Convergence of the Exponentiated Gradient Method with Armijo Line Search

Yen-Huan Li and Volkan Cevher

Laboratory for Information and Inference Systems
École Polytechnique Fédérale de Lausanne

Abstract

Consider the problem of minimizing a convex differentiable function on the probability simplex, spectrahedron, or set of quantum density matrices. We prove that the exponentiated gradient method with Armijo line search always converges to the optimum, if the sequence of the iterates possesses a strictly positive limit point (element-wise for the vector case, and with respect to the Löwner partial ordering for the matrix case). To the best of our knowledge, this is the first convergence result for a mirror descent-type method that only requires differentiability. The proof exploits self-concordant likeness of the log-partition function, which is of independent interest.

1 Introduction

We consider the problem of minimizing a convex differentiable objective function on the probability simplex, spectrahedron, or set of quantum density matrices. Such a convex optimization problem appears in sparse regression, Poisson inverse problems, low-rank matrix estimation, and quantum state tomography, to mention a few [31, 18, 21, 28].

Regarding the structure of the constraint set, a natural approach is the exponentiated gradient (EG) method. In particular, for the probability simplex constraint case, the corresponding iteration rule is computationally cheap—projection is not required. The EG method can be viewed as a special case of mirror descent [24, 7], the interior gradient method [2, 3], and the proximal gradient method [6], with the Bregman divergence induced by Shannon or von Neumann entropy. Because of its close relationship with the multiplicative weights update method (see, e.g., [1]), the EG method was independently discovered by the computer science community: It was studied for the probability simplex constraint in [20, 16], and generalized for the spectrahedron constraint in [33], under the setup of online convex optimization.

Existing covnergence guarantees of the EG method require conditions on the objective function. If the objective function is Lipschitz, standard analysis of mirror descent shows that the exponentiated gradient method converges to the optimum [7]. If the gradient of the objective function is Lipschitz, the EG method converges either with a constant step size or Armijo line search [3]. Recently, the Lipschitz gradient condition was generalized by the notion of relative smoothness in [6, 22]. If the objective function is smooth relative to the negative von Neumann entropy, the EG method converges with a constant step size [6, 12, 22].

Notice that checking the conditions can be highly non-trivial, and there are applications where none of the conditions above hold. In Appendix A, we show that quantum state tomography, an essential task for calibrating quantum computation devices, is one such application. The interior proximal method converges for all convex differentiable objective functions, but its implementation is computationally expensive [14]. For first-order methods, there are indeed convergence guarantees that require mild differentiability conditions, though they are all for gradient descent-type methods. Bertsekas proved that the projected gradient descent with Armijo line search always converges for a differentiable objective function, when the constraint is a box or the positive orthant [8]. Gafni and Bertsekas generalized the previous result for any compact convex constraint [15]. Salzo proved the convergence of proximal variable metric methods with various line search schemes, assuming that the gradient of the objective function is uniformly continuous on any compact set [30].

In this paper, we study convergence of the EG method with Armijo line search, assuming only differentiability of the objective function. We prove that, as long as the sequence of iterates possesses a strictly positive limit point, the EG method with Armijo line search is guaranteed to converge to the optimum. In comparison to existing results, we highlight the following contributions.

- To the best of our knowledge, we give the first convergence guarantee of a mirror descent-type method[1] that only requires differentiability.
- Our convergence analysis exploits the self-concordant likeness of the log-partition function. As a by-product, we improve on the Peierls-Bogoliubov inequality, which is of independent interest; see Remark 2 for the details.

2 Problem Statement and Main Result

We consider the optimization problem

$$f^\star = \min\{ f(\rho) \mid \rho \in \mathscr{D} \}, \tag{P}$$

[1] Here we exclude the very standard projected gradient method.

where f is a convex function differentiable on $\operatorname{int}\operatorname{dom} f$, and $\mathscr{D}$ denotes the set of quantum density matrices, i.e.,

$$\mathscr{D} := \{\rho \in \mathbb{C}^{d\times d} \mid \rho \geq 0, \operatorname{Tr}\rho = 1\},$$

for some positive integer d. We assume that $f^\star > -\infty$.

This problem formulation (P) allows us to address two other constraints simultaneously:

- The probability simplex $\mathscr{P} := \{x \in \mathbb{R}^d_+ \mid \|x\|_1 = 1\}$.
- The spectrahedron $\mathscr{S} := \{X \in \mathbb{R}^{d\times d} \mid X \geq 0, \operatorname{Tr} X = 1\}$.

See Section 4.2 for the details.

Starting with some non-singular $\rho_0 \in \mathscr{D}$, the EG method iterates as

$$\rho_k = C_k^{-1} \exp\left[\log(\rho_{k-1}) - \alpha_k \nabla f(\rho_{k-1})\right], \quad \forall k \in \mathbb{N}, \tag{1}$$

where C_k is a positive real number normalizing the trace of ρ_k, and $\alpha_k > 0$ denotes the step size. Equivalently, one may write

$$\rho_k \in \arg\min\{\alpha_k \langle \nabla f(\rho_{k-1}), \sigma - \rho_{k-1}\rangle + H(\sigma, \rho_{k-1}) \mid \sigma \in \mathscr{D}\}, \tag{2}$$

where H denotes the quantum relative entropy, defined as

$$H(\rho,\sigma) := \begin{cases} \operatorname{Tr}(\rho\log\rho) - \operatorname{Tr}(\rho\log\sigma) - \operatorname{Tr}(\rho-\sigma) & \text{if } \ker\rho \subseteq \ker\sigma, \\ +\infty & \text{otherwise.} \end{cases}$$

The convention $0\log 0 := 0$ is adopted in the definition.

There are various approaches to selecting the step size. In this paper, we will focus on Armijo line search. Let $\bar{\alpha} > 0$ and $r, \tau \in]0,1[$. The Armijo line search procedure outputs $\alpha_k = r^j \bar{\alpha}$, where j is the least non-negative integer that satisfies

$$f(\rho_k) \leq f(\rho_{k-1}) + \tau \langle \nabla f(\rho_{k-1}), \rho_k - \rho_{k-1}\rangle;$$

the dependence on j lies implicitly in ρ_k. Notice that implementing Armijo line search does not require any parameter of the objective function, e.g., the Lipschitz constant of the objective function or its gradient.

Our main result is the following theorem.

Theorem 1 *Suppose that f is differentiable at every non-singular $\rho \in \mathscr{D}$. Then we have:*

1. *The Armijo line search procedure terminates in finite steps.*
2. *The sequence $(f(\rho_k))_{k\in\mathbb{N}}$ is non-increasing.*

3. *For any converging sub-sequence* $(\rho_k)_{k\in\mathscr{K}}$, $\mathscr{K}\subseteq\mathbb{N}$, *it holds that*

$$\liminf\{H(\rho_k(\beta),\rho_k)\mid k\in\mathscr{K}\}=0,\quad \forall\beta>0,$$

where we define

$$\rho_k(\beta):=\tilde{C}_k^{-1}\exp\left[\log(\rho_k)-\beta\nabla f(\rho_k)\right],\quad \forall k\in\mathbb{N};$$ □

the real number $\tilde{C}_k$ *normalizes the trace of* $\rho_k(\beta)$.

