Efficient Zeroth-Order Optimization via Mirror Natural Evolution Strategies: Bridging Gradient-Free and Hessian-Based Methods

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

September 21, 2024

Abstract

We propose a novel approach to zeroth-order optimization by developing the Mirror Natural Evolution Strategy (`MES`), a method that leverages mirror descent to efficiently approximate both gradients and Hessians using only function value queries. Our key innovation is the introduction of a regularized objective function that ensures the covariance matrix converges to the inverse of the Hessian, achieving query-efficient optimization with explicit convergence rates. This framework outperforms existing zeroth-order algorithms by incorporating curvature information without additional query overhead. We provide a rigorous convergence analysis for `MES`, demonstrating an optimal convergence rate of $O(1/k)$ for the covariance matrix. Our results offer new insights into the relationship between natural evolution strategies and derivativefree methods, establishing `MES` as a powerful tool for high-dimensional and noisy optimization problems.

Keywords: Zeroth-Order Optimization; Natural Evolution Strategies (NES); Mirror Descent; Covariance Matrix Adaptation (CMA-ES); Hessian Approximation; Convergence Rate; Query-Efficient Algorithms

1 Introduction

Zeroth-order optimization has become an essential tool in many modern machine learning applications, particularly when the gradient information is either unavailable or expensive to compute. This scenario is prevalent in deep reinforcement learning, hyperparameter tuning for deep networks, and black-box adversarial attacks on neural networks. Traditional zeroth-order methods approximate gradients using only function evaluations, but often struggle with high query complexity and suboptimal convergence rates.

Recent work has explored the use of second-order information to enhance convergence speed. For example, Covariance Matrix Adaptation Evolution Strategies (CMA-ES) and Natural Evolution Strategies (NES) approximate both the gradient and Hessian to accelerate the optimization process. However, obtaining explicit convergence rates, particularly for the covariance matrix, remains challenging in practical NES algorithms, which rely on zeroth-order information. This gap motivates our work.

In this paper, we propose a novel method called *Mirror Natural Evolution Strategy (`MES`)*. Our approach introduces a regularized objective function that allows us to use mirror descent to update the covariance matrix, thereby approximating the Hessian more efficiently. `MES` provides explicit convergence rates for both the objective function and the covariance matrix, offering a new perspective on the use of second-order information in zeroth-order optimization.

Backgrounds. Zeroth-order optimization has a wide range of applications, particularly in machine learning, including deep reinforcement learning [Salimans et al.(2017), Conti et al.(2018)], black-box adversarial attacks on deep neural networks [Ilyas et al.(2018), Tu et al.(2019), Dong et al.(2019), Chen et al.(2019)], and hyperparameter tuning for deep neural networks [Loshchilov and Hutter(2016)]. Simultaneously, the theoretical foundations of zeroth-order optimization have been extensively studied [Nesterov and Spokoiny(2017), Ghadimi and Lan(2013), Duchi et al.(2015)]. It is well-known that the second-order information is helpful to improve the convergence rate of algorithms. A natural question to ask is: ***Can we approximate the Hessian (inverse) by the zeroth-order queries?*** Evolution Strategies such as `CMA-ES` (covariance matrix adaptation evolution strategy) and `NES` (natural evolution strategy) use the zeroth-order information to approximate both the gradient and Hessian matrix. Thus, in practice, `CMA-ES` and `NES` outperform the algorithms which only use the zeroth-order information to approximate the gradient [Hansen(2016)]. However, to obtain the explicit convergence rates of this kinds of algorithms, especially the convergence rates of the covariance matrix, are difficult.

[Wierstra et al.(2014)] propose a reparameterized function $J(\theta)$ with parameters $\theta = (\mu, \Sigma)$. Based on it, [Akimoto(2012)] tries to derive the convergence rates of the covariance matrix of `NES`. However, in the analysis, it requires to use the gradient of $J(\theta)$ with respect to the covariance matrix but not the approximate one which is used in the practical `NES`-type algorithms by zeroth-order queries. Furthermore, the reparameterized function $J(\theta)$ proposed in [Wierstra et al.(2014)] is not good enough to help design and understand zeroth-order algorithms with the both gradient and Hessian being approximated by the zeroth-order queries. For the quadratic function, the derivative of $J(\theta)$ with respect to Σ equals to the Hessian. Thus, we can not obtain the useful information by the first order necessary condition of optimization. What we want to is the solution to the equation $\frac{\partial J(\cdot)}{\partial \Sigma} = 0$ equals to the Hessian inverse.

Thus, it remains open to propose a reparameterized function who can help design and understand zeroth-order algorithms with the both gradient and Hessian being approximated by the zeroth-order queries. It is also an open question whether one can design a zeroth-order algorithm with explicit convergence rates both for the function value and the covariance matrix, which only uses the zeroth-order information to approximate both the gradient and Hessian matrix. In this paper, we give affirmative answers to above open questions and summarize our contributions as follows:

(1) Based on the reparameterized objective function $J(\theta)$ proposed in [Wierstra et al.(2014)], we improve it and propose a regularized objective function $Q_\alpha(\theta)$. We show that the mean vector and the covariance part of the minimizer of $Q_\alpha(\theta)$ are close to the minimum of the objective function and close to the Hessian inverse up to some perturbations up to at most $\mathcal{O}(\alpha)$. These properties of $Q_\alpha(\theta)$ make the mathematical foundation for designing query efficient algorithm.

(2) Based on this new objective function, we propose a novel algorithm called `MES`, which guarantees that the covariance matrix converges to the inverse of the Hessian. Instead of using natural gradient descent, we resort to the mirror descent to update the covariance matrix.

(3) We provide a convergence analysis of `MES` and give the explicit convergence rates of `MES`. Especially, we show the inverse of covariance matrix will converge to the Hessian with a rate $\widetilde{\mathcal{O}}(1/k)$. To the best of our knowledge, this is the first explicit convergence rate for the covariance matrix in this kind of algorithms. Our convergence analysis also shows how the covariance matrix helps the algorithm converge to the minimizer of the objective function.

More Related Literature. Zeroth-order (derivative-free) optimization has a long history [Matyas(1965), Kiefer and Wolfowitz(1952)]. The idea behind these zeroth-order algorithms is to create a stochastic oracle to approximate (first-order) gradients using (zeroth-order) function value difference at a random direction, and then apply the update rule of (sub-)gradient descent [Nesterov and Spokoiny(2017), Ghadimi and Lan(2013), Duchi et al.(2015)]. On the other hand, [Conn et al.(2009)] propose to utilize curvature information in constructing quadratic approximation model under a slightly modified trust region regime. Recently, [Ye et al.(2018)] propose the `ZOHA` algorithm which utilizes curvature information of the objective function. They show that the approximate Hessian can help to improve the convergence rate over zeroth-order algorithms without second-order information. However, the work of [Ye et al.(2018)] requires extra queries to construct the approximate Hessian. In contrast, no extra queries are needed to approximate the Hessian in our work. Thus, our `MES` is more query-efficient.

Evolutionary strategies (`ES`) are another important class of zeroth-order algorithms for optimization problems that only have access to function value evaluations. ES attracts much attention since it was introduced by Ingo Rechenberg and Hans-Paul Schwefel in the 1960s and 1970s [Schwefel(1977)], and many variants have been proposed [Beyer and Deb(2001), Hansen and Ostermeier(2001), Wierstra et al.(2008), Glasmachers et al.(2010)]. `ES` tries to evaluate the fitness of real-valued genotypes in batches, after which only the best genotypes are kept and used to produce the next batch of offsprings. A covariance matrix is incorporated into evolutionary strategies to capture the dependency variables so that independent 'mutations' can be generated for the next generation. In this general algorithmic framework, the most well-known algorithms are the covariance matrix adaptation evolution strategy (`CMA-ES`) [Hansen and Ostermeier(2001)] and natural evolution strategies (`NES`) [Wierstra et al.(2008)].

There have been many efforts to better understand these methods in the literature [Beyer(2014), Ollivier et al.(2017), Auger and Hansen(2016), Malagò and Pistone(2015), Akimoto et al.(2010)]. For example, the authors in [Akimoto et al.(2010)] revealed a connection between `NES` and `CMA-ES`, and showed that `CMA-ES` is a special version of `NES`. `NES` can be derived from information geometry because it employs a natural gradient descent [Wierstra et al.(2014)]. Therefore, the convergence of `NES` and `CMA-ES` has been studied from the information geometry point of view [Beyer(2014), Auger and Hansen(2016)]. [Beyer(2014)] analyzed the convergence of `NES` with infinitestimal learning rate using ordinary differential equation with the objective function being quadratic and strongly convex. However, it does not directly lead to a convergence rate result with finite learning rate. Moreover, [Beyer(2014)] does not show how the covariance matrix converges to the inverse of the Hessian, although this is conjectured in [Hansen(2016)]. [Akimoto(2012)] proves that the covariance matrix of `NES` will converge to the Hessian inverse with a geometric convergence rate under the condition that `NES` takes a natural gradient. However, in practice, `NES` takes a stochastic natural gradient to update the covariance matrix. In contrast, our work gives an explicit convergence rate of the covariance matrix update of `MES` using the stochastic mirror gradient descent obtained by zeroth-order queries. [Glasmachers and Krause(2022)] show that the covariance matrix of Hessian Estimation Evolution Strategies converge to the Hessian inverse but no explicit rate is provided.

Recently, [Auger and Hansen(2016)] showed that for strongly convex functions, the estimated mean value in ES with comparison-based step-size adaptive randomized search (including `NES` and `CMA-ES`) can achieve linear convergence. However, [Auger and Hansen(2016)] have not shown how covariance matrix converges and how covariance matrix affects the convergence properties of the estimated mean vector. Although it has been conjectured that for the quadratic function, the covariance matrix of `CMA-ES` will converge to the inverse of the Hessian up to a constant factor, no rigorous proof has been provided [Hansen(2016)].

The derivative free algorithms and evolutionary strategies are closely related even they are motivated from different ideas and are of totally different forms. For example, to improve the convergence rate of `NES`, [Salimans et al.(2017)] proposed 'antithetic sampling' technique. In this case, `NES` shares the same algorithmic form with derivative free algorithm [Nesterov and Spokoiny(2017)]. `NES` with 'antithetic sampling' is widely used in black-box adversarial attack [Tu et al.(2019), Ilyas et al.(2018)]. Nevertheless, the mathematical relationship between `NES` and derivative free algorithms have not been explored.

Organization. The rest of this paper is organized as follows. In Section 2, we review the relevant background on zeroth-order optimization and natural evolution strategies. Section 3 introduces our novel regularized objective function and demonstrates that its minimizer's mean and covariance are close to the original problem's minimizer and Hessian inverse, respectively. In Section 4, we present the detailed description of the `MES` algorithm and the mirror descent approach for updating the covariance matrix. Section 5 provides a thorough convergence analysis of `MES`. Finally, Section 6 concludes the paper and discusses potential future directions. Detailed proofs are deferred to the appendix.

2 Background and Preliminaries

In this section, we will introduce the natural evolutionary strategies and preliminaries.

2.1 Natural Evolutionary Strategies

The Natural Evolutionary Strategies (`NES`) reparameterize the objective function $f(z)$ ($z \in \mathbb{R}^d$) as follows [Wierstra et al.(2014)]:

$$J(\theta) = \mathbb{E}_{z\sim\pi(\cdot|\theta)}[f(z)] = \int f(z)\pi(z|\theta)\,dz \tag{1}$$

where θ denotes the parameters of density $\pi(z|\theta)$ and $f(z)$ is commonly referred as the fitness function for samples z. Such transformation can help to develop algorithms to find the minimum of $f(z)$ by only accessing to the function value.

Gaussian Distribution and Search Directions. In this paper, we will only investigate the Gaussian distribution, that is,

$$\pi(z|\theta) \sim N(\mu, \bar{\Sigma}) \tag{2}$$

Accordingly, we have

$$z = \mu + \bar{\Sigma}^{1/2}u, \quad u \sim N(0, I_d) \tag{3}$$

where d is the dimension of z. Furthermore, the density function $\pi(z|\theta)$ can be presented as

$$\pi(z|\theta) = \frac{1}{\sqrt{(2\pi)^d \det(\bar{\Sigma})}} \cdot \exp\left(-\frac{1}{2}(z-\mu)^\top \bar{\Sigma}^{-1}(z-\mu)\right)$$

In order to compute the derivatives of $J(\theta)$, we can use the so-called 'log-likelihood trick' to obtain the following [Wierstra et al.(2014)]

$$\nabla_\theta J(\theta) = \nabla_\theta \int f(z)\pi(z|\theta)\, dz = \mathbb{E}_z\big[f(z)\nabla_\theta \log \pi(z|\theta)\big] \tag{4}$$

We also have that

$$\log \pi(z|\theta) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log\det\bar{\Sigma} - \frac{1}{2}(z-\mu)^\top \bar{\Sigma}^{-1}(z-\mu)$$

We will need its derivatives with respect to μ and $\bar{\Sigma}$, that is, $\nabla_\mu \log \pi(z|\theta)$ and $\nabla_{\bar{\Sigma}} \log \pi(z|\theta)$. The first is trivially

$$\nabla_\mu \log \pi(z|\theta) = \bar{\Sigma}^{-1}(z-\mu) \tag{5}$$

while the latter is

$$\nabla_{\bar{\Sigma}} \log \pi(z|\theta) = \frac{1}{2}\bar{\Sigma}^{-1}(z-\mu)(z-\mu)^\top\bar{\Sigma}^{-1} - \frac{1}{2}\bar{\Sigma}^{-1} \tag{6}$$

Let us denote

$$\theta = [\mu^\top,\ \mathrm{vec}(\bar{\Sigma})^\top]^\top$$

where $\theta \in \mathbb{R}^{d(d+1)}$-dimensional column vector consisting of all the elements of the mean vector μ and the covariance matrix $\bar{\Sigma}$. $\mathrm{vec}(\cdot)$ denotes a rearrangement operator from a matrix to a column vector. The Fisher matrix with respect to θ of $\pi(z|\theta)$ for a Gaussian distribution is well-known [Akimoto et al.(2010)],

$$F_\theta = \mathbb{E}_z\left[\nabla_\theta \log \pi(z|\theta)\nabla_\theta \log \pi(z|\theta)^\top\right] = \begin{bmatrix} \bar{\Sigma}^{-1}, & 0 \\ 0, & \frac{1}{2}\bar{\Sigma}^{-1}\otimes\bar{\Sigma}^{-1} \end{bmatrix}$$

where $\otimes$ is the Kronecker product. Therefore, the natural gradient of the log-likelihood of $\pi(z|\theta)$ is

$$F_\theta^{-1}\nabla_\theta \log \pi(z|\theta) = \begin{bmatrix} z-\mu \\ \mathrm{vec}\left((z-\mu)(z-\mu)^\top - \bar{\Sigma}\right) \end{bmatrix} \tag{7}$$

Combining Eqn. (4) and (7), we obtain the estimate of the natural gradient from samples $z_1, \ldots, z_b$ as

$$F_\theta^{-1}\nabla J(\theta) \approx \frac{1}{b}\sum_{i=1}^{b} f(z_i)\begin{bmatrix} z_i-\mu \\ \mathrm{vec}\left((z_i-\mu)(z_i-\mu)^\top - \bar{\Sigma}\right) \end{bmatrix} \tag{8}$$

Therefore, we can obtain the meta-algorithm of `NES` (Algorithm 3 of [Wierstra et al.(2014)] with $F_\theta^{-1}\nabla J(\theta)$ approximated as Eqn. (8))

$$\begin{cases} \mu = \mu - \eta\cdot\dfrac{1}{b}\displaystyle\sum_{i=1}^{b} f(\mu+\bar{\Sigma}^{1/2}u_i)\bar{\Sigma}^{1/2}u_i \\ \bar{\Sigma} = \bar{\Sigma} - \eta\cdot\dfrac{1}{b}\displaystyle\sum_{i=1}^{b} f(\mu+\bar{\Sigma}^{1/2}u_i)\left(\bar{\Sigma}^{1/2}u_iu_i^\top\bar{\Sigma}^{1/2} - \bar{\Sigma}\right) \end{cases}$$

where η is the step size.

2.2 Notions

Now, we introduce some important notions which is widely used in optimization.

L-smooth A function $f(\mu)$ is *L-smooth*, if it holds that, for all $\mu_1, \mu_2 \in \mathbb{R}^d$

$$\|\nabla f(\mu_1) - \nabla f(\mu_2)\| \le L \|\mu_1 - \mu_2\| \tag{9}$$

σ-Strong Convexity A function $f(\mu)$ is *σ-strongly convex*, if it holds that, for all $\mu_1, \mu_2 \in \mathbb{R}^d$

$$f(\mu_1) - f(\mu_2) \ge \langle \nabla f(\mu_2), \mu_1 - \mu_2 \rangle + \frac{\sigma}{2} \|\mu_1 - \mu_2\|^2 \tag{10}$$

γ-Lipschitz Hessian A function $f(\mu)$ *admits γ-Lipschitz Hessians* if it holds that, for all $\mu_1, \mu_2 \in \mathbb{R}^d$, it holds that

$$\left\|\nabla^2 f(\mu_1) - \nabla^2 f(\mu_2)\right\| \le \gamma \|\mu_1 - \mu_2\| \tag{11}$$

Note that L-smoothness and σ-strongly convexity imply $\sigma I \preceq \nabla^2 f(\mu) \preceq LI$.

3 Regularized Objective Function

Conventional `NES` algorithms are going to minimize $J(\theta)$ (Eqn. (1)) [Wierstra et al.(2014)]. Instead, we propose a novel regularized objective function to reparameterize $f(z)$:

$$Q_\alpha(\theta) = J(\theta) - \frac{\alpha^2}{2} \log\det\Sigma \tag{12}$$

where α is a positive constant. Furthermore, we represent $\bar{\Sigma}$ in Eqn. (2) as $\bar{\Sigma} = \alpha^2\Sigma$. Accordingly, $\theta(\mu, \Sigma)$ denotes the parameters of a Gaussian density $\pi(z|\theta) = N(\mu, \alpha^2\Sigma)$. By such transformation, $J(\theta)$ can be represent as

$$J(\theta) = \mathbb{E}_u[f(\mu + \alpha\Sigma^{1/2}u)], \quad \text{with} \quad u \sim N(0, I_d)$$

Then $J(\theta)$ is the *Gaussian approximation function* of $f(z)$ and α plays a role of smoothing parameter [Nesterov and Spokoiny(2017)]. Compared with $J(\theta)$, $Q_\alpha(\theta)$ has several advantages and we will first introduce the intuition we propose $Q_\alpha(\theta)$.

Intuition Behind $Q_\alpha(\theta)$ Introducing the regularization brings an important benefit which can help to clarify the minimizer of Σ. This benefit can be shown when $f(z)$ is a quadratic function where $f(z)$ can be expressed as

$$f(z) = f(\mu) + \langle \nabla f(\mu), z - \mu \rangle + \frac{1}{2}(z - \mu)^\top H(z - \mu) \tag{13}$$

where $H = \nabla^2 f(\mu)$ denotes the Hessian matrix. Note that when $f(z)$ is quadratic, the Hessian matrix is independent on the value of z. In the rest of this paper, we will use H to denote the Hessian matrix of a quadratic function. Since we have $z = \mu + \alpha\Sigma^{1/2}u$ (by Eqn. (3)), $J(\theta)$ can be explicitly expressed as

$$J(\theta) = \mathbb{E}_u\left[f(\mu) + \alpha \langle \nabla f(\mu), \Sigma u \rangle + \frac{\alpha^2}{2} u^\top \Sigma^{1/2} H \Sigma^{1/2} u\right] = f(\mu) + \frac{\alpha^2}{2} \langle H, \Sigma \rangle \tag{14}$$

where $\langle A, B\rangle = \mathrm{tr}(A^\top B)$. By setting $\nabla_\theta Q_\alpha(\theta) = 0$, we can obtain that

$$\frac{\partial Q_\alpha}{\partial \mu} = \nabla_\mu f(\mu) = 0, \qquad \frac{\partial Q_\alpha}{\partial \Sigma} = \frac{\alpha^2}{2} H - \frac{\alpha^2}{2}\Sigma^{-1} = 0$$

Thus, we can obtain that the minimizer μ of Q_α is μ^* — the minimizer of $f(\mu)$ and the minimizer Σ of Q_α is H^{-1} — the inverse of the Hessian matrix. In contrast, without the regularization, $\frac{\partial Q_\alpha}{\partial \Sigma}$ will reduce to $\frac{\partial J(\theta)}{\partial \Sigma}$:

$$\frac{\partial J(\theta)}{\partial \Sigma} = \frac{\alpha^2}{2} H$$

Thus, the $\partial J(\theta)/\partial\Sigma$ does not provide useful information about what covariance matrix is the optimum of $J(\theta)$.

Therefore, in this paper, we will consider the regularized objective function $Q_\alpha(\theta)$. To obtain a concise theoretical analysis of the convergence rate of Σ, we are going to solve the following constrained optimization problem

$$\min_{\mu\in\mathbb{R}^d, \Sigma\in\mathcal{S}} Q_\alpha(\theta(\mu,\Sigma)) \tag{15}$$

with $\mathcal{S}$ is defined as

$$\mathcal{S} = \left\{\Sigma \middle| \zeta^{-1}\cdot I \preceq \Sigma \preceq \tau^{-1}\cdot I\right\} \tag{16}$$

where ζ and τ are positive constants which satisfy $\tau \le \zeta$. Note that, the constraint on Σ is used to keep Σ bounded and this property will be used in the convergence analysis of Σ.

In the rest of this section, we show that the mean vector $\hat{\mu}^*$ of minimizer of Eqn. (15) is close to μ^* which is the minimizer of $f(\mu)$. Furthermore, if ζ and τ in Eqn. (16) are chosen properly, the covariance matrix $\hat{\Sigma}_*$ will be close to the Hessian inverse. As a special case, if $f(\cdot)$ is quadratic, then we show that $\hat{\mu}^*$ is equal to μ^* and $\hat{\Sigma}_*$ is equal to $[\nabla^2 f(\mu^*)]^{-1}$.

3.1 Quadratic Case

We will first investigate the case that function $f(\cdot)$ is quadratic, because the solution in this case is considerably simple.

Theorem 1. *If the function $f(\cdot)$ is quadratic so that $f(z)$ satisfies Eqn.* (13)*. Define $\Pi_{\mathcal{S}}(A)$ be the projection of symmetric A on to $\mathcal{S}$, that is, $\Pi_{\mathcal{S}}(A) = \mathrm{argmin}_{X\in\mathcal{S}}\|A - X\|$ with $\|\cdot\|$ being the Frobenius norm. Let μ^* be the minimizer of $f(\cdot)$, then the minimizer of $Q_\alpha(\theta)$ is*

$$\left(\mu^*, \Pi_{\mathcal{S}}\left(H^{-1}\right)\right) = \operatorname*{argmin}_{\mu,\Sigma\in\mathcal{S}} Q_\alpha(\mu,\Sigma)$$

If $\zeta \ge L$ and $\tau \le \sigma$ in $\mathcal{S}$ (defined in Eqn. (16)), we can observe that the above proposition shows that the optimal covariance matrix Σ is the Hessian inverse matrix.

3.2 General Strongly Convex Function with Smooth Hessian

Next we will consider the general convex function case with its Hessian being γ-Lipschitz continuous. First, we can rewrite $Q_\alpha(\theta)$ as

$$Q_\alpha(\theta) = J(\theta) + R(\Sigma), \quad \text{with} \quad R(\Sigma) = -\frac{\alpha^2}{2}\log\det\Sigma$$

We will show that the value $Q_\alpha(\theta)$ at $\theta = (\mu, \Sigma)$ is close to $f(\mu)$ when α is sufficiently small.

