It's All In the Teacher: Zero-Shot Quantization Brought Closer to the Teacher

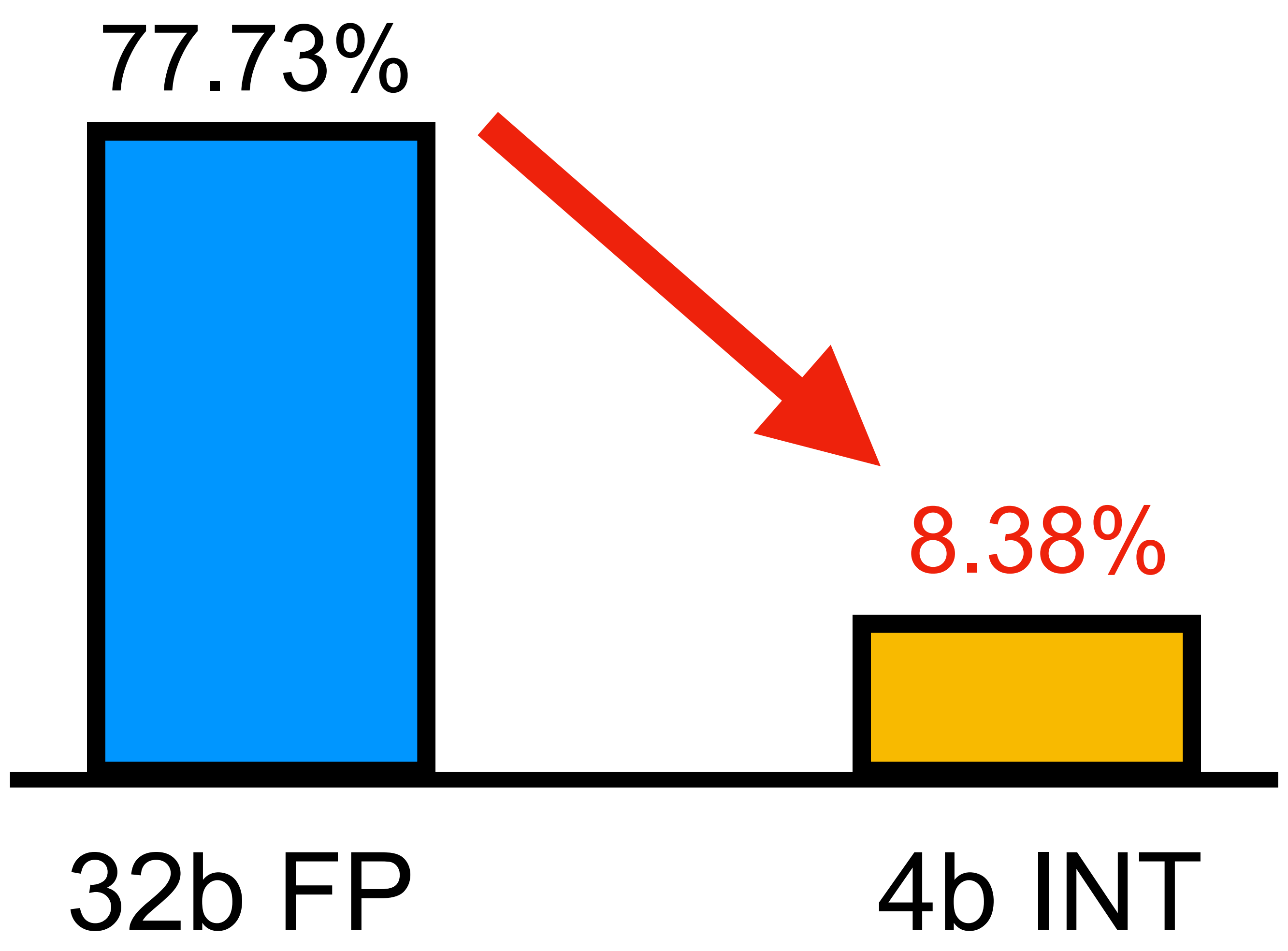
Kanghyun Choi[1] Hye Yoon Lee[1] Deokki Hong[1] Joonsang Yu[2]
Noseong Park[1] Youngsok Kim[1] Jinho Lee[1]

[1]College of Computing, Yonsei University
[2]CLOVA ImageVision, CLOVA AI Lab, NAVER

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

Current Limitations of Quantization

Quantized Network Accuracy
w/o fine-tuning[1]



