HESSIAN BARRIER ALGORITHMS
FOR LINEARLY CONSTRAINED OPTIMIZATION PROBLEMS

Panayotis Mertikopoulos[1]

joint with

Immanuel M. Bomze **Werner Schachinger** **Mathias Staudigl**

[1]French National Center for Scientific Research (CNRS)

Laboratoire d'Informatique de Grenoble

GDO 2019, Cluj-Napoca, April 10, 2019

Outline

Linearly constrained problems

Focus of the talk:

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & x \in \mathcal{X} \equiv \{x \in \mathbb{R}^d : Ax = b,\ x \geq 0\} \end{aligned} \tag{Opt}$$

Primitives:

- Objective function $f \colon \mathbb{R}^d_+ \to \mathbb{R} \cup \{+\infty\}$
- Constraint data $A \in \mathbb{R}^{m \times d}$, $b \in \mathbb{R}^m$

Some background

Applications:

- Imaging science / signal processing
- Machine learning / data science
- Game theory / operations research
- …

Vast literature (can't do justice):

- Quasi-Newton methods
- Interior-point / active-set methods
- Conditional gradient (Franke–Wolfe)
- Mirror descent / Bregman proximal methods
- …

A dynamical systems viewpoint

Gradient flow:

$$\frac{dx}{dt} = -\nabla f(x) \tag{GD}$$

- ✗ Violates nonnegativity constraints
- ✗ Violates equality constraints

A dynamical systems viewpoint

Adjusted gradient flow:

$$\frac{dx}{dt} = -S(x)\,\nabla f(x) \tag{GD}$$

✓ Respects nonnegativity constraints if $S_{ij}(x) = 0$ when $x_i = 0$

✗ Violates equality constraints

A dynamical systems viewpoint

Adjusted projected gradient flow:

$$\frac{dx}{dt} = -P(x)\,S(x)\,\nabla f(x) \tag{GD}$$

✓ Respects nonnegativity constraints if $S_{ij}(x) = 0$ when $x_i = 0$

✓ Respects equality constraints if $\operatorname{im} P(x) = \ker A$

A dynamical systems viewpoint

Adjusted projected gradient flow:

$$\frac{dx}{dt} = -P(x)\,S(x)\,\nabla f(x) \tag{GD}$$

✓ Respects nonnegativity constraints if $S_{ij}(x) = 0$ when $x_i = 0$

✓ Respects equality constraints if $\operatorname{im} P(x) = \ker A$

Is there a principled way to choose S and P?

Riemannian gradient flows

Endow orthant $\mathcal{C} \equiv \mathbb{R}^d_+$ with a Riemannian metric:

$$\langle z_1, z_2 \rangle_x = z_1^\top g(x) z_2 \qquad z_1, z_2 \in \mathbb{R}^d$$

induced by some metric tensor $g(x) \succ 0, \;\; x \in \mathcal{C}$

Riemannian gradient flows

Endow orthant $\mathcal{C} \equiv \mathbb{R}^d_+$ with a Riemannian metric:

$$\langle z_1, z_2 \rangle_x = z_1^\top g(x) z_2 \qquad z_1, z_2 \in \mathbb{R}^d$$

induced by some metric tensor $g(x) \succ 0, \;\; x \in \mathcal{C}$

Principled choices for S and P:

- $S(x) = g(x)^{-1}$ [so $S(x)\nabla f(x) = \operatorname{grad} f(x)$]
- $P(x) = I - g(x)^{-1} A^\top (A g(x)^{-1} A^\top)^{-1} A$ [orthogonal projection to $\ker A$]

Riemannian gradient flows

Endow orthant $\mathcal{C} \equiv \mathbb{R}^d_+$ with a Riemannian metric:

$$\langle z_1, z_2 \rangle_x = z_1^\top g(x) z_2 \qquad z_1, z_2 \in \mathbb{R}^d$$

induced by some metric tensor $g(x) \succ 0, \;\; x \in \mathcal{C}$

Principled choices for S and P:

- $S(x) = g(x)^{-1}$ [so $S(x)\nabla f(x) = \operatorname{grad} f(x)$]
- $P(x) = I - g(x)^{-1} A^\top (A g(x)^{-1} A^\top)^{-1} A$ [orthogonal projection to $\ker A$]

However: well-posedness of (GD) requires blow-up of g near $\operatorname{bd}(\mathcal{C})$

Hessian Riemannian metrics

Generate metric by taking the **Hessian of a Legendre function:**

$$g(x) = \mathrm{Hess}(h(x))$$

where $h: \mathcal{C} \to \mathbb{R} \cup \{+\infty\}$ is:

