Introduction to Probabilistic Machine Learning

Ralf Herbrich, Rainer Schlosser

Graphical Models: Approximate Inference

Overview

1. Distance Measures for Distributions
2. Approximating Marginals: Expectation Propagation
3. Approximating Normalization Constants

Overview

1. **Distance Measures for Distributions**
2. Approximating Marginals: Expectation Propagation
3. Approximating Normalization Constants

Distance Measures: α-Divergence

- **Problem**. We have a non-Gaussian normalized marginal $p(\cdot)$ and would like to approximate it by a Gaussian $q(\cdot) = \mathcal{N}(\cdot; \mu, \sigma^2)$. What is the optimal approximation?
- **Solution**. To define "optimality", we need a distance measure between probability densities $p(\cdot)$ and $q(\cdot)$!
- **α-Divergence (Amari, 1985)**. *Given two probability densities $p(\cdot)$ and $q(\cdot)$ and $\alpha \in \mathbb{R} \setminus \{0,1\}$ the α-divergence $D_\alpha[p,q]$ is defined by*

$$D_\alpha[p,q] = \frac{1}{\alpha(1-\alpha)} \cdot \left(1 - \int_{-\infty}^{+\infty} \left[\frac{p(x)}{q(x)}\right]^\alpha \cdot q(x)\, \mathrm{d}x\right)$$

Expectation of $\left[\frac{p(x)}{q(x)}\right]^\alpha$ over $q(x)$

If $p = q$ then $\left[\frac{p(x)}{q(x)}\right]^\alpha = 1$ and the expectation is 1

  - **Non-Negativity**: If $p = q$ then $D_\alpha[p,q] = 0$; otherwise $D_\alpha[p,q] > 0$
  - **Asymmetry**: $D_\alpha[p,q] \neq D_\alpha[q,p]$
  - **Flexibility**:
    - $\alpha > 1$ gives more weight to regions where $p(x) > q(x)$
    - $\alpha < 1$ gives more weight to regions where $q(x) > p(x)$

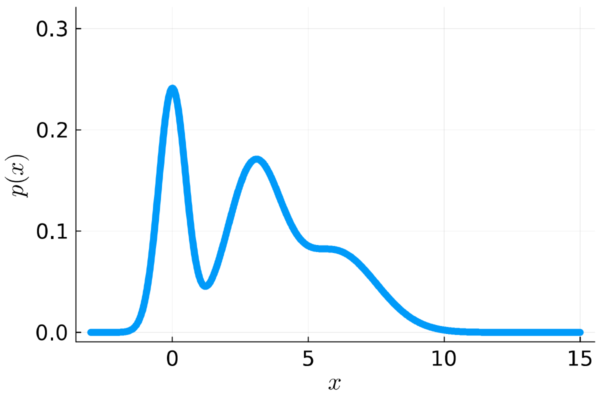
Hasso Plattner Institut
Digital Engineering · Universität Potsdam

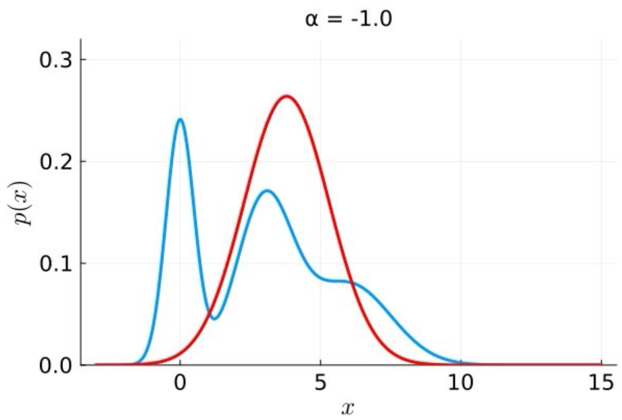
Shun'ichi Amari (甘利 俊一)
(1936)

