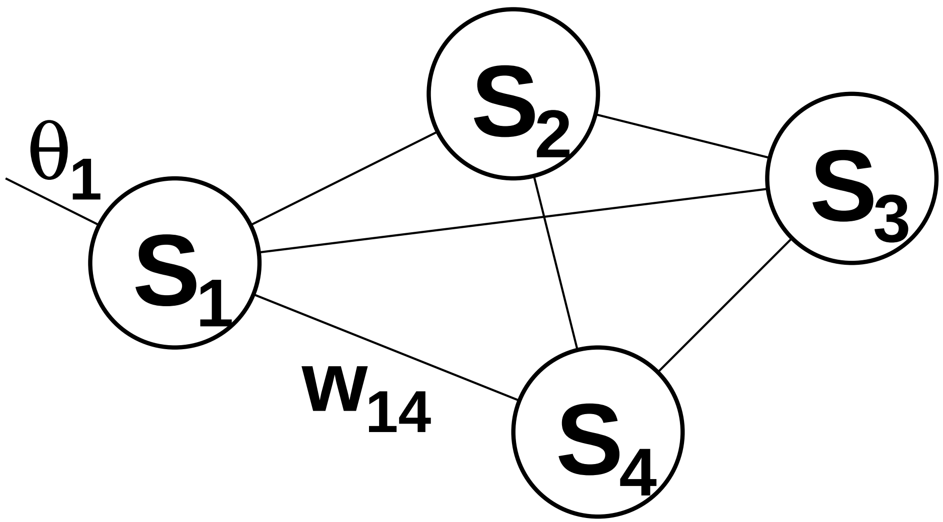
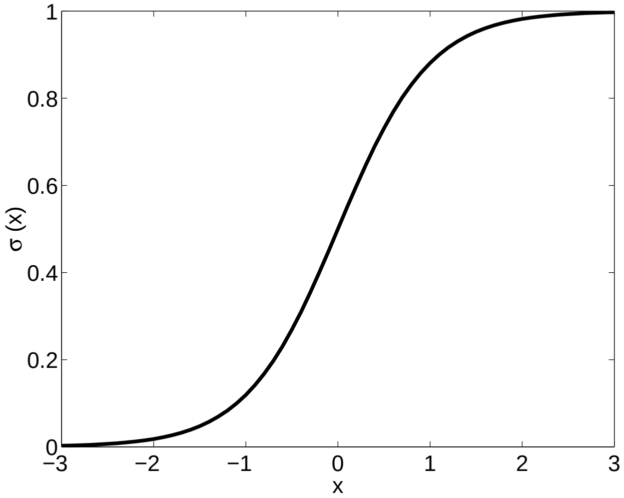
Variational Methods and Bounds

- Introduction to Boltzmann machines
- Simple approximations
- Introduction to variational methods
- The search for bounds
- Summary

The Boltzmann Machine



$$p\left(s_i = 1\right) = \sigma\left(\theta_i + \sum_j w_{ij} s_j\right)$$



The Boltzmann Machine

The *energy* of the Boltzmann machine for a certain *state* is

$$-E\left(\vec{s}\right) = \sum_{i} \theta_i s_i + \frac{1}{2} \sum_{ij} w_{ij} s_i s_j$$

The *probability* to find the BM in state $\vec{s}$:

$$p\left(\vec{s}\right) = \frac{1}{Z} \exp\left(-E\left(\vec{s}\right)\right)$$

The normalizing constant is the *partition function*

$$Z = \sum_{\text{all } \vec{s}} \exp\left(-E\left(\vec{s}\right)\right)$$

Derived Quantities

$$Z = \sum_{\text{all } \vec{s}} \exp\left(-E\left(\vec{s}\right)\right)$$

Means and *correlations*

$$\langle s_i \rangle = \frac{\partial}{\partial \theta_i} \log Z = \frac{1}{Z} \frac{\partial Z}{\partial \theta_i} = \frac{1}{Z} \sum_{\text{all } \vec{s}} \frac{\partial}{\partial \theta_i} \exp\left(-E\left(\vec{s}\right)\right)$$

