

# Low-Precision Stochastic Gradient Langevin Dynamics

Ruqi Zhang

UT Austin/Purdue

Andrew Gordon Wilson

New York University

Christopher De Sa

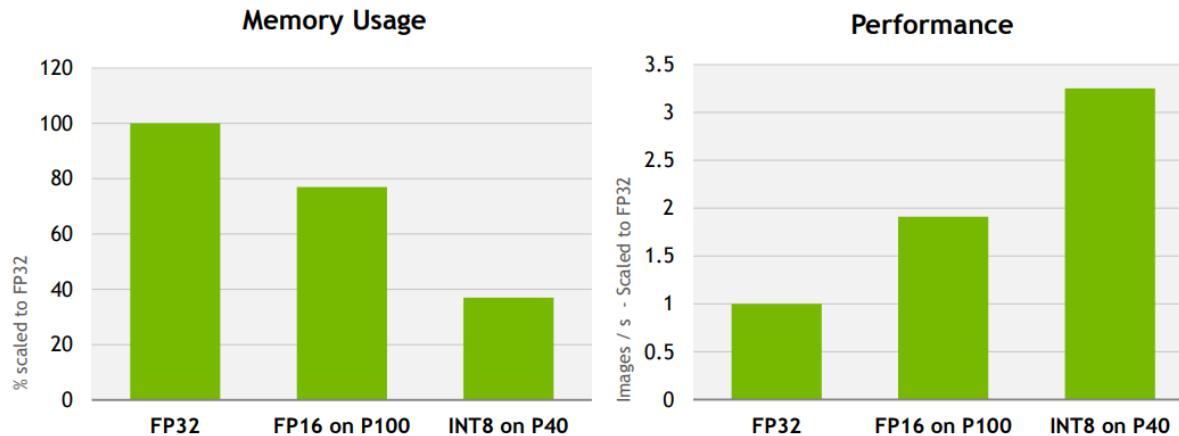
Cornell University

# Low-Precision Optimization

- Use fewer bits to represent numbers, e.g. 8 bits

# Low-Precision Optimization

- Use fewer bits to represent numbers, e.g. 8 bits
- Significantly reduce **memory** and **latency** consumption



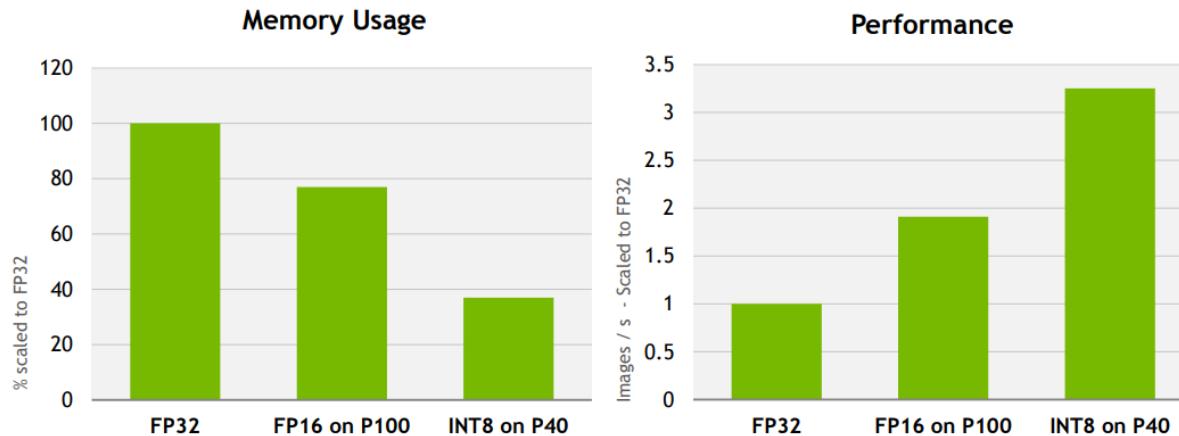
ResNet50 Model, Batch Size = 128, TensorRT 2.1 RC prerelease

[developer.nvidia.com/tensorrt](http://developer.nvidia.com/tensorrt)

19 NVIDIA

# Low-Precision Optimization

- Use fewer bits to represent numbers, e.g. 8 bits
- Significantly reduce **memory** and **latency** consumption



ResNet50 Model, Batch Size = 128, TensorRT 2.1 RC prerelease  
[developer.nvidia.com/tensorrt](http://developer.nvidia.com/tensorrt)

