Approximations and computational tools

Václav Šmídl,

Winter school of machine learning,
Czech Technical University
vasek.smidl@gmail.com

January 21, 2020

Overview

Lecture 1: How to be a Bayesian

Lecture 2: Approximations and computational tools

Lecture 3: Application to Deep Active Learning

Overview

Lecture 1: How to be a Bayesian

Lecture 2: Approximations and computational tools

Lecture 3: Application to Deep Active Learning

Lecture 2:

Motivating Example

- Bayesian average
- analytical solution

Approximations:

- Variational methods
- Monte Carlo

Toy example

We observe two numbers d_1 and d_2 and assume that they are noisy observations of a constant unknown quantity, m,

$$d_1 = m + e_1,$$
$$d_2 = m + e_2,$$

- Noise realizations are independent, $p(e_1, e_2) = p(e_1)p(e_2)$, zero-mean normal distributed

$$p(e_i|s) = \mathcal{N}(0, s),$$

- Compute posterior $p(m, s|d_1, d_2)$

$$p(m, s|d_1, d_2) \quad = \quad \frac{p(m, s, d_1, d_2)}{p(d_1, d_2)}$$

Toy example

We observe two numbers d_1 and d_2 and assume that they are noisy observations of a constant unknown quantity, m,

$$d_1 = m + e_1,$$
$$d_2 = m + e_2,$$

- Noise realizations are independent, $p(e_1, e_2) = p(e_1)p(e_2)$, zero-mean normal distributed

$$p(e_i|s) = \mathcal{N}(0, s), \implies p(e_i + m|m, s) = \mathcal{N}(m, s)$$

- Compute posterior $p(m, s|d_1, d_2)$

$$p(m, s|d_1, d_2) \quad = \quad \frac{p(m, s, d_1, d_2)}{p(d_1, d_2)}$$

Toy example

We observe two numbers d_1 and d_2 and assume that they are noisy observations of a constant unknown quantity, m,

$$d_1 = m + e_1,$$
$$d_2 = m + e_2,$$

- Noise realizations are independent, $p(e_1, e_2) = p(e_1)p(e_2)$, zero-mean normal distributed

$$p(e_i|s) = \mathcal{N}(0, s),$$

- Compute posterior $p(m, s|d_1, d_2)$

$$\begin{aligned} p(m, s|d_1, d_2) &= \frac{p(m, s, d_1, d_2)}{p(d_1, d_2)} \\ &\propto p(d_1|m, s)p(d_2|m, s)p(m|s)p(s) \end{aligned}$$

Toy example

We observe two numbers d_1 and d_2 and assume that they are noisy observations of a constant unknown quantity, m,

$$d_1 = m + e_1,$$
$$d_2 = m + e_2,$$

- Noise realizations are independent, $p(e_1, e_2) = p(e_1)p(e_2)$, zero-mean normal distributed

$$p(e_i|s) = \mathcal{N}(0, s),$$

- Compute posterior $p(m, s|d_1, d_2)$

$$\begin{aligned} p(m, s|d_1, d_2) &= \frac{p(m, s, d_1, d_2)}{p(d_1, d_2)} \\ &\propto p(d_1|m, s)p(d_2|m, s)p(m|s)p(s) \end{aligned}$$

- Priors

$$p(m|s) = \mathcal{N}(0, \tau s), p(s) = \mathcal{G}(\alpha, \beta),$$

Toy example

$$
\begin{aligned}
p(m, s|d_1, d_2) &\propto p(d_1|m, s)p(d_2|m, s)p(m|s)p(s) \\
&\propto \quad s^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(d_1 - m)^2/s\right) \\
&\qquad s^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(d_2 - m)^2/s\right) \\
&\qquad s^{-\frac{1}{2}}\tau^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(0 - m)^2/(\tau s)\right) \\
&\qquad s^{-\alpha-1} \exp(-\beta/s)
\end{aligned}
$$

Toy example

$$
\begin{aligned}
p(m, s|d_1, d_2) &\propto p(d_1|m, s)p(d_2|m, s)p(m|s)p(s) \\
&\propto s^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(d_1 - m)^2/s\right) \\
&\quad s^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(d_2 - m)^2/s\right) \\
&\quad s^{-\frac{1}{2}}\tau^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(0 - m)^2/(\tau s)\right) \\
&\quad s^{-\alpha-1} \exp(-\beta/s) \\
&\propto s^{-(\frac{3}{2}+\alpha)-1} \exp\left(-\frac{1}{2s}(m^2 - 2d_1 m + d_1^2 + m^2 - 2d_2 m + d_2^2 + m^2/\tau + 2\beta\right),
\end{aligned}
$$

Toy example

$$
\begin{aligned}
p(m, s|d_1, d_2) &\propto p(d_1|m, s)p(d_2|m, s)p(m|s)p(s) \\
&\propto s^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(d_1 - m)^2/s\right) \\
&\quad s^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(d_2 - m)^2/s\right) \\
&\quad s^{-\frac{1}{2}}\tau^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(0 - m)^2/(\tau s)\right) \\
&\quad s^{-\alpha-1} \exp(-\beta/s) \\
&\propto s^{-(\frac{3}{2}+\alpha)-1} \exp\left(-\frac{1}{2s}(m^2 - 2d_1 m + d_1^2 + m^2 - 2d_2 m + d_2^2 + m^2/\tau + 2\beta\right), \\
&\propto s^{-(\frac{3}{2}+\alpha)-1} \exp\left(-\frac{1}{2\sigma_m}(m - \mu)^2\right) \exp\left(-\frac{1}{2s}\left(d_1^2 + d_2^2 - \frac{(d_1 + d_2)^2}{2 + 1/\tau} + 2\beta\right)\right)
\end{aligned}
$$

Toy example

$$
\begin{aligned}
p(m,s|d_1,d_2) &\propto p(d_1|m,s)p(d_2|m,s)p(m|s)p(s) \\
&\propto s^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(d_1-m)^2/s\right) \\
&\quad s^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(d_2-m)^2/s\right) \\
&\quad s^{-\frac{1}{2}}\tau^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(0-m)^2/(\tau s)\right) \\
&\quad s^{-\alpha-1}\exp(-\beta/s) \\
&\propto s^{-(\frac{3}{2}+\alpha)-1} \exp\left(-\frac{1}{2s}(m^2-2d_1m+d_1^2+m^2-2d_2m+d_2^2+m^2/\tau+2\beta\right), \\
&\propto s^{-(\frac{3}{2}+\alpha)-1} \exp\left(-\frac{1}{2\sigma_m}(m-\mu)^2\right) \exp\left(-\frac{1}{2s}\left(d_1^2+d_2^2-\frac{(d_1+d_2)^2}{2+1/\tau}+2\beta\right)\right) \\
&= \mathcal{N}(\mu,\sigma_m)\mathcal{G}(\alpha+1,\beta_s) \\
\mu &= \frac{d_1+d_2}{2+1/\tau}, \quad \sigma_m = \frac{s}{2+1/\tau}, \quad \beta_s = \frac{1}{2}\left(d_1^2+d_2^2-\frac{(d_1+d_2)^2}{2+1/\tau}\right)
\end{aligned}
$$

Toy example

$$
\begin{aligned}
p(m,s|d_1,d_2) &\propto p(d_1|m,s)p(d_2|m,s)p(m|s)p(s) \\
&\propto s^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(d_1-m)^2/s\right) \\
&\quad s^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(d_2-m)^2/s\right) \\
&\quad s^{-\frac{1}{2}}\tau^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(0-m)^2/(\tau s)\right) \\
&\quad s^{-\alpha-1}\exp(-\beta/s) \\
&\propto s^{-(\frac{3}{2}+\alpha)-1} \exp\left(-\frac{1}{2s}(m^2-2d_1m+d_1^2+m^2-2d_2m+d_2^2+m^2/\tau+2\beta\right), \\
&\propto s^{-(\frac{3}{2}+\alpha)-1} \exp\left(-\frac{1}{2\sigma_m}(m-\mu)^2\right) \exp\left(-\frac{1}{2s}\left(d_1^2+d_2^2-\frac{(d_1+d_2)^2}{2+1/\tau}+2\beta\right)\right) \\
&= \mathcal{N}(\mu,\sigma_m)\mathcal{G}(\alpha+1,\beta_s) \\
\mu &= \frac{d_1+d_2}{2+1/\tau}, \quad \sigma_m = \frac{s}{2+1/\tau}, \quad \beta_s = \frac{1}{2}\left(d_1^2+d_2^2-\frac{(d_1+d_2)^2}{2+1/\tau}\right)
\end{aligned}
$$

For $\tau \to \infty$, posterior mean approaches arithmetic mean.

- Nice results. Nice choice of priors (conjugate).

