Variational Inference with Hamiltonian Monte Carlo

Christopher Wolf

Technische Universität München

christopher.wolf@tum.de

February 4, 2016

Variational inference

Setting: Probabilistic model $p(x, z)$ with missing or latent variables z

- Quantity of interest: Marginal likelihood $p(x) = \int p(x, z)dz$
- Often not directly tractable
- Use a variational lower bound

$$\begin{aligned}\log p(x) &\geq \log p(x) - D_{KL}\left(q_\theta(z|x)||p(z|x)\right) \\ &= \mathbb{E}_{q_\theta(z|x)}\left[\log p(x, z) - \log q_\theta(z|x)\right] =: \mathcal{L}\end{aligned}$$

Variational inference

Setting: Probabilistic model $p(x, z)$ with missing or latent variables z

- Quantity of interest: Marginal likelihood $p(x) = \int p(x, z)dz$
- Often not directly tractable
- Use a variational lower bound

$$\begin{aligned}\log p(x) &\geq \log p(x) - D_{KL}\left(q_\theta(z|x)||p(z|x)\right) \\ &= \mathbb{E}_{q_\theta(z|x)}\left[\log p(x, z) - \log q_\theta(z|x)\right] =: \mathcal{L}\end{aligned}$$

Problem: Quality of the bound strongly depends on the ability of $q_\theta(z|x)$ to approximate the true posterior

Variational inference

Setting: Probabilistic model $p(x, z)$ with missing or latent variables z

- Quantity of interest: Marginal likelihood $p(x) = \int p(x, z)dz$
- Often not directly tractable
- Use a variational lower bound

$$\begin{aligned}\log p(x) &\geq \log p(x) - D_{KL}\left(q_\theta(z|x)||p(z|x)\right) \\ &= \mathbb{E}_{q_\theta(z|x)}\left[\log p(x, z) - \log q_\theta(z|x)\right] =: \mathcal{L}\end{aligned}$$

Problem: Quality of the bound strongly depends on the ability of $q_\theta(z|x)$ to approximate the true posterior

Aim of this work: Improve the approximation

Markov Chain Monte Carlo (MCMC) methods

Objective: Generate samples from a complicated distribution $f_{\text{target}}(s)$

Idea: Create a Markov chain $(s_t)_{t \in \mathbb{N}}$, which converges to the target distribution

Markov Chain Monte Carlo (MCMC) methods

Objective: Generate samples from a complicated distribution $f_{\text{target}}(s)$

Idea: Create a Markov chain $(s_t)_{t\in\mathbb{N}}$, which converges to the target distribution

Metropolis-Hastings algorithm: Produces such a Markov chain

1. Generate a proposal $\tilde{s}$ from some distribution $f_{\text{prop}}(\tilde{s}|s_{t-1})$ (with $f_{\text{prop}}(t|s) = f_{\text{prop}}(s|t)$)
2. Accept the proposal as the next state s_t with probability

$$p_{\text{accept}} = \min[1, \frac{f_{\text{target}}(\tilde{s})}{f_{\text{target}}(s_{t-1})}],$$

   otherwise set $s_t = s_{t-1}$.

Integrating MCMC into variational inference

Salimans et al. (2015): Interpret the Markov chain obtained in MCMC as a variational approximation

Problem: Original lower bound $\mathcal{L} = \mathbb{E}_{q(z_T|x)} [\log p(x, z_T) - \log q(z_T|x)]$ is now intractable, so it needs to be modified:

Integrating MCMC into variational inference

Salimans et al. (2015): Interpret the Markov chain obtained in MCMC as a variational approximation

Problem: Original lower bound $\mathcal{L} = \mathbb{E}_{q(z_T|x)} [\log p(x, z_T) - \log q(z_T|x)]$ is now intractable, so it needs to be modified:

$$\begin{aligned}
\log p(x) &\geq \mathcal{L} \\
&\geq \mathcal{L} - \mathbb{E}_{q(z_T|x)} \left[D_{KL} \left(q(z_0, \ldots, z_{T-1}|z_T, x) || r(z_0, \ldots, z_{T-1}|z_T, x) \right) \right] \\
&=: \mathcal{L}_{\text{aux}},
\end{aligned}$$