Introduction to Probabilistic Machine Learning

Unit 5 – Graphical Models: Approximate Inference

α-Divergence with a Gaussian in Pictures



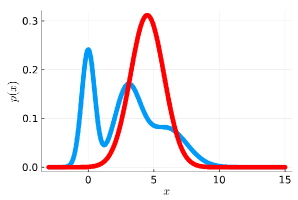


zero-forcing & mode seeking

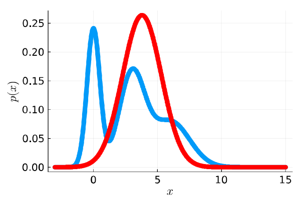
inclusive & support seeking

α

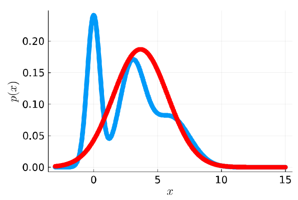
$\alpha = -\infty$



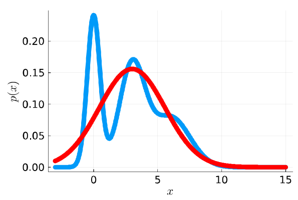
$\alpha = -1$



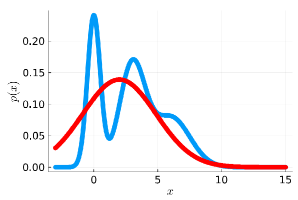
$\alpha = 0$



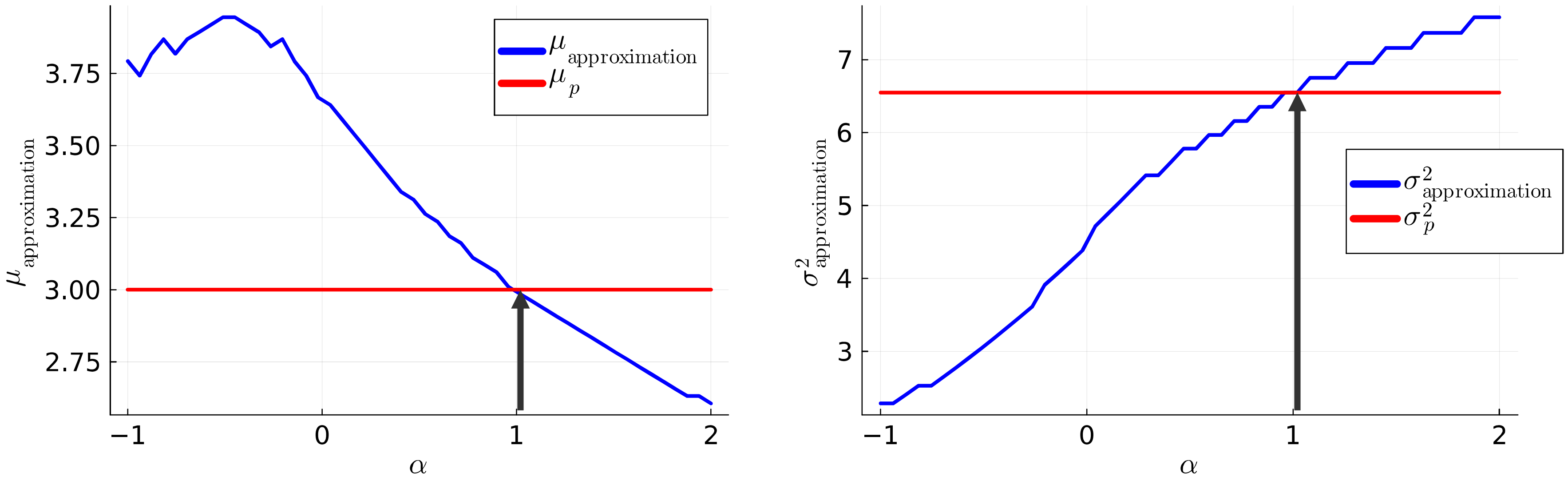
$\alpha = 1$



$\alpha = +\infty$



α-Divergence with a Gaussian and Moment Matching



- Only for $\alpha = 1$ both the first and second moment (that is, mean and variance) are matched with that of the approximation!
- The case $\alpha = 1$ is a limit case.

$\alpha = 1$: KL Divergence

- **Theorem (Limit $\alpha \to 1$).** *Given two probability densities $p(\cdot)$ and $q(\cdot)$ the limit of the α-divergence $D_\alpha[p,q]$ for $\alpha \to 1$ is the Kullback-Leibler divergence*

$$\lim_{\alpha\to 1} D_\alpha[p,q] = \mathrm{KL}[p,q] := \int_{-\infty}^{+\infty} \log\left(\frac{p(x)}{q(x)}\right) \cdot p(x)\, \mathrm{d}x$$

  - **Proof**: Taking limits, we have

$$\lim_{\alpha\to 1} D_\alpha[p,q] = \lim_{\alpha\to 1} \frac{1}{\alpha(1-\alpha)} \cdot \left(1 - \int_{-\infty}^{+\infty} \left[\frac{p(x)}{q(x)}\right]^\alpha \cdot q(x)\, \mathrm{d}x\right)$$

L'Hôpital's rule!

