SIAM J. OPTIM.
Vol. 29, No. 3, pp. 2100–2127

HESSIAN BARRIER ALGORITHMS FOR LINEARLY CONSTRAINED OPTIMIZATION PROBLEMS[*]

IMMANUEL M. BOMZE[†], PANAYOTIS MERTIKOPOULOS[‡], WERNER SCHACHINGER[†], AND MATHIAS STAUDIGL[§]

Abstract. In this paper, we propose an interior-point method for linearly constrained—and possibly nonconvex—optimization problems. The method—which we call the Hessian barrier algorithm (HBA)—combines a forward Euler discretization of Hessian–Riemannian gradient flows with an Armijo backtracking step-size policy. In this way, HBA can be seen as an alternative to mirror descent, and contains as special cases the affine scaling algorithm, regularized Newton processes, and several other iterative solution methods. Our main result is that, modulo a nondegeneracy condition, the algorithm converges to the problem's critical set; hence, in the convex case, the algorithm converges globally to the problem's minimum set. In the case of linearly constrained quadratic programs (not necessarily convex), we also show that the method's convergence rate is $O(1/k\rho)$ for some $\rho \in (0, 1]$ that depends only on the choice of kernel function (i.e., not on the problem's primitives). These theoretical results are validated by numerical experiments in standard nonconvex test functions and large-scale traffic assignment problems.

Key words. Hessian–Riemannian gradient descent, interior-point methods, mirror descent, nonconvex optimization, traffic assignment

AMS subject classifications. 90C51, 90C30, 90C25, 90C26

DOI. 10.1137/18M1215682

1. Introduction.

Consider a linearly constrained optimization problem of the form

$$\text{(Opt)} \qquad \begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && Ax = b,\ x \geq 0. \end{aligned}$$

In this formulation, the primitives of (Opt) are[1]

(i) the problem's *objective function* $f\colon \mathcal{C} \to \mathbb{R} \cup \{+\infty\}$, where $\mathcal{C} \equiv \mathbb{R}^n_+$ denotes the nonnegative orthant of $\mathbb{R}^n$;

(ii) the problem's *feasible region*

$$\mathcal{X} = \{x \in \mathbb{R}^n : Ax = b,\ x \geq 0\}, \tag{1.1}$$

where $A \in \mathbb{R}^{m\times n}$ is a matrix of rank $m \geq 0$ and $b \in \mathbb{R}^m$ is an m-dimensional real vector (both assumed known to the optimizer).

[*] Received by the editors September 25, 2018; accepted for publication (in revised form) May 28, 2019; published electronically August 27, 2019.
https://doi.org/10.1137/18M1215682

Funding: The second author was partially supported by the French National Research Agency (ANR) project ORACLESS (ANR-16-CE33-0004-01). The second and fourth authors were partially supported by the COST Action CA16228 "European Network for Game Theory" (GAMENET).

[†] ISOR, University of Vienna, 1090 Vienna, Austria (immanuel.bomze@univie.ac.at, Werner.Schachinger@univie.ac.at).

[‡] Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38401, Grenoble, France (panayotis.mertikopoulos@imag.fr).

[§] Department of Quantitative Economics, Maastricht University, 6211 Maastricht, The Netherlands (m.staudigl@maastrichtuniversity.nl).

[1] Inequality constraints of the form $Ax \leq b$ can also be accommodated in (Opt) by introducing the corresponding slack variables $s = b - Ax \geq 0$. Despite the slight loss in parsimony, the equality form of (Opt) turns out to be more convenient in terms of notational overhead, so we stick with the standard equality formulation throughout.

Problems of this type are ubiquitous: they arise naturally in data science and machine learning [20, 31], game theory and operations research [16, 43], imaging science and signal processing [10, 11, 36], information theory and statistics [22, 23], networks [12], traffic engineering [28], and many other fields where continuous optimization plays a major role. In addition, (Opt) also covers continuous relaxations of NP-hard discrete optimization problems ranging from the maximum clique problem to integer linear programming [14, 17]. As such, it should come as no surprise that (Opt) has given rise to a thriving literature on iterative algorithmic methods aiming to reach an approximate solution in a reasonable amount of time.

Even though it is not possible to adequately review this literature here, we should point out that it includes methods as diverse as quasi-Newton algorithms, conditional gradient descent (Frank–Wolfe), interior-point and active-set methods, and Bregman proximal/mirror descent schemes. In particular, one very fruitful strategy for solving (Opt) is to take a continuous-time viewpoint and design ordinary differential equations (ODEs) whose solution trajectories are "negatively correlated" with the gradient of f—see, e.g., [3, 4, 5, 13, 21, 38, 48, 54] and references therein. Doing so sheds new light on the properties of many algorithms proposed to solve (Opt), provides Lyapunov functions to analyze their asymptotic behavior, and often leads to new classes of algorithms altogether.

A classical example of this heuristic arises in the study of dynamical systems derived from a *Hessian–Riemannian* (HR) metric, i.e., a Riemannian metric induced by the Hessian of a Legendre-type function [1, 2, 13, 25]. To make this more precise (see subsection 2.2 for the details), the *Hessian–Riemannian gradient descent* (HRGD) dynamics for (Opt) can be stated as

$$\dot{x} = -P(x)H(x)^{-1}\nabla f(x), \tag{HRGD}$$

where

1. $H(x) = \nabla^2 h(x)$ for some convex *barrier function* $h\colon \mathcal{C} \to \mathbb{R} \cup \{+\infty\}$ that satisfies a *steepness* (or *essential smoothness*) condition of the form

$$\lim_{k\to\infty} \|\nabla h(x^k)\|_2 = \infty \tag{1.2}$$

   for every sequence of interior points $x^k \in \operatorname{ri}(\mathcal{C})$ converging to the boundary $\operatorname{bd}(\mathcal{C})$ of $\mathcal{C}$;
2. $P(x)$ is the (Riemannian) projection map for the null space $\mathcal{A}_0 = \ker A \equiv \{x \in \mathbb{R}^n : Ax = 0\}$ of A; concretely, $P(x)$ has the closed-form expression

$$P(x) = I - H(x)^{-1}A^\top(AH(x)^{-1}A^\top)^{-1}A. \tag{1.3}$$

The intuition behind (HRGD) is as simple as it is elegant: to derive an interior-point method for (Opt), the positive orthant $\operatorname{ri}(\mathcal{C})$ is endowed with a Riemannian geometric structure that "blows up" near its boundary (i.e., distances between points increase near the boundary). In so doing, the (unrestricted) Riemannian gradient $\operatorname{grad} f(x) = H(x)^{-1}\nabla f(x)$ of f becomes vanishingly small near the boundary of $\operatorname{ri}(\mathcal{C})$. The projection map $P(x)$ further guarantees that the dynamics evolve in the affine hull $\mathcal{A} \equiv \{x \in \mathbb{R}^n : Ax = b\}$ of $\mathcal{X}$ (assumed throughout to be nonempty); as a result, the solution trajectories of (HRGD) starting in the relative interior $\operatorname{ri}(\mathcal{X})$ of $\mathcal{X}$ remain in $\operatorname{ri}(\mathcal{X})$ for all $t \geq 0$ [1].

If the objective function f is convex, (HRGD) enjoys very robust convergence guarantees [1, 2], and many recent developments in acceleration techniques can also be traced back to this basic scheme—see, e.g., [54] and references therein. However, harvesting the full algorithmic potential of (HRGD) also requires a suitable discretization of the dynamics in order to obtain a bona fide, implementable algorithm. In [6], this was done via a discretization scheme that ultimately gives rise to the *mirror descent* (MD) update rule

$$x^+ = \underset{x' \in \mathcal{X}}{\arg\min}\{\alpha \nabla f(x)^\top (x' - x) + D(x', x)\}, \tag{MD}$$

where $x^+ \in \mathcal{X}$ denotes the algorithm's new state starting from $x \in \mathcal{X}$, α is the method's step size, and D denotes the *Bregman divergence* of h, i.e.,

$$D(x', x) = h(x') - h(x) - \nabla h(x)^\top (x' - x). \tag{1.4}$$

First introduced by Nemirovski and Yudin [40] for nonsmooth problems, the mirror descent algorithm and its variants have met with prolific success in convex programming [9], online and stochastic optimization [47], variational inequalities [41], noncooperative games [19, 39], and many other fields of optimization theory and its applications. Nevertheless, despite the appealing convergence properties of (MD), it is often difficult to calculate the update step from x to x^+ when the problem's feasible region $\mathcal{X}$ is not "prox-friendly," i.e., when there is no efficient oracle for solving the convex optimization problem in (MD) [24]. With this in mind, our main goal in this paper is to provide a convergent, *forward* discretization of (HRGD) which does not require solving a convex optimization problem at each update step.

Our contributions and prior work. Our starting point is to consider an Euler discretization of (HRGD) which we call the *Hessian barrier algorithm* (HBA), and which can be described by the update rule

$$x^+ = x - \alpha P(x) H(x)^{-1} \nabla f(x). \tag{HBA}$$

In the above, $H(x)$ and $P(x)$ are defined as in (HRGD), while the algorithm's step size $\alpha \equiv \alpha(x)$ is determined via an Armijo backtracking rule that we describe in detail in section 3. Before discussing our general results, we provide below a small sample of classical first-order schemes which can be seen as direct antecedents of HBA.

Example 1.1 (Lotka–Volterra systems). Let $m = 0$, so that the feasible region of (Opt) is the nonnegative orthant $\mathcal{C} = \mathbb{R}^n_+$ of $\mathbb{R}^n$. If we set

$$\theta(t) = \begin{cases} t \log t & \text{for } p = 1, \\ \frac{1}{(2-p)(1-p)} t^{2-p} & \text{for } p \in (1, 2), \\ -\log t & \text{for } p = 2, \end{cases} \tag{1.5}$$

and $h(x) = \sum_{i=1}^n \theta(x_i)$, some straightforward algebra gives the Lotka–Volterra rule

$$x_i^+ = x_i - \alpha x_i^p \partial_i f(x), \tag{LV}$$

where we write $\partial_i f(x)$ for the ith partial derivative of f at x (for simplicity, we are also dropping the dependence of $\alpha(x)$ on x). For the convergence analysis of a special case of this system (modulo a regularization term), see [6] and references therein.

Example 1.2 (the replicator dynamics). Let $A = (1, \ldots, 1) \in \mathbb{R}^{1 \times n}$ and $b = 1$, so that the feasible region of (Opt) is the unit simplex $\mathcal{X} = \{x \in \mathbb{R}^n_+ : \sum_i x_i = 1\}$.

If we take the negative entropy function $h(x) = \sum_{i=1}^n x_i \log x_i$ stemming from the choice $p = 1$ above, a direct calculation yields $H(x) = \operatorname{diag}(1/x_1, \ldots, 1/x_n)$ and $P(x) = I - x \cdot (1, \ldots, 1)$. The induced Hessian–Riemannian system is known as the *replicator dynamics* (RD) and the corresponding incarnation of (HBA) takes the form

$$\text{(RD)} \qquad x_i^+ = x_i - \alpha x_i \Big[\partial_i f(x) - \textstyle\sum_{j=1}^n x_j \partial_j f(x)\Big].$$

The continuous-time version of (RD) has a long history in evolutionary game theory [32] and it has been successfully applied to a wide range of relaxations of NP-hard optimization problems [15, 17].

Example 1.3 (affine scaling). Suppose that $f(x) = c^\top x$ for some cost vector $c \in \mathbb{R}^n$. Then, defining $h(x)$ as in Example 1.1, we obtain the *affine scaling* (AS) scheme

$$\text{(AS)} \qquad x^+ = x - \alpha[I - X^p A^\top (AX^p A^\top)^{-1} A] X^p c,$$

where $X = \operatorname{diag}(x_1, \ldots, x_n)$. The origins of (AS) can be traced back to the work of Dikin in the 1960s and Karmarkar in the 1980s; the convergence of the specific incarnation (AS) was established in the seminal paper of Vanderbei, Meketon, and Freedman [51].

Example 1.4 (regularized Newton methods). Suppose that $m = 0$ (so there are no equality constraints), and f is convex and twice continuously differentiable. Setting $h(x) = f(x) + \frac{1}{2}\beta\|x\|_2^2$, we get $H(x) = \beta I + \nabla^2 f(x)$, leading in turn to the *regularized Newton* (RN) update rule

$$\text{(RN)} \qquad x^+ = x - \alpha[\beta I + \nabla^2 f(x)]^{-1} \nabla f(x).$$

If f is *self-concordant* [42], the barrier function $h(x)$ satisfies the steepness requirement (1.2), so (RN) can be seen as a special case of (HBA). The convergence of this method was studied in detail in a recent paper by Polyak [46].

The examples above show that (HBA) is a flexible method that covers several existing algorithms as special cases, and which can be easily tuned to the specifics of the problem at hand. To analyze its asymptotic behavior, we introduce an Armijo backtracking procedure which guarantees "sufficient decrease" of the value of f at each stage. In so doing, we are able to show that the sequence x^k, $k = 0, 1, \ldots$, of the algorithm's generated iterates converges to the set of Karush–Kuhn–Tucker (KKT) points of (Opt) under mild regularity assumptions on f and a full row-rank assumption of the constraint matrix A (cf. Theorem 4.1). As an immediate corollary of this, we show that every limit point of (HBA) is a global minimum of f if the objective function of (Opt) is convex. This global convergence result closes a significant open issue in the asymptotic analysis of Tseng, Bomze, and Schachinger [50] for Armijo methods, where convergence of a replicator-type system is proved modulo a "nonvanishing" step-size hypothesis which cannot be verified directly from the problem's primitives. As we show here, this step-size assumption is by no means harmless, and requires a delicate argument to establish.