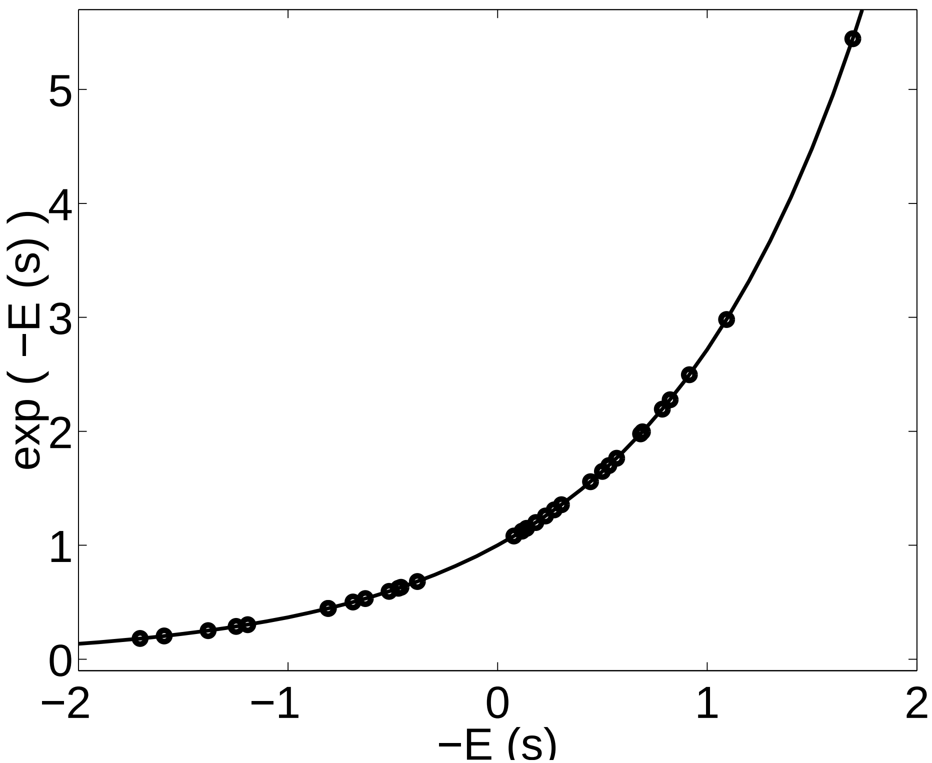
$$= \frac{1}{Z} \sum_{\text{all } \vec{s}} \exp\left(-E\left(\vec{s}\right)\right) s_i = \sum_{\text{all } \vec{s}} p\left(\vec{s}\right) s_i$$

$$\langle s_i s_j \rangle = \frac{\partial}{\partial w_{ij}} \log Z$$

Computing the Partition Function

$$Z = \sum_{\text{all } \vec{s}} \exp\left(-E\left(\vec{s}\,\right)\right)$$

State	exp(Energy)
— — — — —	3.979020
— — — — +	1.586260
— — — + —	3.480777
— — — + +	1.994746
— — + — —	0.566925
— — + — +	0.505540
— — + + —	1.466552
— — + + +	1.879924
— + — — —	1.604714
— + — — +	0.833811
— + — + —	2.247136
— + — + +	1.678465
— + + — —	0.795865
— + + — +	0.924999
— + + + —	3.295664
— + + + +	5.506271
+ — — — —	0.642141
+ — — — +	0.335461
+ — — + —	0.606811
+ — — + +	0.455699
+ — + — —	0.157148
+ — + — +	[illegible]
+ — + + —	0.439139
+ — + + +	0.737664



A Simple Approximation

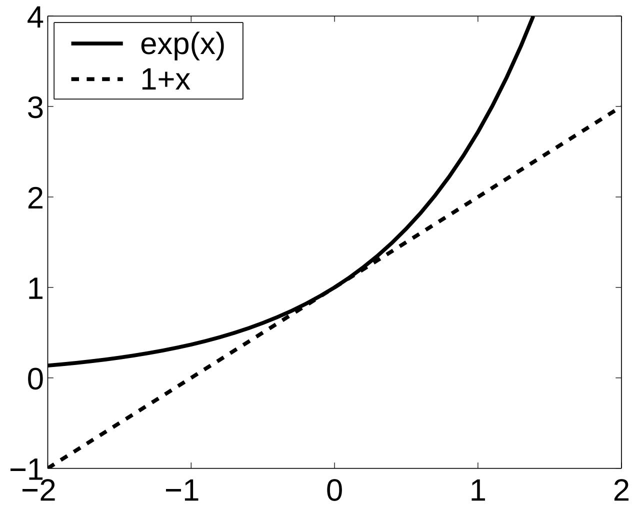
Approximate $\exp(x)$ by $1+x$

Thus

$$Z = \sum_{\text{all } \vec{s}} \exp\left(-E\left(\vec{s}\right)\right) \approx \sum_{\text{all } \vec{s}} \left(1 - E\left(\vec{s}\right)\right)$$

$$= 2^N - \sum_{\text{all } \vec{s}} E\left(\vec{s}\right) = 2^N$$

which is a quite poor approximation.



