A Deterministic Approach to Avoid Saddle Points

Lisa Maria Kreusser
Department of Applied Mathematics and Theoretical Physics
University of Cambridge
L.M.Kreusser@damtp.cam.ac.uk

Stanley J. Osher
Department of Mathematics
University of California, Los Angeles
sjo@math.ucla.edu

Bao Wang
Department of Mathematics
University of California, Los Angeles
wangbaonj@gmail.com

January 22, 2019

Abstract

Loss functions with a large number of saddle points are one of the main obstacles to training many modern machine learning models. Gradient descent (GD) is a fundamental algorithm for machine learning and converges to a saddle point for certain initial data. We call the region formed by these initial values the "attraction region." For quadratic functions, GD converges to a saddle point if the initial data is in a subspace of up to $n-1$ dimensions. In this paper, we prove that a small modification of the recently proposed Laplacian smoothing gradient descent (LSGD) [Osher, et al., arXiv:1806.06317] contributes to avoiding saddle points without sacrificing the convergence rate of GD. In particular, we show that the dimension of the LSGD's attraction region is at most $\lfloor (n-1)/2 \rfloor$ for a class of quadratic functions which is significantly smaller than GD's $(n-1)$-dimensional attraction region.

1 Introduction

Modern machine learning usually involves non-convex optimization. It is typically very difficult to find the global minima, but even local minima are very useful in most machine learning models. Gradient descent (GD), which is a simple and fundamental optimization algorithm for modern machine learning, converges to an ϵ-first-order stationary point $\mathbf{x}$, with $||\nabla f(\mathbf{x})|| \leq \epsilon$, within $O(\frac{1}{\epsilon^2})$ iterations [19, 4]. The convergence rate of GD is independent of the dimension of the underlying problem, which enables GD to be extremely useful in high-dimensional and large-scale optimization problems, e.g., deep learning [24, 2]. Since GD only considers gradient information, it may get stuck at saddle points. The region containing all the initial conditions of a gradient-based algorithm that converge to a saddle point is called the attraction region.

Saddle points are omnipresent in high dimensional non-convex optimization problems, and correspond to highly suboptimal solutions to many machine learning models, including deep learning [12, 7]. This leads to the central open problem of escaping from saddle points efficiently, e.g., by moving past saddle points quickly or avoiding potential slow-down of the algorithm close to saddle points.

For many machine learning models, it is sufficient to find a solution corresponding to a local minimum. In many significant machine learning problems, all local minima are in fact global minima. These include tensor decomposition [9], dictionary learning [25], phase retrieval [26], matrix sensing [21, 23], matrix completion [11], MaxCut [18], Robust PCA [10], and training some neural nets [15]. In deep learning, it is believed that most of the local minima are as good as global minima [5].

In this paper, we follow a deterministic approach on avoiding saddle points: **Can we propose a gradient-based algorithm where the dimension of the attraction region is reduced compared to GD?** By using Laplacian smoothing gradient descent (LSGD) [22] with a strictly monotonic, bounded smoothing parameter σ, we answer the above question affirmatively. We

prove that for a class of quadratic functions on $\mathbb{R}^n$, we can reduce the dimension of the attraction region from $n-1$ for GD to $\lfloor \frac{n-1}{2} \rfloor$ for the modified LSGD.

1.1 Related Work

Lee et al. [16] showed that GD converges to a local minimizer almost surely if the initial point is randomly chosen. The proof is based on the stable manifold theorem and hence no upper bound of the number of steps is provided. How to escape from saddle points efficiently has been a core problem in non-convex optimization. There has been much recent work on developing algorithms which provably avoid saddle points. The existing algorithms use either the perturbed/noisy gradient or partially or fully the Hessian information.

The Hessian is utilized to distinguish saddle point from local minima. By using the Hessian, Nesterov & Polyak [20] designed a cubic regularization algorithm which converges to an ϵ-second-order stationary point in $O(1/\epsilon^{1.5})$ iterations; Curtis et al. [6] developed a class of trust region algorithms and achieved the same complexity. By using only the product of Hessian with any vector, Agarwal et al. [1] and Carmon & Duchi [3] show that it is also possible to find an ϵ-approximate local minimum.

Since the computation of the Hessian is often too expensive in practice, algorithms without second-order information are very desirable. Ge et al. [9] proved that stochastic gradient descent (SGD) can find local minima of strict saddle functions in polynomial time. Levy [17] showed that noisy normalized gradient descent can converge faster than SGD. Jin et al. [13] proposed the perturbed gradient descent (PGD) and showed that, with an additional cost depending poly-logarithmically on the problem dimension, PGD will converge to an ϵ-second-order stationary point of an ℓ-smooth function with arbitrarily high probability. By combining PGD with the heavy ball method, Jin et al. [14] showed that PGD with momentum can converge faster. Du et al. [8] showed that PGD or noisy GD can escape saddle point faster. More exciting work on avoiding saddle points cannot be discussed here, but we summarize some representative work in Table 1.

Table 1: Algorithms and Guarantees

Algorithm	**Guarantee**	**Oracle**
Ge et al. [9]	Converge to an ϵ-approximate local minimum with high probability	Noisy Stochastic Gradient
Levy [17]	Converge to an ϵ-approximate local minimum with high probability	Noisy Normalized Gradient
Jin et al. [13]	Converge to an ϵ-approximate local minimum with high probability	Perturbed Gradient
This Work	Reduce the dimension of the initial points' space converging to saddle	Smoothed Gradient
Agarwal et al. [1]	Converge to an ϵ-approximate local minimum	Hessian-vector
Carmon et al. [3]	Converge to an ϵ-approximate local minimum	Hessian-vector
Nesterov et al. [20]	Converge to an ϵ-approximate local minimum	Hessian
Curtis et al. [6]	Converge to an ϵ-approximate local minimum	Hessian

Osher et al. [22] recently proposed the LSGD based on the theory of Hamilton-Jacobi partial differential equations. LSGD replaces the gradient by the Laplacian smoothed surrogate which can be computed efficiently with the Thomas algorithm or the Fast Fourier Transform (FFT). LSGD can reduce the variance of stochastic gradient on-the-fly, avoid spurious local minima, and leads to better generalization in training a large number of classical machine learning and deep learning models.

1.2 Our Contribution

We propose the first deterministic first-order algorithm (i.e., without noisy gradients), that can avoid saddle points. For a class of quadratic functions, we can quantify the efficacy in avoiding the saddle point. In summary, our contribution is two-fold:

A small modification of LSGD We replace the positive constant σ in the original LSGD by a positive function $\sigma(k)$ that depends on the iteration index k, i.e., we modify

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \eta \left(\mathbf{I} - \sigma \mathbf{L}\right)^{-1} \nabla f(\mathbf{x}^k),$$

to

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \eta \left(\mathbf{I} - \sigma(k)\mathbf{L}\right)^{-1} \nabla f(\mathbf{x}^k)$$

for the objective function $f(\mathbf{x})$, learning rate η and the discrete one-dimensional Laplacian $\mathbf{L}$, defined below. For the analysis, we require $\sigma(k)$ to be a strictly monotonic, bounded function, e.g., $\sigma(k) = \frac{k+1}{k+2}$. We show that the convergence rate of this algorithm is the same as for GD.

Quantifying the efficacy in avoiding saddle points For the class of quadratic functions

$$f(x_1, \cdots, x_n) = \frac{1}{2} c \left(\sum_{i=1}^{n-1} x_i^2 - x_n^2 \right). \tag{1}$$

for any $c > 0$, the attraction regions

$$\mathcal{W}_1 = \left\{ \mathbf{x}_0 | \mathbf{x}^{k+1} = \mathbf{x}^k - \eta \nabla f(\mathbf{x}^k), \mathbf{x}^\infty = \mathbf{0} \right\}$$

and

$$\mathcal{W}_2 = \left\{ \mathbf{x}_0 | \mathbf{x}^{k+1} = \mathbf{x}^k - \eta \left(\mathbf{I} - \sigma(k)\mathbf{L}\right)^{-1} \nabla f(\mathbf{x}^k), \mathbf{x}^\infty = \mathbf{0} \right\}$$

for GD and modified LSGD, respectively, satisfy

$$\dim \mathcal{W}_1 = n - 1, \qquad \dim \mathcal{W}_2 = \left\lfloor \frac{n-1}{2} \right\rfloor,$$

indicating that the set of initial data converging to a saddle point is significantly smaller for the modified LSGD than for GD.

