Weekly Report KARIMI-2021-11-19

My work this week has mainly been towards

1. Fed-LAMB (more experiments)
2. ICLR22 Rebuttal (Lowest score questions)
3. Distributed and Private EBM

1 Fed-LAMB (more experiments)

We included the Adaptive Federated Optimization of [1] to our several experiments. Done for single GPU and currently implementing the new baseline in the distributed settings.

2 ICLR22 Rebuttal (Lowest score questions)

Please refer to the Overleaf projects for the rebuttal

3 Distributed and Private EBM

The distributed aspect of the algorithm is not an issue, either practically (since in practice we actually always train EBM on several GPUs). Yet, the compressed component introduces some issues.

Practical: In practice, the compression obviously implies that natural images are no longer visually appealling. Hence for the training part it does not matter since we only use those negative samples to compute a gradient but for the testing phase, the evaluation of FID of Inception scores is falsified.

Theoretical: In theory, we must assume ergodicity of the MCMC used in Algorithm 3. An issue that is not easy to deal with is that with Jianwen we concluded that using CD-1 (Contrastive divergence with only one iteration, i.e. $K = 1$ in Algorithm 3) is the most interesting. Yet, assuming ergodicity and mixing of the chain with CD-1 is a bit farfetched. So either I do assume it, or I don't have recourse to CD-1.

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

References

[1] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.