Theorem 2. *Let $Q_\alpha(\theta)$ be defined in Eqn. (12). Assuming that function $f(\cdot)$ is L-smooth and its Hessian is γ-Lipschitz, then $Q_\alpha(\theta)$ satisfies that*

$$f(\mu) - \frac{L\alpha^2}{2}\mathrm{tr}(\Sigma) - \alpha^3\phi(\Sigma) + R(\Sigma) \le Q_\alpha(\theta) \le f(\mu) + \frac{L\alpha^2}{2}\mathrm{tr}(\Sigma) + \alpha^3\phi(\Sigma) + R(\Sigma)$$

where $\phi(\Sigma)$ is defined as

$$\phi(\Sigma) = \frac{1}{(2\pi)^{d/2}}\int_u \frac{\gamma}{6}\|\Sigma^{1/2}u\|^3 \exp\left(-\frac{1}{2}\|u\|^2\right) du$$

By the above proposition, we can observe that as $\alpha \to 0$, $\frac{L\alpha^2}{2}\mathrm{tr}(\Sigma)$, $\phi(\Sigma)$ and $R(\Sigma)$ will go to 0. Therefore, instead to directly minimizing $f(\mu)$, we can minimize $Q_\alpha(\theta)$ with $\theta = (\mu, \Sigma)$. Next, we will prove that the μ part of the minimizer of $Q_\alpha(\theta)$ is also close to the solver of $\min f(\mu)$.

Theorem 3. *Let $f(\cdot)$ satisfy the properties in Theorem 2. $f(\cdot)$ is also σ-strongly convex. Let $\hat{\mu}_*$ be the minimizer of $Q_\alpha(\theta)$ under constraint $\Sigma \in \mathcal{S}$. μ^* denotes the optimum of $f(\mu)$. Then, we have the following properties*

$$\begin{aligned} f(\hat{\mu}_*) - f(\mu^*) \le& \frac{dL\alpha^2}{\tau} + \frac{\gamma\alpha^3(d+3)^{3/2}}{3\tau^{3/2}} + \frac{d\alpha^2}{2}\left(\tau^{-1} - \zeta^{-1}\right) \\ \|\mu^* - \hat{\mu}_*\|^2 \le& \frac{2dL\alpha^2}{\sigma\tau} + \frac{2\gamma\alpha^3(d+3)^{3/2}}{3\sigma\tau^{3/2}} + \frac{d\alpha^2}{\sigma}\left(\tau^{-1} - \zeta^{-1}\right) \end{aligned}$$

where d is the dimension of μ.

Finally, we will provide the properties how well Σ approximates the inverse of Hessian matrix.

Theorem 4. *Let $f(\cdot)$ satisfy the properties in Theorem 3. $\hat{\theta} = (\hat{\mu}^*, \hat{\Sigma}_*)$ is the minimizer of $Q_\alpha(\theta)$ under the constraint $\Sigma \in \mathcal{S}$. μ^* is the minimizer of $f(\mu)$ and Σ_* is the minimizer of $Q_\alpha(\theta)$ given $\mu = \mu^*$ under the constraint $\Sigma \in \mathcal{S}$. Assume that α satisfies*

$$\alpha \le \frac{3\tau^{3/2}\sigma}{\gamma\zeta} \cdot \left((d+5)^{5/2} + d(d+3)^{3/2}\right)^{-1} \tag{17}$$

where d is the dimension of μ, then Σ_* has the following properties

$$\left\|\hat{\Sigma}_* - \Pi_{\mathcal{S}}\left(\left(\nabla^2 f(\hat{\mu}^*)\right)^{-1}\right)\right\| \le \frac{\alpha\gamma\zeta}{3\tau^{3/2}\sigma^2} \cdot \left((d+5)^{5/2} + d(d+3)^{3/2}\right)$$

and

$$\begin{aligned} \left\|\Sigma_* - \hat{\Sigma}_*\right\| \le& \frac{2\alpha\gamma\zeta}{3\tau^{3/2}\sigma^2} \cdot \left((d+5)^{5/2} + d(d+3)^{3/2}\right) \\ &+ \frac{\gamma}{\sigma^2} \cdot \left(\frac{2dL\alpha^2}{\sigma\tau} + \frac{2\gamma\alpha^3(d+3)^{3/2}}{3\sigma\tau^{3/2}} + \frac{d\alpha^2}{\sigma}\left(\tau^{-1} - \zeta^{-1}\right)\right)^{1/2} \end{aligned}$$

where $\Pi_{\mathcal{S}}(\cdot)$ is the projection operator which projects a symmetric matrix on to $\mathcal{S}$ with Frobenius norm as distance measure.

Remark. Theorem 4 shows that when α is small, then $\hat{\Sigma}_*$ (the Σ minimizer of problem (15)) is close to $\Pi_{\mathcal{S}}\left(\left(\nabla^2 f(\mu^*)\right)^{-1}\right)$. If $\tau \le \sigma$ and $\zeta \ge L$, then $\hat{\Sigma}_*$ is close to the inverse of the Hessian at μ^*.

The above propositions show that if α is small, then $\left(\hat{\mu}^*, \hat{\Sigma}_*\right)$, which is the minimizer of problem 15, will be close to $\left(\mu^*, \Pi_{\mathcal{S}}\left(\left(\nabla^2 f(\mu^*)\right)^{-1}\right)\right)$. In the next section, we propose a mirror natural evolution strategy to minimize problem (15).

3.3 Proof of Theorem 1

The proof of Theorem 1 is similar to that of Proposition 2.

Proof of Theorem 1. By the definition of $Q_\alpha(\theta)$ and Eqn. (14), we have

$$Q_\alpha(\theta) = f(\mu) + \frac{\alpha^2}{2}\langle H, \Sigma\rangle - \frac{\alpha^2}{2}\log\det\Sigma$$

Then, taking partial derivative of Q_α with respect to μ, we can obtain that

$$\frac{\partial Q_\alpha(\theta)}{\partial\mu} = \nabla_\mu f(\mu).$$

By setting $\frac{\partial Q_\alpha(\theta)}{\partial\mu}$ to zero, we can obtain that Q_α attains its minimum at μ^*.

For the Σ part, we have the following Lagrangian [Lanckriet et al.(2004)],

$$L(\Sigma, A_1, A_2) = \frac{\alpha^2}{2}\langle H, \Sigma\rangle - \frac{\alpha^2}{2}\log\det\Sigma + \frac{\alpha^2}{2}\left\langle A_1, \Sigma - \tau^{-1}I\right\rangle + \frac{\alpha^2}{2}\left\langle A_2, \zeta^{-1}I - \Sigma\right\rangle$$

where A_1 and A_2 are two positive semi-definite matrices and H denotes the Hessian matrix of the quadratic function $f(\cdot)$. The $\partial L(\Sigma, A_1, A_2)/\partial X$ is

$$\frac{\partial L(\Sigma, A_1, A_2)}{\partial\Sigma} = \frac{\alpha^2}{2}\left(H - \Sigma^{-1} + A_1 - A_2\right)$$

By the general Karush-Kuhn-Tucker (KKT) condition [Lanckriet et al.(2004)], we have

$$\begin{aligned}
&\Sigma_*^{-1} = H + A_1 - A_2\\
&A_1\Sigma_* = \tau^{-1}A_1, \quad A_2\Sigma_* = \zeta^{-1}A_2\\
&A_1 \succeq 0, \quad A_2 \succeq 0
\end{aligned}$$

Since the optimization problem is strictly convex, there is a unique solution (Σ_*, A_1, A_2) that satisfy the above KKT condition. We construct such a solution as follows. Let $H = U\Lambda U^\top$ be the spectral decomposition of H, where Λ is diagonal and U is an orthogonal matrix. We define Σ_* as $\Sigma_* = U\bar{\Lambda}U^\top$, where $\bar{\Lambda}$ is a diagonal matrix with $\bar{\Lambda}_{i,i} = \tau^{-1}$ if $\Lambda_{i,i} \le \tau$, $\bar{\Lambda}_{i,i} = \zeta^{-1}$ if $\Lambda_{i,i} \ge \zeta$, and $\bar{\Lambda}_{i,i} = \Lambda_{i,i}^{-1}$ in other cases. A_1 and A_2 are defined as follows, where both $\Lambda_{i,i}^{(1)}$ and $\Lambda_{i,i}^{(2)}$ are diagonal matrices:

$$\begin{aligned}
A_1 =& U\Lambda^{(1)}U^\top \quad \text{with} \quad \Lambda_{i,i}^{(1)} = \max\{\tau - \Lambda_{i,i}, 0\}\\
A_2 =& U\Lambda^{(2)}U^\top \quad \text{with} \quad \Lambda_{i,i}^{(2)} = \max\{\Lambda_{i,i} - \zeta, 0\}
\end{aligned}$$

Next, we will check that A_1, A_2 and Σ_* satisfy the KKT's condition. First, we have

$$H + A_1 - A_2 = U\Lambda U^\top + U\Lambda^{(1)}U^\top - U\Lambda^{(2)}U^\top = U\bar{\Lambda}^{-1}U^\top = \Sigma_*^{-1}$$

Then we also have $A_1\Sigma_* = U\Lambda^{(1)}U^\top U\bar{\Lambda}U^\top = U\Lambda^{(1)}\bar{\Lambda}U^\top = \tau^{-1}U\Lambda^{(1)}U^\top = \tau^{-1}A_1$. Similarly, it also holds that $A_2\Sigma_* = \zeta^{-1}A_2$.

Finally, the construction of A_1 and A_2 guarantees these two matrix are positive semi-definite.

Therefore, Σ_* is the covariance part of the minimizer of $Q_\alpha(\theta)$, and $\left(\mu^*, \Pi_{\mathcal{S}}\left(H^{-1}\right)\right)$ is the optimal solution of $Q_\alpha(\theta)$ under constraint $\mathcal{S}$. $\square$

3.4 Proof of Theorem 2

Proof of Theorem 2. By the Taylor's expansion of $f(z)$ at μ, we have

$$\left| f(z) - \left[f(\mu) + \langle \nabla f(\mu), z - \mu \rangle + \frac{1}{2}(z-\mu)^\top \nabla^2 f(\mu)(z-\mu) \right] \right| \leq \frac{\gamma}{6}\|z-\mu\|_2^3. \tag{18}$$

By $z = \mu + \alpha\Sigma u$ with $u \sim N(0, I_d)$, we can upper bound the $J(\theta)$ as

$$\begin{aligned}
J(\theta) =& \frac{1}{(2\pi)^{d/2}} \int_u f(\mu + \alpha\Sigma^{1/2}u) \exp\left(-\frac{1}{2}\|u\|^2\right) du \\
\overset{(18)}{\leq}& \frac{1}{(2\pi)^{d/2}} \int_u \left(f(\mu) + \left\langle \nabla f(\mu), \alpha\Sigma^{1/2}u \right\rangle + \frac{\alpha^2}{2} u^\top \Sigma^{1/2} \nabla^2 f(\mu) \Sigma^{1/2} u + \frac{\gamma\alpha^3}{6}\|\Sigma^{1/2}u\|_2^3 \right) \exp\left(-\frac{1}{2}\|u\|^2\right) du \\
=& f(\mu) + \frac{\alpha^2}{2}\left\langle \nabla^2 f(\mu), \Sigma \right\rangle + \alpha^3 \phi(\Sigma) \\
\leq& f(\mu) + \frac{L\alpha^2}{2}\mathrm{tr}(\Sigma) + \alpha^3\phi(\Sigma)
\end{aligned}$$

where the last inequality is because of $\|\nabla^2 f(\mu)\|_2 \leq L$, we have $|\mathrm{tr}(\nabla^2 f(\mu)\Sigma)| \leq L\mathrm{tr}(\Sigma)$.

By the fact that $Q_\alpha(\theta) = J(\theta) + R(\Sigma)$, we have

$$Q_\alpha(\theta) \leq f(\mu) + \frac{L\alpha^2}{2}\mathrm{tr}(\Sigma) + \alpha^3\phi(\Sigma) + R(\Sigma)$$

Similarly, we can obtain that

$$J(\theta) \geq f(\mu) - \frac{L\alpha^2}{2}\mathrm{tr}(\Sigma) - \alpha^3\phi(\Sigma)$$

Therefore, we can obtain that

$$f(\mu) - \frac{L\alpha^2}{2}\mathrm{tr}(\Sigma) - \alpha^3\phi(\Sigma) + R(\Sigma) \leq Q_\alpha(\theta)$$

□

3.5 Proof of Theorem 3

Proof of Theorem 3. Let $\hat{\Sigma}_*$ be Σ part the minimizer $\hat{\theta}_*$ of Q_α under the constraint $\Sigma \in \mathcal{S}$. Σ_* denotes the minimizer of Q_α given $\mu = \mu^*$ under the constraint $\Sigma \in \mathcal{S}$. Let us denote $\theta_* = (\mu^*, \Sigma_*)$. By Theorem 2, we have

$$\begin{aligned}
f(\mu^*) + R(\Sigma_*) - \frac{L\alpha^2}{2}\mathrm{tr}(\Sigma_*) - \alpha^3\phi(\Sigma_*) \leq& Q_\alpha(\theta_*) \leq f(\mu^*) + \frac{L\alpha^2}{2}\mathrm{tr}(\Sigma_*) + \alpha^3\phi(\Sigma_*) + R(\Sigma_*) \\
f(\hat{\mu}_*) + R(\hat{\Sigma}_*) - \frac{L\alpha^2}{2}\mathrm{tr}(\hat{\Sigma}_*) - \alpha^3\phi(\hat{\Sigma}_*) \leq& Q_\alpha(\hat{\theta}_*) \leq f(\hat{\mu}_*) + \frac{L\alpha^2}{2}\mathrm{tr}(\hat{\Sigma}_*) + \alpha^3\phi(\hat{\Sigma}_*) + R(\hat{\Sigma}_*)
\end{aligned}$$

By the fact that $\hat{\theta}_*$ is the minimizer of $Q_\alpha(\theta)$ under constraint $\Sigma \in \mathcal{S}$, then we have $Q_\alpha(\hat{\theta}_*) \leq Q_\alpha(\theta_*)$. Thus, we can obtain that

$$\begin{aligned}
& f(\hat{\mu}^*) + R(\hat{\Sigma}_*) - \frac{L\alpha^2}{2}\mathrm{tr}(\hat{\Sigma}_*) - \alpha^3\phi(\hat{\Sigma}_*) \leq f(\mu^*) + \frac{L\alpha^2}{2}\mathrm{tr}(\Sigma_*) + \alpha^3\phi(\Sigma_*) + R(\Sigma_*) \\
\Rightarrow & f(\hat{\mu}_*) - f(\mu^*) \leq \frac{L\alpha^2}{2}\mathrm{tr}(\Sigma_* + \hat{\Sigma}_*) + \alpha^3\left(\phi(\Sigma_*) + \phi(\hat{\Sigma}_*)\right) + R(\Sigma_*) - R(\hat{\Sigma}_*)
\end{aligned}$$

Since μ^* is the solver of minimizing $f(\mu)$, we have $0 \le f(\hat{\mu}_*) - f(\mu^*)$. Thus, we can obtain that

$$0 \le f(\hat{\mu}_*) - f(\mu^*) \le \frac{L\alpha^2}{2}\mathrm{tr}(\Sigma_* + \hat{\Sigma}_*) + \alpha^3\left(\phi(\Sigma_*) + \phi(\hat{\Sigma}_*)\right) + R(\Sigma_*) - R(\hat{\Sigma}_*)$$

Next, we will bound the terms of right hand of above equation. First, we have

$$\frac{L\alpha^2}{2}\mathrm{tr}(\Sigma_* + \hat{\Sigma}_*) \le \frac{dL\alpha^2}{2}(\lambda_{\max}(\Sigma_*) + \lambda_{\max}(\hat{\Sigma}_*)) \le \frac{dL\alpha^2}{\tau} \tag{19}$$

where the last inequality is because Σ_* and $\hat{\Sigma}_*$ are in $\mathcal{S}$. Then we bound the value of $\phi(\Sigma_*)$ as follows.

$$\phi(\Sigma_*) \le \frac{\gamma\left\|\Sigma_*^{1/2}\right\|^3}{6}\left(\mathbb{E}_u\left[\|u\|^3\right]\right) \overset{(56)}{\le} \frac{\gamma(d+3)^{3/2}}{6\tau^{3/2}}$$

where the second inequality follows from Jensen's inequality. Similarly, we have

$$\phi(\hat{\Sigma}_*) \le \frac{\gamma(d+3)^{3/2}}{6\tau^{3/2}}$$

Thus, we obtain that

$$\alpha^3\left(\phi(\Sigma_*) + \phi(\hat{\Sigma}_*)\right) \le \frac{\gamma\alpha^3(d+3)^{3/2}}{3\tau^{3/2}} \tag{20}$$

Finally, we bound $R(\Sigma_*) - R(\hat{\Sigma}_*)$.

$$R(\Sigma_*) - R(\hat{\Sigma}_*) = \frac{\alpha^2}{2}\left(-\log\det\Sigma_* + \log\det\hat{\Sigma}_*\right) \le \frac{d\alpha^2}{2}\left(\lambda_{\max}\left(\hat{\Sigma}_*\right) - \lambda_{\min}\left(\Sigma_*\right)\right) \le \frac{d\alpha^2}{2}\left(\tau^{-1} - \zeta^{-1}\right)$$

Thus, we have

$$R(\Sigma_*) - R(\hat{\Sigma}_*) \le \frac{d\alpha^2}{2}\left(\tau^{-1} - \zeta^{-1}\right) \tag{21}$$

Combining Eqn. (19), (20) and (21), we obtain that

$$f(\hat{\mu}_*) - f(\mu^*) \le \frac{dL\alpha^2}{\tau} + \frac{\gamma\alpha^3(d+3)^{3/2}}{3\tau^{3/2}} + \frac{d\alpha^2}{2}\left(\tau^{-1} - \zeta^{-1}\right)$$

By the property of strongly convex, we have

$$\|\mu^* - \hat{\mu}_*\|^2 \le \frac{2}{\sigma}\left(f(\mu^*) - f(\hat{\mu}_*)\right) \le \frac{2dL\alpha^2}{\sigma\tau} + \frac{2\gamma\alpha^3(d+3)^{3/2}}{3\sigma\tau^{3/2}} + \frac{d\alpha^2}{\sigma}\left(\tau^{-1} - \zeta^{-1}\right)$$

□

3.6 Proof of Theorem 4

First, we give the following property.

Lemma 1. *Letting $u \sim N(0, I)$ and H be a positive semi-definite matrix, then we have*

$$\frac{1}{2}\cdot\mathbb{E}_u\left(u^\top\Sigma^{1/2}H\Sigma^{1/2}u\cdot\left(\Sigma^{-1/2}uu^\top\Sigma^{-1/2} - \Sigma^{-1}\right)\right) = H$$

Proof. Let $J(\theta)$ be defined as Eqn. (1). $f(\cdot)$ is a quadratic function with H as its Hessian matrix. Then $J(\theta)$ can be represented as Eqn. (14). Therefore, we have

$$\frac{\partial J(\theta)}{\Sigma} = \frac{\alpha^2}{2} H$$

Let $\widetilde{G}$ be defined as Eqn. (40) with respect to the quadratic function $f(\cdot)$. $\widetilde{G}$ can further reduce to

$$\mathbb{E}_u\left[\widetilde{G}\right] = \mathbb{E}_u\left[\frac{1}{2}\left(u^\top \Sigma^{1/2} H \Sigma^{1/2} u \cdot \left(\Sigma^{-1/2} u u^\top \Sigma^{-1/2} - \Sigma^{-1}\right)\right)\right] - \frac{1}{2}\Sigma^{-1}$$

By Lemma 4, we can obtain that

$$\mathbb{E}_u\left[\frac{1}{2}\left(u^\top \Sigma^{1/2} H \Sigma^{1/2} u \cdot \left(\Sigma^{-1/2} u u^\top \Sigma^{-1/2} - \Sigma^{-1}\right)\right)\right] = 2\alpha^{-2} \cdot \frac{\partial J(\theta)}{\Sigma} = H$$

This completes the proof. □

Proof of Theorem 4. We will compute $\hat{\Sigma}_*$. First, we have the following Lagrangian

$$L(\Sigma, A_1, A_2) = Q_\alpha(\theta) + \frac{\alpha^2}{2}\left\langle A_1, \Sigma - \tau^{-1} I\right\rangle + \frac{\alpha^2}{2}\left\langle A_2, \zeta^{-1} I - \Sigma\right\rangle$$

where A_1 and A_2 are two positive semi-definite matrices. Next, we will compute $\partial(L)/\partial\Sigma$

$$\partial(L)/\partial\Sigma = \frac{\partial J(\theta)}{\partial\Sigma} - \frac{\alpha^2}{2}\Sigma^{-1} + \frac{\alpha^2}{2}A_1 - \frac{\alpha^2}{2}A_2 \tag{22}$$

Furthermore, by Eqn. (4), (6) and $z = \hat{\mu}^* + \alpha\Sigma^{1/2}u$, we have

$$\begin{aligned}
\frac{\partial J(\theta)}{\partial\Sigma} =& \frac{\partial J(\theta)}{\partial\bar{\Sigma}} \cdot \frac{\partial\bar{\Sigma}}{\partial\Sigma} \\
\overset{(6)}{=}& \mathbb{E}_z\left[f(z)\left(\frac{1}{2}\Sigma^{-1}(z-\hat{\mu}^*)(z-\hat{\mu}^*)^\top\Sigma^{-1}\alpha^{-2} - \frac{1}{2}\Sigma^{-1}\alpha^{-2}\right)\right] \cdot \alpha^2 \\
=& \frac{1}{2} \cdot \mathbb{E}_u\left[f(\hat{\mu}^* + \alpha\Sigma^{1/2}u)\left(\Sigma^{-1/2}uu^\top\Sigma^{-1/2} - \Sigma^{-1}\right)\right]
\end{aligned} \tag{23}$$

We can express $f(\hat{\mu}^* + \alpha\Sigma^{1/2}u)$ using Taylor expansion as follows:

$$f(\hat{\mu}^* + \alpha\Sigma^{1/2}u) = f(\hat{\mu}^*) + \left\langle \nabla f(\hat{\mu}^*), \alpha\Sigma^{1/2}u\right\rangle + \frac{\alpha^2}{2}u^\top\Sigma^{1/2}\nabla^2 f(\hat{\mu}^*)\Sigma^{1/2}u + \tilde{\rho}\left(\alpha\Sigma^{1/2}u\right)$$

where $\tilde{\rho}\left(\alpha\Sigma^{1/2}u\right)$ satisfies that

$$|\tilde{\rho}\left(\alpha\Sigma^{1/2}u\right)| \le \frac{\gamma\alpha^3\left\|\Sigma^{1/2}u\right\|^3}{6} \tag{24}$$

due to Eqn. (18). By plugging the above Taylor expansion into Eqn. (23), we obtain

$$\begin{aligned}
\frac{\partial J(\theta)}{\partial\Sigma} =& \frac{1}{2} \cdot \mathbb{E}_u\left[\left(\frac{\alpha^2}{2}u^\top\Sigma^{1/2}\nabla^2 f(\hat{\mu}^*)\Sigma^{1/2}u + \tilde{\rho}\left(\alpha\Sigma^{1/2}u\right)\right) \cdot \left(\Sigma^{-1/2}uu^\top\Sigma^{-1/2} - \Sigma^{-1}\right)\right] \\
=& \frac{\alpha^2}{2}\nabla^2 f(\hat{\mu}^*) + \Phi(\Sigma)
\end{aligned}$$

where the last equality uses Lemma 1, and $\Phi(\Sigma)$ is defined as

$$\Phi(\Sigma) = \frac{1}{2} \cdot \mathbb{E}_u \left[\tilde{\rho}\left(\alpha \Sigma^{1/2} u\right) \cdot \left(\Sigma^{-1/2} u u^\top \Sigma^{-1/2} - \Sigma^{-1}\right)\right]$$

Replacing $\partial J(\theta)/\partial \Sigma$ to Eqn. (22), we have

$$\frac{\partial L}{\partial \Sigma} = \frac{\alpha^2}{2} \nabla^2 f(\hat{\mu}^*) + \Phi(\Sigma) - \frac{\alpha^2}{2}\Sigma^{-1} + \frac{\alpha^2}{2} A_1 - \frac{\alpha^2}{2} A_2$$

By the KKT condition, we have

$$\begin{aligned} &\hat{\Sigma}_* = \left(\nabla^2 f(\hat{\mu}^*) + 2\alpha^{-2}\Phi(\hat{\Sigma}_*) + A_1 - A_2\right)^{-1} \\ &A_1 \hat{\Sigma}_* = \tau^{-1} A_1, \quad A_2 \hat{\Sigma}_* = \zeta^{-1} A_2 \\ &A_1 \succeq 0, \quad A_2 \succeq 0. \end{aligned} \tag{25}$$

Because the optimization problem is strictly convex, there is a unique solution $(\hat{\Sigma}_*, A_1, A_2)$ that satisfy the above KKT condition. Let $\nabla^2 f(\hat{\mu}^*) + 2\alpha^{-2}\Phi(\hat{\Sigma}_*) = U \Lambda U^\top$ be the spectral decomposition of $\nabla^2 f(\hat{\mu}^*) + 2\alpha^{-2}\Phi(\hat{\Sigma}_*)$, where U is a orthonormal matrix and Λ is a diagonal matrix, then we construct A_1 and A_2 as follows

$$\begin{aligned} A_1 =& U \Lambda^{(1)} U^\top \quad \text{with} \quad \Lambda^{(1)}_{i,i} = \max\{\tau - \Lambda_{i,i}, 0\} \\ A_2 =& U \Lambda^{(2)} U^\top \quad \text{with} \quad \Lambda^{(2)}_{i,i} = \max\{\Lambda_{i,i} - \zeta, 0\} \end{aligned}$$

Substituting A_1 and A_2 in Eqn. (25), we can obtain that

$$\hat{\Sigma}_* = \Pi_{\mathcal{S}}\left(\left(\nabla^2 f(\mu^*) + 2\alpha^{-2}\Phi(\hat{\Sigma}_*)\right)^{-1}\right) \tag{26}$$

Similar to the proof of Theorem 2 and 3, we can check that $\hat{\Sigma}_*$, A_1 and A_2 satisfy the above KKT condition.