Variational Approaches

For all μ we know

$$\exp\left(x\right) \geq \mathrm{e}^{\mu} + \mathrm{e}^{\mu}\left(x - \mu\right) = \mathrm{e}^{\mu}\left(1 + x - \mu\right)$$

Thus

$$\forall_{\mu}\; Z = \sum_{\text{all } \vec{s}} \exp\left(-E\left(\vec{s}\,\right)\right) \geq \sum_{\text{all } \vec{s}} \mathrm{e}^{\mu}\left(1 - E\left(\vec{s}\,\right) - \mu\right)$$
$$= 2^{N}\mathrm{e}^{\mu}\left(1 - \mu\right) = B\left(\mu\right)$$

The best approximation is the maximum of $B\left(\mu\right)$.

$$\frac{\partial}{\partial\mu}B\left(\mu\right) = -\mu\mathrm{e}^{\mu} = 0 \;\Rightarrow\; \mu = 0$$

and again we find

$$Z \geq 2^{N}$$

Variational Approaches

$$\exp(x) \geq \mathrm{e}^{\mu}(1 + x - \mu)$$

Thus

$$\forall_{\mu(\vec{s})}\ Z = \sum_{\text{all } \vec{s}} \exp(-E(\vec{s})) \geq \sum_{\text{all } \vec{s}} \mathrm{e}^{\mu(\vec{s})}(1 - E(\vec{s}) - \mu(\vec{s}))$$

and we choose

$$\mu(\vec{s}) = \mu + \sum_i h_i s_i$$

which can be optimised with respect to μ and h_i.

Other Bounds

The Kullback-Leibler divergence is a bound:

$$K\left(q,p\right)=\sum_{\text{all }\vec{s}} q\left(\vec{s}\right)\log\frac{q\left(\vec{s}\right)}{p\left(\vec{s}\right)}\geq 0$$

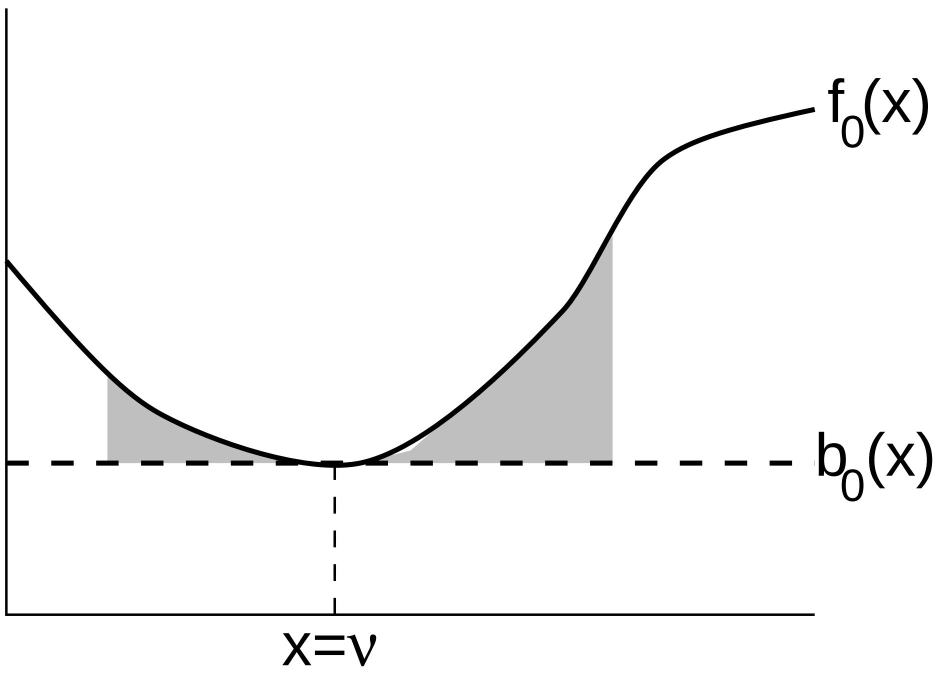
But we have also

$$\log x\leq\frac{x}{\mu}-1+\log\mu$$

and

$$\tanh x\leq\frac{1}{2\sqrt{2}}\left(x-\mu\right)^2+\left(1-\tanh^2\mu\right)\left(x-\mu\right)+\tanh\mu$$

