A Provably Accurate Randomized Sampling Algorithm for Logistic Regression (Supplementary Material)

A Proof of Lemma 5

Lemma A1. *Let matrix* $\mathbf{U}$ *and vector* $\mathbf{x}$ *are as defined in Lemma 5. If the sketching matrix* $\mathbf{S} \in \mathbb{R}^{s\times n}$ *is constructed using Algorithm 1. Then, for* $i = 1, \ldots, d$*, we have*

$$(a)\ \ \mathbb{E}\left((\mathbf{U}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{x})_i\right) = (\mathbf{U}^\mathsf{T}\mathbf{x})_i \tag{21}$$

$$(b)\ \ \mathrm{Var}\left((\mathbf{U}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{x})_i\right) = \frac{1}{s}\sum_{j=1}^{n}\frac{(\mathbf{U}^\mathsf{T})_{ij}^2 x_j^2}{\pi_j} - \frac{1}{s}(\mathbf{U}^\mathsf{T}\mathbf{x})_i^2$$

Proof of Part (a). From the definition of expectation, we have,

$$\begin{aligned}\mathbb{E}\left((\mathbf{U}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{x})_i\right) &= \mathbb{E}\left((\mathbf{U}^\mathsf{T})_{i*}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{x}\right) = \mathbb{E}\left(\sum_{t=1}^{s}\frac{(\mathbf{U}^\mathsf{T})_{ij_t}x_{j_t}}{s\pi_{j_t}}\right) = \sum_{t=1}^{s}\mathbb{E}\left(\frac{(\mathbf{U}^\mathsf{T})_{ij_t}x_{j_t}}{s\pi_{j_t}}\right)\\ &= \sum_{t=1}^{s}\sum_{j=1}^{n}\frac{(\mathbf{U}^\mathsf{T})_{ij}x_j}{s\pi_j}\,\pi_j = \frac{1}{s}\sum_{t=1}^{s}\sum_{j=1}^{n}(\mathbf{U}^\mathsf{T})_{ij}x_j = \sum_{j=1}^{n}(\mathbf{U}^\mathsf{T})_{ij}x_j = (\mathbf{U}^\mathsf{T}\mathbf{x})_i\end{aligned} \tag{22}$$

□

Proof of Part (b). Since the samples are drawn independently (with replacement), from the *additivity of variances* property we have,

$$\begin{aligned}\mathrm{Var}\left((\mathbf{U}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{x})_i\right) &= \mathrm{Var}\left((\mathbf{U}^\mathsf{T})_{i*}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{x}\right) = \mathrm{Var}\left(\sum_{t=1}^{s}\frac{(\mathbf{U}^\mathsf{T})_{ij_t}x_{j_t}}{s\pi_{j_t}}\right) = \sum_{t=1}^{s}\mathrm{Var}\left(\frac{(\mathbf{U}^\mathsf{T})_{ij_t}x_{j_t}}{s\pi_{j_t}}\right)\\ &= \sum_{t=1}^{s}\left(\mathbb{E}\left(\frac{(\mathbf{U}^\mathsf{T})_{ij_t}x_{j_t}}{s\pi_{j_t}}\right)^2 - \mathbb{E}^2\left(\frac{(\mathbf{U}^\mathsf{T})_{ij_t}x_{j_t}}{s\pi_{j_t}}\right)\right)\\ &= \sum_{t=1}^{s}\left(\sum_{j=1}^{n}\left(\frac{(\mathbf{U}^\mathsf{T})_{ij}x_j}{s\pi_j}\right)^2\cdot\pi_j - \left(\sum_{j=1}^{n}\frac{(\mathbf{U}^\mathsf{T})_{ij}x_j}{s\pi_j}\cdot\pi_j\right)^2\right)\\ &= \frac{1}{s}\sum_{j=1}^{n}\frac{(\mathbf{U}^\mathsf{T})_{ij}^2x_j^2}{\pi_j} - \frac{1}{s}(\mathbf{U}^\mathsf{T}\mathbf{x})_i^2\end{aligned} \tag{23}$$