Now we begin to bound the error between $\hat{\Sigma}_*$ and $\Pi_{\mathcal{S}}\left(\left(\nabla^2 f(\hat{\mu}^*)\right)^{-1}\right)$. We have

$$\begin{aligned} &\left\|\hat{\Sigma}_* - \Pi_{\mathcal{S}}\left(\left(\nabla^2 f(\hat{\mu}^*)\right)^{-1}\right)\right\| \\ \overset{(26)}{=} &\left\|\Pi_{\mathcal{S}}\left(\left(\nabla^2 f(\hat{\mu}^*) + 2\alpha^{-2}\Phi(\hat{\Sigma}_*)\right)^{-1}\right) - \Pi_{\mathcal{S}}\left(\left(\nabla^2 f(\hat{\mu}^*)\right)^{-1}\right)\right\| \\ \leq &\left\|\left(\nabla^2 f(\hat{\mu}^*) + 2\alpha^{-2}\Phi(\hat{\Sigma}_*)\right)^{-1} - \left(\nabla^2 f(\hat{\mu}^*)\right)^{-1}\right\| \\ \leq &\left\|\left(\nabla^2 f(\hat{\mu}^*) + 2\alpha^{-2}\Phi(\hat{\Sigma}_*)\right)^{-1}\left(2\alpha^{-2}\Phi(\Sigma_*)\right)\left(\nabla^2 f(\hat{\mu}^*)\right)^{-1}\right\| \\ \leq &\left\|\left(\nabla^2 f(\hat{\mu}^*) + 2\alpha^{-2}\Phi(\hat{\Sigma}_*)\right)^{-1}\right\|_2 \cdot \left\|\left(\nabla^2 f(\hat{\mu}^*)\right)^{-1}\right\|_2 \cdot \left\|2\alpha^{-2}\Phi(\Sigma_*)\right\| \end{aligned} \tag{27}$$

where $\|\cdot\|_2$ is the spectral norm. The first inequality is because the projection operator onto a convex set is non-expansive [Bertsekas(2009)]. The second inequality used the following fact: for any two nonsingular matrices A and B, it holds that

$$A^{-1} - B^{-1} = A^{-1}(B - A) B^{-1} \tag{28}$$

The last inequality is because it holds that $\|AB\| \le \|A\|_2 \|B\|$ for two any consistent matrices A and B.

Now we bound $2\alpha^{-2}\left\|\Phi(\hat{\Sigma}_*)\right\|$ as follows

$$
\begin{aligned}
2\alpha^{-2}\left\|\Phi(\hat{\Sigma}_*)\right\| =& \mathbb{E}_u\left[\tilde{\rho}\left(\alpha\hat{\Sigma}_*^{1/2}u\right)\cdot\left(\hat{\Sigma}_*^{-1/2}uu^\top\hat{\Sigma}_*^{-1/2}-\hat{\Sigma}_*^{-1}\right)\right] \\
\overset{(24)}{\le}&\frac{\alpha\gamma\left\|\hat{\Sigma}_*\right\|_2^{3/2}}{6}\mathbb{E}_u\left[\|u\|^3\cdot\left\|\left(\hat{\Sigma}_*^{-1/2}uu^\top\hat{\Sigma}_*^{-1/2}-\hat{\Sigma}_*^{-1}\right)\right\|\right] \\
\le&\frac{\alpha\gamma\left\|\hat{\Sigma}_*\right\|_2^{3/2}\cdot\left\|\hat{\Sigma}_*^{-1/2}\right\|_2^2}{6}\mathbb{E}_u\left[\|u\|^3\left(\|u\|^2+d\right)\right] \\
\le&\frac{\alpha\gamma\zeta}{6\tau^{3/2}}\mathbb{E}_u\left[\|u\|^5+d\|u\|^3\right]\overset{(56)}{\le}\frac{\alpha\gamma\zeta}{6\tau^{3/2}}\left((d+5)^{5/2}+d(d+3)^{3/2}\right)
\end{aligned}
\tag{29}
$$

where the last inequality follows from the fact that $\hat{\Sigma}_*$ is in the convex set $\mathcal{S}$.

By the condition of α in Eqn. (17), we have $2\alpha^{-2}\left\|\Phi(\hat{\Sigma}_*)\right\| \le \frac{\sigma}{2}$ which implies

$$
\left\|\left(\nabla^2 f(\hat{\mu}^*)+2\alpha^{-2}\Phi(\hat{\Sigma}_*)\right)^{-1}\right\|_2 \le 2\sigma^{-1} \tag{30}
$$

Consequently, we can obtain that

$$
\begin{aligned}
&\left\|\hat{\Sigma}_*-\Pi_{\mathcal{S}}\left(\left(\nabla^2 f(\hat{\mu}^*)\right)^{-1}\right)\right\| \\
\overset{(27)}{\le}&\left\|\left(\nabla^2 f(\hat{\mu}^*)+2\alpha^{-2}\Phi(\Sigma_*)\right)^{-1}\right\|_2\cdot\left\|\left(\nabla^2 f(\hat{\mu}^*)\right)^{-1}\right\|_2\cdot\left\|2\alpha^{-2}\Phi(\Sigma_*)\right\| \\
\overset{(30)}{\le}&2\sigma^{-1}\cdot\sigma^{-1}\cdot\left\|2\alpha^{-2}\Phi(\Sigma_*)\right\|\overset{(29)}{\le}2\sigma^{-1}\cdot\sigma^{-1}\cdot\frac{\alpha\gamma\zeta}{6\tau^{3/2}}\left((d+5)^{5/2}+d(d+3)^{3/2}\right) \\
=&\frac{\alpha\gamma\zeta}{3\tau^{3/2}\sigma^2}\cdot\left((d+5)^{5/2}+d(d+3)^{3/2}\right)
\end{aligned}
$$

Similarly, we have

$$
\left\|\Sigma_*-\Pi_{\mathcal{S}}\left(\left(\nabla^2 f(\mu^*)\right)^{-1}\right)\right\| \le \frac{\alpha\gamma\zeta}{3\tau^{3/2}\sigma^2}\cdot\left((d+5)^{5/2}+d(d+3)^{3/2}\right)
$$

Next, we will bound $\left\|\Sigma_*-\hat{\Sigma}_*\right\|$ as follows

$$
\left\|\Sigma_*-\hat{\Sigma}_*\right\| \le \left\|\Pi_{\mathcal{S}}\left(\left(\nabla^2 f(\mu^*)\right)^{-1}\right)-\Pi_{\mathcal{S}}\left(\left(\nabla^2 f(\hat{\mu}_*)\right)^{-1}\right)\right\| + \frac{2\alpha\gamma\zeta}{3\tau^{3/2}\sigma^2}\cdot\left((d+5)^{5/2}+d(d+3)^{3/2}\right)
$$

We also have

$$
\begin{aligned}
&\left\|\Pi_{\mathcal{S}}\left(\left(\nabla^2 f(\mu^*)\right)^{-1}\right)-\Pi_{\mathcal{S}}\left(\left(\nabla^2 f(\hat{\mu}_*)\right)^{-1}\right)\right\| \le \left\|\left(\nabla^2 f(\mu^*)\right)^{-1}-\left(\nabla^2 f(\hat{\mu}_*)\right)^{-1}\right\| \\
\overset{(28)}{\le}&\left\|\left(\nabla^2 f(\mu^*)\right)^{-1}\right\|\cdot\left\|\left(\nabla^2 f(\hat{\mu}_*)\right)^{-1}\right\|\cdot\left\|\nabla^2 f(\mu^*)-\nabla^2 f(\hat{\mu}_*)\right\| \le \frac{\gamma}{\sigma^2}\|\mu^*-\hat{\mu}_*\| \\
\le&\frac{\gamma}{\sigma^2}\cdot\left(\frac{2dL\alpha^2}{\sigma\tau}+\frac{2\gamma\alpha^3(d+3)^{3/2}}{3\sigma\tau^{3/2}}+\frac{d\alpha^2}{\sigma}\left(\tau^{-1}-\zeta^{-1}\right)\right)^{1/2}
\end{aligned}
$$

The first inequality is because of the property that projection operator onto a convex set is non-expansive [Bertsekas(2009)]. The third inequality is due to $f(\cdot)$ is σ-strongly convex and $\nabla^2 f(\mu)$ is γ-Lipschitz continuous. The last inequality follows from Theorem 3.

Therefore, we can obtain that

$$\begin{aligned}\left\|\Sigma_* - \hat{\Sigma}_*\right\| \le & \frac{2\alpha\gamma\zeta}{3\tau^{3/2}\sigma^2} \cdot \left((d+5)^{5/2} + d(d+3)^{3/2}\right) \\ & + \frac{\gamma}{\sigma^2} \cdot \left(\frac{2dL\alpha^2}{\sigma\tau} + \frac{2\gamma\alpha^3(d+3)^{3/2}}{3\sigma\tau^{3/2}} + \frac{d\alpha^2}{\sigma}\left(\tau^{-1} - \zeta^{-1}\right)\right)^{1/2}\end{aligned}$$

□

4 Mirror Natural Evolution Strategies

In the previous sections, we have shown that one can obtain the minimizer of $f(\mu)$ by minimizing its reparameterized function $Q_\alpha(\theta)$ with $\theta = (\mu, \Sigma)$. Instead of solving the optimization problem (15) by the natural gradient descent, we propose a novel method called MIrror Natural Evolution Strategy (`MES`) to minimize $Q_\alpha(\theta)$. `MES` consists of two main update procedures. It updates μ by natural gradient descent but with 'antithetic sampling' (refers to Eqn. (34)). Moreover, `MES` updates Σ by the mirror descent method. The mirror descent of Σ can be derived naturally because $\nabla_\Sigma R(\Sigma) = -\frac{\alpha^2}{2}\Sigma^{-1}$ is a mirror map widely used in convex optimization [Kulis et al.(2009)].

In the rest of this section, we will first describe our algorithmic procedure in detail. Then we will discuss the connection between `MES` and existing works.

4.1 Algorithm Description

We will give the update rules of μ and Σ respectively.

Natural Gradient Descent of μ The natural gradient of $Q_\alpha(\theta)$ with respect to μ is defined as

$$g(\mu) = F_\mu^{-1}\frac{\partial Q_\alpha}{\partial \mu} \tag{31}$$

where F_μ is the Fisher information matrix with respect to μ. First, by the properties of the Gaussian distribution, we have the following property.

Lemma 2. *Let F_μ be the Fisher information matrix with respect to μ and the natural gradient $g(\mu)$ be defined as Eqn.* (31)*. Then $g(\mu)$ satisfies*

$$g(\mu) = \alpha^2 \cdot \left(\frac{1}{2\alpha} \cdot \mathbb{E}_u\left[(f(\mu + \alpha\Sigma^{1/2}u) - f(\mu - \alpha\Sigma^{1/2}u))\Sigma^{1/2}u\right]\right)$$

Proof. The Fisher matrix F_μ can be computed as follows [Wierstra et al.(2014)]:

$$F_\mu = \mathbb{E}_z\left[\nabla_\mu \log\pi(z|\theta)\nabla_\mu\log\pi(z|\theta)^\top\right] \overset{(5)}{=} \mathbb{E}_u\left[\alpha^{-2}\Sigma^{-1/2}uu^\top\Sigma^{-1/2}\right] = \alpha^{-2}\Sigma^{-1} \tag{32}$$

Note that $\frac{\partial Q_\alpha(\theta)}{\partial\mu} = \frac{\partial J(\theta)}{\partial \mu}$, by Eqn. (4), so we have

$$\frac{\partial Q_\alpha(\theta)}{\partial\mu} = \frac{\partial J(\theta)}{\partial\mu} = \mathbb{E}[f(z)\nabla_\mu \log\pi(z|\theta)]$$

We also have $z = \mu + \alpha\Sigma^{1/2}u$ with $u \sim N(0, I_d)$, that is, $z \sim N(\mu, \alpha^2\Sigma)$. We first consider μ part, by Eqn. (5) with $\bar{\Sigma} = \alpha^2\Sigma$, we have

$$\begin{aligned}\frac{\partial Q_\alpha(\theta)}{\partial \mu} =& \mathbb{E}_z\left[f(z)\nabla_\mu \log \pi(z|\theta)\right] = \mathbb{E}_u\left[f(\mu + \alpha\Sigma^{1/2}u)\left(\alpha^2\Sigma\right)^{-1} \cdot \alpha\Sigma^{1/2}u\right] \\ =& \mathbb{E}_u\left[f(\mu + \alpha\Sigma^{1/2}u)\alpha^{-1}\Sigma^{-1/2}u\right]\end{aligned}$$

Because of the symmetry of Gaussian distribution, by setting $z = \mu - \alpha\Sigma^{1/2}u$, we can obtain

$$\frac{\partial Q_\alpha(\theta)}{\partial \mu} = -\mathbb{E}_u\left[f(\mu - \alpha\Sigma^{1/2}u)\alpha^{-1}\Sigma^{-1/2}u\right]$$

Combining above two equations, we can obtain that

$$\frac{\partial Q_\alpha(\theta)}{\partial \mu} = \frac{1}{2\alpha} \cdot \mathbb{E}_u\left[(f(\mu + \alpha\Sigma^{1/2}u) - f(\mu - \alpha\Sigma^{1/2}u))\Sigma^{-1/2}u\right] \tag{33}$$

With the knowledge of $\partial Q_\alpha(\theta)/\partial\mu$ and F_μ in Eqn. (32), we can obtain the result. □

With the natural gradient $g(\mu)$ at hand, we can update μ by the natural gradient descent as follows:

$$\mu_{k+1} = \mu_k - \eta_1' g(\mu) = \mu_k - \eta_1 \cdot \frac{1}{2\alpha}\mathbb{E}_u\left[(f(\mu + \alpha\Sigma^{1/2}u) - f(\mu - \alpha\Sigma^{1/2}u))\Sigma^{1/2}u\right]$$

where $\eta_1 = \alpha^2\eta_1'$ is the step size. Note that, during the above update procedure, we need to compute the expectations which is infeasible in real applications. Instead, we sample a mini-batch of size b to approximate $\frac{1}{2\alpha}\mathbb{E}\left[(f(\mu + \alpha\Sigma^{1/2}u) - f(\mu - \alpha\Sigma^{1/2}u))\Sigma^{1/2}u\right]$, and we define

$$\tilde{g}(\mu_k) = \frac{1}{b}\sum_{i=1}^{b} \frac{f(\mu_k + \alpha\Sigma_k^{1/2}u_i) - f(\mu_k - \alpha\Sigma_k^{1/2}u_i)}{2\alpha}\Sigma_k^{1/2}u_i \quad \text{with} \quad u_i \sim N(0, I_d) \tag{34}$$

Using $\tilde{g}(\mu_k)$, we update μ as follows:

$$\mu_{k+1} = \mu_k - \eta_1\tilde{g}(\mu_k)$$

Mirror Descent of Σ Recall from the definition of Q_α, we have $Q_\alpha(\theta) = J(\theta) + R(\Sigma)$. With the regularizer $R(\Sigma)$, we can define the Bregman divergence with respect to $R(\Sigma)$ as

$$\begin{aligned}B_R(\Sigma_1, \Sigma_2) =& R(\Sigma_1) - R(\Sigma_2) - \langle \nabla_\Sigma R(\Sigma_2), \Sigma_1 - \Sigma_2\rangle \\ =& -\frac{\alpha^2}{2}\left(\log\det(\Sigma_1\Sigma_2^{-1}) - \left\langle \Sigma_2^{-1}, \Sigma_1\right\rangle + d\right)\end{aligned}$$

The update rule of Σ employing mirror descent is defined as

$$\Sigma_{k+1} = \operatorname*{argmin}_{\Sigma} \eta_2 \left\langle \frac{\partial Q_\alpha(\theta)}{\partial \Sigma_k}, \Sigma \right\rangle + B_R(\Sigma, \Sigma_k) \tag{35}$$

Using $\nabla_\Sigma R(\Sigma)$ as the mapping function, the update rule of above equation can be reduced to

$$\nabla_\Sigma R(\Sigma_{k+1}) = \nabla_\Sigma R(\Sigma_k) - \eta_2\nabla_\Sigma Q_\alpha(\theta_k) \tag{36}$$

In the following lemma, we will compute $\frac{\partial Q_\alpha(\theta)}{\partial \Sigma}$.

Lemma 3. *Let $Q_\alpha(\theta)$ be defined in Eqn. (12). Then it holds that*

$$\frac{\partial Q_\alpha(\theta)}{\partial \Sigma} = \frac{1}{4}\cdot \mathbb{E}_u\left[\left(f(\mu-\alpha\Sigma^{1/2}u)+f(\mu+\alpha\Sigma^{1/2}u)-2f(\mu)\right)\left(\Sigma^{-1/2}uu^\top\Sigma^{-1/2}-\Sigma^{-1}\right)\right]-\frac{\alpha^2}{2}\Sigma^{-1} \tag{37}$$

Proof. Note that $\partial Q_\alpha(\theta)/\partial\Sigma = \partial J(\theta)/\partial\Sigma + \partial R(\Sigma)/\partial\Sigma$. First, we will compute $\frac{\partial J(\theta)}{\partial \Sigma}$. Let $z = \mu+\alpha\Sigma^{1/2}u$ with $u\sim N(0,I_d)$. By Eqn. (6), we can obtain that

$$\begin{aligned}\frac{\partial J(\theta)}{\partial\Sigma} =& \frac{\partial J(\theta)}{\partial\bar{\Sigma}}\cdot\frac{\partial\bar{\Sigma}}{\partial\Sigma} \overset{(6)}{=} \mathbb{E}_z\left[f(z)\left(\frac{1}{2}\Sigma^{-1}(z-\mu)(z-\mu)^\top\Sigma^{-1}\alpha^{-2}-\frac{1}{2}\Sigma^{-1}\alpha^{-2}\right)\right]\cdot\alpha^2\\ =&\frac{1}{2}\cdot\mathbb{E}_u\left[f(\mu+\alpha\Sigma^{1/2}u)\left(\Sigma^{-1/2}uu^\top\Sigma^{-1/2}-\Sigma^{-1}\right)\right]\end{aligned}$$

Because of the symmetry of Gaussian distribution, we can also have $z=\mu-\alpha\Sigma^{1/2}u$. Then, we can similarly derive that

$$\frac{\partial J(\theta)}{\partial\Sigma} = \frac{1}{2}\cdot\mathbb{E}_u\left[f(\mu-\alpha\Sigma^{1/2}u)\left(\Sigma^{-1/2}uu^\top\Sigma^{-1/2}-\Sigma^{-1}\right)\right]$$

Note that, we also have the following identity

$$\mathbb{E}_u\left[f(\mu)\left(\Sigma^{-1/2}uu^\top\Sigma^{-1/2}-\Sigma^{-1}\right)\right]=0$$

Combining above equations, we can obtain that

$$\frac{\partial J(\theta)}{\partial\Sigma} = \frac{1}{4}\cdot\mathbb{E}_u\left[\left(f(\mu-\alpha\Sigma^{1/2}u)+f(\mu+\alpha\Sigma^{1/2}u)-2f(\mu)\right)\left(\Sigma^{-1/2}uu^\top\Sigma^{-1/2}-\Sigma^{-1}\right)\right] \tag{38}$$

Furthermore, by the definition of Q_α in Eqn. (12), we have

$$\frac{\partial Q_\alpha(\theta)}{\partial\Sigma} = \frac{\partial J(\theta)}{\partial\Sigma}-\frac{\alpha^2}{2}\Sigma^{-1}$$

Therefore, we can obtain the result. □

Since, we also have $\nabla_\Sigma R(\Sigma) = -\frac{\alpha^2}{2}\Sigma^{-1}$, substituting $\nabla_\Sigma Q_\alpha(\theta)$ and $\nabla_\Sigma R(\Sigma)$ into Eqn. (36), we have

$$\begin{aligned}-\frac{\alpha^2}{2}\Sigma_{k+1}^{-1} =& -\frac{\alpha^2}{2}\Sigma_k^{-1}-\eta_2\nabla_\Sigma Q_\alpha(\theta_k)\\ \Rightarrow \Sigma_{k+1}^{-1} =& \Sigma_k^{-1}+2\eta_2\alpha^{-2}\frac{\partial Q_\alpha(\theta_k)}{\partial\Sigma}\end{aligned} \tag{39}$$

Similar to the update of μ, we only sample a small batch points to query their values and use them to estimate $\partial Q_\alpha(\theta)/\partial\Sigma$. We can construct the approximate gradient with respect to Σ as follows:

$$\widetilde{G}(\Sigma_k) = \frac{1}{2b\alpha^2}\sum_{i=1}^{b}\left[\left(f(\mu_k-\alpha\Sigma_k^{1/2}u_i)+f(\mu_k+\alpha\Sigma_k^{1/2}u_i)-2f(\mu_k)\right)\left(\Sigma_k^{-1/2}u_iu_i^\top\Sigma_k^{-1/2}-\Sigma_k^{-1}\right)\right]-\Sigma_k^{-1} \tag{40}$$

The following lemma shows that an important property of $\widetilde{G}(\Sigma_k)$.

Lemma 4. *Letting $\widetilde{G}(\Sigma_k)$ be defined in Eqn. (40), then $\widetilde{G}(\Sigma_k)$ is an unbiased estimation of $2\frac{\partial Q_\alpha(\theta)}{\alpha^2\partial\Sigma}$ at Σ_k.*

Proof. By Eqn. (38), we can observe that

$$\bar{G}(\Sigma_k) = \frac{1}{2b\alpha^2}\sum_{i=1}^{b}\left[\left(f(\mu_k - \alpha\Sigma_k^{1/2}u_i) + f(\mu_k + \alpha\Sigma_k^{1/2}u_i) - 2f(\mu_k)\right)\left(\Sigma_k^{-1/2}u_iu_i^\top\Sigma_k^{-1/2}\Sigma_k^{-1/2} - \Sigma_k^{-1}\right)\right] \tag{41}$$

is an unbiased estimation of $2\alpha^{-2}\partial J(\theta)/\partial\Sigma$ at Σ_k. Furthermore, we have

$$\nabla_\Sigma R(\Sigma) = -\frac{\alpha^2}{2}\Sigma^{-1}$$

Therefore, we can conclude that $\widetilde{G}(\Sigma_k)$ is an unbiased estimation of $2\alpha^{-2}\partial Q_\alpha(\theta)/\partial\Sigma$ at Σ_k. $\square$

Replacing $2\alpha^{-2}\partial Q_\alpha(\theta_k)/\partial\Sigma$ with $\widetilde{G}(\Sigma_k)$ in Eqn. (39), we update Σ as follows

$$\Sigma_{k+1}^{-1} = \Sigma_k^{-1} + \eta_2\widetilde{G}(\Sigma_k)$$

where η_2 is the step size.

