Specialization Report - AMP for uncertainty

Adam Avedissian
Supervisors : L. Clarté, L. Zdeborová

August 25, 2025

Contents

1 GAMP in higher dimensions **2**
1.1 BP equations . . . 2
1.2 From BP to r-BP . . . 2
1.3 From r-BP to GAMP . . . 4
1.4 GAMP for Empirical Risk Minimization (ERM) . . . 6
1.5 GAMP for Bayes Optimal (BO) estimation . . . 6

2 State Evolution **7**
2.1 SE for ERM . . . 7

References **7**

A SE equations **9**

B Differentiating the proximal operator **10**

1 GAMP in higher dimensions

In our setting the d-dimensional vectors $\mathbf{w}_0$ and $\mathbf{w}_1$ describe our model and are sampled from the same prior $P_W(w)$, a standard Gaussian. The N data points $\mathbf{x}_\mu$ are also sampled from this prior, and their associated labels are as such : $y_\mu \sim \mathcal{N}\left(\mathbf{x}_\mu^\top \mathbf{w}_0, \sigma(\mathbf{x}_\mu^\top \mathbf{w}_1)\right) \equiv \mathcal{N}\left(z_{\mu,0}, \sigma(z_{\mu,1})\right) \equiv P_{\text{out}}(y_\mu | z_{\mu,0}, z_{\mu,1})$. Our goal is to estimate $\mathbf{w}_0$ and $\mathbf{w}_1$

1.1 BP equations

From this, following the derivation in [1], we may write BP equations for the messages $m_{i\to\mu}(\mathbf{W}_i)$ and $m_{\mu\to i}(\mathbf{W}_i)$ where $\mathbf{W}_i$ is the i^{th} column of $\mathbf{W}$.

$$m_{i\to\mu}(\mathbf{W}_i) \propto P_W(\mathbf{W}_i) \prod_{\gamma\neq\mu} \mathbf{m}_{\gamma\to i}(\mathbf{W}_i) \tag{1.1}$$

$$m_{\mu\to i}(\mathbf{W}_i) \propto \int_{\mathbb{R}^{2(d-1)}} \left(\prod_{j\neq i} \mathbf{m}_{j\to\mu}(\mathbf{W}_j)\ \mathrm{d}\mathbf{W}_j \right) P_{\text{out}}(y_\mu | z_{\mu,0}, z_{\mu,1}) \tag{1.2}$$

1.2 From BP to r-BP

The technical part starts here : first we decompose $\mathbf{z}_\mu = x_{\mu,i}\mathbf{W}_i + \sum_{j\neq i} x_{\mu,j}\mathbf{W}_j$. Now, according to the CLT $\sum_{j\neq i} x_{\mu,j}\mathbf{W}_j$ should be a gaussian with mean $\boldsymbol{\omega}_{\mu\to i} = \sum_{j\neq i} x_{\mu,j}\mathbf{a}_{j\to\mu}$ and covariance $\mathbf{V}_{\mu\to i} = \sum_{j\neq i}(x_{\mu,j})^2 \mathbf{v}_{j\to\mu}$ where :

$$\mathbf{a}_{j\to\mu} = \int_{\mathbb{R}^2} \mathbf{W}_j m_{j\to\mu}(\mathbf{W}_j)\ \mathrm{d}\mathbf{W}_j$$
$$\mathbf{v}_{j\to\mu} = \int_{\mathbb{R}^2} \mathbf{W}_j \mathbf{W}_j^\top m_{j\to\mu}(\mathbf{W}_j)\ \mathrm{d}\mathbf{W}_j - \mathbf{a}_{j\to\mu}\mathbf{a}_{j\to\mu}^\top$$

This allows us to rewrite (1.2) with only one 2D integral as such :

$$m_{\mu\to i}(\mathbf{W}_i) \propto \int_{\mathbb{R}^2} P_{\text{out}}(y_\mu | z_{\mu,0}, z_{\mu,1}) \exp\left[-\frac{1}{2} \left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i} - x_{\mu,i}\mathbf{W}_i\right)^\top \mathbf{V}_{\mu\to i}^{-1} \left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i} - x_{\mu,i}\mathbf{W}_i\right) \right]\ \mathrm{d}\mathbf{z}_\mu \tag{1.3}$$

Now, we may develop the exponential term as follows, using that $V_{\mu\to i}$ is symmetric :

$$\begin{aligned} &\left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i} - x_{\mu,i}\mathbf{W}_i\right)^\top \mathbf{V}_{\mu\to i}^{-1} \left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i} - x_{\mu,i}\mathbf{W}_i\right) \\ &= \left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i}\right)^\top \mathbf{V}_{\mu\to i}^{-1} \left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i}\right) - 2x_{\mu,i} \left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i}\right)^\top \mathbf{V}_{\mu\to i}^{-1}\mathbf{W}_i + (x_{\mu,i})^2 \mathbf{W}_i^\top \mathbf{V}_{\mu\to i}^{-1}\mathbf{W}_i \end{aligned}$$