1.3 Organization

This paper is structured as follows. In section 2, we revisit the LSGD algorithm and motivate the modified LSGD algorithm. For the class of quadratic functions mentioned above, we rigorously prove in section 3 that the modified LSGD can significantly reduce the dimension of the attraction region. We provide the convergence analysis for the modified LSGD for non-convex optimization in section 4. Furthermore, in section 5, we provide numerical results illustrating the advantages of the modified LSGD. Finally, we conclude.

2 Algorithm

2.1 Notation

We use boldface upper-case letters $\mathbf{A}$, $\mathbf{B}$ to denote matrices and boldface lower-case letters $\mathbf{x}$, $\mathbf{y}$ to denote vectors. A_{ij} denotes the $(i,j)^{\text{th}}$ entry of $\mathbf{A}$. For vectors we use $||\cdot||_2$ to denote the ℓ_2-norm, and for matrices we use $||\cdot||_2$ to denote spectral norm, respectively. And $\lambda_{max}(\cdot)$, $\lambda_{min}(\cdot)$, and $\lambda_i(\cdot)$ denote the largest, the smallest, and the i-th largest eigenvalues, respectively. For a function $f : \mathbb{R}^n \to \mathbb{R}$, we use $\nabla f(\cdot)$ and $\nabla^2 f(\cdot)$ to denote its gradient and Hessian, and f^* to denote a local minimum of $f(\cdot)$.

2.2 LSGD Algorithm

Recently, Osher et al. [22] proposed to replace the standard or stochastic gradient vector $\mathbf{v}$ by the following Laplacian smoothed surrogate

$$\left(\mathbf{I} - \sigma\mathbf{L}\right)^{-1} \mathbf{v} \doteq \mathbf{A}_\sigma^{-1} \mathbf{v},$$

where $\mathbf{I}$ and $\mathbf{L}$ are the identity matrix and the discrete one-dimensional Laplacian, respectively. Then,

$$\mathbf{A}_\sigma = \begin{bmatrix} 1+2\sigma & -\sigma & 0 & \dots & 0 & -\sigma \\ -\sigma & 1+2\sigma & -\sigma & \dots & 0 & 0 \\ 0 & -\sigma & 1+2\sigma & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ -\sigma & 0 & 0 & \dots & -\sigma & 1+2\sigma \end{bmatrix} \tag{2}$$

for a positive constant σ and the numerical scheme reads

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \eta \mathbf{A}_\sigma^{-1} \nabla f(\mathbf{x}^k). \tag{3}$$

where GD is recovered for $\sigma = 0$.

This simple Laplacian smoothing can help to avoid spurious minima, reduce the variance of SGD on-the-fly, and leads to better generalization in training neural nets. Computationally, Laplacian smoothing can be achieved either by the Thomas algorithm together with the Sherman-Morrison formula in linear time, or by the FFT in quasi-linear time. For convenience, we use FFT to perform gradient smoothing where

$$\mathbf{A}_\sigma^{-1}\mathbf{v} = \text{ifft}\left(\frac{\text{fft}(\mathbf{v})}{\mathbf{1} - \sigma \cdot \text{fft}(\mathbf{d})}\right),$$

with $\mathbf{d} = [-2, 1, 0, \cdots, 0, 1]^T$.

2.3 Motivation

We apply LSGD with constant $\sigma > 0$ to the quadratic function

$$f(x_1, x_2) = \frac{1}{2}\left(x_1^2 - x_2^2\right).$$

The gradient flow analogy to (3) is given by

$$\begin{bmatrix} \frac{dx_1(t)}{dt} \\ \frac{dx_2(t)}{dt} \end{bmatrix} = \mathbf{A}_\sigma^{-1}\mathbf{B} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \tag{4}$$

where

$$\mathbf{A}_\sigma = \begin{bmatrix} 1+\sigma & -\sigma \\ -\sigma & 1+\sigma \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

in 2D. Since $\mathbf{A}_\sigma^{-1}$ is positive definite, $\mathbf{A}_\sigma^{-1}\mathbf{B}$ has one positive and one negative eigenvalue, denoted by λ^+ and λ^-, respectively. Let the associated eigenvectors be $\mathbf{v}^+$ and $\mathbf{v}^-$, respectively. The solution to Eq. (4) can be written as

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = c_1\mathbf{v}^+ \exp\left(\lambda^+ t\right) + c_2\mathbf{v}^- \exp\left(\lambda^- t\right), \tag{5}$$

where constants c_1 and c_2 depend on the initial data $(x_1(0), x_2(0))$, i.e., the starting point of LSGD. If we choose the starting point to be any point on the line that passes through the origin $(0, 0)$ with the direction along $\mathbf{v}^-$, then the solution to Eq. (4) reduces to

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = c_2\mathbf{v}^- \exp\left(\lambda^- t\right)$$

and $(x_1(t), x_2(t)) \to (0, 0)$ as $t \to \infty$, i.e., LSGD with fixed σ converges to the unique saddle point of f. Hence, the attraction region is given by $\mathcal{W} = \text{span}(\mathbf{v}^-)$ with

$$\mathbf{v}^- = \left(\frac{\sigma + 1 + \sqrt{2\sigma + 1}}{\sigma}, 1\right).$$

It is easy to see that for different σ, the corresponding eigenvector $\mathbf{v}^-$ is different. In particular, this is true for any bounded, strictly monotonic function $\sigma(k)$ of the iteration number k, e.g., $\sigma(k) = \frac{k+1}{k+2}$. Fig. 1 depicts the attraction regions for $\sigma = 0$, $\frac{1+1}{1+2}$, and $\frac{2+1}{2+2}$, respectively. The intersection of any two different attraction regions is only at the origin indicating that starting from any point except $(0, 0)$ LSGD with $\sigma(k) = \frac{k+1}{k+2}$ as the smoothing parameter will not converge to the saddle point $(0, 0)$.

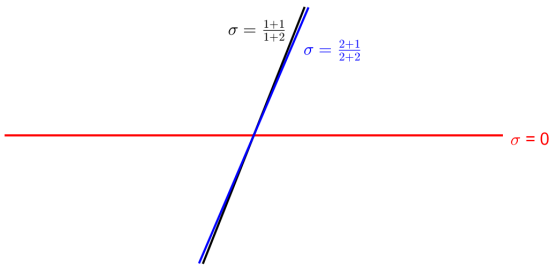


Figure 1: The attraction region when applying LSGD to $x_1^2 - x_2^2$. The black, blue, and red lines plot the attraction regions, respectively, for $\sigma = \frac{1+1}{1+2}$, $\sigma = \frac{2+1}{2+2}$, and 0.

2.4 LSGD with nonconstant σ

Based on the above heuristic analysis, we formulate the modified LSGD algorithm with positive, strictly monotonic, bounded function $\sigma(k)$, e.g., $\sigma(k) = \frac{k+1}{k+2}$. This yields the following scheme

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \eta \left(\mathbf{I} - \sigma(k)\mathbf{L}\right)^{-1} \nabla f(\mathbf{x}^k). \tag{6}$$

Again, $\left(\mathbf{I} - \sigma(k)\mathbf{L}\right)^{-1} \nabla f(\mathbf{x}^k)$ can be obtained by using either the Thomas algorithm or the FFT with exactly the same computational complexity as LSGD with constant σ.

Remark 1. *In Eq. (6), we select an easily computable, positive $\sigma(k)$, any positive, strictly monotonic, bounded function can be a candidate for $\sigma(k)$.*

3 LSGD Can Avoid Saddle Point

Starting from some point $\mathbf{x}^0 \in \mathbb{R}^n$, we apply LSGD with positive, strictly monotonic, bounded smoothing parameter $\sigma(k)$, e.g., $\sigma(k) = \frac{k+1}{k+2}$, to the class of quadratic functions in Eq. (1), resulting in the following iterative scheme

$$\mathbf{x}^{k+1} = (\mathbf{I} - \eta \mathbf{A}_{\sigma(k)}^{-1} B)\mathbf{x}^k, \tag{7}$$

where $\mathbf{A}_{\sigma(k)}$ is obtained from Eq. (2) with $\sigma = \sigma(k)$ and

$$\mathbf{B} = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & -1 \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Lemma 1. *For any $k \in \mathbb{N}$ fixed, the matrix $\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B}$ is diagonalizable, its eigenvectors form a basis of $\mathbb{R}^n$ and the eigenvalues of the matrix $\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B}$ satisfy*

$$\begin{aligned} 1 \geq \lambda_1(\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B}) \geq \ldots \geq \lambda_{n-1}(\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B}) \\ > 0 > \lambda_n(\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B}) \geq -1, \end{aligned}$$

where $\lambda_i(\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B})$ denotes the ith largest eigenvalue of the matrix $\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B}$. In particular, $\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B}$ has exactly one negative eigenvalue.

