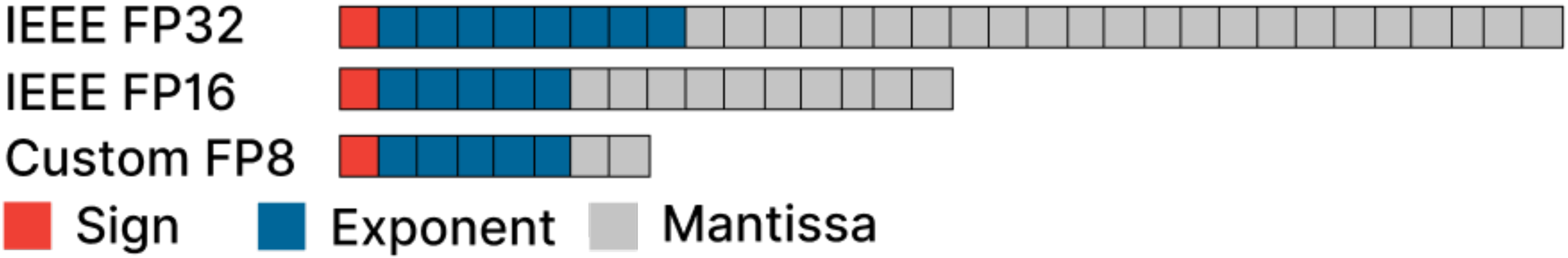
SWALP: Stochastic Weight Averaging in Low-Precision Training

Guandao Yang, Tianyi Zhang, Polina Kirichenko, Junwen Bai,
Andrew Gordon Wilson, Christopher De Sa

Low-precision Computation



IEEE FP32

IEEE FP16

Custom FP8

🟥 Sign 🟦 Exponent ⬜ Mantissa

Problem Statement

We study how to leverage low-precision training to obtain a high-accuracy model.

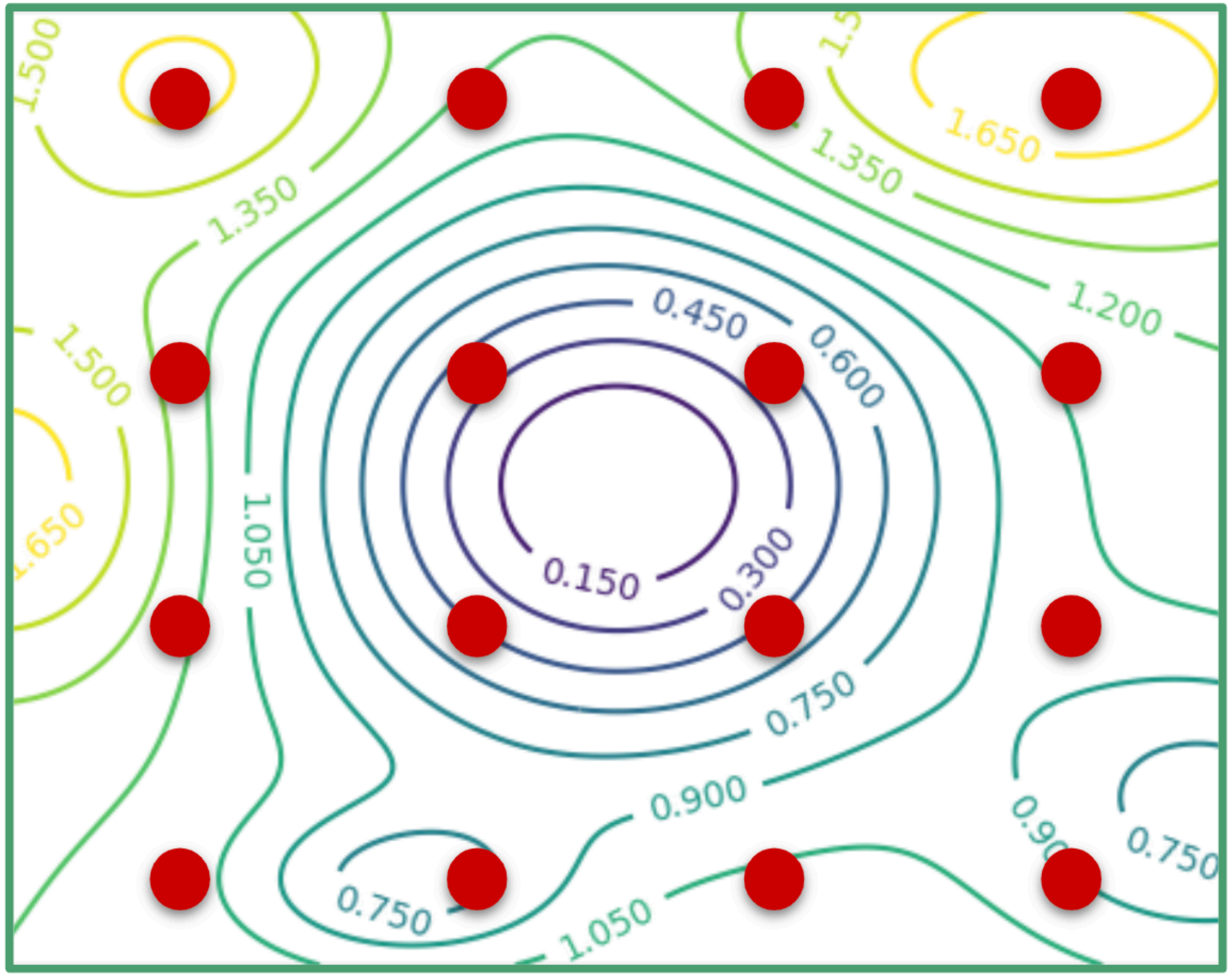
Problem Statement

We study how to leverage low-precision training to obtain a high-accuracy model.