Remark 1 Statement 3 is always meaningful—due to the compactness of $\mathscr{D}$, there exists at least one converging sub-sequence of $(\rho_k)_{k\in\mathbb{N}}$. □

Taking limit, we obtain the following convergence guarantee.

Corollary 1 *If the sequence* $(\rho_k)_{k\in\mathbb{N}}$ *possesses a non-singular limit point, the sequence* $(f(\rho_k))_{k\in\mathbb{N}}$ *monotonically converges to* $f^\star$. □

PROOF Let $(\rho_k)_{k\in\mathscr{K}}$ be a sub-sequence converging to a non-singular $\rho_\infty\in\mathscr{D}$. By Statement 3 of Theorem 1, there exists a sub-sequence $(\rho_k)_{k\in\mathscr{K}'}$, $\mathscr{K}'\subseteq\mathscr{K}$, such that $H(\rho_k(\beta),\rho_k)\to 0$ as $k\to\infty$ in $\mathscr{K}'$. As ρ_∞ is non-singular, we can take the limit and obtain $H(\rho_\infty(\beta),\rho_\infty)=0$, showing that $\rho_\infty(\beta)=\rho_\infty$. Lemma B2 in the appendix then implies that ρ_∞ is a minimizer of f on $\mathscr{D}$. Since the sequence $(f(\rho_k))_{k\in\mathbb{N}}$ is non-increasing and bounded from below by $f^\star$, $\lim_{k\to\infty}f(\rho_k)$ exists. We write

$$f^\star\le\lim_{k\to\infty}f(\rho_k)=\liminf\{f(\rho_k)\mid k\in\mathbb{N}\}\le f(\rho_\infty)\le f^\star.$$ ■

It is currently unclear to us whether convergence to the optimum holds, when there does not exist a non-singular limit point; see Section 4.3 for a discussion. One way to get around is to consider solving

$$f_\lambda^\star=\min\{f(\rho)-\lambda\log\det\rho\mid\rho\in\mathscr{D}\},\tag{P-λ}$$

where λ is a positive real number.

Proposition 1 *It holds that* $\lim_{\lambda\downarrow 0}f_\lambda^\star=f^\star$. □

PROOF Notice that $-\log\det(\cdot)>0$ on $\mathscr{D}$. We write

$$\lim_{\lambda\downarrow 0}f_\lambda^\star=\inf_{\lambda>0}f_\lambda^\star=\inf_{\lambda>0}\inf_{\rho\in\mathscr{D}}f_\lambda(\rho)=\inf_{\rho\in\mathscr{D}}\inf_{\lambda>0}f_\lambda(\rho)=\inf_{\rho\in\mathscr{D}}f(\rho)=f^\star,$$

where $f_\lambda(\rho):=f(\rho)-\lambda\log\det\rho$. ■

Existence of a non-singular limit point can be easily verified in some applications. For example, hedged quantum state tomography corresponds to solving (P) with the objective function

$$f_{\mathrm{HQST}}(\rho):=f_{\mathrm{QST}}(\rho)-\lambda\log\det\rho,$$

for some $\lambda > 0$ [10], where f_{QST} is given in Section A. As discussed above, all limit points of the iterates must be non-singular. Similarly in the probability simplex constraint case, if the optimization problem involves the Burg entropy as in [13], all limit points must be element-wisely strictly positive[2].

Notation

Let g be a convex differentiable function. We denote its (effective) domain by $\mathrm{dom}\, g$, and gradient by ∇g. If g is defined on $\mathbb{R}$, we write g', g'', and g''' for its first, second, and third derivatives, respectively.

Let $A \in \mathbb{C}^{d\times d}$. We denote its largest and smallest eigenvalues by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$, respectively. We denote its Schatten p-norm by $\|A\|_p$. We will only use the Hilbert-Schmidt inner product in this paper; that is, $\langle A, B\rangle := \mathrm{Tr}(A^{\mathrm{H}}B)$ for any $A, B \in \mathbb{C}^{d\times d}$, where A^{H} denotes the Hermitian of A.

The function $\exp(\cdot)$ and $\log(\cdot)$ in (1) are matrix exponential and logarithm functions, respectively. In general, let $X \in \mathbb{C}^{d\times d}$ be Hermitian, and $X = \sum_j \lambda_j P_j$ be its spectral decomposition. Let g be a real-valued function whose domain contains $\{\lambda_j\}$. Then $g(X) := \sum_j g(\lambda_j) P_j$.

Let $\rho, \sigma \in \mathscr{D}$ be non-singular. The negative von Neumann entropy is defined as

$$h(\rho) := \mathrm{Tr}(\rho \log \rho) - \mathrm{Tr}(\rho).$$

It is easily checked that the quantum relative entropy is the Bregman divergence induced by the negative von Neumann entropy. Pinsker's inequality says that [17]

$$H(\rho, \sigma) \geq \frac{1}{2}\|\rho - \sigma\|_1^2.$$

Therefore, $H(\rho, \sigma) = 0$ implies that $\rho = \sigma$.

3 Proof of Theorem 1

The key to our analysis is the following proposition.

Proposition 2 *Let $\rho \in \mathscr{D}$ be non-singular. Suppose that*

$$\Delta := \lambda_{\max}(\nabla f(\rho)) - \lambda_{\min}(\nabla f(\rho)) > 0.$$

Then the mapping

$$\alpha \mapsto \frac{H(\rho(\alpha), \rho)}{\mathrm{e}^{\Delta\alpha}(\Delta\alpha - 1) + 1} \tag{3}$$

[2] For any element-wisely strictly positive vector $v := (v_i)_{1\leq i\leq d}$, the Burg entropy is defined as $b(v) := -\sum_{i=1}^d \log v_i$.

is non-increasing on $]0,+\infty[$. □

Proposition 2 was inspired by a lemma due to Gafni and Bertsekas [15], which says that the mapping

$$\alpha \mapsto \frac{\|\Pi_{\mathscr{D}}(\rho - \alpha\nabla f(\rho)) - \rho\|_{\mathrm{F}}}{\alpha} \tag{4}$$

is non-increasing on $[0,+\infty[$, where $\Pi_{\mathscr{D}}$ denotes projection onto $\mathscr{D}$ with respect to the Frobenius norm $\|\cdot\|_{\mathrm{F}}$. The lemma of Gafni and Bertsekas was proved by an Euclidean geometric argument; see [9] for an illustration. In comparison, we will prove Proposition 2 by exploiting the self-concordant likeness of the log-partition function.

We prove Proposition 2 in Section 3.1. Then we prove the three statements in Theorem 1 separately in the following three sub-sections. To simplify the presentation, we put some necessary technical lemmas in Appendix B.

3.1 Self-concordant Likeness of the Log-Partition Function and Proof of Proposition 2

For any non-singular $\rho \in \mathscr{D}$ and $\alpha > 0$, define

$$\varphi(\alpha;\rho) := \log \mathrm{Tr} \exp\left[\log(\rho) - \alpha\nabla f(\rho)\right],$$

which, in statistical physics, is the log-partition function of the Gibbs state for the Hamiltonian $H_\alpha := -\log(\rho) + \alpha\nabla f(\rho)$ at temperature 1. We will simply write $\varphi(\alpha)$ instead of $\varphi(\alpha;\rho)$, when the corresponding ρ is clear from the context or irrelevant.

