Geodesic Convex Optimization

Jacob Krantz, Rohith Sajith, Vishal Raman

May 8, 2021

1 Geodesic Convexity

1.1 Geodesically Convex Sets

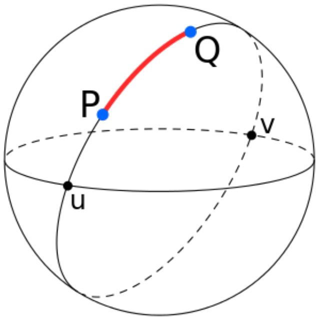
We begin by defining total and geodesic convexity. In all of the following definitions, let (M, g) be a Riemannian manifold.

Definition 1.1 (Total Convexity)**.** A set $K \subseteq M$ is said to be totally convex with respect to g, if for any $p, q \in K$, any geodesic γ_{pq} that joins p to q lies entirely in K.

Note that in the Euclidean case, total geodesic convexity reduces to standard convexity. For a set $K \subset \mathbb{R}^n$, the unique geodesic between two points p, q is given by the straight line between these points. Hence, K containing the geodesic is exactly the condition required for usual convexity.

Definition 1.2 (Geodesic Convexity)**.** A set $K \subseteq M$ is said to be geodesically convex with respect to g, if for any $p, q \in K$, there is a unique minimizing geodesic γ_{pq} contained with K that joins p and q.

Note that total convexity is a stronger notion than geodesic convexity. In the case where there is a unique geodesic joining the points, the definitions are the same. To see that total convexity is more restrictive in general, consider the following example. Let $\mathbb{S}^n = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ with the metric induced by Euclidean distance. Take $P, Q \in \mathbb{S}^n$.



The two geodesics joining P and Q are the two arcs on the corresponding great circle (depicted above). For a subset $K \subseteq M$, in order for K to be geodesically convex, we would require that the red arc is contained in the set. In order for K to be totally convex, it would need to contain both the red arc and the black arc.

1.2 Geodesically Convex Functions

Definition 1.3 (Geodesically Convex Function)**.** Let $K \subseteq M$ be totally convex with respect to g. A function $f : K \to \mathbb{R}$ is geodesically convex with respect to g if for any $p, q \in K$ and for any geodesic $\gamma_{pq} : [0, 1] \to K$ joining p, q, for all $t \in [0, 1]$,

$$f(\gamma_{pq}(t)) \leq (1 - t)f(p) + tf(q).$$

As before, note that convex functions are geodesically convex with respect to the Euclidean metric. This definition is often unwieldy, but with stronger conditions on the function, we gain a more practical definition.

Theorem 1.4 (Higher-Order Characterizations)**.** *Let (M, g) be a Riemannian manifold and let $K \subseteq M$ be an open totally convex set with respect to g.*

- *A differentiable function $f : K \to \mathbb{R}$ is geodesically convex if and only if for any $p, q \in K$ and for any geodesic $\gamma_{pq} : [0, 1] \to K$ joining p and q,*

$$f(p) + \dot{\gamma}_{pq}(f)(p) \leq f(q).$$

- *A twice-differentiable function $f : K \to \mathbb{R}$ is geodesically convex with respect to g if and only if for any $p, q \in K$ and for any geodesic $\gamma_{pq} : [0, 1] \to K$ joining p and q,*

$$\frac{d^2 f(\gamma_{pq}(t))}{dt^2} \geq 0.$$

1.3 Examples

We finish the section with some examples of geodesically convex functions.

Example 1.5. *Take a multivariate polynomial $p(x) = \sum_{\lambda \in \Lambda} c_\lambda x^\lambda$ with $c_\lambda \in \mathbb{R}^+$ where $\Lambda \subseteq \mathbb{N}^m$ is a finite set of multi-indices of order m. We claim that $\log p(x)$ is geodesically convex with respect to the Euclidean metric.*

Proof. We prove this via the second-order characterization of geodesic convexity. The geodesics are of the form $\gamma(t) = \exp(\alpha t + \beta)$ for $\beta, \alpha \in \mathbb{R}^n$. It suffices to show that for all $t \in [0, 1]$, we have

$$\frac{d^2 \log p(\gamma(t))}{dt^2} \geq 0.$$

First, note that

$$p(\gamma(t)) = \sum_{\lambda \in \Lambda} c_\lambda \exp(\langle \lambda, \alpha \rangle t + \langle \lambda, \beta \rangle).$$

The first derivative is given by

$$\frac{d \log p(\gamma(t))}{dt} = \frac{p'(\gamma(t))}{p(\gamma(t))} = \frac{\sum_{\lambda \in \Lambda} c_\lambda \langle \alpha, \lambda \rangle \exp(\langle \lambda, \alpha \rangle t + \langle \lambda, \beta \rangle)}{\sum_{\lambda \in \Lambda} c_\lambda \exp(\langle \lambda, \alpha \rangle t + \langle \lambda, \beta \rangle)}.$$