□

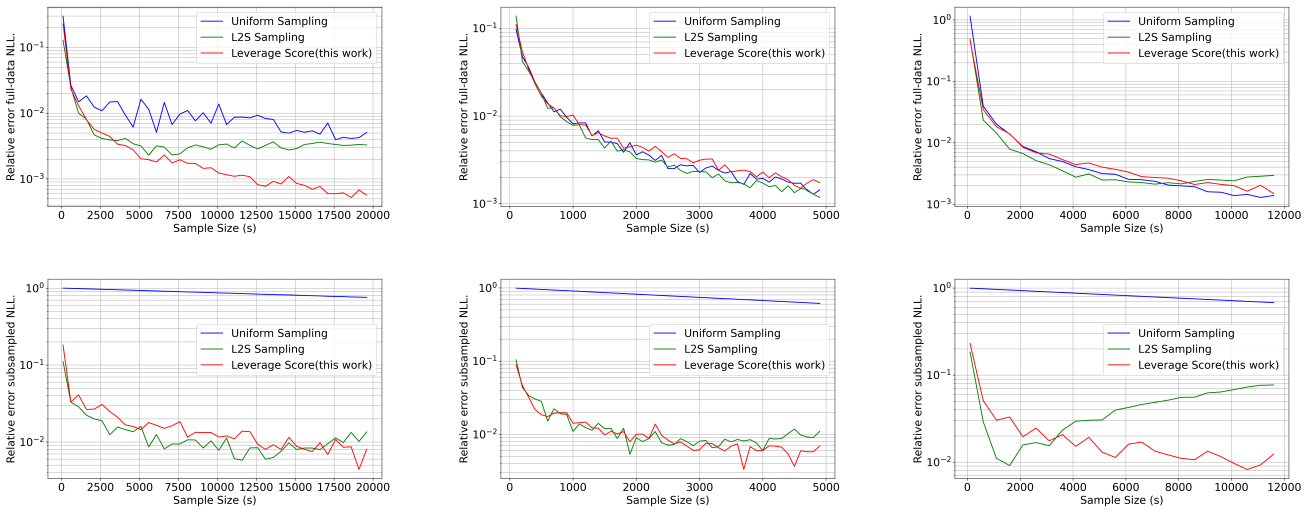
Proof of Lemma 5. Now we are ready to prove Lemma 5. First, from Lemma A1, we already know that $(\mathbf{U}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{x})_i$ is an unbiased estimator of $(\mathbf{U}^\mathsf{T}\mathbf{x})_i$, for all $i = 1, \ldots d$, *i.e.*, $\mathbb{E}\left((\mathbf{U}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{x})_i\right) = (\mathbf{U}^\mathsf{T}\mathbf{x})_i$. Using this fact, we rewrite the left hand side as

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{U}^\mathsf{T}\mathbf{S}^\top\mathbf{S}\mathbf{x} - \mathbf{U}^\mathsf{T}\mathbf{x}\right\|_2^2\right] &= \mathbb{E}\left[\sum_{i=1}^{d}(\mathbf{U}^\mathsf{T}\mathbf{S}^\top\mathbf{S}\mathbf{x} - \mathbf{U}^\mathsf{T}\mathbf{x})_i^2\right] = \sum_{i=1}^{d}\mathbb{E}\left[(\mathbf{U}^\mathsf{T}\mathbf{S}^\top\mathbf{S}\mathbf{x} - \mathbf{U}^\mathsf{T}\mathbf{x})_i\right]^2\\ &= \sum_{i=1}^{d}\mathbb{E}\left[(\mathbf{U}^\mathsf{T}\mathbf{S}^\top\mathbf{S}\mathbf{x})_i - (\mathbf{U}^\mathsf{T}\mathbf{x})_i\right]^2 = \sum_{i=1}^{d}\mathbb{E}\left[(\mathbf{U}^\mathsf{T}\mathbf{S}^\top\mathbf{S}\mathbf{x})_i - \mathbb{E}(\mathbf{U}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{x})_i\right]^2\\ &= \sum_{i=1}^{d}\mathrm{Var}\left((\mathbf{U}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{x})_i\right)\end{aligned} \tag{24}$$

