J. Korean Math. Soc. **56** (2019), No. 4, pp. 1001–1016
https://doi.org/10.4134/JKMS.j180517
pISSN: 0304-9914 / eISSN: 2234-3008

GRADIENT PROJECTION METHODS FOR THE n-COUPLING PROBLEM

Sangho Kum and Sangwoon Yun

Abstract. We are concerned with optimization methods for the L^2-Wasserstein least squares problem of Gaussian measures (alternatively the n-coupling problem). Based on its equivalent form on the convex cone of positive definite matrices of fixed size and the strict convexity of the variance function, we are able to present an implementable (accelerated) gradient method for finding the unique minimizer. Its global convergence rate analysis is provided according to the derived upper bound of Lipschitz constants of the gradient function.

1. Introduction

For probability measures μ and ν on $\mathbb{R}^d$ with finite second moment, the Monge-Kantorovich problem is the minimization problem

$$\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathbb{R}^d\times\mathbb{R}^d}\|x-y\|^2 d\pi(x,y), \tag{1}$$

where $\Pi(\mu,\nu)$ denotes the set of transport plans between μ and ν, i.e., the set of probability measures on $\mathbb{R}^d\times\mathbb{R}^d$ with marginals μ and ν. The distance d_W, the square root of the minimum, on the set of probability measures with finite second moment is called the L^2-*Wasserstein metric*. A coupling achieving the infimum of (1) always exists and is unique when μ is absolutely continuous with respect to the Lebesgue measure. It is determined by $(\mathrm{id}\times T)_*\mu$, where T is Brenier's map between μ and ν, the gradient of a real valued convex function on $\mathbb{R}^d$ satisfying $T_*\mu=\nu$, where $T_*\mu$ denotes the push-forward of μ through T. See also [9, 13, 22].

Received July 31, 2018; Accepted August 29, 2018.
2010 *Mathematics Subject Classification.* Primary 49M30, 90C25.
Key words and phrases. L^2-Wasserstein least squares problem, n-coupling problem, Gaussian measure, positive definite matrix, Nesterov-Todd scaling, gradient projection method.
The work of the first author was supported by Basic Science Research Program through NRF Grant No.NRF-2017R1A2B1002008 and NRF-2014R1A1A2056635. The work of the second author was supported by Science Research Center Program through NRF funded by the Ministry of Science, ICT & Future Planning No.NRF-2016R1A5A1008055 and NRF-2016R1D1A1B03934371.

The multimarginal optimal transport problem reads as

$$\arg\min_{\gamma\in\Pi(\mu_1,\dots,\mu_n)} \int_{(\mathbb{R}^d)^n} \left[\sum_{j=1}^n w_j \| x_j - T(x) \|^2\right] d\gamma(x_1,\dots,x_n), \tag{2}$$

where μ_j's are probability measures with finite second moment and $\omega = (w_1, \dots, w_n)$ is a positive probability vector in $\mathbb{R}^n$, T is the arithmetic barycenter $T(x) = \sum_{j=1}^n w_j x_j$ and $\Pi(\mu_1, \dots, \mu_n)$ is the set of probability measures on $(\mathbb{R}^d)^n$ having $\mu_1, \dots, \mu_n$ as marginals. It is closely related to the least squares problem for the Wasserstein distance d_W; minimizing the averaged sum of squared Wasserstein distances

$$\arg\min_{\mu} \sum_{j=1}^n w_j d_W^2(\mu, \mu_j). \tag{3}$$

In [1], Agueh and Carlier established existence and uniqueness of the solution to (3) under absolutely continuity of some μ_j. This provides a natural notion of Wasserstein barycenter on the space of absolutely continuous probability measures. They also find a sufficient condition for the unique minimizer in terms of Brenier's maps and establish a precise relationship between the multimarginal optimal problem (2) and the least squares problem by $\mu = T_*\gamma$. Finding barycenters between more than three measures is more complicated but interestingly, the Wasserstein barycenter problem recently found natural applications in image processing and statistics [8, 10, 11, 26]. We refer to the books of Villani [29,30] for a modern account of optimal transportation theory.

This paper is concerned with the Wasserstein barycenter of Gaussian measures, which is alternatively known as the *n-coupling problem*, and studying an efficient optimization method for the coupling problem. It has attracted increasing attention for existence [14,27] but uniqueness issue remained [12, 15, 18, 25, 27], until the work of Agueh and Carlier.

For Gaussian measures μ and ν with zero mean and covariance matrices A and B respectively, the Wasserstein distance between μ and ν is explicitly given by Givens and Shortt [15]

$$d_W(A, B) = \sqrt{\mathrm{Tr}(A + B) - 2\mathrm{Tr}(A^{\frac{1}{2}} B A^{\frac{1}{2}})^{\frac{1}{2}}}.$$

The Wasserstein barycenter of Gaussian measures μ_j with zero mean and with positive definite covariance matrices A_j, $j = 1, \dots, n$ respectively, is determined by a positive definite solution of the following nonlinear matrix equation,

$$X = \frac{1}{n} \sum_{j=1}^n (X^{\frac{1}{2}} A_j X^{\frac{1}{2}})^{\frac{1}{2}}. \tag{4}$$

By Agueh and Carlier ([1, Theorem 6.1]), it has a unique positive definite solution for positive definite A_j's. The conjecture on the convergence of the fixed point iteration from (4) has recently been settled [2]. Their proofs in [1,2] depend heavily on nonsmooth analysis, convex duality and optimal transportation theory and do not provide a substantial convergence analysis with numerical methods. As far as a numerical issue of the fixed point iteration in [2] is concerned, a rigorous convergence analysis is not presented even though some numerical experiments are reported. So the iterative method still has a room for improvement from a theoretical point of view, that is, it deserves further research.

Bhatia, Jain and Lim [7] have recently established the strict convexity of the variance function and suggested some potential optimization methods based on their convexity result. They approach the n-coupling problem using only matrix analysis tools from its interpretation on the setting of the convex cone of positive definite matrices.