Necessary Recalibration
with Original Dataset

Inaccessible Dataset Problem

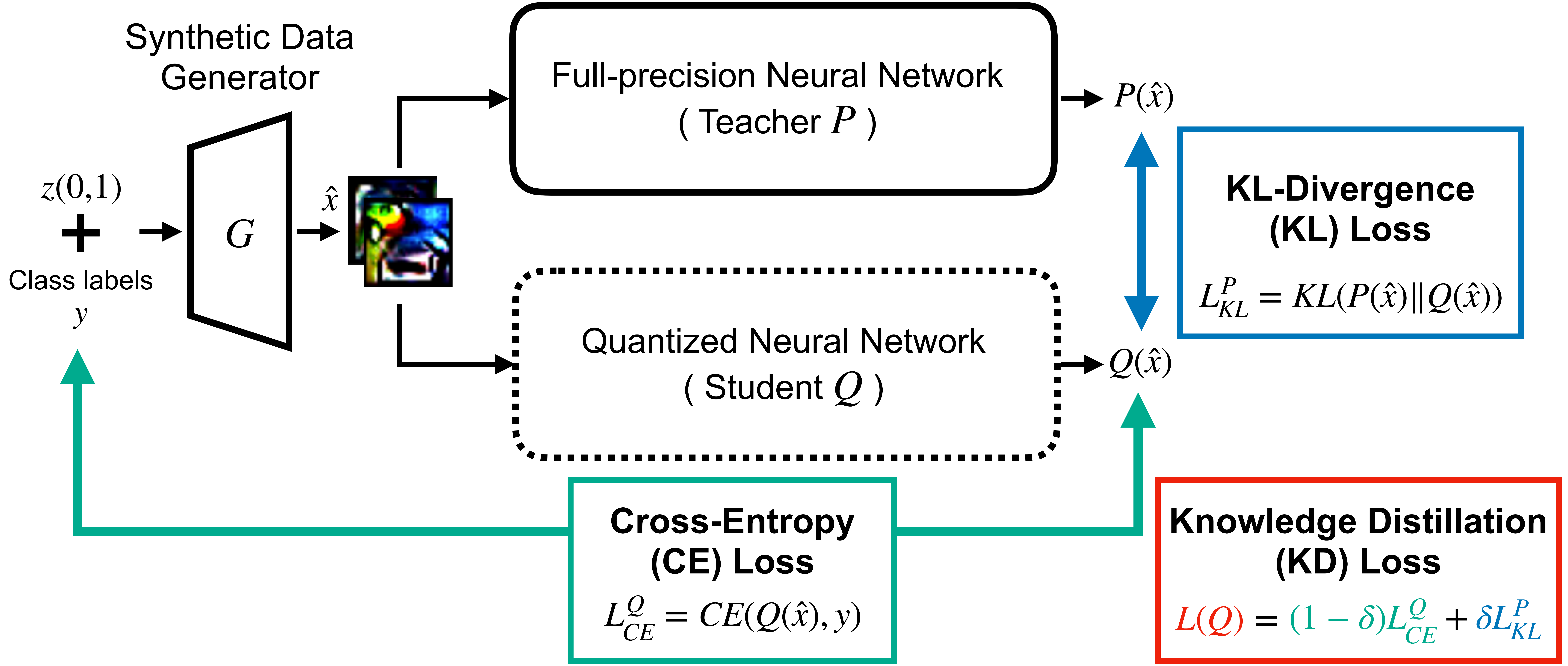
Privacy | Protection | Copyrights

Fine-tuning
without Original Dataset

Zero-shot
Quantization

[1] ResNet50, ImageNet-1k, 4bit quantization

Zero-shot Quantization Overview



Motivation

What loss function have we chosen?

Knowledge Distillation (KD) Loss

$$L(Q) = (1-\delta)L_{CE}^{Q} + \delta L_{KL}^{P}$$

No consideration of synthetic samples, or the quantization

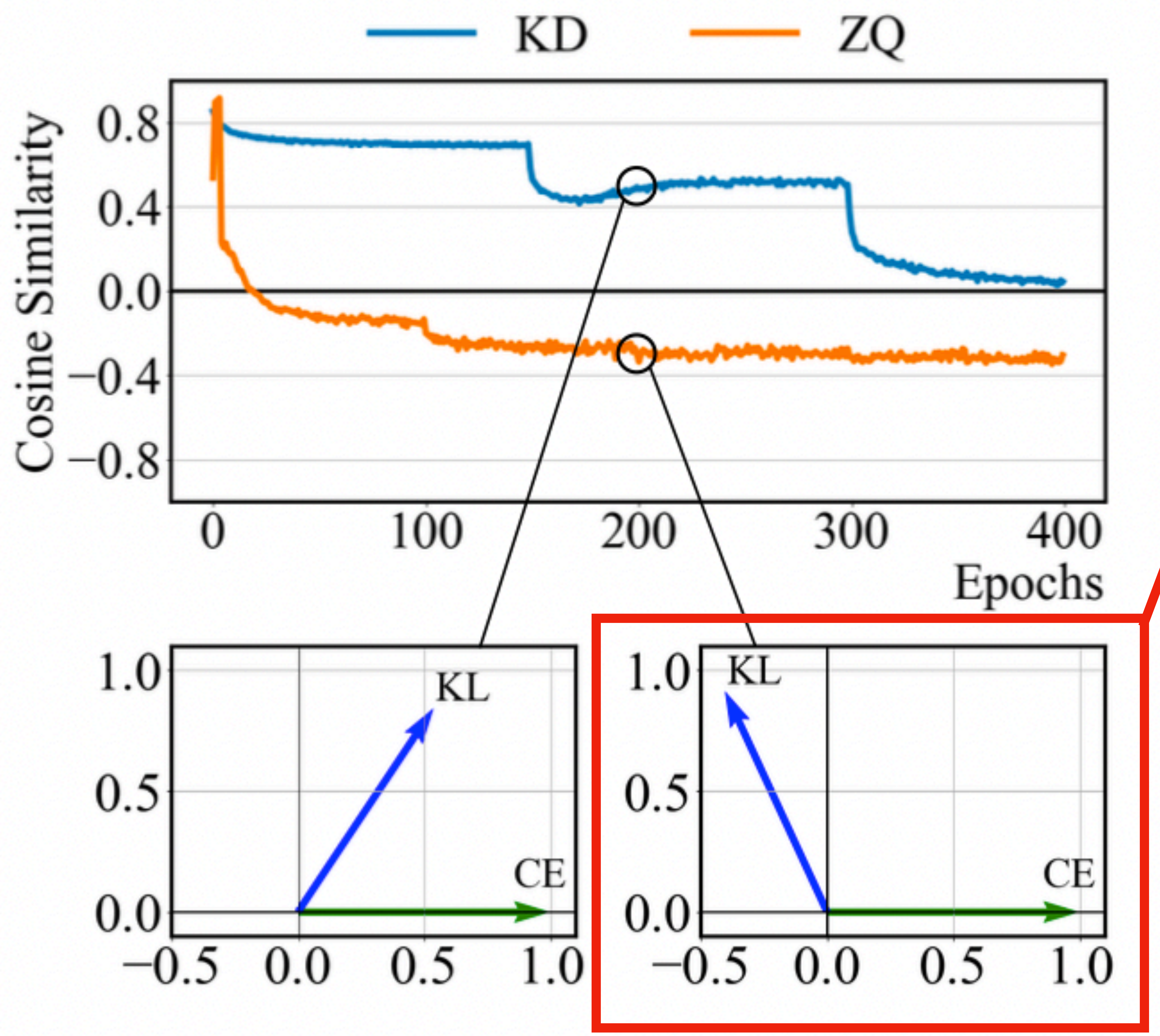
Do KL and CE losses cooperate on ZQ?

Measure **cosine similarity** between gradient of KL & CE

Analysis

Loss Function Analysis

CE-KL Gradient Cosine Similarity Analysis



Discrepancy between the gradient direction

(Cosine Similarity < 0.0)

KL and CE are **not cooperating**

➔ Performance degradation

What loss function do we need to choose?

