The premise of approximate MCMC in Bayesian deep learning

Theodore Papamarkou

N8CIR

Leeds

1 November 2022

Talk structure

1. Problem specification
2. Possible ways forward
3. Future work

Topic and motivation

- Topic: approximate MCMC methods for neural networks
- Motivation: uncertainty quantification for neural networks

Status quo of Bayesian inference in high dimensions

	ML	**Statistics**
Approaches	Approximate models	Transformed models
	Ex: variational inference	Ex: stereographic MCMC
	Experimental validation	Exact inference (convergence)
Suggestions	Decompose exact model	Predictive performance bounds

Understanding what is the problem before solving it

- ML/statistics attack the problem computationaly/theoretically
- But what is the problem really?
- Zero acceptance rates are observed
- Theory: lack of convergence, symmetries, identifiability
- Practically, what are the roots of the problem?

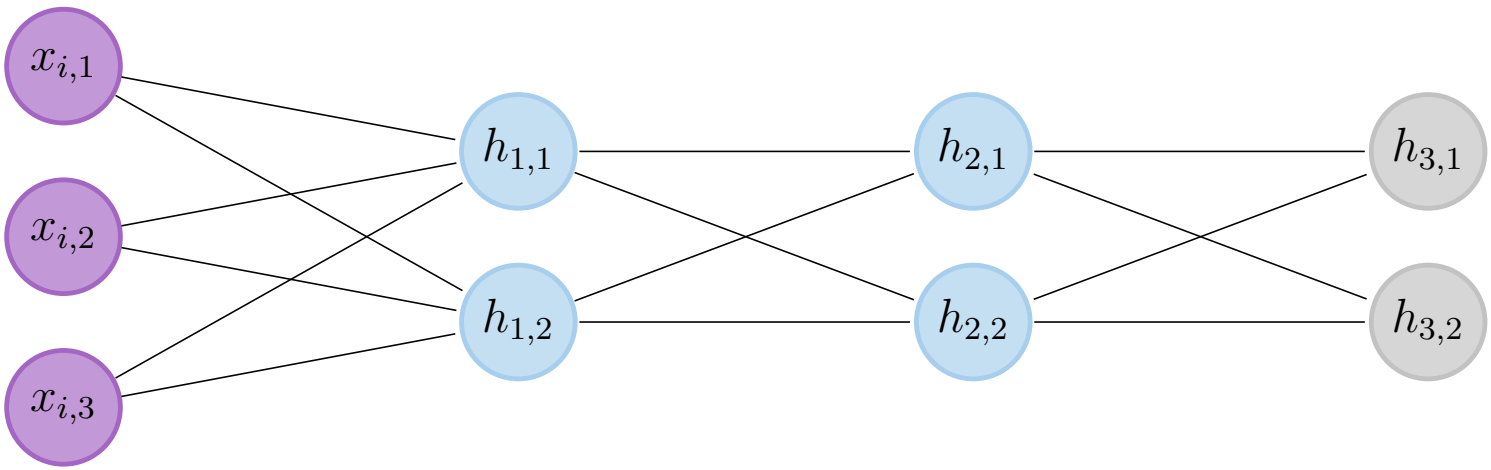
Issue 1: big data (number of samples)

Dataset		
Name	Sample size	
	Training	Test
MNIST	$60,000$	$10,000$
Fashion MNIST (FMNIST)	$60,000$	$10,000$