Proof. For ease of notation, we denote $\sigma(k)$ by σ in the sequel. As a first step, we show that $\mathbf{A}_\sigma^{-1}\mathbf{B}$ is diagonalizable and its eigenvalues $\lambda_i(\mathbf{A}_\sigma^{-1}\mathbf{B})$ are real for $i = 1, \ldots, n$. We prove this by showing that $\mathbf{A}_\sigma^{-1}\mathbf{B}$ is similar to a symmetric matrix. Note that $\mathbf{A}_\sigma^{-1}$ is a real, symmetric, positive definite matrix. Hence, $\mathbf{A}_\sigma^{-1}$ is diagonalizable with $\mathbf{A}_\sigma^{-1} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ for an orthogonal matrix $\mathbf{U}$ and a diagonal matrix $\mathbf{D}$ with eigenvalues $\lambda_i(\mathbf{A}_\sigma^{-1}) > 0$ for $i = 1, \ldots, n$ on the diagonal. This implies that there exists a real, symmetric, positive definite square root $\mathbf{A}_\sigma^{-1/2} = \mathbf{U}\sqrt{\mathbf{D}}\mathbf{U}^T$ with

$\mathbf{A}_\sigma^{-1/2}\mathbf{A}_\sigma^{-1/2} = \mathbf{A}_\sigma^{-1}$ where $\sqrt{\mathbf{D}}$ denotes a diagonal matrix with diagonal entries $\sqrt{\lambda_i(\mathbf{A}_\sigma^{-1})} > 0$. We have

$$\mathbf{A}_\sigma^{1/2}\mathbf{A}_\sigma^{-1}\mathbf{B}\mathbf{A}_\sigma^{-1/2} = \mathbf{A}_\sigma^{-1/2}\mathbf{B}\mathbf{A}_\sigma^{-1/2}$$

where $\mathbf{A}_\sigma^{-1/2}\mathbf{B}\mathbf{A}_\sigma^{-1/2}$ is symmetric due to the symmetry of $\mathbf{A}_\sigma^{-1/2}$ and $\mathbf{B}$. Thus, $\mathbf{A}_\sigma^{-1}\mathbf{B}$ is similar to the symmetric matrix $\mathbf{A}_\sigma^{-1/2}\mathbf{B}\mathbf{A}_\sigma^{-1/2}$. In particular, $\mathbf{A}_\sigma^{-1}\mathbf{B}$ is diagonalizable and has real eigenvalues like $\mathbf{A}_\sigma^{-1/2}\mathbf{B}\mathbf{A}_\sigma^{-1/2}$.

Note that $\det(\mathbf{A}_\sigma^{-1}\mathbf{B}) = \det(\mathbf{A}_\sigma^{-1})\det(\mathbf{B}) < 0$ since $\det(\mathbf{A}_\sigma^{-1}) > 0$ and $\det(\mathbf{B}) = -1$. Since the determinant of a matrix is equal to the product of its eigenvalues and all eigenvalues of $\mathbf{A}_\sigma^{-1}\mathbf{B}$ are real, this implies that $\mathbf{A}_\sigma^{-1}\mathbf{B}$ has an odd number of negative eigenvalues. Next, we show that $\mathbf{A}_\sigma^{-1}\mathbf{B}$ has exactly one negative eigenvalue. Defining

$$\tilde{\mathbf{B}} := \begin{pmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & & -2 \end{pmatrix}$$

we have

$$\mathbf{A}_\sigma^{-1}\mathbf{B} = \mathbf{A}_\sigma^{-1} + \mathbf{A}_\sigma^{-1}\tilde{\mathbf{B}}$$

where the matrix $\mathbf{A}_\sigma^{-1}\tilde{\mathbf{B}}$ has $n-1$-fold eigenvalue 0 and its last eigenvalue is given by $-2[\mathbf{A}_\sigma^{-1}]_{nn}$. We can write $\mathbf{A}_\sigma^{-1} = \frac{1}{\det \mathbf{A}_\sigma}\tilde{\mathbf{C}}$ where $\tilde{C}_{ij} = (-1)^{i+j}M_{ij}$ for the (i,j)-minor M_{ij}, defined as the determinant of the submatrix of $\mathbf{A}_\sigma$ by deleting the ith row and the jth column of $\mathbf{A}_\sigma$. Since all leading principal minors are positive for positive definite matrices, this implies that $\mathbf{M}_{nn} > 0$ due to the positive definiteness of $\mathbf{A}_\sigma$ and hence $[\mathbf{A}_\sigma]_{nn} > 0$, implying that $\lambda_i(\mathbf{A}_\sigma^{-1}\tilde{\mathbf{B}}) = 0$ for $i = 1, \dots, n-1$ and $\lambda_n(\mathbf{A}_\sigma^{-1}\tilde{\mathbf{B}}) < 0$. The eigenvalues of $\mathbf{A}_\sigma^{-1}\mathbf{B}$ can now be estimated by Weyl's inequality for the sum of matrices, leading to

$$\lambda_{n-2}(\mathbf{A}_\sigma^{-1}\mathbf{B}) \geq \lambda_{n-1}(\mathbf{A}_\sigma^{-1}) + \lambda_{n-1}(\mathbf{A}_\sigma^{-1}\tilde{\mathbf{B}}) > 0$$

since the first term is positive and the second term is negative. Since $\mathbf{A}_\sigma^{-1}\mathbf{B}$ has an odd number of negative eigenvalues, this implies that $\lambda_{n-1}(\mathbf{A}_\sigma^{-1}\mathbf{B}) > 0$ and $\lambda_n(\mathbf{A}_\sigma^{-1}\mathbf{B}) < 0$. In particular, $\mathbf{A}_\sigma^{-1}\mathbf{B}$ has exactly one negative eigenvalue.

To estimate upper and lower bounds of the eigenvalues of $\mathbf{A}_\sigma^{-1}\mathbf{B}$, note that the eigenvalues of the Laplacian L are given by $2 - 2\cos(2\pi k/n) \in [0, 4]$ for $k = 0, \dots, n/2$, implying that $\lambda_i(\mathbf{A}_\sigma) \in [1, 1+4\sigma]$ and in particular, we have

$$\lambda_i(\mathbf{A}_\sigma^{-1}) \in \left[\frac{1}{1+4\sigma}, 1\right]$$

for $i = 1, \dots, n$. Besides, we have

$$|\lambda_i(\mathbf{A}_\sigma^{-1}\mathbf{B})| \leq \rho(\mathbf{A}_\sigma^{-1}\mathbf{B}) = \|\mathbf{A}_\sigma^{-1}\mathbf{B}\| \leq \|\mathbf{A}_\sigma^{-1}\|\|\mathbf{B}\| = \rho(\mathbf{A}_\sigma^{-1})\rho(\mathbf{B}) \leq 1$$

all $i = 1, \dots, n$, where $\rho(\mathbf{B})$ denotes the spectral radius of $\mathbf{B}$ and $\|\mathbf{B}\|$ denotes the operator norm of $\mathbf{B}$. □

Since $\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B}$ is diagonalizable by Lemma 1, we can consider the invertible matrix $\mathbf{P}_{\sigma(k)} = (\mathbf{p}_{1,\sigma(k)}, \dots, \mathbf{p}_{n,\sigma(k)})$ whose columns $\mathbf{p}_{i,\sigma(k)}$ denote the eigenvectors of $\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B}$, associated with the eigenvalues $\lambda_i(\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B})$, i.e.

$$\lambda_i(\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B})\mathbf{p}_{i,\sigma(k)} = \mathbf{A}_{\sigma(k)}^{-1}\mathbf{B}\mathbf{p}_{i,\sigma(k)}$$

for all $i = 1, \dots, n$. Without loss of generality, we can assume that $\{\mathbf{p}_{1,\sigma(k)}, \dots, \mathbf{p}_{n,\sigma(k)}\}$ form a basis of unit vectors of $\mathbb{R}^n$. In the following, $p_{i,j,\sigma(k)}$ denotes the ith entry of the jth eigenvector $\mathbf{p}_{j,\sigma(k)}$ of $\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B}$, i.e. $\mathbf{p}_{j,\sigma(k)} = (p_{1,j,\sigma(k)}, \dots, p_{n,j,\sigma(k)})$.

