On Riemannian Gradient Descent Algorithm using gradient averaging

Saugata Purkayastha[*], SAPU00001@STUD.UNI-SAARLAND.DE
Department of Language Science and Technology,
Universität des Saarlandes,
ZUSE School ELIZA
Saarbrücken, Germany-66123

Sukannya Purkayastha[†] SUKANNYA.PURKAYASTHA@NECLAB.EU
TU Darmstadt
Darmstadt, Germany-64289

Abstract

In this work, we introduce the notion of `RGrad-Avg`, a variant of Riemannian Gradient Descent (RGD) algorithm based on gradient averaging, `Grad-Avg` [10]. In the present work, we extend the notion of `Grad-Avg` to Riemannian submanifolds. We further establish that under reasonable assumptions, the value of the objective function decreases with each iteration of `RGrad-Avg` and validate the results on some benchmark datasets. Additionally, our findings suggest that `RGrad-Avg` is comparable to classical Riemannian Gradient Descent for the chosen datasets.

1. Introduction

Let $\mathcal{M}$ be a Riemannian submanifold of $\mathbb{R}^n$ and $f : \mathcal{M} \to \mathbb{R}$ be a smooth function. For any point $p \in \mathcal{M}$, let $T_p(\mathcal{M})$ denote the tangent plane of $\mathcal{M}$ at p and let $\mathrm{grad} f(p)$ stand for the Riemannian gradient of f at p obtained by orthogonally projecting the classical gradient $\nabla f(p)$ to the tangent spaces $T_p(\mathcal{M})$. The iterates of the Riemannian Gradient Descent (RGD) algorithm are given by the following scheme [4]:

$$x_{k+1} = \mathrm{Re}_{x_k}(-\alpha \mathrm{grad} f(x_k)) \tag{1}$$

where $\mathrm{Re}_{x_k} : T_p(\mathcal{M}) \to \mathcal{M}$ is the retraction map at the point x_k and α is the learning rate. On the other hand for a smooth function $f : \mathbb{R}^n \to \mathbb{R}$, the iterates of `Grad-Avg` (based on Heun's method) is given by [10]:

$$x_{k+1} = x_k - \alpha \frac{1}{2}(\nabla f(x_k) + \nabla f(x_k - \alpha \nabla f(x_k))) \tag{2}$$

A natural question to ask at this point is the following: Is it possible to extend the notion of `Grad-Avg` to Riemannian submanifolds? In other words, can we have an equation analogous to (1) based on (2)? We attempt to answer this question in the following subsection.

[*] Saugata Purkayastha is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse School of Excellence in Artificial Intelligence sponsored by the Federal Ministry of Education and Research.
[†] Currently with NEC Laboratories Europe, Heidelberg, Germany

Algorithm 1 Corrected Gradient Average on Riemannian Submanifolds (`RGrad-Avg`)

Input: Initial point $x_0 \in \mathcal{M}$, step size $\alpha > 0$, maximum iterations K, tolerance $\varepsilon > 0$, objective function f

Output: Approximate solution x_K

Set $x \leftarrow x_0$

for $k = 1$ *to* K **do**

  Compute Riemannian gradient: $g_k \leftarrow \operatorname{grad} f(x)$

  Compute predicted point: $\overline{x_{k+1}} \leftarrow \mathrm{Re}_x(-\alpha g_k)$

  Compute gradient at predicted point: $g_{k+1}^{\text{pred}} \leftarrow \operatorname{grad} f(\overline{x_{k+1}})$

  Compute vector $v \leftarrow -\alpha g_k$

  Compute parallel transport: $\tilde{g}_{k+1} \leftarrow P_v^{-1}(g_{k+1}^{\text{pred}})$

  Compute update direction: $u \leftarrow -\frac{1}{2}\alpha(g_k + \tilde{g}_{k+1})$

