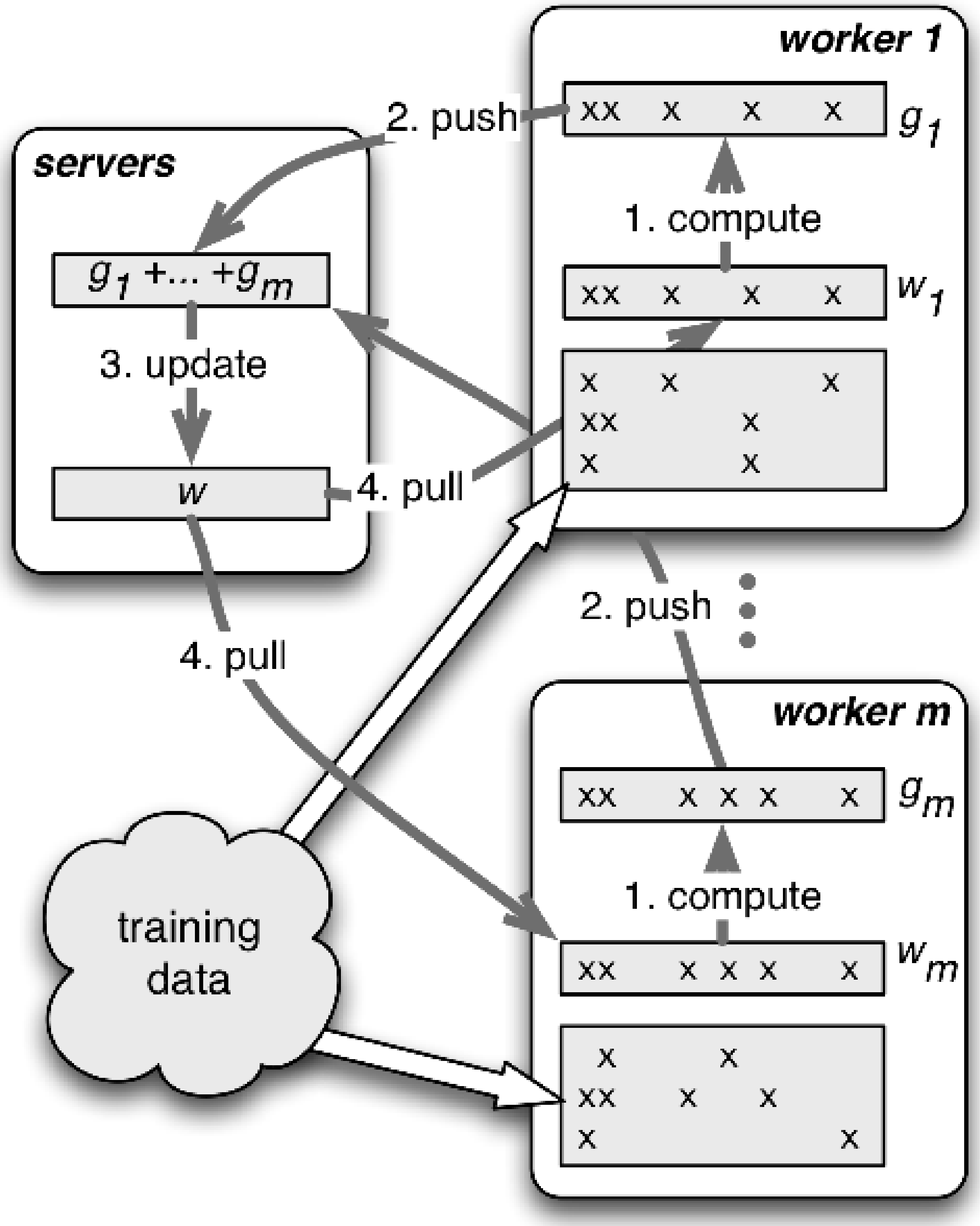
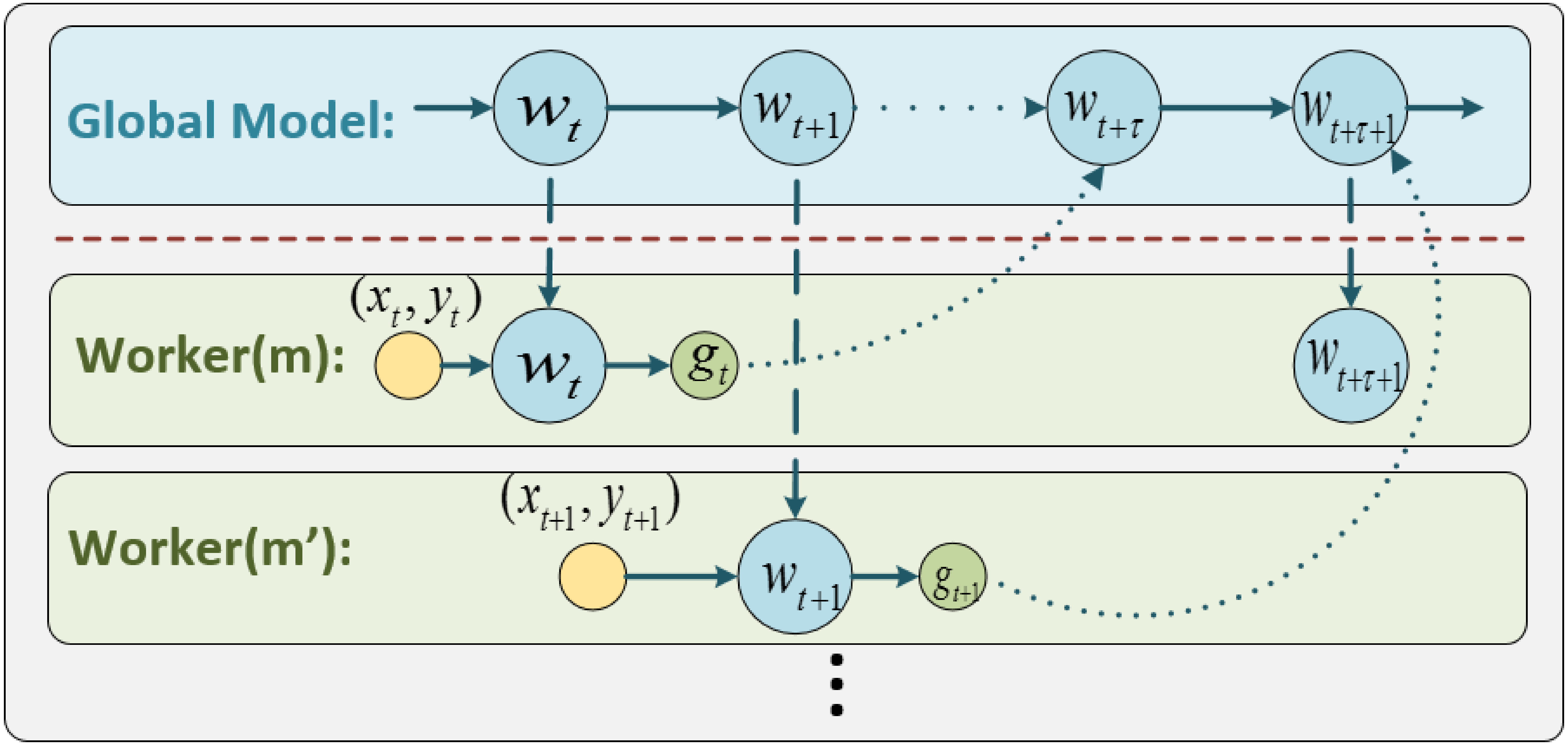
Bayesian Distributed Stochastic Gradient Descent