Then, the second derivative it given by

$$\begin{aligned}\frac{d^2 \log p(\gamma(t))}{dt^2} &= \frac{p(\gamma(t))p''(\gamma(t)) - (p'(\gamma(t)))^2}{(p(\gamma(t)))^2} \\ &= \frac{p''(\gamma(t))}{p(\gamma(t))} - \left(\frac{p'(\gamma(t))}{p(\gamma(t))}\right)^2 \\ &= \frac{\sum_{\lambda,\xi\in\Lambda}(c_\lambda\langle\lambda,\alpha\rangle - c_\xi\langle\xi,\alpha\rangle)^2 \exp(\langle\lambda+\xi,\alpha\rangle t + \langle\lambda+\xi,\beta\rangle)}{\left(\sum_{\lambda\in\Lambda} c_\lambda \exp(\langle\lambda,\alpha\rangle t + \langle\lambda,\beta\rangle)\right)^2} \geq 0.\end{aligned}$$

■

Example 1.6. *Let $\mathcal{S}^d_{++}$ denote the set of positive definite matrices $d \times d$ matrices, with metric $g_X(U,V) = \operatorname{tr}[X^{-1}UX^{-1}V]$. The function $f(X) = \log\det(X)$ is geodesically convex on $\mathcal{S}^d_{++}$.*

Proof. We prove this without appealing to the higher-order characterizations. Let $X, Y \in P$ and take $t \in [0,1]$. The geodesic joining X and Y is given by

$$\gamma(t) = X^{1/2}(X^{-1/2}YX^{-1/2})^t X^{1/2}.$$

It follows that

$$\begin{aligned}\log\det(\gamma(t)) &= \log\det(X^{1/2}(X^{-1/2}YX^{-1/2})^t X^{1/2}) \\ &= 2\log\det(X^{1/2}) + 2t\log\det(X^{-1/2}) + t\log\det Y \\ &= (1-t)\log\det X + t\log\det Y.\end{aligned}$$

In this case, we say that the function is *geodesically linear.* ■

2 Operator Scaling

The operator scaling problem is as follows: given a linear operator $T(X) = \sum_{j=1}^m T_j^\top X T_j$ defined by matrices T_j, find square matrices L and R such that

$$\sum_{j=1}^m \widehat{T}_j^\top \widehat{T}_j = I, \qquad \sum_{j=1}^m \widehat{T}_j \widehat{T}_j^\top = I,$$

where $\widehat{T}_j = LT_jR$.

Definition 2.1 (Operator Capacity)**.** Let $\mathbb{S}^d_{++}$ denote the set of positive definite $d \times d$ matrices. For a linear operator $T : \mathbb{S}^d_{++} \to \mathbb{S}^{d'}_{++}$, define the capacity of T as

$$\operatorname{cap}(T) := \inf_{\det(X)=1} \left(\frac{d}{d'} T(X)\right).$$

In order to reduce the problem to geodesic convex optimization, we introduce the following operator:

Theorem 2.2. *If T is a positive linear operator,* $\operatorname{logcap}(X) := \log\det T(X) - \log\det X$ *is geodesically convex.*

Proof. By Proposition 5.9 of [1], we have that $\log\det T(X)$ is geodesically convex, and we showed in Example 1.6 that $\log\det X$ is geodesically linear. Hence, the difference between the two functions is geodesically convex. ■

2.1 Brascamp-Lieb Constants

The Brascamp-Lieb inequality is an important generalization of Hölder's inequality with broad applications in functional analysis, convex geometry, and computer science. It is stated as follows:

Theorem 2.3 (Brascamp-Lieb Inequality)**.** *Let $m \in \mathbb{N}$, $n \in \mathbb{N}^m$ and take $p \in \mathbb{R}^m_{\geq 0}$. Let $B = (B_j)_{j=1}^m$ to be the concatenation of m surjective linear transformations $B_j : \mathbb{R}^n \to \mathbb{R}^{n_j}$. There exists a constant $C = C(B, p)$ (possibly infinite), so that for any real-valued, non-negative, Lebesgue-measurable functions $f_j : \mathbb{R}^{n_j} \to \mathbb{R}$, we have*

$$\int_{\mathbb{R}^n} \left(\prod_{j=1}^m f_j(B_j(x))^{p_j} \right) dx \leq C \prod_{j=1}^m \left(\int_{\mathbb{R}^{n_j}} f_j(x)\, dx \right)^{p_j}.$$

Definition 2.4. Define $\mathrm{BL}(B, p)$ to be the minimal constant (possibly infinite) satisfying the Brascamp-Lieb inequality for any choice of functions $(f_j)_{j=1}^m$ satisfying the above properties.

