ABRA-GD: ADAPTIVE BREGMAN ACCELERATED GRADIENT DESCENT FOR RELATIVELY SMOOTH AND (STRONGLY-)CONVEX OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose ABrA-GD (Adaptive Bregman Accelerated Gradient Descent), an adaptive algorithm for relatively smooth convex optimization. The algorithm is derived from a computable primal–dual certificate that guarantees progress by comparing the objective value with a lower bound on the minimum of a regularized objective. This certificate enables adaptation to both smoothness and geometry, using only the relative strong convexity constant as a problem-dependent input. We introduce the dual Bregman Length Distortion Factor (BLDF), which measures how dual Bregman lengths change under an anchor shift or rescaling. Under bounded local dual BLDF, ABrA-GD achieves an accelerated $O(1/k^2)$ rate for convex objectives and an accelerated linear rate for relatively strongly convex objectives.

1 SETTING AND INTRODUCTION

We consider optimization problems of the form

$$\phi_\star = \min_{x\in\mathcal{X}} \phi(x) \;:=\; f(x) + \mu D_d(x, x_0), \quad D_d(x,y) := d(x) - d(y) - \langle \nabla d(y), x - y\rangle. \tag{1}$$

The term $D_d(x, y)$ is the Bregman divergence generated by d (Bregman, 1967; Censor & Lent, 1981; Censor & Zenios, 1992). Throughout the paper we work under the following hypotheses.

Assumption 1 (Problem setting)**.** *d is a Legendre function, $\mathcal{X} = \mathrm{cl}(\mathrm{dom}\, d)$, and $x_0 \in \mathrm{int}(\mathrm{dom}\, d)$. The function $f : \mathcal{X} \to \mathbb{R}$ is convex, differentiable, and relatively $(L-\mu)$-smooth with respect to d (Bauschke et al., 2017; Lu et al., 2018), with $0 \le \mu < L$.*

Assumption 1 is equivalent to requiring ϕ to be L-smooth and μ-strongly convex relative to d (appendices A and A.1). The domain of d determines the feasible set. The analysis uses interior minimizers, for which the optimality conditions are stationarity equations. Affine equality constraints are also allowed: after a change of coordinates, the problem satisfies the same assumptions (appendix A.2).

Acceleration in Euclidean optimization. For smooth convex objectives, accelerated gradient methods improve the $O(1/k)$ rate of gradient descent to $O(1/k^2)$ (Nesterov, 2004). Several viewpoints explain this improvement, including estimate sequences, Lyapunov functions, and the coupling of gradient and mirror-descent steps (Allen-Zhu & Orecchia, 2014; Wilson et al., 2021). Primal–dual interpretations express progress through a certificate that combines an objective upper bound with a lower model (Diakonikolas & Orecchia, 2019).

Beyond Euclidean smoothness. Relative smoothness replaces the Euclidean quadratic upper model with a Bregman divergence (Bauschke et al., 2017; Lu et al., 2018). This allows objectives whose gradients are not globally Lipschitz, including log-determinant, entropy, and Poisson-type objectives with singular behavior near the domain boundary. Bregman gradient descent attains an $O(1/k)$ rate in this setting. Unlike the Euclidean case, this rate cannot be improved over the full relatively smooth class (Dragomir et al., 2022). Acceleration therefore requires additional structure.

Conditions for Bregman acceleration. Accelerated Bregman Proximal Gradient (ABPG) methods obtain $O(k^{-\gamma})$ rates under a Triangle Scaling Property (TSP) with exponent γ and gain 1 (Hanzely

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2021; Savchuk et al., 2024). For standard geometries such as entropy and Burg, this exponent is at most 1, so *this guarantee gives no acceleration*. The accelerated variants instead fix $\gamma = 2$ and backtrack on the gain, so their rate depends on the gains accepted during the run. Chen et al. (2026) obtain an accelerated linear rate under a generalized Cauchy–Schwarz condition, whose constant is an input of their method. Liu & Liu (2022) additionally assume relative strong convexity, but their rate does not improve under this assumption.

Contributions

1. **A primal–dual formulation for acceleration (section 2).** We formulate a descent requirement using the Bregman–Moreau envelope and certify it through its dual problem. This framework provides a common starting point for the algorithm and its analysis; appendix B recovers Bregman gradient descent and dual averaging from it.
2. **ABrA-GD (section 3).** We derive ABrA-GD (algorithm 2), in which one gradient serves two steps: A Bregman gradient step gives an upper bound on the objective, and a dual-averaging step gives a lower bound on the minimum of a regularized objective. ABrA-GD backtracks on smoothness and acceleration parameters, and we establish a generic convergence bound in terms of the accepted acceleration parameters (Theorem 2).
3. **A local measure of geometric distortion (section 4).** We introduce the dual Bregman Length Distortion Factor (BLDF) to compare the progress of gradient and dual-averaging steps. We give explicit formulas for standard separable kernels and relate this local condition to the triangle-scaling conditions. The algorithm adapts without evaluating the BLDF.
4. **Accelerated rates under bounded distortion (section 5).** Under bounded local distortion, *a condition implied by a uniform triangle-scaling gain (appendix E.2)*, we obtain an accelerated $O(1/k^2)$ rate for convex objectives and an accelerated linear rate for relatively strongly convex objectives. ABrA-GD's only problem-constant input is μ. To our knowledge, it is the first method with these inputs that has an accelerated linear rate for this class.

Limitations. The accelerated analysis does not cover an additional nonsmooth term, unlike the ABPG analyses of Hanzely et al. (2021); Savchuk et al. (2024). The envelope descent requirement extends to composite objectives (appendix F), but an accelerated BLDF analysis remains open.

2 The Primal–Dual Envelope Framework

For $\eta > 0$, define the regularized objective

$$\phi_\star^\eta := \min_x \phi^\eta(x), \quad \phi^\eta(x) := \phi(x) + \frac{1}{\eta} D_d(x, x_0) = f(x) + \left(\mu + \tfrac{1}{\eta}\right) D_d(x, x_0), \qquad \eta > 0. \tag{2}$$

The value $\phi_\star^\eta$ is the Bregman–Moreau envelope of ϕ anchored at x_0 (Bauschke et al., 2018), which approaches $\phi_\star$ as $\eta \to \infty$. Consider the implicit envelope descent scheme (algorithm 1).

Algorithm 1: Implicit envelope descent (conceptual algorithm, used only as convergence tool)

Input: ϕ, x_0. **Initialization:** $k \leftarrow 0, \eta_0 \leftarrow 0$.

for $k = 0, 1, 2, \ldots$ **do**

  Choose $\eta_{k+1} \geq \eta_k$ and find x_{k+1} such that $\phi(x_{k+1}) \leq \phi_\star^{\eta_{k+1}}$ (defined in (2)).

Each iterate must satisfy $\phi(x_{k+1}) \leq \phi_\star^{\eta_{k+1}}$, but it need not minimize the regularized objective (fig. 1). The scheme does not specify how to compute x_{k+1} or verify this requirement, since the envelope value is itself defined by a minimization. It does, however, give an immediate error bound.

Theorem 1. *Under As. 1, algorithm 1 with inputs ϕ, x_0 satisfies*

$$\phi(x_k) - \phi_\star \leq \eta_k^{-1} D_d(x_\star, x_0). \tag{3}$$

To prove the bound, evaluate the regularized objective at $x_\star$ (appendix C.1). The growth of η_k is therefore the central choice. Faster growth gives a faster rate in (3), but it also lowers $\phi_\star^{\eta_k}$ toward $\phi_\star$, so each iteration must make more progress.

Link with proximal methods. The proximal point method (Lemaréchal & Sagastizábal, 1997) centers each subproblem at the current iterate, and Accelerated Hybrid Proximal Extragradient (Monteiro & Svaiter, 2013) and Catalyst (Lin et al., 2015; 2018) solve these subproblems approximately. In algorithm 1, the anchor x_0 stays fixed and no subproblem is solved.

2.1 Envelope Dual Problem and Lower Bound

The envelope value $\phi_\star^{\eta_{k+1}}$ is defined by a minimization, so the condition $\phi(x_{k+1}) \le \phi_\star^{\eta_{k+1}}$ cannot be checked directly. We certify it through the dual problem and provide a sufficient condition for the requirement of algorithm 1. For any λ, the Fenchel–Young inequality gives the affine lower bound

$$\ell_\lambda(x) := \langle x, \lambda\rangle - f^*(\lambda) \le f(x), \tag{4}$$

where f^* is the Fenchel conjugate of f, with equality if and only if $\lambda = \nabla f(x)$. Adding the regularization defines the dual value and its associated primal point:

$$\mathcal{D}^\eta(\lambda) := \min_x \left\{\ell_\lambda(x) + (\mu + \eta^{-1})D_d(x, x_0)\right\}, \tag{5}$$

$$z^\eta(\lambda) := \arg\min_x \left\{\ell_\lambda(x) + (\mu + \eta^{-1})D_d(x, x_0)\right\} = \nabla d^*\left(\nabla d(x_0) - \frac{\lambda}{\mu + \eta^{-1}}\right). \tag{6}$$

By weak duality, we have $\ell_\lambda \le f$, and therefore $\mathcal{D}^\eta(\lambda) \le \phi_\star^\eta \le \phi^\eta(x)$ for every x. Although $z^\eta(\lambda)$ is computable without evaluating $f^*(\lambda)$, the dual value $\mathcal{D}^\eta(\lambda)$ generally is not. The next section shows how to certify $\phi(x_{k+1}) \le \mathcal{D}^{\eta_{k+1}}(\lambda_{k+1})$ without evaluating f^*.

3 Deriving ABrA-GD and Its Accepted-Parameter Rate

Algorithm 1 is conceptual. The dual formulation (5) says what to certify, $\phi(x_{k+1}) \le \mathcal{D}^{\eta_{k+1}}(\lambda_{k+1})$, but not how to construct x_{k+1}, λ_{k+1} (hence $z_{k+1} = z^{\eta_{k+1}}(\lambda_{k+1})$), or η_{k+1}. ABrA-GD constructs all of them, with $\eta_k \to \infty$, and maintains the invariant

$$\phi_k^{\text{best}} \le \mathcal{D}^{\eta_k}(\lambda_k) \le \phi_\star^{\eta_k}. \tag{7}$$

Each trial of an iteration evaluates one gradient at a point y_k between x_k^{best} and z_k. This gradient gives a primal candidate x_{k+1} and a dual update λ_{k+1}. A computable lower bound then tests whether the invariant holds at step $k+1$; if not, backtracking increases M_k and repeats the trial.

1. Common gradient evaluation. Both model updates use one common gradient evaluation at the point y_k, defined below. Here x_k^{best} is the best retained point and z_k is the current primal map; the analysis specifies τ_k (lemma 7):

$$y_k = (1 - \tau_k)x_k^{\text{best}} + \tau_k z_k. \tag{8}$$

2. Primal upper model. The upper model at y_k is

$$\mathcal{U}(x, y_k) := \phi(y_k) + \langle \nabla\phi(y_k), x - y_k\rangle + L_{\text{cur}} D_d(x, y_k). \tag{9}$$

Since ϕ is L-smooth relative to d (As. 1), $\phi \le \mathcal{U}(\cdot, y_k)$ when $L_{\text{cur}} \ge L$. The *Primal Step* in algorithm 3 minimizes this model:

$$x_{k+1} = \arg\min_x \mathcal{U}(x, y_k). \tag{10}$$

BACKTRACKING-GD (algorithm 4) adjusts L_{cur} until $\phi(x_{k+1}) \le \mathcal{U}(x_{k+1}, y_k)$, and terminates whenever $L_{\text{cur}} \ge L$.

3. Dual step and primal map. The *Dual Averaging Step* updates λ_k with the same gradient:

$$\lambda_{k+1} = (1 - t_k)\lambda_k + t_k \nabla f(y_k), \quad t_k \in [0, 1]. \tag{11}$$

Here $\nabla f(y_k) = \nabla\phi(y_k) - \mu(\nabla d(y_k) - \nabla d(x_0))$, so no additional objective-gradient evaluation is needed. The relation to dual best responses and conditional gradients (Bach, 2015) is discussed in appendix B.3. The *Primal Map* then gives

$$z_{k+1} = z^{\eta_{k+1}}(\lambda_{k+1}). \tag{12}$$

Both updates in eqs. (11) and (12) appear in algorithm 3.

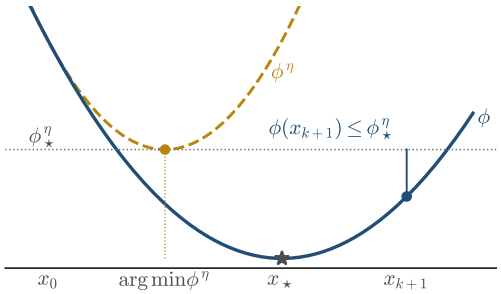

Figure 1: Envelope descent requires that $\phi(x_{k+1}) \le \phi^\eta_\star$, without minimizing the regularized objective ϕ^η.

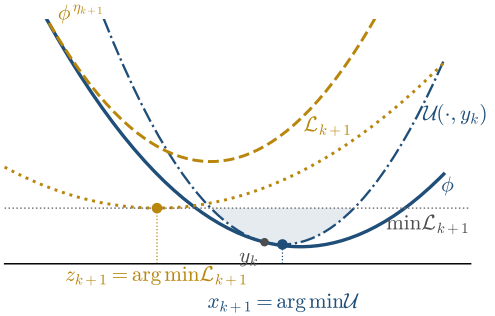

Figure 2: The model minimizers are x_{k+1} and z_{k+1}. An upper model (9) whose minimum value is below $\min_x \mathcal{L}_{k+1}(x)$ (16) certifies envelope descent.

Algorithm 2: ABrA-GD — Adaptive Bregman Accelerated Gradient Descent

Input: x_0, $\mu \ge 0$, ϕ, initial estimates $L_{\text{cur}} = L_0 > 0$ and $M_{-1} > 0$ (default $L_0 = 1$, $M_{-1} = L_0$)

Initialization: $x_0^{\text{best}} = z_0 = x_0$, $\lambda_0 = 0$, $\eta_0 = 0$, $\phi_0^{\text{best}} = \phi(x_0)$, $t_0 = 1$;

for $k = 0, 1, 2, \ldots$ **do**

$M_k \leftarrow \max\{M_{k-1}/4, \mu\}, \quad L_{\text{cur}} \leftarrow \max\{L_{\text{cur}}/2,\ \mu\}, \quad \phi_{k+1}^{\text{best}} \leftarrow \phi_k^{\text{best}}, \quad x_{k+1}^{\text{best}} \leftarrow x_k^{\text{best}}$;

repeat

$$M_k \leftarrow 2M_k, \qquad t_{k\ge 1} = \min\left\{1,\ \frac{2(\eta_k^{-1} + \mu)}{\eta_k^{-1} + \sqrt{\eta_k^{-2} + 4M_k(\eta_k^{-1} + \mu)}}\right\},$$

$$\eta_{k+1}^{-1} = M_k t_k^2 - \mu, \qquad \tau_k = \frac{t_k - \mu/\sqrt{M_k L_{\text{cur}}}}{1 - \mu/\sqrt{M_k L_{\text{cur}}}}.$$

$(x_{k+1}, z_{k+1}, \lambda_{k+1}, L_{\text{cur}}, \phi_k^{\text{low}}) = \text{BREG-PD-STEP}(x_k^{\text{best}}, z_k, \lambda_k, \eta_{k+1}, L_{\text{cur}}, t_k, \tau_k)$.

if $\phi_{k+1}^{\text{best}} > \phi(x_{k+1})$ **then** $\phi_{k+1}^{\text{best}} \leftarrow \phi(x_{k+1}), \quad x_{k+1}^{\text{best}} \leftarrow x_{k+1}$;

until $\phi_{k+1}^{\text{best}} \le \mathcal{L}_{k+1}(z_{k+1}) := (1 - t_k)\phi_k^{\text{best}} + t_k\phi_k^{\text{low}} - (\mu + \eta_{k+1}^{-1})D_d(z_k, z_{k+1})$;

Algorithm 3: BREG-PD-STEP$(x, z, \lambda, \eta, L_{\text{cur}}, t, \tau)$

Input: $x, z, \lambda, \eta, L_{\text{cur}}, t, \tau$

$$y = (1 - \tau)x + \tau z, \qquad \phi^{\text{low}} = \phi(y) + \langle \nabla\phi(y), z - y\rangle + \mu D_d(z, y).$$

Update

$$(x_{\text{new}}, L_{\text{cur}}) = \text{BACKTRACKING-GD}\big(y, \phi(y), \nabla\phi(y), L_{\text{cur}}\big), \qquad \text{(Primal Step)}$$

$$\lambda_{\text{new}} = (1 - t)\lambda + t\Big(\nabla\phi(y) - \mu(\nabla d(y) - \nabla d(x_0))\Big), \qquad \text{(Dual Step)}$$

$$z_{\text{new}} = \arg\min_u \left\{\langle \lambda_{\text{new}}, u\rangle + \left(\mu + \eta^{-1}\right) D_d(u, x_0)\right\}. \qquad \text{(Primal Map)}$$

return $(x_{\text{new}}, z_{\text{new}}, \lambda_{\text{new}}, L_{\text{cur}}, \phi^{\text{low}})$;

Algorithm 4: BACKTRACKING-GD$(y, \phi(y), \nabla\phi(y), L_{\text{cur}})$

Input: y, $\phi(y)$, $\nabla\phi(y)$, trial constant $L_{\text{cur}} > 0$.

Initialization: $L_{\text{new}} = L_{\text{cur}}/2$

repeat

$$L_{\text{new}} \leftarrow 2L_{\text{new}}, \quad x_{\text{new}} = \arg\min_x \left\{\langle \nabla\phi(y), x\rangle + L_{\text{new}} D_d(x, y)\right\},$$

until $\phi(x_{\text{new}}) \le \phi(y) + \langle \nabla\phi(y), x_{\text{new}} - y\rangle + L_{\text{new}} D_d(x_{\text{new}}, y) = \phi(y) - L_{\text{new}} D_d(y, x_{\text{new}})$;

return $(x_{\text{new}}, L_{\text{new}})$;

4. Parameters t_k, η_k, and M_k. For $k \geq 1$, the weight t_k and regularization η_{k+1} satisfy

$$\eta_{k+1}^{-1} = (1 - t_k)\eta_k^{-1}, \qquad \mu + \eta_{k+1}^{-1} = M_k t_k^2. \tag{13}$$

The first identity matches the weight on the previous dual model, enabling the recursive certificate below. Together, the identities give $M_k t_k^2 = (1 - t_k)(\mu + \eta_k^{-1}) + t_k\mu$, the quadratic relation in Nesterov's method (Nesterov, 2004). ABrA-GD selects M_k by backtracking and computes the positive root t_k. The accepted M_k controls the rate; see Theorem 2.

5. A computable certificate. Evaluating $\mathcal{D}^{\eta_{k+1}}(\lambda_{k+1})$ requires $f^*(\lambda_{k+1})$, which is not tractable in general. This step builds a computable lower bound $\mathcal{L}_{k+1}(z_{k+1}) \leq \mathcal{D}^{\eta_{k+1}}(\lambda_{k+1})$. The exit condition of algorithm 2 checks $\phi_{k+1}^{\text{best}} \leq \mathcal{L}_{k+1}(z_{k+1})$, which gives (7) at step $k + 1$.

Lemma 3 gives the construction; we sketch it. The affine bound ℓ_λ of (4) is concave in λ, so the dual update (11) and the first relation in (13) split $\mathcal{D}^{\eta_{k+1}}(\lambda_{k+1})$ into two terms. Relative strong convexity gives the first, and the Fenchel equality with the three-point identity gives the second:

$$\begin{aligned}\mathcal{D}^{\eta_{k+1}}(\lambda_{k+1}) \geq \min_x \Big\{ &(1 - t_k)\boxed{\mathcal{D}^{\eta_k}(\lambda_k) + (\mu + \eta_k^{-1})D_d(x, z_k)} \\ &+ t_k\boxed{\phi(y_k) + \langle \nabla\phi(y_k), x - y_k\rangle + \mu D_d(x, y_k)}\Big\}.\end{aligned} \tag{14}$$

By the invariant (7) at step k, $\mathcal{D}^{\eta_k}(\lambda_k) \geq \phi_k^{\text{best}}$. Replacing $\mathcal{D}^{\eta_k}(\lambda_k)$ by ϕ_k^{best} in the first box of (14) therefore gives $\mathcal{D}^{\eta_{k+1}}(\lambda_{k+1}) \geq \min_x \mathcal{L}_{k+1}(x)$, where $\mathcal{L}_{k+1}$ is the lower model (fig. 2)

$$\mathcal{L}_{k+1}(x) := (1 - t_k)\boxed{\phi_k^{\text{best}} + (\mu + \eta_k^{-1})D_d(x, z_k)} + t_k\boxed{\phi(y_k) + \langle \nabla\phi(y_k), x - y_k\rangle + \mu D_d(x, y_k)}. \tag{15}$$

Both boxes use only ϕ_k^{best}, z_k and one gradient at y_k, so $\mathcal{L}_{k+1}$ is computable. Finally, lemma 3 shows that $\mathcal{L}_{k+1}$ is minimized at z_{k+1}. Writing ϕ_k^{low} for the second term of (15) at $x = z_k$, we obtain the right-hand side of the exit condition of algorithm 2:

$$\mathcal{L}_{k+1}(z_{k+1}) = (1 - t_k)\phi_k^{\text{best}} + t_k\phi_k^{\text{low}} - (\mu + \eta_{k+1}^{-1})D_d(z_k, z_{k+1}). \tag{16}$$

Convergence in terms of the accepted parameters

The exit condition preserves the envelope descent requirement. The following theorem expresses the resulting rate in terms of the accepted acceleration parameters M_k, without imposing an additional condition on the geometry. *The theorem does not bound the accepted parameters themselves, nor show that the inner loop terminates*. Sections 4 and 5 give accelerated rates under bounded local distortion.