Using the fact that $x_{\mu,i}$ is $O(1/\sqrt{d})$, we may expand up to order $O(1/d)$:

$$\begin{aligned} &\exp\left[-\frac{1}{2} \left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i} - x_{\mu,i}\mathbf{W}_i\right)^\top \mathbf{V}_{\mu\to i}^{-1} \left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i} - x_{\mu,i}\mathbf{W}_i\right) \right] \\ &= \exp\left[-\frac{1}{2} \left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i}\right)^\top \mathbf{V}_{\mu\to i}^{-1} \left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i}\right) \right] \\ &\times \left(1 - \frac{1}{2}(x_{\mu,i})^2 \mathbf{W}_i^\top \mathbf{V}_{\mu\to i}^{-1}\mathbf{W}_i + x_{\mu,i}\mathbf{W}_i^\top \mathbf{V}_{\mu\to i}^{-1}\left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i}\right) + \frac{1}{2}(x_{\mu,i})^2 \left(\left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i}\right)^\top \mathbf{V}_{\mu\to i}^{-1}\mathbf{W}_i \right)^2 + O(1/d) \right) \end{aligned}$$

The last term may be simplified slightly by using that for any symmetric matrix A, we have $\mathbf{x}^\top A\mathbf{y} = \mathbf{y}^\top A\mathbf{x}$:

$$\left(\left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i}\right)^\top \mathbf{V}_{\mu\to i}^{-1}\mathbf{W}_i \right)^2 = \mathbf{W}_i^\top \mathbf{V}_{\mu\to i}^{-1}\left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i}\right)\left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i}\right)^\top \mathbf{V}_{\mu\to i}^{-1}\mathbf{W}_i$$

$$= \mathbf{W}_i^\top \left[\mathbf{V}_{\mu\to i}^{-1}\left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i}\right)\right]\left[\mathbf{V}_{\mu\to i}^{-1}\left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i}\right)\right]^\top \mathbf{W}_i$$

It is now in our best interest to define the output function :

$$\mathbf{g}_{\text{out}}(\boldsymbol{\omega}, y, \mathbf{V}) := \frac{\int_{\mathbb{R}^2} P_{\text{out}}(y|z_0, z_1) e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\omega})^\top \mathbf{V}^{-1}(\mathbf{z}-\boldsymbol{\omega})}\left[\mathbf{V}^{-1}(\mathbf{z}-\boldsymbol{\omega})\right]\, \mathrm{d}\mathbf{z}}{\int_{\mathbb{R}^2} P_{\text{out}}(y|z_0, z_1) e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\omega})^\top \mathbf{V}^{-1}(\mathbf{z}-\boldsymbol{\omega})}\, \mathrm{d}\mathbf{z}} \tag{1.4}$$

By making use of the chain rule and the vector calculus identity $\nabla(\psi\mathbf{A}) = (\nabla\psi)\mathbf{A}^\top + \psi\nabla\mathbf{A}$, a few lines of calculation yield :

$$\begin{aligned}&\frac{\int_{\mathbb{R}^2} P_{\text{out}}(y|z_0, z_1) e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\omega})^\top \mathbf{V}^{-1}(\mathbf{z}-\boldsymbol{\omega})}\left[\mathbf{V}^{-1}(\mathbf{z}-\boldsymbol{\omega})\right]\left[\mathbf{V}^{-1}(\mathbf{z}-\boldsymbol{\omega})\right]^\top\, \mathrm{d}\mathbf{z}}{\int_{\mathbb{R}^2} P_{\text{out}}(y|z_0, z_1) e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\omega})^\top \mathbf{V}^{-1}(\mathbf{z}-\boldsymbol{\omega})}\, \mathrm{d}\mathbf{z}} \\ &= \nabla_{\boldsymbol{\omega}}\mathbf{g}_{\text{out}}(\boldsymbol{\omega}, y, \mathbf{V}) + \mathbf{g}_{\text{out}}(\boldsymbol{\omega}, y, \mathbf{V})\mathbf{g}_{\text{out}}^\top(\boldsymbol{\omega}, y, \mathbf{V}) + \mathbf{V}^{-1}\end{aligned} \tag{1.5}$$

We may now use all of this to transform $m_{\mu\to i}(\mathbf{W}_i)$:

$$\begin{aligned}m_{\mu\to i}(\mathbf{W}_i) &\propto 1 - \frac{1}{2}(x_{\mu,i})^2\mathbf{W}_i^\top\mathbf{V}_{\mu\to i}^{-1}\mathbf{W}_i + x_{\mu,i}\mathbf{W}_i^\top\mathbf{g}_{\text{out}(\mu\to i)} \\ &\quad + \frac{1}{2}(x_{\mu,i})^2\mathbf{W}_i^\top\left[\nabla_{\boldsymbol{\omega}}\mathbf{g}_{\text{out}(\mu\to i)} + \mathbf{g}_{\text{out}(\mu\to i)}\mathbf{g}_{\text{out}(\mu\to i)}^\top + \mathbf{V}_{\mu\to i}^{-1}\right]\mathbf{W}_i + O(1/d) \\ m_{\mu\to i}(\mathbf{W}_i) &\propto 1 + x_{\mu,i}\mathbf{W}_i^\top\mathbf{g}_{\text{out}(\mu\to i)} + \frac{1}{2}(x_{\mu,i})^2\mathbf{W}_i^\top\left[\nabla_{\boldsymbol{\omega}}\mathbf{g}_{\text{out}(\mu\to i)} + \mathbf{g}_{\text{out}(\mu\to i)}\mathbf{g}_{\text{out}(\mu\to i)}^\top\right]\mathbf{W}_i + O(1/d) \\ m_{\mu\to i}(\mathbf{W}_i) &\dot{\propto} \exp\left[x_{\mu,i}\mathbf{W}_i^\top\mathbf{g}_{\text{out}(\mu\to i)} + \frac{1}{2}(x_{\mu,i})^2\mathbf{W}_i^\top\nabla_{\boldsymbol{\omega}}\mathbf{g}_{\text{out}(\mu\to i)}\mathbf{W}_i\right]\end{aligned} \tag{1.6}$$

