Randomized Subspace Embeddings for Learning Under Resource Constraints

Rajarshi Saha
Electrical Engineering
Stanford University

Joint work with Mert Pilanci (Stanford) and Andrea Goldsmith (Princeton)

January 24, 2021
More data and Bigger models...

at the expense of Resources: Memory, Computation, Bandwidth, ...

Most intelligence will be at the edge.
Training with large distributed datasets

Two pertinent questions:

1. Given a fixed bandwidth allocated for distributed training purposes, what is the information-theoretic limit on how quickly you can train a model?

2. What is an efficient training algorithm that can train a model as fast as (or nearly as fast as) what those limits dictate?
Deploying large models at the edge

Two more pertinent questions:

1. Given a memory-constraint, what is the information-theoretic limit on the performance when you compress a model?

2. What are some efficient algorithms to compress a model so that the performance of the compressed model deteriorates as little as possible?
Vector Quantization

Distributed Learning under Network Bandwidth constraints:
Quantize (pseudo) gradients.

Compress/Quantize a Model to deploy on Memory-constrained devices

Need a practical and efficient vector quantization scheme!
VQ for Learning: Challenges

• VQ must be agnostic to any distributional information.
  
  o Except for very well-structured problems with several assumptions, statistical information about the vector entries are not known.
  
  o **Fit a distribution?** Computationally intensive. Weights and gradients are constantly changing.

• **Universal Vector Quantization:** Do not want a complicated lattice. Ideally, complexity should be linear in dimension.

• **Lossy Source Coding:** Codebook should be easily available to decoder.

Given a bit-budget of $B \text{ bits per dimension}$, how do we quantize a vector in $\mathbb{R}^d$?
The problem of bit-allocation

• B-bits per dimension $\implies$ dB bits to quantize.

• How to allocate dB bits to d coordinates?

• Is it worth designing a sophisticated bit allocation scheme?
  
  o Vectors are constantly changing.

  o Hardware implementation of non-uniform quantizers is difficult.

\[
\begin{bmatrix}
16 \\
1 \\
0.01 \\
\vdots \\
5
\end{bmatrix}
\]

Orders of magnitude difference.

\[\|x\|_\infty = 1 \text{ and } B \text{ bits per dimension} \implies 2^B \text{ points per coordinate given by } v_i = -1 + (2i - 1)\Delta / 2, \ i = 1, \ldots, M, \ \Delta = 2/M.\]

\[Q(x) = [x'_1, \ldots, x'_N]^{\top}; \ x'_j \triangleq \arg \min_{y \in \{v_1, \ldots, v_M\}} |y - x_j|\]

\[
\sup_{x \in B_\infty^d(1)} \|Q(x) - x\|_2 = \frac{\Delta}{2} \sqrt{d}
\]

Uniform Quantizers
How do Random Embeddings help?

**y**: A vector in $\mathbb{R}^d$ whose coordinates can be arbitrarily large.

**Random embedding from $\mathbb{R}^d$ to $\mathbb{R}^D$**

**x**: A vector in $\mathbb{R}^D$ whose coordinates are equalized.

---

**Gaussian$^3$**

**Gaussian$^5$**

**Student-t (df = 1)**
Quantizing the Random Embeddings

\[ y \in \mathbb{R}^d \xrightarrow{\text{Embedding}} x \in \mathbb{R}^D \quad (D \geq d) \xrightarrow{\text{Uniformly Quantize}} \hat{x} \in \mathbb{R}^D \xrightarrow{\text{Inverse transform}} \hat{y} \in \mathbb{R}^d \]

With randomized embeddings

\[
\sup_{x \in B^d_\infty(1)} \| Q(x) - x \|_2 = O(1) \quad \text{(Computational complexity: O(d^2))}
\]

\[
\sup_{x \in B^d_\infty(1)} \| Q(x) - x \|_2 = O(\sqrt{\log d}) \quad \text{(Computational complexity: O(d log d))}
\]

Worst-case quantization error is dimension-independent or weak-logarithmic dependence!
Part 1
Model Compression
Compressing Linear Models

Observations $\in \mathbb{R}^n \rightarrow X = W\theta + v \rightarrow$ Noise $\in \mathbb{R}^n$

Arbitrary measurement matrix $\in \mathbb{R}^{n \times d}$

Ground-truth model $\in \mathbb{R}^d$

Worker estimates model $\theta$ and can send it to the server using only dB bits.

$$\tilde{\theta} := \arg\min_{s \in S} \|X - Ws\|_2^2$$

$$R(\tilde{\theta}, \theta) = \mathbb{E}_X \left[ \frac{1}{d} \|\tilde{\theta} - \theta\|_2^2 \right]$$

(Risk of any quantized model)
Information-Theoretic Limits

**Definition**

An \((n, d, B)\)-learning code \(Q : \mathbb{R}^n \rightarrow \Theta\) is defined to be the composition of encoder and decoder mappings \(E\) and \(D\), such that for any given data \(X \in \mathbb{R}^n\), \(Q(X) \equiv D(E(X)) \in \Theta\).