$$w = w - \eta \nabla Q(w) = w - \eta \sum_{i=1}^{\infty} \nabla Q_i(w)/n$$

Parameter Server

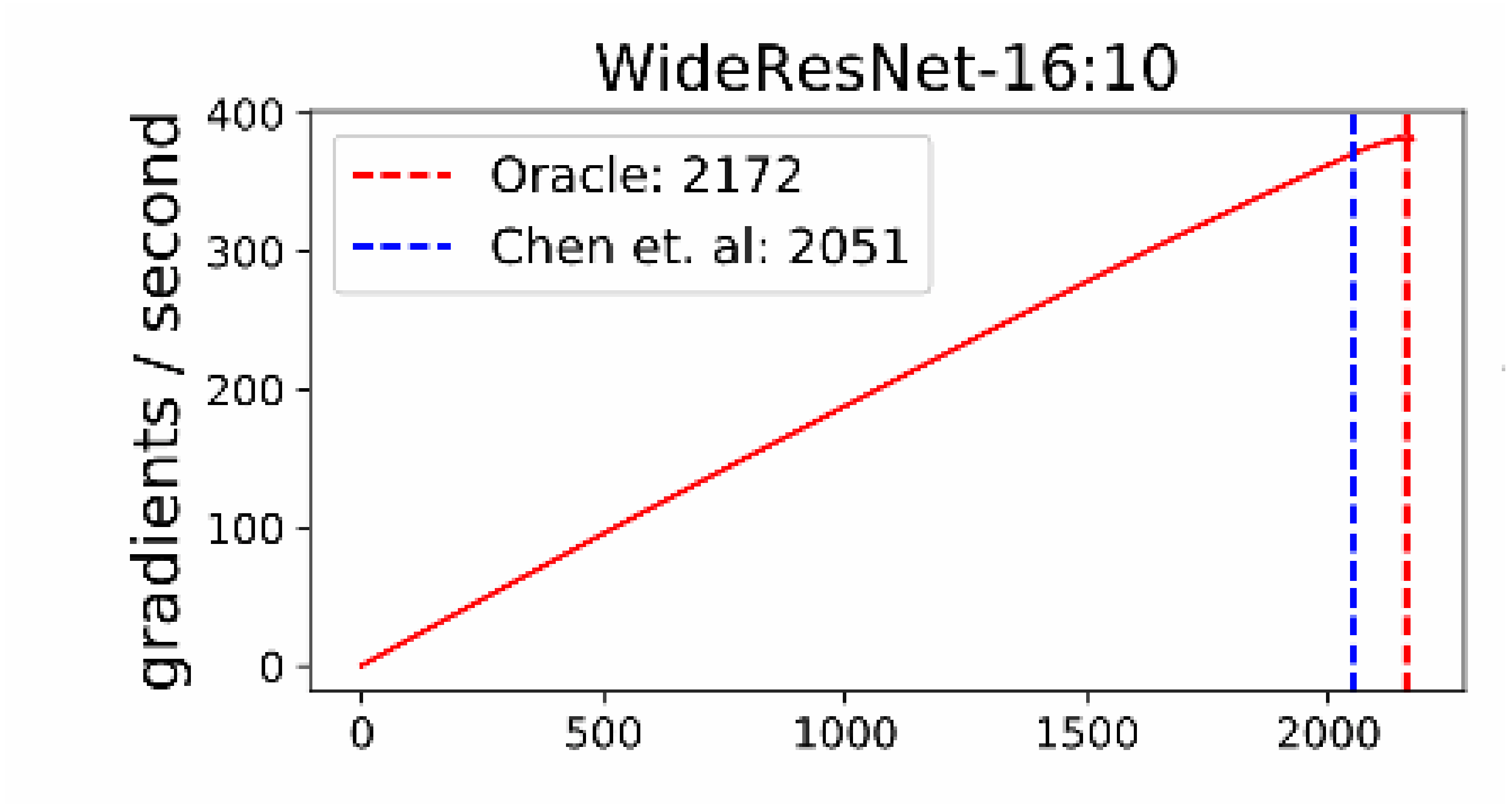


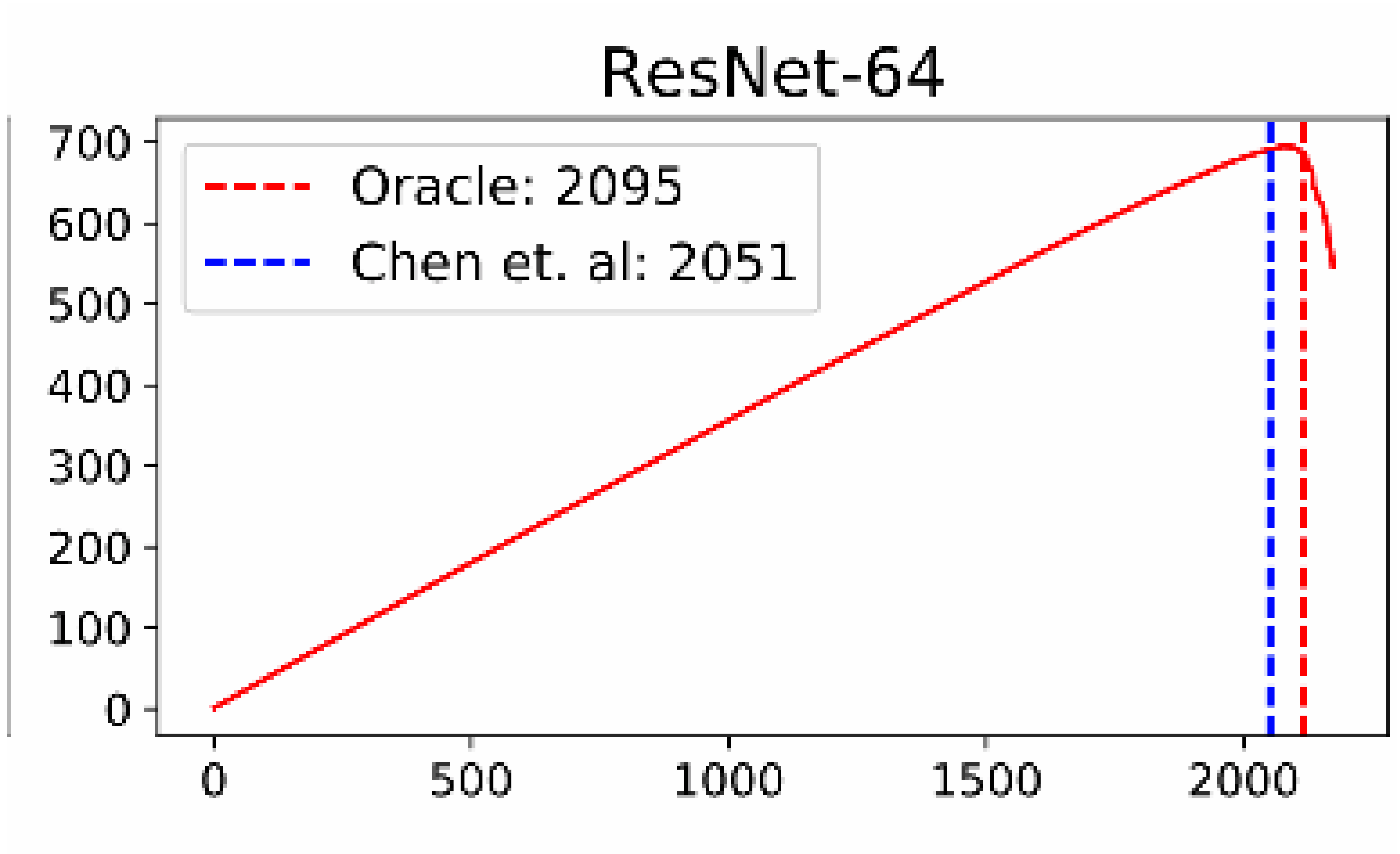
Async SGD or Stale Gradients

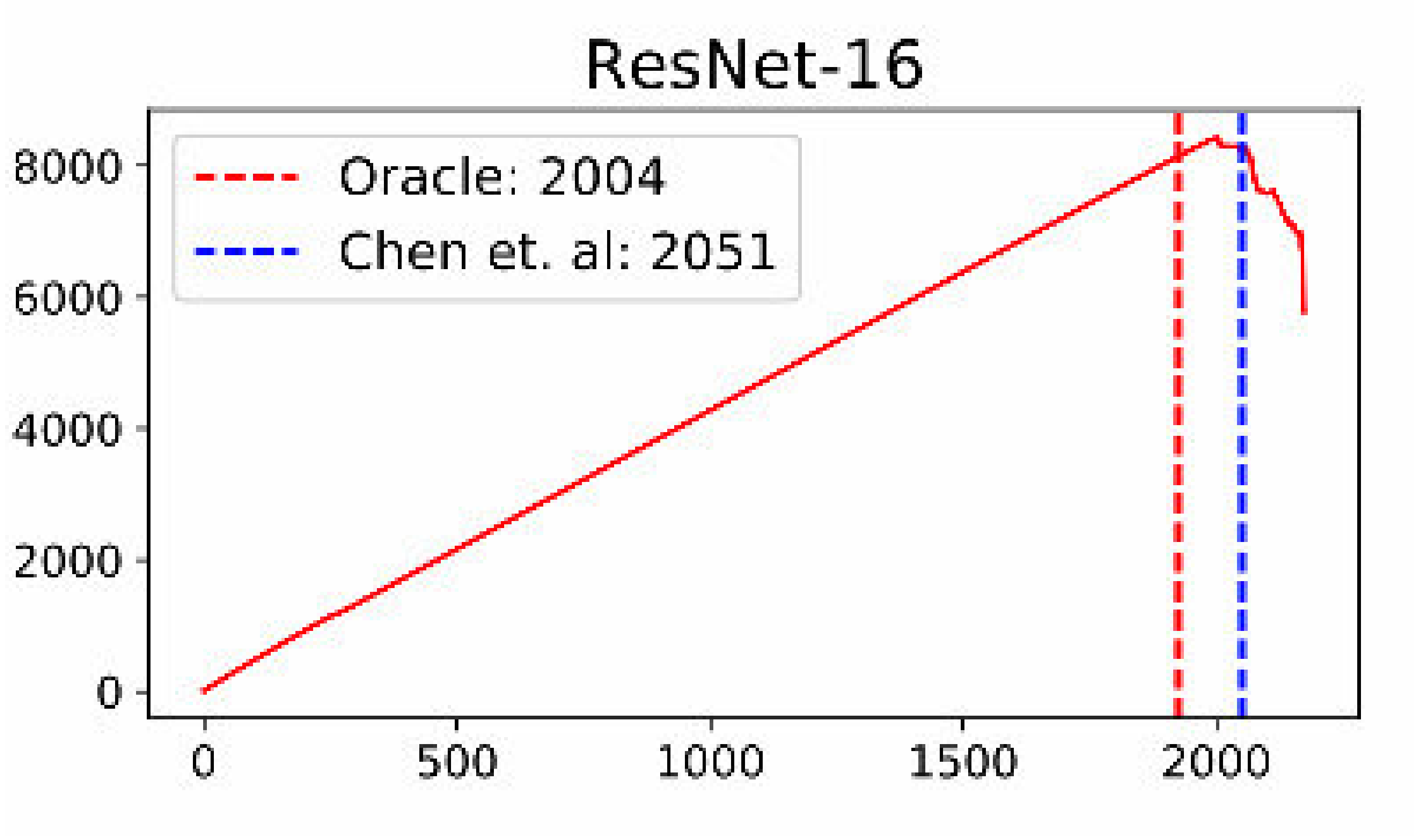