  Retract: $x_{new} \leftarrow \mathrm{Re}_{x_k}(u)$

  if $d(x_{new}, x_k) < \varepsilon$ **then**

    break

  end

  Set $x \leftarrow x_{new}$

end

return x

1.1. The `RGrad-Avg` algorithm

In order to extend the notion of `Grad-Avg` to Riemannian submanifolds, the main hurdle is the sum appearing on the right-hand side of (2). Clearly, for a Riemannian submanifold, the gradients of f at x_k and x_{k+1} reside in two different tangent spaces and thus their sum bears no meaning. However, this issue may be settled using the notion of parallel transport. To be precise, let $\gamma_v : [0,1] \to \mathcal{M}$ such that $\gamma_v(0) = x$, $\gamma_v'(0) = v$ and $\gamma_v(1) = \mathrm{Re}_x(v)$. Further, let P_v denote the parallel transport along $\gamma_v(t)$ from $t = 0$ to $t = 1$ so that P_v^{-1} is from $t = 1$ to $t = 0$ along $\gamma_v(t)$. Keeping this in mind, for a smooth function $f : \mathcal{M} \to \mathbb{R}$, we formally define the iterates of `RGrad-Avg` in the following way:

$$x_{k+1} = \mathrm{Re}_{x_k}(-\frac{\alpha}{2}(\mathrm{grad} f(x_k) + P_v^{-1}(\mathrm{grad} f(\overline{x_{k+1}})))) \tag{3}$$

with

$$\overline{x_{k+1}} = \mathrm{Re}_{x_k}(-\alpha \mathrm{grad} f(x_k)) \tag{4}$$

where $v = -\alpha \mathrm{grad} f(x_k)$. We note that our formulation is consistent with the classical case. Indeed, for $\mathcal{M} = \mathbb{R}^n$ and for a curve $\gamma : [0,1] \to \mathbb{R}^n$ such that $\gamma(0) = x_k$, we claim that the isomorphism $T_{\gamma(0)}(\mathbb{R}^n) \to T_{\gamma(t)}(\mathbb{R}^n)$ for $t \in [0,1]$ induces the identity map as the parallel transport of a vector field $Y \in T_{\gamma(0)}(\mathbb{R}^n)$ along $\gamma(t)$. Since the covariant derivative of $Y \in T_{\gamma(0)}(\mathbb{R}^n)$ in the direction of $X \in T_{\gamma(t)}(\mathbb{R}^n)$ in $\mathbb{R}^n$ is just the directional derivative [5], we get:

$$\begin{aligned} D_X Y(\gamma(0)) &= \lim_{t\to 0} \frac{Y(\gamma(0) + tX) - Y(\gamma(0))}{t} \\ &= \lim_{t\to 0} \frac{Y(\gamma(t) - Y(\gamma(0))}{t} \\ &= 0 \end{aligned}$$

where the second last inequality follows as $\gamma(t) = \gamma(0) + tX$ by virtue of the isomorphism. Thus, the parallel transport of Y keeps Y fixed. The claim now follows by identifying $X = -\alpha \text{grad} f(x_k)$, $Y = \text{grad} f(\overline{x_{k+1}})$, $\gamma(t) = x_k + t(\overline{x_{k+1}} - x_k)$ and finally noting that $\text{Re}_x(s) = x + s$ for $s \in T_x(\mathbb{R}^n)$.

Note that the study of variants of classical RGD is not a new concept in the context of optimization algorithms on Riemannian manifolds. For example, Bonnabel [3] introduces the notion of stochastic gradient descent algorithms, while Zhang et al. [14] studies the manifold proximal gradient algorithm in a non-smooth setting. Becigneul and Ganea [2] proposes an adaptive step-size algorithm (RADAM) for product Riemannian manifolds. For an account of how various optimization algorithms on Riemannian manifolds emerge from their respective counterparts in classical Euclidean space, we refer the readers to Fei et al. [7].

