Memory Efficient EBM Training

Anonymous Author(s)
Affiliation
Address
email

Abstract

To be completed...

1 Introduction

2 Related Work

Energy Based Modeling

Distributed Optimization

Compression and Quantization

3 Distributed and Private EBM Training

3.1 Compression Methods for Distributed and Private Optimization

Definition 1 (Top-k)**.** *For $x \in \mathbb{R}^d$, denote $\mathcal{S}$ as the size-k set of $i \in [d]$ with largest k magnitude $|x_i|$. The **Top-**k compressor is defined as $\mathcal{C}(x)_i = x_i$, if $i \in \mathcal{S}$; $\mathcal{C}(x)_i = 0$ otherwise.*

Definition 2 (Block-Sign)**.** *For $x \in \mathbb{R}^d$, define M blocks indexed by $\mathcal{B}_i$, $i = 1, ..., M$, with $d_i := |\mathcal{B}_i|$. The **Block-Sign** compressor is defined as $\mathcal{C}(x) = [sign(x_{\mathcal{B}_1})\frac{\|x_{\mathcal{B}_1}\|_1}{d_1}, ..., sign(x_{\mathcal{B}_M})\frac{\|x_{\mathcal{B}_M}\|_1}{d_M}]$.*

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

3.2 Main Algorithm

Algorithm 1: Distributed and private EBM

Input: Total number of iterations T, number of MCMC transitions K and of samples M, sequence of global learning rate $\{\eta_t\}_{t>0}$, sequence of MCMC stepsizes $\gamma_{k k>0}$, initial value θ_0, MCMC initialization $\{z_0^m\}_{m=1}^M$. Set of selected devices $\mathcal{D}^t$.

Output: Vector of fitted parameters θ_{T+1}.

Data: $\{x_i^p\}_{i=1}^{n_p}$, n_p number of observations on device p. $n = \sum_{p=1}^P n_p$ total.

---

for $t = 1$ *to* T **do**

/* Happening on distributed devices */

for *For device* $p \in \mathcal{D}^t$ **do**

Draw M negative samples $\{z_K^{p,m}\}_{m=1}^M$ // local langevin diffusion

for $k = 1$ *to* K **do**

$$z_k^{p,m} = z_{k-1}^{p,m} + \gamma_k/2\nabla_z f_{\theta_t}(z_{k-1}^{p,m})^{p,m} + \sqrt{\gamma_k}\mathsf{B}_k^p \ ,$$

where B_k^p denotes the Brownian motion (Gaussian noise).

Assign $\{z_t^{p,m}\}_{m=1}^M \leftarrow \{z_K^{p,m}\}_{m=1}^M$.

Sample M positive observations $\{x_i^p\}_{i=1}^M$ from the empirical data distribution.

Compute the gradient of the empirical log-EBM // local - and + gradients

$$\delta^p = \frac{1}{M}\sum_{i=1}^M \nabla_\theta f_{\theta_t}(x_i^p) - \frac{1}{M}\sum_{m=1}^M \nabla_\theta f_{\theta_t}(z_K^{p,m})$$

Use black box compression operators

$$\Delta^p = \mathcal{C}(\delta^p)$$

Devices broadcast Δ^p to Server

/* Happening on the central server */

Aggregation of devices gradients: $\nabla \log p(\theta_t) \approx \frac{1}{|\mathcal{D}^t|}\sum_{p=1}^{|\mathcal{D}^t|}\Delta^p$.

Update the vector of global parameters of the EBM: $\theta_{t+1} = \theta_t + \eta_t \nabla \log p(\theta_t)$

Output: Vector of fitted parameters θ_{T+1}

---

4 Convergence Guarantees

Recall that the goal of this paper is to train an energy-based model where the data is distributed on P devices. Formally, given a stream of input data noted $x \in \mathbb{R}^d$ such that $x = \{x_i^p\}_{i=1}^{n_p}$, n_p number of observations on device p. $n = \sum_{p=1}^P n_p$ total., the model reads:

$$p(x,\theta) = \prod_{p=1}^P \frac{1}{Z_p(\theta)}\exp(-U_\theta^p(x)) \ , \tag{1}$$

where $\theta \in \Theta \subset \mathbb{R}^d$ denotes the global parameters vector of our model and $Z(\theta) = \prod_{p=1}^P Z_p(\theta) := \prod_{p=1}^P \int_x \exp(-U_\theta^p(x))\mathrm{d}x$ is the normalizing constant with respect to x. $U_\theta^p(x)$ denotes the energy function for device p is parameterized by θ and takes as input an image x.

We now establish a non-asymptotic convergence result for the set of parameters $\{\theta_t\}_{t=1}^T$.

Beforehand, we provide mild assumptions on our model

5 Numerical Experiments

6 Conclusion

A Appendix