Batched Training of Kernel Deep Neural Network

Xiaoran Tong
Qin Lu

Michigan State University

tongxia1@msu.edu
qlu@epi.msu.edu

December 1, 2017

Table of Content

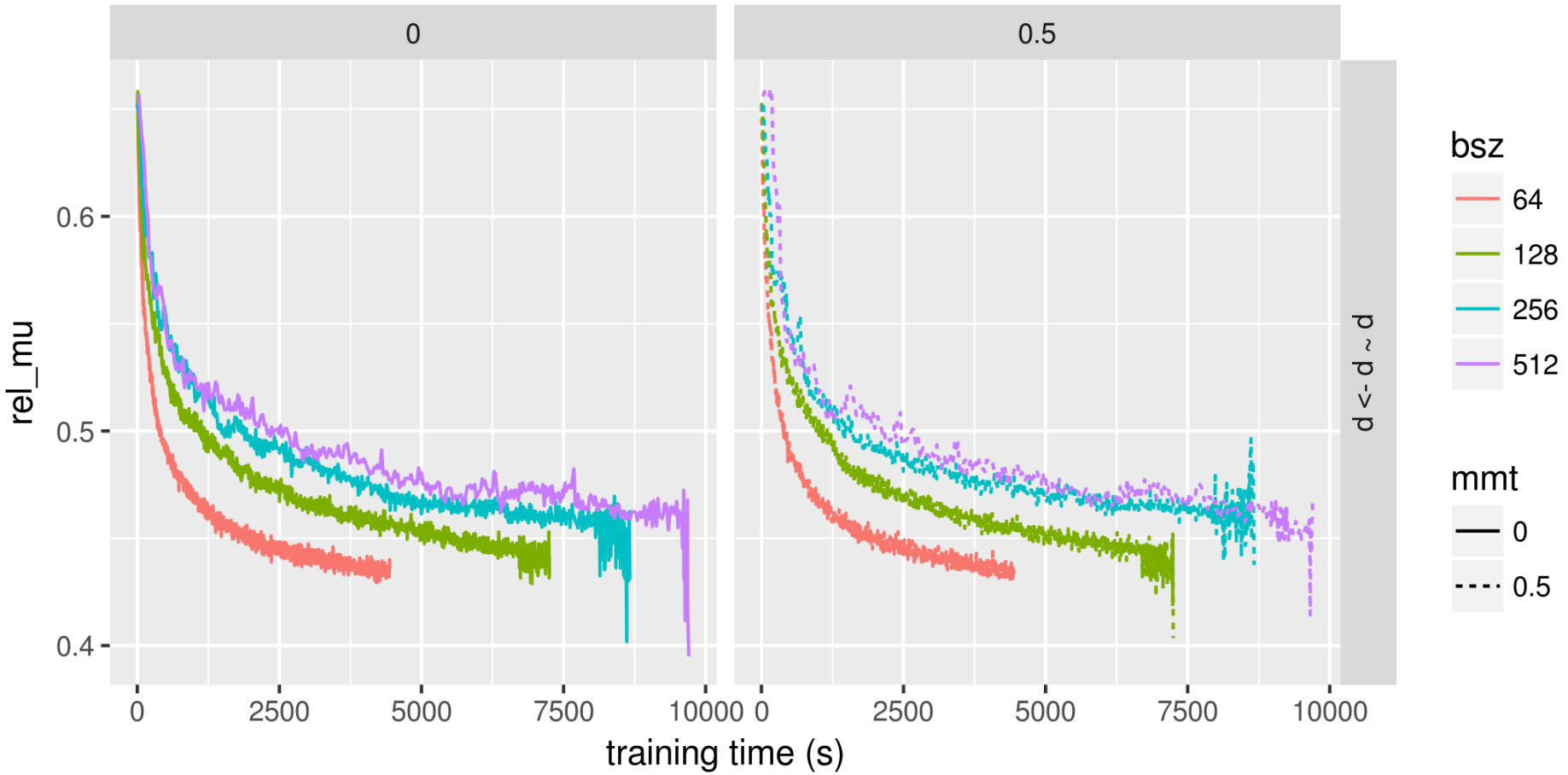
EXP 01: Tied Weight, *linear* $\sim$ *linear*|*linear*

Simulation Settings:

- N = 512, P=2000, FRQ=0.50, ϕ=1.00
- Number of Hidden Units = 256
- True Model: $y \sim \mathcal{N}(0, X'X + \phi I)$
- Batch Size = {512, 256, 128, 64}
- learning rate = 1e-5, 2e-5, 4e-5, 8e-5
- Momentum = {0.0, 0.5}
- Reserved Time = 3 hours
- Reserved Iteration = 1250
- No rendom seed.