Minimax risk:

\[
R_{W, B, \sigma, c} := \liminf_{d \to \infty} \inf_{Q \in \mathcal{Q}_{n, d, B}} \sup_{\theta \in \Theta} R(Q(X), \theta)
\]

**Theorem**

For \(B > 0\), \(\sigma > 0\), \(c > 0\), and \(W \in \mathbb{R}^{n \times d}\) with minimum and maximum singular values as \(\sigma_m\) and \(\sigma_M\) respectively, the asymptotic minimax risk can be lower bounded as:

\[
R_{W, B, \sigma, c} \geq \frac{c^2 \sigma^2}{\sigma^2 + c^2 \sigma_m^2} + \frac{c^4 \sigma_m^2}{\sigma^2 + c^2 \sigma_m^2} \cdot 2^{-2B}.
\]
## Optimally Compressing Linear Models

<table>
<thead>
<tr>
<th>Learning Codes</th>
<th>Performance Guarantee (holds w.h.p.)</th>
<th>Computational Complexity</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Projections on the Unit Sphere</td>
<td>$R(\theta, \hat{\theta}) \leq \frac{c^2 \sigma^2}{\sigma^2 + c^2 \sigma^2_{\min}} + \frac{c^4 \sigma^2_{\max}}{\sigma^2 + c^2 \sigma^2_{\max}} 2^{-2B}$</td>
<td>exp (d)</td>
<td>Tight w.r.t. lower bound.</td>
</tr>
<tr>
<td>Democratic Quantized Estimation</td>
<td>$R(\theta, \hat{\theta}) \leq \frac{2c^2 \sigma^2}{\sigma^2 + c^2 \sigma^2_{\min}} + \frac{16K_u c^4 \sigma^2_{\max}}{\sigma^2 + c^2 \sigma^2_{\max}} 2^{-\frac{2B}{\lambda}}$</td>
<td>O (d^2)</td>
<td>Optimal within constant factors.</td>
</tr>
<tr>
<td>Near-Democratic Quantized Estimation</td>
<td>$R(\theta, \hat{\theta}) \leq \frac{2c^2 \sigma^2}{\sigma^2 + c^2 \sigma^2_{\min}} + \frac{32 \log(2d)c^4 \sigma^2_{\max}}{\sigma^2 + c^2 \sigma^2_{\max}} 2^{-\frac{2B}{\lambda}}$</td>
<td>O (d \cdot \log d)</td>
<td>Near linear-time; Mild logarithmic dependence.</td>
</tr>
</tbody>
</table>
How tight are the Lower and Upper bounds?

$W$: Identity, $\theta$: Gaussian

$W$: Perturbed orthonormal, $\theta$: Gaussian
Compressing Heavy-Tailed Models

**MSE of Quantized Estimation vs. Bit-budget**

- **W**: Perturbed orthonormal, $\theta$: Gaussian
- **W**: Perturbed orthonormal, $\theta$: Student-t (df = 1)
Part 2
Communication-Constrained Distributed Optimization
Iterative First-Order Optimization Protocols

• How to design **efficient algorithms** to achieve the optimal convergence rate when the worker can communicate to the server using **only dB bits per round**?
L - smooth and $\mu$ - strongly convex objectives

Minimax convergence rate:

$$C(B) \triangleq \inf_{\pi \in \Pi_B} \limsup_{T \to \infty} \sup_{f \in \mathcal{F}_{\mu,L,D}} \left( \frac{\|x_T(\pi) - x_f\|_2}{D} \right)^{\frac{1}{T}}$$

Information-theoretic limit

(“Differentially Quantized Gradient Methods”, Chung-Yi Lin and Victoria Kostina and Babak Hassibi, 2021)

$$C(B) \geq \max\{\sigma, 2^{-B}\}$$

<table>
<thead>
<tr>
<th>Optimization Algorithm</th>
<th>Performance Guarantee</th>
<th>Computational Complexity</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQ-PSGD</td>
<td>$\left( \frac{x_T - x_f}{D} \right)^{\frac{1}{T}} \leq \max{\sigma, c_1 \cdot 2^{-B}}$</td>
<td>$O(d^2)$</td>
<td>Optimal within constant factors.</td>
</tr>
<tr>
<td>Near DQ-PSGD</td>
<td>$\left( \frac{x_T - x_f}{D} \right)^{\frac{1}{T}} \leq \max{\sigma, c_2 \sqrt{\log d} \cdot 2^{-B}}$</td>
<td>$O(d \cdot \log d)$</td>
<td>Near linear-time; Mild logarithmic dependence.</td>
</tr>
</tbody>
</table>
General convex and non-smooth objectives

Minimax suboptimality gap: \( E(T, B) \triangleq \inf_{\pi \in \Pi_{T, B}} \sup_{(f, \mathcal{O})} \mathbb{E} f(x(\pi)) - f(x^*) \)

Information-theoretic limit
("Limits on Gradient Compression for Stochastic Optimization" Prathamesh Mayekar and Himanshu Tyagi, 2020)

\[ E(T, B) \geq \frac{cD\sigma}{\sqrt{T} \sqrt{\min\{1, B\}}} \]

<table>
<thead>
<tr>
<th>Optimization Algorithm</th>
<th>Performance Guarantee</th>
<th>Computational Complexity</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQ-PSGD</td>
<td>( E(T, B) \leq \frac{c_1 D\sigma}{\sqrt{T} \sqrt{\min{1, B}}} )</td>
<td>( O(d^2) )</td>
<td>Optimal within constant factors.</td>
</tr>
<tr>
<td>Near DQ-PSGD</td>
<td>( E(T, B) \leq \frac{c_2 D\sigma \sqrt{\log d}}{\sqrt{T} \sqrt{\min{1, B}}} )</td>
<td>( O(d \cdot \log d) )</td>
<td>Near linear-time; Mild logarithmic dependence.</td>
</tr>
</tbody>
</table>
Numerical Results

Least squares: Synthetic data

Least squares: MNIST
Numerical Results (contd..)

Support Vector Machine: Synthetic data

Support Vector Machine: MNIST
General stochastic compression schemes
Thank you!