Sync SGD or Straggled Workerks

Why authers decided to write this paper







Metric to optimize

$$\Omega(c) = \frac{c}{\tilde{x}_{(c)}}$$

$$\tilde{x}_{(1)}, \tilde{x}_{(2)}, \ldots, \tilde{x}_{(n)}$$

$$\arg\max_{c} \Omega(c) = \arg\max_{c} \mathbb{E}[\frac{c}{\tilde{x}_{(c)}}]$$

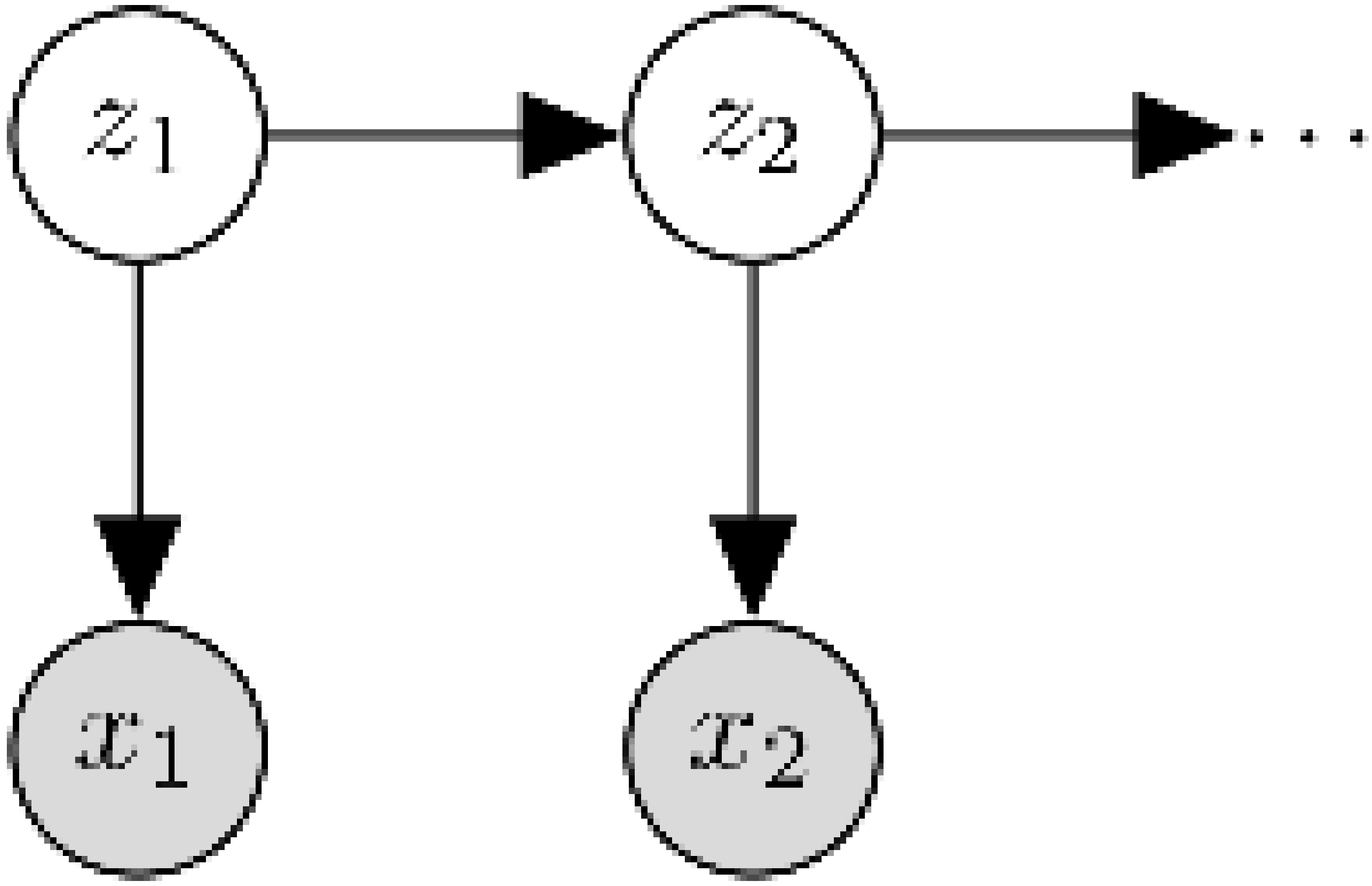
Elfving Cutoff

$$p(\tilde{x}_{(j)}) = Z(n,j) \int_{-\infty}^{\infty} x[\Phi(x)]^{j-1}[1-\Phi(x)]^{n-j} p(x)dx$$