EXP 01: Tied Weight, *linear* ∼ *linear*|*linear*

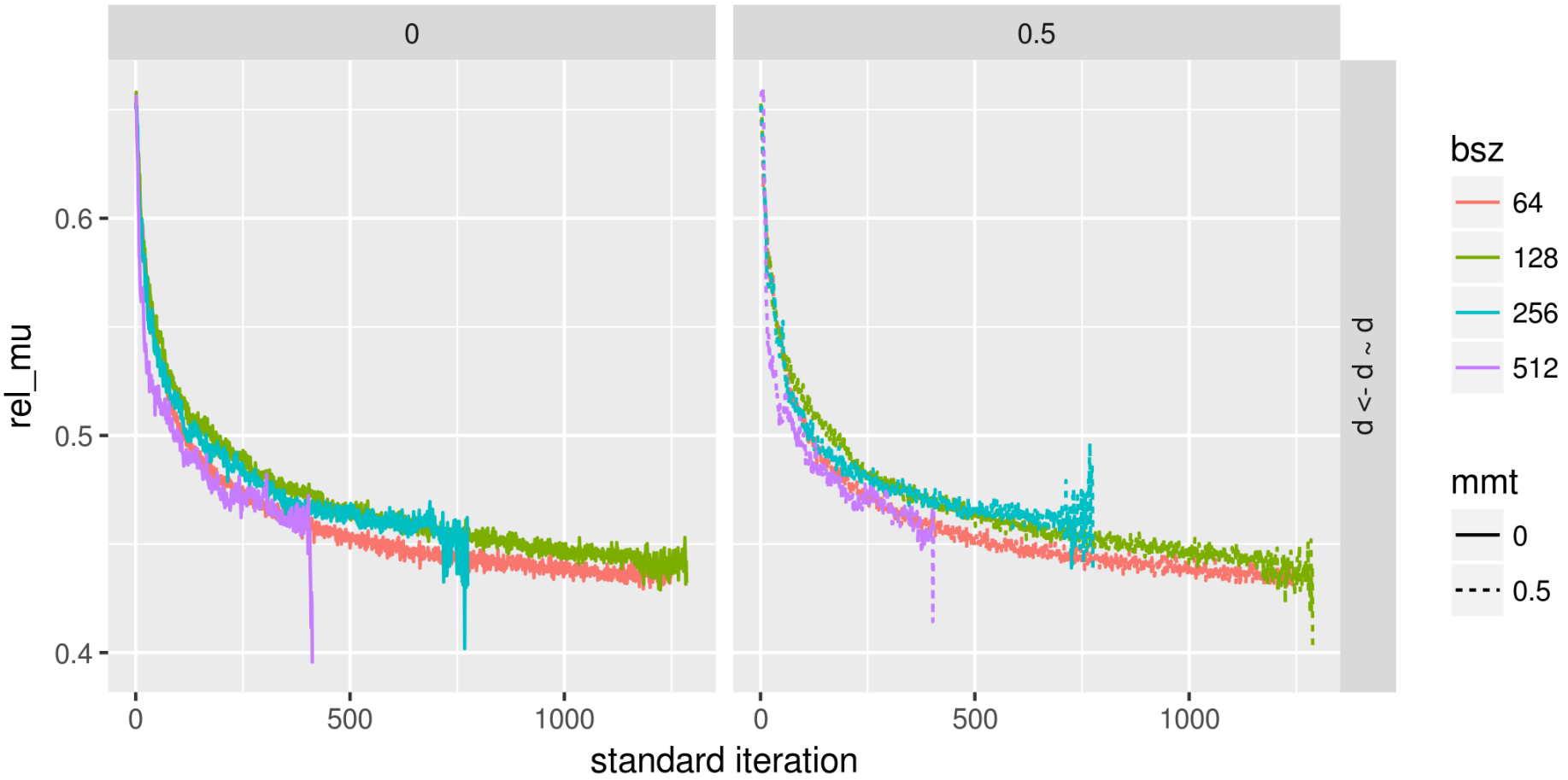
Performance by Time:



- mini-batch exhausts reserved iteration (1250) before reaching time limit (3h)

EXP 01: Tied Weight, *linear* ∼ *linear*|*linear*

Performance by Iteration:



- mini-batch exhausts reserved iteration (1250) before reaching time limit (3h)

EXP 01: Tied Weight, *linear* $\sim$ *linear*|*linear*

Finishing Rate:

	bsz	mmt	rep
1	64.00	0.00	98
2	64.00	0.50	98
3	128.00	0.00	93
4	128.00	0.50	89
5	256.00	0.00	84
6	256.00	0.50	86
7	512.00	0.00	78
8	512.00	0.50	88