Note that $\frac{\mathrm{d}}{\mathrm{d}z}\, b^z = \log(b) \cdot b^z$

$$= \lim_{\alpha\to 1} \frac{1}{1-2\alpha} \cdot \left(- \int_{-\infty}^{+\infty} \log\left(\frac{p(x)}{q(x)}\right) \cdot \left[\frac{p(x)}{q(x)}\right]^\alpha \cdot q(x)\, \mathrm{d}x\right) = \int_{-\infty}^{+\infty} \log\left(\frac{p(x)}{q(x)}\right) \cdot p(x)\, \mathrm{d}x$$

- **Theorem (Moment Matching)**. *Given any distribution $p(\cdot)$ the minimizer $\mu^*, {\sigma^2}^*$ of the KL divergence $\mathrm{KL}[p(\cdot), \mathcal{N}(\cdot;\mu,\sigma^2)]$ to a Gaussian distribution has*

$$\mu^* = E_{X\sim p(\cdot)}[X] \quad \text{and} \quad {\sigma^2}^* = E_{X\sim p(\cdot)}[X^2] - (\mu^*)^2$$

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Solomon Kullback
(1909 – 1994)

Richard Leibler
(1914 – 2003)

Introduction to Probabilistic Machine Learning

Unit 5 – Graphical Models: Approximate Inference

Overview

1. Distance Measures for Distributions
2. **Approximating Marginals: Expectation Propagation**
3. Approximating Normalization Constants

Sum-Product Algorithm Revisited

- The key operation for factor $f(x_1, x_2, \ldots, x_n)$ and variable X_1 is

$$m_{f\to X_1}(x_1) = \sum_{\{x_2\}} \cdots \sum_{\{x_n\}} f(x_1, x_2, \ldots, x_n) \prod_{j=2}^{n} m_{X_j\to f}(x_j)$$

If all $m_{X_j\to f}(x_j)$ are Gaussian, the result **might not be** Gaussian!

- Based on outgoing messages, we can compute both non-normalized marginals $p_X(\cdot)$ and $m_{X\to f}(\cdot)$

$$p_X(x) = \prod_{f\in \mathrm{ne}(X)} m_{f\to X}(x) \qquad m_{X\to f}(x) = \frac{p_X(x)}{m_{f\to X}(x)}$$

If all $m_{X_j\to f}(x_j)$ are Gaussian, the result **must be** Gaussian!

- **Idea**:
  1. We approximate all outgoing messages $m_{f\to X}(\cdot)$ by a Gaussian $\widehat{m}_{f\to X}(\cdot) = \mathcal{N}(\cdot; \mu, \sigma^2)$
  2. We measure the approximation quality in the normalized marginal, **not** the outgoing message

$$\hat{p}(\cdot) = \arg\min_{\mu,\sigma^2} \mathrm{KL}\left[\frac{m_{f\to X}(\cdot)\cdot \widehat{m}_{X\to f}(\cdot)}{\int_{-\infty}^{+\infty} m_{f\to X}(\tilde{x})\cdot \widehat{m}_{X\to f}(\tilde{x})\,\mathrm{d}\tilde{x}}, \frac{\mathcal{N}(\cdot;\mu,\sigma^2)\cdot \widehat{m}_{X\to f}(\cdot)}{\int_{-\infty}^{+\infty} \mathcal{N}(\tilde{x};\mu,\sigma^2)\cdot \widehat{m}_{X\to f}(\tilde{x})\,\mathrm{d}\tilde{x}}\right]$$

True normalized marginal with approximate incoming message

Approximate marginal with approximate incoming message

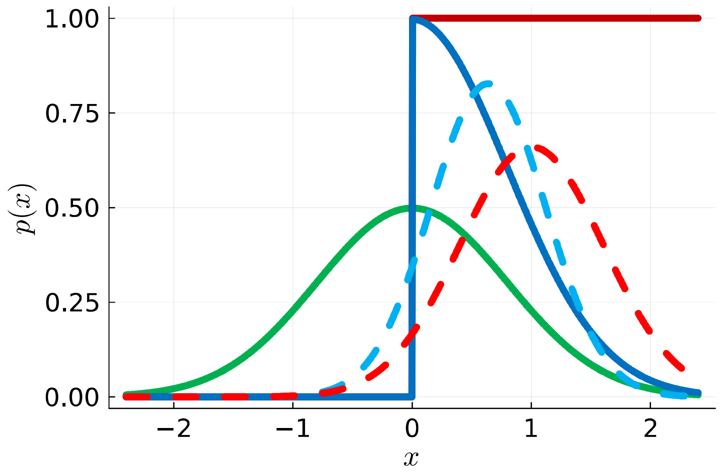
Approximate Message Passing: Example

$$\boxed{f(x) = \mathbb{I}(x > 0)}$$

$$\hat{m}_{X\to f}(x) \propto \frac{\hat{p}_X(x)}{\hat{m}_{f\to X}(x)} \longrightarrow p_X(x) \propto f(x) \cdot \hat{m}_{X\to f}(x)$$

$$\hat{p}_X(x) = \mathcal{N}\left(x; E_{X\sim p_X}[X], \mathrm{var}_{X\sim p_X}[X]\right)$$