Definition 2.5. The pair (B, p) is called feasible if $\mathrm{BL}(B, p) < \infty$, otherwise it is called infeasible. For a fixed m-tuple $(B_j)_{j=1}^m$, we define the set

$$P_B : \{p \in \mathbb{R}^m : \mathrm{BL}(B, p) < \infty\}.$$

2.1.1 Reduction to Operator Scaling

Let (B, p) be as described above. If the exponent $p = (p_j)_{j \in [m]}$ is a rational vector in the sense that each of the components is rational, then we can reduce the problem of calculating $\mathrm{BL}(B, p)$ to operator scaling. In particular, it is possible to construct an operator $T_{(B,p)} : \mathbb{S}^{nc}_{++} \to \mathbb{S}^n_{++}$ such that $\operatorname{cap}(T_{(B,p)}) = 1/\mathrm{BL}(B, p)^2$. To formally construct this operator, we will need to use c_j copies of the matrix B_j for each $j \in [m]$. For ease of notation, define $m' = \sum_{j \in [m]} c_j$ and $\delta : [m'] \to [m] : i \to j$, where j satisfies

$$\sum_{k<j} c_k < i \leq \sum_{k \leq j} c_k.$$

Now, we define Z_{ij} to be an $n_{\delta(i)} \times n$ matrix that is zero if $\delta(i) \neq j$ and $B_{\delta(i)}$ if $\delta(i) = j$, for $(i, j) \in [m'] \times [m']$. Then, for each $j \in [m']$, we can define

$$T_j = \begin{pmatrix} Z_{1j} \\ \vdots \\ Z_{m'j} \end{pmatrix}$$

so that

$$T_{(B,p)} = \sum_{j \in [m']} T_j^\top X T_j$$

is the desired operator. To summarize, we have the following theorem.

Theorem 2.6. *For given (B, p) as described above, the $T_{(B,p)}$ we just defined satisfies* $\operatorname{cap}(T_{(B,p)}) = 1/\operatorname{BL}(B,p)^2$ *[6, Lemma 4.4]. Thus, we can apply operator scaling to solve this problem.*

3 Optimistic Likelihood Calculation

A common problem in machine learning and other disciplines is the following: given a set of Gaussian distributions and points $\{x_i\}$ generated from exactly one of the distributions, find the distribution that maximizes the likelihood (or equivalently, the log-likelihood) that the $\{x_i\}$ belong to it.

Definition 3.1 (Maximum Likelihood Estimation)**.** If we parameterize the distributions $\{(\mu_i, \Sigma_i)\}$, then standard maximum log-likelihood problem can be written as

$$\sigma^* = \operatorname*{argmax}_i -\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_i)^\top \Sigma_i^{-1}(x_n - \mu_i) - \log\det\Sigma_i.$$

Because the distribution parameters are almost always not exactly recoverable, μ_i and Σ_i are usually replace with their empirical counterparts $\widehat{\mu_i}$ and $\widehat{\Sigma_i}$, which can be computed from data in the usual way. For the sake of notational convenience, we will still use μ_i and Σ_i to denote the empirical mean and empirical covariance of the each distribution.

Since methods of sampling points to obtain good estimators of μ_i and Σ_i are typically costly, we modify the maximum log liklelihood problem into the *optimistic* log likelihood problem - instead of maximizing the log likelihood that the data comes from one of the given Gaussians, we maximize over the set of all Gaussians that are sufficiently close to one of the given Gaussians under some distance measure φ.

Definition 3.2 (Optimistic Log-Likelihood)**.** The optimistic log likelihood problem optimizes the same function as the MLE objective, but does so over

$$S = \{\mathbb{P} \in \mathcal{M} : \varphi(\mathbb{P}_i, \mathbb{P}) \le \rho_i\},$$

where $\mathcal{M}$ is the set of all Gaussians over $\mathbb{R}^n$, $\mathbb{P}_i$ is the ith Gaussian in our original set, and ρ_i is some constant associated to each $\mathbb{P}_i$ that indicates how well we know its parameters.

For the following analysis, we assume that all the $\{\mu_i\}$ are the same so that each Gaussian can be parameterized by its covariance.

Definition 3.3 (Fisher-Rao Metric)**.** The Fisher Rao Metric φ for Gaussians with covariances Σ_1 and Σ_2 on the manifold of positive semidefinite matrices is defined as:

$$\varphi(\Sigma_1, \Sigma_2) = \frac{1}{\sqrt{2}}|| \log(\Sigma_2^{-\frac{1}{2}}\Sigma_1\Sigma_2^{-\frac{1}{2}})||_F.$$

The crucial ideas to solving Optimistic Likelihood are that S is a geodesically convex set and that the optimistic log-likelihood objective using the FR metric φ is a geodesically convex function over S.

