Generalization Error Analysis of Quantized Compressive Learning

Xiaoyun Li Ping Li
Department of Statistics, Rutgers University
Cognitive Computing Lab, Baidu Research USA

Random Projection (RP) Method

- Data matrix $X \in \mathbb{R}^{n\times d}$, normalized to unit norm (all samples on unit sphere).
- Save storage by k random projections: $X_R = X \times R$, with $R \in \mathbb{R}^{d\times k}$ a random matrix with i.i.d. $N(0,1)$ entries $\Longrightarrow X_R \in \mathbb{R}^{n\times k}$.
- J-L lemma: approximate distance preservation $\Longrightarrow$ Many applications: clustering, classification, compressed sensing, dimensionality reduction, etc..
- "Projection+quantization": more storage saving. Apply (entry-wise) scalar quantization function $Q(\cdot)$ by $X_Q = Q(X_R)$.
- More applications: MaxCut, SimHash, 1-bit compressive sensing, etc..

Compressive Learning + Quantization

- We can apply learning models to projected data (X_R, Y), where Y is the response or label $\Longrightarrow$ learning in the projected space S_R!
- This is called **compressive learning**. It has been shown that learning in the projected space is able to provide satisfactory performance, while substantially reduce the computational cost, especially for high-dimensional data.
- We go one step further: learning with quantized random projections (X_Q, Y) $\Longrightarrow$ learning in the quantized projected space S_Q!
- This is called **quantized compressive learning**. A relatively new topic, but is practical in applications with data compression.

Paper Summary

- We provide generalization error bounds (of a test sample $x \in \mathcal{X}$) on three quantized compressive learning models:
  - Nearest neighbor classifier
  - Linear classifier (logistic regression, linear SVM, etc.)
  - Linear regression
- **Applications**: we identify the factors that affect the generalization performance of each model, which gives recommendations on the choice of quantizer Q in practice.
- Some experiments are conducted to verify the theory.

Backgrounds

A b-bit quantizer Q_b separates the real line into $M = 2^b$ regions.

- **Distortion**: $D_{Q_b} = E[(Q_b(X) - X)^2] \Longleftrightarrow$ minimized by Lloyd-Max (LM) quantizer.
- **Maximal gap** of Q on interval $[a, b]$: the largest gap between two consecutive boarders of Q on $[a, b]$.
- Indeed, we can estimate the inner product between two samples x_1 and x_2 through the estimator $\hat{\rho}_Q(x_1, x_2) = \frac{Q(x_1^T R) Q(R^T x_2)}{k}$, which might be biased. We define the **debiased variance** of a quantizer Q as the variance of $\hat{\rho}_Q$ after debiasing.
- **Idea:** connection between the generalization of three models and inner product estimates.

Quantized Compressive 1-NN Classifier

- We are interested in the risk of a classifier h, $\mathcal{L}(h) = E[\mathbb{1}\{h(x) \neq y\}]$.
- Assume $(x, y) \sim \mathcal{D}$, with conditional probability $\eta(x) = P(y = 1|x)$. Bayes classifier $h^*(x) = \mathbb{1}\{\eta(x) > 1/2\}$ has the minimal risk.
- $h_Q(x) = y_Q^{(1)}$, where $(x_Q^{(1)}, y_Q^{(1)})$ is the sample and label of nearest neighbor of x in the quantized space S_Q.

Theorem: Generalization of 1-NN Classifier

Suppose (x, y) is a test sample. Q is a uniform quantizer with $\triangle$ between boarders and maximal gap g_Q. Under some technical conditions and with some constants c_1, c_2, with high probability,

$$E_{X,Y}[\mathcal{L}(h_Q(x))] \leq 2\mathcal{L}(h^*(x)) + c_1\left(\frac{\triangle}{g_Q}\sqrt{\frac{1+\omega}{1-\omega}}\right)^{\frac{k}{k+1}}(ne)^{-\frac{1}{k+1}}\sqrt{k} + \frac{c_2\triangle\sqrt{k}}{\sqrt{1-\omega}}.$$

