Riemannian Stochastic Approximation of Tame Functions

Vyacheslav Kungurtsev
Department of Computer Science
FEL
Czech Technical University

July 25, 2024

Joint work with Johannes Aspmann at CTU and Reza Roohi Seraji
This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101084642.

Outline

Background

$$\min_{x \in \mathcal{M}}, \ F(x) := \mathbb{E}_{\xi}[f(x,\xi)]$$

where,

- $\mathcal{M}$ is a Riemannian manifold
- The objective $f(x)$, nor is $F(\cdot,\xi)$ for any ξ, everywhere continuously differentiable, i.e., it is nonsmooth
- More specifically it is *tame*, or its landscape is characterized by the topology of *o-minimal structures*

Background

The canonical algorithm to consider is Retraction-SGD

$$x_{k+1} = R_{x_k}(x_k - \alpha_k g_k),\ g_k \sim \partial F(x_k, \cdot)$$

where,

1. R_{x_k} is a *retraction* that can be considered a projection onto the manifold (and is when the manifold is embedded in Euclidean space)
2. g_k is sampled from some $\xi \in \Xi$ an element of a Clarke subdifferential of F, or an element of a conservative vector field, or an output of an autograd operation.

Conservative Vector Fields on a Manifold

Given $x \in \mathcal{M}$

- Tangent space at x is $T_x\mathcal{M}$
- Tangent bundle $T\mathcal{M}$
- Cotangent space is $T_x^*\mathcal{M}$
- Cotangent bundle $T^*\mathcal{M}$
- There exists an inner product $\langle w, v\rangle$ for $w \in T_x^*\mathcal{M}$, $v \in T_x\mathcal{M}$

Thus g_k is the Riesz representative of the dual of $\partial F(x,\xi) \subset T_x\mathcal{M}$

Conservative Vector Fields on a Manifold

- The metric on $\mathcal{M}$, $g(\cdot,\cdot)$, or $g_x(\cdot,\cdot)$ when evaluated at a point $x \in \mathcal{M}$, induces a norm $\|\cdot_{g_x}\| := \sqrt{g_x(\cdot,\cdot)}$
- The length of a piecewise smooth curve $\gamma : [a,b] \to \mathcal{M}$ is defined as

$$L(\gamma) = \int_a^b \|\dot{\gamma}(t)\|_{g_{\gamma(t)}} \, dt.$$

- For two points $x, y \in \mathcal{M}$, we denote the Riemannian distance from x to y by $d(x,y)$,

$$d(x,y) := \inf\{L(\gamma) \, : \, \gamma \in \mathcal{A}_\infty, \, \gamma(a) = x, \, \gamma(b) = y\},$$

- An *absolutely continuous curve* γ is such that: for all $\varepsilon > 0$ there exists a $\delta > 0$ such that for any $m \in \mathbb{N}$ and any selection of disjoint intervals $\{(a_i, b_i)\}_{i=1}^m$ with $[a_i, b_i] \subseteq I$, whose overall length is $\sum_{i=1}^m |b_i - a_i| < \delta$,

$$\sum_{i=1}^m d(\gamma(b_i), \gamma(a_i)) < \epsilon$$

Conservative Vector Fields on a Manifold

Lemma 2 (Lemma 1 of Bolte and Pauwels (2021)) *Let $D : \mathcal{M} \rightrightarrows T^*\mathcal{M}$ be a set-valued map with nonempty compact values and closed graph. Let $\gamma : [0,1] \to \mathcal{M}$ be an absolutely continuous curve. Then*

$$t \mapsto \max_{v \in D(\gamma(t))} \langle v, \dot{\gamma}(t) \rangle, \tag{14}$$

defined almost everywhere on $[0,1]$, is measurable.

Definition 3 (Conservative set-valued field, cf. Def. 1 of Bolte and Pauwels (2021)) *Let $D : \mathcal{M} \rightrightarrows T^*\mathcal{M}$ be a set-valued map. We call D a* conservative field *whenever it has a closed graph, nonempty compact values and for any absolutely continuous loop $\gamma : [0,1] \to \mathcal{M}$ we have*

$$\int_0^1 \max_{v \in D(\gamma(t))} \langle \dot{\gamma}(t), v \rangle \mathrm{d}t = 0. \tag{15}$$

Equivalently, we could use the minimum in the definition.