2. Assumption

In this section, we mention a key assumption related to the present work. We begin with the following definition [4]: For any $(x, v) \in T(\mathcal{M})$, where $T(\mathcal{M})$ denotes the tangent bundle of $\mathcal{M}$, there exists a unique path γ_v such that $\gamma_v(0) = x$ and $\gamma_v'(0) = v$. We call the map $\text{Exp}_x : T_x(\mathcal{M}) \to M$ to be the exponential retraction map if $\text{Exp}_x(v) = \gamma_v(1)$ with $(x, v) \in \mathcal{O}$ where

$$\mathcal{O} = \{(x, v) \in T(\mathcal{M}) | \gamma_v \text{ is defined on an interval containing (0,1)}\}$$

Throughout the paper, the retraction map will stand for the exponential retraction map i.e. $\text{Re}_x(v) = \text{Exp}_x(v)$ for all $(x, v) \in \mathcal{O}$. We now state the aforesaid assumption [4]:

1. the gradient $\text{grad} f$ of the function $f : \mathcal{M} \to \mathbb{R}$ is Lipschitz continuous i.e. there exists some $L > 0$ (known as the Lipschitz constant) such that

$$\left\| P_v^{-1} \text{grad} f(\text{Exp}_x(v)) - \text{grad} f(x) \right\| < L \left\| v \right\|$$

   for all (x, v) in the domain of the exponential map $\text{Exp}_x(v)$.

Further, assumption (1) leads to the following inequality [4]:

$$f(\text{Exp}_x(v)) \leq f(x) + \langle v, \text{grad} f(x) \rangle + \frac{L}{2} \left\| v \right\|^2 \tag{5}$$

3. Main Result

In this section, we state the main result of the work. Specifically, we show that under suitable assumptions discussed in Theorem 1, the values of the objective function decrease with each iteration of the `RGrad-Avg` algorithm.

3.1. The objective convergence of `RGrad-Avg`

Theorem 1 *Let $\mathcal{M}$ be a Riemannian submanifold and $f : \mathcal{M} \to \mathbb{R}$ be a smooth map such that assumption (1) and inequality (5) hold with constant L. For each $k = 1, 2, \ldots, n$, let (x_k, u) lie in the domain of the exponential retraction map $\text{Exp}_{x_k}(u)$ where u is either $-\alpha \text{grad} f(x_k)$ or $-\frac{\alpha}{2}(\text{grad} f(x_k) + P_v^{-1}(\text{grad} f(\overline{x_{k+1}})))$. Then for $\alpha \leq \frac{2}{25L}$, the following holds:*

$$f(x_{k+1}) - f(x_k) \leq -c \left\| \text{grad} f(x_k) \right\|^2 \tag{6}$$

where c is a positive constant depending upon α.

Proof Putting $u = -\frac{\alpha}{2}(\text{grad}f(x_k) + P_v^{-1}(\text{grad}f(\overline{x_{k+1}})))$ in (5), for $v = -\alpha\text{grad}f(x_k)$ we get-

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \left\langle \text{grad}f(x_k), -\frac{\alpha}{2}(\text{grad}f(x_k) + P_v^{-1}\text{grad}f(\overline{x_{k+1}}) \right\rangle + \frac{L}{2}\left\| \frac{\alpha}{2}(\text{grad}f(x_k) + P_v^{-1}\text{grad}f(\overline{x_{k+1}}) \right\|^2 \\ &= f(x_k) - \frac{\alpha}{2}\left\|\text{grad}f(x_k)\right\|^2 - \frac{\alpha}{2}\langle \text{grad}f(x_k), P_v^{-1}\text{grad}f(\overline{x_{k+1}})\rangle + \frac{L\alpha^2}{8}\left\|\text{grad}f(x_k) + P_v^{-1}\text{grad}f(\overline{x_{k+1}})\right\|^2 \end{aligned} \tag{7}$$

Note that

$$\langle \text{grad}f(x_k), P_v^{-1}\text{grad}f(\overline{x_{k+1}})\rangle = \left\|\text{grad}f(x_k)\right\|^2 + \langle \text{grad}f(x_k), P_v^{-1}\text{grad}f(\overline{x_{k+1}}) - \text{grad}f(x_k)\rangle$$