Motivated by the observations above, we propose a standard optimization method of computing the Wasserstein barycenter of Gaussian measures based on the gradient of the variance function and provide its solid convergence rate analysis via the strict convexity of the objective function on the convex cone of positive definite matrices [7]. This is the main purpose of the present paper. Indeed, we adapt a gradient projection method. Two types of classical gradient projection methods are presented depending on selecting stepsizes using Armijo rule. The first one uses Armijo rule along the feasible direction whereas the second one does Armijo rule along the projection arc. For the second, its global sublinear rate of convergence is well-known. The first method also seems to have global sublinear rate of convergence but a detailed proof has not been found yet. We provide a proof for the global sublinear convergence rate of the first. An accelerated gradient projection method adapted from [28, Algorithm 1] is also presented.

This paper is organized as follows. In Section 2, we briefly review about recent results on the n-coupling problem and the Wasserstein barycenter of Gaussian measures. Convexity of the variance function and its gradient function are explicitly addressed in Section 3 together with a potential use of the gradient-based optimization method. We show in Section 4 the Lipschitz continuity of the gradient function and find an upper bound of the Lipschitz constants. In Section 5, we describe gradient projection methods and analyze its convergence properties. In particular, the global sublinear convergence rate is verified. In Section 6, we briefly describe an accelerated gradient projection method. We report numerical results for finding the Wasserstein barycenter of Gaussian measures on randomly generated matrices with proposed gradient projection methods in Section 7. Concluding remarks are included in Section 8.

2. The n-coupling problem

The realization that the Wasserstein metric can be taken as a reasonable distance between probability distributions first appeared in the paper of Kantorovich and Rubinstein [17]. Although explicit calculation is very difficult for most concrete examples, a successful computation of the Wasserstein metric of Gaussian measures was done by Olkin-Pukelsheim [25], Dowson-Landau [12], Givens-Shortt [15] and Knott and Smith [18]. Identifying a positive definite matrix A with the Gaussian measure with zero mean and with covariance matrix A, the Wasserstein metric between two positive definite matrices is determined explicitly by Givens and Shortt [15]

$$d_W(A,B) = \sqrt{\mathrm{Tr}(A+B) - 2\mathrm{Tr}(A^{\frac{1}{2}}BA^{\frac{1}{2}})^{\frac{1}{2}}}. \tag{5}$$

Definition. Let $\omega = (w_1, \ldots, w_n)$ be a positive probability vector and let $\mathbb{A} = (A_1, \ldots, A_n) \in \mathbb{P}^n$. The ω-weighted Wasserstein barycenter of $\mathbb{A}$ is defined by

$$\Omega(\omega; \mathbb{A}) := \arg\min_{X>0} \sum_{j=1}^{n} w_j d_W^2(X, A_j). \tag{6}$$

By Agueh and Carlier ([1, Theorem 6.1]), it has a unique minimizer. In fact they established the existence and uniqueness of Wasserstein barycenter for absolutely continuous measures by using tools of nonsmooth analysis, convex duality and optimal transportation theory.

The objective function in (6) is $f : \mathbb{P} \to \mathbb{R}$, where

$$f(X) = \sum_{j=1}^{n} w_j \ \mathrm{tr}A_j + \sum_{j=1}^{n} w_j \ \mathrm{tr}\left(X - 2(A_j^{\frac{1}{2}} X A_j^{\frac{1}{2}})^{\frac{1}{2}}\right). \tag{7}$$

In [7], Bahtia-Jain and Lim derived new and important results for the objective function f, the *strict convexity* which follows basically the operator concavity of the square root $X \longmapsto X^{\frac{1}{2}}$, and the formula for its gradient in terms of the matrix geometric mean $A\#B := A^{1/2}(A^{-1/2}BA^{-1/2})^{1/2}A^{1/2}$:

$$\begin{aligned} \nabla f(X) &= \sum_{j=1}^{n} w_j \left(I - A_j \# X^{-1}\right) \\ &= I - \sum_{j=1}^{n} w_j (A_j \# X^{-1}). \end{aligned}$$

(It follows by showing $\nabla g(X) = \frac{1}{2}(A^2 \# X^{-1})$ for the map $g(X) = (AXA)^{\frac{1}{2}}$.) The uniqueness issue of (6) then turns equivalently into the existence of positive definite solution of

$$\nabla f(X) = 0 \tag{8}$$

from its strict convexity. That is, the least squares problem (6) has a unique solution if and only if the following nonlinear matrix equation has a positive definite solution

$$I = \sum_{j=1}^{n} w_j (A_j \# X^{-1}). \tag{9}$$

Existence of positive definite solution of

$$X = \sum_{j=1}^{n} w_j (X^{\frac{1}{2}} A_j X^{\frac{1}{2}})^{\frac{1}{2}} \tag{10}$$

follows from Brouwer's fixed point theorem [1]. The continuous map $X \mapsto \sum_{j=1}^{n} w_j (X^{\frac{1}{2}} A_j X^{\frac{1}{2}})^{\frac{1}{2}}$ is a self map on the Löwner order interval $[\alpha_* I, \beta_* I] := \{X : \alpha_* I \le X \le \beta_* I\}$, where

$$\alpha_* := \left[\sum_{j=1}^{n} w_j \sqrt{\lambda_{\min}(A_j)}\right]^2, \quad \beta_* := \left[\sum_{j=1}^{n} w_j \sqrt{\lambda_{\max}(A_j)}\right]^2$$

and $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalue of A, respectively. Here $X \le Y$ means that $Y - X$ is positive semidefinite. Since the order interval is compact, (8) has a positive definite solution by Brouwer's fixed point theorem.

Although the uniqueness issue of (6) and also (8) is nicely and completely settled by Agueh-Carlier [1] and Bhatia-Jain-Lim [7] in different directions, the computational issue for the minimizer is still problematic and open. An iterative method based on the fixed point equation (8) is recently proved in [2, Theorem 4.2]: $\lim_{k\to\infty} S_k = \Omega(\omega; \mathbb{A})$, where

$$S_{k+1} \quad = \quad S_k^{-\frac{1}{2}} \left[\sum_{j=1}^{n} w_j \ (S_k^{\frac{1}{2}} A_j S_k^{\frac{1}{2}})^{\frac{1}{2}}\right]^2 S_k^{-\frac{1}{2}}, \quad S_0 \in \mathbb{P}.$$

See also [7] for a simplified proof using only matrix analytic techniques. As mentioned in the introduction, this iterative method needs a theoretical strengthening with a proper convergence analysis.