Tough toy example

We observe two numbers d_1 and d_2 and assume that they are noisy observations of a constant unknown quantity, m,

$$d_1 = m + e_1,$$
$$d_2 = m + e_2,$$

- Noise realizations are independent, $p(e_1, e_2) = p(e_1)p(e_2)$, zero-mean normal distributed

$$p(e_i|s) = \mathcal{N}(0, s), \Longrightarrow p(x_i|m, s) = \mathcal{N}(m, s)$$

- Compute posterior $p(m, s|d_1, d_2)$

$$\begin{aligned} p(m, s|d_1, d_2) &= \frac{p(m, s, d_1, d_2)}{p(d_1, d_2)} \\ &\propto p(d_1|m, s)p(d_2|m, s)p(m)p(s) \end{aligned}$$

- Priors

$$p(m) = \mathcal{N}(0, \tau), \qquad p(s) = \mathcal{G}(\alpha, \beta),$$

Tough toy example

We observe two numbers d_1 and d_2 and assume that they are noisy observations of a constant unknown quantity, m,

$$d_1 = m + e_1,$$
$$d_2 = m + e_2,$$

- Noise realizations are independent, $p(e_1, e_2) = p(e_1)p(e_2)$, zero-mean normal distributed

$$p(e_i|s) = \mathcal{N}(0, s), \Longrightarrow p(x_i|m, s) = \mathcal{N}(m, s)$$

- Compute posterior $p(m, s|d_1, d_2)$

$$\begin{aligned} p(m, s|d_1, d_2) &= \frac{p(m, s, d_1, d_2)}{p(d_1, d_2)} \\ &\propto p(d_1|m, s)p(d_2|m, s)p(m)p(s) \end{aligned}$$

- Priors

$$p(m) = \mathcal{N}(0, \tau), \qquad p(s) = \mathcal{G}(\alpha, \beta),$$
$$\cancel{p(m|s) = \mathcal{N}(0, \tau s),}$$

Tough toy example

$$
\begin{aligned}
p(m, s|d_1, d_2) &\propto p(d_1|m, s)p(d_2|m, s)p(m|s)p(s) \\
&\propto s^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(d_1 - m)^2/s\right) \\
&\quad s^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(d_2 - m)^2/s\right) \\
&\quad \tau^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(0 - m)^2/\tau\right) \\
&\quad s^{-\alpha-1} \exp(-\beta/s)
\end{aligned}
$$

Tough toy example

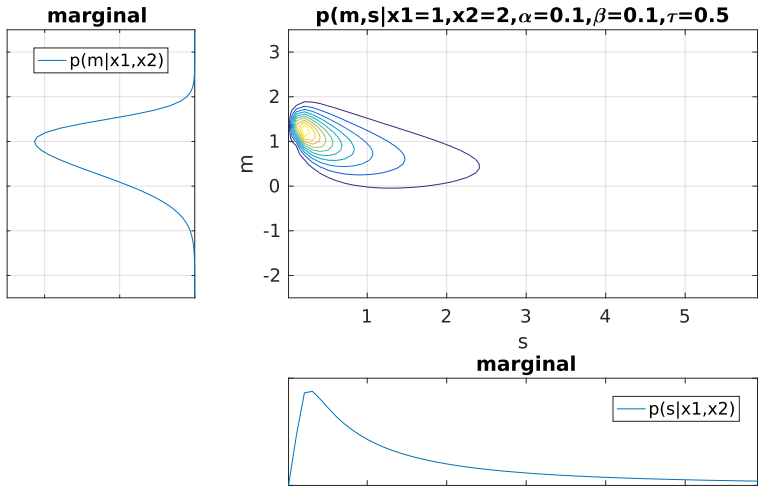
$$
\begin{aligned}
p(m, s|d_1, d_2) &\propto p(d_1|m, s)p(d_2|m, s)p(m|s)p(s) \\
&\propto s^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(d_1 - m)^2/s\right) \\
&\quad s^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(d_2 - m)^2/s\right) \\
&\quad \tau^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(0 - m)^2/\tau\right) \\
&\quad s^{-\alpha-1} \exp(-\beta/s)
\end{aligned}
$$

$$
\propto s^{-(\frac{3}{2}+\alpha)-1} \exp\left(-\frac{1}{2s}(m^2 - 2d_1 m + d_1^2 + m^2 - 2d_2 m + d_2^2 + 2\beta\right) \exp\left(-\frac{1}{2}\frac{m^2}{\tau}\right),
$$

Tough toy example

$$p(m,s|d_1,d_2) \propto p(d_1|m,s)p(d_2|m,s)p(m|s)p(s)$$

$$\begin{aligned}
\propto \quad & s^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(d_1-m)^2/s\right) \\
& s^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(d_2-m)^2/s\right) \\
& \tau^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(0-m)^2/\tau\right) \\
& s^{-\alpha-1} \exp(-\beta/s)
\end{aligned}$$



$$\propto \quad s^{-(\frac{3}{2}+\alpha)-1} \exp\left(-\frac{1}{2s}(m^2 - 2d_1 m + d_1^2 + m^2 - 2d_2 m + d_2^2 + 2\beta)\right) \exp\left(-\frac{1}{2}\frac{m^2}{\tau}\right),$$

Approximations

The need to marginalize and normalize posterior $p(x|d)$ via $p(d) = \int p(d, x)dx$.

Simple:

- Laplace: Taylor expansion of $\log(p(x, d))$ at $\hat{x} = \arg\max p(x, d)$. Taking quadratic terms yields $p(x) \approx \mathcal{N}(\hat{x}, \Sigma)$.

Variational:

- Choose approximating form $q(x)$
- minimize divergence $D(q(x)||p(x|d))$

Monte Carlo:

- Approximate posterior by Dirac mixture $p(x|d) \approx \frac{1}{n} \sum_{i=1}^{n} \delta(x - x^{(i)})$,
- Importance sampling
- Monte Carlo Markov Chain

Divergence

Divergence $D(p||q)$:

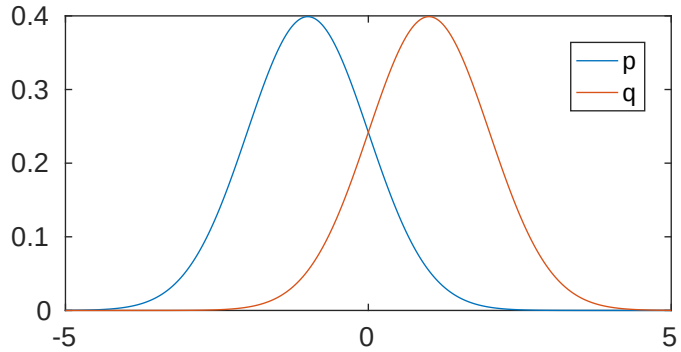
1. $D(q||p) \geq 0$, for all p, q,
2. $D(p||q) = 0$, iff $p = q$.

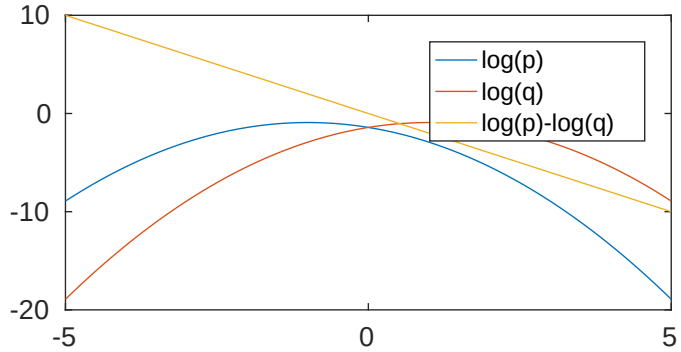
General class f-divergence ($f(1) = 0$)

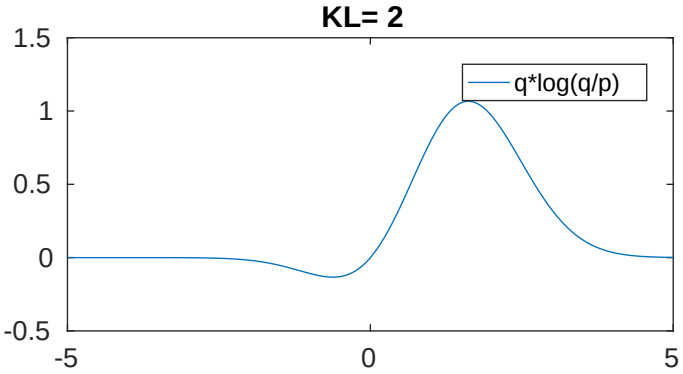
$$D_f(q||p) = \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx.$$

Special case $f(\cdot) = \log(\cdot)$:

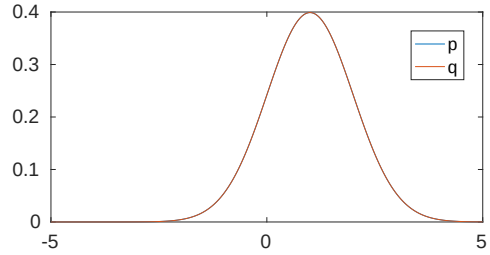
$$KL(q||p) = \int q(x) \log \frac{p(x)}{q(x)}$$