In the special case where f is quadratic (but otherwise possibly nonconvex), we further show that $f(x^k)$ converges at a sublinear rate of $\mathcal{O}(1/k^\rho)$ for some $\rho \in (0, 1]$ depending only on the choice of the method's barrier function. This shows that the chosen barrier function is a key design parameter for the convergence properties of (HBA); we discuss this issue in detail in section 5.

Finally, in section 6, we supplement our theoretical analysis by means of extensive numerical experiments with standard global optimization test functions (such as the Rosenbrock and Beale benchmarks), and we examine the method's observed convergence rate in a large-scale traffic assignment problem.

Notation. For all $x \in \mathbb{R}^n$, we will write $\operatorname{diag}(x) \equiv \operatorname{diag}(x_1, \dots, x_n)$ for the diagonal $n \times n$ matrix with the coordinates of x on the main diagonal. We set $S = \{1, 2, \dots, n\}$, and write $S_x = \{i \in S : x_i \neq 0\}$ for the support of the vector $x \in \mathbb{R}^n$. For $x \in \mathbb{R}^n$ and $J \subset S$, we let $x_J = (x_j)_{j \in J}$ denote the restriction of x to the coordinates in the index set J. Finally, we will write $\mathcal{S}^n$, $\mathcal{S}^n_+$, and $\mathcal{S}^n_{++}$ for the space of real $n \times n$ symmetric, positive-semidefinite, and positive-definite matrices, respectively.

2. Problem setup and preliminaries.

2.1. Definitions and assumptions.

Throughout what follows, we will make the following blanket assumptions for (Opt).

Assumption 1. The objective $f \colon \mathcal{C} \to \mathbb{R} \cup \{+\infty\}$ of (Opt) satisfies the following:

(a) f is proper and lower semicontinuous (l.s.c.) on $\mathcal{C}$, continuously differentiable on $\mathcal{X}$, and ∇f is L-Lipschitz continuous on $\mathcal{X}$.

(b) There exists some $x^0 \in \operatorname{ri}(\mathcal{X})$ such that the sublevel set

$$[f \leq f(x^0)] \equiv \{x \in \mathcal{X} : f(x) \leq f(x^0)\}$$

is bounded.

Assumption 1(b) is trivial when $\mathcal{X}$ is itself bounded; moreover, taken together, Assumptions 1(a) and 1(b) imply that the sublevel set $[f \leq f(x^0)]$ is compact, so f attains its minimum therein. These assumptions are quite standard in interior-point methods and, in particular, affine scaling schemes; for an in-depth discussion, see [30] and references therein.

To formulate the first-order optimality conditions for (Opt), consider the Lagrangian

$$\mathcal{L}(x, y, u) = f(x) - y^\top (Ax - b) - u^\top x, \tag{2.1}$$

where $y \in \mathbb{R}^m$ and $u \in \mathbb{R}^n$ are the Lagrange multipliers respectively corresponding to the problem's equality and inequality constraints. The KKT conditions for (Opt) may then be written as

$$\begin{aligned} \nabla f(x) &= A^\top y + u, \\ Ax &= b, \\ u_i x_i &= 0,\ u_i \geq 0 \text{ for all } i = 1, \dots, n. \end{aligned} \tag{KKT}$$

The set of all points $x^* \in \mathcal{X}$ for which the system (KKT) admits a solution (y, u) will be denoted in what follows by $\mathcal{X}^*$. As all constraints are linear, we do not need any constraint qualifications, and all local minima of f also lie in $\mathcal{X}^*$ by default.

Since the existence of a minimizer is guaranteed by Assumption 1, it follows that $\mathcal{X}^*$ is nonempty. Note also that if $x^* \in \mathcal{X}^*$, then there exists some $y^* \in \mathbb{R}^m$ such that

$$\nabla f(x^*) - A^\top y^* \geq 0, \tag{2.2a}$$

$$\operatorname{diag}(x^*)(\nabla f(x^*) - A^\top y^*) = 0. \tag{2.2b}$$

2.2. Elements of Riemannian geometry. A key notion in our considerations is that of a *Riemannian metric*, i.e., a position-dependent variant of the ordinary (Euclidean) scalar product between vectors. To define it, recall first that a *scalar product* on $\mathbb{R}^n$ is a symmetric, positive-definite bilinear form $\langle\cdot,\cdot\rangle\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$.[2] This product defines a norm in the usual way and it can be represented equivalently via its *metric tensor*, that is, a symmetric, positive-definite matrix $H \in \mathcal{S}^n_{++}$ with components

$$H_{ij} = \langle e_i, e_j\rangle \tag{2.3}$$

in the standard basis $\{e_i\}_{i=1}^n$ of $\mathbb{R}^n$. A *Riemannian metric* on a nonempty open set $\mathcal{U} \subseteq \mathbb{R}^n$ is then defined to be a smooth assignment of scalar products $\langle\cdot,\cdot\rangle_x$ to each $x \in \mathcal{U}$, or, equivalently, a smooth field $H(x)$ of symmetric positive-definite matrices on $\mathcal{U}$.

Given a Riemannian metric on $\mathcal{U}$, the *Riemannian gradient* of a smooth function $\phi\colon \mathcal{U} \to \mathbb{R}$ at $x \in \mathcal{U}$ is defined via the characterization

$$\langle \operatorname{grad}\phi(x), z\rangle_x = \phi'(x; z) \quad \text{for all } z \in \mathbb{R}^n, \tag{2.4}$$

where $\phi'(x;z) = \frac{d}{dt}\big|_{t=0^+} \phi(x+tz)$ denotes the directional derivative of ϕ at x along z. More concretely, by expressing everything in components, it is easy to see that $\operatorname{grad}\phi(x)$ is given by the explicit expression

$$\operatorname{grad}\phi(x) = H(x)^{-1}\nabla\phi(x). \tag{2.5}$$

Bringing the above closer to our setting, let $\mathcal{V}_0 \subseteq \mathbb{R}^n$ be a subspace of $\mathbb{R}^n$ and let $\mathcal{V}$ be an affine translation of $\mathcal{V}_0$ such that $\mathcal{U}_0 \equiv \mathcal{U} \cap \mathcal{V}$ is nonempty. Then, viewing $\mathcal{U}_0$ as an open subset of $\mathcal{V}$, the *gradient of* ϕ *restricted to* $\mathcal{U}_0$ is defined as the unique vector $\operatorname{grad}_{\mathcal{U}_0}\phi(x) \equiv \operatorname{grad}\phi|_{\mathcal{U}_0}(x) \in \mathcal{V}_0$ such that

$$\langle \operatorname{grad}_{\mathcal{U}_0}\phi(x), z\rangle_x = \phi'(x; z) \quad \text{for all } z \in \mathcal{V}_0. \tag{2.6}$$

Hence, specializing all this to the problem at hand, let $H(x)$ be a Riemannian metric on the open orthant $\operatorname{ri}(\mathcal{C}) = \mathbb{R}^n_{++}$ of $\mathbb{R}^n$ and set

$$\begin{aligned} \mathcal{A}_0 &= \ker A = \{x \in \mathbb{R}^n : Ax = 0\},\\ \mathcal{A} &= \{x \in \mathbb{R}^n : Ax = b\}, \end{aligned} \tag{2.7}$$

as in section 1. Then, a straightforward exercise in matrix algebra shows that the gradient of f restricted to $\operatorname{ri}(\mathcal{X}) = \operatorname{ri}(\mathcal{C}) \cap \mathcal{A}$ can be written in closed form as

$$\operatorname{grad}_{\operatorname{ri}(\mathcal{X})} f(x) = P(x)H(x)^{-1}\nabla f(x) \tag{2.8}$$

with $P(x)$ defined as in (1.3), i.e., $P(x) = I - H(x)^{-1}A^\top(AH(x)^{-1}A^\top)^{-1}A$.

To streamline notation for later, we will denote the negative (restricted) gradient of f at $x \in \operatorname{ri}(\mathcal{X})$ as

$$v(x) = -\operatorname{grad}_{\operatorname{ri}(\mathcal{X})} f(x) = -P(x)H(x)^{-1}\nabla f(x). \tag{2.9}$$

Defined this way, $v(x)$ corresponds to the direction of steepest descent of f along $\mathcal{X}$ relative to the metric $H(x)$. In particular, since $v(x) \in \mathcal{A}_0$ for all $x \in C$, it follows that

$$-\nabla f(x)^\top v(x) = \|v(x)\|_x^2, \tag{2.10}$$

where, in obvious notation, we let $\|z\|_x^2 = \langle z, z\rangle_x$ for all $z \in \mathcal{A}_0$.

[2] For a masterful introduction to Riemannian geometry, we refer the reader to [34].

2.3. Hessian–Riemannian metrics. A very important class of Riemannian metrics (and the main focus of our paper) can be generated by taking the Hessian of a smooth convex function. More precisely, we have the following definition.

Definition 2.1. *We say that* $h\colon \mathcal{C} \to \mathbb{R} \cup \{+\infty\}$ *is a* barrier (*or* metric-generating) function *if*

1. h *is twice continuously differentiable on* $\operatorname{ri}(\mathcal{C})$,
2. *the Hessian* $\nabla^2 h$ *of* h *is locally Lipschitz continuous and positive-definite on* $\operatorname{ri}(\mathcal{C})$,
3. $\|\partial_i h(x^k)\|_2 \to \infty$ *for every sequence of interior points* $x^k \in \operatorname{ri}(\mathcal{C})$ *converging to the boundary* $\operatorname{bd}(\mathcal{C})$ *of* $\mathcal{C}$.

If h *is a barrier function as above, the Hessian–Riemannian* (HR) *metric induced by* h *is defined as*

$$H(x) = \nabla^2 h(x) \quad \textit{for all } x \in \operatorname{ri}(\mathcal{C}). \tag{2.11}$$

Remark 1. The systematic study of Hessian–Riemannian metrics dates back at least to Duistermaat [25]. In the context of convex programming, these metrics were popularized by the authors of [1, 2, 13] who introduced the Hessian–Riemannian gradient dynamics (HRGD) discussed in section 1. With regard to terminology, Definition 2.1 essentially follows the setup of [1] with a number of simplifications aimed to take advantage of the specific structure of the nonnegative orthant.

Remark 2. Up to mild differences, the notion of a barrier function essentially coincides with that of a distance generating function (DGF) as used to derive the mirror descent algorithm [40, 41]. A detailed discussion of the connections between Hessian–Riemannian metrics and mirror descent would take us too far afield, so we refer the reader to [1, 8] for a more general treatment.

A systematic way of constructing barrier functions on $\operatorname{ri}(\mathcal{C})$ is to take separable sums of the form

$$h(x) = \sum_{i=1}^{n} \theta_i(x_i), \tag{2.12}$$

where each function $\theta_i\colon (0,\infty) \to \mathbb{R}$ is a barrier function on $(0,\infty) = \mathbb{R}_{++}$ (viewed here as the positive orthant of $\mathbb{R}$). For technical reasons, it will be convenient to assume two further conditions for θ_i, leading to the following definition.

Definition 2.2. *We say that* $\theta\colon (0,\infty) \to \mathbb{R}$ *is a* metric-inducing kernel *if*

(a) θ *is twice continuously differentiable on* $(0,\infty)$, θ'' *is positive and locally Lipschitz continuous on* $(0,\infty)$, *and* $\lim_{t\to 0^+} \theta'(t) = -\infty$;

(b) $\inf_{t>0} \theta''(t) > 0$, *i.e.*, $\theta''(t) \geq \beta$ *for some* $\beta > 0$ *and all* $t \in (0,\infty)$;

(c) $\inf_{t>0} t\theta''(t) > 0$, *i.e.*, $t\theta''(t) \geq \varepsilon$ *for some* $\varepsilon > 0$ *and all* $t \in (0,\infty)$.

Of the above requirements, (a) simply specializes the barrier function requirements of Definition 2.1 to $(0,\infty)$. Requirement (b) strengthens the strict convexity assumption by essentially positing strong convexity over $(0,\infty)$; this assumption can be dropped altogether, but we use it to simplify our arguments later on.[3] Finally, (c)

[3] If $\mathcal{X}$ is compact, it suffices to have $\inf_t \theta''(t) > 0$ on any compact subset of $(0,\infty)$, and this holds trivially by the positivity and continuity of θ''. In the general case, the boundedness requirement of Assumption 1(b) can be used to similar effect because all our analysis takes place in the sublevel set $\{x \in \mathcal{X} : f(x) \leq f(x^0)\}$.

essentially posits that $\theta''(t)$ grows at least as $\mathcal{O}(1/t)$ as $t \searrow 0^+$. This "sufficient growth" requirement plays an important technical role later on in our analysis but is relatively mild otherwise.[4]

For concreteness, we provide some standard examples of kernel functions below.

1. *Regularized Gibbs entropy:* $\theta(t) = \frac{1}{2}\beta t^2 + t\log t$.
2. *Regularized Tsallis entropy:* $\theta(t) = \frac{1}{2}\beta t^2 + \frac{1}{(1-p)(2-p)} t^{2-p}$, $p \in (1,2)$.
3. *Regularized log-barrier (Burg):* $\theta(t) = \frac{1}{2}\beta t^2 - \log t$.

The above examples only provide a snapshot of possible choices; for more examples, see [1, 37]. We should also note that the regularization term $\frac{1}{2}\beta t^2$ is only included to guarantee that $\inf_t \theta''(t) \geq \beta$. As we discussed above, this requirement can be dropped, corresponding to the baseline case $\beta = 0$ (the examples we presented in the introduction were all taken with $\beta = 0$). It is also clear that these functions can be combined to generate mixture functions preserving the defining properties of a metric-inducing kernel. For instance, modulo the regularization term $\frac{1}{2}\beta t^2$, Tseng, Bomze, and Schachinger [50] considered the mixture

$$\theta_\gamma(t) = \frac{1}{2}\beta t^2 + \begin{cases} t\log t - t & \text{if } \gamma = 1/2, \\ \frac{1}{2(1-\gamma)(1-2\gamma)} t^{2(1-\gamma)} & \text{if } \gamma \in (1/2,1), \\ -\log t & \text{if } \gamma = 1, \end{cases} \tag{2.13}$$

which provides a continuous homotopy interpolation of $1/\theta''_\gamma(t)$ between the Gibbs and Burg kernels for $\gamma = 1/2$ and $\gamma = 1$, respectively (the range $0 < \gamma < 1/2$ is not considered here because it violates the steepness requirement $\lim_{t\searrow 0^+} \theta'(t) = -\infty$).