We close this section with an application of the strict convexity of f and its uniqueness of minimizer. This implies the uniqueness of the critical point of f (8), that is, the equation (9) (and also (10)) has a *unique* positive definite solution, the Wasserstein barycenter $\Omega(\omega; \mathbb{A})$. We consider the continuous map $\Gamma : \mathbb{P} \to \mathbb{P}$ from (9)

$$\Gamma(X) = \sum_{j=1}^{n} w_j (A_j \# X^{-1}).$$

Using the Wasserstein barycenter, Γ is a bijection with

$$\Gamma^{-1}(X) = X^{\frac{1}{2}} \Omega(w; X^{-\frac{1}{2}} A_1 X^{-\frac{1}{2}}, \ldots, X^{-\frac{1}{2}} A_n X^{-\frac{1}{2}}) X^{\frac{1}{2}}.$$

Indeed,

$$\begin{aligned} X = \Gamma(Y) &\Longleftrightarrow I = \sum_{j=1}^{n} w_j X^{-\frac{1}{2}}(A_j \# Y^{-1})X^{-\frac{1}{2}} \\ &\Longleftrightarrow I = \sum_{j=1}^{n} w_j (X^{-\frac{1}{2}} A_j X^{-\frac{1}{2}}) \# (X^{-\frac{1}{2}} Y^{-1} X^{-\frac{1}{2}}) \\ &\Longleftrightarrow X^{\frac{1}{2}} Y X^{\frac{1}{2}} = \Omega(w; X^{-\frac{1}{2}} A_1 X^{-\frac{1}{2}}, \ldots, X^{-\frac{1}{2}} A_n X^{-\frac{1}{2}}). \end{aligned}$$

In particular, for a fixed positive definite matrix B, the following nonlinear matrix equation

$$B = \sum_{j=1}^{n} w_j (A_j \# X^{-1}) \tag{11}$$

has the unique positive definite solution $X = \Gamma^{-1}(B)$.

3. Lipschitz continuity

In this section we show the Lipschitz continuity of the gradient function ∇f and find an upper bound of the Lipschitz constants.

Let $\mathbb{H}$ be the Euclidean space of $d \times d$ Hermitian matrices equipped with the inner product $\langle X, Y \rangle := \mathrm{Tr}(XY)$. The *Frobenius norm* $\| \cdot \|_F$ is defined by $\| X \|_F = (\mathrm{tr} X^2)^{1/2}$.

Let $A_1, \ldots, A_n$ be elements of $\mathbb{P}$ and let $w = (w_1, \ldots, w_n)$ be a weight vector; i.e., $w_j > 0$ and $\sum_{j=1}^{n} w_j = 1$. We have seen in the previous section that $\nabla f(X) = I - \sum_{j=1}^{n} w_j (A_j \# X^{-1})$ and $\Omega(\omega; \mathbb{A}) \in [\alpha_* I, \beta_* I]$. Now, we set

$$L_{\min} = \min_{1 \le j \le n} \{\lambda_{\min}(A_j)\}, \quad L_{\max} = \max_{1 \le j \le n} \{\lambda_{\max}(A_j)\}.$$

Now, we establish the following theorem for the Lipschitz continuity of the gradient function.

Theorem 3.1. *For $\alpha_* I \le X, Y \le \beta_* I$ with $X \ne Y$,*

$$\frac{\| \nabla f(X) - \nabla f(Y) \|_F}{\| X - Y \|_F} \le \frac{L_{\max}^2}{2\alpha_*^{\frac{3}{2}} L_{\min}^{\frac{3}{2}}} \le \frac{L_{\max}^2}{2L_{\min}^3}.$$

Proof. We first consider the matrix function $H(X) := X^{-1/2}$ on a Loewner interval $[\alpha I, \beta I]$ with $\alpha > 0$. It is the composition $H = (-F) \circ G$, where

$$\begin{aligned} G &: [\alpha I, \beta I] \to [\sqrt{\alpha} I, \sqrt{\beta} I], & G(X) &= X^{1/2}, \\ F &: [\sqrt{\alpha} I, \sqrt{\beta} I] \to [-\frac{1}{\sqrt{\alpha}} I, -\frac{1}{\sqrt{\beta}} I], & F(X) &= -X^{-1}. \end{aligned}$$

We note that F and G are operator monotone functions on $\mathbb{P}$. According to Theorem X.3.8 of [6]

$$\| G(X) - G(Y) \|_F \le \frac{1}{2\sqrt{\alpha}} \| X - Y \|_F, \quad X, Y \in [\alpha I, \beta I],$$

$$\| F(X) - F(Y) \|_F \leq \frac{1}{\alpha} \| X - Y \|_F, \quad X, Y \in [\sqrt{\alpha} I, \sqrt{\beta} I].$$

Then for $X, Y \in [\alpha I, \beta I]$,

$$\begin{aligned} \| H(X) - H(Y) \|_F = \| F(G(X)) - F(G(Y)) \|_F &\leq \frac{1}{\alpha} \| G(X) - G(Y) \|_F \\ &\leq \frac{1}{\alpha} \frac{1}{2\sqrt{\alpha}} \| X - Y \|_F = \frac{1}{2\alpha^{\frac{3}{2}}} \| X - Y \|_F. \end{aligned}$$

Next, let $X, Y \in [\alpha_* I, \beta_* I]$. Then for all $j = 1, \ldots, n$, $\alpha_* L_{\min} I \leq A_j^{1/2} X A_j^{1/2} \leq \beta_* L_{\max} I$. By the preceding step,

