Information-theoretic Understanding of Population Risk Improvement with Model Compression

Yuheng Bu†, Weihao Gao†, Shaofeng Zou‡ and Venugopal V. Veeravalli†

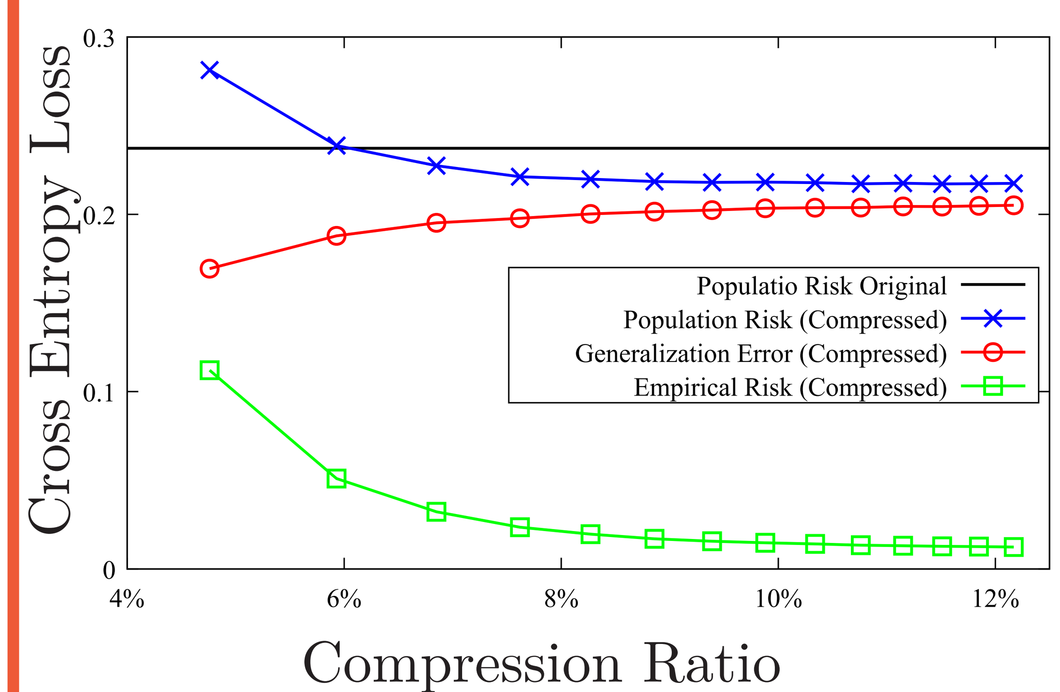
† University of Illinois at Urbana-Champaign ‡ University at Buffalo, the State University of New York

1. Introduction

Compression of deep models is necessary due to limited **storage** and **computational resources**

- Compression $\rightarrow$ worse performance
- Recent work shows population risk can be **improved** after compression

Our contribution: provide information-theoretic explanation by characterizing tradeoff between **generalization error** and **empirical risk**



- **Why?** Model compression serves as *regularizer* to reduce generalization error.
- **How?** We propose new quantization algorithm to reduce generalization error in model compression.

2. Compression Improves Generalizatio

Consider instance space $\mathcal{Z}$, hypothesis space $\mathcal{W}$, loss function $\ell : \mathcal{W} \times \mathcal{Z} \rightarrow \mathbb{R}^+$

- Training dataset: $S = \{Z_1, \cdots, Z_n\}$ from μ
- Population risk: $L_\mu(w) \triangleq \mathbb{E}_{Z\sim\mu}[\ell(w, Z)]$
- Empirical risk: $L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(w, Z_i)$
- Learning algorithm: conditional distribution $P_{W|S}$.
- Generalization error: $\text{gen}(\mu, P_{W|S}) \triangleq \mathbb{E}_{W,S}[L_\mu(W) - L_S(W)]$
- Compression algorithm: conditional distribution $P_{\hat{W}|W}$
  Markov chain $S \rightarrow W \rightarrow \hat{W}$

Theorem. For learning algorithm $P_{W|S}$, and compression algorithm $P_{\hat{W}|W}$, suppose $\ell(\hat{w}, Z)$ is σ-sub-Gaussian under $Z \sim \mu$ for all $\hat{w} \in \hat{\mathcal{W}}$, then

$$|\text{gen}(\mu, P_{\hat{W}|S})| \le \sqrt{\frac{2\sigma^2}{n} I(W; \hat{W})}.$$

$I(W; \hat{W})$ corresponds to **rate** of model compression in rate-distortion theory

3. Distortion on Empirical Risk

In rate-distortion framework,

- Define distortion with empirical risk $d_S(w, \hat{w}) \triangleq L_S(\hat{w}) - L_S(w)$

$D(R) = \min_{I(W;\hat{W})\le R} \mathbb{E}_{S,W,\hat{W}}[d_S(\hat{W}, W)]$ distortion *decreases* as rate increases.

