Approximate Message Passing

Mohammad Emtiyaz Khan
CS, UBC

February 8, 2012

Abstract

In this note, I summarize Sections 5.1 and 5.2 of Arian Maleki's PhD thesis.

1 Notation

We denote scalars by small letters e.g. $a, b, c, \ldots$, vectors by boldface small letters e.g. $\boldsymbol{\lambda}, \boldsymbol{\alpha}, \mathbf{x}, \ldots$, matrices by boldface capital letter e.g. $\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$, (subsets of) natural numbers by capital letters e.g. $N, M, \ldots$. We denote i'th element of a vector $\mathbf{a}$ by a_i and (i,j)'th entry of a matrix $\mathbf{A}$ by A_{ij}. We denote the i'th column (or row) of $\mathbf{A}$ by $\mathbf{A}_{:,i}$ (or $\mathbf{A}_{i,:}$). We use $\mathbf{A}_{a,-i}$ (or $\mathbf{A}-a,i$ to refer to the a'th row (or i'th column) without the element $A_{a,i}$. Also, $\mathbf{A}^T$ denote the transpose of a matrix $\mathbf{A}$.

2 Basis Pursuit Problem

Given measurements $\mathbf{y}$ of length n and matrix $\mathbf{A}$ of size $n \times N$, we wish to compute $\mathbf{s}$ which is the minimizer of Eq. 1. This is known as the basis pursuit problem. Here, $||\cdot||_1$ is the l_1-norm. A version of this problem where we allow for errors in the measurements is called basis pursuit denoising problem (aka LASSO), shown in Eq. 2. Here, $||\cdot||_2$ is the l_2-norm.

$$\text{BP: } \min_{\mathbf{s}} ||\mathbf{s}||_1, \text{ s.t. } \mathbf{y} = \mathbf{A}\mathbf{s} \tag{1}$$

$$\text{BPDN: } \min_{\mathbf{s}} \lambda||\mathbf{s}||_1 + \tfrac{1}{2}||\mathbf{y} - \mathbf{A}\mathbf{s}||_2^2 \tag{2}$$

3 Posterior Distribution

Consider the following posterior distributions in Eq. 3, where the prior distribution $p(s_i)$ is the Laplace distribution and the likelihood $p(y_a|\mathbf{s}, \mathbf{A}_{a,:})$ is the Dirac distribution.

$$p(\mathbf{s}|\mathbf{y}) \propto \prod_{i=1}^{N} p(s_i) \prod_{a=1}^{n} p(y_a|\mathbf{s}, \mathbf{A}_{a,:}) \tag{3}$$

$$= \prod_{i=1}^{N} \exp(-\beta|s_i|) \prod_{a=1}^{n} \delta(y_a = \mathbf{A}_{a,:}\mathbf{s}) \tag{4}$$

As $\beta \to \infty$, mass of this posterior distribution concentrates around the minimizer of BP. This implies that given the marginals of this posterior distribution, solution of BP is immediate. A formal proof is not given in [Mal11]. We give an intuitive explanation in Fig. 3.

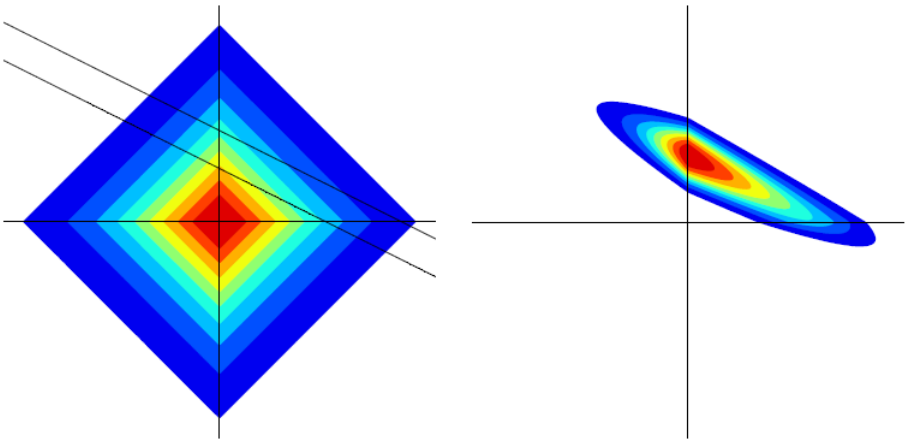

Figure 1: Visualization of the posterior distribution in Eq. 3 for two variables s_1 and s_2. Left figure shows the negative-log of prior distribution which is $\beta(|s_1| + |s_2|)$ and the negative log-likelihood of a single measurement corresponding to a Gaussian likelihood (black lines). Right figure shows the negative-log of the posterior distribution. As $\beta \to \infty$, the posterior become more peaky around the sparse solution where s_1 is zero. We can also see that the marginal of s_1 concentrates around 0, while that of s_2 concentrates around a non-zero value. Figure from [See08].