$$\begin{aligned} \| \nabla f(X) - \nabla f(Y) \|_F &\leq \sum_{j=1}^n w_j \| A_j \# X^{-1} - A_j \# Y^{-1} \|_F \\ &= \sum_{j=1}^n w_j \left\| A_j^{\frac{1}{2}} \left[\left(A_j^{\frac{1}{2}} X A_j^{\frac{1}{2}} \right)^{-\frac{1}{2}} - \left(A_j^{\frac{1}{2}} Y A_j^{\frac{1}{2}} \right)^{-\frac{1}{2}} \right] A_j^{\frac{1}{2}} \right\|_F \\ &\leq \sum_{j=1}^n w_j \left\| A_j^{\frac{1}{2}} \right\|^2 \cdot \left\| \left(A_j^{\frac{1}{2}} X A_j^{\frac{1}{2}} \right)^{-\frac{1}{2}} - \left(A_j^{\frac{1}{2}} Y A_j^{\frac{1}{2}} \right)^{-\frac{1}{2}} \right\|_F \\ &\leq L_{\max} \sum_{j=1}^n w_j \left\| \left(A_j^{\frac{1}{2}} X A_j^{\frac{1}{2}} \right)^{-\frac{1}{2}} - \left(A_j^{\frac{1}{2}} Y A_j^{\frac{1}{2}} \right)^{-\frac{1}{2}} \right\|_F \\ &\leq \frac{L_{\max}}{2\alpha_*^{\frac{3}{2}} L_{\min}^{\frac{3}{2}}} \sum_{j=1}^n w_j \left\| A_j^{\frac{1}{2}} X A_j^{\frac{1}{2}} - A_j^{\frac{1}{2}} Y A_j^{\frac{1}{2}} \right\|_F \\ &= \frac{L_{\max}}{2\alpha_*^{\frac{3}{2}} L_{\min}^{\frac{3}{2}}} \sum_{j=1}^n w_j \left\| A_j^{\frac{1}{2}} (X - Y) A_j^{\frac{1}{2}} \right\|_F \\ &\leq \frac{L_{\max}^2}{2\alpha_*^{\frac{3}{2}} L_{\min}^{\frac{3}{2}}} \| X - Y \|_F. \end{aligned}$$

This completes the proof. □

If $A_j \in [\alpha I, \beta I]$ for all $j = 1, \ldots, n$, then

$$\| \nabla f(X) - \nabla f(Y) \|_F \leq \frac{\beta^2}{2\alpha^3} \| X - Y \|_F.$$

4. Gradient projection methods

In this section, we describe gradient projection methods for solving (6) and analyze their convergence properties. In particular, we provide a proof of the global sublinear convergence rate of the method using the Lipschitz constant of the gradient function obtained from Section 4. Gradient projection methods can be classified into two types according to stepsize selection.

First, we formally describe the algorithmic procedure for the gradient projection method (GPM) below.

Algorithm 1 GPM

Choose $X^0 \in \mathcal{D}$. Initialize $k = 0$. Update $X^{(k+1)}$ from $X^{(k)}$ by the following template:

Step 1.: Find $\bar{X}^{(k)} = [X^{(k)} - s^{(k)}\nabla f(X^{(k)})]^+$,
Step 2.: Select a stepsize $t^{(k)}$,
Step 3.: $X^{(k+1)} = X^{(k)} + t^{(k)}(\bar{X}^{(k)} - X^{(k)})$.

Here $[\cdot]^+$ denotes the projection on the set $\mathcal{D} := [\alpha_* I, \beta_* I]$.

GPM can be classified into two versions [4] depending on how Armijo rule is applied to select stepsizes. The first version is to use Armijo rule along the feasible direction $\bar{X}^{(k)} - X^{(k)}$ where the feasible vector $\bar{X}^k$ is obtained by taking a step $-s^{(k)}\nabla f(X^{(k)})$ along the negative gradient and projecting the result $X^{(k)} - s^{(k)}\nabla f(X^{(k)})$ on $\mathcal{D}$. In this case, we set $s^{(k)} = 1$ for all k and $t^{(k)}$ is chosen by the Armijo rule over the interval $[0, 1]$.

Choose $t^{(k)}_{\text{init}} > 0$ and let $t^{(k)}$ be the largest element of $\{t^{(k)}_{\text{init}}\xi^j\}_{j=0,1,\ldots}$ satisfying

$$f(X^{(k)} + t^{(k)}D^{(k)}) \le f(X^{(k)}) - \sigma t^{(k)}\langle \nabla f(X^{(k)}),\, D^{(k)}\rangle, \tag{12}$$

where $0 < \xi < 1$, $0 < \sigma < 1$, and $D^{(k)} = \bar{X}^{(k)} - X^{(k)}$.

The second version is to use Armijo rule along the projection arc. In this case, the stepsize $t^{(k)} = 1$ for all k and the stepsize $s^{(k)}$ is chosen by the scheme similar to Armijo rule. In other words, $X^{(k+1)}$ is determined by an Armijo-like rule on the projection arc $\{X^{(k)}(s) \mid s > 0\}$, where for all $s > 0$, $X^{(k)}(s)$ is defined by

$$X^{(k)}(s) = [X^{(k)} - s\nabla f(X^{(k)})]^+.$$

Choose $s^{(k)}_{\text{init}} > 0$ and let $s^{(k)}$ be the largest element of $\{s^{(k)}_{\text{init}}\xi^j\}_{j=0,1,\ldots}$ satisfying

$$f(X^{(k)}(s^{(k)})) \le f(X^{(k)}) - \sigma\langle \nabla f(X^{(k)}),\, X^{(k)}(s^{(k)}) - X^{(k)}\rangle, \tag{13}$$

where $0 < \xi < 1$ and $0 < \sigma < 1$.

Note that the projection of the matrix $S \in \mathcal{S}^d$, where $\mathcal{S}^d$ is the set of $d \times d$ symmetric matrices, onto the set $\mathcal{D}$ is to find the solution of the following minimization problem

$$\min_{X \in \mathcal{D}} \| X - S \|_F.$$

The solution of the above problem is

$$[S]^+ = U\text{Diag}(\min(\max(\alpha_*, \lambda_1), \beta^*), \ldots, \min(\max(\alpha_*, \lambda_d), \beta_*)U^T,$$

where $\lambda_1 \geq \cdots \geq \lambda_d$ are the eigenvalues of S and U is a corresponding orthogonal matrix of eigenvalues of S. This result can be found in [21]. It requires an orthogonal diagonalization and so it can be expensive if d is large, say $d \geq 100$. Since GPM with (12) requires one projection per iteration, it can be more efficient than GPM with (13) when d is large. Hence, in our implementation of the gradient projection method, Armijo rule along the feasible direction is used.

In the following, we establish the global convergence of GPM. For the proof, we refer to [4, Propositions 2.3.1 & 2.3.3].