The log-partition function is indeed closely related to the EG method, as shown by the following lemma.

Lemma 1 *For any non-singular $\rho \in \mathscr{D}$ and $\alpha > 0$, it holds that*

$$H(\rho(\alpha),\rho) = \varphi(0) - \left[\varphi(\alpha) + \varphi'(\alpha)(0-\alpha)\right]. \qquad \square$$

PROOF A direct calculation gives

$$\begin{aligned} D(\rho(\alpha),\rho) &= -\alpha\,\mathrm{Tr}(\nabla f(\rho)\rho(\alpha)) - \log \mathrm{Tr}\exp\left[\log(\rho) - \alpha\nabla f(\rho)\right] \\ &= \alpha\varphi'(\alpha) - \varphi(\alpha). \end{aligned}$$

Notice that $\varphi(0) = 0$. ■

We say that a three times continuously differentiable convex function g is *μ-self-concordant like*, if and only if $|g'''(x)| \le \mu g''(x)$ for all x [4, 5, 32].

Lemma 2 *For any non-singular $\rho \in \mathscr{D}$, the function $\varphi(\alpha)$ is Δ-self-concordant like, where $\Delta := \lambda_{\max}(\nabla f(\rho)) - \lambda_{\min}(\nabla f(\rho))$.* □

PROOF Lemma B3 shows that

$$\varphi''(\alpha) = \mathsf{E}\left(\eta_\alpha - \mathsf{E}\,\eta_\alpha\right)^2, \quad \varphi'''(\alpha) = \mathsf{E}\left(\eta_\alpha - \mathsf{E}\,\eta_\alpha\right)^3,$$

where η_α is a random variable taking values in $[-\lambda_{\max}(\nabla f(\rho)), -\lambda_{\min}(\nabla f(\rho))]$. The lemma follows. ■

The following sandwich inequality follows from self-concordant likeness [32]. We defer the proof to Section C.

Lemma 3 *Suppose that $\Delta > 0$. For any non-singular $\rho \in \mathscr{D}$, it holds that*

$$\begin{aligned}\frac{\left(\mathrm{e}^{-\Delta\alpha} + \Delta\alpha - 1\right)}{\Delta^2}\varphi''(\alpha) &\le \varphi(0) - \left[\varphi(\alpha) + \varphi'(\alpha)(0-\alpha)\right] \\ &\le \frac{\left(\mathrm{e}^{\Delta\alpha} - \Delta\alpha - 1\right)}{\Delta^2}\varphi''(\alpha).\end{aligned}$$

□

Remark 2 The lower bound improves upon the Peierls-Bogoliubov inequality [27], which says that

$$0 \le \varphi(0) - \left[\varphi(\alpha) + \varphi'(\alpha)(0-\alpha)\right].$$

Notice that lower bound provided by Lemma 3 is always non-negative. □

Now we are ready to prove Proposition 2.

PROOF (PROPOSITION 2) We look for a differentiable function $\chi :]0, +\infty[\to]0, +\infty[$, such that the mapping

$$g(\alpha) := \frac{H(\rho(\alpha), \rho)}{\chi(\alpha)}$$

is non-increasing on $]0, +\infty[$. Note that g is non-increasing if and only if $g' \le 0$ on $]0, +\infty[$. Applying Lemma 1, a direct calculation gives

$$g'(\alpha) = \frac{\alpha\varphi''(\alpha)\chi(\alpha) - \left\{\varphi(0) - \left[\varphi(\alpha) + \varphi'(\alpha)(0-\alpha)\right]\right\}\chi'(\alpha)}{\left[\chi'(\alpha)\right]^2}.$$

Therefore, $g'(\alpha) \le 0$ if and only if the numerator is negative, i.e.,

$$(\log\chi)'(\alpha) \ge \frac{\alpha\varphi''(\alpha)}{\varphi(0) - \left[\varphi(\alpha) + \varphi'(\alpha)(0-\alpha)\right]},$$

where we have used the fact that $\chi'/\chi = (\log\chi)'$. By Lemma 3, we can set

$$(\log\chi)'(\alpha) = \frac{\Delta^2\alpha}{\mathrm{e}^{-\Delta\alpha} + \Delta\alpha - 1}.$$

Solving the equation gives $\chi(\alpha) := \mathrm{e}^{\Delta\alpha}(\Delta\alpha - 1) + 1$. ■

For convenience, we will apply Proposition 2 via the following corollary.

Corollary 2 *Let $\rho \in \mathscr{D}$ be non-singular and $\bar{\alpha} > 0$. Suppose that $\Delta > 0$. It holds that*

$$\frac{H(\rho(\alpha),\rho)}{\alpha^2} \geq \kappa H(\rho(\bar{\alpha}),\rho), \quad \forall \alpha \in]0,\bar{\alpha}],$$

where $\kappa := \left\{2\left[\mathrm{e}^{\Delta\bar{\alpha}}(\Delta\bar{\alpha}-1)+1\right]\right\}^{-1}\Delta^2$. □

PROOF Define $g(\alpha) := \mathrm{e}^{\Delta\alpha}(\Delta\alpha-1)+1-(\Delta^2/2)\alpha^2$. Then $g(0)=0$, and

$$g'(\alpha) = \alpha\left[\mathrm{e}^{\Delta\alpha}\Delta^2-\Delta^2\right] \geq \alpha(\Delta^2-\Delta^2) = 0, \quad \forall \alpha \in]0,+\infty[.$$

Therefore, $g(\alpha) \geq 0$ on $]0,+\infty[$, i.e.,

$$\mathrm{e}^{\Delta\alpha}(\Delta\alpha-1)+1 \geq \frac{\Delta^2}{2}\alpha^2, \quad \forall \alpha \in]0,+\infty[.$$

By Proposition 2, we write

$$\frac{H(\rho(\alpha),\rho)}{\frac{\Delta^2}{2}\alpha^2} \geq \frac{H(\rho(\alpha),\rho)}{\mathrm{e}^{\Delta\alpha}(\Delta\alpha-1)+1} \geq \frac{H(\rho(\bar{\alpha}),\rho)}{\mathrm{e}^{\Delta\bar{\alpha}}(\Delta\bar{\alpha}-1)+1}, \quad \forall \alpha \in]0,\bar{\alpha}].$$ ■

3.2 Proof of Statement 1

The first statement is a direct consequence of the following proposition.

Proposition 3 *For every non-singular $\rho \in \mathscr{D}$, there exists some $\alpha_\rho > 0$ such that*

$$f(\rho(\alpha)) \leq f(\rho) + \tau\langle \nabla f(\rho), \rho(\alpha)-\rho\rangle, \quad \forall \alpha \in [0,\alpha_\rho]. \tag{5}$$

□

Recall that τ is the parameter in Armijo line search.

