

# A geodesic interior-point method for linear optimization over symmetric cones

Frank Permenter

August 19, 2020

## Abstract

We develop a new interior-point method for symmetric-cone optimization, a common generalization of linear, second-order-cone, and semidefinite programming. Our key idea is updating iterates with a *geodesic* of the *cone* instead of the *kernel* of the *linear* constraints. This approach yields a primal-dual-symmetric, scale-invariant, and line-search-free algorithm that uses just half the variables of a standard primal-dual method. With elementary arguments, we establish polynomial-time convergence matching the standard  $\mathcal{O}(\sqrt{n})$  bound. Finally, we prove global convergence of a long-step variant and compare the approaches computationally. For linear programming, our algorithms reduce to central-path tracking in the log domain.

## 1 Introduction

Let  $\mathcal{J}$  denote a Euclidean Jordan algebra [5] of rank  $n$  with multiplication operator  $\circ : \mathcal{J} \times \mathcal{J} \rightarrow \mathcal{J}$ , identity  $e \in \mathcal{J}$ , and trace inner-product  $\langle x, y \rangle := \text{tr } x \circ y$ . This paper considers the following primal-dual pair of linear optimization problems formulated over the *cone-of-squares*  $\mathcal{K} := \{x \circ x : x \in \mathcal{J}\}$

$$\begin{array}{ll} \text{minimize} & \langle s_0, x \rangle \\ \text{subject to} & x \in \mathcal{K} \cap (x_0 + \mathcal{L}) \end{array} \qquad \begin{array}{ll} \text{minimize} & \langle x_0, s \rangle \\ \text{subject to} & s \in \mathcal{K} \cap (s_0 + \mathcal{L}^\perp), \end{array} \quad (1)$$

where  $(x, s)$  are the primal and dual decision variables,  $(x_0, s_0) \in \mathcal{J} \times \mathcal{J}$  are fixed parameters and  $\mathcal{L} \subseteq \mathcal{J}$  is a linear subspace with orthogonal complement  $\mathcal{L}^\perp \subseteq \mathcal{J}$ . This standard form [7] subsumes linear, second-order-cone, and semidefinite programming [4], in which the affine set  $x_0 + \mathcal{L}$  is often presented as the solution set of linear equations  $Ax = b$ . It is also referred to as a *symmetric-cone* optimization problem given the one-to-one correspondence between such cones and cones-of-squares [5].

A pair  $(x, s)$  is optimal if it satisfies the constraints of (1) and the additional *complementary slackness* condition  $x \circ s = 0$ . Interior-point methods solve a perturbation of these constraints for a decreasing sequence of  $\mu > 0$ :

$$x \in \mathcal{K} \cap (x_0 + \mathcal{L}), \quad s \in \mathcal{K} \cap (s_0 + \mathcal{L}^\perp), \quad x \circ s = \mu e. \quad (2)$$

A unique solution  $(\hat{x}(\mu), \hat{s}(\mu))$  exists for all  $\mu > 0$  if the primal-dual pair (1) satisfies Slater's condition [6, Theorem 2.2], i.e., if there exists feasible  $(x, s)$  in the interior of  $\mathcal{K} \times \mathcal{K}$ . In this case, the limit  $\lim_{\mu \rightarrow 0} (\hat{x}(\mu), \hat{s}(\mu))$  also exists [25, 8] and solves (1). The set  $\{(\hat{x}(\mu), \hat{s}(\mu)) : \mu > 0\}$  is called the *central path*. We will assume Slater's condition holds throughout.

**Assumption 1.** *The primal-dual pair (1) satisfies Slater's condition.*

Primal-dual interior-point methods track the central path by iteratively updating  $(x, s)$  inside *subspaces* of  $\mathcal{J}$  such that the *affine* constraints remain satisfied:

$$x_{i+1} - x_i \in \mathcal{L}, \quad s_{i+1} - s_i \in \mathcal{L}^\perp. \quad (3)$$

In this paper, we take a different approach, leveraging the structure of  $\mathcal{K}$  as a Riemannian manifold [16]. Specifically, we update  $(x, s)$  along *geodesic curves*  $z(t) = Q(z_0^{1/2}) \exp td$  of  $\mathcal{K}$  such that the *complementarity* constraint  $x \circ s = \mu e$  remains satisfied:

$$x_{i+1} = Q(x_i^{1/2}) \exp d_i, \quad s_{i+1} = Q(s_i^{1/2}) \exp(-d_i), \quad (4)$$

where  $\exp : \mathcal{J} \rightarrow \mathcal{K}$  denotes the exponential map,  $Q(w) : \mathcal{J} \rightarrow \mathcal{J}$  denotes the *quadratic representation* of  $w$ , and  $w^{1/2} \in \mathcal{K}$  denotes the square root of  $w \in \mathcal{K}$ . To select  $d_i \in \mathcal{J}$  and the corresponding geodesic, we substitute the first-order Taylor expansions

$$Q(x_i^{1/2}) \exp d_i \approx Q(x_i^{1/2})(e + d_i), \quad Q(s_i^{1/2}) \exp(-d_i) \approx Q(s_i^{1/2})(e - d_i)$$

into the central-path conditions (2) and solve the resulting linear system for  $d_i$ . In total, this process can be interpreted as a *manifold* version of Newton’s method [1, Chapter 6].

While we view  $\mathcal{K}$  as a Riemannian manifold under the metric  $\gamma_x(u, v) := \langle Q(x)^{-1}u, v \rangle$ , we won’t actually need any differential geometry. Instead, all analysis rests on the parametric representation of geodesics  $z(t) = Q(z_0^{1/2}) \exp td$ , which has elementary forms for specific cones. When  $\mathcal{K}$  is the nonnegative orthant, it simplifies to  $z(t) = \exp(\log z_0 + td)$ , i.e., a line segment in log space. When  $\mathcal{K}$  is the cone of positive semidefinite matrices, it takes the form  $Z(t) = Z_0^{1/2} \exp(tD)Z_0^{1/2}$ , where  $\exp(tD)$  denotes the matrix exponential and  $Z_0^{1/2}$  the symmetric square root [3, Chapter 6].

The interior-point literature is immense, and we won’t attempt to cite it completely. Nevertheless, to our knowledge, *all* primal-dual interior-point methods use updates that satisfy (3) given feasible iterates; see, e.g., [26, 23, 29]. While the Riemannian geometry of  $\mathcal{K}$  has been used to analyze the central path [22, 20] and solve non-convex problems [1], to our knowledge no previous algorithm for (1) uses the geodesic update (4). For linear programming, the proposed essentially reduces to Newton’s method when applied to the central-path conditions in the log domain, i.e., to nonlinear equations  $f(v) = 0$  induced by

$$\sqrt{\mu} \exp(v) \in x_0 + \mathcal{L}, \quad \sqrt{\mu} \exp(-v) \in s_0 + \mathcal{L}^\perp.$$

Even this appears unanalyzed in the linear-programming literature—perhaps dismissed upfront because it returns solutions that are only  $\epsilon$ -feasible (due to finite termination). Nevertheless, we’ll see that feasible points can *always* be constructed once the Newton direction satisfies  $\|d_i\|_\infty \leq 1$ .

A desired feature of an interior-point algorithm is primal-dual symmetry [28]. This means that the algorithm does not depend on the labels “primal” and “dual” assigned to the problems of (1), i.e., swapping these labels does not change its output. To obtain symmetry, primal-dual methods (e.g., [21, 2]) separately store and independently update *both* variables  $x$  and  $s$ . We will avoid such drawbacks, exploiting the fact that the geodesic update (4) maintains the complementarity relation  $x = \mu s^{-1}$ . This allows us to *implicitly* perform the symmetric update (4) while *explicitly* storing and updating just *one* variable.

This paper is organized as follows. In Section 2, we give an interior-point method based on the geodesic update (4) and establish its  $\mathcal{O}(\sqrt{n})$  convergence, log-domain interpretation, scale invariance, and relation to the Nesterov-Todd method. We also show its search direction can be found via least-squares methods like many interior-point algorithms [26]. Since this procedure

<b>Procedure</b> <code>shortstep</code> ( $w_0, \mu_0, \mu_f$ ) $w \leftarrow w_0, \mu \leftarrow \mu_0$ <b>while</b> $\mu > \mu_f$ <b>do</b> $\mu \leftarrow \frac{1}{k}\mu$ <b>for</b> $i = 1, 2, \dots, m$ <b>do</b> $d \leftarrow d_N(w, \mu)$ $w \leftarrow Q(w^{1/2}) \exp(d)$ <b>end</b> <b>end</b> <b>return</b> ( $w, \mu$ )	<table border="0" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left; border-bottom: 1px solid black;"><math>\mathcal{K}</math></th> <th style="text-align: left; border-bottom: 1px solid black;">Definition</th> <th style="text-align: left; border-bottom: 1px solid black;">rank</th> </tr> </thead> <tbody> <tr> <td><math>\mathbb{R}_+^n</math></td> <td><math>\{x \in \mathbb{R}^n : x_i \geq 0\}</math></td> <td><math>n</math></td> </tr> <tr> <td><math>\mathbb{S}_+^n</math></td> <td><math>\{X^2 : X \in \mathbb{R}^{n \times n}, X = X^T\}</math></td> <td><math>n</math></td> </tr> <tr> <td><math>\mathbb{L}^{m+1}</math></td> <td><math>\{(x_0, x_1) \in \mathbb{R} \times \mathbb{R}^m : x_0 \geq \ x_1\ \}</math></td> <td>2</td> </tr> </tbody> </table>	$\mathcal{K}$	Definition	rank	$\mathbb{R}_+^n$	$\{x \in \mathbb{R}^n : x_i \geq 0\}$	$n$	$\mathbb{S}_+^n$	$\{X^2 : X \in \mathbb{R}^{n \times n}, X = X^T\}$	$n$	$\mathbb{L}^{m+1}$	$\{(x_0, x_1) \in \mathbb{R} \times \mathbb{R}^m : x_0 \geq \ x_1\ \}$	2
$\mathcal{K}$	Definition	rank											
$\mathbb{R}_+^n$	$\{x \in \mathbb{R}^n : x_i \geq 0\}$	$n$											
$\mathbb{S}_+^n$	$\{X^2 : X \in \mathbb{R}^{n \times n}, X = X^T\}$	$n$											
$\mathbb{L}^{m+1}$	$\{(x_0, x_1) \in \mathbb{R} \times \mathbb{R}^m : x_0 \geq \ x_1\ \}$	2											
	<table border="0" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left; border-bottom: 1px solid black;"><math>\mathcal{K}</math></th> <th style="text-align: left; border-bottom: 1px solid black;"><math>\exp d</math></th> <th style="text-align: left; border-bottom: 1px solid black;"><math>Q(w^{1/2}) \exp d</math></th> </tr> </thead> <tbody> <tr> <td><math>\mathbb{R}_+^n</math></td> <td>element-wise exp.</td> <td><math>\exp(\log w + d)</math></td> </tr> <tr> <td><math>\mathbb{S}_+^n</math></td> <td>matrix exponential</td> <td><math>W^{1/2}(\exp D)W^{1/2}</math></td> </tr> <tr> <td><math>\mathbb{L}^{m+1}</math></td> <td>replace eigenvalues with <math>\exp(d_0 \pm \ d_1\ )</math></td> <td><math>(2zz^T - (\det z)R) \exp d</math></td> </tr> </tbody> </table>	$\mathcal{K}$	$\exp d$	$Q(w^{1/2}) \exp d$	$\mathbb{R}_+^n$	element-wise exp.	$\exp(\log w + d)$	$\mathbb{S}_+^n$	matrix exponential	$W^{1/2}(\exp D)W^{1/2}$	$\mathbb{L}^{m+1}$	replace eigenvalues with $\exp(d_0 \pm \ d_1\ )$	$(2zz^T - (\det z)R) \exp d$
$\mathcal{K}$	$\exp d$	$Q(w^{1/2}) \exp d$											
$\mathbb{R}_+^n$	element-wise exp.	$\exp(\log w + d)$											
$\mathbb{S}_+^n$	matrix exponential	$W^{1/2}(\exp D)W^{1/2}$											
$\mathbb{L}^{m+1}$	replace eigenvalues with $\exp(d_0 \pm \ d_1\ )$	$(2zz^T - (\det z)R) \exp d$											

Figure 1: Short-step algorithm (left) with parameters  $(k, m)$  and implementation details (right) for linear programs ( $\mathbb{R}_+^n$ ), second-order-cone programs ( $\mathbb{L}^{m+1}$ ), and semidefinite programs ( $\mathbb{S}_+^n$ ). In the  $\mathbb{L}^{m+1}$  row, the map  $R$  denotes  $(u_0, u_1) \mapsto (u_0, -u_1)$ , while  $z = w^{1/2}$  and  $\det z = z_0^2 - \|z_1\|^2$ .

conservatively tracks the central path, we refer to it as our *short-step* algorithm [29]. In Section 3, we study connections between geodesic distance and symmetrized Kullback-Leibler divergence, proving key results invoked in our short-step analysis. Leveraging this study, we describe a less conservative *long-step* algorithm in Section 4 and prove its global convergence and scale invariance; we also discuss construction of feasible points and other implementation issues. Finally, Section 5 contains computational results illustrating superior performance of the long-step algorithm, despite its weaker theoretical guarantees. Background material on Jordan algebras appears in Appendix A.