Projection to The Constraint Because of the constraint that $\Sigma_{k+1} \in \mathcal{S}$, we need to project Σ_{k+1} back to $\mathcal{S}$. Since we update Σ^{-1} instead of directly updating Σ, we define another convex set

$$\mathcal{S}' = \left\{\Sigma^{-1}\middle|\Sigma \in \mathcal{S}\right\} \tag{42}$$

It is easy to check that for any $\Sigma^{-1} \in \mathcal{S}'$, then it holds that $\Sigma \in \mathcal{S}$. Taking the extra projection to $\mathcal{S}'$, we modify the update rule of Σ as follows;

$$\begin{cases} \Sigma_{k+0.5}^{-1} = \Sigma_k^{-1} + \eta_2\widetilde{G}(\Sigma_k) \\ \Sigma_{k+1}^{-1} = \Pi_{\mathcal{S}'}\left(\Sigma_{k+0.5}^{-1}\right) \end{cases} \tag{43}$$

The projection $\Pi_{\mathcal{S}'}(\Sigma^{-1})$ is conducted as follows. First, we conduct the spectral decomposition $\Sigma^{-1} = U\Lambda U^\top$, where U is an orthonormal matrix and Λ is a diagonal matrix with $\Lambda_{i,i} = \lambda_i$. Second, we truncate λ_i's. If $\lambda_i > \zeta$, we set $\lambda_i = \zeta$. If $\lambda_i < \tau$, we set $\lambda_i = \tau$, that is

$$\Pi_{\mathcal{S}'}\left(\Sigma^{-1}\right) = U\bar{\Lambda}U^\top, \quad \text{with} \quad \bar{\Lambda}_{i,i} = \begin{cases} \tau^{-1} & \text{if } \lambda_i(\Sigma^{-1}) > \tau^{-1} \\ \zeta^{-1} & \text{if } \lambda_i(\Sigma^{-1}) < \zeta^{-1} \\ \lambda_i(\Sigma^{-1}) & \text{otherwise} \end{cases} \tag{44}$$

where $\bar{\Lambda}$ is a diagonal matrix. It is easy to check the correctness of Eqn. (44). For completeness, we prove it in Proposition 2 in the Appendix.

Algorithmic Summary of `MES` Now, we summarize the algorithmic procedure of `MES`. First, we update μ by natural gradient descent and update Σ by mirror descent as

$$\begin{aligned} \mu_{k+1} =& \mu_k - \eta_{1,k}\tilde{g}(\mu_k) \\ \Sigma_{k+1}^{-1} =& \Pi_{\mathcal{S}'}\left(\Sigma_k^{-1} + \eta_{2,k}\widetilde{G}(\Sigma_k)\right) \end{aligned}$$

where $\tilde{g}(\mu_k)$ and $\widetilde{G}(\Sigma_k)$ are defined in Eqn. (34) and (40), respectively. The detailed algorithm description is in Algorithm 1.

Algorithm 1 Meta-Algorithm MES

1: **Input:** μ_1, Σ_1^{-1}, α and the target iteration number K.
2: **for** $k = 1, \dots, K$ **do**
3: Compute $\tilde{g}(\mu_k) = \frac{1}{b}\sum_{i=1}^{b} \frac{f(\mu_k+\alpha\Sigma_k^{1/2}u_i)-f(\mu_k-\alpha\Sigma_k^{1/2}u_i)}{2\alpha}\Sigma_k^{1/2}u_i$ with $u_i \sim N(0, I_d)$
4: Compute $\widetilde{G}(\Sigma_k)$ as

$$\widetilde{G}(\Sigma_k) = \frac{1}{2b\alpha^2}\sum_{i=1}^{b}\left[\left(f(\mu_k - \alpha\Sigma_k^{1/2}u_i) + f(\mu_k + \alpha\Sigma_k^{1/2}u_i) - 2f(\mu_k)\right)\left(\Sigma_k^{-1/2}u_iu_i^\top\Sigma_k^{-1/2} - \Sigma_k^{-1}\right)\right] - \Sigma_k^{-1}$$

5: Update $\mu_{k+1} = \mu_k - \eta_{1,k}\tilde{g}(\mu_k)$
6: Update $\Sigma_{k+1}^{-1} = \Pi_{\mathcal{S}'}\left(\Sigma_k^{-1} + \eta_{2,k}\widetilde{G}(\Sigma_k)\right)$
7: **end for**

4.2 Relation to Existing Work

First, we compare MES to derivative free algorithms in the optimization literature, which uses function value differences to estimate the gradient. In the work of [Nesterov and Spokoiny(2017)], one approximates the gradient as follows

$$g(\mu_k) = \frac{1}{b}\sum_{i=1}^{b}\frac{f(\mu_k + \alpha u_i) - f(\mu_k - \alpha u_i)}{2\alpha}u_i \quad \text{with} \quad u_i \sim N(0, I_d) \tag{45}$$

and update μ as

$$\mu_{k+1} = \mu_k - \eta_1 g(\mu_k) \tag{46}$$

Comparing $\tilde{g}(\mu)$ to $g(\mu)$, we can observe that the difference lies on the estimated covariance matrix. By utilizing Σ_k to approximate the inverse of the Hessian , $\tilde{g}(\mu_k)$ is an estimation of the natural gradient. In contrast, $g(\mu_k)$ just uses the identity matrix hence it only estimates the gradient of $f(\mu_k)$. Note that if we don't track the Hessian information by updating Σ_k, and set Σ_k to the identity matrix, then MES will reduce to the derivative free algorithm of [Nesterov and Spokoiny(2017)]. This establishes a connection between NES and derivative free algorithms.

Then we compare MES with classical derivative-free algorithm [Conn et al.(2009)]. The algorithm proposed by [Conn et al.(2009)] is also a kind of second order method. Unlike MES, the gradient and Hessian of function $f(\cdot)$ are computed approximately by regression method which takes $O(d^2)$ queries to the function value. Thus, this kind of algorithms is much different from MES and other NES-type algorithms.

We may also compare MES to traditional NES algorithms (including CMA-ES since CMA-ES can be derived from NES [Akimoto et al.(2010)]). There are two differences between MES and the conventional NES algorithms. First, MES minimizes Q_α, while NES minimizes $J(\theta)$ defined in Eqn. (1). Second, the update rule of Σ is different. MES uses the mirror descent to update Σ^{-1}. In comparison, NES uses the natural gradient to update Σ.

5 Convergence Analysis

In this section we analyze the convergence properties of MES. MES has a two-stage convergence convergence properties just as the ones of classical Newton methods [Boyd and Vandenberghe(2004)].

In the first stage, `MES` we set a constant step size and we have the following convergence rate.

Lemma 5. *Let the objective function $f(\cdot)$ be L-smooth and σ-strongly convex. Given $0 < \delta < 1$, define that*

$$c_1 = \left(\sqrt{d} + \sqrt{b} + \sqrt{2\log(2/\delta)}\right)^2, \quad c_2 = b - 2\sqrt{b\log(1/\delta)}, \quad c_3 = 2d + 3\log(1/\delta) \tag{47}$$

Setting the step size $\eta_{1,k} = \frac{b\tau}{4Lc_1}$ in Algorithm 1. Then with a probability at least $1-\delta$, it holds that

$$f(\mu_{k+1}) - f(\mu^*) \le \left(1 - \frac{c_2\tau\sigma}{16c_1L\zeta}\right)\cdot\left(f(\mu_k) - f(\mu^*)\right) + \Delta_{\alpha,1} \tag{48}$$

with $\Delta_{\alpha,1}$ defined as

$$\Delta_{\alpha,1} = \frac{c_3^3\gamma b\alpha^4}{2^5\cdot 3^2\cdot c_1L\zeta\tau^3}\left(1 + \frac{c_1c_3}{2\tau\zeta}\right)$$

After several iterations and entering the local region, then Σ_k^{-1} can be helpful to improve the convergence rate of `MES`. Accordingly, we have the following convergence properties.

Lemma 6. *Assume that $f(\cdot)$ has the properties in Lemma 5 and also admits γ-Lipschitz Hessian. Let $\xi_k\cdot\Sigma_k^{-1} \preceq \nabla^2 f(\mu_k) \preceq \mathscr{L}_k\cdot\Sigma_k^{-1}$ and $f(\mu_k)$ satisfy the following property*

$$f(\mu_k) - f(\mu^*) \le \min\left\{\frac{2^3\cdot 3^2\cdot\mathscr{L}_k^4\tau^4}{L\gamma^2}, \frac{\xi_k^2\sigma^3}{8\gamma^2(L\tau^{-1}+2\xi_k)^2}\right\} \tag{49}$$

Given $0 < \delta < 1$, by setting the step size $\eta_{1,k} = \frac{b}{4\mathscr{L}_kc_1}$, then with a probability at least $1-\delta$, Algorithm 1 has the following property

$$f(\mu_{k+1}) - f(\mu^*) \le \left(1 - \frac{\xi_kc_2}{8\mathscr{L}_kc_1}\right)\cdot\Big(f(\mu_k) - f(\mu^*)\Big) + \Delta_{\alpha,2} \tag{50}$$

where c_1, c_3, c_2 is defined in Eqn. (47) and $\Delta_{\alpha,2}$ is defined as

$$\Delta_{\alpha,2} = \frac{b^{3/2}\zeta^3\gamma^4c_3^6\alpha^6}{2^8\cdot 3^4\cdot\sigma^3c_1^3\tau^6} + \frac{b\zeta\gamma^2c_3^4\alpha^4}{2^5\cdot 3^2\cdot\sigma\tau^3c_1^2} + \frac{b\gamma c_3^3\alpha^4}{2^5\cdot 3^2\cdot\tau^3c_1}$$

Lemma 6 shows that when Σ_k^{-1} is a good preconditioner for $\nabla^2 f(\mu_k)$, that is, $\mathscr{L}_k/\xi_k$ is of small value, then `MES` can achieve a fast convergence rate. For example, if $\mathscr{L}_k/\xi_k = 2$, Eqn. (50) shows that `MES` converges with a rate independent of the condition number L/σ. In this case, `MES` achieves a much faster converge rate than the vanilla zeroth-order algorithm in [Nesterov and Spokoiny(2017)].

Next, we will prove the how Σ_k^{-1} converge to be a good preconditioner for $\nabla^2 f(\mu_k)$ and provide an explicit convergence rate.

Lemma 7. *Given the target iteration number K and $0 < \delta < 1$, we denote that*

$$C_1 := \frac{A_KL(\zeta A_K + d\zeta)}{\tau}\cdot\sqrt{\frac{2}{b}\log\frac{K^2}{\delta}}, \text{ with } A_K = 8\log\frac{1}{b\delta} + 16\log(K+1) + d$$

$$C_2 := \sqrt{\frac{32c_1L\zeta\gamma^2}{c_2\tau\sigma^2}\cdot\Big(\big(f(\mu_1) - f(\mu^*)\big) + (K-1)\max\left\{\Delta_{\alpha,1},\Delta_{\alpha,2}\right\}\Big)},$$

$$C_3 := \frac{c_3^{3/2}\cdot(c_3+1)\cdot\zeta\cdot\sqrt{K-1}}{4\tau^{3/2}d^{1/2}}\cdot\alpha, \qquad C_4 := \frac{2L^2c_3^2\left(c_3\zeta + \sqrt{d}\zeta\right)^2}{b\tau^2} + \frac{d\zeta^2}{2b}$$

with c_1, c_3, c_2 are defined in Eqn. (47), and

$$C = \max\left\{\frac{9(C_1+C_2+C_3)^2}{4} + 3C_4,\ 2C_4 + \frac{L^2}{b},\ \left\|\Sigma_1^{-1} - H^*\right\|^2\right\} \tag{51}$$

Letting the sequence $\{\Sigma_k\}_{k=1}^K$ generated by Algorithm 1 by setting $\eta_{2,k} = 1/k$, then with a probability at least $1-\delta$, it holds that

$$\left\|\Sigma_k^{-1} - \Pi_{\mathcal{S}'}(\nabla^2 f(\mu^*))\right\|^2 \le \frac{C}{k}$$

Above lemma shows that Σ_k^{-1} will converge to the $\nabla^2 f(\mu^*)$ if $\zeta \ge L$ and $\tau \le \sigma$ with a rate $\mathcal{O}(\frac{\log k}{k})$. This rate is the same to the one of the stochastic gradient descent on the strongly convex function shown in [Rakhlin et al.(2012)]. To understand the result in Lemma 7, we only need to consider the quadratic case. Let us consider the following the problem:

$$\min_{\Sigma^{-1}\in\mathcal{S}'} h(\Sigma^{-1}) \triangleq \left\|\Sigma^{-1} - \nabla^2 f(\mu)\right\|^2 \tag{52}$$

It is easy to check that

$$\nabla_{\Sigma^{-1}} h(\Sigma^{-1}) = 2\left(\Sigma^{-1} - \nabla^2 f(\mu)\right) = -2\cdot\mathbb{E}\left[\widetilde{G}(\Sigma)\right]$$

where the last equality is because of Lemma 1 and the definition of $\widetilde{G}$. Thus, we can observe that the update rule of Σ_k in Step 6 of Algorithm 1 is trying to solve the problem (52) with the stochastic gradient descent method.

Next, we will show that the rate of $\widetilde{\mathcal{O}}(1/k)$ for the covariance matrix is almost tight by proving that $\mathbb{E}\left[\left\|\widetilde{G}(\Sigma)\right\|^2\right]$ will not converge to the zero. Without loss of generality, we only consider the case $d=1$ and $b=1$. We have

$$\begin{aligned}
\mathbb{E}\left[\left\|\widetilde{G}(\Sigma)\right\|^2\right] =& \mathbb{E}\left[\left\|\frac{1}{2}\left(u^\top\Sigma^{1/2}H\Sigma^{1/2}u\cdot\left(\Sigma^{-1/2}uu^\top\Sigma^{-1/2} - \Sigma^{-1}\right)\right) - \Sigma^{-1}\right\|^2\right]\\
=& \mathbb{E}\left[\left\|\frac{1}{2}\left(u^\top\Sigma^{1/2}H\Sigma^{1/2}u\cdot\left(\Sigma^{-1/2}uu^\top\Sigma^{-1/2} - \Sigma^{-1}\right)\right)\right\|^2\right] - 2\left\langle\Sigma^{-1}, H\right\rangle + \left\|\Sigma^{-1}\right\|^2\\
=& \frac{1}{4}\mathbb{E}\left[(u^8 + u^4 - 2u^6)H^2\right] - 2\left\langle\Sigma^{-1}, H\right\rangle + \left\|\Sigma^{-1}\right\|^2\\
=& \frac{39}{2}H^2 - 2\Sigma^{-1}H + \left\|\Sigma^{-1}\right\|^2\\
\ge& \frac{37}{2}H^2 > 0
\end{aligned}$$

where the forth equality is because of the moments of standard Gaussian distribution. Since the variance of $\widetilde{G}(\Sigma)$ is always positive constant, the stochastic gradient descent wit $\widetilde{G}$ can not achieve a linear convergence rate. Thus, Lemma 7 can only obtain a rate the same to the one of the stochastic gradient descent on the strongly convex function shown in [Rakhlin et al.(2012)] and this rate is tight.

Combining above results, we can obtain the final convergence properties of `MES`.

Theorem 5. *Let the objective function $f(\cdot)$ be L-smooth and σ-strongly convex and admit γ-Lipschitz Hessians. Given $0 < \delta < 1$, c_1, c_3, c_3 are defined in Eqn. (47) and C is defined in Lemma 7. Setting $\eta_{2,k} = 1/k$ in Algorithm 1, the sequence $\{\mu_k\}$ generated by Algorithm 1 has the following properties*

(1) If $f(\mu_k)$ does not satisfy condition (49), then by setting the step size $\eta_{1,k} = \frac{b\tau}{4Lc_1}$, then Eqn. (48) holds with a probability at least $1-\delta$.

(2) If Condition (49) is satisfied, then by setting the step size $\eta_{1,k} = \frac{b}{4\mathscr{L}_k c_1}$, with a probability at least $1-\delta$, Eqn. (50) holds with

$$\begin{aligned} \mathscr{L}_k =& \min\left\{\frac{L}{\tau}, 1+\tau^{-1}\cdot\left(\sqrt{\frac{C}{k}}+\gamma\sqrt{\frac{2\big(f(\mu_k)-f(\mu^*)\big)}{\sigma}}+\left\|\nabla^2 f(\mu_k)-\Pi_{\mathcal{S}'}\left(\nabla^2 f(\mu_k)\right)\right\|\right)\right\} \\ \xi_k =& \max\left\{\frac{\sigma}{\zeta}, 1-\tau^{-1}\cdot\left(\sqrt{\frac{C}{k}}+\gamma\sqrt{\frac{2\big(f(\mu_k)-f(\mu^*)\big)}{\sigma}}+\left\|\nabla^2 f(\mu_k)-\Pi_{\mathcal{S}'}\left(\nabla^2 f(\mu_k)\right)\right\|\right)\right\} \end{aligned} \tag{53}$$

Remark. Theorem 5 shows that `MES` converges with a slow linear convergence rate before entering the local region defined in Eqn. (49). After entering the local region, Σ_k^{-1} as the preconditioner begins to improve the convergence rate of `MES`. Note that Eqn. (53) implies that the convergence rate in the local region is no slower than the one before entering the local region. As the iteration goes, the terms $\sqrt{C/k}$ and $\sqrt{2\big(f(\mu_k)-f(\mu^*)\big)/\sigma}$ will converge to zero. Furthermore, without loss of generality, we assume that $\zeta \geq L$ and $\tau \leq \sigma$ which implies that $\left\|\nabla^2 f(\mu_k)-\Pi_{\mathcal{S}'}\left(\nabla^2 f(\mu_k)\right)\right\| = 0$. Then, Theorem 5 shows that `MES` will converge faster and faster gradually until achieve a linear rate independent of the condition number of the objective. Though, the convergence rate of `MES` will increase as iteration goes after entering the local region. However, `MES` can not achieve the superlinear convergence rate eventually because the final convergence rate of `MES` is a linear rate $1-\frac{c_2}{8c_1}$ which can be obtain by Eqn. (50) and (53).

Remark. By the definition of c_1 and c_2 in Eqn. (47), the convergence rate of `MES` depend on $\mathcal{O}(b/d)$. For example, before entering the local region, Eqn. (48) shows that `MES` converges with a rate $1-\mathcal{O}\left(\frac{b\tau\sigma}{dL\zeta}\right)$. Thus,, we can effectively improve the convergence rate by increasing the batch size. Similarly, by the definition of C in Eqn. (51) and Eqn. (53), the convergence rate of the covariance matrix can also be improved by increasing the batch size.

5.1 Proof of Theorem 5

Proof. For the case before entering the local region, the result can be obtained by Lemma 5. For the case after entering the local region, we have

$$\begin{aligned} &\left\|\Sigma_k^{1/2}\left(\Sigma_k^{-1}-\nabla^2 f(\mu_k)\right)\Sigma_k^{1/2}\right\|_2 \leq \|\Sigma\|_2\cdot\left\|\Sigma_k^{-1}-\nabla^2 f(\mu_k)\right\|_2 \leq \tau^{-1}\cdot\left\|\Sigma_k^{-1}-\nabla^2 f(\mu_k)\right\| \\ \leq&\tau^{-1}\cdot\left(\left\|\Sigma_k^{-1}-H^*\right\|+\left\|H^*-\nabla^2 f(\mu_k)\right\|\right) \leq \tau^{-1}\cdot\left(\sqrt{\frac{C}{k}}+\gamma\left\|\mu_k-\mu^*\right\|+\left\|\nabla^2 f(\mu_k)-\Pi_{\mathcal{S}'}\left(\nabla^2 f(\mu_k)\right)\right\|\right) \\ \leq&\tau^{-1}\cdot\left(\sqrt{\frac{C}{k}}+\gamma\sqrt{\frac{2\big(f(\mu_k)-f(\mu^*)\big)}{\sigma}}+\left\|\nabla^2 f(\mu_k)-\Pi_{\mathcal{S}'}\left(\nabla^2 f(\mu_k)\right)\right\|\right) \end{aligned}$$

where the last inequality is because of $f(\cdot)$ is σ-strongly convex. Above equation implies that

$$\left(1 - \tau^{-1} \cdot \left(\sqrt{\frac{C}{k}} + \gamma\sqrt{\frac{2\big(f(\mu_k) - f(\mu^*)\big)}{\sigma}} + \left\|\nabla^2 f(\mu_k) - \Pi_{\mathcal{S}'}\left(\nabla^2 f(\mu_k)\right)\right\|\right)\right) \cdot \Sigma_k^{-1} \preceq H_k$$

$$\preceq \left(1 + \tau^{-1} \cdot \left(\sqrt{\frac{C}{k}} + \gamma\sqrt{\frac{2\big(f(\mu_k) - f(\mu^*)\big)}{\sigma}} + \left\|\nabla^2 f(\mu_k) - \Pi_{\mathcal{S}'}\left(\nabla^2 f(\mu_k)\right)\right\|\right)\right) \cdot \Sigma_k^{-1}$$

Furthermore, it holds that $\tau \cdot I \preceq \Sigma_k^{-1} \preceq \zeta \cdot I$ and $\sigma \cdot I \preceq H \preceq L \cdot I$, which implies Eqn. (83). Combining above results with Lemma 6 concludes the proof. $\square$

6 Conclusion

In this paper, we proposed a novel reparameterized objective function, $Q_\alpha(\theta)$, and introduced a new algorithm, `MES` (Mirror Descent Natural Evolution Strategy), to efficiently minimize it. `MES` leverages zeroth-order queries to approximate both first-order and second-order information, offering an explicit convergence rate of $\widetilde{\mathcal{O}}(1/k)$ for the covariance matrix, which converges to the Hessian inverse of the objective function. This marks the first rigorous convergence analysis of a zeroth-order algorithm that approximates the Hessian through zerothorder queries.

Additionally, `MES` bridges traditional derivative-free methods and natural evolution strategies (NES), providing new insights into their theoretical connections. The minimizer, (μ^*, Σ^*), of the reparameterized function is shown to be close to the minimizer and Hessian inverse of the original objective function with small perturbations. We believe this work lays the foundation for designing more efficient evolutionary algorithms and enhances the understanding of the convergence properties of existing zeroth-order methods.

References

[Akimoto(2012)] Y. Akimoto. Analysis of a natural gradient algorithm on monotonic convex-quadratic-composite functions. In *Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation*, GECCO '12, page 1293–1300, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311779.

[Akimoto et al.(2010)] Y. Akimoto, Y. Nagata, I. Ono, and S. Kobayashi. Bidirectional relation between CMA evolution strategies and natural evolution strategies. In *Parallel Problem Solving from Nature - PPSN XI, 11th International Conference, Kraków, Poland, September 11-15, 2010, Proceedings, Part I*, pages 154–163, 2010.

[Auger and Hansen(2016)] A. Auger and N. Hansen. Linear convergence of comparison-based step-size adaptive randomized search via stability of markov chains. *SIAM Journal on Optimization*, 26(3):1589–1624, 2016.

[Bertsekas(2009)] D. P. Bertsekas. *Convex optimization theory.* Athena Scientific Belmont, 2009.

[Beyer(2014)] H. Beyer. Convergence analysis of evolutionary algorithms that are based on the paradigm of information geometry. *Evolutionary Computation*, 22(4):679–709, 2014.

[Beyer and Deb(2001)] H.-G. Beyer and K. Deb. On self-adaptive features in real-parameter evolutionary algorithms. *IEEE Transactions on evolutionary computation*, 5(3):250–270, 2001.

[Boyd and Vandenberghe(2004)] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[Chen et al.(2019)] J. Chen, M. I. Jordan, and M. J. Wainwright. Boundary attack++: Query-efficient decision-based adversarial attack. *arXiv preprint arXiv:1904.02144*, 2019.

[Conn et al.(2009)] A. R. Conn, K. Scheinberg, and L. N. Vicente. Global convergence of general derivative-free trust-region algorithms to first-and second-order critical points. *SIAM Journal on Optimization*, 20(1):387–415, 2009.

[Conti et al.(2018)] E. Conti, V. Madhavan, F. P. Such, J. Lehman, K. Stanley, and J. Clune. Improving exploration in evolution strategies for deep reinforcement learning via a population of novelty-seeking agents. In *Advances in Neural Information Processing Systems*, pages 5027–5038, 2018.

[Dong et al.(2019)] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, and J. Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[Duchi et al.(2015)] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[Ghadimi and Lan(2013)] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[Glasmachers and Krause(2022)] T. Glasmachers and O. Krause. Convergence analysis of the hessian estimation evolution strategy. *Evolutionary Computation*, 30(1):27–50, 2022.