$\rightarrow$ Need to also learn a *reverse* model $r(z_0, \ldots, z_{T-1}|z_T, x)$

Hamiltonian Dynamics

- Reformulation of classical dynamics
- System state given by *position* z and *momentum* v

Hamiltonian Dynamics

- Reformulation of classical dynamics
- System state given by *position* z and *momentum* v
- For HMC: Motion of frictionless particle due to *potential energy* $U(z)$ and *kinetic energy* $K(v)$

Hamiltonian Dynamics

- Reformulation of classical dynamics
- System state given by *position* z and *momentum* v
- For HMC: Motion of frictionless particle due to *potential energy* $U(z)$ and *kinetic energy* $K(v)$
- $U(z) = -\log p(x, z)$, the NLL of $p(z|x)$ upto additive constants

Hamiltonian Dynamics

- Reformulation of classical dynamics
- System state given by *position* z and *momentum* v
- For HMC: Motion of frictionless particle due to *potential energy* $U(z)$ and *kinetic energy* $K(v)$
- $U(z) = -\log p(x, z)$, the NLL of $p(z|x)$ upto additive constants
- $K(v) = (1/2)v^T M^{-1} v + C = -\log N(v|0, M)$, M called the mass matrix

Hamiltonian Dynamics - Numerical Solution

Leapfrog method:

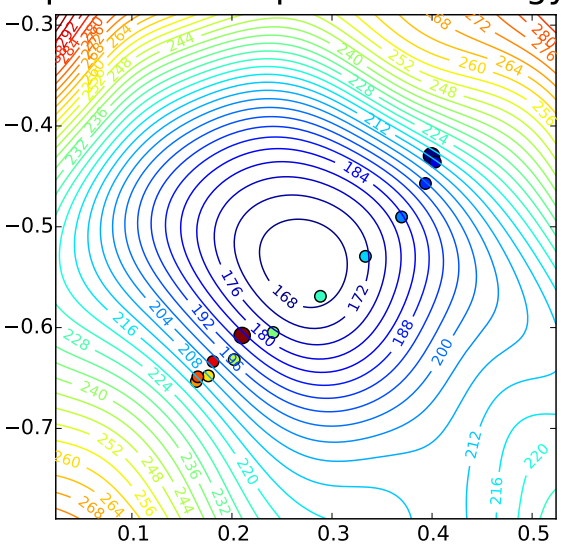
- volume-preserving
- reversible
- approximate conservation of total energy

Hamiltonian Dynamics - Numerical Solution

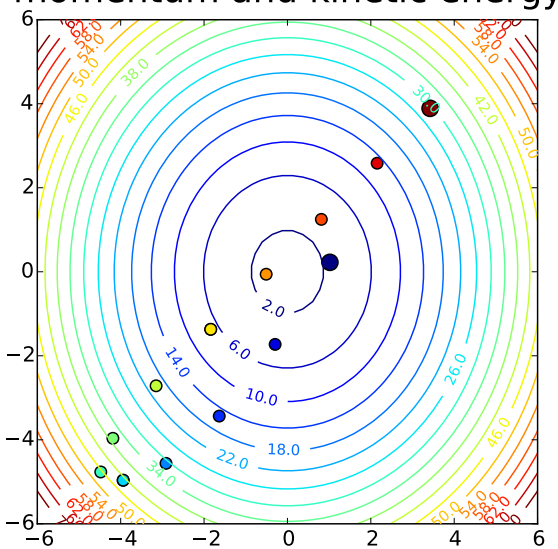
Leapfrog method:

- volume-preserving
- reversible
- approximate conservation of total energy

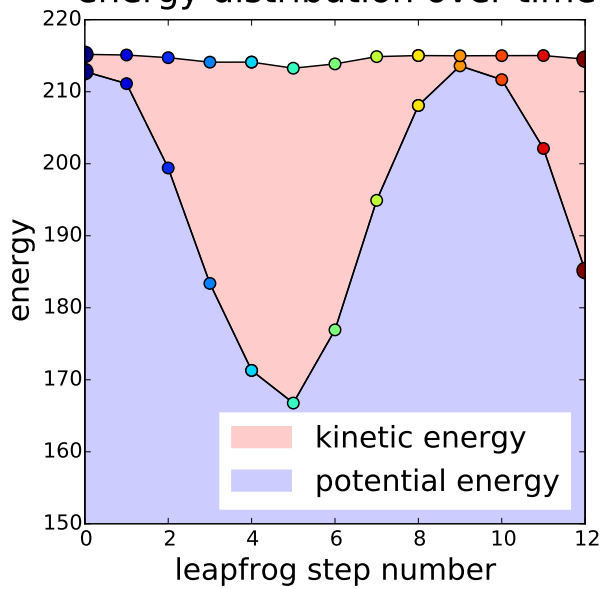
position and potential energy



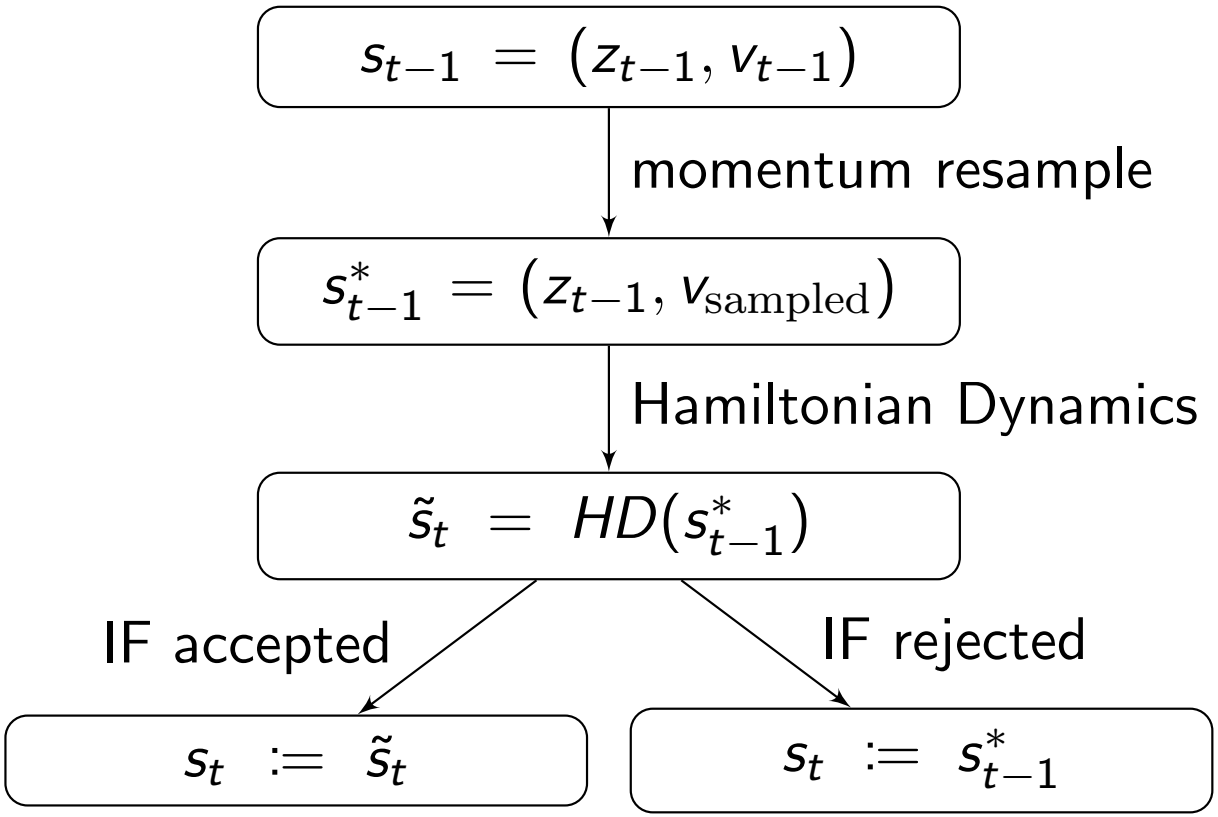
momentum and kinetic energy



energy distribution over time



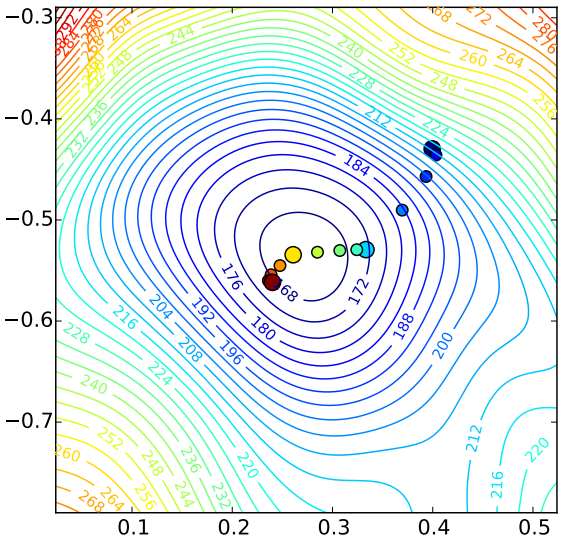
The Hamiltonian Monte Carlo (HMC) algorithm



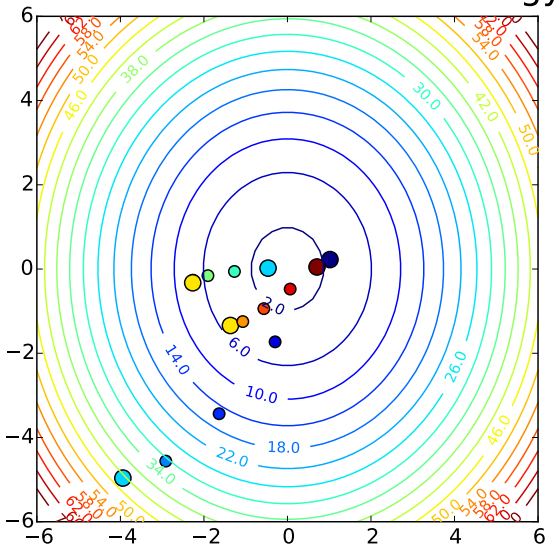
$\rightarrow$ The constructed $(s_t)_{t \in \mathbb{N}}$ converges to $p(z|x) \cdot N(v|0, M)$