Theorem 4.1. *Let $\{X^{(k)}\}$ be the sequence generated by GPM with $s^{(k)} = 1$ for all k and with $t^{(k)}$ chosen by Armijo rule along the feasible direction or with $t^{(k)} = 1$ for all k and with $s^{(k)}$ chosen by Armijo rule along the projection arc. Then every limit point of $\{X^{(k)}\}$ is stationary.*

By Theorem 3.1, f has Lipschitz continuous gradient.

$$\| \nabla f(Y) - \nabla f(Z) \| \leq L \| Y - Z \| \quad \forall Y, Z \in \mathcal{D} \tag{14}$$

for some $L \geq 0$. In this case, we can use a constant stepsize for GPM. That is, $s^{(k)}$ is fixed at some constant, say 1, for all k and $t^{(k)} = \frac{1}{L}$, then we have

$$X^{(k+1)} = X^{(k)} + \frac{1}{L}(\bar{X}^{(k)} - X^{(k)}). \tag{15}$$

Or, for all k, $s^{(k)} = \frac{1}{L}$ and $t^{(k)}$ is fixed at some constant, say 1, we get

$$X^{(k+1)} = [X^{(k)} - \frac{1}{L}\nabla f(X^{(k)})]^+. \tag{16}$$

It is well-known that the above scheme (16) achieves a sublinear rate of convergence. In other words, it can be shown that $f(X^{(k)}) - \inf f \leq O\left(\frac{L}{k}\right)$. To the best of our knowledge, GPM obtained by (15) has also the global sublinear rate of convergence. However, this result has not been proved in detail. Now, we establish an upper bound on the number of iterations for GPM obtained by (15) to achieve ϵ-optimality in the following theorem. In other words, the convergence rate is sublinear.

Theorem 4.2. *Let $\{X^{(k)}\}$ be the sequence generated by GPM with $s^{(k)} = 1$ for all k and with $t^{(k)} = \frac{1}{L}$ and X^* be the optimal solution. Then $f(X^{(k)}) - f(X^*) \leq \epsilon$ if*

$$k \geq \left\lceil \frac{32Ld(\beta_*)^2}{\epsilon} \right\rceil.$$

Proof. Since the set $\mathcal{D}$ is convex and

$$\bar{X}^{(k)} \in \arg\min_{X \in \mathcal{D}} \langle \nabla f(X^{(k)}), X \rangle + \frac{1}{2} \left\| X - X^{(k)} \right\|^2, \tag{17}$$

we have that, for $\alpha \in (0, 1)$,

$$\langle \nabla f(X^{(k)}), \bar{X}^{(k)} \rangle + \frac{1}{2} \left\| \bar{X}^{(k)} - X^{(k)} \right\|^2$$

$$\leq \langle \nabla f(X^{(k)}),\, X^{(k)} + \alpha(\bar{X}^{(k)} - X^{(k)})\rangle + \frac{\alpha^2}{2}\left\| \bar{X}^{(k)} - X^{(k)} \right\|^2.$$

Rearranging terms yields

$$(1-\alpha)\langle \nabla f(X^{(k)}),\, D^{(k)}\rangle \leq -(1-\alpha^2)\frac{1}{2}\left\| D^{(k)} \right\|^2, \qquad (D^{(k)} := \bar{X}^{(k)} - X^{(k)}).$$

Dividing both sides by $1-\alpha > 0$ and then taking $\alpha \uparrow 1$ implies that

$$\langle \nabla f(X^{(k)}),\, D^{(k)}\rangle \leq -\left\| D^{(k)} \right\|^2. \tag{18}$$

We have from the Cauchy-Schwarz inequality that

(19)

$$\begin{aligned}
&f(X^{(k+1)}) - f(X^{(k)}) \\
&= \langle \nabla f(X^{(k)}),\, X^{(k+1)} - X^{(k)}\rangle \\
&\quad + \int_0^1 \langle \nabla f(X^{(k)} + t(X^{(k+1)} - X^{(k)})) - \nabla f(X^{(k)}),\, X^{(k+1)} - X^{(k)}\rangle dt \\
&\leq \langle \nabla f(X^{(k)}),\, X^{(k+1)} - X^{(k)}\rangle \\
&\quad + \int_0^1 \left\| \nabla f(X^{(k)} + t(X^{(k+1)} - X^{(k)})) - \nabla f(X^{(k)}) \right\| \cdot \left\| X^{(k+1)} - X^{(k)} \right\| dt \\
&\leq \langle \nabla f(X^{(k)}),\, X^{(k+1)} - X^{(k)}\rangle + \frac{L}{2}\left\| X^{(k+1)} - X^{(k)} \right\|^2 \\
&= \frac{1}{L}\langle \nabla f(X^{(k)}),\, D^{(k)}\rangle + \frac{1}{2L}\left\| D^{(k)} \right\|^2 \\
&\leq \frac{1}{2L}\langle \nabla f(X^{(k)}),\, D^{(k)}\rangle,
\end{aligned}$$

where the second inequality uses the Lipschitz continuity of the gradient and the third inequality uses (18).

By Fermat's rule applied for (17) we have that for any $X^{(k)} \in \mathcal{D}$,

$$\bar{X}^{(k)} \in \operatorname*{arg\,min}_{X \in \mathcal{D}} \langle \nabla f(X^{(k)}) + \bar{X}^{(k)} - X^{(k)},\, X\rangle. \tag{20}$$

Then we have

$$\langle \nabla f(X^{(k)}) + \bar{X}^{(k)} - X^{(k)},\, \bar{X}^{(k)}\rangle \leq \langle \nabla f(X^{(k)}) + \bar{X}^{(k)} - X^{(k)},\, X^*\rangle.$$

Subtracting $\langle \nabla f(X^{(k)}) + \bar{X}^{(k)} - X^{(k)},\, X^{(k)}\rangle$ from both sides of the above inequality implies that

$$\begin{aligned}
\langle \nabla f(X^{(k)}) + D^{(k)},\, D^{(k)}\rangle &\leq \langle \nabla f(X^{(k)}) + D^{(k)},\, X^* - X^{(k)}\rangle \\
&\leq f(X^*) - f(X^{(k)}) + \langle D^{(k)},\, X^* - X^{(k)}\rangle,
\end{aligned}$$

where the second inequality uses the convexity of f. Thus

$$f(X^{(k)}) - f(X^*) \leq \langle D^{(k)},\, X^* - X^{(k)}\rangle - \langle \nabla f(X^{(k)}) + D^{(k)},\, D^{(k)}\rangle$$