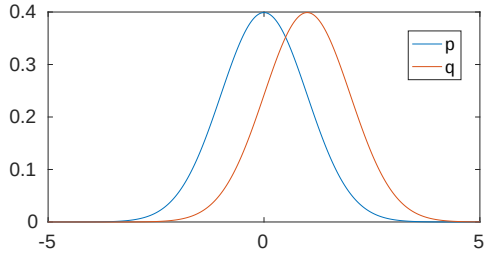


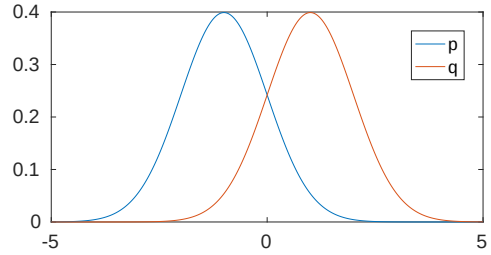


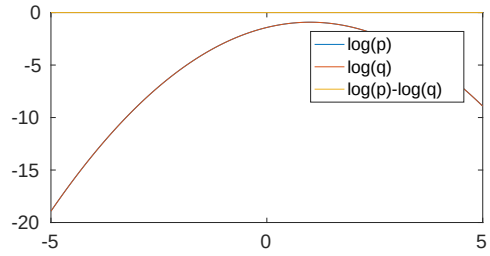


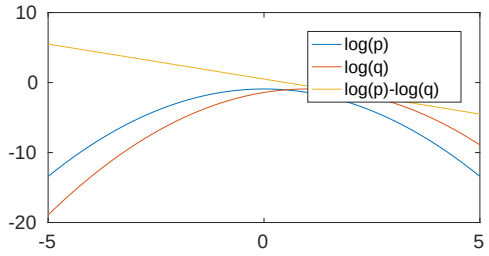
Kullback-Leibler

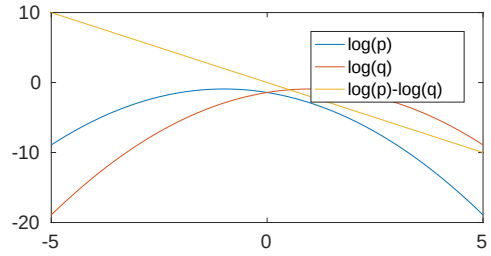


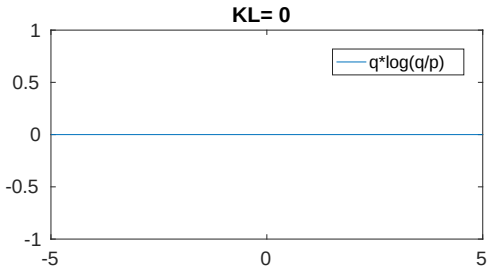


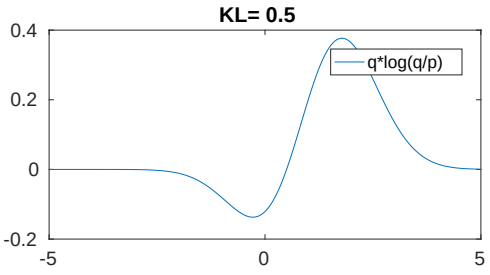


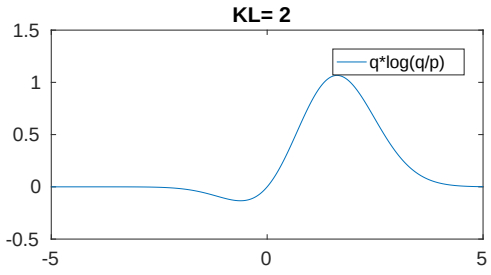




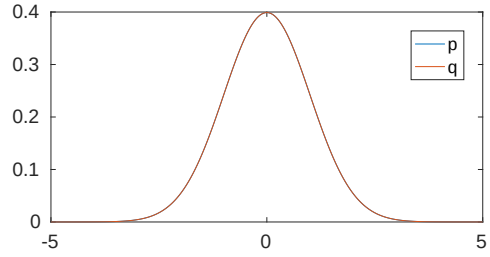


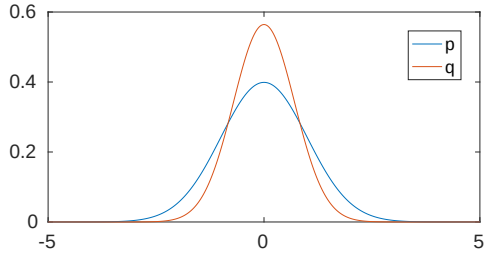


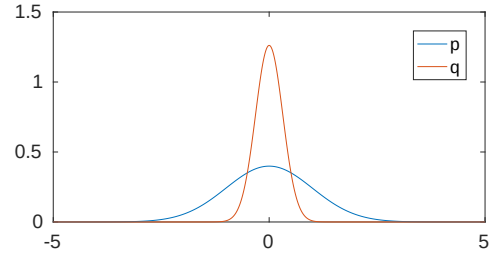


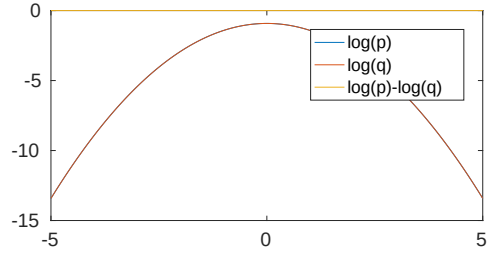


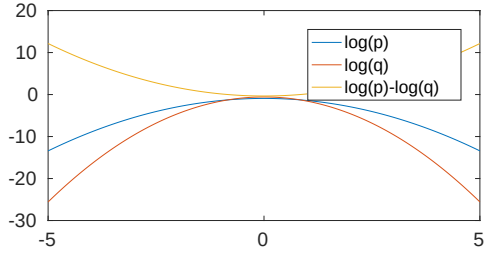
Kullback-Leibler

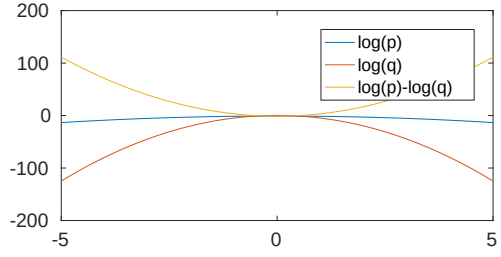


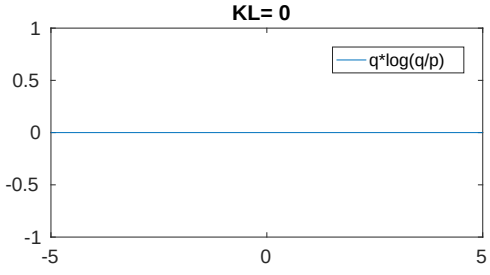


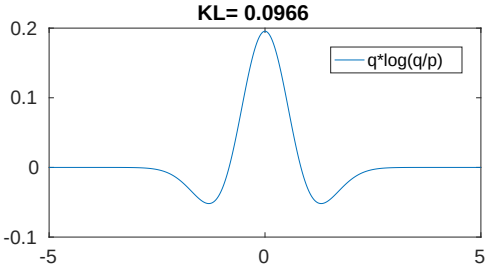


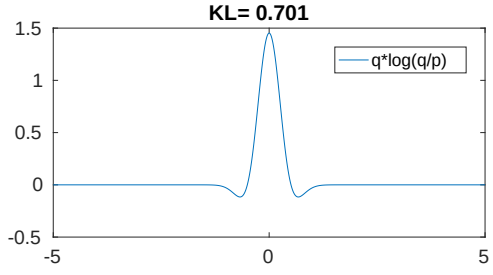












Variational Bayes

Is a divergence minimization technique with

$$q^* = \arg\min_q \mathrm{KL}(q||p) = \arg\min_q \mathrm{E}_q \left(\log \frac{q}{p} \right)$$

Variational Bayes:

1. conditional independence: $q(x_1, x_2) \equiv q(x_1)q(x_2)$.
2. conditions of extremum

$$\frac{\partial \mathrm{KL}}{\partial q_i} = 0, \quad \forall i.$$

ELBO:

1. parametric form of posterior $q(x|\theta)$,
2. use SGD to minimize KL

$$\theta^{new} = \theta^{old} - \eta \nabla_\theta \mathrm{KL}$$

ELBO: Evidence Lower BOund approach

Minimization of KL when

$$p(x|D) \propto p(D|x)p(x)$$

and approximator

$$q(x|\theta)$$

Kullback Leibler

$$\begin{aligned} KL = E_{q(x|\theta)}\left(\log \frac{q(x|\theta)}{p(x|\theta)}\right) &= E_{q(x|\theta)}\left(\log q(x|\theta) - \log p(x|D)\right) \\ &= E_{q(x|\theta)}\left(\log q(x|\theta) - \log\left(p(D|x)p(x)\right) + \log p(D)\right) \\ &= E_{q(x|\theta)}\left(\log q(x|\theta) - \log\left(p(D|x)\right) - \log p(x)\right) \\ &= -E_{q(x|\theta)}\left(\log\left(p(D|x)\right)\right) + KL(q(x|\theta)||p(x)) \end{aligned}$$

Application to tough toy

Our choice (from toy):

$$q(m, s|\theta) = p_{simple}(m, s|d_1, d_2) = \mathcal{N}_m(\mu, \sigma_m)\mathcal{G}_s(\alpha_s, \beta_s)$$

where $\theta = [\mu, \sigma_m, \alpha_s, \beta_s]$ are to be optimized.