$$\mathbb{E}[\tilde{x}_{(j)}] \approx \mu_t + \Phi^{-1}\left(\frac{n-\frac{\pi}{8}}{j-\frac{\pi}{4}+1}; 0, 1\right)\sigma_t$$

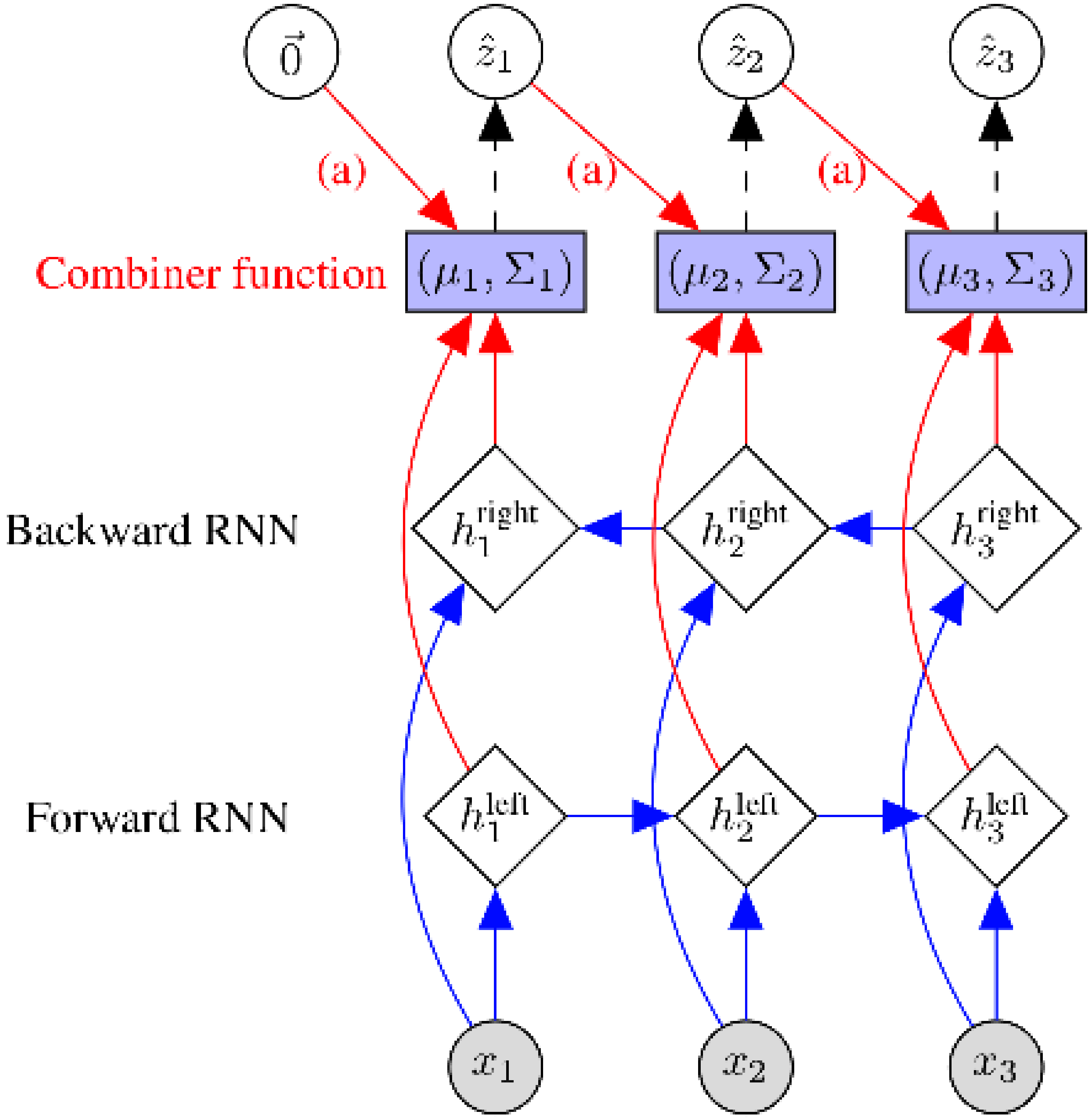
Structured Inference Networks for Nonlinear Gaussian State Space