## 2 Short-step algorithm

We give a procedure `shortstep` (Figure 1) for tracking the central path that employs geodesic updates. To maintain the identity  $s = \mu x^{-1}$  and primal-dual symmetry, it uses a single variable  $w \in \text{int } \mathcal{K}$  satisfying

$$x = \sqrt{\mu}w, \quad s = \sqrt{\mu}w^{-1},$$

where  $\text{int } \mathcal{K}$  denotes the interior of  $\mathcal{K}$ . The inputs are an initial  $w_0 \in \text{int } \mathcal{K}$  and centering parameters  $\mu_0 > \mu_f > 0$ . The output is an approximation of the centered point  $\hat{w}(\mu)$  for  $\mu \leq \mu_f$ , where  $\hat{w}(\mu)$  denotes  $\frac{1}{\sqrt{\mu}}\hat{x}(\mu)$  for  $(\hat{x}(\mu), \hat{s}(\mu))$  on the central path. Behavior depends on a parameter  $k$  that controls how much  $\mu$  decreases at each *outer iteration* and a parameter  $m$  that denotes the number of *inner iterations*, i.e., Newton steps. Like short-step interior-point methods [29], we will conservatively decrease  $\mu$  in our analysis. We will also assume that  $w_0 = \hat{w}(\mu_0)$ .

Iterations apply the update  $w \leftarrow Q(w^{1/2}) \exp(d)$ , or, equivalently,  $w^{-1} \leftarrow Q(w^{-1/2}) \exp(-d)$ , where  $d$  is the *Newton direction* for the current  $(w, \mu)$ . Equivalence of these updates holds given that the quadratic representation  $Q(u)$  satisfies  $[Q(u)v]^{-1} = Q(u^{-1})v^{-1}$  (Appendix A). The Newton direction, denoted  $d_N(w, \mu)$ , is defined by linearizing  $\exp d$  and  $\exp -d$  in these expressions and substituting into the central-path conditions (2). Proposition 2.2 will later prove its uniqueness.

**Definition 2.1.** (*Newton Direction*) For  $w \in \text{int } \mathcal{K}$  and  $\mu > 0$ , the Newton direction  $d_N(w, \mu)$  is the unique  $d \in \mathcal{J}$  satisfying

$$Q(w^{1/2})(e + d) = \frac{1}{\sqrt{\mu}}x_0 + \mathcal{L}, \quad Q(w^{-1/2})(e - d) = \frac{1}{\sqrt{\mu}}s_0 + \mathcal{L}^\perp.$$

Our convergence criterion employs *geodesic distance*  $\delta(z_0, z_1)$ , i.e., the minimum of

$$\int_0^1 \|Q(\gamma^{-1/2}(t))\gamma'(t)\| dt$$

over curves  $\gamma : [0, 1] \rightarrow \text{int } \mathcal{K}$  satisfying  $\gamma(0) = z_0$  and  $\gamma(1) = z_1$ . Among other useful properties, this distance has a closed-form solution.

**Lemma 2.1** (e.g., [14]). *Geodesic distance  $\delta : \text{int}(\mathcal{K} \times \mathcal{K}) \rightarrow \mathbb{R}_+$  has the following properties.*

- (a)  $\delta(z_0, z_1) = \|\log Q(z_0^{-1/2})z_1\|$  for all  $z_0, z_1 \in \text{int } \mathcal{K}$ .
- (b)  $\delta$  is a metric.
- (c)  $\delta(z_0, z_1) = \delta(Tz_0, Tz_1)$  for all  $z_0, z_1 \in \text{int } \mathcal{K}$  and any automorphism  $T$  of  $\mathcal{K}$ , i.e., for any invertible, linear map  $T : \mathcal{J} \rightarrow \mathcal{J}$  satisfying  $\{Tz : z \in \mathcal{K}\} = \mathcal{K}$ .

Our criterion also employs an upper bound  $q : \mathbb{R} \rightarrow \mathbb{R}_+$  of the squaring map  $t \mapsto t^2$  and its nonnegative inverse  $q^{-1} : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ :

$$q(t) := 2(\cosh t - 1), \quad q^{-1}(t) := \cosh^{-1}\left(1 + \frac{1}{2}t\right).$$

Note that  $q(t)$  and  $t^2$  agree to second-order given that  $\cosh t = 1 + \sum_{d=1}^{\infty} \frac{1}{(2d)!} t^{2d}$ .

These definitions and the next two lemmas (to be proven in Section 3) allow us to state and prove our main theorem on the complexity and convergence of **shortstep**. The first lemma bounds the distance between two centered points  $\hat{w}(\mu_0)$  and  $\hat{w}(\mu_1)$  using the rank  $n$  of  $\mathcal{K}$  and the ratio  $k$  of the centering parameters.

**Lemma 2.2** ( $\mu$ -update). *Let  $\mu, k > 0$ . Then,  $\frac{1}{n}\delta\left(\hat{w}(\mu), \hat{w}\left(\frac{1}{k}\mu\right)\right)^2 \leq q\left(\frac{1}{2}\log k\right)$ .*

The second lemma establishes a region of quadratic convergence of the sequence  $w_0, w_1, \dots, w_m$  generated by Newton's method.

**Lemma 2.3** (Centering). *For  $\mu > 0$  and  $w_0 \in \text{int } \mathcal{K}$ , let  $w_{i+1} = Q(w_i^{1/2})d_N(w_i, \mu)$ . If  $\delta(w_0, \hat{w}(\mu)) \leq q^{-1}(\beta)$  for  $\beta \leq \frac{1}{2}$ , then  $\delta(w_i, \hat{w}(\mu))^2 \leq \beta^{2^i}$ .*

Our main result follows.

**Theorem 2.1** (Main Result). *Let **shortstep** (Figure 1) have parameters  $(k, m)$  that satisfy, for some  $\frac{1}{2} \geq \beta > 0$  and  $q^{-1}(\beta) > \epsilon > 0$ , the conditions*

$$\beta^{2^m} \leq \epsilon^2, \quad \frac{1}{2}\log k = q^{-1}\left(\frac{1}{n}\zeta^2\right), \tag{5}$$

where  $\zeta := q^{-1}(\beta) - \epsilon$ . Then, the following statements hold for **shortstep** given input  $(\hat{w}(\mu_0), \mu_0, \mu_f)$ :

- (a) At most  $m \lceil c^{-1}\sqrt{n} \log \frac{\mu_0}{\mu_f} \rceil$  Newton steps execute, where  $c := 2q^{-1}(\zeta^2)$ .
- (b) The output  $(w, \mu)$  satisfies  $\delta(w, \hat{w}(\mu)) \leq \epsilon$  and  $\mu \leq \mu_f$ . Further,

$$\delta(\sqrt{\mu}w, \hat{x}(\mu)) \leq \epsilon, \quad \delta(\sqrt{\mu}w^{-1}, \hat{s}(\mu)) \leq \epsilon,$$

where  $(\hat{x}(\mu), \hat{s}(\mu))$  denotes the solution to the central-path conditions (2).

*Proof.* Let  $\gamma = \log \frac{\mu_0}{\mu_f} \frac{1}{\log k}$ . The number of outer iterations is at most  $\lceil \gamma \rceil$ . To lower bound  $\log k$ , we first note that for  $0 \leq t \leq \alpha$ ,

$$q^{-1}(t) \geq \frac{q^{-1}(\alpha)}{\sqrt{\alpha}} \sqrt{t},$$

since  $\frac{q^{-1}(t)}{\sqrt{t}}$  is a decreasing function. Setting  $\alpha = \zeta^2$  and  $t = \frac{1}{n}\zeta^2$  and using (5) gives

$$\frac{1}{2} \log k = q^{-1}\left(\frac{1}{n}\zeta^2\right) \geq \frac{q^{-1}(\zeta^2)}{\zeta} \frac{\zeta}{\sqrt{n}} = \frac{q^{-1}(\zeta^2)}{\sqrt{n}}.$$

Hence,  $(\log k)^{-1} \leq c^{-1}\sqrt{n}$  for  $c = 2q^{-1}(\zeta^2)$ , proving the first statement.

Let  $w_{\mu,i}$  denote  $w$  at the end of inner iteration  $i$  for the current  $\mu$ . We use induction on  $\mu$ . Suppose that  $\delta(w_{\mu,m}, \hat{w}(\mu)) \leq \epsilon$ . By Lemma 2.2 and our choice of  $k$ ,

$$\delta(\hat{w}(\mu), \hat{w}\left(\frac{1}{k}\mu\right))^2 \leq nq\left(\frac{1}{2}\log k\right) = n\frac{1}{n}\zeta^2 = (q^{-1}(\beta) - \epsilon)^2.$$

From the triangle inequality (Lemma 2.1(b)), we conclude that

$$\delta(w_{\mu,m}, \hat{w}\left(\frac{1}{k}\mu\right)) \leq \delta(w_{\mu,m}, \hat{w}(\mu)) + \delta(\hat{w}(\mu), \hat{w}\left(\frac{1}{k}\mu\right)) \leq \epsilon + q^{-1}(\beta) - \epsilon \leq q^{-1}(\beta).$$

Initializing inner iterations with  $w_{\mu,m}$ , we have, by Lemma 2.3 and  $\beta \leq \frac{1}{2}$ , that

$$\delta(w_{k\mu,m}, \hat{w}\left(\frac{1}{k}\mu\right))^2 \leq \beta^{2m} \leq \epsilon^2,$$

where the last inequality follows from our assumption (5). The base case holds by identical argument. Finally,  $\delta(\sqrt{\mu}w, \sqrt{\mu}\hat{w}(\mu)) \leq \epsilon$  and  $\delta(\sqrt{\mu}w^{-1}, \sqrt{\mu}\hat{w}^{-1}(\mu)) \leq \epsilon$  given that  $\delta(cz, c\hat{z}) = \delta(z, \hat{z})$  for any  $c > 0$  by Lemma 2.1(c).  $\square$

The remainder of this section gives other properties of `shortstep`, namely, a log-space interpretation, an orthogonal decomposition of the Newton direction, and scale invariance. We also discuss connections with an algorithm of Nesterov and Todd.

## 2.1 Log-space interpretation

Suppose that (1) is a primal-dual pair of *linear programs*, i.e., that  $\mathcal{K} = \mathbb{R}_+^n$ . Under this assumption, the algebra  $\mathcal{J}$  is associative. Hence, geodesic distance simplifies to  $\delta(z_0, z_1) = \|\log z_0 - \log z_1\|$ , and the log of the geodesic update  $w \leftarrow Q(w^{1/2}) \exp d$  satisfies

$$\log\left(Q(w^{1/2}) \exp d\right) = \log(w \circ \exp d) = \log(w) + d, \tag{6}$$

i.e., it reduces to addition in the log domain. The Newton direction  $d_N(w, \mu)$  also has a log-domain interpretation: it is precisely the direction one obtains by linearizing  $x(v) := \sqrt{\mu} \exp v$  and  $s(v) := \sqrt{\mu} \exp -v$  at  $v = \log w$  and substituting into the central-path conditions (2).