4 Belief Propagation

Belief propagation can be used to compute the marginal distributions of a posterior distribution. We start by defining a *factor graph* which captures the statistical dependencies between the variables, and then do *message passing*. In this section, we will briefly describe belief propagation for the basis pursuit problem; interested reader should see [Bis06] for a general case. First consider the posterior distribution of Eq. 3 with the prior distribution $p(s_i)$ and likelihood $p(y_a|\mathbf{s}, \mathbf{A}_{a,:})$. We define a bipartite factor graph where $s_1, s_2, \ldots, s_N$ are variables and $y_1, y_2, \ldots, y_n$ are factors. We draw an edge between a variable and a factor if the corresponding measurement depends on the variable (in the BP problem, it will be a dense graph but if $\mathbf{A}$ was sparse then non-zero entries will correspond to an edge). Define, $\mathbb{N}(a)$ to be the neighborhood of a'th factor i.e. the set of variables that are connected to factor a and define $\mathbb{N}(a)\backslash i$ to be the set without the variable i. The messages, defined below, are passed from variables to factors and then factors to variables.

$$m_{i\to a}(s_i) = p(s_i) \prod_{b\in\mathbb{N}(i)\backslash a} m_{b\to i}(s_i) \tag{5}$$

$$m_{a\to i}(s_i) = \int_{\mathbf{s}_{-i}} p(y_a|\mathbf{s}) \prod_{j\in\mathbb{N}(a)\backslash i} m_{j\to a}(s_j) d\mathbf{s}_{-i} \tag{6}$$

Intuitively, the message from a variable i to a factor a contains multiplication of prior belief $p(s_i)$ with all the messages received except the message that was sent by factor a. Similarly, the message from a factor a to a variable i contains multiplication of the likelihood $p(y_a|\mathbf{s})$ with all the messages received except the message that was sent by variable i. The variables other than i are then integrated out of the message. The marginal of a variable is then given by multiplication of all the messages that arrive at that variable along with the local belief as shown below.

$$p(s_i|\mathbf{y}) = p(s_i) \prod_{b\in\mathbb{N}(i)} m_{b\to i}(s_i) \tag{7}$$

We will now give a simple example to show that message passing results in the marginals at each node. Consider two variables s_1, s_2 and s_3 with two measurements y_a and y_b, following a joint distribution which

factorizes as shown below.

$$p(y_a, y_b, s_1, s_2, s_3) = p(y_a|s_1, s_2)p(y_a|s_2, s_3)p(s_1)p(s_2)p(s_3) \tag{8}$$

The statistical dependencies between variable and measurements can be expressed using the following factor graph: $s_1 - y_a - s_2 - y_b - s_3$. Here, y_a depends on s_1, s_2 and y_b depends on s_2, s_3. Using Eq. 5 and 6, we can write down the messages explicitly as shown below.

Messages from factors to variables:

$$m_{a\to 1}(s_1) = \int_{s_2} p(y_a|s_1, s_2)m_{2\to a}(s_2)ds_2 \tag{9}$$

$$m_{a\to 2}(s_2) = \int_{s_1} p(y_a|s_1, s_2)m_{1\to a}(s_1)ds_1 \tag{10}$$

$$m_{b\to 2}(s_2) = \int_{s_3} p(y_b|s_2, s_3)m_{3\to b}(s_3)ds_3 \tag{11}$$

$$m_{b\to 3}(s_3) = \int_{s_2} p(y_b|s_2, s_3)m_{2\to b}(s_2)ds_2 \tag{12}$$

Messages from variables to factors:

$$m_{1\to a}(s_1) = p(s_1) \tag{13}$$

$$m_{2\to a}(s_2) = p(s_2)m_{b\to 2}(s_2) \tag{14}$$

$$m_{2\to b}(s_2) = p(s_2)m_{a\to 2}(s_2) \tag{15}$$

$$m_{3\to b}(s_3) = p(s_3) \tag{16}$$

Now, we establish that this message passing will result in the marginal of s_1, s_2 and s_3. The marginal of s_1 is simplified below in Eq. 22.