19 NVIDIA

- **Supported** on new emerging chips including CPUs, GPUs, TPUs

**Low-precision **sampling** remains largely unexplored**

# Low-precision **sampling** remains largely unexplored

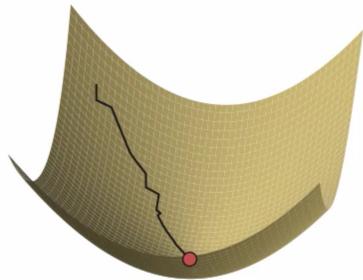
Stochastic gradient decent (SGD)

$$\theta_{k+1} = \theta_k - \alpha \nabla \tilde{U}(\theta_k)$$

# Low-precision **sampling** remains largely unexplored

Stochastic gradient decent (SGD)

$$\theta_{k+1} = \theta_k - \alpha \nabla \tilde{U}(\theta_k)$$



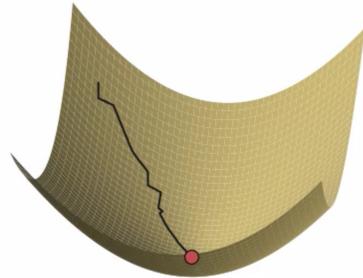
# Low-precision sampling remains largely unexplored

Stochastic gradient decent (SGD)   Stochastic gradient Langevin dynamics (SGLD)

$$\theta_{k+1} = \theta_k - \alpha \nabla \tilde{U}(\theta_k)$$

$$\theta_{k+1} = \theta_k - \alpha \nabla \tilde{U}(\theta_k) + \sqrt{2\alpha} \xi_{k+1}$$

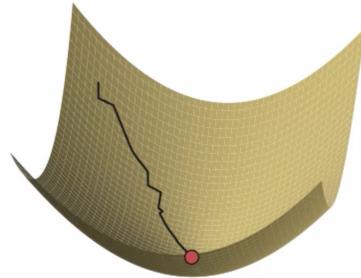
where  $\xi_{k+1} \sim \mathcal{N}(0, I)$



# Low-precision sampling remains largely unexplored

Stochastic gradient decent (SGD)

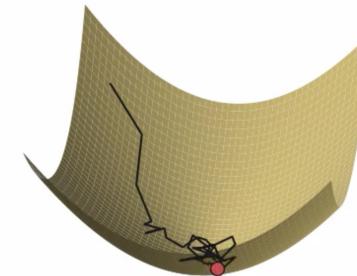
$$\theta_{k+1} = \theta_k - \alpha \nabla \tilde{U}(\theta_k)$$



Stochastic gradient Langevin dynamics (SGLD)

$$\theta_{k+1} = \theta_k - \alpha \nabla \tilde{U}(\theta_k) + \sqrt{2\alpha} \xi_{k+1}$$

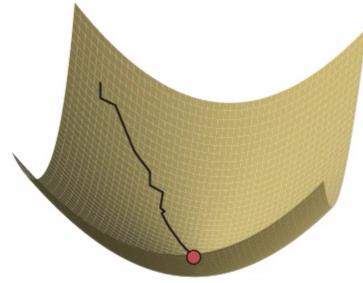
where  $\xi_{k+1} \sim \mathcal{N}(0, I)$



# Low-precision sampling remains largely unexplored

Stochastic gradient decent (SGD)

$$\theta_{k+1} = \theta_k - \alpha \nabla \tilde{U}(\theta_k)$$



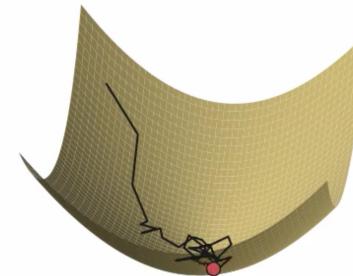
Pros of sampling:

- ✓ Characterize complex and multi-modal DNN posteriors

Stochastic gradient Langevin dynamics (SGLD)

$$\theta_{k+1} = \theta_k - \alpha \nabla \tilde{U}(\theta_k) + \sqrt{2\alpha} \xi_{k+1}$$

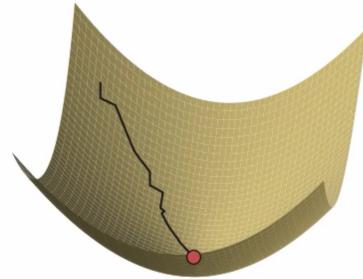
where  $\xi_{k+1} \sim \mathcal{N}(0, I)$