Theorem 2 (A posteriori convergence of ABrA-GD)**.** *Under As. 1, assume that the inner loop of algorithm 2 exits at every iteration. Let $M_k > \mu$ and choose $t_k \in (0, 1]$ by*

$$\mu + \eta_{k+1}^{-1} = M_k t_k^2 \quad (k \geq 0), \qquad \eta_{k+1}^{-1} = (1 - t_k)\eta_k^{-1} \quad (k \geq 1), \qquad t_0 = 1, \quad \eta_1^{-1} = M_0 - \mu.$$

Then, for any $x_\star \in \arg\min_x \phi(x)$,

$$\phi_{k+1}^{\text{best}} \leq \phi_k^{\text{best}}, \qquad \phi_{k+1}^{\text{best}} \leq \phi_\star^{\eta_{k+1}}, \qquad \phi_{k+1}^{\text{best}} - \phi_\star \leq \eta_{k+1}^{-1} D_d(x_\star, x_0).$$

Moreover, the decay of η_k^{-1} is controlled by the acceleration parameters M_k:

Case	***No uniform bound on*** M_k	***If*** $M_i \leq M$ ***for all*** $i \geq K$
$\mu = 0$	$\eta_k^{-1} \leq \dfrac{4}{\left(\sum_{i=0}^{k-1} M_i^{-1/2}\right)^2}$	$\eta_k^{-1} \leq \dfrac{\eta_K^{-1}}{\left(1 + \frac{k-K}{2\sqrt{M\eta_K}}\right)^2}$
$\mu > 0$	$\eta_k^{-1} \leq (M_0 - \mu)\displaystyle\prod_{i=1}^{k-1}\left(1 - \sqrt{\frac{\mu}{M_i}}\right)$	$\eta_k^{-1} \leq \eta_K^{-1}\left(1 - \sqrt{\frac{\mu}{M}}\right)^{k-K}$.

4 Dual Bregman Length Distortion

We identify the local geometric distortion in the comparison between the gradient and dual-averaging steps. We define the dual BLDF, relate it to local norms, and give formulas for standard kernels. Section 5 uses this condition to obtain accelerated rates and compares them with ABPG.

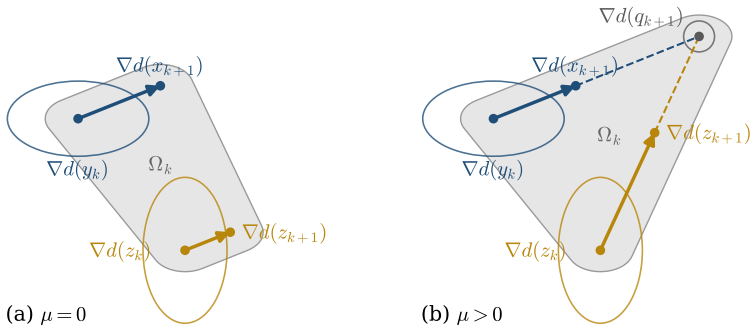

Figure 3: The two steps in dual coordinates $u = \nabla d(x)$ for the Burg kernel. The ellipses are the unit balls $\{\xi : \frac{1}{2}\langle \xi, \nabla^2 d^*(w)\,\xi\rangle \le 1\}$ of the local dual norm (definition 8) at the anchors w, drawn at a common scale; they change across the region, and ρ_{d^*} bounds this change on the shaded convex set, which contains Ω_k (definition 3). **(a)** When $\mu = 0$, the primal and dual steps are parallel displacements along $-\nabla\phi(y_k)$, from different anchors and with different scales. **(b)** When $\mu > 0$, $\nabla d(x_{k+1})$ and $\nabla d(z_{k+1})$ lie on two segments whose extensions meet at $\nabla d(q_{k+1})$; the shaded set contains the triangle they form with $\nabla d(y_k)$ and $\nabla d(z_k)$.

4.1 Comparing the gradient and dual-averaging steps (convex case)

The exit condition requires $\phi(x_{k+1}) \le \mathcal{L}_{k+1}(z_{k+1})$. We compare how far the upper model and the lower bound decrease from the common value $\phi(y_k)$. Here $\mu = 0$ and $\tau_k = t_k$. The primal step (line 3 of algorithm 4) and the exit test (line 5 of algorithm 2) give

$$\phi(x_{k+1}) \le \phi(y_k) - \underbrace{L_{\text{cur}} D_d(y_k, x_{k+1})}_{\text{primal decrease}}, \qquad \mathcal{L}_{k+1}(z_{k+1}) \ge \phi(y_k) - \underbrace{\eta_{k+1}^{-1} D_d(z_k, z_{k+1})}_{\text{dual decrease}},$$

where the second inequality uses $y_k = (1-t_k)x_k^{\text{best}} + t_k z_k$ and convexity, $(1-t_k)\phi_k^{\text{best}} + t_k\phi_k^{\text{low}} \ge \phi(y_k)$ (lemma 8). The exit condition is therefore guaranteed to hold when the upper model decreases at least as much as the lower bound:

$$\eta_{k+1}^{-1} D_d(z_k, z_{k+1}) \le L_{\text{cur}} D_d(y_k, x_{k+1}) \quad \Rightarrow \quad \phi(x_{k+1}) \le \mathcal{L}_{k+1}(z_{k+1}). \tag{17}$$

When $d = \frac{1}{2}\|\cdot\|^2$, both decreases are multiples of $\|\nabla\phi(y_k)\|^2$: $L_{\text{cur}}^{-1}\|\nabla\phi(y_k)\|^2/2$ and $t_k^2\eta_{k+1}\|\nabla\phi(y_k)\|^2/2$. Equality in (17) gives $\eta_{k+1}^{-1} = L_{\text{cur}} t_k^2$, the Nesterov coupling (appendix D). For a general d, the two decreases are Bregman divergences taken at different points and with different step lengths, so they cannot be compared term by term. In dual coordinates, however, with $g_k = \nabla\phi(y_k)$, the two steps are segments along the same direction:

$$\nabla d(x_{k+1}) = \nabla d(y_k) - L_{\text{cur}}^{-1} g_k, \qquad \nabla d(z_{k+1}) = \nabla d(z_k) - t_k\eta_{k+1} g_k.$$

Since $D_d(x, y) = D_{d^*}(\nabla d(y), \nabla d(x))$, the condition in (17) reads

$$\eta_{k+1}^{-1} D_{d^*}\big(\nabla d(z_k) - t_k\eta_{k+1} g_k,\ \nabla d(z_k)\big) \le L_{\text{cur}} D_{d^*}\big(\nabla d(y_k) - L_{\text{cur}}^{-1} g_k,\ \nabla d(y_k)\big).$$

Both sides measure the Bregman length of a segment along the same direction $-g_k$, from different anchors, $\nabla d(z_k)$ and $\nabla d(y_k)$, and with different scales, $t_k\eta_{k+1}$ and L_{cur}^{-1}.

Strongly convex case. When $\mu > 0$, the two dual segments point toward $\nabla d(q_{k+1})$, with q_{k+1} as in lemma 9, and are no longer parallel (fig. 3(b)). The proof of proposition 6 makes two comparisons, each between two segments with the same anchor and direction but different scales: from $\nabla d(y_k)$ to $\nabla d(x_{k+1})$ and to $\nabla d(q_{k+1})$, and likewise from $\nabla d(z_k)$.

4.2 The dual Bregman length distortion factor

In both cases, the comparison is between Bregman lengths of segments along a common direction, after a change of anchor, of scale, or both. The *local dual Bregman length distortion factor* bounds this change.

Definition 1 (Local Dual Bregman Length Distortion Factor (**Dual BLDF**)). *Let $\Omega \subset \operatorname{int}(\operatorname{dom} d^*)$, $s > 0$, and $\gamma > 0$. The* dual BLDF *of d^* on Ω is*

$$c_{d^*}(s, \Omega; \gamma) := \sup_{u_1, u_2, v \neq 0} \left(\frac{D_{d^*}(u_2 + sv, u_2)}{s^\gamma D_{d^*}(u_1 + v, u_1)} \right)^{1/2} \quad \text{s.t.} \quad \begin{cases} u_1 + [0,1]v \subset \Omega, \\ u_2 + [0,s]v \subset \Omega. \end{cases}$$

The numerator uses the displacement sv at anchor u_2, while the denominator uses v at anchor u_1. At $\gamma = 2$, the factor s^2 removes the quadratic scaling, leaving the change due to the geometry. Our convergence analysis uses $\gamma = 2$; other exponents serve only the comparison with triangle scaling (appendix E.2). If d^* is L_{d^*}-smooth and μ_{d^*}-strongly convex on Ω, then $c_{d^*}(s, \Omega; 2) \leq \sqrt{L_{d^*}/\mu_{d^*}}$ (lemma 5), so our rate then matches that of Auslender & Teboulle (2006) up to a constant. If d^* is twice differentiable, the dual BLDF is bounded by the variation of the dual local norm over Ω (proposition 1).

Definition 2 (Dual norm distortion factor). *Let $\Omega \subset \operatorname{int}(\operatorname{dom} d^*)$, and assume that d^* is twice differentiable on Ω. The* dual norm distortion factor *of d^* on Ω is*

$$\rho_{d^*}(\Omega) := \sup_{w_1, w_2 \in \Omega} \sup_{\xi \neq 0} \left(\frac{\langle \xi, \nabla^2 d^*(w_2)\, \xi \rangle}{\langle \xi, \nabla^2 d^*(w_1)\, \xi \rangle} \right)^{1/2} = \underbrace{\sup_{w_1, w_2 \in \Omega} \max_{1 \leq i \leq n} \left(\frac{(d_i^*)''((w_2)_i)}{(d_i^*)''((w_1)_i)} \right)^{1/2}}_{\text{if } d^* \text{ is separable (corollary 2)}}.$$

The factor $\rho_{d^*}(\Omega)$ measures the largest change in local norm across the region, after taking the supremum over directions. It has no scaling parameter and bounds the dual BLDF at $\gamma = 2$.

Proposition 1. *Assume that d^* is twice differentiable on $\Omega \subset \operatorname{int}(\operatorname{dom} d^*)$. Then, for every $s > 0$,*

$$c_{d^*}(s, \Omega; 2) \leq \rho_{d^*}(\Omega).$$

If Ω is open and $\nabla^2 d^*$ is continuous and positive definite on Ω, then $c_{d^*}(s, \Omega; 2) = \rho_{d^*}(\Omega)$ for every $s > 0$. More generally, $c_{d^*}(s, \Omega; \gamma) = s^{1-\gamma/2} \rho_{d^*}(\Omega)$ for every $s > 0$ and $\gamma > 0$ (proposition 4).

The rates of section 5 use ρ_{d^*}. Proposition 1 gives simple expressions for common geometries (table 2). For instance, for the Burg entropy, $\rho_{d^*}(\nabla d(\mathcal{X})) \leq b/a$ whenever $a \leq x \leq b$ for all $x \in \mathcal{X}$.

5 Accelerated Rates under Bounded Distortion

The accepted-parameter bound of Theorem 2 becomes an accelerated rate when the accepted M_k remain bounded. We now bound them using the local distortion ρ_k of each iteration.

Definition 3 (Local distortion). *Let Ω_k be the convex hull of the dual segments of all trials of the M_k-search in iteration k of algorithm 2. The* local distortion *of iteration k is $\rho_k := \rho_{d^*}(\Omega_k)$.*

5.1 Convergence and iteration complexity

Using the dual BLDF, propositions 5 and 6 show that every trial with $M_k \geq \rho_k^2 L$ ($\mu = 0$), or $M_k \geq \rho_k^4 L$ ($\mu > 0$), passes the exit test of algorithm 2. Hence, if $\rho_k \leq \rho$ from some iteration K on, backtracking eventually keeps M_k below $4\rho^2 L$, respectively $4\rho^4 L$.

Theorem 3 (Accelerated rate under bounded distortion). *Under As. 1, assume d^* is twice differentiable, and let ρ_k be as in definition 3. If there exist $K \geq 0$ and ρ such that the inner loop of algorithm 2 exits at every iteration $k < K$ and $\rho_k \leq \rho$ for all $k \geq K$, then it exits at every iteration, and there is a finite $\bar{K} \geq K$ such that, for all $k \geq \bar{K}$, we have $M_k \leq 4\rho^2 L$ ($\mu = 0$), respectively $M_k \leq 4\rho^4 L$ ($\mu > 0$). By Theorem 2, this implies $\phi_k^{\text{best}} - \phi_\star \leq \eta_k^{-1} D_d(x_\star, x_0)$, with*

$$\eta_k^{-1} \leq \frac{\eta_{\bar{K}}^{-1}}{\left(1 + \frac{k - \bar{K}}{4\rho\sqrt{L \eta_{\bar{K}}}}\right)^2} \ (\mu = 0), \qquad \eta_k^{-1} \leq \eta_{\bar{K}}^{-1} \left(1 - \sqrt{\tfrac{\mu}{4\rho^4 L}}\right)^{k - \bar{K}} \ (\mu > 0).$$

Corollary 1 (Iteration complexity). *Under Theorem 3, $\phi_k^{\text{best}} - \phi_\star \leq \varepsilon$ after N iterations from $\bar{K}$:*

$$N = O\left(\rho \sqrt{\frac{L\, D_d(x_\star, x_0)}{\varepsilon}}\right) \ (\mu = 0), \qquad N = O\left(\rho^2 \sqrt{\frac{L}{\mu}} \log \frac{\eta_{\bar{K}}^{-1} D_d(x_\star, x_0)}{\varepsilon}\right) \ (\mu > 0).$$

When $\rho = 1$, these bounds match the iteration complexities of Nesterov's fast gradient method up to an absolute constant; appendix D describes the Euclidean specialization.

Oracle complexity. Backtracking incurs constant amortized oracle overhead per accepted iteration, up to logarithmic terms for underestimated initial parameters. The complexity above therefore translates into an oracle complexity bound with these additional terms; see appendix C.4.

5.2 Comparison with existing methods

Inputs. The rates of Theorem 3 depend on ρ, but ABrA-GD uses neither ρ nor Ω_k: it finds M_k by backtracking on the exit condition, and the only problem constant it takes as input is μ. The accelerated linear rate of Chen et al. (2026) requires the constant of their condition as input. To our knowledge, Theorem 3 gives the first accelerated linear rate for relatively smooth and relatively strongly convex objectives obtained by a method with these inputs. Table 1 summarizes the comparison.

Triangle scaling and a posteriori bounds. ABPG converges at rate $O(k^{-\gamma})$ under a triangle-scaling property with exponent γ and gain 1 (Hanzely et al., 2021; Savchuk et al., 2024). For standard non-Euclidean kernels, this exponent is at most 1: $\gamma = 1$ for jointly convex divergences, and $\gamma \leq 1/2$ for the Burg kernel (appendix E). This guarantee therefore gives no acceleration, and some additional condition is necessary, since $O(1/k)$ is optimal for the full relatively smooth class (Dragomir et al., 2022). ABrA-GD shares this limit: if M_k grows without bound, y_k approaches x_k^{best}, the iterations approach Bregman gradient steps, and the safeguard of appendix G retains the $O(1/k)$ rate.

The accelerated variants instead fix $\gamma = 2$ and backtrack on the gain, so their rate $O(\bar{G}_k/k^2)$ depends on the accepted gains and holds a posteriori, as in Theorem 2. **Bounded distortion is implied by a uniform gain**: if triangle scaling holds with exponent 2 and a gain G uniform over triples, then $\rho_{d^*}^2 \leq G$ on the corresponding dual region (appendix E.2). Hence, whenever ABPG has a uniform gain, Theorem 3 also gives ABrA-GD an accelerated rate.

Method	Smooth. adaptive	Geometry adaptive	Convex rate	Acc. $\mu > 0$
BPG / BPG-LS	✓ (BPG-LS)	–	$O(1/k)$	–
ABPG	–	–	$O(k^{-\gamma})$, global γ	–
ABPG-e	–	✓ (exponent search)	$O(k^{-\gamma_k})$	–
ABPG-g	–	✓ (gain search)	$O(\bar{G}_k/k^2)$	–
Acc-MD (GCS)	–	–	$O(C/k^2)$	✓
ABrA-GD	✓	✓ (via M_k)	$O\Big((\sum_i M_i^{-1/2})^{-2}\Big)$	✓

Table 1: Comparison in the relatively smooth setting. BPG-LS denotes Bregman proximal gradient (-LS: with line search). ABPG-e uses an exponent search while ABPG-g uses a gain search. Acc-MD (Chen et al., 2026) takes the constant C of its generalized Cauchy–Schwarz condition as input.

6 Numerical Experiments

We evaluate ABrA-GD on D-optimal design and Poisson linear inverse problems. The formulations, reference functions, datasets, baselines, and initialization follow Hanzely et al. (2021). Further protocol details are given in appendix G.1, and additional comparisons appear in appendix G. Oracle calls count first-order oracle calls, including those of trials rejected by backtracking.

For D-optimal design, we use $\gamma = 2$ in the ABPG comparisons, following the experimental choices of Hanzely et al. (2021). With $\gamma = 1$, ABPG has only an $O(1/k)$ guarantee and converged more slowly than BPG in their experiments; fig. 5 reproduces this comparison.

Figure 4 shows the housing dataset. In the convex case, ABrA-GD improves on BPG and BPG-LS and performs similarly to the best ABPG variants over the plotted range. With $\mu = 10^{-4}$, its log-residual decreases approximately linearly, and it reaches a lower residual with fewer oracle calls than the displayed baselines. The other three D-optimal design datasets show the same pattern (appendix G). Figures 6 and 7 also plot the accepted M_k values. The D-optimal design instances are scaled so that $L = 1$ (Hanzely et al., 2021), and on these instances L_{cur} stays between $1/4$ and 2.

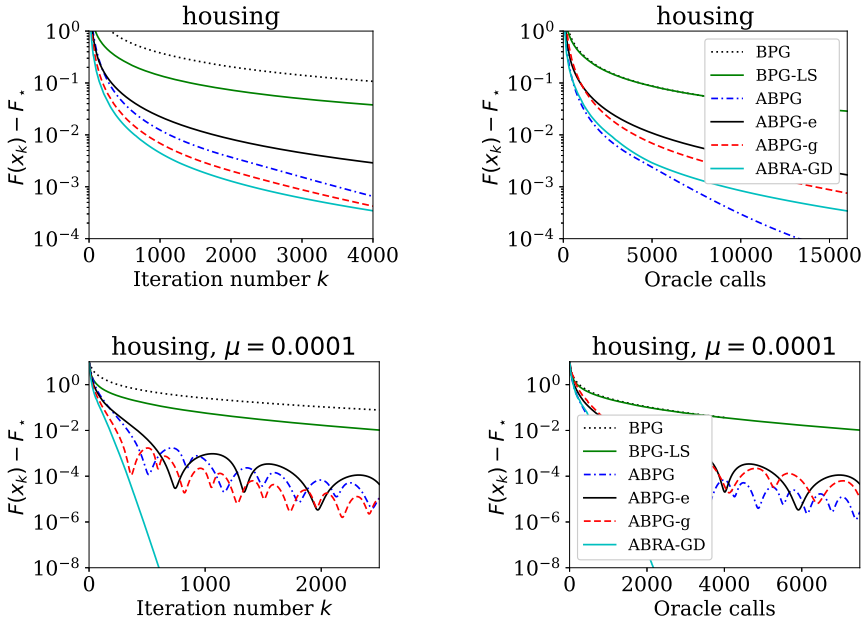

Figure 4: D-optimal design on the housing dataset. Panels show the objective residual $F - F_\star$ against iterations and oracle calls. Methods are BPG, BPG with line search (BPG-LS), ABPG, its exponent-adaptive and gain-adaptive variants (ABPG-e and ABPG-g), and ABrA-GD. ABPG and ABPG-g use $\gamma = 2$. Top: convex case ($\mu = 0$), where ABrA-GD performs similarly to the best ABPG variants. Bottom: relatively strongly convex case ($\mu = 10^{-4}$), where ABrA-GD reaches a lower residual and its log-residual decreases approximately linearly. Additional datasets and accepted parameters are reported in figs. 6 and 7; the $\gamma = 1$ comparison appears in fig. 5.

On the Poisson problems, ABrA-GD improves on the nonaccelerated baselines. Its performance relative to the accelerated methods depends on the instance and variant. We report Poisson L1 with $(m, n) = (200, 100)$ and Poisson L2 with $(m, n) = (100, 1000)$ in appendix G.

7 Conclusion

ABrA-GD combines a Bregman gradient step with a dual lower model to satisfy an envelope descent requirement. Its convergence bound depends on the acceleration parameters accepted during the run. The dual BLDF measures the local geometric distortion in the comparison between the two steps, but the algorithm does not need to evaluate it.

Under bounded distortion, the analysis gives an accelerated $O(1/k^2)$ rate for convex objectives and an accelerated linear rate for relatively strongly convex objectives. In the D-optimal design experiments, ABrA-GD performs similarly to adaptive ABPG in the convex case and reaches lower residuals in the relatively strongly convex case. Performance on the Poisson problems depends on the instance.