Where $\mathbf{g}_{\text{out}}(\boldsymbol{\omega}_{\mu\to i}, y_\mu, \mathbf{V}_{\mu\to i}) \equiv \mathbf{g}_{\text{out}(\mu\to i)}$ is used to shorten the notation and $\dot{\propto}$ is to be seen as proportional in leading order $O(1/d)$.
Finally we define $\mathbf{A}_{\mu\to i}$ and $\mathbf{B}_{\mu\to i}$ as follows :

$$\begin{aligned}\mathbf{A}_{\mu\to i} &= -\nabla_{\boldsymbol{\omega}}\mathbf{g}_{\text{out}(\mu\to i)}(x_{\mu,i})^2 \\ \mathbf{B}_{\mu\to i} &= \mathbf{g}_{\text{out}(\mu\to i)}x_{\mu,i}\end{aligned}$$

Which allows us to rewrite the message as :

$$m_{\mu\to i}(\mathbf{W}_i) \dot{\propto} \exp\left[\mathbf{W}_i^\top\mathbf{B}_{\mu\to i} - \frac{1}{2}\mathbf{W}_i^\top\mathbf{A}_{\mu\to i}\mathbf{W}_i\right]$$

We recognize a gaussian form and we may now explicitly calculate the normalization:

$$\int_{\mathbb{R}^2}\exp\left[\mathbf{W}_i^\top\mathbf{B}_{\mu\to i} - \frac{1}{2}\mathbf{W}_i^\top\mathbf{A}_{\mu\to i}\mathbf{W}_i\right]\, \mathrm{d}\mathbf{W}_i = \frac{2\pi\exp\left[\frac{\mathbf{B}_{\mu\to i}^\top\mathbf{A}_{\mu\to i}^{-1}\mathbf{B}_{\mu\to i}}{2}\right]}{\sqrt{\det\mathbf{A}_{\mu\to i}}}$$

Therefore :

$$\begin{aligned}m_{\mu\to i}(\mathbf{W}_i) &= \frac{\sqrt{\det\mathbf{A}_{\mu\to i}}}{2\pi}\exp\left[\mathbf{W}_i^\top\mathbf{B}_{\mu\to i} - \frac{1}{2}\left(\mathbf{W}_i^\top\mathbf{A}_{\mu\to i}\mathbf{W}_i + \mathbf{B}_{\mu\to i}^\top\mathbf{A}_{\mu\to i}^{-1}\mathbf{B}_{\mu\to i}\right)\right] \\ m_{\mu\to i}(\mathbf{W}_i) &= \mathcal{N}\left(\mathbf{W}_i|\mathbf{A}_{\mu\to i}^{-1}\mathbf{B}_{\mu\to i}, \mathbf{A}_{\mu\to i}^{-1}\right)\end{aligned} \tag{1.7}$$

Now for the second message $m_{i\to\mu}$, using that a product of gaussian distributions is still a gaussian we have :

$$m_{i\to\mu}(\mathbf{W}_i) \propto P_W(\mathbf{W}_i)\prod_{\gamma\neq\mu}\mathcal{N}\left(\mathbf{W}_i|\mathbf{A}_{\gamma\to i}^{-1}\mathbf{B}_{\gamma\to i}, \mathbf{A}_{\gamma\to i}^{-1}\right)$$

$$\propto P_W(\mathbf{W}_i)\mathcal{N}\left(\mathbf{W}_i|\mathbf{R}_{i\to\mu},\boldsymbol{\Sigma}_{i\to\mu}\right)$$

Where :

$$\mathbf{R}_{i\to\mu} = \boldsymbol{\Sigma}_{i\to\mu}\left(\sum_{\gamma\neq\mu}\mathbf{B}_{\gamma\to i}\right)$$
$$\boldsymbol{\Sigma}_{i\to\mu} = \left(\sum_{\gamma\neq\mu}\mathbf{A}_{\gamma\to i}\right)^{-1}$$

Finally we close the equation system iteratively by re-defining $\mathbf{a}_{i\to\mu}$ and $\mathbf{v}_{i\to\mu}$ as a function of $\mathbb{R}_{i\to\mu}$ and $\boldsymbol{\Sigma}_{i\to\mu}$. For this it is practical to define the input function :

$$\mathbf{f}_w(\mathbf{R},\boldsymbol{\Sigma}) := \frac{\int_{\mathbb{R}^2} P_W(\mathbf{W})e^{-\frac{1}{2}(\mathbf{W}-\mathbf{R})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{W}-\mathbf{R})}\mathbf{W}\ \mathrm{d}\mathbf{W}}{\int_{\mathbb{R}^2} P_W(\mathbf{W})e^{-\frac{1}{2}(\mathbf{W}-\mathbf{R})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{W}-\mathbf{R})}\ \mathrm{d}\mathbf{W}} \tag{1.8}$$

Furthermore, similarly to what was done with $\mathbf{g}_{\text{out}}$ we have :

$$\frac{\int_{\mathbb{R}^2} P_W(\mathbf{W})e^{-\frac{1}{2}(\mathbf{W}-\mathbf{R})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{W}-\mathbf{R})}\mathbf{W}\mathbf{W}^\top\ \mathrm{d}\mathbf{W}}{\int_{\mathbb{R}^2} P_W(\mathbf{W})e^{-\frac{1}{2}(\mathbf{W}-\mathbf{R})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{W}-\mathbf{R})}\ \mathrm{d}\mathbf{W}} = \nabla_{\Sigma^{-1}\mathbf{R}}\mathbf{f}_w(\mathbf{R},\boldsymbol{\Sigma}) + \mathbf{f}_w(\mathbf{R},\boldsymbol{\Sigma})\mathbf{f}_w^\top(\mathbf{R},\boldsymbol{\Sigma}) \tag{1.9}$$