Applying Cauchy-Schwarz inequality, we have-

$$\begin{aligned} \langle \text{grad}f(x_k), P_v^{-1}\text{grad}f(\overline{x_{k+1}}) - \text{grad}f(x_k)\rangle &\geq -\left\|\text{grad}f(x_k)\right\| \left\|P_v^{-1}\text{grad}f(\overline{x_{k+1}}) - \text{grad}f(x_k)\right\| \\ &\geq -L\alpha\left\|\text{grad}f(x_k)\right\|^2 \end{aligned}$$

Where the last inequality follows from assumption (1) for $u = v = -\alpha\text{grad}f(x_k)$. Using (4) and assumption (1), we get-

$$\begin{aligned} \langle \text{grad}f(x_k), P_v^{-1}\text{grad}f(\overline{x_{k+1}})\rangle &= \left\|\text{grad}f(x_k)\right\|^2 + \langle \text{grad}f(x_k), P_v^{-1}\text{grad}f(\overline{x_{k+1}}) - \text{grad}f(x_k)\rangle \\ &\geq (1 - L\alpha)\left\|\text{grad}f(x_k)\right\|^2 \end{aligned} \tag{8}$$

From the triangle inequality, we get

$$\left\|\text{grad}f(x_k) + P_v^{-1}\text{grad}f(\overline{x_{k+1}})\right\| \leq \left\|\text{grad}f(x_k)\right\| + \left\|P_v^{-1}\text{grad}f(\overline{x_{k+1}})\right\| \tag{9}$$

Once again using (4) and assumption (1),

$$\begin{aligned} \left\|P_v^{-1}\text{grad}f(\overline{x_{k+1}})\right\| &= \left\|P_v^{-1}\text{grad}f(\text{Exp}_{x_k}(-\alpha\text{grad}f(x_k)))\right\| \\ &= \left\|P_v^{-1}\text{grad}f(\text{Exp}_{x_k}(-\alpha\text{grad}f(x_k))) - \text{grad}f(x_k) + \text{grad}f(x_k)\right\| \\ &\leq \left\|P_v^{-1}\text{grad}f(\text{Exp}_{x_k}(-\alpha\text{grad}f(x_k))) - \text{grad}f(x_k)\right\| + \left\|\text{grad}f(x_k)\right\| \\ &\leq L\left\|-\alpha\text{grad}f(x_k)\right\| + \left\|\text{grad}f(x_k)\right\| \\ &= (L\alpha + 1)\left\|\text{grad}f(x_k)\right\| \end{aligned} \tag{10}$$

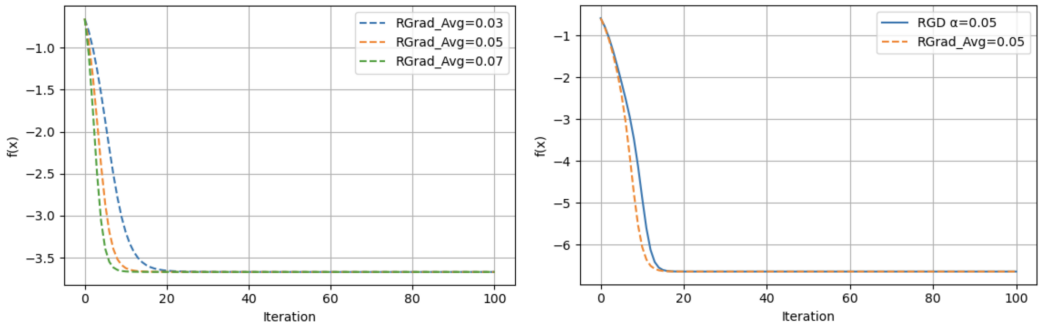
Using (10) in (9) we get

$$\left\|\text{grad}f(x_k) + P_v^{-1}\text{grad}f(\overline{x_{k+1}})\right\| \leq (2 + L\alpha)\left\|\text{grad}f(x_k)\right\| \tag{11}$$

Finally, using (8) and (11) in (7) and simplifying, we get

$$f(x_{k+1}) - f(x_k) \leq \left[-\frac{\alpha}{2}(2 - L\alpha) + \frac{L\alpha^2}{8}(2 + L\alpha)^2\right]\left\|\text{grad}f(x_k)\right\|^2$$

where the term within the parenthesis of the right hand side can be seen to be negative for $\alpha \leq \frac{2}{25L}$ and this completes the proof. ■