Future directions include sharper local distortion bounds, an accelerated analysis for composite objectives (appendix F), and stochastic variants.

AI use statement

We used generative AI tools to challenge and double-check the theoretical statements and their proofs, to draft and edit text, to search for and position related work, and to write code and generate figures. We did not use generative AI tools for the research ideas or to produce the main results: the theorems and their proofs were written by the authors before any AI-assisted checking. We reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

Alfred Auslender and Marc Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16(3):697–725, 2006.

Francis Bach. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1):115–129, 2015.

Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

Heinz H Bauschke, Minh N Dao, and Scott B Lindstrom. Regularizing with bregman–moreau envelopes. *SIAM Journal on Optimization*, 28(4):3208–3228, 2018.

Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

Yair Censor and Arnold Lent. An iterative row-action method for interval convex programming. *Journal of Optimization theory and Applications*, 34(3):321–353, 1981.

Yair Censor and Stavros Andrea Zenios. Proximal minimization algorithm with d-functions. *Journal of Optimization Theory and Applications*, 73(3):451–464, 1992.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011.

Long Chen, Hao Luo, Jingrong Wei, Zeyi Xu, and Yuan Yao. Accelerated mirror descent method through variable and operator splitting. *arXiv preprint arXiv:2601.19038*, 2026.

Jelena Diakonikolas and Lorenzo Orecchia. The approximate duality gap technique: A unified theory of first-order methods. *SIAM Journal on Optimization*, 29(1):660–689, 2019.

Radu-Alexandru Dragomir, Adrien B Taylor, Alexandre d'Aspremont, and Jérôme Bolte. Optimal complexity and certification of bregman first-order methods. *Mathematical Programming*, 194(1): 41–83, 2022.

Filip Hanzely, Peter Richtarik, and Lin Xiao. Accelerated bregman proximal gradient methods for relatively smooth convex optimization. *Computational Optimization and Applications*, 79(2): 405–440, 2021.

Sai Praneeth Karimireddy, Sebastian U. Stich, and Martin Jaggi. Global linear convergence of Newton's method without strong-convexity or Lipschitz gradients. *arXiv preprint arXiv:1806.00413*, 2018.

Claude Lemaréchal and Claudia Sagastizábal. Practical aspects of the moreau–yosida regularization: Theoretical preliminaries. *SIAM journal on optimization*, 7(2):367–385, 1997.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. *Advances in neural information processing systems*, 28, 2015.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. Catalyst acceleration for first-order convex optimization: from theory to practice. *Journal of Machine Learning Research*, 18(212):1–54, 2018.

Jin-Zan Liu and Xin-Wei Liu. A dual bregman proximal gradient method for relatively-strongly convex optimization. *Numerical Algebra, Control and Optimization*, 12(4):679–692, 2022. doi: 10.3934/naco.2021028.

Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

Renato DC Monteiro and Benar Fux Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013.

Yurii Nesterov. *Introductory lectures on convex optimization*. Springer, 2004.

Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

OS Savchuk, MS Alkousa, AS Shushko, AA Vyguzov, FS Stonyakin, DA Pasechnyuk, and AV Gasnikov. Accelerated bregman gradient methods for relatively smooth and relatively lipschitz continuous minimization problems. *arXiv preprint arXiv:2411.16743*, 2024.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2(3), 2008.

Ashia C Wilson, Ben Recht, and Michael I Jordan. A lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34, 2021.

CONTENTS

1 Setting and Introduction **1**

2 The Primal–Dual Envelope Framework **2**

2.1 Envelope Dual Problem and Lower Bound 3

3 Deriving ABrA-GD and Its Accepted-Parameter Rate **3**

4 Dual Bregman Length Distortion **5**

4.1 Comparing the gradient and dual-averaging steps (convex case) 6

4.2 The dual Bregman length distortion factor 6

5 Accelerated Rates under Bounded Distortion **7**

5.1 Convergence and iteration complexity 7

5.2 Comparison with existing methods 8

6 Numerical Experiments **8**

7 Conclusion **9**

A Notations and definitions **13**

A.1 The anchored decomposition 14

A.2 Affine equality constraints 14

B Recovering Bregman Gradient Descent and Dual Averaging **15**

B.1 Mirror Gradient Descent 15

B.2 Saddle formulation of the envelope 15

B.3 Dual Averaging 15

C Technical Lemmas **16**

C.1 Convergence rate of gradient descent and dual averaging 19

C.2 Technical results on BLDF 23

C.3 Technical results on Accelerated Method 26

C.4 Oracle complexity 33

D Euclidean specialization and comparison with Nesterov **34**

E Comparison with ABPG and triangle scaling **36**

E.1 Comparison with ABPG 36

E.2 Relation to the triangle-scaling property 36

F Composite objectives and the subgradient envelope **37**

G Numerical Experiments **38**
G.1 Experimental protocol 38
G.2 Safeguards and ablations 38
G.3 Additional numerical experiments 39
G.3.1 D-optimal design, convex, $\gamma = 1$ for ABPG 40
G.3.2 D-optimal design, convex 41
G.3.3 D-optimal design, strongly convex 42
G.3.4 Poisson problem 43
G.3.5 Restart comparison: D-optimal design, convex 44
G.3.6 Restart comparison: D-optimal design, strongly convex 45
G.3.7 Restart comparison: Poisson problem 46

A Notations and definitions

Definition 4 (Legendre function)**.** *A function d is* Legendre *if it is proper, closed and convex, and is both* essentially smooth *and* essentially strictly convex*: it is differentiable on* $\operatorname{int}(\operatorname{dom} d)$ *with* $\|\nabla d(x_k)\| \to \infty$ *whenever* $x_k \to \partial \operatorname{dom} d$*, and it is strictly convex on every convex subset of* $\operatorname{dom} \partial d$*. Essential strict convexity is what makes* ∇d *injective. In this case,* ∇d *is a bijection between* $\operatorname{int}(\operatorname{dom} d)$ *and* $\operatorname{int}(\operatorname{dom} d^*)$ *with inverse* ∇d^*.

Definition 5 (Bregman Divergence (Bregman, 1967))**.** *The* Bregman divergence *of a Legendre function d is*

$$D_d(x, y) := d(x) - d(y) - \langle \nabla d(y),\, x - y \rangle, \qquad x \in \operatorname{dom} d,\ y \in \operatorname{int}(\mathcal{X}).$$

Definition 6 (Relative smoothness and strong convexity)**.** *We say that f is μ-strongly convex (convex if $\mu = 0$) and L-smooth relative to a Legendre function d if there exists $0 \le \mu \le L$ s.t.*

$$\mu D_d(x, y) \le f(x) - f(y) - \langle \nabla f(y), x - y \rangle \le L D_d(x, y) \qquad \forall x, y \in \operatorname{int}(\mathcal{X}).$$

Definition 7 (Fenchel–Legendre conjugate (Rockafellar, 1997, §12))**.** *Let f be a proper function. Its conjugate is*

$$f^*(\lambda) := \sup_{x \in \mathbb{R}^n} \{\langle \lambda, x \rangle - f(x)\}.$$

Definition 8 (Primal and dual local norms)**.** *Let d be a twice differentiable Legendre function whose Hessian is positive definite on* $\operatorname{int}(\operatorname{dom} d)$*. For $x \in \operatorname{int}(\operatorname{dom} d)$, $w \in \operatorname{int}(\operatorname{dom} d^*)$, $v \in \mathcal{X}$, and $s \in \mathcal{X}^*$, define*

$$\|v\|_x^2 := \tfrac{1}{2}\langle \nabla^2 d(x)\, v,\, v \rangle, \qquad \|s\|_{w,*}^2 := \tfrac{1}{2}\langle s,\, \nabla^2 d^*(w)\, s \rangle.$$

Positive definiteness is a hypothesis, not a consequence of the rest: $d(x) = x^4/4$ is Legendre and twice differentiable on $\mathbb{R}$ with $d''(0) = 0$, so the first form is degenerate at the origin and d^* has no finite second derivative there. Under the hypothesis, $\nabla^2 d^*(w) = \nabla^2 d(\nabla d^*(w))^{-1}$ is positive definite as well and both forms are norms. Every statement below that uses local norms or ratios of Hessians carries this hypothesis on the region where it is applied.

The Bregman divergence admits an integral representation in terms of the local norms.

Proposition 2 (Bregman divergence and local norms)**.** *Let d be a Legendre function of class C^2 with positive definite Hessian on* $\operatorname{int}(\operatorname{dom} d)$*. For any $x_1, x_2 \in \operatorname{int}(\operatorname{dom} d)$, there is a point z on the segment $[x_1, x_2]$ such that*

$$D_d(x_1, x_2) = \|x_1 - x_2\|_z^2.$$

Proof. Since $\operatorname{int}(\operatorname{dom} d)$ is convex, the segment $[x_1, x_2]$ lies in it, and $g(t) := d\big(x_2 + t(x_1 - x_2)\big)$ is twice continuously differentiable on a neighbourhood of $[0, 1]$. Taylor's theorem with the Lagrange remainder gives a $\theta \in (0, 1)$ with $g(1) = g(0) + g'(0) + \frac{1}{2} g''(\theta)$, that is, writing $z := x_2 + \theta(x_1 - x_2)$,

$$d(x_1) - d(x_2) - \langle \nabla d(x_2),\, x_1 - x_2 \rangle = \tfrac{1}{2}\big\langle \nabla^2 d(z)(x_1 - x_2),\, x_1 - x_2 \big\rangle.$$

The left-hand side is $D_d(x_1, x_2)$ and the right-hand side is $\|x_1 - x_2\|_z^2$ by definition 8. □

A.1 THE ANCHORED DECOMPOSITION

For a differentiable g, write

$$D_g(x, y) := g(x) - g(y) - \langle \nabla g(y), x - y \rangle,$$

which is the Bregman divergence already defined for d. In this notation, g is L-smooth relative to d when $D_g \le L D_d$ and μ-strongly convex relative to d when $D_g \ge \mu D_d$, both on $\operatorname{int}(\operatorname{dom} d)$.

The anchor $x \mapsto \mu D_d(x, x_0)$ differs from μd by an affine function, and an affine function has zero Bregman divergence. So the anchor contributes exactly μD_d, and $\phi = f + \mu D_d(\cdot, x_0)$ satisfies

$$D_\phi(x, y) = D_f(x, y) + \mu D_d(x, y), \qquad x, y \in \operatorname{int}(\operatorname{dom} d). \tag{18}$$

Let ϕ be L-smooth and μ-strongly convex relative to d, that is $\mu D_d \le D_\phi \le L D_d$. Strong convexity says precisely that $\phi - \mu D_d(\cdot, x_0)$ is convex, so put

$$f := \phi - \mu D_d(\cdot, x_0).$$

Subtracting μD_d throughout (18) gives $0 \le D_f \le (L - \mu) D_d$: the function f is convex and $(L - \mu)$-smooth relative to d, which is As. 1, and ϕ is back in the form (1).

Conversely, if f satisfies As. 1 then $0 \le D_f \le (L - \mu) D_d$, and (18) returns $\mu D_d \le D_\phi \le L D_d$.

The two hypotheses are therefore the same one. All of the relative strong convexity of ϕ sits in the anchor term, and the anchor is what raises the relative smoothness constant from $L - \mu$ to L.

A.2 AFFINE EQUALITY CONSTRAINTS

The analysis also applies when the feasible set includes affine equality constraints. Let $\mathcal{A} \subseteq \mathbb{R}^n$ be an affine subspace, assume that $x_0 \in \mathcal{A} \cap \operatorname{int}(\operatorname{dom} d)$, and consider the feasible set

$$\mathcal{X} = \operatorname{cl}(\operatorname{dom} d) \cap \mathcal{A}.$$

We can eliminate the equality constraints by choosing coordinates on $\mathcal{A}$.

Let the columns of U form an orthonormal basis for the direction space of $\mathcal{A}$. Every $x \in \mathcal{A}$ has a unique representation $x = x_0 + Uu$. Define

$$\widetilde{f}(u) = f(x_0 + Uu), \qquad \widetilde{d}(u) = d(x_0 + Uu).$$

The Bregman divergence is preserved under this change of variables:

$$D_{\widetilde{d}}(u, v) = D_d(x_0 + Uu, x_0 + Uv).$$

The constrained problem therefore becomes

$$\min_{u \in \operatorname{cl}(\operatorname{dom} \widetilde{d})} \widetilde{f}(u) + \mu D_{\widetilde{d}}(u, 0).$$

The relative smoothness and strong convexity inequalities carry over directly, with the same constants L and μ.

The restricted kernel $\widetilde{d}$ is Legendre. To see this, write $V = \operatorname{range}(U)$. Since $\mathcal{A}$ intersects $\operatorname{int}(\operatorname{dom} d)$, the subdifferential sum rule gives

$$\partial(d + \delta_{\mathcal{A}})(x) = \partial d(x) + V^\perp$$

(Rockafellar, 1997, Thm. 23.8). At a boundary point of the restricted domain, relative to $\mathcal{A}$, the right-hand side is empty because d is Legendre. This gives essential smoothness in the reduced coordinates. Essential strict convexity is inherited from d.

Thus the problem in the variable u satisfies As. 1. All updates and proofs apply in these coordinates, so affine equality constraints require no change to the analysis.

B Recovering Bregman Gradient Descent and Dual Averaging

The implicit envelope of section 2 specializes to the two classical methods. Both derivations are given here: they motivate the accelerated step of algorithm 2 but are not needed to read it.

B.1 Mirror Gradient Descent

Since f is relatively smooth, for all x_k, we have the following upper model $\mathcal{U}(x, x_k)$ for f,

$$f(x) \leq \mathcal{U}(x, x_k) := f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + (L - \mu) D_d(x, x_k).$$

By definition of ϕ, it is sufficient to find an iterate x_{k+1} and a stepsize η_{k+1} such that

$$\mathcal{U}(x_{k+1}, x_k) + \mu D_d(x_{k+1}, x_0) \leq \phi_\star^{\eta_{k+1}}$$

Naturally, we can take the minimum of the upper bound and adjust η_{k+1} accordingly,

$$x_{k+1} = \arg\min_x \mathcal{U}(x, x_k) + \mu D_d(x, x_0) = \nabla d^* \left(\nabla d(x_k) - \frac{1}{L} \nabla \phi(x_k) \right).$$

This is exactly gradient descent, and it fits algorithm 1 with a certain schedule for η_k.

Proposition 3. *The Bregman gradient descent iterates satisfy algorithm 1 with*

$$\eta_k^{-1} = L \min \left\{ \frac{1}{k}, \frac{\kappa}{(1+\kappa)^k - 1} \right\}, \qquad \kappa = \frac{\mu}{L},$$

where the second term is omitted when $\mu = 0$.

The two branches are the known rates of the Bregman gradient method: the $O(1/k)$ rate under relative smoothness (Bauschke et al., 2017; Lu et al., 2018), and the linear rate under relative strong convexity (Lu et al., 2018). What the proposition adds is the translation into the envelope schedule, that is, the η_k for which the method satisfies the requirement of algorithm 1.

B.2 Saddle formulation of the envelope

Define

$$S^\eta(x, \lambda) := \langle x, \lambda \rangle - f^*(\lambda) + \left(\mu + \frac{1}{\eta} \right) D_d(x, x_0). \tag{19}$$

The Fenchel representation of f and the definition of $\mathcal{D}^\eta$ give

$$\phi^\eta(x) = \sup_\lambda S^\eta(x, \lambda), \qquad \mathcal{D}^\eta(\lambda) = \min_x S^\eta(x, \lambda).$$

Weak duality therefore yields

$$\sup_\lambda \mathcal{D}^\eta(\lambda) \leq \min_x \sup_\lambda S^\eta(x, \lambda) = \phi_\star^\eta.$$

For $\lambda = \nabla f(y)$, Fenchel equality gives

$$S^\eta(x, \nabla f(y)) = f(y) + \langle \nabla f(y), x - y \rangle + \left(\mu + \frac{1}{\eta} \right) D_d(x, x_0),$$

With $\eta^{-1} = 0$, this is the second term of the lower model $\mathcal{L}_{k+1}$ of (15), at $y = y_k$.

This formulation is helpful for proving the convergence rate of Dual Averaging, and therefore, AbrA-GD.

B.3 Dual Averaging

Even when $S^\eta(x, \lambda)$ cannot be evaluated explicitly, we can still update the dual variable through the best-response (**BR**) maps

$$z^\eta(\lambda) \in \arg\min_x S^\eta(x, \lambda) = \nabla d^* \left(\nabla d(x_0) - \frac{1}{\mu + \eta^{-1}} \lambda \right), \tag{Primal BR}$$

$$\lambda(x) \in \arg\max_\lambda S^\eta(x, \lambda) = \nabla f(x). \tag{Dual BR}$$

The dual-averaging update is the damped best-response-to-best-response iteration

$$\begin{aligned}\lambda_{k+1} &= (1-t_k)\lambda_k + t_k\lambda(z_k) = (1-t_k)\lambda_k + t_k\nabla f(z_k),\\ z_{k+1} &= z^{\eta_{k+1}}(\lambda_{k+1}).\end{aligned} \tag{Dual Avg.}$$

This can be viewed as a damped alternating min-max scheme. Its main advantage is that it avoids computing $\nabla f^*(\lambda)$, which is usually intractable.

Theorem 4. *The update (Dual Avg.), initialized with $\eta_0 = 0$, $\lambda_0 = 0$ and $z_0 = x_0$, and using*

$$t_k = \frac{\mu + \eta_k^{-1}}{L + \eta_k^{-1}}, \qquad \eta_{k+1}^{-1} = (1-t_k)\eta_k^{-1},$$

satisfies algorithm 1 with

$$x_k \in \operatorname*{arg\,min}_{1\le i\le k} \phi(z_i), \quad \eta_k^{-1} \le L\min\left\{\frac{1}{k}, \frac{\kappa}{(1+\kappa)^k - 1}\right\}, \qquad \kappa = \frac{\mu}{L},$$

where the second term is omitted when $\mu = 0$.

C Technical Lemmas

Lemma 1 (Three-point property of Tseng (Tseng, 2008))**.** *Let $g : \mathcal{X} \to \mathbb{R}\cup\{+\infty\}$ be a convex function, and let $D_d(\cdot,\cdot)$ be the Bregman divergence induced by d. For a given $z \in \mathcal{X}$, define*

$$z_+ \;:=\; \operatorname*{arg\,min}_{x\in\mathcal{X}} \big\{g(x) + D_d(x,z)\big\}.$$

Then, for all $x \in \mathcal{X}$,

$$g(x) + D_d(x,z) \;\ge\; g(z_+) + D_d(z_+,z) + D_d(x,z_+).$$

Note that when the minimization is unconstrained and g is affine, the inequality holds with equality,

$$g(x) + D_d(x,z) \;=\; g(z_+) + D_d(z_+,z) + D_d(x,z_+). \tag{20}$$

Lemma 2 (Growth around the minimizer)**.** *Let g be differentiable and α-strongly convex relative to d, and let $z := \arg\min_x g(x)$ lie in the interior of the domain of d. Then, for all x,*

$$g(x) \;\ge\; g(z) + \alpha D_d(x,z).$$

If $g - \alpha d$ is affine, the inequality holds with equality.

Proof. The function $h := g - \alpha d$ is convex, so $h(x) \ge h(z) + \langle \nabla h(z), x - z\rangle$, with equality when h is affine. Since z is an interior minimizer of g, $\nabla g(z) = 0$ and $\nabla h(z) = -\alpha\nabla d(z)$. Adding $\alpha d(x) = \alpha d(z) + \alpha\langle \nabla d(z), x - z\rangle + \alpha D_d(x,z)$ gives the result. □

The following generic lemma is important in the proof of Dual Averaging and Accelerated Gradient Descent.