By Proposition II.1.4 in [8], the manifold $\mathbb{S}^n_{++}$ is a Hadamard space, which means that balls of radius less than $\frac{D_0}{2} = \infty$ in a Hadamard space are geodesically convex, where D_0 is the diameter of a sphere in $\mathbb{S}^n_{++}$ with zero curvature. Since the set S consists of a ball of finite radius, it is geodesically convex.

Theorem 3.4 (Geodesic Convexity of Optimistic MLE Loss Function)**.** *The following function is geodesically convex over* $S = \{\mathbb{P} \in \mathcal{M} : \varphi(\mathbb{P}_i, \mathbb{P}) \leq \rho_i\}$:

$$L(\Sigma) = -\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_i)^\top \Sigma^{-1}(x_n - \mu_i) - \log\det\Sigma.$$

Proof. From Lemma II.1 in [9], for a continuous function on $\mathbb{S}^n_{++}$, midpoint geodesic convexity is equivalent to geodesic convexity. Therefore, it suffices to show that for endpoints Σ_1 and Σ_2 that

$$L(\Sigma_1) + L(\Sigma_2) \geq 2L(\Sigma_1^{1/2}(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})^{1/2}\Sigma_1^{1/2}),$$

where the argument of L on the right hand side is the midpoint of the geodesic connecting Σ_1 and Σ_2.

Since L is composed of a quadratic form and the log of a determinant, midpoint geodesic convexity relies on the following two lemmas.

Lemma 3.5. *For* $\Sigma_1, \Sigma_2 \in \mathbb{S}^n_{++}$,

$$\log(\det(\Sigma_1)) + \log(\det(\Sigma_2)) = 2\log(\det(\Sigma_3)),$$

where Σ_3 *is the midpoint of the geodesic connecting* Σ_1, Σ_2

Proof. By the equation $\gamma(t)$ for a geodesic, we can plug in $t = 1/2$ to get a closed form for Σ_3. Using this, we get

$$\Sigma_3\Sigma_1^{-1}\Sigma_3 = \Sigma_1^{1/2}(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})^{1/2}\Sigma_1^{1/2}\Sigma_1^{-1}\Sigma_1^{1/2}(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})^{1/2}\Sigma_1^{1/2} = \Sigma_2$$

after performing some cancellations. Using the product rule for determinants, one can swiftly see that $\det(\Sigma_1)\det(\Sigma_2) = \det(\Sigma_3)^2$, and taking the logarithm of both sides gives the desired result. ■

Lemma 3.6. *For* $\Sigma_1, \Sigma_2 \in \mathbb{S}^n_{++}$ *and* Σ_3 *defined similarly,*

$$\log(x^\top\Sigma_1 x) + \log(x^\top\Sigma_2 x) \geq 2\log(x^\top\Sigma_3 x).$$

Proof. If we define an appropriate change of coordinates so that $x^\top\Sigma_1 x = y^\top y$, then we can write $x^\top\Sigma_2 x = y^\top\Sigma_0 y$ and $x^\top\Sigma_3 x = y^\top\Sigma_0^{1/2}y$, the inequality (if we get rid of the logarithms) we want to prove is

$$\sum_{i=1}^{k}\lambda_i y_i^2 \sum_{i=1}^{k} y_i^2 \geq \left(\sum_{i=1}^{k}\lambda_i^{1/2}y_i^2\right)^2,$$

where $\{\lambda_i\}$ is the eigenvalue decomposition of Σ_0. This is clearly true by Cauchy-Schwartz. ■

Since the sum of two geodesically convex functions is convex, Lemmas 3.2 and 3.3 imply geodesic convexity for L. ■

Since we have now shown that optimistic likelihood is geodesically convex, we can apply geodesic gradient descent to the problem, as illustrated in section 4.3.

4 Sampling and Optimization on Manifolds

In this section, we present methods to sample from convex subsets of Riemannian manifolds with curvature constraints. This will allow us to develop a stochastic algorithm to estimate global optima for geodesically convex functions. We also present a geodesic gradient descent algorithm.

4.1 Geodesic Random Walk

In order to sample from a convex subset K, we construct a random walk procedure whose stationary distribution is uniform on K.