(*a*) Objective function plot (*b*) Comparison plot

Figure 1: (a) Comparison of objective function values for different values of α and (b) comparison of `RGrad-Avg` and RGD for breast cancer dataset

3.2. Choice of manifold and Objective Function

Motivated by [11] and [13], We consider the *unit sphere manifold* $\mathcal{M} = \mathbb{S}^{n-1} = \{x \in \mathbb{R}^n \mid \|x\|_2 = 1\}$, which is a (compact) Riemannian submanifold of $\mathbb{R}^n$ with Euclidean metric. The tangent space at $x \in \mathbb{S}^{n-1}$ is given by $T_x\mathbb{S}^{n-1} = \{\xi \in \mathbb{R}^n \mid x^\top \xi = 0\}$ and the parallel transport is given by [4]

$$P_v(u) = \left[I + (\cos\ \|v\| - 1)\frac{vv^\top}{\|v\|^2} - \sin\|v\|\frac{xv^\top}{\|v\|}\right]u$$

for (x, v) lying in the domain of the exponential map. Our objective is $f(x) = -\frac{1}{2}x^\top Cx$, where C is the empirical covariance matrix which is symmetric and positive semidefinite by construction. Maximizing $x^\top Cx$ yields the principal eigenvector of C (first principal component). This is the Rayleigh quotient optimization subject to $\|x\| = 1$, i.e. on $\mathbb{S}^{n-1}$. Note that our choice of $\mathcal{M}$ and f satisfies the conditions of Theorem 1. See Appendix A for further discussion.

3.3. Numerical results and discussions

We empirically validate our findings from Theorem 1 for the case where $L = 1$. We conduct experiments on the Breast Cancer Classification Dataset[1] [12]. We report the result after running 100 iterations over a grid of learning rates $\alpha \in \{0.03, 0.05, 0.07\}$. We plot the value of the objective function against the number of iterations in Figure $1(a)$, and compare the performance of `RGrad-Avg` and RGD in Figure $1(b)$ for $\alpha = 0.05$.

From the objective function plot (Figure $1(a)$) , it is clear that the objective function value decreases uniformly as the number of iterations increases for both datasets. This is in line with our findings in Theorem 1.

From Figure $1(b)$, we observe that the objective function value decreases with a slightly faster rate for `RGrad-Avg` as compared to that for the RGD algorithm for breast cancer dataset. In Appendix B, we compare the performance of `RGrad-Avg` and RGD for different values of alpha for a range of other datasets namely digits [1], iris [8] and and wine dataset [9]. The empirical findings once again show that the performance of `RGrad-Avg` is comparable with the RGD algorithm.

1. We provide the details of the datasets we use in our study in Appendix B

4. Conclusion and future work

In this work, we propose a new optimizer `RGrad-Avg` and study the conditions under which the objective function value decreases for each iteration of the proposed optimizer. The empirical results are found to be consistent with the findings of Theorem 1. Furthermore, our experiments reveal that the performance of the proposed optimizer is comparable to that of RGD within the specified range of learning rate. We acknowledge that while there are other robust optimizers such as RADAM [2], we do not consider them in the present work as our objective is to introduce a new optimizer and compare its performance with RGD, following [10]. As a future line of work, it would be interesting to examine whether Theorem 1 holds for a broader range of values of learning rates under additional assumptions. Moreover, it would be worthwhile to investigate the conditions under which Theorem 1 holds for general retraction maps.

References

[1] F. Alimoglu and E. Alpaydin. Optical recognition of handwritten digits data set, 1998. URL https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits. Retrieved from [6].

[2] Gary Becigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1eiqi09K7.

[3] Silvère Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013. doi: 10.1109/TAC.2013.2254619.

[4] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023. doi: 10.1017/9781009166164. URL https://www.nicolasboumal.net/book.

[5] Manfredo Perdigao Do Carmo and J Flaherty Francis. *Riemannian geometry*, volume 2. Springer, 1992.