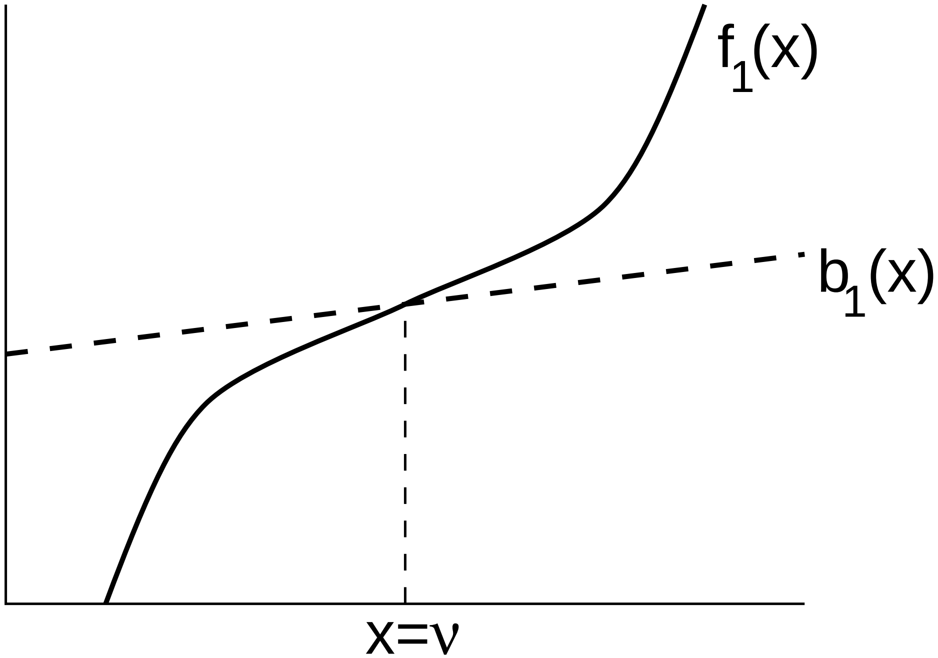
Improving Bounds



$$f_0(x) \geq b_0(x)$$

$$f_1(x) = \int_{x'=\nu}^{x'=x} f_0(x')\, \mathrm{d}x' \quad \begin{array}{cc} \geq & (x \geq \nu) \\ \leq & (x \leq \nu) \end{array} \quad \int_{x'=\nu}^{x'=x} b_0(x')\, \mathrm{d}x' = b_1(x)$$

Improving Bounds



$$f_1(x) \lesseqgtr b_1(x)$$

$$f_2(x) = \int_{x'=\nu}^{x'=x} f_1(x')\ \mathrm{d}x' \qquad \geq \qquad \int_{x'=\nu}^{x'=x} b_1(x')\ \mathrm{d}x' = b_2(x)$$

Improving Bounds

Given $f_0\,(x) \geq b_0\,(x)$ we can derive

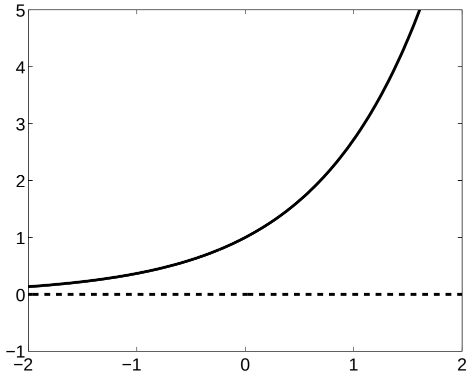
- $\boxed{f_1 = \int f_0}$ and $\boxed{b_1 = \int b_0}$ with $\boxed{f_1\,(\nu) = b_1\,(\nu)}$ for some ν
- $\boxed{f_2 = \int f_1}$ and $\boxed{b_2 = \int b_1}$ with $\boxed{f_2\,(\nu) = b_2\,(\nu)}$ for that ν

Then we know

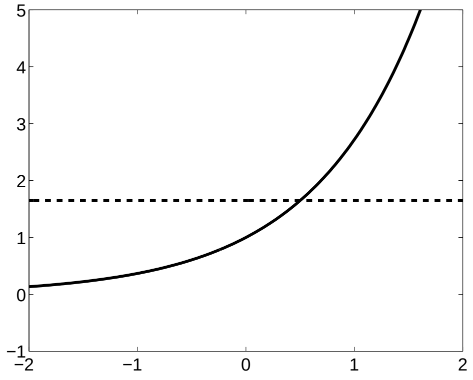
- $\forall_{x,\nu} \quad f_2\,(x) \geq b_2\,(x)$