# Low-precision sampling remains largely unexplored

Stochastic gradient decent (SGD)

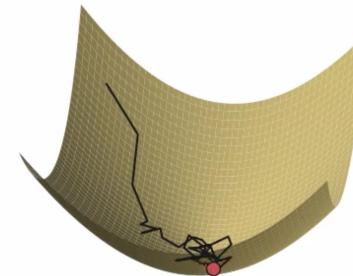
$$\theta_{k+1} = \theta_k - \alpha \nabla \tilde{U}(\theta_k)$$



Stochastic gradient Langevin dynamics (SGLD)

$$\theta_{k+1} = \theta_k - \alpha \nabla \tilde{U}(\theta_k) + \sqrt{2\alpha} \xi_{k+1}$$

where  $\xi_{k+1} \sim \mathcal{N}(0, I)$



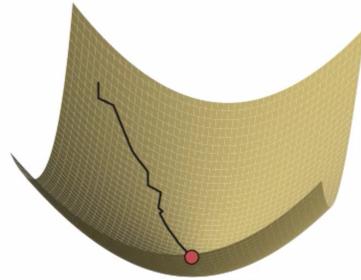
Pros of sampling:

- ✓ Characterize **complex** and **multi-modal** DNN posteriors
- ✓ Provide state-of-the-art generalization **accuracy** and **calibration**

# Low-precision sampling remains largely unexplored

Stochastic gradient decent (SGD)

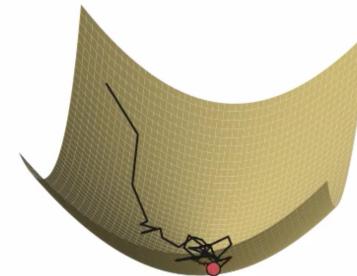
$$\theta_{k+1} = \theta_k - \alpha \nabla \tilde{U}(\theta_k)$$



Stochastic gradient Langevin dynamics (SGLD)

$$\theta_{k+1} = \theta_k - \alpha \nabla \tilde{U}(\theta_k) + \sqrt{2\alpha} \xi_{k+1}$$

where  $\xi_{k+1} \sim \mathcal{N}(0, I)$



Pros of sampling:

- ✓ Characterize **complex** and **multi-modal** DNN posteriors
- ✓ Provide state-of-the-art generalization **accuracy** and **calibration**
- ✓ **Robust to system noise, especially suited for low-precision!**

# Low-precision sampling remains largely unexplored

Stochastic gradient decent (SGD)

$$\theta_{k+1} = \theta_k - \alpha \nabla \tilde{U}(\theta_k)$$

Stochastic gradient Langevin dynamics (SGLD)

$$\theta_{k+1} = \theta_k - \alpha \nabla \tilde{U}(\theta_k) + \sqrt{2\alpha} \xi_{k+1}$$

where  $\xi_{k+1} \sim \mathcal{N}(0, I)$

Goal: a comprehensive study for low-precision SGLD

Pros of sampling:

- ✓ Characterize complex and multi-modal DNN posteriors
- ✓ Provide state-of-the-art generalization accuracy and calibration
- ✓ Robust to system noise, especially suited for low-precision!

# SGLD with full-precision gradient accumulators

SGLDLP-F: a full-precision weight buffer to accumulate gradient updates

$$\theta_{k+1} = \theta_k - \alpha Q_G \left( \nabla \tilde{U}(Q_W(\theta_k)) \right) + \sqrt{2\alpha} \xi_{k+1}$$

# SGLD with full-precision gradient accumulators

SGLDLP-F: a full-precision weight buffer to accumulate gradient updates

$$\theta_{k+1} = \theta_k - \alpha Q_G \left( \nabla \tilde{U}(Q_W(\theta_k)) \right) + \sqrt{2\alpha} \xi_{k+1}$$



**Theorem** (informal): On strongly log-concave distributions, SGLDLP-F converges under the 2-Wasserstein distance to the target distribution

# SGLD with full-precision gradient accumulators

SGLDLP-F: a full-precision weight buffer to accumulate gradient updates

$$\theta_{k+1} = \theta_k - \alpha Q_G \left( \nabla \tilde{U}(Q_W(\theta_k)) \right) + \sqrt{2\alpha} \xi_{k+1}$$



**Theorem** (informal): On **strongly log-concave** distributions, *SGLDLP-F converges under the 2-Wasserstein distance to the target distribution*