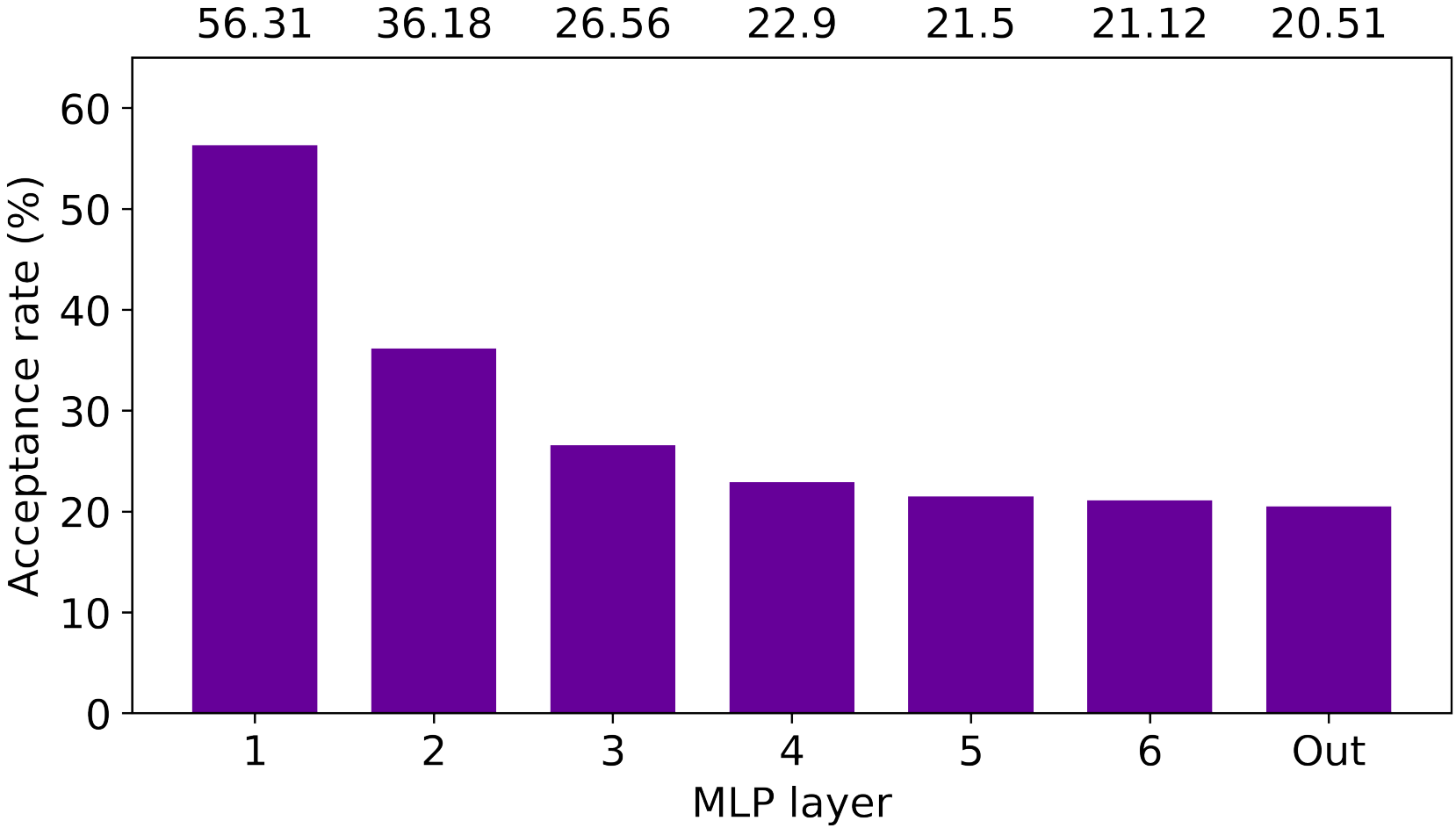
Issue 2: big models (number of parameters)

Neural network	
Architecture	# parameters
MLP$(784, 10, 10, 10, 10)$	$8,180$

Issue 3: hierarchical model structure (part I)



Issue 3: hierarchical model structure (part II)