PROOF If ρ is a minimizer, by Lemma B2, we have $\rho(\alpha)=\rho$ for all $\alpha \in [0,+\infty[$, and the proposition follows. Suppose that ρ is not a minimizer in the rest of this proof. By Lemma B2, we have $H(\rho(\alpha),\rho) > 0$ for all $\alpha \in]0,+\infty[$. By the mean-value theorem, we write

$$f(\rho(\alpha)) - f(\rho) = \langle \nabla f(\sigma), \rho(\alpha)-\rho\rangle,$$

for some σ in the line segment joining $\rho(\alpha)$ and ρ. Then (5) can be equivalently written as

$$\langle \nabla f(\sigma) - \nabla f(\rho), \rho(\alpha)-\rho\rangle \leq -(1-\tau)\langle \nabla f(\rho), \rho(\alpha)-\rho\rangle, \quad \forall \alpha \in [0,\alpha_\rho]. \tag{6}$$

By Lemma B1, (6) holds if

$$\langle \nabla f(\sigma) - \nabla f(\rho), \rho(\alpha)-\rho\rangle \leq \frac{(1-\tau)H(\rho(\alpha),\rho)}{\alpha}, \quad \forall \alpha \in [0,\alpha_\rho]. \tag{7}$$

Consider two cases.

- If $\lambda_{\max}(\nabla f(\rho)) = \lambda_{\min}(\nabla f(\rho))$, then $\nabla f(\rho)$ is a multiple of the identity. We have

$$\langle \nabla f(\rho), \sigma - \rho \rangle = 0, \quad \forall \sigma \in \mathscr{D};$$

showing that ρ is a minimizer. By Lemma B2, the proposition follows for every $\alpha_\rho > 0$.

- Otherwise, set $\alpha_\rho \leq \bar{\alpha}$. By Corollary 2, there exists some $\kappa > 0$, such that

$$\frac{H(\rho(\alpha), \rho)}{\alpha} \geq \sqrt{H(\rho(\alpha), \rho)} \sqrt{\kappa H(\rho(\bar{\alpha}), \rho)}, \quad \forall \alpha \in [0, \alpha_\rho].$$

Applying Hölder's inequality and Pinsker's inequality, we write

$$\begin{aligned} \langle \nabla f(\sigma) - \nabla f(\rho), \rho(\alpha) - \rho \rangle &\leq \|\nabla f(\sigma) - \nabla f(\rho)\|_\infty \|\rho(\alpha) - \rho\|_1 \\ &\leq \|\nabla f(\sigma) - \nabla f(\rho)\|_\infty \sqrt{2 H(\rho(\alpha), \rho)}. \end{aligned}$$

Then (7) holds if

$$\|\nabla f(\sigma) - \nabla f(\rho)\|_\infty \sqrt{2} \leq (1 - \tau) \sqrt{\kappa H(\rho(\bar{\alpha}), \rho)}, \quad \forall \alpha \in [0, \alpha_\rho]$$

Recall that a convex differentiable function is continuously differentaible [29]. Notice that $\rho(\alpha)$ is continuous in α. As the right-hand side is a strictly positive constant by Lemma B2, the proposition follows for a small enough α_ρ. ■

3.3 Proof of Statement 2

By the definition of Armijo line search and Lemma B1, we have

$$f(\rho_k) \leq f(\rho_{k-1}) + \tau \langle \nabla f(\rho_{k-1}), \rho_k - \rho_{k-1} \rangle \leq f(\rho_{k-1}) - \frac{\tau H(\rho_k, \rho_{k-1})}{\alpha_k}.$$

As the quantum relative entropy D is always non-negative, it follows that the sequence $(f(\rho_k))_{k \in \mathbb{N}}$ is non-increasing.

3.4 Proof of Statement 3

If ρ_k is a minimizer for some $k \in \mathbb{N}$, by Lemma B2, it holds that $\rho_{k'} = \rho_k$ for all $k' > k$, and the statement trivially follows. In the rest of this sub-section, we assume that ρ_k is not a minimizer for all k; then by Lemma B2, it holds that $\rho_k \neq \rho_{k-1}$ for all $k \in \mathbb{N}$.

Let $(\rho_k)_{k \in \mathscr{K}}$ be a sub-sequence converging to a limit point $\rho_\infty \in \mathscr{D}$, which exists due to the compactness of $\mathscr{D}$. Then ρ_∞ must be non-singular; otherwise, monotonicity of the sequence $(f(\rho_k))_{k \in \mathbb{N}}$ (Statement 2 of Theorem 1) cannot hold. As f is continuously differentiable, it holds that

$$\frac{\Delta_\infty}{2} \leq \lambda_{\max}(\nabla f(\rho_k)) - \lambda_{\min}(\nabla f(\rho_k)) \leq 2\Delta_\infty, \tag{8}$$

for large enough $k \in \mathscr{K}$, where $\Delta_\infty := \lambda_{\max}(\nabla f(\rho_\infty)) - \lambda_{\min}(\nabla f(\rho_\infty)$.

Lemma 4 *If* $\Delta_\infty = 0$*, then* $\liminf\{ H(\rho_k(\beta), \rho_k) \mid k \in \mathscr{K} \} = 0$ *for every* $\beta \in [0, +\infty)$. □

PROOF Define $\Delta_k := \lambda_{\max}(\nabla f(\rho_k)) - \lambda_{\min}(\nabla f(\rho_k))$; then $\Delta_k \to \Delta_\infty = 0$. Define $\varphi_k : \alpha \mapsto \varphi(\alpha; \rho_k)$. By Lemma 3 and Corollary B1, we have

$$\begin{aligned}\varphi_k(0) - \left[\varphi_k(\beta) + \varphi_k'(\beta)(0-\beta)\right] &\leq \frac{\left(\mathrm{e}^{\Delta_k \beta} - \Delta_k \beta - 1\right)}{\Delta_k^2} \varphi_k''(\beta) \\ &\leq \frac{\left(\mathrm{e}^{\Delta_k \beta} - \Delta_k \beta - 1\right)}{4}.\end{aligned}$$

By Lemma 1, we obtain

$$\begin{aligned}0 &\leq \liminf\{ H(\rho_k(\beta), \rho_k) \mid k \in \mathscr{K} \} \\ &= \liminf\{ \varphi_k(0) - \left[\varphi_k(\beta) + \varphi_k'(\beta)(0-\beta)\right] \mid k \in \mathscr{K} \} \\ &\leq \frac{\mathrm{e}^0 - 0 - 1}{4} = 0.\end{aligned}$$

■

Suppose that $\Delta_\infty > 0$. We have the following analogy of Corollary 2 for large enough $k \in \mathscr{K}$:

Corollary 3 *Suppose that* $\Delta_\infty > 0$ *and* ρ_k *is not a minimizer for every* $k \in \mathscr{K}$*. There exists some* $\kappa > 0$*, such that*

$$\frac{H(\rho_k(\alpha), \rho_k)}{\alpha^2} \geq \kappa H(\rho_k(\bar{\alpha}), \rho_k), \quad \forall \alpha \in]0, \bar{\alpha}],$$

for large enough $k \in \mathscr{K}$. □

PROOF Recall that (8) provides both upper and lower bounds of $\lambda_{\max}(\nabla f(\rho_k)) - \lambda_{\min}(\nabla f(\rho_k))$, for large enough $k \in \mathscr{K}$. With regard to Corollary 2, it suffices to set

$$\kappa = \frac{\Delta_\infty^2}{4\left[\mathrm{e}^{2\Delta_\infty \bar{\alpha}}(2\Delta_\infty \bar{\alpha} - 1) + 1\right]}.$$

■

Based on Corollary 3, we prove the following proposition.