⇧

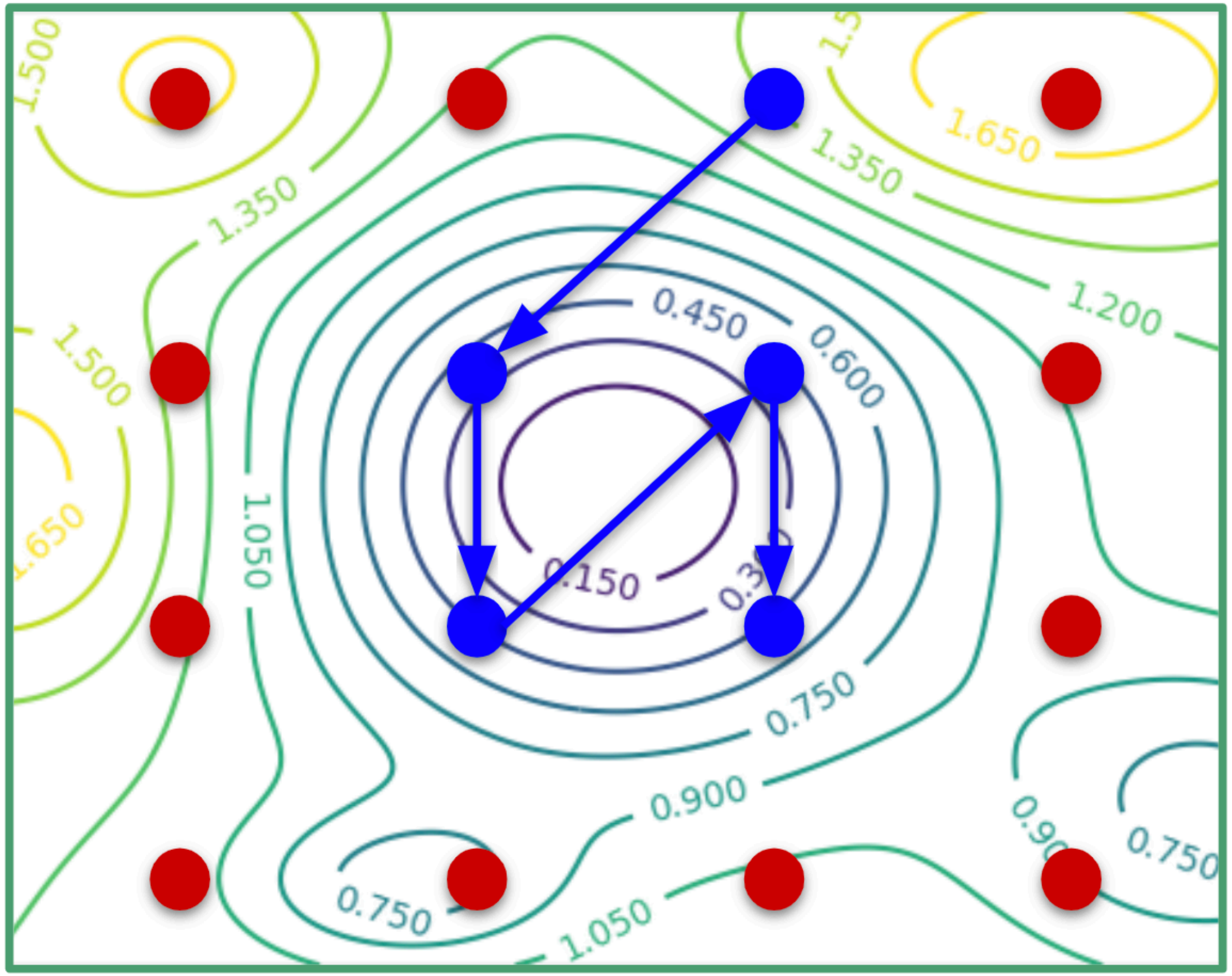
Output model can be higher-precision.