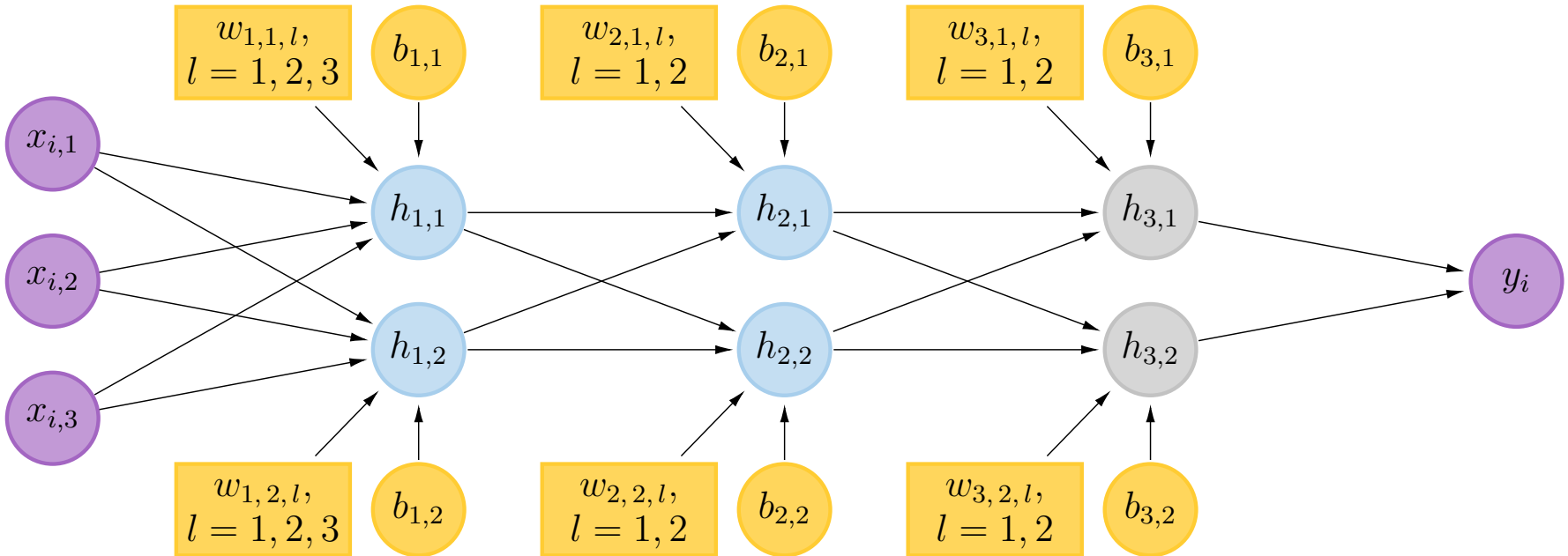
Summary of issues

	Issue
1	Number of samples
2	Number of parameters
3	Hierarchical model

Possible solutions

	Issue	**Solution**
1	Number of samples	Minibatch sampling
2	Number of parameters	Block sampling
3	Hierarchical model	Depth $\nearrow \Rightarrow$ Step $\searrow$

Training (part I)



Training (part II)

Algorithm 1 Metropolis-within-blocked-Gibbs sampling

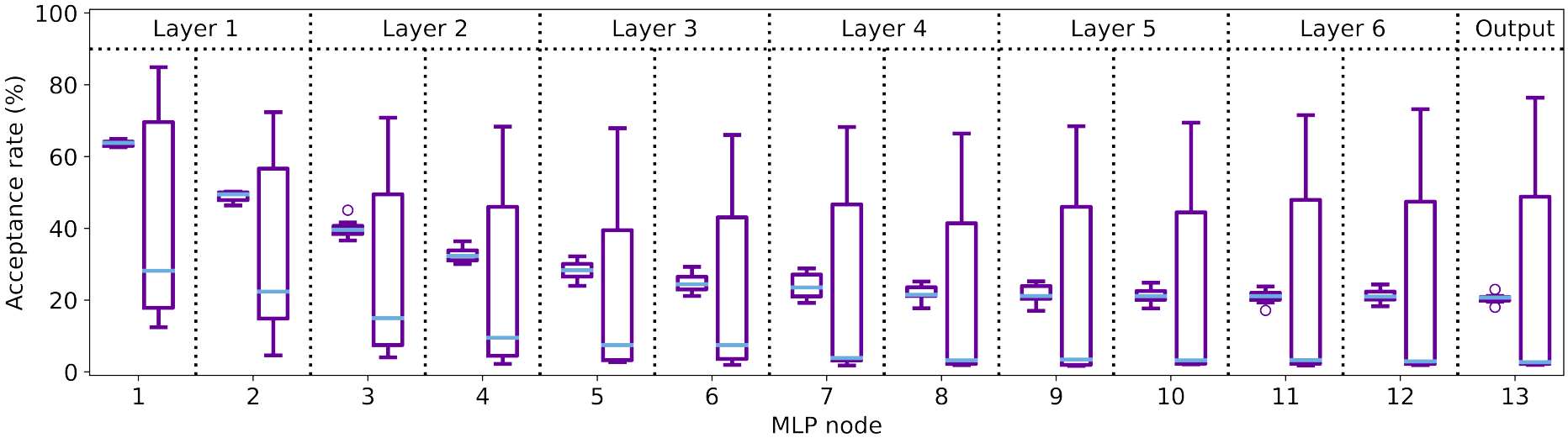
Input: training dataset $\mathcal{D}_{1:s}$
Input: initial state $\theta^{(0)}_{z(1):z(m)}$
Input: proposal variances $(\sigma_1^2, \ldots, \sigma_m^2)$ across blocks

