EstImAgg: A Learning Framework for Groupwise Aggregated Data: Supplement

Avradeep Bhowmik [*] Minmin Chen [†] Zhengming Xing [‡] Suju Rajan [§]

1 Generalised Linear Models and Bregman Divergences

A generalized linear model or GLM [10, 9] is a generalization of linear regression that subsumes various models like Poisson regression, logistic regression, etc. as special cases[1]. It assumes that the response variables, y are generated from a distribution in the exponential family centred around a mean parameter that is related to a linear function of the predictor $\mathbf{x}$ via a monotonic link function often denoted as $(\nabla\phi)^{-1}(\cdot)$. Here, $\boldsymbol{\phi}$ is a convex function that depends on the specific exponential family distribution used [2]. Specifically, given a predictor $\mathbf{x}$, a parameter $\boldsymbol{\beta}$ and a probability distribution $P_{\boldsymbol{\phi}}$ from the exponential family, the target y is obtained according to the distribution $P_{\boldsymbol{\phi}}$ such that

$$y|\mathbf{x} \sim P_{\boldsymbol{\phi}}(\eta_{\mathbf{x}}), \quad \text{where} \quad \eta_{\mathbf{x}} = E_{P_{\boldsymbol{\phi}}}(y|\mathbf{x}) = (\nabla\phi)^{-1}(\mathbf{x}^\top\boldsymbol{\beta})$$

Appropriate GLMs can be used to model a wide variety of data types– Gaussian for real-valued, Poisson regression for integer valued, logistic for binary, log-Normal for non-negative reals, etc. They have been successfully deployed in a wide variety of fields including machine learning[3, 1], biological surveys [11], image segmentation and reconstruction [12], analysis of medical trials [4], studying species-environment relationships in ecological sciences [7], virology [5] and estimating mortality from infectious diseases [6], among many others, and are widely prized for the interpretability of their results and the extendability of their methods in a plethora of domain specific variations [13]. They are easy to use and implement and many off-the-shelf software packages are available for most major programming platforms.

The matching loss functions associated with learning GLM parameters are distance-like functions called Bregman divergences, which are generalisations of square loss. Let $\boldsymbol{\phi} : \Theta \mapsto \mathbb{R}$ be a strictly convex, closed function on a convex domain $\Theta \subseteq \mathbb{R}^n$, that is differentiable on $\text{int}(\Theta)$. Then, for any $\mathbf{a}, \mathbf{b} \in \Theta$, the Bregman divergence $D_{\boldsymbol{\phi}}(\cdot\|\cdot)$ between $\mathbf{a}$ and $\mathbf{b}$ corresponding to the function $\boldsymbol{\phi}$ is defined as

$$D_{\boldsymbol{\phi}}(\mathbf{a}\|\mathbf{b}) \triangleq \boldsymbol{\phi}(\mathbf{a}) - \boldsymbol{\phi}(\mathbf{b}) - \langle \nabla\boldsymbol{\phi}(\mathbf{b}), \mathbf{a} - \mathbf{b}\rangle$$

where $\nabla\phi$ is the gradient of the function $\boldsymbol{\phi}$, applied elementwise. Bregman divergences are convex in their first argument. Although strictly speaking they are not a distance metric, they satisfy many properties of metrics, for example $D_{\boldsymbol{\phi}}(\mathbf{a}\|\mathbf{b}) \geq 0$ for any $\mathbf{a}, \mathbf{b}$, and $D_{\boldsymbol{\phi}}(\mathbf{a}\|\mathbf{b}) = 0$ if and only if $\mathbf{a} = \mathbf{b}$. Many standard distance-like functions like Square loss and KL-divergence are members of this family (see Table 1 in the Appendix).

Bregman Divergences have a very close relationship with generalized linear models. In particular, there is a one-to-one correspondence between each GLM and each Bregman divergence via the convex function $\boldsymbol{\phi}(\cdot)$ that is also closely related to the specific exponential family distribution associated with the GLM.

Specifically, for our work we use the fact that MLE parameter estimation in a GLM with given object features $\mathbf{X}$ and target variable $\mathbf{y}$ is equivalent to minimising $D_{\boldsymbol{\phi}}\left(\mathbf{y}\|(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta})\right)$ over $\boldsymbol{\beta}$, that is, the optimal parameter is the minimiser $\widehat{\boldsymbol{\beta}}$ for the following optimisation problem,