Visualizations of HMC

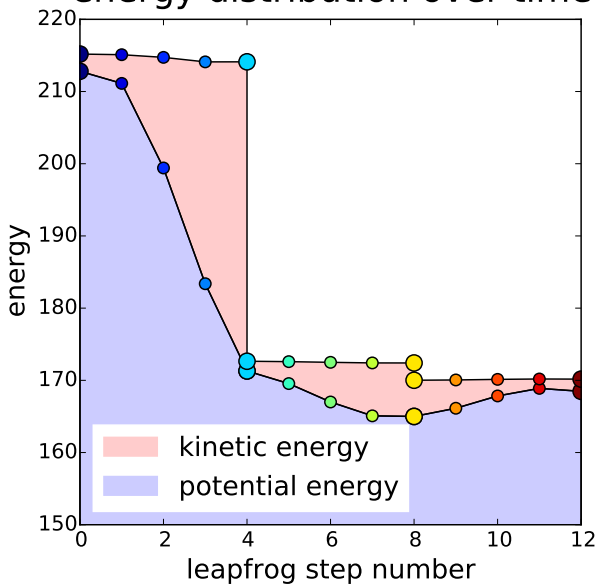
position and potential energy



momentum and kinetic energy



energy distribution over time



→ Momentum resampling leads to changes in total energy

→ Hamiltonian Dynamics explore the state spaces

Contributions

1. Include the acceptance step of the HMC algorithm in VI
2. Integrate partial momentum updates into lower bound
3. Make the kinetic energy input-dependent