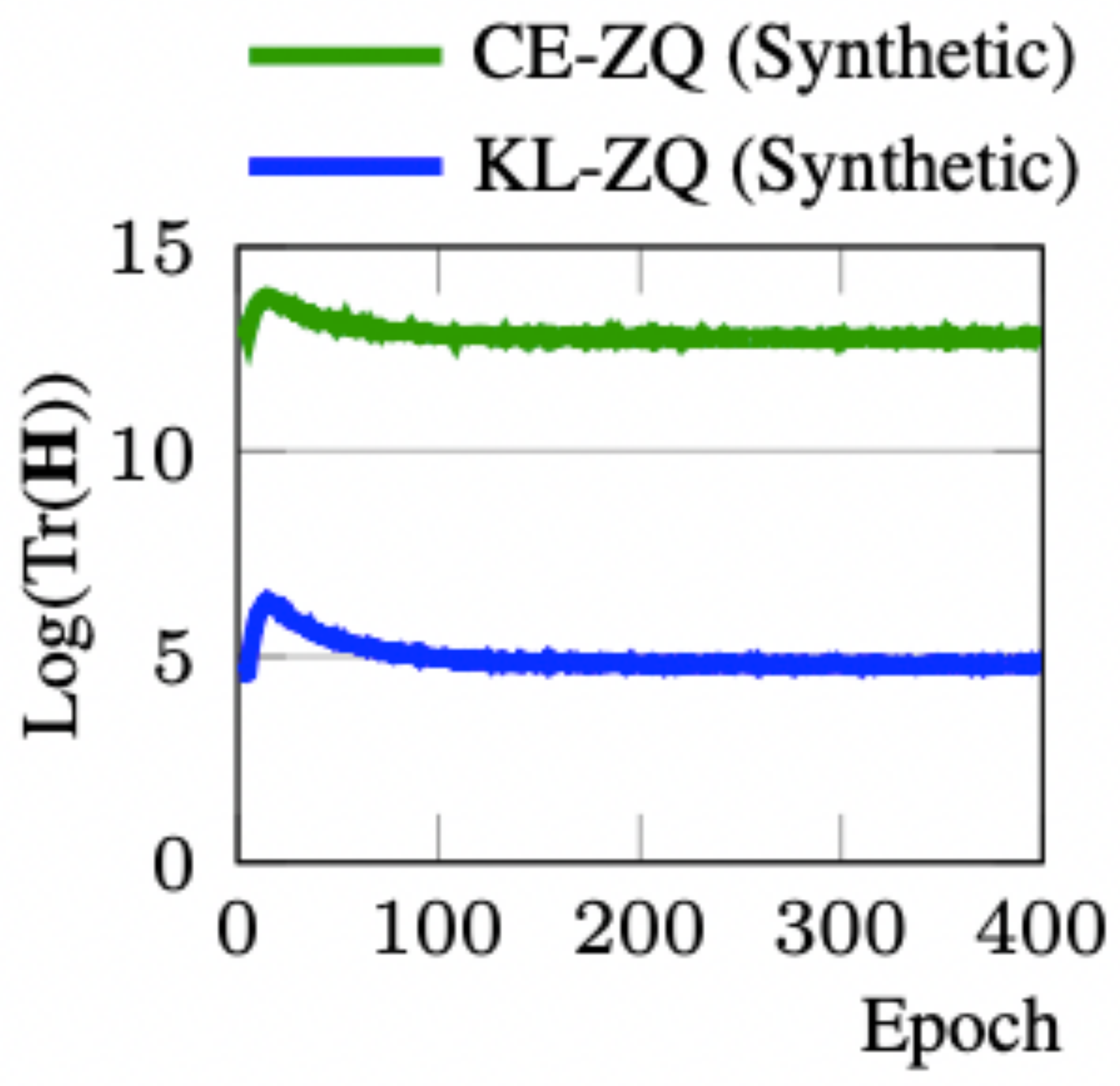
A : Loss function

with better generalizability

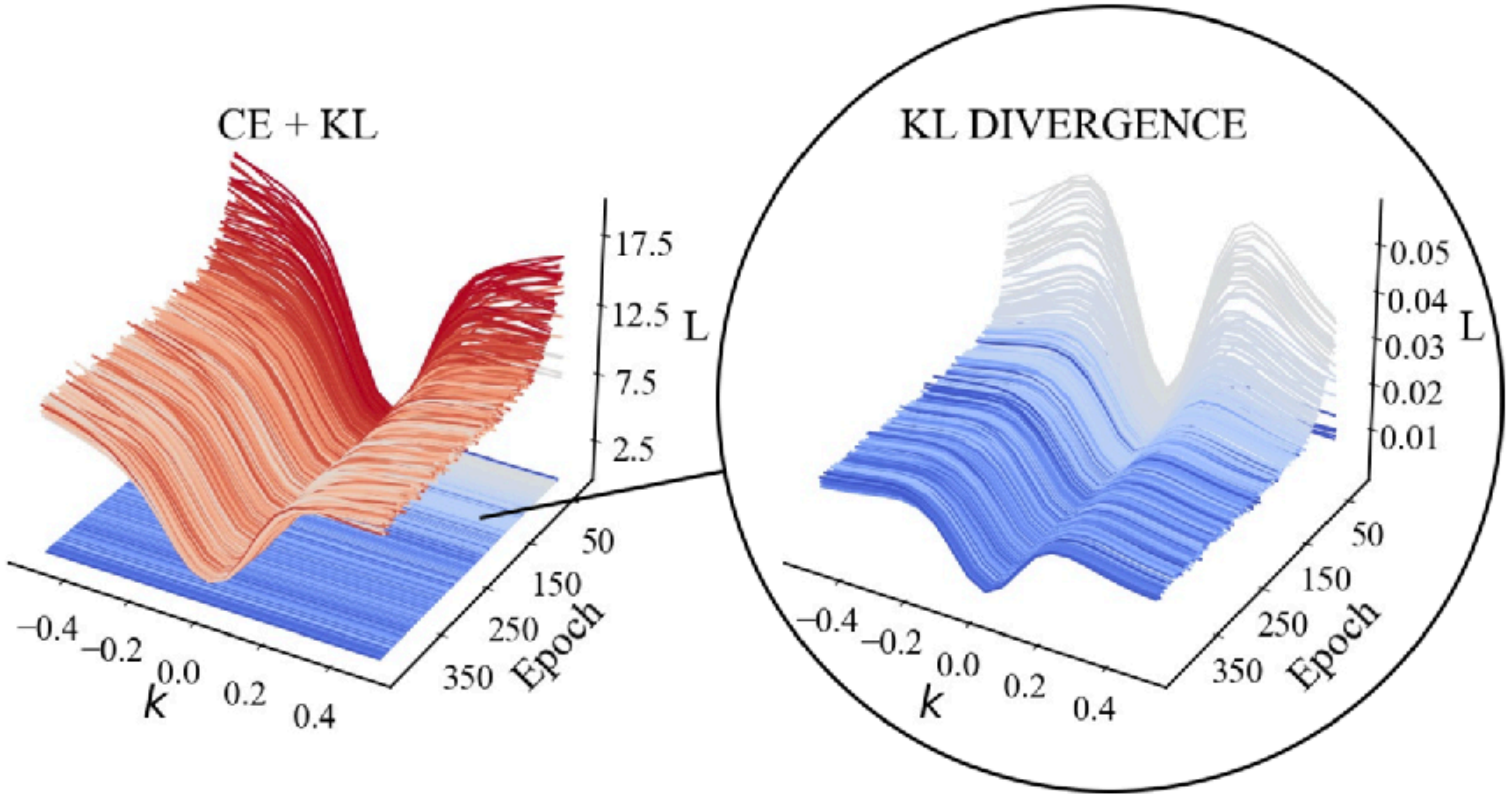
Analysis

Loss Surface and Generalization

Better generalizability ≈ Flatter local minima on loss surface ≈ Smaller trace of Hessian matrix



Trace of Hessian matrix (CE/KL)



Visualization of Loss Surface (CE/KL)

Loss Surface Flatness
KL-Divergence > Cross-Entropy

Use KL-Divergence loss only

Analysis

KL-Divergence Only Training

Dataset	Cifar-10	Cifar-100	ImageNet		
Model	ResNet-20	ResNet-20	ResNet-18	ResNet-50	MobileNetV2
Baseline (GDFQ)	90.25	63.39	60.60	52.12	59.43
KL-only	90.06	58.93	58.49	42.64	47.03
	-0.19%p	-4.46%p	-2.11%p	-9.48%p	-12.40%p