[6] Dheeru Dua and Casey Graff. Uci machine learning repository. http://archive.ics.uci.edu/ml, 2019.

[7] Yanhong Fei, Yingjie Liu, Chentao Jia, Zhengyu Li, Xian Wei, and Mingsong Chen. A survey of geometric optimization for deep learning: from euclidean space to riemannian manifold. *ACM Computing Surveys*, 57(5):1–37, 2025.

[8] R. A. Fisher. Iris data set, 1936. URL https://archive.ics.uci.edu/ml/datasets/iris. Retrieved from [6].

[9] M. Forina, C. Armanino, S. Castellano, and M. Ubigli. Wine data set, 1991. URL https://archive.ics.uci.edu/ml/datasets/Wine. Retrieved from [6].

[10] Saugata Purkayastha and Sukannya Purkayastha. A variant of gradient descent algorithm based on gradient averaging. *ArXiv*, abs/2012.02387, 2020. URL https://api.semanticscholar.org/CorpusID:227305491.

[11] Steven Thomas Smith. Optimization techniques on riemannian manifolds. *arXiv preprint arXiv:1407.5965*, 2014.

[12] W. Nick Street, W. H. Wolberg, and O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In Raj S. Acharya and Dmitry B. Goldgof, editors, *Biomedical Image Processing and Biomedical Visualization*, volume 1905, pages 861 – 870. International Society for Optics and Photonics, SPIE, 1993. doi: 10.1117/12.148698. URL https://doi.org/10.1117/12.148698.

[13] Lei-Hong Zhang. Riemannian newton method for the multivariate eigenvalue problem. *SIAM Journal on Matrix Analysis and Applications*, 31(5):2972–2996, 2010. doi: 10.1137/100788975. URL https://doi.org/10.1137/100788975.

[14] Zhuan Zhang, Shuisheng Zhou, Dong Li, and Ting Yang. Riemannian proximal stochastic gradient descent for sparse 2dpca. *Digital Signal Processing*, 122:103320, 2022.

Appendix A.

As discussed in section 3.2, we have $f(x) = -\frac{1}{2}x^\top Cx$ and thus $\text{grad}f = -(I - xx^\top)Cx$ which is Lipschitz continuous. Thus in view of [4, corollary 10.48] Assumption 1 holds. However, for the sake of completeness, we provide the following proof as a particular instance of [4, corollary 10.48]:

Theorem 2 *Let $f : \mathcal{M} \to \mathbb{R}$ be such that $\text{grad}f$ is Lipschitz continuous with Lipschitz constant L. Then for all (x, v) in the domain of the exponential map Exp, the following holds:*

$$\left\| P^{-1}_{\text{Exp}_x(v)\to x}\text{grad}f(\text{Exp}_x(v)) - \text{grad}f(x) \right\| < L\,\|v\|$$

Proof Let

$$g(t) = P^{-1}_{\gamma(t)\to x}\text{grad}(f \circ \gamma_v)(t)$$

for $\gamma_v : [0, 1] \to \mathcal{M}$ such that $\gamma_v(0) = x$ and $\gamma_v'(0) = v$. Consequently, $g(t) \in T_x(\mathcal{M})$. Following the construction of [4, Proposition 10.46] we choose a basis $e_1, e_2, \ldots, e_d \in T_x(\mathcal{M})$ so that their parallel transports are given by $E_i(t) = P_{0\to t}(e_i)$. In that case, we have $\text{grad}(f \circ \gamma_v)(t) = \sum_{i=1}^d v_i(t)E_i(t)$, where v_i's are continuously differentiable functions. From the chain rule property of covariant derivatives [4, Theorem 8.67], it follows that

$$\sum_{i=1}^{d} v_i'(t)E_i(t) = \frac{D}{dt}(\text{grad}(f \circ \gamma_v)(t)) = \nabla_{\gamma_v'(t)}\text{grad}f$$

where ∇ denotes the connection on $\mathcal{M}$. Further,

$$\begin{aligned}
g(1) - g(0) &= P^{-1}_{\text{Exp}_x(v)\to x}\text{grad}f(\text{Exp}_x(v)) - \text{grad}f(x) \\
&= \sum_{i=1}^{d}(v_i(1) - v_i(0))e_i \\
&= \sum_{i=1}^{d}\int_0^1 (v_i'(t))dte_i \\
&= \int_0^1 P_{t\to 0}\left(\sum_{i=1}^{d}(v_i'(t))E_i(t)\right)dt \\
&= \int_0^1 P_{t\to 0}\nabla_{\gamma_v'(t)}\text{grad}f
\end{aligned}$$