Lemma 2. *For any $k \in \mathbb{N}$ fixed, the matrix $\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B}$ has $\lfloor n/2 \rfloor$ eigenvectors $\mathbf{p}_{j,\sigma(k)} = (p_{1,j,\sigma(k)}, \dots, p_{n,j,\sigma(k)})$, associated with positive eigenvalues. The lth entry $p_{l,j,\sigma(k)}$ of the jth eigenvector $\mathbf{p}_{j,\sigma(k)}$ satisfies:*

$$p_{l,j,\sigma(k)} \begin{cases} = p_{n-l,j,\sigma(k)}, & l = 1, \dots, n-1, \\ \neq 0, & l = n. \end{cases}.$$

There are $\lfloor (n-1)/2 \rfloor$ eigenvectors associated with positive eigenvalues where the entry $p_{l,j,\sigma(k)}$ of eigenvector $\mathbf{p}_{j,\sigma(k)}$ satisfies

$$p_{l,j,\sigma(k)} = b \sin(l\theta_j), \quad l = 1, \dots, n \tag{8}$$

for $b \in \mathbb{R}\backslash\{0\}$ and $\theta_j = \frac{2\pi m_j}{n}$ for some $m_j \in \mathbb{Z}$, implying

$$p_{l,j,\sigma(k)} = \begin{cases} -p_{n-l,j,\sigma(k)}, & l = 1, \dots, n-1, \\ 0, & l = n. \end{cases}$$

Besides, the eigenvector $\mathbf{p}_{n,\sigma(k)}$ associated with the unique negative eigenvalue $\lambda_n(\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B})$ satisfies

$$p_{l,j,\sigma(k)} \begin{cases} = p_{n-l,j,\sigma(k)}, & l = 1, \dots, n-1, \\ \neq 0, & l = n. \end{cases}.$$

.

Proof. Since $\sigma(k)$ is fixed in the entire proof, we consider $\sigma = \sigma(k)$ throughout the proof. Besides, we simplify the notation by dropping the index $\sigma = \sigma(k)$ in the notation of the eigenvectors $\mathbf{p}_{j,\sigma(k)} = (p_{1,j,\sigma(k)}, \dots, p_{n,j,\sigma(k)})$, and we write $\mathbf{p}_j = (p_{1,j}, \dots, p_{n,j})$ for $j = 1, \dots, n$.

Since $\mathbf{A}_\sigma^{-1}\mathbf{B}$ and $\mathbf{B}\mathbf{A}_\sigma$ have the same eigenvectors and their eigenvalues are reciprocals, we can consider $\mathbf{B}\mathbf{A}_\sigma$ for determining the eigenvectors $\mathbf{p}_j$ for $j = 1, \dots, n$. Note that the $n-1$ eigenvectors $\mathbf{p}_j$ of $\mathbf{B}\mathbf{A}_\sigma$ for $j = 1, \dots, n-1$ are associated with positive eigenvalues $\lambda_j(\mathbf{B}\mathbf{A}_\sigma)$ of $\mathbf{B}\mathbf{A}_\sigma$, while the eigenvector $\mathbf{p}_n$ is associated with the only negative eigenvalue $\lambda_n(\mathbf{B}\mathbf{A}_\sigma)$. By introducing a slack variable $p_{0,j}$ we rewrite the eigenequation for the jth eigenvalue $\lambda_j(\mathbf{B}\mathbf{A}_\sigma)$, given by

$$(\mathbf{B}\mathbf{A}_\sigma - \lambda_j(\mathbf{B}\mathbf{A}_\sigma))\mathbf{p}_j = \mathbf{0},$$

as:

$$\begin{aligned} -\sigma p_{k-1,j} + (1 + 2\sigma - \lambda_j(\mathbf{B}\mathbf{A}))p_{k,j} - \sigma p_{k+1,j} = 0, \\ k = 1, \dots, n-1, \end{aligned} \tag{9}$$

with boundary conditions

$$\sigma p_{1,j} + \sigma p_{n-1,j} - (1 + 2\sigma + \lambda_j(\mathbf{B}\mathbf{A}_\sigma))p_{n,j} = 0 \tag{10}$$

and

$$p_{0,j} = p_{n,j}. \tag{11}$$

Equation (9) is a difference equation which can be solved by making the ansatz $p_{k,j} = r^k$. Plugging this ansatz into (9) results in the quadratic equation

$$1 - \frac{1 + 2\sigma - \lambda_j(\mathbf{B}\mathbf{A}_\sigma)}{\sigma} r + r^2 = 0$$

with solutions $r_{+/-} = d \pm \sqrt{d^2 - 1}$ where

$$d := \frac{1 + 2\sigma - \lambda_j(\mathbf{B}\mathbf{A}_\sigma)}{2\sigma}.$$

Note that $r_+ r_- = d^2 - (d^2 - 1) = 1$ and $2d = r_+ + r_- = r_+ + (r_+)^{-1}$.

Let us consider the eigenvector $\mathbf{p}_n$ first. Since $\lambda_n(\mathbf{B}\mathbf{A}_\sigma) < 0$ this implies that $d > 1$ and in particular $r_+ \neq r_-$. We set $r := r_+$, implying $r_- = 1/r$, and obtain the general solution of the form

$$p_{k,n} = b_1 r^k + b_2 r^{-k}, \quad k = 0, \dots, n$$

for scalars $b_1, b_2 \in \mathbb{R}$ which have to be determined from the boundary conditions (10),(11). From (11) we obtain

$$b_1 + b_2 = b_1 r^n + b_2 r^{-n},$$

implying $b_1(1-r^n) = b_2 r^{-n}(1-r^n)$ and in particular $b_1 = b_2 r^{-n}$ since $r = r_+ > 1$. Hence, we obtain

$$p_{k,n} = b_1(r^k + r^{n-k}), \quad k = 0, \dots, n. \tag{12}$$

For non-trivial solutions for the eigenvector $\mathbf{p}_n$ we require $b_1 \neq 0$. Note that (12) implies that $p_{k,n} = p_{n-k,n}$ for $k = 0, \dots, n$. Further note that the boundary condition (10) implies that $p_{n,n} \neq 0$ is necessary for non-trivial solutions.

Next, we consider the $n-1$ eigenvectors $\mathbf{p}_j$ of $\mathbf{B}\mathbf{A}_\sigma$ associated with positive eigenvalues $\lambda_j(\mathbf{B}\mathbf{A}_\sigma) > 0$ for $j = 1, \dots, n-1$. Note that all positive eigenvalues of $\mathbf{B}\mathbf{A}_\sigma$ are in the interval $[1, 1+4\sigma]$ since $\lambda_j(\mathbf{A}_\sigma) \in [1, 1+4\sigma]$ and

$$\lambda_j(\mathbf{B}\mathbf{A}_\sigma) = \frac{1}{\lambda_j(\mathbf{A}_\sigma^{-1}\mathbf{B})} \geq 1$$

by Lemma 1. Hence, $\lambda_j(\mathbf{B}\mathbf{A}_\sigma) \leq \rho(\mathbf{B}\mathbf{A}_\sigma) = \|\mathbf{B}\mathbf{A}_\sigma\| \leq \|\mathbf{B}\|\|\mathbf{A}_\sigma\| = \rho(\mathbf{B})\rho(\mathbf{A}_\sigma) \leq 1 + 4\sigma$. Thus, it is sufficient to consider three different cases $\lambda_j(\mathbf{B}\mathbf{A}_\sigma) = 1$, $\lambda_j(\mathbf{B}\mathbf{A}_\sigma) = 1 + 4\sigma$ and $\lambda_j(\mathbf{B}\mathbf{A}_\sigma) \in (1, 1+4\sigma)$.

