2025年7月15日

老师的任务：

0. Ok. 你写一下系统搭建和抓数据的方法 以后我们开源和写论文需要 我们backtrack一下，先总结一下： 1) GPU之间通信的数据是什么格式，每个数据有多少bits; 2) 你存下来是什么格式，每个数据多少bits; 3) 你熵编码以后的数据平均每个数据长度多少？ 4) 数据的分布的shape parameter是什么？理论entropy多少，你算出来的entropy多少？ 你系统部署和数据存储的过程的说明在写了么？
1. entropy 的理论值（根据你的shape parameter）和实际数据你算出来的结果尽快对比一样 entropy是由概率分布定义的，你有分布怎么算不出entropy? What are you talking about? 这里我们关心的是量化GGD以后的离散entropy，你去看看1999年trans it的那个paper -- [哪一个paper?]

从7月9日未完成的工作恢复：

```
0. 你把压缩以后的平均长度和entropy对比一下
    -- [已经支持对发生负优化的bucket进行分析，但是对于目前所有的bucket还暂无支持]
    -- [2025年7月15日：已开展对于所有可用数据的计算]
0. 你说的压缩以后反而变大的case对应的分布的entropy和shape是什么?
    -- [shape部分发生了overflow错误，无法收敛，仍需修复。但是entropy计算完毕，两者似乎
没有关系，详见附上的scatter plot]
    -- [2025年7月15日：overflow问题已经修复，已开展对于所有可用数据的计算]
1. 修好pt2h5的逻辑
    -- [未完成]
    -- [2025年7月15日：暂时搁置]
2. 更换模型为gemma或者其他模型(您建议是？)
    -- [未完成]
    -- [2025年7月15日：详见今日安排]
4. 尝试模仿pytorch的 bf16 compression hook，将EG Compression加入训练流程？  --
[0,1,2,3完成后再做]
```

今天的安排：

0. matlab与python配合问题： matlab调用python -- 搁置 python调用matlab -- 等待官方指令：

   (base) [zg2598@hpclogin zg2598]$ cd "/gpfsnyu/spack/opt/spack/linux-rhel8-icelake/contribute/matlab/2023b/parallel-matlab" (base) [zg2598@hpclogin parallel-matlab]$ ls appdata bin cefclient derived extern help interprocess java license_agreement.txt licenses patents.txt platform polyspace resources rtw simulink sys toolbox trademarks.txt ui VersionInfo.xml (base) [zg2598@hpclogin parallel-matlab]$ ^C (base) [zg2598@hpclogin parallel-matlab]$ cd "/gpfsnyu/spack/opt/spack/linux-rhel8-icelake/contribute/matlab/2024a/parallel-matlab/extern/engines/python" (base) [zg2598@hpclogin python]$ ls dist pyproject.toml setup.py (base) [zg2598@hpclogin python]$ conda activate matlab (matlab) [zg2598@hpclogin python]$ python setup.py install --user 【这是最后一步】

1. 修复shape parameter的overflow问题 -- 不to_int的话，会发生overflow -- 进行了to_int后，就不会发生overflow了（至少在28MB数量级上没问题。在130MB的数量级上也没有任何问题！那么我的思路是可行

的） -- 所以是数据太小了吗？ -- 观察到scaling后的数据不会影响shape paramter (gamma) 的拟合结果，其他的参数会线性放大。 -- 观察到是否只对1-99百分位数拟合对于结果是有影响的，如果filter了，那么shape就是1.2，反之为0.8左右。 -- 结论，开两个session，一个session负责filtered，另一个不filtered -- 很久很久以前的上一次（6月份）的尝试中发现，如果只取10%的数据，拟合出来的结果相差无几（目视，没有数字依据） -- 要不，我们尝试下，加快进程？

2. 计算各个bucket的shape parameter -- [已开展]
3. 阅读文章，开始计算entropy -- [等待具体是哪一篇文章，以便计算离散熵]
4. 更换模型为gemma —— 有官方教程，学起来方便 -- [2025年7月15日14:31:05：google/gemma-3-4b-pt的下载已经开始；预计完成时间：1hour] -- [2025年7月15日15:25:16：google/gemma-3-4b-pt的下载已经完成，开始上传至HPC，预计完成时间20min] -- [2025年7月15日15:39:34：教程中的数据集已经下载完毕] -- [开始调试……]

附注：

matlab的报错：

```
pyenv('Version', '/gpfsnyu/home/zg2598/.conda/envs/matlab/bin/python')
```

ans =

PythonEnvironment with properties:

```
       Version: "3.10"
    Executable: "/gpfsnyu/home/zg2598/.conda/envs/matlab/bin/python"
       Library: "/gpfsnyu/home/zg2598/.conda/envs/matlab/lib/libpython3.10.so"
          Home: "/gpfsnyu/home/zg2598/.conda/envs/matlab"
        Status: NotLoaded
ExecutionMode: OutOfProcess
```

```
pyenv
```

ans =

PythonEnvironment with properties:

```
       Version: "3.10"
    Executable: "/gpfsnyu/home/zg2598/.conda/envs/matlab/bin/python"
       Library: "/gpfsnyu/home/zg2598/.conda/envs/matlab/lib/libpython3.10.so"
          Home: "/gpfsnyu/home/zg2598/.conda/envs/matlab"
        Status: NotLoaded
ExecutionMode: OutOfProcess
```

```
torch = py.importlib.import_module('torch'); torch
```

torch =

sys:1: UserWarning: TypedStorage is deprecated. It will be removed in the future and UntypedStorage will be the only storage class. This should only matter to you if you are using storages directly. To access UntypedStorage directly, use tensor.untyped_storage() instead of tensor.storage() Python module with properties: 省略了

```
<module 'torch' from '/gpfsnyu/home/zg2598/.conda/envs/matlab/lib/python3.10/site-packages/torch/__init__.py'>
```

```
pt_path = '/gpfsnyu/scratch/zg2598/Qwen/OUT/COMMUNICATION_LOG/R_1_E_0_S_9_B_79.pt';
pt_path
```

pt_path =

```
'/gpfsnyu/scratch/zg2598/Qwen/OUT/COMMUNICATION_LOG/R_1_E_0_S_9_B_79.pt'
```

```
pt_path_py = py.str(pt_path)
```

pt_path_py =

Python str with no properties.

```
/gpfsnyu/scratch/zg2598/Qwen/OUT/COMMUNICATION_LOG/R_1_E_0_S_9_B_79.pt
```

```
some_tensor = torch.load(pt_path_py) Error using serialization>init Python Error: RuntimeError: PytorchStreamReader failed reading zip archive: invalid parameter
```

Error in serialization>load (line 1486)

疑问：

1. 我们是否需要抽选单个bucket中一部分的数据来进行GGD拟合？