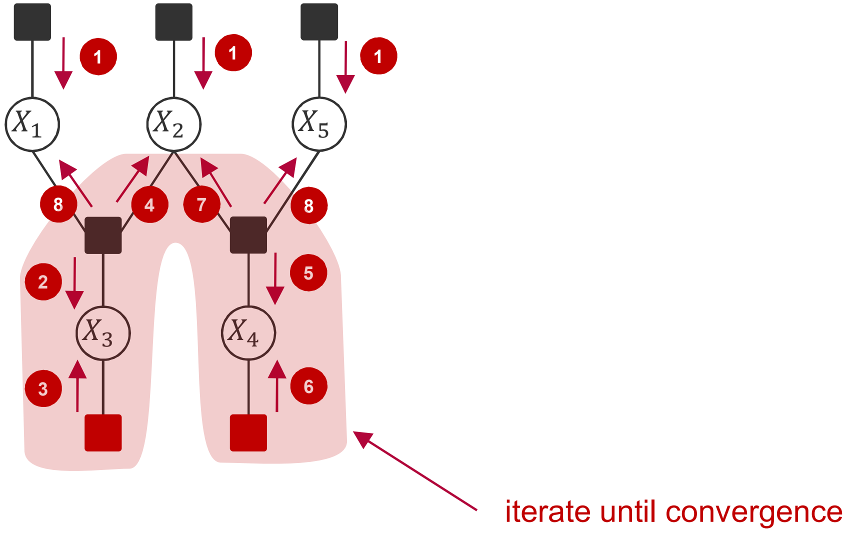
$$\hat{m}_{f\to X}(x) \propto \frac{\hat{p}_X(x)}{\hat{m}_{X\to f}(x)}$$



Expectation Propagation

- **Idea**: If we have factors in the factor graph that require approximate messages, we keep iterating on the whole path between them until convergence minimizing $\mathrm{KL}\left(p(\cdot)|\mathcal{N}(\cdot;\mu,\sigma^2)\right)$ locally for the affected marginals of the approximate factor.
- **Theorem (Minka, 2003)**: *The approximate message passing algorithm using the Kullback-Leibler divergence will always converge if the approximating distribution is in the exponential family!*

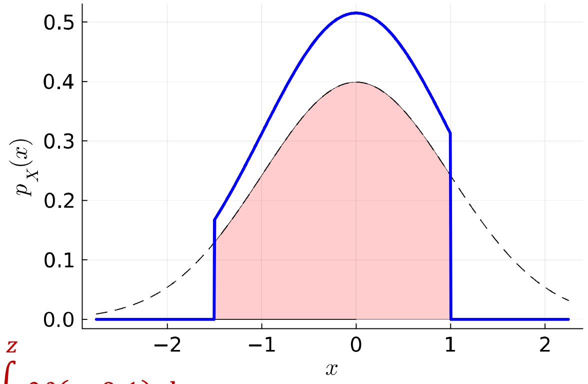
Tom Minka



Doubly-Truncated Gaussians

- **Doubly-Truncated Gaussian**. *Given* $l, u \in \mathbb{R}$, $\mu \in \mathbb{R}$ *and* $\sigma^2 \in \mathbb{R}^+$, *a random variable* X *has a* doubly-truncated Gaussian distribution *if*

$$p_X(x) \propto \mathbb{I}(l < x < u) \cdot \mathcal{N}(x; \mu, \sigma^2)$$



- **Moments of Doubly-Truncated Gaussian**. *Given a random variable* X *that has a doubly-truncated Gaussian distribution and* $t_a \coloneqq a/\sigma$, *we know*

$$E[X^0] = \Phi(t_{u-\mu}) - \Phi(t_{l-\mu})$$

$$\Phi(z) \coloneqq \int_{-\infty}^{z} \mathcal{N}(x; 0, 1)\, \mathrm{d}x$$

$$E[X^1] = \mu + \sigma \cdot \frac{\mathcal{N}(t_{l-\mu}) - \mathcal{N}(t_{u-\mu})}{\Phi(t_{u-\mu}) - \Phi(t_{l-\mu})}$$

Additive correction that goes to zero as $u \to \infty$ and $l \to -\infty$

$$E[X^2] = \mu^2 + \sigma^2 \cdot \left[1 - \frac{t_{u+\mu} \cdot \mathcal{N}(t_{u-\mu}) - t_{l+\mu} \cdot \mathcal{N}(t_{l-\mu})}{\Phi(t_{u-\mu}) - \Phi(t_{l-\mu})}\right]$$

Multiplicative correction that goes to one as $u \to \infty$ and $l \to -\infty$

Doubly-Truncated Gaussians (ctd)

- Using the variance decomposition theorem we see

$$E[X] = \mu + \sigma \cdot v_{\frac{l}{\sigma},\frac{u}{\sigma}}\left(\frac{\mu}{\sigma}\right) \qquad \mathrm{var}[X] = \sigma^2 \cdot \left[1 - w_{\frac{l}{\sigma},\frac{u}{\sigma}}\left(\frac{\mu}{\sigma}\right)\right]$$