Lemma 3 (Dual lower-bound recursion)**.** *Under As. 1, so that ϕ is L-smooth and μ-strongly convex relative to d, assume that, for some b_k,*

$$b_k \le \mathcal{D}^{\eta_k}(\lambda_k), \qquad \lambda_{k+1} = (1-t_k)\lambda_k + t_k\nabla f(y_k), \qquad \eta_{k+1}^{-1} = (1-t_k)\eta_k^{-1},$$

and let

$$z_k \in \operatorname*{arg\,min}_x S^{\eta_k}(x,\lambda_k), \qquad z_{k+1} \in \operatorname*{arg\,min}_x S^{\eta_{k+1}}(x,\lambda_{k+1}),$$

where S^η is the saddle point problem defined in (19). Define

$$\phi_k^{\mathrm{low}} := \phi(y_k) + \langle \nabla\phi(y_k), z_k - y_k\rangle + \mu D_d(z_k, y_k).$$

Then

$$\mathcal{D}^{\eta_{k+1}}(\lambda_{k+1}) \ge (1-t_k)b_k + t_k\phi_k^{\mathrm{low}} - \big(\mu + \eta_{k+1}^{-1}\big)\, D_d(z_k, z_{k+1}).$$

If $t_k = 1$, the same conclusion holds with the term $(1-t_k)b_k$ omitted, for any $\eta_{k+1} > 0$, and without the hypothesis $b_k \le \mathcal{D}^{\eta_k}(\lambda_k)$ or the recursion $\eta_{k+1}^{-1} = (1-t_k)\eta_k^{-1}$:

$$\mathcal{D}^{\eta_{k+1}}(\lambda_{k+1}) \ge \phi_k^{\mathrm{low}} - \big(\mu + \eta_{k+1}^{-1}\big)\, D_d(z_k, z_{k+1}).$$

Proof. By concavity of $S^\eta(x,\cdot)$, evaluated at $x = z_{k+1}$,

$$\mathcal{D}^{\eta_{k+1}}(\lambda_{k+1}) = S^{\eta_{k+1}}(z_{k+1},\lambda_{k+1}) \geq (1-t_k)S^{\eta_{k+1}}(z_{k+1},\lambda_k) + t_k S^{\eta_{k+1}}(z_{k+1},\nabla f(y_k)).$$

Using

$$S^{\eta_{k+1}}(x,\lambda_k) = S^{\eta_k}(x,\lambda_k) + (\eta_{k+1}^{-1} - \eta_k^{-1})D_d(x,x_0),$$
$$S^{\eta_k}(x,\lambda_k) \geq b_k + (\mu+\eta_k^{-1})D_d(x,z_k),$$

where the last inequality follows directly from the optimality of z_k. Indeed, with $\alpha_k := \mu + \eta_k^{-1}$,

$$\begin{aligned} S^{\eta_k}(x,\lambda_k) - S^{\eta_k}(z_k,\lambda_k) &= \big\langle \lambda_k + \alpha_k\big(\nabla d(z_k) - \nabla d(x_0)\big), x - z_k\big\rangle + \alpha_k D_d(x,z_k) \\ &\geq \alpha_k D_d(x,z_k), \end{aligned}$$

and $S^{\eta_k}(z_k,\lambda_k) = \mathcal{D}^{\eta_k}(\lambda_k) \geq b_k$. This is the only use of the hypothesis on b_k, and it enters the conclusion through the factor $(1-t_k)$. Finally, since $\eta_{k+1}^{-1} = (1-t_k)\eta_k^{-1}$, the coefficients of $D_d(x,x_0)$ sum to

$$(1-t_k)(\eta_{k+1}^{-1} - \eta_k^{-1}) + t_k(\mu + \eta_{k+1}^{-1}) = t_k\mu.$$

The identity $D_d(x,x_0) = D_d(y_k,x_0) + \langle \nabla d(y_k) - \nabla d(x_0), x - y_k\rangle + D_d(x,y_k)$ turns $t_k\big(f(y_k) + \langle \nabla f(y_k), x - y_k\rangle + \mu D_d(x,x_0)\big)$ into $t_k\big(\phi(y_k) + \langle \nabla\phi(y_k), x - y_k\rangle + \mu D_d(x,y_k)\big)$. Hence

$$\mathcal{D}^{\eta_{k+1}}(\lambda_{k+1}) \geq (1-t_k)b_k + t_k\phi(y_k) + H_k(z_{k+1}),$$

where

$$H_k(x) := (1-t_k)(\mu+\eta_k^{-1})D_d(x,z_k) + t_k\mu D_d(x,y_k) + t_k\langle \nabla\phi(y_k), x - y_k\rangle.$$

We claim that $z_{k+1} \in \arg\min_x H_k(x)$. Indeed, using

$$\lambda_k + (\mu+\eta_k^{-1})(\nabla d(z_k) - \nabla d(x_0)) = 0,$$
$$\lambda_{k+1} = (1-t_k)\lambda_k + t_k\nabla f(y_k), \qquad \nabla f(y_k) = \nabla\phi(y_k) - \mu(\nabla d(y_k) - \nabla d(x_0)),$$

the optimality condition of

$$z_{k+1} \in \arg\min_x S^{\eta_{k+1}}(x,\lambda_{k+1})$$

is equivalent to

$$0 = (1-t_k)(\mu+\eta_k^{-1})\big(\nabla d(z_{k+1}) - \nabla d(z_k)\big) + t_k\mu\big(\nabla d(z_{k+1}) - \nabla d(y_k)\big) + t_k\nabla\phi(y_k),$$

which is exactly the optimality condition for $z_{k+1} \in \arg\min_x H_k(x)$.

Moreover, H_k has Bregman curvature

$$(1-t_k)(\mu+\eta_k^{-1}) + t_k\mu = \mu + \eta_{k+1}^{-1}.$$

Therefore H_k is an affine function plus a weighted sum of Bregman divergences in its first argument, with total weight $\mu + \eta_{k+1}^{-1}$, and z_{k+1} is its unconstrained minimizer. (20) applies and gives, at $x = z_k$,

$$H_k(z_{k+1}) = H_k(z_k) - (\mu+\eta_{k+1}^{-1})D_d(z_k,z_{k+1}).$$

The three-point property of lemma 1 would give this inequality in the opposite direction; here it holds as an equality because the minimization is unconstrained. Since

$$H_k(z_k) = t_k\mu D_d(z_k,y_k) + t_k\langle \nabla\phi(y_k), z_k - y_k\rangle,$$

we obtain

$$\begin{aligned} \mathcal{D}^{\eta_{k+1}}(\lambda_{k+1}) \geq (1-t_k)b_k + t_k\phi(y_k) + t_k\langle \nabla\phi(y_k), z_k - y_k\rangle \\ + t_k\mu D_d(z_k,y_k) - (\mu+\eta_{k+1}^{-1})D_d(z_k,z_{k+1}). \end{aligned}$$

By definition,

$$\phi_k^{\mathrm{low}} = \phi(y_k) + \langle \nabla\phi(y_k), z_k - y_k\rangle + \mu D_d(z_k,y_k),$$

and therefore

$$\mathcal{D}^{\eta_{k+1}}(\lambda_{k+1}) \geq (1-t_k)b_k + t_k\phi_k^{\mathrm{low}} - (\mu+\eta_{k+1}^{-1})D_d(z_k,z_{k+1}).$$

For the case $t_k = 1$, argue directly. Then $\lambda_{k+1} = \nabla f(y_k)$, so the concavity step above is unnecessary and neither λ_k nor η_k appears. Writing $\alpha_{k+1} = \mu + \eta_{k+1}^{-1}$ and using that the conjugate is attained at y_k,

$$\mathcal{D}^{\eta_{k+1}}(\lambda_{k+1}) = S^{\eta_{k+1}}(z_{k+1}, \nabla f(y_k)) = f(y_k) + \langle \nabla f(y_k), z_{k+1} - y_k \rangle + \alpha_{k+1} D_d(z_{k+1}, x_0).$$

Substituting $\nabla f(y_k) = \nabla \phi(y_k) - \mu(\nabla d(y_k) - \nabla d(x_0))$ and $f(y_k) = \phi(y_k) - \mu D_d(y_k, x_0)$, and expanding $\mu D_d(z_{k+1}, x_0)$ by the three-point identity around y_k, the terms anchored at x_0 combine into $\mu D_d(z_{k+1}, y_k)$:

$$\mathcal{D}^{\eta_{k+1}}(\lambda_{k+1}) = \phi(y_k) + \widetilde{H}_k(z_{k+1}), \qquad \widetilde{H}_k(x) := \eta_{k+1}^{-1} D_d(x, x_0) + \mu D_d(x, y_k) + \langle \nabla \phi(y_k), x - y_k \rangle.$$

The optimality condition for $z_{k+1} \in \arg\min_x S^{\eta_{k+1}}(x, \lambda_{k+1})$ is exactly the one for $z_{k+1} \in \arg\min_x \widetilde{H}_k(x)$, and $\widetilde{H}_k$ is affine plus Bregman divergences of total weight $\eta_{k+1}^{-1} + \mu = \alpha_{k+1}$, so (20) at $x = z_k$ gives $\widetilde{H}_k(z_{k+1}) = \widetilde{H}_k(z_k) - \alpha_{k+1} D_d(z_k, z_{k+1})$. Since $\phi(y_k) + \widetilde{H}_k(z_k) = \phi_k^{\mathrm{low}} + \eta_{k+1}^{-1} D_d(z_k, x_0)$ and $D_d \geq 0$,

$$\mathcal{D}^{\eta_{k+1}}(\lambda_{k+1}) = \phi_k^{\mathrm{low}} + \eta_{k+1}^{-1} D_d(z_k, x_0) - \alpha_{k+1} D_d(z_k, z_{k+1}) \geq \phi_k^{\mathrm{low}} - \alpha_{k+1} D_d(z_k, z_{k+1}).$$

At $k = 0$ the initialization gives $z_0 = x_0$, so the middle term is zero and the bound holds with equality. □

C.1 Convergence rate of gradient descent and dual averaging

Proof of Theorem 1. By construction, algorithm 1 produces x_k satisfying $\phi(x_k) \le \phi_\star^{\eta_k}$. Bounding the minimum in (2) by its value at $x^\star$,

$$\phi_\star^{\eta_k} = \min_x \left\{ \phi(x) + \frac{1}{\eta_k} D_d(x, x_0) \right\} \;\le\; \phi(x^\star) + \frac{1}{\eta_k} D_d(x^\star, x_0) = \phi_\star + \frac{D_d(x_\star, x_0)}{\eta_k}.$$

Chaining the two inequalities gives (3). □

Proof of proposition 3. The optimality condition of the Bregman gradient step is

$$\nabla\phi(x_k) + L\big(\nabla d(x_{k+1}) - \nabla d(x_k)\big) = 0.$$

The objective of this step is $\langle \nabla f(x_k), x\rangle + L D_d(x, w_k)$ up to a constant, with $\nabla d(w_k) = (1-\kappa)\nabla d(x_k) + \kappa \nabla d(x_0)$ and $\kappa = \mu/L$. By relative $(L-\mu)$-smoothness of f, convexity of f, and (20), we obtain

$$\phi(x_{k+1}) \le f(x) + \mu D_d(x, x_0) + (L-\mu) D_d(x, x_k) - L D_d(x, x_{k+1}). \tag{21}$$

Let

$$r := \frac{L}{L-\mu} = \frac{1}{1-\kappa}.$$

Multiplying (21) by r^{k+1} and using

$$(L-\mu) r^{k+1} = L r^k,$$

gives

$$r^{k+1}\phi(x_{k+1}) \le r^{k+1}\big(f(x) + \mu D_d(x, x_0)\big) + L\big(r^k D_d(x, x_k) - r^{k+1} D_d(x, x_{k+1})\big).$$

Summing from $i = 0$ to k yields

$$\sum_{i=0}^{k} r^{i+1}\phi(x_{i+1}) \le \left(\sum_{i=0}^{k} r^{i+1}\right)\big(f(x) + \mu D_d(x, x_0)\big) + L\big(D_d(x, x_0) - r^{k+1} D_d(x, x_{k+1})\big).$$

Dropping the last negative term and using monotonicity of $\phi(x_i)$, we get

$$\phi(x_{k+1}) \le f(x) + \left(\mu + \frac{L}{\sum_{i=0}^{k} r^{i+1}}\right) D_d(x, x_0).$$

Therefore x_{k+1} satisfies algorithm 1 with

$$\eta_{k+1}^{-1} = \frac{L}{\sum_{i=0}^{k} r^{i+1}}.$$

If $\mu = 0$, then $r = 1$, so

$$\eta_{k+1}^{-1} = \frac{L}{k+1}.$$

If $0 < \mu < L$, then

$$\eta_{k+1}^{-1} = L\frac{\kappa}{r^{k+1} - 1} = L\frac{\kappa(1-\kappa)^{k+1}}{1 - (1-\kappa)^{k+1}}.$$

Since $r = (1-\kappa)^{-1} \ge 1 + \kappa$, we have

$$\eta_{k+1}^{-1} \le L\frac{\kappa}{(1+\kappa)^{k+1} - 1}.$$

Thus the simpler valid rate is

$$\eta_{k+1}^{-1} \le L \min\left\{ \frac{1}{k+1}, \frac{\kappa}{(1+\kappa)^{k+1} - 1} \right\}.$$

□

Lemma 4 (Dual-averaging one-step recursion)**.** *Let*

$$\mathcal{D}_k := \mathcal{D}^{\eta_k}(\lambda_k), \qquad z_k = z^{\eta_k}(\lambda_k),$$

and assume

$$\lambda_{k+1} = (1-t_k)\lambda_k + t_k \nabla f(z_k), \qquad z_{k+1} = z^{\eta_{k+1}}(\lambda_{k+1}).$$

Choose

$$t_k = \frac{\mu + \eta_k^{-1}}{L + \eta_k^{-1}}, \qquad \eta_{k+1}^{-1} = (1-t_k)\eta_k^{-1}.$$

Then

$$\mathcal{D}_{k+1} \geq (1-t_k)\mathcal{D}_k + t_k \phi(z_{k+1}).$$

Moreover, if $t_k \in [0,1]$, $x_{k+1} \in \arg\min\{\phi(x_k), \phi(z_{k+1})\}$, *and*

$$\phi(x_k) \leq \mathcal{D}_k,$$

then

$$\phi(x_{k+1}) \leq \mathcal{D}_{k+1}.$$

Proof. By concavity of $S^\eta(x,\cdot)$, for every x,

$$S^{\eta_{k+1}}(x, \lambda_{k+1}) \geq (1-t_k) S^{\eta_{k+1}}(x, \lambda_k) + t_k S^{\eta_{k+1}}(x, \nabla f(z_k)).$$

Using

$$S^{\eta_{k+1}}(x, \lambda_k) = S^{\eta_k}(x, \lambda_k) + (\eta_{k+1}^{-1} - \eta_k^{-1}) D_d(x, x_0),$$

and $z_k = z^{\eta_k}(\lambda_k)$, we have

$$S^{\eta_k}(x, \lambda_k) \geq \mathcal{D}_k + (\mu + \eta_k^{-1}) D_d(x, z_k).$$

By relative $(L-\mu)$-smoothness of f,

$$f(z_k) + \langle \nabla f(z_k), x - z_k \rangle \geq f(x) - (L-\mu) D_d(x, z_k),$$

hence

$$S^{\eta_{k+1}}(x, \nabla f(z_k)) \geq \phi(x) + \eta_{k+1}^{-1} D_d(x, x_0) - (L-\mu) D_d(x, z_k).$$

Combining gives

$$\begin{aligned} S^{\eta_{k+1}}(x, \lambda_{k+1}) \geq{}& (1-t_k)\mathcal{D}_k + t_k \phi(x) \\ &+ \Big[(1-t_k)(\mu + \eta_k^{-1}) - t_k(L-\mu)\Big] D_d(x, z_k) \\ &+ \Big[\eta_{k+1}^{-1} - (1-t_k)\eta_k^{-1}\Big] D_d(x, x_0). \end{aligned}$$

With the chosen t_k and η_{k+1}, both Bregman coefficients vanish, so

$$S^{\eta_{k+1}}(x, \lambda_{k+1}) \geq (1-t_k)\mathcal{D}_k + t_k \phi(x).$$

Taking $x = z_{k+1} = z^{\eta_{k+1}}(\lambda_{k+1})$ gives

$$\mathcal{D}_{k+1} \geq (1-t_k)\mathcal{D}_k + t_k \phi(z_{k+1}).$$

Now assume $\phi(x_k) \leq \mathcal{D}_k$. Since

$$x_{k+1} \in \arg\min\{\phi(x_k), \phi(z_{k+1})\},$$

and $t_k \in [0,1]$,

$$\phi(x_{k+1}) \leq (1-t_k)\phi(x_k) + t_k \phi(z_{k+1}) \leq (1-t_k)\mathcal{D}_k + t_k \phi(z_{k+1}) \leq \mathcal{D}_{k+1}.$$

□

Proof of Theorem 4. Let

$$\mathcal{D}_k := \mathcal{D}^{\eta_k}(\lambda_k).$$

The updates in the statement are the hypotheses of lemma 4, whose dual step is taken at z_k. Its remaining hypothesis $z_k = z^{\eta_k}(\lambda_k)$ holds for $k \geq 1$ because that is how (Dual Avg.) produced z_k, and at $k = 0$ because $\eta_0 = 0$ and $\lambda_0 = 0$ give $z^{\eta_0}(\lambda_0) = x_0 = z_0$. It gives

$$\mathcal{D}_{k+1} \geq (1 - t_k)\mathcal{D}_k + t_k\phi(z_{k+1}).$$

With the limiting initialization $\eta_0 = 0$, we have $t_0 = 1$, so

$$\mathcal{D}_1 \geq \phi(z_1).$$

Since $x_1 \in \arg\min_{1 \leq i \leq 1} \phi(z_i)$, this gives

$$\phi(x_1) \leq \mathcal{D}_1.$$

Assume now that

$$\phi(x_k) \leq \mathcal{D}_k.$$

Because

$$x_{k+1} \in \operatorname*{arg\,min}_{1 \leq i \leq k+1} \phi(z_i),$$

we have

$$\phi(x_{k+1}) \leq (1 - t_k)\phi(x_k) + t_k\phi(z_{k+1}).$$

Thus

$$\phi(x_{k+1}) \leq (1 - t_k)\mathcal{D}_k + t_k\phi(z_{k+1}) \leq \mathcal{D}_{k+1}.$$

By induction,

$$\phi(x_k) \leq \mathcal{D}_k \qquad \forall k \geq 1.$$

Finally,

$$\mathcal{D}_k = \min_x S^{\eta_k}(x, \lambda_k) \leq \phi_\star^{\eta_k},$$

so

$$\phi(x_k) \leq \phi_\star^{\eta_k}.$$

Thus the update satisfies algorithm 1.

It remains to estimate η_k^{-1}. If $\mu = 0$, then

$$t_k = \frac{\eta_k^{-1}}{L + \eta_k^{-1}}, \qquad \eta_{k+1}^{-1} = \frac{L\eta_k^{-1}}{L + \eta_k^{-1}}.$$

Equivalently,

$$\eta_{k+1} = \eta_k + \frac{1}{L}.$$

With $\eta_0 = 0$, this gives

$$\eta_k^{-1} = \frac{L}{k}.$$

If $\mu > 0$, then, with $\kappa := \mu/L$,

$$\eta_{k+1}^{-1} = (1 - t_k)\eta_k^{-1} = \frac{(L - \mu)\eta_k^{-1}}{L + \eta_k^{-1}}.$$

Hence

$$\frac{L}{\eta_{k+1}^{-1}} = \frac{1}{1 - \kappa}\left(\frac{L}{\eta_k^{-1}} + 1\right).$$

With the limiting initialization $\eta_0^{-1} = +\infty$, this yields

$$\frac{L}{\eta_k^{-1}} = \sum_{i=1}^{k}(1 - \kappa)^{-i} = \frac{(1 - \kappa)^{-k} - 1}{\kappa}.$$

Thus

$$\eta_k^{-1} = L\frac{\kappa}{(1-\kappa)^{-k}-1} \leq L\frac{\kappa}{(1+\kappa)^k-1},$$

where we used

$$(1-\kappa)^{-1} \geq 1+\kappa.$$

Each term of the sum $\sum_{i=1}^{k}(1-\kappa)^{-i}$ above is at least 1, so $L/\eta_k^{-1} \geq k$, that is,

$$\eta_k^{-1} \leq \frac{L}{k}.$$

Therefore

$$\eta_k^{-1} \leq L\min\left\{\frac{1}{k}, \frac{\kappa}{(1+\kappa)^k-1}\right\}.$$

□

Geometry	$d_i(x)$	**Dual form**	**Primal form** ($\Omega = \nabla d(\mathcal{X}')$)
Euclidean	$\frac{1}{2}x^2$	1	1
KL	$x\log x - x$	$\sup_{w_1,w_2\in\Omega} \max_i \left(\frac{e^{(w_2)_i}}{e^{(w_1)_i}}\right)^{1/2}$	$\sup_{x_1,x_2\in\mathcal{X}'} \max_i \left(\frac{(x_2)_i}{(x_1)_i}\right)^{1/2}$
Burg	$-\log x$	$\sup_{w_1,w_2\in\Omega} \max_i \frac{(w_1)_i}{(w_2)_i}$	$\sup_{x_1,x_2\in\mathcal{X}'} \max_i \frac{(x_2)_i}{(x_1)_i}$
Power-type	$\frac{\lvert x\rvert^p}{p}$	$\sup_{w_1,w_2\in\Omega} \max_i \left(\frac{\lvert(w_2)_i\rvert}{\lvert(w_1)_i\rvert}\right)^{\frac{2-p}{2(p-1)}}$	$\sup_{x_1,x_2\in\mathcal{X}'} \max_i \left(\frac{\lvert(x_1)_i\rvert}{\lvert(x_2)_i\rvert}\right)^{\frac{p-2}{2}}$

Table 2: Dual norm distortion factor $\rho_{d^*}(\Omega)$ for standard separable geometries, in dual and primal coordinates, where $\mathcal{X}' \subset \operatorname{int}(\operatorname{dom} d)$ is the primal region with $\nabla d(\mathcal{X}') = \Omega$. The factor is finite when the coordinate ratios over Ω are bounded, but diverges as a coordinate approaches the boundary of the domain for the KL and Burg geometries, or the origin for the power-type geometry. In the Euclidean case, it equals 1 on every Ω.