Gaussian State Space and Deep Markov Models



$$z_t \sim \mathcal{N}(G_\alpha(z_{t-1}, \Delta_t), S_\beta(z_{t-1}, \Delta_t)) \quad \textit{(Transition)}$$
$$x_t \sim \Pi(F_\kappa(z_t)) \quad \textit{(Emission)}$$

Structured Inference Network



Back to BDSGD

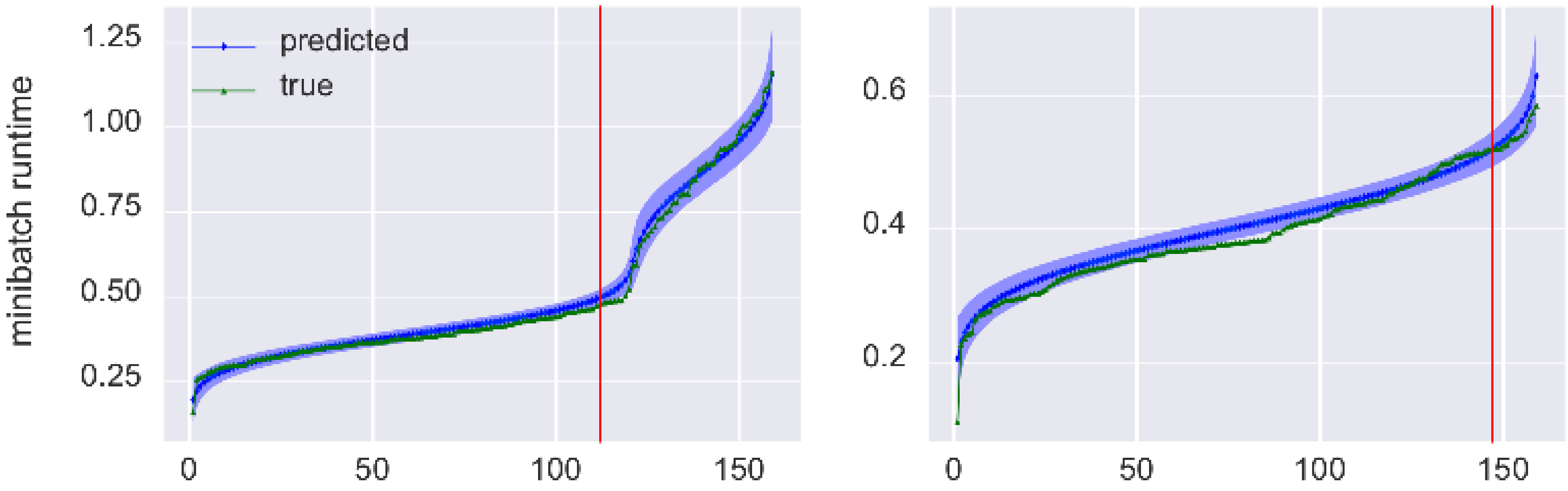
$$p_\theta(\boldsymbol{x}_{T-\ell:T}, \boldsymbol{z}_{T-\ell:T}) = \prod_{i=T-\ell}^{T} p_\theta(\boldsymbol{z}_i|\boldsymbol{z}_{i-1}) \prod_{i=T-\ell}^{T} p_\theta(\boldsymbol{x}_i|\boldsymbol{z}_i)$$

$$p(\boldsymbol{x}_{T+1}|\boldsymbol{x}_{T-\ell:T}) = \int p_\theta(\boldsymbol{x}_{T+1}|\boldsymbol{z}_{T+1}) p_\theta(\boldsymbol{z}_{T+1}|\boldsymbol{z}_T) p(\boldsymbol{z}_{T-\ell:T}|\boldsymbol{x}_{T-\ell:T}) d\boldsymbol{z}_{T-\ell:T+1}$$

$$\text{ELBO} = \mathbb{E}_{q_\phi(\boldsymbol{z}_{T-\ell:t}|\boldsymbol{x}_{T-\ell:T})} \log \left(\frac{p_\theta(\boldsymbol{x}_{T-\ell:t}, \boldsymbol{z}_{T-\ell:t})}{q_\phi(\boldsymbol{z}_{T-\ell:t}|\boldsymbol{x}_{T-\ell:T})} \right)$$