The benefit of using a metric-inducing kernel as above is that the resulting Hessian–Riemannian metric takes the convenient diagonal form

$$H(x) = \operatorname{diag}(\theta''_1(x_1), \ldots, \theta''_n(x_n)), \tag{2.14}$$

which leads to the straightforward expression $H(x)^{-1} = \operatorname{diag}(1/\theta''_1(x_1), \ldots, 1/\theta''(x_n))$. By Definition 2.2(c), the inverse matrix $H(x)^{-1}$ can be extended continuously to the boundary $\operatorname{bd}(\mathcal{C})$ of $\mathcal{C}$ in the obvious way, and its explicit diagonal form greatly facilitates our analysis in the next sections. Unless explicitly mentioned otherwise, all Hessian–Riemannian metrics in what follows will be assumed to come from a kernel function as above; for a more general treatment, see [1].

3. The Hessian barrier algorithm.

Viewed abstractly, the Hessian barrier algorithm can be formulated as a recursive update rule of the general form

$$x^+ = x + \alpha z. \tag{3.1}$$

Specifically, given an input state $x \in \mathcal{X}$, a new state $x^+ \in \mathcal{X}$ is produced by taking a step along the tangent search direction $z \in \mathcal{A}_0$, properly scaled by the step size $\alpha > 0$. In the rest of this section, we discuss in detail the definition of the search direction z and the step size α.

3.1. The search direction.

The algorithm's search direction will be determined, given a Hessian–Riemannian metric $H(x) \equiv \nabla^2 h(x)$ on $\operatorname{ri}(\mathcal{X})$, by solving a

[4] Coupled with the requirement $\lim_{t\to 0^+} \theta'(t) = -\infty$, the growth condition (c) only fails for fringe examples such as $\theta''(t) = 1/(t\log t)$ and the like.

quadratic optimization problem of the form

$$\begin{aligned} &\text{minimize} && \nabla f(x)^\top z + \frac{1}{2}\|z\|_x^2 \\ &\text{subject to} && Az = 0, \end{aligned} \tag{3.2}$$

with the norm $\|\cdot\|_x$ prescribed by some Hessian–Riemannian metric on $\operatorname{ri}(\mathcal{C})$ as in the previous section. Heuristically, the linear term $\nabla f(x)^\top z \equiv f'(x; z)$ simply captures the corresponding first-order change in the value of f along z; analogously, the quadratic term in (3.2) can be interpreted as a "cost of motion" along z. As such, (3.2) identifies the direction of steepest descent modulo the cost of taking said step.[5]

From an algebraic standpoint, a standard calculation shows that the solution of (3.2) is simply the (negative) Hessian–Riemannian gradient of f at x, i.e., it is equal to

$$v(x) \equiv -\operatorname{grad}_{\mathcal{X}} f(x) = -P(x)H(x)^{-1}\nabla f(x). \tag{3.3}$$

Perhaps more intuitively, this search direction also coincides with the solution of the trust-region problem

$$\begin{aligned} &\text{minimize} && \nabla f(x)^\top z \\ &\text{subject to} && Az = 0,\ \|z\|_x \leq r \end{aligned} \tag{3.4}$$

when $r > 0$ is large enough.[6] The above shows that a search vector chosen in this way maximizes the first-order decrease in the value of f over all vectors with bounded norm. In turn, this exhibits the close connection of Hessian–Riemannian descent methods to interior-point trust-region methods as in [18, 31]; we will return to this point later.

We close this section with the straightforward observation that the zeros of the search direction $v(x)$ correspond precisely to the critical points of (Opt).

LEMMA 3.1. *For all $x \in \operatorname{ri}(\mathcal{X})$, we have $v(x) = 0$ if and only if $\nabla f(x) \in \mathcal{A}_0^\perp \equiv \operatorname{im}(A^\top)$.*

The proof of Lemma 3.1 is an elementary consequence of the definition of $v(x)$, so we omit it. We only mention this result here to highlight the fact that the update rule (3.1) with search direction $v(x)$ remains stationary if the input state x is a zero of $v(x)$. In what follows, we use this fact freely without referring to it explicitly.

3.2. The method's step size. The main challenge in setting the method's step size is twofold: (a) we need to guarantee that x^+ is feasible for all input states $x \in \operatorname{ri}(\mathcal{X})$, and (b) the method should exhibit "sufficient decrease" in the sense that $f(x^+)$ is sufficiently smaller than $f(x)$ at each step.

We begin with the issue of feasibility. To this end, adopting terminology which is common in the affine scaling literature, consider the "dual variable"

$$y(x) = (AH(x)^{-1}A^\top)^{-1}AH(x)^{-1}\nabla f(x) \tag{3.5}$$

and the "reduced cost"

$$r(x) = \nabla f(x) - A^\top y(x) = -H(x)v(x). \tag{3.6}$$

[5] For a game-theoretic analogue of this idea, see [37].

[6] In particular, it suffices to take r equal to the minimum value of (3.2).

Since the Hessian $H(x)$ is diagonal by construction, we can use the reduced cost vector $r(x)$ to rewrite the update rule (3.1) in components as

$$x_i^+ = x_i - \alpha(x)\frac{r_i(x)}{\theta_i''(x_i)} = x_i\left(1 - \frac{\alpha(x)r_i(x)}{x_i\theta_i''(x_i)}\right). \tag{3.7}$$

Consequently, we will have $x_i^+ > 0$ if either $r_i(x) \leq 0$ or else

$$\alpha(x) < \frac{x_i\theta_i''(x_i)}{r_i(x)}. \tag{3.8}$$

Hence, to guarantee feasibility, it suffices to take $\alpha(x) < \alpha_0(x)$, where

$$\alpha_0(x) = \min_{i=1,\dots,n}\{x_i\theta_i''(x_i)/r_i(x) : r_i(x) > 0\}, \tag{3.9}$$

with the usual convention $\min \varnothing = \infty$.

Now, to decrease the value of the objective function at each step of the algorithm, our starting point will be the well-known descent inequality [45]

$$f(x') - f(x) \leq \nabla f(x)^\top (x' - x) + \frac{L}{2}\|x' - x\|_2^2, \tag{3.10}$$

which holds for all $x, x' \in \mathcal{X}$. Then, taking $x' = x + \lambda v(x)$ in (3.10) and using the angle relation (2.10), we get

$$\begin{aligned} f(x + \lambda v(x)) - f(x) &\leq -\lambda\|v(x)\|_x^2 + \frac{1}{2}\lambda^2 L\|v(x)\|_2^2 \\ &\leq -\beta\lambda\|v(x)\|_2^2 + \frac{1}{2}\lambda^2 L\|v(x)\|_2^2 \\ &= -\beta\lambda\left(1 - \frac{\lambda L}{2\beta}\right)\|v(x)\|_2^2, \end{aligned} \tag{3.11}$$

where, in the second line, we used the fact that $\|z\|_x^2 = z^\top H(x) z \geq \beta z^\top z = \beta\|z\|_2^2$.

In view of the above, feasibility and descent are both guaranteed as long as the step size $\alpha(x)$ of the method at the point $x \in \mathcal{X}$ is less than $\min\{\alpha_0(x), 2\beta/L\}$. To proceed, we will further employ an Armijo backtracking procedure to guarantee *sufficient decrease*, i.e., that

$$f(x^+) \leq f(x) - \mu \cdot \alpha(x)\|v(x)\|_x^2 \tag{3.12}$$

for some $\mu \in (0, 1)$. To achieve this, we bootstrap the process with the step size

$$\underline{\alpha}(x) = \min\{\alpha_0(x), 2\beta/L\}. \tag{3.13}$$

If (3.12) is satisfied with $\alpha(x) = \underline{\alpha}(x)$, we will accept the iterate x^+ generated from (3.1); otherwise, we shrink the step size $\underline{\alpha}(x)$ by a factor of $\delta \in (0, 1)$, and we keep backtracking until (3.12) is satisfied.[7] Formally, this means that the step size of the method will be of the form $\alpha(x) = \delta^\ell \underline{\alpha}(x)$, where $\ell \geq 0$ is the first nonnegative integer such that

$$f(x + \delta^\ell \underline{\alpha}(x) v(x)) - f(x) \leq -\mu\delta^\ell \underline{\alpha}(x)\|v(x)\|_x^2. \tag{3.14}$$

Lemma 4.5 in the next section shows that this backtracking process terminates after a finite number of steps. In this way, we obtain a well-defined step-size policy which simultaneously guarantees feasibility and sufficient decrease.

[7] In practice, μ is chosen very small (around 10^{-4}), while typical values for δ lie in the range between 0.1 and 0.5 [44].

Algorithm 1 Hessian barrier algorithm (HBA).

```
Require: sufficient decrease factor $\mu \in (0,1)$, shrink factor $\delta \in (0,1)$
initialize $x \in \mathcal{X}$                                  # initialization
while stopping criterion not satisfied do
    $v \leftarrow -\operatorname{grad}_{\mathcal{X}} f(x)$                     # search direction
    $\alpha \leftarrow \min\{\alpha_0(x), 2\beta/L\}$                        # set step-size
    $x^+ \leftarrow x + \alpha v$                                  # set test point
    while $f(x^+) > f(x) - \mu\alpha\|v\|_x^2$ do                 # suff. decrease?
        $\alpha \leftarrow \delta\alpha$                                  # shrink step-size
        $x^+ \leftarrow x + \alpha v$                              # update test point
    end while
    $x \leftarrow x^+$                                             # new state
end while
return $x$
```

3.3. The Hessian barrier algorithm. Combining all of the above, the *Hessian barrier algorithm* (HBA) can be stated in recursive form as

$$x^{k+1} = x^k - \alpha(x^k)P(x^k)H(x^k)^{-1}\nabla f(x^k), \tag{HBA}$$

where

1. $k = 0, 1, \ldots$ is the algorithm's iteration counter;
2. x^k denotes the state of the algorithm at step k; the algorithm is initialized at a point x^0 satisfying Assumption 1(b);
3. $\alpha(x)$ is the algorithm's step size at state x, defined implicitly via the Armijo backtracking process described in the previous section;
4. $P(x)$ and $H(x)$ are determined by a Hessian–Riemannian metric chosen by the optimizer (cf. subsection 2.3).

For a pseudocode implementation of (HBA), see Algorithm 1.

Importantly, even though (HBA) looks similar to the interior gradient methods of [7, 8], the actual update steps performed are fundamentally different. Specifically, the gradient method of Auslender and Teboulle [8] performs at each iteration a prox-step using a Bregman function to ensure that the algorithm's iterates remain in the problem's feasible region—recall the definition of (MD). This approach implicitly assumes that the problem's constraint set is sufficiently "simple" for the Bregman proximal step to be performed in a computationally efficient way; (HBA) does not require a prox-step, so it is more lightweight in that respect.

4. Global convergence analysis.

To present our convergence analysis, two more definitions are required. Specifically, if x^k, $k = 0, 1, \ldots,$ is the sequence of iterates generated by (HBA), we write

$$\mathcal{L} = \{\hat{x} \in \mathcal{X} : \text{some subsequence } x^{k_r} \text{ of } x^k \text{ converges to } \hat{x}\} \tag{4.1}$$

for the set of limit points of the algorithm, and we let

$$\Lambda = \Big\{\hat{x} \in \mathcal{X} : \lim_{k\to\infty} f(x^k) = f(\hat{x}) \text{ and } \operatorname{diag}(\hat{x})r(\hat{x}) = 0\Big\}. \tag{4.2}$$

Our main convergence result may then be stated as follows.

THEOREM 4.1. *With notation as above, we have the following.*

(a) *The sequence x^k is bounded and $f(x^k)$ is nonincreasing.*

(b) *Every point $x^* \in \mathcal{L}$ satisfies complementarity in the sense that $r_i(x^*) = 0$ whenever $x_i^* > 0$. In particular, $\mathcal{L} \subseteq \Lambda$, so $f(x^k)$ converges.*

(c) *Every limit point of x^k is a KKT point of f, provided one of the following conditions holds:*

  (1) *f is convex; in this case x^k converges to* $\arg\min f$.

  (2) *Λ consists of isolated points.*

  (3) *Every point in Λ satisfies strict complementarity, i.e., $x_i + r_i(x) > 0$ for all $i \in S = \{1, \ldots, n\}$.*

Theorem 4.1 can be seen as the bona fide, algorithmic analogue of the continuous-time analysis of Alvarez, Bolte, and Brahic [1] of Hessian–Riemannian gradient flows. To the best of our knowledge, the closest result of this type in the literature is the convergence analysis of Tseng, Bomze, and Schachinger [50] for a replicator-type descent algorithm applied to quadratic programs in standard form. However, the results of [50] rely crucially on the assumption that the algorithm's step size does not become vanishingly small in the limit: this assumption is a major obstacle to the applicability of the analysis of [50], as there is no way to verify it from the problem's primitives. Dropping this assumption requires a delicate—and intricate—argument which takes up the first part of the remainder of this section.

4.1. Step-size analysis.

As stated above, our main goal in what follows is to show that the algorithm's step-size sequence $\alpha^k \equiv \alpha(x^k)$ is bounded away from zero. We begin with a trivial upper bound which we state only for completeness.