C.2 Technical results on BLDF

Lemma 5 (BLDF under smoothness and strong convexity)**.** *Assume that d^* is L_{d^*}-smooth and μ_{d^*}-strongly convex on a convex set $\Omega \subset \operatorname{int}(\operatorname{dom} d^*)$. Then, for every $s > 0$,*

$$c_{d^*}(s, \Omega; 2) \le \sqrt{\frac{L_{d^*}}{\mu_{d^*}}}.$$

In particular, if $d^*(u) = \frac{1}{2}\|u\|^2$, then

$$c_{d^*}(s, \Omega; 2) = 1.$$

Proof. By L_{d^*}-smoothness and μ_{d^*}-strong convexity, for all admissible u_1, u_2, v,

$$D_{d^*}(u_2 + sv, u_2) \le \frac{L_{d^*}}{2}\|sv\|^2 = s^2\frac{L_{d^*}}{2}\|v\|^2,$$

and

$$D_{d^*}(u_1 + v, u_1) \ge \frac{\mu_{d^*}}{2}\|v\|^2.$$

Therefore

$$\frac{D_{d^*}(u_2 + sv, u_2)}{s^2 D_{d^*}(u_1 + v, u_1)} \le \frac{L_{d^*}}{\mu_{d^*}}.$$

Taking the square root and then the supremum over all admissible u_1, u_2, v gives

$$c_{d^*}(s, \Omega; 2) \le \sqrt{\frac{L_{d^*}}{\mu_{d^*}}}.$$

For $d^*(u) = \frac{1}{2}\|u\|^2$,

$$D_{d^*}(u_2 + sv, u_2) = \frac{1}{2}s^2\|v\|^2, \qquad D_{d^*}(u_1 + v, u_1) = \frac{1}{2}\|v\|^2,$$

so the ratio is exactly 1, hence

$$c_{d^*}(s, \Omega; 2) = 1.$$

□

Corollary 2 (Separable case)**.** *Assume that d^* is twice differentiable and separable on Ω, i.e.,*

$$d^*(u) = \sum_{i=1}^n d_i^*(u_i).$$

Then the dual norm distortion factor satisfies

$$\rho_{d^*}(\Omega) = \sup_{w_1,w_2\in\Omega} \max_{1\le i\le n} \left(\frac{(d_i^*)''((w_2)_i)}{(d_i^*)''((w_1)_i)}\right)^{1/2}.$$

Proof. For separable d^*, definition 8 gives

$$\|\xi\|_{w,*}^2 = \tfrac{1}{2}\sum_{i=1}^{n}(d_i^*)''(w_i)\,\xi_i^2,$$

and the factor $\frac{1}{2}$ cancels in the ratio below. Hence, for any $w_1, w_2 \in \Omega$,

$$\sup_{\xi\neq 0}\frac{\|\xi\|_{w_2,*}}{\|\xi\|_{w_1,*}} = \max_{1\le i\le n}\left(\frac{(d_i^*)''((w_2)_i)}{(d_i^*)''((w_1)_i)}\right)^{1/2},$$

and the result follows by taking the supremum over $w_1, w_2 \in \Omega$. □

Proposition 1. *Assume that d^* is twice differentiable on $\Omega \subset \operatorname{int}(\operatorname{dom} d^*)$. Then, for every $s > 0$,*

$$c_{d^*}(s, \Omega; 2) \le \rho_{d^*}(\Omega).$$

Proof. Fix admissible $u_1, u_2 \in \Omega$ and $v \neq 0$. Since $u_1 + [0,1]v \subset \Omega$ and $u_2 + [0,s]v \subset \Omega$, the integral representation of the Bregman divergence gives

$$D_{d^*}(u_1 + v, u_1) = \int_0^1 (1-r)\, v^\top \nabla^2 d^*(u_1 + rv)\, v\, dr,$$

and this holds under pointwise twice differentiability alone. Indeed, $g(r) := d^*(u_1 + rv)$ is convex, so g' is nondecreasing, and g' is differentiable everywhere on $[0,1]$ with $g'' \ge 0$ and $\int_0^1 g'' \le g'(1) - g'(0) < \infty$; a function differentiable everywhere with integrable derivative is absolutely continuous, so the fundamental theorem of calculus applies to g', and one integration by parts produces the weight $(1-r)$. Note also that $D_{d^*}(u_1 + v, u_1) > 0$: d is Legendre, hence so is d^*, hence d^* is strictly convex on the convex set $\Omega \subseteq \operatorname{int}(\operatorname{dom} d^*)$, so the denominators below do not vanish. Explicitly,

$$D_{d^*}(u_1 + v, u_1) = \int_0^1 (1-r)\, v^\top \nabla^2 d^*(u_1 + rv)\, v\, dr,$$

and

$$D_{d^*}(u_2 + sv, u_2) = s^2 \int_0^1 (1-r)\, v^\top \nabla^2 d^*(u_2 + rsv)\, v\, dr.$$

Both integrands are nonnegative and $\int_0^1 (1-r)\, dr = \frac{1}{2}$, so each divergence lies between the extreme values of $\|v\|_{\cdot,*}^2$ on its segment,

$$\inf_{w\in u_1+[0,1]v}\|v\|_{w,*}^2 \;\le\; D_{d^*}(u_1 + v, u_1) \;\le\; \sup_{w\in u_1+[0,1]v}\|v\|_{w,*}^2,$$

and likewise $D_{d^*}(u_2 + sv, u_2) \in s^2[\inf, \sup]$ over $u_2 + [0,s]v$. No continuity of $\nabla^2 d^*$ is required. Therefore,

$$\left(\frac{D_{d^*}(u_2 + sv, u_2)}{s^\gamma D_{d^*}(u_1 + v, u_1)}\right)^{1/2} \le s^{1-\gamma/2}\left(\frac{\sup_{w_2\in u_2+[0,s]v}\|v\|_{w_2,*}^2}{\inf_{w_1\in u_1+[0,1]v}\|v\|_{w_1,*}^2}\right)^{1/2} \le s^{1-\gamma/2}\rho_{d^*}(\Omega),$$

the last step because both segments lie in Ω. Taking the supremum over all admissible u_1, u_2, v yields the claim. □

With the following additional regularity assumptions, the two factors are equal.

Proposition 4. *Assume that Ω is open and that $\nabla^2 d^*$ is continuous and positive definite on Ω. Then*

$$c_{d^*}(s, \Omega; 2) = \rho_{d^*}(\Omega) \quad \textit{for every } s > 0, \qquad c_{d^*}(s, \Omega; \gamma) = s^{1-\gamma/2}\rho_{d^*}(\Omega) \quad \textit{for every } \gamma > 0.$$

Proof. The proof of proposition 1 gives $\le$ at every γ, although the proposition is stated at $\gamma = 2$; only the reverse direction is at issue. Fix $u_1, u_2 \in \Omega$ and $\xi \neq 0$, and take $v = \varepsilon\xi$. Since Ω is open, both $u_1 + [0,1]v$ and $u_2 + [0,s]v$ lie in Ω once ε is small enough, so the pair is admissible. Here $\nabla^2 d^*$ is continuous, so the weighted mean-value theorem applies to the integral representation used

in the proof of proposition 1 and gives points $\bar{w}_1(\varepsilon) \in u_1 + [0,\varepsilon]\xi$ and $\bar{w}_2(\varepsilon) \in u_2 + [0,s\varepsilon]\xi$ with $D_{d^*}(u_1+v,u_1) = \|v\|^2_{\bar{w}_1(\varepsilon),*}$ and $D_{d^*}(u_2+sv,u_2) = s^2\|v\|^2_{\bar{w}_2(\varepsilon),*}$. Both points converge to u_1 and u_2 as $\varepsilon \downarrow 0$, and $\nabla^2 d^*$ is positive definite there, so

$$\left(\frac{D_{d^*}(u_2+sv,u_2)}{s^\gamma D_{d^*}(u_1+v,u_1)}\right)^{1/2} = s^{1-\gamma/2}\frac{\|\xi\|_{\bar{w}_2(\varepsilon),*}}{\|\xi\|_{\bar{w}_1(\varepsilon),*}} \xrightarrow[\varepsilon\downarrow 0]{} s^{1-\gamma/2}\frac{\|\xi\|_{u_2,*}}{\|\xi\|_{u_1,*}}.$$

The left-hand side is bounded above by $c_{d^*}(s,\Omega;\gamma)$ for every ε, so the limit is too. Taking the supremum over $u_1,u_2 \in \Omega$ and $\xi \neq 0$ gives $c_{d^*}(s,\Omega;\gamma) \geq s^{1-\gamma/2}\rho_{d^*}(\Omega)$. □

C.3 Technical results on Accelerated Method

Throughout this subsection, $\alpha_k := \mu + \eta_k^{-1}$.

Lemma 6 (Implicit recursive accelerated bound)**.** *Assume that $t_k = 1$, or that*

$$\phi_k^{\text{best}} \le \mathcal{D}^{\eta_k}(\lambda_k).$$

If the inner loop of Algorithm 2 exits, i.e.,

$$\phi_{k+1}^{\text{best}} \le (1 - t_k)\phi_k^{\text{best}} + t_k\phi_k^{\text{low}} - \left(\mu + \eta_{k+1}^{-1}\right) D_d(z_k, z_{k+1}),$$

then

$$\phi_{k+1}^{\text{best}} = \phi(x_{k+1}^{\text{best}}) \le \phi_\star^{\eta_{k+1}}.$$

Proof. Apply lemma 3 with $b_k = \phi_k^{\text{best}}$; when $t_k = 1$, apply instead its second conclusion, which carries no hypothesis on ϕ_k^{best} and whose right-hand side is the same because $(1 - t_k) = 0$. The exit condition gives

$$\phi_{k+1}^{\text{best}} \le \mathcal{D}^{\eta_{k+1}}(\lambda_{k+1}).$$

Since $S^{\eta_{k+1}}(x, \lambda_{k+1}) \le \phi^{\eta_{k+1}}(x)$ for all x,

$$\mathcal{D}^{\eta_{k+1}}(\lambda_{k+1}) = \min_x S^{\eta_{k+1}}(x, \lambda_{k+1}) \le \min_x \phi^{\eta_{k+1}}(x) = \phi_\star^{\eta_{k+1}}.$$

Therefore

$$\phi_{k+1}^{\text{best}} \le \phi_\star^{\eta_{k+1}}.$$

□

Theorem 2 (A posteriori convergence of ABrA-GD)**.** *Under As. 1, assume that the inner loop of algorithm 2 exits at every iteration. Let $M_k > \mu$ and choose $t_k \in (0, 1]$ by*

$$\mu + \eta_{k+1}^{-1} = M_k t_k^2 \quad (k \ge 0), \qquad \eta_{k+1}^{-1} = (1 - t_k)\eta_k^{-1} \quad (k \ge 1), \qquad t_0 = 1, \quad \eta_1^{-1} = M_0 - \mu.$$

Then, for any $x_\star \in \arg\min_x \phi(x)$,

$$\phi_{k+1}^{\text{best}} \le \phi_k^{\text{best}}, \qquad \phi_{k+1}^{\text{best}} \le \phi_\star^{\eta_{k+1}}, \qquad \phi_{k+1}^{\text{best}} - \phi_\star \le \eta_{k+1}^{-1} D_d(x_\star, x_0).$$

Moreover, the decay of η_k^{-1} is controlled by the acceleration parameters M_k:

Case	***No uniform bound on*** M_k	***If*** $M_i \le M$ ***for all*** $i \ge K$
$\mu = 0$	$\eta_k^{-1} \le \dfrac{4}{\left(\sum_{i=0}^{k-1} M_i^{-1/2}\right)^2}$	$\eta_k^{-1} \le \dfrac{\eta_K^{-1}}{\left(1 + \frac{k-K}{2\sqrt{M\eta_K}}\right)^2}$
$\mu > 0$	$\eta_k^{-1} \le (M_0 - \mu) \displaystyle\prod_{i=1}^{k-1} \left(1 - \sqrt{\frac{\mu}{M_i}}\right)$	$\eta_k^{-1} \le \eta_K^{-1} \left(1 - \sqrt{\frac{\mu}{M}}\right)^{k-K}$

Proof. By lemma 6, the invariant and the exit condition imply

$$\phi_{k+1}^{\text{best}} \le \phi_\star^{\eta_{k+1}}.$$

Thus the iterates satisfy the implicit descent condition. Applying Theorem 1 with $x_{k+1} = x_{k+1}^{\text{best}}$ gives

$$\phi_{k+1}^{\text{best}} - \phi_\star \le \eta_{k+1}^{-1} D_d(x_\star, x_0).$$

The monotonicity

$$\phi_{k+1}^{\text{best}} \le \phi_k^{\text{best}}$$

follows from the definition of the best value.

It remains to estimate η_k^{-1}. First assume $\mu = 0$. Then

$$\eta_{k+1}^{-1} = M_k t_k^2, \qquad t_k = 1 - \frac{\eta_k}{\eta_{k+1}}.$$

Hence

$$\eta_{k+1} = M_k(\eta_{k+1} - \eta_k)^2,$$

so

$$\eta_{k+1} - \eta_k = \sqrt{\frac{\eta_{k+1}}{M_k}}.$$

Therefore

$$\sqrt{\eta_{k+1}} - \sqrt{\eta_k} = \frac{\eta_{k+1} - \eta_k}{\sqrt{\eta_{k+1}} + \sqrt{\eta_k}} \geq \frac{1}{2\sqrt{M_k}}.$$

With the limiting initialization $\eta_0 = 0$, summing gives

$$\sqrt{\eta_k} \geq \frac{1}{2} \sum_{i=0}^{k-1} M_i^{-1/2},$$

hence

$$\eta_k^{-1} \leq \frac{4}{\left(\sum_{i=0}^{k-1} M_i^{-1/2}\right)^2}.$$

If $M_i \leq M$ for all $i \geq K$, then

$$\sqrt{\eta_k} \geq \sqrt{\eta_K} + \frac{k-K}{2\sqrt{M}},$$

and thus

$$\eta_k^{-1} \leq \frac{\eta_K^{-1}}{\left(1 + \frac{k-K}{2\sqrt{M\eta_K}}\right)^2}.$$

Now assume $\mu > 0$. Since

$$\mu + \eta_{k+1}^{-1} = M_k t_k^2,$$

we have

$$t_k \geq \sqrt{\frac{\mu}{M_k}}.$$

Thus

$$\eta_{k+1}^{-1} = (1 - t_k)\eta_k^{-1} \leq \left(1 - \sqrt{\frac{\mu}{M_k}}\right) \eta_k^{-1}.$$

With the limiting initialization $\eta_0 = 0$, we have $t_0 = 1$, hence

$$\eta_1^{-1} = M_0 - \mu.$$

Iterating from $i = 1$ to $k - 1$ gives

$$\eta_k^{-1} \leq (M_0 - \mu) \prod_{i=1}^{k-1} \left(1 - \sqrt{\frac{\mu}{M_i}}\right).$$

If $M_i \leq M$ for all $i \geq K$, then

$$1 - \sqrt{\frac{\mu}{M_i}} \leq 1 - \sqrt{\frac{\mu}{M}},$$

and therefore

$$\eta_k^{-1} \leq \eta_K^{-1} \left(1 - \sqrt{\frac{\mu}{M}}\right)^{k-K}.$$

□

Lemma 7 (Generic accelerated bound)**.** *Assume that $t_k = 1$, or that*

$$\phi(x_k) \leq \mathcal{D}^{\eta_k}(\lambda_k), \qquad \lambda_{k+1} = (1 - t_k)\lambda_k + t_k \nabla f(y_k), \qquad \eta_{k+1}^{-1} = (1 - t_k)\eta_k^{-1},$$

and set

$$y_k = (1 - \tau_k)x_k + \tau_k z_k.$$

Let

$$\mathcal{L}_{k+1}(z_{k+1}) := (1 - t_k)\phi(x_k) + t_k \phi_k^{\text{low}} - (\mu + \eta_{k+1}^{-1}) D_d(z_k, z_{k+1}),$$

with ϕ_k^{low} as in lemma 3. This is the right-hand side of the exit condition of algorithm 2 with $\phi_k^{\mathrm{best}} = \phi(x_k)$, and $\mathcal{L}_{k+1}(z_{k+1}) \le \mathcal{D}^{\eta_{k+1}}(\lambda_{k+1})$ by lemma 3. Then x_{k+1} in Algorithm 3 satisfies

$$\begin{aligned}\mathcal{L}_{k+1}(z_{k+1}) \ge \phi(x_{k+1}) &+ LD_d(y_k, x_{k+1}) - (\mu + \eta_{k+1}^{-1})D_d(z_k, z_{k+1}) \\ &+ (1-t_k)\mu D_d(x_k, y_k) + t_k \mu D_d(z_k, y_k) \\ &+ (\tau_k - t_k)\langle \nabla\phi(y_k), x_k - z_k\rangle.\end{aligned}$$

Proof. The bound $\mathcal{L}_{k+1}(z_{k+1}) \le \mathcal{D}^{\eta_{k+1}}(\lambda_{k+1})$ is lemma 3 with $b_k = \phi(x_k)$, or its second conclusion when $t_k = 1$. When $t_k = 1$, every term below carrying the factor $(1 - t_k)$ — including the strong convexity substitution for $\phi(x_k)$ — is absent. Expanding ϕ_k^{low} gives

$$\begin{aligned}\mathcal{L}_{k+1}(z_{k+1}) = (1-t_k)\phi(x_k) &+ t_k\phi(y_k) + t_k\langle \nabla\phi(y_k), z_k - y_k\rangle \\ &+ t_k\mu D_d(z_k, y_k) - (\mu + \eta_{k+1}^{-1})D_d(z_k, z_{k+1}).\end{aligned}$$

By relative μ-strong convexity of ϕ,

$$\phi(x_k) \ge \phi(y_k) + \langle \nabla\phi(y_k), x_k - y_k\rangle + \mu D_d(x_k, y_k).$$

Substituting this lower bound yields

$$\begin{aligned}\mathcal{L}_{k+1}(z_{k+1}) \ge \phi(y_k) &+ (1-t_k)\mu D_d(x_k, y_k) + t_k\mu D_d(z_k, y_k) \\ &+ \langle \nabla\phi(y_k), (1-t_k)(x_k - y_k) + t_k(z_k - y_k)\rangle \\ &- (\mu + \eta_{k+1}^{-1})D_d(z_k, z_{k+1}).\end{aligned}$$

Since

$$y_k = (1-\tau_k)x_k + \tau_k z_k,$$

we have

$$(1-t_k)(x_k - y_k) + t_k(z_k - y_k) = (\tau_k - t_k)(x_k - z_k).$$

Finally, the descent step gives

$$\phi(x_{k+1}) \le \phi(y_k) - LD_d(y_k, x_{k+1}),$$

which gives the claim. □

Lemma 8 (Convex specialization)**.** *Assume $\mu = 0$ and choose $\tau_k = t_k$. Then*

$$\mathcal{L}_{k+1}(z_{k+1}) \ge \phi(x_{k+1}) + LD_d(y_k, x_{k+1}) - \alpha_{k+1}D_d(z_k, z_{k+1}).$$

Proof. Apply lemma 7. Since $\mu = 0$, the two terms multiplied by μ vanish. Since $\tau_k = t_k$, the scalar-product term also vanishes. □

Proposition 5 (Convex one-step bound under dual norm distortion)**.** *Under As. 1, assume $\mu = 0$. Let $\rho_k \ge 1$ be such that the relevant dual segments of iteration k are contained in a convex set Ω_k satisfying*

$$\rho_{d^*}(\Omega_k) \le \rho_k.$$

Let $M_k \ge \rho_k^2 L$ and choose

$$\tau_k = t_k, \qquad \eta_{k+1}^{-1} = M_k t_k^2, \qquad \eta_{k+1}^{-1} = (1-t_k)\eta_k^{-1}.$$

Then the accelerated step satisfies

$$\mathcal{L}_{k+1}(z_{k+1}) \ge \phi(x_{k+1}).$$

Hence the exit condition holds at every acceleration parameter $M_k \ge \rho_k^2 L$, with $\mathcal{L}_{k+1}(z_{k+1})$ as in lemma 7.