Take $X_0 \in K$ as the starting point, η as the step size, and N as the number of steps. At step $t+1$, we sample from the tangent space at the point X_t and move in this direction along the corresponding geodesic. If it lands outside of K, we reset the point $X_{t+1} = X_t$. More precisely, the algorithm is as follows:

Algorithm 4.1 Geodesic Walk on a Manifold

```
function GeodesicWalk($n, K, X_0, \eta, N$)
    for $t < N$ do
        Choose $u_{t+1}$ from $N(0, I_n)$ on $T_{X_t}M$.
        if $y = \exp_{X_t}(\eta u_{t+1}) \in K$ then
            Set $X_{t+1} = y$.
        else Set $X_{t+1} = X_t$.
    return $X_N$
```

To measure how close distributions are to each other, we introduce the following definitions,

Definition 4.2 (H-warm)**.** Given random variables X, Y on the same space, we say that X is H-warm with respect to Y if

$$\sup_A \frac{\mathbb{P}[X \in A]}{\mathbb{P}[Y \in A]} \le H.$$

Definition 4.3. For two distributions X, Y with corresponding densities f_X, f_Y, define the distance

$$\|X/Y\| = \int \left(\frac{f_X(t)}{f_Y(t)}\right)^2 f_Y(t)\,dt = \int \left(\frac{f_X(t)}{f_Y(t)}\right) f_X(t)\,dt.$$

Using the above definition, we can obtain the following bounds on the algorithm:

Theorem 4.4 (Geodesic Random Walk)**.** *Let (M, g) be an n-dimensional Riemannian manifold with non-negative Ricci curvature and $K \subset M$ be a strongly convex subset satisfying the following conditions:*

- *M has non-negative sectional curvature,*
- *The Riemann curvature tensor is bounded from above;* $\max \|R_M\|_F \leq R$.
- *K contains a ball of radius r and has diameter D.*

If our starting distribution is H-warm with respect to the uniform distribution on K, then in

$$t = O\left(\frac{H^2 D^2 n^3 (R+1)}{r^2 \epsilon^2} \log(H/\epsilon)\right)$$

steps, the Geodesic Random Walk algorithm returns a distribution that is ϵ-close to the uniform distribution on K.

4.2 Simulated Annealing

Finally, we present a stochastic optimization algorithm on Riemannian Manifolds called Simulated Annealing. This is named after annealing in metallurgy, where metals are heated and cooled in a controlling setting to remove imperfections. In particular, the processes mimics the activity of the misplaced atoms under the temperature changing procedure.

Let (M, g) be a Riemannian Manifold. We will find global optima for a geodesically convex function $f : K \to \mathbb{R}$, where $K \subset M$ is a strongly convex subset satisfying the conditions of Theorem 4.3. First, we define a probability density $\pi_{f,T} \sim e^{-f/T}$. Note that the density is concentrated about points where f is small. Moreover, as T gets smaller, the contribution of the minima of f to the density grow exponentially.

Now, we modify the Geodesic Random Walk algorithm to incorporate the probability density. Namely, when setting the value of X_{t+1}, we introduce a filter that accepts the new value of X_{t+1} with probability $\min\left(1, e^{\frac{f(y)-f(X_t)}{T}}\right)$. Otherwise, we keep $X_{t+1} = X_t$. By making this change, the point we return at the end of the algorithm $X_N \in K$ is sampled approximately proportional to $e^{-f/T}$. More explicitly, the algorithm is as follows:

Algorithm 4.5 Adapted Geodesic Walk on a Manifold

```
function GeodesicWalk($n, K, f, X_0, \eta, N, T$)
    for $t < N$ do
        Choose $u_{t+1}$ from $N(0, I_n)$ on $T_{X_t}M$.
        if $y = \exp_{X_t}(\eta u_{t+1}) \in K$ then
            Set $X_{t+1} = y$ with probability $\min\left(1, e^{\frac{-f(y)+f(X_t)}{T}}\right)$.
            Otherwise, set $X_{t+1} = X_t$.
        else Set $X_{t+1} = X_t$.
    return $X_N$
```

For a fixed value of T, the optimization problem is just as difficult as sampling the optima from the uniform distribution. To resolve this issue, we progressively lower the temperature so that the concentration about the minima increase and it is easier to sample from the new distribution. Now, we can perform simulated annealing acoording to an annealing schedule $\{T_i\}$ as follows:

Algorithm 4.6 Simulated Annealing on a Manifold

```
function ANNEALING($n, K, f, X_0, N, \{T_i\}$)
    for $t < N$ do
        Sample $X_t$ according to the distribution $\pi_{f,T_t}$ using geodesic walk with $X_0 = X_t$
    return $X_N$
```

In order to ensure that the distributions π_{f,T_i} and $\pi_{f,T_{i+1}}$ are sufficiently close and that the length of the sequence of temperatures is small, we explicitly choose

$$T_{i+1} = \left(1 - \frac{1}{\sqrt{n}}\right) T_i.$$

Namely, we have that $\|\pi_{f,T_i}/\pi_{f,T_{i+1}}\| \leq 5$(see Theorem 11.9 in [5]). This leads to the following runtime bound:

Theorem 4.7. *Starting from a uniform sample from K, the simulated annealing algorithm runs in*

$$O\left(\frac{D^2 n^{7.5}(R+1)L^2}{r^2\varepsilon^2\delta^6} \log\left(\frac{n}{\delta}\log\left(\frac{T_0(n+1)}{\varepsilon\delta}\right)\right)\log^5\left(\frac{T_0 n}{\varepsilon\delta}\right)\right)$$

steps and returns x_* such that

$$f(x_*) - \inf_{x\in K} f(x) \leq \varepsilon$$

with probability $1-\delta$.