We start by showing that all eigenvalues satisfy in fact $\lambda_j(\mathbf{B}\mathbf{A}_\sigma) \in (1, 1+4\sigma)$. For this, assume that there exists $\lambda_j(\mathbf{B}\mathbf{A}_\sigma) = 1$ for some $j \in \{1, \dots, n-1\}$, implying that we have a single root $r_+ = r_- = d = 1$. The general solution to the difference equation (9) with boundary conditions (10), (11) reads

$$p_{k,j} = (b_{1,j} + b_{2,j}k)r^k = b_{1,j} + b_{2,j}k, \quad k = 0, \dots, n$$

for constants $b_{1,j}, b_{2,j} \in \mathbb{R}$. Summing up all equations in (9) and subtracting (10) implies that $2p_{n,j} = 0$, i.e. $p_{n,j} = 0$. Hence, (11) implies $p_{0,j} = p_{n,j}$ and our ansatz yields $0 = p_{n,j} = p_{0,j} = b_{1,j}$. This results in $p_{k,j} = b_{2,j}k$ and $p_{n,j} = 0 = b_{2,j}n$ implies $b_{2,j} = 0$. In particular, there exists no non-trivial solution and hence $\lambda_j(\mathbf{B}\mathbf{A}_\sigma) \neq 1$ for all $j = 1, \dots, n-1$. Next, we show that $\lambda_j(\mathbf{B}\mathbf{A}_\sigma) \neq 1 + 4\sigma$ for all $j = 1, \dots, n-1$ by contradiction and assume that there exists $j \in \{1, \dots, n-1\}$ such that $\lambda_j(\mathbf{B}\mathbf{A}_\sigma) = 1 + 4\sigma$, implying that $r_+ = r_- = d = -1$. Due to the single root, the general solution is of the form

$$p_{k,j} = (b_{1,j} + b_{2,j}k)r^k = (b_{1,j} + b_{2,j}k)(-1)^k, \quad k = 0, \dots, n.$$

For n even, (11) yields $b_{1,j} = p_{0,j} = p_{n,j} = b_{1,j} + b_{2,j}n$, implying $b_{2,j} = 0$. Hence, $p_{k,j} = b_{1,j}$, but the constant solution does not satisfy the boundary condition (10) unless $b_{1,j} = 0$, resulting in the trivial solution. Similarly, we obtain for n odd that $b_{1,j} = p_{0,j} = p_{n,j} = -b_{1,j} - b_{2,j}n$, implying $b_{2,j} = -2b_{1,j}/n$, i.e. $p_{k,j} = b_{1,j}(1 - 2k/n)(-1)^k$. Plugging this into the boundary condition (10) yields $b_{1,j} = 0$ since $\sigma > 0$ and $n \geq 2$. In particular, there exists no non-trivial solution and the positive eigenvalues satisfy $\lambda_j(\mathbf{B}\mathbf{A}_\sigma) < 1 + 4\sigma$ for all $j = 1, \dots, n-1$. Hence, we can now assume that $\lambda_j(\mathbf{B}\mathbf{A}_\sigma) \in (1, 1+4\sigma)$. This implies that $d \in (-1, 1)$ and $r_{+/-} = d \pm i\sqrt{1-d^2}$ leads to two distinct roots. Setting $r := r_+$ with $|r| = 1$, we can introduce an angle θ and write $r = \exp(i\theta) = \cos\theta + i\sin\theta$, implying $d = \cos\theta$ and $r^k = \exp(ik\theta)$. Due to the distinct roots we consider the ansatz

$$p_{k,j} = b_{1,j}r^k + b_{2,j}r^{-k}, \quad k = 0, \dots, n.$$

The boundary condition (11) implies $b_{1,j}r^n(1-r^n) = b_{2,j}(1-r^n)$ resulting in the two cases $r^n = 1$ and $b_{1,j}r^n = b_{2,j}$.

In the case $r^n = \cos(n\theta) + i\sin(n\theta) = 1$, we can conclude that $\theta = 2\pi m/n$ for some $m \in \mathbb{Z}$, implying

$$p_{k,j} = (b_{1,j} + b_{2,j})\cos(k\theta) + i(b_{1,j} - b_{2,j})\sin(k\theta), \quad k = 0, \dots, n,$$

and we obtain

$$\begin{aligned} p_{1,j} &= (b_{1,j} + b_{2,j})\cos(\theta) + i(b_{1,j} - b_{2,j})\sin(\theta), \\ p_{n-1,j} &= (b_{1,j} + b_{2,j})\cos(\theta) - i(b_{1,j} - b_{2,j})\sin(\theta), \\ p_{n,j} &= b_{1,j} + b_{2,j}. \end{aligned}$$

Boundary condition (10) yields

$$2\sigma(b_{1,j}+b_{2,j})\cos(\theta)-(1+2\sigma+\lambda_j(\mathbf{B}\mathbf{A}_\sigma))(b_{1,j}+b_{2,j})=0,$$

implying $b_{1,j}+b_{2,j}=0$ or $2\sigma\cos(\theta)=1+2\sigma+\lambda_j(\mathbf{B}\mathbf{A}_\sigma)$. Since $\lambda_j(\mathbf{B}\mathbf{A}_\sigma)>0$, the second case cannot be satisfied and we conclude $b_{1,j}+b_{2,j}=0$, resulting in the general solution of the form $p_{k,j}=2ib_{1,j}\sin(k\theta)$ for $k=0,\dots,n$ for $b_{1,j}\in\mathbb{C}$, i.e., $\mathbf{p}=2ib_{1,j}(\sin(\theta),\dots,\sin(n\theta))$. Rescaling by $1/(2i)$ results in the real eigenvectors $\mathbf{p}_j=(p_{1,j},\dots,p_{n,j})$ whose entries are of the form (8) where $b\in\mathbb{R}$ can be chosen such that $\|\mathbf{p}_j\|=1$, for instance. Here, $p_{k,j}=-p_{n-k,j}$ for $k=1,\dots,n-1$ and $p_{n,j}=0$. Further note that $p_{n/2,j}=0$ for n even. By writing θ as $\theta_j=(2\pi m_j)/n$ for some $m_j\in\mathbb{Z}$, we can construct $(n-1)/2$ linearly independent eigenvectors for n odd and $(n-2)/2$ for n even, resulting in $\lfloor(n-1)/2\rfloor$ linearly independent eigenvectors for any $n\in\mathbb{N}$. Since the matrix $\mathbf{A}_\sigma^{-1}\mathbf{B}$ is diagonalizable, there exist exactly $\lfloor(n-1)/2\rfloor$ normalized eigenvectors of the form (8).

For $b_{1,j}r^n=b_{2,j}$, we obtain

$$p_{k,j}=b_{1,j}(r^k+r^{n-k})=p_{n-k,j},\quad k=0,\dots,n,$$

i.e. the entries of $\mathbf{p}_j$ are arranged in the same way as the entries of $\mathbf{p}_n$. Further note that we can always set $p_{n,j}\neq 0$, and additionally $p_{k,j}$ with $k=1,\dots,n/2$ for n even and $p_{k,j}$ with $k=1,\dots,(n-1)/2$ for n odd, resulting in a space of dimension $\lfloor n/2\rfloor+1$. Since $\mathbf{p}_1,\dots,\mathbf{p}_n$ form a basis of $\mathbb{R}^n$, there are $\lfloor n/2\rfloor$ eigenvectors of this form, associated with positive eigenvalues. □

Lemma 2 implies that for the matrix $\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B}$ we have one eigenvector $\mathbf{p}_{n,\sigma(k)}$ associated with the unique negative eigenvalue, $\lfloor(n-1)/2\rfloor$ eigenvectors of the form (8) and $\lfloor n/2\rfloor$ eigenvalues associated with certain positive eigenvalues which are of the same form as $\mathbf{p}_{n,\sigma(k)}$. Note that $1+\lfloor(n-1)/2\rfloor+\lfloor n/2\rfloor=n$ for any $n\in\mathbb{N}$.