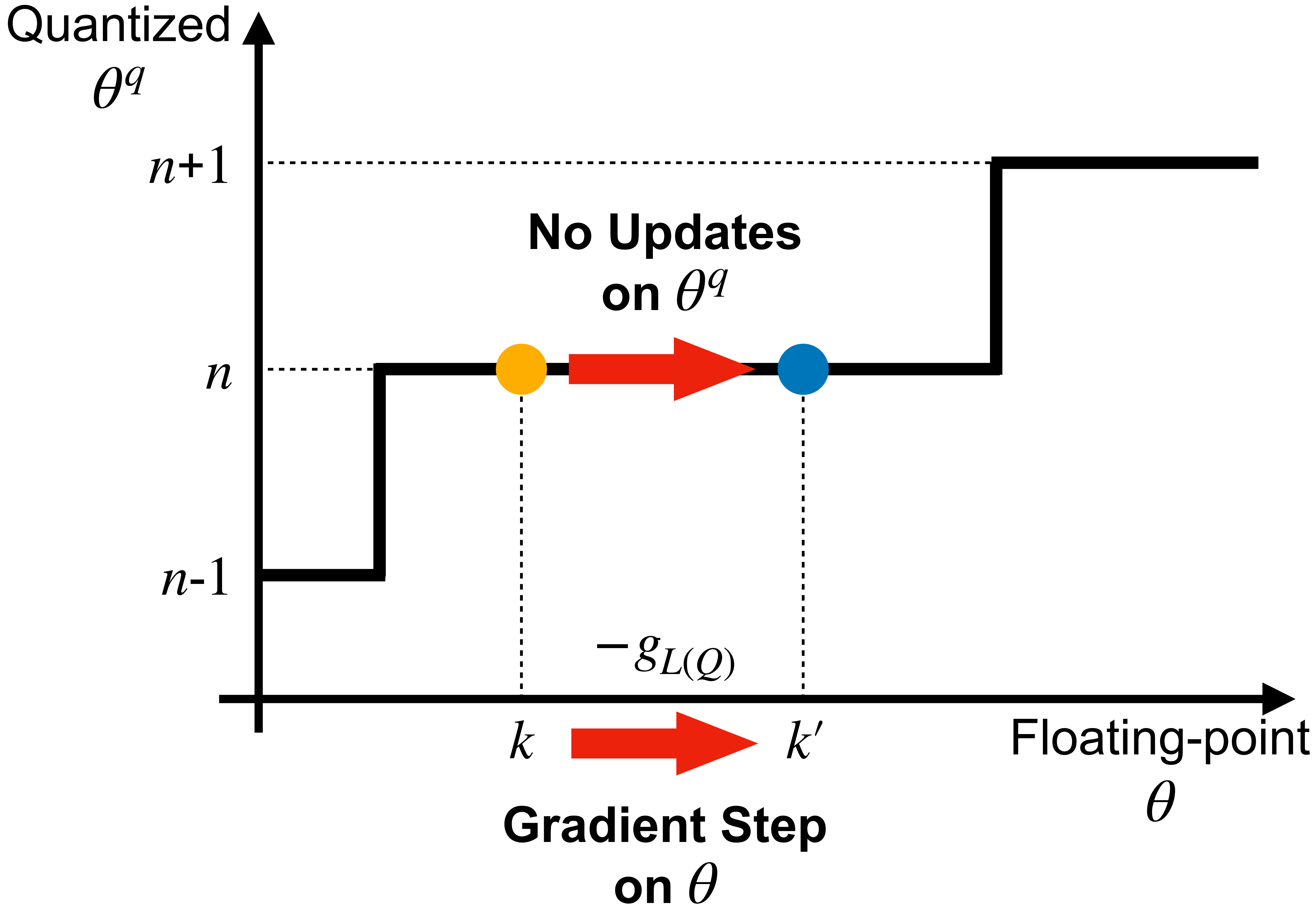
Huge accuracy degradation w/ KL-only training

Why?

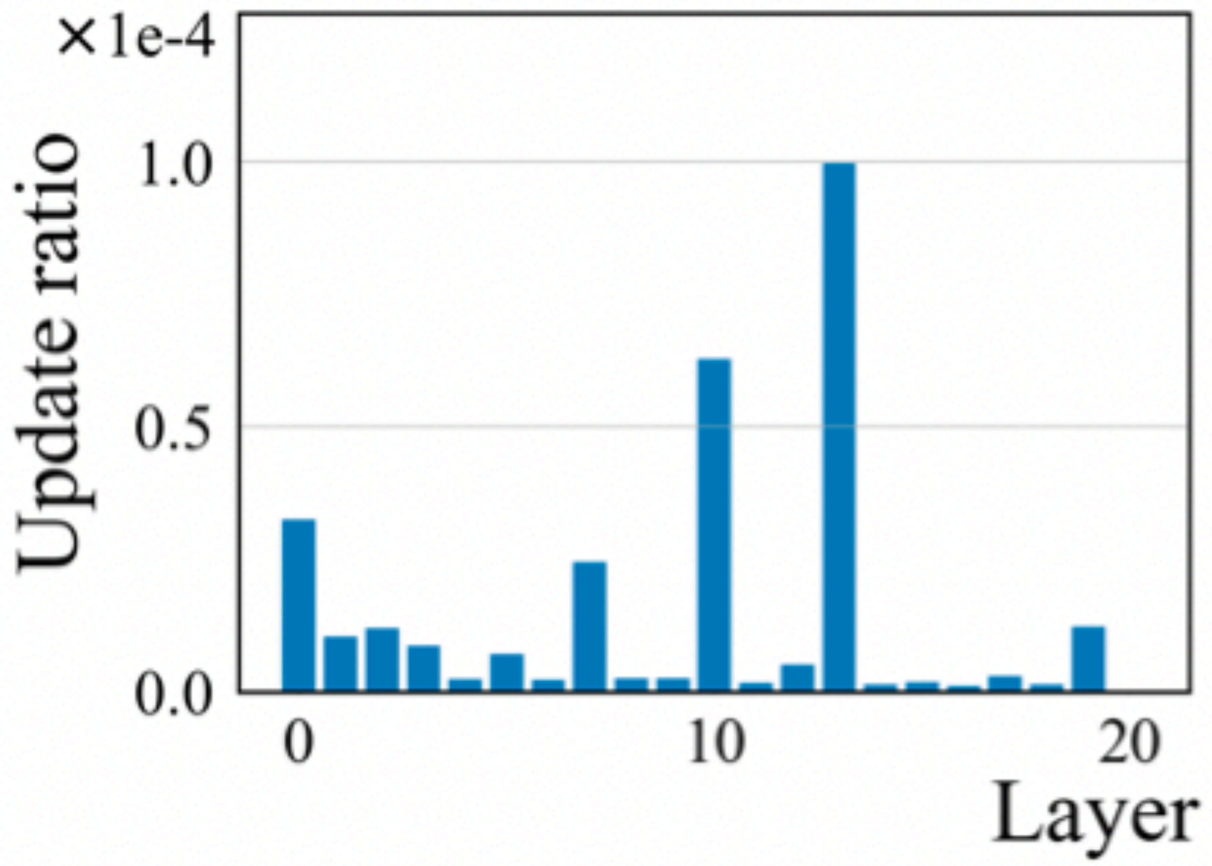
Quantization!