## Takeaway

- SGLDLP-F is **convergent** and can be **safely** used in practice
- SGLDLP-F is more **robust** to the quantization error than its SGD counterpart

# SGLD with low-precision gradient accumulators

SGLDLP-L: the weight is always represented in low-precision

$$\theta_{k+1} = Q_W \left( \theta_k - \alpha Q_G \left( \nabla \tilde{U}(\theta_k) \right) + \sqrt{2\alpha} \xi_{k+1} \right)$$

# SGLD with low-precision gradient accumulators

SGLDLP-L: the weight is always represented in low-precision

$$\theta_{k+1} = Q_W \left( \theta_k - \alpha Q_G \left( \nabla \tilde{U}(\theta_k) \right) + \sqrt{2\alpha} \xi_{k+1} \right)$$

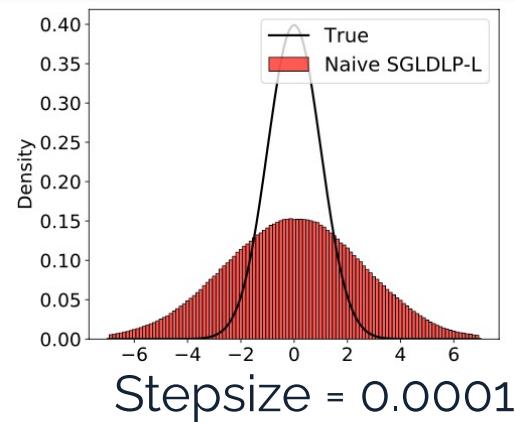
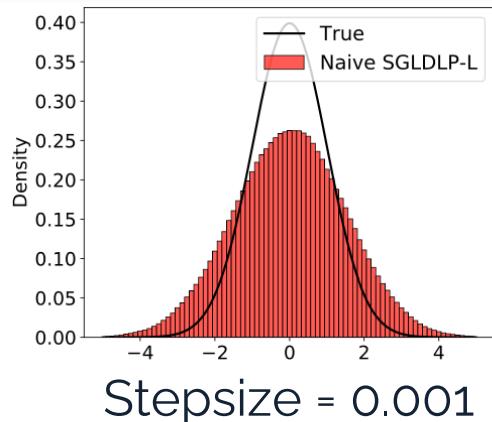
 **Theorem** (informal): As the stepsize **decreases**, the  $\text{W}_2$  distance between the stationary distribution of SGLDLP-L and the target distribution may **increase**

# SGLD with low-precision gradient accumulators

SGLDLP-L: the weight is always represented in low-precision

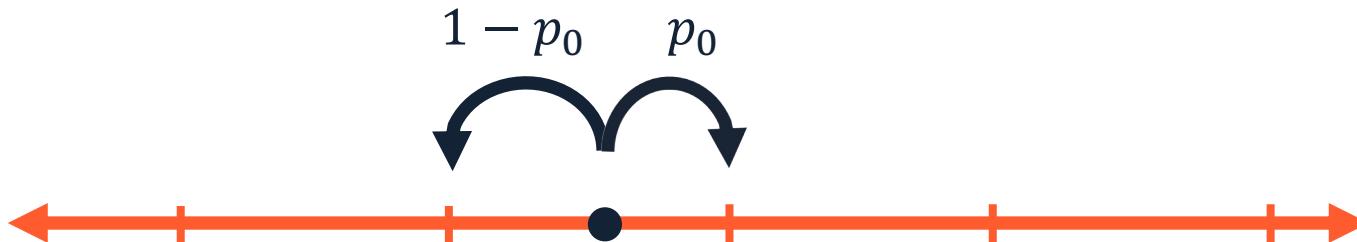
$$\theta_{k+1} = Q_W \left( \theta_k - \alpha Q_G \left( \nabla \tilde{U}(\theta_k) \right) + \sqrt{2\alpha} \xi_{k+1} \right)$$

 **Theorem** (informal): As the stepsize **decreases**, the  $\text{W}_2$  distance between the stationary distribution of SGLDLP-L and the target distribution may **increase**