$$\begin{aligned}
&\leq \left\| D^{(k)} \right\| \left\| X^* - X^{(k)} \right\| - \langle \nabla f(X^{(k)}),\, D^{(k)} \rangle \\
&\leq \left\| D^{(k)} \right\| \left(\| X^* \| + \left\| X^{(k)} \right\| \right) - \langle \nabla f(X^{(k)}),\, D^{(k)} \rangle \\
&\leq 2\sqrt{d}\beta_* \left\| D^{(k)} \right\| - \langle \nabla f(X^{(k)}),\, D^{(k)} \rangle \\
&\leq 2\sqrt{d}\beta_* \sqrt{-\langle \nabla f(X^{(k)}),\, D^k \rangle} - \langle \nabla f(X^{(k)}),\, D^{(k)} \rangle, \qquad (21)
\end{aligned}$$

where the fourth inequality is obtained from the fact that $X^{(k)}$ and X^* are in $\mathcal{D}$ and the fifth inequality uses (18). Combining (21) with (19) yields

$$e^{(k)} \leq 2\sqrt{2dL}\beta_* \sqrt{e^{(k)} - e^{(k+1)}} + 2L(e^{(k)} - e^{(k+1)}), \qquad (22)$$

where $e^{(k)} = f(X^{(k)}) - f(X^*)$.

There are two cases that the difference between the objective values with consecutive iterates takes. The first case is $e^{(k)} - e^{(k+1)} \geq \frac{2d(\beta_*)^2}{L}$ and the second case is $e^{(k)} - e^{(k+1)} < \frac{2d(\beta_*)^2}{L}$. For the first case, the reduction of the objective function value at the iteration k is at least $\frac{2d(\beta_*)^2}{L}$. But the reduction is less than $\frac{2d(\beta_*)^2}{L}$ for the second case. The second case is the worst one and so we only consider this case.

From (22), we have $e^{(k)} \leq 4\sqrt{2dL}\beta_* \sqrt{e^{(k)} - e^{(k+1)}}$, and rearranging terms yields

$$e^{(k+1)} \leq e^{(k)} - \frac{1}{32dL} \left(\frac{e^{(k)}}{\beta_*} \right)^2. \qquad (23)$$

We may assume $e^{(k)} > 0$ for all $k \geq 0$ (otherwise, $e^{(k)} \leq \epsilon$) in (23). Then we consider the reciprocals $\xi^{(k)} = 1/e^{(k)}$. By (23) and $e^{(k)} > 0$, we have

$$0 \leq C_1 e^{(k)} < 1,$$

where $C_1 = \frac{1}{32dL(\beta_*)^2}$.

Thus (23) yields

$$\xi^{(k+1)} - \xi^{(k)} \geq \frac{1}{e^{(k)}(1 - C_1 e^{(k)})} - \frac{1}{e^{(k)}} = \frac{C_1}{1 - C_1 e^{(k)}} \geq C_1.$$

This implies that if $e^{(k)} - e^{(k+1)} \geq \frac{2d(\beta_*)^2}{L}$ for all $k \geq 0$, then

$$\xi^{(k)} = \xi^{(0)} + \sum_{i=0}^{k-1} (\xi^{(i+1)} - \xi^{(i)}) \geq C_1 k$$

and consequently

$$e^{(k)} = \frac{1}{\xi^{(k)}} \leq \frac{1}{C_1 k}.$$

It follows that $e^{(k)} \leq \epsilon$ if

$$k \geq \left\lceil \frac{32dL(\beta_*)^2}{\epsilon} \right\rceil. \qquad \square$$

From Theorem 4.2, the complexity bound on the number of iterations for achieving ϵ-optimality can be $O\left(\frac{L}{k}\right)$. In other words, $f(X^{(k)}) - \inf f \leq O\left(\frac{L}{k}\right)$.

5. Accelerated gradient projection methods

In this section, we briefly describe an accelerated gradient projection method based on [16, 24, 28] for solving (6) and provide its faster global sublinear convergence rate using the Lipschitz constant of the gradient function obtained from Section 4.

There are several versions of accelerated methods; see [16, 24, 28] and references therein. Since the projection can be expensive when the matrix dimension is large, we adapt Algorithm 1 in [28] with $h(\cdot) = \frac{1}{2}\|\cdot\|_F^2$ and $x^{k+1} = \hat{x}^{k+1}$, that requires only one projection. The algorithmic procedure for the accelerated gradient projection method (AGPM) is formally given below.

Algorithm 2 AGPM

Choose $X^{(0)}$, $Z^{(0)} \in \mathcal{D}$, and $t^{(0)} \in (0, 1]$. Initialize $k = 0$. Update $X^{(k+1)}$ and $Z^{(k+1)}$ from $X^{(k)}$ and $Z^{(k)}$ by the following template:

Step 1.: Find $Z^{(k+1)} = \left[Z^{(k)} - \frac{1}{t^{(k)}L}\nabla f\left(X^{(k)} + t^{(k)}(Z^{(k)} - X^{(k)})\right)\right]^+$,

Step 2.: $X^{(k+1)} = (1 - t^{(k)})X^{(k)} + t^{(k)}Z^{(k+1)}$.

Step 3.: $t^{(k+1)} = \frac{\sqrt{(t^{(k)})^4 + 4(t^{(k)})^2} - (t^{(k)})^2}{2}$.

In the next theorem, we provide the iteration complexity for AGPM. Its proof can be easily induced from [28, Proposition 1].

Theorem 5.1. *Let $\{X^{(k)}\}$, $\{Z^{(k)}\}$, and $\{t^{(k)}\}$ be the sequences generated by AGPM with $t^{(0)} = 1$ and X^* be the optimal solution. Then*

$$f(X^{(k)}) - f(X^*) \leq 2L(t^{(k-1)})^2 \left\| Z^{(0)} - X^* \right\|^2.$$

Note that the stepsize $t^{(k)}$ with $t^{(0)} = 1$ satisfies $t^{(k)} \leq \frac{2}{k+2}$ [28] and so

$$f(X^{(k)}) - f(X^*) \leq \epsilon \quad \text{whenever } k \geq \sqrt{\frac{2L\left\| Z^{(0)} - X^* \right\|}{\epsilon}} - 1.$$

Thus it is shown that $f(X^{(k)}) - \inf f \leq O\left(\frac{L}{k^2}\right)$.