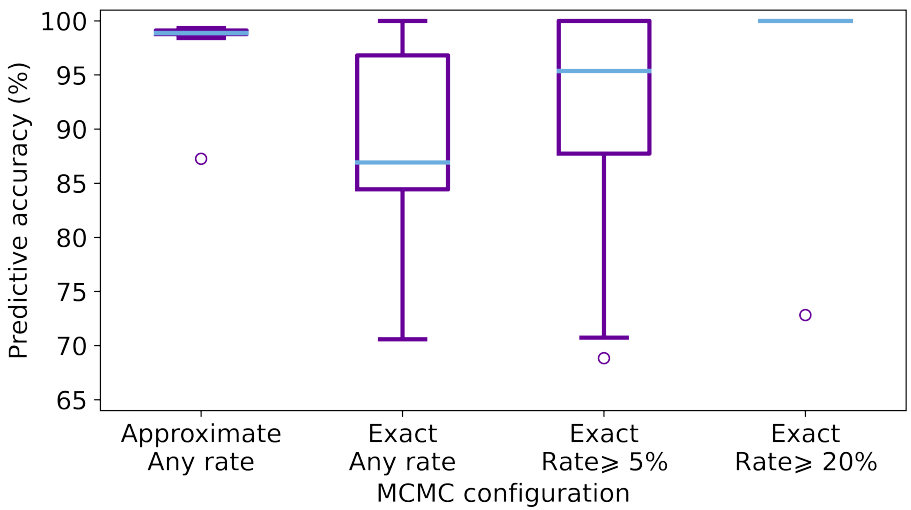
for $t = 1, \ldots, v$ **do**
    for $q = 1, \ldots, m$ **do**
        $\theta^{\star}_{z(q)} \sim \mathcal{N}(\theta^{(t-1)}_{z(q)}, \sigma_q^2 I_q)$
        $$a(\theta^{\star}_{z(q)}, \theta^{(t-1)}_{z(q)}) = \min \left\{ \frac{\pi(\theta^{\star}_{z(q)})}{\pi(\theta^{(t-1)}_{z(q)})} \exp\left(\mathcal{E}(\theta^{(t-1)}, \mathcal{D}_{1:s}) - \mathcal{E}(\theta^{\star}, \mathcal{D}_{1:s}) \right), 1 \right\}$$
        $u \sim \mathcal{U}(0, 1)$
        if $u \leq a(\theta^{\star}_{z(q)}, \theta^{(t-1)}_{z(q)})$ **then**
            Set $\theta^{(t)}_{z(q)} = \theta^{\star}_{z(q)}$
        else
            Set $\theta^{(t)}_{z(q)} = \theta^{(t-1)}_{z(q)}$
        end if
    end for
end for

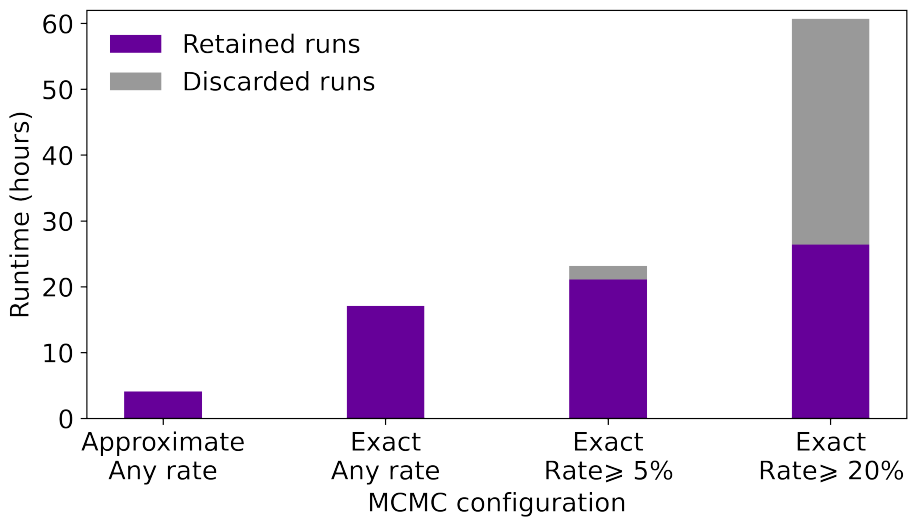
Testing (Bayesian marginalization)

$$\underbrace{p(y|x, D_{1:s})}_{\text{Predictive distribution}} = \int \underbrace{p(y|x, \theta)}_{\text{Likelihood}} \underbrace{p(\theta|D_{1:s})}_{\text{Parameter posterior}} d\theta$$