Analysis

Parameter Updates on Quantization

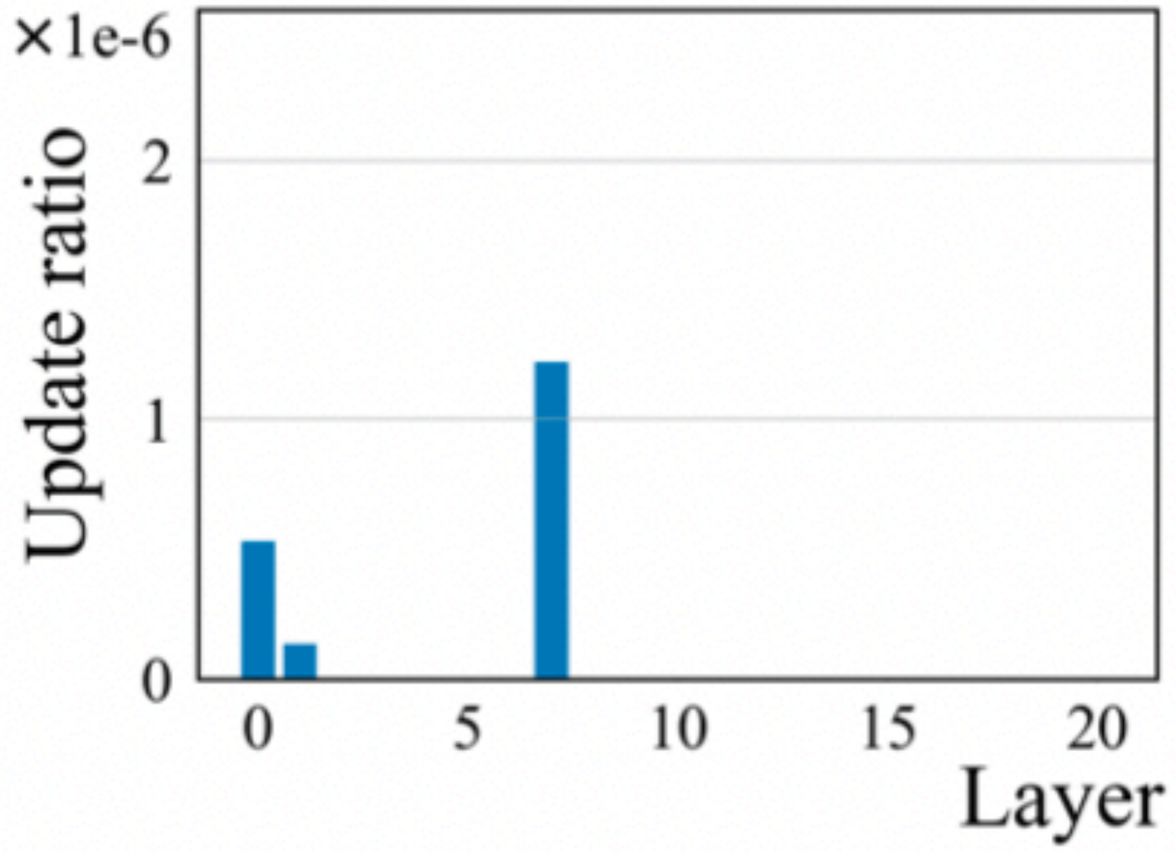


KL-only Training



Early Epochs
(60/400)

Imbalanced Updates



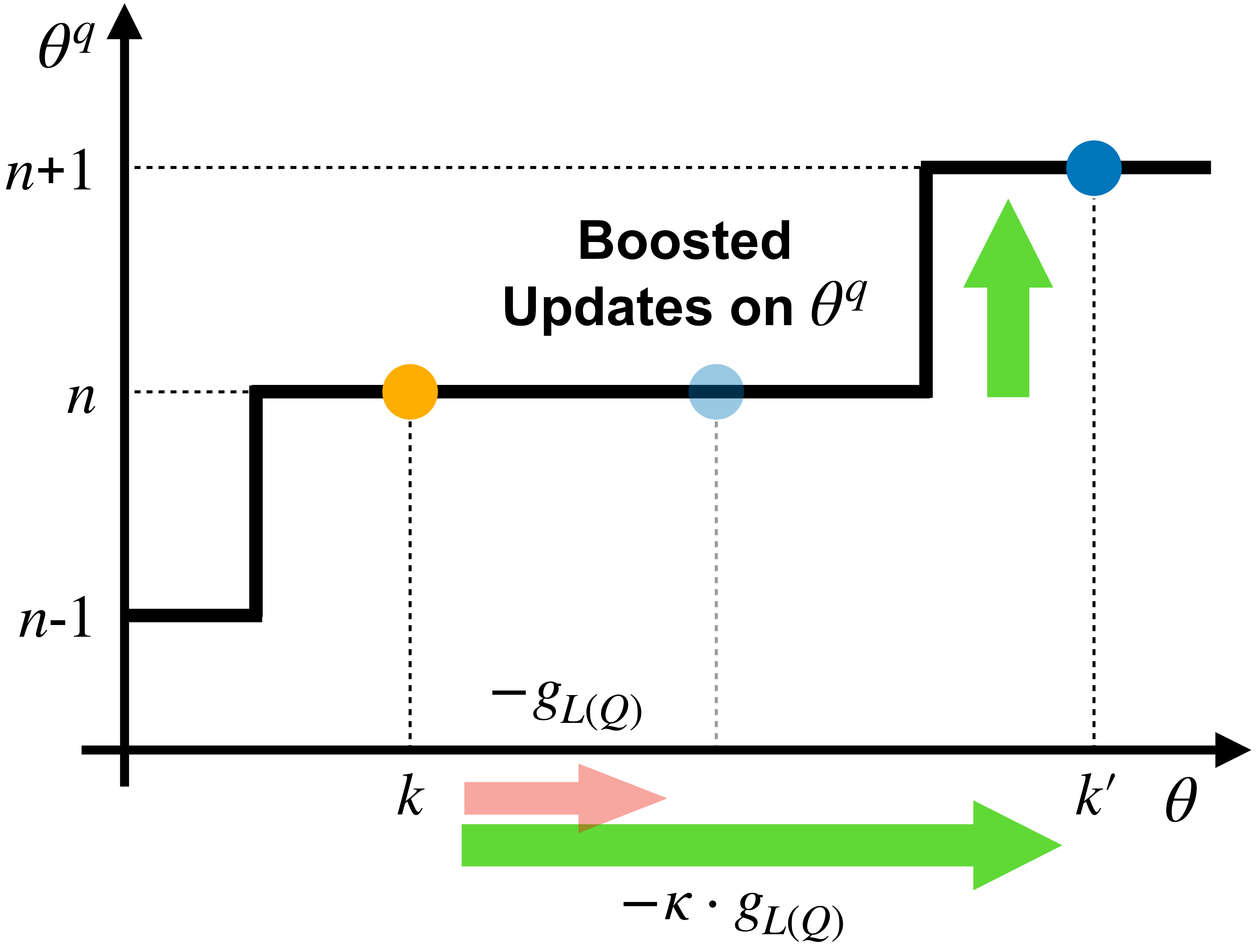
Later Epochs
(350/400)