This yields :

$$\mathbf{a}_{j\to\mu} = \mathbf{f}_w(\mathbf{R}_{j\to\mu},\boldsymbol{\Sigma}_{j\to\mu}) \tag{1.10}$$
$$\mathbf{v}_{j\to\mu} = \nabla_{\boldsymbol{\Sigma}^{-1}\mathbf{R}}\mathbf{f}_w(\mathbf{R}_{j\to\mu},\boldsymbol{\Sigma}_{j\to\mu}) := \mathbf{f}_c(\mathbf{R}_{j\to\mu},\boldsymbol{\Sigma}_{j\to\mu}) \tag{1.11}$$

The r-BP equations are thus :

Algorithm 1 r-BP in higher dimensions

Initialize $\mathbf{a}_i$ and $\mathbf{v}_i$
while Convergence criterion not satisfied **do**
  $\boldsymbol{\omega}_{\mu\to i} \leftarrow \sum_{j\neq i} x_{\mu,j}\mathbf{a}_{j\to\mu}$
  $\mathbf{V}_{\mu\to i} \leftarrow \sum_{j\neq i}(x_{\mu,j})^2\mathbf{v}_{j\to\mu}$
  $\mathbf{A}_{\mu\to i} \leftarrow -\nabla_{\boldsymbol{\omega}}\mathbf{g}_{\text{out}}(\boldsymbol{\omega}_{\mu\to i}, y_\mu, \mathbf{V}_{\mu\to i})(x_{\mu,i})^2$
  $\mathbf{B}_{\mu\to i} \leftarrow \mathbf{g}_{\text{out}}(\boldsymbol{\omega}_{\mu\to i}, y_\mu, \mathbf{V}_{\mu\to i})x_{\mu,i}$
  $\boldsymbol{\Sigma}_{i\to\mu} \leftarrow \left(\sum_{\gamma\neq\mu}\mathbf{A}_{\gamma\to i}\right)^{-1}$
  $\mathbf{R}_{i\to\mu} \leftarrow \boldsymbol{\Sigma}_{i\to\mu}\left(\sum_{\gamma\neq\mu}\mathbf{B}_{\gamma\to i}\right)$
  $\mathbf{a}_{i\to\mu} \leftarrow \mathbf{f}_w(\mathbf{R}_{i\to\mu},\boldsymbol{\Sigma}_{i\to\mu})$
  $\mathbf{v}_{i\to\mu} \leftarrow \mathbf{f}_c(\mathbf{R}_{i\to\mu},\boldsymbol{\Sigma}_{i\to\mu})$
end while

1.3 From r-BP to GAMP

What is left is to transform the r-BP equations into a GAMP algorithm. For this we first define the following :

$$\boldsymbol{\omega}_\mu := \sum_i x_{\mu,i}\mathbf{a}_{i\to\mu}$$

$$\mathbf{V}_\mu := \sum_i (x_{\mu,i})^2 \mathbf{v}_{i\to\mu}$$
$$\mathbf{\Sigma}_i := \left(\sum_\mu \mathbf{A}_{\mu\to i}\right)^{-1}$$
$$\mathbf{R}_i := \mathbf{\Sigma}_i \left(\sum_\mu \mathbf{B}_{\mu\to i}\right)$$

First we notice that :

$$\begin{aligned}\mathbf{V}_\mu &= (x_{\mu,i})^2 \mathbf{v}_{i\to\mu} \\ &\approx (x_{\mu,i})^2 \mathbf{v}_i\end{aligned}$$

Then, expliciting $\mathbf{A}$ and $\mathbf{B}$ in terms of their definitions and expanding in leading order :

$$\begin{aligned}\mathbf{\Sigma}_i^{-1} &= -\nabla_{\mathbf{W}}\mathbf{g}_{\text{out}}(\boldsymbol{\omega}_{\mu\to i}, y_\mu, \mathbf{V}_{\mu\to i})(x_{\mu,i})^2 \\ &\approx -\nabla_{\mathbf{W}}\mathbf{g}_{\text{out}}(\boldsymbol{\omega}_\mu, y_\mu, \mathbf{V}_\mu)(x_{\mu,i})^2 \\ \mathbf{R}_i &= \mathbf{\Sigma}_i\mathbf{g}_{\text{out}}(\boldsymbol{\omega}_{\mu\to i}, y_\mu, \mathbf{V}_{\mu\to i})x_{\mu,i} \\ &\approx \mathbf{\Sigma}_i\mathbf{g}_{\text{out}}(\boldsymbol{\omega}_\mu, y_\mu, \mathbf{V}_\mu)x_{\mu,i} - \mathbf{\Sigma}_i\nabla_{\boldsymbol{\omega}}\mathbf{g}_{\text{out}}(\boldsymbol{\omega}_\mu, y_\mu, \mathbf{V}_\mu)\,(x_{\mu,i})^2\,\mathbf{a}_{i\to\mu} \\ &\approx \mathbf{\Sigma}_i\mathbf{g}_{\text{out}}(\boldsymbol{\omega}_\mu, y_\mu, \mathbf{V}_\mu)x_{\mu,i} - \mathbf{\Sigma}_i\nabla_{\boldsymbol{\omega}}\mathbf{g}_{\text{out}}(\boldsymbol{\omega}_\mu, y_\mu, \mathbf{V}_\mu)\,(x_{\mu,i})^2\,\mathbf{a}_i \\ &\approx \mathbf{a}_i + \mathbf{\Sigma}_i\mathbf{g}_{\text{out}}(\boldsymbol{\omega}_\mu, y_\mu, \mathbf{V}_\mu)x_{\mu,i}\end{aligned}$$