Including the acceptance step

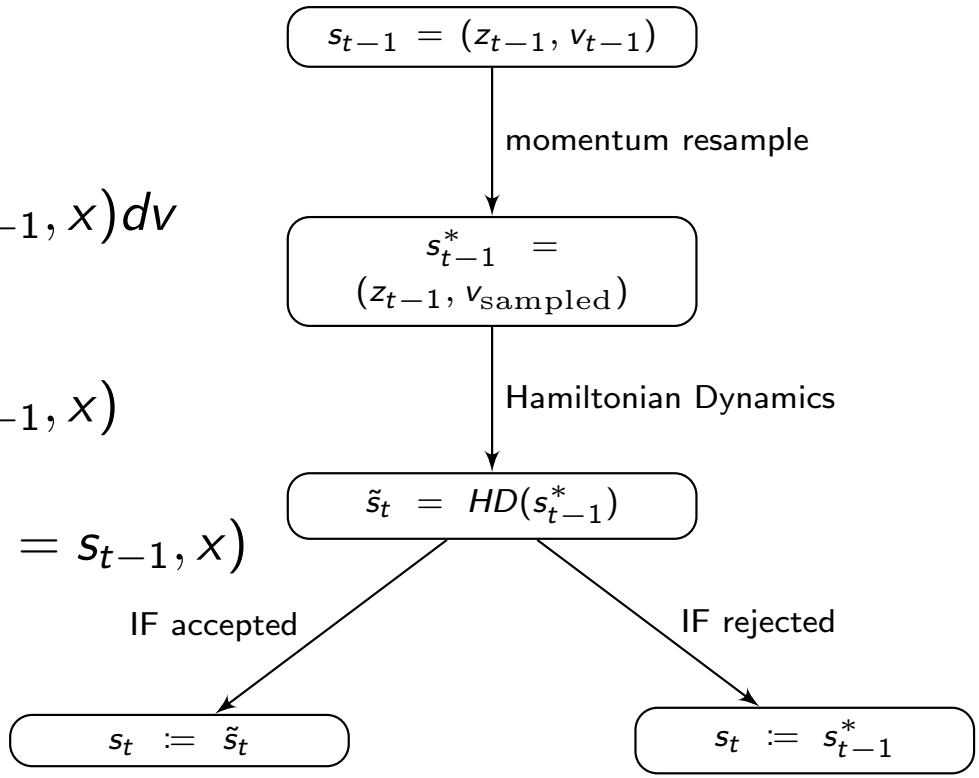
$$f_{S_t|S_{t-1},X}(s_t|s_{t-1},x)$$

$$= \sum_{a=0}^{1} \int f_{S_t,A,V_{\text{sampled}}|S_{t-1},X}(s_t,a,v|s_{t-1},x)dv$$

$$= \sum_{a=0}^{1} \int f_{S_t|A,V_{\text{sampled}},S_{t-1},X}(s_t|a,v,s_{t-1},x)$$

$$\cdot \mathbb{P}(A=a|V_{\text{sampled}}=v,S_{t-1}=s_{t-1},x)$$

$$\cdot f_{V_{\text{sampled}}|S_{t-1},X}(v|s_{t-1},x)dv$$



$\rightarrow$ With acceptance step: Asymptotic convergence is guaranteed

Partial momentum updates

In resampling step: Use weighted sum of current momentum and newly sampled momentum

Partial momentum updates

In resampling step: Use weighted sum of current momentum and newly sampled momentum

Effects:

- Particle tends to maintain its direction in the resampling step
- Avoids random walk behaviour
- Reduction of total energy is slower

$\rightarrow$ For sampling with HMC: Shown to be beneficial for short trajectories

Adaptive kinetic energy

Idea: Allow mass matrix M in the kinetic energy to depend on x

Adaptive kinetic energy

Idea: Allow mass matrix M in the kinetic energy to depend on x

Relevance: For trajectories of Hamiltonian Dynamics:

Change of mass matrix $\iff$ Rescaling of z-space

- Scaling too small $\rightarrow$ Large numerical errors
- Scaling too large $\rightarrow$ Slow exploration of the state space