Low-precision Training



● Representable Points in Low Precision

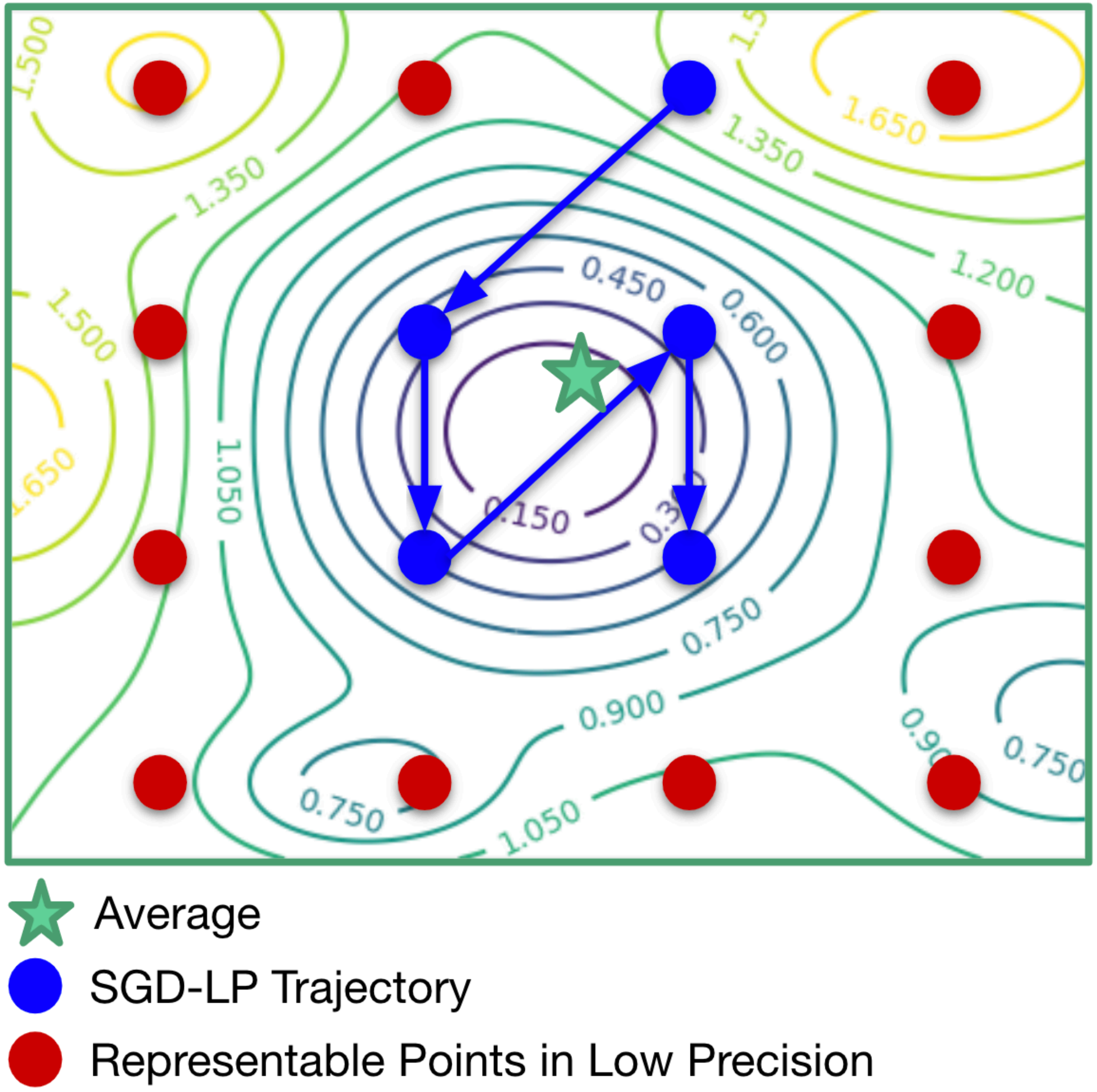
Low-precision SGD



● SGD-LP Trajectory

● Representable Points in Low Precision

Weight Averaging



SWALP

SWALP
model

SGD-LP
model

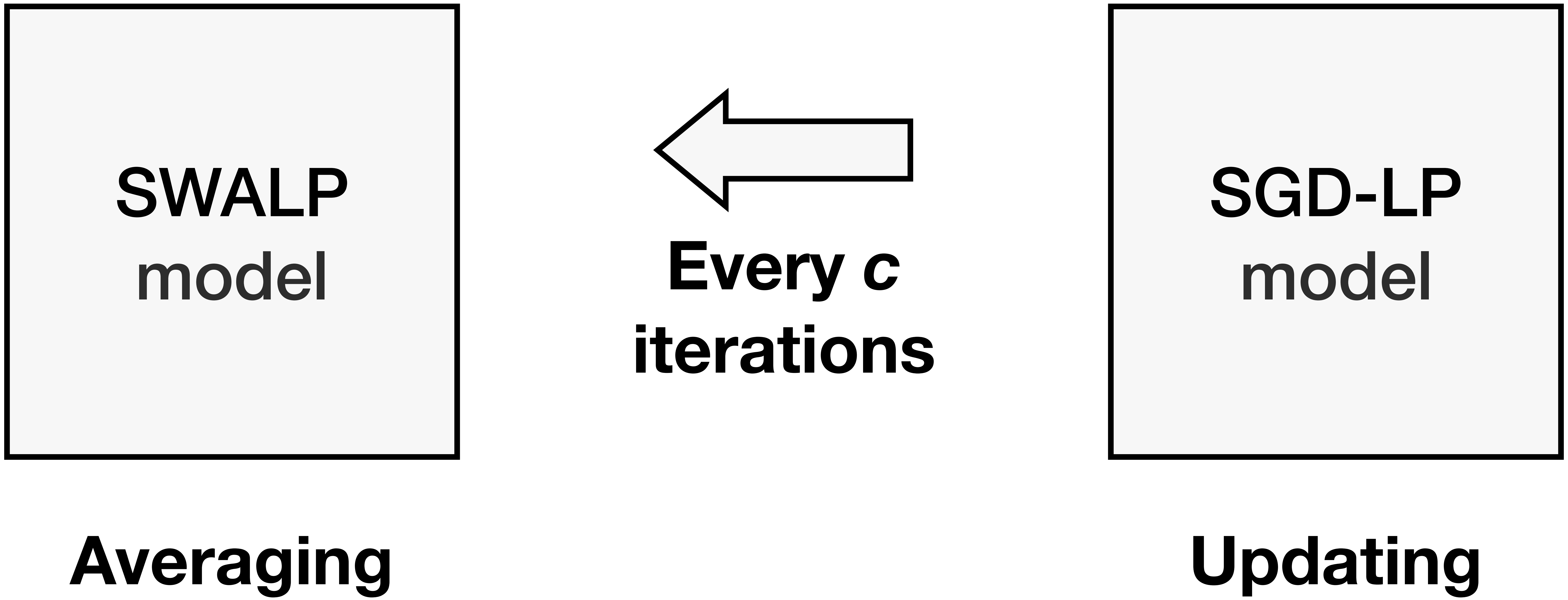
SWALP

SWALP model

SGD-LP model

Updating

SWALP