LEMMA 4.2. *The step-size sequence $\alpha^k \equiv \alpha(x^k)$ of* (HBA) *satisfies* $\sup_k \alpha^k < \infty$.

To get a lower bound for the algorithm's step size, we begin by showing that the "bootstrap" step size $\underline{\alpha}(x)$ of (3.13) is itself bounded away from zero. In the context of affine scaling algorithms for linear programming, similar results have been proven in the special case where the Riemannian geometry is generated by the log-barrier kernel (the Burg entropy); see [33] for an early result in this direction.[8] This kernel gives rise to very convenient closed-form expressions that greatly simplify the calculations; however, for the general framework considered here, we need a fairly intricate analysis that cannot be handled by the derivations of [33]. We present the relevant calculations below.

LEMMA 4.3. *We have* $\inf\{\underline{\alpha}(x) : x \in \operatorname{ri}(\mathcal{X}),\ f(x) \leq f(x^0)\} > 0$.

Proof. Since $\underline{\alpha}(x) = \min\{\alpha_0(x), 2\beta/L\}$, it suffices to show that

$$\inf\{\alpha_0(x) : x \in \operatorname{ri}(\mathcal{X}),\ f(x) \leq f(x^0)\} > 0.$$

In turn, by the definition (3.9) of the function $\alpha_0(x)$, it suffices to look at points x for which $r_i(x) > 0$ for some $i = 1, \ldots, n$. We thus have to bound the quantity $x_i \theta_i''(x_i)/r_i(x)$ away from zero, which, by Definition 2.2, boils down to showing that $r_i(x)$ is bounded from above.

Since $r(x) = \nabla f(x) - A^\top y(x)$, this is achieved once we have an upper bound for the "dual variable" $y(x) = (AH(x)^{-1}A^\top)^{-1}AH(x)^{-1}\nabla f(x)$ defined in (3.5). To achieve this, define the matrix $M_x = AH(x)^{-1/2} \in \mathbb{R}^{m \times n}$, so that $y(x)$ is the unique solution to the linear system

$$M_x M_x^\top y = M_x H(x)^{-1/2} \nabla f(x). \tag{4.3}$$

[8] We thank an anonymous referee for mentioning this reference to us.

By Cramer's rule, we can explicitly compute the ith coordinate of the vector $y(x)$ via the formula

$$y_i(x) = \frac{\det\left((M_xM_x^\top)^1, \dots, M_xH(x)^{-1/2}\nabla f(x), \dots, (M_xM_x^\top)^m\right)}{\det(M_xM_x^\top)}. \tag{4.4}$$

This can be simplified by some straightforward, albeit tedious, algebraic manipulations. Indeed, for a matrix $A \in \mathbb{R}^{m\times n}$ let

$$A^{k_1,\dots,k_m} = (a^{k_1}, \dots, a^{k_m}), \qquad 1 \le k_1 < k_2 < \dots < k_m \le n, \tag{4.5}$$

denote the $m \times m$ matrix obtained from the columns $a^{k_1}, \dots, a^{k_m}$ of A. By the Cauchy–Binet formula, we can compute the denominator as

$$\det(M_xM_x^\top) = \sum_{1\le k_1<\dots<k_m\le n} \det\left(M_x^{k_1,\dots,k_m}\right)^2. \tag{4.6}$$

Since the Hessian matrix $H(x)$ is diagonal, it is immediate that

$$M_x = \left[\theta_1''(x_1)^{-1/2}a^1, \dots, \theta_n''(x_n)^{-1/2}a^n\right], \tag{4.7}$$

implying in turn that M_x can be extended continuously to the entire orthant $\mathbb{R}^n_+$ via the convention $1/\infty = 0$. We thus get

$$\det\left(M_xM_x^\top\right) = \sum_{1\le k_1<\dots<k_m\le n} \frac{1}{\theta_{k_1}''(x_{k_1})\cdots\theta_{k_m}''(x_{k_m})} \det(A^{k_1,\dots,k_m})^2. \tag{4.8}$$

In a similar fashion we can express the numerator as the determinant of a matrix product between the matrices

$$A_x = [a^1/\theta_1''(x_1), \dots, a^n/\theta_n''(x_n)] \quad \text{and} \quad B_{x,i}^\top = [a_1, \dots, \nabla f(x), \dots, a_m].$$

Then, applying the Cauchy–Binet formula again, we obtain

$$y_i(x) = \frac{\sum_{1\le k_1<\dots<k_m\le n} \theta_{k_1}''(x_{k_1})^{-1}\cdots\theta_{k_m}''(x_{k_m})^{-1}\det(A^{k_1,\dots,k_m})\det(B_{x,i}^{k_1,\dots,k_m})}{\sum_{1\le k_1<\dots<k_m\le n} \theta_{k_1}''(x_{k_1})^{-1}\cdots\theta_{k_m}''(x_{k_m})^{-1}\det(A^{k_1,\dots,k_m})^2}. \tag{4.9}$$

Since A has full rank, the above is well defined.

To establish an upper bound for this last expression, we use the simple inequality

$$\frac{|\sum_{i=1}^n b_i|}{\sum_{i=1}^n \sigma_i} \le \max_i \frac{|b_i|}{\sigma_i} \tag{4.10}$$

for $\sigma_i > 0$. We then get

$$|y_i(x)| \le \max\left|\frac{\det(B_{x,i}^{k_1,\dots,k_m})}{\det(A^{k_1,\dots,k_m})}\right| =: \omega_i(x), \tag{4.11}$$

where the maximum is taken over all tuples $1 \le k_1 < \dots < k_m \le n$ for which the denominator in the above expression does not vanish (which, again, is possible thanks to A being full rank). By Assumption 1(b), we have

$$K_0 \equiv \sup\{\|\nabla f\|_\infty : x \in \mathrm{ri}(\mathcal{X}),\ f(x) \le f(x^0)\} < \infty$$

so $\omega_i(x)$ is bounded in norm for all i and all $x \in \mathcal{X}$ and

$$\|r(x)\|_\infty \leq \|\nabla f(x)\|_\infty + \|A^\top y(x)\|_\infty \leq K_0 + \|A\|_* \|\omega(x)\|_\infty, \tag{4.12}$$

where $\|A\|_* = \max_{1\leq i \leq n} |\sum_{j=1}^m a_{ji}|$. This gives

$$\alpha_0(x) \geq \varepsilon/(K_0 + \|A\|_* \|\omega(x)\|) \geq \varepsilon/K_0,$$

i.e., $\inf\{\alpha_0(x) : x \in \mathrm{ri}(\mathcal{X}),\ f(x) \leq f(x^0)\} > 0$, as claimed. □

Our next result shows that the Armijo step-size rule (3.14) terminates after finitely many iterations.

LEMMA 4.4. *Suppose that $v(x) \neq 0$, i.e., x is not a KKT point of f. Then*
1. *the process* (3.14) *is well defined at x,*
2. *$\alpha(x) \geq \min\{2(1-\mu)\beta\delta/L, \underline{\alpha}(x)\}$.*

Our proof builds on a classical line of reasoning as in [8], but the algorithm's non-Euclidean nature necessitates some extra care.

Proof of Lemma 4.4. Suppose that the Armijo backtracking process carries on without terminating at $x \in \mathrm{ri}(\mathcal{X})$. Then, setting $x^+(\lambda) = x + \lambda v(x)$ for all $\lambda > 0$, and writing $\alpha \equiv \alpha(x)$ and $\underline{\alpha} \equiv \underline{\alpha}(x) = \min\{\alpha_0(x), 2\beta/L\}$ for concision, we get

$$f(x^+(\delta^\ell \underline{\alpha})) - f(x) > \mu \nabla f(x)^\top (x^+(\delta^\ell \underline{\alpha}) - x) \tag{4.13}$$

for all $\ell \in \mathbb{N}$. Then, by the mean value theorem, there exists $\xi^\ell \in (x, x^+(\delta^\ell \underline{\alpha}))$ such that

$$\nabla f(\xi^\ell_j)^\top (x^+(\delta^\ell \underline{\alpha}) - x) = f(x^+(\delta^\ell \underline{\alpha})) - f(x) > \mu \nabla f(x)^\top (x^+(\delta^\ell \underline{\alpha}) - x). \tag{4.14}$$

Clearly, we also have $\xi^\ell \to x$ as $\ell \to \infty$. Hence, passing to the limit and recalling that $\mu \in (0,1)$, we get

$$-\|v(x)\|_x^2 \geq -c\|v(x)\|_x^2 \iff v(x) = 0 \iff \nabla f(x) \in \mathcal{A}_0^\perp, \tag{4.15}$$

a contradiction.

For our second claim, suppose that the Armijo criterion (3.14) is first satisfied at x after $\ell \geq 1$ steps, i.e., $\alpha/\delta = \delta^{\ell-1}\underline{\alpha}$. By assumption, this means that we don't yet have sufficient decrease at the $(\ell-1)$th step of the backtracking process, i.e.,

$$f(x^+(\alpha/\delta)) - f(x) > \mu \nabla f(x)^\top (x^+(\alpha/\delta) - x). \tag{4.16}$$

Since ∇f is L-Lipschitz continuous relative to $\|\cdot\|_2$, the descent inequality (3.10) for an arbitrary step size $\lambda > 0$ becomes

$$f(x^+(\lambda)) - f(x) \leq -\lambda \|v(x)\|_x^2 + \frac{\lambda^2 L}{2}\|v(x)\|_2^2. \tag{4.17}$$

Thus, since $\|z\|_x^2 = z^\top H(x) z \geq \lambda_{\min}(H(z))\|z\|_2^2 \geq \beta \|z\|_2^2$ for all $z \in \mathbb{R}^n$, we get

$$\begin{aligned} f(x^+(\lambda)) - f(x) &\leq -\lambda \|v(x)\|_x^2 + \frac{\lambda^2 L}{2\beta}\|v(x)\|_x^2 = -\lambda\left(1 - \frac{\lambda L}{2\beta}\right)\|v(x)\|_x^2 \\ &= \left(1 - \frac{\lambda L}{2\beta}\right)\nabla f(x)^\top (x^+(\lambda) - x), \end{aligned} \tag{4.18}$$

where we used the angle condition (2.10) and the definition of $x^+(\lambda)$. Hence, setting $\lambda = \alpha/\delta$, we get

$$f(x^+(\alpha/\delta)) - f(x) \le \left(1 - \frac{L\alpha}{2\beta\delta}\right) \nabla f(x)^\top (x^+(\alpha/\delta) - x), \tag{4.19}$$

which, combined with (4.16), implies that $1 - \alpha L/(2\beta\delta) \le \mu$, i.e., $\alpha \ge 2\beta\delta(1-\mu)/L$.

On the other hand, if the Armijo criterion (3.14) is already satisfied at x with step size $\underline{\alpha}$ (i.e., after $\ell = 0$ shrinkage steps), we will have $\alpha = \underline{\alpha}$. Thus, combining all of the above, we get $\alpha \ge \min\{\underline{\alpha}, 2(1-\mu)\beta\delta/L\}$, as claimed. □

We are finally in a position to show that the algorithm's step size is nonvanishing in the limit.

Lemma 4.5. *The algorithm's step-size sequence* $\alpha^k \equiv \alpha(x^k)$ *has* $\inf_k \alpha^k > 0$.

Proof of Lemma 4.5. Since $f(x^k)$ is weakly decreasing (by the Armijo rule (3.14)), it follows that $x^k \in [f \le f(x^0)]$ for all k. Lemma 4.3 further guarantees that

$$\inf\{\underline{\alpha}(x) : \mathcal{X} \in \mathrm{ri}(\mathcal{X}),\ f(x) \le f(x^0)\} > 0,$$

so our claim follows from Lemma 4.4. □

4.2. Iterate analysis. We now turn to the long-run behavior of the iterates x^k generated by (HBA). The arguments are partly based on general facts on descent methods and extend the analysis of [50] to a considerably richer algorithmic framework. We start with a simple observation.

Lemma 4.6. *Let* $x^0 \in \mathrm{ri}(\mathcal{X})$ *be an initial condition satisfying Assumption* 1(b). *Then the sequence of iterates* x^k *of* (HBA) *is bounded.*

Proof. By the definition of (HBA), we have

$$f(x^{k+1}) \le f(x^k) - \mu\alpha^k \|v(x^k)\|^2_{x^k}, \tag{4.20}$$

showing that $f(x^k)$ is nonincreasing. Our claim then follows trivially. □

The next result is actually a standard result for descent methods—see, e.g., [3].

Lemma 4.7. *With notation as in Theorem* 4.1*, we have that*
(a) *the limit set* $\mathcal{L}$ *of* (HBA) *is nonempty, compact, and connected;*
(b) $\lim_{k\to\infty} \mathrm{dist}(x^k, \mathcal{L}) = 0$*;*
(c) *the objective function* f *is constant on* $\mathcal{L}$.

With this lemma at hand, we proceed to show that the iterate change vanishes.

Lemma 4.8. *With notation as in Theorem* 4.1*, we have* $\lim_{k\to\infty}(x^{k+1} - x^k) = 0$.