$\rightarrow$ Input-dependent kinetic energy should improve convergence

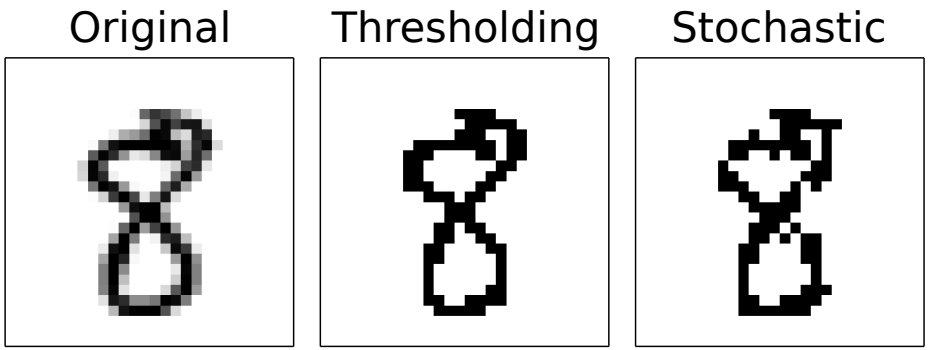
Experimental setup

Variational Auto-Encoders on stochastically binarized MNIST digits

Experimental setup

Variational Auto-Encoders on stochastically binarized MNIST digits

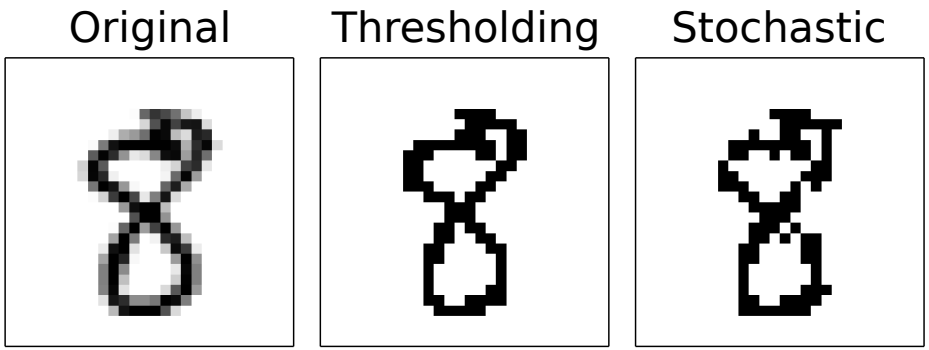
- Binarization: Want bilevel image to match Bernoulli modelling assumption



Experimental setup

Variational Auto-Encoders on stochastically binarized MNIST digits

- Binarization: Want bilevel image to match Bernoulli modelling assumption



- Stochastic binarization: Randomly set pixel to 1 with probability given by its pixel value

- Effectively much larger dataset; similar effect to dropout regularization

Experimental results

Name	#HMC	#LF	Partial	M	Accept	$-\mathcal{L}_{\text{aux}}$	$-\log(p(x)) \approx$
Basic VI	0	0	-	-	-	92.35	88.27
HMCVI 1	1	12	-	Global	-	89.77	87.77
HMCVI 2	2	6	-	Global	-	89.83	87.53
HMCVI 3	3	4	-	Global	-	90.24	87.56
HMCVI 4	3	4	Yes	Global	-	90.15	87.49
HMCVI 5	3	4	-	NN	-	90.23	87.30
HMCVI 6	3	4	Yes	NN	-	89.72	87.44
HMCVI 7	3	4	-	Global	Yes	91.40	87.28
HMCVI 8	3	4	-	NN	Yes	???	???

$\rightarrow$ VI with HMC outperforms basic VI

Experimental results

Name	#HMC	#LF	Partial	M	Accept	$-\mathcal{L}_{\text{aux}}$	$-\log(p(x)) \approx$
Basic VI	0	0	-	-	-	92.35	88.27
HMCVI 1	1	12	-	Global	-	89.77	87.77
HMCVI 2	2	6	-	Global	-	89.83	87.53
HMCVI 3	3	4	-	Global	-	90.24	87.56
HMCVI 4	3	4	Yes	Global	-	90.15	87.49
HMCVI 5	3	4	-	NN	-	90.23	87.30
HMCVI 6	3	4	Yes	NN	-	**89.72**	87.44
HMCVI 7	3	4	-	Global	Yes	91.40	87.28
HMCVI 8	3	4	-	NN	Yes	???	???