Conservative Vector Fields on a Manifold

Lemma 5 (Chain rule, cf. Lemma 2 of Bolte and Pauwels (2021)) *Let $D : \mathcal{M} \rightrightarrows T^*\mathcal{M}$ be a locally bounded, graph closed set-valued map and $f : \mathcal{M} \to \mathbb{R}$ a locally Lipschitz continuous function. Then D is a conservative field for f if and only if, for any absolutely continuous curve $\gamma : [0,1] \to \mathcal{M}$, the function $t \mapsto f(\gamma(t))$ satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t)) = \langle v, \dot{\gamma}(t) \rangle, \qquad \forall v \in D_f(\gamma(t)), \tag{17}$$

for almost all $t \in [0,1]$.

Theorem 6 (Cf. Theorem 1 in Bolte and Pauwels (2021)) *Consider a conservative field $D : \mathcal{M} \rightrightarrows T^*\mathcal{M}$ for the potential $f : \mathcal{M} \to \mathbb{R}$. Then $D = \{\mathrm{d}f\}$ almost everywhere.*

Conservative Vector Fields on a Manifold

Definition 7 (Generalized directional derivative and Clarke subdifferential) *Let f $\mathcal{M} \to \mathbb{R}$ be a locally Lipschitz function and (U, φ) a chart at $x \in \mathcal{M}$. The* generalized directional derivative *of f at x in the direction $v \in T_x\mathcal{M}$, denoted $f^\circ(x; v)$, is then defined by*

$$f^\circ(x;v) \coloneqq \limsup_{y \to x, t \searrow 0} \frac{f \circ \varphi^{-1}(\varphi(y) + t d\varphi(x)(v)) - f \circ \varphi^{-1}(\varphi(y))}{t}. \tag{18}$$

The Clarke subdifferential of f at x, denoted $\partial f(x)$, is furthermore the subset of $T_x^*\mathcal{M}$ whose support function is $f^\circ(x; \cdot)$.

Theorem 8 (Cf. Corollary 1 in Bolte and Pauwels (2021)) *Let $f : \mathcal{M} \to \mathbb{R}$ allowing a conservative field $D : \mathcal{M} \rightrightarrows T^*\mathcal{M}$. Then ∂f is a conservative field for f, and for all $x \in \mathcal{M}$*

$$\partial f(x) \subset conv(D(x)). \tag{19}$$

Retraction Operation

For a smooth curve $\gamma : I \to \mathcal{M}$, we denote the parallel transport along γ from $\gamma(a)$ to $\gamma(b)$, for every $a, b \in I$, as $P^{\gamma}_{\gamma(a)\gamma(b)}$. It is defined by

$$P^{\gamma}_{\gamma(a)\gamma(b)}(v) := V(\gamma(b)), \qquad \text{for every } v \in T_{\gamma(a)}\mathcal{M}, \tag{1}$$

where V is the unique parallel vector field along γ with $V(\gamma(a)) = v$.
When γ is a unique minimizing geodesic between x and y we simply write P_{xy}.
The exponential map $\exp_x : T_x\mathcal{M} \to \mathcal{M}$ projects a vector from the tangent space to the manifold along a geodesic.

Retraction Operation

Assumption 2.1 (Geodesic completeness) *$\mathcal{M}$ is a connected geodesically complete Riemannian manifold. This makes the exponential map well-defined over the tangent bundle $T\mathcal{M}$.*

Definition 1 (Retraction, Def. 2 in Shah (2021)) *A retraction on $\mathcal{M}$ is a smooth mapping $\mathcal{R} : T\mathcal{M} \to \mathcal{M}$ such that*

1. *$\mathcal{R}_x(0_x) = x$, where $\mathcal{R}_x$ is the restriction of the retraction to $T_x\mathcal{M}$ and 0_x denotes the zero element of $T_x\mathcal{M}$;*
2. *with the canonical identification $T_{0_x}T_x\mathcal{M} \cong T_x\mathcal{M}$, $\mathcal{R}_x$ satisfies*

$$D\mathcal{R}_x(0_x) = Id_{T_x\mathcal{M}}, \tag{10}$$

where $Id_{T_x\mathcal{M}}$ denotes the identity operator on $T_x\mathcal{M}$.