SWALP model

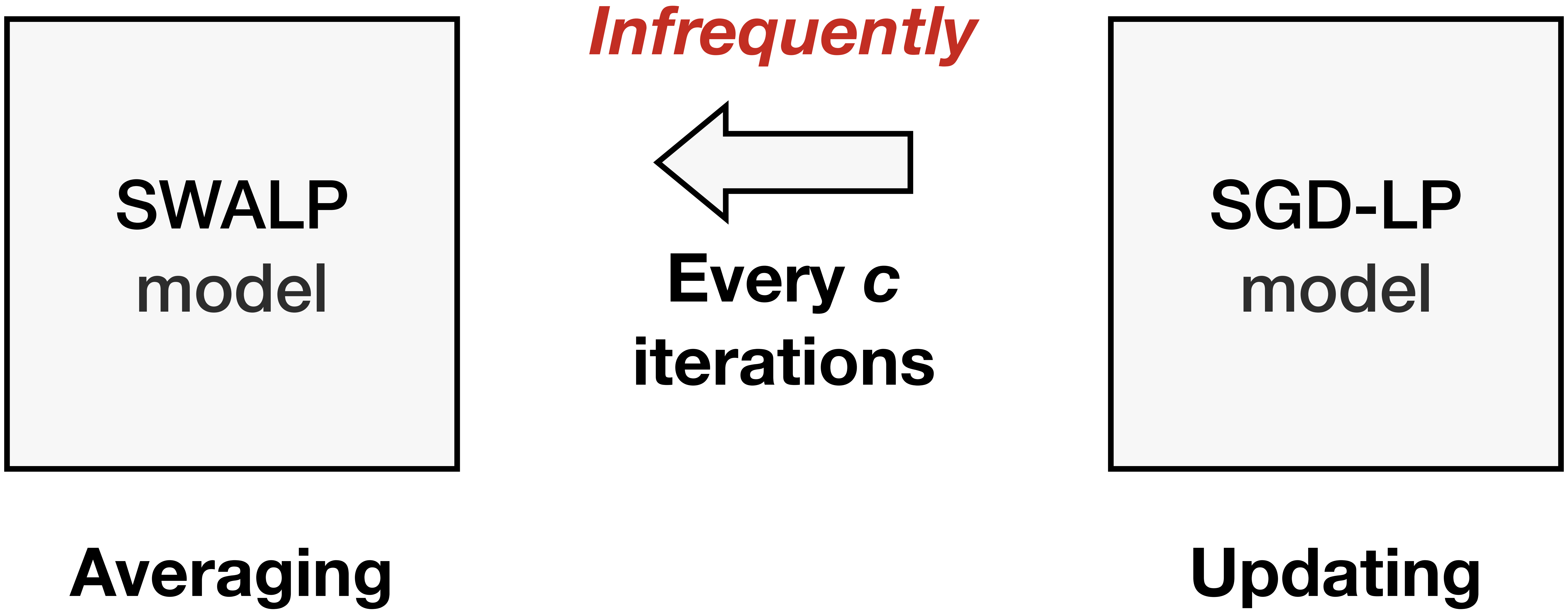
Every *c* iterations

SGD-LP model

Averaging

Updating

SWALP



SWALP model

Infrequently

Every *c* iterations

SGD-LP model

Averaging

Updating

Convergence Analysis

Let T be the number of iterations.

Theorem 1 (quadratic)
SWALP converges to the optimal solution at a O(1/T) rate.