- Strictly convex (+ proper, lsc) on $\mathcal{C}$
- Smooth on $\mathcal{C}^\circ$
- Steep at the boundary of $\mathcal{C}$ (i.e., $\operatorname{dom} \partial h = \mathcal{C}^\circ$)

Long history:

- Physics: thermodynamic fluctuation theory, integrable space-times,…
  [Shima, 1977; Ruppeiner, 1979;…]
- Diff. geometry: characterization of umbilical points, pinching,…
  [Duistermaat, 2001;…]
- **Optimization:** *Hessian Riemannian gradient flows*
  [Bolte & Teboulle, 2003; Alvarez &al., 2004;…]

Hessian Riemannian gradient flows

Hessian Riemannian gradient descent:

$$\frac{dx}{dt} = -\underbrace{\left[I - H(x)^{-1}A^\top \left(AH(x)^{-1}A^\top\right)^{-1}A\right]}_{\text{projection to } \ker A} \underbrace{H(x)^{-1}\nabla f(x)}_{\text{HR gradient}} \qquad \text{(HRGD)}$$

with $H(x) = \operatorname{Hess}(h(x))$

Hessian Riemannian gradient flows

Hessian Riemannian gradient descent:

$$\frac{dx}{dt} = -\underbrace{[I - H(x)^{-1}A^\top(AH(x)^{-1}A^\top)^{-1}A]}_{\text{projection to } \ker A}\underbrace{H(x)^{-1}\nabla f(x)}_{\text{HR gradient}} \qquad \text{(HRGD)}$$

with $H(x) = \mathrm{Hess}(h(x))$

Examples

1. Simplex constraints + Shahshahani metric / entropic regularization:
   $A = (1, \dots, 1)$ and $h(x) = \sum_{i=1}^d x_i \log x_i$ leads to the replicator dynamics

$$\frac{dx_i}{dt} = -x_i\left[\partial_i f(x) - \sum_{i=1}^d x_i \partial_i f(x)\right] \qquad \text{(RD)}$$

2. Affine scaling (Dikin, Karmarkar,...):
   General A, $h(x) = -\sum_{i=1}^d \log x_i$, gives the affine scaling dynamics

$$\frac{dx}{dt} = -[I - \mathrm{diag}(x)A^\top(A\,\mathrm{diag}(x)A^\top)^{-1}A]\,\mathrm{diag}(x)\nabla f(x) \qquad \text{(AS)}$$

Properties

Energy / Lyapunov functions:

- The objective itself (f)
- If f is $\{\cdots\}$-convex, **Bregman divergence** to global minimizer

$$D(p, x) = h(p) - h(x) - \langle \nabla h(x) \,|\, p - x \rangle$$

Properties

Energy / Lyapunov functions:

- The objective itself (f)
- If f is $\{\cdots\}$-convex, **Bregman divergence** to global minimizer

$$D(p, x) = h(p) - h(x) - \langle \nabla h(x) \,|\, p - x \rangle$$

Theorem (Bolte & Teboulle, 2003; Alvarez &al, 2004)

If: *f is $\{\cdots\}$-convex (+ technical conditions for h).*

Then: *any interior solution trajectory of (HRGD) converges to a solution of (Opt).*

Outline

Background

The Hessian barrier algorithm

Analysis and results

From flows to algorithms

General dynamics

$$\dot{x} = V(x) \tag{D}$$

[Here: $V(x) = -P(x)H(x)^{-1}\nabla f(x)$]

Obtain algorithm via **discretization:**

1. **Implicit:**

$$x^{+} = x + \alpha V(x^{+})$$

$\Longrightarrow$ Leads to **mirror descent**

[Nemirovski and Yudin, 1983; Attouch, Bolte, Teboulle + too many to list]

2. **Explicit:**

$$x^{+} = x + \alpha V(x)$$

[**this talk**]

The Hessian barrier algorithm

We consider a general explicit method:

$$x^+ = x + \alpha(x)V(x)$$

with

- Search direction given by projected HR gradient

$$V(x) = -P(x)H(x)^{-1}\nabla f(x)$$

- Variable step-size given by Armijo backtracking

$$f(x^+) \leq f(x) - \mu\alpha(x)\|V(x)\|_x^2 \quad \text{for some } \mu \in (0,1)$$