Log-likelihood:

$$\log p(d_1, d_2|s, m) = \left(-(\frac{3}{2} + \alpha) - 1\right) \log s - \frac{1}{2s} \sum_{i=1}^{2} (m - d_i)^2 - \frac{1}{2} \left(\frac{m^2}{\tau}\right)$$

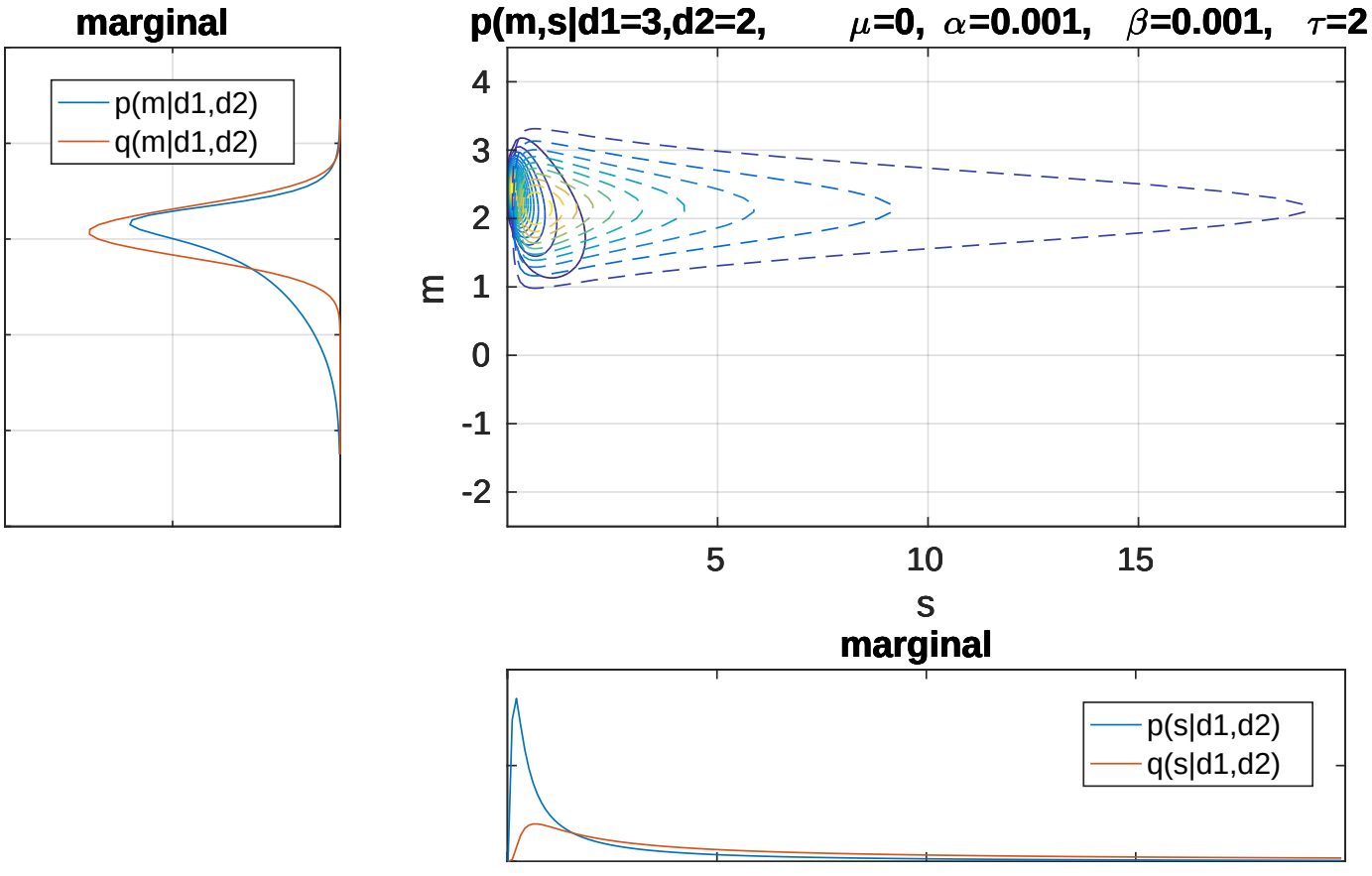
$$\mathrm{E}_q \left\{ \left(-(\frac{3}{2} + \alpha) - 1\right) \log s \right\} = \left(-(\frac{3}{2} + \alpha) - 1\right) (\log \beta - \psi(\alpha)),$$

$$\mathrm{E}_q \{1/s\} = \alpha/\beta,$$

$$\mathrm{E}_q \left\{m^2\right\} = \mu^2 + \sigma_m,$$

$$\mathrm{E}_q \left\{(m - d_i)^2\right\} = (\mu - d_i) + \sigma_m$$

Results


marginal
p(m|d1,d2)
q(m|d1,d2)
p(m,s|d1=3,d2=2, μ=0, α=0.001, β=0.001, τ=2
m
s
4
3
2
1
0
-1
-2
5
10
15
marginal
p(s|d1,d2)
q(s|d1,d2)

Monte Carlo: Mixture of Dirac

Probability density

$$p(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x^{(i)})$$

Cumulative function

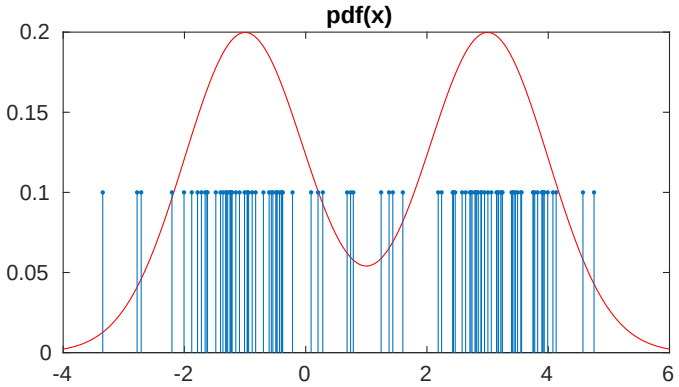
$$P(y) = \sum_{i=1}^{j} \frac{1}{n}, j = \sum_{i=1}^{n} \left[x^{(i)} < y \right]$$

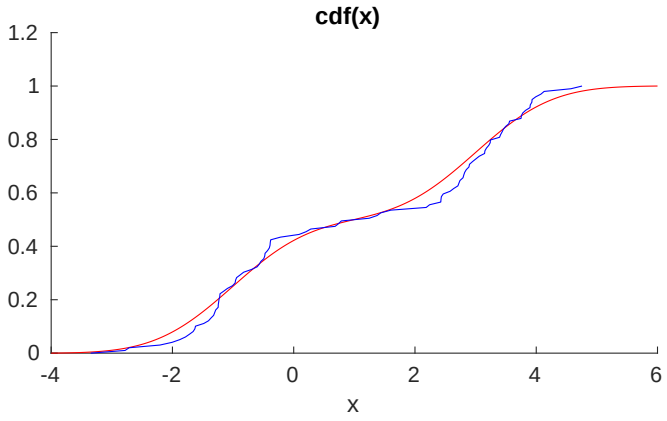
Expected value:

$$\mathrm{E}_{p(x)}(g(x)) = \frac{1}{n} \sum_{i=1}^{n} g(x^{(i)}),$$

Quantiles:

$$Q(p) = \inf \{x : p \leq P(x)\}.$$





Monte Carlo

Methods approximating unknown $p(x)$ by mixtures of Dirac.

$$p(x|D) \approx \frac{1}{n} \sum_{i=1}^{n} \delta(x - x^{(i)}).$$

Challenge: we can not sample from $p(x)$, we need a substitute.

1. Importance sampling, using proposal $q(x)$
   - 1.1 Adaptive importance sapling
   - 1.2 Population Monte Carlo
2. Monte Carlo Markov Chain (MCMC), using transition kernel,
   - 2.1 Metropolis-Hastings (Gibbs sampler)
   - 2.2 Hybrid MC (Hamiltonian Monte Carlo)

Convergence assured under mild conditions, different convergence rate.