- numerical stability improves over reduction in batch size

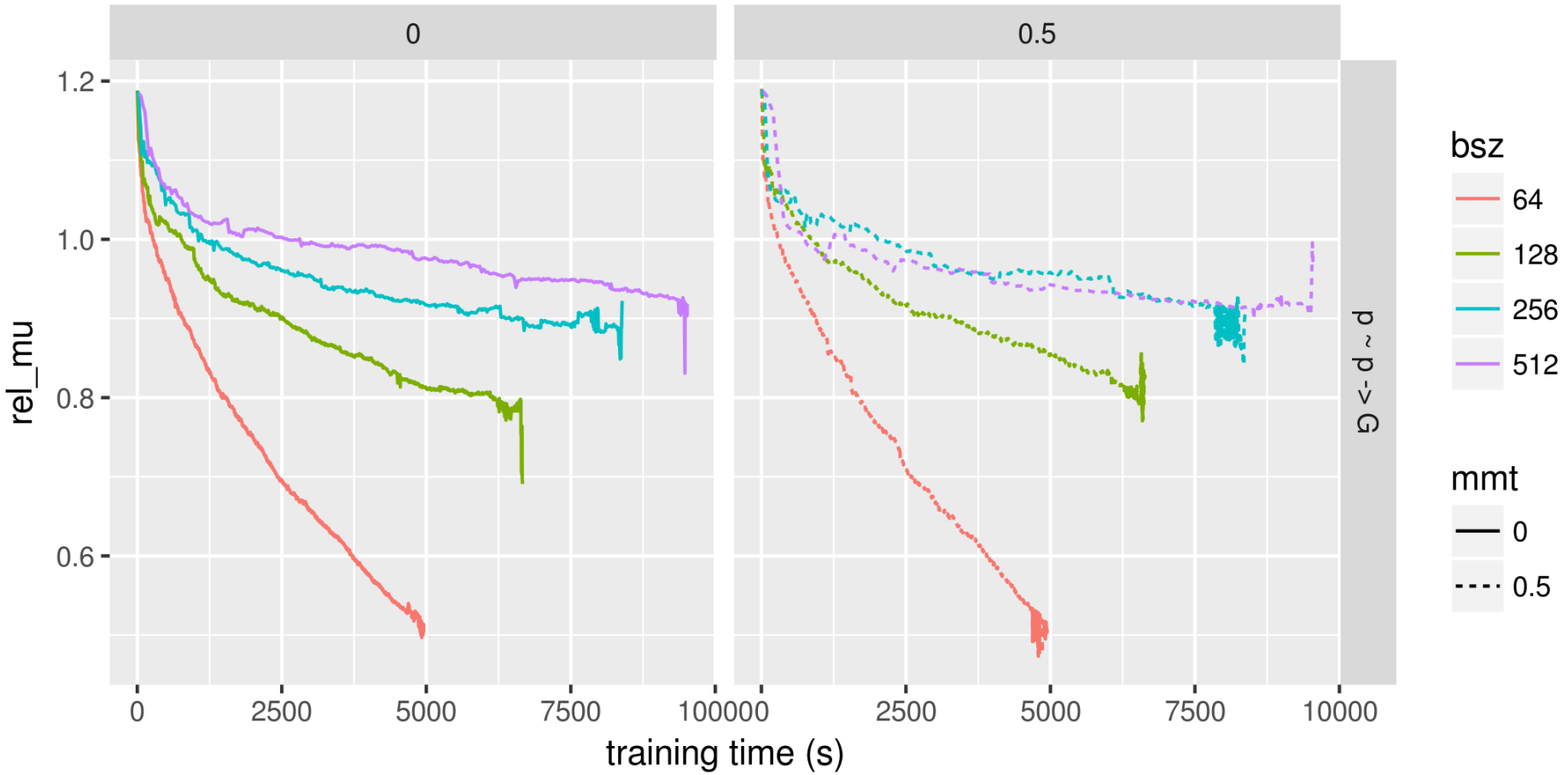
EXP 02: Tied Weight, *linear* $\sim$ *linear*|*Gaussian*

Simulation Settings:

- N = 512, P=2000, FRQ=0.50, ϕ=1.00
- Number of Hidden Units = 256
- True Model: $y \sim \mathcal{N}(0, X'X + \phi I)$
- Batch Size = {512, 256, 128, 64}
- learning rate = 1e-5, 2e-5, 4e-5, 8e-5
- Momentum = {0.0, 0.5}
- Wall Time = 1e4 (s)
- Reserved Iteration = 1250
- No rendom seed.