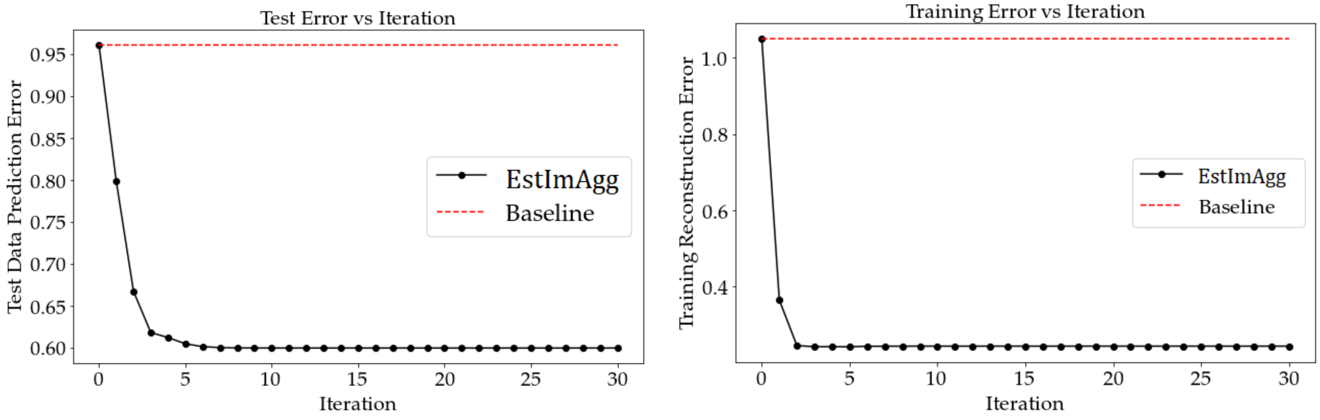
$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{(\mathbf{x},y)} D_{\boldsymbol{\phi}}\left(y\|(\nabla\phi)^{-1}(\mathbf{x}^\top\boldsymbol{\beta})\right) + \lambda\mathcal{R}(\boldsymbol{\beta})$$

where $\boldsymbol{\phi}(\cdot)$ is the convex function associated with the particular GLM used. For example, maximum likelihood for a Gaussian model or standard linear regression corresponds to square loss, for Poisson the corresponding divergence is generalized I-divergence (GI-divergence) and for Binomial, the corresponding divergence is the Kullback-Leibler or KL divergence. We refer the reader to [2] for a detailed exposition on the relationship between Bregman Divergences and GLM's.

In particular, we note that the only aspect of our framework that is affected by generalising linear

[*] The University of Texas at Austin, Austin, TX
[†] Google, Mountain View, CA
[‡] Criteo, Palo Alto, CA
[§] Criteo, Palo Alto, CA

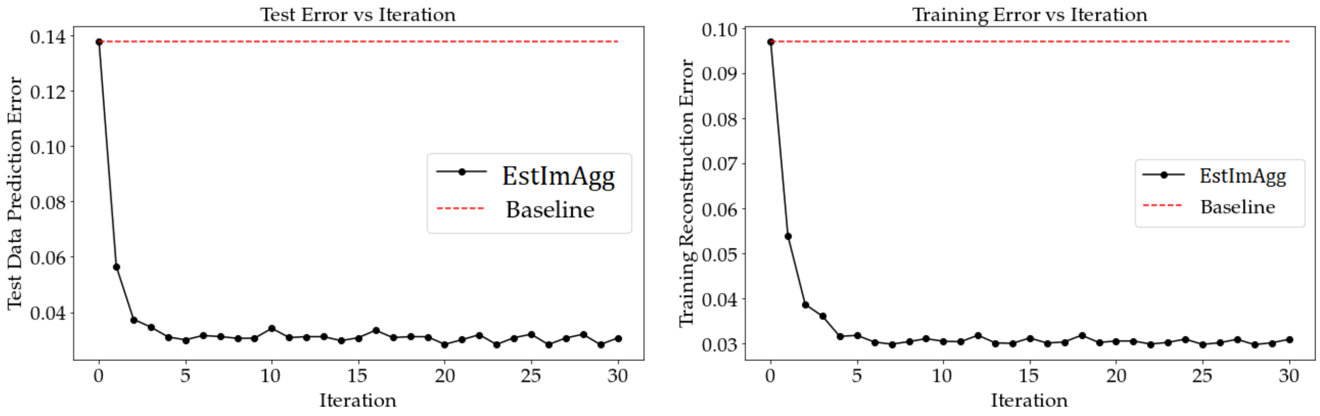
[1] see [9] or [8] for a detailed discussion on GLMs



(a) Test Prediction Error

(b) Train Reconstruction Error

Figure 1: Real Data: Estimation of Medicare Reimbursement Using CMS Data: Error on predictions for test data and error in reconstructed training data plotted vs iteration, as estimated using a Gaussian Model. Our model outperforms the baseline and converges within very few iterations, with a reasonably faithful reconstruction of the training data



(a) Test Prediction Error

(b) Train Reconstruction Error

Figure 2: Real Data: Estimation of Texas State Hospital Charges: Error on predictions for test data and error in reconstructed training data plotted vs iteration, as estimated using a Poisson Regression Model. Our model outperforms the baseline and converges within very few iterations, with a reasonably faithful reconstruction of the training data

regression to GLMs is in the loss function, which now changes from a square loss to a general Bregman Divergence.