**Proposition 2.1.** *Let  $\mathcal{J}$  be associative. For  $\mu > 0$  and  $w \in \text{int } \mathcal{K}$ , let  $d = d_N(w, \mu)$ . Then,*

$$\exp v + J(v)d \in \frac{1}{\sqrt{\mu}}x_0 + \mathcal{L}, \quad \exp -v - J(-v)d \in \frac{1}{\sqrt{\mu}}s_0 + \mathcal{L}^\perp,$$

where  $v = \log w$  and  $J(v) : \mathcal{J} \rightarrow \mathcal{J}$  is the Jacobian of  $\exp v$ .

*Proof.* Under our associativity assumption, we observe that  $J(v)d = (\exp v) \circ d$  and

$$Q(w^{1/2})(e + d) = w + w \circ d, \quad Q(w^{-1/2})(e - d) = w^{-1} - w^{-1} \circ d.$$

Substituting  $w = \exp v$  and  $w^{-1} = \exp -v$  and using Definition 2.1 proves the claim.  $\square$

In total, we can reinterpret the inner iterations of `shortstep` as simply Newton's method applied to the central-path equations in log space. Though this is a totally elementary algorithm, we could not find its analysis in the literature.

Observe that when  $\mathcal{J}$  is *not* associative, this interpretation fails because the identity (6) fails. For semidefinite programming, failure of (6) reduces to the fact that for matrices  $W \succ 0$  and  $D$ ,

$$\log \left( W^{1/2}(\exp D)W^{1/2} \right) \neq \log(W) + D,$$

since, in general,  $\exp(A + B) \neq \exp A \exp B$  for the matrix exponential.

## 2.2 Newton direction via orthogonal projection

We next derive an orthogonal, direct-sum decomposition of the Newton direction with respect to the subspaces  $\mathcal{L}_w := \{Q(w^{-1/2})x : x \in \mathcal{L}\}$  and  $\mathcal{L}_w^\perp = \{Q(w^{1/2})s : s \in \mathcal{L}^\perp\}$ . This decomposition establishes both its claimed uniqueness (Definition 2.1) and a formula for its construction via orthogonal projection.

**Proposition 2.2.** *For  $\mu > 0$  and  $w \in \text{int } \mathcal{K}$ , let*

$$d_1 = \text{proj}_{\mathcal{L}_w^\perp} \left( Q(w^{-1/2}) \left( \frac{1}{\sqrt{\mu}} x_0 - w \right) \right), \quad d_2 = \text{proj}_{\mathcal{L}_w} \left( Q(w^{1/2}) \left( \frac{1}{\sqrt{\mu}} s_0 - w^{-1} \right) \right).$$

*Then the Newton direction  $d_N(w, \mu)$  satisfies  $d_N(w, \mu) = d_1 - d_2$ .*

*Proof.* Let  $r_p = Q(w^{-1/2}) \left( \frac{1}{\sqrt{\mu}} x_0 - w \right)$  and  $r_d = Q(w^{1/2}) \left( \frac{1}{\sqrt{\mu}} s_0 - w^{-1} \right)$ . By the identity  $Q(z^{1/2})e = z$  (Lemma A.1 of Appendix A), the conditions of Definition 2.1 are equivalent to

$$w + Q(w^{1/2})d \in \frac{1}{\sqrt{\mu}} x_0 + \mathcal{L}, \quad w^{-1} - Q(w^{-1/2})d \in \frac{1}{\sqrt{\mu}} s_0 + \mathcal{L}^\perp.$$

Rearranging and using  $Q(z^{-1}) = Q(z)^{-1}$ , we conclude that  $d \in r_p + \mathcal{L}_w$  and  $d \in -r_d + \mathcal{L}_w^\perp$ , i.e.,

$$d \in \left( \text{proj}_{\mathcal{L}_w^\perp}(r_p) + \mathcal{L}_w \right) \cap \left( \text{proj}_{\mathcal{L}_w}(-r_d) + \mathcal{L}_w^\perp \right),$$

since any affine set  $z_0 + \mathcal{S}$  satisfies  $z_0 + \mathcal{S} = \text{proj}_{\mathcal{S}^\perp}(z_0) + \mathcal{S}$ . Hence,  $d$  has the following direct-sum decompositions with respect to  $\mathcal{L}_w$  and  $\mathcal{L}_w^\perp$ :

$$d = \text{proj}_{\mathcal{L}_w^\perp}(r_p) + d_{\mathcal{L}_w}, \quad d = \text{proj}_{\mathcal{L}_w}(-r_d) + d_{\mathcal{L}_w^\perp}.$$

Since such decompositions are unique,  $d_{\mathcal{L}_w} = \text{proj}_{\mathcal{L}_w}(-r_d)$ , proving the claim.  $\square$

This decomposition has immediate practical implications: one can use any algorithm for orthogonal projection, e.g., the Gram-Schmidt process or a least-squares method, to find  $d_N$ . Section 4.3 gives more details.

### 2.3 Scale invariance

For an automorphism  $T : \mathcal{J} \rightarrow \mathcal{J}$  of  $\mathcal{K}$ , consider the transformed primal-dual pair:

$$\begin{array}{ll} \text{minimize} & \langle (T^{-1})^* s_0, x \rangle \\ \text{subject to} & x \in \mathcal{K} \cap T(x_0 + \mathcal{L}) \end{array} \quad \begin{array}{ll} \text{minimize} & \langle T x_0, s \rangle \\ \text{subject to} & s \in \mathcal{K} \cap (T^{-1})^*(s_0 + \mathcal{L}^\perp), \end{array} \quad (7)$$

where  $(T^{-1})^* : \mathcal{J} \rightarrow \mathcal{J}$  denotes the adjoint of  $T^{-1} : \mathcal{J} \rightarrow \mathcal{J}$ . We next show the following: if **shortstep** (Figure 1) maps input  $w_0$  to output  $\bar{w}$  for the primal-dual pair (1), then it maps input  $T w_0$  to output  $T \bar{w}$  for the transformed pair (7). In other words, it is *scale invariant* in the sense of [28]. To show this, we first establish that the Newton direction  $d_{N,T}(w, \mu)$  for the transformed problem satisfies  $d_{N,T}(T w, \mu) = M d_N(w, \mu)$  for an automorphism  $M$ , dependent on  $T$  and  $w$ , that is also *orthogonal*, i.e.,  $M^{-1} = M^*$ . Scale invariance will follow, leveraging the fact that  $\exp M d = M \exp d$  for any such  $M$  (Lemma A.2).

To give a formula for  $M$  and to establish its key properties, we use the decomposition  $d_N(w, \mu) = d_1(w, \mu) - d_2(w, \mu)$  of the Newton direction from Proposition 2.2. We similarly decompose the transformed direction as  $d_{N,T}(w, \mu) = d_{1,T}(w, \mu) - d_{2,T}(w, \mu)$ .

**Lemma 2.4.** *Let  $M = Q(Tw)^{-1/2} T Q(w)^{1/2}$  for  $w \in \text{int } \mathcal{K}$  and an automorphism  $T : \mathcal{J} \rightarrow \mathcal{J}$  of  $\mathcal{K}$ . The following statements hold.*

(a)  *$M$  is an orthogonal automorphism of  $\mathcal{K}$ .*

(b)  *$M = Q(Tw)^{1/2} (T^{-1})^* Q(w)^{-1/2}$ .*

(c) *For all  $\mu > 0$ , the Newton directions  $(d_N, d_{N,T})$  and their direct summands  $(d_i, d_{i,T})$  satisfy*

$$d_{N,T}(T w, \mu) = M d_N(w, \mu), \quad d_{1,T}(T w, \mu) = M d_1(w, \mu), \quad d_{2,T}(T w, \mu) = M d_2(w, \mu).$$

*Proof.* That  $M$  is an automorphism follows because it is a composition of automorphisms. We next verify orthogonality, i.e., that  $M^{-1} = M^*$ :

$$M^* M = Q(w)^{1/2} T^* Q(Tw)^{-1} T Q(w)^{1/2} = Q(w)^{1/2} T^* (T Q(w) T^*)^{-1} T Q(w)^{1/2} = I,$$

where we've used the identities  $Q(Tw) = T Q(w) T^*$  and  $Q(w)^{1/2} Q(w)^{-1} Q(w)^{1/2} = I$  (Lemma A.1). Since by construction  $M^* Q(Tw)^{1/2} (T^{-1})^* Q(w)^{-1/2} = I$ , orthogonality implies the next statement.

By definition of  $M$  and the second property, we conclude that  $M Q(w)^{-1/2} = Q(Tw)^{-1/2} T$  and  $M Q(w)^{1/2} = Q(Tw)^{1/2} (T^{-1})^*$ . Combining this with  $M e = e$  (Lemma A.2) yields

$$d_{N,T}(T w, \mu) \in M \left( \frac{1}{\sqrt{\mu}} Q(w^{-1/2}) x_0 - e + Q(w^{-1/2}) \mathcal{L} \right) \cap M \left( e - Q(w^{1/2}) \frac{1}{\sqrt{\mu}} s_0 + Q(w^{1/2}) \mathcal{L}^\perp \right),$$

showing that  $d_{N,T}(T w, \mu) = M(d_1 - d_2)$ . By uniqueness of direct-sum decompositions, we also conclude that  $d_{1,T} = M d_1$  and  $d_{2,T} = M d_2$ .  $\square$

We use this lemma to show scale invariance of the update  $w \leftarrow Q(w^{1/2}) \exp(\alpha(d_1, d_2)d)$ , where  $d = d_N$  and  $\alpha : \mathcal{J} \times \mathcal{J} \rightarrow \mathbb{R}$  is a step-size rule invariant under transformation by  $M$ .

**Proposition 2.3.** *Let  $\alpha : \mathcal{J} \times \mathcal{J} \rightarrow \mathbb{R}$  be a function satisfying  $\alpha(d_1, d_2) = \alpha(M d_1, M d_2)$  for any orthogonal automorphism  $M : \mathcal{J} \rightarrow \mathcal{J}$ . Then, for any automorphism  $T : \mathcal{J} \rightarrow \mathcal{J}$  and  $w \in \text{int } \mathcal{K}$ ,*

$$Q(\hat{w}^{1/2}) \exp(\alpha(\hat{d}_1, \hat{d}_2)\hat{d}) = T Q(w^{1/2}) \exp(\alpha(d_1, d_2)d),$$

where  $\hat{w} = T w$ ,  $d = d_N(w, \mu)$ ,  $\hat{d} = d_{N,T}(\hat{w}, \mu)$ ,  $d_i = d_i(w, \mu)$ , and  $\hat{d}_i = d_{i,T}(\hat{w}, \mu)$  for  $i \in \{1, 2\}$ .

*Proof.* Let  $M = Q(Tw)^{-1/2}TQ(w)^{1/2}$ . By Lemma 2.4,  $M$  is an orthogonal automorphism. Hence,  $\exp Mx = M \exp x$  for all  $x$  (Lemma A.2). Combining this with Lemma 2.4(c) yields

$$Q(\hat{w}^{1/2}) \exp \alpha(\hat{d}_1, \hat{d}_2)\hat{d} = Q(\hat{w}^{1/2}) \exp \alpha(Md_1, Md_2)Md = Q(\hat{w}^{1/2})M \exp \alpha(Md_1, Md_2)d.$$

But  $\alpha(Md_1, Md_2) = \alpha(d_1, d_2)$  by assumption and  $Q(\hat{w}^{1/2})M = TQ(w^{1/2})$  by definition of  $M$  and the identity  $Q(w)^{1/2} = Q(w^{1/2})$  (Lemma A.1).  $\square$

Scale invariance of `shortstep` follows by invoking this result at each iteration with the step-size  $\alpha = 1$ . We will use a nontrivial step-size rule in our long-step algorithm (Section 4).

## 2.4 Comparison with the Nesterov-Todd algorithm

The celebrated algorithm of Nesterov and Todd (NT) [21, Section 6], which extends the linear programming algorithms of Kojima et al. [11] and Monteiro and Adler [19], shares key properties with `shortstep` (Figure 1): it is scale invariant, it executes  $\mathcal{O}(\sqrt{n})$  iterations, it is primal-dual symmetric, and finding its search direction reduces to a least-squares problem. This suggests a fundamental connection with `shortstep`. In general, iterations of the NT algorithm do *not* satisfy  $x = \mu s^{-1}$ . However, if this relation holds, then the NT search direction coincides with our Newton direction. Further, its  $(x, s)$ -update is a first-order approximation of our geodesic update.