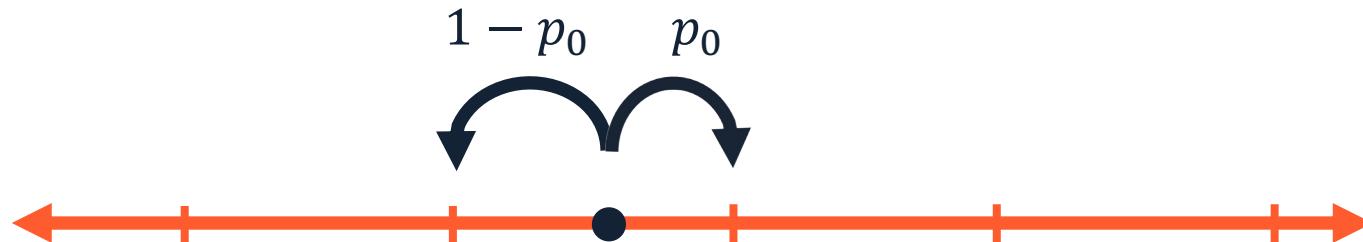
# Variance-Corrected Quantization Function

- Reason: wrong variance of  $\theta_{k+1} = Q_W \left( \theta_k - \alpha Q_G \left( \nabla \tilde{U}(\theta_k) \right) + \sqrt{2\alpha} \xi_{k+1} \right)$

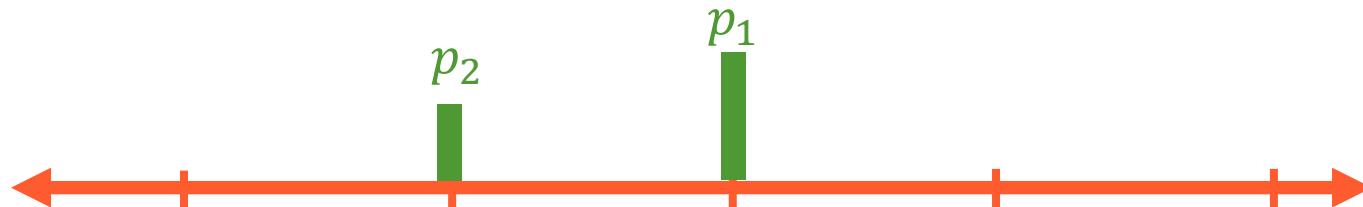


# Variance-Corrected Quantization Function

- Reason: wrong variance of  $\theta_{k+1} = Q_W \left( \theta_k - \alpha Q_G \left( \nabla \tilde{U}(\theta_k) \right) + \sqrt{2\alpha} \xi_{k+1} \right)$



- Our solution: new quantization function to keep correct mean and variance

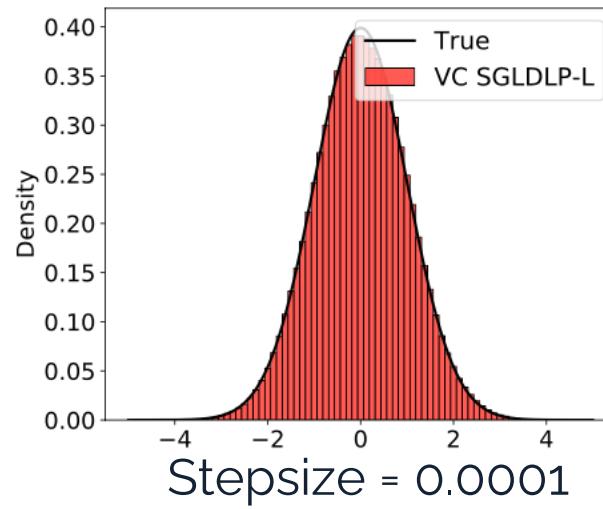
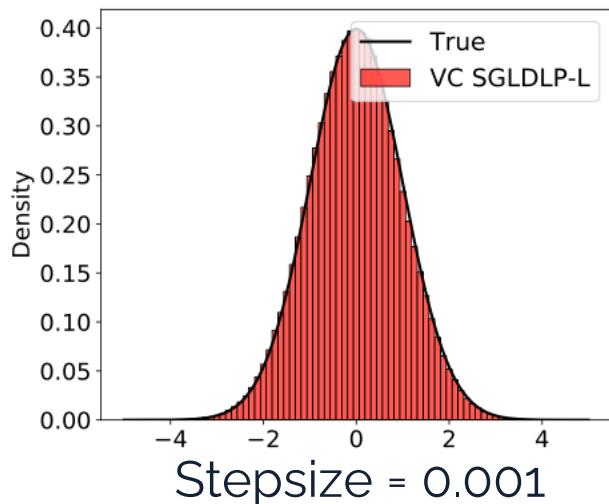


# Variance-Corrected SGLD (VC SGLDLP-L)

 **Theorem** (informal): On **strongly log-concave** distributions, VC SGLDLP-L converges under the ***2-Wasserstein distance*** to the target distribution

# Variance-Corrected SGLD (VC SGLDLP-L)

😊 **Theorem** (informal): On **strongly log-concave** distributions, VC SGLDLP-L converges under the ***2-Wasserstein distance*** to the target distribution



# Experimental Results: Prediction and Calibration

Table 1. Test errors (%).