Quantized Compressive 1-NN Classifier: Asymptotics

Theorem: Asymptotic Error of 1-NN Classifier

Let the cosine estimator $\hat{\rho}_Q = \frac{Q(x_1^T R)Q(R^T x_2)}{k}$, assume $\forall x_1, x_2$, $E[\hat{\rho}_Q(x_1, x_2)] = \alpha \rho_{x_1,x_2}$ for some $\alpha > 0$. As $k \to \infty$, we have

$$E_{X,Y,R}[\mathcal{L}(h_Q(x))] \leq E_{X,Y}[\mathcal{L}(h_S(x))] + r_k,$$

$$r_k = E[\sum_{i:x_i \in \mathcal{G}} \Phi(\frac{\sqrt{k}(\cos(x, x_i) - \cos(x, x^{(1)}))}{\sqrt{\xi^2_{x,x_i} + \xi^2_{x,x^{(1)}} - 2Corr(\hat{\rho}_Q(x, x_i), \hat{\rho}_Q(x, x^{(1)}))\xi_{x,x_i}\xi_{x,x^{(1)}}}})],$$

with $\xi^2_{x,y}/k$ the debiased variance of $\hat{\rho}_Q(x, y)$ and $\mathcal{G} = X/x^{(1)}$. $\mathcal{L}(h_S(x))$ is the risk of data space NN classifier, and $\Phi(\cdot)$ is the CDF of $N(0, 1)$.

- Let $x^{(1)}$ be the nearest neighbor of a test sample x. Under mild conditions, smaller debiased variance around $\rho = \cos(x, x^{(1)})$ leads to smaller generalization error.

Quantized Compressive Linear Classifier with (0,1)-loss

- H separates the space by a hyper-plane: $H(x) = \mathbb{1}\{h^T x > 0\}$.
- ERM classifiers: $\hat{H}(x) = \mathbb{1}\{\hat{h}^T x > 0\}, \hat{H}_Q(x) = \mathbb{1}\{\hat{h}_Q^T Q(R^T x) > 0\}$.

Theorem: Generalization of linear classifier

Under some technical conditions, with probability $(1 - 2\delta)$,

$$Pr[\hat{H}_Q(x) \neq y] \leq \hat{\mathcal{L}}_{(0,1)}(S, \hat{h}) + \frac{1}{\delta n} \sum_{i=1}^{n} f_{k,Q}(\rho_i) + C_{k,n,\delta},$$

where $f_{k,Q}(\rho_i) = \Phi(-\frac{\sqrt{k}|\rho_i|}{\xi_{\rho_i}})$, with ρ_i the cosine between training sample x_i and ERM classifier $\hat{h}$ in the data space, and $\xi_{\rho_i}^2/k$ the debiased variance of $\hat{\rho}_Q = \frac{Q(x_1^T R) Q(R^T x_2)}{k}$ at ρ_i.

- Small debiased variance around $\rho = 0$ lowers the bound.

Quantized Compressive Least Squares (QCLS) Regression

- Fixed design: $Y = X^T\beta + \epsilon$, with x_i fixed, ϵ i.i.d. $N(0,\gamma)$
- $L(\beta) = \frac{1}{n}E_Y[\|Y - X\beta\|^2]$, $L_Q(\beta_Q) = \frac{1}{n}E_{Y,R}[\|Y - Q(XR)\beta_Q\|^2]$.
- $\hat{L}(\beta) = \frac{1}{n}\|Y - X\beta\|^2$, $\hat{L}_Q(\beta_Q) = \frac{1}{n}\|Y - \frac{1}{\sqrt{k}}Q(XR)\beta_Q\|^2$. (given R)

Theorem: Generalization of QCLS

Let $\hat{\beta}^* = \underset{\beta\in\mathbb{R}^d}{\operatorname{argmin}}\, \hat{L}(\beta)$ and $\hat{\beta}_Q^* = \underset{\beta\in\mathbb{R}^k}{\operatorname{argmin}}\, \hat{L}_Q(\beta)$. Let $\Sigma = X^TX/k$, $k < n$. D_Q is the distortion of Q. Then we have

$$E_{Y,R}[L_Q(\hat{\beta}_Q^*)] - L(\beta^*) \leq \gamma\frac{k}{n} + \frac{1}{k}\|\beta^*\|_\Omega^2, \tag{1}$$

where $\Omega = [\frac{\xi_{2,2}-1+D_Q}{(1-D_Q)^2} - 1]\Sigma + \frac{1}{1-D_Q}I_d$, with $\|w\|_\Omega = \sqrt{w^T\Omega w}$ the Mahalanobis norm.

- Smaller distortion lowers the error bound.

Implications

- **1-NN classification**: In most applications, we should choose the quantizer with small debiased variance of inner product estimator $\hat{\rho}_Q = \frac{Q(R^T x)^T Q(R^T y)}{k}$ in high similarity region. $\Longrightarrow$ **Normalizing the quantized random projections (X_Q)** may help, see ref

  Xiaoyun Li and Ping Li, Random Projections with Asymmetric Quantization, NeurIPS 2019.