Now, combining *Part (b)* of Lemma A1 and eqn. (24), we further have

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{U}^\mathsf{T}\mathbf{S}^\top\mathbf{S}\mathbf{x} - \mathbf{U}^\mathsf{T}\mathbf{x}\right\|_2^2\right] &= \sum_{i=1}^{d}\sum_{j=1}^{n}\left[\frac{1}{s}\frac{(\mathbf{U}^\mathsf{T})_{ij}^2x_j^2}{\pi_j} - \frac{1}{s}(\mathbf{U}^\mathsf{T}\mathbf{x})_i^2\right] = \sum_{j=1}^{n}\left[\frac{1}{s}\frac{x_j^2\sum_{i=1}^{d}(\mathbf{U}^\mathsf{T})_{ij}^2}{\pi_j} - \frac{1}{s}\sum_{i=1}^{d}(\mathbf{U}^\mathsf{T}\mathbf{x})_i^2\right]\\ &= \sum_{j=1}^{n}\left[\frac{1}{s}\frac{x_j^2\|(\mathbf{U}^\mathsf{T})_{*j}\|_2^2}{\pi_j} - \frac{1}{s}\|\mathbf{U}^\mathsf{T}\mathbf{x}\|_2^2\right] \le \sum_{j=1}^{n}\frac{\|\mathbf{U}_{j*}\|_2^2\cdot x_j^2}{s\,\pi_j},\end{aligned}$$

where the last inequality follows from the fact that $\frac{1}{s}\|\mathbf{U}^\mathsf{T}\mathbf{x}\|_2^2 \ge 0$. This concludes the proof. □



(a) cardio (b) churn (c) default

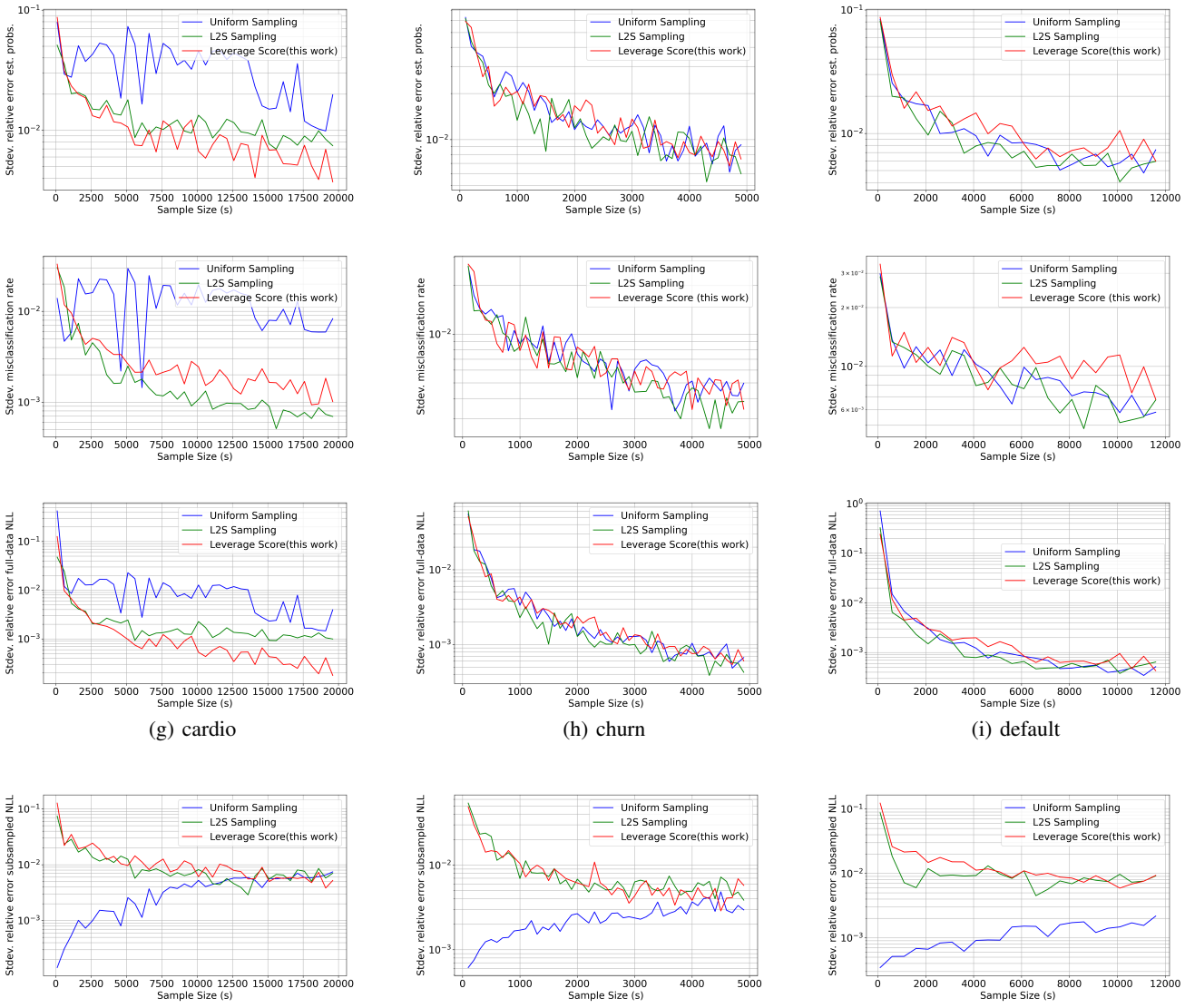
Figure 2: Relative error full-data negative log-likelihood (top row) and subsampled negative log-likelihood (bottom row) for all three datasets. Errors are in log-scale.