	CIFAR-10	CIFAR-100	IMDB
32-BIT FLOATING POINT			
SGLDFP	$4.65 \pm 0.06$	$22.58 \pm 0.18$	$13.43 \pm 0.21$
SGDFP	$4.71 \pm 0.02$	$22.64 \pm 0.13$	$13.88 \pm 0.29$
cSGLDFP	$4.54 \pm 0.05$	$21.63 \pm 0.04$	$13.25 \pm 0.18$
8-BIT FIXED POINT			
NAÏVE SGLDLP-L	$7.82 \pm 0.13$	$27.25 \pm 0.13$	$16.63 \pm 0.28$
VC SGLDLP-L	$7.13 \pm 0.01$	$26.62 \pm 0.16$	$15.38 \pm 0.27$
SGDLP-L	$8.53 \pm 0.08$	$28.86 \pm 0.10$	$19.28 \pm 0.63$
SGLDLP-F	$5.12 \pm 0.06$	$23.30 \pm 0.09$	$15.40 \pm 0.36$
SGDLP-F	$5.20 \pm 0.14$	$23.84 \pm 0.12$	$15.74 \pm 0.79$
8-BIT BLOCK FLOATING POINT			
NAÏVE SGLDLP-L	$5.85 \pm 0.04$	$26.38 \pm 0.13$	$14.64 \pm 0.08$
VC SGLDLP-L	$5.51 \pm 0.01$	$25.22 \pm 0.18$	$13.99 \pm 0.24$
SGDLP-L	$5.86 \pm 0.18$	$26.19 \pm 0.11$	$16.06 \pm 1.81$
SGLDLP-F	$4.58 \pm 0.07$	$22.59 \pm 0.18$	$14.05 \pm 0.33$
SGDLP-F	$4.75 \pm 0.05$	$22.9 \pm 0.13$	$14.28 \pm 0.17$
VC cSGLDLP-L	$4.97 \pm 0.10$	$22.61 \pm 0.12$	$13.09 \pm 0.27$
cSGLD-F	$4.32 \pm 0.07$	$21.50 \pm 0.14$	$13.13 \pm 0.37$

Table 2. ECE  $\downarrow$  (%).

	CIFAR-10	CIFAR-100
32-BIT FLOATING POINT		
SGLD	1.11	3.92
SGD	2.53	4.97
cSGLDFP	0.66	1.38
8-BIT FIXED POINT		
VC SGLDLP-L	0.6	3.19
SGLDLP-L	3.4	10.38
SGLDLP-F	1.12	4.42
SGDLP-F	3.05	6.80
8-BIT BLOCK FLOATING POINT		
VC SGLDLP-L	0.6	5.82
SGLDLP-L	4.23	12.97
SGLDLP-F	1.19	3.78
SGDLP-F	2.76	5.2
VC cSGLDLP-L	0.51	1.39
cSGLD-F	0.56	1.33

- Low-precision SGLD matches full-precision SGLD with only **8 bits**
- Significantly **outperforms** low-precision SGD

# Summary

- We give the **first** comprehensive study of low-precision SGLD

# Summary

- We give the **first** comprehensive study of low-precision SGLD
- We prove that **SGLDLP-F** is **convergent** and is more **robust** to quantization noise than SGD counterpart

# Summary

- We give the **first** comprehensive study of low-precision SGLD
- We prove that **SGLDLP-F** is **convergent** and is more **robust** to quantization noise than SGD counterpart
- We reveal the **issue** of naive **SGLDLP-L** and propose a **new** quantization function to fix it

# Summary

- We give the **first** comprehensive study of low-precision SGLD
- We prove that **SGLDLP-F** is **convergent** and is more **robust** to quantization noise than SGD counterpart
- We reveal the **issue** of naive **SGLDLP-L** and propose a **new** quantization function to fix it
- **Empirical** demonstration on Bayesian deep learning benchmarks

 arXiv.org <https://arxiv.org/abs/2206.09909>

 <https://github.com/ruqizhang/low-precision-sgld>