Combine with generalization error bound

Theorem. Suppose assumptions in previous Theorem hold, and $I(W; \hat{W}) = R$, then

$$\min_{P_{\hat{W}|W}: I(W;\hat{W})=R} \mathbb{E}_{S,W,\hat{W}}[L_\mu(\hat{W}) - L_S(W)] \le \sqrt{\frac{2\sigma^2}{n}R} + D(R).$$

Compression controls trade-off between empirical risk and generalization error and improves population risk

4. Example: Linear Regression

Model: $Y_i = X_i w^* + \varepsilon_i$, for $i = 1, ..., n$.

- $X_i \in \mathbb{R}^d$ i.i.d. Gaussian $\mathcal{N}(0, \Sigma_X)$
- ε_i i.i.d. Gaussian $\mathcal{N}(0, \sigma'^2)$
- Diameter of weights space $\hat{\mathcal{W}}$ is $C(\hat{\mathcal{W}})$
- Consider ERM: $W = (XX^T)^{-1}XY$.

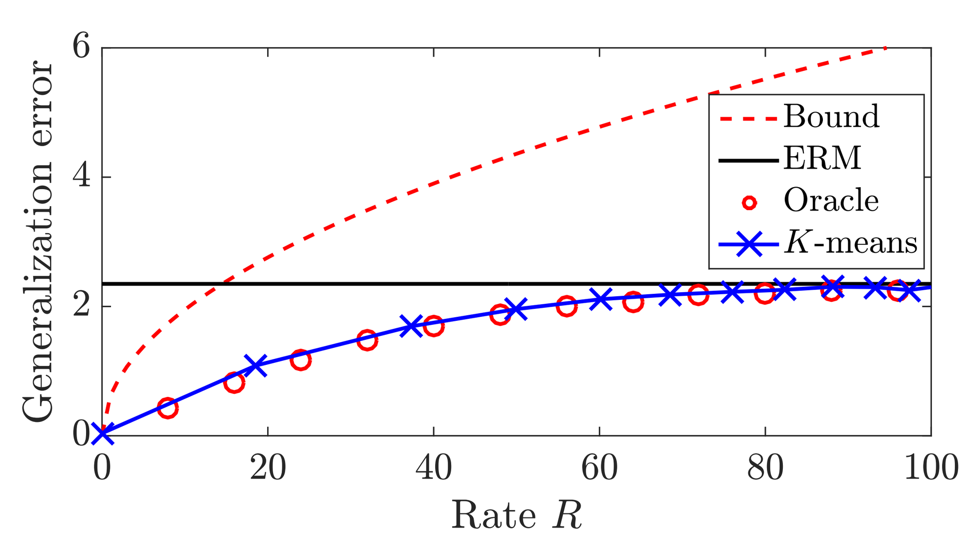
Theorem. Upper bound on gen

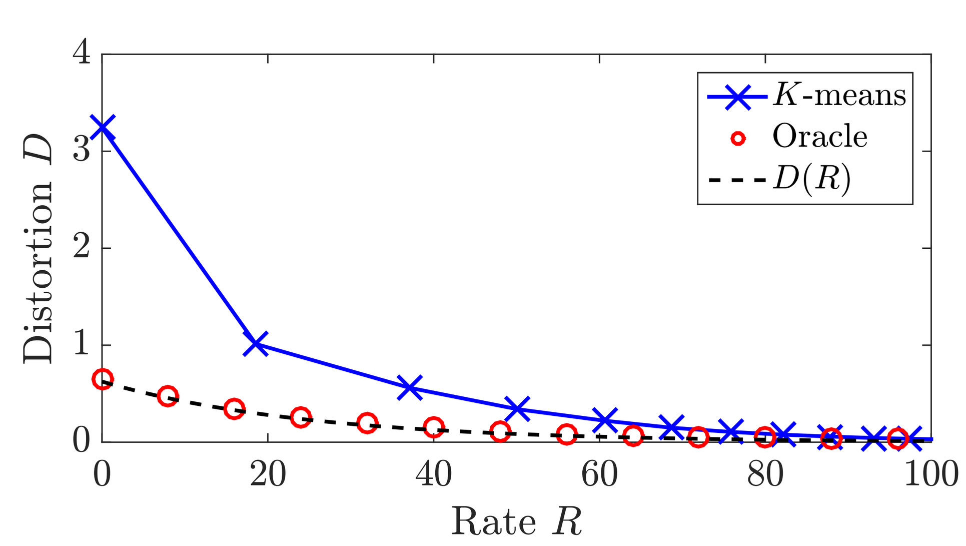
$$\text{gen}(\mu, P_{\hat{W}|S}) \le 2\sigma_\ell^{*2}\sqrt{\frac{I(W; \hat{W})}{n}}.$$