Example: Exponential function

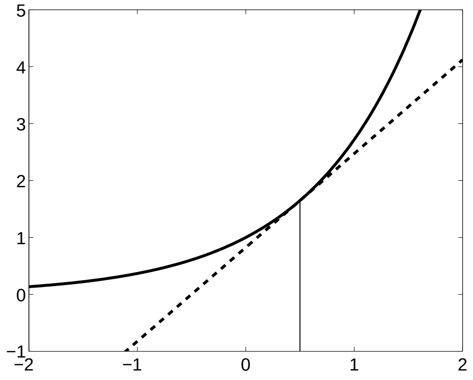
$$
\begin{aligned}
& f_0(x) = \mathrm{e}^x &\geq& \quad 0 = b_0(x) \\
& f_1(x) = \mathrm{e}^x &\lesseqgtr& \quad \mathrm{e}^\nu = b_1(x) \\
\forall_{x,\nu} \quad & f_2(x) = \mathrm{e}^x &\geq& \quad \mathrm{e}^\nu\left(1 + x - \nu\right) = b_2(x)
\end{aligned}
$$



$f_0\left(x\right) \geq b_0\left(x\right)$



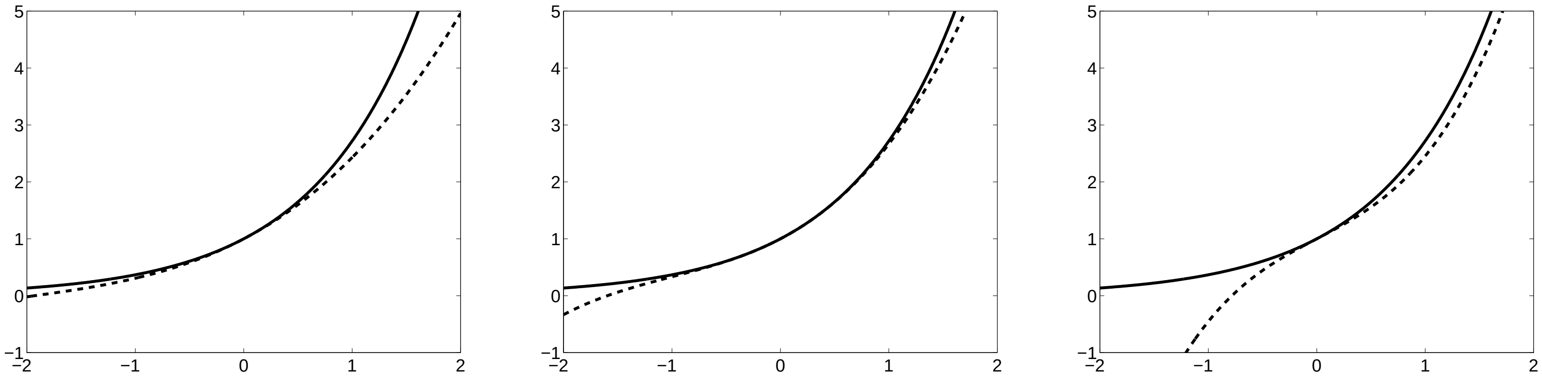
$f_1\left(x\right) \lesseqgtr b_1\left(x\right)$



$f_2\left(x\right) \geq b_2\left(x\right)$

Example: Exponential function

$$\forall_{x,\mu,\lambda} \quad \mathrm{e}^{x} \geq \mathrm{e}^{\mu}\left\{1 + x - \mu + \mathrm{e}^{\lambda}\left(\frac{1-\lambda}{2}\,(x-\mu)^2 + \frac{1}{6}\,(x-\mu)^3\right)\right\}$$



$\mu = 0$ $\quad$ $\mu = 0$ $\quad$ $\mu = 0$

$\lambda = -1$ $\quad$ $\lambda = 0$ $\quad$ $\lambda = +1$

Summary

- You have a function that is intractable (e.g. $\sum_{\text{all } \vec{s}} \exp\left(-E\left(\vec{s}\right)\right)$)
- You can derive a (large) class of bounding functions ($f\left(x\right) \geq b\left(x, \mu\right)$)
- These functions are parametrized by μ, the *variational parameters*
- The larger this class is, the better your estimate
- Optimize the class of bounding function with respect to μ to find the tightest bound

Take Home Message

1. I have seen a way to do approximate computations for a Boltzmann machine. But $p\,(\text{I will use a BM}) = 0$.
2. I have an intuition of what can be done with variational methods and the Boltzmann machine was just one of the applications.