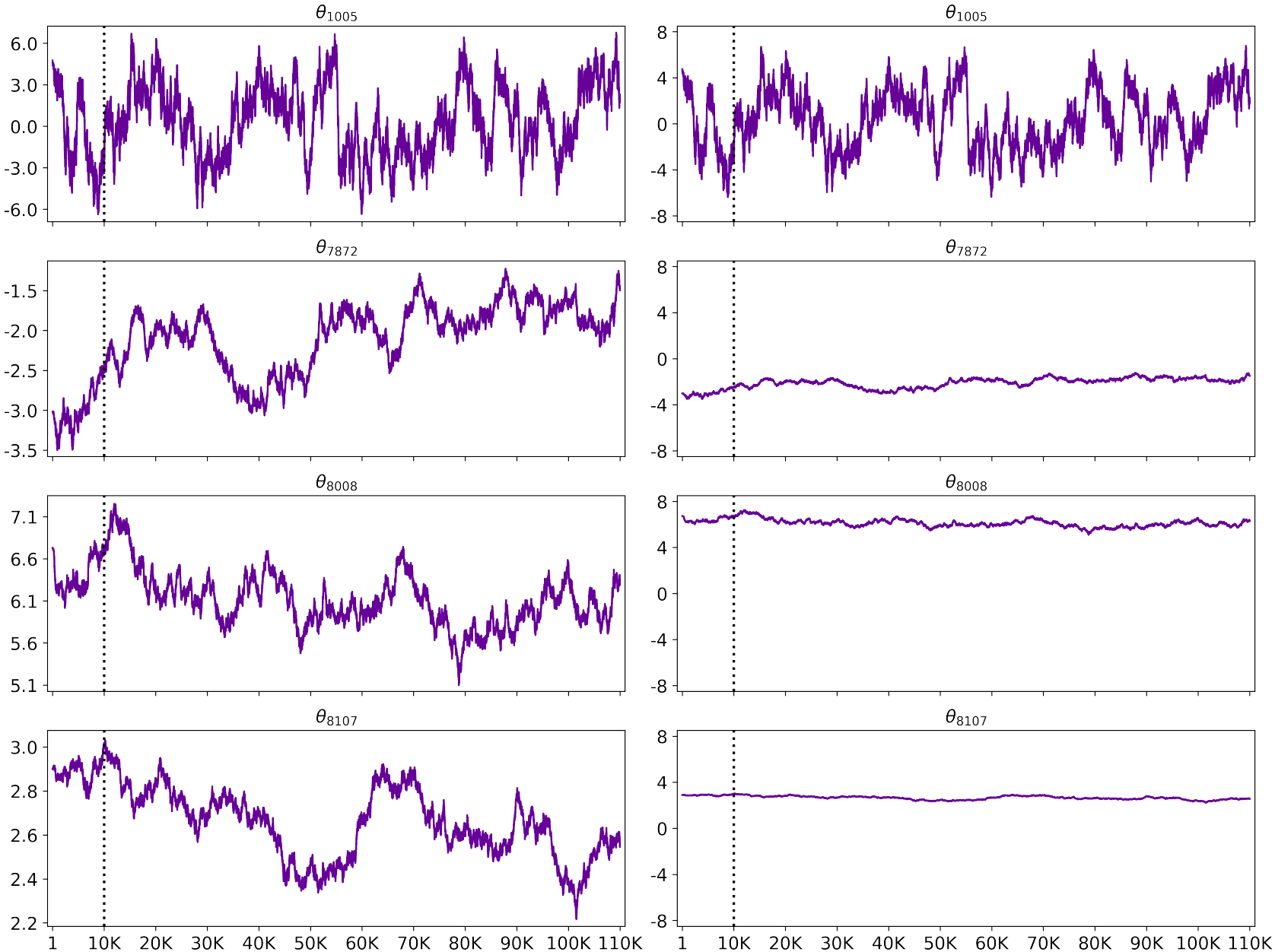
Revisiting issue 1: exact vs minibatch MCMC







Revisiting issue 2: finer node-blocked sampling (part I)



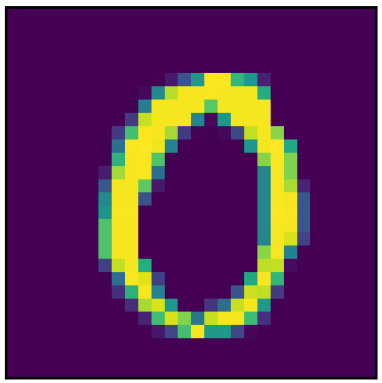
Revisiting issue 2: finer node-blocked sampling (part II)