Importance Sampling

To represent

$$p(\theta|\cdot) \approx \frac{1}{N}\sum_{i=1}^{N}\delta(\theta-\theta^{(i)}). \tag{1}$$

an ideal sampler should sample $\theta^{(i)} \sim p(\theta|\cdot)$, which is not available.

Using

$$p(\theta|D) = p(\theta|D)\frac{q(\theta)}{q(\theta)},$$

we can approximate $q(\theta)$ by (1) by sampling $\theta^{(i)} \sim q(\theta)$.

$$\begin{aligned} p(\theta) &\propto \frac{p(\theta)}{q(\theta)}\frac{1}{N}\sum_{i=1}^{N}\delta(\theta-\theta^{(i)}), \\ &\propto \sum_{i=1}^{N}\tilde{w}_i\delta(\theta-\theta^{(i)}), && \tilde{w}_i = \frac{p(\theta^{(i)})}{q(\theta^{(i)})} \\ &= \sum_{i=1}^{N} w_i\delta(\theta-\theta^{(i)}) && w_i = \frac{\tilde{w}_i}{\sum_{i=1}^{N}\tilde{w}_i} \end{aligned}$$

Algebra of weighted empirical distribution

Moments:

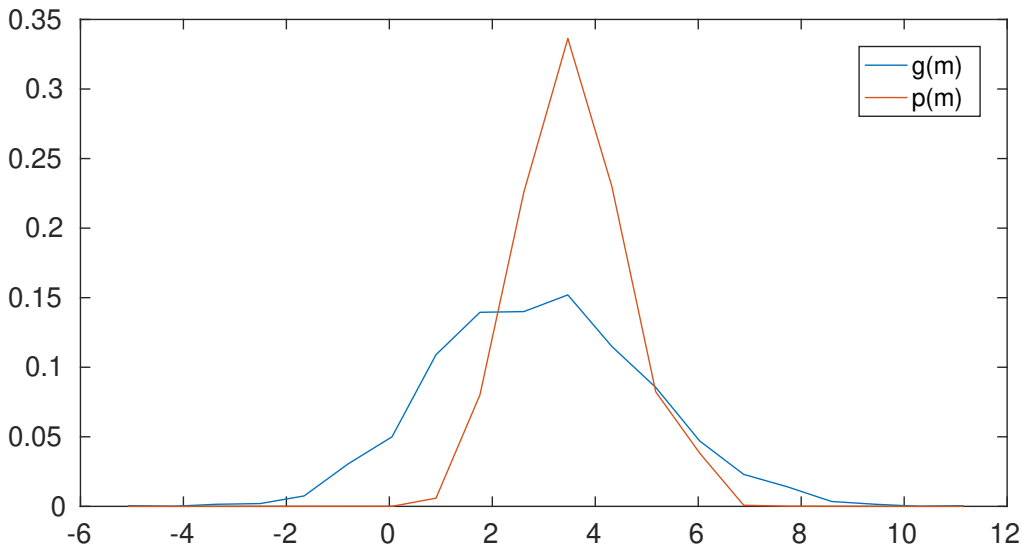
$$\mathrm{E}(g(\theta)) = \sum_{i=1}^{N} w_i g(\theta^{(i)})$$

Histogram:

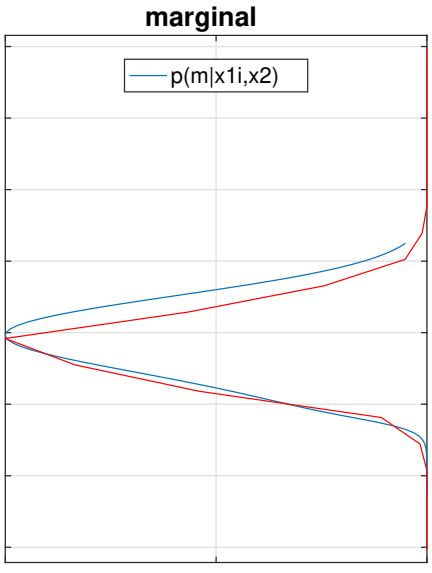
$$c_i = \sum_{i:x_i \in (l_i, u_i]} 1$$

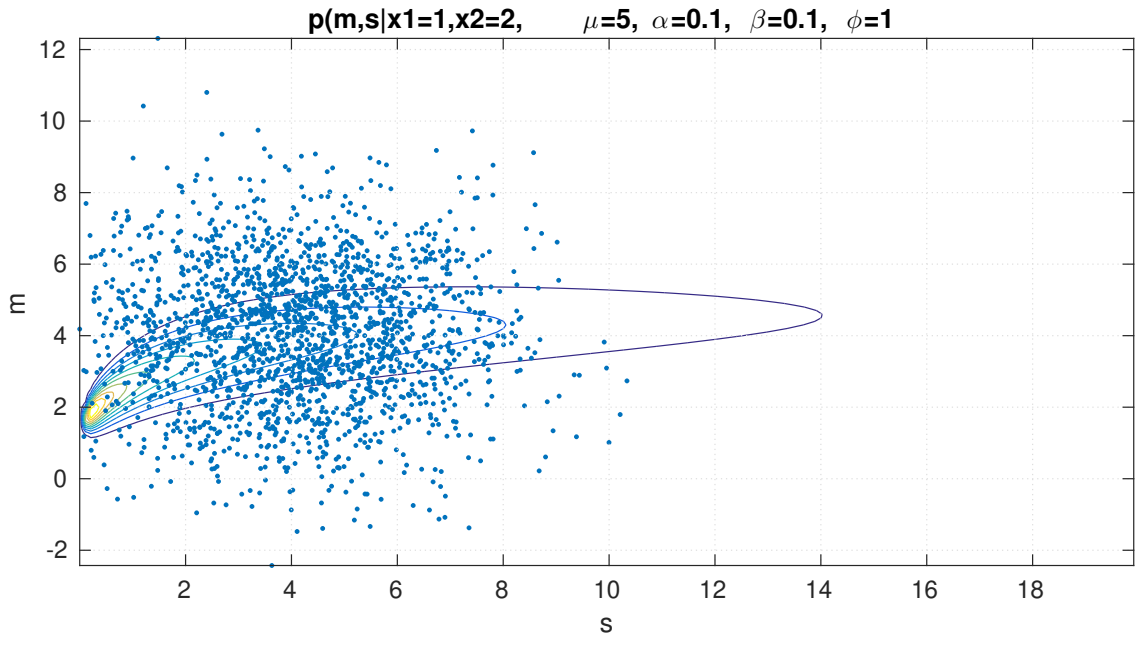
Weighted histogram:

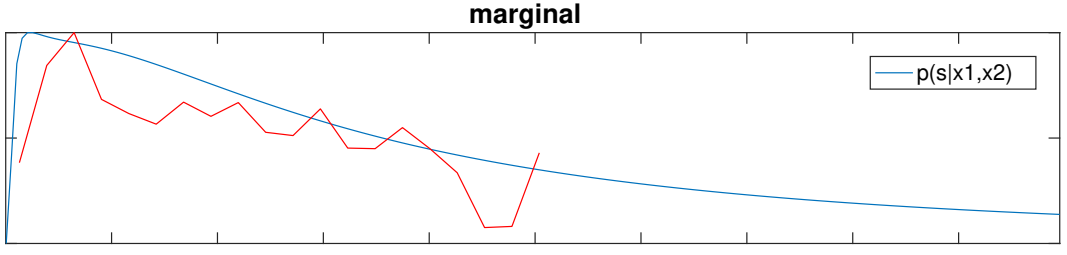
$$c_i = \sum_{i:x_i \in (l_i, u_i]} w_i$$



Toy: Importance sampling $N = 2000$







- Sample from heavy tailed distributions...

Adaptive Importance sampling

What if $q(\theta)$ is too far from $p(\theta)$?

Adaptive Importance sampling

What if $q(\theta)$ is too far from $p(\theta)$? Move it. Choose $q(\theta|\phi)$ and find $\hat{\phi}$.

Adaptive Importance sampling

What if $q(\theta)$ is too far from $p(\theta)$? Move it. Choose $q(\theta|\phi)$ and find $\hat{\phi}$.

Population MC: [Cappé, O., Guillin, A., Marin, J. M., & Robert, C. P. (2004).]

- Sample one generation
- compute weights, estimate parameter
- Sample next generation

Adaptive Importance sampling

What if $q(\theta)$ is too far from $p(\theta)$? Move it. Choose $q(\theta|\phi)$ and find $\hat{\phi}$.

Population MC: [Cappé, O., Guillin, A., Marin, J. M., & Robert, C. P. (2004).]