To see this, note that the NT direction is, in the framework of Jordan algebras [30, Section 3.4], the unique  $(d_x, d_s) \in \mathcal{J} \times \mathcal{J}$  satisfying

$$x + \sqrt{\mu}Q(p^{1/2})d_x \in x_0 + \mathcal{L}, \quad s + \sqrt{\mu}Q(p^{-1/2})d_s \in s_0 + \mathcal{L}, \quad d_x + d_s = v^{-1} - v, \quad (8)$$

where  $p$  is the *scaling point*, defined as  $Q(x^{1/2})(Q(x^{1/2})s)^{-1/2}$ , and  $v := \frac{1}{\sqrt{\mu}}Q(p^{-1/2})x$ . Given  $(d_x, d_s)$ , the NT algorithm updates  $(x, s)$  to  $(x', s')$ , where

$$x' := x + \sqrt{\mu}Q(p^{1/2})d_x, \quad s' := s + \sqrt{\mu}Q(p^{-1/2})d_s. \quad (9)$$

These objects have the following relationships with our variable  $w := \frac{1}{\sqrt{\mu}}x$ , Newton direction  $d_N$ , and geodesic updates  $x \leftarrow \sqrt{\mu}Q(w^{1/2}) \exp d$  and  $s \leftarrow \sqrt{\mu}Q(w^{-1/2}) \exp -d$ .

**Proposition 2.4.** *Let  $x, s \in \text{int } \mathcal{K}$  satisfy  $x = \mu s^{-1}$  for  $\mu > 0$ . Let  $w = \frac{1}{\sqrt{\mu}}x$  and  $d = d_N(w, \mu)$ . Then,*

- (a)  $p = w$ , where  $p$  is the scaling point  $Q(x^{1/2})(Q(x^{1/2})s)^{-1/2}$ .
- (b)  $d_x = d$  and  $d_s = -d$  where  $(d_x, d_s)$  is the NT direction (8).
- (c)  $x' = \sqrt{\mu}Q(w^{1/2})(e + d)$  and  $s' = \sqrt{\mu}Q(w^{-1/2})(e - d)$ , where  $(x', s')$  is the NT update (9) and  $e + d$  and  $e - d$  are the first-order Taylor-expansions of  $\exp(d)$  and  $\exp(-d)$  at  $d = 0$ .

*Proof.* If  $x = \mu s^{-1}$ , then the definitions of  $w$  and the scaling point  $p$  easily imply that  $p = w$  and  $v = e$ . We also conclude that  $d_x + d_s = v^{-1} - v = 0$ . Combining these identities with  $w = Q(w^{1/2})e$ ,  $w^{-1} = Q(w^{-1/2})e$ , and (8) yields

$$\sqrt{\mu}Q(w^{1/2})(e + d_x) \in x_0 + \mathcal{L}, \quad \sqrt{\mu}Q(w^{-1/2})(e - d_x) \in s_0 + \mathcal{L},$$

which are the defining conditions of  $d_N(w, \mu)$  given by Definition 2.1. Hence,  $d = d_x$ . Finally, the claimed formula for  $(x', s')$  holds because  $x = \sqrt{\mu}Q(w^{1/2})e$  and  $s = \sqrt{\mu}Q(w^{-1/2})e$ .  $\square$



Note with the stronger assumption that  $x = s^{-1}$ , we can similarly interpret algorithms based on the so-called H.K.M direction since, in this case, it coincides with the NT direction [27]. It was introduced independently by Helmberg et al. [10], Kojima et al. [12] and Monteiro [18]. Also note that even if  $x = \mu s^{-1}$  fails, the scaling point  $p$  still has a Riemannian interpretation: it is precisely the midpoint of the geodesic connecting  $x$  and  $s$ , or, equivalently, their geometric mean [15].

### 3 Geodesics and divergence

The goal of this section is to prove the  $\mu$ -update and centering lemmas used in the analysis of `shortstep` (Figure 1). Towards this, we first study a proxy for geodesic distance  $\delta(z_0, z_1)$  that is easier to bound during the course of Newton's method. This proxy generalizes the *symmetric Kullback-Leibler divergence*  $h(Z_0, Z_1) := \text{Tr}(Z_0 Z_1^{-1} + Z_0^{-1} Z_1 - 2I)$  of two positive definite matrices, also known as the *Jeffrey divergence* [9, 17]. We hence call this proxy *divergence*. We define it using the fact that  $n = \text{tr } e$ , where we recall that  $n$  denotes the rank of  $\mathcal{K}$ .

**Definition 3.1** (Divergence). *For  $z_0, z_1 \in \text{int } \mathcal{K}$ , let  $h(z_0, z_1) := \langle z_0, z_1^{-1} \rangle + \langle z_0^{-1}, z_1 \rangle - 2n$ .*

Divergence is symmetric  $h(z_0, z_1) = h(z_1, z_0)$ . Further,  $h(z_0, z_1) = 0$  if and only if  $z_0 = z_1$ . However, unlike geodesic distance  $\delta$ , it is *not* a metric, as the triangle inequality can fail.

Recall from Lemma 2.1 that geodesic distance satisfies  $\delta(z_0, z_1) = \|\log Q(z_1^{-1/2})z_0\|$ . Equivalently,  $\delta(z_0, z_1)^2 = \sum_{\lambda \in S} \lambda^2$ , where  $S$  denotes the multiset of eigenvalues of  $\log Q(z_1^{-1/2})z_0$ . This formula holds for divergence if we replace  $\lambda^2$  with the upper bound  $q(\lambda) := 2(\cosh \lambda - 1)$  introduced in Section 2.

**Lemma 3.1.** *For all  $z_0, z_1 \in \text{int } \mathcal{K}$ , the divergence satisfies  $h(z_0, z_1) = \sum_{\lambda \in S} q(\lambda)$ , where  $S$  is the multiset of eigenvalues of  $\log Q(z_1^{-1/2})z_0$ .*

This enables us to prove the following bounds relating divergence to geodesic distance.

**Lemma 3.2.** *Let  $z_0, z_1 \in \text{int } \mathcal{K}$ . Then,  $\delta(z_0, z_1)^2 \leq h(z_0, z_1) \leq q(\delta(z_0, z_1))$ .*

*Proof.* Let  $\lambda \in \mathbb{R}^n$  denote the vector of eigenvalues of  $\log Q(z_1^{-1/2})z_0$ . The lower bound follows from Lemma 3.1 and Lemma 2.1(a) given that  $q(\lambda_i) \geq \lambda_i^2$ . To prove the upper bound, it suffices to show that  $\sum_{i=1}^n (\cosh(\lambda_i) - 1) \leq \cosh(\|\lambda\|) - 1$ . To begin, consider the upper bound

$$\sum_{i=1}^n (\cosh(\lambda_i) - 1) \leq \sup_{\|v\|=\|\lambda\|} \sum_{i=1}^n (\cosh(v_i) - 1).$$

Let  $v$  achieve the supremum. Then it must be a critical point, which implies existence of  $\gamma \in \mathbb{R}$  satisfying  $\gamma v + \sinh v = 0$ . We conclude that  $v_i = 0$  or  $v_i = c$  for a constant  $c$ . We now claim that  $v_i \neq 0$  and  $v_j \neq 0$  implies  $i = j$ . Suppose otherwise. Then we don't change  $\|v\|$  by setting  $v_i = 0$  and  $v_j = \frac{1}{\sqrt{2}}2c$ . Further, we increase  $\sum_{i=1}^n \cosh(v_i) - 1$  given that  $\cosh(\frac{1}{\sqrt{2}}2c) - 1 \geq 2(\cosh(c) - 1)$ , contradicting our assumption that  $v$  attains the supremum.  $\square$

#### 3.1 Divergence along the central path

Divergence has the following utility: we can calculate it *exactly* for two centered points  $\hat{w}(\mu_0)$  and  $\hat{w}(\mu_1)$  even if we do not know these points explicitly. Instead, all we need is the ratio of the centering parameters  $\mu_0$  and  $\mu_1$  and the rank  $n$  of  $\mathcal{K}$ .

**Theorem 3.1.** *Let  $\mu_0, \mu_1 > 0$ . Then,  $\frac{1}{n}h(\hat{w}(\mu_0), \hat{w}(\mu_1)) = q(\frac{1}{2} \log \frac{\mu_0}{\mu_1})$ .*

*Proof.* Let  $x = \hat{w}(\mu_1)$ ,  $y = \hat{w}(\mu_0)$  and  $k = \sqrt{\frac{\mu_0}{\mu_1}}$ . Then, by definition,

$$x - ky \in \mathcal{L}, \quad x^{-1} - ky^{-1} \in \mathcal{L}^\perp.$$

Hence,  $0 = \langle x - ky, x^{-1} - ky^{-1} \rangle = (1 + k^2)n - k\langle x, y^{-1} \rangle - k\langle y, x^{-1} \rangle$ . Rearranging, we conclude that

$$\langle y, x^{-1} \rangle + \langle x, y^{-1} \rangle = n \frac{1 + k^2}{k} = n(k + \frac{1}{k}) = 2n(\cosh \log(k)).$$

Hence,  $h(x, y) = 2n(\cosh \log(k) - 1) = nq(\log k)$ . Since  $\log k = \frac{1}{2} \log \frac{\mu_0}{\mu_1}$ , the claim follows.  $\square$

Combining this theorem with the bounds relating divergence and geodesic distance (Lemma 3.2) lets us prove the  $\mu$ -update lemma, which we reproduce below.

**Lemma 2.2** ( $\mu$ -update). *Let  $\mu, k > 0$ . Then,  $\frac{1}{n}\delta\left(\hat{w}(\mu), \hat{w}(\frac{1}{k}\mu)\right)^2 \leq q(\frac{1}{2} \log k)$ .*

*Proof.* From Theorem 3.1, we conclude that  $\frac{1}{n}h(\hat{w}(\mu), \hat{w}(\frac{1}{k}\mu)) = q(\frac{1}{2} \log k)$ . Since  $\delta(\hat{w}(\mu), \hat{w}(\frac{1}{k}\mu))^2 \leq h(\hat{w}(\mu), \hat{w}(\frac{1}{k}\mu))$  by Lemma 3.2, the claim follows.  $\square$

**Remark 1.** *Since geodesic distance is invariant under inversion and positive rescaling [5, Theorem III.5.3], the lengths  $L_x$  and  $L_s$  of the primal or dual central path also upper bound  $\delta(\hat{w}(\mu_0), \hat{w}(\mu_1))$ , where*

$$L_x = \int_{\mu_0}^{\mu_1} \left\| \frac{d}{d\mu} \hat{x}(\mu) \right\|_{\hat{x}(\mu)} d\mu, \quad L_s = \int_{\mu_0}^{\mu_1} \left\| \frac{d}{d\mu} \hat{s}(\mu) \right\|_{\hat{s}(\mu)} d\mu$$

and  $\|z\|_u^2 := \langle Q(u)^{-1}z, z \rangle$ . Bounds on  $L_x$  in terms of  $\log \frac{\mu_0}{\mu_1}$  and the (generally unknown) values of the barrier function  $\log \det u$  at  $\hat{x}(\mu_0)$  and  $\hat{x}(\mu_1)$  appear in [20, Lemma 4.1].

### 3.2 Divergence along geodesics

Fix  $\mu > 0$ ,  $w \in \text{int } \mathcal{K}$ , and nonzero  $d \in \mathcal{J}$ , and define the function  $f : \mathbb{R} \rightarrow \mathbb{R}$

$$f(t) = h\left(Q(w^{1/2}) \exp td, \hat{w}(\mu)\right).$$

That is, let  $f(t)$  return the divergence between the centered point  $\hat{w}(\mu)$  and points on the geodesic induced by  $(w, d)$ . Though we don't know  $\hat{w}(\mu)$  and hence cannot evaluate  $f$ , we can still establish crucial properties, such as its strict convexity.

**Lemma 3.3.** *The function  $f$  is strictly convex.*

*Proof.* Let  $a := Q(w^{1/2})\hat{w}(\mu)^{-1}$  and let  $\sum_{i=1}^n \lambda_i e_i$  denote the spectral decomposition of  $d$ . Then,

$$f(t) + 2n = \langle a, \exp td \rangle + \langle a^{-1}, \exp -td \rangle = \sum_{i=1}^n \exp(t\lambda_i) \langle a, e_i \rangle + \exp(-t\lambda_i) \langle a^{-1}, e_i \rangle.$$

But  $\langle a, e_i \rangle > 0$  and  $\langle a^{-1}, e_i \rangle > 0$  since  $a, a^{-1} \in \text{int } \mathcal{K}$  and  $e_i \in \mathcal{K}$ , proving the claim.  $\square$

We can also bound  $f(t)$  for full Newton steps, i.e., for  $d = d_N(w, \mu)$  and  $t = 1$ .