Proof. By lemma 8,

$$\mathcal{L}_{k+1}(z_{k+1}) \geq \phi(x_{k+1}) + LD_d(y_k, x_{k+1}) - \eta_{k+1}^{-1} D_d(z_k, z_{k+1}).$$

Since $\mu = 0$,

$$\nabla d(x_{k+1}) = \nabla d(y_k) - \frac{1}{L}\nabla\phi(y_k), \qquad \nabla d(z_{k+1}) = \nabla d(z_k) - \frac{t_k}{\eta_{k+1}^{-1}}\nabla\phi(y_k).$$

Thus the two dual displacements are parallel, with relative scaling

$$\frac{Lt_k}{\eta_{k+1}^{-1}}.$$

Using *proposition* 1 and $\rho_{d^*}(\Omega_k) \leq \rho_k$,

$$D_d(z_k, z_{k+1}) \leq \rho_k^2 \left(\frac{Lt_k}{\eta_{k+1}^{-1}}\right)^2 D_d(y_k, x_{k+1}).$$

Therefore

$$LD_d(y_k, x_{k+1}) - \eta_{k+1}^{-1} D_d(z_k, z_{k+1}) \geq \left(L - \frac{\rho_k^2 L^2 t_k^2}{\eta_{k+1}^{-1}}\right) D_d(y_k, x_{k+1}) \geq 0,$$

where the last inequality uses $\eta_{k+1}^{-1} = M_k t_k^2 \geq \rho_k^2 L t_k^2$. Hence

$$\mathcal{L}_{k+1}(z_{k+1}) \geq \phi(x_{k+1}).$$

□

Lemma 9 (Strongly convex specialization)**.** *Assume $\mu > 0$. Define*

$$q_{k+1} := \nabla d^*\left(\nabla d(y_k) - \frac{1}{\mu}\nabla\phi(y_k)\right).$$

Assume that $\nabla d(y_k) - \frac{1}{\mu}\nabla\phi(y_k) \in \operatorname{int}(\operatorname{dom} d^*)$, so that q_{k+1} is well defined, and that $\tau_k < 1$. Then, if

$$t_k \leq 1, \qquad \frac{t_k - \tau_k}{1 - \tau_k} \leq t_k,$$

we have

$$\mathcal{L}_{k+1}(z_{k+1}) \geq \phi(x_{k+1}) + \left[LD_d(y_k, x_{k+1}) - \mu\frac{t_k - \tau_k}{1 - \tau_k} D_d(y_k, q_{k+1})\right] + \left[\mu\frac{t_k - \tau_k}{1 - \tau_k} D_d(z_k, q_{k+1}) - \alpha_{k+1} D_d(z_k, z_{k+1})\right].$$

Proof. By definition of q_{k+1},

$$\nabla\phi(y_k) = \mu\big(\nabla d(y_k) - \nabla d(q_{k+1})\big).$$

Since

$$y_k = (1 - \tau_k)x_k + \tau_k z_k,$$

we have

$$x_k - z_k = \frac{y_k - z_k}{1 - \tau_k}.$$

Therefore

$$(\tau_k - t_k)\langle\nabla\phi(y_k), x_k - z_k\rangle = \frac{\mu(\tau_k - t_k)}{1 - \tau_k}\langle\nabla d(y_k) - \nabla d(q_{k+1}), y_k - z_k\rangle.$$

By the three-point identity,

$$\langle\nabla d(y_k) - \nabla d(q_{k+1}), y_k - z_k\rangle = D_d(y_k, q_{k+1}) + D_d(z_k, y_k) - D_d(z_k, q_{k+1}).$$

Hence

$$\begin{aligned}(\tau_k - t_k)\langle \nabla\phi(y_k), x_k - z_k\rangle = &-\mu\frac{t_k-\tau_k}{1-\tau_k}D_d(y_k, q_{k+1})\\ &-\mu\frac{t_k-\tau_k}{1-\tau_k}D_d(z_k, y_k)\\ &+\mu\frac{t_k-\tau_k}{1-\tau_k}D_d(z_k, q_{k+1}).\end{aligned}$$

Substituting this identity into lemma 7 and writing $\alpha_{k+1} = \mu + \eta_{k+1}^{-1}$ gives

$$\begin{aligned}\mathcal{L}_{k+1}(z_{k+1}) \geq \phi(x_{k+1}) &+ LD_d(y_k, x_{k+1}) - \alpha_{k+1}D_d(z_k, z_{k+1})\\ &+ (1-t_k)\mu D_d(x_k, y_k) + t_k\mu D_d(z_k, y_k)\\ &- \mu\frac{t_k-\tau_k}{1-\tau_k}D_d(y_k, q_{k+1}) - \mu\frac{t_k-\tau_k}{1-\tau_k}D_d(z_k, y_k)\\ &+ \mu\frac{t_k-\tau_k}{1-\tau_k}D_d(z_k, q_{k+1}).\end{aligned}$$

Rearranging,

$$\begin{aligned}\mathcal{L}_{k+1}(z_{k+1}) \geq \phi(x_{k+1}) &+ \left[LD_d(y_k, x_{k+1}) - \mu\frac{t_k-\tau_k}{1-\tau_k}D_d(y_k, q_{k+1})\right]\\ &+ \left[\mu\frac{t_k-\tau_k}{1-\tau_k}D_d(z_k, q_{k+1}) - \alpha_{k+1}D_d(z_k, z_{k+1})\right]\\ &+ (1-t_k)\mu D_d(x_k, y_k) + \mu\left(t_k - \frac{t_k-\tau_k}{1-\tau_k}\right)D_d(z_k, y_k).\end{aligned}$$

The last two terms are nonnegative under

$$t_k \leq 1, \qquad \frac{t_k-\tau_k}{1-\tau_k} \leq t_k.$$

Dropping them gives the result. □

Proposition 6 (Strongly convex one-step bound under dual norm distortion)**.** *Under As. 1, assume $\mu > 0$ and that q_{k+1} of lemma 9 is well defined. Let $\rho_k \geq 1$ be such that the relevant dual segments of iteration k, including the dual segments joining $\nabla d(y_k)$ to $\nabla d(q_{k+1})$ and joining $\nabla d(z_k)$ to $\nabla d(q_{k+1})$, are contained in a convex set Ω_k satisfying*

$$\rho_{d^*}(\Omega_k) \leq \rho_k.$$

Let $M_k \geq \rho_k^4 L$, and let $t_k \in (0, 1]$ be chosen by

$$\mu + \eta_{k+1}^{-1} = M_k t_k^2, \qquad \eta_{k+1}^{-1} = (1-t_k)\eta_k^{-1}.$$

Equivalently, with $\alpha_k := \mu + \eta_k^{-1}$, t_k is the positive root of

$$M_k t_k^2 + (\alpha_k - \mu)t_k - \alpha_k = 0.$$

Set

$$\tau_k = \frac{t_k - \mu/\sqrt{M_k L}}{1 - \mu/\sqrt{M_k L}}, \qquad y_k = (1-\tau_k)x_k + \tau_k z_k.$$

Throughout, L is the constant the gradient step of algorithm 3 actually uses, which is the same constant that forms τ_k. Then the accelerated step satisfies

$$\mathcal{L}_{k+1}(z_{k+1}) \geq \phi(x_{k+1}).$$

Hence the exit condition holds at every acceleration parameter $M_k \geq \rho_k^4 L$, with $\mathcal{L}_{k+1}(z_{k+1})$ as in lemma 7.

Proof. Assume first $t_k < 1$; the case $t_k = 1$, which the initialization imposes at $k = 0$, is treated at the end.

Set

$$\alpha_{k+1} := \mu + \eta_{k+1}^{-1}, \qquad \delta_k := \frac{t_k - \tau_k}{1 - \tau_k}.$$

By construction,

$$\delta_k = \frac{\mu}{\sqrt{M_k L}}.$$

Moreover, $\alpha_{k+1} = M_k t_k^2$.

By lemma 9,

$$\mathcal{L}_{k+1}(z_{k+1}) \geq \phi(x_{k+1}) + \Big[L D_d(y_k, x_{k+1}) - \mu\delta_k D_d(y_k, q_{k+1}) \Big] + \Big[\mu\delta_k D_d(z_k, q_{k+1}) - \alpha_{k+1} D_d(z_k, z_{k+1}) \Big],$$

provided $t_k \leq 1$ and $\delta_k \leq t_k$.

We first show that the two brackets are nonnegative. The two uses of the dual BLDF are separate and each compares displacements along the same dual ray. For the first bracket, set

$$u_y := \nabla d(y_k), \qquad v_y := \nabla d(x_{k+1}) - \nabla d(y_k).$$

The primal gradient step and the definition of q_{k+1} give

$$\nabla d(x_{k+1}) = u_y + v_y, \qquad \nabla d(q_{k+1}) = u_y + \frac{L}{\mu} v_y.$$

Hence

$$D_d(y_k, x_{k+1}) = D_{d^*}(u_y + v_y, u_y), \qquad D_d(y_k, q_{k+1}) = D_{d^*}\left(u_y + \frac{L}{\mu} v_y, u_y \right).$$

Applying the dual BLDF on this same ray, with scaling L/μ, gives

$$D_d(y_k, x_{k+1}) \geq \frac{1}{\rho_k^2} \left(\frac{\mu}{L} \right)^2 D_d(y_k, q_{k+1}),$$

hence

$$L D_d(y_k, x_{k+1}) - \mu\delta_k D_d(y_k, q_{k+1}) \geq \left(\frac{\mu^2}{\rho_k^2 L} - \mu\delta_k \right) D_d(y_k, q_{k+1}) \geq 0,$$

since $\sqrt{M_k L} \geq \rho_k^2 L$. For the second bracket, set

$$u_z := \nabla d(z_k), \qquad v_z := \nabla d(q_{k+1}) - \nabla d(z_k).$$

Using the optimality conditions of the dual map and $\alpha_{k+1} = \mu + \eta_{k+1}^{-1}$, we get

$$\nabla d(z_{k+1}) = u_z + \frac{t_k \mu}{\alpha_{k+1}} v_z.$$

Thus

$$D_d(z_k, z_{k+1}) = D_{d^*}\left(u_z + \frac{t_k \mu}{\alpha_{k+1}} v_z, u_z \right), \qquad D_d(z_k, q_{k+1}) = D_{d^*}(u_z + v_z, u_z),$$

and the dual BLDF on this ray gives

$$D_d(z_k, z_{k+1}) \leq \rho_k^2 \left(\frac{t_k \mu}{\alpha_{k+1}} \right)^2 D_d(z_k, q_{k+1}),$$

so

$$\mu\delta_k D_d(z_k, q_{k+1}) - \alpha_{k+1} D_d(z_k, z_{k+1}) \geq \left(\mu\delta_k - \frac{\rho_k^2 t_k^2 \mu^2}{\alpha_{k+1}} \right) D_d(z_k, q_{k+1}) \geq 0,$$

where the last inequality uses $\delta_k = \mu/\sqrt{M_k L}$ and $\alpha_{k+1} = M_k t_k^2$, so that

$$\mu\delta_k - \frac{\rho_k^2 t_k^2 \mu^2}{\alpha_{k+1}} = \mu^2 \left(\frac{1}{\sqrt{M_k L}} - \frac{\rho_k^2}{M_k} \right) \geq 0 \quad \text{since } M_k \geq \rho_k^4 L.$$

Thus both brackets are nonnegative, and therefore

$$\mathcal{L}_{k+1}(z_{k+1}) \geq \phi(x_{k+1}).$$

It remains to check the side conditions of *lemma* 9. Since

$$\alpha_{k+1} = M_k t_k^2 = \mu + \eta_{k+1}^{-1} \geq \mu,$$

we have

$$t_k \geq \sqrt{\frac{\mu}{M_k}} \geq \sqrt{\frac{\mu}{M_k}} \sqrt{\frac{\mu}{L}} = \frac{\mu}{\sqrt{M_k L}} = \delta_k,$$

where we used $\mu \leq L$. Also $t_k \leq 1$: the polynomial

$$p(t) = M_k t^2 + (\alpha_k - \mu)t - \alpha_k$$

satisfies $p(0) = -\alpha_k < 0$ and

$$p(1) = M_k - \mu \geq 0,$$

so its positive root lies in $(0, 1]$. Hence the assumptions of *lemma* 9 hold.

The case $t_k = 1$. This case is outside lemma 9, which requires $\tau_k < 1$: the definition of τ_k gives $\tau_k = 1$ when $t_k = 1$, and $\delta_k = (t_k - \tau_k)/(1 - \tau_k)$ is then $0/0$. It arises only at $k = 0$, where algorithm 2 imposes $t_0 = 1$ from outside the root formula above, and it does not need the virtual point q_{k+1} at all. The initialization sets $x_0^{\text{best}} = z_0 = x_0$, so $y_0 = x_0$ for every τ_0, and $\nabla f(y_0) = \nabla\phi(x_0)$. The two dual displacements therefore leave the single anchor $\nabla d(x_0)$ along $-\nabla\phi(x_0)$:

$$\nabla d(x_1) = \nabla d(x_0) - \tfrac{1}{L}\nabla\phi(x_0), \qquad \nabla d(z_1) = \nabla d(x_0) - \tfrac{1}{\alpha_1}\nabla\phi(x_0),$$

which is the configuration of proposition 5. At $y_0 = z_0 = x_0$, the $t_k = 1$ case of lemma 3 gives $\mathcal{D}^{\eta_1}(\lambda_1) \geq \min_x\{\phi(x_0) + \langle\nabla\phi(x_0), x - x_0\rangle + \alpha_1 D_d(x, x_0)\}$. The minimizer is z_1, and $\nabla\phi(x_0) = \alpha_1(\nabla d(x_0) - \nabla d(z_1))$ gives $\langle\nabla\phi(x_0), z_1 - x_0\rangle = -\alpha_1\big(D_d(z_1, x_0) + D_d(x_0, z_1)\big)$. Hence

$$\mathcal{D}^{\eta_1}(\lambda_1) \geq \phi(x_0) - \alpha_1 D_d(x_0, z_1) = \phi(x_0) - \alpha_1 D_d(z_0, z_1) = \mathcal{L}_1(z_1),$$

since $\phi_0^{\text{low}} = \phi(x_0)$ at $y_0 = z_0 = x_0$. The same computation with L in place of α_1 and x_1 in place of z_1, together with the relative smoothness of ϕ, gives $\phi(x_1) \leq \phi(x_0) - L D_d(x_0, x_1)$. One comparison of the two displacements at the scale $s = L/\alpha_1$, through definition 1 with $\rho_{d^*}(\Omega_0) \leq \rho_0$, gives $\alpha_1 D_d(z_0, z_1) \leq \rho_0^2 L^2 \alpha_1^{-1} D_d(y_0, x_1)$. The exit condition $\mathcal{L}_1(z_1) \geq \phi(x_1)$ follows whenever $\alpha_1 \geq \rho_0^2 L$, and the choice in the statement gives $\alpha_1 = M_0 t_0^2 = M_0 \geq \rho_0^4 L \geq \rho_0^2 L$ since $\rho_0 \geq 1$. The first iteration therefore passes the exit condition at every acceleration parameter $M_0 \geq \rho_0^4 L$. □

Theorem 3 (Accelerated rate under bounded distortion)**.** *Under As. 1, assume d^* is twice differentiable, and let ρ_k be as in definition 3. If there exist $K \geq 0$ and ρ such that the inner loop of algorithm 2 exits at every iteration $k < K$ and $\rho_k \leq \rho$ for all $k \geq K$, then it exits at every iteration, and there is a finite $\bar{K} \geq K$ such that, for all $k \geq \bar{K}$, we have $M_k \leq 4\rho^2 L$ ($\mu = 0$), respectively $M_k \leq 4\rho^4 L$ ($\mu > 0$). By Theorem 2, this implies $\phi_k^{\text{best}} - \phi_\star \leq \eta_k^{-1} D_d(x_\star, x_0)$, with*

$$\eta_k^{-1} \leq \frac{\eta_{\bar{K}}^{-1}}{\big(1 + \frac{k - \bar{K}}{4\rho\sqrt{L\eta_{\bar{K}}}}\big)^2} \ (\mu = 0), \qquad \eta_k^{-1} \leq \eta_{\bar{K}}^{-1}\Big(1 - \sqrt{\tfrac{\mu}{4\rho^4 L}}\Big)^{k - \bar{K}} \ (\mu > 0).$$

Proof. We prove the convex case; the strongly convex case is identical, replacing $\rho^2 L$ by $\rho^4 L$.

For every $k \geq K$, proposition 5 with $\rho_k \leq \rho$ and $L_{\text{cur}} \leq 2L$ implies that any trial value $M \geq 2\rho^2 L$ is sufficient for acceptance. Since each trial doubles M_k, the inner loop exits at every $k \geq K$; with the hypothesis for $k < K$, it exits at every iteration, so Theorem 2 applies. By the update rule in algorithm 2, the search starts from a reduced value of the previously accepted M_{k-1} and then doubles until acceptance. Hence, if the search starts below $2\rho^2 L$, the accepted value is at most $4\rho^2 L$. If it starts above $2\rho^2 L$, the first trial is accepted, and the next iteration again starts from a fixed reduction

of that accepted value. Thus any overestimate inherited from earlier iterations decreases geometrically until it falls below $4\rho^2 L$. Therefore there exists $\bar{K} \geq K$ such that

$$M_k \leq 4\rho^2 L \qquad \forall k \geq \bar{K}.$$

Applying Theorem 2 from iteration $\bar{K}$ with the uniform bound $M_k \leq 4\rho^2 L$ gives

$$\eta_k^{-1} \leq \frac{\eta_{\bar{K}}^{-1}}{\left(1 + \frac{k-\bar{K}}{2\sqrt{4\rho^2 L \eta_{\bar{K}}}}\right)^2} = \frac{\eta_{\bar{K}}^{-1}}{\left(1 + \frac{k-\bar{K}}{4\rho\sqrt{L\eta_{\bar{K}}}}\right)^2}.$$

The strongly convex case follows similarly from $M_k \leq 4\rho^4 L$, yielding

$$\eta_k^{-1} \leq \eta_{\bar{K}}^{-1} \left(1 - \sqrt{\frac{\mu}{4\rho^4 L}}\right)^{k-\bar{K}}.$$

□

Corollary 1 (Iteration complexity)**.** *Under Theorem 3, $\phi_k^{\text{best}} - \phi_\star \leq \varepsilon$ after N iterations from $\bar{K}$:*

$$N = O\left(\rho\sqrt{\frac{L\, D_d(x_\star, x_0)}{\varepsilon}}\right) \quad (\mu = 0), \qquad N = O\left(\rho^2 \sqrt{\frac{L}{\mu}} \log \frac{\eta_{\bar{K}}^{-1} D_d(x_\star, x_0)}{\varepsilon}\right) \quad (\mu > 0).$$

Proof. Write $D := D_d(x_\star, x_0)$ and let $k \geq \bar{K}$.

Case $\mu = 0$. By Theorem 3 it suffices that

$$\frac{\eta_{\bar{K}}^{-1} D}{(1+s)^2} \leq \varepsilon, \qquad s := \frac{k-\bar{K}}{4\rho\sqrt{L\eta_{\bar{K}}}},$$

that is $1 + s \geq \sqrt{\eta_{\bar{K}}^{-1} D/\varepsilon}$. Dropping the 1 on the left and substituting s gives the sufficient condition

$$k - \bar{K} \;\geq\; 4\rho\sqrt{L\eta_{\bar{K}}}\sqrt{\frac{\eta_{\bar{K}}^{-1} D}{\varepsilon}} \;=\; 4\rho\sqrt{\frac{LD}{\varepsilon}},$$

in which $\eta_{\bar{K}}$ cancels.

Case $\mu > 0$. Set $q := \sqrt{\mu/(4\rho^4 L)} \in (0,1)$. By Theorem 3 it suffices that $\eta_{\bar{K}}^{-1} D(1-q)^{k-\bar{K}} \leq \varepsilon$, i.e.

$$(k - \bar{K}) \log \frac{1}{1-q} \;\geq\; \log \frac{\eta_{\bar{K}}^{-1} D}{\varepsilon}.$$

Since $\log \frac{1}{1-q} \geq q$, the condition $k - \bar{K} \geq q^{-1} \log\!\left(\eta_{\bar{K}}^{-1} D/\varepsilon\right)$ is sufficient, and $q^{-1} = 2\rho^2\sqrt{L/\mu}$.

□

C.4 Oracle complexity

Backtracking uses a variable number of oracle calls per accepted iteration. Let L denote the relative smoothness constant, let L_0 and M_{-1} be the initial estimates, and let

$$M_{\max} := \max_{0 \leq i < K} M_i$$

be the largest accepted value in the first K outer iterations. Each iteration reduces the current estimates before the searches double them. This gives constant amortized overhead, with logarithmic terms for underestimated initial values. Writing $(a)_+ := \max\{a, 0\}$, the total number of trials satisfies

$$N_{\text{trials}} \leq 4K + \left\lceil \log_2\left(\frac{M_{\max}}{M_{-1}}\right)_+ \right\rceil + \left\lceil \log_2\left(\frac{L}{L_0}\right)_+ \right\rceil.$$

This is the standard backtracking tradeoff (Nesterov & Polyak, 2006): the additional trials allow the method to use local estimates instead of a fixed worst-case constant.

D EUCLIDEAN SPECIALIZATION AND COMPARISON WITH NESTEROV

We show that, in the Euclidean geometry, ABrA-GD reduces to the standard estimate-sequence form of Nesterov acceleration. Let

$$d(x) = \frac{1}{2}\|x\|^2, \qquad D_d(x,y) = \frac{1}{2}\|x-y\|^2, \qquad \nabla d(x) = x.$$

Assume for this comparison that the line searches accept the sharp constants, namely the smoothness estimate is L and the accepted acceleration parameter is $M_k = L$. Then the BLDF is equal to one, so no geometric distortion is introduced by the Bregman geometry.

One difference from the ordinary trajectory remains. Algorithm 2 passes x_k^{best}, not x_k, into the next step, so the identities below are the classical estimate-sequence identities with x_k read as x_k^{best}. The correspondence is therefore between the updates. The two trajectories coincide whenever the gradient step never increases the objective, and ABrA-GD is otherwise a monotone variant of the classical method.

Convex case. When $\mu = 0$, one has $\tau_k = t_k$ and $\eta_{k+1}^{-1} = Lt_k^2$. Writing $g_k = \nabla\phi(y_k)$, the ABrA-GD updates become

$$y_k = (1-t_k)x_k + t_k z_k, \qquad x_{k+1} = y_k - \frac{1}{L}g_k, \qquad z_{k+1} = z_k - \frac{1}{Lt_k}g_k.$$

Therefore

$$x_{k+1} = (1-t_k)x_k + t_k z_{k+1},$$

which is exactly the usual accelerated estimate-sequence coupling. Moreover, from

$$\eta_{k+1}^{-1} = Lt_k^2, \qquad \eta_{k+1}^{-1} = (1-t_k)\eta_k^{-1},$$

we recover the standard Nesterov/FISTA scalar recursion

$$t_k^2 = (1-t_k)t_{k-1}^2.$$

Theorem 2 gives

$$\phi(x_k^{\mathrm{best}}) - \phi_\star \le \eta_k^{-1} D_d(x_\star, x_0) \le \frac{4L}{k^2} D_d(x_\star, x_0) = \frac{2L}{k^2}\|x_\star - x_0\|^2,$$

which matches the classical accelerated $O(L\|x_\star - x_0\|^2/k^2)$ rate up to the standard indexing convention.