Proof. Observe that for all $k = 0, 1, \dots,$ we have

$$\|v(x^k)\|^2_{x^k} = \frac{1}{(\alpha^k)^2} \|H(x^k)^{1/2}(x^{k+1} - x^k)\|^2 \ge \frac{\beta}{(\underline{\alpha}^k)^2} \|x^{k+1} - x^k\|^2. \tag{4.21}$$

Choose a convergent subsequence $\{x^k\}_{k\in\mathcal{K}}$, so that $\lim_{k\to\infty, k\in\mathcal{K}} x^k = x^*$. Since $f(x^k)$ is nonincreasing, we readily get $f(x^k) \downarrow f(x^*) \le f(x^0)$, and also

$$\lim_{k\to\infty, k\in\mathcal{K}} [f(x^{k+1}) - f(x^k)] = 0.$$

Then, from (3.14), it follows that $\mu\alpha^k\|v(x^k)\|^2_{x^k} \leq f(x^k) - f(x^{k+1})$ and hence, $\lim_{k\to\infty,k\in\mathcal{K}} \alpha^k\|v(x^k)\|^2_{x^k} = 0$. We thus get

$$\limsup_{k\to\infty} \alpha^k\|v(x^k)\|^2_{x^k} = \liminf_{k\to\infty} \alpha^k\|v(x^k)\|^2_{x^k} = 0.$$

In turn, Lemma 4.4 implies that $\inf_{k\in\mathbb{N}} \alpha^k > 0$, so $\lim_{k\to\infty}\|v(x^k)\|_{x^k} = 0$. ∎

LEMMA 4.9. $\mathcal{L} \subset \Lambda$.

Proof. Let $\{x^k\}_{k\in\mathbb{N}}$ be a convergent subsequence (we omit the relabeling). Since $v(x) = -H(x)^{-1}r(x)$, we conclude from the above that

$$0 = \lim_{k\to\infty}\langle H(x^k)v(x^k), v(x^k)\rangle = \lim_{k\to\infty}\|H(x^k)^{-1/2}r(x^k)\|^2. \tag{4.22}$$

Therefore, for all $i \in S$, we will have $\lim_{k\to\infty}|r_i(x^k)\theta_i''(x_i^k)^{-1/2}| = 0$. Hence, if $i \in S_{x^*}$, we must have $\lim_{k\to\infty} r_i(x^k) = r_i(x^*) = 0$. Now, for all k, the linear system

$$(\nabla f(x^k) - A^\top y)_i = r_i(x^k), \quad i \in S_{x^*}, \tag{4.23}$$

admits the solution $y^k = y(x^k) \in \mathbb{R}^m$. Set $y^* = y(x^*) = \lim_{k\to\infty} y(x^k)$, by continuity. Hence, passing to the limit in (4.23) gives $(\nabla f(x^*) - A^\top y^*)_i = 0$ for all $i \in S_{x^*}$. We thus conclude that $\operatorname{diag}(x^*)(\nabla f(x^*) - A^\top y^*) = 0$, i.e., $x^* \in \Lambda$. ∎

4.3. Proof of Theorem 4.1.

We now combine all the above established preliminary facts to prove the main results on the global convergence of (HBA). Parts (a) and (b) of Theorem 4.1 follow from Lemmas 4.6 and 4.9. The remainder of this section is concerned with establishing claims (1)–(3) of Theorem 4.1. For this we have to show that $r(x^*) \geq 0$ for all $x^* \in \mathcal{L}$ holds under each of the conditions described in Theorem 4.1. The fact that x^* is a KKT point is then a consequence of Lemma 4.9, showing also that complementarity slackness holds.

Proof of Theorem 4.1(1). Assume that f is convex. Let $x^* \in \mathcal{L}$, and define $\bar{J} = \{i \in S : r_i(x^*) = 0\}$ and $\bar{J}^c = \{i \in S : r_i(x^*) \neq 0\}$. Moreover, define the set

$$\Omega = \arg\min\{f(x) : x \in \mathcal{X},\ x_{\bar{J}^c} = 0\}. \tag{4.24}$$

Since f is continuous and convex, the set Ω is closed and convex. x^* is a feasible point for the convex program (4.24), satisfying the KKT condition $\operatorname{diag}(x^*)r(x^*) = 0$ (Lemma 4.9). Hence, $\Omega = \{x \in \mathcal{X} : f(x) = f(x^*),\ x_{\bar{J}^c} = 0\}$, and therefore f is constant on Ω. By convexity, $\nabla f(x) = \nabla f(x^*)$ for all $x \in \Omega$. We next prove that the reduced cost $r(x)$ is constant on Ω, and in fact must be nonnegative, showing that $x^* \in \mathcal{X}^*$.

LEMMA 4.10. *For all $x \in \Omega$ we have $r(x) = r(x^*)$.*

Proof. Let $x \in \Lambda$ be arbitrary. We have

$$\begin{aligned} r(x) &= \nabla f(x) - A^\top y(x) \\ &= \nabla f(x^*) - A^\top(AH(x)^{-1}A^\top)^{-1}AH(x)^{-1}\nabla f(x) \\ &= [I - A^\top(AH(x)^{-1}A^\top)^{-1}AH(x)^{-1}]\nabla f(x^*) \\ &= [I - A^\top(AH(x)^{-1}A^\top)^{-1}AH(x)^{-1}](r(x^*) + A^\top y(x^*)) \\ &= r(x^*) - A^\top(AH(x)^{-1}A^\top)^{-1}AH(x)^{-1}r(x^*) \\ &= r(x^*). \end{aligned} \tag{4.25}$$

The first line is the definition of $r(x)$, the second line is the definition of $y(x)$ and uses the constancy of the gradient mapping on Λ. The third line is then again the definition of $r(x^*)$. In the last line we have used the fact that $H(x)^{-1}r(x^*) = (r_j(x^*)/\theta''(x_j^*))_{j\in S} = 0$, which holds because if $i \in \bar{J}^c$ then $1/\theta_i''(x_i^*) = 0$, and the dual variable is bounded. ∎

We next prove that all accumulation points of (HBA) are contained in Ω. To that end, for fixed $\eta > 0$, we define

$$\Omega_\eta = \mathcal{L} \cap \{x \in \mathbb{R}^n : \mathrm{dist}(x, \Omega) < \eta\}. \tag{4.26}$$

Observe that this set is nonempty since $x^* \in \Omega$. We will use this set to localize the limit points of the trajectory $\{x^k\}_{k\in\mathbb{N}}$.

LEMMA 4.11. *If $\hat{x} \in \mathcal{L}$ then $\hat{x} \in \Omega$ or $\hat{x} \notin \Omega_\eta$.*

Proof. The proof follows via an argument by contradiction. Assume that $\hat{x} \notin \Omega$ and $\hat{x} \in \Omega_\eta$. Therefore, there exists a point $\tilde{x} \in \Omega$ such that $\|\hat{x} - \tilde{x}\| < \eta$. Since $f(x^*) = f(\hat{x})$ (Lemma 4.7), there must exist $j \in \bar{J}^c$ such that $\hat{x}_j > 0$. $r : \mathcal{X} \to \mathbb{R}^n$ is continuous and bounded. $[-\infty < f \le f(x^0)]$ is compact by assumption. Hence, r is uniformly continuous on $[-\infty < f \le f(x^0)]$, guaranteeing the existence of a scalar $\eta > 0$ such that

$$\|r(x) - r(z)\| \le \min_{i\in\bar{J}^c} r_i(x^*)/2 \tag{4.27}$$

whenever $f(x), f(z) \le f(x^0)$ and $\|x - z\| \le \eta$. In particular, the uniform continuity of the dual variable guarantees that

$$|r_j(\tilde{x}) - r_j(\hat{x})| \le |r_j(x^*)|/2 \tag{4.28}$$

for some $j \in \bar{J}^c$. Since $r_j(\tilde{x}) = r_j(x^*)$ by Lemma 4.10, this implies that $r_j(\hat{x}) \ge |r_j(x^*)|/2 > 0$. Hence $\mathrm{diag}(\hat{x})r(\hat{x}) \neq 0$, contradicting the conclusion $\hat{x} \in \mathcal{L} \subset \Lambda$ of Lemma 4.9. ∎

LEMMA 4.12. $\mathcal{L} \subseteq \Omega$.

Proof. Assume there exists an accumulation point $\hat{x} \notin \Omega$. From Lemma 4.11, we deduce that $\hat{x} \notin \Omega_\eta$. Since the limit set is connected, it follows that the sequence $\{x^k\}_{k\in\mathbb{N}}$ must have accumulation points in $\Omega_\eta\backslash\Omega$. Hence, there exists $\tilde{x} \in \mathcal{L}\cap(\Omega_\eta\backslash\Omega)$. In particular, $\tilde{x} \in \mathcal{L}$, so that $f(\tilde{x}) = f(x^*)$. Furthermore, $\tilde{x} \notin \Omega$, so there exists $j \in \bar{J}^c$ such that $\tilde{x}_j > 0$. From this we derive the same contradiction as in Lemma 4.11. ∎

This shows that for every converging subsequence x^{k_q}, we have $\lim_{q\to\infty} r(x^{k_q}) = r(x^*)$. Suppose now that $r_j(x^*) \equiv \bar{r}_j < 0$ for some $j \in S$. Then, by the complementarity condition $\mathrm{diag}(x^*)r(x^*) = 0$, we have $j \in \bar{J}^c$. By continuity, we know that there exists a $\kappa \in \mathbb{N}$ such that $r_j(x^k) < 0$ for all k far along the subsequence, say all $k \ge \kappa$. Therefore, for all $k \ge \kappa$ we conclude that

$$x_j^{k+1} = x_j^k - \alpha^k r_j(x^k)/\theta_j''(x_j^k) > x_j^k. \tag{4.29}$$

By induction, we conclude that $x_j^k > x_j^\kappa \ge 0$ for all $k \ge \kappa$, a contradiction. Theorem 4.1(1) now follows from the KKT conditions (2.2a) and (2.2b). ∎

Proof of Theorem 4.1(2). We know that $\mathcal{L}$ is a connected set. From Lemma 4.9, we know that $\mathcal{L} \subset \Lambda$. Since the iterate change goes to zero (Lemma 4.8), this implies that the entire sequence converges. Hence, $\mathcal{L} = \{x^*\} \in \mathcal{X}$, with x^* depending only on the initial condition. Since $\operatorname{diag}(x^*)r(x^*) = 0$ by complementarity, the same contradiction argument used in the previous paragraph rules out the possibility that $r_i(x^*) < 0$ for some $i \in \bar{J}^c$. Hence, x^* is a KKT point and our claim follows. □

Proof of Theorem 4.1(3). Let $x^* \in \mathcal{L}$ and let $\bar{J}_0 = \{i \in S : r_i(x^*) = 0\}$, $\bar{J}_+ = \{i \in S : r_i(x^*) > 0\}$, $\bar{J}_- = \{i \in S : r_i(x^*) < 0\}$. Now, define the set

$$\bar{\Lambda} = \{x \in \Lambda : r_{\bar{J}_0}(x) = 0,\ r_{\bar{J}_+}(x) > 0,\ r_{\bar{J}_-}(x) < 0\} \tag{4.30}$$

and let $\mathcal{B} = \{x \in \mathbb{R}^n : \|x\| \leq 1\}$ be the unit ball in $\mathbb{R}^n$. By the primal nondegeneracy assumption and strict complementarity, $\bar{\Lambda}$ is isolated from the rest of Λ. Hence, there exists $\delta > 0$ such that $(\bar{\Lambda} + \delta\mathcal{B}) \cap \Lambda = \bar{\Lambda}$. Since $\mathcal{L}$ is connected and contained in Λ, we conclude that $\mathcal{L} \cap (\bar{\Lambda} + \delta\mathcal{B}) \subseteq \Lambda \cap (\bar{\Lambda} + \delta\mathcal{B}) = \bar{\Lambda}$. Hence, for every $j \in \bar{J}_-$ we have $r_j(x^k) < 0$ for all k sufficiently large. Repeating the argument we used to prove part (1) of the theorem, we again arrive at a contradiction. We conclude that $\bar{J}_- = \emptyset$, i.e., $r(x^*) \geq 0$. □

5. Convergence rate.

In this section, we establish an estimate of the value convergence rate of (HBA) in the special case where f is quadratic, i.e.,

$$f(x) = \frac{1}{2}x^\top Qx + c^\top x \tag{5.1}$$

for some symmetric $Q \in \mathcal{S}^n$ and $c \in \mathbb{R}^n$. When Q is the zero matrix, we recover a linear programming problem. In the rest of this section, we will focus on the challenging case where Q has at least one negative eigenvalue, in which case (Opt) is NP-complete [52].

Our proof establishes sublinear convergence of the sequence $f(x^k)$ to a KKT point. This result generalizes and extends previous work of Tseng [49] and Tseng, Bomze, and Schachinger [50]. Our results are based on techniques developed in [50]; however, the introduction of a Riemannian metric necessitates a series of intricate estimates in order to establish a rate of convergence. Specifically, our analysis requires some mild additional control on the metric-inducing kernels close to the boundary of the feasible set, which we call *moderate steepness*.

Assumption 2. A kernel function $\theta\colon (0,\infty) \to \mathbb{R}$ is *moderately steep* at 0 if there exist some $\varepsilon_i \in (0,1)$, $\omega \geq 1/2$, and $m, M > 0$ such that

$$\frac{m}{s} \leq \theta''(s) \leq \frac{M}{s^{2\omega}} \quad \text{for all } s \in (0,\varepsilon). \tag{5.2}$$

We verify below that the kernels described in subsection 2.3 satisfy this condition.

1. $\theta(t) = \frac{1}{2}\beta t^2 + t\log t$ for $t \geq 0$. Then $\theta'(t) = \beta + 1/t$, and (5.2) is satisfied with $\omega = 1/2$, $m = 1$, and $M = 1 + \beta\varepsilon$.
2. $\theta(t) = \frac{1}{2}\beta t^2 + \frac{1}{(1-p)(2-p)}t^{2-p}$, $p \in (1,2)$. Then $\theta''(t) = \beta + 1/t^p$, so (5.2) is satisfied with $m = p\varepsilon^{p-1}$, $M = \beta\varepsilon^{2\omega} + p\varepsilon^{p+2(\omega-1)}$, and $\omega = 1$.
3. $\theta(t) = \frac{1}{2}\beta t^2 - \log t$. Then $\theta''(t) = \beta + \frac{1}{t^2}$, and (5.2) is satisfied with $m = \frac{1}{\varepsilon}$ and $M = \beta\varepsilon^{2\omega} + \varepsilon^{2(\omega-1)}$, and $\omega = 1$.