Too Few Updates

Per-epoch
Updated θ^q Ratio

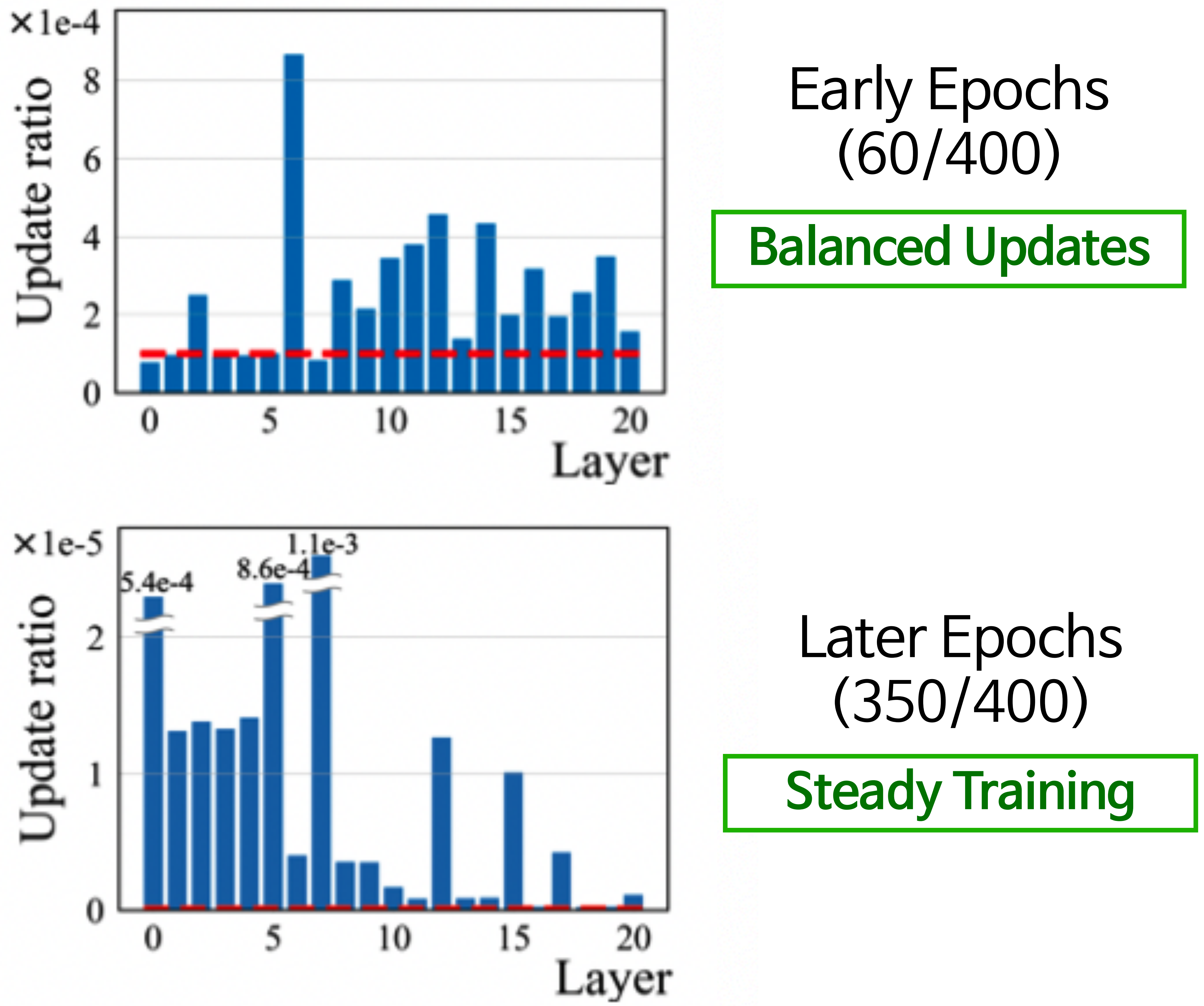
Proposed Method

Gradient Inundation (GI)



Gradient Inundation Step on θ

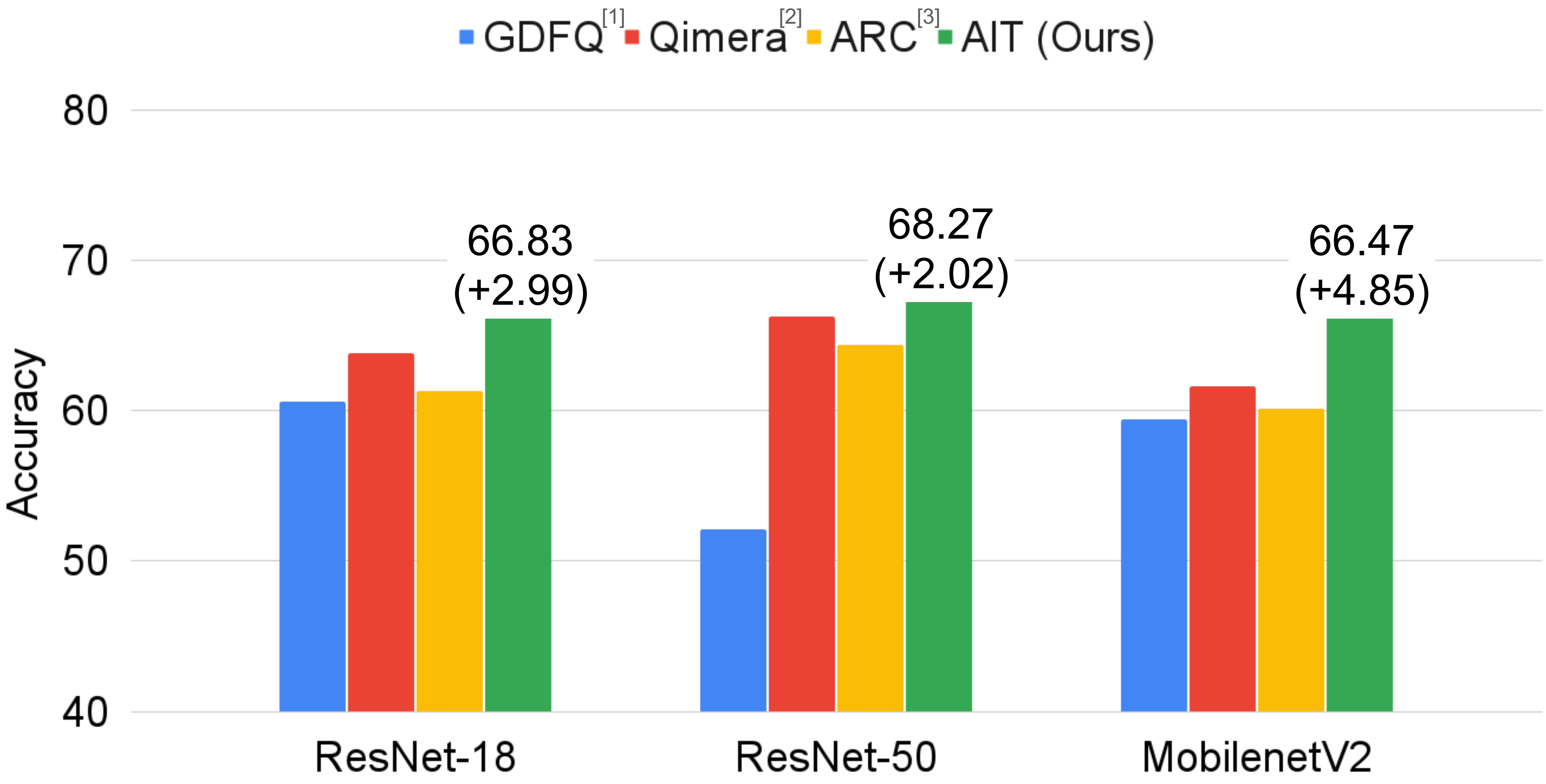
KL−only + **GI**
(proposed)