Hessian barrier algorithm

$$x_{t+1} = x_t - \alpha(x_t)P(x_t)H(x_t)^{-1}\nabla f(x_t) \tag{HBA}$$

The method's step-size

Key challenges for the HBA step-size:

1. **Feasibility:**

$$x_t \text{ feasible} \implies x_{t+1} \text{ feasible}$$

2. **Sufficient decrease:**

$$f(x_{t+1}) \leq f(x_t) - \mu\alpha(x_t)\|V(x_t)\|_{x_t}^2 \quad \text{for some } \mu \in (0,1)$$

3. **No early stops:**

$$\sum_{t=1}^{\infty} \alpha(x_t) = \infty$$

Feasibility

Focus on separable regularizers

$$h(x) = \sum_{i=1}^{d} \theta(x_i)$$

Then:

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

where $r(x) = -H(x)V(x)$ is the "reduced cost"

Feasibility guarantee:

$$\alpha(x) < \alpha_0(x) \equiv \min_{i=1,\dots,d} \{x_i \theta_i''(x_i)/r_i(x) : r_i(x) > 0\}$$

Sufficient decrease

Descent inequality for L-smooth f:

$$f(x^+) = f(x + \alpha(x)V(x)) \leq f(x) - \beta\alpha(x)\left[1 - \frac{\alpha(x)L}{2\beta}\right]\|V(x)\|_2^2$$

provided that $\theta''(z) \geq \beta$

Sufficient decrease:

$$f(x^+) \leq f(x) - \mu\alpha(x)\|V(x)\|_x^2 \quad \text{for some } \mu \in (0,1)$$

Armijo backtracking:

- Bootstrap: $\underline{\alpha}(x) = \min\{\alpha_0(x), 2\beta/L\}$
- Backtrack: shrink step-size by δ until suff. decrease satisfied

Sufficient decrease

Descent inequality for L-smooth f:

$$f(x^+) = f(x + \alpha(x)V(x)) \leq f(x) - \beta\alpha(x)\left[1 - \frac{\alpha(x)L}{2\beta}\right]\|V(x)\|_2^2$$

provided that $\theta''(z) \geq \beta$

Sufficient decrease:

$$f(x^+) \leq f(x) - \mu\alpha(x)\|V(x)\|_x^2 \quad \text{for some } \mu \in (0,1)$$

Armijo backtracking:

- Bootstrap: $\underline{\alpha}(x) = \min\{\alpha_0(x), 2\beta/L\}$
- Backtrack: shrink step-size by δ until suff. decrease satisfied

But does this terminate?

Early stops

Key lemma (Bomze, M, Schachinger, Staudigl, 2018): if $\inf_{z>0} z\theta''(z) > 0$, then

$$\inf_{x} \underline{\alpha}(x) > 0$$

Early stops

Key lemma (Bomze, M, Schachinger, Staudigl, 2018): if $\inf_{z>0} z\theta''(z) > 0$, then

$$\inf_x \underline{\alpha}(x) > 0$$

Key consequence:

$$\inf_t \alpha(x_t) > 0$$

Early stops

Key lemma (Bomze, M, Schachinger, Staudigl, 2018): if $\inf_{z>0} z\theta''(z) > 0$, then

$$\inf_x \underline{\alpha}(x) > 0$$

Key consequence:

$$\inf_t \alpha(x_t) > 0$$

HBA is feasible, guarantees sufficient decrease, and does not stop prematurely

Outline

The algorithm

Algorithm 1 The Hessian barrier algorithm

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

Hypotheses on primitives

Blanket assumption

The objective function of (Opt) *satisfies the following:*

1. **Regularity:** f *is proper, lsc, and* L*-smooth on* $\mathcal{X}$
2. **Level set boundedness:** $\{x \in \mathcal{X} : f(x) \leq f(x_0)\}$ *is bounded for some* $x_0 \in \mathcal{X}$
3. **Finite value:** $\min_{x \in \mathcal{X}} f(x) > -\infty$

Main convergence result

Theorem (Bomze, M, Schachinger, Staudigl, 2018)

1. *The sequence x_t is bounded and $f(x_t)$ is non-increasing.*
2. *Every limit point $\hat{x}$ of* (HBA) *satisfies reduced cost complementarity (RCC), i.e., $\hat{x}_i r_i(\hat{x}) = 0$ for all $i = 1, \dots, d$*
3. *Every limit point $\hat{x}$ of* (HBA) *is a KKT point of f if any of the following holds*
   - 3.1 *f is convex*
   - 3.2 *RCC points are isolated*
   - 3.3 *RCC points satisfy strict complimentarity, i.e., $\hat{x}_i + r_i(\hat{x}) > 0$ for all $i = 1, \dots, d$*