$\rightarrow$ VI with HMC outperforms basic VI

$\rightarrow$ Extensions to the HMC algorithm improve the performance

Experimental results

Name	#HMC	#LF	Partial	M	Accept	$-\mathcal{L}_{\text{aux}}$	$-\log(p(x)) \approx$
Basic VI	0	0	-	-	-	92.35	88.27
HMCVI 1	1	12	-	Global	-	89.77	87.77
HMCVI 2	2	6	-	Global	-	89.83	87.53
HMCVI 3	3	4	-	Global	-	90.24	87.56
HMCVI 4	3	4	Yes	Global	-	90.15	87.49
HMCVI 5	3	4	-	NN	-	90.23	87.30
HMCVI 6	3	4	Yes	NN	-	**89.72**	87.44
HMCVI 7	3	4	-	Global	Yes	91.40	87.28
HMCVI 8	3	4	-	NN	Yes	???	**???**

$\rightarrow$ VI with HMC outperforms basic VI

$\rightarrow$ Extensions to the HMC algorithm improve the performance

$\rightarrow$ Acceptance step improves the NLL

Future work

- More flexible reverse model
- Understand role of the auxiliary reverse model
- Speed up algorithm
- Allow other parameters to be input-dependent

Thank you for your attention

References

Neal, R. M. (2011). "MCMC using Hamiltonian dynamics". In: *Handbook of Markov Chain Monte Carlo*, pp. 113–162.

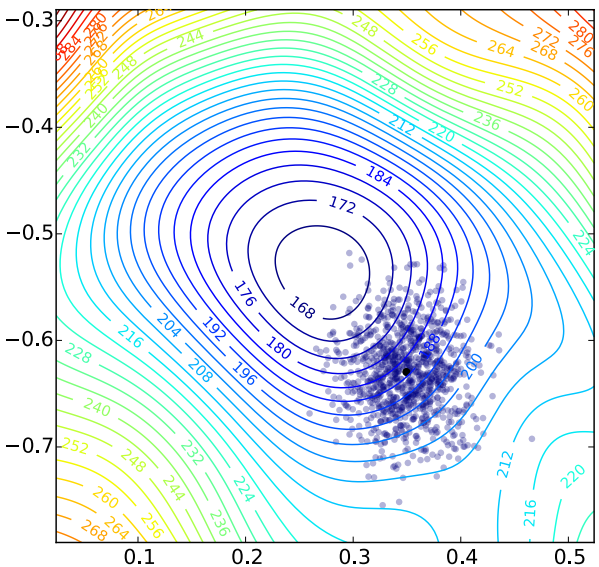
Salimans, T., D. P. Kingma, and M. Welling (2015). "Markov Chain Monte Carlo and Variational Inference: Bridging the Gap". In: *ICML 2015*.

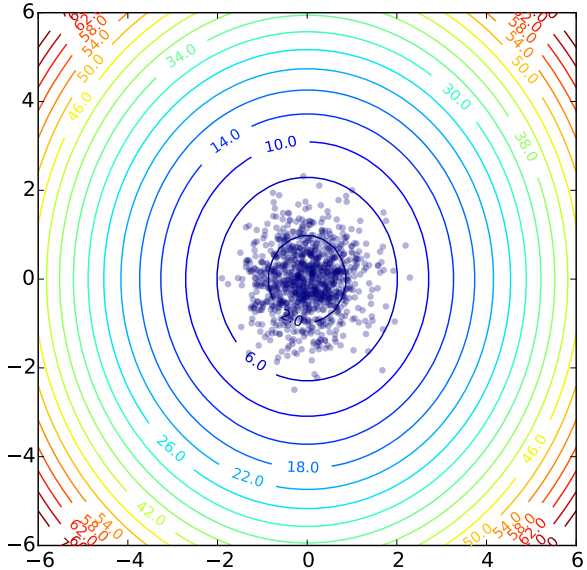
Visualizations of HMC (2)