Proposition 4 *Suppose that* $\Delta_\infty > 0$ *and* ρ_k *is not a minimizer for every* $k \in \mathscr{K}$*. It holds that* $\liminf\{ H(\rho_k(\bar{\alpha}), \rho_k) \mid k \in \mathscr{K} \} = 0$. □

The proof of Proposition 4 can be found in Section D, which essentially follows the strategy of Gafni and Bertsekas [15] with necessary modifications.

To summarize, we have proved that for any converging sub-sequence $(\rho_k)_{k \in \mathscr{K}}$, there exists some $\gamma > 0$ such that

$$\liminf\{ H(\rho_k(\gamma), \rho_k) \mid k \in \mathscr{K} \} = 0.$$

For the case where ρ_k is a minimizer for some $k \in \mathscr{K}$ or $\Delta_\infty = 0$, γ can be any strictly positive real number. Otherwise, we set $\gamma = \bar{\alpha}$ by Proposition 4.

By Lemma 1 and Lemma 3, it holds that

$$\begin{aligned} 0 &\leq \liminf \left\{ \frac{\left(\mathrm{e}^{-(1/2)\Delta_\infty \gamma} + (1/2)\Delta_\infty \gamma - 1\right)}{\gamma^2} \varphi_k''(\gamma) \;\middle|\; k \in \mathscr{K} \right\} \\ &\leq \liminf \{ H(\rho_k(\gamma), \rho_k) \mid k \in \mathscr{K} \} = 0, \end{aligned}$$

showing that $\liminf \{ \varphi_k''(\gamma) \mid k \in \mathscr{K} \} = 0$. Applying Lemma 1 and Lemma 3 again, we obtain

$$\begin{aligned} 0 &\leq \liminf \{ H(\rho_k(\beta), \rho_k) \mid k \in \mathscr{K} \mid k \in \mathscr{K} \} \\ &\leq \liminf \left\{ \frac{\left(\mathrm{e}^{2\Delta_\infty \beta} - 2\Delta_\infty \beta - 1\right)}{\beta^2} \varphi_k''(\beta) \;\middle|\; k \in \mathscr{K} \right\} = 0, \end{aligned}$$

for any $\beta \in]0, +\infty[$. This proves Statement 3 of Theorem 1.

4 Discussions

We give three remarks regarding the convergence result and its proof.

4.1 Importance of Self-Concordant Likeness

With regard to (4), one may suspect whether it suffices, for the convergence analysis, to prove the following: There exists some $\epsilon > 0$, such that the mapping $\alpha \mapsto \alpha^{-\epsilon} H(\rho(\alpha), \rho)$ is non-increasing on $]0, \bar{\alpha}]$ for every non-singular $\rho \in \mathscr{D}$. Indeed, following the proof strategy for Proposition 2, we obtain the following result *without self-concordant likeness*.

Proposition 5 *Let $\rho \in \mathscr{D}$ be non-singular. Define*

$$M := \sup \{ \varphi''(\alpha; \rho) \mid \alpha \in]0, \bar{\alpha}[\}, \quad m := \inf \{ \varphi''(\alpha; \rho) \mid \alpha \in]0, \bar{\alpha}[\}.$$

Suppose that $m > 0$. Then the mapping $\alpha \mapsto \alpha^{-\epsilon} H(\rho(\alpha), \rho)$ is non-increasing on $]0, \bar{\alpha}[$, where $\epsilon := 2M/m$. □

Remark 3 For the case where $m = 0$, Lemma B3 implies that ∇f must be a multiple of the identity. Then it is easily checked that ρ is a minimizer as it verifies the optimality condition. □

Then in the proof of Proposition 3, for example, the condition we need to verify becomes:

$$\|\nabla f(\sigma) - \nabla f(\rho)\|_\infty \sqrt{2} \leq (1 - \tau) \alpha^{\epsilon/2 - 1} \sqrt{\frac{H(\rho(\bar{\alpha}), \rho)}{\bar{\alpha}^2}}, \quad \forall \alpha \in [0, \alpha_\rho].$$

Notice that $\epsilon \geq 2$ by definition. Both sides can converge to zero as $\alpha \to 0$, so in general, there does not exist a small enough α_ρ that verifies the condition. Moreover, because $\alpha^\epsilon \leq \alpha^2$ for $\alpha \in [0,1]$, it is impossible to obtain an analogue of Corollary 2.

The point in our analysis is to show that there exists some $\chi(\alpha)$, bounded from below by α^2 for every α close to zero, such that the mapping $\alpha \mapsto H(\rho(\alpha),\rho)/\chi(\alpha)$ is non-increasing. This is where self-concordant likeness of the log-partition function comes into play.

4.2 Extensions for the Probability Simplex and Spectrahedron Constraints

The EG method can be extended for the spectrahedron and probability simplex constraints; in fact, the EG method is arguably better known for these two cases [3, 7, 20, 33]. For the former case, the iteration rule writes exactly the same as (1), and is equivalent to (2) with $\mathscr{D}$ replaced by the spectrahedron $\mathscr{S}$. For the latter case, the iteration rule becomes element-wise (see, e.g., [7]) and is equivalent to (2), with $\mathscr{D}$ replaced by the probability simplex $\mathscr{P}$, and the quantum relative entropy replaced by the (classical) relative entropy. The Armijo line search rule applies without modification.

It is easily checked that our proof holds without modification for the spectrahedron constraint. As a vector in $\mathbb{R}^d$ is equivalent to a diagonal matrix in $\mathbb{R}^{d\times d}$, it is easily checked that the statements in Theorem 1 applies to the probability simplex constraint. Corollary 1 also holds true for these two constraints with slight modification—for the probability simplex constraint, non-singularity should be replaced by element-wise strict positivity.

4.3 Convergence with Possibly Singular Limit Points

Corollary 1 requires existence of at least one non-singular limit point. Can this condition be removed?

Suppose that the sequence $(\rho_k)_{k\in\mathbb{N}}$ has a possibly singular limit point ρ_∞, around which ∇f is locally L-Lipschitz continuous with respect to the Schatten 1-norm. Let $(\rho_k)_{k\in\mathscr{K}}$, $\mathscr{K} \subseteq \mathbb{N}$, be a sub-sequence converging to ρ_∞. Then following the proof of the second part of Proposition 4, it is easily checked that $\liminf\{\alpha_k \mid k \in \mathscr{K}\} = 0$ implies

$$\alpha_k \geq \frac{L}{r(1-\tau)},$$

a contradiction; hence, $\liminf\{\alpha_k \mid k \in \mathscr{K}\}$ must be strictly positive. Then following the proof in [3], it holds that the sequence $(f(\rho_k))_{k\in\mathbb{N}}$ monotonically converges to the optimal value.

In general without the local Lipschitz gradient condition, we conjecture that convergence to the optimum cannot be guaranteed. However, we have not found a counter-example.

5 Conclusions

Assuming only differentiability of the objective function, we have proved that the EG method with Armijo line search monotonically converges to the optimum, if the sequence of iterates possesses a non-singular limit point. Our proof exploits the self-concordant likeness of the log-partition function, which is of independent interest; in particular, Lemma 3 improves upon the Peierls-Bogoliubov inequality. Our result extends for the probability simplex and spectrahedron constraints. If a non-singular limit point may not exist, we conjecture that convergence cannot be guaranteed without additional condition on the objective function.