Relatively strongly convex case. Let $\mu > 0$ and set

$$\alpha_k := \mu + \eta_k^{-1}.$$

With $M_k = L$, the parameters satisfy

$$\alpha_{k+1} = Lt_k^2, \qquad \alpha_{k+1} = (1-t_k)\alpha_k + t_k\mu,$$

and

$$\tau_k = \frac{t_k - \mu/L}{1 - \mu/L}.$$

Define the Euclidean mirror point

$$q_{k+1} := y_k - \frac{1}{\mu}\nabla\phi(y_k).$$

The dual/primal-averaging update can be written as

$$z_{k+1} = \left(1 - \frac{t_k\mu}{\alpha_{k+1}}\right) z_k + \frac{t_k\mu}{\alpha_{k+1}} q_{k+1}.$$

The specific choice of τ_k gives again

$$x_{k+1} = (1-t_k)x_k + t_k z_{k+1}.$$

Thus the Euclidean strongly convex specialization is the standard Nesterov estimate-sequence update written in the primal–dual variables. The generic rate becomes

$$\phi(x_k^{\mathrm{best}}) - \phi_\star \le \eta_k^{-1} D_d(x_\star, x_0), \qquad \eta_k^{-1} \le \eta_K^{-1} \left(1 - \sqrt{\frac{\mu}{L}}\right)^{k-K},$$

once the accepted parameter is $M_k = L$. This recovers the classical accelerated linear contraction $1 - \sqrt{\mu/L}$, up to constants and indexing. If the line search accepts values $M_k > L$, the updates no longer coincide with Nesterov's, and the rate is the one Theorem 2 states in terms of the accepted M_k.

E Comparison with ABPG and triangle scaling

E.1 Comparison with ABPG

ABPG controls acceleration through a triangle-scaling exponent or gain. The adaptive variants of Hanzely et al. (2021) search over the exponent or gain, while Savchuk et al. (2024) fix the exponent and backtrack the relative smoothness constant. ABrA-GD backtracks both a smoothness parameter and an acceleration parameter. Its search over M_k tests the primal–dual bound directly and responds to both smoothness and local geometric distortion.

Both approaches give bounds in terms of quantities accepted during the run (table 1). For ABrA-GD, the same framework also gives a conditional accelerated linear rate when $\mu > 0$. The cited TSP-based analyses do not establish such a rate. Hanzely et al. (2021) report faster convergence with restarted ABPG on a D-optimal design instance described as relatively strongly convex, but give no accelerated linear guarantee for that experiment.

For gain-adaptive ABPG at $\gamma = 2$, the rate is $O(\bar{G}_k/k^2)$, where $\bar{G}_k$ depends on the accepted gains. A uniform bound on the gain can be obtained from a Hessian ratio over the full region $\mathcal{X}$ (Hanzely et al., 2021, §2.2, eq. (16)). ABrA-GD uses the corresponding local quantity $\rho_{d^*}(\Omega_k)$ to bound its dual divergence comparisons. For twice-differentiable geometries, proposition 1 relates these quantities, and table 2 gives formulas for standard kernels.

E.2 Relation to the triangle-scaling property

A uniform triangle-scaling gain at exponent 2 bounds the dual norm distortion. Let $\mathcal{X}$ be convex with d twice differentiable and $\nabla^2 d \succ 0$, let $\Omega = \nabla d(\mathcal{X})$, and suppose the triangle-scaling property holds on $\mathcal{X}$ with exponent $\gamma = 2$ and a gain G *uniform over triples*. Writing both divergences in integral form along the common direction, letting $\theta \to 0$ and shrinking the pair to a point q gives

$$G \;\geq\; \sup_{x,q\in\mathcal{X}} \lambda_{\max}\big(\nabla^2 d(q)^{-1}\nabla^2 d(x)\big) \;=\; \rho_{d^*}(\Omega)^2,$$

using $\nabla^2 d^*(\nabla d(x)) = \nabla^2 d(x)^{-1}$. Thus $\rho_{d^*}^2 \leq G$ on the corresponding dual region. This Hessian ratio appears in the triangle-scaling bound of Hanzely et al. (2021, §2.2, eq. (16)) and in the Hessian stability condition of Karimireddy et al. (2018). Here, it directly controls the local divergence comparisons, including those needed when $\mu > 0$.

The implication requires a gain uniform over triples. It does not concern the intrinsic triangle-scaling exponent defined by a pointwise limit and used by adaptive variants in practice.

For a jointly convex Bregman divergence, $\gamma = 1$ is a uniform triangle-scaling exponent at gain 1, giving an $O(1/k)$ rate. The entropy kernel has this property. For $d = \frac{1}{p}\|x\|^p$ with $p > 2$, the global exponent can also be below 1. The kernel-specific constructions of Hanzely et al. (2021) modify the reference function and restrict the domain. The formulas in table 2 express the corresponding local distortion through coordinate ratios: approaching the origin can make the distortion unbounded, while bounded ratios give finite distortion.

The Burg kernel. The Burg kernel $d(x) = -\sum_i \log x_i$ is not jointly convex, and Hanzely et al. (2021) show that it has no uniform exponent $\gamma > 1/2$ at gain 1. The global guarantee of ABPG for this kernel is therefore at best $O(k^{-1/2})$, which is slower than the $O(1/k)$ rate of Bregman gradient descent. By table 2, the dual distortion factor of the Burg kernel is

$$\rho_{d^*}(\Omega) = \sup_{x_1,x_2\in\mathcal{X}} \max_i \frac{(x_2)_i}{(x_1)_i}, \qquad \Omega = \nabla d(\mathcal{X}),$$

which is finite whenever the coordinate ratios are bounded on $\mathcal{X}$. For example, let $\mathcal{X}_{a,b} = \{x : a \leq x_i \leq b \text{ for all } i\}$ with $0 < a < b$. Its image $\nabla d(\mathcal{X}_{a,b})$ is a box, hence convex, and $\rho_{d^*}(\nabla d(\mathcal{X}_{a,b})) = b/a$. If, from some iteration on, the points that define the dual segments of definition 3 lie in $\mathcal{X}_{a,b}$, then $\rho_k \leq b/a$. Theorem 3 then gives an $O(1/k^2)$ rate, and corollary 1 gives $O\big((b/a)\sqrt{L\,D_d(x_\star, x_0)/\varepsilon}\big)$ iterations when $\mu = 0$ and an accelerated linear rate when $\mu > 0$. For the Burg kernel, the distortion factor thus gives an accelerated rate where the uniform triangle-scaling exponent gives a rate slower than $O(1/k)$.

The comparison with a uniform gain at exponent 2 is closer. On $\mathcal{X}_{a,b}$, the display above gives $G \geq (b/a)^2$, so the $O(G/k^2)$ bound of a uniform gain is not smaller than the $O(\rho^2/k^2)$ bound of Theorem 3, up to absolute constants. The distortion factor ρ_k is measured on the region of iteration k only, so it can be smaller than b/a when the iterates concentrate. The guarantee of Theorem 3 is conditional: it requires the bound on ρ_k to hold along the iterates, and the paper does not prove this bound a priori.

F Composite objectives and the subgradient envelope

The implicit envelope viewpoint does not require differentiability of the whole objective. Consider the composite problem

$$\Phi_\star = \min_{x \in \mathcal{X}} \Phi(x) := f(x) + r(x) + \mu D_d(x, x_0),$$

where f is differentiable and relatively smooth, while r is a proper closed convex term for which a Bregman proximal step is available. For $\eta > 0$, define

$$\Phi_\star^\eta := \min_x \left\{ f(x) + r(x) + \left(\mu + \eta^{-1}\right) D_d(x, x_0) \right\}.$$

If an algorithm produces iterates satisfying the requirement

$$\Phi(x_k) \leq \Phi_\star^\eta,$$

then the same envelope argument gives

$$\Phi(x_k) - \Phi_\star \leq \eta^{-1} D_d(x_\star, x_0), \qquad x_\star \in \arg\min_x \Phi(x),$$

because

$$\Phi_\star^\eta \leq \Phi^\eta(x_\star) = \Phi_\star + \eta^{-1} D_d(x_\star, x_0).$$

Thus the envelope descent requirement survives unchanged in the composite setting.

A corresponding lower model can be written either with a primal nonsmooth term or with a subgradient. Dualizing only the smooth part gives

$$S_r^\eta(x, \lambda) := \langle \lambda, x \rangle - f^*(\lambda) + r(x) + \left(\mu + \eta^{-1}\right) D_d(x, x_0) \leq \Phi^\eta(x),$$

and therefore

$$\mathcal{D}_r^\eta(\lambda) := \min_x S_r^\eta(x, \lambda) \leq \Phi_\star^\eta.$$

Equivalently, for any $s_y \in \partial r(y)$, convexity of r gives the computable subgradient lower model

$$\begin{aligned} \Phi_y^{\mathrm{low}}(z) &:= \Phi(y) + \left\langle \nabla f(y) + s_y + \mu\big(\nabla d(y) - \nabla d(x_0)\big), z - y \right\rangle + \mu D_d(z, y) \\ &\leq \Phi(z). \end{aligned}$$

The proximal primal step also remains well defined through

$$x^+ \in \arg\min_x \left\{ \langle \nabla f(y), x - y \rangle + r(x) + L D_d(x, y) + \mu D_d(x, x_0) \right\}.$$

This shows that the envelope and lower-model mechanisms extend naturally to nonsmooth composite terms. What does not follow automatically is the accelerated BLDF proof in the main text: the cancellation there uses a single continuous gradient-generated dual direction, whereas the composite model introduces a set-valued subgradient or a nonsmooth proximal residual. A full accelerated composite analysis would therefore require additional control of this residual.

G Numerical Experiments

G.1 Experimental protocol

The code is the toolbox of Hanzely et al. (2021)[1] with an implementation of ABrA-GD added as one more method; the feasible domain and its boundary handling, the datasets, the baseline methods, and the initialization are theirs, unchanged. The objectives, reference functions and initial points below are stated from that source.

D-optimal design. For $H \in \mathbb{R}^{m\times n}$ with $m < n$,

$$f(x) = -\log\det\left(H\,\mathrm{diag}(x)H^\top\right), \qquad \mathcal{X} = \left\{x \geq 0 :\ \textstyle\sum_i x^{(i)} = 1\right\},$$

and d is the Burg entropy $d(x) = -\sum_i \log x^{(i)}$ restricted to that simplex. Here f is 1-smooth relative to d (Lu et al., 2018), so $L = 1$. The simplex constraint is the affine equality of appendix A.2. The restriction makes $\mathrm{dom}\, d$ bounded, so $\mathrm{dom}\, d^* = \mathbb{R}^n$ and the virtual point q_{k+1} of lemma 9 exists at every iterate on these instances. Every method starts at the simplex center $x_0 = (1/n)\mathbf{1}$. The instances are built from the LIBSVM regression datasets abalone ($n = 4177$, $m = 8$), bodyfat (252, 14), mpg (392, 7) and housing (506, 13) (Chang & Lin, 2011).

The relatively strongly convex instances are formed by the anchor term already in (1): $\phi(x) = f(x) + \mu D_d(x, x_0)$, with the same f, d, and x_0 as the convex instance and $\mu = 10^{-4}$.

Poisson linear inverse problem. $f(x) = D_{\mathrm{KL}}(b, Ax)$ on $\mathcal{X} = \mathbb{R}^n_+$, with $A \in \mathbb{R}^{m\times n}$ drawn entrywise uniform on $[0, 1]$ and column-normalized, and $b = A\bar{x} +$ noise for a sparse nonnegative $\bar{x}$. The reference function is again the Burg entropy, now on the orthant, and f is $\|b\|_1$-smooth relative to it (Bauschke et al., 2017).

Parameters. ABrA-GD starts every run at $L_0 = 1$ and $M_{-1} = L_0$ on both problems. This differs from Hanzely et al. (2021), who supply the theoretical constant: $L = 1$ for D-optimal design and $L = \|b\|_1$ for the Poisson problem. The baselines are run through their own interface with the constants that interface expects, as in Hanzely et al. (2021): BPG and ABPG receive L, and BPG-LS, ABPG-e and ABPG-g adjust a gain online starting from it. No relative strong convexity constant is supplied to any baseline, because that interface takes no μ; on the $\mu > 0$ instances the anchor term enters through ϕ itself.

G.2 Safeguards and ablations

Restart strategies. As in adaptive ABPG (Hanzely et al., 2021), we test two standard restart heuristics as ablations. The first is a function-value restart, triggered when the candidate objective value increases. The second is a gradient restart, triggered when the accelerated direction becomes misaligned with the current descent direction. After a restart, we keep the current best primal point and reset the accelerated state by setting

$$z_k = x_k = x_k^{\mathrm{best}},$$

resetting the dual averaging state, and continuing with the same adaptive searches for L_{cur} and M_k.

These restart rules are heuristic and are not part of ABrA-GD's convergence analysis. They are included only to test whether explicit momentum resets add anything beyond the exit condition of algorithm 2. The experiments below show that restarts do not provide a consistent advantage; in several cases they increase the variability of the accepted acceleration parameter M_k or reduce oracle-call efficiency.

Gradient-descent safeguard. The accelerated step may become inefficient when the accepted acceleration parameter M_k is very large. To make the method never worse than Bregman gradient descent, one can compare the certificate progress of the accelerated step with the nonaccelerated dual-averaging update. This safeguard uses the true relative smoothness constant L only as a theoretical

[1] https://github.com/linxiaolx/accbpg

comparison. Define

$$t_k^{\mathrm{gd}} = \frac{\mu + \eta_k^{-1}}{L + \eta_k^{-1}}, \qquad \bar{\eta}_{k+1}^{-1} = (1 - t_k^{\mathrm{gd}})\eta_k^{-1}.$$

After the accelerated line search has produced a candidate

$$\eta_{k+1}^{-1} = M_k t_k^2 - \mu,$$

we accept it only if

$$\eta_{k+1}^{-1} \le \bar{\eta}_{k+1}^{-1}.$$

Otherwise, we perform the certified nonaccelerated primal–dual step obtained by setting

$$y_k = z_k, \qquad t_k = t_k^{\mathrm{gd}}, \qquad \eta_{k+1}^{-1} = \bar{\eta}_{k+1}^{-1}.$$

This fallback is exactly a Bregman gradient step in primal variables, but written in the primal–dual variables so that the envelope certificate is preserved. Consequently,

$$\phi(x_k^{\mathrm{best}}) - \phi^\star \le \eta_{k,\mathrm{gd}}^{-1} D_d(x^\star, x_0),$$

so the safeguarded method has at least the convergence rate of Bregman gradient descent while retaining accelerated steps whenever their certificate progress is better.

In all reported experiments, the safeguard was never triggered: every accepted ABrA-GD step already produced at least as much certificate progress as the nonaccelerated update. Thus, the safeguard is included only to make the worst-case comparison with Bregman gradient descent explicit and does not affect the numerical results.

G.3 Additional numerical experiments

We report additional experiments on D-optimal design and Poisson linear inverse problems. For D-optimal design, we test both the convex case and a relatively strongly convex variant with $\mu = 10^{-4}$. For the Poisson inverse problem, we test two variants, denoted Poisson L1 and Poisson L2.

Across the convex D-optimal design instances, ABrA-GD improves on nonaccelerated Bregman gradient methods and matches the best ABPG variants over the plotted range. In the relatively strongly convex setting, ABrA-GD's log-residual decreases approximately linearly over the plotted range and reaches a lower residual than the competing methods. On the Poisson problems, ABrA-GD again improves on the nonaccelerated baselines, but which method reaches the lowest residual depends more strongly on the instance and the variant. The restart comparisons below show that explicit restarts do not improve on the plain method in these instances.

G.3.1 D-OPTIMAL DESIGN, CONVEX, $\gamma = 1$ FOR ABPG

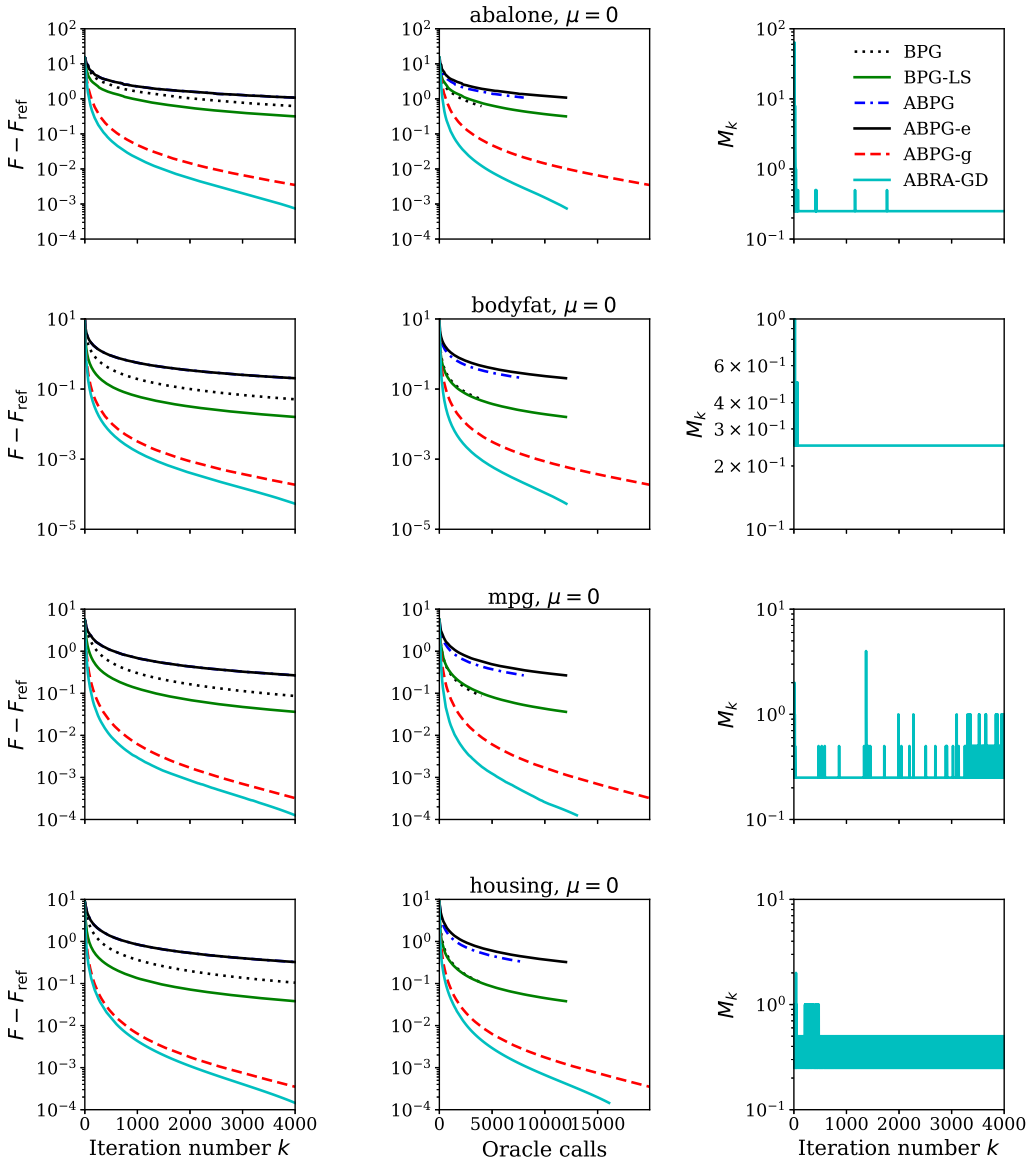

Figure 5: D-optimal design on the housing dataset with ABPG run at the triangle-scaling exponent $\gamma = 1$, one point in Hanzely et al. (2021)'s own sweep over $\gamma \in \{1.0, 1.5, 2.0, 2.2\}$; they report that at $\gamma = 1$ ABPG converges at $O(k^{-1})$ but is slower than BPG. They fix $\gamma = 2$ for every ABPG variant in their own comparisons, which is why we use $\gamma = 2$ in all other comparisons.