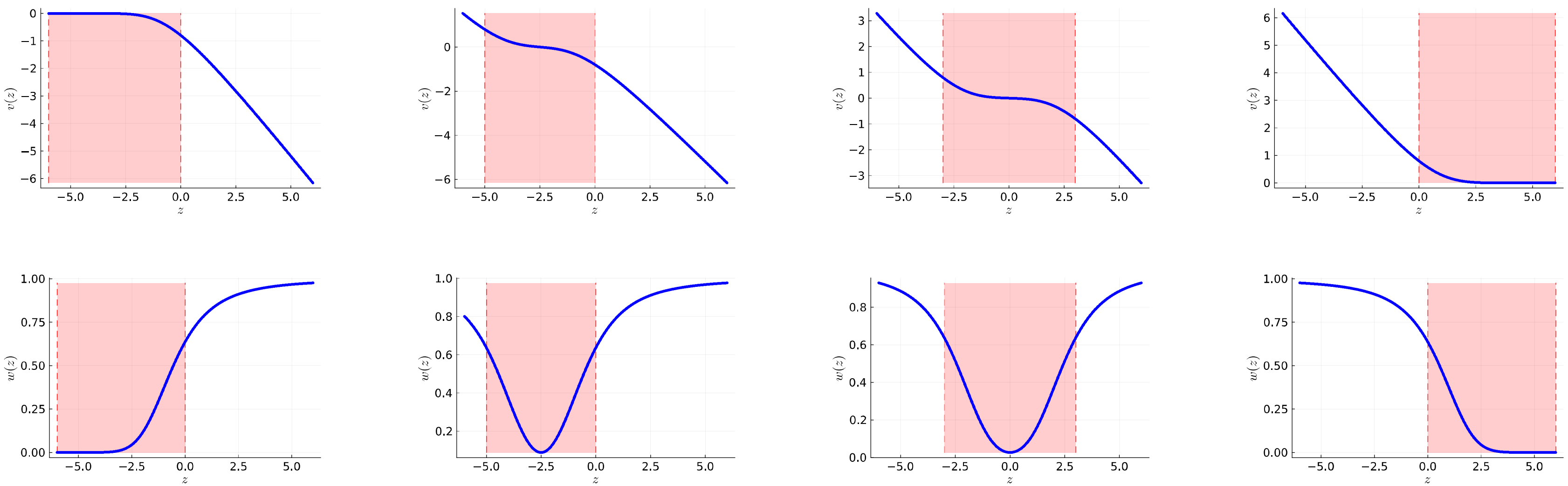
$$v_{l,u}(z) := \frac{\mathcal{N}(l-z) - \mathcal{N}(u-z)}{\Phi(u-z) - \Phi(l-z)}$$

$$w_{l,u}(z) := \frac{(u+z)\cdot\mathcal{N}(u-z) - (l+z)\cdot\mathcal{N}(l-z)}{\Phi(u-z) - \Phi(l-z)} + v_{l,u}(z)\cdot\left[2z + v_{l,u}(z)\right]$$

$\boldsymbol{x < 0}$ | $\boldsymbol{-5 \le x < 0}$ | $\boldsymbol{-3 \le x < 3}$ | $\boldsymbol{0 \le x}$

$\boldsymbol{v(t)}$

$\boldsymbol{w(t)}$



Overview

1. Distance Measures for Distributions
2. Approximating Marginals: Expectation Propagation
3. **Approximating Normalization Constants**

Normalization Constant

- **Normalization Constant**: *Given a factor graph with factors $f_1, \ldots, f_m$, each over a subset of n variables $X_1, X_2, \ldots, X_n$, the* normalization constant Z *is defined as the sum over all variables*

$$Z = \sum_{\{x_1\}} \cdots \sum_{\{x_n\}} f_1\left(\boldsymbol{x}_{\mathrm{ne}(f_1)}\right) \cdot f_2\left(\boldsymbol{x}_{\mathrm{ne}(f_2)}\right) \cdots f_m\left(\boldsymbol{x}_{\mathrm{ne}(f_m)}\right)$$

- **Inference in Factor Graphs**: In order to learn from data D we

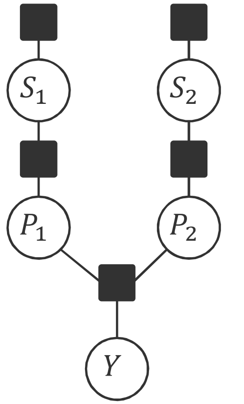
1. **Modelling**: Formulate a joint model $p(\boldsymbol{\theta}, D)$ of parameters $\boldsymbol{\theta} = \theta_1, \ldots, \theta_n$ and data D

$$Z = \sum_{\{\theta_1\}} \cdots \sum_{\{\theta_n\}} \sum_{\{D\}} p(\boldsymbol{\theta}, D) = 1$$

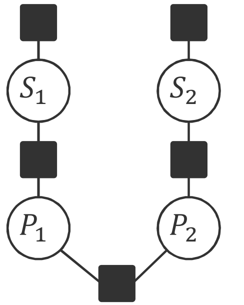
2. **Conditioning**: Remove the variables that represent data D from the factor graph

$$Z = \boxed{\sum_{\{\theta_1\}} \cdots \sum_{\{\theta_n\}} p(\boldsymbol{\theta}, D) = p(D)} \longleftarrow \text{Law of total probability}$$

   - The normalization constant is the probability of the data D and measures how good our probabilistic model explains the observed training set D (*model evidence*)!