In the sequel, we number the eigenvectors as follows. By $\mathbf{p}_{j,\sigma(k)}$ for $j=1,\dots,\lfloor(n-1)/2\rfloor$ we denote the $\lfloor(n-1)/2\rfloor$ eigenvectors of the form (8). For $j=\lfloor(n-1)/2\rfloor+1,\dots,n-1$, $\mathbf{p}_{j,\sigma(k)}$ denotes the $\lfloor n/2\rfloor$ eigenvectors of the form $p_{l,j,\sigma(k)}=p_{n-l,j,\sigma(k)}$ for $l=1,\dots,n-1$ and $j=\lfloor(n-1)/2\rfloor+1,\dots,n-1$, associated with positive eigenvalues, and $\mathbf{p}_{n,\sigma(k)}$ denotes the eigenvector associated with the unique negative eigenvalue. Similarly, we relabel the eigenvalues so that eigenvalue $\lambda_j(\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B})$ is associated with eigenvector $\mathbf{p}_{j,\sigma(k)}$. Using this basis of eigenvectors, we can write $\mathbb{R}^n=\mathcal{V}\oplus\mathcal{W}$ with

$$\begin{aligned}\mathcal{V}&:=\left\{x_k=x_{n-k},k=1,\dots,\left\lfloor\frac{n-1}{2}\right\rfloor\right\},\\ \mathcal{W}&:=\left\{x_k=-x_{n-k},k=1,\dots,\left\lfloor\frac{n-1}{2}\right\rfloor;x_n=0\right\},\end{aligned}$$

where $\mathcal{V}$ is a space of dimension $\lfloor(n-1)/2\rfloor+2=\lfloor n/2\rfloor+1$ for n even and of dimension $\lfloor(n-1)/2\rfloor+1=\lfloor n/2\rfloor+1$ for n odd. Note that $\mathcal{W}=\operatorname{span}\{\mathbf{p}_{1,\sigma(k)},\dots,\mathbf{p}_{\lfloor(n-1)/2\rfloor,\sigma(k)}\}$ and $\mathcal{V}=\operatorname{span}\{\mathbf{p}_{\lfloor(n-1)/2\rfloor+1,\sigma(k)},\dots,\mathbf{p}_{n,\sigma(k)}\}$ where $\mathcal{V}$, $\mathcal{W}$ are orthogonal spaces and their definition is independent of $\sigma=\sigma(k)$ for any $k\in\mathbb{N}$. For ease of notation, we introduce the set of indices $\mathcal{I}_\mathcal{V}:=\{\lfloor(n-1)/2\rfloor+1,\dots,n\}$ and $\mathcal{I}_\mathcal{W}:=\{1,\dots,\lfloor(n-1)/2\rfloor\}$ so that for all $k\in\mathbb{N}$ we obtain $\mathbf{p}_{i,\sigma(k)}\in\mathcal{V}$ for all $i\in\mathcal{I}_\mathcal{V}$ and $\mathbf{p}_{i,\sigma(k)}\in\mathcal{W}$ for all $i\in\mathcal{I}_\mathcal{W}$. In particular, $\mathbf{p}_{i,\sigma(k)}$ for $i\in\mathcal{I}_\mathcal{W}$ is independent of σ.

We have all the preliminary results to proof the main statement of this paper now:

Theorem 1. *For any $n\geq 2$, the iterative scheme (7) converges to the minimizer of f, provided $\mathbf{x}^0\notin\mathcal{W}$.*

Proof. Since $\sigma=\sigma(k)$ is a bounded, strictly increasing or decreasing function, this implies that $\sigma(k)$ converges as $k\to\infty$ and $\sigma(k)$ is a Cauchy sequence, i.e. for any $\epsilon>0$ there exists $k_0\in\mathbb{N}$ such that $|\sigma(k_1)-\sigma(k_2)|<\epsilon$ for all $k_1,k_2\geq k_0$. Given $\epsilon>0$ and $k\geq k_0$, we can write $\mathbf{x}^k$ with respect to the basis $\mathbf{p}_{j,\sigma(k)}$ for $j=1,\dots,n$ yielding $\mathbf{x}^k=\sum_{j=1}^n\alpha_{j,k}\mathbf{p}_{j,\sigma(k)}$ where, in fact, $\mathbf{p}_{j,\sigma(0)}=\mathbf{p}_{j,\sigma(k)}$ independent of $\sigma(k)$ for $j\in\mathcal{I}_\mathcal{W}$. From the LSGD scheme (7) we obtain

$$\begin{aligned}\mathbf{x}^{k+1}&=(\mathbf{I}-\eta\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B})\mathbf{x}^k\\&=\sum_{j=1}^n\alpha_{j,k}(1-\eta\lambda_j(\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B}))\mathbf{p}_{j,\sigma(k)}.\end{aligned}$$

By Taylor expansion, we have

$$\mathbf{A}_{\sigma(k+1)}^{-1} \approx \mathbf{A}_{\sigma(k)+\epsilon}^{-1} = \mathbf{A}_{\sigma(k)}^{-1} + \mathcal{O}(\epsilon),$$

resulting in

$$\begin{aligned}\mathbf{x}^{k+2} &= (\mathbf{I} - \eta \mathbf{A}_{\sigma(k+1)}^{-1}\mathbf{B})\mathbf{x}^{k+1} \\ &= \sum_{j=1}^{n} \alpha_{j,k}(1 - \eta\lambda_j(\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B}))^2 \mathbf{p}_{j,\sigma(k)} + \mathcal{O}(\epsilon),\end{aligned}$$

and iterative application yields

$$\mathbf{x}^{k+l} = \sum_{j=1}^{n} \alpha_{j,k}(1 - \eta\lambda_j(\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B}))^l \mathbf{p}_{j,\sigma(k)} + \mathcal{O}(\epsilon) \tag{13}$$

for any $l \geq 0$. Note that $|1-\eta\lambda_j(\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B})| < 1$ for all $j = 1, \dots, n-1$ and $|1-\eta\lambda_n(\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B})| > 1$ by Lemma 1. Hence, $\mathbf{x}^{k+l}$ is unbounded as $l \to \infty$ if and only if $\alpha_{n,k} \neq 0$. It remains to show that for $\mathbf{x}^0 \notin \mathcal{W}$ there exists $k \geq k_0$ such that $\mathbf{x}^k = \sum_{j=1}^n \alpha_{j,k}\mathbf{p}_{j,\sigma(k)}$ with $\alpha_{n,k} \neq 0$.

Let $\mathbf{x}^k = \sum_{i=1}^n \alpha_{i,k}\mathbf{p}_{i,\sigma(k)}$ which can be rewritten as $\mathbf{x}^k = \mathbf{w}^k + \mathbf{v}^k$ where

$$\mathbf{v}^k := \sum_{i\in\mathcal{I}_\mathcal{V}} \alpha_{i,k}\mathbf{p}_{i,\sigma(k)} \in \mathcal{V}, \mathbf{w}^k := \sum_{i\in\mathcal{I}_\mathcal{W}} \alpha_{i,k}\mathbf{p}_{i,\sigma(k)} \in \mathcal{W}$$

for $\mathcal{V}, \mathcal{W}$ defined independently of σ with $\mathbb{R}^n = \mathcal{V} \oplus \mathcal{W}$. We consider the sequence $\mathbf{x}^{k+1} = (\mathbf{I} - \eta\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B})\mathbf{x}^k = (\mathbf{I}-\eta\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B})\mathbf{w}^k + (\mathbf{I}-\eta\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B})\mathbf{v}^k$. We obtain $\mathbf{w}^{k+1} = (\mathbf{I}-\eta\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B})\mathbf{w}^k \in \mathcal{W}$ and $\mathbf{v}^{k+1} = (\mathbf{I} - \eta\mathbf{A}_{\sigma(k)}^{-1}\mathbf{B})\mathbf{v}^k \in \mathcal{V}$. This results in $\mathbf{w}^k \to 0$ as $k \to \infty$ since $\mathbf{w}^0 := \sum_{i\in\mathcal{I}_\mathcal{W}} \alpha_{i,0}\mathbf{p}_{i,\sigma(0)}$ yields $\mathbf{w}^k = \sum_{i\in\mathcal{I}_\mathcal{W}} \alpha_{i,0}(1 - t\lambda_i(\mathbf{A}_\sigma^{-1}\mathbf{B}))^k\mathbf{p}_{i,\sigma(k)}$ for any $k \geq 0$ where the eigenvalues $\lambda_i(\mathbf{A}_\sigma^{-1}\mathbf{B})$ satisfy $|1 - t\lambda_i(\mathbf{A}_\sigma^{-1}\mathbf{B})| < 1$ for $i \in \mathcal{I}_\mathcal{W}$. For showing the unboundedness of $\mathbf{x}^k$ with $\mathbf{x}^0 \notin \mathcal{W}$ as $k \to \infty$ it is hence sufficient to show that $\mathbf{v}^k$ is unbounded as $k \to \infty$ where $\mathbf{v}^0 \neq \mathbf{0}$. We show that for any $\tilde{\mathbf{v}}^{k_0} \in \mathcal{V}\backslash\{\mathbf{0}\}$ we obtain $\mathbf{v}^k = \sum_{i\in\mathcal{I}_\mathcal{V}} \alpha_{i,k}\mathbf{p}_{i,\sigma(k)}$ with $\alpha_{i,k} \neq 0$ after finitely many steps.