Under the assumption that all the metric-inducing kernels satisfy the moderate steepness property, we are able to obtain the announced sublinear convergence rate of the function value sequence.

THEOREM 5.1. *Assume f is of the form* (5.1) *for some $Q \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}^n$. Suppose that* (HBA) *is run with metric-inducing kernels $\theta_1(x), \dots, \theta_n(x)$, satisfying Assumption 2, and generating the sequence $(x^k)_{k\geq 0}$. Then $f(x^k)$ converges to some $f_\infty \in \mathbb{R}$ and*

$$f(x^k) - f_\infty = \mathcal{O}(k^{-\rho}), \tag{5.3}$$

where $\bar{\omega} = \max\{1, \omega\}$ and $\rho = 1/(2\bar{\omega} - 1)$.

Proof. Let $r^k \equiv r(x^k)$, $y^k \equiv y(x^k)$, and set

$$\eta^k := H(x^k)^{1/2} v(x^k) = -H(x^k)^{-1/2} r^k.$$

Since $\lim_{k\to\infty}(f(x^{k+1}) - f(x^k)) = 0$, and Armijo backtracking guarantees sufficient decrease by

$$f(x^{k+1}) \leq f(x^k) + \mu\alpha^k \nabla f(x^k)^\top v(x^k) = f(x^k) - \mu\alpha^k \|H(x^k)^{1/2} v(x^k)\|_2^2, \tag{5.4}$$

it follows that $\eta^k \to 0$. For $J \in 2^S$, define

$$\mathcal{K}_J = \{k \in \mathbb{N}_0 : \theta_j''(x_j^k)^{-1/2} \leq |\eta_j^k|^{1/2}\ \forall j \in J \text{ and } |r_j^k| \leq |\eta_j^k|^{1/2}\ \forall j \in J^c\}. \tag{5.5}$$

Since $|\eta_j^k| = |r_j^k \theta_j''(x_j^k)^{-1/2}|$ by definition, it follows that either $|r_j^k| \leq |\eta_j^k|^{1/2}$ or $\theta_j''(x_j^k)^{-1/2} \leq |\eta_j^k|^{1/2}$. Hence, for every $k \in \mathbb{N}_0$, there exists at least one $J \in 2^S$ such that $k \in \mathcal{K}_J$. Since 2^S is finite, there is at least one set J for which $\mathcal{K}_J$ is infinite. Fix such a set J. For all $k \in \mathcal{K}_J$, consider the system of linear inequalities defining a point $(p, z) \in \mathbb{R}^n \times \mathbb{R}^m \cong \mathbb{R}^{n+m}$, given by

$$\begin{aligned} &p_j = x_j^k, \quad q_j^\top p - a_j^\top z = -c_j + r_j^k \quad \text{for all } j \in J^c \equiv S \setminus J,\\ &p \geq 0, \quad Ap = b. \end{aligned} \tag{5.6}$$

Let $\mathcal{P}_k$ be the polyhedron defined by these inequalities. Since (x^k, y^k) satisfies these inequalities, we have $\mathcal{P}_k \neq \emptyset$ for all $k \in \mathcal{K}_J$. Moreover, for all $j \in J$, we have $\lim_{k\to\infty, k\in\mathcal{K}_J} \theta_j''(x_j^k) = \infty$, implying in turn that $\lim_{k\to\infty, k\in\mathcal{K}_J} x_j^k = 0$. Therefore, for all $k \in \mathcal{K}_J$ sufficiently large, Assumption 2 yields for $M_* = \max\{M_1, \dots, M_n\}$ the bound

$$(x_j^k)^\omega \leq M_j^{1/2} |\eta_j^k|^{1/2} \leq M_*^{1/2} |\eta_j^k|^{1/2} \quad \text{for all } j \in J, \tag{5.7a}$$

$$|r_j^k| \leq |\eta_j^k|^{1/2} \quad \text{for all } j \in J^c. \tag{5.7b}$$

If $\omega \in [1/2, 1)$, then $|\eta_j^k|^{\frac{1}{2\omega}} \leq |\eta_j^k|^{\frac{1}{2}}$, and therefore

$$x_j^k \leq C_1^{1/2} |\eta_j^k|^{1/2} \quad \text{for all } j \in J, \tag{5.8a}$$

$$|r_j^k| \leq C_1^{1/2} |\eta_j^k|^{1/2} \quad \text{for all } j \in J^c, \tag{5.8b}$$

for all $k \in \mathcal{K}_J$ sufficiently large, where we set $C_1 = \max\{1, M_*, M_*^{1/\omega}\}$.

If $\omega \geq 1$, then $|\eta_j^k|^{1/(2\omega)} \geq |\eta_j^k|^{1/2}$, and therefore, for $k \in \mathcal{K}_J$ sufficiently large, we get

$$x_j^k \leq C_1^{1/2} |\eta_j^k|^{1/(2\omega)} \quad \text{for all } j \in J, \tag{5.9a}$$

$$|r_j^k| \leq C_1^{1/2} |\eta_j^k|^{\frac{1}{2\omega}} \quad \text{for all } j \in J^c. \tag{5.9b}$$

Setting $\bar{\omega} = \max\{1, \omega\}$, the previous two estimates yield

$$x_j^k \leq C_1^{1/2} |\eta_j^k|^{1/(2\bar{\omega})} \quad \text{for all } j \in J, \tag{5.10a}$$

$$|r_j^k| \leq C_1^{\frac{1}{2}} |\eta_j^k|^{1/(2\bar{\omega})} \quad \text{for all } j \in J^c, \tag{5.10b}$$

and hence

$$\|(x_j^k, r_{J^c}^k)\|_{2\bar{\omega}}^{2\bar{\omega}} \leq C_1^{\bar{\omega}} \|\eta^k\|_1 \tag{5.11}$$

for all $k \in \mathcal{K}_J$ sufficiently large.

Hence, since $\eta^k \to 0$, we see that $\{(x_j^k, r_{J^c}^k)\}_{k \in \mathcal{K}_J} \to 0$. This implies that the right-hand side defining the polyhedron $\mathcal{P}_k$ is uniformly bounded. Since $\{(x^k, y^k)\}_{k \in \mathcal{K}_J}$ is bounded (see the proof of Lemma 4.3), any cluster point of this sequence must satisfy

$$p_j = 0, \quad q_j^\top p - a_j^\top z = -c_j \quad \text{for all } j \in J^c, \tag{5.12a}$$

$$p \geq 0 \quad Ap = b. \tag{5.12b}$$

Call $\mathcal{P}_J$ the polyhedron defined by the above linear inequalities. Let $(\bar{x}^k, \bar{y}^k)$ denote the Euclidean projection of (x^k, y^k) onto $\mathcal{P}_J$. Since $(x^k, y^k) \in \mathcal{P}_k$ for all $k \in \mathcal{K}_J$, Hoffman's error bound [26, Corollary 3.2.5] implies that

$$\|(\bar{x}^k, \bar{y}^k) - (x^k, y^k)\|_2 \leq C_2 \|(x_j^k, r_{J^c}^k)\|_{2\bar{\omega}} \quad \forall k \in \mathcal{K}_J, \tag{5.13}$$

where C_2 is a constant that depends only on $\bar{\omega}$, Q, A, and J. Combining this with (5.11) shows that

$$\|(\bar{x}^k, \bar{y}^k) - (x^k, y^k)\|_2 \leq C_2 C_1^{\frac{1}{2}} \|\eta^k\|_1^{\frac{1}{2\bar{\omega}}} \quad \forall k \in \mathcal{K}_J \text{ sufficiently large.} \tag{5.14}$$

We next claim that f is constant on $\mathcal{P}_J$. To see this, let $(p, z), (p', z') \in \mathcal{P}_J$ be arbitrary. Then,

$$\begin{aligned} f(p) - f(p') &= \frac{1}{2}(p - p')^\top Q (p - p') + (c + Qp')^\top (p - p') \\ &= \frac{1}{2}(p - p')^\top Q (p - p') + (c + Qp' - Az')^\top (p - p') \\ &= \frac{1}{2}(p - p')^\top Q (p - p'), \end{aligned} \tag{5.15}$$

where the second equality follows from the fact that $A(p - p') = 0$, and the third equality follows from the definition of $\mathcal{P}_J$. Similarly $f(p') - f(p) = \frac{1}{2}(p - p')^\top Q (p - p')$, resulting in $f(p') = f(p)$.

Next, observe that

$$\begin{aligned} (Q\bar{x}^k + c)^\top (x^k - \bar{x}^k) &= (Q\bar{x}^k + c - A^\top \bar{y}^k)^\top (x^k - \bar{x}^k) \\ &= \sum_{j \in J} (q_j^\top \bar{x}^k + c_j - a_j^\top \bar{y}^k) x_j^k \\ &= \sum_{j \in J} (q_j^\top (\bar{x}^k - x^k) - a_j^\top (\bar{y}^k - y^k) - r_j^k) x_j^k. \end{aligned} \tag{5.16}$$

From this, we compute

$$\begin{aligned}|f(x^k)-f(\bar{x}^k)| &= \left|\frac{1}{2}(x^k-\bar{x}^k)^\top Q(x^k-\bar{x}^k)+(Q\bar{x}^k+c)^\top(x^k-\bar{x}^k)\right|\\ &\le \frac{1}{2}\lambda_{\max}(Q)\,\|x^k-\bar{x}^k\|_2^2+\left|\sum_{j\in J}(q_j^\top(\bar{x}^k-x^k)-a_j^\top(\bar{y}^k-y^k)+r_j^k)x_j^k\right|\\ &\le \frac{1}{2}\lambda_{\max}(Q)\,\|x^k-\bar{x}^k\|_2^2\\ &\quad+\sum_{j\in J}\big[\|(q_j,-a_j)\|\cdot\|(\bar{x}^k,\bar{y}^k)-(x^k,y^k)\|_2x_j^k+x_j^k|r_j^k|\big].\end{aligned}\tag{5.17}$$

Collecting all the information from the previous estimates, we can bound each of these terms, for $k\in\mathcal{K}_J$ sufficiently large and $j\in J$, as follows:

- $\|x^k-\bar{x}^k\|_2^2\le\|(x^k,y^k)-(\bar{x}^k,\bar{y}^k)\|_2^2\le C_2^2C_1\|\eta^k\|_1^{1/\bar{\omega}}$.
- $x_j^k\le C_1^{1/2}|\eta_j^k|^{1/(2\bar{\omega})}$.
- $x_j^k|r_j^k|\le C_3|\eta_j^k|^{1/\bar{\omega}}$.

To see the last relation, observe that if $\omega\in[1/2,1)$, we have

$$x_j^k|r_j^k|=x_j^k|\eta_j^k|\theta_j''(x_j^k)^{1/2}\le C_1^{1/2}(x_j^k)^{1-\omega}|\eta_j^k|\le C_1^{1-\omega/2}|\eta_j^k|^{(3-\omega)/2}\le C_1|\eta_j^k|.\tag{5.18}$$

The first equality uses the identity $r_j^k=-\eta_j^k\theta_j''(x_j^k)^{1/2}$. The first inequality uses Assumption 2, and the second inequality is a consequence of relation (5.8a). The final inequality follows since $\eta_j^k\to0$ as $\mathcal{K}_J\ni k\to\infty$. Now assume that $\omega\ge1$. We first deduce from Assumption 2 the inequality $(x_j^k)^\omega|r_j^k|\le C_1^{1/2}|\eta_j^k|$, and then

$$\big(x_j^k|r_j^k|\big)^\omega\le C_4(x_j^k)^\omega|r_j^k|\le C_4C_1^{1/2}|\eta_j^k|,\tag{5.19}$$

where $C_4=\max_{k\ge1}|r_j^k|^{\omega-1}<\infty$. Departing from this relation, we obtain

$$x_j^k|r_j^k|\le C_4^{1/\omega}C_1^{1/(2\omega)}|\eta_j^k|^{1/\omega}.$$

To combine the two cases, set $C_3:=\max\{C_1,C_4^{1/\omega}C_1^{1/(2\omega)}\}$, and recall that $\bar{\omega}=\max\{1,\omega\}$.