2 General Commments

While the motivation for our work was predictive modelling with aggregated CPC in online advertising, we re-emphasise that our methods and framework are very general, and can be applied to any domain where data is available as aggregated targets. We defer to future work the other variations of this problem with alternative aggregation paradigms– where the targets are available at individual level but (some of) the features are aggregated, or when both targets and features are only available in aggregated form.

It is instructive to examine the mathematical form predicated by the imputation step update for the targets for the simplified version of our algorithm. Specifically, recall that given a particular value of parameter $\boldsymbol{\beta}^+$, the optimal imputed targets are obtained as the following update step

$$\forall \;\; i \in \mathcal{I}_k \;\; : \;\; y_i^+ = \mathbf{x}_i^\top \boldsymbol{\beta}^+ - \gamma_k$$

For non-overlapping aggregation, this is equivalent to a piecewise shift (that depends on the aggregation group)

$\phi(\mathbf{a})$	$D_{\phi}(\mathbf{a}\Vert\mathbf{b})$
$\frac{1}{2}\Vert\mathbf{a}\Vert^2$	$\frac{1}{2}\Vert\mathbf{a}-\mathbf{b}\Vert^2$
$\sum_i (a^{(i)} \log a^{(i)})$ $\mathbf{a} \in$ Prob. Simplex	$\mathrm{KL}(\mathbf{a}\Vert\mathbf{b}) =$ $\sum_i \left(a^{(i)} \log(\frac{a^{(i)}}{b^{(i)}})\right)$
$\sum_i \left(a^{(i)} \log a^{(i)} - a^{(i)}\right)$ $\mathbf{a} \in \mathbb{R}^n_+$	$\mathrm{GI}(\mathbf{a}\Vert\mathbf{b}) =$ $\sum_i a^{(i)} \log(\frac{a^{(i)}}{b^{(i)}}) - a^{(i)} + b^{(i)}$

Table 1: Examples of Bregman Divergences

applied to the previously estimated values of each target, before using these new estimates for learning a new parameter. Intuitively, this essentially imposes a “shape” on the imputed targets that depends on the learned parameter and the learning model, but the “value” is determined by constraints as set by their known group-wise aggregated values. This suggests a form of implicit “structural” regularisation– the model prefers imputed targets that adhere better to the linear relationship with the feature vectors.

This suggests a form of implicit “structural” regularisation– the model prefers imputed targets that adhere better to the linear relationship with the feature vectors. It remains to be seen whether this can be used to design variations of this algorithm where the modelling framework or domain expertise suggests a different structural formalism– we defer further investigations in this direction to future work. Other potential directions worth pursuing would involve a rigorous statistical analysis of our algorithms including asymptotic consistency properties and sample complexity for finite sample error analyses.

3 Extra Plots

Figures 1 and 2 show the results for test data estimation error and training data reconstruction error for the DESynPUF and TxID datasets respectively. For both these datasets, the plots show that our algorithm significantly outperforms the baseline with respect to either metric. Furthermore, our framework reaches a reasonably steady-state solution fairly rapidly within a few iterations of the algorithm. Results for other similar values of ρ were similar and are omitted for space constraints.

References

[1] S. Acharyya, O. Koyejo, and J. Ghosh. Learning to rank with bregman divergences and monotone retargeting. *arXiv preprint arXiv:1210.4851*, 2012.

[2] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *The Journal of Machine Learning Research*, 6:1705–1749, 2005.

[3] A. Bhowmik, J. Ghosh, and O. Koyejo. Generalized Linear Models for Aggregated Data. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 93–101, 2015.

[4] S. Dias, A. J. Sutton, A. Ades, and N. J. Welton. Evidence synthesis for decision making 2 a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making*, 33(5):607–617, 2013.

[5] J. J. Gart. The analysis of poisson regression with an application in virology. *Biometrika*, 51(3/4):pp. 517–521, 1964.

[6] P. Hardelid, R. Pebody, and N. Andrews. Mortality caused by influenza and respiratory syncytial virus by age group in England and Wales 1999–2010. *Influenza and other respiratory viruses*, 7(1):35–45, 2013.

[7] T. Jamil, W. A. Ozinga, M. Kleyer, and C. J. ter Braak. Selecting traits that explain species–environment relationships: a generalized linear mixed model approach. *Journal of Vegetation Science*, 24(6):988–1000, 2013.

[8] P. McCullagh and J. A. Nelder. Generalized linear models. 1989.

[9] J. A. Nelder and R. Baker. *Generalized linear models.* Wiley Online Library, 1972.

[10] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):pp. 370–384, 1972.

[11] A. Nicholls. How to make biological surveys go further with generalised linear models. *Biological Conservation*, 50(1):51–75, 1989.

[12] G. Paul, J. Cardinale, and I. F. Sbalzarini. Coupling image restoration and segmentation: a generalized linear model/bregman perspective. *International Journal of Computer Vision*, 104(1):69–93, 2013.

[13] L. Song, P. Langfelder, and S. Horvath. Random generalized linear model: a highly accurate and interpretable ensemble predictor. *BMC bioinformatics*, 14(1):5, 2013.