**Theorem 3.2.** *Suppose that  $d = d_N(w, \mu)$  and  $\|d\|_\infty^2 \leq 2$ . Then,  $f(1) \leq \frac{1}{2}\|d\|_\infty^2 f(0)$ .*

To prove this theorem, we'll first provide the derivatives of  $f$  and a descent condition on  $t$  for arbitrary  $d$ . We then specialize results to the Newton direction  $d_N(w, \mu)$ .

### 3.2.1 Derivatives and descent condition

The derivatives of  $f(t)$  have a concise form thanks to the role of the exponential function in its definition. Interpreting  $f(t)$  as the trace of a particular point in  $\mathcal{K}$  also allows us to bound *even* derivatives using just  $d$  and  $f(t)$ .

**Lemma 3.4.** *Let  $a(t) = Q(\exp td/2)Q(w^{1/2})\hat{w}(\mu)^{-1}$ . Then,  $f$  and its derivatives  $f^{(m)}$  satisfy*

$$(a) \quad f(t) = \text{tr}(a(t) + a(t)^{-1} - 2e), \text{ where } a(t) + a(t)^{-1} - 2e \in \mathcal{K}.$$

$$(b) \quad f^{(m)}(t) = \langle a(t) + (-1)^m a(t)^{-1}, d^m \rangle$$

$$(c) \quad f^{(2m)}(t) \leq \|d\|_\infty^{2m} f(t) + 2\langle e, d^{2m} \rangle$$

*Proof.* By definition of  $f$  and divergence (Definition 3.1),

$$f(t) = \langle Q(w^{1/2}) \exp td, \hat{w}(\mu)^{-1} \rangle + \langle Q(w^{-1/2}) \exp -td, \hat{w}(\mu) \rangle - 2n.$$

Substituting  $\exp td = Q(\exp td/2)e$  and  $\exp -td = Q(\exp -td/2)e$  shows the trace formula of the first statement. That  $a(t) + a(t)^{-1} - 2e \in \mathcal{K}$  follows because each eigenvalue has form  $\lambda + \frac{1}{\lambda} - 2$  for some  $\lambda \geq 0$ , which is always nonnegative.

For statement (b), we have that  $\frac{d^m}{dt^m} \exp td = d^m \exp td = Q(\exp t/2d)d^m$ . This implies that

$$\begin{aligned} \frac{d^m}{dt^m} \langle e, a(t) \rangle &= \langle Q(w^{1/2})\hat{w}(\mu)^{-1}, \frac{d^m}{(dt)^m} \exp td \rangle \\ &= \langle Q(\exp td/2)Q(w^{1/2})\hat{w}(\mu)^{-1}, d^m \rangle \\ &= \langle a(t), d^m \rangle. \end{aligned}$$

By similar argument,  $\frac{d^m}{dt^m} \langle e, a(t)^{-1} \rangle = (-1)^m \langle a(t)^{-1}, d^m \rangle$ . We conclude for all integers  $m \geq 1$  that  $f^{(m)}(t) = \langle a(t) + (-1)^m a(t)^{-1}, d^m \rangle$ . For statement (c), we have, since  $a(t) + a^{-1}(t) - 2e \in \mathcal{K}$ ,

$$\begin{aligned} f^{(2m)}(t) &= \langle a(t) + a^{-1}(t) - 2e, d^{2m} \rangle + 2\langle e, d^{2m} \rangle \\ &\leq \|a(t) + a^{-1}(t) - 2e\|_1 \|d\|_\infty^{2m} + 2\langle e, d^{2m} \rangle \\ &= \text{tr}(a(t) + a^{-1}(t) - 2e) \|d\|_\infty^{2m} + 2\langle e, d^{2m} \rangle \\ &= \|d\|_\infty^{2m} f(t) + 2\langle e, d^{2m} \rangle. \end{aligned}$$

□

Assuming  $f'(0) < 0$ , we now use these derivatives to provide a descent condition on  $t$ , i.e., we establish an interval on which  $f(t) \leq f(0)$ . Our analysis rests on Taylor's theorem, convexity of  $f$ , and the bound on  $f''(t)$  provided by the previous lemma.

**Lemma 3.5.** *Suppose that  $f'(0) < 0$ . Then,  $f(t) \leq f(0)$  if  $t \leq \frac{-2f'(0)}{\|d\|_\infty^2 f(0) + 2\|d\|^2}$ .*

*Proof.* By Taylor's theorem,  $f(t) = f(0) + f'(0)t + \frac{1}{2}f''(\zeta)t^2$  for some  $\zeta \in [0, t]$ . Further,

$$f''(\zeta) \leq \|d\|_\infty^2 f(\zeta) + 2\|d\|^2 \leq \max_{u \in \{0, t\}} (\|d\|_\infty^2 f(u) + 2\|d\|^2),$$

where the first inequality is Lemma 3.4(c) and the second inequality uses convexity of  $f(t)$ . Hence,

$$f(t) \leq f(0) + f'(0)t + \frac{1}{2} \max_{u \in \{0, t\}} (\|d\|_\infty^2 f(u) + 2\|d\|^2)t^2. \quad (10)$$

Now, let  $\hat{t}$  be the smallest  $t > 0$  for which  $f(\hat{t}) = f(0)$ . Then

$$f(0) \leq f(0) + \hat{t}f'(0) + \frac{1}{2}(\|d\|_\infty^2 f(0) + 2\|d\|^2)\hat{t}^2,$$

which implies that

$$\hat{t} \geq \frac{-2f'(0)}{\|d\|_\infty^2 f(0) + 2\|d\|^2}.$$

Since  $f(t) \leq f(0)$  for all  $t \leq \hat{t}$ , the claim follows.  $\square$

### 3.2.2 Newton direction

Suppose now that  $d = d_N(w, \mu)$ . For this direction, the divergence  $f(t)$  rapidly decreases at a rate lower bounded by  $f(0)$ . This follows from the orthogonal, direct-sum decomposition of  $d$  (Proposition 2.2) induced by  $\mathcal{L}_w := \{Q(w^{-1/2})x : x \in \mathcal{L}\}$  and  $\mathcal{L}_w^\perp = \{Q(w^{1/2})s : s \in \mathcal{L}^\perp\}$ .

**Lemma 3.6.** *Suppose that  $d = d_N(w, \mu)$ . Then  $f'(0) = -(f(0) + \|d\|^2)$ . Further,  $f(t) \leq f(0)$  if*

$$t \leq \frac{2(f(0) + \|d\|^2)}{\|d\|_\infty^2 f(0) + 2\|d\|^2}.$$

*Proof.* Let  $r_1(t) = a(t)^{-1} - e$  and  $r_2(t) = a(t) - e$ , where  $a(t)$  is as in Lemma 3.4. Then, by Lemma 3.4(b),

$$-f' = \langle a^{-1} - a, d \rangle = \langle a^{-1} - e + e - a, d \rangle = \langle r_1 - r_2, d \rangle.$$

Setting  $t = 0$  and substituting  $d = \text{proj}_{\mathcal{L}_w^\perp} r_1(0) - \text{proj}_{\mathcal{L}_w} r_2(0)$  using Proposition 2.2 gives

$$-f'(0) = \langle r_1 - r_2, d \rangle = -\langle r_1, r_2 \rangle + \|\text{proj}_{\mathcal{L}_w^\perp} r_1\|^2 + \|\text{proj}_{\mathcal{L}_w} r_2\|^2 = -\langle r_1, r_2 \rangle + \|d\|^2.$$

But  $f(t) = -\langle r_1(t), r_2(t) \rangle$  by Lemma 3.4(a), proving the first claim. The descent condition (Lemma 3.5) specialized to the Newton direction  $d = d_N$  proves the second claim.  $\square$

### 3.2.3 Proof of Theorem 3.2

We can now prove the claimed decrease in  $f(t)$  for a full Newton step, assuming  $\|d\|_\infty^2 \leq 2$ . By Lemma 3.6, we have that  $f(t) \leq f(0)$  for  $t = 1$  when  $\|d\|_\infty^2 \leq 2$ . Combining this with the quadratic upper bound (10) and Lemma 3.6 yields

$$f(1) \leq f(0) - (f(0) + \|d\|^2) + \frac{1}{2}(\|d\|_\infty^2 f(0) + 2\|d\|^2) = \frac{1}{2}\|d\|_\infty^2 f(0),$$

which is precisely the claim of Theorem 3.2.

### 3.3 Divergence bounds

Though the centered point  $\hat{w}(\mu)$  is unknown, the Newton direction  $d_N(w, \mu)$  provides a lower bound  $h_{lb}$  of the divergence  $h(w, \hat{w}(\mu))$  for any  $w \in \text{int } \mathcal{K}$  and  $\mu > 0$ . Under a norm condition, we also obtain an upper bound  $h_{ub}$  and relative-error estimates; precisely, we obtain  $h_{ub}$  and  $k \geq 1$  satisfying

$$h(w, \hat{w}(\mu)) \geq h_{lb} \geq \frac{1}{k}h(w, \hat{w}(\mu)) \quad h(w, \hat{w}(\mu)) \leq h_{ub} \leq k \cdot h(w, \hat{w}(\mu)). \quad (11)$$

These bounds use the direct-sum decomposition  $d_N = d_1 - d_2$  from Proposition 2.2 induced by the subspaces  $\mathcal{L}_w := \{Q(w^{-1/2})x : x \in \mathcal{L}\}$  and  $\mathcal{L}_w^\perp = \{Q(w^{1/2})s : s \in \mathcal{L}^\perp\}$ .

**Theorem 3.3.** For  $\mu > 0$  and  $w \in \text{int } \mathcal{K}$ , let  $d = d_N(w, \mu)$ ,  $d_1 = \text{proj}_{\mathcal{L}_w^\perp} d$ , and  $d_2 = -\text{proj}_{\mathcal{L}_w} d$ . The following statements hold:

(a)  $h(w, \hat{w}(\mu)) \geq h_{lb}$  for  $h_{lb} := \frac{\|d\|^2}{1 + \|d_1 + d_2\|_\infty}$ .

(b) If  $\|d_1 + d_2\|_\infty < 1$ , then  $h(w, \hat{w}(\mu)) \leq h_{ub}$  for  $h_{ub} := \frac{\|d\|^2}{1 - \|d_1 + d_2\|_\infty}$ . Further, the relative-error estimates (11) hold for  $k = \frac{1 + \|d_1 + d_2\|_\infty}{1 - \|d_1 + d_2\|_\infty}$ .

*Proof.* Let  $a = Q(w^{1/2})\hat{w}(\mu)^{-1}$ ,  $z = a + a^{-1} - 2e$  and  $g = a - a^{-1}$ . In this notation, we have by Proposition 2.2 that  $d_1 = \text{proj}_{\mathcal{L}_w^\perp}(a^{-1} - e)$  and  $d_2 = \text{proj}_{\mathcal{L}_w}(a - e)$ . We conclude that

$$\text{proj}_{\mathcal{L}_w^\perp}(g + 2d) = \text{proj}_{\mathcal{L}_w^\perp}(a - a^{-1} + 2(a^{-1} - e)) = \text{proj}_{\mathcal{L}_w^\perp}(a + a^{-1} - 2e) = \text{proj}_{\mathcal{L}_w^\perp} z,$$

and, similarly, that  $\text{proj}_{\mathcal{L}_w}(g + 2d) = -\text{proj}_{\mathcal{L}_w} z$ . This implies that  $\langle g + 2d, d \rangle = \langle z, d_1 + d_2 \rangle$ . Hence,

$$-\|z\|_1 \|d_1 + d_2\|_\infty \leq -\langle g + 2d, d \rangle \leq \|z\|_1 \|d_1 + d_2\|_\infty.$$

But from Lemma 3.6, we also have that  $-\langle g + 2d, d \rangle = h(w, \hat{w}(\mu)) - \|d\|^2$ . Hence,

$$-\|z\|_1 \|d_1 + d_2\|_\infty \leq h(w, \hat{w}(\mu)) - \|d\|^2 \leq \|z\|_1 \|d_1 + d_2\|_\infty.$$

Using the fact that  $\|z\|_1 = h(w, \hat{w}(\mu))$  from Lemma 3.4(a) and rearranging these inequalities gives

$$h(w, \hat{w}(\mu))(1 + \|d_1 + d_2\|_\infty) \geq \|d\|^2 \geq h(w, \hat{w}(\mu))(1 - \|d_1 + d_2\|_\infty)$$

Dividing by  $1 + \|d_1 + d_2\|_\infty$  proves the formula and error estimate for  $h_{lb}$ . Dividing by  $1 - \|d_1 + d_2\|_\infty$  proves the same for  $h_{ub}$ .  $\square$

Observe that we also obtain valid bounds by replacing  $\|d_1 + d_2\|_\infty$  with  $\|d_N(w, \mu)\|$  given that  $\|d_1 + d_2\|_\infty \leq \|d_1 + d_2\| = \|d_1 - d_2\| = \|d_N(w, \mu)\|$ . This in turn allows us to bound the size of Newton steps assuming bounds on divergence.