Where the transition from $\mathbf{a}_{i\to\mu}$ to $\mathbf{a}_i$ in the above is done using :

$$\begin{aligned}\mathbf{a}_{i\to\mu} &= \mathbf{f}_w(\mathbf{R}_{i\to\mu}, \mathbf{\Sigma}_{i\to\mu}) \\ &\approx \mathbf{f}_w(\mathbf{R}_i, \mathbf{\Sigma}_i) - \nabla_{\mathbf{R}}\mathbf{f}_w(\mathbf{R}_i, \mathbf{\Sigma}_i)\mathbf{\Sigma}_i\mathbf{B}_{\mu\to i} \\ &\approx \mathbf{a}_i - \mathbf{\Sigma}_i^{-1}\mathbf{v}_i\mathbf{\Sigma}_i\mathbf{g}_{\text{out}}(\boldsymbol{\omega}_\mu, y_\mu, \mathbf{V}_\mu)x_{\mu,i}\end{aligned}$$

Furthermore we have :

$$\begin{aligned}\boldsymbol{\omega}_\mu &= x_{\mu,i}\mathbf{a}_{i\to\mu} \\ &\approx \mathbf{a}_i x_{\mu,i} - \mathbf{\Sigma}_i^{-1}\mathbf{v}_i\mathbf{\Sigma}_i\mathbf{g}_{\text{out}}(\boldsymbol{\omega}_\mu, y_\mu, \mathbf{V}_\mu)\,(x_{\mu,i})^2\end{aligned}$$

In the above, $\approx$ is to be treated as equal in leading order, and Einstein convention for summation is used.
Finally we may use these equations to write a GAMP algorithm :

Algorithm 2 GAMP in higher dimensions

Initialize $\mathbf{a}_i$, $\mathbf{v}_i$ and $\mathbf{g}_{\text{out},\mu}$
while Convergence criterion not satisfied **do**
  $\mathbf{V}_\mu \leftarrow (x_{\mu,i})^2 \mathbf{v}_i$
  $\omega_\mu \leftarrow \mathbf{a}_i x_{\mu,i} - \mathbf{\Sigma}_i^{-1}\mathbf{v}_i\mathbf{\Sigma}_i\mathbf{g}_{\text{out},\mu}\,(x_{\mu,i})^2$
  $\mathbf{g}_{\text{out},\mu} \leftarrow \mathbf{g}_{\text{out}}(\boldsymbol{\omega}_\mu, y_\mu, \mathbf{V}_\mu)$
  $\mathbf{\Sigma}_i \leftarrow \left(-\sum_\mu \nabla_{\boldsymbol{\omega}}\mathbf{g}_{\text{out}}(\boldsymbol{\omega}_\mu, y_\mu, \mathbf{V}_\mu)(x_{\mu,i})^2\right)^{-1}$
  $\mathbf{R}_i \leftarrow \mathbf{a}_i + \mathbf{\Sigma}_i\mathbf{g}_{\text{out},\mu}x_{\mu,i}$
  $\mathbf{a}_i \leftarrow \mathbf{f}_w(\mathbf{R}_i, \mathbf{\Sigma}_i)$
  $\mathbf{v}_i \leftarrow \mathbf{f}_c(\mathbf{R}_i, \mathbf{\Sigma}_i)$
end while

1.4 GAMP for Empirical Risk Minimization (ERM)

We must now calculate the input and output functions to apply the GAMP algorithm in different cases, starting with ERM. For which we have :

$$P_W(\mathbf{W}) = e^{-\beta r(\mathbf{W})} \tag{1.12}$$

$$P_{\text{out}}(y|\mathbf{z}) = e^{-\beta l(y,\mathbf{z})} \tag{1.13}$$

Where β is a parameter sent to $+\infty$ during the gradient descent. This yields for the $\mathbf{f}_w$ and $\mathbf{g}_{\text{out}}$:

$$\mathbf{f}_w(\mathbf{R}, \boldsymbol{\Sigma}) = \text{prox}_{(\mathbf{R},\boldsymbol{\Sigma})}\left(r(\cdot)\right) \tag{1.14}$$

$$\mathbf{g}_{\text{out}}(\boldsymbol{\omega}, y, \mathbf{V}) = \mathbf{V}^{-1}\left(\text{prox}_{(\boldsymbol{\omega},\mathbf{V})}\left(l(y,\cdot)\right) - \boldsymbol{\omega}\right) \tag{1.15}$$

Where we defined

$$\text{prox}_{(\boldsymbol{\mu},\boldsymbol{\Omega})}\left(f(\cdot)\right) := \underset{\mathbf{x}}{\text{argmin}}\left(\frac{(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Omega}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2} + f(\mathbf{x})\right) \tag{1.16}$$

In the case of l_2 regularization and the loss associated to the log likelihood of our model, we can close one of these equations :

$$\mathbf{f}_w(\mathbf{R}, \boldsymbol{\Sigma}) = \left(\boldsymbol{\Sigma}^{-1} + \lambda\mathbb{I}\right)^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{R} \tag{1.17}$$

$$\mathbf{f}_c(\mathbf{R}, \boldsymbol{\Sigma}) = \left(\boldsymbol{\Sigma}^{-1} + \lambda\mathbb{I}\right)^{-1} \tag{1.18}$$