[Glasmachers et al.(2010)] T. Glasmachers, T. Schaul, S. Yi, D. Wierstra, and J. Schmidhuber. Exponential natural evolution strategies. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, pages 393–400. ACM, 2010.

[Hansen(2016)] N. Hansen. The cma evolution strategy: A tutorial. *arXiv preprint arXiv:1604.00772*, 2016.

[Hansen and Ostermeier(2001)] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–195, 2001.

[Ilyas et al.(2018)] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin. Black-box adversarial attacks with limited queries and information. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2137–2146, Stockholmsmässan, Stockholm Sweden, 2018.

[Kiefer and Wolfowitz(1952)] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pages 462–466, 1952.

[Kulis et al.(2009)] B. Kulis, M. A. Sustik, and I. S. Dhillon. Low-rank kernel learning with bregman matrix divergences. *Journal of Machine Learning Research*, 10(Feb):341–376, 2009.

[Lanckriet et al.(2004)] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine learning research*, 5(Jan):27–72, 2004.

[Laurent and Massart(2000)] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.

[Loshchilov and Hutter(2016)] I. Loshchilov and F. Hutter. Cma-es for hyperparameter optimization of deep neural networks. *arXiv preprint arXiv:1604.07269*, 2016.

[Malagò and Pistone(2015)] L. Malagò and G. Pistone. Information geometry of the gaussian distribution in view of stochastic optimization. In *Proceedings of the 2015 ACM Conference on Foundations of Genetic Algorithms XIII*, pages 150–162. ACM, 2015.

[Matyas(1965)] J. Matyas. Random optimization. *Automation and Remote control*, 26(2):246–253, 1965.

[Nesterov and Spokoiny(2017)] Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, Apr 2017.

[Ollivier et al.(2017)] Y. Ollivier, L. Arnold, A. Auger, and N. Hansen. Information-geometric optimization algorithms: A unifying picture via invariance principles. *Journal of Machine Learning Research*, 18: 18:1–18:65, 2017.

[Rakhlin et al.(2012)] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1571–1578, 2012.

[Salimans et al.(2017)] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

[Schwefel(1977)] H.-P. Schwefel. *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie.(Teil 1, Kap. 1-5)*. Birkhäuser, 1977.

[Tu et al.(2019)] C.-C. Tu, P. Ting, P.-Y. Chen, S. Liu, H. Zhang, J. Yi, C.-J. Hsieh, and S.-M. Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 742–749, 2019.

[Vershynin(2010)] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[Wainwright(2019)] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[Wierstra et al.(2008)] D. Wierstra, T. Schaul, J. Peters, and J. Schmidhuber. Natural evolution strategies. In *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, pages 3381–3387. IEEE, 2008.

[Wierstra et al.(2014)] D. Wierstra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhuber. Natural evolution strategies. *The Journal of Machine Learning Research*, 15(1):949–980, 2014.

[Ye et al.(2018)] H. Ye, Z. Huang, C. Fang, C. J. Li, and T. Zhang. Hessian-aware zeroth-order optimization for black-box adversarial attack. *arXiv preprint arXiv:1812.11377*, 2018.

A Convexity of $\mathcal{S}$ and $\mathcal{S}'$

First, we will show that $\mathcal{S}$ and $\mathcal{S}'$ are convex.

Proposition 1. *The sets $\mathcal{S}$ and $\mathcal{S}'$ defined in Eqn.* (16) *and* (42) *are convex.*

Proof. Letting Σ_1 and Σ_2 belong to $\mathcal{S}$ and given $0 \le \beta \le 1$, then we have

$$\beta\Sigma_1 + (1-\beta)\Sigma_2 \preceq \beta\tau^{-1}I + (1-\beta)\tau^{-1}I = \tau^{-1}I$$

Similarly, we have $\beta\Sigma_1 + (1-\beta)\Sigma_2 \succeq \zeta^{-1}I$. Therefore, $\mathcal{S}$ is a convex set. The convexity of $\mathcal{S}'$ can be proved similarly. $\square$

Proposition 2. *Let A be a symmetric matrix. $A = U\Lambda U^\top$ is the spectral decomposition of A. The diagonal matrix $\bar{\Lambda}$ is defined as*

$$\bar{\Lambda}_{i,i} = \begin{cases} \tau^{-1} & \textit{if } \Lambda_{i,i} > \tau^{-1} \\ \zeta^{-1} & \textit{if } \Lambda_{i,i} < \zeta^{-1} \\ \Lambda_{i,i} & \textit{otherwise} \end{cases}$$

Let $\Pi_{\mathcal{S}}(A)$ be the projection of symmetric A on to $\mathcal{S}$ defined in Eqn. (16)*, that is,* $\Pi_{\mathcal{S}}(A) = \operatorname{argmin}_{X\in\mathcal{S}} \|A - x\|$ *with $\|\cdot\|$ being Frobenius norm, then we have*

$$\Pi_{\mathcal{S}'}(A) = U\bar{\Lambda}U^\top$$

Proof. We have the following Lagrangian [Lanckriet et al.(2004)]

$$L(X, A_1, A_2) = \|A - X\|^2 + 2\left\langle A_1, X - \tau^{-1}I\right\rangle + 2\left\langle A_2, \zeta^{-1}I - X\right\rangle$$

where A_1 and A_2 are two positive semi-definite matrices. The partial derivative $\partial L(X, A_1, A_2)/\partial X$ is

$$\frac{\partial L(X, A_1, A_2)}{\partial X} = 2(X - A + A_1 - A_2)$$

By the general Karush-Kuhn-Tucker (KKT) condition [Lanckriet et al.(2004)], we have

$$\begin{aligned}
&X_* = A - A_1 + A_2\\
&A_1X_* = \tau^{-1}A_1, \quad A_2X_* = \zeta^{-1}A_2\\
&A_1 \succeq 0, \quad A_2 \succeq 0
\end{aligned}$$

Since the optimization problem is strictly convex, there is a unique solution (X_*, A_1, A_2) that satisfy the above KKT condition. Let $A = U\Lambda U^\top$ be the spectral decomposition of A. We construct A_1 and A_2 as follows:

$$A_1 = U\Lambda^{(1)}U^\top \quad \text{with} \quad \Lambda^{(1)}_{i,i} = \max\{\Lambda_{i,i} - \tau^{-1}, 0\} \tag{54}$$

$$A_2 = U\Lambda^{(2)}U^\top \quad \text{with} \quad \Lambda^{(2)}_{i,i} = \max\{\zeta^{-1} - \Lambda_{i,i}, 0\} \tag{55}$$

X_* is defined as $X_* = U\bar{\Lambda}U^\top$. We can check that $A - A_1 + A_2 = U(\Lambda - \Lambda^{(1)} + \Lambda^{(2)})U^\top = U\bar{\Lambda}U^\top = X_*$. The construction of A_1 and A_2 in Eqn. (54), (55) guarantees these two matrix are positive semi-definite. Furthermore, we can check that A_1 and A_2 satisfy $A_1X_* = \tau^{-1}A_1$ and $A_2X_* = \zeta^{-1}A_2$. Thus, A_1, A_2 and X_* satisfy the KKT's condition which implies $X_* = U\bar{\Lambda}U^\top$ is the projection of A onto $\mathcal{S}$. □

B Some Useful Lemmas

Lemma 8 ([Nesterov and Spokoiny(2017)])**.** *Let $p \geq 2$, u be from $N(0, I_d)$, then we have the following bound*

$$d^{p/2} \leq \mathbb{E}_u\left[\|u\|^p\right] \leq (p + d)^{p/2} \tag{56}$$

Lemma 9 (χ^2 tail bound [Laurent and Massart(2000)])**.** *Let $q_1, \ldots, q_n$ be independent χ^2 random variables, each with one degree of freedom. For any vector $\gamma = (\gamma_1, \ldots, \gamma_n) \in \mathbb{R}^n_+$ with non-negative entries, and any $t > 0$,*

$$\mathbb{P}\left[\sum_{i=1}^n \gamma_i q_i \geq \|\gamma\|_1 + 2\sqrt{\|\gamma\|_2^2\, t} + 2\|\gamma\|_\infty t\right] \leq \exp(-t),$$

$$\mathbb{P}\left[\sum_{i=1}^n \gamma_i q_i \leq \|\gamma\|_1 - 2\sqrt{\|\gamma\|_2^2\, t}\right] \leq \exp(-t)$$

where $\|\gamma\|_1 = \sum_{i=1}^n |\gamma_i|$.

Lemma 10 (Corollary 5.35 of [Vershynin(2010)]). *Let A be an $N \times n$ matrix whose entries are independent standard normal random variables. Then for every $\tau \geq 0$, with probability at least $1 - 2\exp(-\tau^2/2)$, the largest singular value $s_{\max}(A)$ satisfies*

$$s_{\max}(A) \leq \sqrt{N} + \sqrt{n} + \tau \tag{57}$$

Lemma 11. *Letting $u \sim N(0, I)$ be a d-dimensional Gaussian vector, then with probability at least $1 - \delta$, it holds that*

$$\|u\|^2 \leq 2d + 3\log(1/\delta) \tag{58}$$

Proof. By Lemma 9, we have

$$\|u\|^2 \leq d + 2\sqrt{d\log(1/\delta)} + 2\log(1/\delta) \leq 2d + 3\log(1/\delta)$$

□

Lemma 12 (Hoeffding's inequality). *For bounded random variables $X_i \in [a_i, b_i]$, where $X_1, \dots, X_n$ are independent, then $S_n = \sum_{i=1}^n X_i$ satisfies that*

$$\mathbb{P}\left(S_n - \mathbb{E}\left[S_n\right] \geq \psi\right) \leq \exp\left(-\frac{-2\psi^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Lemma 13 ([Rakhlin et al.(2012)]). *Let a positive sequence $\{a_t\}_{t=1}^\infty$ satisfy that $a_t \leq \frac{a_0}{t}$ for $t = 1, \dots, k$ and*

$$a_{k+1} \leq \frac{b}{k(k-1)}\sqrt{\sum_{t=2}^{k}(t-1)^2 \cdot a_t} + \frac{c}{k}, \qquad \text{with} \qquad b, c > 0$$

Then if $a_0 \geq 9b^2/4 + 3c$, then it holds that $a_{k+1} \leq a_0/(k+1)$.

C Properties of $\tilde{g}(\mu)$ and $\widetilde{G}(\Sigma)$

In this section, we will give several important properties related to $\tilde{g}(\mu)$ and $\widetilde{G}(\Sigma)$ that will be used in the proof of the convergence rates of Algorithm 1.

Lemma 14. *Denoting $\tilde{u}_i \sim N(0, \Sigma)$ with $\Sigma^{-1} \in \mathcal{S}$, and $\tilde{U} = [\tilde{u}_1, \dots, \tilde{u}_b] \in \mathbb{R}^{d \times b}$, we can represent $\tilde{g}(\mu)$ defined in Eqn.(34) as*

$$\tilde{g}(\mu) = \frac{1}{b}\sum_{i=1}^{b}\left(\nu_i \cdot \tilde{u}_i + \tilde{u}_i\tilde{u}_i^\top \nabla f(\mu)\right) = \frac{1}{b}\left(\tilde{U}\nu + \tilde{U}\tilde{U}^\top \nabla f(\mu)\right) \tag{59}$$

with

$$\nu_i = \frac{f(\mu + \alpha\tilde{u}_i) - f(\mu - \alpha\tilde{u}_i) - \langle \nabla f(\mu), 2\alpha\tilde{u}_i\rangle}{2\alpha}, \qquad \text{and} \qquad \nu = [\nu_1, \nu_2, \dots, \nu_b]^\top$$

Given $0 < \delta < 1$, if $f(x)$ is γ-Hessian smooth, then, with probability at least $1 - \delta$, it holds that

$$|\nu_i| \leq \frac{\gamma\alpha^2(2d + 3\log(1/\delta))^{3/2}}{6\tau^{3/2}} \tag{60}$$

Proof. By the definition of $\tilde{g}$, we have

$$\begin{aligned}\tilde{g}(\mu) =& \frac{1}{b}\sum_{i=1}^{b}\frac{f(\mu+\alpha\tilde{u}_i)-f(\mu-\alpha\tilde{u}_i)}{2\alpha}\cdot\tilde{u}_i\\ =& \frac{1}{b}\sum_{i=1}^{b}\Big(\frac{f(\mu+\alpha\tilde{u}_i)-f(\mu-\alpha\tilde{u}_i)-\langle\nabla f(\mu),2\alpha\tilde{u}_i\rangle}{2\alpha}\cdot\tilde{u}_i+\tilde{u}_i\tilde{u}_i^\top\nabla f(\mu)\Big)\\ =& \frac{1}{b}\sum_{i=1}^{b}\Big(\nu_i\cdot\tilde{u}_i+\tilde{u}_i\tilde{u}_i^\top\nabla f(\mu)\Big) = \frac{1}{b}\Big(\tilde{U}\nu+\tilde{U}\tilde{U}^\top\nabla f(\mu)\Big)\end{aligned}$$

If $f(\mu)$ is γ-Hessian smooth, denoting $a_{\tilde{u}_i}(\alpha) = f(\mu+\alpha\tilde{u}_i)-f(\mu)-\langle\nabla f(\mu),\alpha\tilde{u}_i\rangle-\frac{\mu^2}{2}\left\langle\nabla^2 f(\mu)\tilde{u}_i,\tilde{u}_i\right\rangle$, then we have

$$\begin{aligned}|\nu_i| =& \frac{a_{\tilde{u}_i}(\alpha)-a_{\tilde{u}_i}(-\alpha)}{2\alpha}\overset{(18)}{\leq}\frac{\gamma\left\|\alpha\tilde{u}_i\right\|^3}{6\alpha}\leq\frac{\gamma\alpha^2\left\|\Sigma^{3/2}\right\|\cdot\|u\|^3}{6}\\ \overset{(42)}{\leq}& \frac{\gamma\alpha^2\cdot\|u\|^3}{6\tau^{3/2}}\overset{(58)}{\leq}\frac{\gamma\alpha^2(2d+3\log(1/\delta))^{3/2}}{6\tau^{3/2}}\end{aligned}$$

□

Lemma 15. *Letting ν and $\tilde{U}$ be defined in Lemma 14, given $0<\delta<1$, then with a probability at least $1-\delta$, it holds that*

$$\|\nu\|^2 \leq \frac{\gamma b\alpha^4(2d+3\log(1/\delta))^3}{36\tau^3},\ \textit{and}\ \left\|\tilde{U}\nu\right\|^2\leq\frac{\gamma b\alpha^4(2d+3\log(1/\delta))^4}{36\tau^4} \tag{61}$$

Proof. First, by the definition of ν, we have

$$\|\nu\|^2 = \sum_{i}^{b}\nu_i^2\overset{(60)}{\leq}\frac{\gamma b\alpha^4(2d+3\log(1/\delta))^3}{36\tau^3}$$

Furthermore, we have

$$\begin{aligned}\left\|\tilde{U}\nu\right\|^2 =& \nu^\top\tilde{U}^\top\tilde{U}\nu = \nu^\top U^\top\Sigma U\nu\leq\|\Sigma\|\cdot\nu^\top U^\top U\nu\leq\|\Sigma\|\cdot\max_i|\nu_i|^2\cdot\|U\mathbf{1}\|^2\\ \leq& \frac{\gamma\alpha^4(2d+3\log(1/\delta))^3}{36\tau^4}\cdot\|U\mathbf{1}\|^2\end{aligned}$$

where $\mathbf{1}=[1,\dots,1]^\top\in\mathbb{R}^b$ and $U\in\mathbb{R}^{d\times b}$ with each entry being of standard normal distribution. Note that each entry of $U\mathbf{1}$ satisfies that $[U\mathbf{1}]_i\sim N(0,b)$. By Lemma 9, with a probability at least $1-\delta$, it holds that

$$\|U\mathbf{1}\|^2 = \sum_{i=1}^{d}[U\mathbf{1}]_i^2\leq b(2d+3\log(1/\delta)) \tag{62}$$

Therefore, we obtain the final result about $\left\|\tilde{U}\nu\right\|^2$ which concludes the proof. □

Lemma 16. *Letting $\tilde{U}$ be defined in Lemma 14 and given $0 < \delta < 1$, then with a probability at least $1-\delta$, it holds that*

$$\left\|\tilde{U}^\top \nabla f(\mu)\right\|^2 \geq \left(b - 2\sqrt{b\log(1/\delta)}\right)\left\|\Sigma^{1/2}\nabla f(\mu)\right\|^2 \tag{63}$$

Proof. By the definition of $\tilde{U}$, we have $\tilde{U}^\top \nabla f(\mu) = U^\top \Sigma^{1/2}\nabla f(\mu)$ with $U \in \mathbb{R}^{d\times b}$ whose entries are of standard normal distribution. Note that each entry of $U^\top\Sigma^{1/2}\nabla f(\mu)$ satisfies that $[U^\top\Sigma^{1/2}\nabla f(\mu)]_i \sim N(0, \left\|\Sigma^{1/2}\nabla f(\mu)\right\|^2)$ with $i = 1,\dots,b$. By Lemma 9, with probability at least $1-\delta$, it holds that

$$\left\|\tilde{U}^\top \nabla f(\mu)\right\|^2 = \sum_{i=1}^{b}[U^\top\Sigma^{1/2}\nabla f(\mu)]_i^2 \geq \left(b - 2\sqrt{b\log(1/\delta)}\right)\left\|\Sigma^{1/2}\nabla f(\mu)\right\|^2$$

□

Lemma 17. *Assume that $\nabla^2 f(\mu) \preceq \mathscr{L}\cdot\Sigma^{-1} \preceq \mathscr{L}\tau^{-1}\cdot I$ and $\nabla^2 f(\mu)$ is γ-Lipschitz continuous. Denote $U = [u_1, u_2, \dots, u_b] \in \mathbb{R}^{d\times b}$ with $u_i \sim N(0, I_d)$. Given $0 < \delta < 1/2$, with probability $1-\delta$, it holds that*

$$\begin{aligned}
&\|\tilde{g}(\mu)\|^2_{\nabla^2 f(\mu)}\\
\leq&\frac{2\mathscr{L}}{b^2}\left(\nu^\top U^\top U\nu + \nabla^\top f(\mu)\tilde{U}U^\top U\tilde{U}^\top\nabla f(\mu)\right)\\
\leq&\frac{2\mathscr{L}\left(\sqrt{d}+\sqrt{b}+\sqrt{2\log(2/\delta)}\right)^2}{b^2}\left\|U^\top\Sigma^{1/2}\nabla f(\mu)\right\|^2 + \frac{\mathscr{L}\gamma^2\alpha^4(2d+3\log(1/\delta))^4}{18b\tau^3}
\end{aligned} \tag{64}$$

Proof. It holds that $\tilde{U} = \Sigma^{1/2}U$ by the definition of $\tilde{U}$ in Lemma 14. Using H to denote $\nabla^2 f(\mu)$, we have

$$\begin{aligned}
\|\tilde{g}(\mu)\|_H^2 =&\tilde{g}^\top H\tilde{g} = \frac{1}{b^2}\left(\nu^\top\tilde{U}^\top H\tilde{U}\nu + 2\nu^\top\tilde{U}^\top H\tilde{U}\tilde{U}^\top\nabla f(\mu) + \nabla^\top f(\mu)\tilde{U}\tilde{U}^\top H\tilde{U}\tilde{U}^\top\nabla f(\mu)\right)\\
\leq&\frac{2}{b^2}\left(\nu^\top\tilde{U}^\top H\tilde{U}\nu + \nabla^\top f(\mu)\tilde{U}\tilde{U}^\top H\tilde{U}\tilde{U}^\top\nabla f(\mu)\right)
\end{aligned}$$

where the last inequality is because of the Cauchy's inequality.

We will bound above two terms in order. First, we have

$$\begin{aligned}
\nabla^\top f(\mu)\tilde{U}\tilde{U}^\top H\tilde{U}\tilde{U}^\top\nabla f(\mu) =&\nabla^\top f(\mu)\Sigma^{1/2}UU^\top\Sigma^{-1/2}H\Sigma^{1/2}UU^\top\Sigma^{1/2}\nabla f(\mu)\\
\leq&\mathscr{L}\cdot\nabla^\top f(\mu)\Sigma^{1/2}UU^\top UU^\top\Sigma^{1/2}\nabla f(\mu)
\end{aligned}$$

where the last inequality is because of the assumption that $\nabla^2 f(\mu) \preceq \mathscr{L}\cdot\Sigma^{-1}$. Based on above inequality, we can furthermore obtain that

$$\begin{aligned}
\nabla^\top f(\mu)\tilde{U}\tilde{U}^\top H\tilde{U}\tilde{U}^\top\nabla f(\mu) \leq&\mathscr{L}\cdot\|U\|^2\cdot\left\|U^\top\Sigma^{1/2}\nabla f(\mu)\right\|^2\\
\overset{(57)}{\leq}&\mathscr{L}\left(\sqrt{d}+\sqrt{b}+\sqrt{2\log(2/\delta)}\right)^2\cdot\left\|U^\top\Sigma^{1/2}\nabla f(\mu)\right\|^2
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
\nu^\top\tilde{U}^\top H\tilde{U}\nu =&\nu^\top U^\top\Sigma^{1/2}H\Sigma^{1/2}U\nu \leq \mathscr{L}\cdot\nu^\top U^\top U\nu \leq \mathscr{L}\max_i\{|\nu_i|^2\}\cdot\|U\mathbf{1}\|^2\\
\overset{(62)}{\leq}&\mathscr{L}b(2d+3\log(1/\delta))\cdot\max_i\{|\nu_i|^2\} \overset{(60)}{\leq} \frac{b\mathscr{L}\gamma^2\alpha^4(2d+3\log(1/\delta))^4}{36\tau^3}
\end{aligned}$$

Combining above results, we can obtain the final result. □

Lemma 18. *Let objective function $f(\cdot)$ be γ-Hessian smooth and $\Sigma^{-1} \preceq \tau^{-1} \cdot I$. Given $0 < \delta < 1/2$, it holds with probability at least $1 - \delta$ that*

$$\|\tilde{g}(\mu)\| \le \frac{\gamma\alpha^2(2d + 3\log(1/\delta))^2}{6b^{1/2}\tau^2} + \frac{\left(\sqrt{d} + \sqrt{b} + \sqrt{2\log(2/\delta)}\right)}{b\tau^{1/2}} \cdot \left\|U^\top \Sigma^{1/2} \nabla f(x)\right\| \tag{65}$$

Proof. By Eqn. (59), we have

$$\|\tilde{g}\| \le \frac{1}{b}\left(\left\|\tilde{U}\sigma\right\| + \left\|\tilde{U}\tilde{U}^\top \nabla f(x)\right\|\right)$$

We will bound above terms in order. First, we have

$$\begin{aligned}\left\|\tilde{U}\sigma\right\| \le& \left\|\Sigma^{1/2}\right\| \|U\mathbf{1}\| \cdot \max_i\{\sigma_i\} \overset{(60)(62)}{\le} \frac{b^{1/2}\gamma\alpha^2(2d + 3\log(1/\delta))^2 \left\|\Sigma^{1/2}\right\|}{6\tau^{3/2}} \\ \le& \frac{b^{1/2}\gamma\alpha^2(2d + 3\log(1/\delta))^2}{6\tau^2}\end{aligned}$$

We also have

$$\begin{aligned}\left\|\tilde{U}\tilde{U}^\top \nabla f(x)\right\|^2 =& \nabla^\top f(x)\Sigma^{1/2}UU^\top \Sigma UU^\top \Sigma^{1/2}\nabla f(x) \\ \le& \|\Sigma\| \cdot \|U\|^2 \cdot \left\|U^\top \Sigma^{-1/2}\nabla f(x)\right\|^2 \\ \overset{(57)}{\le}& \left(\sqrt{d} + \sqrt{b} + \sqrt{2\log(2/\delta)}\right)^2 \|\Sigma\| \cdot \left\|U^\top \Sigma^{1/2}\nabla f(x)\right\|^2 \\ \overset{(16)}{\le}& \tau^{-1}\left(\sqrt{d} + \sqrt{b} + \sqrt{2\log(2/\delta)}\right)^2 \cdot \left\|U^\top \Sigma^{1/2}\nabla f(x)\right\|^2\end{aligned}$$