Main convergence result

Theorem (Bomze, M, Schachinger, Staudigl, 2018)

1. *The sequence x_t is bounded and $f(x_t)$ is non-increasing.*
2. *Every limit point $\hat{x}$ of (HBA) satisfies reduced cost complementarity (RCC), i.e., $\hat{x}_i r_i(\hat{x}) = 0$ for all $i = 1, \dots, d$*
3. *Every limit point $\hat{x}$ of (HBA) is a KKT point of f if any of the following holds*
   3.1 *f is convex*
   3.2 *RCC points are isolated*
   3.3 *RCC points satisfy strict complimentarity, i.e., $\hat{x}_i + r_i(\hat{x}) > 0$ for all $i = 1, \dots, d$*

Corollary (Immediate take-away)

If f is $\{\cdots\}$-convex, x_t converges to $\arg\min f$.

Applications to quadratic programming

Important case of interest:

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

for some symmetric $Q \in \mathbb{R}^{d \times d}$, $c \in \mathbb{R}^d$

Theorem (Bomze, M, Schachinger, Staudigl, 2018)

If: *HBA is run with a moderately steep kernel*

$$\frac{m}{z} \leq \theta''(z) \leq \frac{M}{z^{2\omega}} \quad \text{for some } \omega \geq 1/2,\ z \text{ suff. small}$$

Then: $f(x_t) - f_\infty = \mathcal{O}(1/t^\rho)$ *with* $\rho = (2\max\{1, \omega\} - 1)^{-1}$.

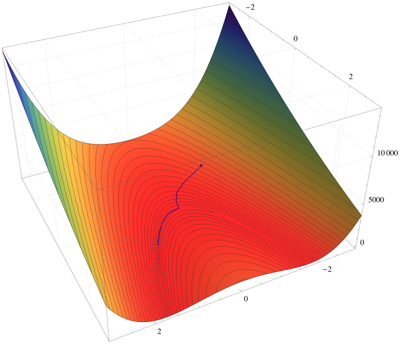
[Best choice: $\theta(x) = x \log x$, $\rho = 1$]

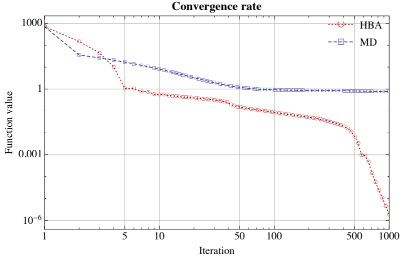
Numerical experiments

The Rosenbrock benchhmark:

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

$[-3 \le x_{1,2} \le 3]$



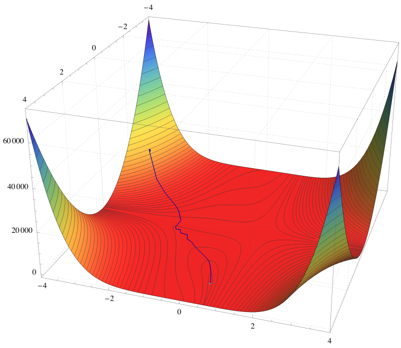


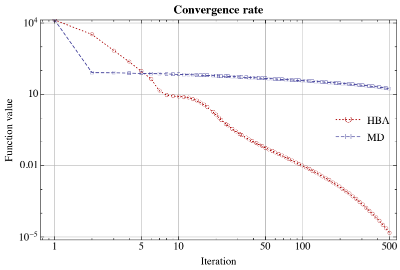
Numerical experiments

The Beale benchhmark:

$$f(x_1, x_2) = (1.5 - x_1 + x_1 x_2)^2 + (2.25 - x_1 + x_1 x_2^2)^2 + (2.625 - x_1 + x_1 x_2^3)^2$$

$[-4 \leq x_{1,2} \leq 4]$





Numerical experiments

Traffic routing:

$$\begin{aligned} &\text{minimize} && f(x) = \sum_{e\in\mathcal{E}} x_e c_e(x_e) && \text{[aggregate delay]} \\ &\text{subject to} && x_e \geq 0 && \text{[nonneg. loads]} \\ & && x_e = \sum_{i=1}^{N} \sum_{p\in\mathcal{P}_i, p\ni e} x_{ip} && \text{[loads induced by traffic]} \\ & && \sum_{p\in\mathcal{P}_i} x_{ip} = m_i && \text{[total inflow of an O/D pair]} \end{aligned}$$