$p(S_1, S_2, P_1, P_2, Y)$



$p(S_1, S_2, P_1, P_2, Y = 1)$

Normalization Constant via Message Passing

- The non-normalized marginal $p_{X_j}(\cdot)$ is defined as

$$p_{X_j}(x_j) = \sum_{\{x_1\}} \cdots \sum_{\{x_{j-1}\}} \sum_{\{x_{j+1}\}} \cdots \sum_{\{x_n\}} f_1\left(\boldsymbol{x}_{\mathrm{ne}(f_1)}\right) \cdot f_2\left(\boldsymbol{x}_{\mathrm{ne}(f_2)}\right) \cdots f_m\left(\boldsymbol{x}_{\mathrm{ne}(f_m)}\right)$$

$$m_{f_i \to X_j}(x_j) \cdot m_{X_j \to f_i}(x_j)$$

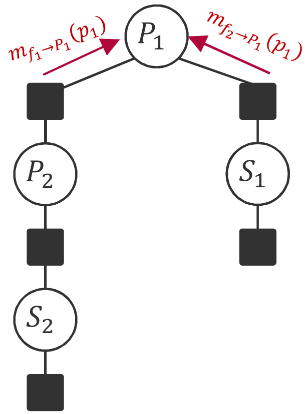
$$\prod_{i \in \mathrm{ne}(X_j)} m_{f_i \to X_j}(x_j)$$

$$Z = \sum_{\{x_j\}} p_{X_j}(x_j)$$

$$\sum_{\{\boldsymbol{x}_{\mathrm{ne}(f_i) \setminus \{j\}}\}} f\left(\boldsymbol{x}_{\mathrm{ne}(f_i)}\right) \prod_{k \in \mathrm{ne}(f_i) \setminus \{j\}} m_{X_k \to f_i}(x_k)$$

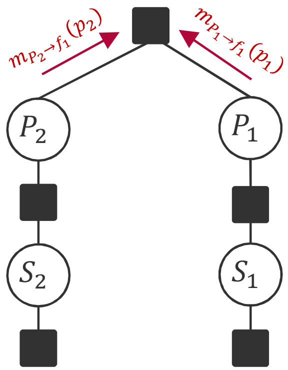
Variable node root

$$Z = \sum_{\{x_j\}} \prod_{i \in \mathrm{ne}(X_j)} m_{f_i \to X_j}(x_j)$$



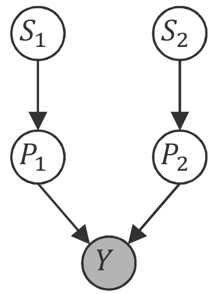
Factor node root

$$Z = \sum_{\{\boldsymbol{x}_{\mathrm{ne}(f_i)}\}} f\left(\boldsymbol{x}_{\mathrm{ne}(f_i)}\right) \prod_{k \in \mathrm{ne}(f_i)} m_{X_k \to f_i}(x_k)$$



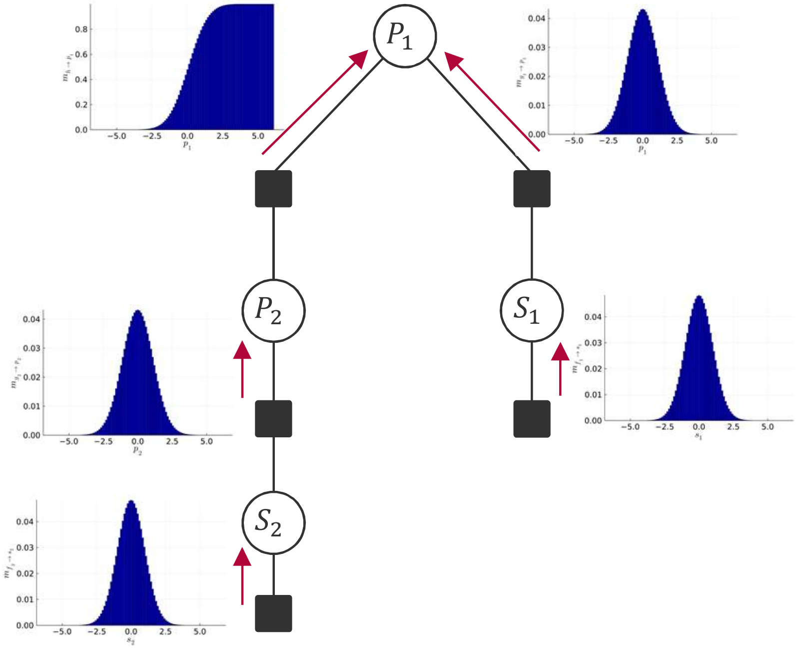
Normalization Constant via Message Passing: Example

$$\mathcal{N}(s_1;\mu_1,\sigma_1^2)\cdot\mathcal{N}(s_2;\mu_2,\sigma_2^2)\cdot\mathcal{N}(p_1;s_1,\beta^2)\cdot\mathcal{N}(p_2;s_2,\beta^2)\cdot\mathbb{I}(y(p_1-p_2)>0)$$