However the second one remains unchanged :

$$\mathbf{g}_{\text{out}}(\boldsymbol{\omega}, y, \mathbf{V}) = \mathbf{V}^{-1}\left(\text{prox}_{(\boldsymbol{\omega},\mathbf{V})}\left(l(y,\cdot)\right) - \boldsymbol{\omega}\right) \tag{1.19}$$

$$l(y,\mathbf{z}) := \frac{(y-z_0)^2}{2\sigma(z_1)} + \frac{\ln(\sigma(z_1))}{2}$$

1.5 GAMP for Bayes Optimal (BO) estimation

For BO we have :

$$P_W(\mathbf{W}) = \mathcal{N}(\mathbf{W}|0,\mathbb{I}) \tag{1.20}$$

$$P_{\text{out}}(y|\mathbf{z}) = \mathcal{N}(y|z_0, \sigma(z_1)) \tag{1.21}$$

Then we may write the input function and its derivative in a simple closed form :

$$\mathbf{f}_w(\mathbf{R}, \boldsymbol{\Sigma}) = \left(\boldsymbol{\Sigma}^{-1} + 1\right)^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{R} \tag{1.22}$$

$$\mathbf{f}_c(\mathbf{R}, \boldsymbol{\Sigma}) = \left(\boldsymbol{\Sigma}^{-1} + 1\right)^{-1} \tag{1.23}$$

However the output function doesn't have a simple closed form, and one integral still remains :

$$\mathbf{g}_{\text{out}}(\boldsymbol{\omega}, y, \mathbf{V}) := \frac{\int_{\mathbb{R}} \mathcal{N}\left(\tilde{\mathbf{z}}|\boldsymbol{\omega}, \tilde{\mathbf{V}}(z_1)\right)\begin{pmatrix} \frac{y\det(\mathbf{V}) + \omega_0 V_1 \sigma(z_1) + (z_1-\omega_1)V\sigma(z_1)}{\det(\tilde{\mathbf{V}})} \\ z_1 \end{pmatrix} \mathrm{d}z_1}{\int_{\mathbb{R}} \mathcal{N}\left(\tilde{\mathbf{z}}|\boldsymbol{\omega}, \tilde{\mathbf{V}}(z_1)\right)\,\mathrm{d}z_1}$$

With :

$$\tilde{\mathbf{z}} := \begin{pmatrix} y \\ z_1 \end{pmatrix}, \qquad \boldsymbol{\omega} := \begin{pmatrix} \omega_0 \\ \omega_1 \end{pmatrix}, \qquad \mathbf{V} := \begin{pmatrix} V_0 & V \\ V & V_1 \end{pmatrix}, \qquad \tilde{\mathbf{V}}(z_1) := \mathbf{V} + \begin{pmatrix} \sigma(\mathbf{z}_1) & 0 \\ 0 & 0 \end{pmatrix}$$

2 State Evolution

The GAMP equations can be further refined to translate into State Evolution (SE) equations, allowing us to obtain the quantities of interest in a less computationally expensive manner. By taking the limit $d \to +\infty$ and preoccupying ourselves only with the overlaps $\mathbf{q} := \mathbf{w}\mathbf{w}^\top/d$ and $\mathbf{m} := \mathbf{w}^*\mathbf{w}^\top/d$ in which $\mathbf{w}$ denotes the estimates and $\mathbf{w}^*$ the truth. A rigorous derivation of these equations can be found in [2] and their general form is detailed in appendix A. When implementing these equations, the multi-dimensional expectations are taken using Monte-Carlo methods. Samples from $\boldsymbol{\omega}, \mathbf{z}, A, \mathbf{w}^*, \xi$ are generated and an empirical mean is used to estimate the expectations.

2.1 SE for ERM

Using the results from the GAMP section, we have the following for the ERM part.
First the simpler part, the input functions :

$$\mathbf{f}_w(\mathbf{R}, (\hat{\boldsymbol{\chi}})^{-1}) = (\hat{\boldsymbol{\chi}} + \lambda\mathbb{I})^{-1}\,\hat{\boldsymbol{\chi}}\mathbf{R} \tag{2.1}$$

$$\mathbf{f}_c(\mathbf{R}, (\hat{\boldsymbol{\chi}})^{-1}) := \partial_{\hat{\boldsymbol{\chi}}\mathbf{R}}\mathbf{f}_w(\mathbf{R}, (\hat{\boldsymbol{\chi}})^{-1}) = (\hat{\boldsymbol{\chi}} + \lambda\mathbb{I})^{-1} \tag{2.2}$$

