

---

# Memory Efficient EBM Training

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 To be completed...

## 2 1 Introduction

## 3 2 Related Work

### 4 Energy Based Modeling

### 5 Distributed Optimization

### 6 Compression and Quantization

## 7 3 Distributed and Private EBM Training

### 8 3.1 Compression Methods for Distributed and Private Optimization

9 **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  
10  $|x_i|$ . The **Top- $k$**  compressor is defined as  $\mathcal{C}(x)_i = x_i$ , if  $i \in \mathcal{S}$ ;  $\mathcal{C}(x)_i = 0$  otherwise.

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

### 13 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  $\theta_0$ , MCMC initialization  $\{z_0^m\}_{m=1}^M$ . Set of selected devices  $\mathcal{D}^t$ .

**Output:** Vector of fitted parameters  $\theta_{T+1}$ .

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

---

```

1
2 for  $t = 1$  to  $T$  do
3     /* Happening on distributed devices */
4     for For device  $p \in \mathcal{D}^t$  do
5         Draw  $M$  negative samples  $\{z_K^{p,m}\}_{m=1}^M$  // local langevin diffusion
6         for  $k = 1$  to  $K$  do
7              $z_k^{p,m} = z_{k-1}^{p,m} + \gamma_k/2\nabla_z f_{\theta_t}(z_{k-1}^{p,m})^{p,m} + \sqrt{\gamma_k}B_k^p$ ,
8             where  $B_k^p$  denotes the Brownian motion (Gaussian noise).
9             Assign  $\{z_t^{p,m}\}_{m=1}^M \leftarrow \{z_K^{p,m}\}_{m=1}^M$ .
10            Sample  $M$  positive observations  $\{x_i^p\}_{i=1}^M$  from the empirical data distribution.
11            Compute the gradient of the empirical log-EBM // local - and + gradients
12
13            
$$\delta^p = \frac{1}{M} \sum_{i=1}^M \nabla_{\theta} f_{\theta_t}(x_i^p) - \frac{1}{M} \sum_{m=1}^M \nabla_{\theta} f_{\theta_t}(z_K^{p,m})$$

14
15            Use black box compression operators
16
17            
$$\Delta^p = \mathcal{C}(\delta^p)$$

18
19            Devices broadcast  $\Delta^p$  to Server
20
21            /* Happening on the central server */
22
23            Aggregation of devices gradients:  $\nabla \log p(\theta_t) \approx \frac{1}{|\mathcal{D}^t|} \sum_{p=1}^{|\mathcal{D}^t|} \Delta^p$ .
24
25            Update the vector of global parameters of the EBM:  $\theta_{t+1} = \theta_t + \eta_t \nabla \log p(\theta_t)$ 
26
27 Output: Vector of fitted parameters  $\theta_{T+1}$ 

```

---

### 15 4 Convergence Guarantees

16 Recall that the goal of this paper is to train an energy-based model where the data is distributed on  
17  $P$  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  
18 of observations on device  $p$ .  $n = \sum_{p=1}^P n_p$  total., the model reads:

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

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

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

23 Beforehand, we provide a mild assumption on our model

24 **Assumption 1.** (*Smoothness of the energy function*)

25 We also formalize some properties of the sampling procedure used to obtain synthetic samples dur-  
26 ing training.

27 **Assumption 2.** (*Bounded MC noise*)

28 **Assumption 3.** (*Geometric ergodicity of CD-1*)

29 **5 Numerical Experiments**

30 **6 Conclusion**