Proof. First, we choose the initial temperature so that π_{f,T_0} is sufficiently close to the uniform distribution on K. In particular, if X denotes the uniform distribution on K, we assert that $\|X/\pi_{f,T_0}\| \leq C$ for some constant $C > 0$. After $I = \sqrt{n}\log\left(\frac{T_0(n+1)}{\varepsilon\delta}\right)$ iterations of the algorithm, we are at the temperature $\frac{\varepsilon\delta}{n+1}$.

Note the following lemma:

Lemma 4.8. *Let $K \subset M$ be a strongly convex subset of a Riemannian manifold M with non-negative Ricci curvature. Let $g : K \to \mathbb{R}$ be a convex function with minimum vlaue zero. Let X be sampled from $\pi_{g,T}$. Then,*

$$E_{\pi_{g,t}}[g(X)] \leq T(n+1) + \min_{x\in K} f(x).$$

Applying this here, we have that $\mathbb{E}[f(z)] \leq \min_x f(x) + \varepsilon\delta$. By Markov's Inequality, it follows that

$$\mathbb{P}\left(f(z) - \min_x f(x) \geq \varepsilon\right) \leq \frac{\varepsilon\delta}{\varepsilon} = \delta.$$

So we are within ϵ of the minimum value with probability $1-\delta$.

Now, we analyze the runtime. In each phase i, we sample so that the approximate distribution is $\frac{\delta}{100I}$-close to the desired distribution. Then, adapting Theorem 4.3 to account for the new distribution, this requires

$$L_i = O\left(\frac{H^2 I^2 D^2 n^3 (1+R) L^2}{r^2 T_i^2 \delta^2} \log\left(\frac{HI}{\delta}\right)\right)$$

steps when we start from an H-warm start. In particular, we can choose the distribution so that $H = 2000I/\delta$ with probability $1 - \frac{\delta}{2000I}$. [$\|\pi_{g,T_i}/\pi_{g,T_{i+1}}\| \leq 5$] It follows that

$$L_i = O\left(\frac{I^4 D^2 n^3 (1+R) L^2}{r^2 T_i^2 \delta^4} \log\left(\frac{I}{\delta}\right)\right).$$

Finally, since $T_i \geq \frac{\varepsilon\delta}{2n}$ and $\|\pi_{f,T_i}/\pi_{f,T_{i+1}}\| \leq 5$, we achieve the result

$$O\left(\frac{D^2 n^{7.5}(R+1)L^2}{r^2\varepsilon^2\delta^6} \log\left(\frac{n}{\delta}\log\left(\frac{T_0(n+1)}{\varepsilon\delta}\right)\right)\log^5\left(\frac{T_0 n}{\varepsilon\delta}\right)\right)$$

which completes the proof. ■

4.3 Geodesic Gradient Descent

We breifly detail a geodesic gradient descent algorithm that can be used to solve optimistic likelihood and other related problems.

Algorithm 4.9 Projected Geodesic Gradient Descent

```
function GRADIENT DESCENT($\Sigma_0, S, L, T$)
    for $t < T$ do
        $G_t \leftarrow \nabla L(\Sigma_t)$, where $\nabla$ is the Riemannian Gradient operator.
        $\Sigma_{t+1/2} \leftarrow \exp_{\Sigma_t}(-\alpha_t G_t) = \Sigma_t^{1/2}\exp(-\alpha_t \Sigma_t^{-1/2} G_t \Sigma_t^{-1/2})\Sigma_t^{1/2}$
        Project $\Sigma_{t+1/2}$ back into the feasible space, using $P_S$ as a projection operator.
        $\Sigma_t \leftarrow P_S\left(\Sigma_{t+1/2}\right)$
    return $\Sigma_T$
```

In particular, the gradient for the optimistic likelihood problem is listed below - equipped with a closed form for the projection operator P_S and the fact that the Riemannian Gradient is bounded, Theorem 2.7 in [7] shows a sublinear convergence rate. It is also possible to prove the geodesic smoothness and geodesic strong convexity (analogous to similar concepts on twice-differentiable Euclidean functions) of L in the optimistic likelihood case, and these conditions can provide tighter bounds.

Lemma 4.10 (Riemmanian Gradient for the Optimistic Likelihood Problem)**.** *Using the loss L defined in Section* 3*,*

$$\nabla L(\Sigma) = 2(\Sigma - C),$$

where $C = -\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_i)(x_n - \mu_i)^\top$ *is an estimator of the covariance matrix from samples.*

Appendix A: Topology

Throughout this section, let X be a set. See [LeeSM] for more details.