- Sample one generation
- compute weights, estimate parameter
- Sample next generation

AMIS: [CORNUET, J. M., MARIN, J. M., Mira, A., & Robert, C. P. (2012)]

- Consider each generation to be a component in deterministic mixture

$$q(\theta) = \sum_{g=1}^{G} q_g(\theta)$$

IMIS: [Steele, R. J., Raftery, A. E., & Emond, M. J. (2006).]

- build mixture at sample with high weight

MCMC: Metropolis Hastings

Instead of fixed distribution, we define a Markov chain that converges to the true distribution.

1. choose transition kernel $q(x|x^{(i)})$,
2. generate sample $x^* \sim q(x|x^{(i)})$,
3. With probability
$$\min\left(1, \frac{p(x^*)q(x^{(i)}|x^*)}{p(x^{(i)})q(x^*|x^{(i)})}\right)$$
accept ($i = i + 1$,$x^{(i)} = x^*$), else **reject**; goto 2.

MCMC: Metropolis Hastings

Instead of fixed distribution, we define a Markov chain that converges to the true distribution.

1. choose transition kernel $q(x|x^{(i)})$,
2. generate sample $x^* \sim q(x|x^{(i)})$,
3. With probability
$$\min\left(1, \frac{p(x^*)q(x^{(i)}|x^*)}{p(x^{(i)})q(x^*|x^{(i)})}\right)$$
accept ($i = i + 1$,$x^{(i)} = x^*$), else **reject**; goto 2.

How to choose the kernel:

- Random walk (Gaussian), $q(x|x^{(i)}) = \mathcal{N}(x^{(i)}, \phi I)$
  - small spread ϕ: correlated samples
  - high spread ϕ: low acceptance, too many rejections!
- Use known properties: conditionals

MCMC: Gibbs sampler

Special case of MH for mutidimensional distributions.

$$p(\theta_1, \theta_2, \ldots, \theta_k)$$

with MH probability of acceptance equal to **one**.

1. generate sample $\theta_1^{(i+1)} \sim p(\theta_1 | \theta_2^{(i)}, \ldots \theta_k^{(i)})$,
2. generate sample $\theta_2^{(i+1)} \sim p(\theta_2 | \theta_1^{(i+1)}, \ldots \theta_k^{(i)})$,

   $\vdots$
3. generate sample $\theta_k^{(i+1)} \sim p(\theta_2 | \theta_1^{(i+1)}, \ldots \theta_{k-1}^{(i+1)})$,

Suitable when these distributions are tractable.

MCMC: Gibbs sampler

Special case of MH for mutidimensional distributions.

$$p(\theta_1, \theta_2, \ldots, \theta_k)$$

with MH probability of acceptance equal to **one**.

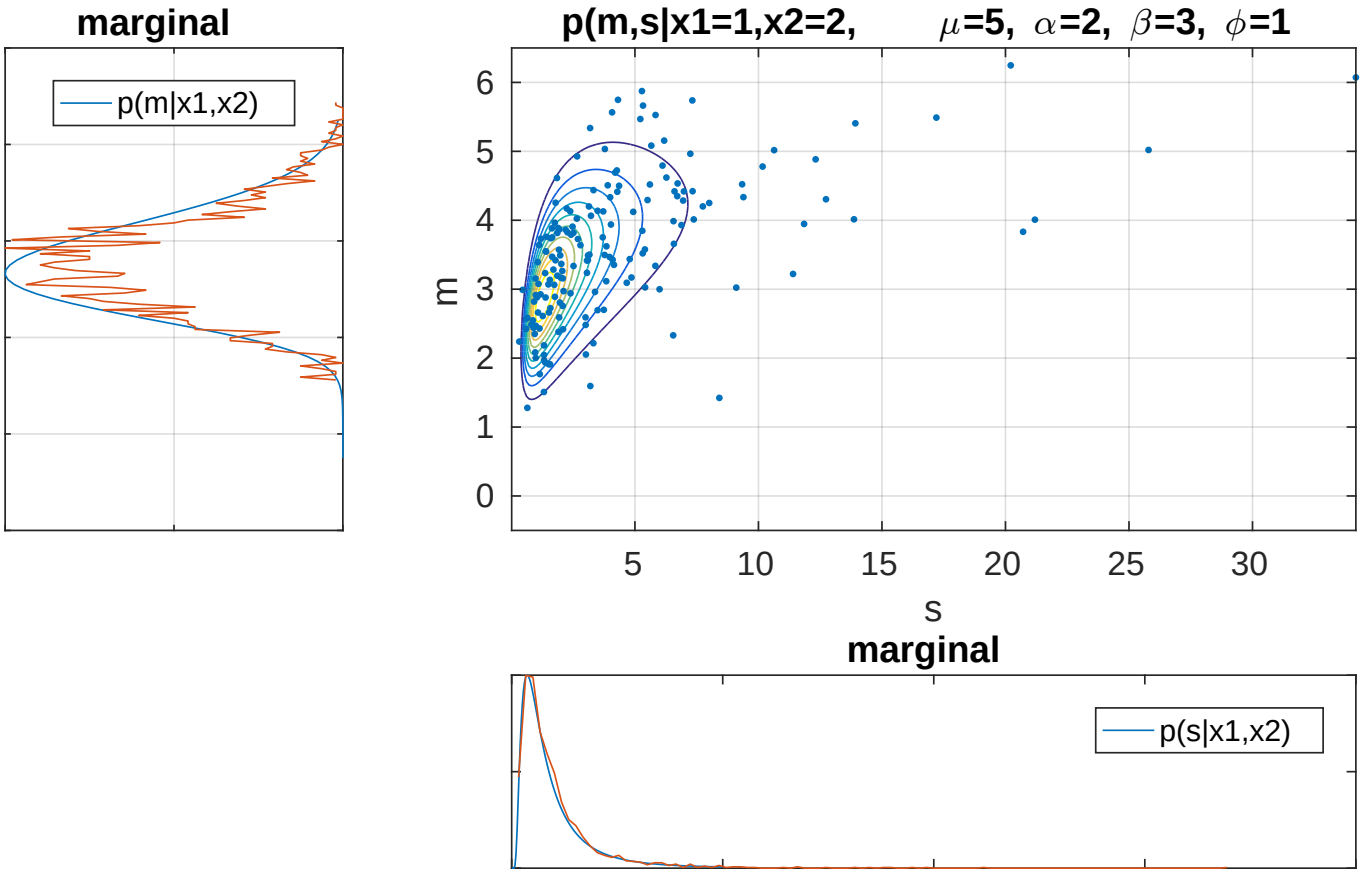
1. generate sample $\theta_1^{(i+1)} \sim p(\theta_1 | \theta_2^{(i)}, \ldots \theta_k^{(i)})$,
2. generate sample $\theta_2^{(i+1)} \sim p(\theta_2 | \theta_1^{(i+1)}, \ldots \theta_k^{(i)})$,

   $\vdots$
3. generate sample $\theta_k^{(i+1)} \sim p(\theta_2 | \theta_1^{(i+1)}, \ldots \theta_{k-1}^{(i+1)})$,

Suitable when these distributions are tractable.