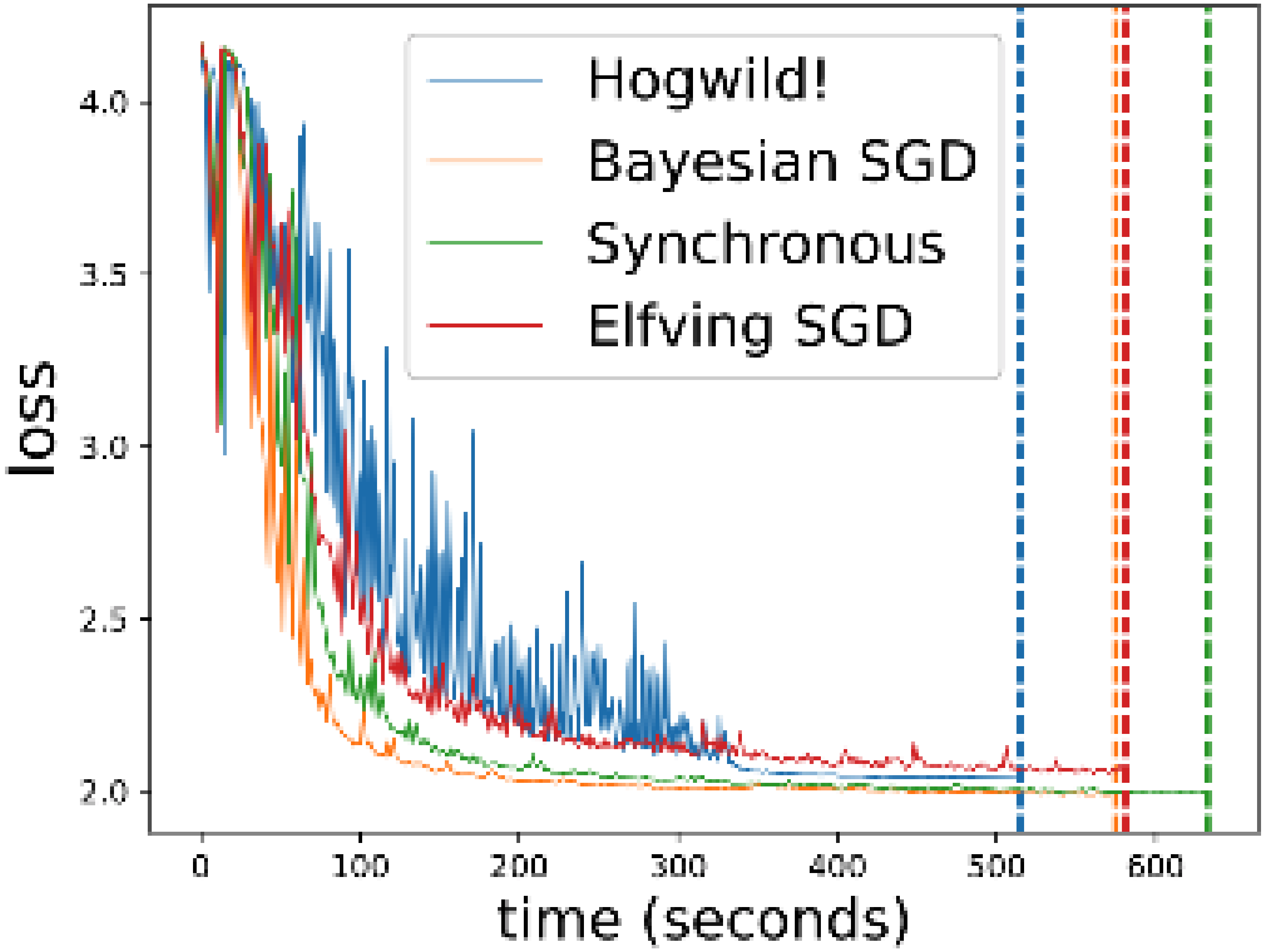
$$q_\phi(\boldsymbol{z}_{T-\ell:t}|\boldsymbol{x}_{T-\ell:T}) = \prod_{t=T-\ell}^{T} q_\phi(\boldsymbol{z}_t|\boldsymbol{z}_{T-\ell:t}, \boldsymbol{x}_{T-\ell:T}).$$

Results



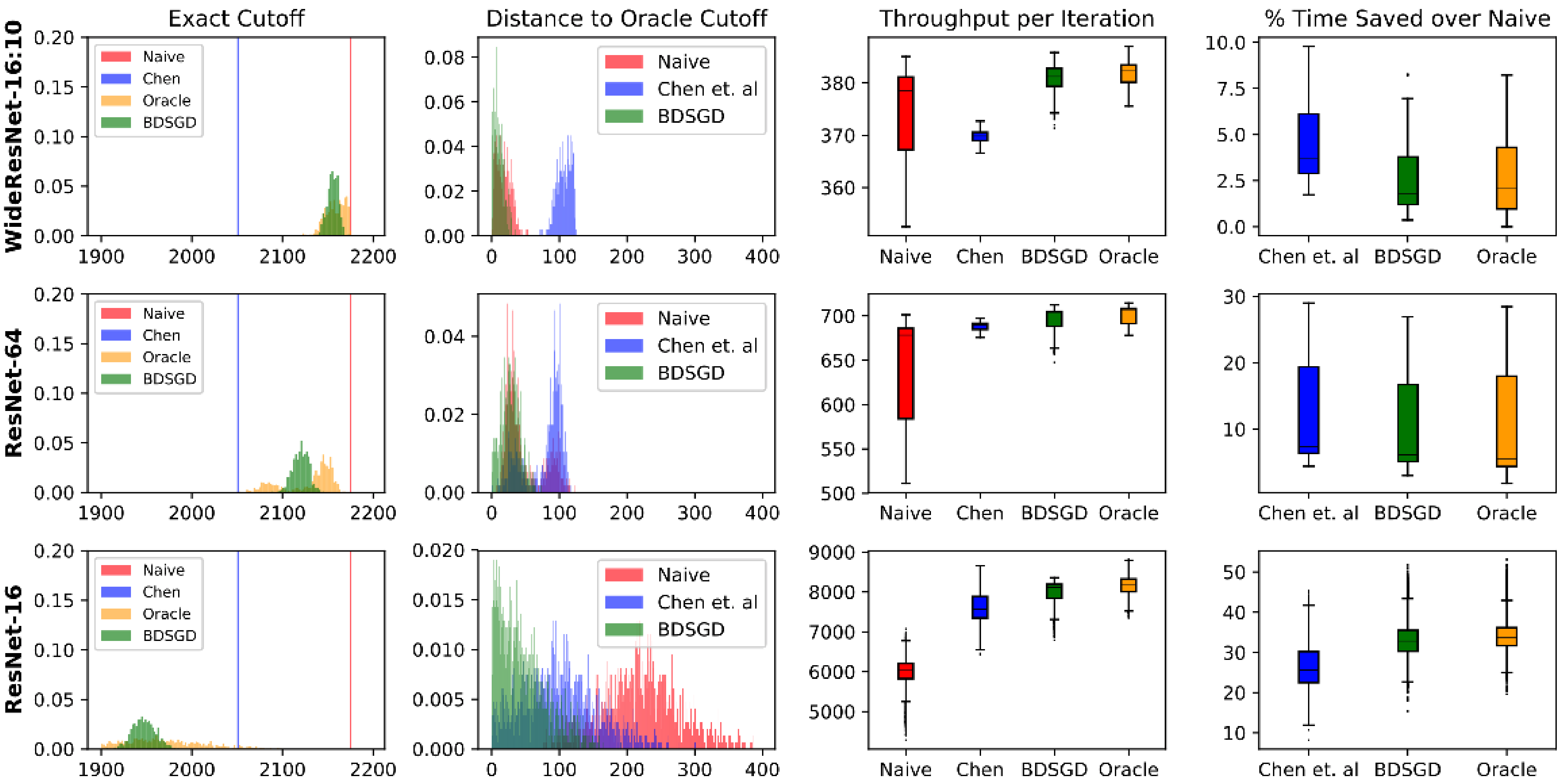
(a) SGD runtime profiles for two iterations