B Additional Experiments

We also compared our method in terms of (i) $|\ell(\hat{\boldsymbol{\beta}})-\ell(\boldsymbol{\beta}^*)|/-\ell(\boldsymbol{\beta}^*)$, the relative error of the full-data negative log likelihood (Figure 2 top row), which is a more common metric in other recent works and (ii) $|\bar{\ell}(\hat{\boldsymbol{\beta}})-\ell(\boldsymbol{\beta}^*)|/-\ell(\boldsymbol{\beta}^*)$, the relative error of the subsampled negative log-likelihood with respect to the full data negative log likelihood (see Figure 2 bottom row). For (ii), note that the first term on the numerator is $\bar{\ell}(\hat{\boldsymbol{\beta}})$ *i.e.*, eqn. (9) evaluated at the output of Algorithm 2 when the sampling matrix $\mathbf{S}$ is constructed using uniform sampling, leverage score sampling, and L2S method of (Munteanu et al. 2018).

In the top row of Figure 2, we present the relative error of the full data negative log-likelihoods in Figure 2. The trends closely resemble those shown in the first row of Figure 1 (in terms of relative error of estimated probabilities). While the error due to leverage score sampling for the *cardiovascular disease* dataset outperforms the other two sampling strategies, in the other two datasets (namely, *Bank customer churn prediction* and *Default of credit card clients*), the performance of all three sampling schemes is quite close (with L2S slightly better), and the errors decrease with the increase in s. In the bottom row of Figure 2, we evaluate the relative error of the subsampled negative log-likelihood. Across all three datasets, both leverage score sampling and L2S exhibit significantly lower errors compared to uniform sampling. In the last column, our approach gets much better performance as s increases. Similarly, for the other two datasets, our method outperforms L2S as s gets larger.

In Figure 3, we plot the standard deviations from the 20 runs for each of the experiments conducted.



(g) cardio (h) churn (i) default

Figure 3: Standard deviations for all the metrics in Figures 1 and 2. All the errors are in log-scale.