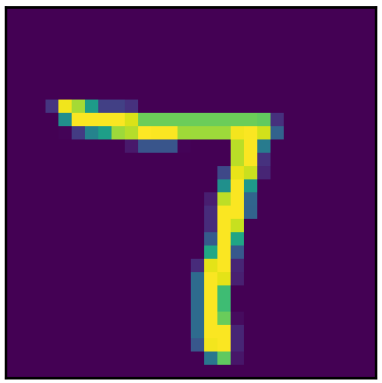
Dataset	Chain length			
	1K	10K	20K	30K
MNIST	88.31	90.75	91.12	91.20
FMNIST	78.93	80.89	81.36	81.53

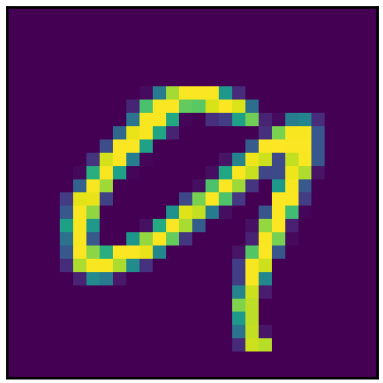
Revisiting issue 3: acceptance rates and depth (MNIST)

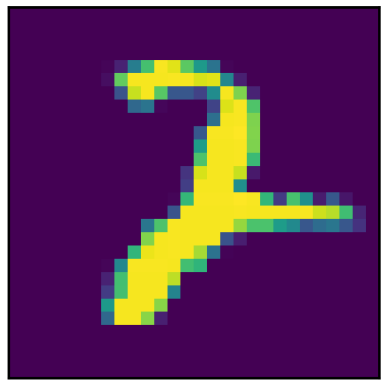
Layer		σ	Rate
Batch size = 1800 (3%)			
Hidden	1^{st}	$2 \cdot 10^{-2}$	41.41
	2^{nd}	$2 \cdot 10^{-4}$	30.68
	3^{rd}	$2 \cdot 10^{-4}$	31.92
Output		$2 \cdot 10^{-5}$	35.66
Batch size = 3000 (5%)			
Hidden	1^{st}	10^{-2}	54.95
	2^{nd}	10^{-4}	45.73
	3^{rd}	10^{-4}	44.98
Output		10^{-5}	51.54

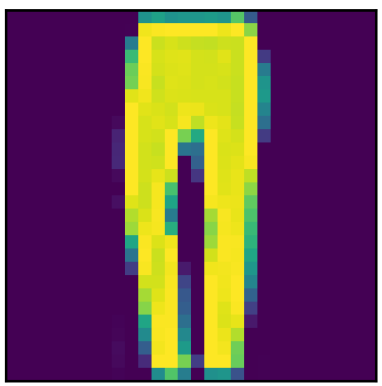
Uncertainty quantification

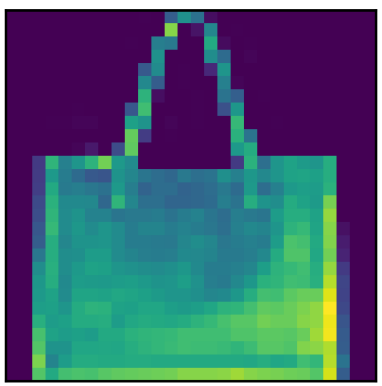


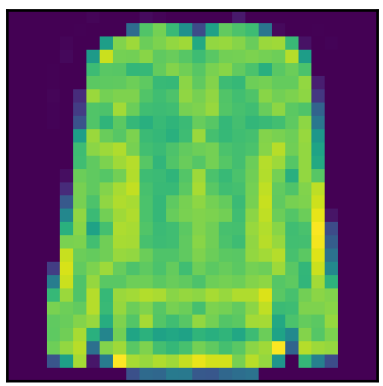


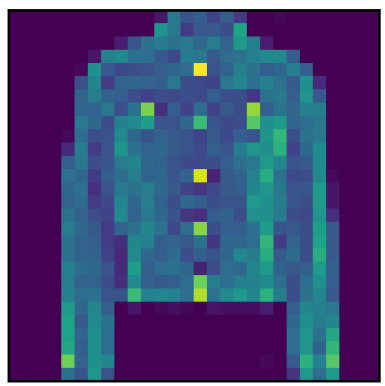












Future work

- Automate tuning of proposal variances via adaptive Gibbs
- Examine alternative parameter blocks
- Derive lower bounds of predictive accuracy