- not suitable for parallel computing

Toy: Gibbs sampler



Hamiltonian(Hybrid) Monte Carlo

- view log-probability as potential energy

$$U(x) = -\log p(\theta) = \frac{\theta^2}{2},$$

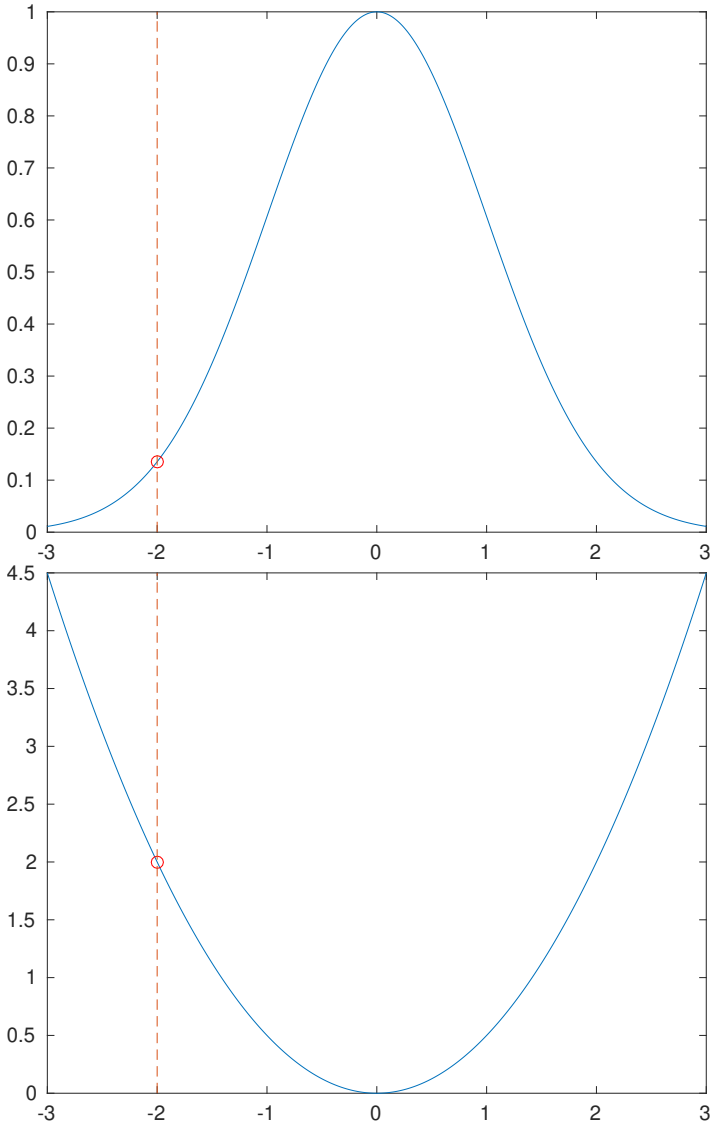
- add kinetic energy in variable p

$$K(p) = \frac{p^2}{2},$$

- define Hamiltonian

$$\frac{d\theta}{dt} = p, \quad \frac{dp}{dt} = -\theta$$

- simulate *differential equation* for selected $t = 0 \ldots t_s$
- resulting sample θ' is independent of p'
- Asymptotically converges to samples from true density.



Hamiltonian(Hybrid) Monte Carlo

- view log-probability as potential energy

$$U(x) = -\log p(\theta) = \frac{\theta^2}{2},$$

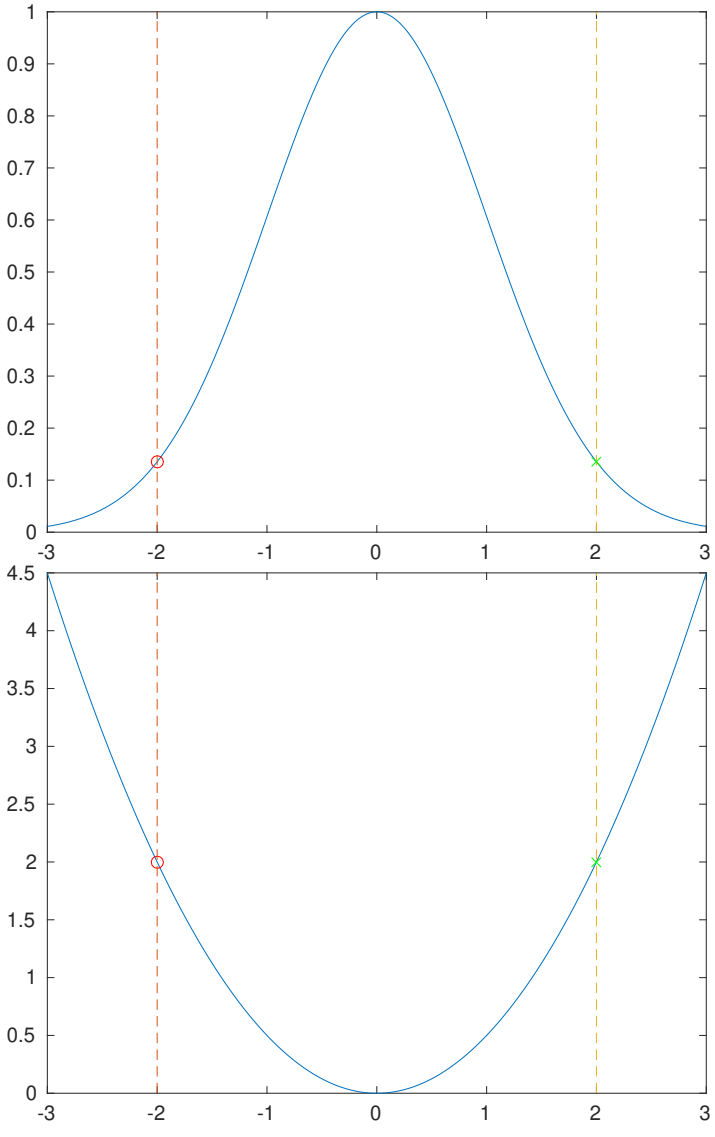
- add kinetic energy in variable p

$$K(p) = \frac{p^2}{2},$$

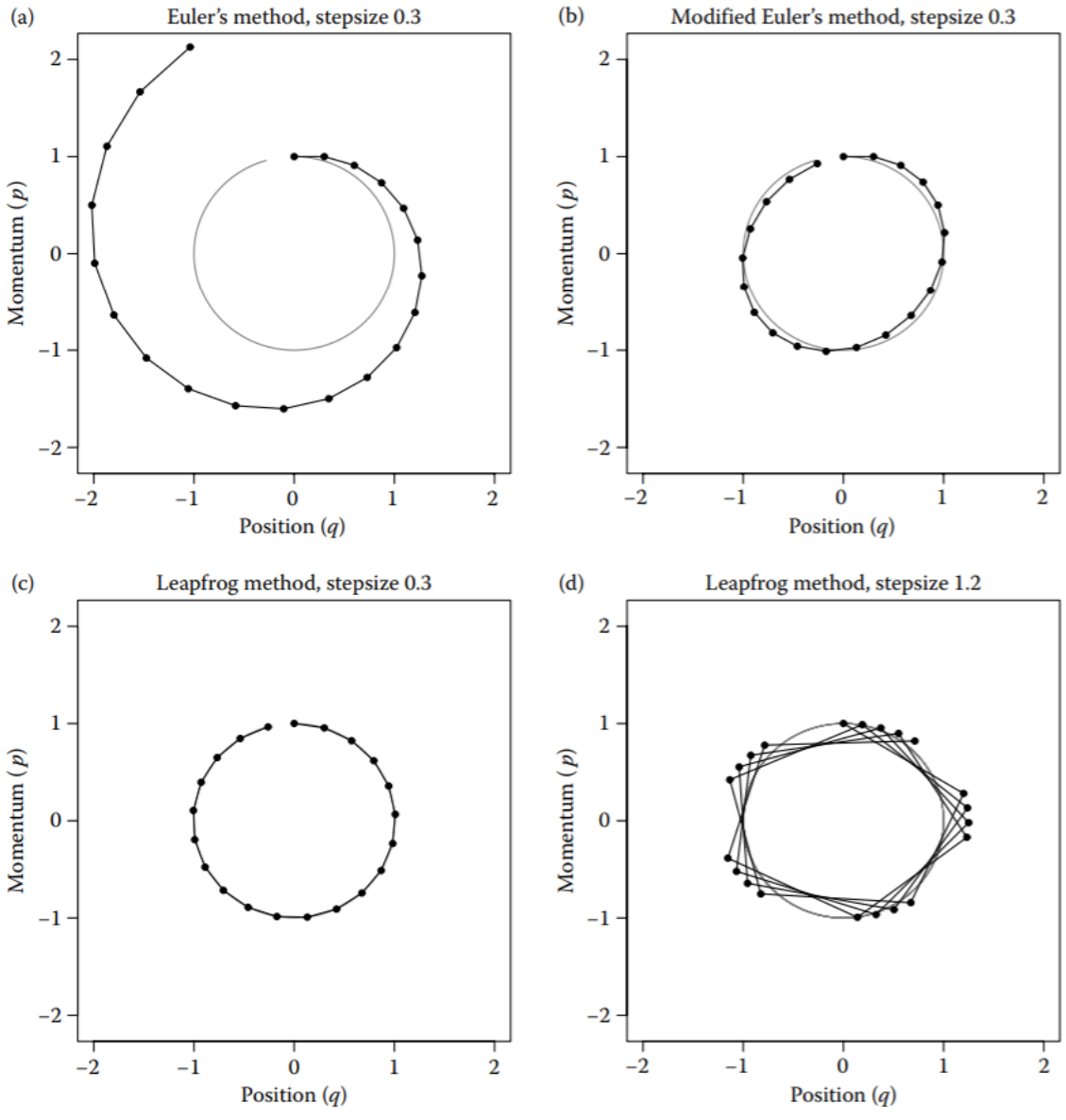
- define Hamiltonian

$$\frac{d\theta}{dt} = p, \quad \frac{dp}{dt} = -\theta$$

- simulate *differential equation* for selected $t = 0 \ldots t_s$
- resulting sample θ' is independent of p'
- Asymptotically converges to samples from true density.



