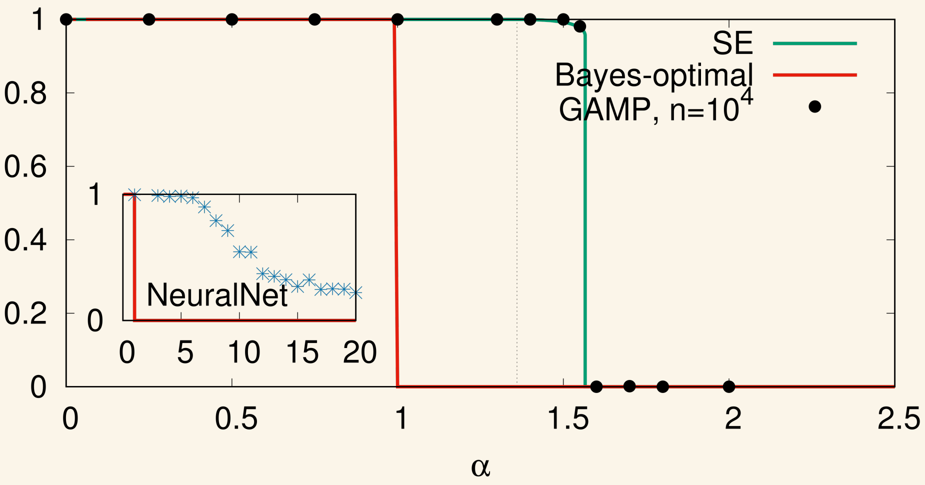
Regression: phase retrieval

$$\mathbf{Y} = \left|\mathbf{\Phi X}^*\right|, \qquad P_0 = \tfrac{1}{2}\delta_{-1} + \tfrac{1}{2}\delta_{+1}$$



Classification: the symmetric door

$$\mathbf{Y} = \mathbf{1}\big(\mathbf{\Phi}\mathbf{X}^* \in [-K, K]\big), \qquad P_0 = \tfrac{1}{2}\delta_{-1} + \tfrac{1}{2}\delta_{+1}$$



Thank you for your attention.

Any questions?

References I

- Barbier, Jean et al. (2016). “The mutual information in random linear estimation”. In: *arXiv preprint arXiv:1607.02335*.
- Donoho, David L, Arian Maleki, and Andrea Montanari (2009). “Message-passing algorithms for compressed sensing”. In: *Proceedings of the National Academy of Sciences* 106.45, pp. 18914–18919.
- Gardner, Elizabeth and Bernard Derrida (1989). “Three unfinished works on the optimal storage capacity of networks”. In: *Journal of Physics A: Mathematical and General* 22.12, p. 1983.
- Györgyi, Géza (1990). “First-order transition to perfect generalization in a neural network with binary synapses”. In: *Physical Review A* 41.12, p. 7097.
- Javanmard, Adel and Andrea Montanari (2013). “State evolution for general approximate message passing algorithms, with applications to spatial coupling”. In: *Information and Inference*, iat004.
- Kabashima, Yoshiyuki (2008). “Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels”. In: *Journal of Physics: Conference Series*. Vol. 95. 1. IOP Publishing, p. 012001.

References II

- Korada, Satish Babu and Nicolas Macris (2010). "Tight bounds on the capacity of binary input random CDMA systems". In: *IEEE Transactions on Information Theory* 56.11, pp. 5590–5613.
- Krzakala, Florent et al. (2012). "Statistical-physics-based reconstruction in compressed sensing". In: *Physical Review X* 2.2, p. 021005.
- Mezard, Marc (1989). "The space of interactions in neural networks: Gardner's computation with the cavity method". In: *Journal of Physics A: Mathematical and General* 22.12, p. 2181.
- Montanari, Andrea (2008). "Estimating random variables from random sparse observations". In: *European Transactions on Telecommunications* 19.4, pp. 385–403.
- Rangan, Sundeep (2011). "Generalized approximate message passing for estimation with random linear mixing". In: *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*. IEEE, pp. 2168–2172.
- Reeves, Galen and Henry D Pfister (2016). "The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact". In: *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE, pp. 665–669.
- Seung, HS, Haim Sompolinsky, and Naftali Tishby (1992). "Statistical mechanics of learning from examples". In: *Physical review A* 45.8, p. 6056.