6. Numerical results

We report the performance of GPM with Armijo rule (GPM-A), GPM with constant stepsize (GPM-C), and AGPM on n randomly generated matrices of the size $d \times d$. The random matrices we use for our test are generated by MATLAB code as follows:

```
for  i = 1 : n
     [Q, ] = qr(randn(d));
     a{i} = Q * diag(eiglb + eigub * rand(d, 1)) * Q';
```

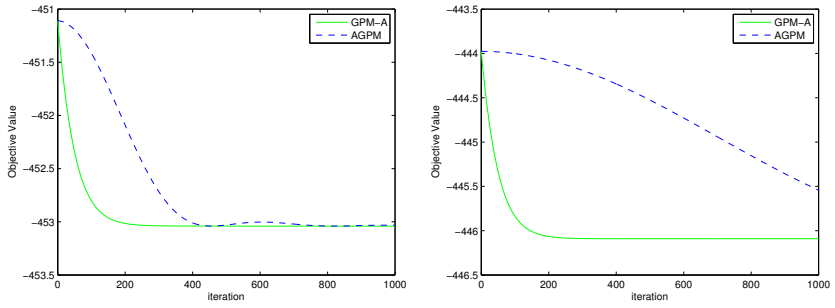
TABLE 1. Test results of the final objective values and CPU time in seconds for three methods GPM-A, GPM-C, and AGPM on 5 random data sets.

		eiglb = 1, eigub = 99			eiglb = 0.1, eigub = 99.9		
		GPM-A	GPM-C	AGPM	GPM-A	GPM-C	AGPM
1	obj	-455.1073	-453.3906	-455.1008	-447.6930	-445.5805	-446.5237
	CPU	387.5469	131.6250	130.3594	392.0469	129.3750	132.2188
2	obj	-453.0401	-451.3390	-453.0314	-446.0910	-443.9959	-445.5426
	CPU	390.4375	137.0469	131.0469	394.0938	133.0000	129.6250
3	obj	-436.2434	-433.5503	-436.2426	-428.7903	-425.5745	-426.6294
	CPU	393.9375	129.2031	129.6406	381.9063	129.1406	131.7188
4	obj	-474.9703	-474.0734	-474.9663	-468.0361	466.9026	-467.8266
	CPU	394.7969	131.0469	132.8281	386.6250	129.2031	131.4844
5	obj	-457.9968	-456.5869	-457.9919	-450.8314	-449.0858	-449.6617
	CPU	389.9219	130.1094	133.0000	388.7188	132.7031	130.5781

Note that the eigenvalues of generated matrices are randomly distributed in the interval $[\text{eiglb}, \text{eiglb} + \text{eigub}]$. In our experiments, we set $n = 100$ and $d = 10$. To observe the effect of the Lipschitz constant, we set $\text{eiglb} = 1, \text{eigub} = 99$ and $\text{eiglb} = 0.1, \text{eigub} = 99.9$. Hence, the Lipschitz constant of the latter case is roughly 1000 times that of the former one. Note that we estimate the Lipschitz constant by using the tight bound in Theorem 3.1. We terminate all the algorithms when the number of iterations reaches 1000.

All runs are performed on a Laptop with Intel Core i7-3537U CPU (2.00GHz) and 8GB Memory, running 64-bit windows 10 and MATLAB (Version 8.3). Throughout the experiments, we choose the initial iterate to be $X^{(0)} = 0.5(L_{\min} + L_{\max})I$ for all algorithms and $Z^{(0)} = X^{(0)}$, $t^{(0)} = 1$ for AGPM. And we set $\xi = 0.5$ and $\sigma = 0.1$ for GPM-A. We report in Table 1 our numerical results, showing final objective value(obj) and total time(CPU) in seconds.

From Table 1, the objective value of APGM is less than that of GPM-C. This supports that APGM has faster convergence rate than GPM-C. The objective value of GPM-A is less than that of GPM-C and AGPM since GPM-A can take larger stepsizes and so the objective values of GPM-A are reduced faster than the other algorithms. But GPM-A takes more CPU time than GPM-C and AGPM since it requires function evaluations at each iteration. In Table 1, the estimated Lipschitz constant for the test matrices with eiglb = 0.1 and eigub = 99.9 is more than 1000 times bigger than that for the test matrices with eiglb = 1 and eigub = 99. Then it is observed that the estimated Lipschitz constant is smaller, the gap between the objective values of AGPM and that of GPM-A is bigger. This is also shown in Figure 1.



(a) (b)

FIGURE 1. (a) Objective value versus iteration with eiglb = 1, eigub = 99. (b) Objective value versus iteration with eiglb = 0.1, eigub = 99.9. The estimated Lipschitz constant in (b) is more than 10 times bigger than that in (a). It is observed that the estimated Lipschitz constant is smaller, the gap between the objective value of AGPM and that of GPM-A is bigger.

7. Concluding remarks and future work

In this paper, we have proposed standard optimization methods, i.e., the gradient projection method and its accelerated version, for computing the Wasserstein barycenter of Gaussian measures and analyzed their convergence properties. By using the estimation of the Lipschitz constant for the gradient, we established the global sublinear rate of convergence of GPM that uses the projection to find the feasible direction. When the estimated Lipschitz constant is relatively small, GPM with Armijo rule seems to be more efficient than AGPM.

There is another important barycenter on $\mathbb{P}$ (different from the Wasserstein barycenter) equipped with the Cartan-Hadamard trace metric $\langle X, Y\rangle_A = \mathrm{tr}(A^{-1}XA^{-1}Y)$. The Cartan barycenter has been studied extensively the past several years by many authors as a multivariable extension of the two variable matrix geometric mean [19, 20]. At present, the proximal point method with sublinear convergence [3] seems to be the only global optimization method for the Cartan barycenter. The optimization problem in this paper can be cast as the corresponding one on the Riemannian manifold $\mathbb{P}$ with the Riemannian metric

$$\langle X, Y\rangle_A = \mathrm{Tr}(XAY) = \mathrm{Tr}((XA^{\frac{1}{2}})(XA^{\frac{1}{2}})^*)$$

for $X, Y \in T_A(\mathbb{P}) \equiv \mathbb{H}$, the tangent space at $A \in \mathbb{P}$ even though our gradient-based algorithm is performed under the classical Euclidean setting rather than the Riemannian space $\mathbb{P}$.