Probability Measures on a Manifold

Let $\mathcal{L}(\mathcal{M})$ be the Lebesgue σ-algebra on $\mathcal{M}$. A subset $A \subset \mathcal{M}$ is in $\mathcal{L}(\mathcal{M})$ if, for any chart (U, φ), $\varphi(A \cap U)$ is a Lebesgue-measurable subset of $\mathbb{R}^m$. Note that $\mathcal{L}(\mathcal{M}) \supseteq \mathcal{B}(\mathcal{M})$, the Borel sigma algebra on $\mathcal{M}$. For any set $A \subset U$, with $A \in \mathcal{L}(\mathcal{M})$, we have a unique measure defined by

$$\lambda(A) = \int_{\varphi(A)} \sqrt{g} d\lambda_L,$$

where $g = \det g_{ij}$ is the determinant of the metric in local coordinates and λ_L is the Lebesgue measure on $\mathbb{R}^m$. Since this induces a volume element for each tangent space, we also get a measure on the whole manifold $\mathcal{M}$, which we denote $\lambda := \lambda(\mathcal{M})$. We can then define a probability space $(\Omega, \mathcal{B}, \mu)$ on $\mathcal{M}$.

Probability Measures on a Manifold

A random primitive on $\mathcal{M}$ is a Borelian function X from Ω to $\mathcal{M}$, with probability density function, p_X defined by

$$\begin{aligned} \mu(X \in \mathcal{X}) &= \int_{\mathcal{X}} p_X(y) d\lambda(y), \\ \mu(\mathcal{M}) &= \int \mathcal{M} p_X(y) d\lambda(y) = 1, \end{aligned} \tag{2}$$

for all $\mathcal{X}$ in the Borelian tribe of $\mathcal{M}$.

Probability Measures on a Manifold

There is some subtlety regarding the choice of metric to use when defining the pdf on a manifold. For a Borelian real valued function $\phi(x)$ on $\mathcal{M}$ we calculate the expectation value by

$$\mathbb{E}[\phi(X)] = \int_{\mathcal{M}} \phi(y) p_X(y) d\lambda(y). \tag{3}$$

We further define the variance of a random primitive X as

$$\sigma_X^2(y) = \int_{\mathcal{M}} d(x,y)^2 p_X(z) d\lambda(z), \tag{4}$$

where y is now a fixed primitive.

o-minimal Structures on a Manifold

Definition 10 (Analytic-geometric category, van den Dries and Miller (1996)) *An analytic-geometric category, $\mathcal{C}$, is given if each manifold $\mathcal{M}$ is equipped with a collection $\mathcal{C}(\mathcal{M})$ of subsets of $\mathcal{M}$ such that the following conditions hold for each manifolds $\mathcal{M}$ and $\mathcal{N}$:*

1) *$\mathcal{C}(\mathcal{M})$ is a boolean algebra of subsets of $\mathcal{M}$, with $\mathcal{M} \in \mathcal{C}(\mathcal{M})$;*
2) *if $A \in \mathcal{C}(\mathcal{M})$, then $A \times \mathbb{R} \in \mathcal{C}(\mathcal{M} \times \mathbb{R})$;*
3) *if $f : \mathcal{M} \to \mathcal{N}$ is a proper analytic map and $A \in \mathcal{C}(\mathcal{M})$, then $f(A) \in \mathcal{N}$;*
4) *if $A \subseteq \mathcal{M}$ and $\{U_i\}_{i \in I}$ is an open covering of $\mathcal{M}$, then $A \in \mathcal{C}(\mathcal{M})$ iff $A \cap U_i \in \mathcal{C}(U_i)$ for all $i \in I$;*
5) *every bounded set in $\mathcal{C}(\mathbb{R})$ has finite boundary.*

Definition 11 (Whitney stratification) *A Whitney C^k stratification $M = \{M_i\}_{i \in I}$ of a set A is a partition of A into finitely many non-empty C^k submanifolds, or strata, satisfying:*

- ***Frontier condition:*** *For any two strata M_i and M_j, the following implication holds,*

$$\overline{M}_i \cap M_j \neq \emptyset \implies M_j \subset \overline{M}_i. \tag{21}$$

- ***Whitney condition (a):*** *For any sequence of points x_k in a stratum M_i converging to a point x in a stratum M_j, if the corresponding normal vectors $v_k \in N_{M_i}(x_k)$ converge to a vector v, then the inclusion $v \in N_{M_j}(x)$ holds.*

o-minimal Structures on a Manifold

Definition 12 (Variational stratification) *Let $f : \mathcal{M} \to \mathbb{R}$ be locally Lipschitz continuous, $D : \mathcal{M} \rightrightarrows T^*\mathcal{M}$ a set-valued map and let $k \geq 1$. We say that (f, D) has a C^k* variational stratification *if there exists a C^k Whitney stratification M of $\mathcal{M}$ such that f is C^k on each stratum and for all $x \in \mathcal{M}$:*

$$Proj_{T_x M_x} D(x) = \{d_x f(x)\}, \tag{22}$$

where $d_x f(x)$ is the differential of f restricted to the active strata M_x containing x.