Results

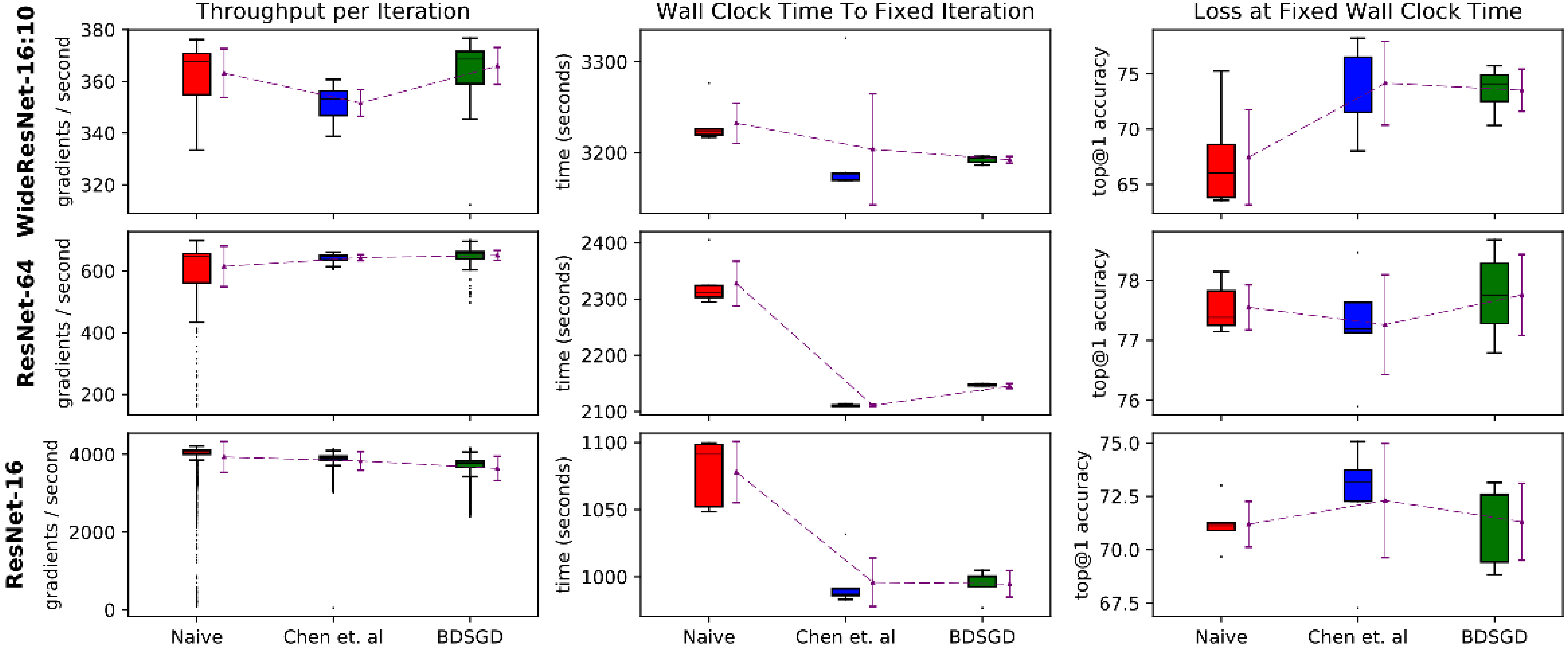


(b) MNIST full training run

Results


Exact Cutoff
Distance to Oracle Cutoff
Throughput per Iteration
% Time Saved over Naive
WideResNet-16:10
ResNet-64
ResNet-16
Naive
Chen
Oracle
BDSGD
Chen et. al
0.20
0.15
0.10
0.05
0.00
1900
2000
2100
2200
0.08
0.06
0.04
0.02
0
100
200
300
400
380
370
360
10.0
7.5
5.0
2.5
0.0
700
650
600
550
500
30
20
10
0.020
0.015
0.010
0.005
0.000
9000
8000
7000
6000
5000
50
40

Results


Throughput per Iteration
Wall Clock Time To Fixed Iteration
Loss at Fixed Wall Clock Time
WideResNet-16:10
ResNet-64
ResNet-16
gradients / second
time (seconds)
top@1 accuracy
380
360
340
320
3300
3200
75
70
65
600
400
200
2400
2300
2200
2100
78
77
76
4000
2000
0
1100
1050
1000
75.0
72.5
70.0
67.5
Naive
Chen et. al
BDSGD

Interesting Facts

VAE inference is also called *amortised inference*

Authers use DMM for predicting straggled workers time

Papers

BDSGD

https://papers.nips.cc/paper/7874-bayesian-distributed-stochastic-gradient-descent

Structured Inference Network

https://arxiv.org/abs/1609.09869