Definition (Topological Space)**.** Let $\mathcal{T}$ be a collection of subsets of X. If

1. An arbitrary union of elements of $\mathcal{T}$ is an element of $\mathcal{T}$
2. A finite union of elements of $\mathcal{T}$ is an element of $\mathcal{T}$
3. $\varnothing, X \in \mathcal{T}$

then we can call each element of $\mathcal{T}$ an open set and $\mathcal{T}$ a topology. $(X, \mathcal{T})$ is then called a topological space. We will now omit the topology and call X a topological space.

Definition (Continuous Function)**.** A function $f : X \to Y$ is said to be continuous if for every open set, U, in Y, $f^{-1}(U)$ is an open set in X.

Definition (Hausdorff)**.** For every pair of distinct points $p, q \in X$, there are disjoint open subsets $U, V \subseteq X$ such that $p \in U$ and $q \in V$.

Definition (Basis)**.** A subset, $\mathcal{B} \subseteq \mathcal{T}$, is called a basis for $\mathcal{T}$ if for each $U \in \mathcal{T}$, we have that U is the union of some elements of $\mathcal{B}$.

Definition (Second Countable)**.** There exists a countable basis for the topology of X.

Definition (Homeomorphism)**.** A bijective continuous function with a continuous inverse.

Appendix B: Smooth Manifolds

See [LeeSM] for more details.

Definition (Topological Manifold)**.** Suppose M is a topological space. We say that M is a **manifold of dimension** n if it satisfies the following properties:

- M is a Hausdorff space.
- M is second-countable: there exists a countable basis for the topology of M.
- M is locally Euclidean of dimension n: each point of M has a neighborhood that is homeomorphic to an open subset of $\mathbb{R}^n$.

The most important property that we will consider is the local Euclidean property. More precisely, it means that for each $p \in M$, we have

- an open neighborhood $U \subseteq M$ containing p,
- an open neighborhood $V \subseteq \mathbb{R}^n$,
- a bijective function $\varphi : U \to V$ so that φ is continuous and φ^{-1} is continuous.

From now on, let M be a topological manifold of dimension n.

Definition (Coordinate Chart)**.** A pair, (U, φ), where U is an open subset of M, $\widehat{U}$ is an open subset of $\mathbb{R}^n$, and $\varphi : U \to \widehat{U}$ is a homeomorphism.

Definition (Transition Map)**.** If (U, φ) and (V, ψ) are charts on M such that $U \cap V \neq \varnothing$, we say that $\psi \circ \varphi^{-1} : \varphi(U \cap V) \to \psi(U \cap V)$ is the transition map from φ to ψ.

Definition (Smooth Compatibility of Transition Maps)**.** The (U, φ) and (V, ψ) of the above definition are said to be smoothly compatible if $\psi \circ \varphi^{-1}$ is smooth (i.e. infinitely differentiable in each coordinate because this is a map on $\mathbb{R}^n$). Note also that if $U \cap V = \varnothing$ then (U, φ) and (V, ψ) are said to be smoothly compatible.

Definition (Atlas)**.** A collection of charts such that the union of all of the open sets cover M. If all of the charts are smoothly compatible, then the atlas is said to be a smooth atlas. A smooth atlas is said to be maximal if it is not properly contained in another smooth atlas, and a maximal smooth atlas is called a smooth structure on M.

Definition (Smooth Manifold)**.** A topological manifold paired with a smooth structure.

Definition (Smooth Map)**.** We say that a map $f : M \to \mathbb{R}^k$ is smooth if at each $p \in M$, there is a chart (U, φ) with $p \in U$ such that $f \circ \varphi^{-1}$ is infinitely differentiable in each coordinate. The set of smooth maps from M to $\mathbb{R}^k$ is denoted $C^\infty(M; \mathbb{R}^k)$. One says that a map between two smooth manifolds, $f : M \to N$, is smooth if for every point $p \in M$, there are charts (U, φ) containing p and (V, ψ) containing $f(p)$ such that $\psi \circ f \circ \varphi^{-1}$ is smooth and $f(U) \subseteq V$.

Definition (Tangent Space)**.** The tangent space at $p \in M$ is defined to be the subset, T_pM, of $\{v : C^\infty(M) \to \mathbb{R}\}$ that is linear and satisfies the Leibniz rule:

$$v(fg) = f(p)v(g) + g(p)v(f).$$

Definition (Tangent Bundle)**.** The tangent bundle is the disjoint union

$$TM = \bigsqcup_{p \in M} T_pM$$

equipped with the map $\pi : (p, v) \mapsto p$.

Definition (Differential)**.** Given a smooth map $F : M \to N$, one defines the differential, $dF : TM \to TN$, at $p \in M$, by $dF_p(v)(f) = v(f \circ F)$ for $v \in T_pM$ and $f \in C^\infty(N)$.