Theorem 13 (Variational stratification for definable conservative fields) *Let $D : \mathcal{M} \rightrightarrows T^*\mathcal{M}$ be a definable conservative field having a definable potential $f : \mathcal{M} \to \mathbb{R}$. Then (f, D) has a C^k variational stratification.*

The Whitney stratifiability of the $\mathcal{C}$-maps allows us to make some important claims. The following will be important:

Theorem 14 (Non-smooth Morse-Sard, cf. Theorem 5 in Bolte and Pauwels (202 *Let $D : \mathcal{M} \rightrightarrows T^*\mathcal{M}$ be a conservative field for $f : \mathcal{M} \to \mathbb{R}$ and assume that f and D are definable. Then the set of D-critical values, $\{f(x) : x \in \mathcal{M}$ is D-critical for $f\}$, is finite.*

Diminishing Stepsize Stochastic Approximation

Now we consider

$$x_{k+1} = R_{x_k}(x_k - \alpha_k g_k),\ g_k \sim \partial F(x_k, \cdot)$$

with

$$\alpha_k \to 0$$

in a *Stochastic Approximation* framework

Consider the metric space with the distance of uniform convergence on the set of continuous functions $C(\mathbb{R}, \mathcal{M}, d_C)$ endowed with the metric of uniform convergence on compact sets,

$$d_C(x(t), y(t)) := \sum_{k=1}^{\infty} \frac{1}{2^k} \min\left(\int_{-k}^{k} d(x(t), y(t))dt, 1\right)$$

Given a set-valued map $G : \mathcal{M} \rightrightarrows T\mathcal{M}$, we call an absolutely continuous curve $\gamma : [0, a] \to \mathcal{M}$ a solution to the differential inclusion

$$\dot{\gamma}(t) \in G(\gamma(t)), \qquad x_0 \in \mathcal{M}, \tag{23}$$

Diminishing Stepsize Stochastic Approximation

Assumption 4.1 *1. The steps $\{\alpha_k\}_{n\in\mathbb{N}^*}$ form a sequence of non-negative numbers such that*

$$\lim_{k\to\infty}\alpha_k=0,\quad \sum_k \alpha_k=\infty\quad \sum_k \alpha_k^2<\infty. \tag{28}$$

2. For all $T>0$ and any $x\in\mathcal{M}$

$$\limsup_{n\to\infty}\Big\{\sum_{i=n}^{k-1}\alpha_{i+1}g(\iota G(x_{i+1}),\iota g_{i+1}):\\ k=n+1,\dots,m(\tau_n+T)\Big\}=0, \tag{29}$$

with

$$m(t)=\sup\{k\geq 0\,:\,t\geq\tau_k\},\quad \tau_n=\sum_{i=1}^n\alpha_i, \tag{30}$$

and $\tau_0=0$.

3. $\sup_n d(x_n,z)<\infty$ for any point $z\in\mathcal{M}$.

The equation

$$\Theta^t(\gamma)(s)=\gamma(s+t) \tag{31}$$

defines a translation flow $\Theta^t\,:\,C(\mathbb{R},\mathcal{M})\times\mathbb{R}\to C(\mathbb{R},\mathcal{M})$. We call a continuous curve $\zeta\,:\,\mathbb{R}_+\to\mathcal{M}$ an *asymptotic pseudo trajectory* (APT) for Φ if

$$\lim_{t\to\infty}d_C(\Theta^t(\zeta),\mathcal{S}_{\zeta(t)})=0. \tag{32}$$

Diminishing Stepsize Stochastic Approximation

Theorem Let $f : \mathcal{M} \to \mathbb{R}$ be a locally Lipschitz C^k-stratifiable function. Consider the iterates $\{x_k\}_{k\geq 1}$ produced by the diminishing stepsize Stochastic Approximation process with $G = -\iota(\text{conv}(D_f))$, where ι is the musical isomorphism $T\mathcal{M}^* \to T\mathcal{M}$. Then every limit point of the iterates $\{x_k\}_{k\geq 1}$ is critical for f and the function values $\{f(x_k)\}_{k\geq 1}$ converges.