Note that we have $|\mathcal{I}_\mathcal{V}| = n - \lfloor (n-1)/2 \rfloor$ parameters in the definition of $\mathbf{v}^{k_0} = \sum_{i\in\mathcal{I}_\mathcal{V}} \alpha_{i,k_0}\mathbf{p}_{i,\sigma(k_0)}$ where $\alpha_{n,k_0} = 0$. Another parameter in the linear combination for $\mathbf{v}^{k_0}$ can be regarded as a scaling parameter and thus, it can be set as any constant. This results in $n - \lfloor (n-1)/2 \rfloor - 2$ further parameters which can be adjusted in such a way that $\mathbf{v}^k = \sum_{i\in\mathcal{I}_\mathcal{V}} \alpha_{i,k}\mathbf{p}_{i,\sigma(k)}$ with $\alpha_{n,k} = 0$ for $k = k_0, \dots, k_e$ with $k_e := k_0 + n - \lfloor (n-1)/2 \rfloor - 2$. We can determine these $n - \lfloor (n-1)/2 \rfloor - 1$ parameters from $n - \lfloor (n-1)/2 \rfloor - 1$ conditions, resulting in a linear system of $n - \lfloor (n-1)/2 \rfloor - 1$ equations. However, the additional condition

$$\begin{aligned}\mathbf{v}^{k_0+k_e+1} &= \Pi_{i=0}^{k_e}(\mathbf{I} - \eta\mathbf{A}_{\sigma(k_0+i)}^{-1}\mathbf{B})\mathbf{v}^{k_0} \\ &= \sum_{i\in\mathcal{I}_\mathcal{V}} \alpha_{i,k_0+k_e+1}\mathbf{p}_{i,\sigma(k_0+k_e)}\end{aligned}$$

with $\alpha_{n,k_0+k_e+1} = 0$ leads to the unique trivial solution of the full linear system of size $n - \lfloor (n-1)/2 \rfloor$, i.e., the assumption $\mathbf{v}^{k_0} \neq \{0\}$ is not satisfied. This implies that for any $\mathbf{v}^{k_0} \in \mathcal{V}\backslash\{0\}$ a vector $\mathbf{v}^k = \sum_{i\in\mathcal{I}_\mathcal{V}} \alpha_{i,k}\mathbf{p}_{i,\sigma(k)}$ with $\alpha_{n,k} \neq 0$ is reached in finitely (after at most $k_e + 1 = n - \lfloor (n-1)/2 \rfloor - 1$) steps. □

Remark 2. *Theorem 1 implies that for the modified LSGD the dimension of the attraction region $\mathcal{W}$ is 0 for the class of functions in (1), i.e., $\mathcal{W} = \{0\}$. For any quadratic function with saddle points $f(x_1, x_2) = c_1x_1^2 + c_2x_2^2$, i.e., $c_1 > 0 > c_2$ after a potential change of variables, the proof of Theorem 1 can easily be adapted and $\mathcal{W} = \{0\}$ for any quadratic function in contrast to GD's 1D attraction region.*

4 Convergence Rate of the modified LSGD

In this section, we will discuss the convergence rate of LSGD with non-constant σ for ℓ-smooth functions, which follows the standard convergence analysis.

Definition 1. *A differentiable function f is ℓ-smooth (or ℓ-gradient Lipschitz) if all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ satisfy*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}) + \frac{\ell}{2}||\mathbf{x} - \mathbf{y}||^2.$$

Definition 2. *For a differentiable function f, we say that $\mathbf{x}$ is an ϵ-first-order stationary point if $||\nabla f(\mathbf{x})|| \leq \epsilon$.*

Theorem 2. *Assume that the function f is ℓ-smooth. Then, the modified LSGD with $\sigma(k) = \frac{k+1}{k+2}$, step size $\eta = \frac{1}{\ell}$ and termination condition $||\nabla f(\mathbf{x})|| \leq \epsilon$ converges to an ϵ-first-order stationary point within*

$$\frac{50\ell(f(\mathbf{x}^0) - f^*)}{9\epsilon^2}.$$

iterations where f^* denotes a local minimum of f.

Proof. First, we will establish an estimate for $f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k)$ for all $k \geq 0$. By the ℓ-smoothness of f and the LSGD scheme (7), we have

$$\begin{aligned}
&f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k) \\
&\leq \langle \nabla f(\mathbf{x}^k), \left(\mathbf{x}^{k+1} - \mathbf{x}^k\right) \rangle + \frac{\ell}{2}||\mathbf{x}^{k+1} - \mathbf{x}^k||_2^2 \\
&= \langle \nabla f(\mathbf{x}^k), -\frac{1}{\ell}\mathbf{A}_{\sigma(k)}^{-1} \nabla f(\mathbf{x}^k) \rangle + \frac{1}{2\ell}||\mathbf{A}_{\sigma(k)}^{-1} \nabla f(\mathbf{x}^k)||_2^2 \\
&= \frac{1}{2\ell} \left\| \left(\mathbf{A}_{\sigma(k)}^{-1} - \mathbf{I} \right) \nabla f(\mathbf{x}^k) \right\|_2^2 - \frac{1}{2\ell}||\nabla f(\mathbf{x}^k)||_2^2 \\
&\leq \frac{1}{2\ell} \left\| \mathbf{I} - \mathbf{A}_{\sigma(k)}^{-1} \right\|_2^2 ||\nabla f(\mathbf{x}^k)||_2^2 - \frac{1}{2\ell}||\nabla f(\mathbf{x}^k)||_2^2.
\end{aligned}$$

To estimate $\|\mathbf{I} - \mathbf{A}_{\sigma(k)}^{-1}\|$ note that $\mathbf{A}_{\sigma(k)}^{-1}$ is diagonalizable, i.e. there exist an orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$ and a diagonal matrix Λ with diagonal entries $\lambda_j(\mathbf{A}_{\sigma(k)}^{-1}) \in [1, 1 + 4\sigma(k)]$ such that $\mathbf{A}_{\sigma(k)}^{-1} = \mathbf{Q}^T \mathbf{\Lambda} \mathbf{Q}$. For $\sigma(k) = \frac{k+1}{k+2}$, we have

$$||\mathbf{I} - \mathbf{A}_{\sigma(k)}^{-1}||_2^2 = ||\mathbf{I} - \mathbf{\Lambda}||_2^2 \leq \left(1 - \frac{1}{1 + 4\sigma(k)}\right)^2 \leq \frac{16}{25}.$$

Plugging this estimate into the previous estimate yields

$$f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k) \leq -\frac{9}{50\ell}||\nabla f(\mathbf{x}^k)||_2^2.$$

Based on the above estimate the function value of the iterates decays by at least $\frac{9}{50\ell}||\nabla f(\mathbf{x}^k)||_2^2 \geq \frac{9}{50\ell}\epsilon^2$ in each iteration before an an ϵ-first-order stationary point with function value f^* is reached. After $\frac{50\ell(f(\mathbf{x}^0) - f^*)}{9\epsilon^2}$ iterations, LSGD is guaranteed to converge to an ϵ-first-order stationary point with function value f^*.

□

The above convergence rate for non-convex optimization is consistent with the gradient descent [19].

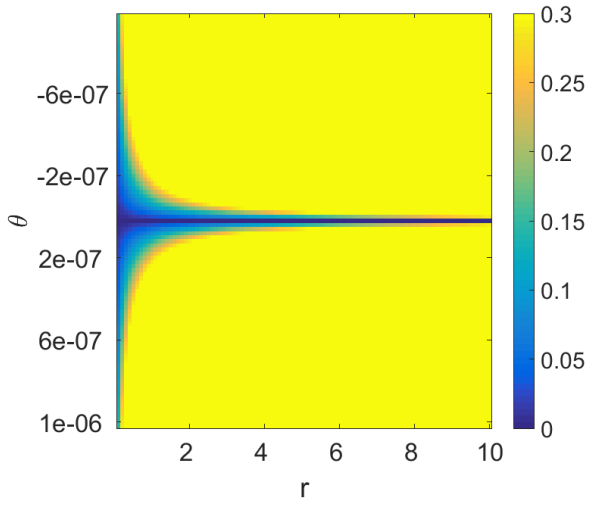
5 Numerical Examples

We run 100 iterations of both GD and LSGD with learning rate $\eta = 0.1$ for the function $f(x_1, x_2) = \frac{1}{2}\left(x_1^2 - x_2^2\right)$. For GD, the attraction region is $\{(x_1, x_2) | x_1 \in \mathbb{R}, x_2 = 0\}$. To demonstrate GD's behavior in terms of saddle points, we start GD from any point in the set $\{(x_1, x_2) | x_1 = r\cos\theta, x_2 = r\sin\theta, r \in [0.1, 10], \theta \in [-1\mathrm{e}-6^\circ, 1\mathrm{e}-6^\circ)\}$, with a grid spacing of 0.1 and $2\mathrm{e}-8^\circ$ for r and θ, respectively. As shown in Fig. 2(a), the distance to the saddle point $(0, 0)$ is 0 after 100 iterations of GD for any starting point with $\theta = 0$. For starting points with small r and θ, the iterates get very close to the saddle point with distances of less than 0.1 after 100 GD iterations.