Per−epoch
Updated θ^q Ratio

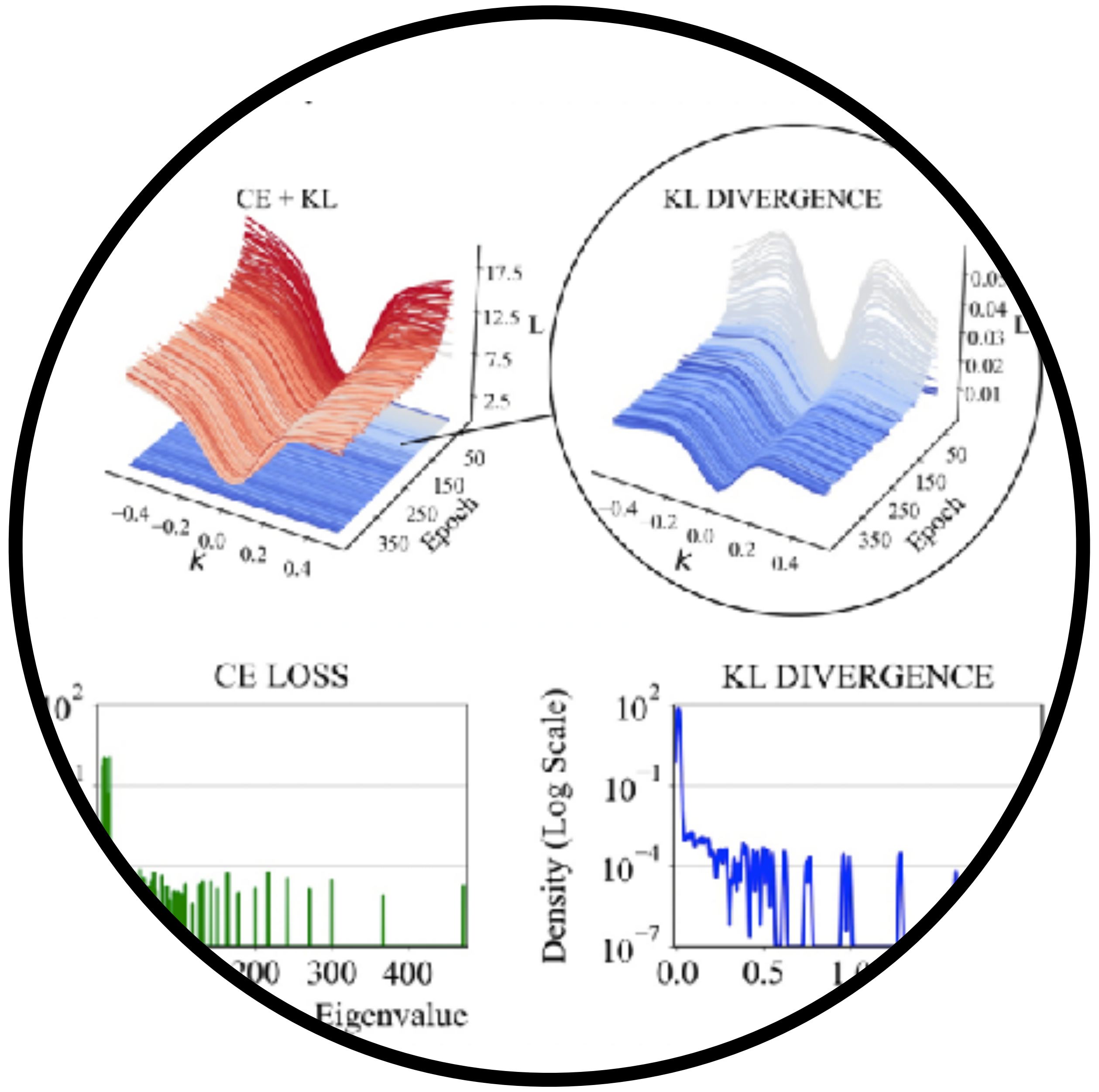
Experiment Results

ImageNet Classification (4-bit Quantization)