Constant Stepsize Markov Process

Now we consider

$$x_{k+1} = R_{x_k}(x_k - \alpha g_k),\ g_k \sim \partial F(x_k, \cdot)$$

with α constant, inducing a *Markovian* analysis, with ergodicity results.

Definition 20 (Almost everywhere gradients, Def. 1 in Bianchi et al. (2022)) *Assume that $f(\cdot, s)$ is locally Lipschitz continuous for every $s \in \Omega$. A function $\phi : \mathcal{M} \times \Omega \to T\mathcal{M}$ is called an almost everywhere (a.e.) gradient of f if $\phi = \nabla f$ $\lambda \otimes \mu$-almost everywhere.*

The following proposition makes this a relevant definition for us.

Proposition 21 (Prop. 1 in Bianchi et al. (2022)) *Assume that for any $s \in \Omega$, $f(\cdot, s)$ is locally Lipschitz, path differentiable, and is a potential of a conservative field $D_s : \mathcal{M} \rightrightarrows T\mathcal{M}$. Consider a $\mathcal{B}(\mathcal{M}) \otimes \mathcal{J}/\mathcal{B}(\mathcal{M})$-measurable function $\phi : \mathcal{M} \times \Omega \to T\mathcal{M}$ satisfying $\phi(x, z) \in D_s(x)$ for all $(x, s) \in \mathcal{M} \times \Omega$. Then ϕ is an a.e. gradient function for f.*

Definition 22 (SGD sequence, Def. 2 in Bianchi et al. (2022)) *Let f be $\mathcal{B}(\mathcal{M}) \otimes \mathcal{J}/\mathcal{B}(\mathcal{M})$-measurable, and assume $f(\cdot, s)$ is locally Lipschitz for any $s \in \Omega$. A sequence $\{x_n\}_{n \in \mathbb{N}^*}$ of functions on $\tilde{\Omega} \to \mathcal{M}$ is called an SGD sequence for f with steps $\alpha_n > 0$ if there exists an a.e. gradient ϕ of f such that*

$$x_{n+1} = \exp_{x_n}\left[\alpha_n \phi(x_n, \xi_{n+1})\right], \quad \forall n \geq 0. \tag{34}$$

Constant Stepsize Markov Process

Assumption 4.2 *We make the following assumptions on the function $f : \mathcal{M} \times \Omega \to \mathbb{R}$ having an SGD sequence.*

1. *There exists a measurable function $\kappa : \mathcal{M} \times \Omega \to \mathbb{R}_+$ such that for each $x \in \mathcal{M}$ we have $\int \kappa(x,s)\mu(\mathrm{d}s) < \infty$ and there exists an $\varepsilon > 0$ for which*

$$\forall y, z \in B(x, \varepsilon),\ \forall s \in \Omega,\ |f(y,s) - f(z,s)| \leq \kappa(x,s)d(y,z). \tag{37}$$

   i.e., $f(\cdot, s)$ is geodesically $\kappa(\cdot, s)$-Lipschitz for all $s \in \Omega$.

2. *For all $x \in \mathcal{M}$, $f(x, \cdot)$ is μ-integrable.*
3. *There exists a $0_{\mathcal{M}} \in \mathcal{M}$ and a constant $K \geq 0$ such that $\int \kappa(x,s)\mu(\mathrm{d}s) \leq Kd(0_{\mathcal{M}}, x)$ for all $x \in \mathcal{M}$.*
4. *For each compact set $\mathcal{K} \subset \mathcal{M}$, $\sup_{x \in \mathcal{K}} \int \kappa(x,s)^2 \mu(\mathrm{d}s) < \infty$.*

Constant Stepsize Markov Process

Theorem 23 *Let the above assumptions hold true (in fact we only need 1 and 2). Consider* $\alpha \in \Gamma$ *and* $\nu \in \mathcal{P}_{abs.}(\mathcal{M}) \cap \mathcal{P}_1(\mathcal{M})$*. Let* $\{x_k\}_{k \in \mathbb{N}^*}$ *be an SGD sequence for* f *with steps* α*. Then, for any* $k \in \mathbb{N}$*, it holds* $\mathbb{P}^\nu$*-a.e. that*

1. F*,* $f(\cdot, \xi_{k+1})$ *and* $f(\cdot, s)$ *(for* μ*-a.e.* s*) are differentiable at* x_k*, with* F *as above;*
2. $x_{k+1} = \exp_{x_k}[\alpha \operatorname{grad} f(x_k, \xi_{k+1})]$*;*
3. $\mathbb{E}_k[x_{k+1}] = \exp_{x_k}[\alpha \operatorname{grad} F(x_k)]$.