Combining above results, we can obtain the final result. □

Lemma 19. *Let $\Sigma^{-1} \in \mathcal{S}'$ and $\bar{G}(\Sigma)$ be defined in Eqn. (41). Assume that $\nabla^2 f(\mu)$ is γ-Lipschitz continuous. Given $0 < \delta < 1$, then with a probability at least $1 - \delta$, it holds that*

$$\begin{aligned}&\left\|\left(f(\mu - \alpha\Sigma^{1/2}u) + f(\mu + \alpha\Sigma^{1/2}u) - 2f(\mu)\right)\left(\Sigma^{-1/2}uu^\top\Sigma^{-1/2} - \Sigma^{-1}\right)\right\| \\ \le& 2L\tau^{-1}\alpha^2(2d + 2\log(1/\delta))\left(\zeta(2d + 2\log(1/\delta)) + \sqrt{d}\zeta\right)\end{aligned} \tag{66}$$

and

$$\begin{aligned}&\left\|\bar{G}(\Sigma) - \frac{1}{2b}\sum_{i=1}^{b} u_i^\top \Sigma^{1/2}\nabla^2 f(\mu)\Sigma^{1/2}u_i\left(\Sigma^{-1/2}u_iu_i^\top\Sigma^{-1/2} - \Sigma^{-1}\right)\right\|_2 \\ \le& \frac{(2d + 3\log(1/\delta))^{3/2} \cdot (2d + 3\log(1/\delta) + 1) \cdot \zeta\gamma}{4\tau^{3/2}} \cdot \alpha\end{aligned} \tag{67}$$

Proof. Using the Taylor's expansion, we have

$$\begin{aligned}f(\mu - \alpha\Sigma^{1/2}u) =& f(\mu) + \left\langle \nabla f(\mu), -\alpha\Sigma^{1/2}u\right\rangle + \alpha^2 u^\top \Sigma^{1/2}\left[\int_0^1 \nabla^2 f(\mu - t\alpha\Sigma^{1/2}u)\,dt\right]\Sigma^{1/2}u \\ f(\mu + \alpha\Sigma^{1/2}u) =& f(\mu) + \left\langle \nabla f(\mu), \alpha\Sigma^{1/2}u\right\rangle + \alpha^2 u^\top \Sigma^{1/2}\left[\int_0^1 \nabla^2 f(\mu + t\alpha\Sigma^{1/2}u)\,dt\right]\Sigma^{1/2}u\end{aligned}$$

Using above Taylor's expansions, we have

$$
\begin{aligned}
&\left\|\left(f(\mu-\alpha\Sigma^{1/2}u)+f(\mu+\alpha\Sigma^{1/2}u)-2f(\mu)\right)\left(\Sigma^{-1/2}uu^\top\Sigma^{-1/2}-\Sigma^{-1}\right)\right\| \\
=&\alpha^2\left\|u^\top\Sigma^{1/2}\left(\int_0^1\left[\nabla^2 f(\mu+t\alpha\Sigma^{1/2}u)+\nabla^2 f(\mu-t\alpha\Sigma^{1/2}u)\right]\,dt\right)\Sigma^{1/2}u\cdot\left(\Sigma^{-1/2}uu^\top\Sigma^{-1/2}-\Sigma^{-1}\right)\right\| \\
\le&\alpha^2\left\|\Sigma^{1/2}u\right\|^2\cdot\int_0^1\left\|\nabla^2 f(\mu+t\alpha\Sigma^{1/2}u)+\nabla^2 f(\mu-t\alpha\Sigma^{1/2}u)\right\|\,dt\cdot\left\|\Sigma^{-1/2}uu^\top\Sigma^{-1/2}-\Sigma^{-1}\right\| \\
\le&2L\alpha^2\cdot\left\|\Sigma^{1/2}u\right\|^2\cdot\left(\left\|\Sigma^{-1/2}u\right\|^2+\left\|\Sigma^{-1}\right\|\right) \\
\le&2L\tau^{-1}\alpha^2(2d+2\log(1/\delta))\left(\zeta(2d+2\log(1/\delta))+\sqrt{d}\zeta\right)
\end{aligned}
$$

where the second inequality is because of $f(\cdot)$ is L-smooth and the last inequality is because of Eqn. (16) and Eqn. (58).

Using above Taylor's expansions again, we can obtain that

$$
\begin{aligned}
&\frac{f(\mu-\alpha\Sigma^{1/2}u_i)+f(\mu-\alpha\Sigma^{1/2}u_i)-2f(\mu)}{2\alpha^2}-u_i^\top\Sigma^{1/2}\nabla^2 f(\mu)\Sigma^{1/2}u_i \\
=&u_i^\top\Sigma^{1/2}\left[\int_0^1\frac{\nabla^2 f(\mu-t\alpha\Sigma^{1/2}u_i)+\nabla^2 f(\mu+t\alpha\Sigma^{1/2}u_i)}{2}\,dt\right]\Sigma^{1/2}u_i-u_i^\top\Sigma^{1/2}\nabla^2 f(\mu)\Sigma^{1/2}u_i \\
\le&u_i^\top\Sigma^{1/2}\left[\int_0^1\frac{\left\|\nabla^2 f(\mu-t\alpha\Sigma^{1/2}u_i)+\nabla^2 f(\mu+t\alpha\Sigma^{1/2}u_i)-2\nabla^2 f(\mu)\right\|}{2}\,dt\right]\Sigma^{1/2}u_i \\
\overset{(11)}{\le}&u_i^\top\Sigma u_i\cdot\int_0^1\gamma t\alpha\left\|\Sigma^{1/2}u_i\right\|\,dt=u_i^\top\Sigma u_i\cdot\frac{\alpha\gamma\left\|\Sigma^{1/2}u_i\right\|}{2}\overset{(16)(58)}{\le}\frac{\gamma(2d+2\log(1/\delta))^{3/2}}{2\tau^{3/2}}\cdot\alpha
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\left\|\bar{G}-\frac{1}{2b}\sum_{i=1}^b u_i^\top\Sigma^{1/2}\nabla^2 f(\mu)\Sigma^{1/2}u_i\left(\Sigma^{-1/2}u_iu_i^\top\Sigma^{-1/2}-\Sigma^{-1}\right)\right\|_2 \\
\le&\frac{1}{2b}\sum_{i=1}^b\left|\frac{f(\mu-\alpha\Sigma^{1/2}u_i)+f(\mu-\alpha\Sigma^{1/2}u_i)-2f(\mu)}{2\alpha^2}-u_i^\top\Sigma^{1/2}\nabla^2 f(\mu)\Sigma^{1/2}u_i\right| \\
&\cdot\left\|\Sigma^{-1/2}u_iu_i^\top\Sigma^{-1/2}-\Sigma^{-1}\right\|_2 \\
\le&\frac{\gamma(2d+2\log(1/\delta))^{3/2}}{4\tau^{3/2}}\cdot\alpha\cdot\left(\left\|\Sigma^{-1/2}u_iu_i^\top\Sigma^{-1/2}\right\|_2+\left\|\Sigma^{-1}\right\|_2\right) \\
\overset{(16)(58)}{\le}&\frac{(2d+2\log(1/\delta))^{3/2}\cdot(2d+2\log(1/\delta)+1)\cdot\zeta\gamma}{4\tau^{3/2}}\cdot\alpha
\end{aligned}
$$

□

Lemma 20. *Let $\Sigma^{-1}\in\mathcal{S}'$ and $\widetilde{G}(\Sigma)$ be defined in Eqn. (40). Given $0<\delta<1$, denoting $H^*=$*

$\Pi_{\mathcal{S}'}(\nabla^2 f(\mu^*))$, *then with a probability at least* $1-\delta$, *it holds that*

$$\left\|\widetilde{G}(\Sigma_k)\right\|^2 \le \frac{2L^2(2d+3\log(1/\delta))^2\left(\zeta(2d+3\log(1/\delta))+\sqrt{d}\zeta\right)^2}{b\tau^2}+\frac{d\zeta^2}{2b} \tag{68}$$

$$\left\|\widetilde{G}(\Sigma_k)-\left(H^*-\Sigma_k^{-1}\right)\right\|^2 \le \frac{2L^2(2d+3\log(1/\delta))^2\left(\zeta(2d+3\log(1/\delta))+\sqrt{d}\zeta\right)^2}{b\tau^2}+\frac{d\zeta^2}{2b} \tag{69}$$

Proof. By the definition of $\widetilde{G}(\Sigma)$ in Eqn. (40), we have

$$\begin{aligned}
&\left\|\widetilde{G}(\Sigma_k)\right\|^2\\
=&\left\|\frac{1}{2b\alpha^2}\sum_{i=1}^{b}\left[\left(f(\mu_k-\alpha\Sigma_k^{1/2}u_i)+f(\mu_k+\alpha\Sigma_k^{1/2}u_i)-2f(\mu_k)\right)\left(\Sigma_k^{-1/2}u_iu_i^\top\Sigma_k^{-1/2}-\Sigma_k^{-1}\right)\right]-\Sigma_k^{-1}\right\|^2\\
\le&\frac{1}{4b^2\alpha^4}\sum_{i=1}^{b}\left\|\left(f(\mu_k-\alpha\Sigma_k^{1/2}u_i)+f(\mu_k+\alpha\Sigma_k^{1/2}u_i)-2f(\mu_k)\right)\left(\Sigma_k^{-1/2}u_iu_i^\top\Sigma_k^{-1/2}-\Sigma_k^{-1}\right)-\alpha^2\Sigma_k^{-1}\right\|^2\\
\le&\frac{1}{2b^2\alpha^4}\left(\sum_{i=1}^{b}\left\|\left(f(\mu_k-\alpha\Sigma_k^{1/2}u_i)+f(\mu_k+\alpha\Sigma_k^{1/2}u_i)-2f(\mu_k)\right)\left(\Sigma_k^{-1/2}u_iu_i^\top\Sigma_k^{-1/2}-\Sigma_k^{-1}\right)\right\|^2+b\alpha^4\left\|\Sigma_k^{-1}\right\|^2\right)\\
&\overset{(66)(42)}{\le}\frac{2L^2(2d+3\log(1/\delta))^2\left(\zeta(2d+3\log(1/\delta))+\sqrt{d}\zeta\right)^2}{b\tau^2}+\frac{d\zeta^2}{2b}
\end{aligned}$$

Similarly, it holds that

$$\begin{aligned}
&\left\|\widetilde{G}(\Sigma_k)-\left(H^*-\Sigma_k^{-1}\right)\right\|^2\\
=&\left\|\frac{1}{2b\alpha^2}\sum_{i=1}^{b}\left[\left(f(\mu_k-\alpha\Sigma_k^{1/2}u_i)+f(\mu_k+\alpha\Sigma_k^{1/2}u_i)-2f(\mu_k)\right)\left(\Sigma_k^{-1/2}u_iu_i^\top\Sigma_k^{-1/2}-\Sigma_k^{-1}\right)\right]-H^*\right\|^2\\
\le&\frac{2L^2(2d+3\log(1/\delta))^2\left(\zeta(2d+3\log(1/\delta))+\sqrt{d}\zeta\right)^2}{b\tau^2}+\frac{d\zeta^2}{2b}
\end{aligned}$$

where the last inequality is because of the objective function $f(\cdot)$ is L-smooth. □

D Auxiliary Proofs for Section 5

D.1 Proof of Lemma 5

Proof. By the update rule of Algorithm 1, we have

$$\begin{aligned}
f(\mu_{k+1})=&f(\mu_k-\eta_1\tilde{g}(\mu_k))\overset{(9)}{\le} f(\mu_k)-\eta_1\left\langle\nabla f(\mu_k),\tilde{g}(\mu_k)\right\rangle+\frac{L\eta_1^2}{2}\left\|\tilde{g}(\mu_k)\right\|^2\\
\overset{(59)}{=}&f(\mu_k)-\frac{\eta_1}{b}\left\langle\nabla f(\mu_k),\tilde{U}_k\nu+\tilde{U}_k\tilde{U}_k\nabla f(\mu_k)\right\rangle+\frac{L\eta_1^2}{2b^2}\left\|\tilde{U}_k\nu+\tilde{U}_k\tilde{U}_k^\top\nabla f(\mu_k)\right\|^2\\
\le&f(\mu_k)-\frac{\eta_1}{2b}\left\langle\nabla f(\mu_k),\tilde{U}_k\tilde{U}_k^\top\nabla f(\mu_k)\right\rangle+\frac{\eta_1\left\|\nu\right\|^2}{2b}\\
&+\frac{L\eta_1^2}{b^2}\left(\left\|\tilde{U}_k\tilde{U}_k^\top\nabla f(\mu_k)\right\|^2+\left\|\tilde{U}_k\nu\right\|^2\right)
\end{aligned}$$

where the last inequality is because of Cauchy's inequality and the fact that $2ab \le a^2 + b^2$.

Furthermore, with a probability at least $1-\delta$, it holds that

$$
\begin{aligned}
&\left\|\tilde{U}_k\tilde{U}_k^\top \nabla f(\mu_k)\right\|^2 = \nabla^\top f(\mu_k)\tilde{U}_k\tilde{U}_k^\top \tilde{U}_k\tilde{U}_k^\top \nabla f(\mu_k) = \nabla^\top f(\mu_k)\tilde{U}_k U\Sigma_k U^\top \tilde{U}_k^\top \nabla f(\mu_k)\\
\le& \left\|\Sigma_k\right\|_2 \left\|U\right\|^2 \cdot \left\|\tilde{U}_k^\top \nabla f(\mu_k)\right\|^2 \overset{(16)(57)}{\le} \tau^{-1}\left(\sqrt{d}+\sqrt{b}+\sqrt{2\log(2/\delta)}\right)^2 \cdot \left\|\tilde{U}_k^\top \nabla f(\mu_k)\right\|^2
\end{aligned}
$$

By replacing the value of $\eta_{1,k}$, we have

$$
\begin{aligned}
f(\mu_{k+1}) - f(\mu^*) \le& f(\mu_k) - f(\mu^*) - \frac{1}{16L\zeta}\left\|\tilde{U}_k^\top \nabla f(\mu_k)\right\|^2 + \frac{\|\nu\|^2}{8Lc_1\zeta} + \frac{\left\|\tilde{U}_k\nu\right\|^2}{16L\zeta^2}\\
\overset{(63)(61)}{\le}& f(\mu_k) - f(\mu^*) - \frac{c_2\tau}{16c_1L}\left\|\Sigma^{1/2}\nabla f(\mu_k)\right\|^2 + \frac{c_3^3\gamma b\alpha^4}{2^5\cdot 3^2\cdot c_1L\zeta\tau^3}\left(1+\frac{c_1c_3}{2\tau\zeta}\right)\\
\le& f(\mu_k) - f(\mu^*) - \frac{c_2\tau}{16c_1L\zeta}\left\|\nabla f(\mu_k)\right\|^2 + \Delta_{\alpha,1}\\
\le& f(\mu_k) - f(\mu^*) - \frac{c_2\tau\sigma}{16c_1L\zeta}\left(f(\mu_k) - f(\mu^*)\right) + \Delta_{\alpha,1}\\
=& \left(1-\frac{c_2\tau\sigma}{16c_1L\zeta}\right)\cdot\left(f(\mu_k) - f(\mu^*)\right) + \Delta_{\alpha,1}
\end{aligned}
$$

where the last inequality is because of $f(\cdot)$ is σ-strongly convex. □

D.2 Proof of Lemma 6

Before the proof, we first give a lemma which constructs connection between $\|\nabla f(\mu_k)\|_{\Sigma_k}^2$ and $f(\mu_k) - f(\mu^*)$.

Lemma 21. *Let $f(x)$ satisfy the properties described in Lemma 6 and the covariance matrix Σ_k satisfy $\xi_k\cdot\Sigma_k^{-1} \preceq \nabla^2 f(\mu_k) \preceq \mathscr{L}_k\cdot\Sigma_k^{-1}$. Then $\|\nabla f(\mu_k)\|_{\Sigma_k}^2$ satisfies that*

$$
-\left\|\nabla f(\mu_k)\right\|_{\Sigma_k}^2 \le -\xi_k\left(f(\mu_k) - f(\mu^*)\right), \quad \text{if} \quad f(\mu_k) - f(\mu^*) \le \frac{\xi_k^2\sigma^3}{8\gamma^2(L\tau^{-1}+2\xi_k)^2} \tag{70}
$$

Proof. Now, we begin to give the connections between $\|\nabla f(\mu_k)\|_{\Sigma_k}^2$ and $f(\mu_k) - f(\mu^*)$. First, by the Taylor's expansion, we have

$$
\nabla f(\mu^*) = \nabla f(\mu_k) + \nabla^2 f(\mu_k)(\mu^* - \mu_k) + \int_0^1 \left(\nabla^2 f(\mu_k - s(\mu_k - \mu^*)) - \nabla^2 f(\mu_k)\right)(\mu^* - \mu_k)
$$

Combining with the fact that $\nabla f(\mu^*) = 0$, we can obtain that

$$
\nabla f(\mu_k) = \nabla^2 f(\mu_k)(\mu_k - \mu^*) + \int_0^1 \left(\nabla^2 f(\mu_k - s(\mu_k - \mu^*)) - \nabla^2 f(\mu_k)\right)(\mu_k - \mu^*)\, ds
$$

Let us denote that

$$
\Delta_1 = \int_0^1 \left(\nabla^2 f(\mu^* - s(\mu_k - \mu^*)) - \nabla^2 f(\mu^*)\right)(\mu_k - \mu^*)\, ds
$$

which can be bounded as follows

$$\|\Delta_1\| \overset{(11)}{\leq} \gamma \int_0^1 s \|\mu_k - \mu^*\|^2 \ ds = \frac{\gamma}{2} \|\mu_k - \mu^*\|^2 \tag{71}$$

Let us denote $H^* = \nabla^2 f(\mu^*)$ and $H = \nabla^2 f(\mu_k)$. We have

$$\begin{aligned}
-\|\nabla f(\mu_k)\|_{\Sigma_k}^2 =& -\|H(\mu_k - \mu^*) + \Delta_1\|_{\Sigma_k}^2 \\
\leq& -\|H(\mu_k - \mu^*)\|_{\Sigma}^2 + 2\langle H(\mu_k - \mu^*), \Sigma_k \Delta_1\rangle \\
\leq& -\xi_k \|\mu_k - \mu^*\|_H^2 + 2\langle H(\mu_k - \mu^*), \Sigma_k \Delta_1\rangle
\end{aligned}$$

where the last inequality is because of the condition that $\xi_k \cdot \Sigma_k \preceq \nabla^2 f(\mu_k)$. Furthermore, we have

$$\begin{aligned}
f(\mu_k) \overset{(18)}{\leq}& f(\mu^*) + \langle \nabla f(\mu^*), \mu_k - \mu^*\rangle + \frac{1}{2}\langle \nabla^2 f(\mu^*)(\mu_k - \mu^*), \mu_k - \mu^*\rangle + \frac{\gamma}{6}\|\mu_k - \mu^*\|^3 \\
=& f(\mu^*) + \frac{1}{2}\|\mu_k - \mu^*\|_{H^*}^2 + \frac{\gamma}{6}\|\mu_k - \mu^*\|^3 \\
=& f(\mu^*) + \frac{1}{2}\|\mu_k - \mu^*\|_H^2 + \frac{1}{2}\|\mu_k - \mu^*\|_{H^* - H}^2 + \frac{\gamma}{6}\|\mu_k - \mu^*\|^3 \\
\overset{(11)}{\leq}& f(\mu^*) + \frac{1}{2}\|\mu_k - \mu^*\|_H^2 + \frac{2\gamma}{3}\|\mu_k - \mu^*\|^3
\end{aligned}$$

Hence, we have

$$-\frac{1}{2}\|\mu_k - \mu^*\|_H^2 \leq -\left(f(\mu_k) - f(\mu^*)\right) + \frac{2\gamma}{3}\|\mu_k - \mu^*\|^3 \tag{72}$$

Then we begin to bound $\langle H(\mu_k - \mu^*), \Sigma\Delta_1\rangle$. First, we have

$$\langle H(\mu_k - \mu^*), \Sigma\Delta_1\rangle \leq \|\mu_k - \mu^*\| \|\Delta_1\| \|H\Sigma\| \overset{(71)}{\leq} \frac{\gamma}{2}\|\mu_k - \mu^*\|^3 \|H\Sigma\| \overset{(9)(16)}{\leq} \frac{\gamma L}{2\tau}\|\mu_k - \mu^*\|^3$$

Therefore, we obtain that

$$\begin{aligned}
-\|\nabla f(\mu_k)\|_{\Sigma_k}^2 \leq& -\xi_k \|\mu_k - \mu^*\|_H^2 + \gamma L \tau^{-1}\|\mu_k - \mu^*\|^3 \\
\overset{(72)}{\leq}& -2\xi_k\left(f(\mu_k) - f(\mu^*)\right) + \frac{4\xi_k\gamma}{3}\|\mu_k - \mu^*\|^3 + \gamma L\tau^{-1}\|\mu_k - \mu^*\|^3 \\
\leq& -2\xi_k\left(f(\mu_k) - f(\mu^*)\right) + \gamma(L\tau^{-1} + 2\xi_k)\|\mu_k - \mu^*\|^3 \\
\leq& -2\xi_k\left(f(\mu_k) - f(\mu^*)\right) + \gamma(L\tau^{-1} + 2\xi_k) \cdot \left(\frac{2}{\sigma}\right)^{3/2}\left(f(\mu_k) - f(\mu^*)\right)^{3/2},
\end{aligned}$$

where the last inequality is because of the σ-strong convexity of $f(x)$. Replacing the condition on $f(\mu_k) - f(\mu^*)$ to above inequality, we can obtain that

$$-\|\nabla f(\mu_k)\|_{\Sigma_k}^2 \leq -2\xi_k\left(f(\mu_k) - f(\mu^*)\right) + \xi_k\left(f(\mu_k) - f(\mu^*)\right) = -\xi_k\left(f(\mu_k) - f(\mu^*)\right)$$

□

Proof of Lemma 6. Taking a random step from μ_k, we have

$$
\begin{aligned}
& f(\mu_{k+1}) = f(\mu_k - \eta_{1,k}\tilde{g}(\mu_k)) \\
\overset{(18)}{\leq} & f(\mu_k) - \eta_{1,k}\left\langle \nabla f(\mu_k), \tilde{g}(\mu_k)\right\rangle + \frac{\eta_{1,k}^2}{2}\left\|\tilde{g}(\mu_k)\right\|_H^2 + \frac{\eta_{1,k}^3\gamma}{6}\left\|\tilde{g}(\mu_k)\right\|^3 \\
\overset{(59)(64)}{\leq} & f(\mu_k) - \frac{\eta_{1,k}}{b}\left\langle \nabla f(\mu_k), \tilde{U}\nu + \tilde{U}\tilde{U}^\top \nabla f(\mu_k)\right\rangle + \frac{\eta_{1,k}^3\gamma}{6}\left\|\tilde{g}(\mu_k)\right\|^3 \\
& + \frac{\eta_{1,k}^2\mathscr{L}_k c_1}{b^2}\left\|U^\top \Sigma_k^{1/2}\nabla f(\mu_k)\right\|^2 + \frac{\mathscr{L}_k\eta_{1,k}^2\gamma^2\alpha^4 c_3^4}{18b\tau^3} \\
\leq & f(\mu_k) - \frac{\eta_{1,k}}{2b}\nabla^\top f(\mu_k)\tilde{U}\tilde{U}^\top\nabla f(\mu_k) + \frac{\eta_{1,k}}{2b}\left\|\nu\right\|^2 + \frac{\eta_{1,k}^3\gamma}{6}\left\|\tilde{g}(\mu_k)\right\|^3 \\
& + \frac{\eta_{1,k}^2\mathscr{L}_k c_1}{b^2}\left\|U^\top \Sigma_k^{1/2}\nabla f(\mu_k)\right\|^2 + \frac{\eta_{1,k}^2\mathscr{L}_k\gamma^2\alpha^4 c_3^4}{18b\tau^3} \\
\overset{(65)(61)}{\leq} & f(\mu_k) - \frac{\eta_{1,k}}{2b}\nabla^\top f(\mu_k)\tilde{U}\tilde{U}^\top\nabla f(\mu_k) + \frac{\eta_{1,k}^2\mathscr{L}_k c_1^{1/2}}{b^2}\left\|U^\top \Sigma_k^{1/2}\nabla f(\mu_k)\right\|^2 \\
& + \frac{2\gamma c_1^{3/2}\eta_{1,k}^3}{3b^3\tau^{3/2}}\left\|U^\top \Sigma_k^{1/2}\nabla f(\mu_k)\right\|^2 + \frac{\eta_{1,k}^3\gamma^4\alpha^6 c_3^6}{2^2\cdot 3^4\cdot b^{3/2}\tau^6} + \frac{\mathscr{L}_k\eta_{1,k}^2\gamma^2\alpha^4 c_3^4}{18b\tau^3} + \frac{\eta_{1,k}\gamma c_3^3\alpha^4}{2^3\cdot 3^2\cdot\tau^3} \\
= & f(\mu_k) - \frac{1}{4\mathscr{L}_k c_1}\left\|U^\top \Sigma_k^{1/2}\nabla f(\mu_k)\right\|^2 + \frac{\gamma}{2^5\cdot 3\cdot \mathscr{L}_k^3 c_1^{3/2}\tau^{3/2}}\left\|U^\top \Sigma_k^{1/2}\nabla f(\mu_k)\right\|^3 \\
& + \frac{b^{3/2}\gamma^4 c_3^6\alpha^6}{2^8\cdot 3^4\cdot \mathscr{L}_k^3 c_1^3\tau^6} + \frac{b\gamma^2 c_3^4\alpha^4}{2^5\cdot 3^2\cdot\mathscr{L}_k\tau^3 c_1^2} + \frac{b\gamma c_3^3\alpha^4}{2^5\cdot 3^2\cdot\tau^3 c_1}
\end{aligned}
$$

where the third inequality follows from Cauchy's inequality and the last equality uses the value of the step size η_k.

