

Probabilistic DIGing

I. BRIEF DESCRIPTION

The optimization problem is given by

$$\textbf{(P1)} \quad \Theta^* := \arg \min_{\Theta} \sum_{n=1}^N f_n(\Theta), \quad (1)$$

where $\Theta \in \mathbb{R}^{d \times 1}$ is the model parameter and $f_n : \mathbb{R}^d \rightarrow \mathbb{R}$ is a local function composed of data stored at worker n .

A. Linear Regression

1) *Loss Function*: In this case, the local cost function at worker n is explicitly given by

$$f_n(\theta) = \frac{1}{2} \|\mathbf{X}_n \theta - \mathbf{y}_n\|^2, \quad (2)$$

where $\mathbf{X}_n \in \mathbb{R}^{s \times d}$ and $\mathbf{y}_n \in \mathbb{R}^{s \times 1}$ are private for each worker $n \in \mathcal{V}$ where s represents the size of the data at each worker.

2) *Datasets*: In this task, we will consider the following datasets (see Table 2 of <http://cacr.uwaterloo.ca/techreports/2019/cacr2019-05.pdf>):

- Boston: https://github.com/benchopt/benchmark_ols/tree/master/datasets
- Wine Quality: <https://archive.ics.uci.edu/ml/datasets/wine+quality>

B. Logistic Regression

1) *Loss Function*: In this subsection, we consider the L_2 -regularized binary logistic regression task. We assume that each worker n owns a data matrix $\mathbf{X}_n = (\mathbf{x}_{n,1}, \dots, \mathbf{x}_{n,s})^T \in \mathbb{R}^{s \times d}$ along with the corresponding labels vector $\mathbf{y}_n = (y_{n,1}, \dots, y_{n,s}) \in \{-1, 1\}^s$. The local cost function for worker n is then given by

$$f_n(\theta) = \frac{1}{s} \sum_{j=1}^s \log(1 + \exp(-y_{n,j} \mathbf{x}_{n,j}^T \theta)) + \frac{\lambda}{2} \|\theta\|_2^2, \quad (3)$$

where λ is the regularization parameter.

2) *Datasets*: In this task, we will consider the following datasets:

- a1a: https://github.com/konstmish/opt_methods/tree/master/datasets
- mushrooms: https://github.com/konstmish/opt_methods/tree/master/datasets
- w8a: https://github.com/konstmish/opt_methods/tree/master/datasets
- madelon: https://github.com/benchopt/benchmark_logreg_l2/tree/master/datasets
- covtype: https://github.com/benchopt/benchmark_logreg_l2/tree/master/datasets

Dataset	Task	Model Size (d)	Number of Instances	Number of Workers (N)
	linear regression	50	1200	
	linear regression			
	logistic regression			
	logistic regression			

TABLE I: List of datasets used in the numerical experiments.

C. Websites for datasets

- <https://www.openml.org/>
- <https://archive.ics.uci.edu/ml/datasets.php>

D. Tasks

- Dataset preparation: download at least 4 datasets for each task (regression/classification) in the format X and y (as .npz files if possible, otherwise any format you are comfortable with).
- Strongly convex case
 - Start simple: implement DGD and DIGing for the linear regression + logistic regression cases.
 - Implement the probabilistic DIGing and experiment with different choice of the probability p .
 - (i) $p^k = \frac{a}{a+k}$, $a > 0$.
 - (ii) $p^k = \exp\left(-\frac{k}{T}\right)$, $T > 0$.
 - Plots (need to think more about it):
 - (i) train/test loss (accuracy) or residuals vs. number of iterations.
 - (ii) train/test loss (accuracy) or residuals vs. cumulative communication cost. How to define the cumulative communication cost in our context?

$$C_T^k = \sum_{n=1}^N d_n c^k \quad (4)$$

where c^k is the size of the variables exchanged at iteration k .

- Proof attempt.
- Stochastic version of above (double-check the literature).
- Non-convex case.