EXP 02: Tied Weight, *linear* $\sim$ *linear*|*Gaussian*

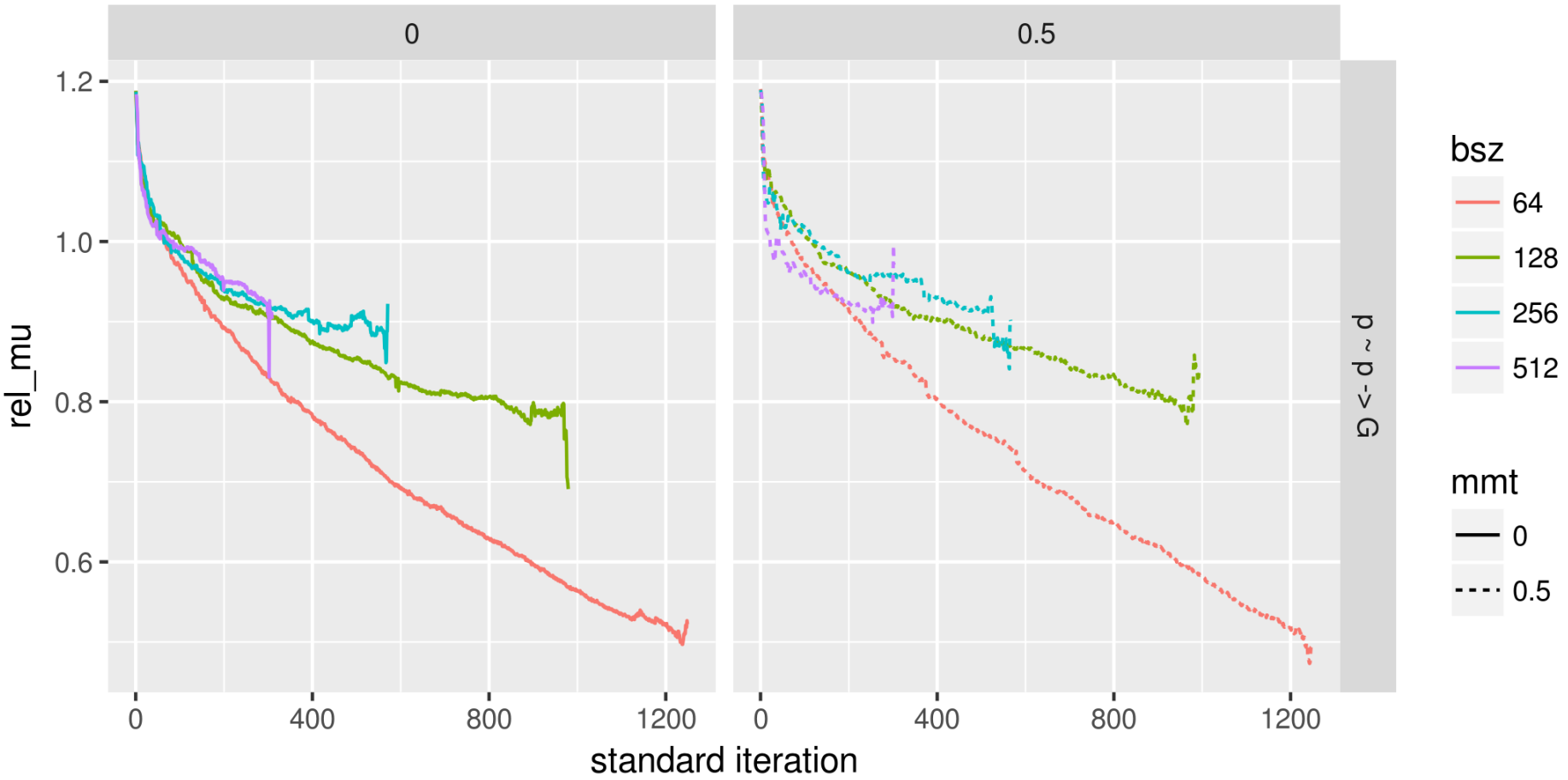
Performance by Time:



- mini-batch exhausts reserved iteration (1250) before reaching time limit (3h)

EXP 02: Tied Weight, *linear* $\sim$ *linear*$|$*Gaussian*

Performance by Iteration:



- mini-batch exhausts reserved iteration (1250) before reaching time limit (3h)

EXP 02: Tied Weight, *linear* $\sim$ *linear*|*Gaussian*

Finishing Rate:

	bsz	mmt	rep
1	64.00	0.00	88
2	64.00	0.50	84
3	128.00	0.00	81
4	128.00	0.50	82
5	256.00	0.00	80
6	256.00	0.50	85
7	512.00	0.00	80
8	512.00	0.50	88

- batches of intermediate size performed worst;
- by user time, batch out performed whole data;
- small batch coupled with large number of hidden units improve numerical stability.
- for Gaussian inner kernel
  - non-standardized Gaussian kernel resemble the Identity kernel;
  - it artificially boost the performance of KDNN;
  - the issue seems persist even after standardization.