$$p(s_1|y_a, y_b) \propto p(s_1, y_a, y_b) \tag{17}$$

$$= \int_{s_2}\int_{s_3} p(s_1, s_2, s_3, y_a, y_b)ds_3ds_2 \tag{18}$$

$$= \int_{s_2}\int_{s_3} p(y_a, y_b|s_1, s_2, s_3)p(s_1, s_2, s_3)ds_3ds_2 \tag{19}$$

$$= \int_{s_2}\int_{s_3} p(y_a|s_1, s_2)p(y_b|s_2, s_3)p(s_1)p(s_2)p(s_3)ds_3ds_2 \tag{20}$$

$$= p(s_1)\int_{s_2}\int_{s_3} p(y_a|s_1, s_2)p(y_b|s_2, s_3)p(s_2)p(s_3)ds_3ds_2 \tag{21}$$

$$= p(s_1)\int_{s_2} p(y_a|s_1, s_2)p(s_2)\int_{s_3} p(y_b|s_2, s_3)p(s_3)ds_3ds_2 \tag{22}$$

We see that after the following 4 message passes $3 \to b, b \to 2, 2 \to a, a \to 1$, we get the marginal of s_1.

$$p(s_1|y_a, y_b) \propto \overbrace{p(s_1)\int_{s_2} p(y_a|s_1, s_2)}^{m_{a\to 1}(s_1)} \underbrace{p(s_2)\overbrace{\int_{s_3} p(y_b|s_2, s_3)\underbrace{p(s_3)}_{m_{3\to b}(s_3)}ds_3}^{m_{b\to 2}(s_2)}}_{m_{2\to a}(s_2)} ds_2 \tag{23}$$

Similarly, marginal of s_2 can be written as follows,

$$p(s_2|y_a, y_b) \propto p(s_2)\int_{s_1} p(y_a|s_1, s_2)p(s_1)ds_1 \int_{s_3} p(y_b|s_2, s_3)p(s_3)ds_3ds_2 \tag{24}$$

and after the following 4 message passes $3 \to b, b \to 2, 1 \to a, a \to 2$, we get the marginal of s_2.

$$p(s_2|y_a, y_b) \propto p(s_2) \overbrace{\int_{s_1} p(y_a|s_1, s_2) \underbrace{p(s_1)}_{m_{1\to a}(s_1)} ds_1}^{m_{a\to 1}(s_1)} \overbrace{\int_{s_3} p(y_b|s_2, s_3) \underbrace{p(s_3)}_{m_{3\to b}(s_3)} ds_3}^{m_{b\to 2}(s_2)} \tag{25}$$

5 Approximate Message Passing

Our goal is to compute the marginal distribution of the following posterior distribution,

$$p_1(\mathbf{s}|\mathbf{y}) \propto \prod_{i=1}^{N} \exp(-\beta|s_i|) \prod_{a=1}^{n} \delta(y_a = \mathbf{A}_{a,:}\mathbf{s}) \tag{26}$$

We define a factor graph with $\{s_i\}_{i=1}^N$ as variables and $\{y_a\}_{a=1}^n$ as factors. From the posterior distribution, it is easy to see that every y_a depends on all s_i's. Therefore, in the factor graph each y_a is connected to all the s_i's, i.e. the factor graph is a fully connected bipartite graph where each factor is connected to all variables. Using the belief propagation algorithm, we can compute marginal distributions of all variables s_i. A direct application of Eq. 5 and 6, however, is not possible because of the following reasons:

1. The marginal distributions $p(s_i|\mathbf{y})$ are not Gaussians since the likelihood $p(\mathbf{y}|\mathbf{s})$ is not conjugate to the prior distribution $p(\mathbf{s})$. Similarly, messages are also non-Gaussian and it is not clear how to parameterize them.

2. Number of messages that need to be propagated every iteration is in $O(nN)$ since every variable sends n messages to every factor (and vice-versa).

Problem (1) can be solved by approximating the messages by Gaussians using Lemma 5.1, 5.2 and 5.3. Problem (2) can be solved by using Lemma 5.4, which makes more approximations on messages to make them independent of the sink of the messages. We will now describe these lemmas briefly. We will leave the exact description of "approximations" in these lemmas and focus on intuitive explanations; please see [Mal11] for a detailed description.

For problem (1), it turns out that if the third moment of a message is bounded then a Gaussian approximation is a reasonable one. This is shown in next two lemmas. The following lemma assumes that if messages from variables to factor have their third moment bounded, then messages from factor to variables can be approximated by Gaussians. This lemma can be proved by using Eq. 6 and applying the Berry-Eseen central limit theorem.