- **Linear classification**: we should choose the quantizer with small debiased variance of inner product estimate $\hat{\rho}_Q = \frac{Q(R^T x)^T Q(R^T y)}{k}$ at around $\rho = 0$. $\Longrightarrow$ First choice: Lloyd-Max quantizer.

- **Linear regression**: we should choose the quantizer with small distortion D_Q. $\Longrightarrow$ First choice: Lloyd-Max quantizer.

Experiments

Dataset	# samples	# features	# classes	Mean 1-NN ρ
BASEHOCK	1993	4862	2	0.6
orlraws10P	100	10304	10	0.9

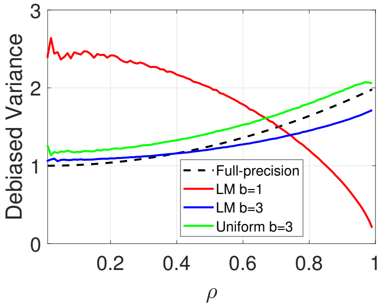

Figure 1: Empirical debiased variance of three quantizers.

Mean 1-NN ρ is the estimated $\cos(x, x^{(1)})$ from training set.

Quantized Compressive 1-NN Classification

Claim: smaller debiased variance at around $\rho = \cos(x, x^{(1)})$ is better.

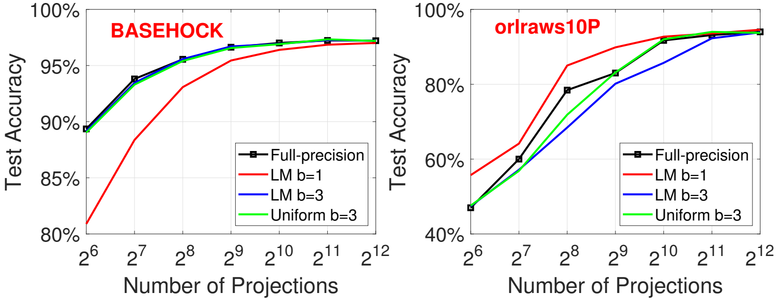

Figure 2: Quantized compressive 1-NN classification.

- Target ρ should be around:
  BASEHOCK: 0.6, where 1-bit quantizer has largest debiased variance.
  Orlraws10P: 0.9, where 1-bit quantizer has smallest debiased variance.

1-bit quantizer may generalize better than using more bits!

Quantized Compressive Linear SVM

Claim: smaller debiased variance at $\rho = 0$ is better.

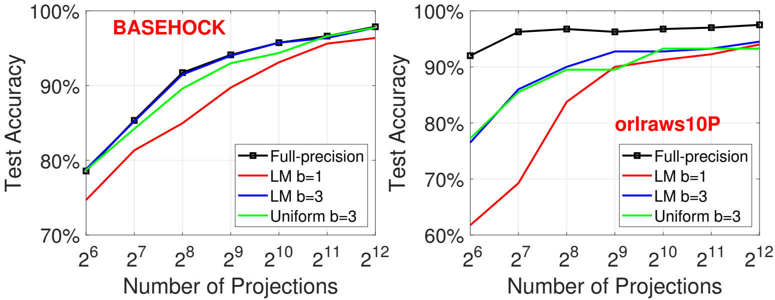

Figure 3: Quantized compressive linear SVM.

- At $\rho = 0$, red quantizer has much larger debiased variance than others $\Longrightarrow$ Lowest test accuracy on both datasets.

Quantized Compressive Linear Regression

Claim: smaller distortion is better.

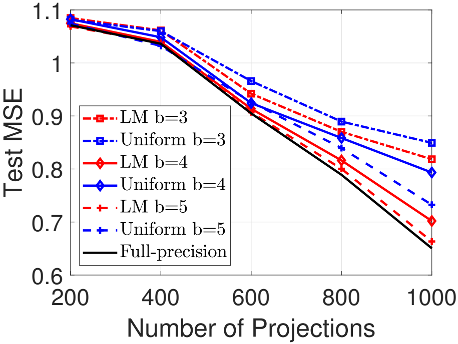

Figure 4: Test MSE of QCLS.

Blue: uniform quantizers. Red: Lloyd-Max (LM) quantizers.

- LM quantizer always outperforms uniform quantizer.
- **The order of test error agrees with the order of distortion.**