Taking norm on both the sides of the last equation and using the isometry of parallel transports [4, Proposition 10.33], we get-

$$\|g(1) - g(0)\| = \left\| P^{-1}_{\text{Exp}_x(v)\to x}\text{grad}f(\text{Exp}_x(v)) - \text{grad}f(x) \right\| \leq L\,\|v\|$$

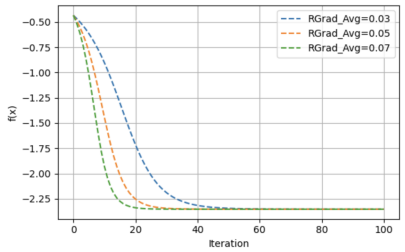
where the inequality follows from [4, Proposition 10.46]. The proof now follows. ■

Appendix B.

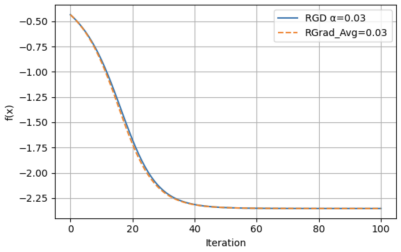
We provide the details of the datasets in Table 1. We show the performance of `RGrad-Avg` algorithm for $\alpha \in \{0.03, 0.05, 0.07\}$ and comparison of `RGrad-Avg` with RGD for the for Wine, iris and digits datasets in Figure 2 below:

Dataset	Samples	Features	Classes
Iris [8]	150	4	3
Wine [9]	178	13	3
Breast Cancer (Diagnostic) [12]	569	30	2
Digits (Optical Recognition) [1]	1797	64 (8×8 images)	10

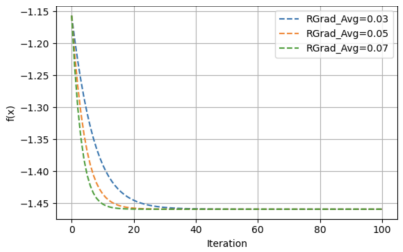
Table 1: Summary of benchmark datasets used in this work.



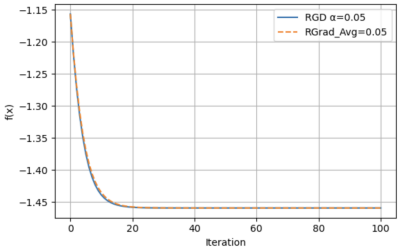
(*a*) Objective function plot (Wine)



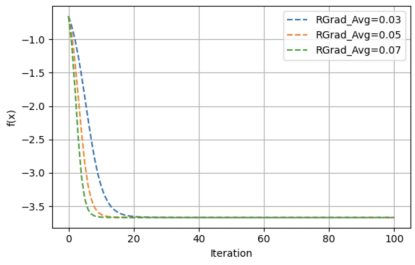
(*b*) Comparison plot (Wine)



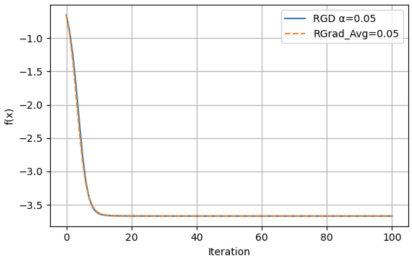
(*c*) Objective function plot (Iris)



(*d*) Comparison plot (Iris)



(*e*) Objective function plot (Digits)



(*f*) Comparison plot (Digits)

Figure 2: Comparison of objective function values for different values of α (left) and comparison of `RGrad-Avg` and RGD (right) for different datasets