Lemma 5.1. *Let us denote the mean and variance of the messages $m_{j\to a}(s_j)$ by X_{ja} and T_{ja}/β and assume that their third moment is bounded, then messages $m_{a\to i}(s_i)$ are "close" to the Gaussian distribution given in Eq. 27, defined through the mean parameter M_{ai} and variance parameter V_{ai} given in Eq. 28 and 29.*

$$m_{a\to i} \approx \mathcal{N}\left(\frac{M_{ai}}{A_{ai}}, \frac{V_{ai}}{\beta A_{ai}^2}\right) \tag{27}$$

$$M_{ai} := y_a - \mathbf{A}_{a,-i}\mathbf{X}_{-i,a} \tag{28}$$

$$V_{ai} := \mathbf{A}_{a,-i} diag(\mathbf{T}_{-i,a}) \mathbf{A}_{a,-i}^T \tag{29}$$

The following lemma shows that if messages from factors to variables are Gaussians, then message from variables to factors will follow a simple distribution. This lemma can be proved by a direct application of 5.

Algorithm 1 Message passing algorithm for the basis-pursuit problem

Require: Measurements $\mathbf{y}$ and matrix $\mathbf{A}$
Ensure: Marginals of the distribution Eq. 3
  $X_{ai} \leftarrow 0, \forall a, i$ and $v = 1$
  repeat
    for $a = 1, 2, \ldots, n$ **do**
      for $i = 1, 2, \ldots, N$ **do**
        $M_{ai} \leftarrow y_a - \mathbf{A}_{a,-i}\mathbf{X}_{-i,a}$
        $v \leftarrow \frac{v}{N} \sum_{i=1}^{N} \eta'\left(\mathbf{A}_{:,i}^T \mathbf{M}_{:,i}, v\right)$
      end for
    end for
    for $a = 1, 2, \ldots, n$ **do**
      for $i = 1, 2, \ldots, N$ **do**
        $X_{ia} \leftarrow \eta\left(\mathbf{A}_{-a,i}^T \mathbf{M}_{-a,i}, v\right)$
      end for
    end for
  until convergence

Lemma 5.2. *Assuming that each $m_{a\to i}(s_i)$ follows the Gaussian distribution defined in Eq. 27, the messages $m_{i\to a}(s_i)$ follow a distribution given in Eq. 30 which is defined through a distribution defined in Eq. 31.*

$$m_{i\to a}(s_i) \approx p_\beta\left(s_i | \mathbf{A}_{-a,i}^T \mathbf{M}_{-a,i}, V_{ai}\right) \tag{30}$$

$$p_\beta(s|\mu, \sigma^2) \propto \exp\left[-\beta|s| - \frac{\beta}{2\sigma^2}(s-\mu)^2\right] \tag{31}$$

A simple algorithm is to represent these messages by only first two moments. We can start the distribution from variables to factor $m_{j\to a}$ to a standard Gaussian, i.e. $X_{ja} = 0$ and $T_{ja} = 1$, then iterate as follows:

$$M_{ai} \leftarrow y_a - \mathbf{A}_{a,-i}\mathbf{X}_{-i,a} \tag{32}$$

$$V_{ai} \leftarrow \mathbf{A}_{a,-i}\mathrm{diag}(\mathbf{T}_{-i,a})\mathbf{A}_{a,-i}^T \tag{33}$$

$$X_{ia} \leftarrow \mathrm{Mean}\left[p_\beta\left(s_i | \mathbf{A}_{-a,i}^T \mathbf{M}_{-a,i}, V_{ai}\right)\right] \tag{34}$$

$$T_{ia} \leftarrow \mathrm{Variance}\left[p_\beta\left(s_i | \mathbf{A}_{-a,i}^T \mathbf{M}_{-a,i}, V_{ai}\right)\right] \tag{35}$$

This algorithm can be simplified further by assuming that V_{ai} is equal to a constant v for all a, i, then replacing $\mathbf{A}_{-a,i}^T \mathbf{M}_{-a,i}$ by $\mathbf{A}_{:,i}^T \mathbf{M}_{:,i}$ in Eq. 35 and then approximating Eq. 33 by a sample average.

$$M_{ai} \leftarrow y_a - \mathbf{A}_{a,-i}\mathbf{X}_{-i,a} \tag{36}$$

$$v \leftarrow \frac{1}{N}\sum_{i=1}^{N} \mathrm{Variance}\left[p_\beta\left(s_i | \mathbf{A}_{:,i}^T \mathbf{M}_{:,i}, v\right)\right] \tag{37}$$

$$X_{ia} \leftarrow \mathrm{Mean}\left[p_\beta\left(s_i | \mathbf{A}_{-a,i}^T \mathbf{M}_{-a,i}, v\right)\right] \tag{38}$$

Next lemma shows that in the limit as $\beta \to \infty$, computation of mean and variance can be done by a simple soft-thresholding function.