Definition (Smooth Vector Field)**.** A smooth map, X, which assigns a point, p, on the manifold to a tangent vector, X_p, at that point. The set of smooth vector fields will be denoted $\mathrm{Vect}(M)$. Note that if $X \in \mathrm{Vect}(M)$ and $f \in C^\infty(M; \mathbb{R})$, then $(Xf)(p) = X_p(f)$.

Appendix C: Riemannian Geometry

Throughout this section, let M be a smooth manifold. See [LeeRM] for more details.

Definition (Metric Tensor)**.** A family of inner products, $\{g_p : T_pM \times T_pM \to \mathbb{R}\}_{p \in M}$.

Definition (Riemannian Manifold)**.** If g_p is a smooth function of p, then (M, g) is said to be a Riemannian manifold.

Definition (Connection on the Tangent Bundle)**.** A map $\nabla : \mathrm{Vect}(M) \times \mathrm{Vect}(M) \to \mathrm{Vect}(M) : (X, Y) \mapsto \nabla_X Y$ such that

- $\nabla_{fX+gY} Z = f\nabla_X Z + g\nabla_Y Z$ for $f, g \in C^\infty(M; \mathbb{R})$ and $X, Y, Z \in \mathrm{Vect}(M)$
- $\nabla_X(aY + bZ) = a\nabla_X Y + b\nabla_X Z$, for $a, b \in \mathbb{R}$ and $X, Y, Z \in \mathrm{Vect}(M)$
- $\nabla_X(fY) = f\nabla_X Y + (Xf)Y$, for $X, Y \in \mathrm{Vect}(M)$.

Definition (Levi-Civita Connection)**.** The unique connection on the tangent bundle that satisfies $\nabla_X Y - \nabla_Y X = [X, Y]$.

Definition (Geodesic)**.** A curve, γ, is said to be a geodesic with respect to a connection, ∇, if $\nabla_{\dot\gamma}\dot\gamma = 0$, where $\dot\gamma = \frac{d\gamma(t)}{dt}$.

Definition (Exponential Map)**.** Given a point and a direction, $(p, v) \in M \times T_pM$, it is a general fact that there is a unique geodesic that satisfies $\gamma(0) = p$ and $\gamma'(0) = v$, when M is sufficiently nice. One then defines the exponential map as $\exp_p(v) = \gamma(1)$.

Definition (Riemann curvature tensor)**.** The tensor R which satisfies

$$R(X, Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z$$

where $X, Y, Z \in \mathrm{Vect}(M)$.

Definition (Ricci Curvature Tensor)**.** The tensor Ric, which is defined as

$$\mathrm{Ric}(X, Y) = \mathrm{tr}(Z \mapsto R(Z, X)Y).$$

Definition (Sectional Curvature)**.** For $p \in M$ and linearly independent $v, w \in T_pM$, the sectional curvature of the plane that v and w span is

$$\mathrm{Sec}(v, w) = \frac{g_p(R(v, w)w, v)}{g_p(v, v)g_p(w, w) - g_p(v, w)^2}.$$

Definition (Gradient)**.** The vector field $\mathrm{grad}\, f$ which satisfies

$$df_p(v) = g_p(\mathrm{grad}\, f|_p, v).$$

5 References

[1] Geodesic Convex Optimization: Differentiation on Manifolds, Geodesics, and Convexity Nisheeth K. Vishnoi

[2] Sampling and Optimization on Convex Sets in Riemannian Manifolds of Non-Negative Curvature Navin Goyal Abhishek Shetty

[3] On Geodesically Convex Formulations for the Brascamp-Lieb Constant Nisheeth K. Vishnoi and Ozan Yıldız École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

[4] Zeyuan Allen-Zhu, Ankit Garg, Yuanzhi Li, Rafael Oliveira, Avi Wigderson, Operator Scaling via Geodesically Convex Optimization, Invariant Theory and Polynomial Identity Testing.

[5] Navin Goyal and Abhishek Shetty, Sampling and Optimization on Convex Sets in Riemannian Manifolds of Non-Negative Curvature

[6] Ankit Garg, Leonid Gurvits, Rafael Oliveira, Avi Wigderson, Algorithmic and optimization aspects of Brascamp-Lieb inequalities, via Operator Scaling

[7] Viet Anh Nguyen, Soroosh Shafieezadeh-Abadeh, Man-Chung Yue, Daniel Kuhn, Wolfram Wiesemann, Calculating Optimistic Likelihoods Using (Geodesically) Convex Optimization

[8] M. R. Bridson and A. Haefliger. Metric Spaces of Non-Positive Curvature. Springer, 2013.

[9] Teng Zhang, Robust subspace recovery by geodesically convex optimization

[LeeSM] John M. Lee, Introduction to Smooth Manifolds

[LeeRM] John M. Lee, Introduction to Riemannian Manifolds