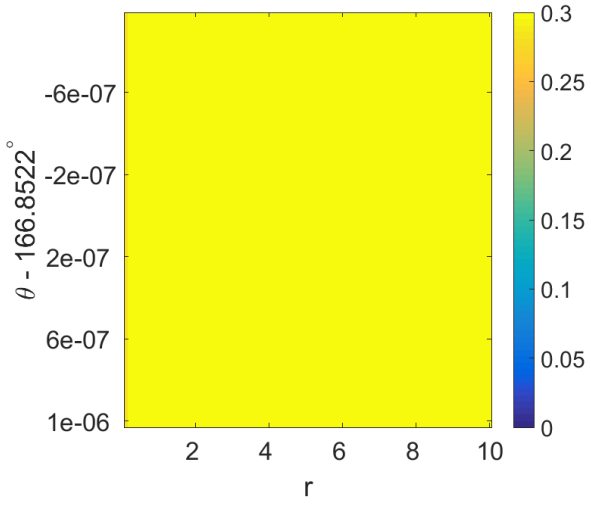
For LSGD, the attraction region of the saddle point $(0, 0)$ is of dimension 0 for the 2D problem, see Theorem 1. To verify this, we consider the starting points in $\{(x_1, x_2) | x_1 = r\cos\theta, x_2 =$

$r \sin\theta, r \in [0.1, 1], \theta \in [-180°, 180°)\}$ with a grid spacing of 0.1 and $1e-6°$, respectively, for r and θ. We observe that the minimum distance to $(0,0)$ is achieved when we initially start with $r = 0.1$ and $\theta = 166.8522°$. Then, we perform a finer grid search around $\theta = 166.8522°$ with grid spacing $\Delta\theta = 2e-8°$. Figure. 2(b) shows the same region as in (a), but with θ centered at $166.8522°$. After 100 LSGD iterations, the iterates do not converge to the saddle point $(0,0)$. If $r = 0.1$ the distance to the saddle point is less than 0.3, while the distance is larger than 0.3 for any other starting point.

These numerical results verify that for the 2D problem the modified LSGD does not converge to the saddle point for any starting point. Moreover, there is a region of starting points, resulting in a slow escape from the saddle points for GD, while this slow-down does not occur for LSGD. These results are consistent with the dimension $\lfloor (n-1)/2 \rfloor = 0$ of the attraction region for the modified LSGD according to Theorem 1.



(a)



(b)

Figure 2: Plot of distant field to the saddle point $(0,0)$ after running 100 iterations of GD (a) and LSGD (b) with learning rate $\eta = 0.1$ for the function $f(x_1, x_2) = x_1^2 - x_2^2$. The coordinates of each pixel denote starting points. For both (a) and (b) we show the same region of starting points for GD and LSGD, and the distances to the saddle point after 100 iterations.

6 Concluding Remarks

In this paper, we present a simple modification of the Laplacian smoothing gradient descent (LSGD) to avoid saddle points. For a class of quadratic functions, we prove that LSGD has a significantly lower dimensional attraction region than GD. To the best of our knowledge, our algorithm is the first one that only uses first-order information without any perturbation or noise to avoid saddle points deterministically. Since our approach is completely different to existing perturbed or noisy gradient-based approaches for avoiding saddle points, it is of great interest to investigate the efficacy of a combination of these approaches, such as Laplacian smoothing perturbed/noisy gradient descent, in terms of escaping and circumventing saddle points.

Due to additional technical difficulties, we only prove the efficacy in avoiding saddle points for a class of quadratic functions, giving valuable insights into the modified LSGD. Currently, we prepare a generalization of our analysis to all quadratic and even more general functions. It is worth mentioning that the current version of LSGD fails to reduce the dimension of the attraction region for the function $f(x_1, x_2) = x_1 x_2$ since the attraction region has the same values for x_1 and x_2 and hence, LSGD cannot perturb the eigenvectors to avoid saddle points. A possible way to fix this issue is to perform perturbations of the Laplacian.

Acknowledgments

This material is based on research sponsored by the Air Force Research Laboratory and DARPA under agreement number FA8750-18-2-0066, by the U.S. Department of Energy, Office of Science and by National Science Foundation, under Grant Numbers DOE-SC0013838 and DMS-1554564, (STROBE), and by the NSF DMS-1737770 and the Simons foundation. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. LMK was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/L016516/1, the RISE project NoMADS, and the German National Academic Foundation (Studienstiftung des Deutschen Volkes).

References

[1] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma. Finding approximate local minima faster than gradient descent. In *Symposium on Theory of Computing (STOC 2017)*, 2017.

[2] Y. Bengio. Learning deep architectures for ai. *Foundations and trends in Machine Learning*, 2(1), 2009.

[3] Y. Carmon and J. C. Duchi. Gradient descent efficiently finds the cubic-regularized non-convex newton step. *ArXiv:1612.00547*, 2016.

[4] C. Cartis, N. Gould, and P. L. Toint. On the complexity of steepest descent and regularized newtonâĂŹs methods for nonconvex unconstrained optimization problems. *Siam journal on optimization*, 20(6), 2010.

[5] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun. The loss surface of multi-layer networks. *ArXiv:1412.0233*, 2014.

[6] F. E. Curtis, D. P. Robinson, and M. Samadi. A trust region algorithm with a worst-case iteration complexity of $\mathcal{O}(\epsilon^{-3/2})$ for nonconvex optimization. *Mathematical Programming*, 2014.

[7] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems (NIPS 2016)*, 2014.

[8] S. Du, C. Jin, J. D. Lee, M. I. Jordan, B. Poczos, and A. Singh. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems (NIPS 2017)*, 2017.

[9] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points – online stochastic gradient for tensor decomposition. In *Conference on Learning Theory (COLT 2015)*, 2015.

[10] R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *ArXiv:1704.00708*, 2017.

[11] R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems (NIPS 2016)*, 2016.

[12] P. Jain, C. Jin, S. M. Kakade, and P. Netrapalli. Computing matrix square root via non convex local search. *ArXiv:1507.05854*, 2015.

[13] C. Jin, R. Ge, P. Netrapalli, S. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, 2017.

[14] C. Jin, P. Netrapalli, and M. I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference on Learning Theory (COLT 2018)*, 2018.

[15] K. Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems (NIPS 2016)*, 2016.

[16] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent converges to minimizers. In *Conference on Learning Theory (COLT 2016)*, 2016.

[17] K. Y. Levy. The power of normalization: Faster evasion of saddle points. *ArXiv:1611.04831*, 2016.

[18] S. Mei, T. Misiakiewicz, A. Montanari, and R. I. Oliveira. Solving sdps for synchronization and maxcut problems via the grothendieck inequality. In *Conference on Learning Theory (COLT 2017)*, 2017.

[19] Y. Nesterov. Introductory lectures on convex programming volume i: Basic course. *Lecture Notes*, 1998.

[20] Y. Nesterov and B. T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1), 2006.

[21] S. Nhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. *ArXiv:1605.07221*, 2016.

[22] S. Osher, B. Wang, P. Yin, X. Luo, M. Pham, and A. Lin. Laplacian smoothing gradient descent. *ArXiv:1806.06317*, 2018.

[23] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. *ArXiv:1609.03240*, 2016.

[24] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 1998.

[25] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2), 2016.

[26] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. In *In Information Theory (ISIT), 2016 IEEE International Symposium on. IEEE*, 2016.