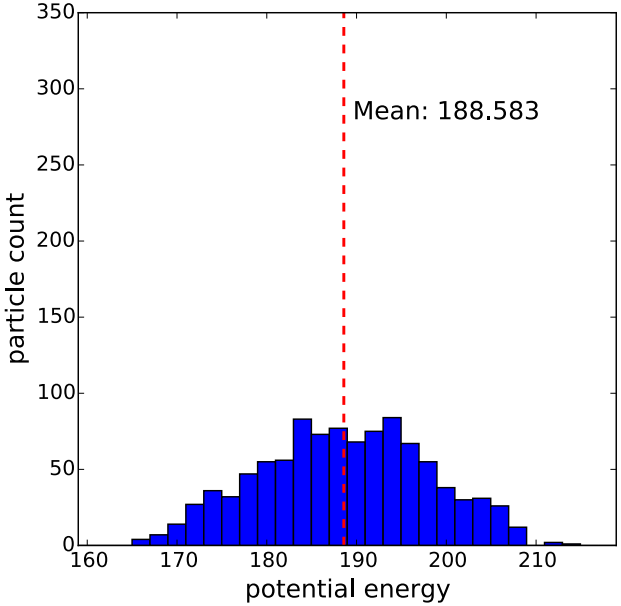
position and potential energy

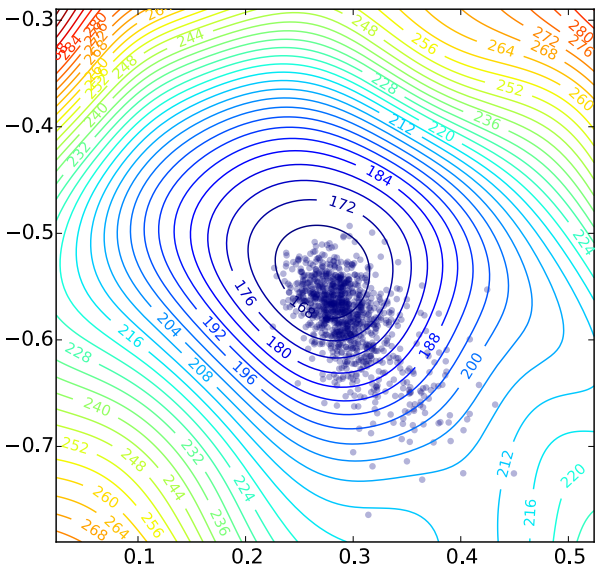
momentum and kinetic energy

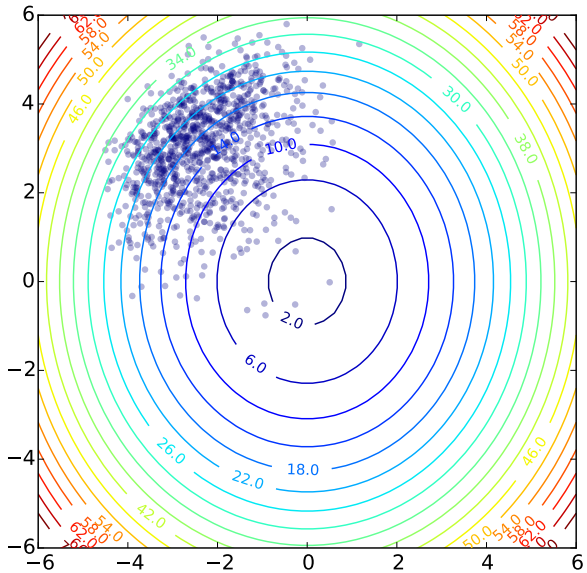
potential energy distribution

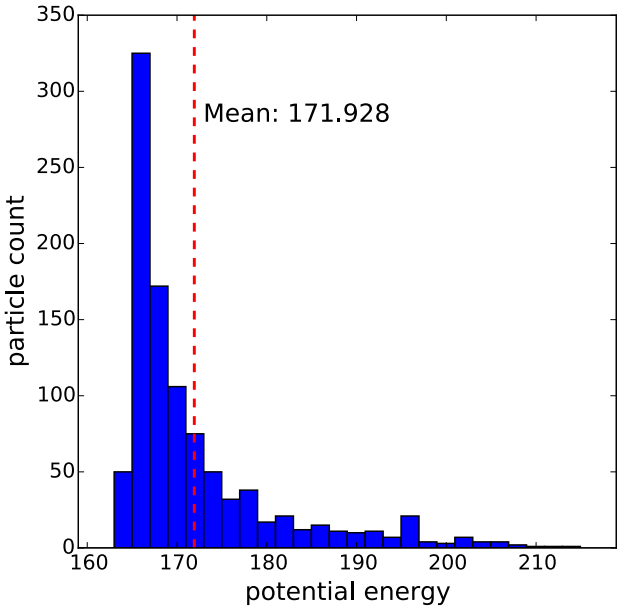












Potential energy of VAEs

Learnt latent representation of training data