Using all these bounds, we conclude that there exists a constant $C_J>0$ such that

$$|f(x^k)-f(\bar{x}^k)|\le C_J\|\eta^k\|_1^{1/\bar{\omega}}\tag{5.20}$$

for all $k\in\mathcal{K}_J$ sufficiently large. Let C_* be the maximum of C_J over all $J\in2^S$ for which $\mathcal{K}_J$ is infinite. Thus, there exists an index $\bar{k}\in\mathbb{N}$ sufficiently large so that for all $k\ge\bar{k}$ we have

$$|f(x^k)-f(\bar{x}^k)|\le C_*\|\eta^k\|_1^{1/\bar{\omega}}.\tag{5.21}$$

The sequence $f(x^k)$ is bounded and decreasing, so there exists $f_\infty>-\infty$ such that $f(x^k)\downarrow f_\infty$. Since $\{\bar{x}^k\}_{k\in\mathcal{K}_J}\subset\mathcal{P}_J$, it follows from the constancy of f on $\mathcal{P}_J$ that $f(\bar{x}^k)=f_\infty$ for all $k\in\mathcal{K}_J$, and thus for all $k\ge\bar{k}$. Hence, (5.21) becomes

$$f(x^k)-f_\infty\le C_*\|\eta^k\|_1^{1/\bar{\omega}}\quad\forall k\ge\bar{k}.\tag{5.22}$$

Set $d^k := f(x^k) - f_\infty$. Armijo backtracking then gives

$$f(x^{k+1}) - f(x^k) \leq -\mu\alpha^k \|\eta^k\|_2^2 \leq -\mu\alpha^k C_5 \|\eta^k\|_1^2 \leq -C_6\|\eta^k\|_1^2. \tag{5.23}$$

Here the constant C_5 captures the equivalence of the norms $\|\cdot\|_1$ and $\|\cdot\|_2$, and the constant C_6 incorporates the boundedness of the step-size sequence $\{\alpha^k\}_k$. Hence,

$$f(x^k) - f(x^{k+1}) = d^k - d^{k+1} \geq C_6\|\eta^k\|_1^2. \tag{5.24}$$

Combining this with (5.22), we conclude that

$$C_6^{-1/(2\bar{\omega})}(d^k - d^{k+1})^{1/(2\bar{\omega})} \geq \|\eta^k\|_1^{1/\bar{\omega}} \geq d^k/C_*. \tag{5.25}$$

Hence, for $\kappa = C_*/C_6^{1/(2\bar{\omega})}$, we obtain the recursion

$$\kappa\left(d^k - d^{k+1}\right)^{1/(2\bar{\omega})} \geq d^k. \tag{5.26}$$

This can be rearranged to yield the equivalent expression

$$d^{k+1} \leq d^k - \left(d^k/\kappa\right)^{2\bar{\omega}} \tag{5.27}$$

for $k \geq \bar{k}$. Now, write $\phi(d^k)$ for the right-hand side of the above, and observe that the function ϕ is strictly increasing on the interval $[0, \tilde{x}]$, where $\tilde{x} = (\kappa^{2\bar{\omega}}/(2\bar{\omega}))^{1/(2\bar{\omega}-1)}$. Then, fix $\rho = 1/(2\bar{\omega}-1) \in (0,1]$ and choose constants $C > 0$ and $K \in \mathbb{N}$, $K \geq \bar{k}$ such that $C \geq \kappa^{2\bar{\omega}/(2\bar{\omega}-1)}$ and $d^K \leq CK^{-\rho} \leq \tilde{x}$.

Such a choice of constants C, K is indeed possible: first look for $K \geq \bar{k}$ such that $d^k \leq \tilde{x}$ for all $k \geq K$ and choose $C = \kappa^{2\bar{\omega}/(2\bar{\omega}-1)}$. If $d^K \leq CK^{-\rho} \leq \tilde{x}$ holds, there is nothing further to do. If $d^K > CK^{-\rho}$, increase C such that $CK^{-\rho} = \tilde{x}$ holds. If $CK^{-\rho} > \tilde{x}$, increase K to achieve $CK^{-\rho} \leq \tilde{x}$ and then again increase C such that $CK^{-\rho} = \tilde{x}$ holds. Then, $(C/\kappa)^{2\bar{\omega}} \geq C$, and therefore we have

$$\left(\frac{C}{k^\rho \kappa}\right)^{2\bar{\omega}} \geq Ck^{-2\rho\bar{\omega}}. \tag{5.28}$$

We will now prove by induction the claim that $d^k \leq Ck^{-\rho} \leq \tilde{x}$ holds for all $k \geq K$. The base case $k = K$ holds by construction of C and K. Assume now $k \geq K$ and $d^k \leq Ck^{-\rho} \leq \tilde{x}$. Then we obtain

$$d^{k+1} \leq d^k - \left(d^k/\kappa\right)^{2\bar{\omega}} = \phi(d^k) \leq \phi(Ck^{-\rho}) \leq Ck^{-\rho} - Ck^{-2\rho\bar{\omega}} \leq \frac{C}{(k+1)^\rho}, \tag{5.29}$$

where we used the fact that $\rho \leq 1$ and $1 - 1/k \leq (1-1/k)^\rho$. This shows that

$$f(x^k) - f_\infty \leq Ck^{-\rho} \quad \forall k \geq K, \tag{5.30}$$

so our proof is complete. □

6. Numerical experiments.

In this section, we validate the theoretical analysis of the previous sections via a series of numerical experiments and practical applications.

6.1. Experiments with common benchmarks. As a first illustration of the convergence of (HBA), we focus on two low-dimensional test functions that are widely used in the global optimization literature.

1. The Rosenbrock function:

$$f(x_1, x_2) = 100(x_2 - x_{1^2})^2 + (1 - x_1)^2, \tag{6.1}$$

   with input domain $x_1, x_2 \in [-3, 3]$.

2. The Beale function:

$$\begin{aligned} f(x_1, x_2) &= (1.5 - x_1 + x_1 x_2)^2 \\ &\quad + (2.25 - x_1 + x_1 x_2^2)^2 + (2.625 - x_1 + x_1 x_2^3)^2, \end{aligned} \tag{6.2}$$

   with input domain $x_1, x_2 \in [-4, 4]$.

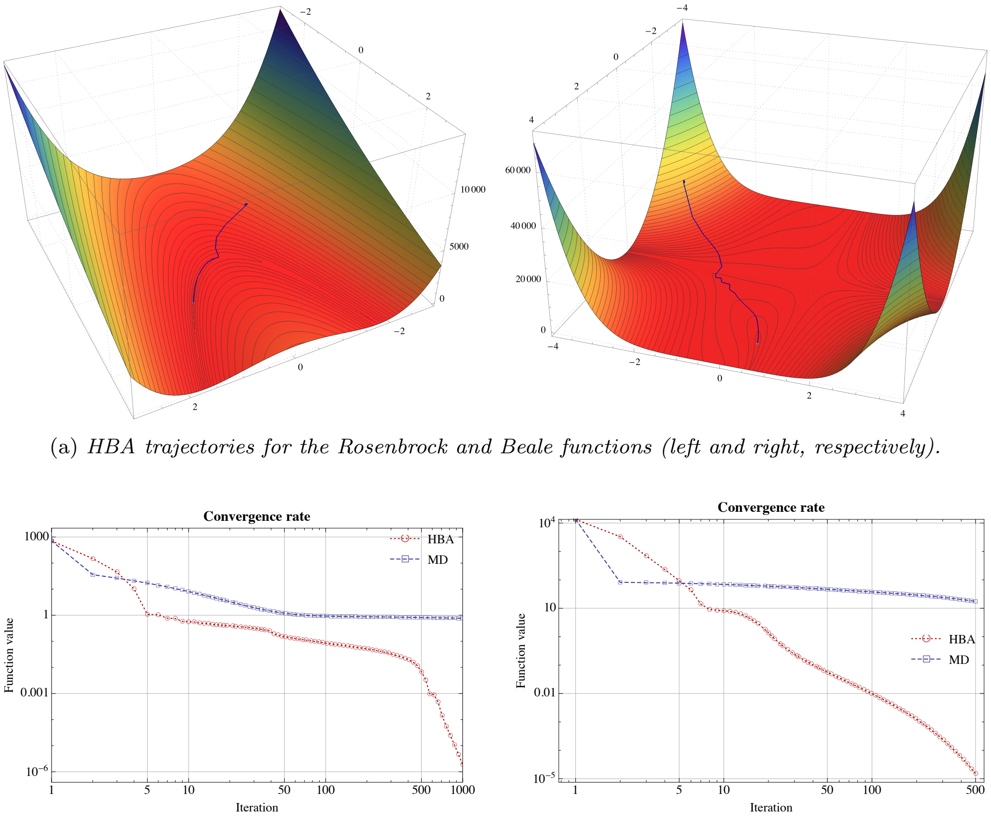
The Rosenbrock function is a nonconvex unimodal function with a unique global minimum located at the lowest point of a very flat and thin parabolic valley which is notoriously difficult for first-order methods to traverse. The Beale function is a nonconvex multimodal function with very sharp peaks at the corners of the input domain which cause considerable difficulties to aggressive step-size policies.

In Figure 1, we plot two test runs of the Hessian barrier algorithm (Algorithm 1) with the negative entropy kernel $\theta(x) = x \log x$ and a random initialization. For benchmarking purposes, we also ran the corresponding mirror descent algorithm (MD) with the same initialization, step size, and kernel function. The sample HBA trajectories are shown in Figure 1(a) and are seen to converge to a solution of (Opt). Subsequently, the value convergence rate of the algorithm is plotted in Figure 1(b): the log-log scale of the plot indicates a monotonic decrease following a power law convergence rate, consistent with the theoretical predictions of Theorem 5.1 (the nonuniformity of the algorithm's speed has to do with the very flat valleys/plateaus that the algorithm needs to traverse in order to approach a solution).

Even though we do not report the results here, a similar behavior was observed in all the common benchmarks (Himmelblau, Styblinski–Tang, etc.) and kernels (Burg, Hellinger, etc.) that we tested. We find this feature of the Hessian barrier algorithm particularly appealing for practical applications, especially for objectives with a complex landscape.

6.2. Applications to traffic routing. As a concrete application of our results, we focus below on the *traffic assignment problem* (TAP), a key problem in transportation and network science that concerns the optimal selection of paths between origins and destinations in traffic networks. Referring the reader to [12, 43] for a detailed discussion, the main ingredients of the problem are as follows: First, let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed multigraph with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. Assume further that there is a finite set of *origin-destination* (O/D) pairs indexed by $i \in \mathcal{N}$, each with an individual *traffic demand* $m^i \geq 0$ that is to be routed from the pair's *origin node* $o^i \in \mathcal{V}$ to its *destination* $d^i \in \mathcal{V}$. To route this traffic, the ith O/D pair employs a set $\mathcal{P}^i$ of *paths* joining o^i to d^i, with each path $p \in \mathcal{P}^i$ comprising a sequence of edges that meet head-to-tail in the usual way.[9]

[9] Specifically, we do not assume that $\mathcal{P}^i$ is necessarily the set of *all* paths joining o^i to d^i, but only some subset thereof. This distinction is important in packet-switched networks where, typically, only a set of paths with minimal hop count are used for traffic routing.



(a) *HBA trajectories for the Rosenbrock and Beale functions (left and right, respectively).*

(b) *Convergence rate for the Rosenbrock and Beale functions (left and right, respectively).*

FIG. 1. *Convergence of* (HBA) *in the case of the Rosenbrock and Beale test functions (*(6.1) *and* (6.2)*, respectively). The convergence rate of* (HBA) *is compared to that of a standard mirror descent algorithm; all experiments were run with the entropic kernel* $\theta(x) = x \log x$.

Now, writing $\mathcal{P} \equiv \bigcup_{i\in\mathcal{N}} \mathcal{P}^i$ for the ensemble of all such paths, the set of feasible *routing flows* $x = (x_p)_{p\in\mathcal{P}}$ in the network is defined as

$$\mathcal{X} = \left\{ x \in \mathbb{R}_+^{\mathcal{P}} : \textstyle\sum_{p\in\mathcal{P}^i} x_p = m^i \text{ for all } i \in \mathcal{N} \right\}. \tag{6.3}$$

In turn, a routing flow $x \in \mathcal{X}$ induces a *load* on each edge $e \in \mathcal{E}$ as

$$w_e = \sum_{p \ni e} x_p, \tag{6.4}$$

and we write $w = (w_e)_{e\in\mathcal{E}}$ for the corresponding *load profile* on the network. Given all this, the delay (or latency) experienced by an infinitesimal traffic element traversing edge e is determined by a nondecreasing continuous *cost function* $c_e \colon [0,\infty) \to [0,\infty)$: more precisely, if $w = (w_e)_{e\in\mathcal{E}}$ is the load profile induced by a feasible routing flow $x = (x_p)_{p\in\mathcal{P}}$, the incurred delay on edge $e \in \mathcal{E}$ is $c_e(w_e)$. Hence, with a slight abuse of notation, the associated cost of path $p \in \mathcal{P}$ will be

$$c_p(x) = \sum_{e\in p} c_e(w_e), \tag{6.5}$$

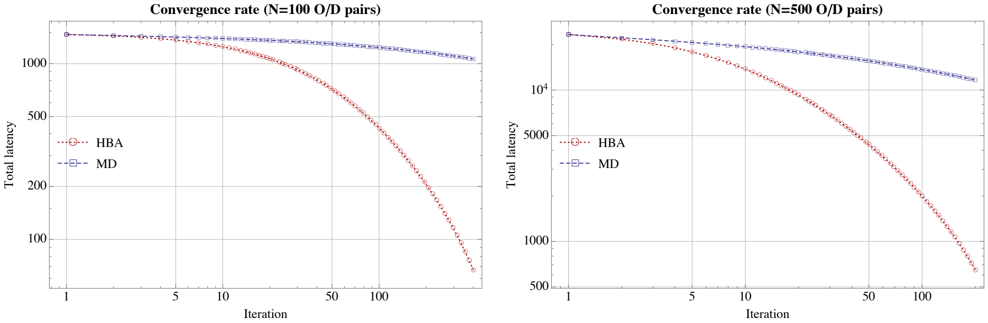


FIG. 2. *Convergence of Algorithm* 1 *in the traffic assignment problem* (TAP). *The base network is a randomly drawn Barabasi–Albert graph with $|\mathcal{V}| = 50$ nodes and $N = 100$ or $N = 500$ O/D pairs (left and right, respectively). In both cases, Algorithm* 1 *exhibits a very fast rate of convergence relative to standard mirror descent methods.*

and the aggregate latency in the network will be given by

$$C(x) = \sum_{i\in\mathcal{N}} \sum_{p^i \in \mathcal{P}^i} c_{p^i}(x) = \sum_{p\in\mathcal{P}} c_p(x). \tag{6.6}$$

Accordingly, with all this at hand, the goal of the traffic assignment problem is to identify a flow profile that minimizes the aggregate latency in the network, i.e., solve the continuous, nonlinear problem

$$\text{(TAP)} \qquad \begin{aligned} \text{minimize} \quad & C(x) \\ \text{subject to} \quad & x \in \mathcal{X}. \end{aligned}$$

Since (TAP) is a linearly constrained problem, the proposed HBA algorithm can be applied essentially "off the shelf." To do so, we consider an experimental setup consisting of a Barabasi–Albert random graph with $|\mathcal{V}| = 50$ nodes and N origin-destination pairs chosen uniformly at random from the generated graph. Subsequently, we used a variant of Dijkstra's algorithm to pick out $|\mathcal{P}^i| = 20$ minimal hop count paths per O/D pair, and we drew the corresponding traffic demands m^i, $i \in \mathcal{N}$, uniformly at random from $[0, 1]$. Concretely, in our experiments, we took $N = 100$ and $N = 500$, implying in turn that the dimensionality $n = \sum_{i\in\mathcal{N}} |\mathcal{P}^i|$ of the resulting traffic assignment problem is $n = 1000$ or $n = 2500$, respectively. The network's edge cost functions were also drawn randomly following a straightforward linear model of the form $c_e(w) = a_e + b_e w_e$, with a_e and b_e drawn uniformly at random from $[0, 10]$ and $[0, 1]$, respectively.