A Inapplicability of Existing Convergence Guarantees to Quantum State Tomography

Quantum state tomography is the task of estimating the state of a quantum systems, which is essential to calibrating quantum computation devices [28, 19]. Numerically, it corresponds to solving (P) with the objective function

$$f_{\mathrm{QST}}(\rho) := -\sum_{i=1}^{n} \log \mathrm{Tr}(M_i \rho),$$

where M_i are positive semi-definite matrices given by the experimental data.

The following proposition shows that existing convergence guarantees for the EG method do not apply to quantum state tomography.

Proposition 6 *The function f_{QST} is not Lipschitz, its gradient is not Lipschitz, and it is not smooth relative to the negative von Neumann entropy.* □

PROOF Consider the two-dimensional case, where $\rho = (\rho_{i,j})_{1\le i,j\le 2} \in \mathbb{C}^{2\times 2}$. Define $e_1 := (1,0)$ and $e_2 := (0,1)$. Suppose that there are only two summands, with $M_1 = e_1 \otimes e_1$ and $M_2 = e_2 \otimes e_2$. Then we have $f(\rho) = -\log \rho_{1,1} - \log \rho_{2,2}$. It suffices to disprove all properties for this specific f on the set of diagonal density matrices. Hence, we will focus on the function $g(x,y) := -\log x - \log y$, defined for any $x, y > 0$ such that $x + y = 1$.

As either x or y can be arbitrarily close to zero, g cannot be Lipschitz continuous in itself or its gradient due to the logarithmic functions. Define the entropy function

$$h(x,y) := -x \log x - y \log y + x + y,$$

with the convention $0\log 0 = 0$. Then g is L-smooth relative to the relative entropy, if and only if $-Lh - g$ is convex. It suffices to check the positive semi-definiteness of the Hessian of $-Lh - g$. A necessary condition for the Hessian to be positive semi-definite is that

$$-L\frac{\partial^2 h}{\partial x^2}(x,y) - \frac{\partial^2 g}{\partial x^2}(x,y) = \frac{L}{x} - \frac{1}{x^2} \geq 0,$$

for all $x \in]0,1[$, which cannot hold for $x < (1/L)$, for any fixed $L > 0$. ■

We note that similar objective functions can be found in positive linear inverse problems, positron emission tomography, portfolio selection, and Poisson phase retrieval [11, 23, 26, 34].

B Technical Lemmas Necessary for Section 3

Define

$$\rho(\alpha) := C_\rho^{-1} \exp\left[\log(\rho) - \alpha \nabla f(\rho)\right],$$

for every non-singular $\rho \in \mathscr{D}$ and $\alpha \geq 0$, where C_ρ is the positive real number normalizing the trace of $\rho(\alpha)$.

Lemma B1 *For every non-singular $\rho \in \mathscr{D}$ and $\alpha > 0$, it holds that*

$$\langle \nabla f(\rho), \rho(\alpha) - \rho \rangle \leq -\frac{H(\rho(\alpha), \rho)}{\alpha}.$$ □

PROOF The equivalent formulation of the EG method, (2), implies that

$$\alpha \langle \nabla f(\rho), \rho(\alpha) - \rho \rangle + H(\rho(\alpha), \rho) \leq \alpha \langle \nabla f(\rho), \rho - \rho \rangle + H(\rho, \rho) = 0.$$ ■

Lemma B2 *Let $\rho \in \mathscr{D}$ be non-singular. If ρ is a minimizer of f on $\mathscr{D}$, then $\rho(\alpha) = \rho$ for all $\alpha \geq 0$. If $\rho(\alpha) = \rho$ for some $\alpha > 0$, then ρ is a minimizer of f on $\mathscr{D}$.* □

PROOF The optimality condition says that $\rho \in \operatorname{int}\mathscr{D}$ is a minimizer, if and only if

$$\langle \nabla f(\rho), \sigma - \rho \rangle \geq 0, \quad \forall \sigma \in \mathscr{D}.$$

For any $\alpha > 0$, we can equivalently write

$$\langle \alpha \nabla f(\rho) + \left[\nabla h(\rho) - \nabla h(\rho)\right], \sigma - \rho \rangle \geq 0, \quad \forall \sigma \in \mathscr{D}, \tag{9}$$

where h denotes the negative von Neumann entropy function, i.e.,

$$h(\rho) := \operatorname{Tr}(\rho \log \rho) - \operatorname{Tr}\rho.$$

Notice that the quantum relative entropy H is the Bregman divergence induced by the negative von Neumann entropy. It is easily checked, again by the optimality condition, that (9) is equivalent to

$$\rho = \arg\min\{\alpha \langle \nabla f(\rho), \sigma - \rho \rangle + H(\sigma, \rho) \mid \sigma \in \mathscr{D}\} = \rho(\alpha).$$ ■

For every non-singular $\rho \in \mathscr{D}$ and $\alpha \geq 0$, define

$$G := -\nabla f(\rho), \quad H_\alpha := \log \rho + \alpha G.$$

Let $G = \sum_j \lambda_j P_j$ be the spectral decomposition of G. Define η_α as a random variable satisfying

$$\mathsf{P}\left(\eta_\alpha = \lambda_j\right) = \frac{\mathrm{Tr}\left(P_j \exp(H_\alpha)\right)}{\mathrm{Tr}\exp(H_\alpha)}; \tag{10}$$

it is easily checked that $\mathsf{P}\left(\eta_\alpha = \lambda_j\right) > 0$ for all j, and the probabilities sum to one.

Lemma B3 *For any $\alpha \in \mathbb{R}$, it holds that*

$$\varphi'(\alpha) = \mathsf{E}\,\eta_\alpha, \quad \varphi''(\alpha) = \mathsf{E}\left(\eta_\alpha - \mathsf{E}\,\eta_\alpha\right)^2, \quad \varphi'''(\alpha) = \mathsf{E}\left(\eta_\alpha - \mathsf{E}\,\eta_\alpha\right)^3. \qquad \square$$

PROOF Notice that

$$\mathsf{E}\,\eta_\alpha^n = \frac{\mathrm{Tr}(G^n \exp(H_\alpha))}{\mathrm{Tr}\exp(H_\alpha)},$$

for any $n \in \mathbb{N}$. Define $\sigma_\alpha := \exp(H_\alpha)/\mathrm{Tr}\exp(H_\alpha)$. A direct calculation gives

$$\begin{aligned}
\varphi'(\alpha) &= \mathrm{Tr}(G\sigma_\alpha), \quad \varphi''(\alpha) = \mathrm{Tr}(G^2\sigma_\alpha) - (\mathrm{Tr}(G\sigma_\alpha))^2, \\
\varphi'''(\alpha) &= \mathrm{Tr}(G^3\sigma_\alpha) - 3\,\mathrm{Tr}(G^2\sigma_\alpha)\,\mathrm{Tr}(G\sigma_\alpha) + 2\,(\mathrm{Tr}(G\sigma_\alpha))^3.
\end{aligned}$$

The lemma follows. ■

Since η_α is a bounded random variable, it follows that φ'' is bounded from above.