**Corollary 3.1.** Let  $\mu > 0$  and  $w \in \text{int } \mathcal{K}$ . If  $h(w, \hat{w}(\mu)) \leq \frac{1}{2}$ , then  $\|d_N(w, \mu)\| \leq 1$ .

*Proof.* Substituting  $\|d_1 + d_2\|_\infty$  with  $\|d_N(w, \mu)\|$  into the Theorem 3.3 lower bound yields

$$h(w, \hat{w}(\mu)) \geq \frac{\|d_N(w, \mu)\|^2}{1 + \|d_N(w, \mu)\|}, \tag{12}$$

which proves the claim.  $\square$

We now have all the ingredients needed to show convergence of Newton's method.

### 3.4 Convergence of Newton's method

We have seen that the Newton direction bounds the reduction in divergence (Theorem 3.2). Divergence in turn bounds the size of a full Newton step (Corollary 3.1). Combining these results proves quadratic convergence of the sequence  $w_0, w_1, \dots, w_m$  generated by Newton's method.

**Theorem 3.4.** For  $\mu > 0$  and  $w_0 \in \text{int } \mathcal{K}$ , let  $w_{i+1} = Q(w_i^{1/2})d_N(w_i, \mu)$ . If  $h(w_0, \hat{w}(\mu)) \leq \beta \leq \frac{1}{2}$ , then  $h(w_i, \hat{w}(\mu)) \leq \beta^{2^i}$ .

<p><b>Algorithm</b> <code>longstep</code>(<math>w_0, \mu_0, \mu_f, \epsilon</math>)</p> <pre> <math>\mu \leftarrow \mu_0, w \leftarrow w_0</math> <b>while</b> <math>\mu &gt; \mu_f</math> <b>do</b>     <math>w \leftarrow \text{center}(w, \mu, \alpha)</math>     <math>\mu \leftarrow \inf\{\mu &gt; 0 : h_{ub}(w, \mu) \leq \beta\}</math> <b>end</b> <b>return</b> <code>center</code>(<math>w, \mu, \epsilon</math>) </pre>	<p><b>Procedure</b> <code>center</code>(<math>w_0, \mu, \epsilon</math>)</p> <pre> <math>w \leftarrow w_0</math> <b>while</b> <math>h_{ub}(w, \mu) &gt; \epsilon</math> <b>do</b>     <math>d \leftarrow d_N(w, \mu)</math>     <math>t \leftarrow \gamma t_{\max}(w, \mu)</math>     <math>w \leftarrow Q(w^{1/2}) \exp(td)</math> <b>end</b> <b>return</b> <math>w</math> </pre>
--	--

Figure 2: A long-step algorithm (left) and a globally convergent centering procedure (right). The parameters  $\beta > \alpha > 0$  control distance to the central path and  $1 > \gamma > 0$  the size of Newton steps.

*Proof.* Let  $h_i = h(w_i, \hat{w}(\mu))$  and  $d_i = d_N(w_i, \mu)$ . Then,

$$h_{i+1} \leq \frac{1}{2} h_i \|d_i\|_\infty^2 \leq \frac{1}{2} h_i \|d_i\|^2 \leq \frac{1}{2} h_i (\|d_i\| + 1) h_i,$$

where the first inequality is Theorem 3.2 and the last is (12). Since  $\|d_i\| \leq 1$  by Corollary 3.1, we conclude that  $h_{i+1} \leq h_i^2$ . Hence,  $h_i \leq (h_0)^{2^i}$ , proving the claim.  $\square$

Combining this with our previous bounds relating divergence and geodesic distance (Lemma 3.2) leads to a proof of the centering lemma, reproduced below.

**Lemma 2.3** (Centering). *For  $\mu > 0$  and  $w_0 \in \text{int } \mathcal{K}$ , let  $w_{i+1} = Q(w_i^{1/2})d_N(w_i, \mu)$ . If  $\delta(w_0, \hat{w}(\mu)) \leq q^{-1}(\beta)$  for  $\beta \leq \frac{1}{2}$ , then  $\delta(w_i, \hat{w}(\mu))^2 \leq \beta^{2^i}$ .*

*Proof.* By Lemma 3.2, we conclude that  $h(w, \hat{w}(\mu)) \leq \beta \leq \frac{1}{2}$ . By Theorem 3.4, this implies that  $h(w_i, \hat{w}(\mu)) \leq \beta^{2^i}$ , which, since  $\delta(w_i, \hat{w}(\mu))^2 \leq h(w_i, \hat{w}(\mu))$ , proves the claim.  $\square$

## 4 Long-step algorithm

When proving the convergence of `shortstep` (Figure 1), we established results that together suggest an alternative algorithm (Figure 2). This alternative uses our divergence bounds (Theorem 3.3) to loosely track the central path and our descent condition (Lemma 3.6) to pick the size of Newton steps. We will show that this alternative is both scale invariant and *globally convergent*, i.e., it returns an  $\epsilon$ -approximation of a centered point  $\hat{w}(\mu)$  with  $\mu \leq \mu_f$ , given *any* initial  $w_0 \in \text{int } \mathcal{K}$  and centering parameters  $\mu_0 > \mu_f > 0$ . We call this algorithm `longstep` in reference to long-step interior-point methods [29], which also loosely track the central path. Crucial to `longstep` is a procedure `center` that is also globally convergent.

Quantities employed by `longstep` and `center` include the divergence bounds  $h_{lb}$  and  $h_{ub}$  from Theorem 3.3 and a maximum step-size  $t_{\max}$ .

**Definition 4.1.** *For  $\mu > 0$  and  $w \in \text{int } \mathcal{K}$ , let*

$$h_{lb}(w, \mu) = \frac{\|d\|^2}{1 + \|d_1 + d_2\|_\infty}, \quad h_{ub}(w, \mu) = \begin{cases} \frac{\|d\|^2}{1 - \|d_1 + d_2\|_\infty} & \|d_1 + d_2\|_\infty < 1 \\ \infty & \text{otherwise,} \end{cases}$$

$$t_{\max}(w, \mu) = 2 \frac{h_{lb}(w, \mu) + \min\{\|d\|^2, 2k\}}{\|d\|_\infty^2 (h_{lb}(w, \mu) + 2k)},$$

where  $d = d_N(w, \mu)$ ,  $d_1 = \text{proj}_{\mathcal{L}_w^\perp} d$ ,  $d_2 = -\text{proj}_{\mathcal{L}_w} d$  and  $k = (\frac{\|d\|}{\|d\|_\infty})^2$ .

Recall that  $h_{lb}$  and  $h_{ub}$  use the orthogonal decomposition (Proposition 2.2) of the Newton direction  $d_N(w, \mu) = d_1 - d_2$  that is induced by  $\mathcal{L}_w := \{Q(w^{-1/2})x : x \in \mathcal{L}\}$  and  $\mathcal{L}_w^\perp = \{Q(w^{1/2})s : s \in \mathcal{L}^\perp\}$ . The step size  $t_{\max}$  arises by lower bounding our descent condition (Lemma 3.6) with an increasing function of divergence  $h$  and then substituting  $h$  with the lower bound  $h_{lb}$ .

**Lemma 4.1.** *For  $w_0 \in \text{int } \mathcal{K}$  and  $\mu > 0$ , let  $d = d_N(w_0, \mu)$  and  $w(t) = Q(w_0^{1/2}) \exp(td)$ . Then for all  $0 \leq t \leq t_{\max}(w_0, \mu)$ , the divergence satisfies  $h(w(t), \hat{w}(\mu)) \leq h(w_0, \hat{w}(\mu))$ .*

*Proof.* From our descent condition (Lemma 3.6), the claim holds if  $t_{\max} \leq t_*$ , where

$$t_* = 2 \frac{h(w_0, \hat{w}(\mu)) + \|d\|^2}{\|d\|_\infty^2 (h(w_0, \hat{w}(\mu)) + 2k)}.$$

But this holds by observing that  $h_{lb} \leq h$  and noting that for a rational function with  $a, b > 0$ :

$$\frac{x_2 + a}{x_2 + b} \geq \frac{x_1 + \min\{a, b\}}{x_1 + b}, \text{ if } x_2 \geq x_1 \geq 0.$$

□

We can now establish convergence results. The main insight is that the sublevel sets of divergence  $h$  are compact, which implies positive lower bounds on certain progress measures. Unrelated to convergence, we also show scale invariance (Section 2.3), leveraging the fact that  $t_{\max}$  and  $h_{ub}$  depend only on the eigenvalues of  $d_1 + d_2$  (which allows us to invoke Proposition 2.3).

**Theorem 4.1.** *The algorithm `longstep` and its subroutine `center` (Figure 2) have the following properties.*

- (a) *For all inputs  $w_0 \in \text{int } \mathcal{K}$  and  $(\mu_0, \mu_f, \epsilon) > 0$ , `longstep` terminates and returns  $w$  satisfying  $h(w, \hat{w}(\mu)) \leq \epsilon$  for  $\mu \leq \mu_f$ . Further, it monotonically decreases  $\mu$ .*
- (b) *For all inputs  $w_0 \in \text{int } \mathcal{K}$  and  $(\mu, \epsilon) > 0$ , `center` terminates and returns  $w$  satisfying  $h(w, \hat{w}(\mu)) \leq \epsilon$ . Further, it monotonically decreases  $h(w, \hat{w}(\mu))$ .*
- (c) *Both `center` and `longstep` are scale invariant.*

*Proof.* We first prove statement (b). Let  $S(\zeta) = \{(w, \mu) : h(w, \hat{w}(\mu)) \leq \zeta, \mu_f \leq \mu \leq \mu_0\}$ . Then  $S(\zeta)$  is compact because it is closed and contained in a sublevel set of geodesic distance  $\delta(w, \hat{w}(\mu))$ :

$$S(\zeta) \subseteq \{(w, \mu) : \delta(w, \hat{w}(\mu))^2 \leq \zeta, \mu_f \leq \mu \leq \mu_0\}.$$

This holds given that  $h \geq \delta^2$  (Lemma 3.2).

Now, let  $w_0$  denote the initialization point of `center` and let  $\zeta = h(w_0, \hat{w}(\mu))$ . Define  $g : S(\zeta) \rightarrow \mathbb{R}$  as the decrease in  $h$  after one Newton step from  $w$ , i.e.,  $g(w, \mu) = h(\hat{w}(\mu), w) - h(\hat{w}(\mu), w')$  where  $w' = Q(w^{1/2}) \exp td$ . By Lemma 4.1 and strict convexity of  $h$  (Lemma 3.3), we have that  $g(w, \mu) > 0$  if  $w \neq w(\mu)$ . Since  $g(w, \mu)$  is continuous, the infimum

$$g_* := \inf\{g(w, \mu) : h_{ub}(\hat{w}(\mu), w) \geq \epsilon, (w, \mu) \in S(\zeta)\}$$

is attained. Moreover,  $g_* > 0$  given that  $w = \hat{w}(\mu)$  implies  $0 = h_{ub}(\hat{w}(\mu), w) < \epsilon$ . It follows that  $h(w, \hat{w}(\mu)) \leq h(w_0, \hat{w}(\mu)) - g_* N$  after  $N$  iterations. Since  $h \geq 0$ , `center` terminates for some  $N$  satisfying  $g_* N \leq h(w_0, \hat{w}(\mu)) - h(w, \hat{w}(\mu))$ .