Our results are shown in Figure 2. In detail, since the problem's feasible region is a high-dimensional simplex (or, rather, a product thereof), we focused on the negative entropy kernel $\theta(x) = x \log x$, which is known to achieve a (nearly) dimension-free convergence rate for mirror descent [9, 40]. Subsequently, we ran both the Hessian barrier algorithm and mirror descent with the uniform traffic assignment initialization $x^i_p = m^i / |\mathcal{P}^i|$, $p \in \mathcal{P}^i$, $i \in N$, which is standard in the traffic assignment literature [12, 53]. In both cases, the HBA algorithm exhibits great gains in total latency after no more than a few hundred iterations: specifically, we observe a total latency reduction of over 95% relative to uniform traffic assignment, and over 90% relative to mirror descent after the same number of iterations. Given the problem's dimensionality of a few thousand control variables, this represents a gain that is particularly encouraging for other applications of the algorithm to large-scale optimization problems.

7. Conclusion. In this paper, we presented a class of first-order methods that includes as special cases several widely used numerical schemes for solving (possibly nonconvex) smooth optimization problems with linear constraints. Motivated by the continuous-time Hessian–Riemannian gradient dynamics of [1], we construct a computationally efficient algorithm which avoids the need for a prox-step. We call this method the *Hessian barrier algorithm* (HBA). We show that HBA, accompanied with a line search procedure based on Armijo backtracking, yields convergence to KKT points. In the case of quadratic programming, we also provide a sublinear value convergence rate. Interestingly, the rate depends on the employed metric, highlighting its importance as a design choice.

There are several interesting and challenging open questions left for future research. A first step concerns the extension of HBA methods to nonsmooth problems: in particular, the key driver in proving global convergence is the lower bound on the algorithm's step-size sequence. From the proof of Lemma 4.3, it is clear that we can actually weaken the smoothness assumption made on the objective function significantly in that regard. We therefore conjecture that it is possible to extend our arguments to problems in which the objective function f is not smooth, which would allow us to apply (HBA) to important applications in statistics and signal processing [31].

To better assess the method's total oracle complexity, it is important to make a distinction between gradient and function evaluations. With regard to the former, a key extension of our work would be to an accelerated version of (HBA): recently, [29] introduced a gradient method for nonconvex optimization problems, raising the question of whether this method can be combined with Hessian–Riemannian gradient steps. On the other hand, to estimate the number of function evaluations per iteration/gradient call, one would need to establish a bound on the number of Armijo backtracking steps per iteration. Given the highly nonlinear dependence of the bootstrap step size $\alpha_0(x)$ on the problem's primitives, this question seems to be a fairly challenging technical exercise which we leave for future work.

Finally, we should mention that we have presented (HBA) as a generic template for first-order methods: the search direction $v(x)$ can be changed to other data structures, such as a statistical estimator for the gradient, or the profile of individual gradients in a game-theoretic problem. This opens the door to analyzing (HBA) in the context of stochastic optimization and/or variational inequalities. This would provide a unifying framework for the recent results of [31, 35]; we delegate this technically challenging question to future work.

Acknowledgments. The authors are indebted to the associate editor and two anonymous referees for their detailed suggestions and remarks. Mathias Staudigl would also like to thank the University of Vienna for its hospitality while finishing this paper.

REFERENCES

[1] F. Alvarez, J. Bolte, and O. Brahic, *Hessian Riemannian gradient flows in convex programming*, SIAM J. Control Optim., 43 (2004), pp. 477–501.

[2] H. Attouch, J. Bolte, P. Redont, and M. Teboulle, *Singular Riemannian barrier methods and gradient-projection dynamical systems for constrained optimization*, Optimization, 53 (2004), pp. 435–454.

[3] H. Attouch, J. Bolte, and B. F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting and regularized Gauss–Seidel method*, Math. Program., 137 (2013), pp. 91–129.

[4] H. Attouch, X. Goudou, and P. Redont, *The heavy ball with friction method,* I. *The con-*

tinuous dynamical system: Global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system, Commun. Contemp. Math., 2 (2000), pp. 1–34.

[5] H. Attouch and J. Peypouquet, *The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $1/k^2$*, SIAM J. Optim., 26 (2016), pp. 1824–1834.

[6] H. Attouch and M. Teboulle, *Regularized Lotka–Volterra dynamical system as continuous proximal-like method in optimization*, J. Optim. Theory Appl., 121 (2004), pp. 541–570.

[7] A. Auslender, P. J. S. Silva, and M. Teboulle, *Nonmonotone projected gradient methods based on barrier and Euclidean distances*, Comput. Optim. Appl., 38 (2007), pp. 305–327, https://doi.org/10.1007/s10589-007-9025-0.

[8] A. Auslender and M. Teboulle, *Interior gradient and proximal methods for convex and conic optimization*, SIAM J. Optim., 16 (2006), pp. 697–725, https://doi.org/10.1137/S1052623403427823.

[9] A. Beck and M. Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Oper. Res. Lett., 31 (2003), pp. 167–175.

[10] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[11] M. Bertero, P. Boccacci, G. Desiderà, and G. Vicidomini, *Image deblurring with Poisson data: From cells to galaxies*, Inverse Problems, 25 (2009), 123006.

[12] D. P. Bertsekas and R. Gallager, *Data Networks*, 2nd ed., Prentice Hall, Englewood Cliffs, NJ, 1992.

[13] J. Bolte and M. Teboulle, *Barrier operators and associated gradient-like dynamical systems for constrained minimization problems*, SIAM J. Control Optim., 42 (2003), pp. 1266–1292.

[14] I. M. Bomze, *Evolution towards the maximum clique*, J. Global Optim., 10 (1997), pp. 143–164.

[15] I. M. Bomze, *Global escape strategies for maximizing quadratic forms over a simplex*, J. Global Optim., 11 (1997), pp. 325–338, https://doi.org/10.1023/A:1008297330705.

[16] I. M. Bomze, *Regularity versus degeneracy in dynamics, games, and optimization: A unified approach to different aspects*, SIAM Rev., 44 (2002), pp. 394–414, https://doi.org/10.1137/S00361445003756.

[17] I. M. Bomze, W. Schachinger, and R. Ullrich, *The complexity of simple models—a study of worst and typical hard cases for the standard quadratic optimization problem*, Math. Oper. Res., 43 (2017), pp. 651–674, https://doi.org/10.1287/moor.2017.0877.

[18] J. Bonnans and C. Pola, *A trust region interior point algorithm for linearly constrained optimization*, SIAM J. Optim., 7 (1997), pp. 717–731, https://doi.org/10.1137/S1052623493250639.

[19] M. Bravo, D. S. Leslie, and P. Mertikopoulos, *Bandit learning in concave N-person games*, in Adv. Neural Inf. Process. Syst. 32, Curran Associates, Red Hook, NY, 2018, pp. 5666–5676.

[20] S. Bubeck, *Convex optimization: Algorithms and complexity*, Found. Trends Mach. Learn., 8 (2015), pp. 231–358.

[21] A. Cabot, H. Engler, and S. Gadat, *On the long time behavior of second order differential equations with asymptotically small dissipation*, Trans. Amer. Math. Soc., 361 (2009), pp. 5983–6017.

[22] E. Candès and T. Tao, *The Dantzig selector: Statistical estimation when p is much larger than n*, Ann. Statist., 35 (2007), pp. 2313–2351.

[23] E. J. Candès and T. Tao, *Decoding by linear programming*, IEEE Trans. Inf. Theory, 51 (2005), pp. 4203–4215.

[24] B. Cox, A. Juditsky, and A. Nemirovski, *Dual subgradient algorithms for large-scale nonsmooth learning problems*, Math. Program., 148 (2014), pp. 143–180, https://doi.org/10.1007/s10107-013-0725-1.

[25] J. J. Duistermaat, *On Hessian Riemannian structures*, Asian J. Math., 5 (2001), pp. 79–91.

[26] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Vol. I, Springer Ser. Oper. Res. Financ. Eng., Springer, New York, 2003.

[27] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Vol. II, Springer Ser. Oper. Res. Financ. Eng., Springer, New York, 2003.

[28] M. Fukushima, *A modified Frank–Wolfe algorithm for solving the traffic assignment problem*, Transp. Res. B, 18 (1984), pp. 169–177.

[29] S. Ghadimi and G. Lan, *Accelerated gradient methods for nonconvex nonlinear and stochastic programming*, Math. Program., 156 (2016), pp. 59–99, https://doi.org/10.1007/s10107-015-0871-8.

[30] O. Güler, D. den Hertog, C. Roos, T. Terlaky, and T. Tsuchiya, *Degeneracy in interior point methods for linear programming: A survey*, Ann. Oper. Res., 46 (1993), pp. 107–138,

https://doi.org/10.1007/BF02096259.

[31] G. Haeser, H. Liu, and Y. Ye, *Optimality Condition and Complexity Analysis for Linearly-Constrained Optimization Without Differentiability on the Boundary*, preprint, https://arxiv.org/abs/1702.04300, 2017.

[32] J. Hofbauer and K. Sigmund, *Evolutionary game dynamics*, Bull. Amer. Math. Soc. (N.S.), 40 (2003), pp. 479–519.

[33] J. Lagarias and R. Vanderbei, *I. I. Dikin's convergence result for the affine scaling algorithm*, Contemp. Math., 114 (1990), pp. 109–119.

[34] J. M. Lee, *Introduction to Smooth Manifolds*, Grad. Texts in Math. 218, Springer, New York, 2003.

[35] H. Liu, T. Yao, R. Li, and Y. Ye, *Folded concave penalized sparse linear regression: Sparsity, statistical performance, and algorithmic theory for local solutions*, Math. Program., 166 (2017), pp. 207–240, https://doi.org/10.1007/s10107-017-1114-y.

[36] P. Mertikopoulos, E. V. Belmega, R. Negrel, and L. Sanguinetti, *Distributed stochastic optimization via matrix exponential learning*, IEEE Trans. Signal Process., 65 (2017), pp. 2277–2290.

[37] P. Mertikopoulos and W. H. Sandholm, *Riemannian game dynamics*, J. Econom. Theory, 177 (2018), pp. 315–364.

[38] P. Mertikopoulos and M. Staudigl, *On the convergence of gradient-like flows with noisy gradient input*, SIAM J. Optim., 28 (2018), pp. 163–197, https://doi.org/10.1137/16M1105682.

[39] P. Mertikopoulos and Z. Zhou, *Learning in games with continuous action sets and unknown payoff functions*, Math. Program., 173 (2019), pp. 465–507.

[40] A. S. Nemirovski and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley, New York, 1983.

[41] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, Math. Program., 120 (2009), pp. 221–259.

[42] Y. Nesterov and A. S. Nemirovskii, *Interior Point Polynomial Methods in Convex Programming*, SIAM, Philadelphia, PA, 1994.

[43] N. Nisan, T. Roughgarden, É. Tardos, and V. V. Vazirani, eds., *Algorithmic Game Theory*, Cambridge University Press, Cambridge, 2007.

[44] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., Springer, New York, 2000.

[45] B. T. Polyak, *Introduction to Optimization*, Transl. Ser. in Math. Eng., Optimization Software, New York, 1987.

[46] R. A. Polyak, *Regularized Newton method for unconstrained convex optimization*, Math. Program., 120 (2009), pp. 125–145, https://doi.org/10.1007/s10107-007-0143-3.

[47] S. Shalev-Shwartz, *Online learning and online convex optimization*, Found. Trends Mach. Learn., 4 (2011), pp. 107–194.

[48] W. Su, S. Boyd, and E. J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, in Adv. Neural Inf. Process. Syst. 27, Curran Associates, Red Hook, NY, 2014, pp. 2510–2518.

[49] P. Tseng, *Convergence properties of Dikin's affine scaling algorithm for nonconvex quadratic minimization*, J. Global Optim., 30 (2004), pp. 285–300.

[50] P. Tseng, I. M. Bomze, and W. Schachinger, *A first-order interior-point method for linearly constrained smooth optimization*, Math. Program., 127 (2011), pp. 399–424, https://doi.org/10.1007/s10107-009-0292-7.

[51] R. J. Vanderbei, M. S. Meketon, and B. A. Freedman, *A modification of Karmarkar's linear programming algorithm*, Algorithmica, 1 (1986), pp. 395–407.

[52] S. A. Vavasis, *Quadratic programming is in NP*, Inform. Process. Lett., 36 (1990), pp. 73–77.

[53] L. Vigneri, G. Paschos, and P. Mertikopoulos, *Large-scale network utility maximization: Countering exponential growth with exponentiated gradients*, in Proceedings of the 38th IEEE International Conference on Computer Communications (INFOCOM '19), IEEE Press, Piscataway, NJ, 2019, pp. 1630–1638.

[54] A. Wibisono, A. C. Wilson, and M. I. Jordan, *A variational perspective on accelerated methods in optimization*, Proc. Natl. Acad. Sci. USA, 113 (2016), pp. E7351–E7358, https://doi.org/10.1073/pnas.1614734113.