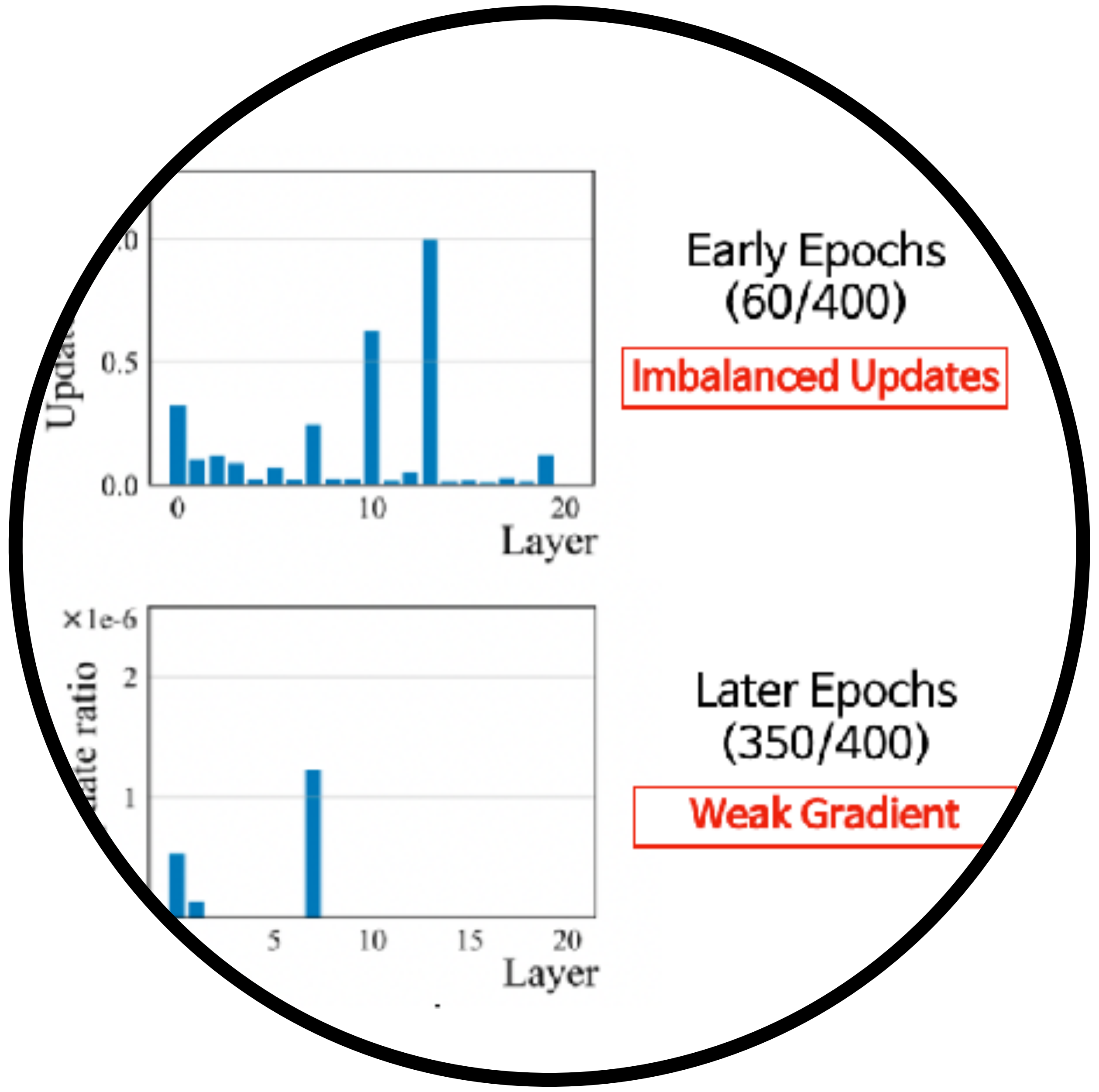


[1] Xu, Shoukai, et al. "Generative low-bitwidth data free quantization." ECCV 2020.
[2] Choi, Kanghyun, et al. "Qimera: Data-free Quantization with Synthetic Boundary Supporting Samples." NeurIPS 2021.
[3] Zhu, Baozhou, et al. "AutoReCon: Neural Architecture Search-based Reconstruction for Data-free Compression." IJCAI 2021

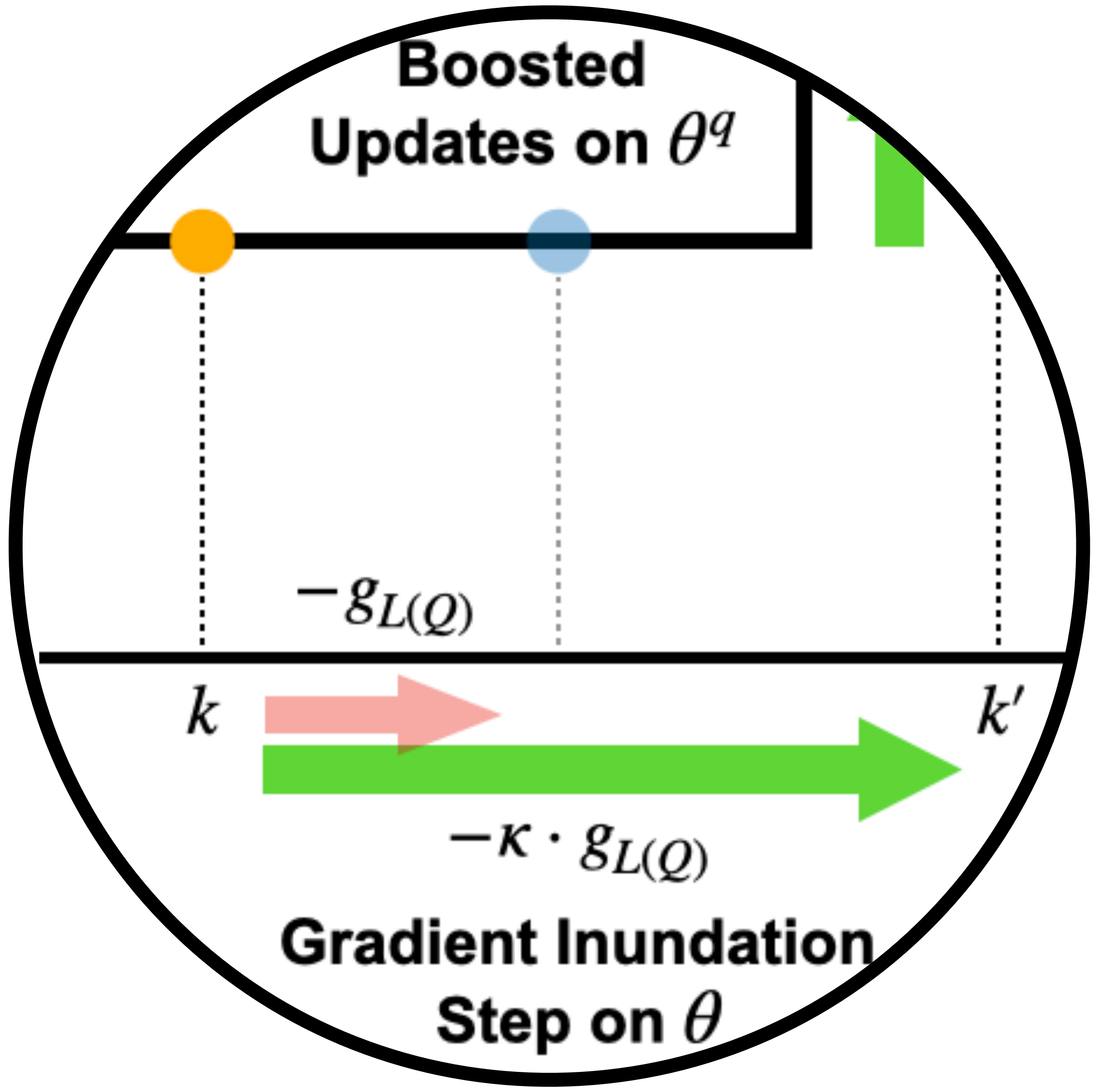
Conclusion



Analyzed loss functions
from multiple perspectives
and emphasized flatter loss minima



Inspected current limitations of
zero-shot quantization training



Proposed method that ensures
balanced weight updates
among all layers

For the details and more analysis, refer to the paper or find us on poster session 2.2.