DIFFERENTIALLY PRIVATE GEODESIC REGRESSION

Aditya Kulkarni, Carlos Soto

University of Massachusetts, Amherst

One of the most foundational tools in statistical analyses is linear regression. In its simplest form, linear regression learns a linear relationship between an independent variable, the predictor, and a dependent variable, the response. Typically, these variables are both assumed to lie in a flat, Euclidean space. However, in modern statistical practices, it is common to encounter data that inherently live in curved, non-Euclidean spaces, such as spherical data (e.g., directional wind data), symmetric positive definite matrices (e.g., covariance matrices, brain tensors). For this reason, there have been many different extensions of regression to non-linear spaces, such as Geodesic Regression [Fletcher, 2011] that captures the relationship between Euclidean predictors and response variables that lie on a Riemannian manifold, and many more. Regardless of where the data resides, the learned relationship between variables relies on data captured from individuals; these individuals may be concerned about safeguarding their sensitive data. To protect one's data, differential privacy [DP, Dwork et al., 2006] has emerged as a leading standard for data sanitisation. Surprisingly, differential privacy for a methodology as fundamental as linear regression is not so straightforward.

In this paper, we consider the problem of privately estimating the parameters of geodesic regression, regression with a Euclidean predictor and a response variable on a Riemannian manifold. One can estimate the regression parameters by minimising the least squared energy, $E(p, v) = \frac{1}{2n}\sum_{i=1}^{n} d(\mathrm{Exp}(p, x_i\, v), y_i)^2$. Our method sanitises the regression parameters sequentially with the Riemannian manifold extension of the KNG mechanism. One can sample private footpoint $\tilde{p}$ and shooting vector $\tilde{v}$ using the KNG mechanism by sampling from the density: $f(p; D) \propto \exp\left\{-\frac{\|\nabla_p E(p;D)\|_p}{\Delta_p/\epsilon_p}\right\}$, and $f(v; D) \propto \exp\left\{-\frac{\|\nabla_v E(v;D)\|_v}{\Delta_v/\epsilon_v}\right\}$, where Δ_p, Δ_v are the footpoint and shooting vector sensitivities, respectively, and ϵ_p, ϵ_v are the privacy budgets. To bound the sensitivities, we assume that the sectional curvature of the manifold is bounded everywhere by (κ_l, κ_h). We also assume that the data is bounded as $D \subseteq B_r(m_0)$ where $r \leq \frac{\pi}{8\sqrt{\kappa_h}}$ for Riemannian manifolds with positive curvature and $r < \tau_m$ for Riemannian manifolds with negative curvature. Further, the least-squares geodesic is $\tau-$close to the data, i.e, for $\tau > 0$, we have $\sup_D d(y_i, \mathrm{Exp}(\hat{p}, x_i\hat{v})) \leq \tau, \forall i$. We find the following bounds on the sensitivities.

Table 1. Bounds on Δ_p and $\Delta_{\vec{v}}$.

Sensitivity	$\kappa_l \geq 0$	$\kappa_l < 0$
Δ_p	$\frac{2\tau}{n}$	$\frac{2\tau}{n}\cosh\big(2\sqrt{-\kappa_l}(\tau_m + \tau)\big)$
$\Delta_{\vec{v}}$	$\frac{2\tau}{n}$	$\frac{\tau}{n}\frac{\sinh(2\sqrt{-\kappa_l}(\tau_m + \tau))}{\sqrt{-\kappa_l}(\tau_m + \tau)}$

We aim to measure how the private estimates affect the energy $E(p, v; D)$. To do this, we sample 100 pairs of $(\tilde{p}, \tilde{v})$ for given privacy budgets (ϵ_p, ϵ_v). As one of the examples, we analyze corpus callosum shapes from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset processed by Cornea et al. [2017]. By using 50 uniformly sampled boundary landmarks, each image lies on the Kendal shape space $\mathbb{C}P^{k-2}$. The sectional curvature of Kendall shape space with landmarks ≥ 4 is bounded between $[1, 4]$, resulting in $\kappa_l = 1$. We get the sensitivities of the footpoint and shooting vector as, $\Delta_p = \frac{2\tau}{n}, \quad \Delta_v = \frac{2\tau}{n}$. For Kendall's shape space, we examine the log average MSE under both unequal and equal budget allocations, as shown in Figure 1. Figure 2 provides a visual comparison of predictions obtained from non-private and differentially private regression parameters under a budget split of $\epsilon_p = \epsilon_v = 0.3$. For this choice of privacy allocation, the predicted shapes from the private parameters align closely with those from the non-private regression, with only minor deviations visible across ages.