Convergence Analysis

Let T be the number of iterations.

Theorem 1 (quadratic)
SWALP converges to the optimal solution at a O(1/T) rate.

SWALP has the same convergence rate as full precision SGD.

Convergence Analysis

Let δ be the quantization gap.

Theorem 2 (strongly convex)
The expected distance between SWALP solution and the optimal one is bounded by O(δ^2).

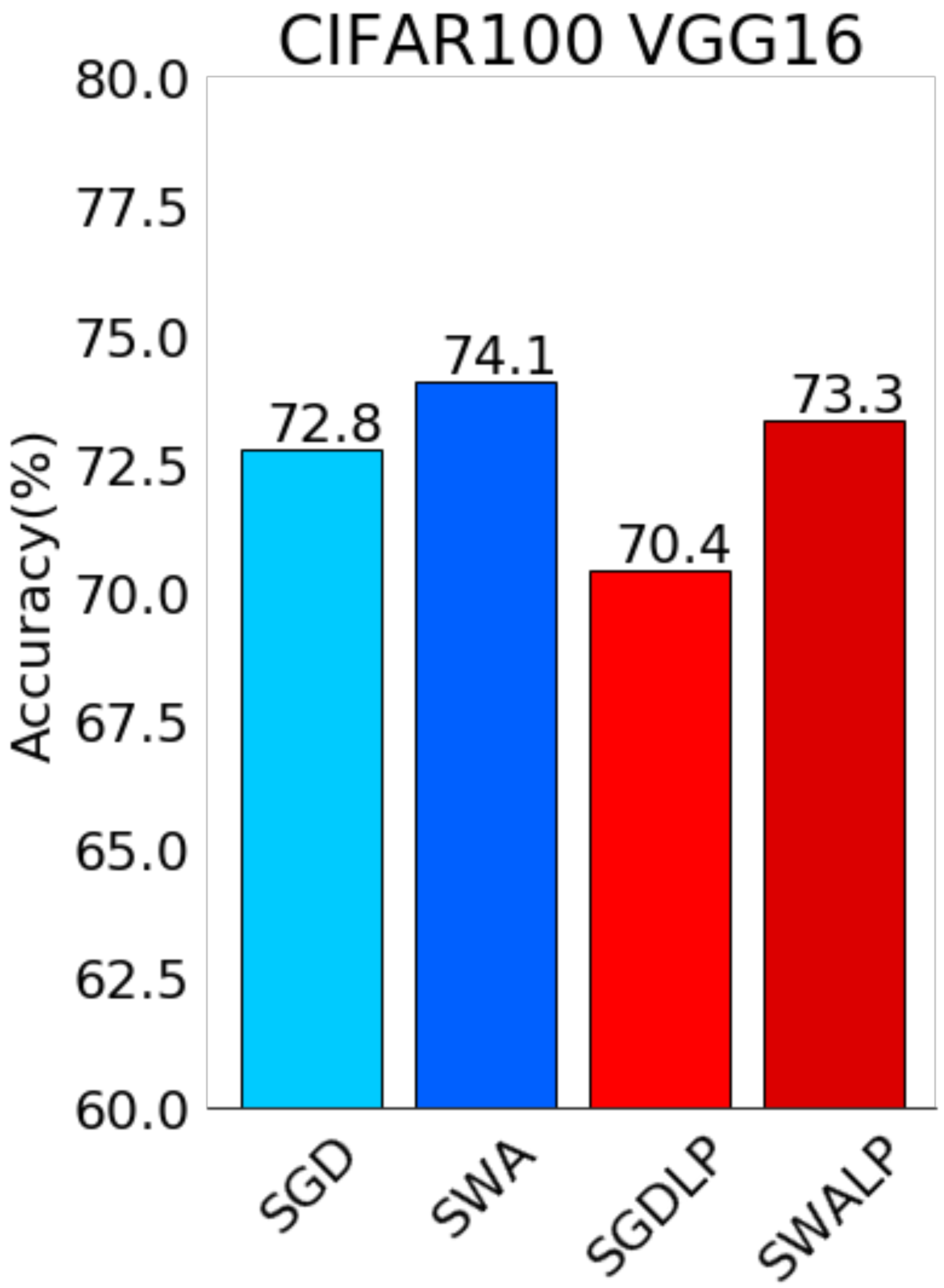
Convergence Analysis

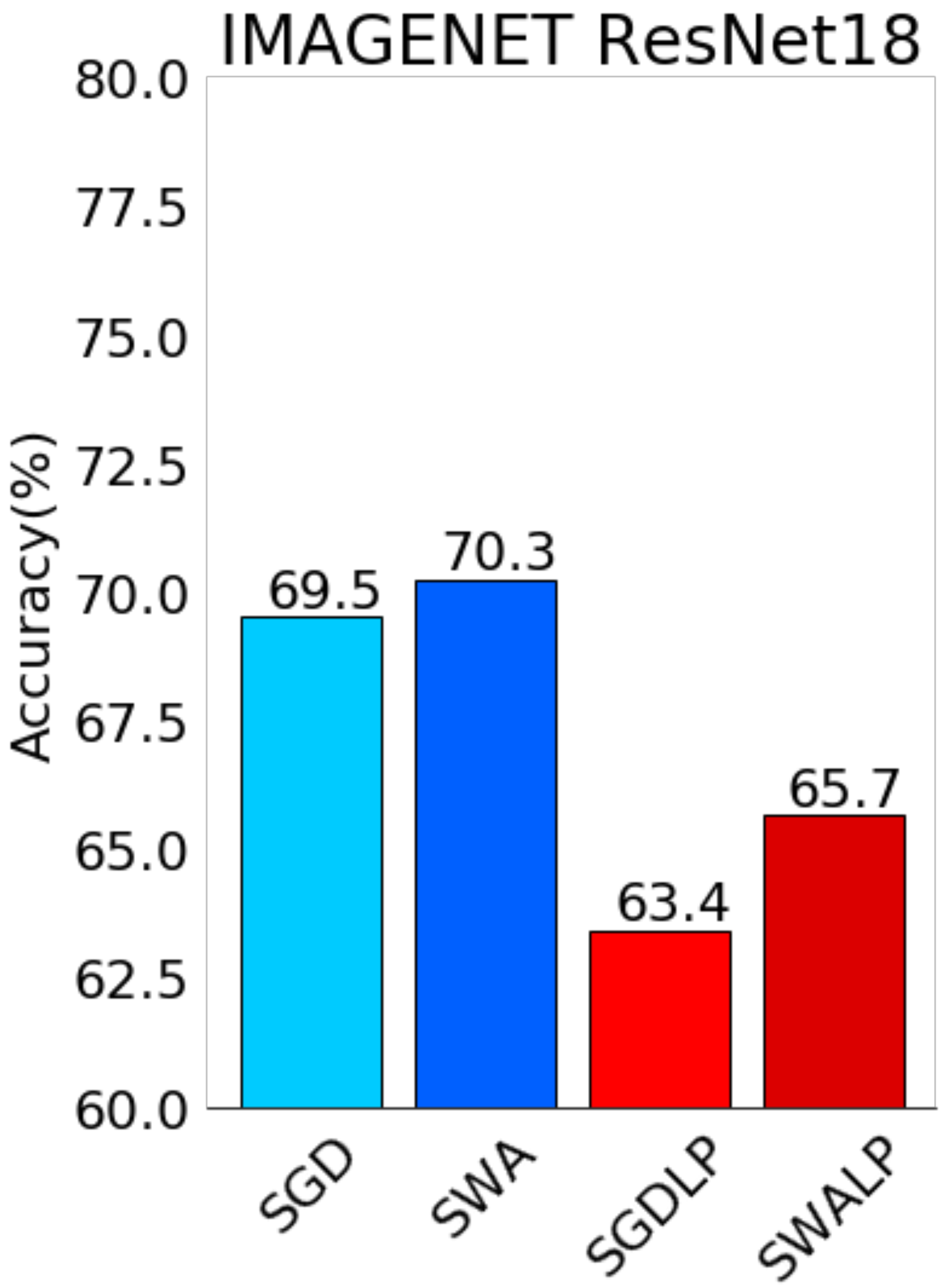
Let δ be the quantization gap.

Theorem 2 (strongly convex)
The expected distance between SWALP solution and the optimal one is bounded by O(δ^2).