The output functions are slightly more complex, as derivatives of the proximal are required for them. This step is detailed in appendix B. They thus yield:

$$\mathbf{g}_{\text{out}}(\boldsymbol{\omega}, \varphi(\mathbf{z}, A), \boldsymbol{\sigma}) = \boldsymbol{\sigma}^{-1}\left(\text{prox}_{(\boldsymbol{\omega},\boldsymbol{\sigma})}\left(l(\varphi(\mathbf{z}, A), \cdot)\right) - \boldsymbol{\omega}\right) \tag{2.3}$$

$$\begin{aligned}\nabla_{\mathbf{z}}\mathbf{g}_{\text{out}}(\boldsymbol{\omega}, \varphi(\mathbf{z}, A), \boldsymbol{\sigma}) &= \boldsymbol{\sigma}^{-1}\left(\nabla_{\varphi}\,\text{prox}_{(\boldsymbol{\omega},\mathbf{V})}\left(l(\varphi(\mathbf{z}, A), \cdot)\right)\right)\left(\nabla_{\mathbf{z}}\varphi(\mathbf{z}, A)\right)^\top \\ &= -\left(\boldsymbol{\sigma}^{-1} + \nabla_{\mathbf{z}}\nabla_{\mathbf{z}} l(\varphi(\mathbf{z}, A), \mathbf{z})\right)^{-1}\left(\nabla_{\varphi}\nabla_{\mathbf{z}} l(\varphi(\mathbf{z}, A), \mathbf{z})\right)\left(\nabla_{\mathbf{z}}\varphi(\mathbf{z}, A)\right)^\top\end{aligned} \tag{2.4}$$

$$\begin{aligned}\nabla_{\boldsymbol{\omega}}\mathbf{g}_{\text{out}}(\boldsymbol{\omega}, \varphi(\mathbf{z}, A), \boldsymbol{\sigma}) &= \boldsymbol{\sigma}^{-1}\left(\nabla_{\boldsymbol{\omega}}\,\text{prox}_{(\boldsymbol{\omega},\mathbf{V})}\left(l(\varphi(\mathbf{z}, A), \cdot)\right) - \mathbb{I}_2\right) \\ &= \boldsymbol{\sigma}^{-1}\left(\left(\boldsymbol{\sigma}^{-1} + \nabla_{\mathbf{z}}\nabla_{\mathbf{z}} l(\varphi(\mathbf{z}, A), \mathbf{z})\right)^{-1}\boldsymbol{\sigma}^{-1} - \mathbb{I}_2\right)\end{aligned} \tag{2.5}$$

Where :

$$\begin{aligned}
l(\varphi(\mathbf{z}, A), \mathbf{z}) &:= \frac{(\varphi(\mathbf{z}, A) - z_0)^2}{2\sigma(z_1)} + \frac{\ln(\sigma(z_1))}{2} \\
\nabla_{\mathbf{z}}\nabla_{\mathbf{z}} l(\varphi(\mathbf{z}, A), \mathbf{z}) &= \begin{pmatrix} \frac{1}{\sigma(z_1)} & \frac{(y-z_0)\sigma'(z_1)}{\sigma(z_1)^2} \\ \frac{(y-z_0)\sigma'(z_1)}{\sigma(z_1)^2} & \frac{\sigma(z_1)\sigma''(z_1)\left(\sigma(z_1)-(y-z_0)^2\right)+\sigma'(z_1)^2\left(2(y-z_0)^2-\sigma(z_1)\right)}{2\sigma(z_1)^2} \end{pmatrix} \\
\nabla_{\varphi}\nabla_{\mathbf{z}} l(\varphi(\mathbf{z}, A), \mathbf{z}) &= \begin{pmatrix} -\frac{1}{\sigma(z_1)} \\ \frac{\sigma'(z_1)(z_0-y)}{\sigma(z_1)^2} \end{pmatrix}
\end{aligned}$$

References

[1] Lenka Zdeborová and Florent Krzakala. "Statistical physics of inference: thresholds and algorithms". In: *Advances in Physics* 65.5 (2016), pp. 453–552. DOI: 10.1080/00018732.2016.1211393. eprint: https://doi.org/10.1080/00018732.2016.1211393. URL: https://doi.org/10.1080/00018732.2016.1211393.

[2] Benjamin Aubin et al. “The committee machine: computational to statistical gaps in learning a two-layers neural network*”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.12 (Dec. 2019), p. 124023. DOI: 10.1088/1742-5468/ab43d2. URL: https://dx.doi.org/10.1088/1742-5468/ab43d2.

[3] Stephen Gould et al. *On Differentiating Parameterized Argmin and Argmax Problems with Application to Bi-level Optimization*. July 2016. DOI: 10.48550/arXiv.1607.05447.

A SE equations

Here are the detailed generic SE equations as presented in [2].

$$
\begin{aligned}
\mathbf{m}^{t+1} &= \underset{\mathbf{w}^\star,\xi}{\mathbb{E}}\, \mathbf{f}_w\left(\mathbf{R}^t[\mathbf{w}^\star,\xi], (\hat{\boldsymbol{\chi}}^t)^{-1}\right)(\mathbf{w}^*)^\top \\
\mathbf{q}^{t+1} &= \underset{\mathbf{w}^\star,\xi}{\mathbb{E}}\, \mathbf{f}_w\left(\mathbf{R}^t[\mathbf{w}^\star,\xi], (\hat{\boldsymbol{\chi}}^t)^{-1}\right)\left(\mathbf{f}_w\left(\mathbf{R}^t[\mathbf{w}^\star,\xi], (\hat{\boldsymbol{\chi}}^t)^{-1}\right)\right)^\top \\
\boldsymbol{\sigma}^{t+1} &= \underset{\mathbf{w}^\star,\xi}{\mathbb{E}}\, \mathbf{f}_c\left(\mathbf{R}^t[\mathbf{w}^\star,\xi], (\hat{\boldsymbol{\chi}}^t)^{-1}\right) \\
\hat{\mathbf{q}}^t &= \alpha \underset{\boldsymbol{\omega},\mathbf{z},A}{\mathbb{E}}\, \mathbf{g}_{\text{out}}\left(\boldsymbol{\omega}, \varphi(\mathbf{z},A), \boldsymbol{\sigma}^t\right)\left(\mathbf{g}_{\text{out}}\left(\boldsymbol{\omega}, \varphi(\mathbf{z},A), \boldsymbol{\sigma}^t\right)\right)^\top \\
\hat{\mathbf{m}}^t &= \alpha \underset{\boldsymbol{\omega},\mathbf{z},A}{\mathbb{E}}\, \nabla_{\mathbf{z}} \mathbf{g}_{\text{out}}\left(\boldsymbol{\omega}, \varphi(\mathbf{z},A), \boldsymbol{\sigma}^t\right) \\
\hat{\boldsymbol{\chi}}^t &= -\alpha \underset{\boldsymbol{\omega},\mathbf{z},A}{\mathbb{E}}\, \nabla_{\boldsymbol{\omega}} \mathbf{g}_{\text{out}}\left(\boldsymbol{\omega}, \varphi(\mathbf{z},A), \boldsymbol{\sigma}^t\right)
\end{aligned}
$$