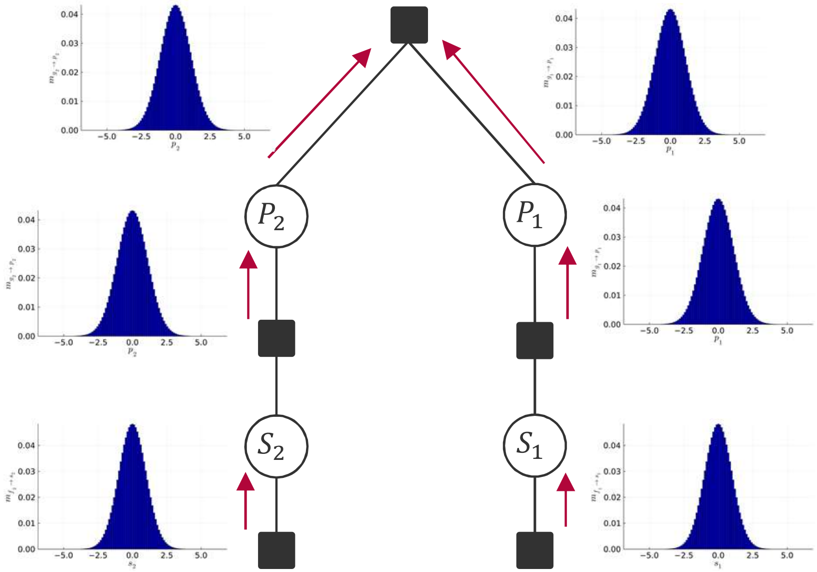
Variable node root

$$Z=\sum_{\{x_j\}}\prod_{i\in\mathrm{ne}(X_j)} m_{f_i\to X_j}(x_j)$$



Factor node root

$$Z=\sum_{\{\boldsymbol{x}_{\mathrm{ne}(f_i)}\}} f\left(\boldsymbol{x}_{\mathrm{ne}(f_i)}\right)\prod_{k\in\mathrm{ne}(f_i)} m_{X_k\to f_i}(x_k)$$

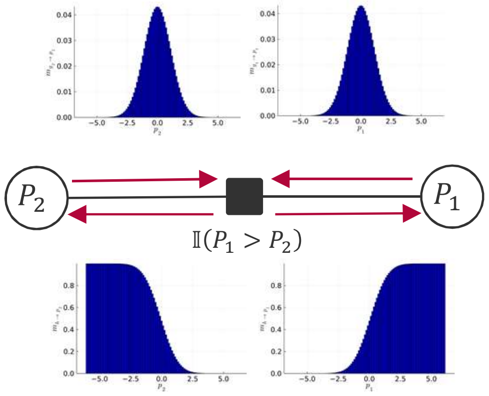


Introduction to Probabilistic Machine Learning

Unit 5 – Graphical Models: Approximate Inference

Normalization Constant via Approximate Message Passing

- **Challenge**: Computing the normalization constant Z requires
  1. Tracking the normalization constants for all messages
  2. Choosing a variable X_j or factor f_i as a root of the factor tree
- **Observation**: Tracking the normalization constant for all messages is not always possible when approximating messages!
- ***Theorem**: Given a factor tree with renormalized messages $\tilde{m}_{X_j \to f_i}(\cdot) = \beta_{j,i} \cdot m_{X_j \to f_i}(\cdot)$ and $\tilde{m}_{f_i \to X_j}(\cdot) = \alpha_{i,j} \cdot m_{f_i \to X_j}(\cdot)$ the normalization constant Z is*



$$Z = \left(\prod_{i=1}^{m} \boxed{Z_{f_i}}\right) \cdot \left(\prod_{j=1}^{n} \boxed{Z_{X_j}}\right)$$

Factor Normalization

$$Z_{f_i} = \frac{\sum_{\left\{\boldsymbol{x}_{\mathrm{ne}(f_i)}\right\}} f\left(\boldsymbol{x}_{\mathrm{ne}(f_i)}\right) \prod_{k \in \mathrm{ne}(f_i)} \tilde{m}_{X_k \to f_i}(x_k)}{\sum_{\left\{\boldsymbol{x}_{\mathrm{ne}(f_i)}\right\}} \prod_{k \in \mathrm{ne}(f_i)} \tilde{m}_{f_i \to X_k}(x_k) \cdot \tilde{m}_{X_k \to f_i}(x_k)}$$