G.3.2 D-OPTIMAL DESIGN, CONVEX

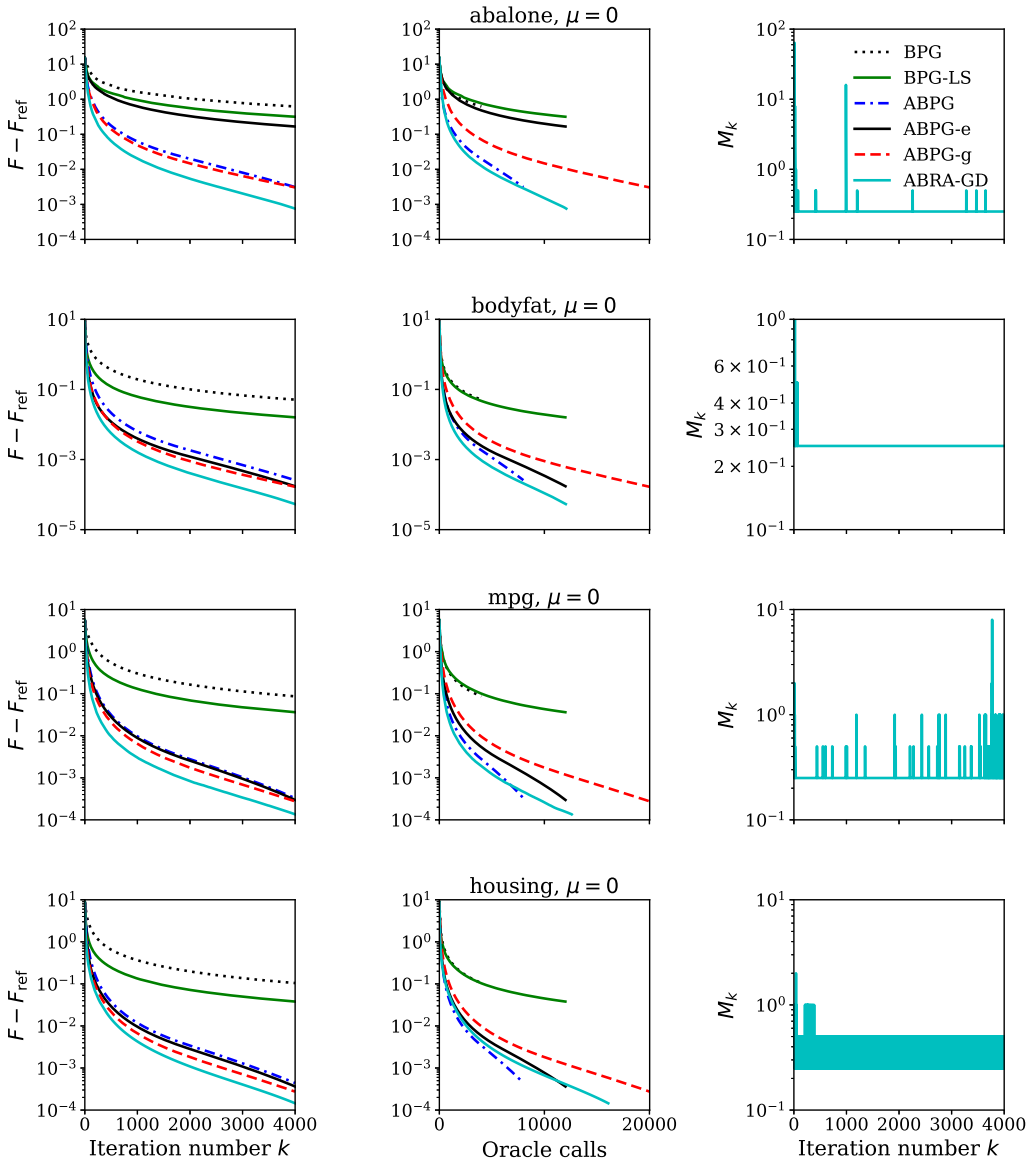

Figure 6: Convex D-optimal design experiments on the abalone, bodyfat, mpg, and housing datasets. For each dataset, we report the residual $F - F_\star$ versus iterations, the same residual versus oracle calls, and the accepted values of M_k for ABrA-GD. ABrA-GD improves over BPG and BPG-LS on all datasets and matches the best accelerated ABPG variants over the plotted range, while adapting M_k locally during the run.

G.3.3 D-OPTIMAL DESIGN, STRONGLY CONVEX

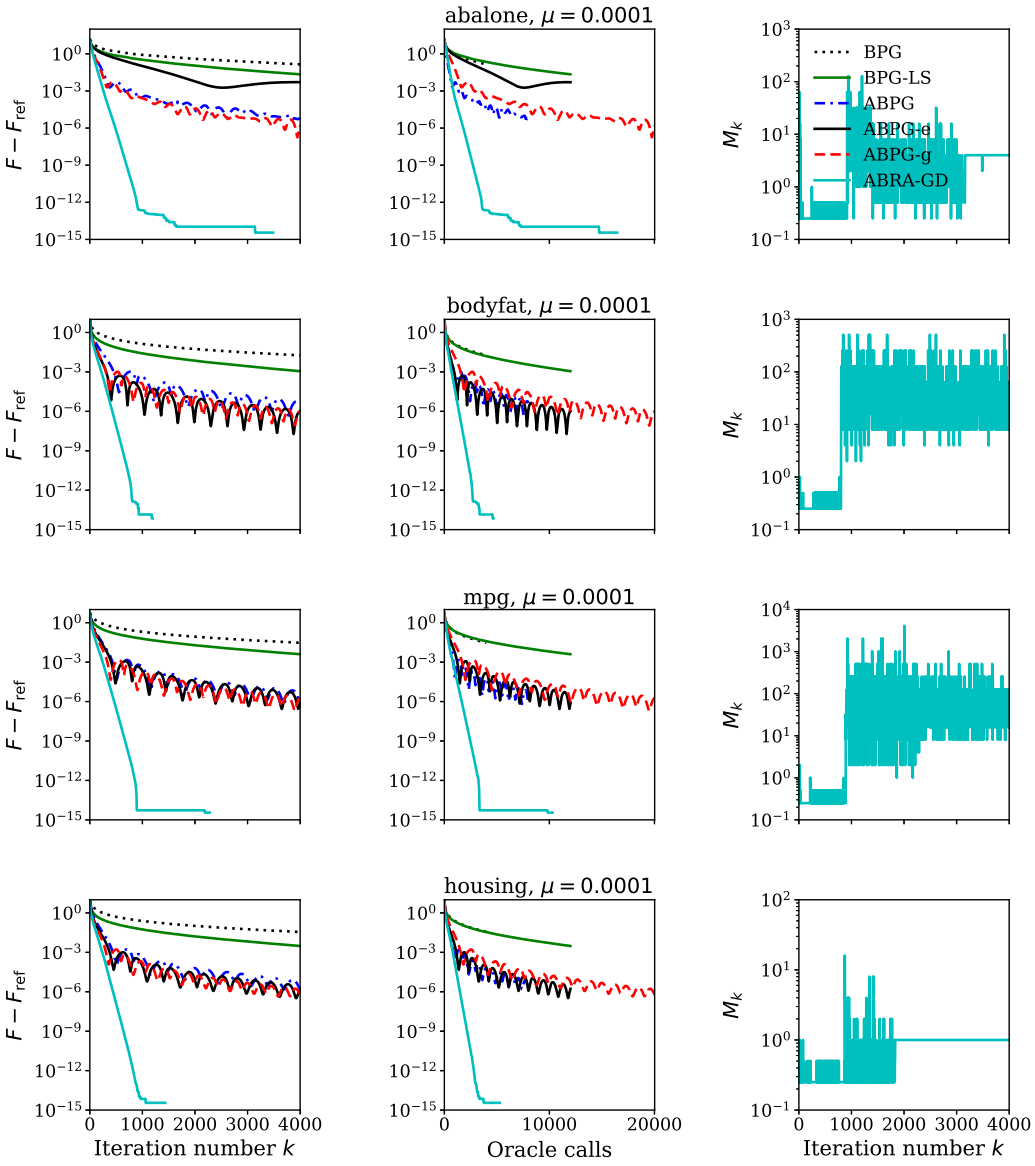

Figure 7: Relatively strongly convex D-optimal design experiments with $\mu = 10^{-4}$. The left and middle columns show convergence in iterations and oracle calls, while the right column reports the accepted values of M_k. In this regime, ABrA-GD's log-residual decreases approximately linearly over the plotted range and reaches a lower residual than BPG, BPG-LS, and the ABPG variants. The accepted values of M_k can fluctuate; ϕ_k^{best} is non-increasing by construction (algorithm 2) regardless.

G.3.4 POISSON PROBLEM

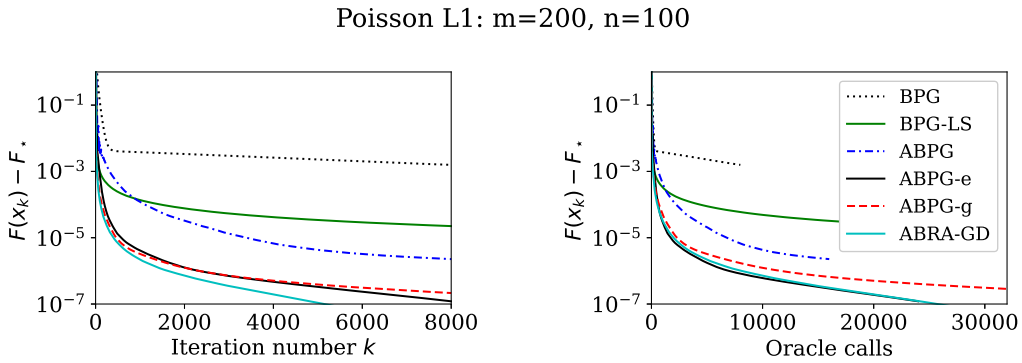

Figure 8: Poisson linear inverse problem, L1 variant $(m = 200, n = 100)$. ABrA-GD improves on the nonaccelerated Bregman methods and matches the best ABPG variants over the plotted range, in both iteration count and oracle-call count.

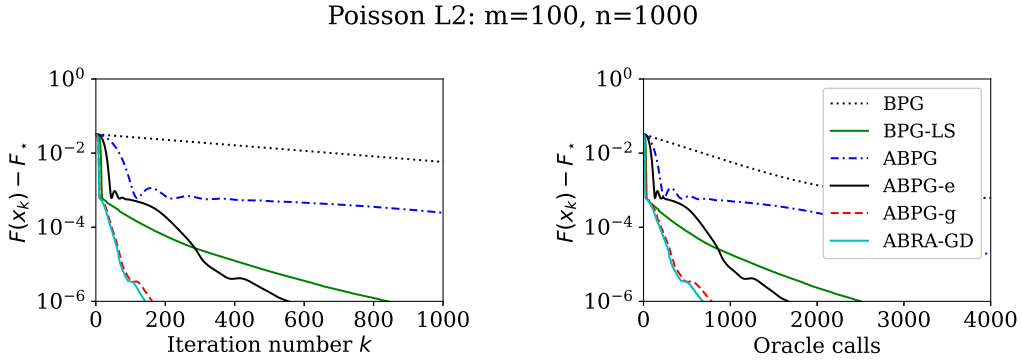

Figure 9: Poisson linear inverse problem, L2 variant $(m = 100, n = 1000)$. The accelerated methods reach a lower residual than BPG and BPG-LS. ABrA-GD tracks the best accelerated baseline over the plotted range; which method reaches the lowest final residual is more sensitive to the instance and the adaptive strategy than in the D-optimal design experiments.

G.3.5 RESTART COMPARISON: D-OPTIMAL DESIGN, CONVEX

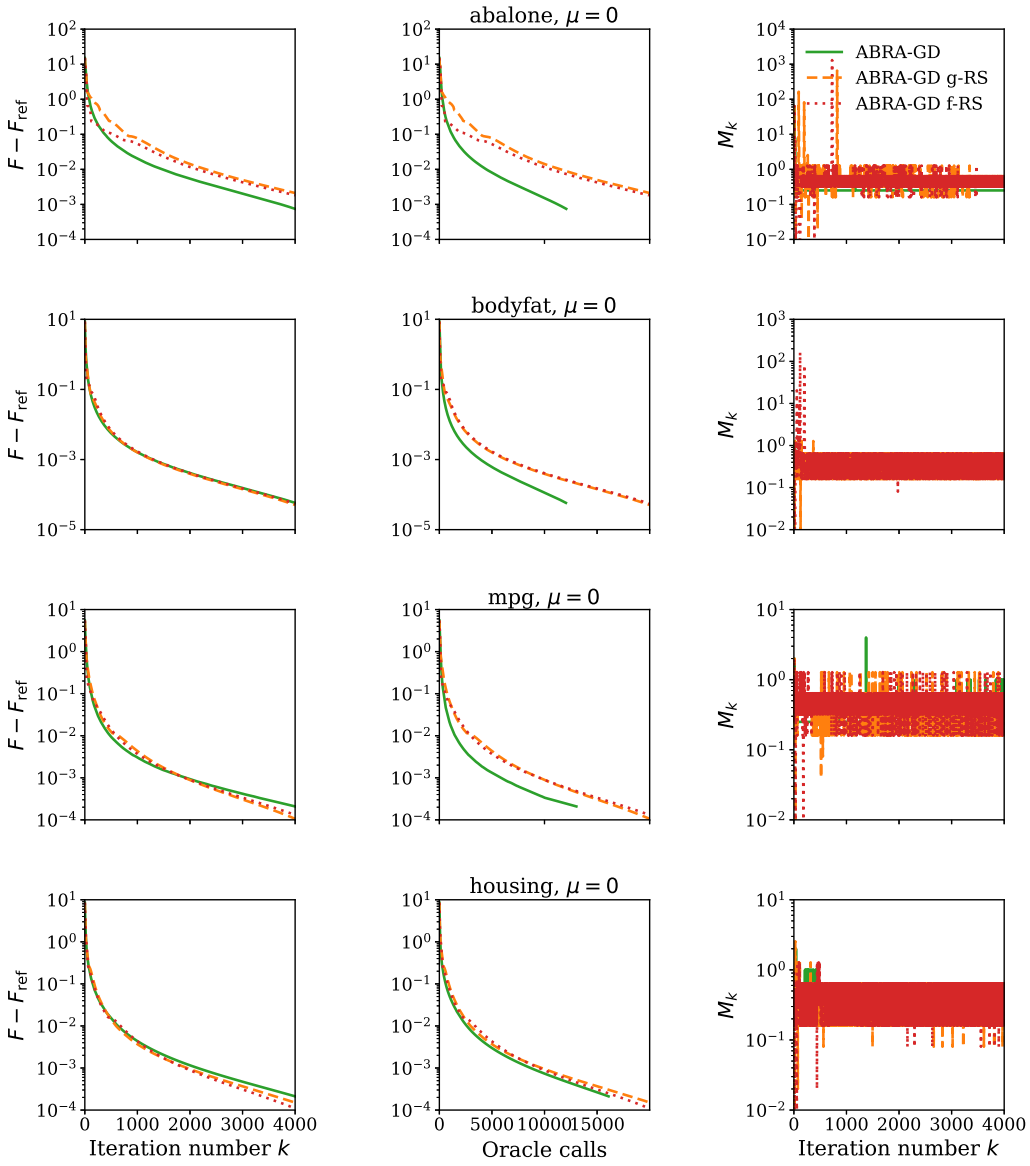

Figure 10: Restart comparison for convex D-optimal design. We compare ABrA-GD without restart, with gradient restart (g-RS), and with function-value restart (f-RS). Restarts do not improve performance consistently: the plain ABrA-GD variant is usually comparable or better in oracle calls, while the restarted variants can introduce additional variability in M_k.

G.3.6 RESTART COMPARISON: D-OPTIMAL DESIGN, STRONGLY CONVEX

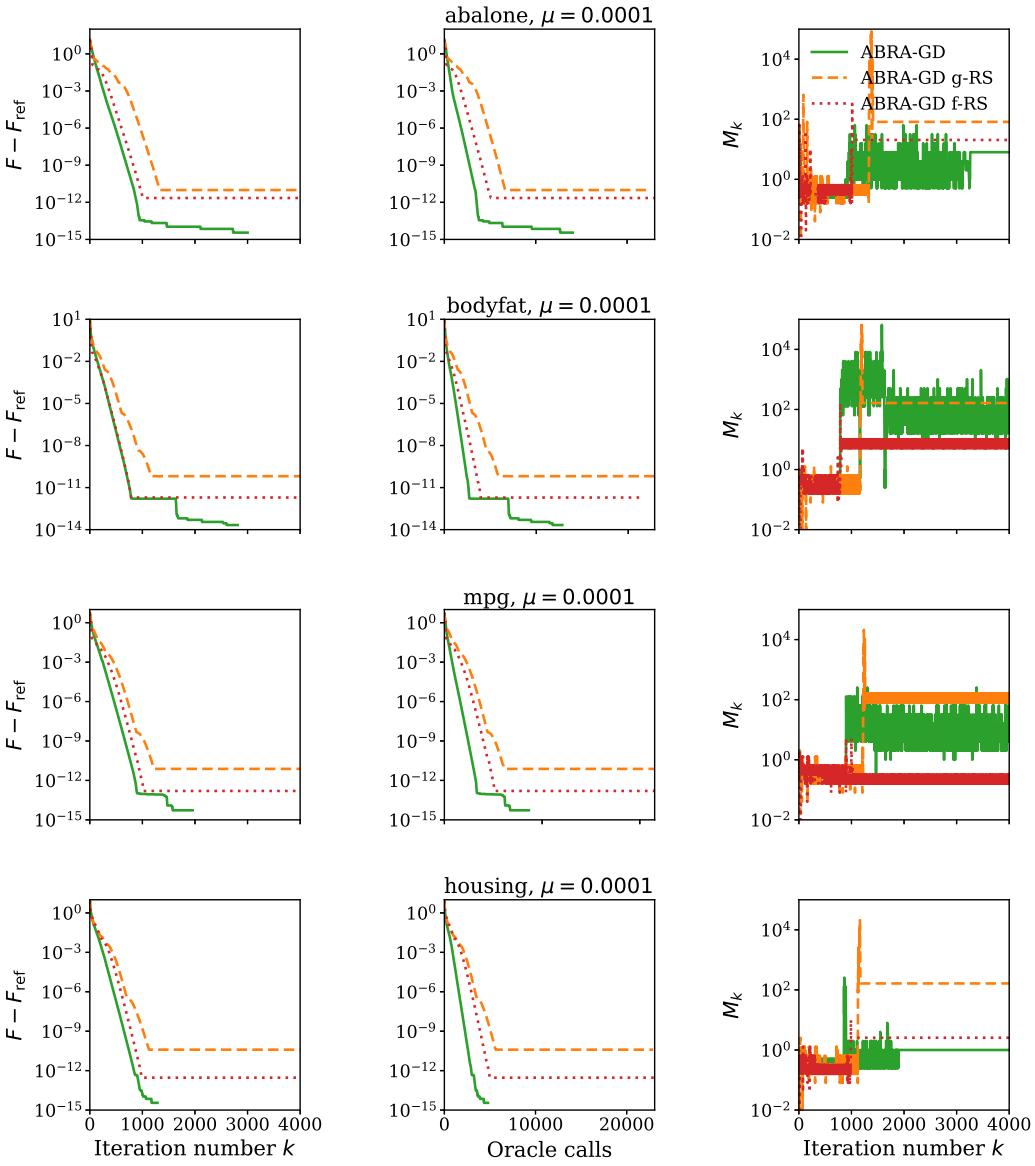

Figure 11: Restart comparison for relatively strongly convex D-optimal design with $\mu = 10^{-4}$. The non-restarted ABrA-GD variant typically reaches the lowest final residual across the datasets. The restarted variants often plateau earlier or require larger accepted values of M_k.

G.3.7 RESTART COMPARISON: POISSON PROBLEM

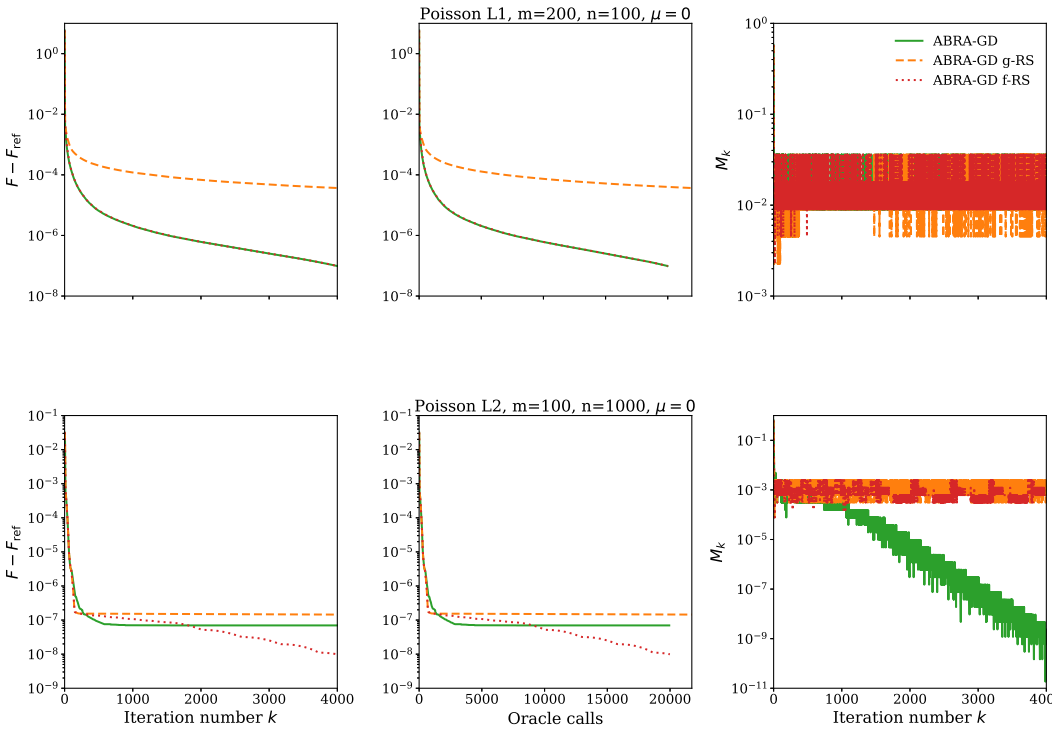

Figure 12: Restart comparison for the Poisson linear inverse problems. Restart is neutral or worse on both instances: the plain ABrA-GD update is not improved upon by either restarted variant on either instance. The restarted variants both stall on the Poisson L2 instance.