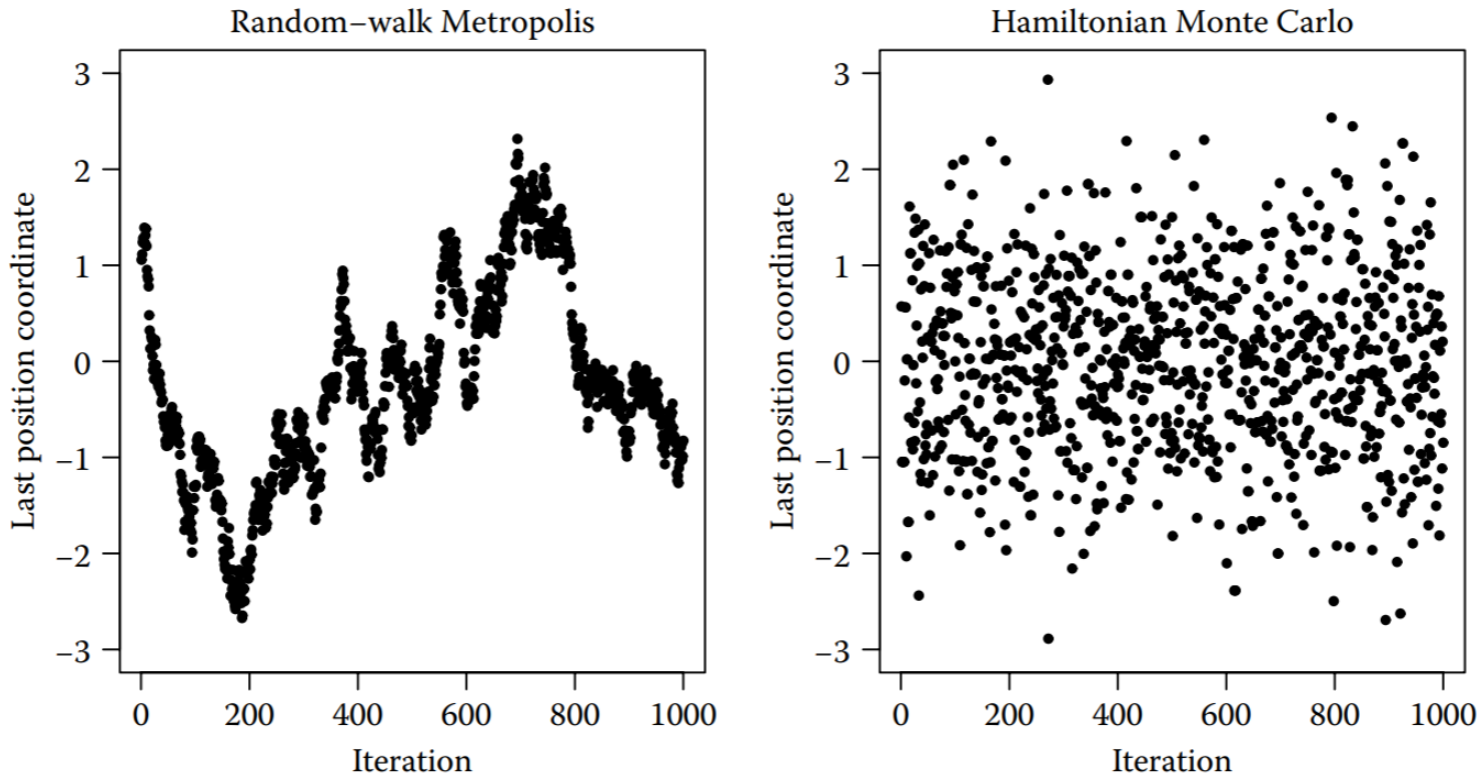
Numerical issues: leapfrog algorithm



[Neal, 2011]

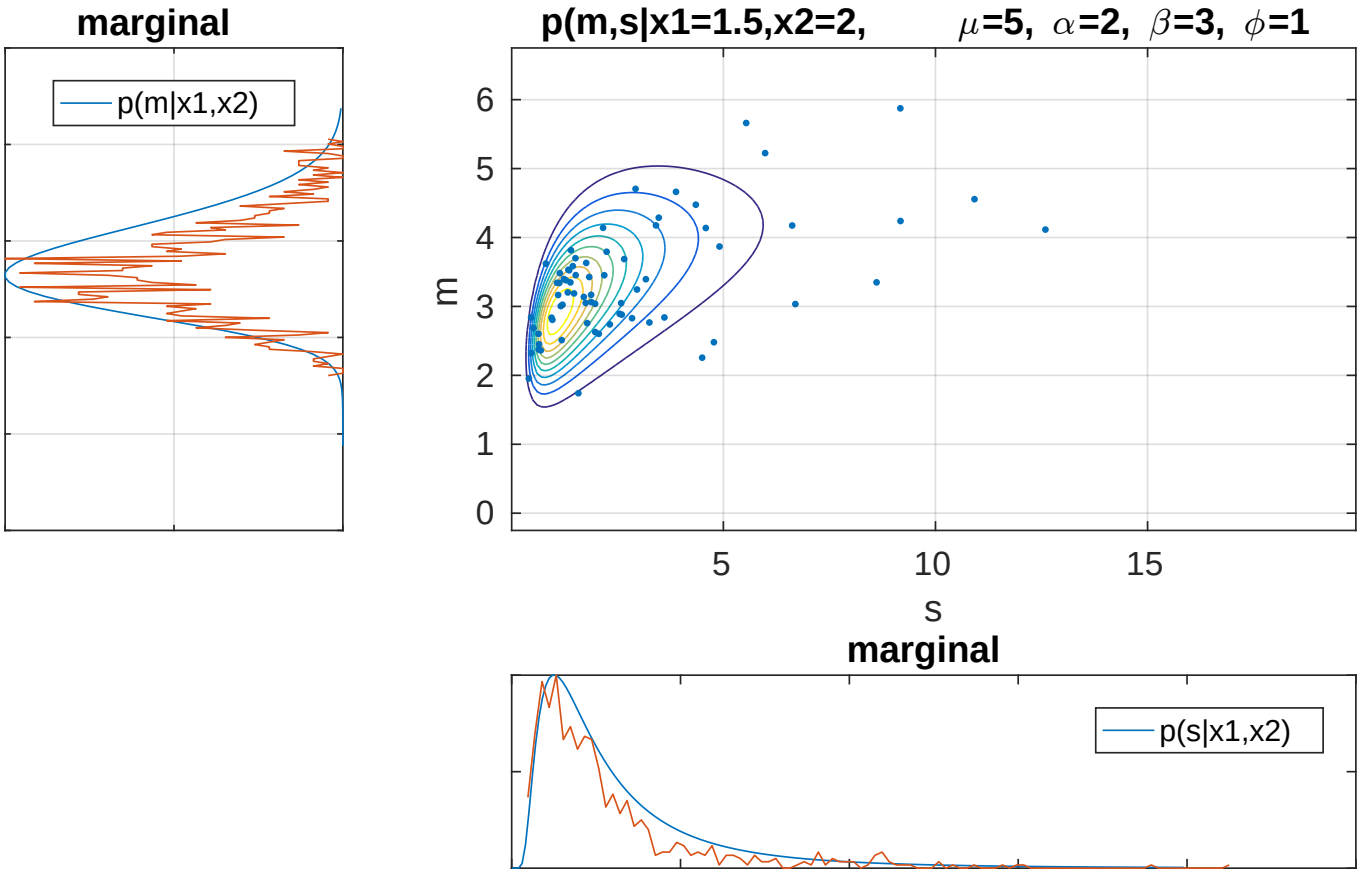
HMC advantages and disadvantages

- Able to use information about gradient
  - troubles with discrete variables
- Generated samples are not excessively correlated (check autocorrelation)



- Much faster exploraton of the space
  - at computational cost (doubles the number of variables
- How to choose stepsize and number of leapfrogs
  - NUTS, DynamicHMC, etc.

Results


marginal
p(m|x1,x2)
p(m,s|x1=1.5,x2=2, μ=5, α=2, β=3, ϕ=1
m
s
0
1
2
3
4
5
6
5
10
15
marginal
p(s|x1,x2)

Probabilistic programming

Universal nature of HMC gave rise to automatic tools:

STAN: https://mc-stan.org/

- HMC, NUTS
- Variational inference
- Matlab, R, Mathematica, Python, ...

Turing.jl: https://github.com/TuringLang/Turing.jl

- HMC, NUTS, SMC, ...
- Julia

Model development almost too easy

```
5   @model gdemo(x) = begin
6     s ~ InverseGamma(2,3)
7     m ~ Normal(0, sqrt(s))
8     x[1] ~ Normal(m, sqrt(s))
9     x[2] ~ Normal(m, sqrt(s))
10    return s, m
11  end
12
13  chain = sample(gdemo([1.5, 2.0]), SGLD(10000, 0.5))
```

- Non-conjugate priors
  - log-normal instead of inverse gamma
- automatic chain rule, differentiation
- Hard part: analyze results
- High dimensions?

Take Home

- Apart from simple problems, posterior distribution is intractable
- Trade-off between approximation quality and computational cost
- Variational methods
  - relatively cheap
  - requires good choice of posterior
- Monte Carlo
  - nice asymptotic properties,
  - expensive
  - hard analysis of results.