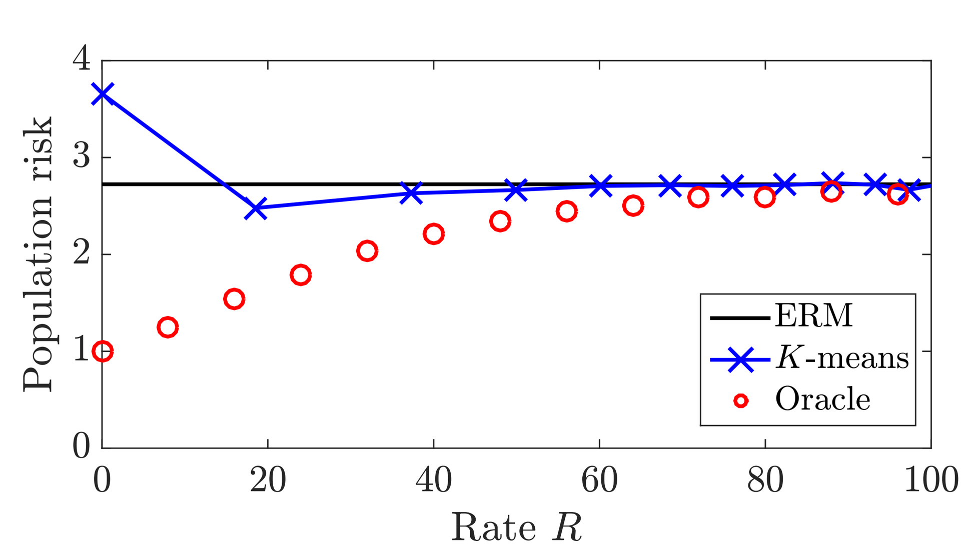
where $\sigma_\ell^{*2} \triangleq C(\hat{\mathcal{W}})\|\Sigma_X\| + \sigma'^2$.

Theorem. Upper bound population risk

$$\min_{P_{\hat{W}|W}: I(W;\hat{W})=R} \mathbb{E}_{S,W,\hat{W}}[L_\mu(\hat{W}) - L_S(W)] \le 2\sigma_\ell^{*2}\sqrt{\frac{R}{n}} + \frac{d\sigma'^2}{n-d-1}e^{-\frac{2R}{d}}, \quad R \ge 0.$$







5. From Theory to Algorithms

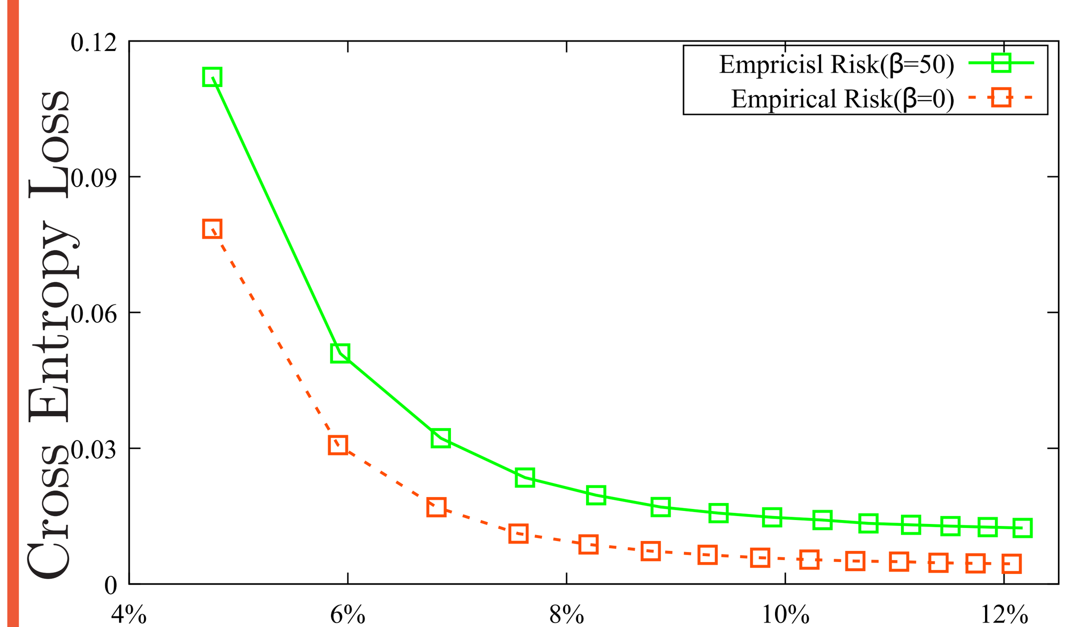
Improving **quantization** algorithm:

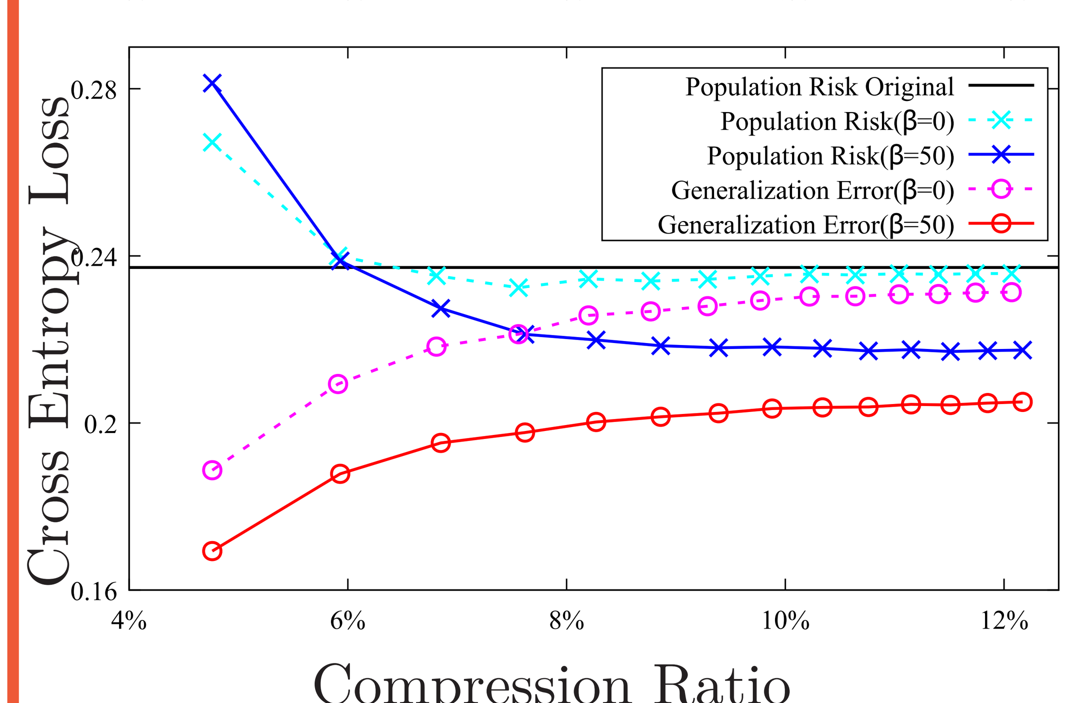
- Network parameters $w = \{w_1, \cdots, w_d\}$
- k clusters, choose centroids $\{c^{(1)}, \dots, c^{(k)}\}$ and assignments.
- Approximated distortion for ERM $d_S(\hat{w}, w) \approx \sum_{j=1}^d h_j(w_j - \hat{w}_j)^2$
- $I(W; \hat{W}) \le \log k$.
- Diameter of compressed weights can be approximated by: $\max_{k_1,k_2} |c^{(k_1)} - c^{(k_2)}|^2$.

Diameter-regularized Hessian-weighted K-means algorithm:

$$\min \left\{ \sum_{k=1}^K \sum_{w_j \in C^{(k)}} h_j |w_j - c^{(k)}|^2 + \beta \max_{k_1,k_2} |c^{(k_1)} - c^{(k_2)}|^2 \right\}$$

- Dataset: MNIST and CIFAR10
- Model: retrained models with part of training set
- Compare with original Hessian-weighted K-means ($\beta = 0$)





6. Related work

- Minimizing empirical risk of compressed model [Gao et.al. 2018]
- Bounding generalization error using small complexity of compressed model [Zhou et.al. 2018]

Reference

Y. Bu, S. Zou, V. V. Veeravalli. "Tightening Mutual Information Based Bounds on Generalization Error", in Proc. IEEE International Symposium on Information Theory (ISIT), Paris, France, 2019