We now prove statement (a), first noting that  $\hat{S}(\alpha) := \{(w, \mu) \in S(\alpha) : h_{ub}(w, \mu) \leq \alpha\}$  contains  $(w, \mu)$  when the  $\mu$ -update step is reached. Next, we let  $f : \hat{S}(\alpha) \rightarrow \mathbb{R}$  return the  $k > 1$  that satisfies  $h_{ub}(w, \frac{1}{k}\mu) = \beta$ . The set  $\hat{S}(\alpha)$  is compact because, as established,  $S(\alpha)$  is compact. We conclude that  $f$  attains its infimum  $f_* > 1$  on  $\hat{S}(\alpha)$ , which implies that  $\mu \leq f_*^{-M}\mu_0$  after  $M$  outer iterations. This in turn implies termination of **longstep**.

Finally, statement (c) follows from Proposition 2.3 and Lemma 2.4, given that  $t_{\max}$  and  $h_{ub}$ , viewed as functions of  $d_1$  and  $d_2$ , are invariant under transformation by an orthogonal automorphism  $M$ , i.e.,  $t_{\max}(d_1, d_2) = t_{\max}(Md_1, Md_2)$  and  $h_{ub}(d_1, d_2) = h_{ub}(Md_1, Md_2)$ .  $\square$

We close this section with practical matters related to implementation. Namely, how to efficiently evaluate the divergence bound  $h_{ub}(w, \mu)$  for fixed  $w$ , how to find the Newton direction using a least-squares technique, and how to construct feasible points for the primal-dual pair (1) after early termination.

#### 4.1 Evaluating divergence for $\mu$ -selection

For fixed  $w$ , the divergence bound  $h_{w,ub}(\mu) := h_{ub}(w, \mu)$  has a simple formula that admits efficient selection of  $\mu$  at each iteration of the long-step algorithm. To evaluate the formula, we only need to know  $\mu$  and quantities involving the vector

$$g_w := \underset{\mathcal{L}_w^\perp}{\text{proj}} Q(w^{-1/2})x_0 + \underset{\mathcal{L}_w}{\text{proj}} Q(w^{1/2})s_0,$$

where we recall that  $\mathcal{L}_w := \{Q(w^{-1/2})x : x \in \mathcal{L}\}$  and  $\mathcal{L}_w^\perp = \{Q(w^{1/2})s : s \in \mathcal{L}^\perp\}$ .

**Proposition 4.1.** *For  $w \in \text{int } \mathcal{K}$ , let  $g_w$  have minimum and maximum eigenvalues  $\lambda_{\min}$  and  $\lambda_{\max}$ . Let  $k(\mu) = \min(\frac{1}{\sqrt{\mu}}\lambda_{\min}, 2 - \frac{1}{\sqrt{\mu}}\lambda_{\max})$ . Then,*

$$h_{w,ub}(\mu) = \begin{cases} \frac{\frac{1}{\mu}\|g_w\|^2 - 2\frac{1}{\sqrt{\mu}}\text{tr } g_w + n}{k(\mu)} & k(\mu) > 0 \\ \infty & \text{otherwise.} \end{cases}$$

*Proof.* Let  $d, d_1$  and  $d_2$  be as Definition 4.1, and suppose that  $h_{ub}$  is finite, i.e.,  $1 - \|d_1 + d_2\|_\infty > 0$ . Then, we have that

$$h_{ub} = \frac{\|d\|^2}{1 - \|d_1 + d_2\|_\infty}, \quad d_1 + d_2 = \frac{1}{\sqrt{\mu}}g_w - e.$$

Hence,  $\|d_1 + d_2\|_\infty$  is the max of  $\frac{1}{\sqrt{\mu}}\lambda_{\max} - 1$  and  $1 - \frac{1}{\sqrt{\mu}}\lambda_{\min}$ . The claimed denominator  $k(\mu)$  follows using the identity

$$1 - \max(1 - a, b - 1) = 1 + \min(a - 1, 1 - b) = \min(a, 2 - b).$$

The identity for  $\|d\|^2$  follows by expanding  $\|\frac{1}{\sqrt{\mu}}g_w - e\|^2$  and observing that  $\|d\| = \|d_1 + d_2\|$ .  $\square$

#### 4.2 Feasible points

Since the presented algorithms update  $w$  along geodesics, the point  $\sqrt{\mu}(w, w^{-1})$  only satisfies the affine constraints of the primal-dual pair (1) in the limit. Nevertheless, under a norm condition, we can *always* produce a feasible  $(x, s)$  from the Newton direction  $d_N(w, \mu)$ .



**Proposition 4.2.** For  $w \in \text{int } \mathcal{K}$  and  $\mu > 0$ , let  $d = d_N(w, \mu)$  and

$$x = \sqrt{\mu}Q(w^{1/2})(e + d), \quad s = \sqrt{\mu}Q(w^{-1/2})(e - d).$$

If  $\|d\|_\infty \leq 1$ , then  $(x, s)$  is feasible for (1).

*Proof.* By definition of the Newton direction (Definition 2.1), it holds that  $x \in x_0 + \mathcal{L}$  and  $s \in s_0 + \mathcal{L}^\perp$ . Further, since  $\|d\|_\infty \leq 1$ , we have that  $e \pm d \in \mathcal{K}$ . Finally,  $x, s \in \mathcal{K}$  given that  $Q(z)y \in \mathcal{K}$  for all  $z \in \mathcal{J}$  and  $y \in \mathcal{K}$ .  $\square$

### 4.3 Newton direction via least squares

Interior-point methods typically find search directions by solving least-squares problem of the form

$$\text{minimize}_y \frac{1}{2}y^T A^*W(x, s)Ay - f^T y \text{ subject to } By = g,$$

where  $W(x, s)$  is a positive-definite weighting matrix induced by the current iterate  $(x, s)$  and  $(A, B, f, g)$  are parameters induced by the affine constraints  $x_0 + \mathcal{L}$  and  $s_0 + \mathcal{L}^\perp$ . Equivalently, they solve linear systems of the form

$$\begin{bmatrix} A^*W(x, s)A & B^* \\ B & 0 \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}$$

for which specialized algorithms exist (e.g., [13]). Such a system can also yield the Newton direction  $d_N(w, \mu)$ . This, of course, is not surprising given its construction via orthogonal projection (Proposition 2.2). Nevertheless, we give this system explicitly for affine constraints of the form:

$$s_0 + \mathcal{L}^\perp = \{c - Ay : By = g, y \in \mathbb{R}^m\}, \quad x_0 + \mathcal{L} = \{x \in \mathcal{J} : \exists z \in \mathbb{R}^d A^*x + B^*z = b\},$$

where  $(y, z) \in \mathbb{R}^m \times \mathbb{R}^d$  denote additional variables,  $A : \mathbb{R}^m \rightarrow \mathcal{J}$  and  $B : \mathbb{R}^m \rightarrow \mathbb{R}^d$  are linear maps with adjoint operators  $A^* : \mathcal{J} \rightarrow \mathbb{R}^m$  and  $B^* : \mathbb{R}^d \rightarrow \mathbb{R}^m$ , and  $(b, g, c) \in \mathbb{R}^m \times \mathbb{R}^d \times \mathcal{J}$  are fixed parameters.

In this notation, the Newton direction becomes the  $d$  that for some  $(y, z)$  solves

$$A^*(Q(w^{1/2})(e + d)) = \frac{1}{\sqrt{\mu}}b - B^*z, \quad Q(w^{-1/2})(e - d) = \frac{1}{\sqrt{\mu}}c - Ay, \quad By = \frac{1}{\sqrt{\mu}}g. \quad (13)$$

Eliminating  $d$  and using  $Q(w) = Q(w^{1/2})Q(w^{-1/2})^{-1}$  yields a system with the desired form. Note by modifying the right-hand-side of this system, we can also construct the direct-summands  $d_1$  and  $d_2$  of Proposition 2.2.

**Proposition 4.3.** For  $w \in \text{int } \mathcal{K}$  and  $\mu > 0$ , let  $(y, z) \in \mathbb{R}^m \times \mathbb{R}^d$  solve the least-squares system

$$\begin{bmatrix} A^*Q(w)A & B^* \\ B & 0 \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{\mu}}(b + A^*Q(w)c) - 2A^*w \\ \frac{1}{\sqrt{\mu}}g \end{bmatrix}.$$

Then, the Newton direction satisfies  $d_N(w, \mu) = e - Q(w^{1/2})(\frac{1}{\sqrt{\mu}}c - Ay)$ .

*Proof.* From the second equation of (13), we conclude that  $d = e - Q(w^{1/2})(\frac{1}{\sqrt{\mu}}c - Ay)$ . Substituting into the first equation yields

$$\begin{aligned} \frac{1}{\sqrt{\mu}}b - B^*z &= A^*Q(w^{1/2})\left(2e - Q(w^{1/2})\left(\frac{1}{\sqrt{\mu}}c - Ay\right)\right) \\ &= 2A^*w - \frac{1}{\sqrt{\mu}}A^*Q(w)c + A^*Q(w)Ay. \end{aligned}$$

Rearranging terms proves the claim.  $\square$

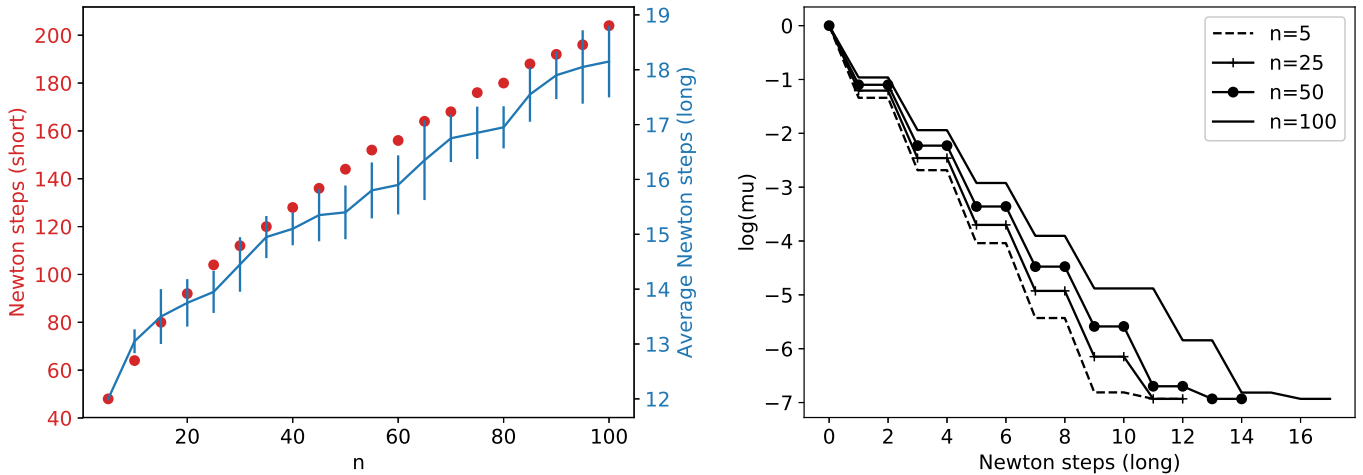


Figure 3: Total Newton steps vs  $n$  for `shortstep` and `longstep` (left). (Note the different scales.) Typical decrease in centering parameter  $\mu$  for `longstep` (right).

## 5 Computational results

We compare the number of Newton steps executed by the algorithms `shortstep` and `longstep` (Figures 1-2). We also illustrate global convergence of the centering procedure. Results show superior performance of `longstep`, mirroring the performance gap between short- and long-step interior-point methods [24]. We also observe distinct convergence phases for `center`. Test problems are randomly generated semidefinite programs. We choose the matrix logarithms of the parameters  $(x_0, s_0)$  and a basis for the subspace  $\mathcal{L}$  with the operations  $X \leftarrow \text{randn}(n, n)$  and  $X \leftarrow \frac{1}{2}(X + X^T)$ . Here,  $\text{randn}(n, n)$  denotes an  $n \times n$  matrix with entries drawn independently from the normal distribution (with zero mean and unit variance). In all examples,  $\dim \mathcal{L} = 10$ .