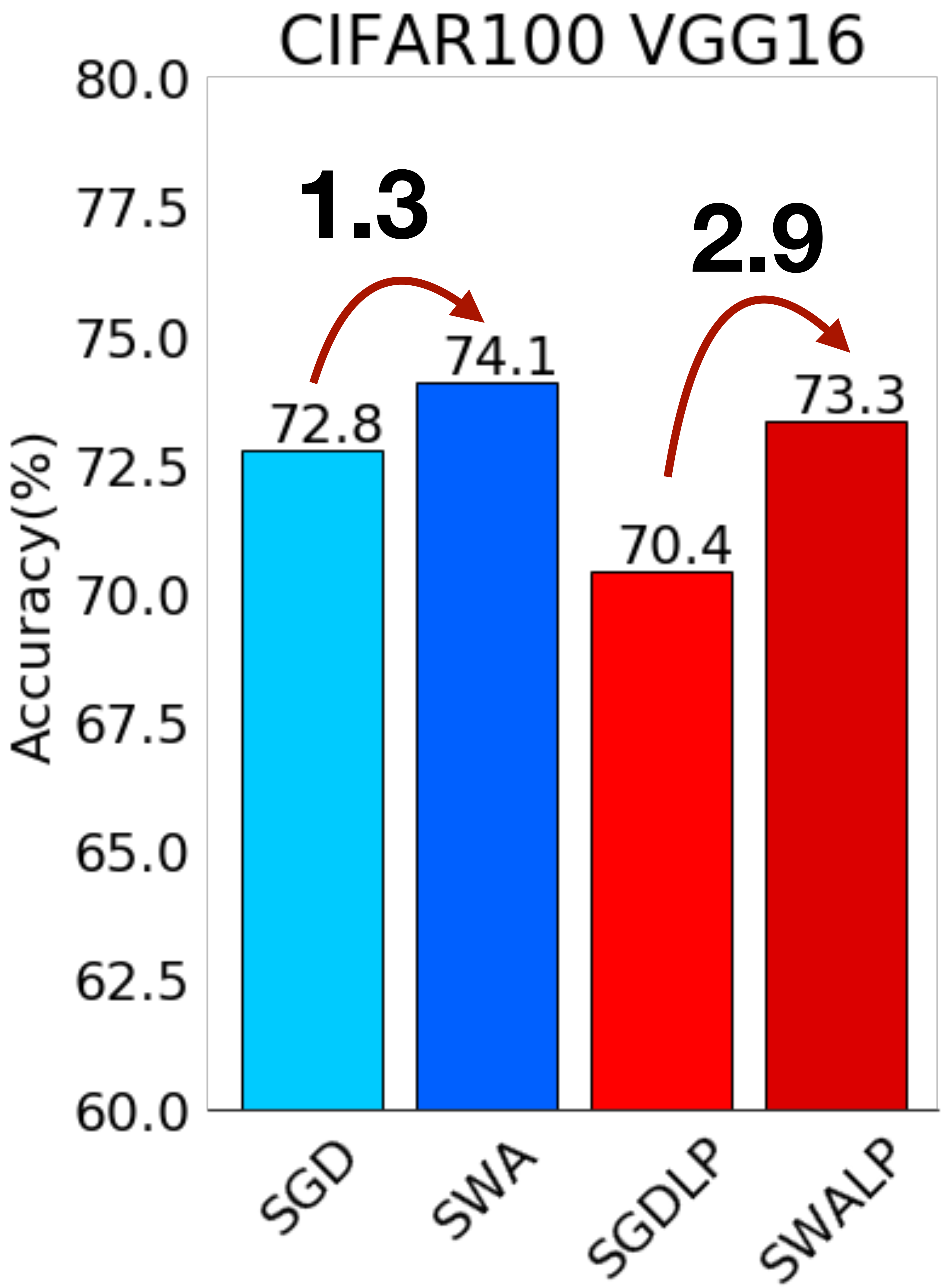
- The best bound for SGD-LP is O(δ) (Li et al, NeurIPS 2017).
- SWALP requires half the number of bits to reduce the noise ball by the same factor.