Variable Normalization

$$Z_{X_j} = \sum_{\{x_j\}} \prod_{i \in \mathrm{ne}(X_j)} \tilde{m}_{f_i \to X_j}(x_j)$$

Proof of Normalization Theorem

- Simplifying the variable normalization Z_{X_j} and factor normalization Z_{f_i}

Variable normalization

$\alpha_{i,j}\cdot m_{f_i\to X_j}(x_j)$

$$Z_{X_j} = \sum_{\{x_j\}} \prod_{i\in \mathrm{ne}(X_j)} \tilde{m}_{f_i\to X_j}(x_j)$$

Z

$$= \left(\prod_{i\in \mathrm{ne}(X_j)} \alpha_{i,j}\right)\cdot\left(\sum_{\{x_j\}} \prod_{i\in \mathrm{ne}(X_j)} m_{f_i\to X_j}(x_j)\right)$$

$$= \left(\prod_{i\in \mathrm{ne}(X_j)} \alpha_{i,j}\right)\cdot Z$$

Factor normalization

$\beta_{k,i}\cdot m_{X_k\to f_i}(x_k)$

$$Z_{f_i} = \frac{\sum_{\{\mathbf{x}_{\mathrm{ne}(f_i)}\}} f\left(\mathbf{x}_{\mathrm{ne}(f_i)}\right) \prod_{k\in\mathrm{ne}(f_i)} \tilde{m}_{X_k\to f_i}(x_k)}{\sum_{\{\mathbf{x}_{\mathrm{ne}(f_i)}\}} \prod_{k\in\mathrm{ne}(f_i)} \tilde{m}_{f_i\to X_k}(x_k)\cdot \tilde{m}_{X_k\to f_i}(x_k)}$$

Z

$$= \frac{\cancel{\left(\prod_{k\in\mathrm{ne}(f_i)} \beta_{k,i}\right)}\cdot\left(\sum_{\{\mathbf{x}_{\mathrm{ne}(f_i)}\}} f\left(\mathbf{x}_{\mathrm{ne}(f_i)}\right) \prod_{k\in\mathrm{ne}(f_i)} m_{X_k\to f_i}(x_k)\right)}{\cancel{\left(\prod_{k\in\mathrm{ne}(f_i)} \beta_{k,i}\right)}\cdot\left(\prod_{k\in\mathrm{ne}(f_i)} \alpha_{i,k}\right)\cdot\left(\sum_{\{\mathbf{x}_{\mathrm{ne}(f_i)}\}} \prod_{k\in\mathrm{ne}(f_i)} m_{f_i\to X_k}(x_k)\cdot m_{X_k\to f_i}(x_k)\right)}$$

$$= \frac{Z}{\left(\prod_{k\in\mathrm{ne}(f_i)} \alpha_{i,k}\right)\cdot Z^{|\mathrm{ne}(f_i)|}}$$

$$\prod_{k\in\mathrm{ne}(f_i)} \sum_{\{x_k\}} m_{f_i\to X_k}(x_k)\cdot m_{X_k\to f_i}(x_k)$$

$$\left(\prod_{j=1}^{n} Z_{X_j}\right)\cdot\left(\prod_{i=1}^{m} Z_{f_i}\right) = \left(\prod_{j=1}^{n}\prod_{i\in\mathrm{ne}(X_j)} \alpha_{i,j}\right)\cdot Z^n\cdot\frac{Z^m}{\left(\prod_{i=1}^{m}\prod_{k\in\mathrm{ne}(f_i)} \alpha_{i,k}\right)\cdot\prod_{i=1}^{m} Z^{|\mathrm{ne}(f_i)|}}$$

identical

Z^{n+m-1} because the factor graph is a tree with $n+m$ nodes!

Summary

1. **Distance Measures for Distributions**
   - α-divergences are a general class of distance measures between distributions
   - For $\alpha = 1$, the α-divergence becomes the Kullback-Leibler divergence where the minimizer for Gaussian approximating distributions matches mean and variance
   - Minimizers of α-divergences range from mode-seeking to support-seeking
2. **Approximate Message Passing and Expectation Propagation**
   - Approximations will always be done on the marginals, **not** the messages
   - When the Kullback-Leibler divergence is used as distance, all moments get preserved
   - In case of doubly-truncated Gaussians, the moments are closed form
3. **Approximating Normalization Constants**
   - If factor graphs represent joint probabilities of data and parameters with data variables only, the normalization constant equals the probability of data (under the model)
   - If messages can be explicitly represented and normalized, efficient updates can be done starting at any factor or variable
   - There is a general distributed algorithm for approximating the normalization constant

Hasso Plattner Institut

Digital Engineering · Universität Potsdam

See you next week!