Where :

- $\alpha := N/d$,
- $\mathbf{f}_w, \mathbf{f}_c$ and $\mathbf{g}_{\text{out}}$ are as defined for the GAMP,
- $\mathbf{R}^t[\mathbf{w}^\star,\xi] := (\hat{\boldsymbol{\chi}}^t)^{-1}\left(\hat{\mathbf{m}}^t\mathbf{w}^* + (\hat{\mathbf{q}}^t)^{1/2}\xi\right)\right)$,
- $\varphi(\mathbf{z},A)$ is the labeling process (in our case $\varphi(\mathbf{z},A) = \mathbf{z}_0 + A\sqrt{\sigma(\mathbf{z}_1)}$),
- $\mathbf{w}^*$ is a 2-vector where each of its components are distributed as those in $\mathbf{w}_0$ and $\mathbf{w}_1$,
- $A \sim \mathcal{N}(0,1)$,
- $\xi \sim \mathcal{N}(0,\mathbb{I}_2)$,
- the 4-vector $(\boldsymbol{\omega}, z)$ is distributed as $\mathcal{N}\left(0, \begin{matrix} \mathbb{I}_2 & \mathbf{m}^t \\ \mathbf{m}^t & \mathbf{q}^t \end{matrix}\right)$

B Differentiating the proximal operator

Based of Lemma 3.2 of [3] we have :

Let

$$f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R} \in C^2(\mathbb{R}^n \times \mathbb{R}^m)$$

be an objective function and

$$\mathbf{g}(\mathbf{x}) := \underset{\mathbf{y}\in\mathbb{R}^n}{\text{argmin}} f(\mathbf{x}, \mathbf{y})$$

Then we have

$$\nabla_{\mathbf{x}} \mathbf{g}(\mathbf{x}) = -\left(\nabla_{\mathbf{y}}\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{g}(\mathbf{x}))\right)^{-1} \nabla_{\mathbf{x}}\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{g}(\mathbf{x}))$$

Where

$$\begin{aligned}
&\nabla_{\mathbf{y}}\nabla_{\mathbf{y}} f \in \mathbb{R}^{n\times n}, \left(\nabla_{\mathbf{y}}\nabla_{\mathbf{y}} f\right)_{i,j} = \partial_{y_i}\partial_{y_j} f, \\
&\nabla_{\mathbf{x}}\nabla_{\mathbf{y}} f \in \mathbb{R}^{n\times m}, \left(\nabla_{\mathbf{x}}\nabla_{\mathbf{y}} f\right)_{i,j} = \partial_{y_i}\partial_{x_j} f
\end{aligned}$$

While the proximal depends on more than one parameter, the derivative by more than one at a time is never needed. Thus we may use this formula to obtain all the necessary derivatives for the SE. First, writing $f(\boldsymbol{\omega}, y, \boldsymbol{\sigma}, \mathbf{z}) := \frac{(\mathbf{z}-\boldsymbol{\omega})^\top \boldsymbol{\sigma}^{-1}(\mathbf{z}-\boldsymbol{\omega})}{2} + l(y, \mathbf{z})$. Then defining $f_1(\boldsymbol{\omega}, \mathbf{z}) := f(\boldsymbol{\omega}, y, \boldsymbol{\sigma}, \mathbf{z})$ for fixed vales of y and $\boldsymbol{\sigma}$ as well as $f_2(y, \mathbf{z}) := f(\boldsymbol{\omega}, y, \boldsymbol{\sigma}, \mathbf{z})$ for fixed vales of $\boldsymbol{\omega}$ and $\boldsymbol{\sigma}$ allows us to proceed as defined above. This yields the following :

$$\begin{aligned}
\nabla_{\boldsymbol{\omega}} \operatorname{prox}_{(\boldsymbol{\omega}, \mathbf{V})}\left(l(y, \cdot)\right) &= \left(\boldsymbol{\sigma}^{-1} + \nabla_{\mathbf{z}}\nabla_{\mathbf{z}} l(y, \mathbf{z})\right)^{-1} \boldsymbol{\sigma}^{-1} \\
\nabla_{y} \operatorname{prox}_{(\boldsymbol{\omega}, \mathbf{V})}\left(l(y, \cdot)\right) &= -\left(\boldsymbol{\sigma}^{-1} + \nabla_{\mathbf{z}}\nabla_{\mathbf{z}} l(y, \mathbf{z})\right)^{-1} \nabla_{y}\nabla_{\mathbf{z}} l(y, \mathbf{z})
\end{aligned}$$