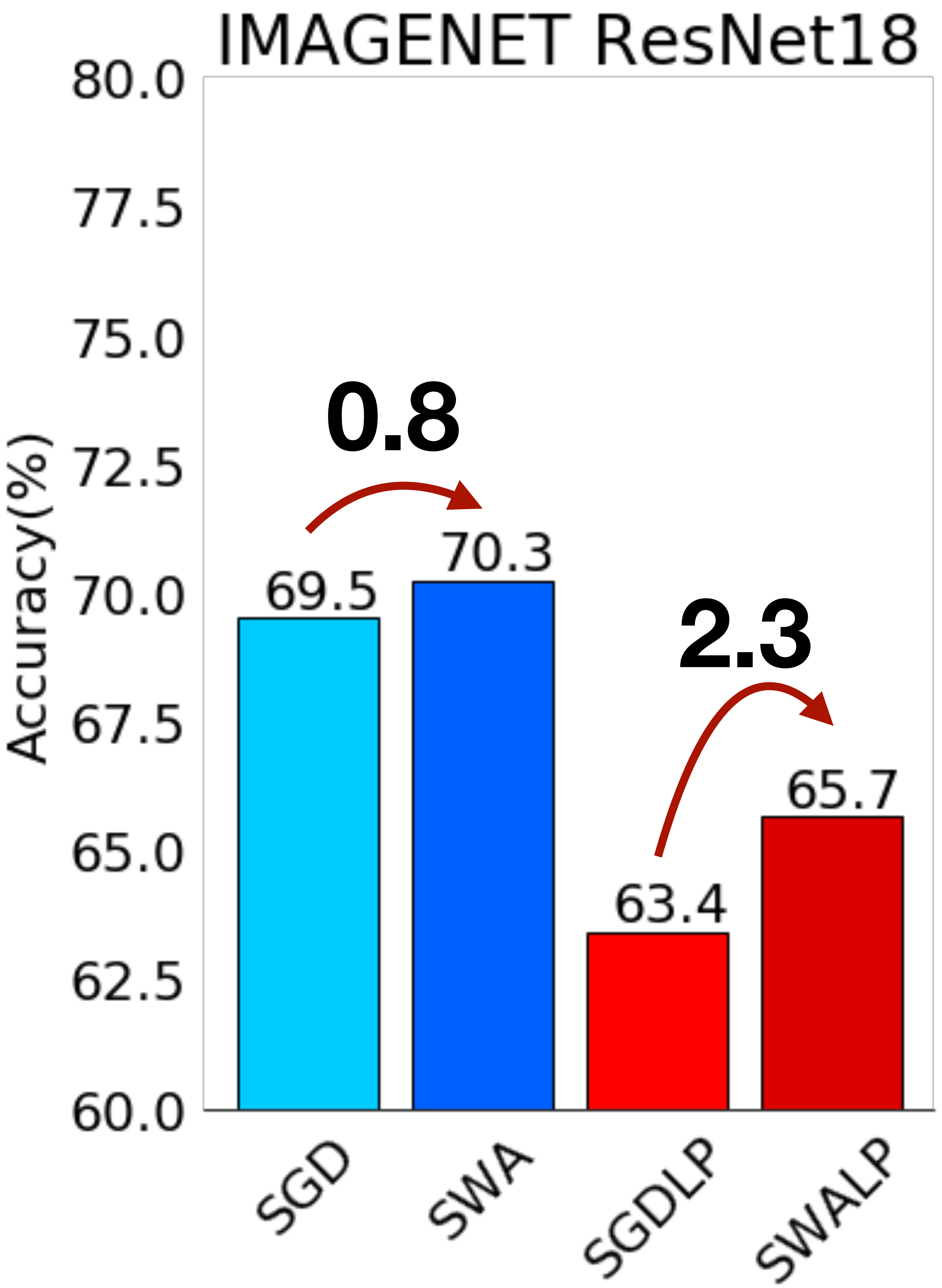
Experiments



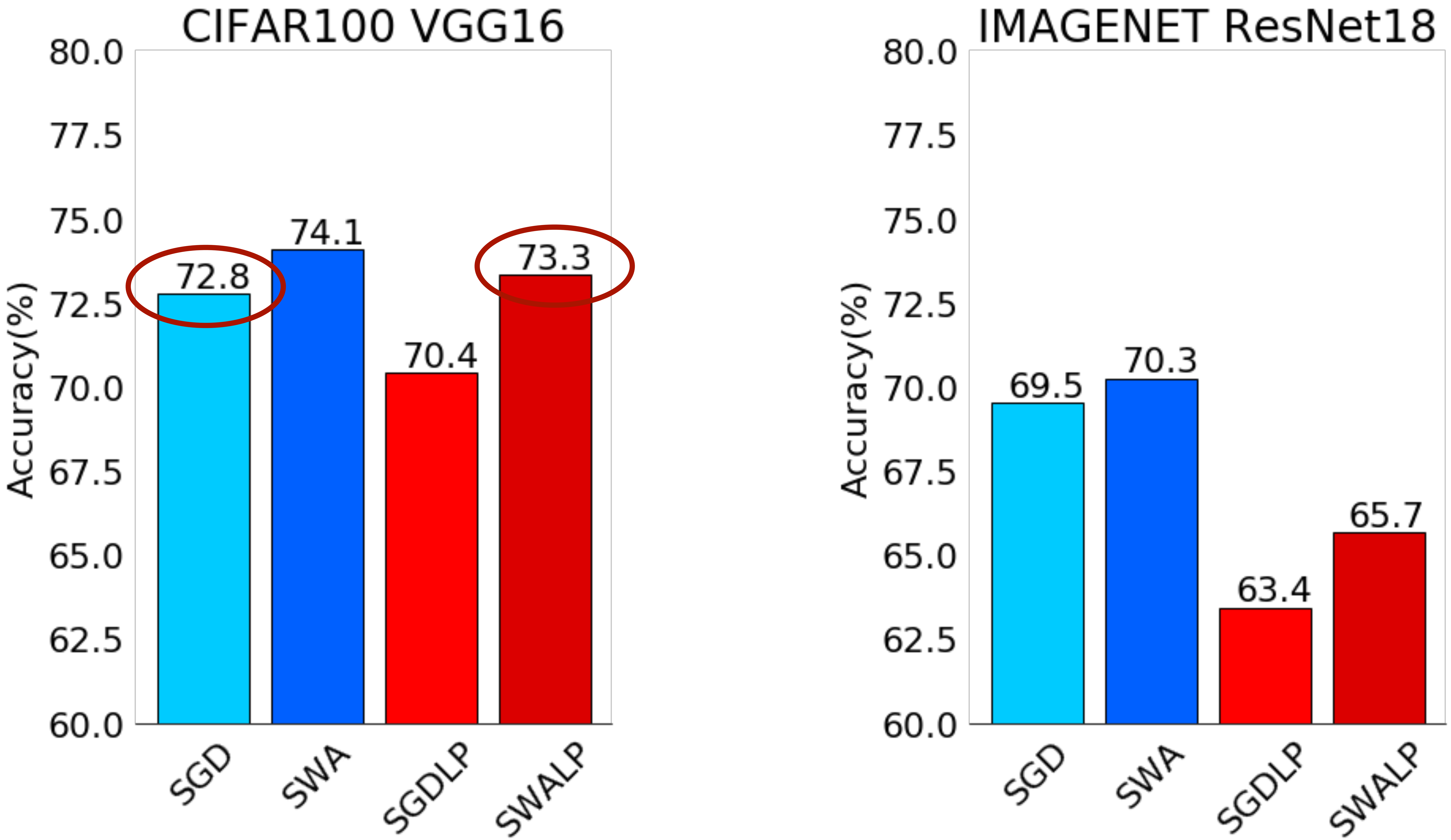


Experiments





Experiments



Poster @ Pacific Ballroom #58

SWALP Codes

QPyTorch:
A Low-Precision
Framework