Lemma 5.3. *For bounded μ and σ^2,*

$$\lim_{\beta\to\infty} \mathit{Mean}\left[p_\beta(s|\mu, \sigma^2)\right] = \eta(\mu, \sigma^2) \tag{39}$$

$$\lim_{\beta\to\infty} \mathit{Variance}\left[p_\beta(s|\mu, \sigma^2)\right] = \sigma^2 \eta'(\mu, \sigma^2) \tag{40}$$

Algorithm 2 Approximate message passing algorithm for the basis-pursuit problem

Require: Measurements $\mathbf{y}$ and matrix $\mathbf{A}$
Ensure: Marginals of the distribution Eq. 3
  $\mathbf{x} \leftarrow 0, \mathbf{m} \leftarrow 0$ and $v = 1$
  repeat
    $\mathbf{x}^{old} \leftarrow \mathbf{x}$
    $v^{old} \leftarrow v$
    $\mathbf{t} \leftarrow \mathbf{A}^T\mathbf{m}$
    $\mathbf{x} \leftarrow \eta(\mathbf{t} + \mathbf{x}, v)$
    $v \leftarrow \frac{v}{\delta} \langle \eta' (\mathbf{t} + \mathbf{x}, v) \rangle$
    $\mathbf{m} \leftarrow \mathbf{y} - \mathbf{A}\mathbf{x} + \frac{1}{\delta}\mathbf{m}. * \left\langle \eta' \left(\mathbf{t} + \mathbf{x}^{old}, v^{old}\right) \right\rangle$
  until convergence

where $\eta(\mu, v)$ is the soft-threshold function where takes a value $\mu - v$ if $\mu > v$ or $\mu + v$ if $\mu < -v$ and zero elsewhere, $\eta'(\mu, v)$ is the derivative of $\eta(\mu, v)$.

Using this, we get the following message passing algorithm shown in Algorithm 1. Although this algorithm is simple, we still have too many messages. Each of these steps require matrix multiplication which needs to be done for all variables and factors. The following lemma shows that given a certain asymptotic behavior, a message can be approximated by another message that is independent of the sink, i.e. independent of the variable/factor that the message is sent to. This lemma can be derived by simply substituting the assumptions of Eq. 41 and 42 in the message passing iterations of Algorithm 1, and then simplifying by removing the term which are $O(1/N)$.

Lemma 5.4. *Denote the messages at k'th iteration with a subscript (k). Let us assume that the messages at k'th iteration follow the following asymptotic behavior:*

$$X_{ia}^{(k)} = x_i^{(k)} + \delta X_{ia}^{(k)} + O(1/N) \tag{41}$$

$$M_{ai}^{(k)} = m_a^{(k)} + \delta M_{ai}^{(k)} + O(1/N) \tag{42}$$

with $\delta X_{ia}^{(k)}, \delta M_{ai}^{(k)} = O(1/N)$, then variable $x_i^{(k)}$ and $m_a^{(k)}$ satisfy the following,

$$x_i^{(k)} = \eta \left(\mathbf{A}_{:,i}^T \mathbf{m}^{(k-1)} + x_i^{(k-1)}, v^{(k-1)} \right) + o_N(1) \tag{43}$$

$$m_a^{(k)} = y_a - \mathbf{A}_{a,:}\mathbf{x}^{(k)} + \frac{1}{\delta} m_a^{(k-1)} \left\langle \eta' \left(\mathbf{A}^T \mathbf{m}^{(k-1)} + \mathbf{x}^{(k-1)}, v^{(k-1)} \right) \right\rangle + o_N(1) \tag{44}$$

$$v^{(k)} = \frac{v^{(k-1)}}{\delta} \left\langle \eta' \left(\mathbf{A}^T \mathbf{m}^{(k-1)} + \mathbf{x}^{(k)}, v^{(k-1)} \right) \right\rangle \tag{45}$$

where $o_N(1)$ terms vanish as $N, n \to \infty$.

Using this lemma, we can simplify Algorithm 1 to obtain Algorithm 2.

References

[Bis06] C. Bishop. *Pattern recognition and machine learning.* Springer, 2006.

[Mal11] A. Maleki. *Approximate message passing algorithms for compressed sensing.* PhD thesis, Stanford University, 2011.

[See08] M. Seeger. Bayesian Inference and Optimal Design in the Sparse Linear Model. *J. of Machine Learning Research*, 9:759–813, 2008.