Corollary B1 *It holds that $\varphi''(\alpha) \leq (1/4)\Delta^2$, where $\Delta := \lambda_{\max}(\nabla f(\rho)) - \lambda_{\min}(\nabla f(\rho))$.* □

PROOF Recall that the variance of a random variable taking values in $[a,b]$ is bounded from above by $(b-a)^2/4$. ■

C Proof of Lemma 3

Recall the random variable η_α defined in (10). Suppose that $\varphi''(\alpha) = 0$ for some $\alpha \in [0, +\infty[$. Then we have $\eta_\alpha = 0$ almost surely, but this implies that $\Delta = 0$, a contradiction. Therefore, we have $\varphi''(\alpha) > 0$ for all $\alpha \in [0, +\infty[$.

We prove a general result. Let $\psi : \mathbb{R} \to \mathbb{R}$ be a μ-self-concordant like function. Suppose that $\psi''(t) > 0$ for all t. Consider the function $\chi(t) := \log\left(\psi''(t)\right)$. We write, by the self-concordant likeness of ψ, that

$$|\chi'(t)| = \frac{|\psi'''(t)|}{\psi''(t)} \leq \mu, \quad \forall\, t \in \mathbb{R}.$$

Then, for any $t_1, t_2 \in \mathbb{R}$, we have

$$|\chi(t_1) - \chi(t_2)| = \left|\log\left(\psi''(t_1)\right) - \log\left(\psi''(t_2)\right)\right| \le \mu|t_2 - t_1|;$$

that is,

$$\mathrm{e}^{-\mu|t_2-t_1|}\psi''(t_2) \le \psi''(t_1) \le \mathrm{e}^{\mu|t_2-t_1|}\psi''(t_2).$$

Applying the Newton-Leibniz formula, we obtain

$$\begin{aligned}
\psi'(t_2) - \psi'(t_1) &= \int_0^1 \psi''(t_1 + \tau(t_2 - t_1))(t_2 - t_1)\,\mathrm{d}\tau \\
&\le \int_0^1 \mathrm{e}^{\mu\tau|t_2-t_1|}\psi''(t_1)(t_2 - t_1)\,\mathrm{d}\tau \\
&= \left(\frac{\mathrm{e}^{\mu|t_2-t_1|} - 1}{\mu|t_2 - t_1|}\right)\psi''(t_1)(t_2 - t_1);
\end{aligned}$$

similarly, we obtain

$$\psi'(t_2) - \psi'(t_1) \ge -\left(\frac{\mathrm{e}^{-\mu|t_2-t_1|} - 1}{\mu|t_2 - t_1|}\right)\psi''(t_1)(t_2 - t_1).$$

Applying the Newton-Leibniz formula again, we obtain

$$\begin{aligned}
\psi(t_2) - \psi(t_1) &= \int_0^1 \psi'(t_1 + \tau(t_2 - t_1))(t_2 - t_1)\,\mathrm{d}\tau \\
&= \psi'(t_1)(t_2 - t_1) + \int_0^1 \left(\psi'(t_1 + \tau(t_2 - t_1)) - \psi'(t_1)\right)(t_2 - t_1)\,\mathrm{d}\tau \\
&\le \psi'(t_1)(t_2 - t_1) + \int_0^1 \left(\frac{\mathrm{e}^{\mu\tau|t_2-t_1|} - 1}{\mu\tau|t_2 - t_1|}\right)\psi''(t_1)\tau(t_2 - t_1)^2\,\mathrm{d}\tau \\
&= \psi'(t_1)(t_2 - t_1) + \frac{\left(\mathrm{e}^{\mu|t_2-t_1|} - \mu|t_2 - t_1| - 1\right)}{\mu^2}\psi''(t_1);
\end{aligned}$$

similarly, we obtain

$$\psi(t_2) - \psi(t_1) \ge \psi'(t_1)(t_2 - t_1) + \frac{\left(\mathrm{e}^{-\mu|t_2-t_1|} + \mu|t_2 - t_1| - 1\right)}{\mu^2}\psi''(t_1).$$

Lemma 3 follows from setting $\psi = \varphi$, $\mu = \Delta$, $t_2 = 0$, and $t_1 = \alpha$.

D Proof of Proposition 4

Suppose that $\underline{\alpha} := \liminf\{\alpha_k \mid k \in \mathcal{K}\} > 0$. We write

$$\begin{aligned} f(\rho_k) - f(\rho_{k+1}) &\geq -\tau \langle \nabla f(\rho_k), f(\rho_{k+1}) - f(\rho_k) \rangle \\ &\geq \tau \alpha_k^{-1} H(\rho_{k+1}, \rho_k) \\ &= \tau \alpha_k \alpha_k^{-2} H(\rho_k(\alpha_k), \rho_k) \\ &\geq \tau \underline{\alpha} \kappa H(\rho_k(\bar{\alpha}), \rho_k) \\ &\geq 0, \end{aligned}$$

for large enough $k \in \mathcal{K}$, where the first inequality follows from the Armijo line search rule, the second follows from Lemma B1, and the third follows from Corollary 2. Taking limit, we obtain that $H(\rho_k(\bar{\alpha}), \rho_k) \to 0$ as $k \to \infty$ in $\mathcal{K}$.

Suppose that $\liminf\{\alpha_k \mid k \in \mathcal{K}\} = 0$. Let $(\alpha_k)_{k \in \mathcal{K}'}$, $\mathcal{K}' \subseteq \mathcal{K}$, be a sub-sequence converging to zero. According to the Armijo rule, we have

$$f(\rho_k(r^{-1}\alpha_k)) - f(\rho_k) > \tau \langle \nabla f(\rho_k), \rho(r^{-1}\alpha_k) - \rho_k(\alpha_k) \rangle, \tag{11}$$

for large enough $k \in \mathcal{K}$. The mean value theorem says that the left-hand side equals $\langle \nabla f(\sigma), \rho_k(r^{-1}\alpha_k) - \rho_k \rangle$ for some σ in the line segment jointing $\rho_k(r^{-1}\alpha_k)$ and ρ_k. Then (11) can be equivalently written as

$$\langle \nabla f(\sigma) - \nabla f(\rho_k), \rho_k(r^{-1}\alpha_k) - \rho_k \rangle > -(1-\tau) \langle \nabla f(\rho_k), \rho_k(r^{-1}\alpha_k) - \rho_k(\alpha_k) \rangle. \tag{12}$$

By Pinsker's inequality and Hölder's inequality, we obtain

$$\begin{aligned} \|\nabla f(\sigma) - \nabla f(\rho_k)\|_\infty \sqrt{2H(\rho_k(r^{-1}\alpha_k), \rho_k)} &\geq \|\nabla f(\sigma) - \nabla f(\rho_k)\|_\infty \|\rho_k(r^{-1}\alpha_k), \rho_k\|_1 \\ &\geq \langle \nabla f(\sigma) - \nabla f(\rho_k), \rho_k(r^{-1}\alpha_k) - \rho_k \rangle. \end{aligned} \tag{13}$$

for large enough $k \in \mathcal{K}$. Notice that $r^{-1}\alpha_k \leq \bar{\alpha}$ for large enough $k \in \mathcal{K}$. By Lemma B1 and Corollary 3, we obtain

$$\begin{aligned} -\langle \nabla f(\rho_k), \rho_k(r^{-1}\alpha_k) - \rho_k(\alpha_k) \rangle &\geq \frac{H(\rho_k(r^{-1}\alpha_k), \rho_k)}{r^{-1}\alpha_k} \\ &\geq \sqrt{\kappa H(\rho_k(\bar{\alpha}), \rho_k)} \sqrt{H(\rho_k(r^{-1}\alpha_k), \rho_k)}, \end{aligned} \tag{14}$$

for large enough $k \in \mathcal{K}$. Since $H(\rho_k(r^{-1}\alpha_k), \rho_k)$ is strictly positive for all $k \in \mathcal{K}'$ by assumption, (12), (13), and (14) imply

$$\|\nabla f(\sigma) - \nabla f(\rho_k)\|_\infty > (1-\tau) \sqrt{\frac{\kappa H(\rho_k(\bar{\alpha}), \rho_k)}{2}} \geq 0.$$

Taking limits, we obtain that $H(\rho_k(\bar{\alpha}), \rho_k) \to 0$ a $k \to \infty$ in $\mathcal{K}'$.