We show that the sensitivity of each parameter is tied to their respective the curvature of the manifold. We demonstrate our methodology on Kendall's planar shape space; however, it is general to any Riemannian manifold. Using the

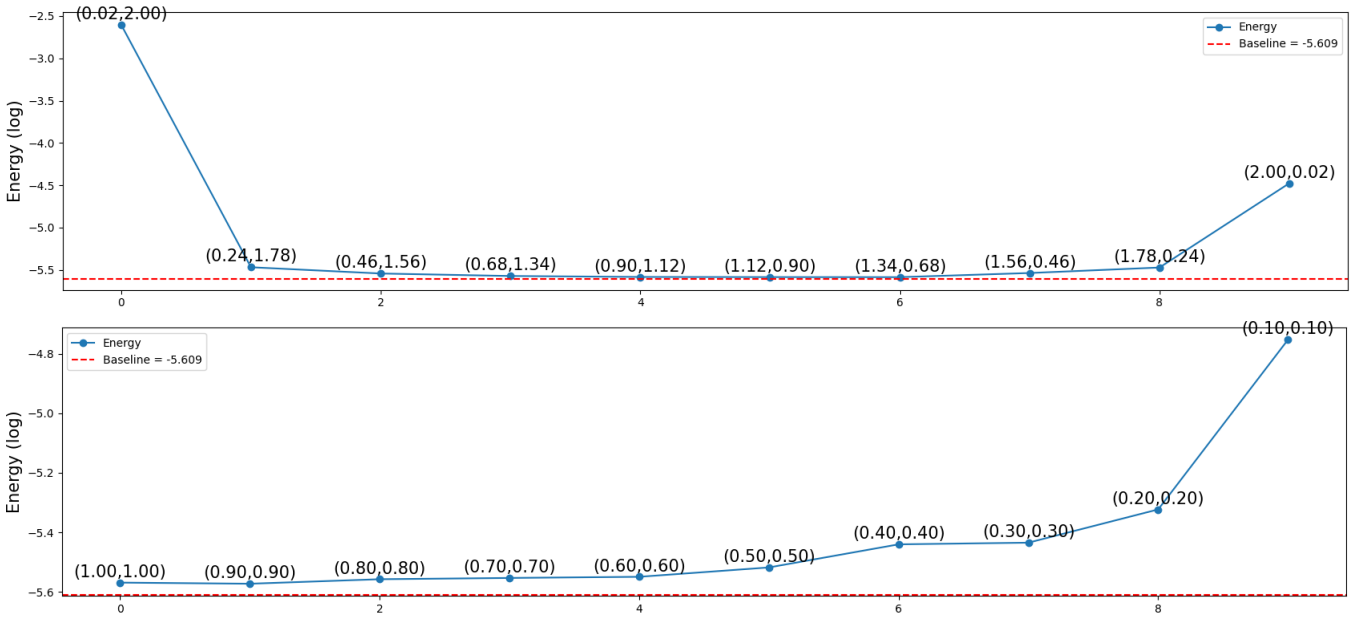

Figure 1. Log average MSE, $\ln \bar{\mathcal{E}}$, for 100 data points on $\mathbb{C}P^{k-2}$. Dotted line is the log energy without privatisation. Top: unequal budget splits $\epsilon_p \in [0.02, 2.0]$, $\epsilon_v \in [2.0, 0.02]$, total $\epsilon = 2.02$ and Bottom: Equal budget splits with varying total budget $\epsilon \in [0.2, 2.0]$.

sanitised $(\tilde{p}, \tilde{v})$ we can get a DP geodesic which can be used for predictions.

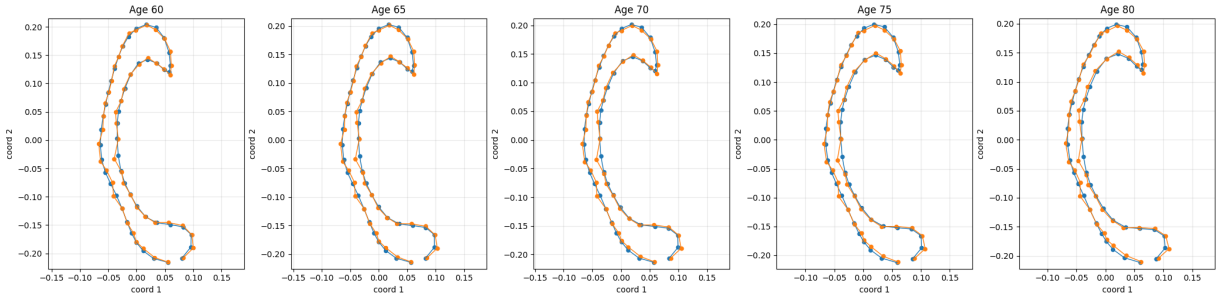

Figure 2. Corpus callosum shapes at ages 60, 65, 70, 75, and 80. Blue curves denote predictions from the non-private regression parameters, while orange curves correspond to predictions from the differentially private parameters with $\epsilon_p = \epsilon_v = 0.3$.

References

E. Cornea, H. Zhu, P. Kim, J. G. Ibrahim, and A. D. N. Initiative. Regression models on riemannian symmetric spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):463–482, 2017.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

T. Fletcher. Geodesic regression on riemannian manifolds. In *Proceedings of the Third International Workshop on Mathematical Foundations of Computational Anatomy-Geometrical and Statistical Methods for Modelling Biological Shape Variability*, pages 75–86, 2011.