Because of $\sigma\cdot I \preceq \nabla^2 f(\mu) \preceq L\cdot I$ and $\tau\cdot I \preceq \Sigma_k^{-1} \preceq \zeta\cdot I$, it holds that $\frac{\sigma}{\zeta} \leq \mathscr{L}_k$. Thus, we can upper bound that

$$
\begin{aligned}
& \frac{b^{3/2}\gamma^4 c_3^6\alpha^6}{2^8\cdot 3^4\cdot \mathscr{L}_k^3 c_1^3\tau^6} + \frac{b\gamma^2 c_3^4\alpha^4}{2^5\cdot 3^2\cdot\mathscr{L}_k\tau^3 c_1^2} + \frac{b\gamma c_3^3\alpha^4}{2^5\cdot 3^2\cdot\tau^3 c_1} \\
\leq & \frac{b^{3/2}\zeta^3\gamma^4 c_3^6\alpha^6}{2^8\cdot 3^4\cdot \sigma^3 c_1^3\tau^6} + \frac{b\zeta\gamma^2 c_3^4\alpha^4}{2^5\cdot 3^2\cdot\sigma\tau^3 c_1^2} + \frac{b\gamma c_3^3\alpha^4}{2^5\cdot 3^2\cdot\tau^3 c_1} := \Delta_{\alpha,2}
\end{aligned}
$$

Furthermore,

$$
\left\|\tilde{U}^\top\nabla f(\mu_k)\right\| \leq \left\|U\right\|\cdot\left\|\Sigma^{1/2}\nabla f(\mu_k)\right\| \overset{(57)}{\leq} c_1^{1/2}\cdot\left\|\Sigma^{1/2}\nabla f(\mu_k)\right\| \leq \sqrt{2c_1 L\tau^{-1}(f(\mu_k)-f(\mu^*))}
$$

Combining above equation with the condition on $f(\mu_k)-f(\mu^*)$, we can obtain

$$
\frac{\gamma\left\|\tilde{U}^\top\nabla f(\mu_k)\right\|}{2^5\cdot 3\cdot c_1^{3/2}\cdot\tau^{3/2}\mathscr{L}_k^3} \leq \frac{1}{8\mathscr{L}_k c_1}
$$

Therefore

$$
\begin{aligned}
f(\mu_{k+1}) - f(\mu^*) \leq & f(\mu_k) - f(\mu^*) - \frac{1}{8\mathscr{L}_k c_1}\left\|U^\top \Sigma_k^{1/2}\nabla f(\mu_k)\right\|^2 + \Delta_{\alpha,2} \\
\overset{(63)}{\leq} & f(\mu_k) - f(\mu^*) - \frac{b - 2\sqrt{b\log(1/\delta)}}{8\mathscr{L}_k c_1}\left\|\Sigma_k^{1/2}\nabla f(\mu_k)\right\|^2 + \Delta_{\alpha,2} \\
\overset{(70)}{\leq} & f(\mu_k) - f(\mu^*) - \frac{\xi_k c_2}{8\mathscr{L}_k c_1}\Big(f(\mu_k) - f(\mu^*)\Big) + \Delta_{\alpha,2} \\
= & \left(1 - \frac{\xi_k c_2}{8\mathscr{L}_k c_1}\right)\cdot\Big(f(\mu_k) - f(\mu^*)\Big) + \Delta_{\alpha,2}
\end{aligned}
$$

□

D.3 Proof of Lemma 7

The proof of Lemma 7 consists of several lemmas. The main idea is the same to the convergence analysis of stochastic gradient descent in [Rakhlin et al.(2012)] since our update of Σ_k^{-1} can be viewed as a kind of stochastic gradient descent just as discussed in Section 5.

Lemma 22. *Let us denote $H^* = \Pi_{\mathcal{S}'}\left(\nabla^2 f(\mu^*)\right)$ and set $\eta_{2,k} = \frac{1}{k}$. Given $0 < \delta < 1$, then Σ_k generated by Algorithm 1 has the following property with a probability at least $1-\delta$*

$$
\left\|\Sigma_{k+1}^{-1} - H^*\right\|^2 \leq \frac{1}{k(k-1)}\sum_{t=2}^{k}(t-1)\left\langle \bar{G}(\Sigma_t) - H^*, \Sigma_t^{-1} - H^*\right\rangle + \frac{1}{k(k-1)}\sum_{t=2}^{k}\frac{t-1}{t}\left\|\widetilde{G}(\Sigma_t)\right\|^2 \tag{73}
$$

Specifically, it also holds that

$$
\left\|\Sigma_2^{-1} - H^*\right\|^2 \leq \frac{4L^2c_3^2\left(c_3\zeta + \sqrt{d}\zeta\right)^2}{b\tau^2} + \frac{d\zeta^2}{b} \tag{74}
$$

$\bar{G}(\Sigma)$ and c_3 are defined in Eqn. (41) *and* (47) *respectively.*

Proof. By the update rule of Σ_k, we have

$$
\begin{aligned}
& \left\|\Sigma_{k+1}^{-1} - H^*\right\|^2 \\
= & \left\|\Pi_{\mathcal{S}'}(\Sigma_k^{-1} + \eta_{2,k}\widetilde{G}(\Sigma_k)) - H^*\right\|^2 \leq \left\|\Sigma_k^{-1} + \eta_{2,k}\widetilde{G}(\Sigma_k) - H^*\right\|^2 \\
= & \left\|\Sigma_k^{-1} - H^*\right\|^2 + 2\eta_{2,k}\left\langle \Sigma_k^{-1} - H^*, \widetilde{G}(\Sigma_k)\right\rangle + \eta_{2,k}^2\left\|\widetilde{G}(\Sigma_k)\right\|^2 \\
= & \left\|\Sigma_k^{-1} - H^*\right\|^2 + 2\eta_{2,k}\left\langle \Sigma_k^{-1} - H^*, H^* - \Sigma_k^{-1}\right\rangle \\
& + 2\eta_{2,k}\left\langle \Sigma_k^{-1} - H^*, \widetilde{G}(\Sigma_k) - \left(H^* - \Sigma_k^{-1}\right)\right\rangle + \eta_{2,k}^2\left\|\widetilde{G}(\Sigma_k)\right\|^2 \\
= & \left(1 - \frac{2}{k}\right)\left\|\Sigma_k^{-1} - H^*\right\|^2 + \frac{2}{k}\left\langle \widetilde{G}(\Sigma_k) - \left(H^* - \Sigma_k^{-1}\right), \Sigma_k^{-1} - H^*\right\rangle + \frac{1}{k^2}\left\|\widetilde{G}(\Sigma_k)\right\|^2
\end{aligned}
$$

where the first inequality is because of the non-expansiveness of projection onto convex set.

Unwinding this recursive inequality till $k=2$, we get that for any $k \geq 2$,

$$
\begin{aligned}
&\left\|\Sigma_{k+1}^{-1} - H^*\right\|^2 \\
\leq & 2\sum_{t=2}^{k} \frac{1}{t}\left(\prod_{j=t+1}^{k}\left(1-\frac{2}{j}\right)\right)\left\langle \widetilde{G}(\Sigma_t) - \left(H^* - \Sigma_t^{-1}\right), \Sigma_t^{-1} - H^*\right\rangle \\
&+ \sum_{t=2}^{k} \frac{1}{t^2}\left(\prod_{j=t+1}^{k}\left(1-\frac{2}{j}\right)\right)\left\|\widetilde{G}(\Sigma_t)\right\|^2 \\
=&\frac{1}{k(k-1)}\sum_{t=2}^{k}(t-1)\left\langle \widetilde{G}(\Sigma_t) - \left(H^* - \Sigma_t^{-1}\right), \Sigma_t^{-1} - H^*\right\rangle + \frac{1}{k(k-1)}\sum_{t=2}^{k}\frac{t-1}{t}\left\|\widetilde{G}(\Sigma_t)\right\|^2 \\
=&\frac{1}{k(k-1)}\sum_{t=2}^{k}(t-1)\left\langle \bar{G}(\Sigma_t) - H^*, \Sigma_t^{-1} - H^*\right\rangle + \frac{1}{k(k-1)}\sum_{t=2}^{k}\frac{t-1}{t}\left\|\widetilde{G}(\Sigma_t)\right\|^2
\end{aligned}
$$

Furthermore, for $k=1$, we have

$$
\begin{aligned}
&\left\|\Sigma_2^{-1} - H^*\right\|^2 \\
\leq&\left(1-\frac{2}{1}\right)\left\|\Sigma_1^{-1} - H^*\right\|^2 + \frac{2}{1}\left\langle \widetilde{G}(\Sigma_1) - \left(H^* - \Sigma_1^{-1}\right), \Sigma_1^{-1} - H^*\right\rangle + \frac{1}{1^2}\left\|\widetilde{G}(\Sigma_1)\right\|^2 \\
\leq& -\left\|\Sigma_1^{-1} - H^*\right\|^2 + 2\left\|\widetilde{G}(\Sigma_1) - \left(H^* - \Sigma_1^{-1}\right)\right\| \cdot \left\|\Sigma_1^{-1} - H^*\right\| + \left\|\widetilde{G}(\Sigma_1)\right\|^2 \\
\leq& -\left\|\Sigma_1^{-1} - H^*\right\|^2 + \left\|\widetilde{G}(\Sigma_1) - \left(H^* - \Sigma_1^{-1}\right)\right\|^2 + \left\|\Sigma_1^{-1} - H^*\right\|^2 + \left\|\widetilde{G}(\Sigma_1)\right\|^2 \\
=&\left\|\widetilde{G}(\Sigma_1) - \left(H^* - \Sigma_1^{-1}\right)\right\|^2 + \left\|\widetilde{G}(\Sigma_1)\right\|^2 \overset{(68)(69)}{\leq} \frac{4L^2c_3^2\left(c_3\zeta + \sqrt{d}\zeta\right)^2}{b\tau^2} + \frac{d\zeta^2}{b}
\end{aligned}
$$

□

Next, we mainly to bound the value $\frac{1}{k(k-1)}\sum_{t=2}^{k}(t-1)\left\langle \bar{G}(\Sigma_t) - H^*, \Sigma_t^{-1} - H^*\right\rangle$. We first decompose this term into several terms.

Lemma 23. *Lettting us denote that $H_t = \Pi_{\mathcal{S}'}\left(\nabla^2 f(\mu_t)\right)$, it holds that*

$$
\begin{aligned}
\left\langle \bar{G}(\Sigma_t) - H^*, \Sigma_t^{-1} - H^*\right\rangle =& \left\langle \frac{1}{2b}\sum_{i=1}^{b} u_i^\top \Sigma_t^{1/2} H_t \Sigma_t^{1/2} u_i \left(\Sigma_t^{-1/2} u_i u_i^\top \Sigma_t^{-1/2} - \Sigma_t^{-1}\right) - H_t, \Sigma_t^{-1} - H^*\right\rangle \\
&+ \left\langle \bar{G} - \frac{1}{2b}\sum_{i=1}^{b} u_i^\top \Sigma_t^{1/2} H_t \Sigma_t^{1/2} u_i \left(\Sigma_t^{-1/2} u_i u_i^\top \Sigma_t^{-1/2} - \Sigma_t^{-1}\right), \Sigma_t^{-1} - H^*\right\rangle + \left\langle H_t - H^*, \Sigma_t^{-1} - H^*\right\rangle
\end{aligned}
$$

Proof. It holds that

$$
\begin{aligned}
&\left\langle \bar{G} - H^*, \Sigma_t^{-1} - H^* \right\rangle = \left\langle \bar{G} - H_t + H_t - H^*, \Sigma_t^{-1} - H^* \right\rangle \\
=& \left\langle \frac{1}{2b} \sum_{i=1}^{b} u_i^\top \Sigma_t^{1/2} H_t \Sigma_t^{1/2} u_i \left(\Sigma_t^{-1/2} u_i u_i^\top \Sigma_t^{-1/2} - \Sigma_t^{-1} \right) - H_t, \Sigma_t^{-1} - H^* \right\rangle \\
&+ \left\langle \bar{G} - \frac{1}{2b} \sum_{i=1}^{b} u_i^\top \Sigma_t^{1/2} H_t \Sigma_t^{1/2} u_i \left(\Sigma_t^{-1/2} u_i u_i^\top \Sigma_t^{-1/2} - \Sigma_t^{-1} \right), \Sigma_t^{-1} - H^* \right\rangle \\
&+ \left\langle H_t - H^*, \Sigma_t^{-1} - H^* \right\rangle
\end{aligned}
$$

□

Lemma 24. *It holds that*

$$
\begin{aligned}
&\sum_{t=2}^{k} (t-1) \left\langle \bar{G} - \frac{1}{2b} \sum_{i=1}^{b} u_i^\top \Sigma_t^{1/2} H_t \Sigma_t^{1/2} u_i \left(\Sigma_t^{-1/2} u_i u_i^\top \Sigma_t^{-1/2} - \Sigma_t^{-1} \right), \Sigma_t^{-1} - H^* \right\rangle \\
&+ \sum_{t=2}^{k} (t-1) \left\langle H_t - H^*, \Sigma_t^{-1} - H^* \right\rangle \\
\leq& \sqrt{\sum_{t=2}^{k} \left\| H_t - H^* \right\|^2} \cdot \sqrt{\sum_{t=2}^{k} (t-1)^2 \left\| \Sigma_t^{-1} - H^* \right\|^2} \\
&+ \frac{c_3^{3/2} \cdot (c_3+1) \cdot \zeta\gamma \cdot \sqrt{k-1}}{4\tau^{3/2} d^{1/2}} \cdot \alpha \cdot \sqrt{\sum_{t=2}^{k} (t-1)^2 \left\| \Sigma_t^{-1} - H^* \right\|^2}
\end{aligned}
\tag{75}
$$

Proof. By the Cauchy' inequality, we have

$$
\begin{aligned}
&\sum_{t=2}^{k} (t-1) \left\langle H_t - H^*, \Sigma_t^{-1} - H^* \right\rangle \leq \sum_{t=2}^{k} (t-1) \left\| H_t - H^* \right\| \cdot \left\| \Sigma_t^{-1} - H^* \right\| \\
\leq& \sqrt{\sum_{t=2}^{k} \left\| H_t - H^* \right\|^2} \cdot \sqrt{\sum_{t=2}^{k} (t-1)^2 \left\| \Sigma_t^{-1} - H^* \right\|^2}
\end{aligned}
$$

Similarly, we can obtain that

$$
\begin{aligned}
&\sum_{t=2}^{k} (t-1) \left\langle \bar{G} - \frac{1}{2b} \sum_{i=1}^{b} u_i^\top \Sigma_t^{1/2} H_t \Sigma_t^{1/2} u_i \left(\Sigma_t^{-1/2} u_i u_i^\top \Sigma_t^{-1/2} - \Sigma_t^{-1} \right), \Sigma_t^{-1} - H^* \right\rangle \\
\leq& \sum_{t=2}^{k} (t-1) \left\| \bar{G} - \frac{1}{2b} \sum_{i=1}^{b} u_i^\top \Sigma_t^{1/2} H_t \Sigma_t^{1/2} u_i \left(\Sigma_t^{-1/2} u_i u_i^\top \Sigma_t^{-1/2} - \Sigma_t^{-1} \right) \right\| \cdot \left\| \Sigma_t^{-1} - H^* \right\| \\
\overset{(67)}{\leq}& \frac{c_3^{3/2} \cdot (c_3+1) \cdot \zeta}{4\tau^{3/2} d^{1/2}} \cdot \alpha \cdot \sum_{t=2}^{k} (t-1) \left\| \Sigma_t^{-1} - H^* \right\| \\
\leq& \frac{c_3^{3/2} \cdot (c_3+1) \cdot \zeta \cdot \sqrt{k-1}}{4\tau^{3/2} d^{1/2}} \cdot \alpha \cdot \sqrt{\sum_{t=2}^{k} (t-1)^2 \left\| \Sigma_t^{-1} - H^* \right\|^2}
\end{aligned}
$$

Combining above results, we can obtain the finaly result. □

In the next five lemmas, we try to bound the value

$$\sum_{t=2}^{k}\left\langle \bar{G}-\frac{1}{2b}\sum_{i=1}^{b}u_i^\top\Sigma_t^{1/2}H_t\Sigma_t^{1/2}u_i\left(\Sigma_t^{-1/2}u_iu_i^\top\Sigma_t^{-1/2}-\Sigma_t^{-1}\right),\Sigma_t^{-1}-H^*\right\rangle.$$

Lemma 25. *Let the eigenvalue decomposition of $\Sigma_t^{1/2}H_t\Sigma_t^{1/2}$ be defined as follows:*

$$\Sigma_t^{1/2}H_t\Sigma_t^{1/2}=V_t\Lambda_tV_t^\top,\ \textit{with } V_tV_t^\top=V_t^\top V_t=I,\Lambda_t=\mathrm{diag}\{\lambda_1^{(t)},\dots,\lambda_d^{(t)}\}$$

Denote that $\bar{u}_{t,i}=V_t^\top u_i$ and $E_{t,i}\triangleq\bar{u}_{t,i}^\top\Lambda_t\bar{u}_{t,i}\left(\bar{u}_{t,i}\bar{u}_{t,i}^\top-I\right)-2\Lambda_t$. It holds that

$$\begin{aligned}&\left\langle\frac{1}{2}u_i^\top\Sigma_t^{1/2}H_t\Sigma_t^{1/2}u_i\left(\Sigma_t^{-1/2}u_iu_i^\top\Sigma_t^{-1/2}-\Sigma_t^{-1}\right)-H_t,\Sigma_t^{-1}-H^*\right\rangle\\=&\frac{1}{2}\left\langle E_{t,i},V_t^\top\Sigma_t^{-1/2}\left(\Sigma_t^{-1}-H^*\right)\Sigma_t^{-1/2}V_t\right\rangle\end{aligned}\tag{76}$$

Proof. First, we have

$$\begin{aligned}&\frac{1}{2}u_i^\top\Sigma_t^{1/2}H_t\Sigma_t^{1/2}u_i\left(\Sigma_t^{-1/2}u_iu_i^\top\Sigma_t^{-1/2}-\Sigma_t^{-1}\right)-H_t\\=&\frac{1}{2}\Sigma_t^{-1/2}V_t\left(u_i^\top V_t\Lambda_tV_t^\top u_i\left(V_t^\top u_iu_i^\top V_t-I\right)-2\Lambda_t\right)V_t^\top\Sigma_t^{-1/2}\\=&\frac{1}{2}\Sigma_t^{-1/2}V_t\left(\bar{u}_{t,i}^\top\Lambda_t\bar{u}_{t,i}\left(\bar{u}_{t,i}\bar{u}_{t,i}^\top-I\right)-2\Lambda_t\right)V_t^\top\Sigma_t^{-1/2}\\=&\frac{1}{2}\Sigma_t^{-1/2}V_tE_{t,i}V_t^\top\Sigma_t^{-1/2}\end{aligned}$$

where the second equality is because we denote $\bar{u}_{t,i}=V_t^\top u_i$ and the last equality is due to the definition of $E_{t,i}$. Using the properties of inner product of matrices, we can obtain the final result. □

Lemma 26. *Letting $E_{t,i}$ be defined in Lemma 25 and given a parameter $A_t>0$, we can represent $E_{t,i}$ that*

$$E_{t,i}=\left(X_{t,i}-\mathbb{E}[X_{t,i}]\right)+\left(Y_{t,i}-\mathbb{E}[Y_{t,i}]\right)\tag{77}$$

with

$$X_{t,i}=\bar{u}_{t,i}^\top\Lambda_t\bar{u}_{t,i}\left(\bar{u}_{t,i}\bar{u}_{t,i}^\top-I\right)\mathbb{1}_{\|\bar{u}_{t,i}\|^2\le A_t},\qquad\textit{and}\qquad Y_{t,i}=\bar{u}_{t,i}^\top\Lambda_t\bar{u}_{t,i}\left(\bar{u}_{t,i}\bar{u}_{t,i}^\top-I\right)\mathbb{1}_{\|\bar{u}_{t,i}\|^2>A_t}$$

Proof. First, it holds that

$$\bar{u}_{t,i}=\bar{u}_{t,i}\mathbb{1}_{\|\bar{u}_{t,i}\|^2\le A_t}+\bar{u}_{t,i}\mathbb{1}_{\|\bar{u}_{t,i}\|^2>A_t}$$

Accordingly, we can obtain that

$$\begin{aligned}\bar{u}_{t,i}^\top\Lambda_t\bar{u}_{t,i}\left(\bar{u}_{t,i}\bar{u}_{t,i}^\top-I\right)=&\bar{u}_{t,i}^\top\Lambda_t\bar{u}_{t,i}\left(\bar{u}_{t,i}\bar{u}_{t,i}^\top-I\right)\mathbb{1}_{\|\bar{u}_{t,i}\|^2\le A_t}+\bar{u}_{t,i}^\top\Lambda_t\bar{u}_{t,i}\left(\bar{u}_{t,i}\bar{u}_{t,i}^\top-I\right)\mathbb{1}_{\|\bar{u}_{t,i}\|^2>A_t}\\=&X_{t,i}+Y_{t,i}\end{aligned}$$

Furthermore, since it holds that $\bar{u}_{t,i} = V_t^\top u_i$, $u_i \sim N(0, I)$ and V_t is orthonormal, it holds that $\bar{u} \sim N(0, I)$. Combining Lemma 1, we can obtain that

$$\mathbb{E}\left[\bar{u}_{t,i}^\top \Lambda_t \bar{u}_{t,i}\left(\bar{u}_{t,i}\bar{u}_{t,i}^\top - I\right)\right] = 2\Lambda_t = \mathbb{E}\left[X_{t,i} + Y_{t,i}\right]$$

Thus,

$$\begin{aligned} E_{t,i} =& \bar{u}_{t,i}^\top \Lambda_t \bar{u}_{t,i}\left(\bar{u}_{t,i}\bar{u}_{t,i}^\top - I\right) - 2\Lambda_t = X_{t,i} + Y_{t,i} - \mathbb{E}\left[X_{t,i} + Y_{t,i}\right] \\ =& \Big(X_{t,i} - \mathbb{E}[X_{t,i}]\Big) + \Big(Y_{t,i} - \mathbb{E}[Y_{t,i}]\Big) \end{aligned}$$

□

Lemma 27. *Let A_t and Λ_t be defined in Lemma 26 and Lemma 25 respectively. We define that*

$$\Gamma_t := A_t \left\|\Lambda_t\right\|_2 \left(\zeta A_t + \operatorname{tr}\left(\Sigma_t^{-1}\right)\right) \cdot \left\|\Sigma_t^{-1} - H^*\right\|_2$$

Given sequence $\{z_{t,k}\}_{t=1}^k$ with $z_{t,k} \geq 0$, $0 < \delta < 1$ and $X_{t,i}$ being defined in Lemma 26, it holds that with a probability at least $1 - \frac{\delta}{k^2}$ that

$$\left|\sum_{t=2}^k z_{t,k}\left\langle X_t - \mathbb{E}\left[X_t\right], \frac{1}{2}V^\top \Sigma_t^{-1/2} B_t \Sigma_t^{-1/2} V\right\rangle\right| \leq \sqrt{\frac{2}{b}\log\frac{k^2}{\delta} \cdot \sum_{t=2}^k z_{t,k}^2 \Gamma_t^2}$$

Proof. By the definition of $X_{t,i}$, we have

$$\begin{aligned} &\left|\left\langle X_{t,i}, \frac{1}{2}V_t^\top \Sigma_t^{-1/2}\left(\Sigma_t^{-1} - H^*\right)\Sigma_t^{-1/2} V_t\right\rangle\right| \\ =& \frac{1}{2}\left|\left\langle \bar{u}_{t,i}^\top \Lambda_t \bar{u}_{t,i}\left(\bar{u}_{t,i}\bar{u}_{t,i}^\top - I\right)\mathbb{1}_{\|\bar{u}_{t,i}\|^2 \leq A_t}, V_t^\top \Sigma_t^{-1/2}\left(\Sigma_t^{-1} - H^*\right)\Sigma_t^{-1/2} V_t\right\rangle\right| \\ \leq& \frac{1}{2}\left|\left\langle \bar{u}_{t,i}^\top \Lambda_t \bar{u}_{t,i} \cdot \bar{u}_{t,i}\bar{u}_{t,i}^\top \cdot \mathbb{1}_{\|\bar{u}_{t,i}\|^2 \leq A_t}, V_t^\top \Sigma_t^{-1/2}\left(\Sigma_t^{-1} - H^*\right)\Sigma_t^{-1/2} V_t\right\rangle\right| \\ &+ \frac{1}{2}\left|\left\langle \bar{u}_{t,i}^\top \Lambda_t \bar{u}_{t,i} \cdot \mathbb{1}_{\|\bar{u}_{t,i}\|^2 \leq A_t} \cdot I, V_t^\top \Sigma_t^{-1/2}\left(\Sigma_t^{-1} - H^*\right)\Sigma_t^{-1/2} V_t\right\rangle\right| \\ \leq& \frac{A_t^2\left\|\Lambda_t\right\|_2}{2} \cdot \left\|\Sigma_t^{-1/2}\left(\Sigma_t^{-1} - H^*\right)\Sigma_t^{-1/2}\right\|_2 + \frac{A_t\left\|\Lambda_t\right\|_2}{2}\left|\operatorname{tr}\left(\Sigma_t^{-1/2}\left(\Sigma_t^{-1} - H^*\right)\Sigma_t^{-1/2}\right)\right| \\ \leq& \frac{\zeta A_t^2\left\|\Lambda_t\right\|_2}{2} \cdot \left\|\Sigma_t^{-1} - H^*\right\|_2 + \frac{A_t\left\|\Lambda_t\right\|_2 \cdot \operatorname{tr}(\Sigma_t^{-1})}{2} \cdot \left\|\Sigma_t^{-1} - H^*\right\|_2 = \frac{\Gamma_t}{2} \end{aligned}$$

where the last inequality is because of Eqn. (16) and the von Neumann's trace theorem.