13

Constant Stepsize Markov Process

Theorem 24 *Under the standing assumptions, let* $\{(x_k^\alpha)_{k\in\mathbb{N}^*} : \alpha \in (0, \alpha_0]\}$ *be a collection of SGD sequences of steps* $\alpha \in (0, \alpha_0]$. *Define* x^α *iteratively to be:*

$$x^\alpha(t) = \gamma^k(t/\alpha - k), \forall t \in [k\alpha, (k+1)\alpha)$$

where $\gamma^k : [0,1] \to \mathcal{M}$ *is the geodesic curve with constant velocity* $\|\dot{\gamma}^k\|$ *from* $\gamma^k(0) = x_k$ *to* $\gamma^k(1) = x_{k+1}$.

It holds that for every compact set $\mathcal{K} \subset \mathcal{M}$,

$$\forall \epsilon > 0, \lim_{\alpha\to 0, \alpha\in\Gamma} \left(\sup_{\nu\in\mathcal{P}_{abs}(\mathcal{K})} \mathbb{P}^\nu(d_C(x^\alpha, \mathcal{S}_{-\partial F}(\mathcal{K})) > \epsilon) \right) = 0$$

Moreover, the family of distributions $\{\mathbb{P}^\nu(x^\alpha)^{-1} : \nu \in \mathcal{P}_{abs}(\mathcal{K}), 0 < \alpha < \alpha_0, \alpha \in \Gamma\}$ *is tight.*

Constant Stepsize Markov Process

Theorem 25 (Convergence – constant step size) *Let the standing assumptions hold true. Let $\{(x_n^\alpha)_{n\in\mathbb{N}} : \alpha \in (0, \alpha_0]\}$ be a collection of SGD sequences of step size α. Then, the set $\mathcal{Z} := \{x : 0 \in \partial F(x)\}$ is nonempty and for all $\nu \in \mathcal{P}(\mathcal{M})$ and all $\epsilon > 0$,*

$$\limsup_{n\to\infty} \mathbb{P}^\nu \left(\boldsymbol{d}(x_n^\alpha, \mathcal{Z}) > \epsilon \right) \Longrightarrow_{\alpha\to 0, \alpha\in\Gamma} 0. \tag{41}$$

Numerical Results

Sparse PCA

$$\min_{X\in\mathcal{M}} \quad -tr(X^TA^TAX)+\rho\|X\|_1$$
$$\mathcal{M} := \{X\in\mathbb{R}^{n\times p},\, X^TX=I_p\}$$

In order to consider the problem as stochastic, at each iteration, we sample a subset of rows of A, i.e.,

$$A=\mathbb{E}[A(\xi)]=n\begin{pmatrix}\mathbf{1}_p(1)a_1\\ \mathbf{1}_p(2)a_2\\ \cdots\end{pmatrix}$$

where with probability $1/n$ we sample $p\in[n]$.