Acknowledgements

We thank Ya-Ping Hsieh for his comments. This work was supported by SNF 200021-146750 and ERC project time-data 725594.

This is a pre-print of an article published in the Journal of Optimization Theory and Applications. The final authenticated version is available online at: `https://doi.org/10.1007/s10957-018-1428-9`.

References

[1] ARORA, S., HAZAN, E., AND KALE, S. The multiplicative weights update method: A meta-algorithm and applications. *Theory Comput. 8* (2012), 121–164.

[2] AUSLENDER, A., AND TEBOULLE, M. Interior gradient and epsilon-subgradient descent methods for constrained convex minimization. *Math. Oper. Res. 29*, 1 (2004), 1–26.

[3] AUSLENDER, A., AND TEBOULLE, M. Interior gradient and proximal methods for convex and conic optimization. *SIAM J. Optim. 16*, 3 (2006), 697–725.

[4] BACH, F. Self-concordant analysis for logistic regression. *Electron. J. Stat. 4* (2010), 384–414.

[5] BACH, F. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *J. Mach. Learn. Res. 15* (2014), 595–627.

[6] BAUSCHKE, H. H., BOLTE, J., AND TEBOULLE, M. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Math. Oper. Res. 42*, 2 (2017), 330–348.

[7] BECK, A., AND TEBOULLE, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett. 31* (2003), 167–175.

[8] BERTSEKAS, D. P. On the Goldstein-Levitin-Polyak gradient projection method. *IEEE Trans. Automat. Contr. AC-21*, 2 (1976), 174–184.

[9] BERTSEKAS, D. P. *Nonlinear Programming*, 3rd ed. Athena Sci., Belmont, MA, 2016.

[10] BLUME-KOHOUT, R. Hedged maximum likelihood quantum state estimation. *Phys. Rev. Lett. 105* (2010).

[11] BYRNE, C., AND CENSOR, Y. Proximity function minimization using multiple Bregman projections, with application to split feasibility and Kullback-Leibler distance minimization. *Ann. Oper. Res. 105* (2001), 77–98.

[12] COLLINS, M., GLOBERSON, A., KOO, T., CARRERAS, X., AND BARTLETT, P. L. Exponentiated gradient algorithms for conditional random fields and max-margin Markov networks. *J. Mach. Learn. Res. 9* (2008), 1775–1822.

[13] DECARREAU, A., HILHORST, D., LEMARÉCHAL, C., AND NAVAZA, J. Dual methods in entropy maximization. application to some problems in crystallography. *SIAM J. Optim. 2*, 2 (1992), 173–197.

[14] DOLJANSKY, M., AND TEBOULLE, M. An interior proximal algorithm and the exponential multiplier method for semidefinite programming. *SIAM J. Optim. 9*, 1 (1998), 1–13.

[15] GAFNI, E. M., AND BERTSEKAS, D. P. Convergence of a gradient projection method. LIDS-P-1201, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, 1982.

[16] HELMBOLD, D. P., SHAPIRE, R. E., SINGER, Y., AND WARMUTH, M. K. On-line portfolio selection using multiplicative updates. *Math. Finance 8*, 4 (1998), 325–347.

[17] HIAI, F., OHYA, M., AND TSUKADA, M. Sufficiency, KMS condition and relative entropy in von Neumann algebras. *Pac. J. Math. 96*, 1 (1981), 99–109.

[18] HOHAGE, T., AND WERNER, F. Inverse problems with Poisson data: statistical regularization theory, applications and algoithms. *Inverse Probl. 32* (2016).

[19] HRADIL, Z. Quantum-state estimation. *Phys. Rev. A 55*, 3 (1997).

[20] KIVINEN, J., AND WARMUTH, M. K. Exponentiated gradient versus gradient descent for linear predictors. *Inf. Comput. 132* (1997), 1–63.

[21] KOLTCHINSKII, V. von Neumann entropy penalization and low-rank matrix estimation. *Ann. Stat. 39*, 6 (2011), 2936–2973.

[22] LU, H., FREUND, R. M., AND NESTEROV, Y. Relatively-smooth convex optimization by first-order methods, and applications. arXiv:1610.05708v1.

[23] MACLEAN, L. C., THORP, E. O., AND ZIEMBA, W. T., Eds. *The Kelly Capital Growth Investment Criterion*. World Sci., Singapore, 2012.

[24] NEMIROVSKY, A. S., AND YUDIN, D. B. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons, Chichester, 1983.

[25] NESTEROV, Y., AND NEMIROVSKII, A. *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM, Philadelphia, PA, 1994.

[26] ODOR, G., LI, Y.-H., YURTSEVER, A., HSIEH, Y.-P., EL HALABI, M., TRAN-DINH, Q., AND CEVHER, V. Frank-Wolfe works for non-Lipschitz continuous gradient objectives: Scalable Poisson phase retrieval. In *IEEE Int. Conf. Acoustics, Speech and Signal Processing* (2016), pp. 6230–6234.

[27] OHYA, M., AND PETZ, D. *Quantum Entropy and Its Use*. Springer, Berlin, 1993.

[28] PARIS, M., AND ŘEHÁČEK, J., Eds. *Quantum State Estimation*. Springer, Berlin, 2004.

[29] ROCKAFELLAR, R. T. *Convex Analysis*. Princeton Univ. Press, Princeton, NJ, 1970.

[30] SALZO, S. The variable metric forward-backward splitting algorithms under mild differentiability assumptions. *SIAM J. Optim. 27*, 4 (2017), 2153–2181.

[31] TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., Ser. B 58*, 1 (1996), 267–288.

[32] TRAN-DINH, Q., LI, Y.-H., AND CEVHER, V. Composite convex minimization involving self-concordant-like cost functions. In *Model. Comput. & Optim. in Inf. Syst. & Manage. Sci.* (Cham, 2015), Springer, pp. 155–168.

[33] TSUDA, K., RÄTSCH, G., AND WARMUTH, M. K. Matrix exponentiated gradient updates for on-line learning and Bregman projection. *J. Mach. Learn. Res. 6* (2005), 995–1018.

[34] VARDI, Y., SHEPP, L. A., AND KAUFMAN, L. A statistical model for positron emission tomography. *J. Am. Stat. Assoc. 80*, 389 (1985), 8–20.