Memory Efficient EBM Training

Anonymous Author(s)
Affiliation
Address
email

Abstract

To be completed...

1 Introduction

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 $\mathcal{C}(x) = [sign(x_{\mathcal{B}_1})\frac{\|x_{\mathcal{B}_1}\|_1}{d_1}, ..., sign(x_{\mathcal{B}_M})\frac{\|x_{\mathcal{B}_M}\|_1}{d_M}]$.*

Algorithm 1 EFF-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 $\{z_0^m\}_{m=1}^M$ and observations $\{x_i\}_{i=1}^n$.

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

  Draw M samples $\{z_t^m\}_{m=1}^M$ from the objective potential via Langevin diffusion:

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

    Use black box compression operators:

$$\tilde{g}_{k-1}^m = \mathcal{C}(\nabla_z f_{\theta_t}(z_{k-1}^m))$$

    Construct the Markov Chain as follows:

$$z_k^m = z_{k-1}^m + \gamma_k/2\tilde{g}_{k-1}^m + \sqrt{\gamma_k}\mathsf{B}_k \ , \tag{1}$$

    where B_t denotes the Brownian motion (Gaussian noise).

  end for

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

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

  Compute the gradient of the empirical log-EBM:

$$\nabla \log p(\theta_t) = \mathbb{E}_{p_{\text{data}}}\left[\nabla_\theta f_{\theta_t}(x)\right] - \mathbb{E}_{p_\theta}\left[\nabla_{\theta_t} f_\theta(z_t)\right] \approx \frac{1}{n}\sum_{i=1}^n \nabla_\theta f_{\theta_t}\left(x_i\right) - \frac{1}{m}\sum_{i=1}^m \nabla_\theta f_{\theta_t}\left(z_t^m\right) \ .$$

  Update the vector of global parameters of the EBM:

$$\theta_{t+1} = \theta_t + \eta_t \nabla \log p(\theta_t) \ .$$

end for

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

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

2 Conclusion

A Appendix