Numerical Results

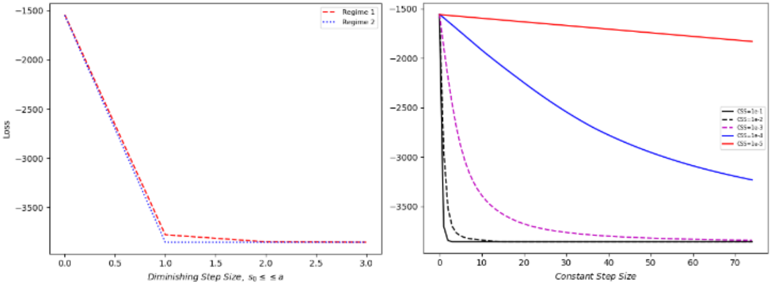

Figure 1: The Loss of Objective Function in RSGD for Sparse PCA

Numerical Results

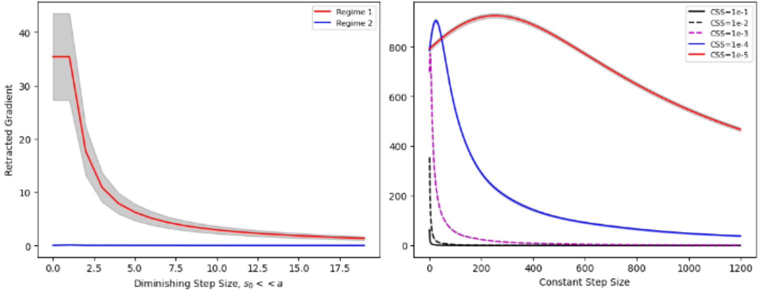

Figure 2: Methods Convergence and Margin of Errors in RSGD for Sparse PCA

Numerical Results

Low Rank Matrix Completion

$$\min_{X \in \mathcal{M}} \quad \sum_{i,j} |A_{ij} - X_{ij}|$$
$$\mathcal{M} := \{X \in \mathbb{R}^{m \times n},\ rank(X) = p\}$$

Numerical Results

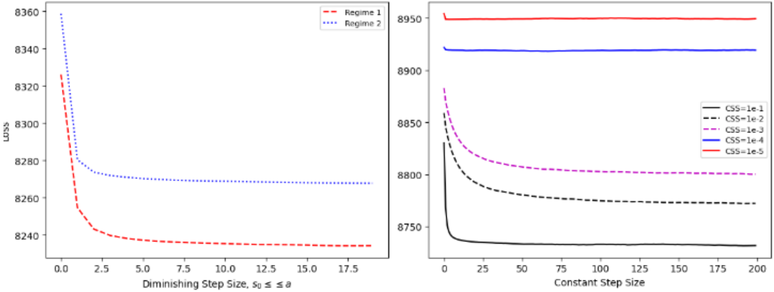

Figure 4: The Loss of Objective Function in RSGD: Low Rank Matrix Completion

Numerical Results

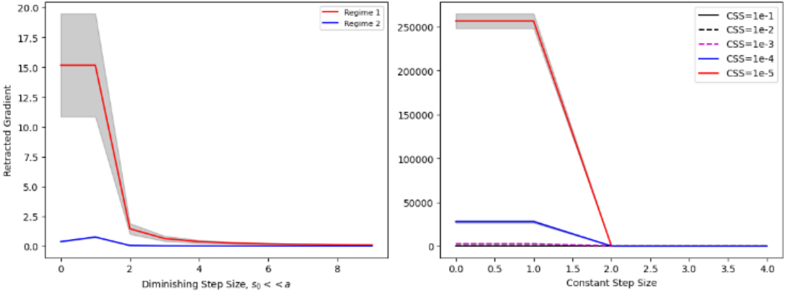

Figure 5: Methods Convergence and Margin of Errors in RSGD: Low Rank Matrix Completion

ReLU Neural Network with Batch Normalization

$$\min_{w\in\mathcal{M}} \quad \frac{1}{N}\sum_{i=1}^{N} |\hat{y}(x_i; w) - y_i|$$
$$\mathcal{M} := \{x \in \mathbb{S}^{n_1} \times \mathbb{S}^{n_2} \times \cdots \times \mathbb{S}^{n_L} \times \mathbb{R}^{n_o}\}$$

Numerical Results

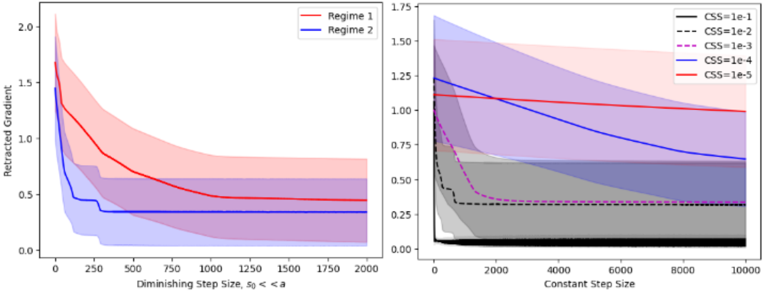

Figure 7: Methods Convergence and Margin of Error in RSGD: ReLU Neural Network with Batch Normalization

Conclusion

- On arxiv, soon to be updated
- Fascinating interplay of topology, measure theory, and differential geometry
- Many possible extensions to consider
- Pymanopt Alternative Suggestions?