### 5.1 Newton iterations

We compare (Figure 3) the total number of Newton steps `longstep` and `shortstep` execute to update an initial centered point  $\hat{w}(\mu)$  to  $\hat{w}(\frac{1}{1024}\mu)$ . For each  $n$ , we compute the average number of steps executed by `longstep` over twenty random problems. The number of steps executed by `shortstep` is independent of the problem instance, so no averaging is necessary. As shown, `longstep` provides a significant constant-factor improvement over `shortstep`. We also see `longstep` only weakly depends on the specific instance, as the standard deviation of iterations is less than one. Finally, `longstep` evidently reduces the centering parameter  $\mu$  at a constant rate that decreases with increasing  $n$ . The algorithm `shortstep` has this property by design.

We selected parameters for the algorithms as follows. For `longstep`, we arbitrarily chose a divergence bound of  $\beta = 100$ , a recentering tolerance of  $\alpha = 10$ , and a final centering tolerance of  $\epsilon = \frac{1}{10000}$ . For the step size, we took  $\gamma = \frac{1}{2}$ , which corresponds to minimizing the 2nd-order Taylor expansion of divergence (Section 3.2.1). For `shortstep`, we selected the centering-parameter update  $k$  and the number of inner iterations  $m$  using Theorem 2.1 with the parameter values  $(\beta, \epsilon) = (\frac{1}{2}, \frac{1}{10000})$ .

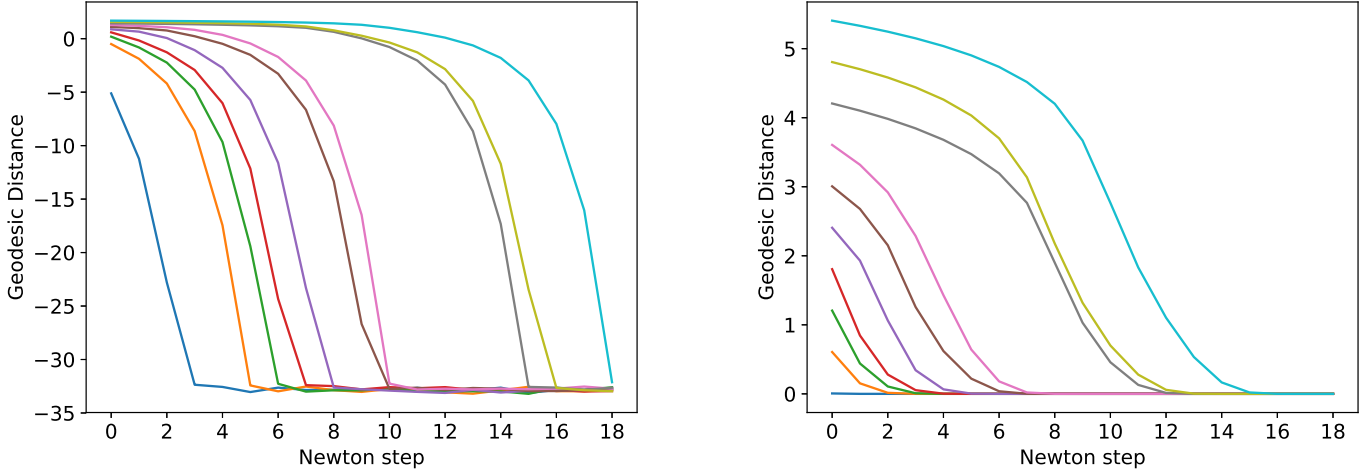


Figure 4: Convergence of the centering procedure for different initialization points in log (left) and linear (right) scalings.

## 5.2 Global convergence

The procedure `center` used by `longstep` globally converges. That is, it always returns  $\hat{w}(\mu)$  given an *arbitrary* initial point  $w_0 \in \text{int } \mathcal{K}$  and centering parameter  $\mu > 0$ . For a fixed problem instance, we plot convergence behavior for different initial conditions (Figure 4). We observe that the rate transitions from linear to quadratic as the geodesic distance  $\delta(w_i, \hat{w}(\mu))$  decreases. The latter phase is expected (Theorem 3.4). For step sizes, we took  $\gamma = \frac{1}{2}$ .

## References

- [1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.
- [2] F. Alizadeh, J.-P. A. Haeberly, and M. L. Overton. Primal-dual interior-point methods for semidefinite programming: convergence rates, stability and numerical results. *SIAM Journal on Optimization*, 8(3): 746–768, 1998.
- [3] R. Bhatia. *Positive definite matrices*. Princeton University Press, 2009.
- [4] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2009.
- [5] J. Faraut and A. Korányi. *Analysis on symmetric cones*. Oxford University Press, 1994.
- [6] L. Faybusovich. Euclidean Jordan algebras and interior-point algorithms. *Positivity*, 1(4):331–357, 1997.
- [7] L. Faybusovich. Linear systems in Jordan algebras and primal-dual interior-point algorithms. *Journal of computational and applied mathematics*, 86(1):149–175, 1997.
- [8] M. Halická, E. de Klerk, and C. Roos. On the convergence of the central path in semidefinite optimization. *SIAM Journal on Optimization*, 12(4):1090–1099, 2002.
- [9] M. Harandi, M. Salzmann, and R. Hartley. Dimensionality reduction on SPD manifolds: The emergence of geometry-aware methods. *IEEE transactions on pattern analysis and machine intelligence*, 40(1): 48–62, 2017.

- [10] C. Helmberg, F. Rendl, R. J. Vanderbei, and H. Wolkowicz. An interior-point method for semidefinite programming. *SIAM Journal on Optimization*, 6(2):342–361, 1996.
- [11] M. Kojima, S. Mizuno, and A. Yoshise. A primal-dual interior point algorithm for linear programming. In *Progress in mathematical programming*, pages 29–47. Springer, 1989.
- [12] M. Kojima, S. Shindoh, and S. Hara. Interior-point methods for the monotone semidefinite linear complementarity problem in symmetric matrices. *SIAM Journal on Optimization*, 7(1):86–125, 1997.
- [13] C. L. Lawson and R. J. Hanson. *Solving least squares problems*. SIAM, 1995.
- [14] H. Lee and Y. Lim. Metric and spectral geometric means on symmetric cones. *Kyungpook Mathematical Journal*, 47(1), 2007.
- [15] Y. Lim. Geometric means on symmetric cones. *Archiv der Mathematik*, 75(1):39–45, 2000.
- [16] Y. Lim. Riemannian and Finsler structures of symmetric cones. *Trends in Mathematics*, 4(2):111–118, 2001.
- [17] M. Moakher and P. G. Batchelor. Symmetric positive-definite matrices: From geometry to applications and visualization. In *Visualization and Processing of Tensor Fields*, pages 285–298. Springer, 2006.
- [18] R. D. Monteiro. Primal–dual path-following algorithms for semidefinite programming. *SIAM Journal on Optimization*, 7(3):663–678, 1997.
- [19] R. D. Monteiro and I. Adler. Interior path following primal-dual algorithms. Part I: Linear programming. *Mathematical programming*, 44(1-3):27–41, 1989.
- [20] Y. Nesterov and A. Nemirovski. Primal central paths and Riemannian distances for convex sets. *Foundations of Computational Mathematics*, 8(5):533–560, 2008.
- [21] Y. E. Nesterov and M. J. Todd. Primal-dual interior-point methods for self-scaled cones. *SIAM Journal on optimization*, 8(2):324–364, 1998.
- [22] Y. E. Nesterov, M. J. Todd, et al. On the Riemannian geometry defined by self-concordant barriers and interior-point methods. *Foundations of Computational Mathematics*, 2(4):333–361, 2002.
- [23] J. Peng, C. Roos, and T. Terlaky. *Self-regularity: a new paradigm for primal-dual interior-point algorithms*, volume 22. Princeton University Press, 2009.
- [24] F. A. Potra and S. J. Wright. Interior-point methods. *Journal of Computational and Applied Mathematics*, 124(1-2):281–302, 2000.
- [25] H. Ramírez and D. Sossa. On the central paths in symmetric cone programming. *Journal of Optimization Theory and Applications*, 172(2):649–668, 2017.
- [26] M. J. Todd. A study of search directions in primal-dual interior-point methods for semidefinite programming. *Optimization methods and software*, 11(1-4):1–46, 1999.
- [27] M. J. Todd, K.-C. Toh, and R. H. Tütüncü. On the Nesterov–Todd direction in semidefinite programming. *SIAM Journal on Optimization*, 8(3):769–796, 1998.
- [28] L. Tunçel. Primal-dual symmetry and scale invariance of interior-point algorithms for convex optimization. *Mathematics of Operations Research*, 23(3):708–718, 1998.
- [29] S. J. Wright. *Primal-dual interior-point methods*. SIAM, 1997.
- [30] M. Zangiabadi, G. Gu, and C. Roos. Full Nesterov-Todd step primal-dual interior-point methods for second-order cone optimization. *J. Optim. Theory Appl*, 158(3), 2008.

## A Appendix

This section contains background results about the Euclidean Jordan algebra  $\mathcal{J}$  and cone-of-squares  $\mathcal{K}$  that we referenced without proof. The first establishes properties of the quadratic representation  $Q(w)z := 2w \circ (w \circ z) - (w \circ w) \circ z$ .

**Lemma A.1** ([5]). *The following statements hold.*

1.  $Q(w)^{-1} = Q(w^{-1})$  for all invertible  $w \in \mathcal{J}$ .
2.  $(Q(w)z)^{-1} = Q(w^{-1})z^{-1}$  for all invertible  $w, z \in \mathcal{J}$ .
3.  $Q(Tw) = TQ(w)T^*$  for all  $w \in \mathcal{J}$  and automorphisms  $T : \mathcal{J} \rightarrow \mathcal{J}$  of  $\mathcal{K}$ .
4.  $Q(w)^2 = Q(w^2)$  for all  $w \in \mathcal{J}$ .
5.  $Q(w)e = w^2$  for all  $w \in \mathcal{J}$ .
6.  $Q(w)$  is self-adjoint, i.e.,  $\langle Q(w)x, y \rangle = \langle x, Q(w)y \rangle$  for all  $w, x, y \in \mathcal{J}$ .

*Proof.* The first properties are Propositions II.3.1., II.3.3, III.5.2, p. 55, and p. 48 of [5]. The last is evident from the definition of  $Q(w)$  and the fact that Jordan multiplication is self-adjoint, i.e.,  $\langle x \circ y, z \rangle = \langle y, x \circ z \rangle$ .  $\square$

The next establishes properties of orthogonal automorphisms of  $\mathcal{K}$ . They trivially follow from the fact that such automorphisms are precisely the Jordan-algebra automorphisms of  $\mathcal{J}$  given our use of the trace inner-product [5, p. 56].

**Lemma A.2.** *Let  $M : \mathcal{J} \rightarrow \mathcal{J}$  be an orthogonal automorphism of  $\mathcal{K}$ . Then, the following statements hold for all  $w \in \mathcal{J}$ .*

1. If  $w$  is an idempotent, i.e.,  $w \circ w = w$ , then  $Mw$  is an idempotent.
2. If  $w$  has spectral decomposition  $\sum_{i=1}^n \lambda_i e_i$ , then  $Mw$  has spectral decomposition  $\sum_{i=1}^n \lambda_i M e_i$ .
3.  $\exp Mw = M \exp w$ .

Further,  $Me = e$ .

*Proof.* By use of the trace inner-product,  $M$  is also an automorphism of  $\mathcal{J}$  [5, p. 56] and hence satisfies  $(Mx) \circ (My) = M(x \circ y)$ . Hence,  $(Mw) \circ (Mw) = M(w \circ w) = Mw$ , showing the first statement. The second statement is immediate from the first: if  $w$  has spectral decomposition  $\sum_{i=1}^n \lambda_i e_i$ , then  $Mw$  has decomposition  $\sum_{i=1}^n \lambda_i M e_i$ , since the  $M e_i$  are idempotent and pairwise orthogonal, i.e.,  $\langle M e_i, M e_j \rangle = \langle e_i, M^* M e_j \rangle = \langle e_i, e_j \rangle = 0$ . The third is immediate from the second:

$$\exp Mw = \sum_{i=1}^n \exp(\lambda_i) M e_i = \sum_{i=1}^n M \exp(\lambda_i) e_i = M \exp w.$$

Finally,  $Me = e$  given that  $e = \exp M0 = M \exp 0 = Me$ .  $\square$