Applying the Hoeffding's inequality (refer to Lemma 12), we can obtain that

$$\mathbb{P}\left(\sum_{t=1}^k z_{t,k} \cdot \frac{1}{b}\sum_{i=1}^b \left\langle X_{t,i} - \mathbb{E}\left[X_{t,i}\right], \frac{1}{2}V^\top \Sigma_t^{-1/2}\left(\Sigma_k^{-1} - H^*\right)\Sigma_t^{-1/2} V\right\rangle \geq \psi\right) \leq \exp\left(-\frac{b\psi^2/2}{\sum_{t=1}^k z_{t,k}^2 \Gamma_t^2}\right)$$

For any $\delta \in (0, 1)$, letting the right hand side of above equation be δ/k^2, then with probability at least $1 - \delta/k^2$, it holds that

$$\left|\sum_{t=2}^k z_{t,k}\left\langle X_t - \mathbb{E}\left[X_t\right], \frac{1}{2}V^\top \Sigma_t^{-1/2} B_t \Sigma_t^{-1/2} V\right\rangle\right| \leq \sqrt{\frac{2}{b}\log\frac{k^2}{\delta} \cdot \sum_{t=2}^k z_{t,k}^2 \Gamma_t^2}$$

□

Lemma 28. *Given parameter $0 < \delta < 1$, define $A_t = 8\log\frac{1}{b\delta} + 16\log(t+1) + d$. Given sequence $\{z_{t,k}\}_{t=1}^k$ with $z_{t,k} \geq 0$ and $Y_{t,i}$ being defined in Lemma 26, with a probability at least $1-\delta$, it holds that*

$$\left|\sum_{t=0}^{k} z_{t,k}\cdot\frac{1}{b}\sum_{i=1}^{b}\left\langle Y_{t,i}-\mathbb{E}\left[Y_{t,i}\right], \frac{1}{2}V_t^\top\Sigma_t^{-1/2}\left(\Sigma_t^{-1}-H^*\right)\Sigma_t^{-1/2}V_t\right\rangle\right|$$
$$\leq \sum_{t=0}^{k} z_{t,k}\cdot\frac{3\mathrm{tr}(\Lambda_t)}{d}\cdot\frac{\delta\left(A_t^2+16A_t+144\right)}{(t+1)^2}\cdot\mathrm{tr}(\Sigma_t^{-1})\cdot\left\|\Sigma_t^{-1}-H^*\right\|$$

Proof. By the standard Chi-square concentration inequality (refer to Eqn. (2.18) and Example 2.11 of [Wainwright(2019)]), it holds that

$$\mathbb{P}\left(\|\bar{u}_{t,i}\|^2 \geq (\beta+1)d\right) \leq \exp\left(-d\beta/8\right), \forall\, \beta\geq 1, \bar{u}_{t,i}\sim N(0,I_d) \tag{78}$$

Combining the definition of A_t, we have

$$\mathbb{P}\left(\|\bar{u}_{t,i}\|^2 \geq A_t\right) \leq \frac{\delta}{b(t+1)^2} \tag{79}$$

Thus, we can obtain that

$$\mathbb{P}\left(\exists\, t\geq 0, 1\leq i\leq b, \|\bar{u}_{t,i}\|^2\geq A_t\right) \leq \sum_{t=1}^{\infty}\sum_{i=1}^{b}\mathbb{P}\left(\|\bar{u}_{t,i}\|^2\geq A_t\right) \leq \sum_{t=1}^{\infty}\frac{\delta}{(t+1)^2} = \frac{\pi^2\delta}{6}$$

That is,

$$\mathbb{P}\Big(Y_{t,i}=0, \forall t\geq 0, 1\leq i\leq b\Big) \geq 1-\frac{\pi^2\delta}{6} \tag{80}$$

Next, we will bound the value of $\mathbb{E}[Y_{t,i}]$ and $\sum_{t=0}^k z_{t,k}\mathbb{E}[Y_{t,i}]$. Noting that for $\bar{u}\sim N(0,I_d)$, we have $\bar{u} = \|\bar{u}\|\cdot\bar{u}'$, where $\bar{u}'\sim \mathrm{Unif}(\mathcal{S}^{d-1})$ and $\mathrm{Unif}(\mathcal{S}^{d-1})$ means that uniform distribution on sphere of dimension d. Furthermore, it has the property that $\|\bar{u}\|$ is independent of $\bar{u}'$.

Then, we can obtain that

$$\begin{aligned}\mathrm{tr}(\Lambda)\cdot I_d + 2\Lambda =& \mathbb{E}\left[\bar{u}^\top\Lambda\bar{u}\cdot\bar{u}\bar{u}^\top\right] = \mathbb{E}\left[\|\bar{u}\|^4\right]\cdot\mathbb{E}\left[\bar{u}'^\top\Lambda\bar{u}'\cdot\bar{u}'\bar{u}'^\top\right] = (d^2+2d)\cdot\mathbb{E}\left[\bar{u}'^\top\Lambda\bar{u}'\cdot\bar{u}'\bar{u}'^\top\right]\\ \mathrm{tr}(\Lambda)\cdot I_d =& \mathbb{E}\left[\bar{u}^\top\Lambda\bar{u}\cdot I_d\right] = \mathbb{E}\left[\|\bar{u}\|^2\right]\cdot\mathbb{E}\left[\bar{u}'^\top\Lambda\bar{u}'\cdot I_d\right] = d\cdot\mathbb{E}\left[\bar{u}'^\top\Lambda\bar{u}'\cdot I_d\right]\end{aligned}$$

Thus, we have

$$\begin{aligned}\mathbb{E}\left[Y_{t,i}\right] =& \mathbb{E}\left[\bar{u}_{t,i}^\top\Lambda_t\bar{u}_{t,i}\left(\bar{u}_{t,i}\bar{u}_{t,i}^\top - I\right)\mathbb{1}_{\|\bar{u}_{t,i}\|^2>A_t}\right]\\ =& \mathbb{E}\left[\left(\|\bar{u}_{t,i}\|^4\cdot\bar{u}_{t,i}'^\top\Lambda_t\bar{u}_{t,i}'\cdot\bar{u}_{t,i}'\bar{u}_{t,i}'^\top - \|\bar{u}_{t,i}\|^2\,\bar{u}_{t,i}'^\top\Lambda_t\bar{u}_{t,i}'\right)\mathbb{1}_{\|\bar{u}_{t,i}\|^2>A_t}\right]\\ =& \mathbb{E}\left[\left(\frac{\mathrm{tr}(\Lambda_t)\cdot I_d + 2\Lambda_t}{d^2+2d}\cdot\|\bar{u}_{t,i}\|^4 - \frac{\mathrm{tr}(\Lambda_t)}{d}\cdot I_d\right)\mathbb{1}_{\|\bar{u}_{t,i}\|^2>A_t}\right]\end{aligned}$$

which leads to

$$\|\mathbb{E}\left[Y_{t,i}\right]\| \leq \max\left\{\frac{3\mathrm{tr}(\Lambda_t)}{d^2+2d}\mathbb{E}\left[\|\bar{u}_{t,i}\|^4\,\mathbb{1}_{\|\bar{u}_{t,i}\|^2>A_t}\right], \frac{\mathrm{tr}(\Lambda_t)}{d}\mathbb{E}\left[\|\bar{u}_{t,i}\|^2\,\mathbb{1}_{\|\bar{u}_{t,i}\|^2>A_t}\right]\right\}$$

Furthermore,

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{u}_{t,i}\|^4 \mathbb{1}_{\|\bar{u}_{t,i}\|^2>A_t}\right] \\
=&\int_0^\infty \mathbb{P}\left(\|\bar{u}_{t,i}\|^4 \mathbb{1}_{\|\bar{u}_{t,i}\|^2>A_t} > s\right)\, ds = \int_0^\infty \mathbb{P}\left(\|\bar{u}_{t,i}\|^2 > \max\{\sqrt{s}, A_t\}\right)\, ds \\
=&\int_0^{A_t^2} \mathbb{P}\left(\|\bar{u}_{t,i}\|^2 > A_t\right)\, ds + \int_{A_t^2}^\infty \mathbb{P}\left(\|\bar{u}_{t,i}\|^2 > \sqrt{s}\right)\, ds \\
\overset{(78)(79)}{\leq}&\int_0^{A_t^2} \frac{\delta}{(t+1)^2}\, ds + \int_{A_t^2}^\infty \exp\left(-\frac{\sqrt{s}-d}{8}\right)\, ds = \frac{\delta A_t^2}{(t+1)^2} + \int_{A_t}^\infty 2x\exp\left(-\frac{x-d}{8}\right)\, dx \\
=&\frac{\delta A_t^2}{(t+1)^2} + 16(A_t+8)\exp\left(-\frac{A_t-d}{8}\right) \leq \frac{\delta(A_t^2+16A_t+144)}{(t+1)^2}
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\mathbb{E}\left[\|\bar{u}_{t,i}\|^2 \mathbb{1}_{\|\bar{u}_{t,i}\|^2>A_t}\right] =& \int_0^\infty \mathbb{P}\left(\|\bar{u}_{t,i}\|^2 \mathbb{1}_{\|\bar{u}_{t,i}\|^2>A_t} > s\right)\, ds = \int_0^\infty \mathbb{P}\left(\|\bar{u}_{t,i}\|^2 > \max\{s, A_t\}\right)\, ds \\
\overset{(78)(79)}{\leq}& \int_0^{A_t} \frac{\delta}{(t+1)^2}\, ds + \int_{A_t}^\infty \exp\left(-\frac{s-d}{8}\right)\, ds \leq \frac{\delta(A_t^2+8)}{(t+1)^2}
\end{aligned}$$

Thus, we have

$$\|\mathbb{E}\left[Y_{t,i}\right]\| \leq \frac{3\mathrm{tr}(\Lambda_t)}{d} \cdot \frac{\delta\left(A_t^2+16A_t+144\right)}{(t+1)^2}$$

Thus, with a probability at least $1-\delta$, it holds that

$$\begin{aligned}
&\left|\sum_{t=0}^k z_{t,k} \cdot \frac{1}{b}\sum_{i=1}^b \left\langle Y_{t,i} - \mathbb{E}\left[Y_{t,i}\right], \frac{1}{2}V_t^\top \Sigma_t^{-1/2}\left(\Sigma_t^{-1} - H^*\right)\Sigma_t^{-1/2} V_t\right\rangle\right| \\
\overset{(80)}{=}&\left|\sum_{t=0}^k z_{t,k} \left\langle \mathbb{E}\left[Y_{t,i}\right], \frac{1}{2}V_t^\top \Sigma_t^{-1/2}\left(\Sigma_t^{-1} - H^*\right)\Sigma_t^{-1/2} V_t\right\rangle\right| \\
\leq&\sum_{t=0}^k z_{t,k} \cdot \|\mathbb{E}[Y_t]\| \cdot \sum_{i=1}^d \left|\lambda_i\left(\Sigma_t^{-1/2}\left(\Sigma_t^{-1} - H^*\right)\Sigma_t^{-1/2}\right)\right| \\
\leq&\sum_{t=0}^k z_{t,k} \cdot \frac{3\mathrm{tr}(\Lambda_t)}{d} \cdot \frac{\delta\left(A_t^2+16A_t+144\right)}{(t+1)^2} \cdot \mathrm{tr}(\Sigma_t^{-1}) \cdot \left\|\Sigma_t^{-1} - H^*\right\|
\end{aligned}$$

where the first equality is also because $Y_{t,i}$'s have the independent and identical distribution and the last two inequalities are because of the von Neumann's trace theorem. ☐

Lemma 29. *It holds that*

$$\begin{aligned}
&\sum_{t=2}^k \frac{t-1}{k(k-1)} \left\langle \frac{1}{2b}\sum_{i=1}^b u_i^\top \Sigma_t^{1/2} H_t \Sigma_t^{1/2} u_i \left(\Sigma_t^{-1/2} u_i u_i^\top \Sigma_t^{-1/2} - \Sigma_t^{-1}\right) - H_t, \Sigma_t^{-1} - H^*\right\rangle \\
\leq&\frac{A_k L(\zeta A_k + d\zeta)}{k(k-1)\tau} \cdot \sqrt{\frac{2}{b}\log\frac{k^2}{\delta}} \cdot \sqrt{\sum_{t=2}^k (t-1)^2 \left\|\Sigma_t - H^*\right\|^2}
\end{aligned} \tag{81}$$

Proof.

$$
\begin{aligned}
&\frac{1}{k(k-1)}\sum_{t=2}^{k}(t-1)\left\langle \frac{1}{2b}\sum_{i=1}^{b} u_i^\top \Sigma_t^{1/2} H_t \Sigma_t^{1/2} u_i \left(\Sigma_t^{-1/2} u_i u_i^\top \Sigma_t^{-1/2} - \Sigma_t^{-1}\right) - H_t, \Sigma_t^{-1} - H^* \right\rangle \\
\overset{(76)}{=}&\frac{1}{k(k-1)}\sum_{t=2}^{k}(t-1)\left\langle \frac{1}{2}\sum_{i=1}^{b} E_{t,i}, V_t^\top \Sigma_t^{-1/2}\left(\Sigma_t^{-1} - H^*\right)\Sigma_t^{-1/2} V_t \right\rangle \\
\overset{(77)}{=}&\sum_{t=2}^{k}\frac{(t-1)}{k(k-1)}\left\langle \frac{1}{2}\sum_{i=1}^{b}\left[\Big(X_{t,i} - \mathbb{E}[X_{t,i}]\Big) + \Big(Y_{t,i} - \mathbb{E}[Y_{t,i}]\Big)\right], V_t^\top \Sigma_t^{-1/2}\left(\Sigma_t^{-1} - H^*\right)\Sigma_t^{-1/2} V_t \right\rangle
\end{aligned}
$$

Let $z_{t,k} = \frac{(t-1)}{k(k-1)}$ and $A_t = 8\log\frac{1}{b\delta} + 16\log(t+1) + d$ just as set in Lemma 28. Using the facts $\|\Lambda_t\|_2 \le \|\Sigma_t\|_2 \cdot \|H_t\|_2 \le L/\tau$ and $\mathrm{tr}(\Sigma_t^{-1}) \le d\zeta$, resorting to Lemma 27, then with a probability at least $1 - \frac{\delta}{k^2}$, it holds that

$$
\begin{aligned}
&\sum_{t=2}^{k}\frac{(t-1)}{k(k-1)}\left\langle \frac{1}{2}\sum_{i=1}^{b}\left(X_{t,i} - \mathbb{E}[X_{t,i}]\right), V_t^\top \Sigma_t^{-1/2}\left(\Sigma_t^{-1} - H^*\right)\Sigma_t^{-1/2} V_t \right\rangle \\
\le&\sqrt{\frac{2}{b}\log\frac{k^2}{\delta}\cdot\sum_{t=2}^{k}\frac{(t-1)^2}{k^2(k-1)^2}\cdot\left(A_t\|\Lambda_t\|_2\left(\zeta A_t + \mathrm{tr}\left(\Sigma_t^{-1}\right)\right)\cdot\left\|\Sigma_t^{-1} - H^*\right\|_2\right)^2} \\
\le&\sqrt{\frac{2}{b}\log\frac{k^2}{\delta}\cdot\Big(A_k L\tau^{-1}(\zeta A_k + d\zeta)\Big)^2\cdot\sum_{t=2}^{k}\frac{(t-1)^2\left\|\Sigma_t^{-1} - H^*\right\|_2^2}{k^2(k-1)^2}} \\
=&\frac{A_k L(\zeta A_k + d\zeta)}{k(k-1)\tau}\cdot\sqrt{\frac{2}{b}\log\frac{k^2}{\delta}}\cdot\sqrt{\sum_{t=2}^{k}(t-1)^2\left\|\Sigma_t - H^*\right\|^2}
\end{aligned}
$$

□

Lemma 30. *Given $k \ge 2$ and letting $H_t = \Pi_{\mathcal{S}'}(\nabla^2 f(\mu_t))$ and $H^* = \Pi_{\mathcal{S}'}(\nabla^2 f(\mu^*))$, then it holds that*

$$
\sum_{t=2}^{k}\|H_t - H^*\|^2 \le \frac{2\gamma^2}{\rho_1\sigma}\cdot\Big(\big(f(\mu_1) - f(\mu^*)\big) + (k-1)\max\{\Delta_{\alpha,1}, \Delta_{\alpha,2}\}\Big) \tag{82}
$$

with $\rho_1 = \frac{c_2\tau\sigma}{16Lc_1\zeta}$, $\Delta_{\alpha,1}$ and $\Delta_{\alpha,2}$ being defined in Lemma 5 and 6.

Proof. Since when $\tau \cdot I \preceq \Sigma^{-1} \preceq \zeta \cdot I$, then, it is easy to check that

$$
\frac{\sigma}{\zeta}\cdot\Sigma_t^{-1} \preceq \nabla^2 f(\mu_t) \preceq \frac{L}{\tau}\cdot\Sigma_t^{-1} \tag{83}
$$

Thus, Eqn. (50) shows that

$$
f(\mu_{k+1}) - f(\mu^*) \le (1-\rho_1)\Big(f(\mu_k) - f(\mu^*)\Big) + \Delta_{\alpha,2}
$$

Combining with Lemma 5, we can obtain that for $k = 1, 2, \ldots$, it holds that

$$
f(\mu_{k+1}) - f(\mu^*) \le (1-\rho_1)\Big(f(\mu_k) - f(\mu^*)\Big) + \max\{\Delta_{\alpha,1}, \Delta_{\alpha,2}\}
$$

By the γ-Lipschitz Hessian assumption and σ-strong convexity, we can obtain that

$$\begin{aligned}
&\|H_t - H^*\|^2 \le \left\|\nabla^2 f(\mu_k) - \nabla^2 f(\mu^*)\right\|^2 \le \gamma^2 \|\mu_k - \mu^*\|^2 \le \frac{2\gamma^2}{\sigma}\Big(f(\mu_k) - f(\mu^*)\Big) \\
\le&\frac{2\gamma^2}{\sigma}\left((1-\rho_1)^{k-1}\big(f(\mu_1) - f(\mu^*)\big) + \max\{\Delta_{\alpha,1}, \Delta_{\alpha,2}\} \cdot \sum_{i=0}^{k-2}(1-\rho_1)^i\right)
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sum_{t=2}^{k}\left\|\nabla^2 f(\mu_t) - \nabla^2 f(\mu^*)\right\|^2 \le \frac{2\gamma^2}{\sigma}\sum_{t=2}^{k}\left((1-\rho_1)^{t-1}\big(f(\mu_1) - f(\mu^*)\big) + \max\{\Delta_{\alpha,1}, \Delta_{\alpha,2}\}\sum_{i=0}^{t-2}(1-\rho_1)^i\right) \\
=&\frac{2\gamma^2}{\sigma}\cdot\frac{(1-\rho_1)}{\rho_1}\left(\left(1-(1-\rho_1)^{k-1}\right)\big(f(\mu_1) - f(\mu^*)\big) + \left((k-1) - \frac{1-(1-\rho_1)^{k-1}}{\rho_1}\right)\max\{\Delta_{\alpha,1}, \Delta_{\alpha,2}\}\right) \\
\le&\frac{2\gamma^2}{\rho_1\sigma}\cdot\Big(\big(f(\mu_1) - f(\mu^*)\big) + (k-1)\max\{\Delta_{\alpha,1}, \Delta_{\alpha,2}\}\Big)
\end{aligned}$$

□

Proof of Lemma 7. Combining Lemma 23 with Eqn. (73), (68), (75), (81) (82), we have

$$\begin{aligned}
&\left\|\Sigma_{k+1}^{-1} - H^*\right\|^2 \\
\le&\frac{1}{k(k-1)}\sum_{t=2}^{k}(t-1)\left\langle \bar{G}(\Sigma_t) - H^*, \Sigma_t^{-1} - H^*\right\rangle + \frac{1}{k(k-1)}\sum_{t=2}^{k}\frac{t-1}{t}\left\|\widetilde{G}(\Sigma_t)\right\|^2 \\
\le&\frac{C_1 + C_2 + C_3}{k(k-1)}\cdot\sqrt{\sum_{t=2}^{k}(t-1)^2\|\Sigma_t - H^*\|^2} + \frac{C_4}{k}
\end{aligned} \tag{84}$$

Next, we will prove the result by induction. We assume that $\|\Sigma_t - H^*\|^2 \le C/t$ holds for all $t = 1, \ldots, k$. First, for the case $t = 1$, we have

$$\left\|\Sigma_1^{-1} - H^*\right\|^2 \le \frac{\left\|\Sigma_1^{-1} - H^*\right\|^2}{1} \le \frac{C}{1}$$

and

$$\left\|\Sigma_2^{-1} - H^*\right\|^2 \overset{(74)}{\le} \frac{4L^2(2d + 3\log(1/\delta))^2\left(\zeta(2d + 3\log(1/\delta)) + \sqrt{d}\zeta\right)^2}{b\tau^2} + \frac{d\zeta^2}{b} \le \frac{C}{2}$$

Next, we will prove the case $t = k+1$. By Lemma 13, Eqn. (84) and the definition of C, it holds that

$$\left\|\Sigma_{k+1}^{-1} - H^*\right\|^2 \le \frac{C}{k+1}$$

Thus, the induction result holds for $t = k+1$ which concludes the proof.

□