Many other questions remain open. One would like to show the strong convexity of the objective function for more efficient optimization methods. In the one-dimensional case, this is true. Moreover, our minimization problem is of the form

$$\arg\min_{X\in\mathcal{D}} \sum_{j=1}^{n} f_j(X) + g_j(X),$$

where $f_j(X) = w_j \mathrm{tr}(A_j + X)$ and $g_j(X) = -2w_j \mathrm{tr}(A_j^{\frac{1}{2}} X A_j^{\frac{1}{2}})^{\frac{1}{2}}$ are convex functions on the compact and convex set $\mathcal{D}$. The above problem is a classical form for applying incremental methods especially when n is large [5, 23]. So we can adapt incremental (gradient) methods to deal with the minimization problem (6). The key ingredient of the Wasserstein barycenter is the matrix valued map $(A, B) \mapsto (A^{\frac{1}{2}} B A^{\frac{1}{2}})^{\frac{1}{2}}$. We believe that there is a class of differentiable functions $F : \mathbb{P} \times \mathbb{P} \to \mathbb{P}$ satisfying strict convexity for the first or second variable. Then the corresponding minimization problem arises naturally with $g_j(X) = \mathrm{tr}F(A_j, X)$. A unified approach will be studied in our forthcoming papers.

References

[1] M. Agueh and G. Carlier, *Barycenters in the Wasserstein space*, SIAM J. Math. Anal. **43** (2011), no. 2, 904–924.

[2] P. C. Álvarez Esteban, E. del Barrio, J. Cuesta-Albertos, and C. Matrán, *A fixed-point approach to barycenters in Wasserstein space*, J. Math. Anal. Appl. **441** (2016), no. 2, 744–762.

[3] M. Bačák, *Computing medians and means in Hadamard spaces*, SIAM J. Optim. **24** (2014), no. 3, 1542–1566.

[4] D. P. Bertsekas, *Nonlinear Programming*, second edition, Athena Scientific Optimization and Computation Series, Athena Scientific, Belmont, MA, 1999.

[5] ———, *Incremental gradient, subgradient, and proximal methods for convex optimization: A survey*, Preprint arXiv:1507.01030, 2015.

[6] R. Bhatia, *Matrix Analysis*, Graduate Texts in Mathematics, **169**, Springer-Verlag, New York, 1997.

[7] R. Bhatia, T. Jain, and Y. Lim, *On the bures-wasserstein distance between positive definite matrices*, preprint.

[8] J. Bigot and T. Klein, *Consistent estimation of a population barycenter in the wasserstein space*, Preprint arXiv:1212.2562, 2012.

[9] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Comm. Pure Appl. Math. **44** (1991), no. 4, 375–417.

[10] G. Carlier and I. Ekeland, *Matching for teams*, Econom. Theory **42** (2010), no. 2, 397–418.

[11] G. Carlier, A. Oberman, and E. Oudet, *Numerical methods for matching for teams and Wasserstein barycenters*, ESAIM Math. Model. Numer. Anal. **49** (2015), no. 6, 1621–1642.

[12] D. C. Dowson and B. V. Landau, *The Fréchet distance between multivariate normal distributions*, J. Multivariate Anal. **12** (1982), no. 3, 450–455.

[13] W. Gangbo and R. J. McCann, *The geometry of optimal transportation*, Acta Math. **177** (1996), no. 2, 113–161.

[14] W. Gangbo and A. Święch, *Optimal maps for the multidimensional Monge-Kantorovich problem*, Comm. Pure Appl. Math. **51** (1998), no. 1, 23–45.
[15] C. R. Givens and R. M. Shortt, *A class of Wasserstein metrics for probability distributions*, Michigan Math. J. **31** (1984), no. 2, 231–240.
[16] M. Ito and M. Fukuda, *A family of subgradient-based methods for convex optimization problems in a unifying framework*, Optim. Methods Softw. **31** (2016), no. 5, 952–982.
[17] L. Kantorovich and G. Rubinstein, *On a space of completely additive functions*, Vestn. Lenin. Univ. **13** (1958), 52–59.
[18] M. Knott and C. S. Smith, *On a generalization of cyclic monotonicity and distances among random vectors*, Linear Algebra Appl. **199** (1994), 363–371.
[19] J. Lawson and Y. Lim, *Monotonic properties of the least squares mean*, Math. Ann. **351** (2011), no. 2, 267–279.
[20] ———, *Weighted means and Karcher equations of positive operators*, Proc. Natl. Acad. Sci. USA **110** (2013), no. 39, 15626–15632.
[21] A. S. Lewis and J. Malick, *Alternating projections on manifolds*, Math. Oper. Res. **33** (2008), no. 1, 216–234.
[22] R. J. McCann, *A convexity principle for interacting gases*, Adv. Math. **128** (1997), no. 1, 153–179.
[23] A. Nedic and D. P. Bertsekas, *Incremental subgradient methods for nondifferentiable optimization*, SIAM J. Optim. **12** (2001), no. 1, 109–138.
[24] Yu. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program. **103** (2005), no. 1, Ser. A, 127–152.
[25] I. Olkin and F. Pukelsheim, *The distance between two random vectors with given dispersion matrices*, Linear Algebra Appl. **48** (1982), 257–263.
[26] J. Rabin, G. Peyré, J. Delon, and M. Bernot, *Wasserstein Barycenter and Its Application to Texture Mixing*, 435–446, Springer, Berlin, Heidelberg, 2012.
[27] L. Rüschendorf and L. Uckelmann, *On the n-coupling problem*, J. Multivariate Anal. **81** (2002), no. 2, 242–258.
[28] P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization* Preprint http://www.mit.edu/∼dimitrib/PTseng/papers/apgm.pdf, 2008.
[29] C. Villani, *Topics in Optimal Transportation*, Graduate Studies in Mathematics, **58**, American Mathematical Society, Providence, RI, 2003.
[30] ———, *Optimal Transport*, Grundlehren der Mathematischen Wissenschaften, **338**, Springer-Verlag, Berlin, 2009.

Sangho Kum
Department of Mathematics Education
Chungbuk National University
Cheongju 28644, Korea
Email address: shkum@cbnu.ac.kr

Sangwoon Yun
Department of Mathematics Education
Sungkyunkwan University
Seoul 03063, Korea
Email address: yswmathedu@skku.edu